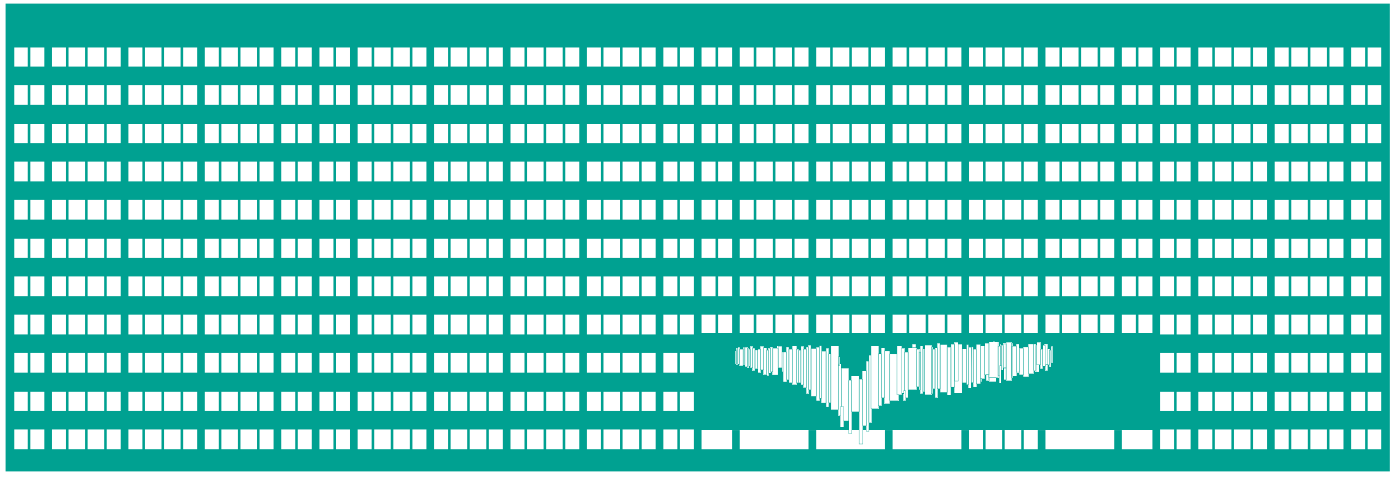
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

Algoritmy pro Bioinformatiku

Vyhledávání v databázích - BLAST

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

7. listopadu 2023

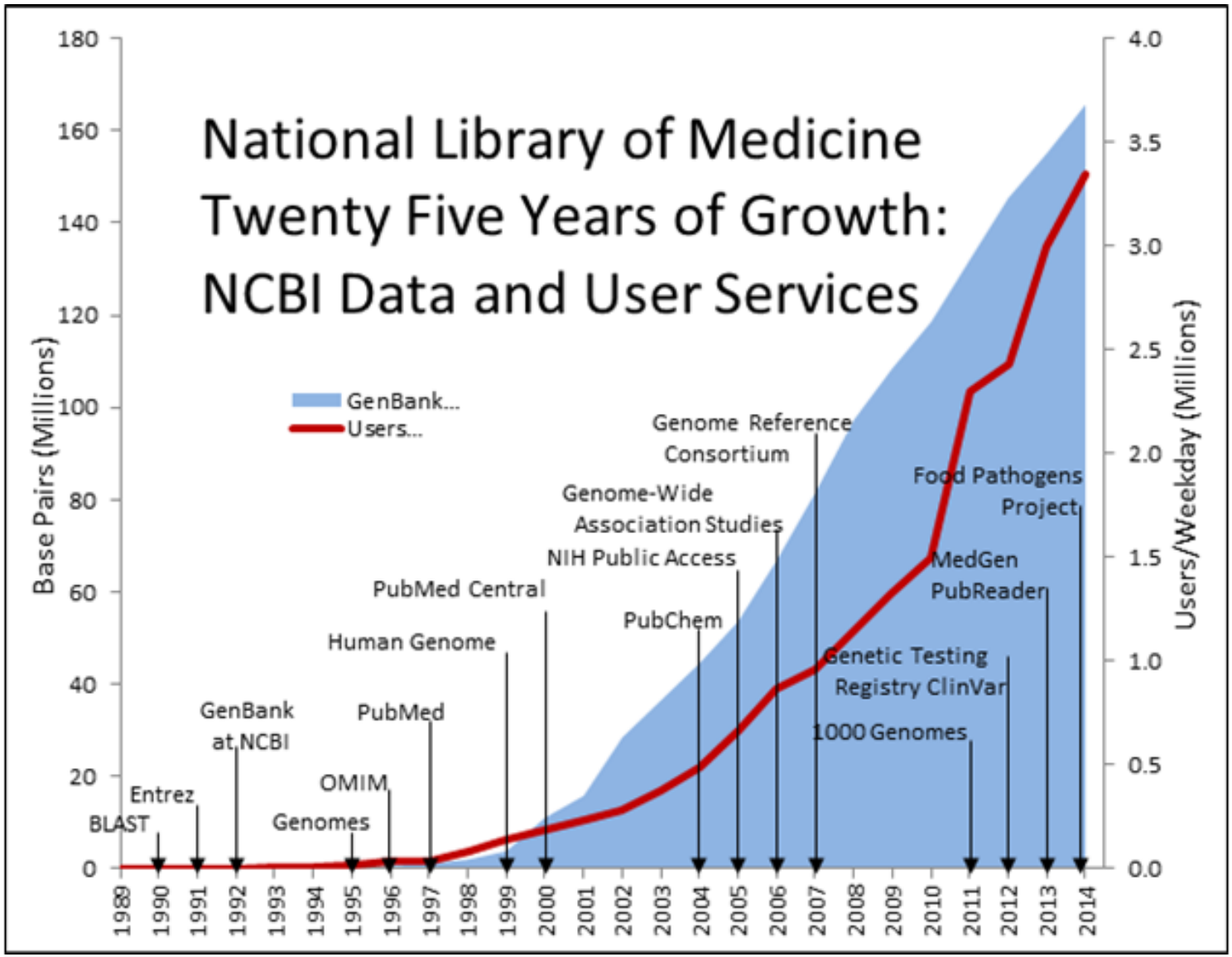
VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

BLAST

Databáze biologických dat

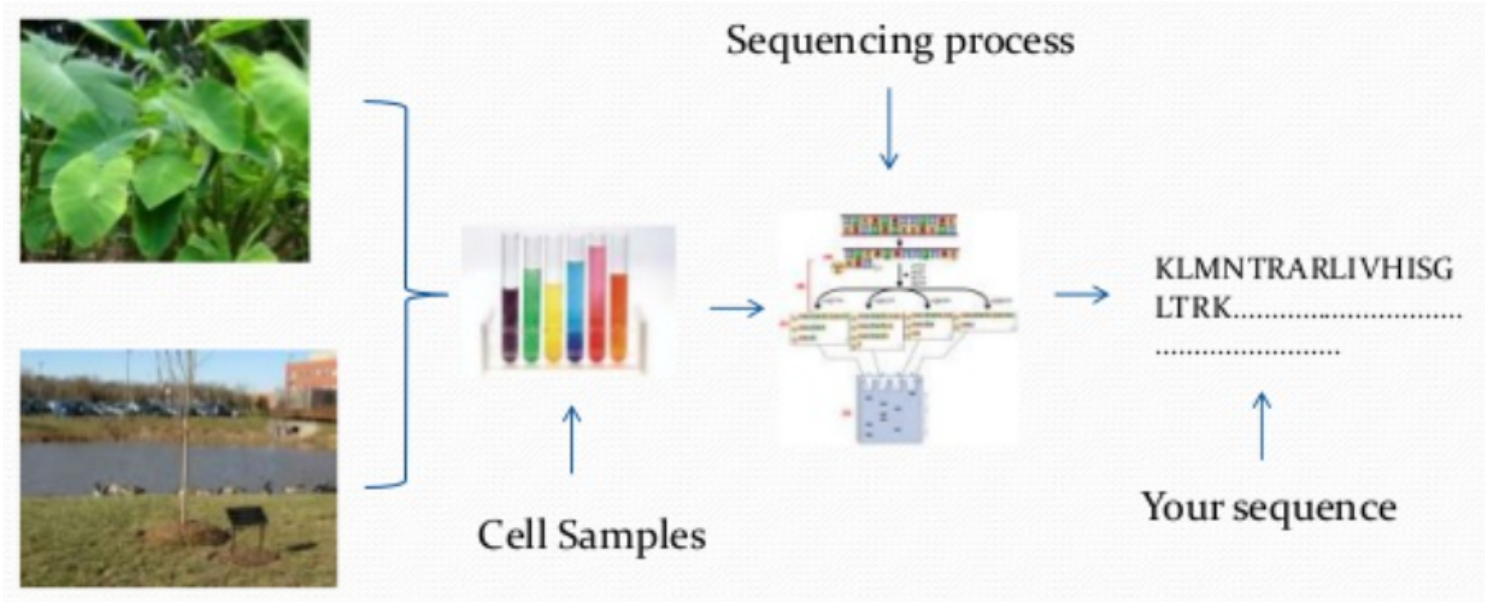
- Množství genomických dat se každých 15 měsíců zdvojnásobí.
- Rostoucí počet dotazů na databáze.
- Potřeba efektivních vyhledávacích metod pro genomická data.

Databáze biologických dat



Databáze biologických dat - použití

- Laboratoř odsekvenovala sekvenci nukleotidů nebo aminokyselin nějakého proteinu.
- Hledáte sekvenci která má podobnou funkci jako jiná sekvence => hledáte podobnou sekvenci.
- Zjištění podobnosti umožňuje odvodit strukturu, funkce a evoluční vztahy hledané sekvence.



Databáze biologických dat - komponenty

Tři hlavní komponenty

- Ohodnocovací funkce - určení podobnosti
- Vyhledávací algoritmus - nalezení záznamu v databázi
- Statistický model - přiřazení významnosti k nalezenému záznamu

- Komponenty berou v potaz především rychlost vykonání dotazu.

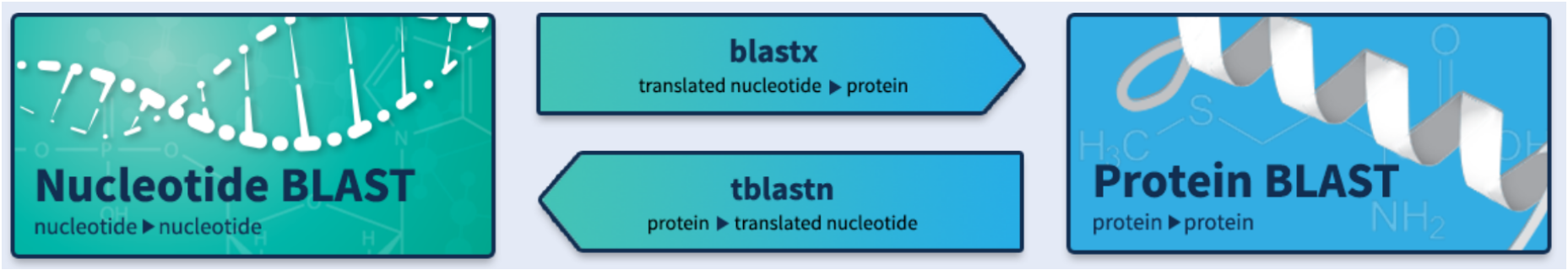
Alignment - lokální vs. globální

Alignment sekvencí nukleotidů či aminokyselin je vzájemné přiřazení dvou sekvencí.

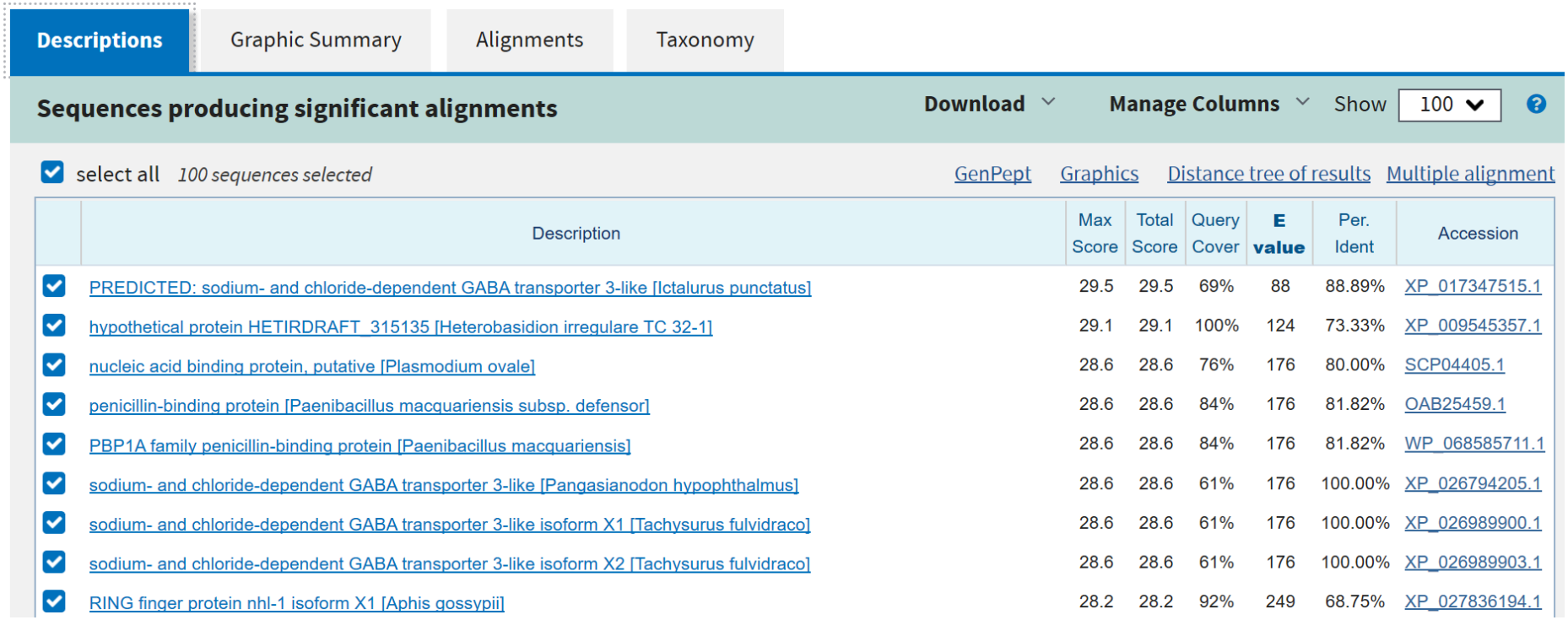
```
Global  FTFTALILLAVAV
        F__TAL_LLA_AV
 Local  FTFTALILLAVAV
        __FTAL_LLAAV__
```

BLAST

- BLAST je nástroj používaný pro porovnávání sekvencí s knihovnou či databází sekvencí.
- Využívá lokální alignment.
- Používá heuristického přístupu založeného na statistických metodách.
- Hlavní cíl je umožnit rychlé vyhledávání v databázích.



BLAST Rozhraní



Descriptions | Graphic Summary | Alignments | Taxonomy

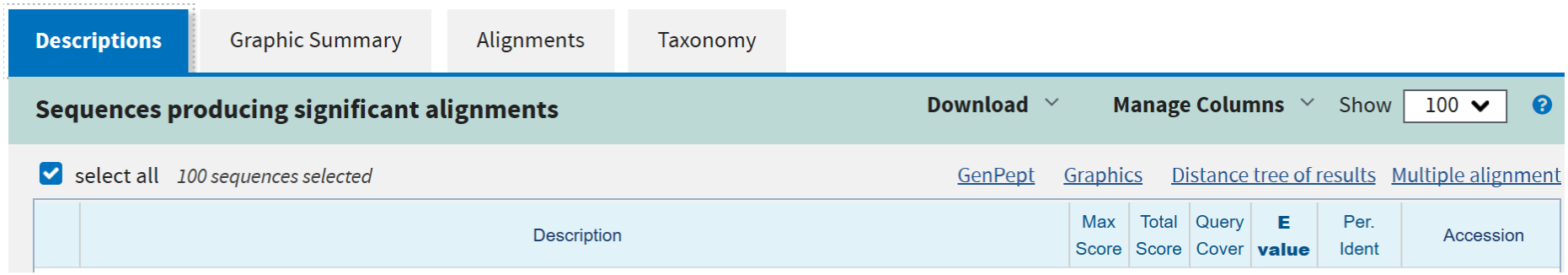
Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
PREDICTED: sodium- and chloride-dependent GABA transporter 3-like [Ictalurus punctatus]	29.5	29.5	69%	88	88.89%	XP_017347515.1
hypothetical protein HETIRDRAFT_315135 [Heterobasidion irregulare TC 32-1]	29.1	29.1	100%	124	73.33%	XP_009545357.1
nucleic acid binding protein, putative [Plasmodium ovale]	28.6	28.6	76%	176	80.00%	SCP04405.1
penicillin-binding protein [Paenibacillus macquariensis subsp. defensor]	28.6	28.6	84%	176	81.82%	OAB25459.1
PBP1A family penicillin-binding protein [Paenibacillus macquariensis]	28.6	28.6	84%	176	81.82%	WP_068585711.1
sodium- and chloride-dependent GABA transporter 3-like [Pangasianodon hypophthalmus]	28.6	28.6	61%	176	100.00%	XP_026794205.1
sodium- and chloride-dependent GABA transporter 3-like isoform X1 [Tachysurus fulvidraco]	28.6	28.6	61%	176	100.00%	XP_026989900.1
sodium- and chloride-dependent GABA transporter 3-like isoform X2 [Tachysurus fulvidraco]	28.6	28.6	61%	176	100.00%	XP_026989903.1
RING finger protein nhl-1 isoform X1 [Aphis gossypii]	28.2	28.2	92%	249	68.75%	XP_027836194.1

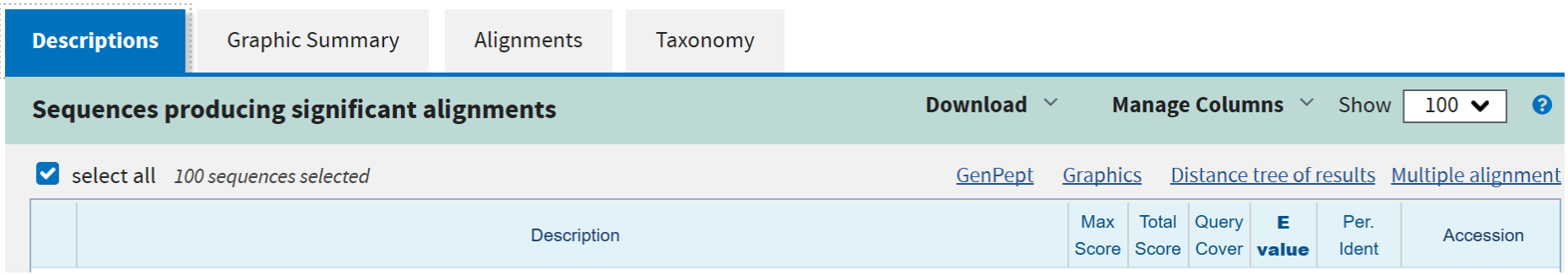
BLAST Rozhraní - pojmy

- Max Score - skóre alignmentu jednoho z úseků nalezené sekvence, který se povedlo namapovat na sekvenci z databáze. V případě, že je nalezená sekvence je namapována sekvenci z databáze v celé délce, hodnota maximálního skóre bude totožná celkovému skóre(viz. níže).
- Total Score - součet skóre všech nekontinuálních částí lokálních namapování mezi hledanou a nalezenou sekvencí z databáze.
- Query Cover - jaká část hledané sekvence je porovnána s nalezenou sekvencí.

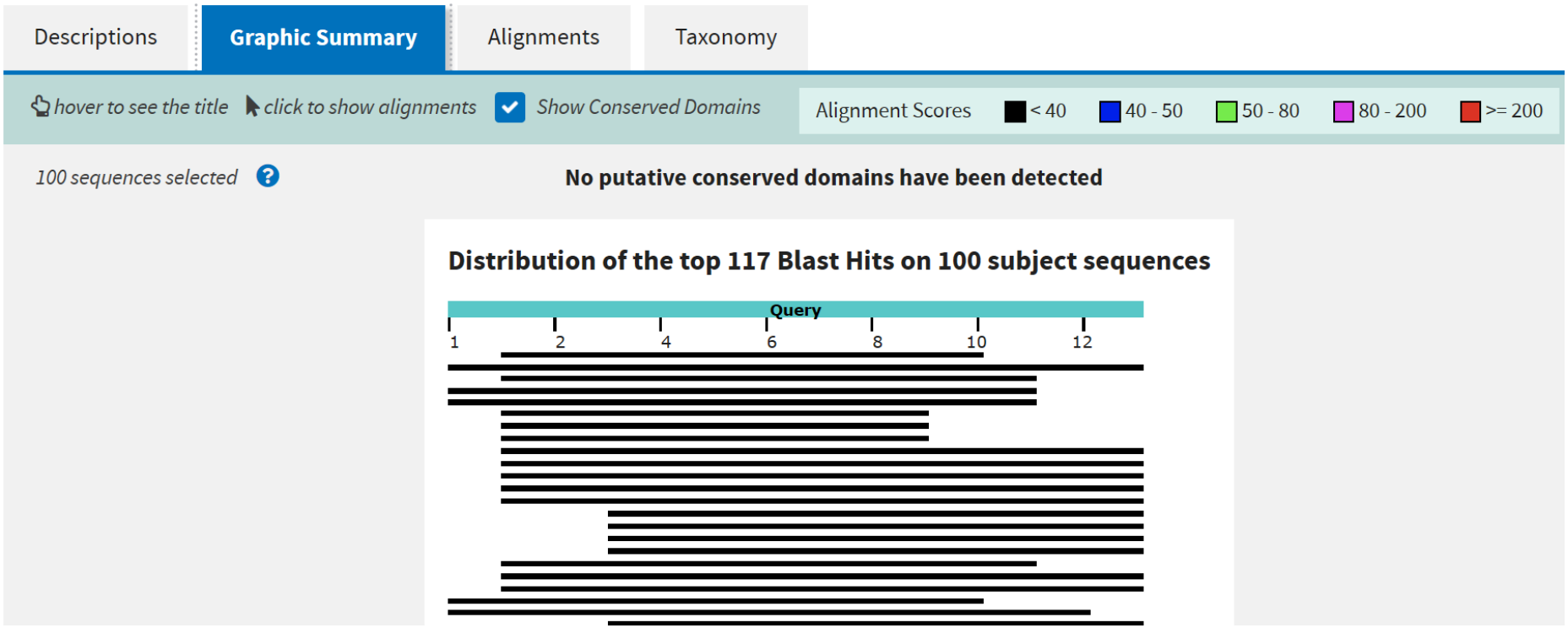


BLAST Rozhraní - pojmy

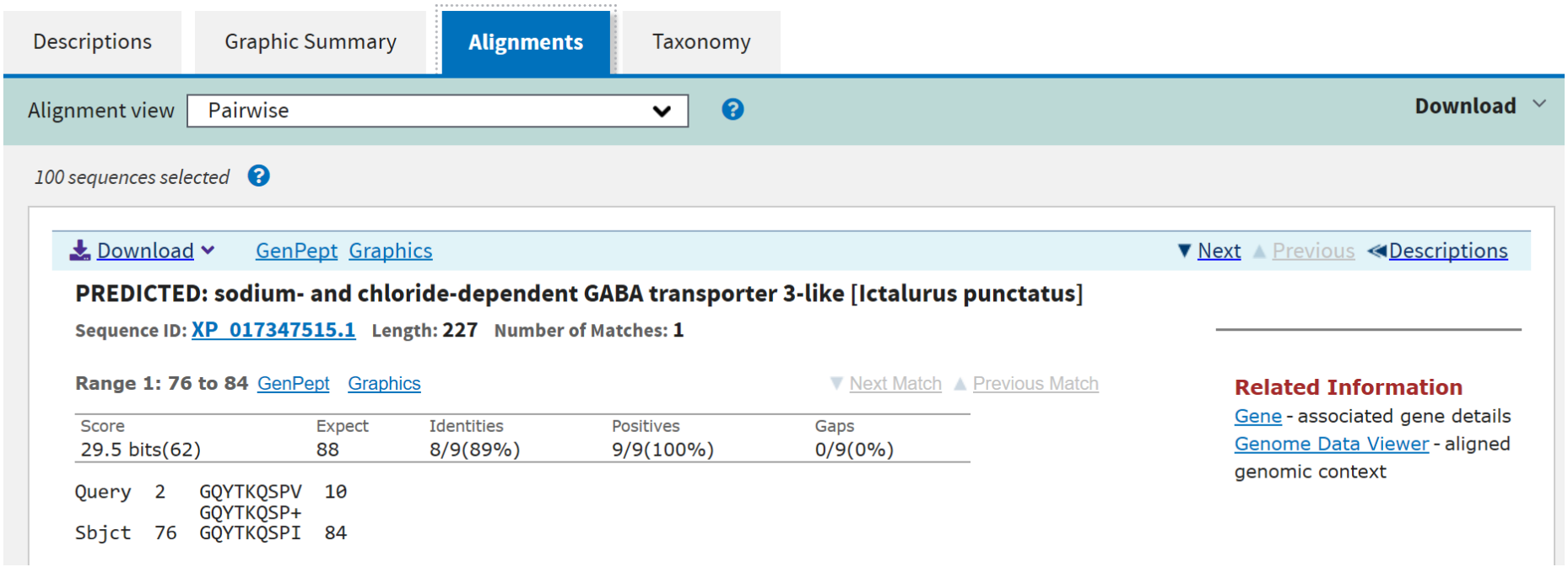
- E value - kolik krát je možné v prohledávané databázi očekávat sekvenci se stejným skóre jako má nalezená sekvence vůči hledanéé. náhodou. E-hodnota by měla být co nejmenší, ideálně blízko nuly.
- Per. Ident. - kolik procent stejných nukleotidových bází/aminokyselin se nachází v nalezené sekvenci.
- Acc. Len - délka sekvence se kterou jsme nalezli shodu.
- Accession - identifikátor sekvence, obvykle odkaz na organismus na stránky NCBI.



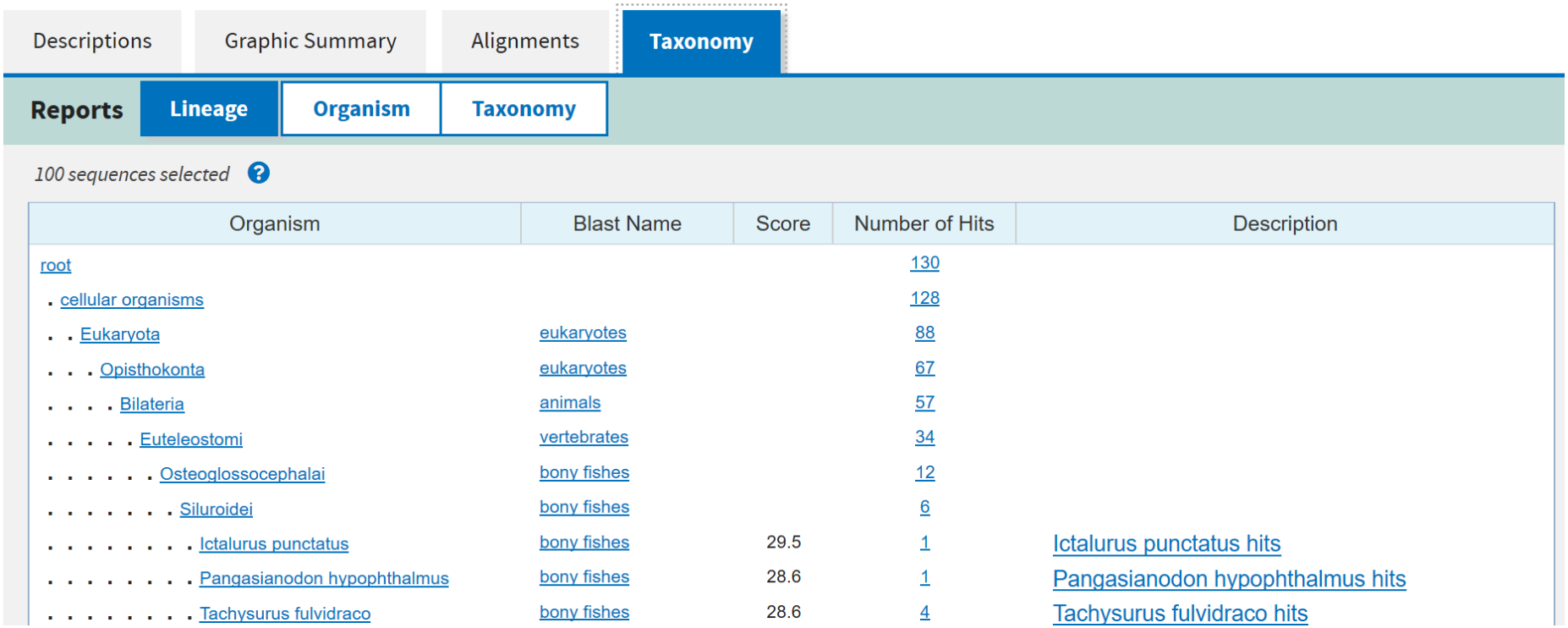
BLAST Rozhraní



BLAST Rozhraní



BLAST Rozhraní



Descriptions | Graphic Summary | Alignments | **Taxonomy**

Reports | **Lineage** | **Organism** | **Taxonomy**

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
root			130	
. cellular organisms			128	
. . Eukaryota	eukaryotes		88	
. . . Opisthokonta	eukaryotes		67	
. . . . Bilateria	animals		57	
. . . . . Euteleostomi	vertebrates		34	
. . . . . Osteoglossocephalai	bony fishes		12	
. . . . . Siluroidei	bony fishes		6	
. . . . . Ictalurus punctatus	bony fishes	29.5	1	Ictalurus punctatus hits
. . . . . Pangasianodon hypophthalmus	bony fishes	28.6	1	Pangasianodon hypophthalmus hits
. . . . . Tachysurus fulvidraco	bony fishes	28.6	4	Tachysurus fulvidraco hits

Historie BLAST algoritmů

- 1990: algoritmus BLAST1
  - Velmi rychlý algoritmus k nalezení podobných sekvencí, neumí pracovat s insercemi a delecemi.
  - Altschul et al, Basic local alignment search tool. Jeden z nejvíce citovaných vědeckých článků v bioinformatice.
- 1996/1997 návrh algoritmu BLAST2
  - BLAST2 umožňuje zpracování indelů.
  - Vyvinuty nezávisle dvě verze:
    - 1996: WU-BLAST2 (Washington University)
    - 1997: NCBI-BLAST2 (National Center for Biotechnology Information)

BLAST1 - hlavní myšlenka

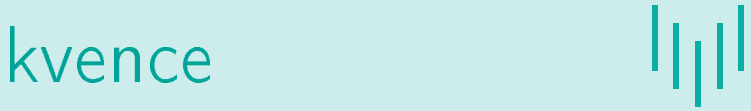
Princip

Hlavní myšlenkou v algoritmu BLAST je, že homologní(podobné) sekvence budou s velkou pravděpodobností obsahovat krátké, velmi podobné oblasti. Každá taková oblast se pak stane počátkem od kterého začneme rozšiřovat shodu na obě strany(tzv. Seed and Extend přístup).

BLAST1

- Heuristická metoda, která vyhledává lokální podobnosti bez indelů.
- V základu čtyři kroky:
  1. Předzpracování hledané sekvence.
  2. Vyhledání sekvence v databázi.
  3. Rozšíření nalezených výsledků.
  4. Ohodnocení nalezených sekvencí.

BLAST1 - předzpracování hledané sekvence



- Pro každou pozici p v hledané sekvenci nalezněte sekvenci o délce w, takovou, že její ohodnocení podobnosti je vyšší nebo rovna mezní hodnotě T.
- Těmto podobným sekvencím se říká neighbors, sousedé sekvence.
- Podobnost sekvencí určujeme pomocí matice BLOSUM nebo PAN.
- Pro DNA stanovujeme $w = 11$.
- Pro proteiny stanovujeme $w = 3$ a $T = 13$.

BLOSUM62

Some amino acid substitutions have a smaller impact on structure & function than others.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Rare; larger effect on structure/function

Common; modest effect on structure/function

negative — positive

Matrix is symmetric

Ohodnocení nukleotidy - BLAST1

BLAST Nucleotide Matrix ("Ungapped Alignment")

	A	T	C	G
A	5			
T	-4	5		
C	-4	-4	5	
G	-4	-4	-4	5

BLAST Nucleotide Matrix ("Gapped Alignment")

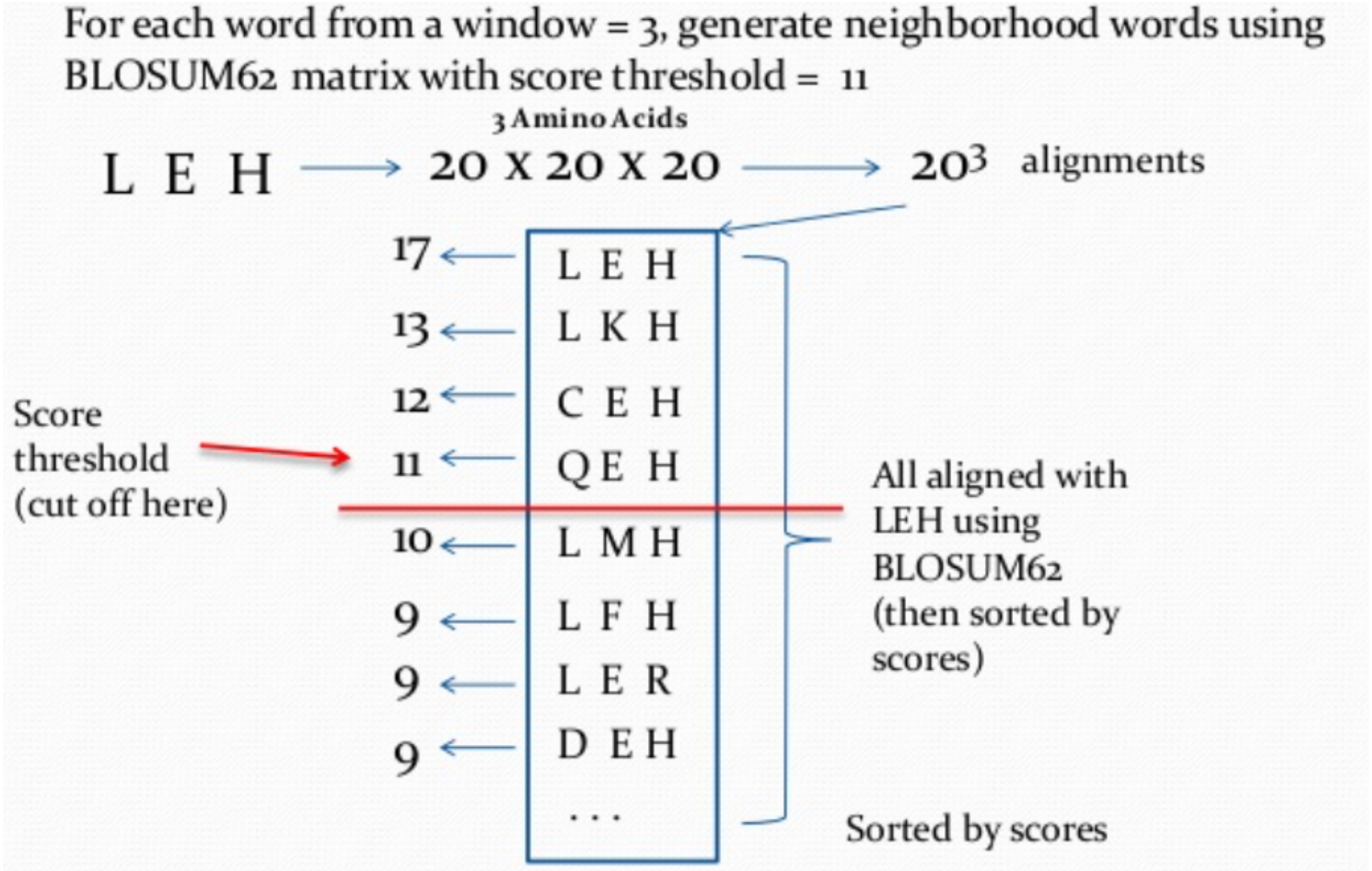
	A	T	C	G
A	1			
T	-3	1		
C	-3	-3	1	
G	-3	-3	-3	1

Rozdělení proteinové sekvence na seed words

- Hledaná sekvence má délku 11.
- Pro $w = 3$ vygenerujeme 9 slov.

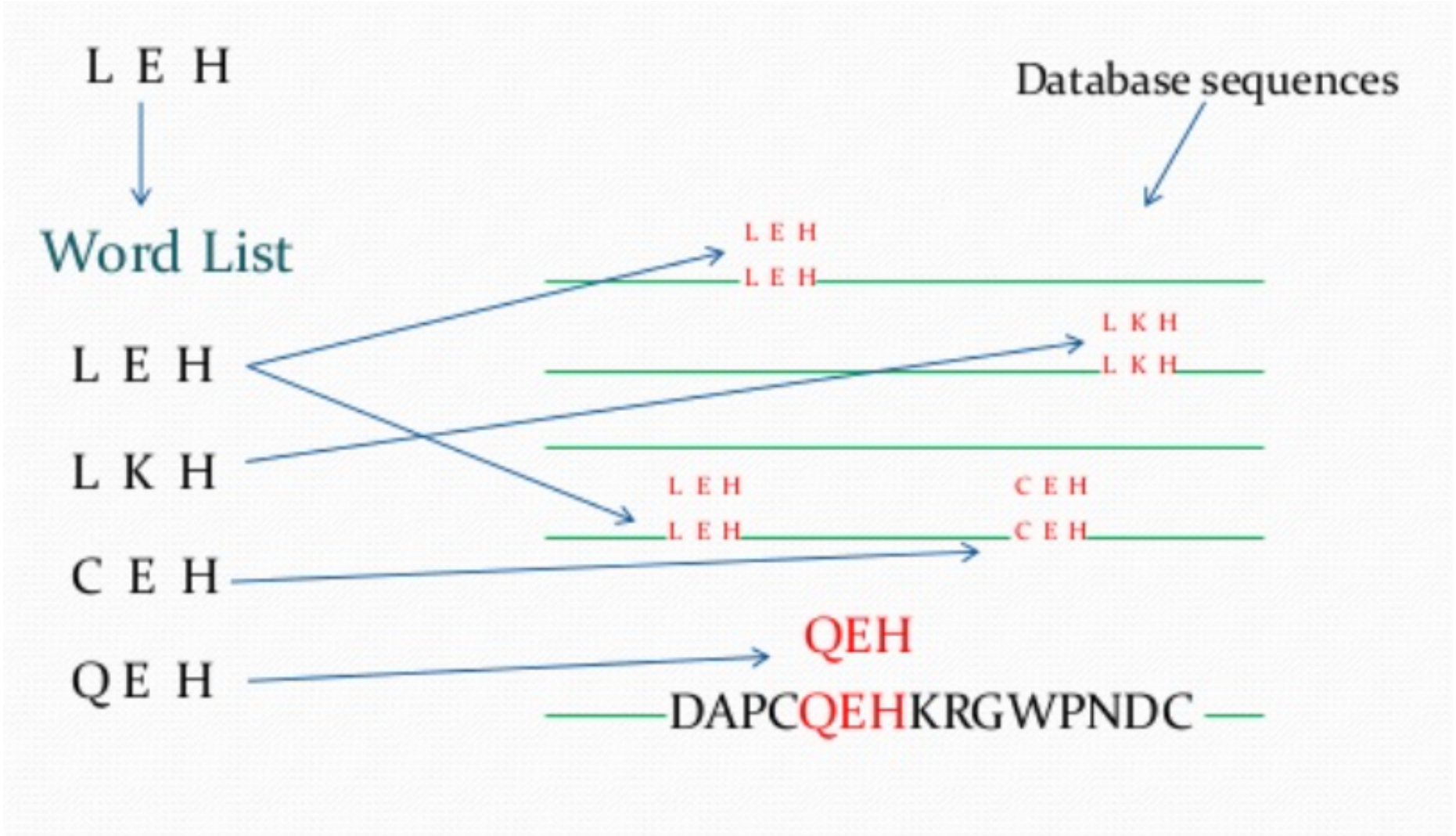
```
YANCLEHKMGS
YAN
  ANC
    NCL
      CLE
        LEH
          EHK
            HKM
              KMG
                MGS
```

BLAST1 - setřízení



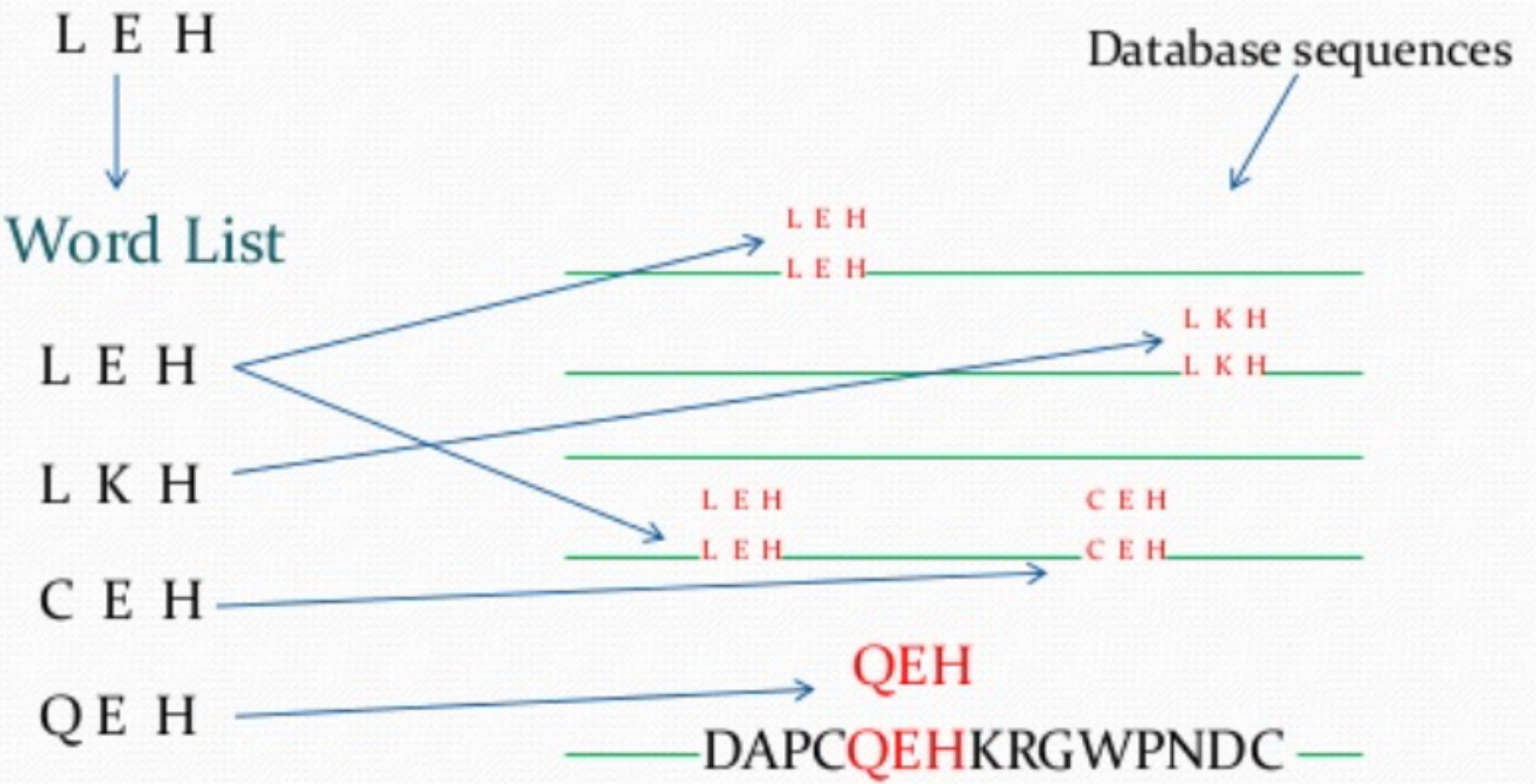
Nalezení přesné shody s databází

- Sekvence v databázi si obvykle předzpracováváme.
- Lze uložit do hashovací tabulky nebo vyhledávacího stromu.



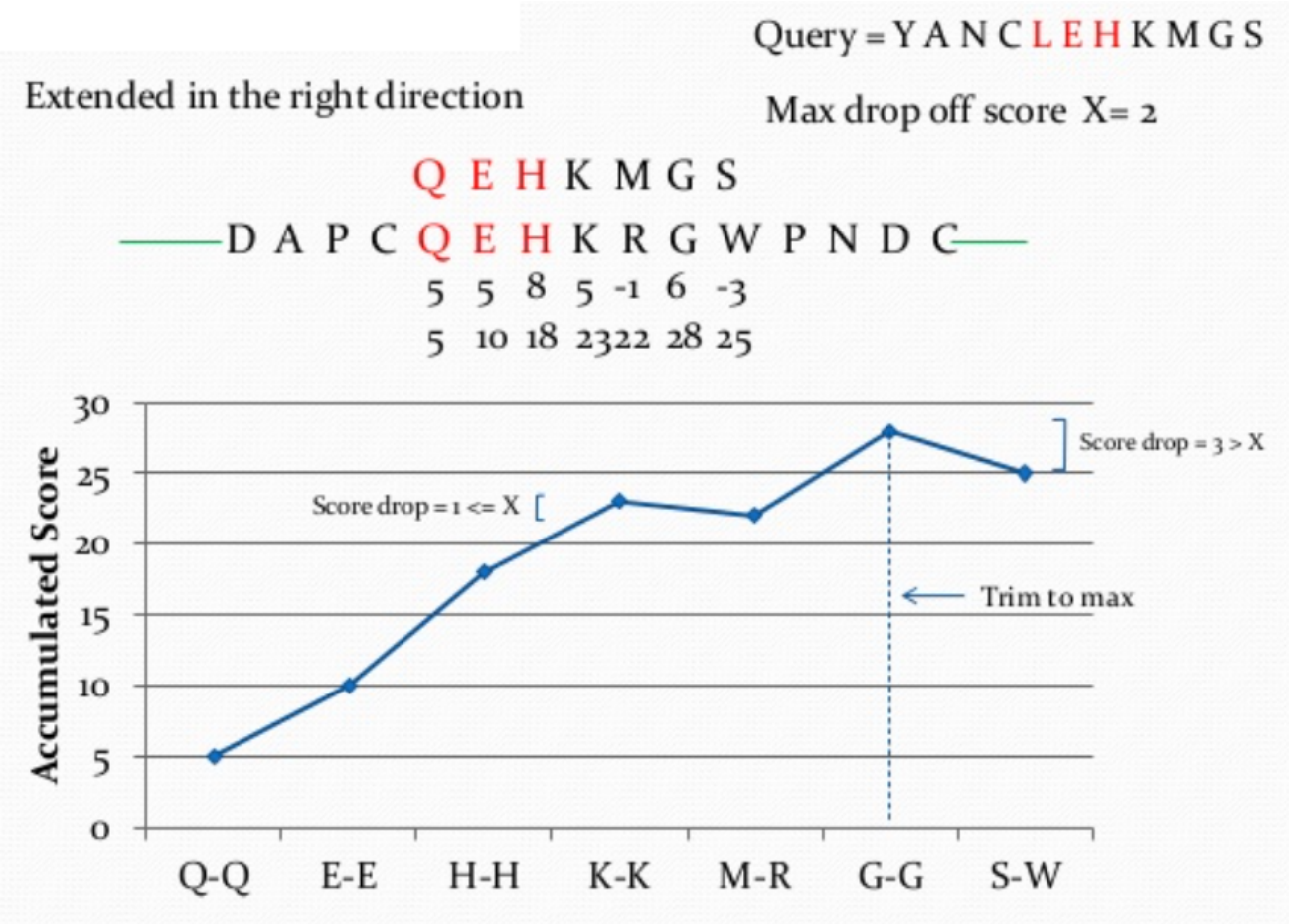
Nalezení přesné shody s databází

- Pro každou sousední sekvenci hledané sekvence, která má ohodnocení větší než T nalezneme všechny výskyty v databázi.



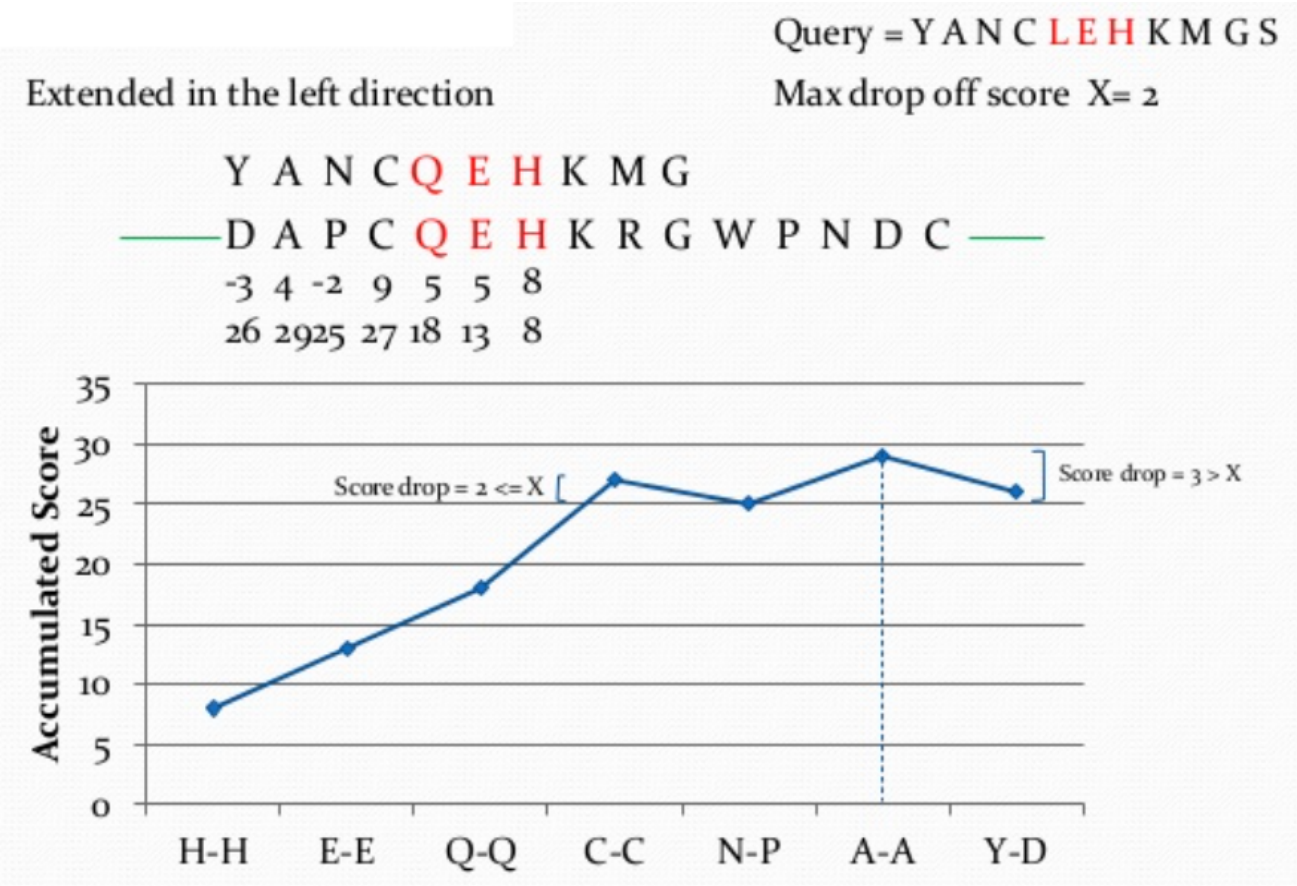
Rozšíření alignmentu

- Pro každou shodu se pokusíme o oboustranné rozšíření vůči sekvenci z databáze. X je uživatelem volená hodnota, určuje situaci, kdy zastavujeme rozširování.



Rozšíření alignmentu

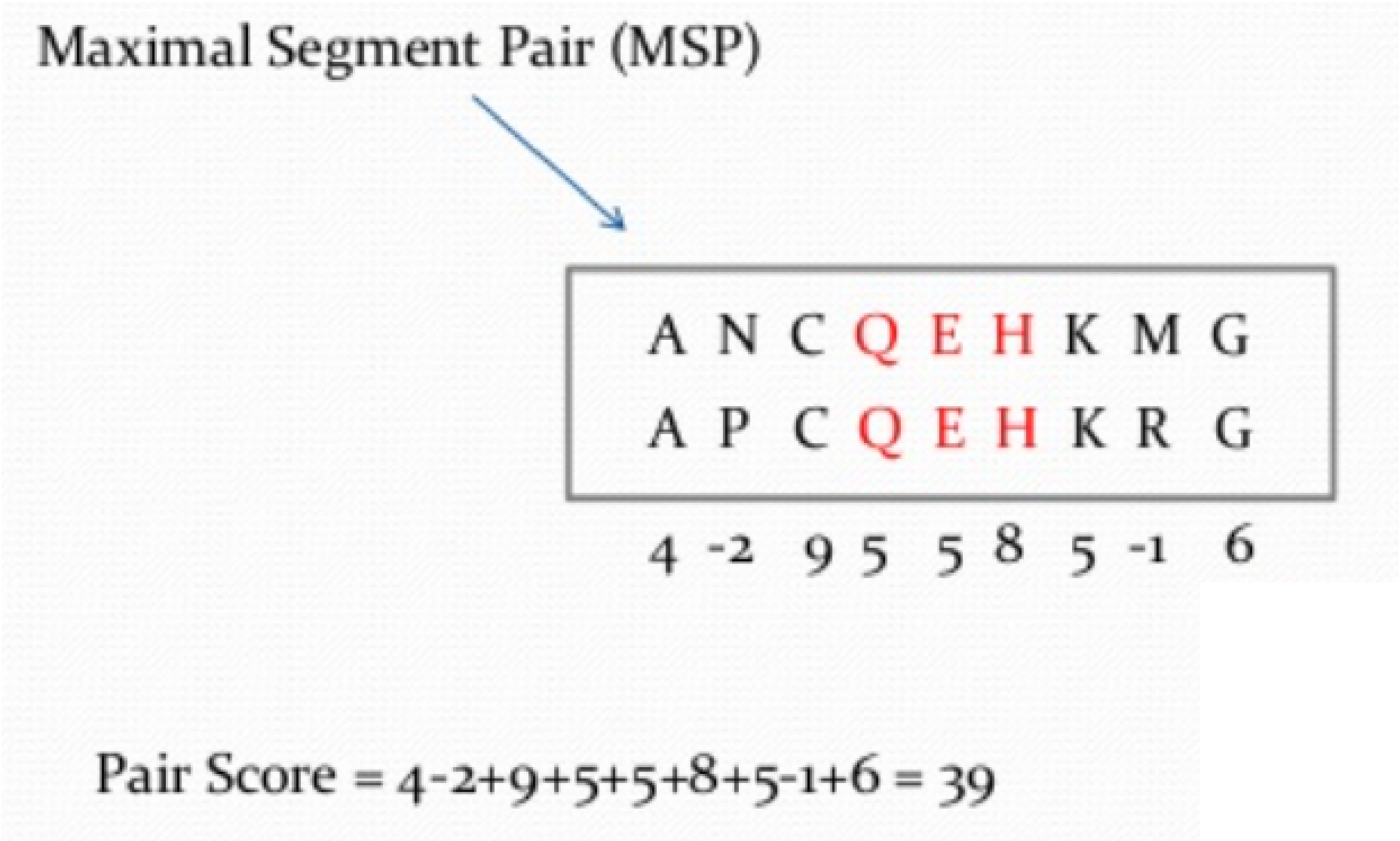
- Pro každou shodu se pokusíme o oboustranné rozšíření vůči sekvenci z databáze. X je uživatelem volená hodnota, určuje situaci, kdy zastavujeme rozšiřování.



Maximal segment pairs

- Rozšířená sekvence, jejíž skóre je vyšší než hodnota S je vrácena jako výsledek.
- Tyto sekvence označujeme anglickým termínem High Scoring Segment Pair (HSP), dvojice sekvencí s vysokým skóre.
- HSP s nejvyšším skóre je potom označen jako maximální dvojice segmentů - Maximal segment pair (MSP).

Maximal segment pairs



Určení statistické významnosti

Základní otázky

- Je skóre dostatečně vysoké abychom mohli usuzovat o homologii sekvenci?
- Jsou skóre získané namapováním s náhodnou sekvencí vyšší, než naše nalezené skóre?
- Jaký je očekávávaný počet náhodných sekvencí z databáze se skóre vyšším než naše skóre?

Otázka

Mějme nukleotidovou sekvenci s_1 o délce n. Mějme databázi sekvencí s celkovou délkou m. Pokud budeme předpokládat, že každý nukleotid se v databázi vyskytuje nezávisle a stejnou pravděpodobností, kolikrát bychom očekávali, že danou sekvenci najdeme.

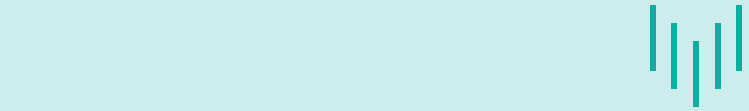
Otázka

Mějme nukleotidovou sekvenci s_1 o délce n. Mějme databázi sekvencí s celkovou délkou m. Pokud budeme předpokládat, že každý nukleotid se v databázi vyskytuje nezávisle a stejnou pravděpodobností, kolikrát bychom očekávali, že danou sekvenci najdeme.

$$p(s_1) = (\frac{1}{4})^n = \frac{1}{4^n}$$

$$E(s_1) = p(s_1) \times m$$

Intuitivní přístup



Otázka

Mějme nukleotidovou sekvenci s_1 o délce n. Pokud budeme předpokládat, že každý nukleotid se v databázi vyskytuje nezávisle a stejnou pravděpodobností, jak velkou databázi potřebujeme abychom měli očekávaný počet sekvencí roven 1?

Otázka

Mějme nukleotidovou sekvenci s_1 o délce n. Pokud budeme předpokládat, že každý nukleotid se v databázi vyskytuje nezávisle a stejnou pravděpodobností, jak velkou databázi potřebujeme abychom měli očekávaný počet sekvencí roven 1?

$$m = \frac{1}{p(s_1)}$$

E-Value

Počet MSP s podobným nebo vyšším skóre, který bychom očekávali, že nalezeneme na základě pravděpodobnosti, když procházíme databázi určité velikosti.

Například: pokud je E-Value rovna 1 pro určitou MSP se skóre S, můžeme říci, že v databázi o určité velikosti bychom očekávali, že nalezneme 1 shodu s podobným skóre.

Základní vlastnosti

- Pokud známe statistické rozdělení hodnot skóre vzájemně nesouvisejících sekvencí, pak budeme schopni vyhodnotit statickou významnost.

Předpoklady

- Ohodnocovací tabulka musí být zkostruována tak, aby náhodně namapované sekvence měly záporné skóre.
- Délky mapovaných sekvencí jsou dostatečně veliké
- Rozdělení nukleotidů/aminokyselin v sekvenci a databázi je shodné
- Musí být možné obdržet kladné skóre. Alespoň jeden prvek v ohodnocovací tabulce musí být kladný.

E-Value výpočet - bit score

$$S' = \frac{\lambda S - \log(K)}{\log(2)}$$

- S - skóre z alignmentu.
- λ, K - hodnoty závislé na ohodnocovacím schématu a konkrétních sekvencí v databázi.
- $\log(2)$ ve jmenovateli rovnice převádí výslednou hodnotu na logaritmus o základu 2. Tedy veličina se měří v bitech.

Lze interpretovat, jako potřebnou velikost databáze, ve které jsme schopni najít shodu s hledanou sekvencí na základě náhody. Pokud je databáze dosti velká, pak danou sekvenci s vysokou pravděpodobností najdeme. S každým nárůstem velikosti bitového skóre se potřebná velikost databáze zdvojnásobí. Bit skóre je porovnatelné mezi různými dotazy i nad jinými databázemi.

E-Value výpočet

$$EV \approx mn2^{-S'} = m\frac{n}{2^{S'}}$$

- S' - bit score.
- m - délka vstupní sekvence.
- n - součet délek všech sekvencí v databázi.

E-Value interpretace

- Hodnoty E-Value menší než 10^{-4} značí významnou podobnost.
- Hodnoty E-Value větší než 10^{-4} by měly být ručně překontrolovány.

Děkuji za pozornost

Michal Vašinek

VŠB – Technická univerzita Ostrava

FEI/EA404

michal.vasinek@vsb.cz

7. listopadu 2023

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY