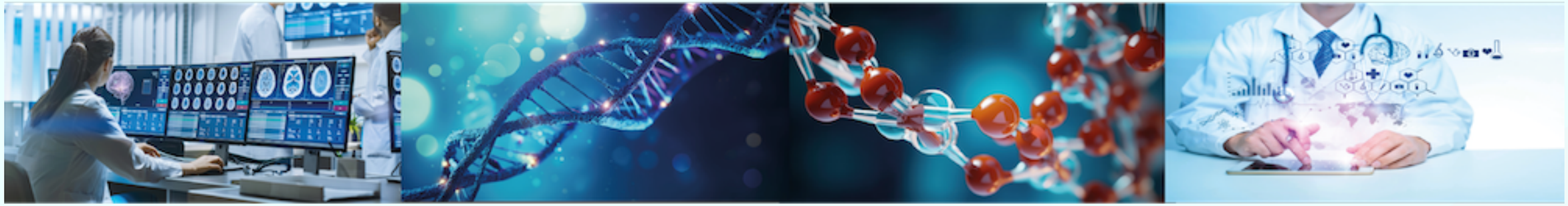


Programming for Biomedical Informatics

Revision Session 1

Lecture 18 - “Course Lecture Review & Short Answer Questions”

https://github.com/tisimpson/pbi

Ian Simpson
ian.simpson@ed.ac.uk

THE UNIVERSITY of EDINBURGH
informatics

Exam Information

Monday 9th December 13:00 - 15:00

(Patersons Land - G.43)

Structure

Question 1 (25 marks)

and choice of

Question 2 (25 marks)

or

Question 3 (25 marks)

Question 1

- short answer questions worth 1 or 2 marks covering material from whole course

Questions 2 &3

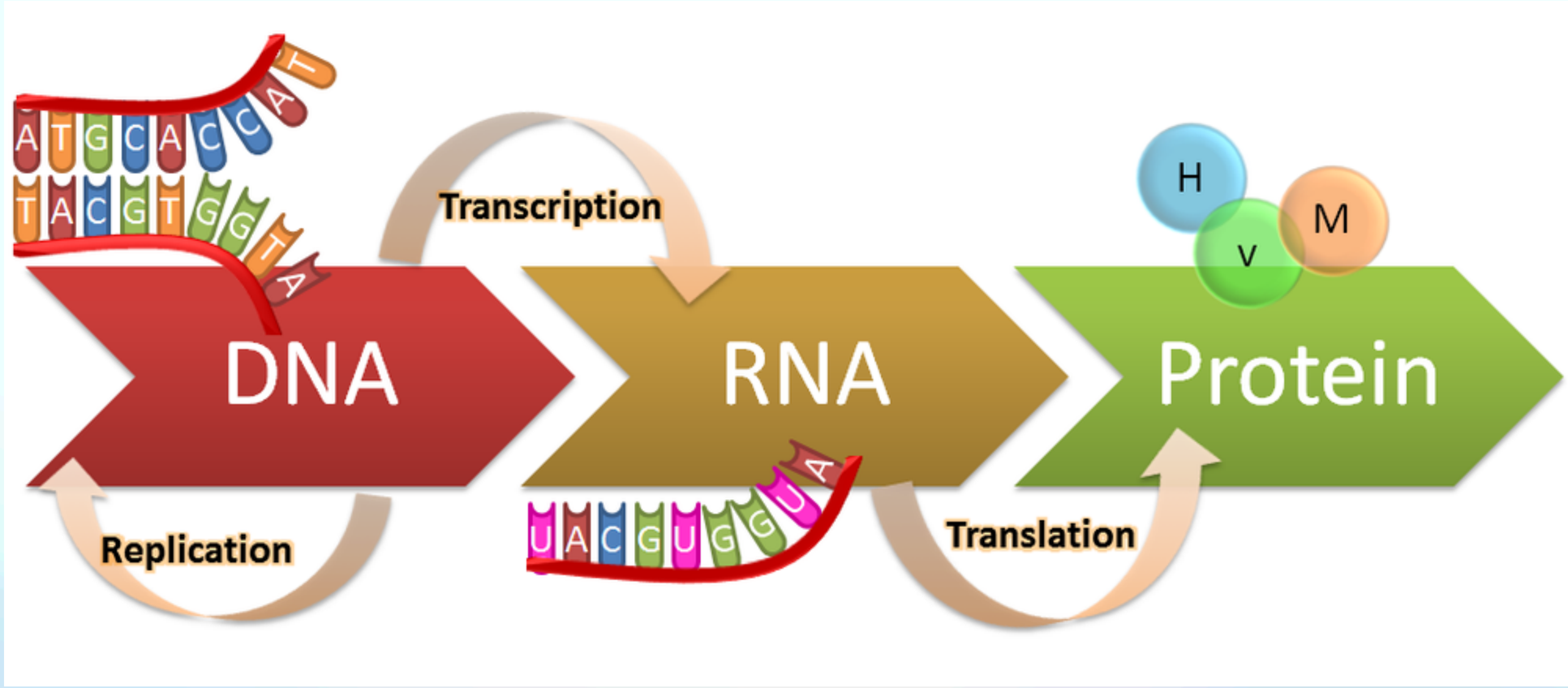
- example based (a research question/ scenario)
- small number of short questions worth a total of 5 marks related to the scenario
- 2x pseudo-coding questions worth 10 marks each

Overview of Lecture Materials

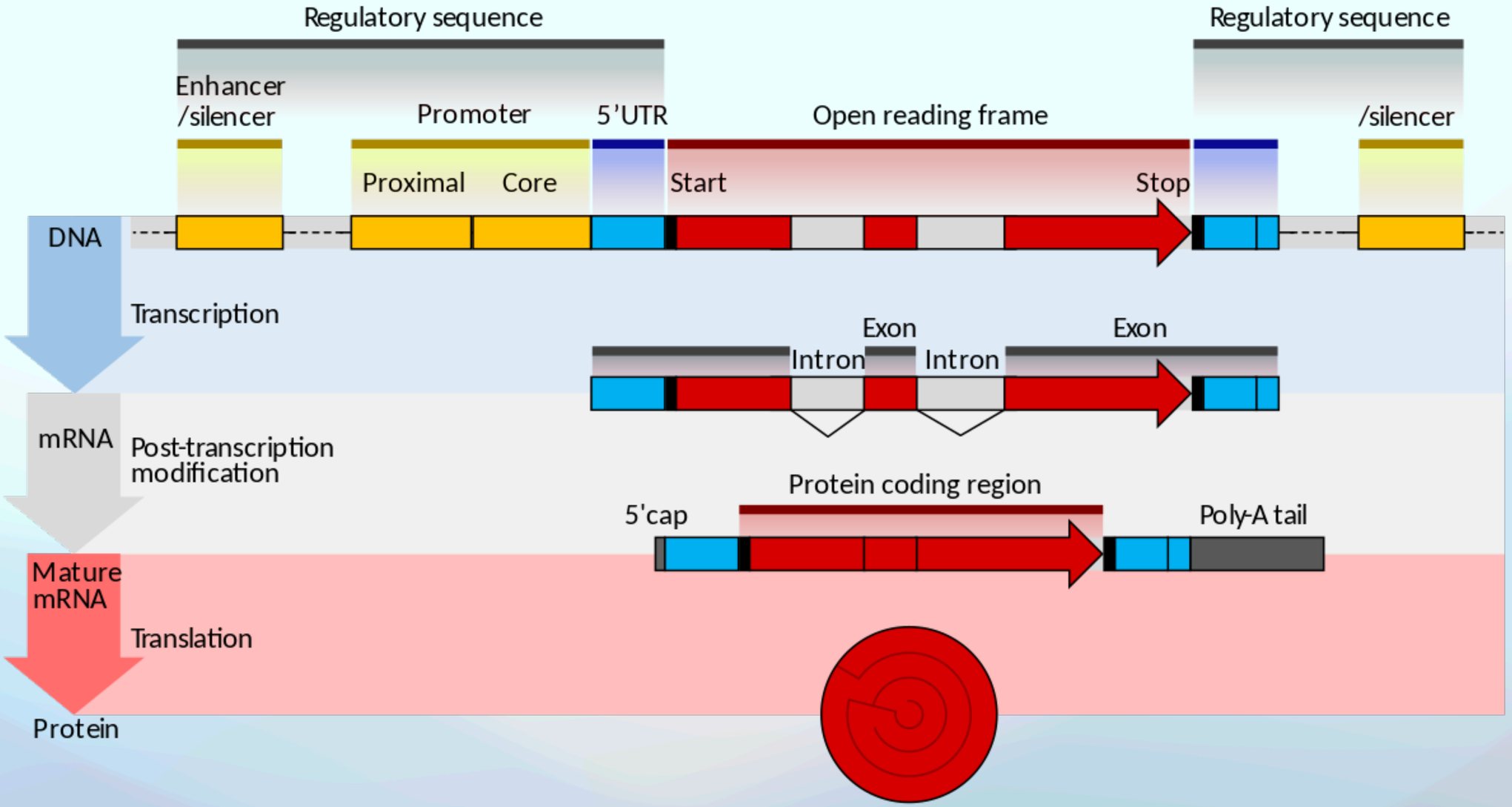
Lectures

- Week 1 - Welcome & Getting Started (lecture 1) & Extra Slides (lecture 1 extra), Working with Notebooks & Git (lecture 2)
- Week 2 - Introduction to the Biomedical Dataverse (lecture 3) and Bulk FTP Download & APIs (lecture 4)
- Week 3 - Mapping & Harmonisation (lecture 5) and Data Integration & Summary Analysis (lecture 6)
- Week 4 - Biomedical Evidence (lecture 7) and Mining & Analysing the Biomedical Literature (lecture 8)
- Week 5 - October Break - No lectures or assignment this week
- Week 6 - Measuring Gene Expression (lecture 9) and Differential Gene Expression Analysis (lecture 10)
- Week 7 - Biological Networks (lecture 11) and Network Construction Techniques (lecture 12)
- Week 8 - Essential Network Methods (lecture 13) and Network Analysis in Practice (lecture 14)
- Week 9 - Structuring Biomedical Data with Ontologies (lecture 15) and Working with Ontologies (lecture 16)
- Week 10 - Working with Multiple Data Modalities (lecture 17) and Revision Session 1 (lecture 18)
- Week 11 - Revision Session 2 (lecture 19)

The “Central Dogma”



The “Central Dogma”



Programmatic Bulk Download

We can use a number of Python approaches to achieve bulk download programmatically rather than having to deal with manual downloading and saving of data:

URL Fetching Libraries

- `requests` - `requests.get(url).content` followed by saving the content to file.
- `urllib` - `urllib.request.urlretrieve(url, filename)` this downloads to file.
- `wget` (wrapper for GNU wget) - `wget.download(url)` this downloads to file.

Threading of multiple downloads simultaneously

- `Scrapy` - large-scale web crawling and bulk data extraction, efficient for downloading data from multiple URLs by defining spiders.
- `pycurl` - curl interface. good for handling multiple simultaneous HTTP/FTP transfers.

Handling Authentication & Session Management

- `requests.Session` to persist certain parameters across requests, can handle login and cookie management for downloading from authenticated sources.
- Selenium for Web Scraping, especially for websites that load with JavaScript, can simulate a browser to fetch and download files and interact with web elements.

Application Programming Interfaces (APIs)

Abstraction - APIs provide a simple interface to more complex underlying systems. They abstract the internal workings exposing only what is necessary for the user.

Standardisation - APIs are designed with standard protocols and rules, ensuring consistency

Security - APIs enforce security, can limit access by requiring authentication, and ensure data integrity and confidentiality through encryption.

Scalability - APIs allow systems to handle increased volumes of data or interactions.

Efficiency - Data exchange through APIs can be optimised to reduce the amount of data that needs to be sent.

Documentation - APIs commonly come with comprehensive documentation that details available endpoints, data formats, and methods.

Versioning - APIs often employ version control, allowing them to introduce new features or deprecate old ones without disrupting existing code developed using them.

Customisation - APIs commonly allow the user to specify parameters associated with their query to tailor data retrieval to they specific needs.

The NCBI-NLM eUtilities

- **EInfo (database statistics)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi
  - Provides the number of records indexed in each field of a given database, the date of the last update of the database, and the available links from the database to other Entrez databases.
- **ESearch (text searches)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi
  - Responds to a text query with the list of matching UIDs in a given database (for later use in ESummary, EFetch or ELink), along with the term translations of the query.
- **EPost (UID uploads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/epost.fcgi
  - Accepts a list of UIDs from a given database, stores the set on the History Server, and responds with a query key and web environment for the uploaded dataset.
- **ESummary (document summary downloads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi
  - Responds to a list of UIDs from a given database with the corresponding document summaries.
- **EFetch (data record downloads)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi
  - Responds to a list of UIDs in a given database with the corresponding data records in a specified format.
- **ELink (Entrez links)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi
  - Responds to a list of UIDs in a given database with either a list of related UIDs (and relevancy scores) in the same database or a list of linked UIDs in another Entrez database; checks for the existence of a specified link from a list of one or more UIDs; creates a hyperlink to the primary LinkOut provider for a specific UID and database, or lists LinkOut URLs and attributes for multiple UIDs.
- **EGQuery (global query)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi
  - Responds to a text query with the number of records matching the query in each Entrez database.
- **ESpell (spelling suggestions)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/espell.fcgi
  - Retrieves spelling suggestions for a text query in a given database.
- **ECitMatch (batch citation searching in PubMed)** - eutils.ncbi.nlm.nih.gov/entrez/eutils/ecitmatch.cgi
  - Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.

eUtils Mapping Review

- We used a selection of combinations of eUtils tools to recover information that allows for mapping between sources
- eSearch - can take mixed queries, typically terms or phrases e.g. gene symbols “Pax6”
- eSummary - can take identifiers returned by eSearch to recover basic summary information. This is often what you need as it includes internal numerical NCBI identifiers and basic meta-data (descriptions, names...)
- eLink - can be used to map between NCBI databases e.g. from “nucleotide” to “gene”. It returns a LinkSet in which individual Link items contain the cross-mapped information (NM_12345 -> GeneID:678)
- eFetch - can be used to pull full entries. These are typically extremely long and complex records with deeply nested elements and/or long lists of features - there are options to filter these

eUtils Mapping Review

- We can use the BioPython Entrez module for more accessible use of eUtils and it can be extended to more complex implementations
- Once familiar with the basic function of the eUtils it is more flexible to begin using the requests library to make API calls using URLs that you have built
- This makes it easier to take advantage of eUtils features like “WebEnv & QueryKey” and to serialise large requests that are much more difficult to retrieve efficiently through other means
- Key to the requests approach is an understanding of how you parse XML/JSON response texts in Python to access the features you are looking for.
- Typically this requires that you first print out an XML response for a small query so that you can correctly format your XML element search.
- Once this has been perfected you can readily scale up your analysis efficiently.

eLink “link-names”

Entrez Link Descriptions:									
PubMed	Protein	Nucleotide	Structure	Genome	Books	CancerChromosomes	Conserved Domains	3D Domains	Gene
Genome Project	GENSAT	GEO Profiles	GEO DataSets	HomoloGene	MeSH	NCBI Site Search	OMIA	OMIM	
PMC	PopSet	Probe	PubChem BioAssay	PubChem Compound	PubChem Substance	SNP	Taxonomy	UniGene	UniSTS

Link name (used in Entrez URLs)	Label in "Links" menu	Label in "Display" menu	
assembly_assembly_diploid	Linked assembly from diploid	Linked assembly from diploid	Genome assembly that is the other (ps
assembly_bioproject	BioProject	BioProject	BioProject
assembly_bioproject	BioProject	BioProject	BioProject
assembly_bioproject	BioProject		
assembly_biosample	BioSample		
assembly_biosample	BioSample		
assembly_biosample	BioSample		
assembly_genome	Genome		
assembly_genome	Genome		
assembly_genome	Genome		
assembly_nuccore	Nucleotide		
assembly_nuccore_insdc	Nucleotide INSDC		
assembly_nuccore_refseq	Nucleotide RefSeq		
assembly_nuccore_wgsmaster	WGS Master		
assembly_nuccore_wgscontig	WGS contigs		
assembly_nuccore_insdc	Nucleotide INSDC		
assembly_nuccore_refseq	Nucleotide RefSeq		
assembly_nuccore_wgsmaster	WGS Master		
assembly_nuccore_insdc	Nucleotide INSDC		
assembly_nuccore_refseq	Nucleotide RefSeq		
assembly_nuccore_wgsmaster	WGS Master		
assembly_pubmed	PubMed		
assembly_pubmed	PubMed		
assembly_pubmed	PubMed		
assembly_sra	Sra		
assembly_sra	Sra		
assembly_taxonomy	Taxonomy		
assembly_taxonomy	Taxonomy		
assembly_taxonomy	Taxonomy		
biocollections_biosample	BioSample		
biocollections_nuccore	Nucleotide		
biocollections_nuccore	Nucleotide		
biocollections_nucgss	GSS		
biocollections_nucgss	GSS		
biocollections_popset	PopSet		
biocollections_popset	PopSet		
biocollections_protein	Protein		
biocollections_protein	Protein		
bioconcepts_bioconcepts_child	children		
bioconcepts_bioconcepts_parent	parent		
bioconcepts_bioconcepts_sibling	sibling		
bioconcepts_bioconcepts_child	children		
bioconcepts_bioconcepts_parent	parent		
bioconcepts_bioconcepts_sibling	sibling		
bioconcepts_bioconcepts_child	children		
bioconcepts_bioconcepts_parent	parent		
bioconcepts_bioconcepts_sibling	sibling		

pubmed_assembly	Assembly	Assembly	Assembly
pubmed_assembly	Assembly	Assembly	Assembly
pubmed_assembly	Assembly	Assembly	Assembly
pubmed_bioproject	Related Project	Project Links	Related Projects
pubmed_bioproject	Related Project	Project Links	Related Projects
pubmed_bioproject	Related Project	Project Links	Related Projects
pubmed_biosample	BioSample	BioSample Links	BioSample links
pubmed_biosample	BioSample	BioSample Links	BioSample links
pubmed_biosample	BioSample	BioSample Links	BioSample links
pubmed_biosystems	BioSystems	BioSystem Links	Pathways and biological systems (BioSystems) that cite the current articles. Citations are from the BioSystems source data
pubmed_biosystems	BioSystems	BioSystem Links	BioSystems
pubmed_biosystems	BioSystems	BioSystem Links	BioSystems
pubmed_books_refs	Cited in Books	Cited in Books	NCBI Bookshelf books that cite the current articles.
pubmed_books_refs	Cited in Books	Cited in Books	PubMed links associated with Books
pubmed_books_refs	Cited in Books	Cited in Books	PubMed links associated with Books
pubmed_cancerchromosomes	Cancer Chromosomes	CancerChrom Links	Cancer chromosome records that cite the current articles.
pubmed_cdd	Conserved Domains	Conserved Domain Links	Conserved Domain Database (CDD) records that cite the current articles. Citations are obtained from the CDD source data
pubmed_cdd	Conserved Domains	Conserved Domain Links	Link to related CDD entry
pubmed_cdd	Conserved Domains	Conserved Domain Links	Link to related CDD entry
pubmed_clinvar	ClinVar	ClinVar	Clinical variations associated with publication
pubmed_clinvar_calculated	ClinVar (calculated)	ClinVar (calculated)	Clinical variations calculated to be associated with publication
pubmed_clinvar	ClinVar	ClinVar	Clinical variations associated with publication
pubmed_clinvar_calculated	ClinVar (calculated)	ClinVar (calculated)	Clinical variations calculated to be associated with publication
pubmed_clinvar	ClinVar	ClinVar	Clinical variations associated with publication
pubmed_clinvar_calculated	ClinVar (calculated)	ClinVar (calculated)	Clinical variations calculated to be associated with publication
pubmed_dbvar	dbVar	dbVar	Link from PubMed to dbVar
pubmed_dbvar	dbVar	dbVar	Link from PubMed to dbVar
pubmed_dbvar	dbVar	dbVar	Link from PubMed to dbVar
pubmed_domains	3D Domains	3D Domain Links	Structural domains in the NCBI Structure database that are parts of the 3D structures reported in the current articles.
pubmed_epigenomics	Epigenomics	Epigenomics Links	Related Epigenomics records
pubmed_epigenomics_experiment	Epigenomics (experiments)	Epigenomics (experiments)	Links to experimental data from epigenomic studies
pubmed_epigenomics_sample	Epigenomics (samples)	Epigenomics (samples)	Links to descriptions of biological samples used in epigenomic studies
pubmed_epigenomics_study	Epigenomics (study)	Epigenomics (study)	Links to overviews of epigenomic studies
pubmed_epigenomics_experiment	Epigenomics (experiments)	Epigenomics (experiments)	Links to experimental data from epigenomic studies
pubmed_epigenomics_sample	Epigenomics (samples)	Epigenomics (samples)	Links to descriptions of biological samples used in epigenomic studies
pubmed_epigenomics_study	Epigenomics (study)	Epigenomics (study)	Links to overviews of epigenomic studies
pubmed_gap	dbGaP	dbGaP Links	Genotypes and Phenotypes (dbGaP) studies that cite the current articles.
pubmed_gap	dbGaP	dbGaP Links	Related dbGaP record
pubmed_gap	dbGaP	dbGaP Links	Related dbGaP record
pubmed_gcassembly	Assembly for publication.	Assembly for publication.	Links from Pubmed to Assembly.
pubmed_gds	GEO DataSets	GEO DataSet Links	Gene expression and molecular abundance data reported in the current articles that are also included in the curated Gene
pubmed_gds	GEO DataSets	GEO DataSet Links	Related GEO DataSets
pubmed_gds	GEO DataSets	GEO DataSet Links	Related GEO DataSets
pubmed_gene	Gene	Gene Links	Gene records that cite the current articles. Citations in Gene are added manually by NCBI or imported from outside public
pubmed_gene_rif	Gene (GeneRIF)	Gene (GeneRIF) Links	Gene records that have the current articles as Reference into Function citations (GeneRIFs). NLM staff reviewing the litera
pubmed_gene_citedinomim	Gene (OMIM)	Gene (OMIM) Links	Gene records associated with Online Mendelian Inheritance in Man (OMIM) records that cite the current articles in their ref
pubmed_gene_bookrecords	Gene (from Bookshelf)	Gene (from Bookshelf)	Gene records in this citation
pubmed_gene_pmc_nucleotide	Gene (nucleotide/PMC)	Gene (nucleotide/PMC)	Records in Gene identified from shared sequence and PMC links.
pubmed_gene	Gene	Gene Links	Link to related Genes
pubmed_gene_bookrecords	Gene (from Bookshelf)	Gene (from Bookshelf)	Gene records in this citation

DESeq2 for Differential Gene Expression Analysis

DESeq2 is a popular tool implemented in R and Python for analysing count-based data like RNA-seq. It provides methods to test for differential expression by using negative binomial generalised linear models. The core of DESeq2's approach lies in its ability to accurately estimate variance-mean dependencies in count data, and to test for differential expression based on these estimates.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8

- Data is input as genes (rows) x samples (columns) matrix, along with sample information (conditions, replicates etc.).
  - **Filtering**: Low count genes are removed and gross normalisation for library size and sequencing depth are made
  - **Size Factor Calculation**: each gene's count is divided by the geometric mean of the counts for that gene across all samples
  - **Median Ratio Method**: size factors are calculated by taking the median of these ratios for each sample and these are used to normalise across samples.

- Dispersion is then estimated.
  - **Gene-wise**: each gene's dispersion is estimated from its mean and variance
  - **Shrinkage**: A Bayesian method is used to shrink the gene-wise dispersion estimates to moderate the values. This is especially important when the number of replicates is small.
  - **Fitting Dispersion-Mean Relationship**: A fit is made of dispersion estimates against mean expression to balance out gene-wise estimates and provide stable measures for all genes.

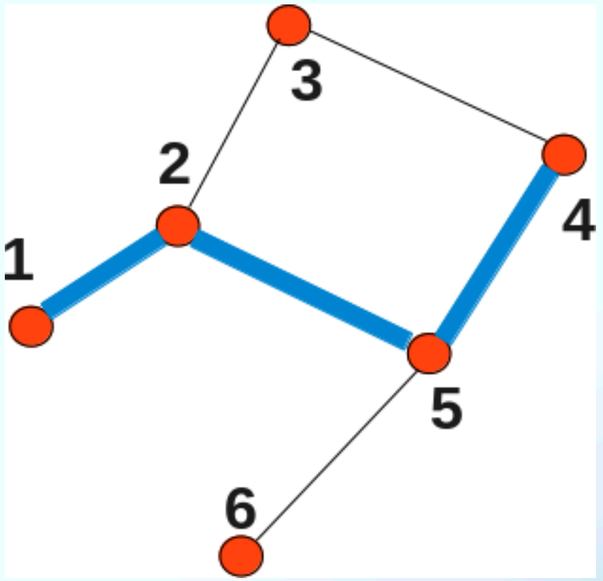
- Statistical Modelling & Hypothesis Testing
  - **Model Design**: A generalised linear model is used to incorporate information about experimental design, such as treatment conditions.
  - **Wald's Test or Likelihood Ratio Test**: DESeq2 then performs a statistical test (typically Wald's test) for each gene to determine whether it is significantly differentially expressed across conditions.
  - **Multiple Testing Correction**: Methods including Benjamini-Hochberg procedure are used to adjust p-values (often presented as an FDR - False Discovery Rate).

Network - Representations

$A_{ij} \in \mathbb{R}$ (unweighted = 0,1)

$A_{ij} = A_{ji}$ (undirected)

$A_{ii} = 2$ (if self-edge)



Adjacency Matrix

$$\mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 0 \\ 5 & 0 & 1 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$k_i = \sum_{j=1}^{n} A_{ij}$$ (degree of node i)

(edges in network)

$$\frac{1}{2}\sum_{i=1}^{n} k_i = 6$$

Degree Matrix

$$\mathbf{D} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 2 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 2 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 3 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Laplacian Matrix

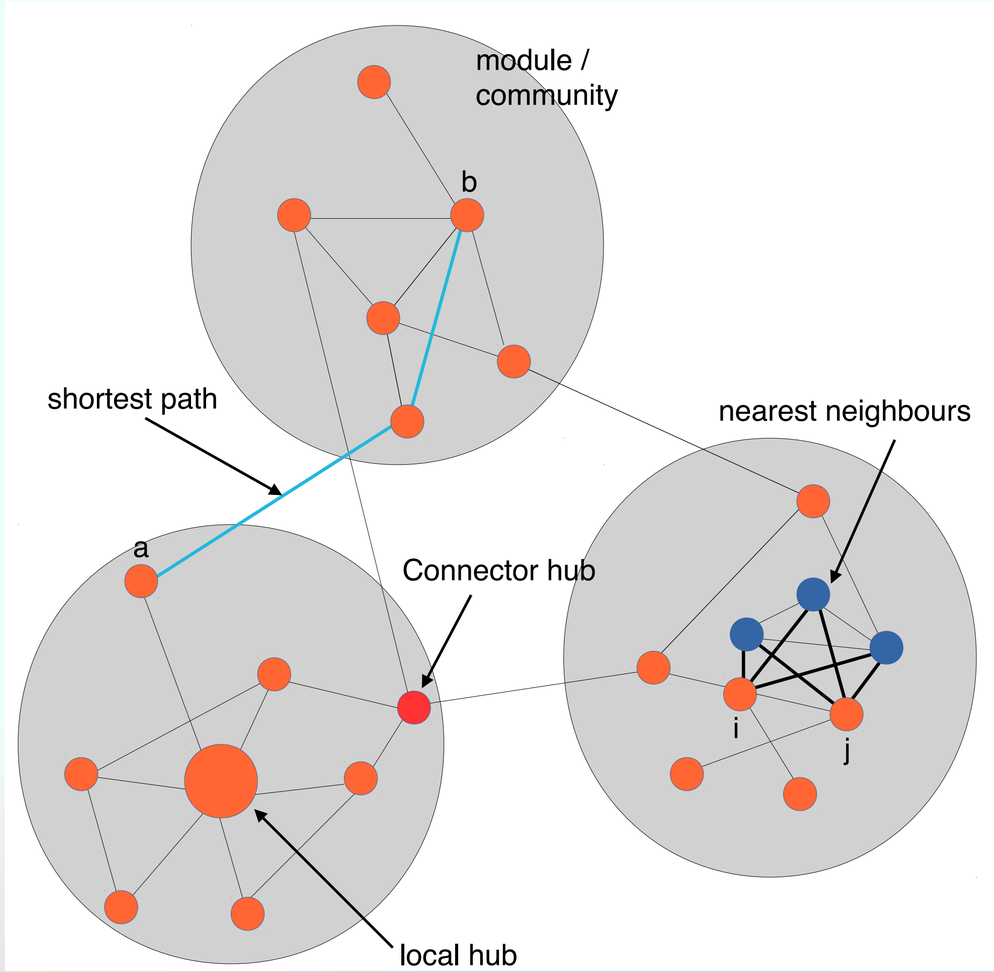
$$\mathbf{L} = \mathbf{D} - \mathbf{A} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 2 & -1 & 3 & -1 & 0 & -1 & 0 \\ 3 & 0 & -1 & 2 & -1 & 0 & 0 \\ 4 & 0 & 0 & -1 & 2 & -1 & 0 \\ 5 & 0 & -1 & 0 & -1 & 3 & -1 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Network - Measures

Measures:
- From our mathematical representation, we can define many concepts which reveal a networks underlying character.

Centrality:
- “What are the most important or central nodes (i.e. influencer, disease hub) in our network.”
- Degree, Betweenness Google’s PageRank, Semi-local, Closeness...

Similarity:
- “How similar two nodes are to each other.”
- how many shared neighbours.
- Pearson correlation, Modularity.



Clustering Networks - Community Clustering

General Approach

Graph Representation
- represent the data as a graph, where nodes represent entities and edges represent connections or relationships between them

Initialise Clusters
- start with each node as its own individual cluster or use an initial partitioning method to assign nodes to initial clusters

Calculate Modularity
- compute the initial modularity score, which quantifies the density of connections within clusters compared to connections between clusters

Optimise Modularity Locally
- iteratively merge or reassign nodes or small groups of nodes to maximise modularity
- for each iteration, assess the change in modularity resulting from moves or merges

Evaluate Modularity Gain
- accept moves or merges that improve the modularity score, and reject those that do not

Repeat Until Convergence
- continue optimising the modularity score until no further significant improvements can be made

Refine Clusters
- optionally, reapply the process at different resolutions (e.g., hierarchical clustering) to refine the clusters

Output Final Clusters
- when modularity optimisation reaches a stable maximum, finalise and output the clusters

Parameters

Resolution Parameter
- adjusts the size of the communities detected
- higher values of gamma often yield smaller localised communities, lower values yield larger, broader communities
- important for tuning clusters to the desired level of granularity

Initial Partitioning (Starting Clusters)
- determines the starting point for the modularity optimisation process
- some methods start with each node as its own cluster, while others use pre-defined partitions to initialise the process
- can affect the convergence and final clustering results

Edge Weights
- defines the weight or strength of the connections between nodes, where available
- higher weights indicate stronger connections, which can increase likelihood of nodes staying in the same community
- edge weights are particularly useful for networks with varying strengths of relationships, such as social networks

Maximum Iterations
- sets a limit on the number of iterations for the optimisation algorithm
- ensures that the algorithm does not run indefinitely and helps manage computational cost

Stopping Criterion (Convergence Threshold)
- specifies the minimum change in modularity required to continue iterations
- when the change in modularity between iterations falls below this threshold, the algorithm stops

Random Seed (for stochastic methods)
- controls the random initialisation for algorithms that have a stochastic component, such as the Louvain method
- ensuring a consistent random seed can make results reproducible

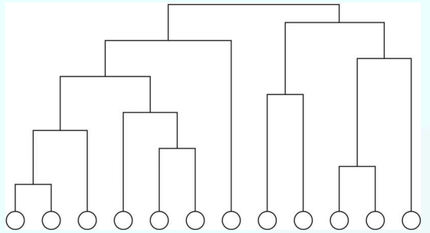
Algorithm Type
- choice of modularity-based algorithm, such as Louvain, Leiden, or Girvan-Newman
- different algorithms have unique approaches to modularity optimisation, affecting speed, accuracy, and scalability

Modularity Function Type
- specifies the formula for modularity, as there can be variations
- the choice depends on the network structure and the specific goals of the clustering

Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826

Clustering Networks - Hierarchical Clustering

agglomerative - (bottom-up)



classic hierarchical clustering

General Approach

Initialise clusters
- treat each data point as its own cluster (single cluster for each data point)

Compute initial distances
- calculate the distance between every pair of clusters
- with average linkage, the distance between clusters is the average distance between each point in one cluster and each point in the other

Merge closest clusters
- identify the two clusters with the smallest average linkage distance and merge them to form a new cluster

Update distances
- recalculate the distances between the new cluster and all remaining clusters using average linkage

Repeat merging
- continue merging the closest clusters and updating distances until only one cluster remains or until a desired number of clusters is reached

Build dendrogram
- track the order of merges to create a dendrogram, a tree-like structure that shows the hierarchical relationships between clusters

Determine clusters
- based on the dendrogram, decide on the final clustering by “cutting” the dendrogram at the desired level

Distance Metrics

Euclidean Distance
- measures the straight-line distance between two points in Euclidean space
- commonly used in continuous, numeric data

Manhattan Distance
- computes the sum of absolute differences between points
- useful when dealing with high-dimensional data

Cosine Similarity
- measures the cosine of the angle between two non-zero vectors, indicating how aligned they are
- often used in text analysis and high-dimensional spaces

Jaccard Similarity
- measures the proportion of common elements in two sets over the union of the sets.
- commonly used for binary and categorical data, especially in set-based data.

Pearson Correlation
- measures the linear relationship between two variables, ranging from -1 to 1
- useful for continuous data where the focus is on correlation rather than direct distance

Distance
- counts the number of positions at which two binary strings differ
- ideal for categorical or binary data, such as in DNA sequence analysis

Over Representation Analysis (ORA)

- consideration of foreground and background lists is crucial
  - genome ?
  - array content ?
  - mappable elements ?

contingency table	foreground	background	Row Total
genes with Term in list	a 10	b 50	60
genes with Term not in list	c 190	d 13950	14140
Column Total	200	14000	n 14200

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

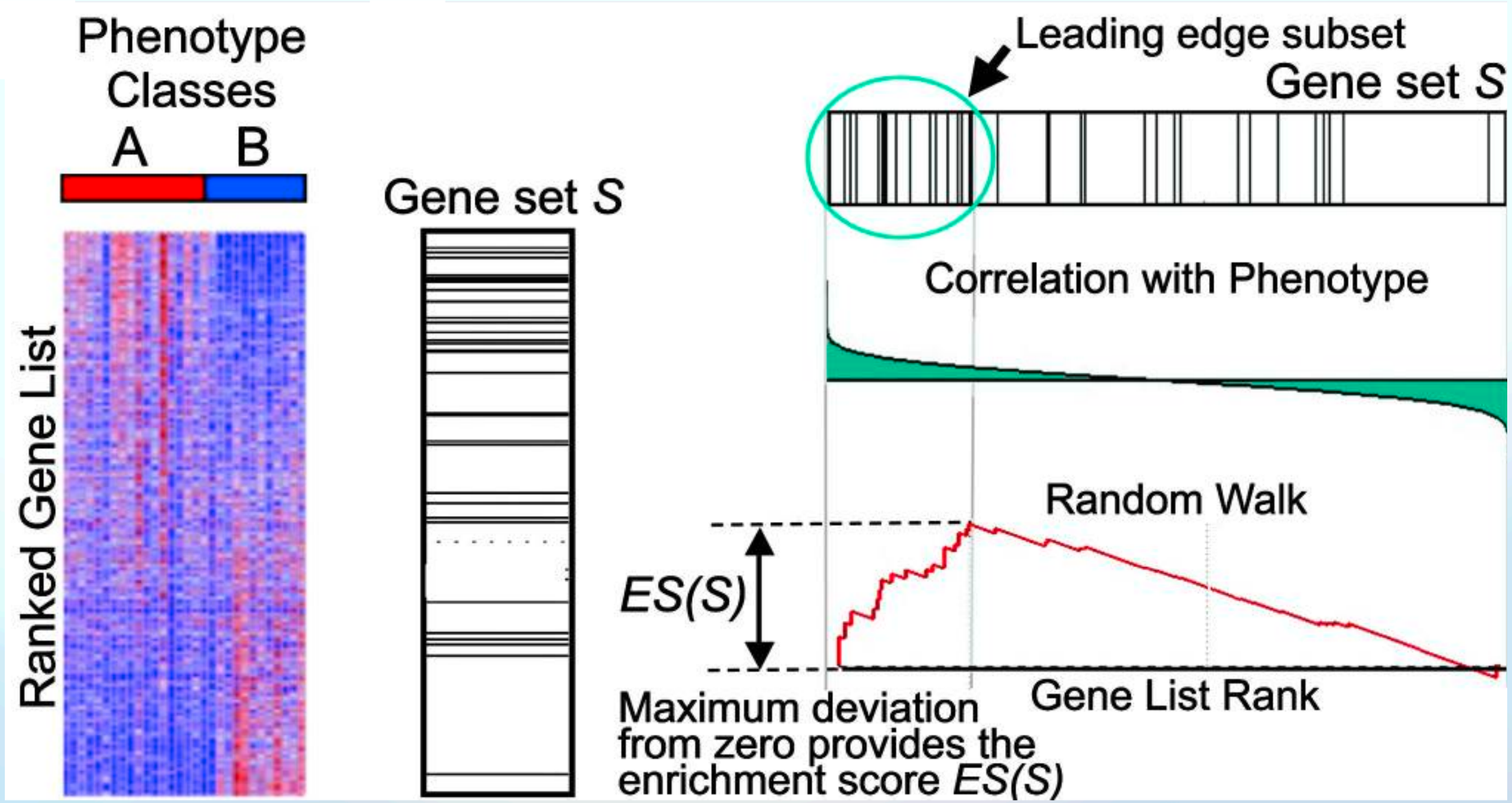
probability given by the hypergeometric distribution

in order to calculate significance we need to

- one-tailed test
  sum all probabilities as extreme or more extreme
- two-tailed test
  sum as above but also all that are as extreme or more extreme in both directions

p-value = 1.003×10^{-8} odds ratio = 14.68

Gene Set Enrichment Analysis (GSEA)



Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP.
Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

Ontologies or Terminologies

Terminologies

Controlled Vocabulary
defined list of terms that are used consistently within a specific domain, ensuring standardised naming.

Synonyms and Lexical Variants
provides different names or terms that refer to the same concept, enabling matching of varied terms to the same meaning.

Hierarchical Structure
terms may be organised in a hierarchy (e.g., disease categories) with broader and narrower terms to aid in understanding relationships.

Coding Systems
often assigned alphanumeric codes (e.g., ICD-10, SNOMED CT) for easy identification and reference.

Versioning and Updates
terminologies are periodically updated to reflect advances in medicine and changes in terminology.

Cross-Referencing with Other Terminologies
terms often have mappings to other terminologies, enabling data integration across systems (e.g., mapping ICD codes to SNOMED CT).

Ontologies

Formal Representation of Knowledge
define and formalise concepts, relationships, and rules within a domain, providing a deeper semantic structure than terminologies

Conceptual Hierarchies
often have structured hierarchies that describe relationships (e.g., "is-a," "part-of") among concepts, supporting logical inferences.

Logical Axioms and Rules
define rules and logical constraints for how concepts relate, enabling reasoning engines to deduce new information.

Interoperability
are designed for integration, allowing different systems to interpret and use the same data consistently

Linking and Reuse of Concepts
often link to other ontologies (e.g., linking human disease ontology with chemical ontology), promoting data integration across disciplines.

Defined Semantic Relationships
Unlike terminologies, ontologies provide explicit definitions of relationships between terms (e.g., "causes," "treats," "affects").

Semantic Annotation
allow for tagging or annotating data with well-defined concepts, facilitating searches and analysis based on these annotations.

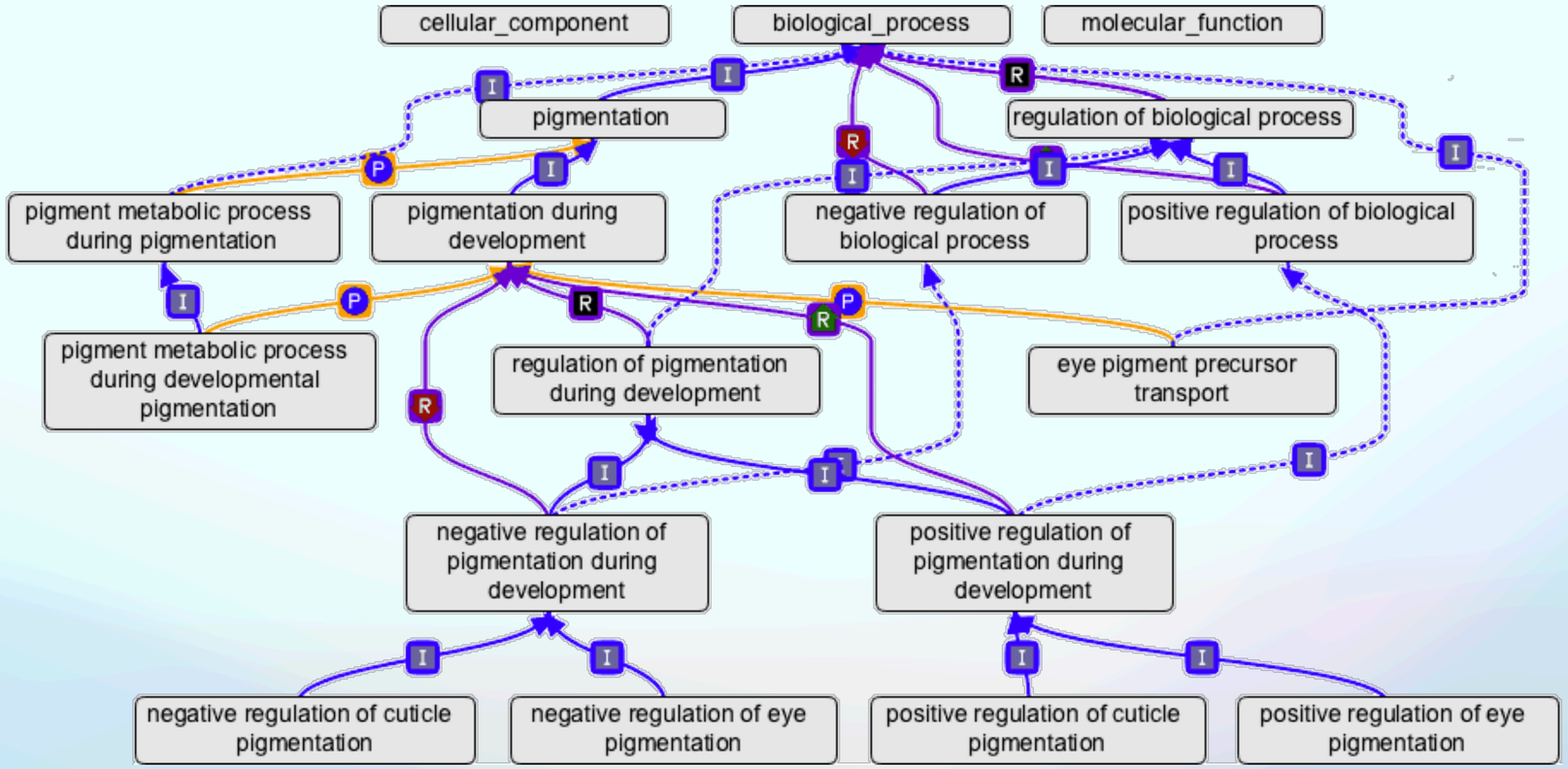
Inference Capabilities
support automated reasoning, enabling the system to infer new knowledge by applying rules to the defined relationships.

Structure of an Ontology

- nodes are “Terms”
- edges are “Relations”

low specificity roots

parent

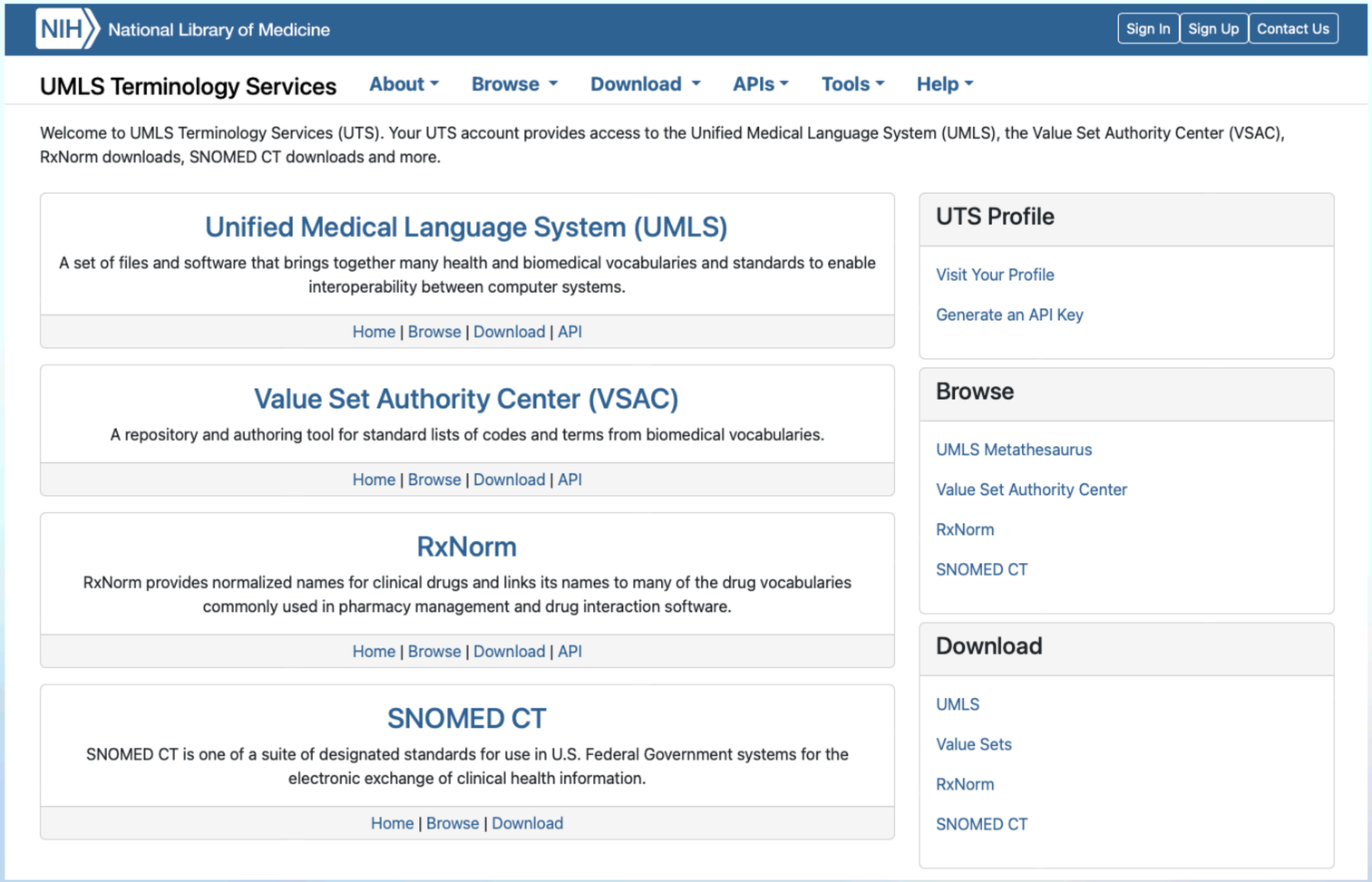


child

high specificity leaves

Unified Medical Language System UMLS

https://uts.nlm.nih.gov/



Metathesaurus: Terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. Hierarchies, definitions, and other relationships and attributes.

Semantic Network: Broad categories (semantic types) and their relationships (semantic relations).

SPECIALIST Lexicon and Lexical Tools: A large syntactic lexicon of biomedical and general English and tools for normalising strings, generating lexical variants, and creating indexes.

Current Release Statistics

Concepts: 3,426,877
concept names (AUIs): 16,709,195
concept names (SUIs): 13,775,220
normalised concept names (LUIs): 12,578,717
Number of sources: 170
Number of languages contributing concept names: 28

Opportunities & Challenges in Biomedical Informatics

Opportunities

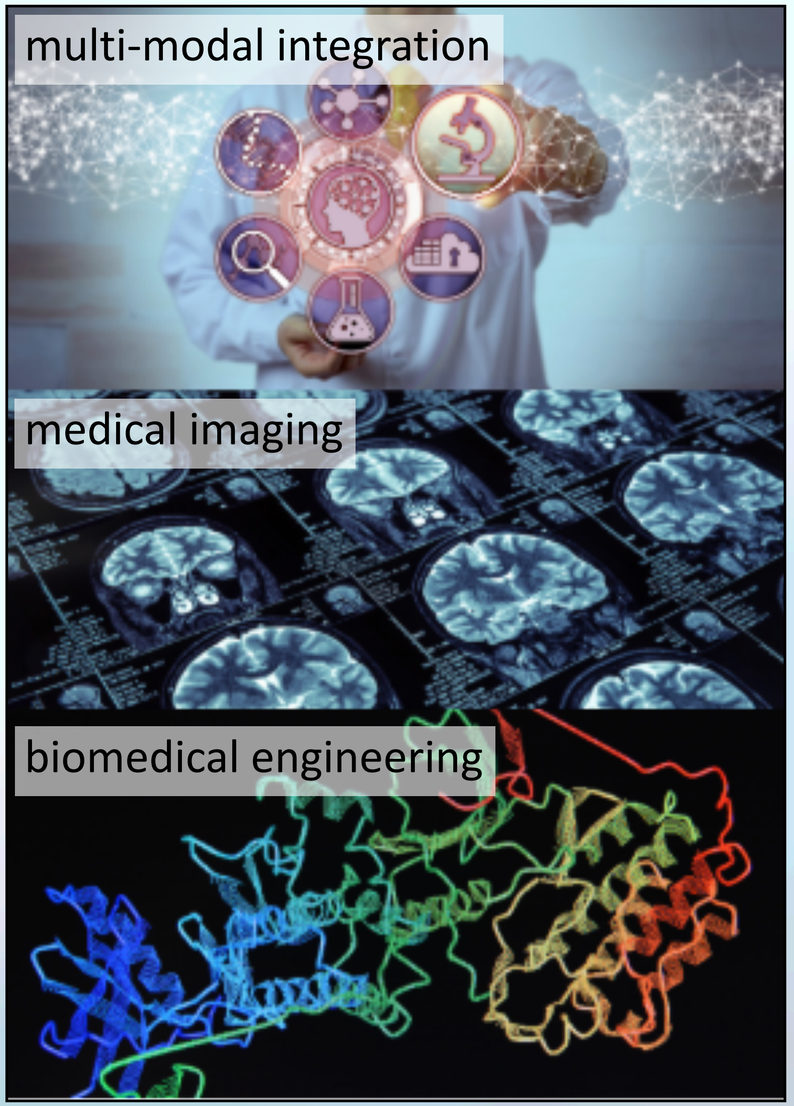
Clinical & Health
- Administration Support
- Decision Support
- Patient Engagement
- Synthetic Data Generation
- Clinical Trial Design & Monitoring
- Population Level Modelling
- Professional Education

Biomedical Science
- Drug Discovery and Design
- Protein Structure Prediction
- Biomedical Image Synthesis
- Patient Data Generation
- Drug Response Prediction
- Biological Sequence Generation
- Medical Text Generation
- Biomedical Signal Generation
- Disease Progression Modeling

Challenges

Technical Challenges
- Unlabelled & Unstructured Data
- Missing Values
- Model & Data Bias
- Poor Longitudinal Coverage
- Scaling Problems
- Lack of Realistic Evaluation Benchmarks
- Explainability
- Data Availability & Inter-Operability

Societal & Health Systems
- Clincal safety, Efficacy, & Reliability
- Evaluation, Regulation, & Certification
- Privacy
- Copyright & Ownership
- Implementation & Adoption



Why Integrate Data?

Advantages

Holistic View of Biological Systems
Combines diverse data types (e.g., genomics, proteomics, and clinical data) to provide a comprehensive understanding of complex biological processes.

Improved Disease Mechanism Discovery
Links molecular-level data (e.g., gene expression) with phenotypic data (e.g., disease symptoms) to uncover disease pathways and biomarkers.

Enhanced Predictive Models
Increases the accuracy and robustness of predictive models by incorporating multi-dimensional data.

Cross-Validation of Findings
Enables verification of results across different datasets, increasing confidence in findings.

Identification of Novel Patterns
Facilitates the discovery of relationships and patterns not evident within individual datasets.

Personalised Medicine
Supports tailored healthcare approaches by integrating genomic, clinical, and lifestyle data for individualized treatment.

Efficient Use of Resources
Leverages existing datasets, reducing the need for redundant experiments.

Support for Multi-Scale Analysis
Bridges scales of biology, from molecular interactions to organ-level and population-wide studies.

Disadvantages

Data Heterogeneity
- Different data types (e.g., genomic vs. clinical) may have varying formats, scales, and resolutions, making integration complex.
- Integrated data may contain errors or missing values from individual sources, compounding the problem in the combined dataset.
- Lack of universal standards for data collection and annotation complicates integration across studies or institutions.
- Integrating data may introduce artificial patterns or average out real biological signal.

Complexity of Interpretation
Multi-modal data can produce results that are difficult to interpret, requiring domain expertise across multiple fields.

Data Availability and Access
Limited access to proprietary or sensitive datasets can restrict integration efforts.

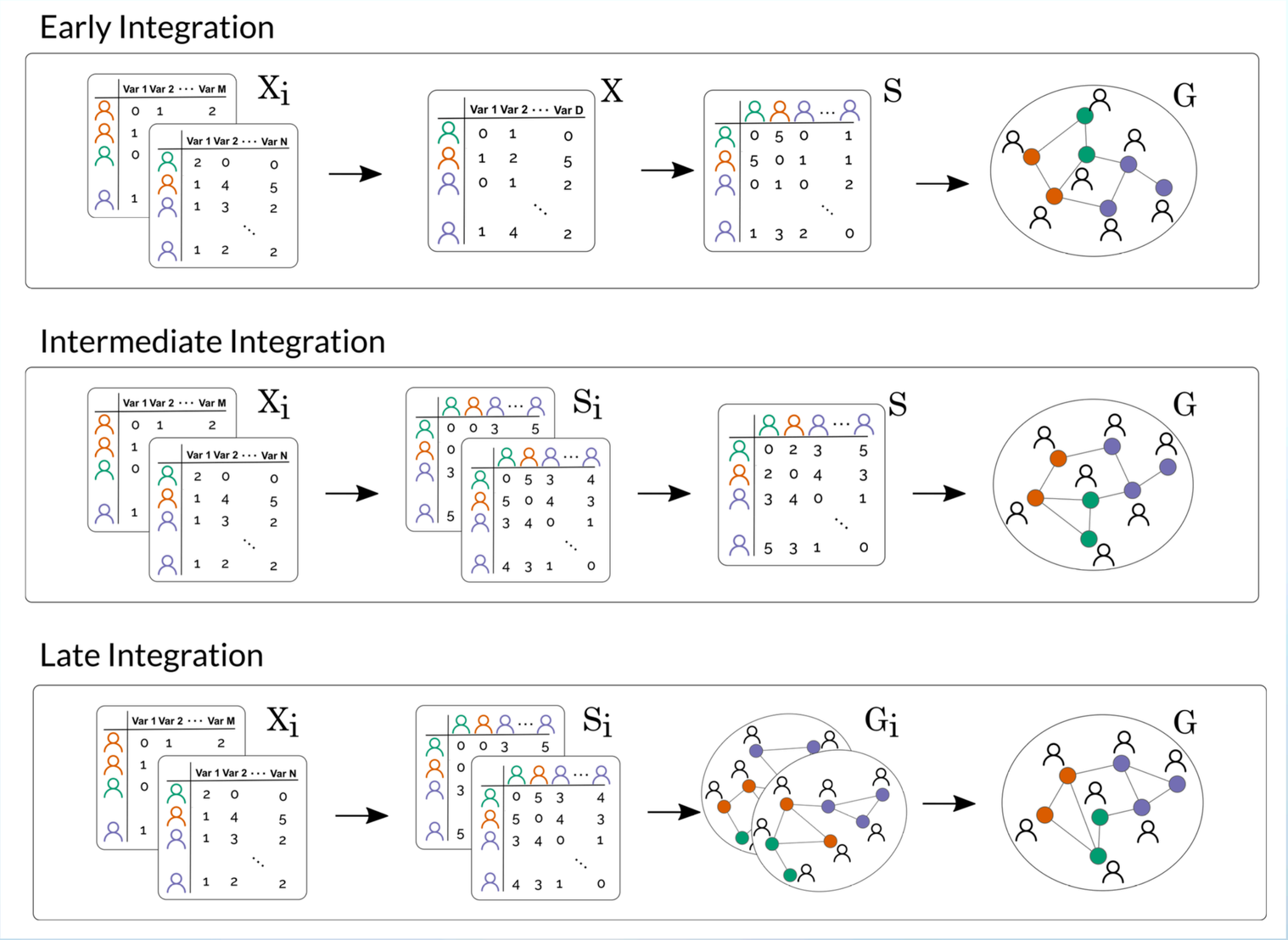
Computational Challenges
Requires significant computational power and advanced algorithms to handle large, multi-dimensional datasets.

Bias
Integration may amplify biases inherent in individual datasets (e.g., population-specific biases).

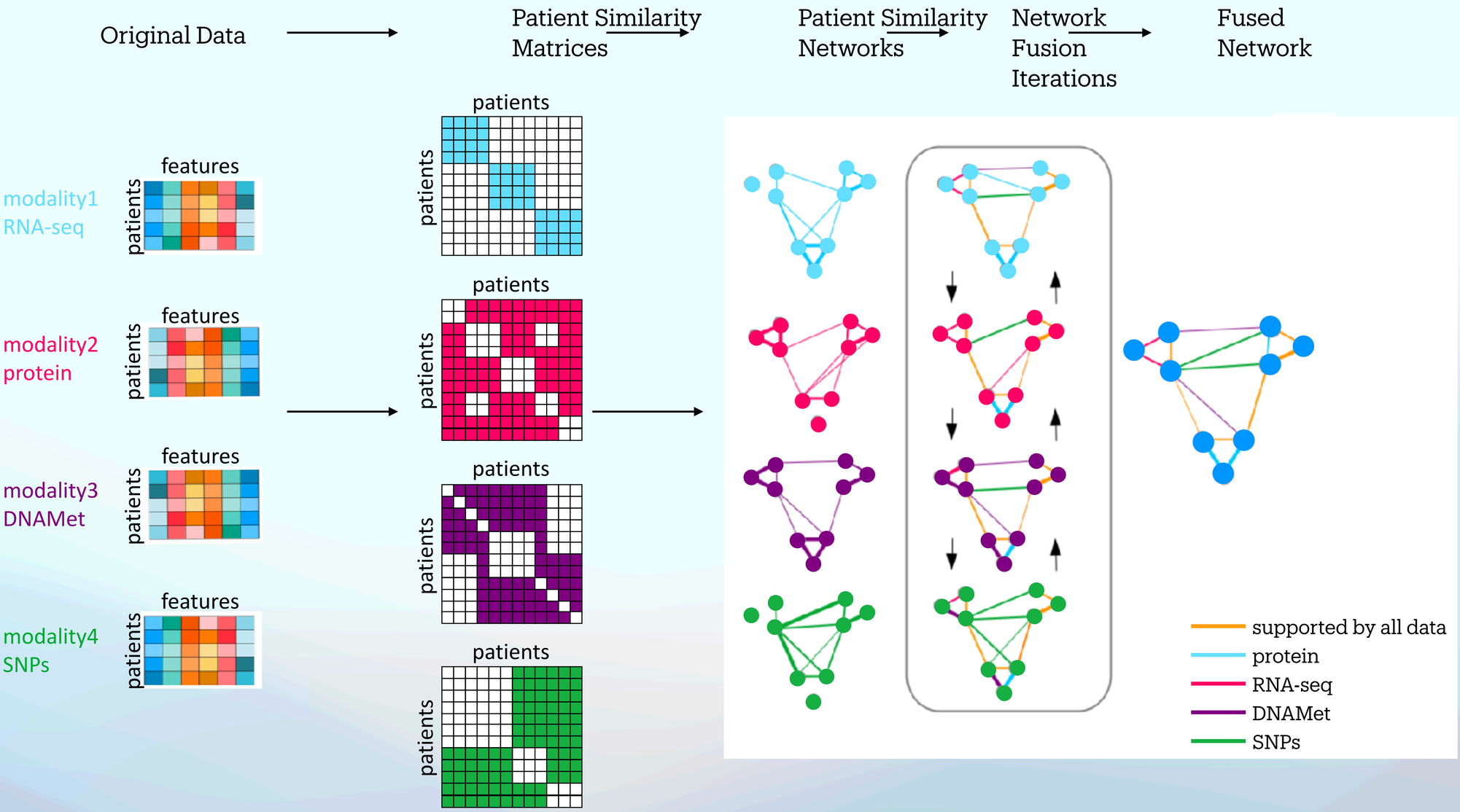
Privacy and Ethical Concerns
Combining sensitive data (e.g., genomic and clinical information) increases the risk of privacy breaches and ethical issues.

Cost and Resource Intensive
Collecting, processing, and integrating diverse datasets often requires significant financial and human resources.

Strategies for Multi-Modal Network Integration



Multi-Modal Integration Through Similarity Network Fusion

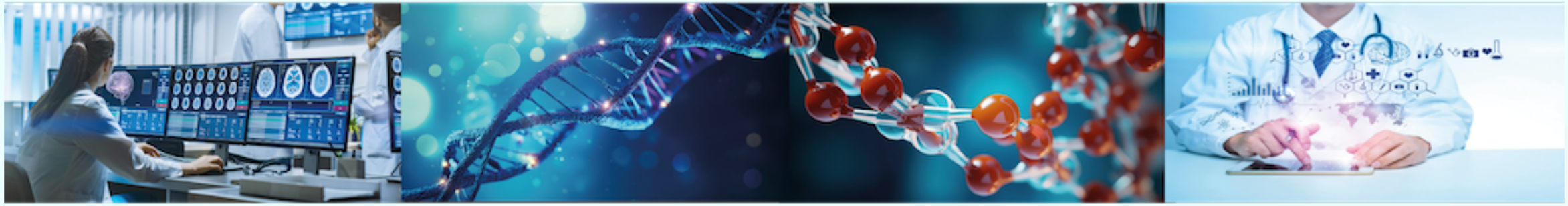


after Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11, 333–337 (2014).

Example Short Answer Questions

1. Why might you consider Gene Set Enrichment Analysis (GSEA) to be a better approach to use than standard Over Representation Analysis (ORA)? (2 marks)
2. What does normalisation of gene expression data achieve? (2 marks)
3. Name two properties that ontologies have that make them well suited as reference structures for storing prior knowledge about biological systems? (1 mark)
4. What is the difference between the OBO-Foundry and the NCBO BioPortal? (1 mark)
5. What is BioMart? What methods can you use to access it programmatically? (2 marks)

NB that 1 mark doesn't necessary mean a single fact answer. Question answers will generally need a small number of lines of explanatory text.



Programming for Biomedical Informatics

Final session next Tuesday (26th)

Revision Session 2

"Guidance on Pseudo-coding and Final Q&A"

Ask Questions on the EdStem Discussion Board

https://github.com/tisimpson/pbi

THE UNIVERSITY of EDINBURGH
informatics