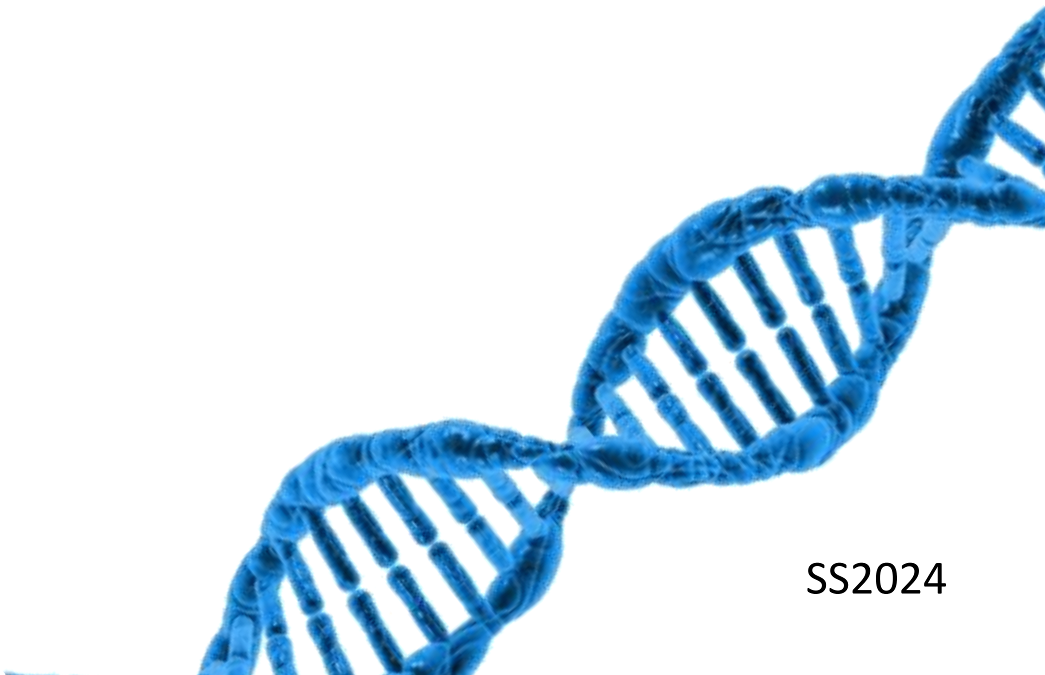
Algorithms and Tools in Bioinformatics

Data, Tools and Technologies in Bioinformatics

Julia Vetter

julia.vetter@fh-hagenberg.at



SS2024

Course Content

(1) Overview

(2) Standard Datasets/Modern File Formats

(3) Databases/Platforms

(4) Data (Pre-) Processing

(5) Tools

(6) Machine Learning

(4) Data (Pre-) Processing

“Real-Life” Examples:

~~Population Genetics~~

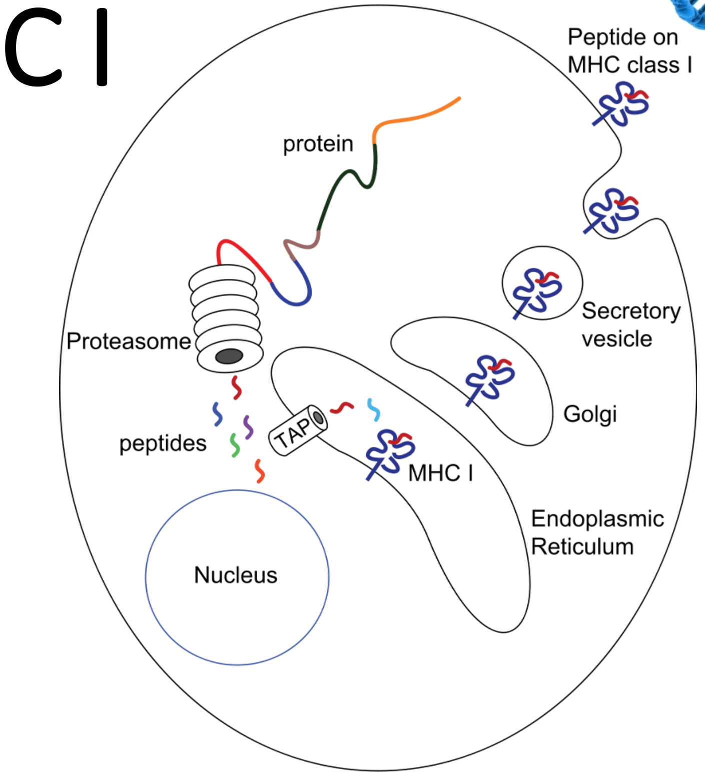
~~NGS Data~~

~~Genome Data~~

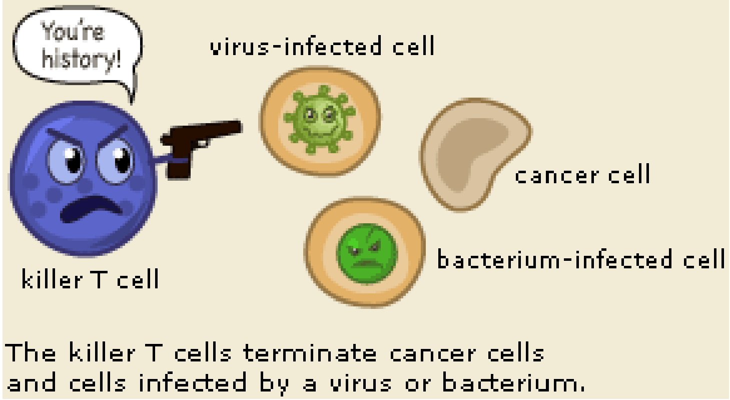
Protein Data

Example: Natural Killer Cells & MHC I

- MHC Class I:
  - = major histocompatibility complex molecule
  - On cell surface of all nucleated cells (in vertebrates) (also on platelets, but not on red blood cells)
  - Present peptides from cytosolic proteins
  - In humans known as **HLA system** (HLA-A, HLA-B, HLA-C are linked to MHC I)
  - (Associate with β2-microglobulin)
- Natural Killer Cells:
  - Belong to the lymphocytes
  - Recognize abnormal cells (tumor or virus infected cells)
  - Innate immune system
  - Detect MHC I
- PDB-ID: 7DUU

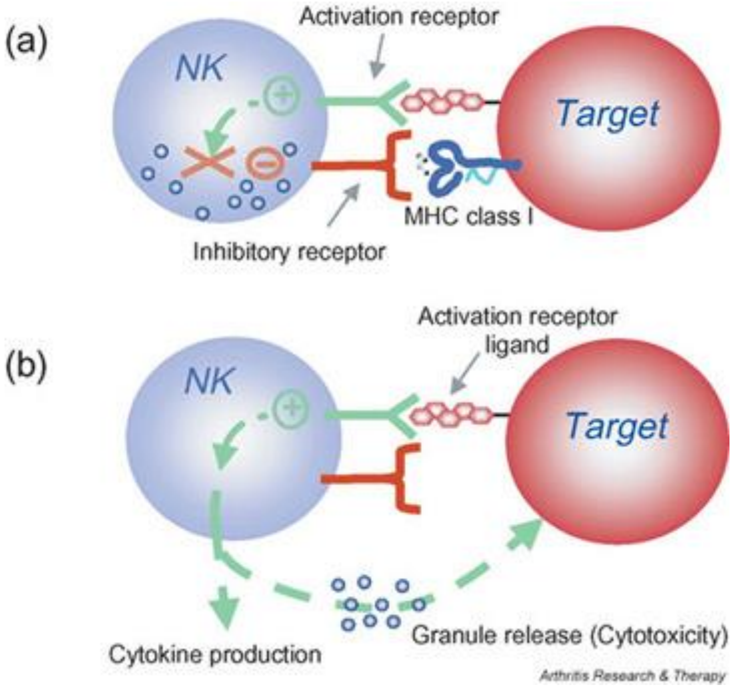


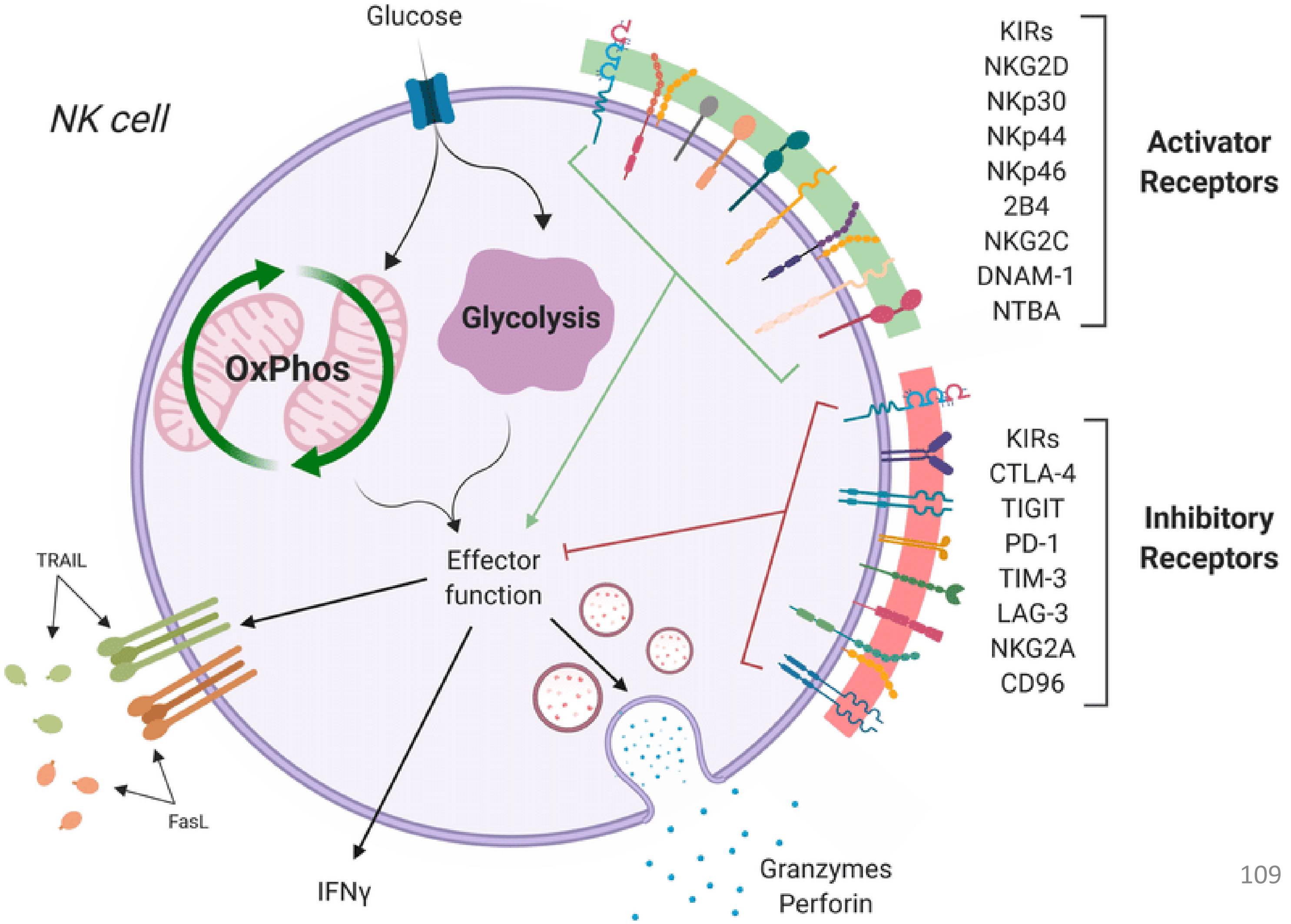
Source: Wikipedia



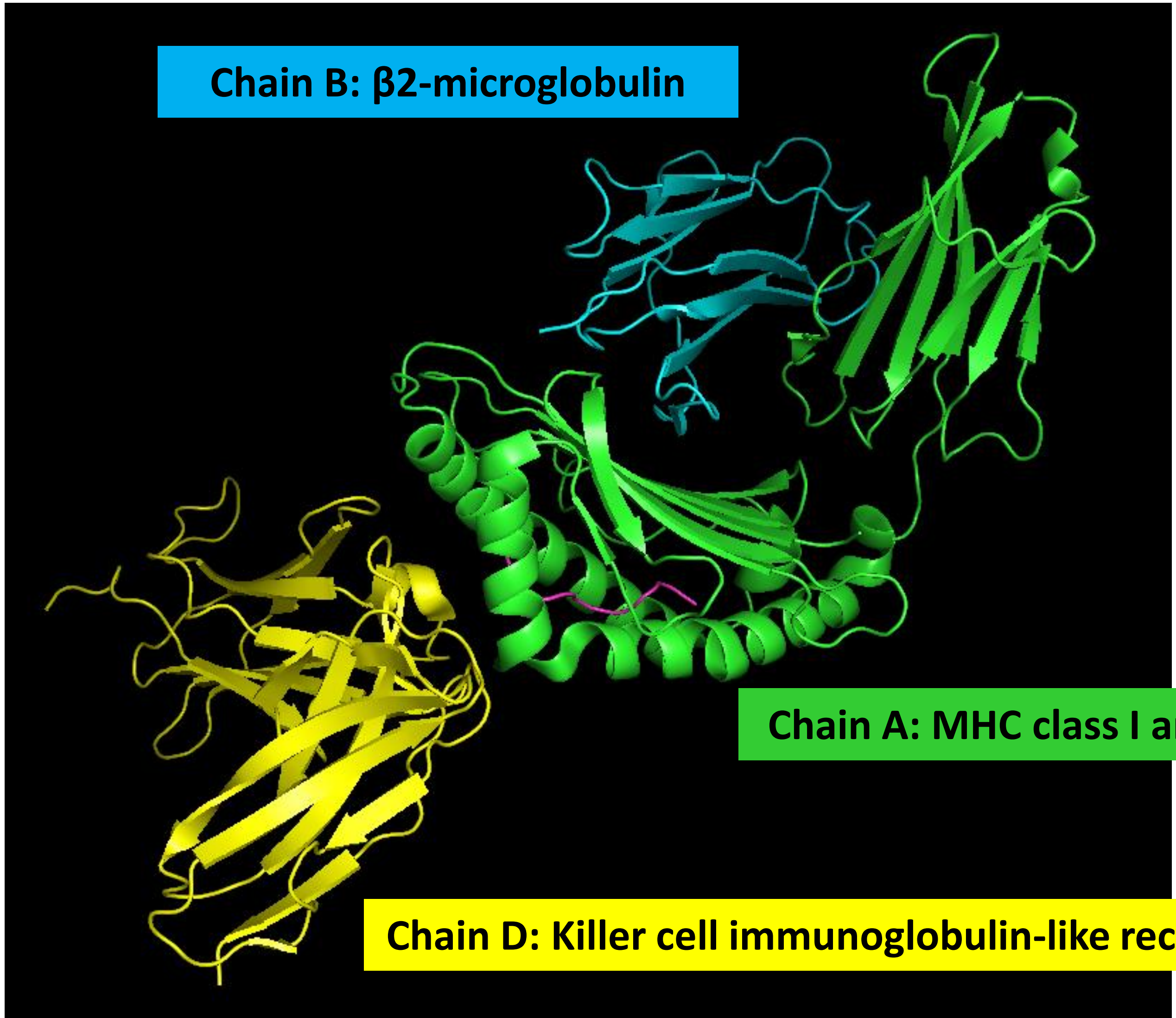
The killer T cells terminate cancer cells and cells infected by a virus or bacterium.

Example: Natural Killer Cells & MHC I





7DUU: Crystal structure of HLA molecule with an KIR receptor



Chain B: β2-microglobulin

Chain A: MHC class I antigen

Chain D: Killer cell immunoglobulin-like receptor (KIR) DS2

Hands-on...

```
Colab: ATBI_3

# 1. load file
# 2. color chains
# 3. remove solvents
# 4. save structure
# 5. show image
# 6. get FASTA sequences
# 7. use AlphaFold2 for structure prediction
     here: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb
     (one sequence)
     or here: https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb
     (multiple sequences)
# 8. load one predicted structure
# 9. align predicted and "true" structure and calculate root-mean-square deviation (RMSD)
     of atomic positions
```

(5) Tools

Tools

- .NET Bio
- AMPHORA
- Anduril
- Ascalaph Designer
- AutoDock
- Avogadro
- BEDtools
- Bioclipse
- Bioconductor
- BioJava
- BioJS
- BioMOBY
- BioPerl
- BioPHP
- Biopython
- BioRuby
- BLAST
- Bowtie2
- BWA
- Clustal Omega
- CP2K
- EMBOSS
- Ensembl
- FastQC
- Galaxy
- GATK
- GenePattern
- Geworkbench
- GMOD
- GenGIS
- Genomespace
- GENtle
- GROMACS
- IGV
- InterMine
- LabKey Server
- LAMMPS
- Mothur
- Orange
- ORFfinder
- PathVisio
- Picard
- pyMOL
- SAMtools
- SOAP Suite
- Staden Package
- Taverna workbench
- UGENE
- Unipept
- VOTCA
- Wordom

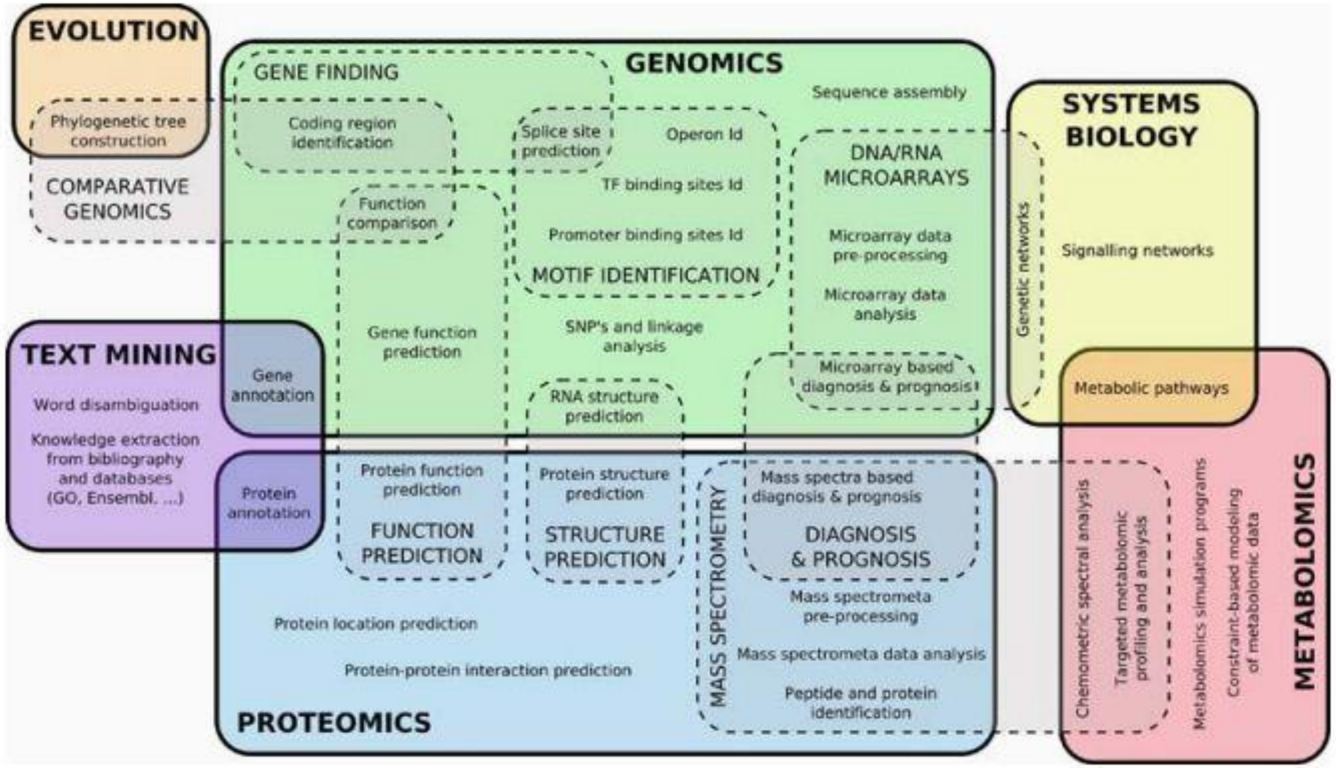
Adapted from https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software

(6) Machine Learning

Feature Selection, Classification, Regression, Clustering

Why Machine Learning in Bioinformatics?

- Gene finding
- Motif identification
- Microarray data analysis/diagnosis/prognosis
- RNA structure prediction
- Protein structure prediction
- Protein function prediction
- Protein-protein interaction prediction
- Knowledge extraction


EVOLUTION
Phylogenetic tree construction
COMPARATIVE GENOMICS
GENE FINDING
GENOMICS
Coding region identification
Splice site prediction
Operon Id
Sequence assembly
Function comparison
TF binding sites Id
Promoter binding sites Id
MOTIF IDENTIFICATION
DNA/RNA MICROARRAYS
Microarray data pre-processing
Microarray data analysis
Genetic networks
SYSTEMS BIOLOGY
Signalling networks
TEXT MINING
Word disambiguation
Knowledge extraction from bibliography and databases (GO, Ensembl, ...)
Gene annotation
Gene function prediction
SNP's and linkage analysis
RNA structure prediction
Microarray based diagnosis & prognosis
Metabolic pathways
Protein annotation
Protein function prediction
FUNCTION PREDICTION
Protein structure prediction
STRUCTURE PREDICTION
MASS SPECTROMETRY
Mass spectra based diagnosis & prognosis
DIAGNOSIS & PROGNOSIS
Mass spectrometa pre-processing
Mass spectrometa data analysis
Peptide and protein identification
Protein location prediction
Protein-protein interaction prediction
PROTEOMICS
Chemometric spectral analysis
Targeted metabolomic profiling and analysis
Metabolomics simulation programs
Constraint-based modeling of metabolomic data
METABOLOMICS

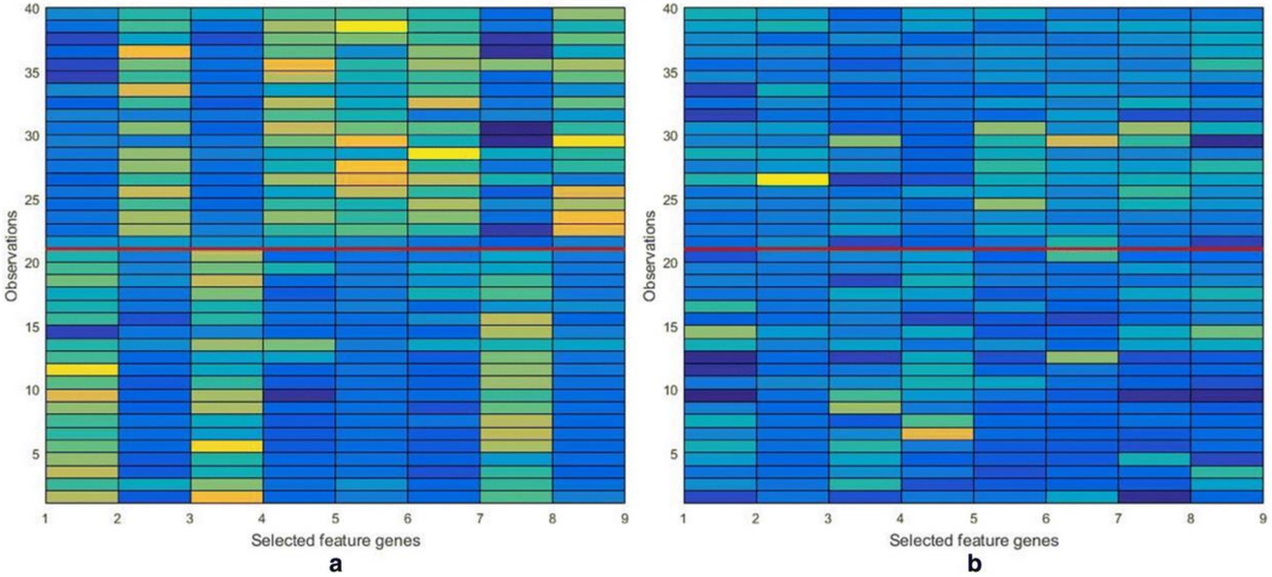
https://omicstutorials.com/introduction-to-machine-learning-bioinformatics/

Feature Selection

- = process of removing features from the data set that are irrelevant with respect to the task
- Techniques:
  - Filter-based
    - Selection of optimum feature subset by using statistical metrics
    - e.g., t-test feature selection, correlation-based feature selection
  - Wrappers
    - Building classification models to determine feature importances
    - e.g., genetic algorithms (GA)
  - Embedded
    - Are found within the learners themselves
    - e.g., SVM method of recursive feature elimination (RFE), Random Forest (RF)

Fig. 6

From: Feature selection of gene expression data for Cancer classification using double RBF-kernels



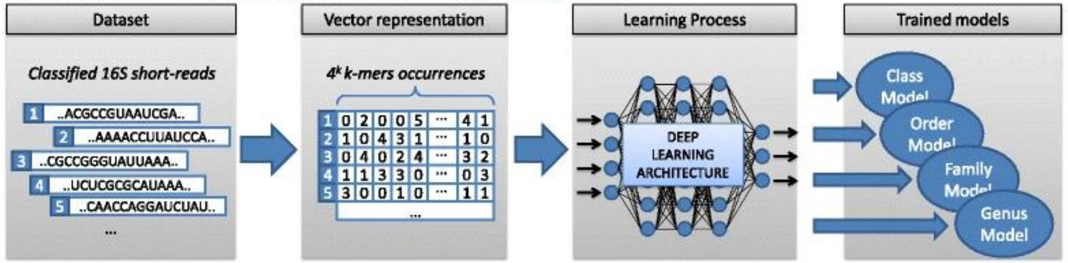
The colormap of the expression profiles for nine most significant genes selected by DKBCGS (**a**) and for 9 randomly chosen genes (**b**). The red line distinguishes between cancer samples and normal samples

Liu et al. 2018

Classification

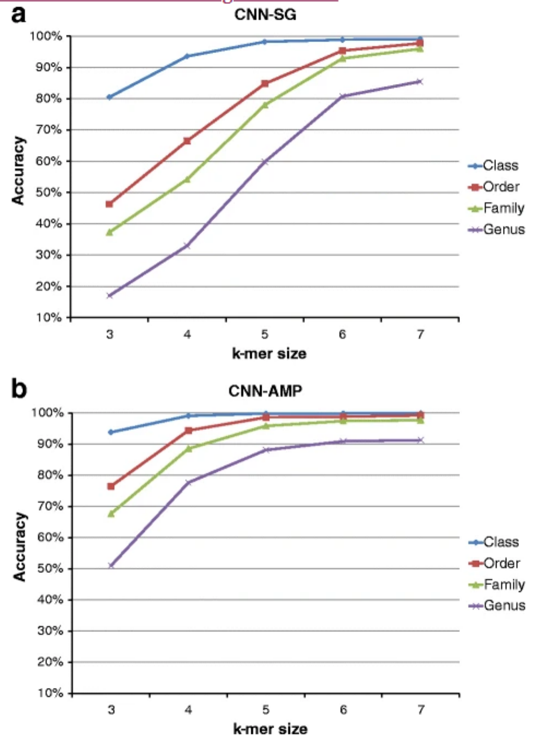
- Labeling genomic data – taxonomic classification

From: Deep learning models for bacteria taxonomic classification of metagenomic data



Proposed training process. Starting from 16S reads, we proposed a vector representation and a deep learning architecture to obtain trained models for taxonomic classification

From: Deep learning models for bacteria taxonomic classification of metagenomic data

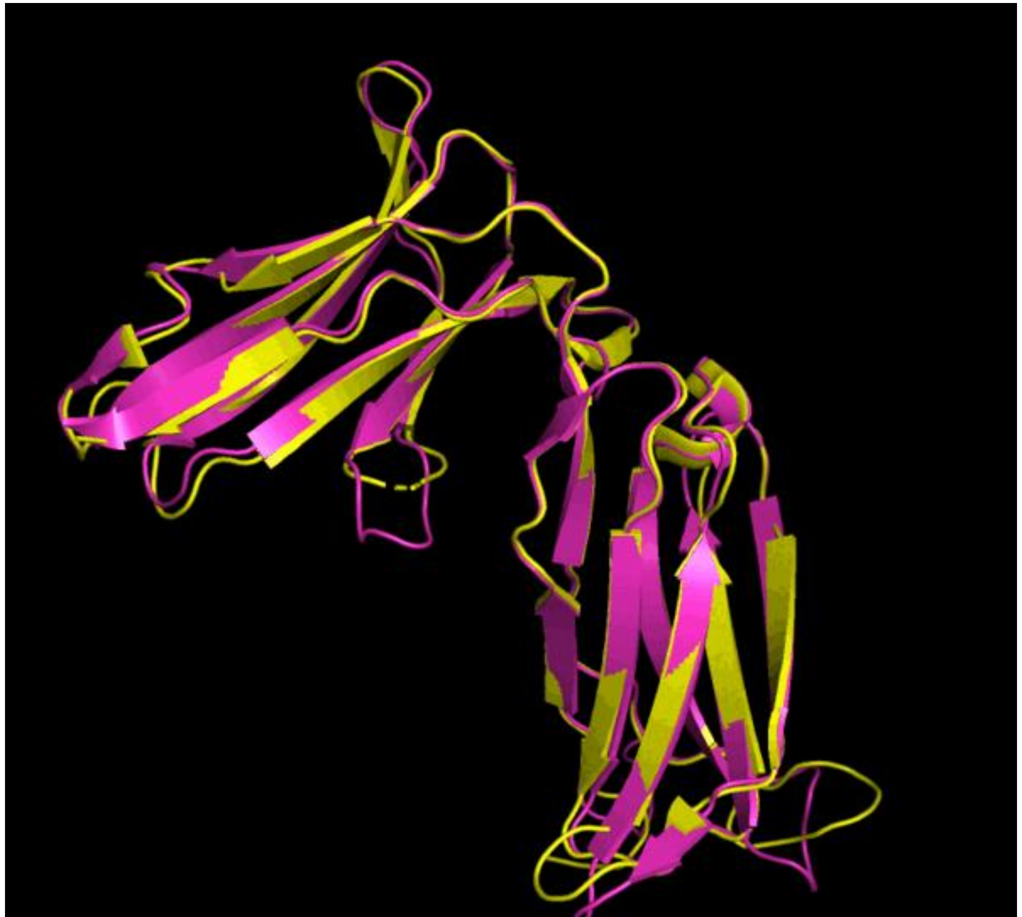


Fiannaca et al., 2018

- Classify biological sequences by learning sequences and structures
- Diseased vs. healthy
- etc.

Regression (Predictions)

- Protein structure prediction
- Protein-Protein interactions prediction
- Protein-Ligand interaction prediction
- Microarray data prognosis
- Therapy outcome prediction

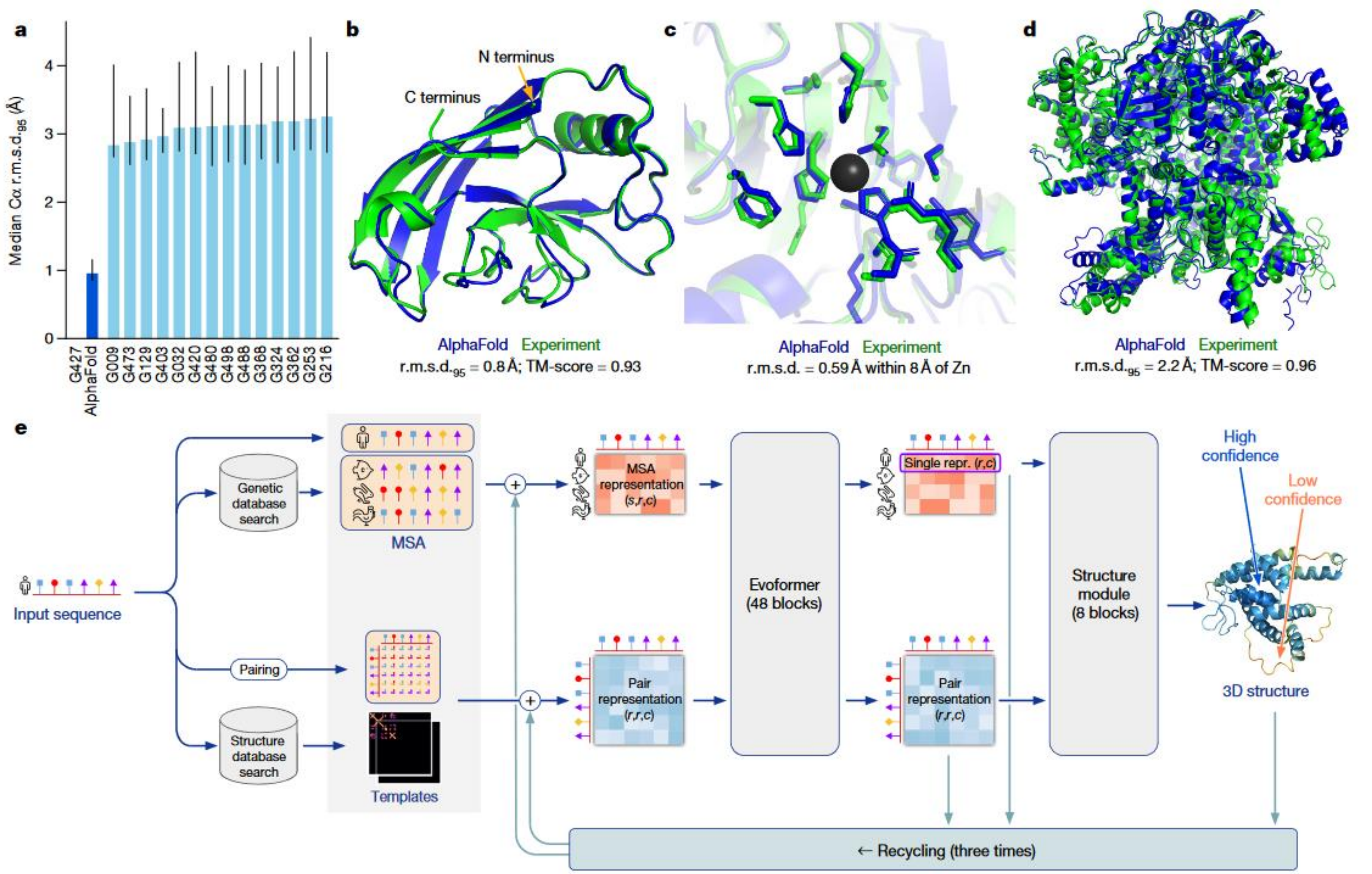


7DDU:D – predicted and measuresd

480238_1_En_Print.indd - 202

CASP3 (19
CASP2 (19
CASP1 (19
Initiatives
Data Archi
Proceeding
CASP Meas
Feedback
Assessors
People
Community
Job Fair

Article

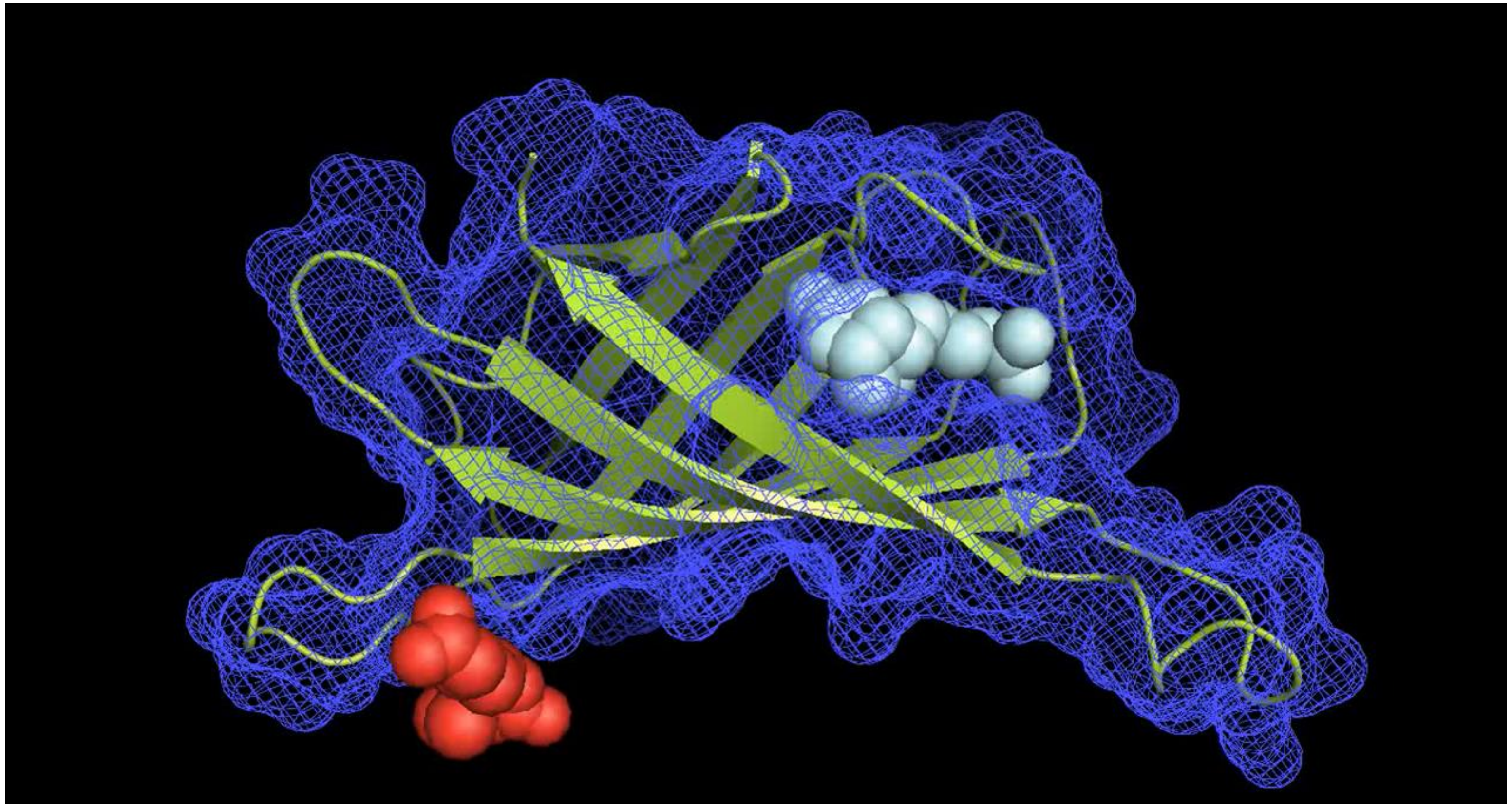


a
Median Cα r.m.s.d.$_{95}$ (Å)
0
1
2
3
4
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.$_{95}$ = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.$_{95}$ = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

core	Rank AVG Zscore (>0.0)
	1
	2
	3
	4
	5
	7
	8

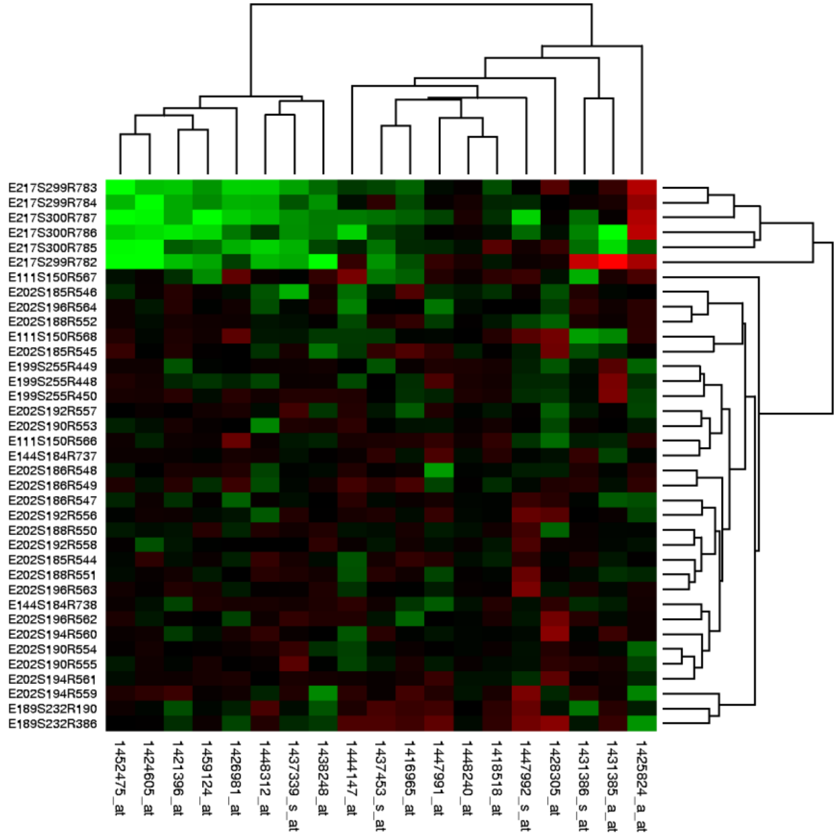
Protein/Ligand Docking

- Rosetta docking
- AutoDock
- ClusPro
- HADDOCK
- ZDOCK
- etc.



Clustering

- Identify patient subtypes
- Phylogenetic analysis
- Sequence clustering
- Quality check – do replicates cluster together
- Cluster genes/proteins to identify functions
- Multiple sequence alignment



https://upload.wikimedia.org/wikipedia/commons/4/48/Heatmap.png