STM JOURNALS

International Journal of
Bioinformatics and Computational Biology

ISSN: 2583-9977
Volume 4, Issue 1, 2026
January—June
DOI (Journal): 10.37591/IJBCB

https://journals.stmjournals.com/ijbcb

Review IJBCB

# Digital Frontiers in Life Sciences: The Transformative Role of Computing in Modern Biology

V. Basil Hans*

***Abstract***

*The integration of computers in the biological sciences has revolutionized research and experimentation, facilitating advancements in areas such as genomics, bioinformatics, systems biology, and ecological modeling. The ability to process vast amounts of biological data efficiently has transformed how scientists study complex biological systems and phenomena. Computational tools enable the analysis of DNA sequences, protein structures, metabolic pathways, and ecological dynamics, which were previously beyond the reach of traditional laboratory methods. This article explores the role of computers in biological sciences, discussing key technologies, software, and algorithms that support modern biological research. It highlights the importance of computational biology in predictive modeling, drug design, personalized medicine, and the development of more sustainable agricultural practices. As biological data continue to expand, the synergy between computational tools and biological research will remain essential in addressing global challenges in health, ecology, and biotechnology.*

**Keywords:** DNA sequences, software, biotechnology, genomics, genetic data, computational biology

## INTRODUCTION TO COMPUTATIONAL BIOLOGY

These notes aim to introduce the scope of computational biology and the most common tasks dealt within that line of work, the primary considerations for selecting a programming language, the tools used for biological data handling and numerical computations, and other related software resources. For some, this introduction may contain well-known information, in which case it will be swiftly reviewed. If the material is unfamiliar, on the other hand, only parts of it should be digested. In any event, the exposition moves quickly toward the point where some programming must be done and a more hands-on text used side-by-side with these notes will be necessary. For those readers who require more than the brief introduction to programming that this primer provides–or who struggle with the tasks at hand and wish to bone up on the subject of writing instructions for computers–several seminal texts are recommended. Two excellent references that are also freely available on-line are https://www.ncbi.nlm.nih.gov [1].

***Author for Correspondence**
V. Basil Hans
E-mail: vhans2011@gmail.com

[1]Research Professor, Department of Commerce & Management and Humanities & Social Sciences, Srinivas University, Mangalore, Karnataka, India

Received Date: May 10, 2025
Accepted Date: September 24, 2025
Published Date: January 16, 2026

**Citation:** V. Basil Hans. Digital Frontiers in Life Sciences: The Transformative Role of Computing in Modern Biology. International Journal of Bioinformatics and Computational Biology . 2026; 4(1): 1–20p.

## HISTORY OF COMPUTING IN BIOLOGY

The dream ticket! A computer that tells you what's wrong with an ailing limb, what you should have for dinner, the web-cam choice of TV programs, the best holiday destination tailored to the wintry Arctic breath of a better half and, the answer to the daunting prospect of a crowded dance floor. The computer should also suggest how one can persuade other halves less enchanted by the prospect of dancing! Advertised intelligently as the 'hassle free' approach, it guarantees that one could leaven more time contemplating the meaning of

 

life. Or half an hour in the bath! This after all is the computer age, the new dawn that emerged not so long ago when the biologists found they could teach electrons to sing the song of life. Is already here in the guise of silicon chips printed with the DNA sequence of a parasite that has scientists scratching their heads [2]. Whatever computers might make of life one aspect has been beneficial in the biological sciences: the technology has revolutionized the power of the experimentalist and made for efficiency in the analysis of reams of raw data. This in turn has allowed the biological experiments to be more ambitious and generate more refined hypotheses.

It is scarcely representative to delve into the realms of mathematical biology to elaborate that point, as it could be by considering the principles that govern what it is that a biological experiment can hope to infer about the net of nature. There are almost a universal set of guiding principles that embrace experimental endeavors in all the sciences but they can be obscured at the bench face [1]. Subtle errors of interpretation of experimental observation can destroy the purity of thought so essential to the formulation of theories that apply generally. Such errors are perhaps more prevalent in biology than other sciences where the investigator must confront entropic systems whose perturbation yields results far removed from those expected on the basis of much simpler physical models. It may be that the experimental study of biology has matured sufficiently of late for not only the experimental workers but also the theorists to place gross errors aside and to confront the basic issues. On the other hand, it could be argued that biologists have long followed this dictum–hence the rich panoply of sub-disciplines. Given that it is also true (but less readily acknowledged within the fraternity) that much of modern biology makes a leap of faith when it applique findings from one sub-discipline to another. The upshot of this rather inconclusive debate is that high-powered computation methods have made a major dent recent years of the biological sciences whose subject matter is the what is loosely classified as the related "life-sciences".

As the Post-Genome Age progresses the amount and diversity of biological sequences is increasing exponentially. Apart from purely DNA processing research, life scientists are beginning to study life processes as dynamic systems, and this in turn has led to an exponential increase in the volume and complexity of data which must be analyzed by the techniques of mathematical and computational modelling. Bioinformatics is concerned with the recording, storage and retrieval of nucleotide sequences and amino acid residues of protein structures. Biologists use a variety of programs and tools to aid in their research, especially in the conversion of raw data in a form that can be analyzed. Typically these tools run on different platforms, and are by different manufacturers. To minimize the investment in time and cost of hardware and software, one may consider using the resources that are shared either by or available to the institution on a network. Desktops will be by far the most common medium of interaction with the systems; it is within this environment that a number of X-Windows–PC implementations are available. Faced with a simple menu and mouse click interfaces much familiar to users of the standard window GUI's, they make working from the desktop a straightforward process. For those who need to access command line programs, then Telnet programs are also available. Split screen interfaces allow the user to both view online help pages and execute the program via a menu of pull-down or tab functions.

## KEY CONCEPTS IN BIOINFORMATICS

The purpose of this lesson is to introduce Computational Biology in the study of molecular biology and evolution; more specifically through the analysis of Poliovirus culled from various dating locales in the Soviet Republics. A two-part tutorial begins with the use of Polio. Among other things, the computational hypothesis emerges that the 1980 Turkmenistan outbreak of Poliovirus originated in materials from the same time period. In general Bioinformatics may be defined as the application of computational techniques to problems in molecular biology, typically problems involving large data sets and analytical techniques from statistics and computer science. It is often used in the effort to map and sequence the human genome. Bioinformatics research has produced many such successful bioinformatic problems.

 

Contiguity analzses for large-scale whole-genome assembly of human sequence data. Clustering based on expression patterns in microarray data. These problems require new or improved algorithms to increase computational efficiency and/or statistical yield. Initial sequencing research was conducted by automating solutions originating in classical biostatistics, including the use of discriminant functions to train the computer to differentiate between sequences and non-sequences. This strategy, however, is often computationally expensive. Modern bioinformatic projects include the Finding Algorithm and the Neuroscience Essay Model. At the same time, biomedical informatics projects seek to make IT for the whole population and to close the loop between care and research. Notable too are the Bioinformatics Response Initiative and the Framework for Support of Nanotechnology and Nanoscience it also includes the project.

**Genomics**

Biological computing is an important branch of scientific research in the upcoming years, mostly emphasizing on bioinformatics in computer science, whereas progress in the field of biotechnology plays a role in this research. In this paper, an overview of the advances in general biology that are followed by the advances in computing is given. Furthermore, some convenient guidelines to those topics within bioinformatics are given that are appropriate for basic and computer sciences as well as the availability of the biotechnological laboratory.

Gene expression under different conditions can be studied by mapping and measuring the messenger RNAs that are transcribed (transcriptomics). Some vector systems are used to produce large quantities of mRNAs. The letter method is sensitive (about 1000 molecules), but it is an in vitro method. However, a recent technique can do this kind of measurement in vivo and also among single cell data with cellular resolution. Very large-scale transcriptomics can be performed on microarrays, but only known genes can be probed. Chromatin immunoprecipitation-sequencing (ChIP-Seq) can be used to map DNA-protein interactions, for example, of transcription factors. From these advances, many possibilities of DNA-binding protein factors can also be found that lead to many targets for gene knockouts. Furthermore, advances creative at transcription factor binding events related to both transcriptional activation and repression also has been presented. Computational approaches can also elucidate relationships between targets and cellular processes in ways not easily deduced by unaided human observation. None on mutation analzses.

Computers have become an indispensable tool in biological sciences research. The amount of data produced has grown exponentially in recent years, and checking all this data by hand is not possible. Moreover, it has been established that manual validation is very difficult to reproduce by different researchers, since it depends on a series of studies that are not usually reported (Nijveen, 2013). Because manual analysis is unfeasible despite the small scale of the networks, the data used in this work is publicly available.

**Proteomics**

Shotgun proteomics enables comprehensive identification and quantification of numerous proteins in biological samples. This is necessary to discover the underlying molecular mechanisms for the condition of interest. The proteins extracted from a laboratory culture or an environmental sample are digested into a complex peptide mixture by proteases. Peptides are then separated using liquid chromatography and analyzed by a mass spectrometer. To increase throughput, acquisitions are usually performed in a data-dependent manner. Algorithms have been devised for peptide identification from the spectra and further for protein inference. The acquired tandem mass spectral data (MS/MS) is then searched against a protein database to identify the peptides in the peptide mixture [3]. The relative abundances of the identified proteins can be estimated across samples from multiple biological conditions. This requires a pre-planned design of sample preparation and the mass spectrometry experiment itself to ensure that any systematic bias is minimized. The design and the subsequent extraction of biological samples should be based on biological- and technical-driven replicates. Except for a focused hypothesis, protein quantification can be performed in a label-free or stable isotope

 

labeling manner, and various concepts have been proposed to model and quantify the relative abundance of proteins. On the other hand, there are numerous software tools for proteomics data analysis. For example, pFind3 can support both open and restricted search modes for database searching, along with multiple types of quantification. Wili can estimate peptide-spectrum match (PSM)-level false discovery rate and also support multiple types of precursor ion quantification. FragPipe can combine various post-processing methods, including database search, to better leverage their respective strengths, and provide table summaries that can be readily imported by other tools. There are also proteomics tools available on Galaxy, a web-based platform for research analysis. MaxQuant is a quantitative proteomics software package for analyzing large-scale mass spectrometric data sets. MaxQuant can be connected to Galaxy for data analysis using tools available in the toolshed; however, installation itself may be problematic as it requires both IT and biological knowledge. A conda software package for the analysis of proteomics data, enriched with the most popular standalone bioinformatics tools, can also be used in Galaxy. To overcome these challenges and advance analyzes residing on an HPC cluster, a platform called CloudProteoAnalyzer was developed. On the one side, CloudProteoAnalyzer is a server-based solution that integrates multiple standalone bioinformatics tools and well-established pipelines for protein identification and quantification assembled in Portable Encapsulated Projects (PEP) format. On the other side, this platform provides at the same time a user-friendly web interface, which empowers users to easily select and customize all the parameters for protein identification and quantification.

## Metabolomics

### *Analysis*

Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind – the study of small molecules (typically less than 800 Da). The metabolome represents the complete set of small-molecule metabolites found within a biological sample. As with other omics, metabolomics can be carried out in a hypothesis-generating manner using untargeted techniques and then validated using hypothesis-driven targeted methods.

Metabolomics can give an instant snapshot of the physiology of a cell. Several forms of the technology already on the market can produce interpretable data in a matter of minutes. The speed of response means that unexpected changes, possibly brought about by external agencies, can be traced in real time, even in vivo. It does not require permanent modification of the micro-organism, cell or tissue being studied, so the same cells can be monitored repeatedly over time. If transient, the effects can still be quantified. Metabolomics is a much more sensitive technology than most of the alternatives; it is not unusual for a two- to threefold change in a metabolite to be reported. And applied to food and bacteria, metabolomics can reveal how our body processes food and how food can affect health. The diagnostic potential of metabolomics is clear, though realization of this potential requires the understanding of healthy variation alongside disease variation [4].

## COMPUTATIONAL TOOLS AND SOFTWARE

Computational biology has come a long way in the last 50 years. It originated through the requirement of broadly collecting biological data and later developed and grew over the years by going through numerous challenging events. These requirements also brought the need for developing the computational tools that have since been immensely increased. Originally a computational biologist had to be the developer and end user of these tools. However, in the light of the increasing demands and workload in biophysics, bioinformatics, molecular biology, genomics, and cell biology better tools to handle the forecasting workload were being developed by both commercial and open-source developers. While they are still inadequate, they now are tools that can be implemented to a generic model workflow for biological sciences.

Computational biology evolved through the need of systematic and wide collection, cultivation of biological data. In the earlier years this was done through basic modeling, but with the growing amount and increasing systemic resolution of data, the need for more sophisticated methods was growing promptly. Chemical and physical processes within biological systems have been well described for single molecules or simple systems. It was a natural consequence of this situation for researchers to

 

focus on fundamental units, proteins or genes, and related these fundamental units through their well-understood interactions. Schedules have doubled in every 2 years since the first process of genomic sequencing, therefore, developing and using these high-throughput screening techniques created an exponential increase in the biological data available for computation. While we try to make sense of this data, initially somewhat narrow attempts have been made; most of them have an inherent interest in describing coherence between these effects and biological stochastic distribution and pattern formation in huge ensemble interaction datasets. The requirement for a sophisticated dynamical description has unavoidably grounded beginning experimental biology to emerge and grow, but somehow the inherently huge dynamics and relative constraint on the measurements implies the strong need of accurate, sensitive, and efficient complementary computational tools. From which, a fair amount of research works with specific aspects have been adapted, containing a thorough dataset curations. Subsection 2 of the adaptive generic progression model workflow through it, general components that made this progression started will also be explained.

**Sequence Alignment Tools**

Sequence Alignment Tools convene a fundamental part of any immunoinformatic study as MHC molecules conventionally show a curved peptide shape that is implemented by the sequence and structural properties of the peptide. Encephalitogenic antigenic determinants of proteolipid protein (PLP) concerned in semisyngeneic retinal autoimmune uveitis (SAU) in the model rat (PVG.RAO) have been identified by T-cell lines iterative with synthetic overlapping nonamer peptides, and their unrestricted proliferative responses to synthetic peptides have been analyzed. As now well known, the MHC restriction element in Lewis rats is RT1.B u, whereas in PVG.R have orthologues for other MHC-II molecules, such as the earliest PVG.R ancestors or as the PVG.R.3 (RT1.B^c)-controlled SAU-arousing uveitogenic epitope p116-130, a 15mer peptide, showed that uveitic immune responses were mounted in MHC-compatible but not unrelated regular animals [5].

The multiple sequence alignment (MSA) of the peptides with the most elegiac and protracted reaction was undertaken, but by means of a notable exception, the top and bottom residues, graded by the immune reactivities were not aligned. Alternatively, the two centralmost completely conserved six and seven residues of the 15mer sequences, required in the MHC complexes, were strictly fixed during the MSA concept. Species predilections in the spitefulness of the uveitic reactions in restriction element-appropriate animals and in vivo analysis of selected synthetic peptides guided the leukocyte epitopes into three overlapping nonamers, designated p116-124, p119-127, and p122-130, lying near the N-terminus in the centre, and at the C-terminus, respectively, of the p116-130 peptide. A fundamental part of the primary findings, as well as of the subsequent investigation of the binding characteristics of PLP, have been bracketed in speedily written work notes. This, together with essential time logistics while examining beforehand unimproved antigenic determinants, as the MSA of the protective and the pathogenic determinants in the same antigen required much toil and numerous preparatory washing of the tissue cultures, have delayed the preparation of this work [6].

**Structural Biology Software**

Computational activities of scientists pursuing studies in biological sciences are now an indispensable part of their daily work. Super computers, computer clusters and sophisticated software tools are widely used to help work with ever-improving computational models and large experimental datasets. This fact inspired to prepare a paper devoted to the software side of structural biology, comprising general remarks about two most widely used computational methods for biomacromolecules and multi-molecular systems. This article starts with the description of a number of problems one encounters while dealing with experimental data at its various stages including macromolecular structure determination. Next, problems more specific to biological sciences are discussed and finally problems still waiting to be understood are tackled [7].

Four complementary papers in this issue present improvements on key elements traditionally used by structural biologists: the object being studied, the probe and the means of interrogation. Effort has been made to provide the software for each of these areas along with some critical analysis of its limitations and insights into potential research directions that could lead to further advances.

 

Determination of a unique molecular structure of bio-macromolecules, or at least a small complex of them, is essential to understand whatever process or mechanism occurs in living organism. Many aspects of this intriguing process are not understood yet and the role of computation in this field seems to be even more significant than for others as this understanding will most probably be in silico [8].

### Statistical Analysis Tools

Statistical analyzes of biological data are often complex, involving repeated measures and more than one dependent variable. The presentation is at a introductory level, maximum emphasis being placed on the identification of appropriate t-tests and associated issues. More advanced techniques (analysis of covariance, multiple comparisons, etc.) should be used only when it is clear they are appropriate [9] Again, the emphasis should be placed on the judicious use of these techniques in the biological context. Relevant output should be attached as an appendix. Intriguingly, ninety-five uncorrected p values were less than 0.05, which is 80 (0.05*160). Thirteen uncorrected p values are less than 0.0001. All of these are considered biologically significant. Uncorrected p values less than 0.05 are considered to be biologically significant, only if they are significant in the appropriate FDR adjustments. In order to better understand these increasingly detailed protocols, it is desirable to have a good feel for how each of them works. The program Basic Statistics generates simple, univariate statistics on a variable specified by either position or name for a simple one-dimension biological data matrix, as used here. The variance and standard error of the mean are generated using built-in subroutines. Up to 5000 items may be operated on. This restriction arises from limitations to the testing to 5000 for (i – 1)** 2 > i, in a room where is the number of cards in the room. Outputs are restricted to XKP card printer and XKP print tape [10].

## DATA MANAGEMENT IN BIOLOGICAL RESEARCH

How to efficiently manage your research data and literature is a question with which any biologist has to deal. Modern biological science is an experimental, evidence based science. More than ever modern biology is complex, interdisciplinary and data driven, requiring robust data management infrastructure [11]. Still many research groups in this field maintain data in spreadsheets, lab journals, and personal computers leading to conflicts between group members, wasted resources, lost data or – even worse – non-reproducible research results. It is suggested to use a research data management system that works well with all standard file formats. Data have to be linked together and to be annotated by metadata. A common practice should be used for the names of things. Experiments and analyzes should be designed and described before they are executed. Upon describing all these technical details, your data management system can transform them into plans and check whether they make sense. Similar considerations as to the design of the research data management system also apply to your literature data. It should enable import, export and search of information about the contained scientific literature and support the management of literature types other than just PDFs.

### Data Storage Solutions

While terabyte-scale data storage solutions have become increasingly affordable and widely available, life scientists are now generating data at rates too high to be easily managed on single hard drives. In addition, many bioinformatics applications are memory-limited and too slow with slower changing between data sources. This combination of requirements has led to a group of dedicated would-be bioinformaticians developing and purchasing "cluster" computer facilities to meet their computing needs. Long-term storage of data has been a problem for biologists, but there is increasing need to look after large data sets including much important unpublished material. This is a brief description of the most recent local innovation in computing provision to aid those short on funds and with much data to handle.

### Data Sharing and Collaboration

The Pulsnet / Collaborative Data Platform is a web-based system designed to simplify the environment for data sharing and collaboration. On the most basic level, it provides an organized database structure and an interface for data retrieval and entry. However, a number of more advanced

 

features have been incorporated that make it an extensible representation of collaborative data for systems biology research. Among the key features are a fast full-text search engine, a user-friendly, data domain-specific representation of data, and a flexible, XML-based data format that provides high format extensibility [12].

The Pulsnet / Collaborative Data Platform provides a flexible representation of collaborative data based on XML. It is designed for research collaborations, which involve the heterogeneous 'omics' data along with functional data from validation experiments, genetic and phenotypic data. The representation of the data is completely generic and is based upon a single, annotated. Thus, the introduction of new data types or the modification of existing data types based on the user's collaborative data can be easily accomplished. Of critical importance is the format extensibility feature of the representation schema, which is important as the addition of new data types should be anticipated over the course of the collaboration. The representation of the collaborative data is implemented along with platform-independent parsing methods. In addition, several helper applications that operate on top of the collaborative data representation have been developed. These helper applications include an interactive Pulsnet / Collaborative Data Platform viewer for data-specific, multi-collaborator analysis. Suitable documentation and guidelines of programming interfaces of the representation format as well as helper applications are also provided. The developed software components, including the collaborative data representation format, the parser and the helper applications, are open-source developments and have been designed in close collaboration with network partners involved in experimental systems biological research. As open-source developments, they can be easily adopted by other research consortia involved in research on systems biology.

## MACHINE LEARNING IN BIOLOGY

Biological research has been one of the most exciting and rapidly evolving scientific research areas. High-throughput data, such as microarray and next-generation sequencing data, have facilitated the studies of complex mechanisms underlying biological processes. Due to the massive amount of high dimensional data generated in modern biological research, machine learning approaches have been extensively used in biology. It can help uncover new insights into complex biological systems that are difficult to discover using traditional statistical or mathematical methods. There have been a wide range of machine learning methods developed. These approaches include clustering analysis, different types of discriminant analysis, regression analysis (both linear and nonlinear), neural networks, support vector machines, random forest methods, and boosting methods. This computational analysis was shown to give high predictive accuracies for a wide range of datasets, some of them containing data of a very complex structure.

However, successful prediction is not always associated with an equivalent result in interpretability. In the case of biological data analysis, the model interpretation may be even more important than the model accuracy in various contexts. Were the genes or markers selected during model fitting biologically interpretable? What genes/markers were biologically associated with a given cluster? What are the possible biological reasons behind the prediction/classification made by a particular model? This issue becomes even more important in the era of the big data and machine learning, when the model of data can be very complex and difficult to follow even from the mathematical point of view. In response to these problems, the model interpretation tools gained attention and were intensively developed during the last decade. Most of these tools aimed to increase the transparency of the model, and therefore either could be directly applied to very simple models or aimed to provide some locally interpretable, results for more complex approaches. Still, the results obtained in biological studies, especially for really complex data structures, could be difficult to interpret even in the framework of the model interpretation and require better adaptation to the existing biological knowledge.

### Applications of Machine Learning

Interest in the development and application of machine learning methods in the biological sciences has grown rapidly over the past decade due to the fruitfulness of these methods in analyzing biological

 

data [13]. A number of freeware and user-friendly software are being made available for applying machine learning to biological data analysis. One category of machine learning is supervised learning, a type of regression and classification analysis that builds models of outcomes of interest as a function of input variables or predictors. The goal of supervised learning is to make accurate predictions for new, unseen data. Supervised learning models may be simple or complicated, from linear regression to deep learning. Another category is unsupervised learning, which includes analysis like principal component analysis, clustering, and association rule mining. These unsupervised learning methods do not predict an outcome but rather characterize relationships among the input data. A regularized regression analysis combines complexity with interpretability. In regularized regression (also called penalized regression) the relationship between the independent (input) and dependent (output) variables is estimated by fitting a model to the data and using a penalty factor to control model complexity. Regularized regression varies from simple linear regression in that the ordinary least squares are augmented by a penalty on the size of the coefficients, which shrinks the coefficients toward zero. Depending on the study the Lasso or the Elastic Net may be useful in building models to analyze the complex biological dataset. Each predictor variable is associated with a regression coefficient that must be estimated along with other parameters of the model. The goal of the Lasso-type methods is often to achieve sparse models, that is, to allow individual coefficients to be exactly zero. As a result, interpretable and concise models can be obtained. Additionally, regularization will reduce overfitting and simplify the interpretation by pushing down the coefficients of some predictors. Regularized regression models make inference easier, and the fitted parameters can be boiled down to a minimum number of important features that are associated with the response variable. The coefficients in the regularized regression models are selected by cross-validation of the final model to keep the best hyperparameters for the penalty factors. Due to the requirement of a linear structure, the machine learning models using linear combinations of the input variables to predict the output variable may not have enough prediction power and robustness for a complex biological dataset. For this reason, nonlinear regressions are also employed. Neuronal nets with one hidden layer tend to perform better than different variants such as neuronal nets with one or two units in a single layer. Random forests and regularized regression methods based on neural networks have generally performed the best. Highly correlated dependent variables can impede robust modeling of prediction or classification. Genomewide transcriptional profiling often sets out to find patterns in the genes whose expression significantly changes in response to stimuli. These machine learning analyses identify sets of features that distinguish two given classes of the dataset. A routine procedure is to link the identified genes to RNA Likely classes of that. A library for computational biology, containing more than 2,000 pre-defined gene sets. In the gene sets used, resultant parameters are commonly assigned to groups of genes that are known to act together in some way. Experimentally: elicited both transcriptional and metabolic responses in parallel treating Vitis vinifera cv. Corvina cell suspension cultures with acetaldehyde, a compound involved in the response to both abiotic and biotic stresses. Despite the observation of severe acute effects on cellular respiration and comparisons of time-related changes in both omics data, the relationships between transcripts and metabolites were markedly different. One of the powerful aspects of machine learning methods in biological sciences is the ability to build and refine models, which can be used to predict the behavior of the complex biological system. Built models can be used as computational simulators of dynamic changes in biological systems and can be further used to design experiments and optimize parameters to make the desired output. Model selection often relies on traditional statistical approaches, which are not necessarily able to capture the complex relationships characteristic of biological systems, even after having identified all relevant parameters. This approach yields models that are simpler and, on occasion, more informative than those derived from statistical methods through a combination of literature search, hypothesis-driven network models, and data. These hand-built models could then be used to guide the experimental effort. Models that describe the multivariate relationship among variables are necessarily simplistic, and they may not capture higher order or nonlinear dynamics which are often characteristic of biological systems. Data-driven methods may be better suited for exploratory analysis since they do not require a priori knowledge of the topology of the relationships among variables and may yield a more comprehensive view. Modeling methods are increasingly used to analyze biological data, including statistical inference,

 

network reconstruction, and optimization. The latter models describe how the behavior of the system is determined by a set of controllable variables. Findings indicate that machine learning optimization techniques are very flexible and could be tuned to encode the specifics of a given system. Such models could potentially exploit the full potential of machine learning and, as a consequence, lead to a deeper understanding of complex biological systems.

### Challenges and Limitations

For decades, researchers have been applying computer simulation to address problems in biology. Many of the grand challenges, however, most notably in a transfer from accuracy to how integration across multiple scales, remained unsolved. In recent years, however, the increased power of available computational resources is making it possible to address a new generation of problems, and major inroads have been made even in seemingly intractable areas. An overview of some of these areas is presented, and a discussion of the challenges of both the analytical frameworks–atomic-level force fields and solvation models–and techniques–sampling and enhanced sampling regimes–that are produced by the models. A biophysical approach to a variety of problems in biology faces two major barriers: the accuracy of the input models and the challenge of sampling the essential degrees of freedom of the system.

The unavoidable limitations in models manifest themselves in myriad complicated ways, all contributing to reduced confidence in the results. However, many of the most fundamental limitations that can be drawn directly from the inputs consist of the force fields, the solvation model, and the parameters are swept under the rug within these fields.

Significant stumbling blocks are also the limitations in sampling. The grand challenge of sampling, as reformulated here, is the challenge to efficiently explore the entire equilibrium or stationary-state phase space of a system. The sampling problem is far too great than the current capacity, and dated even a decade ago, it has continued to grow at a remarkable rate. This decade-long trend has resulted in a far greater growth than even Moore's Law, resulting in the acceleration of rough timescales for certain classes of simulation that have grown far faster than comparable algorithmic or computational advances in the area of the broad field of computational biology.

## ETHICS IN COMPUTATIONAL BIOLOGY

Researchers working with large datasets, even those generated from completely public sources, must consider the ethics of their work, particularly now that it is quite easy to re-identify individuals from their data. With the scale of information collected in genomics and other "omic endeavors" it is easy to fool oneself into thinking that there truly is such a thing as anonymized data. With 34 fitness tracker companies making the bulk of their data public in an effort to foster secondary (academic) research, this issue is only going to grow [1] understanding the basic mechanisms underlying population genetics offers a lens through which to view computational biology and bioinformatics. Basic concepts introduce the field and allow for a more in-depth discussion of more complex ideas. Examining case studies on human populations, cattle, and dogs shows how utilizing these concepts in a computational biology or bioinformatics setting can lead to better results. Basic concepts include allele frequency, linkage disequilibrium, migration, drift, population structure, and the infinite allele model (IAM). Indicators of these mechanisms can then be assessed such as Fst, Wright’s Fst, Robertson and Hill’s Fst, F’st, nodes, ancestral alleles/haplotypes, and Runs of Homozygosity (ROH). Applying these indicators to case studies of human populations, cattle, and dogs illustrates how the use of these concepts can lead to a greater ability to correctly interpret the results of such analyses.

### Data Privacy Concerns

Science, of course, informs case investigation and epidemiology, basic and applied biological research, characterization of biologically interesting molecules and sequences, taxonomic conundrums, clinical trials, foundation of public and animal health policies by health departments and the World Health Organization, and last, but not least, peer into the mechanisms of garish, serendipitous birth and often unmerited success of ventures commercial, speculative, and wondrous.

 

Yet-though a gem of wisdom occurs – how do we go after it? First thoughts may turn to bush beating – bullying a grant-funded underling or spending an age fiddling with databases. Yet, another place to begin might be the computer, in some form or other. Computers now run a range of lab equipment, from sequencers to pipettes; and turbo-boost countless program code-crunchers: BLASTing up orthologs, motif finding, gene-finding, simulation modeling. Raw data come in through A-D converters, or via online biological databases, and once you have this piece of biotechnology, what has been found out, what specimen studied, what gene sequence determined? Nowadays, best ask its code, its locus number, and studies cited by the papers sitting in its bibliography, and don't hold on for a conversation [14].

The comparatively few who have the numerical wizardry can run molecular mechanics equations for Nano-tick anti-infection devices or whatever the fantasia de jour happens to be. Most resort to off-the-shelf simulations packages for molecular dynamics and clustering analysis. Beyond these basics, systems for text-mining and bio-logical ontology construction are becoming increasingly common, and could well in time significantly affect the nature of projects and purity of data involved in scholarly publication in the biosciences.

**Ethical Use of Genetic Data**

SPECIAL CONSIDERATIONS: A dataset with high information value has been prepared to serve as target for the manuscript [15]. The data consists of a set of 10000 CDS's and the 9003 possible pairwise comparisons between them. Also the performance of a naive expensive approach and that of Parasail as a function of the number of sequences in the database will be assessed. An essential tool for the biological sciences is providing a dataset that is a perfect match for the biological asks at hand, tailored and trimmed properly in such a way that the main biological questions can be answered with a high level of success. The rise of big data in the biological sciences does not change this situation–if big data is used (after proper de-noising and trimming) the biological questions can be addressed on a large scale. So in the sense that big data is a data-type particularly useful in the biological sciences, it is very much compatible with, and supportive of, science. Rest assured, current work on bioinformatic support for transcriptomics and proteomics aims to provide researchers in those areas of the biological sciences with the best-fitting datasets possible, or protocols to set them up.

There are two primary ways in which individuals began working with datasets that were too large to move, and will generally continue to do so until the data can be made to fit in a laptop. In the first approach, interactive data analysis tools can be run directly on a secured server where the data resides such as Rstudio server, Jupyter notebooks with a python kernel, or the galaxy bioinformatics platform. In this scheme, code operates on the data where they live, and only the tiny percentage of it required to display a few numbers and plots is returned to the analyst's computer. It is expected that most local IT organizations or ELIXIR nodes will have run existing containers with popular analytical tools on their servers, to accommodate researchers working with big data.

## CASE STUDIES IN COMPUTATIONAL BIOLOGY

Computational biology and bioinformatics are distinct, yet related domains in science. Computational biology employs mathematics and algorithms to understand and analyze biomolecular data, whereas bioinformatics can be thought of as the automation of this approach using informatics tools. This chapter explores questions regularly presented to faculty teaching the subject of computational biology and offers experiences and advice from lecturers at different institutions worldwide, each with a unique set of challenges [1]. This chapter discusses lectures and workshops, dealing with the use of examples from the literature, exercise sets, and practical-intensive case studies to illustrate key concepts, and these ideas are further developed here.

As the biological sciences gain increasing access to extensive numerical data types there is an expected parallel increase in the need to use computational resources to analyze and manipulate this data. The "computational" demands of this research encompass coding, data interrogation and

 

processing, design simulation, and data visualization. However, a major problem is that Biologists lack a sufficiently developed knowledge of computing technologies and practices. This ineptness is born from a lack of exposure to computers and a societal structure that does not reward or expect any training or aptitude in these areas. Instead non-theoretical sciences have privileged the development of knowledge and practical expertise in laboratory methodologies. It is not uncommon, therefore, for PhD students to have minimal or no computing support and come from a discipline-oriented education system. To try and encourage an independent approach to using computers this section attempts to answer the most common queries students have when starting out and suggests several packages that will help with statistics, simulation, and data visualization.

**Cancer Genomics**

The word Bioinformatics originated by the expansion of the terms Biology and Informatics and was first coined in early seventies. It refers to the creation and development of advanced information and computational methods that provide new ways to understand and solve biological problems such as those in the domain of developmental biology or behavior. Bioinformatics has brought to light a different and complex perspective for biological analysis. There are three significant approaches–data variety, data size and parameterizing methods. Bioinformatics also tries to learn the complex unmeasured variables by studying the observed variables, determining the dependency of those unmeasured variables on the observed variables.

There are different categories used for Bioinformatics; most notably, for genomics, proteomics, and systems biology. For genomics analysis, the biological insight involves gene identification, sequence comparison, functional and comparative genomics, and genome evolution [16]. Nevertheless, biological research also need to deal with other levels of this information (proteomic and systemic levels), so the interest keeps growing since the completeness of such knowledge allows a better understanding of several aspects of life (embryogenesis, pathologies, etc.). For proteins it involves the computational study of general structure determination, structure-function relationship, structure prediction and modeling, interactive graphics, principal biological functions, sequence motif, applications, and implications for drug design and study of conformational changes. For systems biology, the biological insight involves network properties behind biological systems, control analysis, metabolic control diagnosis, culminating in an understanding of the robustness of biological networks to genetic, environmental changes, and parameter changing of the network, of the spectrum of the network.

**Microbiome Analysis**

With more possibilities came the demand for sufficient resources and the corresponding approach to benefit from them. The computing architecture is comprised of workstations, servers and a storage space. Users can access softwares for writing scripts, analyzing and summarizing results, saving and sharing files, and controlling server operations via a browser. Microbiome research is evolving rapidly and has produced a profusion of experiment possibilities that are available with different resources. Defining and characterizing bacterial communities from numerous ecological and health care circumstances is now less costly and easier. Consequently, the number of 16S rRNA gene sequence optimized for bacteria has increased drastically [17]. The laboratory tasks are more accelerated and defined with generally fast and accurate sequencers while one is left with an assortment of bioinformatics literature on the web tools for the microbiome. CoMA (Comparative Microbiome Analysis) simplifies basic and typical analyzing tasks and describes an extensive evaluation of the methods offered. The pipeline is a client-server application offering a GUI with a local and a server part. CoMA runs on Windows and Linux and is free software. Being comprehensive, CoMA goes beyond a more formal conclusion sensitive to the local situation in microbiome benchmarking to suggest both general and specific recommendations for the application of modern benchmarking standards [18].

 

## FUTURE TRENDS IN COMPUTATIONAL BIOLOGY

### Hardware and Software Nowadays

Surely, elaboration of bioinformatics is closely connected with computers that used both in computational and laboratory sides. That is related to data generation and interpretation in genomics. Development of bioinformatics computer systems essentially determines the successfulness of research. Intelligent decision making in bioinformatics would provide researchers with marketable models and abundant information. Beyond understanding the bottom fabric of the living creatures, researchers also expect to discover ways to prevent or cure illnesses, or to improve the quality of life, and even increase life expectancy. This bioinformatics computer system is also capable to prepare data in laboratory studies. A bioinformatics computer system aims features like high performance, creative solving, low-profit design, reliability, flexibility, implementation and ease of use.

The main topic in that context is the subject that bioinformatics approximates to is for biological sciences. A bioinformatics computer system that dedicated to the biological sciences domains should feature seamless integration of algorithms, visualization, database, and experiments, components, be user-friendly, yet providing opportunities for the customization and augmentation. The user-friendly bioinformatics system should hide algorithmic complexity, high-level interfaces, and visualization that should ease analysis of the processed data. Above all, computational parameters are relieved upon from the UI such as for similarity search algorithms, clustering, or ensemble predictors. Input data, thereinforth, enters the bioinformatics computer system in uniform file formats predefined for computational data. Possible attempt to compute a novel ambulance from the scratch is not taken into account. The user in biological sciences is expected to be an experimental researcher.

Bioinformatics apparatus software is a number of bioinformatics programs running in a personal computer, strategies, WINDOWS NT, Windows2000, Windows 9x (including scientific data systems Waterloo NT, IBM9800, Silicon Graphics, DELL, COMPAQ, HP, and so on). That hardware is affordable to many research groups. Special emphasis is given to recent development and emerging choices in bioinformatics software. Just from a note to the extent that combinations of platforms, massive hardware nullite use of the hardware, improved, aesthetic, numbering and quality of the components is considered. The upcoming features pertain educative commitment and zymotic research output. Additional, a complex bioinformatics reference apparatus is maintained. All beginning or experienced bioinformatics computer scientists is invited to participate in enhancement, modification, and widening of the bioinformatics hardware archives. Bioinformatics hardware-software is placed in the bench as fragments in a complex, interrelated set of hardware and software in the network. Look for the descriptions of the others. A bioinformatics computer system, where the same observation can be evinced for commercially available software. This is justice. The purpose is rather to evaluative the prospective prospective and emerging approaches to software development in bioinformatics. Comments relate to those programs, which are anticipated to be of most concerns for the bioinformatics computer systems daily work, rather than those that are deemed to the irrelevant or inadequate [19].

### Artificial Intelligence Integration

Biomedical researchers are increasingly exploring artificial intelligence (AI) for the analysis of their data to find patterns of specific interest. Bioinformatics primarily focuses on how to efficiently store, retrieve, and analyze biological data. Emphasis is also given to the development of tools for data analysis. Most of the current AI libraries support data analysis after it is prepared in a particular file format. This process involves several steps, including downloading the data from an appropriate data source, converting the data to a desired file format, and then finally analyzing it. Here, EndToPrepML, a web-based, end-to-end pipeline is proposed that can analyze raw data. The library is capable of preprocessing, training, evaluating, and visualizing a variety of AI models in a user-friendly way without manual interventions or any prior coding expertise. A wide range of tasks, such as recognizing, classifying, clustering, or predicting events solely from observation of multimodal, multisensor data, is addressed. This library is tested on several data sets and shows how drug discovery, pathogen classification, medical diagnostics can be performed [20]. Biomedical researchers and companies have

 

produced an enormous number of data in the field of life-science research, medicine, and pharmacy. The data reported in a digital form of scientific articles, patents, and detailed information on chemicals and proteins. It has not been possible to follow the remarkable increase in new knowledge by reading the scientific literature anymore. Modeling and simulation can aid understanding physical and biological systems. By connecting models to data, powerful analytic tools can result. There is a wide variety of mathematical techniques available to build models and compare them to data. In the emerging field of systems biology, dynamic models play a crucial role. However, for the biologically less educated people, these models are often hard to understand. BioUML is an open-source integrated platform that provides homogeneous access to resources and services for a wide audience ranging from modeling to powerful pathway visualizer [21].

**Advancements in Sequencing Technologies**

With the TRF labeling method, the read quality can be increased to a high value, and the maximum read length can reach 2.4 kb. High-quality data like TR80–100 can support the assembly of 2.33 Gb *E. coli* DSM1103. The TRF good reads and better reads assembled N50 values are 154.4 kb and 54.4 kb, respectively [22] The *E. coli* genome was sequenced using the Illumina GA II platform. It was demonstrated that up to 400-bp-long unitigs could be obtained with the trun900clp following 546× genome sequencing coverage. The v25 assembler was then applied in the final genome assembly. Single-molecule real-time (SMRT) sequencing technology was also used to sequence the bacteria O. formigenes.

There are two attempts in bioinformatics analysis methodology development. The first was to develop the only one TRF raw read analysis software and analyze the TRF read-based metagenomic data. The topics of this approach include TRF read filtering, sequence trimming and clustering (de novo assembly), community composition interpretation, and statistical analysis of sequencing coverage of assembled sequences. The bioinformatics analysis flow was developed to complete the above analyses. The second approach was experimental design and bioinformatics methodology development. The experimental conditions can heavily influence the sequencing data quality and characteristics. A total of 20 different Illumina libraries were prepared with the extracted rice shoot and root DNA, and 29 sequencing data sets were generated [23]. The effects of different enzymes used in the preparation of the TruSeq and Nonlinear libraries were compared using the sequencing data. The total number of bases, as well as the GC content of cleaned reads, was significantly different. rpkm saturation analysis was applied to help the guidance of the experimental design.

## INTERDISCIPLINARY COLLABORATION

Modern biology, including medicine, comprises an increasingly large and complex body of knowledge, much credited to the steady stream computer-assisted experimentations and discoveries during the last sixty years. Even though legislative definitions consider a computer system as a set of elements perfectly integrated in order to reach a specific objective, computer hardware, the technological bricks of this ever-growing edifice, keeps evolving at an amazing pace, constantly presenting new structures, new concepts, and new materials and techniques. Biologists and medical researchers, who are also continuously confronted with a complex and often difficult-to-understand technology, increasingly have to complement their usual teaching and training with computer-related education in order to fully exploit the vast potentialities of computing. Quantitative biology issues are pervasively having a field-transforming effect on many specialized branches of life sciences: genetics and genomics, ecological modeling, multiscale bioimaging and biosensors, systems and population biology. On the other hand, life sciences education tends to be very specialized and traditionally split into separate departments, making the exchange of ideas across this rigid institutional border a challenging task. Besides a common technological ground of computers and digital data, there is a theoretical foundation coming from the common ground of stochastic fields and processes shared by many physical phenomena and by cell biology. There is, therefore, a general analytical framework and interpretation that can be openly shared by modelers in very different fields. However, such framework is often depth enough to be out of the reach of biologists and bioengineers traditionally trained in a

 

simplifying, trial-and-error, phenomenological approach. Moreover, detectives and studying dynamics is a very broad field, so that shared discoveries in different fields tend to stay confined in their original domains. While quantitative biology is bringing most of the innovation, often feeling unprepared to use this new weapon effectively, or are present only as partially informed end-users.

### Biologists and Computer Scientists

Biologists and computer scientists have been jointly exploring the frontier between their respective disciplines for decades. All the areas of biological and medical research have become so data rich that only powerful computers can handle them. Meanwhile, the computing revolution has entered a stage where a long-range search for novel applications of this versatile technology is well justified. Understand and make use of the amazing developments in computer technology now seem to call for extensive computer experiences along with the infinite knowledge of biological nature. Tips and tricks various programs and application notes from the experience are required to ease the life of those biological scientists willing to do more with the computer. Such a collection is presented to meet these needs. It should be helpful for experiment designers and research labs either without long practical computer experiences, or without contacts to bioinformatics and programming help. This manuscript is written as a readers' guide, helping in quickly finding what someone needs, when they want to analyze and interpret their biochemical experiments with the help of the computer. It treats 25 selected tasks that can be done using 8 different programs that perform standard or more unusual bioinformatics analyses creatively [24].

### Cross-Disciplinary Research Initiatives

Computational demands of life scientists have grown substantially since the 1980s: molecular biology transitioned to biochemistry, then genomics brought a deluge of sequence data. Data generation remains a challenge, but advances since 2005 in multiple fields are now creating biological big-data. This, most daunting for field and other environmental studies, is collected to provide insights into deep histories and systems-level interactions. Life scientists, mathematically the least computer-savvy, reveal now that an increasing portion of their productivity is constrained by computational needs. Rapidly sinking costs of high-throughput machines have facilitated their acquisition by small, often single-investigator, labs. Simultaneously, growth in computing power has enabled a second throng of life scientists to acquire complex systems and population-scale data from commercial and public sources that they are ill-prepared to analyze. Many appreciate current and imminent needs and are adapting accordingly: purchasing computing solutions, using cloud resources, and consulting with professional bioinformaticians. However, a significant number of either process or interpret data, or use these to direct experimental work. The integration of big-data to study ecology and evolution is a main goal of a number of programs at the National Science Foundation (NSF). Parallel initiatives have been proposed widely to the various funding bodies, and the near-future growth in the complexity of the resulting questions reflects the far-reaching data acquisition of attempts to date. Thus, what must be addressed is physical requirements that would dovetail computing resources with biological equipment and lead to common standards of data management. Towards this objective, this forthcoming fleet of programs will be drafted with specific intent, and actions proposed to arrive at workable solutions will likewise aim at collaborative platforms able to enrich downstream studies [25].

## EDUCATIONAL RESOURCES FOR COMPUTATIONAL BIOLOGY

### Online Courses

Computer literacy for life sciences: helping the digital-era biology undergraduates face today's research. The ubiquity of computers in our lives is undisputed. For every modern-age college student it is virtually a must to own or have access to a computer for him or her to succeed in course work. Almost every assignment in today's classroom requires some computer work. Unfortunately, this does not seem to be true. Many students seem daunted by more sophisticated uses, such as data manipulation and analysis, for the purpose of scientific research. It is widely believed that this is at least partially due to the insufficient preparation of the students in terms of basic mathematical skills, and especially logic and statistics, which are indispensable in life sciences research [26]. This article describes an innovative

 

undergraduate-level course, titled Computer Literacy for Life Sciences, which aims at partially remedying this situation. This course is intended to provide the students with some basic proficiency needed for scientific research in the digital era. The overall purpose of Computer Literacy for Life Sciences is for students to develop a hands-on working experience in using standard computer software tools as well as computer techniques and methodologies used in life sciences research. Scientists must be able to efficiently manage and query the data resulting from wet lab experiments as well as those derived from computer simulations. Such data are then subject to statistical analysis, or other means of data mining, to answer the underlying scientific questions. Finally, researchers should be able to present their discoveries in the form of reports or scientific articles and audiovisual presentations. To fulfill all of the above-mentioned objectives, the students in the Computer Literacy for Life Sciences course are mentored throughout the process of scientific pursuit, which is implemented in the form of several group virtual research projects to be completed with the use of state-of-the-art computer techniques and applications. Each virtual project includes the following components: background literature research, data manipulation, formulation of scientific questions and verifiable hypotheses, statistical data analysis, and presentation of scientific discoveries. The first installment of Computer Literacy for Life Sciences was taught at Emory University in fall semester 2009.

Introduction to Modeling and Simulation for the Life Sciences. This open-access book presents a wide variety of computational issues, from principles to applications. The selection of topics has been inspired by the teaching of modeling and simulation classes within graduate programs in the Biological Sciences, but the book will also be useful to undergraduates as well as other interested individuals or researchers. The chapters have been written to be self-contained, allowing flexibility for the reader. Each chapter presents a computer science topic in combination with an application from the Biological Sciences. This volume includes topics ranging from programming languages to modeling, and from chaotic behavior to cellular automata. Interested readers can use the book without the computational language.

**Workshops and Conferences**

Several conferences and workshops of interest have been or will be held soon. These are either specifically relevant to computational biology, or have been recommended by the Editorial Board as having potentially useful techniques, software or concepts.

Bioinformatics and Medicine is probably the biggest meeting concerning biological simulations within the next months, here is the call for participation: With the completion of the Human Genome Project and a number of other high throughput functional genomics projects the modelling and simulation of complex biological processes from cell dynamics to development becomes an inevitable task. A start towards this has been made by the industrial and nuclear physics earlier; many of the moved mathematical, algorithmic and software technics are now available also for biologists. But this range of projects needs much faster development of the above. Simple porting of some existing simulation algorithms to a new field fails: a detailed and complex mathematical model of the simulated phenomenon and a considerable number of experiments are needed. Simulations must go to the most part of the biotechnological experiments here. To coordinate and elaborate these efforts some new events are needed.

The Industrialization Workshop Series is just sprouted for this, aiming to promote and discuss integration, automation, simulation, quality, availability and standards in the high-throughput life sciences. The first three topics are conventionally adopted in the caBIG pilot node, it is necessary here to avoid very formal and traditional presentations just on them. Interesting information about the likely events occurring in bioinformatics within the next years will be very useful too. The more complete text of previous caBIG-related messages is welcome also.

## FUNDING AND GRANTS FOR COMPUTATIONAL RESEARCH

**For Principal Investigators**

Individual laboratories often independently fund their own infrastructure and perform systems administration for clusters or the cloud. Some funding agencies offer grant mechanisms that support

 

computing hardware or cloud-based computing; others support computation costs on existing hardware or the cloud. Such grants can be written entirely in collaboration with existing computational expertise within a university.

**For Computational Biologists**

The development, debugging and benchmarking of scripts and software for computational projects can take as much time as wet-lab work. However, most grant applications prioritize data generation and biological research, not bioinformatics. Computational biologists should be vigilant about which software tools are needed for proposed projects and apply for assistance before the grant application deadline. Funding can be requested for co-development of computational methods or pipelines with an in-house bioinformatician, core facility or just to contract a bioinformatician.

**Recommendations to Institutions and Funding**

Biological research continues shifting towards large data analysis. Fund most individuals in big-data experimental biology projects, not just those running sequencing and metabolomics experiments. Ensure bioinformaticians are involved from the grant writing stages. Salary support for graduate students to additionally spend significant time on the development of computational tools would benefit both the bioinformatician's projects and the training of the grad student. Many public funding agencies now have guidelines stating that public funds must generate public results.

Most guidelines suggest that data and code should be made available. Compliance with these guidelines should be evaluated and mechanisms that facilitate and support the data publishing process should be created. Best practices should be publicly available.

**Government Funding Opportunities**

Recent efforts in the United States to expand access to high performance computing (HPC) for small- to mid-size academic institutions hold great promise for broadening the demographic of life science researchers who are able to take advantage of capabilities [27]. Ambitions of spring-boarding HPC use at these institutions can seem seductive to PIs ultimately seeking long-term secure resources. Meanwhile, pipelines for the rapidly expanding number of professionals trained to analyze such data are not in place. This and other discrepancies between the biological sciences, data-intensive computing, and attitudes and funding rates of Federal agencies responsible for stewarding this harmonization are discussed. Although a limited number of publications on the topic appeared prior to 2010, funding agencies face a seismic shift in the data requirements needed to do reproducible bioinformatics. But how to anticipate what needs are on the horizon when the existing suite of problems is still not recognized or understood across the academic research spectrum. While leveraging existing resources is encouraged, Federal funding agencies also need to contemplate how to support such analyses when the community lacks the resources or expertise to develop the types of programs necessary for taking advantage of existing computational resources.

**Private Sector Partnerships**

Scientists who were not trained in computation are learning it now. Some train themselves, whereas others are trained by students who have acquired skills either in courses or CUREs. Many of these biologists, primarily young ones, are learning computation in the form of analyzing massive biological data sets: genomics, proteomics, metabolomics, fluxomics, and population modeling, as well as the integration of these data. Of interest were the unmet needs regarding work in these areas, defined on the basis of the 704 National Science Foundation Biological Sciences principal investigators (BIO PIs) who appeared to be, and stated that they expected to be, most engaged in these kinds of research in the near future [27]. Among the needs are mundane basics: most biologists now store gigabyte data sets without the benefit of information technology support; their data are quickly outstripping their computational ability. But looming are problems in integrating data from different kinds of experiments done on different computational platforms from different labs intending to yield overlapping insight into complex biological processes from the subcellular to ecosystem levels. These modest statistical

 

and analytical tasks must be based on data that are exceedingly more complex than anything BIO PIs had handled in the past. Moreover, both the data and the results of analyses must be shared with other labs, many funded by different institutions, most using different data management platforms. A growing problem that will be shared by all stakeholders is how to integrate many diverse, even competing, data sets.

## CHALLENGES IN COMPUTATIONAL BIOLOGY

For decades, researchers have employed computer simulation to address problems in the biological sciences, although countless grand challenges in the field have yet to be convincingly solved. For example, a major open question in molecular biophysics is how proteins fold. The fundamental principles governing protein folding are now well established: a protein chain will spontaneously find a single 3-D structure, the native state, because this state has extremely low energy [28]. This energy landscape, however, is exponentially complex. Consequently, finding the native structure of a protein is formally an NP-hard problem. Thus, the principles behind how to solve such a simple-sounding computational biology problem are well understood–it is not feasible. One of the first attempted computational approaches to folding was lattice-gas models. Although simplified, these models demonstrated that simple polypeptide models captured some aspects of the folding process by forming compact, but incorrect structures, globules. At about the same time, a master equation was adapted for the protein folding problem as a random walk on the lattice to the native state. Since then, the field of computing and biology has exploded and become more diverse. However, many other efforts, including many analytical theories and isolated simulations, were doomed because the problem of protein folding is so highly complex.

For many of these biological grand challenges, there remains a large gap between understanding in a qualitative, often hand-waving sense, and proof via quantitative prediction. To rigorously demonstrate that one can learn enough from a gene sequence, for instance, to be able to build the organism from scratch, one must be able to in silico exactly simulate the real biological system with high accuracy across all relevant time scales. Much recent research has attempted to narrow this gap by designing novel algorithms to efficiently and accurately investigate these grand challenges in new ways. Contrary to this approach, in the spirit of certain foundational theories, so-called shockingly naive yet highly successful methods are instead employed to fundamentally speed such previously intractable computational problems.

### Data Overload

A revolution that caught many of us by surprise has crept quietly into the natural sciences and some engineering fields. Pegasus, which can carry the weight of a desktop computer on its back, is but one example. More generally, scientists and engineers have developed a tremendous data gathering and processing capability remarkable for its size, extent, reach, and variety. A magnetic resonance imagery scanner might serve as another example of the size of the revolution. This instrument produces tens of gigabytes for a single subject and illness scanned. If the scanner runs information technology systems that allows easy capture and persistent data treatment in a broadly accessible, searchable fashion it becomes a formidable knowledge discovery tool. As more and more data become available this tool provides increasingly rich insights for understanding patterns of disease. The Cancer Radiation Therapy Network captures roughly twelve million claims for about 500,000 Medicare patients. This data is used for investigating the comparative effectiveness of different cancer therapies. The data consist of measures of clinical encounters, medical procedures, practice setting, cost of care, providers' specialty, patient demographics, etc., and spans 1991 and 2002. The Health Insurance Portability and Accountability Act of 1996 makes this data possible.

The researchers who decided to capture this data have three plans. First, the bulk of the data will be retained by the Centers for Medicare & Medicaid Services, where it will be used for several investigations. For example, it will be used to study variation in treatment patterns between hospitals treating cancer. Second, the Medical Outsourcing Research Network will use part of the data in order

 

to test results with independent methods and people. Finally, the same group that collected the data will build another computer system with specific aims to facilitate data mining and posts results online [29].

**Reproducibility Issues**

The literature is replete with papers complaining about the lack of reproducibility in current computational research; while the experimentalists having traditionally considered advances found in the literature to be a "promise" in computing have found reason to be suspicious [30]. The attempts to reproduce and use results from computer-based models require "sets of low-level instructions simple enough that the execution machinery can properly follow them" and manually scrutinizing thousands of lines of code efficiently falls well outside this threshold. However, the detailed examination of the math (which is the high level description of these computations) is a common denominator in every published paper. Working in many areas, codes are almost never available, sometimes not even for the manufacturer. Yet there are computational models that are not implemented by the investigators, rather they are "experimented" computationally by submitting it to software which is pre-compiled or otherwise kept secret. In those instances where available content is publishable, code archives do exist, but to date academics tend to elect proprietary or bespoke software for code development versus the more robust commercial alternatives. The latter, of course, involve trade secrets if not third-party ownership. Domestically known software may attract patent trolling and to this day, a large fraction of the top-flight non-classified analysis shops refuse to work in a fully registered mode. This is largely from policy and social, rather than right could, grounds. Micro-CT is an adaptable synchrotron technique which is a very fast and non-invasive method for imaging the internal morphology of large samples. It can image a wide range of sample types such as teeth, fossils, wood, engineering samples, human limbs, animal muscles, and rocks. Despite this versatility, there is no general and effective protocol for the pre-experiment data processing and all works involve local code or method developments. This lack of standardization has been the main limiting factor in the wider adaptation of micro-CT radiation to new areas of research, measurably slowing the arrival of new scientific methods and inhibiting the speedy production of results in an era where rapid turn-around publication is imperative for impact. This is an example where it is not the algorithm, but the data preparation which is obstructive to reproduce and expertise lost to the void. More publications are likely forthcoming.

**RESULTS**

The application of computational tools in biological sciences has led to several notable advancements, including:

- *Genomic and Proteomic Analysis:* The development of algorithms and software, such as BLAST (Basic Local Alignment Search Tool) and tools for genome assembly, has significantly accelerated the analysis of large-scale genomic data. These tools allow for efficient sequencing, gene annotation, and identification of genetic variants, providing valuable insights into genetic disorders and evolutionary patterns.
- *Bioinformatics and Data Integration:* The use of databases like GenBank, Protein Data Bank (PDB), and KEGG (Kyoto Encyclopedia of Genes and Genomes) has enabled researchers to integrate biological data across various platforms, creating comprehensive datasets that inform biological research. Advanced statistical methods and machine learning algorithms have further enhanced the ability to predict protein interactions and cellular pathways.
- *Systems Biology and Predictive Modeling:* Computational modeling has played a crucial role in simulating biological systems, from metabolic networks to entire ecosystems. This has led to more accurate predictions of cellular behavior, the impact of environmental factors, and disease progression, ultimately guiding experimental design and therapeutic strategies.
- *Drug Discovery and Development:* Computational methods, such as molecular docking and molecular dynamics simulations, have revolutionized drug discovery by predicting how small molecules interact with biological targets. These techniques have accelerated the identification of potential drug candidates and optimized their effectiveness and safety profiles.
- *Ecological Modeling:* Computational tools have enabled the modeling of complex ecological systems, aiding in the understanding of biodiversity, climate change impacts, and conservation

 

strategies. Models that incorporate species interactions, environmental variables, and ecological data help predict ecosystem responses to various stressors.

- *Personalized Medicine:* The growing use of computational approaches in genomics and clinical data analysis has paved the way for personalized medicine, where treatments can be tailored to an individual's genetic makeup. This has shown promise in improving patient outcomes and reducing adverse drug reactions.

In conclusion, computational methods in biological sciences have vastly improved the speed, accuracy, and scope of biological research. The continued development of these tools will undoubtedly enhance our understanding of biology and foster innovation in healthcare, agriculture, and environmental management.

## CONCLUSION

Peering into this abyss of molecular biology? A useful starting point is usually beginning to write computer programs to analyze biological data. Though this processing has never been favorable for biology scholars, 'Biology is becoming more quantitative, and quantitative skills are a prerequisite for work in modern biology'. Many experimentalists agree and argue that Big Data challenges of biology 'require creative bioinformatics'. Unfortunately the importance of this has yet to reach most students, who are clustered in traditional life sciences. Schools of higher education should consequently course prompt 'the convergence of biological and computer sciences'. Indeed, not only there is a need for biologists who quickly understand complex software, but sophisticated statistical skills . Mainstream life sciences schools should accordingly begin to incorporate "programming skills and the basics of discrete mathematics and statistics". There is however wider courses existence in this area, largely in medical schools or as part of more general software initiatives. As modern biological research has become an advanced discipline from technological and methodological points of view, it is desired to introduce basic research and knowledge subject for undergraduates before they start the professional major courses in medical schools.

Computers and computer programming are going to become a necessary skill to do biology. There are the obvious applications of data processing such as "microarray clustering or digital PCR". With downturn of prices, Next-Gen sequencing is becoming more and more used. A high-throughput de novo sequencing of the genome and re-sequencing of evolved lines is a trivial task with the new technologies and might very well be expected to find bioinformatics tools to mine the data. Commercial programs exist for meeting most of these needs, but understandably they are not tailored to the hundreds particular listening of biology. Ergo with the exhausting use of rather universal algorithm, the biological problems end up being forced into a format that suits the software and much data and possibly knowledge get dispatched away with the wrappers in which they come.

## REFERENCES

1. Ekmekci B, McAnany CE, Mura C. An introduction to programming for bioscientists: A Python-based primer. 2016. Available from: https://www.ncbi.nlm.nih.gov
2. Akula B, Cusick J. Biological computing fundamentals and futures. 2009.
3. Nijveen H. Applications in computer-assisted biology. Wageningen: Wageningen University, 2013. 106 p. doi: 10.18174/282764.
4. Li J, Xiong Y, Feng S, Pan C, Guo X. CloudProteoAnalyzer: Scalable processing of big data from proteomics using cloud computing. 2024. Available from: https://www.ncbi.nlm.nih.gov.
5. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: A web server for metabolomic data analysis and interpretation. 2009. Available from: https://www.ncbi.nlm.nih.gov
6. Prasad M. Fast program for sequence alignment using partition function posterior probabilities. 2011.
7. Morrison D. Multiple sequence alignment is not a solved problem. 2018. [PDF].
8. Schlick T. Biomolecular structure and modeling: Problem and application perspective. 2010. Available from: https://www.ncbi.nlm.nih.gov.

 

9. Morris C. Towards a structural biology work bench. 2013. Available from: https://www.ncbi.nlm.nih.gov.
10. Allison DB, Brand PLJ, Edwards JW, Gadbury GL, Kim K, Mehta T, et al. A platform-independent software suite for statistical analysis of high dimensional biology data. 2005.
11. Hillerud MJ. BIΘSTAT: A computer program providing simple statistics for biological samples. 1973.
12. Mayer G. Data management in systems biology I – Overview and bibliography. 2009. [PDF].
13. Dreher F, Kreitler T, Hardt C, Kamburov A, Yildirimman R, Schellander K, et al. DIPSBC – Data integration platform for systems biology collaborations. 2012. Available from: https://www.ncbi.nlm.nih.gov.
14. Xu C, Jackson SA. Machine learning and complex biological data. 2019. Available from: https://www.ncbi.nlm.nih.gov.
15. Goldstein ND, Sarwate AD. Privacy, security, and the public health researcher in the era of electronic health record research. 2016. Available from: https://www.ncbi.nlm.nih.gov.
16. Vía M. Big data in genomics: Ethical challenges and risks. 2017.
17. Vazquez M, de la Torre V, Valencia A. Chapter 14: Cancer genome analysis. 2012. Available from: https://www.ncbi.nlm.nih.gov.
18. Ibal J, Park YJ, Park MK, Lee J, Kim MC, Shin JH. Review of the current state of freely accessible web tools for the analysis of 16S rRNA sequencing of the gut microbiome. 2022. Available from: https://www.ncbi.nlm.nih.gov.
19. Hupfauf S, Etemadi M, Fernández-Delgado Juárez M, Gómez-Brandón M, Insam H, Podmirseg SM. CoMA – An intuitive and user-friendly pipeline for amplicon-sequencing data analysis. 2020. Available from: https://www.ncbi.nlm.nih.gov.
20. Berger Leighton B, Daniels N, Yu WY. Computational biology in the 21st century. 2018.
21. Pillai N, Ram Das A, Ayoola M, Gireesan G, Nanduri B, Ramkumar M. EndToEndML: An open-source end-to-end pipeline for machine learning applications. 2024.
22. Place JF, Truchaud A, Ozawa K, Pardue H, Schnipelsky P. Use of artificial intelligence in analytical systems for the clinical laboratory. 1995. Available from: https://www.ncbi.nlm.nih.gov.
23. Chen P, Sun Z, Wang J, Liu X, Bai Y, Chen J, et al. Portable nanopore-sequencing technology: Trends in development and applications. 2023. Available from: https://www.ncbi.nlm.nih.gov.
24. Pareek CS, Smoczyński R, Tretyn A. Sequencing technologies and genome sequencing. 2011.
25. Navlakha S, Bar-Joseph Z. Algorithms in nature: The convergence of systems biology and computational thinking. 2011. Available from: https://www.ncbi.nlm.nih.gov.
26. Carpenter AE, Greene CS, Carnici P, Carvalho BS, de Hoon M, Finley S, et al. A field guide to cultivating computational biology. 2021.
27. Smolinski TG. Computer literacy for life sciences: Helping the digital-era biology undergraduates face today's research. 2010. Available from: https://www.ncbi.nlm.nih.gov.
28. Barone L, Williams J, Micklos D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. 2017.
29. Larson SD, Snow CD, Shirts M, Pande VS. Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. 2009.
30. White RR, Munch K. Handling large and complex data in a photovoltaic research institution using a custom laboratory information management system. 2014.

 