Analýza genomických a proteomických dat

Mgr. Eva Budinská, Ph.D.

RNDr. Ivana Ihnatová, Ph.D.

Jaro 2019

Osnova přednášek

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 dat hmotnostní spektrometrie (proteomika) (přednáška 11)

XII. Analýza arrayCGH mikročipů (přednáška 12)

XIII. Meta-analýza (přednáška 13)

Požadavky

- Individuální projekt (**20 bodů** - 50% z celkového hodnocení zkoušky, 83% hodnocení zápočtu)
- Písemná zkouška (**16 bodů**) – 40% z celkového hodnocení zkoušky, 10% hodnocení zápočtu
- Aktivita a přítomnost na přednáškách (**4 body**) – 10% z celkového hodnocení zkoušky,
- Úspěšné absolvování:
  - **Ukončení zkouškou: 21 bodů**, z toho min **10** z projektu a **min 8** ze zkoušky
  - **Ukončení zápočtem**: **14 bodů**

Požadavky

	Projekt	Aktivita, účast	Zkouška	Min počet bodů
Zkouška	20b (50%)	4b (10%)	16b (40%)	21 b, z toho min 10 za projekt a min 8 za zkoušku
Zápočet	20b (87%)	4b (13%)	-	14 b

Projekt

- Zpracovává se samostatně
- Možnost zpracovávat vlastní data nebo data z veřejné dostupných databází
- Výběr ze stanovených projektů, vlastní téma nutno schválit předem – **nejzazší termín výběru projektu**: 6.3.2019
- Projekt nutno odevzdat před zkouškou, pouze po odevzdání a obdržení 10 bodů z projektu je možné přihlásit sa na zkušební termín, kde se projekt pak ústně obhajuje
- Nejzazší termín odevzdání projektů:
  - pro udělení zápočtu: 25.6.2019
  - pro kontrolu počtu bodů: 7 dní před zkušebním termínem
- 1x možnost opravy projektu

Požadavky vypracování projektu

- 2 soubory:
  - popis projektu ve formátu pdf
  - .R soubor se skriptem analýzy od načtení dat po finální grafy
- Struktura popisu projektu:
  - Název
  - Úvod – co je cílem projektu, přesně definované hypotézy
  - Data – přesně definovaný typ dat, odkaz na stažení dat, počet vzorků, typ platformy ze které byly data získány, kolik bylo na platformě sond, kolik genů reprezentovaly, v případě dvoukanálového experimentu jasná definice vzorků v jednotlivých kanálech
  - Metodika – jaké metody zpracování dat od úpravy až po finální interpretaci byly použity a proč
  - Výsledky – výsledková část rozdeleňá na
    - a. Předzpracování a normalizace základních dat (popis, grafy, interpretace výsledků vzhledem k dalším analýzám)
    - b. Statistická analýza a data mining – rozděleno dle typu analýzy, popis nejdůležitějších výsledků a jejich sumarizace, grafy (např. venovy diagramy, heatmapa, volcano plot, forest plot…), sumární tabulky výsledků, odkazy na tabulky s podrobnými výsledky
    - c. Biologická interpretace
    -
- Struktura .R souboru se skriptem analýzy:
  - Skript rozdělený do kapitol podle analýzy dat s podrobným komentářem jednotlivých kroků

Kapitola I.

Současné výzvy genomiky a proteomiky

Význam studia genomiky a proteomiky

- V biologii jsme se znalostmi dostali na nejmenší jednotky, které mají komplexní biologický význam

- GENY a PROTEINY, dále jsou už jen nukleotidy a aminokyseliny a ještě níž jsou jen menší molekuly a atomy a … subatomární částice.

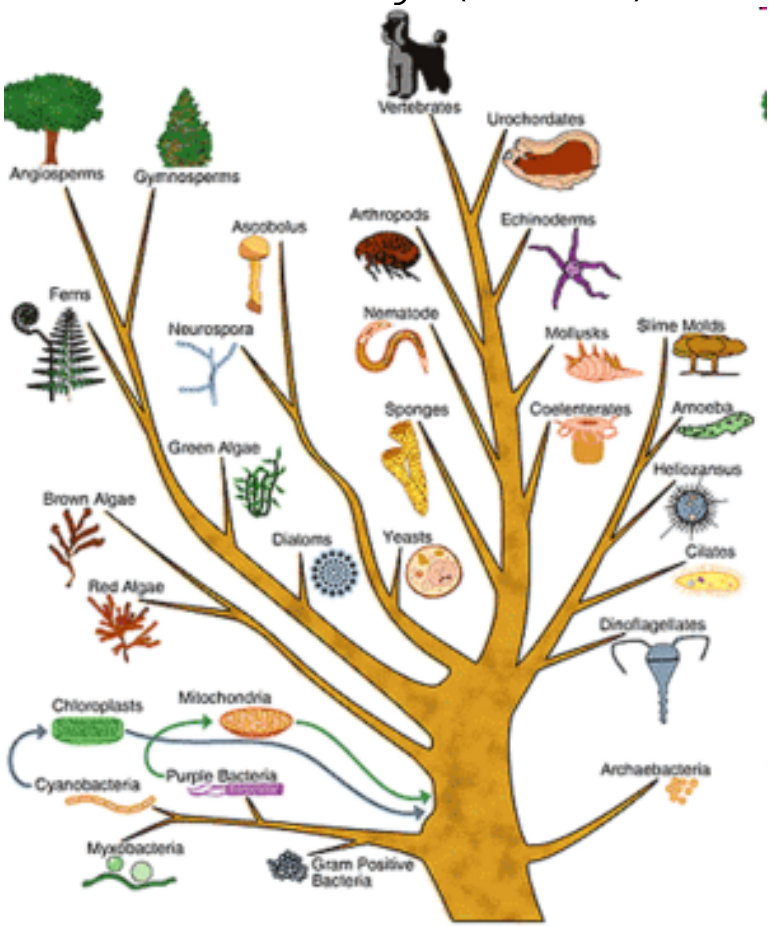
- Studujeme složení molekul a hlavně jejich funkcí v organismu

Genomika je věda zabývající se studiem souboru genů v buňce (genom)

Proteomika je věda zabývající se studiem souboru proteinů v buňce (proteom)

Geny

- Geny podmiňují fyzický vzhled organismu a jeho schopnost adaptace na prostředí, ve kterém žije a jeho pomalé i náhlé změny (stres).



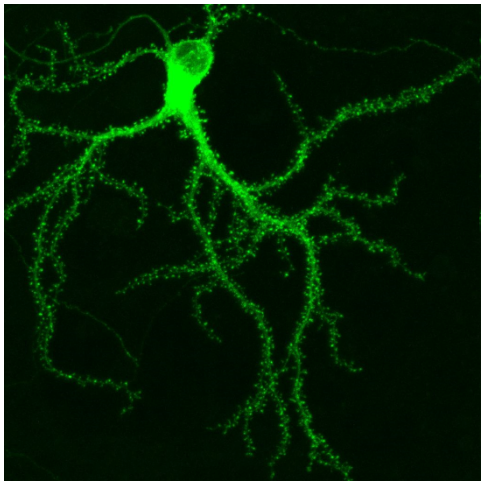
Adaptace na prostředí

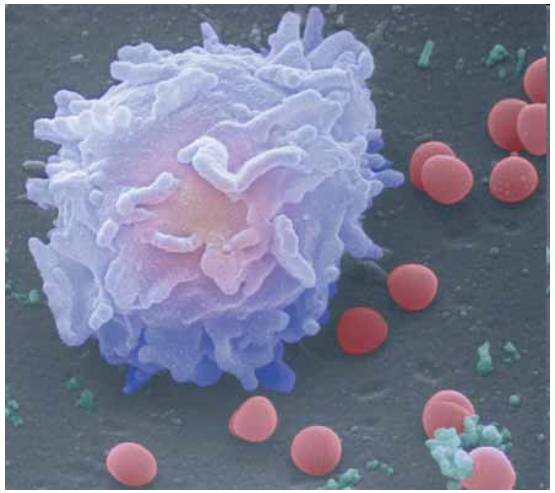
- Odolnost bakterií na antibiotika podmíněná mutacemi.
- Adaptace na extrémní podmínky - život ve vesmíru, v sopce, sirných pramenech, vařících pramenech a mrazech do -70

Rozdíly mezi organismy jsou podmíněné rozdíly v genomu (kompletní sada genů obsažená v každé buňce organismu).

Geny II.

Jak je možné, že se navzájem liší i buňky v rámci jednoho organismu, když mají stejnou sadu genů?





Tyto rozdíly jsou důsledkem odlišné aktivity genů a jejich produktů, proteinů a funkčních RNA molekul.

Genomika a proteomika v BIOLOGII

Dekódování genomu u různých druhů

⇩

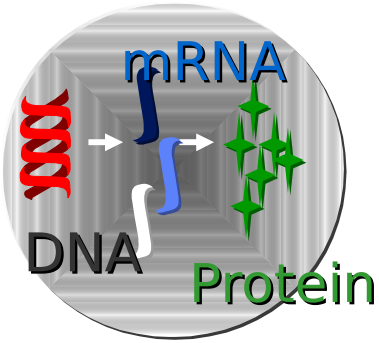
Můžeme studovat

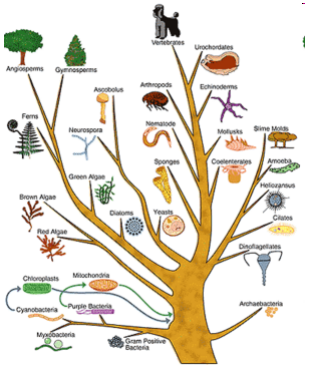
⇩ ⇩

Rozdíly v genomu/proteomu jednotlivých druhů

⇩

studovat tak evoluční propojení a vytvářet fylogenetické stromy





aktivitu genů a proteinů organizmů v rozdílných podmínkách

⇩

Můžeme pochopit správně parazity, abychom odhalili mechanismy jejich přizpůsobení se hostiteli, případně studovat bakterie a jejich mechanismy přizpůsobení se extrémním podmínkám …

Genomika a proteomika v MEDICÍNĚ

Studium genetické podstaty dědičných i získaných onemocnění

⇩

⇩ Můžeme studovat ⇩

Genetické mutace, a jiné genetické/genomické aberace způsobující choroby

Rozdílnou aktivitu genů a proteinů u konkrétních chorob v porovnání se zdravým organismem

⇩

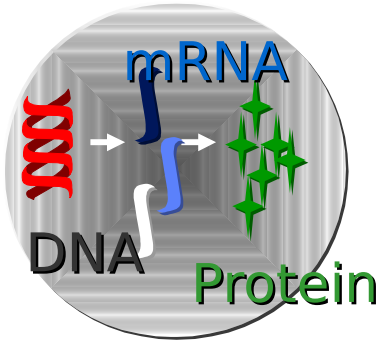
Jsme schopní
korelovat funkci produktů jednotlivých genů s **onemocněním**
NEMOC ⇔ GEN (Y)

⇩

Pochopit **podstatu** onemocnění

⇩

Najít nejvhodnější způsob léčby (cílená léčba), prevence a diagnostiky



Geny a onemocnění I. - příčiny

- Downův syndrom, hemofilie, cystická fibróza, svalová dystrofie, rakovina…
- Dědičné i získané, u některých stačí jediná *mutace* v patřičném genu a vzniká choroba, u jiných je zapotřebí více genetických změn

1. Změny ve struktuře DNA:
   - Mutace ve struktuře jednoho genu (jednonukleotidové polymorfizmy, delece, inzerce, amplifikace nukleotidů)
   - Aberace celého genu a nebo části chromozomu (delece, translokace, inzerce, amplifikace)
   - Aberace celých chromozomů
2. Změny v expresi a aktivitě genů a jejich produktů
3. Změny v posttranslačních úpravách proteinů

Geny a onemocnění II. - mutace

- Buňky v organismu se stále obnovují a dělí - při každém dělení replikují celý genom na nukleotid přesně. Tento proces není při velikosti lidského genu (3.2 bilionu nukleotidů) jednoduché.

- Proto existuje mnoho kontrolních mechanismů:
  - na opravu poškozené časti DNA
  - pro správnou distribuci chromozomů v procese mitózy/meiózy
  - pro případnou apoptózu (regulovanou smrt buňky) v případě nezvratných změn
  - apod....

- Genetické aberace vznikají selháním kontrolních mechanismů

Geny a onemocnění III. – aktivita genů

- Nejen mutace, ale i *nesprávná aktivita* genů může vést ke vzniku onemocnění.
- V lidské buňce probíhá každou chvíli obrovské množství procesů, přepisují se stovky genů a neustále se vytvářejí proteiny na základě vnitřních a venkovních podnětů.
- Tyto podněty jsou regulované stovkami regulačních mechanismů, které jsou opět založené na proteinech.
- Chyba v jednom z mechanismů může také skončit vyvinutím onemocnění.

Geny a onemocnění IV. - shrnutí

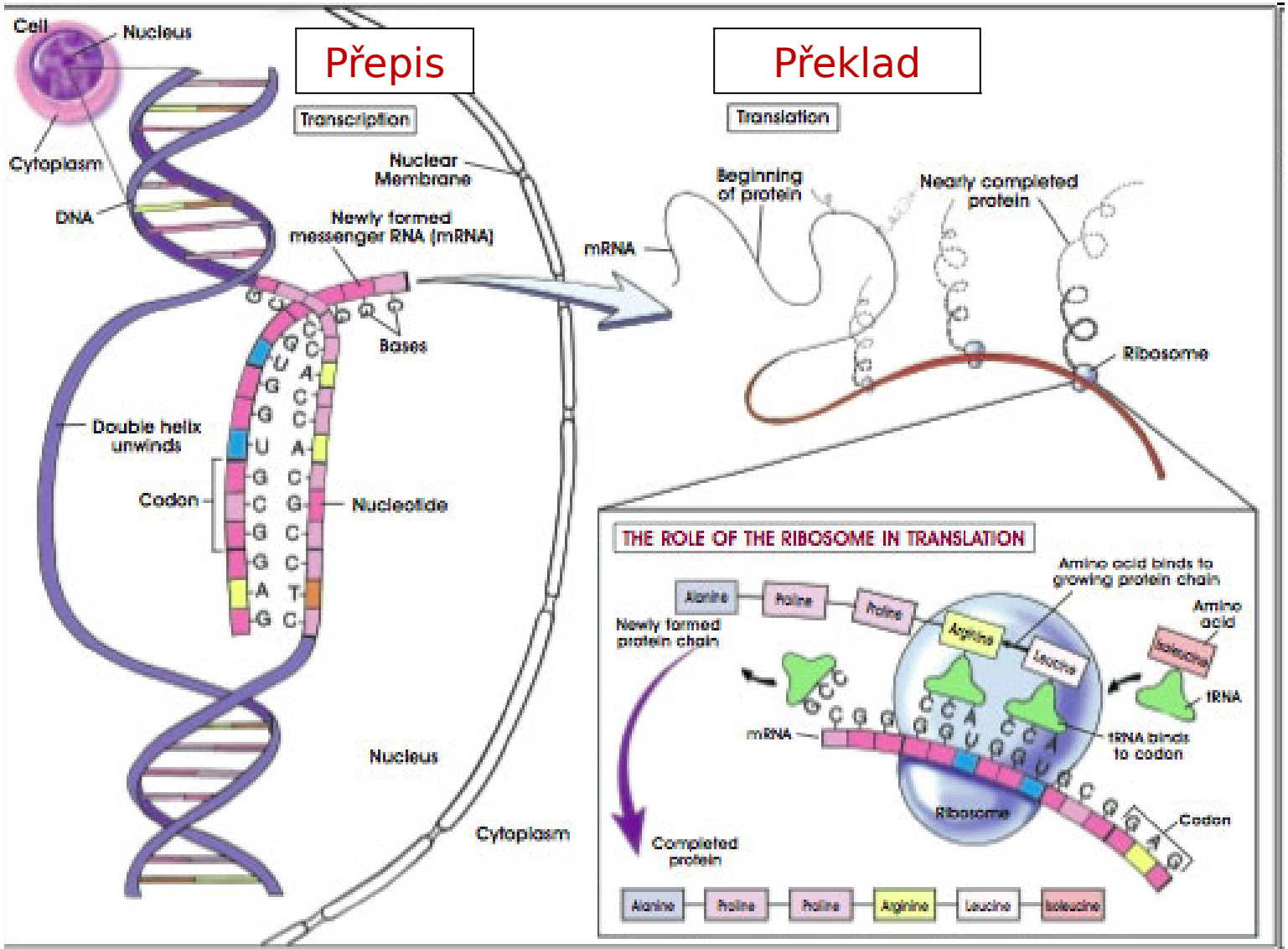
- Co způsobuje onemocnění – **proteiny a jiné funkční molekuly**, které mají změněnou svojí funkčnost, nebo expresi.

- Příčiny nesprávné funkce:
  - Mutace v příslušném genu, způsobující v důsledku změnu v sekvenci aminokyselin proteinu a tím jeho:
    - nefunkčnost
    - nadměrnou aktivitu
  - Změny v mechanismech kontroly exprese daného proteinu, který je následně produkovaný
    - v nedostačujícím množství
    - v nadměrném množství
  - Změny v postranslačních úpravách a sekundární/terciární struktuře proteinu

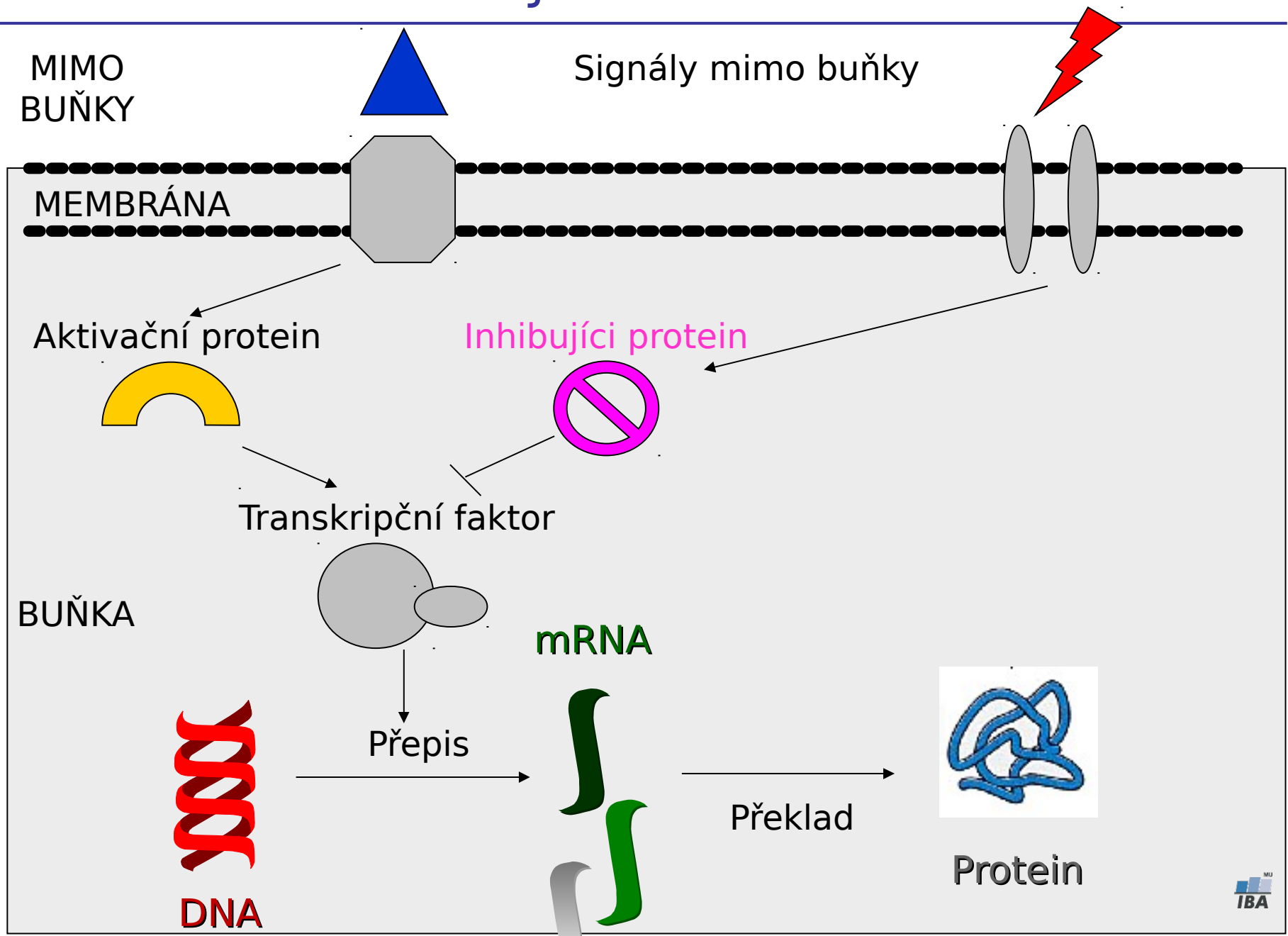
Ústřední dogma molekulární biologie

Přepis Překlad

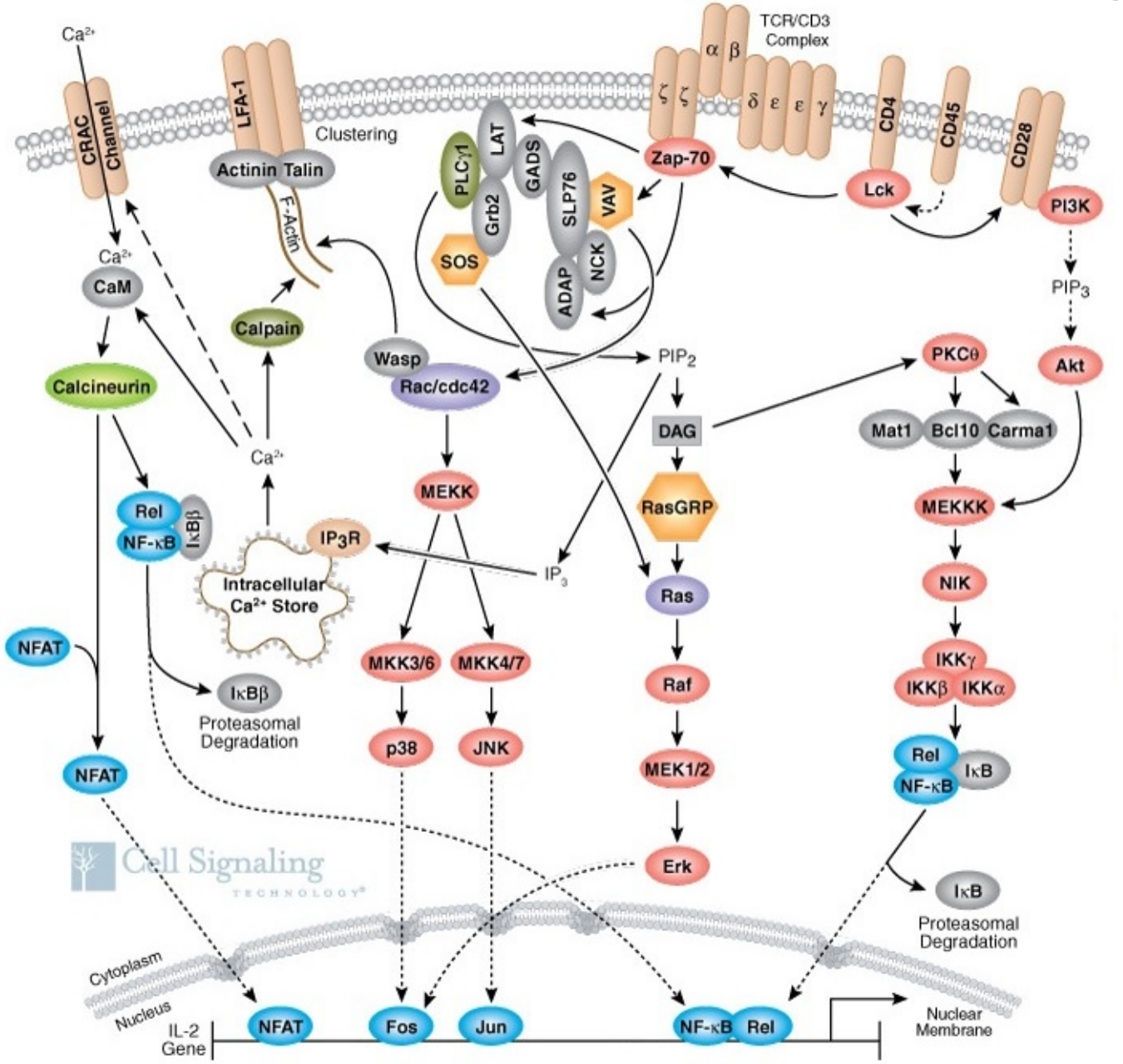
DNA -> mRNA -> protein



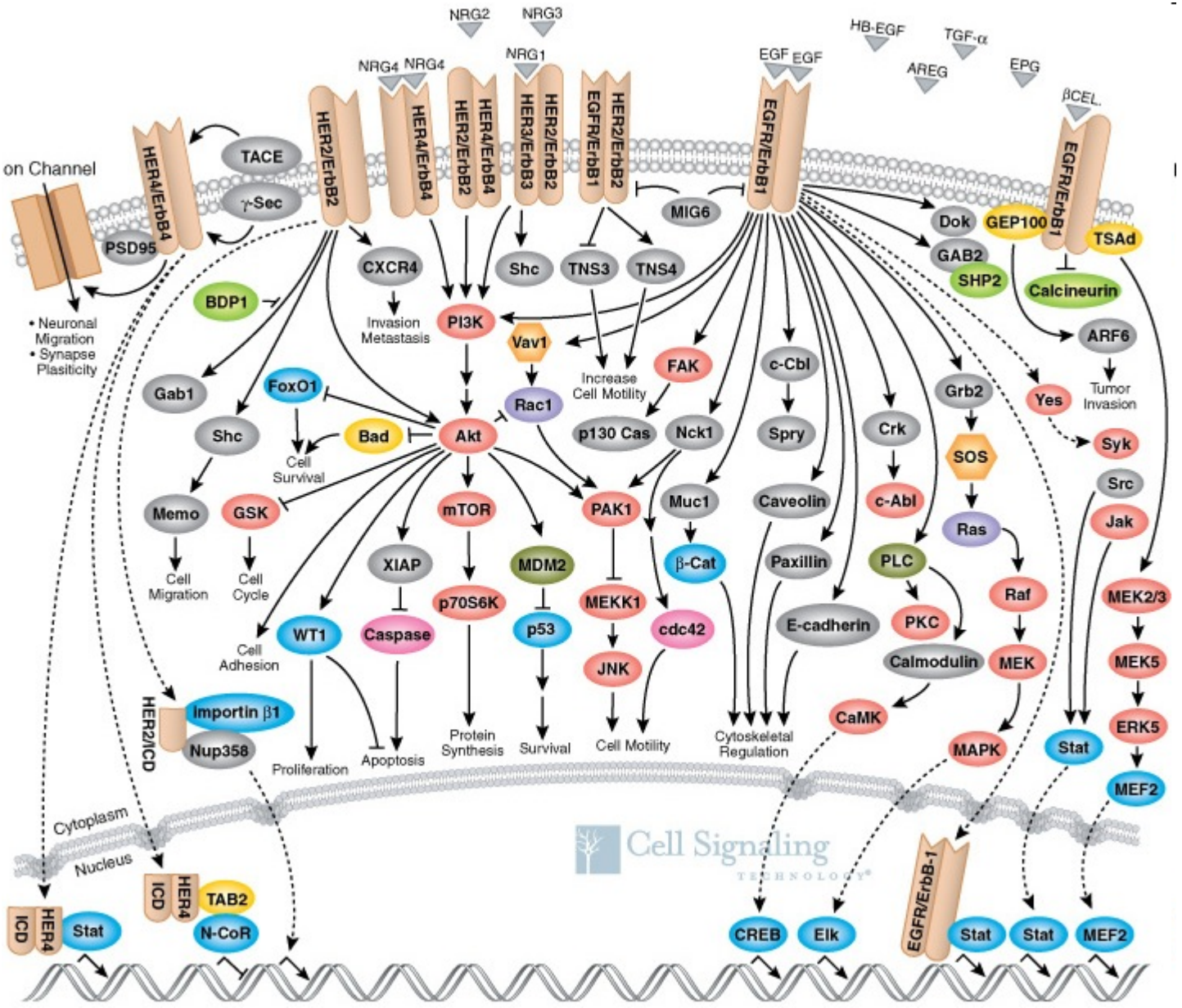
Co ještě víme



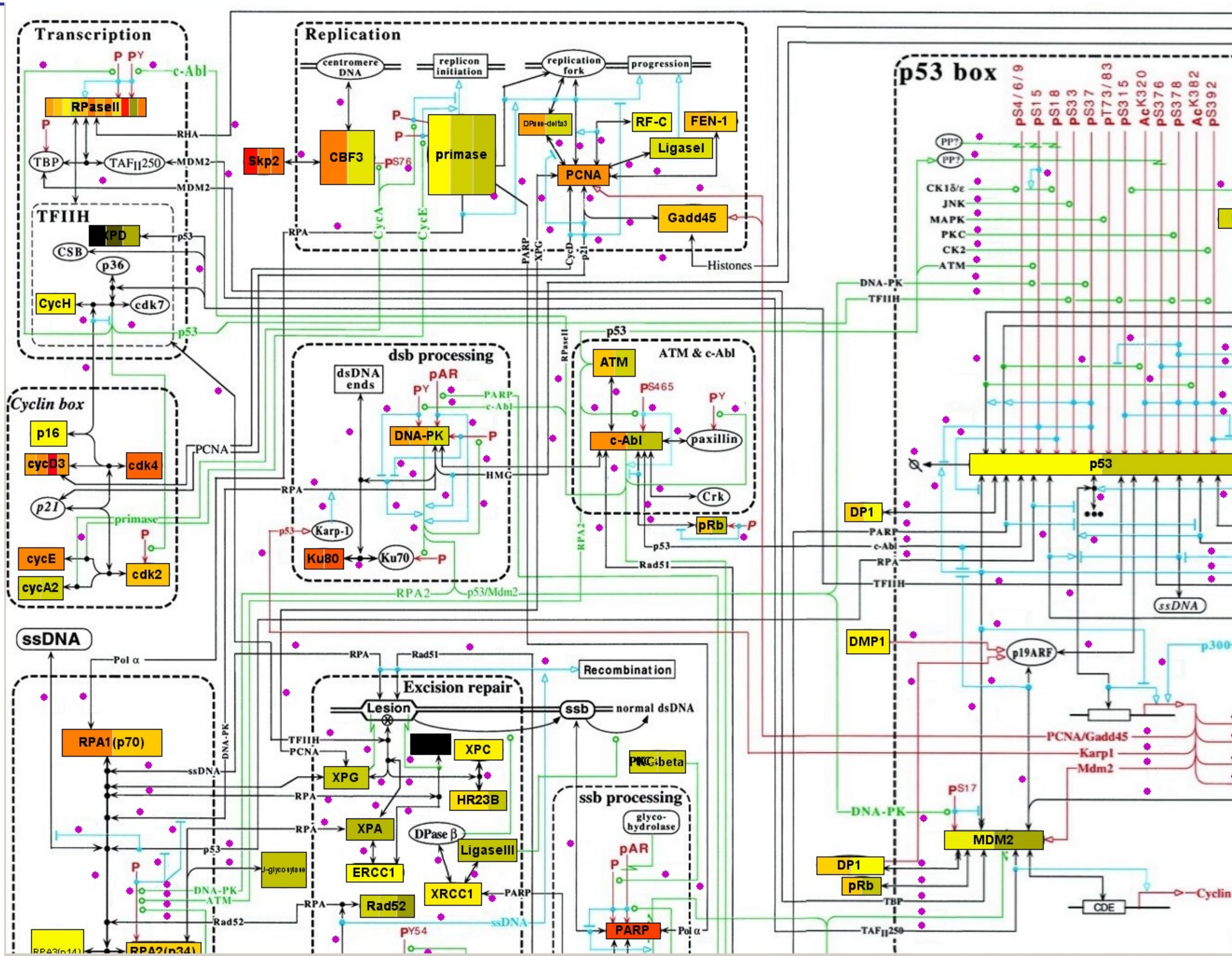
Ale víme ještě víc



..a ještě víc...

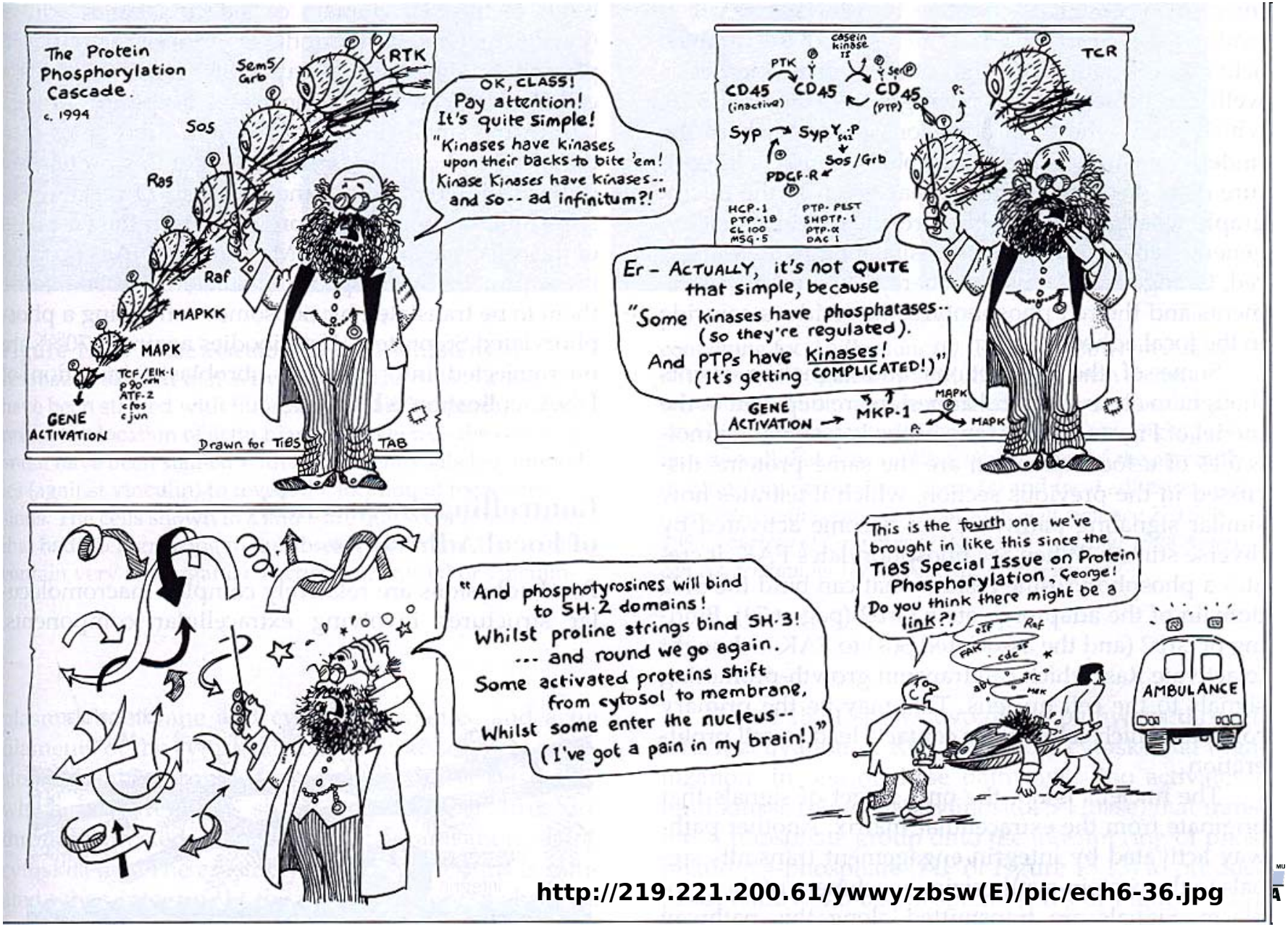

NRG2
NRG3
NRG1
NRG4
NRG4
HB-EGF
TGF-α
AREG
EPG
EGF
EGF
βCEL.
HER2/ErbB2
HER4/ErbB4
HER2/ErbB2
HER4/ErbB4
HER3/ErbB3
HER2/ErbB2
EGFR/ErbB1
HER2/ErbB2
EGFR/ErbB1
EGFR/ErbB1
on Channel
TACE
γ-Sec
PSD95
HER4/ErbB4
MIG6
Dok
GEP100
GAB2
SHP2
TSAd
Calcineurin
ARF6
Tumor Invasion
• Neuronal Migration
• Synapse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

…a ještě víc…



Transcription
RPaseII
TBP
TAFII250
RHA
MDM2
c-Abl
TFIIH
XPD
CSB
p36
CycH
cdk7
p53
Replication
centromere DNA
replicon initiation
replication fork
progression
Skp2
CBF3
primase
DPase-delta3
RF-C
FEN-1
LigaseI
PCNA
Gadd45
Histones
RPA
CycA
CycE
PARP
XPG
CycD
p21
Cyclin box
p16
cycD3
cdk4
p21
primase
cycE
cdk2
cycA2
dsb processing
dsDNA ends
pAR
DNA-PK
HMG
Karp-1
Ku80
Ku70
ATM & c-Abl
ATM
c-Abl
paxillin
Crk
pRb
Rad51
p53/Mdm2
RPA2
p53 box
pS4/6/9
pS15
pS18
pS33
pS37
pT73/83
pS315
AcK320
pS376
pS378
AcK382
pS392
CK1δ/ε
JNK
MAPK
PKC
CK2
ATM
DNA-PK
TFIIH
p53
DP1
PARP
c-Abl
RPA
ssDNA
DMP1
p19ARF
p300
PCNA/Gadd45
KarpI
Mdm2
pS17
MDM2
DP1
pRb
TBP
TAFII250
CDE
Cyclin A
ssDNA
Pol α
RPA1(p70)
RPA2(p34)
RPA3(p14)
Rad52
U-glycosylase
Excision repair
Lesion
ssb
normal dsDNA
Recombination
TFIIH
XPG
XPC
HR23B
XPA
DPase β
LigaseIII
ERCC1
XRCC1
Rad52
pY54
PKC-beta
ssb processing
glyco-hydrolase
PARP

…ale je velmi obtížné to vše propojit a interpretovat



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

Co zkoumáme v genomice a proteomice

- U genů můžeme zkoumat jejich
  - **Strukturu a její změny** – sekvence nukleotidů A, C, G, T
  - **Množství** – zda jsou a nebo nejsou přítomné a v jakém počtu kopií
  - **Aktivitu** – zda se gen přepisuje do mRNA a v jakém množství

- U proteinů zkoumáme
  - **Složení** – z jakých aminokyselin
  - **Strukturu** – jak jsou řetězce peptidů uspořádané do 3D struktur
  - **Množství** – zda jsou a nebo nejsou přítomné a v jakém množství
  - **Funkci** – modelování, identifikace aktivních vazebných míst

- Další fáze je **modelování** komplexních buněčných systémů – proteinové interakce, buněčné dráhy, regulační a metabolické sítě …

Metody studia genomu a proteomu

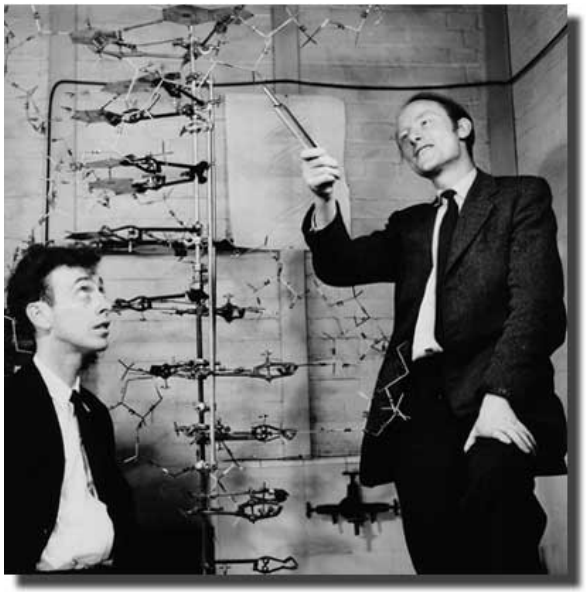
- *Klasické metody* molekulární biologie a cytogenetiky:
  - Metody zkoumající jen jeden nebo několik genů a proteinů v jednom experimentu:
    - PCR, RT-PCR, real-time PCR
    - FISH (fluorescence in-situ hybridization)
    - gelová elektroforéza, …

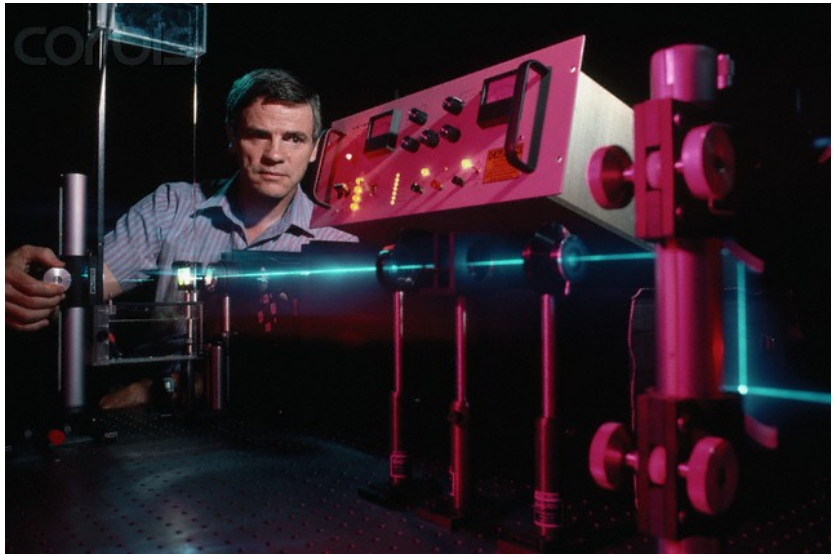
- *Vysokopokryvné metody* molekulární biologie:
  - schopné zkoumat tisíce molekul v jednom experimentu….

… jak vznikly?

Od Watsona & Cricka po Leroya Hooda

- Na začátku byl dvoušroubovicový model DNA…

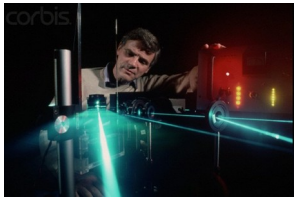




a na konci byly:

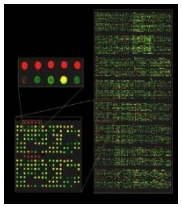
- automatické **sekvenátory** DNA a proteinů
- automatické **syntetizátory** DNA a proteinů

Nové možnosti

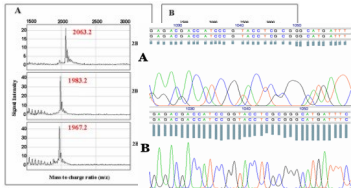


Sekvenátory umožnily rychle dekódovat sekvenci genů a proteinů

Znalost přesné sekvence umožnila navrhovat *specifické genové sondy* a syntetizátor umožňoval jejich rychlou a *automatickou výrobu*.



Otevřely se dveře pro nové, vysokopokryvné technologie, schopné analyzovat tisíce genů/proteinů v jednom experimentu!



Vysokopokryvné metódy I.

- Analýza genomu (od nukleotidových sekvencí po úplně anotovaný genom) a **transkriptom**
  - Analýza struktury
  - Analýza exprese
  - Srovnávací genomika
  - Regulace genomu

- Analýza proteomu (od hmostnostních spekter – přes komplexní struktury proteinových shluků - po analýzu funkce proteinů)
  - Analýza struktury
  - Analýza exprese
  - Analýza funkce

- Modelování komplexních systémů – proteínové interakce, buněčné dráhy, regulační a metabolické sítě...

Analýza genomu

- Od nukleotidových sekvencí po úplně anotovaný genom
- Analýza **struktury**
  - DNA sekvenace, Chip-seq, WES (whole exome sequencing), WGS
  - Srovnávací genomika – aCGH čipy, SNP polymorfismy, alternative splicing arrays, fingerprinting
- Analýza **aktivity** (exprese) – Mikročipy, SAGE, MPSS, Expressed sequence tags (ESTs), RNA-seq, …
- Regulace genomu
  - Chip-on-chip
  - Epigenetika (mikročipy, metylace…)

Analýza proteomu

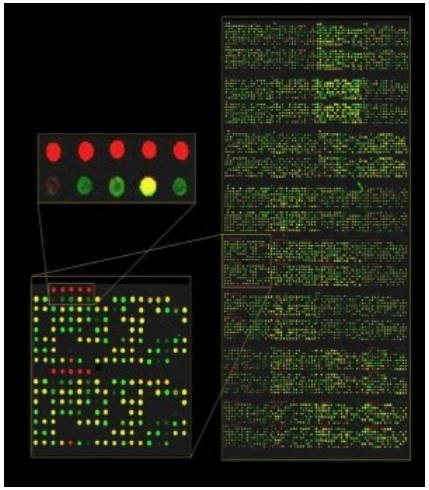
Od hmostnostních spekter – přes komplexní struktury proteinových shluků - po analýzu funkce proteinů

- Analýza **struktury**: Proteínová sekvenace
- Analýza **expresee**: Hmotnostní spektrometrie, Proteínové mikročipy...
- Analýza **funkce**: Modelování makromolekulárních systémů – odvození vlastností z atomových interakcí

Data vysokopokryvních metod I.

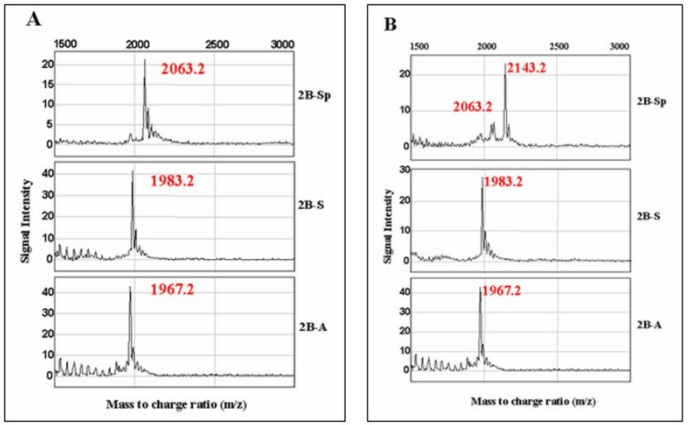
- Moderní vysokopokryvní technologie produkují obrovské tabulky komplexních dat

Mikročipy

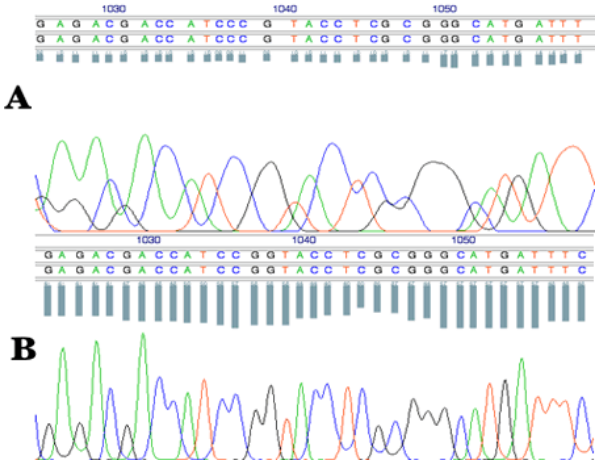


☑ Exprese 10 000 – 100 000 transkriptů u 100 – 1000 vzorků

MASS – hmotnostní spektrometrie



Sekvenace DNA



☑ Tisíce spekter proteinů – GB datové soubor

☑ Genom s bilióny nukleotidů

Dáta vysokopokryvných metód II.

Dátový súbor z vysokopokryvných experimentů– pohled biologa

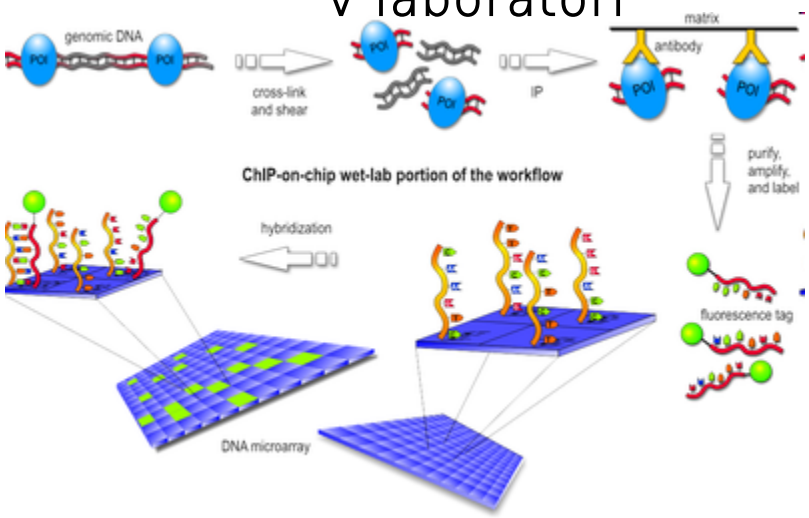
"In principle, the string of genetic bits holds long-sought secrets of human development, physiology and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate".

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

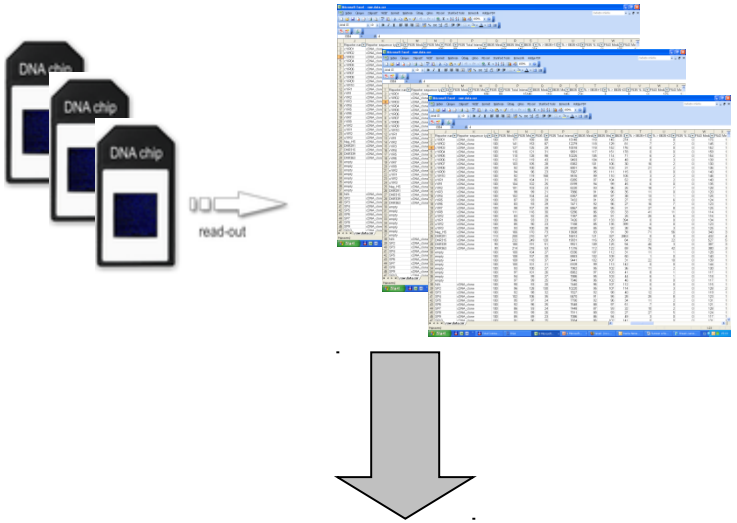
Hledání jehly v kupce sena?

Datový soubor z vysokopokryvných experimentů– pohled matematického biologa

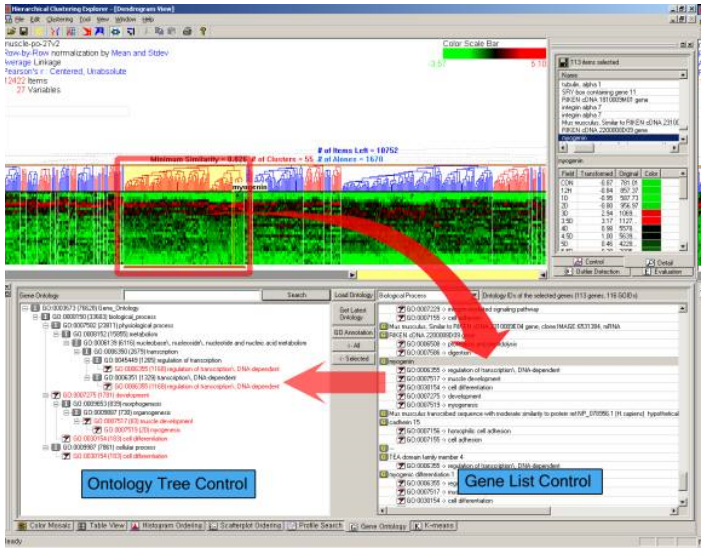
1. Příprava a provedení experimentu v laboratoři



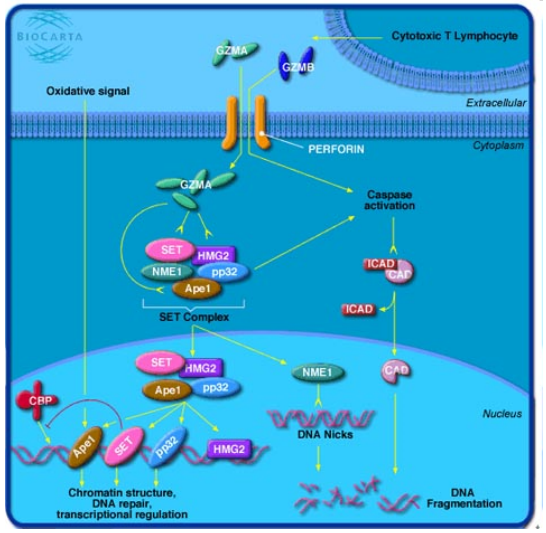
2. Extrakce a úprava dat



3. Statistická analýza dát



4. Biologická a klinická interpretáce



Důležité!

- 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 NGS dát – v samostatnom predmete Bi5444

Vysokopokryvné metódy – čo si priblížime

Podrobnejšie si predstavíme technológie:

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

- Hmotnostná spektrometria

Shrnutí první části

- Velká část rozmanitosti života včetně onemocnění se dá zřejmě obsáhnout studiem **funkce genomu a proteomu**
- V současnosti existují speciální **vysokopokryvné metody (high-density methods)**, které umožňují zkoumat tisíce genů a proteinů v jednom vzorku a jednom experimentu
- Tyto metody produkují **obrovské množství dat** a vyžadují specializovanou statistickou analýzu