1 Introduction

"Grundlagen der Bioinformatik"
Summer 2018

Prof. Daniel Huson

Office hours: Thursdays 17-18h (Sand 14, C310a)

1.1 Time, place and problem sessions

Lectures:

Mondays	10:15-12h	Hörsaal 2, Sand 6/7
Wednesdays	10:15-12h	Hörsaal 2, Sand 6/7

Groups:

Tutor	Time	Where
Benjamin Albrecht	Wed 12-14	Hörsaal 1, Sand 6/7
Sina Beier	Wed 12-14	C311, Sand 14
Klara Reichard	Wed 8-10	Hörsaal 2, Sand 6/7

We will use the university *Ilias* learning platform to publish the script and assignments.

Also, you will upload your assignment solutions there.

A link is provided on this website:

http://ab.inf.uni-tuebingen.de/teaching/sose2018/gbi

1.2 How to get credit

This course is worth 9 ECTS credits. Hence, you will be expected to invest approximately $9 \times 30 = 270$ hours in it.

Each week, you will be assigned a set of practical and theoretical problems (Übungsblatt). Your solutions will be graded and you will be asked to present them in the tutorials (Übungsgruppen). Participation in the tutorial is mandatory, if you miss more than two sessions, then you will not pass the course.

Your grade will be based on an exam. To be admitted to the exam, you must achieve at least 50% of all assignment points.

The exam will be taken in two parts:

- First part of exam (covering first half of course): June 6^{st}
- Second part of of exam (covering second half of course): July 25^{th}

If you miss or fail a part of the exam, then the make-up dates are:

- Makeup part 1: Wed June-20, 8-10h, Hörsaal 2, Sand 6/7.
- Makeup part 2: Fri Sep-28, 10-12h, Hörsaal 2, Sand 6/7.

1.3 Script and assignments

At the beginning of each lecture, a script covering the current content of the course will be handed out. The script will also be made available online.

Assignment sheets will usually be handed out and posted online on Mondays.

You must hand in your solutions on the following Monday before the beginning of the lecture.

1.4 Bioinformatics

The following characterization is from NCBI (Nationl Center for Biotechnology Information):

Bioinformatics: *Major advances in the field of molecular biology and advances in genomic technologies have led to an explosive growth in the biological information generated by the scientific community.*

This huge amount of genomic information has led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

1.4.1 What is a biological database?

A *biological database* is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.

For example, the NCBI-nr database contains 149 million different (nr = "non-redundant") protein sequences, approximately 54 billion letters in total, that is, 54 GB of sequence (March 2018).

A simple database might be a single file containing many records, each of which includes the same set of information. For example, a record associated with a nucleotide sequence database typically contains information such as contact name, the input sequence with a description of the type of molecule, the scientific name of the source organism from which it was isolated, and often, literature citations associated with the sequence.

For researchers to benefit from the data stored in a database, two additional requirements must be met:

- easy access to the information
- a method for extracting only that information needed to answer a specific biological question

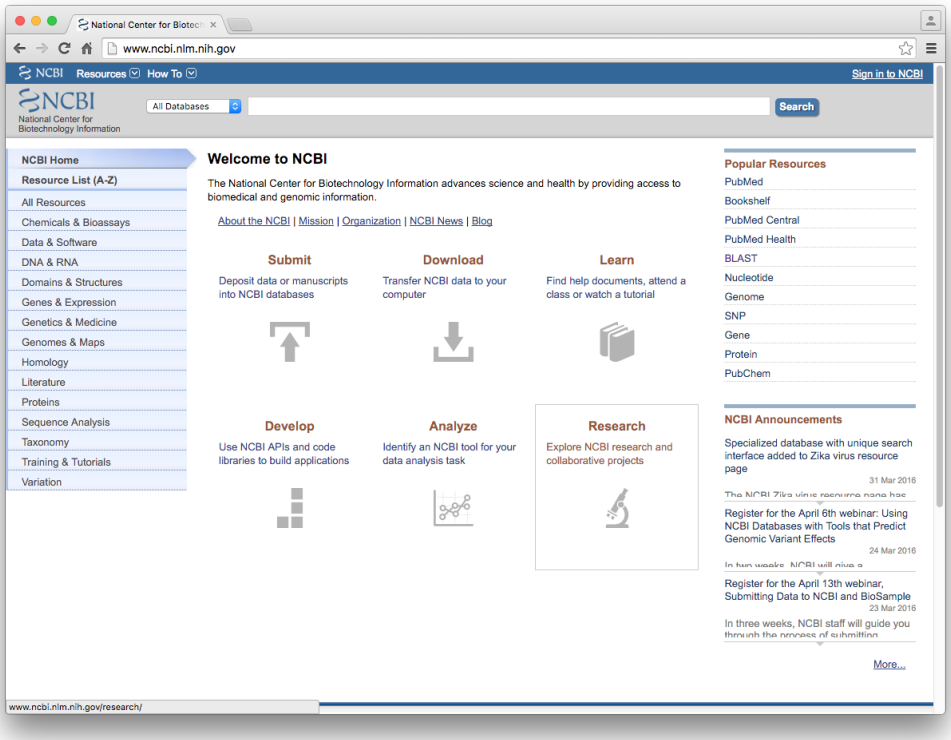
NCBI provides the largest publicly available database at: http://www.ncbi.nlm.nih.gov. [1]

NCBI hosts the GenBank sequence database, which is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.

The NCBI website links many different databases through a single search and retrieval system.

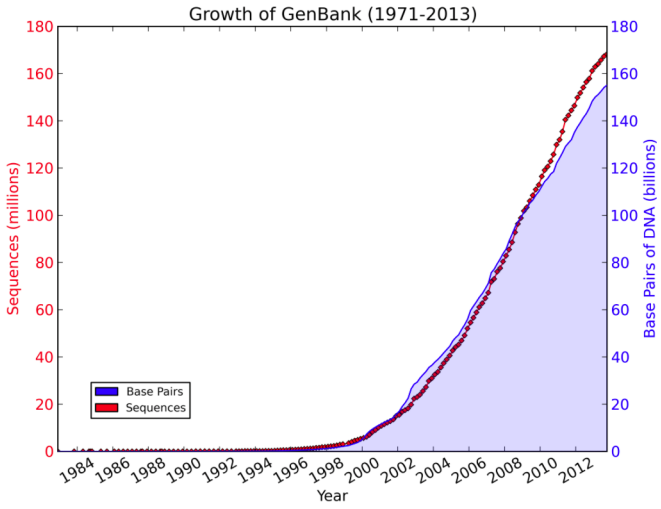
For example, the NCBI-nr *protein* database is cross-linked to the NCBI *taxonomy* database. This allows a researcher to find taxonomic information for the species from which a protein sequence was derived. (Taxonomy is a division of the natural sciences that deals with the classification of animals and plants.)

[1] NCBI Resource Coordinators (2013). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013 Jan;41(Database issue):D8-D20. https://doi.org/10.1093/nar/gks1189



(http://www.ncbi.nlm.nih.gov, accessed 1-Apr-16)

The number of sequences available from NCBI is growing at an exponential rate:



(Image due to: Mark A. Pauley, University of Nebraska at Omaha, 2013)

1.4.2 What is Bioinformatics?

NCBI: *Bioinformatics* is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.[2]

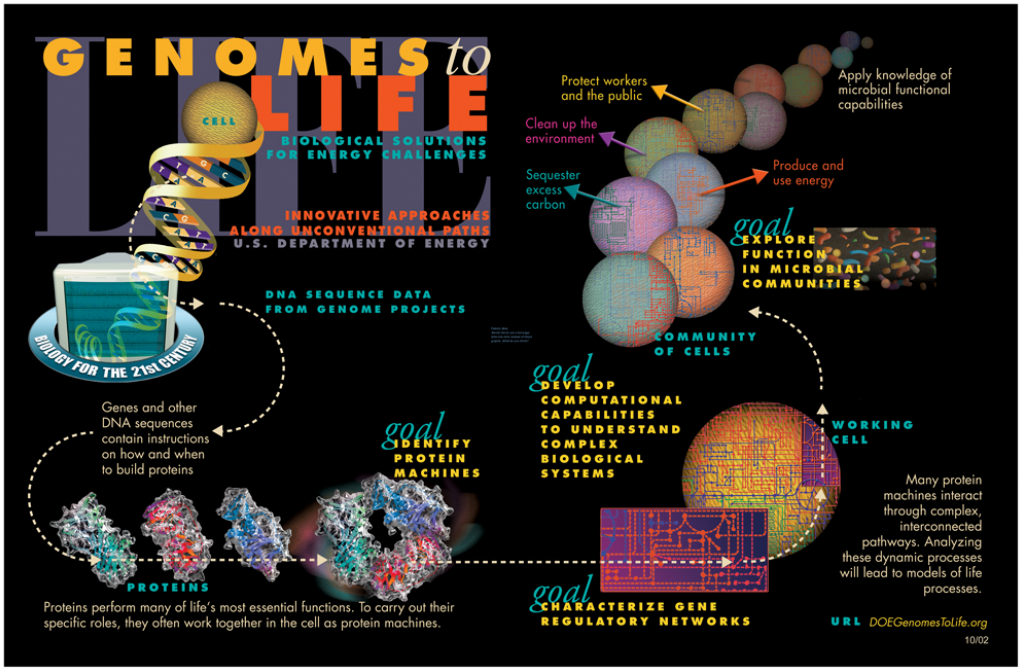
[2] https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html accessed 15.4.2018

At the beginning of the "genomic revolution", a major bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences.

Development of this type of database involved both design issues and the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

Ultimately, however, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states.

Therefore, the field of bioinformatics has evolved such that the focus is on *developing* and *applying* methods for the analysis and interpretation of various types of data, including 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:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information, and
- the development of new mathematical models, algorithms and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

1.5 Overview

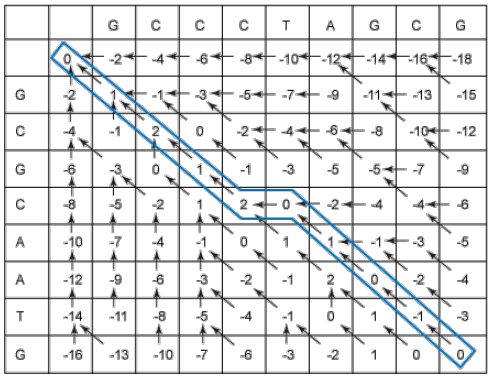
Here is an overview of the course:

Introduction

Pairwise alignment

The comparison of two sequences is perhaps the most important computation addressed by bioinformatics. How to solve this efficiently? How to score an alignment? Which variants of the problem exist?

```
GCCCTAGCG       GCCCTAGCG
          ->    || | |  |
GCGCAATG        GCGC-AATG
```



(http://www.ibm.com/developerworks/opensource/library/j-seqalign/index.html, accessed 1-Apr-16)

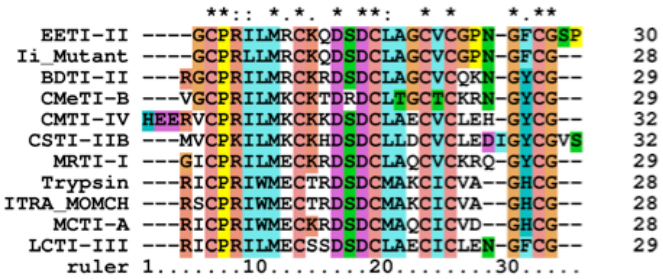
BLAST

Searching for similar sequences in large databases requires a fast heuristic for pairwise alignment. We will discuss BLAST, the most widely-used bioinformatics tool.

Multiple alignment

"Two homologous sequences whisper . . . a full multiple sequence alignment shouts out loud." (Arthur Lesk)

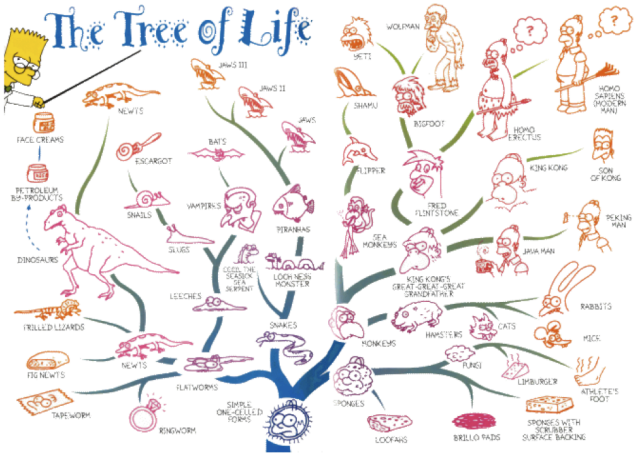
Although obtaining an optimal multiple alignment is computationally hard, biologists often require multiple sequence alignments. We will discuss some of the basic heuristics.



Phylogenetic analysis

"Nothing in biology makes sense except in the light of evolution" (Theodosius Dobzhansky)

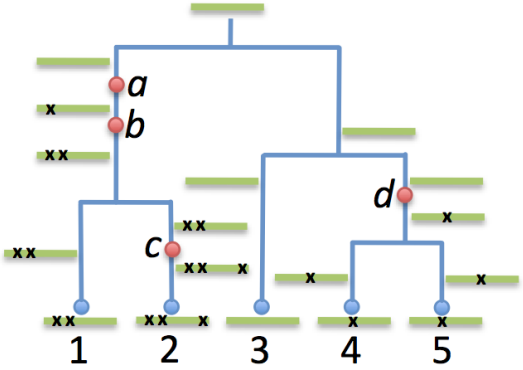
Phylogenetic analysis based on molecular sequences is important for understanding many aspects of biology. We will look at the most important distance- and character-based methods.



Bart Simpson's Tree of Life, © Matt Groening, Bart Simpson's Guide to Life (1993, Harper Collins)

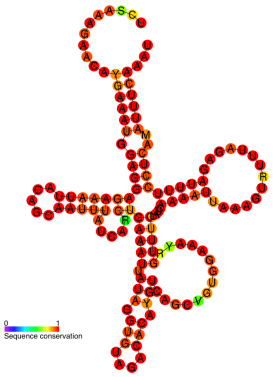
Population genetics

Given a current-day population of individuals, represented by a collection of genes, the goal of population genetics is to infer details of the evolutionary processes that produced the population.



RNA secondary structure

An RNA sequence is single stranded and folds into a characteristic secondary structure by forming base pairs with itself. How to computationally predict the secondary structure of an RNA sequence?



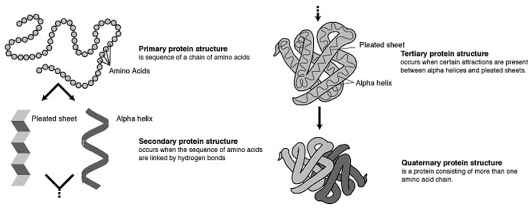
(http://en.wikipedia.org/wiki/File:HAR1F_RF00635_rna_secondary_structure.jpg)

Protein secondary structure

What are the secondary structure elements of a protein? How do they form? Given only the primary sequence of a protein, how to predict its secondary structure?

Protein tertiary structure

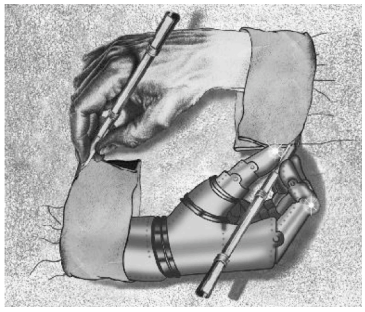
The tertiary structure of a protein determines its function. How to predict the tertiary structure of a protein computationally (a) when the structure of a homologous protein is already known, and (b) when no such information is available.



(http://www.umass.edu/molvis/workshop/prot1234.htm, accessed 1-Apr-16)

Machine-learning methods

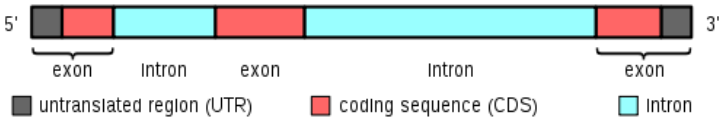
One important class of techniques for analysing biological data are machine-learning approaches. We will discuss some examples of uses of HMMs or SVMs.



(http://www.di.unito.it/WWW/ICML96/home.html, accessed 1-Apr-16)

Gene finding

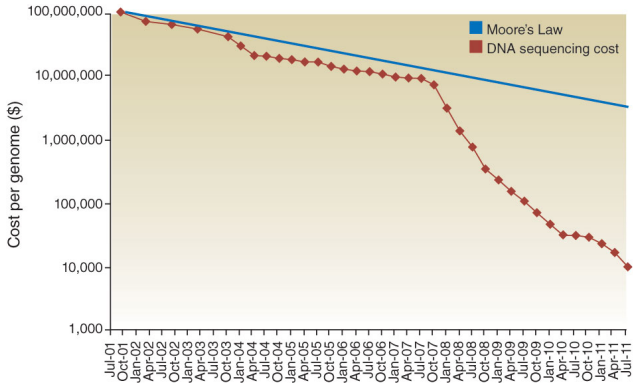
The goal of computational sequence analysis is to find all genes, transcription factor binding sites and other important sequence features in a genome. We will discuss some of the basic techniques.



(Image source: wikimedia.org)

Sequencing, assembly and resequencing of genomes

The most fundamental data associated with an organism is its genome. How is DNA sequenced? What is genome assembly? What is resequencing?



(Nature Biotechnology 30, 20-25 (2012))

Human genomics

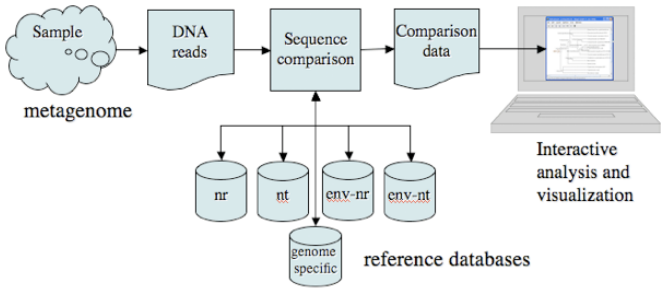
Increasingly, bioinformatics is being used to address questions of medical relevance.

1000 Genomes Project:
"The goal ... is to find most genetic variants that have frequencies of at least 1% in the populations studied." See http://www.1000genomes.org.

Human Microbiome Project:
"The broad goal of this five year effort is to catalog and characterize the microbes living in and on the human body (the microbiome)." See http://precedings.nature.com/collections/human-microbiome-project, accessed 1-Apr-16.

Microbiome analysis

Ultra high-throughput sequencing technologies paired with huge computational resources allow one to study the communities of microbes contained in samples from water, soil, the human body or other environments. We will look at some of the main approaches used.



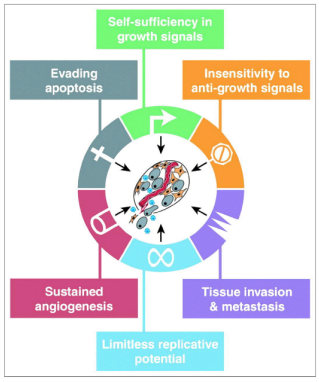
Cancer Genomics

Cancer is the second most frequent cause of death in western societies, after cardiovascular disease.

An understanding of cancer pathlogy and progression is slowly emerging.

Cancer takes on many different forms and involves many different types of anomalies.

In "personalized medicine", one idea is to use DNA sequencing and assembly to analyse cancer cell genomes of a patient so as to develop a targeted treatment.



(Image: Hanahan and Weinberg (2000))