Homology

Bio5488

Ting Wang

1/25/15, 1/27/15

```
……ACGTTGCCACTTTCCGGGCCACCTGGCCACCTTATTTTCGGAAATATACCGGGCCTTTTTTT……
          ||||||x||||x||||||||
          CTTTCCCGGCCTCCTGGCCA
```

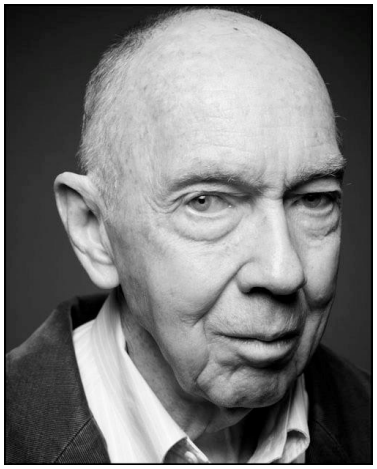
```
match: +1
mismatch: -1
matching score = 16
```

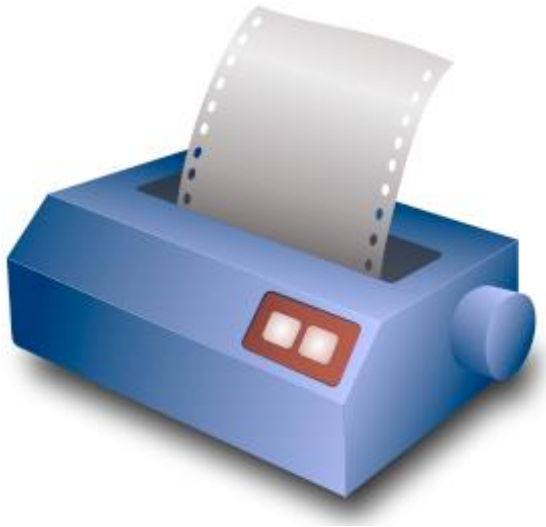
- **How to align them?**

- **Why we can align them?**
- **Why +1 for match, and -1 for mismatch?**
- **What does the score mean?**
- **Is 16 a good score?**

Outline

- Nobel-price-worthy work on homology
- What is homology?
- How to detect homology?
- How to quantify homology?
- How to use homology?
- Homology beyond sequence analysis
- Next-gen sequencing alignment



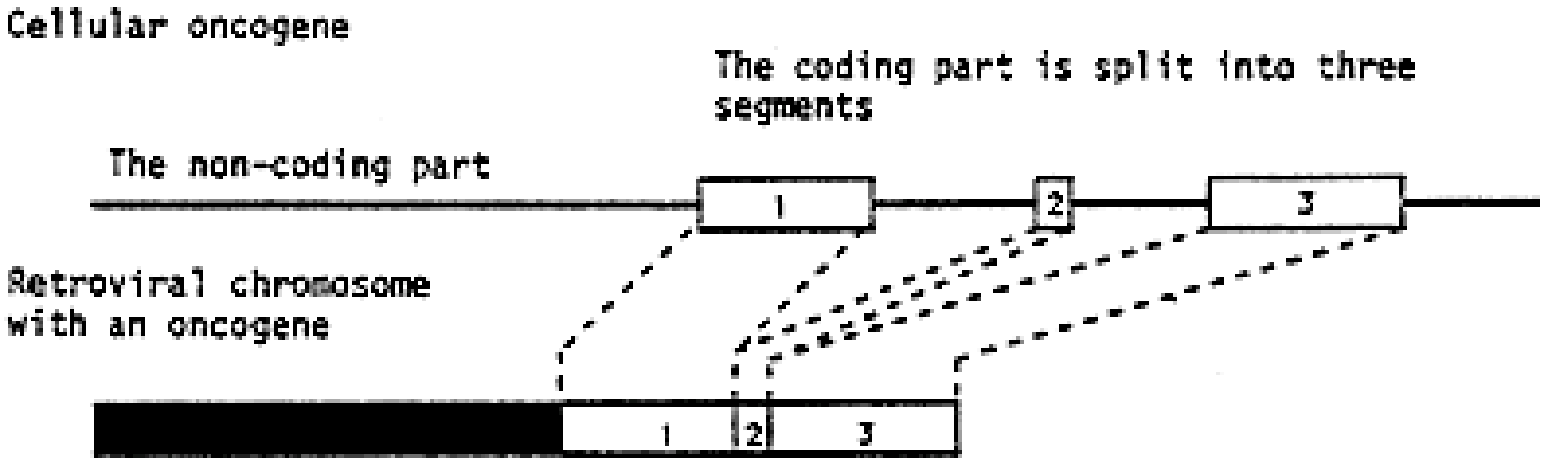


Russell Doolittle
(Bishop and Varmus)

```
p28sis        1   MTLTWQGDPIPEELYKMLSGHSIRSFDDLQRLLQGDSGKEDGAELDLNMT     50


p28sis       51   RSHSGGELESLARGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTN    100
PDGF-2        1                  SLGSLTIAEPAMIAECKTREEVFCICRRL?DR??     34
PDGF-1        1                        SIEEAVPAVCKTRIVIYEISRRELD???     28

p28sis      101   ANFLVWPPCVEVQRCSGCCNNRNVQCRPTQVQLRPVQVRKIEIVRKKPIF    150
PDGF-2       35   ??????PPCVEVKRCTGCCNNRNVKCRPSQVQLRP?QVRKIEIVRK[        80
PDGF-1       29   ANFL[                                                  32

p28sis      151   KKATVTLEDHLACKCEIVAAARAVTRSPGTSQEQRAKTTQSRVTIRTVRV    200
PDGF-2
PDGF-1

p28sis      201   RRPPKGKHRKCKHTHDKTALKETLGA                         226
PDGF-2                                     ]
PDGF-1                                     ]
```

Simian Sarcoma Virus onc Gene, v-sis, is Derived from the Gene (or Genes) Encoding a Platelet-Derived Growth Factor
Author(s): Russell F. Doolittle, Michael W. Hunkapiller, Leroy E. Hood, Sushilkumar G. Devare, Keith C. Robbins, Stuart A. Aaronson, Harry N. Antoniades
Source: *Science*, New Series, Vol. 221, No. 4607 (Jul. 15, 1983), pp. 275-277

Bishop and Varmus strategy (Nobel price 1989)



Doolittle strategy (could be the first Nobel price for computational biology)

tween these proteins. This similarity was discovered by one of us (R.F.D.) during a search for sequence homology between the PDGF amino-terminal sequences and the other protein sequences in the Newat sequence data base at the University of California, San Diego (*19*). Subsequent-

base searched included 145,581 amino acid residues comprising 684 individual sequences in the Newat list and 121,098 residues from 1081 sequences in the 1978 Dayhoff collection [*Protein*

What is the significance?

A few Definitions

Homologs: genes/sequences sharing a common origin

Orthologs: genes originating from a single ancestral gene in the last common ancestor of the compared genomes; genes related via speciation

Paralogs: genes related via duplication

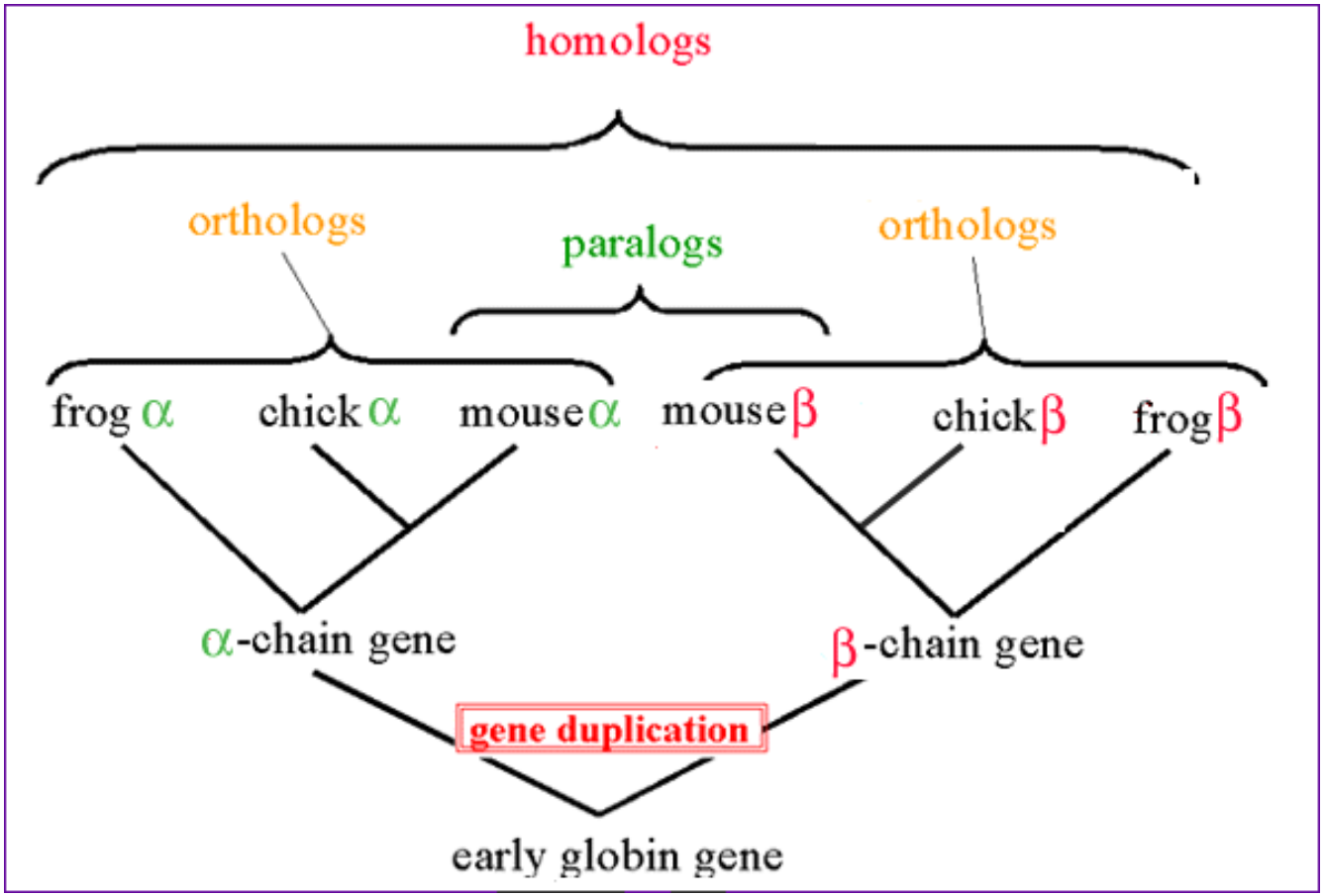
Xenolog: sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

Co-orthologs: two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage due to a lineage-specific duplication(s)

Outparalogs: paralogous genes resulting from a duplication(s) preceding a given speciation event

Inparalogs: paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event

Relation of sequences



Need ancestral sequences to distinguish orthologs and paralogs

(exercise on board)

Similarity versus Homology

- Similarity refers to the likeness or % identity between 2 sequences
- Similarity means sharing a statistically significant number of bases or amino acids
- Similarity does not imply homology
- Similarity can be quantified
- It is ok to say that two sequences are X% identical
- It is ok to say that two sequences have a similarity score of Z
- It is generally incorrect to say that two sequences are X% *similar*

- Homology refers to shared ancestry
- Two sequences are homologous if they are derived from a common ancestral sequence
- Homology usually implies similarity
- Low complexity regions can be highly similar without being homologous
- Homologous sequences are not always highly similar
- A sequence is either homologous or not.
- Never say two things are X% homologous

Why Compare Sequences?

- Sequence comparisons lie at the heart of all bioinformatics
- Identify sequences
  - What is this thing I just found?
- Compare new genes to known ones
- Compare genes from different species
  - information about evolution
- Guess functions for entire genomes full of new gene sequences
  - Metagenomics
- What does it matter if two sequences are similar or not?
  - Globally similar sequences are likely to have the same biological function or role
  - Locally similar sequences are likely to have some physical shape or property with similar biochemical roles
  - If we can figure out what one does, we may be able to figure out what they all do

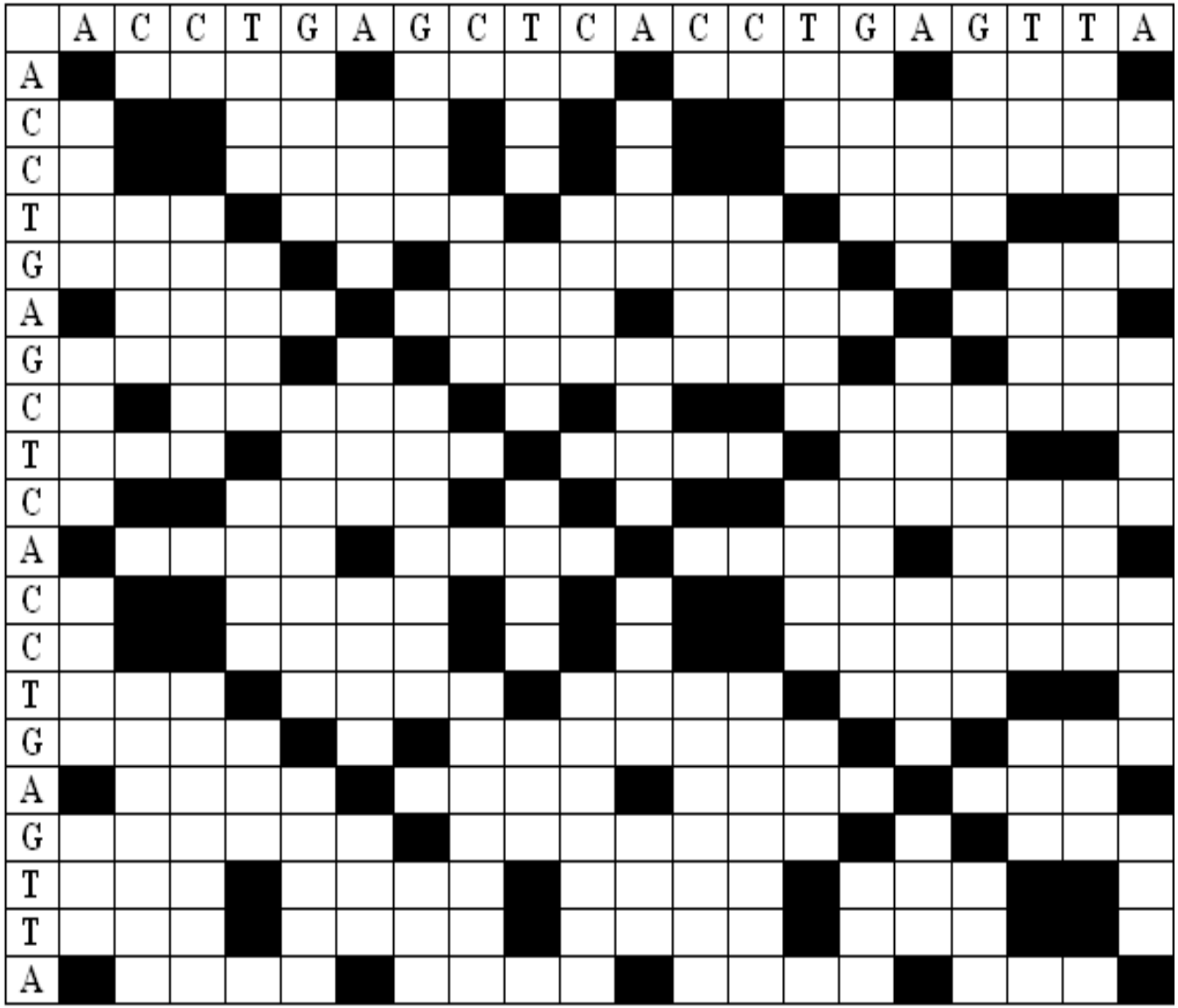
Sequence alignment

- How to optimally align two sequences
  - Dot plots
  - Dynamic programming
    - Global alignment
    - Local alignment
- How to score an alignment
- Fast similar sequence search
  - BLAST
  - BLAT
  - More recent development: short read alignment
- Determine statistical significance
- Using information in multiple sequence alignment to improve sensitivity

Visual Alignments (Dot Plots)

- Build a comparison matrix
  - Rows: Sequence #1
  - Columns: Sequence #2

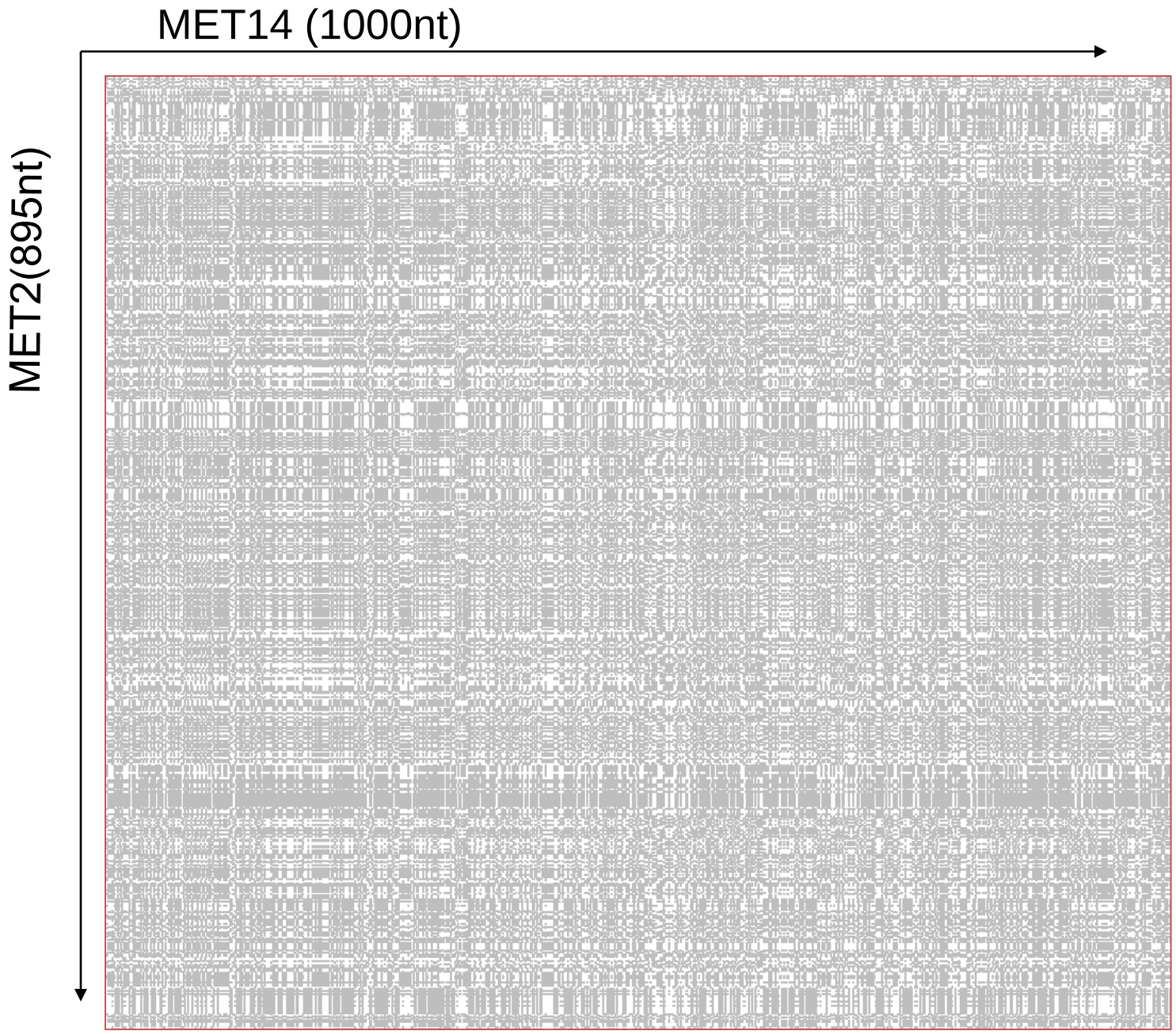
- Filling
  - For each coordinate, if the character in the row matches the one in the column, fill in the cell
  - Continue until all coordinates have been examined


A C C T G A G C T C A C C T G A G T T A
A
C
C
T
G
A
G
C
T
C
A
C
C
T
G
A
G
T
T
A

Noise in Dot Plots

- Nucleic Acids (DNA, RNA)
  - 1 out of 4 bases matches at random
- Windowing helps reduce noise
  - Can require >X bp match before plotting
  - Percentage of bases matching in the window is set as threshold

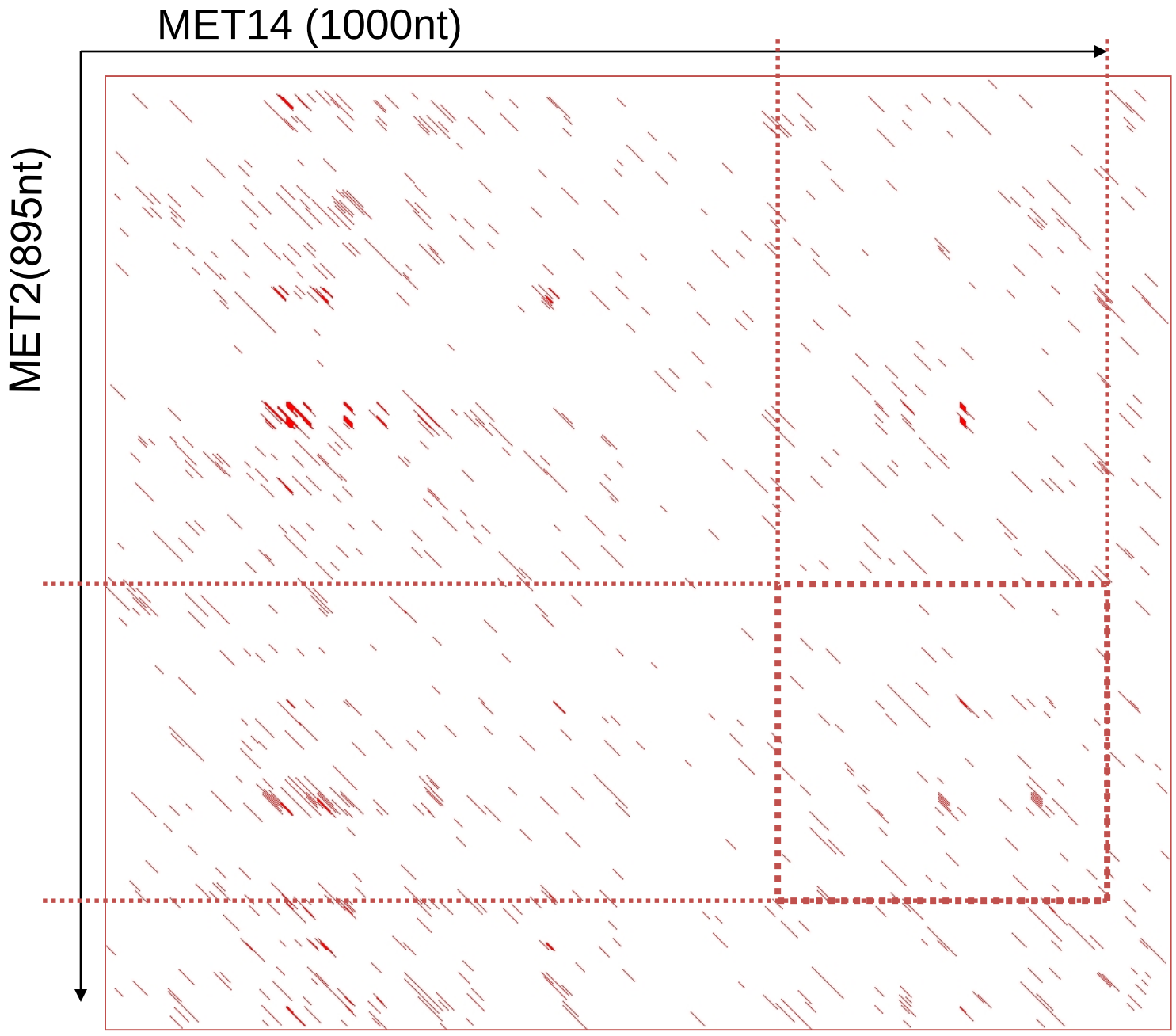
Met14 vs Met2 "DotPlot"



Match = 1
Mismatch =-1
Gray: 1

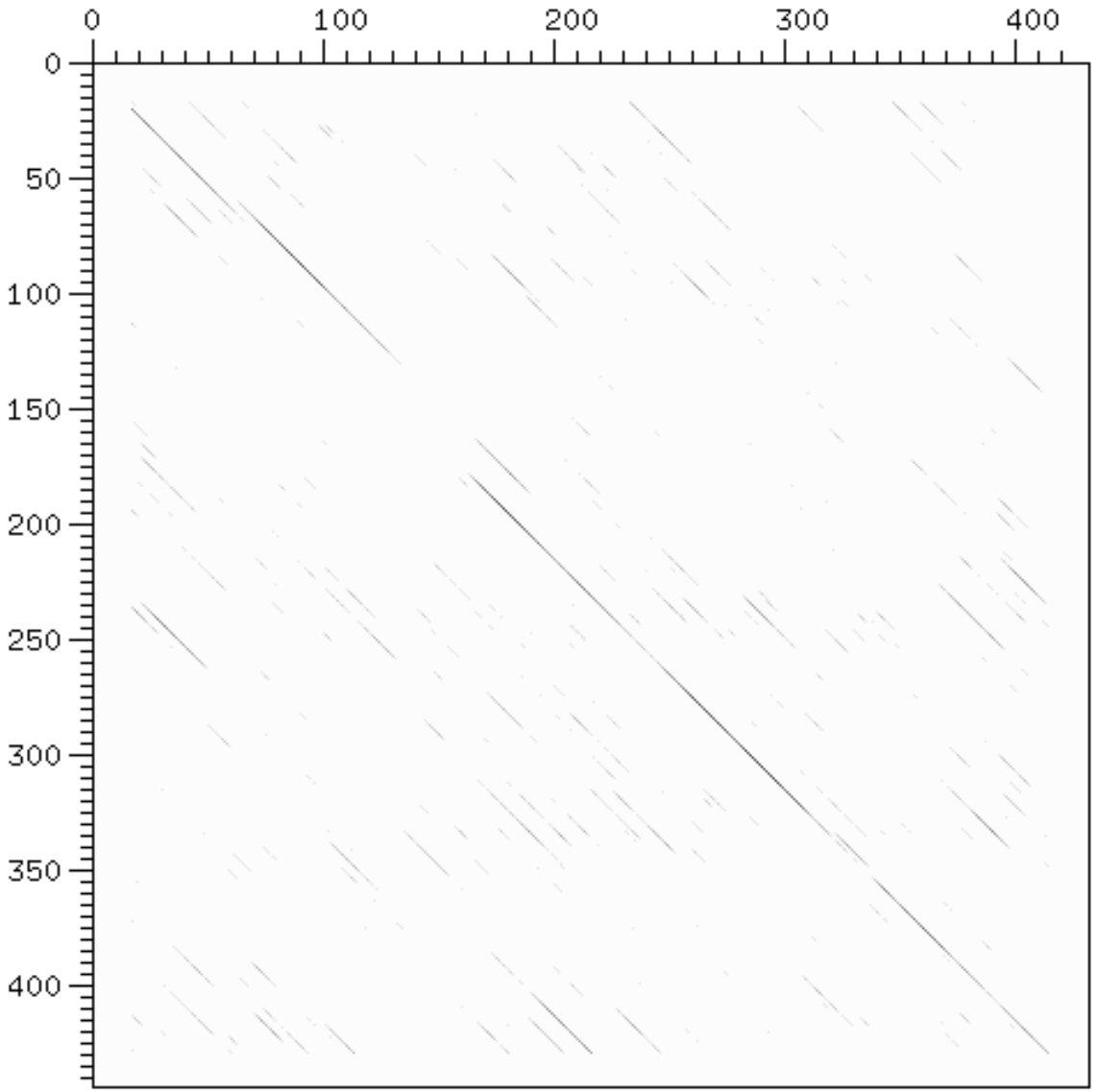
Met14 vs Met2

Red: >5

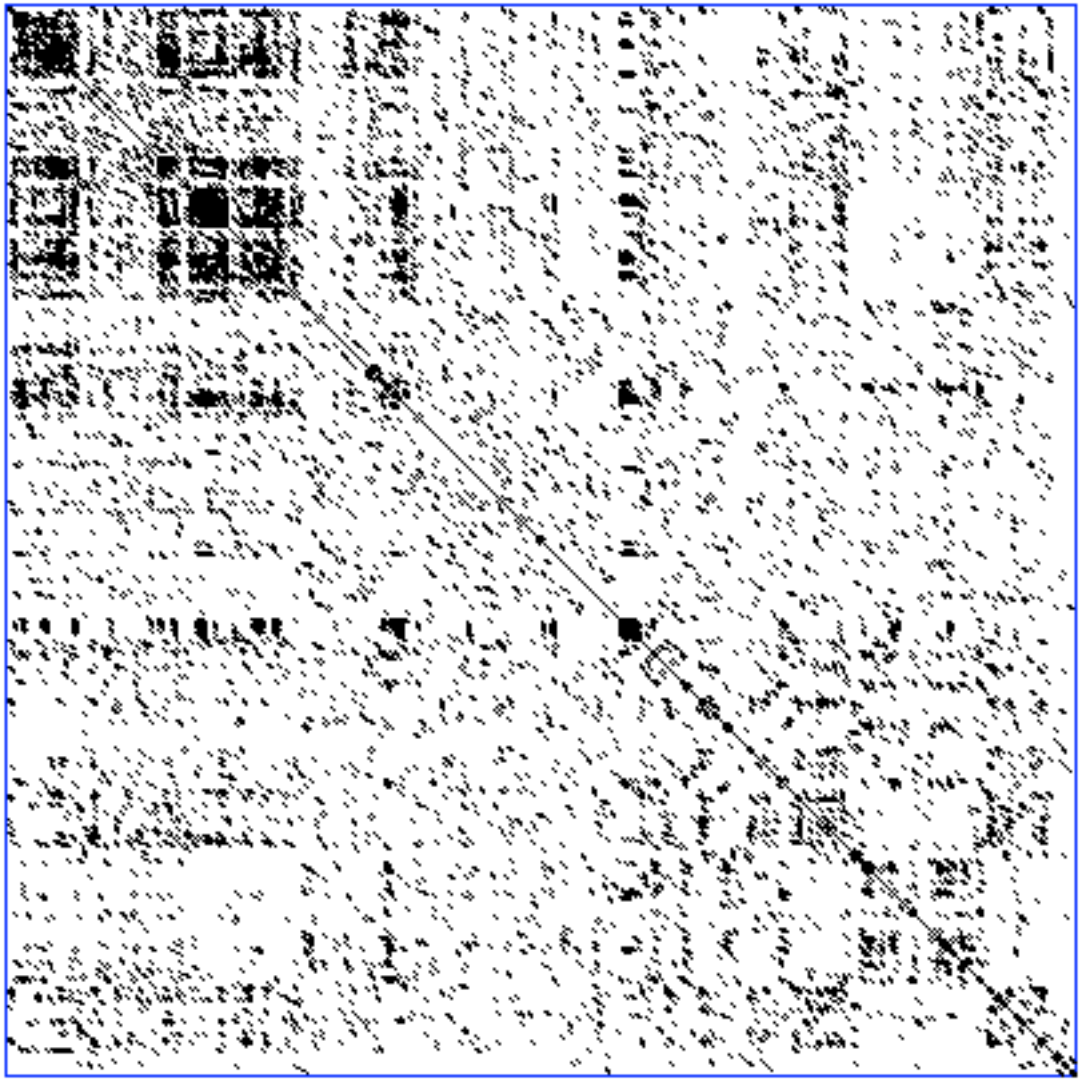


α chain of human hemoglobin

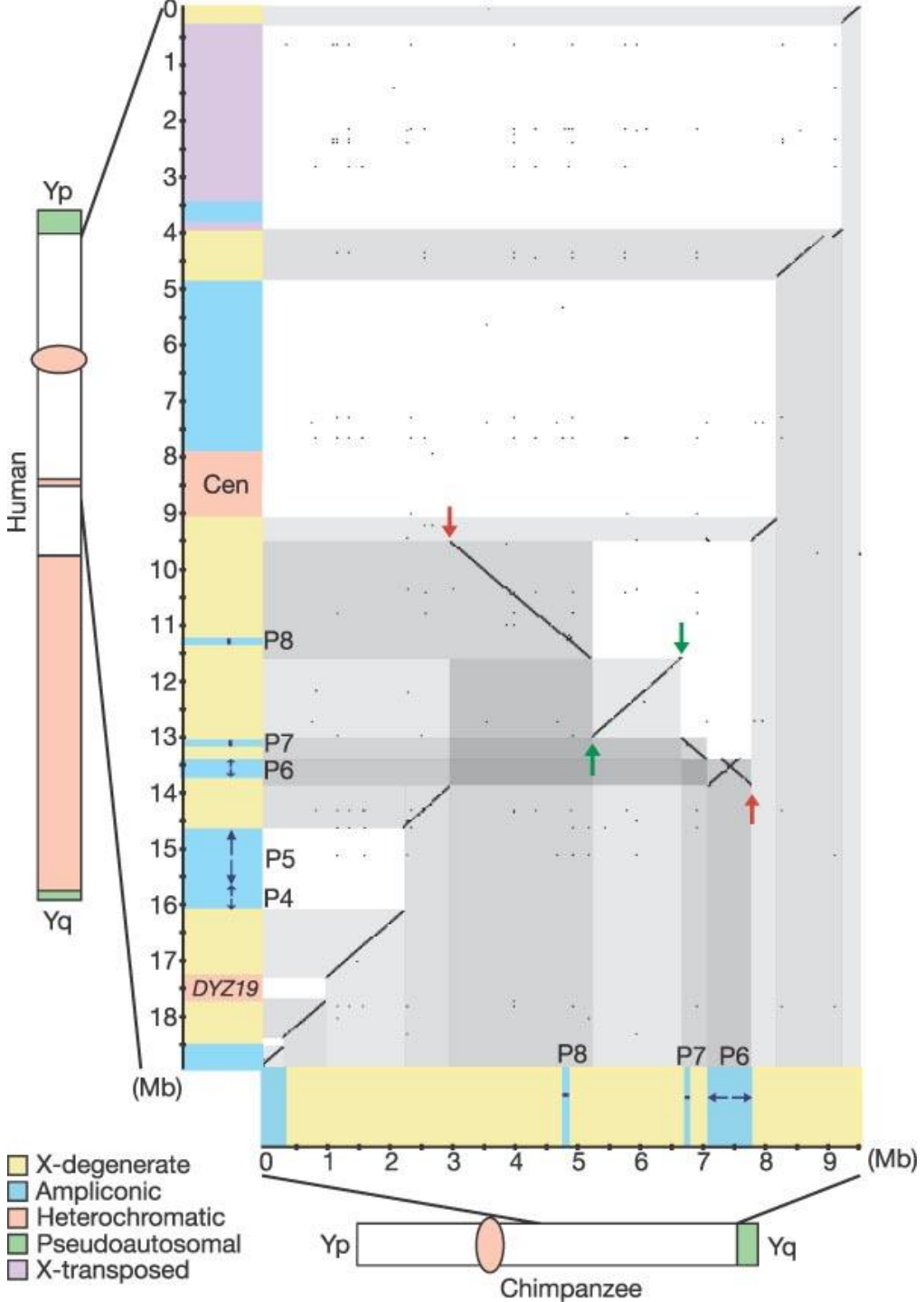
β chain of human hemoglobin



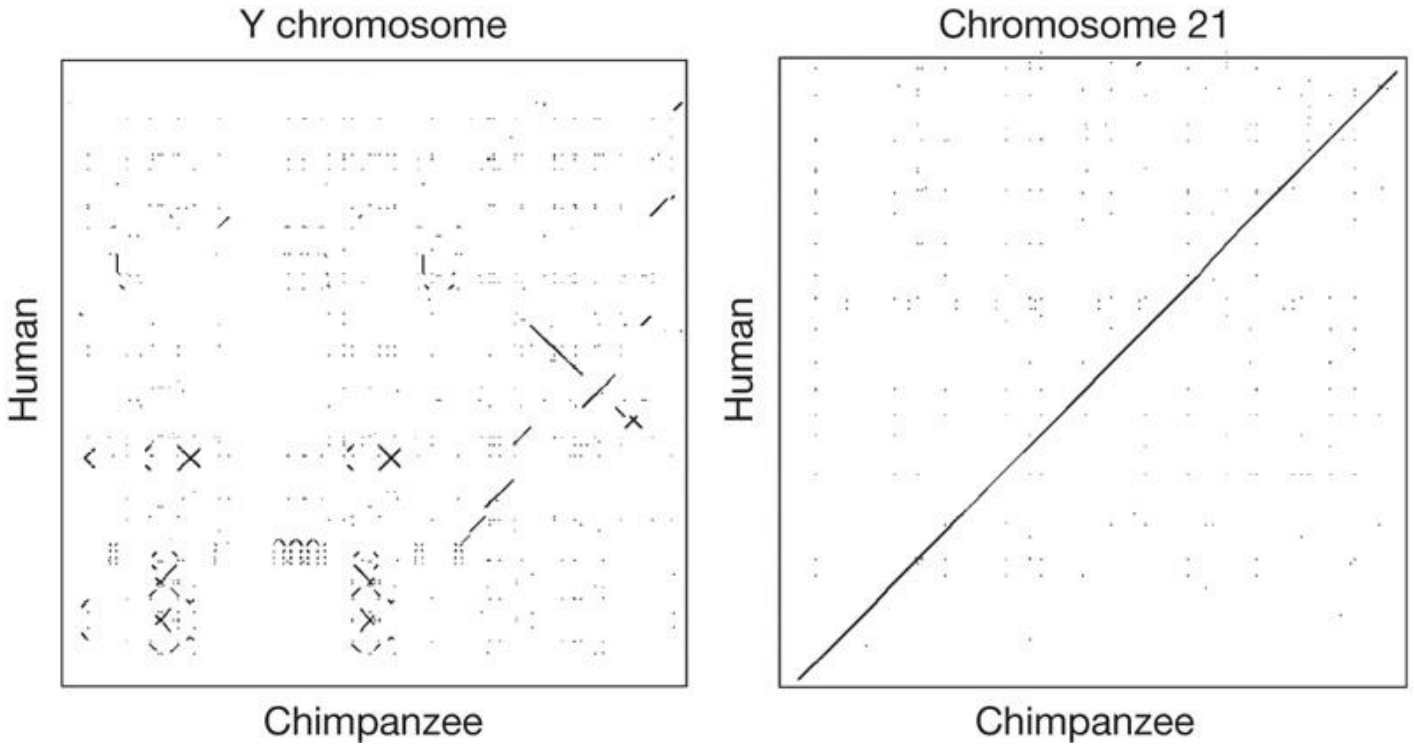
MAZ: Myc associated zinc finger isoform 1 self alignment



Human vs Chimp Y chromosome comparison



Dot plots of DNA sequence identity between chimpanzee and human Y chromosomes and chromosomes 21.



JF Hughes *et al. Nature* **000**, 1-4 (2010) doi:10.1038/nature08700

Aligning sequences by residue

- Match: award
- Mismatch (substitution or mutation): penalize
- Insertion/Deletion (INDELS – gaps): penalize (gap open, gap extension)

```
A L I G N M E N T
  | | |   | | | |
- L I G A M E N T
```

More than one solution is possible

- Which alignment is best?

```
A T C G G A T - C T
|   |   |       | |
A - C - G G - A C T
```

```
A T C G G A T C T
|   | | |     | |
A - C G G - A C T
```

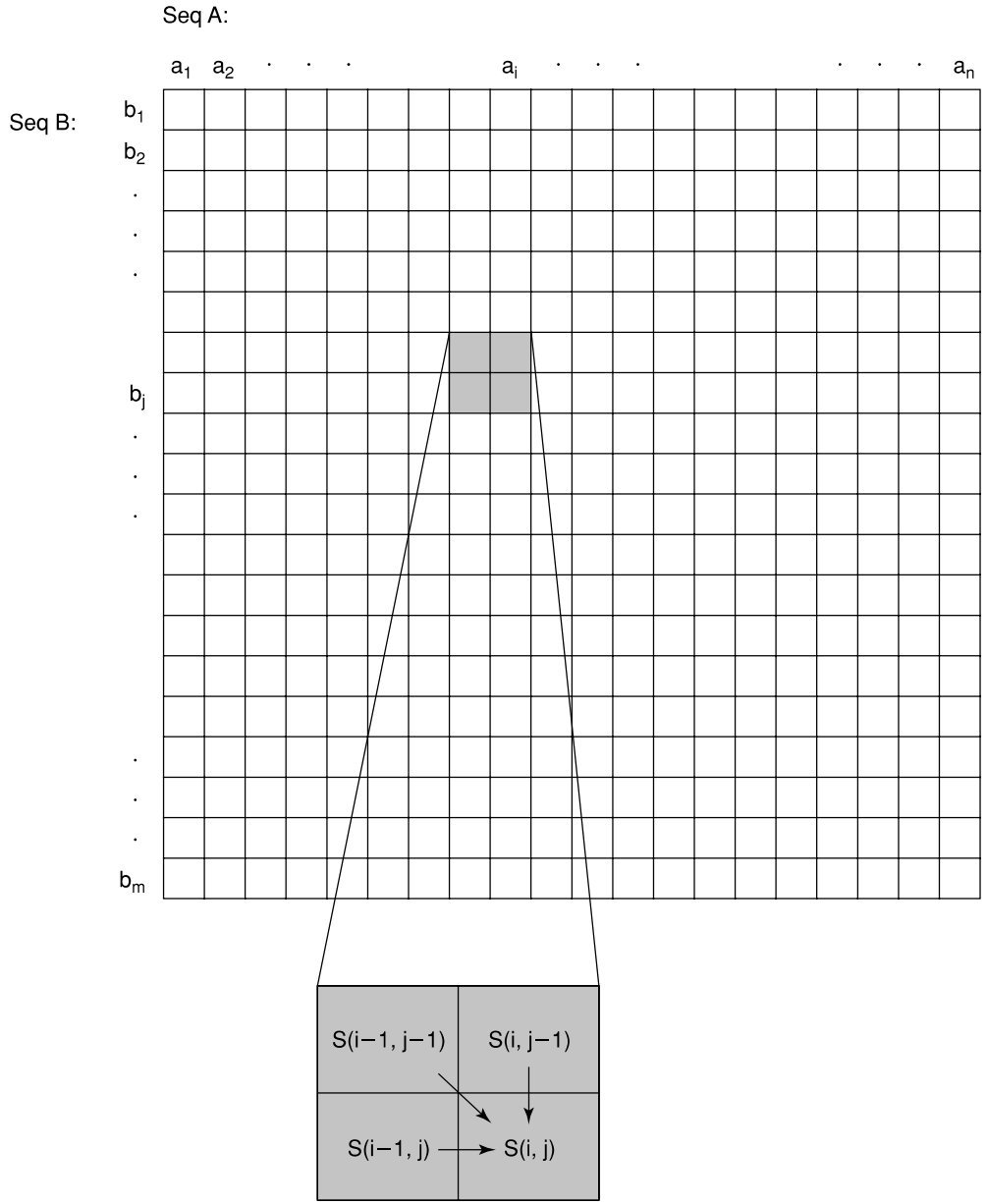
Alignment Scoring Scheme

- Possible scoring scheme:
  - match: +2
  - mismatch: -1
  - indel –2

- Alignment 1: 5*2 + 1*-1 + 4*-2 = 10 – 1 – 8 = 1
- Alignment 2: 6*2 + 1*-1 + 2*-2 = 12 – 1 – 4 = 7

Dynamic Programming

- Global Alignments:
  - Needleman S.B. and Wunsch C.D. (1970) *J. Mol. Biol.* 48, 443-453
- Local Alignments:
  - Smith T.F. and Waterman M.S. (1981) *J. Mol. Biol.*147, 195-197
  - One simple modification of Needleman/Wunsch: when a value in the score matrix becomes negative, reset it to zero (begin of new alignment)
- Guaranteed to be mathematically optimal:
  - Given two sequences (and a scoring system) these algorithms are guaranteed to find the very best alignment between the two sequences!
- Slow N^2 algorithm
- Performed in 2 stages
    - Prepare a scoring matrix using recursive function
    - Scan matrix diagonally using traceback protocol



$$S(i,j) = max \begin{cases} S(i-1, j-1) + score(a_i, b_j) \\ S(i, j-1) + \delta \\ S(i-1, j) + \delta \end{cases}$$

Dynamic Programming

	G	E	N	E	T	I	C	S
G	10	0	0	0	0	0	0	0
E	0	10	0	10	0	0	0	0
N	0	0	10	0	0	0	0	0
E	0	0	0	10	0	0	0	0
S	0	0	0	0	0	0	0	10
I	0	0	0	0	0	10	0	0
S	0	0	0	0	0	0	0	10

	G	E	N	E	T	I	C	S
G	60	40	30	20	20	0	10	0
E	40	50	30	30	20	0	10	0
N	30	30	40	20	20	0	10	0
E	20	20	20	30	20	10	10	0
S	20	20	20	20	20	0	10	10
I	10	10	10	10	10	20	10	0
S	0	0	0	0	0	0	0	10

```
G  E  N  E  T  I  C  S
|  |  |  |  *  |     |
G  E  N  E  S  I     S
```

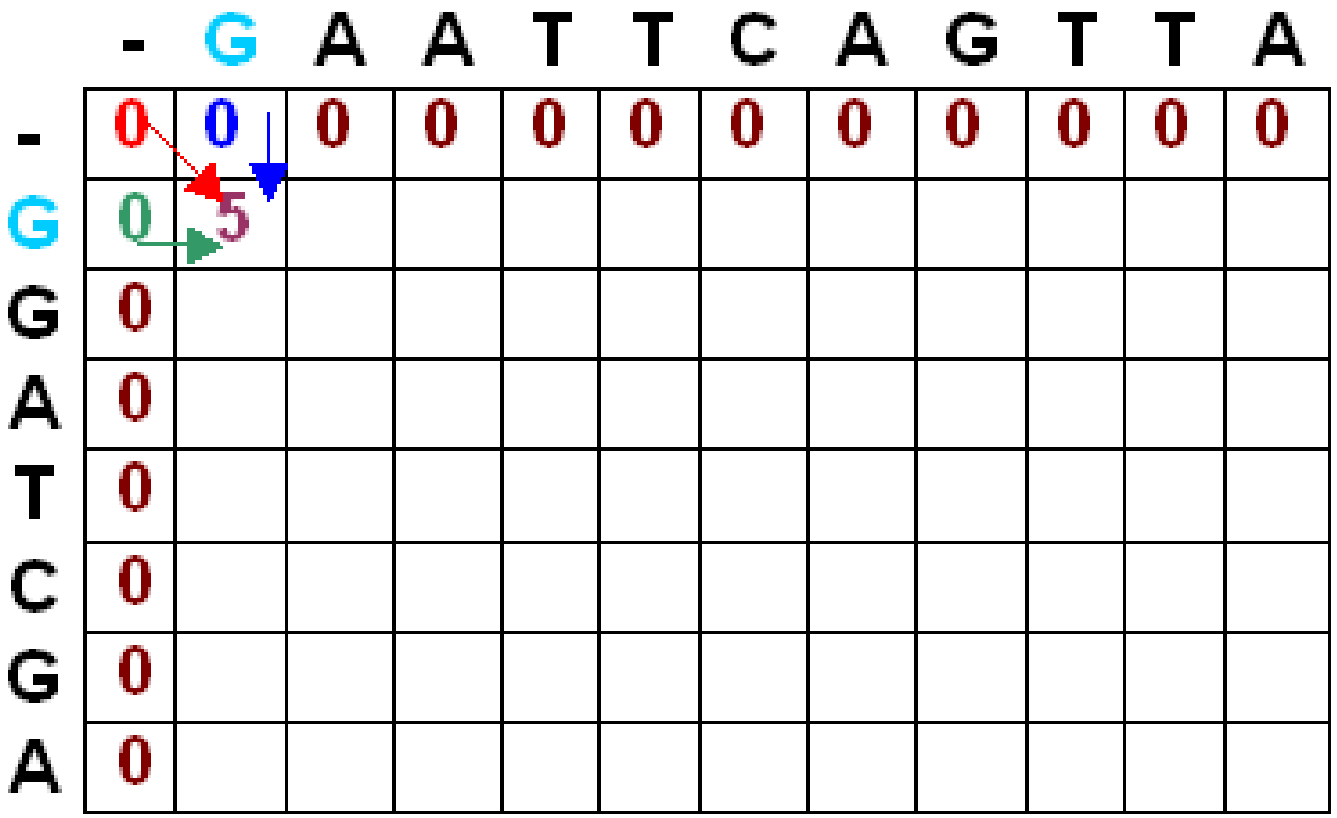
DP (demo)

- Match=5, mismatch=-3, gap=-4

	-	G	A	A	T	T	C	A	G	T	T	A
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

DP (demo)

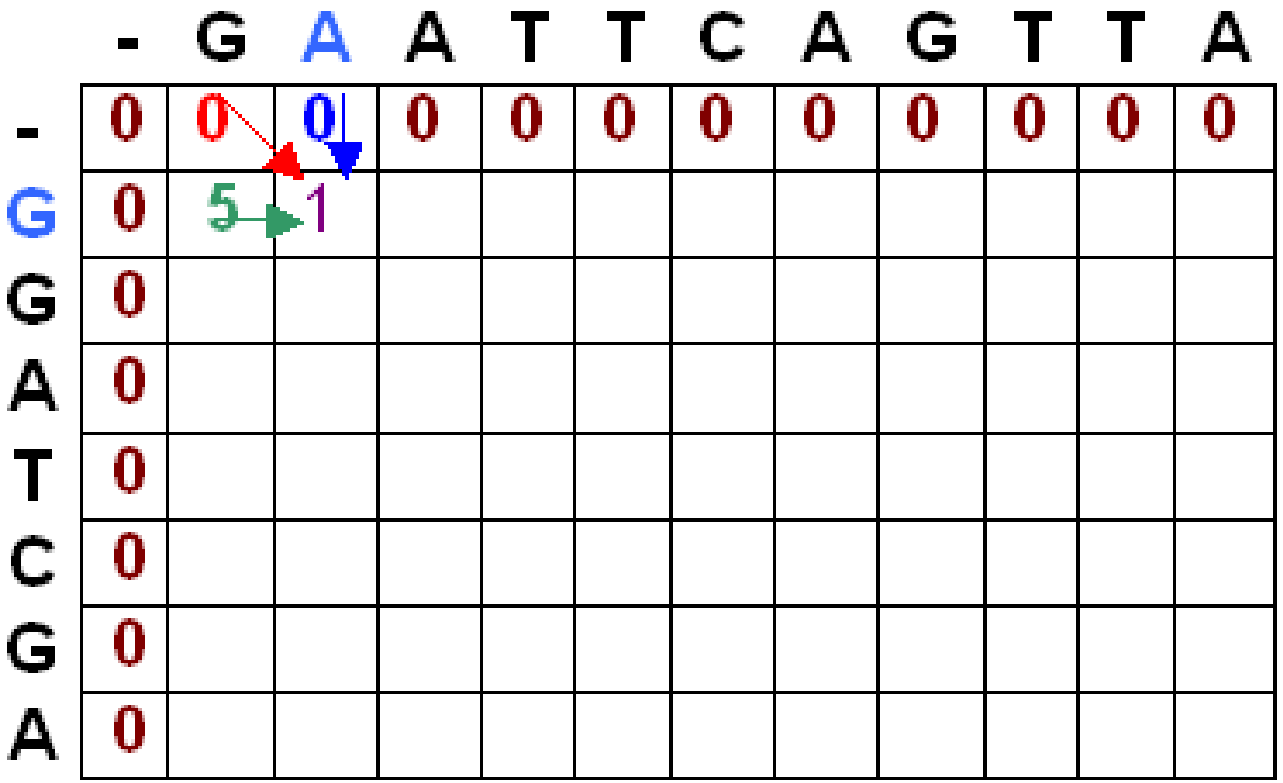
$S_{1,1} = MAX\{S_{0,0} + 5, S_{1,0} - 4, S_{0,1} - 4, 0\} = MAX\{5, -4, -4, 0\} = 5$



DP (demo)

$S_{1,2} = MAX\{S_{0,1} - 3, S_{1,1} - 4, S_{0,2} - 4, 0\} = MAX\{0 - 3, 5 - 4, 0 - 4, 0\} =$

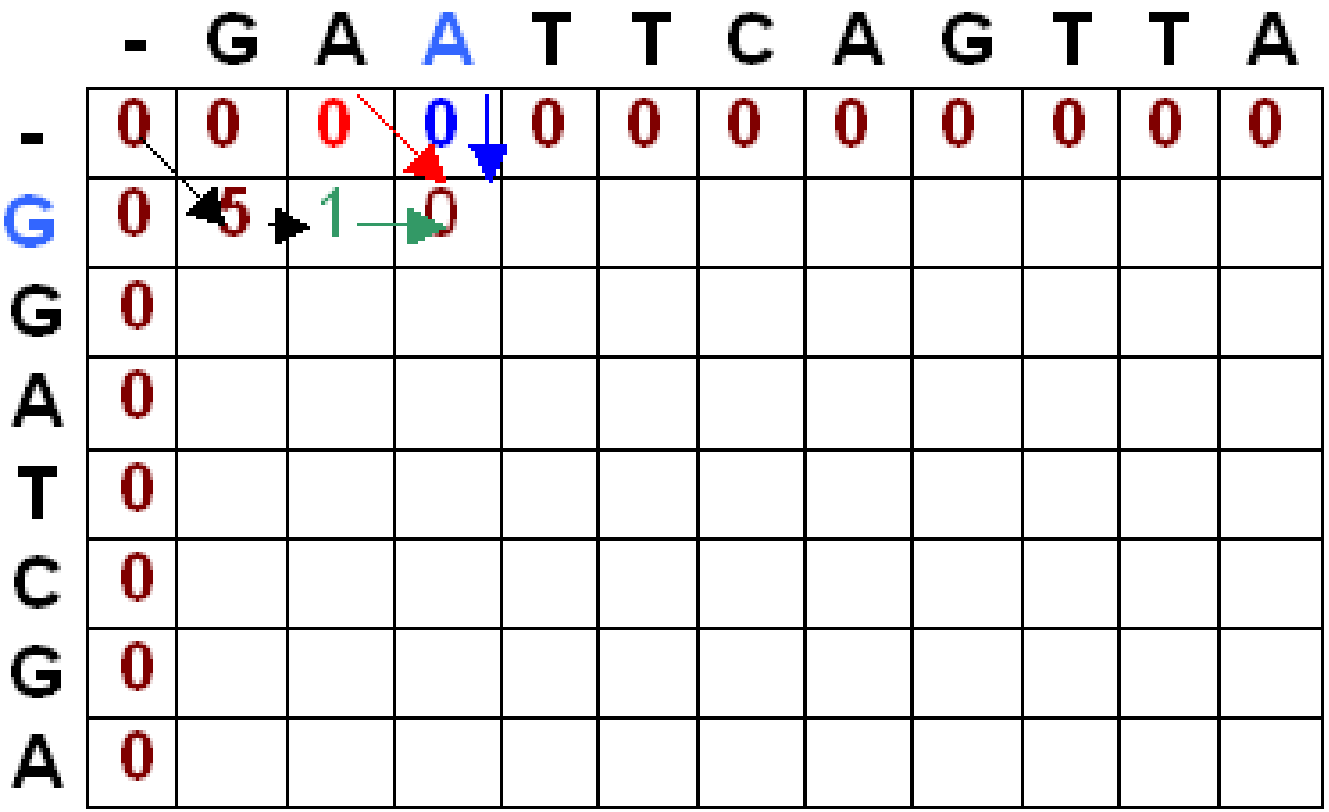
$MAX\{-3, 1, -4, 0\} = 1$



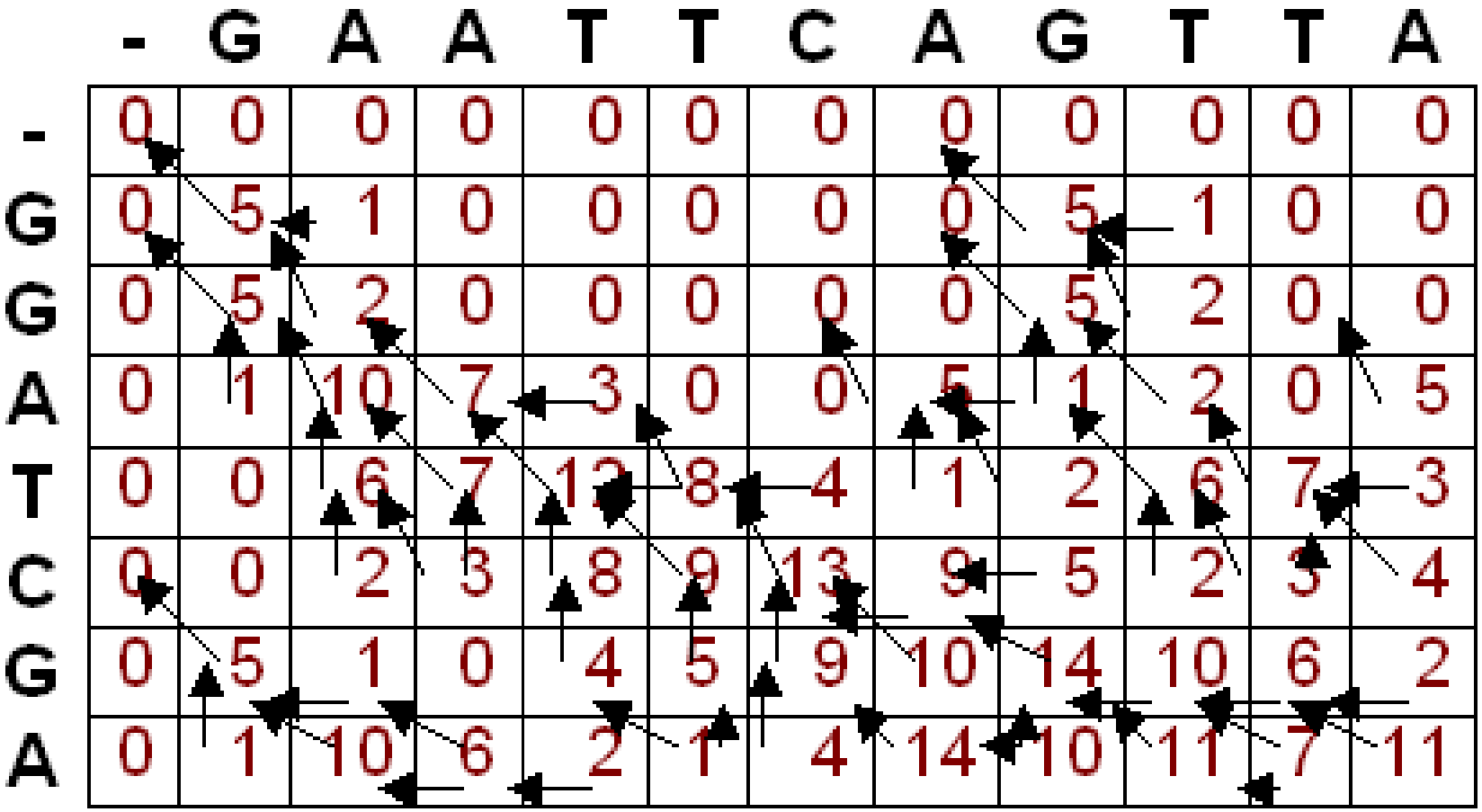
DP (demo)

$S_{1,3} = MAX\{S_{0,2} - 3, S_{1,2} - 4, S_{0,3} - 4, 0\} = MAX\{0 - 3, 1 - 4, 0 - 4, 0\} =$

$MAX\{-3, -3, -4, 0\} = 0$



DP (demo)


- G A A T T C A G T T A
-
G
G
A
T
C
G
A
0 0 0 0 0 0 0 0 0 0 0 0
0 5 1 0 0 0 0 0 5 1 0 0
0 5 2 0 0 0 0 0 5 2 0 0
0 1 10 7 3 0 0 5 1 2 0 5
0 0 6 7 12 8 4 1 2 6 7 3
0 0 2 3 8 9 13 9 5 2 3 4
0 5 1 0 4 5 9 10 14 10 6 2
0 1 10 6 2 1 4 14 10 11 7 11

Trace Back (Local Alignment)

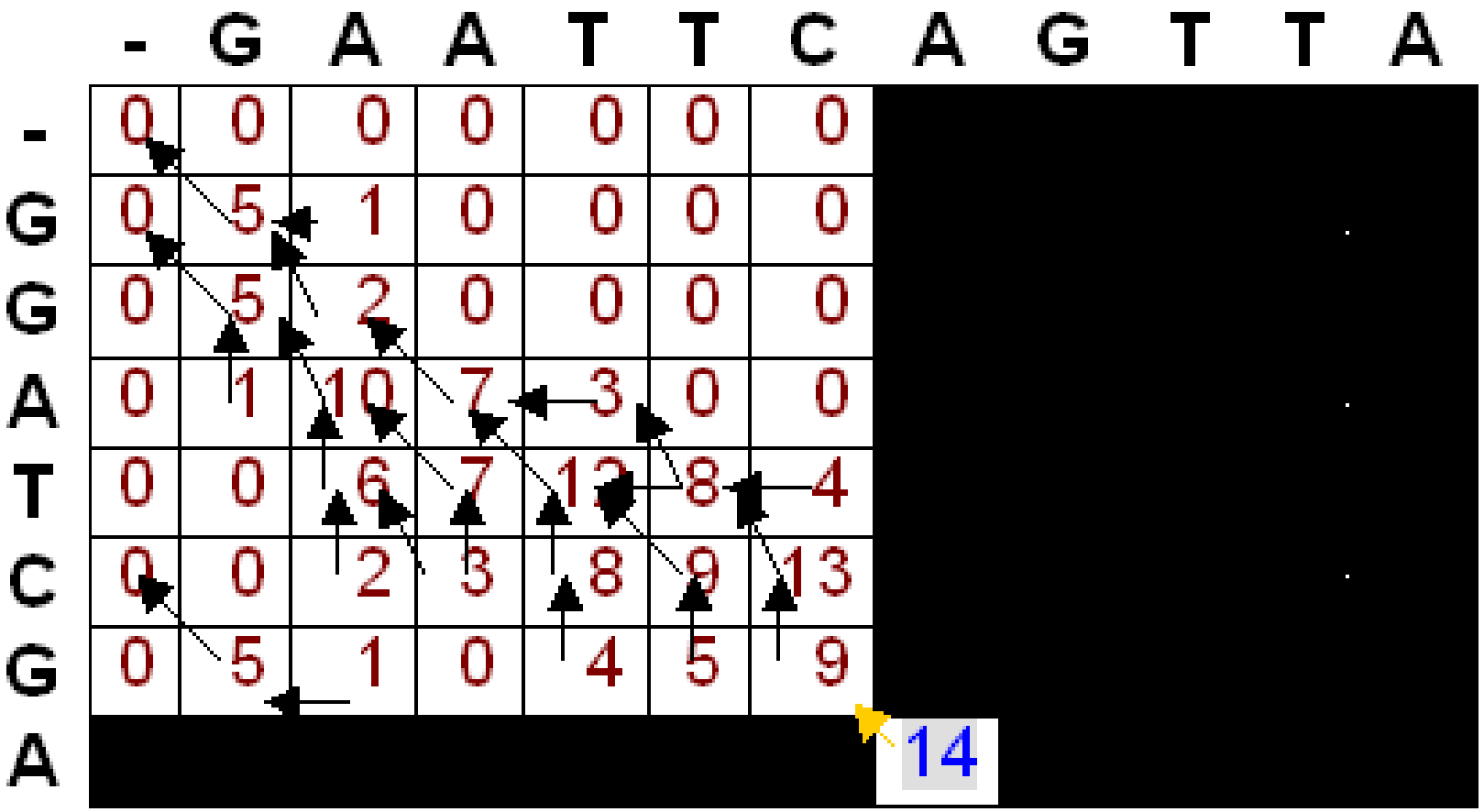
- Maximum local alignment score is the highest score anywhere in the matrix (14 in this example)

- 14 is found in two separate cells, indicating two possible multiple alignments producing the maximal local alignment score

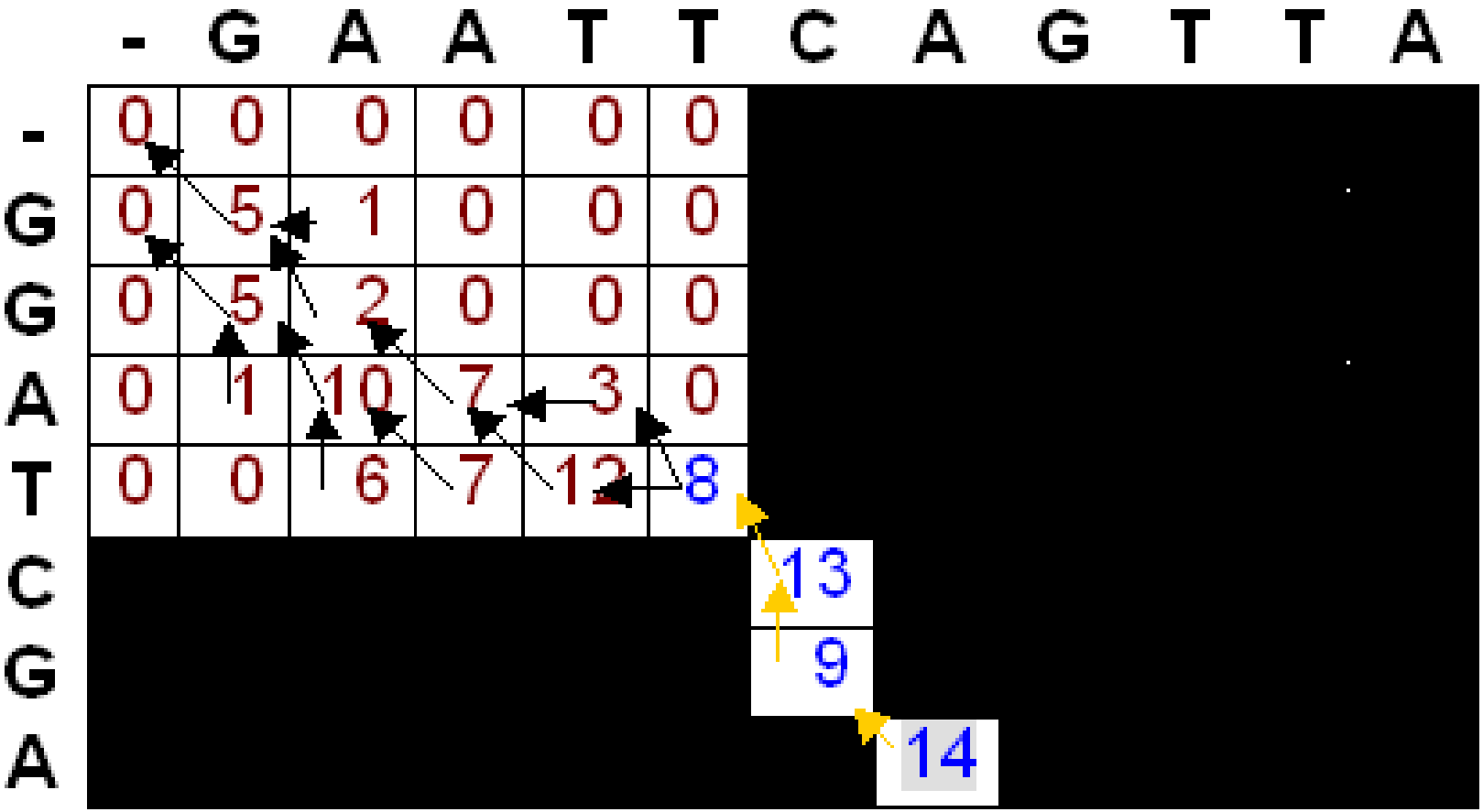
Trace Back (Local Alignment)

- Traceback begins in the position with the highest value.
- At each cell, we look to see where we move next according to the pointers
- When a cell is reached where there is not a pointer to a previous cell, we have reached the beginning of the alignment

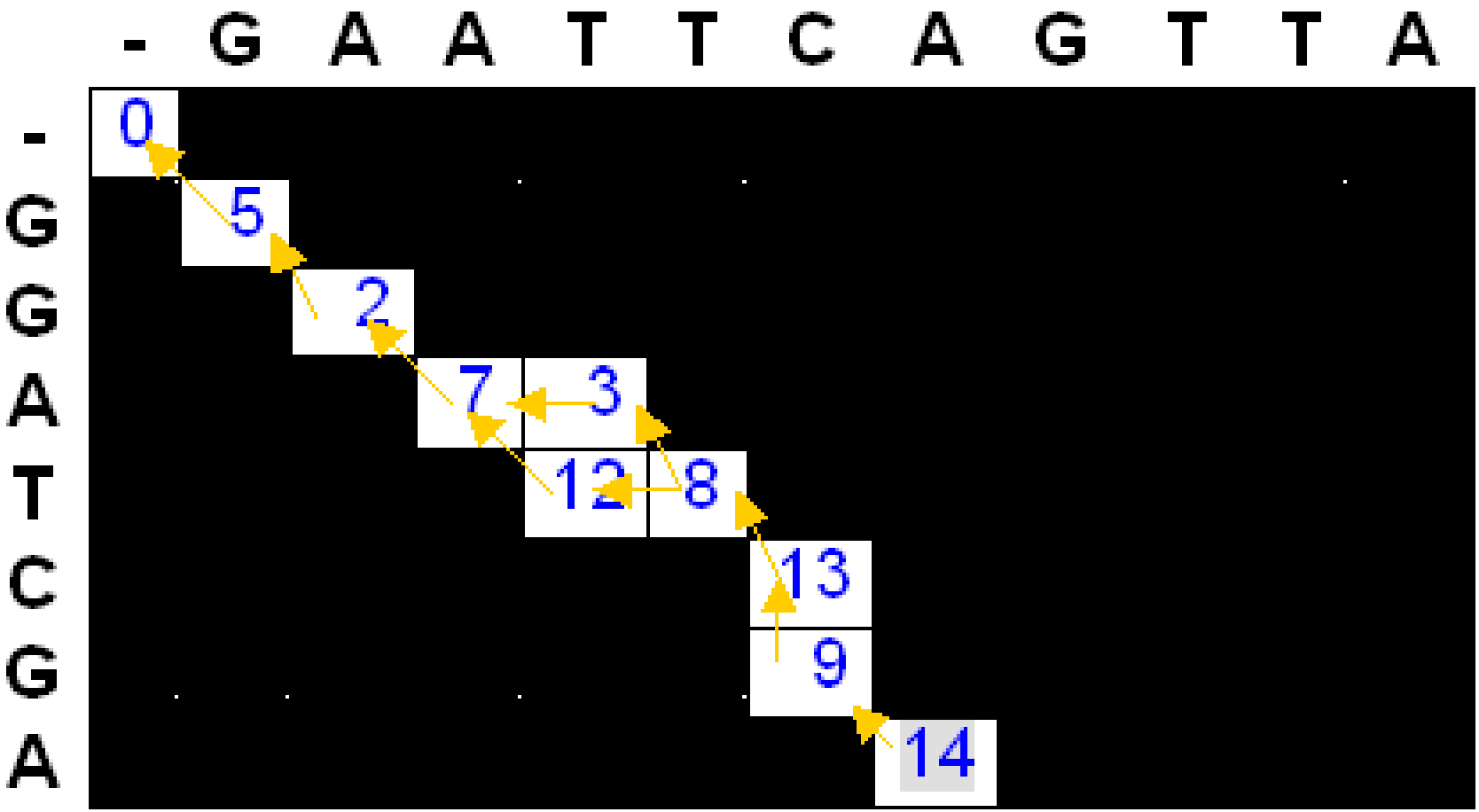
Trace Back Demo



Trace Back Demo



Trace Back Demo



Maximum Local Alignment

```
G A A T T C - A
|   | |   |   |
G G A T - C G A

+ - + + - + - +
5 3 5 5 4 5 4 5
=14
```

```
G A A T T C - A
|   |   | |   |
G G A - T C G A

+ - + - + + - +
5 3 5 4 5 5 4 5
=14
```

Linear vs. Affine Gaps

- So far, gaps have been modeled as linear
- More likely contiguous block of residues inserted or deleted
  - 1 gap of length k rather than k gaps of length 1
- Can create scoring scheme to penalize big gaps relatively less
  - Biggest cost is to open new gap, but extending is not so costly

Affine Gap Penalty

$$w_x = g + r(x-1)$$

- w_x : total gap penalty
- g: gap open penalty
- r: gap extend penalty
- x: gap length

- gap penalty chosen relative to score matrix

Scoring Alignments

- Pick a scoring matrix
  - BLOSUM62
  - PAM250
  - Match=5, mismatch=-4
- Decide on gap penalties
  - -gap opening penalty (-8)
  - -gap extension penalty (-1)
- Assume every position is independent
- Sum scores at each position
  - [log(x*y)=logx+logy]

Scoring Matrices

$$S_{ij} = \frac{\log(\frac{q_{ij}}{p_i p_j})}{\lambda}$$

- An empirical model of evolution, biology and chemistry all wrapped up in a 20 X 20 (or 4 X 4) table of numbers
- Structurally or chemically similar residues should ideally have high diagonal or off-diagonal numbers
- Structurally or chemically dissimilar residues should ideally have low diagonal or off-diagonal numbers
- What does the score mean: The likelihood of seeing two residues align (preserved) than random expected.

Scoring Alignments

Blosum62 Scoring Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1

BLOSUM substitution matrices

Developed for distantly related proteins

Substitutions only from multiple alignments of conserved regions of protein families, hand curated, constitute the known homologous blocks

Identity threshold to define conserved blocks can be varied, e.g. 62% idenitity gives BLOSUM62

Scores calculated from frequency of amino acids in aligned pairs compared to what would be expected due to abundance alone, given all sequences

Henikoff and Henikoff (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915-19

Blosum Matricies

What score should we give to a ser residue aligned with a thr residue?

$$\text{score}(\text{S}:\text{T}) \propto \log_2 \frac{P(\text{S}:\text{T}\,|\,\text{homology})}{P(\text{S}:\text{T}\,|\,\text{random})}$$

example of deriving Blosum scores for S:S, S:T, and T:T

Database of known alignments

SDHIP	HK**S**A	WMFE**T**	R**T**QC
SDHLP	HR**T**A	WMFD**T**	R**T**NC
SDHIP	HK**S**G	WLFD**T**	K**T**QC
SEHLP			K**S**QC
SEHLP			K**T**QC

Homology Model (consider each pair of sequences separately)

S:S pairs in alignments = 11
S:T pairs in alignments = 6
T:T pairs in alignments = 9

P(S:S|homology) = 11/117 =.094
P(S:T|homology) = 6/117 =.051
P(T:T|homology) = 9/117 =.078

Total pairs in alignments = 117

example of deriving Blosum scores for S:S, S:T, and T:T

Database of known alignments

SDHIP	HK**S**A	WMFE**T**	R**T**QC
SDHLP	HR**T**A	WMFD**T**	R**T**NC
SDHIP	HK**S**G	WLFD**T**	K**T**QC
SEHLP		K**S**QC	
SEHLP		K**T**QC	

Random Model

Number of S residues = 8
Number of T residues = 8
Total residues = 72

$P(S:S|random)=P(S)P(S)=(8/72)^2=.012$

$P(S:T|random)=2*P(S)P(T)=2*(8/72)^2=.024$

$P(T:T|random)=P(T)P(T)=(8/72)^2=.012$

example of deriving Blosum scores for S:S, S:T, and T:T

$$\text{score}(\text{S}:\text{S}) = \log_2 \frac{\text{P}(\text{S}:\text{S}\,|\,\text{homology})}{\text{P}(\text{S}:\text{S}\,|\,\text{random})} = \log_2 \frac{.094}{.012} = 2.96$$

$$\text{score}(\text{S}:\text{T}) = \log_2 \frac{\text{P}(\text{S}:\text{T}\,|\,\text{homology})}{\text{P}(\text{S}:\text{T}\,|\,\text{random})} = \log_2 \frac{.051}{.024} = 1.09$$

$$\text{score}(\text{T}:\text{T}) = \log_2 \frac{\text{P}(\text{T}:\text{T}\,|\,\text{homology})}{\text{P}(\text{T}:\text{T}\,|\,\text{random})} = \log_2 \frac{.078}{.012} = 2.70$$

BLOSUM and PAM

BLOSUM 45	BLOSUM 62	BLOSUM 90
PAM 250	PAM 160	PAM 100
More Divergent		*Less Divergent*

- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.
- PAM matrices: point accepted mutation

Scoring Matrices Take Home Points

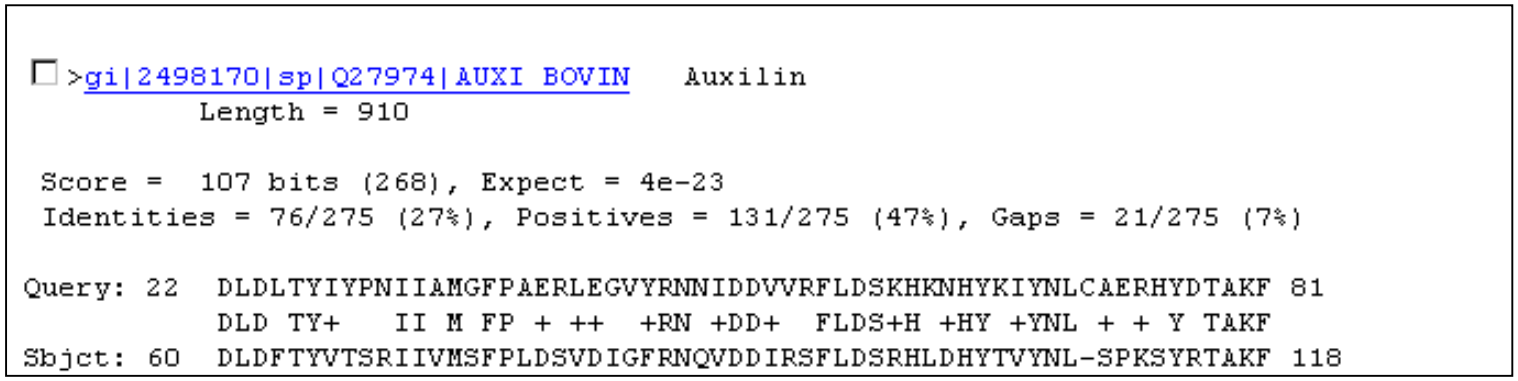
- Based on log odds scores
  - Ratios>1 give positive scores, ratios<1 give negative scores
  - Because log(x*y)=logx+logy the score of an alignment is the sum of the scores for each pair of aligned residues
- Assume independence of adjacent residues when scoring
- Introduced the concept that the frequency of a residue in a multiple alignment is informative

Fast Similar Sequence Search

- Can we run Smith-Waterman between query and every DB sequence?
- Yes, but too slow!
- General approach
  - Break query and DB sequence to match subsequences
  - Extend the matched subsequences, filter hopeless sequences
  - Use dynamic programming to get optimal alignment

BLAST

- Basic Local Alignment Search Tool
- Altschul et al. *J Mol Biol.* 1990
- One of the most widely used bioinformatics applications
  - Alignment quality not as good as Smith-Waterman
  - But much faster, supported at NCBI with big computer cluster
- For tutorials or information: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html



```
>gi|2498170|sp|Q27974|AUXI_BOVIN    Auxilin
          Length = 910

 Score =  107 bits (268), Expect = 4e-23
 Identities = 76/275 (27%), Positives = 131/275 (47%), Gaps = 21/275 (7%)

Query: 22  DLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKF 81
           DLD TY+   II M FP + ++  +RN +DD+  FLDS+H +HY +YNL + + Y TAKF
Sbjct: 60  DLDFTYVTSRIIVMSFPLDSVDIGFRNQVDDIRSFLDSRHLDHYTVYNL-SPKSYRTAKF 118
```

BLAST Algorithm Steps

- Query and DB sequences are optionally filtered to remove low-complexity regions
  - E.g. ACACACACA, TTTTTTTTT

BLAST Algorithm Steps

- Query and DB sequences are optionally filtered to remove low-complexity regions
- Break DB sequences into k-mer words and hash their locations to speed later searches
  - k is usually 11 for DNA/RNA and 3 for protein

LPPQGLL
LPP
PPQ
PQG
QGL
GLL

BLAST Algorithm Steps

- Query and DB sequences are optionally filtered to remove low-complexity regions
- Break DB sequences into k-mer words and hash their locations to speed later searches
- Each k-mer in query find possible k-mers that matches well with it
  - “well” is evaluated by substitution matrices

BLAST Algorithm Steps

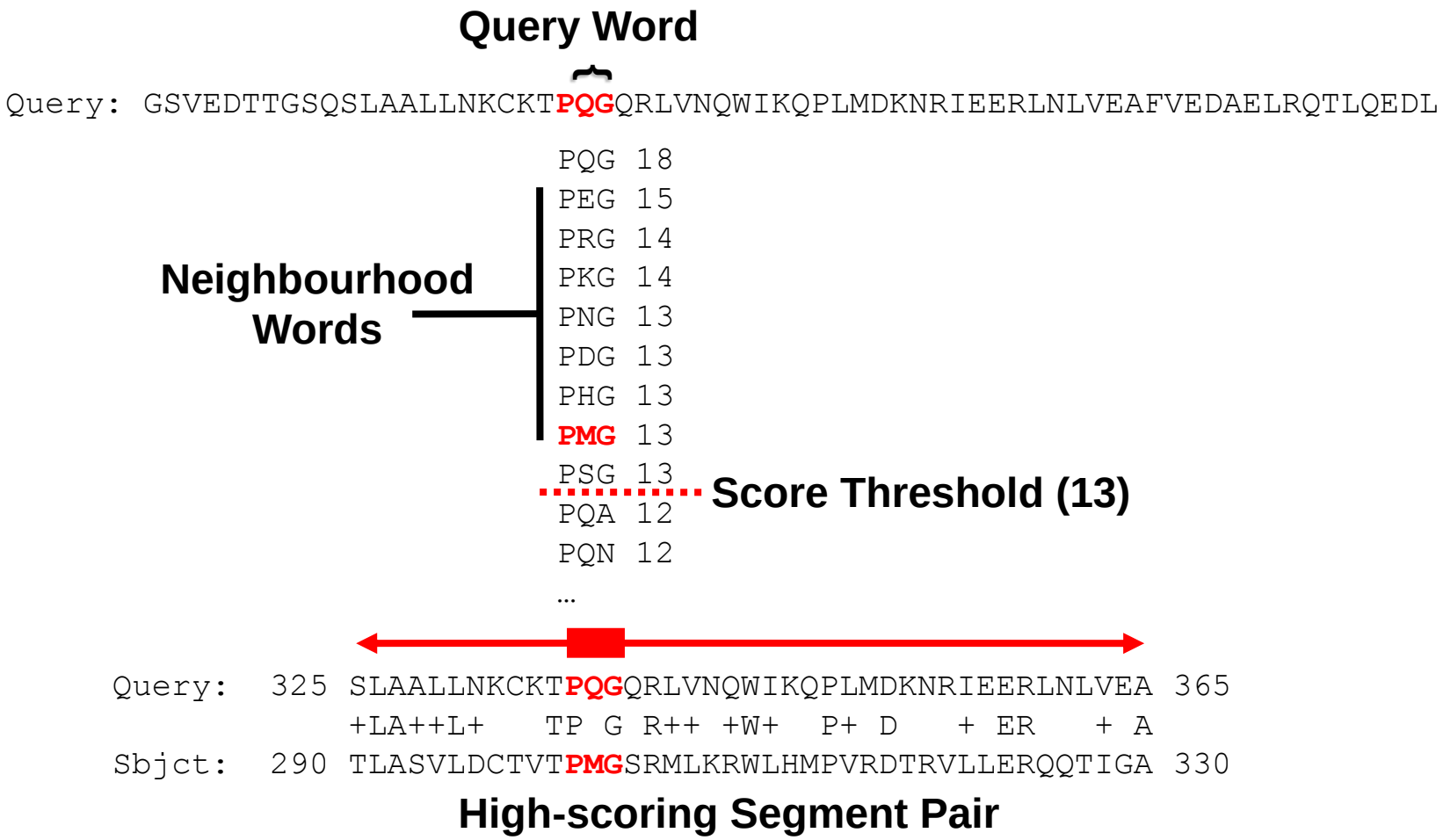
- Only words with ≥ T cutoff score is kept
  - T is usually 11-13, ~ 50 words make T cutoff
  - Note: this is 50 words at every query position
- For each DB sequence with a high scoring word, try to extend it in both ends

  Query: **LP PQG LL**

  DB seq: **MP PEG LL**

  HSP score 9 + 15 + 8 = 32
  - Form HSP (High-scoring Segment Pairs)
  - Use BLOSUM to score the extended alignment
  - No gaps allowed

The BLAST Search Algorithm

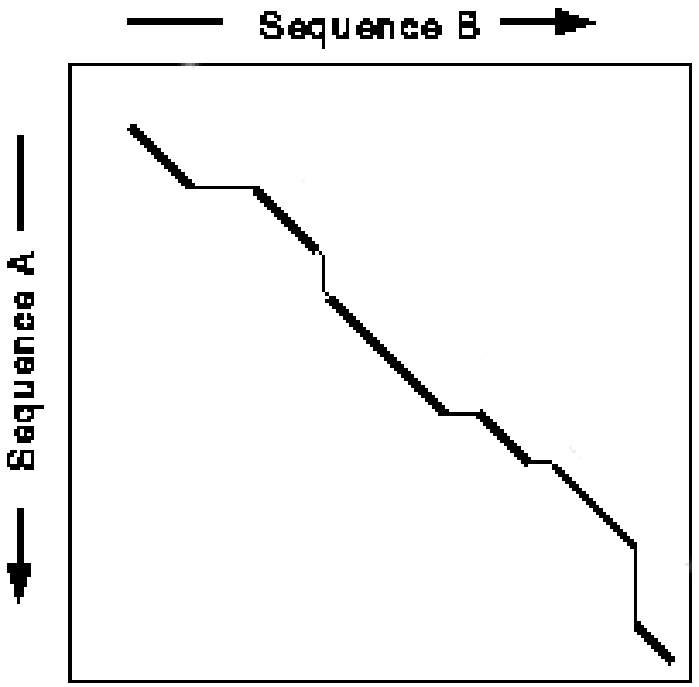


BLAST Algorithm Steps

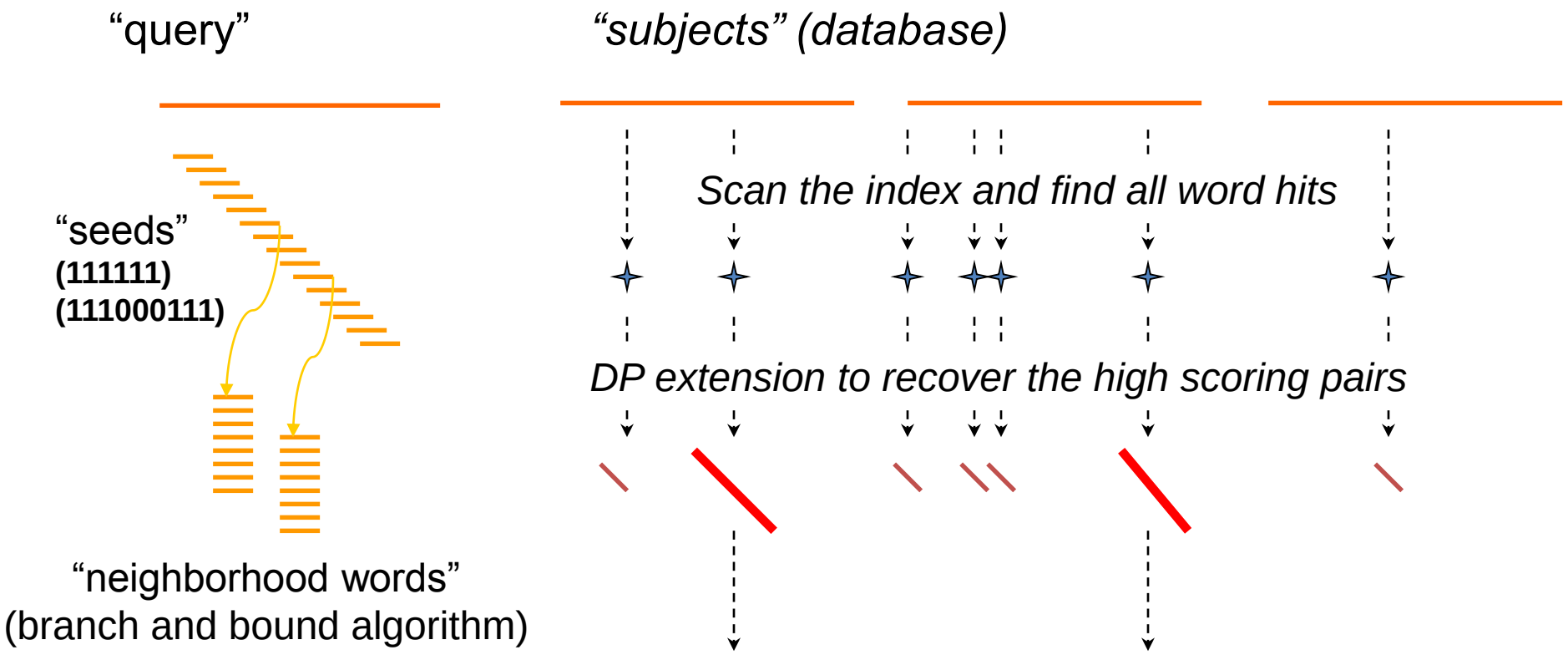
- Keep only statistically significant HSPs
  - Based on the scores of aligning 2 random seqs
- Use Smith-Waterman algorithm to join the HSPs and get optimal

  alignment
  - Gaps are allowed

    default (-11, -1)



BLAST algorithm summary



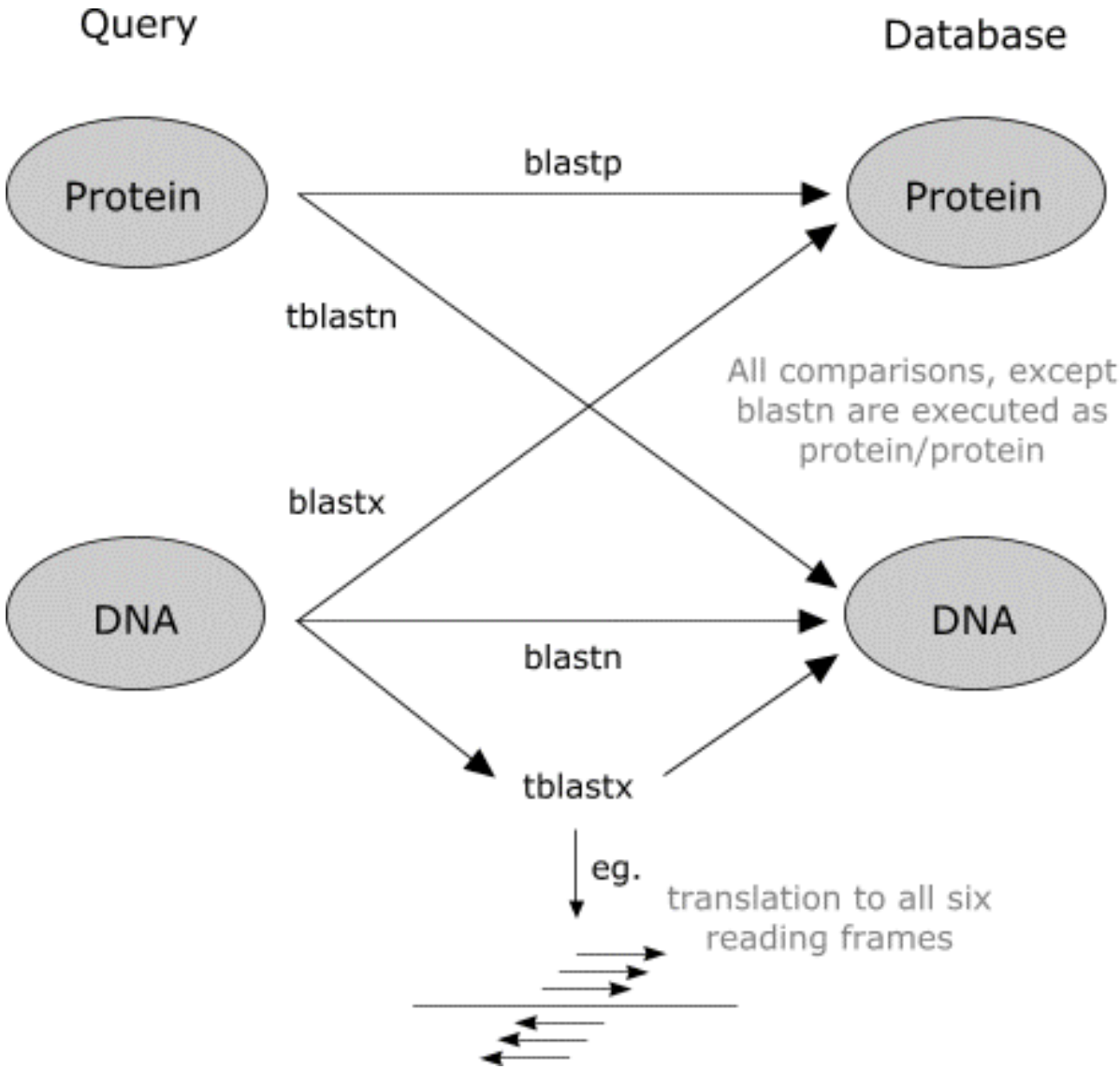
Different BLAST Programs

BLAST DB:

- nr (non-redundant):
  - GenBank, RefSeq, EMBL…
- est:
  - expressed sequences (cDNA), redundant
- Swissprot and pdb:
  - protein databases

If query is DNA, but known to be coding (e.g. cDNA)

- Translate cDNA into protein
- Zero gap-extension penalty

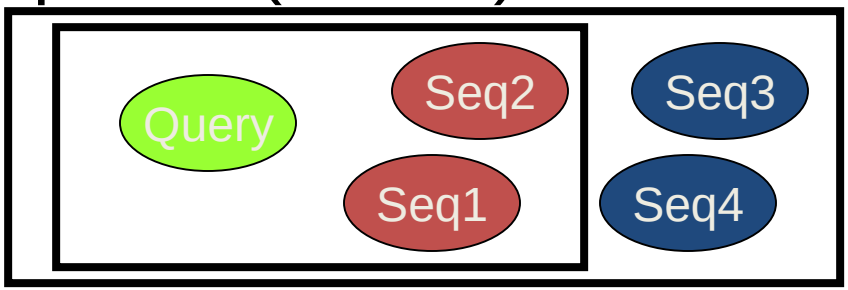


Different BLAST Programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.

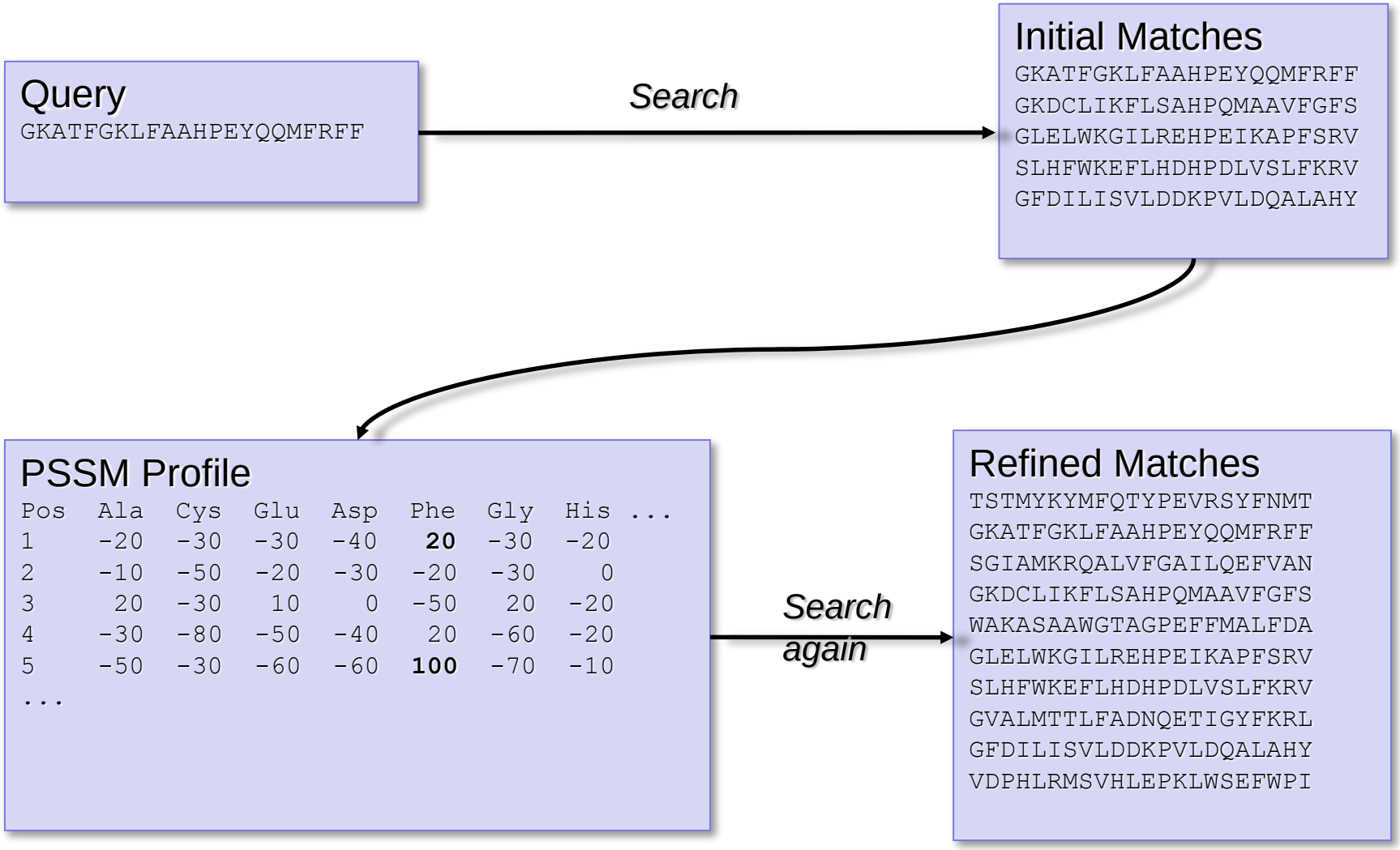
PSI-BLAST

- Position Specific Iterative BLAST
  - Align high scoring hits in initial BLAST to construct a profile for the hits
  - Use profile (PSSM) for next iteration BLAST



- Find remote homologs or protein families
- FP sequences can degrade search quickly

PSI-BLAST



http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

http://www.ebi.ac.uk/blastpgp/

Reciprocal Blast

- Search for orthologous sequences between two species
  - GeneA in Species1 BLAST Species2 ➔ GeneB
  - GeneB in Species2 BLAST Species1 ➔ GeneA
  - GeneA 

GeneB
- Also called bi-directional best hit

BLAT

- BLAST-Like Alignment Tool
  - Compare to BLAST, BLAT can align much longer regions (MB) really fast with little resources
  - E.g. can map a sequence to the genome in seconds on one Linux computer
  - Allow big gaps (mRNA to genome)
  - Need higher similarity (> 95% for DNA and 80% for proteins) for aligned sequences
- Basic approach
  - Break long sequence into blocks
  - Index k-mers, typically 8-13
  - Stitch blocks together for final alignment

BLAT: Indexing

Genome: cacaattatcacgaccgc

3-mers: cac aat tat cac gac cgc

Index:
```
aat 3      gac 12
cac 0,9    tat 6
cgc 15
```

cDNA (mRNA -> DNA): aattctcac

3-mers: aat att ttc tct ctc tca cac

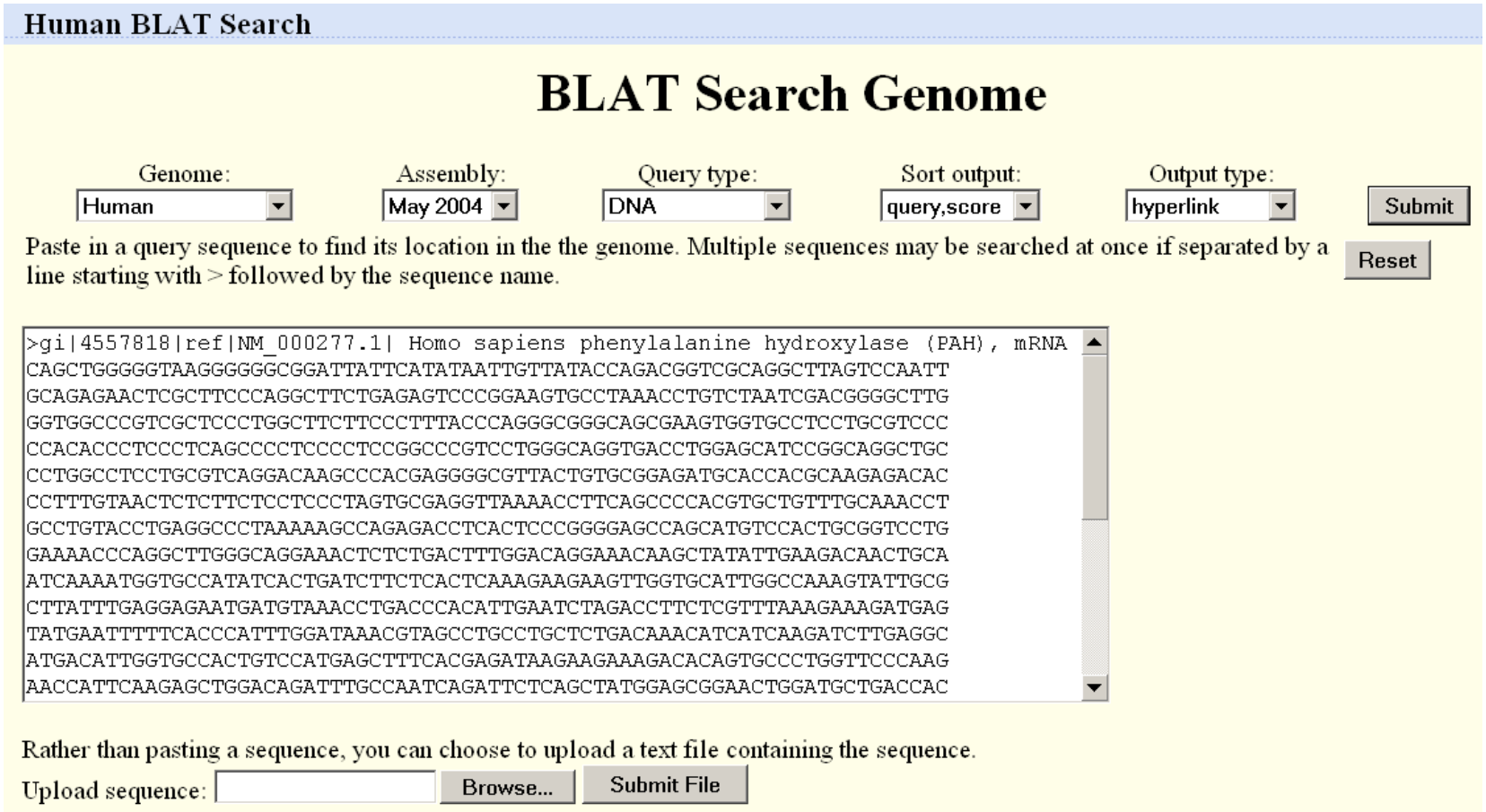
```
 0   1   2   3   4   5   6
```

hits:
```
aat 0,3  -3
cac 6,0   6
cac 6,9  -3
```

clump:
```
cacAATtatCACgaccgc
   |||   |||
   aattctcac
```

BLAT Example

- Enter sequence and parameters



BLAT Example

- Get result instantly!!

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	NM_000277.1	1734	1	1750	1750	99.9%	12	-	101736682	101813848	77167
browser details	NM_000277.1	634	1	641	1750	99.7%	12_random	-	281957	286767	4811
browser details	NM_000277.1	186	639	824	1750	100.0%	12_random	+	116309	116494	186
browser details	NM_000277.1	32	1456	1567	1750	97.1%	1	-	21612213	21612486	274
browser details	NM_000277.1	24	210	233	1750	100.0%	22	-	20374187	20374210	24
browser details	NM_000277.1	24	1377	1407	1750	88.5%	9	+	30738626	30738655	30
browser details	NM_000277.1	24	753	789	1750	96.2%	16	+	3484390	3484428	39
browser details	NM_000277.1	22	1009	1030	1750	100.0%	7	-	30429927	30429948	22
browser details	NM_000277.1	22	827	860	1750	82.4%	11	-	2145950	2145983	34
browser details	NM_000277.1	22	208	229	1750	100.0%	2	+	95471146	95471167	22
browser details	NM_000277.1	22	1564	1585	1750	100.0%	16	+	56691362	56691383	22
browser details	NM_000277.1	22	153	175	1750	100.0%	1	+	37402859	37402882	24
browser details	NM_000277.1	21	1095	1115	1750	100.0%	X	+	5847999	5848019	21
browser details	NM_000277.1	17	253	283	1750	77.5%	17	+	61370186	61370216	31

Summary of Fast Search

- Fast sequence similarity search
  - Break seq, hash DB sub-seq, match sub-seq and extend, use DP for optimal alignment
  - *BLAST, most widely used, many applications with sound statistical foundations
  - *BLAT, align sequence to genome, fast yet need higher similarity

BLAST score and significance

- Report DB sequences above a threshold
  - E value: Number (instead of probability → pvalue) of matches expected merely by chance

$$E = Kmn\, e^{-\lambda S}$$

$$p(s \geq x) \approx 1 - \exp[-e^{-x}]$$

  - m, n are query and DB length
  - K, □ are constants
  - Smaller E, more stringent

Are these proteins homologs?

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P    W L   Y N    Y C     L
SEQ 2: QFFPLMPPAPYWILATDYENLPLVYSCTTFFWLF
```

Probably not (score = 9)

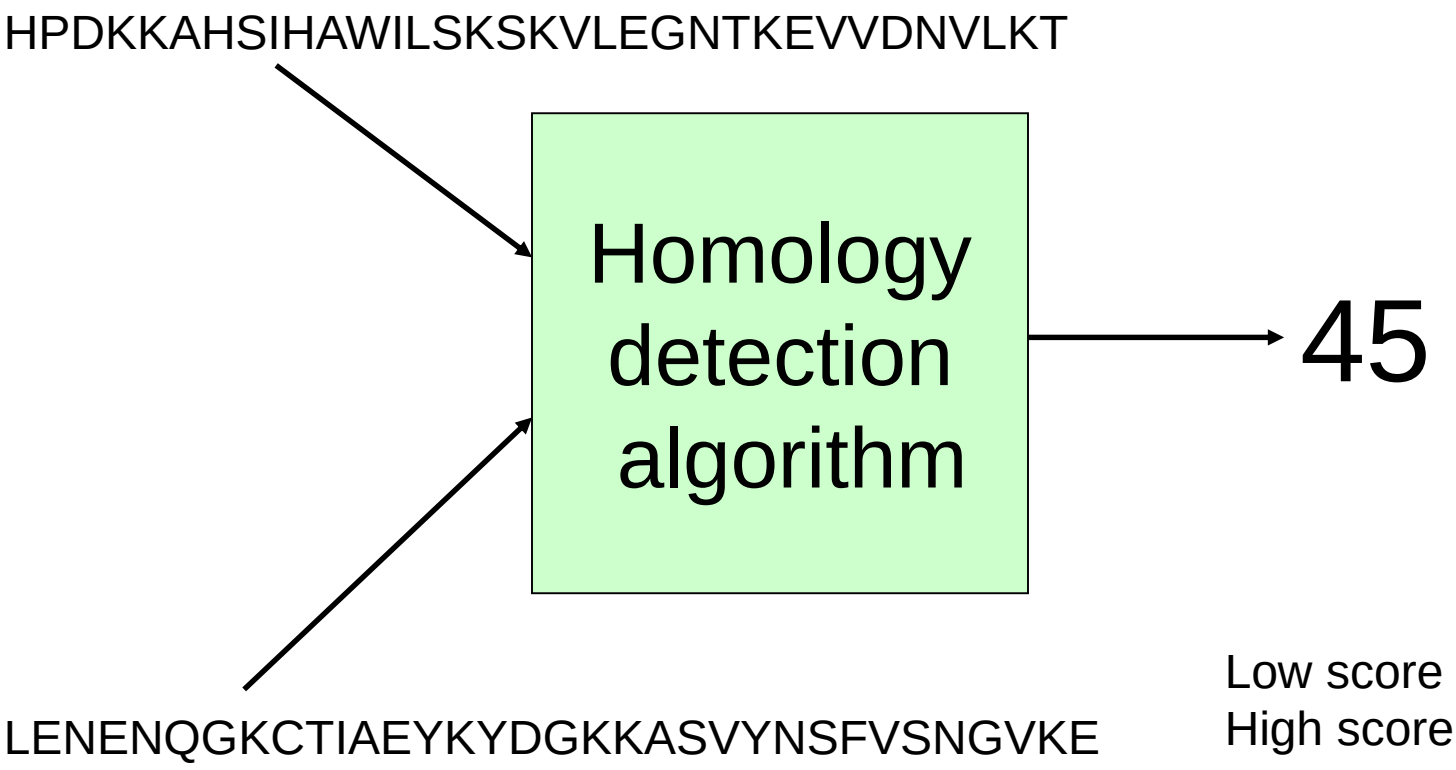
```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
           L P    W LDATYKNYA  Y C     L
SEQ 2: QFFPLMPPAPYWILDATYKNYALVYSCTTFFWLF
```

MAYBE (score = 15)

```
SEQ 1: RVVNLVPS--FWVLDATYKNYAINYNCDVTYKLY
       RVV L PS   W LDATYKNYA  Y CDVTYKL
SEQ 2: RVVPLMPSAPYWILDATYKNYALVYSCDVTYKLF
```

Most likely (score = 24)

Significance of scores



Low score = unrelated
High score = homologs

How high is high enough?

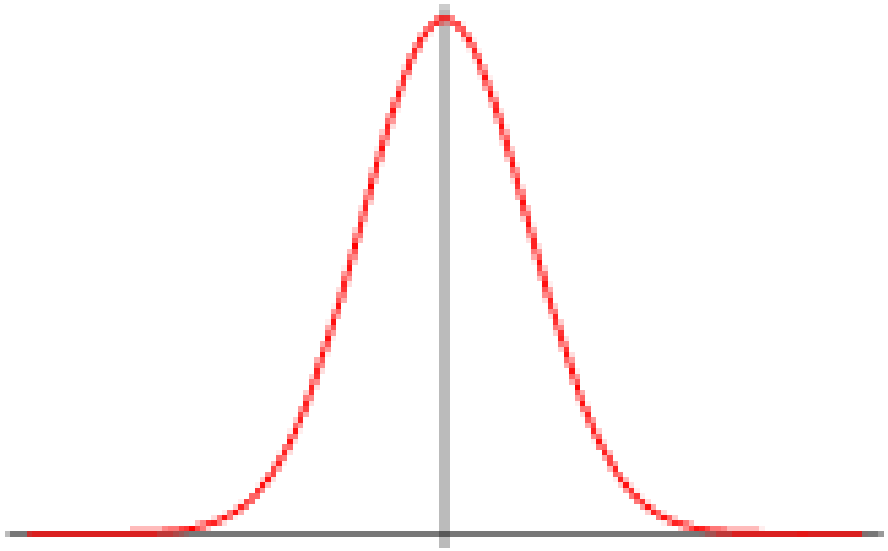
Other significance questions

- Pairwise sequence comparison scores
- Microarray expression measurements
- Sequence motif scores
- Functional assignments of genes
- Call peaks from ChIP-seq data

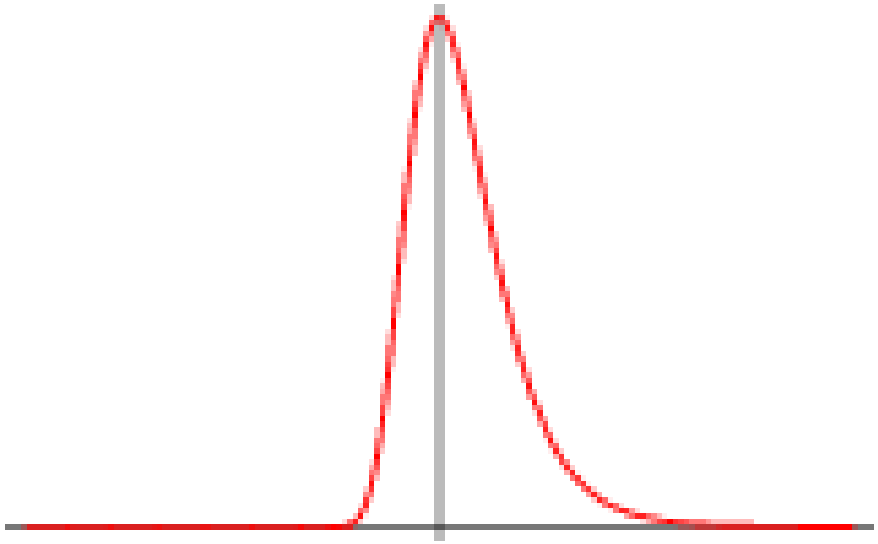
The null hypothesis

- We are interested in characterizing the distribution of scores from sequence comparison algorithms.
- We would like to measure how surprising a given score is, *assuming that the two sequences are not related.*
- The assumption is called the null hypothesis.
- The purpose of most statistical tests is to determine whether the observed results provide a reason to reject the hypothesis that they are merely a product of chance factors.

Gaussian vs. Extreme Value Distribution (EVD)

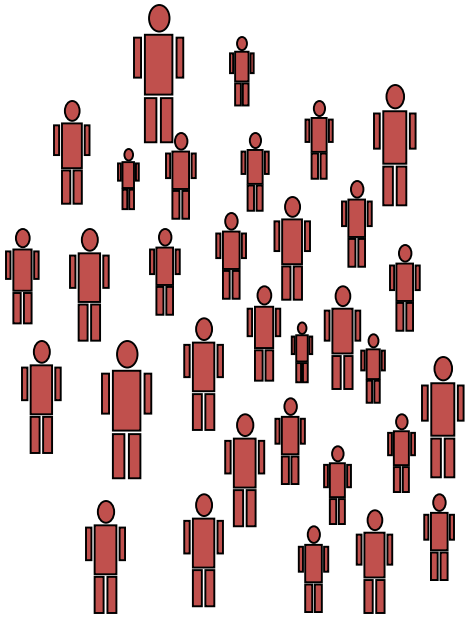


Gaussian (Normal)

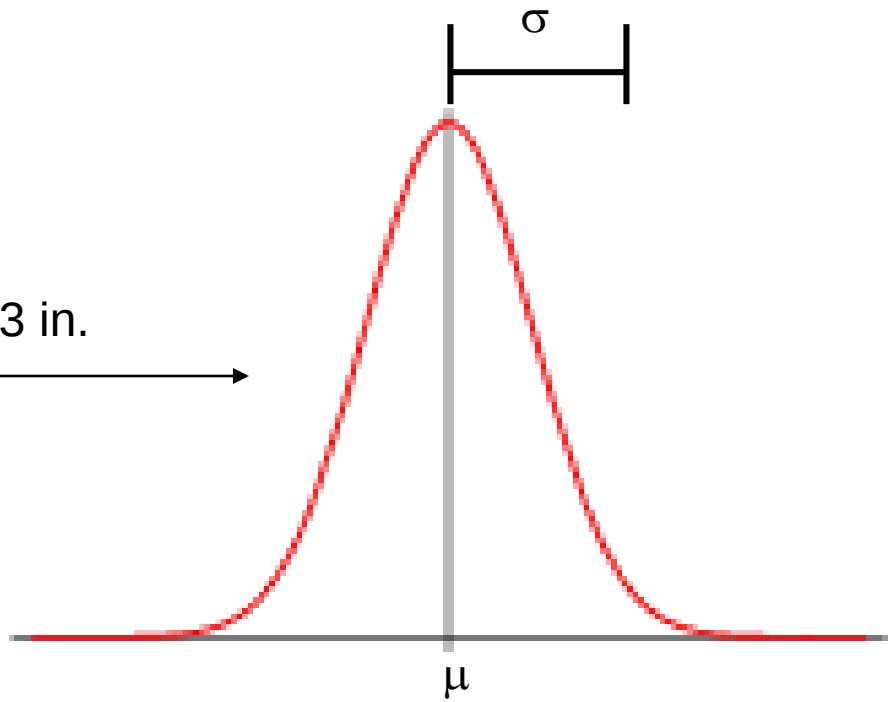


Extreme value

Gaussian



μ= 68 in.,σ=3 in.

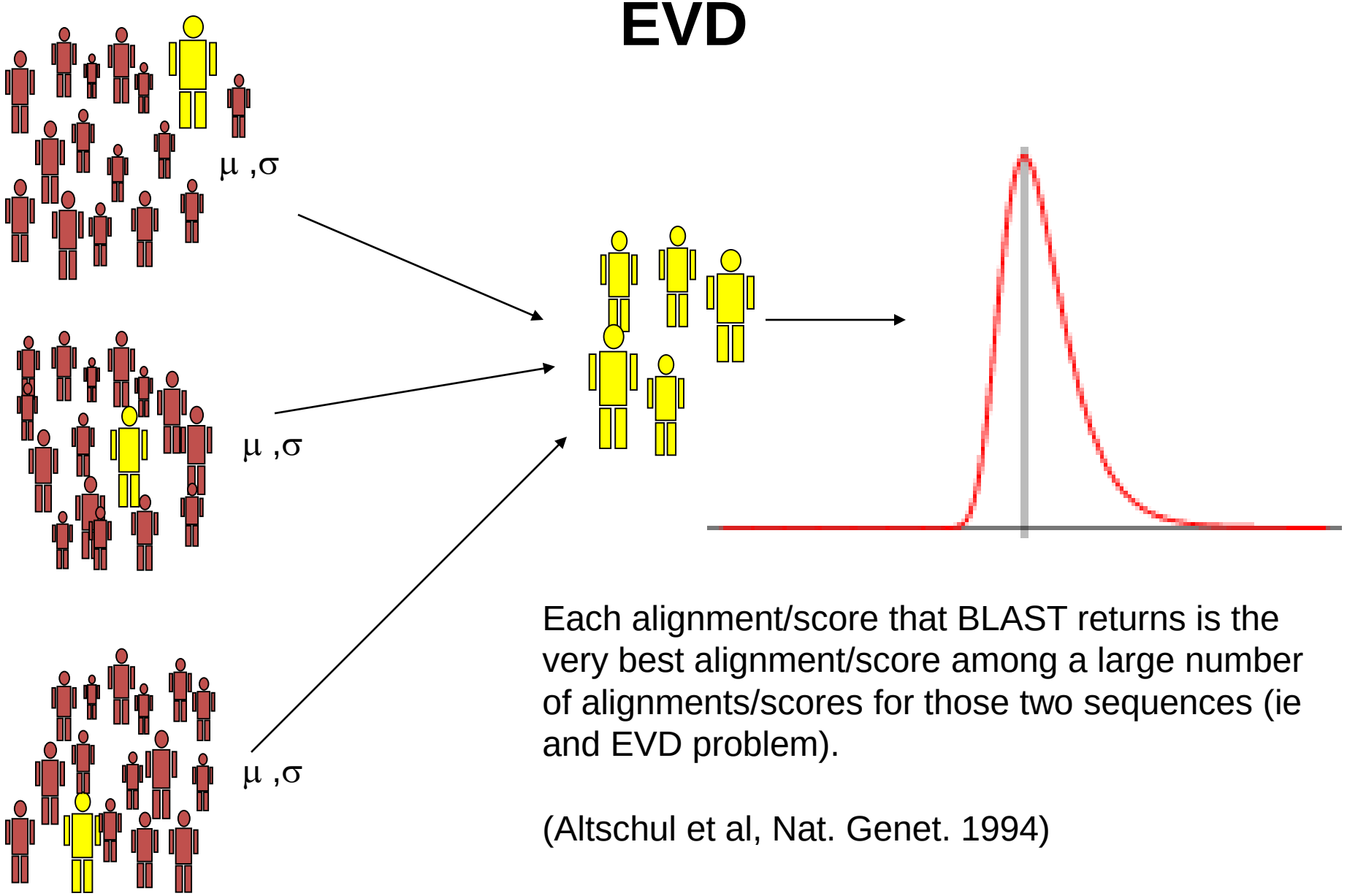


What is the chance of picking a person at least 75 in. tall P(X≥75)?

$$z_{score}(x) = \frac{x-\mu}{\sigma} = \frac{75-68}{3} = 2.33$$

From Table:
z=2.33 → *P*=0.01

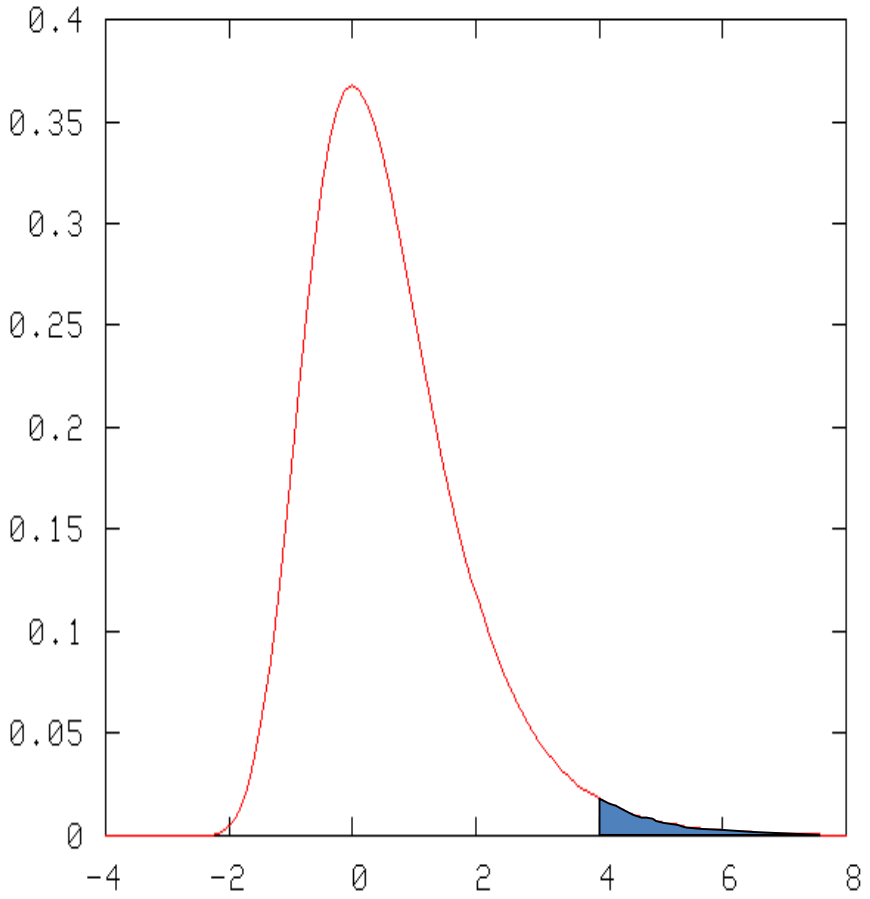
EVD



Each alignment/score that BLAST returns is the very best alignment/score among a large number of alignments/scores for those two sequences (ie and EVD problem).

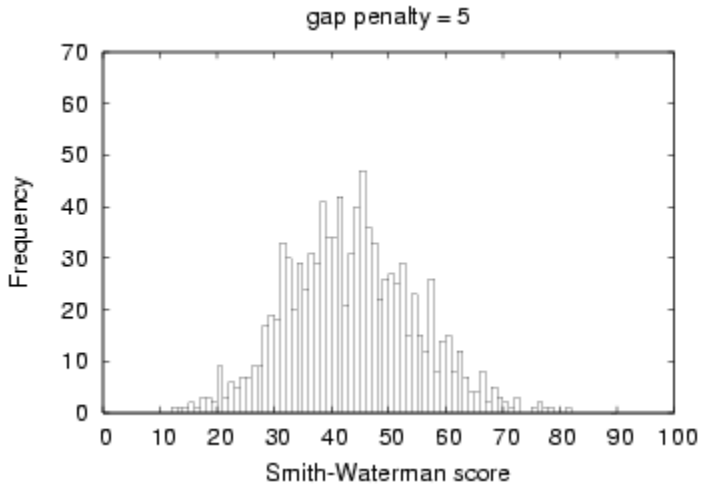
(Altschul et al, Nat. Genet. 1994)

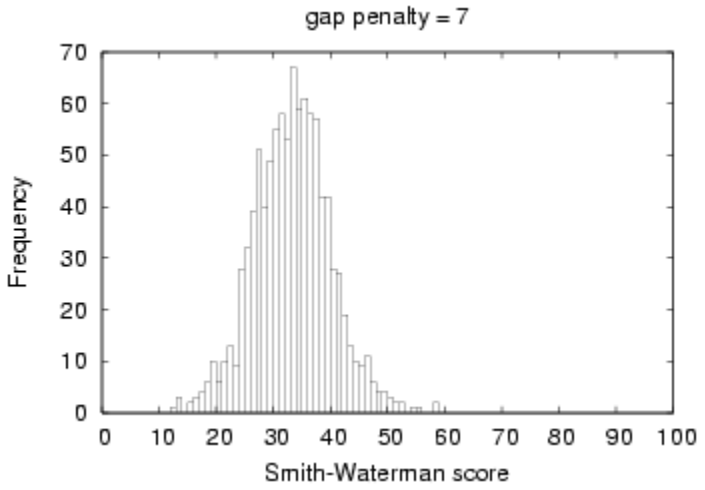
Computing a p-value



- The probability of observing a score >4 is the area under the curve to the right of 4.
- This probability is called a p-value.
- p-value = Pr(data|null)

Scaling the EVD





- An extreme value distribution derived from, e.g., the Smith-Waterman algorithm will have a characteristic mode μ and scale parameter λ.

$$P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\mu)}\right]$$

- These parameters depend upon the size of the query, the size of the target database, the substitution matrix and the gap penalties.

An example

You run BLAST and get a score of 45. You then run BLAST on a shuffled version of the database, and fit an extreme value distribution to the resulting empirical distribution. The parameters of the EVD are μ = 25 and λ = 0.693. What is the p-value associated with 45?

$$P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\mu)}\right]$$

$$P(S \geq 45) = 1 - \exp\left[-e^{-0.693(45-25)}\right]$$

$$= 1 - \exp\left[-e^{-13.86}\right]$$

$$= 1 - \exp\left[-9.565 \times 10^{-7}\right]$$

$$= 1 - 0.999999043$$

$$= 9.565 \times 10^{-7}$$

Summary of statistical significance

- A distribution plots the frequency of a given type of observation.
- The area under the distribution is 1.
- Most statistical tests compare observed data to the expected result according to the null hypothesis.
- Sequence similarity scores follow an extreme value distribution, which is characterized by a larger tail.
- The p-value associated with a score is the area under the curve to the right of that score.

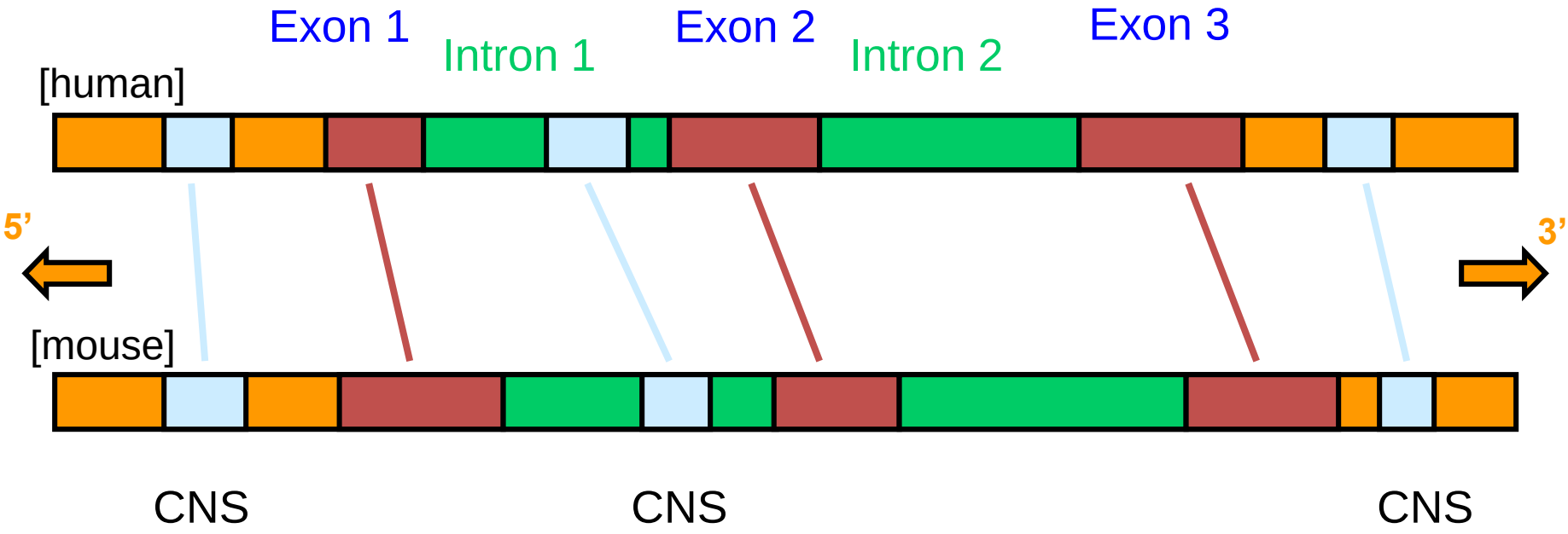
Applying homology: concept and technology

- Genome evolution
  - Mammalian genome evolution
  - Human genome variation
  - Cancer genome evolution
- Gene finding
  - Comparative approaches
  - Ab initio approaches
    - Hidden Markov Model
- Protein structure
  - Threading
- Regulatory motif finding
  - Profile comparison
- Pathway/Network comparison
  - PathBLAST
- Conservation
  - Ultra conserved elements
  - Human accelerated regions

Gene prediction

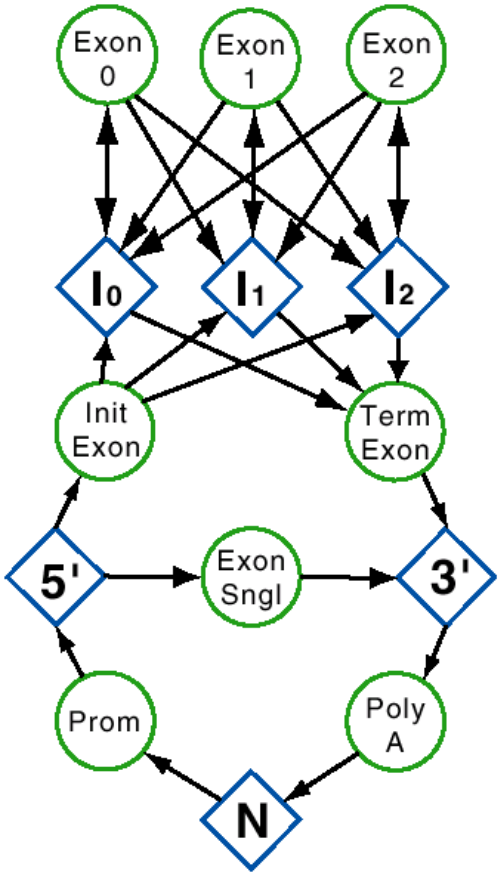
- Comparing to a known gene from a different species
- Using EST evidence (aligning transcript to genome)
- Predicting from sequence (HHM)
- Using conservation
  - Signature of coding potential
  - What about RNA gene?
- Using other genomics signals
  - Specific epigenetic marks of promoters and gene bodies

Modeling gene features



Genscan (Burge and Karlin, 1998)

- Dramatic improvement over previous methods
- Generalised HMM
- Different parameter sets for different GC content regions (intron length distribution and exon stats)



Predicting non-coding RNA?

- From sequence?
  - Not clear which properties can be exploited
  - Sequence features such as promoters are too weak
- Histone modifications + conservation worked

Vol 458|12 March 2009|doi:10.1038/nature07672 nature

LETTERS

Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals

Mitchell Guttman[1,2], Ido Amit[1], Manuel Garber[1], Courtney French[1], Michael F. Lin[1], David Feldser[3], Maite Huarte[1,6], Or Zuk[1], Bryce W. Carey[2,8], John P. Cassady[2,8], Moran N. Cabili[7], Rudolf Jaenisch[2,8], Tarjei S. Mikkelsen[1,4], Tyler Jacks[2,3], Nir Hacohen[1,9], Bradley E. Bernstein[1,10,11], Manolis Kellis[1,5], Aviv Regev[1,2], John L. Rinn[1,6,11]* & Eric S. Lander[1,2,7,8]*

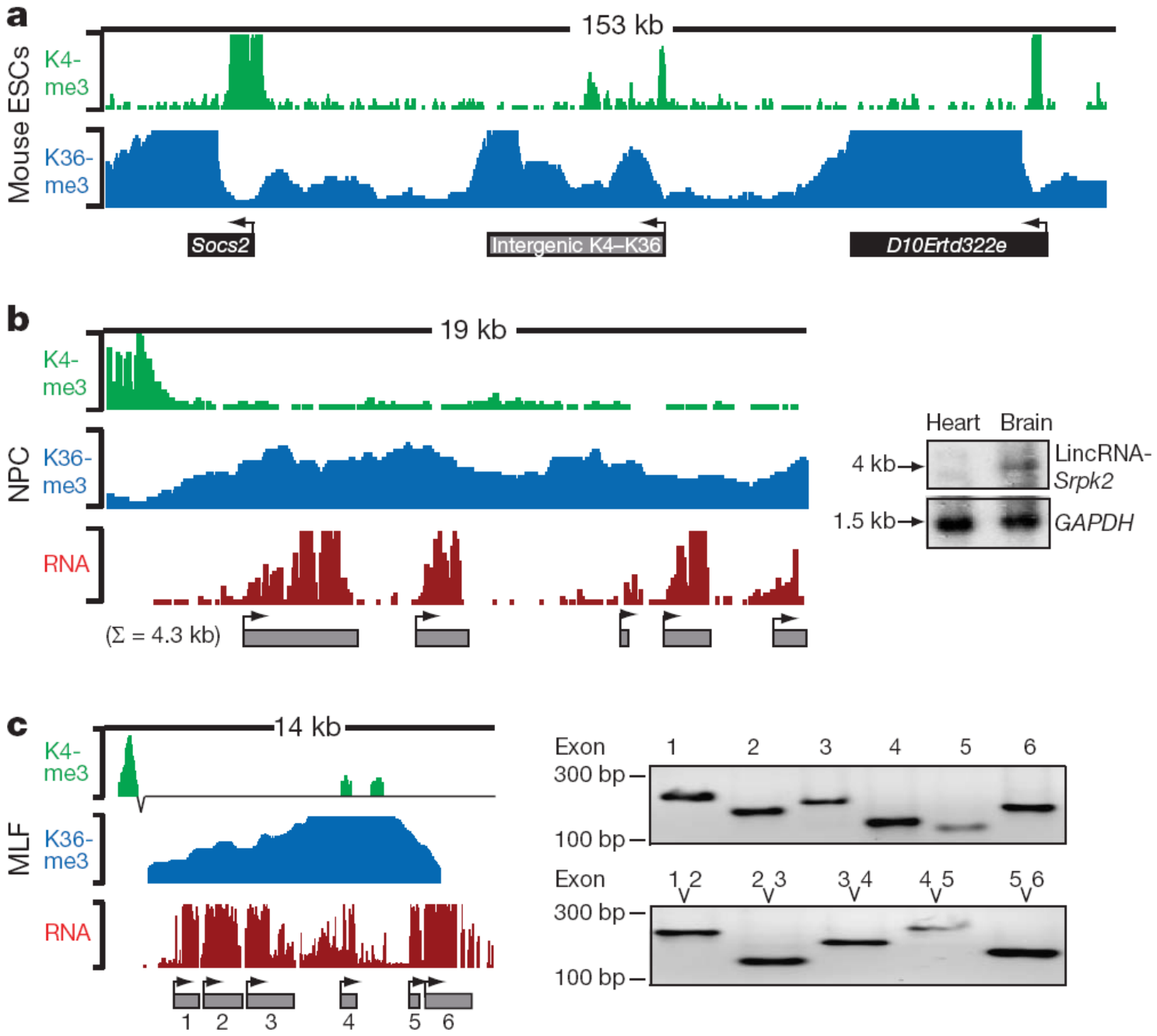
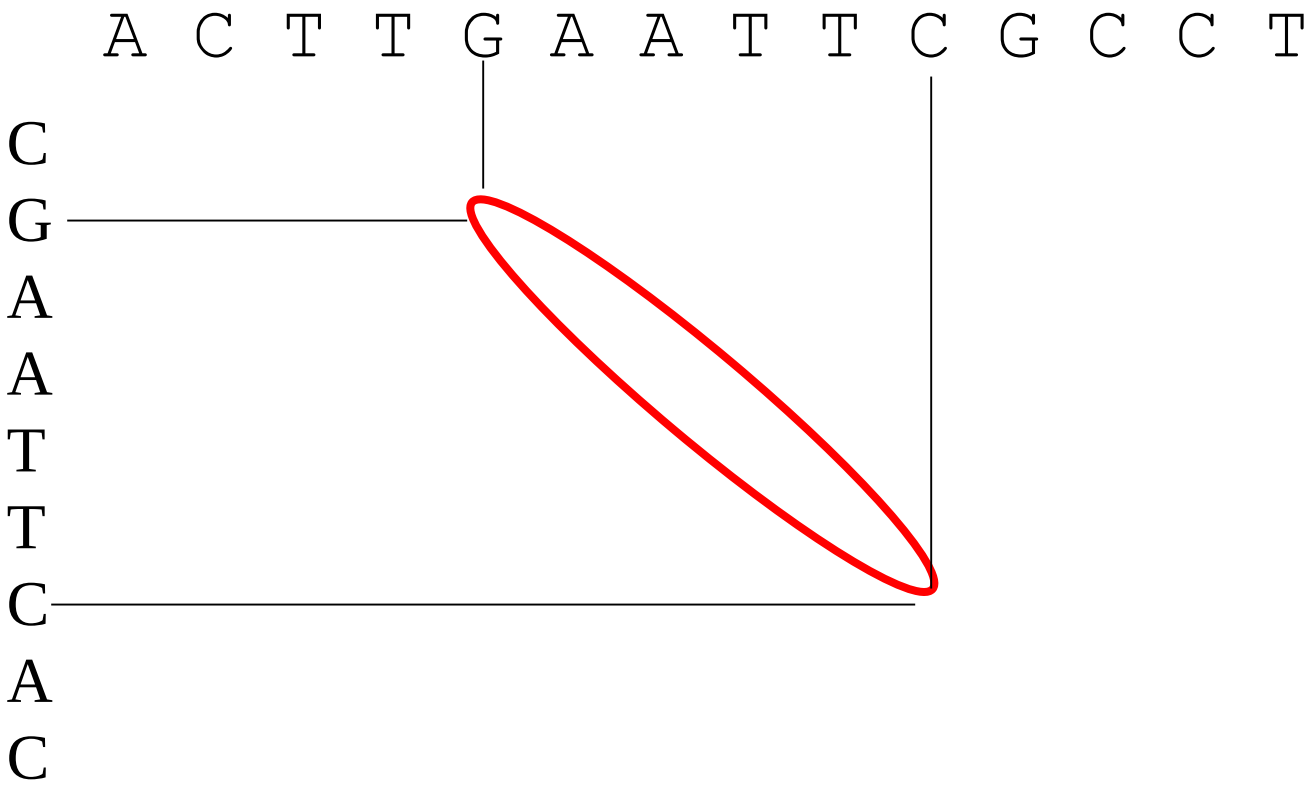
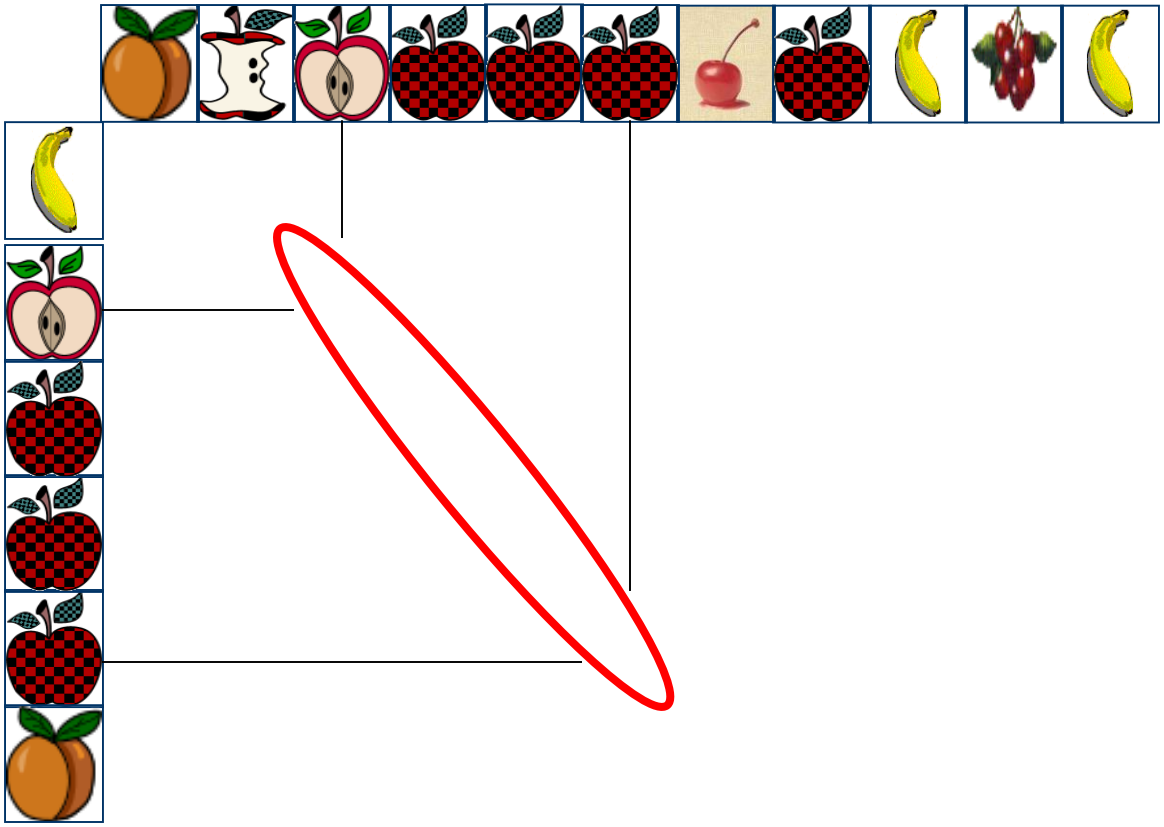


Figure 1 | Intergenic K4–K36 domains produce multi-exonic RNAs.

So far, only linear sequence comparison



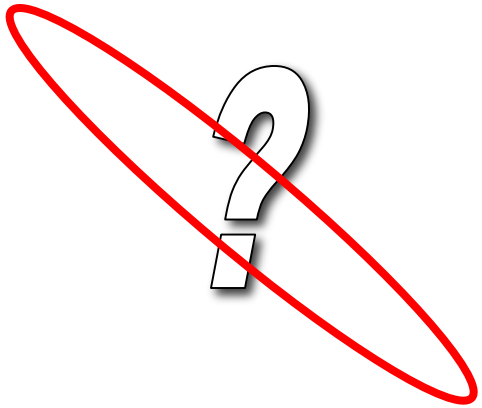
Expanding the idea of a sequence



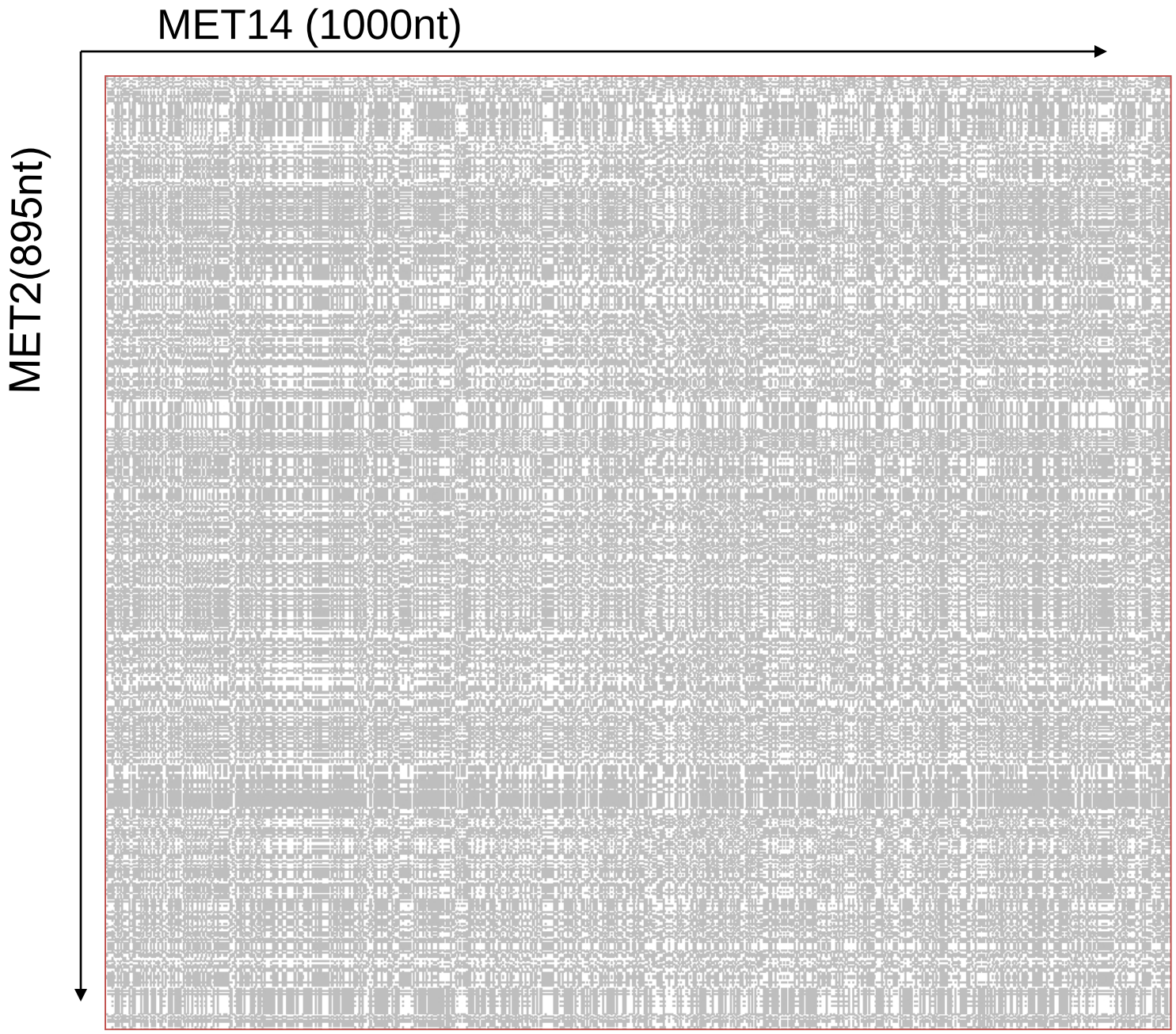
Central theme of the new algorithm – compare profiles

A	|	6	6	1	0	6	5	0	0
C	|	0	0	1	0	0	0	1	5
G	|	0	0	4	6	0	1	0	1
T	|	0	0	0	0	0	0	5	0

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



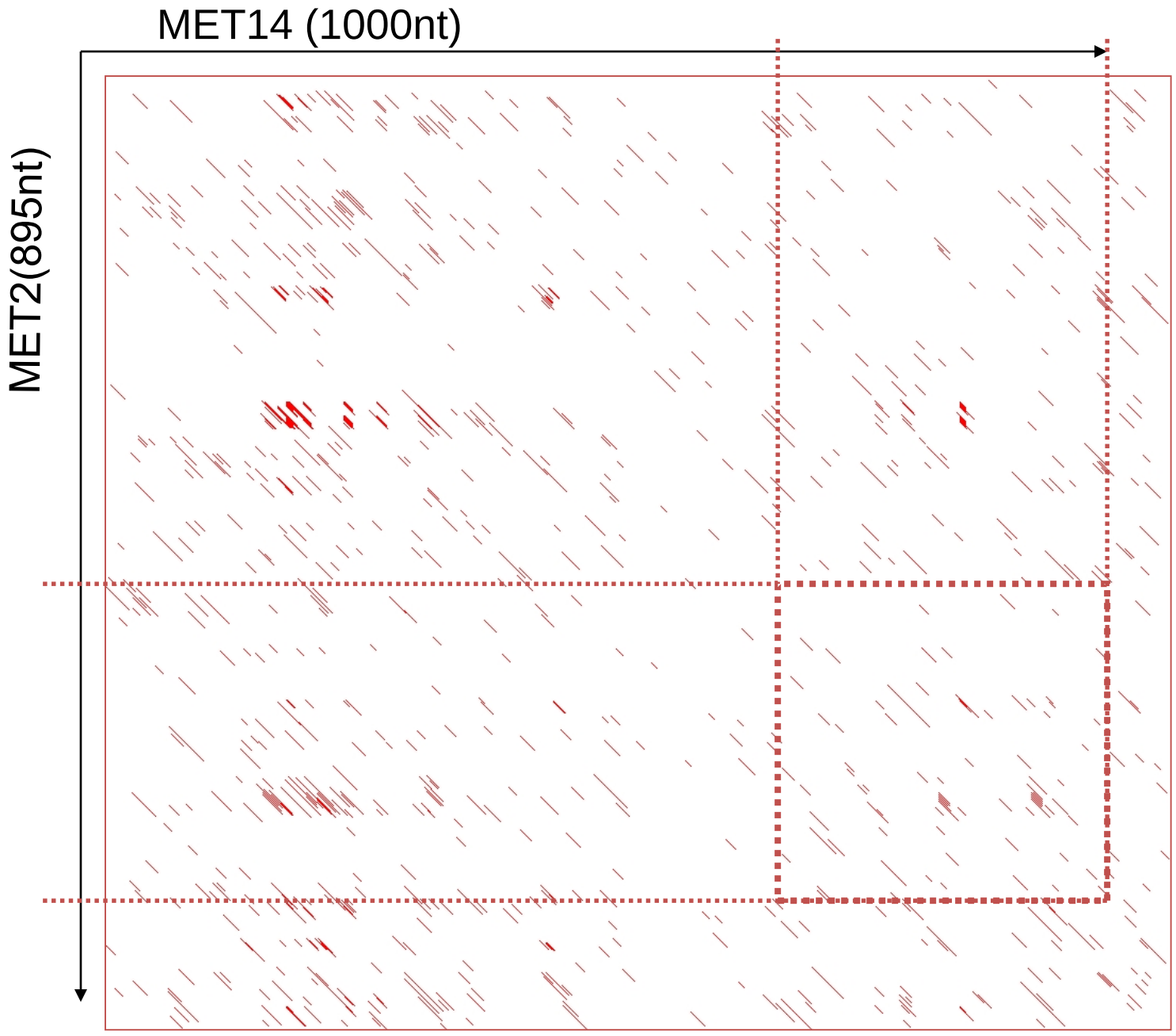
Met14 vs Met2 "DotPlot"



Match = 1
Mismatch =-1
Gray: 1

Met14 vs Met2

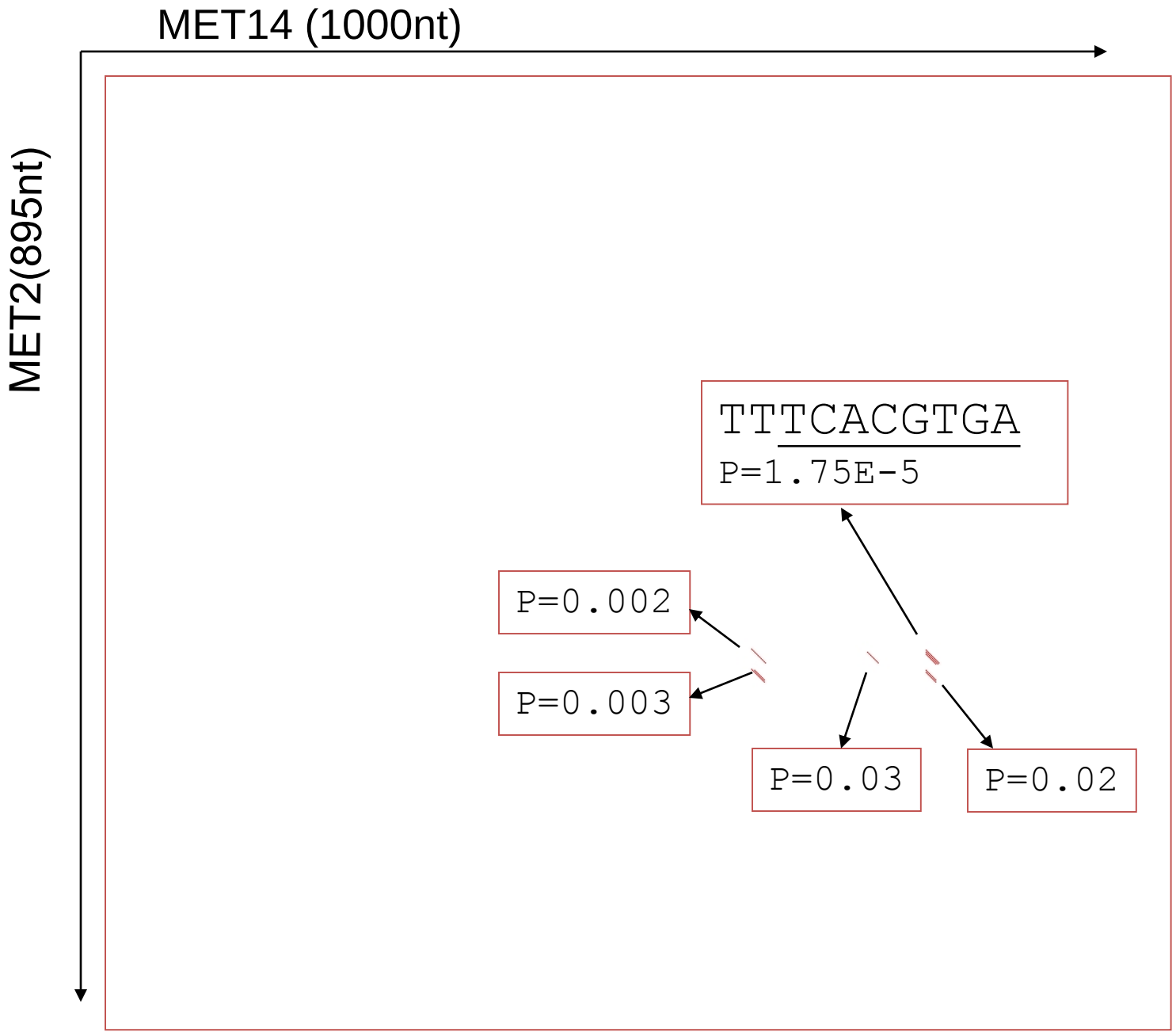
Red: >5



Met14 vs Met2

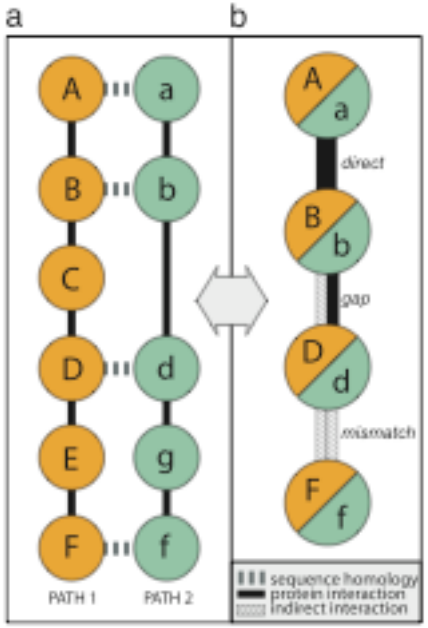
PhyloNet

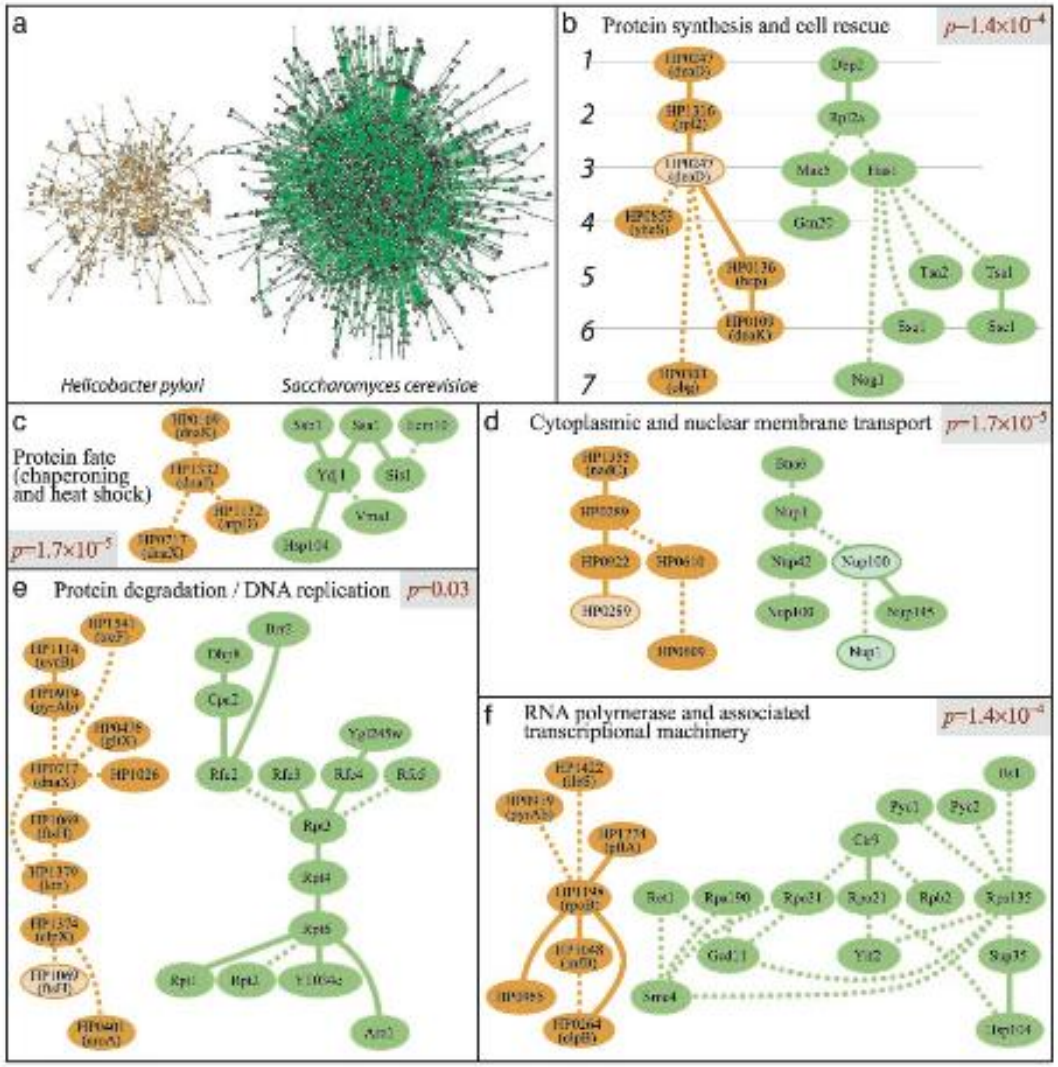
HSPs:
E < 0.1



PathBlast, NetworkBlast

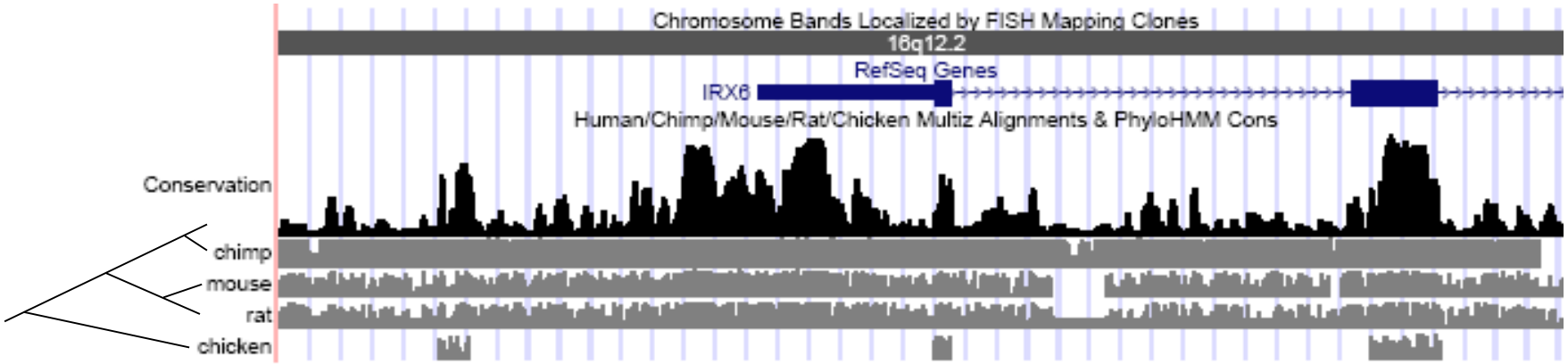
Trey Ideker



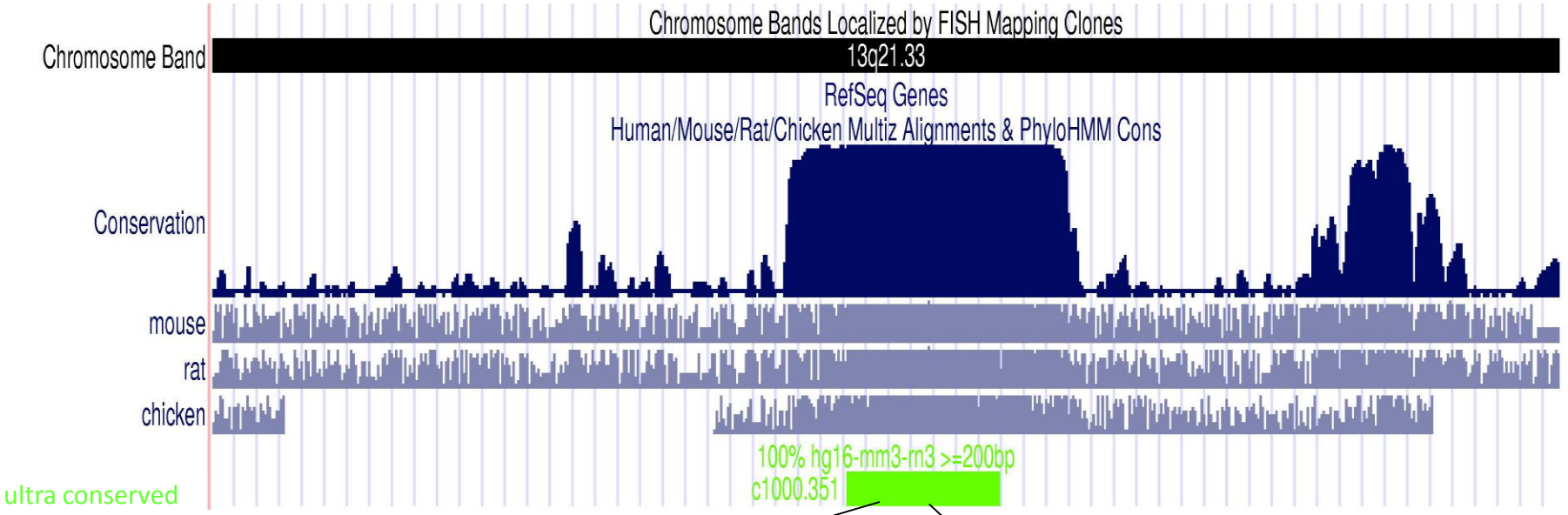


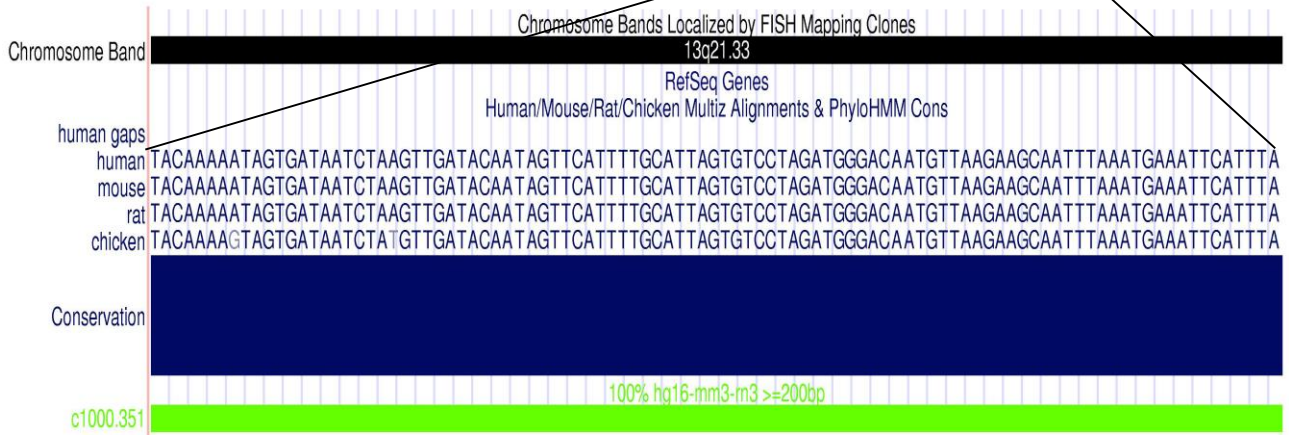
Comparative Genomics

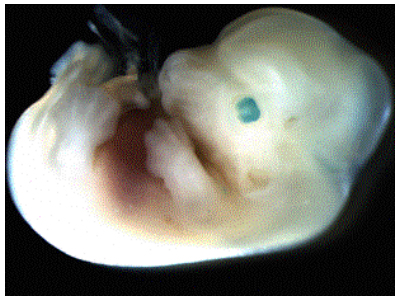
- Functional DNA often evolves slower than neutral DNA.
- To detect functional elements:
  - align genomes of related species,
  - and find regions of high conservation.
- The difference between conservation and constraint.



Ultra conserved elements

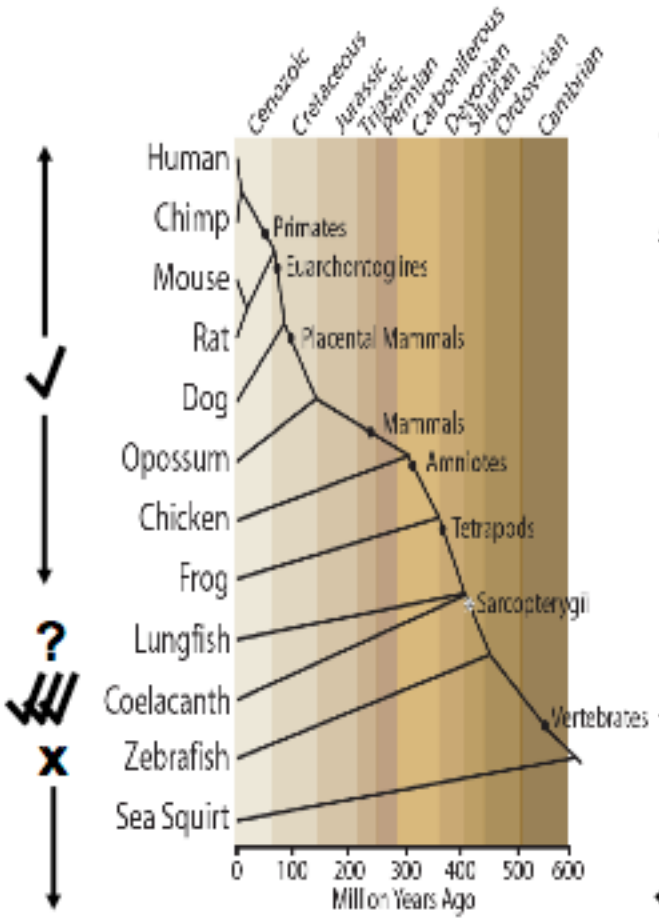






e.d 12.5

Uniquely Abundant in Coelacanth



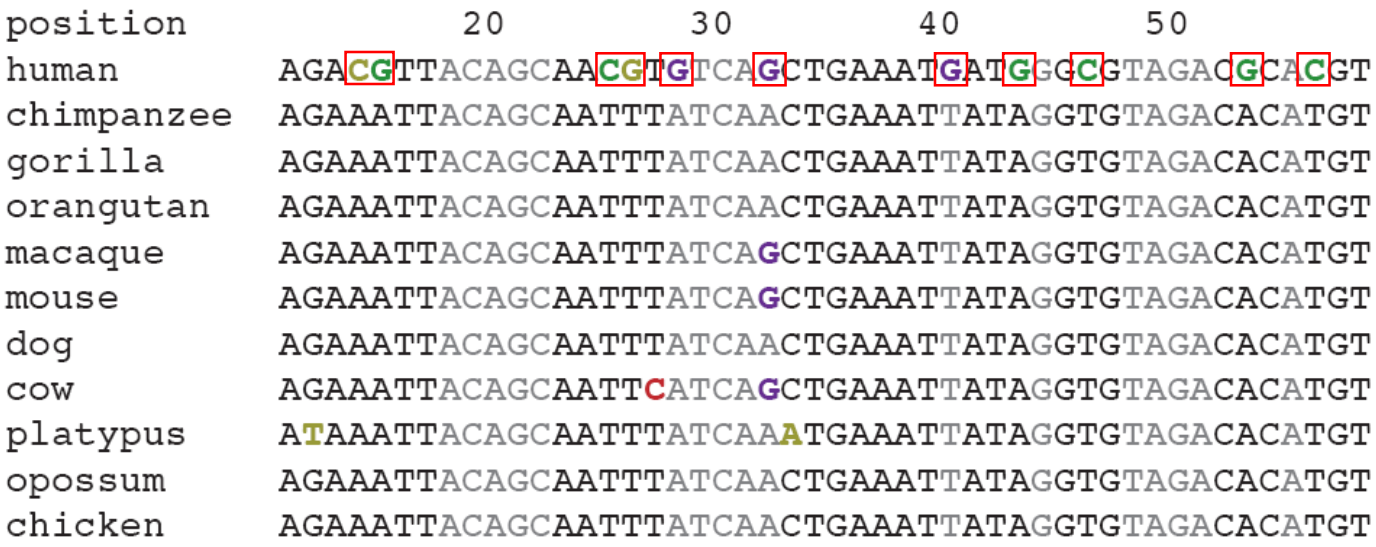
Upto 80%id between Coelacanth instances and some human instances, inc uc.338.

Species	UCSC Assembly	LF-SINE Detected	Species	UCSC Assembly	LF-SINE Detected
Homo sapiens	hg17	Yes	*Danio rerio*	danRer2	No
Pan troglodytes	panTro1	Yes	*Tetraodon nigroviridis*	tetNig1	No
Macaca mulatta	rheMac1	Yes	*Takifugu rubripes*	fr1	No
Mus musculus	mm6	Yes	*Ciona intestinalias*	ci1	No
Rattus norvegicus	rn3	Yes	*Strongylocentrotus purpuratus*	strPur1	No
Canis familiaris	canFam1	Yes	*Drosophila melanogaster*	dm2	No
Bos taurus	bosTau1	Yes	*Anopheles gambiae*	anoGam1	No
Monodelphis domestica	monDom1	Yes	*Caenorhabditis elegans*	ce2	No
Gallus gallus	galGal2	Yes	*Saccharomyces cerevisiae*	sacCer1	No
Xenopus tropicalis	xenTro1	Yes			

✓ 100 diverged copies in a Gigabase

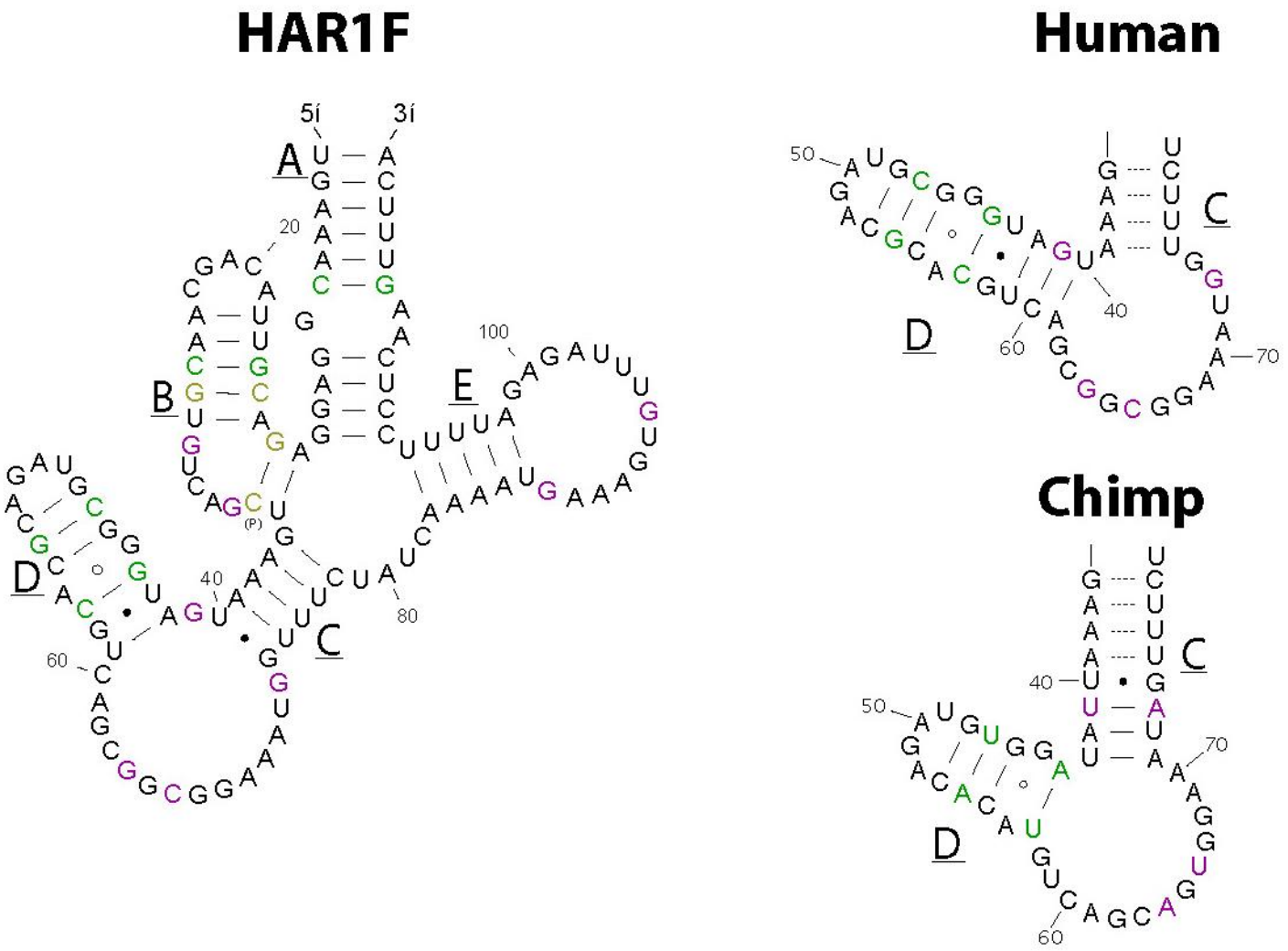
✓✓✓ 60 highly similar copies in a Megabase

HARs: Human accelerated regions



- 118 bp segment with 18 changes between the human and chimp sequences
- Expect less than 1

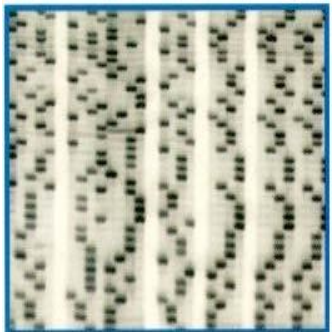
Human HAR1F differs from the ancestral RNA stucture



Aligning Short Reads

0 and 1st generation sequencing

Pre-1992
"old fashioned way"



S35 ddNTPs
Gels
Manual loading
Manual base calling

1992-1999
ABI 373/377

Fluorescent ddNTPs*
Gels
Manual loading
Automated base calling*

1999
ABI 3700

Fluorescent ddNTPs
Capillaries*
Robotic loading*
Automated base calling
Breaks down frequently

2003
ABI 3730XL

Fluorescent ddNTPs
Capillaries
Robotic loading
Automated base calling
Reliable*

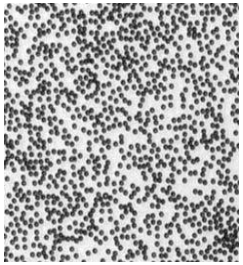
Next or 2nd-generation sequencing

454/Roche GS-20/FLX

(Oct 2005)

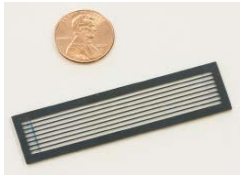
ABI SOLiD

(Oct 2007)

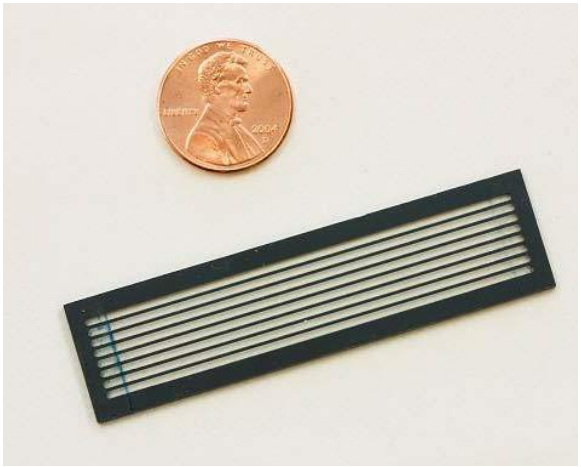


Illumina/Solexa

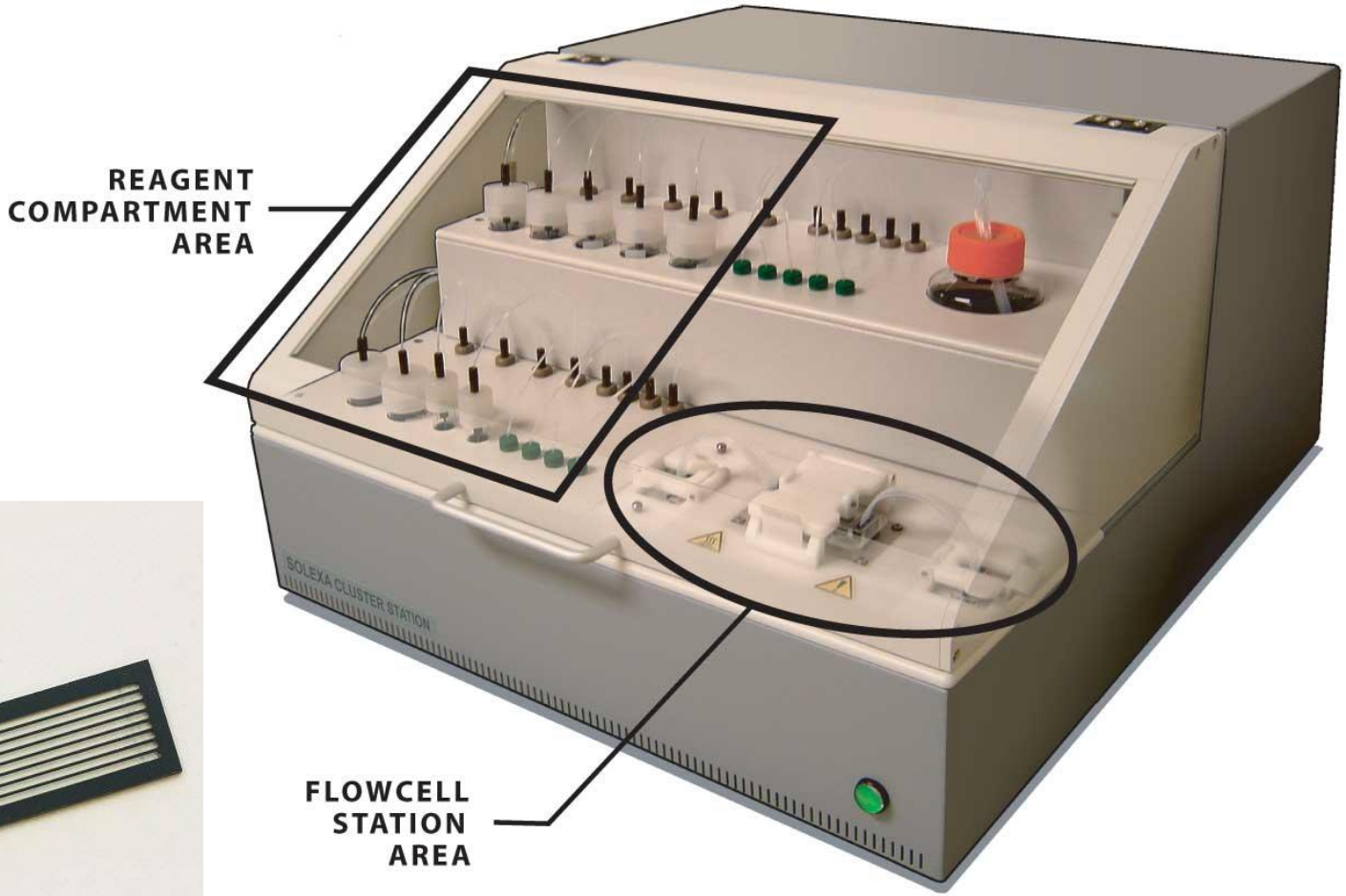
1G Genetic Analyser (Feb 2007)



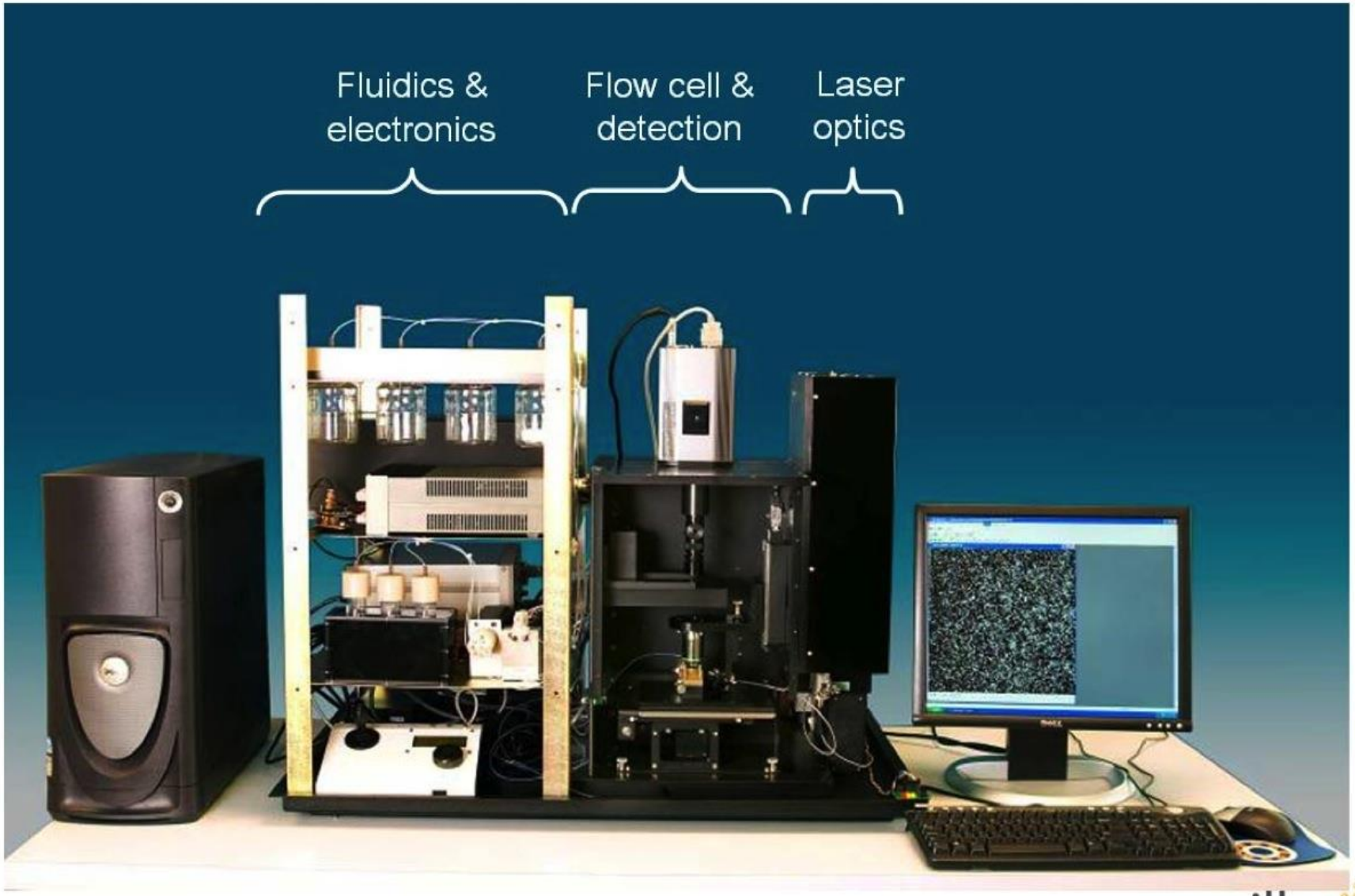
Cluster generation



8 channels (lanes)

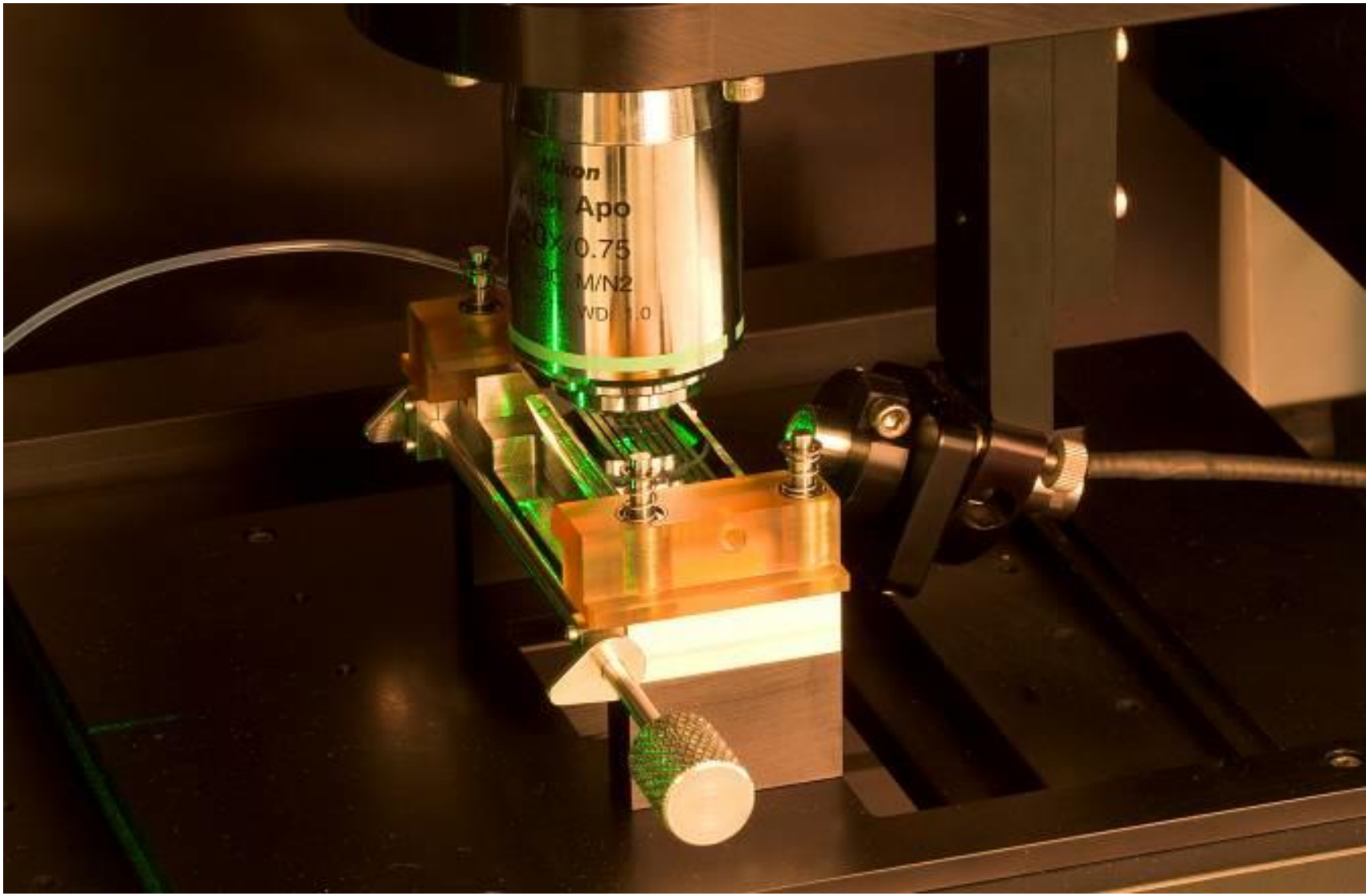


IGA without cover



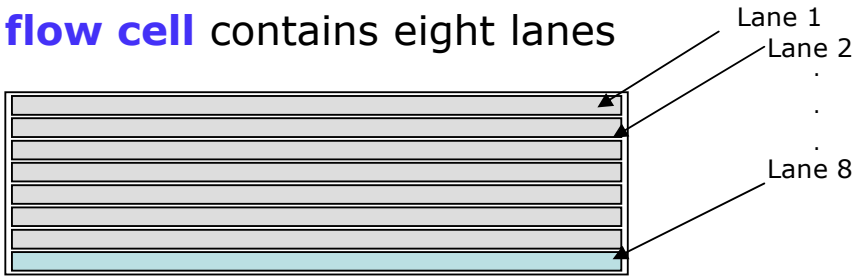
illumina

Flow cell imaging

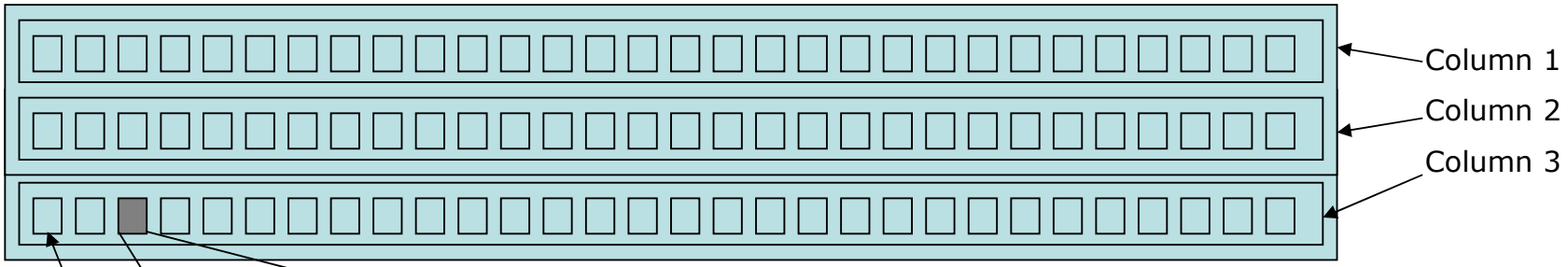


A flow cell

A **flow cell** contains eight lanes

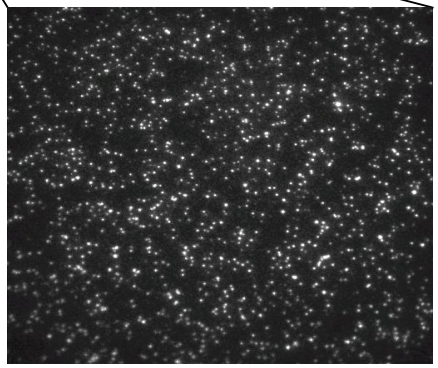


Each **lane/channel** contains **three columns** of tiles



Each **column** contains **100 tiles**

Tile



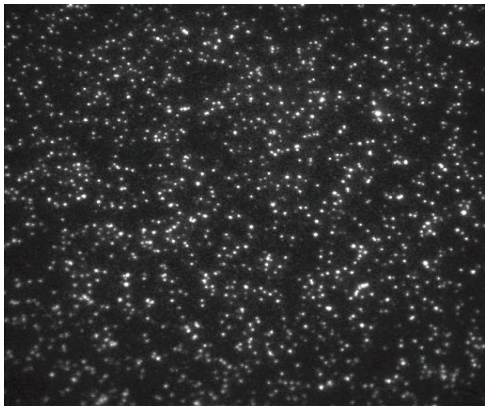
20K-30K Clusters

350 X 350 µm

Each tile is imaged four times per cycle – one image per base.

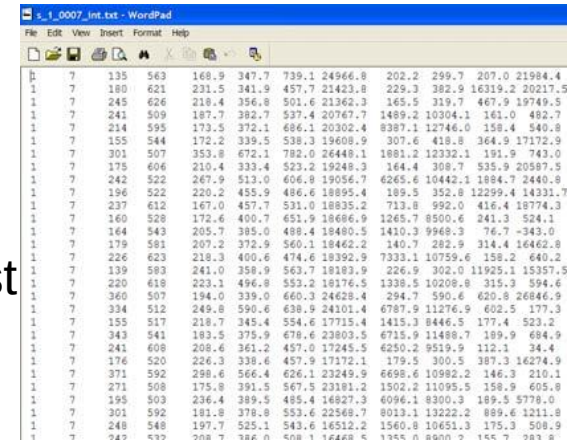
345,600 images for a 36-cycle run

Data analysis pipeline



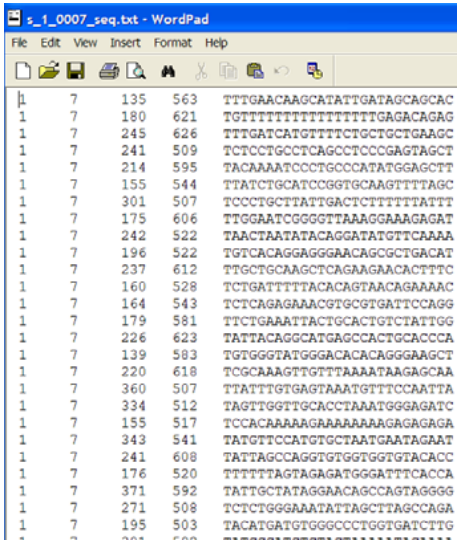
tiff image files
(345,600)

Firecrest



intensity files

Bustard



Sequence files

Eland

Alignment to Genome

Additional
Data Analysis

Primary tools and analysis tasks

- Image processing
  - (unique to each manufacturer)
- Basecalling
  - (unique to each manufacturer)
- Align sequence reads to reference genome
- Assemble contigs and whole genomes using quality scores and/or paired-end information
- Peak finding for Chip-Seq applications
  - (and statistics to validate, map to regulated genes, etc)
- SNP calling/genotyping
- Transcript profiling
  - measure gene expression, identifying alternative splicing, etc.

NGS: Sequence alignment

- Map the large numbers of short reads to a reference genome
  - In a broader sense: Identify similar sequences (DNA, RNA, or protein) in consequence of functional, structural, or evolutionary relationships between the them
  - Applications: Genome assembly, SNP detection, homology search, etc

- large ⇒ faster search speed
  short ⇒ greater search sensitivity.

Mapping Reads Back

- Hash Table (Lookup table)
  - FAST, but requires perfect matches
- Array Scanning
  - Can handle mismatches, but not gaps
- Dynamic Programming (Smith Waterman, Forward, Viterbi)
  - Indels
  - Mathematically optimal solution
  - Slow (most programs use Hash Mapping as a prefilter)
- Burrows-Wheeler Transform (BW Transform)
  - FAST (memory efficient)
  - But for gaps/mismatches, it lacks sensitivity

Many short read aligners

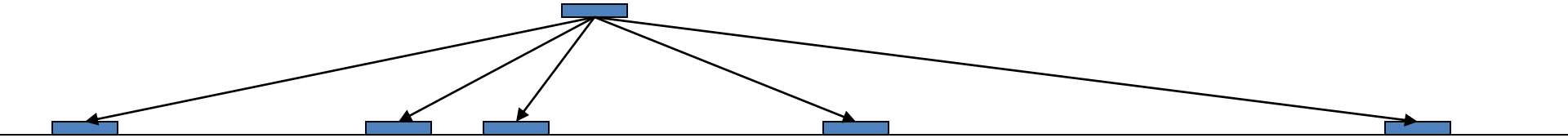
- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- .....

Short read mapping

- Input:
  - A reference genome
  - A collection of many 25-100bp tags (reads)
  - User-specified parameters
- Output:
  - One or more genomic coordinates for each tag

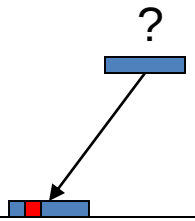
- In practice, only 70-75% of tags successfully map to the reference genome. Why?

Multiple mapping



- A single tag may occur more than once in the reference genome.
- The user may choose to ignore tags that appear more than *n* times.
- As *n* gets large, you get more data, but also more noise in the data.

Inexact matching



- An observed tag may not exactly match any position in the reference genome.
- Sometimes, the tag *almost* matches one or more positions.
- Such mismatches may represent a SNP or a bad read-out.
- The user can specify the maximum number of mismatches, or a phred-style quality score threshold.
- As the number of allowed mismatches goes up, the number of mapped tags increases, but so does the number of incorrectly mapped tags.

Using base qualities to evaluate

```
READ: AGGTCCGGGATACCGGGGAC
```

BETTER!

Q: 30

```
CHR1: CGGTCCGGGATACCGGGGAC
```

```
CHR2: AGGTCCGGGATACCGGGGGT
```

BETTER!

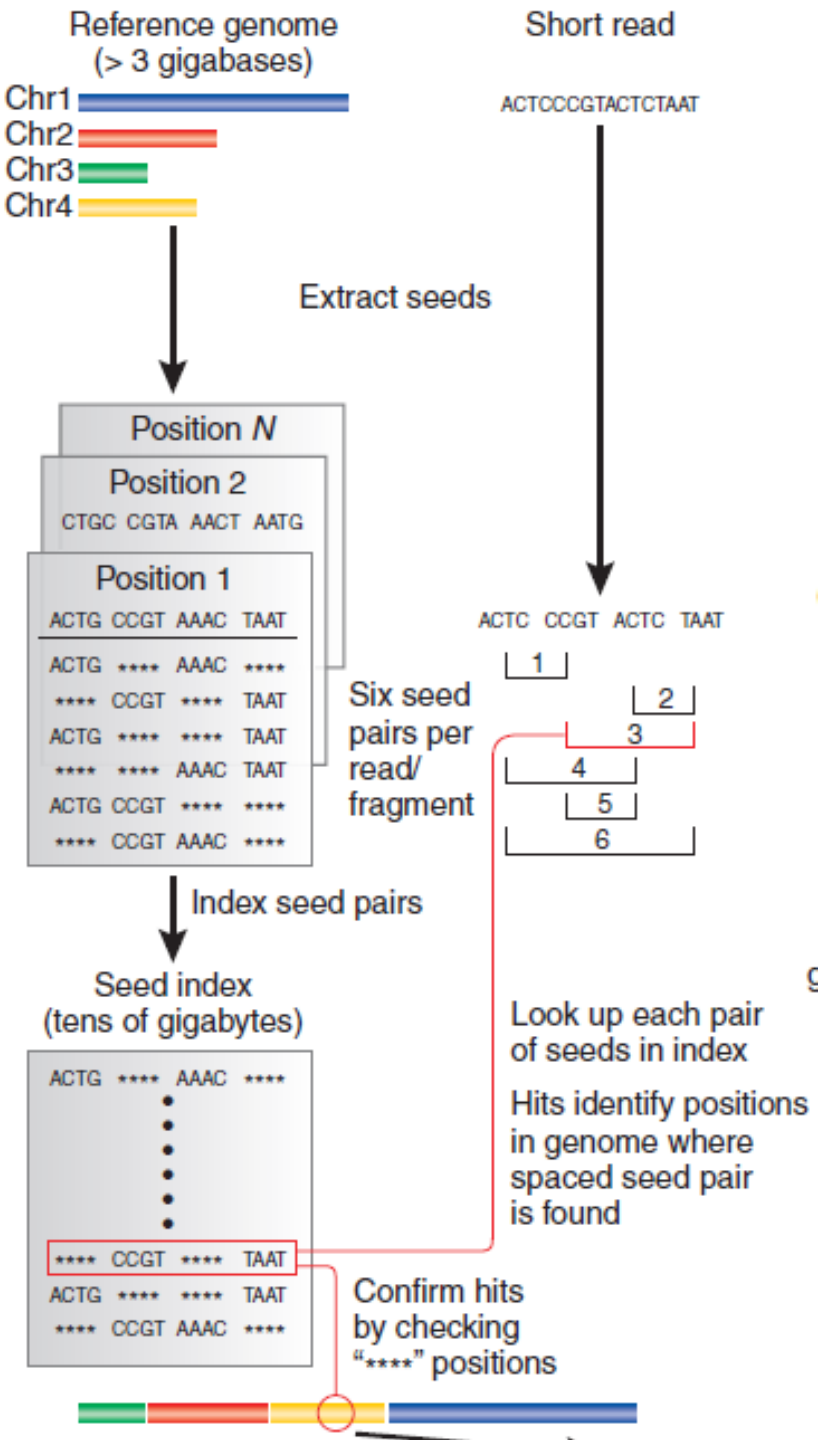
Q: 10+10

Hash table (Eland, SOAP)

- Main idea: preprocess genome to speed up queries
  - Hash every substring of length k
  - K is a tiny constant
- For each query p, can easily retrieve all suffixes of the genome that start with p1, p2, … pk.
- Easy to implement.
- Significant speed up in practice.
- Large memory consumption.
- Inexact match is difficult.
  - Need multiple hash tables
  - More memory

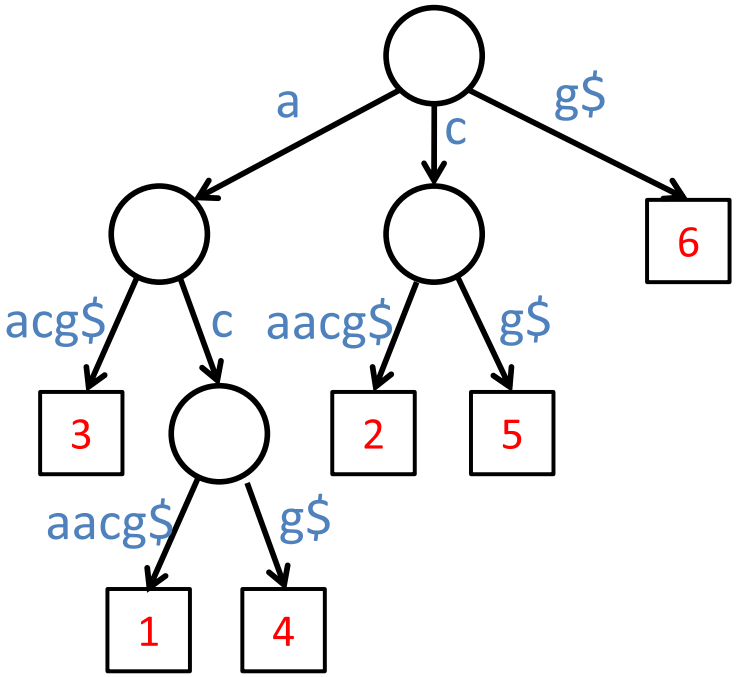
Spaced seed alignment (MAQ)

- Tags and tag-sized pieces of reference are cut into small “seeds.”
- Pairs of spaced seeds are stored in an index.
- Look up spaced seeds for each tag.
- For each “hit,” confirm the remaining positions.
- Report results to the user.

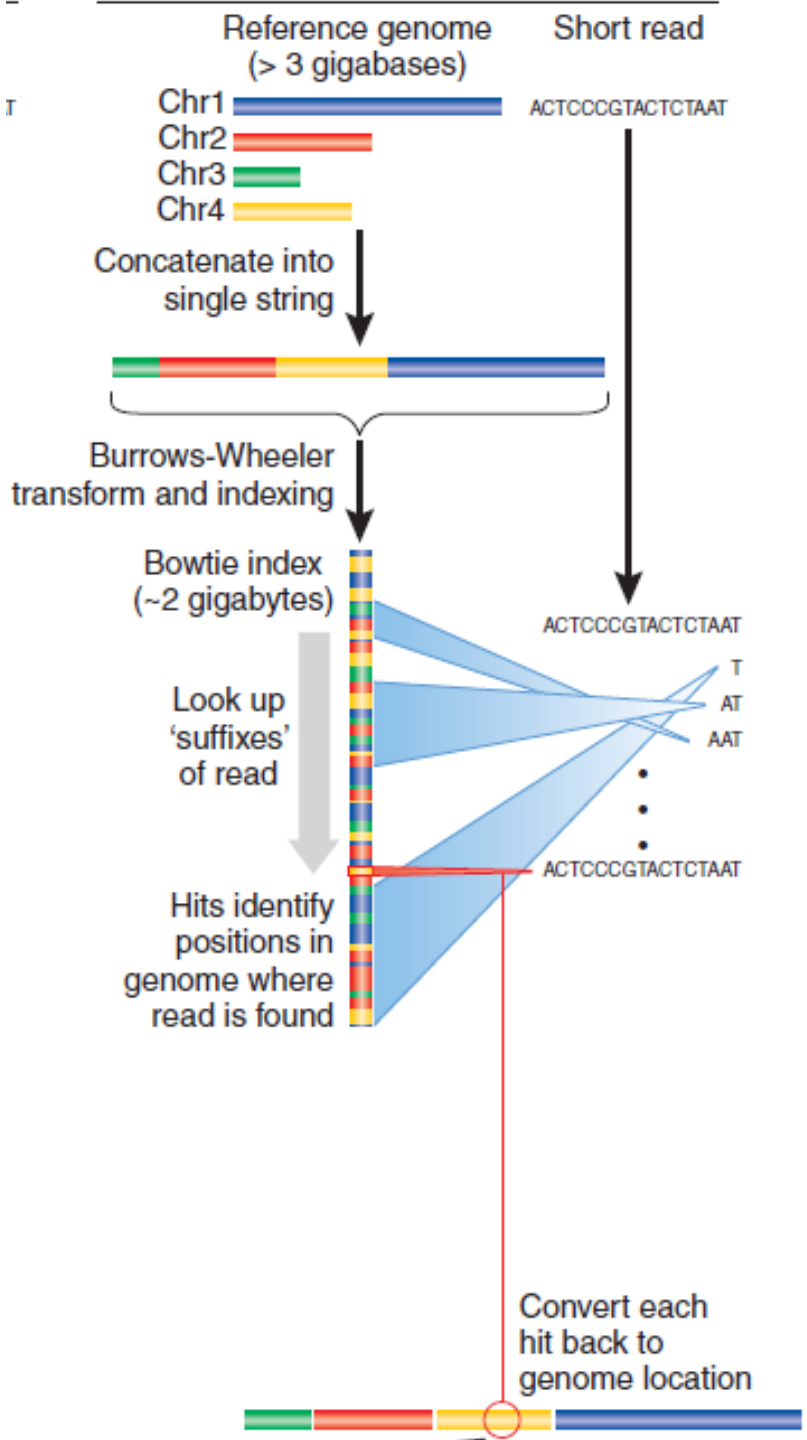


Index the reference genome: Suffix Tree

1	2	3	4	5	6	7
a	c	a	a	c	g	$



- Each suffix corresponds to exactly one path from the root to a leaf
- Edges spell non-empty strings
- Construction: linear time and space
- Check if a string of length m is a substring
- Each substring is a prefix of a suffix!



Burrows-Wheeler (Bowtie, BWA)

- Store entire reference genome.
- Align tag base by base from the end.
- When tag is traversed, all active locations are reported.
- If no match is found, then back up and try a substitution.

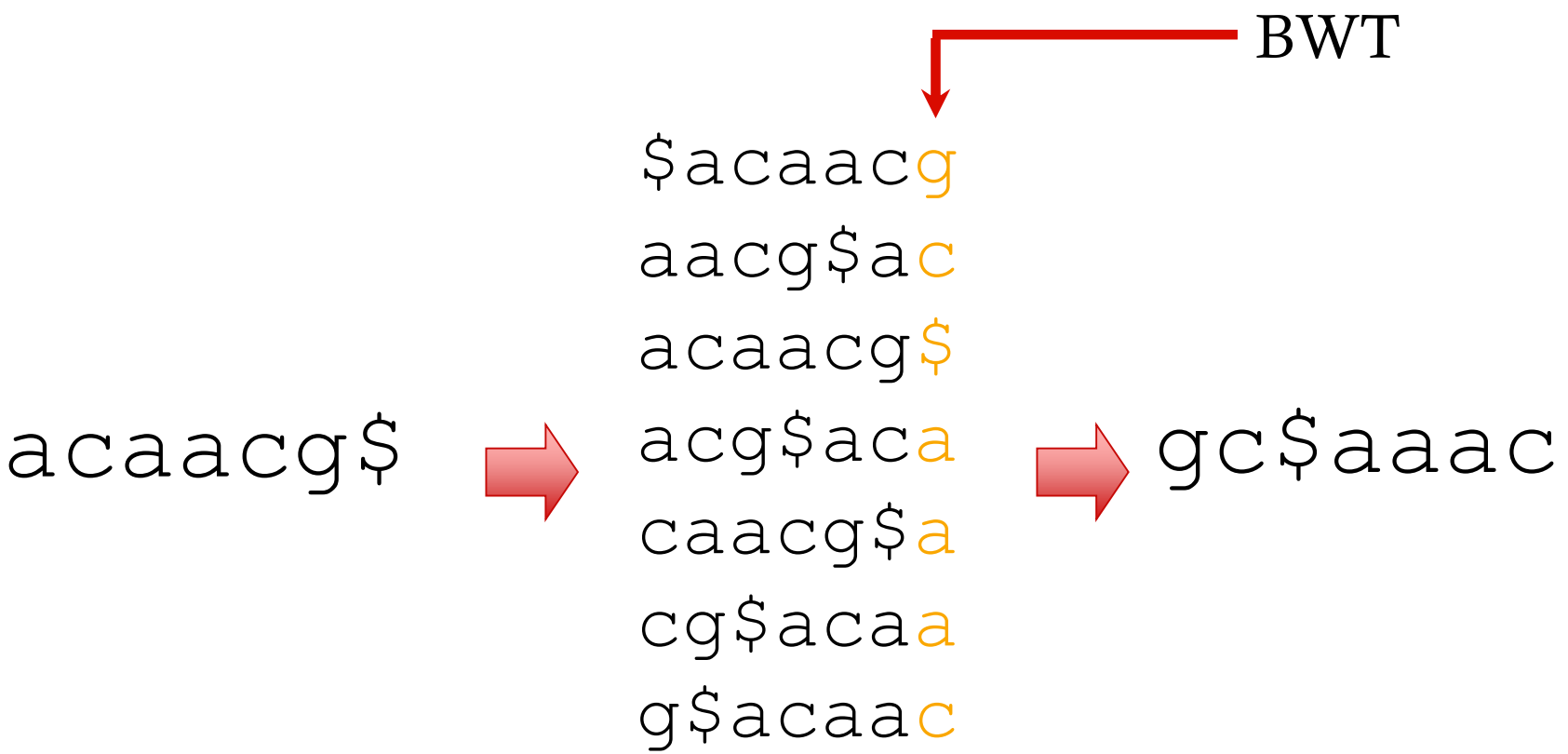
Why Burrows-Wheeler?

BWT very compact:

- Approximately ½ byte per base
- As large as the original text, plus a few "extras"
- Can fit onto a standard computer with 2GB of memory

- Linear-time search algorithm
  - proportional to length of query for exact matches

Burrows-Wheeler Transform (BWT)



Burrows-Wheeler Matrix (BWM)

Key observation

$a^1c^1a^2a^3c^2g^1\$^1$

“last first (LF) mapping”

The *i*-th occurrence of character X in the last column corresponds to the same text character as the *i*-th occurrence of X in the first column.

$^1\$acaacg^1$

$^2aacg\$ac^1$

$^1acaacg\$^1$

$^3acg\$aca^2$

$^1caacg\$a^1$

$^2cg\$acaa^3$

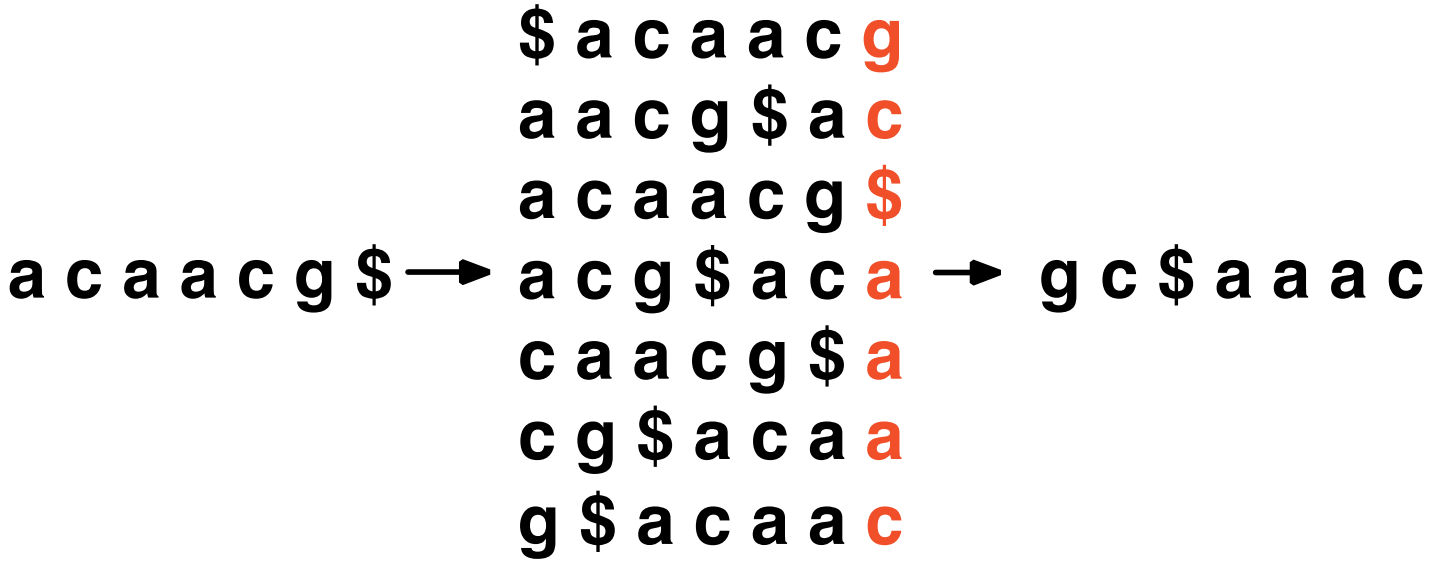
$^1g\$acaac^2$

Burrows-Wheeler Matrix

	$acaacg
3	aacg$ac
1	acaacg$
4	acg$aca
2	caacg$a
5	cg$acaa
6	g$acaac

See the suffix array?

Burrows-Wheeler Transform

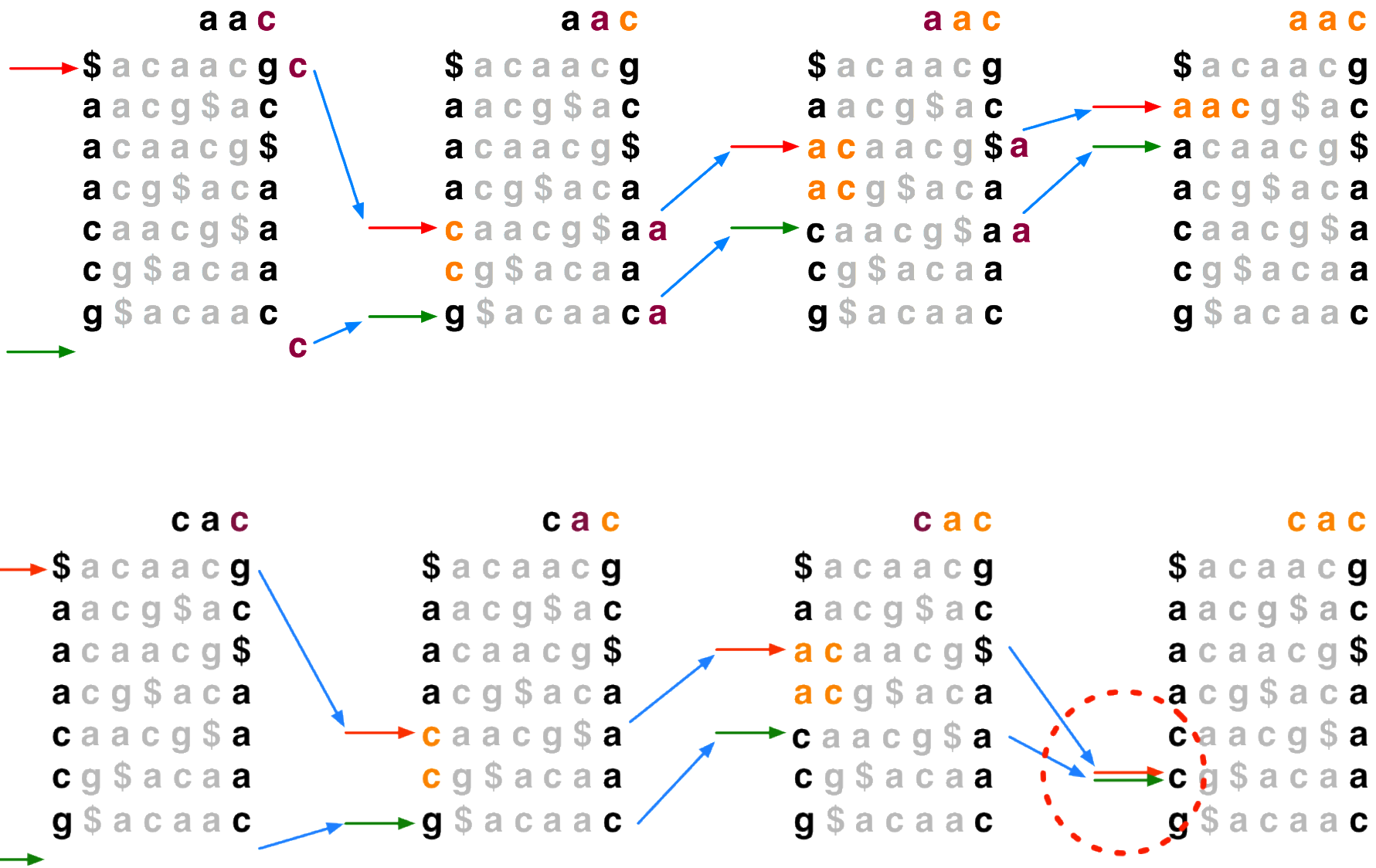


- Originally designed for data compression for large text
- Burrows-Wheeler matrix: sort lexicographically all cyclic rotations of S$
- BWT(S): the last column of Burrows-Wheeler matrix
- Compression: runs of repeated characters are easy to compress using move-to-front transform and run-length encoding, etc.
- BWT(S) is a reversible permutation of S

Reverse Burrows-Wheeler Transform

- BW Matrix Property: Last-First (LF) Mapping
- The ith occurrence of character X in the last column correspond to the same text character as the ith occurrence of X in the first column

Searching BWT



Searching BWT

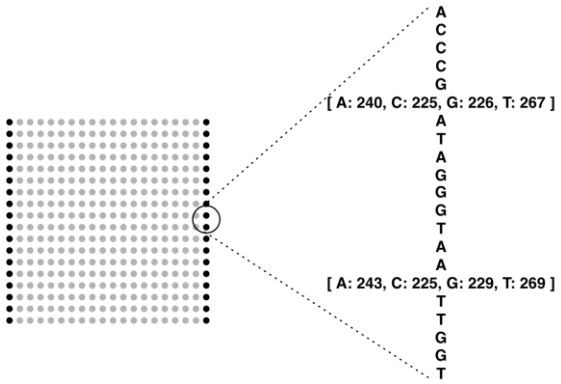
BWT(agcagcagact) = tgcc$ggaaaac

Search for pattern: gca

gca	gca	gca	gca
$agcagcagact	$agcagcagact	$agcagcagact	$agcagcagact
act$agcagcag	act$agcagcag	act$agcagcag	act$agcagcag
agact$agcagc	agact$agcagc	agact$agcagc	agact$agcagc
agcagact$agc	agcagact$agc	agcagact$agc	agcagact$agc
agcagcagact$	agcagcagact$	agcagcagact$	agcagcagact$
cagact$agcag	cagact$agcag	cagact$agcag	cagact$agcag
cagcagact$ag	cagcagact$ag	cagcagact$ag	cagcagact$ag
ct$agcagcaga	ct$agcagcaga	ct$agcagcaga	ct$agcagcaga
gact$agcagca	gact$agcagca	gact$agcagca	gact$agcagca
gcagact$agca	gcagact$agca	gcagact$agca	gcagact$agca
gcagcagact$a	gcagcagact$a	gcagcagact$a	gcagcagact$a
t$agcagcagac	t$agcagcagac	t$agcagcagac	t$agcagcagac

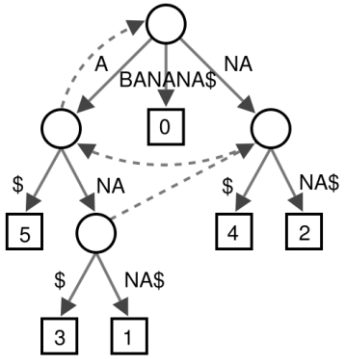
Human genome memory footprint

Bowtie Index



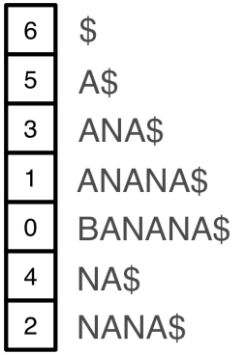
1.3 gigabytes

Suffix Tree



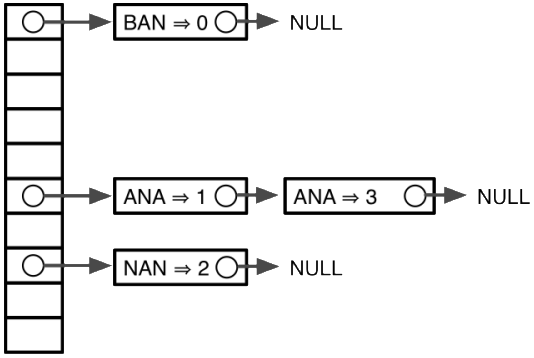
>35 gigabytes

Suffix Array



>12 gigabytes

Hash Tables



>12 gigabytes