Conditional Random Fields
a ich použitie na anotáciu biologických sekvencií

Broňa Brejová

Katedra informatiky FMFI UK

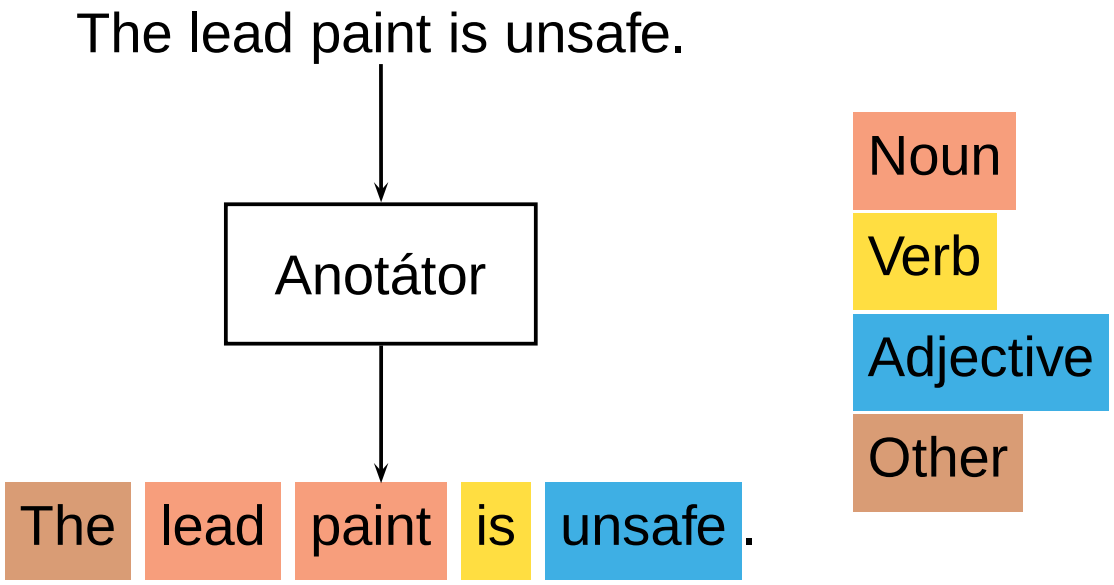
Conditional Random Fields
a ich použitie na anotáciu biologických sekvencií

- Anotácia sekvencií
- Skryté Markovove modely
- Conditional random fields (CRF) [Lafferty, McCallum, Pereira 2001]
- Použitie CRF na hľadanie génov [DeCaprio et al 2007]
- Náš výskum

Anotácia sekvencií/reťazcov

Označ (ofarbi) každý symbol na vstupe podľa jeho fukcie/významu.

Príklad: určovanie slovných druhov (part-of-speech tagging)



Zdroj: Andrew McCallum

Použitie v biológii: hľadanie génov

Vstup: DNA sekvencia $X = x_1, \ldots, x_n$

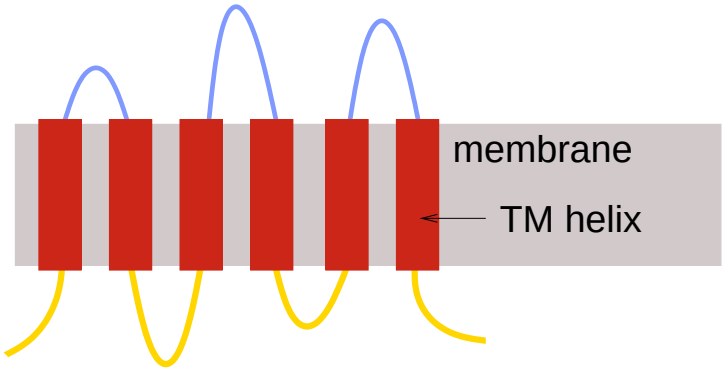
```
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
```

Výstup: anotácia (postupnosť farieb) $A = a_1, a_2, \ldots, a_n$

Červená: kóduje proteín, žltá: intrón, modrá: medzigénová oblasť

```
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
```

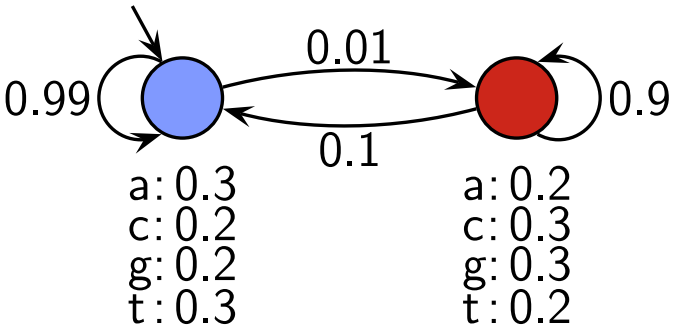
Použitie v biológii: topológia transmembránových proteínov



Červená: TM hélix, žltá: vnútri bunky, modrá: vonku

MNYSLHLAFVCLSLFTERMCIQGSQFNVEVGRSDKLSLPGFENLTAGYNKFLRPNFGGEP
VQIALTLDIASISSISESNMDYTATIYLRQRWMDQRLVFEGNKSFTLDARLVEFLWVPDT
YIVESKKSFLHEVTVGNRLIRLFSNGTVLYALRITTTVACNMDLSKYPMDTQTCKLQLES
WGYDGNDVEFTWLRGNDSVRGLEHLRLAQYTIERYFTLVTRSQQETGNYTRLVLQFELRR
NVLYFILETYVPSTFLVVLSWVSFWISLDSVPARTCIGVTTVLSMTTLMIGSRTSLPNTN
CFIKAIDVYLGICFSFVFGALLEYAVAHYSSLQQMAAKDRGTTKEVEEVSITNIINSSIS
SFKRKISFASIEISSDNVDYSDLTMKTSDKFKFVFREKMGRIVDYFTIQNPSNVDHYSKL
LFPLIFMLANVFYWAYYMYF

Skryté Markovove modely (hidden Markov models, HMMs)



Sekvencia: $\{a, c, g, t\}^*$

$$X = x_1, x_2, \ldots, x_n$$

Anotácia: $\{\square, \blacksquare\}^*$

$$A = a_1, a_2, \ldots, a_n$$

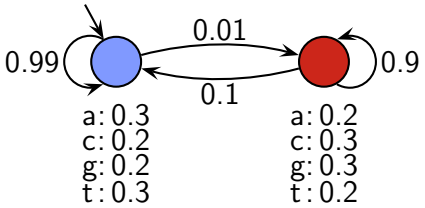
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta

HMM definuje $P(X, A)$ pre sekvencie X a anotácie A:

$$\begin{aligned} P(X, A) &= s(a_1)e(a_1, x_1)t(a_1, a_2)e(a_2, x_2)\cdots \\ &\quad t(a_{n-1}, a_n)e(a_n, x_n) \end{aligned}$$

Viterbiho algoritmus [Forney 1973]

Vstup: reťazec $X = x_1, \ldots, x_n$ a HMM




```
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
```

Cieľ: nájdi **najpravdepodobnejšiu anotáciu** $A^* = \arg\max_A P(X, A)$

```
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
```

Jednoduché dynamické programovanie

$O(n|E|)$ čas (n = dĺžka sekvencie, $|E|$ = počet prechodov)

Trénovanie HMM z anotovaných dát

Kritérium maximálnej vierohodnosti (maximum likelihood)

Poznáme (X, A)

Chceme parametre θ také, že $P_\theta(X, A)$ je maximálna

$$\theta^* = \arg\max_\theta P_\theta(X, A)$$

Jednoduché počítanie frekvencií, napr.

$$t(\blacksquare, \blacksquare) = \frac{|\{i \mid a_i = \blacksquare \wedge a_{i+1} = \blacksquare\}|}{n}$$

Trénovanie HMM z neanotovaných dát

Poznáme X

Chceme parametre θ^* také, že

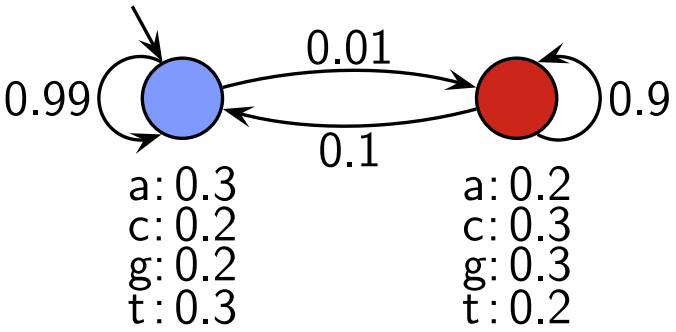
$$\theta^* = \arg\max_{\theta} P_\theta(X) = \arg\max_{\theta} \sum_{A} P_\theta(X, A)$$

Iteratívny Baum-Welchov algoritmus (verzia EM)

Každá iterácia dynamické programovanie $O(n|E|)$, zvýši $P_\theta(X)$

Nemusí konvergovať do globálneho optima

Zhrnutie HMM



Sekvencia: $\{a, c, g, t\}^*$

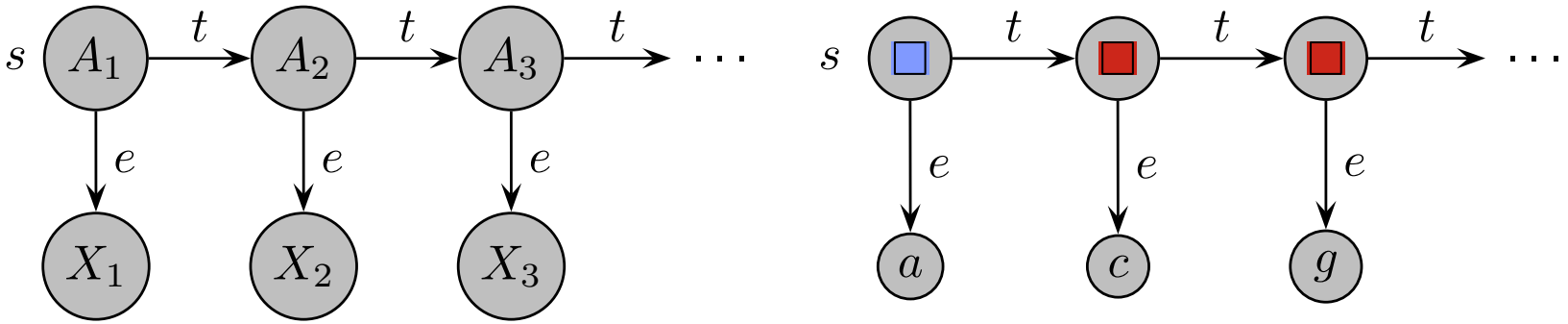
$$X = x_1, x_2, \ldots, x_n$$

Anotácia: $\{\square, \square\}^*$

$$A = a_1, a_2, \ldots, a_n$$

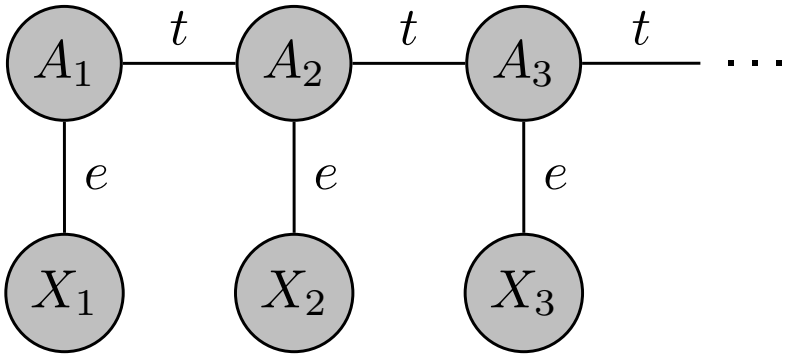
- Určíme stavový automat HMM
- Pravdepodobnosti trénujeme na anotovaných alebo neanotovaných dátach
- Efektívny algoritmus na hľadanie najpravdepodobnejšej anotácie

HMM ako bayesovská sieť



$$P(X, A) = s(a_1)e(a_1, x_1)t(a_1, a_2)e(a_2, x_2)\cdots$$
$$t(a_{n-1}, a_n)e(a_n, x_n)$$

Prvý krok k CRF: zmažeme šípky



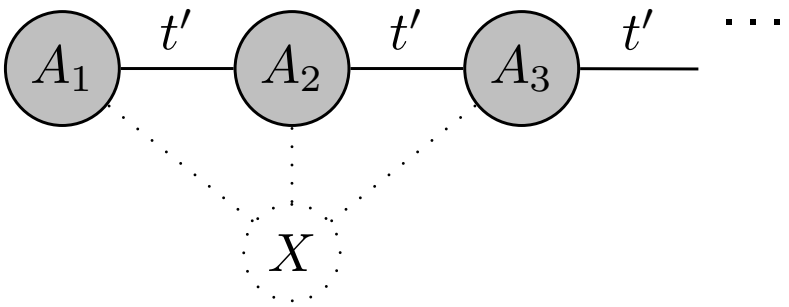
$$P(X, A) = \frac{1}{Z} e(a_1, x_1) t(a_1, a_2) e(a_2, x_2) \cdots t(a_{n-1}, a_n) e(a_n, x_n)$$

$$Z = \sum_{X', A'} P(X', A')$$

Markovovská sieť (neorientovaný grafový model)

$e(a, x)$ a $t(a, a')$ nenormalizované, ľubovoľné nezáporné čísla

Druhý krok k CRF: “zabudneme” na X



$$P(A|X) = \frac{1}{Z(X)} t'(1, a_1, a_2, X) \cdots t'(n-1, a_{n-1}, a_n, X)$$

$$Z(X) = \sum_{A'} P(A'|X)$$

Sekvencia X je daná, nepotrebujeme modelovať jej pravdepodobnosť.

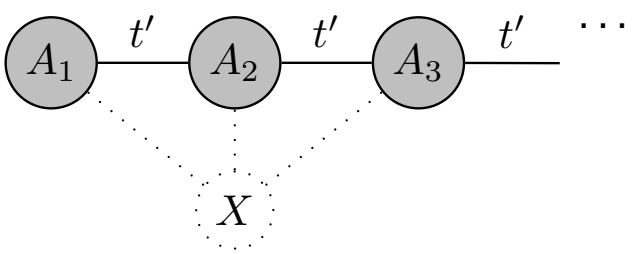
Ovplyvňuje však pravdepodobnosť anotácie A.

Funkcia $t'(j, a, a', X)$ závisí od dvoch susedných stavov a celého X.

Typický zápis CRF

$$P(A|X) = \frac{1}{Z(X)} t'(j, a_1, a_2, X) \cdots t'(j, a_{n-1}, a_n, X)$$

$$Z(X) = \sum_{A'} P(A'|X)$$



Váhovaný súčet lokálnych atribútov

$$P(A|X) = \frac{1}{Z(X)} exp\left(\sum_{i=1}^{k} \lambda_i \sum_{j=1}^{n-1} f_i(j, a_j, a_{j+1}, X)\right)$$

Atribút (feature) $f_i(j, a, a', X)$

Váha atribútu λ_i

$t'(j, a, a', X) = \prod_{i=1}^{k} (e^{f_i(j,a,a',X)})^{\lambda_i}$

HMM sú špeciálnym prípadom CRF

$$P(A|X) = \frac{1}{Z(X)} exp\left(\sum_{i=1}^{k} \lambda_i \sum_{j=1}^{n-1} f_i(j, a_j, a_{j+1}, X)\right)$$

$$Z(X) = \sum_{A'} P(A'|X)$$

Parametre HMM ako atribúty, jednotkové váhy:

$f_1(j, a, a', X) = \ln e(a, x_j)$

$f_2(j, a, a', X) = \ln t(a, a')$

$f_3(j, a, a', X) = \ln s(a)$ ak $j = 1$; inak 0

$\lambda_1 = \lambda_2 = \lambda_3 = 1$

HMM sú špeciálnym prípadom CRF

$$P(A|X) = \frac{1}{Z(X)} exp\left(\sum_{i=1}^{k} \lambda_i \sum_{j=1}^{n-1} f_i(j, a_j, a_{j+1}, X)\right)$$

$$Z(X) = \sum_{A'} P(A'|X)$$

Parametre HMM ako váhy, indikátorové atribúty:

$$f_{e,b,y}(j, a, a', X) = \delta(b, a)\delta(y, x_j)$$

$$f_{t,b,b'}(j, a, a', X) = \delta(b, a)\delta(b', a')$$

$$f_{s,b}(j, a, a', X) = \delta(b, a)\delta(j, 1)$$

$$\lambda_{e,b,y} = \ln e(b, y)$$

$$\lambda_{t,b,b'} = \ln t(b, b')$$

$$\lambda_{s,b} = \ln s(b)$$

Tretí krok k CRF: zmeníme trénovacie kritérium

Atribúty volíme ručne, trénujeme váhy λ_i

Pri HMM maximalizujeme vierohodnosť $\arg\max_\theta P(A, X)$

Pri CRF **maximalizujeme podmienenú vierohodnosť** $\arg\max_\lambda P(A|X)$

Trénovanie CRF:

- Nepoznáme uzavretý stav na výpočet váh z anotovaných dát
- Iteratívne algoritmy (iteratívne škálovanie, gradientové metódy)
- V každej iterácii dynamické programovanie na trénovacích dátach
- Logaritmus podmienej vierohodnosti je konkávny

  – konvergujeme ku globálnemu maximu

Zhrnutie CRF

$$P(A|X) = \frac{1}{Z(X)} exp\left(\sum_{i=1}^{k} \lambda_i \sum_{j=1}^{n-1} f_i(j, a_j, a_{j+1}, X)\right)$$

- Nemodelujú pravdepodobnosť známych dát X
- Môžeme si voliť ľubovoľné atribúty závisiace od celej sekvencie, prípadne ďalších dostupných dát
- Pri trénovaní maximalizujeme podmienenú pravdepodobnosť anotácie
- Trénovanie náročnejšie (iteratívna optimalizácia namiesto počítania frekvencií)
- Parametre nemajú intuitívny význam
- Použitie natrénovaného modelu stále Viterbiho algoritmom

Použitie CRF na hľadanie génov

Vstup: DNA + ďalšie zdroje informácii

Výstup: označenie oblastí kódujúcich proteíny

Bohatá história používania zovšeobecnených HMM

Program Conrad [DeCaprio et al 2007] založený na CRF

Používa semi-Markov CRF

– atribúty závisia od celého úseku s rovnakým stavom

– čas výpočtu potom kvadratický od n

– umožňuje modelovať dĺžky úsekov

Atribúty v Conrade

- Parametre HMM rozdelené do niekoľkých skupín, každá zvláštu váhu

  Napr. emisia v stave b: $f_b(j, a, a', X) = \delta(a, b) \ln e(b, x_j)$

  Emisie v rôznych stavov môžu mať rôznu váhu

- Použitie ďalších dát: porovnanie s genómami príbuzných druhov, experimentálne overenie transkripcie (indikátorové atribúty)

Zlepšenie presnosti prechodom na CRF

Training method	No. of species	Additional features	Nucl. sens.	Nucl. spec.	Exon sens.	Exon spec.	Gene sens.	Gene spec.
MEA-splice	5	Gap/Foot/EST	98.8	99.4	95.0	96.7	86.0	86.0
CML	2	—	97.9	98.2	85.6	90.2	61.3	61.8
MEA-Nucleotide	2	—	98.0	98.1	86.0	90.0	61.1	61.0
MEA-splice	2	—	97.8	98.1	85.8	89.9	60.8	60.8
GHMM	2	—	94.7	98.4	82.3	87.0	52.7	54.1

Adaptácia Conradu na mitochondriálne genómy kvasiniek

Diplomová práca Juraja Meštánka 2010

Mierna modifikácia základného modelu

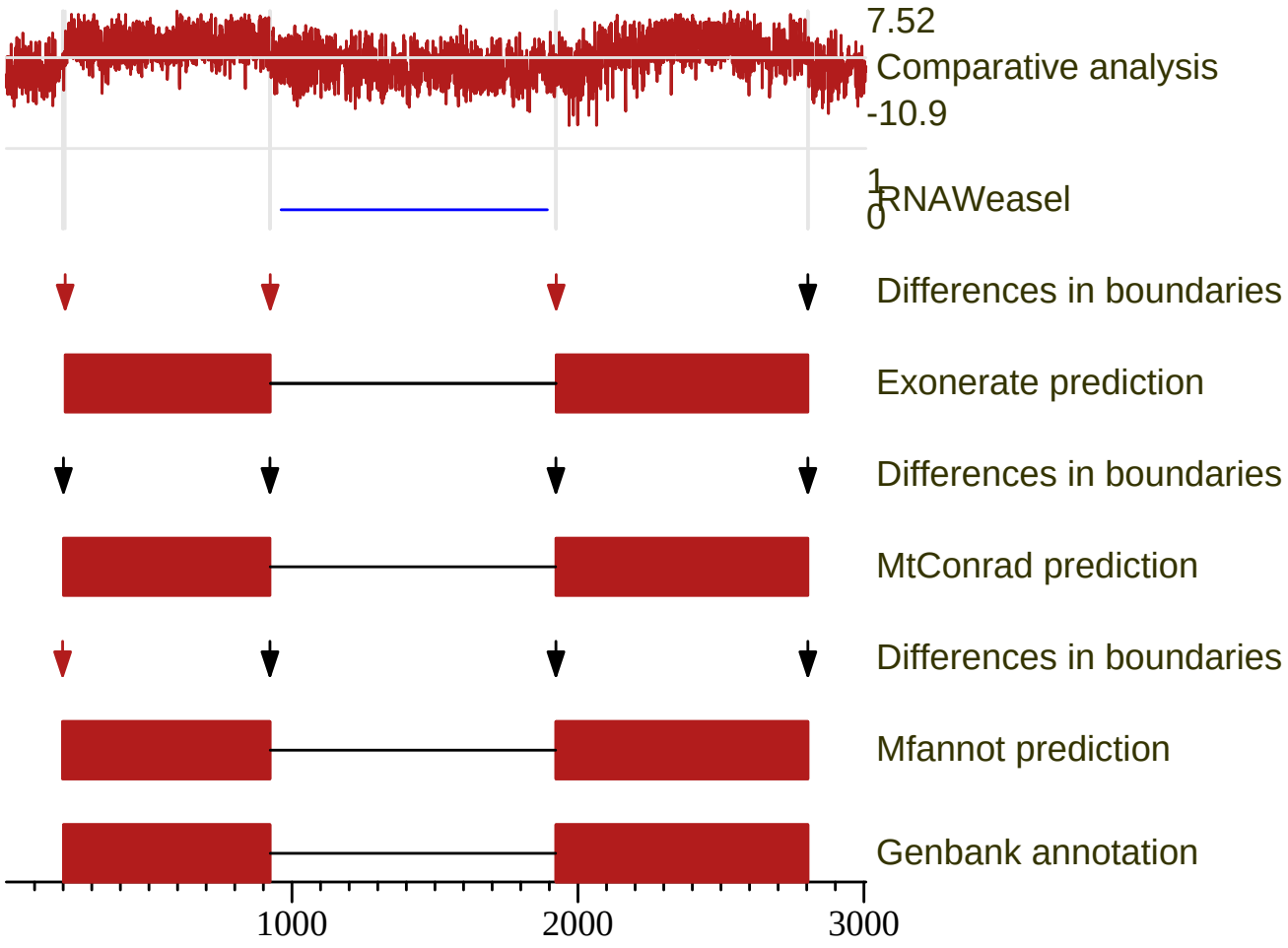
Využitie viacerých druhov prídavnej informácie

Zostavenie trénovacích a testovacích dát

Porovnanie presnosti

Tool	Perfect	Exon		Intron		Coding nucl.	
		sens	spec	sens	spec	sens	spec
HMM	58%	45%	27%			96%	99%
MtConrad	**78%**	70%	74%	55%	59%	**99%**	99%
MFannot	74%	**77%**	**81%**	**82%**	**82%**	94%	**99.8%**

Ukážka jedného génu



Červené šípky nesprávne okraje.

Ďalšie plány

- Implementácia CRF trénovania do nášho hľadača génov ExonHunter (prototyp diplomant Lukáš Bača)
- Oproti Conradu ExonHunter lepšie spracováva dĺžky úsekov, má systém na trénovanie na nových organizmoch.
- Rozšírenie HMM resp. CRF na zložitejšie typy prídavnej informácie (diplomant Marcel Kucharík)
- Rôzne kritériá odvodzovania v HMM resp. CRF (doktorand Michal Nánási)