CSE 527
Computational Biology
Autumn 2009

3: BLAST, Alignment score significance;

PCR and DNA sequencing

Outline

BLAST

Scoring

Weekly Bio Interlude: PCR & Sequencing

BLAST:

Basic Local Alignment Search Tool

Altschul, Gish, Miller, Myers, Lipman, J Mol Biol 1990

The most widely used comp bio tool

Which is better: long mediocre match or a few nearby, short, strong matches with the same total score?

- score-wise, exactly equivalent
- biologically, later may be more interesting, & is common
- at least, if must miss some, rather miss the former

BLAST is a heuristic emphasizing the later

- speed/sensitivity tradeoff: BLAST may miss former, but gains greatly in speed

BLAST: What

Input:

- A query sequence (say, 300 residues)
- A data base to search for other sequences similar to the query (say, 10^6 - 10^9 residues)
- A score matrix $\sigma(r,s)$, giving cost of substituting r for s (& perhaps gap costs)
- Various score thresholds & tuning parameters

Output:

- “All” matches in data base above threshold
- “E-value” of each

BLAST: How

Idea: most interesting parts of DB are those with a good ungapped match to some short subword of the query

Break query into overlapping words w_i of small fixed length (e.g. 3 aa or 11 nt)

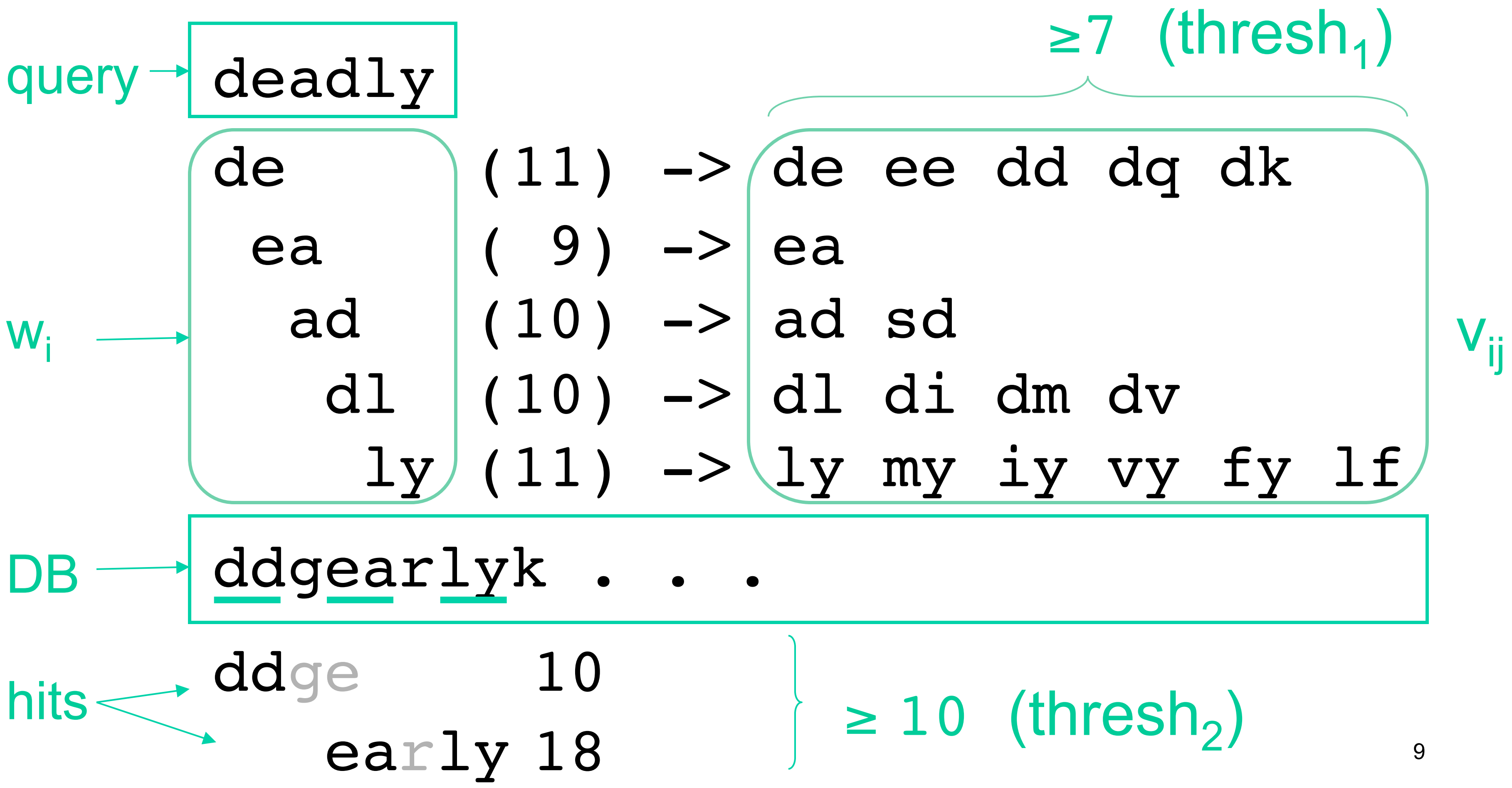
For each w_i, find (empirically, ~50) “neighboring” words v_{ij} with score $\sigma(w_i, v_{ij}) > \text{thresh}_1$

Look up each v_{ij} in database (via prebuilt index) -- i.e., exact match to short, high-scoring word

Extend each such “seed match” (bidirectional)

Report those scoring > thresh_2, calculate E-values

BLAST: Example



query → deadly

≥7 (thresh$_1$)

w_i

```
de        (11) -> de ee dd dq dk
 ea       ( 9) -> ea
  ad      (10) -> ad sd
   dl     (10) -> dl di dm dv
    ly    (11) -> ly my iy vy fy lf
```

v_{ij}

DB → ddgearlyk . . .

hits

```
ddge      10
   early  18
```

≥ 10 (thresh$_2$)

BLOSUM 62

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

BLAST Refinements

“Two hit heuristic” -- need 2 nearby, nonoverlapping, gapless hits before trying to extend either

“Gapped BLAST” -- run heuristic version of Smith-Waterman, bi-directional from hit, until score drops by fixed amount below max

PSI-BLAST -- For proteins, iterated search, using “weight matrix” pattern from initial pass to find weaker matches in subsequent passes

Many others

Significance of Alignments

Is “42” a good score?

Compared to what?

Usual approach: compared to a specific “null model”, such as “random sequences”

Hypothesis Testing: A Very Simple Example

Given: A coin, either fair (p(H)=1/2) or biased (p(H)=2/3)

Decide: which

How? Flip it 5 times. Suppose outcome D = HHHTH

Null Model/Null Hypothesis M_0: p(H)=1/2

Alternative Model/Alt Hypothesis M_1: p(H)=2/3

Likelihoods:

$P(D \mid M_0) = (1/2)\ (1/2)\ (1/2)\ (1/2)\ (1/2) = 1/32$

$P(D \mid M_1) = (2/3)\ (2/3)\ (2/3)\ (1/3)\ (2/3) = 16/243$

Likelihood Ratio: $\frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{16/243}{1/32} = \frac{512}{243} \approx 2.1$

I.e., alt model is ≈ 2.1x more likely than null model, given data

Hypothesis Testing, II

Log of likelihood ratio is equivalent, often more convenient

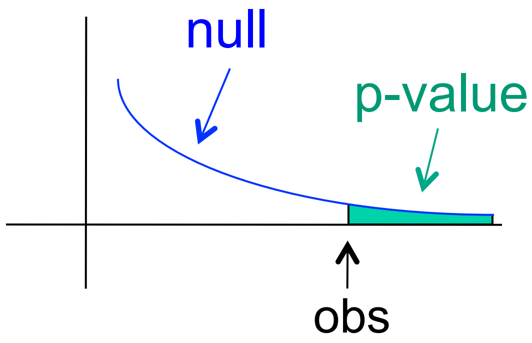
add logs instead of multiplying…

“Likelihood Ratio Tests”: reject null if LLR > threshold

LLR > 0 disfavors null, but higher threshold gives stronger evidence against

Neyman-Pearson Theorem: For a given error rate, LRT is as good a test as any (subject to some fine print).

p-values



The *p-value* of such a test is the probability, assuming that the null model is true, of seeing data as extreme or more extreme than what you actually observed

E.g., we observed 4 heads; p-value is prob of seeing 4 or 5 heads in 5 tosses of a fair coin

Why interesting? It measures *probability that we would be making a mistake in rejecting null*.

Can analytically find p-value for simple problems like coins; often turn to simulation/permutation tests (introduced earlier) or to approximation (coming soon) for more complex situations

Usual scientific convention is to reject null only if p-value is < 0.05; sometimes demand p << 0.05 (esp. if estimates are inaccurate)

A Likelihood Ratio

Defn: two proteins are *homologous* if they are alike because of shared ancestry; similarity by descent

Suppose among proteins overall, residue x occurs with frequency p_x

Then in a random alignment of 2 random proteins, you would expect to find x aligned to y with prob $p_x p_y$

Suppose among *homologs*, x & y align with prob p_{xy}

Are seqs X & Y homologous? Which is more likely, that the alignment reflects chance or homology? Use a *likelihood ratio test.*

$$\sum_i \log \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

Non-*ad hoc* Alignment Scores

Take alignments of homologs and look at frequency of *x*-*y* alignments *vs* freq of *x, y* overall

Issues

- biased samples
- evolutionary distance

BLOSUM approach

- Large collection of trusted alignments
  - (the BLOCKS DB)
- Subset by similarity
  - BLOSUM62 ⇒ ≥ 62% identity
- e.g. http://blocks.fhcrc.org/blocks-bin/getblock.pl?IPB013598

$$\frac{1}{\lambda}\log_2 \frac{p_{x\ y}}{p_x p_y}$$

ad hoc Alignment Scores?

Make up any scoring matrix you like

Somewhat surprisingly, under pretty general assumptions[**], it is *equivalent* to the scores constructed as above from some set of probabilities p_{xy}, so you might as well understand what they are

NCBI-BLAST: +1/-2

WU-BLAST: +5/-4

[**] e.g., average scores should be negative, but you probably want that anyway, otherwise local alignments turn into global ones, and some score must be > 0, else best match is empty

BLOSUM 62

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

Alignment Scores vs Test Statistic

Alignment alg *works hard* to contort data into a high-scoring alignment

Goal of test statistic is to discriminate good/bad ones

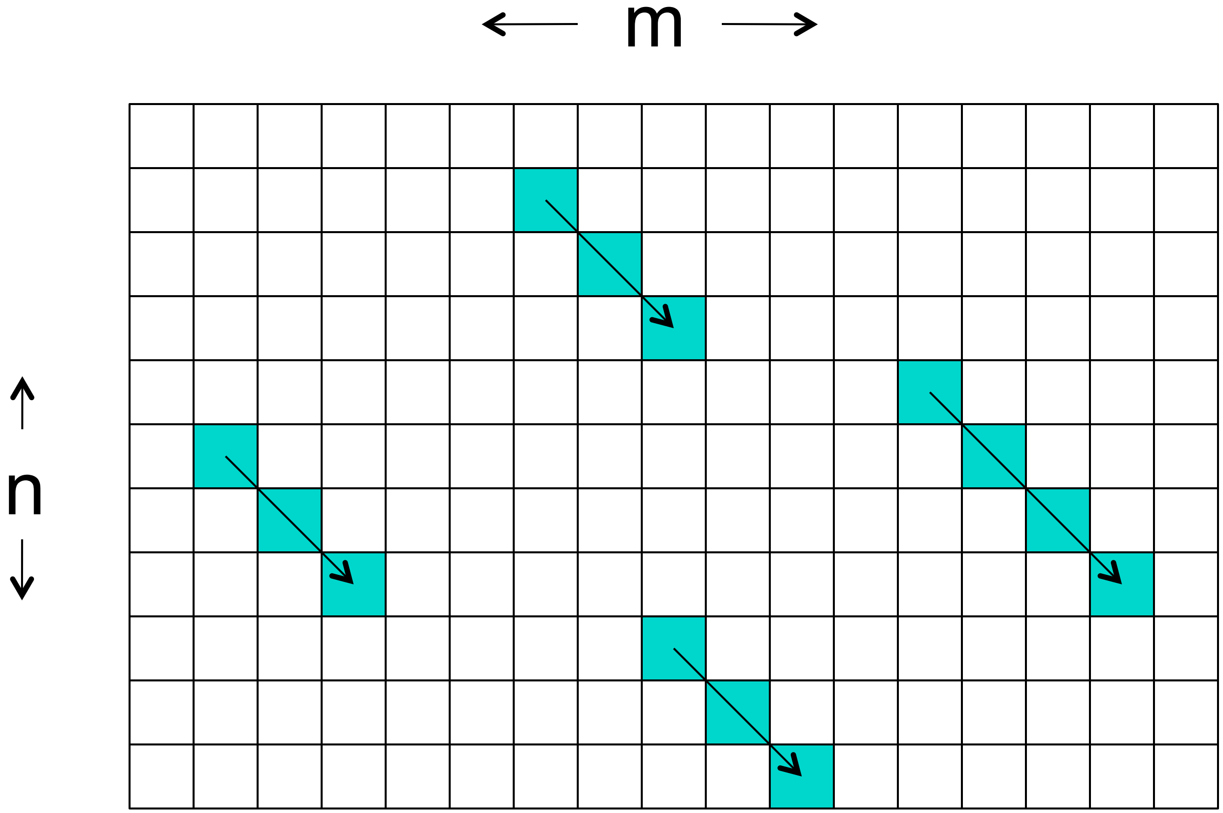
Why use same score? Doesn’t a better alg just push up scores? Maybe better to test via an *independent* criterion?

A: Yes, better alg may raise background scores. *But***,** want best discrimination in both phases, so use best possible score/test statistic, with appropriate threshold, rather than an indp. criterion

Note: best random match looks like real match (e.g. same matching-letter frequencies), except for score.

One reason to score/test differently–if score is too expensive for search, might try search w/ approx score, look at multiple hits

Random (ungapped) local alignment



it’s *max* of m*n ~indp random scores

Overall Alignment Significance, I
A Theoretical Approach: EVD

Let X_i, $1 \le i \le N$, be indp. random variables drawn from some (non-pathological) distribution

Q. what can you say about distribution of $y = sum\{X_i\}$?

A. y is approximately *normally* distributed

Q. what can you say about distribution of $y = max\{X_i\}$?

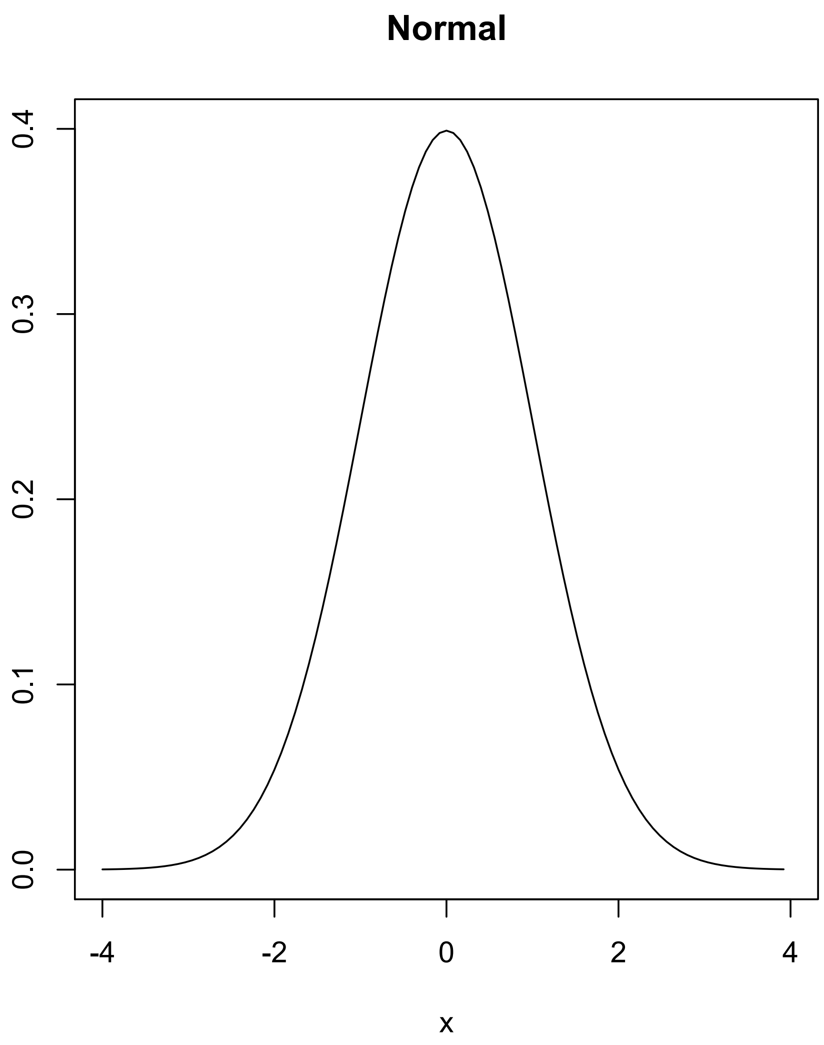
A. it's approximately an *Extreme Value Distribution (EVD)*

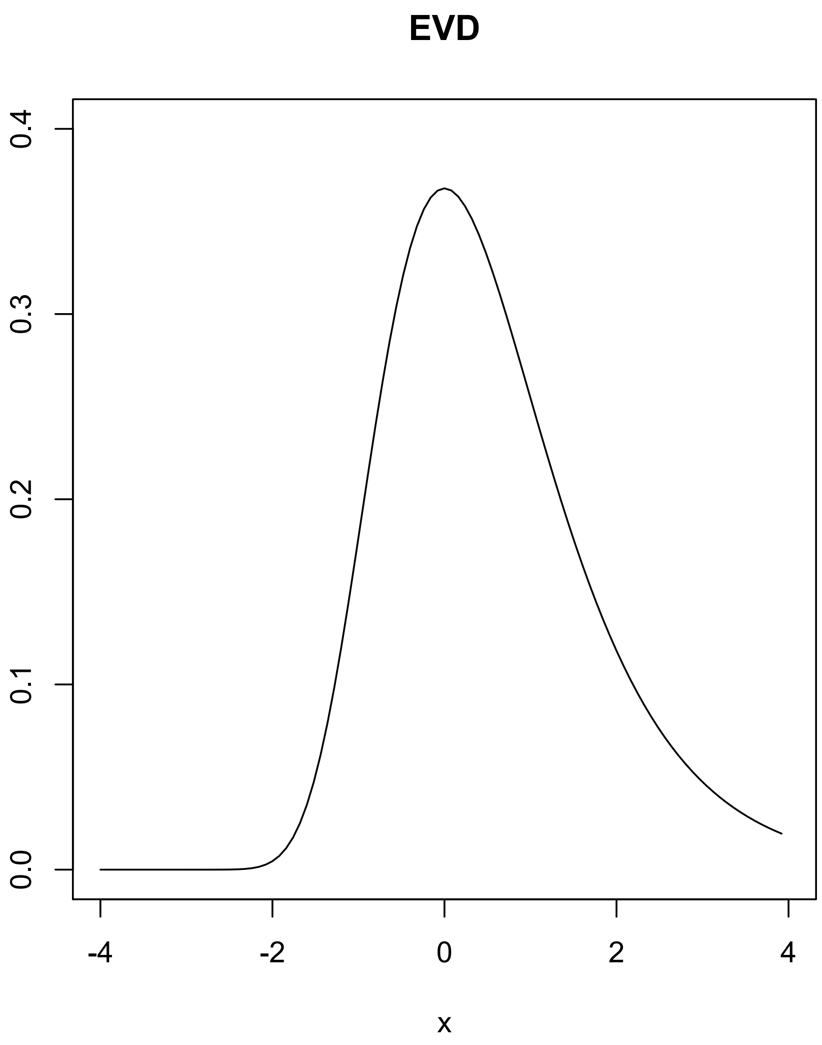
[one of only 3 kinds; for our purposes, the relevant one is:]

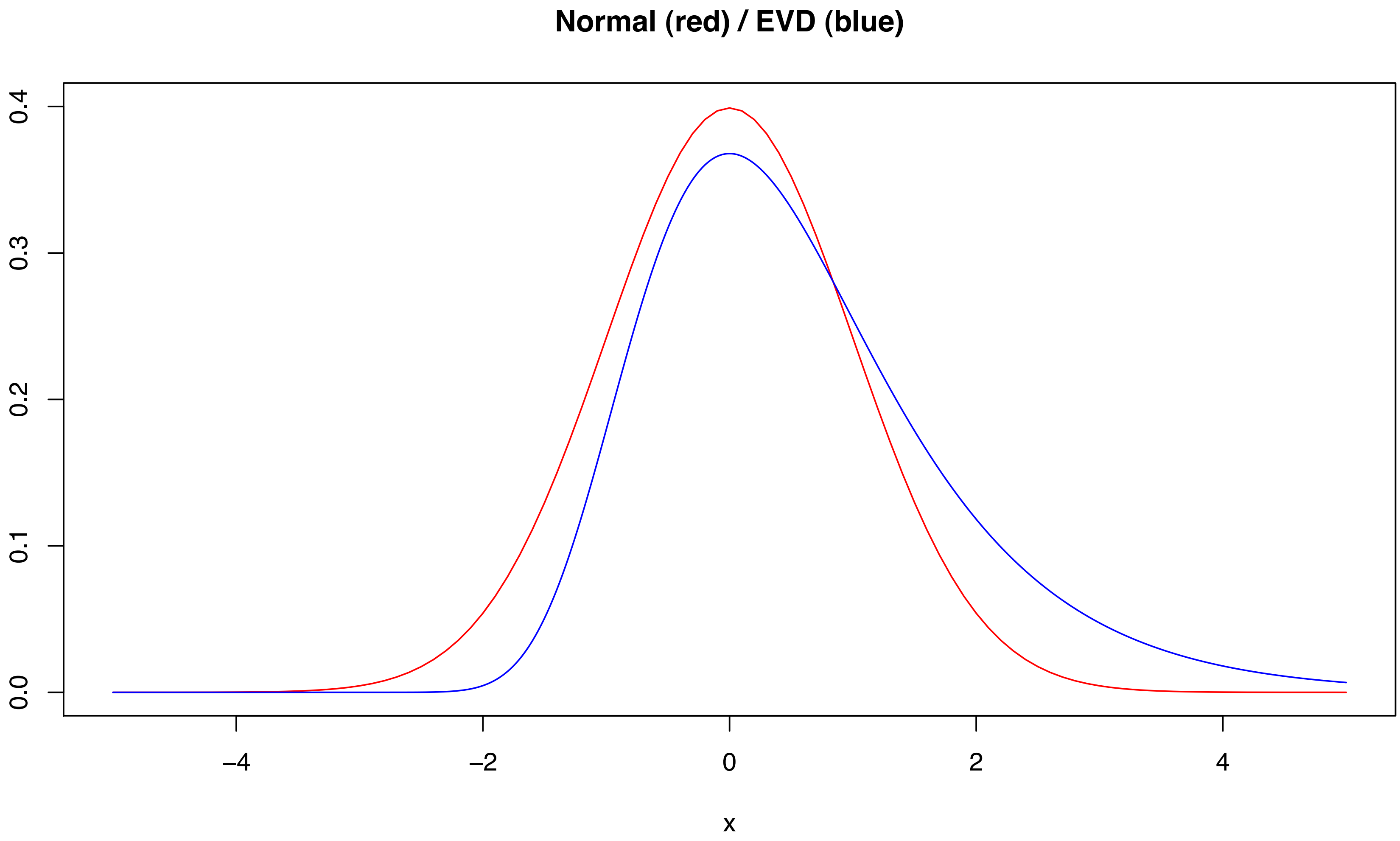
$$P(y \le z) \approx \exp(-KNe^{-\lambda(z-\mu)}) \qquad (*)$$

For ungapped local alignment of seqs x, y, N ~ |x|*|y|

λ, K depend on scores, etc., or can be estimated by curve-fitting random scores to (*). (cf. reading)


Normal
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x


EVD
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x


Normal (red) / EVD (blue)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

EVD Pro/Con

Pro:

- Gives p-values for alignment scores

Con:

- It’s only approximate
- Parameter estimation
- Theory may not apply. E.g., NOT proven to hold for gapped alignments (although strong empirical support).

Overall Alignment Significance, II
Empirical (via randomization)

Generate N random sequences (say N = 10^3 - 10^6)

Align *x* to each & score

If *k* of them have better score than alignment of *x* to *y*, then the (empirical) probability of a chance alignment as good as observed *x:y* alignment is (*k*+1)/(*N*+1)

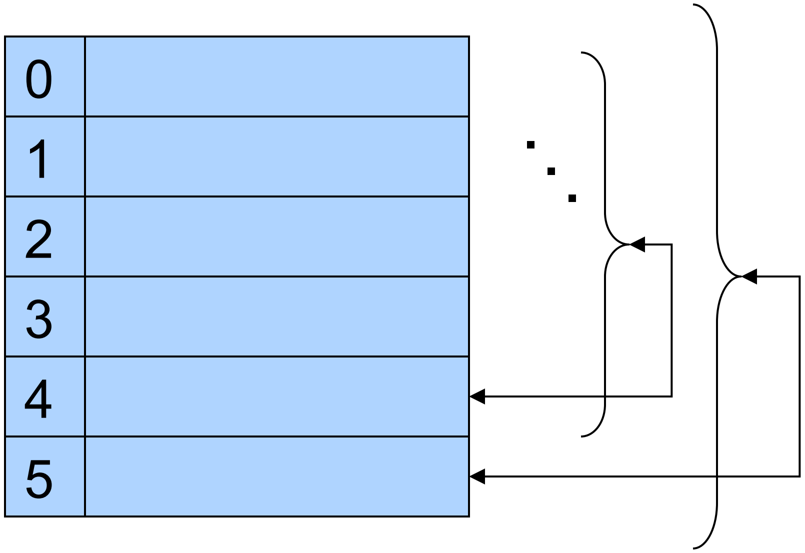
- e.g., if 0 of 99 are better, you can say "estimated p < .01"

How to generate "random" sequences?

- Scores are often sensitive to sequence composition
- So uniform 1/20 or 1/4 is a bad idea
- Even background p_i can be dangerous
- Better idea: *permute* y N times

Generating Random Permutations

```
for (i = n-1; i > 0; i--){
    j = random(0..i);
    swap X[i] <-> X[j];
}
```



All n! permutations of the original data equally likely: A specific element will be last with prob 1/n; given that, a specific other element will be next-to-last with prob 1/(n-1), …; overall: 1/(n!)

Permutation Pro/Con

Pro:

- Gives empirical p-values for alignments with characteristics like sequence of interest, e.g. residue frequencies
- Largely free of modeling assumptions (e.g., ok for gapped…)

Con:

- Can be inaccurate if your method of generating random sequences is unrepresentative
- E.g., probably better to preserve di-, tri-residue statistics and/or other higher-order characteristics, but increasingly hard to know exactly what to model & how
- Slow
- Especially if you want to assess low-probability p-values

E-values

“p-value”: *probability* of a score more extreme than observed in a given random target data base

E-value: expected *number* of matches that good or better in a random data base of the given size & composition

Related: $P = 1 - \exp(-E)$

$E = 5 \iff P = .993$

$E = 10 \iff P = .99995$

$E = .01 \iff P = E - E^2/2 + E^3/3! \; \ldots \approx E$

both equally valid; E-value is perhaps a more intuitively interpretable quantity, & perhaps makes role of data base size more explicit

Issues

What if the model is wrong?

E.g., are adjacent positions really independent?

Summary

BLAST is a highly successful search/alignment heuristic. It looks for alignments anchored by short, strong, ungapped “seed” alignments

Assessing statistical significance of alignment scores is crucial to practical applications

- Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model
- For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments
- Permutation tests are a simple (but brute force) alternative

Weekly Bio(tech) Interlude

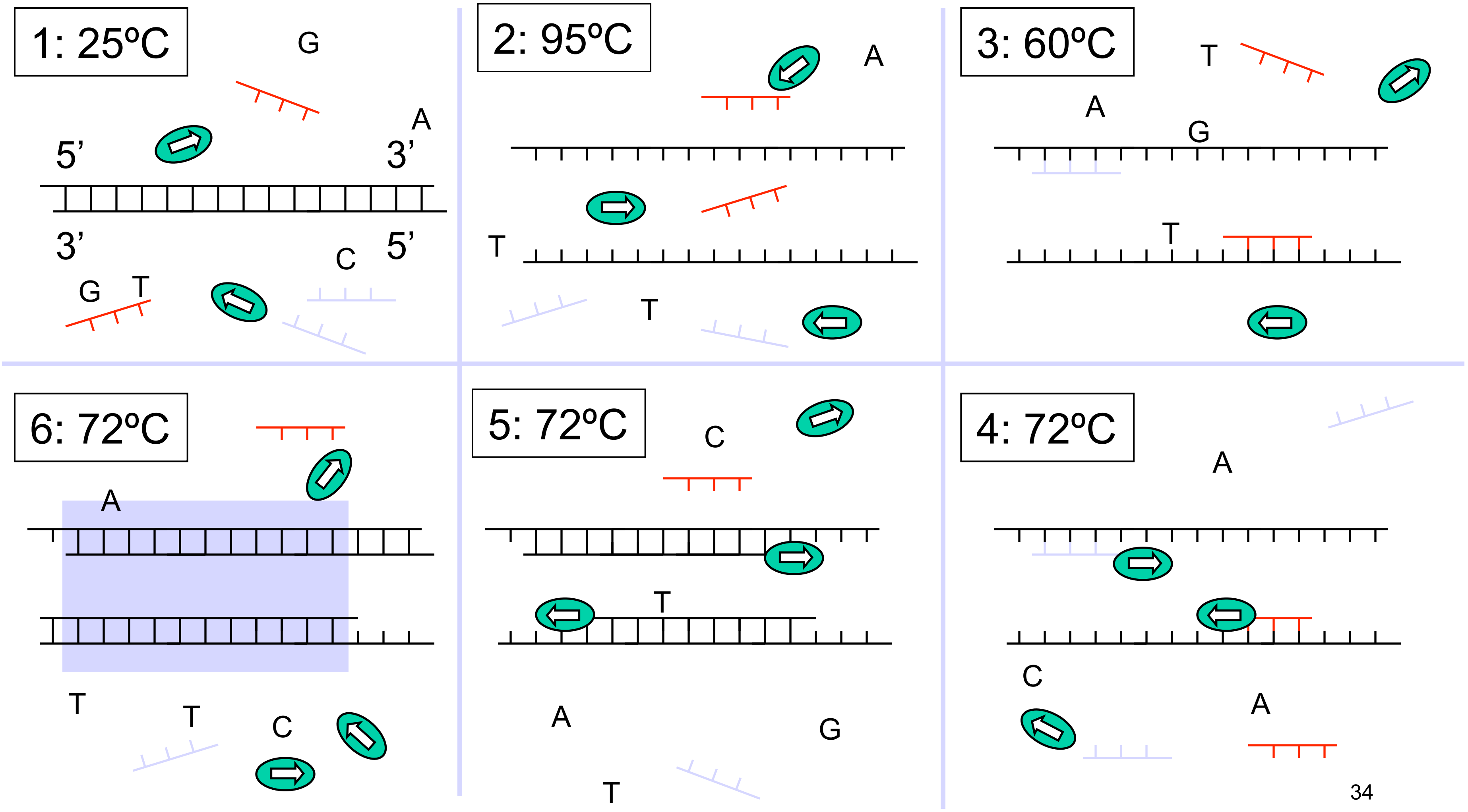
3 Nobel Prizes:

PCR: Kary Mullis, 1993

Electrophoresis: A.W.K. Tiselius, 1948

DNA Sequencing: Frederick Sanger, 1980

PCR



1: 25ºC

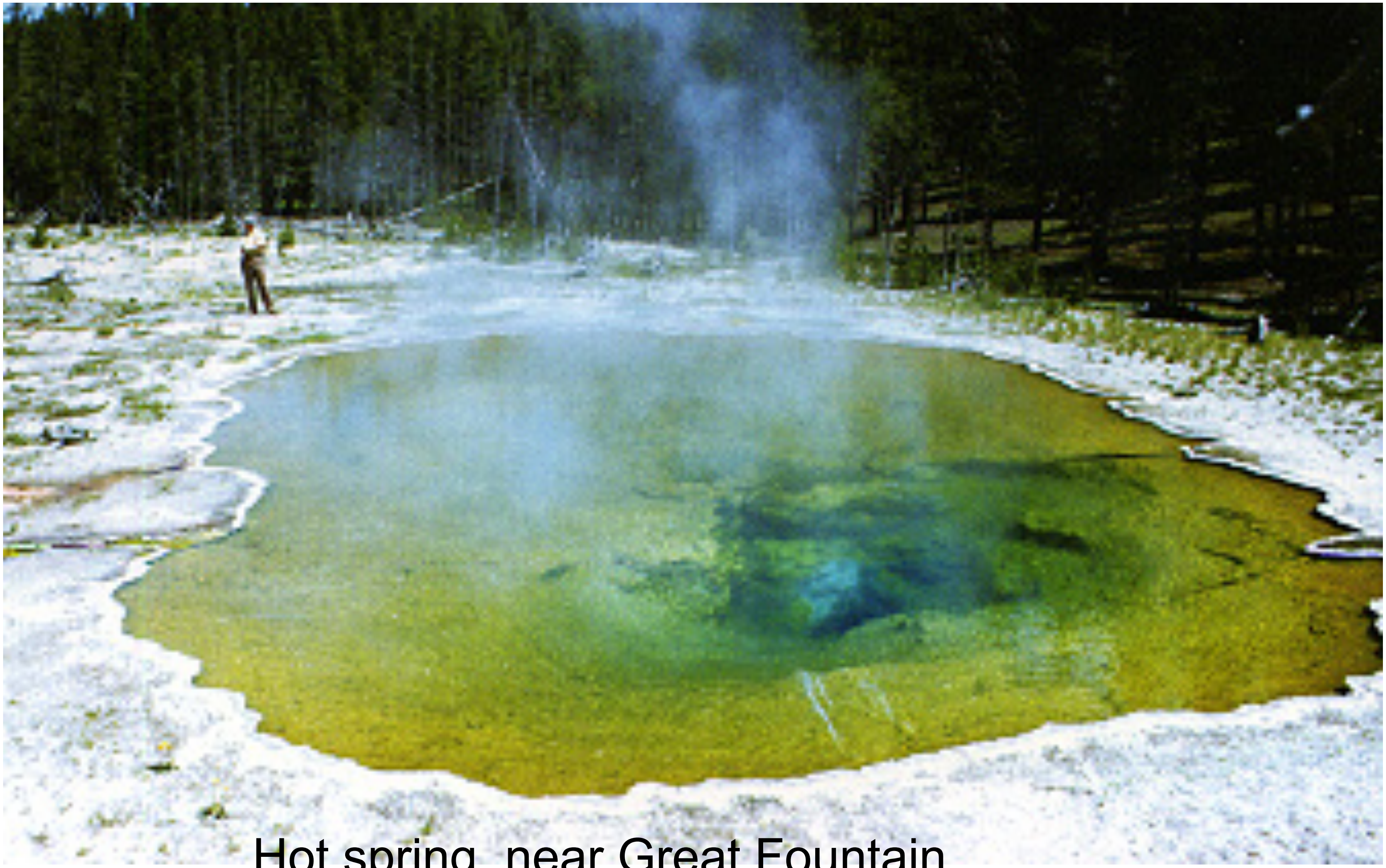
2: 95ºC

3: 60ºC

4: 72ºC

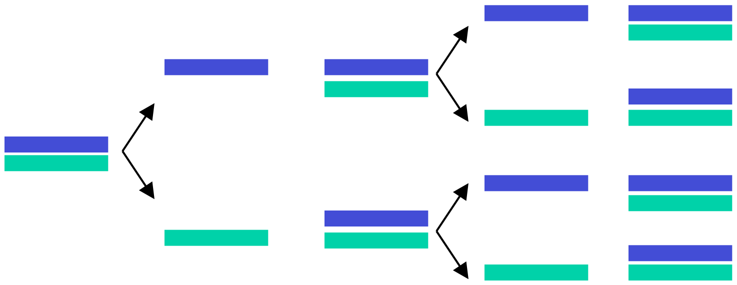
5: 72ºC

6: 72ºC



Hot spring, near Great Fountain Geyser, Yellowstone National Park

PCR



Ingredients:

- many copies of deoxy nucleotide triphosphates
- many copies of two primer sequences (~20 nt each)
  - readily synthesized
- many copies of Taq polymerase (*Thermus aquaticus*),
  - readily available commercialy
- as little as 1 strand of template DNA
- a programmable “thermal cycler”

Amplification: million to billion fold

Range: up to 2k bp routinely; 50k with other enzymes & care

Very widely used; forensics, archeology, cloning, sequencing, …

DNA Forensics

E.g. FBI “CODIS” (combined DNA indexing system) data base

picked 13 short, variable regions of human genome

amplify each from, e.g., small spot of dried blood

measure product lengths (next slides)

PCR is important for all the reasons that filters and amplifiers are important in electronics, e.g., sample size is reduced from grams of tissue to a few cells, can pull out small signal amidst “noisy” background

Gel Electrophoresis

DNA/RNA backbone is negatively charged (they’re acids)

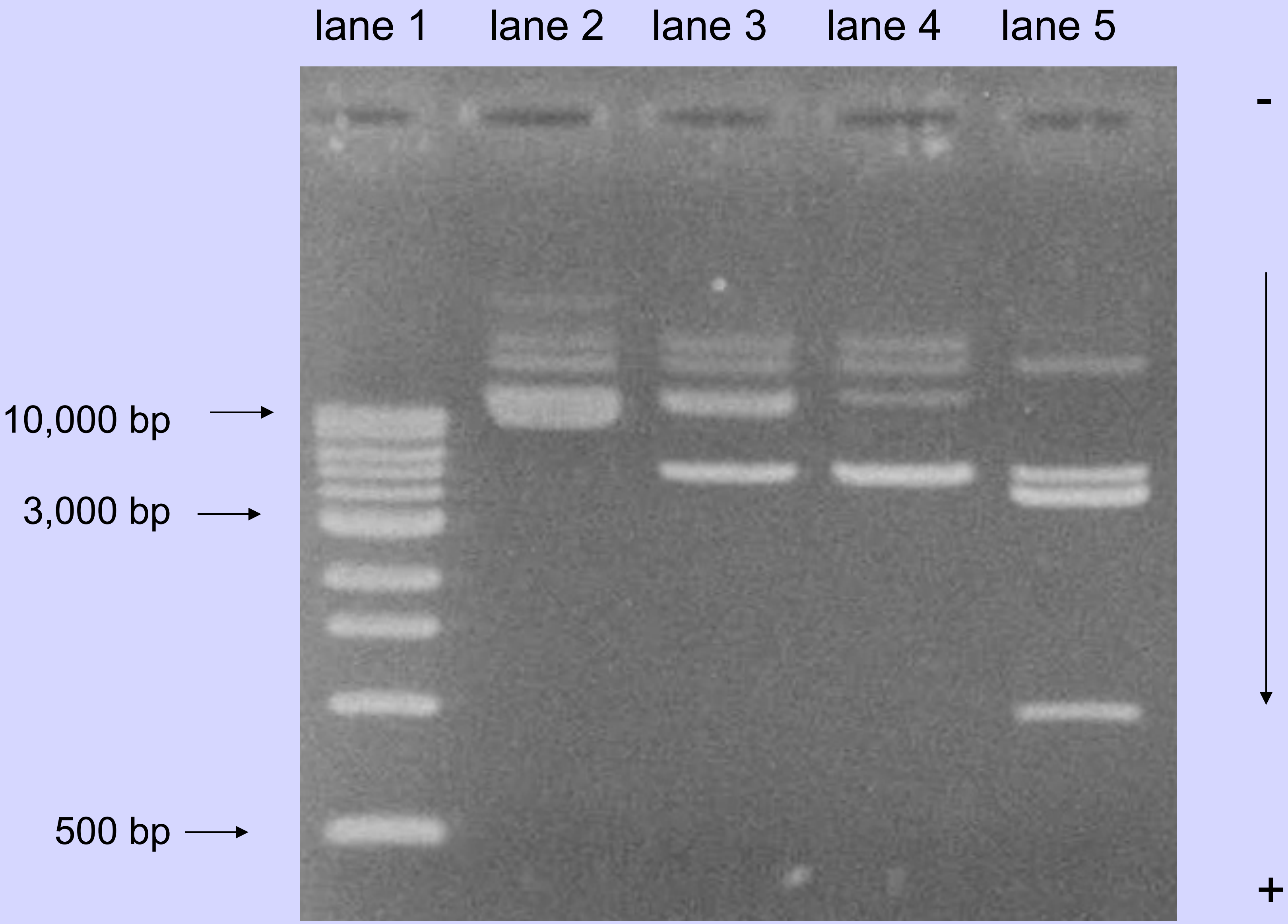
Molecules moves slowly in gels under an electric field

- agarose gels for large molecules
- polyacrylamide gels for smaller ones

Smaller molecules move faster

So, you can *separate DNAs & RNAs by size*

Nobel Chem prize, 1948 Arne Wilhelm Kaurin Tiselius



lane 1 lane 2 lane 3 lane 4 lane 5

-

10,000 bp

3,000 bp

500 bp

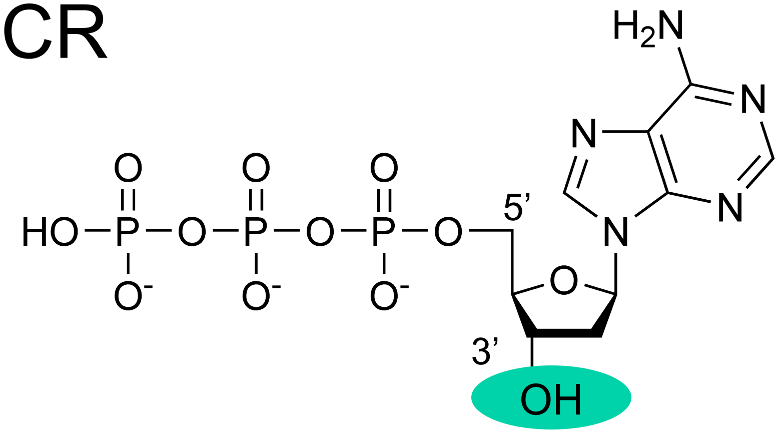
+

DNA Sequencing

Like one-cycle, one-primer PCR

Suppose 0.1% of A’s:

- are *di*-deoxy adenosine’s; backbone can’t extend
- carry a green florescent dye

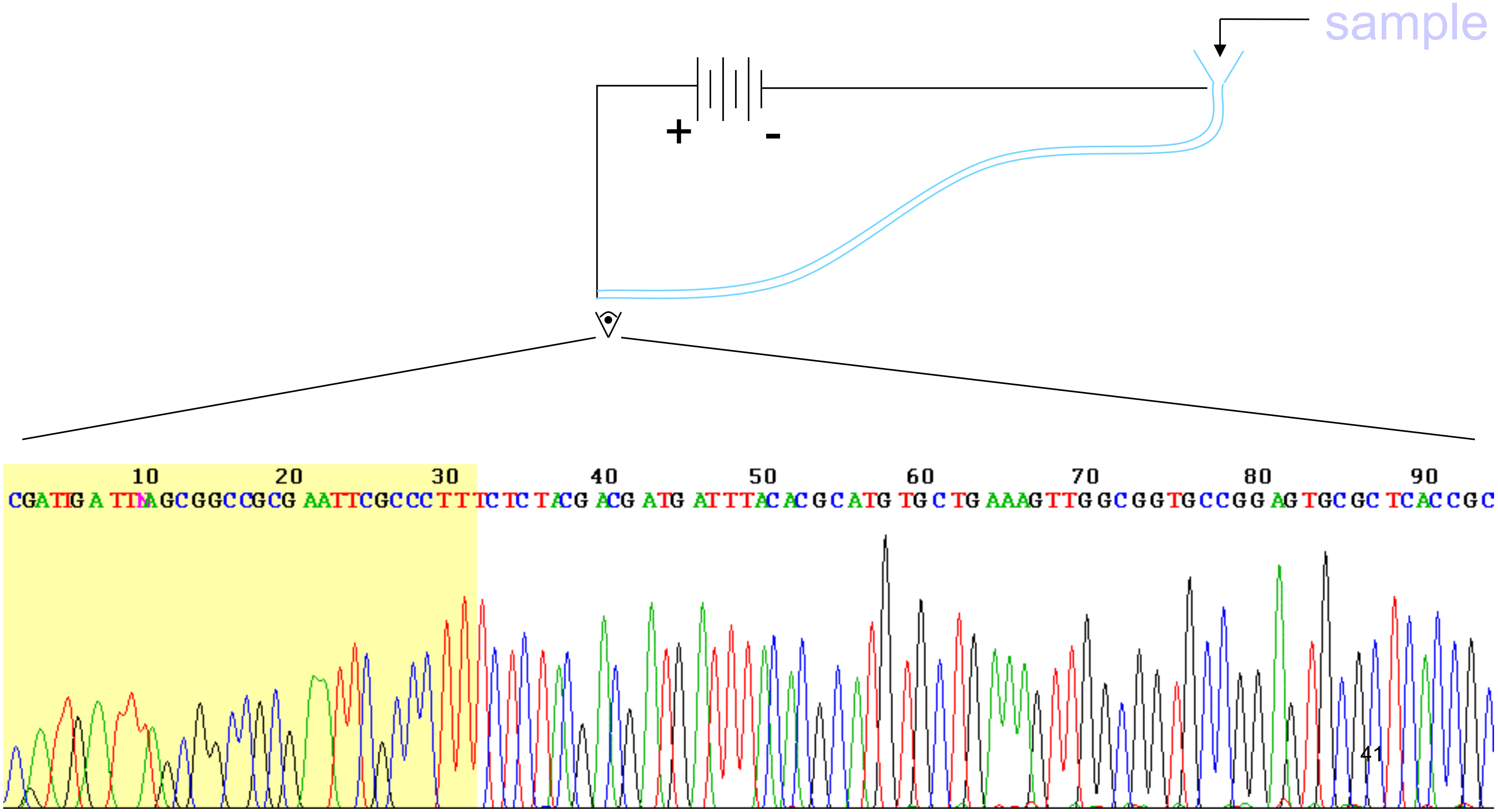


Separate by capillary gel electrophoresis

If frags of length 42, 49, 50, 55 … glow green, those positions are A’s

Ditto C’s (blue), G’s (yellow), T’s (red)

DNA Sequencing


sample
+
-
10
20
30
40
50
60
70
80
90
CGATTGATTNAGCGGCCGCGAATTCGCCCTTTTCTCTACGACGATGATTTACACGCATGTGCTGAAAGTTGGCGGTGCCGGAGTGCGCTCACCGC

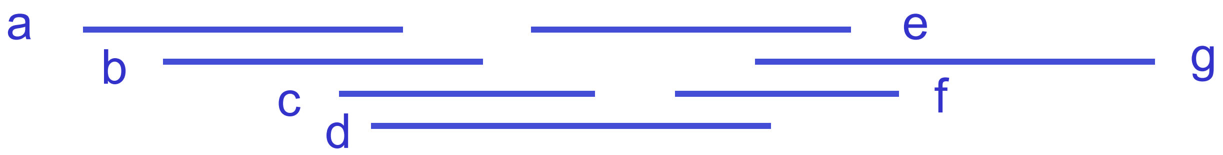
DNA Sequencing

Highly automated

Typically can “read” about 600 nt in one run

“Whole Genome Shotgun” approach:

- cut genome randomly into ~ G / 600 x 10 fragments
- sequence each
- reassemble by computer



Complications: repeated region, missed regions, sequencing errors, chimeric DNA fragments, …

But overall accuracy ~10^{-4}, if careful

“Next Generation” Sequencing

40 million microscopic PCR “colonies” on 1x2” slide

“read” ~50 bp of sequence from end of each

Automated

takes 2-3 days

costs a few thousand dollars

generates ~ a few terabytes of data (mostly images)

that’s ~ 1x human genome (but you need 5x-50x to assemble)

Other approaches: long reads, single molecules,…
Technology is changing rapidly!

Personal Genomes

2001: ~$2.7 billion (Human Genome Project)

2003: ~$300 million

2007: ~$1 million

2008: ~$60 thousand

2009: ~$4400 *bioinformatics not included…*

Summary

PCR allows simple *in vitro* amplification of minute quantities of DNA (having pre-specified boundaries)

Sanger sequencing uses

- a PCR-like setup with modified chemistry to generate varying length prefixes of a DNA template with the last nucleotide of each color-coded
- gel electrophoresis to separate DNA by size, giving sequence

Sequencing random overlapping fragments allows genome sequencing

“Next Gen” sequencing: throughput up, cost down (lots!)