Statystyczna analiza danych molekularnych: wykład 1

Anna Gambin

Instytut Informatyki, Uniwersytet Warszawski

Warszawa, 21 luty 2007

Grupa Biologii Obliczeniowej

laboratorium: *Statystyczna analiza danych w języku R*, **Janusz Dutkowski** – czwartek 14.15.

Organizacja wykładu

- Statystyczna analiza danych molekularnych:
  - wykład: teoria 80% (NOTATKI) + motywacje biologiczne 20% (SLAJDY),
  - ćwiczenia (zadania: informatyka, statystyka, prezentacja projektu),
  - laboratorium – projekt (informatyka w biomedycynie);

Reguły zaliczeń

- wykład: egzamin ustny.
- ćwiczenia: prezentacja projektu.
- laboratorium: wykonany projekt + krótki opis.

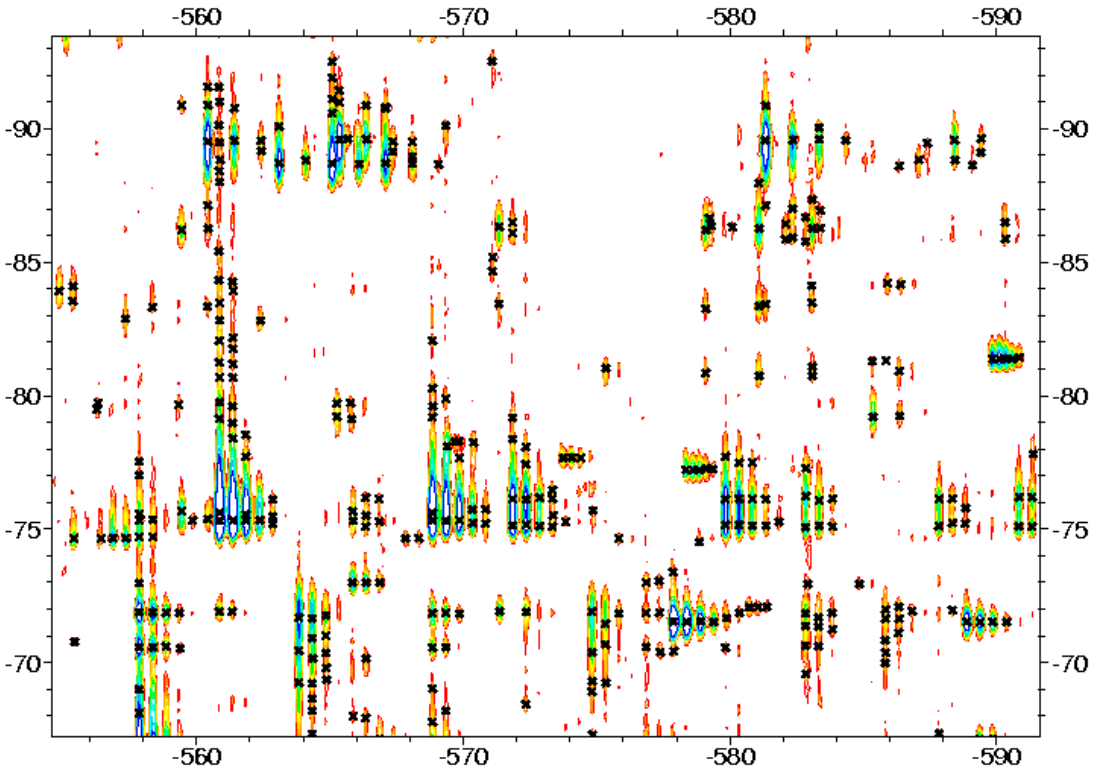
PROJEKTY

- mikromacierze (Centrum Onkologii - Instytut im. Marii Skłodowskiej-Curie):
  - rak tarczycy: dane prof. Barbary Jarząb.
  - przełyk Baretta: dane prof. Jerzego Ostrowskiego.
- spektrometria masowa (Środowiskowe Laboratorium Spektrometrii Mas, PAN):
  - mukowiscydoza: dane prof. Michała Dadleza.
  - rak jelita grubego: dane prof. Jerzego Ostrowskiego.

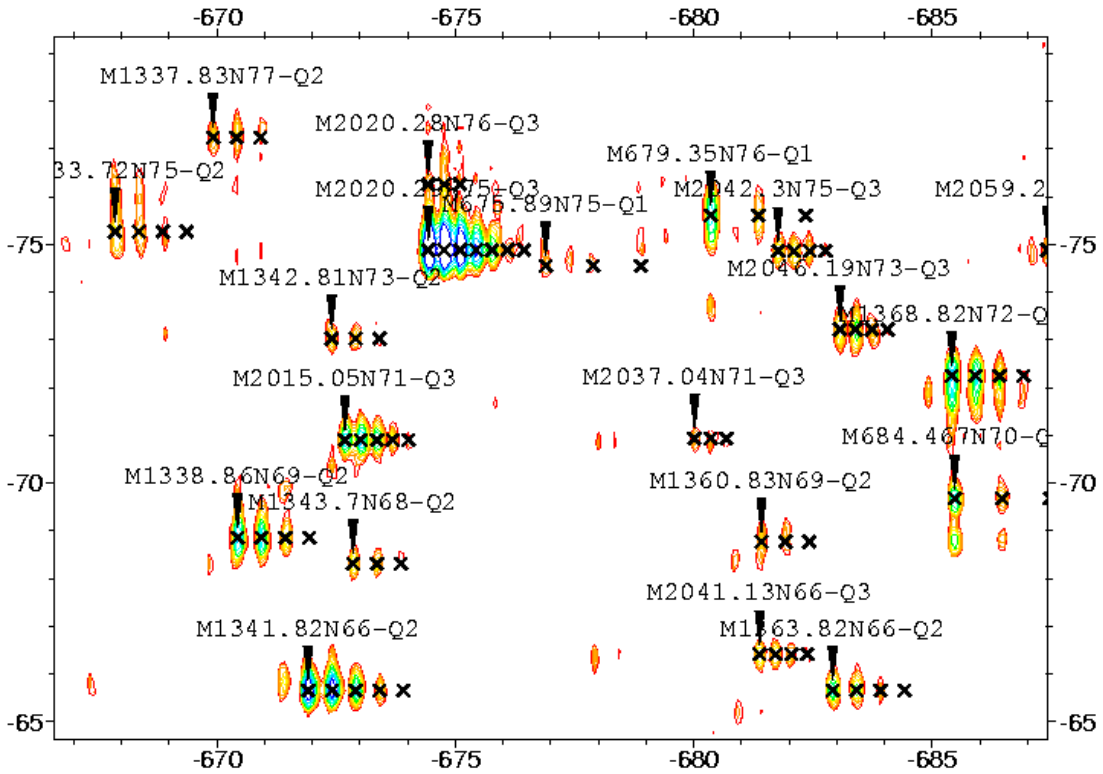
PROJEKTY, c.d.

- preprocessing (metody standardowe, bez wgłębiania się).
- klasteryzacja (uliniowienie),
- selekcja cech (biomarkerów), redukcja wymiaru,
- ocena statystycznej istotności cech dyskryminujących,
- klasyfikacja (test diagnostyczny).

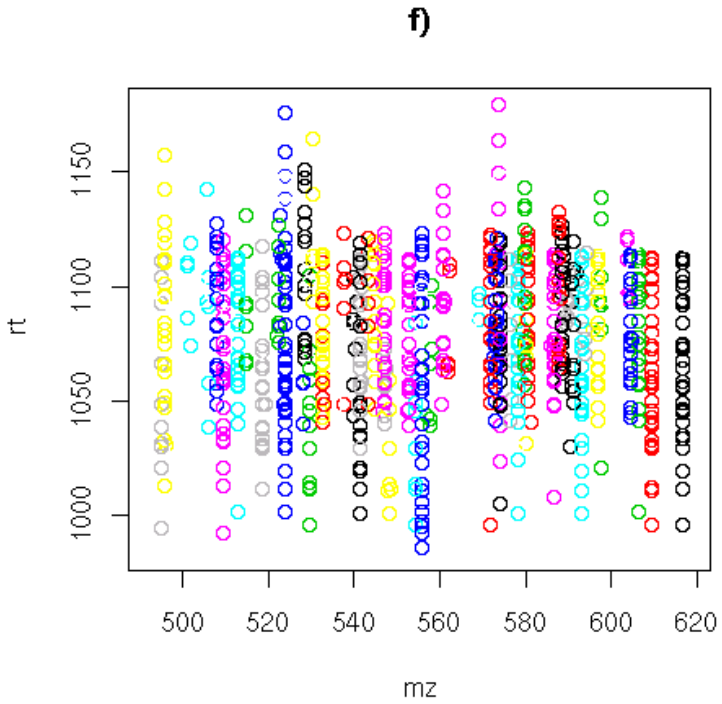
dane



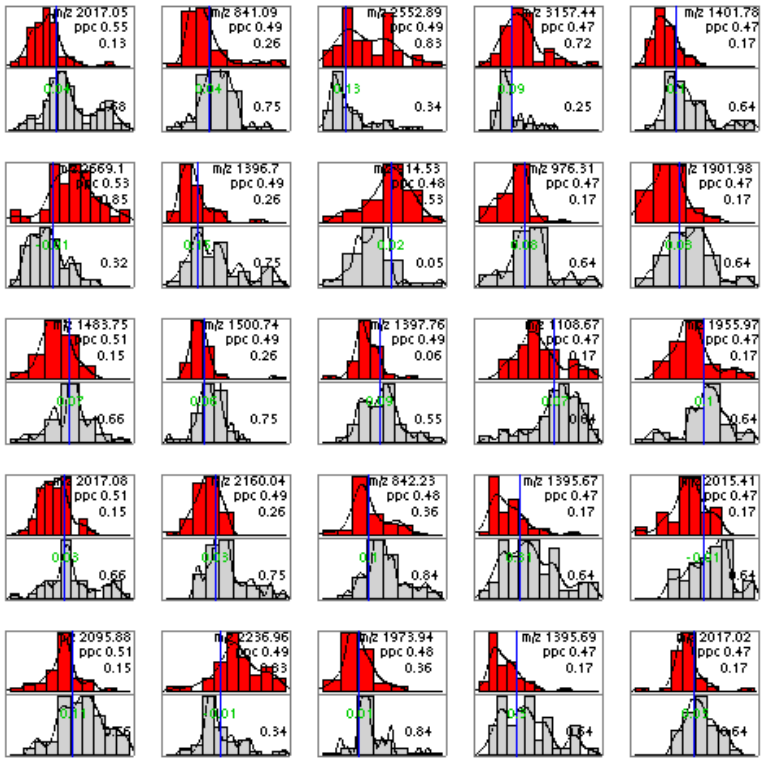
preprocessing



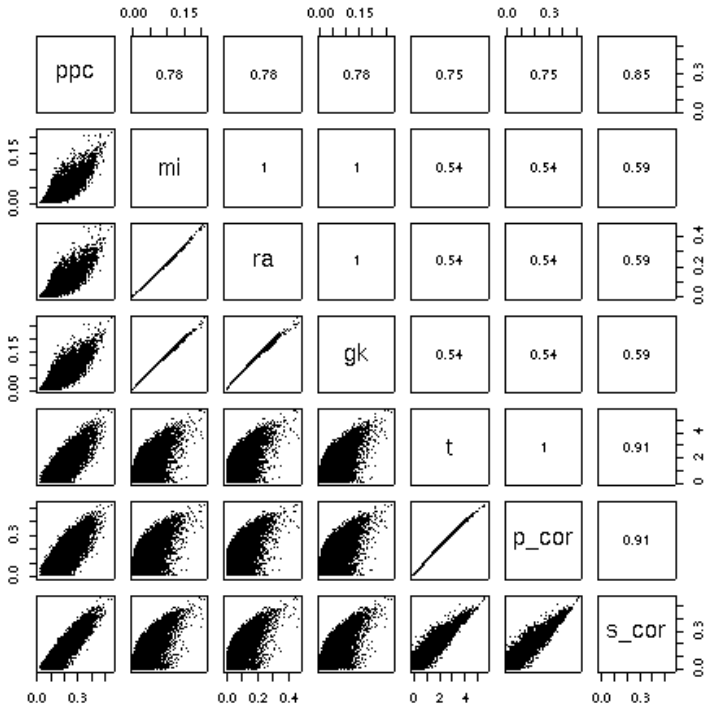
klasteryzacja



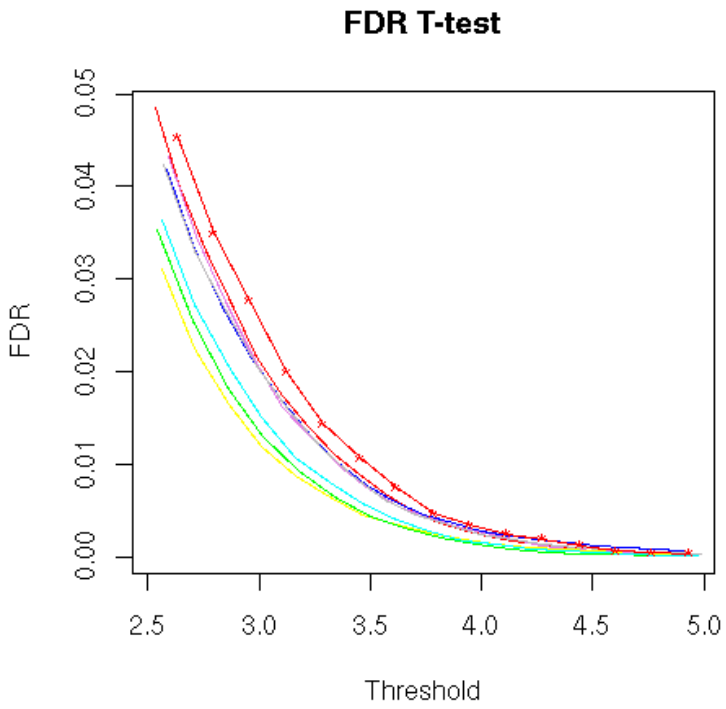
selekcja cech



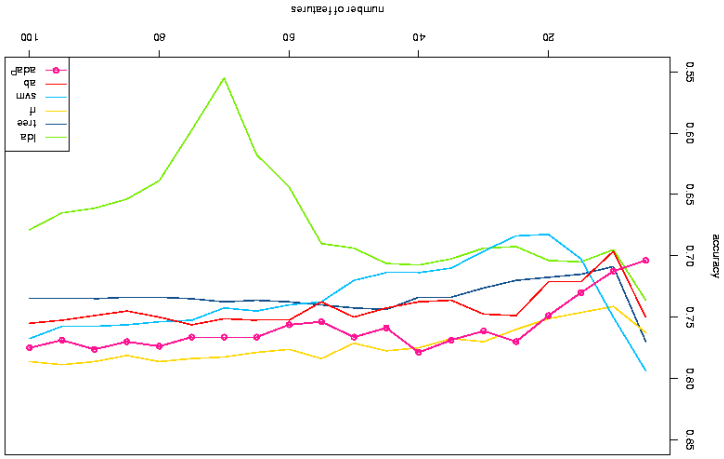
selekcja cech, cd



ocena cech


FDR T-test
FDR
0.00
0.01
0.02
0.03
0.04
0.05
2.5
3.0
3.5
4.0
4.5
5.0
Threshold

klasyfikacja



Zgrubny plan wykładów

- wykład 1: krótki wstęp biologiczny.
- wykład 2: wstęp do technologii mikromacierzy i spektormetrii mas, zastosowania w biologii i medycynie
- wykład 3 – 5: klasteryzacja i biklasteryzacja danych: alg. k-średnich, podejścia grafowe, hierarchiczne, mieszanki gaussowskie.
- wykład 6: zagadnienie redukcji wymiaru danych.
- wykład 7: selekcja cech (biomarkerów), ocena jakości cech,
- wykład 8 – 12 klasyfikacja - metody najbliższych sąsiadów, klasyfikator bayesowski, LDA, QDA, boosting, drzewa decyzyjne, metoda SVM, ocena jakości klasyfikatorów.

Literatura

- Trevor Hastie, Robert Tibshirani, Jerome Friedman The Elements of tatistical Learning:Data Mining, Inference, and Prediction, Springer-Verlag, 2001.
- Statistical Methods in Bioinformatics Warren J. Ewens, Gregory R. Grant, Springer-Verlag, 2001.
- Bioinformatics and Computational Biology Solutions Using R and Bioconductor (2005). Edited by R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit. Springer Series in Statistics for Biology and Health.

Plan na dziś

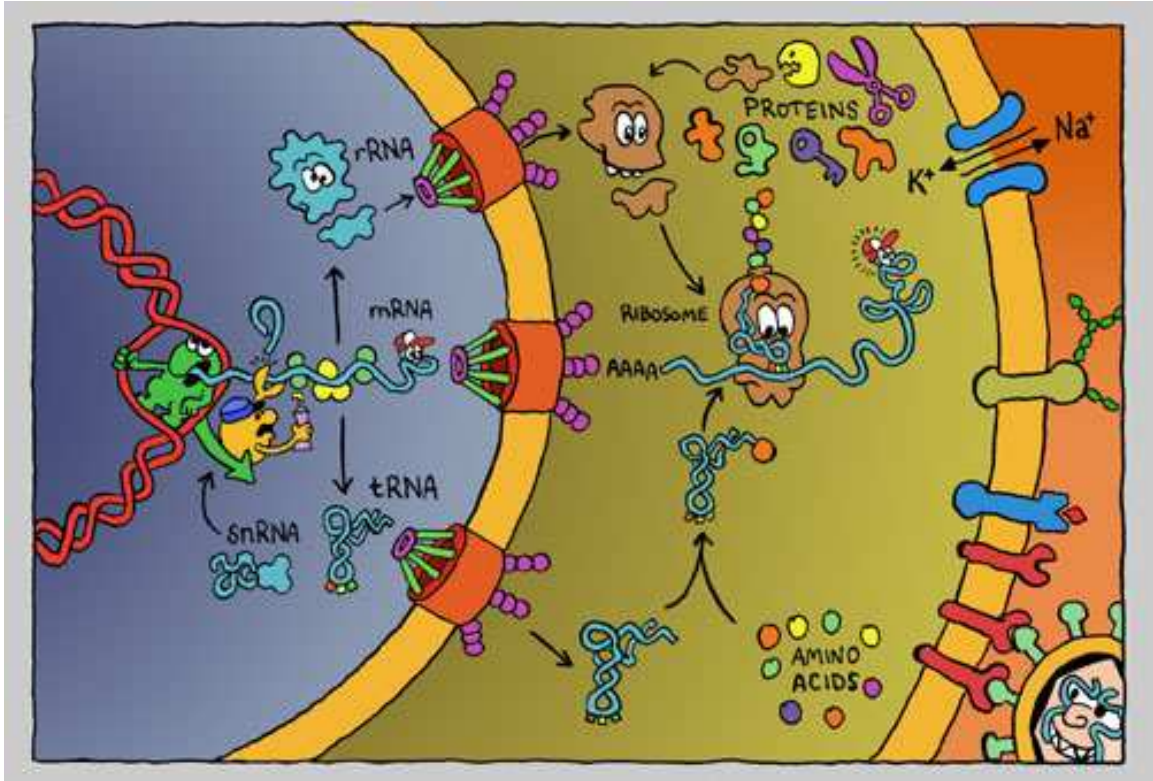
- transkrypcja, translacja...
- struktura, funkcja białka...
- organizacja chromosomów...
- co to jest alternatywny *splicing*?
- czego dowiedzieliśmy się z HGP ?
- genomika fukcjonalna, biologia systemów: czym się to je?

Centralny dogmat ...

$$\text{DNA} \rightsquigarrow \text{RNA} \rightsquigarrow \text{białko}$$

- przejście od DNA do RNA – **transkrypcja**,
- przejście od RNA do białka – **translacja**,
- geny prokariotyczne: spójne fragmenty DNA przepisywane na mRNA (*messenger RNA*),
- geny eukariotyczne: poprzedzielane niekodującymi fragmentami (**intronami**); fragmenty kodujące – **eksony**.
- komplementarne DNA (**cDNA**) otrzymuje się z mRNA przy pomocy enzymu **odwrotnej transkryptazy**, cDNA nie zawiera juz intronów.

centralny dogmat



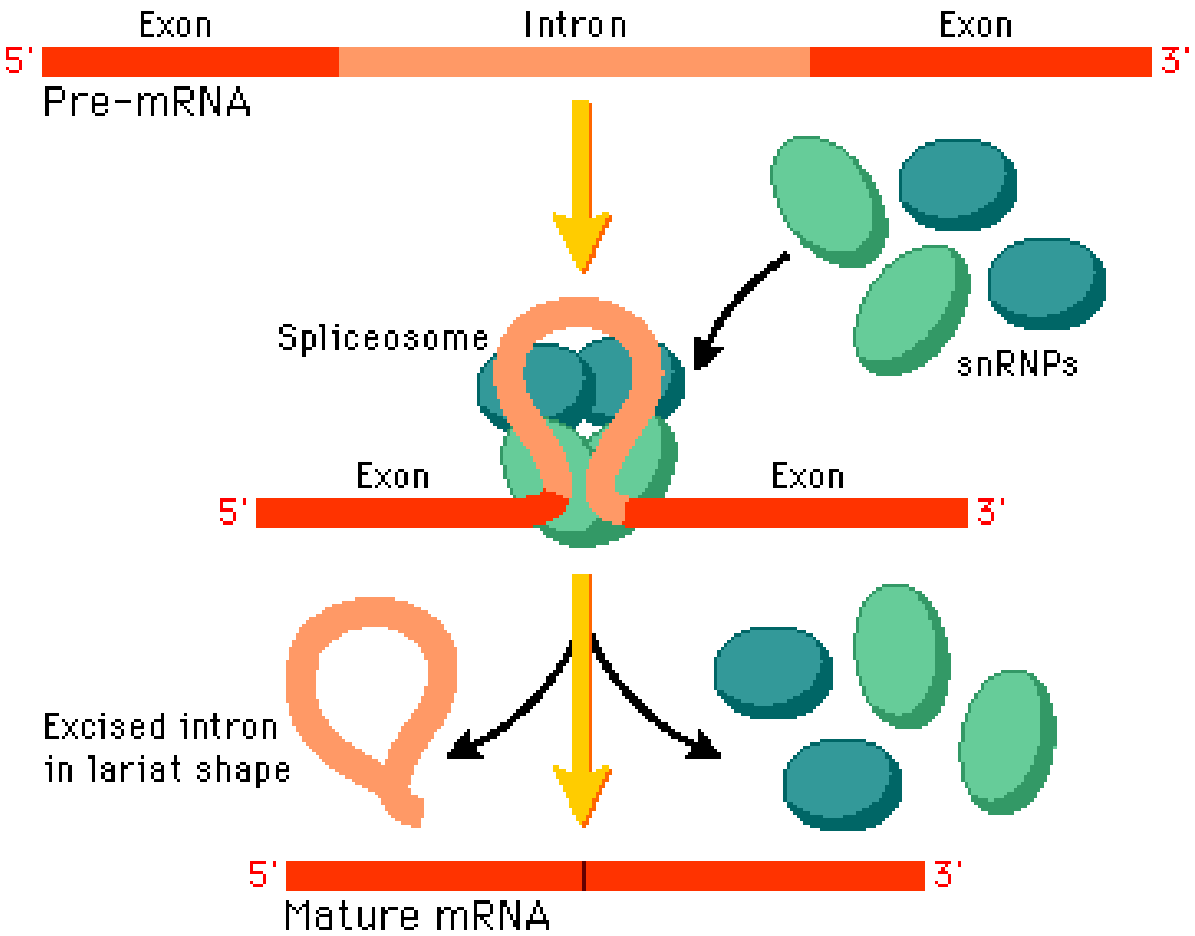
genetyka i historia

- **1**866: eksperymenty Grzegorz Mendla.
- **1**944: Avery i McCarty wskazują na wiodąca rolę DNA w dziedziczeniu.
- **1**953: podwójna spirala Watsona i Cricka.
- **2**001: pierwszy opublikowany draft ludzkiego genomu: jednocześnie firma prywatna (Celera) i publiczny projekt (HGP).

transkrypcja: prokariota a eukariota

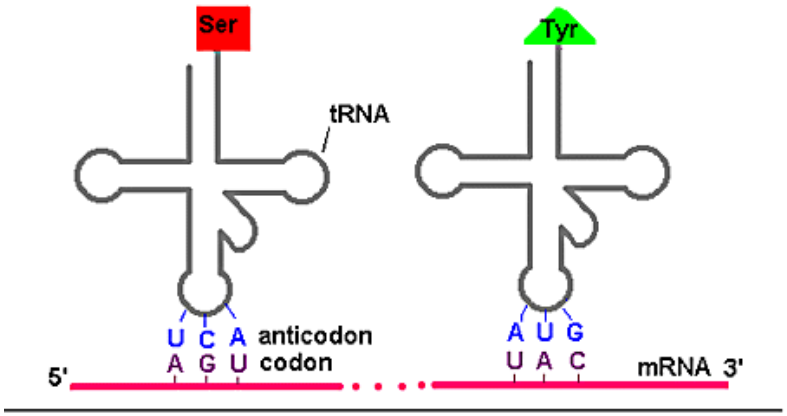
- prokariota: jeden etap DNA $\rightsquigarrow$ mRNA; czasami wspólnie regulowana transkrypcja całego **operonu** (zbioru genów odpowiedzialnego za jeden cel metaboliczny).
- eukariota: transkrypcja dwu-etapowo:
  - DNA $\rightsquigarrow$ pre-mRNA (zarówno introny jak i eksony);
  - pre-mRNA $\rightsquigarrow$ (*splicing* = wycinanie i sklejanie) mRNA.

splicing


Exon
Intron
Exon
5'
3'
Pre-mRNA
Spliceosome
snRNPs
Exon
Exon
5'
3'
Excised intron
in lariat shape
5'
3'
Mature mRNA

translacja: kod genetyczny

- aminokwas kodowany przez sekwencję 3 nukleotydów (czyli **kodon**).
- 20 rodzajów aminokwasów.
- kod jest uniwersalny i jednoznaczny.



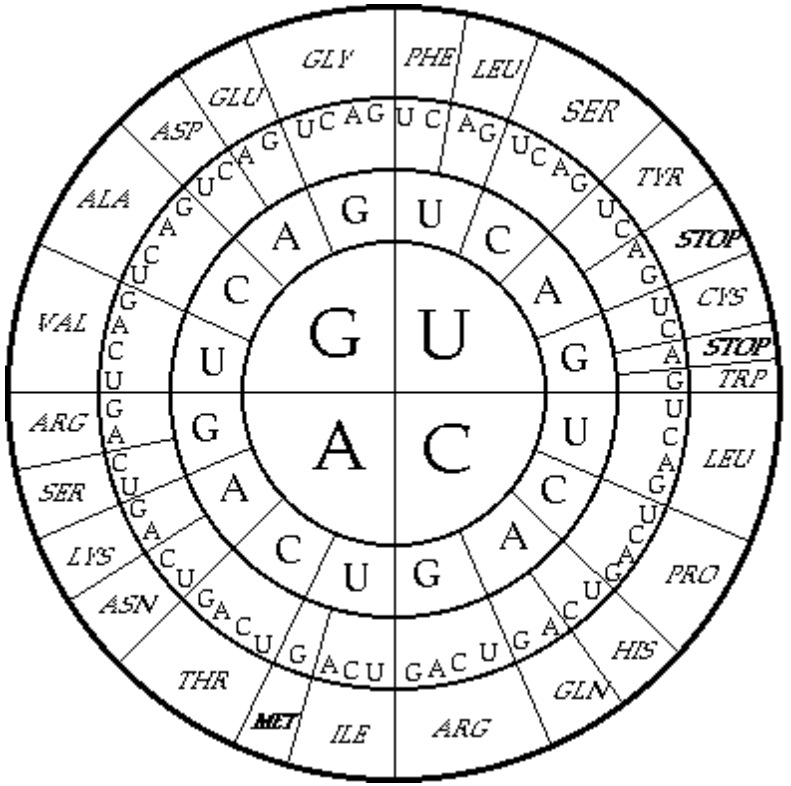
2nd base in codon

1st base in codon	U	C	A	G	3rd base in codon
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

The Genetic Code

kod genetyczny

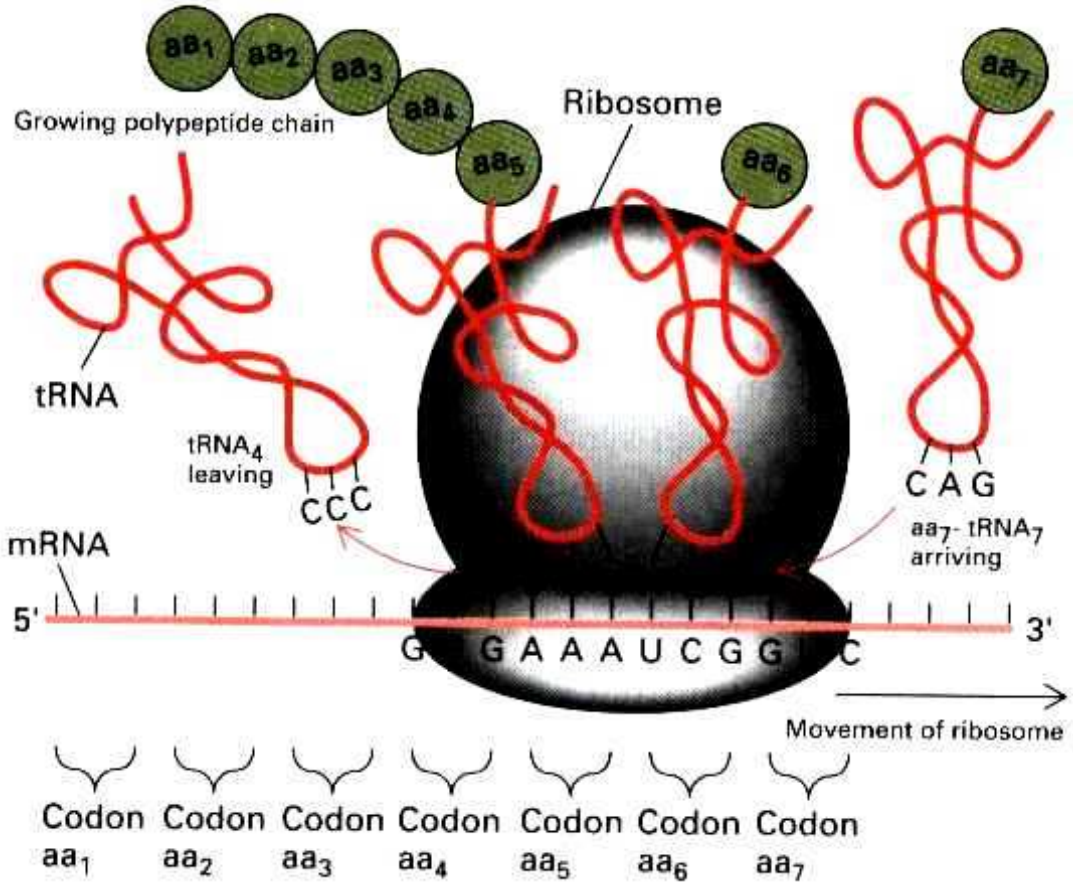
- Kod jest redundantny:
  np. **leucyna** ma 6 kodonów.
- kodon START = AUG
  (czyli **metionina**).
- STOP = UAA | UGA | UAG
  (nie kodują żadnego
  aminokwasu).



ORFy i translacja

- OFR (*open reading frame*) czyli **ramka odczytu** to ciąg nukleotydów rozpoczynający się kodonem START a kończący kodonem STOP.
- w translację zaangażowane: **mRNA**, **tRNA** (*transfer RNA*), **rRNA** (rybosomowe RNA).
- prędkość budowania białka przez **rybosom**: $3 - 5$ aminokwasów na sekundę.
- przeciętne białko ma $100 - 300$ aminokwasów (są też większe np. 30000)

translacja


aa1
aa2
aa3
aa4
aa5
aa6
aa7
Growing polypeptide chain
Ribosome
tRNA
tRNA4 leaving
CCC
CAG
aa7- tRNA7 arriving
mRNA
5'
3'
G GAAAUCGG C
Movement of ribosome
Codon aa1
Codon aa2
Codon aa3
Codon aa4
Codon aa5
Codon aa6
Codon aa7

struktura białka

- struktura pierwszorzędowa: ciąg aminokwasów.
- struktura drugorzędowa: alfa helisa, beta kartka (lokalne oddziaływania).
- struktura trzeciorzędowa: całkowity konformacja (*folding*).
- struktura czwartorzędowa: kryształy.

struktura białka

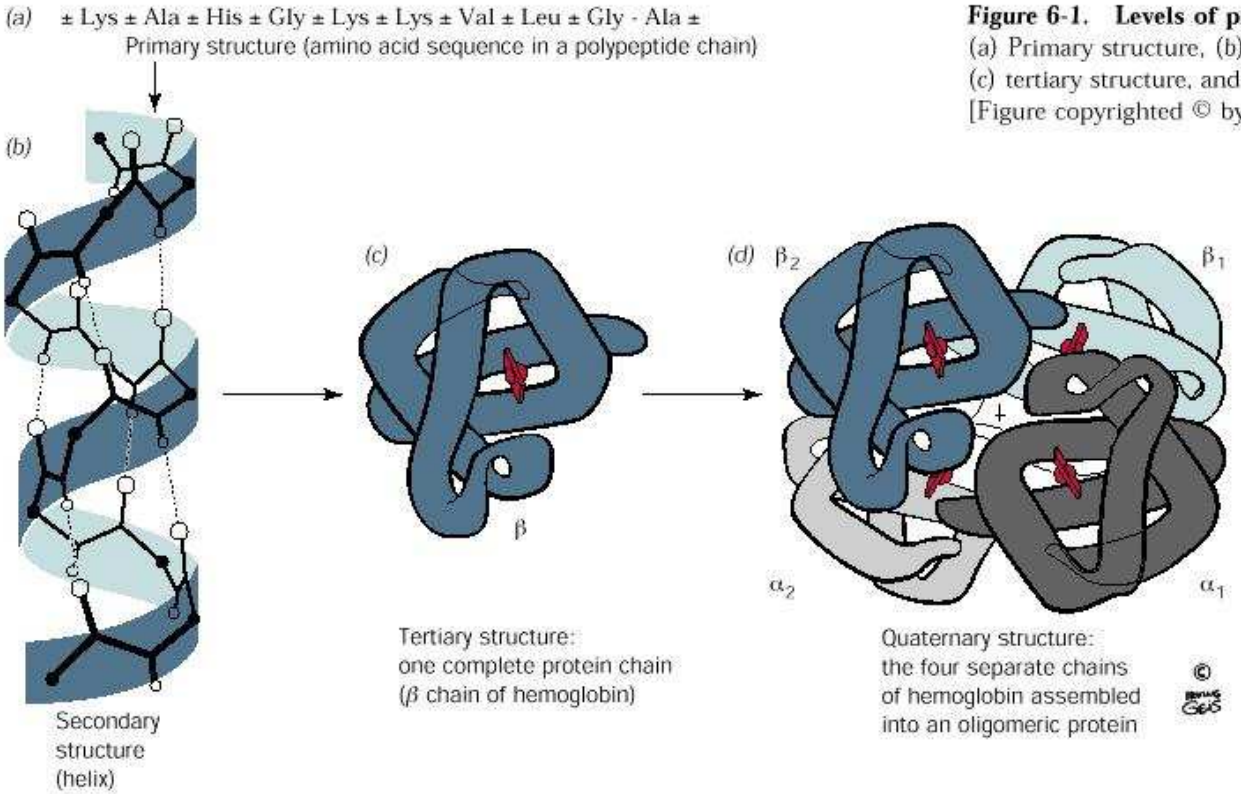
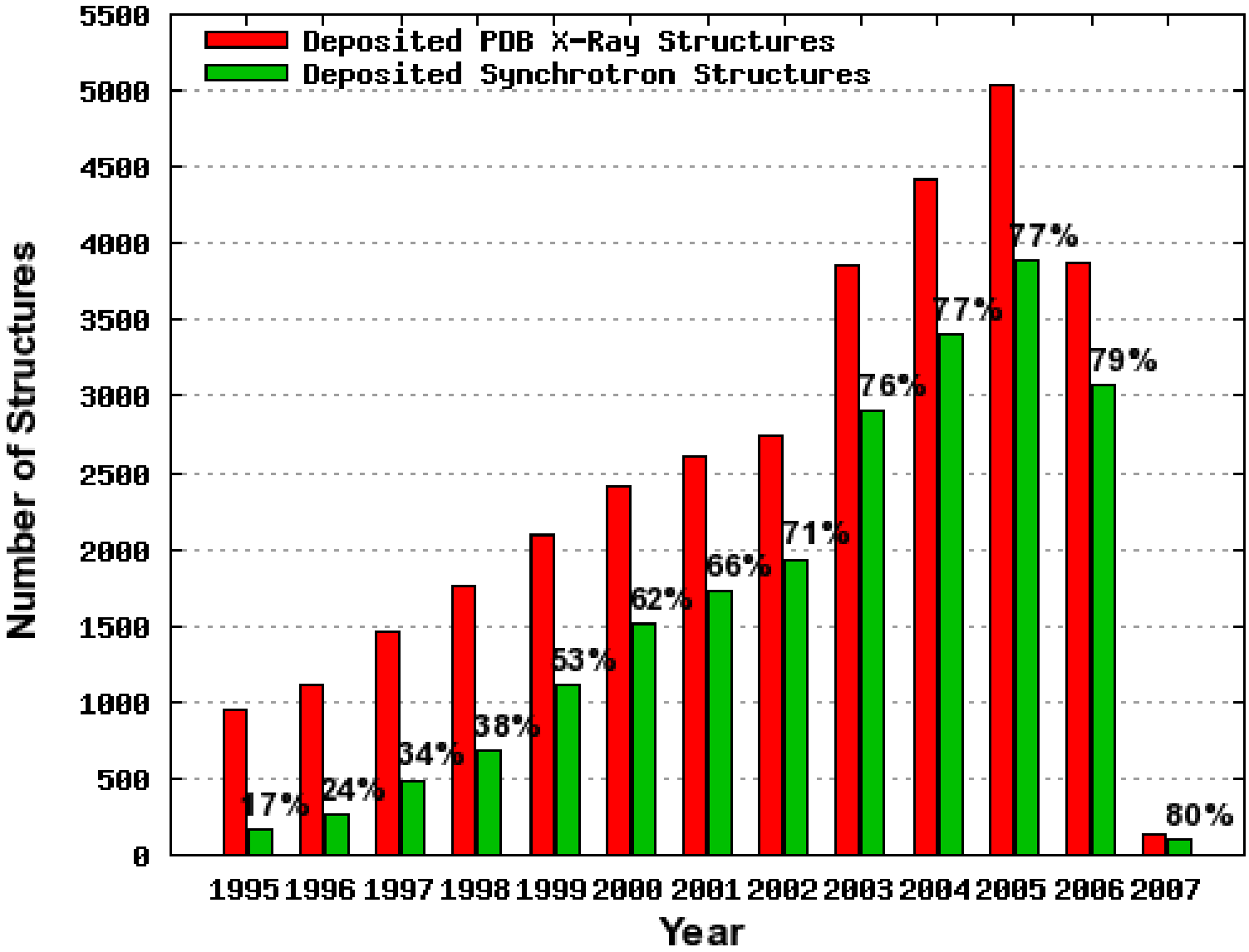


Figure 6-1. Levels of protein structure.
(a) Primary structure, (b) secondary structure, (c) tertiary structure, and (d) quaternary structure.
[Figure copyrighted © by Irving Geis.]

baza struktur



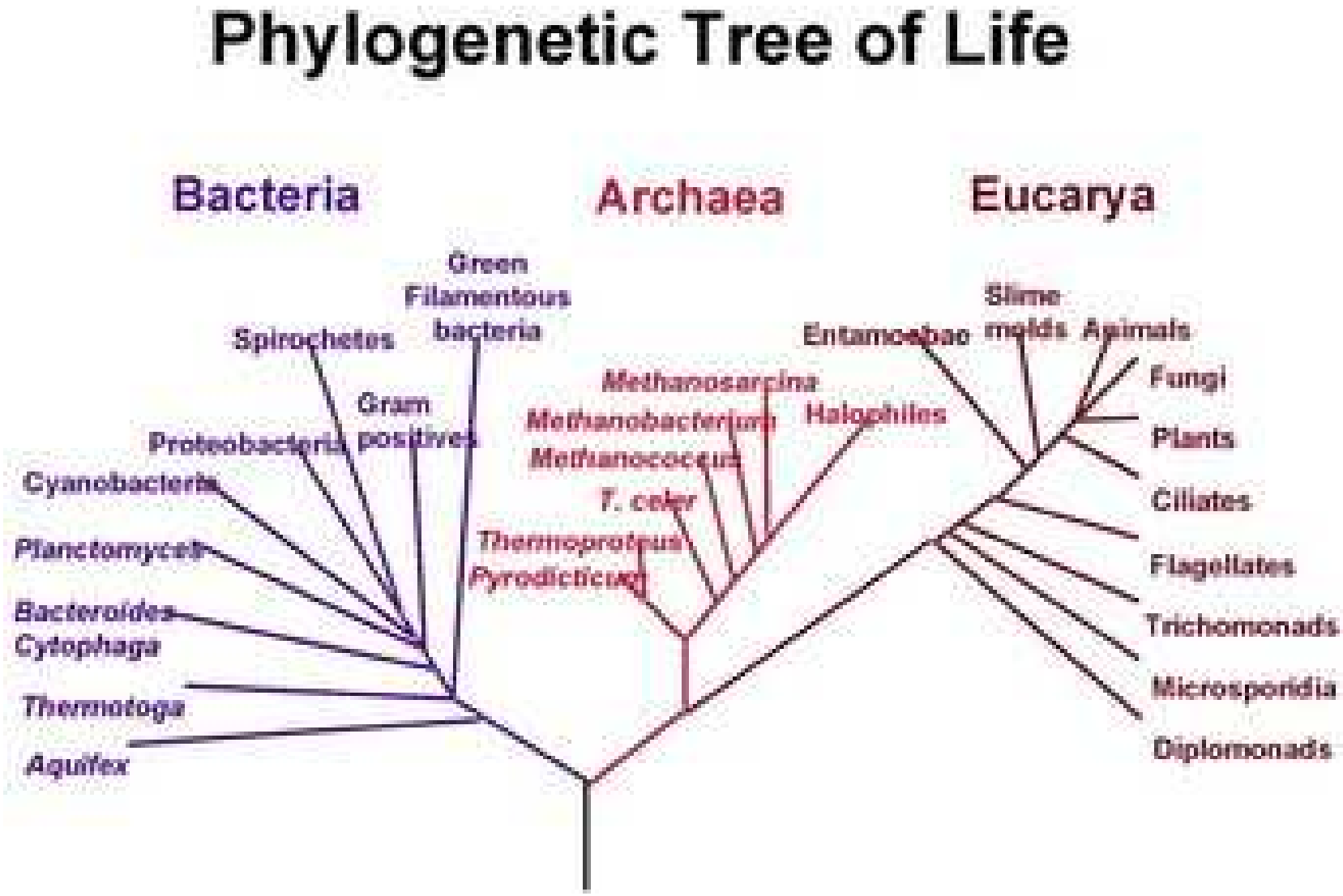
ORF (gen) $\rightsquigarrow$ struktura $\rightsquigarrow$ funkcja

- białka: budulec, enzymy, hormony, transport, przeciwciała, kontrola procesu, translacji i transkrypcji.
- struktura przestrzenna ma ogromny wpływ na funkcję białka.
- znane struktury w bazie PDB (*Protein Data Bank*).
- podobieństwo sekwencji vs podobieństwo struktur.
- centralny problem biologii obliczeniowej: przewidywanie struktury białka.

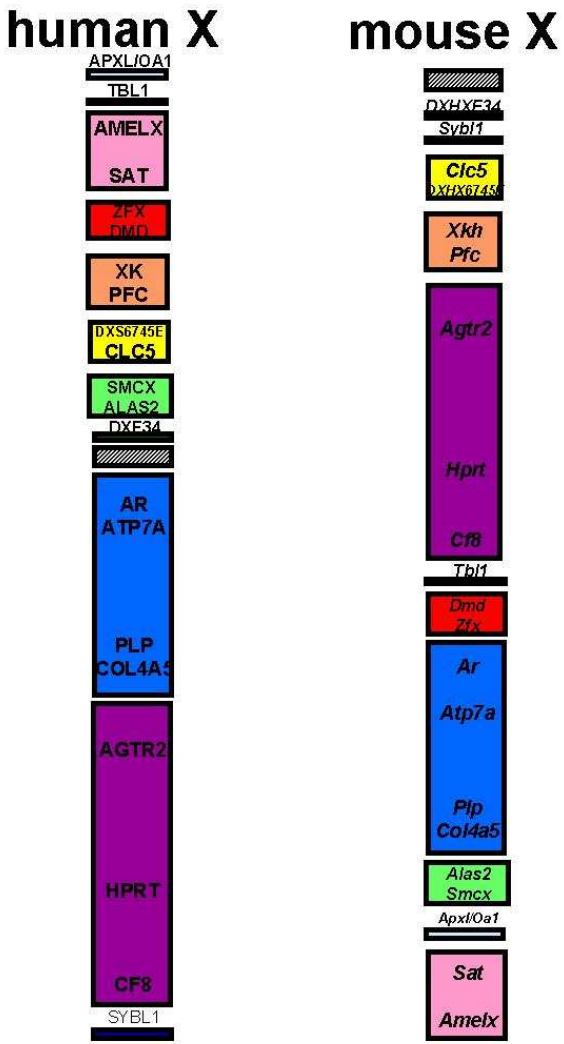
konserwacja białek

- drzewa **filogenetyczne** ilustrują:
  - zależności ewolucyjne genów lub organizmów,
  - różnice strukturalne białek.
- białka pochodzące od wspólnego przodka: **homologi**
- w tym samym genomie: **paralogi**,
- u różnych gatunków: **ortologi**.

drzewo życia

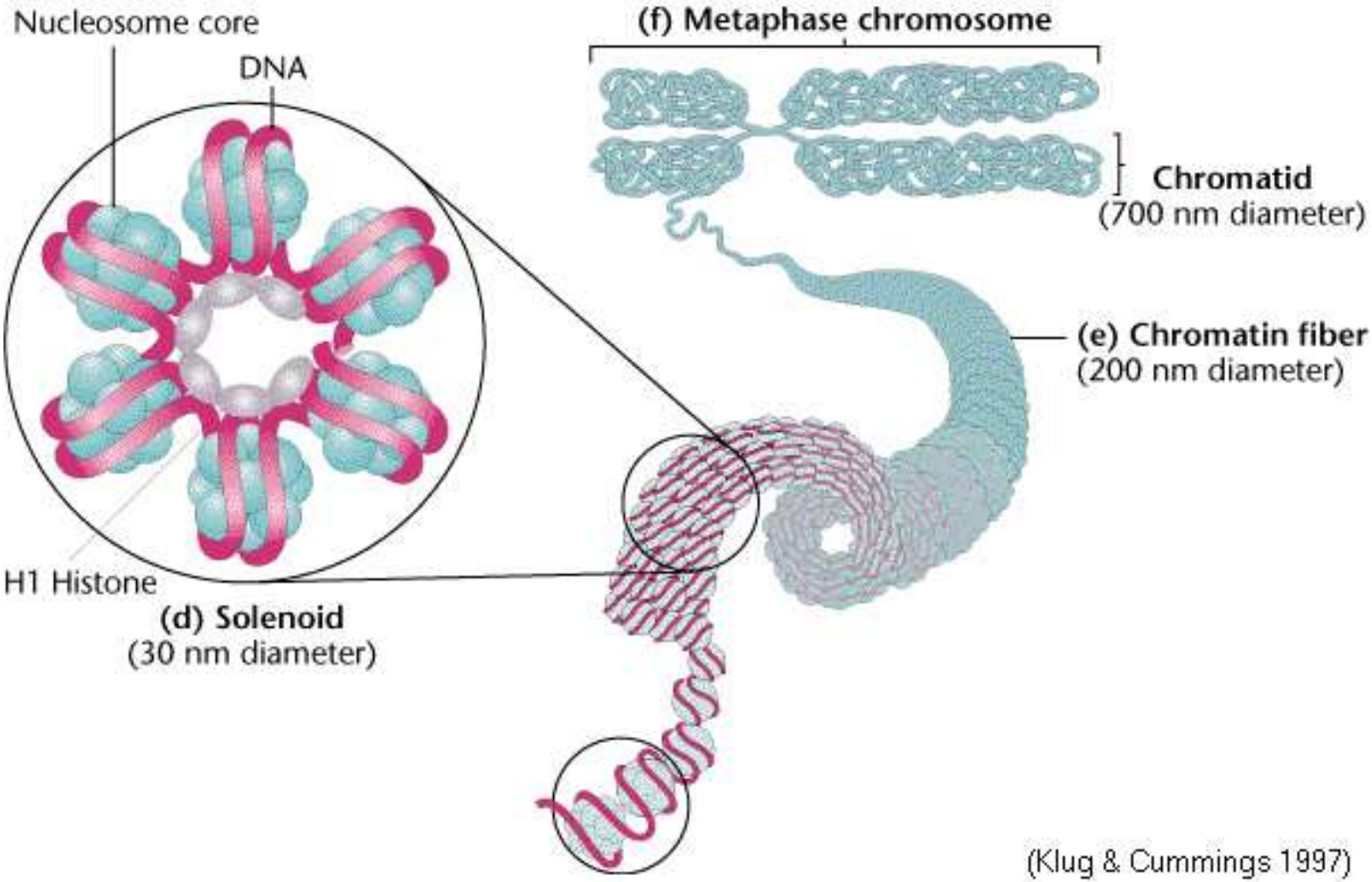


chromosomy



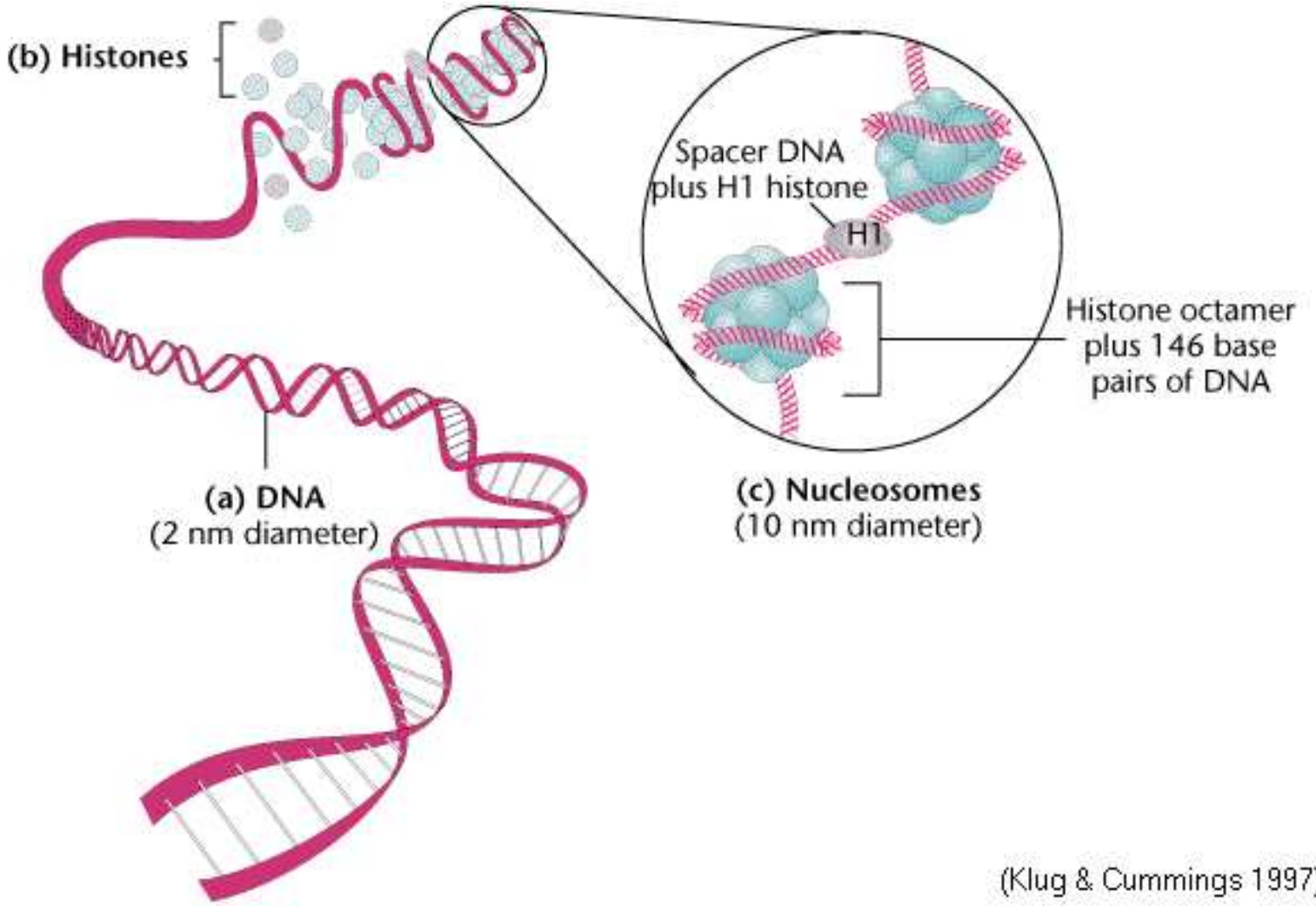
- człowiek ma 23 pary chromosomów.
- ziemniak: 24.
- pies: 39.
- karp: 52.

organizacja chromosomu



(Klug & Cummings 1997)

nukleosom



(Klug & Cummings 1997)

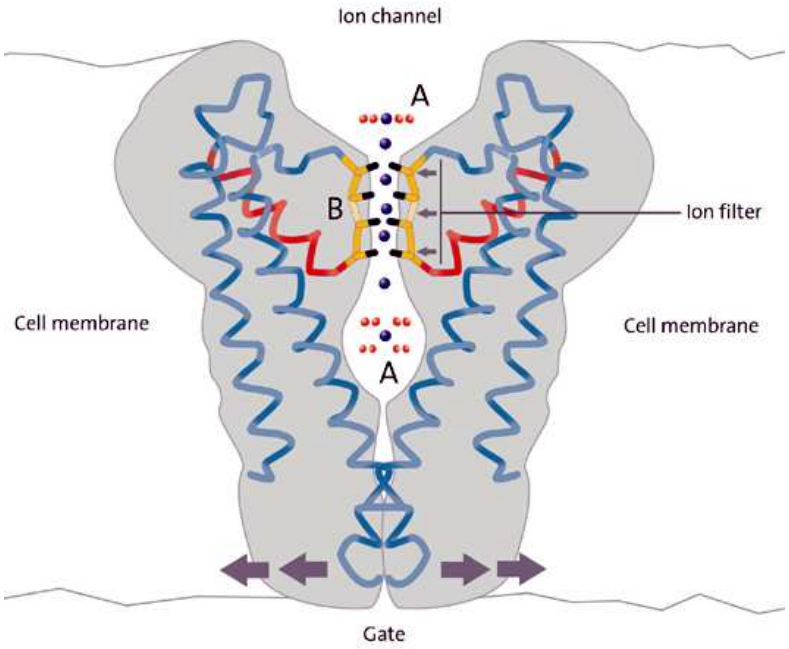
polowanie na geny

- co to jest gen ?
  - dawniej: każdy fragment DNA podlegający **transkrypcji**.
  - dziś: każdy **funkcjonalny** fragment DNA.
  - a nawet: każdy konserwowany fragment DNA.
- promotor, sekwencja kodujaca, kodon STOPu.
- 225 bp OFR / 1 kb DNA nie podlegającego transkrypcji.
- **pseudogen**: niefunkcjonalny (z powodu mutacji) OFR.
- GeneMark, GeneScan, Glimmer: automatyczne polowanie na geny.

alternatywny splicing, przykład

- gen Nox1 ulokowany w ludzkim chromosomie X generuje 3 różne mRNA.
- wersja dłuższa: NADPH (fosforan dinukleotydu nikotynoamidoadeninowego zredukowany :) oksydaza (enzym katalizujący reakcję utleniania),
- wersja krótsza (nie zawiera pewnych eksonów): kanał jonowy (wodorowy).
- zależy od tkanki, które białko powstanie.

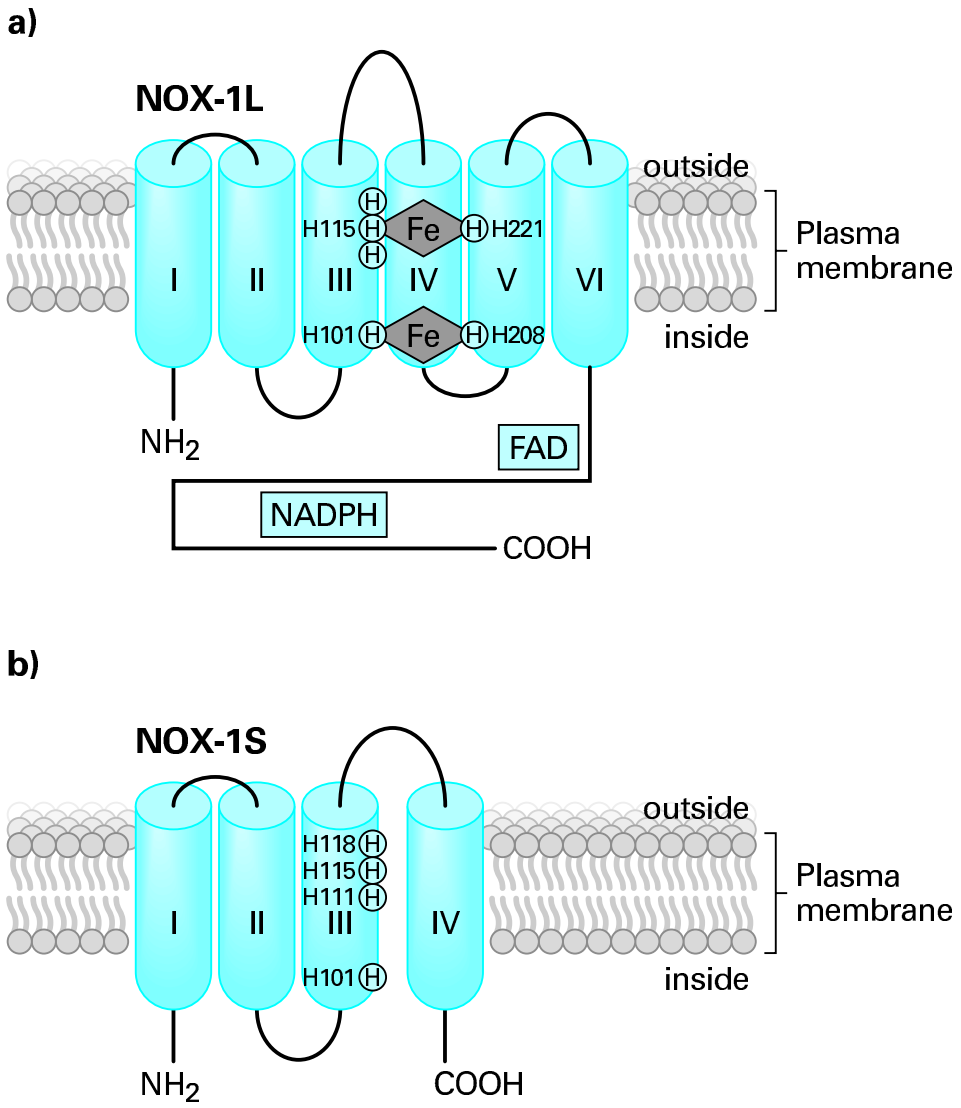
kanał jonowy



- badanie kanałow jonowych: 3 x Nobel,
- selektywne filrty,
- zaburzenia w działaniu (mutacje białka) prowadzą do ciężkich chorób (mukowiscydoza).

alternatywny splicing, c.d.

FIGURE 1.4 • Two proteins produced by alternative splicing of RNA. **a)** The full-length mRNA was translated to produce the larger protein (NOX-1L) that is an NADPH oxidase. **b)** Several exons were excised to produce a smaller protein (NOX-1S) that functions as a H^+ channel.



HGP luty 2001

- czyje DNA zostało zsekwencjonowne ? (HGP:8 dawców, 1 dawczyni :)
- 1 błąd na 10000 bp, DNA średnio 4 razy zsekwencjonowane.
- człowiek ma około 35000 genów.
- szybkość mutacji u mężczyzn około 2 razy większa niż u kobiet.
- rekombinacja częstsza bliżej *telomerów* i na krótszych ramionach chromosomów.
- podejrzenie **transferu horyzonatalnego** od bakterii w przypadku niektórych genów (niesłuszne).

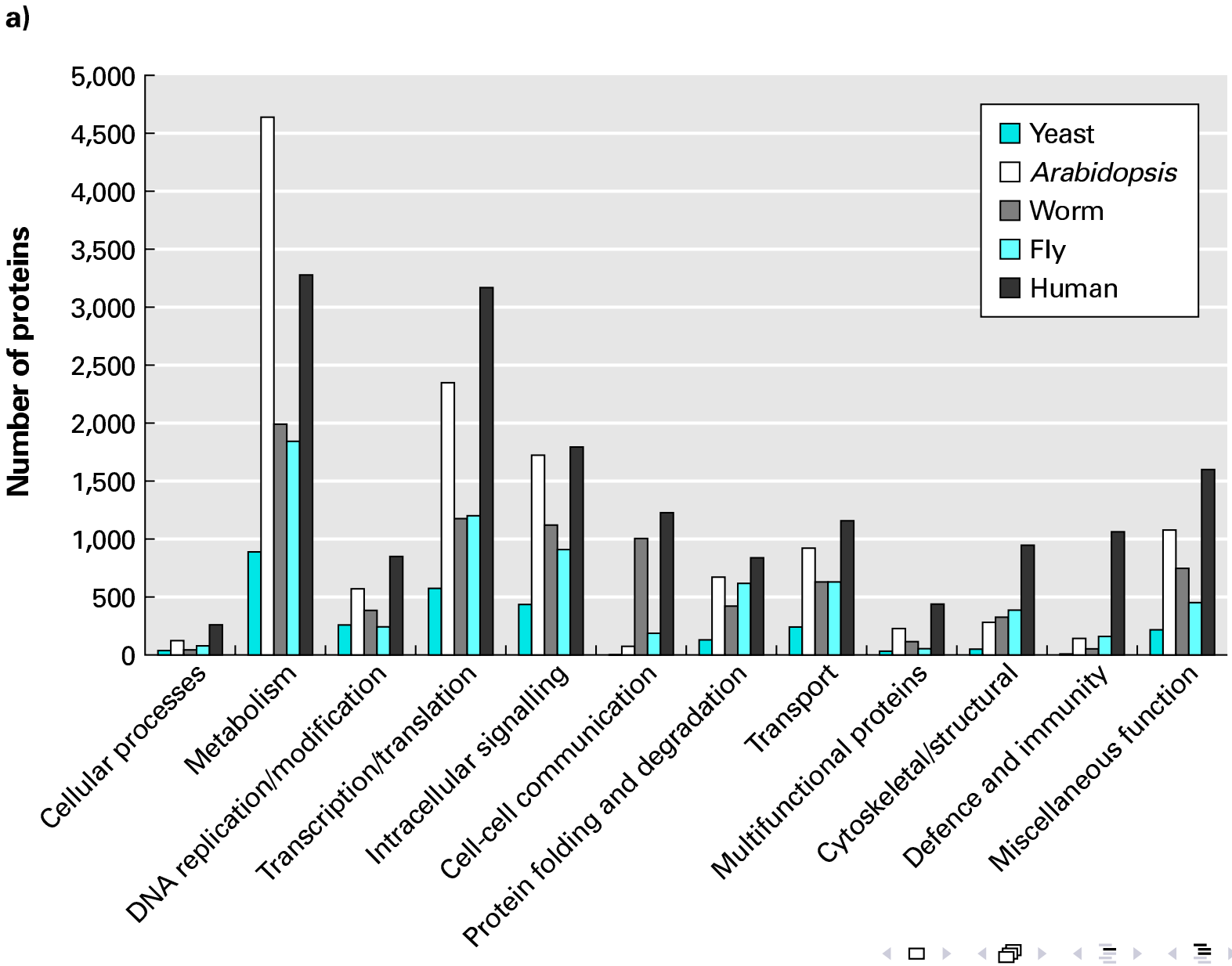
HGP podsumowanie, c.d.

- sklasyfikowano 5 klas powtarzających się sekwencji, np. VRNT (*variable number of tandem repeats*).
- 4 główne kategorie elementów przestawialnych (45% genomu): mogą posłużyć do kalibracji **zegara molekularnego**.
- chromosom X ma największe zagęszczenie przestawialnych elementów (*transpozonów*).
- chromosom Y ma najdłuższy przestawiony fragment czyli jest najmłodszy.
- 50 ludzkich genów pochodzi z transpozonów.

typowy (średni) ludzki gen

- ma 7 krótkich (50 bp) eksonów, przedzielonych długimi (3.3 kpb) intronami,
- 5′ UTR (*untranslated region*) ma długość 300 bp,
- 3′ UTR ma długość 770 bp,
- region kodujący 1340 bp koduje 447 aminokwasy; całość rozrzucona na długości 27000 bp.
- gdzie ukryta nasza złożoność (odp: w proteomie, ok 60% genów produkuje więcej niz jedno mRNA - średnio 2.6).
- bardziej wyrafinowana *ekspresja genów*, czyli więcej *czynników transkrypcyjnych*.

funkcje białek



Quo vadis ?

- genomika funkcjonalna:
  - zrozumieć funkcje genów/ białek
  - znamy funkcje dla ok. 5000 ludzkich białek.
- Zrozumieć regulację ekspresji genów (gąsiennica i motyl).
  - badania transkryptomu (mikromacierze).
  - badania proteomu (spektrometria).
- biologia systemów (bada jak białka oddziaływują ze sobą, nawet bez wiedzy dot. funkcji poszczególnych składowych).

genomika funkcjonalna

- podejście redukcjonistyczne: w jednym eksperymencie weryfikujemy sformułowaną wcześniej hipotezę dotyczącą funkcji jednego genu (nieadekwatne przy badaniu złożonych organizmów).
- podejście holistyczne: analizujemy dostępne dane (najczęściej dostarczane przez biotechnologie o dużej przepustowości jak mikromacierze czy spektrometria) i generujemy hipotezy dot. współregulacji genów czy zaangażowania w pewien biologiczny proces.