DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

AIBC 2024

17th Annual International Biocuration Conference India

Enabeling the path to data to knowledge

Jointly organized by:
Indian Biological Data Centre of Department of Biotechnology, GoI housed at RCB, NCR Biotech cluster, Faridabad
&
University of Delhi, South Campus, New Delhi

MARCH 5-8, 2024

CONFERENCE BOOKLET

United Nations Educational, Scientific and Cultural Organization

Regional Centre for Biotechnology

ISB International Society for Biocuration

Welcome Note

Dear Esteemed Participants,

It is with great pleasure and excitement that we extend our warmest greetings to you for the 17th Annual International Biocuration Conference (AIBC-2024), proudly hosted for the first time in India. Organized jointly by the Indian Biological Data Centre (IBDC) of Department of Biotechnology, GoI housed at RCB, NCR Biotech cluster, Faridabad and the Department of Plant Molecular Biology, University of Delhi South Campus, this year's conference promises to be a landmark event in the field of biocuration. Mentored by the esteemed International Society of Biocuration (ISB), AIBC-2024 offers a unique platform for curators, developers, and data stakeholders from across the globe to converge, interact, and exchange cutting-edge ideas and expertise. In an era where data science permeates every facet of life science research, the evolution of the data curation community towards inclusivity and diversity is not just desirable but imperative. With its distinctive 'flavour', this year's conference is particularly special. As India's incredibly diverse biodiversity - encompassing flora, fauna, and even ethnic human populations - continues to yield an abundance of data, the synergy between the global biocuration community and national stakeholders promises exciting new avenues of collaboration and discovery.

We cordially invite participants from academia, government, and industry to join us in this extraordinary gathering. Your presence will enrich the vibrant exchange of knowledge and ideas that defines the essence of the Annual International Biocuration Conference. Together, let us chart a course towards a future where biocuration not only meets the demands of contemporary research but also embraces the richness of global diversity.

With enthusiasm and anticipation
17th Annual International Biocuration Conference (AIBC-2024)
Organizing Committee

AIBC 2024 PATRONS

Dr. Rajesh S. Gokhale
Secretary
Department of Biotechnology
(DBT)

Prof. Yogesh Singh
Vice Chancellor
University of Delhi

Prof. Arvind K. Sahu
Executive Director,
Regional Centre for
Biotechnology
(RCB)

Dr. Ramesh V. Sonti,
Director,
The International
Centre for Genetic
Engineering and
Biotechnology (ICGEB)

Dr. Debasisa Mohanty
Director
National Institute of
Immunology
(NII)

AIBC 2024 ORGANISING COMMITTEE

Dr. Guy Cochrane
Global Biodata Coalition, EMBL-EBI, UK

Dr. Charles Tapley Hoyt Northeastern University, USA

Dr. Rajendra Joshi
Centre for Development of Advanced Computing (C-DAC), India

Dr. Chuck Cook
Global Biodata Coalition UK

Dr. Dinesh Gupta
The International Centre for Genetic Engineering and Biotechnology (ICGEB), India

Dr. Harpreet Singh
Indian Council of Medical Research (ICMR), India

Dr. Rama Balakrishnan,
Genentech, USA

Ms. Mary Ann Tuli
GigaScience Press, Hong Kong

Dr. Saloni Mathur
National Institute of Plant Genome Research (NIPGR), India

Prof. Mukesh Jain,
Jawaharlal Nehru University (JNU), India

Prof. D Sundar,
Indian Institute of Technology Delhi (IIT), India

Prof. Deepak Nair,
Regional Centre for Biotechnology (RCB), India
(Co-Convener)

Prof. Saurabh Raghuvanshi,
Department of Plant Molecular Biology, UDSC, India
(Convener)

AIBC 2024 SCIENTIFIC COMMITTEE

Dr. Sushma Naithani
Oregon State University
USA

Prof. D Sundar,
Indian Institute of Technology (IIT), India

Prof. Mukesh Jain,
Jawaharlal Nehru University (JNU), India

Prof. Saurabh Raghuvanshi,
Department of Plant Molecular Biology, UDSC, India
(Convener)

AIBC 2024 LOCAL ORGANIZING COMMITTEE

Dr. Sonia Balyan
IBDC

Dr. Shivani Sharma
IBDC

Sanjay Deshpande
IBDC

Dr. Pawan Kumar
IBDC

Vipul Adhana
IBDC

Gautam Kanwal
IBDC

Dr. Shivani Kansal
UDSC

Dr. Utkarsh Raghuvanshi
UDSC

ORGANIZING PARTNERS

सत्यमेव जयते
Department of Biotechnology
Ministry of Science and Technology
Government of India

United Nations
Educational, Scientific and
Cultural Organization

क्षेत्रीय जैव प्रौद्योगिकी केन्द्र
Regional Centre
for Biotechnology

UNIVERSITY OF DELHI
निष्ठा धृतिः सत्यम्

ISB International Society for Biocuration

ACADEMIC SPONSORS

SPONSORS

aws

LOCUZ

Hewlett Packard Enterprise

Redington

RITTAL

ddn

Netweb TECHNOLOGIES

Tyrone

premas life sciences

REPLICA BIOTECH

PROGRAM SCHEDULE

Pre-Conference Workshop

Venue: Univ. of Delhi South Campus, New Delhi

Tuesday, 5th March 2024			
Pre-conference workshop: GlyGen & CFDE Biocuration 2024			
10:00-10:15		Welcome note	Prof. Saurabh Raghuvanshi, Univ. of Delhi South Campus
10:15-11:15	Workshop	Introduction and overview of GlyGen and NIH-funded Common Fund Data Ecosystem (CFDE)	Prof. Raja Mazumder Mr. Jeet Vora Dr. Suvarna Nadendla Dr. Mike Tiemeyer Dr. Rene Ranzinger GlyGen, CFDE
11:15-11:45		Scientific Question Co-Development	
11:45-12:15		Tea break	
12:15-13:15		Hands-On Session and Q&A	
13:15-15:00	**Lunch Break**		
15:00-15:35	Lecture	EMBL-European Nucleotide Data Archive	**Dr. Guy Cochrane**, Head, European Nucleotide Archive, EMBL-EBI and Executive Director, Global Biodata Coalition
15:35-16:10	Lecture	DNA Databank of Japan	**Prof. Masanori Arita**, Head, Bioinformation & DDBJ Center, National Institute of Genetics, Japan
16:10-16:45	Lecture	NCBI Resources for Sharing of Sequence Data	**Dr. Ilene Karsch Mizrachi**, Program Head, National Institutes of Health, NCBI, USA
16:45-17:20	Lecture	The Human Glycome Atlas Project to catalogue a Reference Human Glycome	**Prof. Kiyoko F. Aoki-Kinoshita**, Institute for Glycobiology and Integrative Biosystems, Soka University, Japan

Main Conference: Day 1

Venue: Indian Biological Data Centre of Department of Biotechnology, GoI housed at RCB, NCR Biotech cluster, Faridabad

Wednesday, 6th March 2024			
8:00-8:45	**AIBC-2024, Registration and Kit Distribution**		
Session I	**International Nucleotide Sequence Database Collaboration (INSDC)**		
Chairs	**Dr. Guy Cochrane and Prof. Saurabh Raghuvanshi**		
Hall	**M.K.Bhan Auditoruim**		
9:00-9:10	Welcome Note		**Prof. Saurabh Raghuvanshi** (Convener), Univ. of Delhi South Campus
9:10-9:20	International Society for Biocuration Greetings		**Dr. Mary Ann Tuli (ISB)**
9:20-9:50	Keynote lecture	The International Nucleotide Sequence Database	**Dr. Guy Cochrane** Head of European Nucleotide Archive, EMBL-EBI
9:50-10:10	Session talk	Coping with benefit sharing from digital sequence information	**Prof. Masanori Arita** Head, Bioinformation & DDBJ Center, National Institute of Genetics, Japan
10:10-10:30	Session talk	NCBI Resources for Sharing of Sequence Data	**Dr. Ilene Karsch Mizrachi** Program Head National Institutes of Health, NCBI, USA
10:30-11:00	**Tea Break**		
11:00-12:00	**Inauguration ceremony** **Chief Guest: Dr. Jitendra Singh** Hon'ble Minister of State (IC), Ministry of Science & Technology; Minister of State, PMO; Minister of Personnel, Public Grievances and Pensions; Department of Atomic Energy; and Department of Space, Govt. of India		
12:00-12:20	Session talk	The European Nucleotide Archive: Sequence Data Resource and Submissions Brokering Network	**Ms. Zahra Waheed** Senior Bioinformatician, Data Coordination and Archiving Team, EMBL-European Bioinformatics Institute
Session II	**Global biodata resources: a crucial infrastructure underpinning life science research**		
Chair	**Dr. Chuck Cook**		
12:20-12:40	Session talk	Importance of Data Persistence (GBC session)	**Prof. Masanori Arita** Head, Bioinformation & DDBJ Center, National Institute of Genetics, Japan
12:40-14:00	Lunch Break		

14:00-14:30	Keynote lecture	The global biodata infrastructure and the drive for sustainability	**Dr. Guy Cochrane** Executive Director, Global Biodata Coalition
14:30-14:50	Session talk	Super Computing platform and Data Resources derived from Omics Data in Agriculture	**Dr. K.K. Chaturvedi** Principal Scientist and In-Charge IT Unit, ICAR-Indian Agricultural Statistics Research Institute, India
14:50-15:10	Session talk	Health Research Data Ecosystem – Challenges and Opportunities	**Dr. Harpreet Singh** Scientist and Chief Data Officer, Indian Council of Medical Research (ICMR), India
15:10-15:30	Session talk	Pillars of resource sustainability for the Human Disease Ontology Knowledgebase (DO-KB)	**Dr. Lynn M. Schriml** Associate Professor, University of Maryland, USA
15:30-15:50	Session talk	Bgee, GCBR for animal transcriptomics: sustainability and prospects	**Dr. Frederic Bastian** Associate Director, Swiss Institute of Bioinformatics (SIB), Switzerland
15:50-16:20	**Tea Break**		
Session III	**Indian Biological Data Center**		
Chairs	**Dr. Ilene Karsch Mizrachi and Dr. Saurabh Raghuvanshi**		
16:20-16:50	Keynote lecture	The Indian Biological Data Centre: Past, Present and Future	**Prof. Deepak T. Nair** Head, Indian Biological Data Centre, India
16:50-17:10	Session talk	Indian SARS-CoV-2 Genomics Consortium (INSACOG)	**Dr. Nidhan Biswas** Associate Professor, National Institute of Biomedical Genomics (NIBMG), India
17:10-17:25	Flash talk	Nucleotide Data Resources at Indian Biological Data Centre	**Mr. Sanjay Deshpande** Data Curator, IBDC
17:25-17:40	Flash talk	Indian Crop Phenome Database (ICPD) at IBDC	**Dr. Sonia Balyan** Scientist, IBDC
17:40-17:55	Flash talk	Indian Metabolome Data Archive	**Dr. Shivani Sharma** Data Curator, IBDC
17:55-18:10	Flash talk	Indian Structural Portal at IBDC	**Dr. Pawan Kumar** Data Curator, IBDC
18:10-19:00	**Poster Session-I (Even Numbers) and Sponsors Exhibition**		
19:00-20:00	**Cutural Program**		
20:00 onwards	**Gala Dinner**		

Main Conference: Day 2

Venue: Indian Biological Data Centre of Department of Biotechnology, GoI housed at RCB, NCR Biotech cluster, Faridabad

Thursday, 7th March 2024			
Session IV	**Data curation & Standards (I)**		
Chairs	**Prof. Raja Mazumder and Dr. Pascale Gaudet**		
Hall	**M.K.Bhan Auditoruim**		
9:00-9:10	**Presentation of the 2023 Exceptional Contribution to Biocuration Award to Dr. Pascal Gaudet**		
9:10-9:40	ISB Lifetime Awarde e lecture	Maintaining curated resources current in an evolving knowledge landscape	**Dr. Pascale Gaudet** ISB Lifetime Awardee 2023, GO Central & SIB Swiss Institute of Bioinformatics, Switzerland
9:40-10:10	Keynote lecture	Biocuration for ensuring national food security	**Prof. Saurabh Raghuvanshi** Department of Plant Molecular Biology, Univ. of Delhi South Campus, India
10:10-10:30	Session talk	Biomarker-centric data modeling and knowledge exploration for translational research	**Prof. Raja Mazumder** Biochemistry and Molecular Medicine, George Washington University, USA
10:30-10:50	Session talk	Assembly and Reasoning over Semantic Mappings at Scale	**Dr. Charles Tapley Hoyt** Northeastern University, USA
10:50-11:10	Session talk	Empowering the translation of semantic Disease Ontology (DO) data to knowledge in DO-KB	**Dr. Lynn M. Schriml** Associate Professor, University of Maryland, USA
11:10-11:40	**Tea Break**		
11:40-12:00	Session talk	Standardized naming of microbiome samples in Genomes OnLine Database (GOLD)	**Dr. TBK Reddy** Genomic Standards Group Lead, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, USA
12:00-12:20	Session talk	Unlocking a Century of Biological Insights: Connecting the ATCC Genome Portal to a Historical Collection	**Scott V. Nguyen** Senior Biocuration Scientist, American Type Culture Collection, USA
12:20-12:40	Session talk	Biocuration, diversity in knowledge work, and epistemic justice	**Dr. Sarah R Davies** Professor of Technosciences, University of Vienna, Austria
12:40-13:00	Session talk	Trialling DataSeer - streamlining the pre-review of manuscripts submitted to GigaScience Press	**Mary Ann Tuli** Data Editor, GigaDB, GigaScience Journal at GigaScience, UK
13:00-13:20	Session talk	Data standards and data diversity: Lessons and questions from curating cancer data in COSMIC	**Dr. Rachel Lyne** Scientist, COSMIC, Wellcome Sanger Institute, UK
13:20-14:30	Lunch Break		

Session VI	**Data Curation & Standards (II)**		
Chairs	**Dr. Frederic Bastion and Dr. Sushma Naithani**		
Hall	**M.K. Bhan Auditoruim**		
14:30-15:00	Keynote lecture	Plant Reactome Knowledgebase: Decoding Plant Pathway Networks and Unraveling the Genotype-to-Phenotype Connection	**Dr. Sushma Naithani** Associate Professor, Oregon State University, USA
15:00-15:20	Session talk	Semantic Web technologies enabling integration of glycoscience data in the GlyCosmos Portal	**Dr. Kiyoko F. Aoki-Kinoshita** Professor, Institute for Glycobiology and Integrative Biosystems, Soka University, Japan
15:20- 15:40	Session talk	Enabling FAIR access to single-cell RNA-Seq data for reproducible analyses	**Dr. Frederic Bastian** Associate Director, Swiss Institute of Bioinformatics (SIB), Switzerland
15:40-16:00	Session talk	GlyGen: Advancing Glycoscience through Integrated Knowledge Discovery	**Mr. Jeet Vora** Research Associate, The George Washington University, Washington DC, USA
16:00-16:15	Flash Talk	Leveraging Wikidata to improve OBO Foundry ontologies	**Tiago Lubiana** University of São Paulo, Brazil
16:15-16:30	Flash talk	Towards Synthetic Cell Types: curating genes, cells, and functions for open-source innovation in direct reprogramming and industrial cell lines	**Thea Gayatri Fennell** University of Cambridge, UK
16:30-17:00	**Tea Break**		
17:00-17:15	Flash talk	Increasing data evidence by text-fragment references using nanopublications	**Dr. Ulrike Wittig** Scientist, Heidelberg Institute for Theoretical Studies, Germany
17:15-17:30	Flash talk	ChatGPT usage in Reactome curation process	**Krishna Tiwari** Scientist, EMBL-EBI
17:30-17:45	Flash talk	An integrated literature search, triage and extraction workflow for biocuration	**Matt Jeffryes** Biomedical text mining researcher, EMBL-EBI, UK
17:45-19:30	**Poster Session- II (Odd Numbers) and Sponsors Exhibition**		
	Visit to Data Center (IBDC)		
19:30 onwards	**Dinner**		

Thursday, 7th March 2024 (Concurrent Session)			
Session V	**Biocuration and Human Health (Seminar hall 2)**		
Chairs	**Dr. Prof. Pravat Kumar Mandal and Prof. K Thangaraj**		
11:30-12:00	Keynote lecture	GenomeIndia: A population genomics endeavor	**Prof. K Thangaraj** J C Bose Fellow, CSIR-Centre for Cellular & Molecular biology, Telangana, India
12:00-12:20	Session talk	SWADESH: a multimodal multi-disease brain imaging and neuropsychological database and data analytics platform	**Prof. Pravat Kumar Mandal** Scientist, National Brain Research Institute, India
12:20-12:40	Session talk	Digital Information-Systems For Research In Hypoxic-Systems	**Dr. Pankaj Khurana** Defence Institute of Physiology and Allied Sciences, New Delhi, India
12:40-13:00	Session talk	DBT-Centre for Microbial Informatics (DBT-CMI): Genomic exploration of the Indian microbial landscape	**Dr. H. A. Nagarajaram** Professor and Head, Dept. of Systems and Computational Biology, School of Life Sciences, UOH, India
13:00-14:30	Lunch Break		
Session VII	**AI and ML in Action (Seminar hall 2)**		
Chairs	**Prof. G.P.S Raghava and Dr. Dinesh Gupta**		
14:30-15:00	Keynote lecture	Establishing and Sustaining Public Databases in Our Group	**Prof. G.P.S Raghava** Head (CB), Indraprastha Institute of Information Technology Delhi (IIIT-Delhi), India
15:00-15:20	Session talk	The Genomic Standards Consortium shift to using Link-ML	**Dr. Christopher Ian Hunter** Director of GigaDB at GigaScience, UK
15:20- 15:40	Session talk	Curation of imaging datasets and databases	**Dr. Dinesh Gupta** Team Lead, International Centre for Genetic Engineering and Biotech (ICGEB), India
15:40-16:00	Session talk	Deep Learning the Terpenome: Data Semantification using AI over NLP	**Dr. Gitanjali Yadav** Scientist, National Institute of Plant Genome Research, New Delhi, India
16:00-16:20	Session talk	Integration, harmonization and extrapolation of legacy gene expression data sets using deep learning	**Prof. Shandar Ahmad** Principal Investigator, Jawaharlal Nehru University, New Delhi, India
16:20-16:40	Session talk	Unlocking Bio-Curation: Harnessing Text Highlighting for Neural Named Entity Models of SABIO-RK database	**Dr. Ulrike Wittig** Scientist, Heidelberg Institute for Theoretical Studies, Germany
16:40-17:00	**Tea Break**		
17:00-17:20	Session talk	Classification of Brain Cancer Gene Expression Using RNN Approach	**Dr. Heba M.Afify** Associate Professor, University of Cairo, Egypt
17:20-17:35	Flash talk	Cholangiocarcinoma-associated risk factors identified by ontology term enrichment analysis	**Anuwat Pengput** Department of Biomedical Informatics, University at Buffalo, USA
17:45-19:30	**Poster Session- II (Odd Numbers) and Sponsors Exhibition**		
	Visit to Data Center (IBDC)		
19:30 onwards	**Dinner**		

Main Conference: Day 3

Venue: Indian Biological Data Centre of Department of Biotechnology, GoI housed at RCB, NCR Biotech cluster, Faridabad

Friday, 8th March 2024			
SESSION VIII	**Protein Structural Bioinformatics**		
Chairs	**Dr. Debasisa Mohanty and Prof. M. S. Madhusudhan**		
Hall	**M.K.Bhan Auditoruim**		
9:00-9:30	Keynote lecture	Evaluation of AI/ML based methods for prediction of novel folds, oligomeric complexes of proteins and binding affinity of protein-ligand complexes	**Dr. Debasisa Mohanty** Director, National Institute of Immunology, New Delhi, India
9:30-9:50	Session talk	Reference Proteomes in UniProtKB	**Dr. Dushyanth Jyothi** Project Leader, EMBL-EBI, UK
9:50-10:10	Session talk	PDB: Improved and enriched biodata repository serving many millions of users worldwide	**Dr. Brinda Vallat** Associate Research Professor Rutgers University & Scientific Software Developer at RCSB Protein Data Bank, USA
10:10-10:30	Session talk	Curating the biological functions of proteins using the AlphaFold Protein Structure Database	**Maxim Tsenkov** Bioinformatician, EMBL-EBI
10:30-10:50	Session talk	Molecular interactions in the context of Rare diseases: Annotation rich dataset from the IMEx consortium	**Dr. Kalpana Panneerselvam** Senior Scientific Database Curator, EMBL-EBI UK
10:50-11:30	**Tea Break**		
11:30-11:45	Flash talk	Standardisation of protein modification data in the Protein Data Bank (PDB) archive	**Marcus Bage** Bioinformatician, EMBL-EBI, UK
11:45-12:00	Flash talk	Improved findability of small molecule data in the PDB	**Ibrahim Roshan Kunnakkattu** Bioinformatician, PDBe, EMBL-EBI, UK
12:00-12:15	Flash talk	DrugMechDB: A Curated Database of Drug Mechanisms	**Dr. Umasri Sankarlal** Scientist, The Scripps Research Institute, Department of Integrative Structural and Computational Biology, USA
12:15-12:35	Session talk	Design considerations for securely storing and efficiently accessing large biological datasets on AWS cloud	**Mainak Chakraborty** Senior Solutions Architect at Amazon Web Services (AWS)
12:35-14:30	**Lunch Break**		

SESSION IX	**Translating Biocuration: Applications in Agriculture and Healthcare**		
Chairs:	**Dr. Rajendra Joshi and Prof. Mukesh Jain**		
14:00-14:30	Keynote lecture	Integrated Computing Environment for Next Generation Biology: Cloud based HPC and Big Data Platform	**Dr. Rajendra Joshi** Senior Director, Centre for Development of Advanced Computing (C-DAC), India
14:30-14:50	Session talk	Stress Combinations and their Interactions in Plants Database	**Dr. Senthil-Kumar Muthappa** Scientist, National Institute of Plant Genome Research, New Delhi, India
14:50-15:10	Session talk	Genomic tools and resources for knowledge-based functional and translational genomics research in crop plants	**Prof. Mukesh Jain** Professor, School of Computational and Integrative Sciences, JNU, New Delhi, India
15:10-15:30	Session talk	Development of a genomics-assisted coreset in sesame from the collections being conserved at the National Genebank of India	**Dr. Rangan P** Scientist, National Bureau of Plant Genetic Resources, New Delhi, India
15:30-15:50	Session talk	Indian Breast Cancer Genome Atlas: Emerging India-specific Molecular Attributes	**Dr. Shantanu Chowdhury** CSIR-Institute of Genomics and Integrative Biology, New Delhi, India
15:50-16:10	Session talk	A review of biocuration efforts and alternative approaches in transcriptomics, meta analysis and other bioinformatic research, across 15 years.	**Dr. Kshitish K Acharya** Professor, Institute of Bioinformatics and Applied Biotechnology, Bengaluru, India
16:10-16:40	**Tea Break**		
16:40-17:00	Session talk	Cancer Imaging Biobank	**Dr. Swapnil Rane** Professor (F) of Pathology, Tata MemorialCentre (TMC), Mumbai, India
17:00-17:20	Session talk	Reading Genomes	**Prof. M. S. Madhusudhan** Professor, Indian institute of Science, Education and Research, Pune, India
17:20-19:30	**Panel Discussion**		
	Award Ceremony and Announcement of ISB 2025 with concluding remarks		
	Vote of Thanks		
19:30 onwards	**Dinner**		

CONTACT DETAILS

Email: biocuration2024@ibdc.rcb.res.in

X platform: @aibc2024

indradhanush

by: aayam institute of performing arts

Chair: Ms. Sindhu Mishra

1. Shivashtakam – Group Bharatanatyam

2. Bhavai – Gujrat Folk

3. Tillana – Duet Bharatanatyam

4. Chari – Rajasthan Folk with fire on head

5. A version of Terah Taali of Rajasthan, performed with Manjira

6. Cherao bamboo dance – Mizoram

7. kaalbeliya – Rajasthan folk

8. Pung Chollam – Manipuri

9. Bhangra – Punjab

INDEX

Pre-conference workshop: GlyGen and Common Fund Data Ecosystem (CFDE) Bioinformatics Workshop

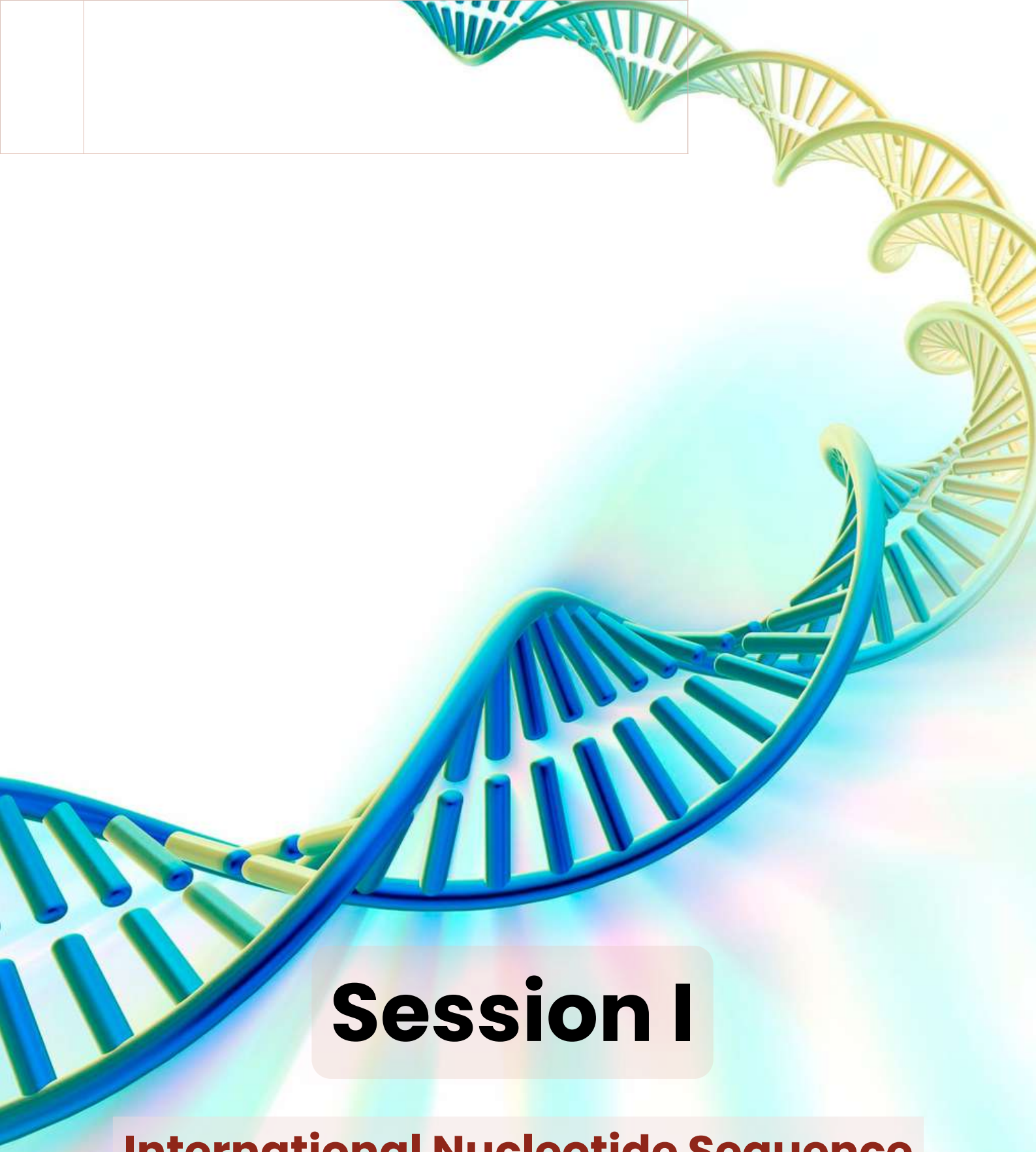
Organizing Committee:
Raja Mazumder (GlyGen, CFDE), Jeet Vora (GlyGen, CFDE), Suvarna Nadendla (CFDE), Saurabh Raghuvanshi (Univ. of Delhi), Sonia Balyan (IBDC), Utkarsh Raghuvanshi (UDSC), Mike Tiemeyer (GlyGen, CFDE, remote), Rene Ranzinger (GlyGen, CFDE, remote)

Abstract

An interactive 3-hour in-person bioinformatics workshop is scheduled for March 5th, 2024, from 10 AM to 1 PM. Blending training sessions with hands-on activities the workshop is designed for students, post-docs, and early investigators interested in bioinformatics analysis and knowledgebase development. Attendees will be able to participate using their cell phones, mobile devices, or laptops. This workshop marks a continuation of the Glycan Function Workshop series initiated last year during Biocuration 2023 in Padua, Italy. The focus of the workshop will be on bioinformatics analysis use cases that span exploration of glycan and protein function, gene mutations, gene expression, and how biocuration along with large-scale analysis, querying, and filtering of data can reveal insights into biomolecule function. NIH's Common Fund Data Ecosystem will also be presented.

This workshop aims to provide attendees with practical skills in bioinformatics analysis. Best practices in knowledgebase development and biocuration will also be discussed, fostering seamless integration into existing resources worldwide and making data findable, accessible, interoperable, and reusable. There will also be a feedback session to identify gaps in current resources.

https://wiki.glygen.org/GlyGen_CFDE_Biocuration2024_Workshop



Session I

International Nucleotide Sequence Database Collaboration (INSDC)

Keynote lecture

AIBC-ABS-122

Title: The International Nucleotide Sequence Database

Dr. Guy Cochrane, European Nucleotide Archive, EMBL-EBI

The International Nucleotide Sequence Database Collaboration (INSDC) has for a period spanning five decades captured, preserved, curated, and presented the world's sequence data. The data services provided by its member organisations continue to provide a foundation within and beyond the life sciences. With the rapid advance of sequencing technology, the ever-broadening application of sequencing and the ubiquity of the method across and beyond the life sciences, the databases of the INSDC have continuously developed to serve new platforms, applications, and user constituencies. In the presentation, I will introduce the INSDC and outline its operations. I will highlight a number of important developments over its history, including the relationships with scientific publishers, engagement with data standards organisations, and developments in relation to emerging technologies. Finally, I will outline areas of current work and address our approach to our growing and diversifying the global user base.

Session talk

AIBC-ABS-123

Title: Coping with benefit sharing from digital sequence information

Masanori Arita, National Institute of Genetics, Japan

In the recent CBD (convention on biological diversity) discussion on the benefit sharing from digital sequence information in Geneva, national perspectives remained divergent and showed no indication of convergence, let alone agreement. This has been seen over the past 30 years, and the factor of time must be considered. It is best to start some types of benefit sharing before new technological innovations change the society, as we have seen with AI. Japan proposes a voluntary mechanism where businesses and organizations donate to the global multilateral fund to receive the proper recognition and credits from the community. This legally non-binding plan reduces international conflict and is implementable within the existing framework. Examples include the Lion's Share Fund, which requests 0.5% of advertising expenses to protect wild animals. For the sake of basic science and public benefit, DDBJ is in favor of such volunteer efforts to collect and preserve permanent record of nucleotide sequences and their metadata.

Session talk

AIBC-ABS-124

Title: NCBI Resources for Sharing of Sequence Data

Ilene Karsch Mizrachi, Sequence Submission and Archives, Program Head, NCBI, NLM, NIH

The COVID-19 pandemic demonstrated the need for rapid open sharing of genomic sequence data. The US National Library of Medicine's National Center for Biotechnology Information (NLM-NCBI) is host to numerous resources for facilitating the deposition, dissemination, and analysis of genomic sequence data from pathogens. Raw and assembled genomic sequences derived from both isolated and environmental samples can be submitted to the Sequence Read Archive (SRA) and GenBank, respectively, to become available for download and analysis.Metadata describing the source of a sequence, and which is essential to data analysis, is also captured in these submissions and validated via various quality assurance checks. Improvements have been made to our search and retrieval tools that allow users to query for sequence data based on standardized metadata fields or organismal content for analysis both on the web and on commercial clouds. Data submitted to NCBI is exchanged with other International Nucleotide Sequence Archive (INSDC) member databases, the European Nucleotide Archive (ENA; European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)) and the DNA Database of Japan (DDBJ; Research Organization of Information and Systems, National Institute of Genetics (ROIS-NIG)) for more global reach. We will present an overview of these resources and discuss plans to engage with global users to ensure more equitable data access.

Session talk

AIBC-ABS-125

Title: The European Nucleotide Archive: Sequence Data Resource and Submissions Brokering Network

Zahra Waheed, Senior Bioinformatician, Data Coordination and Archiving Team, EMBL-European Bioinformatics Institute.

Session II

Global biodata resources: a crucial infrastructure underpinning life science research

Session talk

AIBC-ABS-83

Importance of Data Persistence

Masanori Arita, National Institute of Genetics, Japan

In the era of open science, several guidelines such as FAIR and CARE have been proposed. These principles are worth consideration especially by public repositories, but their prime mission is the preservation or data or persistence. Since biology data require updates to keep them FAIR; this issue alone invokes various issues different from those that arise in art or natural history museums. Such topics are not solvable at each institution, and global alliance is the right place to cope with.

Keynote lecture

AIBC-ABS-126

The global biodata infrastructure and the drive for sustainability

Dr. Guy Cochrane, European Nucleotide Archive, EMBL-EBI

Progress in life and biomedical sciences depends absolutely on biodata resources—databases comprising biological data and services around those databases. Supporting scientists in data operations and spanning management, analysis and publication of newly generated data and access to pre-existing reference data, these biodata resources together comprise a critical infrastructure for the domain. Unlike other scientific infrastructures, biodata resources are globally distributed and lack any kind of central coordination. While this configuration supports innovation, it lends itself poorly to the long-term sustainability of individual biodata resources and the infrastructure as a whole. The Global Biodata Coalition (GBC) brings together life science research funding organisations that recognise these challenges and acknowledge the threat that the lack of sustainability poses. They agree to work together to find ways to improve sustainability.

In the presentation I will present an overview of the work that GBC has carried out to understand and classify biodata resources and the entire biodata resource infrastructure. I will address the challenges of ensuring sustained long-term support for the individual resources within the infrastructure. Finally, I will outline the GBC's work with funders to identify more robust methods to build sustainability for the biodata resource infrastructure.

Session talk

AIBC-ABS-127

Super Computing platform and Data Resources derived from Omics Data in Agriculture

K.K. Chaturvedi, ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Doubling crop production per unit area to meet food demand of burgeoning population in India by 2050 is a major challenge under the existing constraints like reduced inputs and agriculture land, global warming and extreme weather conditions, increasing biotic and abiotic stresses. Solution to address some of these problems can be addressed to some extent by omics research, through bioinformatics and computational biology. The major omics data are genomics, proteomics, transcriptomics and metagenomics. With the advent of High-Performance Computing (HPC), Artificial Intelligence (AI) and advances in Next Generation Sequencing (NGS) technologies, it is feasible to support biologists and other application scientists by utilizing the advanced knowledge of HPC, parallel and distributed computing, and machine learning techniques to analyze and decode this diverse omics data. Apart from this, extraction, integration and storage of these omics data further encompasses need of large scale computational infrastructure to meet the need of ever growing omics data generation. In ICAR, significant efforts have been made by establishing supercomputing facility, called Advanced Supercomputing Hub for Omics Knowledge in Agriculture (ASHOKA) to analyze the data generated in various labs established in the NARES system. In addition to this, several secondary data bases, web servers and algorithms were also developed and made available to the biologists. This talk will briefly cover such computational initiatives in ICAR including the National Agricultural Biocomputing Portal and Omics Data Warehouse.

Session talk

AIBC-ABS-128

Health Research Data Ecosystem – Challenges and Opportunities

Dr. Harpreet Singh, Indian Council of Medical Research, India

This presentation explores into the health research data ecosystem, a dynamic network of stakeholders including researchers, healthcare professionals, and patients, all interconnected through health data and technology. It emphasizes the ecosystem's role in driving innovation, enhancing data reuse, improving research quality and efficiency, and advancing health equity and access via informed public health policies. However, the establishment of such an ecosystem faces significant challenges, such as data diversity, interoperability, and access to program-generated data. The Indian Council of Medical Research (ICMR) is at the forefront of addressing these issues, launching initiatives like the Strategic Vision Document – 2030, a comprehensive policy on health research data management, a multi-institutional effort to create thematic gold standard reference data banks, and the development of an integrated research data platform. These efforts mark critical steps towards realizing a robust health research data ecosystem.

Session talk

AIBC-ABS-129

Pillars of resource sustainability for the Human Disease Ontology Knowledgebase (DO-KB)

Dr. Lynn M. Schriml, University of Maryland School of Medicine, USA

Sustainability, in the context of a biomedical data knowledgebase, is measured by the pillars of longevity, reliability, and availability. The Human Disease Ontology (DO) and DO Knowledgebase (DO-KB) have focused, since 2003, on the preservation and long-term availability of FAIR and TRUST-worthy disease knowledge. Fulfilling the need for an authoritative, comprehensive disease classification resource, the DO established trustworthy governance, lifecycle and usage metrics along with rigorous best practices to become the 'go to' resource for new disease terms and disease classifications. Further, the DO-KB has been recognized as a critical resource among the global infrastructure of interconnected resources, as a Global Biodata Coalition (GBC) Global Core BioData Resource. Facilitating disease data standardization across the disease data ecosystem, the DO has been incorporated into over 390 biomedical resources (including the model organism databases: FlyBase, MGD, RGD, SGD, WormBase, Xenbase, ZFIN) and used in thousands of individual research projects. Sustainability is incremental, as building the reputation of your resource takes time and the development of key collaborations. Community resources must demonstrate their commitment to serving the needs of their communities, their reliability for integrating new knowledge, and their openness to discourse in order to deliver scientific impact to the communities that they serve.Our diversified funding model (e.g., R01s; Community Resource funding U41, U24; Common Fund) has provided opportunities for new collaborations and areas of knowledge expansion (e.g., Human FMA, CFDE, Wikidata, CIViC). Strategic growth into areas of evolving disease research (e.g., molecular variants of pediatric cancers; RNA-associated diseases; rare diseases) has been instrumental in enhancing DO's utility and engaging new stakeholders. Whereas, moving to a CC0 license opened the DO for clinical and pharmaceutical research opportunities, while the DO-KB's Open Linked Data SPARQL endpoint expanded the potential for cross-resource data querying and retrieval of ML-ready dataset.'

Session talk

AIBC-ABS-130

Bgee, GCBR for animal transcriptomics: sustainability and prospects

Dr. Frederic Bastion, Swiss Institute of Bioinformatics (SIB), Switzerland

Bgee is a database to retrieve and compare gene expression data across experiments and species. It notably provides high-quality annotations of multiple transcriptomics technologies, integrated in one place, including single-cell and bulk RNA-Seq data. This work involves highly labor-intensive manual curation, supported by data wrangling tools. Distributing this very large amount of data through multiple access points (SPARQL endpoint, R package, JSON API, website) to make them FAIR also requires a significant investment in personal and infrastructure material.

Bgee is part of the portfolio of resources of the SIB Swiss Institute of Bioinformatics. The SIB has a long-term strategy of providing stable support for the development of bioinformatics resources, and for providing high-quality curation of biological data. This context has been unique for many years, and as a result, Switzerland has the highest density of biocurators in the world, notably thanks to the establishment of SIB resources such as Swiss-Prot.

We will present here the sustainability aspects and prospects for Bgee in this quite unique environment for Biodata Resources: long-term support from SIB to bioinformatics resources, strategic plan from Swiss universities to support Open Research Data. We will also present how, despite these opportunities, retaining a stable staff to produce the Bgee resource can be highly challenging.

Session III

Indian Biological Data Center

Keynote lecture

AIBC-ABS-131

Indian Biological Data Centre (IBDC)

Dr Deepak T. Nair, Professor, Regional Centre for Biotechnology and Head, Indian Biological Data Centre, India

The Indian Biological Data Centre (IBDC), represents the first national digital data repository in India devoted to archive and disseminate life sciences data. The centre is located at the Regional Centre for Biotechnology and is supported by the Government of India through the Department of Biotechnology. IBDC is a joint project of the Regional Centre for Biotechnology (RCB), the National Institute of Immunology (NII), the National Informatics Centre (NIC) and the International Centre for Genetic Engineering and Biotechnology (ICGEB), New Delhi. A brief history of IBDC, the current status and future plans will be presented.

Session talk

AIBC-ABS-132

Indian SARS-CoV-2 Genomics Consortium (INSACOG)

Dr. Nidhan Biswas, National Institute of Biomedical Genomics (NIBMG), India

The Indian SARS-CoV-2 Genomics Consortium (INSACOG) is a network of 67 laboratories across India that conducts whole genome sequencing of the SARS-CoV-2 virus to monitor genomic variations of the virus over temporal and spatial space. Initiated by the Union Health Ministry and the Department of Biotechnology, and the Department of Biotechnology (DBT) with the Council for Scientific & Industrial Research (CSIR) and the Indian Council of Medical Research (ICMR), it aims to deepen understanding of virus spread and evolution for effective public health response. On behalf of INSACOG, the National Institute of Biomedical Genomics (NIBMG) has taken the initial role in establishing the INSACOG DataHub portal to efficiently capture data from all participating institutes within the consortium. The INSACOG DataHub served as a centralized repository for storing the metadata and the FASTA sequences, facilitating the seamless exchange of genomic information among various research institutions. Notably, the portal was equipped with robust analytical capabilities with updated community standard tools, allowing for near real-time data analysis. It was designed to generate comprehensive reports promptly, providing valuable insights into the genomic landscape of the virus, and the quick turnaround time, with reports delivered within a day, contributed significantly to the consortium's ability to respond swiftly to emerging viral variants. Recognizing the need for further scalability, sustainability, and integration with other data types such as raw fastq files, NIBMG transitioned the entire data workflow to the Indian Biological Data Centre (IBDC). This move aimed to enhance the efficiency, accessibility, and collaborative potential of the genomic data analysis process and by transferring the workflow to IBDC, NIBMG and INSACOG collectively bolstered their capabilities in understanding and responding to the evolving nature of the SARS-CoV-2 virus, thereby contributing to national and global efforts in managing the ongoing pandemic. As of 29-02-2024, IBDC has total submission of 2,79,138 sequences of SARS-CoV-2 from INSACOG partnered institutes.

Flash talk

AIBC-ABS-109

Nucleotide Data Resources at Indian Biological Data Centre

Sanjay Deshpande, Kalpanath Paswan, Indu Kumari, Arvind Yadav, Himanshu Bhusan Samal, Deepak T. Nair, Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad

Nucleotide portals at IBDC have been developed to archive and disseminate nucleotide data. The portals have been designed keeping the formats set by International Nucleotide Sequence Database Collaboration (INSDC) guidelines. The database serves as a comprehensive archival platform for Raw data, assembly, and annotated sequences. The nucleotides portals are divided in 2 modules called INDA and INDA-CA. INDA archives the data, which is open access in nature, with data brokered to European Nucleotide Archive (ENA) to get INSDC accessions. INDA-CA is a portal with a specific focus on accommodating data that may be sensitive in nature, like the data from the Indian SARS-CoV-2 Genomic Consortium (INSACOG), INSACOG-Sewage Surveillance, and GenomeIndia initiatives. The databases contribute to the advancement of genomic research by providing a scalable, secure, and user-centric platform for the archival and analysis of genomic data. It enables seamless data integration, fostering collaboration and knowledge sharing among researchers while maintaining the highest data integrity and security standards.

Flash talk

AIBC-ABS-120

Indian Crop Phenome Database (ICPD) at IBDC

Sonia Balyan, Mayuri Jain, Nivedita Yadav, Isha Saini, Deepak T. Nair, Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad

The Indian Crop Phenome Database (ICPD), housed within the Indian Biological Data Centre (IBDC), plays a pivotal role in advancing agricultural research. Its primary purpose is to streamline the digitization, management, storage, analysis, and exchange of crop phenotyping data. Given India's prominence as a global agricultural powerhouse, a substantial volume of biological data is generated through diverse agricultural trials. This data provides crucial insights that guide breeders in selecting cultivars with desirable traits. However, a significant challenge lies in the underutilization of plant phenotype data due to its non-uniform structure and the absence of discipline-specific repositories. To address this, the ICPD adheres to the FAIR principles—ensuring that data is Findable, Accessible, Interoperable, and Re-usable. By aligning with international data standards, the ICPD establishes a standardized and efficient framework for handling crop phenome data. Additionally, each submitted dataset receives a valid accession, allowing users to contribute phenotype data using ontology terms related to traits, tissues, development stages, and methods across more than 30 crop species.

In summary, the Indian Crop Phenome Database serves as a valuable resource for researchers, scientists, and policymakers, empowering them to explore and harness crop phenotype information effectively in their work.

Flash talk

AIBC-ABS-23

Indian Metabolome Data Archive

Shivani Sharma, Mayuri Jain, Deepak T Nair, IBDC, Regional Centre for Biotechnology

Indian Metabolome Data Archive (IMDA) is a database by Indian Biological Data Center (IBDC), Regional Centre for Biotechnology, Faridabad to catalogue metabolomics data and their associated experimental metadata with various parameters including (experimental conditions, sample details and instrument characteristics) of mass spectrometer (MS) and Nuclear Magnetic Resonance (NMR) techniques. IMDA accepts targeted and un-targeted metabolomics data as well as their reference spectra and metabolite structures obtained from MS and NMR techniques in metabolomics experiment. IMDA database supports raw (d, raw, idb, netcdf, wiff, scan, dat etc.) as well as derived (mzml, nmrml, mztab, mzxml, mzdat) file formats of metabolomics studies. The raw data can be uploaded in the form of binary files and processed data in the form of quantitated metabolite concentrations, MS peak height/area values, LC retention time, NMR binned areas etc. A unique and persistent IBDC accession will be assigned on data submission to IMDA.

Flash talk

AIBC-ABS-108

Indian Structural Portal at IBDC

Pawan Kumar, Himanshu Bhusan Samal, Mayank Mamgaain, Kalpanath Paswan, Deepak T. Nair, IBDC, RCB

ISDA (Indian Structural Data Archive) constitutes a comprehensive repository of biological 3D-macromolecular structures. The database is in sync with the wwPDB to weekly retrieve the data for major updates. All biological entries in mmCIF format are used to extract the summary, experimental and functional information along with other information such as publication, etc. to update the MongoDB database. The database is also focused on providing critical insights into the function and biology of structure, ligand and other factors. Consequently, each entry in the database is further enriched with different level functional annotations as Pfam, SCOP, CATH, GO, Enzyme, etc. On a weekly basis, we are also manually curating and updating the Indian origin structural entries and classifying them on author affiliation states. All entries are grouped on the basis of five major taxonomic categories to empathize with the quantitative number of entries belonging to each source organism. To navigate regularly updated Covid-19 disease related entries, a curated "Covid-19 Resource" page is designed which provides direct access to the Covid-19 protein-wise experimental structures categorized over experiment techniques. We also incorporated 3D molecular visualizers for chemical and macromolecular visualization in the database using JSMol and Mol*. The download is available for validation report and the 3D-structural information in. cif, .pdb and its zipped (gz) format.

In future, ISDA will host an analysis portal for water conservation analysis, protein network analysis, protein surface map characterized by electrostatic and hydrophobic properties as well as ensemble analysis for the same class of the protein to probe global conformational changes.

Session IV

Data curation & Standards (I)

Keynote lecture

AIBC-ABS-90

Maintaining curated resources current in an evolving knowledge landscape

Pascale Gaudet, GO Central & SIB Swiss Institute of Bioinformatics

Biomedical ontologies have been a major tool in bioinformatics, enabling the classification and analysis of genomics, transcriptomics and proteomics data. The Gene Ontology is the first of the biomedical ontologies to have been developed, staring over 20 years ago, and is one of the most widely used. GO currently contains over 42,000 terms connected by 85,000 relations. There are more than a million GO annotations derived from experimental data, and several order of magnitude more obtained from automated methods and semi-automated methods.

As knowledge evolves, it is essential that the ontology and associated annotations remain up to date, to ensure that the data analysis supported by these tools continue to provide useful results. As the ontology and the annotation corpus grow, this task becomes both increasingly complex and increasingly indispensable. In addition to term creation to represent newly discovered processes, functions and components, ontology refactoring is a central part of ontology development. I will explain the main areas of ontology refactoring in the Gene Ontology, illustrating why these changes are necessary.

Another approach that contributes to improving consistency of the annotation corpus is phylogenetic annotation, which allows to annotate a large number of genes linked via evolutionary relationships. We have curated all Panther protein families containing at least one human gene and that contained experimental data that can be used to infer some functional information. Interesting insights about protein function evolution will be described.

Keynote lecture

AIBC-ABS-133

Biocuration for ensuring national food security

Saurabh Raghuvanshi, University of Delhi South Campus

Increasing population poses a constant pressure on the agricultural output to ensure food security at national level. Agricultural output primarily depends on three major research activities: trait-gene association, Precision Breeding and Precision agriculture. While the first two activities target crop improvement, precision agriculture attempts to ensure crop management in an ecologically and economically sustainable manner. Data science is at the very core of these activities and thus Biocuration plays a pivotal role. While biocuration has always been a critical factor in trait-gene association studies, however, most curation efforts have concentrated on model plant systems. Since agricultural output of a nation depends on a number of non-model crop systems, huge amount of data is getting generated in non-model systems. Consequently, biocuration activities in non-model crops (e.g. millets, pulses, horticulture crops etc.) must keep pace. Further, crop breeding (not genetic-engineering) still remains the major tool for crop improvement. A major bottle-neck in breeding approaches is the extensive and laborious step of plant phenotyping. AI-ML data models for trait prediction can ensure 'Genomic Selection' thereby drastically reducing the magnitude of phenotyping. Biocuration and data integration in non-model crop systems can further increase the accuracy of these prediction models (which are largely based on genetic variability data). On the other hand, 'Precision agriculture' includes disease prediction, soil status and crop health monitoring. Highly curated and crop-specific reference image datasets are crucial for developing crop-specific data models for predicting disease and monitoring crop health.

Thus, biocuration activities at national level needs to be boosted extensively in order to ensure national food security in an economically and environmentally sustainable manner. As more and more life science 'Big data' starts to accumulate from non-model crop systems biocuration activities must keep pace. An 'Agri-grid' at national level which can provide benchmark curated data sets as well as real-time agricultural monitoring data from across the nation is envisaged.

Session talk

AIBC-ABS-75

Biomarker-centric data modeling and knowledge exploration for translational research

Raja Mazumder, George Washington University

Biomarkers are a fundamental focus in many clinical inquiries and research. Emerging as crucial tools in the paradigm shift towards Predictive, Preventive, Personalized, and Participatory (P4) medicine, biomarkers not only serve as valuable indicators in drug development and surrogate endpoints in clinical trials but also contribute to the evolving landscape of personalized healthcare. Current research endeavors often result in the extensive accumulation of distributed data sets encompassing diverse biological data types (e.g., genes, proteins, glycans, metabolites, cells, etc.), shedding light on nuanced alterations in cellular functions that connect human biology with disease. Although NIH's Common Fund (CF) research has generated a substantial data repository offering contextual insight, provenance, and quantitative inference for biomarker data types, challenges persist in the systematic harmonization, organization of biomarker data, and establishment of robust connections to CF data. The CF-supported biomarker project aims to address these gaps by proposing a two-fold approach—initially harmonizing and subsequently mapping biomarkers from public sources (e.g., NCBI, EBI, SIB, publications) to and across CF and Electronic Health Records (EHR) data elements. This strategic mapping will foster knowledge integration across multiple CF Data Coordination Centers (DCCs) and diverse biomedical disciplines represented by the DCCs and CFDE. This presentation introduces the core biomarker data model and outlines a comprehensive strategy for integrating, harmonizing, and adding value to biomarker data derived from public resources and publications, while also incorporating additional contextual summary information from EHRs. Examples will be provided, showcasing the integration of biomarker data into GlyGen (a glycoinformatic knowledgebase). Additional information at: https://hivelab.biochemistry.gwu.edu/biomarker-partnership and https://glygen.org/.

Session talk

AIBC-ABS-103

Assembly and Reasoning over Semantic Mappings at Scale

Charles Tapley Hoyt and Benjamin M. Gyori, Northeastern University

Scientific discovery in biomedicine relies on the integration of data and knowledge from diverse sources. However, this is encumbered by the inconsistent usage of partially overlapping nomenclature resources (e.g., ontologies, databases) that results in unresolved redundancies. Semantic mappings are crucial for establishing equivalence and other relations (broad, narrow, related, etc.) across resources, enabling consistent integration. Such mappings are often made available by individual resources, aggregators such as BridgeDB, or independent mapping repositories like Biomappings. However, they remain difficult to assemble at scale because of the variety of storage formats (e.g., ontologies, SSSOM, ad-hoc formats), the ways they are produced (e.g., biocuration, rule-based inference, lexical matching), and the availability of metadata (e.g., precise mapping relations, curator confidence).

To address these issues, we introduce the Semantic Mapping Reasoning Assembler (SeMRA), a configurable workflow for automatically assembling mappings at scale. SeMRA represents mapping sets as a directed graph, and provides functionality to infer indirect mappings based on graph traversal then determine an associated confidence. Importantly, it allows for customizing prioritization order to merge equivalent concepts during data integration in a consistent way. We demonstrate SeMRA in three scenarios: 1) an automated assessment of the landscape of cell and cell line resources, 2) preparing a coherent dictionary to support lexical matching methods in named entity recognition, and 3) automated assembly of biomedical knowledge at scale. SeMRA is available as an open-source Python package (https://github.com/biopragmatics/semra) and provides a comprehensive mapping database in the SSSOM exchange format.

Session talk

AIBC-ABS-20

Empowering the translation of semantic Disease Ontology (DO) data to knowledge in DO-KB

Lynn M. Schriml, Claudia Sanchez-Beato Johnson, J. Allen Baron, Institute for Genome Sciences, UMSOM

The new Disease Ontology Knowledgebase, DO-KB (https://disease-ontology.org/do-kb/), empowers users to identify and connect data across Linked Open Data resources. The DO-KB SPARQL service and Faceted Search Interface tools transform the Disease Ontology's (DO) semantically defining "disease to disease" and "disease to feature" relationships (e.g., phenotype, symptom, age of onset, anatomy, genetic and environmental driver) into ML-ready datasets. Curators, resource developers, and data stakeholders are invited to identify and export slices and dices of the DO, for example, downloading all of the DO terms in the DO_cancer_slim report or the list of diseases with MeSH and OMIM cross-references. DO-KB is empowering diverse data discovery through the addition of DO-specific and federated SPARQL query options each month. These query options provide alternatives for mining disease-related, semantically-defined data. For example, DO-KB users can download disease-associated proteoforms (Protein Ontology). Connecting disease related data will be further enhanced with the upcoming additions of federated queries to explore pathway (Wikipathways), rare disease (Wikipathways/Orphanet), protein (UniProt), and disease-gene (DisGeNet) datasets.Traversing the path from data to knowledge is a global undertaking that requires coordinated effort among database stakeholders, expert curators and ontology developers. The DO's global user community informs the DO's ongoing, targeted areas of enrichment. In 2023/2024 these include RNA-associated diseases, overlap syndromes, COPD, soft-tissue cancers, sarcoma, and diseases of glycosylation. Additionally, the DO team works to enhance data interoperability through ongoing curation of new diseases in Orphanet, OMIM, GARD and diseases represented in UniProt.

Session talk

AIBC-ABS-10

Standardized naming of microbiome samples in Genomes OnLine Database (GOLD)

TBK Reddy, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

The Genomes OnLine Database (GOLD) is a metadata management system hosting curated metadata for sequencing projects from around the world for the last 25 years. GOLD serves as a gateway to the Integrated Microbial Genomes & Microbiomes (IMG/M) where users can define their projects before submitting assemblies for annotation and comparative analysis. Broader availability of different types of sequencing methodologies has resulted in an explosive growth in the number of sequencing projects. Metagenome projects aim to understand the diversity of complex microbial environments. Unlike isolate genomes with organism names governed by established principles of phylogeny, metagenomes lack systematic naming standards. The interdisciplinary nature of microbiome research and the lack of established standards for naming metagenomes have led to inconsistent names being used for microbiome samples. As a result, many metagenomes in public repositories have cryptic, esoteric names that hinder findability and comparative analysis efforts. To address this challenge, we developed a metagenome naming approach to use distinct metadata features of a sample to construct a standardized name. A combination of habitat, community, location, and distinct identifier makes up a GOLD metagenome name to capture the entire contextual information of the sample. To date we have manually curated over 198,000 microbiome samples with standardized names. We will discuss the process with select examples and encourage the broader microbiome community to adopt this naming system when depositing data to public repositories and publishing manuscripts. This approach makes microbiome data more Findable, Accessible, Interoperable and Reusable (FAIR).

Session talk

AIBC-ABS-35

Unlocking a Century of Biological Insights: Connecting the ATCC Genome Portal to a Historical Collection

Scott V. Nguyen, Nikhita P. Puthuveetil, Joseph R. Petrone, Jade L. Kirkland, Kaitlyn Gaffney, Corina L. Tabron, Noah Wax, James Duncan, Stephen King, Robert Marlow, Amy L. Reese, David A. Yarmosh, Hannah H. McConnell, Ana S. Fernandes, John Bagnoli, Briana Benton, Jonathan L. Jacobs, American Type Culture Collection (ATCC)

The ATCC Genome Portal (AGP, https://genomes.atcc.org/) is a database of authenticated and traceable whole genome sequence assemblies for bacteria, fungi, protists, and viruses held in ATCC's biorepository. Assemblies of ATCC strains sequenced by the scientific community are not directly traceable to isolates in the collection and public NCBI assemblies may represent strains that have accumulated genetic drift arising from laboratory domestication or derived from cultures shared between laboratories. Prior research into these strains show that genetic drift can have a marked impact on phenotypes, posing issues for research reproducibility. The authenticated genomes within the AGP provide researchers a resource to track genetic pedigree and trace the genetic stability of these strains. AGP assemblies are sequenced with both Illumina and Oxford Nanopore platforms to generate hybrid de novo assemblies. Assemblies must pass sequencing quality metrics prior to publishing. AGP assemblies are cross checked against reference taxonomic databases and original depositors' accessioning records, spanning nearly a century's worth of data. As many of these strains were deposited prior to modern genomic identification methods, taxonomic classification of these strains are revised through the AGP initiative. The AGP also prioritizes strains for which there are no publicly available genomes, which will impact microbial taxonomy and provide valuable insight into understudied genera. Access to data in the AGP is available through a REST-API (https://github.com/ATCC-Bioinformatics/genome_portal_api) and through the website.
Full documentation of methods and protocols are available at https://docs.onecodex.com/en/collections/2314023-the-atcc-genome-portal.

Session talk

AIBC-ABS-107

Biocuration, diversity in knowledge work, and epistemic justice

Sarah R Davies & Constantin Holmer, University of Vienna

This paper reports on findings from an exploratory qualitative study of careers in and experiences of biocuration. Based on 15 semi-structured interviews with biocurators, as well as ethnographic engagement with biocuration resources, documents, and activities, we outline some of the key challenges that interviewees report, and analyse these using the concept of epistemic justice. In particular we show that the precarity of biocuration work is a central concern for interviewees, alongside its relative invisibility and under-recognition in the wider research landscape of the biosciences. We connect these experiences to biocuration's framing as a service- and detail-oriented activity, one that involves supporting other research and researchers and that is oriented to collaboration. We mobilise the notion of epistemic justice – developed in the context of the need to acknowledge indigenous and other traditional knowledges – as a means of reflecting on the diverse kinds of activity required to create and manage knowledge in the sciences, and to argue for the need, in science policy, evaluation, and funding, to value and reward diverse forms of knowledge work. Better support for biocuration is thus urgent both for scientific reasons and for those connected to efforts to ensure more just research cultures.

Session talk

AIBC-ABS-71

Trialling DataSeer - streamlining the pre-review of manuscripts submitted to GigaScience Press

Mary Ann Tuli, Yannan Fan, Chris Armit, Christopher I Hunter, GigaDB, GigaScience Press, Hong Kong.

Background

GigaScience Press publishes award winning Open Science journals, which all have the goal of making scientific communication reach researchers and communities around the globe. The Press also runs a data publishing platform (GigaDB) with an in-house data curation team to link published articles with data and other research objects.

Methods

Prior to submitted manuscripts being peer reviewed, GigaDB curators carry out a 'pre-review'. This process ensures the principles of supporting reproducibility, reusability and the FAIR principles of data stewardship are maintained i.e. that all data and tools required to reproduce manuscripts are available to the reviewers and ultimately the readers. To aid GigaDB curators in this step we have been trialling a collaboration with DataSeer. DataSeer uses AI and NLP to scan scientific texts for sentences describing data collection, then gives best-practice advice for sharing that type of data.

Results

During the trial period ~50 papers have been scanned, and curators have compared their own manual pre-review findings with DataSeer's. DataSeer could potentially enhance GigaDB's pre-review stage; not only in reducing the time taken (a significant metric in publishing), but more importantly in identifying data and software that curators may have overlooked.

Conclusion

Regular meetings between GigaDB and DataSeer curators has enabled a deeper understanding of the scanning process and to suggest ways in which it can be enhanced and improved.

Session talk

AIBC-ABS-60

Data standards and data diversity: Lessons and questions from curating cancer data in COSMIC

Rachel Lyne, Madiha Ahmed, Joanna Argasinska, Denise Carvalho-Silva, Alex Holmes, Madhumita, Amaia Sangrador-Vegas and Sari Ward, COSMIC, Wellcome Sanger Institute

COSMIC, the Catalogue of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk), is the world's largest source of expert manually curated somatic mutation information relating to human cancers. Here we focus on two distinct challenges we face - data diversity and data quality.

Ethnic variation in tumour somatic mutations is poorly understood. Greater understanding of such differences and their origins could help interpretation of disease aetiology and influence diagnostic choices. COSMIC data curation focusses on somatic mutations together with a wealth of metadata, that whenever possible includes environmental variables and ethnicity. We prioritise papers that have data from rarer ethnic groups or geographical areas, and we record this data in as much detail as the paper gives. However, problems with both lack of data from globally distinct areas and recording of such data within publications mean creating a truly diverse database is challenging. Despite our efforts, less than 20% of samples in COSMIC have ethnicity data and only around 10% an environmental datapoint. Although we capture over 150 ethnic subgroups, many are still underrepresented.

In addition, around 30% of papers that we triage for curation are found to be un-curatable due to some consistent problems across the biological literature - including not publishing the underlying data and inappropriate formatting.

These are problems faced not just by COSMIC but across the biocuration community. What further steps can we, as a community, take to address these challenging issues? We will highlight such problems and present ideas from the perspective of COSMIC curation.

Session V

Biocuration and Human Health (Seminar hall 2)

Keynote lecture

AIBC-ABS-134

GenomeIndia: A population genomics endeavor

K. Thangaraj, CSIR-Centre for Cellular and Molecular Biology, Hyderabad (On behalf of the GenomeIndia consortium)

The contemporary India is a region of remarkable cultural, linguistic, and genetic diversity with over 4,500 anthropologically well-defined groups. Most of these groups follow endogamy, resulting in a large number of genetic variations, some of which are deleterious and population-specific. Although several countries have initiated genomic studies of their populations, more than decades ago, we recently got an opportunity to sequence 99 diverse populations of India to; create an exhaustive catalog of genetic variations; construct a reference genome for the Indian populations; design genome wide arrays; and create a biobank for future research. Thanks to the Department of Biotechnology, Ministry of Science and Technology, Government of India for supporting this efforts. We have successfully completed whole-genome sequencing (WGS) of 10,000 individuals. Of which, 7800 WGS data is securely archived at IBDC. We performed joint variant calling for 5750 samples and found more than 135 million genetic variations, mostly comprising biallelic single nucleotide polymorphisms (SNVs) and short insertions- deletions (INDELs). Post quality control, we observed several putatively functional variants; more than 1.4 million are missense, frameshift or splice variants, or affect the untranslated regions. We also found about 7 million novel SNVs, which are not found in the global catalogue (dbSNP build156, gnomADv4, 1000Gph3, GAsia). Data obtained from this study would help us in understanding population history, health and diseases; not only for Indian populations, but also for the global community.

Session talk

AIBC-ABS-135

SWADESH: a Multimodal Multi Disease Brain Imaging and Neuropsychological Database and Data Analytics Platform

Pravat K. Mandal *, Komal Jindal, Yashika Arora and Shallu Sharma, National Brain Research Centre, Gurgaon, India

Multimodal neuroimaging data of various brain disorders provides valuable information to understand brain function in health and disease. Various brain imaging databases mainly consist of volumetric magnetic resonance imaging (MRI) data. We present the comprehensive web-based neuroimaging platform "SWADESH" for hosting multi-disease, multimodal neuroimaging, and neuropsychological data along with analytical pipelines. This novel initiative includes neurochemical and magnetic susceptibility data for healthy and diseased conditions, acquired using MR spectroscopy (MRS) and quantitative susceptibility mapping (QSM) respectively. The SWADESH architecture also provides a neuroimaging database which includes MRI, MRS, functional MRI (fMRI), diffusion weighted imaging (DWI), QSM, neuropsychological data and associated data analysis pipelines. Our final objective is to provide a master database of major brain disease states (neurodegenerative, neuropsychiatric, neurodevelopmental, and others) and to identify characteristic features and biomarkers associated with such disorders.

Session talk

AIBC-ABS-136

Digital Information-Systems For Research In Hypoxic-Systems

Dr. Pankaj Khurana, Defence Institute of Physiology and Allied Sciences

Technological advancements in modern biological sciences, has made it a data-rich science. Investigations at cellular and biomolecular level, generates massive high throughput data which is huge and multifaceted in volume, variety and varsity. This rapid explosion in data generation in biological sciences with its inherent increasing complexity necessitates the research to move from hypothesis-driven to a data-driven approach. This exponential growth also necessitates the need for data management, analysis and accessibility.

Biocuration, i.e. organizing, representing and making biological information accessible to both humans and computers, has thus become an essential part of biological discovery in biomedical research. This approach involves integration of large volumes of complex data from varied sources and their analysis, to derive biologically meaningful insights and solving practical problems. Biomolecular research in extreme environment physiology, has generated large volumes of complex data-sets in hypoxia-environments, thus necessitating demand for data integration and management tools. Data resources that enable efficient access, management, storage and retrieval of biological data from experimental studies and equipped with user-friendly data-visualization toolkits which enhances user-experience and augments knowledge generation have been developed for hypoxic-systems. HypoxiaDB, HAHmiRDB, MyomirDB, mirFFLDB, and HighaltitudeomicsDB, the repositories developed for hypoxic-stress systems will be discussed. The various data-analytics and prediction features offered by the resources will also be presented.

Session talk

AIBC-ABS-137

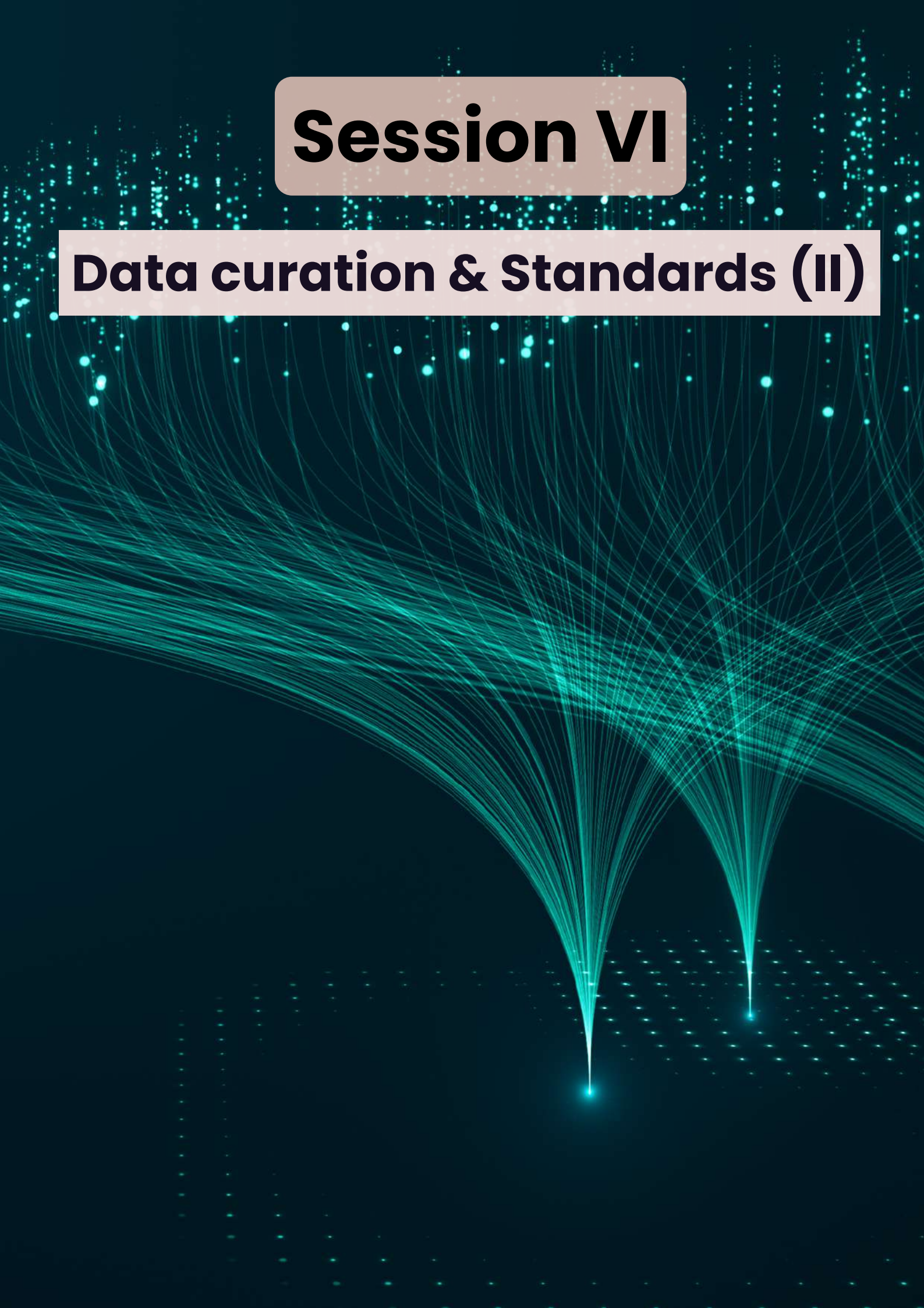
DBT-Centre for Microbial Informatics (DBT-CMI): Genomic exploration of the Indian microbial landscape

Prof H A Nagarajaram, DBT-CMI, School of Life Sciences University of Hyderabad, Hyderabad, Telangana

In view of our centre's objectives to establish a stable bioinformatics platform and provide comprehensive training, our work delves into the exploration of India's unique microbial diversity. We are focused on unveiling the environment-based variations both in the public datasets as well as those generated through collaborations, providing valuable insights into the intricate web of microbial life in diverse ecosystems. Here, we present our initial steps on data collection and curation, along with our proposed roadmap to utilize the industry standard tools and pipelines to explore the genomic intricacies of microorganisms across different geographical regions of the country. Our research aligns with the mission to serve the nation through microbial information and contributes to the development of a skilled workforce in the field.

Session VI

Data curation & Standards (II)



Keynote lecture

AIBC-ABS-64

Plant Reactome Knowledgebase: Decoding Plant Pathway Networks and Unraveling the Genotype-to-Phenotype Connection

Sushma Naithani, Parul Gupta, Justin Elser, Peter D'Eustachio and Pankaj Jaiswal, Oregon State University

Analyzing and connecting genotype to phenotype data at the organismal level poses a major challenge, requiring integration and connecting heterogeneous datasets and their visualization. The Plant Reactome (https://plantreactome.gramene.org) provides a platform that integrates heterogeneous data sources (literature, transcriptome, proteome, metabolome, orthology-based projections) for in silico modeling of system-level plant pathway networks, encompassing metabolic pathways, plant development, reproduction, and genetic-regulatory mechanisms under stress conditions. Plant Reactome serves as a valuable framework for understanding how genes or genotypic differences contribute to phenotypes and impacted by environmental stresses and stinuli. It fosters generation of data-driven hypotheses for understanding intra- and inter-species differences as well as for translational research aimed at crop improvement, precision agriculture, and biodiversity conservation. Our recent efforts focus on using Omics data, particularly rice datasets, to enhance gene/gene family functional annotations and biocuration of gene-gene networks. The platform is freely accessible, facilitating precision breeding, biodiversity conservation, and long-term sustainability research. The Plant Reactome is currently funded by the National Aeronautics and Space Administration (NASA), USA [80NSSC22K0891], USDA-ARS, National Science Foundation, USA [2029854], and in-kind support in the curation and hosting of the web server is provided by the Human Reactome, funded by the NIH NHRGI [U24 HG012198].

Session talk

AIBC-ABS-98

Semantic Web technologies enabling integration of glycoscience data in the GlyCosmos Portal

Kiyoko F. Aoki-Kinoshita, Glycan and Life Systems Integration Center, Soka University

The GlyCosmos Glycoscience Portal has been developed using Semantic Web technologies to integrate the vast amount of heterogeneous data related to glycans, or carbohydrate sugar chains. Because glycans are found both in free form but also attached to lipids, proteins, and other small molecules, they form complex glycoconjugates. These glycoconjugates are found on all cell surfaces, and their variety is important in various biological processes. For example, they are known to change their glycosylation patterns based on their environment. In order to capture the variety and function of glycoconjugates, we have developed ontologies and RDFized a plethora of data, resulting in the GlyCosmos portal, which is available at https://glycosmos.org. User interfaces and APIs are also available, and we have also recently improved the consistency of the data that we have integrated. Moreover, we have developed a new repository called GlyComb such that glycopeptides and glycoproteins can also be identified with accession numbers. Using this, we could identify common glycoproteins across different samples.

Session talk

AIBC-ABS-72

Enabling FAIR access to single-cell RNA-Seq data for reproducible analyses

Frederic Bastian, SIB Swiss Institute of Bioinformatics

Access to single-cell RNA-Seq (scRNA-Seq) data is currently unstandardized, and presents significant challenges. Key information such as cell barcodes or cell type labeling is often difficult to retrieve, hindering our ability to access, validate, and reproduce findings in single-cell biology. These major problems are being tackled by the Bgee database and the scFAIR consortium, for enabling FAIR access to single-cell RNA-Seq data.

Bgee is a Global Core Biodata Resource (GCBR) that addresses these challenges by offering expertly curated and FAIR scRNA-Seq data, to facilitate comparative studies of gene expression between animal species. Bgee leads the scFAIR consortium, that has the aim to standardize access to scRNA-Seq data for reproducible analyses. This consortium includes participants from major resources, such as CELLxGENE, Flybase, ASAP, Wellcome Trust Sanger Institute.

The consortium's initiatives include: advertising the existence of expertly annotated scRNA-Seq datasets in different resources through public APIs; coordinating annotations to prevent duplication of curation work; and establishing common ontologies and metadata schema for curation across animal species. scFAIR also focuses on capturing provenance information about cell labels, from, e.g., the marker genes used, to the analysis parameters allowing the clustering of cells.

This presentation will showcase the advantages of the scFAIR approach for overcoming current barriers in scRNA-Seq data access, and invite collaboration in this community effort.
Bgee is available at https://www.bgee.org/. The scFAIR consortium project description is accessible at https://sc-fair.org/.

Session talk

AIBC-ABS-114

GlyGen: Advancing Glycoscience through Integrated Knowledge Discovery

Jeet Vora, The George Washington University

Background:
Glycans and glycosylation significantly influence protein structure, stability, and function, impacting essential cellular processes, health and physiological development whereas their dysregulation is linked to diseases, emphasizing their biomedical importance. However, the complex structures of glycans and scattered glycobiology data present exploration challenges. To bridge this gap, GlyGen, a glycoinformatics knowledgebase, is curating and integrating diverse data types, including glycan structures, glycoproteins, and enzymes, into its portal.

Methods:
GlyGen collaborates with international databases like the EBI, NCBI, PDB, and UniProt to retrieve data. Additionally, it employs a multi-step approach for data curation, including manual extraction of glycosylation data from peer-reviewed publications and manual identification of glycan structures using the SNFG nomenclature. Furthermore, a literature mining tool is being developed that utilizes machine-learning and natural-language-processing methodologies to automatically curate glycosylation data from peer-reviewed publications.

Results:
GlyGen facilitates glycan and glycoconjugate data exploration via its user-friendly portal (https://glygen.org), while APIs (https://api.glygen.org) and the SPARQL endpoint (https://sparql.glygen.org) facilitate both browsing and programmatic access. Processed datasets at (http://data.glygen.org) include BioCompute Objects with detailed processing documentation. GlyGen's curation efforts have produced glycan and glycoprotein datasets whereas its by-products such as the Glycan dictionary, GlyGen Sandbox, and GNome are supporting data harmonization, biocuration, aiding in text mining, enhancing search functionalities, and contributing to mapping glycan structures to function and disease, benefiting the community.

Conclusion:
GlyGen, a pioneering knowledgebase in glycoscience, offers diverse data and a feature rich portal. It is actively implementing innovative methods like literature mining and machine learning for data curation, transforming the field.

Flash talk

AIBC-ABS-29

Leveraging Wikidata to improve OBO Foundry ontologies

Tiago Lubiana,University of São Paulo

IThe ontologies in the OBO Foundry play a pivotal role in biological data annotation, but keeping them up to date is an enormous task. For example, many terms for cell types needed by the growing scRNA-seq community are not yet available in the Cell Ontology. Besides missing terms, OBO Foundry ontologies also have limited mapping to external resources and lack labels in other languages than English.

In this work, we explored how Wikidata can improve OBO Foundry ontologies. Wikidata is an openly editable knowledge graph with tight links to Wikipedia, with support for over 500 languages and harboring over 100 million concepts. With over 20.000 monthly editors, Wikidata provides a socio-technical environment for fast collection of structured knowledge.

We evaluate the interplay between Wikidata and biomedical ontologies, detailing the coverage on Wikidata and analyzing the prevalence of Wikipedia cross-references in ontologies. Taking the Cell Ontology (CL) as an example, we cataloged over 6000 cell types in Wikidata and manually curated over 2600 mappings on Wikidata to all CL classes. Via SPARQL queries, we retrieved a list of candidate terms for curation in CL, as well as queries to check the modeling consistency between platforms. We also discuss how Wikidata can be leveraged by semi-automatic workflows to add Wikipedia links and provide multi language coverage for CL. We present the work as a path for ontologies to evolve a symbiotic relation with Wikidata, benefitting from its crowd-curated content.

Flash talk

AIBC-ABS-57

Towards Synthetic Cell Types: curating genes, cells, and functions for open-source innovation in direct reprogramming and industrial cell lines

Thea Gayatri Fennell, MRC Laboratory of Molecular Biology, University of Cambridge

Transcriptome reprogramming is direct conversion between cell types, typically by transcription factor (TF) overexpression. TFs can be selected by algorithms (e.g. Mogrify) but such tools presuppose natural, well-characterised cell types – precluding synthetic types, here defined as cells with non-natural combinations of function. Limitations arise with reliance on regulatory networks from select species (e.g. mouse, human) and on target cell RNA-seq data as input. Overcoming these has industrial value; our study was initiated to redesign the popular, hamster-derived CHOK1 line for better antibody secretion.

Addressing cross-species limitations, we piloted solutions on hamster marrow transcriptomes (10X). Through curation of orthologues from reciprocal best-hits, integrating multiple sequence types, half of hamster protein-coding genes were mapped. This facilitated reference-based cell annotation using Tabula Muris Senis (SAMap), highlighting conserved cell types and markers, e.g. Pax5+ B-cells – and enabling Mogrify predictions for conversions from CHOK1 to mouse, human, and hamster plasma cells. Result comparison revealed a 70% overlap in top predicted TFs but preserved biological difference in members' relative importance within a key family of immune regulators. Independently, functional module analysis (scMiko) across hamster bone marrow produced a literature-consistent plasma cell profile. This suggests the possibility of replacing target cell RNA-seq with inputs constructed from non-natural combinations of naturally occurring co-expression modules.

We conclude that transcriptomic reprogramming can be generalised outside model species, given sufficient data for inter-species triangulation. We further propose that synthetic cell types could be designed, with reference to a curated, open-source atlas charting functional module distribution across cell types in relevant species.

Flash talk

AIBC-ABS-101

Increasing data evidence by text-fragment references using nanopublications

Ulrike Wittig1, Mihail Anton 2, Alexandre Flament 3, Matt Jeffryes 4, Luana Licata 5, Patrick Ruch 3, Vincent Emonet 6, Toshiaki Katayama 7, Adel Bouhraoua 8,1 Heidelberg Institute for Theoretical Studies, Heidelberg, Germany, 2 Department of Life Sciences, National Bioinformatics Infrastructure Sweden, SciLifeLab, Chalmers University of Technology, Gothenburg, Sweden, 3 HES-SO & SIB Swiss Institute of Bioinformatics, Geneva, Switzerland, 4 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), UK, 5 Department of Biology, University of Rome Tor Vergata, 00133 Rome, Italy, 6 Institute of Data Science, Maastricht University, Netherlands, 7 BioData Science Initiative & Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan, 8 Department of Biomedical Sciences, University of Padova, Italy

Biocuration includes the extraction and integration of data from published literature into databases. A critical requirement of biocuration is linking evidentiary statements to their source. Currently, databases either refer to the whole publication or manually insert literature statements. Instead, an increased value could be realized by pointing to specific text fragments within and across articles, especially if this link can be packaged in a reusable and attributable way. We analysed the current state of biocuration requirements of different biological databases for referencing to the original sources of information.

To allow references to text-fragments of full text publications we are combining two existing technologies: URL text fragments and nanopublications. URL text fragments are a form of ‘deep linking’ supported by most of the web browsers. They link to specific sections in a web page by extending the URL of the page with the selected text. The link of the URL text fragment jumps directly to the selected position of the page. Nanopublications allow publication of scientific data in a granular and machine-readable format. They consist of three parts: assertion, provenance, and publication information. Persistent identifiers are assigned to create a traceable record and to allow sharing, citing, and reusing of the data published in nanopublications.

Here we demonstrate the advantages of text-fragment based biocuration to assist curators in maximising the reliability and utility of the data provided in the databases. For database users it enhances the accuracy and findability of data and their corresponding references.

Flash talk

AIBC-ABS-50

ChatGPT usage in Reactome curation process

Krishna Tiwari1,5, Lisa Matthews2, Bruce May3, Veronica Shamovsky2, Marija Milacic3, Karen Rothfels3, Eliot Ragueneau1,5, Chuqiao Gong1,5, Ralf Stephan3, Nancy Li3, Guanming Wu4, Lincoln Stein3, Peter D'Eustachio2, Henning Hermjakob1,5,1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK 2 NYU Grossman School of Medicine, New York, NY 10016, USA 3 Ontario Institute for Cancer Research, Toronto, Ontario, M5G 0A3, Canada 4 Oregon Health and Science University, Portland, OR 97239, USA 5 Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

Background: Reactome data curation is a meticulous manual, time-intensive process involving data gathering and annotation in a defined data model. To align with the rapid application of large language models, we did a pilot study to explore the potential of using LLMs, particularly ChatGPT for the Reactome curation process.

Methods: The study aimed to test ChatGPT/GPT4 usage in typical Reactome curation tasks like enriching pathway information (Circadian clock), identifying new protein participants, and generating text summaries for pathways. We prompted ChatGPT with questions and generated answers which were verified manually by expert curators. We additionally took a manual approach to extract similar information and compared the two approaches to estimate the added value of using ChatGPT.

Results: Upon manual verification of the ChatGPT-generated content, we found 40% to 45% accuracy with existing protein functions with 54% accuracy of new protein functional details. Over 90% of cited references were fabricated and 55% to 95% of UniProt IDs were inaccurate. The time taken to extract usable details for a protein by both approaches was very similar (ChatGPT = 42 v/s manual = 46 (minutes)). Generated pathway summaries were incomplete with inaccurate citations until in-text citations were provided.

Conclusion: We gained new annotable proteins through ChatGPT but both approaches took a similar amount of time and effort to generate accurate information with the right literature references. The identified limitation with literature citations and accuracy of the information in this study restricts its direct implementation in the Reactome curation process at present.

Flash talk

AIBC-ABS-28

An integrated literature search, triage and extraction workflow for biocuration

Matt Jeffryes, Henning Hermjakob, Melissa Harrison,EMBL-EBI

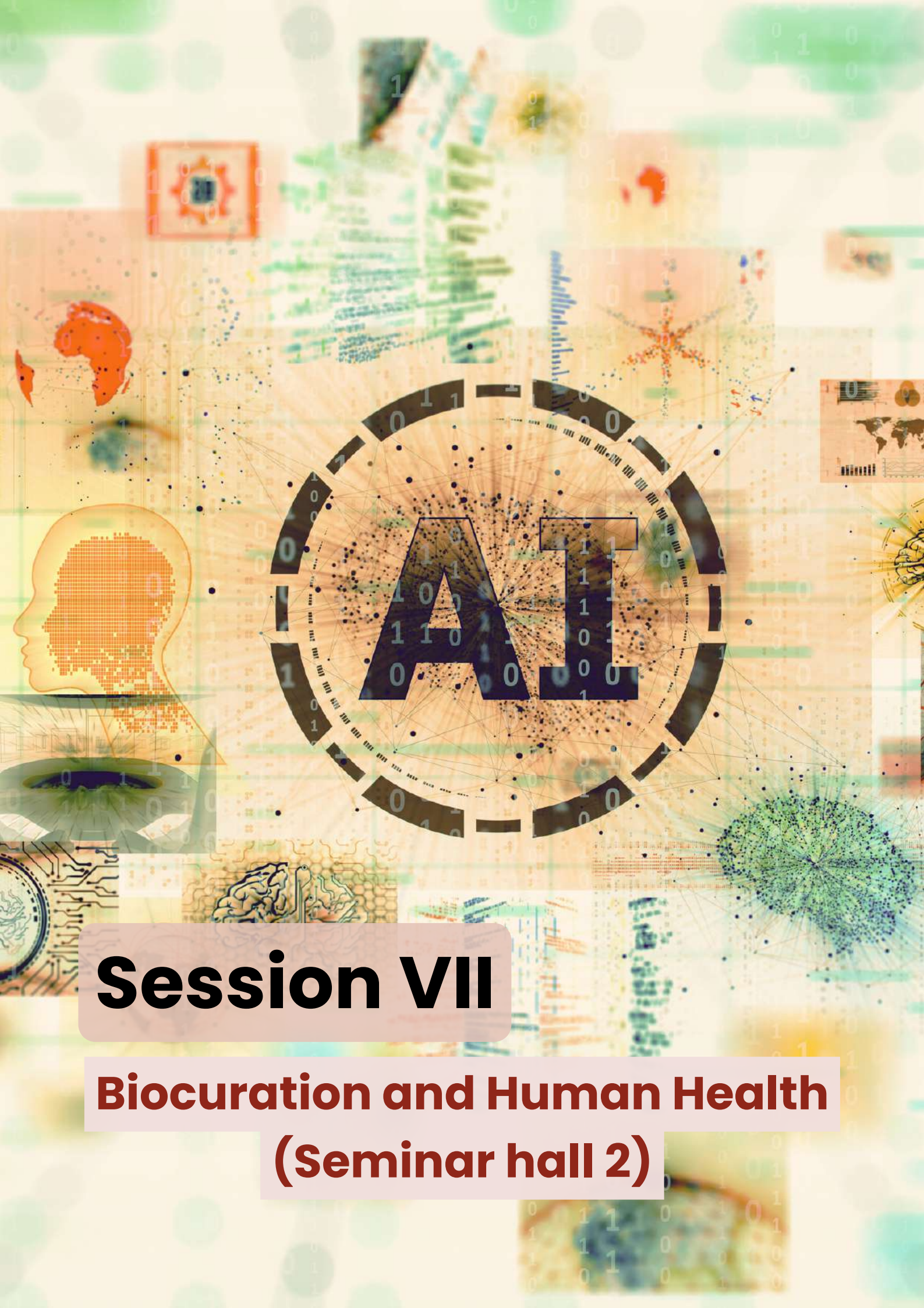
Biocurators spend a large amount of time searching for, reading, and extracting information from the biomedical literature. This process typically involves multiple different websites or applications. We have developed a 'Biocuration toolbox' which combines search with triage and extraction in an integrated workflow.

As well as providing search across titles, abstracts and any full text articles available in Europe PMC, filtering can be performed using Europe PMC's collection of annotations, for entities such as genes or proteins, species or disease names. Searching can be optionally performed across the entire Europe PMC database, covering PubMed, PubMed Central, abstracts from 31 preprint servers, and ~55,000 full text preprints from our Covid-19 corpus and Europe PMC funders.

We have developed the filtering features in collaboration with the curators of the IntAct molecular interaction database, enabling them to filter out articles which report interactions between proteins which are already covered in the database. But the feature is flexible enough to permit filtering in or out based on any of the annotations present in Europe PMC.

Articles can be selected or rejected using a highly customisable lists feature, and a reading view allows curators to highlight and add notes to the article, similar to features available in PDF readers. Integration of these features in a single application offers the potential of improvements in efficiency.

Based on feedback from biocurators, we intend to add further filtering options, and to enable sharing of highlights and notes within curation teams.



Session VII

Biocuration and Human Health

(Seminar hall 2)

Keynote lecture

AIBC-ABS-138

Establishing and Sustaining Public Databases in Our Group

Gajendra P. S. Raghava, Department Computational Biology, IIIT, Delhi, India

Over the past two decades, our group has made significant strides in the realm of biomedical sciences by developing over 70 meticulously curated databases. Each of these databases represents a laborious effort in manual curation, where we meticulously extract information from a multitude of sources including literature, public domain repositories, and various other resources. What sets our databases apart is not just their creation, but their active usage and citation within the scientific community. They serve as invaluable resources, facilitating groundbreaking research and fostering collaboration among researchers worldwide. During my presentation, I will discuss some of our major repositories, providing detailed insights into their structure, content, and significance within the scientific landscape. Additionally, I will shed light on challenges in maintaining and updating these repositories. While the initial creation of these databases is undoubtedly a monumental task, it is very small in comparison to the ongoing challenge of ensuring their updation and accessibility, particularly within the Indian environment. Navigating through the ever-evolving landscape of biological data presents a host of challenges, ranging from technological constraints to the constant influx of new information. Furthermore, the dynamic nature of scientific research demands continuous updates and enhancements to keep pace with the latest discoveries and advancements. Despite these challenges, we strive to uphold the integrity and utility of our databases, thereby contributing to the advancement of biomedical sciences on a global scale.

Further Reading

https://webs.iiitd.edu.in/raghava/webservices.html

Session talk

AIBC-ABS-58

The Genomic Standards Consortium shift to using Link-ML

Chris Hunter on behalf of the Genomic Standards Consortium, Genomic Standards Consortium

The Genomic Standards Consortium (GSC) is an open-membership working body formed in 2005. The aim of the GSC is to make genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards.

The GSC devises and supports a suite of standards, referred to as MIxS (pronounced Mix-ess), which stands for Minimum Information about any Sequence. This suite currently consists of 6 checklists that can be combined with any one of 22 extensions depending on the sampling environment.

Historically the MIxS checklists and extensions have been maintained and distributed in Excel files. Recently, we have made the jump to utilising a more technical approach; the Linked Data Modeling Language (Link-ML). The LinkML implementation means that the checklists are now fully machine readable and implementable and the standards are available in multiple formats (e.g. csv, yaml, owl). As part of the transition, every individual term created in the MIxS suite now also has a unique digital object ID via w3id.org, meaning that they can be integrated into RDF representations.

Here I will present the advantages of the recent transition and outline some of the future progressions that the Link-ML model will afford the GSC.

Session talk

AIBC-ABS-139

Curation of Imaging datasets and databases

Dr Dinesh Gupta, Translational Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), New Delhi, India.

Recent developments in high throughput data generation of biological images, cheaper computation power, and storage have led to challenges in data curation of image datasets. The problem of curation of image datasets is compounded by the diversity of image datasets due to the variety of sources and instruments used to acquire images. No single expert or committee of experts can efficiently curate all kinds of images. Moreover, currently, there is no single standard or pipeline for curating images deposited in various databases. The recent research on the use of Artificial intelligence and data science for image classification has established tremendous potential for developing practical applications. However, the efficiency and utility of these applications will depend on the associated metadata, quality, explanatory potential, reliability, level of annotation and curation of the datasets being used. The talk will summarize the current challenges in image data curation and review the current standards and guidelines followed and proposed by various research groups and database curators.

Session talk

AIBC-ABS-140

Deep Learning the Terpenome: Data Semantification using AI over NLP

Dr. Gitanjali Yadav, National Institute of Plant Genome Research, New Delhi, India

For thousands of years plant-derived natural products have been harvested for their medicinal properties in an effort that is now called Bioprospecting. The distinctive chemical repertoire can also make plants successful in long-distance dispersal to regions where plants have not previously occurred, followed by rapid multiplication and range expansion. Terpenes represent one of the largest classes of functionally and structurally diverse phytochemicals, derived from five-carbon isoprene units assembled and modified in thousands of ways such that the full complement of genes involved in terpene biosynthesis is now called the 'Terpenome'. Understanding the mechanism by which few initial substrates are converted into a tremendous chemical arsenal holds promise for improving agro-biochemical and pharmacological potential of plants. My talk will focus on how we unpack plant chemical arsenals, using a combination of Genomics, Natural language Processing and headspace chromatography. Integration of chemical ecology, biogeography and phytochemistry helps us associate these factors to rapid adaptive changes in DNA/RNA/amino acids, in order to understand whether, and to what extent, changes in the terpenome may regulate plant adaptation.

The Talk will underscore Open Notebook Science is a powerful accelerant for finding solutions to our most pressing global challenges, enabling worldwide collaboration to overcome barriers of cost, culture, and policy — that keep knowledge from flowing among scientists and the society. Towards this, we welcome everyone to join our citizen science movement, #semanticClimate, led by Indian students, who have built an open python-based toolkit that converts PDFs into semantic, hypermedia form, embedded in the Global Knowledge Graph. #seman+cClimate is designed so the barrier to entry is zero. We believe in making Knowledge freely accessible to all and are catalysing the building of a Global Semantic Knowledge Commons for phytochemical data, a self-improving network of semantic tools. This is a very powerful idea and can transform the way the world collects and uses knowledge.

Session talk

AIBC-ABS-141

Integration, harmonization and extrapolation of legacy gene expression data sets using deep learning

Shandar Ahmad, Shruti Gupta and Ajay Kumar Verma, SciWhyLab, School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi

Legacy gene expression data deposited in global repositories such as Gene Expression Omnibus (GEO) and Array Express hold millions of gene expression profiles, harboring answers to complex biological questions which have never been asked in the context and quantity. Stray meta-analyses on these data sets are cumbersome and often produce incoherent results as they are limited by their inconsistent quantification and incomplete or ambiguous contents. A pre-integrated and systematically expanded universal data set from public sources can give a headstart to gene expression based data-driven biology. We introduce an integrated multi-platform gene expression matrix (MPGEM) sourced from GEO with a rigorous expansion of profiles to a comprehensive gene set in the different considered. MPGEM not only harmonizes, integrates and uniformly quantifies expression profiles with novel robust algorithms but also predicts the expression values for gene sets not quantified in reported experiments using deep learning, thereby providing a unified database of gene expressions at unprecedented scales and accuracy.

Session talk

AIBC-ABS-36

Unlocking Bio-Curation: Harnessing Text Highlighting for Neural Named Entity Models of SABIO-RK database.

Sucheta Ghosh, Maja Rey, Ulrike Wittig, Wolfgang Müller, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Automatic bio-named entity recognition (BioNER) could be vital in developing any biochemical database, like the SABIO-RK database (https://sabiork.h-its.org/).

Background: Although the State-Of-The-Art performance for this task has reached above 0.90 F1-score using BERT-based large-language-models, there is a variation among collections, and the overall performance needs to be improved to rely on automatic methods without humans in the loop. Many best-performing models depend on pruning strategies that remove unrelated semantic information. In this work, we use text chunks highlighted in papers during the manual curation process to focus on the part of the article that is crucial for this BioNER task. Only the most important fraction of the article is highlighted, leaving most of it without highlights.

Methods: Here, we pose this problem as a sequential labeling task. First, we fine-tune the best-performing BERT-based model from the BioCreative challenge for Chemical NER detection using BioNER data from the SABIO-RK database to achieve the baseline model. Next, the highlighted text-span feature is on top of the baseline feature set. The highlighted text-span on an article is subsidiary information because it is helpful for a collaborative curation to understand the purpose. Here, we utilize a multi-head attention mechanism in our transformer framework to include the highlighting information.

Results: Within this work, we investigated whether highlighting that is helpful for human readers is also beneficial for training NER systems. It is a novel approach to leveraging traces of curation work for improving AI systems. We will present the approach together with the first results.

Session talk

AIBC-ABS-51

Classification of Brain Cancer Gene Expression Using RNN Approach

Heba M.Afify, Associate professor

The major contributions of this work are based on the development of RNN and Bayesian optimization hyperparameters for the classification of five classes of brain cancer genes.

This proposed model is the first challenge using this combination of (Bayesian optimization + CNN + RNN) on CuMiDa brain cancer gene expression data. Then, the evaluation of the performance of the suggested model for categorizing brain cancer classes using various criteria was achieved. Thus, gene expression data are used to create a deep learning classification-based-hybrid model that will hold senior promise in the treatment of brain cancer.

Flash talk

AIBC-ABS-45

Cholangiocarcinoma-associated risk factors identified by ontology term enrichment analysis

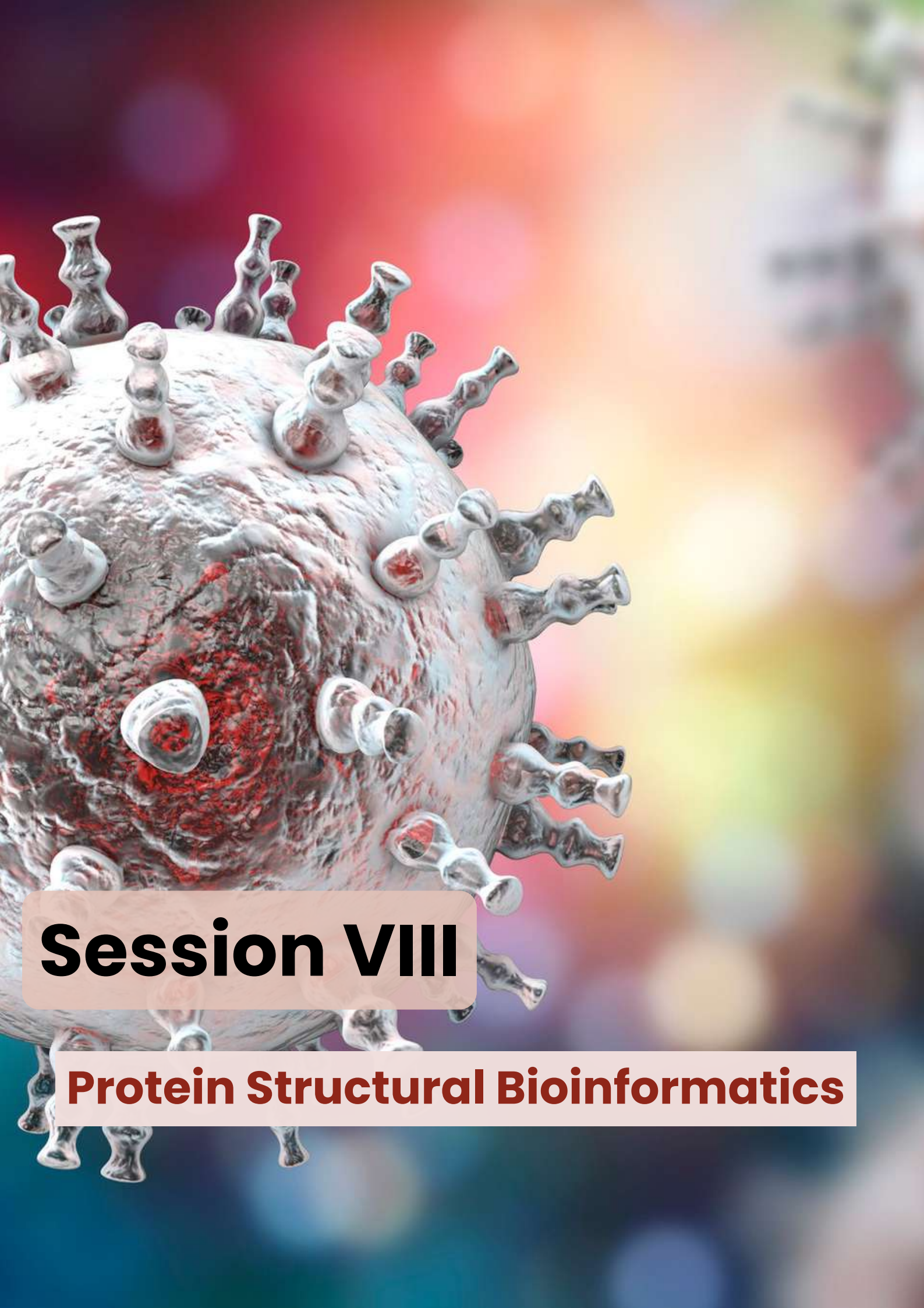
Anuwat Pengput, MPH, MSc, and Alexander D. Diehl, PhD, Department of Biomedical Informatics, University at Buffalo, Buffalo, New York, USA

Background: Cholangiocarcinoma (CCA) is a major public health problem in Southeast Asia. Screening of a research cohort and their associated electronic health records from Thailand provides data about patient demographics and clinical phenotypes with features that have complex relationships. This work integrates and analyzes data related to CCA to investigate risk factors.

Methods: Cholangiocarcinoma Ontology (CCAO) has been developed in OWL format using Protégé based on data about demographics, ultrasound findings, diagnoses and treatments, and post-operative outcomes from areas of Thailand where CCA is endemic due to Opisthorchis viverrini. CCAO also represents CCA diagnosis data from EHRs. Datasets from Thailand were annotated with CCAO and analyzed using term enrichment to compare specific subpopulations to the participant group as a whole.

Results: CCAO has more than 430 classes specific for the CCA, including classes from external ontologies such as BFO, OGMS, and Uberon. Using term enrichment, we have identified significant risk factors in patients with suspected CCA. The results confirmed the main indicators for suspected CCA including findings about dilated bile ducts and periductal fibrosis. We found additional significant echography results that are overrepresented in patients with suspected CCA including findings about hepatic mass, thickening of wall of gall bladder, and jaundice. Moreover, our analysis revealed demographic and lifestyle risk factors from verbal screening of participants.

Conclusions: Our term enrichment results can be tested using other statistical methods and serve as the basis for new studies. The findings can be used to focus and monitor populations at risk for suspected CCA.



Session VIII

Protein Structural Bioinformatics

Keynote lecture

AIBC-ABS-142

Evaluation of AI/ML based methods for prediction of novel folds, oligomeric complexes of proteins and binding affinity of protein-ligand complexes

Sanjeet Kumar Mahtha, Sapna Pal, Ankita Pal and **Debasisa Mohanty**, National Institute of Immunology, New Delhi

The realm of Artificial Intelligence (AI) and Machine Learning (ML) has made remarkable progress in tackling complex biological problems. This spans a spectrum of applications, encompassing sequence analysis, clustering, protein structure predictions, and the estimation of protein-ligand binding affinities. The release of deep learning-based methods, such as AlphaFold, RoseTTAFold, ESMFold etc has significantly advanced structural biology by offering novel solutions to the protein folding challenge as well as in the prediction of protein complexes. Similarly, recent developments in machine learning (ML) have opened up opportunities for developing ML-based scoring functions for binding affinity prediction based on training on available datasets of crystal structures of protein-ligand complexes with known binding affinity values. Nevertheless, there remains a need to investigate the benefits and limitations of current machine-learning methods. In this study, we have evaluated the performance of AI/ML methods like AlphaFold and the language model-based ESMFold on protein sets lacking suitable templates. In addition, we have also developed and benchmarked a novel ML-based scoring function for binding affinity prediction based on training on available datasets of crystal structures of protein-ligand complexes with known binding affinity values.

Our results showed that AlphaFold accurately predicted 84% of monomer structures and ESMFold correctly predicted 74% of monomers. Similarly, in the case of dimers, Alphafold was able to predict 47% of dimers correctly, compared to 23% in the case of ESMFold. For the ML-based scoring function, using extended connectivity interaction fingerprints (ECIF), we were able to achieve a Pearson correlation coefficient (PCC) of 0.85 between experimental and predicted affinities in the CASF-2016 benchmarking test dataset. Overall, our findings demonstrate the capability of AI-based methods in predicting novel protein structures. In addition, we find the ML-based scoring function helps in improving the structure-based virtual screening.

Session talk

AIBC-ABS-88

Reference Proteomes in UniProtKB

Dushyanth Jyothi, European Bioinformatics Institute

The number of sequenced genomes is growing rapidly due to the advancements in the sequencing technologies, and emerging projects aimed at sequencing representatives of each species, meta-genome and pan-genome projects alike.

Organising, identifying and annotating proteomes of a wide variety of important organisms is a huge challenge in the ever-growing “proteome” space. UniProt addresses this by identifying a set of "Reference Proteomes” as "landmarks" considering various factors including well-studied model organisms, organisms important for biomedical & biotechnological research. We work closely with the scientific community to manually select reference proteomes as well as using computational methods to identify proteomes representing broad taxonomic diversity.

To manage the escalating volume of data and uphold the delivery of high-quality, well-annotated, non-redundant proteomes while preserving taxonomic diversity, UniProt is planning to reorganise the UniProt Knowledgebase (UniProtKB). Presently, UniProtKB serves as a comprehensive repository of protein sequences and their functional information. A majority (~96%) of these proteins are linked to proteomes, ~33% are specifically associated with reference proteomes, and the remaining 4% are individual proteins.

UniProt aims to transform the knowledgebase into a more concise resource, offering a high-quality and most useful dataset. This entails focusing on proteins from reference proteomes, manually curated proteins, and other pivotal proteins featuring structures, references, studies, etc. The envisioned transformation anticipates a reduction of at least ~65% in the current knowledgebase. Access to all other proteins not represented in UniProtKB will still be available through UniParc. We are seeking feedback on how the suggested reorganization may impact the curation community.

Session talk

AIBC-ABS-68

PDB: Improved and enriched biodata repository serving many millions of users worldwide

Brinda Vallat, Zukang Feng, Vladimir Guranovic, Ezra Peisach, Dennis W. Piehl, Chenghua Shao, Jasmine Y. Young, Stephen K. Burley, and the wwPDB Consortium, Research Collaboratory for Structural Bioinformatics Protein Data Bank and the Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

The Protein Data Bank (PDB) was established in 1971 as the first open access digital data resource in biology. It has grown more than 30,000-fold to become the single global archive of experimentally-determined structures of proteins, nucleic acids, macromolecular machines, and viruses. The PDB is managed according to the Findable, Accessible, Interoperable, and Reusable (FAIR) Principles by the Worldwide Protein Data Bank partnership (wwPDB; wwpdb.org), an international consortium that collaboratively oversees deposition, validation, biocuration, and open access dissemination of three-dimensional (3D) macromolecular structure data. These data are central to understanding processes that make life possible, because function follows form in biology. Knowledge of form, shape, or 3D structure at the atomic-level enables functional understanding of biological macromolecules, facilitates design of vaccines, therapeutic antibodies, and other biologics, discovery of small-molecule drugs, and development of medical diagnostics.

Herein, we report recent improvements and enrichment of the PDB archive, including remediation efforts to improve data quality, creation of a versioned PDB archive to preserve the historical record, development of the next generation PDB archive (PDB NextGen) to provide access to frequently updated external annotations from trusted sources, and future plans to support a revised format for PDB accession codes. Together, these developments will enhance and strengthen the growing PDB archive as it continues to deliver 3D biostructure information to many millions of users worldwide at no charge and with no limitations on data usage.

Session talk

AIBC-ABS-68

Curating the biological functions of proteins using the AlphaFold Protein Structure Database

Maxim Tsenkov, Protein Data Bank in Europe (PDBe) at EMBL's European Bioinformatics Institute (EBI)

The advent of AI-based tools like AlphaFold has led to a flood of almost a billion predicted protein structure models hosted in data resources such as the AlphaFold Protein Structure Database (AFDB) and the ESM Metagenomic Atlas. This massive influx of data poses significant challenges; chief amongst them is characterising and functionally interpreting these structures.

To tackle this challenge, we used a novel algorithm (Foldseek Cluster), in collaboration with the Steinegger group, to cluster all the 214 million+ predicted protein structures in AFDB based on sequence and structure similarity. We also added structure-based search against AFDB and the Protein Data Bank, powered by Foldseek and sequence-based search, powered by BLAST, to our services.

These new, similarity-based tools aid biocurators in efficiently transferring annotations from functionally well-characterised proteins to proteins with unknown functions, especially when conserved domains are interrupted by insertions and deletions. In addition, structure-based search can uncover remote homology and help annotators explore structural diversity over evolutionary timescales, aiding functional inference amidst low sequence similarity.

Enriching the AlphaFold database with structure similarity-based clusters and similarity search functionalities created a robust platform for biocurators. Core data resources such as InterPro, PDBe and UniProt already take advantage of this platform to facilitate the functional characterisation of proteins. Moving forward, we are committed to empowering the broader community of biocurators in unveiling the functions of the protein universe.

Session talk

AIBC-ABS-24

Molecular interactions in the context of Rare diseases: Annotation rich dataset from the IMEx consortium

Kalpana Panneerselvam, European Bioinformatics Institute

Background: Molecular interaction networks provide maps to explore cellular processes from a systems perspective. Access to the experimental evidence and contextual metadata which influence the interaction outcome are critical for accurate interpretation of interaction dynamics. Method: The IMEx consortium (www.imexconsortium.org) is an international database collaboration that exists to represent molecular interaction data to full experimental detail from scientific literature and make it freely accessible to the public, using an open data model. We are providing a Rare disease dataset of 12,000 molecular interactions with a focus on mutations associated with rare diseases. Result: The dataset contains both binary and n-ary associations (where mutant bait binds differently with multiple prey compared to the wild-type). Host organisms, where the interactions get detected are mapped to ontologies representing tissue/cell lines, diseases and developmental stages. Interactors involve proteins, complexes, nucleic acid and small molecules, categorised according to their experimental and biological roles.

Each binary interaction is assigned a confidence score based on interaction type, detection methods and publications reporting the interaction. Additionally, factors affecting interactions such as mutations, binding regions, PTMs, agonists and antagonists are captured from evidence. For quality control, we run pipelines during every release to ensure mutations and binding domains are in sync with latest UniProtKB sequence updates.

Conclusion: IMEx's mutations are integrated with UniProt, Ensembl's VEP and more recently with AlphaFold predictive models. The resource is an invaluable tool with applications in the study of variation impact on the interactome, interaction interfaces and previously un-annotated variants.

Flash talk

AIBC-ABS-17

Standardisation of protein modification data in the Protein Data Bank (PDB) archive

Marcus Bage, Deborah Harrus, wwPDB Consortium, EMBL-EBI

The PDB archive contains over 200,000 macromolecular structures, of which >80,000 contain a protein modification. These modifications are highly diverse and result from both natural and synthetic processes. The variety of chemical groups ranges from small compounds (e.g. methylation) to very large polymeric compounds (e.g. glycosylation). Due to this variety, it is challenging to represent all protein modifications consistently within the PDB archive. The standardisation of protein modifications will make the data more FAIR and support the enrichment of protein modification data in the PDB archive.

The wwPDB consortium is working on creating a standardised approach to annotating protein modifications. This approach involves introducing a rules-based annotation procedure to ensure the consistent handling of new protein modifications deposited in the PDB archive. Protein modifications are grouped into distinct modification categories, which dictate their handling. Backbone atoms for all peptide residues have been annotated and modifications are cross-referenced to the UniProt PTM controlled vocabulary, enabling the identification of natural post-translational modifications. Using controlled vocabulary and standard representation of protein modification information across all the PDB entries will make protein modification data Findable, Accessible, Interoperable and Reusable (FAIR).

The PTM remediation project is a wwPDB collaborative project carried out mainly by the Protein Data Bank in Europe (PDBe) at the EMBL-EBI and is funded by the Biotechnology and Biological Sciences Research Council (BBSRC), UK.

Flash talk

AIBC-ABS-32

Improved findability of small molecule data in the PDB

Ibrahim Roshan Kunnakkattu , Preeti Choudhary, Lukas Pravda, Nurul Nadzirin, Oliver S. Smart, Qi Yuan, Stephen Anyango, Sreenath Nair, Mihaly Varadi, Sameer Velankar,1 Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

The Protein Data Bank in Europe (PDBe) is a founding member of the Worldwide Protein Data Bank (wwPDB). PDBe actively curates, validates, archives, and disseminates macromolecular structure data. It provides enriched PDB data, tools and services for easy data access and visualisation of macromolecular structures and their interaction with small molecules. Here, we describe improvements to the representation of complex ligands in the PDB and additional data added to the wwPDB small molecule reference dictionaries for improving findability.

The wwPDB Chemical Component Dictionary (CCD) represents the reference data for all unique small molecule components in the PDB. Due to current wwPDB annotation practices, some of the ligands in the PDB are split into smaller chemical components for data archival purposes. This fragmentation results in incomplete ligand representation, limiting their functional interpretation. We addressed this limitation by identifying fragmented ligands across the PDB archive, creating their complete representations, assigning unique identifiers and consolidating them in a complementary reference dictionary, the Covalently Linked Components (CLC). By October 2023, we released 2,731 CLCs represented in 10,236 PDB entries.

The PDBe team provides additional data for both CCDs and newly defined CLCs. This includes 3D conformers, 2D coordinates, physicochemical properties, matching scaffolds and fragments and cross-references to common cheminformatic resources. The scientific community can access this enhanced data from the PDBe FTP area at https://ftp.ebi.ac.uk/pub/databases/msd/pdbechem_v2/. These advancements improve the completeness and accuracy of ligand representations in the PDB and enhance the accessibility of data.

Flash talk

AIBC-ABS-33

DrugMechDB: A Curated Database of Drug Mechanisms

Adriana Carolina Gonzalez-Cavazos, Anna Tanska, Michael D Mayers, Denise Carvalho-Silva, Brindha Sridharan, Patrik A Rewers, Umasri Sankarlal, Lakshmanan Jagannathan, Andrew I Su,The Scripps Research Institute, Department of Integrative and Structural Biology, 10550 N Torrey Pines Rd. La Jolla, CA, 92037, USA

DrugMechDB is a comprehensive manually curated resource that can be employed as a benchmark dataset for assessing computational drug repurposing models or as a valuable resource for training such models. This database is created with a detailed graph representation of mechanisms of action that serves as a reference point to evaluate the mechanistic accuracy of predictions made by repositioning models and is a useful resource for understanding drug pharmacodynamics. Data standardization in DrugMechDB adopts the Biolink Model that enables the mapping of concepts and relationships to a common vocabulary thus allowing interoperability between various data sources.

Session talk

AIBC-ABS-143

Design considerations for securely storing and efficiently accessing large biological datasets on AWS cloud

Mainak Chakraborty, AWS

AWS matches the needs of biocurators and research organizations with innovative technologies to provide scalable, secure, and cost-effective tools that accelerate genomic discoveries. From solutions to migrate and securely store genomic data on AWS, to tools that accelerate secondary and tertiary analysis, to services that integrate genomic data into multi-modal datasets, AWS for Lifesciences offers solutions across the genomics workflows. For almost a decade, industry leaders such as Illumina, Genomics England, University of Chicago, UK Biobank and others have leveraged AWS to do more with their data. In this talk, will discuss design considerations for securely storing and efficiently accessing large biological datasets on AWS cloud using examples from UK Biobank and University of Chicago.

Session IX

Translating Biocuration: Applications in Agriculture and Healthcare

Keynote lecture

AIBC-ABS-144

Integrated Computing Environment for Next Generation Biology: Cloud based HPC and Big Data Platform

Sandeep Malviya, Vivek Gavane, E.P. Ramakrishnan, Renu Gadhari, Neeraj Bharti, Palash Pullarwar, Prachi Barkale, Preet Jamsandekar, Pallavi Niturkar, Sunitha Manjari Kasibhatla, Uddhavesh Sonawane, **Rajendra Joshi***, HPC-M&BA Group, Centre for Development of Advanced Computing, Innovation Park, Panchwati, Pashan, Pune-411008

Advances in DNA sequencing technology have enabled whole-genome sequencing become affordable and scalable for extensive use in pathogen surveillance and for tackling public health challenges globally. Vast amounts of data from various life sciences domains, including genomics and molecular dynamics simulations are being generated at a high-throughput scale. The utilization of a centralized compute and storage facility on a cloud platform fosters collaborative research, facilitating further progress in these fields. This centralized facility serves as a hub for secure data storage and knowledge sharing. The current challenge has shifted from data generation to secure storage and analysis of this extensive data within a computational environment.

To address these challenges, we have developed the 'Integrated Computing Environment' tailored for storage and computation of next-generation biological data on cloud and big data platforms. The solution is constructed based on a container orchestration framework, specifically Kubernetes. Operating on microservices architecture, this solution ensures scalability, high availability, data security, and fault tolerance. The storage module incorporates role-based access control, facilitating secure data sharing. The compute module enables users to execute any publicly available Docker containers on this platform. Additionally, a Kubernetes-native application has been seamlessly integrated to support large-scale comparisons of variant files generated by population genomics projects. The VCF analysis module allows dynamic comparisons of samples from different populations, identifying both unique and common variants. This comprehensive approach addresses the evolving landscape of computational biology, emphasizing secure and efficient data storage and analysis for cutting-edge research.

Session talk

AIBC-ABS-145

Stress Combinations and their Interactions in Plants Database

Muthappa Senthil-Kumar, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India.

The Stress Combinations and their Interactions in Plants Database (SCIPDb; http://www.nipgr.ac.in/scipdb.php) is a valuable platform for comprehending plant responses to various stresses. This resource, a team effort, facilitates the exploration of plant phenome, transcriptome, metabolome, and gene trait data, providing a holistic view of stress interactions. In the current era of data-centric agricultural science research, SCIPDb underscores the importance of inclusive and diverse data curation.

SCIPDb, a result of meticulous manual curation, incorporates high-quality information from 123 different stress combinations across 408 organisms and three omics categories. This data is sourced from over 1000 peer-reviewed papers by experienced experts, who conducted thorough literature searches to ensure precision and quality. The manual curation foundation sets the stage for future integration with AI and ML-based tools, forming a crucial part of SCIPDb's evolutionary roadmap.

The integration of phenomics datasets into SCIPDb involved implementing a user-friendly interface using HTML5, CSS, and PHP. The dataset was integrated and presented as an HTML page based on a three-level-dropdown selection designed in JavaScript for each plant species. The back-end schema was developed using MySQL, and Bootstrap 4 and jQuery were utilized for an interactive interface. Visualizations, created using Tableau Public Desktop, included an interactive Treemap/stress matrix illustrating positive or negative outcomes within various stress combinations. Radial trees were developed using Flourish Studio, and various interactive visualization tools such as Chord diagrams and Sankey Diagrams were harnessed. Additionally, an interactive geographical map was generated using Google My Maps.

On behalf of the team, in this talk, I will delve into SCIPDb's capabilities, showcasing its potential to uncover new genes and pathways associated with plant interactions under biotic and abiotic stresses. As an indispensable tool in plant biology and agricultural sciences, SCIPDb paves the way for advancements under changing climates, emphasizing the critical role of stress interaction research in addressing challenges to reach UN-sustainable development goal-SDG13. SCIPDb also serves as an introductory resource for students studying plant-environment interactions, offering videos, podcasts, and updates through its YouTube channel.

Reference:

Priya P, Patil M, Pandey P, Singh A, Babu V, Senthil-Kumar M. (2023). Stress Combinations and their Interactions in Plants Database: A one-stop resource on combined stress responses in plants. The Plant Journal, 116(4) 1097-1117. https://doi.org/10.1111/tpj.16497

Session talk

AIBC-ABS-146

Genomic tools and resources for knowledge-based functional and translational genomics research in crop plants

Mukesh Jain, School of Computational and Integrative Sciences (SCIS), Jawaharlal Nehru University, New Delhi - 110067

Advancements in genomics technologies have revolutionized the landscape of plant biology research to unravel the complexities of plant genomes and accelerate the functional genomics and translation research. I shall present the highlights of diverse genomic tools and resources generated by my group that are pivotal for advancing knowledge-based functional and translational genomics research in rice and chickpea. We developed a machine-learning based tool, Plant Long Non-Coding RNA Prediction by Random fOrests (PLncPRO), for prediction of lncRNAs in plants using transcriptome data. PLncPRO was found to have better prediction accuracy as compared to other existing tools and is well-suited for plants. Consensus models for dicots and monocots were also developed to facilitate prediction of lncRNAs in non-model/orphan plants. Using PLncPRO, we discovered high-confidence lncRNAs in different plants using the transcriptome data in different biological contexts, and studied their regulatory aspects. In addition, we have generated several genomic resources for important agronomic traits, including seed weight/size determination and abiotic stress responses in chickpea and rice, which highlighted transcriptional regulatory networks and important genes/pathways. These tools and resources collectively contribute to unraveling the genetic basis of traits and accelerating crop improvement efforts.

Session talk

AIBC-ABS-89

Development of a genomics-assisted coreset in sesame from the collections being conserved at the National Genebank of India

Ruperao P, Yadav R, Angamuthu M, Subramani R, Tiwari K, Rai V, Maurya R, Batra T, Kumar A, Pradheep K, Govindasamy S, Jayaraman A, Rathore A, Singh R, Singh K, Singh GP, Angadi UB, Mayes S, Odeny DA, **Rangan P***,1 Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India2 ICAR-National Bureau of Plant Genetic Resources, PUSA Campus, New Delhi 110012, India
3 TNAU-Regional Research Station, Vriddhachalam-606001, India
4 Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar 385506, India
5 ICAR-National Institute of Plant Biotechnology, PUSA Campus, New Delhi 110012, India
6 ICAR-NBPGR Regional Station, Thrissur 680656, India
7 Excellence in Breeding Platform, CIMMYT, Hyderabad 500078, India
8 Genebank, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad 502324, India
9 ICAR-Indian Agricultural Statistical Research Institute, New Delhi 110012, India
10 Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia 4072, Australia

The germplasm that are being conserved at the global Genebanks are undergoing a transformational shift from a primary focus on conservation to one of utilization.This is being mediated through the availability and easy access to Next generation sequencing (NGS). The establishment of a core collection comprises of the minimal set of accessions that down-size the complete collection (through minimizing redundancy and maximizing the genetically diverse representative sets) are key objectives. Here, using genomics-assisted tools, a diverse representative set comprising 1,029 accessions was established from a set of 5,856 sesame accessions being conserved at the national Genebank, ICAR-NBPGR. The complete set of 5,856 accessions were grouped based on various metrics to quantify diversity between the accessions using double-digest restriction-site associated DNA-sequencing (ddRAD-seq) approaches. Computational tools like heterozygosity, k-mer analysis, SNP counts and presence-absence variation, sequence-based genetic distance, PCA, Corehunter3, Powercore, and GenoCore were used to assess the diversity parameters. Firstly, this diverse set of 1,029 accessions was manually curated, and 32 accessions representing wild progenitors and released cultivars were excluded. This brings it to 997 accessions. Secondly, 196 promising accessions expressing various traits of interest were added manually, hence forming an iterative coreset of 1,193 accessions. Further, a composite coreset to augment this genomics-assisted iterative coreset will be developed to represent the phenotype-based core that will also include trait-specific accessions representing for various biotic and abiotic stresses.

Session talk

AIBC-ABS-147

Indian Breast Cancer Genome Atlas: Emerging India-specific Molecular Attributes

Shantanu Chowdhury, Institute of Genomics and Integrative Biology (CSIR–IGIB), New Delhi

This multi-centre program aims genomic profiling of about 1000 breast cancer patient tumors (and matched normal tissue), with 3-year follow-up to characterize potential genomic signature(s) associated with response to therapy. Data from the study comprise whole genome sequencing at relatively high depth, transcriptome and epigenome sequencing. Overall, the program seeks to not only create India-specific cancer genomic resources, but also aid in identifying actionable molecular signatures of clinical significance. In about two years, the program has built a network of more than 10 primary hospitals across India, and profiled 350+ breast tumors along with clinical follow up. Interesting trends in molecular attributes of Indian breast cancers are emerging. These might constitute ethnicity-based signatures of significance in understanding drug tolerance among Indian patients.

Session talk

AIBC-ABS-96

A review of biocuration efforts and alternative approaches in transcriptomics, meta-analysis and other bioinformatic research, across 15 years

Kshitish K Acharya, Institute of Bioinformatics and Applied Biotechnology (IBAB)

Background:
Discrepancies in transcriptomic studies have impeded the potential benefits of meta-analyzing gene expression profiles across various studies. To overcome these limitations, manual curation of datasets proves to be instrumental. Additionally, a manual comparative analysis of bioinformatic resources can contribute to making more objective decisions when selecting the most suitable resources for downstream transcriptomic analysis.

Method:
We initiated a meticulous search and compilation of mammalian gene expression datasets to facilitate efficient meta-analysis of gene expression patterns under specific conditions. Throughout the biocuration process, we closely monitored error rates, sought solutions to address the slow pace of manual curation, and explored alternative and supporting methods. During this exploration, we recognized the necessity for a comparative analysis of the utility of databases and software for literature search or functional analysis of differentially expressed genes. Consequently, we conducted a systematic comparative analysis of bioinformatic resources across multiple areas.

Results:
Curators initially encountered varying degrees of errors during the early stages. However, over time, we successfully developed mammalian gene expression databases that optimized meta-analysis by incorporating enhanced sample sizes. Our studies underscore the meticulous efforts required for manual curation while also providing valuable insights into the advantages of this process. The comparative analysis of bioinformatic resources involved the curation of 'golden sets' for analysis, offering valuable guidance for biologists in objectively selecting databases and software. Overall, our results and observations aim to assist in data collection and subsequent analysis, particularly in the context of transcriptomics.

Session talk

AIBC-ABS-148

Cancer Imaging Biobank

Dr. Swapnil Rane, Tata Memorial Centre (TMC), Mumbai, India

The Cancer Imaging Biobank, is a DBT funded multi-institutional project led by Tata Memorial Centre along with IIT-Bombay, RGCIRC-New Delhi, AIIMS-New Delhi, and PGIMER-Chandigarh. The purpose of this project is to establish resource of anonymized, longitudinally collected Radiology and Pathology Images along with their linked clinical information, outcome data, treatment and other metadata from the point of diagnosis to the point of first progression/recurrence, along with long term survival information such as death. The data is primarily focussed from creating a resource for AI algorithm training, validation, and testing on Indian patients, however there are many use-cases beyond these applications. The project has initially focussed on building the repository for Head Neck Cancers and Lung Cancers (with minimum 1000 patients of each cancer type).

In addition to creating the database, multiple exemplar AI algorithms are being trained and tested using the data being collected. The AI/ML algorithms focus on medically relevant tasks such as Lymph node metastases screening, Nucleus segmentation and classification, biomarker prediction (e.g. HPV in oropharyngeal and EGFR in lung cancer), therapy response prediction, in addition to developing novel methods for crowd sourcing annotations from pathologists and building the pipelines for clinical deployment such as “Slide quality control” and “Cautious AI”.

The portal and project is inspired by the TCGA, TCIA and other similar databases which have been drivers of AI research on Pathology and Radiology Images. Unlike TCGA database, which hosts extensive genetic information, CAIB database focusses on clinical data and images and only hosts routine genetic test information conducted as a part of clinical testing.

The data will eventually be made publicly available alongside marker publications and exemplar AI algorithm results.

References:

1. Cancer Imaging Biobank-AI ready health data: https://caib.actrec.gov.in

Session talk

AIBC-ABS-149

Reading genomes

M.S. Madhusudhan, Indian Institute of Science Education and Research, Pune

Interacting proteins usually bind 5-6 base pairs of DNA. We looked at distributions of 5/6-mer DNA motifs, in whole chromosomes and in smaller sections, across the whole genome to get insights into how proteins read genomic sequences. The distribution of motifs in the genome is non-random as established by their observed to expected (OE) ratios at all examined length scales. We correlated the motif distributions in promoter regions of genes to one another and found evidence for translocations, gene regulatory networks and even data pertaining to spatial proximity of regions between genes. In general, correlating genomic regions by motif distribution comparisons alone is rife with functional information.

POSTERS

Poster

AIBC-ABS-7

Archetype Glycans: a Novel Representation to Organise Glycan Data

Thomas Masding, Akihiro Fujita, Kiyoko Aoki-Kinoshita,Soka University Japan

Glycans play vital roles in biology and disease. However, structural complexity remains a significant challenge in glycan research. Powerful methods capable of describing a broad range of carbohydrates structures have been developed, such as WURCS 2.0 used in the international glycan repository GlyTouCan (glytoucan.org), that assigns globally unique identifiers to any glycan, serving as the only freely available uncurated registry for glycan structures. This has resulted in the registration of many almost identical glycans that have the same chemical structures but differ only at the reducing terminal (e.g. open ring form, anomeric differences, etc.), with each having their own unique GlyTouCan ID. As these groups of similar structures are currently unlinked in GlyTouCan this may lead to uncertainty regarding correct glycan selection and may impact research and analysis. Furthermore, as multiple data resources now use these GlyTouCan identifiers, such as Reactome, ChEBI and PubChem, it places great importance with the organization of these groups of nearly identical structures. This study proposes the new archetype glycan concept which defines a representative glycan for glycans with identical structures but which differ only at their reducing end. The user is thus able to search individual glycans by their archetype representation. This study has used computational methods to calculate each archetype glycan representation, where 137,530 were successfully identified from 217,913 total registered glycans. This new representation, which is being incorporated in GlyTouCan and GlyCosmos (glycosmos.org), aims to help improve correct glycan selection, literature scope and data mining analysis for researchers using these resources.

Poster

AIBC-ABS-8

Unraveling Muscular Dystrophy: A Knowledge Graph Approach to Biomarker Research

Maitreyi Meshram, Tanuj Singh Shekhawat, Parthiban Srinivasan, Indian Institute of Science Education and Research Bhopal

Biomarkers and Knowledge Graphs both play a very important role in today's biomedical world. As of now, there is no such thing that connects biomarkers to a knowledge graph so here we have made a knowledge graph for biomarkers by considering An example of one of a rare diseases is Muscular Dystrophy which can help us link biomarkers to a knowledge graph. Here we consider the 2 main types of Muscular Dystrophy which are Congenital Muscular Dystrophy and Duchenne Muscular Dystrophy. The primary biomarker for both of these diseases is Creatine Kinase (CK). Further, this Biomarker is connected with 12 other relationships and the knowledge graph is made which gives us an idea of linking biomarkers with the knowledge graph The Biomarkers may be used to see how well the body responds to a treatment for a disease or condition or biological processes. On the other hand, Knowledge Graph represents a piece of structured information by integrating and analyzing data and directing them to the appropriate database. And connecting both biomarkers and knowledge graph helps to integrate and analyze various biomedical data and direct various new research and helps in a better understanding of biomarker discovery and identification.

Poster

AIBC-ABS-9

Generative AI and PROTAC concepts to address Sickle Cell Disease

Manoj Kumar, Parthiban Srinivasan, Indian Institute of Science Education and Research, Bhopal

In recent years, drug discovery has embraced two groundbreaking approaches: PROTACs and Molecular Glue, offering hope for treating once-considered undruggable diseases. These methods show promise in rare diseases like sickle cell and cancer.

A PROTAC molecule, consisting of ligands for the target protein and E3 ubiquitin ligase, along with a linking component, selectively marks the target for degradation. Molecular Glue, functioning as a bridge, accommodates weaker binding between the target protein and E3 ligase, enabling protein destruction.

Artificial intelligence, particularly deep generative models, contributes to the creation of novel molecules. This work explores two methods, using Gated Graph Neural Network and a combination of Deep Generative models and Reinforcement Learning, to predict PROTACs.

For a case study, a de novo PROTAC was designed to degrade the BCL11A protein, offering a potential solution for sickle cell disease. The synergy of AI and drug discovery holds promise for targeted therapies in challenging medical conditions, demonstrating the evolving landscape of pharmaceutical research.

Poster

AIBC-ABS-14

FAIRsharing for curators: community engagement and assistant

Allyson, Milo, Ramon, Prakhyat, Susanna, University of Oxford

FAIRsharing is a curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies, across all disciplines. It guides consumers to discover, select and use these resources with confidence, producers to make their resource more discoverable, more widely adopted and cited, and powers third party tools by providing trustworthy content to promote standards and databases. FAIRsharing has two new initiatives that are of direct relevance to the biocuration community: the Community Curation Programme and the FAIRsharing Assistant.

Launched in the summer of 2022, and supported by the RDA FAIRsharing WG and the RDA/EOSC-Future Ambassadorship Programme, the FAIRsharing the Community Champions Programme shows the importance of cultivating and sharing the collective knowledge on standards, databases and policies to map this complex landscape that enables the FAIR Principles. The Champions are a thriving community of domain and discipline experts who:

1) act as advocates to promote the value of standards, databases and policies for digital objects (incl. data, software).
2) create educational material describing these resources helping researchers and other stakeholders to find, use and adopt them.
3) enrich the content of FAIRsharing, adding and enhancing the description and discoverability of these resources.

The FAIRsharing Assistant, a Q& A-style tool, has been designed using VueJs, Javascript and the FAIRsharing API. It offers personalised guidance to those in research and in research-support roles (e.g. researchers, developers, curators, policymakers, data stewards, librarians, standards development organisations, societies) to discover standards and repositories in FAIRsharing.

Poster

AIBC-ABS-16

Advancing Drug Discovery through Knowledge Graphs and Natural Language Processing

Tanuj Singh Shekhawat, Maitreyi Meshram, Dr Parthiban Srinivasan, Indian Institute of Science Education and Research, Bhopal

We have established an extensive repository containing more than one billion records, aggregating essential attributes related to human diseases using web scraping and a variety of ontologies. We extracted meaningful insights from related scientific articles using Name Entity Recognition (NER) and Relationship Extraction (RE) techniques. Cypher queries are executed on the knowledge graph, facilitating the extraction and visualization of valuable information.

Our research focuses on innovative solutions to address the challenges of drug repurposing and safety assessment. We employ graph algorithms and advanced Natural Language Processing (NLP) techniques. Specifically, our work centres on muscular dystrophy, serving as a test case for our novel methodologies. We aim to enhance drug discovery, deepen our understanding of disease mechanisms, and identify key opinion leaders within the medical research community.

Poster

AIBC-ABS-27

How to actionably leverage the Disease Ontology in biomedical research

Claudia Marie Sánchez-Beato Johnson, Lynn Schriml and J. Allen Baron, University of Maryland School of Medicine, Institute of Genome Sciences

The Disease Ontology (DO) stands as the gold standard for disease classification. The DO enables researchers to leverage a standardized foundation of disease data and enrich their projects with high-quality and FAIR knowledge. Citations inform the usage of the DO across diverse areas of research (https://disease-ontology.org/community/use-cases) and knowledge building. Citations inform the usage of the DO across diverse areas of research and knowledge building (https://disease-ontology.org/community/use-cases).

Examining how the DO is utilized, exemplifies alternative options for leveraging the DO to:

(1) Query DO obo or owl files or the DO-KB faceted search interface to identify the classification of a new disease, determining the age of onset, mode of inheritance.

(2) Identify the set of related diseases pertinent for your area of work.

(3) Unify disease annotation and curation efforts across related resources, e.g., Alliance of Genome Resources.

(4) Determine mappings between sets of clinical vocabularies, e.g., SNOMED to MeSH for metabolic disorders, or download one of the DO xref Reports in GitHub (https://github.com/DiseaseOntology/HumanDiseaseOntology/tree/main/DOreports).

(5) Programmatically extract ML-ready datasets of the DO via DO's SPARQL service, e.g, all syndromes where the age of onset has been determined.

(6) Request a new disease for inclusion in the DO (https://github.com/DiseaseOntology/HumanDiseaseOntology/issues).

Utilizing alternative methods to contribute to or mine the DO's semantic disease knowledge for your research projects ensures a high-quality, standardized representation of diseases, providing a reliable foundation for your work.

Poster

AIBC-ABS-31

Knowledge Graph-Based Drug Repurposing for Duchenne Muscular Dystrophy: A Promising Alternative for Rare Disease Treatment

Kavya Katara, Parthiban Srinivasan, Indian Institute of Science Education and Research, Bhopal

Duchenne Muscular Dystrophy (DMD) is a rare and devastating muscle-wasting disease caused by mutations in the DMD gene, leading to the absence of the crucial dystrophin protein. The development of novel drugs for rare diseases like DMD is an arduous and lengthy process, often plagued by extensive time and capital requirements. Drug repurposing offers a promising alternative by reevaluating existing drugs, making alterations to their original formulations and indications. This approach could provide newfound hope for DMD patients, improving their quality of life and potentially extending their survival.

This research paper outlines a novel strategy for drug repurposing in DMD. We propose the creation of a comprehensive knowledge graph, rich in multi-dimensional relationships associated with Duchenne Muscular Dystrophy. The knowledge graph synthesizes data on drug compounds, genetic factors, clinical outcomes etc. to identify potential candidates for repurposing.

Our work is a step towards realizing the potential of drug repurposing for DMD and other rare diseases. We aim to provide a path towards rapid, cost-effective, and more accessible treatments that could significantly improve the lives of patients suffering from these conditions. This research reinforces the notion that existing pharmaceutical compounds, with slight modifications, can be repurposed to address the unique and pressing needs of rare disease patients.

Poster

AIBC-ABS-34

(Re-)bridging the anatomy ontologies with SSSOM

Damien Goutte-Gattat, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge, CB2 3DY, UK

Comparing high-throughput datasets (such as single-cell RNA-sequencing datasets) across species requires, among other things, that said datasets be annotated with ontological terms that must themselves be comparable. Ideally, this in turn requires that a single consistent ontology be used, containing all the terms needed to annotate anatomical structures and cell types across all the commonly used species.

The Uberon (for anatomical structures) and CL (for cell types) ontologies, used together, aim to provide such a consistent ontology that can be used across a wide range of species. But importantly, they do not attempt to single-handedly describe the anatomy of all commonly used species; instead, they leverage existing species-centric anatomy ontologies to construct a single integrated multi-species ontology called composite-metazoan.

The aim of this project is to overhaul the integration mechanism between Uberon/CL and the species-centric ontologies by using the newly devised SSSOM ("Simple Standard for Sharing Ontology Mappings") standard to represent the mappings between Uberon/CL and the species-centric ontologies. Among other benefits, this will: a) improve the traceability of said mappings; b) make it possible to reuse and combine mappings provided by third-party sources; c) ease the maintenance and the evolution of Uberon's composite-metazoan product; d) facilitate the creation of ad-hoc cross-ontology bridges.

Poster

AIBC-ABS-50

ChatGPT usage in Reactome curation process

Krishna Tiwari1,5, Lisa Matthews2, Bruce May3, Veronica Shamovsky2, Marija Milacic3, Karen Rothfels3, Eliot Ragueneau1,5, Chuqiao Gong1,5, Ralf Stephan3, Nancy Li3, Guanming Wu4, Lincoln Stein3, Peter D'Eustachio2, Henning Hermjakob1,5,1 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK 2 NYU Grossman School of Medicine, New York, NY 10016, USA 3 Ontario Institute for Cancer Research, Toronto, Ontario, M5G 0A3, Canada 4 Oregon Health and Science University, Portland, OR 97239, USA 5 Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

Background: Reactome data curation is a meticulous manual, time-intensive process involving data gathering and annotation in a defined data model. To align with the rapid application of large language models, we did a pilot study to explore the potential of using LLMs, particularly ChatGPT for the Reactome curation process.

Methods: The study aimed to test ChatGPT/GPT4 usage in typical Reactome curation tasks like enriching pathway information (Circadian clock), identifying new protein participants, and generating text summaries for pathways. We prompted ChatGPT with questions and generated answers which were verified manually by expert curators. We additionally took a manual approach to extract similar information and compared the two approaches to estimate the added value of using ChatGPT.

Results: Upon manual verification of the ChatGPT-generated content, we found 40% to 45% accuracy with existing protein functions with 54% accuracy of new protein functional details. Over 90% of cited references were fabricated and 55% to 95% of UniProt IDs were inaccurate. The time taken to extract usable details for a protein by both approaches was very similar (ChatGPT = 42 v/s manual = 46 (minutes)). Generated pathway summaries were incomplete with inaccurate citations until in-text citations were provided.

Conclusion: We gained new annotable proteins through ChatGPT but both approaches took a similar amount of time and effort to generate accurate information with the right literature references. The identified limitation with literature citations and accuracy of the information in this study restricts its direct implementation in the Reactome curation process at present.

Poster

AIBC-ABS-52

wwPDB biocuration: Strategies for Managing the Growth of the Protein Data Bank (PDB)

Deepti Gupta, PDBe, EMBL-EBI

The PDB serves as an open-access repository, providing structural biologists with atomic-level insights into 200,000+ biological macromolecules. These structures are fundamental in deepening our comprehension of the intricate interplay between biological structure and function.

The Worldwide Protein Data Bank (wwPDB) partnership, supported by expert biocurators, upholds the data-sharing commitment of the structural biology community. Through the OneDep software system, wwPDB biocurators standardise, validate, and curate incoming structure data, ensuring they adhere to the FAIR principles.

As of late 2023, the PDB archive exhausted its list of existing three-character Chemical Component IDs (CCD IDs). Consequently, we have transitioned into using five-character alphanumeric accession codes for CCD IDs within the OneDep system. Entries featuring the new five-character ID codes are now exclusively distributed in PDBx/mmCIF and PDBML formats due to legacy PDB file format constraints.

With the increasing number of depositions, convenient access for depositors is essential. By utilising ORCiDs for accessing OneDep, authors can now securely log-in and access their entries without the need for deposition ID and password combinations, simplifying the access to their depositions. We are also developing and testing a Deposition API, a REST API offering secure and programmatic access to OneDep, with the aim to automate deposition management efficiently. The objective is to seamlessly integrate the API within structure determination software in a language-agnostic manner, enhancing the streamlined creation of depositions.

As science and structural biology evolve, the PDB faces challenges in handling a growing volume of complex depositions. To meet these challenges, we are enhancing metadata accuracy through automated data harvesting, improving biocuration efficiency via automation, and partnering with task forces to expand data models and enhance data validation, in line with emerging technologies and methodologies.

Poster

AIBC-ABS-55

Global distribution and abundance of Mycobacterium spp. in marine water from WGS metagenomic studies

Samridhi Verma1#, Indu Kumari2 and Sandeep Swargam1*,1Centre for Computational Biology and Bioinformatics, School of Life Sciences, Central University of Himachal Pradesh, Shahpur and Dharmashala, India 2 Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad, India
*Corresponding author: Dr. Sandeep Swargam (Email: swargams@hpcu.ac.in)
#Presenting author

Marine biology is significant “hotspot” region for clinical anti-microbial resistance (AMR) genes in different microbes. Tuberculous and non-tuberculous Mycobacterium (MTB) spp. are one of the unbeaten bacteria that infect human, reptiles and fish etc. The Marine water genomic studies revealed that Mycobacterium marinum, M. fortuitum, M. gordonae, and M. chelonae infects large number of fish spp. In this study, total 129 metagenomic studies were retrieved from ENA till 14/07/2023 to explore the Mycobacterial spp. diversity and abundance across different oceans focused on whole genomic sequence (WGS) data. Only 34 marine/brackish/estuary metagenomic WGS studies considered for quality check and trimming. The mycobacterial abundance and taxonomic assignment were done by Kraken2 tool (Lu et al., 2022). Re-estimation and filtration done through Bracken. In-depth downstream analysis and visualization was done through Pavian and Kraken Tools. The study unravelled the high abundance of the Mycobacterium canettii in Pacific Ocean in mesopelagic, Bathopelagic, Abyssal and Hadal zone and hydrothermal vent systems northeast Pacific. M. intracellulare and M. canettii observed in Black Sea and sediment-water interphase of marine methane seep near the isle of Elba Italy in the Mediterranean Sea and sediment core from Pearl River estuary. Interestingly, Mycolicibacterium was observed in sediment core from Pearl River estuary. This study has documented Mycobacterial spp. that has not been explored in the earlier studies and fills the knowledge gap of its global abundance with distribution. Further the ecological roles of MTB and AMR genomic patterns can be unravelled to understand different drug escaping genes and its mechanism.

Keywords: Marine Biology, Fish, Mycobacterium spp, Metagenomics and WGS

Poster

AIBC-ABS-56

Evidence and Conclusion Ontology: 2024 Update

Suvarna Nadendla, Rebecca Jackson, James Munro, Dustin Olley, Michelle G Giglio, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, USA

Evidence forms the foundation of scientific knowledge. Every statement or conclusion made by a researcher should be based on supporting evidence that can be shared with, and examined by, the wider scientific community. The Evidence and Conclusion Ontology (ECO) provides terms to capture the types of evidence that support biomedical annotations and assertions. Use of ECO allows tracking of annotation provenance, establishment of quality control measures, and evidence-based data mining. ECO is in use by a wide array of data repositories with both specific (e.g the DisProt resource for intrinsically disordered proteins) and general (e.g. UniProt) areas of focus. Currently, ECO contains >2000 terms to capture experimental, computational, and inferential types of evidence. ECO engages in continuous collaboration with biocuration resources as well as other ontologies to ensure that ECO meets the needs of their respective communities. Recent collaborations have resulted in expansion of ECO through the addition of terms in numerous existing branches as well as creation of new branches including one for machine learning evidence. Here we report on recent updates and expansions to ECO and associated resources. Products of the ECO project are freely available for download from the project website (https://evidenceontology.org/)GitHub (https://github.com/evidenceontology/evidenceontology). ECO is released into the public domain under a CC0 1.0 Universal license.

Poster

AIBC-ABS-62

ViCEKb: Creation and analysis of a curated knowledgebase on vitiligo-triggering chemicals to link exposome and health

Nikhil Chivukula(a,b, presenting author), Kundhanathan Ramesh(a), Ajay Subbaroyan(a,b), Ajaya Kumar Sahoo(a,b), Gokul Balaji Dhanakoti(a), Janani Ravichandran(a,b), Areejit Samal(a,b, corresponding author),(a) The Institute of Mathematical Sciences, Chennai; (b) Homi Bhabha National Institute, Mumbai

Vitiligo is a complex disease wherein the environmental factors, in conjunction with the underlying genetic predispositions, trigger the autoimmune destruction of melanocytes, ultimately leading to depigmented patches on the skin. Apart from being susceptible to other autoimmune disorders, the affected patients may face social stigmatization leading to a decreased quality in life. While genetic factors have been extensively studied, the knowledge on environmental triggers remains sparse and less understood. Therefore, a comprehensive understanding of the environmental triggers will not only explain the complex etiology of the disease, it will also enable the prioritization of potential toxic chemicals in human chemical exposome. Towards this, we present the first comprehensive resource on vitiligo triggering chemicals namely, Vitiligo-linked Chemical Exposome Knowledgebase (ViCEKb). ViCEKb involved an extensive and systematic manual effort in curation of 113 unique chemical triggers of vitiligo from published literature and categorized them based on their evidence and source of exposure. ViCEKb catalogues various chemical information, including a wide range of metrics necessary for different toxicological evaluations. Moreover, extensive cheminformatics-based analysis of the ViCEKb chemical space highlighted its diversity and uniqueness in comparison to skin specific chemical regulatory lists. Additionally, a transcriptomics-based analysis of ViCEKb chemical perturbations in skin cell samples highlighted the commonality in their linked biological processes. ViCEKb is freely available for academic research at: https://cb.imsc.res.in/vicekb.

Poster

AIBC-ABS-65

Protein-Protein Interactions Between Human Host and Gut Pathogens

Monika Tiwari, Amresh Kumar Sharma, Anup Som*, Centre of Bioinformatics, Institute of Interdisciplinary Studies, University of Allahabad, Prayagraj - 211002, India

In combating infectious diseases, the emergence of drug-resistant pathogen necessitates a deep
understanding of molecular interactions between human and gut-pathogens. Experimental identification of these interactions is hindered by challenges, leading to sparse data in existing databases. To address this, researchers used a homology-based approach to predict protein-protein interactions in host-pathogen dynamics. In this poster, we focused on the homology-based approach as a predictive tool for host-pathogen interactions (HPIs). By identifying similarities between known and unknown proteins, this method offers a computational workaround to address the scarcity of experimental data, enabling a broader exploration of HPI landscape. Our poster synthesizes findings from computational prediction using homology-based approach. Through extrapolation from known interactions, we present an overview of protein-protein interactions shaping human-gut pathogen dynamics. This analysis enriches our understanding of infection mechanisms. Protein-Protein interactions emerge as pivotal players in infectious diseases, particularly in the context of the human-gut interface. The homology-based approach proves instrumental in predicting HPIs, providing valuable insights for targeted therapeutic interventions against drug-resistant pathogens.

Keywords: Human-Gut Pathogens; Host-Pathogen Interactions; Protein-Protein Interactions;
Homology-Based Approach.

Poster

AIBC-ABS-66

An integrative data-centric approach to derivation and characterization of an adverse outcome pathway network for cadmium-induced toxicity

Ajaya Kumar Sahoo (a,b,*), Nikhil Chivukula (a,b), Kundhanathan Ramesh (a), Jasmine Singha (c), Shambanagouda Rudragouda Marigoudar (c), Krishna Venkatarama Sharma (c), Areejit Samal (a,b),(a) - The Institute of Mathematical Sciences (IMSc), Chennai, India; (b) - Homi Bhabha National Institute (HBNI), Mumbai, India; (c) National Centre for Coastal Research, Ministry of Earth Sciences, Government of India, Pallikaranai, Chennai, India; (*) - Presenting author

Cadmium is a prominent toxic heavy metal that contaminates both terrestrial and aquatic environments. Owing to its high biological half-life and low excretion rates, cadmium causes a variety of adverse biological outcomes. Adverse outcome pathway (AOP) networks were envisioned to systematically capture toxicological information to enable risk assessment and chemical regulation. Here, we leveraged AOP-Wiki and integrated heterogeneous data from four other exposome-relevant resources to build the first AOP network relevant for inorganic cadmium-induced toxicity. From AOP-Wiki, we filtered 309 high confidence AOPs, identified 312 key events (KEs) associated with inorganic cadmium, and thereafter, curated 30 cadmium relevant AOPs (cadmium-AOPs), using a data-centric approach. By constructing the undirected AOP network, we identified a large connected component of 18 cadmium-AOPs. Further, we analyzed the directed network of 59 KEs and 82 key event relationships (KERs) in the largest component using graph-theoretic approaches. Subsequently, we mined published literature using artificial intelligence-based tools to provide auxiliary evidence of cadmium association for all KEs in the largest component. Finally, we performed case studies to verify the rationality of cadmium-induced toxicity in humans and aquatic species. Overall, cadmium-AOP network constructed in this study will aid ongoing research in systems toxicology and chemical exposome.

Poster

AIBC-ABS-67

Biomarkers and the underlying mechanisms of Alzheimer's disease

Abhay Raj Kori, Piyush Mishra, Anup Som, Centre of Bioinformatics, Institute of Interdisciplinary Studies, University of Allahabad, Prayagraj – 211002, India

Alzheimer's disease (AD) is a progressive neurological illness that mostly affects the brain, causing memory loss and cognitive deterioration. The buildup of aberrant protein deposits in the brain, particularly tau tangles and beta-amyloid plaques, is one of the defining characteristics of Alzheimer's disease. Considerable work has been done in the last few years to find reliable biomarkers for AD, where researchers identified specific biomarkers, including RNA molecules (such as mRNA, miRNA, and lncRNA), that can serve as indicators of disease occurence, progression, or response to treatment. Exploring the interaction amongst these biomarkers may reveal a deeper understanding of the fundamental processes behind AD, which will ultimately lead to design better diagnostic and therapeutic approaches. This work synthesizes current knowledge on AD pathology and diagnostic biomarkers, outlining how transcriptomics contribute to unveiling reliable biomarkers for clinical AD management.

Keywords: Alzheimer's disease; Biomarkers; Molecular interactions; Transcriptomics; Network biology.

Poster

AIBC-ABS-69

Identification of activity cliffs in structure-activity landscape of chemicals binding to endocrine receptors

Shanmuga Priya Baskaran(a,b,1), Ajaya Kumar Sahoo(a,b,1), R. P. Vivek-Ananth(a,b,1), Nikhil Chivukula(a,b), Kishan Kumar(a), Janani Ravichandran(a,b) and Areejit Samal(a,b),a-The Institute of Mathematical Sciences (IMSc), Chennai - 600113; b-Homi Bhabha National Institute (HBNI), Mumbai - 400094; 1-Joint-First authors

Activity cliffs constitute chemicals that are structurally similar but exhibit very different activity against a biological target. The presence of such activity cliffs in chemical datasets present a challenge towards development of predictive models, including Quantitative Structure Activity Relationship (QSAR) models. In other words, systematic analysis of the structure-activity relationship including identification of activity cliffs is important for the development of robust computational toxicity models. To this end, we have performed detailed cheminformatics analysis of the structure activity relationship in curated datasets of chemicals binding to human endocrine receptors. Notably, we have employed several approaches to identify and characterize activity cliffs. These approaches include Structure-Activity Similarity (SAS) map, Structure-Activity Landscape Index (SALI), Matched molecular pair (MMP), Mechanism of action (MOA), Triple-activity difference (TAD) map and Dual-activity difference (DAD) map. Lastly, we identify and characterize activity cliffs from the perspective of a single (receptor) target as well as multiple (receptors) targets.

Poster

AIBC-ABS-70

Curating Somatic Variants in Rare Skin Cancers in COSMIC

Madiha Ahmed, Joanna Argasinska, Denise Carvalho-Silva, Alex Holmes, Rachel Lyne, Amaia Sangrador-Vegas and Sari Ward, Wellcome Sanger Institute

COSMIC, the Catalogue of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk), is the world's largest source of expert manually curated somatic mutation information relating to human cancers. Common skin cancers, such as basal cell carcinoma, squamous cell carcinoma and melanoma are overrepresented in the scientific literature. To better represent rarer skin cancers in COSMIC our recent (V98) release has focussed on curation of rare skin cancers. We searched for publications about adnexal tumours, Merkel cell carcinoma, Kaposi sarcoma, dermatofibrosarcoma protuberans, sebaceous carcinoma and extramammary Pagets disease. This resulted in 36 curatable publications, 776 new samples and 25,236 new variants related to rare skin tumours. In addition, 17 new skin tumour types or subtypes were added to the COSMIC histology classification system. Data from whole genome studies, large next generation sequencing panels and case reports of individual cases were included.

Despite some recent successes in drugs for the treatment of rare skin cancers, many still have very poor prognosis. A greater understanding of the mutational landscape of these tumours will hopefully lead to the development of new targeted therapies and better outcomes for patients. Here we will present both the successes and challenges in curating such a dataset in the context of the changes in biocuration over the last 20 years and looking to the future to how we can improve on data extraction, presentation and becoming a FAIRer resource

Poster

AIBC-ABS-73

Evaluation of the Effects of Oxidative Stress and Biomarkers of Inflammation in Cardiomyopathy Sufferers

Shivam Tiwari, Om Shankar, Royana Singh, Umesh Choudhary, Institute of Medical Sciences, Banaras Hindu University Varanasi

Introduction: Cardiovascular disease can develop and worsen as a result of inflammation and oxidative stress. The current research looked into the relationship between oxidative damage and biomarkers of inflammation in individuals with cardiomyopathy.

Methods: Particular kits for ELISA were used to measure the serum concentrations of CRP, TNF-α, IL-6, and NT-proBNP. These specific ELISA kits are based on sandwich enzyme immunoassay techniques whose results are quantitative. The accuracy of the tests was established by comparing them to control sera that were included in the kits and had known quantities of the analytes.

Results: When compared to individuals without cardiomyopathy (control group), we found that cardiomyopathy patients had significantly higher blood C-reactive protein concentrations (P =<0.0001). When compared between control case and cardiomyopathy patients, then find that cardiomyopathy patients had significantly higher concentrations of tumour necrosis factor-alpha (TNF-alpha) (P =<0.0001). TNF- α associated favourably with malondialdehyde (P =<0.0001, r =0.4524) and glutathione peroxidase (GPX) (P =<0.0001, r =0.8311) in Cardiomyopathy patients. Interleukin-6 was not significantly linked with GPX (P =0.0001, r =-0.1194) in cardiomyopathy patients. In those with cardiomyopathy, there was a strong association (P=<0.0001, r=0.4826) between malondialdehyde and NT-proBNP. Furthermore, we observed that the activity of glutathione peroxidase (GPX) exhibited a significant connection with NT-proBNP (P =<0.0001, r = 0.6084) in all cardiomyopathy patients.

Conclusions: In cardiomyopathy patients, but not in normal cases, there is a correlation between inflammatory and oxidative stress indicators. These findings imply intricate cross-talk between the two cellular processes in late-stage cardiomyopathy.

Keywords: Cardiomyopathy; Inflammatory Biomarker; Cardiac Biomarker.

Poster

AIBC-ABS-74

Anti-Microbial Peptide Database version 1 (AMPDB v1): A Manually Curated Comprehensive & Exhaustive Resource of AMPs

Rajat Kumar Mondal#, Sintu Kumar Samanta*, Biochemistry and Bioinformatics Laboratory, Department of Applied Sciences, Indian Institute of Information Technology Allahabad (IIIT-A), Uttar Pradesh, Devghat, Jhalwa, Prayagraj, 211012, India. #Presenter, *Supervisor

In the rising tide of microbial resistance, antimicrobial peptides (AMPs) can be the most potent therapeutic alternatives to conventional drugs. However, existing AMP databases struggle to keep pace with the expanding information landscape. To bridge this gap, we introduce the Anti-Microbial Peptide Database version 1 (AMPDB v1), a meticulously curated comprehensive, and exhaustive resource aimed at addressing the limitations of current databases in antimicrobial research. In developing AMPDB v1, we derived information from the NCBI-Protein and EMBL-EBI database and employed the latest technology for database structure, UI, and tools. The database comprises 59,122 entries classified into 88 classes, each curated manually for accuracy and reliability. Sequence alignment (includes BLASTp, MUSCLE, N&W global, and S&W local sequence alignment) and protein feature calculation (includes composition, physicochemical properties, CTD and QSAR descriptors) tools are integrated as web applications, ensuring real-time response and user-friendly interactions. AMPDB v1 boasts a significant size and a more robust classification system than any other generalized database to date. The database offers multiple browsing and searching options, from straightforward text searches to a customizable advanced search page. The web-based interface, accessible at https://bblserver.org.in/ampdb/, further enhances user experience and accessibility. AMPDB v1 represents a substantial advancement in AMP databases. Through meticulous curation, advanced tools integration, and user-friendly features, it addresses the limitations of existing resources. Researchers and practitioners in antimicrobial research can benefit from the comprehensive and reliable information provided by AMPDB v1, fostering further advancements in the understanding and application of AMPs.

Poster

AIBC-ABS-76

In-silico investigation on tryptic and Lys-C peptides in the context of proteomics: Analysis of amino acid composition

V. S. Gowri [1], V. Sabareesh [2],[1] PG and Research Department of Chemistry, Auxilium College (Autonomous), Gandhi Nagar, Vellore - 632006, Tamil Nadu, India.[2] Centre for Bio-Separation Technology (CBST), Vellore Institute of Technology (VIT), Vellore - 632014, Tamil Nadu, India.

Background: Significant developments in genome sequencing of various organisms led to the emergence of 'proteomics', meaning, high-throughput sequencing of proteins. In most of the proteomic studies 'trypsin' has been utilized that has resulted in successful identification of proteins, as inferred from the mass spectrometry (MS) based identification of tryptic peptides, which was popularly called 'bottom-up (BU)' approach. In another approach termed 'middle-down (MD)', proteases such as Lys-C, Asp-N, etc., were used that yielded peptides longer than the typical tryptic peptides. Ample evidence has shown that MD approach yields better sequence coverage of peptides than BU.

Length and net charge of the proteolytic peptides depend on the specificity of the protease. Moreover, length and the charge state of the precursor ion strongly influence mass spectrometric identification of peptides, particularly with regard to electrospray ionization - tandem MS. Therefore, in this study, we aim to analyze the amino acid composition of the in-silico generated tryptic and Lys-C peptide sequences from model proteins, in order to examine the correlation between net charge and length of the proteolytic peptides, if any.

Methods: In-silico trypsin and Lys-C proteolysis was carried out on some model proteins selected from five different organisms, wherein Xxx-Pro sites were not subjected to proteolysis. In-house suite of scripts written in perl and shell were used.

Results: The impact of histidine, arginine and lysine on the length and net charge of tryptic and Lys-C peptides shall be presented in detail. The influence of missed cleavages also shall be delineated.

Poster

AIBC-ABS-77

Curating a Phytochemical Gold Standard for Plant Invasion Biology

Abhishek Prasad; Dr. Gitanjali Yadav, National Institute of Plant Genome Research

Invasive plants pose significant threats to biodiversity, ecosystems, environmental quality and human health. Allelopathy is a prominent weapon of invasive plants. Despite invasive species being a major threat, there exists no comprehensive global resource for qualitative & quantitative information on the phytochemical weaponry used by invasives in new habitats. To address this gap, we investigated phytochemical dynamics during invasion across diverse geo-spatial locations. This was done by extraction and analysis of composition profiles of essential oils from published records of 27 invasive plants across 39 countries, spanning both native and invaded habitats. These emission profiles were curated from hundreds of research articles in more than 60 scientific journals. Detailed information of biotic and abiotic variables as well as corresponding metadata from each research article was compiled. The data was directly collected from PDFs by extracting the data tables. The extracted data was processed and normalised to remove semantic errors and duplications, followed by assignment of unique ids to each phytochemical record, each profile and each source article to enable a comprehensive and quantifiable analysis. A universal phytochemical reader for literature is being curated in this manner for being trained on a gold-standard of more than 16,000 records of invasive species emissions encompassing almost 1,600 chemical compounds. In summary, this study represents a pioneering effort in the bio-curation of invasive plant phytochemistry, offering a holistic exploration of phytochemical dynamics on a global scale as well as paving the way for deciphering the roles of phytochemicals in plant invasion.

Poster

AIBC-ABS-78

A proposal for yet another search service for life science ontologies

Yasunori Yamamoto, Database Center for Life Science (DBCLS), ROIS-DS

There are widely used services to look up life science ontologies such as BioPortal, Ontology Lookup Service, and Ontobee. After trying them to develop a new ontology and RDF data with it, I found a function that could make them more useful. It is to search for properties relevant to a given class and vice versa, that is, to search for classes relevant to a given property. We are planning to develop an ontology search service that provides this function. I have developed Triple Data Profiler, a tool to retrieve relationships among classes and properties used in a given RDF dataset. We assumed the data obtained with this tool can be used to provide the aforementioned service and surveyed them which collected from 41 endpoints and 1003 graphs. Based on this survey, we confirmed the feasibility of this service.

Poster

AIBC-ABS-79

BIG DATA BIOCURATION: The Earth MetaPhenome

Priti; Dr. Gitanjali Yadav, National Institute of Plant Genome Research

Plants constantly synthesis diverse metabolites, each playing a distinct role in defense or communication, collectively now considered as unique “chemical spectra specific fingerprints”. Bioprospecting for Plant secondary metabolites has multifaceted applications apart from helping to unravel the functions and significance of such metabolites. However, a significant wealth of information towards natural products research remains locked in the vast repository of published data spanning centuries, due to copyrighted scholarly records, hindering accessibility and integration across scientific domains.

We are addressing this challenge by means of a citizen science management and student authorized toolkit to establish a comprehensive and collaborative platform. That is being used to curate and normalize data from 107 million research articles, this data is available as n-grams of plants and chemistry and we have ingested into a searchable database. In all we have identified 617,263 chemical-plant pairs, sourced from 383,793 unique DOI numbers, featuring 6,647 Plants and 1,725 Chemicals. Additionally, incorporated available data on Physico-chemical properties of chemical compounds and Plant taxonomy, enhancing data accuracy. Integration with Wikidata IDs has enabled in to embed this knowledge into the “Global Knowledge Graph” strengthening accessibility, interoperability, and invites collaborative data curation within the community, ushering in novel publication models.

Our aim is to seamlessly integrate biology and chemistry, Profiling the metabolic content of all currently known Plants on our planet and openly sharing this wealth of information on a unified platform. This initiative holds immense potential to revolutionize natural products research and extend its applications beyond the confines of the field.

Poster

AIBC-ABS-78

PlotS: web-based application for data visualization and analysis

Ringyao Jajo, Shivani Kansal, Sonia Balyan, Saurabh Raghuvanshi, University of Delhi, Department of Plant Molecular Biology, New Delhi; University of Delhi, Department of Plant Molecular Biology, New Delhi; Indian Biological Data Centre, Faridabad; University of Delhi, Department of Plant Molecular Biology, New Delhi

Advancements in technology have made data representation using graphs more transparent and interpretable. However, existing freely available visualization tools have limitations in handling multivariate data, adding metainformation, and most of it fail to provide comprehensive support for statistical analysis. PlotS is a web-based application that allows the integration of visualization and statistical analysis into a single workflow. It supports multivariate data visualization and incorporation of metainformation in graphs using side graphs, inset graph, faceting, or adding layers with or without secondary y-axis. The current version has eight types of graphs (bar, box, density, frequency polygon, histogram, line, scatter and violin plot) and four statistical methods (T-test, ANOVA, Wilcoxon test and Krushkal-Wallis test). It is an interactive application with numerous customization options for visualization. It can handle variety of data formats with or without replicates. Results for inferential statistical analysis are annotated in the graph. PlotS is readily accessible online, serving as a valuable and accessible resource for researchers seeking to enhance their data analysis and visualization capabilities.

Poster

AIBC-ABS-81

IDIC: An Integrated Database of Ion Channels

Kiran Bharat Gaikwad1,2,#, K. T. Shreya Parthasarathi1,2,#, Meh Jabeen1, Akhilesh Pandey3,4,5 and Jyoti Sharma1,2*,1 Institute of Bioinformatics, International Technology Park, Bangalore 560066, India 2 Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India3 Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA 4 Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA # Authors contributed equally

Background

Ion channels play critical roles in a variety of physiological processes. Their alterations may cause numerous diseases. These are the second largest class of membrane proteins considered as drug targets.

Methods

This database contains ion channel annotations including protein signature sequences, protein domains organization, nuclear localization signals, protein folds. Furthermore, IDIC contains details about sequence alterations and altered expressions of ion channels across various types of tumors and genetic disorders. Database interface and middleware were built using Hyper Text Markup Language and Django Python web framework.

Results

Overall, the database contains 31 unique signature sequences predicted in 176 ion channels. 103 nuclear localization signals were identified in 80 ion channels. Moreover 131 domains were predicted in 328 ion channels whereas 127 unique folds were predicted for 276 ion channels. 496,075 variations were obtained across tumor types from COSMIC. Approximately, 134 genetic disorders were found to be associated with 101 ion channels using data mining approach.

Conclusion

Using interactive graphical displays and updatable features, “IDIC” could serve as a knowledge-based resource to understand ion channels and their involvement in various human diseases.

Poster

AIBC-ABS-82

Machine learning-based approach for the classification of antibiotic resistant bacteria

K. T. Shreya Parthasarathi (1,2), Kiran Bharat Gaikwad (1,2), and Jyoti Sharma (1,2,*),(1)Manipal Academy of Higher Education (MAHE), Manipal, Karnataka 576104, India. (2) Institute of Bioinformatics, International Technology Park, Bangalore 560066, India

Background

Microorganisms, crucial for ecosystem balance, can be harmful, causing life-threatening infections. Characterizing microbial communities traditionally is expensive and time-consuming. High-throughput sequencing generates vast genomic data, presenting challenges in analysis and interpretation. Machine learning offers a data-driven solution, aiding classification and interpretation issues. Integrating genomic data and machine learning into clinical practice is a feasible, reproducible, and robust strategy that may produce clinician-friendly outputs.

Methods

k-mers were generated from nucleotide sequences of pathogenic bacteria resistant to antibiotics and were clustered into groups of bacteria sharing similar genomic features using the Affinity propagation algorithm. Thereafter, a classification model based on support vector machine algorithm was developed to classify the bacterial strains into resistant and non-resistant strains against specific antibiotics.

Results

The bacterial strains belonging to 7 different genera were clustered into 13 groups with a silhouette coefficient of 0.81. The bacterial strains were classified into resistant and non-resistant strains against multiple antibiotics including tetracycline, penicillin, ampicillin and imipenem.

Conclusion

Segregating pathogenic bacteria based on genomic similarities is a valuable approach for assessing the severity of diseases caused by new bacterial strains. Through clustering, pathogenic bacteria with shared genomic features were identified, facilitating computational prediction analyses. Notably, machine learning may prove advantageous in uncovering antibiotic resistance (ABR) determinants. This method could enhance our understanding of ABR patterns, paving the way for more informed and effective treatment options.

Poster

AIBC-ABS-84

Intersecting in vivo, in silico approach to assess the cognitive patterns of vascular dementia

Nymphaea Arora1,3; Anil Kumar Rana2,3; Damanpreet Singh2,3*; Vishal Acharya1,3*,1 Artificial Intelligence for Computational Biology (AICoB) Lab, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, HP, India; 2 Pharmacology & Toxicology Laboratory, Dietetics and Nutrition Technology Division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, HP, India; 3 Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India

A report by WHO states that 55 million people are currently living with dementia, and this number is predicted to increase tremendously by 2050. Around 10 million new cases of dementia are being diagnosed every year, foisting a huge burden on their families. Cerebrovascular disease pathologies may lead to reduction (ischaemic stroke) or rupture (haemorrhagic stroke) of blood supply and are the major contributors to vascular dementia. Chronic ischaemia, constituting 62% of all stroke cases; and cognitive impairment are reported to be 50% associated in patients younger than 50 years. Stroke is third nationally and second globally highest in terms of mortality and third highest worldwide in terms of Disability Adjusted Life Years (DALYs). Post-stroke causes neuronal damage responsible for cognitive deficits which can get to severe levels causing dementia. This condition makes the patients incapable of living independently, which without proper treatment, results in high morbidity and mortality. The study of pathogenesis is still scarce and needs an investigatory interdisciplinary approach.

Our research focuses mainly on the two regions; cortex and hippocampus, which are known to be involved in cognitive functioning. This study recognises the use of RNA-Seq high-throughput sequencing for identifying and measuring gene expression. The interactome plays an unavoidable role in the pathogenesis of disease. Thus, the protein-protein interaction analysis assesses the significance of disease signatures in neurodegenerative pathways. Further, analyzing the expression and modulation of the targeted proteins substantiates their significant role in cognitive impairment, which may be of great importance to patients with vascular dementia.

Poster

AIBC-ABS-85

Network-based deep learning approach to uncover the plant immune system in response to biotic stresses

Ravi Kumar1,2; Vishal Acharya1,2*,1Artificial Intelligence for Computational Biology Lab (AICoB), Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India; 2Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India

Identification and interpretation of the interactomes involved during host–pathogen interaction events might advance our understanding of the defense mechanisms of plants. Previously, several studies were conducted to examine the host gene's network response to the pathogens. However, how plants use their defense-responsive gene circuitry to adapt to biotic challenges is still uncertain. Deep-learning models are plausible for future applications owing to their sensitivity to other traditional methods. However, developing deep-learning algorithms to understand the host response to pathogens is difficult due to a paucity of omics data for a specific pathogen. We developed a deep-learning model that uses the expression profile and network information for the classification and constructs an immune-related gene network to avoid the problem of fewer samples than the features. Validation studies on pathogenic responses show that the DL-based model outperforms other machine learning and traditional methods.

There is lack of collaborative study on how the intertwined networks of defense-related genes in rice contribute to typical pattern-triggered immunity (PTI) and effector-triggered immunity (ETI). Apart from PTI/ETI, further study is needed to explore within-ETI differences between pathogenic lifestyles in rice. We use the omics data of rice in response to bacterial Xanthomonas oryzae and hemibiotrophic fungus Magnaporthe oryzae. Our finding indicates the compactness of the PTI network and the rise of independent modules in the ETI network. Results also show more independent network modules and minimum structural disorderliness in ETI-M. oryzae than the ETI-Xoo model, revealing the different adaptational strategies of rice plants to evade pathogen effectors.

Poster

AIBC-ABS-86

A hybrid explainable ensemble transformer-based approach for prediction of miRNA-lncRNA interaction in plants

Dipali1,2; Vishal Acharya1,2*,1. Artificial Intelligence for Computational Biology (AICOB) Lab, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India; 2. Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India

The interactions between miRNAs and lncRNAs play crucial roles in regulating various biological process such as development, biotic and abiotic stress, flowering time, grain yield and more. The experimental details of obtaining lncRNA-miRNA interaction are still lacking; however, few in silico approaches such as CIRNN, PmliPred, PmliPEMG, PmliHFM were previously developed in model plants. The aforesaid state-of-the-art (SOTA) methods tried to improve the prediction accuracy but still have some limitations, such as manually extraction of features and long-distance dependency, as these methods relied on gated recurrent neural networks (GRNNs) such as long-short term memories (LSTMs) and gated recurrent units (GRUs) with added attention. Therefore, these methods are unable to determine the exact positions of bases present in the long sequences resulting in higher false positives. We will try to implement hybrid algorithm in the form of transformer, the advanced architecture in the natural language processing (NLP) with attention mechanism rather than recurrent alone. Self attention implemented in the transformer is used to compute similarity scores between words in a sentence. Our method while incorporating the modified architecture can provide necessary but stringent data that will have additional advantage to train our model to overcome overfitting. To conclude all, our study will help to uncover the interplay of miRNA-lncRNA by predicting highly confident interactions which leads to understand the role of these non-coding RNAs in different biological processes in plants.

Poster

AIBC-ABS-87

Rethinking drug repositioning and development with artificial intelligence, machine learning, and transcriptomics

Bibhu Prasad Behera, V Badrireenath Konkimalla, NISER

Background: The current drug discovery paradigm, burdened by high costs and protracted timelines, demands a transformative approach. We propose a novel strategy integrating artificial intelligence (AI), deep learning, and transcriptomics to accelerate drug repurposing for infectious diseases, providing an efficient response to emerging health threats.

Methodology: This study utilizes gene expression of infectious diseases from the GEO database. The common gene expression patterns and their molecular functions were analyzed by performing transcriptomic analysis with different Bioconductor packages (limma, clusterProfiler, and DESeq2). Further, identified gene markers with substantial fold changes were processed through recurrent neural networks (RNNs) trained on large datasets of drug-gene interactions to navigate a pre-existing library of approved drugs. These RNNs predict potential repurposing candidates with high target specificity towards the predicted dysregulated genes. Molecular dynamics simulations further scrutinized promising candidates identified through virtual screening. Lastly, replica exchange Monte Carlo and meta-dynamics were used to explore the ligand-protein binding energy landscape.

Results:

Conclusion: Based on the transcriptomic data, this study suggested a significant potential for drug repurposing, given the availability, quality, and selection of the disease dataset. Also, the resultant genes and the availability of drugs for that particular gene will help in drug repurposing with favorable pharmacokinetic and safety profiles.

Note: This ongoing project intentionally excludes the results section from the abstract.

Poster

AIBC-ABS-91

Heat induced changes in epigenetic and transcriptomic landscape shape the stress response in tomato inflorescence

Monika Shrivastava, Shivani Kansal, Sonia Balyan, Saurabh Raghuvanshi, Saloni Mathur, National Institute of Plant Genome Research

Heat stress (HS) is one of the major adversity impacting the growth, yield and quality of tomato. The inflorescence transcriptome data of a heat-tolerant (CLN) and heat-sensitive (CA4) tomato cultivar-pair showed that most of the differentially expressing genes (DEGs) follow cultivar-biased expression pattern; 410 genes (215/195 down/up) have HS-responsive differential expression only in CLN while 1728 genes (770/958 down/up) exhibit CA4-specific HS-response. There are 569 conserved DEGs between the cultivars while only 16 genes exhibit antagonistic HS-response. The whole genome re-sequencing variant call analysis using 'Genome Analysis Tool Kit' highlighted 2,474,357 SNPs in CLN, 236,668 insertions and 167,306 deletions as compared to 1,963,832 SNPs, 214,576 insertions and 141,379 deletions in CA4. Whole-genome-bisulfite-sequencing showed that the methylated Cytosines increase under stress to two-folds in CLN and 1.25-folds in CA4 and the methylation level of 'CpG' context is highest followed by the 'CHG' context and the lowest in the 'CHH' context. The silent genes exhibited constant methylation in and around gene-body but as expression level increases, a clear distinction between gene-body methylation and flanking regions in 'CpG' and 'CHG' context is evident. The two cultivars showed differences in the demethylation and de-novo methylation patterns under stress. Moreover, Differentially Methylated Cytosines and Differentially Methylated Regions showed context biasness wherein, CLN had greater methylation level than CA4 in 'CpG' and 'CHG' contexts while CA4 exhibited preference for 'CHH' context. The study brings to light the inherent differences between the two contrasting varieties in terms of genomic sequence, transcriptome and methylome landscape in HS.

Poster

AIBC-ABS-92

Curating a Gold Standard for Hyperspectral Phytochemical Data

Mandeep; Dr. Gitanjali Yadav, National Institute of Plant Genome Research

Plants produce a vast diversity of Biologically active Volatile organic compounds (BVOCs) in different environmental conditions. These emissions have traditionally been recorded by means of GS-MS approaches. The resulting spectral peaks are deciphered for identifying chemical composition profiles by experts and published in literature, but often in journal specific formats. This lack of a common data format or nomenclature is a severe detriment to large scale meta analyses or pattern searches. We aim to liberate this knowledge, using information that is currently locked in over 500 million pages of the published scholarly corpus, for the benefit of research and climate action.

BVOC emissions can vary based on several factors such as geographical location, stage of development, different tissues, season etc. Thus, we are collecting very comprehensive contextual information about Phytochemical emissions, including released compound names, taxonomic source species names, any reported conditions or environment at the time of emission, as well as extraction methodology and exact amounts of emission, apart from spatio-temporal data about each record. Since each journal has its own format for spectrophotometric data publishing, we are building journal specific automated readers in parallel for this work. The manually curated gold standard is focused on data from we the Journal of Essential Oil Research (JEOR) which contains tabular data on GC-MS composition profiles. The current version of the gold standard includes detailed composition profiles of essential oils 200 plants across 40 countries and about 150 locations, spanning 2000 compounds. This gold standard is intended to serve as a reference for testing the accuracy of the universal phytochemical reader that we are creating in parallel. In addition it will help to quantify plant emissions towards deep learning the plant metabolome.

Poster

AIBC-ABS-93

A Systematic Comparative Analysis of Pathway Databases from a User's Perspective

Moushumi Goswami, Divya Dosemane, Sravanthi Davuluri, Akhilesh Bajpai, Kavitha Thirumurgan, Kshitish K Acharya,Institute of Bioinformatics and Applied Biotechnology (IBAB), Bengaluru; Manipal Academy of Higher Education (MAHE), Manipal; Shodhaka Life Sciences Pvt. Ltd, Bengaluru; Structural Biology Lab, Centre for Biomedical Research, School of Bio Sciences & Technology (SBST), Vellore Institute of Technology (VIT) University, Vellore

Background: Pathway DataBases (PathDBs) are online tools that provide organized information about biological pathways and processes. However, multiple aspects of these PathDBs, including the mode of information collection, varies. For example, some of them gather data via manual curation, while others follow automated methods. The variations stemming from multiple reasons create a challenge for the users in making objective selection among the PathDBs. Hence, there is a need to compile these databases and systematically compare them from a user's perspective.

Method: We first compiled the PathDBs via thorough literature search, and shortlisted those suitable for human pathway studies. We designed a query set consisting of genes, pathways and conditions, and performed a quantitative comparison of output from these shortlisted PathDBs. Additionally, a list of PDBs were queried with MAPK pathway, and the gene coverage was compared. Furthermore, to test the update frequency among the PathDBs, potential novel genes involved in the MAPK pathway were curated and compared with information from PathDBs.

Result: For the quantitative comparison of the PathDBs, PathDIP yielded the highest number of pathways when queried with gene and condition names, followed by others. However, when queried for pathway names, Reactome yielded the highest number of pathways, followed by others. For the pathway specific query, Reactome gives the highest number of genes followed by PathDIP, KEGG, and others. PathDIP additionally, also provides predicted information for certain pathways.

Poster

AIBC-ABS-94

A systematic comparison of literature search engines

Goswami Moushumi, Dosemane Divya, Davuluri Sravanthi, Bajpai Akhilesh, Jayakumar N Nandhini, Kurimindla Pawan, Saha Satabdi, Belakeri Bindurani, BM Sanjana, VallikkadanKandi Gopika, Sharon Lizzi, Acharya K Kshitish,Institute of Bioinformatics and Applied Biotechnology (IBAB), Electronic City, Phase I, Bengaluru – 560100, Karnataka, India (www.ibab.ac.in), Research Scholar, Manipal Academy of Higher Education (MAHE), Manipal – 576104, Karnataka, India, Shodhaka Life Sciences Pvt. Ltd., Electronic City, Phase I, Bengaluru – 560100, Karnataka, India(www.shodhaka.net)

Background: Literature search is an essential component of research at various phases, including 'effective text mining' or bio-curation. Bio-literature search engines differ in their ability to scan various sections of articles for user-provided search terms. Periodic scientific analysis of major bibliographic tools is necessary. Without such thorough comparative studies, the selection of search engines remains subjective.

Method: Over a period of about 2 years, we conducted an extensive analysis of almost 200 existing bio-literature search engines, and systematically narrowed down to most useful ones. Several research topics were selected for examination. After each search, we noted the number of hits initially. A more systematic comparison was done in the final phase, where we carefully assessed the relevance of retrieved citations manually - one article at a time.

Results: The study presents a comprehensive list of the bio-literature search tools, and a comparative account mainly in terms of their efficiency of obtaining relevant citations. Many top-performers noted in our previous study, carried over a three-year period, had stopped working when the current study began. As before, the search engines provided varying total number of relevant citations as relative ratio of such hits. To enhance the recall efficiency, using multiple search tools seems to be the only option. The detailed observations from this study can guide researchers and health professionals in making more objective selections among available search engines.

Poster

AIBC-ABS-95

Covid-Kosha, a model pandemic database: curated findings and associated primary data simplified and categorised for expediting the propagation of scientific information for all

Acharya K Kshitish, Umesh Akkash, Thayyil V Vishnu, Thalathoti S Lizzi, Parikh Payal, Biswas Lipsa, Pai S Poojitha, Gupta Anubha, Agarwal Stuti, Saifudeen Ashikha, Sowmya Anantha Krishnan, Kolli Sarika, Pabla Jaspreet, Balasubramanian Janaki, Pillai Anushka, Nirali Mehta, Tripathi Nisha, Donthi S Saakshi, Dipti Shaw, Harshita Miyar, Lazar Louis Anto Prettina, Balagannavar Govind, Vasudevan Anjali, Saji Linda, Abdul Basit Khan, Mathuria Kanika, Jitendra Kumar,Institute of Bioinformatics and Applied Biotechnology (IBAB), Biotech Park Electronics City Phase I, Bengaluru 560 100, Karnataka, India; (www.ibab.ac.in),Shodhaka Life Sciences Pvt. Ltd., Electronics City Phase I, Bengaluru (https://www.shodhaka.net), Bangalore Bioinnovation Centre, Biotech Park, Electronics City Phase I, Bengaluru (https://www.bioinnovationcentre.com),Biological data Analyzers' Association (BdataA) – a virtual organization (https://startbioinfo.org/BdataA).

Background:
The need to enhance the pace of research on Covid-19 related topics was realized by many scientists and some databases have been created in an attempt to address the urgent needs. But there were limitations with such efforts.

Method: We used careful literature search, prioritized selection of articles and manual curation of key findings to create a new database. Attempts were made to use help from volunteer researchers, mainly students and unemployed youth across India for the manual curation. MySQL and python scripts were used.

Results: A multilingual model database has been created. It has a) collection of other relevant research publications that are categorized more systematically than before; b) auto-summaries of research findings in two-three short sentences generated using carefully selected NLPAs along with significant amount of manually curated research findings; c) an option for pre-qualified 'qualified' scientists to disclose their findings even before writing their first draft of a complete report. It also attempts to cover many other Covid-19 related scientific and educational information. The database facilitates the effective dissemination of research outcomes to the public. It can counter the propagation and impact of false information among the public, and possibly support uniform and effective policy decisions across the globe. The database serves as a useful source for Covid-19 information. But more importantly, it offers a model to develop a future-ready online platform that can help the world fight any future pandemics better.

Poster

AIBC-ABS-97

A comparative analysis of AI based tools and manual compilation of biological information extraction

Saifudeen Ashikha, Goswami Moushumi, Jayakumar N Nandhini, Thalathoti S Lizzi, Umesh Akkash, Srivatsan Harini, Biswas Lipsa, Acharya K Kshitish, 3Institute of Bioinformatics and Applied Biotechnology (IBAB), Bengaluru 560 100, Karnataka (www.ibab.ac.in); Manipal Academy of Higher Education (MAHE), Manipal – 576104, Karnataka, India; Shodhaka Life Sciences Pvt. Ltd., Bengaluru (https://www.shodhaka.net); Bangalore Bioinnovation Centre, Biotech Park, Electronics City Phase I, Bengaluru (https://www.bioinnovationcentre.com); Biological data Analyzers' Association (BdataA)(https://startbioinfo.org/BdataA)

Background: The advent of artificial intelligence (AI) has impacted, and has a potential future impact, the efficiency of literature search and curation. Many common search engines like Google have started incorporating AI. Several AI-powered chatbots also indicate efficient auto-information retrieval. Hence, there is a need to specifically study the efficiency of information retrieval from AI-powered search engines and chatbots from the perspective of biocurators and general biologists. We have also made efforts to compare the efficiencies of auto-summarization generated via natural language processing (NLPAs), with manual summarization.

Methods: AI-based search engines (SEs) were listed by using multiple search engines, and shortlisted for their suitability based on different types of queries. Four cases of literature search/data compilation were carefully identified and the efficiency of the shortlisted tools were compared by analyzing the results. Auto summarization results using multiple NLPAs were also compared with manual summaries.

Results: A total of 25 AI based search engines were listed. Twelve of them were found appropriate for the types of research exercises considered, which represented the diversity of specific for life science-based searches or general searches. The results indicated high diversity among the tools and manual curation results were always superior, even though this process was remarkably slow.

Poster

AIBC-ABS-99

Fusion Transcriptome of Cicer arietinum: New insight into Nature's Genetic Mosaic

Fiza Hamid, Jawaharlal Nehru University

Due to advancements in high-throughput sequencing, a large amount of biological data is available, which facilitates the in-depth study of transcriptome diversity in various organisms. This study focuses on exploring the transcriptome complexity in Cicer arietinum due to fusion events.

Fusion transcripts (FTs) are chimeric RNA formed by the joining of individual genes either at the DNA level or at the RNA level. Here, we present fusion transcriptome profiles for five organs of Cicer and at two abiotic stress conditions using paired-end Illumina data. From all the detected FTs, we selected high-confidence 32 FTs predicted by at least two fusion-detection tools and randomly validated 7 FTs via PCR followed by Sanger sequencing, which are the first FTs described in Cicer. Further, RT-PCR was performed to study their expression in different tissues and abiotic stresses.

Using 181 publicly available RNA-seq data, we performed in silico validation of FTs to check their frequency of occurrence. Without further assembly of Illumina reads, full-length FTs were identified using PacBio long-read data. Junction-sequence analysis reveals the presence of canonical splice sites at the breakpoint, which indicate their generation via trans-splicing. To further support the trans-splicing model of fusion formation, we evaluated the hybridization potential of parental transcripts involved in fusion by RNA-RNA interaction analysis, assessing the possibility of complementary base-pairing that could bring these transcripts into proximity and thereby facilitate their formation via trans-splicing.

Overall, our study provides detailed insight into the fusion landscape in Cicer and their expression profiles to identify potential regulatory FTs.

Poster

AIBC-ABS-100

Identification of Fusion Transcripts in Arabidopsis thaliana: Unveiling Novel Insights into Molecular Dynamics

Simran, Jawaharlal Nehru University

With advancements in high-throughput sequencing technologies, the exploration of fusion transcripts originating from DNA or RNA splicing events, has emerged as a pivotal avenue in unraveling the intricacies of plant genome complexity. While initially recognized in the context of cancer, where chromosomal rearrangements often lead to oncogenic fusion genes, recent research has uncovered their presence and significance in various biological processes, including those in plants.

In this study, we profiled a total of 80170 uniquely expressed fusion transcripts from 1280 publicly available Illumina paired-end RNA-Seq datasets of non-mutants under various tissues and stress-specific conditions using multiple fusion detection callers. The most frequently occurring fusion events identified in this profiling were subsequently subjected to validation through RT-PCR and Sanger sequencing. Further analysis of fusion transcripts showed that they often come from genes involved in essential processes, suggesting their importance in different tissues and stress conditions. Moreover, to investigate the occurrence of these fusion transcripts under particular stress and tissue conditions, we employed the same analytical pipeline to our library to identify the overlapping fusion events between our in-house constructed library and publicly available RNA-seq data, providing additional insights into the reproducibility and consistency of fusion transcript identification across different experimental setups.

Despite the identification of a substantial number of fusion transcripts, their functional significance remains largely unexplored. This study lays the groundwork for future investigations to unravel the specific roles and regulatory mechanisms associated with these fusion events, highlighting a crucial avenue for further exploration within the intricate landscape of plant molecular biology.

Poster

AIBC-ABS-102

Mining of novel microRNA: target module(s) in tomato (Solanum lycopersicum L.)

Adesh Kumar, Sonia Balyan, Apoorva Gupta, Saloni Mathur, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi

MicroRNAs (miRNAs) are short non-coding RNA molecules, typically 20-24 nucleotides in length, in eukaryotes. They play pivotal roles in regulating plant growth, development and responses to biotic and abiotic stresses. Despite their significance, many miRNAs in various plants remain unidentified due to their low expression levels or species-specific features like fleshy fruit in tomato versus dehiscent fruits in Arabidopsis. The advent of next-generation sequencing technologies has facilitated the generation of extensive data, leading to a surge in the discovery of miRNAs in diverse dicot and monocot model plants, as well as various crop species. The miR-PREFeR prediction tool, that leverages its prediction on miRNA expression and strict plant miRNA annotation criteria, was employed for miRNA prediction from 78 small RNA datasets representing different developmental stages in tomatoes, comprising nearly 530 million reads. The analysis unveiled 923 unique miRNAs coded by 1153 precursors. Twenty-two miRNAs, exhibiting a mapping efficiency above 0.8 on precursors and high tag numbers, were selected and screened for precursor stem-loop secondary structure formation. Further, the targets were identified using CleaveLand bioinformatics pipeline on degradome libraries obtained from the PmiREN miRNA web server, and additional prediction using the plant small RNA target analysis server, psRNATarget. Putative novel miRNAs were cloned and validated through sequencing. To establish miRNA:target pairs, transient effector:reporter assays, expression profiling of the module in different tomato tissues and RNA ligase–mediated rapid amplification of cDNA ends to confirm miRNA-mediated cleavage of the targets is being employed.

Poster

AIBC-ABS-104

Vajra- the osteo AI, the AI-ML based disease risk prediction tool

Abhishek Singh1, Abhishek Kem2, Ringyao Jajo1, Farooque Kamran3, Ujjwala Sharma3, Shivani Kansal1, Shailendra Vyas2, Saurabh Raghuvanshi1, Vijay Sharma3*,1. University of Delhi, South Campus, Delhi 2. Bioheaven360, New Delhi 3. All India Institute of Medical Sciences, New Delhi. *Corresponding author.

Osteoporosis, a prevalent skeletal disorder characterized by reduced bone density and increased fracture risk, poses a significant public health challenge. Traditional risk assessment methods often rely on demographic and clinical factors, potentially overlooking complex interactions that contribute to the disease. The diagnosis is based on an expensive DEXA test that involves radioactivity, potentially limiting its use on a patient. This study, named as Vajra- the osteo AI, introduces a pioneering application of machine learning (ML) in the context of osteoporosis risk assessment in Indian cohort of 1529 individuals. Vajra leverages machine learning algorithms to enhance the accuracy and individualization of osteoporosis risk prediction. The methodology employed involves collection of a diverse dataset comprising lifestyle factors, medical history, genetic, and other bone health indicators which were used for detailed and methodical testing of the machine learning algorithms such as Random Forests, Support Vector Machines, and XGboost.Among these, we found XGBoost to be the best suited for our prediction model. The ML model built showcases a sensitivity of 67% while specificity of 78%. The inclusion of sensitivity and specificity metrics not only enhances the predictive performance of the models but also helps in making informed healthcare decisions for more effective preventive measures and early interventions.

Poster

AIBC-ABS-105

Targeted Approach Towards Genomic Selection for Rice Breeding by Trait Phenotype Prediction Based on an AI Tool

Shivani Kansal, Utkarsh Raghuvanshi, Abhishek Singh and Saurabh Raghuvanshi, Department of Plant Molecular Biology, University of Delhi South Campus

With increasing population and changing climate, sustainable agriculture along with improved crop yield is the pressing priority among the scientific as well as breeder community. To achieve the agricultural output goals breeders constantly need to produce new high yielding varieties that are tolerant to the various abiotic stresses as well as resistant to biotic pathogens that can potentially damage the crop. To solve this problem, one approach that exploits molecular genetic markers to design novel breeding programs is Genomic Selection. It provides opportunities to increase genetic gain of complex traits per unit time and cost by evaluating the presented candidate varieties and predicting their Genomic Estimate of Breeding Value or GEBV thus accelerating the process of selection of superior genotypes and the subsequent breeding cycle. GS involves three steps- prediction model training, prediction of breeding value and finally selection of favorable individual based on the predicted GEBV. The first two steps fundamentally make use of Artificial Intelligence for building and training of the prediction model. The current study involves prediction of rice crop phenotype for grain and panicle related traits which are all agronomically important traits required by breeders and farmers alike for assessing the value of a variety. The model has been built based on XGBoost algorithm and its further modifications. It would eventually be integrated into a web-based tool wherein any user could submit their genotype and predict the phenotype. The advantage herein is that this pipeline uses only the genotype data to predict valuable agronomic traits without going to the field for labor-intensive and time-consuming phenotyping exercise saving loads of time as well as cost.

Poster

AIBC-ABS-106

FAIRsharing curation and community

Ramon Granell, Allyson Lister, Milo Thurston, Delphine Dauga, Prakhyat Gailani, Susanna-Assunta Sansone, University of Oxford

FAIRsharing is a scientific and educational resource on data and metadata standards (1714 records), inter-related to databases (2074 records) and data policies (172 records). FAIRsharing promotes FAIR data through the description of these resources across all disciplines (77% of records are related to Life Science).

Each record has more than 20 attributes including organisations, licence, access points and type-specific fields (e.g policies have 15 unique fields). All the data in FAIRsharing is manually curated, with the help of computing tools, by:

1) Record Maintainers, who are the maintainers/developers of a standard, database or policy. They come to FAIRsharing to create/update the records describing their resources (e.g. a researcher creates a database she helped to develop or a journal editor creates a data policy for his journal) using the FAIRsharing web edit system. They created 264 records in 2023.

2) Champions, who are volunteers that perform curation tasks related to their domain of expertise, having access to edit all FAIRsharing records. There are currently 18 active Champions. They updated and created 331 and 47 records, respectively, in 2023.

3) FAIRsharing curators, who check updates performed by the previous two groups, using a dedicated curation dashboard to make decisions (e.g. accept/reject changes or maintainer requests). NLP tools are also used to enrich the content using data from other repositories, e.g. ROR. They updated and created 2159 and 58 records, respectively, in 2023.

FAIRsharing curators and Champions also work together to create educational material that helps all FAIRsharing curators and users.

Poster

AIBC-ABS-110

IBIA: Indian Biological Images Archive

Arun Sharma, Deepak T. Nair, Indian Biological Data Centre, Regional Centre for Biotechnology, NCR Biotech Science Cluster, 3rd Milestone, Faridabad-Gurugram Expressway, Faridabad, India

Imaging biological entities is essential to explore and understand the nature of complexity present in living organisms. In modern era, the images play a vital role in the diagnosis and treatment of diseases, identification of cellular processes, estimation of crop yields, identification of pests, nutritional deficiencies in plants, etc. Thus, it is imperative to systematically capture, store and analyze these images on digital platforms. Globally, efforts have been made by the researchers in this direction. The publicly available resources such as BioImage Archive (BIA), BioStudies, The Cancer Imaging Archive (TCIA), Image Data Resource (IDR), Electron Microscopy Public Image Archive (EMPIAR), are originated due to such efforts. The actual utility of biological resources depends on the consideration and implementation of Findability, Accessibility, Interoperability, and, Reusability (FAIR) principles, during their development. However, the common standards are not available to accept all types of imaging modalities data. The progress is being made to achieve this goal. Some community suggested guidelines such as Recommended Metadata for Biological Images (REMBI) and standards are becoming available to standardize rules for the collection of biological images and associated metadata. Motivated by the community efforts and considering their importance, we are developing an online Indian Biological Images Archive (IBIA). The archive will store all types of biological images such as histopathology, digital X-ray, microscopy, CT, PET, ultrasound, MRI, etc. IBIA will be a comprehensive resource to manage biological images and associated metadata in a systematic manner through five well-defined sections namely, Project, Study, Sample, Experiment, and Run Upload.

Poster

AIBC-ABS-111

Indian Crop Phenome Database (ICPD)

Nivedita Yadav, Isha Saini, Mayuri Jain, Sonia Balyan, Deepak T. Nair, Indian Biological Data Centre, Regional Centre for Biotechnology Faridabad-121001 INDIA

Phenotypic data is the characteristics or traits of an organism and is useful for knowing how genes manifest in observable traits and how these traits interact with the environment. This data is critical for crop improvement, as it provides insights into the performance, adaptability, and resilience of different crop varieties under varying environmental conditions.

Currently, there's no dedicated submission database for crop phenome data. While repositories exist for genetic and genomic data. As a leading agricultural nation, INDIA generates a vast amount of phenotypic data from agricultural trials. Many phenotypic datasets, generated through field trials, may remain unpublished or inaccessible to the broader scientific community due to the absence of a suitable platform for deposition and sharing and resulting in loss of data. The lack of a standardized platform for crop phenome data submission leads to inconsistencies in formats, metadata, and quality standards, hindering data integration and knowledge advancement in crop research.

In addressing these challenges, the Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad, INDIA, developed the Indian Crop Phenome Database (ICPD) for the digitization of crop phenome data. ICPD would serve as a single-stop user-friendly centralized repository where researchers, breeders, and stakeholders can deposit, access, share and analyze phenotypic data across different crops, following FAIR (Findable, Accessible, Interoperable, and Re-usable) data principles.

The present work highlights the establishment of a dedicated submission database for crop phenome data. ICPD facilitates the efficient management and sharing of phenotypic information as well as fosters collaboration and accelerates progress toward addressing global challenges in food security, climate resilience, and sustainable agriculture. We assign unique and persistent IBDC accessions to data submitted to ICPD and the data portal is accessible at https://ibdc.rcb.res.in/icpd/

KEYWORDS: Crops, Phenotypic Data, Data Submission, FAIR data principles

Poster

AIBC-ABS-112

Enhancing Data Carpentry for Fungal Drug Resistance with fundamental machine learning

Aakriti Jain, Neelja Singhal, and Manish Kumar, Department of Biophysics, University of Delhi South Campus, New Delhi

The development of drug resistance in fungal pathogens results in therapeutic failures. Studies targeted at understanding the underlying mechanisms of drug resistance have indicated involvement of both quantitative and/or qualitative alteration of drug targets. Broadly, qualitative alterations bring about changes in the sequence of proteins and quantitative ones change their gene expression levels. These sequential changes or mutations are thus, important biomarkers for evaluating drug efficacy and testing drug competency.

With the objective of enhancing analytical insights of such data we have developed a generalized natural language processing (NLP) based data-mining pipeline for systematic extraction of mutation-drug target(gene/protein)-drug relations and an R based visualization tool for holistic analysis of mutation and drug-resistance, named as Mute-Drug-Spot1 (Fig 1 & 2). We used our pipeline to extract these relations for antifungal drugs and tested a total of 13,516 research articles, extracted 6786 favorable articles, yielding 3691 unique mutation-antifungal drug relations belonging to 32 fungal species. We have implemented this data in the form of a database, AFRbase2 (Jain et al., 2023). This pipeline can thus enable a comprehensive analysis of mutations yielding powerful insights about fundamental factors concerning drug design and target selection.

Coherent analysis of mutation, drug targets and drug relationships could be useful in designing novel drugs and evaluating the potential of the existing ones against novel targets. Additionally, the significance of this pipeline is not limited to fungal molecular biomarkers but can also be applied in the advancement of precision medicine.

Poster

AIBC-ABS-113

IPD: Indian Proteome Databank

Abhisek Kumar Behera, Pawan Kumar, Nitu Kumari, Deepak T. Nair, IBDC, RCB

Mass spectrometry-based proteomics has emerged as a powerful tool for exploring the critical molecular players in basic and translational research that enables a growing number of protein identification datasets. The Proteomics Society, India (PSI) conducts annual conferences and workshops for Indian proteomic researchers to share ideas and research collaborations in a common platform. To facilitate the ever-growing demands of proteomics researchers, we present a novel and integrated proteomics database, IPD, that aims to streamline data access, analysis, and interpretation. The database has several unique features, including high-speed file uploading and a flexible file management system, and prioritizes data quality and standardization, ensuring reliability and reproducibility across experiments. So, IPD is an initiative of IBDC to compile, organize, and disseminate proteomic data derived from various tissues, cells, and biological fluids of individuals across different regions and ethnicities in India and around the globe.

Poster

AIBC-ABS-115

Promoting DOME-ML annotations in GigaScience Press

Christopher I Hunter, GigaDB, GigaScience Press, Hong Kong.

Background

Machine learning is increasingly applied to OMICS data, and there is a need for sufficient detail to enable a researcher to understand the machine learning approach used in a research study. To incentivise the generation of sufficiently detailed annotation, GigaScience Press has partnered with the DOME Consortium with the goal of encouraging authors to follow the DOME (Data, Optimisation, Model, Evaluation) recommendations.

Methods

The DOME Consortium has generated the DOME Wizard (https://dome.ds-wizard.org/) which enables researchers to submit their DOME annotations to a central repository and share them with reviewers. The GigaScience DataBase (GigaDB) team scans submitted manuscripts for machine learning content, requests supporting DOME annotations to be submitted using the DOME Wizard, and performs checks to ensure that DOME annotations in support of GigaScience and GigaByte manuscripts are sufficiently complete.

Results

To increase the visibility of the supporting DOME annotation, at the time of manuscript publication, a link to the supporting DOME annotation is included in the supporting GigaDB dataset that accompanies a GigaScience or GigaByte manuscript.

Conclusion

The DOME annotations are a great asset to the peer review, providing the necessary high-level overview to properly understand a machine learning study. We recommend that other journals follow our example in encouraging DOME annotations to be submitted early in the publication process and prior to peer-review.

Poster

AIBC-ABS-116

GigaDB dataset curation as a means to increase transparency and trust

Yannan Fan, Mary Ann Tuli, Chris Armit, Christopher I Hunter, GigaDB, GigaScience Press, Hong Kong.

Background

GigaScience Press now publishes two journals, GigaScience and GigaByte, both have extremely high standards for transparency and reproducibility. In order to maintain those standards, we have a team of data curators (biocurators) whose job is to ensure the manuscripts are as transparent and reproducible as possible before publishing. Currently, we follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles to ensure scientific data transparency and complement that with the CARE (Collective benefit, Author to control, Responsibility, Ethics) principles to encourage consideration of both people and purpose in the publications. Furthermore, the TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles are implemented with FAIR to allow users to benefit directly through our implementation of best practices in digital preservation.

Methods

The data curation team is exposed to the scientific manuscripts at two points in the submission process; Initially at the beginning, just after the editorial team has agreed that a paper is within the scope of the journal; and then again after the peer-reviewers have recommended its acceptance.

Results

At the initial stage, the curator advises the authors on what data should be made available and how, along with advice and assistance on submission to public repositories. At the later stage, the data curators ensure that all the input data, output data, and required methodologies are both available and suitably annotated, again offering advice and guidance to the authors. Finally, through our association with DataCite, each dataset in GigaDB is assigned a Digital Object Identifier (DOI) that can be used as a standard citation for future use of these data in other articles by the authors and other researchers.

Conclusion

In this way GigaScience Press maintains an extremely high standard of transparency in the research it publishes, which we believe increases the level of trust our readers can have in the research we publish.

Poster

AIBC-ABS-117

Unravelling the Mycobacterium tuberculosis associated bacterial species from meta-analysis of lung microbiomes

Sucheta Patila, Ruma Banerjeea, Sunitha Manjaria, Uddhavesh Sonavanea, Rajendra Joshia*, High Performance Computing - Medical and Bioinformatics Applications, Centre for Development of Advanced Computing (C-DAC), Panchavati, Pashan, Pune

Tuberculosis (TB) remains a global threat in respiratory infections. The traditional knowledge on disease physiology is extensive however, there exists a gap in understanding microbial modulation in TB. Recent metagenomic studies highlight the crucial association between TB and microbial imbalance but their inconsistent findings demand for a large sample size covering maximum human diversity. A careful meta-analysis of these studies is essential to establish a consensus on the signature species associated with Mycobacterium tuberculosis (MTB).

In the present investigation, we included amplicon-metagenomic studies comprising 1819 samples representing diverse sources including sputum, bronchoalveolar lavage (BAL), biopsy, and stool from healthy and TB infected candidates. Raw data from public repositories was processed using Qiime2 to obtain amplicon sequence variants (ASVs) with taxonomic annotation.

We recovered a robust dataset of 3549 taxa from 1155 samples after careful merging and filtering of taxa. Diversity analysis displayed a clear clustering of samples by their source type than any other variables like disease-status, reference study, or the variable regions used for sample preparation. This suggests a strong influence of tissue environment in determining microbial community and highlights the unbiased merging of data. Differential abundance analyses identified a number of significantly differentially abundant bacterial species along with Mycobacterium, in MTB-positive BAL and sputum samples.

Our investigation, with consistent pre-processing across various amplicon-metagenomics studies, offers interesting insight into signature species associated with MTB. It also provides a robust and scalable method for subsequent exploration.

Poster

AIBC-ABS-117

Unravelling the Mycobacterium tuberculosis associated bacterial species from meta-analysis of lung microbiomes

Sucheta Patila, Ruma Banerjeea, Sunitha Manjaria, Uddhavesh Sonavanea, Rajendra Joshia*, High Performance Computing - Medical and Bioinformatics Applications, Centre for Development of Advanced Computing (C-DAC), Panchavati, Pashan, Pune

Tuberculosis (TB) remains a global threat in respiratory infections. The traditional knowledge on disease physiology is extensive however, there exists a gap in understanding microbial modulation in TB. Recent metagenomic studies highlight the crucial association between TB and microbial imbalance but their inconsistent findings demand for a large sample size covering maximum human diversity. A careful meta-analysis of these studies is essential to establish a consensus on the signature species associated with Mycobacterium tuberculosis (MTB).

In the present investigation, we included amplicon-metagenomic studies comprising 1819 samples representing diverse sources including sputum, bronchoalveolar lavage (BAL), biopsy, and stool from healthy and TB infected candidates. Raw data from public repositories was processed using Qiime2 to obtain amplicon sequence variants (ASVs) with taxonomic annotation.

We recovered a robust dataset of 3549 taxa from 1155 samples after careful merging and filtering of taxa. Diversity analysis displayed a clear clustering of samples by their source type than any other variables like disease-status, reference study, or the variable regions used for sample preparation. This suggests a strong influence of tissue environment in determining microbial community and highlights the unbiased merging of data. Differential abundance analyses identified a number of significantly differentially abundant bacterial species along with Mycobacterium, in MTB-positive BAL and sputum samples.

Our investigation, with consistent pre-processing across various amplicon-metagenomics studies, offers interesting insight into signature species associated with MTB. It also provides a robust and scalable method for subsequent exploration.

ORGANIZING PARTNERS

सत्यमेव जयते
Department of Biotechnology
Ministry of Science and Technology
Government of India

United Nations
Educational, Scientific and
Cultural Organization

क्षेत्रीय जैव प्रौद्योगिकी केन्द्र
Regional Centre
for Biotechnology

ISB International Society for Biocuration

SPONSORS

iNSERB DIA

aws

LOCUZ

Hewlett Packard Enterprise

Redington

RITTAL

ddn

Netweb TECHNOLOGIES

Tyrone

premas life sciences

REPLICA BIOTECH