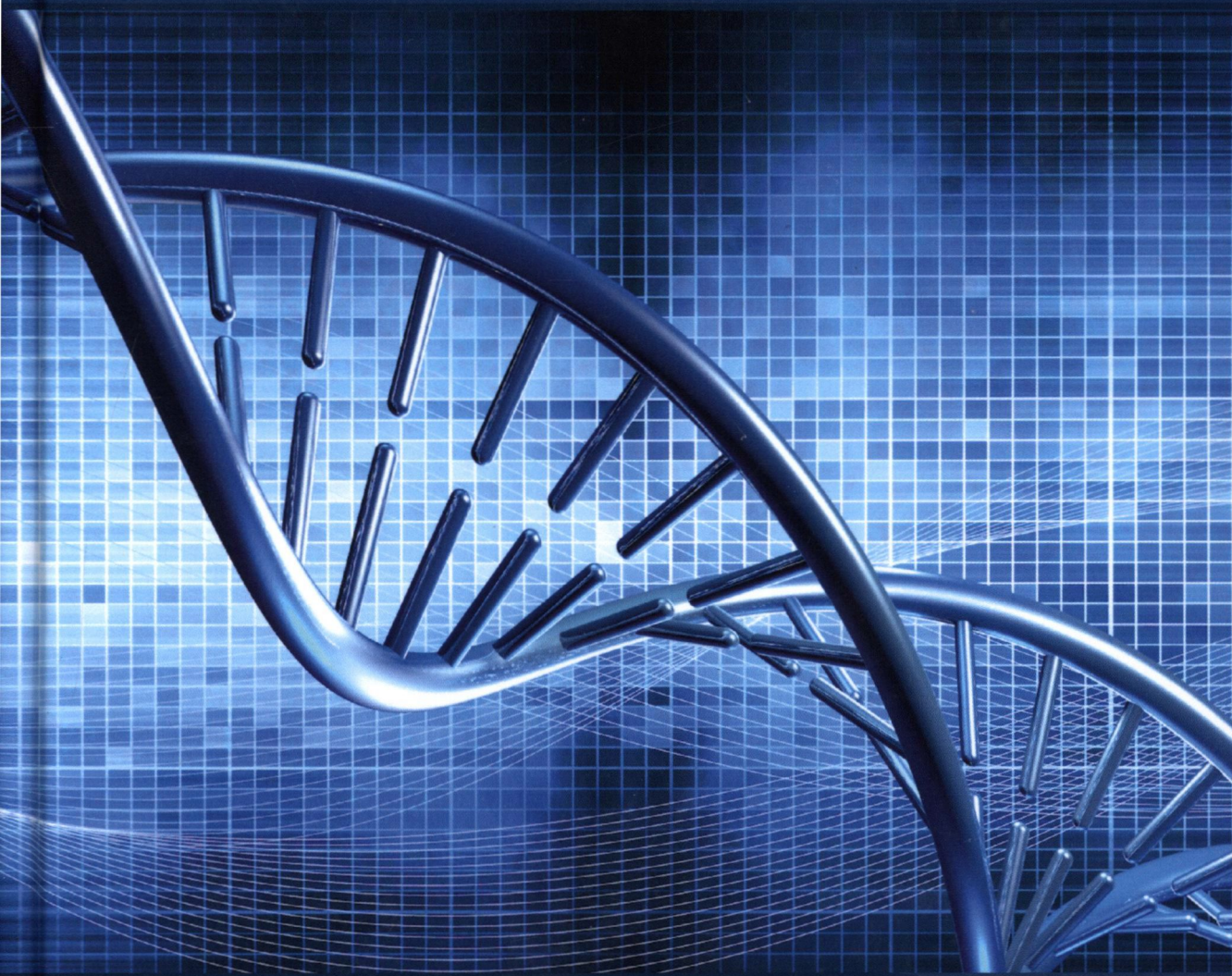



# An Introduction to **Bioinformatics**



**Regan Knight**

 Larsen & Keller



# An Introduction to Bioinformatics

Bioinformatics is an amalgamation of mathematics, engineering, computer sciences and statistics. It refers to the practice of using software tools to understand biological data. This book explores all the important aspects of bioinformatics in the present day scenario. It elaborates the different branches related to the subject and their applications. While understanding the long-term perspectives of the topics, the book makes an effort in highlighting their impact as a modern tool for the growth of bioinformatics. This text, with its detailed analyses and data, will prove immensely beneficial to students involved in this area at various levels. It will be of great help to those in the fields of genetics, forensic science and evolutionary biology.

**Regan Knight** holds a Bachelor's degree in Bioengineering from The University of Illinois, Chicago and a Master's degree in Biotechnology and Bioinformatics from the La Trobe University, Melbourne, Australia. His interest areas of academic research include DNA Sequencing and comparative genomics. Knight has presented over 25 papers at international symposiums and conferences and has numerous papers, chapters and book articles to his credit. His works have been published in various books as reference materials for students.

 **Larsen & Keller**  
www.larsen-keller.com

ISBN 978-1-63549-045-9



DBS-BM



09L1473327

# Knights and Dragons: A History of the English Language

by Larsen & Keller



# An Introduction to Bioinformatics

Edited by  
Regan Knight

An Introduction to Bioinformatics  
Edited by Regan Knight  
ISBN: 978-1-63549-045-9 (Hardback)

© 2017 Larsen & Keller



Published by Larsen and Keller Education,  
5 Penn Plaza,  
19th Floor,  
New York, NY 10001, USA

**Cataloging-in-Publication Data**

An introduction to bioinformatics / edited by Regan Knight.  
p. cm.  
Includes bibliographical references and index.  
ISBN 978-1-63549-045-9  
1. Bioinformatics. 2. Biology--Data processing. 3. Computational biology.  
I. Knight, Regan.  
QH324.2 .I58 2017  
570.28--dc23

This book contains information obtained from authentic and highly regarded sources. All chapters are published with permission under the Creative Commons Attribution Share Alike License or equivalent. A wide variety of references are listed. Permissions and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the vailidity of all materials or the consequences of their use.

Trademark Notice: All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

The publisher’s policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Printed and bound in China.

For more information regarding Larsen and Keller Education and its products, please visit the publisher’s website [www.larsen-keller.com](http://www.larsen-keller.com)

# An Introduction to Bioinformatics



# Preface

Bioinformatics is an amalgamation of mathematics, engineering, computer sciences and statistics. It refers to the practice of using software tools to understand biological data. This book explores all the important aspects of bioinformatics in the present day scenario. It elaborates the different branches related to the subject and their applications. While understanding the long-term perspectives of the topics, the book makes an effort in highlighting their impact as a modern tool for the growth of bioinformatics. This text, with its detailed analyses and data, will prove immensely beneficial to students involved in this area at various levels. It will be of great help to those in the fields of genetics, forensic science and evolutionary biology.

To facilitate a deeper understanding of the contents of this book a short introduction of every chapter is written below:

Chapter 1- Bioinformatics is a field of study that helps in the development of methods and software tools for developing a better understanding of biological data. It combines subjects such as computer science, statistics, mathematics and engineering. The chapter on bioinformatics offers an insightful focus, keeping in mind the complex subject matter.

Chapter 2- Sequence analysis is the process that clarifies the sequence of DNA, RNA or peptide sequence. The methods used in this are sequence alignment, searches against biological databases etc. This section also focuses on the techniques used in DNA sequencing such as Sanger sequencing, Illumina dye sequencing and ion semiconductor sequencing. The section elucidates the crucial theories and principles of sequence analysis.

Chapter 3- The study of the genome is known as genomics. It is a discipline of genetics and applies methods such as recombinant DNA and DNA sequencing. Personal genomics, oncogenomics, comparative genomics and the genome project have been explained in this section. This chapter is an overview of the subject matter incorporating all the major aspects of genome analysis.

Chapter 4- Computational biology involves the application of the theoretical methods and data analytical methods to the study of biology and social systems. This involves subjects such as computer science, statistics, chemistry, molecular biology, ecology and visualization. This chapter will provide an integrated understanding of computational biology.



Chapter 5- The software used in bioinformatics range from simple tools to complex graphical programs. The types of bioinformatics softwares elucidated in the following section are Biopython, bioconductor, BioPerl, BioJava, BioRuby and EMBOSS. Bioinformatics software is best understood in confluence with the major topics listed in the following chapter.

Chapter 6- Bioinformatics has diverse aspects; some of these are structural bioinformatics, modelling biological systems, protein-protein interaction prediction, interactome, flow cytometry bioinformatics etc. The topics discussed in the section are of great importance to broaden the existing knowledge on bioinformatics.

I owe the completion of this book to the never-ending support of my family, who supported me throughout the project.

**Editor**

# Table of Contents

<b>Preface</b>	<b>VII</b>
<b>Chapter 1 Introduction to Bioinformatics</b>	<b>1</b>
<b>Chapter 2 Sequence Analysis: A Comprehensive Study</b>	<b>16</b>
a. Sequence Analysis	16
b. Sequence Alignment	17
c. Multiple Sequence Alignment	26
d. Sequence Assembly	41
e. DNA Sequencing	48
<b>Chapter 3 Genome Analysis: An Overview</b>	<b>88</b>
a. Genomics	88
b. Personal Genomics	100
c. Oncogenomics	105
d. Comparative Genomics	114
e. Genome Project	119
f. Genome-wide Association Study	123
<b>Chapter 4 Computational Biology: An Integrated Study</b>	<b>131</b>
a. Computational Biology	131
b. Gene Prediction	136
c. Modelling Biological Systems	142
d. Computational Genomics	146
e. Computational and Statistical Genetics	148
<b>Chapter 5 Types of Bioinformatics Software</b>	<b>165</b>
a. Biopython	165
b. Bioconductor	170
c. BioPerl	174
d. BioJava	177
e. BioJS	192
f. BioRuby	194
g. Bioclipse	202
h. EMBOSS	203
i. GenoCAD	205
<b>Chapter 6 Diverse Aspects of Bioinformatics</b>	<b>209</b>
a. Structural Bioinformatics	209
b. Modelling Biological Systems	209

c. Protein–protein Interaction Prediction	213
d. Interactome	217
e. Flow Cytometry Bioinformatics	227
f. Biodiversity Informatics	240

**Permissions**

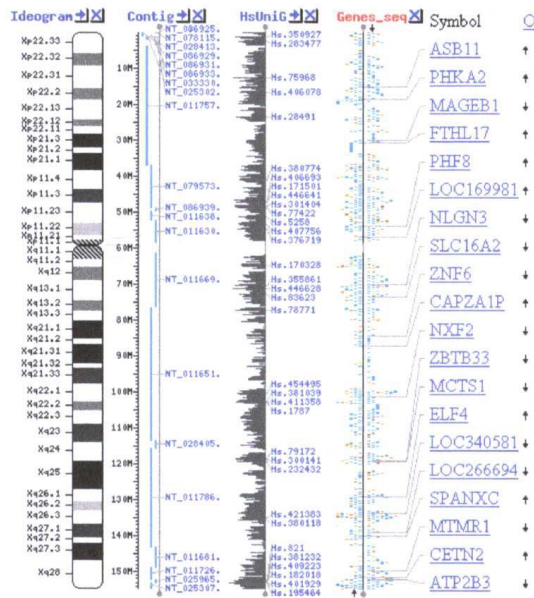
**Index**



# Introduction to Bioinformatics

Bioinformatics is a field of study that helps in the development of methods and software tools for developing a better understanding of biological data. It combines subjects such as computer science, statistics, mathematics and engineering. The chapter on bioinformatics offers an insightful focus, keeping in mind the complex subject matter.

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics has been used for *in silico* analyses of biological queries using mathematical and statistical techniques.



Map of the human X chromosome (from the National Center for Biotechnology Information website).

Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis “pipelines” that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidate genes and nucleotides (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organisational principles within nucleic acid and protein sequences, called proteomics.

## Introduction

Bioinformatics has become an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques such as image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, proteins as well as biomolecular interactions.

## History

Historically, the term *bioinformatics* did not mean what it means today. Paulien Hogeweg and Ben Hesper coined it in 1970 to refer to the study of information processes in biotic systems. This definition placed bioinformatics as a field parallel to biophysics (the study of physical processes in biological systems) or biochemistry (the study of chemical processes in biological systems).

## Sequences

```
5'ATGACGTGGGGA3'  
3'TACTGCACCCCT5'
```

Sequences of genetic material are frequently used in bioinformatics and are easier to manage using computers than manually.

Computers became essential in molecular biology when protein sequences became available after Frederick Sanger determined the sequence of insulin in the early 1950s. Comparing multiple sequences manually turned out to be impractical. A pioneer in the field was Margaret Oakley Dayhoff, who has been hailed by David Lipman, director of the National Center for Biotechnology Information, as the “mother and father of bioinformatics.” Dayhoff compiled one of the first protein sequence databases, initially published as books and pioneered methods of sequence alignment and molecular evolution. Another early contributor to bioinformatics was Elvin A. Kabat, who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991.

## Goals

To study how normal cellular activities are altered in different disease states, the bi-



ological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- Development and implementation of computer programs that enable efficient access to, use and management of, various types of information
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

## Relation to Other Fields

Bioinformatics is a science field that is similar to but distinct from biological computation, while it is often considered synonymous to computational biology. Biological computation uses bioengineering and biology to build biological computers, whereas bioinformatics uses computation to better understand biology. Bioinformatics and



computational biology involve the analysis of biological data, particularly DNA, RNA, and protein sequences. The field of bioinformatics experienced explosive growth starting in the mid-1990s, driven largely by the Human Genome Project and by rapid advances in DNA sequencing technology.

Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from graph theory, artificial intelligence, soft computing, data mining, image processing, and computer simulation. The algorithms in turn depend on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, and statistics.

Sequence Analysis

```

A8ASC3.1 14 SIKLPPSPQTRRLVERMANNLST..PSIFTRK..YGLSKEEAEENAKQIEEVACSTAHQ.....HYEKEPDGCGSAVQLYAKESKRLLEVLK 101
B4F917.1 13 SIKLPPSPQTRRLVERMANNLST..PSIFTRK..YGLSKEEAEENAKQIEEVACSTAHQ.....HYEKEPDGCGSAVQLYAKESKRLLEVLK 100
A9S1V2.1 23 VFKLPPSPQTRREAVRKMALKLSS..ACFESSG..VRIELADAGEHARRAIEEVAFGAQE.....ADSGGDKTGSRAVVMYAKHASKLMLETR 109
B9GSN7.1 13 SIKLPPSPQTRRLVERMANNLST..PSIFTRK..YGLSKEEAEENAKQIEEVACSTAHQ.....HYEKEPDGCGSAVQLYAKESKRLLEVLK 100
QBH056.1 30 SFSIUPPTQTRDRAVVRRLVDTLGG..DTILCKR..YGVAFPADEAPARRGIEAEAFDAAA..SGEAAATASVEEIGIKALQYSKEVSRRLDQVVK 120
Q0D423.2 44 SLSIUPPTQTRDRAVVRRLVDTLVA..PSILSKR..YGVAFPEAGRAHRAEAEAYATES..SSAAHAPASVEDGIEVLQYKESVSRRLLEAK 135
B9NMJ9.1 56 SFSIUPPTQTRDRAVVRRLVDTLST..TVLSKR..YGTIPKESEASRAIEEAFSGAST.....VASSEKDLQVLYKESVSRRLLEAK 141
Q0IYC5.1 29 SFAVUPPTQTRDRAVVRRLVAVLSGDTTALRKRYRYGVAFPADEAEARRAVEAQAFAASA..SSSSSSSVEDGIEVLQYSKEVSRRLAFVR 121
A9NM46.1 13 SIKLPPSPQTRRLVERMANNLST..PSIFTRK..YGLSKEEAEENAKQIEEVACSTAHQ.....HYEKEPDGCGSAVQLYAKESKRLLEVLK 100
Q9C500.1 57 SLSIUPPTQTRDRAVVRRLVDTLST..PSILSKR..YGTILSDDATTYVKLIEEAYGVASH.....AVSSDDGKILELYSKEISKRMLLEAK 142
Q2HR17.1 25 WYSIUPPTQTRDRAVVRRLVDTLST..PSVLSKR..YGTMSADEASRAIIEEAFSAVNA.....SSSTSHDVTLELYSKEISKRMLLEAK 110
Q9M7N3.1 28 SFSIUPPTQTRDRAVVRRLVDTLST..PSVLSKR..YGVTFEEDATSARRIEEAFSAVSV..ASAASGGRPEDEIIEVLHYSQIEIXQVWESAK 119
Q9M7N6.1 25 SFSIUPPTQTRDRAVVRRLVDTLST..PSILSKR..YGTLPDEASETHRIIEEAFSAAGS.....TASDADGIEILQVYSKEISKRMLDQVVK 110
Q9LEB2.1 14 SIKLPPSPQTRRLVERMANNLST..PSIFTRK..YGLSKEEAEENAKQIEEVACSTAHQ.....HYEKEPDGCGSAVQLYAKESKRLLEVLK 101
Q9M651.2 13 SIKLPPSPQTRRLVERMANNLST..PSVLSKR..YGTISDEAESARRIEEAFSAVNA.....QFEREPDGGGSAVQLYAKESKRLLEVLK 100
B9K749.1 48 SLSIUPPTQTRDRAVVRRLVDTLST..PSVLSKR..YGTISDEAESARRIEEAFSAVNA.....ATSHEDGIEILQVYSKEISKRMLDQVVK 133
```

The sequences of different genes or proteins may be aligned side-by-side to measure their similarity. This alignment compares protein sequences containing WPP domains.

Since the Phage  $\Phi$ -X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode proteins, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs such as BLAST are used daily to search sequences from more than 260 000 organisms, containing over 190 billion nucleotides. These programs can compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself.

DNA Sequencing

Before sequences can be analyzed they have to be obtained. DNA sequencing is still a non-trivial problem as the raw data may be noisy or afflicted by weak signals. Algorithms have been developed for base calling for the various experimental approaches to DNA sequencing.

Sequence Assembly

Most DNA sequencing techniques produce short fragments of sequence that need to be assembled to obtain complete gene or genome sequences. The so-called shotgun



sequencing technique (which was used, for example, by The Institute for Genomic Research (TIGR) to sequence the first bacterial genome, *Haemophilus influenzae*) generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly usually contains numerous gaps that must be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

## Genome Annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of sequencing has ceased to pose a bottleneck. Annotation is made possible by the fact that genes have recognisable start and stop regions, although the exact sequence found in these regions can vary between genes.

The first genome annotation software system was designed in 1995 by Owen White, who was part of the team at The Institute for Genomic Research that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. White built a software system to find the genes (fragments of genomic sequence that encode proteins), the transfer RNAs, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA, such as the GeneMark program trained and used to find protein-coding genes in *Haemophilus influenzae*, are constantly changing and improving.

Following the goals that the Human Genome Project left to achieve after its closure in 2003, a new project developed by the National Human Genome Research Institute in the U.S appeared. The so-called ENCODE project is a collaborative data collection of the functional elements of the human genome that uses next-generation DNA-sequencing technologies and genomic tiling arrays, technologies able to generate automatically large amounts of data with lower research costs but with the same quality and viability.

## Computational Evolutionary Biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists by enabling researchers to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational population genetics models to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

Future work endeavours to reconstruct the now more complex tree of life.

The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are not necessarily related.

## Comparative Genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

Many of these studies are based on the homology detection and protein families computation.

## Pan Genomics

Pan genomics is a concept introduced in 2005 by Tettelin and Medini which eventually took root in bioinformatics. Pan genome is the complete gene repertoire of a particular taxonomic group: although initially applied to closely related strains of a species, it can be applied to a larger context like genus, phylum etc. It is divided in two parts- The Core