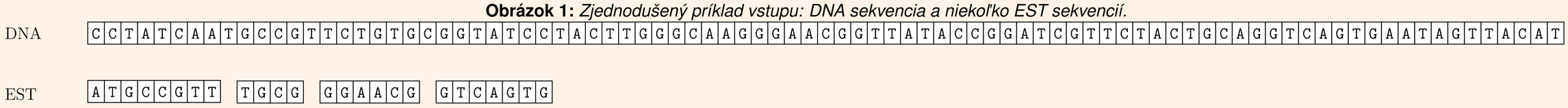
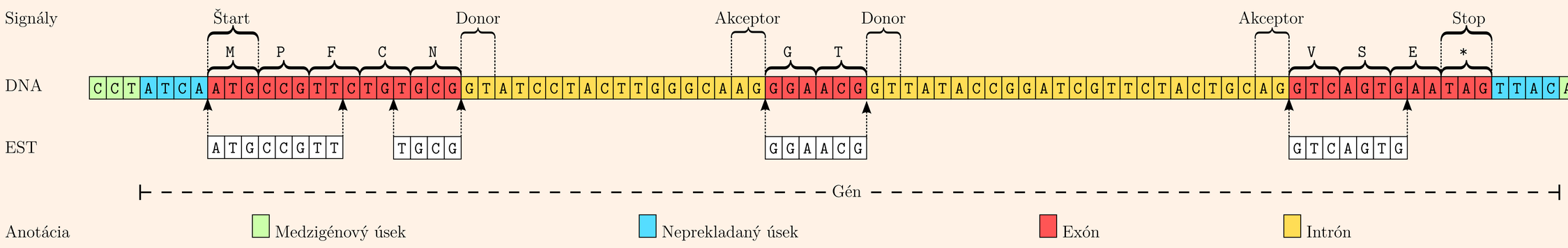
Spracovanie externých informácii pri hľadaní génov

Peter Kováč[1]
Školiteľ: Bronislava Brejová[1]
[1] Katedra informatiky, FMFI UK, Mlynská Dolina, 842 48 Bratislava

Obrázok 1: *Zjednodušený príklad vstupu: DNA sekvencia a niekoľko EST sekvencií.*



Obrázok 2: *Zjednodušený príklad výslednej štruktúry génu. Zarovnaním EST sekvencií k DNA sekvencii sa určia približné hranice exónov a potom pomocou signálov sa určia presné hranice a poradie exónov a intrónov. Gén na obrázku kóduje sekvenciu aminokyselín MPFCNGTVSE. Dĺžka skutočných ľudských exónov je niekoľko stoviek nukleotidov, dĺžka intrónov je niekoľko tisíc nukleotidov.*



Úvod do biológie

Hľadanie (predikcia) génov v DNA sekvenciách je jedným zo základných problémov bioinformatiky. **DNA** sa nachádza v bunkách živých organizmov a je nosičom genetickej informácie. DNA sekvencia je reťazec molekúl zvaných **nukleotidy** (*adenozín, cytozín, guanín, tymín*). **Gén** je súvislý úsek DNA. Zložitým molekulárnym aparátom sa z génu tvorí jeden alebo viacero *proteínov* (reťazcov *aminokyselín*), molekúl nevyhnutných pre život organizmu. Gény pozostávajú z:

- **neprekladaných úsekov** – krátke úseky na začiatku a konci génu
- **exónov** – úseky kódujúce reťazce aminokyselín; skladajú sa z trojíc nukleotidov nazývaných *kodóny*, každá trojica je kódom pre jednu aminokyselinu; na začiatku prvého exónu je *štart kodón* (ATG), na konci posledného exónu je *stop kodón* (TAA, TAG, alebo TGA)
- **intrónov** – úseky medzi exónmi, počas prepisu DNA do proteínu sa vystrihnú
- **signálov** – špeciálne miesta v génoch, napríklad:
  - *donor splice site* – začiatok intrónu, zvyčajne GT
  - *acceptor splice site* – koniec intrónu, zvyčajne AG

Okrem DNA sa pri hľadaní génov využívajú aj tzv. externé dáta - EST sekvencie, proteínové sekvencie, zarovnania viacerých DNA sekvencií atď. **EST** (*expressed sequence tag*) je krátky úsek DNA získaný sekvenovaním exprimovaných génov - to znamená, že už **neobsahuje introny a medzigénové úseky**. Zarovnaním EST k DNA vieme určiť približnú polohu exónov a intrónov v DNA sekvencii (obr. 2).

Hľadanie génov s externými informáciami

Pre informatikov je DNA sekvencia reťazcom nad abecedou $\{A, C, G, T\}$ a hľadanie génov znamená označenie (*anotáciu*) každého písmena prvkom z množiny

$$G = \{\text{exón}, \text{intrón}, \text{neprekladaný úsek}, \text{medzigénový úsek}, ...\}.$$

Problém sa často rieši Viterbiho algoritmom na skrytých Markovovských modeloch (HMM). Zahrnutie externých informácií môže významne zlepšiť výsledok.
Program **ExonHunter** (EH) [Brejová et al., 2005] okrem HMM obsahuje aj mechanizmus na spracovanie EST dát. Pôvodná implementácia pomocou programu *sim4* [Florea et al., 1998] nevyhovovala kvôli celkovému času, potrebnému na predikciu. Preto sme vymenili sim4 za program *BLAT* [Kent, 2002], ktorý zarovnania nájde rýchlejšie.

Výsledky predikcií

Testovali sme predikciu na sekvencii mušky *Drosophila melanogaster* (tab. 2). Testovacia DNA mala 16 MB (1796 génov, 8626 exónov), EST sekvencie 62 MB. Porovnanie EH s inými programami (tab. 3) ukázalo veľké rozdiely v presnosti a čase. Sú spôsobené rôznymi HMM a optimálnosťou natrénovaných parametrov týchto modelov. Časy behov jednotlivých programov sú len orientačné, namerané na počítačoch s 2 až 2.27 GHz CPU, 16 GB RAM.

Tabuľka 1: *Senzitivita a špecifickosť EST zarovnaní*

	sim4	BLAT
Exón Sn.	84.49%	86.46%
Exón Sp.	22.19%	58.96%

Tabuľka 2: *Senzitivita a špecifickosť predikcie*

	EH	EH+sim4	EH+BLAT
Gén Sn.	41.98%	54.34%	51.95%
Gén Sp.	47.63%	50.28%	49.29%
Exón Sn.	72.03%	77.22%	77.37%
Exón Sp.	72.48%	73.30%	73.11%
Čas behu	120 min	282 min	147 min

Tabuľka 3: *Iné programy: A - Augustus, A+EST - Augustus s EST dátami, GI - GeneID, GM - GeneMark.*

	A	A+EST	GI	GM
Gén Sn.	49.50%	63.86%	32.32%	46.71%
Gén Sp.	51.99%	65.45%	32.86%	37.60%
Exón Sn.	69.12%	82.52%	65.58%	71.72%
Exón Sp.	79.81%	82.64%	66.08%	64.23%
Čas behu	52 min	1707 min	3 min	82 min

Literatúra

[Brejová et al., 2005] Brejová, B., Brown, D., Li, M., and Vinař, T. (2005). Exonhunter: a comprehensive approach to gene finding. *Bioinformatics*, 21:i57–i65.

[Florea et al., 1998] Florea, L., Hartzell, G., Zhang, Z., Rubin, G., and Miller, W. (1998). A computer program for aligning a cdna sequence with a genomic dna sequence. *Genome Res.*, 8(9):967–974.

[Kent, 2002] Kent, W. J. (2002). Blat: The blast like alignment tool. *Genome Res.*, 12:656–664.