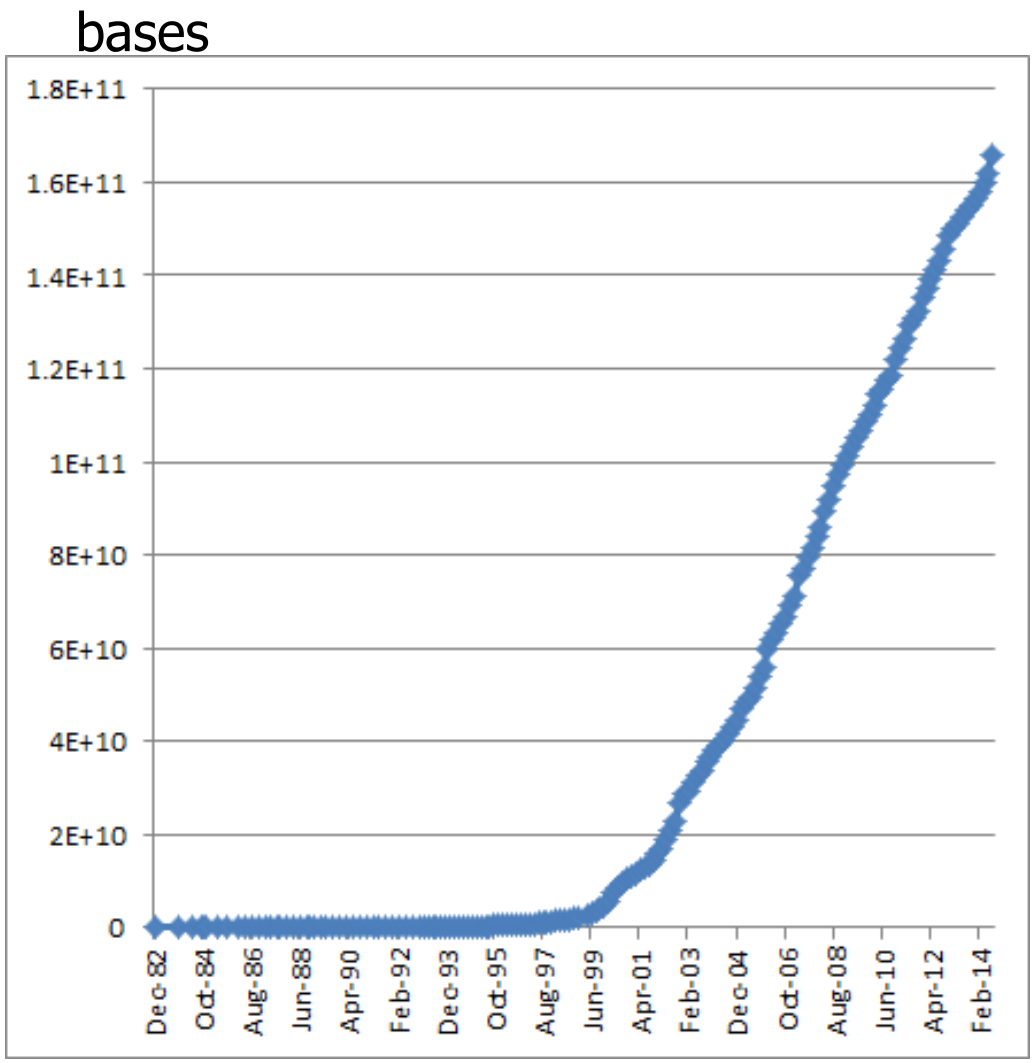
CS3225v: Combinatorial Methods in Computation Biology

Searching biological database

Wing-Kin Sung, Ken 宋永健
ksung@comp.nus.edu.sg

Biological databases

- Biological data increases rapidly.
- Searching methods also need to scale-up to the large datasets



Problem definition

- Consider a database D of genomic sequences (or protein sequences)
- Given a query string Q,
  - we look for a string S in D which is the closest match to the query string Q
  - There are two meanings for closest match:
    - S and Q has a semi-global alignment (forgive the spaces on the two ends of Q)
    - S and Q have a local alignment

Measurement of the goodness of a search algorithm

- Sensitivity
  - Ability to detect “true positive”.
  - Sensitivity can be measured as the probability of finding the match given the query and the database sequence has only x% similarity.
- Specificity
  - Ability to reject “false positive”
  - Specificity is related to the efficiency of the algorithm.
- A good search algorithm should be both sensitive and specific

Different approaches

- Exhaustive approach
  - Smith-Waterman Algorithm
- Heuristic methods
  - BLAST and BLAT
  - PatternHunter
- Filter and refine approaches
  - LSH

- Note: many approaches are local alignment!
- There are other searching algorithms. We don’t have enough time to cover them.

Smith-Waterman Algorithm

- Input:
  - the database D (total length: n) and
  - the query Q (length: m)
- Output: all closest matches (based on local alignment)

Algorithm

- For every sequences S in the database,
  - Use Smith-Waterman algorithm to compute the best local alignment between S and Q
- Return all alignments with the best score

- Time: O(nm)
- This is a brute force algorithm. So, it is the most sensitive algorithm.

What is BLAST?

- BLAST = Basic Local Alignment Search Tool
- Input:
  - A database D of sequences
  - A sequence s
- Aim of BLAST:
  - Compare s against all sequences in D faster based on heuristics.
- Disadvantage of BLAST:
  - To be fast, it scarifies the accuracy. Thus, less sensitive

NCBI Blast - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&ALIGNMENT_VIEW=Pairwise&CLIENT=web&DATAB. Go

NCBI

nucleotide–nucleotide BLAST

Nucleotide Protein Translations Retrieve results for an RID

Search

```
gatggcccaggagaaccccaagatgcacaactcggagatcagcaagcgcctgggcgccg
a
```

Set subsequence From: To:

Choose database nr

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Choose filter ☑ Low complexity ☐ Human repeats ☐ Mask for lookup table only ☐ Mask lower case

Expect 10

Internet

RID=1124972832-31271-29540375534.BLASTQ3, - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi Go

```
Query  241   CTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGCCAGCT  280
             ||||||||||||||||||||||||||||||||||||||||
Sbjct  1136  CTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGCCAGCT  1175


>gi|854181|emb|Z31560.1|HSSOX2G  U E G  H.sapiens sox-2 mRNA (partial)
          Length=1085

 Score =  555 bits (280),  Expect = 2e-155
 Identities = 280/280 (100%), Gaps = 0/280 (0%)
 Strand=Plus/Plus

Query  1     ATGGACAGTTACGCGCACATGAACGGCTGGAGCAACGGCAGCTACAGCATGATGCAGGAC  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   ATGGACAGTTACGCGCACATGAACGGCTGGAGCAACGGCAGCTACAGCATGATGCAGGAC  540

Query  61    CAGCTGGGCTACCCGCAGCACCCGGGCCTCAATGCGCACGGCGCAGCGCAGATGCAGCCC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   CAGCTGGGCTACCCGCAGCACCCGGGCCTCAATGCGCACGGCGCAGCGCAGATGCAGCCC  600

Query  121   ATGCACCGCTACGACGTGAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACCTAC  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   ATGCACCGCTACGACGTGAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACCTAC  660

Query  181   ATGAACGGCTCGCCCACCTACAGCATGTCCTACTCGCAGCAGGGCACCCCTGGCATGGCT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   ATGAACGGCTCGCCCACCTACAGCATGTCCTACTCGCAGCAGGGCACCCCTGGCATGGCT  720

Query  241   CTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGCCAGCT  280
             ||||||||||||||||||||||||||||||||||||||||
Sbjct  721   CTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGCCAGCT  760


>gi|21591813|gb|AC117415.7|  D  Homo sapiens 3 BAC RP11-43F17 (Roswell Park Cancer Institute
             Human BAC Library) complete sequence
          Length=161090
```

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=49457815&dopt=GenBank Internet

RID=1124972832-31271-29540375534.BLASTQ3, - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi Go

```
>gi|30024607|dbj|AB108673.1| U E G Mus musculus Sox2 mRNA for transcriptional factor SOX2, complete
          cds
        Length=1008

 Score =  436 bits (220),  Expect = 1e-119
 Identities = 265/280 (94%), Gaps = 0/280 (0%)
 Strand=Plus/Plus

Query  1    ATGGACAGTTACGCGCACATGAACGGCTGGAGCAACGGCAGCTACAGCATGATGCAGGAC  60
            |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  508  ATGGACAGCTACGCGCACATGAACGGCTGGAGCAACGGCAGCTACAGCATGATGCAGGAG  567

Query  61   CAGCTGGGCTACCCGCAGCACCCGGGCCTCAATGCGCACGGCGCAGCGCAGATGCAGCCC  120
            ||||||||||||||||||||||||||||||| || |||||||| || |||||||| ||
Sbjct  568  CAGCTGGGCTACCCGCAGCACCCGGGCCTCAACGCTCACGGCGCGGCACAGATGCAACCG  627

Query  121  ATGCACCGCTACGACGTGAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACCTAC  180
            |||||||||||||||||| |||||||||||||||||||||||||||||||||||||||||
Sbjct  628  ATGCACCGCTACGACGTCAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACCTAC  687

Query  181  ATGAACGGCTCGCCCACCTACAGCATGTCCTACTCGCAGCAGGGCACCCCTGGCATGGCT  240
            |||||||||||||||||||||||||||||||||||||||||||||||||| || |||||
Sbjct  688  ATGAACGGCTCGCCCACCTACAGCATGTCCTACTCGCAGCAGGGCACCCCCGGTATGGCG  747

Query  241  CTTGGCTCCATGGGTTCGGTGGTCAAGTCCGAGGCCAGCT  280
            || |||||||||||| || |||||||||||||||||||||
Sbjct  748  CTGGGCTCCATGGGCTCTGTGGTCAAGTCCGAGGCCAGCT  787


>gi|20068540|emb|AL606746.17| D Mouse DNA sequence from clone RP23-423J10 on chromosome 3, complete
          sequence
        Length=203345

 Score =  436 bits (220),  Expect = 1e-119
 Identities = 265/280 (94%), Gaps = 0/280 (0%)
```

Internet

NCBI Sequence Viewer v2.0 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=Nucleotide&list_uids=3419872&dopt=GenBank Go

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Nucleotide for Go Clear

Limits Preview/Index History Clipboard Details

Display GenBank Show 20 Send to

Range: from begin to end Reverse complemented strand Features: SNP graph CDD MGC HPRD STS tRNA Refresh

1: U31967. Reports Mus musculus high...[gi:3419872] Links

```
LOCUS       MMU31967                 2283 bp    mRNA    linear   ROD 14-AUG-1998
DEFINITION  Mus musculus high mobility group box protein (sox2) mRNA, complete
            cds.
ACCESSION   U31967
VERSION     U31967.1  GI:3419872
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 2283)
  AUTHORS   Yuan,H., Corbi,N., Basilico,C. and Dailey,L.
  TITLE     Developmental-specific activity of the FGF-4 enhancer requires the
            synergistic action of Sox2 and Oct-3
  JOURNAL   Genes Dev. 9 (21), 2635-2645 (1995)
   PUBMED   7590241
REFERENCE   2  (bases 1 to 2283)
  AUTHORS   Basilico,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-JUL-1995) Claudio Basilico, Microbiology, NYU Medical
            Center, 550 First Avenue, New York, NY 10016, USA
COMMENT     On Aug 14, 1998 this sequence version replaced gi:1101763
```

Internet

History of BLAST

- 1990: Birth of BLAST1
  - It is very fast and dedicate to the search of local similarities **without** gaps
  - Altschul et al, Basic local alignment search tool. J. Mol. Biol., 215(3):403-410, 1990.
  - The most highly cited paper in 1990 and the third most highly cited paper in 1983-2002.
- 1996-1997: Birth of BLAST2
  - BLAST2 allows insertion of gaps
  - BLAST2 have two versions. Developed by two groups of authors independently
    - 1997: NCBI-BLAST2 (National Center for Biotechnology Information)
    - 1996: WU-BLAST2 (Washington University)

BLAST1

- A heuristic method which searches for local similarity without gap
- It can be divided into four steps:
  - Step 1: Query preprocessing
  - Step 2: Scan the database for hits
  - Step 3: Extension of hits

Step 1: Query preprocessing

- For every position p of the query, insert the w-tuple (w=11 default) at position p into the hash table.

Q=TCATCATG

w-tuple	positions
ATCA	3
CATC	2
CATG	5
TCAT	1, 4

Step 2: Generation of hits

- Scan every sequence in the database DB.
  - For each position q in the sequence, if there is an exact match between the w-tuple at position q and a w-tuple in the query, a hit is made.
- A hit is characterized by the positions in both query and DB sequences.

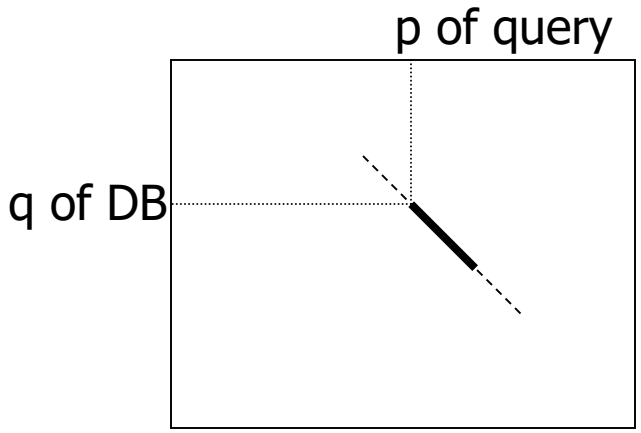
>seq1
CCGCTCATGATGATCA

The list of hits:
- (5 of DB, 1 of query)
- (5 of DB, 4 of query)
- (13 of DB, 3 of query)

W-tuple	positions
ATCA	3
CATC	2
CATG	5
TCAT	1, 4

Step 3: Extension of hits (I)

- For every hit, extend it in both directions, without gap.
- The extension is stopped as soon as the score decreases by more than X(parameter of the program) from the highest value reached so far.



Step 3: Extension of hits (II)

- If the extended segment pair has score better than or equal to S(parameter of the program), it is called an HSP (High scoring segment pair). Then, they will be reported.
- For every sequence in the database, the best scoring HSP is called the MSP (Maximal segment pair).

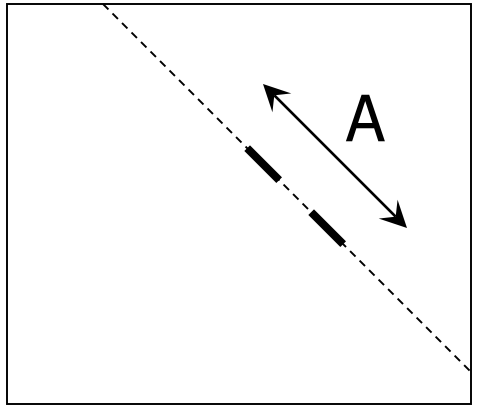
NCBI-Blast2

- Allows local alignment with gaps.

- The first 2 steps are the same as BLAST1.
- Two major differences:
  - Two-hits requirement (implemented for protein)
  - Gapped extension

Step 3: Two-hits requirement

- To extend a hit, we require that there is another hit on the same diagonal within a distance smaller than A
- By default, A=40
- Note: Two-hits requirement is implemented for protein sequences (not DNA).

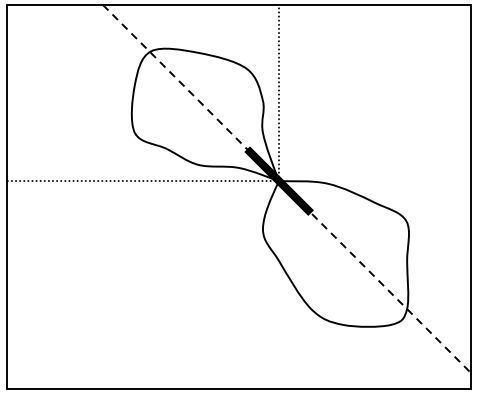


Step 4: Gapped extension (I)

- For hits satisfying the two-hits requirement, extend them similar to Step 3 of BLAST1
- Among the generated HSP, we perform gapped extension for those with score > some threshold

Step 4: Gapped extension (II)

- Gapped extension is a modified Smith-Waterman algorithm
  - Explore the dynamic programming starting from the middle of the hit
  - When the alignment score drops off by more than X_g, stop



BLAST1 vs. NCBI-BLAST2

- BLAST1 spends 90% of its time on extension
- For NCBI-BLAST2, due to the two-hits requirement, the number of extensions is reduced.
  - NCBI-BLAST2 is about 3 times faster than BLAST1.

BLAST program options

Program	Query	Database	Alignment type
blastn	DNA	DNA	Search DNA query sequence in DNA database
blastp	Protein	Protein	Search protein query sequence in protein database
blastx	DNA	Protein	Convert DNA query sequence into protein sequences in all 6 reading frames. Search these translated proteins in protein database
tblastn	Protein	DNA	Search protein query sequence again protein sequences generated from the 6 reading frames of the DNA sequences in the DNA database
tblastx	DNA	DNA	Convert DNA query sequence into protein sequences in all 6 reading frames. Search these translated protein query sequence again protein sequences generated from the 6 reading frames of the DNA sequences in the DNA database

Statistics for local alignment

- A local alignment without gaps consists simply of a pair of equal length segments.
- BLAST finds the local alignments whose score cannot be improved by extension. Such local alignments are called high-scoring segment pairs or HSPs.

- To determine the significant of the local alignments, BLAST gives E-value and bit score. Below, we give a brief discussion on them.

- Assumption: We required the expected score for aligning a random pair of residues/bases to be negative.
  - Otherwise, the longer the alignment, the higher is the score independent of whether the segments aligned are related or not.

Raw Score for BLAST

- Raw score = 8*2 – 3 – (5+2*3) = 2.

```
ACGTGGGACTCC
ACTT---ACTCC
```

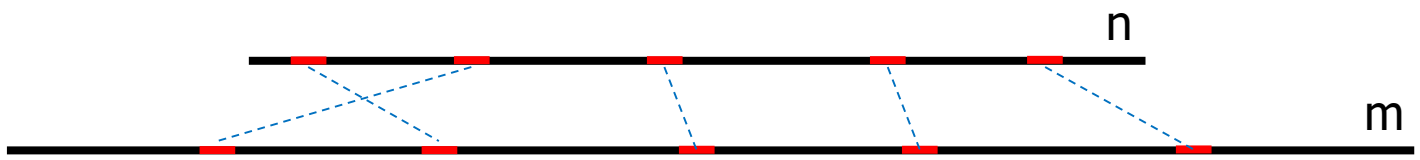
$g(q)=5+2q$

	A	C	G	T
A	2	-3	-3	-3
C	-3	2	-3	-3
G	-3	-3	2	-3
T	-3	-3	-2	2

BLAST Matrix

E-value

- E-value is the expected number of alignments having raw score > S totally at random.
- Let m and n be the lengths of the query sequence and the database sequence.
- Intuition:
  - Double the length of either sequence will double the expected number of HSPs. (i.e. $E \propto nm$)
  - Double the score S will exponentially reduce the expected number of HSPs. (i.e. $E \propto e^{-\lambda S}$)

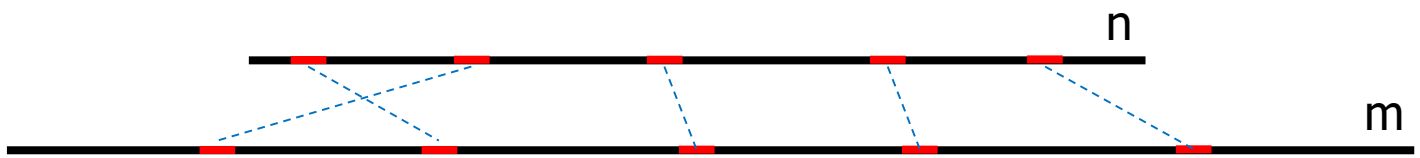


E-value (II)

- Mathematically, when both m and n are sufficiently long,
  - the expected number E of HSPs with score at least S follows the extreme distribution (Gumbel distribution). We have
    - $E = Kmne^{-\lambda S}$

      for some parameters K and λ which depends on the scoring matrix δ and the expected frequencies of the residues/bases.

- Hence, when E-value is small, the HSP is significant.



E-value (III)

- For more information on estimating K and λ, please read
  - http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html
  - http://oreilly.com/catalog/blast/chapter/ch04.pdf

Bit score

- The raw score S of an alignment depends on the scoring system.
- Without knowing the scoring system, the raw score is meaningless.

- The bit score is defined to normalize the raw score, which is defined as follows.

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- Note that $E = Kmne^{-\lambda S}$. By definition of S', $E = mn2^{-S'}$.
- Hence, when S' is big, the HSP is significant.

P-value

- The number of random HSPs with score ≥ S follows a Poisson distribution.
- Pr(exactly x HSPs with score ≥ S) $= \frac{e^{-E}E^{x}}{x!}$
  - where $E = Kmne^{-\lambda S}$ is the E-score
- Hence, p-value = Pr(at least 1 HSPs with score ≥ S) $= 1 - e^{-E}$.
- Note:
  - when E increases, p-value is approaching 1.
  - When E=3, p-value is $1\text{-}e^{-3} = 0.95$.
  - When E=10, p-value is $1\text{-}e^{-10} = 0.99995$
  - when E<0.01, $1\text{-}e^{-E} \approx E$.
- Hence, in BLAST, p-value is not shown since we expect p-value and E-value are approximately the same when E<0.01 while p-value is almost 1 when E>10.

Local alignment with gaps

- There is no solid theoretical foundation for local alignment with gaps.
- Moreover, experimental results suggested that the theory for ungapped local alignment can be applied to the gapped local alignment as well.

Completeness of BLAST (I)

- BLAST is the most popular solution for finding local alignments. It is well-known that BLAST is heuristics and it will miss solution.
- We would like to check how many good alignments are missed by BLAST.

- We extracted 2000 mRNA sequences from each of the 4 different species. We aligned them on human genome. Then, we checked how many significant alignments are missed by BLAST.

Completeness of BLAST (II)

	Chimpanzee	Mouse	Chicken	Zebrafish	All 4 species
E-value ($\leq$)	Missing %	Missing %	Missing %	Missing %	Missing %
1.0×10^{-16}	0.00	0.03	0.05	0.06	0.01
1.0×10^{-15}	0.00	0.03	0.05	0.06	0.02
1.0×10^{-14}	0.00	0.04	0.06	0.06	0.02
1.0×10^{-13}	0.00	0.03	0.07	0.14	0.02
1.0×10^{-12}	0.01	0.04	0.10	0.17	0.03
1.0×10^{-11}	0.02	0.05	0.11	0.28	0.05
1.0×10^{-10}	0.02	0.07	0.13	0.39	0.06
1.0×10^{-9}	0.03	0.09	0.16	0.60	0.08
1.0×10^{-8}	0.05	0.11	0.25	0.77	0.12
1.0×10^{-7}	0.10	0.19	0.31	0.81	0.18
1.0×10^{-6}	0.17	0.31	0.45	1.08	0.28
1.0×10^{-5}	0.32	0.47	0.70	1.45	0.45
1.0×10^{-4}	0.57	0.88	0.99	1.81	0.75
1.0×10^{-3}	0.99	1.36	1.25	2.25	1.17
1.0×10^{-2}	1.69	2.11	1.68	2.61	1.84
1.0×10^{-1}	2.70	2.97	2.33	2.86	2.76

- 2000 queries for each species.
- BLAST only missed 0.06% of those 8000 queries (with E-value smaller than 1.0×10^{-10}).
- In conclusion, BLAST is accurate enough in most cases, yet the few alignments missed could be critical for biological research.

Variation of BLAST

- MegaBLAST
- BLAT
- PatternHunter
- PSI-BLAST

MegaBLAST

- Only for DNA
- For DNA, in BLAST, w = 11 by default.
- To improve efficiency, MegaBLAST uses longer w-tuples (by default, w=28).
- The cost is the reduction in sensitivity.

BLAT

- Only for DNA.
- By default, BLAT uses w=11 and two-hit.

- BLAT is very fast.
  - The main trick is to index the database and put the index in the main memory
  - Note that BLAT is less sensitive than BLAST, but more sensitive than MegaBLAST.

Main trick of BLAT

- BLAST cannot build index of human genome since it it big.
- BLAT's index stores the positions of non-overlapping w-tuples in memory.

Database = ACTTGTACTTGTACTTGTA

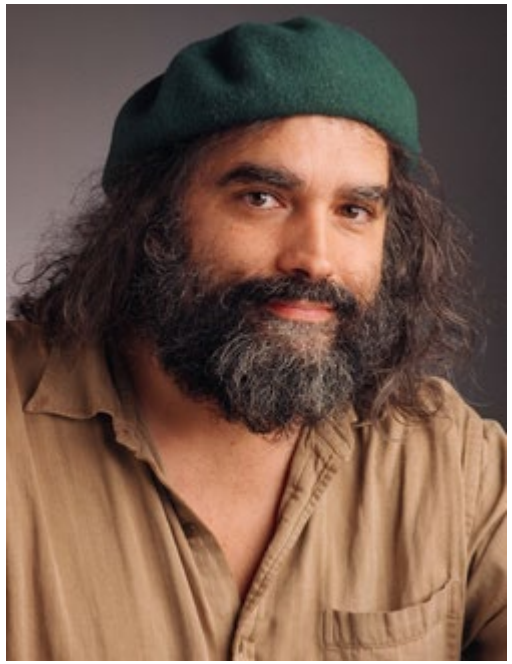
Index of all w-mers

w-mer	positions
ACTT	1, 7, 13
CTTG	2, 8, 14
GTAC	5, 11
TACT	6, 12
TGTA	4, 10
TTGT	3, 9, 15
TGTA	16

Index of w-mers at positions iw+1

w-mer	positions
ACTT	1, 13
GTAC	5
TTGT	9

About the inventor: Jim Kent



- Education: University of California, Santa Cruz
- Awards: Overton Prize, Benjamin Franklin Award

PatternHunter

- PatternHunter can only apply to DNA
- PatternHunter is similar to BLAST. Moreover, it uses gapped w-tuple.
  - For w=11, they use 111010010100110111
  - Example,

```
111010010100110111
ACTCCGATATGCGGTAAC
|||-|--|-|--||-|||
ACTTCACTGTGAGGCAAC
```

- They found that gapped w-tuple can increase the sensitivity while increase the efficiency.

Advantage of gapped w-tuple (I)

- Increase sensitivity
  - Gapped w-tuples are more independent.
  - Examples:
    - Two adjacent ungapped 11-tuples share 10 symbols
      - `11111111111`
        `11111111111`

        1/4 chances to have 2nd hit next to the 1st hit
    - Two adjacent gapped 11-tuples share 5 symbols
      - `11101001010011011 1`
        `11101001010011011 1`

        $1/4^6$ chances to have 2nd hit next to the 1st hit
  - If the w-tuples are more independent, the probability of having at least one hit in a homologous region is higher.

Advantage of gapped w-tuple (II)

- Reduce the number of hits.
  - For the same query length (says, 64),
    - It covers by 54 ungapped 11-tuples
    - It covers by 47 gapped 11-tuples
  - So, the number of hits is smaller.
- Thus, the efficiency is increased!

PatternHunter I

Ma et al., *Bioinformatics* 18:440-445, 2002

Proposition. The expected number of hits of a weight-W length-M model within a length-L region of similarity p is $(L - M + 1) * p^W$

Proof.

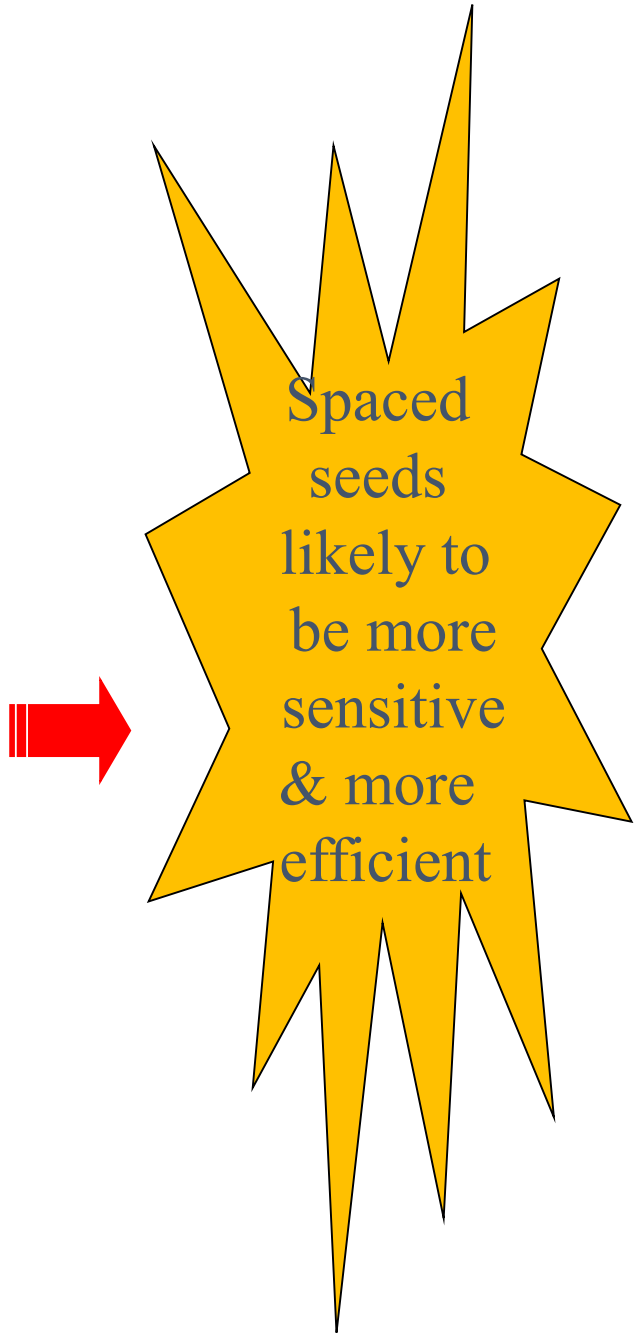
For any fixed position, the prob of a hit is p^W.

There are $L - M + 1$ candidate positions.

The proposition follows.

Implication

- For $L = 1017$
  - BLAST seed expects $(1017 - 11 + 1) * p^{11} = 1007 * p^{11}$ hits
  - But ~$1/4$ of these overlap each other. So likely to have only ~$750 * p^{11}$ distinct hits
  - Our example spaced seed expects $(1017 - 18 + 1) * p^{11} = 1000 * p^{11}$ hits
  - But only $1/4^6$ of these overlap each other. So likely to have ~$1000 * p^{11}$ distinct hits



Sensitivity of PatternHunter I

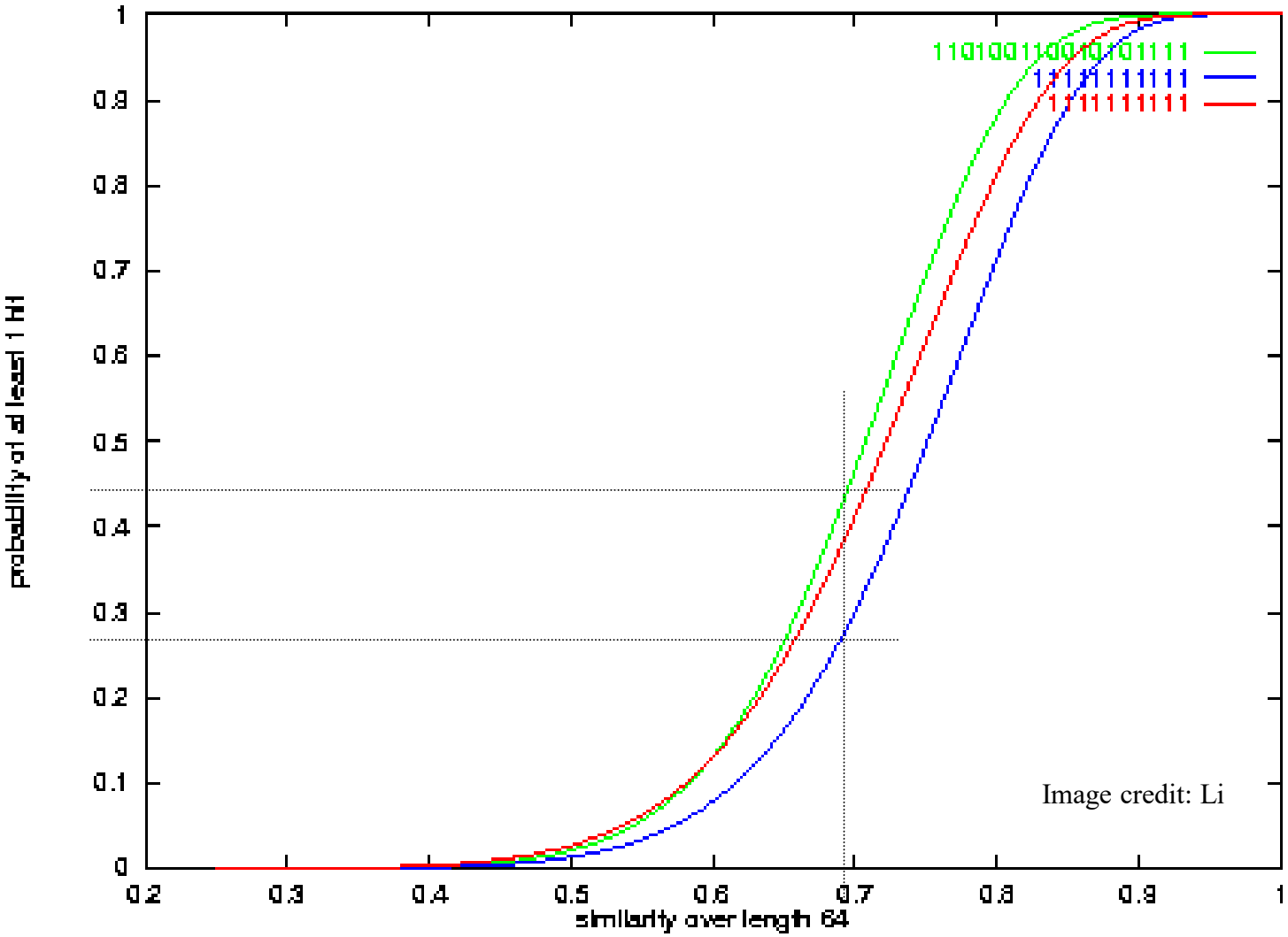


Image credit: Li

More for PatternHunter

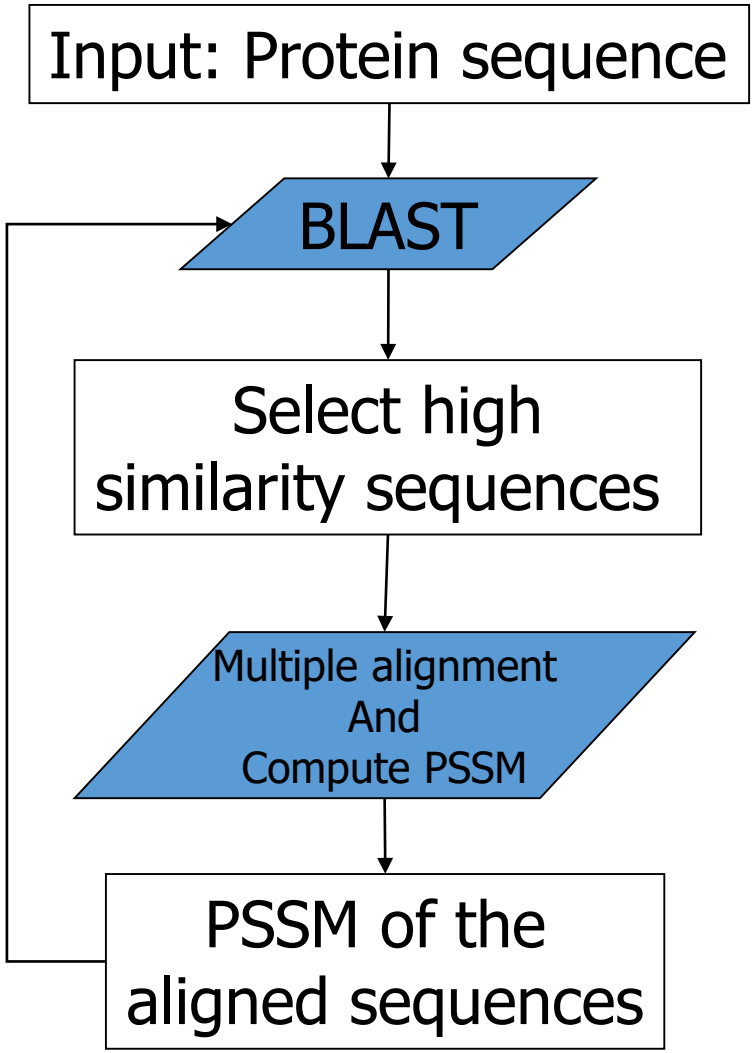
- To further improve the efficiency,
  - PatternHunter uses a variety of advanced data structures including priority queues, a variation of red-black tree, queues, hash tables.
  - PatternHunter also uses a new method of sequence alignment.
- To further improve the accuracy,
  - PatternHunter II suggested to use multiple gapped seeds.
  - They show that the accuracy can approach smith-waterman algorithm while the speed 3000 times faster than smith-waterman.
- PatternHunter II is both faster and sensitive than BLAST, MegaBLAST.

About the Inventor: Ming Li

- Ming Li
  - Canada Research Chair Professor of Bioinformatics, University Professor, Univ of Waterloo
  - Fellow, Royal Society of Canada. Fellow, ACM. Fellow, IEEE.

PSI-BLAST (Position Specific Iterated BLAST)

- PSI-BLAST is an implementation of BLAST for finding protein families. It allows us to detect distant homology.
- Input: a protein sequence
  - Using BLAST, we get a set of sequences that align with the query protein with E-score below a threshold, 0.01 (by default).
  - Align the selected sequences
  - Generate a PSSM profile from the multiple alignment
  - Iterate until no significant alignment found,
    - Using a modified BLAST, search the database with the PSSM profile.
    - Align the selected sequences
    - Generate a PSSM from the multiple alignment
- This version automatically combines statistically significant alignments produced by BLAST into a position-specific score matrix.
- It is much more sensitive to weak but biologically relevant sequence similarities

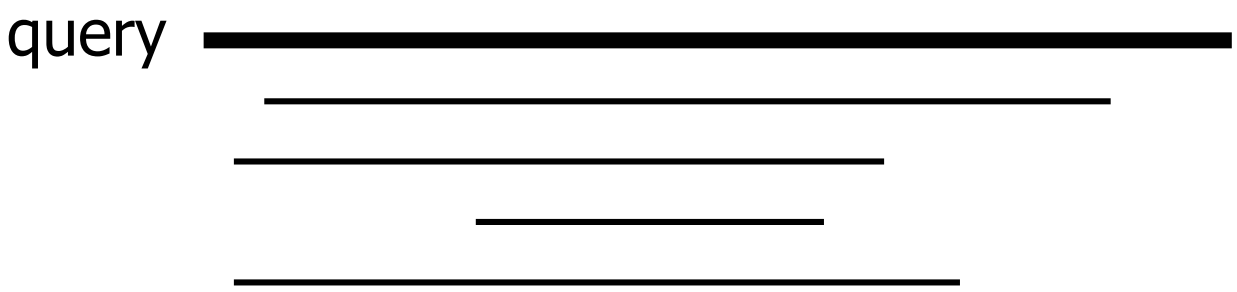


Find a set of sequences similar to the query

- Using BLAST 2.0, we get a set of sequences that align with the query protein with E-score below a threshold, 0.01 (by default).

Multiple sequence alignment of the selected sequences

- Using the query sequence as the template, we aligned the selected sequences.
- All gap characters inserted into the query sequence are ignored.
- Note:
  - the length of the alignment is the same as the query sequence.
  - Some columns of the multiple sequence alignment may include nothing except the query.



Generate a PSSM profile from the alignment

- Given the multiple alignment of length n,
  - We generate the position-specific score matrix (PSSM) profile, which is a 20xn matrix.
  - For each column and each residue a in the profile, we generate a log-odds score $\log(O_{ia}/P_a)$.
    - where O_{ia} is the observed frequency of residue a at position i and P_a is the expected frequency respectively of the residue a.
- Since number of sequences may be small, data-dependent pseudo frequency is added to O_{ia}.

Find a set of sequences similar to the PSSM profile

- We apply a modified BLAST to the PSSM profile.
  - Basically, when we compare a position of the PSSM and a residue in the database, we use the corresponding log-odds score in that position.

- Repeat until we satisfy.

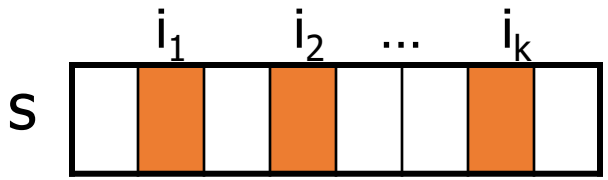
Locality-Sensitive Hashing (LSH)

LSH-ALL-PAIRS

- Input: biosequence database D
- Aim: find pairs of w-mers that differ by at most d substitutions (ungapped local alignment) in a collection of biosequences D.

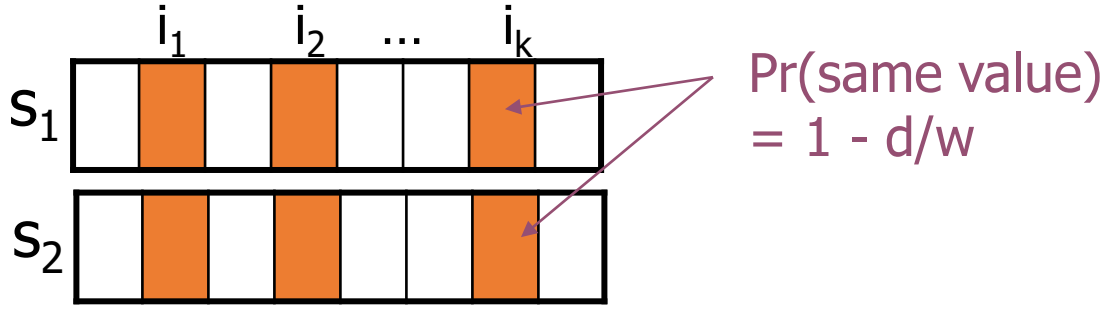
Locality-sensitive hash function

- Consider an w-mers s,
  - choose k indices $i_1, i_2, \ldots, i_k$ uniformly from the set {1, 2, …, w}
  - Define $\pi(s) = (s[i_1], s[i_2], \ldots, s[i_k])$. This function is called the locality-sensitive hash function



Property of locality-sensitive hash function (I)

- Consider two w-mers s_1 and s_2,
  - the more similar are they, the higher probability that $\pi(s_1) = \pi(s_2)$.
- More precisely, if the hamming distance of s_1 and s_2 = d,
  - $\Pr[\pi(s_1) = \pi(s_2)] = \prod_{j=1,\ldots,k} \Pr[s_1[i_j]=s_2[i_j]]$
    $= (1 - d/w)^k$



Property of locality-sensitive hash function (II)

- Hence, s_1 and s_2 are similar if
  - $\pi(s_1) = \pi(s_2)$
- However, we may have false positive and false negative
  - False positive: s_1 and s_2 are dissimilar but $\pi(s_1) = \pi(s_2)$.
    - False positive can be distinguished from true positive by computing hamming distance between s_1 and s_2
  - False negative: s_1 and s_2 are similar but $\pi(s_1) \neq \pi(s_2)$.
    - We cannot detect false negative.
    - We can only reduce the number of false negative by repeating the test using different $\pi()$ functions

LSH-ALL-PAIRS

Algorithm:

1. Generate m random locality-sensitive hash functions $\pi_1(\ )$, $\pi_2(\ )$, …, $\pi_m(\ )$.
2. For every w-mer s in the database, compute $\pi_j(s)$ for $1 \leq j \leq m$.
3. For every pair of w-mers s and t such that $\pi_j(s) = \pi_j(t)$ for some j,
   - If hamming distance(s, t) < d, report (s, t)-pair.

Conclusion

- This lecture presents some database searching methods.
- In fact, there are many other methods. For examples:
  - CAFÉ, FLASH, RAMdb, FD, suffix tree, suffix array, compressed suffix array

More information

- The list of database used by blast
  - ftp://ftp.ncbi.nlm.nih.gov/blast/db/