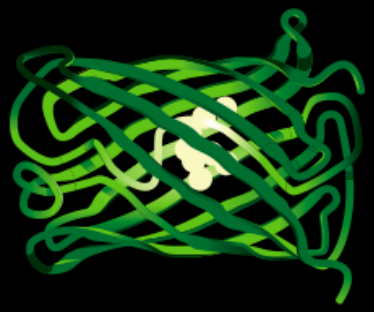
2008 Nobel Prize in Chemistry: GFP

Osamu Shimomura (Woods Hole, & Boston U)
- GFP from *Aequorea victoria*

Martin Chalfie (Columbia)
- used as a biomarker

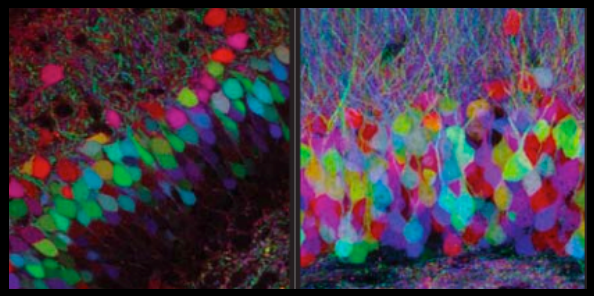
Roger Y. Tsien (UCSD)
- GFP photochemistry & new colors

Shimomura “never interested in applications" – just wanted to figure out how they glowed



Green fluorescent protein (GFP) consists of 238 amino acids. This chain folds up into the shape of a beer can. Inside the beer can structure the amino acids 65, 66 and 67 form the chemical group that absorbs UV and blue light, and fluoresces green.

Livet et al (2007) Nature 450, 56-63



CSEP 590A
Computational Biology
Autumn 2008

Lecture 3:
BLAST
Alignment score significance
PCR and DNA sequencing

8

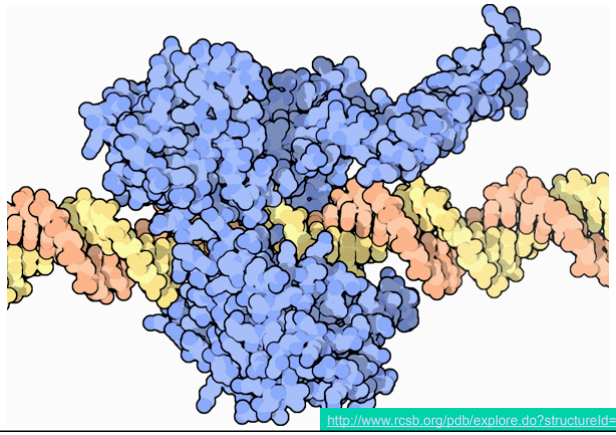
Tonight's plan

BLAST
Scoring
Weekly Bio Interlude: PCR & Sequencing

9

A Protein Structure:
(Dihydrofolate Reductase)

Topoisomerase I



http://www.rcsb.org/pdb/explore.do?structureId=1a36

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

13

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

14

BLAST: How

Idea: only parts of data base worth examining are those near a good 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

15

BLAST: Example

query → deadly

≥ 7 (thresh$_1$)

```
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
```

DB → ddgearlyk . . .

hits → ddge 10

early 18

≥ 10 (thresh$_2$)

16

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

18

Significance of Alignments

Is "42" a good score?

Compared to what?

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

19

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

20

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).

21

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 that 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*.

Usual scientific convention is to reject null only if p-value is < 0.05; sometimes demand p << 0.05

Can analytically find p-value for simple problems like coins; often turn to simulation/permutation tests for more complex situations; as below

22

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.*

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

23

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)
- subsetted by similarity, e.g. BLOSUM62 => 62% identity
- e.g. http://blocks.fhcrc.org/blocks-bin/getblock.pl?IPB013598

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

24

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

25

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**

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 }?

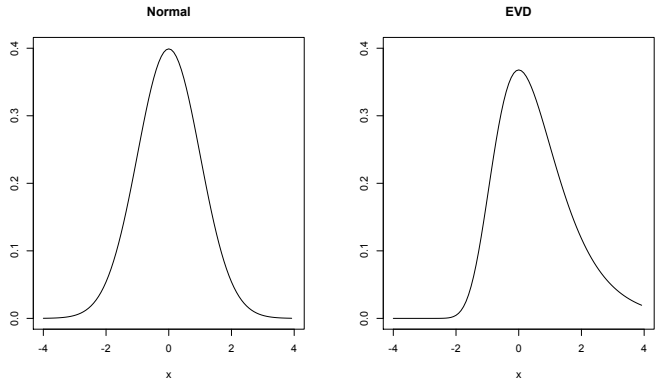
A. it's approximately an *Extreme Value Distribution (EVD)*

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

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)

28



29

EVD Pro/Con

Pro:

- gives p-values for alignment scores

Con:

- It's only approximate
- parameter estimation
- theory may not apply. E.g., it is NOT known to hold for gapped alignments (although empirically it seems to work pretty well).

30

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"

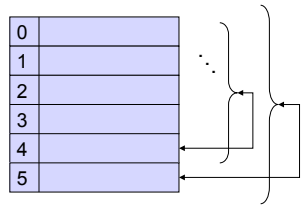
How to generate "random" sequences?

- Alignment scores 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

31

Generating Random Permutations

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



32

Permutation Pro/Con

Pro:

- Gives empirical p-values for alignments with characteristics like sequence of interest, e.g. residue frequencies

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

33

Issues

What if the model is wrong?

E.g., are adjacent positions really independent?

36

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 <--> P = .993
- E = 10 <--> P = .99995
- E = .01 <--> 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

35

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 seems ok for gapped alignments, too
- permutation tests are a simple (but brute force) alternative

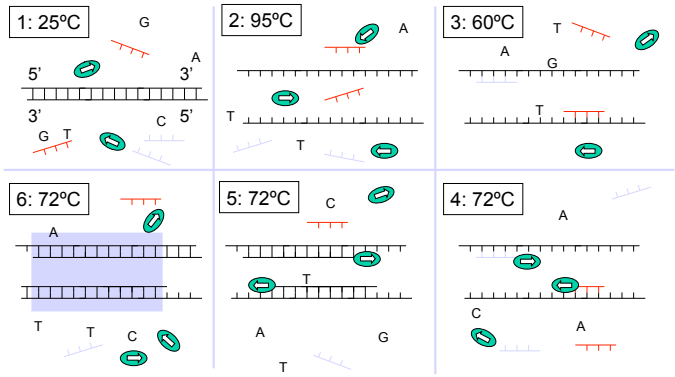
37

Weekly Bio(tech) Interlude

3 Nobel Prizes:
PCR: Kary Mullis, 1993
Electrophoresis: A.W.K. Tiselius, 1948
DNA Sequencing: Frederick Sanger, 1980

38

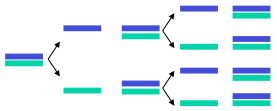
PCR



Hot spring, near Great Fountain
Geyser, Yellowstone National Park

40

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, …

41

DNA Forensics

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

pick 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 amplifiers are important in electronics, e.g., sample size is reduced from grams of tissue to a few cells

42

Gel Electrophoresis

DNA/RNA backbone is negatively charged

Molecules moves slowly in gels under an electric field

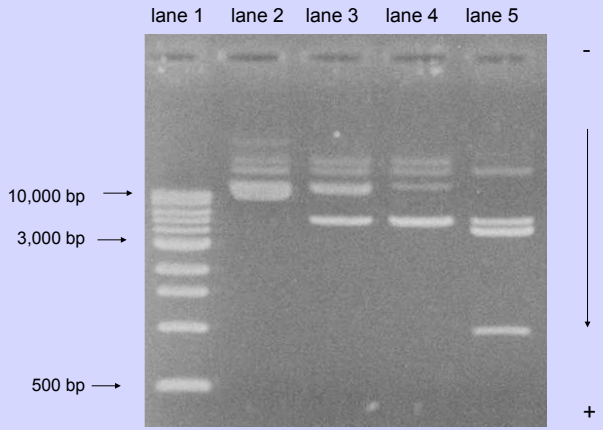
- agarose gels for large molecules
- polyacrylamide gels for smaller ones

Smaller molecules move faster

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

Nobel Chem prize, 1948 Arne Wilhelm Kaurin Tiselius

43

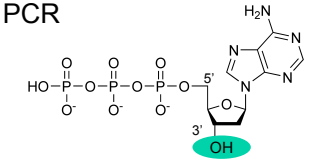


DNA Sequencing

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



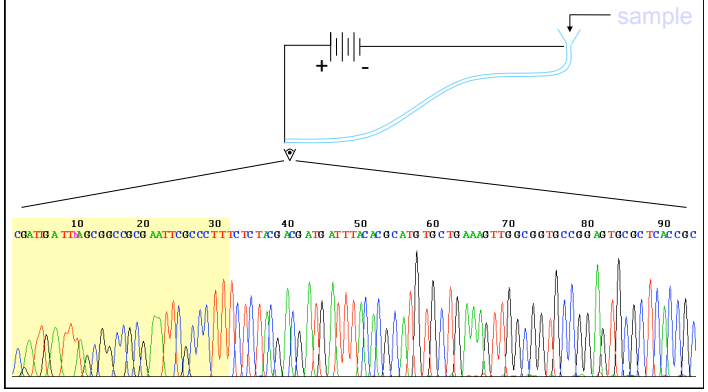
Separate by capillary gel electrophoresis

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)

45

DNA Sequencing



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

47

"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 ~ terabyte of data (mostly images)

that's ~ ½ of a human genome

other approaches: long reads, single molecules

48

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 (a lot)

49