Review Article

An Overview of Bioinformatics and Computational Genomics in Modern Plant Science

Abstract

Bioinformatics plays monumental role in decoding the genomic landscape of plant species for agricultural advancements. While only a few plant genomes have been fully sequenced, the pan-genomic approach, involving the study of both the core and accessory genomes, is necessary to fully understand a species' genomic content. Bioinformatics, at the intersection of computer science and biology, handles the vast biological data generated during plant genomics research, facilitating data analysis, organization, and interpretation. By developing algorithms and tools, bioinformatics aids in gene prediction, functional annotation, sequence alignment, phylogenetic analysis, and more. It enables researchers to uncover genetic diversity, identify adaptive traits, and improve crops for disease resistance, stress tolerance, and nutritional value. With its computational models and data analysis techniques, bioinformatics accelerates discoveries and contributes to sustainable agriculture, guiding the second green revolution towards resilient food systems.

Key words: Bioinformatics, Genomics, Computational genomics, Pan-genome.

Introduction

Bioinformatics is an emerging scientific discipline that amalgamates computational power, mathematical algorithms, and statistical methods with life science concepts to address complex biological quandaries. Its scope encompasses diverse activities such as the collection and storage of biological data, data mining, database retrieval, analysis and interpretation, modelling, and product design (Degrave *et al*., 2002; Xiong, 2009; Jayaram and Dhingra, 2010). As an interdisciplinary science, bioinformatics leverages information technology and computer science to derive meaningful insights from biological data. The term "bioinformatics" was coined in 1970 by Paulien Hogeweg and Ben Hesper, who focused on elucidating informatics processes in biological systems. Hogeweg, a Dutch theoretical biologist and complex system researcher, investigates biological systems as dynamic entities. Recent technological advancements have precipitated a scientific revolution, yielding copious amounts of "omic" data.

However, the sheer volume and availability of this information in public databases pose significant challenges for professionals across various domains (Yates *et al*., 2015). Within the realm of biology, the principal challenge resides in comprehending the vast structural data and sequences generated at multiple levels of biological systems (Pevsner, 2015). In the field of bioinformatics, the development of requisite tools, encompassing statistical and computational methodologies, assumes paramount importance in elucidating the underlying mechanisms that underpin biological inquiries (Pevsner, 2015). It is imperative to acknowledge that this reductionist view oversimplifies the inherent complexity of science. The advent of a "new biology" epoch has witnessed the concomitant rise and progression of bioinformatics and computational biology, providing an integrated interface with molecular biology. These interdependent disciplines have significantly impacted the existing body of knowledge. Accordingly, this review aims to present a concise overview of bioinformatics and genomics, elucidating foundational principles that undergird bioinformatics. The focus encompasses the following aspects: i) types of biological information and databases, ii) sequence analysis and molecular modelling, iii) genomic analysis, and iv) systems biology. While acknowledging the broad scope of these areas, our objective is to underscore novel techniques and furnish analytical tools for data analysis and interpretation of results derived from these cutting-edge technologies.

Origin and theoretical vision of bioinformatics

Bioinformatics, a field that emerged prior to the feasibility of DNA sequencing, has a historical backdrop characterized by significant milestones. Key moments in its development include the seminal publication of the DNA structure by Watson and Crick in 1953, as well as the accumulation of data and knowledge in biochemistry and protein structure through the works of Pauling, Coren, and Ramachandran in the 1960s (Hagen, 2000). Regarded as the pioneer in organizing knowledge of protein three-dimensional (3D) structure, Margaret O. Dayhoff is often hailed as the matriarch of bioinformatics. Her contributions were instrumental in developing computers capable of determining peptide sequences, creating programs for structure recognition and display in X-ray crystallography, and devising computational methods for protein sequence comparison, enabling inferences about evolutionary connections across kingdoms (Hagen, 2000). Among her notable achievements, Dr. Dayhoff authored the influential "Atlas of Protein Sequence and Structure," a seminal work that played a pivotal role in systematizing and disseminating information. Beyond Dr. Dayhoff's contributions, numerous other researchers have also significantly contributed to the advancement of bioinformatics. Notably, these advancements owe much to the evolution of computing technology. Present-day achievements are primarily attributable to improvements in computational power and genome projects encompassing sequencing, annotation, data processing, and analysis. The development of large-scale capillary DNA sequencers and the introduction of fluorescence-labeled dideoxynucleotides in the 1990s facilitated the generation of vast quantities of data (Prosdocimi, 2010). However, the advent of next-generation sequencing technologies (NGS) has further accelerated the growth of complete genome sequences and

the overall volume of data. Bioinformatics has a rich history shaped by important scientific breakthroughs. Its origins predate DNA sequencing, with notable milestones including the elucidation of DNA structure and advancements in biochemistry and protein structure analysis. Margaret O. Dayhoff's contributions, especially in organizing protein 3D structure knowledge, have had a lasting impact. Moreover, the field's progress owes much to advancements in computing power and the continuous expansion of genome projects. The advent of NGS technologies has further propelled bioinformatics, leading to an exponential growth in complete genome sequences and the accompanying data volume.

Why is bioinformatics important?

The molecular biology community faces a significant challenge in effectively understanding the vast amount of data generated by genome sequencing projects. Traditionally, molecular biology research primarily took place in laboratory settings. However, the era of genomics has brought about a substantial increase in data volume, necessitating the integration of computers into the research process. This recognition has led to the development of new tools and databases in molecular biology, enabling research not only at the genomic level but also at the proteome, transcriptome, and metabolome levels.

Presently, the bioinformatics community grapples with the task of intelligently and efficiently storing the large quantities of generated data, as well as providing accessible and reliable means of accessing this data. It is crucial to develop concise computational tools capable of extracting meaningful biological information from the vast datasets. In the pharmaceutical industry, a growing trend involves leveraging bioinformatics tools to expedite the process and reduce costs associated with developing molecular markers and drug discovery.

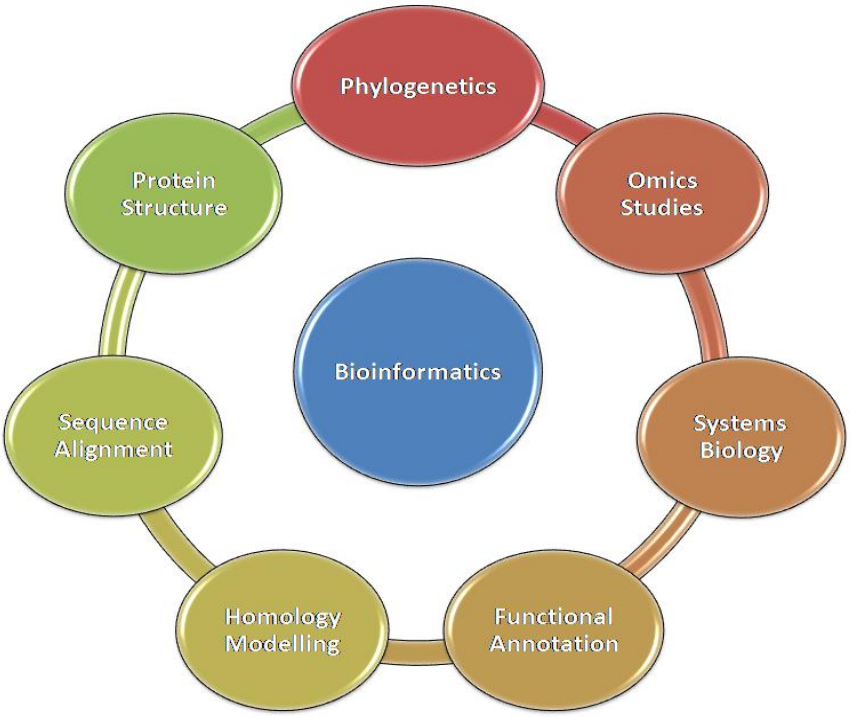


Figure 1. Importance of bioinformatics in different field of plant science

The central challenge for molecular biologists lies in effectively harnessing and interpreting the immense volume of data generated through genome sequencing projects. Incorporating bioinformatics and computational tools has become essential for handling these data and conducting research across multiple levels of biological systems. Developing efficient data storage methods and accessible tools for data analysis is critical to deriving meaningful insights from the wealth of biological information available. The application of bioinformatics tools also holds great potential in accelerating processes within the pharmaceutical industry, particularly in the development of molecular markers and drug discovery.

Organization of information: types of information and data bases

The exponential growth of data in the field of bioinformatics necessitates effective organization and storage strategies. As a response to this challenge, numerous biological databases have been established to store and process a vast amount of biological information, enabling access for the scientific community (Luscombe *et al*., 2001; Prosdocimi, 2010). The proliferation of data has been accompanied by the emergence of a substantial number of biological databases, with the Nucleic Acids Research journal taking the responsibility for compiling, updating, and disseminating these resources. According to the latest update in January 2017, there were 1739 biological databases available.

Bioinformatics relies on various sources of information, including raw DNA sequences, protein sequences, macromolecular structures, and genome sequencing data. Public databases play a crucial

role in storing and sharing large volumes of information, typically classified as primary or secondary databases. Primary databases primarily consist of experimental data without extensive analysis of previous publications. In contrast, secondary databases undergo a comprehensive content curation process involving compilation and interpretation of data (Prosdocimi, 2010). Additionally, functional databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) enable the analysis and interpretation of metabolic pathways.

Notably, the primary databases GenBank at the National Center for Biotechnology Information (NCBI), DNA Database of Japan (DDBJ), and European Molecular Biology Laboratory (EMBL) serve as prominent repositories for nucleotide sequences and proteins (Pevsner, 2015). These databases are integral members of the International Nucleotide Sequence Database Collaboration (INSDC), facilitating the daily exchange of deposited information among them. As for secondary databases, examples include the Protein Information Resource (PIR), UniProtKB/Swiss-Prot, Protein Data Bank (PDB), Structural Classification of Proteins2 (SCOP), and Prosite. These curetted databases focus specifically on protein-related information, providing details on protein structure, domains, function, and classification.

Table 1. Classification of databases in the 2004 edition of the Molecular Biology Database Collection.

Category	Number of Databases
Genomic	164
Protein sequences	87
Human/vertebrate genomes	77
Human genes and diseases	77
Structures	64
Nucleotide sequences	59
Microarray/gene expression	39
Metabolic and signalling pathways	33
RNA sequences	32
Proteomics	6
Other	6

(**Source:** D, Vassilev, J, Leunissen, A, Atanassov, A, Nenov, G. Dimov., 2005. Application of bioinformatics in plant breeding, biotechnology and biotechnology equipment, 19: sup3, 139152).

Analysis of biological sequences

The advent of next-generation sequencing (NGS) technologies has significantly increased the availability of biological data, consequently making sequence alignment a widely used and essential tool for comparison (Daugelaite *et al*., 2013). Sequence alignment involves the comparison of two or more nucleotide sequences (DNA or RNA) or amino acid sequences (peptides or proteins) to identify shared characters or patterns arranged within the sequences (Manohar and Shailendra, 2012).

The question arises: Why is sequence comparison important? This procedure serves various applications, including the assessment of evolutionary relationships between organisms, individuals, and genes, prediction of protein functions and structures, and more. Moreover, alignment techniques play a crucial role in whole genome analysis, enabling comparisons between different genomes or within the same species to identify sequence variations associated with specific phenotypes.

By comparing sequences, researchers can gain insights into the functional, evolutionary, and structural aspects of biological molecules. It allows for the inference of evolutionary relationships, unravelling the shared ancestry or divergence between different organisms or genes. Additionally, sequence comparison aids in predicting the functions of newly discovered genes or proteins based on similarities to known sequences with annotated functions. Furthermore, by aligning genomes from different individuals or species, variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, or structural rearrangements can be identified, providing valuable information about the genetic basis of phenotypic traits.

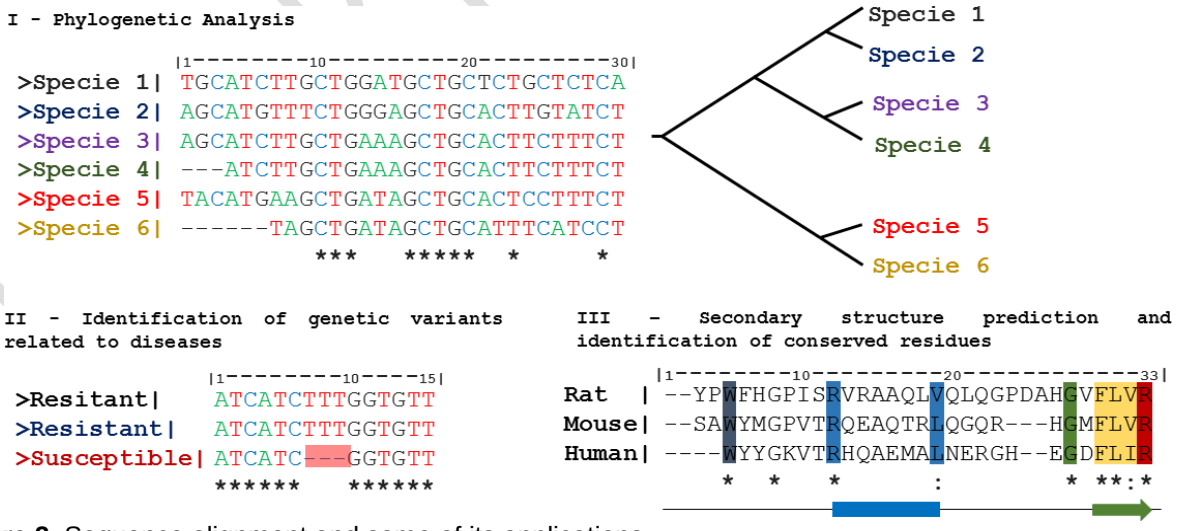
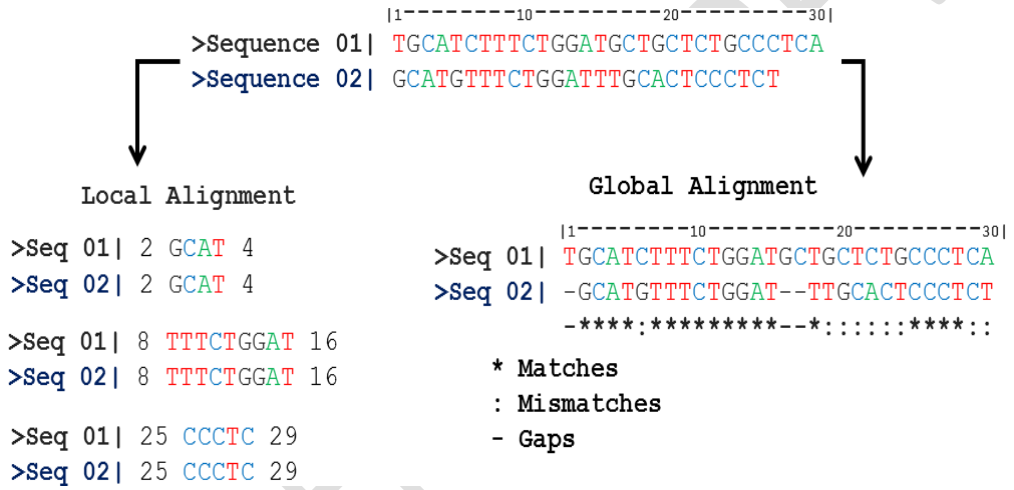


Figure 2. Sequence alignment and some of its applications.

Protein structure alignment is a prominent bioinformatics tool that plays a crucial role in analyzing the similarities and differences between protein structures. This process involves determining the equivalent amino acids between two or more protein structures. Comparing protein structures is essential for understanding their functional and evolutionary relationships. Alignment of protein structures can be categorized based on the number of sequences being compared, which includes simple and multiple alignments. Simple alignments focus on illustrating the similarity relationship between two sequences, whereas multiple alignments involve the comparison of three or more sequences. Furthermore, alignments can be further classified based on their scope, either as global or local alignments. Global alignments consider the entire length of the sequences, while local alignments specifically search for small regions of similarity.

Protein structure alignment provides valuable insights into the structural conservation and divergence among proteins. It aids in identifying common structural elements, conserved motifs, and functional domains across different proteins. Moreover, it facilitates the prediction of protein function and



aids in understanding the impact of structural variations on protein stability, dynamics, and interactions. The advancement of computational algorithms and tools for protein structure alignment has greatly enhanced the efficiency and accuracy of this analysis. Various methods, such as pair wise alignment algorithms (e.g., Needleman-Wunsch, Smith-Waterman) and multiple sequence alignment algorithms (e.g., ClustalW, MUSCLE), are widely employed for aligning protein structures.

Figure 3. Global and local alignment of amino acid sequences.

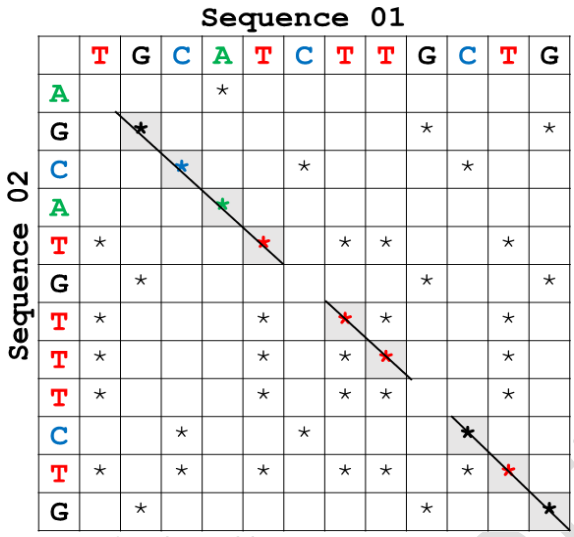


Figure 4. Simple alignment of amino acid sequences.

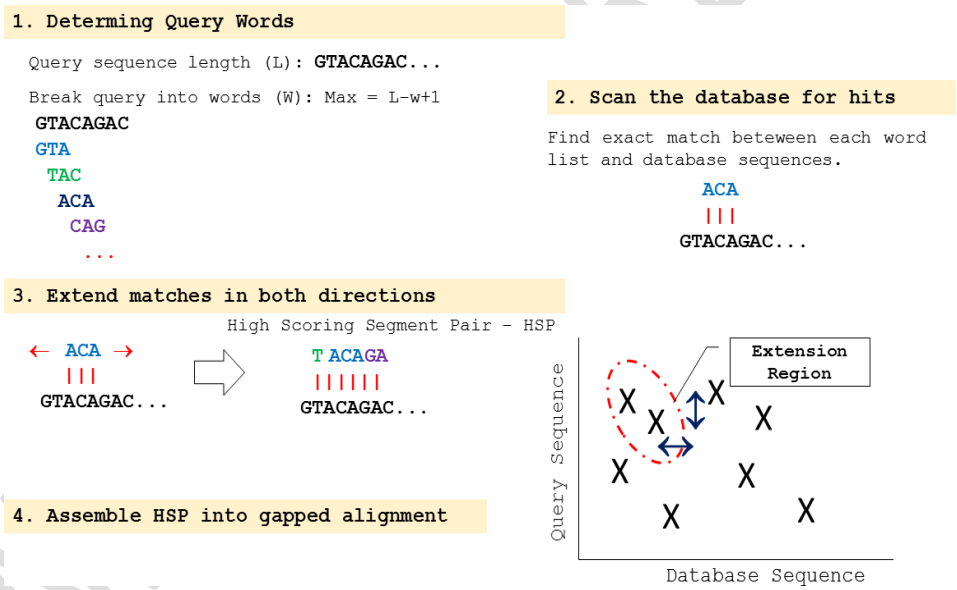


Figure 5. Process of BLAST operation (Multiple alignment). Figure modified from (http://steipe.biochemistry.utoronto.ca/abc/index.php/BLAST).

Comparative molecular modelling

Homology modelling, also known as comparative modelling, is a computational technique used to predict the three-dimensional structure of a protein based on the structure of a related protein whose structure has been experimentally determined (Capriles *et al*., 2014). This method relies on the observation that evolutionarily related protein sequences tend to adopt similar folding patterns and tertiary

structures (Calixto, 2013). Determining the three-dimensional structure of a protein is crucial for understanding its function, dynamics, and interactions, as well as for predicting its functional properties and identifying specific targets (Madhusudhan *et al*., 2005).

Although experimental techniques such as X-ray diffraction crystallography and nuclear magnetic resonance (NMR) can be used to determine protein structures, there are limitations to their application. Therefore, computational methods, including ab initio modelling and homology modelling, are often employed. Ab initio protein modelling uses physical and chemical principles to calculate the most energetically favourable conformation. On the other hand, homology modelling tends to yield more accurate results (Wang, 2010). However, the accuracy of homology modelling is influenced by the degree of similarity between the target protein and the template protein used for modelling (Capriles *et al*., 2014). While minimum identity values of 25 to 30% are generally considered acceptable, higher identity values tend to result in better quality predicted models (Calixto, 2013; Capriles *et al*., 2014).

The process of protein structure prediction through homology modelling typically involves five main steps (Figure 6): 1) identification of suitable reference proteins; 2) selection of appropriate templates with known structures; 3) alignment of the target sequence with the template sequence; 4) construction of a three-dimensional model based on the alignment; and 5) validation of the generated model.

Homology modelling plays a crucial role in bridging the gap between the increasing number of protein sequences available from genome sequencing projects and the limited availability of experimentally determined protein structures. It enables researchers to gain insights into the structure-function relationships of proteins, thereby facilitating drug discovery, protein engineering, and various other applications in biotechnology and molecular biology.

Overall, homology modelling provides a valuable approach for predicting the three-dimensional structure of proteins and offers insights into their functional properties. By leveraging the evolutionary conservation of protein structures, this computational method contributes to our understanding of protein biology and supports numerous scientific and applied research endeavours.

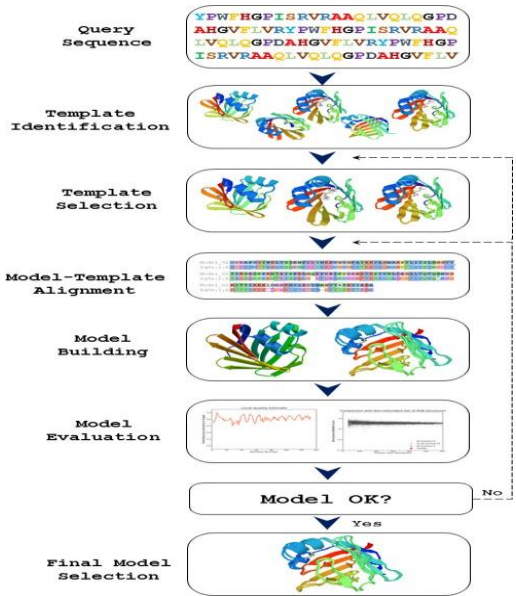


Figure 6. Prediction stage of 3D structures by comparative modelling.

Genome wide analyzes from genome to proteome

DNA sequencing has revolutionized molecular biology by providing unprecedented insights into the structure and function of genomes (Zhou *et al*., 2010). The advent of next-generation sequencing (NGS) technologies has enabled the analysis of DNA at an unprecedented scale and speed, paving the way for numerous applications. In the field of bioinformatics, these technological advancements have led to a focus on comprehensive genome-wide analyses, encompassing the genome, transcriptome, and proteome.

Genome-wide analyses involve the examination of entire genomes to uncover patterns, variations, and functional elements within the DNA sequence. These analyses provide valuable information about gene content, regulatory regions, non-coding RNAs, and genetic variations that contribute to phenotypic diversity and disease susceptibility. Bioinformatics tools and algorithms are employed to extract meaningful information from the vast amount of genomic data generated by NGS technologies.

In addition to genomics, transcriptomics plays a crucial role in understanding gene expression and regulation. Transcriptome analysis involves studying the complete set of RNA transcripts produced by an organism or a specific cell type under different conditions. It provides insights into the dynamic nature of gene expression, alternative splicing, post-transcriptional modifications, and non-coding RNA functions. Bioinformatics tools are utilized to analyze and interpret transcriptomic data, allowing researchers to unravel the complex regulatory networks governing gene expression.

Furthermore, proteomics, the study of the entire set of proteins expressed by an organism or a particular cell type, has gained significant attention in recent years. Proteome analysis provides insights into protein functions, post-translational modifications, protein-protein interactions, and cellular pathways. Bioinformatics approaches are employed to process, analyze, and integrate proteomic data, facilitating the identification and characterization of proteins, as well as the exploration of their biological roles.

Integrating genome-wide analyses across the genome, transcriptome, and proteome levels offers a comprehensive understanding of the molecular mechanisms underlying biological processes. These analyses provide insights into the interplay between genetic information, gene expression, and protein functions, enabling the discovery of novel biomarkers, therapeutic targets, and diagnostic tools.

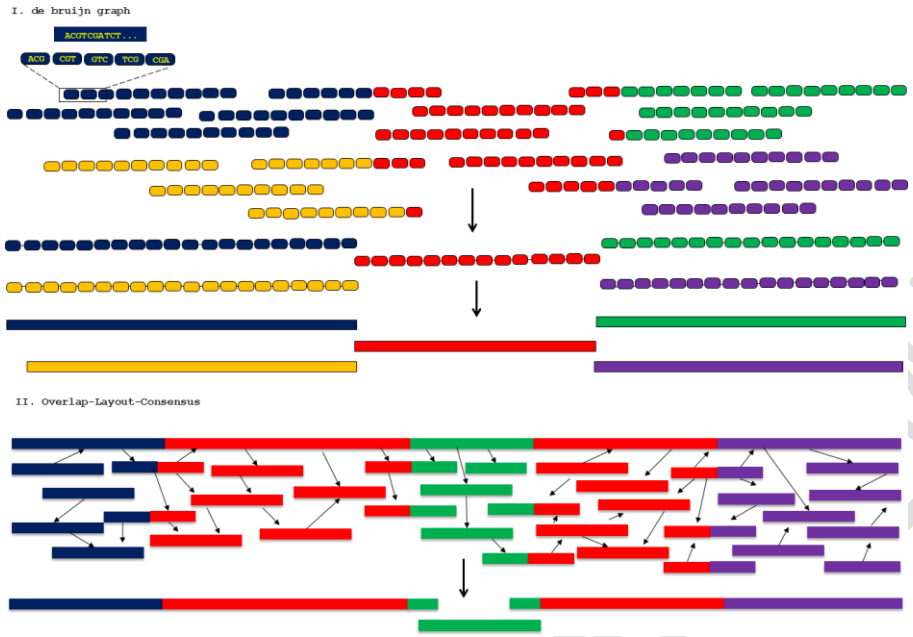


Figure 7. Strategies for assembling genomes.

Table 2. List of some published plant genome:

Species Name	Size (~Mb)*	No. of Chromosome
Arabidopsis thaliana (mouse ear cress)	115	5
Bracypodium distachyon	355	5
Brassica rapa (Chinese cabbage)	284	10
Cajanus cajan (pigeonpea)	883	11
Carica papaya (papaya)	372	9
Cucumis sativus (cucumber)	203	-
Fragaria vesca (woodland strawberry)	240	7
Glycine max (soybean)	975	20
Medicago truncatula (barrel me dic)	241	8
Malus domestica (apple)	881.3	-
Oryza sativa (rice, japonica)	372	12
Panicum virgatum (switchgrass)	1,230	-
Populus trichocarpa (poplar)	422.9	19
Ricinus communis (castor bean)	400	-
Pinus taeda (loblolly pine)	22,180	127
Solanum tuberosum (potato)	800	12

Sorghum bicolor (sorghum)	730	10
Theobroma cacao (cacao)	346	-
Vitis vinifera (grapevine)	487	19

Table 3. Integrative databases in plants:

Database name	Species	URL
TAIR	Arabidopsis	http://www.arabidopsis.org/
SIGnAL	Arabidopsis	http://signal.salk.edu/
RARGE	Arabidopsis	http://rarge.psc.riken.jp/
Rice Genome Annotation Project	Rice	http://rice.plantbiology.msu.edu/
RAP-DB	Rice	http://rapdb.dna.affrc.go.jp/
SOL genomics network	Solanaceae	http://solgenomics.net/
Gramene	Gramineae	http://www.gramene.org/
GrainGenes	Triticeae and Avena	http://wheat.pw.usda.gov/GG2/index.shtml
SoyBase	Soybean	http://www.soybase.org/
MazieGDB	Maize	http://www.maizegdb.org/
CyanoBase	Cyanobacteria	http://genome.kazusa.or.jp/cyanobase/
GDR (Genome Database for Rosaceae)	Rosaceae	http://www.bioinfo.wsu.edu/gdr/
Brassica Genome Gateway	Brassica	http://brassica.bbsrc.ac.uk/
Cucurbit Genomics Database	Cucurbitaceae	http://www.icugi.org/
Phytozome	Plant species (whole genome data available)	http://www.phytozome.net/
PlantGDB	Plant species (whole genome and/or large-scale EST data available)	http://www.plantgdb.org/
EnsemblPlants	Plant species (whole genome data available)	http://plants.ensembl.org/index.html
ChloroplastDB	Plant species (Chloroplast genome data available)	http://chloroplast.cbio.psu.edu/
KEGG PLANT	Plant species (whole genome	http://www.genome.jp/kegg/plant/

	and/or large-scale EST data available)	

Transcriptomics

DNA sequencing and hybridization technologies have revolutionized the study of transcriptomes, allowing researchers to infer and quantify gene expression patterns (Wang *et al*., 2010). Although approaches such as real-time PCR (qPCR) and DNA microarrays have contributed significantly to transcriptomic research, they have certain limitations (Marioni *et al*., 2008; Wang *et al*., 2010). In contrast, Next-Generation Sequencing (NGS) platforms have emerged as a powerful alternative for global expression profiling (Montgomery *et al*., 2010).

One of the widely used methods in transcriptomics is RNA sequencing (RNA-seq), which enables the mapping of reads and precise quantification of transcript levels in a high-throughput manner, offering improved accuracy and cost-effectiveness compared to other technologies (Wang *et al*., 2010). To generate an RNA-seq dataset, RNA from the target conditions is first extracted, purified, and fragmented. Reverse transcriptase is then used to convert the RNA fragments into complementary DNA (cDNA). Adapters are ligated to the cDNA fragments, and the desired size range is selected. Finally, the cDNAs are sequenced using NGS technologies (Figure 8).

RNA-seq has revolutionized transcriptomics by providing comprehensive and quantitative insights into gene expression, alternative splicing, fusion transcripts, and non-coding RNA transcripts. It has enabled the identification of novel transcripts, detection of low-abundance transcripts, and characterization of transcript is forms. Furthermore, RNA-seq data analysis involves aligning the sequenced reads to a reference genome or transcriptome, estimating transcript abundance, and performing differential expression analysis.

The application of RNA-seq has greatly expanded our understanding of gene regulation, biological processes, and disease mechanisms. It has found applications in various fields, including biomedical research, agriculture, and environmental sciences. The continuous advancements in NGS technologies, data analysis algorithms, and computational resources are further enhancing the capabilities of RNA-seq and driving breakthrough discoveries in transcriptomics.

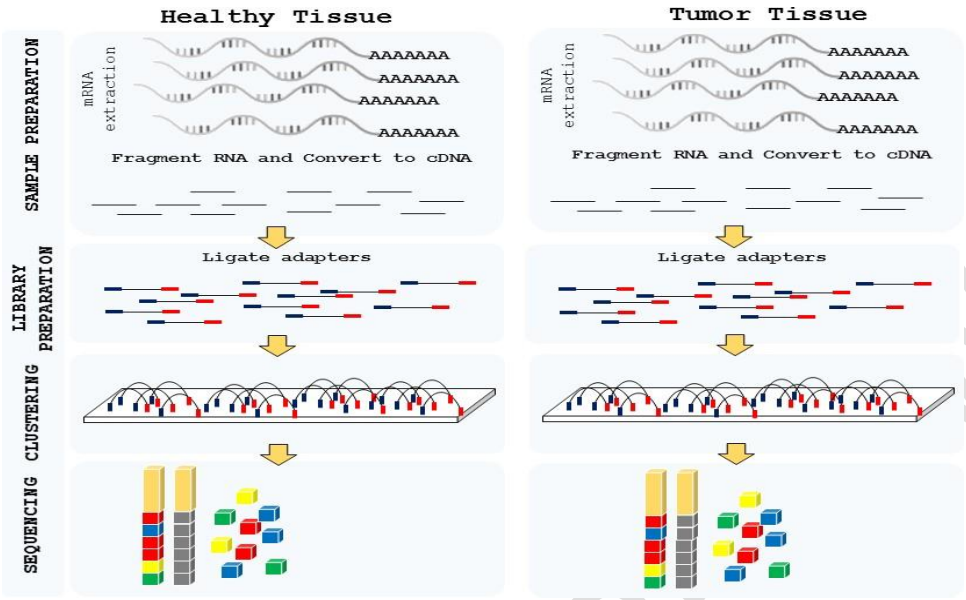


Figure 8. Steps to data generation for RNA-seq.

Table 4. Transcription factor database in plants:

Database	URL	Species	References
RARTF	http://rarge.gsc.riken.jp/rartf/	Arabidopsis	Iida *et al*., (2005)
AGRIS, AtTFDB	http://arabidopsis.med.ohio-state.edu/AtTFDB/	Arabidopsis	Palaniswamy *et al*., (2006)
DATF	http://datf.cbi.pku.edu.cn/	Arabidopsis	Gao *et al*., (2006)
DRTF	http://drtf.cbi.pku.edu.cn/	Rice	Gao *et al*., (2006)
DPTF	http://dptf.cbi.pku.edu.cn/	Poplar	Gao *et al*., (2006)
TOBFAC	http://compsysbio.achs.virginia.edu/tobfac/	Tobacco	Rushton *et al*., (2008)
SoybeanTFDB	http://soybeantfdb.psc.riken.jp/	Soybean	Mochida *et al*., (2009c)
PlnTFDB	http://plntfdb.bio.uni-potsdam.de/v3.0/	20 plant	Riano-Pachon

		species	*et al*., (2007)
GRASSIUS, GrassTFDB	http://grassius.org/grasstfdb.html	Maize, rice, sorghum, sugarcane	Yilmaz *et al*., (2009)

Proteomics

Proteomics plays a crucial role in unravelling the intricate molecular processes underlying cellular physiology by identifying, quantifying, and characterizing all the proteins within a cell (Schmidt *et al*., 2014). It has rapidly evolved as a field aiming to systematize the study of protein structure, function, interactions, and dynamics in both spatial and temporal dimensions. Figure 9 illustrates some of the diverse applications of proteomics.

Protein identification can be achieved through three main approaches: i) direct protein sequencing, ii) electrophoresis gel-based methods, and iii) mass spectrometry (MS). Mass spectrometry has revolutionized proteomics by enabling the sensitive identification of proteins in complex mixtures, facilitating expression quantification and characterization of post-translational modifications (Pevsner, 2015). It involves several key components, including i) an ionization technique such as electro spray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI), ii) one or more types of mass analyzers, such as time-of-flight (TOF) or ion trap analyzers, and iii) a detector.

The application of mass spectrometry-based proteomics has led to significant advancements in our understanding of protein function, cellular signalling pathways, and disease mechanisms. It allows for the identification of thousands of proteins simultaneously and provides insights into their quantitative changes under different biological conditions. Moreover, mass spectrometry facilitates the characterization of protein post-translational modifications, such as phosphorylation, glycosylation, and acetylation, which play crucial roles in protein regulation and cellular processes.

Proteomics has widespread applications in various fields, including biomedical research, drug discovery, and biomarker identification. It aids in the study of protein-protein interactions, protein-ligand binding, and protein dynamics within cellular compartments. The integration of proteomics data with other omics technologies, such as genomics and transcriptomics, enables a comprehensive understanding of biological systems and their complexity. As mass spectrometry techniques continue to advance in terms of sensitivity, resolution, and throughput, proteomics is poised to make even greater contributions to our understanding of cellular processes, disease mechanisms, and the development of personalized medicine. By unravelling the intricate web of protein interactions and dynamics, proteomics holds the key to deciphering the complexities of living systems at the molecular level.

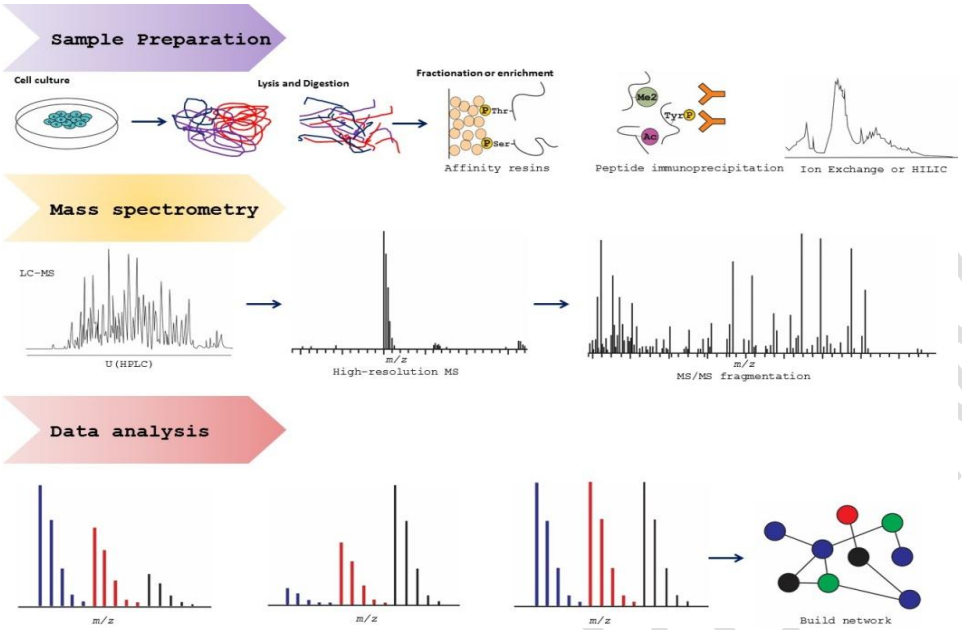


Figure 9. Generalized approach of proteomics based on mass spectrometry.

Importance of supercomputer in bioinformatics

Supercomputers play a crucial role in bioinformatics by enabling complex computational analyses of biological data. Bioinformatics involves the application of computational methods and algorithms to analyze, interpret, and model biological data, such as genomic sequences, protein structures, and molecular interactions. Here are some supercomputers in india that has been used for bioinformatical studies:

- Param siddhi located at IIT Kanpur, it has been used for various bioinformatical research and computational biology applications.
- SahasraT located at the Centre for Development of Advanced Computing (C-DAC) in Pune, India. It is designed to cater to bioinformatics and life sciences research and has been used for genomics, proteomics, and molecular modeling studies.
- Bhaskara is a supercomputer located at the Centre for Development of Advanced Computing (C-DAC) in Bengaluru, India.
- Annapurna is a supercomputer located at the Centre for Development of Advanced Computing (C-DAC) in Pune, India.
- Pratyush is a series of supercomputers located at the Indian Institute of Tropical Meteorology (IITM) in Pune, India. While primarily focused on weather and climate research.

- PARAM Shivay, the first supercomputer assembled indigenously, was installed in IIT (BHU), followed by PARAM Shakti, PARAM Brahma, PARAM Yukti, PARAM Sanganak at IIT-Kharagpur IISER, Pune, JNCASR, Bengaluru and IIT Kanpur, IIT Hyderabad, NABI Mohali, CDAC Bengaluru respectively.

Applications of bioinformatics in plant science

Bioinformatics plays a crucial role in agricultural research by facilitating the analysis of various types of large agricultural data. Its application in plant genetic resources management and data analysis enables the development of improved crops with enhanced resistance to drought, diseases, and insects, as well as improved livestock quality and health.

1. *Crops*

In the field of crop research, comparative genomics allows for the study of gene interrelationships between model and non-model plants, enabling the transfer of knowledge from model plant systems to other crops. For instance, complete genomes such as Arabidopsis and Oryza sativa (rice) have provided valuable insights (Proost *et al*., 2009).

2. *Renewable Energy*

Renewable energy production from plant-based biomass, such as ethanol, has gained significance. Biomass-based crop species and lignocellulosic species have become important resources for bio fuel production. Genomics and bioinformatics aid in the detection of sequence variants in biomass-based plants, leading to optimized biomass production and tolerance. The decoding of the Eucalyptus grandis genome has provided valuable information on the mechanisms involved in sugar conversion, contributing to advancements in biomass component production (Boyle *et al*., 2004; Betz *et al*., 2000).

3. *Insect resistance*

Insect resistance can be achieved through the incorporation of Bacillus thuringiensis (Bt) genes into crops, leading to reduced pesticide usage. This genetic modification enhances plants' ability to resist insect invasion, benefiting agricultural practices.

4. *Improve Nutritional Quality*

Improving the nutritional quality of crops is another area where bioinformatics has made an impact. By inserting genes into crops, scientists have successfully increased levels of essential micronutrients such as Vitamin A and iron, addressing deficiencies and reducing associated health issues (Nierman *et al*., 2005; Fraser *et al*., 2009).

5. *Grow in Poorer Soils and Drought Resistant*

Bioinformatics aids in the development of crop varieties that can thrive in poor soil conditions, withstand drought, and tolerate reduced water quality. By analyzing and interpreting the vast amount of data generated from research, bioinformatics helps address these challenges.

6. *Plant Breeding*

Plant genomics contributes to the understanding of genetic and molecular processes in plants, providing valuable insights for breeding new varieties with improved quality and reduced environmental impact. The study of gene expression enables a better understanding of how plants respond to internal and external stimuli, guiding future breeding decisions.

7. *Agriculturally Important Microorganism*

In the field of plant disease management, bioinformatics plays a critical role in understanding the genetic structure of microorganisms and pathogens. By investigating host-pathogen interactions and identifying disease-causing microorganisms, bioinformatics assists in the development of disease-resistant plants and the implementation of effective management strategies (Berg, 2009; Schenck, 2012).

8. *Accelerate Crop Improvement in a Changing Climate*

With climate change and population growth posing challenges to food production, bioinformatics combined with genomics offers the potential to accelerate the breeding of climate-resilient crops. By linking genomic data with climate-related agricultural traits, bioinformatics can contribute to the development of climate-friendly crops, ensuring food security in the face of changing environmental conditions (Batley and Edwards, 2016).

Future perspectives

With the proliferation of sequencing projects, bioinformatics continues to make significant advances in biology by providing scientists with access to genomic information. Cloud-based services on the Internet give scientists free access to such vast amounts of biological information, enabling advances in scientific discoveries in agriculture. In the coming decades, computer models of system-wide characteristics are expected to make another major leap in the field of bioinformatics, where they can serve as the basis for experimentation and discovery. This does more than just understand exactly how plants determine specific traits, discover disease causality, and predict their response to changes in the environment. This can lead to disease prevention and targeted treatment, improved food production and environmental protection.

Conclusion

Bioinformatics and computational genomics have revolutionized modern plant breeding, offering powerful tools and methodologies to address the challenges faced by breeders. The integration of genetic and genomic data with advanced computational techniques has greatly accelerated the breeding process,

allowing for the development of improved crop varieties with enhanced traits. From data management and analysis to genetic mapping, marker-assisted breeding, GWAS, genomic selection, and crop modelling, bioinformatics has become an integral part of plant breeding programs worldwide. The continued advancements in bioinformatics and computational genomics will further enhance our understanding of plant genomes, facilitating the development of climate-resilient crops, sustainable agricultural practices, and ensuring global food security in the face of environmental challenges.

References

Berg, G. 2009. Plant–microbe interactions promoting plant growth & health: perspectives for controlled use of microorganisms in agriculture. Applied microbiology & biotechnology. 84: 11-18.

Betz, R. C., Lee, Y. A., Bygum, A., F., Bernal, A. I., Toribio, J. & Nothen, M. M. 2000. A gene for hypotrichosis simplex of the scalp maps to chromosome 6p21. 3. The American Journal of Human Genetics. 66(6): 1979-1983.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. & Sherlock, G. 2004. GO:: TermFinder—open source software for accessing Gene Ontology information & finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 20(18): 3710-3715.

Calixto, P. H., Bitar, M., Ferreira, K. A., Abrahao, O., Lages-Silva, E., Franco, G. R. & Pedrosa, A. L. 2013. Gene identification & comparative molecular modeling of a Trypanosoma rangeli major surface protease. *Journal of molecular modeling*. 19: 3053-3064.

Capriles, V. D. & Areas, J. A. G. 2014. Novel approaches in gluten- free breadmaking: interface between food science, nutrition, & health. *Comprehensive Reviews in Food Science & Food Safety*. 13(5): 871-890.

D, Vassilev, J, Leunissen, A, Atanassov, A, Nenov, G. Dimov. 2005. Application of bioinformatics in plant breeding, biotechnology & biotechnology equipment. 19: 139-152

Daugelaite, J., O'Driscoll, A. & Sleator, R. D. 2013. An overview of multiple sequence alignments & cloud computing in bioinformatics. *International Scholarly Research Notices*.

Degrave, W. M., Vargas, R., Alvarez, F., Collado-Vides, J., Nunez, L. & Luis, J. 2002. REVIEW COPY. *Applied Bioinformatics*, 1(1): 53-56.

Fraser, D. & Kaern, M. 2009. A chance at survival: gene expression noise & phenotypic diversification strategies. Molecular microbiology. 71(6): 1333-1340.

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W. & Luo, J. 2006. DRTF: a database of rice transcription factors. *Bioinformatics*. 22(10):1286-1287.

Hagen, J. B. 2000. The origins of bioinformatics. *Nature Reviews Genetics*. 1(3):231-236.

Hogeweg, P. & Hesper, B. 1978. Interactive instruction on population interactions. *Computers in biology & medicine*. 8(4): 319-327.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T. & Shinozaki, K. 2005. RARTF: database & tools for complete sets of Arabidopsis transcription factors. *DNA Research*. 12(4):247-256.

Jayaram, B., Dhingra, P., Mishra, A., Kaushik, R., Mukherjee, G., Singh, A. & Shekhar, S. 2014. Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC bioinformatics*. 15(16):1-12.

Luscombe, N. M., Greenbaum, D. & Gerstein, M. 2001. What is bioinformatics? A proposed definition & overview of the field. *Methods of information in medicine*. 40(04):346-358.

Madhusudhan, M. S., Marti-Renom, M. A., Eswar, N., John, B., Pieper, U., Karchin, R. & Sali, A. 2005. Comparative protein structure modeling. *The proteomics protocols h&book*. 831-860.

Manohar, P. & Shailendra, S. 2012. Protein sequence alignment: A review. *World Applied Program*. 2:141-145.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. 2008. RNA-seq: an assessment of technical reproducibility & comparison with gene expression arrays. *Genome research*. 18(9):1509-1517.

Mochida, K. & Shinozaki, K. 2010. Genomics & bioinformatics resources for crop improvement. *Plant & Cell Physiology*. 51(4): 497-523.

Montgomery, D. C. & Runger, G. C. 2010. *Applied statistics & probability for engineers*. John wiley & sons.

Nierman, W. C., Pain, A., &erson, M. J., Wortman, J. R., Kim, H. S., Arroyo, J., & Denning, D. W. 2005. Genomic sequence of the pathogenic & allergenic filamentous fungus Aspergillus fumigatus. Nature. 438(7071): 1151-1156.

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V. & Grotewold, E. 2006. AGRIS & AtRegNet. a platform to link cis-regulatory elements & transcription factors into regulatory networks. *Plant physiology*. 140(3):818-829.

Pevsner, J. 2015. *Bioinformatics & functional genomics*. John Wiley & Sons.

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y., & V&epoele, K. 2009. PLAZA: a comparative genomics resource to study gene & genome evolution in plants. The Plant Cell. 21(12): 3718-3731.

Prosdocimi, f., bittencourt, d. D. C., silva, f., motta, p. & rech filho, e. L. 2010. Molecular, behavioral & anatomical sophistication in spider webs: insights from spinning gl& RNA-seq experiments in primitive & modern spiders. In: International conference of the brazilian association for bioinformatics & computational biology. 67.

Riano-Pachon, D. M., Ruzicic, S., Dreyer, I. & Mueller-Roeber, B. 2007. PlnTFDB: an integrative plant transcription factor database. *BMC bioinformatics*. 8(1):1-10.

Rushton, P. J., Bokowiec, M. T., Han, S., Zhang, H., Brannock, J. F., Chen, X. & Timko, M. P. 2008. Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant physiology*. 147(1):280-295.

Schmidt, M., Pedersen, L. & Sorensen, H. T. 2014. The Danish Civil Registration System as a tool in epidemiology. *European journal of epidemiology*. 29:541-549.

Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*. 26(1): 136-138.

Watson, J. D. & Crick, F. H. 1953, January. The structure of DNA. In *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press. 18:123-131.

Xiong, F., Shu, L., Zeng, H., Gan, X., He, S. & Peng, Z. 2022. Methodology for fish biodiversity monitoring with environmental DNA metabarcoding: the primers, databases & bioinformatic pipelines. *Water Biology & Security*. 1(1): 100007.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. & Flicek, P. 2015. The Ensembl REST API: Ensembl data for any language. *Bioinformatics*. 31(1): 143-145.

Yilmaz, A., Nishiyama Jr, M. Y., Fuentes, B. G., Souza, G. M., Janies, D., Gray, J. & Grotewold, E. 2009. GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant physiology. 149(1): 171-180.

Zhou, J. J., Vieira, F. G., He, X. L., Smadja, C., Liu, R., Rozas, J. & Field, L. M. 2010. Genome annotation & comparative analyses of the odorant- binding proteins & chemosensory proteins in the pea aphid Acyrthosiphon pisum. Insect molecular biology. 19: 113-122.