

Bioinformatyka

Michał Przyłuski

Plan prezentacji

- Wstęp biologiczny
- Biologia molekularna
 - genetyka
- Bioinformatyka
- Przykłady zastosowań:
 - sekwencjonowanie
 - mikromacierze
 - filogenetyka

Czemu wstęp biologiczny?

- **Informatyk** nie ma pojęcia o biologii
- Chcemy współpracować z **biolożkami**, a nie przetwarzać suche ATGCATCGGCTA...
- Informatyk może poznać podstawy biologii, bo już nauczył się tylu dziwnych rzeczy
- Zagubiona biolożka, sama w labie...
 - nie postawi klastra Oracle'owego
 - nie zakoduje sobie programu do analizy swoich danych
 - nie zna Mathematica'i, Matlaba, albo R'a...
 - nawet nie ma FireFox'a!
- Dlatego chcemy się nauczyć podstaw biologii, aby pomóc biolożkom w obsłudze komputera :-)

Informatyka dla biologii

- Wykorzystanie informatyki w biologii
 - a w szczególności: w biologii molekularnej
- Gdzie?
 - aparatura
 - software: analiza danych, doświadczenia *in silico*, projektowanie doświadczeń *in vivo* (liczne zalety)
- Nauka doświadczalna! Nie ma prawd objawionych, aksjomatów, i niczego na pewno.

Genetyka - początki

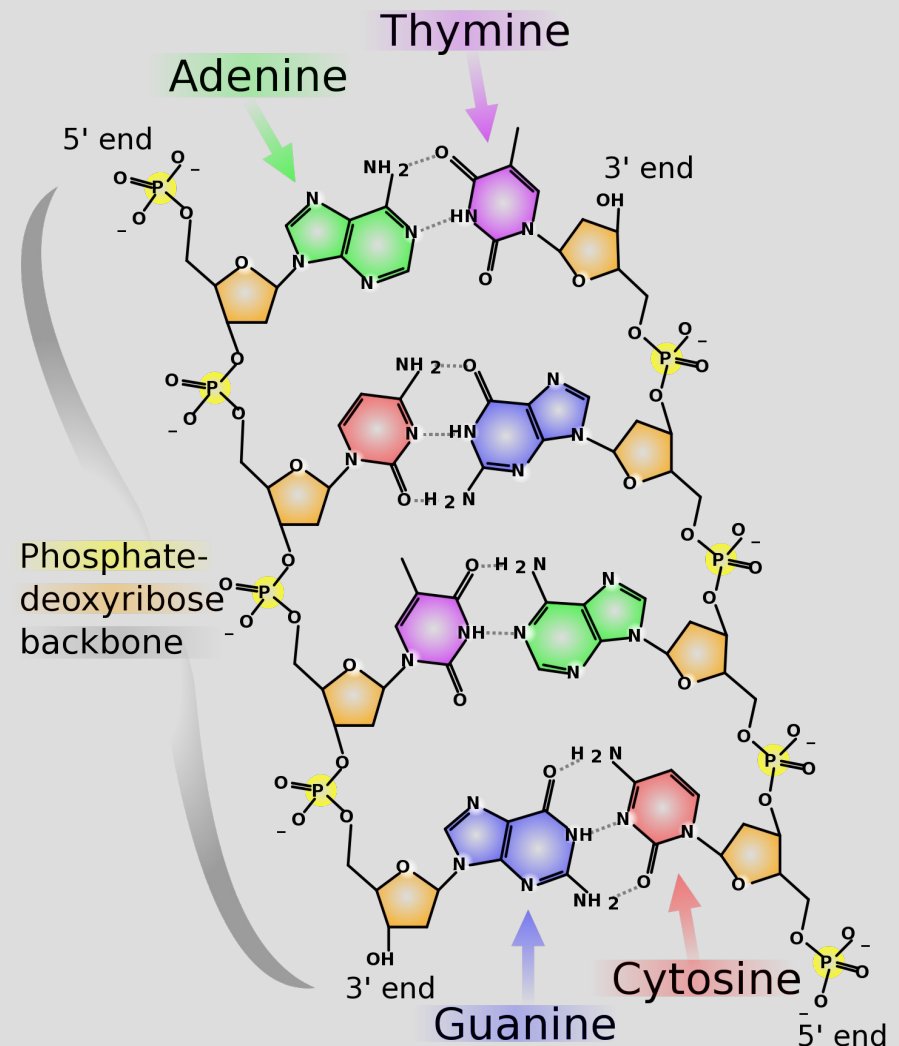
- Mendel
 - Brno
 - groszki (~1865)
 - odtworzenie praw na podstawie obserwacji
 - prawo czystości gamet
 - prawo niezależnej segregacji cech
- Morgan
 - chromosomowa teoria dziedziczności (1915)
- Watson, Crick i panna Franklin (1953)
 - dyfrakcja promieni X

Biologia Molekularna

- Biologia na poziomie molekularnym
 - białka
 - kwasy nukleinowe
- Związana z biochemią i genetyką
- Wiele technik pomiarowych, analitycznych, ilościowych i jakościowych
- Trudno jest przeanalizować "dane" bez odpowiedniego wsparcia komputera

Genetyka - DNA

- Informacja genetyczna głównie zawarta w **DNA** (=kwas deoksyrybonukleinowy)
- Informacja = czyste dane
- wymagania od nośników danych
- zasady azotowe
- A-T, G-C



Genetyka - DNA

- Nadmiarowy (ECC)

		Second base in codon				
		U	C	A	G	
First base in codon	U	Phe	Ser	Tyr	Cys	U
		Phe	Ser	Tyr	Cys	C
		Leu	Ser	STOP	STOP	A
		Leu	Ser	STOP	Trp	G
	C	Leu	Pro	His	Arg	U
		Leu	Pro	His	Arg	C
		Leu	Pro	Gln	Arg	A
		Leu	Pro	Gln	Arg	G
	A	Ile	Thr	Asn	Ser	U
		Ile	Thr	Asn	Ser	C
		Ile	Thr	Lys	Arg	A
		Met	Thr	Lys	Arg	G
	G	Val	Ala	Asp	Gly	U
		Val	Ala	Asp	Gly	C
		Val	Ala	Glu	Gly	A
		Val	Ala	Glu	Gly	G

Third base in codon

Genetyka - co się dzieje z DNA?

Pierwsze przybliżenie:

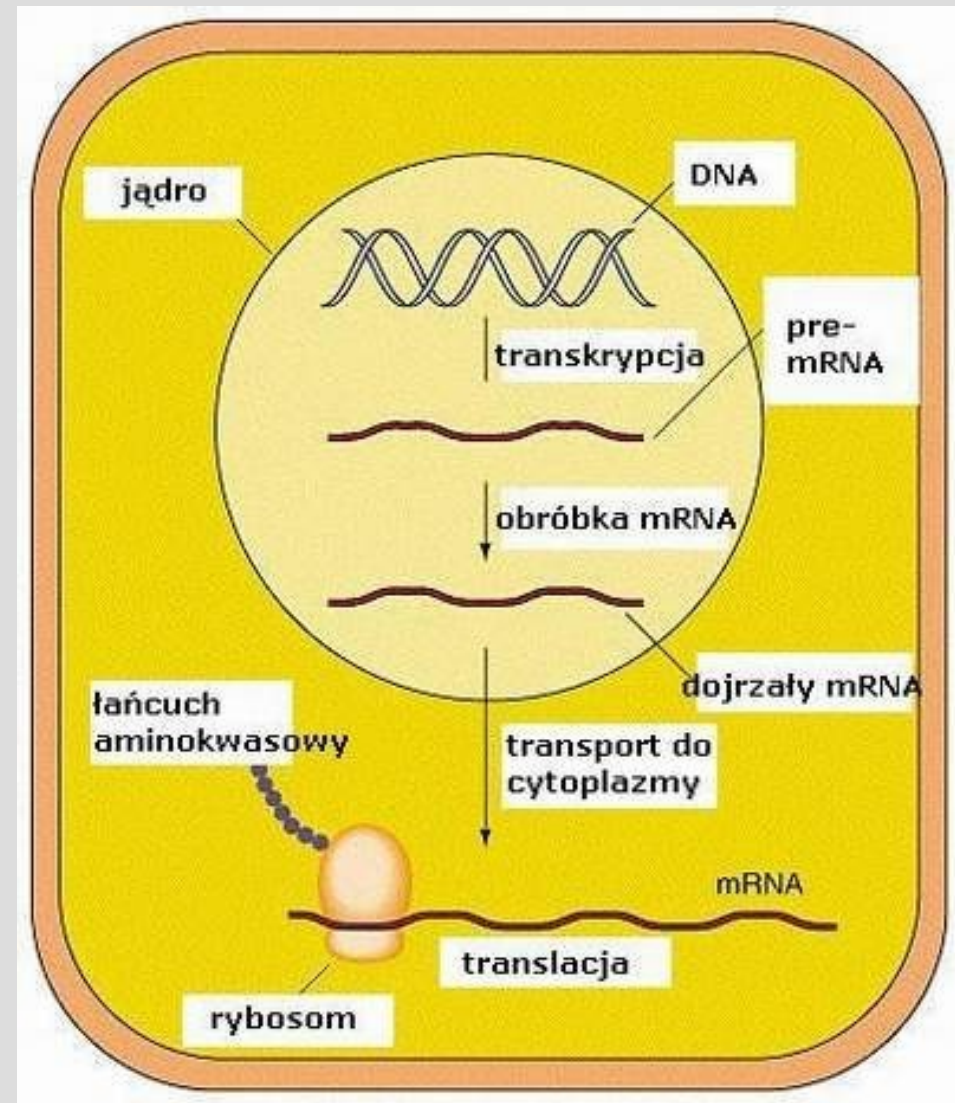
- DNA poddane **transkrypcji** -> RNA
 - polimeraza
- RNA poddane **translacji** -> białko
 - rybosom
- czasem też replikacja (mitoza/mejoza)

W rzeczywistości nie takie proste.

- bardzo liczne wyjątki, dużo wątpliwości

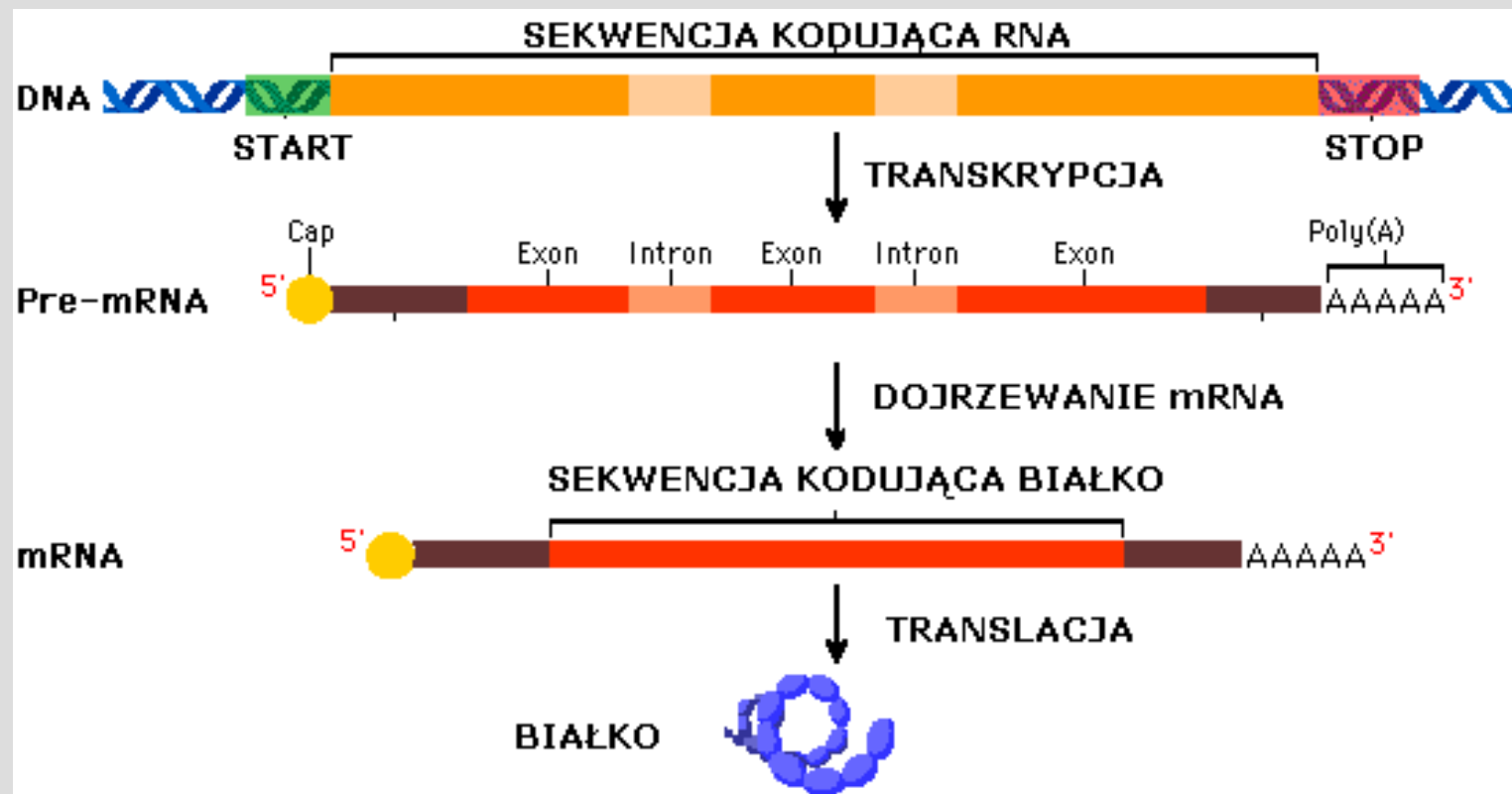
Od DNA do białka

- Wydaje się dość proste
 - liczne znaczniki (start, stop)
 - wiązanie polimerazy
 - wiązanie rybosomu
 - wierność procesu
 - korekcja błędów



Organizacja DNA

- Introny i eksony
- Nie jest to trywialny proces:
 - splicing GU / AG



Organizacja DNA

- Najczęściej w chromosomach
- Czasem (np. u wirusów) rolę nośnika pełni RNA... zorganizowany w kołko
 - i bardzo dobrze, bo dzięki temu mamy wartościowy enzym

Replikacja

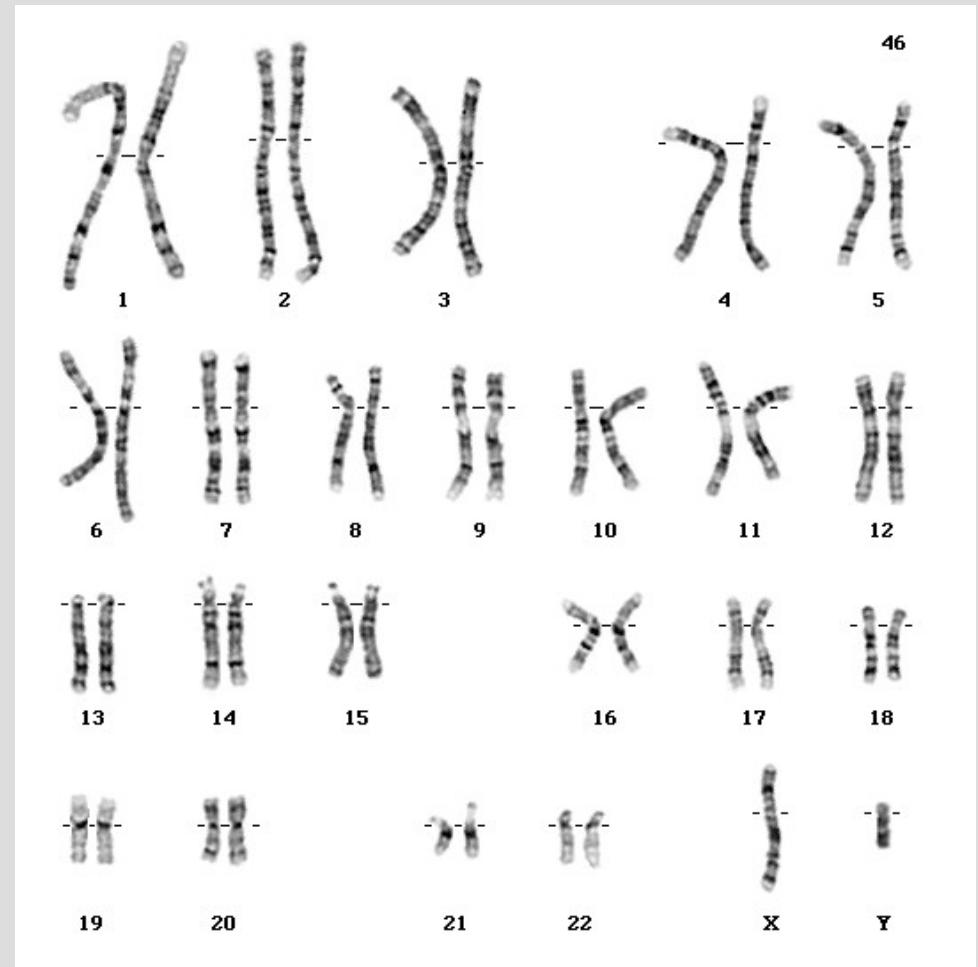
Dopiero w 1958 ustalono jak przebiega replikacja DNA

Doświadczenie Meselsona-Stahla:

- replikacja półkonserwatywna
- izotop azotu ^{15}N
- *Escherichia coli*

Kariotyp człowieka

- 22 pary + XY/XX
- znane od bardzo wielu lat
- mało informacji o zawartości

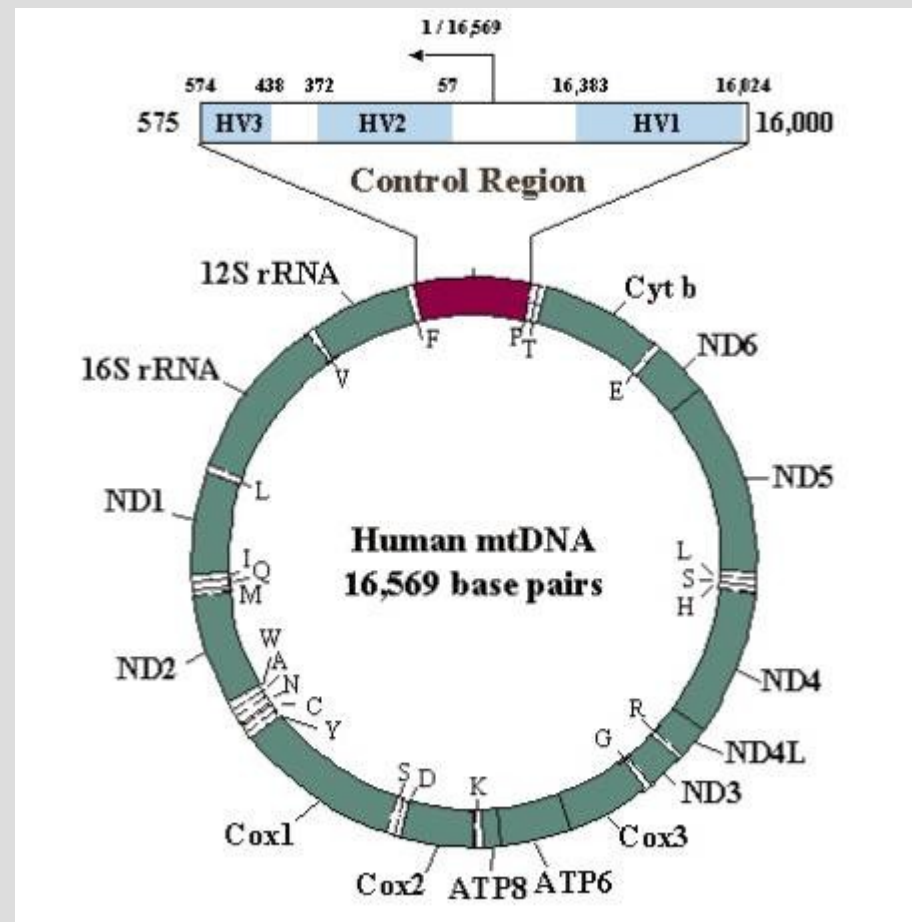


Genom

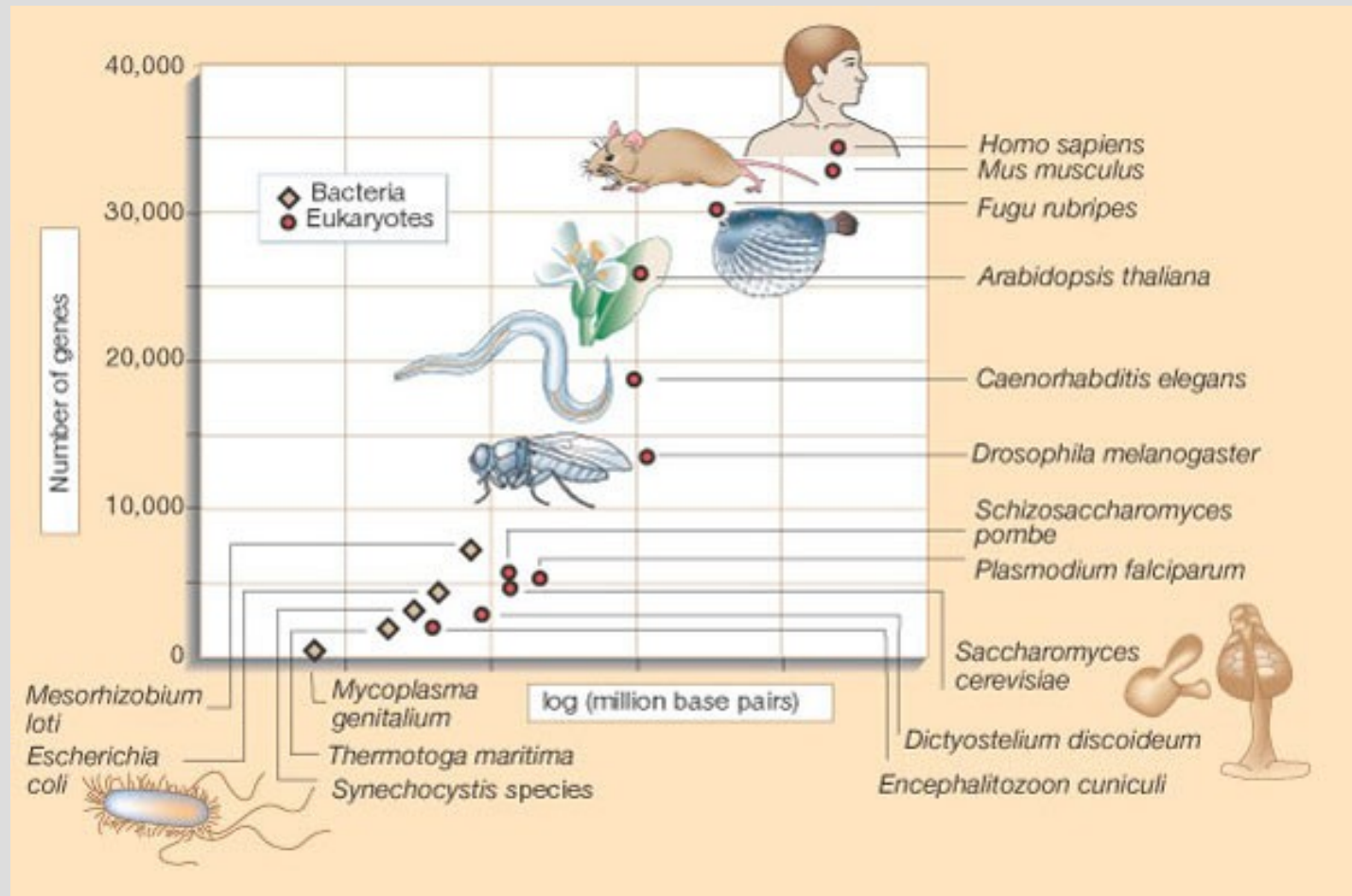
- *Saccharomyces cerevisiae*: 12 Mbp
 - zsekwencjonowany: 24 kwiecień 1996 (w 7 lat)
 - < 6000 genów
- Człowiek (osiągnięcie bioinformatyczne)
 - 3 mld USD
 - 14 kwiecień 2003
 - ok. 3300 Mbp, 20-25k genów
- Znamy genomy wielu "modelowych" organizmów: *E. coli*, *D. melanogaster*, *C. elegans*, *A. thaliana*
- dużo kodu genetycznego, którego nie rozumiemy (powtórzenia, frag. niekodujące)

Mapa Genomu

- lokalizacja odpowiednich genów na chromosomach
- bioinformatyka: przetwarzanie tych danych
- kryminalistyka i nie tylko



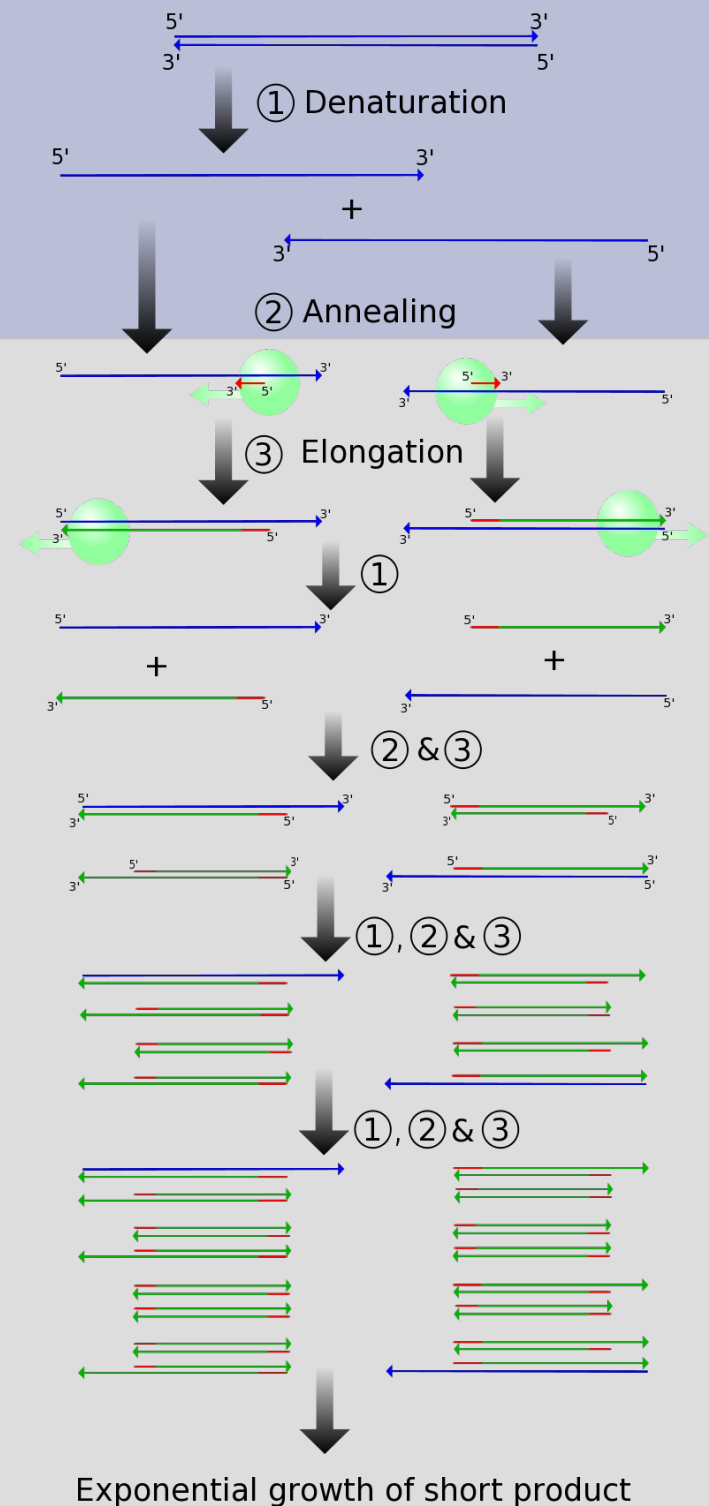
Liczba genów a rozmiar genomu



Techniki Biologii Molekularnej

PCR - łańcuchowa reakcja polimerazy

- Kary Mullis ~ 1983
- Pomysł prosty:
 - denaturacja
 - przyłączanie primerów
 - elongacja
- 25-35 razy
- Informatyka: projektowanie primerów
- Działa dla DNA, ale...



Techniki Biologii Molekularnej

Elektroforeza

- Analiza - rozdzielanie DNA, RNA, białek
- Żel - agarozowy, poliakrylamidowy
- Pomysł równie prosty jak PCR
 - denaturacja
 - zanurzyć (np.) białko w żelu
 - przyłożyć różnicę potencjałów
 - poczekać
- Uzyskujemy rozdział względem masy
- Barwienie, utrwalanie, fotka

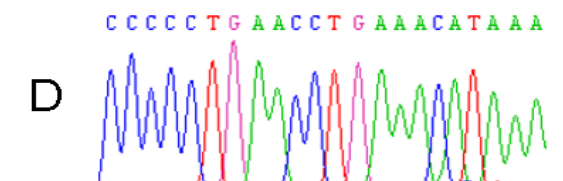
Zastosowania - Sekuencjonowanie

- Sekwencjonowanie DNA
- Dwa podejścia:
 - cięcie
 - synteza
- Materializacja:
 - metoda Maxama-Gilberta
 - metoda Sangera
- Bioinformatyczna metoda:
 - shotgun (*H. influenzae*)

A 5' CCTCCCACACCTCCCCCTGAACCTGAAACATAAA
3' GGAGGGTGTGGAGGGGGACTTGGACTTTGTATTT

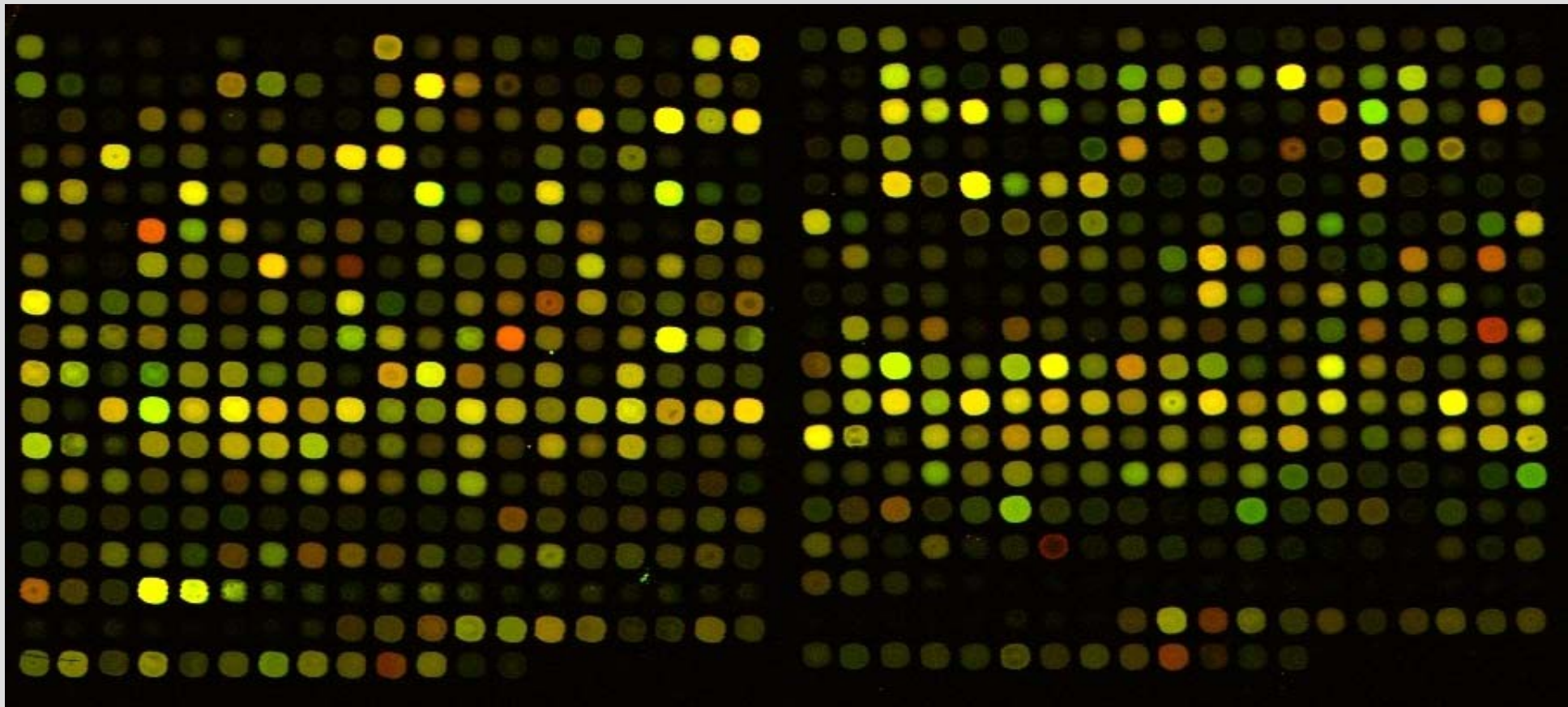
B 5' CCTCCCACACCT----->----->----->
3' GGAGGGTGTGGAGGGGGACTTGGACTTTGTATTT

C 5' CCTCCCACACCTC
5' CCTCCCACACCTCC
5' CCTCCCACACCTCCC
5' CCTCCCACACCTCCCC
5' CCTCCCACACCTCCCCC
5' CCTCCCACACCTCCCCCT
5' CCTCCCACACCTCCCCCTG
5' CCTCCCACACCTCCCCCTGA
5' CCTCCCACACCTCCCCCTGAA
5' CCTCCCACACCTCCCCCTGAAC
5' CCTCCCACACCTCCCCCTGAACC
5' CCTCCCACACCTCCCCCTGAACCT
5' CCTCCCACACCTCCCCCTGAACCTGAA
5' CCTCCCACACCTCCCCCTGAACCTGAA
5' CCTCCCACACCTCCCCCTGAACCTGAAAC
5' CCTCCCACACCTCCCCCTGAACCTGAAACAT
5' CCTCCCACACCTCCCCCTGAACCTGAAACATA
5' CCTCCCACACCTCCCCCTGAACCTGAAACATAA
5' CCTCCCACACCTCCCCCTGAACCTGAAACATAAA



Zastosowania - Mikromacierze

- Badanie ekspresji genów
- Pomysł: dużo, szybko, tanio



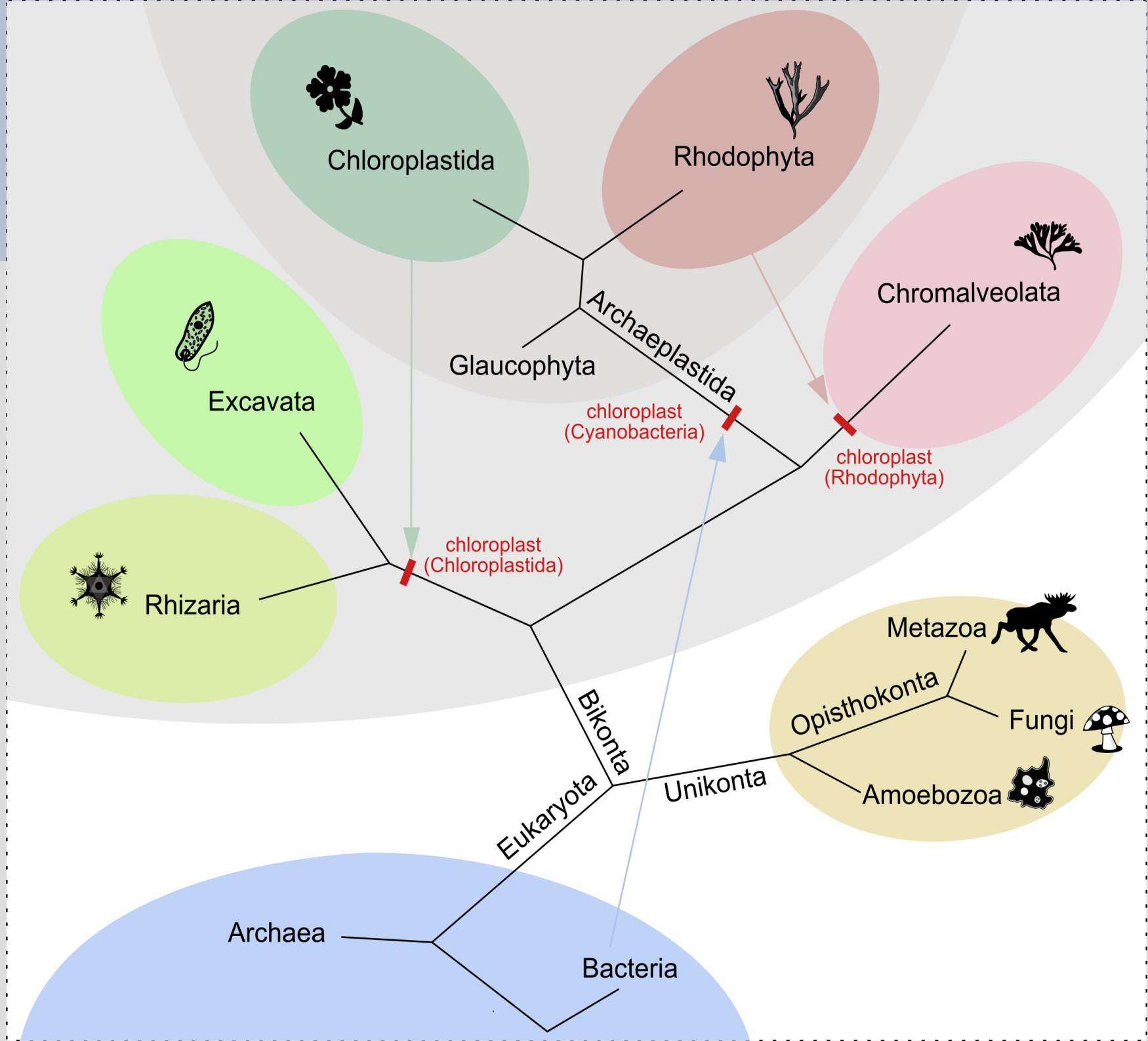
Mikromacierze

- Problem dla informatyków: analiza dużej ilości danych
- Wiele sond (do 6M na mikromacierzy)
- Statystyka
- Eksploracja danych

Dużo danych dostępne w bazach, pozostaje je przeanalizować, i wyciągnąć wnioski.

Zastosowania - Filogenetyka

- Analiza sekwencji DNA/białek w celu ustalenia ewolucji między organizmami
- Najczęściej: analiza zmienności kilku (!) genów między pewną liczbą grup taksonomicznych



Zastosowania - Filogenetyka

- Analiza sekwencji DNA/białek w celu ustalenia ewolucji między organizmami
- Najczęściej: analiza zmienności kilku (!) genów między pewną liczbą grup taksonomicznych
- Tajemnicze strzałki na obrazku?
 - endosymbiozy
- Szczególny przykład poziomego transferu genów (HGT)
- W tym, wykrywanie transpozonów

Bioinformatyka

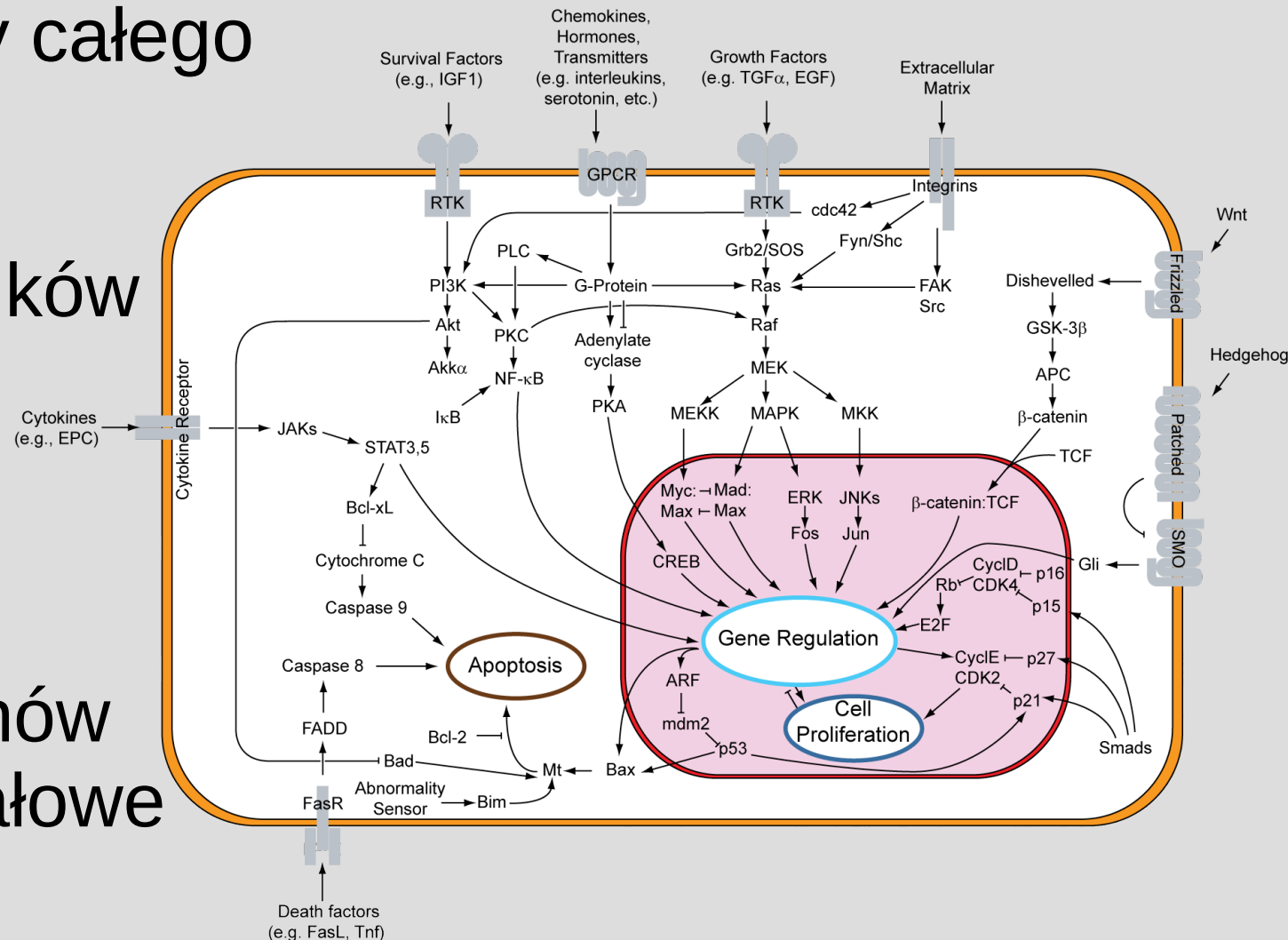
- Analiza sekwencji, ich dopasowywanie, porównywanie, określanie podobieństwa
- Sekwencjonowanie nadal wyzwaniem
- Obliczeniowa biologia ewolucyjna:
 - budujemy modele organizmów (ich DNA)
 - "wirtualizujemy" całą populację
 - uruchamiamy symulację
- Pomiary bioróżnorodności, analiza populacji, zachowanie genomu zagrożonych gatunków

Bioinformatyka

- Analiza ekspresji i regulacji genów
- Przewidywanie przestrzennej struktury białek
- Analiza rearanżacji genomu w komórkach nowotworowych
- Wysoko-wydajne przetwarzanie obrazów
- Łączenie białko-białko

Biologia systemów

- próba analizy całego organizmu
- i wielu czynników na niego działających
- metabolizm
- regulacja genów
- ścieżki sygnałowe



Skąd się biorą dane?

- (Bio)informatyk operuje na danych
- NCBI - dużo baz...
- ... oraz dużo algorytmów
 - BLAST
 - FASTA (jest też format pod tą samą nazwą)
 - i ich liczne odmiany
- Cel: porównywanie ciągów literek?
- Nie tylko: projektowanie primerów, miejsc restrykcyjnych...
- **BARDZO DUŻO DANYCH**

Przykład

>YKL218C Chr 11

```
ATGATTGTTCCCACTTATGGAGACGTTTTGGATGCCAGCAACAGAATTAAAGAATATGTA
AATAAAACACCGGTTCTCACTTCACGAATGCTTAATGATCGACTTGGAGCACAAATATAC
TTTAAAGGTGAGAATTTCCAACGAGTGGGAGCGTTCAAGTTTCGTGGAGCAATGAATGCT
GTTTCAAAATTAAGTGATGAAAAAAGAAGTAAGGGAGTAATTGCCTTCTCATCAGGAAAC
CATGCACAGGCTATTGCTCTTAGTGCAAACTATTAAATGTACCTGCAACAATTGTTATG
CCCGAGGATGCGCCCCGCCCTTAAAGTTGCTGCTACAGCCGGTTACGGAGCACATATCATA
AGGTATAACAGGTATACTGAAGATCGCGAGCAGATTGGGCGTCAACTAGCAGCCGAACAT
GGTTTTGCATTGATTCCGCCCTACGATCATCCTGATGTTATTGCAGGGCAAGGTACGTCG
GCAAAAGAGCTATTAGAAGAGGTTGGACAACCTTGATGCATTATTTGTTCTTTGGGCGGT
GGTGGGCTCCTTTTCAGGATCTGCACTTGCCGCTAGAAGCCTTTCTCCAGGCTGCAAAATT
TTTGGGGTTGAACCTGAAGCTGGTAACGATGGACAACAATCCTTCAGATCGGGTTCCATT
GTTTCATATCAATACGCCAAAACTATCGCAGATGGCGCTCAAACACAACACCTCGGTGAG
TACACATTTGCCATTATTCGCGAAAATGTCGATGATATTTTAACGGTTAGCGACCAAGAG
CTAGTAAAATGCATGCACTTTCTTGCGGAACGCATGAAGGTGGTTGTTGAGCCACAGCT
TGTTTGGGATTTGCAGGTGCACTTCTAAAAAAGGAAGAGCTAGTTGGGAAGAAAGTAGGC
ATAATACTAAGTGGAGGTAATGTAGACATGAAGAGATATGCTACTTTAATCTCTGGGAAG
GAAGATGGCCCAACGATTTAG
```

dehydrataza 3-hydroksyaspertazy, odpowiedzialna za deaminację
L-threo-3-hydroksyaspertatu do oksaloacetatu oraz amoniaku

...

Literatura

- P. Węgleński i in.: Genetyka molekularna, PWN 2006
- R.F. Doolittle: *The parasite genome, the grand assault*, Nature 419 (493-494, 3 Oct 2002)
- <http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html>
- <http://www.ncbi.nlm.nih.gov/>
- obrazki: google image search

Pytania?

?

Pytania?

?

Poza tym, gdzie jest Wydział Biologii UW ;-)