CSEP 527
Computational Biology
Spring 2016

3: BLAST, Alignment score significance;
PCR and DNA sequencing
Outline

Scoring
BLAST
Weekly Bio Interlude: PCR & Sequencing
Significance of alignment scores

http://dericbownds.net/uploaded_images/god_face2.jpg
Significance of Alignments

Is “42” a good score?

*Compared to what?*

Usual approach: compared to a specific “null model”, such as “random sequences”
Brief Review of Probability
**random variables**

Discrete random variable: takes values in a finite or countable set, e.g.
- \( X \in \{1, 2, \ldots, 6\} \) with equal probability
- \( X \) is a positive integer \( i \) with probability \( 2^{-i} \)

Continuous random variable: takes values in an uncountable set, e.g.
- \( X \) is the weight of a random person (a real number)
- \( X \) is a randomly selected point inside a unit square
- \( X \) is the waiting time until the next packet arrives at the server
pdf and cdf

f(x) : the *probability density function* (or simply “density”)

\[ F(a) = \int_{-\infty}^{a} f(x) \, dx \]

\( P(X < a) = F(x) \): the *cumulative distribution function*

\( P(a < X < b) = F(b) - F(a) \)

Need \( f(x) \geq 0 \), \( \int_{-\infty}^{+\infty} f(x) \, dx = F(+\infty) = 1 \)

A key relationship:

\[ f(x) = \frac{d}{dx} F(x) \], since \( F(a) = \int_{-\infty}^{a} f(x) \, dx \),
Densities are *not* probabilities; e.g. may be > 1

\[ P(x = a) = 0 \]

\[ P(a - \varepsilon/2 \leq X \leq a + \varepsilon/2) = F(a + \varepsilon/2) - F(a - \varepsilon/2) \approx \varepsilon \cdot f(a) \]

I.e., the probability that a continuous random variable falls *at* a specified point is zero.

The probability that it falls *near* that point is proportional to the density; in a large random sample, expect more samples where density is higher (hence the name “density”).
X is a normal (aka Gaussian) random variable \( X \sim N(\mu, \sigma^2) \)

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

\[
E[X] = \mu \quad \text{Var}[X] = \sigma^2
\]

The Standard Normal Density Function
changing $\mu, \sigma$

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

density at $\mu$ is $\approx .399/\sigma$
Z-scores

\[ Z = \frac{(X-\mu)}{\sigma} = \frac{(X - \text{mean})}{\text{standard deviation}} \]

e.g.

\[ Z = +3 \text{ means “}3 \text{ standard deviations above the mean”} \]

Applicable to any distribution, and gives a rough sense of how usual/unusual the datum is.

If \( X \) is normal\((\mu, \sigma^2)\) then \( Z \) is normal\((0,1)\), and you can easily calculate (or look up in a table) just how unusual E.g., if normal, \( P(Z\text{-score} \geq +3) \approx 0.001 \)
Central Limit Theorem

If a random variable $X$ is the sum of many independent random variables, then $X$ will be approximately normally distributed.
Hypothesis Tests and P-values
Hypothesis Tests

Competing models might explain some data
E.g., you’ve flipped a coin 5 times, seeing HHHTH

Model 0 (The “null” model): \( P(H) = 1/2 \)
Model 1 (The “alternate” model): \( P(H) = 2/3 \)

Which is right?
A possible decision rule: reject the null if you see 4 or more heads in 5 tries
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 $p \ll 0.05$ (esp. if estimates are inaccurate, and/or big data).
p-values: controversial

p-values are commonly misused/misinterpreted
Most importantly, it is not the probability that the null is true, nor the 1 minus he prob that the alternate is true

Nevertheless, p-values are very widely used
Many resources, e.g.:

• https://en.wikipedia.org/wiki/P-value
• http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values
• http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html
Alignment Scores
Overall Alignment Significance, I
Empirical p-values (via randomization)

You just searched with x, found “good” score for x:y
Generate N random “y-like” 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 your 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 y-like” seqs? Scores depend on:
    Length, so use same length as y
    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 un-representative
  - 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
Theoretical Distribution of Alignment Scores?

A straw man: suppose I want a simple null model for alignment scores of, say MyoD versus random proteins of similar lengths. Consider this: Write letters of MyoD in one row; make a random alignment by filling 2nd row with random permutation of the other sequence plus gaps.

MELLSPPLR...
uv---wxyz...

Score for column 1 is a random number from the M row of BLOSUM 62 table, column 2 is random from E row, etc.

By central limit theorem, total score would be approximately normal
Histogram for scores of 20k Smith-Waterman alignments of MyoD vs permuted versions of C. elegans Lin32.

Looks roughly normal!

And real Lin32 scores well above highest permuted seq.
And, we can try to estimate p-value: from mean/variance of the data, true Lin32 has z-score = 7.9, corresponding p-value is $1.4 \times 10^{-15}$.

**But something is fishy:**

a) Histogram is skewed w.r.t. blue curve, and, especially,

b) Is *above* it in right tail (e.g. 111 scores $\geq 80$, when only 27 expected; highest permuted score is $z=5.7$, $p = 6 \times 10^{-9}$, very unlikely in only 20k samples)
Rethinking score distribution

Strawman above is ok: random permutation of letters & gaps *should* give normally distributed scores.

But S-W doesn’t stop there; *it then slides the gaps around so as to maximize score, in effect taking the maximum over a huge number of alignments with same sequence but different gap placements, and furthermore trims ends to find the max local score.*
Overall Alignment Significance, II

A Theoretical Approach: EVD

Let $X_i, 1 \leq i \leq N$, 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 (central limit theorem)

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:]

\[
P(y \leq z) \approx \exp(-KNe^{-\lambda(z-\mu)}) \quad (\ast)
\]

For ungapped local alignment of seqs $x, y$, $N \sim |x|^*|y|$
\(\lambda, K\) depend on score table, and can be estimated by curve-fitting random scores to (\(\ast\)), even with gaps. (cf. reading)
Both mean 0, variance 1; EVD skewed & has “fat right tail”
Red curve is approx fit of EVD to score histogram – fit looks better, esp. in tail. Max permuted score has probability $\sim 10^{-4}$, about what you’d expect in $2 \times 10^4$ trials.

True score is still moderately unlikely, $< \text{one tenth the above.}$
EVD Pro/Con

Pro:
  Gives p-values for alignment scores

Con:
  It’s only approximate
  You must estimate parameters
  Theory may not apply. E.g., known to hold for ungapped local alignments (like BLAST seeds). It is NOT proven to hold for gapped alignments, although there is strong empirical support.
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 (below)

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 and broadly applicable (but brute force) alternative
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: demo

E.g.

http://expasy.org/sprot
(or http://www.ncbi.nlm.nih.gov/blast/)
look up MyoD
go to blast tab
paste in ID or seq for human MyoD
set params (gapped=yes, blosum62,…)
get top 100 (or 1000) hits
**BLAST: How**

*Idea: most interesting parts of the DB have 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, $\sim$50) “similar” words $v_{ij}$ with score $\sigma(w_i, v_{ij}) > \text{thresh}_1$ (say, 1, 2, … letters different)

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

Grow each such “seed match” bidirectionally

Report those scoring $> \text{thresh}_2$, calculate E-values
BLAST: Example

query -> deadly

\[ \begin{align*}
\text{de} & \quad (11) \rightarrow \quad \text{de} \quad \text{ee} \quad \text{dd} \quad \text{dq} \quad \text{dk} \\
\text{ea} & \quad (9) \rightarrow \quad \text{ea} \\
\text{ad} & \quad (10) \rightarrow \quad \text{ad} \quad \text{sd} \\
\text{dl} & \quad (10) \rightarrow \quad \text{dl} \quad \text{di} \quad \text{dm} \quad \text{dv} \\
\text{ly} & \quad (11) \rightarrow \quad \text{ly} \quad \text{my} \quad \text{iy} \quad \text{vy} \quad \text{fy} \quad \text{lf}
\end{align*} \]

DB -> ddgearlyk ...

hits

\[ \begin{align*}
\text{ddge} & \quad 10 \\
\text{early} & \quad 18
\end{align*} \]

\[ \geq 10 \quad \text{(thresh}_1) \]

\[ \geq 7 \quad \text{(thresh}_1) \]

\[ \geq 10 \quad \text{(thresh}_2) \]
### BLOSUM 62 (the “σ” scores)

|     | 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  |     | 0   | -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  | 1   | -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” (next week?) pattern from initial pass to find weaker matches in subsequent passes

Many others
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, it's ok in other contexts, too, e.g., for gapped alignments.
- Permutation tests are a simple (but brute force) alternative.
Bio(tech) Interlude

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

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.

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

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

As of 1/2013, over 10,142,600 offender profiles

Picked 13 “short tandem repeats”, i.e., variable-length regions of human genome flanked by (essentially) invariant sequences (primer targets), several alleles common at each locus, of which you have 2

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

Measure product lengths (next slides)

http://www.fbi.gov/about-us/lab/biometric-analysis/codis
http://www.dna.gov/solving-crimes/cold-cases/howdatabasesaid/codis/
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
DNA Sequencing – Sanger Method

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
Sanger with capillary electrophoresis
Sequencing A Genome

Highly automated
Typical Sanger “read” about 600 nt
“Whole Genome Shotgun” approach:
  randomly fragment (many copies of) genome
  sequence many, enough to cover each base 10x or more times
  reassemble by computer

Complications: repeated region, missed regions, sequencing errors, chimeric DNA fragments, …
But overall accuracy ~10^{-4}, if careful

E.g., human genome project:
≈ 30Gbases and
≈ 3\times10^9/600\times10
= 5\times10^7 reads
“Next Generation” Sequencing

Many technical improvements to Sanger approach over many years, culminating in highly automated machines used for the HGP

Since then, many innovative new ideas/products:

• Helicos: single molecule fluorescence tethered to flow cell
• Illumina: colony PCR; reversible dye terminator
• Ion Torrent: semiconductor detection of ions released by polymerase
• Roche 454: emulsion PCR; pyro sequencing
• Oxford Nanopore
• Pacific Biosciences: single tethered polymerases in “zero mode waveguide” nano-wells, circularized DNA, “real time”
• ABI SOLiD: emulsion PCR, sequence by ligation, “color-space”
• Complete Genomics: rolling circle replication/DNA nanoballs

Technology is changing rapidly!
“Next Generation” Sequencing

~1 billion microscopic PCR “colonies” on 1x2” slide
“Read” ~50-150bp of sequence from (1 or 2) ends of each
Ends fluorescently labeled, blocked, chemically cycled
Automated: takes a few days; ~ 100 G bases/day
Costs a few thousand dollars
Generates terabytes of data (mostly images)
I,e., ~ 30x human genome/day (you need 25x-50x to assemble)

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

~1 billion microscopic PCR “colonies” on 1x2” slide
“Read” ~50-150bp of sequence from (1 or 2) ends of each
Reversible dye terminators
Automated: takes a few days; ~ 100 G bases/day
Costs a few thousand dollars
Generates terabytes of data (mostly images)
I,e., ~ 30x human genome/day
(you need 25x-50x to assemble)
(equal to all of pre-2008 Genbank)
In contrast with other platforms, therefore, the sequencing by synthesis is asynchronous in that some features may get represented by individual beads. Like the HeliScope (discussed below), the sequencing is 'asynchronous' in that some features may get represented by individual beads. The pattern of detected incorporation events reveals the sequence of template molecules present. In productive emulsion compartments (where both a bead and template molecule are present), PCR amplicons are captured to the surface of the bead. After breaking the emulsion, PCR amplicons are amplified, but both primers densely coat the surface of a solid substrate, attached at their 5'-termini by a flexible linker. As a consequence, amplification products originating from any given member of the template library remain locally tethered near the point of origin. At the conclusion of the PCR, amplification products from each clonal cluster are mixed together and the template library is used as a template for subsequent rounds of amplification. Clonal sequencing features. In brief, an in vitro–constructed adaptor-flanked shotgun library is PCR amplified to distinguishable locations on an array using primer pairs that rely on emulsion PCR (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an in vitro–constructed adaptor-flanked shotgun library is PCR amplified to distinguishable locations on an array using primer pairs that rely on emulsion PCR (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an in vitro–constructed adaptor-flanked shotgun library is PCR amplified to distinguishable locations on an array using primer pairs that rely on emulsion PCR (aka 'cluster PCR') to amplify clonal sequencing features. In brief, an in vitro–constructed adaptor-flanked shotgun library is PCR amplified to distinguishable locations on an array using primer pairs that rely on emulsion PCR (aka 'cluster PCR') to amplify clonal sequencing features.
## Illumina HiSeq (1500/2500, as of Spring 2013)

**HIGH OUTPUT RUN MODE***

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Dual Flow Cell (2500 only)</th>
<th>Single Flow Cell (1500 or 2500)</th>
<th>Dual Flow Cell Run Time</th>
<th>Dual Flow Cell Run Time (2500 only)</th>
<th>Single Flow Cell Run Time (1500 or 2500)</th>
<th>Dual Flow Cell Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 36</td>
<td>95-105 Gb</td>
<td>47-52 Gb</td>
<td>2 days</td>
<td>18-22 Gb</td>
<td>9-11 Gb</td>
<td>7 hr</td>
</tr>
<tr>
<td>2 x 50</td>
<td>270-300 Gb</td>
<td>135-150 Gb</td>
<td>5.5 days</td>
<td>50-60 Gb</td>
<td>25-30 Gb</td>
<td>16 hr</td>
</tr>
<tr>
<td>2 x 100</td>
<td>540-600 Gb</td>
<td>270-300 Gb</td>
<td>11 days</td>
<td>100-120 Gb</td>
<td>50-60 Gb</td>
<td>27 hr</td>
</tr>
<tr>
<td>2 x 150</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>150-180 Gb</td>
<td>75-90 Gb</td>
<td>40 hr</td>
</tr>
</tbody>
</table>

**RAPID RUN MODE***

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Dual Flow Cell (2500 only)</th>
<th>Single Flow Cell (1500 or 2500)</th>
<th>Dual Flow Cell Run Time</th>
<th>Dual Flow Cell Run Time (2500 only)</th>
<th>Single Flow Cell Run Time (1500 or 2500)</th>
<th>Dual Flow Cell Run Time</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Reads Passing Filter</th>
<th>Up to 3 billion single reads or 6 billion paired-end reads</th>
<th>Up to 1.5 billion single reads or 3 billion paired-end reads</th>
<th>Up to 600 million single reads or 1.2 billion paired-end reads</th>
<th>Up to 300 million single reads or 600 million paired-end reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality</td>
<td>&gt; 85% of bases above Q30 at 2 x 50 bp</td>
<td>&gt; 80% of bases above Q30 at 2 x 100 bp</td>
<td>&gt; 85% of bases above Q30 at 2 x 50 bp</td>
<td>&gt; 80% of bases above Q30 at 2 x 100 bp</td>
</tr>
</tbody>
</table>

*Install specifications based on Illumina PhiX control library at supported cluster densities (between 610-678 K clusters/mm2 passing filter using TruSeq v3 Kits or 700-820 clusters/mm2 passing filter using TruSeq Rapid Kits. Run times for rapid run mode correspond to on-board cluster generation (1.5 hr) and sequencing; for high output mode, run times correspond to sequencing only. Performance may vary based on sample quality, cluster density, and other experimental factors. Early HiSeq 2000 instruments will run slightly slower when upgraded to a HiSeq 2500.*

Modern DNA Sequencing

A table-top box the size of your oven (but costs a bit more … ;-) can generate ~100 billion BP of DNA seq/day; i.e.
= 2008 genbank,
= 30x your genome
Pacific Biosciences

SMRT® Technology

The PacBio® RS II is based on novel Single-Molecule, Real-Time (SMRT) technology which enables the observation of natural DNA synthesis by a DNA polymerase in real time. Sequencing occurs on SMRT Cells, each containing thousands of Zero-Mode Waveguides (ZMWs) in which polymerases are immobilized. The ZMWs provide a window for watching the DNA polymerase as it performs sequencing by synthesis.

SMRT® Cells

- Phospholinked nucleotides
- Zero-Mode Waveguides

Library Preparation Instrument Run Data Analysis

- DNA Template Prep Kit
- DNA Polymerase Binding Kit
- MagBead Kit
- PacBio RS II with touch screen
- RS Remote for run design
- SMRT Cells
- DNA Sequencing Kit
- SMRT Analysis
- SMRT Portal
- SMRT View

Sequencing time

30 to 120 min per SMRT Cell

Results in as few as 10 hours

No amplification required

- Open source,
- open standards
Advantages:
- single molecules
- long reads
- direct CH$_3$ detection

Disadvantages:
- throughput
- error rate; (circularize?)

**Read Length:**
- Average: 4,606 bp
- 95$^{th}$ Percentile: 11,792 bp
- Maximum: 23,297 bp

**Throughput per SMRT$^\text{®} $ Cell:**
- 216 Mb
  - 47,197 reads

Based on data from 11 kb plasmid library using a 120 minute movie

http://www.pacificbiosciences.com/img/assets/smrt_sequencing_advantage_readlength_lg.png
Oxford Nanopore

Prerelease claims ≈ 100k read lengths, 150Mb in 6 hrs, $1000
Personal Genomes

2001: ~$2.7 billion (Human Genome Project)
2003: ~$300 million
2007: ~$1 million
2008: ~$60 thousand
2009: ~$4400
2015: ~$1000

*bioinformatics not included…*
Figure 3: Illumina Sequencing Technology Outpaces Moore’s Law for the Price of Whole Human Genome Sequencing
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 (and many other applications)

“Next Gen” sequencing: many innovations
- throughput up, cost down (lots!)
More on p-values and hypothesis testing
P-values & E-values

p-value:  $P(s,n) = \text{probability}$ of a score more extreme than $s$ in a random target data base of size $n$

E-value: $E(s,n) = \text{expected number}$ of such matches

They Are Related:

- $E(s,n) = p n$ (where $p = P(s,1)$)
- $P(s,n) = 1 - (1 - p)^n = 1 - (1 - 1/(1/p))^{(1/p)(pn)} \approx 1 - \exp(-pn) = 1 - \exp(-E(s,n))$

E big (say, $\gg 1$) $\iff$ P big ($\to 1$)
  
  - $E = 5 \iff P \approx 0.993$
  - $E = 10 \iff P \approx 0.99995$

E small $\iff$ P small (both near 0)
  
  - $E = 0.01 \iff P \approx E - E^2/2 + E^3/3! \ldots \approx E$

Both equally valid; E-value is perhaps more intuitively interpretable
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 \(D = \text{HHHTH}\)

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

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

Likelihoods:

\[
P(D \mid M_0) = (1/2) \times (1/2) \times (1/2) \times (1/2) \times (1/2) = 1/32
\]
\[
P(D \mid M_1) = (2/3) \times (2/3) \times (2/3) \times (1/3) \times (2/3) = 16/243
\]

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., given data is \(\approx 2.1\)x more likely under alt model than null model
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).
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 $\Rightarrow$ $\geq$ 62% identity
- e.g. http://blocks.fhcrc.org/bl...
Scores: formula above, rounded

|     | 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  | -3  | -2  | -2  | -3  | -1  | -1  |
| R   | -1  | 5   | 0   | -2  | -3  | 0   | 0   | 0   | 0   | -3  | -2  | -1  | -3  | -2  | -1  | -1  | -3  | -2  | -2  | -3  | -1  | -1  |
| N   | -2  | 0   | 6   | 1   | -3  | 0   | 0   | 0   | 0   | 1   | -3  | -2  | -3  | -2  | -1  | 0   | -4  | -2  | -3  | -3  | -3  | -3  |
| D   | -2  | -2  | 1   | 6   | -3  | 0   | 2   | -1  | -1  | -1  | -3  | -3  | -1  | -1  | -3  | -1  | -1  | -2  | -2  | -2  | -2  | -3  |
| C   | 0   | -3  | -3  | -3  | 9   | -3  | -4  | -3  | -3  | -1  | -1  | -1  | -2  | -3  | -1  | -1  | -2  | -2  | -2  | -1  | -1  | -1  |
| Q   | -1  | 1   | 0   | 0   | -3  | 5   | 2   | -2  | 0   | -3  | -2  | 1   | 0   | -3  | -1  | 0   | -1  | -2  | -2  | -2  | -3  | -2  |
| E   | -1  | 0   | 0   | 2   | -4  | 2   | 5   | 2   | 0   | -3  | -3  | 1   | -2  | -3  | -1  | 0   | -1  | -3  | -2  | -2  | -2  | -2  |
| G   | 0   | -2  | 0   | -1  | -3  | -2  | -2  | 6   | -2  | -4  | -4  | -2  | -3  | -3  | -2  | 0   | -2  | -2  | -3  | -3  | -3  | -3  |
| H   | -2  | 0   | 1   | -1  | -3  | 0   | 0   | 2   | 8   | -3  | -3  | -1  | -2  | -1  | -2  | -1  | -2  | -2  | -2  | -2  | -2  | -3  |
| I   | -1  | -3  | -3  | -3  | -1  | -3  | -3  | -4  | -3  | 4   | 2   | 3   | 1   | 0   | -3  | -2  | -3  | -1  | -3  | -1  | -1  | -1  |
| L   | -1  | -2  | -3  | -4  | -1  | -2  | -3  | -4  | -3  | 2   | 4   | 2   | 2   | 0   | -3  | -2  | -1  | -2  | -1  | -2  | -2  | -1  |
| K   | -1  | 2   | 0   | -1  | -3  | 1   | 1   | 2   | -1  | -3  | -2  | 5   | -1  | -3  | -1  | -1  | -3  | -2  | -2  | -2  | -2  | -2  |
| M   | -1  | -1  | -2  | -3  | -1  | 0   | -2  | -3  | -2  | 1   | 2   | -1  | 5   | 0   | -2  | -1  | -1  | -1  | -1  | -1  | -1  | -1  |
| F   | -2  | -3  | -3  | -3  | -2  | -3  | -3  | -3  | -1  | 0   | 0   | 3   | 5   | 6   | -4  | 0   | -3  | 1   | -1  | -1  | -1  |
| P   | -1  | -2  | -2  | -1  | -3  | -1  | -1  | -2  | -2  | -3  | -3  | -1  | -2  | -2  | -4  | 7   | -1  | -1  | -4  | -3  | -2  | -2  |
| S   | 1   | -1  | 1   | 0   | -1  | 0   | 0   | 0   | -1  | -2  | -2  | 0   | -1  | -2  | -1  | 4   | 1   | -3  | -2  | -2  | -2  | -2  |
| T   | 0   | -1  | 0   | -1  | -1  | -1  | -1  | -1  | -2  | -2  | -1  | -1  | -1  | -1  | -1  | 1   | 1   | -3  | -2  | 0   | -2  | -2  |
| W   | -3  | -3  | -4  | -4  | -2  | -2  | -3  | -2  | -2  | -3  | -2  | -3  | -1  | 1   | 4   | -3  | -2  | 11  | 2   | -3  | -2  | -3  |
| Y   | -2  | -2  | -2  | -3  | -2  | -1  | -2  | -3  | 2   | -1  | -1  | -2  | -1  | 3   | -3  | -2  | -2  | 2   | 7   | -1  | -2  | -2  |
| V   | 0   | -3  | -3  | -3  | -1  | -2  | -2  | -3  | -3  | 3   | 1   | -2  | 1   | -1  | -2  | -2  | -3  | -1  | 4   | -3  | -1  | 4   |
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 tuned for ~ 95% sequence identity
- WU-BLAST: +5/-4 tuned for ~ 66% identity ("twilight zone")

** 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
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 (below)

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 and broadly applicable (but brute force) alternative

Looking at residue substitutions in a large set of “trusted” alignments provides a sound basis for defining the score tables