

# Indexy a stromy

INDEX TYPU POLE

(vhodný pro "bohatý" index, fixní délku, časově kritické aplikace)

AAAA	12	325	65987	124589
AAAC	13	4278	23001	
AAAG	326	9989	44032	199235
...				

INDEX TYPU STROM

(vhodný pro "děravý" index, variabilní délku, je flexibilní)

		===	AAAC
AAA		===	AAAG
		===	AAAA

# Sufixové pole

```
0 1 2 3 4 5 6 7 8 9
C G T G T G A C G *
```

sekvence

sufixy (přípony)

pozice

```

                                * 9
                                G * 8
                                C G * 7
                                A C G * 6
                                G A C G * 5
                                T G A C G * 4
                                G T G A C G * 3
                                T G T G A C G * 2
                                G T G T G A C G * 1
                                C G T G T G A C G * 0
```

# Sufixové pole

0	1	2	3	4	5	6	7	8	9	
C	G	T	G	T	G	A	C	G	*	sekvence
3	7	9	6	8	5	1	2	4	0	poradi v sufixovém poli

sufixy abecedně	pozice
-----------------	--------

								*	9										
						A	C	G	*	6									
							C	G	*	7									
C	G	T	G	T	G	A	C	G	*	0									
								G	*	8									
						G	A	C	G	*	5								
							G	T	G	A	C	G	*	3					
								G	T	G	T	G	A	C	G	*	1		
										T	G	A	C	G	*	4			
											T	G	T	G	A	C	G	*	2

9	6	7	0	8	5	3	1	4	2	sufixové pole
---	---	---	---	---	---	---	---	---	---	---------------

# Sufixové pole

0 1 2 3 4 5 6 7 8 9

C G T G T G A C G \*      sekvence

3 7 9 6 8 5 1 2 4 0      poradi v sufixovém poli

sufixy abecedně

pozice



*	9	0
A C G *	6	0
C G *	7	2
C G T G T G A C G *	0	0
G *	8	1
G A C G *	5	1
G T G A C G *	3	3
G T G T G A C G *	1	0
T G A C G *	4	2
T G T G A C G *	2	0

9 6 7 0 8 5 3 1 4 2      sufixové pole

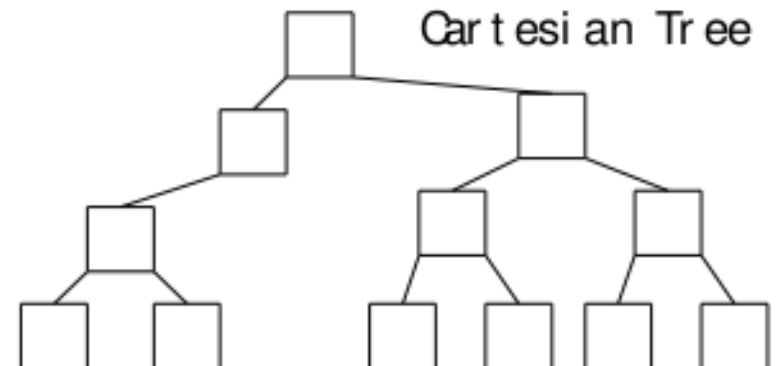
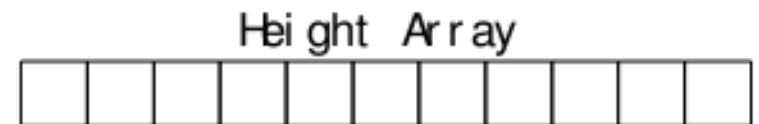
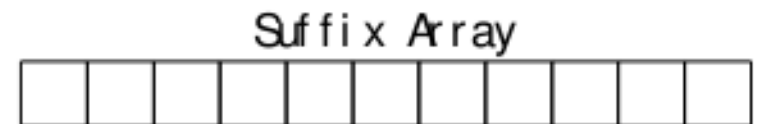
0 0 2 0 1 1 3 0 2 0      delka spol. prefixu (height)

# Sufixové pole

	m	i	s	s	i	s	s	i	p	p	i
i										0	0
p									0	0	1
p									0	0	0
i								0	0	1	0
s					0	0	1	1	2	1	
s				0	0	0			2	3	2
i			0	0	1				0	1	2
s		0	0	1	0				1	1	2
s	0	0	0						0	2	2
i	0	0	1		0	1			0	1	2
m	0	1	1	2	1	1	2	1	1	2	3

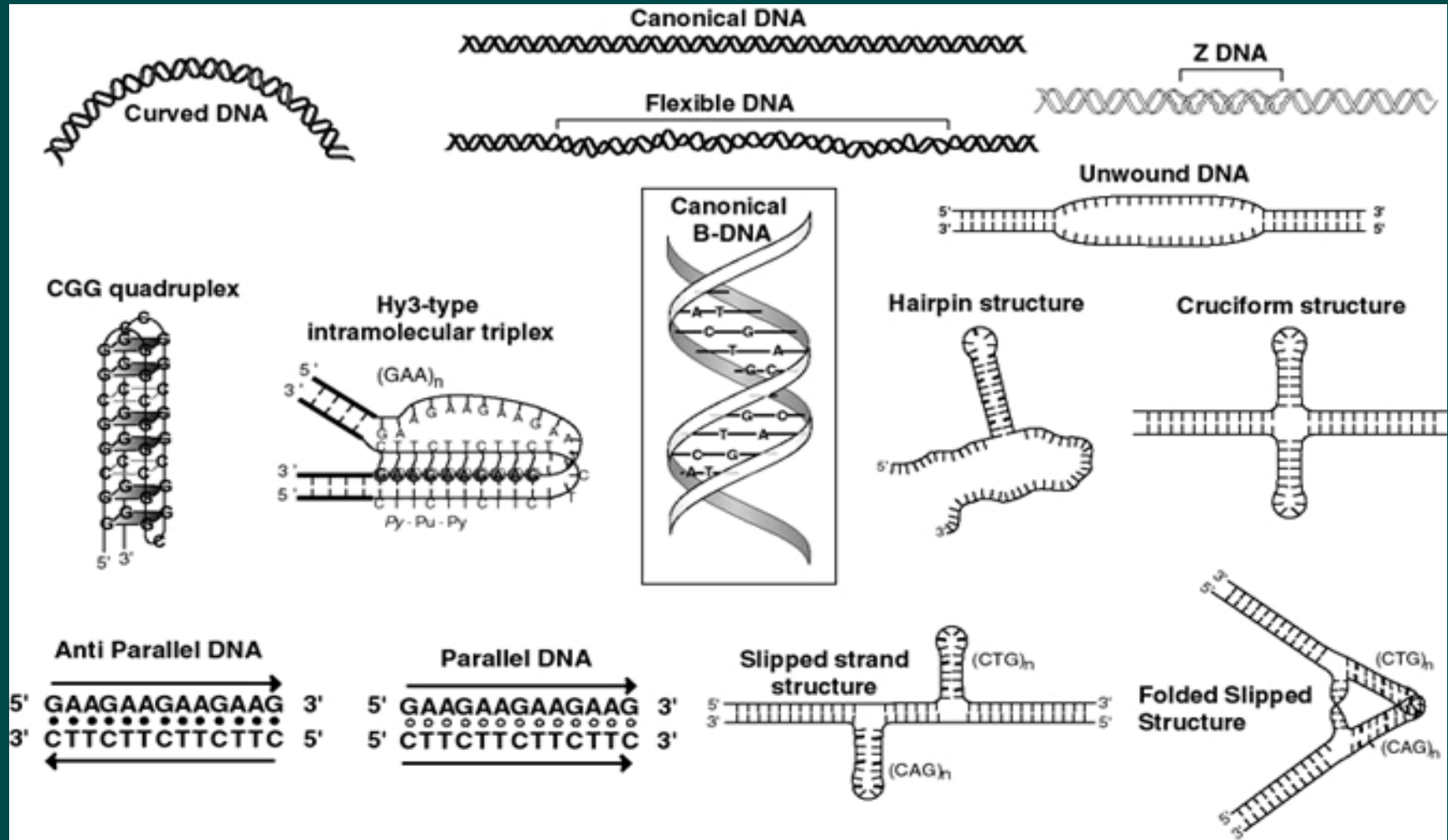
-  Longest Common Prefix
-  Neighboring cells with worse scores

a)



b)

# Sufixové pole



# Stupeň podobnosti dvou sekvencí

SEKVENCE A:       MASAQSFYLL

SEKVENCE B:       MASGQWLLAS

Které oblasti A a B jsou si nejvíce podobné ?

Jsou si A a B víc podobné než A a C ?

Která ze sekvencí  $X_1, \dots, X_n$  je nejpodobnější A ?

Jaká je pravděpodobnost výskytu dané podobnosti v náhodné sekvenci ?

Existují známé homology sekvence A ?

Jaká je funkce genu/proteinu A ?

# Stupeň podobnosti dvou sekvencí

## IDENTITA

MASAQSFYLL

|||||

MASAQSFYLL

## SUBSTITUTE

MASAQSFYLL

|||||:|

MASAQSWYLL

MASAQSFYLL

||||| ||

MASAQSTYLL

## INZERCE/DELECE

MASAQSFYLL

||||| ||

MASAQS-YLL

## TRANSPOZICE

MASAQSFYLL

|||| ||

MASAQFSYLL





# Stupeň podobnosti dvou sekvencí

Netriviální hodnocení substitucí  
u proteinů  
(matice PAM250)

A	2																				
R	-2	6																			
N	0	0	2																		
D	0	-1	2	4																	
C	-2	-4	-4	-5	12																
Q	0	1	1	2	-5	4															
E	0	-1	1	3	-5	2	4														
G	1	-3	0	1	-3	-1	0	5													
H	-1	2	2	1	-3	3	1	-2	6												
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5											
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6										
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5									
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6								
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9							
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6						
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2					
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3				
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17			
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10		
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	

# Stupeň podobnosti dvou sekvencí

		A	G	A	T	A
	0	-2	-4	-6	-8	-10
A	-2	2	0	-2	-3	-5
G	-4	-1	?			
T	-6					
C	-8					
A	-10					

INDEL=-2 IDENT=2 SUBST=-1

AGATA  
|| | S=4  
AGTCA

# BLAST (basic local alignment search tool)

Co když jsou sekvence dlouhé a máme jich několik milionů ?

DP nestačí, výpočty trvají příliš dlouho. Alternativou výpočtu by byl předpočítaný soubor podobnosti různých slov v databázi. Problémem indexu je, že je pro dlouhá slova nezvladatelný objemově. Existuje např.

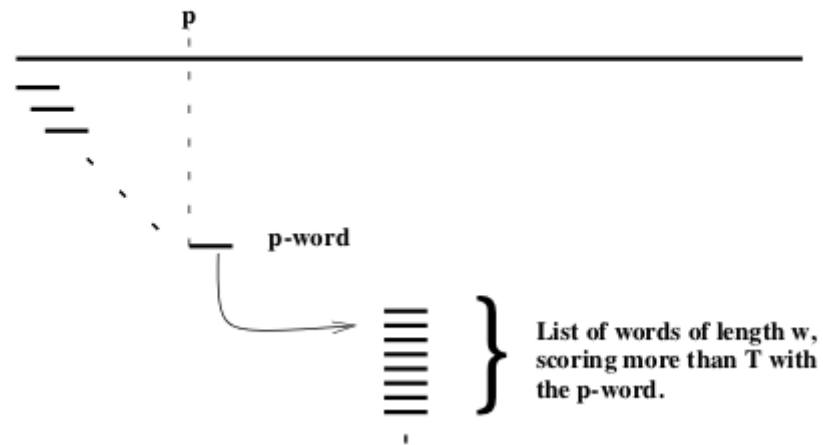
$$20^8 = 25\,600\,000\,000$$

různých uspořádání osmi aminokyselin v řetězci, několik způsobů hodnocení podobnosti atd.

Kompromisem je heuristické řešení. Nalezení tzv. “seeds”, výskytu krátkých řetězců (2-4 aminokyseliny, 7-11 nukleotidů) a hledání podobnosti algoritmem DP jenom v jejich blízkosti.

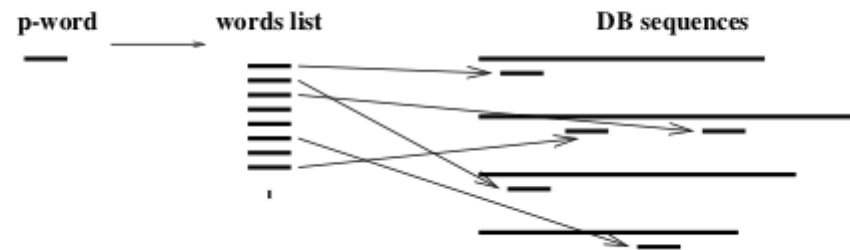
# BLAST (basic local alignment search tool)

**A: For each position  $p$  of the query, find the list of words of length  $w$  scoring more than  $T$  when paired with the word starting at  $p$ :**



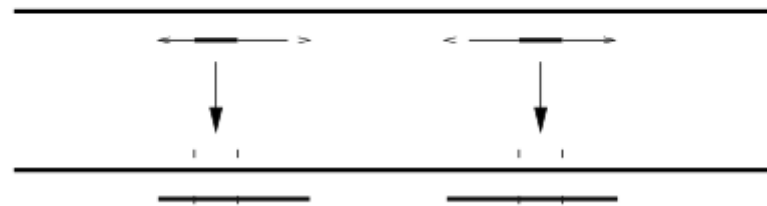
# BLAST (basic local alignment search tool)

**B: For each words list, identify all exact matches with DB sequences:**

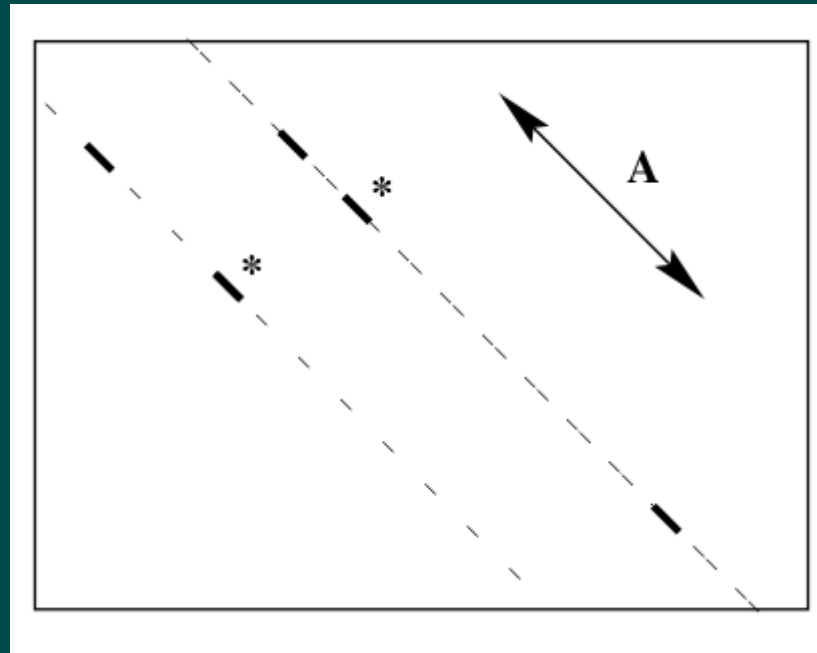


# BLAST (basic local alignment search tool)

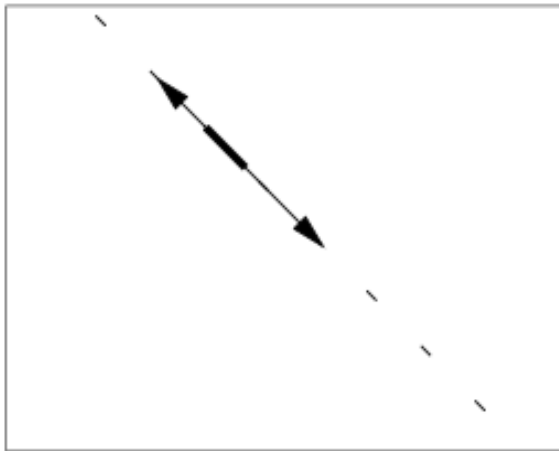
**C:** For each word match («hit»), extend ungapped alignment in both directions. Stop when  $S$  decreases by more than  $X$  from the highest value reached by  $S$ .



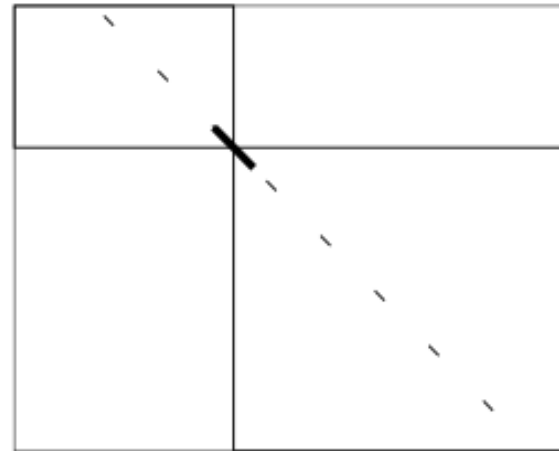
# BLAST (basic local alignment search tool)



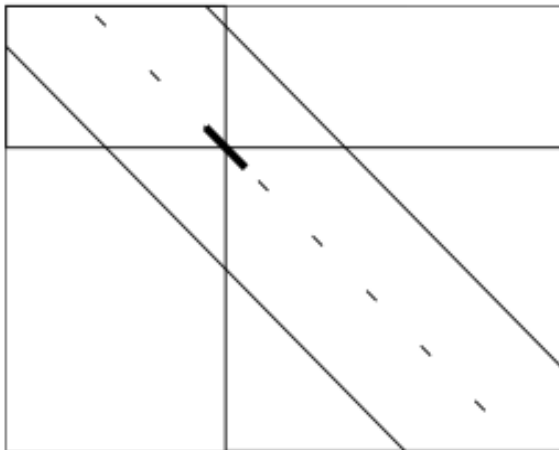
# BLAST (basic local alignment search tool)



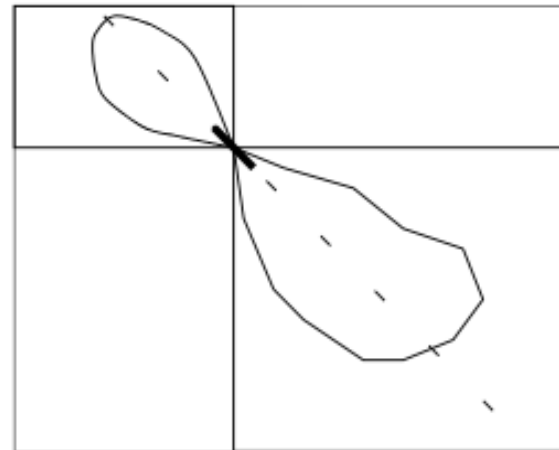
**Ungapped extension**



**Gapped extension by full DP**



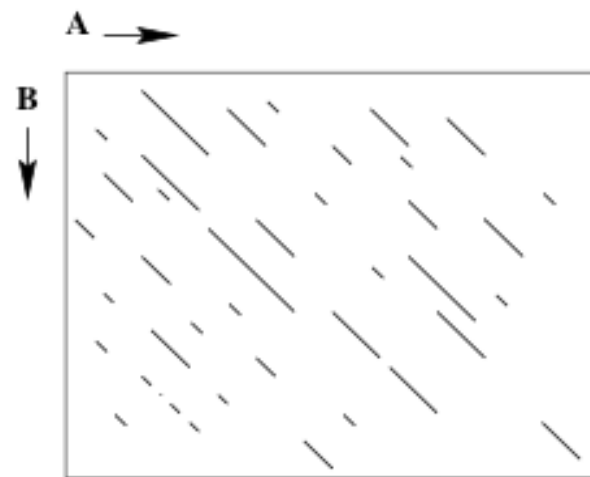
**Gapped extension by «banded DP»**



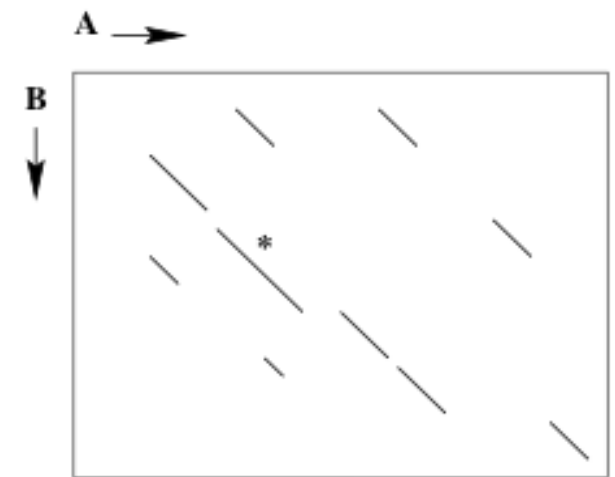
**Gapped extension by «score-limited DP»**



# FASTA

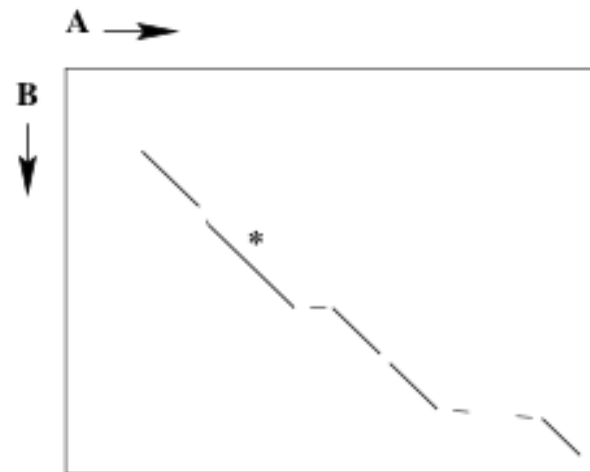


Identify all k-tuple matches



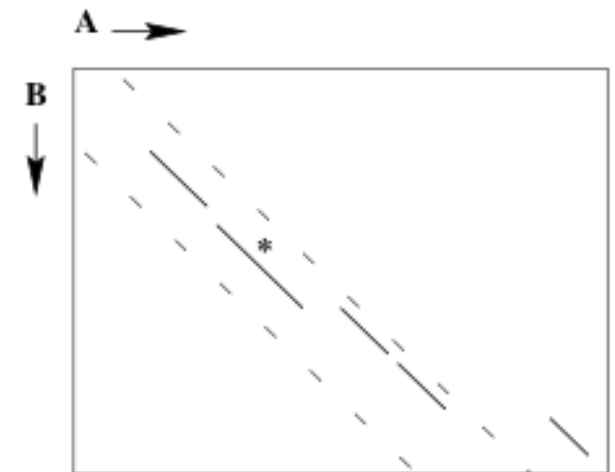
Re-score the 10 best scoring regions using a scoring matrix

→ Init1 score



Apply joining procedure

→ Initn score



Apply limited DP

→ Opt score

# BLAST (basic local alignment search tool)

P-VALUE . . .  $P(\text{score} > S)$

PRAVDĚPODOBNOST VÝSKYTU PODOBNOSTI VĚTŠÍ  
NEŽ S V NÁHODNÝCH SEKVENCÍCH URČITÉ DÉLKY

$$P(\text{MSP}(M, N) > S) = 1 - \exp(-Kmn \cdot \exp(-\lambda \cdot S))$$

E-VALUE

OČEKÁVANÝ POČET PODOBNOSTÍ KDE  $\text{score} > S$

$$Kmn \cdot \exp(-\lambda \cdot S)$$

# BLAST (basic local alignment search tool)

## PAM150

Percent Accepted Mutations

Substituční matice odvozena z předpokladu 150 mutací na 100 pozic sekvence

## BLOSUM65

BLOck SUBstitution Matrix

Substituční matice odvozena ze sekvencí se 65% identitou

PAM120  $\Leftrightarrow$  BLOSUM80

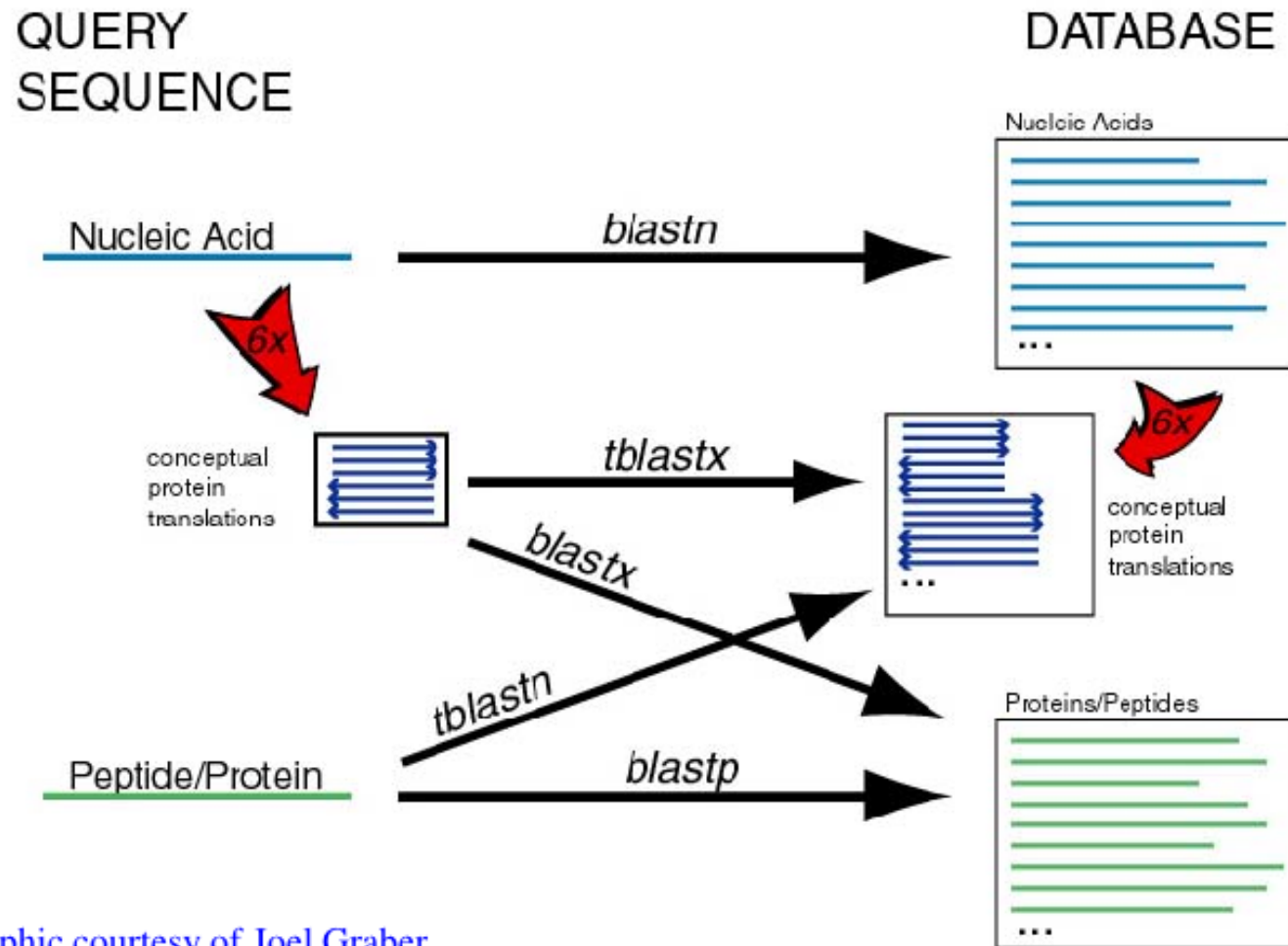
PAM250  $\Leftrightarrow$  BLOSUM45

---

---

# BLAST (basic local alignment search tool)

## Types of BLAST:



Graphic courtesy of Joel Graber.

# BLAST (basic local alignment search tool)

```
>gi|50757596|ref|XP_425354.1|similar to protein kinase  
Length = 613
```

```
Score = 50.4 bits (119), Expect(2) = 2e-17
```

```
Identities = 26/54 (48%), Positives = 36/54 (66%), Gaps = 1/54 (1%)
```

```
Query: 740 YVMVLEYANEGNLREYLEK-KFDTLQWENKIQMALDITRGLLCLHSRNI IHRDL 582  
      Y +V EY +EG+LR YL K + +L + I ALDI RG+ +HS+ +IHRDL  
Sbjct: 250 YCVVTEYLSSEGLRAYLHKLERKSLPLQKLI AFALDIARGMEYIHSQGV IHRDL 303
```

# BLAST (basic local alignment search tool)

BLAST (NCBI-BLAST WU-BLAST)  
BLASTN BLASTP BLASTX TBLASTN TBLASTX  
MEGABLAST  
PSI-BLAST  
PHI-BLAST  
SNPBLAST  
BLASTZ

# Pattern Hunter

**PATTERN HUNTER**

**BLAST**

MODEL SLOVA

110100110010101111

111111111111

ZÁVISLOST POZIC

110100110010101111

111111111111

110100110010101111

111111111111

5 10

110100110010101111

111111111111

5 9

110100110010101111

111111111111

5 8

110100110010101111

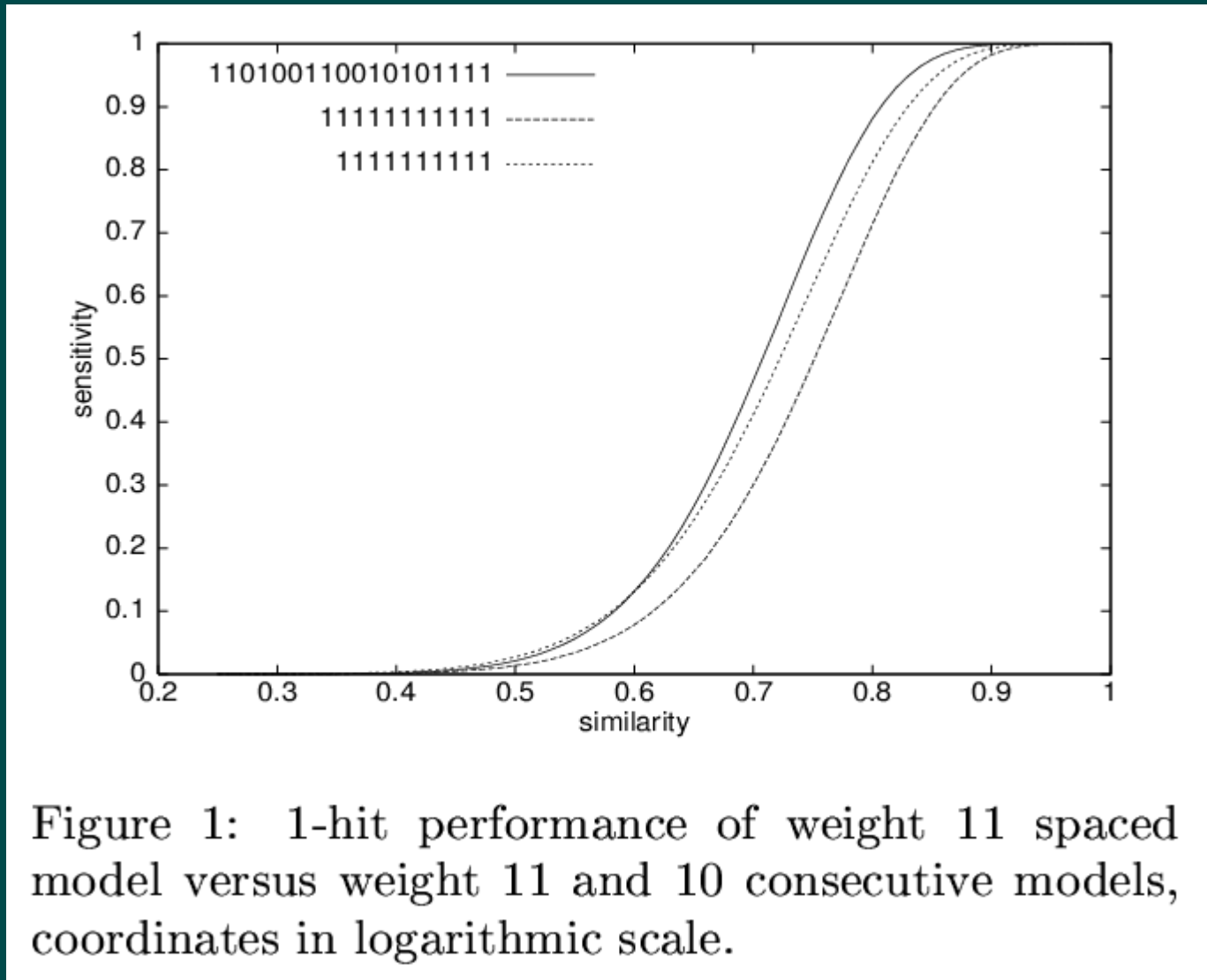
111111111111

4 7

PŘEKRYV



# Pattern Hunter





# Pattern Hunter

Seq1	Size	Seq2	Size	PH	PH2	MB28	Blastn
<i>M. pneumoniae</i>	828K	<i>M. genitalium</i>	589K	10s/65M	4s/48M	1s/88M	47s/45M
<i>E. coli</i>	4.7M	<i>H. influenza</i>	1.8M	34s/78M	14s/68M	5s/561M	716s/158M
<i>A.thaliana</i> chr 2	19.6M	<i>A.thaliana</i> chr 4	17.5M	5020s/279M	498s/231M	21720s/1087M	∞
<i>H. sapiens</i> chr 22	35M	<i>H. sapiens</i> chr 21	26.2M	14512s/419M	5250s/417M	∞	∞

# Pattern Hunter

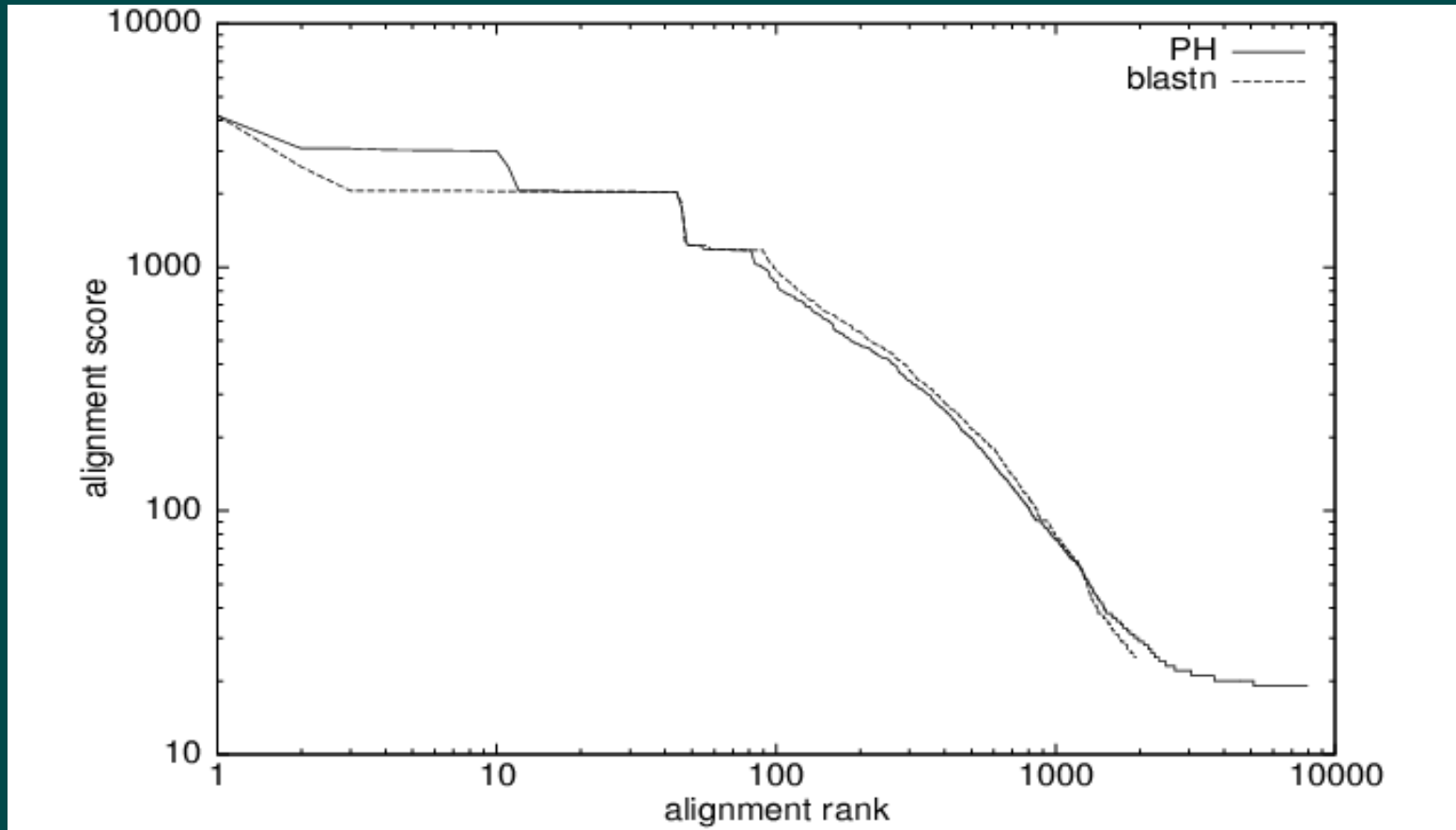


Figure 5: Input: *H. influenza* and *E. coli*. Run times are shown in Figure 7. PatternHunter produces better quality output than Blastn while running 20 times faster.

# Pattern Hunter

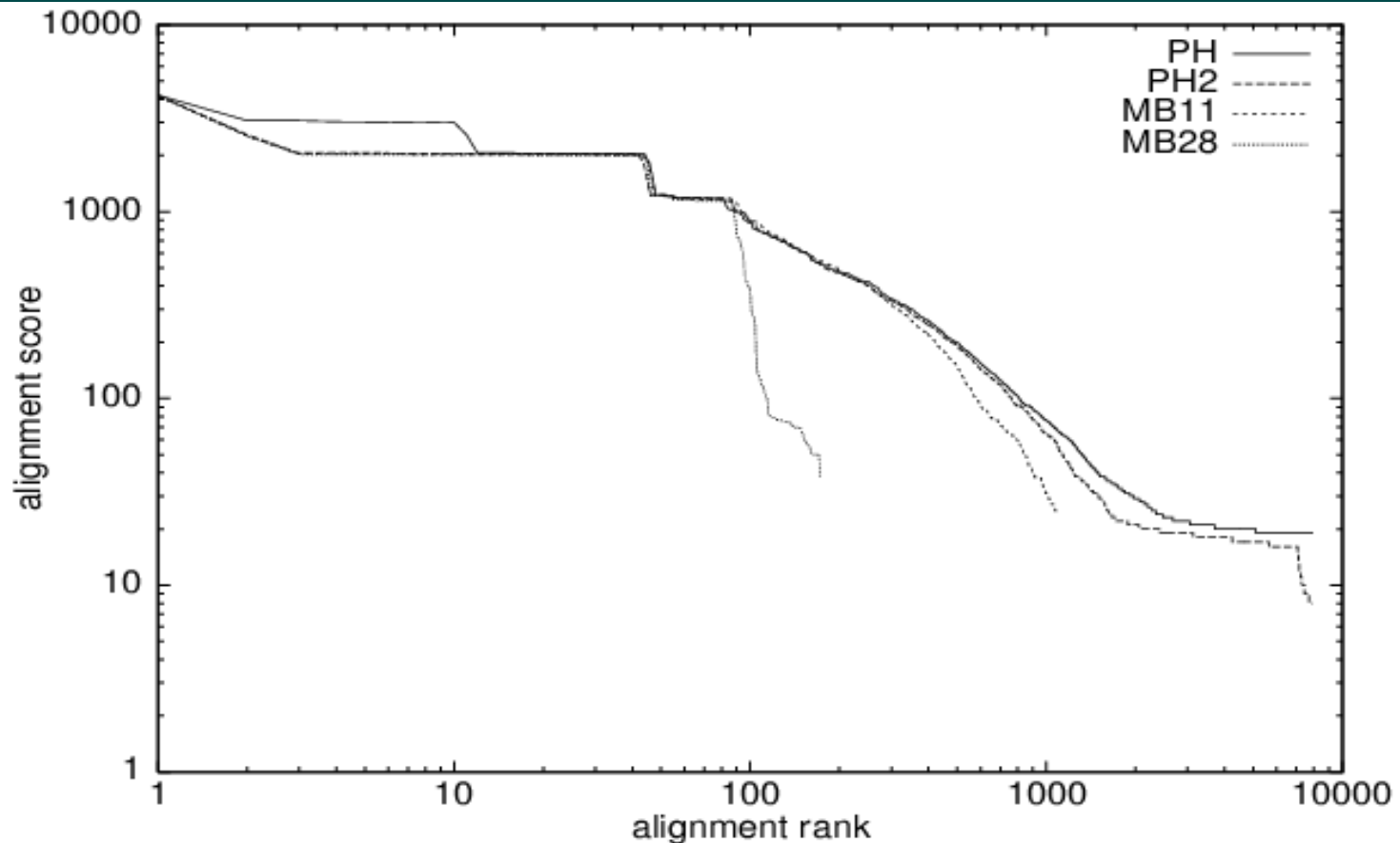


Figure 4: Input: *H. influenza* and *E. coli*. Run times are shown in Figure 7. Score is plotted as a function of the rank of the alignment, with both axes logarithmic. MegaBlast (MB28) misses over 700 alignments

# Primex

INDEX

$4^w + n \rightarrow 150\text{Mbps}$  ( $w=12$ )  $\rightarrow 167\text{M}$  ( $\sim 650\text{MB}$ )

FILTRACE

**ACGAGATGACGATGACGATGCGAT**

DP (N-W)

Probíhá na omezeném vzorku sekvencí

# Primex

**ATAGTAGGTCCGTCGATA**

18 bp -> 3 slova po 6bp

nastavením  $m1=1$  chyby na slovo  
nalezneme spolehlivě v nejhorším

**GTATTAGGTACGTTGACA**

i řetězec s 5 chybami

Sekvence se 6 chybami už nemusí být nalezena:

**GTATTAGATACGTTGACA**

nenalezena

**GTATTAGGTACGTTGACG**

nalezena

Rozdělení vyhledávacího řetězce na menší segmenty urychlí výpočty tím, že u kratších lze efektivně využít rychlého indexu celé genomové databázy.

# Primex

PROGRAM	TOTAL	MISMATCHES						SEARCH TIME	
		0	1	2	3	4	5		6
BLAST	162	1	0					12 s	
BLAST-O	2	1						5 s	
FASTA	72	1	0	4	13	23	17	12	53 s
FASTA-O	1	1							19 s
BLAT	0	0							80 s
SSAHA	0	0							71 s
SSAHA-O	3	1							10 s
CGC FP-O	1	1							10 s
EMBOSS	983	1	0	5	104	86		8	18 s
EMBOSS-O	1	1							14 s
TACG	1	1							49 s
AGREP	1204	1	0	5	100	779	3632		34 s
AGREP-O	1	1							2 s
PRIMEX	214	1	0	14	199				56 s
PRIMEX-S (2,5)	12140	1	0	14	199	1686	10240		19 s
PRIMEX-S (2,4)	1900	1	0	14	199	1686			4 s
PRIMEX-S (2,3)	214	1	0	14	199				1 s
PRIMEX-S (2,2)	15	1	0	14					< 1 s
PRIMEX-S (2,1)	1	1	0						<< 1 s
PRIMEX-S0 (0,0)	1	1							<< 1 s

Výkony vybraných vyhledávacích programů při hledání výskytu sekvence oligonukleotidu AAAAATGATCAATTACAT v genomu Arabidopsis thaliana (cca 100Mbp). Příponou -O jsou označeny programy, které byly nastaveny s nejmenší citlivostí. Přípona -S označuje programy, které v momentě dotazu běžely jako server.

# Primex + Virtual PCR

Vyhledané oligonukleotidy lze použít pro simulaci PCR reakce

VPCR Virtual PCR



# BLAT

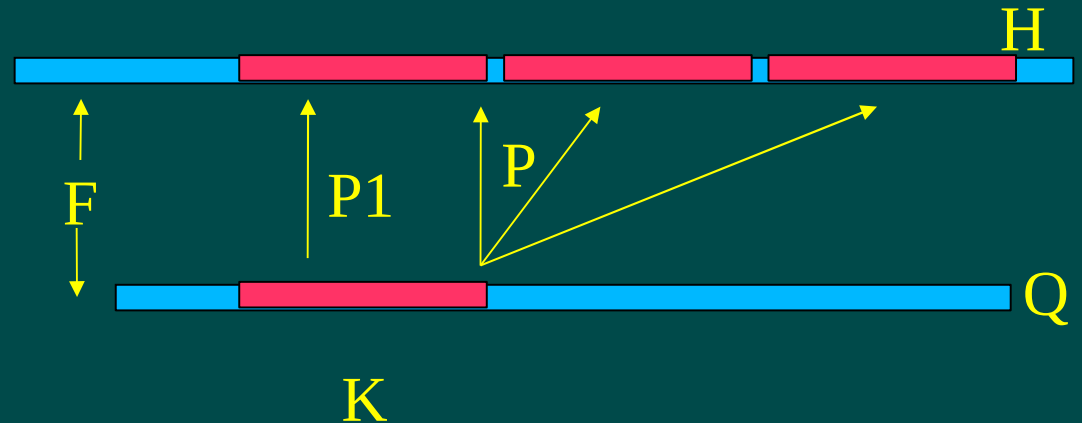
$$P1 = M^K$$

$$T = \text{int}(H/K)$$

$$P = 1 - (1 - P1)^T$$

$$P = 1 - (1 - M^K)^T$$

$$F = (Q - K + 1) * (G/K) * (1/A)^K$$



P1 – pravdepodobnosť shody s k-merem

P – pravdepodobnosť existence aspon jednej takej shody

F – pravdepodobnosť nahodného vyskytu shody medzi H a Q

K – dĺžka k-meru

H – dĺžka úseku podobnosti (několik 100 bp)

M – zhoda medzi sekvencami (%identity/100)

G – veľikost databázy

Q – veľikost vyhľadávacj sekvence

A – veľikost abecedy



# BLAT

**Table 3.** Sensitivity and Specificity of Single Perfect Nucleotide K-mer Matches as a Search Criterion

	7	8	9	10	11	12	13	14
<b>A.</b> 81%	0.974	0.915	0.833	0.726	0.607	0.486	0.373	0.314
83%	0.988	0.953	0.897	0.815	0.711	0.595	0.478	0.415
85%	0.996	0.978	0.945	0.888	0.808	0.707	0.594	0.532
87%	0.999	0.992	0.975	0.942	0.888	0.811	0.714	0.659
89%	1.000	0.998	0.991	0.976	0.946	0.897	0.824	0.782
91%	1.000	1.000	0.998	0.993	0.981	0.956	0.912	0.886
93%	1.000	1.000	1.000	0.999	0.995	0.987	0.968	0.957
95%	1.000	1.000	1.000	1.000	0.999	0.998	0.994	0.991
97%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
<b>B. K</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>
F	1.3e+07	2.9e+06	635783	143051	32512	7451	1719	399

(A) Columns are for K sizes of 7–14. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated from equation 3 assuming a homologous region of 100 bases. The larger the value of K, the fewer homologies are detected.

(B) K represents the size of the perfect match. F shows how many perfect matches of this size expected to occur by chance according to equation 4 in a genome of 3 billion bases using a query of 500 bases.

# BLAT

**Table 5.** Sensitivity and Specificity of Single Near-Perfect (One Mismatch Allowed) Nucleotide K-mer Matches as a Search Criterion

	12	13	14	15	16	17	18	19	20	21	22
<b>A.</b> 81%	0.945	0.880	0.831	0.721	0.657	0.526	0.465	0.408	0.356	0.255	0.218
83%	0.975	0.936	0.904	0.820	0.770	0.649	0.591	0.535	0.480	0.361	0.318
85%	0.991	0.971	0.954	0.900	0.865	0.767	0.719	0.669	0.619	0.490	0.445
87%	0.997	0.990	0.983	0.954	0.935	0.867	0.833	0.796	0.757	0.634	0.591
89%	1.000	0.997	0.995	0.984	0.976	0.939	0.920	0.897	0.872	0.775	0.741
91%	1.000	1.000	0.999	0.996	0.994	0.979	0.971	0.962	0.950	0.890	0.869
93%	1.000	1.000	1.000	0.999	0.999	0.996	0.994	0.991	0.988	0.963	0.954
95%	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.994	0.992
97%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
<b>B.</b> K	12	13	14	15	16	17	18	19	20	21	22
F	275671	68775	17163	4284	1070	267	67	17	4.2	1.0	0.3

(A) Columns are for K sizes of 12–22. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated by equation 6 assuming a homologous region of 100 bases. (B) K represents the size of the near-perfect match. F shows how many perfect matches of this size expected to occur by chance according to equation 7 in a genome of 3 billion bases using a query of 500 bases.

# PSST

Struktura sekvence znázorněna vektorem přítomnosti jednotlivých “slov”  $v=(1,0,0,1,\dots,1)$ , kde každá pozice odpovídá určitému “slovu”.  $p_i$  je normalizovaná frekvence výskytu daného slova a podobnost

$$S(q,t) = \text{suma } (q_i * t_i * 1/p_i)$$

Vektory proteinů bez jakékoliv společné podsekvence na dané úrovni jsou ortogonální,  $S(q,t)=0$ .

Pro identické proteiny  $S(q,q) = \text{suma}(1/p_i)$ . Pokud by všechny slova měla stejnou pravděpodobnost výskytu, pak by to bylo

$$N * (1/N) = 1$$

kde  $N$  je počet sledovaných nebo existujících slov.

---

---

