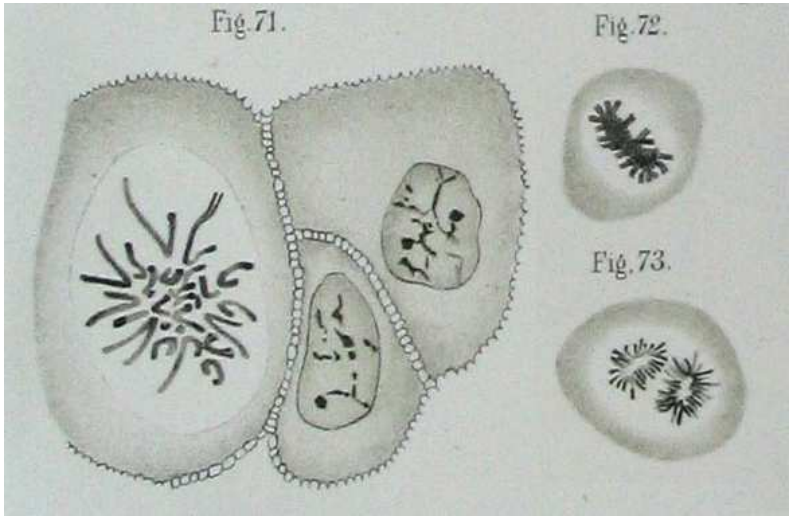
Biológia pre informatikov

Broňa Brejová

25.9.2014



Walther Flemming, 1881

Ďalšie informácie

- Tutoriál `http://www.rothamsted.ac.uk/notebook/index.php`
- NCBI books
  `http://www.ncbi.nlm.nih.gov/sites/entrez?db=books`
- Vysokoškolské učebnice molekulárnej biológie
- Anglická wikipédia, ...
- Zvelebil, Baum: Understanding Bioinformatics, kap. 1

Hlavné postavy

Deoxyribonukleová kyselina (DNA)

Obsahuje genetickú informáciu prenášanú z generácie na generáciu.
Dlhý reťazec nukleotidov z množiny $\{A, C, G, T\}$
(adenín, cytozín, guanín, tymín).
Informácia uložená v symbolickej, digitálnej forme.

Ribonukleová kyselina (RNA)

Blízka príbuzná DNA, tymín T nahradený uracylom U

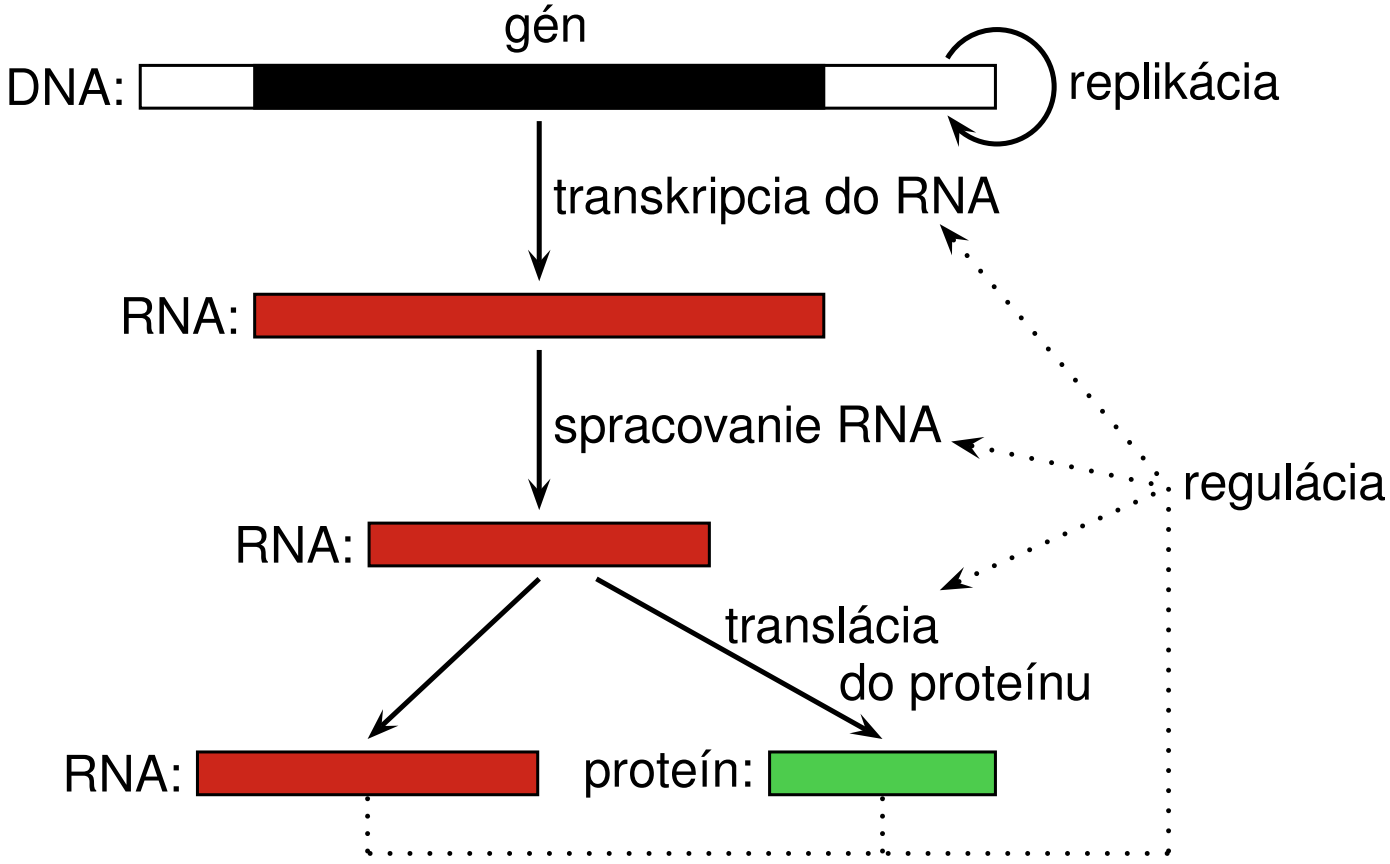
Proteíny (bielkoviny)

Katalyzujú biochemické reakcie v bunke (enzýmy),
prenášajú signály v rámci bunky/medzi bunkami,
sú dôležité pre stavbu bunky a pohyb.
Reťazec aminokyselín (20 rôznych aminokyselín).

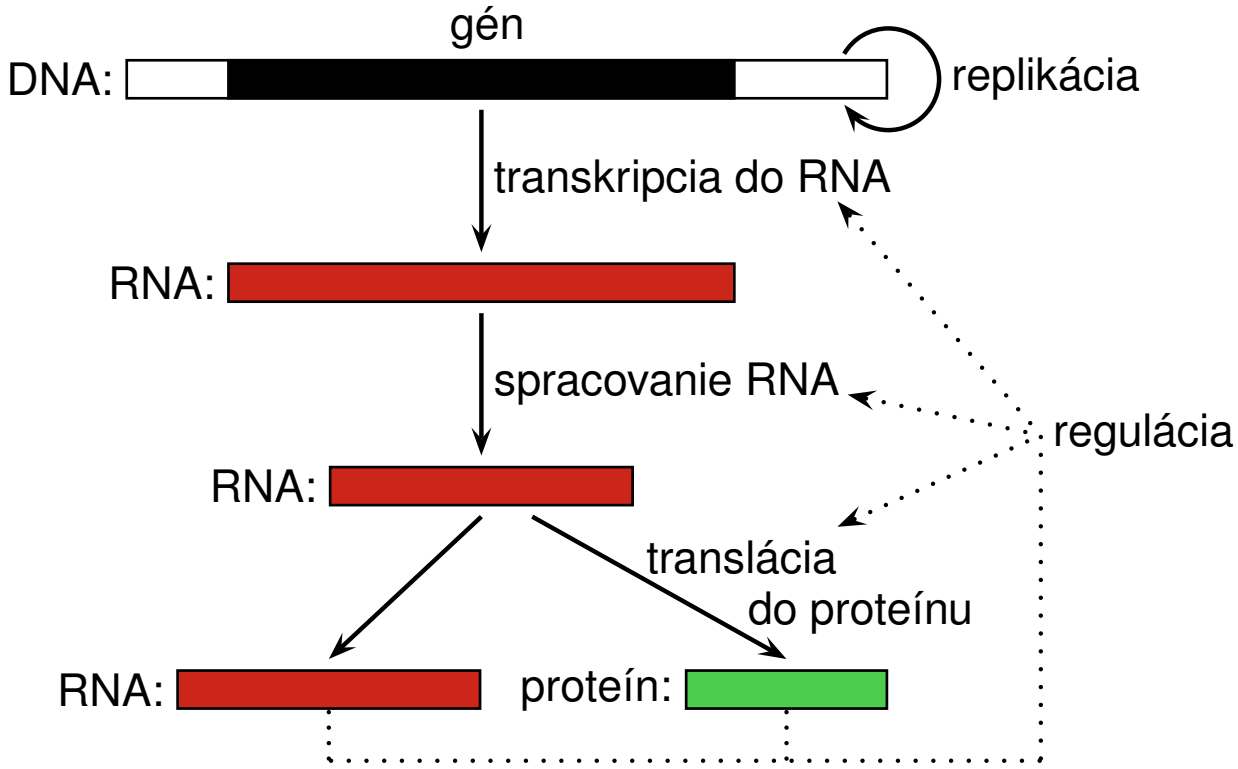
Aká informácia je uložená v DNA?

Gény: Predpisy na tvorbu proteínov a funkčných RNA molekúl.

Riadenie ich expresie: kedy a koľko sa má tvoriť.



Centrálna dogma (Francis Crick 1958,1970)

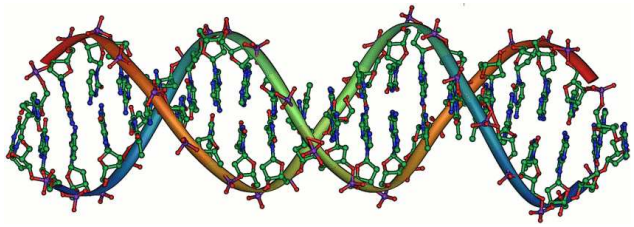


"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."

DNA, chromozómy

DNA: dve komplementárne vlákna, strands (páry A-T, C-G),
v opačnej orientácii (konce sa nazývajú 5' a 3').
Napr. ACCATG je komplementárny s CATGGT.

Tvar dvojitej špirály:



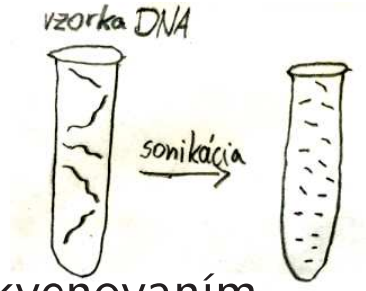
Dvojvláknová štruktúra poskytuje redundanciu, možnosť opravy pri poškodení jedného vlákna.
Pri delení bunky sa dvojvláknová DNA rozdelí a ku každému vláknu sa doplní komplement (DNA replikácia).

Chromozóm: Súvislý úsek dvojvláknovej DNA a podporných proteínov.

Ľudský genóm má 22 párov chromozómov plus dva pohlavné,
spolu 3GB.

Technológia: sekvenovanie DNA

- Postup na zisťovanie poradia báz v chromozómoch genómu.
- Zložitý proces:
  chromozómy sa nasekajú na krátke kúsky,
  z každého sa vyrobí veľa kópií,
  každý sa nasekvenuje zvlášť napr. Sangerovým sekvenovaním.
  – využíva prírodné enzýmy, napr. DNA polymerázu



- **Bioinformatický problém:** skladanie celej sekvencie z kúskov.
- Dostupnosť genómov umožňuje
  katalogizovať gény a iné funkčné úseky,
  hľadať podobnosti a rozdiely medzi druhmi a jedincami.

Sangerovo sekvenovanie (Sanger sequencing)

Sekvenujeme AGCTAGGACT (zobrazená sprava doľava)

Primer AGT + enzýmy + nukleotidy + modifikované ofarbené nukleotidy

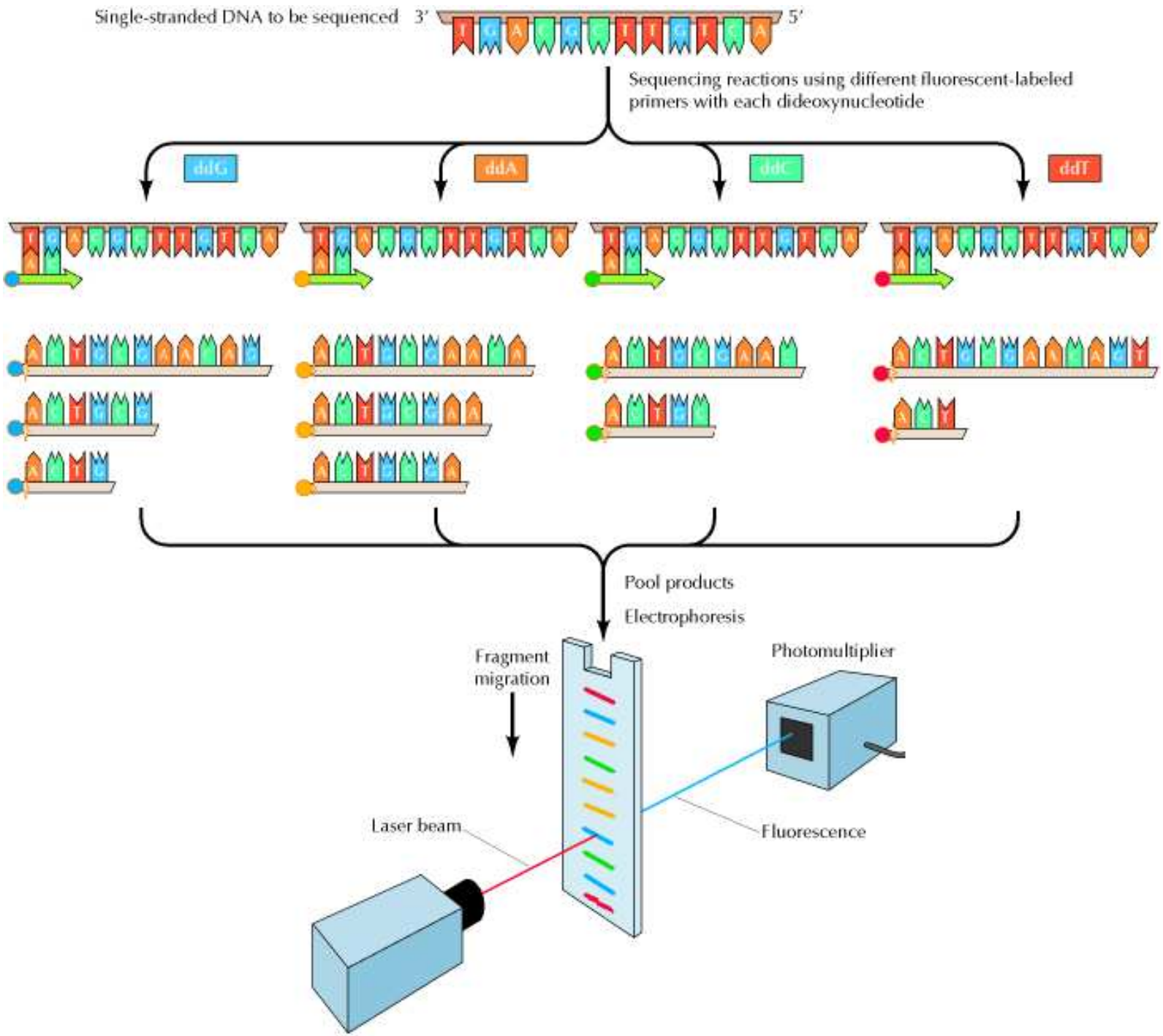
Výsledky sekvenovacej reakcie:

```
TCAGGATCGA
AGTCCTAGC

          TCAGGATCGA
          AGTCCTA

        TCAGGATCGA
        AGTCCTAGCT

                  TCAGGATCGA
                  AGTCCT

TCAGGATCGA
AGTCC

                TCAGGATCGA
                AGTCCTAG

      TCAGGATCGA
      AGTC
```

Na géli zoradíme podľa dĺžky:

```
AGTCCTAGCT
AGTCCTAGC
AGTCCTAG
AGTCCTA
AGTCCT
AGTCC
AGTC
```

Odčítaním farieb dostaneme komplementárne vlákno: AGTCCTAGCT

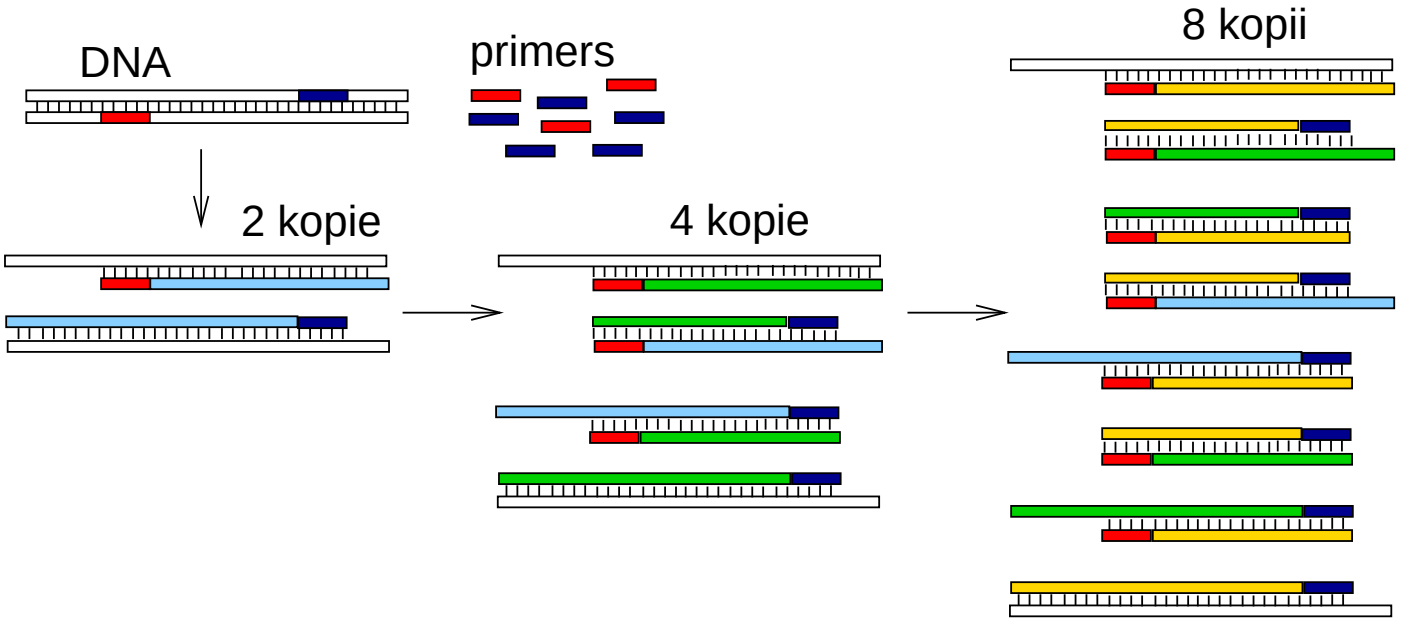

Single-stranded DNA to be sequenced
3'
5'
Sequencing reactions using different fluorescent-labeled primers with each dideoxynucleotide
ddG
ddA
ddC
ddT
Pool products
Electrophoresis
Fragment migration
Photomultiplier
Laser beam
Fluorescence

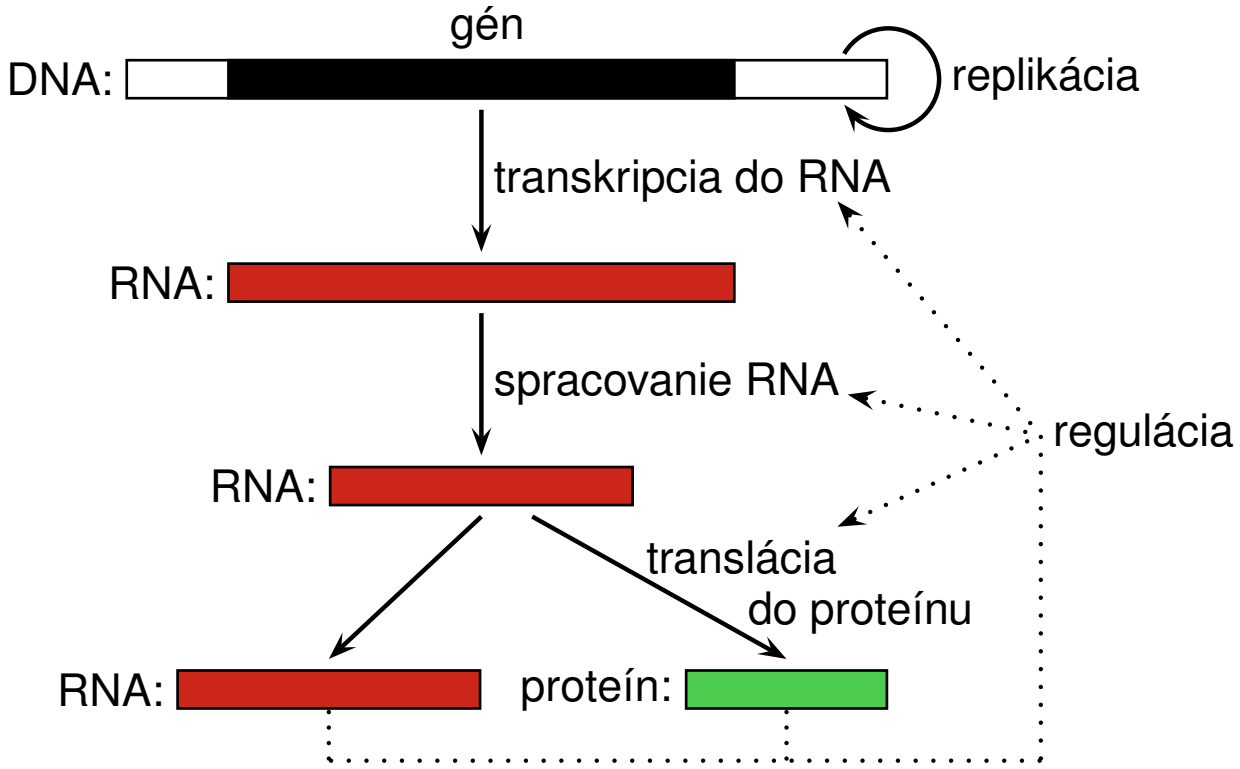
PCR (polymerase chain reaction)

Zvolíme si dva krátke úseky DNA (primers)

PCR testuje či sú v DNA blízko seba (stovky, tisíce báz)

Ak áno, namnoží úsek medzi nimi

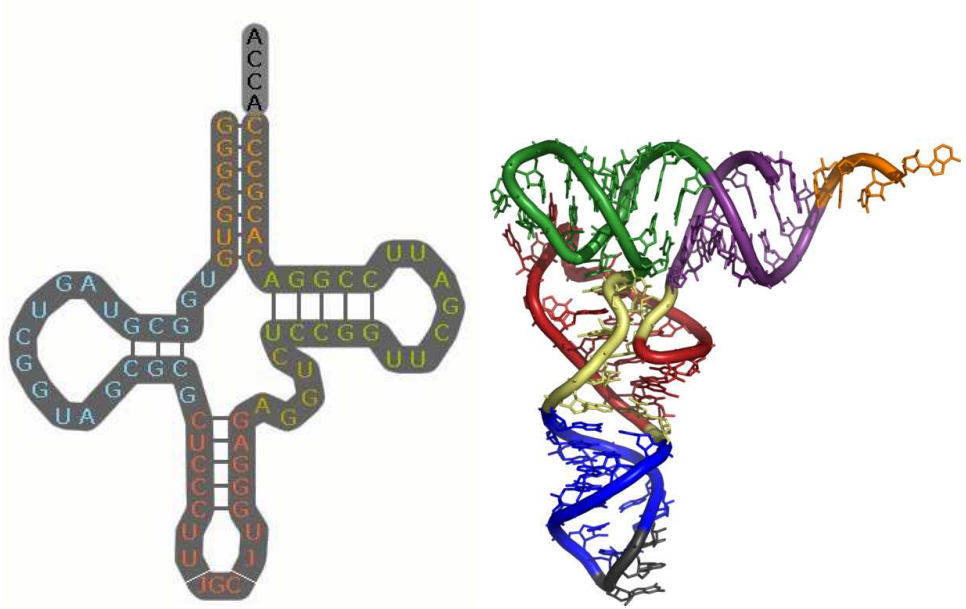



gén
DNA:
replikácia
transkripcia do RNA
RNA:
spracovanie RNA
regulácia
RNA:
translácia
do proteínu
RNA:
proteín:

RNA

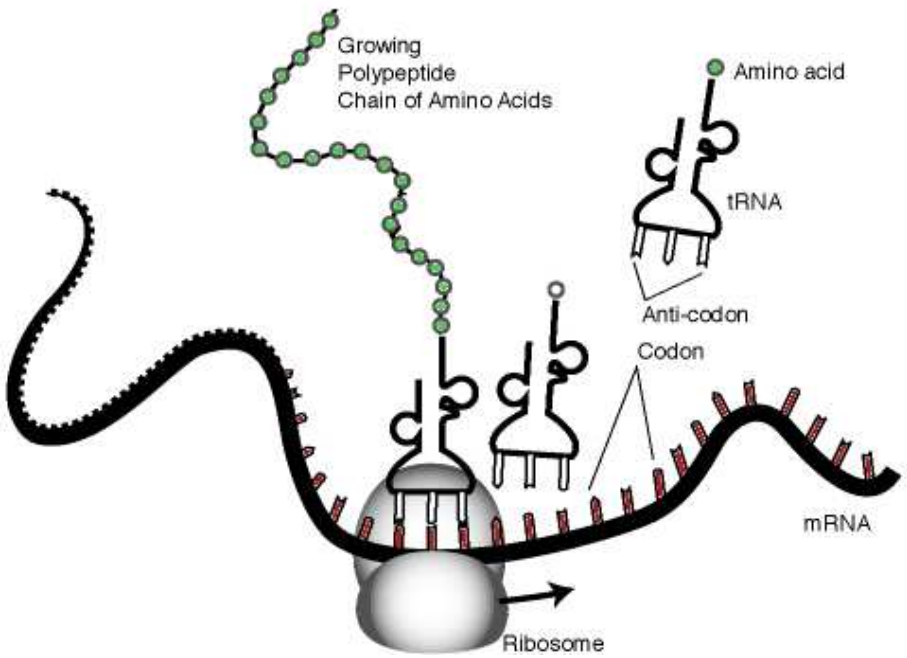
Ako sa líši od DNA?

- obsahuje ribózu namiesto deoxyribózy
- obsahuje uracil namiesto tymínu (bázy A,C,G,U)
- jednovláknové reťazce, zvyčajne kratšie
- zložitá sekundárna štruktúra: spárované komplementárne úseky

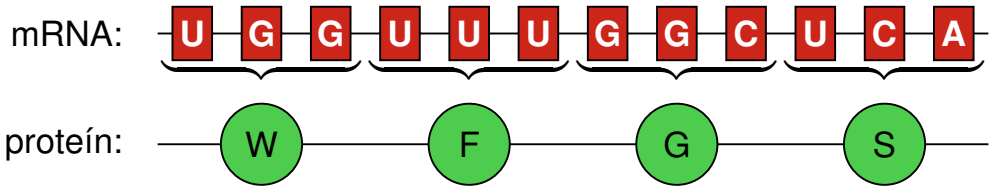


transferová RNA (tRNA)

Translácia



Kodón (trojica nukleotidov) určuje 1 aminokyselinu



Genetický kód

Alanine (A)
GC*

Cysteine (C)
TGC
TGT

Aspartic acid (D)
GAC
GAT

Glutamic acid (E)
GAA
GAG

Phenylalanine (F)
TTC
TTT

Glycine (G)
GG*

Histidine (H)
CAC
CAT

Isoleucine (I)
ATA
ATC
ATT

Lysine (K)
AAA
AAG

Leucine (L)
CT*
TTA
TTG

Methionine (M)
ATG

Asparagine (N)
AAC
AAT

Proline (P)
CC*

Glutamine (Q)
CAA
CAG

Arginine (R)
CG*
AGA
AGG

Serine (S)
TC*
AGT
AGC

Threonine (T)
AC*

Valine (V)
GT*

Tryptophan (W)
TGG

Tyrosine (Y)
TAC
TAT

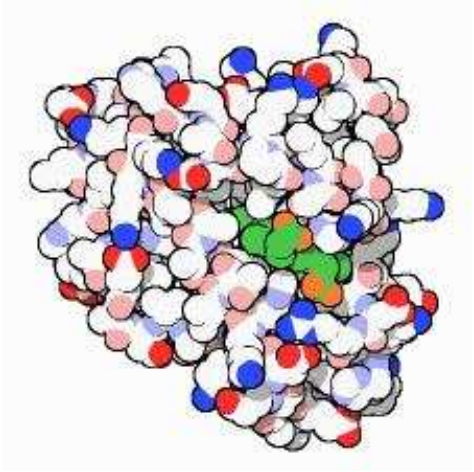
Stop codon (*)
TAA
TAG
TGA

Proteíny

Reťazce 20 rôznych aminokyselín s rôznymi chemickými vlastnosťami:

Amino Acid	Side chain	Hydrophobic	Polar	Charged	
Alanine (A)	-CH3	X	-	-	-
Arginine (R)	-(CH2)3NH-C(NH)NH2	-	X	basic	-
Asparagine (N)	-CH2CONH2	-	X	-	-
Aspartic acid (D)	-CH2COOH	-	X	acidic	-
Cysteine (C)	-CH2SH	X	-	acidic	-
Glutamic acid (E)	-CH2CH2COOH	-	X	acidic	-
Glutamine (Q)	-CH2CH2CONH2	-	X	-	-
Glycine (G)	-H	-	-	-	-
Histidine (H)	-CH2-C3H3N2	-	X	weak basic	Aromatic
Isoleucine (I)	-CH(CH3)CH2CH3	X	-	-	Aliphatic
Leucine (L)	-CH2CH(CH3)2	X	-	-	Aliphatic
Lysine (K)	-(CH2)4NH2	-	X	basic	-
Methionine (M)	-CH2CH2SCH3	X	-	-	-
Phenylalanine (F)	-CH2C6H5	X	-	-	Aromatic
Proline (P)	-CH2CH2CH2-	X	-	-	-
Serine (S)	-CH2OH	-	X	-	-
Threonine (T)	-CH(OH)CH3	-	X	weak acidic	-
Tryptophan (W)	-CH2C8H6N	X	-	-	Aromatic
Tyrosine (Y)	-CH2-C6H4OH	X	X	-	Aromatic
Valine (V)	-CH(CH3)2	X	-	-	Aliphatic

Štruktúra proteínov

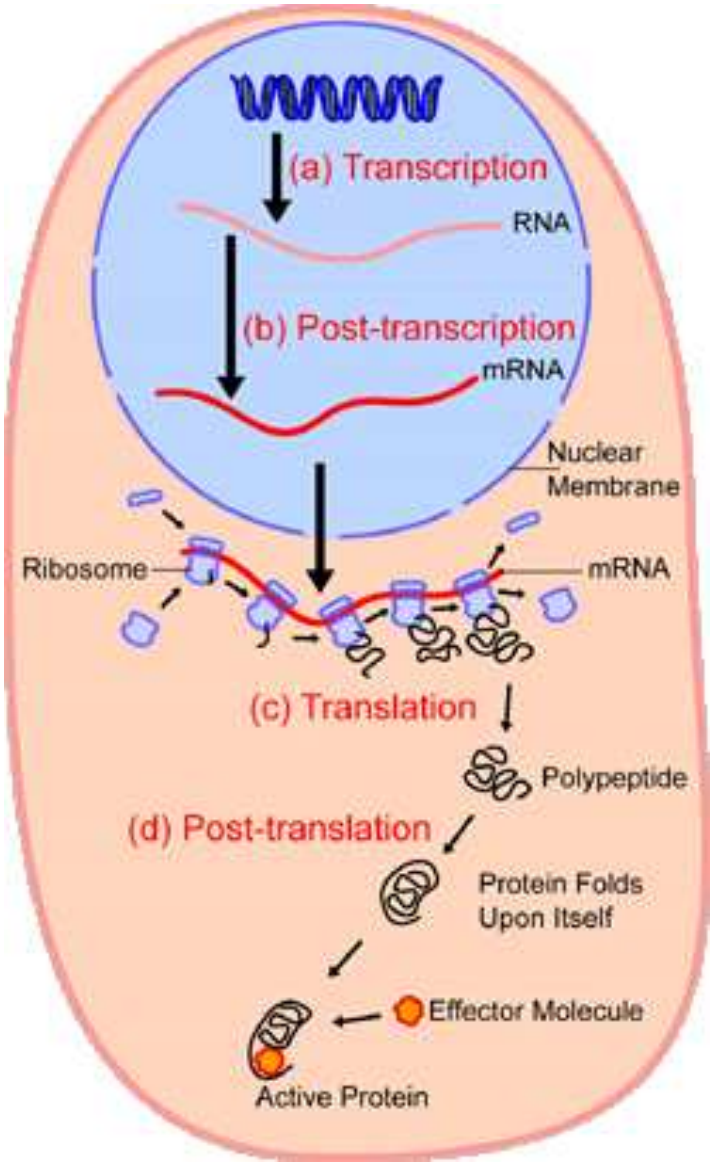


Myoglobín, prvý proteín so známou štruktúrou.

Proteíny sa vyskytujú poskladané v určitej stabilnej štruktúre, prípadne prechádzajú medzi niekoľkými stavmi.

Hydrofóbne aminokyseliny neinteragujú s vodou, zväčša sa vyskytujú vo vnútri štrukúry.

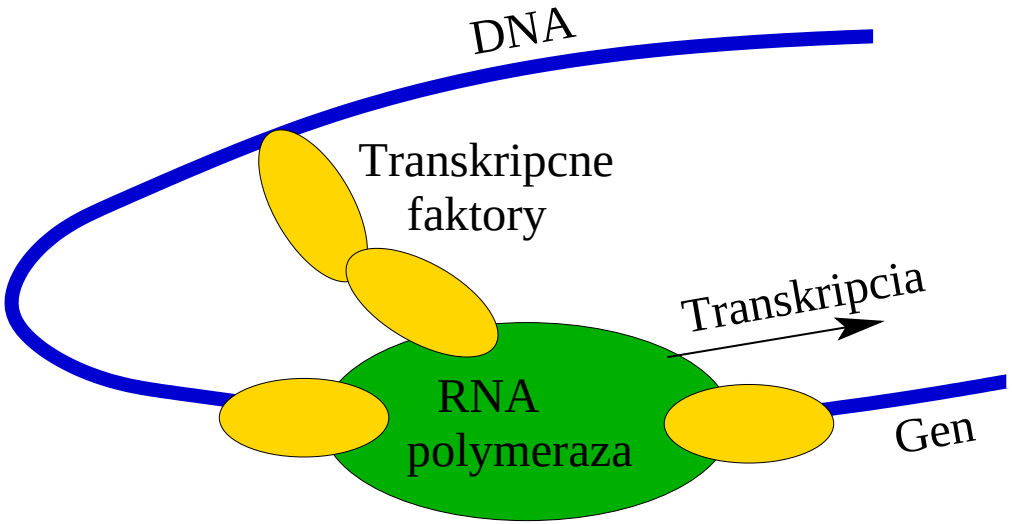
Štruktúra proteínu určuje jeho funkciu.


(a) Transcription
RNA
(b) Post-transcription
mRNA
Nuclear Membrane
Ribosome
mRNA
(c) Translation
Polypeptide
(d) Post-translation
Protein Folds Upon Itself
Effector Molecule
Active Protein

Regulácia expresie

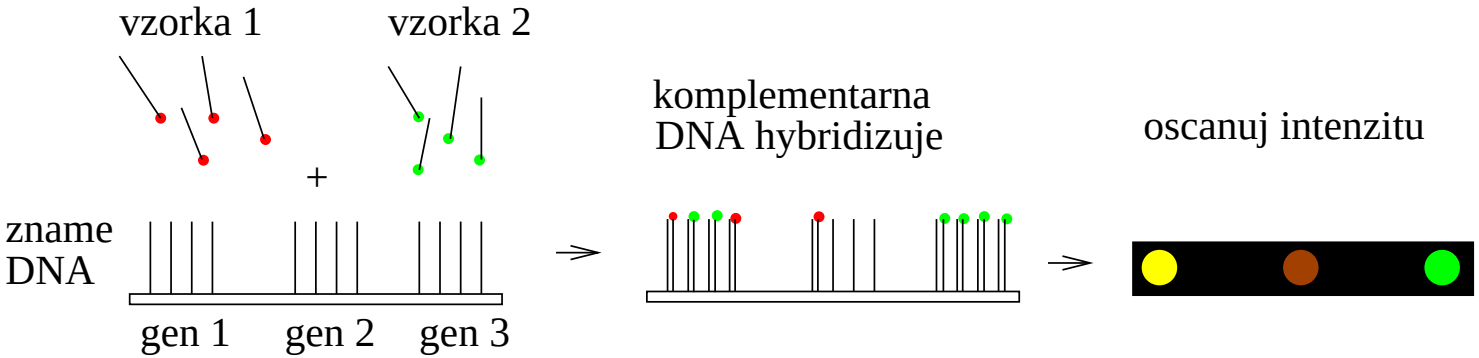
Bunky v rôznych tkanivách toho istého organizmu zdieľajú ten istý genóm, vyzerajú a fungujú však veľmi rôzne.
Niektoré proteíny sa tvoria len za určitých okolností, alebo v premenlivom množstve.

Regulácia začatia transkripcie pomocou transkripčných faktorov:



Bioinformatický problém: zisti, ktoré faktory ovplyvňujú ktorý gén, kde presne sa viažu.

Technológia: microarray

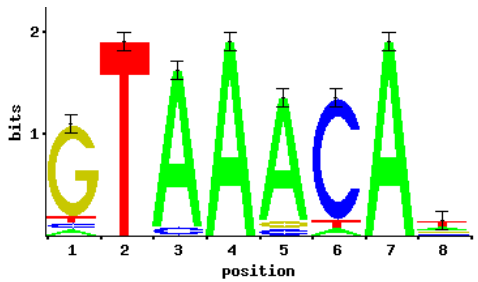


Meranie množstva mRNA prítomnej v bunke pre **veľa génov** naraz.

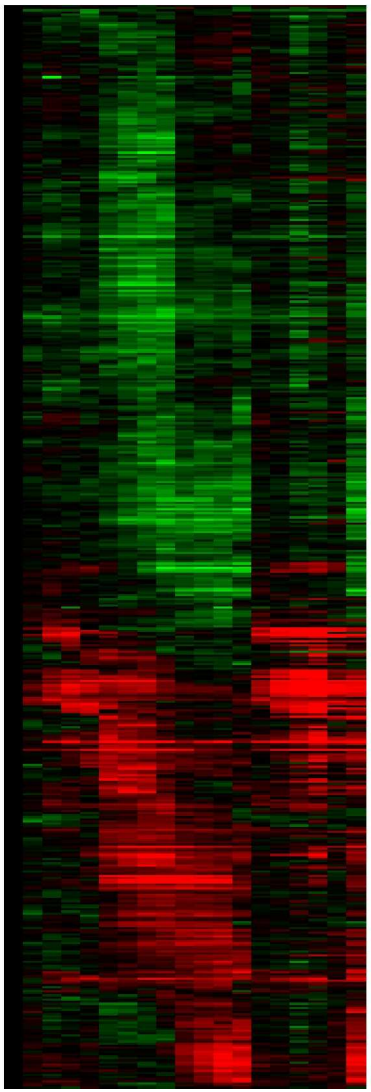
Zopakujeme za rôznych podmienok, študujeme korelácie medzi génmi.

Môžu byť dôsledkom spoločného regulátora (transkripčného faktoru).

Bioinformatický problém: niekoľko koregulovaných génov, nájdi motív, ku ktorému sa môže viazať spoločný transkripčný faktor **(motif finding)**



Príklad microarray dát



Mutácie DNA

V DNA občas dochádza k zmenám, mutáciám
(napr. pod vplyvom prostredia, či chybou pri replikácii).

Typy mutácií:

substitúcia, substitution (jedna báza sa zmení na inú),
inzercia, insertion (vloží sa niekoľko nových báz),
delécia, deletion (vynechá sa niekoľko báz),
zmeny väčšieho rozsahu (napr. translokácie).

Bioinformatické problémy:

Ktoré sekvencie vznikli z spoločného predka mutovaním?
(hľadanie homológov, homology search)
Ktoré bázy v dvoch príbuzných sekvenciách si navzájom zodpovedajú?
(sequence alignment, zarovnávanie sekvencií)

Populačná genetika

Mutácie sa šíria v populácii z rodičov na potomkov.
Nebezpečné mutácie rýchlejšie vymiznú, prospešné sa rýchlejšie ujmú (prírodný výber, natural selection).

Polymorfizmus: genetický rozdiel medzi organizmami v rámci druhu.
Vedie k rozdielnosti vo fenotype, napr. výzor, dedičné choroby.
Sekvenovaním viacerých jedincov toho istého druhu získame prehľad o polymorfizme.

Bioinformatický problém:
Izoluj polymorfizmus zodpovedný za určitý znak (napr. chorobu).

Evolúcia

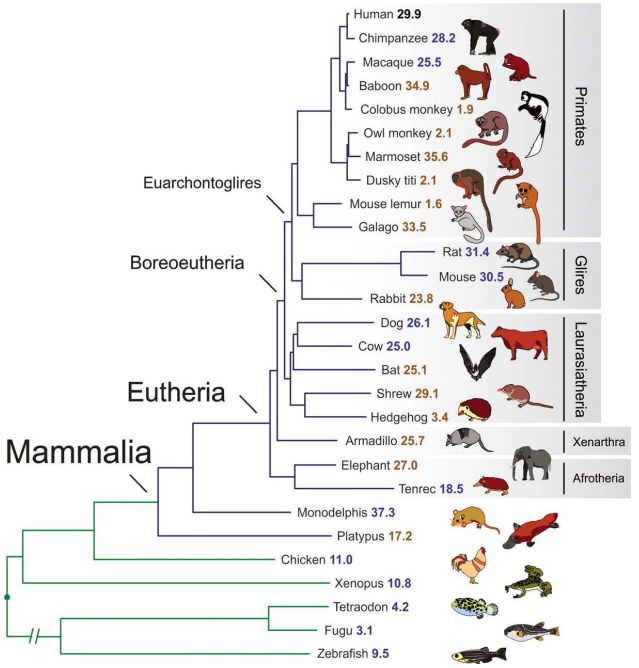
Vznik nových druhov (speciation):

Po rozdelení populácie na viacero oddelených častí nedochádza k výmene genetického materiálu.

Hromadia sa zmeny až kým nie je možné párenie: vznik nových druhov.

Bioinformatický problém:

Na základe dnešných sekvencií urči strom reprezentujúci vývoj druhov (fylogenetický strom, phylogenetic tree)



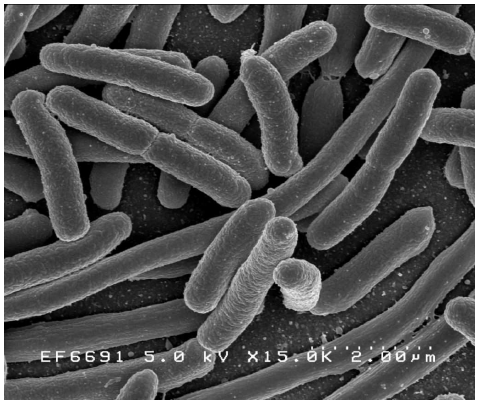
Prokaryotické vs. eukaryotické organizmy

Prokaryoty: baktérie, jednoduché jednobunkové organizmy.
Nemajú jadro (DNA priamo v cytoplazme),
majú kruhový chromozóm (a prípadné kratšie plasmidy),
jednoduchšia štruktúra génu atď.

Eukaryoty: živočíchy, rastliny, huby, niektoré jednobunkové organizmy.
Bunka obsahuje jadro s DNA, viacero organel.
Mitochondrie a chloroplasty sú pohltené prokaryoty, ktoré sa stali časťou eukaryotickej bunky.
Dlhší genóm v niekoľkých lineárnych chromozómoch.

Modelové organizmy

Dôležité pre biologický výskum, vieme o nich viac než o príbuzných druhoch. Poznatky širšie aplikovateľné.



Escherichia coli: baktéria žijúca v črevách. Jednoduchá manipulácia, delenie každých 20 min. Štúdium základných životných procesov: DNA replikácia, expresia génov, atď. Genóm s 4000 génmi, 4.6MB.

Saccharomyces cerevisiae: pekárske droždie. Jednoduchý eukaryotický organizmus. Genóm s 6000 génmi, 13MB. Delenie každé 2 hodiny. Štúdium špecificky eukaryotických javov.

Modelové organizmy

Arabidopsis thaliana: malá kvitnúca rastlina, 6-týždňový životný cyklus. Skúmanie javov špecifických pre rastliny.

Caenorhabditis elegans: malý červ, nematód, žijúci v pôde. Štúdium vývinu (ontogenéza, development), diferenciácie buniek.

Drosophila melanogaster: vínna muška. Štúdium genetiky, gény riadiace vývin jedinca.

Stavovce: žaba Xenopus laevis (veľké, ľahko manipuľovateľné vajíčka), akvarijná ryba Danio rerio (priehľadné embryá), myš Mus musculus (existuje veľa plemien so špeciálnymi vlastnosťami).

Dostupné dáta

- DNA sekvencie: celé genómy, ich časti
- Ich anotácia: súradnice génov a iných funkčných častí
- Sekvencie RNA, ich štruktúra
- Sekvencie proteínov, ich funkcia a štruktúra
- Merania množstva RNA/proteínu v bunke
- ...

Dáta založené na experimetoch alebo výsledky výpočtových metód
Veľa chýb (v oboch prípadoch)

Úvod do dynamického programovania, proteomika

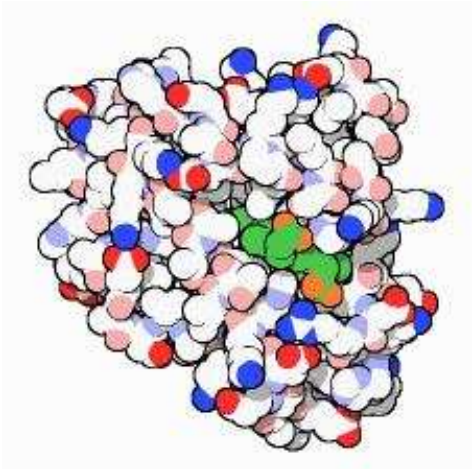
Tomáš Vinař

2.10.2014

Proteomika

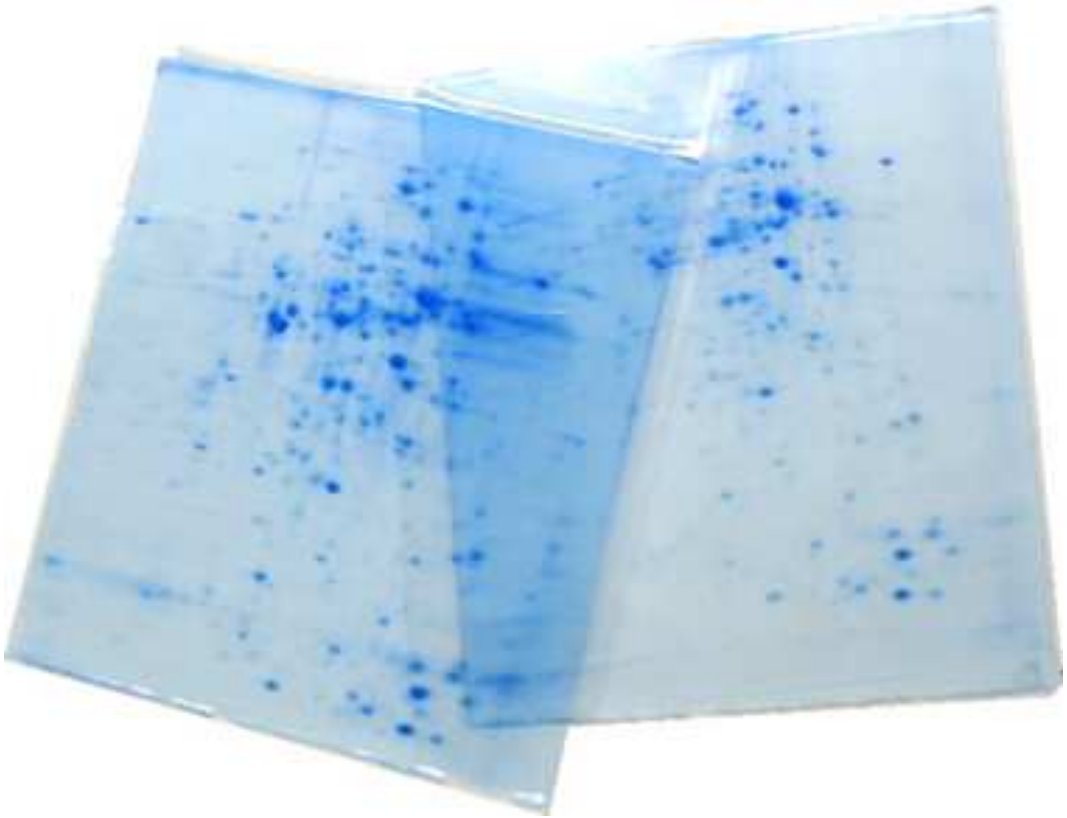
Proteín: sekvencia pozostáva z 20 rôznych aminokyselín

```
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
```



Z bunky sme izolovali určitý proteín, chceme zistiť jeho sekvenciu.

Proteomika: Gélová elektroforéza (gel electrophoresis)



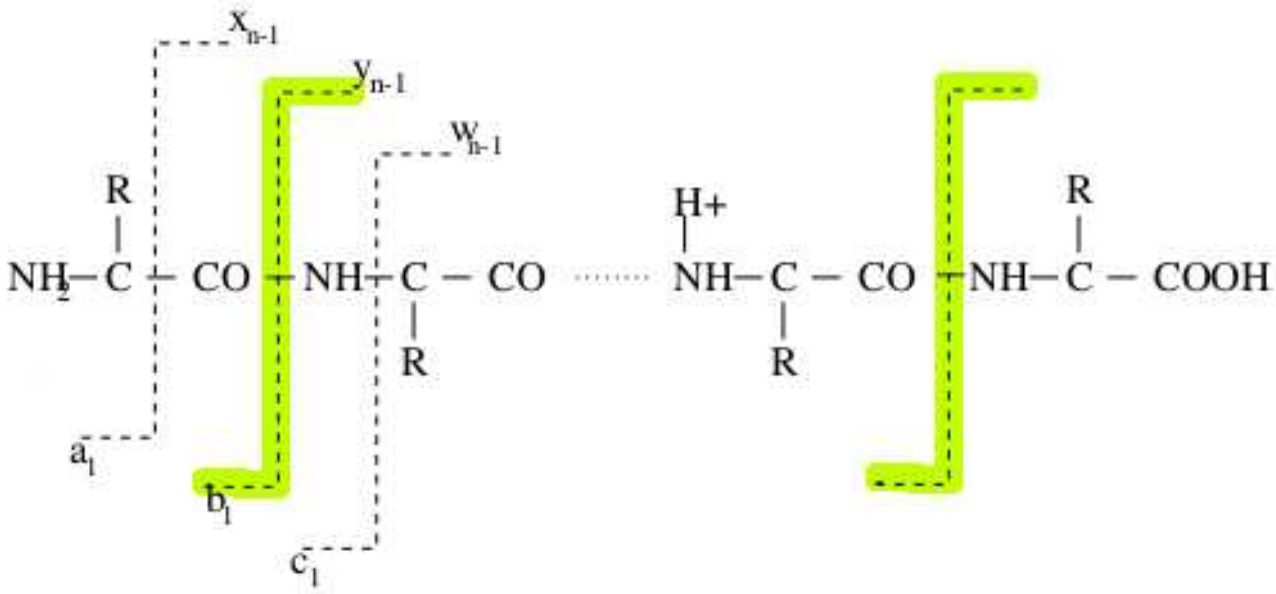
Alternatíva: kvapalinová chromatografia (liquid chromatography)

Hmotnostná spektrometria (mass spectometry)

- Meria pomer hmostnosť/náboj molekúl vo vzorke
- Používa sa na identifikáciu proteínov, napr. z 2D gélu.
- Proteín nasekáme enzýmom trypsín (seká na `[KR]{P}`) na peptidy
- Meriame hmostnosť kúskov, porovnáme s databázou proteínov.
- Tandemová hmotnostná spektrometria (MS/MS) ďalej fragmentuje každý kúsok a dosiahne podrobnejšie spektrum, ktoré obsahuje viac informácie
- V niektorých prípadoch tak vieme sekvenciu proteínu určiť priamo z MS/MS, bez databázy proteínov

Tandemová hmotnostná spektrometria MS/MS

Štiepenie peptidu na prefixy a sufixy



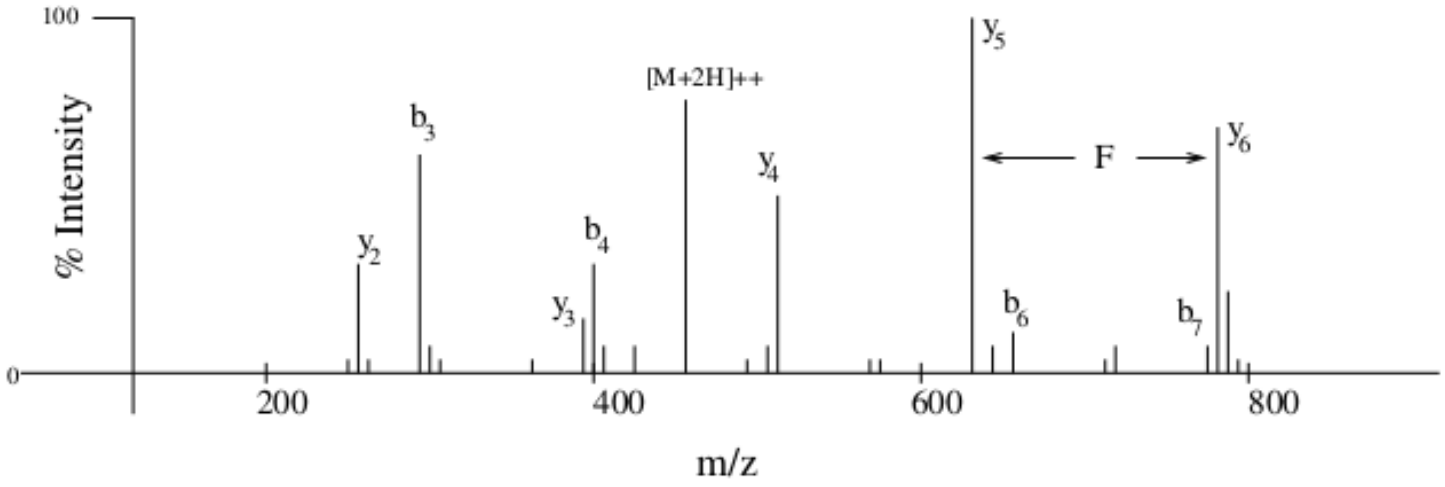
zdroj: Bafna and Reinert

b-ióny: prefixy

y-ióny: sufixy

Tandemová hmotnostná spektrometria MS/MS

88	145	292	405	534	663	778	924	b-ions
S	G	F	L	E	E	D	K	
924	837	780	633	520	391	262	141	y-ions



zdroj: Bafna and Reinert

Sekvenovanie peptidov pomocou MS/MS

Pre začiatok berieme do úvahy iba b-ióny (prefixy)

Všetky hmostnosti celé čísla

Vstup:

celková hmotnosť peptidu M,

hmotnosti aminokyselín $a[1], \ldots, a[20]$,

spektrum ako tabuľka $f[0], \ldots, f[M]$, ktorá hmotnosti b-iónu určí skóre podľa signálu v okolí príslušného bodu grafu

Pre postupnosť aminokyselín $p_1 \ldots p_k$ nech $m(p_1 \ldots p_k)$ je jej hmotnosť, t.j. $m(p_1 \ldots p_k) = \sum_{j=1}^{k} a[p_j]$

Výstup:

postupnosť aminokyselín $p_1 \ldots p_k$, taká, že

$m(p_1 \ldots p_k) = M$ a

$\sum_{i=1}^{k} f[m(p_1 \ldots p_i)]$ je maximálna možná

Sekvenovanie peptidov pomocou MS/MS

Pre peptid $p = p_1, \ldots, p_k$ uvažujme množinu hmotností $\mathcal{M}(p)$, ktorá obsahuje hmostnosti všetkých jeho prefixov a sufixov

$$\mathcal{M}(p) = \{m(p_1 \ldots p_i) \mid i = 1 \ldots k\} \cup \{m(p_i \ldots p_k) \mid i = 1 \ldots k\}$$

Cieľ: maximalizujeme $\sum_{m \in \mathcal{M}(p)} f[m]$
alebo maximalizujeme $\sum_{m \in \mathcal{M}(p), m \leq M/2} f'[m]$
kde $f'[m] = f[m] + f[M - m]$
(pre $m = M/2$ vezmeme $f'[m] = f[m]$)

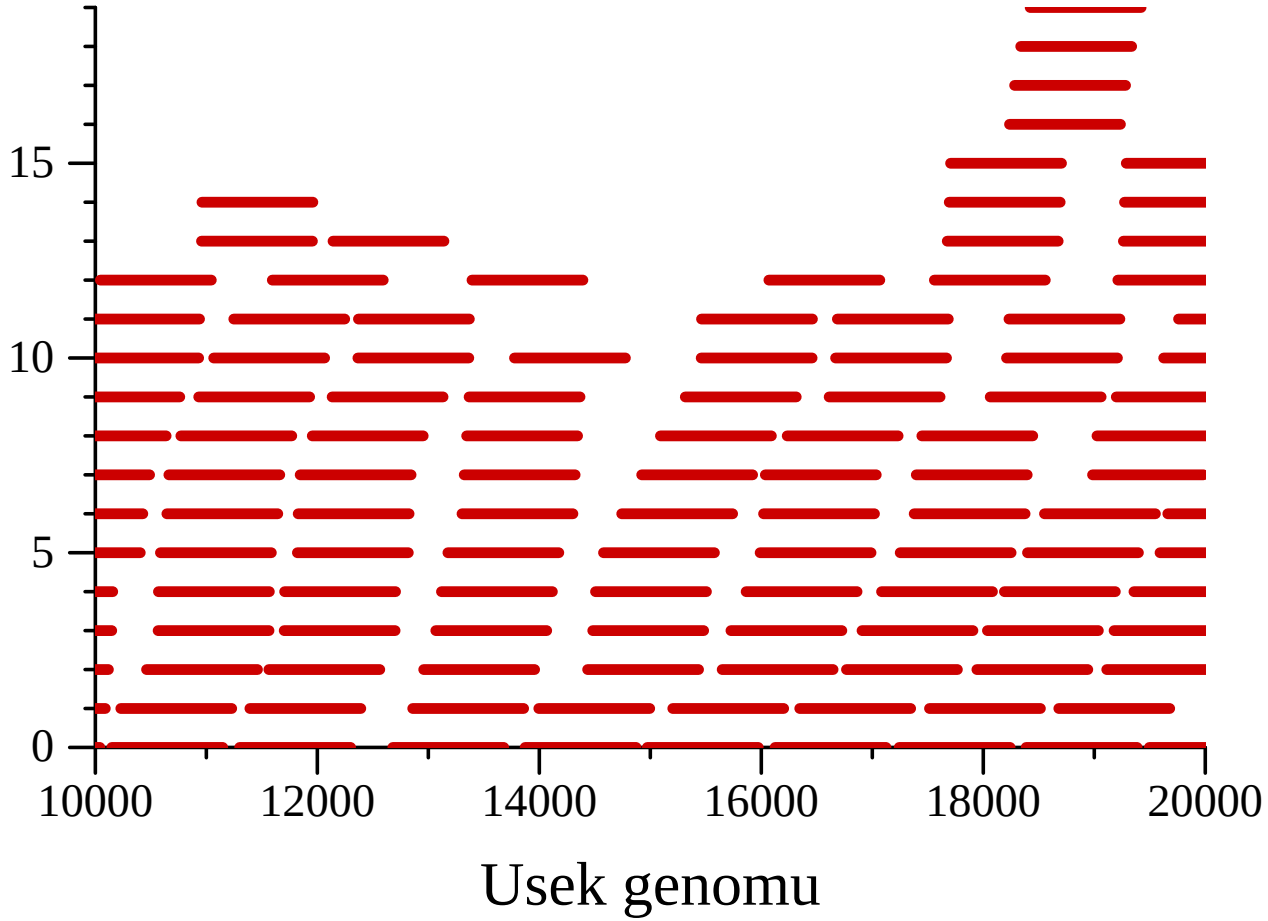
Úvod do pravdepodobnosti, sekvenovanie genómov (cvičenie)

Tomáš Vinař

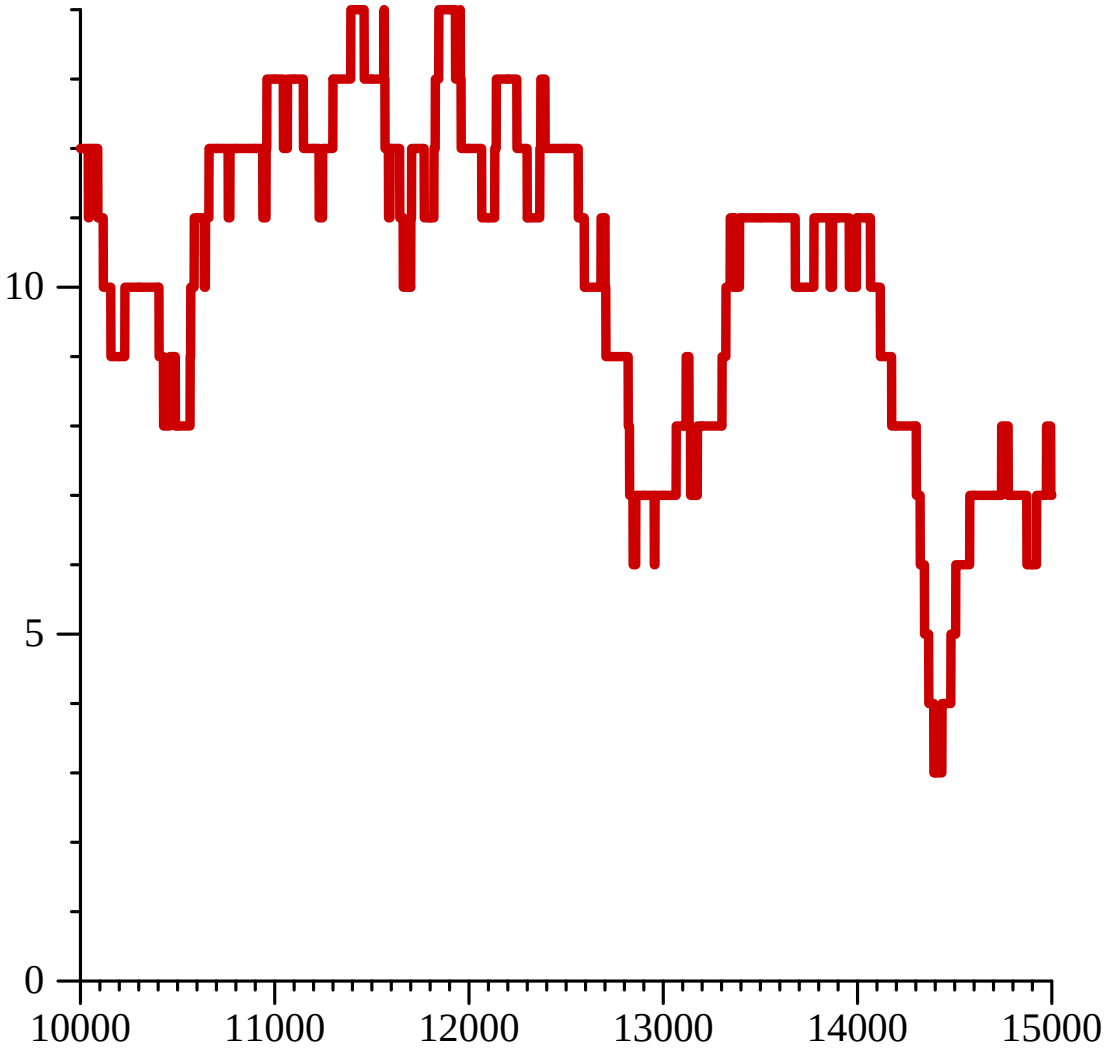
23.10.2014

- $G =$ dĺžka genómu, napr. 1 000 000
- $N =$ počet segmentov (readov), napr. 10 000
- $L =$ dĺžka segmentu, napr. 1000
- $T =$ potrebná dĺžka prekryvu, napr. 50

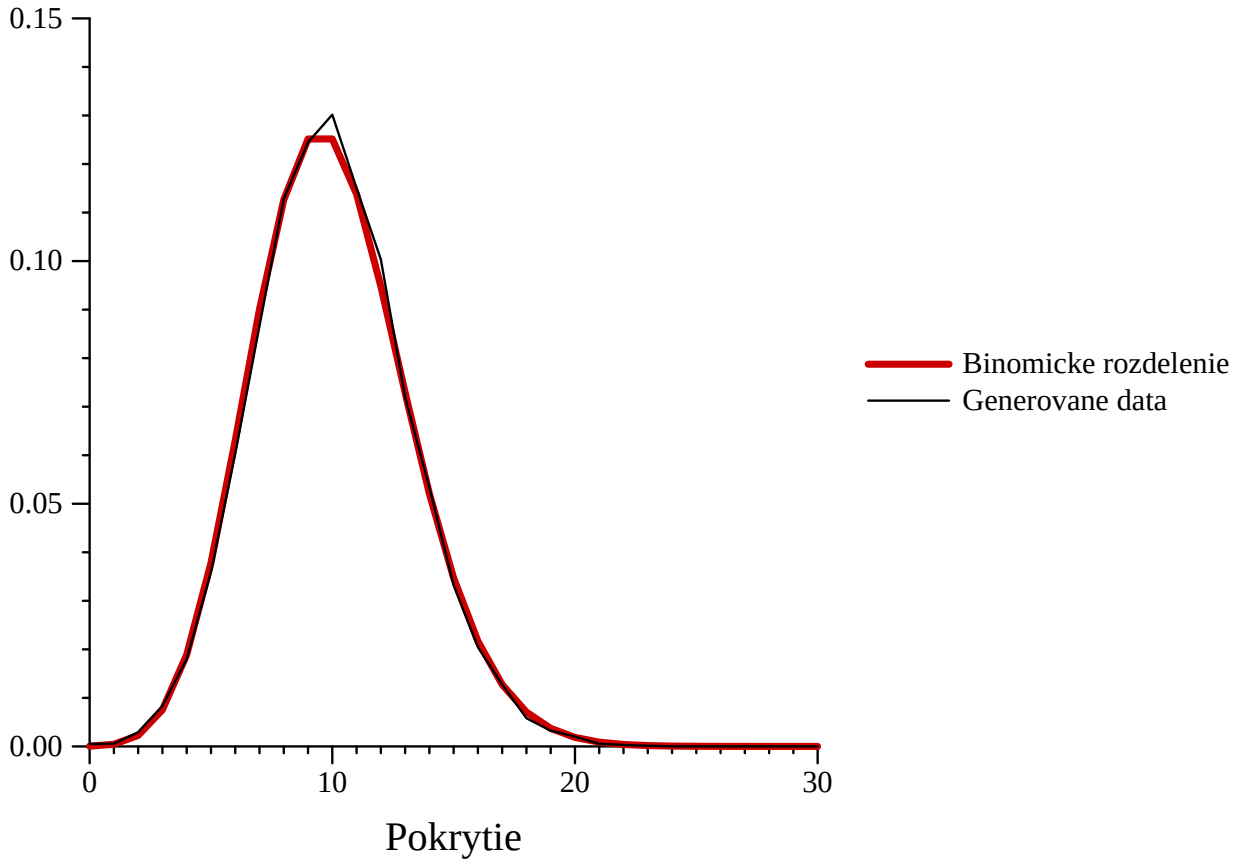
Náhodne generované segmenty



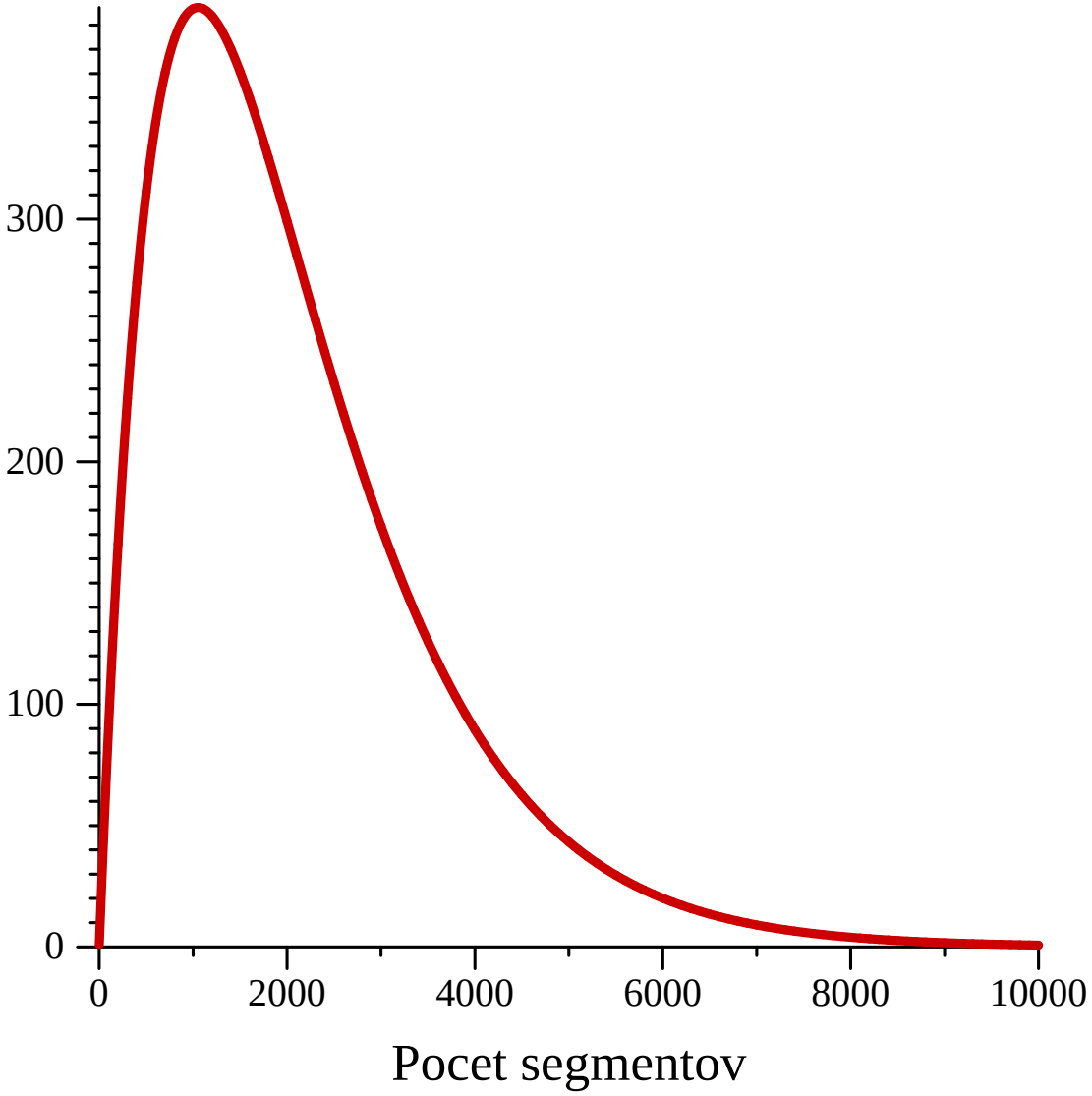
Pokrytie jednotlivých báz



Počet báz s určitým pokrytím



Predpokladaný počet kontigov od počtu segmentov



```
nepokr: 0 koncov: 0          nepokr: 0 koncov: 0    nepokr: 0 koncov:
nepokr: 274 koncov: 2     nepokr: 282 koncov: 1    nepokr: 0 koncov: 0
nepokr: 0 koncov: 0     nepokr: 0 koncov: 0    nepokr: 8 koncov: 1
nepokr: 0 koncov: 0     nepokr: 12 koncov: 1    nepokr: 0 koncov: 0
nepokr: 122 koncov: 1     nepokr: 135 koncov: 1    nepokr: 111 koncov: 1
nepokr: 13 koncov: 1     nepokr: 1 koncov: 1    nepokr: 56 koncov: 1
nepokr: 265 koncov: 1     nepokr: 0 koncov: 0    nepokr: 10 koncov: 1
nepokr: 0 koncov: 0     nepokr: 0 koncov: 0    nepokr: 130 koncov: 1
nepokr: 217 koncov: 1     nepokr: 3 koncov: 1    nepokr: 0 koncov: 0
nepokr: 0 koncov: 0     nepokr: 0 koncov: 0    nepokr: 86 koncov: 1
nepokr: 139 koncov: 2     nepokr: 0 koncov: 0    nepokr: 0 koncov: 0
nepokr: 76 koncov: 1     nepokr: 221 koncov: 1    nepokr: 26 koncov: 1
nepokr: 0 koncov: 0     nepokr: 1 koncov: 1    nepokr: 0 koncov: 0
nepokr: 0 koncov: 0     nepokr: 0 koncov: 0    nepokr: 0 koncov: 0
nepokr: 0 koncov: 0     nepokr: 0 koncov: 0    nepokr: 12 koncov: 1
nepokr: 103 koncov: 2     nepokr: 0 koncov: 0    nepokr: 71 koncov: 1
nepokr: 69 koncov: 1     nepokr: 0 koncov: 0
```

Jadrá zarovnaní

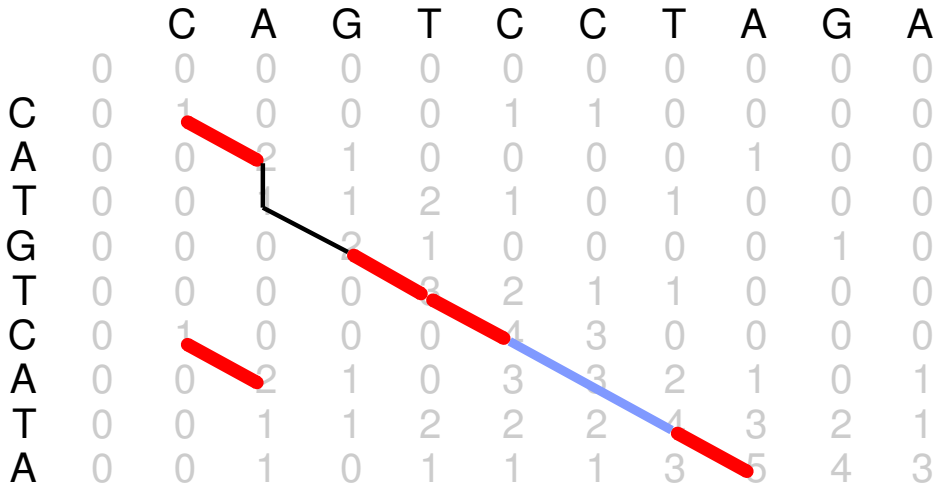
Tomáš Vinař

30.10.2014

Opakovanie: Heuristické lokálne zarovnávanie, BLAST

Príklad: $w = 2$ (začíname z jadier dĺžky 2).

(V praxi sa používa $w = 10$ a viac.)



1. nájdi zhodné úseky
2. rozšír bez medzier
3. spoj medzerami

Senzitivita heuristického algoritmu

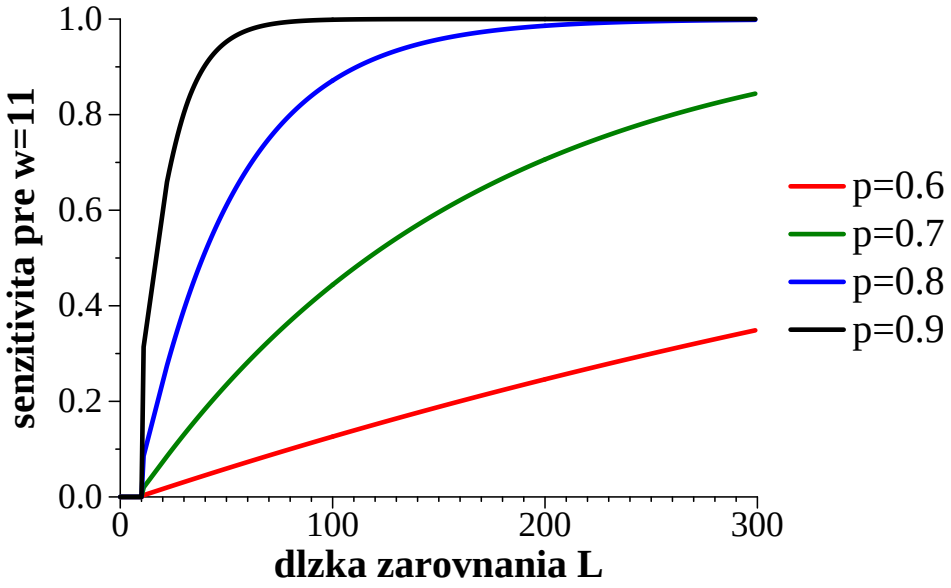
Odhad senzitivity:

Predpokladáme zarovnanie bez medzier, dĺžky L

Každá pozícia je zhoda s pravdepodobnosťou p

Senzitivita:

$$f(L,p) = \Pr(\text{zarovnanie obsahuje } w \text{ zhôd za sebou})$$



Jadrá z medzerami, spaced seeds

PatternHunter [Ma, Tromp, Li 2002]

Jadro s medzerami: vyžadovaná konfigurácia zhôd

Príklad:

"match—match—don't care—match" značíme ako `1101`

```
GTGGTGCTCTCTGACAAAGCC
 |  ||  ||    |||  ||||
ATTGTTCTTAATGAGAAAGAA
   1101      1101
                1101
```

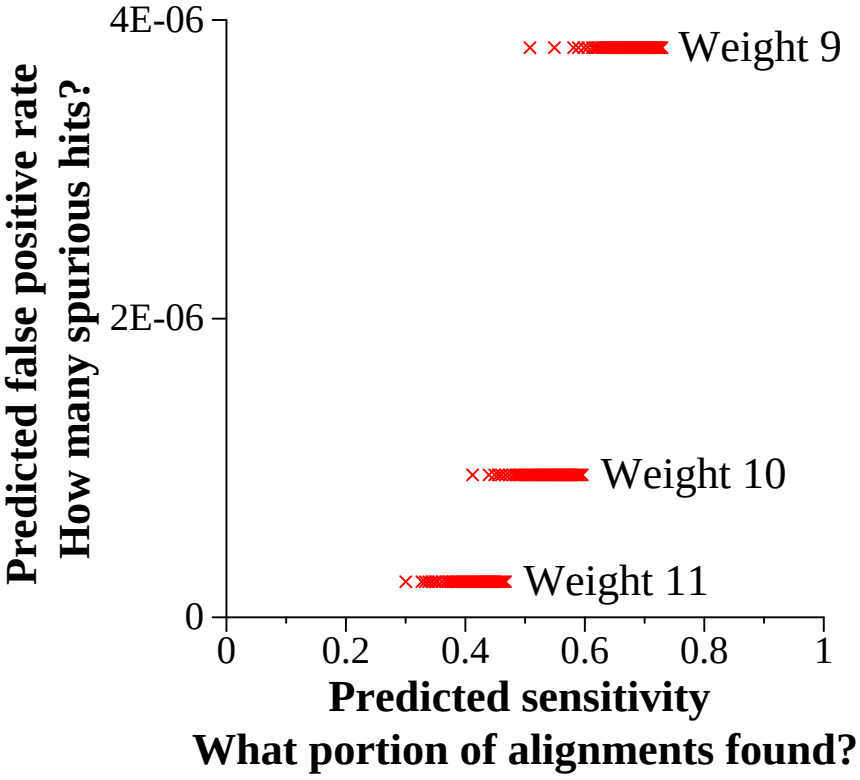
BLASTN jadro (11 za sebou idúcich zhôd)

ekvivalentné jadru `11111111111`

Nie všetky jadrá sú rovnaké

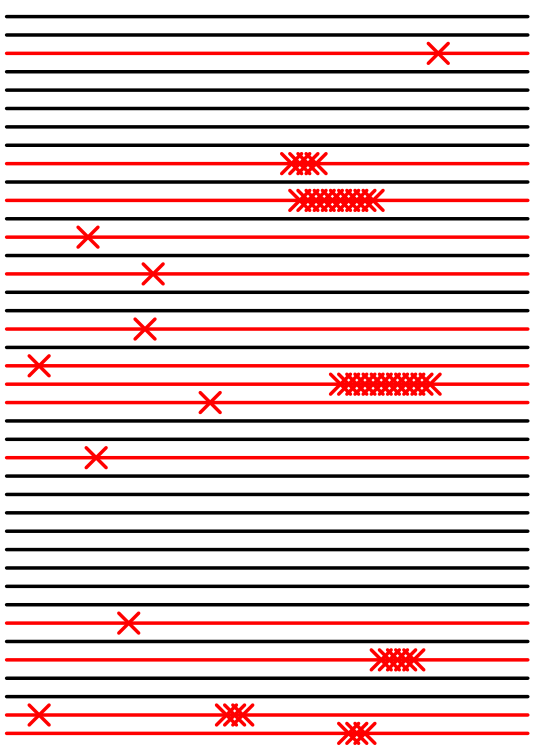
Váha jadra: počet vyžadovaných zhôd

Každý krížik: senzitivita vs. čas pre jedno jadro v pravdepodobnostnom modeli



Prečo sú jadrá s medzerami senzitívnejšie?

Príklad: dĺžka zarovnania $L = 64$,
pravdepodobnosť zhody $p = 0.7$ na každej pozícii
40 náhodných zarovnaní, výskyty jadra



11111111111
Sn.: 14/40, hits: 46

11101001010011011 1
Sn.: 18/40, hits: 35

Prečo sú jadrá s medzerami senzitívnejšie?

Príklad: dĺžka zarovnania $L = 64$,
pravdepodobnosť zhody $p = 0.7$ na každej pozícii

Bez medzier	**S medzerami**
11111111111	111010010100110111
Stredná hodnota počtu výskytov v zarovnaní:	
$54 \cdot 0.7^{11} = 1.1$	$47 \cdot 0.7^{11} = 0.9$
Pravdepodobnosť výskytu na poz. $i+1$ ak výskyt na i:	
0.7	$0.7^6 = 0.12$
1111111111 1111111111	111010010100110111 111010010100110111
Výskyty často vedľa seba	Výskyty "nezávislejšie"
Senzitivita (pravdepodobnosť aspoň jedného výskytu):	
0.30	0.47

Ďalšie hašovacie stratégie

Nukleotidový BLAST: 10 zhôd za sebou

Jadro s medzerami: povoľuje nezhody na 8 z 18 pozícií

BLAT [Kent 2002]: povoľuje 1 nezhodu na ľub. z 11 pozícií

BLASTP: 3 amino kyseliny so skóre aspoň 13 v matici BLOSUM62

Výskyt:

```
N I R
N L R

6+2+5=13
```

Nie výskyt:

```
A I L
A I L

4+4+4=12
```

Vektorové jadrá: kombinácia jadier s medzerami a BLAT/BLASTP

Viaceré výskyty: začni rozširovať iba ak viac výskytov blízko seba na tej istej uhlopriečke

Viaceré jadrá: zober zjednotenie výskytov

Záleží na modeli zarovnania

Pravdepodobnosť zhody kolíše v rámci kodónu:

Poloha v kodóne:	prvá	druhá	tretia
Pravdepodobnosť zhody:	0.67	0.77	0.40

Senzitivita na testovacej vzorke exónov kódujúcich proteíny:

Jadro		Človek vs. Drosophila	Človek vs. myš
Optimálne pre dáta	**110 110 000 110 110 11**	86%	92%
Optimalne pre kodónový model	**110 110 010 110 010 11**	86%	91%
WABA [Kent, Zahler 2000]	**110 110 110 110 11**	80%	90%
Optimálne pre i.i.d. model	**1110010010010101 11**	60%	86%
BLAST	**1111111111**	43%	81%
Najhoršie	**1010101010101011**	39%	79%

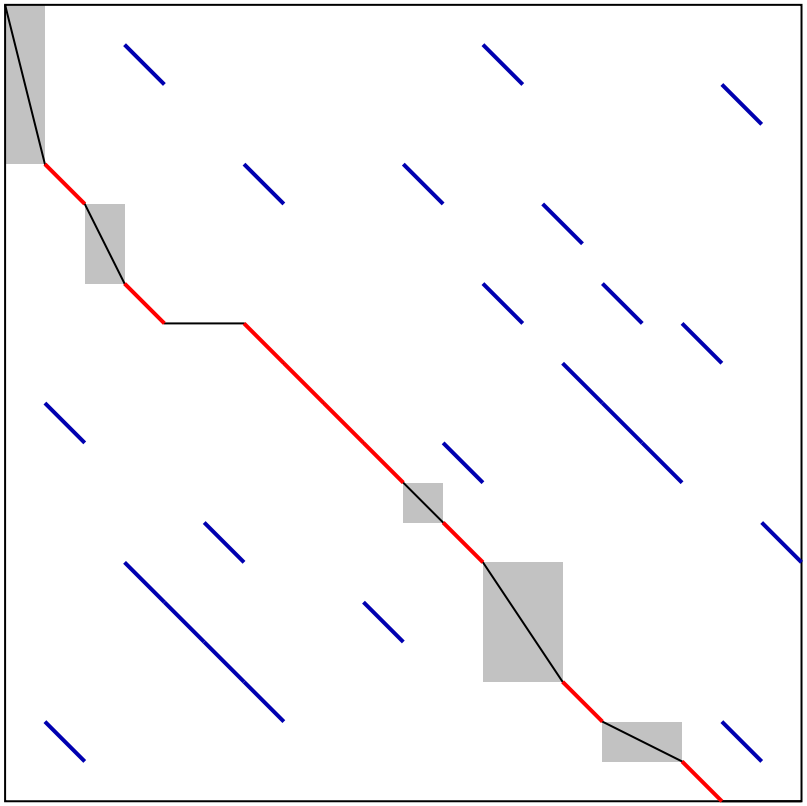
A čo globálne zarovnania?

Ukotvené zarovnanie (Anchored alignment)

- Nájdime lokálne zarovnania (alebo výskyty nejakého jadra) – možné **ukotvenia**
- Zvoľ konzistentnú množinu ukotvení (monotónna postupnosť)
- Zarovnaj časti sekvencií medzi ukotveniami (pomocou dyn. prog. alebo rekurzívne ďalším kotvením)

MUMMER [Delcher 1999]
GLASS [Batzoglou et al 2000]
AVID [Bray et al 2003]

Ukotvené zarovnanie



Modré: nezvolené ukotvenia

Červené: zvolené ukotvenia

Sivé: riešime dyn. prog.

Čierne: globálne zarovnanie

Znova protichodné vplyvy:

málo spoľahlivých ukotvení – dobrá kvalita, pomalé

veľa slabších ukotvení – rýchle (malá sivá plocha), viac chýb v ukotvení

Algoritmy pre HMM a phyloHMM

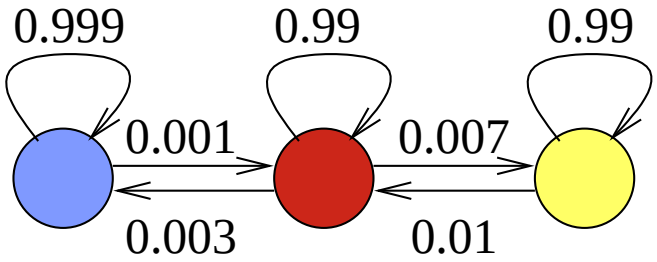
Tomáš Vinař

20.11.2014

Opakovanie: HMM (skrytý Markovov model)

→ náhodná DNA sekvencia S, náhodná anotácia A
(podobné na ozajstnú DNA)

$\Pr(S, A)$ – pravdepodobnosť, že model vygeneruje pár (S, A).



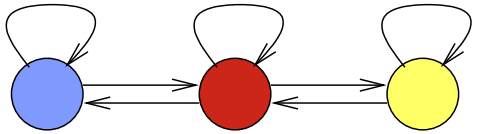
a:0.27	a:0.24	a:0.26
c:0.23	c:0.27	c:0.22
g:0.23	g:0.28	g:0.22
t:0.27	t:0.21	t:0.30

Predpokladajme, že model vždy začína v modrom stave.

$\Pr(\text{aca}) = 0.27 \cdot 0.001 \cdot 0.27 \cdot 0.99 \cdot 0.24 = 0.000017$

$\Pr(\text{aca}) = 0.27 \cdot 0.999 \cdot 0.23 \cdot 0.999 \cdot 0.27 = 0.017$

Parametre HMM (označenie)



Sekvencia $S_1, \ldots, S_n$

Anotácia $A_1, \ldots, A_n$

Parametre modelu:

Prechodová pravdepodobnosť $a(u, v) = \Pr(A_{i+1} = v | A_i = u)$,

Emisná pravdepodobnosť $e(u, x) = \Pr(S_i = x | A_i = u)$,

Počiatočná pravdepodobnosť $\pi(u) = \Pr(A_1 = u)$.

a			
	0.99	0.007	0.003
	0.01	0.99	0
	0.001	0	0.999

e	a	c	g	t
	0.24	0.27	0.28	0.21
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

Výsledná pravdepodobnosť: $\Pr(A_1, \ldots, A_n, S_1, \ldots, S_n) = \pi(A_1)e(A_1, S_1) \prod_{i=2}^{n} a(A_{i-1}, A_i)e(A_i, S_i)$

Viterbiho algoritmus

Pre danú sekvenciu S nájde najpravdepodobnejšiu anotáciu
$A = \arg\max_A \Pr(A|S)$

Dynamické programovanie v čase $O(nm^2)$

Podproblém $V[i,u]$: pravdepodobnosť najpravdepodobnejšej cesty končiacej po i krokoch v stave u, pričom vygeneruje $S_1 S_2 \ldots S_i$

Rekurencia:

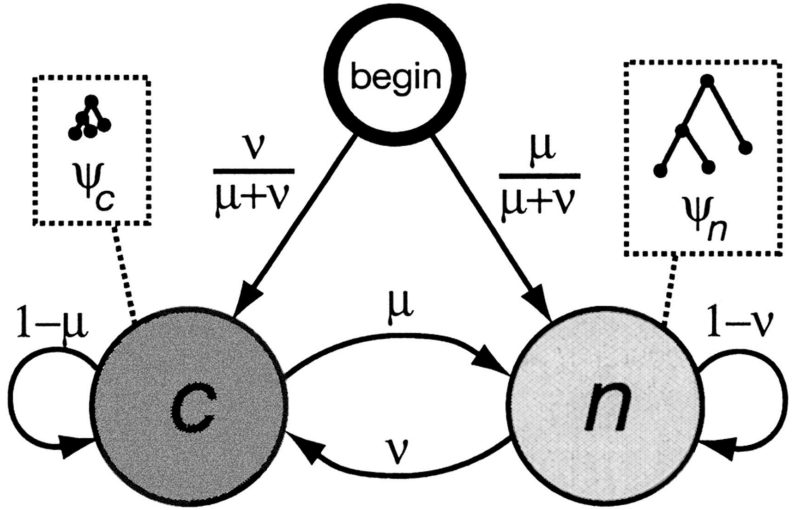
$V[1,u] = \pi_u \cdot e_{u,S_1}$
$V[i,u] = \max_w V[i-1,w] \cdot a_{w,u} \cdot e_{u,S_i}$

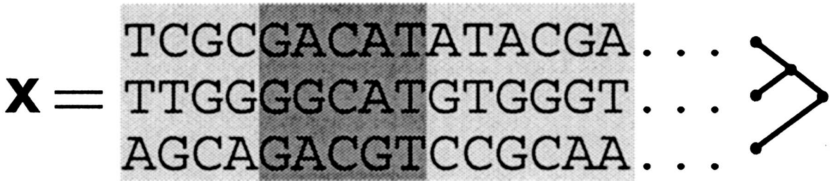
Algoritmus:

Inicializuj $V[1,*]$
for $i = 2 \ldots n$ (n=dĺžka reťazca)
    for $u = 1 \ldots m$ (m =počet stavov)
        vypočítaj $V[i,u]$
Maximálne $V[n,j]$ je pravdepodobnosť najpravdepodobnejšej cesty

PhyloHMM: kombinácia HMM a fylogenetického stromu

PhastCons: detekcia dobre zachovaných sekvencií





- Dva stavy: zachovaná sekv., neutrálna sekv.
- V každom stave generujeme celý stĺpec zarovnania
- Zachovaná sekvencia má kratšie hrany stromu

Dopredný algoritmus

Počíta celkovú pravdepodobnosť, že vygenerujeme sekvenciu S
$\Pr(S) = \sum_A Pr(A, S)$

Podproblém $F[i, u]$**:** pravdepodobnosť, že po i krokoch vygenerujeme $S_1, S_2, \ldots S_i$ a dostaneme sa do stavu u.

$F[i, u] = \Pr(A_i = u \wedge S_1, S_2, \ldots, S_i) =$
$\sum_{A_1, A_2, \ldots, A_i = u} \Pr(A_1, A_2, ..., A_i \wedge S_1, S_2, ..., S_i)$

Rekurencia:

$F[1, u] = \pi_u \cdot e_{u,S_1}$
$F[i, u] = \sum_v F[i-1, v] \cdot a_{v,u} \cdot e_{u,S_i}$

Celková pravdepodobnosť $\Pr(S) = \sum_u F[n, u]$

Spätný algoritmus

Obdoba dopredného algoritmu

Dopredný algoritmus: $F[i,u] = \Pr(A_i = u \wedge S_1, \ldots, S_i)$

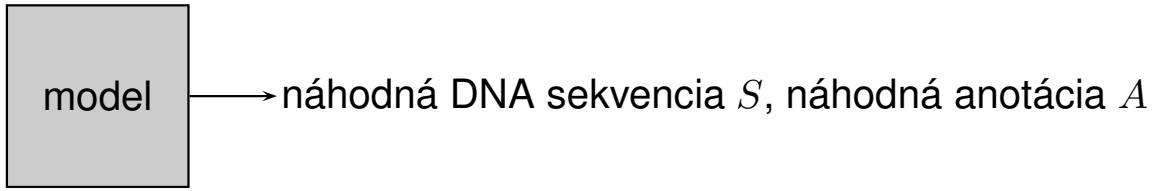
$F[1,u] = \pi_u \cdot e_{u,S_1}$

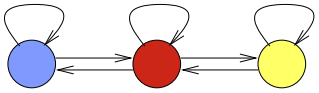
$F[i,u] = \sum_v F[i-1,v] \cdot a_{v,u} \cdot e_{u,S_i}$

$\Pr(S) = \sum_u F[n,u]$

Spätný algoritmus: $B[i,u] = \Pr(S_{i+1} \ldots, S_n | A_i = u)$

Hľadanie génov s HMM



- **Určenie stavov a prechodov v modeli:** ručne, na základe poznatkov o štruktúre génu. 
- **Trénovanie parametrov:** pravdepodobnosti určíme na základe sekvencií so známymi génmi (**trénovacia množina**).
  Model zostavíme tak, aby páry (S, A) s vlastnosťami podobnými skutočným génom mali veľkú pravdepodobnosť $\Pr(S, A)$
- **Použitie:** pre novú sekvenciu S nájdi najpravdepodobnejšiu anotáciu $A = \arg\max_A \Pr(A|S)$ Viterbiho algoritmom

Trénovanie HMM

- Stavový priestor + povolené prechody väčšinou ručne
- Parametre (pravdepodobnosti prechodu, emisie a počiatočné) automaticky z trénovacích sekvencií
- Čím zložitejší model a viac parametrov máme, tým potrebujeme viac trénovacích dát, aby nedošlo k preučeniu, t.j. k situácii, keď model dobre zodpovedá nejakým zvláštnostiam trénovacích dát, nie však ďalším dátam.
- Presnosť modelu testujeme na zvláštnych testovacích dátach, ktoré sme nepoužili na trénovanie.

Trénovanie HMM z anotovaných sekvencií

Vstup: topológia modelu a niekoľko trénovacích párov $S^{(i)}, A^{(i)}$

Cieľ: nastaviť π_u, $e_{u,x}$, $a_{u,v}$ tak, aby $\prod_i \Pr(S^{(i)}, A^{(i)})$ bola čo najväčšia

Dosiahneme jednoduchým počítaním frekvencií

Napr. $a_{u,v}$: nájdeme všetky výskyty stavu u a zistíme, ako často za nimi ide stav v

Trénovanie HMM z neanotovaných sekvencií

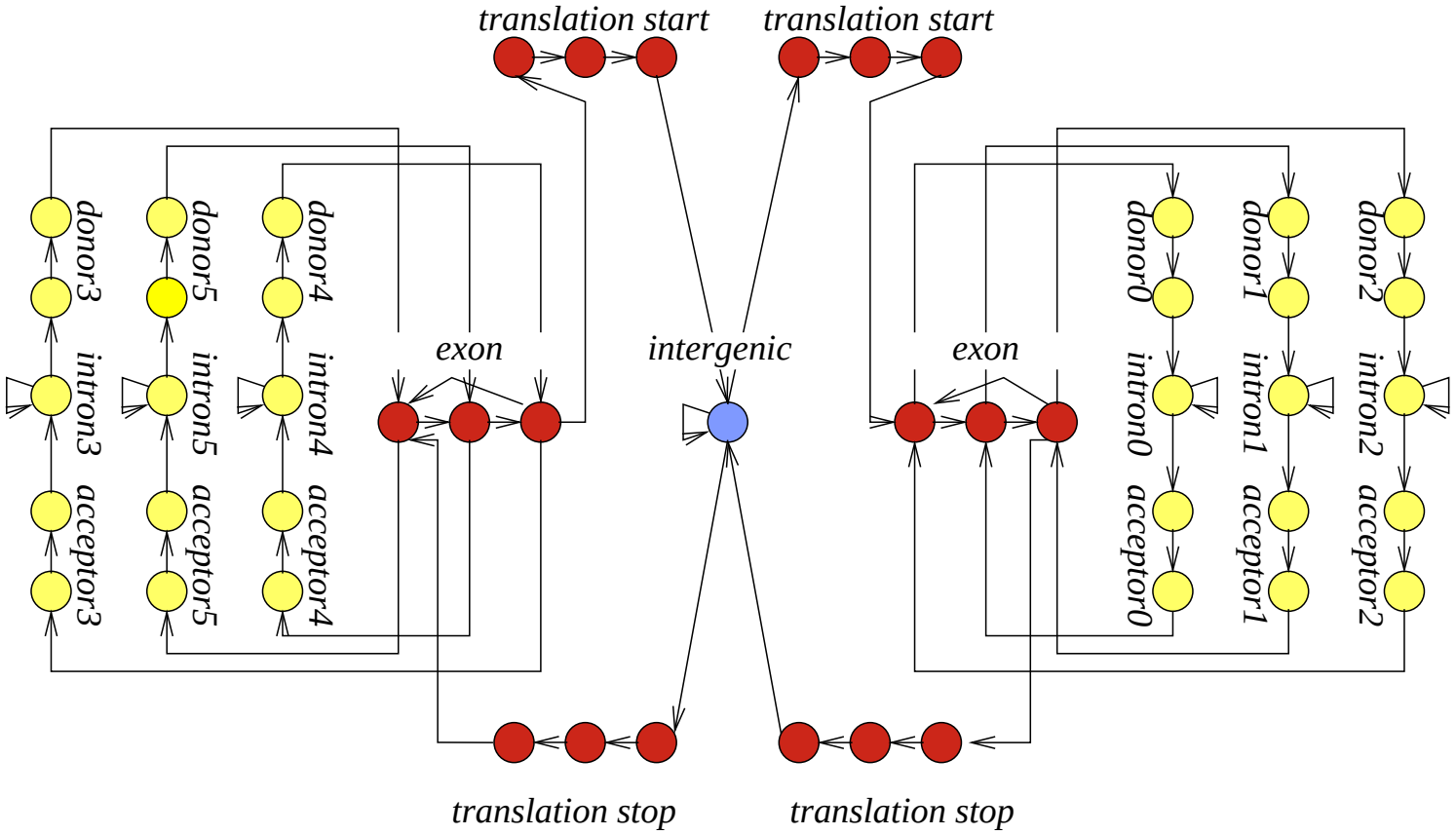
Vstup: topológia modelu a niekoľko trénovacích sekvencií $S^{(i)}$ anotácie $A^{(i)}$ nepoznáme

Cieľ: nastaviť π_u, $e_{u,x}$, $a_{u,v}$ tak, aby $\prod_i \Pr(S^{(i)})$ bola čo najväčšia

Používajú sa heuristické iteratívne algoritmy, napr. Baum-Welchov, ktorý je verziou všeobecnejšieho algoritmu EM (expectation maximization).

Tvorba stavového priestoru modelu

Príklad HMM na hľadanie génov



Cvičenia pre informatikov, 11.12.2014
Zhrnutie semestra

Pravdepodobnostné modely

- Skryté Markovove modely (hľadanie génov, konzervovaných oblastí, fylogenetické HMM, profilové HMM)
- Fylogenetické stromy a substitučné modely
- Stochastické bezkontextové gramatiky
- Metóda maximálnej vierohodnosti
- Expectation maximization (EM)

Štatistické metódy

- Pojem štatistickej významnosti
- E-value a P-value
- Test na pozitívny výber
- Linkage disequilibrium, mapovanie asociácií

Precvičenie dynamického programovania

- Zarovnávanie sekvencií
  (globálne, lokálne, afínne medzery)
- Skryté Markovove modely (Viterbiho algoritmus)
- Výpočty na stromoch
  (úspornosť, vierohodnosť - Felsensteinov algoritmus)
- Hmotnostná spektrometria (MS/MS)
- Sekundárna štruktúra RNA

Iné

- Integer linear programming
- deBruijnove grafy
- Zhlukovanie a klasifikácia

Ako modelovať problémy reálneho sveta

- Rozmyslieť si, aké máme dáta, čo by sme chceli ako výsledok
- Sformulovať ako informatický problém (napr. optimalizácia nejakého skóre)
- Pravdepodobnostné modely nám často dovolia zvoliť skórovaciu schému systematickým spôsobom
- Výsledný problém často NP ťažký
  - Heuristiky, aproximačne algoritmy
  - ILP a iné techniky na presné riešenie
  - Nedá sa problém trochu preformulovať?
- Testovanie: sú výpočtové výsledky relevantné v danej doméne? (bola formulácia dostatočne realistická?)

Ďalšie predmety

- **Strojové učenie** 2-INF-150, Vinař/Petrovič (ZS, 4 hodiny, 6 kreditov)
- **Grafové modely v strojovom učení** 2-INF-238, Vinař (LS, 4 hodiny, 6 kreditov)
- **Vybrané partie z dátových štruktúr** 2-INF-237, Brejová (LS, 4 hodiny, 6 kreditov) predtým Vyhľadávanie v texte
- **Biológia** N-bCXX-055, Tomáška (ZS, 2 hodiny, 2 kredity)
- **Všeobecná biológia** N-bCXX-085, Tomáška (LS, 2 hodiny, 2 kredity)
- **Genomika** N-mCBI-303, Nosek (LS, 2 hodiny, 3 kredity)
- http://compbio.fmph.uniba.sk/vyuka/