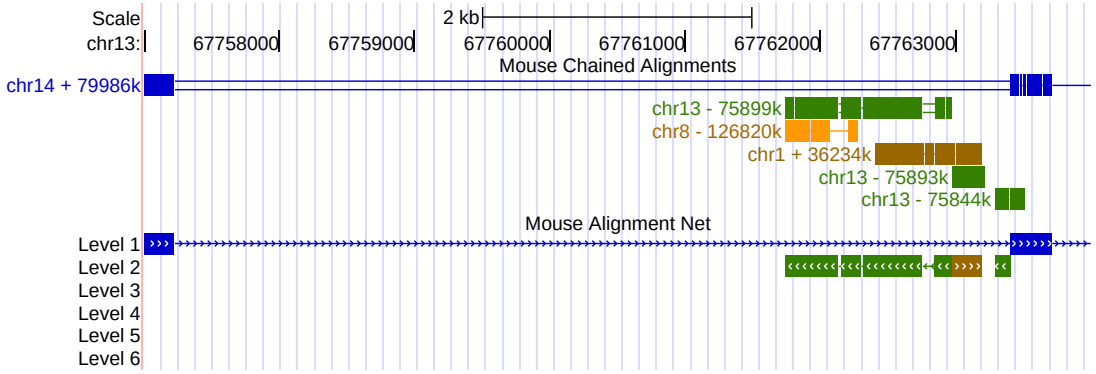
Announcements

- Submit your preferences for journal club papers using the form at the website until next Wednesday, Oct. 20 22:00
- Homework 1 will be published on the website, submit until Tuesday November 9 22:00 (pdf via Moodle, guests by e-mail)
- You are are allowed to discuss homework questions with classmates, but do not take notes during discussions and do not show your solutions to others. Everybody should write their homework submission independently, do not copy from classmates or other sources.
- Please use MS Teams for questions regarding homeworks, quizzes and the course in general.
- However, any questions involving your ideas about solving the questions should be sent privately to instructors by email.

Sequence alignment 2/2

Tomáš Vinař

October 14 2021



Summary from the last lecture

- **Global and local alignment problem**
  Input: sequences $X = x_1 x_2 \ldots x_n$ and $Y = y_1 y_2 \ldots y_m$.
  Output: alignment of X and Y with the highest score
  or alignment of **substrings** $x_i \ldots x_j$ a $y_k \ldots y_\ell$ with the highest score
- **Correct algorithms** using dynamic programming
- **Realistic scoring schemes**

We have dynamic programming, what else do we need?

Running time: $O(n^2)$ on two sequences of length n

How much is that in practice?

(simple implementation, random sequences, desktop computer)

n	time
100	0.0008s
1,000	0.08s
10,000	8s
100,000	13m (*)
1,000,000	22h (*)
10,000,000	3months (*)
100,000,000	25years (*)

We need a more efficient algorithm, particularly for comparative genomics

Memory: basic implementation $O(n^2)$, but can be done in $O(n)$

Heuristic alignment

- Trade sensitivity for speed (some alignments not found)
- Reduce the search to "promising" parts of the matrix

Heuristic local alignment

BLASTN [Altschul et al 1990], FASTA [Pearson 1988]

- Find short exact matches of length w (**seeds**)
- Extend hits along diagonals to ungapped alignments
- Connect alignments on nearby diagonals to gapped alignment
- Possibly optimize by dynamic programming

How to find short exact matches?

- Create a **dictionary** of short substrings of length w from the first sequence.
- Search for all substring from the second sequence in the dictionary

Exmple: CAGTCCTAGA vs CATGTCATA

Dictionary:

```
AG 2, 8
CA 1
CC 5
CT 6
GA 9
GT 3
TA 7
TC 4
```

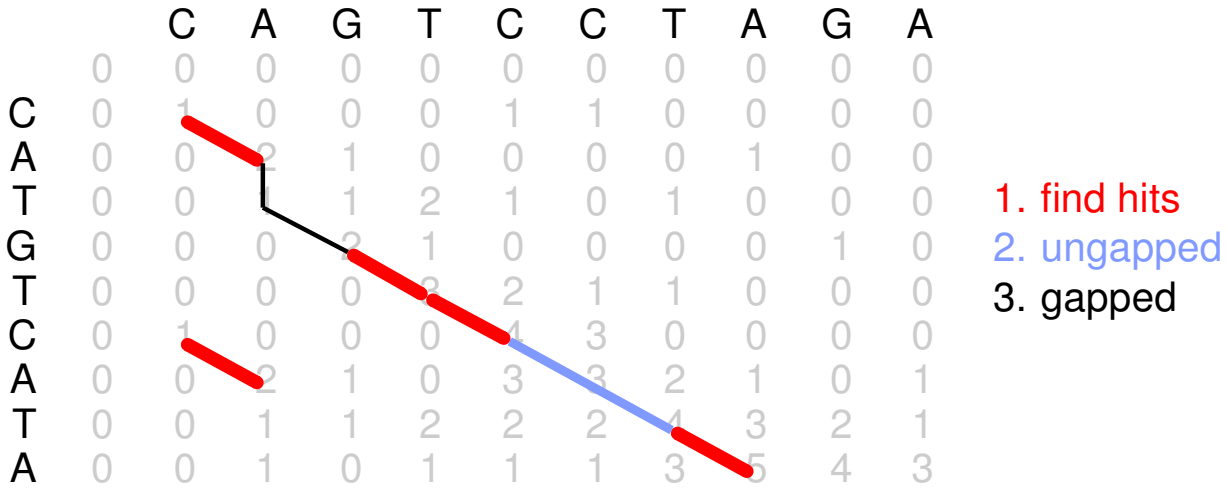
Search for:

```
CA → 1
AT → -
TG → -
GT → 3
TC → 4
CA → 1
AT → -
TA → 7
```

Heuristic local alignment

Example: start from **seeds** of length $w = 2$

(in practice we would use $w = 11$ or more)



Running time of heuristic local alignment

Algorithm

- Find seeds (short exact matches of length w)
- **Expensive step:** extend/connect seeds to longer alignments

Random seeds of length w: not part of any high-scoring alignment. These are filtered in the extension step, but they slow down the program

How many random hits?

Two unrelated nucleotides match with probability $1/4$

We have w matches in a row with probability 4^{-w}

Expected number of false positives roughly $nm4^{-w}$

Increase of w by 1 means cca 4-fold decrease of spurious seeds

Sensitivity of heuristic local alignment

Algorithm

- Find seeds (short exact matches of length w)
- **Expensive step:** extend/connect seeds to longer alignments

Some alignments not found: high score but no seed of length w

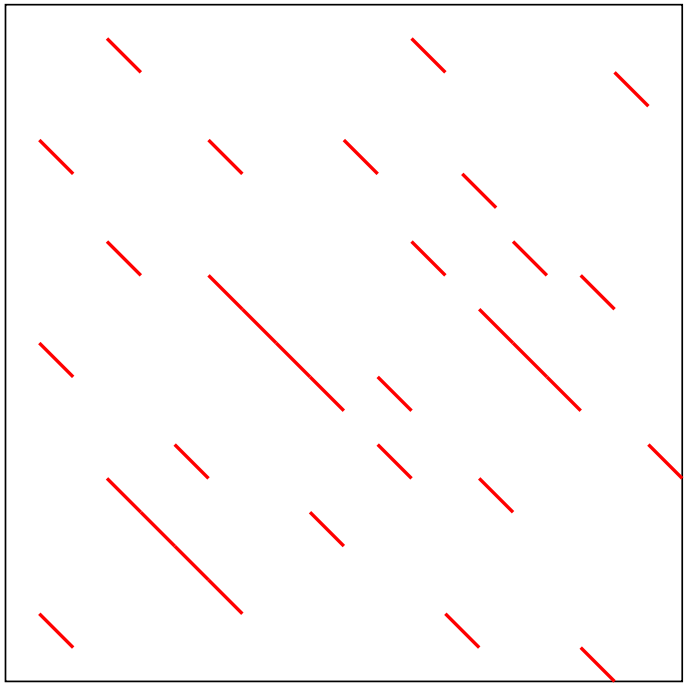
Example: `CA-GTCCTA` no seed of length $w \geq 4$
`CATGTCATA`

Sensitivity: fraction of real alignments containing a seed of length w

Sensitivity vs. running time

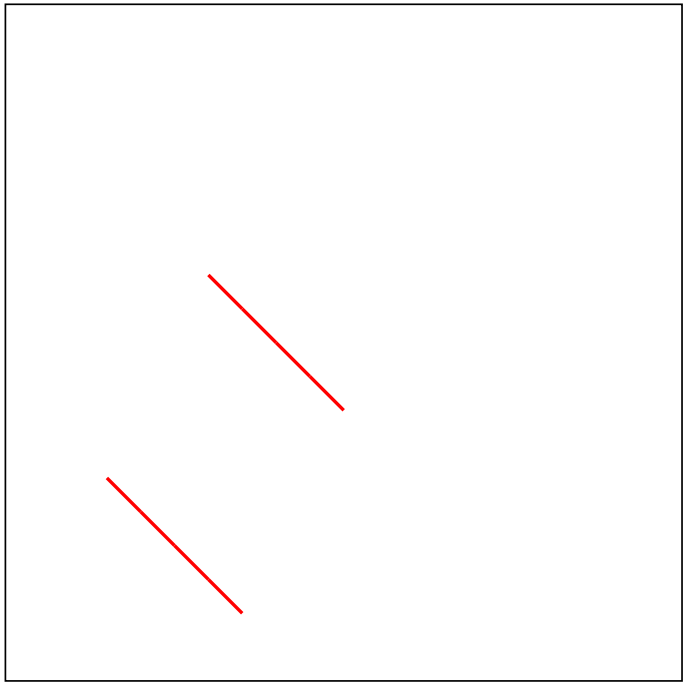
Small w

many spurious seeds, slow



Large w

many alignments not found

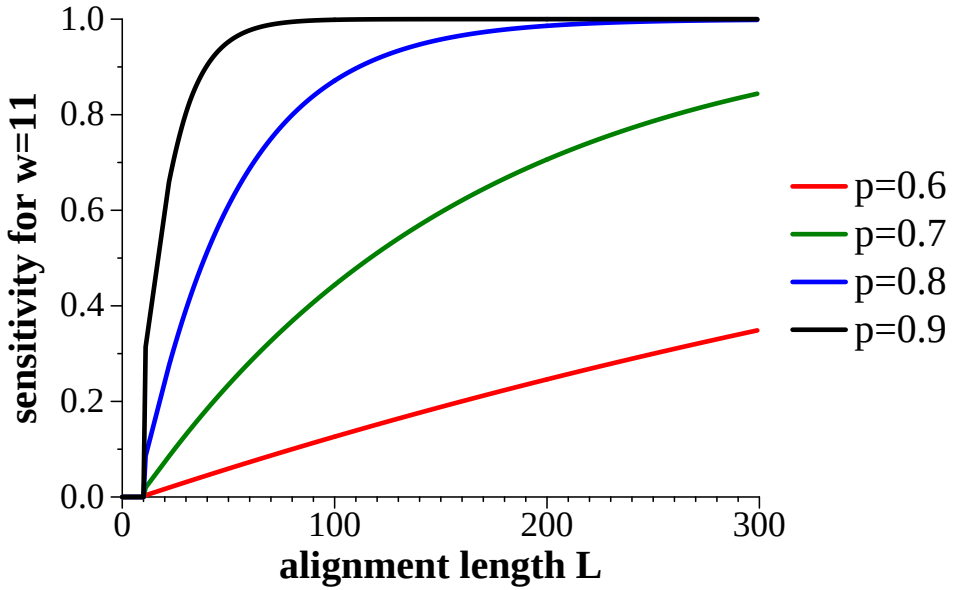


Can we estimate the sensitivity?

Assume random ungapped alignment of length L

Every position match with probability p

Sensitivity $f(L, p) = \Pr(\text{alignment contains } w \text{ consecutive matches})$



(human-mouse: $p \approx 0.7$)

Protein BLAST

BLOSUM62 scoring matrix for proteins

```
   A  R  N  D  C  Q  E  G  H  I ...
A  4 -1 -2 -2  0 -1 -1  0 -2 -1
R -1  5  0 -2 -3  1  0 -2  0 -3
N -2  0  6  1 -3  0  0  0  1 -3
D -2 -2  1  6 -3  0  2 -1 -1 -3
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1
Q -1  1  0  0 -3  5  2 -2  0 -3
E -1  0  0  2 -4  2  5 -2  0 -3
G  0 -2  0 -1 -3 -2 -2  6 -2 -4
H -2  0  1 -1 -3  0  0 -2  8 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
```

Instead of exact match of length w, protein BLAST requires 3 amino acids with score at least 13

Hit:

```
N I R
N L R
```

6+2+5=13

Not a hit:

```
A I L
A I L
```

4+4+4=12

Examples of software tools for various tasks

NCBI BLAST: blastn for DNA/RNA, blastp for proteins, tblastx translates DNA to proteins and uses blastp

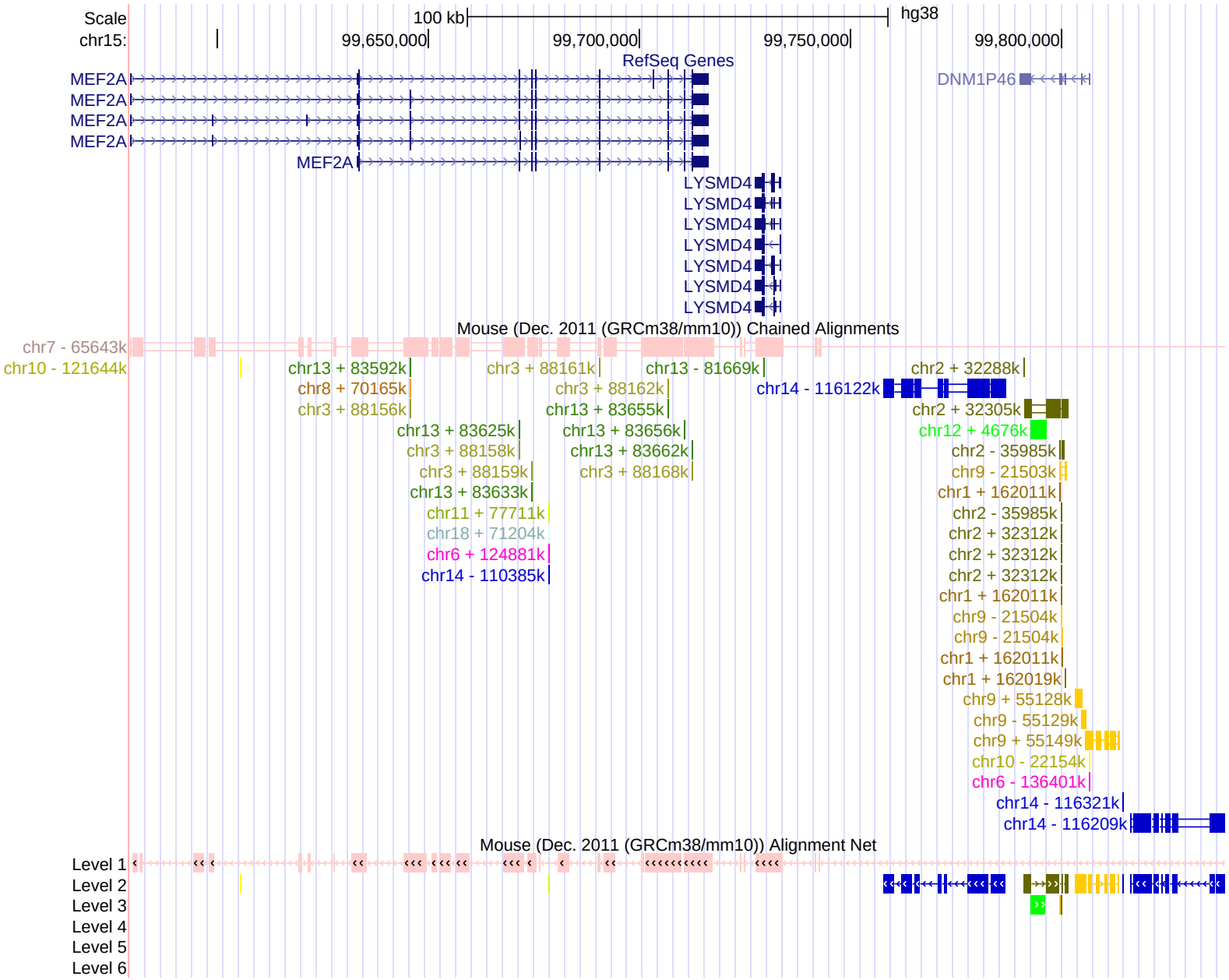
UCSC Blat: very fast search for very similar sequences, i.e. aligning sequencing reads to the genome

- uses very large values of w
- can split alignments with big gaps (aligning transcripts with introns)

Whole-genome alignments

For each section of human genome find closest section from mouse, dog, chicken, etc. (see e.g. UCSC genome browser)

- Local alignments will cover protein coding exons and other conserved parts
- Sections that diverged too much cannot be aligned
- If there was a duplication, we need to decide which pairs belong together
- **Synteny principle:** if two similar sections (local alignments) are present in the same order and orientation in two genomes, they likely evolved from the same common ancestor (orthologs)


Scale
chr15:
100 kb
hg38
99,650,000
99,700,000
99,750,000
99,800,000
RefSeq Genes
MEF2A
MEF2A
MEF2A
MEF2A
MEF2A
DNM1P46
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
LYSMD4
Mouse (Dec. 2011 (GRCm38/mm10)) Chained Alignments
chr7 - 65643k
chr10 - 121644k
chr13 + 83592k
chr8 + 70165k
chr3 + 88156k
chr3 + 88161k
chr3 + 88162k
chr13 + 83655k
chr13 - 81669k
chr13 + 83625k
chr3 + 88158k
chr3 + 88159k
chr13 + 83633k
chr11 + 77711k
chr18 + 71204k
chr6 + 124881k
chr14 - 110385k
chr13 + 83656k
chr13 + 83662k
chr3 + 88168k
chr2 + 32288k
chr14 - 116122k
chr2 + 32305k
chr12 + 4676k
chr2 - 35985k
chr9 - 21503k
chr1 + 162011k
chr2 - 35985k
chr2 + 32312k
chr2 + 32312k
chr2 + 32312k
chr1 + 162011k
chr9 - 21504k
chr9 - 21504k
chr1 + 162011k
chr1 + 162019k
chr9 + 55128k
chr9 - 55129k
chr9 + 55149k
chr10 - 22154k
chr6 - 136401k
chr14 - 116321k
chr14 - 116209k
Mouse (Dec. 2011 (GRCm38/mm10)) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6

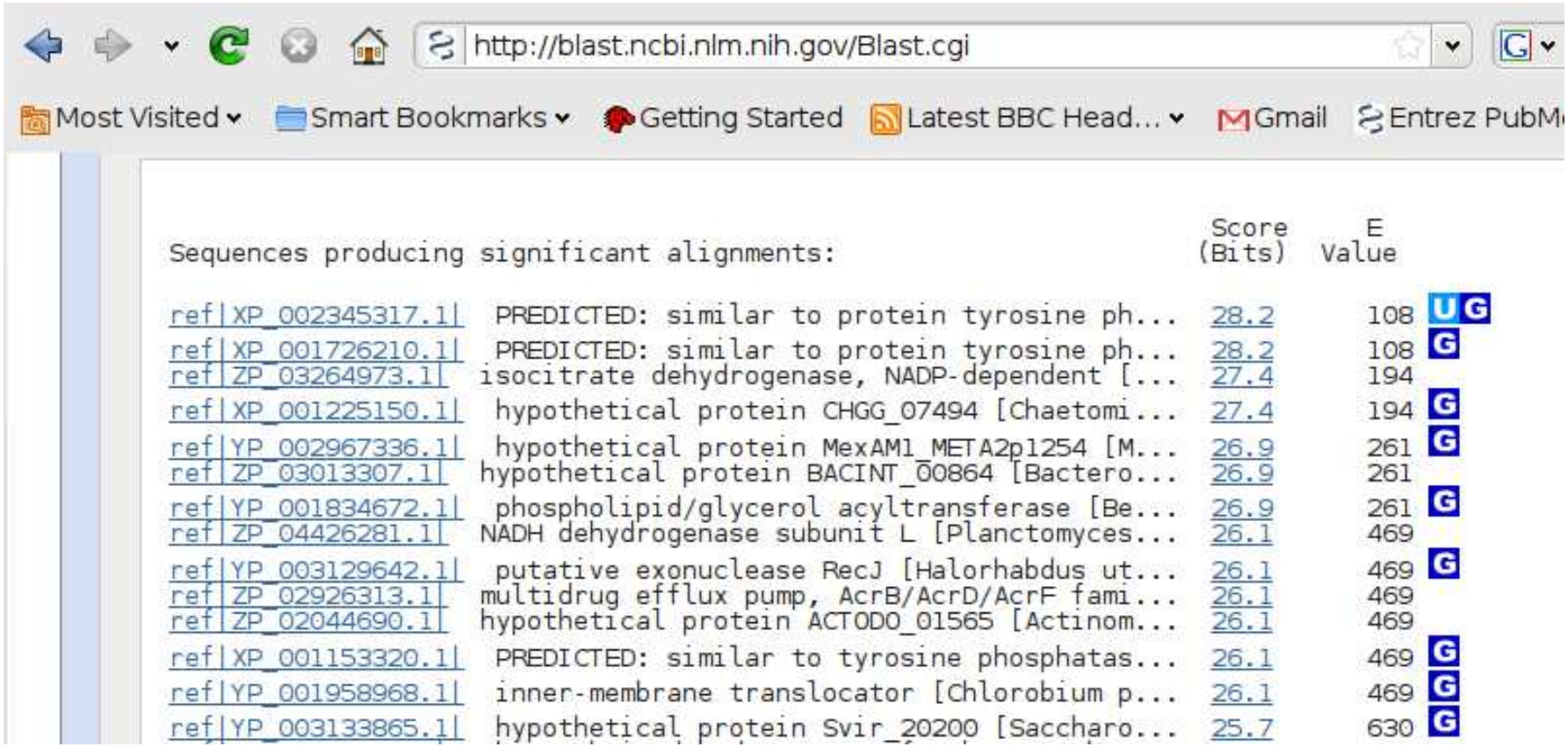
Multiple sequence alignment

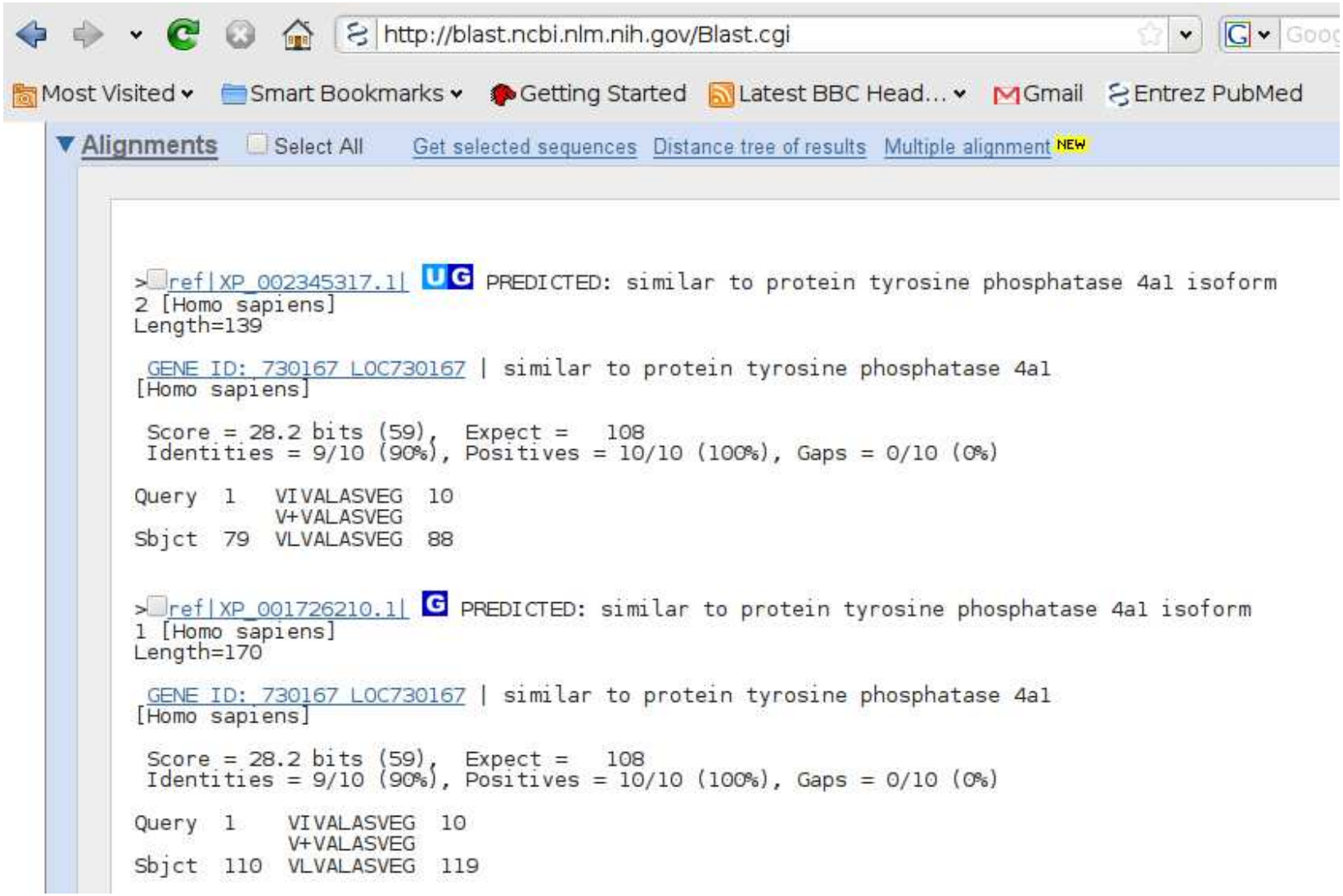
Running time: $O(2^k n^k)$ for k sequences of length n

For general k NP-hard.

```
      Human  ctccatagcaatgt-cagagatagggcagagcggat------ggtggtgac
     Rhesus  ctccatggcaatgt-cagagatagggcagagcggat------gctggtgac
      Mouse  ttt--tgacaaca--tagagac-tgagatagaaaat-------atgctgac
        Dog  -tccccgctaatgtacaaagatggggcag-gaaga--a----tgtgctgaa
      Horse  -tccacggcaatac-tggagatggggcagagcaga--agat-ggtgatgaa
  Armadillo  ctgcatagaaatct-cagagatgggggaaagcaga------agacattcat
    Opossum  atccatggaaacat-cagaagtgggagaaatagaaga----tggcaatga-
   Platypus  acccggggaagggg-aagaggaagggccggccg-----------------
```

Heuristic algorithms, e.g. CLUSTAL-W [Higgins et al., 1996], MUSCLE [Edgar, 2004] and TBA [Blanchette et al., 2004].


http://blast.ncbi.nlm.nih.gov/Blast.cgi
Most Visited
Smart Bookmarks
Getting Started
Latest BBC Head...
Gmail
Entrez PubM
Sequences producing significant alignments:
Score (Bits)
E Value
ref|XP_002345317.1| PREDICTED: similar to protein tyrosine ph... 28.2 108 U G
ref|XP_001726210.1| PREDICTED: similar to protein tyrosine ph... 28.2 108 G
ref|ZP_03264973.1| isocitrate dehydrogenase, NADP-dependent [... 27.4 194
ref|XP_001225150.1| hypothetical protein CHGG_07494 [Chaetomi... 27.4 194 G
ref|YP_002967336.1| hypothetical protein MexAM1_META2p1254 [M... 26.9 261 G
ref|ZP_03013307.1| hypothetical protein BACINT_00864 [Bactero... 26.9 261
ref|YP_001834672.1| phospholipid/glycerol acyltransferase [Be... 26.9 261 G
ref|ZP_04426281.1| NADH dehydrogenase subunit L [Planctomyces... 26.1 469
ref|YP_003129642.1| putative exonuclease RecJ [Halorhabdus ut... 26.1 469 G
ref|ZP_02926313.1| multidrug efflux pump, AcrB/AcrD/AcrF fami... 26.1 469
ref|ZP_02044690.1| hypothetical protein ACTODO_01565 [Actinom... 26.1 469
ref|XP_001153320.1| PREDICTED: similar to tyrosine phosphatas... 26.1 469 G
ref|YP_001958968.1| inner-membrane translocator [Chlorobium p... 26.1 469 G
ref|YP_003133865.1| hypothetical protein Svir_20200 [Saccharo... 25.7 630 G


http://blast.ncbi.nlm.nih.gov/Blast.cgi
Most Visited
Smart Bookmarks
Getting Started
Latest BBC Head...
Gmail
Entrez PubMed
Alignments
Select All
Get selected sequences
Distance tree of results
Multiple alignment
NEW
>ref|XP_002345317.1| PREDICTED: similar to protein tyrosine phosphatase 4a1 isoform 2 [Homo sapiens]
Length=139
GENE ID: 730167 LOC730167 | similar to protein tyrosine phosphatase 4a1 [Homo sapiens]
Score = 28.2 bits (59), Expect = 108
Identities = 9/10 (90%), Positives = 10/10 (100%), Gaps = 0/10 (0%)
Query 1 VIVALASVEG 10
V+VALASVEG
Sbjct 79 VLVALASVEG 88
>ref|XP_001726210.1| PREDICTED: similar to protein tyrosine phosphatase 4a1 isoform 1 [Homo sapiens]
Length=170
GENE ID: 730167 LOC730167 | similar to protein tyrosine phosphatase 4a1 [Homo sapiens]
Score = 28.2 bits (59), Expect = 108
Identities = 9/10 (90%), Positives = 10/10 (100%), Gaps = 0/10 (0%)
Query 1 VIVALASVEG 10
V+VALASVEG
Sbjct 110 VLVALASVEG 119

How to distinguish when the alignment is "real"?

Query length m. Database length n.
Alignment with score S.

P-value: Probability that a random query of length m in a random database of length n yields alignment of score at least S

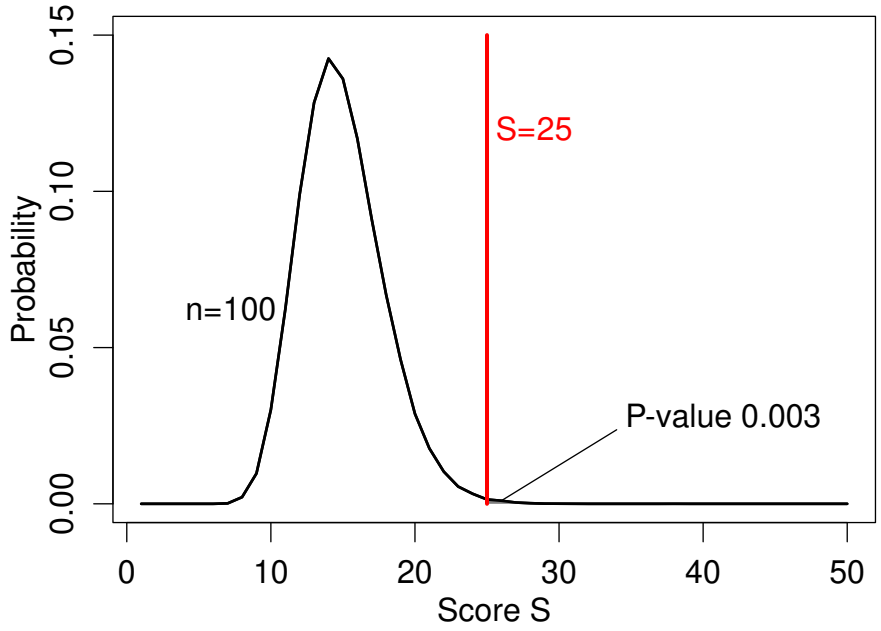
E-value: Expected number of alignments with the score of at least S when searching for a random query of length m in a random database of length n

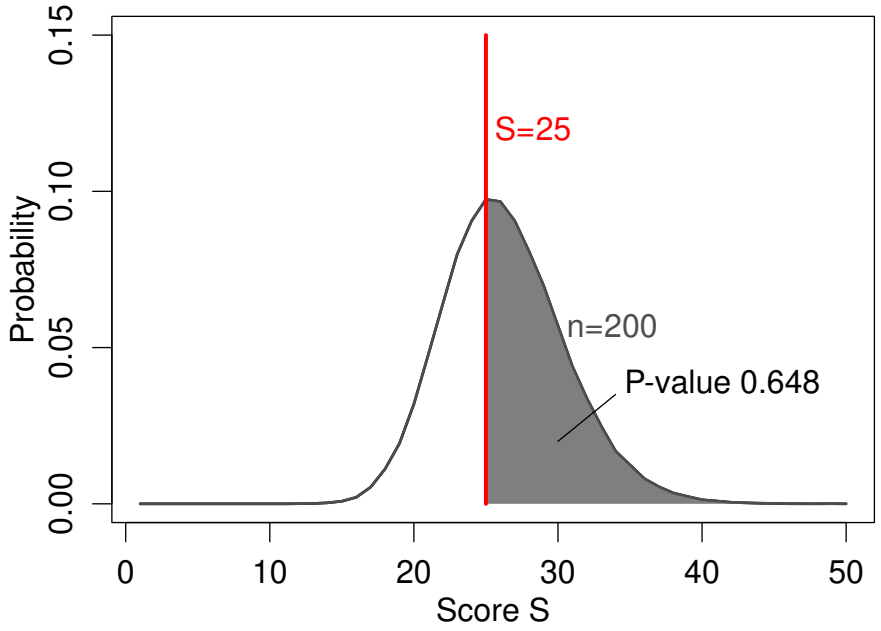
Note: $P = 1 - e^{-E} \Rightarrow$ for very small values of E, $P \approx E$

[Karlin and Altschul, 1990, Dembo et al., 1994]

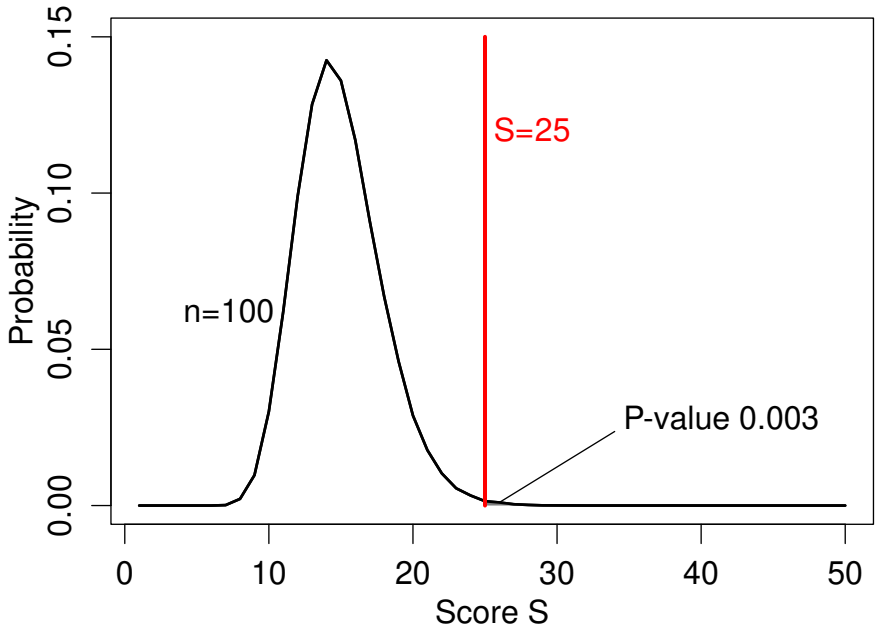
Computing P-values by simulation

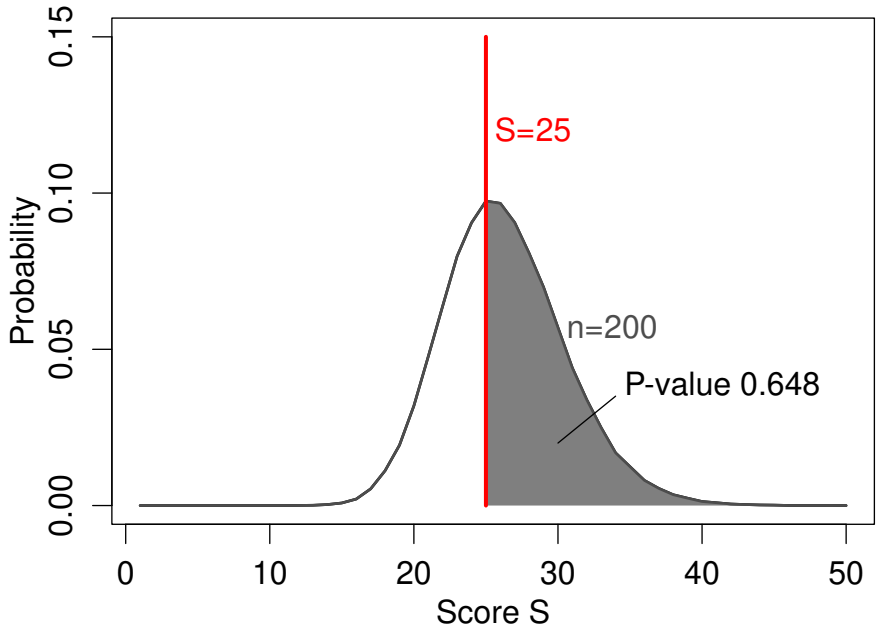
- Generate a random query and a random database of length n
- Compute best local alignment (+1/-1 scheme)
- Record the resulting score
- Repeat many times





Computing P-values by simulation (cont)





P-value for score 25:

How many alignments have score 25 or higher?

(In practice, simulations are slow, but we have mathematical estimates of how these distributions look like.)

Summary

- Sequence alignment is the essential bioinformatics tool
- Problem formulation: defining a scoring scheme
- Problem solution: either slow and exact algorithms, or fast heuristics that can miss some alignments
- There are specialized tools for various tasks related to the sequence alignment
- Estimation of statistical significance (P-values) is an important tool in distinguishing real alignments from those that occur just by chance