

BLAST (basic local alignment search tool)

Vyhledává podobné sekvence v databázích.

Stal se nástrojem pro všechno.

Určitou dobu kolektiv autorů držel krok s potřebami uživatelů úpravami kódu (BLAST1.0 – BLAST2.0 – PSI-BLAST – MEGABLAST – BLASTZ).

V současnosti se zdá, že lepších výsledků můžeme mnohdy dosáhnout použitím nových postupů šitých na míru daného problému.



Suboptimální nasazení programu BLAST

Použití:

Běžné vyhledávání.

Problém:

Blast nevyužívá plně možnosti zpracovat si předem databázy novými způsoby indexování

Řešení:

Použití programů, které používají asociativní pole nebo stromové struktury pro indexaci databázových sekvencí (SSAHA, MUMER, QUASAR)

Indexy a stromy

INDEX TYPU POLE

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

| | | | | |
|------|-----|------|-------|--------|
| AAA | 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 |

Suboptimální nasazení programu BLAST

Použití:

Běžné vyhledávání.

Problém:

Blast nepracuje s optimální strukturou slova,
čím trpí poměr citlivost/rychlost.

Řešení:

Použití programu PATTERN HUNTER

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

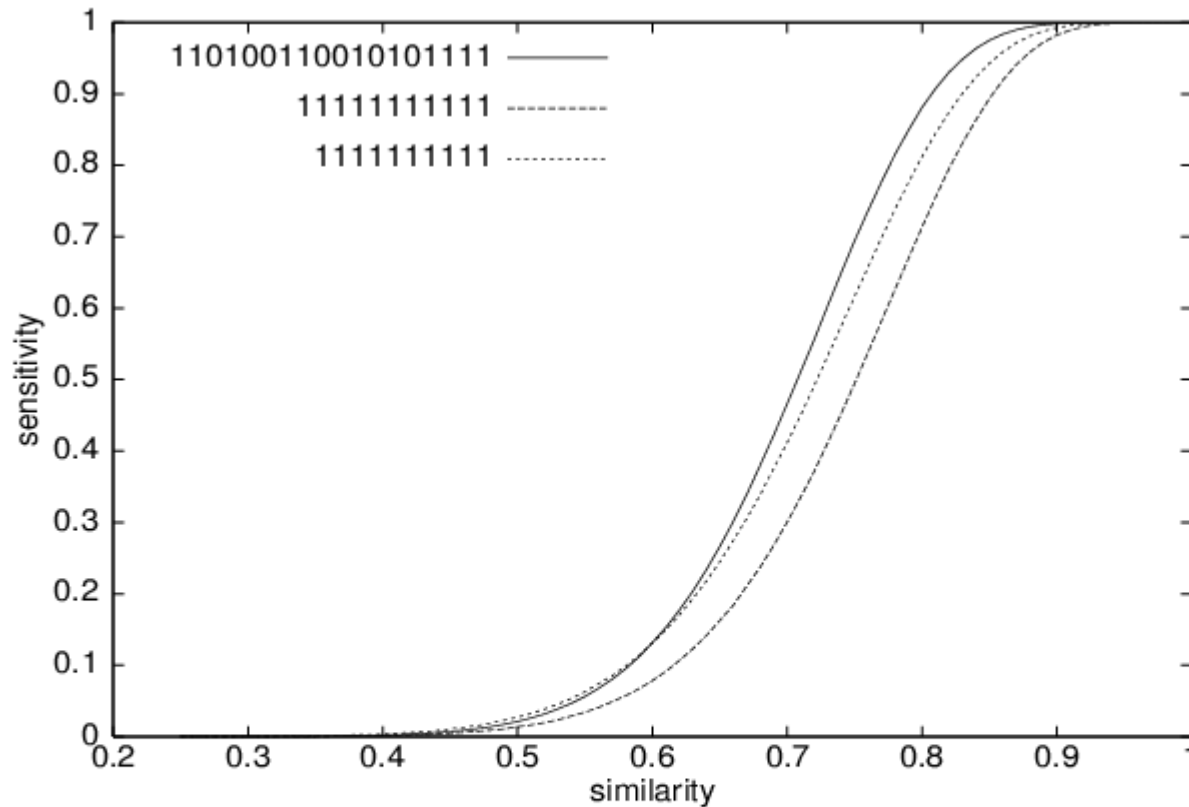


Figure 1: 1-hit performance of weight 11 spaced model versus weight 11 and 10 consecutive models, coordinates in logarithmic scale.

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

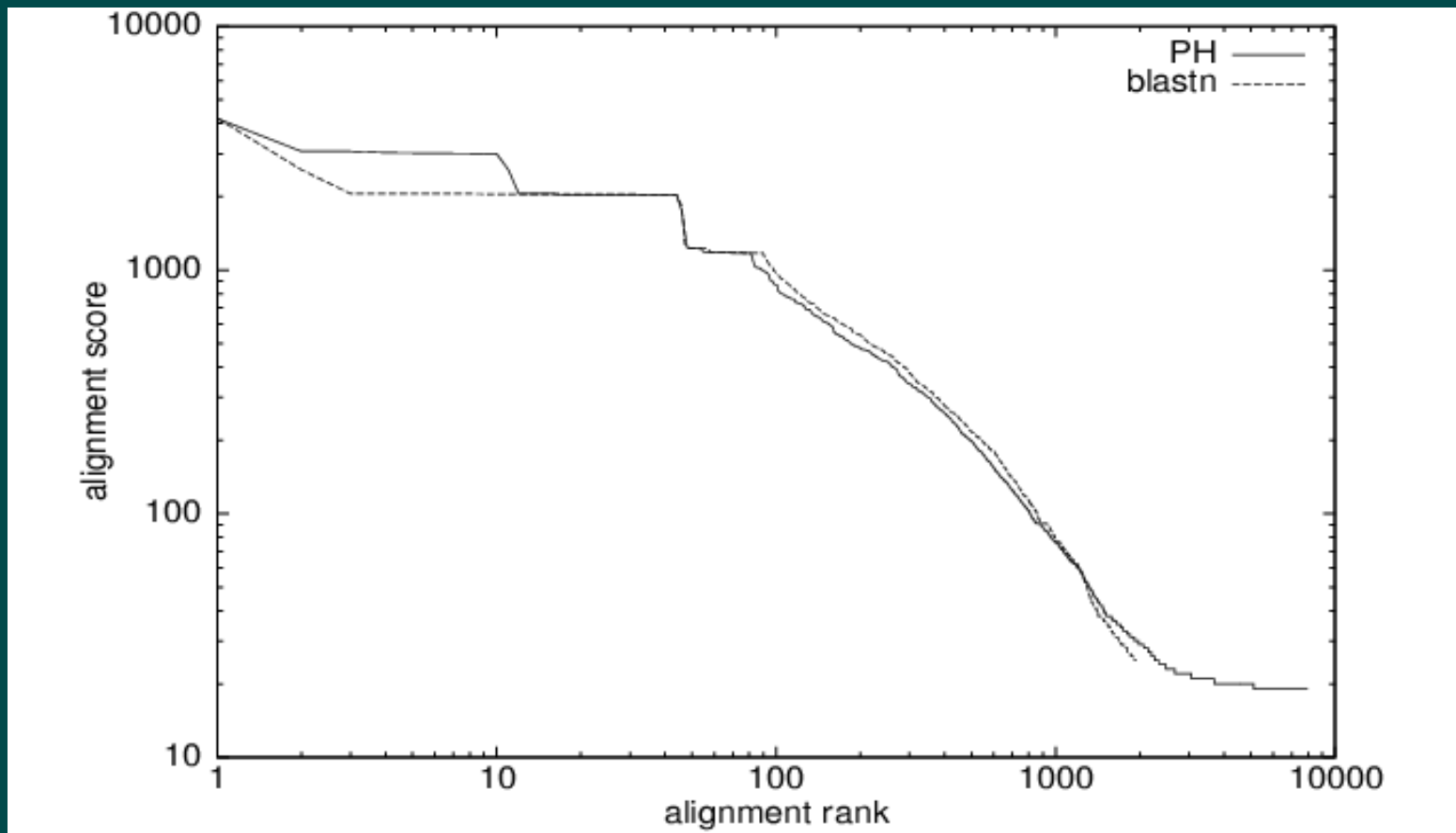
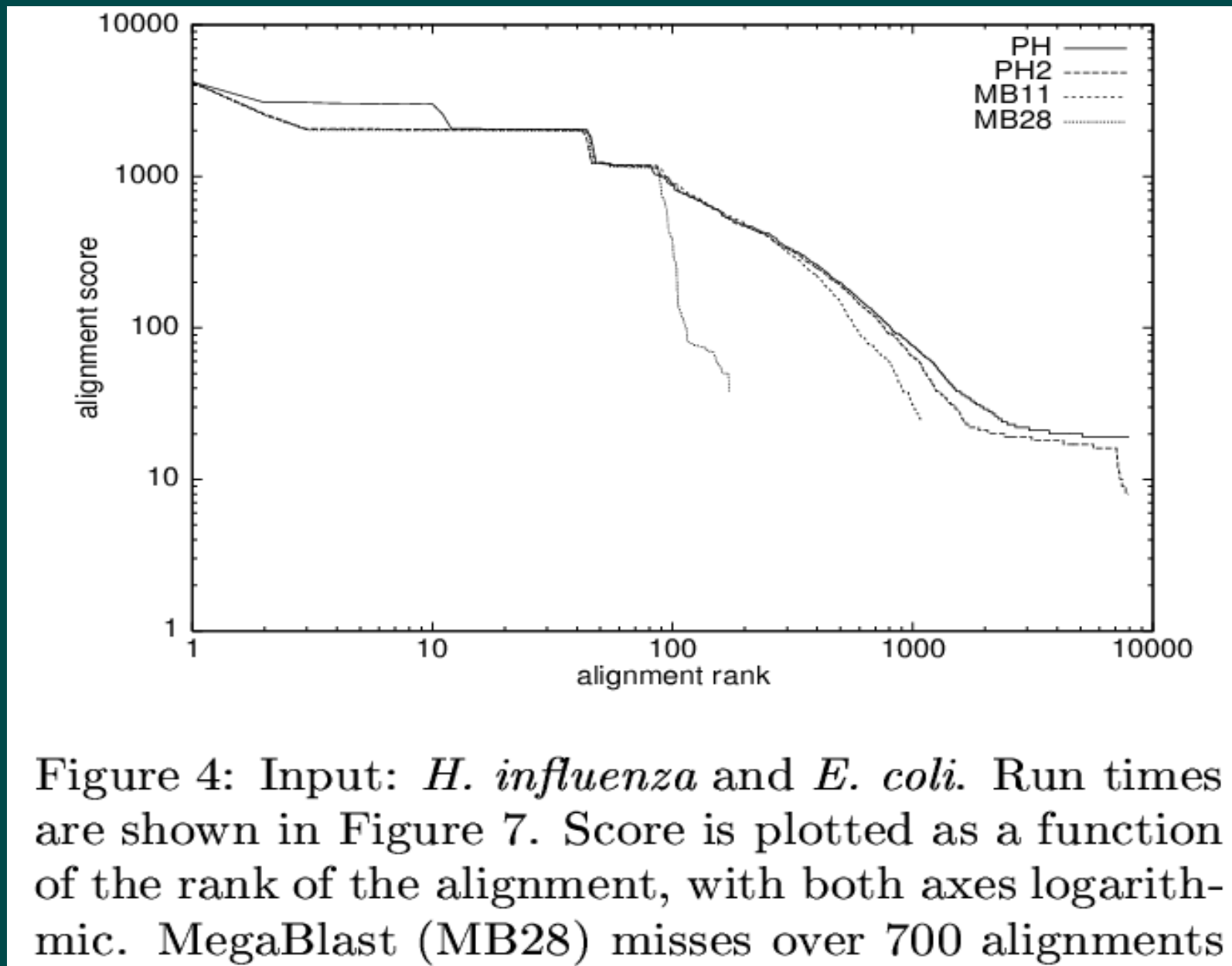


Figure 5: Input: *H. influenza* and *E. coli*. Run times are shown in Figure 7. PatternHunter produces better quality output than Blastn while running 20 times faster.

Pattern Hunter



Suboptimální nasazení programu BLAST

Použití:

Hledání primerů pro sekvenování a PCR (10–40bp)
nebo oligonukleotidů pro microarray (25–100bp)

Problém:

BLAST prohledává DNA se slovy o velikosti 7 a víc. Při použití extrémně krátkých slov nalézá příliš velký počet irelevantních podobností a jejich spracování trvá neúnosně dlouho. Bez předpočítaného indexu je hledání velkého počtu podobností pomalé.

Řešení:

Použití programu PRIMEX

Primex

INDEX

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

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 spolehliv v nejhorším

GTATTAGGTACGTTGACA

i et zec 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-SO (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



Suboptimální nasazení programu BLAST

Použití:

Hledání pozice intronů a exonů (srovnávání mRNA s genomovou DNA) nebo vyhledávání domén proteinů

Problém:

BLAST vyhodnocuje příliš mnoho podobností, i když nás zajímají jenom ty, které jsou typické pro RNA/DNA. Pokud to ošetříme zvýšením parametrů X a S, stratíme krátké podobnosti. Nesnaží se určit přesnou hranici intronů a exonů nebo domén.

Řešení:

Použití programu BLAT, SIM4

BLAT

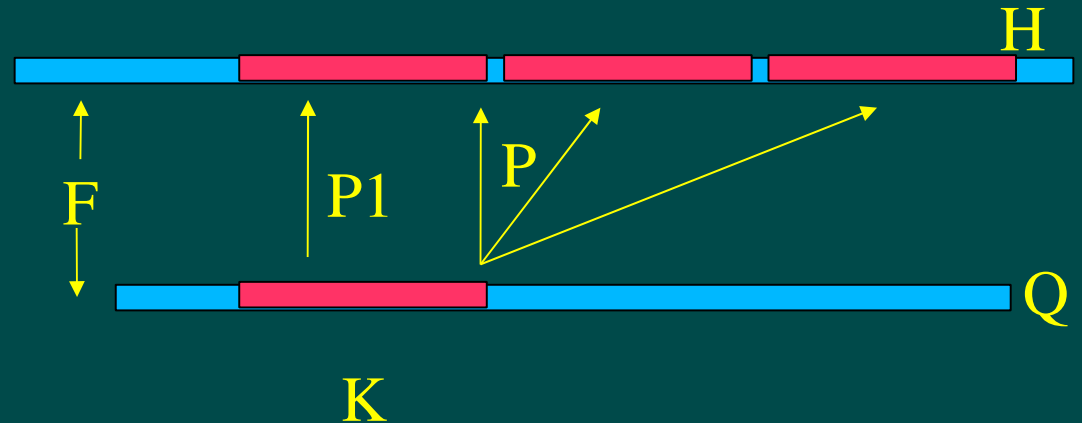
$$P1 = M^K$$

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

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

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

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



P1 – pravdepodobnosť shody s k-merem

P – pravdepodobnosť existencie 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 – veľkosť databázy

Q – veľkosť vyhľadávacej sekvencie

A – veľkosť 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 |
|---------------|----------|----------|----------|-----------|-----------|-----------|-----------|-----------|
| A. 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. K | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 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 |
|---------------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A. 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. 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.

BLAT

Table 7. Sensitivity and Specificity of Multiple (2 and 3) Perfect Nucleotide K-mer Matches as a Search Criterion

| | 2,8 | 2,9 | 2,10 | 2,11 | 2,12 | 3,8 | 3,9 | 3,10 | 3,11 | 3,12 |
|---------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A. 81% | 0.681 | 0.508 | 0.348 | 0.220 | 0.129 | 0.389 | 0.221 | 0.112 | 0.051 | 0.021 |
| 83% | 0.790 | 0.638 | 0.475 | 0.326 | 0.208 | 0.529 | 0.339 | 0.193 | 0.099 | 0.045 |
| 85% | 0.879 | 0.762 | 0.615 | 0.460 | 0.318 | 0.676 | 0.487 | 0.313 | 0.180 | 0.093 |
| 87% | 0.942 | 0.866 | 0.752 | 0.611 | 0.461 | 0.809 | 0.649 | 0.470 | 0.305 | 0.177 |
| 89% | 0.978 | 0.940 | 0.868 | 0.761 | 0.625 | 0.910 | 0.801 | 0.648 | 0.476 | 0.314 |
| 91% | 0.994 | 0.980 | 0.947 | 0.884 | 0.787 | 0.969 | 0.914 | 0.815 | 0.673 | 0.505 |
| 93% | 0.999 | 0.996 | 0.986 | 0.962 | 0.912 | 0.993 | 0.976 | 0.933 | 0.851 | 0.722 |
| 95% | 1.000 | 1.000 | 0.998 | 0.993 | 0.979 | 0.999 | 0.997 | 0.987 | 0.961 | 0.902 |
| 97% | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | 0.999 | 0.997 | 0.987 |
| B. N,K | 2,8 | 2,9 | 2,10 | 2,11 | 2,12 | 3,8 | 3,9 | 3,10 | 3,11 | 3,12 |
| F | 524 | 27 | 1.4 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |

(A) Columns are for N sizes of 2 and 3 and K sizes of 8–12. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated by equation 10. (B) N and K represent the number and size of the near-perfect matches, respectively. F shows how many perfect clustered matches expected to occur by chance according to equation 14 in a translated genome of 3 billion bases using a query of 167 amino acids.

Suboptimální nasazení programu BLAST

Použití:

Hledání podobnosti mezi extrémně dlouhými sekvencemi

Problém:

BLAST jde rovnou na věc a hledá všechny HSP

Řešení:

Použití programů s indexy, které přistupují k analýze hierarchicky, nejprve rychle vyhledávají s nízkou citlivostí, později dpočítávají detaily (SSAHA, BLAT, MUMER)

Suboptimální nasazení programu BLAST

Použití:

Hledání podobnosti mezi vzdálenými proteinovými sekvencemi

Problém:

Proteiny jsou třírozměrné a jejich struktura a funkce často závisí na prostorově se doplňujících motivech sekvence a ne na existenci dlouhého lineárního řetězce aminokyselin

Řešení:

Použití programů, které nepracují s podobností na bázi Smith-Waterman, ale analyzují např. výskyt krátkých slov (PSST)

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.