

Analýza genomických a proteomických dat

Mgr. Eva Budinská, Ph.D.

RNDr. Ivana Ihnatová, Ph.D.

Jaro 2018

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 na přednáškách včetně prezentace projektu (**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 a prezentace projektu	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
- Každý provádí výběr vlastní témy, kterou musíme nejdříve schválit
- Projekt nutno odevzdat před zkouškou, pouze po odevzdání a obdržení 10 bodů z projektu je možné přihlásit sa na zkušební termín
- Nejzazší termín odevzdání projektů:
 - pro udělení zápočtu: **25.6.2018**
 - pro kontrolu počtu bodů: **7 dní** před zkušebním termínem
- **1x** možnost opravy projektu

Kapitola I.

Současné výzvy genomiky a proteomiky

Význam studia genomiky a proteomiky

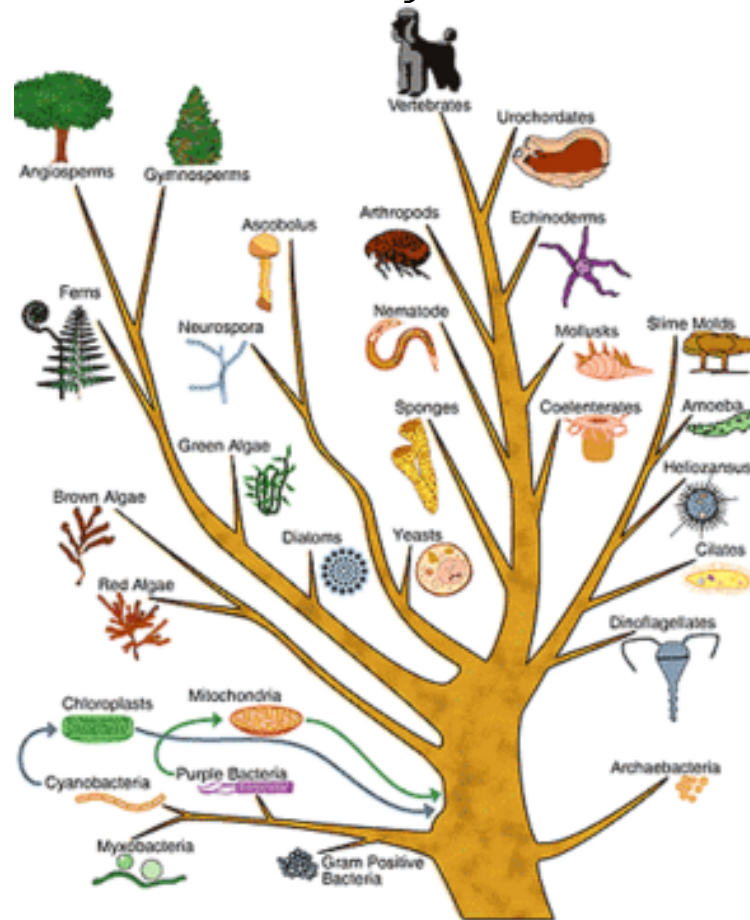
- V biologii jsme se znalostmi dostali na nejmenší jednotky, které mají komplexní biologický význam
- GENY a PROTEINY, dále jsou už jen nukleotidy a aminokyseliny a ještě níž jsou jen menší molekuly a atomy a ... subatomární částice.
- Studujeme složení molekul a hlavně jejich funkcí v organismu

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

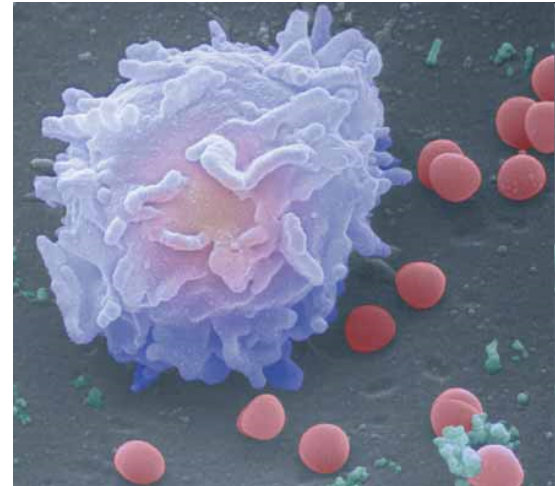
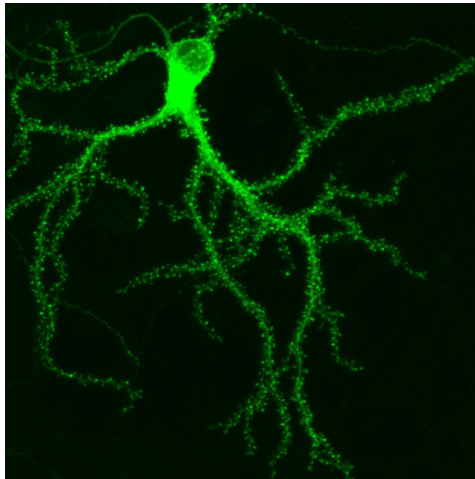
- Odolnost bakterií na antibiotika podmíněná mutacemi.
- Adaptace na extrémní podmínky - život ve vesmíru, v sopce, sírných pramenech, vařících pramenech a mrazech do -70



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



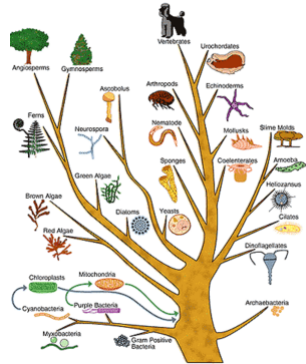
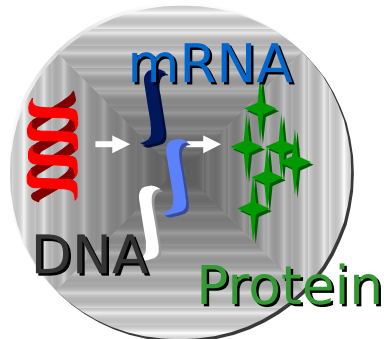
studovat tak evoluční propojení a vytvářet fylogenetické stromy



aktivitu genů a proteinů organismů 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

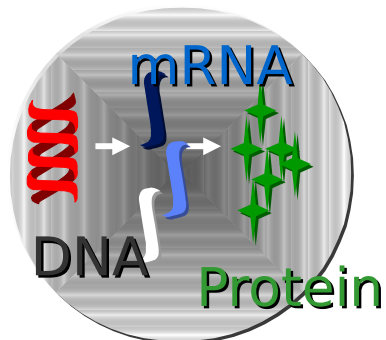
NEMOC ↔ GEN (Y)



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, inserce, amplifikace nukleotidů)
- Aberace celého genu a nebo části chromozomu (delece, translokace, inserce, 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 části 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

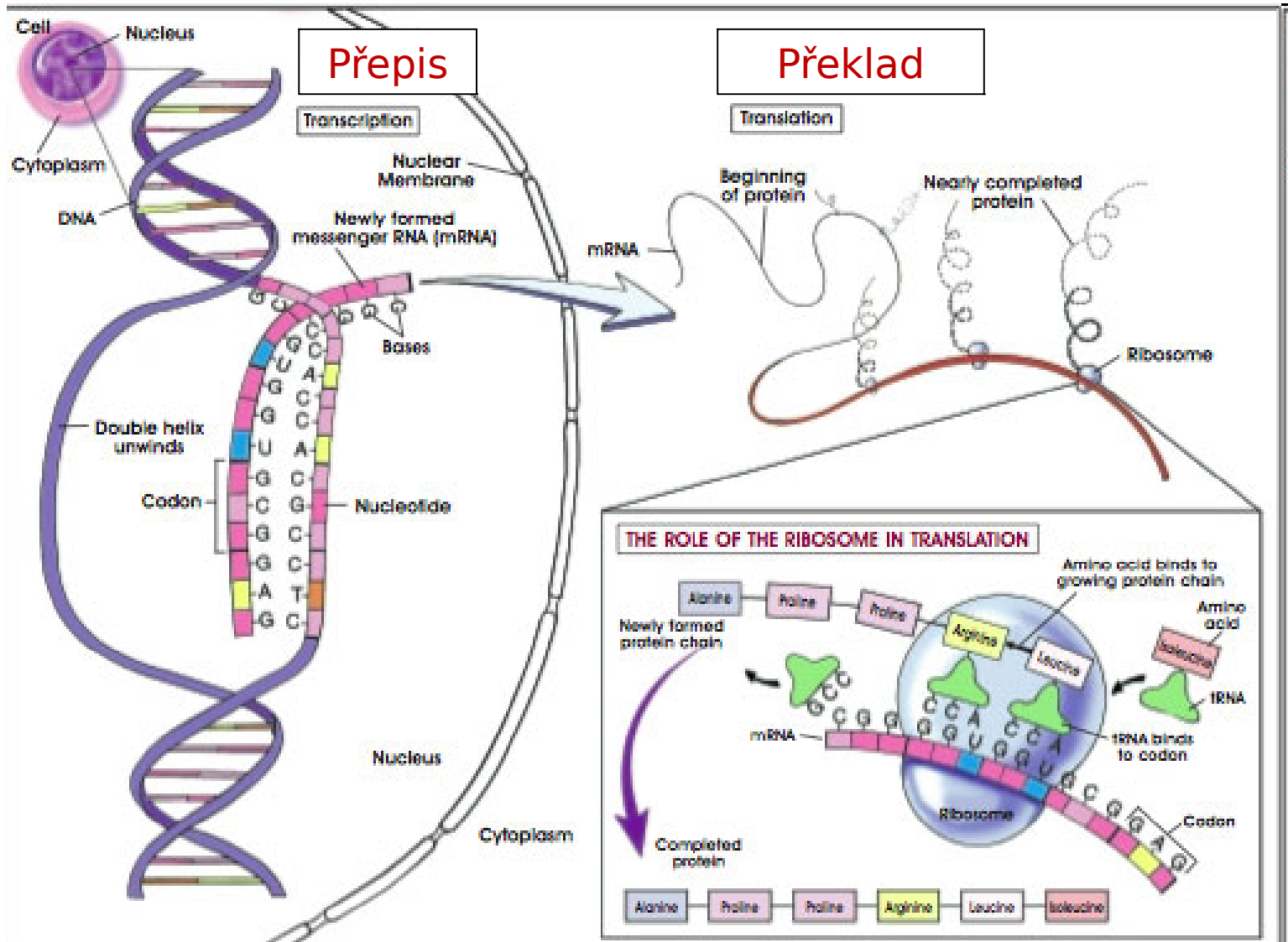
- Nejen mutace, ale i *nesprávná aktivita* genů může vést ke vzniku onemocnění.
- V lidské buňce probíhá každou chvíli obrovské množství procesů, přepisují se stovky genů a neustále se vytvářejí proteiny na základě vnitřních a venkovních podnětů.
- Tyto podněty jsou regulované stovkami regulačních mechanismů,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

Geny a onemocnění IV. - shrnutí

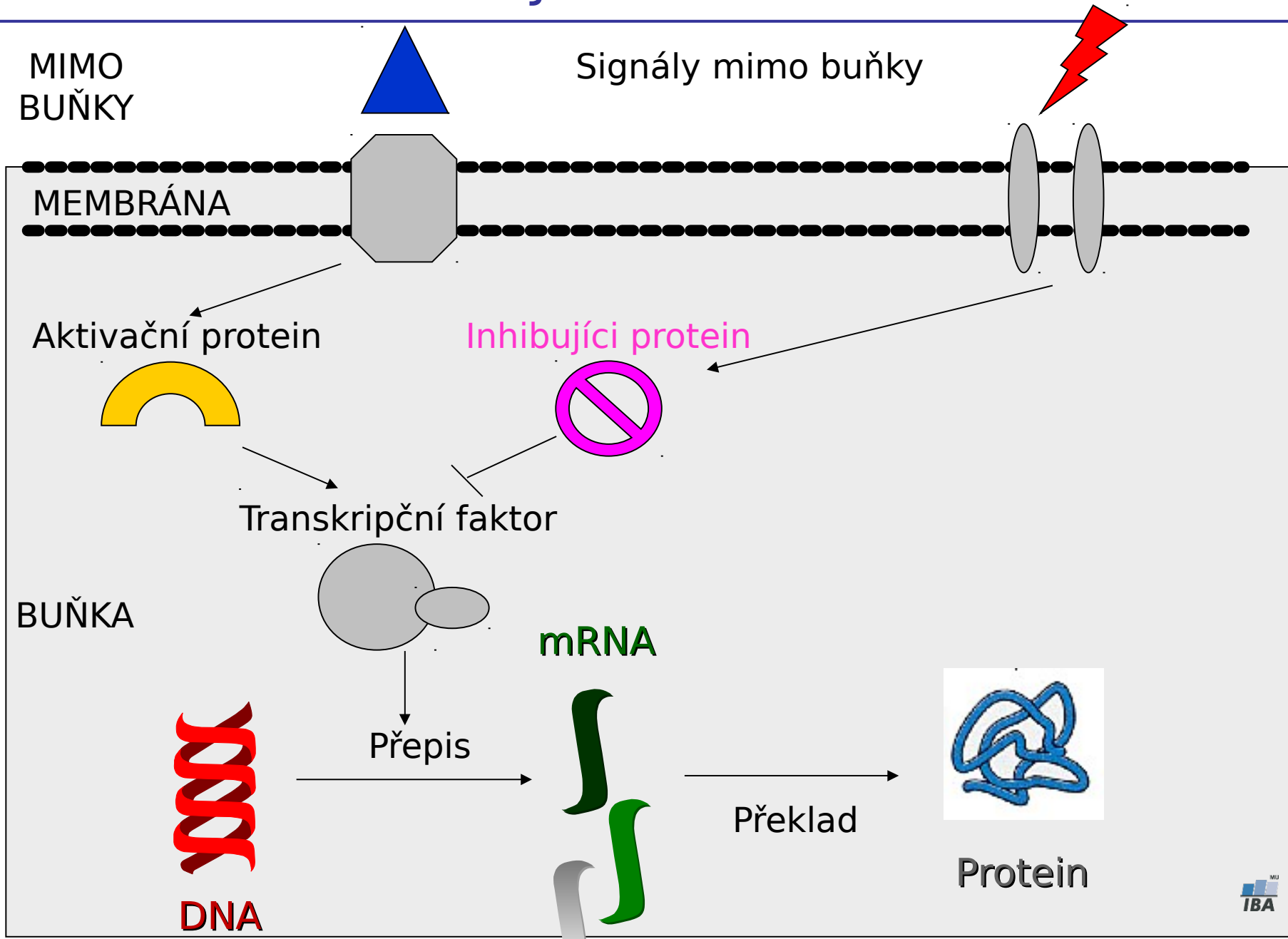
- Co způsobuje onemocnění – **proteiny a jiné funkční molekuly**, které mají změněnou svojí funkčnost, nebo expresi.
- Příčiny nesprávné funkce:
 - **Mutace v příslušném genu**, způsobující v důsledku změnu v sekvenci aminokyselin proteinu a tím jeho:
 - nefunkčnost
 - nadměrnou aktivitu
 - **Změny v mechanismech kontroly exprese daného proteinu**, který je následně produkován
 - v nedostačujícím množství
 - v nadměrném množství
 - **Změny v postranlačních úpravách** a sekundární/terciární struktuře **proteinu**

Ústřední dogma molekulární biologie

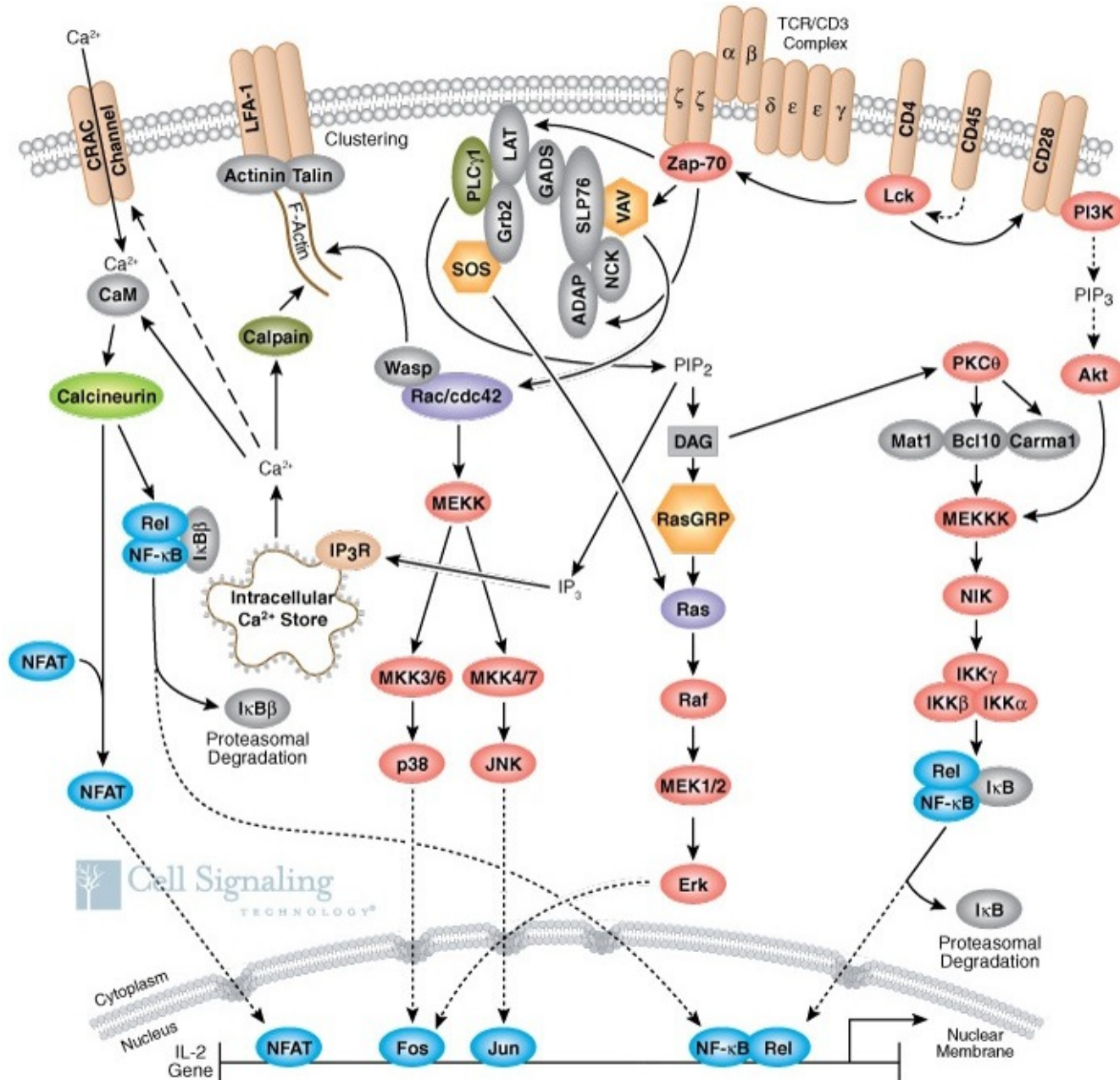
Přepis Překlad
DNA -> mRNA -> protein



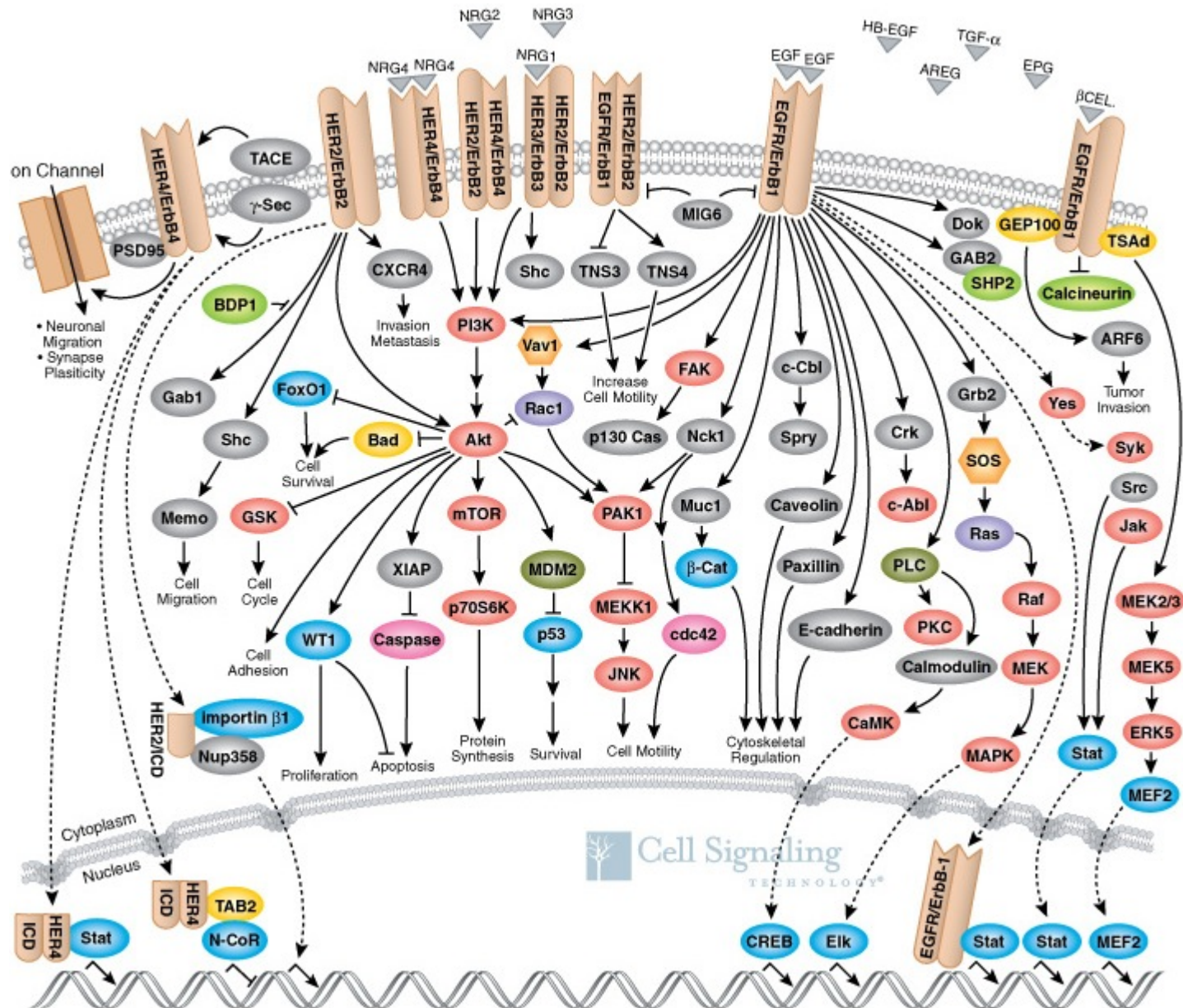
Co ještě víme



Ale víme ještě víc

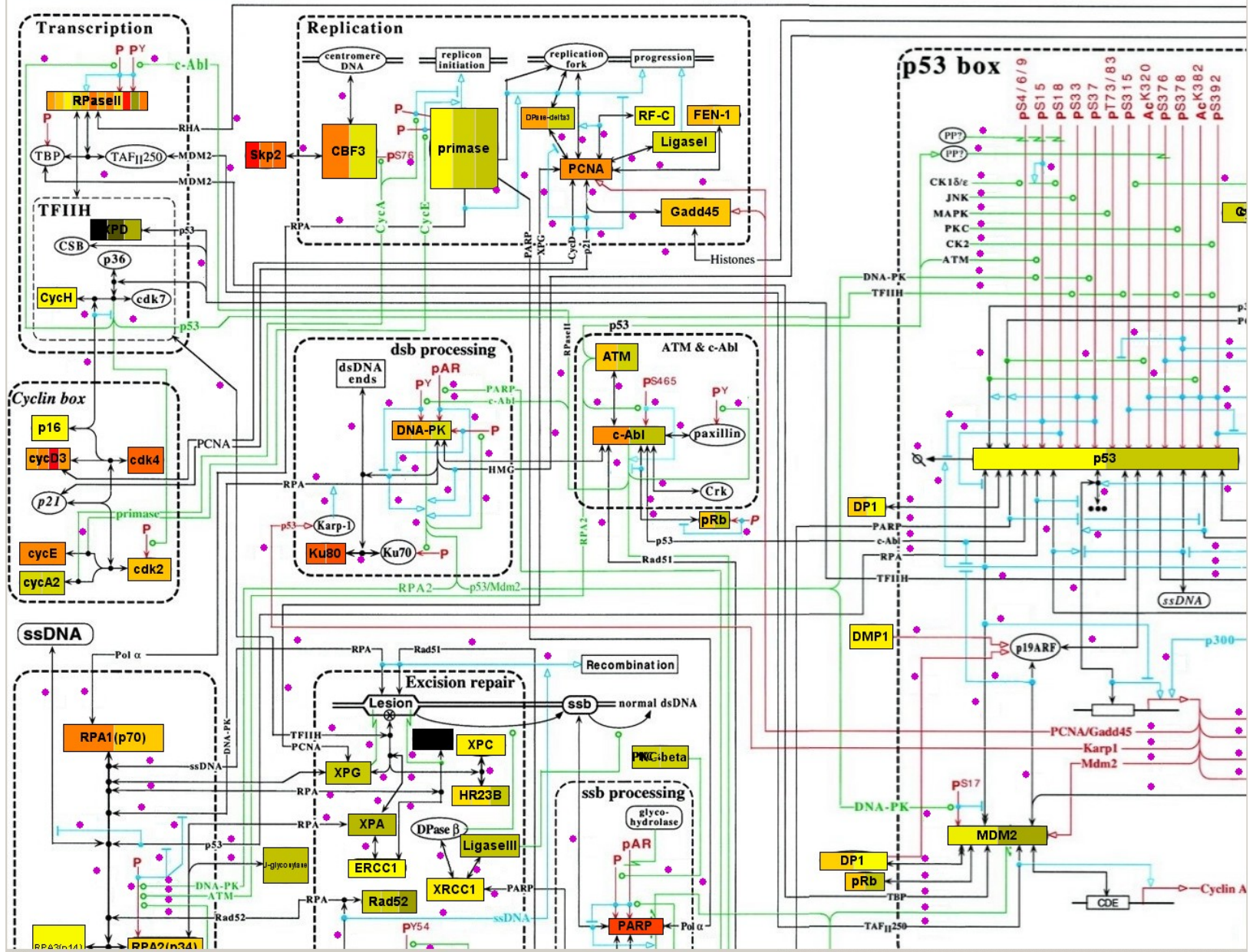


..a ještě víc...

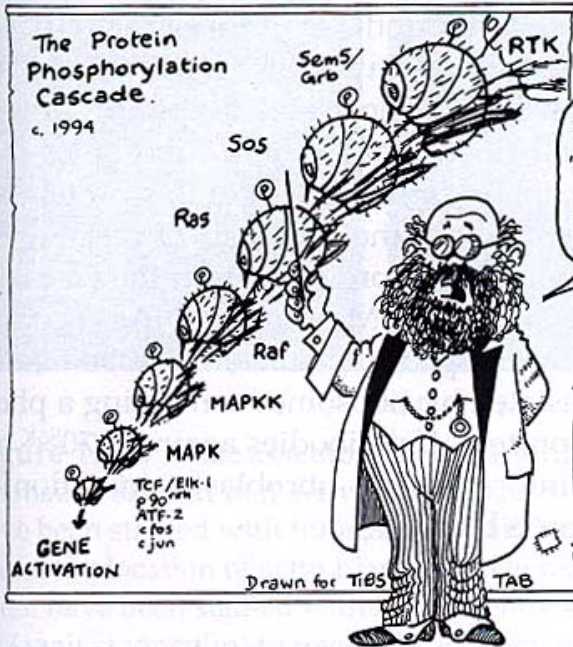


Cell Signaling
TECHNOLOGY®

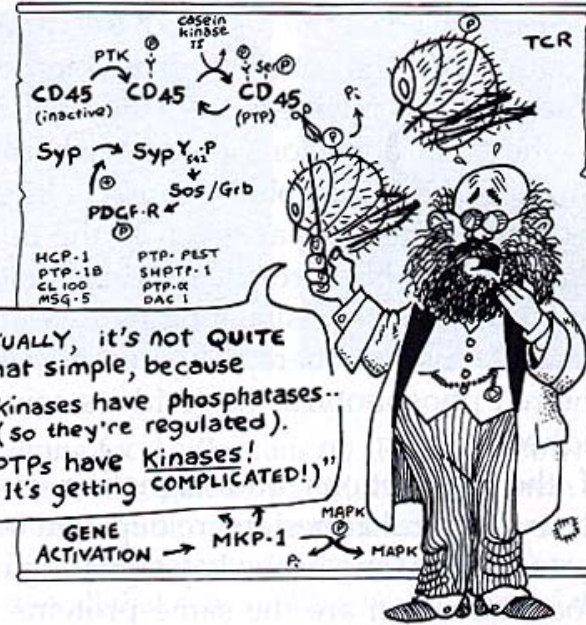
...a ještě víc...



...ale je velmi obtížné to vše propojit a interpretovat



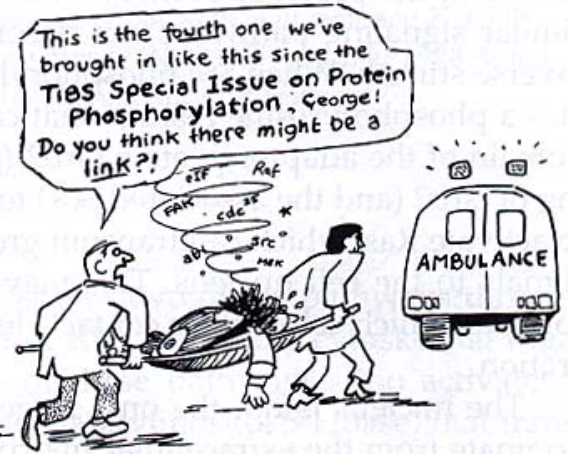
OK, CLASS!
Pay attention!
It's quite simple!
"Kinases have kinases
upon their backs to bite 'em!
Kinase Kinases have kinases--
and so-- ad infinitum?!"



Er - ACTUALLY, it's not QUITE
that simple, because
"Some kinases have phosphatases--
(so they're regulated).
And PTPs have kinases!
(It's getting COMPLICATED!)"



"And phosphotyrosines will bind
to SH-2 domains!
Whilst proline strings bind SH-3!
... and round we go again.
Some activated proteins shift
from cytosol to membrane,
Whilst some enter the nucleus--
(I've got a pain in my brain!)"



Co zkoumáme v genomice a proteomice

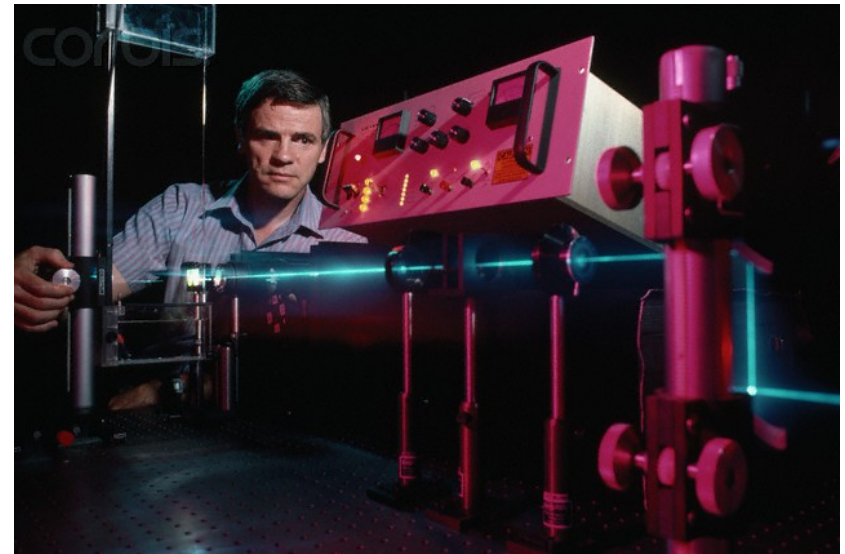
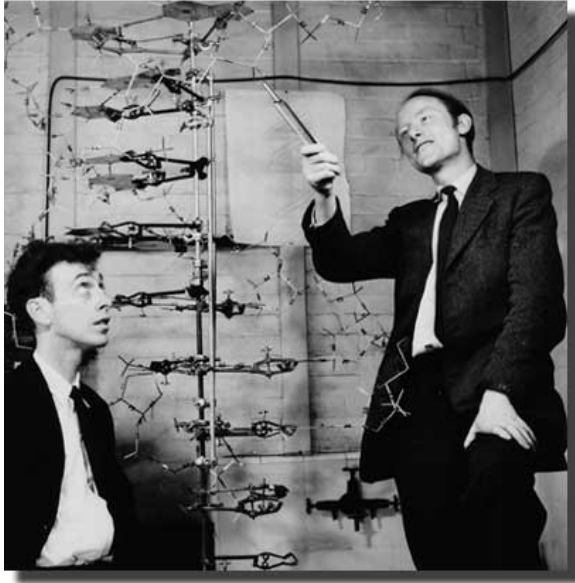
- U **genů** můžeme zkoumat jejich
 - **Strukturu a její změny** – sekvence nukleotidů A, C, G, T
 - **Množství** – zda jsou a nebo nejsou přítomné a v jakém počtu kopií
 - **Aktivitu** – zda se gen přepisuje do mRNA a v jakém množství
- U **proteinů** zkoumáme
 - **Složení** – z jakých aminokyselin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

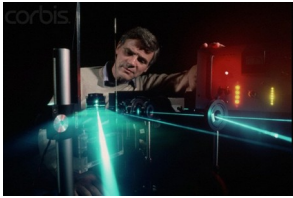
- Na začátku byl dvoušroubovicový model DNA...



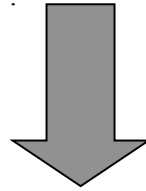
a na konci byly:

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

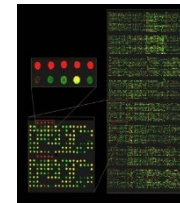
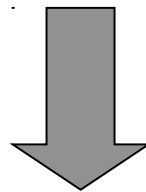
Nové možnosti



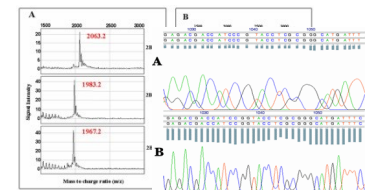
Sekvenátory umožnily rychle dekodovat sekvenci genů a proteinů



Znalost přesné sekvence umožnila navrhnout *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 **genómu** (od nukleotidových sekvencií po úplne anotovaný genóm) a **transkriptómu**
 - Analýza štruktúry
 - Analýza expresie
 - Porovnávacía genomika
 - Regulácia genómu
- Analýza **proteómu** (od hmostnostných spektier – cez komplexné štruktúry proteínových zhlukov - po analýzu funkcie proteínov)
 - Analýza štruktúry
 - Analýza expresie
 - Analýza funkcie
- Modelovanie **komplexných systémov** – proteínové interakcie, bunkové dráhy, regulačné a metabolické siete...

Analýza genómu

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

Analýza proteómu

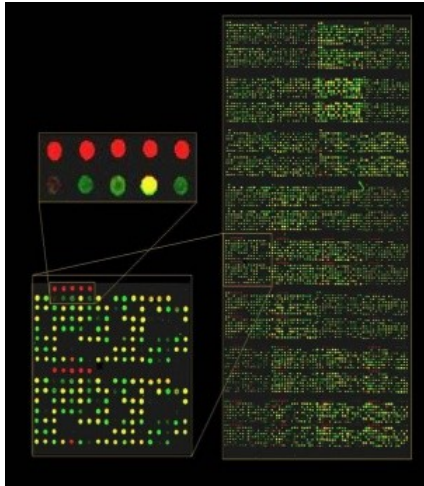
- Od hmotnostných spektier – cez komplexné štruktúry proteínových zhlukov - po analýzu funkcie proteínov
 - Analýza **štruktúry**: Proteínová sekvenácia
 - Analýza **expresie**: Hmotnostná spektrometria, Proteínové mikročipy...
 - Analýza **funkcie**: Modelovanie makromolekulárnych systémov – odvodzovanie vlastností z atómových interakcií

Dáta vysokopokryvných metód I.

- Moderné vysokopokryvné technológie produkujú obrovské tabuľky komplexných dát

Mikročipy

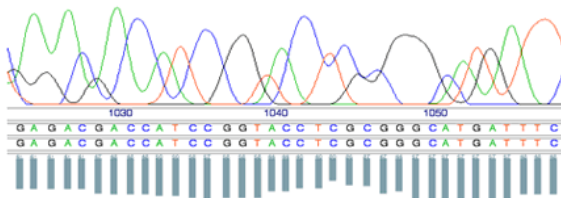
- Expresia 10 000 – 100 000 transkriptov u 100 – 1000 vzoriek



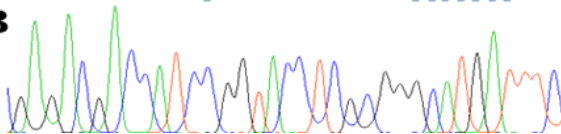
Sekvenácia DNA



A

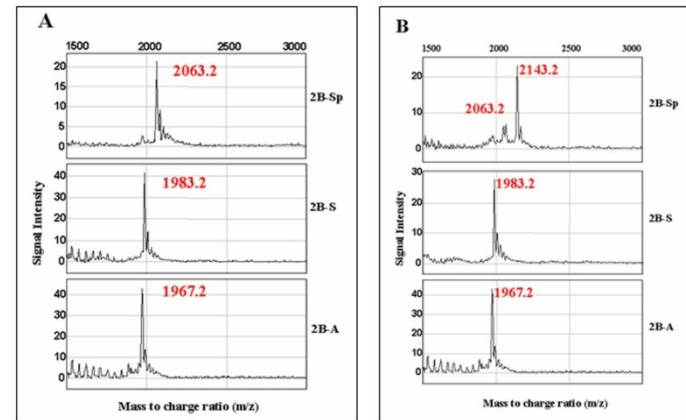


B



- Genóm s biliónmi nukleotidov

MASS – hmotnostná spektrometria



- Tisíce spektier proteínov – GB datové súbo

Dáta vysokopokryvných metod II.



Datový súbor z vysokopokrytých 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)

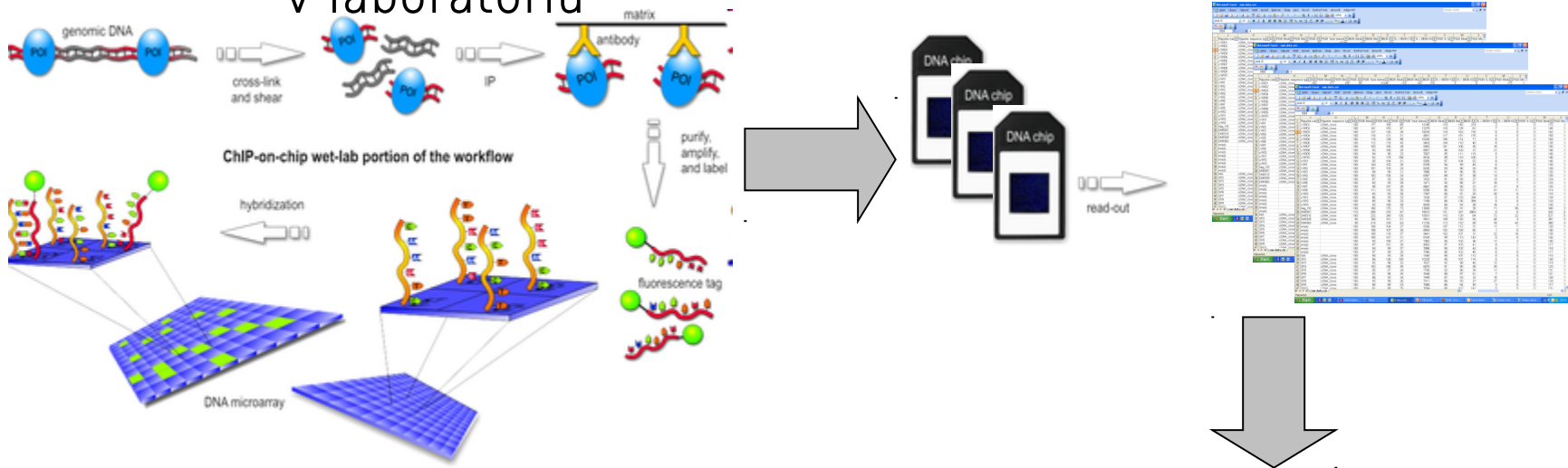


Hľadanie ihiel v kopách sena?

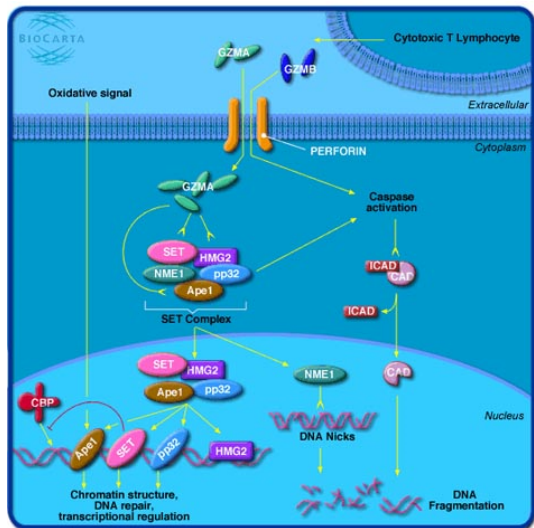


Dátový súbor z vysokopokryvných experimentov- pohľad matematického biológa

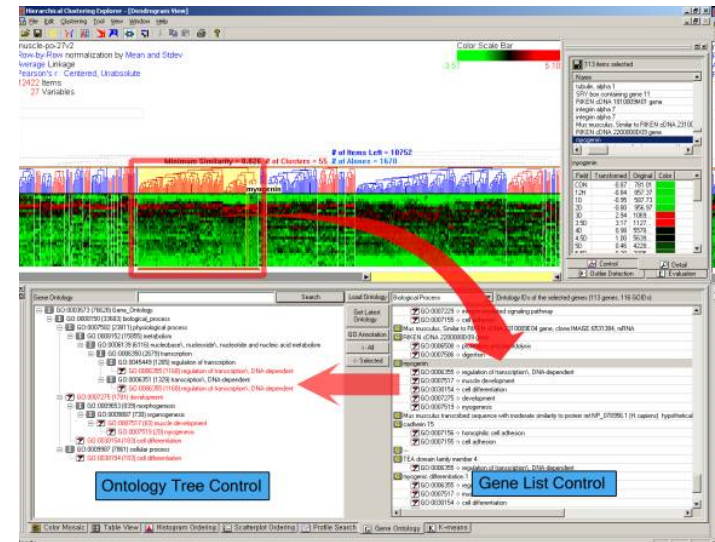
1. Príprava a vykonanie experimentu
2. Extrakcia a úprava dát v laboratóriu



4. Biologická a klinická interpretácia



3. Štatistická analýza dát



Dôležité!

- Veľká časť rozmanitosti života vrátane ochorení sa zrejme dá obsiahnuť 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