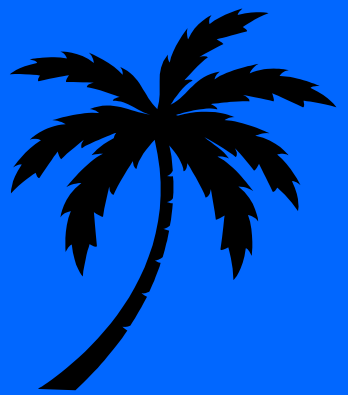
Sequence Database Search Techniques I: Blast and PatternHunter tools

Zhang Louxin

National University of Singapore



Outline

1. Database search
2. BLAST (and filtration technique)
3. PatternHunter (empowered with spaced seeds)
4. Good spaced seeds

1. Sequence database search

- Problem: Find all highly similar segments (called homologies) in the query sequence and sequences in a database, which are listed as local alignments.

>gi|19111785|gb|AC060809.7| _ Homo sapiens chromosome 15, clone RP11-404B13, complete sequence Length = 172123

```
Query: 116      g a a -- t t t t a c a c t t t  c a a a  -- g  136
                 | | |  |||  | | | | |||  | | | |    |
Sbjct: 131078   g a a a t t t g a c a c t t t c a a a g  g  131098
```

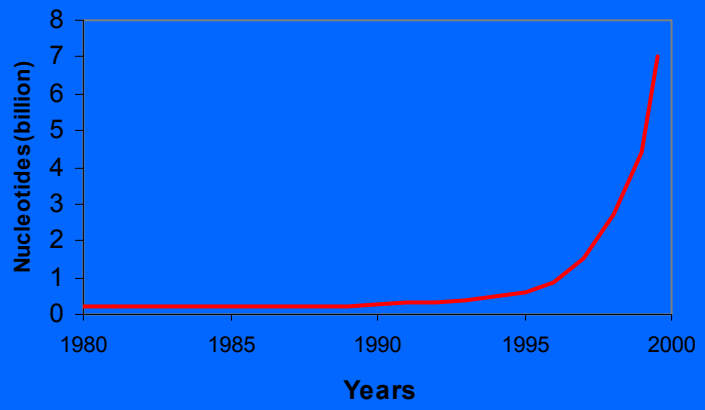
Local alignment

- Mathematically, the local alignment problem is to find a local alignment with maximum score.

- Smith-Waterman Algorithm
  --- dynamic programming algorithm
  --- output optimal local alignments
  --- quadratic time O(mn), and so not scalable.

Scalability is critical

- The genetic data grows exponentially.



30 billions
in 2005.

- Genomes: Human, Mouse, Fly, and etc.
- To meet this demand, many programs were created:

  Blast (MegaBlast, WU- Blast, psi-Blast),

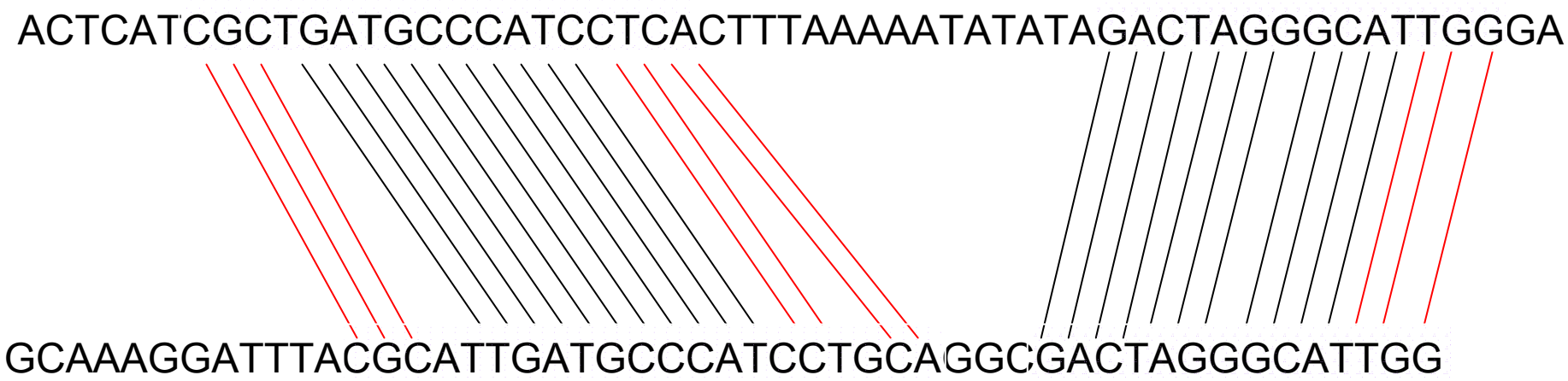
  FSATA,SENSEI, MUMmer, BLAT, etc.

2. Blast Family

- Based on filtration technique.

  1. Filtering stage: identify short matches of length k (=11) in both query and target sequences.
  2. Alignment stage: extend each match found in Stage 1 into a gapped alignment, and report it if significance.

- Its running time is linear time c (m+n), where constant factor c depends on k.


ACTCATCGCTGATGCCCATCCTCACTTTAAAAATATATAGACTAGGGCATTGGGA
GCAAAGGATTTACGCATTGATGCCCATCCTGCAGGCGACTAGGGCATTGG

Dilemma

- Increasing match size k speeds up the program, but loses sensitivity (i.e. missing homology region that are highly similar but do not contain k consecutive base matches).

- Decreasing size k gains sensitivity but loses speed.

Can the di
We need t

>gi|19111785|gb|AC060809.7| _ Homo sapiens chromosome 15, clone RP11-404B13, complete sequence Length = 172123

```
Query: 116      g a a -- t t t t a c a c t t t  c a a a  -- g   136
                  | | |   | | |   | | | | | | | | |    |
Sbjct: 131078   g a a a t t t g a c a c t t t c a a a g  g   131098
```

Dilemma

- Increasing match size k speeds up the program, but loses sensitivity (i.e. missing homology regions that are highly similar but do not contain k consecutive base matches).

- Decreasing match size k gains sensitivity but loses speed.

Can the dilemma be solved?
We need to have both sensitivity and speed.

3. Spacing out matching positions
--- PatternHunter's approach
(Ma, Tromp, and Li, Bioinformatics, 2002)

- Filtering stage: looks for matches in k(=11) noncontinous positions specified by

  an optimal pattern, for example,

  1**11*1

  or several patterns.

  Such a pattern is called a spaced seed.
- Alignment stage: same as Blast.

```
GCAATTGCCGGATCTT      GCAATTGCCGGATCTT
   I I I I I  I          I I  I I  I
GCGATTGCTGGCTCTA      GCGATTGCTGGCTCTA
```

Simple idea makes a big difference

- A good spaced seed not only increases hits in homology regions, but also reduces running time.

  In a region of length 64 with similarity 70%, PH has probability of 0.466 to hit vs Blast 0.3,

  50% increase.

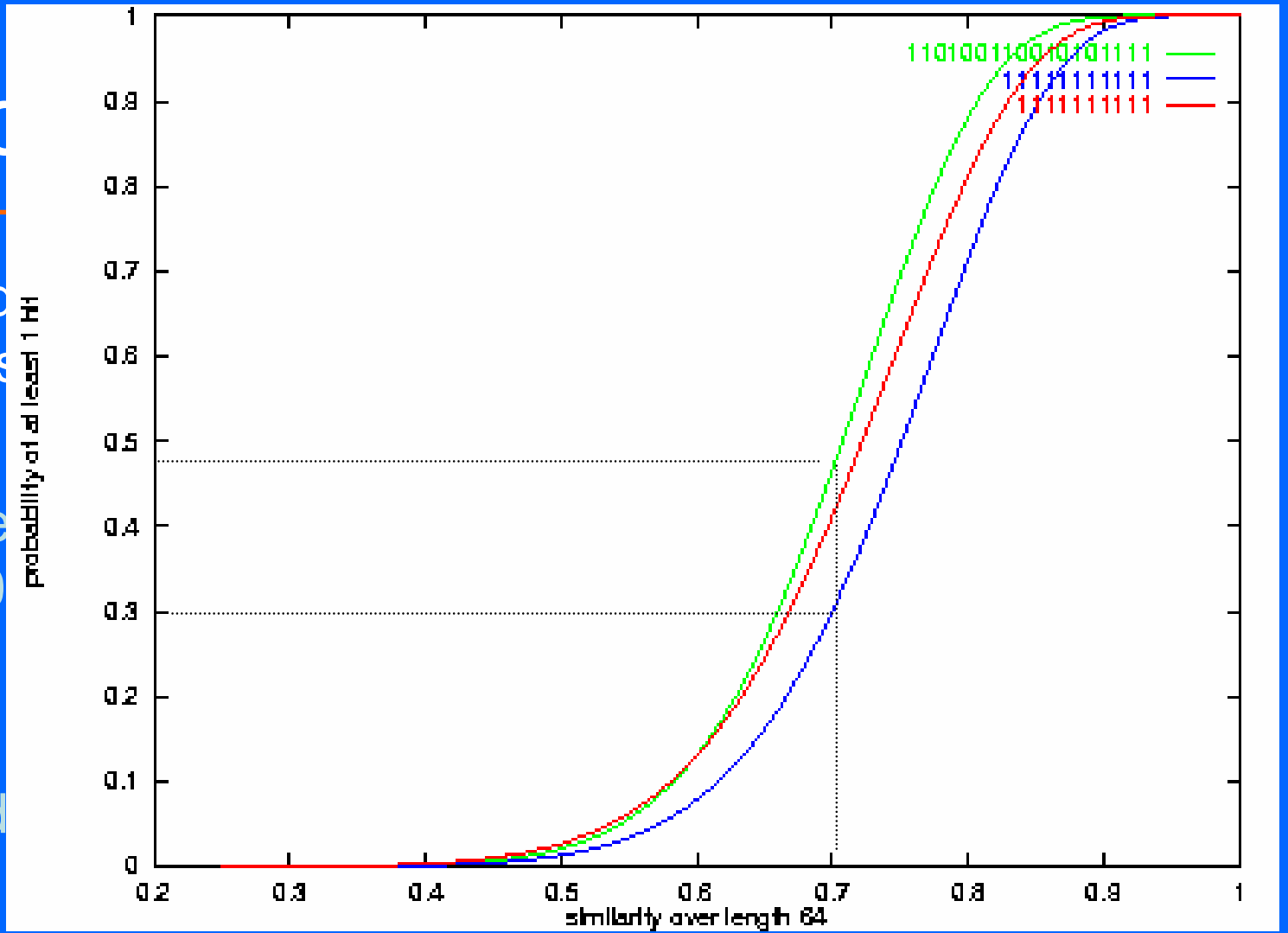
  Time reduction comes from that the average number of matches found in Stage 1 decreases.

- Adopted by BLASTZ, MegaBlast progams. Used by Mouse Genome Consortium.

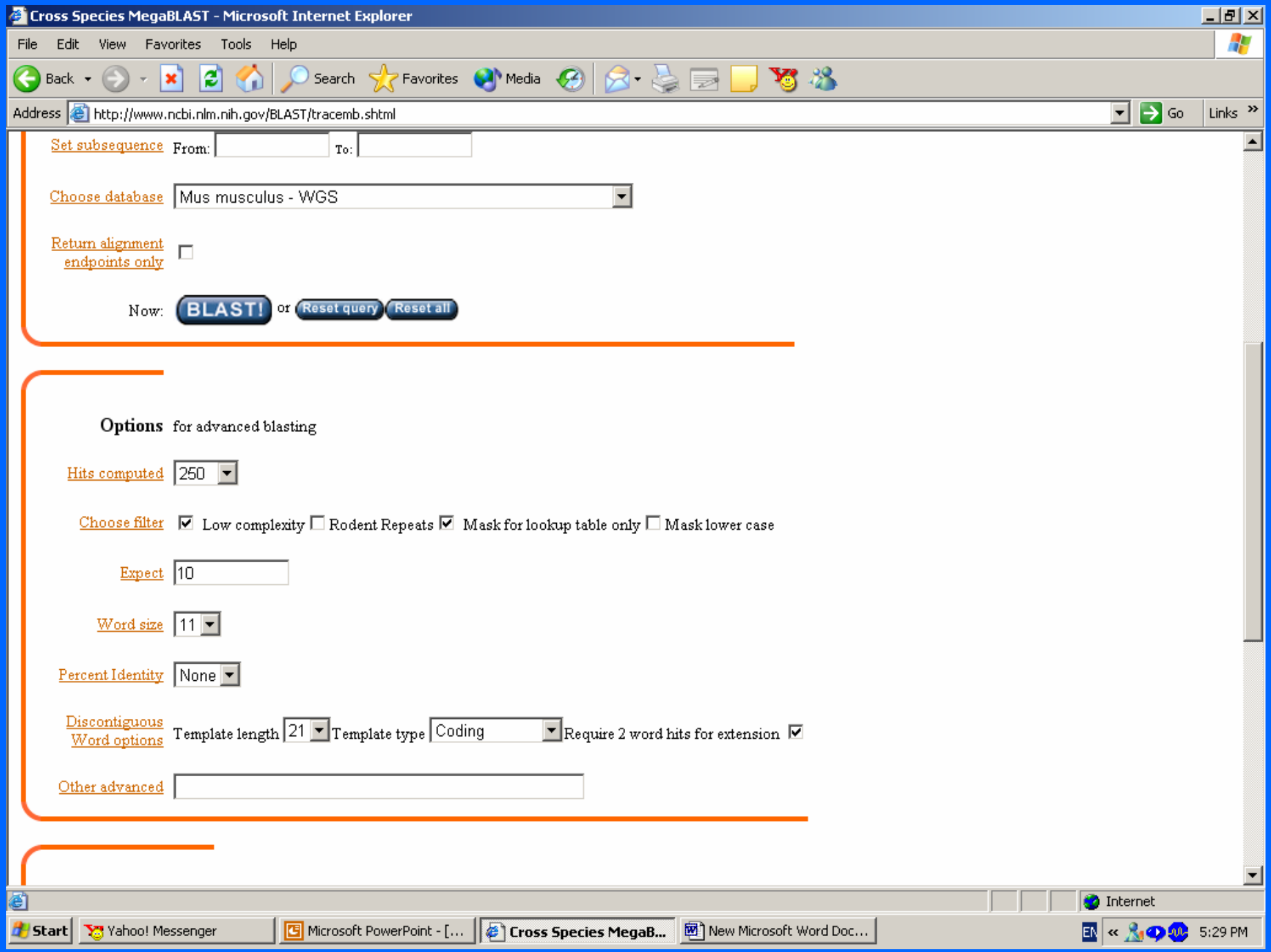
Simple ic

- A good spaced
  regions, but als
  - ■

  In a region of le
  probability of 0
  50% increase.
  Time reduction
  matches found



- Adopted by BLASTZ, MegaBlast progams. Used by Mouse Genome Consortium.


Cross Species MegaBLAST - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/BLAST/tracemb.shtml
Go Links
Set subsequence From: To:
Choose database Mus musculus - WGS
Return alignment endpoints only
Now: BLAST! or Reset query Reset all
Options for advanced blasting
Hits computed 250
Choose filter Low complexity Rodent Repeats Mask for lookup table only Mask lower case
Expect 10
Word size 11
Percent Identity None
Discontiguous Word options Template length 21 Template type Coding Require 2 word hits for extension
Other advanced
Internet
Start Yahoo! Messenger Microsoft PowerPoint - [... Cross Species MegaB... New Microsoft Word Doc... 5:29 PM

Not just spaced seed

- PatternHunter uses a variety of advanced data structures including priority queues, red-black tree , queues, hash tables.
- Several other algorithmic improvements.

Comparison with Blastn, MegaBlast

(A slide from M. Li)

- On Pentium III 700MH, 1GB

	Blastn	MB	PH
E.coli vs H.inf	*716s*	*5s/561M*	*14s/68M*
Arabidopsis 2 vs 4	*--*	*21720s/1087M*	*498s/280M*
Human 21 vs 22	*--*	*--*	*5250s/417M*
61M vs 61M			*3hr37m/700M*
100M vs 35M			*6m*
Human vs Mouse			*20 days*

- All with filter off and identical parameters
- 16M reads of Mouse genome against Human genome for Whiteheads & UCSC. Best Blast program takes 19 years at the same sensitivity (seed length 11).

Questions

- Why is the PH seed better than Blast consecutive seed (11111111111 for weight 11) of the same weight?

  PH spaced seed is `less regular' than Blast seed;
  A random sequence should contain more `less regular' patterns

- Are all spaced seeds better than Blast seed of the same weight ?

  1*1*1*1*1*1*1 is worse than 1111111

- Which spaced seeds are optimal?

  A difficult problem. No polynomial-time algorithm is known for finding them

4. Identifying Good Spaced Seed

--- Ungapped alignment model

- Given two DNA sequences S', S'' with similarity p, we assume the events that they have a base-match at each position are jointly independent, each with probability p.

Under this model, an ungapped alignment between S' and S'' corresponds to a 0-1 random sequence S in which 0 and 1 appear in each position with probability 1-p and p respectively.

Example

```
G C A A T T G C C G G A T C T T
I I   I I I I I I   I I   I I I
G C G A T T G C T G G C T C T A
```

Translate a match to 1 and a mismatch to 0

1 1 0 1 1 1 1 1 0 1 1 0 1 1 1 0

If spaced seed 1**11*1 is used, there are two seed matches in the alignment.

```
G C A A T T G C C G G A T C T T
    I I I I I I   I
G C G A T T G C T G G C T C T A
```

1 1 0 1 1 1 1 1 0 1 1 0 1 1 1 0

```
G C A A T T G C C G G A T C T T
          I I   I I   I
G C G A T T G C T G G C T C T A
```

1 1 0 1 1 1 1 1 0 1 1 0 1 1 1 0

Definitions

- Under the model of similarity p,

(the sensitivity of a spaced seed Q)
||
(the prob. of Q hitting a random 0-1 sequence of a fixed length N=64)

- Sensitivity depends on the similarity p if N is fixed.
- Optimal spaced seed is the one with largest sensitivity, over all the seeds of same weight..

Computing Sensitivity Q_n

1. Consecutive Seed B=1111…1 of weight w.

Let S be a random 0-1sequence of length n

Let A_i be the event of the seed B occurring at position k <= n.

For n>=w+1,

$$
\begin{aligned}
Q_n &= P[A_1 \cup A_2 \cup \cdots \cup A_n] \\
&= P[(A_1 \cup A_2 \cup \cdots \cup A_{n-1}) \cup \overline{A_1}\,\overline{A_2} \cdots \overline{A_{n-1}} A_n] \\
&= Q_{n-1} + P[\overline{A_1}\,\overline{A_2} \cdots \overline{A_{n-1}} A_n] \\
&= Q_{n-1} + p^w (1-p)(1-Q_{n-w-1})
\end{aligned}
$$

Let $\overline{Q_n} = 1 - Q_n$ $\qquad \overline{Q_n} = \overline{Q}_{n-1} - p^w (1-p)\overline{Q}_{n-w-1}$

$$Q_n = P[\cup_{1 \le i \le n} A_i]$$

and the probability, f_n, that the spaced seed first hits S at position n is

$$f_n = P[\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-1} A_n].$$

Obviously,

$$Q_L = p^w = f_L, \quad Q_n = f_n = 0, \quad 1 \le n < L$$

where p is the probability that 1 occurs at a position in S, and

$$Q_n = Q_{n-1} + f_n, \quad n \ge 1. \tag{4}$$

Furthermore, f_n can be computed recursively as follows.

Let

$$W_Q = \{w_1, w_2, \cdots, w_m\}$$

be the set of all $m := 2^{L-w}$ distinct strings w_j obtained from the seed Q by filling 1 in the 'care' positions i_k $(1 \le k \le w)$, i.e. $w_j[i_k] = 1$, and 0 or 1 in the 'don't care' positions. For example, for seed $Q = 1 * 1 * 1$,

$$W_Q = \{10101, 11101, 10111, 11111\}.$$

The seed Q hits at position n if and only if there is an $w_j \in W_Q$ occurs at n. For each j, we use $A_n^{(j)}$ to denote the event that the word w_j occurs at n. Then, $A_n = \cup_{1 \le j \le m} A_n^{(j)}$, and $A_n^{(j)}$'s are disjoint (i.e., $A_n^{(j)} A_n^{(k)} = \emptyset$ for $1 \le j \ne k \le m$). Setting

$$f_n^{(j)} = P[\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-1} A_n^{(j)}], \quad 1 \le j \le m,$$

we have

$$f_n = \sum_{1 \le j \le m} f_n^{(j)}. \tag{5}$$

We use $w_j[a, b]$ to denote the substring of $w_j \in W_Q$ from position a to position b inclusively. For example, $w_j[1, L] = w_j$, where L is the length of the seed Q. Since

$$\begin{aligned}
&\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-1} A_n^{(j)} \\
&= \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-L} A_n^{(j)} \setminus \cup_{i=1}^{L-1} [\bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-i-1} A_{n-i} A_n^{(j)}] \\
&= \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-L} A_n^{(j)} \setminus \cup_{i=1}^{L-1} (\cup_{k=1}^{m} \bar{A}_1 \bar{A}_2 \cdots \bar{A}_{n-i-1} A_{n-i}^{(k)} A_n^{(j)})
\end{aligned}$$

and the joint event $A_{n-i}^{(k)} A_n^{(j)}$ implies that substrings $w_k[i+1, L]$ and $w_j[1, L-i]$ are identical, we obtain that

$$f_n^{(j)} = (1 - Q_{n-L}) P[w_j] - \sum_{i=1}^{L-1} (\sum_{k \in \Gamma_{i,j}} f_{n-i}^{(k)}) P[w_j[L-i+1, L]], \tag{6}$$

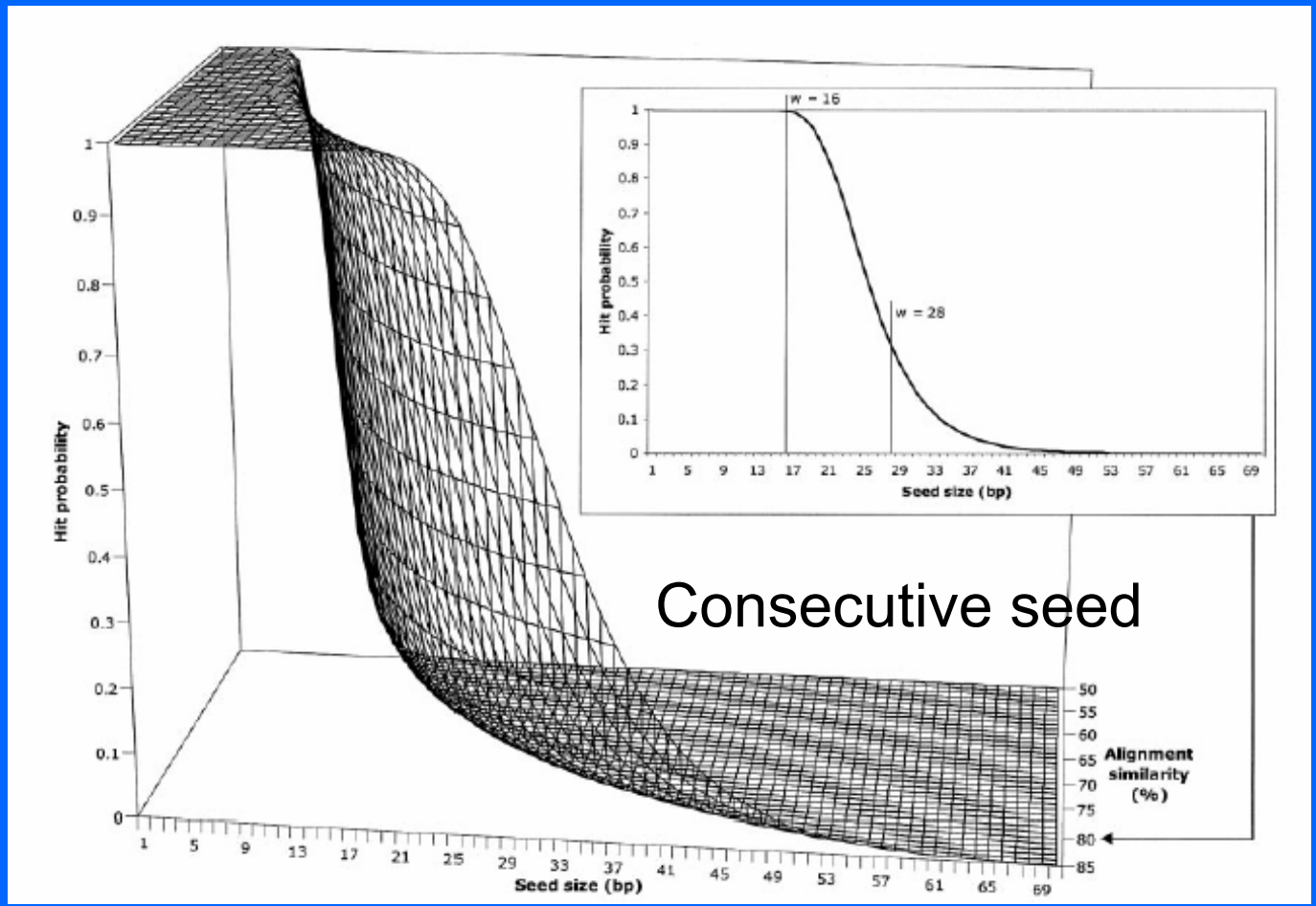
where $P[w_j]$ is the probability that the word w_j occurs at position L and

$$\Gamma_{i,j} = \{k \mid w_k[i+1, L] = w_j[1, L-i]\}.$$

(1-Sensitivity) grows exponentially with length

$$\overline{Q}_n \approx \beta\lambda^n$$

(Buhler et al.)



Expected Number of Exact Matches

Let Q be a spaced seed of weight *w* and length *L*.

Under our model, the expected number E of the exact matches found in an ungapped alignment of length n is equal to the expected value of times T the seed Q occurs in an 0-1 random sequence of length n.

Consider a length-N ungapped alignment with similarity *p* and hence a length-N 0-1 random sequence in which 1 appears with probability *p* in each position. Let A_j denote the event that seed Q occurs at position *j* ($L \le j \le N$) and I_j be the indicator function of A_j:

I_j =1 *if the event* A_i *occurs*

I_j =0 *if the event* A_i *does not occur*

Then $T = \sum_{L \le j \le N} I_j$

$$E(T) = \sum_{L \le j \le N} E(I_j) = \sum_{L \le j \le N} P[A_j] = (N - L + 1) p^w$$

Good Spaced Seeds

- We identified good spaced seeds of weight from 9 to 18 in terms of their optimum span (i.e. the similarity interval in which the seed is optimal).

Weights	Good spaced seeds	Rank under a similarity 65%	70%	75%	80%	85%	90%
9	11∗11∗1∗1∗∗∗111	1	1	1	1	1	1
	11∗1∗11∗∗∗1∗111	2	2	2	2	2	3
	11∗11∗∗1∗1∗∗111	4	4	4	4	4	4
10	11∗11∗∗∗11∗1∗111	1	1	1	1	1	1
	111∗∗1∗1∗∗11∗111	2	2	4	6	8	9
	11∗11∗∗1∗1∗1∗∗111	8	6	2	2	2	5
11	111∗1∗∗1∗1∗∗11∗111	1	1	2	2	2	3
	111∗∗1∗11∗∗1∗1∗111	2	2	1	1	1	1
	11∗1∗1∗11∗∗1∗∗1111	6	3	3	5	5	6
12	111∗1∗11∗1∗∗11∗111	1	1	1	1	1	1
	111∗1∗∗11∗1∗11∗111	2	2	2	5	3	2
	111∗∗1∗1∗1∗∗11∗1111	6	3	3	2	4	4
13	111∗1∗11∗∗11∗∗1∗1111	2	1	1	2	2	2
	111∗1∗∗11∗1∗∗111∗111	7	2	2	1	1	1
	111∗11∗11∗∗1∗1∗1111	6	3	3	3	4	4
14	111∗111∗∗1∗11∗∗1∗1111	2	1	1	1	1	1
	1111∗1∗∗11∗∗11∗1∗1111	5	2	2	3	3	6
	1111∗∗11∗11∗∗1∗1∗1111	6	4	3	4	7	10
15	1111∗∗1∗1∗1∗11∗∗11∗1111	–	5	1	1	1	1
	111∗111∗∗1∗11∗∗1∗11111	14	1	2	5	5	4
	1111∗1∗∗11∗1∗∗111∗1111	17	3	5	8	8	7
16	1111∗11∗∗11∗1∗1∗11∗1111	7	1	2	6	13	20
	1111∗∗11∗1∗1∗11∗∗11∗1111	–	7	1	1	1	3
	1111∗1∗∗11∗1∗1∗∗111∗1111	–	–	5	2	2	1
17	1111∗1∗1∗111∗∗11∗11∗1111	6	1	2	4	4	5
	1111∗1∗11∗∗11∗∗11∗1∗11111	–	–	1	1	1	1
	1111∗111∗∗11∗11∗1∗11111	1	3	–	–	–	–
18	1111∗11∗∗111∗1∗1∗11∗11111	–	1	1	2	3	2
	111∗1111∗1∗∗111∗1∗∗11∗1111	–	–	4	3	1	1
	1111∗111∗111∗∗1∗11∗11111	1	4	–	–	–	–

Experimental Validation

- We conducted genomic sequence comparison with PatternHunter on

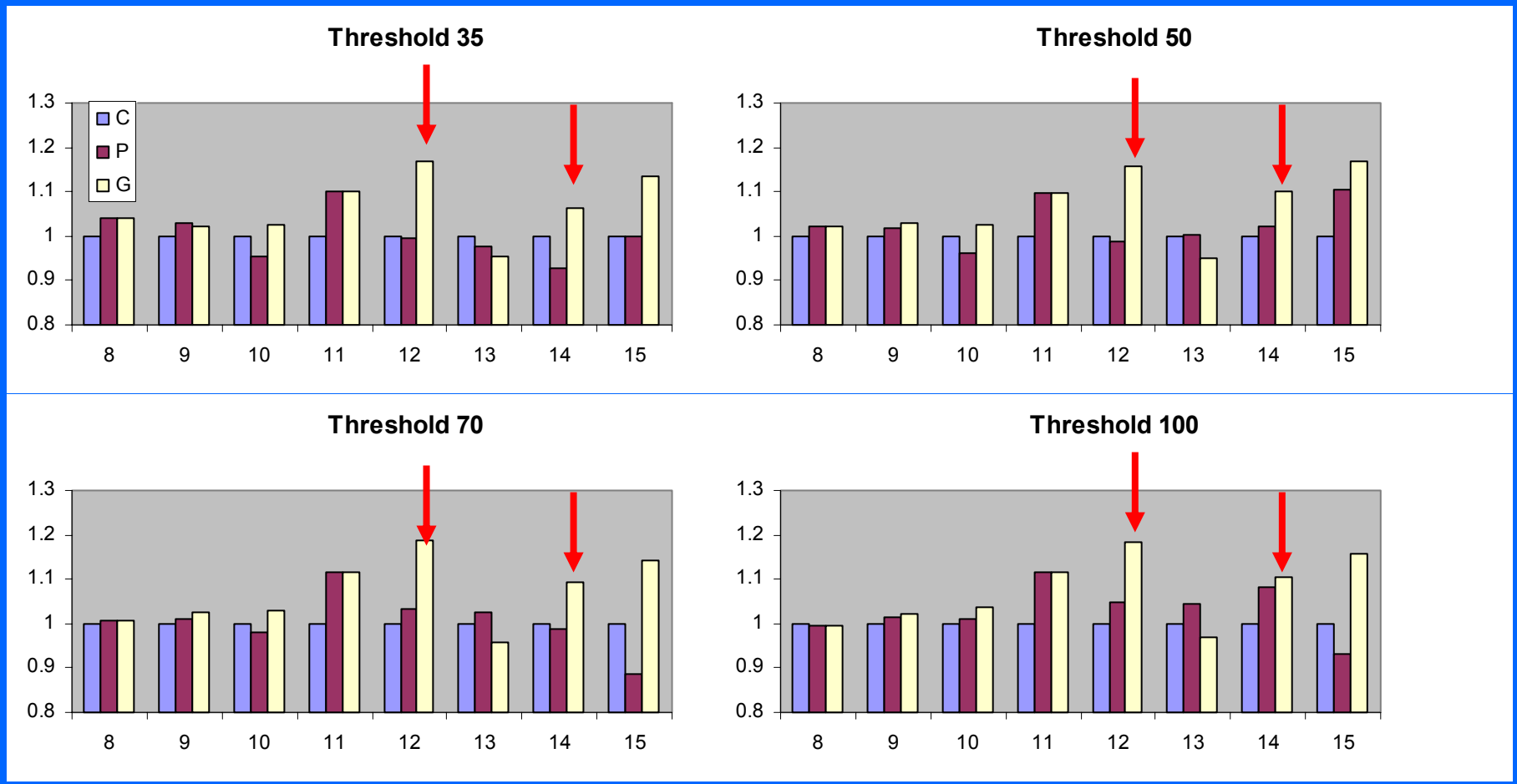
  (i) H. influenza (1.83Mbp) and E. coli (4.63Mbp)

  (ii) A 1.7Mbp segment in mouse ChX
  and a 1Mbp segment in human ChX

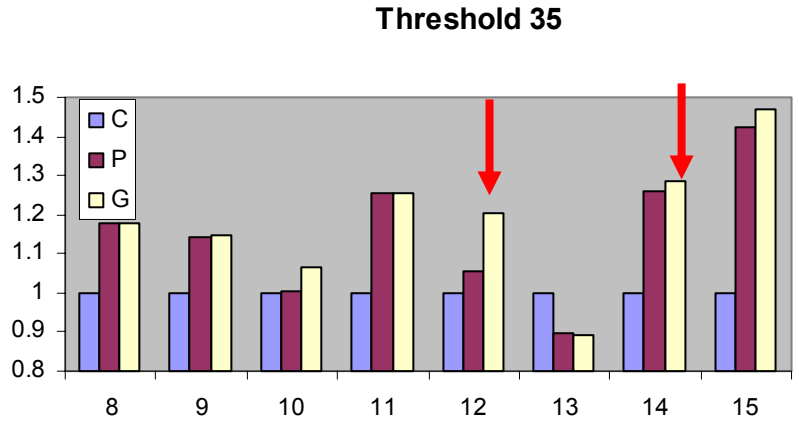
  (iii) A 1.3Mbp segment in mouse Ch10
  and a 2Mbp segment in human Ch19

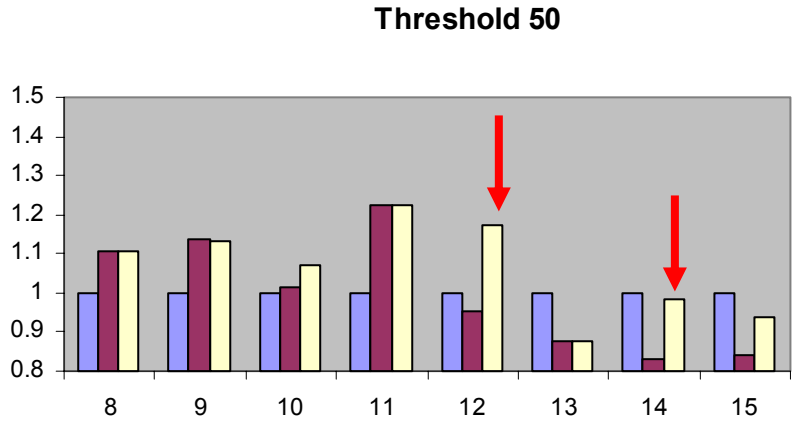
- We evaluate a spaced seed by relative performance. (setting the performance of Blast seed as 1)

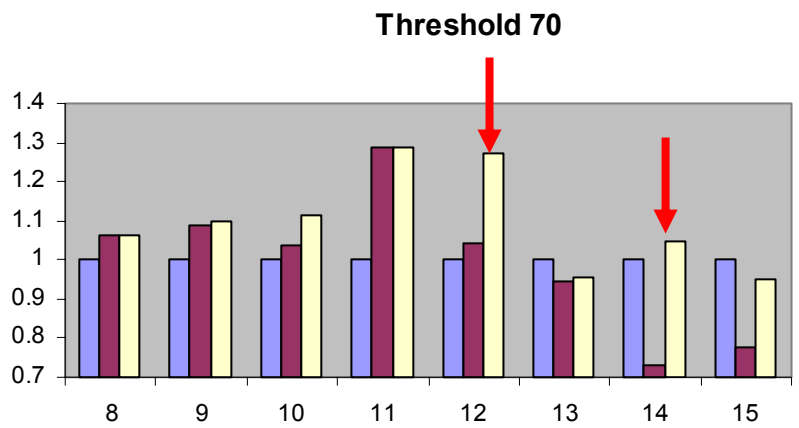
H. Influenza vs E. coli

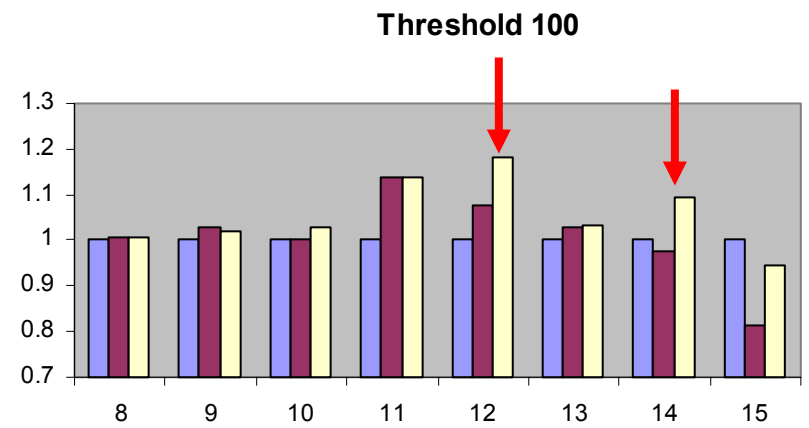


Human X vs Mouse X

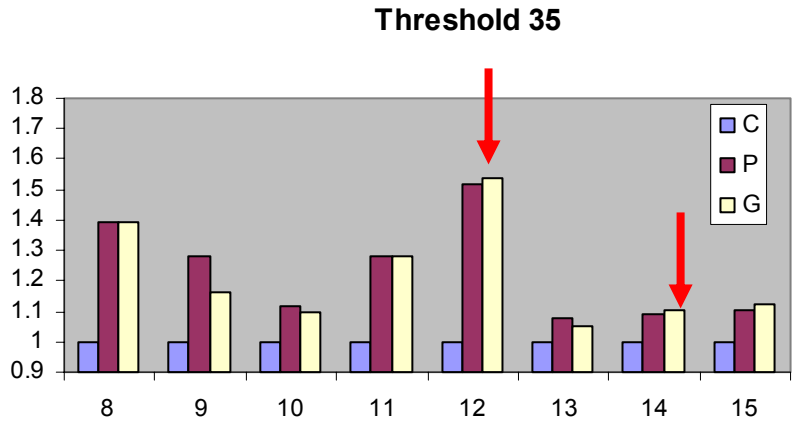


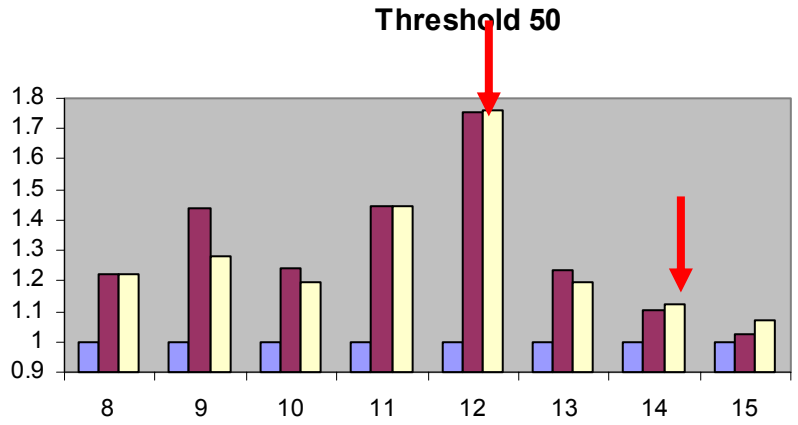


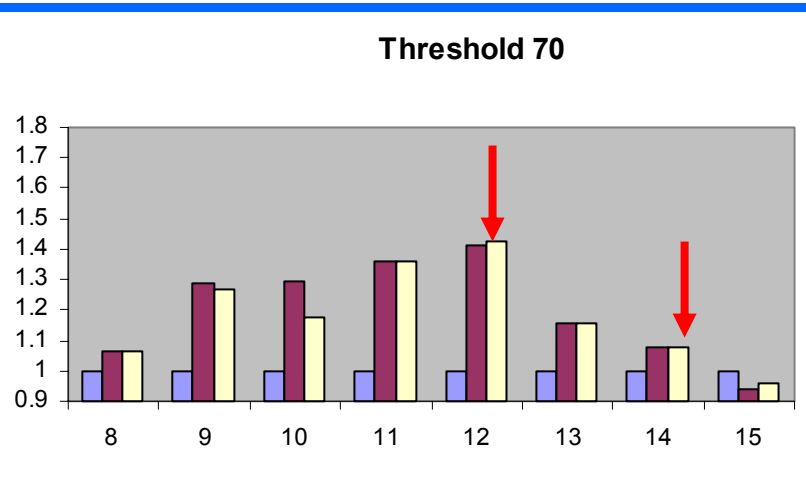


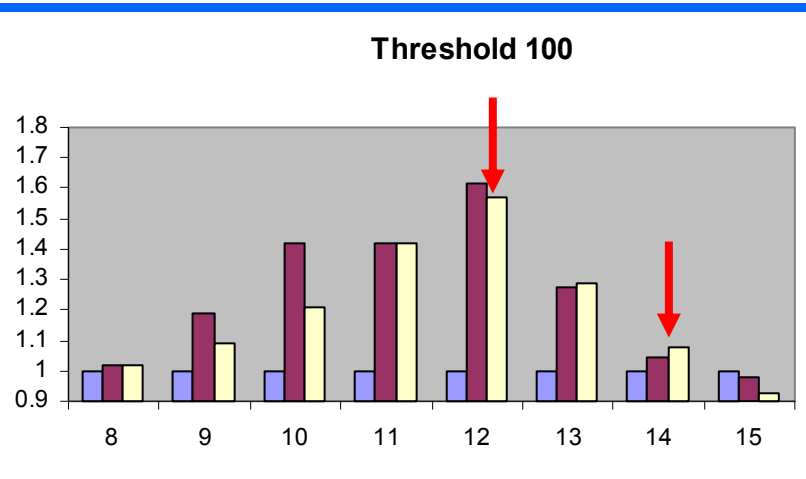


Human 19 vs Mouse 10









Some Recommendations

- There are two competing seeds of weight 11

	Optimum Interval
111*1**1*1**11*111 (PH seed)	[61%, 73%]
111**1*11**1*1*111 (Buhler et al)	[74%, 96%]

PH seed is good for distant homology search, while the latter one for aligning sequences with high similarity.

Recommendations (con't)

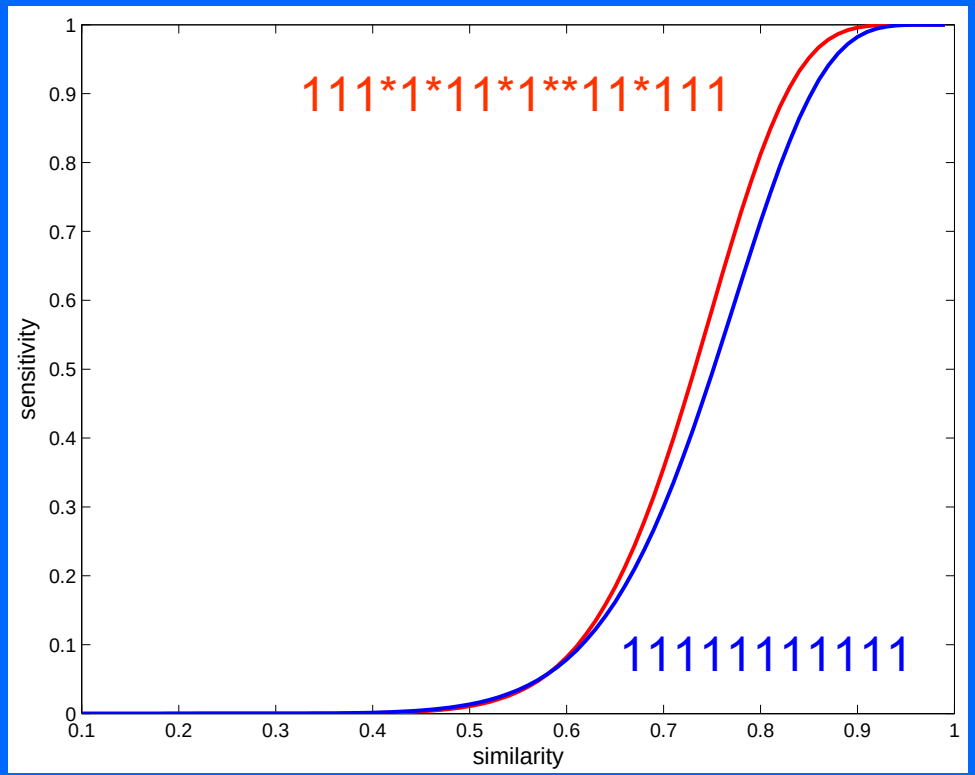
- Spaced seed 111*1*11*1**11*111 of weight 12 is probably the best for fast genomic database search

(1) faster and more sensitive than the Blast default seed weight 11

(2) widest optimum interval [59%, 96%]

(3) good for aligning coding regions since it contains 4 repeats of 11* in its 6-codon span in a reverse direction:

11 1*1 *11 *1* *11 *11 1



Recommendations (con't)

- The larger the weight of a spaced seed, the narrower its optimum interval. So, for database search, a larger weight spaced seed should be carefully selected.

References

Altschul, S.F. et al. "Basic local alignment search tool."
J. Mol. Biol. 1990; 215:403-410.

Altschul, S.F. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 1997; 25:3389-3402.

Ma, B., Tromp, J., Li, M., "PatternHunter: faster and more sensitive homology search", Bioinformatics 2002;18:440-5

Choi, K.P. Zeng F. and Zhang L.X., “Good Spaced Seeds for Homology search”, Bioinformatics 2004 (to appear).
http://www.math.nus.edu.sg/~matzlx/papers/Bio2003_105.pdf

Filtration-based Homology Search Algorithm

- Filtering stage: look for matches in k(=11) noncontinous positions specified by a spaced seed such as 1**11*1 or several seeds.

- Alignment stage: Extend matches found in above stage into an (ungapped or gapped) alignment.

```
G C A A T T G C C G G A T C T T      G C A A T T G C C G G A T C T T
    I I I I I   I                           I I   I I   I
G C G A T T G C T G G C T C T A      G C G A T T G C T G G C T C T A
```

Optimal spaced seeds have larger hitting prob., but smaller expected number of hits found in stage 1 in an alignment.

Protein Sequence Database Search

1. BLASTP

Assume T is a database sequence and S a query sequence.

1. With a fixed length w (from 3 to 5 for protein sequences) and a fixed threshold t, BLASTP finds all length-w subsequences of T that align to a length-w subsequence of S with ungapped-alignment score above t.

2. Each such a subsequence (called a 'hit') is then extended to a maximal alignment, which is reported if the alignment score is above C.

The choices of the score matrix, w and t are critical to the efficiency and effectiveness of BLASTP. For example, lowering t will report more hits but increase the running time.

Amino Acid Substitution Score Matrixes

The theory is fully developed for scores used to find ungapped local alignments.

Let F be a class of protein sequences in which amino acid *i* has background frequency q_i, and each match of residue *i* verse residue *j* has target frequency q_{ij}.

For finding local sequence comparison in F, up to a constant scaling factor, every appropriate substitution score matrix (s_{ij}) is uniquely determined by $\{ q_i , q_{ij} \}$:

$$s_{ij} = \ln\left(\frac{q_{ij}}{q_i q_j}\right)$$

Idea:

In scoring a local alignment, we would like to assess how strong the length-n alignment of x and y can be expected from chance alone

$$P = \frac{\text{Prob[x and y have common ancestor]}}{\text{Prob[x and y are aligned by chance]}}$$

Let $x = x_1 x_2 \cdots x_n, y = y_1 y_2 \cdots y_n$

$$P = \frac{\prod_{1 \le i \le n} q_{x_i y_i}}{\prod_{1 \le i \le n} (q_{x_i} q_{y_i})} = \prod_{1 \le i \le n} \frac{q_{x_i y_i}}{q_{x_i} q_{y_i}} \qquad \ln P = \sum_{1 \le i \le n} \ln \left(\frac{q_{x_i y_i}}{q_{x_i} q_{y_i}} \right)$$

PAM and BLOSUM Substitution Matrices

Hence, all substitution matrices are implicitly of log-odds form.
But how to estimate the target frequencies?
Different methods have been used for the task and result in two series of substitution matrices.

PAM Matrices: the target frequencies were estimated from the observed residue replacements in closely related proteins within a given evolutionary distance (Dayhoff et al. 1978).

BLOSUM Matrices: the target frequencies were estimated from multiple alignments of distantly related protein regions directly (Henikoff & Henikoff, 1992).

DNA vs. Protein Comparison

If the sequences of interest are code for protein,
it is almost always better to compare the protein translations
than to compare the DNA sequences directly.

The reason is (1) many changes in DNA sequences do
not change protein, and (2) substitution matrices for
amino acids represents more biochemical information.

Statistics of Local Ungapped Alignment

It is well understood. The theory is based on the following simple alignment model:

i). All the amino acid appear in each position independently with specific background probabilities.

ii). the expected score for aligning a random pair of amino acid is required to be negative:

$$\sum_{1 \le i, j \le 20} p_i p_j s_{ij} < 0$$

Statistics 2: E-value

The BLAST program was designed to find all the *maximal* local ungapped alignments whose scores cannot be improved by extension or trimming. These are called high-scoring segment pairs (HSPs).

In the limit of sufficiently large sequence lengths *m* and *n*, the expected number (E-value) of HSPs with score at least *S* is given by the formula:

$$E = Kmne^{-\lambda S}$$

Where *K* and *lambda* can be considered as scales for the database size and the scoring system respectively.

Statistics 3: P-Value

E-Value: $E = Kmne^{-\lambda S}$

The number of random HSPs with score >= *S* can be described by a Poisson distribution. This means that the probability of finding exactly *x* HSPs with score >=*S* is given by:

$$e^{-E}\frac{E^x}{x!}$$

By setting x=0, the probability of finding at least one such HSP is

$$1 - e^{-E}$$

This is called the *P-value* associated with score S.

References

1. Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes."
Proc. Natl. Acad. Sci. USA 87:2264-2268.

2. Dembo, A., Karlin, S. & Zeitouni, O. (1994) "Limit distribution of maximal non-aligned two-sequence segmental score." Ann. Prob. 22:2022-2039.

3. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) "A model of evolutionary change in proteins." In "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3 (ed. M.O. Dayhoff), pp. 345-352. Natl. Biomed. Res. Found., Washington, DC.

4. Henikoff, S. & Henikoff, J.G. (1992) "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA 89:10915-10919. (PubMed)