MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

University of Cyprus

MAI643: Artificial Intelligence in Medicine

Kalia Orphanou

January – May 2023

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Translational Bioinformatics

(Some material drawn from Hersh, WR, 2022. Health Informatics: Practical Guide, 8th Edition, slides).

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Translational Bioinformatics

1. Introduction to translational bioinformatics and the relevant terminology
2. Gene expression and other gene-association problem tasks
3. Phenotyping
4. AI in translational bioinformatics
   - Types of data in bioinformatics
   - Examples of databases
   - Bioinformatics data processing
   - Supervised and Unsupervised learning
4. Biomarkers
5. Precision Medicine
6. Genetic Disorders
   - Security and privacy issues in bioinformatics

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

INTENDED LEARNING OUTCOMES

Upon completion of this unit on Medical image informatics and interpretation, students will be able to:

1. Define translational bioinformatics
2. Point out the key translational bioinformatics problem that AI methods are positioned to solve
3. Apply AI techniques and processing methods to a translational bioinformatics problem
4. Know some important “-omic” databases that can be used to interpret and validate translational bioinformatics
5. Discuss how precision medicine differs from public health
6. Describe the different types of data that can be used in translational bioinformatics
7. Appreciate the role of biomarkers in precision medicine
8. Understand the difference between simple and complex genetic disorders and know few examples of each category.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Omics (Genetic) Data

- **Gen**omics (sequence annotation - proteins in an organism)
- **Transcript**omics(microarray - DNA transcribed)
- **Prote**omics (proteome databases)
- **Metabol**omics (enzyme annotation - molecules involved in metabolism)
- **Flux**omics(isotopic tracing, metabolic pathways)
- **Phen**omics(biomarkers - traits of an organism)
- **Epigen**omics(epigenetic modifications)
- **Microbi**omics(microorganisms inhabiting an individual)
- **Lipid**omics(pathways of cellular lipids)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bioinformatics Terminology

- **Bioinformatics:** the field of computational science that search for patterns within DNA sequences and similarities between sequences within the same organism or between organisms.
- **Translational Bioinformatics:** the field of Bioinformatics applied to humans. Concerns the development of novel techniques for processing.
- **Computational Biology:** application of computational methods to discover biological principles.

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bioinformatics Terminology (ctd.)

- **Genome:** Total collection of genetic data (DNA) for a single person or organism.
- **Genes:** Discrete units encoded in DNA sequence
  - they are copied into **ribonucleic acid (RNA),** which has a composition very similar to DNA.
  - they are copied into **messenger RNA (mRNA)** and a majority of mRNA sequences are translated into protein.
- **Proteins:** 3-D structure, transform information hidden in genes.
- **Biomarker:** Measurable indicator of molecular, histologic, radiographic, or physiologic characteristics used to assess normal or pathologic process.

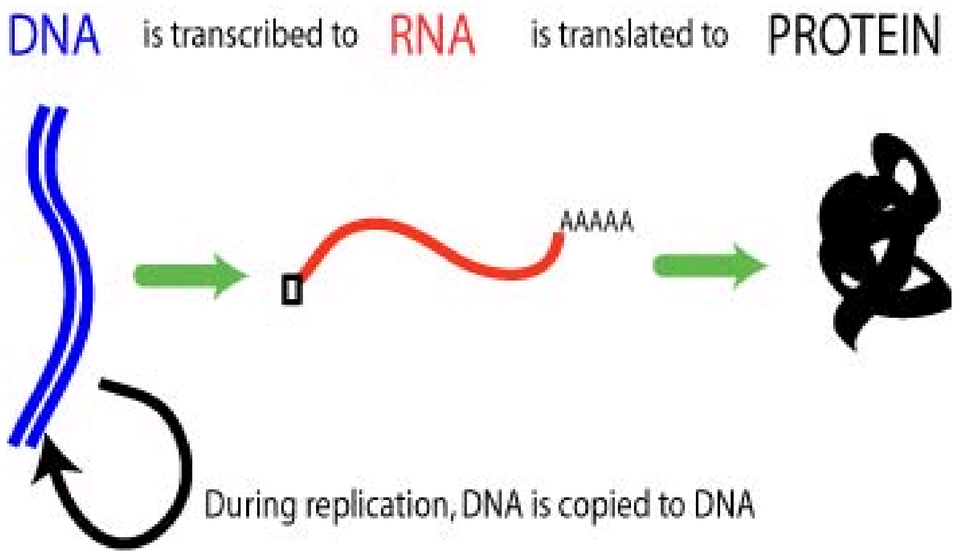
From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Central Dogma of Biology

- DNA is transcribed to RNA
- RNA is translated to protein
- Gene expression refers to how often or when proteins are created from the instructions within the genes.
- Many regulatory processes control these steps



From Shortliffe et al., 2021

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Central Dogma is no Longer so Simple

- Genes are not contiguous on DNA
- Epigenetics influences expression of genes (do not change the DNA sequence, but the body reads a DNA sequence).
- While genetic changes can alter which protein is made, **epigenetic changes** affect gene expression to turn genes “on” and “off.”
- **But**
  - Growing discovery of associations of genes with disease
  - Increasing understanding of the role of genetics in many diseases, such as cancer

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Translational Bioinformatics

"**Translational bioinformatics** research integrates information about molecular entities (DNA, RNA, proteins, small molecules, and lipids) with information about clinical entities (patients, diseases, symptoms, laboratory tests, pathology reports, clinical images, and drugs) to improve patient care." (Altman, 2012)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

History of Translational Bioinformatics

- Started in 1996
  - ➢ Organization of biomedical data
- In 2003, advancement in this area after the announcement of the Human Genome Project
  - ➢ Genomic analysis was added to that field
- Since 2005 in Europe and 2009 in the United States, EHR have been widely used in translational bioinformatics.
- Today, translational bioinformatics is a **multidisciplinary field**, extending from the molecular level (genes, proteins, and other molecular entities below the cell) to the population level (collections of living subjects).

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Primary Data Categories in Translational Bioinformatics

- **Genomic data**: genes, proteins, miRNAs, metabolites
- **Clinomic data:** any quantitative biomedical data that are useful for medical research and interventions
- **Phenotypic data**: diseases and abnormalities affecting each subject i.e. cancer, diabetes. Directly derived from clinomic observations.

From shortliffe et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Phenotyping

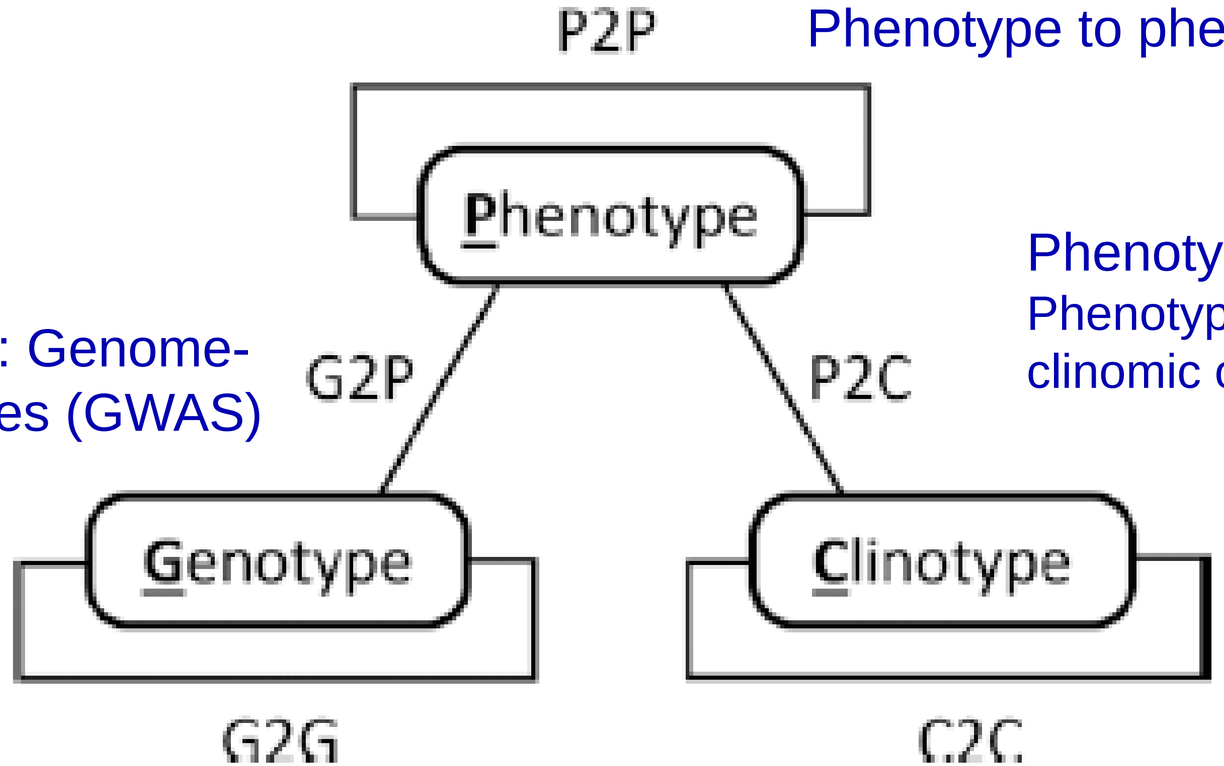
- Subjects' categorization and definition as assigned by biomedical experts. i.e. **'cell proliferation'** and **'chemotherapy resistance',**
- Phenotypic data are directly derived from experts' observations based on the subject's biomedical data.
- Major source of data is EHR: http://www.rethinkingclinicaltrials.org/resources/ehr-phenotyping/

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Categorizing AI Translational Bioinformatics



Phenotype to phenotype

Genomic to phenotypic: Genome-Wide Association Studies (GWAS)

Phenotype to clinotype: Phenotypic data are derived from clinomic observations.

Genomic to genomic: Associations between genes

Clinotype to clinotype: Correlation among laboratory tests

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Gene Expression

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Gene Expression

- Measures actual expression of genes, and not just presence in DNA
- Cells are different because of differential gene expression (i.e. in different body organs different genes are expressed).
- Initial technology was **microarrays**, which have mostly been supplanted by RNA-Seq
- Early success was in predicting prognosis in breast cancer
  - Developed into commercial test, Mammaprint
  - Other tests developed as well, e.g., Octotype DX

From Hersh et al. 2022

- Gene regulation: how a cell controls which genes are expressed.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Uses of Gene Expression

- Use of **Mammaprint** identified women with breast cancer at high clinical but low genomic risk and found could safely forego chemotherapy (Cardoso, 2016)
- In patients with advanced (Stages II-III) colon cancer, non-expression of CDX2 gene associated with better response to adjuvant chemotherapy (Dalerba, 2016)
- **Impact on expression** – in older adults, 596 genes with expression related to aging were found to have profile of aging “reversed” by resistance exercise (Melov, 2007)

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Genome-wide association studies (GWAS)

- **Aim:** to find the genetic variants associated with a specific disease or phenotype.
- Data represented by a **binary matrix** which represents whether a patient's genome has a specific variant.
- Growing collection of electronic clinical data allow increasing measure of phenotype, setting the stage for **genome-wide association studies (GWAS)** that **associate genotype and phenotype**.
- Early and well-known effort to use EHR data for GWAS came **from Electronic Medical Records and Genomics (eMERGE) Project**: https://emerge-network.org/
- Leading to new methods of diagnosis and treatment.

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Other Applications of Genomics

- Polygenic risk scores – quantifying clinical risk based on scores for presence of genes and variants, e.g.,
  - Prediction of coronary artery disease more accurate than traditional risk factors (Inouye, 2018)
  - Risk of breast cancer and subtypes (Mavaddat, 2019)
  - Role in clinical care, including actionability, to be determined (Hunter, 2019)
  - To predict from EHR, need precise and accurate data (Li, 2020)
  - Polygenic Score Catalog: https://www.pgscatalog.org/

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Annotating the Function of the Genome

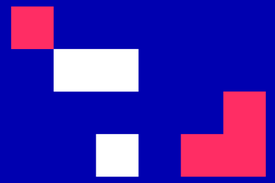
- Most widely used is Gene Ontology (GO; GO Consortium): http://geneontology.org/
- Gene function is classified by:
  - Molecular function – what gene product does, e.g., adenylate cyclase enzyme
  - Biological process – biological objective of gene, e.g., pyrimidine metabolism
  - Cellular component – where function and process take place, e.g., rough endoplasmic reticulum

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

BIOMARKERS



This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Biomarkers

- The use of domain knowledge to validate the findings requires features extraction.
- Model-explicit features that differentiate the sample classes.
- These features are then forwarded to pathway, gene set, and gene ontology analysis to reveal which biological mechanisms are involved.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Types of Biomarkers

- Indication of an underlying physiological state.
- **Early biomarkers**: fever, increased respiratory rate, or a rash.
- **Multi-dimensional biomarkers** - may consist of not just one but many different characteristics, which together give insight into underlying states or processes.
  - Analysis of gene expression
- **Predictive biomarkers** can facilitate decision making.
  - The presence of a biomarker indicating poor prognosis suggest a different treatment selection
  - Lifestyle changes i.e. gene sequence might indicating that weight loss is likely to improve insulin resistance

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Biomarkers Discovery

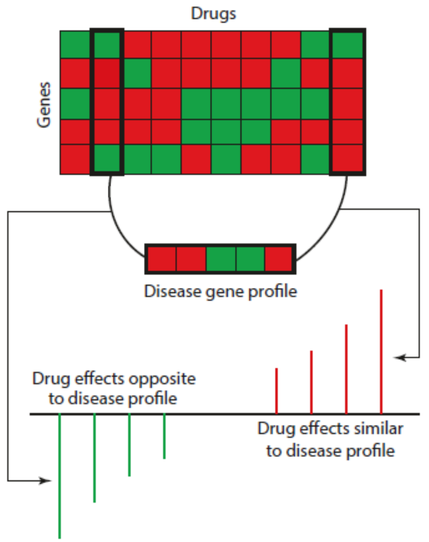
- To categorize samples or patients:
  - Cancerous samples versus normal tissues, good versus poor prognosis, bacterial versus viral infection.
  - Supervised learning
- To predict the likelihood of a heart attack based on gene expression.
- In clinical trials, identifying biomarkers to predict progression reduce the total sample and consequently the cost of trial

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Biomarkers for Drug Repurposing

- Identifying existing drugs that may be useful for indications other than those for which they were initially approved.
- Overlapping symptoms may suggest a potential match between one disease area and another.
  - ➢ **Biomarker discovery**, possible to use underlying molecular pathway signatures to suggest new uses for existing drugs.
  - ➢ Comparison of **gene expression signature patterns** observed in different diseases.
- Combination of pattern mining and sequence comparison/analysis algorithms can be used.



A computational approach to candidate selection for drug repurposing. Sirota et al. 2011, first generated genomic signatures representing both diseases and drug exposure. For each disease signature, they compared it to the panel of drug signatures and assigned a drug-disease score based on profile similarity. Drugs whose pattern were most significantly *dis*similar to the disease state were ranked as lead candidates to treat the disease of interest. Shortliffe et al., 2021.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

AI in Bioinformatics

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Applications Tasks of AI in Bioinformatics

- Medical diagnosis (or personalized diagnosis)
  - Detect genetic markers from a patient that are associated with a disease
- Treatment selection
  - Detect genetic information that affect a patient's response to a drug
- Prognosis of a disease
  - Using gene expression analysis
- Identify effects of known and potential treatments
  - Use of microarray expression data - effects related to specific molecules

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Applications of AI Techniques in Bioinformatics

- **Association studies:** mining for novel relationships among different biomedical entities
- **Clustering:** dividing patients and samples into different groups.
- **Modeling and knowledge representation:** mathematically representing the associations and cause-effect relations among different biomedical entities
- **Simulation:** mathematically representing the changes observed in biomedical subjects by a system of dynamic equations
- **Spatial visualization:** visualizing biomedical datapoints in 2D or 3D space.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Properties of Data in Bioinformatics

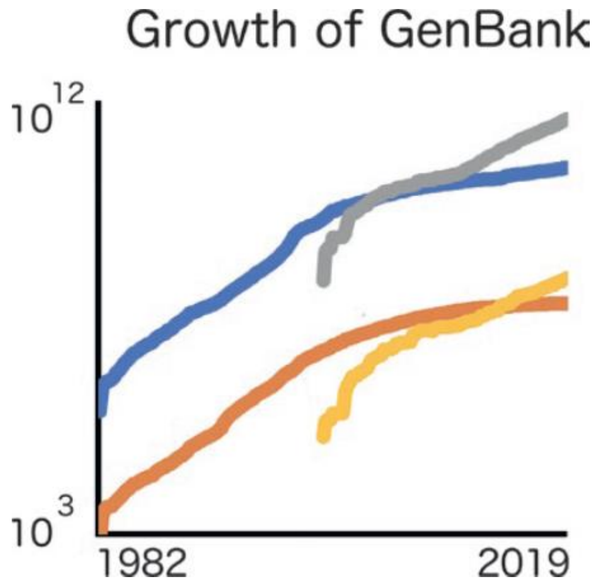
- **Measure:** refers to the type of molecular entities that are collected, counted, or observed.
  - Microarray
  - Copy-number variation and mutation
  - RNA sequencing and transcript count
  - Protein binding affinity
- **Resolution:** refers to whether the measures are collected from the tissue (bulk) sample, which is a collection of cells, at the single-cell level, or at the sub-cellular molecular level.
- The results from analyzing -omic data by researchers from multiple fields are carefully curated and organized into annotated catalogs.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bioinformatics Databases

- **GenBank** (more than 211 million sequences): Genetic sequence database from NCBI
- **Protein Data Bank (PDB)** (three-dimensional structural data for over 45,538 distinct protein structures)
- **EMBL-EBI:** Nucleotide Sequence Database
- **UniProt:** Protein sequence database
- **GEO Database:** Gene expression profiles from NCBI (over 2.8 million arrayed samples)
- **Expression Atlas:** Gene expression across species and biological conditions



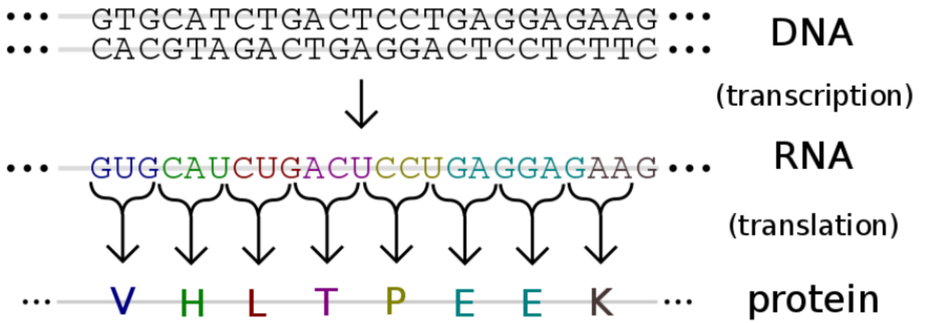
Source: Shortliffe et al., 2021

Co-financed by the European Union
Connecting Europe Facility

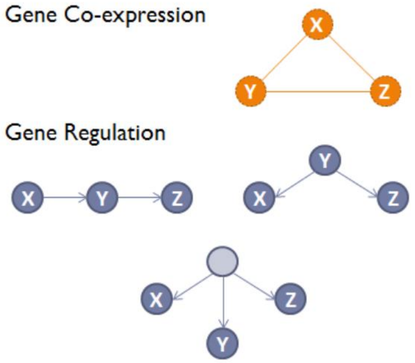
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Bioinformatics Data Types for AI Systems

- **Gene sequences:**
  - Text and characters – values
  - Nucleic acid sequences – ACGT – namely, Adenine, Cytosine, Guanine and Thymin



- **Gene expressions:**
  - Numerical values
  - Represent the expression of a gene at a certain time point



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Data Processing

- Format data in a matrix
- Dimensionality reduction
- Data scaling and normalization
- Batch effect correction is applied when the same type of dataset is generated at different rounds of experiments
- Embedded data visualization

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Preprocessing Pipelines to build the Data Matrix

Name	Data Type	Input Data	Output Data
RAMPAGE (Encode project)	Bulk RNA sequencing	Fastq file: data and reference genome	– gene alignment – gene quantification (matrix)
miRNA-seq (Encode project)	microRNA sequencing	Fastq files	– miRNA quantification
CellRanger	Single-cell RNA sequencing	Fastq files	– gene-cell expression matrix – simple analysis of single-cell data
Stanford CoreNLP	Medical text	Medical text collection	– medical term quantification (matrix)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Machine Learning (ML) in Bioinformatics

- Protein structure prediction
- Optimization methods
- Analysis of **microarray data (matrix)**
  - Expression pattern identification
  - Classification
    - Probabilistic graphical models
    - Genetic algorithms

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Supervised Learning

- Gene identification/prediction
  - Bayesian networks
  - Support Vector Machines (SVM)
  - Logistic regression
- Gene expression analysis
  - Random forest
  - Support Vector Machines (SVM)
  - Neural networks

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Supervised Learning

- DNA sequence comparison
  - Comparison of between different genomes (sequence alignment)
  - Optimization methods
  - Dynamic Programming Algorithms (DPA)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Classification/Prediction of Gene Expression Data

- **Features (Columns):** No of genes and their value
- **Samples (Rows):** No of individuals
- **Methods:**
  - k-nearest-neighbor (KNN) - “curse of dimensionality”
    - Dimensionality reduction methods: feature selection or feature extraction
  - Function approximation: Bayesian modeling, logistic regression and Support Vector Machines
  - Rule-based: decision trees, random forests, or covering rules
  - Ensemble methods: bagged and boosted decision trees

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Clustering of Gene Expression Data

- To determine which genes are being expressed similarly.
- Genes that are associated with similar expression profiles are often **functionally associated.**
  - Identification of genes associated with neoplastic progression
  - Distance metric must be determined to compare a gene's profile with another gene's profile i.e. euclidean distance, correlation distances.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Unsupervised Learning

- Build of the genes based on their expression profiles
- Identifying new disease subtypes from genomic data
- Drug repurposing
- Identify and characterize new cell subtypes in single-cell omics data
- **Methods:**
  - ➢ Clustering
  - ➢ Expectation-maximization methods
  - ➢ Non-negative matrix factorization

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Gene-expression in Breast Cancer

- Breast cancer [van ‘t veer et al. Nature 2002]
  - Predicting outcome of breast cancer patients using gene expression produced by microarray technology.
  - 70 gene signature
- Validated in follow-up paper
  - Signature of 70 genes to predict breast cancer prognosis.
  - Some issues with validation due to overlap of training & testing.

Co-financed by the European Union
Connecting Europe Facility

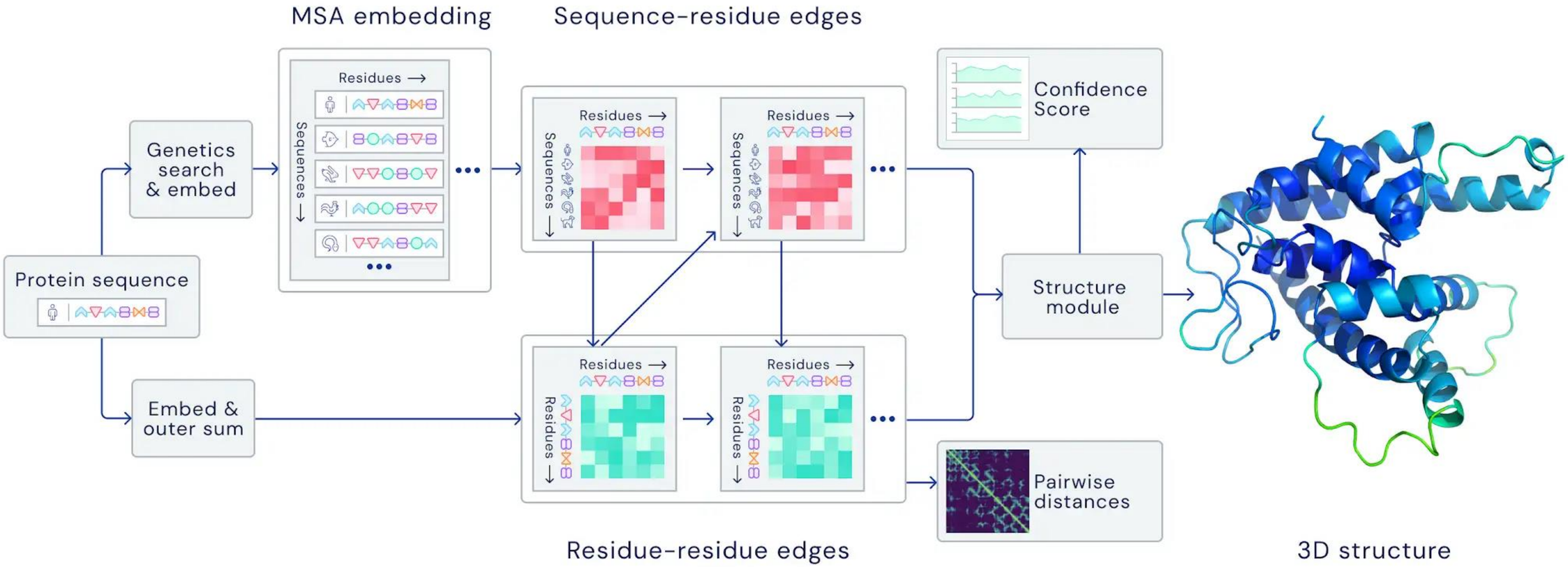
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Deep Learning in Bioinformatics

- Predicting the structure of the associated molecules of a newly determined DNA sequence (Protein structure)
  - the use of convolutional neural networks
  - by DeepMind Inc. called AlphaFold (Senior, et al., 2020)
  - Protein folding problem
  - DeepBind using CNN (Alibanahi et al., 2015)
  - Recurrent Neural Networks (RNN) – long-short-term memory - DeeperBind
- Gene regulation
  - CNN

Co-financed by the European Union
Connecting Europe Facility

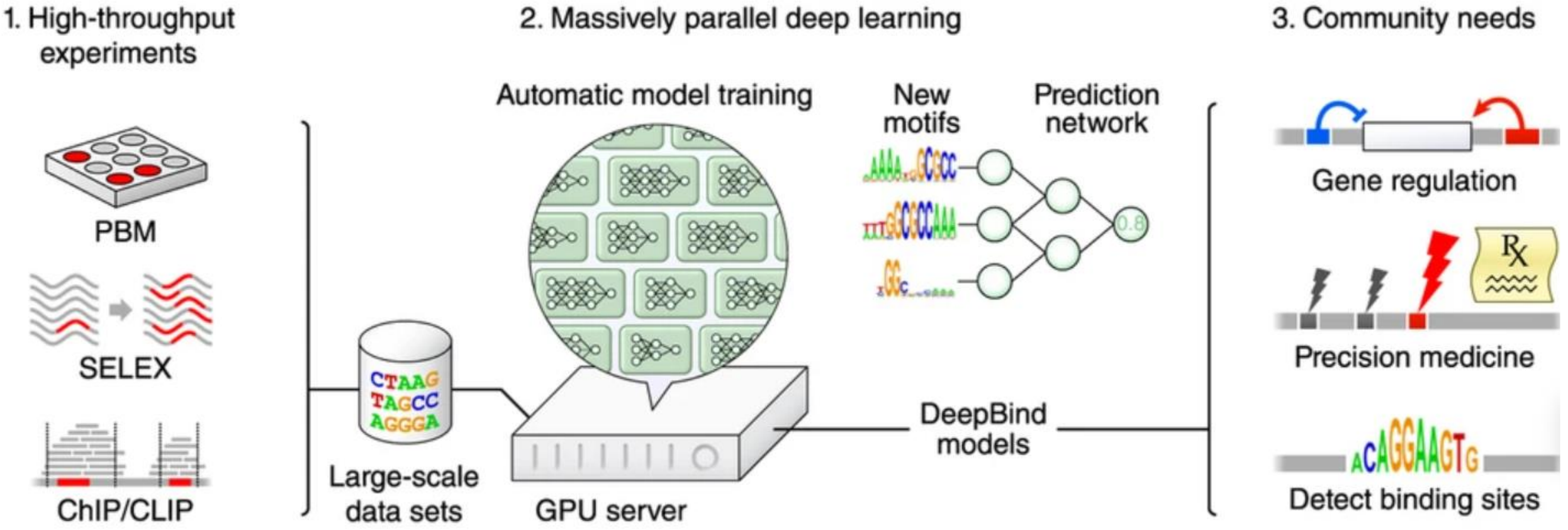
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423



AlphaFold: An overview of the main neural network model architecture. The model operates over evolutionarily related protein sequences as well as amino acid residue pairs, iteratively passing information between both representations to generate a structure. Source: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology .

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423



1. The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including PBM, SELEX, and ChIP- and CLIP-seq techniques. 2. DeepBind captures these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. 3. The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations. **Source:** Alipanahi et al., Nature Biotechnology, 2015.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Probabilistic Graphical Models in Bioinformatics

- Modelling of DNA sequences
  - Gene finding process – Hidden Markov models
- Pattern recognition in microarray data
- Inference of genetic networks
  - Handle uncertainty and noise using probability theory
  - Structure – interrelations between genes
  - Conditional probabilities—encoding the strength of the interdependencies between the variables
  - Inference – reasoning using exact or approximate algorithms
- The models are biologically interpretable given the observational data

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Precision Medicine and Genetic Disorders

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Precision Medicine

- Diagnosis and treatment targeted to individual patients on the basis of genetic, biomarker, phenotypic, or psycho-social characteristics that distinguish them from other patients with similar clinical presentations.
- Requires new “taxonomy” of medicine (IOM, 2011)
- Growing amount of patient and consumer data in electronic and personal health records (the “phenome”).
- Potential to reduce unnecessary overtreatment, especially in cancer.

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

From Clinical Genetics and Genomics to Precision Medicine

- Gene expression
- Genome-wide association studies
- Monogenic disorders
- Complex genetic disorders
- Pharmacogenomics

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Genetic Disorders

- Involvement of one or more genes in human disease
- Many new technologies for determining **genome (genotype),** combined with growing amounts of **clinical data (phenotype)**, provide opportunity to uncover genetic causes of disease and lead **to precision medicine**.
- Catalog of human genetic disorders (On-line Mendelian Inheritance in Man, OMIM): https://www.omim.org/
- Over 25,000 entries from over 16,000 gene loci

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Monogenetic Disorders

- If the DNA that codes for a protein contains an error, or mutation, it might result in an altered structure and, as a result, altered function
- Caused by variation in a single gene
- Examples:
  - Normal vs. sickle cell hemoglobin
  - Huntington's Chorea (Huntington's Disease)

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Sickle Cell Anemia

- **Homozygotes (Hgb SS)** have “sickling” of blood cells causes painful crises.
- **Heterozygotes (Hgb AS)** have normal blood cells and increased resistance to malaria, which is one reason why gene persists in population.

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Huntington's Chorea

- Progressive neurological disorder characterized by facial and/or extremity twitching as well as behavioral disorders
  - Onset usually after age 35 (i.e., after reproductive years)
- Cause of disease is an **abnormal protein (huntingtin)** that accumulates in brain
  - Autosomal dominant gene with incomplete penetrance
- Risk of disease (penetrance) associated with **number of repeats in gene for huntingtin protein** on chromosome 4
  - <10-15 repeats is "normal"
  - >35 leads to symptoms of disease likely to present

From Hersh et al. 2022

- Can be detected by test, but ethical dilemmas over if and when to have test

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Complex Genetic (polygenic) Disorders

- In many disorders, multiple genes causes disorder
- In other disorders, there is pleiotropy, i.e. one gene may cause >1 disease
- In both single- and multi-gene disorders, they are may be locus heterogeneity, i.e., **variable expression** within the trait
- There may also be epistasis, which is the interaction between the genes at two or more loci, such that the **phenotype differs from what would be expected if the loci were expressed independently**

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Examples of Complex Genetic Disorders

- BRCA1/BRCA2
  - Increased risk in multiple types of cancer but related to family history
- Color blindness
  - Recall that Y chromosome determines gender
    - Females are XX and males are XY
  - Most genes related to color vision reside on X chromosome
    - There are many genes, which influence vision of one or more colors
  - Males are more likely to have color blindness (7% of men in US) because they only have one X chromosome

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Pharmacogenomics

- Study of the genetic basis for variation in drug response
- Genes influence
  - Drug concentrations – how rapidly they are metabolized
  - Drug target – how individuals respond
- Genetic variation plays a significant role in variable response to drug therapy
  - Every clinician knows that patients respond differently to medications, sometimes along ethnic lines, which implies genetic relationship

From Hersh et al. 2022

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Security and Privacy Issues in Bioinformatics

- **Population Bias:** Race demographics and ethnicities are under-sampled
- **Data Ownership:** Consent from study participants, which allows using their data for research must be obtained.
- **Privacy Protection:** Identifiable information must be removed

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

SUMMARY

- Bioinformatics is the field of computational science applied to biologic data in order to detect patterns, find similarities/differences between omics sequences which can diagnose or predict a disease or even to predict the effects of an intervention.
- Translational bioinformatics is applied to human health using the identification/prediction of biomarkers combining with other biomedical data and phenotype.
- Phenotype is the identification of an abnormality/ a disease
- Common uses of bioinformatics are for gene identification, structure prediction, gene expression analysis, detect similarities/differences between DNA sequences.
- AI has been applied in bioinformatics using both supervised and unsupervised learning
- Bioinformatics data are often represented in a binary matrix.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

SUMMARY

- ML such as logistic regression, ensemble trees, random forest, SVM, neural networks and probabilistic graphical models are widely applied in bioinformatics.
- Deep learning techniques such as CNN and RNN are also widely used in bioinformatics especially due to the complexity of the data.
- Biomarkers are important for precision medicine
- Association studies are also used to detect association patterns between a set of genes and a disease.
- Genetic disorders are divided into simple and complex based on the number of genes that are associated to the human disease.

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Discussion

- Which do you think are the major challenges of applying AI on Translational Bioinformatics?
- What is the difference between precision medicine and public health?

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

References

- Jaskaran Kaur, Understanding Bioinformatics As A Beginner In Data Science, Opinions, December 10, 2020, https://analyticsindiamag.com/understanding-bioinformatics-as-a-beginner-in-data-science/
- Alipanahi, B., Delong, A., Weirauch, M. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33, 831–838 (2015). https://doi.org/10.1038/nbt.3300
- Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
- Seonwoo Min, Byunghan Lee, Sungroh Yoon, Deep learning in bioinformatics, Briefings in Bioinformatics, Volume 18, Issue 5, September 2017, Pages 851–869, https://doi.org/10.1093/bib/bbw068
- Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." Journal of computational biology 7.3-4 (2000): 601-620.
- Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, Victor Robles, Machine learning in bioinformatics, Briefings in Bioinformatics, Volume 7, Issue 1, March 2006, Pages 86–112, https://doi.org/10.1093/bib/bbk007
- Shortliffe, Edward H., and James J. Cimino, eds. Biomedical informatics: Computer applications in health care and biomedicine. Springer Science & Business Media, 2021.
- Hersh, WR, 2022. Health Informatics: Practical Guide, 8th Edition

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423