Activity

1 Read the proposed scenario.

2 Identify all types of data that are required in order to complete the analysis.

3 Assign a data type to each member of the group and keep the implementation details of each data source independent and secret until all of them have been completed.

4 For each data type, describe an imaginary data source. This can be a published data set or online database/service. If you are already familiar with an appropriate, real data source, feel free to describe it instead of an imaginary one. Do not worry about specifying technical details. Avoid the temptation to produce a data source that is exactly tailored to your needs!

Please consider the following questions about the data source.

- Is it from a single publication or is it a repository of multiple data sets?
  - If it is from a single publication, in what year was it published?
  - If it is a repository, how frequently is it updated? When was the last update?
- Is the data produced experimentally, computationally or both? If multiple data types are provided, specify which are experimental and which are computationally predicted.
- Which species is or are represented in the data?
- What is the main entity or entities of the data source? That is, what would the user primarily search for in order to find the data that they need (*e.g.*, genes, proteins, domains, etc.)?
- How are these entities identified? For example, if it were a genome database, what would the individual gene identifiers (unique names) look like? Is this identifier format unique to this data source? Does the data source provide mappings to other common identifier formats?

5 With the data sources now completed, work together to **draw a schematic of how your analysis "pipeline" would look**.

6 Are there any incompatibilities between the data sources? How would you resolve them?

7 Do the data from any of the data sources need to be transformed in order to be useful for your analysis?

Challenges

1 Imagine that one of your data sources has changed its identifier format recently. Some of the other sources have been updated to refer to these new identifiers, but other sources still refer to the old ones. How would you handle this situation?

2 What assumptions do this integration procedure and your eventual analysis make about the data you have collected? How would you verify that the assumptions are not violated?

Scenario 1

Protein phosphorylation is the reversible addition of a phosphate group to an amino acid, typically a serine, threonine or tyrosine. It is catalyzed by protein kinases, while its reverse reaction, dephosphorylation, is catalyzed by protein phosphatases. It is perhaps the most common form of post-translational protein modification. Phosphorylation is known to play a key role in protein signaling cascades and thus its disregulation may be an important factor in many diseases.

Your group aims to discover phosphorylation sites (phosphosites) whose loss by mutation results in signal disregulation and disease. You would like to use bioinformatics techniques in order to prioritize phosphosites for targeted, experimental analysis. To this end, you would like to focus on phosphosites that have been found to be conserved (that is, the same site has been found to be phosphorylated on orthologous proteins) between humans and other species, as their conservation likely implies important functionality. You would also like to give priority to phosphosites that are predicted to be exposed on the protein's surface, since these are more likely to play a role in dynamic signaling. Finally, you are interested to know if any of these phosphosites map to locations of previously described disease-causing mutations.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

Scenario 2

RNA-seq is a powerful means to assess the quantity of RNA in a cell. By determining the "transcriptome" in cells at a given time or under particular conditions, we can uncover a great amount of information about the cells' activities. Furthermore, by comparing quantifications for different conditions, we can potentially identify the genes that are involved in mechanisms that are important in the response to these changing conditions.

Your group is trying to understand the development of pancreatic cancer cells in order to develop targeted treatments. As a first step, you will find data on how gene expression is affected in these cells compared to healthy cells. You would also like to know whether any mutations have occurred in transcription factors or in their binding sites in pancreatic cancer cells. Finally, in order to assess the possible functional implications of this disregulation of expression, you will determine whether the affected transcripts encode proteins that interact with each other.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

Scenario 3

Mass spectrometry is a method for accurately determining the presence of molecules in a sample. In particular, it can be used to identify peptides in a sample, which can then be used to infer which proteins were present. Various sample-labeling techniques exist that allow for the direct comparison of multiple samples, providing a means to measure relative quantities of peptides between different conditions.

Your group is investigating the effects of osmotic shock on yeast. You would like to identify important molecular systems involved in the cell's response to this stress, which you will later attempt to optimize through targeted mutagenesis. You first would like to know how the cell reconfigures its proteome once introduced into a high-salinity environment. Next, you will determine whether the annotations of proteins showing changes in their abundances are enriched for particular functionalities. Similarly, you would like to know if they tend to cluster together in the yeast protein-protein interaction network. Finally, you will search for known mutations that produce an osmotic shock phenotype and map these to the affected proteins.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

Scenario 4

Complex diseases typically involve the interactions between several genes and the environment. Thus, it is often difficult to identify the causative mechanism(s) of the disease. Furthermore, it can be difficult and time-consuming to test candidate treatments, given so many "moving parts". Fortunately, emerging technologies and knowledge-bases are improving our ability to perform early-stage drug validation through computational means.

Your group is studying inflammatory bowel disease (IBD), a complex disease. You would like to assess the extent to which known compounds can potentially treat the disease. To this end, you will first need a list of all genes and allelic variants that have been associated with IBD. Next, you will check the relative expression profiles of these genes in affected patients compared to healthy individuals. For disease-associated genes that do not show decreased expression, you will then search for chemical compounds that have been found to act as antagonists or activators of the proteins encoded by these genes. Finally, you would like to verify that these compounds will still bind to allelic variants of the proteins associated with IBD. For this, you will find existing structural models of the proteins in order to predict new models with the amino acid changes of the allelic variants; these variant models will then be used in computational docking experiments.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

Scenario 5

G-protein signaling serves as the foundation of a broad range of cellular signaling pathways. The key components are a G-protein coupled receptor (GPCR); a heterotrimeric G-protein, consisting of α, β and γ subunits; and a downstream effector. While the receptor is highly specific to its ligand, the components comprising the G-protein are often interchangeable with no significant effect on signaling dynamics. However, disruption of the interaction between the G-protein and the receptor, or any other malfunction, is a common cause of disease.

Your group is studying the diversity of human heterotrimeric G-proteins. Because the subunits are frequently interchangeable, you would like to know which combinations are likely. You will first collect a broad array of data from expression experiments in order to estimate co-expression patterns for G-protein subunits. For pairs of subunits that tend to co-express, you will then search for prior evidence of physical interaction between them. You will next assess the genetic diversity (polymorphisms) of the subunits across humans and you will determine whether any of the likely heterotrimers also show related polymorphisms (e.g. co-varying across human populations), to strengthen the claim of co-functionality. Lastly, you will assign your predicted heterotrimers to their most-likely signaling pathways according to the previously annotated participation of one or more of the subunits in those pathways.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

Bruno Alvarez, Jan Izquierdo, David Marquez, Adrià Navarro, Eloi Vilella

ASSESSMENT 9 DATABASES

Scenario 2
RNA-seq is a powerful means to assess the quantity of RNA in a cell. By determining the “transcriptome” in cells at a given time or under particular conditions, we can uncover a great amount of information about the cells’ activities. Furthermore, by comparing quantifications for different conditions, we can potentially identify the genes that are involved in mechanisms that are important in the response to these changing conditions. Your group is trying to understand the development of pancreatic cancer cells in order to develop targeted treatments. As a first step, you will find data on how gene expression is affected in these cells compared to healthy cells. You would also like to know whether any mutations have occurred in transcription factors or in their binding sites in pancreatic cancer cells. Finally, in order to assess the possible functional implications of this dysregulation of expression, you will determine whether the affected transcripts encode proteins that interact with each other. You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis. Assume that all the required data is publicly available.

As a first step, you will find data on how gene expression is affected in these cells*(pancreatic cancer cells)* compared to healthy cells.

You would also like to know whether any mutations have occurred in transcription factors or in their binding sites in pancreatic cancer cells.

Finally, in order to assess the possible functional implications of this dysregulation of expression, you will determine whether the affected transcripts encode proteins that interact with each other.

You must now determine what data you would need and how you would integrate that data from disparate sources in order to complete your analysis.

Assume that all the required data is publicly available.

1. **Read the scenario**
2. **Identify all types of data that are required in order to complete the analysis.**
3. **Assign a data type to each member of the group and keep the implementation details of each data source independent and secret until all of them have been completed**

4. **For each data type, describe an imaginary data source. This can be a published data set or online database/service. If you are already familiar**

with an appropriate, real data source, feel free to describe it instead of an imaginary one. Do not worry about specifying technical details. Avoid the temptation to produce a data source that is exactly tailored to your needs!
Please consider the following questions about the data source.

- **Is it from a single publication or is it a repository of multiple data sets?**
  - **If it is from a single publication, in what year was it published?**
  - **If it is a repository, how frequently is it updated? When was the last update?**
- **Is the data produced experimentally, computationally or both? If multiple data types are provided, specify which are experimental and which are computationally predicted.**
- **Which species is or are represented in the data?**
- **What is the main entity or entities of the data source? That is, what would the user primarily search for in order to find the data that they need (e.g., genes, proteins, domains, etc.)?**
- **How are these entities identified? For example, if it were a genome database, what would the individual gene identifiers (unique names) look like? Is this identifier format unique to this data source? Does the data source provide mappings to other common identifier formats?**

- **Transcriptomic Data:**
  - **Type:** RNA-seq data for pancreatic cancer cells and healthy cells.
  - **Imaginary Data Source:** This data source for this would come from a single publication, published in 2020. This data was produced both experimentally and computationally. For this publication, the only species represented were humans, mice and horses. The main entity of the data source were genes, whose unique identifiers are the official gene symbol provided by HGNC (for example BRCA1, TP53 or APOE) as can be found in other databases like NCBI. The output is a RNA nucleotide sequence with uracil instead of thymine.

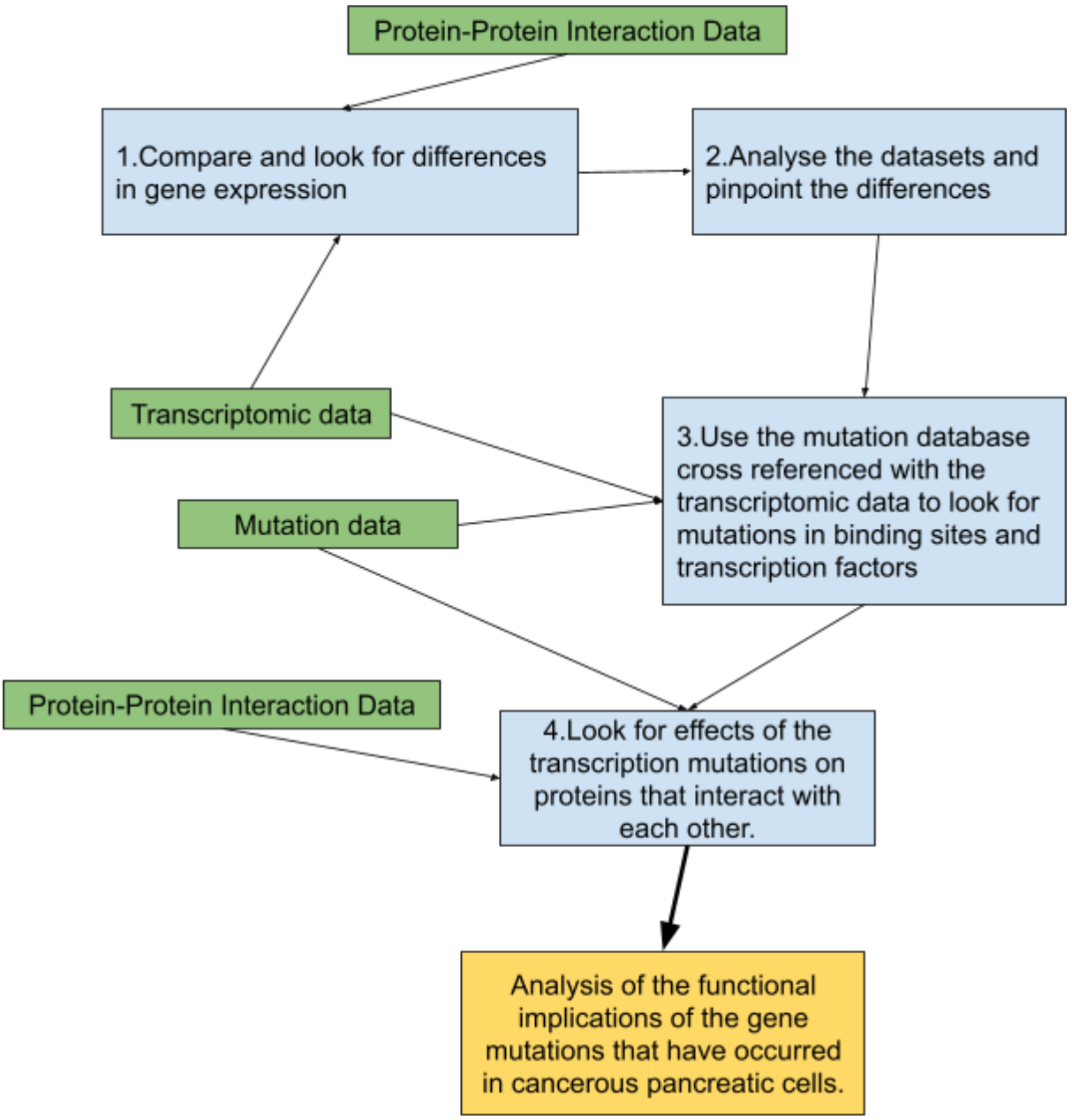
- **Mutation Data:**
  - **Type:** Information about mutations in transcription factors or their binding sites in pancreatic cancer cells.
  - **Imaginary Data Source:** This data source would come from a repository of multiple data sets, updated every six months, whose last update was June 2023. As this is a repository with many data sources,

there are data sources that are obtained experimentally, computationally or both. There are many species represented in this imaginary repository. The identifiers used are gene symbols in a style of their own creation. provided the entry of every gene symbol is the gene suffering the mutation explained in that entry. The imaginary database offers mapping to other databases that describe the gene and the protein encoded in both the wild type and the mutated kind.

- **Protein-Protein Interaction Data:**
  - **Type:** Data on protein-protein interactions.
  - **Imaginary Data Source:** This data source would be a public repository containing multiple data sets. This repository is updated monthly as many data is introduced daily, the last update then was last month. In this repository the only data accepted is the data that is provided both experimentally and computationally predicted, as this ensures the most confident data. There's many species represented. The main entity of the data source are proteins, and its identifiers are UniProt ID of every protein, thus the database is mapped to UniProt directly. The output is an amino acid sequence

5. **With the data sources now completed, work together to draw a schematic of how your analysis "pipeline" would look.**

   The schematic would look like:



The databases(green) used in each step(blue) are indicated by the arrows pointing from those databases to that step. The last step is in the yellow box

6. **Are there any incompatibilities between the data sources? How would you resolve them?**

Indeed, there are incompatibilities between the data sources, primarily arising from Mutated data, which possesses its own set of IDs and gene symbols. However, if we incorporate a mechanism within the databases that allows us to translate the IDs or gene symbols to the specific ones used in the Mutated data, the issue can be effectively resolved. This translation feature would facilitate seamless integration and correlation between the various datasets.

7. **Do the data from any of the data sources need to be transformed in order to be useful for your analysis?**

   Yes, the data from Transcriptomic data, for instance, is an RNA sequence, so we would need to transform it into a DNA sequence in order to have thymines instead of uracils, so it can be analysed as a DNA sequence. This in order to analyse it with other DNA sequences
   Also, in the Protein-Protein interaction data, the data is an amino acid sequence, so we would need to transform it into nucleotides in order to analyse the sequence as DNA. This in order to analyse it with other DNA sequences