

Bioinformatyka

Andrzej Łyskowski, dr inż.
Katedra Biotechnologii i Bioinformatyki
andrzej.lyskowski@prz.edu.pl
H-237

analiza sekwencji

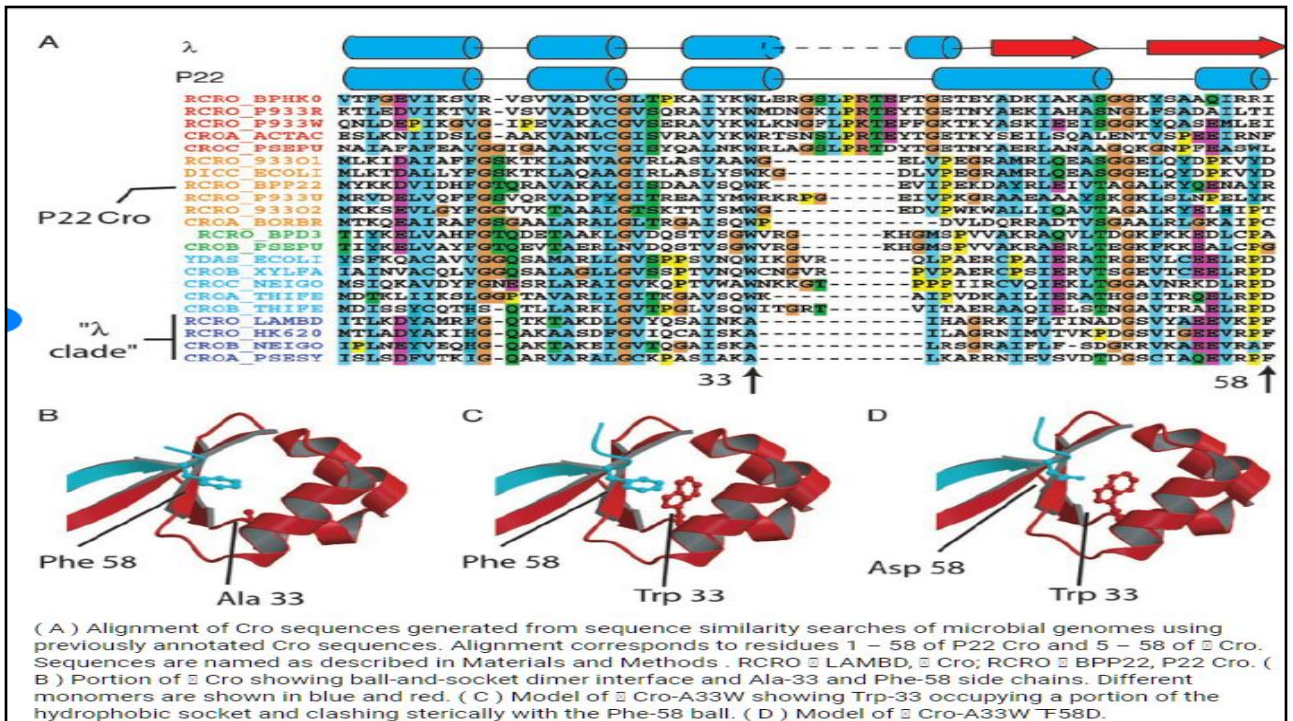
- wyszukiwanie sekwencji podobnych
- przyrównanie/dopasowanie sekwencji
- algorytmy wykorzystywane w przyrównaniu/dopasowaniu sekwencji
- parametry oceny dopasowań sekwencji
- ocena zgodności sekwencji
- podobieństwo i homologia sekwencji

wyszukiwanie sekwencji podobnych

Wyszukiwanie sekwencji w bazach danych pozwala odnaleźć sekwencje podobne do sekwencji bazowej (*ang.: query*).

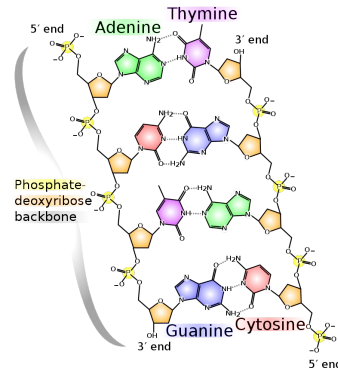
Zidentyfikowane sekwencje na zasadzie podobieństwa pozwalają na przewidywanie funkcji i struktury nieznanej sekwencji.

Przewidywanie właściwości na podstawie podobieństwa dwóch lub więcej sekwencji stanowi potężne narzędzie i podstawę analizy bioinformatycznej.



przyrównanie/dopasowanie sekwencji

- Każda para sekwencji DNA będzie wykazywać pewien stopień dopasowania.
- W trakcie analizy konieczne jest rozróżnienie dopasowania losowego od faktycznej ewolucyjnej zgodności sekwencji



dopasowanie sekwencji

- Pierwszym krokiem analizy jest dopasowanie lub przyrównanie sekwencji względem siebie
- Przyrównanie odczytuje się od lewej do prawej
- Interpretacja wymaga oznaczenia
 - Pozycji zgodnych
 - Pozycji niezgodnych
 - Przerw
- Interpretację przeprowadza się w kontekście zmian ewolucyjnych

Unaligned sequences

Human	ACA	TATGGACA	GTAA	TAAGTAAAAA	ACATAT
Chimpanzee	ACA	TATGGACA	GTAA	TAAGTAAAAA	ACATAT
Macaque	ATA	ACATTACG	ACAG	TAAGTAAAAA	ACAT

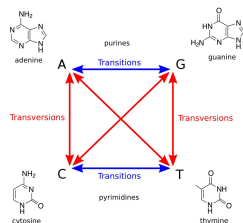
Aligned sequences

Human	ACA		TATGGACA	GTAA	TAAGTAAAAA	ACATATT
Chimpanzee	ACA		TATGGACA	GTAA	TAAGTAAAAA	ACATATT
Macaque	ATA	TACA	TACG	ACAG	TAAGTAAAAA	ACAT

zmiany ewolucyjne

Insercja (interkalacja)

najczęściej spontaniczna mutacja genu polegająca na wstawieniu krótkiej sekwencji DNA w obrębie pojedynczego genu albo wstawieniu dłuższego fragmentu chromosomu.



Delecja

zmiana w materiale genetycznym polegająca na utracie jego fragmentu. Ubytek taki może obejmować od jednego lub kilku nukleotydów (mutacja punktowa), poprzez ubytek większych fragmentów genów, całych genów lub ich grup (strukturalna aberracja chromosomowa), aż po całe chromosomy (liczbowa aberracja chromosomowa).

Unaligned sequences

Human	A	C	A	T	T	A	T	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	A	C	A	T	A	T	T	
Chimpanzee	A	C	A	T	T	A	T	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	A	C	A	T	A	T	T	
Macaque	A	T	A	T	A	C	A	T	T	A	C	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	C	A	T	T

Aligned sequences

Human	A	C	A		T	T	A	T	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	A	A	C	A	T	A	T	T
Chimpanzee	A	C	A		T	T	A	T	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	A	A	C	A	T	A	T	T
Macaque	A	T	A	T	A	C	A		T	T	A	C	G	G	A	C	A	G	G	T	A	A	G	T	A	A	A	A	A	C	A	T

Dopasowanie sekwencji

Warianty:

- identyczne sekwencje
- sekwencje z błędami
(ang. *match & mismatch*)
- insercje i delecje
- konserwacja

```
ATGGCATGCCTGATC
|||||
ATGGCATGCCTGATC
```

```
MSTPAGSDQERMILV
|||||
MSTPAGSDQERMILV
```

```
ACGGCTTACCTGGCC
|||||
ATGGCATGCCTGATC
```

```
MGTPAASQERMSTV
|||||
MSTPAGSDQERMILV
```

```
ATGGC-TGCCTGATC
|||||
ATGGCATGCC-GATC
```

```
MSTPA-SDWERMILV
|||||
MSTPAGSDQE-MILV
```

```
MSKMAGSNVERMILV
|||.|||.:.|||.
MSRVAGSDIERMIMV
```

parametry oceny dopasowań sekwencji ocena zgodności sekwencji

- Interpretacja wymaga oznaczenia
 - Pozycji zgodnych
 - Pozycji niezgodnych
 - Przerw

1|2 31/31 - 100% zgodności

1|3 22/31 - 70% zgodności

Kluczowym elementem oceny jest przyznanie
odpowiedniej punktacji przerwom:

- Za otwarcie
- Za wydłużenie lub
- Proporcjonalnie do długości

Unaligned sequences

Human	ACA	TATGGACAG	GTAAGTAAAAACATAT
Chimpanzee	ACA	TATGGACAG	GTAAGTAAAAACATAT
Macaque	ATA	ACATTACG	ACAGTAAAGTAAAAACA

Aligned sequences

Human	ACA	TATGGACAGGTAASTAAAAACATATT
Chimpanzee	ACA	TATGGACAGGTAASTAAAAACATATT
Macaque	ATA	TACAGGACAGGTAASTAAAAACAT

Ocena dopasowania



AGGCTCATGGT
GCCTCTGTGGAA

Nucleotide alignment 1



	A	G	T	C
A	1	0	0	0
G	0	1	0	0
T	0	0	1	0
C	0	0	0	1

Identity matrix



Score: 3

AGGCTCATGGT
GCCTCTGTGGAA

Nucleotide alignment 2



Score: 7

b

- Interpretacja wymaga oznaczenia
 - Pozycji zgodnych
 - Pozycji niezgodnych
 - Przerw
- Kluczowym elementem oceny jest przyznanie odpowiedniej punktacji przerwom:
 - Za otwarcie
 - Za wydłużenie lub
 - Proporcjonalnie do długości

algorytmy wykorzystywane w przyrównaniu/dopasowaniu sekwencji

- Odnalezienie odpowiedniego dopasowania sekwencji wymaga zastosowanie zaawansowanych technik obliczeniowych, np.: programowania dynamicznego.

Programowanie dynamiczne opiera się na podziale rozwiązywanego problemu na podproblemy względem kilku parametrów. W odróżnieniu od techniki dziel i zwyciężaj podproblemy w programowaniu dynamicznym nie są rozłączne, ale musi je cechować własność optymalnej podstruktury. Zagadnienia odpowiednie dla programowania dynamicznego cechuje również to, że zastosowanie do nich metody siłowej (ang. brute force) prowadzi do ponadwielomianowej liczby rozwiązań podproblemów, podczas gdy sama liczba różnych podproblemów jest wielomianowa.

algorytm Needlemana-Wunscha

oparty na programowaniu dynamicznym, umożliwiający znalezienie optymalnego globalnego dopasowania dwóch sekwencji.

Jest często wykorzystywany w bioinformatyce jako jedno z narzędzi do poszukiwania ułiniowienia sekwencji nukleotydowych lub aminokwasowych.

Został stworzony przez Saul B. Needleman'a i Christian D. Wunsch'a oraz opublikowany w roku 1970. Dzieli on większy problem obliczeniowy (np. całą sekwencję) na mniejsze problemy, i używa rozwiązań mniejszych problemów do znalezienia optymalnego rozwiązania dużego problemu.

Needleman-Wunsch

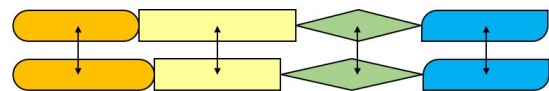
match = 1 mismatch = -1 gap = -1

	G	C	A	T	G	C	U	
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	-1	-2	-3	-4
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

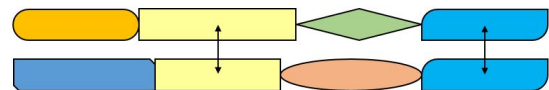
algorytm Smith-Waterman

bazujący na programowaniu dynamicznym umożliwiający poszukiwanie optymalnych lokalnych dopasowań sekwencji.

Algorytm został opracowany przez Temple F. Smith i Michael S. Waterman w 1981 r.

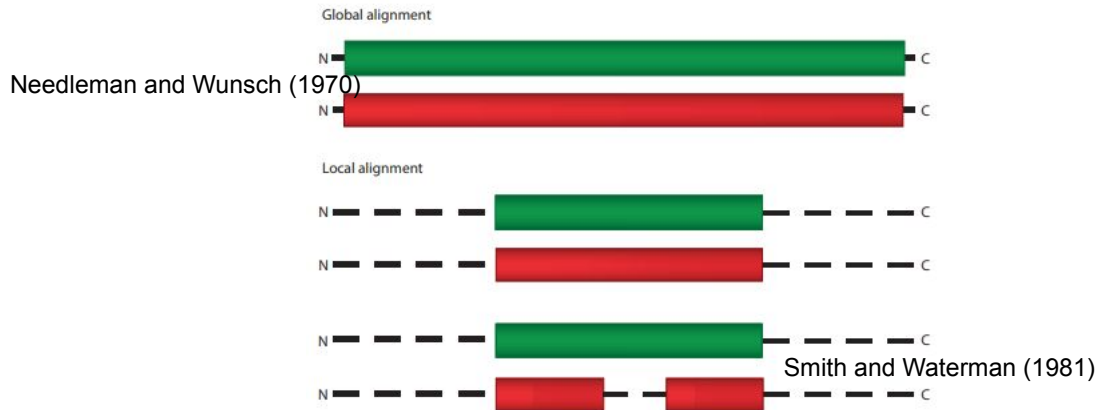


Global Alignment



Local Alignment

Typy dopasowań



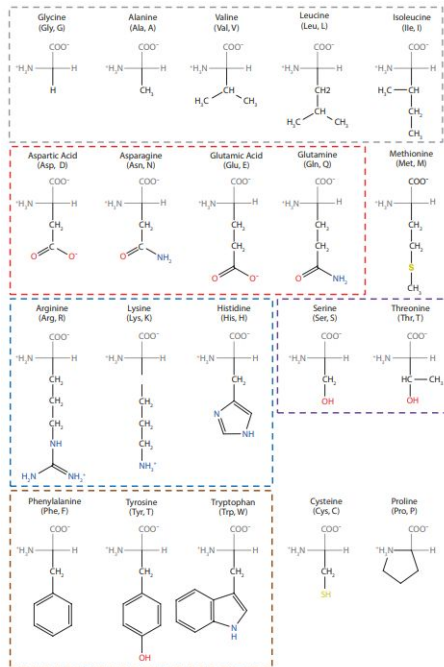
podobieństwo i homologia sekwencji

- Każda para sekwencji będzie wykazywać pewien poziom podobieństwa.
- Podobieństwo nie świadczy o wspólnym pochodzeniu lub właściwościach – może mieć ono charakter losowy.

- Sekwencje homologiczne to sekwencje pochodzące od wspólnego przodka.

*Dwie lub więcej sekwencji nigdy nie będzie homologiczne w 50%.
Dwie lub sekwencji może posiadać 50 % podobieństwa co może świadczyć o homologii pomiędzy nimi.*

- Ortologi
dwa homologiczne geny występujące w odrębnych gatunkach posiadające te same funkcje.
- Paralogi
dwa homologiczne geny u tego samego (lub odrębnego) gatunku posiadające różne funkcje.



Dopasowanie sekwencji aminokwasowych

GMIDLITARCAYPSWTGH
IEVRLAKCAYPGWSGHY

GMIDLITARCAYPSWTGH
IEVRLAKCAYPGWSGHY

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	8	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3	
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	2	2	4	3	0	-1	-2	
V	-1	-2	-2	-3	0	-3	-3	-3	-2	-2	-3	-2	1	3	1	4	-1	-1	-3	
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	-2	-2	-1	-1	-1	-1	3	7	2	
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-3	-2	-3	-3	-1	-3	-2	-3	1	2	11

Score: 65

BLOSUM62 matrix

Score: 19

Identyfikacja homologów

Homologia przejawia się często znaczącym podobieństwem w:

- sekwencji nukleotydowej;
- sekwencji aminokwasowej;
- strukturze przestrzennej.

Problem: Identyfikacja w oparciu o którą cechę będzie najwydajniejsza?

Table 1.1 The size of genomes

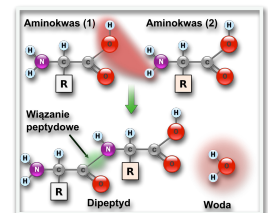
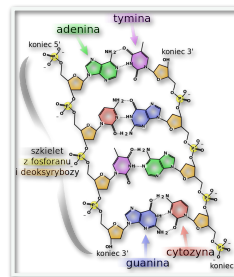
Species	Genome size (10 ⁹ nucleotides)	Number of genes
<i>Escherichia coli</i>	4,7	4300
<i>Saccharomyces cerevisiae</i>	12	6700
<i>Drosophila melanogaster</i>	169	13,900
<i>Danio rerio</i>	1500	26,000
<i>Homo sapiens</i>	3200	21,000
<i>Zea mays</i>	3200	63,000
<i>Oryza sativa</i>	488	57,000

Source: The Ensembl Genome Browser (www.ensembl.org) April 2012.
Table 1.1 Practical Bioinformatics (© Garland Science 2013)

Przyrównywanie sekwencji...

Klasyfikacja przyrównań (*ang. sequence alignment*):

- nukleotydowej:
cztery elementy porównawcze w tripletach.
- sekwencji aminokwasowej:
20 jednostek porównawczych.
- strukturze przestrzennej:
struktura I, II, III, IV rządowa definiująca przestrzenne ułożenie atomów.



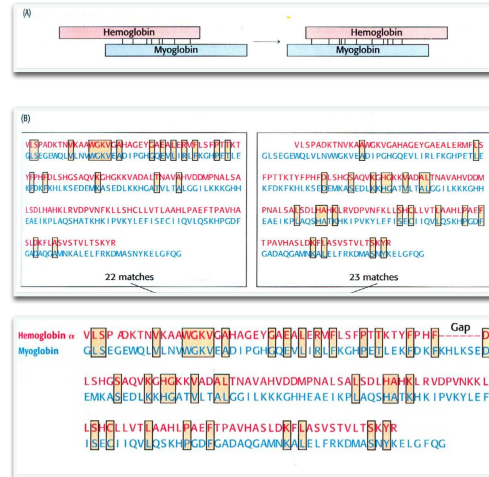
Przyrównanie doskonałe

1:1

Przesunięcie sekwencji względem siebie i obliczenie punktacji przyrównania (*ang. score*).

1:1 + przerwa (*ang. gap*)

Celem lepszego dopasowania wprowadza się w sekwencję przerwy celem kompensacji insercji lub delecji nukleotydów.



Zmiany ewolucyjne

Mutacje (zmiana tożsamości, usunięcie lub dodanie aminokwasu/domeny) na poziomie genetycznym zachodzą tylko w określonych warunkach i gdy nie wpływają na strukturę lub funkcję kodowanego białka

Problem: jak ocenić podobieństwo sekwencji uwzględniając zmiany ewolucyjne?

Kod DNA

Cechy kodu DNA:

- Trójkowy
- Niezachodzący
- Bezprzecinkowy
- Zdegenerowany
różne kodony umożliwiają kodowanie tego samego aminokwasu, tzn. prawie wszystkie aminokwasy mogą być zakodowane na kilka sposobów.
Część zmian w informacji genetycznej w wyniku mutacji nie znajduje swojego odbicia w sekwencji aminokwasów.
- Jednoznaczny
- Kolinearny
- Uniwersalny

AMINO ACIDS AND THEIR SYMBOLS	CODONS	MEMORY AID
A Ala Alanine	GCA GCC GCG GCT	Alanine
C Cys Cysteine	TGC TGT	Cysteine
D Asp Aspartic acid	GAC GAT	Aspartic acid
E Glu Glutamic acid	GAA GAG	Glutamic acid
F Phe Phenylalanine	TTC TTT	Phenylalanine
G Gly Glycine	GGA GGC GGG GGT	Glycine
H His Histidine	CAC CAT	Histidine
I Ile Isoleucine	ATA ATC ATT	Isoleucine
K Lys Lysine	AAA AAG	Lysine
L Leu Leucine	TTA TTG CTA CTC CTG CTT	Leucine
M Met Methionine	ATG	Methionine
N Asn Asparagine	AAC AAT	Asparagine
P Pro Proline	CCA CCC CCG CCT	Proline
Q Gln Glutamine	CAA CAG	Glutamine
R Arg Arginine	AGA AGG CGA CGC CGG CGT	Arginine
S Ser Serine	AGC AGT TCA TCC TCG TCT	Serine
T Thr Threonine	ACA ACC ACG ACT	Threonine
V Val Valine	GTA GTC GTG GTT	Valine
W Trp Tryptophan	TGG	Tryptophan
Y Tyr Tyrosine	TAC TAT	Tyrosine
X Any amino acid		

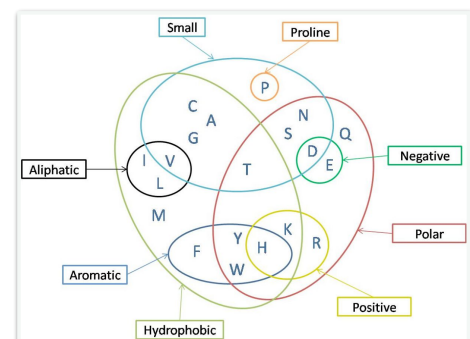
Figure 4.4 Practical Bioinformatics (© Garland Science 2013)

Ewolucja białek

Zmiany ewolucyjne mogą zachodzić w dwóch obszarach sekwencji:

- *kluczowym*, czyli obszarze odpowiedzialnym za konkretną funkcję biologiczną;
- *powłoce*, czyli obszarze odpowiedzialnym za uformowanie struktury białka ale nie biorącym bezpośredniego udziału w katalizie.

Zachodzące mutacje w obszarze kluczowym mają tendencję do zamiany aminokwasów na inne o podobnych właściwościach fizykochemicznych. Np. zamiana hydrofobowych aminokwasów w rdzeniu białka (Leu, Ile, Val).



Macierze substytucji

Problem: Jak przyrównać sekwencje zawierające mutacje?

Ocenie jakości przyrównania sekwencji białek zawierających mutacje służą macierze substytucji (*ang. substitution matrices*).

Służą one do oceny podobieństwa pomiędzy aminokwasami i obliczenia punktacji dla konkretnego wariantu przyrównania.

Uwaga: przerwa w sekwencji ma zawsze ujemną wartość punktową!

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	0	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-2	-1	1	5					
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	1	3	-3	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Figure 4.9 Practical Bioinformatics (© Garland Science 2013)

Typowe macierze

PAM (*ang. Accepted Point Mutation*)

PAM250 to macierz zawierająca punktację dla drogi ewolucyjnej równej 250 akceptowalnym mutacjom punktowym na 100 reszt aminokwasowych.

Macierze PAM oblicza się przez przyrównanie i porównanie zmian ewolucyjnych w blisko spokrewnionych białkach a następnie ekstrapoluje się te wartości na dalsze 'odległości' ewolucyjne.

pam (point accepted mutation)

[illegible]

PAM jest miarą zmian ewolucyjnych.

W praktyce, 1 PAM jest miarą zmian ewolucyjnych, w trakcie których 1% aminokwasów białkowego produktu analizowanego genu uległ zmianom.

Macierz PAM zostały wprowadzone przez Margaret Dayhoff w1978.

Wartości zawarte w macierzy zostały ustalone w oparciu o obserwację 1572 mutacji występujących w drzewach filogenetycznych 71 rodzin blisko spokrewnionych białek. Badane białka zostały wybrane jako w oparciu o ich wysokie podobieństwo ze swoimi przodkami. Wykorzystane dopasowania białek miały ponad 85% identyczności.

Typowe macierze

BLOSUM (*ang. BLOcks SUBstitutio Matrix*)

Punktację macierzy BLOSUM ustala się w oparciu o przyrównanie i porównanie segmentów sekwencji ewolucyjnie spokrewnionych białek.

Im mniejszy numer identyfikacyjny macierzy BLOSUM tym dłuższa odległość ewolucyjna.

blosum (blocks substitution matrix)

Macierze BLOSUM zostały opracowane w oparciu o analizę lokalnych dopasowań spokrewnionych fragmentów sekwencji.

W wyniku analizy spokrewnionych i zakonserwowanych sekwencji nie zawierających przerw w bazie danych BLOCKS ustalono częstotliwość występujących mutacji. Następnie dokonano obliczenia odpowiednich parametrów dla wszystkich potencjalnych 210 zmian dla wszystkich standardowych 20 aminokwasów.

Macierze BLOSUM zostały wprowadzone przez Steven Henikoff i Jorja Henikoff w 1992 r.

Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-1	0	0	-3	0	6							
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	7						
Ser	1	-1	1	0	-1	0	0	-1	-2	0	-1	-2	0	-1	4					
Thr	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
Trp	-3	-3	-4	-4	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Tyr	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

pam vs. blosum

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM62
PAM200	BLOSUM50
PAM250	BLOSUM45

BIOINFORMATYKA.1



wyszukiwanie danych w bazach danych

- przeszukiwanie baz danych
- BLAST i FASTA
- parametry oceny wyników wyszukiwania

przeszukiwanie baz danych

Wyszukiwanie oparte o homologię sekwencji jest metodą ekstrakcji danych ze specjalistycznych baz bioinformatycznych mające na celu odnalezienie i wyznaczenie statystycznie znaczącego podobieństwa pomiędzy sekwencją analizowaną a zgromadzonymi danymi.

W przeciwieństwie do algorytmów dynamicznych celem jest identyfikacja podobieństwo pomiędzy więcej niż dwoma sekwencjami.

fasta & blast

FASTA oraz BLAST pośrednie rozwiązanie problemu przeszukiwania dużych baz danych.

Obydwa pakiety oprogramowania nie gwarantują równie dokładnego jak w przypadku algorytmów dynamicznych dopasowania sekwencji ale pozwalają w zadowalającym czasie odnaleźć sekwencje homologiczne.

Zasadą działania jest:

- Podział sekwencji analizowanej na mniejsze fragmenty, słowa (ang.: words).
- Wyszukiwanie zgodności możliwie największej liczby słów pochodzących z analizowanej sekwencji z sekwencjami w bazie danych.
- Rozbudowę dopasowania poprzez rozszerzenie wstępnych przyrównań.

fasta

FASTA (ang.: FAST-All)

Umożliwia wyszukiwanie w bazach danych w parach:

- Białko:białko
- DNA:DNA
- Tłumaczona sekwencja białkowa:DNA

FASTA wykorzystuje zadaną sekwencję nukleotydową lub białkową do przeszukania bazy danych w celu znalezienia lokalnych dopasowań.

- Identyfikacja regionów o wysokim dopasowaniu. Kluczowym parametrem jest *kmer*.
- Ponowne skanowanie i dopasowywanie.
- Wstępne dopasowanie (*cutoff*).
- Finalne dopasowanie z wykorzystaniem algorytmów dynamicznych.

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLVSVNLLCQGVSTPVCNPGNGNCQVSLRDLFDRAVMVSHYIHDLSSEMFNEFDKRYAQKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSMNDPLYHLVTEVRGMKGAPDAILSRATEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDEDARYSAFYNLLHCLRRDSSKIDTYLKLNCRIIYNNNC*
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQTAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTAEALQDMINEVDADGNGTIDFPEFLTM*MARKMKDTSDEEIEAERFVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK*
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLVEWINGGFSVDKATLNRFFAFHILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYTIKDFLGLLILILLLLLLALLSPDMLGDPDNIHMPADPLNTPLHIKPEWYFLFAYAILRSPVNIKLGVLALFLSIVILGLMPFLHTSKHRSMMRLPLSQALFWLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPAGXIENY
```

blast

BLAST (ang.: basic local alignmet search tool)
pozwala na wyszukiwanie w parach:

- DNA:DNA (blastn)
- białko:białko (blastp)
- DNA tłumaczenie:białko (blastx)
- DNA tłumaczenie: DNA tłumaczenie (tblastx)
- białko: DNA tłumaczenie (tblastn)
- megablast
- PSI-BLAST

Algorytm BLAST:

- Analiza sekwencji w celu usunięcia regionów o niskim skomplikowaniu
- Wygenerowanie listy słów o zadanej długości *k*
- Wygenerowanie listy *high-scoring word(s)*
Główna różnica w stosunku do algorytmu FASTA.
- Sortowanie i analiza listy słów
- Identyfikacja *high-scoring pairs*
- Wygenerowanie i analiza HSP
- Łączenie fragmentów HSP i finalne dopasowanie

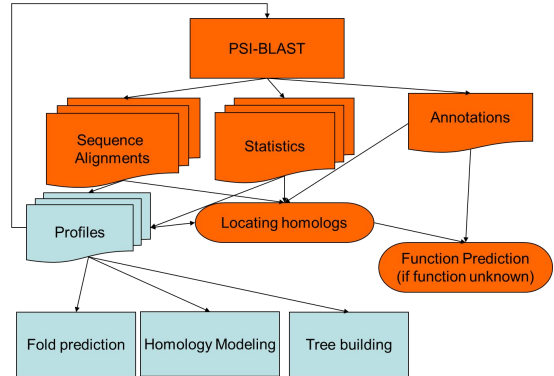


psi-blast

PSI-BLAST (ang.: position specific iterative BLAST)

W wyniku działania PSI-BLAST powstaje:

- „mini baza danych” zawierająca listę spokrewnionych sekwencji
- Zestaw dopasowań sekwencji
- Profile sekwencji
- Informacje statystyczne



parametry oceny wyników wyszukiwania

Parametr p - prawdopodobieństwo testowe
prawdopodobieństwo, że zależność jaką
zaobserwowano w losowej próbie z populacji mogła
wystąpić przypadkowo, wskutek losowej zmienności
prób, choć w populacji wcale nie występuje.
Jest to narzędzie służące jedynie do podstawowej
kontroli błędów, i świadczy o wartości dowodowej
danych jedynie pośrednio.

Według wielu przeglądów i komentarzy błędne interpretacje wartości p są w naukach powszechne

Wartość E odpowiada oczekiwanej liczbie sekwencji z bazy danych, których ocena dopasowania z sekwencją w zapytaniu (kwerendą) jest większa od obserwowanej oceny określonego dopasowania lub równa tej ocenie.

dopasowanie wielosekwencyjne

Dopasowanie większej liczby sekwencji (ang.: multiple sequence alignment) wykonywane jest zazwyczaj z użyciem metod progresywnego dopasowania wielosekwencyjnego opartego o algorytmy dynamiczne wzbogacone o niezbędne reguły rekurencyjne.

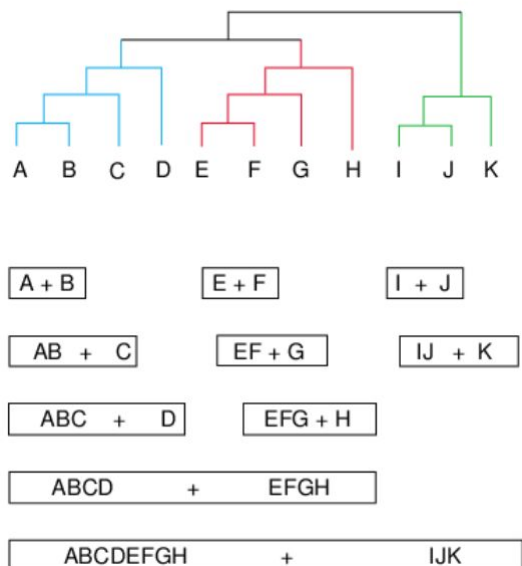
Procedura dopasowania rozpoczyna się od wygenerowania tzw. drzewa przewodniego (ang. guide tree).

[illegible]

drzewo przewodnie

Dla zbioru wszystkich sekwencji poddawanych dopasowaniu generowane są przyrównania lokalne.

Algorytm działa na zasadzie obliczania wyników podobieństwa jako liczby dopasowań k-tup między dwiema sekwencjami, biorąc pod uwagę ustaloną karę za przerwy. Im bardziej podobne sekwencje, tym wyższy wynik, im bardziej rozbieżne, tym niższe wyniki. Gdy sekwencje zostaną ocenione, generowany jest dendrogram w celu przedstawienia kolejności dopasowania wielu sekwencji. Wyższe uporządkowane zestawy sekwencji są wyrównane jako pierwsze, a następnie pozostałe w kolejności malejącej.



msa można wykonać na więcej niż jeden sposób

```
1 10 20 30 40 50 60 70 80 90 100 104
|-----|
sp102C4I21Z_CWYR  HNTKTKDRIYQNSDPTNTSPALLDQIEPSLYGDAWRCDFADTALNLSNAYLCLKLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
sp108L4I21Z_CPYR  HNGSKSKESAPICPNTUPEETEFRRIRNIVSLYGRYNKCCDFADKALIKSDIAYLCLRLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
sp108U711Z_GYVY  HNGSKSKESAPICPNTUPEETEFRRIRNIVSLYGRYNKCCDFADKALIKSDIAYLCLRLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
sp108U711Z_JUNIN  HNGSKSKESAPICPNTUPEETEFRRIRNIVSLYGRYNKCCDFADKALIKSDIAYLCLRLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
sp108U711Z_IPPY  HNGSKSKESAPICPNTUPEETEFRRIRNIVSLYGRYNKCCDFADKALIKSDIAYLCLRLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
Consensus  HNGSKSKESAPICPNTUPEETEFRRIRNIVSLYGRYNKCCDFADKALIKSDIAYLCLRLNTHLRSHLCOTGEEELPTTITVPEEPSPLPQR
```

Do dopasowania wielu sekwencji można wykorzystać wiele programów:

- CLUSTAL Omega
- MAFFT
- MUSCLE
- T-Coffe

Uzyskane wyniki co do zasady będą porównywalne ale każdy z programów ze względu na specyficzny algorytm może wprowadzać różnice w ostatecznym przyrównaniu. Również prezentacja wyników będzie dla każdego programu inna.

jak wykorzystać wyniki msa

- Można wnioskować o homologii sekwencji. Najlepiej zachowane | dopasowane fragmenty sekwencji będą zazwyczaj świadczyć o wysokim podobieństwie funkcji biologicznych.
- Zmiany | różnice pomiędzy sekwencjami świadczą o zdarzeniach ewolucyjnych.
- W inżynierii genetycznej MSA pozwala zaprojektować statery zdegenerowane.
- Pozwala generować drzewa filogenetyczne pozwalające wnioskować o przebiegu zmian ewolucyjnych, pokrewieństwie sekwencji i w efekcie organizmów.
- W przypadku nieznanych sekwencji możemy wnioskować o ich funkcjach biologicznych.

ocena wyniku msa

[illegible]

Dla dopasowania dwóch sekwencji możliwe jest obliczenie parametru będącego oceną jakości przyrównania.

W przypadku dopasowania wielu sekwencji jest to nie możliwe!

Nie ma dopasowania idealnego | poprawnego.
Jest tylko dopasowanie **optymalne** .

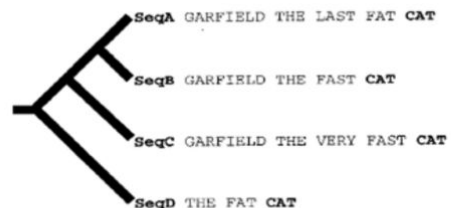
Ocena jakości dopasowania wielu sekwencji jest w gestii badacza.

Czy kolejność wprowadzenia sekwencji w dopasowaniu będzie mieć znaczenia?

czy kolejność wprowadzenia sekwencji w dopasowaniu będzie mieć znaczenia?

Im więcej sekwencji tym większy 'ciężar' poprawnie dopasowanych fragmentów sekwencji.

Bez interwencji badacza automatyczna korekta dopasowanie jest niezwykle trudna.



```
SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD ----- THE ---- FA-T CAT
```

historia msa

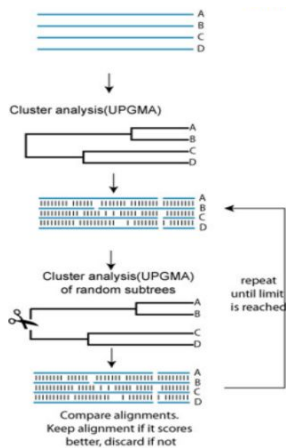
CLUSTALW, protoplasta CLUSTAL Omega powstał w 1994 roku i jest de facto standardem dla MSA.

Inne programy dołączyły później i oferują dodatkowe, wartościowe cechy dla swoich dopasowań.

- T-Coffee (2000)
- MAFFT (2002, multiple alignment using fast Fourier transform)
- MUSCLE (2004, multiple sequence comparison by log-expectation)



muscle



- Dopasowanie sekwencji w parach;
- Poprawa jakości dopasowania poprzez dodatkowe cykle optymalizujące dopasowanie;
- Drzewo przewodnie jest dzielone na dwie części w sposób losowy;
- Kolejna runda optymalizacji wewnątrz lokalnych gałęzi drzew;
- Dopasowanie gałęzi;
- Ocena:
 - Jeżeli ocena dopasowania dla gałęzi jest lepsze niż wstępna ocena drzewo jest porzucane i dopasowanie jest generowane od nowa;
 - Jeżeli ocena jest gorsza, dopasowanie jest odrzucane i następuje optymalizacji starego drzewa.

jak postępować...

- Im więcej sekwencji tym lepiej...
- Im mniej podobne sekwencje tym lepiej...
- Im bardziej zgodna pod względem długości pula sekwencji tym bardziej wiarygodne dopasowanie...

Input data	Recommendation
A small number of unusually long sequences (>20,000 residues)	Use CLUSTALW. Other program may run out of memory, causing an abort.
2-100 sequences of typical protein length (maximum around 10,000 residues)	Use T-Coffee, MAFFT or MUSCLE
100-500 sequences that are globally alignable	Use MUSCLE or MAFFT
>500 sequences	Use MUSCLE or MAFFT with a faster option

filogenetyka

Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species [...].

Joseph Felsenstein, *Inferring Phylogenies* (2004)

filogenetyka

Dział biologii zajmujący się badaniem drogi rozwojowej (filogenezą) organizmów.

Przedmiotem zainteresowania filogenetyki są organizmy żyjące współcześnie oraz kopalne, ich pochodzenie i relacje pokrewieństwa.

W swoich badaniach korzysta z osiągnięć paleontologii, genetyki i innych nauk przyrodniczych.

Aby wyjaśnić jak powstają drzewa filogenetyczne można spojrzeć na ich genezę z dwóch perspektyw:

- Teoretyczne podejście;
- Podejście eksperymentalne.

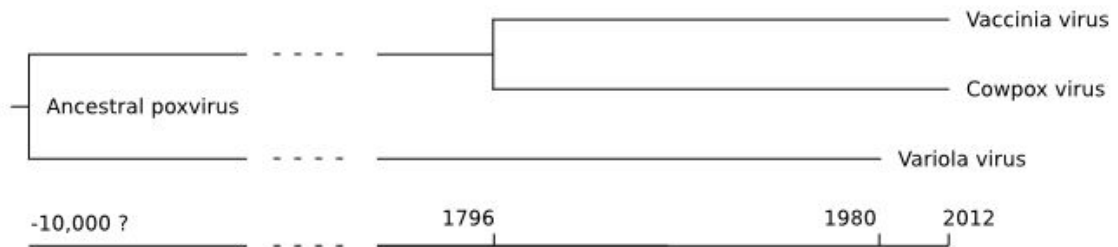
Jak przykład wykorzystania teorii drzew filogenetycznych spojrzymy na historię ewolucji wirusa ospy prawdziwej.

drzewo filogenetyczne pozwala prześledzić zdarzenia w historii ewolucji gatunków

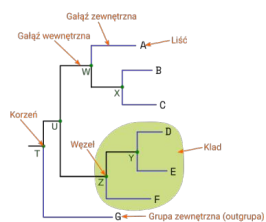
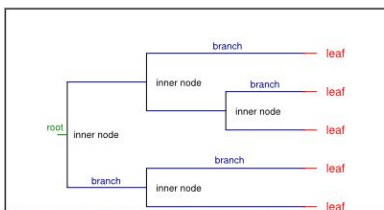
Prawdopodobna historia ospy prawdziwej:

- Na początku był tylko jeden rodzaj wirusa;
- Około 10 000 lat temu nastąpił rozdział linii na dwa oddzielne szczepy
 - krowinkę (łac. variola vaccinia – wirusowa choroba zakaźna występująca u bydła domowego i świń)
 - ospę prawdziwą (łac. variola vera - dawne nazwy: ospa naturalna, czarna ospa (łac. variola nigra) – wirusowa choroba zakaźna o ostrym przebiegu wywoływana przez jedną z dwóch odmian wirusa ospy prawdziwej (variola minor lub variola maior))
- W 1796 roku Edward Jenner używa wirusa krowinki jako szczepionki u ludzi
- W 1980 roku uznaje się iż wirus variola został w wyniku szczepień całkowicie wyeliminowany

graficzna reprezentacja zmian ewolucyjnych



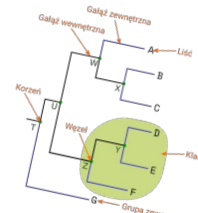
drzewo filogenetyczne



- Root | korzeń drzewa filogenetycznego
- Branch | konar lub gałąź drzewa filogenetycznego
- Leaf | liść

- Inne elementy:
- Gałąź zewnętrzna
- Gałąź wewnętrzna
- Kład
- Grupa zewnętrzna

drzewo filogenetyczne | topologia

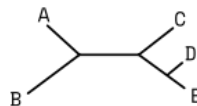


- Gałęzie pokazują związki pomiędzy nimi.
Ich długość może (w zależności od rodzaju drzewa) odpowiadać zmianom w sekwencjach nagromadzonych podczas ewolucji.
Można wyróżnić gałęzie wewnętrzne prowadzące do węzłów i gałęzie zewnętrzne zakończone liśćmi.
 - Węzły to miejsca łączenia się gałęzi - reprezentują jednostki taksonomiczne (gatunki, osobniki, odmiany itd.).
 - Węzły wewnętrzne (nie będące liśćmi) reprezentują hipotetycznego wspólnego przodka kładu.
 - Liście są końcowymi (terminalnymi) węzłami, odpowiadają badanym sekwencjom/taksonom.
-
- Drzewa nieukorzone przedstawiają wzajemne podobieństwa ale nie pozwalają określić w jakiej kolejności poszczególne taksony się od siebie oddzielały.
 - Drzewa ukorzone posiadają węzeł, który odpowiada ostatniemu wspólnemu przodkowi badanych taksonów.

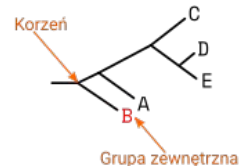
drzewo filogenetyczne | topologia

W drzewach ukorzonych można wyznaczyć grupę zewnętrzną, zwaną także outgrupą (ang. outgroup). Jest to takson (lub grupa taksonów), który jest dalej spokrewniony z pozostałymi badanymi, niż one między sobą.
Innymi słowy, oddzielił się on najwcześniej podczas ewolucji.

Drzewo nieukorzone

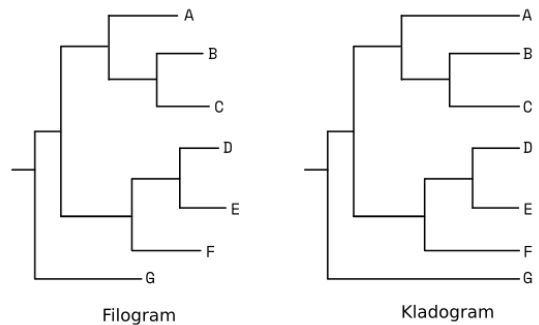


Drzewo ukorzone

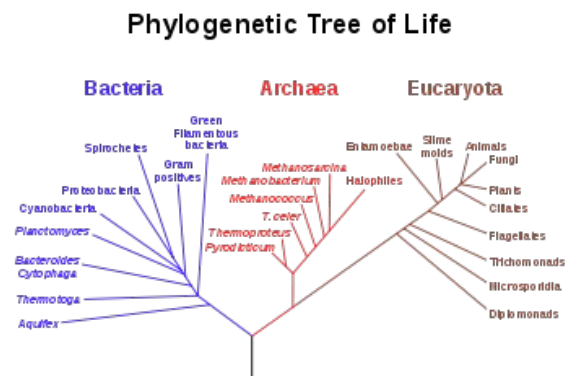
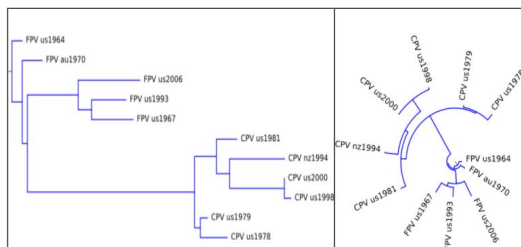


drzewo filogenetyczne | topologia

- Gdy długość gałęzi drzewa odzwierciedla odległość ewolucyjną badanych sekwencji, wtedy drzewo nazywamy **filogramem**.
- **Kladogram** natomiast pokazuje jedynie pokrewieństwa między badanymi taksonami.



drzewa filogenetyczne | reprezentacja



wykorzystanie drzew filogenetycznych

Aby powstało drzewo filogenetyczne konieczna jest analiza wiele cech gatunków lub innych grup, np.:

- morfologię zewnętrzną (kształt/wygląd),
- anatomię wewnętrzną,
- zachowania,
- szlaki biochemiczne,
- sekwencje DNA
- białek

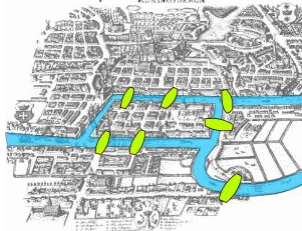
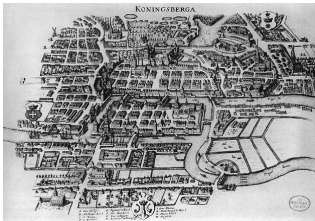
- charakterystykę skamieniałości.

Drzewa filogenetyczne mają charakter hipotezy, a nie ostatecznej odpowiedzi.

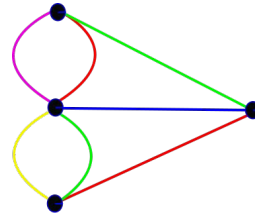
Drzewa filogenetyczne są korygowane i aktualizowane, gdy pojawiają się nowe informacje dotyczące pokrewieństwa gatunków.

Analiza grafów

- Analiza sieci biologicznych | teoria grafów i mediów społecznościowych.



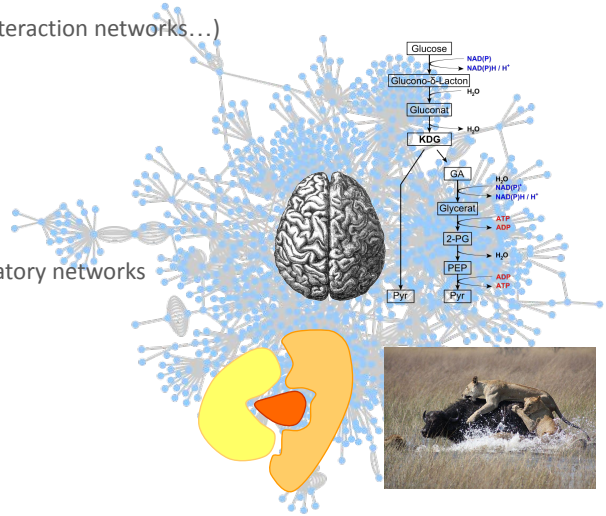
- Zagadnienie mostów królewieckich to problem matematyczny, który rozwiązał w XVIII wieku Leonhard Euler. Przez Królewiec przepływała rzeka Pregola, w której rozwidleniach znajdowały się dwie wyspy. Czy można było przejść po każdym z 7 mostów (na rzece Pregola) dokładnie raz? Okazało się, że nie da się tego zrobić. Spójny graf jest grafem Eulera wtedy i tylko wtedy, gdy każdy wierzchołek ma parzysty stopień.



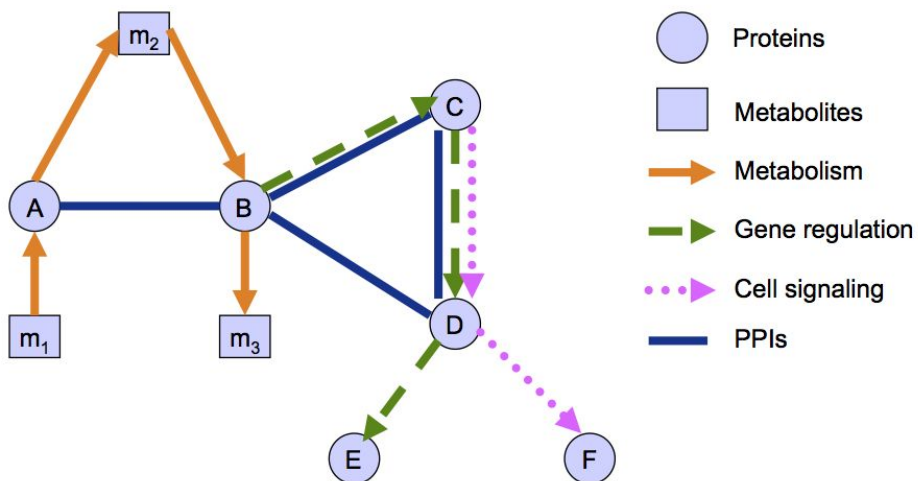
Analiza grafów w biologii

Złożone zagadnienia biologiczne mogą być przedstawione za pomocą grafów:

- Ekologia | Ecological networks (food webs, species interaction networks...)
- Neurologia | Neurological networks
- Metabolomika | Metabolic networks
- Przekazywanie sygnałów | Signalling networks
- Genetyka | Genetic interaction networks, gene regulatory networks
- Białka | Protein-protein interaction networks
- ...

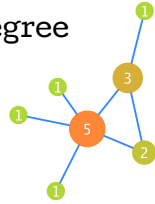


Teoria grafów | podsumowanie

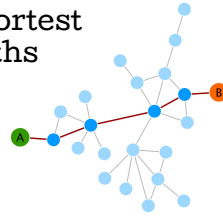


Teoria grafów definicje

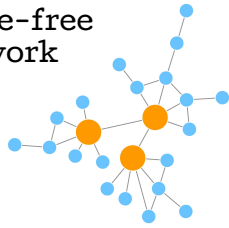
Degree



Shortest paths



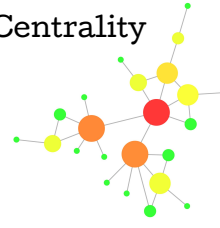
Scale-free network



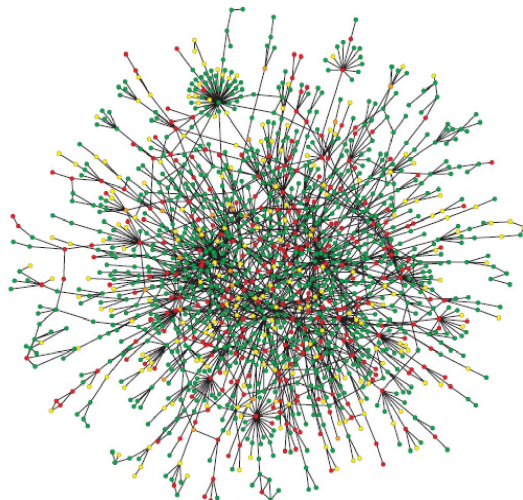
Transitivity



Centrality



Przykład

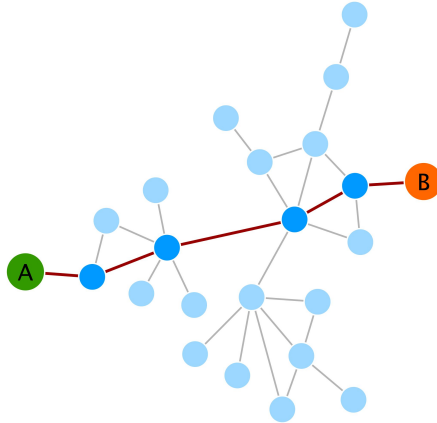


>80% białek skoncentrowane w 'centrum' sieci

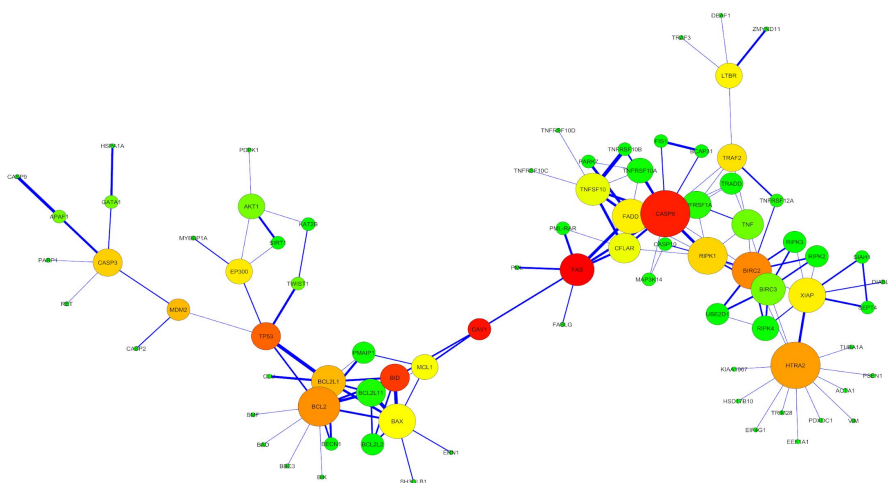
Średnia odległość pomiędzy węzłami: 6 (duży świat, bliskie oddziaływania)

- Węzły – białka
- Krawędzie – interakcje

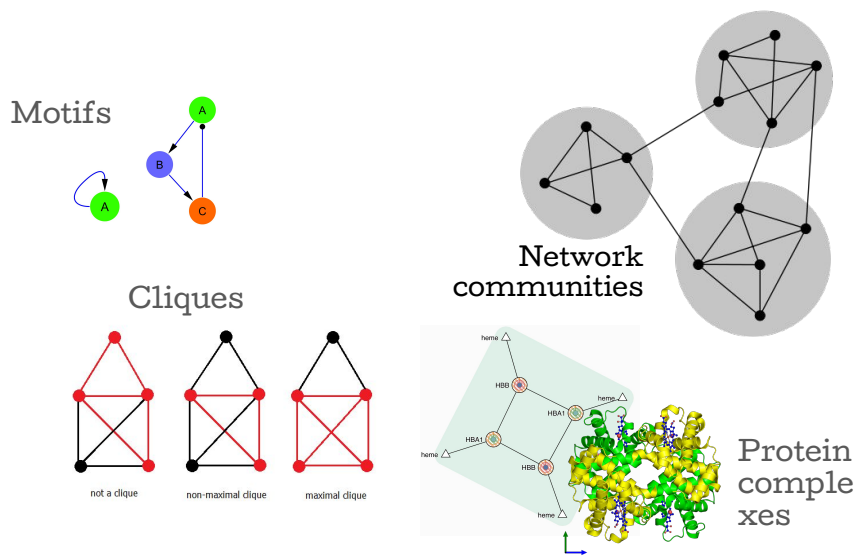
PPI: efekt krótkiej ścieżki oddziaływań



PPI: efekt centralizacji



PPI: grupowanie



<https://commons.wikimedia.org/w/index.php?curid=17125894>

<http://www.ebi.ac.uk/pdbe/entry/pdb/2dn1>

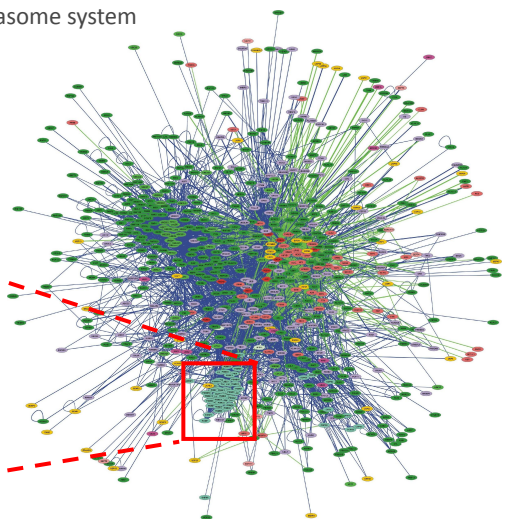
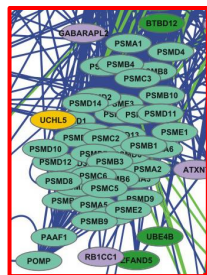
<https://www.ebi.ac.uk/complexportal/complex/CPX-2158>

PPI: grupowanie

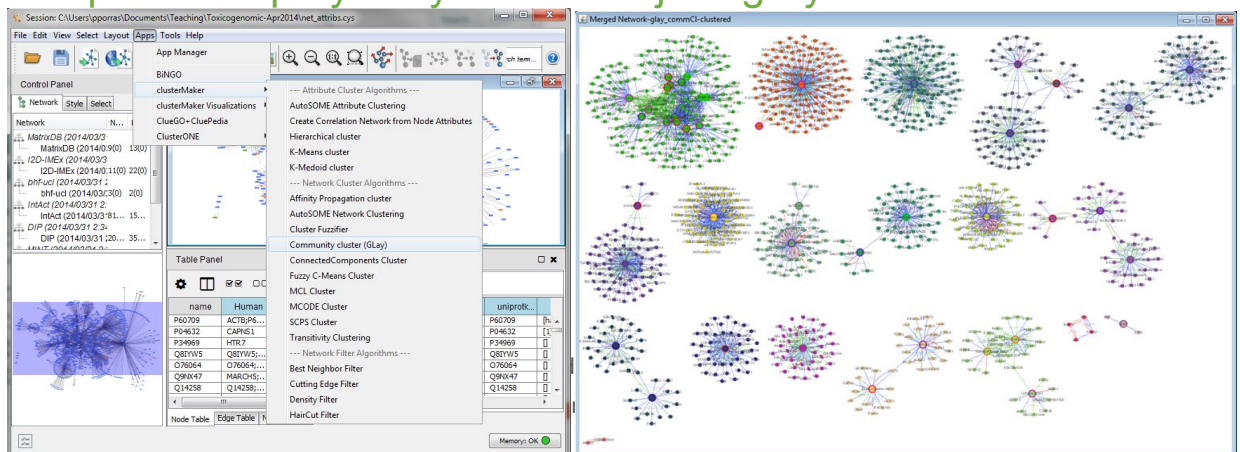
- Functions are likely to be carried out in a highly modular manner.
- High clustering => High modularity

Ubiquitin-proteasome system network

Proteasome cluster



Grupowanie: przykłady wizualizacji i algorytmów



- ✓ Weighting: gives a higher score to those nodes whose neighbours are more interconnected.
- ✓ Molecular complex prediction: starting with the highest-weighted node (seed), recursively move out, adding nodes to the complex that are above a given threshold.
- ✓ Post-processing, which applies filters to improve the cluster quality (haircut and fluff).
- ✓ Check <http://www.youtube.com/watch?v=7wA4ZEoFGI8> and <http://baderlab.org/Software/MCODE/UsersManual>.

filogenetyka

Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species [...].

Joseph Felsenstein, Inferring Phylogenies (2004)

filogenetyka

Dział biologii zajmujący się badaniem drogi rozwojowej (filogenezą) organizmów.

Przedmiotem zainteresowania filogenetyki są organizmy żyjące współcześnie oraz kopalne, ich pochodzenie i relacje pokrewieństwa.

W swoich badaniach korzysta z osiągnięć paleontologii, genetyki i innych nauk przyrodniczych.

Aby wyjaśnić jak powstają drzewa filogenetyczne można spojrzeć na ich genezę z dwóch perspektyw:

- Teoretyczne podejście;
- Podejście eksperymentalne.

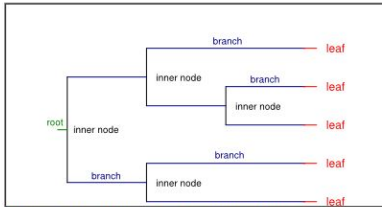
Jak przykład wykorzystania teorii drzew filogenetycznych spojrzymy na historię ewolucji wirusa ospy prawdziwej.

drzewo filogenetyczne pozwala prześledzić zdarzenia w historii ewolucji gatunków

Prawdopodobna historia ospy prawdziwej:

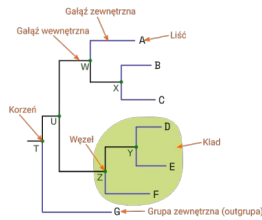
- Na początku był tylko jeden rodzaj wirusa;
- Około 10 000 lat temu nastąpił rozdział linii na dwa oddzielne szczepy
 - krowinkę (łac. variola vaccinia – wirusowa choroba zakaźna występująca u bydła domowego i świń)
 - ospę prawdziwą (łac. variola vera - dawne nazwy: ospa naturalna, czarna ospa (łac. variola nigra) – wirusowa choroba zakaźna o ostrym przebiegu wywoływana przez jedną z dwóch odmian wirusa ospy prawdziwej (variola minor lub variola maior))
- W 1796 roku Edward Jenner używa wirusa krowinki jako szczepionki u ludzi
- W 1980 roku uznaje się iż wirus variola został w wyniku szczepień całkowicie wyeliminowany

drzewo filogenetyczne



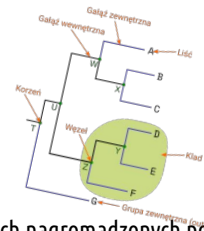
- Root | korzeń drzewa filogenetycznego
- Branch | konar lub gałąź drzewa filogenetycznego
- Leaf | liść

- Inne elementy:
 - Gałąź zewnętrzna
 - Gałąź wewnętrzna
 - Klad
 - Grupa zewnętrzna



drzewo filogenetyczne | topologia

- Gałęzie pokazują związki pomiędzy nimi.
Ich długość może (w zależności od rodzaju drzewa) odpowiadać zmianom w sekwencjach nagromadzonych podczas ewolucji.
Można wyróżnić gałęzie wewnętrzne prowadzące do węzłów i gałęzie zewnętrzne zakończone liśćmi.
 - Węzły to miejsca łączenia się gałęzi - reprezentują jednostki taksonomiczne (gatunki, osobniki, odmiany itd.).
 - Węzły wewnętrzne (nie będące liśćmi) reprezentują hipotetycznego wspólnego przodka kladu.
 - Liście są końcowymi (terminalnymi) węzłami, odpowiadają badanym sekwencjom/taksonom.
-
- Drzewa nieukorzenione przedstawiają wzajemne podobieństwa ale nie pozwalają określić w jakiej kolejności poszczególne taksony się od siebie oddzielały.
 - Drzewa ukorzenione posiadają węzeł, który odpowiada ostatniemu wspólnemu przodkowi badanych taksonów.



wykorzystanie drzew filogenetycznych

Aby powstało drzewo filogenetyczne konieczna jest analiza wiele cech gatunków lub innych grup, np.:

- morfologię zewnętrzną (kształt/wygląd),
- anatomię wewnętrzną,
- zachowania,
- szlaki biochemiczne,
- sekwencje DNA
- białek

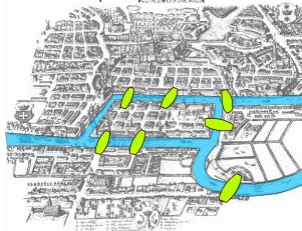
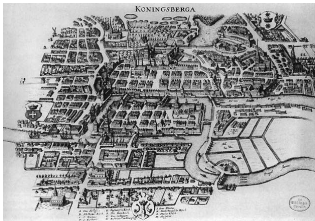
- charakterystykę skamieniałości.

Drzewa filogenetyczne mają charakter hipotezy, a nie ostatecznej odpowiedzi.

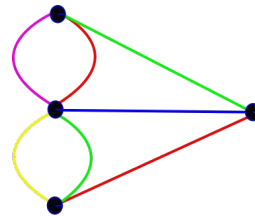
Drzewa filogenetyczne są korygowane i aktualizowane, gdy pojawiają się nowe informacje dotyczące pokrewieństwa gatunków.

Analiza grafów

- Analiza sieci biologicznych | teoria grafów i mediów społecznościowych.

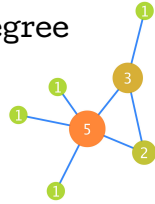


- Zagadnienie mostów królewieckich to problem matematyczny, który rozwiązał w XVIII wieku Leonhard Euler. Przez Królewiec przepływała rzeka Pregola, w której rozwidleniach znajdowały się dwie wyspy. Czy można było przejść po każdym z 7 mostów (na rzece Pregola) dokładnie raz? Okazało się, że nie da się tego zrobić. Spójny graf jest grafem Eulera wtedy i tylko wtedy, gdy każdy wierzchołek ma parzysty stopień.

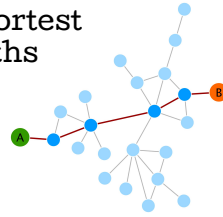


Teoria grafów definicje

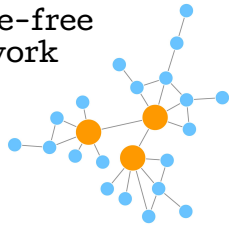
Degree



Shortest paths



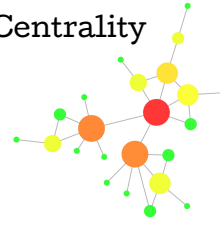
Scale-free network



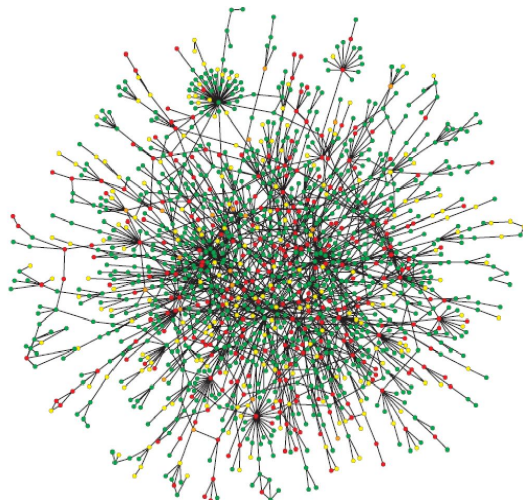
Transitivity



Centrality



Przykład

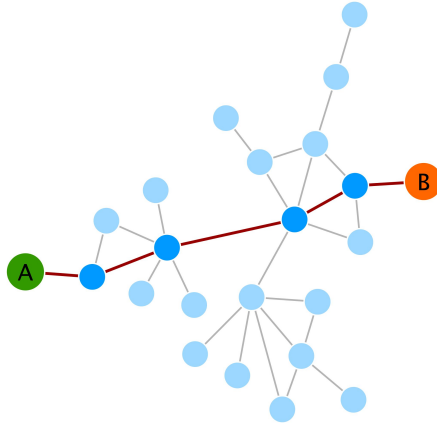


>80% białek skoncentrowane w 'centrum' sieci

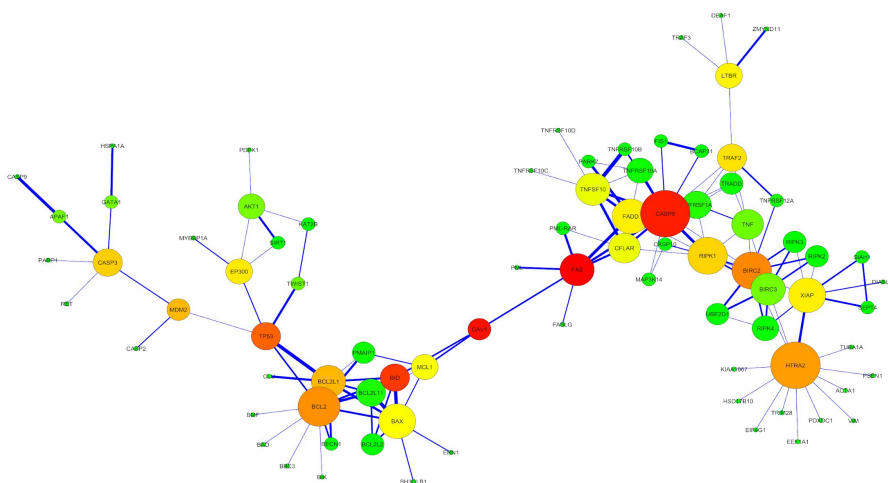
Średnia odległość pomiędzy węzłami: 6 (duży świat, bliskie oddziaływania)

- Węzły – białka
- Krawędzie – interakcje

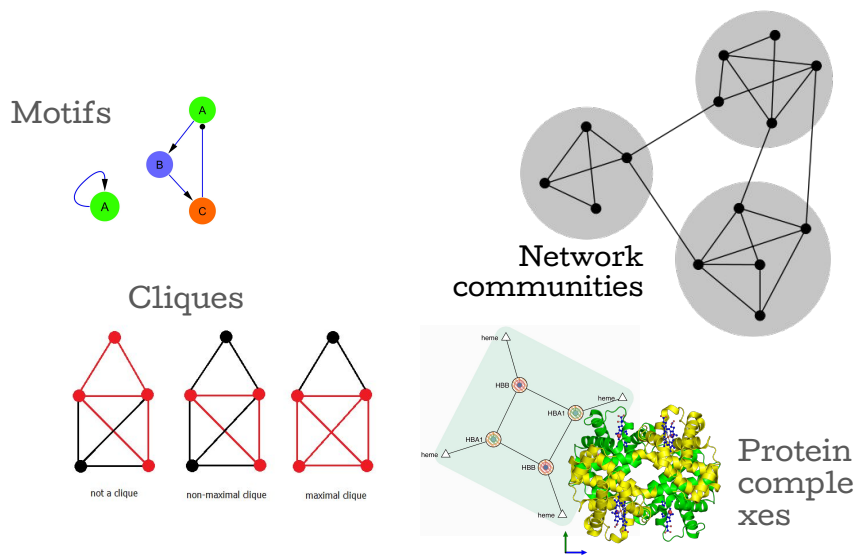
PPI: efekt krótkiej ścieżki oddziaływań



PPI: efekt centralizacji



PPI: grupowanie



<https://commons.wikimedia.org/w/index.php?curid=17125894>

<http://www.ebi.ac.uk/pdbe/entry/pdb/2dn1>

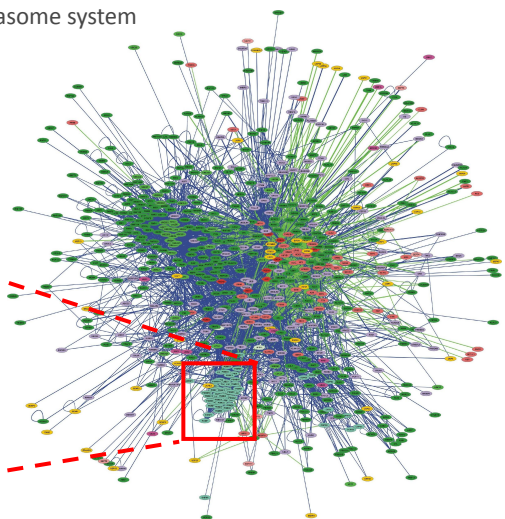
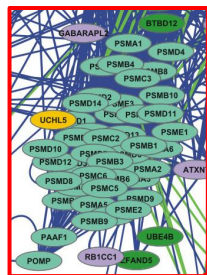
<https://www.ebi.ac.uk/complexportal/complex/CPX-2158>

PPI: grupowanie

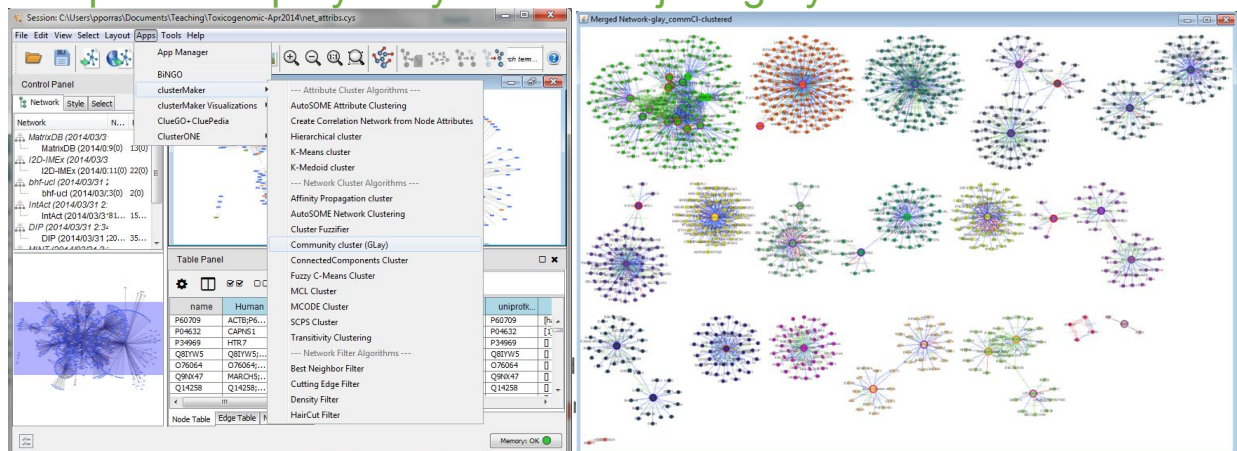
- Functions are likely to be carried out in a highly modular manner.
- High clustering => High modularity

Ubiquitin-proteasome system network

Proteasome cluster

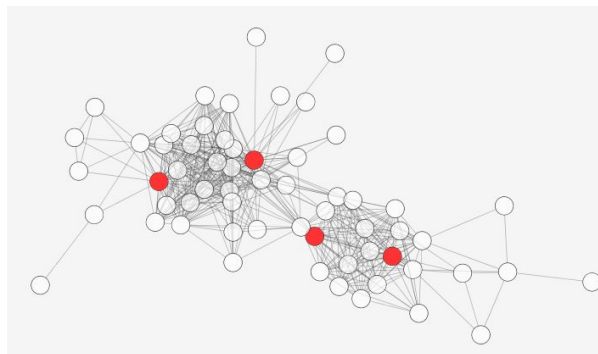


Grupowanie: przykłady wizualizacji i algorytmów

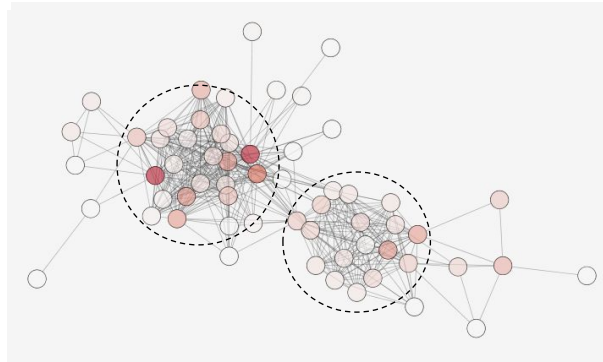


- ✓ Weighting: gives a higher score to those nodes whose neighbours are more interconnected.
- ✓ Molecular complex prediction: starting with the highest-weighted node (seed), recursively move out, adding nodes to the complex that are above a given threshold.
- ✓ Post-processing, which applies filters to improve the cluster quality (haircut and fluff).
- ✓ Check <http://www.youtube.com/watch?v=7wA4ZEoFGI8> and <http://baderlab.org/Software/MCOE/UsersManual>.

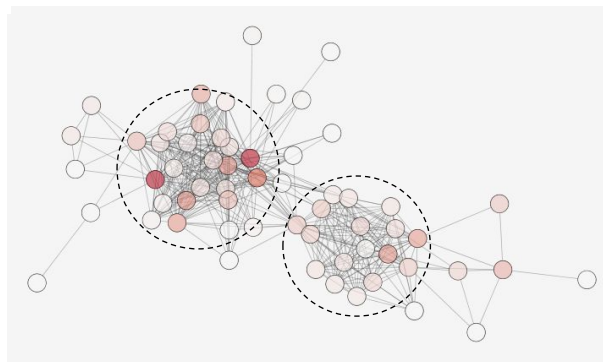
PPI Network analysis: Network propagation + module detection



PPI Network analysis: Network propagation + module detection

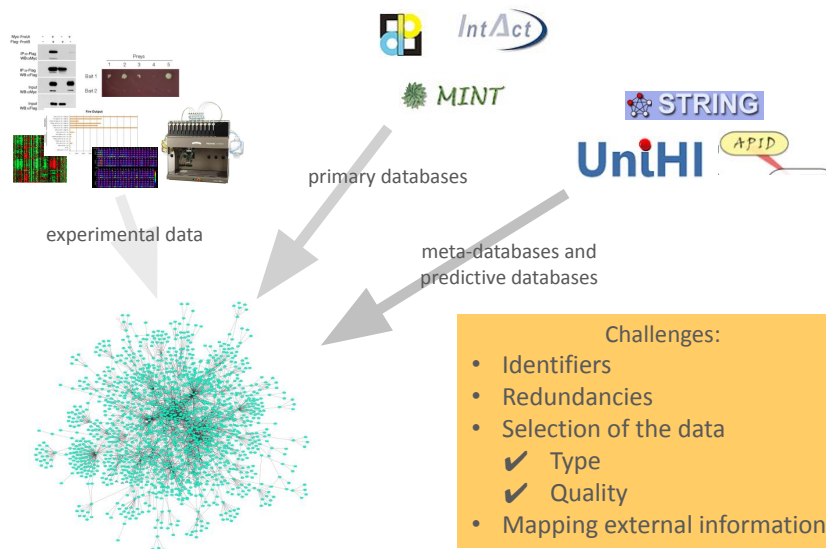


PPI Network analysis: Network propagation + module detection

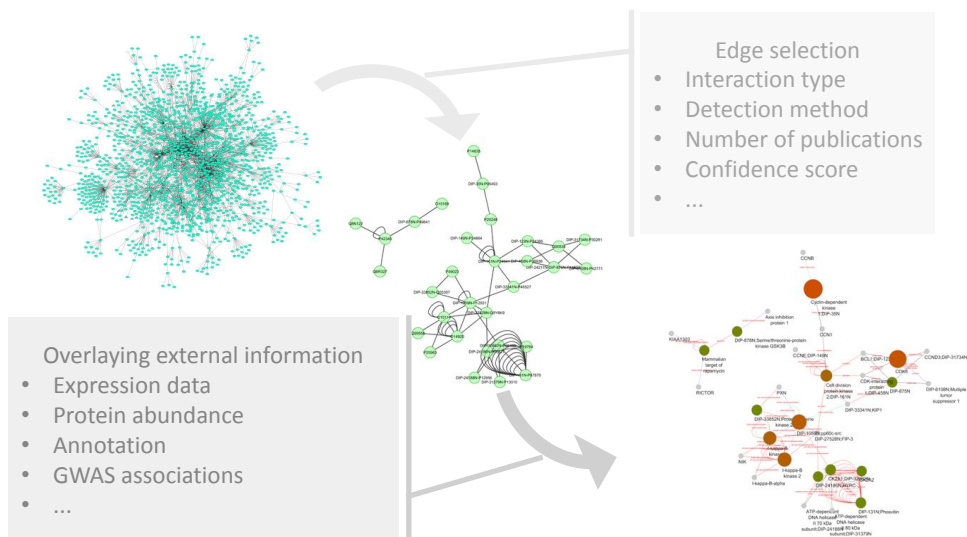


- Network propagation algorithms:
 - PageRank (Google)
 - Heat diffusion
- Implementation of propagation / diffusion through Cytoscape:
 - Carlin et al. PLoS Comput Biol 2017 (PMID:29023449)

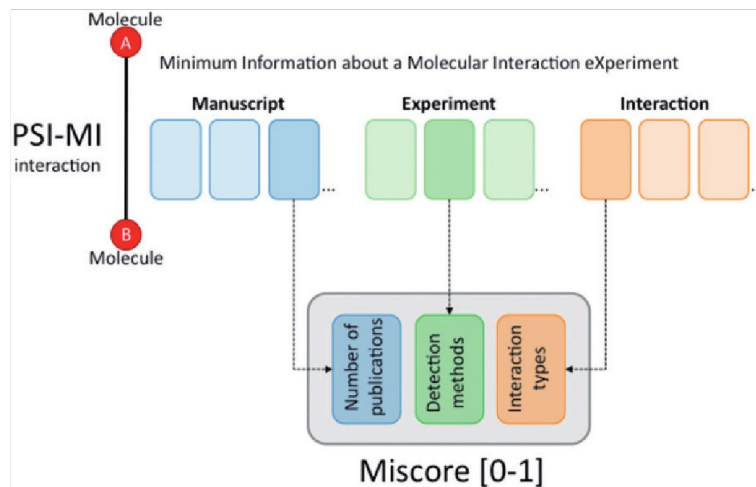
PPI: budowanie sieci



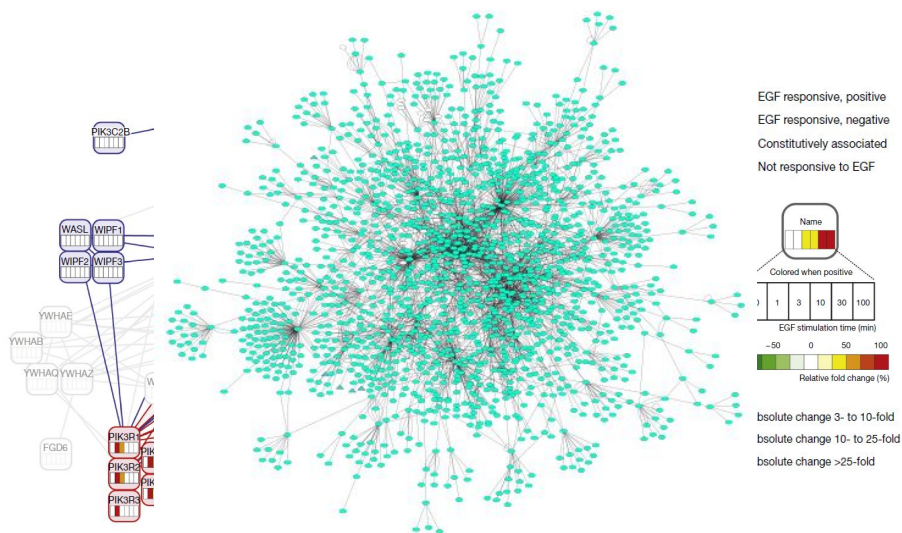
PPI: budowanie sieci



PPI: budowanie sieci, jakość danych



Tworzenie wizualizacji | wyzwania



Definicje

Protein-protein interactions (PPIs):

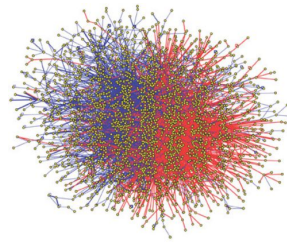
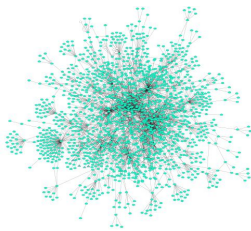
- oparte o dane związane z fizycznym oddziaływaniem białek
- *physical and selective contacts that happen between pairs of proteins, in certain molecular regions and in a defined biological context.*

Interactome:

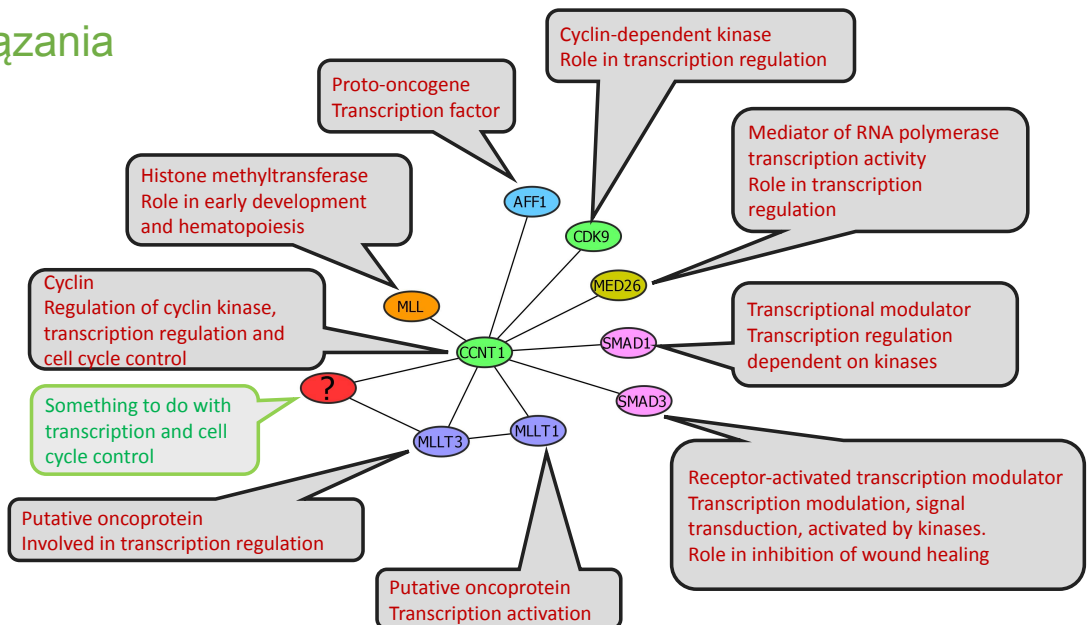
- oparte o dane na temat biologicznych oddziaływań
- *the totality of PPIs that happen in a cell / in an organism / in a specific biological context...*

Protein-protein interaction network:

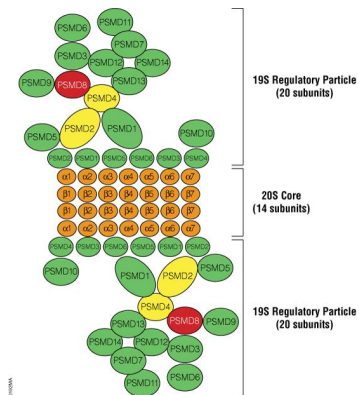
- graficzna interpretacja
- *graphical representation of a group of PPIs in which proteins are represented as nodes and interactions as edges.*



Powiązania



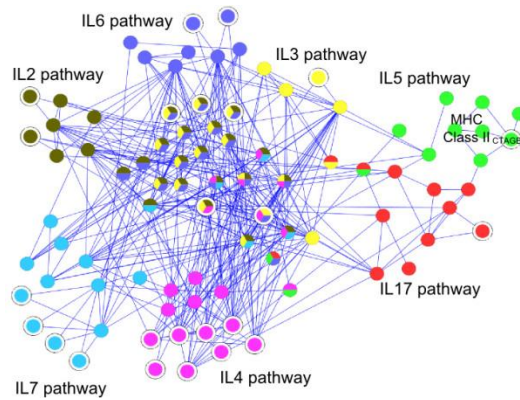
Characterization of protein complexes and pathways



Hook, B. and Schagat, T. [Internet] 2011.

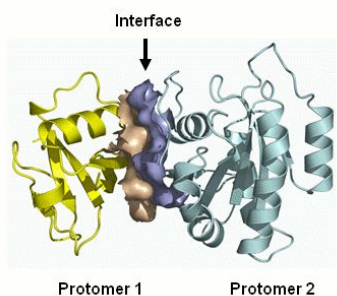
Available from:

www.promega.com/resources/articles/pubhub/functiona-l-proteomics-techniques-to-isolate-and-characterize-the-human-proteasome/

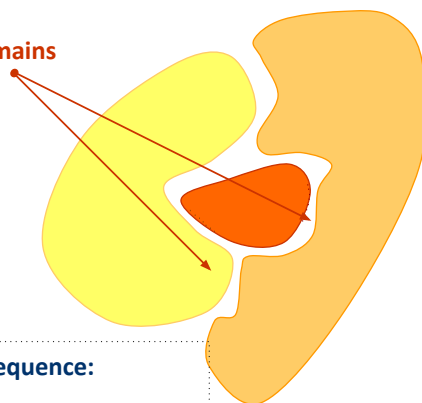


Glaab et al., BMC Bioinformatics, PMID: 21144022.

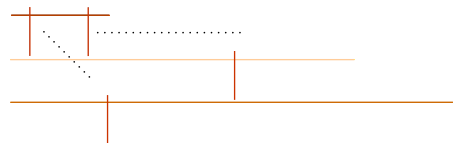
Representing protein-protein interactions: interacting domains



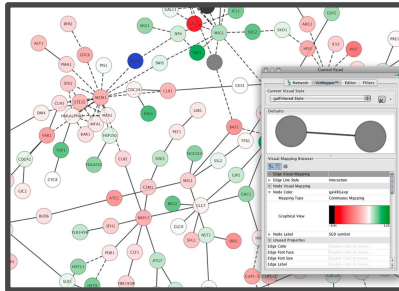
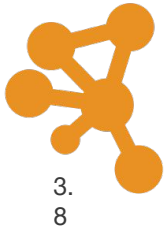
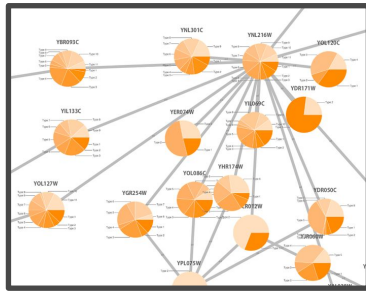
Interacting domains



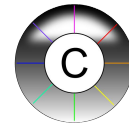
Overlay of Ranges on sequence:



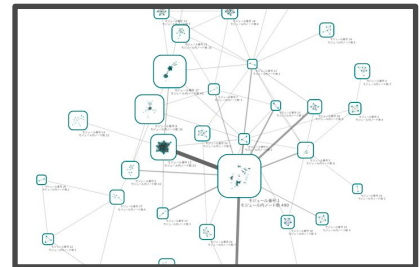
Cytoscape



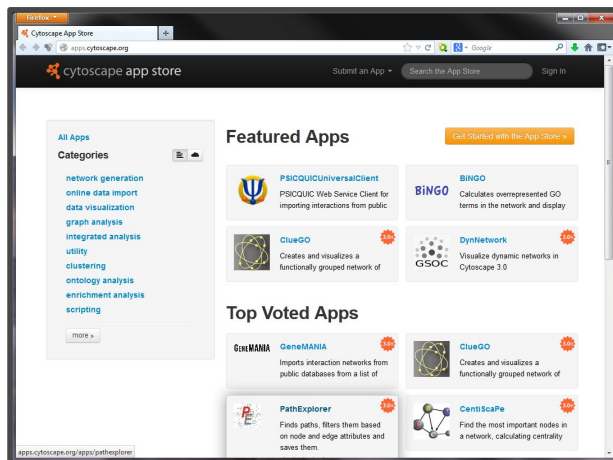
www.cytoscape.org



2.8



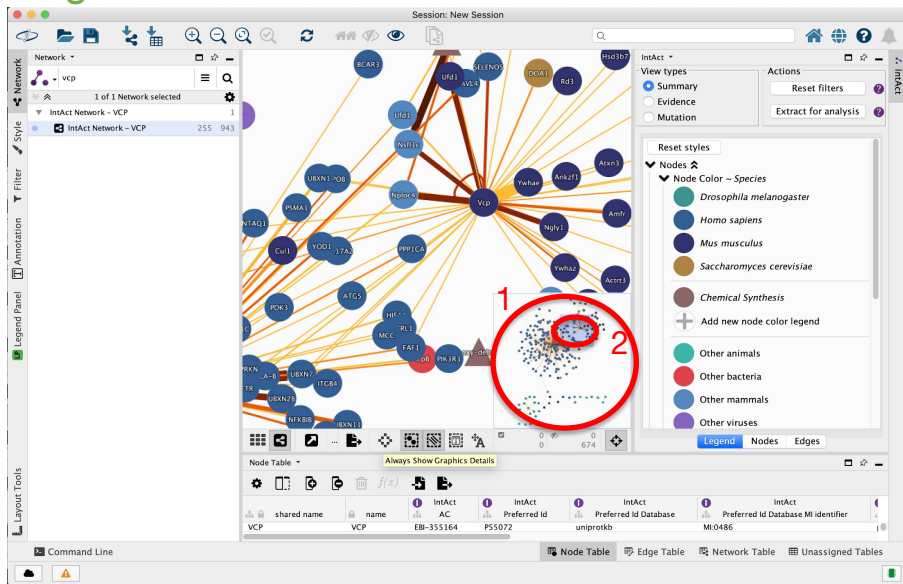
Cytoscape apps/plugins



<https://apps.cytoscape.org>

- Large variety of apps/plugins (~370)
- Great flexibility, adaptable to multiple types of analysis, in various domains of knowledge: bioinformatics, social network analysis, semantic web...
- Possibility to create your own.
- Some deprecated or obsolete
- Different functionality depending on the Cytoscape version

Visualizing the network



Visualizing the network: view types

