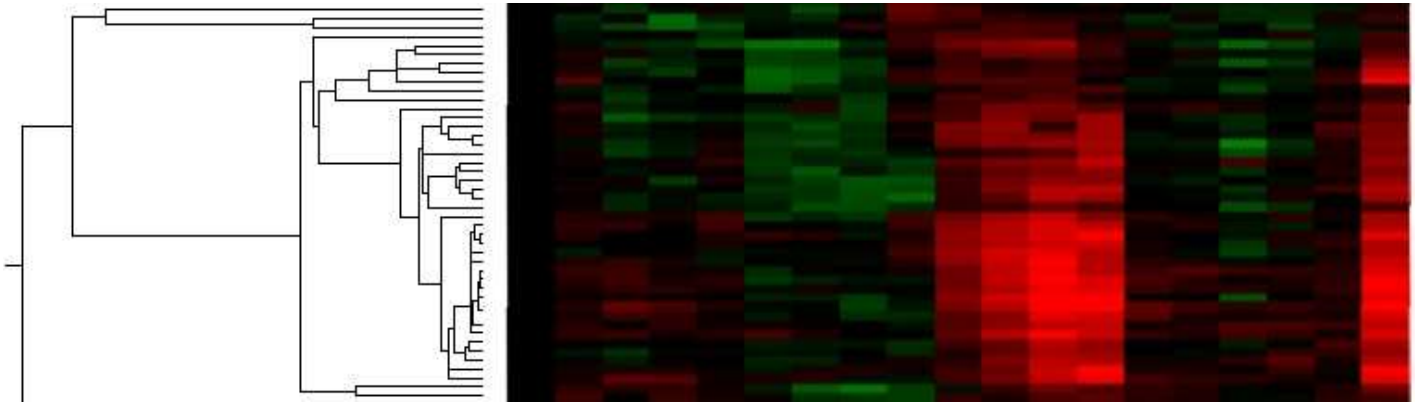
Oznamy

- DÚ 2 na stránke, odovzdať do 6.12.
  (na začiatku cvičení, resp. prednášky)
  Informatici zdrojový kód odovzdávajú v systéme Moodle
- Ďakujem za správy o stretnutiach journal clubu,
  dajte vedieť, ak potrebujete konzultácie

Regulácia génovej expresie

Broňa Brejová

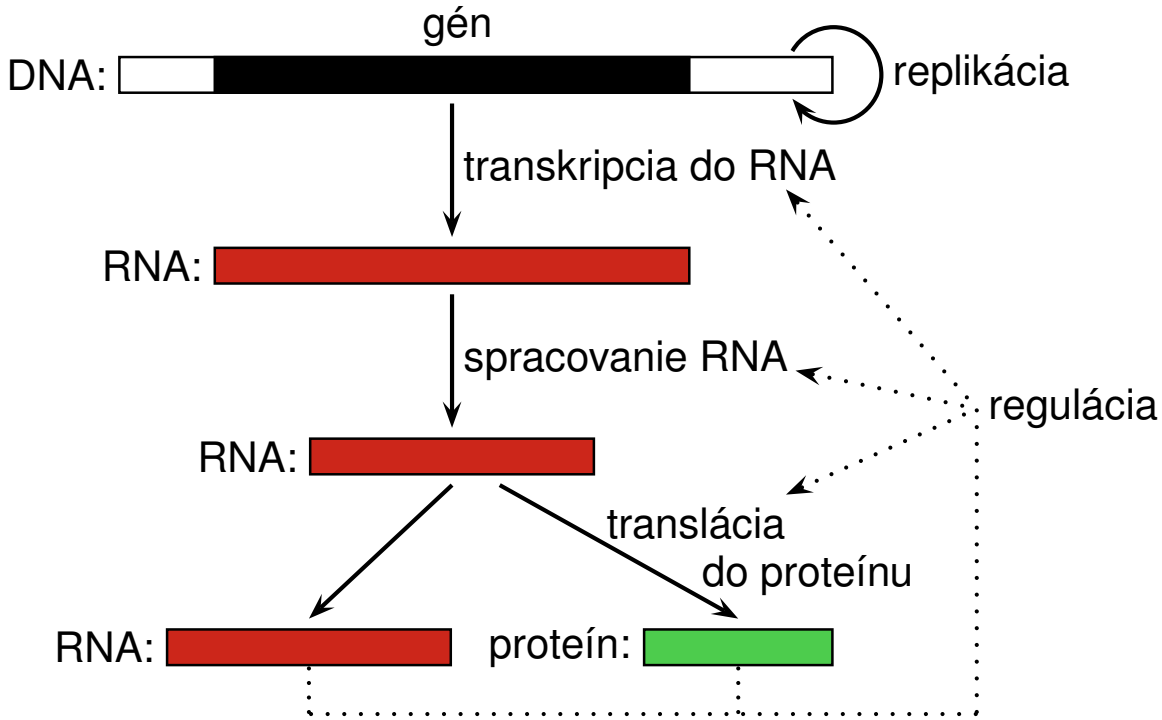
22.11.2018



Aká informácia je uložená v DNA?

Gény: Predpisy na tvorbu proteínov a funkčných RNA molekúl.

Riadenie ich expresie: kedy a koľko sa má tvoriť.

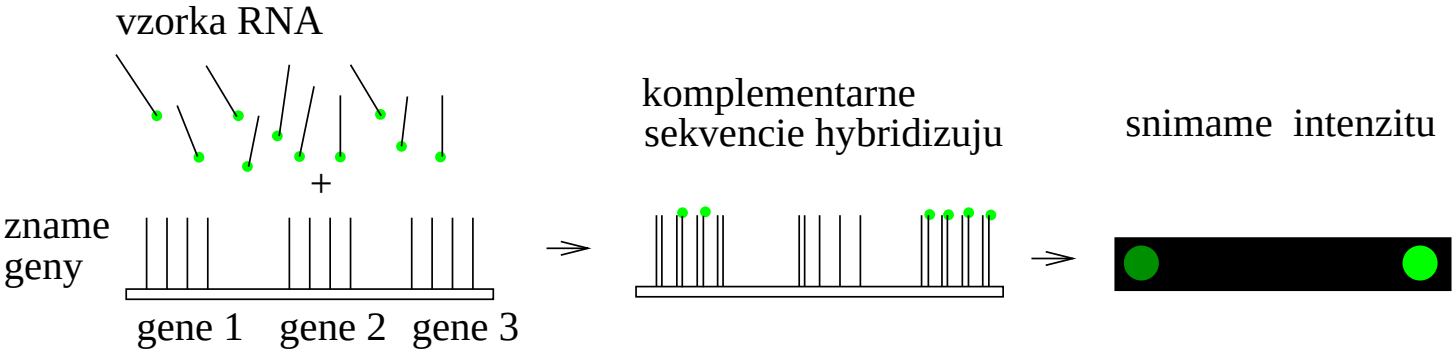


Regulácia na úrovni transkripcie, spracovania, translácie, posttranslačných modifikácií, . . .

Ciele

- Zistiť, za akých podmienok je daný gén exprimovaný
  (súvisí s funkciou génu)
- Ktoré gény ho regulujú
- Detaily regulačného mechanizmu
  (väzobné miesta, zmeny v množstve expresie, . . .)

Technológia: expression array, microarray

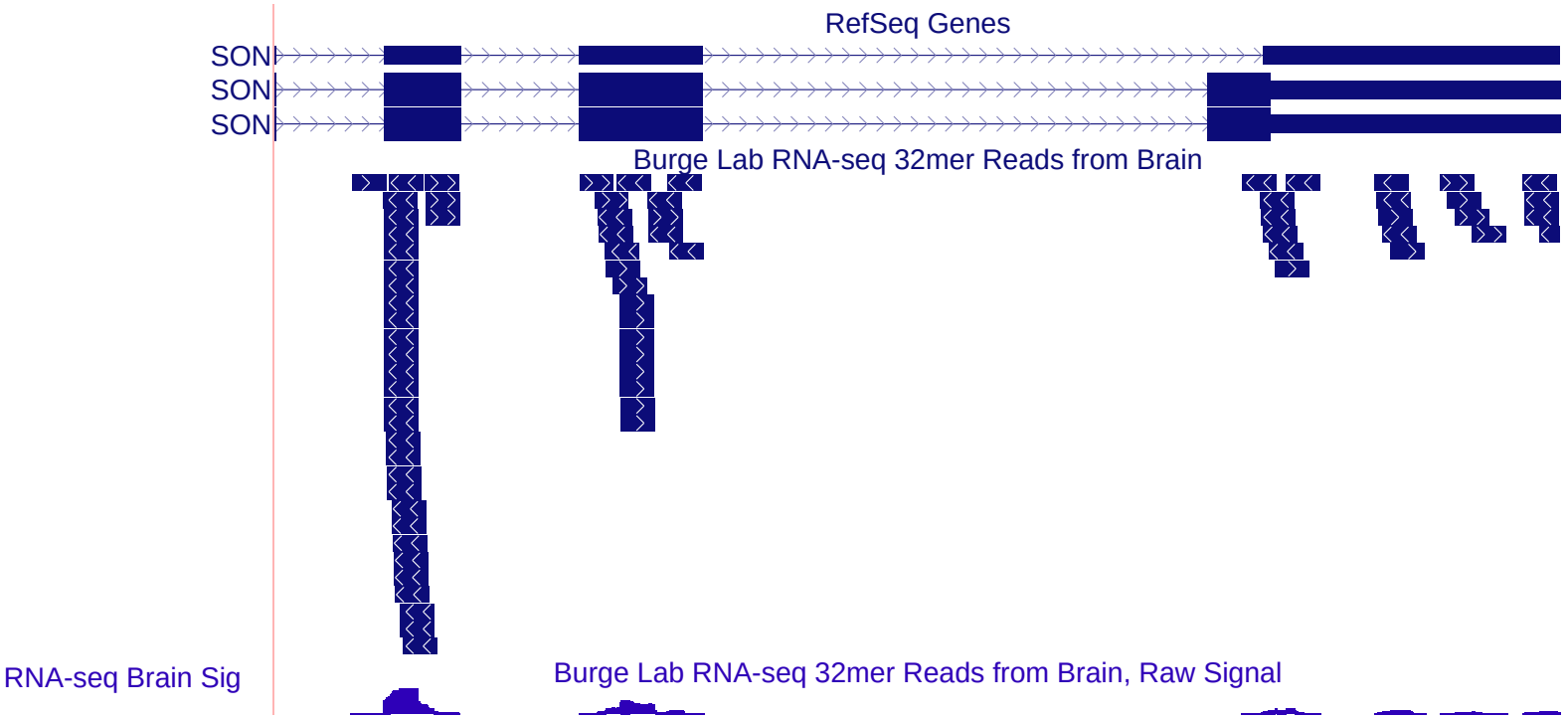


Meranie množstva mRNA prítomnej v bunke pre **veľa génov** naraz.

Zopakujeme za rôznych podmienok.

Technológia: RNA-seq

sekvenujeme RNA extrahovanú z bunky NGS technológiami,
mapujeme na genóm, hĺbka pokrytia zodpovedá úrovni expresie



Príklad microarray dát

Pomer expresie génu v meranej a kontrolnej vzorke fg/bg

	15min	30min	1hod	2hod	4hod	...
W95909	0.72	0.1	0.57	1.08	0.66	
AA045003	1.58	1.05	1.15	1.22	0.54	
AA044605	1.1	0.97	1	0.9	0.67	
W88572	0.97	1	0.85	0.84	0.72	
AA029909	1.21	1.29	1.08	0.89	0.88	
AA059077	1.45	1.44	1.12	1.1	1.15	

...

Iyer et al 1999 The Transcriptional Program in the Response of Human Fibroblasts to Serum

Fibroblast: bunky generujúce zložky medzibunkovej hmoty

pre delenie potrebujú rastové faktory dodávané ako "fetal bovine serum"

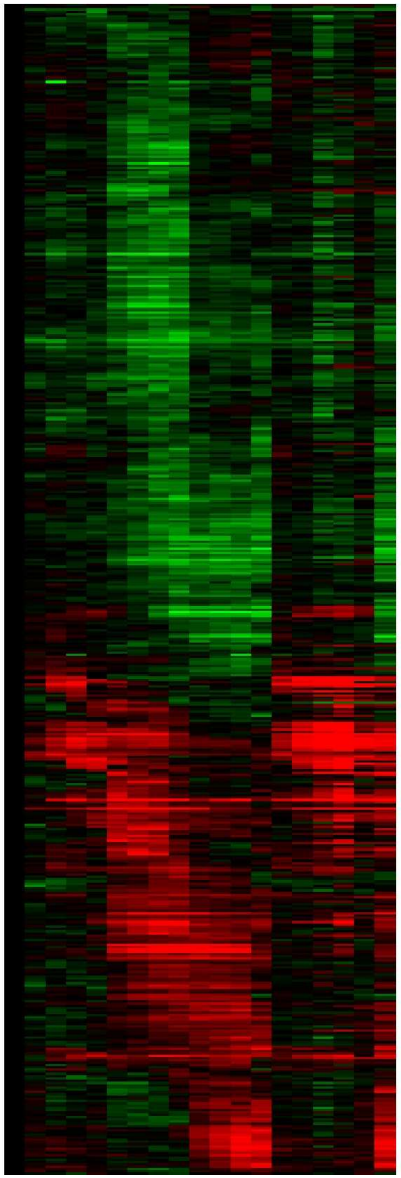
Vizualizácia

Červená: fg>bg

Zelená: fg<bg

517 génov (z 8600)

19 experimentov



Dnes: iný typ dát

- tabuľka čísel
- typické dáta v štatistike
- možno použiť všeobecné metódy štatistiky, strojového učenia

Všetky ostatné prednášky: pracujeme so sekvenciami

- zostavovanie genómov
- zarovnávanie sekvencií
- hľadanie génov
- fylogenetické stromy, populačná a komparatívna genomika
- štruktúra a funkcia proteínov a RNA

Prvá sada problémov: predspracovanie dát

- Zo scanovaných obrázkov určiť intenzitu, odhaliť zlé merania
- Agregácia dát z viacerých meraní pre jeden gén
- Použitie kontrolných meraní
- Normalizácia, aby sme mali porovnateľné výsledky z rôznych experimentov

Merania expresie nie veľmi presné, veľa šumu, rôzne zdroje chýb

Jednoduchý výsledok:

zoznam výrazne podexprimovaných/nadexprimovaných génov
napr. fg/bg> 2, resp. fg/bg< 0.5
často na ďalšiu analýzu používame iba tieto

Zhlukovanie (clustering)

Cieľ: nájsť skupiny génov s podobným profilom expresie.
Ak veľa génov v skupine má rovnakú funkciu,
ďalšie gény asi robia to isté

Meranie podobnosti profilov: napr. Pearsonov korelačný koeficient
Profil génu 1: $x_1, x_2, \ldots, x_n$, priemer $\overline{x}$
Profil génu 2: $y_1, y_2, \ldots, y_n$, priemer $\overline{y}$

$$C(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

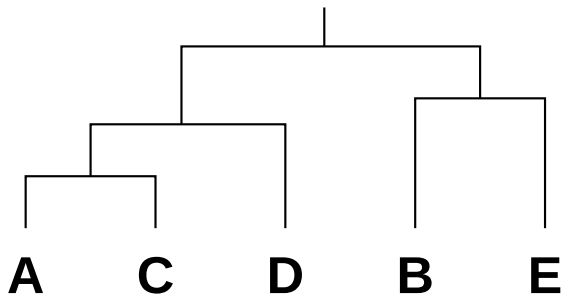
Číslo od -1 do 1, 1 pre lineárne korelované dáta
Vzdialenosť $d(x,y) = 1 - C(x,y)$

Aj iné možnosti, napr. Euklidovská vzdialenosť

Hierarchické zhlukovanie

- Podobné na metódu spájania susedov vo fylogenetických stromoch
- Začneme s každým génom v samostatnej skupinke
- Nájdeme dve najbližšie skupinky a spojíme ich do jednej
- Opakujeme, kým nie sú všetky gény spolu
- Vzdialenosť skupiniek: napr. vzdialenosť najbližších génov z jednej a druhej, alebo priemer vzdialeností cez všetky páry
- Výsledkom je strom zobrazujúci postupnosť spájania

	A	B	C	D	E
gén A	0	0.6	0.1	0.3	0.7
gén B	0.6	0	0.5	0.5	0.4
gén C	0.1	0.5	0	0.6	0.6
gén D	0.3	0.5	0.6	0	0.8
gén E	0.7	0.4	0.6	0.8	0

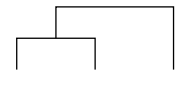


Hierarchické zhlukovanie - príklad

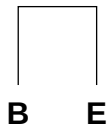
Vzdialenosť skupiniek ako vzdialenosť najbližších génov z jednej a druhej (single linkage clustering)

	A	B	C	D	E
gén A	0	0.6	0.1	0.3	0.7
gén B	0.6	0	0.5	0.5	0.4
gén C	0.1	0.5	0	0.6	0.6
gén D	0.3	0.5	0.6	0	0.8
gén E	0.7	0.4	0.6	0.8	0

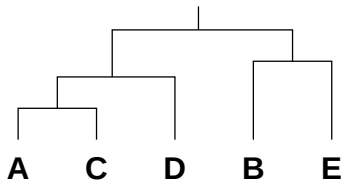
	A+C	B	D	E
A+C	0	0.5	0.3	0.6
B	0.5	0	0.5	0.4
D	0.3	0.5	0	0.8
E	0.6	0.4	0.8	0



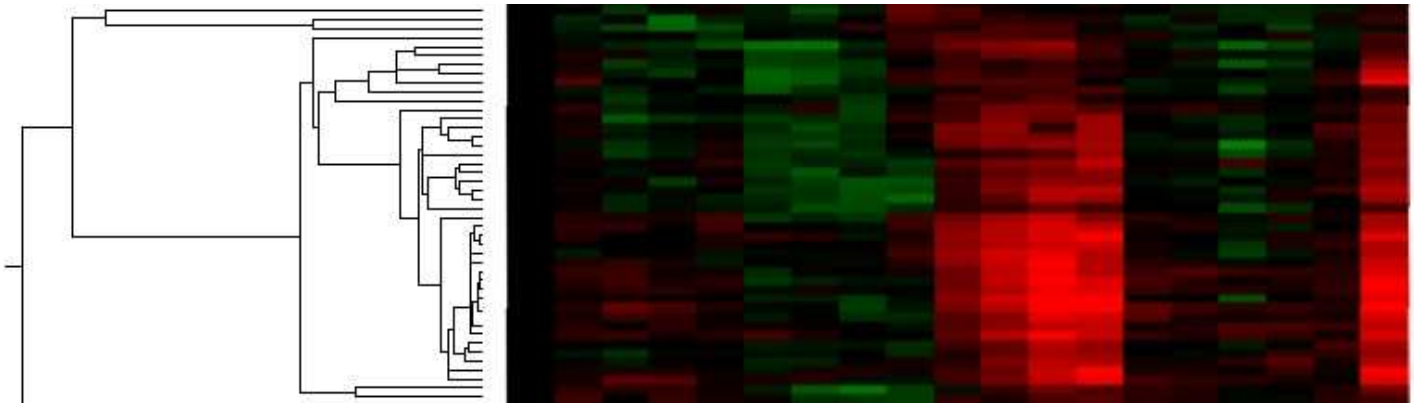
	A+C+D	B	E
A+C+D	0	0.5	0.6
B	0.5	0	0.4
E	0.6	0.4	0



	A+C+D	B+E
A+C+D	0	0.5
B+E	0.5	0

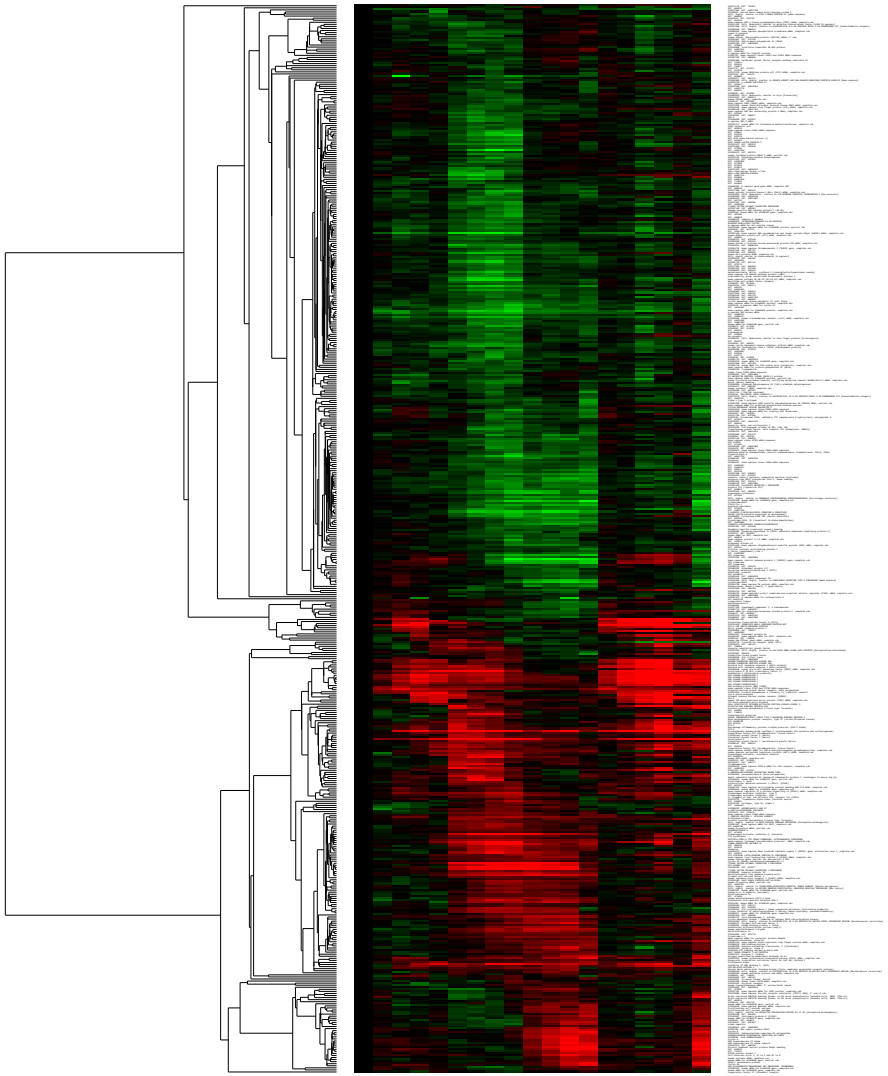


Príklad: časť mikroarray dát



Zhlukovanie tiež pomáha vizualizácii dát,

podobné gény sa dostanú ku sebe

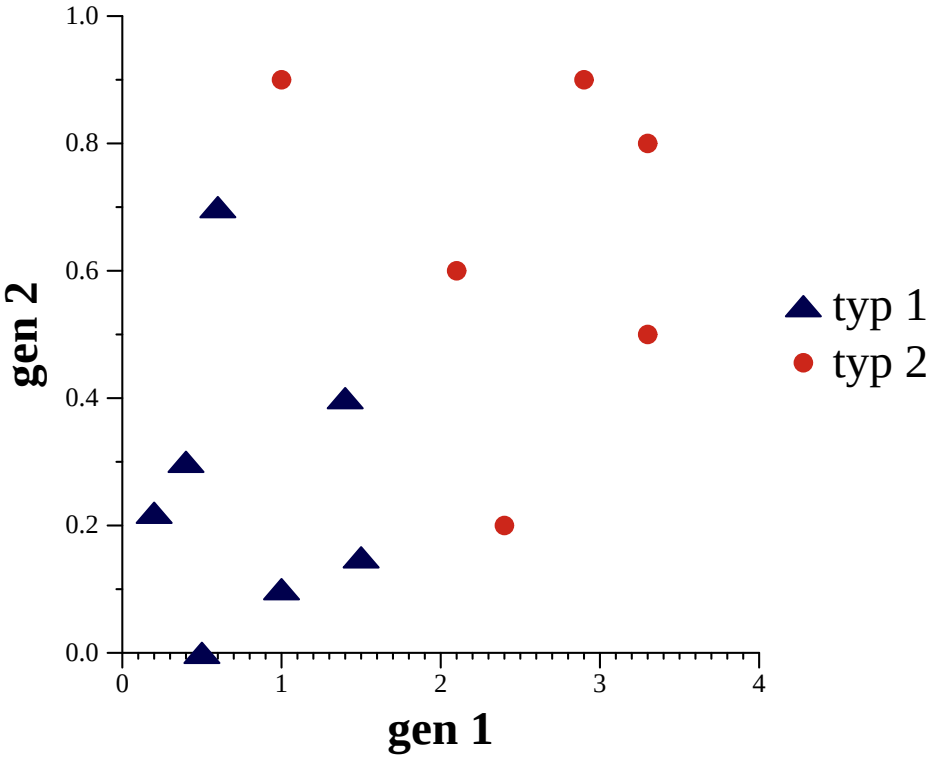


Klasifikácia

- Typický problém v strojovom učení
- Chceme odlíšiť napr. rôzne typy tumorov podľa expresie génov
- Máme nejaké príklady, kde vieme expresiu aj typ tumoru
- Chceme napr. nájsť vzorec, ktorý nám z expresie vyráta záporné číslo pre typ 1, kladné číslo pre typ 2.
- Vopred si vyberieme si typ vzorca s neznámymi parametrami (trieda hypotéz)
- Na trénovacích dátach hľadáme hodnoty parametrov, pre ktoré vzorec najlepšie funguje
- Fungovanie vzorca testujeme na testovacích dátach (nepoužité na trénovanie)
- Hotový vzorec použijeme na dáta s neznámym typom

Jednoduchý príklad: expresia 2 génov

Trénovacie dáta so známym typom:



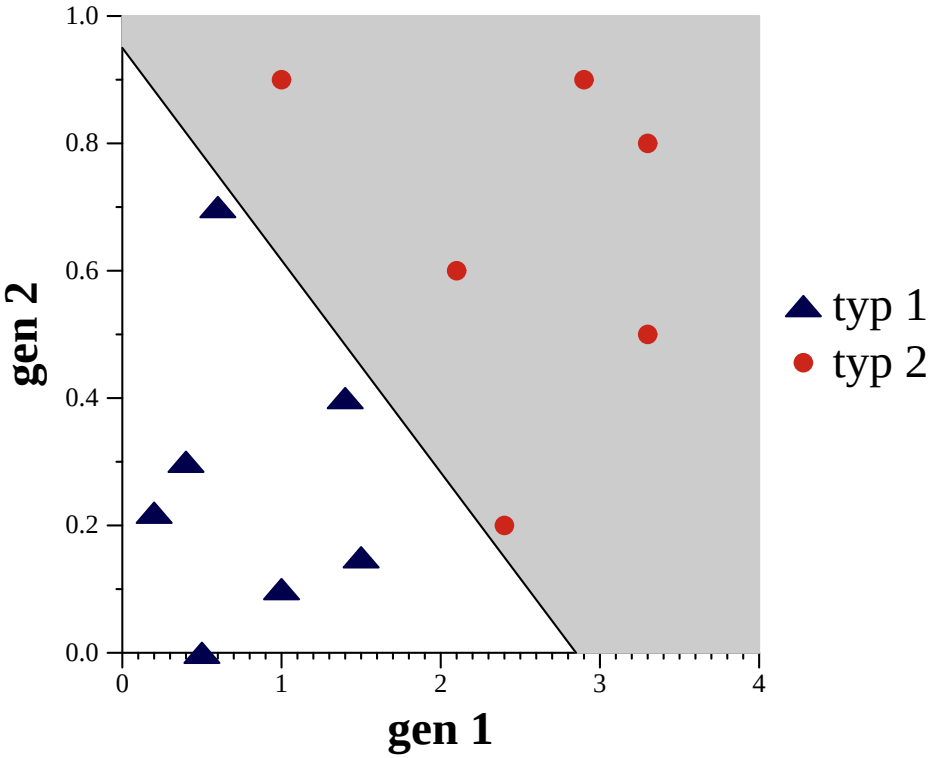
Typ vzorca: lineárne funkcie (lineárny diskriminant)

tumor typu 1 ak $ax + by + c < 0$

Hľadáme a, b, c také, aby na trénovacích dátach predpovedal dobre

Jednoduchý príklad: expresia 2 génov

Výsledný vzorec:



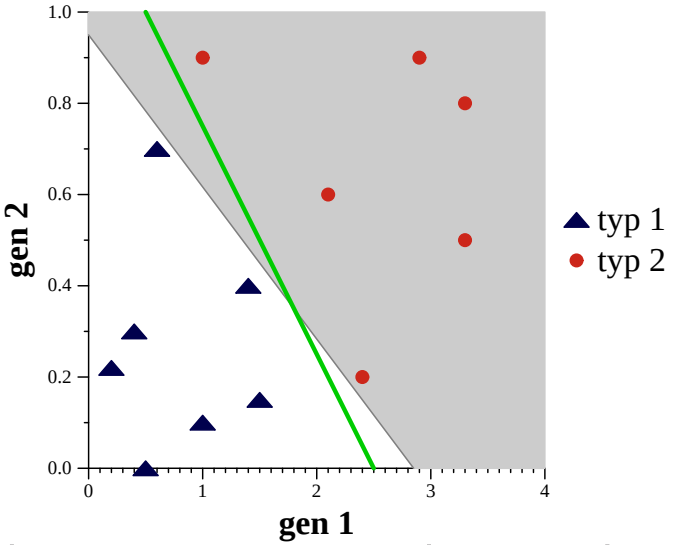
$a = 1, b = 3, c = -2.85$

tumor typu 1 ak $x + 3y - 2.85 < 0$

Populárne techniky na klasifikáciu

Logistic regression, logistická regresia:
lineárny diskriminátor, vracia pravdepodobnosť jednotlivých tried, dobre známa štatistická metóda.

Support vector machines (SVM): hľadanie lineárneho diskriminátora s nulovou trénovacou chybou, ktorý je najďalej od všetkých trénovacích dát.



Dá sa zovšeobecniť na nelineárne funkcie priemetom vektorov do väčšieho priestoru.

Populárne techniky na klasifikáciu

Neurónové siete:

"neuróny" poprepájané "synapsami",

každý neurón na výstupe váhovaný priemer vstupov.

Bayesovské siete:

pravdepodobnostný model generujúci náhodné expresie

typ tumoru je tiež náhodná premenná, ktorej hodnotu nepoznáme

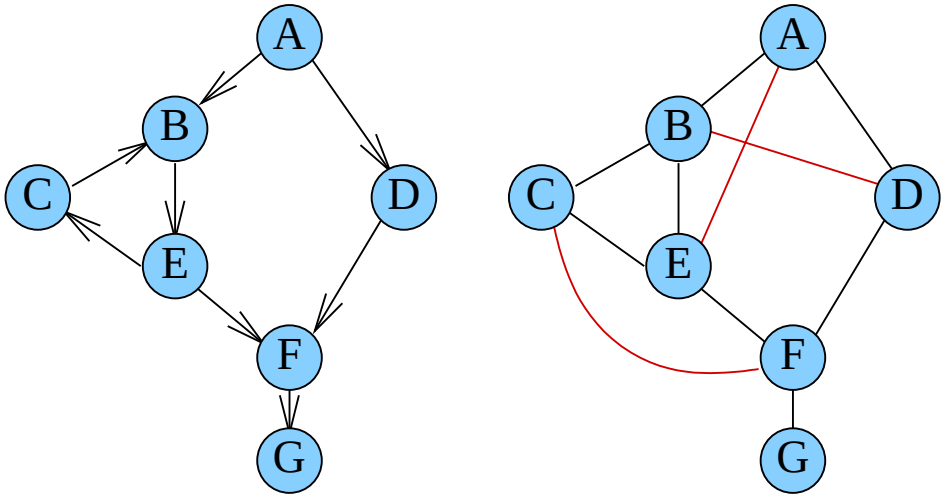
podobne ako stav v HMM

Regulačné siete z profilov expresie

Vstup: Profily expresie génov (napr. séria microarray/RNA-seq experimentov),
možno so známymi podmienkami (časové rady, delečný mutant)

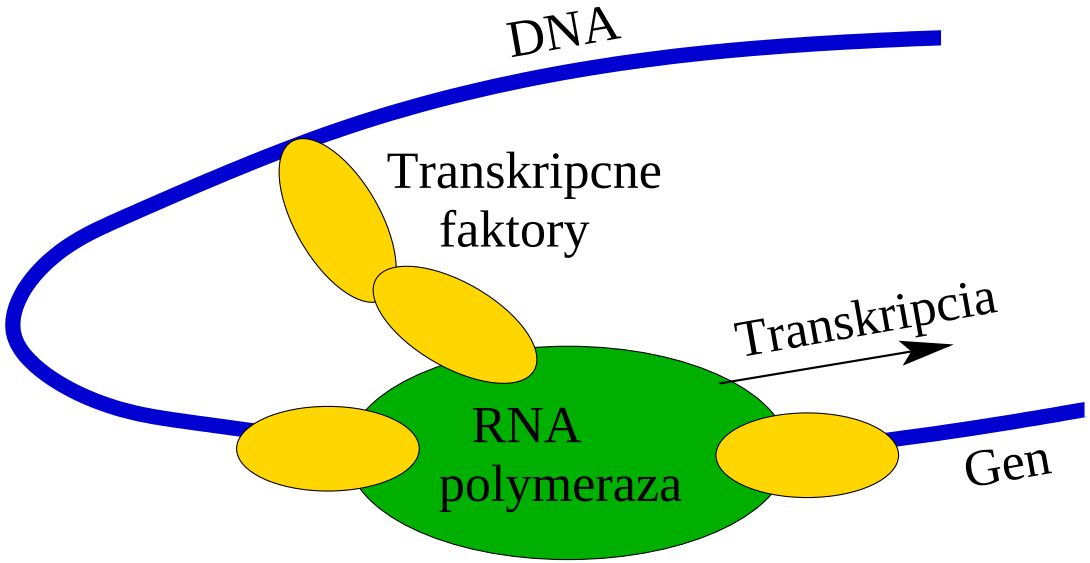
Výstup: regulačná sieť, vrcholy sú gény,
orientovaná hrana $A \to B$, ak A reguluje B

Podobnosť profilov expresie nám môže dať neorientované hrany.
Chceme vylúčiť hrany, ktoré vznikli tranzitivitou
a správne orientovať hrany (ťažký problém)



Transkripčné faktory (TF)

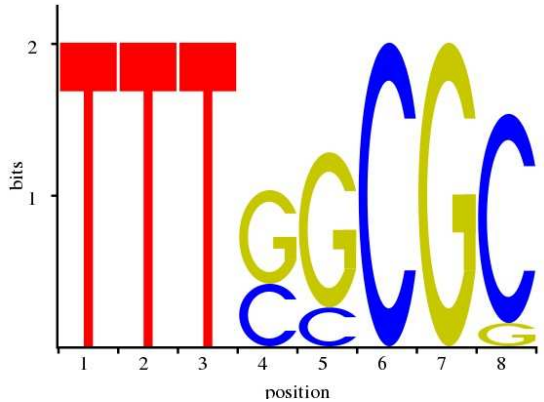
Regulácia začatia transkripcie pomocou transkripčných faktorov:
proteíny viažúce DNA, pomáhajú pritiahnuť RNA polymerázu

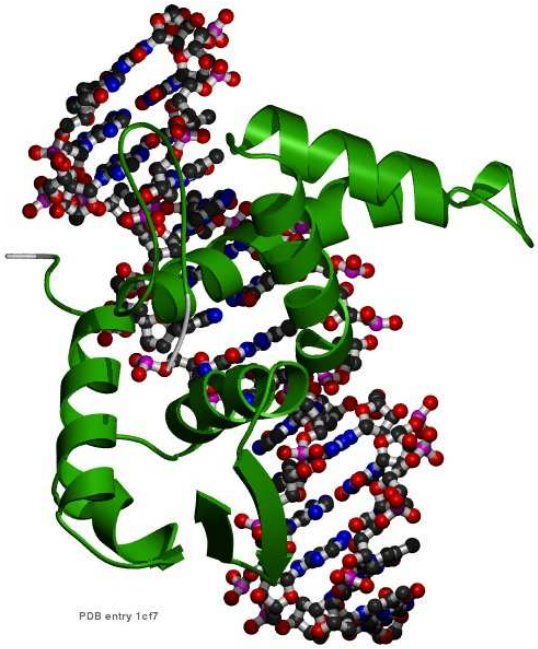


Človek má vyše 2000 TF-ov
Môžu zvyšovať alebo znižovať mieru expresie,
fungovať v skupinách

Príklad: transkripčný faktor E2F1

- Reguluje bunkový cyklus
- Viaže TTTCCCGC alebo TTTCGCGC, prípadne ďalšie varianty





- Sekvencie DNA, na ktoré sa viaže určitý TF chceme **reprezentovať** ako sekvenčný **motív** a hľadať **ďalšie výskyty** v genóme

Reprezentácia väzobných motívov

Reťazec s nezhodami (konsenzus):

motív je reťazec, výskyty môžu mať vopred ohraničený počet nezhôd

Príklad: motív TTTGGCGC + 1 nezhoda
TTTGGCGC, TT**A**GGCGC, TTTG**C**CGC sú výskyty motívu
TTT**CC**CGC nie je výskyt

Zostavenie motívu: napr. vezmi najčastejšie písmeno na každej pozícii

A	0	0	0	0	0	0	0	0
C	0	0	0	4	2	10	0	9
G	0	0	0	6	8	0	10	1
T	10	10	10	0	0	0	0	0

Reprezentácia väzobných motívov 2

Regulárny výraz:

niektoré pozície motívu dovoľujú výber z viacej možností

[GC] znamená pozíciu, na ktorej môže byť G alebo C

N znamená hociktorú bázu

Príklad: motív TTT[CG][CG]CGC

TTTGGCGC, TTT**CC**CGC, TTTG**C**CGC sú výskyty motívu

TT**A**GGCGC nie je výskyt

Zostavenie motívu: povoľ najčastejšie bázy na každej pozícii

```
A  0  0  0  0  0  0  0  0
C  0  0  0  4  2 10  0  9
G  0  0  0  6  8  0 10  1
T 10 10 10  0  0  0  0  0
```

Reprezentácia väzobných motívov 3

Position specific scoring matrix (PSSM, PWM):

skórovacia matica, skóre pre každú bázu na každej pozícii

Výskyty dosahujú skóre väčšie ako číslo T

Príklad: $T = 8$

```
A  -2.0  -2.0  -2.0 -2.0  -2.0  -2.0  -2.0  -2.0
C  -1.6  -1.6  -1.6  0.6   0.0   1.5  -1.6   1.4
G  -1.6  -1.6  -1.6  1.0   1.3  -1.6   1.5  -0.5
T   1.1   1.1   1.1 -2.0  -2.0  -2.0  -2.0  -2.0
```

TTT**CC**CGC je výskyt: 1.1+1.1+1.1+0.6+0.0+1.5+1.5+1.4=8.3

TTTGGCG**G** je výskyt: 1.1+1.1+1.1+1.0+1.3+1.5+1.5-0.5=8.1

TT**A**GGCGC nie je: 1.1+1.1-2.0+1.0+1.3+1.5+1.5+1.4=6.4

Zostavenie matice z frekvencií: budúca prednáška

Hľadanie výskytov v genóme

- Hľadanie motívu v genóme: skús každú pozíciu, či je výskytom
- Väčšinou veľa falošných výskytov
- Vieme spočítať E-hodnotu: koľko výskytov očakávame v náhodnej sekvencii
- Napr. TTT[CG][CG]CGC sa vyskytuje v priemere raz za 30 000 báz
- Na zlepšenie špecifickosti hľadáme
  - zhluky väzobných miest,
  - miesta podporené experimentálne,
  - evolučne zachované
- Databázy motívov, napr. TRANSFAC, JASPAR

Ako nájsť väzobné miesta experimentálne?

Chromatin immunoprecipitation (ChIP)

Pomocou protilátky (antibody) na špecifický transkripčný faktor zistí, kde približne sa tento faktor viaže.

- Väzba medzi TF a DNA sa spevní formaldehydom
- DNA sa naseká na kusy
- Kusy, na ktorých je TF, sa zachytia na protilátke
- DNA sa izoluje a sekvenuje pomocou NGS (**ChIP-seq**)
  alebo deteguje pomocou microarray (**ChIP-chip**)

Problém: zistíme len približnú polohu väzobného miesta

Ako nájsť motívy výpočtovými metódami?

. . . ak nemáme niekoľko príkladov väzobného miesta

- Máme skupinu sekvencií, kde každá obsahuje väzobné miesto toho istého TF, ale väzobné preferencie TF nie sú známe
- Snažíme sa nájsť **čo najšpecifickejší** motív, ktorý sa vyskytuje vo všetkých týchto sekvenciách
  resp. sa vyskytuje častejšie, ako by sme očakávali.
- **Pôvodne:** zoberieme skupinu génov s podobným profilom expresie a teda možno regulovaných tým istým TF, hľadáme motív v oblastiach pred týmito génmi
- **V súčasnosti:** zoberieme oblasti detegované pomocou ChIP-seq okolo väzobných miest, nájdený motív použijeme na presnejšie určenie polohy väzby TF

Príklad: Consensus Pattern Problem (CPP)

Jednoduchá formulácia problému hľadania motívov

Vstup: dĺžka motívu L, reťazce (sekvencie) $S_1, S_2, \ldots, S_k$

Výstup: motív (reťazec) M dĺžky L
a výskyt motívu v každom S_i (reťazec s_i dĺžky L)
také, že celkový počet nezhôd medzi M a s_i je najmenší možný

Príklad:
Vstup: CAAACAT, AGTAGC, TAACCA, TCTCCTC, $L = 4$
Výstup: motív TAAC
výskyty a nezhody AAAC 1, TAGC 1, TAAC 0, TCTC 2
celkový počet nezhôd 4

Riešenie CPP

NP-ťažký problém

- **Idea 1:** Vyskúšaj všetky možné motívy dĺžky L

  Problém: Nepraktické — prečo?

- **Idea 2:** Vyskúšaj všetky možné podreťazce dĺžky L reťazcov $S_1, \ldots, S_k$

  Problém: Nemusí fungovať — prečo?
  Ale dá sa dokázať, že cena riešenia bude najviac dvojnásobok optima (2-aproximačný algoritmus)

- **Ďalšie vylepšenie:** Skúšame všetky konsenzus sekvencie ℓ podreťazcov. PTAS (polynomial-time approximation scheme)

Praktickejší prístup k hľadaniu motívov

Pravdepodobnostný model generujúci sekvenciu S pomocou matice frekvencií báz v motíve W a frekvencie báz q mimo motívu

```
A 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01
C 0.01 0.01 0.01 0.39 0.19 0.97 0.01 0.01 0.89
G 0.01 0.01 0.01 0.59 0.79 0.01 0.97 0.97 0.09
T 0.97 0.97 0.97 0.01 0.01 0.01 0.01 0.01 0.01
```

$q(A) = 0.3$, $q(C) = 0.2$, $q(G) = 0.2$, $q(T) = 0.3$

Pozícia motívu v S sa zvolí náhodne, každá báza sa vygeneruje z q alebo z jedného stĺpca W

Tento model definuje rozdelenie $\Pr(S \mid W)$.

Hľadanie motívov cez pravdepodobnostné modely

Vstup: dĺžka motívu L, sekvencie $S_1, S_2, \ldots, S_k$, frekvencie q

Výstup: spoločný motív ako matica frekvencií W maximalizujúca vierohodnosť dát $\Pr(S_1|W) \cdot \ldots \cdot \Pr(S_k|W)$

- Ťažký problém, používajú sa heuristické algoritmy
- Napríklad EM (expectation maximalization)
- Lokálna optimalizácia, ktorá konverguje k lokálnemu maximu vierohodnosti
- Softvér: MEME

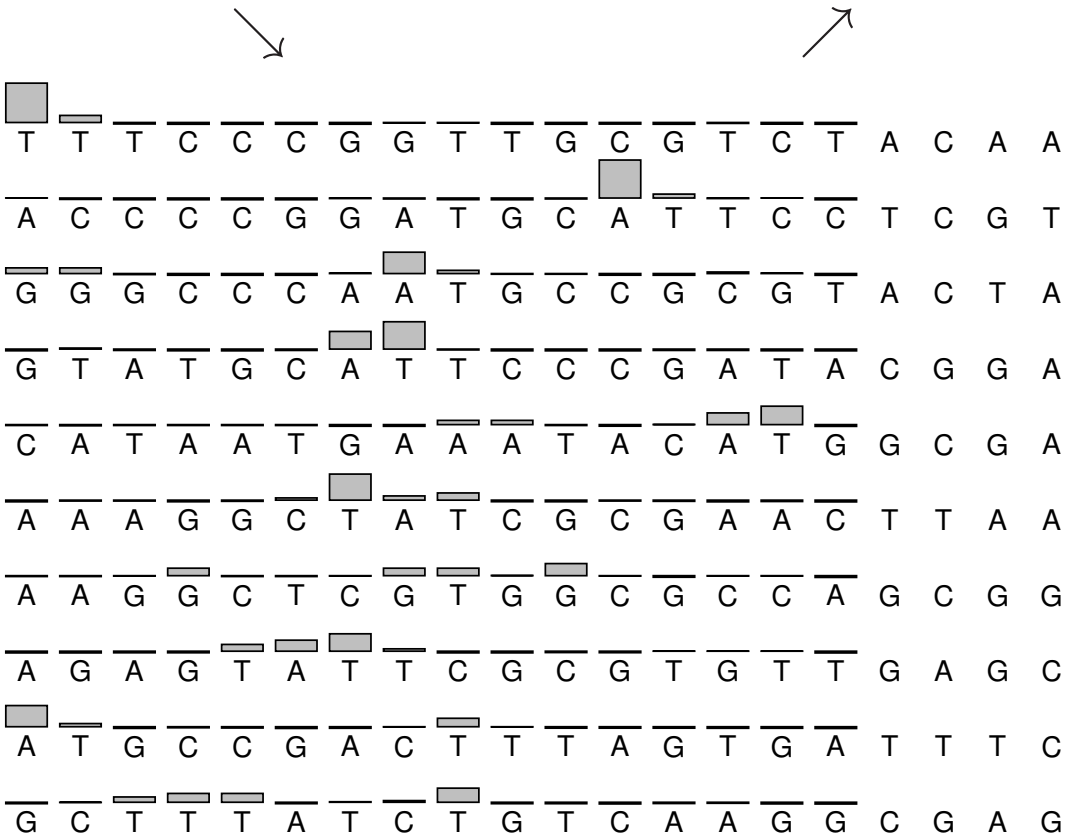
Schéma algoritmu EM

- **Inicializácia:**
  Zvoľ si počiatočnú maticu W
  (napr. zostavenú podľa jedného okna dĺžky L)
- **Iterácia:**
  1. Priraď každej pozícii j v sekvencii S_i váhu $p_{i,j}$, ktorá zodpovedá pravdepodobnosti, že na pozícii $S_i[j]$ začína výskyt motívu W
  2. Spočítaj W zo všetkých možných výskytov v $S_1, \ldots, S_k$ váhovaných podľa $p_{i,j}$

Iterácie zvyšujú vierohodnosť dát, kým nedôjde ku konvergencii.
Skúšame veľakrát z rôznych počiatočných W

Príklad algoritmu EM

A	0.10	0.10	0.10	0.10	0.10	A	0.31	0.14	0.06	0.07	0.07
C	0.10	0.10	0.10	0.70	0.70	C	0.06	0.10	0.19	0.71	0.61
G	0.10	0.10	0.10	0.10	0.10	G	0.12	0.17	0.29	0.14	0.25
T	0.70	0.70	0.70	0.10	0.10	T	0.51	0.60	0.46	0.08	0.07

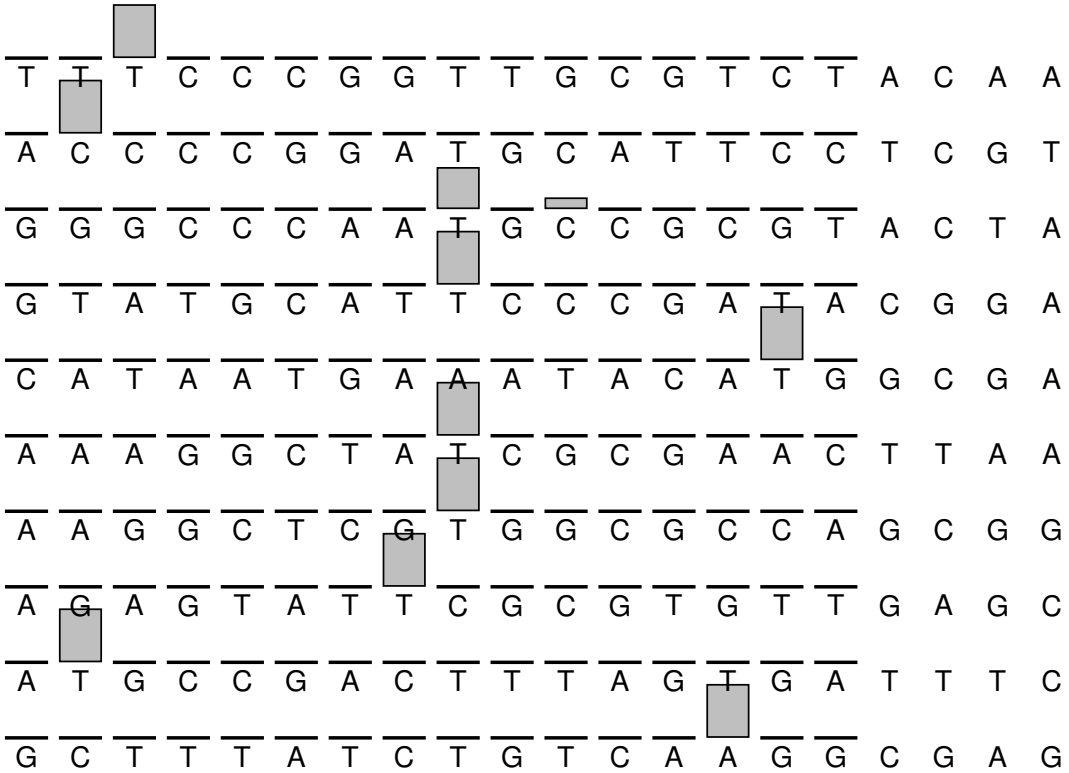


Príklad algoritmu EM: ďalšia iterácia

A	0.31	0.14	0.06	0.07	0.07	A	0.47	0.09	0.01	0.02	0.03
C	0.06	0.10	0.19	0.71	0.61	C	0.02	0.11	0.20	0.80	0.58
G	0.12	0.17	0.29	0.14	0.25	G	0.08	0.22	0.48	0.15	0.35
T	0.51	0.60	0.46	0.08	0.07	T	0.42	0.58	0.30	0.03	0.03

Príklad algoritmu EM: po 20 iteráciách

A	0.10	ϵ	ϵ	ϵ	ϵ
C	0.12	0.52	0.48	$1-3\epsilon$	ϵ
G	ϵ	0.48	0.52	ϵ	$1-3\epsilon$
T	0.78	ϵ	ϵ	ϵ	ϵ



Zhrnutie

- Microarray alebo RNA-seq merajú úroveň expresie pre veľa génov naraz, ale v dátach veľa šumu
- Zhlukovanie (clustering) nájde podobné gény,
  nepotrebujeme o dátach vopred nič vedieť (unsupervised learning)
- Klasifikácia môže rozlišovať napr. choroby podľa expresie,
  potrebuje dáta so známou odpoveďou (supervised learning)
- Dáta o expresii pomáhajú zostaviť regulačné siete
- Väzobné motívy môžeme reprezentovať rôznym spôsobom
  (reťazec, regulárny výraz, skórovacia matica)
- Tieto motívy nie sú dosť špecifické, preto sa ťažko rozpoznávajú
  ich výskyty v genóme
- EM algoritmus na hľadanie nových motívov v sekvenciách