

Uczenie maszynowe w bioinformatyce

Wykład 11: analiza funkcjonalna

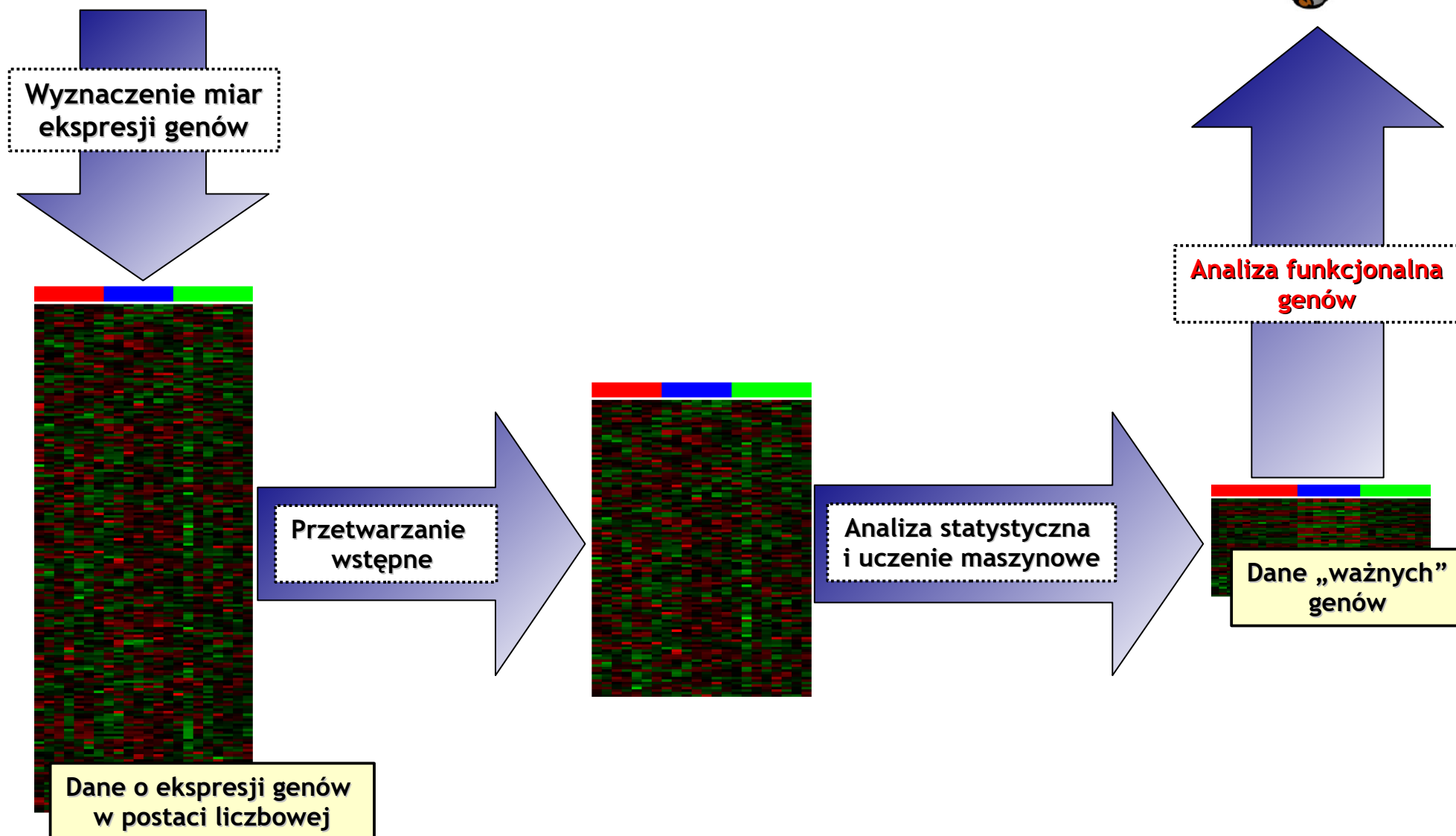
Tymon Rubel

**Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW**

Przetwarzanie danych z mikromacierzy ekspresyjnych

Określenie funkcji biologicznej zbioru genów wybranych na podstawie wcześniejszych etapów analizy danych z pomiarów mikromacierzowych.

Odbyna się ona przy wykorzystaniu baz danych gromadzących informacje o genach i białkach oraz w oparciu o dane literaturowe i wiedzę ekspercką.

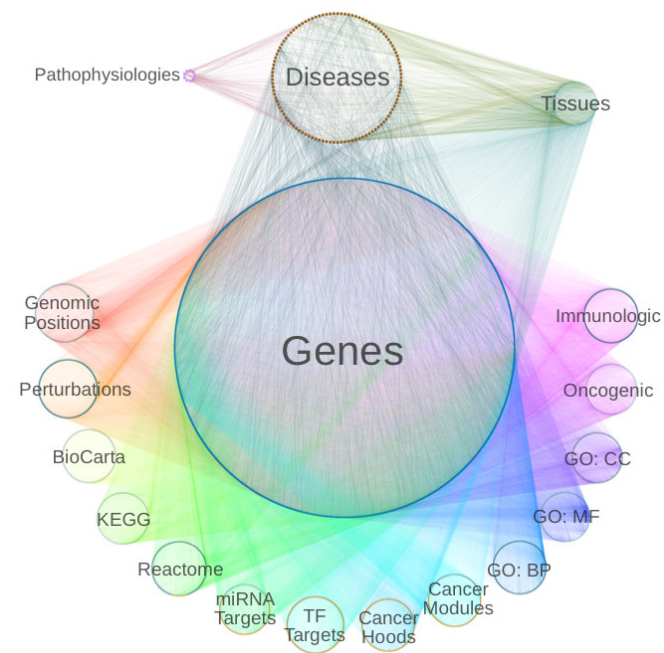


Analiza funkcjonalna list genów

W kontekście technik wielkoskalowych, w tym mikromacierzy, **analiza funkcjonalna** (*functional analysis*) może być definiowana jako ogół czynności automatyzujących i ułatwiających nadanie interpretacji biologicznej wynikom wcześniejszych etapów przetwarzania danych.

Punktem wyjścia dla tej analizy są listy genów lub białek uznanych za "interesujące" z punktu widzenia układu eksperymentalnego, zaś jej celem może być m. in.:

- grupowanie genów zgodnie z ich funkcją bądź udziałem produktów ich ekspresji w wybranych procesach biochemicznych, ścieżkach metabolicznych, itd.;
- wyznaczanie nadreprezentowanych grup funkcjonalnie powiązanych genów;
- tworzenie, wizualizacja i interpretacja sieci interakcji pomiędzy genami;
- ...



Analiza funkcjonalna list genów: bioinformatyczne bazy danych

Podstawą analizy funkcjonalnej wyników badań mikromacierzowych są informacje gromadzone przez liczne **bioinformatyczne bazy danych**.

Sekwencje nukleotydowych DNA (genów lub całych genomów):

- **GenBank** (www.ncbi.nlm.nih.gov/nucleotide)
- **DDBJ** (www.ddbj.nig.ac.jp)
- ...

Sekwencje i struktury cząstek RNA:

- **Rfam** (rfam.sanger.ac.uk)
- ...

Sekwencje i struktury białek:

- **UniProt** (www.uniprot.org)
- **PDB** (www.ebi.ac.uk/pdbe)
- ...

Metabazy (serwisy integrujące informacje z wielu baz danych):

- **Entrez** (www.ncbi.nlm.nih.gov/sites/gquery)
- **GeneCards** (www.genecards.org)
- ...

Analiza funkcjonalna list genów: bioinformatyczne bazy danych

Podstawą analizy funkcjonalnej wyników badań mikromacierzowych są informacje gromadzone przez liczne **bioinformatyczne bazy danych**.

Słowniki kontrolowane, ontologie, mapy szlaków metabolicznych:

- **Gene Ontology (GO)** (www.geneontology.org)
- **KEGG PATHWAY** (www.genome.jp/kegg/pathway.html)
- ...

Narzędzia tłumaczące pomiędzy identyfikatorami różnych baz danych:

- **UniProt ID mapping** (www.uniprot.org, zakładka ID Mapping)
- **DAVID** (david.abcc.ncifcrf.gov/conversion.jsp)
- ...

Bazy danych publikacji:

- **PubMed** (www.ncbi.nlm.nih.gov/pubmed/)
- ...

Analiza funkcjonalna list genów: bioinformatyczne bazy danych

Podstawą analizy funkcjonalnej wyników badań mikromacierzowych są informacje gromadzone przez liczne **bioinformatyczne bazy danych**.

Słowniki kontrolowane, ontologie, mapy szlaków metabolicznych:

- **Gene Ontology (GO)** (www.geneontology.org)
- **KEGG PATHWAY** (www.genome.jp/kegg/pathway)
- ...

Znaczna część baz danych gromadzi możliwie najpełniejsze informacje o pojedynczych obiektach (genach, białkach,...), w mniejszym zaś stopniu skupia się na występujących pomiędzy nimi relacjach.

Narzędzia tłumaczące pomiary

- **UniProt ID mapping** (www.uniprot.org)
- **DAVID** (david.abcc.ncifcrf.gov)
- ...

Istnieją jednak specjalizowane bazy danych, nastawione na modelowanie związków funkcjonalnych pomiędzy produktami ekspresji genów (np. ich wspólnego uczestnictwa w szlakach metabolicznych, kolokalizacji w komórkach, ...).

Bazy danych publikacji:

- **PubMed** (www.ncbi.nlm.nih.gov/pubmed/)
- ...

Uczenie maszynowe w bioinformatyce

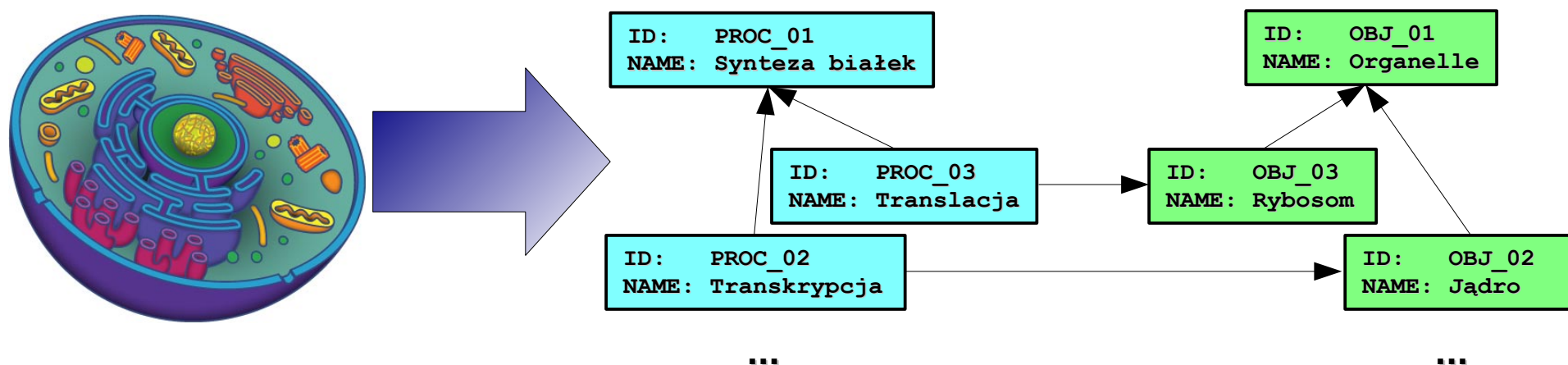
Wykład 12: ontologie i adnotacje genów

Tymon Rubel

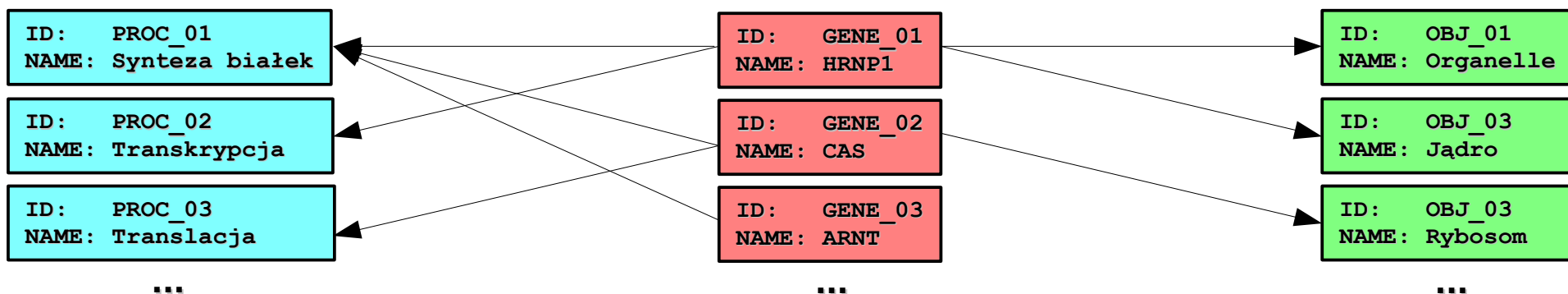
**Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW**

Analiza funkcjonalna list genów: ontologie i adnotacje

Ontologia (w znaczeniu bioinformatycznym) to uporządkowany zestaw zrozumiałych dla komputera definicji obiektów i procesów biologicznych oraz zachodzących pomiędzy nimi relacji.



Do wykonania analizy funkcjonalnej w sposób automatyczny, oprócz ontologii, konieczne są jeszcze **adnotacje** określające powiązania produktów ekspresji genów ze zdefiniowanymi wcześniej obiektami i procesami.



Analiza funkcjonalna list genów: Gene Ontology

Bazą danych ściśle ukierunkowaną na analizę funkcjonalną genów i białek jest **Gene Ontology (GO)**, która definiuje ontologię umożliwiającą opis obiektów biologicznych oraz procesów zachodzących na poziomie komórkowym z uwzględnieniem występujących pomiędzy nimi relacji.

Rekordami bazy danych (dostępnej pod adresem www.geneontology.org) są tzw. **kategorie GO (GO terms)**, z których każda definiowana jest poprzez:

- unikalny identyfikator (`id`);
- nazwę (`name`);
- opis słowny (`def`);
- subontologię, do której należy (`namespace`);
- rodzaj relacji z innymi kategoriami GO (`is_a`, `part_of`, `regulates`).

Przykładowy rekord bazy GO (format OBO)

[Term]

id: GO:0032452

name: histone demethylase activity

namespace: molecular_function

def: "Catalysis of the removal of a methyl group from a histone." [GOC:mah]

is_a: GO:0032451 ! demethylase activity

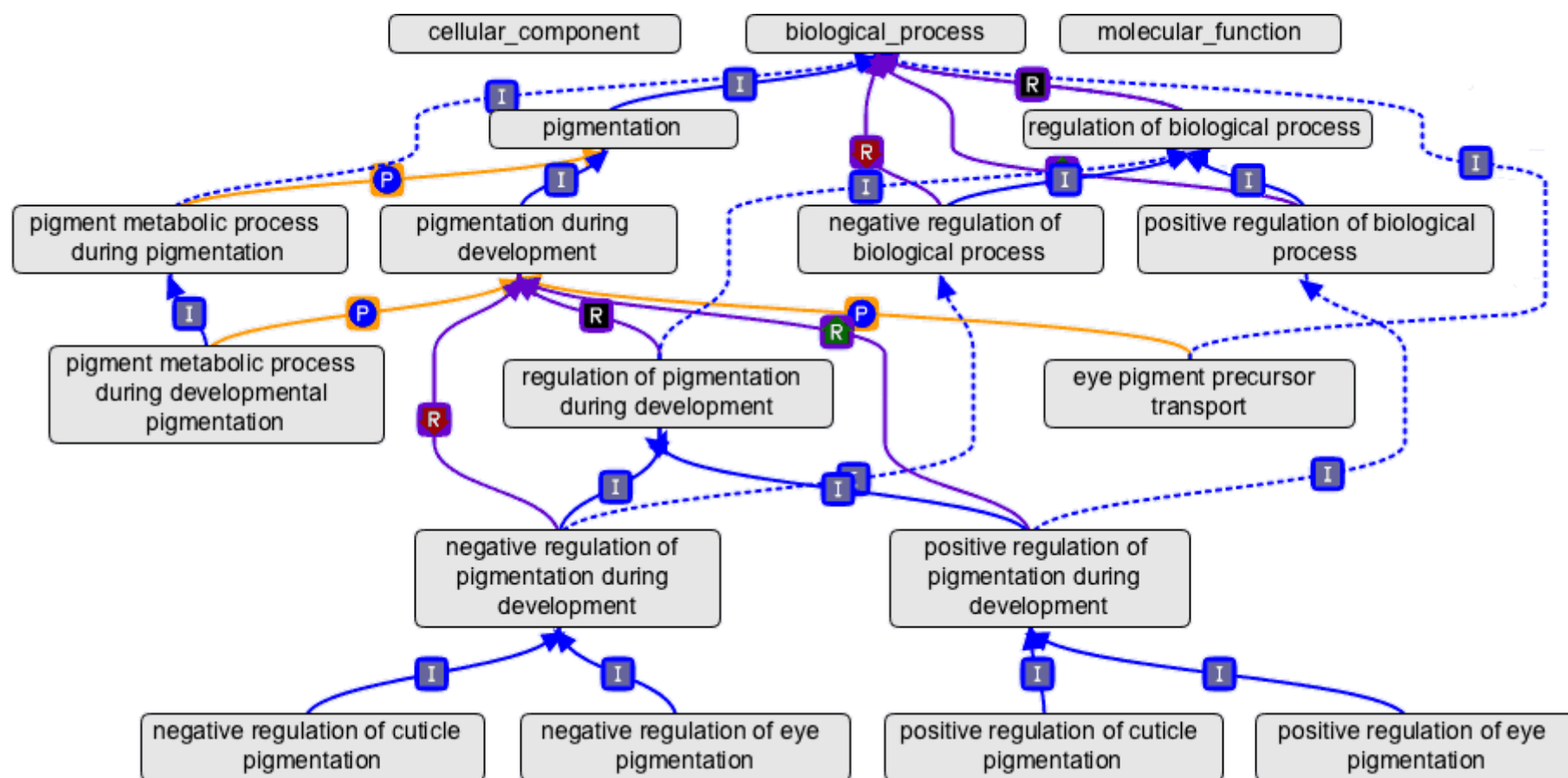
Analiza funkcjonalna list genów: Gene Ontology

Zdefiniowane w GO kategorie podzielone są na trzy główne **domeny (subontologie)**, reprezentujące zbiory pojęć niezbędnych do opisu właściwości produktów genów w sposób niezależny od rodzaju organizmu:

- funkcja produktu ekspresji genu (**MF – Molecular Function**), rozumiana jako rodzaj jego aktywności biochemicznej lub sposób działania (np. enzym, cząsteczka transportująca);
- proces biologicznych (**BP – Biological Process**), w którym bierze udział produkt genu (np. regulacja cyklu komórkowego, przekazywanie sygnałów). W procesie może uczestniczyć jednocześnie wiele produktów genów pełniących różne funkcje molekularne;
- lokalizacja komórkowa (**CC – Cellular Component**), czyli obszar występowania produktów genów w obrębie komórki (np. jądro, cytoplazma).

Analiza funkcjonalna list genów: Gene Ontology

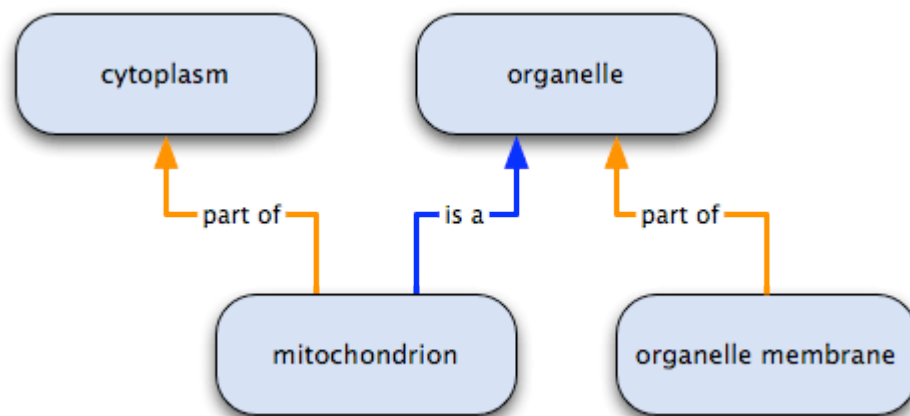
Każda z domen zorganizowana jest w sposób hierarchiczny, jako **acykliczny graf skierowany**, którego **węzły reprezentują kategorie GO**, a **krawędzie oznaczają relacje** występujące pomiędzy nimi. Każdy z węzłów może mieć jednego lub większą liczbę przodków (z wyjątkiem głównych węzłów domen, które nie mają przodków) i jednocześnie może – choć nie musi – posiadać dowolną liczbę węzłów potomnych.



Analiza funkcjonalna list genów: Gene Ontology

Podstawowymi rodzajami relacji pomiędzy kategoriami GO są:

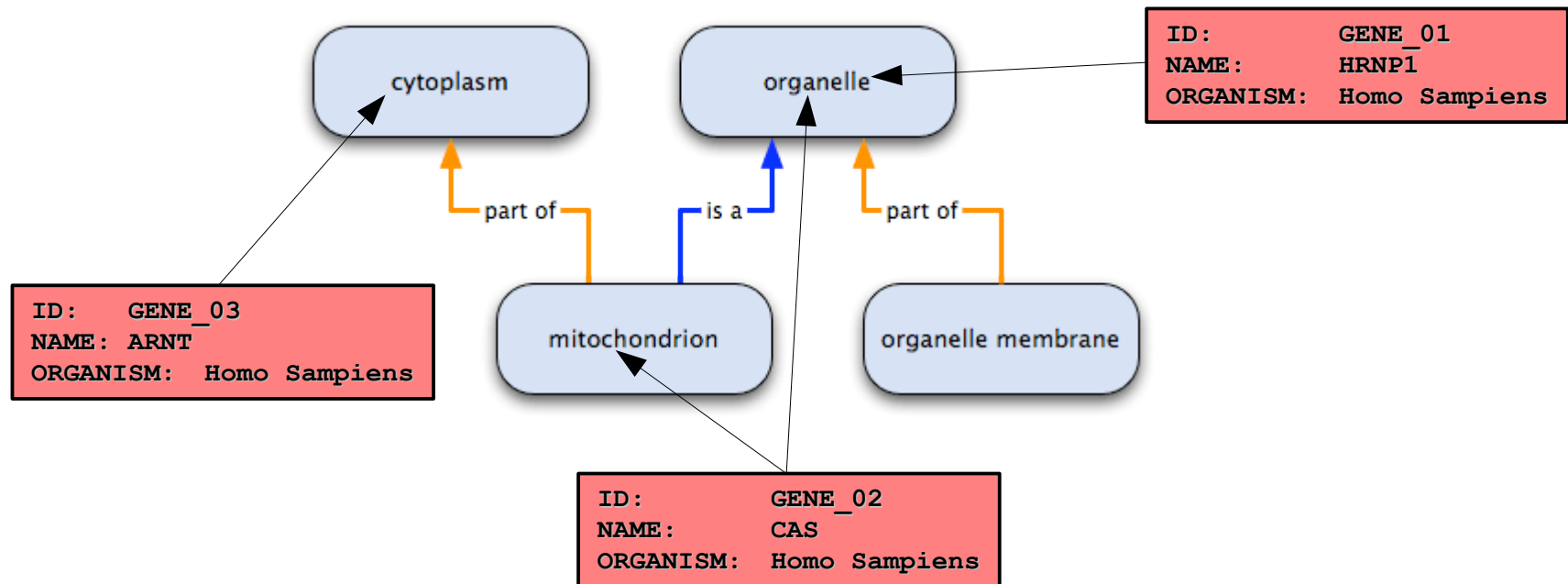
- **is_a (jest podtypem)**: relacja **A is_a B** oznacza, że kategoria **A** jest podtypem kategorii **B**. Przykładowo: kategoria „mitochondrium” jest podtypem kategorii „organelle komórkowe”;
- **part_of (jest częścią)**: relacja **A part_of B** oznacza, że **A** jest zawsze częścią **B** (**A** nie może istnieć bez **B**, ale nie wszystkie **B** muszą koniecznie zawierać **A**). Na przykład „widelki replikacyjne” zawsze występują w „chromosomach”, jednak nie wszystkie „chromosomy” muszą aktualnie posiadać „widelki replikacyjne”;



- **regulates (reguluje)**: relacja **A regulates B** oznacza, że **A** zawsze bezpośrednio wpływa na **B** (jednak **B** nie zawsze musi być regulowany przez **A**), np. „regulacja apoptozy” reguluje „śmierć komórki”;

Analiza funkcjonalna list genów: Gene Ontology

Oprócz samej ontologii, w ramach projektu Gene Ontology tworzone są również adnotacje genów do poszczególnych kategorii GO.

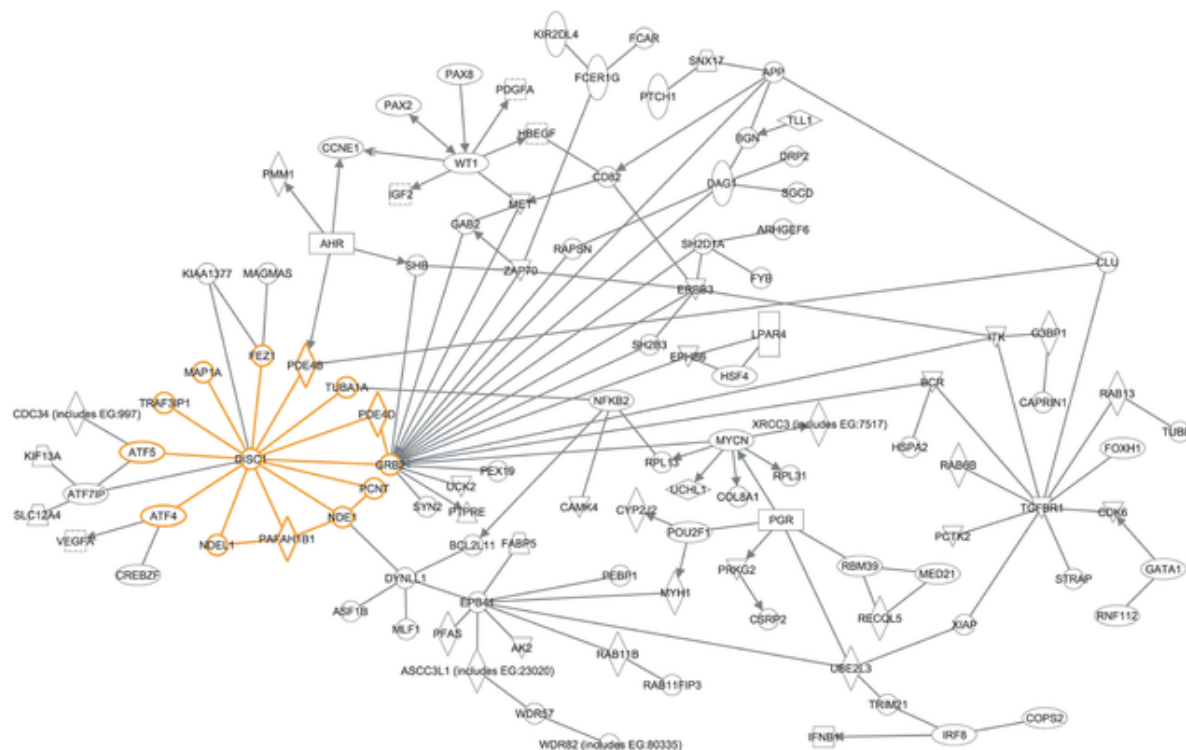


Pliki adnotacji tworzone są osobno dla poszczególnych gatunków organizmów i mogą zawierać definicje genów zgodne z różnymi bazami danych. To właśnie te pliki umożliwiają wykorzystanie informacji zawartych w Gene Ontology podczas analizy funkcjonalnej wyników eksperymentów mikromacierzowych.

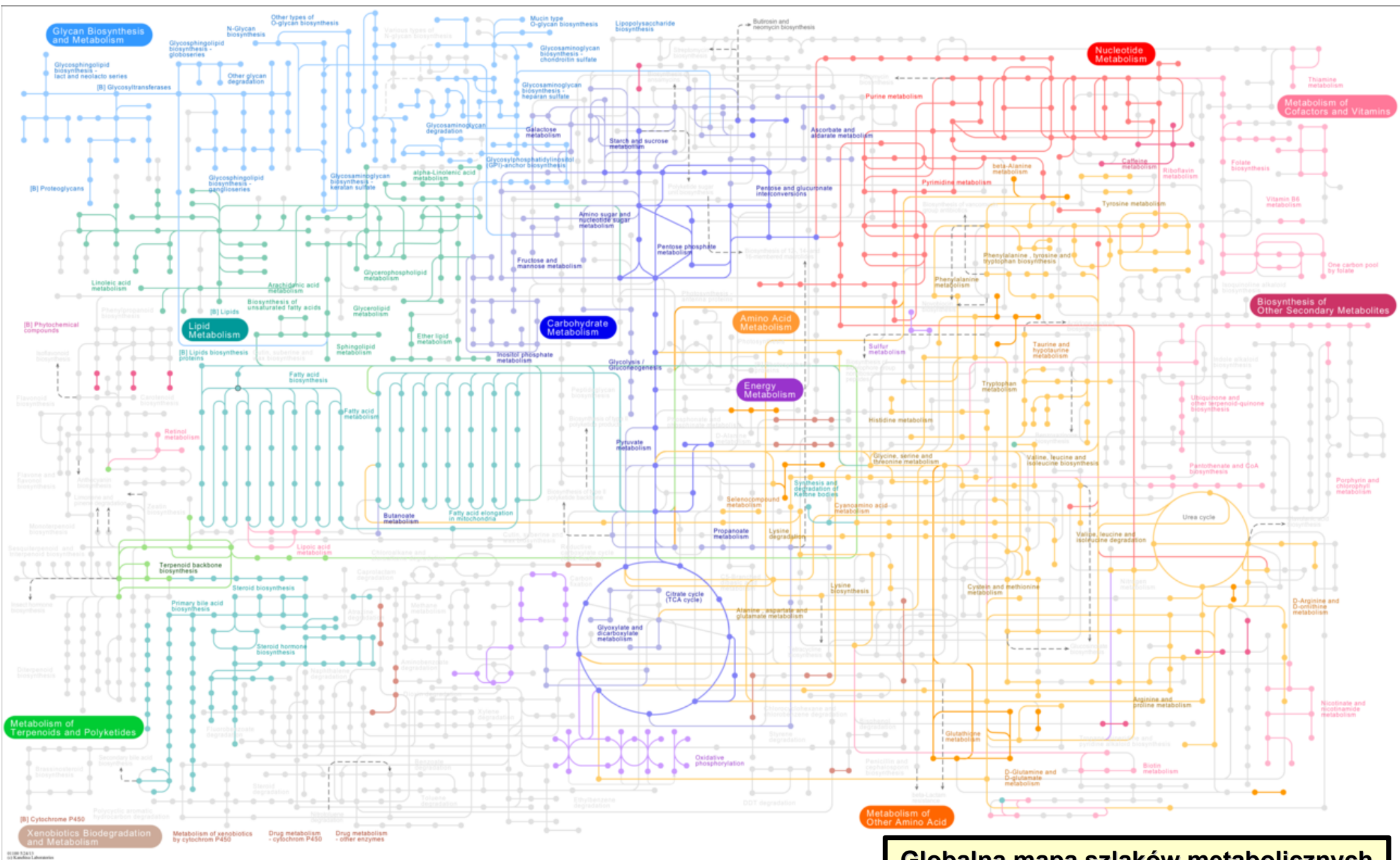
Analiza funkcjonalna list genów: baza danych KEGG PATHWAY

Innym przykładem bazy danych często wykorzystywanej w analizie funkcjonalnej jest **KEGG PATHWAY** (www.genome.jp/kegg/pathway.html). Stanowi ona część zbioru biologicznych baz danych KEGG (*Kyoto Encyclopedia of Genes and Genomes*), powstałych w ramach japońskiego programu sekwencjonowania ludzkiego genomu.

Informacje zgromadzone w bazie KEGG PATHWAY są odzwierciedleniem aktualnego stanu wiedzy na temat **sieci interakcji** pomiędzy produktami ekspresji genów zachodzących w ramach szlaków metabolicznych, podczas przetwarzania informacji genetycznej, sygnalizacji komórkowej lub w innych procesach molekularnych. Baza zawiera również sieci interakcji dla wielu schorzeń i leków.



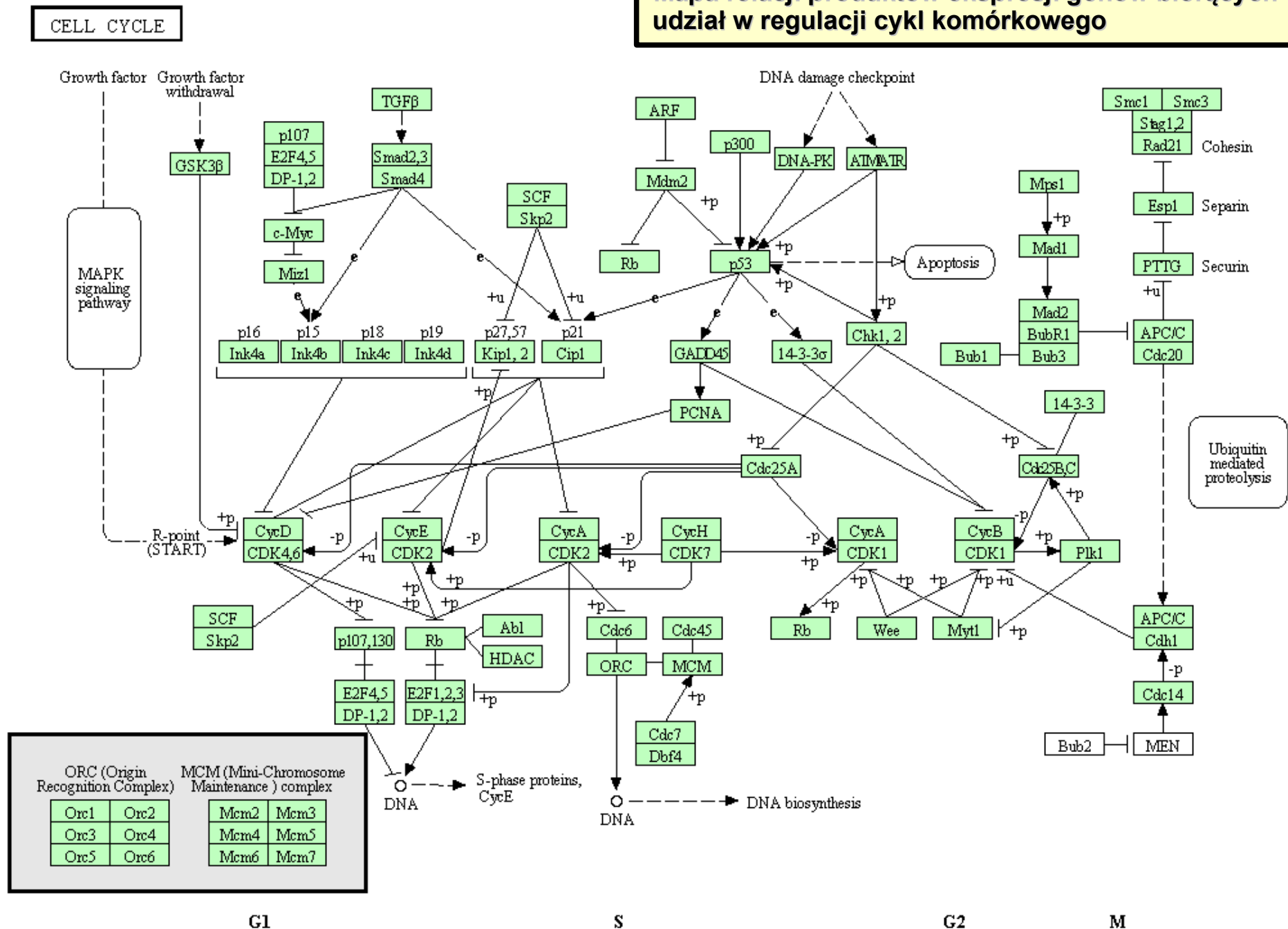
Analiza funkcjonalna list genów: baza danych KEGG PATHWAY



Globalna mapa szlaków metabolicznych w komórkach ludzkich

Analiza funkcjonalna list genów: baza danych KEGG PATHWAY

Mapa relacji produktów ekspresji genów biorących udział w regulacji cyklu komórkowego



Uczenie maszynowe w bioinformatyce

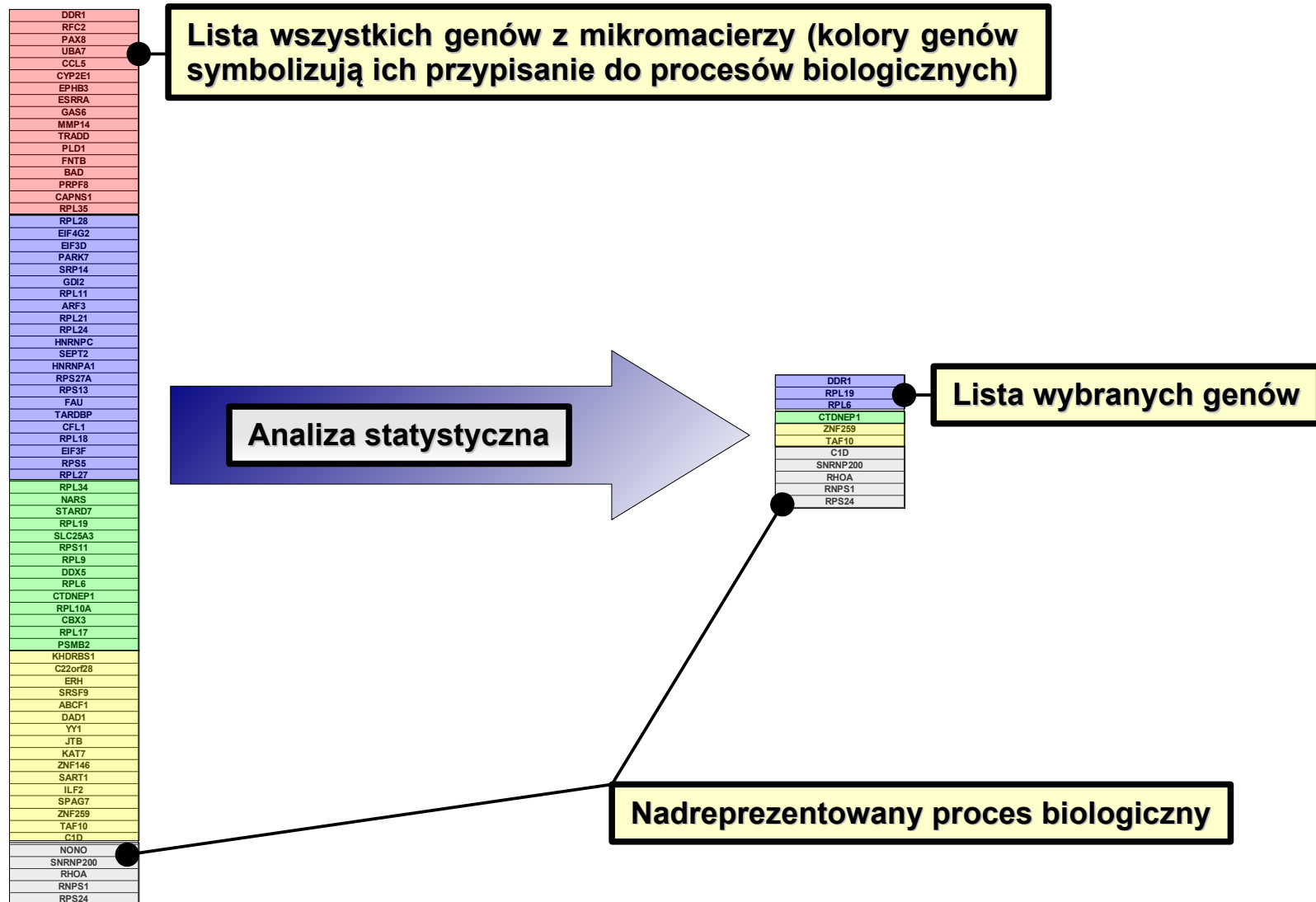
Wykład 11: badanie nadreprezentacji genów lub białek

Tymon Rubel

Zakład Elektroniki Jądrowej i Medycznej
Instytut Radioelektroniki i Technik Multimedialnych PW

Analiza funkcjonalna list genów: badanie nadreprezentacji grup genów

W przypadku mikromacierzy typowym celem analizy jest wyznaczenie grup powiązanych ze sobą funkcjonalnie genów (np. uczestniczących w tym samym procesie biologicznym), które są w istotny sposób nadreprezentowane w zbiorze genów wytypowanych na wcześniejszym etapie przetwarzania danych (np. wybranych technikami selekcji cech różnicujących badane grupy).



Analiza funkcjonalna list genów: badanie nadreprezentacji grup genów

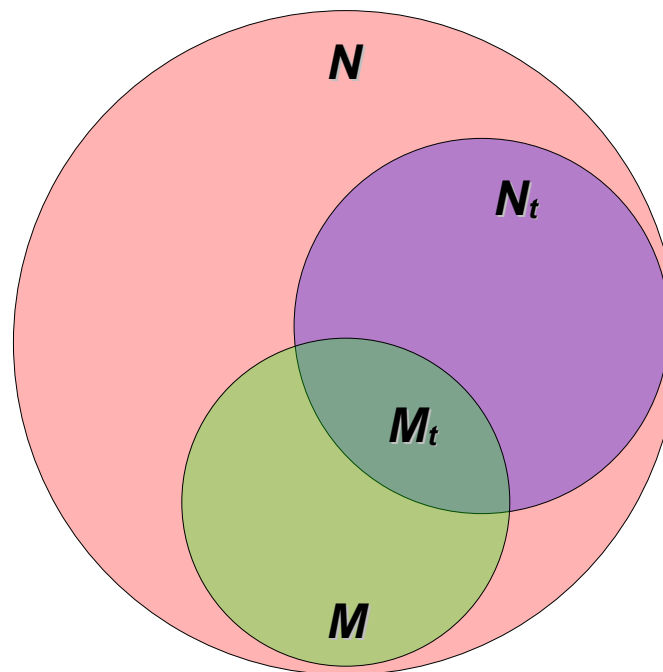
Dla każdego z badanych procesów biologicznych (np. opisywanych przez kolejne kategorie GO) można wprowadzić oznaczenia:

N – liczebność całkowitej populacji genów, czyli zbioru genów reprezentowanych przez wszystkie zestawy sondy mikromacierzy

M – liczba wytypowanych genów (np. genów różnicujących dwie grupy próbek)

N_t – liczba genów związanych z danym procesem biologicznym (kategorią GO) w populacji całkowitej

M_t – liczba genów związanych z danym procesem biologicznym (kategorią GO) w zbiorze tych wytypowanych (różnicujących)



Analiza funkcjonalna list genów: badanie nadreprezentacji grup genów

Badana jest hipoteza zerowa H_0 mówiąca, że dany proces biologiczny (np. opisywany przez wybraną kategorię GO) nie jest nadreprezentowany wśród genów wybranych jako różnicujące, wobec alternatywnej H_1 , która zakłada nadreprezentację genów związanych z danym procesem w zbiorze genów różnicujących.

Przy prawdziwości hipotezy zerowej problem może być opisany jako losowanie bez zwracania M genów z populacji o liczebności N , w której N_t genów jest przypisanych do rozpatrywanego procesu biologicznego (kategorii GO). Prawdopodobieństwo wylosowania dokładnie M_t genów związanych z tym procesem dane jest z rozkładu hipergeometrycznego o parametrach N , N_t i M :

$$P(x = M_t) = \frac{\binom{N_t}{M_t} \binom{N - N_t}{M - M_t}}{\binom{N}{M}} \qquad \binom{N}{M} = \frac{N!}{M!(N-M)!}$$

Hipotezę zerową odrzucamy (uznajemy, że nie można uzyskać losowo tak dużej wartości M_t) jeżeli prawdopodobieństwo wylosowania co najmniej M_t genów związanych z danym procesem:

$$P(x \geq M_t) = \sum_{i=M_t}^{\min(N_t, M)} \frac{\binom{N_t}{i} \binom{N - N_t}{M - i}}{\binom{N}{M}}$$

okaże się mniejsze od wybranego poziomu istotności α .