Recommendations for developing predictive and systems medicine for drug discovery in India

Surat Parvatam* and Sham Bharadwaj

Centre for Predictive Human Model Systems, Atal Incubation Centre, Centre for Cellular and Molecular Biology, Hyderabad 500 039, India

Biological phenomena often emerge based on the interaction between pathways, cells and tissues, rather than a single set of genes or proteins. This has led to the emergence of systems medicine. Predictive medicine is another emerging field that aims to predict the disease onset, progression, deterioration, risk and treatment strategies. In this article, we review how systems and computational tools are being used globally in the drug discovery pipeline. With increase in the amount of biological data being generated, data integration is also a critical aspect in systems biology. Towards this, we describe the use of various data integration frameworks. We also analyse the global and local funding patterns, regulations and challenges and propose recommendations to enable India as a key player in this area.

Keywords: Adverse outcome pathways, computational tools, drug discovery, predictive medicine, systems biology.

Incorporating precision and systems paradigm into drug discovery

ADVERSE drug reactions in patients contribute to more than 2 million cases of hospitalization and 100,000 deaths every year in USA. Premarketing studies of drugs are based on preclinical studies in animal models and around 500–3000 human participants for relatively short duration during clinical trials[1]. However, the drug response can vary between individuals based on disease heterogeneity, as well as environmental and genetic factors. For example, substantial frequency differences in genetic variants of drug-metabolizing enzymes and transporters have been observed across various geographical regions[2]. These can lead to changes in local and systemic drug exposure and/or the target leading to variabilities in drug response. This realization has led to increasing discontentedness with the 'one size fits all' paradigm currently being followed in the drug discovery pipeline, which is primarily dependent on a homogenously raised (in bred) test species to understand both drug safety and efficacy before being extrapolated to a highly diverse and heterogenous human population. Dosing is often based on mean values which disregard individual variabilities, including differences in body surface area. These limitations are also reflected in the failure of significant increase in the number of drugs approved by the US Food and Drug Administration (USFDA) every year despite witnessing a linear increase in amount of funding in the drug discovery process. Also, out of the 302 drugs approved during 2008–17 by USFDA, majority were anticancer drugs (17.54%) and biologics (15.56%) with a drop in cardiovascular, neurological, antibiotic and antiviral drugs, indicating huge gaps in the drug discovery process[3].

While India is a single country, studies that have mapped high-density single nucleotide polymorphism (SNP) arrays and genome-wide genotype data specifically from India[4] and southeast Asia[5] show that Indian populations have large amounts of genetic variation that needs to be studied and documented further. This indicates the need for precision methodologies that can account for these genetic variations which in turn may translate to heterogeneities in drug response.

Precision medicine attempts to customize medical treatments and decisions based on the genetic heterogeneities in individuals. Various aspects of precision medicine include probing and identifying genomic variation in a population, leveraging and analysing demographic and clinical data, and high-throughput holistic functional phenotypic profiling which goes beyond genomic, transcriptomic, proteomic and metabolomic differences to measure individual-level phenotypic differences, and sub-grouping of diseases[6].

Apart from the 'one size fits all' approach, disease biology research also usually follows a reductionist approach for understanding various components of a system. However, the paradigm of systems medicine attempts a complementary approach to analyse the interactions between different components within one biological organizational level, and then between the different levels (molecular, cellular, tissue and organism)[7]. This is also reflective of the native biological phenomenon where various components from different pathways often crosstalk and regulate each other. Thus, there is a need to develop methodologies that promote pathway-based approaches to disease biology[8].

Future of medicine: predictive paradigm in precision and systems medicine

Predictive medicine is a relatively new field that seeks to predict the onset, deterioration or reduction in disease progression in an individual, risk associated with a particular disease outcome and its treatment. For example, identification of

*For correspondence. (e-mail: surat.parvatam@ccmb.res.in)

biomarkers, a biological event which can be measured accurately and reproducibly, is one methodology to predict the disease onset or progression and therapeutic outcome. Whereas systems medicine is a complementary approach that analyses interactions between the various components within a biological organization (genome, transcriptome, proteome) and then between the different organizational levels[7].

However, one of the key elements of both predictive precision medicine and pathway-based systems medicine is informatics. Thus, methodologies to analyse, integrate and interpret biological data along with simulation and visualization methods are key for building mathematical models of biological processes, including cellular interaction networks in these fields[9]. Several tools and techniques are being currently developed in the various areas of informatics, including data collection (such as data mining), data analysis (such machine learning) and predictive modelling.

The last decade there has seen a rise in complex heterogeneous data, including omics data, such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, interactomics, pharmacogenomics; biomedical and clinical data. Big data are often described using the six Vs, viz. value, volume, velocity, variety, veracity and variability[10]. This has also led to the parallel evolution of big data analytics and data science to analyse and interpret large and complex datasets[11]. Apart from analysis, there has also been a development of platforms to collect, clean, store, transform, transfer and visualize the data in appropriate and user-friendly formats[12]. Big data analytics has immense potential in tackling long-standing challenges in drug development, including therapeutic target discovery, prioritization of candidate drugs, clinical toxicity and machine-learning methods. For example, an analysis of genomic biomarkers in sporadic breast cancer indicated a 70-gene combinatorial signature that served as a biomarker with a 83% accuracy to predict poor prognosis[13,14]. Such analysis would be extremely difficult using conventional methods of investigation.

Machine learning and deep learning are also being employed in various aspects of the drug discovery process. For example, quantitative structure–activity relationship (QSAR) is a commonly used tool to predict on-and off-target activities, and QSAR datasets consist of a large number of compounds (>100,000) and descriptors (>1000). This leads to computational challenges for prioritizing drug compounds. Thus, various machine learning methods have been applied to QSAR to achieve a good prediction rate[15,16].

However, traditional algorithms often have difficulties with processing raw data, which then requires manual extraction of data features for representation. Deep learning algorithms are providing solutions in this arena where they can automatically extract data features from raw data[17].

It is also well accepted that in contrast to single-candidate approaches, diseases are often outcomes of networks of genes or pathways. Informatics also assists in pathway-based systems biology approaches where disease networks attempt to connect diseases to biological pathways via overlapping genes[18]. Network-based cluster approaches are also used for drug repurposing to identify biological networks that share similar properties, or for discovering novel relationships between networks, sub-networks or groups[19–22].

Need for data integration frameworks

While data informatics and analytics form the basis of predictive modelling and systems medicine, highly structured information is one of the main steps of informatics. The advancements in bioinformatics, omics and other high-content and high-throughput technologies are leading to huge amounts of data being generated every year at various levels of biological organization (molecular, cellular, tissue and organ level), life states (gestational, neonatal, early and late development), gender and model organisms. The information at each of these steps or stages is invaluable however, a systems understanding of human biology will rely on developing frameworks that can integrate this information. While tools for generating new and more resolved data are being rapidly developed, there needs to be parallel evolution of tools that could assist in integrating or structuring vast amounts of data in varied format and biological organization.

The adverse outcome pathway (AOP) framework promoted by Organisation for Economic Co-operation and Development (OECD) is one such open-access, crowd-sourced framework that attempts to integrate the information that currently exists for various levels of biological organization (molecular, cellular, tissue, organ and organism) and developmental stages to provide mechanistic and pathway-based understanding of various adverse events. The AOP framework was initially proposed by the US Environmental Protection Agency (EPA) as a conceptual framework to support ecotoxicology research and risk assessment[23]. The OECD AOP framework captures existing biological information related to the linear sequence of how a stressor interacts at a the molecular level, how molecular perturbation leads to a measurable change at the cellular and tissue levels, and how the tissue-level changes cause an adverse effect at the level of organism or population[24].

The key steps of AOP include the molecular initiating event (MIE) that captures the molecular-level interaction of a stressor; key events (KEs) that capture essential and measurable cellular and tissue-level changes due to molecular perturbation and finally, the adverse outcome that usually has regulatory significance (Figure 1). The causal relationship between the two KEs is captured by the key event relationships (KERs).

A single AOP depicts a linear sequence of events that connects a molecular-level interaction to a population- or organism-level adverse outcome. However, real-world situations may often involve single stressors acting via multiple pathways or multiple stressors inducing the same adverse outcomes. Thus, single linear AOPs cannot be considered in isolation while understanding biological phenomena and the interactions between pathways and KEs need to be incorporated to understand the effects resulting from a single or multiple stressors. In this light, AOP networks which are an assembly of two or more AOPs with common KEs, are viewed as the most likely unit of predicting biological events. In an AOP network, single linear AOPs interact with each other, where the KEs and MIEs are shared between different AOPs (Figure 2). This network, that more closely resembles biology, represents the functional unit of a biological system[25].

The role of AOPs in the development of computational predictive models

The AOPs are also subject to weight of evidence (WoE) analysis to assess their maturity and level of confidence, and are accordingly scored 'high', 'moderate' or 'low'. The OECD AOP guidance document consists of a template for WoE evaluation that is based on the three Bradford hill criteria of biological plausibility, essentiality of KEs and associated empirical support[26]. These WoE evaluations provide a qualitative assessment of feasibility of various outcomes that are based on MIEs and KEs, and the predictivity of the associated adverse outcomes[27,28].

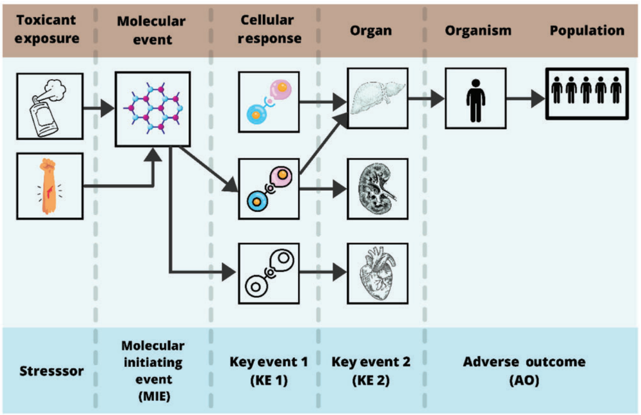


Figure 1. Adverse outcome pathways (AOPs).

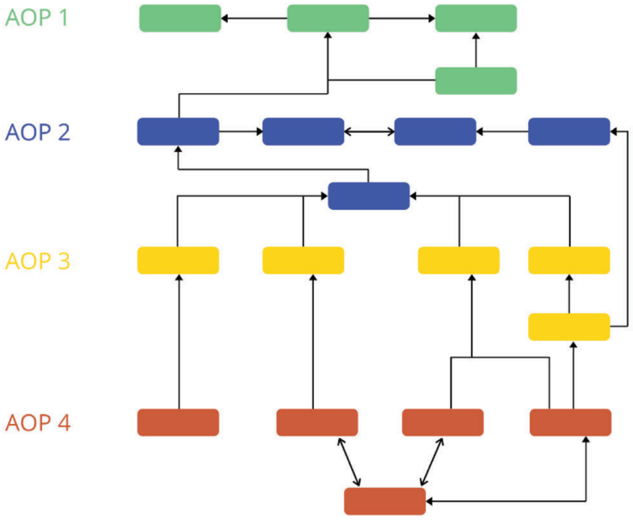


Figure 2. Representation of an AOP network. Each colour box represents various events of an AOP and the connecting arrows show cross-talk between AOPs.

The structured knowledge that is depicted in an AOP can assist in reducing the overwhelming complexity of a biological phenomenon to its essential elements; and this reductionist approach can be extremely useful during the development of a predictive model. The identification of the MIE in an AOP can also provide insightful information regarding the initial molecular interactions with respect to QSAR models and chemical categories that may have relevance in the context of a particular adverse outcome. An AOP also provides the context of each KE and level of biological organization, developmental stage, gender, etc. associated with it, indicating and suggesting the scope or the boundary conditions under which a model may operate. Additionally, each AOP also provides information regarding the methodology through which a particular KE may be measured. This can be critical to determine the available data and how additional data to inform a predictive model may be generated[29].

For instance, several cellular and molecular events are known to be critical for the functioning of the central nervous system (CNS) and peripheral nervous system (PNS). However, the high degree of biological complexity in CNS and PNS has led to challenges in understanding and establishing causative relationships between chemical exposures to an adverse outcome in the nervous system. A recent study proposed the use of potential or putative AOPs for developmental or adult neurotoxicity towards developing predictive models of neutotoxicity[30].

While qualitative AOPs provide a framework for hazard assessment and indicate which hazards can possibly be connected with a biological perturbation, they may not be sufficient to predict the probability of an adverse outcome under a specified exposure situation. Thus, quantitative adverse outcomes are also being developed to provide quantitative understanding of the transition from one KE to the next, critical factors for modulating such relationships, and quantitative prediction of the probability or severity of the adverse outcome[31,32]. Needless to say, such quantitative information can be invaluable during the development of quantitative predictive models.

National and global status of data integration frameworks

Global status of development and promotion of the frameworks

OECD launched the AOP knowledge base (KB) to serve as a hub for AOP-related information (Figure 3). The KB includes

AOP-Wiki, AOPXplorer, Effectopedia and Intermediate Effects Database. The main entry point into the AOP KB is the eAOP portal (https://aopkb.oecd.org/). However, the primary repository of qualitative information for the international AOP development efforts is AOP-Wiki (https://aopwiki.org), which is a user-friendly platform to support browsing and searching for AOPs, KEs, KERs and stressors. AOP-Wiki provides off-line access via creation of each AOP in html and pdf formats. AOP-Wiki is a joint effort between the European Commission (EU-DG Joint Research Centre) and the EPA. In 2013, the Society for the Advancement of Adverse Outcome Pathways (SAAOP) was formed for hosting AOP-Wiki and promoting the development of AOPs.

Currently, there are 330 AOPs in AOP-Wiki (as on 6 May 2021) and they span adverse events in various areas of biology. We mapped the most prevalent and represented areas of biology in AOP-Wiki and found that reproductive disorders, cancer and mortality constitute three of the widely covered adverse end-points in it (Table 1). This also indicates the untapped potential of many of the other highly relevant human diseases and end-points. We also examined the geographical distribution of AOPs that have been submitted to AOP-Wiki. This was done via mapping the current geographical location of the workplace of authors who have contributed to AOP-Wiki. We found that the distribution was highly skewed towards, the US and Europe, with few or no AOPs submitted from countries such as India (Figure 4).

This data points towards the need for AOP outreach for including more researchers from biologically and geographically diverse areas to contribute to the growing database of AOP-Wiki.

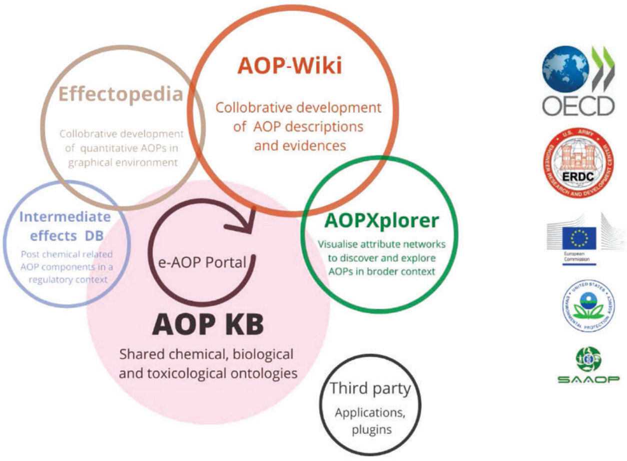

AOP-Wiki
Collobrative development of AOP descriptions and evidences
Effectopedia
Collobrative development of quantitative AOPs in graphical environment
Intermediate effects DB
Post chemical related AOP components in a regulatory context
e-AOP Portal
AOP KB
Shared chemical, biological and toxicological ontologies
AOPXplorer
Visualise attribute networks to discover and explore AOPs in broder context
Third party
Applications, plugins
OECD
ERDC
European Commission
SAAOP

Figure 3. AOP knowledge base (KB).

Table 1. Mapping of biological end-points currently covered in AOP-Wiki

Broad subject area	Number of related adverse outcome pathways (AOPs) in the AOP-Wiki
Reproductive dysfunction	51
Cancer	46
Miscellaneous	45
Mortality	29
Population increase/decline	26
Liver dysfunction	22
Colony loss/failure	21
Growth and development dysfunction	19
Endocrine dysfunction	15
Lung dysfunction	12
Neuroscience	9
Kidney dysfunction	8
Cardiac dysfunction	6
Learning and memory impairment	5
Neurodegeneration	5
Gastric disorders	4
Increased predation	3
Immune dysfunction	3
Hereditary mutations	1

Status of development of the frameworks in India

In a public–private partnership, the Department of Biotechnology (DBT) under the Ministry of Science and Technology, Government of India (GoI) along with Persistent Systems Pvt Ltd, a technology services company in India, the ambitious MANAV ATLAS Project in 2019. This is a crowd-sourced citizen science project to annotate and collate the human biological data that exist in public databases, and map how changes at the molecular level affect the organ and human body. Such frameworks are powerful tools for biological organization. In addition, the project also regularly conducts webinars on topics such as 'How to read scientific literature', various aspects of data science and its applications, from astronomy to biology and public health, etc. It also provides training to students for understanding and extracting relevant information from the scientific literature using digital annotation tools.

However, apart from a few initiatives, research and development for the use and promotion of frameworks that can integrate existing data is still in its infancy in India.

Global and national status on the use of systems and computational biology

Research

The pathway-based information that is being generated using the array of new tools, such as omics and high-throughput approaches is being increasingly tackled employing systems and network-based methods to derive insights on human development and disease. Recently, EPA's chemical safety sustainability research is developing a 'virtual embryo' and 'virtual thyroid' model using systems biology-based tools. In these models different data types, including *in vitro*, *in vivo* and *in silico* are integrated and used to simulate key steps during foetal development which would help regulators better understand the developmental risks

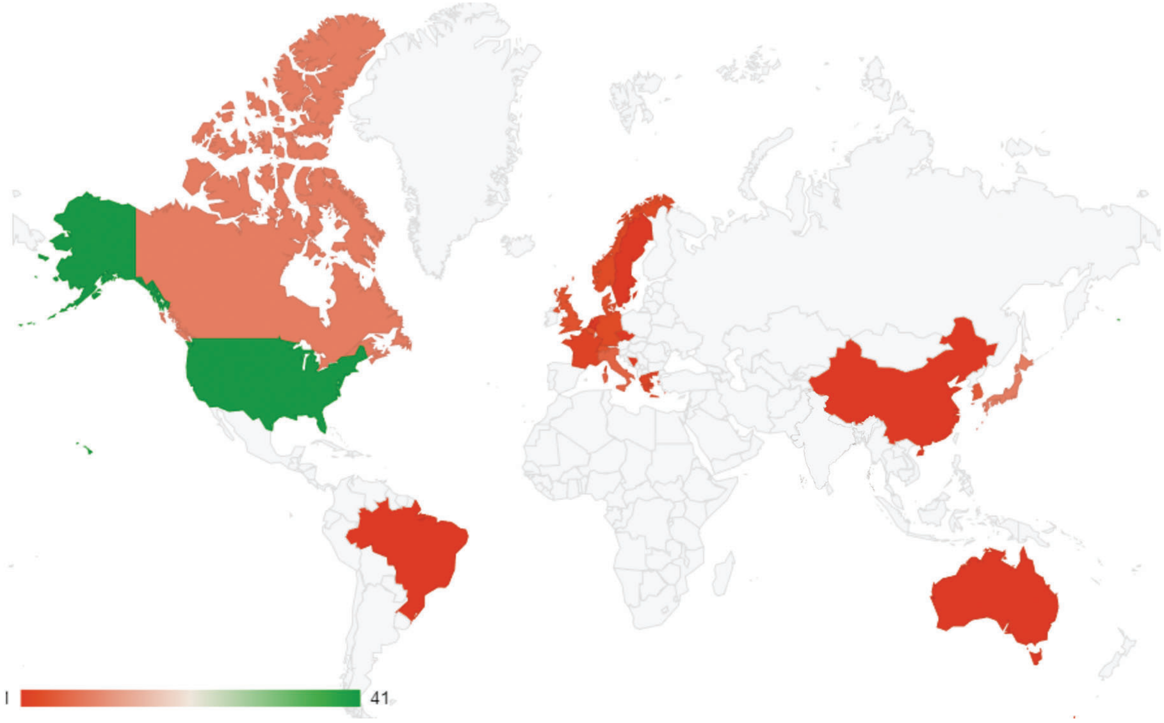


Figure 4. Geographical mapping of the workplace of authors contributing to AOP-Wiki.

posed by chemicals and other environmental stressors[33]. The Virtual Mouse Brain is another opensource system that allows neuroscientists to automatically extract structural and functional connectomes using diffusion-weighted MRI and fMRI data, and use various methods for image processing, tractography reconstruction and connectome generation[34]. Apart from simulating individual organs, another area where systems and computational tools have a significant footprint is the understanding of molecular networks and processes, and how they interact with each other. For example, The Cancer Genome Atlas Program[35], the flagship program of the National Cancer Institute, National Institutes of Health (NIH), USA, has characterized over 20,000 primary cancer samples and matched normal samples from 33 cancer types, leading to the generation of 2.5 petabytes of genomic, epigenomic, transcriptomic and proteomic data. This program aims to integrate several layers of molecular information, including genome, transcriptome, proteome and metabolome, thus providing multilayered insights to cancer biology.

The Disease Maps Project is a community-driven resource with a focus on knowledge-based representation of disease mechanisms[36]. It includes data involving disease-related signalling, metabolic and gene regulatory processes, evidence towards pathophysiological causes, clinical data and outcomes. Such multi-scale management of knowledge can be used to develop computational disease models and advanced data interpretation tools.

Systems biology approaches have also been used for understanding various facets of infectious diseases, including discovery (data collection and analysis), modelling and visualizing complex datasets and the interpretation and prediction of outcomes[37]. Machine learning methods that incorporate electronic health and clinical data have also been recently used to predict disease outcome and severity[38].

Global market growth and patterns of investment

The global computational biology market size which was valued at USD 2.9 billion in 2018 is expected to see a Compound Annual Growth Rate (CAGR) of 21.3% over the forecast period and reach USD 13.6 billion by 2026. Various factors have contributed to this expected growth, including rise in the R&D for drug discovery and predictive models, population-based sequencing projects such as the human genome project and increased funding. The computational biology market can be further categorized into drug discovery and disease modelling, cellular and biological simulation, pre-clinical research, clinical trials and human body simulation software. While cellular and biological simulation leads the market share, drug discovery and disease modelling is projected to be the fastest growing sub-area during the forecast period[39].

USA: The US Department of Energy (DoE) provides majority of the funding in this space in USA. To develop a predictive understanding of complex biological, earth and environmental system, the Biological and Environmental Research (BER) programme of DoE supports transformative science and scientific user facilities. In 2020, the Systems Science Division of DoE received US$ 404.8 million funding to understanding complex interactions that determine the

Table 2. Major global funding initiatives in systems and computational biology

Country	Initiatives	Period	Funding
USA	DoE: Systems Science Division	FY 2020	US$ 404.8 million
USA	NIGMS: Biophysics, Biomedical, Technology and Computational Biosciences	FY 2020	US$ 572 million
EU	IBISBA	2017–21	€ 5 million
EU	Horizon 2020 Future and Emerging Technologies (FET) Proactive – Boosting Emerging Technologies Program	2018–20	€ 184 million
EU	Infrastructure for Systems Biology for EU (ISBE)	2012–15	€ 4.74 million
EU	BioS: Digital Skills on Computational Biology	2018–20	Multi-stakeholder project EU contribution – € 999.6 million
EU	Casym: Coordinating Systems Medicine across Europe	2012–16	€ 2.9 million
EU	ERASysAPP	2013–15	€ 17.79 million
EU	Innovative Systems Toxicology for Alternatives to Animal Testing (InnoSysTox)	2015–18	€ 2.9 million
EU	ERACoSysMed Collaboration on Systems Medicine	2015–20	€ 4.9 million
EU (Germany)	LiSyM – Research Network Systems Medicine of the Liver	2016–20	€ 20 million
EU (Germany)	MED – Systems Medicine	2015–18	€ 200 million
UK	EPSCRC – Biological Informatics	2019	€ 1.4 million
UK	BBSRC – Systems Biology	2019	€ 19.6 million
UK	BBSRC – Bioinformatics and Biological Resource (BBR) Fund	2019	€ 2 million

function of biological systems. Around US$ 572 million extramural funds were provided by the Biophysics, Biomedical Technology and Computational Biosciences Division of the National Institute of General Medical Sciences (NIGMS) under NIH in 2020. One of the major NIH-supported projects for the research and development of innovative tools in using big data and data science in biomedical research is ‘Big Data to Knowledge’[40]. This programme has helped in the development of more than 200 software tools for tackling the challenges associated with funding and accessing biomedical datasets. Another initiative of NIH includes the ‘The Brain Research through Advancing Innovative Neurotechnologies (BRAIN)’ project that aims to map how cells and complex networks interact in time and space using innovative tools, thus revolutionizing our understanding of the human brain. Under this project, single-nucleus RNA-sequencing analysis was performed on various cell types in temporal gyrus of the human cortex, which revealed extensive differences between the cell types of mouse and human, such as differences in proportions, distributions, gene expression and morphology of brain cell types[41]. Such studies also help in resolving and providing mechanistic understanding behind the failure of animal models.

EU: One of the biggest flagship programme’s of the European Union Research and Innovation initiative is the Horizon 2020 Project available over 7 years (2014–20). In 2020, € 184 million was allocated to Future and Emerging Technologies (FETs), one of the initiatives under this umbrella and the funding areas under this project include AI and cognition, Bio Neuro-ICT (information and communication technologies), complexity, human–computer interactions, etc. Another EU programme is the € 4.74 million infrastructure for Systems Biology in Europe (ISBE) designed to meet the infrastructural needs of European systems biology. ‘BioS: Digital Skills on Computational Biology is a multi-stakeholder initiative of which € 999.6 million was contributed by EU, which was approved in the European Framework of Erasmus+/Sector Skills Alliances Programme. This initiative aims to advance the digital skills of medical doctors through the design, development and delivery of new tools associated with computational biology. Another project launched by the EU is the IBISBA programme for which € 5 million was allocated. This programme aims to provide infrastructure to carry out research, development and innovation activities to various stakeholders of the industrial biotechnology sector. ERASysApp–ER NET for Systems Biology Network was one of the programmes launched to enhance and improve research opportunities in the field of systems biology; a total of € 17.79 million was allocated to this scheme from 2013 to 2015.

UK: The funding towards research in the UK can be primarily divided into two programmes – Biotechnology and Biological Sciences Research Council (BBSRC), and Engineering and Physical Sciences Research Council (EPSRC). The latter has a Biological Informatics Division that is involved with the understanding of information processing in biological systems, such as development of novel computational methods for analysing data and modelling of biological systems. The division provided a total of £ 1.4 million in grants in 2019. Funding of £ 19.6 million was provided by BBSRC in the field of systems biology and £ 2 million was provided to bioinformatics and biological resource fund.

Table 2 highlights some of the main funding initiatives in USA, EU and UK.

Government funding and regulatory initiatives in India

The computational and systems biology ecosystem of India has seen rapid growth with an exponential increase in the information technology (IT) sector. There are nearly 40,000

IT companies in India, which indicates a population with computational skills. Steps to advance this field had begun as early as in 1986 in India when a nation-wide bioinformatics system was initiated. This programme assisted in the transfer and exchange of information, knowledge and technology within the country. India was the first country in the world to establish a Biotechnology Information System Network (BTISNet) in 1987. This initiative, started by DBT, Ministry of Science and Technology, GoI, provided infrastructure, education, manpower and tools in bioinformatics. In addition, seven Centres of Excellence (CoEs), 11 Distributed Information Centres (DICs), and 70 Bioinformatics Infrastructure Facilities for Biology Teaching through Bioinformatics (BTBIs) were supported by BTISNet. A Virtual Public Network with high speed and high bandwidth, named as BioGrid India, was also established by DBT to allow exchange of databases and software created or acquired by individual nodes of BTIS.

Department of Biotechnology

DBT constitutes one of the largest funding bodies of science in India and sanctions projects in various categories, including plant biotechnology, animal biotechnology, basic research in modern biology, biomedical engineering, biosciences, medical biotechnology, neuroscience and nanotechnology in biology, public health, food and nutrition, theoretical and computational biology, and biosystems and bioprocessing technologies every year. In 2019, artificial intelligence was added to the above list. DBT also announced a Call for Proposals in 2019 for building Bioinformatics Centres (BCs) across the country, where each Centre would be based on a core theme spanning various areas of computational biology.

A task force on 'Bio-informatics, Computational and Systems Biology' has been set up by DBT for developing computation-based tools, conducting data-driven R&D, developing algorithms in biological sciences, enhancing big data analysis skills, capacity building; and for encouraging networking and collaboration. The Flagship Consortium Project on TBRice Bioinformatics, Mango Database, Interactive Visual Diagnostic software to check nutrient deficiency in crops, etc. are some of the projects supported by DBT.

We estimated the yearly funding allocated to systems and computational biology projects by DBT during 2013–20, and around 12.75% of the total funding was allotted to systems/computational projects during that period (Figure 5). The data were collected from the DBT website which provides year-wise details of projects and funding amounts approved.

Department of Science and Technology

The Science and Engineering Research Board (SERB) was set up under the Department of Science and Technology (DST) in 2008 by an Act of Indian parliament. This body was initiated to allocate funds for dedicated projects in the areas of chemical sciences, earth atmospheric sciences, engineering sciences, life sciences, mathematical sciences and physical sciences. We collected the list of projects that have been approved by SERB during 2015–19 to assess the number of projects and amount allocated under the umbrella of 'systems and computational biology'. We found that an average of 1.6% of the SERB funds was allocated to systems/computational biology projects during 2015–19 (Figure 5).

The SERB grant scheme 'MATRIX' is geared towards the areas of theoretical sciences, including all areas of science and engineering (other than mathematical and allied areas). This scheme includes a grant of ₹ 2 lakhs per annum plus overheads for a period of three years.

Indian Council of Medical Research

The Indian Council of Medical Research (ICMR), one of the oldest medical research bodies in the world, is the apex body in India for formulating, coordinating and promoting biomedical research. The Informatics, Systems, and Research Management (ISRM) Division was set up by ICMR in 2017 to nucleate, promote and support biomedical informatics through services, focused programmes and CoEs. This Division received around 6% of the total ICMR funds for the year 2017–18.

While DBT, SERB and ICMR have set up specific divisions and task forces to strengthen these fields in India in the past few years, there is still a huge scope for improvement in the percentage of funds allocated towards research, development and training in these areas.

Status of research in India

A comparison of model systems used in India for research has shown that *in silico* model systems are the second most popular ones (mice and rat models being the first), where almost 24% of the published studies in India use *in silico* methodologies[42]. A further analysis of sub-divisions of computational biology research in the country showed high frequency of computational research was focused on omics, sequence analysis and databases (Figure 6).

Disease phenotypes often manifest due to malfunctioning of many genes rather than one or two key genes, and these group of genes are referred to as 'disease modules'. A research team in the Indian Institute of Technology (IIT) Madras, Chennai, proposed a detection algorithm to identify the core modules of disease phenotypes using heterogeneous datasets of genes/proteins[43]. Identification of such disease modules can help in understanding the critical nodes of a disease and these nodes can act as points of focus for therapeutic targets[43].

While annotation of protein functions is key to understanding molecular events in our body, performing experimental

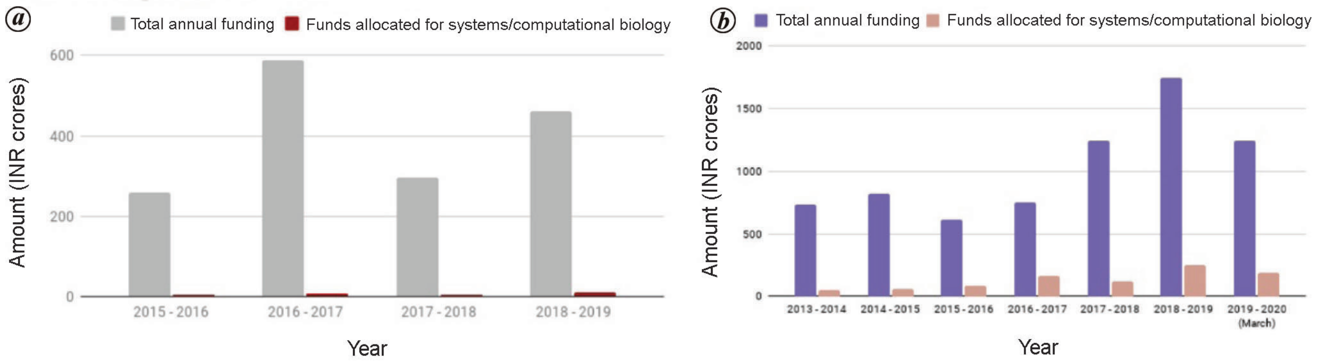
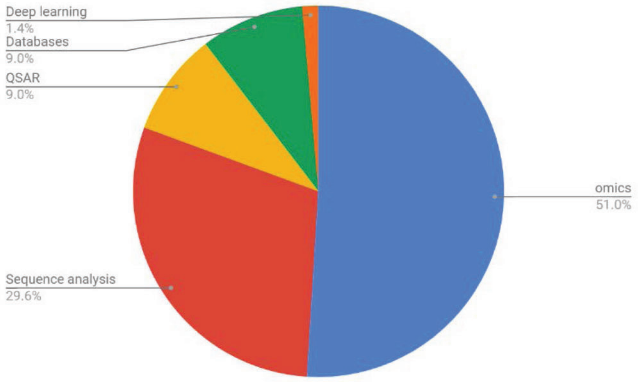


Figure 5. ***a***, Funding granted by the Science and Engineering Board, Department of Science and Technology towards systems and computational biology (2015–19). ***b***, Funding granted by the Department of Biotechnology, Ministry of Science and Technology, Government of India, towards systems and computational biology (2013–2020).



Area of publication	No. of papers
Network analysis	2628
Omics	36,508
Sequence analysis	21,195
QSAR	6443
Databases	6461
Deep learning	1003
Total since 2000	74,238

Figure 6. Mapping of sub-areas of computational biology research in India.

studies for all DNA-binding proteins (DBPs) under all biological contexts will be an expensive and time-consuming exercise. Thus, many potential protein–DNA interactions remain unknown. In a recent study, researchers from the Jawaharlal Nehru University, New Delhi, utilized deep convolutional neural networks to predict DNA-binding proteins using sequences which showed higher accuracy compared to other models with similar profiles[44].

Network analysis has also been performed to study gene expression networks in the post-mortem brain for various neurodegenerative disorders such as schizophrenia and bipolar disorder compared to normal adult. A research group from Tezpur University, Assam used RNAseq data from databases and discovered unique and overlapping gene expression networks for various disorders[45].

Recommendations to develop systems and predictive biology in India

Encouraging collaborations between biologists, computational scientists and clinicians

To encourage novel research in this interdisciplinary field of systems and computational tools, more creative collaborations are required between scientists trained in various fields, such as pharmacology, systems biology, bioengineering, pharmaceutical science, cell and molecular biology, chemical biology, genetics and bioinformatics. Many research institutions in India, barring a few such as the various Indian Institutes of Science Education and Research (IISERs), IITs, Indian Institute of Science (IISc), Bengaluru and Tata Institute of Fundamental Research (TIFR), Mumbai, have been established for predominantly one discipline, such as life sciences, chemistry or physics. This reduces the interaction of scientists from different domains of science, limits crosstalk and collaboration. Thus, we must focus on establishing more interdisciplinary Centres of Higher Education and Research.

Including end-users in the model/framework development

More models of engagement with the end-users of a technology, such as the pharma company, need to be established during the technology development process. This would help in understanding their concerns and needs, which could be incorporated during the development process. This step in essential to scale the technology once it is developed.

Increasing dialogue between regulatory agencies, funding bodies and academia

At various stages of development of a technology, a forum where the new advances can be shared, discussed and critiqued by various stakeholders is needed. This is required to keep the regulatory agencies and government bodies involved and updated about the current advancements in the field. It would also help in addressing their concerns regarding the technology itself, and ease the path towards regulatory approval.

Developing and maintaining Indian clinical databases

Databases on biological data from the Indian population can provide the foundation to perform meta-analysis, and build and validate quantitative models. While such databases are currently lacking in the country, DBT has recently released the Draft Biological Data Storage, Access, and Sharing Policy of India[46]. This Policy states that 'Data generated from publicly-funded projects should be shared openly for public good, with few restrictions and in a timely manner, safeguarding the ethical issues that may arise out of shared data'. The types of data covered in the Policy include DNA and RNA sequence data, genotype data, epigenomic data, microbiome data, protein structure, mass spectrometry, flow cytometry and imaging data.

However, preclinical animal data and clinical data are currently not included in the Policy. Creation and analysis of large-scale and managed access omic datasets of patients and their treatment history can help in understanding disease and drug responses in various Indian populations. Such databases can also assist in building and training *in silico* models and subsequently validating them.

Training programmes and workshops to create awareness and expertise

There are currently few on-line or off-line short-term training programmes for either students or faculty in the area of systems and computational tools and technologies. A core set of skills should be defined to pursue specific domains of this field, for example, systems pharmacology. In addition, creating a national database of such training programmes can help in raising awareness among students.

Encouraging interdisciplinary education and research

Most colleges and institutions in India still provide the binary choice between mathematics and biology. Systems biology is an interdisciplinary field with an overlap of various fields, such as engineering, mathematics, statistics, computer science and biology. Some institutions in the country, like IISERs and IITs can provide interdisciplinary training as they include biology, chemistry, mathematics and physics disciplines in their curriculum. However, most institutions in India provide an option of either biology or engineering. This makes the students ill-trained to cope up with the nuances of both engineering and biology. Thus, a well-structured interdisciplinary curriculum should be designed for students to understand interdisciplinary subjects.

Conclusion

India currently has several programmes and initiatives that are geared towards enabling research in various fields of systems and computational biology. The Council of Scientific and Industrial Research (CSIR), GoI recently launched 'IndiGen', an ambitious project aimed to sequence the whole genomes of 10,000 individuals of diverse ethnicity from across the country over the course of the next three years (2019–22)[47]. Thus, India is at an opportune moment where government initiatives and research are tuned to further empower and support the generation of novel technologies. We hope that these analyses and recommendations will further enable the development of emerging technologies associated with systems and computational biology to change the landscape of biomedical research and drug discovery in the country.

1. Berlin, J. A., Glasser, S. C. and Ellenberg, S. S., Adverse event detection in drug development: recommendations and obligations beyond phase 3. *Am. J. Public Health*, 2008, **98**, 1366–1371.
2. Ahmed, S., Zhou, Z., Zhou, J. and Chen, S.-Q., Pharmacogenomics of drug metabolizing enzymes and transporters: relevance to precision medicine. *Genomics, Proteom. Bioinform*., 2016, **14**, 298–313.
3. Batta, A., Kalra, B. S. and Khirasaria, R., Trends in FDA drug approvals over last 2 decades: an observational study. *J. Family Med. Primary Care*, 2020, **9**, 105–114.
4. Xing, J. *et al.*, Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol*., 2010, **11**, R113.
5. Tätte, K. *et al.*, The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci. Rep*., 2019, **9**, 3818.
6. Morgan, A. A., Mooney, S. D., Aronow, B. J. and Brenner, S. E., Precision medicine: data and discovery for improved health and therapy. *Pac. Symp. Biocomput*., 2016, **21**, 243–248.
7. Cardinal-Fernández, P., Nin, N., Ruíz-Cabello, J. and Lorente, J. A., Systems medicine: a new approach to clinical practice. *Arch. Bronconeumol*., 2014, **50**, 444–451.
8. Marshall, L. J., Austin, C. P., Casey, W., Fitzpatrick, S. C. and Willett, C., Recommendations toward a human pathway-based approach to disease research. *Drug Discov. Today*, 2018, **23**, 1824–1832.
9. Berikol, G. B. and Berikol, G., Predictive models in precision medicine. In *Artificial Intelligence in Precision Health* (ed. Barh, D.), Academic Press, USA, 2020, pp. 177–188; doi:10.1016/B978-0-12-817133-2.00007-0.
10. Andreu-Perez, J., Poon, C. C. Y., Merrifield, R. D., Wong, S. T. C. and Yang, G.-Z., Big data for health. *IEEE J. Biomed. Health Informat*., 2015, **19**, 1193–1208.
11. Ristevski, B. and Chen, M., Big data analytics in medicine and healthcare. *J. Integr. Bioinformat*., 2018, **15**.
12. Lillo-Castellano, J. M., Mora-Jiménez, I., Santiago-Mozos, R., Chavarría-Asso, F., Cano-González, A., García-Alberola, A. and Rojo-Álvarez, J. L., Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services. *IEEE J. Biomed. Health Informat*., 2015, **19**(4), 1253–1263.
13. van de Vijver, M. J. *et al.*, A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 2002, **347**, 1999–2009.

14. van't Veer, L. J. *et al.*, Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 2002, **415**, 530–536.
15. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. and Svetnik, V., Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model*., 2015, **55**(2), 263–274.
16. Newby, D., Freitas, A. A. and Ghafourian, T., Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur. J. Med. Chem.*, 2015, **90**, 751–765.
17. Hinton, G. E., Osindero, S. and Teh, Y.-W., A fast learning algorithm for deep belief nets. *Neural Comput*., 2006, **18**, 1527–1554.
18. Li, Y. and Agarwal, P., A pathway-based view of human diseases and disease relationships. *PLoS ONE*, 2009, **4**, e4346.
19. Sander, J., Ester, M., Kriegel, H.-P. and Xu, X., Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min. Knowl. Discov.*, 1998, **2**, 169–194.
20. Chen, B., Ding, Y. and Wild, D. J., Assessing drug target association using semantic linked data. *PLoS Comput. Biol*., 2012, **8**(7), e1002574.
21. Xue, H., Li, J., Xie, H. and Wang, Y., Review of drug repositioning approaches and resources. *Int. J. Biol. Sci.*, 2018, **14**, 1232–1244.
22. Qian, T., Zhu, S. and Hoshida, Y., Use of big data in drug development for precision medicine: an update. *Expert Rev. Precis. Med. Drug Dev.*, 2019, **4**, 189–200.
23. Ankley, G. T. *et al.*, Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem*., 2010, **29**, 730–741.
24. Hecker, M. and LaLone, C. A., Adverse outcome pathways: moving from a scientific concept to an internationally accepted framework. *Environ. Toxicol. Chem*., 2019, **38**, 1152–1163.
25. Knapen, D. *et al.*, Adverse outcome pathway networks I: development and applications. *Environ. Toxicol. Chem.*, 2018, **37**, 1723–1733.
26. OECD, Guidance document on the reporting of defined approaches and individual information sources to be used within integrated approaches to testing and assessment (IATA) for skin sensitisation, Organization for Economic Cooperation Development, 2017; doi:10.1787/9789264279285-en.
27. Tollefsen, K. E. *et al.*, Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regul. Toxicol. Pharmacol*., 2014, **70**, 629–640.
28. LaLone, C. A. *et al.*, Weight of evidence evaluation of a network of adverse outcome pathways linking activation of the nicotinic acetylcholine receptor in honey bees to colony death. *Sci. Total Environ*. 2017, **584–585**, 751–775.
29. Wittwehr, C. *et al.*, How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology. *Toxicol. Sci*., 2017, **155**, 326–336.
30. Bal-Price, A. *et al.*, Developing and applying the adverse outcome pathway concept for understanding and predicting neurotoxicity. *NeuroToxicology*, 2017, **59**, 240–255.
31. Conolly, R. B. *et al.*, Quantitative adverse outcome pathways and their application to predictive toxicology. *Environ. Sci. Technol.*, 2017, **51**, 4661–4672.
32. Perkins, E. J. *et al.*, Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environ. Toxicol. Chem.*, 2019, **38**, 1850–1865.
33. Shipman, M., EPA high-tech 'virtual embryo project' will target developmental risk. Inside EPA's Risk Policy Report 15, no. 2, 2008, pp. 1–6; https://www.jstor.org/stable/26727372.
34. Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A. and Lippert, T., The human brain project: creating a European research infrastructure decode the human brain. *Neuron*, 2016, **96**(3), 574–581; doi:10.1016/j.neuron.2016.10.046; PMID: 27809997.
35. Wang, Z., Jensen, M. A. and Zenklusen, J. C., A practical guide to The Cancer Genome Atlas (TCGA). *Methods Mol. Biol.*, 2016, **1418**, 111–141.
36. Mazein, A. *et al.*, Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl*., 2018, **4**, 1–10.
37. Eckhardt, M., Hultquist, J. F., Kaake, R. M., Hüttenhain, R. and Krogan, N. J., A systems approach to infectious disease. *Nature Rev. Genet.*, 2020, **21**, 339–354.
38. Fisher, C. K., Smith, A. M. and Walsh, J. R., Machine learning for comprehensive forecasting of Alzheimer's disease progression. *Sci. Rep*., 2019, **9**, 13622.
39. Computational biology market size worth $13.6 billion by 2026. March 2019; https://www.grandviewresearch.com/press-release/global-computational-biology-market (accessed on 17 March 2020).
40. Margolis, R. *et al.*, The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Informat. Assoc.*, 2014, **21**, 957–958.
41. Hodge, R. D. *et al.*, Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 2019, **573**, 61–68.
42. Parvatam, S. *et al.*, The need to develop a framework for human-relevant research in India: towards better disease models and drug discovery. *J. Biosci*., 2020, **45**, 144.
43. Tripathi, B. *et al.*, Adapting community detection algorithms for disease module identification in heterogeneous biological networks. *Front. Genet.*, 2019, **10**, 164.
44. Chauhan, S. and Ahmad, S., Enabling full-length evolutionary profiles based deep convolutional neural network for predicting DNA-binding proteins from sequence. *Proteins*, 2020, **88**, 15–30.
45. Sahu, A. *et al.*, Integrative network analysis identifies differential regulation of neuroimmune system in schizophrenia and bipolar disorder. *Brain, Behav. Immun. – Health*, 2020, **2**, 100023.
46. DBT, Biological data storage, access and sharing policy of India – draft 1, Department of Biotechnology, New Delhi, 2019; https://www.nhp.gov.in/NHPfiles/Draft1-Biological_Data_Policy.pdf (accessed on 14 February 2021).
47. Koshy, J., What is 'IndiGen' project that is sequencing Indian genes? *The Hindu*, 3 November 2019; https://www.thehindu.com/sci-tech/sequencing-indian-genes/article29865310.ece#:%7E:text=The%20project%20ties%20in%20with,every%20State%20and%-20diverse%20ethnicities (accessed on 25 May 2021).

Received 7 October 2021; accepted 13 October 2022

doi: 10.18520/cs/v123/i11/1317-1326