



Databases - 2

BS3033 Data Science for Biologists

Dr Wilson Goh

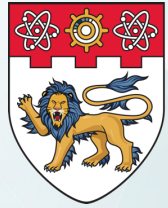
School of Biological Sciences

Learning Objectives

By the end of this topic, you should be able to:

- Explain the design principles for Gene Ontology.
- Apply inferential relations to biological reasoning.
- Explain why biology is considered big data for which data integration is particularly important.





**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

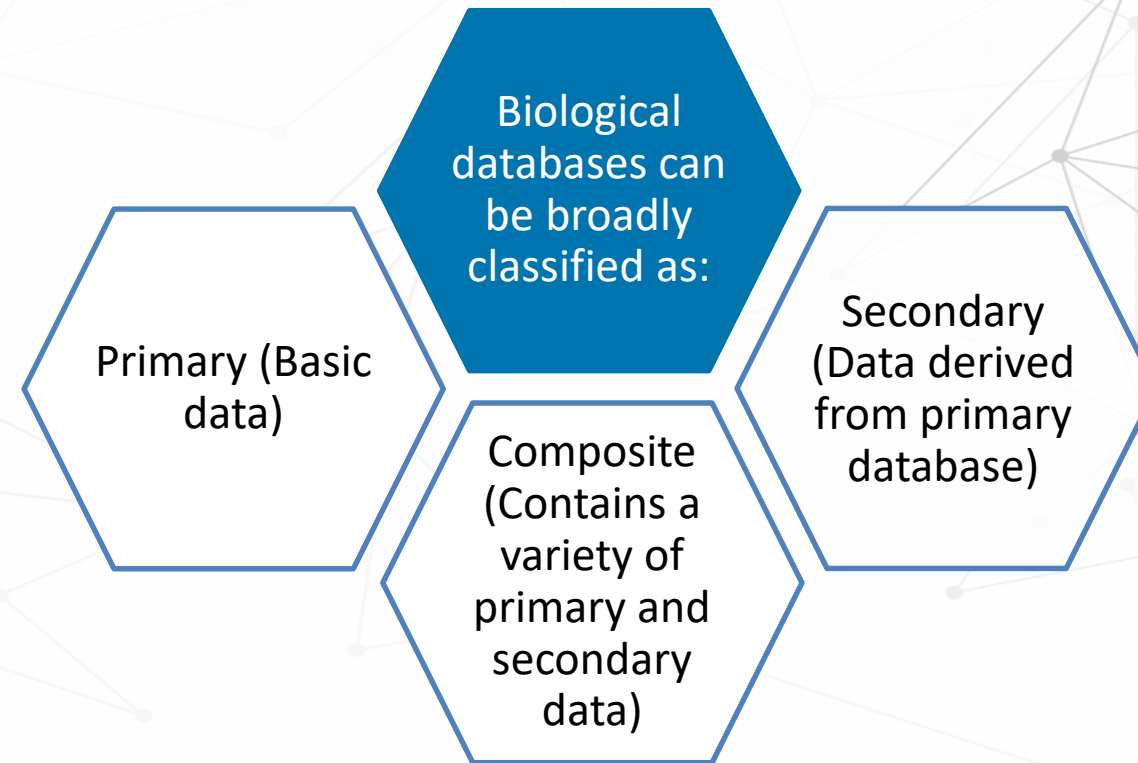
Biological Databases

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

Kinds of Biological Databases



Primary Databases

Primary databases contain information for sequence or structure only:

- Swiss-Prot and PIR for protein sequences
- GenBank and DDBJ for genome sequences
- Protein Databank for protein structures

Secondary Databases

- Secondary databases contain information derived from primary databases.
- Secondary databases store information such as conserved sequences, active site residues, and signature sequences:
 - SCOP and CATH for structural classification of proteins.
 - PROSITE for protein domains.

Composite Databases

- Composite databases contain a variety of primary databases, which eliminates the need to search each one separately.
- Each composite database has different search algorithms and data structures.
- Best known examples include NCBI (<https://www.ncbi.nlm.nih.gov/>) and ENSEMBL (<https://www.ensembl.org/>).

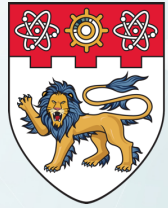
Kinds of Biological Databases

Biological databases can be classified by data type/ focus:

- Gene Sequence (e.g NCBI, EMBL)
- Protein Sequence (Uniprot, SwissProt)
- Genome Assembly (ENSEMBL, SGD, TAIR)
- Bibliographic (Pubmed and Web of Science)
- Disease (OMIM)
- Metabolic pathways (KEGG, WikiPathways, IPA)
- Experimental/Expression (GEO, PRIDE)

Invaluable Resource

- Every year, Nucleic Acids Research publishes an update on newly created biological databases, updates on existing ones, and also information on databases previously published in other journals.
- Access at <https://academic.oup.com/nar/issue/45/D1>.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Overview on Ontologies

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

What is an Ontology?

An ontology is a formal representation of a body of knowledge, within a given domain. A domain can be a field or area e.g. biology is a domain.

Ontologies usually consist of a set of classes or terms with relations that operate between them.

Why do we need Ontologies in Biology?

- Biology is complex.
- Reasoning about biological knowledge, information and concepts is useful at arriving at a common understanding.
 - Common understanding leads to common vocabulary and data structures (human readable and understandable).
 - Leading towards machine-readable applications.

Gene Ontology (GO)

Gene Ontology (GO) is concerned with knowledge organisation on the function and organisation of genes.

The domains that GO represents are biological processes, functions and cellular components.

It is constantly revised and expanded as biological knowledge accumulates.

Gene Ontology

GO describes function with respect to three aspects:

Molecular function (molecular-level activities performed by gene products).

Cellular component (the locations relative to cellular structures in which a gene product performs a function).

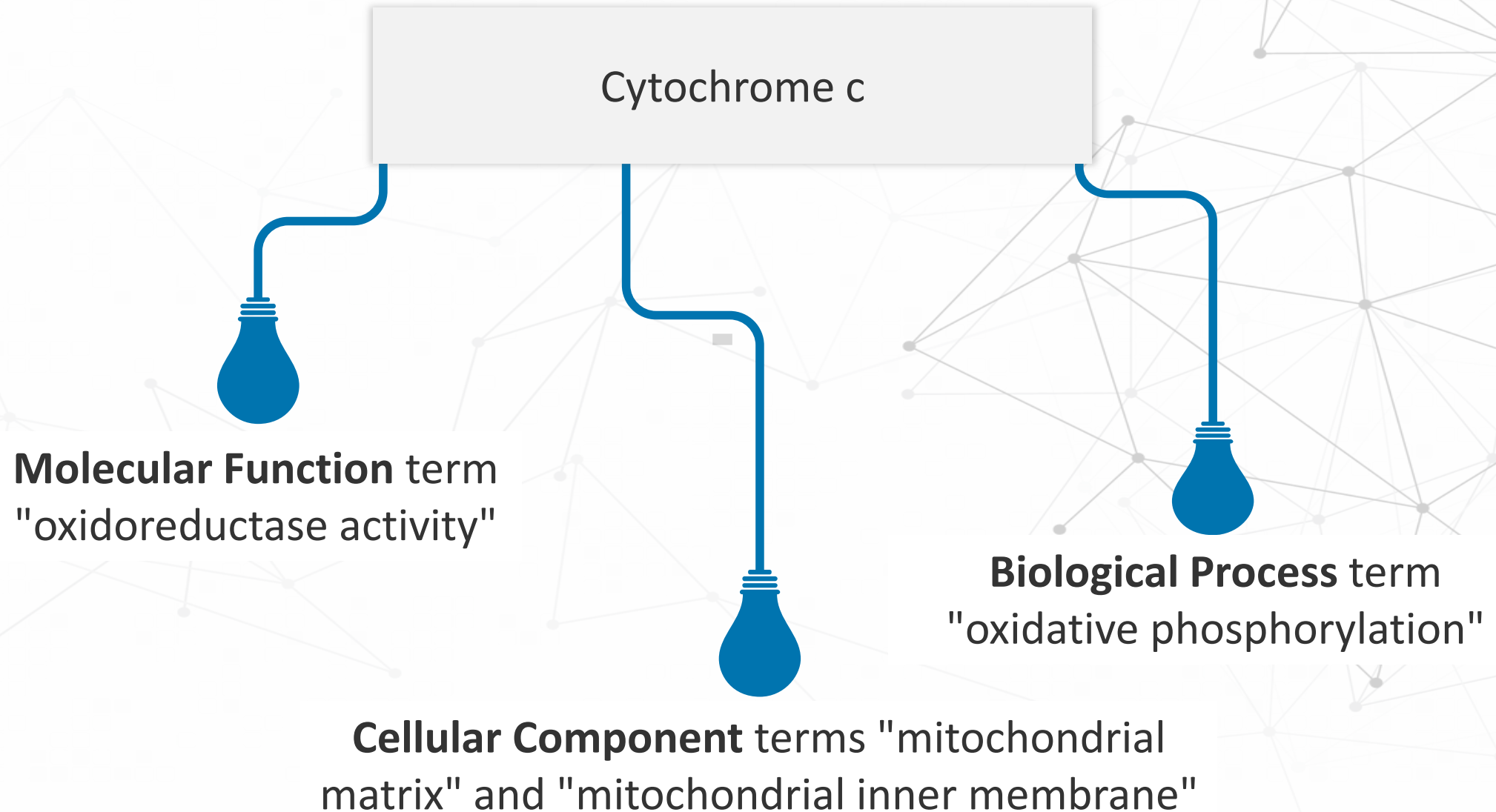
Biological process (the larger processes, or 'biological programs' accomplished by multiple molecular activities).

What do you do?
(Secretary, Engineer, Programmer)

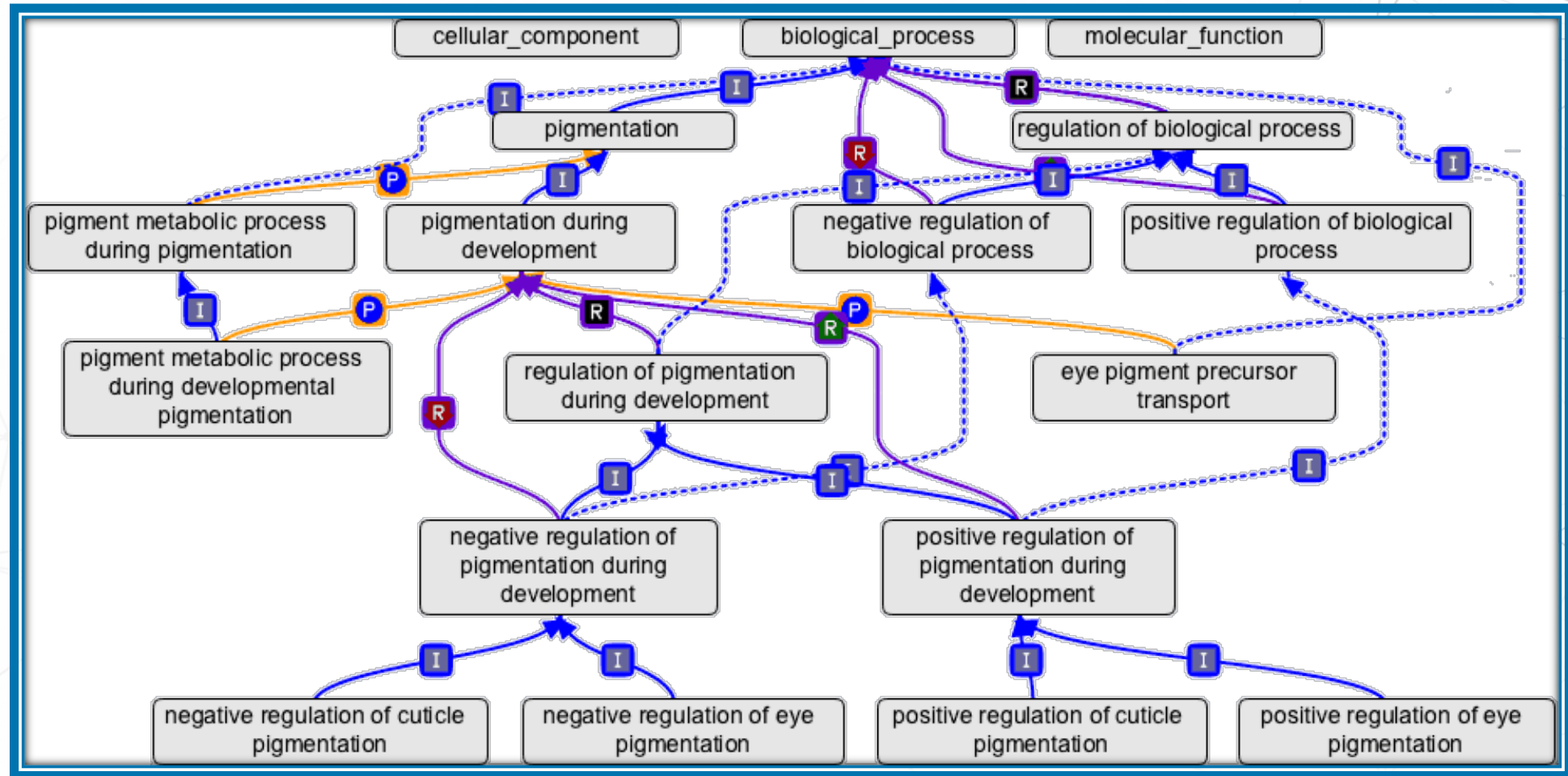
Where do you do it?
(Office, Field, Sea, Land)

What is your department?
(Accounts, Human Resources)

Gene Ontology



Representation of GO Terms as a Graph



Representation of GO Terms as a graph (with arrows as relations)

One Ontology... or Three?

The three GO domains (cellular component/cc, biological process/bp, and molecular function/mf) are each represented by a root ontology term.

All terms in a domain can trace their parentage to the root term, although there may be numerous different paths.

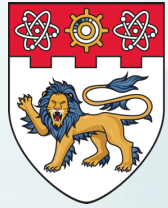
The three root nodes are unrelated and **do not** have a common parent node, and hence GO is in fact...three ontologies.

One Ontology...or Three?

The three GO ontologies are *is_a* disjoint, meaning that no *is_a* relations operate between terms from the different ontologies.

However, other relationships such as *part_of* and *regulates* do operate between the GO ontologies.

For example the molecular function term 'cyclin-dependent protein kinase activity' is *part_of* the biological process 'cell cycle'.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Elements of a GO Term

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

GO Term Structure Elements

Essential Elements:

Unique identifier and term name

Namespace (which of the three sub-ontologies—cc, bp or mf—the term belongs to)

Definition (description of what the term is)

Relationships to other terms

GO Term Structure Elements

Optional Elements:

Secondary IDs

Synonyms (words or phrases closely related in meaning to the term name; related to definition)

Database cross-references (dbxrefs, refer to identical or very similar objects in other databases)

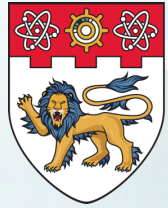
Comment (any other information of interest)

Subset (the term belongs to a designated subset of terms, e.g. one of the GO slims)

Obsolete tag (term has been deprecated and should not be used)

Sample GO Term

- id: GO:0016049
- name: cell growth
- namespace: biological_process
- def: "The process in which a cell irreversibly increases in size over time by accretion and biosynthetic production of matter similar to that already present." [GOC:ai]
- subset: goslim_generic
- subset: goslim_plant
- subset: gosubset_prok
- synonym: "cell expansion" RELATED []
- synonym: "cellular growth" EXACT []
- synonym: "growth of cell" EXACT []
- is_a: GO:0009987 ! cellular process
- is_a: GO:0040007 ! growth
- relationship: part_of GO:0008361 ! regulation of cell size



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Relations in GO

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Understanding Relations in GO

- An ontology also contains relationship information.
- GO is structured as a graph (terms are nodes) and the relations between the terms (edges).
- Just as each term is defined, the relations between GO terms are also categorised and defined.
- *is a (is a subtype of); part of; has part; regulates, negatively regulates and positively regulates.*

Representing Relations in GO

There are a number of ways of referring to and representing logical relations. The GO relations documentation uses the following conventions:

A *node* is used to refer to GO terms.

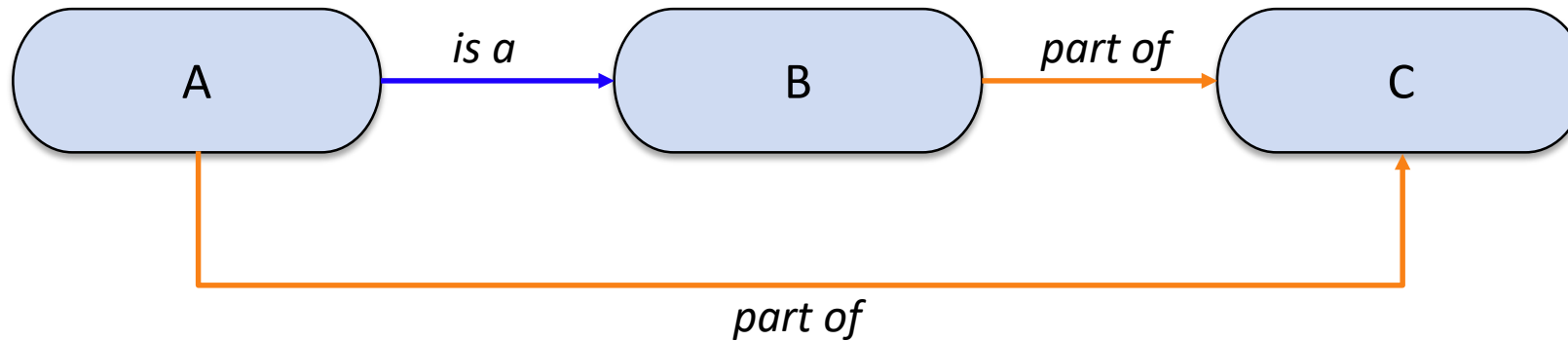
Where it is appropriate to talk about a parent-child relationship between nodes, *parent* refers to the node closer to the root(s) of the graph, and *child* to that closer to the leaf nodes; for the relations *is_a* and *part_of* the *parent* would be a broader GO term, and the *child* would be a more specific term.

The arrowhead indicates the direction of the relationship.

Dotted lines represent an inferred relationship, i.e. one that has not been expressly stated.

Representing Relations in GO

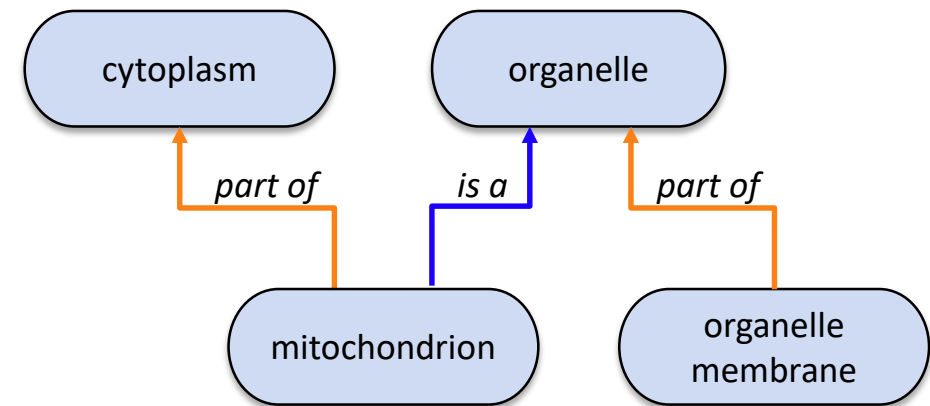
This diagram would be interpreted as per below:



- *A is a B.*
- *B is part of C.*
- Therefore, we can infer that *A is part of C.*
- What form of logical reasoning is this?

GO Relations in Action

- GO nodes can have any number and type of relationships to other nodes.
- Like hierarchies—for example, a family tree or a taxonomy of species—a node may have connections to more than one child (more specific) node, but unlike them, it can also have more than one parent (broader) node, and different relations to its different parents; for example, a node may have a *part of* relationship to one node, and an *is a* relationship to another.

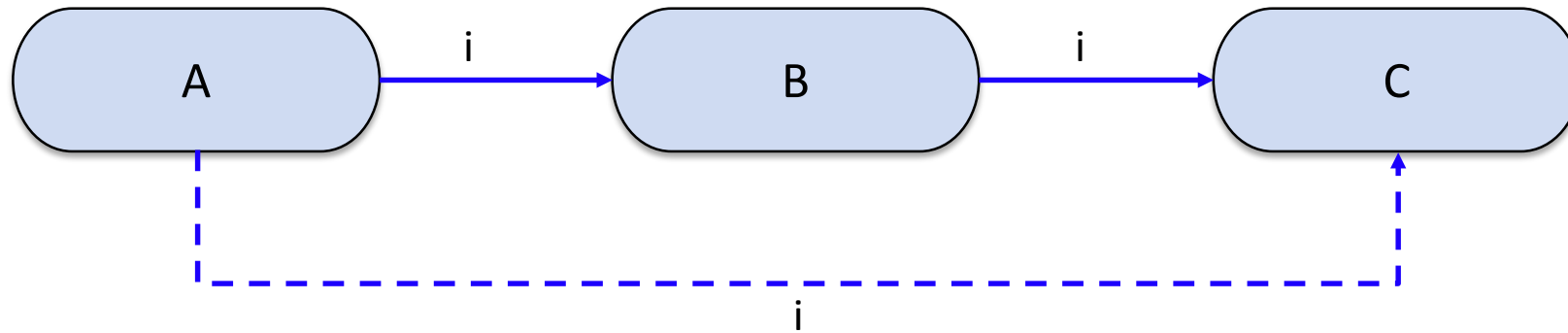


- mitochondrion has two parents: it *is* an organelle and it is *part of* the cytoplasm.
- organelle has two children: mitochondrion *is an* organelle, and organelle membrane is *part of* organelle.

The *is a* Relation

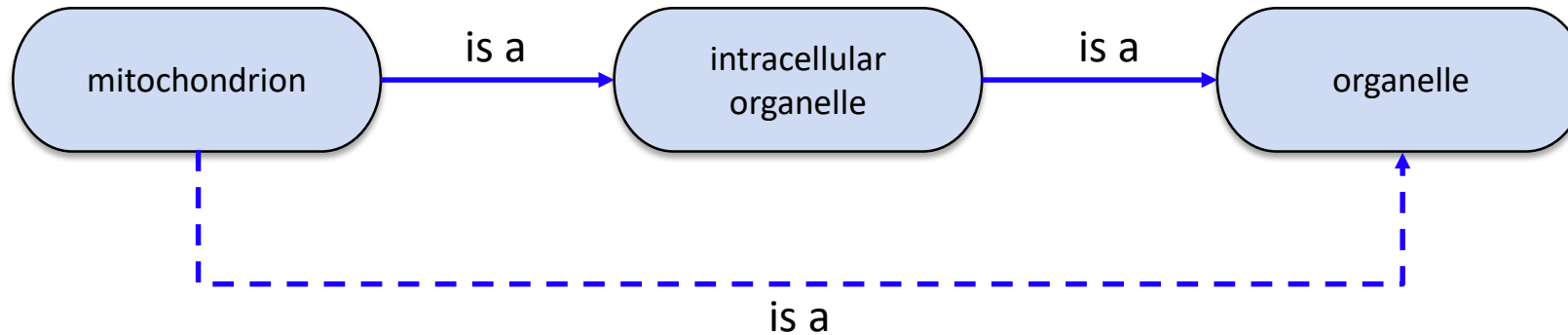
- The *is a* relation forms the basic structure of GO. If we say *A is a B*, we mean that node *A is a subtype of* node B.
- For example, mitotic cell cycle *is a* cell cycle, or lyase activity *is a* catalytic activity.
- It should be noted that *is a* does not mean 'is an instance of'.
- An 'instance', ontologically speaking, is a specific example of something; e.g. a cat *is a* mammal, but Garfield is an *instance* of a cat, rather than a subtype of cat.
- Remember object instantiation from a constructor class in OOP?

The *is a* Relation



The *is a* relation is *transitive*, which means that if *A is a B*, and *B is a C*, we can infer that *A is a C*.

The *is a* Relation

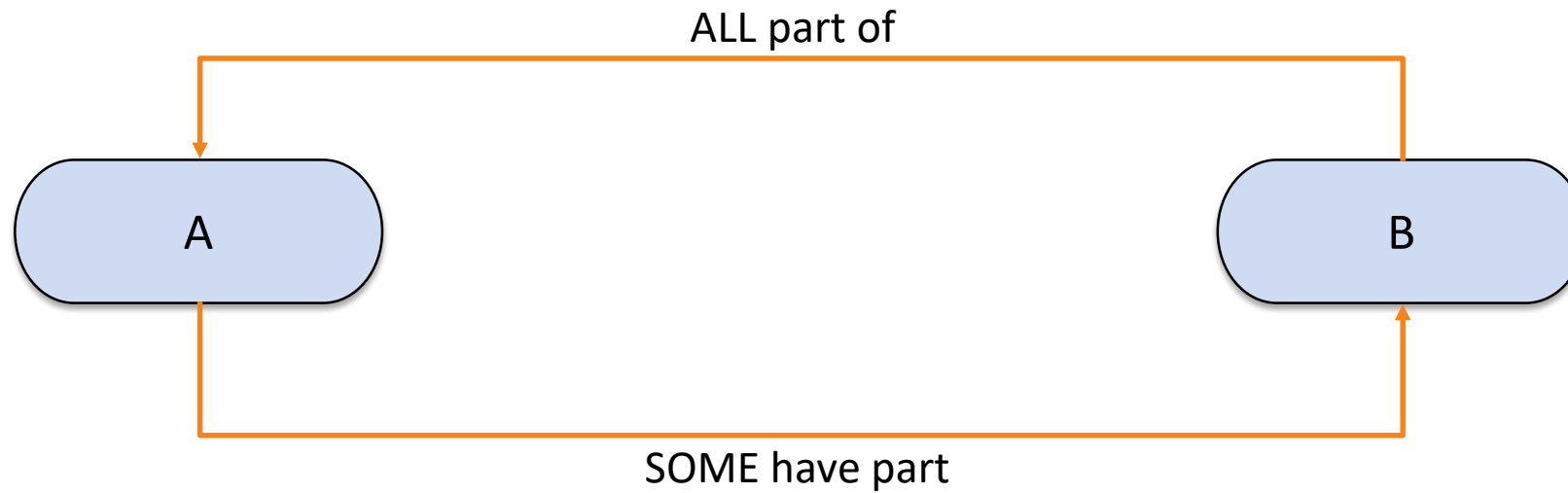


- mitochondrion *is an* intracellular organelle AND intracellular organelle *is an* organelle.
- Therefore mitochondrion *is an* organelle (inferred)
- This is transitive reasoning, and therefore, inductive.
- “Is a” relations are considered strong. So the inferred relationship can be used for subgrouping.
- For example, it is safe to say, a mitochondria is an organelle.

The *part of* Relationship

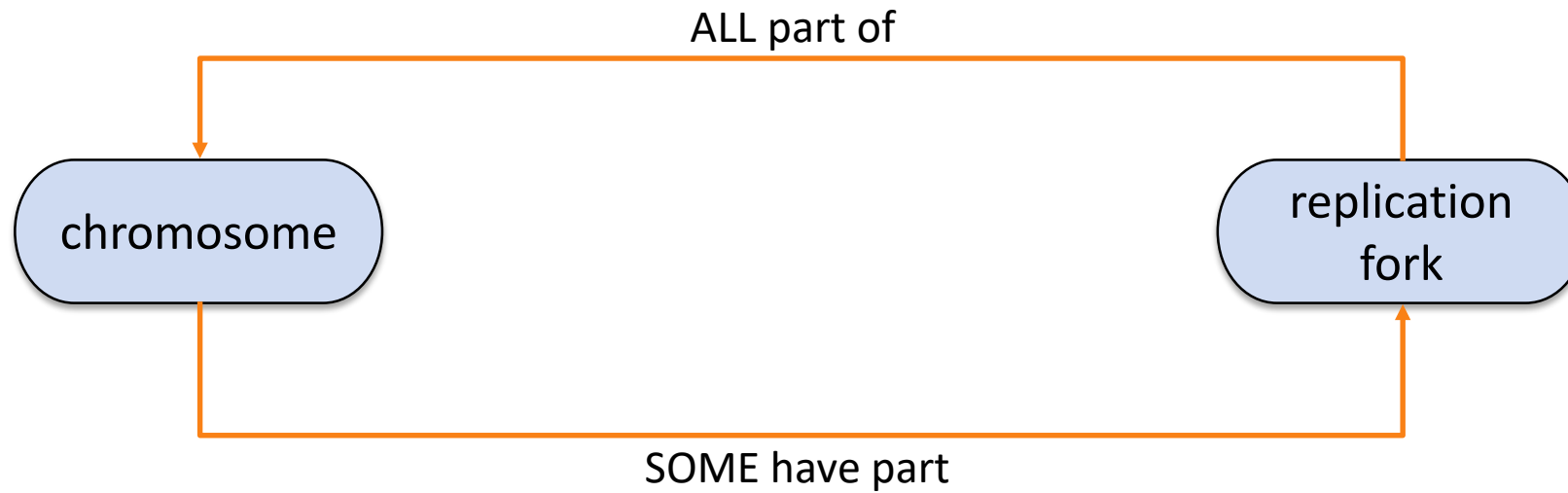
- The relation *part of* is used to represent part-whole relationships in GO.
- *part of* has a specific meaning: exists between A and B
 - if B is *necessarily part of* A: wherever B exists it is as part of A
 - and the presence of B implies the presence of A.
 - However, given the occurrence of A, we cannot say for certain that B exists.

The *part of* Relationship



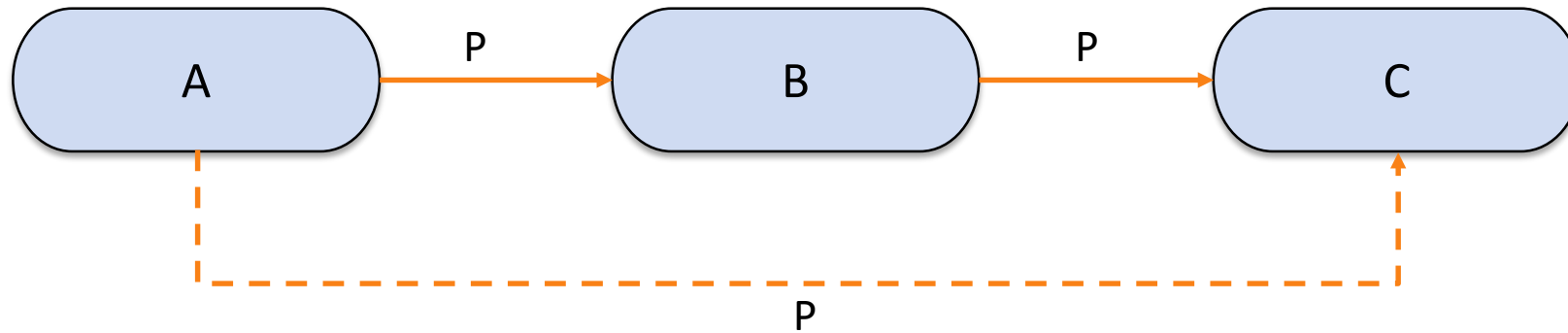
All B are *part of* A; some A *have part* B.

The *part of* Relationship



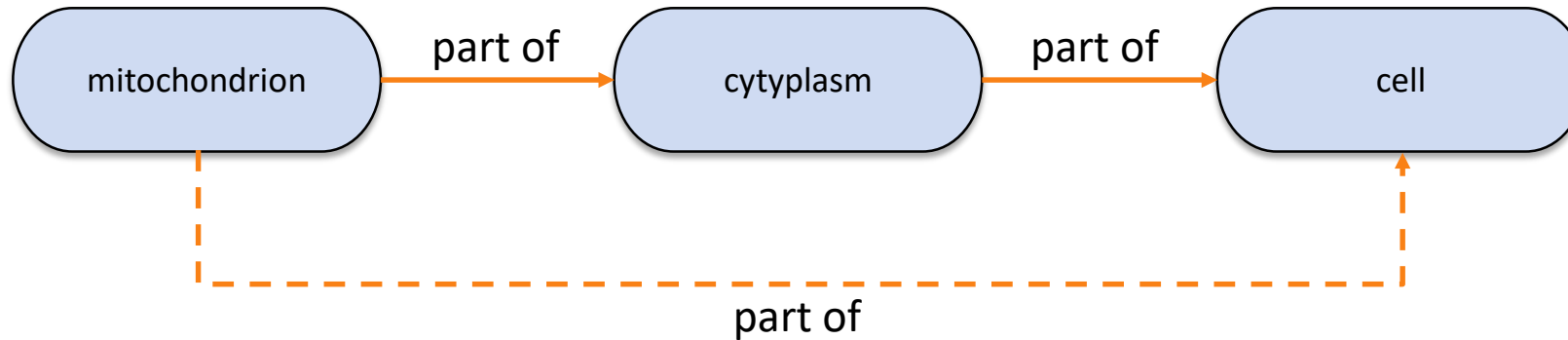
All replication fork are *part of* chromosome; some chromosome *have replication fork*.

The *part of* Relationship



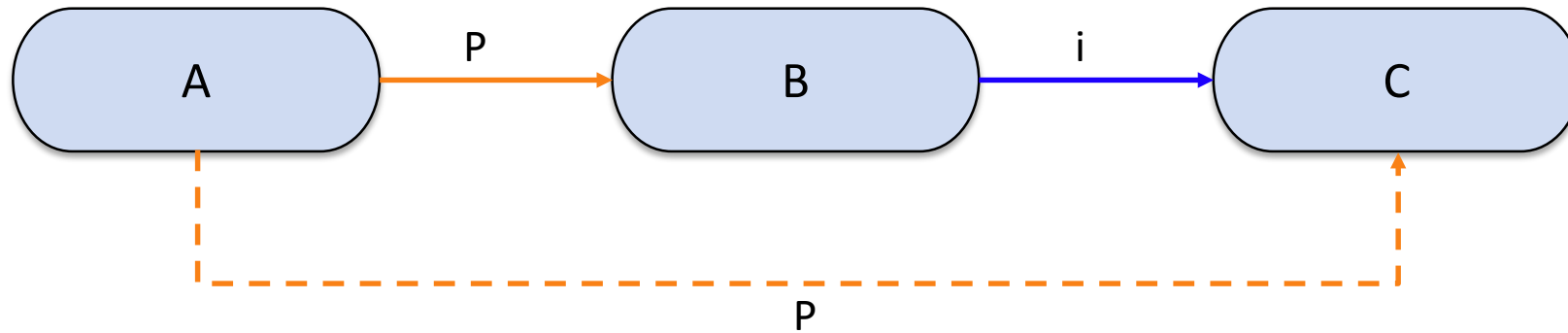
Like *is a*, *part of* is transitive: if A *part of* B *part of* C then A *part of* C.

The *part of* Relationship



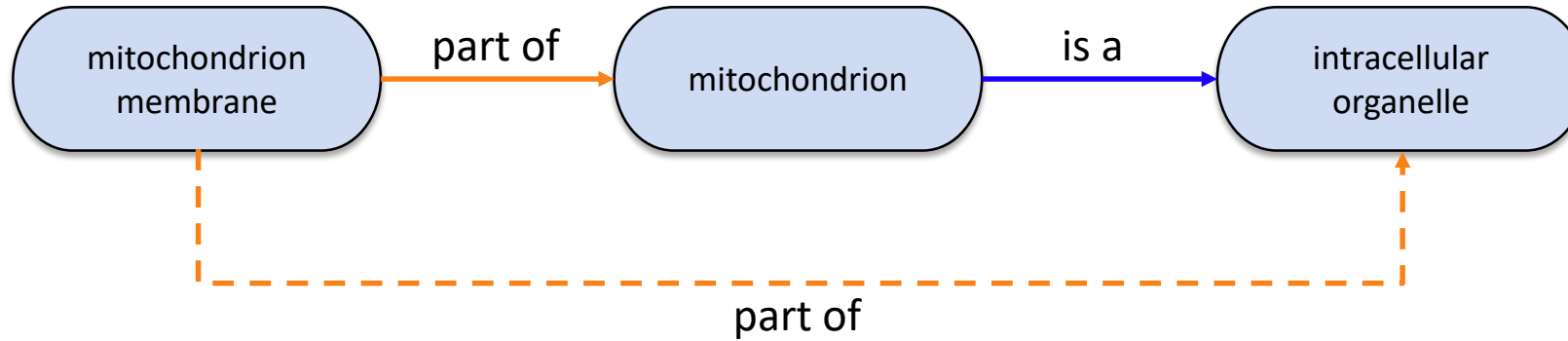
- mitochondrion is *part of* cytoplasm and cytoplasm is *part of* cell therefore mitochondrion is *part of* cell.
- “Part of” relations are considered strong. So the inferred relationship can be used for subgrouping.
- For example, it is safe to say, a mitochondria is part of a cell.

Combining *part of* and *is a* Relations



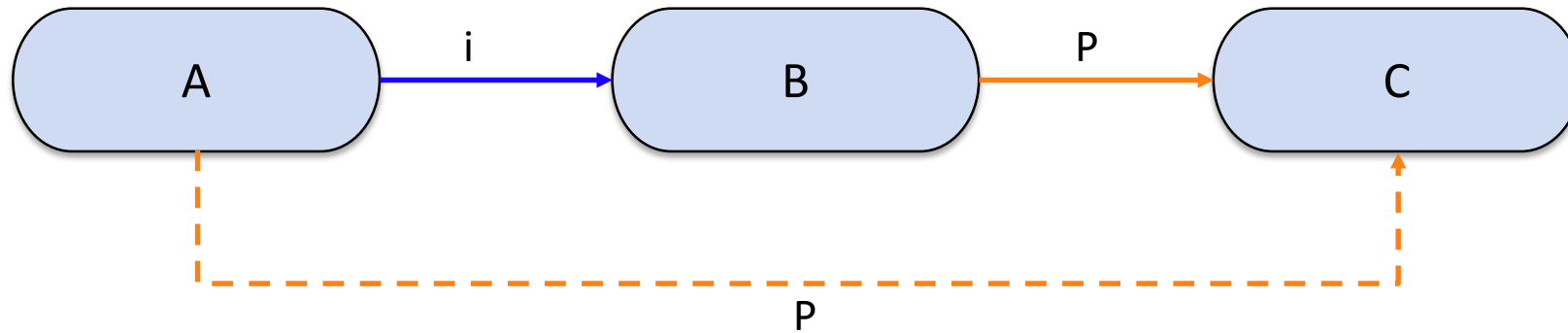
If a *part of* relation is followed by an *is a* relation, it is equivalent to a *part of* relation; if A is *part of* B, and B *is a* C, we can infer that A is *part of* C.

Combining *part of* and *is a* Relations



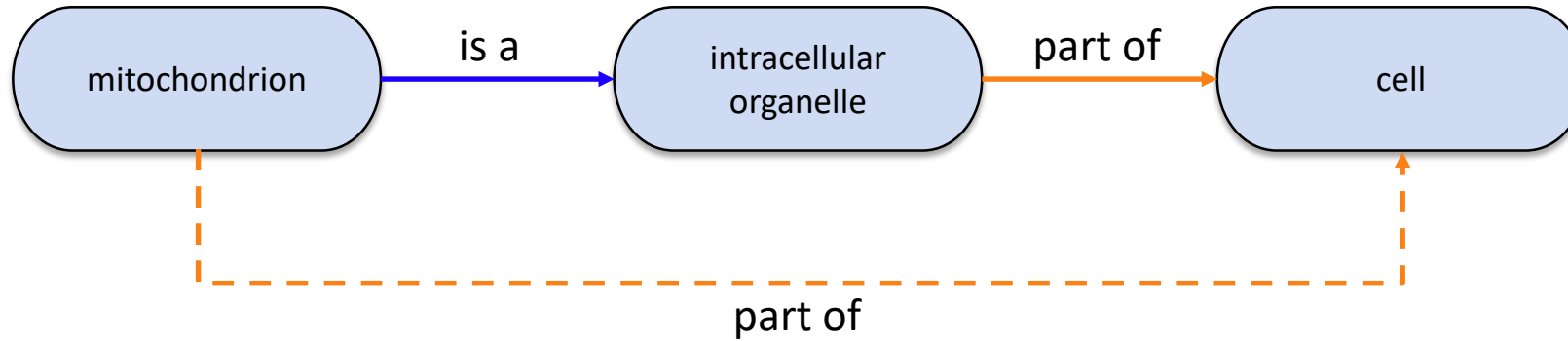
- mitochondrial membrane is *part of* mitochondrion, and mitochondrion is *an* intracellular organelle.
- **Therefore** mitochondrial membrane is *part of* intracellular organelle.

Combining *part of* and *is a* Relations



If the order of the relationships is reversed, the result is the same; if *A is a B*, and *B is part of C*, *A is part of C*.

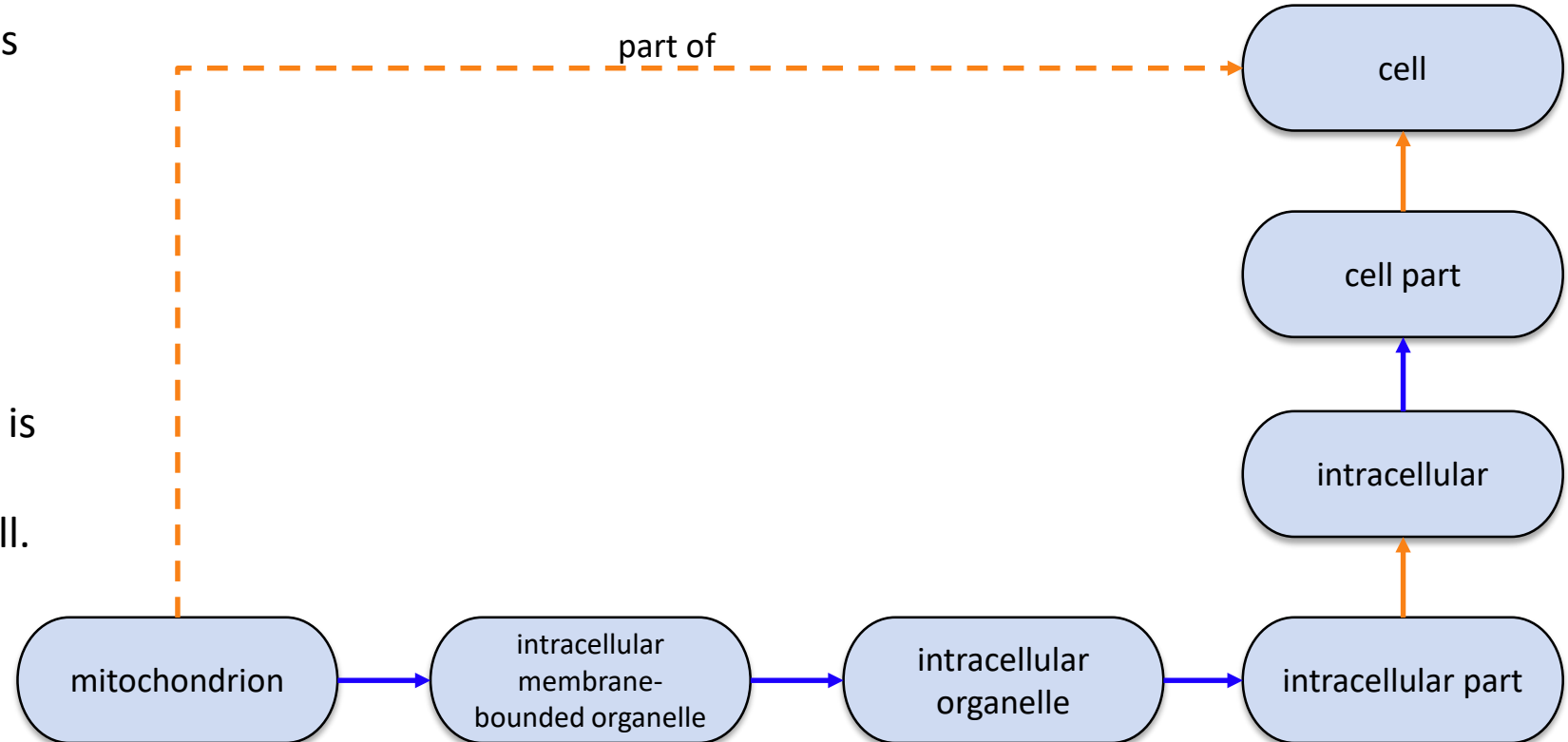
Combining *part of* and *is a* Relations



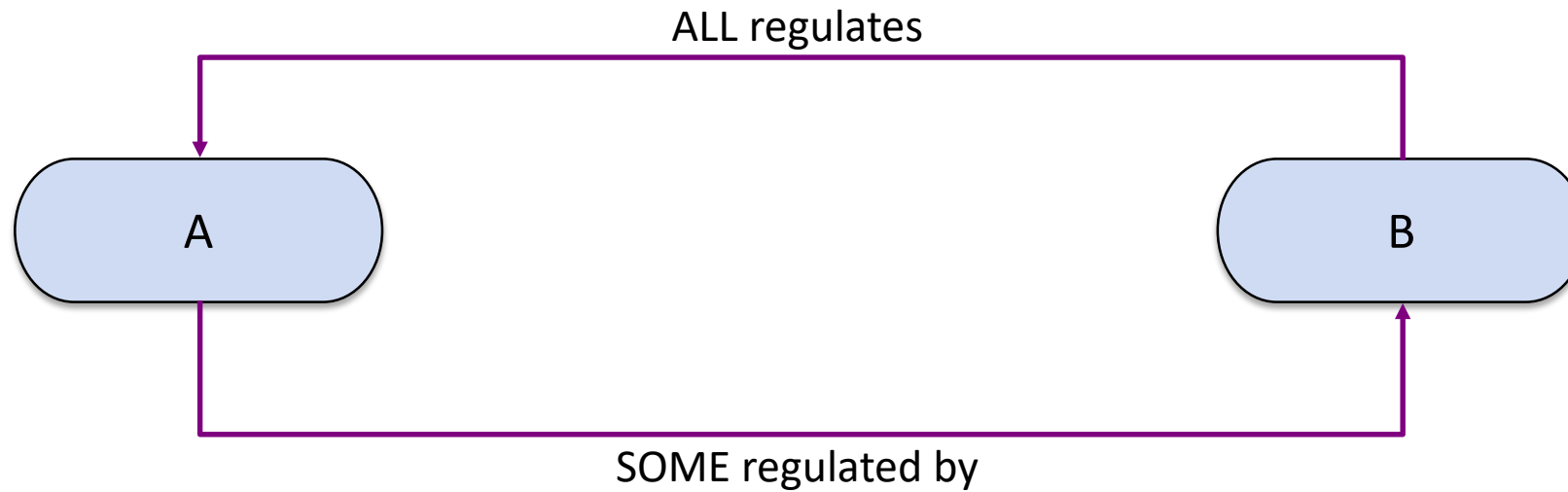
mitochondrion *is a* intracellular organelle and intracellular organelle is *part of* cell therefore mitochondrion is *part of* cell.

Combining *part of* and *is a* Relations

The logical rules regarding the *part of* and *is a* relations hold no matter how many intervening *is a* and *part of* relations there are. Here the nodes between mitochondrion and cell are connected by both *is a* and *part of* relations; this is equivalent to saying mitochondrion is *part of* cell.

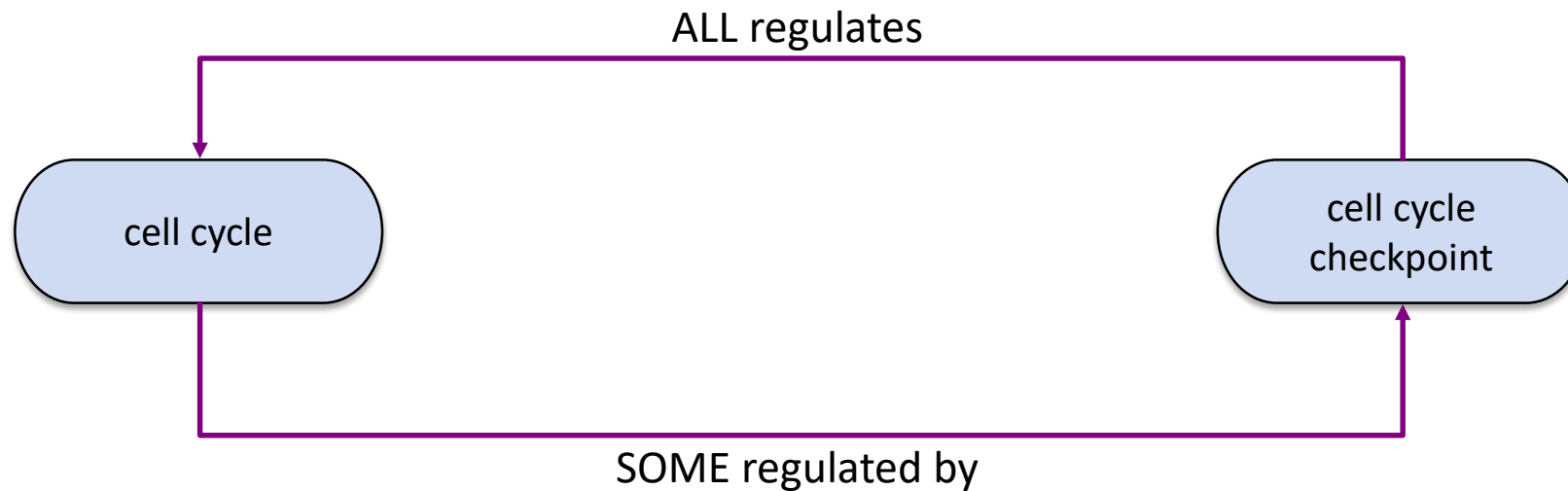


The *regulates* Relationship



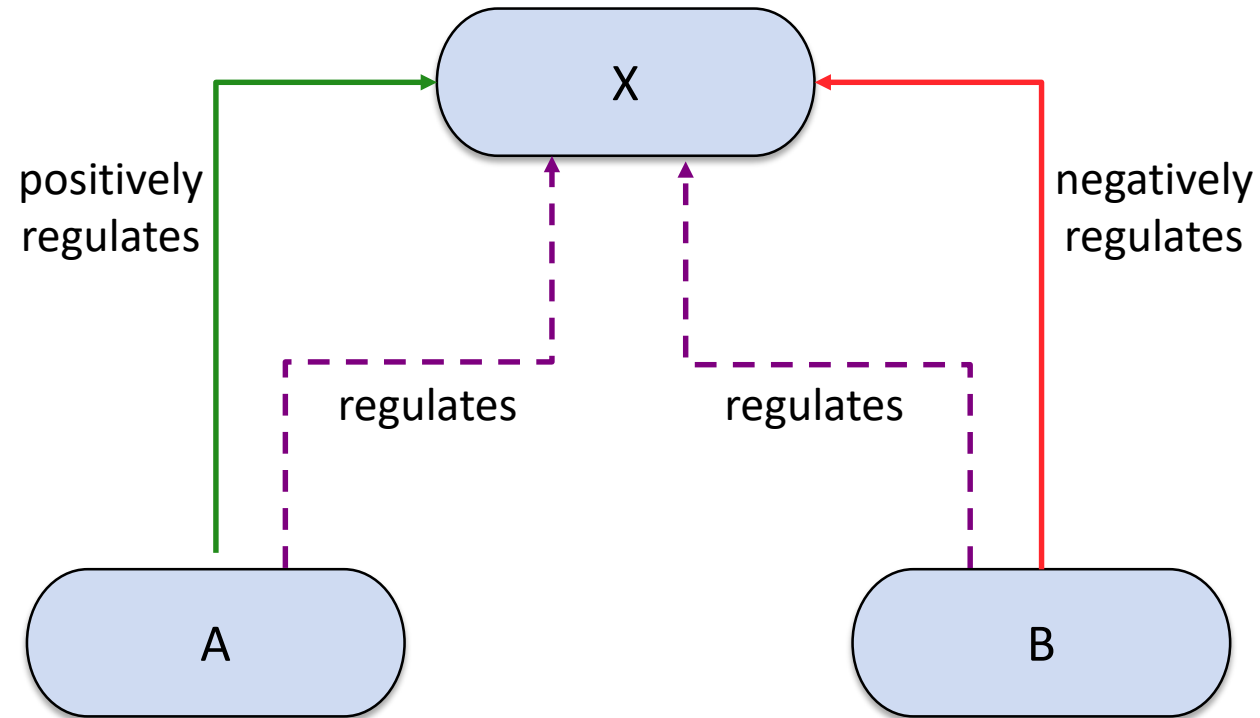
If both A and B are present, B *always regulates* A, but A may not always be *regulated by* B.

The *regulates* Relationship



- Non-reciprocity: whenever a cell cycle checkpoint occurs, it always *regulates* the cell cycle.
- However, the cell cycle is not solely *regulated by* cell cycle checkpoints; there are also other processes that regulate it.

Positive and Negative Regulation



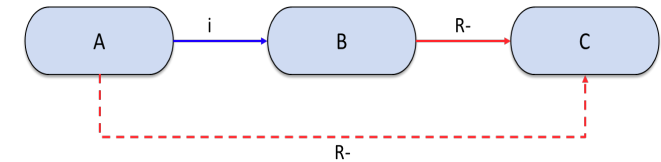
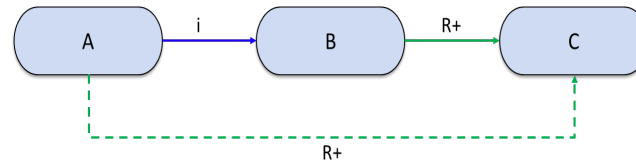
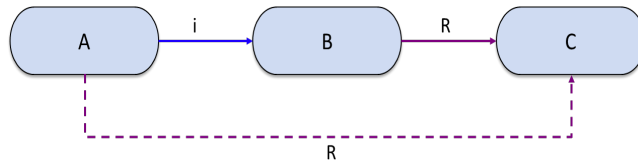
A positively regulates X, so it also regulates X; B negatively regulates X, so it also regulates X.

The *regulates* Relationship

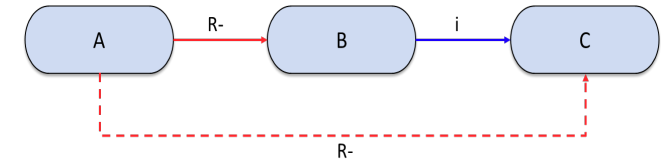
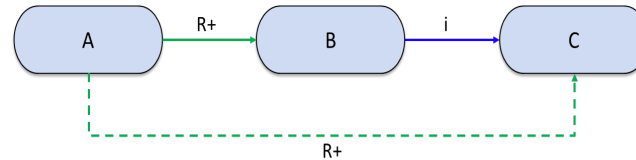
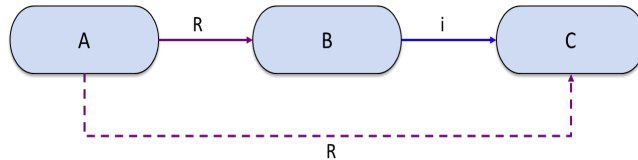
Unlike *is a* and *part of*, grouping annotations to gene products grouped via *regulates* changes the relationship between the GO term and the gene product.

E.g. If an annotation on gene product X records that it is involved in a process that regulates glycolysis, it would not be correct to conclude that X is involved in glycolysis.

Combining *regulates* and *is a*

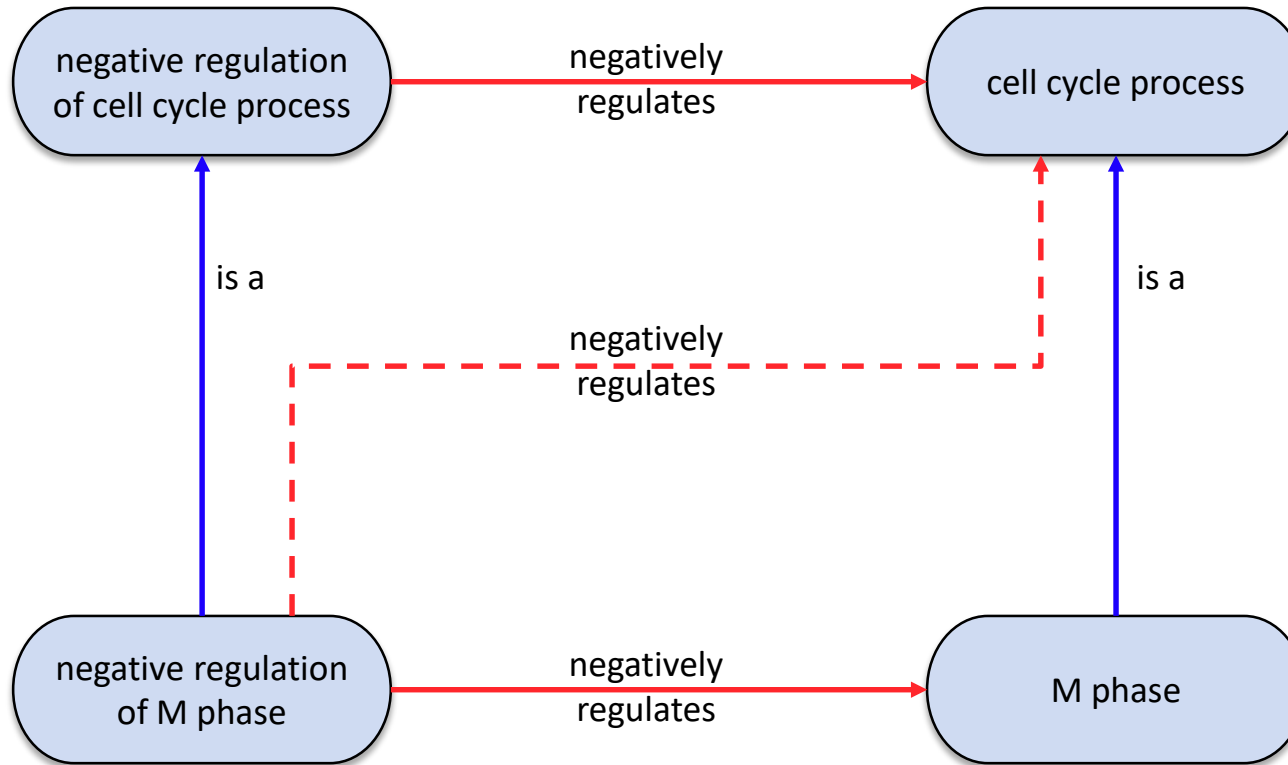


If A *is a* B, and B *regulates* C, we can infer that A *regulates* C. This rule is true for *positively regulates* and *negatively regulates*.

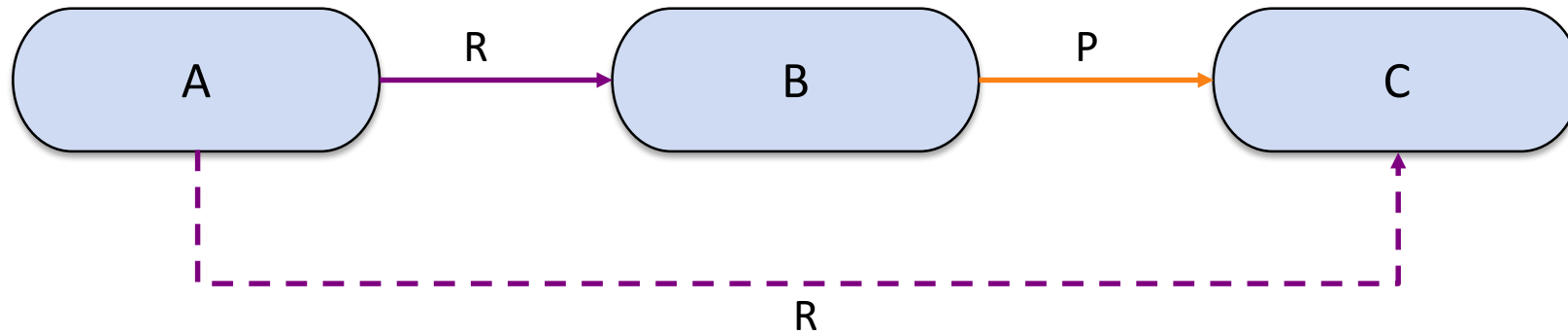


If we switch the relations around, so that A *regulates* B, and B *is a* C, we can again infer that A *regulates* C. This rule also holds true for the *positively regulates* and *negatively regulates* relations.

Combining *regulates* and *is a*

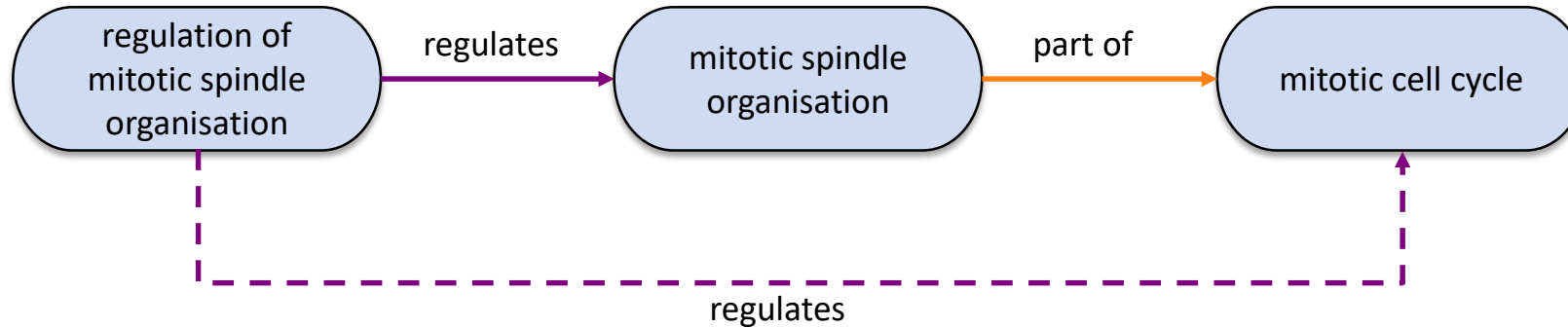


Combining *regulates* and *part of*



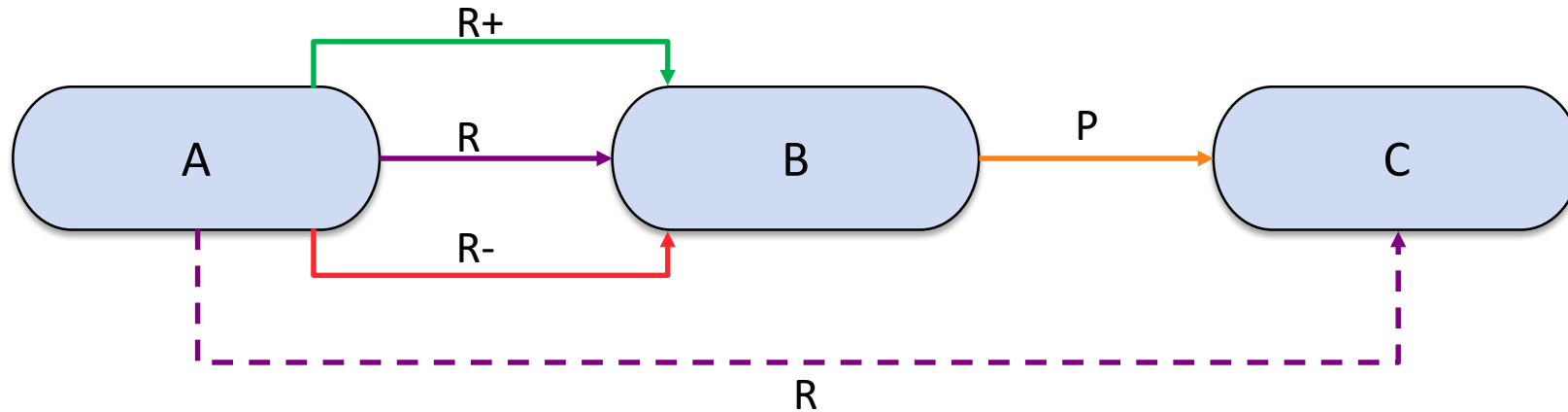
If B is *part of* C, any A that *regulates* B also *regulates* C.

Combining *regulates* and *part of*



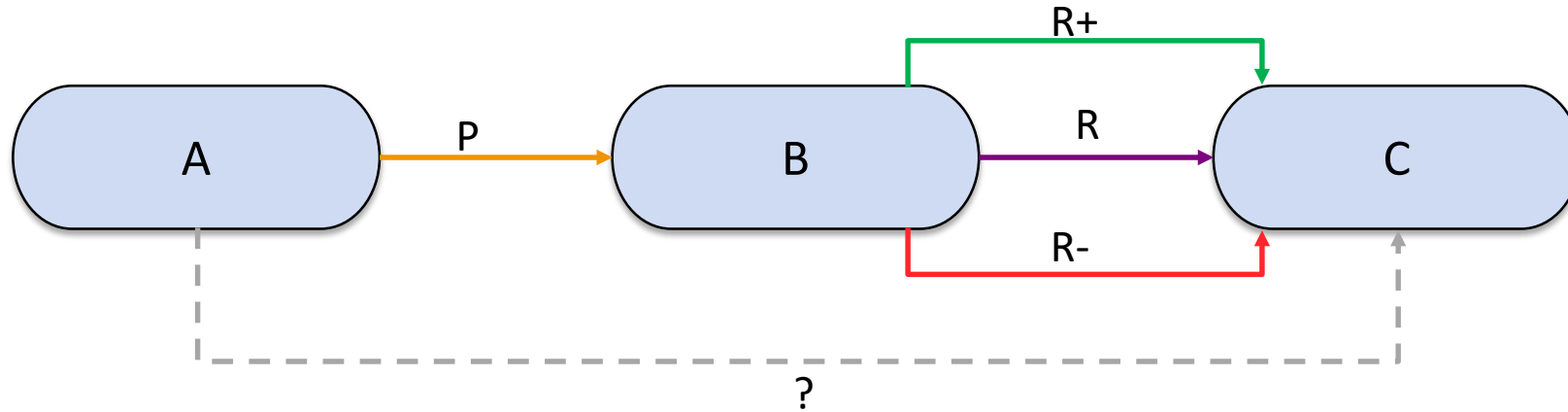
“regulation of mitotic spindle organisation” *regulates* “mitotic spindle organisation” and “mitotic spindle organisation” is *part of* the mitotic cell cycle **therefore** “regulation of mitotic spindle organisation” *regulates* the “mitotic cell cycle”.

Combining *regulates* and *part of*



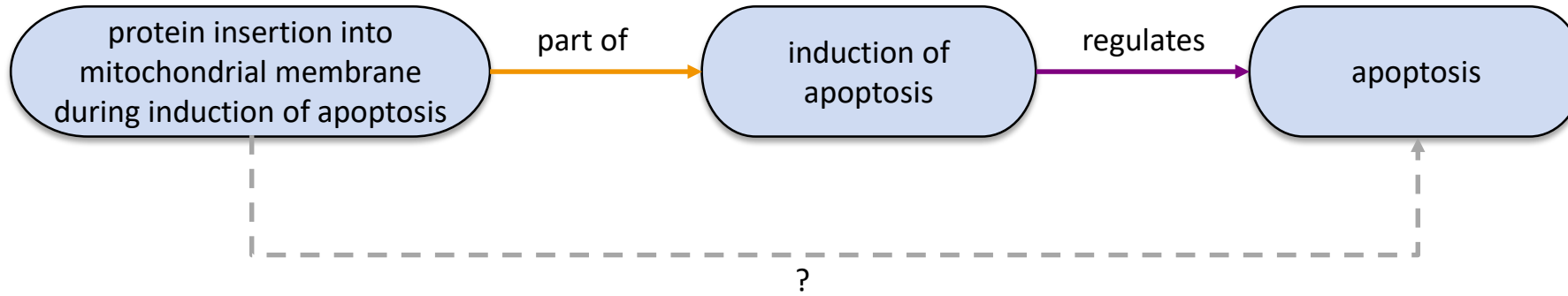
If the relation between A and B is *positively* or *negatively regulates*, and B is *part of* C, we can infer that A *regulates* C—*positively regulates* is a sub-relation of the *regulates* relation, and as previously stated, A *regulates* B *part of* C is equivalent to A *regulates* C—but we cannot be more specific than that.

Combining *regulates* and *part of*



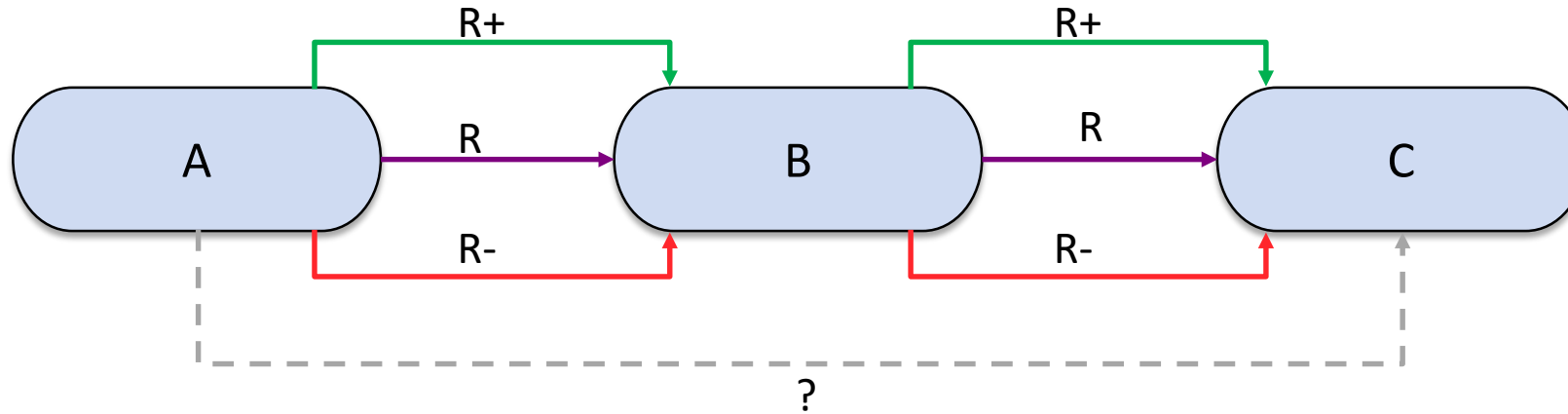
If A is *part of* B, and B *regulates/positively regulates/negatively regulates* C, we cannot make any inferences about the relationship between A and C.

Combining *regulates* and *part of*



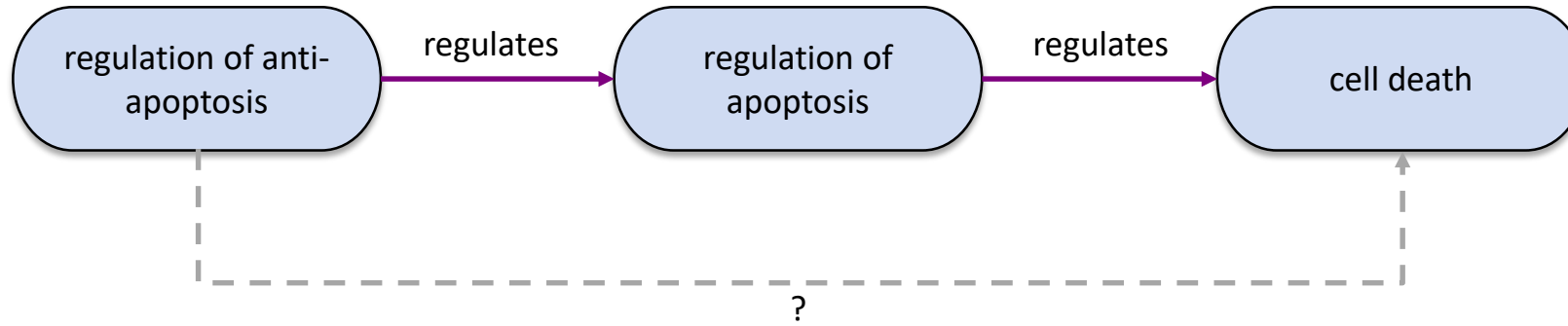
- “protein insertion into mitochondrial membrane occurs during...” is *part of* “induction of apoptosis”, which *regulates* “apoptosis”.
- But we can make no inferences on the relationship of “protein insertion into mitochondrial membrane during induction apoptosis” to “apoptosis”.

Combining *regulates* and *regulates*

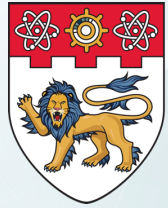


No inference is possible when a *regulates* relation is followed by a second *regulates* relation. This is also true for *positively regulates* and *negatively regulates*.

Combining *regulates* and *regulates*



Regulation of anti-apoptosis *regulates* regulation of apoptosis, which, in turn, *regulates* cell death, but we cannot draw any conclusions from these statements about the relationship between regulation of anti-apoptosis and cell death.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

GO Slims

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

What is GO Slim?

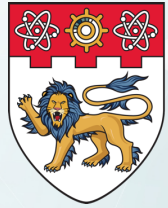
GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms.

What is GO Slim?

GO slims are particularly useful for giving a summary of the results of GO annotation of a genome, microarray, or cDNA collection when broad classification of gene product function is required.

User-created: GO slims are created by users according to their needs, and may be specific to species or to particular areas of the ontologies.

Generic: GO provides a generic GO slim which, like the GO itself, is not species specific, and which should be suitable for most purposes.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Uses of GO

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Purpose of GO

Human readable

- Understanding how concepts and terms are related to each other systematically.

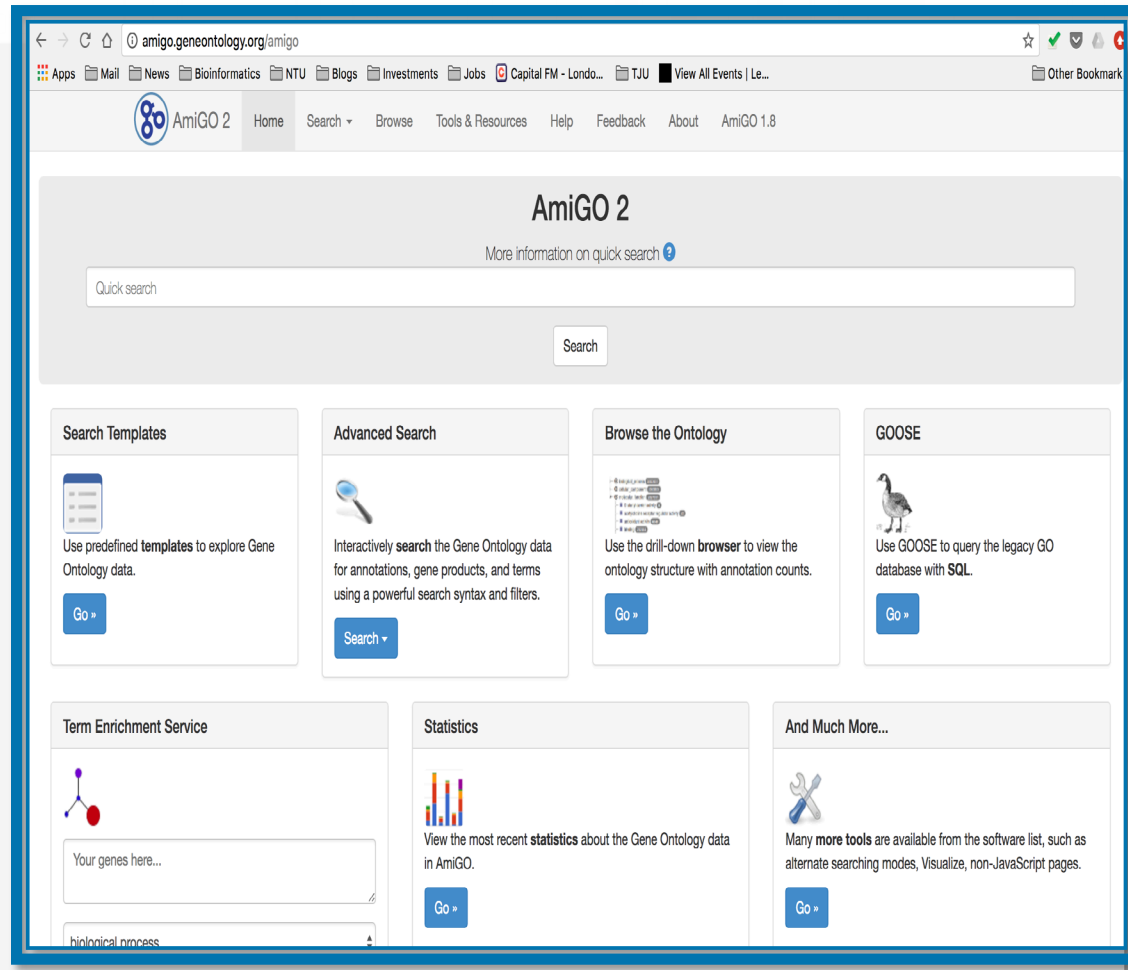
Machine readable

- Enabling tools to access the data and perform tasks and analyses that would be time-consuming and work intensive for humans.

Interpretation of large-scale molecular biology experiments

- Identifies groups of genes that work together, transforming thousands of genes to a few enriched biological functions.

AMIGO/2

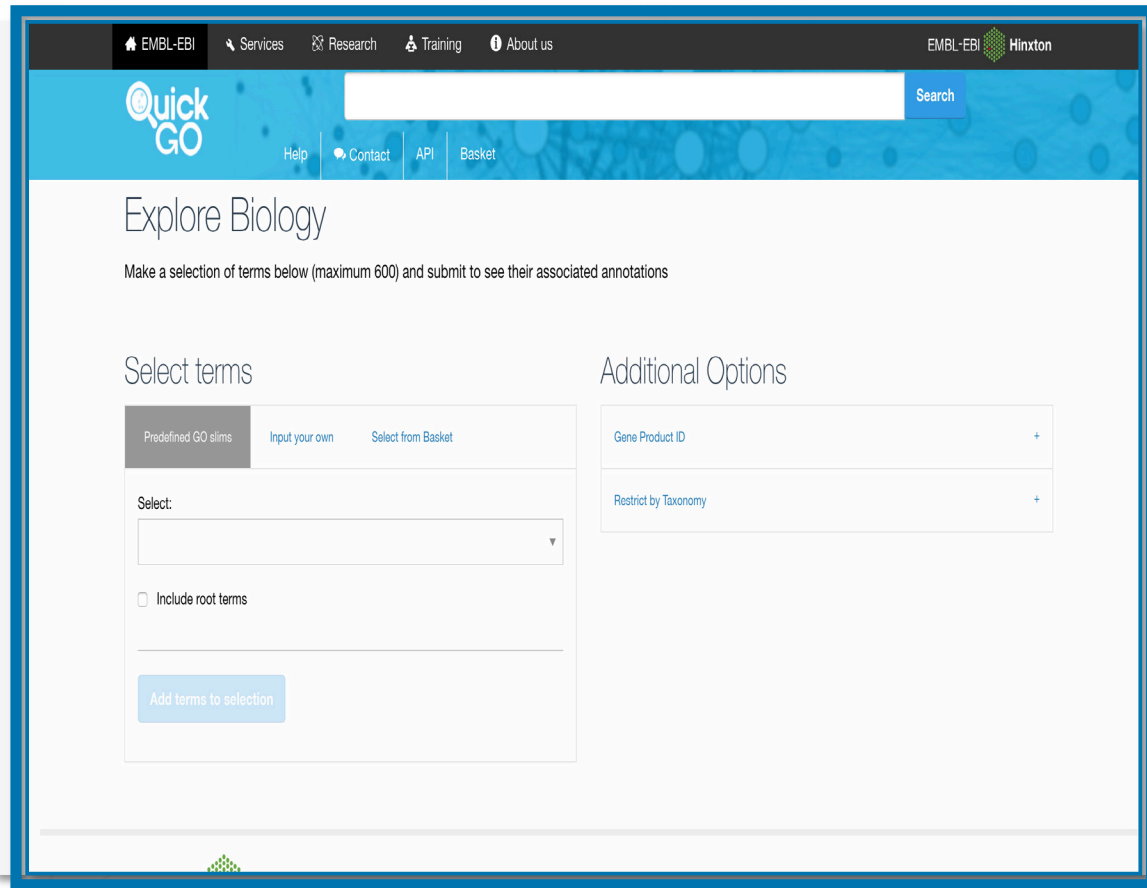


AMIGO/2 is a web-based browser of the Gene Ontology and Gene Ontology annotation data. It has the following features:

- Browse GO annotations.
- View GO-ontology structure.
- GO term enrichment analysis given gene sets.
- Provides access to legacy SQL-based GO-terms database.

Source: <http://amigo.geneontology.org/amigo>

QuickGO



The screenshot shows the QuickGO web interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, and About us. Below this is a blue header with the QuickGO logo and a search bar. The main content area is titled "Explore Biology" and includes a sub-header "Make a selection of terms below (maximum 600) and submit to see their associated annotations". The interface is divided into two main sections: "Select terms" and "Additional Options". The "Select terms" section has tabs for "Predefined GO slims", "Input your own", and "Select from Basket". Under "Predefined GO slims", there is a "Select:" dropdown menu and a checkbox for "Include root terms". An "Add terms to selection" button is at the bottom of this section. The "Additional Options" section has two input fields: "Gene Product ID" and "Restrict by Taxonomy", each with a plus sign to its right.

QuickGO is a fast web-based browser of the Gene Ontology and Gene Ontology annotation data. It has the following features:

- Browse GO annotations.
- Bulk downloads of GO annotation data.
- Provide API for interfacing with computer programs.

Source: <https://www.ebi.ac.uk/QuickGO/>

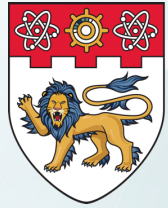
DAVID

The screenshot shows the DAVID Bioinformatics Resources 6.8 website. The header includes the DAVID logo and the text "DAVID Bioinformatics Resources 6.8" and "Laboratory of Human Retrovirology and Immunoinformatics (LHRI)". A navigation bar contains links: Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. A welcome message states: "*** Welcome to DAVID 6.8 ***" and "*** If you are looking for DAVID 6.7, please visit our development site. ***". A "Shortcut to DAVID Tools" sidebar on the left lists: Functional Annotation, Gene Functional Classification, Gene ID Conversion, and Gene Name Batch Viewer. The main content area says "Welcome to DAVID 6.8" and "2003 - 2018". It includes a search bar and a section "What's Important in DAVID?" with links to "Cite DAVID", "IDs of Affy Exon and Gene arrays supported", "Novel Classification Algorithms", "Pre-built Affymetrix and Illumina backgrounds", "User's customized gene background", and "Enhanced calculating speed". Below this is a "Statistics of DAVID" section with a bar chart titled "DAVID Citations (2003-2017)" showing an increasing trend in citations over time. A list of features with checkboxes is also present: Identify enriched biological themes, Discover enriched functional-related gene groups, Cluster redundant annotation terms, Visualize genes on BioCarta & KEGG pathway maps, Display related many-genes-to-many-terms on 2-D view, Search for other functionally related genes not in the list, and List interacting proteins.

DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. It allows you to:

- Performs over-representation analysis (ORA) on gene lists to test for enrichment per GO term.
- Perform ID conversion and gene list subgrouping (based on shared functional terms).

Source: <https://david.ncicrf.gov/>



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Similarities and Differences between Databases and Ontologies

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Similarities between a Database and an Ontology

Database

Can store data.

Requires reasoning about data (to meet functional requirement).

Can be represented as graph (ERD).

Can interface with a program (DBMS).

Can be modified and expanded in future (new tables and entries).

Ontology

Can store data.

Requires reasoning about data (to meet philosophical requirement).

Can be represented as a graph.

Can interface with a program (e.g. Ontology Web Language).

Can be modified and expanded in future (merging and retiring terms).

Similarities between a Database and an Ontology

Database

Focus on collecting, storing and retrieving data.

Multiple options for design (including graph-based ERD).

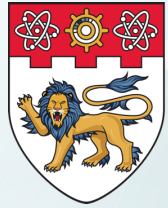
Use of normal forms-decomposition to avoid redundancy.

Ontology

Focus on reasoning about nature and relationships of knowledge.

Primarily uses graph-based reasoning.

Use of inferential reasoning to infer relationships amongst other nodes.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

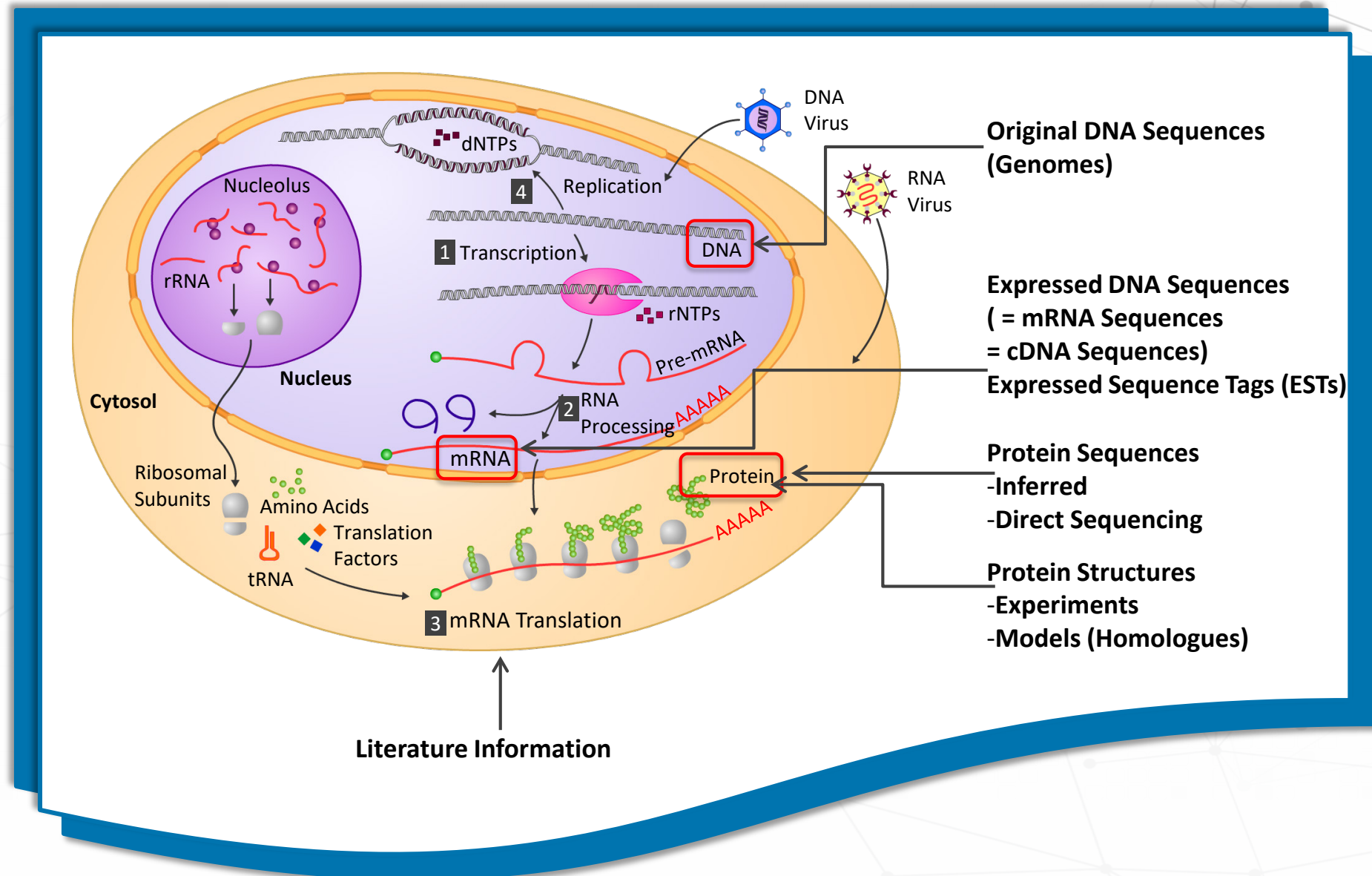
Why Biology is Big Data (and needs integration)?

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

Central Dogma and Biological Data



Biology is Big Data

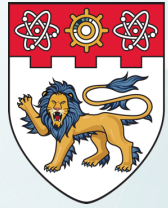
- Because of high-performance computational platforms, biological databases have become important in providing the infrastructure needed for biological research, from data preparation to data extraction.
- The simulation of biological systems also requires computational platforms, which further underscores the need for biological databases.

Biology is Big Data

- In terms of research, bioinformatics tools should be streamlined for analysing the growing amount of data generated from genomics, metabolomics, proteomics, and metagenomics.
- Another future trend will be the annotation of existing data and better integration of databases.

Future Lies is Integration

- With a large number of biological databases available, the need for integration, advancements, and improvements in bioinformatics is paramount.
- Bioinformatics will steadily advance when problems about nomenclature and standardisation are addressed.
- The growth of biological databases will pave the way for further studies on proteins and nucleic acids, impacting therapeutics, biomedical, and related fields.



**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

Summary

BS3033 Data Science for Biologists

Dr Wilson Goh

School of Biological Sciences

Key Takeaways from this Topic

1. Gene ontology (GO) is a major bioinformatics initiative to unify the standardisation and representation of biological knowledge.
2. With the advent of high-performance computational platforms, biological databases have become more important than ever in providing the infrastructure needed for biological research, from data preparation to data extraction.

