Analýza genomických a proteomických dát

Ivana Ihnatová, Barbora Hanáková

Jaro 2016

Osnova prednášok

I. Současné výzvy a technologie genomiky a proteomiky (přednáška 01)

II. Princip a analýza obrazu DNA mikročipů (přednáška 02)

III. Úprava a normalizace dat cDNA mikročipů (přednáška 03)

IV. Úprava a normalizace dat oligonukleotidových mikročipů (přednáška 04)

V. Společné principy analýzy genomických a proteomických dat (přednáška 05)

VI. Porovnávání tříd (přednáška 06)

VII. Predikce tříd (přednáška 07)

VIII. Objevování tříd (přednáška 08)

IX. Analýza přežití a další regrese (přednáška 09)

X. Analýza genových sad a genových sítí (přednáška 10)

XI. Analýza hmotnostní spektrometrie (přednáška 11)

XII. Analýza arrayCGH mikročipů (přednáška 12)

XIII. Meta-analýza (přednáška 13)

Požiadavky

- úlohy v priebehu semestra (5 bodov)
- skupinový projekt (15 bodov)
- písomná skúška (20 bodov)

- úspešné absolvovanie: 21 bodov z toho 10 z projektu

Kapitola I.

Současné výzvy genomiky a proteomiky

Význam štúdia genomiky a proteomiky

- V biológii sme sa znalosťami dostali na najmenšie jednotky, ktoré majú komplexný biologický význam

- GÉNY a PROTEÍNY, ďalej sú už len nukleotidy a aminokyseliny a ešte nižšie už len menšie molekuly a atómy a ... subatomárne častice.

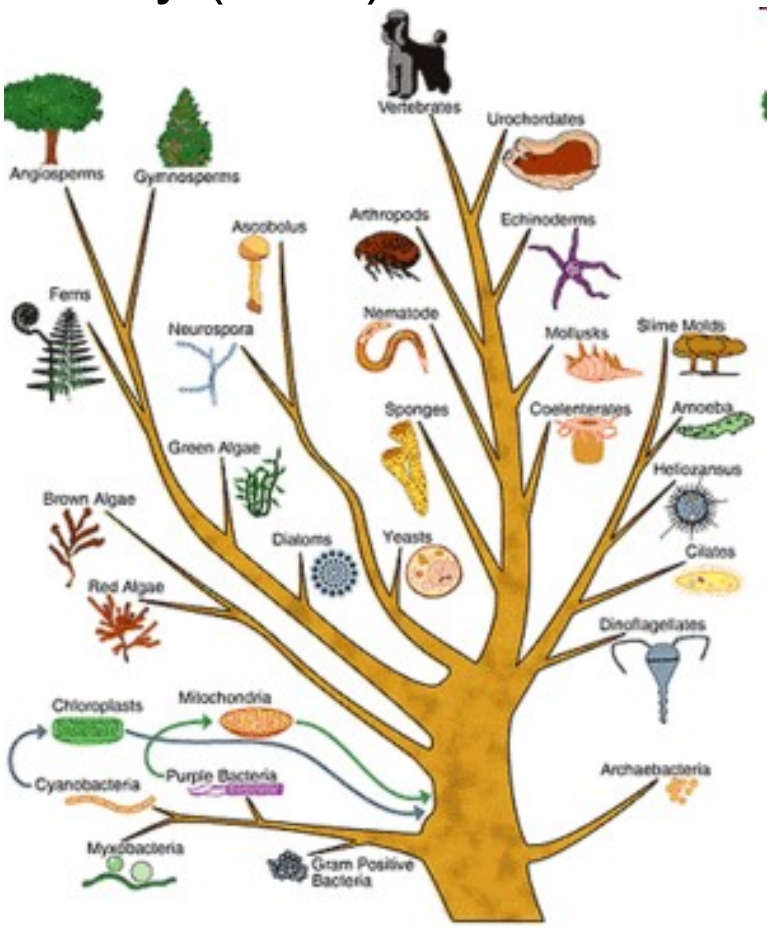
- Študujeme **zloženie** molekúl, ale hlavne ich **funkcie** v organizme

Genomika je veda zaoberajúca sa štúdiom súboru génov v bunke (genóm)

Proteomika je veda zaoberajúca sa štúdiom súboru proteínov v bunke (proteóm)

Gény

- Gény podmieňujú fyzický vzhľad organizmu a jeho schopnosť adaptácie na prostredie v ktorom žije a jeho pomalé i náhle zmeny (stres).



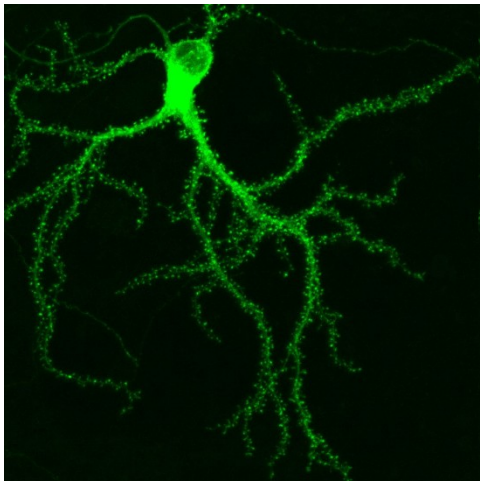
Adaptácia na prostredie

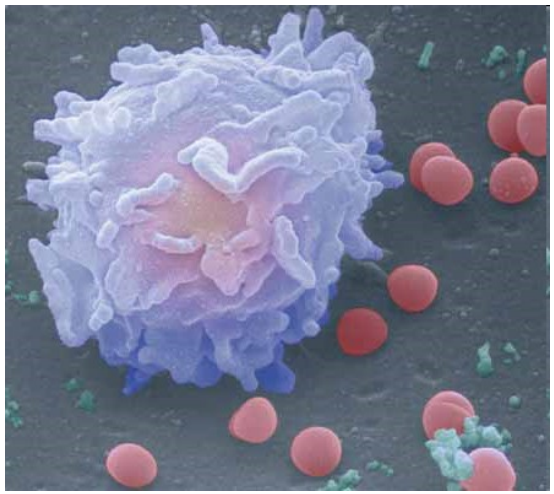
- Odolnosť baktérií na antibiotiká podmienená mutáciami.
- Adaptácia na extrémne podmienky - život vo vesmíre, v sopke, sírnych prameňoch, vriacich prameňoch a mrazoch do -70

Rozdiely medzi organizmami sú podmienené rozdielmi v genóme (kompletná sada génov obsiahnutá v každej bunke organizmu)

Gény II.

Ako to potom, že sa líšia aj bunky v rámci jedného organizmu navzájom, keď majú tú istú sadu génov?





Tieto rozdiely sú dôsledkom odlišnej aktivity génov a ich produktov, proteínov a funkčných RNA molekúl

Genomika a proteomika v BIOLÓGII

Dekódovanie genómu u rôznych druhov

⇩

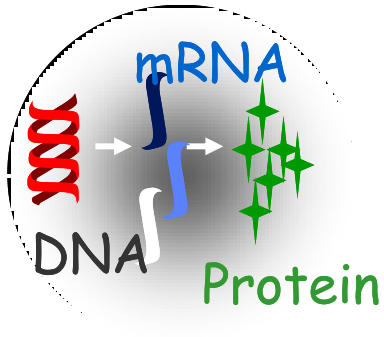
Môžeme študovať

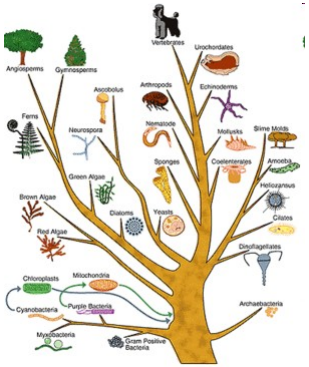
⇩

rozdiely v genóme/proteóme jednotlivých druhov

⇩

študovať tak evolučné prepojenia a vytvárať fylogenetické stromy





⇩

aktivitu génov a proteínov organizmov v rozličných podmienkach

⇩

Môžeme pochopiť správanie parazitov aby sme odhalili mechanizmy ich prispôsobenia hostiteľovi, prípadne študovať baktérie a ich mechanizmy prispôsobenia extrémnym podmienkam ...

Genomika a proteomika v MEDICÍNE

Štúdium genetickej podstaty dedičných i získaných chorôb

⇩

Môžeme študovať

⇩ ⇩

Genetické mutácie, a iné genetické/genomické aberácie spôsobujúce choroby

Rozdielnu aktivitu génov a proteínov u konkrétnych chorôb v porovnaní so zdravým organizmom

⇩

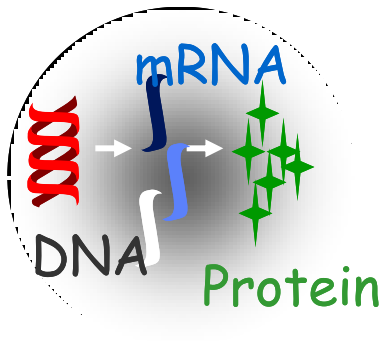
Sme schopní
korelovať funkciu produktov jednotlivých génov s **ochoreniami**
CHOROBA ⇔ GÉN (Y)

⇩

Pochopiť **podstatu** ochorenia

⇩

Nájsť najvhodnejší spôsob liečby (cielená liečba),
prevencie a diagnostiky ochorení



Gény a ochorenia I. - príčiny

- Downov syndróm, hemofília, cystická fibróza, svalové dystrofie, rakovina...
- Dedičné i získané, u niektorých stačí jediná *mutácia* v patričnom géne a vzniká choroba, u iných potrebné viaceré genetické zmeny:

1. Zmeny v štruktúre DNA:
   - Mutácie v štruktúre jedného génu (jednonukleotidové polymorfizmy, delecie, inzercie, amplifikácie nukleotidov)
   - Aberácie celého génu, alebo časti chromozómu (delecie, translokácie, inzercie, amplifikácie)
   - Aberácie celých chromozómov
2. Zmeny v expresii a aktivite génov a ich produktov
3. Zmeny v posttranslačných úpravách proteínov

Gény a ochorenia II. - mutácie

- Bunky v organizme sa stále obnovujú a delia, replikujúc zakaždým celý genóm na nukleotid presne. To nie je pri veľkosti ľudského genómu 3.2 biliónov nukleotidov jednoduché.

- Preto existuje mnoho kontrolných mechanizmov:
  - na opravu poškodenej časti DNA
  - pre správnu distribúciu chromozómov v procese mitózy/meiózy
  - pre prípadnú apoptózu (regulovanú smrť bunky) v prípade nezvratných zmien
  - a pod…

- Genetické aberácie vznikajú zlyhaním kontrolných mechanizmov

Gény a ochorenia III. – aktivita génov

- Nielen mutácie, ale aj *nesprávna aktivita génov* môže viesť k vzniku chorôb.
- V ľudskej bunke každú chvíľu prebieha obrovské množstvo procesov, prepisujú sa stovky génov a neustále sa vytvárajú proteíny, na základe vnútorných a vonkajších podnetov.
- Tieto podnety sú regulované *stovkami regulačných mechanizmov*, zas založených na proteínoch.
- Chyba v jednom z mechanizmov môže takisto skončiť vyvinutím choroby.

Gény a ochorenia IV. - zhrnutie

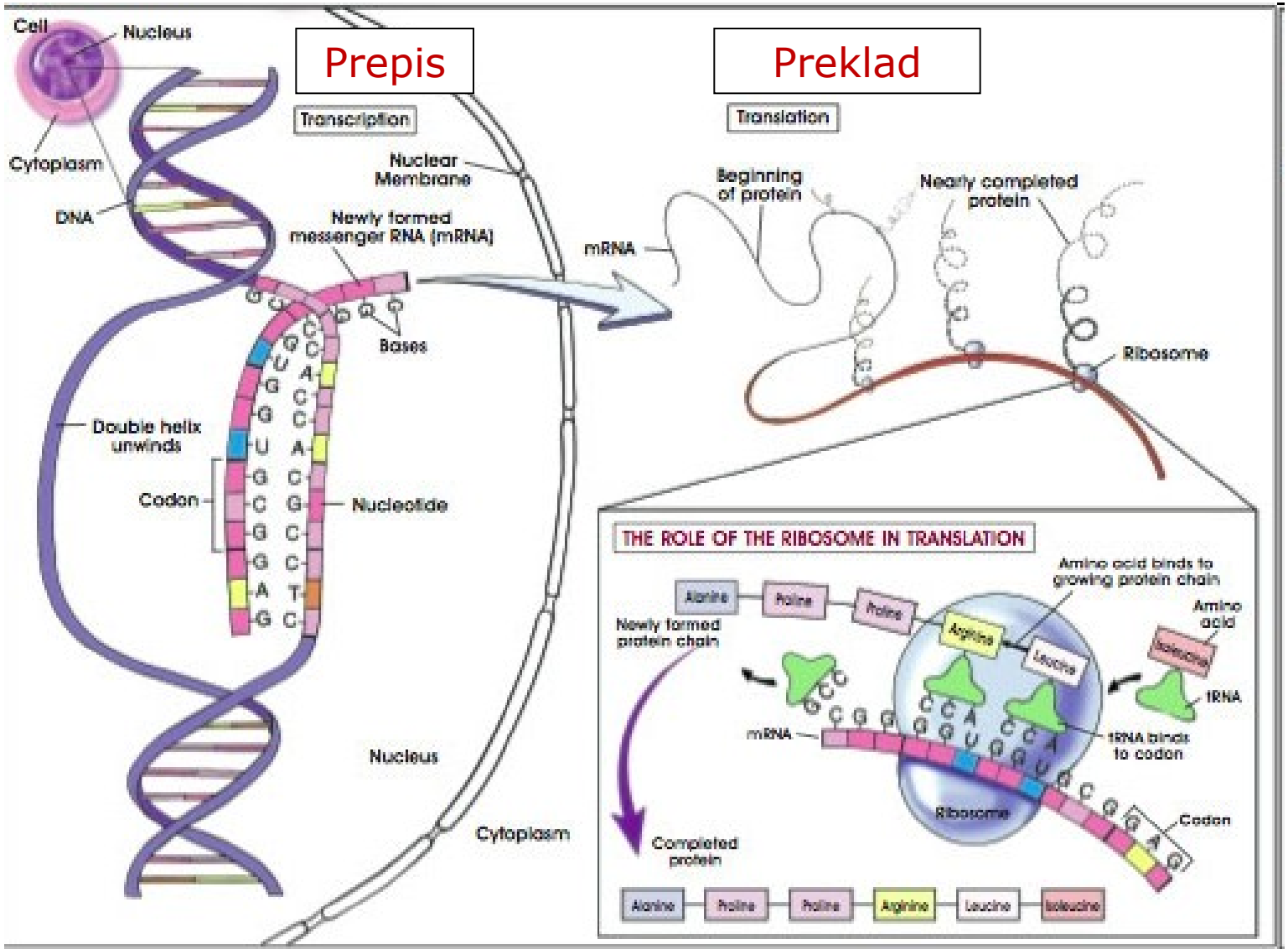
- Čo spôsobuje ochorenia – **proteín(-y)** a **iné funkčné molekuly**, ktoré majú zmenenú svoju funkčnosť!

- Príčiny nesprávnej funkcie:
  - mutácia v príslušnom géne, spôsobujúca v dôsledku zmenu v sekvencii aminokyselín proteínu a tým jeho:
    - nefunkčnosť
    - nadmernú aktivitu
  - zmeny v mechanizmoch kontroly expresie daného proteínu, ktorý je následne produkovaný
    - v nedostačujúcom množstve
    - v nadmernom množstve
  - zmeny v postranslačných úpravách a sekundárnej/terciárnej štruktúre proteínu

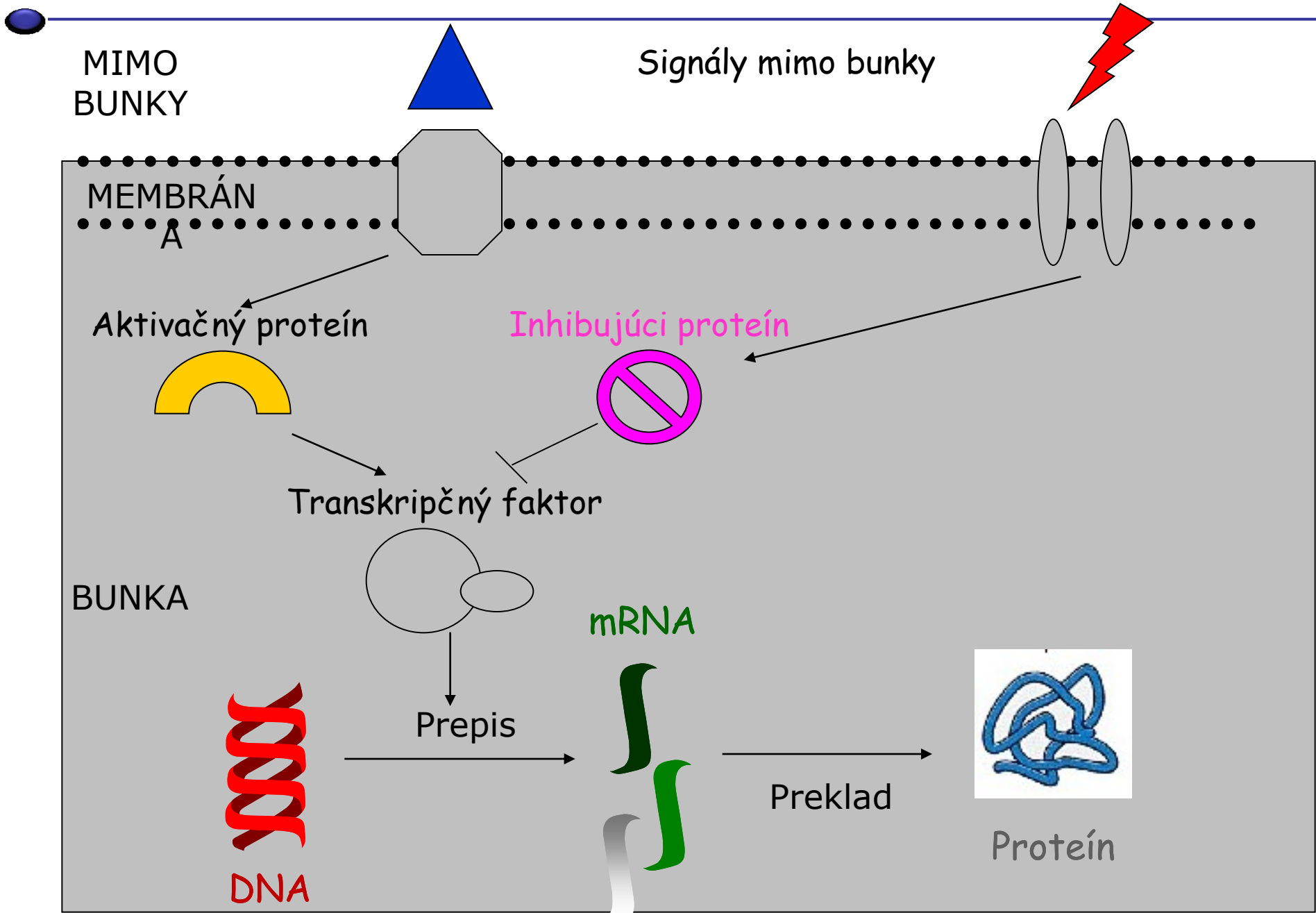
Ústredná dogma molekulárnej biológie

Prepis Preklad

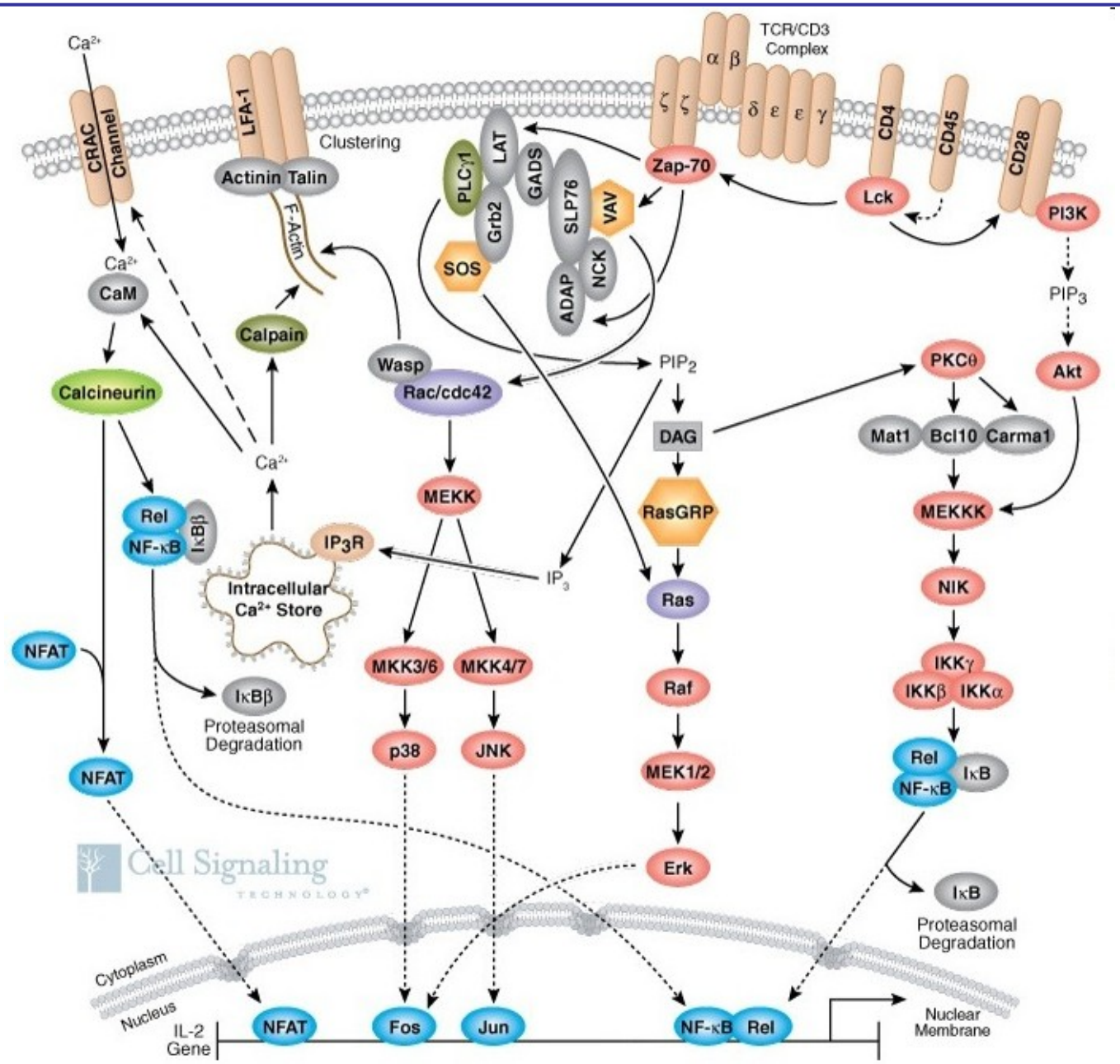
DNA -> mRNA -> proteín



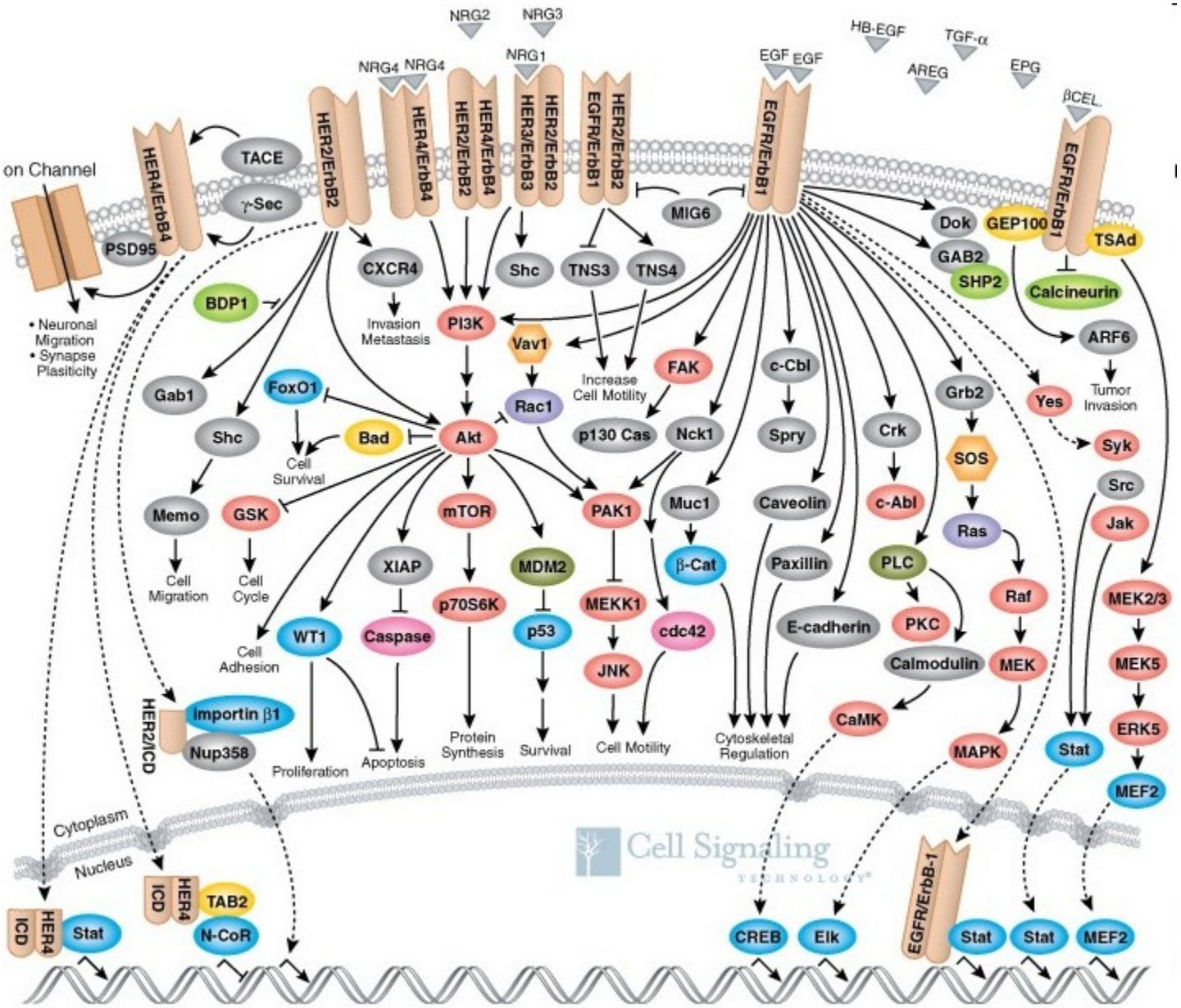
Čo ešte vieme



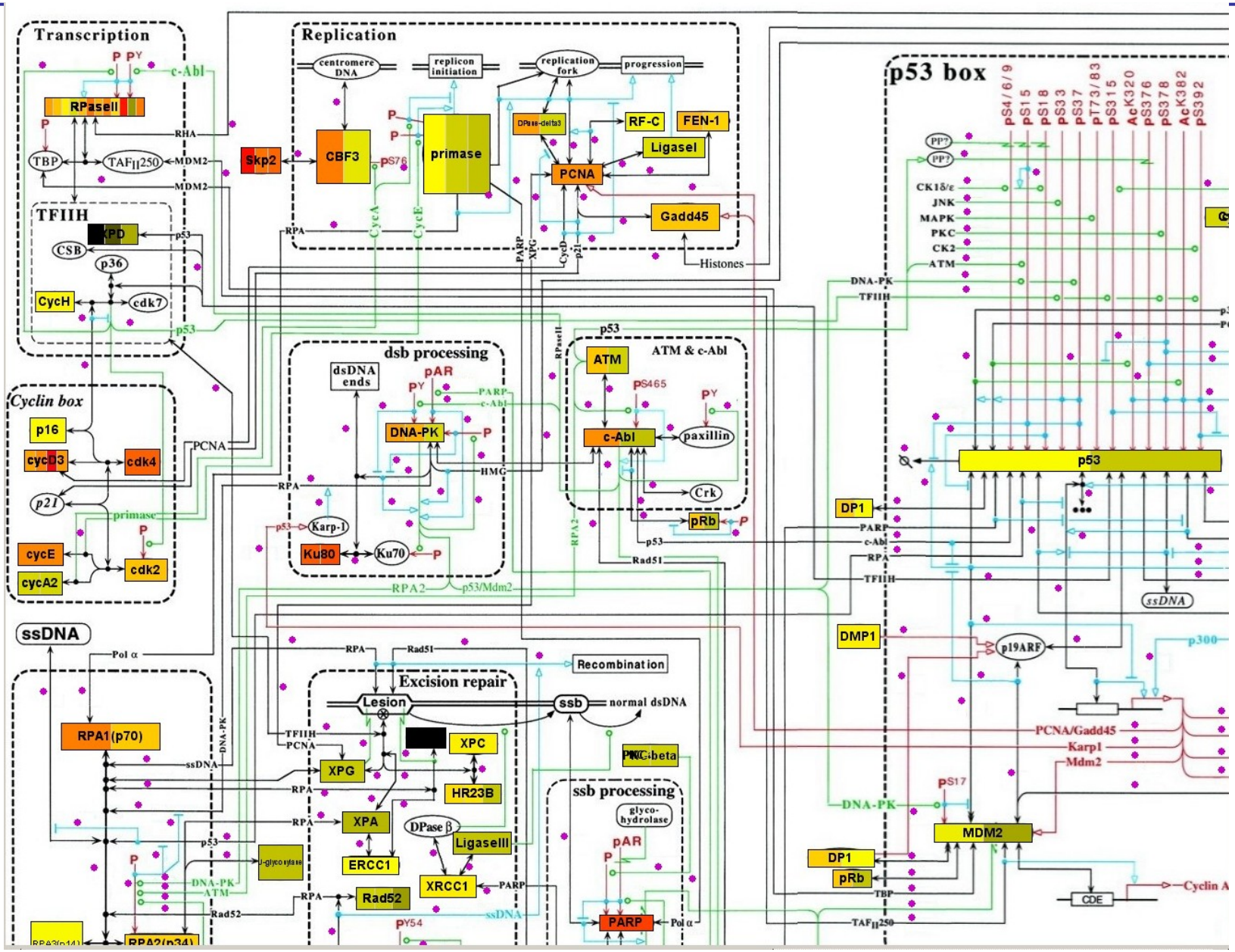
Ale vieme ešte viac...



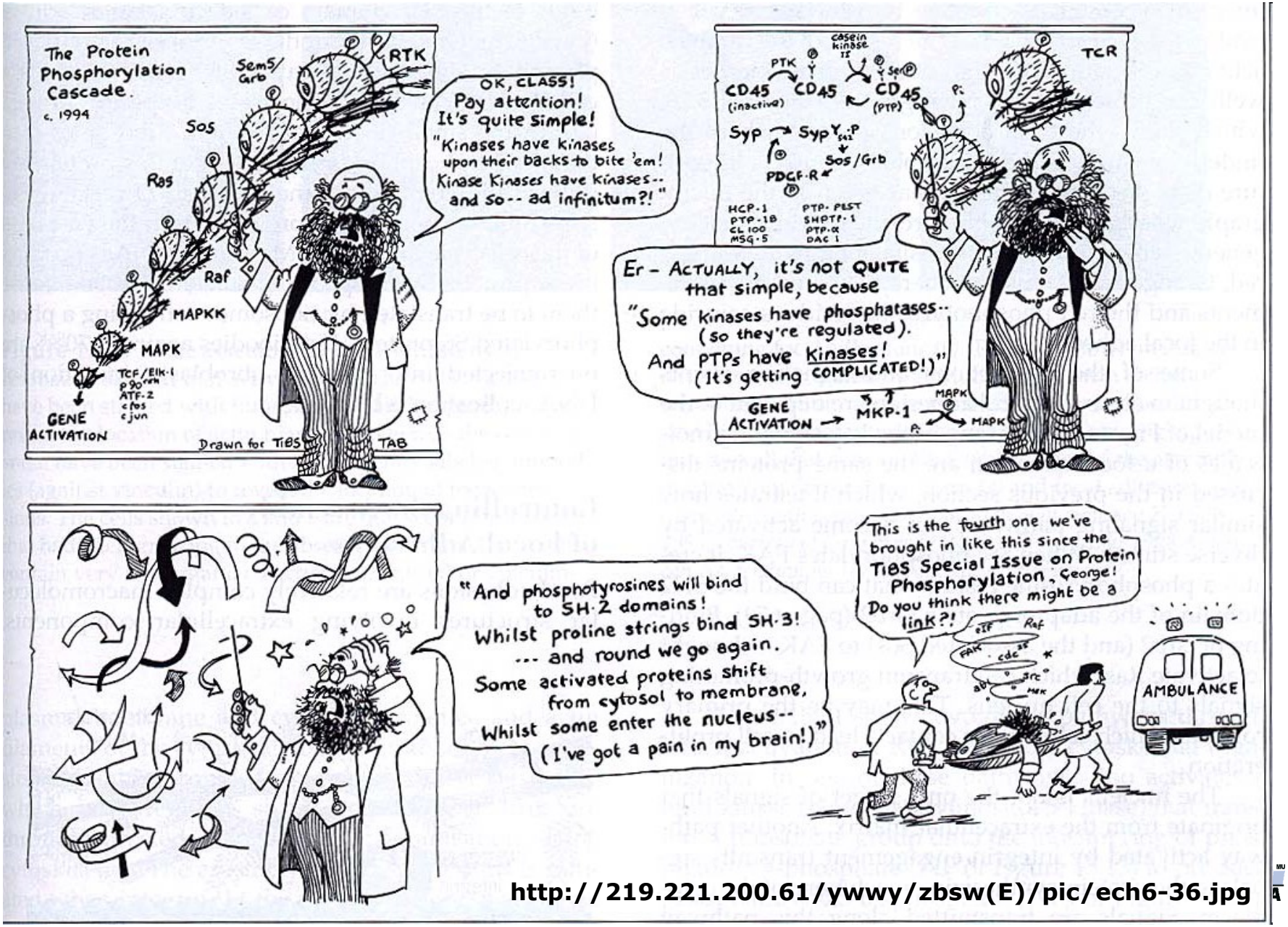
..a ešte viac...


NRG2
NRG3
NRG1
NRG4
NRG4
HB-EGF
TGF-α
AREG
EPG
EGF
EGF
βCEL.
HER2/ErbB2
HER4/ErbB4
HER2/ErbB2
HER4/ErbB4
HER3/ErbB3
HER2/ErbB2
EGFR/ErbB1
HER2/ErbB2
EGFR/ErbB1
EGFR/ErbB1
on Channel
HER4/ErbB4
TACE
γ-Sec
PSD95
MIG6
Dok
GEP100
GAB2
SHP2
TSAd
Calcineurin
CXCR4
Shc
TNS3
TNS4
BDP1
PI3K
Vav1
ARF6
• Neuronal Migration
• Synapse Plasticity
Invasion Metastasis
Increase Cell Motility
FAK
c-Cbl
Grb2
Yes
Tumor Invasion
Gab1
FoxO1
Rac1
Shc
Bad
Akt
p130 Cas
Nck1
Spry
Crk
SOS
Syk
Cell Survival
Src
Memo
GSK
mTOR
PAK1
Muc1
Caveolin
c-Abl
Ras
Jak
Cell Migration
Cell Cycle
XIAP
MDM2
β-Cat
Paxillin
PLC
Raf
MEK2/3
WT1
Caspase
p70S6K
p53
MEKK1
cdc42
E-cadherin
PKC
MEK
MEK5
Cell Adhesion
JNK
Calmodulin
ERK5
HER2/ICD
Importin β1
Nup358
CaMK
MAPK
Stat
MEF2
Proliferation
Apoptosis
Protein Synthesis
Survival
Cell Motility
Cytoskeletal Regulation
Cytoplasm
Nucleus
Cell Signaling
TECHNOLOGY®
ICD
HER4
Stat
ICD
HER4
TAB2
N-CoR
CREB
Elk
EGFR/ErbB-1
Stat
Stat
MEF2

...a ešte viac...


Transcription
Replication
p53 box
Cyclin box
dsb processing
ATM & c-Abl
ssDNA
Excision repair
ssb processing
Recombination
RPAseII
TFIIH
TBP
TAFII250
MDM2
Skp2
CBF3
primase
PCNA
RF-C
FEN-1
LigaseI
Gadd45
Histones
CSB
p36
CycH
cdk7
p16
cycD3
cdk4
p21
cycE
cycA2
cdk2
dsDNA ends
DNA-PK
Karp-1
Ku80
Ku70
ATM
c-Abl
paxillin
Crk
pRb
p53
DP1
p19ARF
DMP1
p300
MDM2
CDE
Cyclin A
RPA1(p70)
RPA2(p34)
XPC
XPG
HR23B
XPA
ERCC1
Rad52
XRCC1
LigaseIII
DPase β
glyco-hydrolase
PARP
Pol α
Lesion
ssb
normal dsDNA
PCNA/Gadd45
Karp1
Mdm2

...ale je veľmi obtiažne to všetko prepojiť a interpretovať



http://219.221.200.61/ywwy/zbsw(E)/pic/ech6-36.jpg

Čo skúmame v genomike a proteomike

- U **génov** môžeme skúmať ich
  - **Štruktúru a zmeny v nej** – sekvencia nukleotidov A, C, G, T
  - **Množstvo** – či sú alebo nie sú prítomné a v akom počte kópií
  - **Aktivitu** – či sa gén prepisuje do mRNA a v akom množstve

- U **proteínov** skúmame
  - **Zloženie** – z akých aminokyselín
  - **Štruktúru** – ako sú reťazce peptidov usporiadané do 3D štruktúr?
  - **Množstvo** – či sú alebo nie sú prítomné a v akom množstve
  - **Funkciu** – modelovanie, identifikácia aktívnych väzobných miest

- Ďalšia fáza je **modelovanie komplexných bunkových systémov** – proteínové interakcie, bunkové dráhy, regulačné a metabolické siete...

Metódy štúdia genómu a proteómu

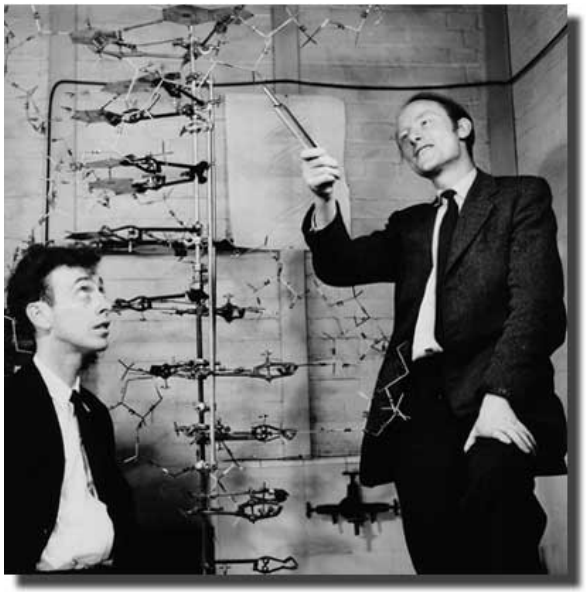
- *Klasické metódy* molekulárnej biológie a cytogenetiky:
  - metódy skúmajúce len jeden alebo niekoľko génov a proteínov v jednom experimente:
    - PCR, RT-PCR, real-time PCR
    - FISH (fluorescence in-situ hybridization)
    - gélová elektroforéza, ...

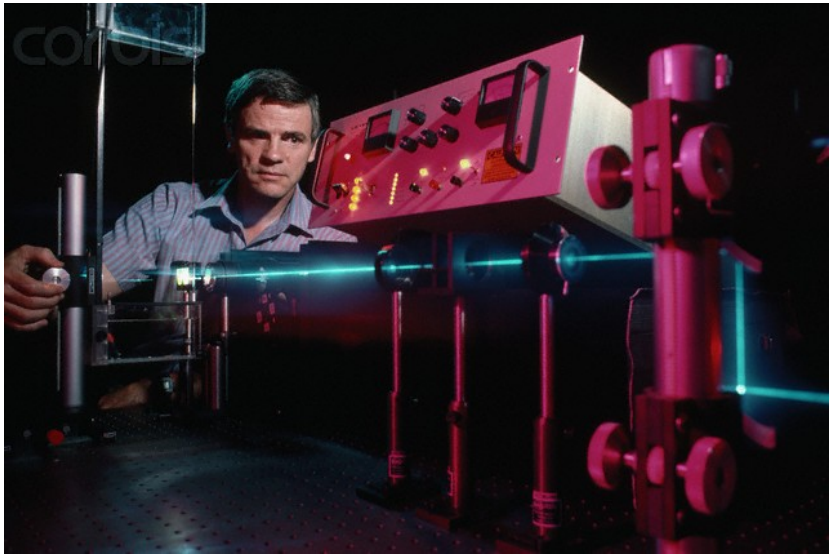
- *Vysokopokryvné metódy* molekulárnej biológie:
  - schopné skúmať tisíce molekúl v jednom experimente....

... ako vznikli?

Od Watsona & Cricka po Leroya Hooda

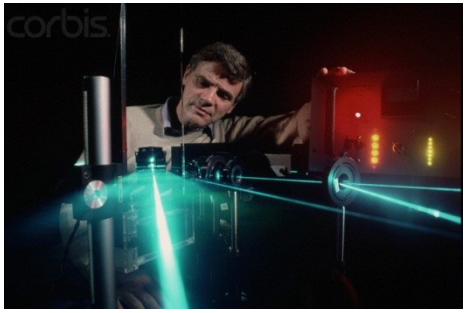
- Na začiatku bol dvojšrobovicový model DNA...





- a na konci boli:
  - automatické **sekvenátory** DNA a proteínov
  - automatické **syntetizátory** DNA a proteínov

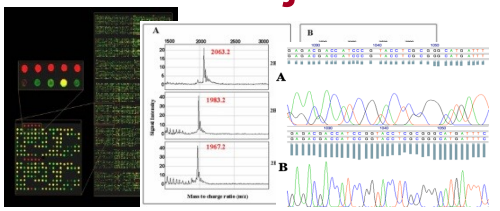
Nové možnosti



Sekvenátory umožnili rýchlo dekódovať sekvenciu génov a proteínov

Znalosť presnej sekvencie umožnila navrhovať *špecifické génové sondy* a syntetizátor umožňoval ich rýchlu a *automatickú výrobu*.

Otvorili sa dvere pre nové, vysokopokryvné technológie, schopné analyzovať tisíce génov/proteínov v jednom experimente!



Zhrnutie prvej časti

- Veľká časť rozmanitosti života vrátane ochorení sa zrejme dá obsiahnúť štúdiom **funkcie genómu a proteómu**
- V súčasnosti existujú špeciálne **vysokopokryvné metódy (high-density methods)**, ktoré umožňujú skúmať tisíce génov a proteínov v jednej vzorke a jednom experimente
- Tieto metódy produkujú **obrovské množstvá dát** a vyžadujú špecializovanú štatistickú analýzu

Kapitola II.

Technológie študujúce genomiku a proteomiku

Vysokopokryvné metódy I.

- Analýza genómu (od nukleotidových sekvencií po úplne anotovaný genóm)
  - Analýza štruktúry
  - Analýza expresie
  - Porovnávacia genomika
  - Regulácia genómu

- Analýza proteómu (od hmostnostních spektier – cez komplexné štruktúry proteinových zhlukov - po analýzu funkcie proteínov)
  - Analýza štruktúry
  - Analýza expresie
  - Analýza funkcie

- Modelovanie komplexných systémov – proteínové interakcie, bunkové dráhy, regulačné a metabolické siete...

Vysokopokryvné metódy II.

- Analýza genómu (od nukleotidových sekvencií po úplne anotovaný genóm)
  - Analýza štruktúry – DNA sekvenácia, Chip-seq. WES (whole exome sequencing)
  - Analýza expresie – Mikročipy, SAGE, MPSS, Expressed sequence tags (ESTs), RNA-seq, ...
  - Porovnávacia genomika – aCGH čipy, SNP polymorfizmy, alternative splicing arrays, fingerprinting
  - Regulácia genómu – CHip-on-chip

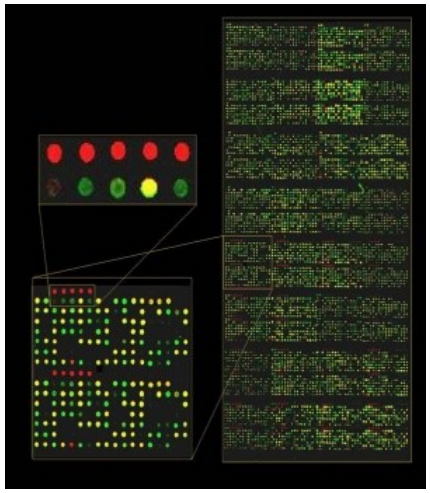
- Analýza proteómu (od hmostnostních spektier – cez komplexné štruktúry proteinových zhlukov - po analýzu funkcie proteínov)
  - Analýza štruktúry: Proteínová sekvenácia
  - Analýza expresie: Hmotnostná spektrometria, Proteínové microarraye,
  - Analýza funkcie: Modelovánie makromolekulárnych systémov – odvodzovanie vlastností z atómových interakcií

- Modelovanie komplexných systémov – proteínové interakcie, bunkové dráhy, regulačné a metabolické siete...

Dáta vysokopokryvných metód I.

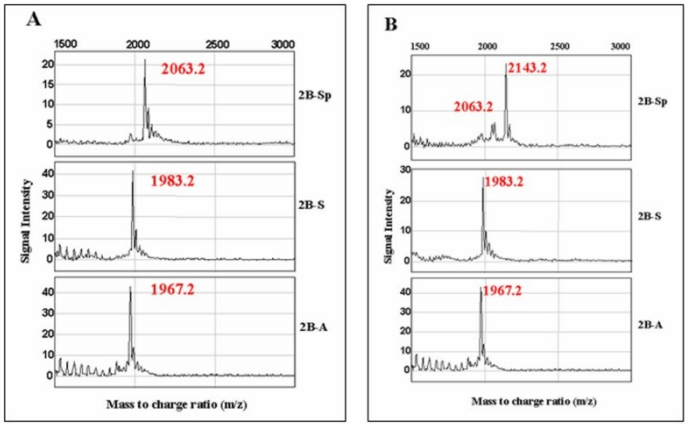
- Moderné vysokopokryvné technológie produkujú obrovské tabuľky komplexných dát

Mikročipy

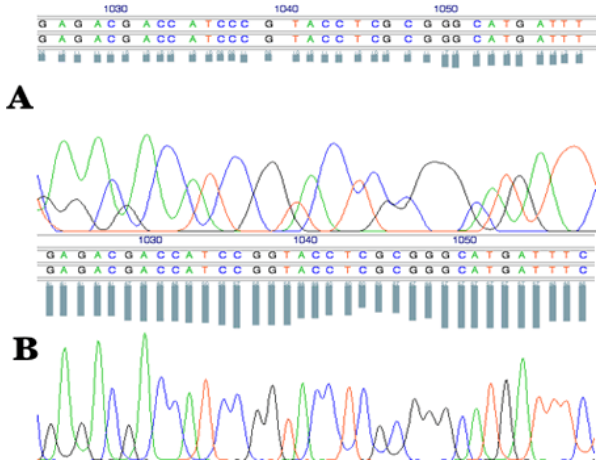


☑ **Expresia 10 000 – 100 000 transkriptov u 100 – 1000 vzoriek**

MASS – hmotnostná spektrometria



Sekvenácia DNA



☑ **Tisíce spektier proteínov – GB datové súbory**

☑ **Genóm s biliónmi nukleotidov**

Dáta vysokopokryvných metód II.

Dátový súbor z vysokopokryvných experimentov – pohľad biológa

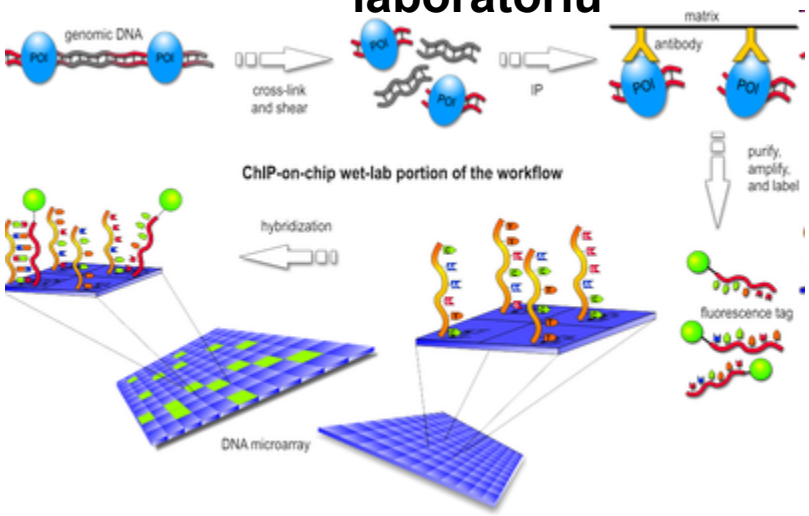
"In principle, the string of genetic bits holds long-sought secrets of human development, physiology and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate".

The Genome International Sequencing Consortium, "Initial sequencing and analysis of the human genome," *Nature* 409**:** 860-921 (2001)

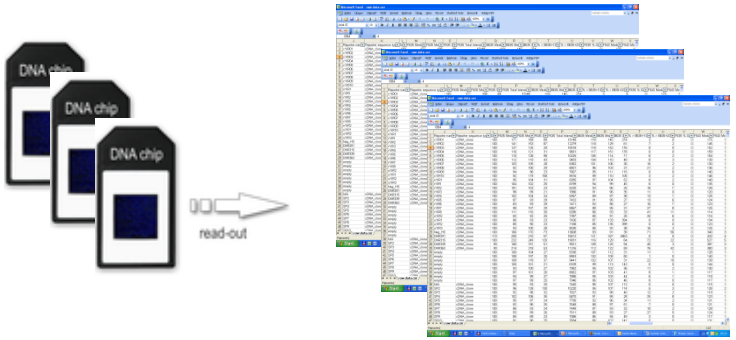
Hľadanie ihiel v kopách sena?

Dátový súbor z vysokopokryvných experimentov– pohľad matematického biológa

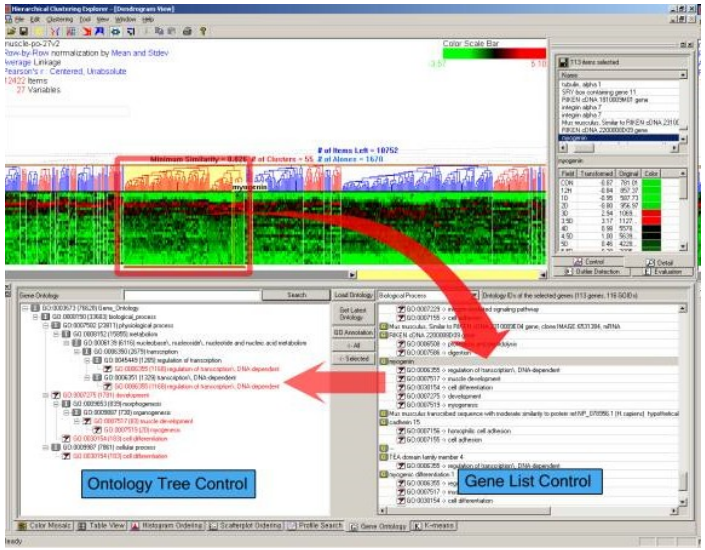
1. Príprava a vykonanie experimentu v laboratóriu



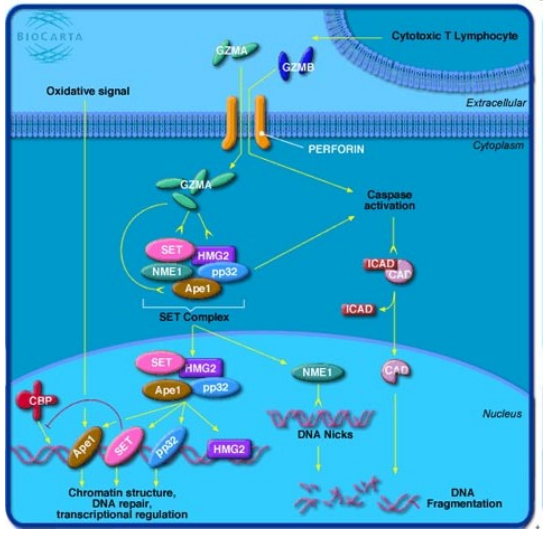
2. Extrakcia a úprava dát



3. Štatistická analýza dát



4. Biologická a klinická interpretácia



Dôležité!

- Veľká časť rozmanitosti života vrátane ochorení sa zrejme dá obsiahnúť štúdiom funkcie genómu a proteómu a ich vzťahu
- Biológia a medicína sa v súčasnosti nezaobíde bez štúdia genetiky a proteomiky
- V súčasnosti existujú špeciálne vysokopokryvné metódy, ktoré umožňujú skúmať tisíce génov a proteínov v jednej vzorke a jednom experimente
- Biológovia a lekári produkujú v súčasnosti obrovské množstvá genomických a proteomických dát, ktoré vyžadujú špeciálne metódy analýzy
- Biológovia a lekári sú špecialisti vo svojom obore ale táto práca im zaberá všetok čas. Obvykle nemajú čas študovať štatistiku a analyzovať svoje dáta
- Databázy sú plné genomických a proteomických dátových súborov, ale je relatívne málo odborníkov, čo ich analyzujú

Vysokopokryvné metódy – čo si priblížime

Podrobnejšie si predstavíme technológie:

- Mikročipy:
  - Expresné: cDNA, Affymetrix, Illumina
  - aCGH čipy
- Hmotnostná spektrometria
- Analýza RNA-seq – v samostatnom kurze

Vysokopokryvné metódy – čo si priblížime

Podrobnejšie si predstavíme technológie:

- Micročipy:
  - Expresné: cDNA a Affymetrix
  - aCGH čipy

- Hmotnostná spektrometria

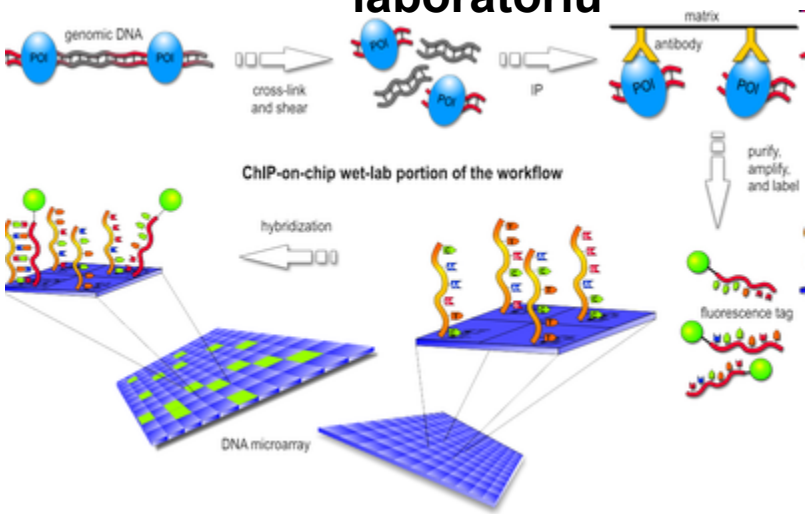
Kapitola II.1.

Technológie študujúce genomiku a proteomiku

Mikročipy (microarrays)

Dátový súbor z high-density experimentov– pohľad matematického biológa

1. **Príprava a vykonanie experimentu v laboratóriu**

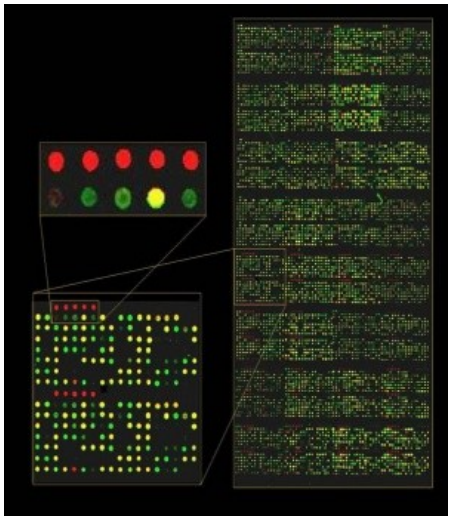


Technika mikročipov

- **Mikročipy**– biotechnológia simultánne porovnávajúca biologické objekty (molekuly, tkanivá) na základě ich imobilizácie na jediný **podklad** do oblastí (spotov) ktoré jsou pravidelne usporiadané do riadkov a stĺpcov.

- **Podklad**: sklo, gel, parafín, ...

- Mikročipy pre genóm alebo proteóm:
  - DNA mikročipy
  - Proteínové mikročipy

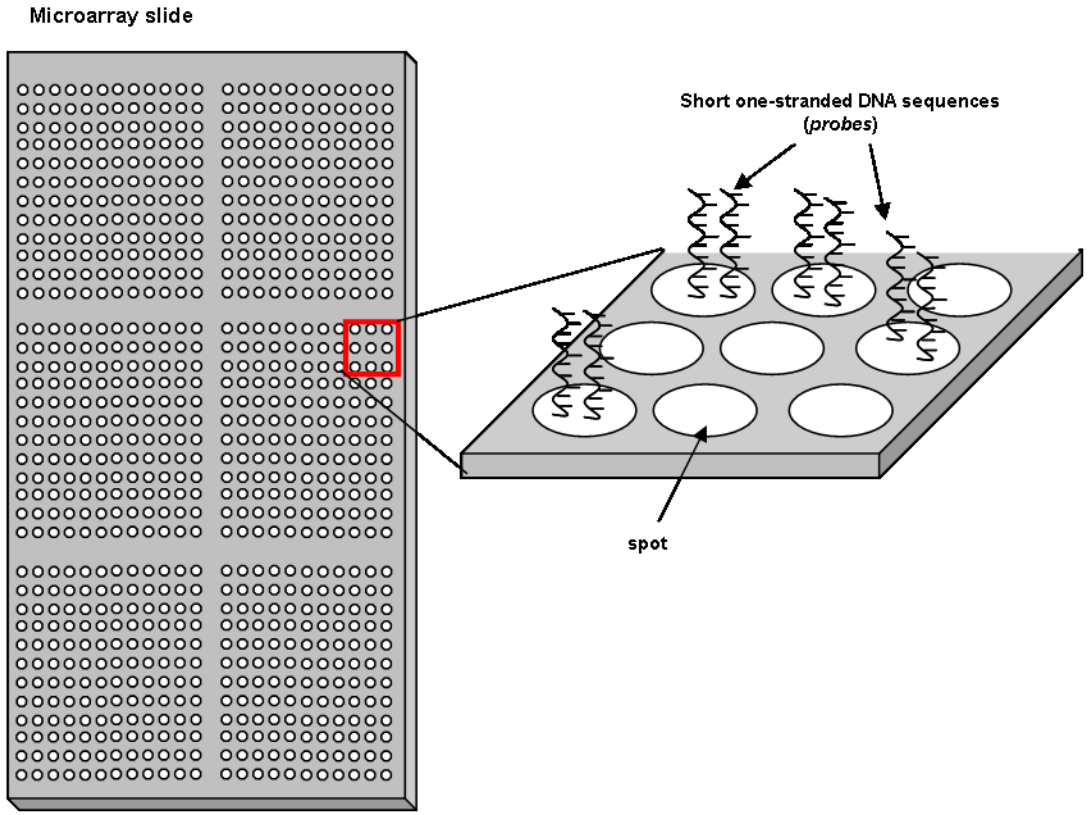


DNA mikročipy

- Séria **krátkych DNA sekvencií** imobilizovaných rovnomerne na podklad, používaná na detekciu DNA alebo RNA (obvykle vo forme cDNA) vo vzorkách. Najčastejšie aplikovaná na:
  - meranie zmien v hladinách génovej expresie (gene expression profiling, detekcia RNA - cDNA) - **expresné arraye**
  - detekciu štruktúrnych zmien genómu (SNPs- jednonukleotidové polymorfizmy alebo zmeny v počte kópií génov) – **arrayCGH, SNP arrays**
- Taktiež sa úspešne používa na detekciu väzbových miest proteínov na genóme (**ChIP-on-chip**), detekciu alternatívneho zostrihu (**exon junction arrays**) a takisto na presnú detekciu neznámych a nepredikovaných transkriptov alebo alternatívnych foriem zostrihu (**tiling arrays**)

Sonda (probe)

- **Krátke DNA sekvencie (oligonukleotidy)** na microarray sklíčku sa nazývajú ***sondy***, anglicky *probes*
- Každá oblasť DNA (obvykle gén), ktorú chceme skúmať
- Sondy sú navrhnuté tak, aby boli pre daný gén/oblasť čo najšpecifickejšie



Základný princíp

- Fragmenty DNA/cDNA zo vzorky sa **spárujú** s **komplementárnymi** sondami na microarray sklíčku a tým sa **imobilizujú**.

- Imobilizované molekuly DNA, ktoré boli predtým označené **fluorescenčným farbivom** sa potom dajú detekovať pomocou **UV skenera** a kvantifikovať tak množstvo mRNA/DNA s danou sekvenciou prítomnej vo vzorke.