History of the Genetic Code
Administrative notes

• Final paper / project due by 5:00 pm on Friday, Dec. 7.

Remember:

• 30-minute meetings later this week, as per schedule.
• 15-minute in-class presentations next week, as per schedule.

• CSE Dept. seminar on "Electronic Voting: Dangers and Opportunities" by J. Alex Halderman, a Ph.D. student at Princeton, on Tuesday, Nov. 27th at 4:00 pm in PL 466. Reception starting at 3:30 pm in lobby of Packard Lab.

• CSE Dept. seminar on "Ben Franklin Racing Team and the DARPA Urban Challenge" by Prof. John Spletzer, on Thursday, Nov. 29th at 4:00 pm in PL 101. Reception starting at 3:30 pm in lobby of Packard Lab.

It's interesting to look back and see what (very smart) people were thinking in mid-1950's, just after double helix structure of DNA was unraveled but we still had no idea how it all worked.

These early ideas had a strong computer science "flavor."

To understand theories of the time, most of which sounded good but ultimately proved wrong, we must forget almost everything we know about molecular biology ...
### Genetic Code timeline #1

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1865</td>
<td>Gregor Mendel, working alone in Austrian monastery, discovers that some characteristics are inherited in ‘units’.</td>
</tr>
<tr>
<td>1870</td>
<td>Friedrich Miescher isolates chemicals from cell nucleus, including ‘nucleic acids’. However, most people are more interested in proteins in nucleus.</td>
</tr>
<tr>
<td>1879</td>
<td>Walter Flemming describes behavior of chromosomes during cell division, implicating these nuclear structures in inheritance.</td>
</tr>
<tr>
<td>1900</td>
<td>Hugo DeVries and others rediscover Mendel’s work and establish first laws of inheritance.</td>
</tr>
<tr>
<td>1909</td>
<td>Wilhelm Johannsen coins term ‘gene’.</td>
</tr>
<tr>
<td>1911</td>
<td>Thomas Hunt Morgan is first to show that genes are arranged in linear fashion along chromosomes.</td>
</tr>
</tbody>
</table>

Early work based on studying phenotypes. “Chromosome” is abstract concept – no one knows exactly what it is.

http://www.wellcome.ac.uk/en/fourplus/DNA_timeline.html
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1928</td>
<td>Frederick Griffith uses chemical extract to convert harmless pneumonia bacteria into pathogenic forms, but nature of ‘inheritance factor’ is unknown.</td>
</tr>
<tr>
<td>1929</td>
<td>Phoebus Levene discovers that a sugar, deoxyribose, is present in nucleic acids. Later identifies that DNA is made up of nucleotides, a chemical unit comprising a deoxyribose sugar, a phosphate group and one of four small organic molecules known as bases.</td>
</tr>
<tr>
<td>1941</td>
<td>George Beadle &amp; Edward Tatum show genes direct production of proteins.</td>
</tr>
<tr>
<td>1943</td>
<td>William Astbury makes first X-ray diffraction images of DNA.</td>
</tr>
<tr>
<td>1944</td>
<td>Building on Griffith’s work, Oswald Avery &amp; colleagues show that DNA can ‘transform’ cells, cementing link between DNA and genes.</td>
</tr>
<tr>
<td>1950</td>
<td>Edwin Chargaff discovers patterns in amounts of four bases in DNA: amounts of G and C, and of A and T, are always same.</td>
</tr>
<tr>
<td>1951</td>
<td>Rosalind Franklin takes her first X-ray diffraction pictures.</td>
</tr>
<tr>
<td>1953</td>
<td>James Watson &amp; Francis Crick publish first paper proposing double helix structure for DNA.</td>
</tr>
</tbody>
</table>

http://www.wellcome.ac.uk/en/fourplus/DNA_timeline.html
The years 1952-1953

- Sony sells a miniature transistor radio.
- Univac projects the winner of the presidential election on CBS.
- Telephone area codes.
- Acoustic suspension loudspeaker invented by Henry Kloss.
- RCA’s Bizmac has first computer database.
- Claude Shannon uses electric mouse and maze to prove computers can learn.
- Hemingway's The Old Man and the Sea, the lonely struggle of an old fisherman.
- John Steinbeck's novel East of Eden, a Biblical tale set in California.
- Grace Hopper develops the first computer compiler.

1952

- Sony sells a miniature transistor radio.
- Univac projects the winner of the presidential election on CBS.
- Telephone area codes.
- Acoustic suspension loudspeaker invented by Henry Kloss.
- RCA’s Bizmac has first computer database.
- Claude Shannon uses electric mouse and maze to prove computers can learn.
- Hemingway's The Old Man and the Sea, the lonely struggle of an old fisherman.
- John Steinbeck's novel East of Eden, a Biblical tale set in California.
- Grace Hopper develops the first computer compiler.

1953

- TV Guide; initial press run is 1.5 million copies.
- Magnetic core memory is installed in a computer, the Whirlwind.
- Nobel Prize in Literature: Winston Churchill.
- Arthur Miller's play, The Crucible, a parable of the McCarthy witch hunts.
- FCC adopts NTSC color standard developed by RCA; drops CBS standard.
- With the 701, IBM starts building commercial computers.
- Bill Haley records first rock hit, "Crazy Man Crazy".
- One American, two Russians figure out how to harness what will be the laser.
What was known in 1953?

- DNA composed of four nucleotides, A, C, G, T, forming double-stranded helix.
- A binds with T, C binds with G, hence, strands are reverse complements.
- DNA copies itself during cell division (replication).

... and ...

- Proteins composed of 20 amino acids.
- Protein production controlled by genes.
- DNA seems to be the genetic material.

But what is the connection???
What wasn't known in 1953?

- No DNA sequences (none had been sequenced yet).
- Fragmentary information about protein sequences (insulin).
- No concept of RNA (neither mRNA nor tRNA).
- No Genetic Code – mapping from a four symbol alphabet to a 20 symbol alphabet – or how it might be implemented.
Back in 1953

Pretend you're back in 1953 ...

- You know that DNA is a double helix made up of two strands, each over a four symbol alphabet.
- Likewise, proteins are sequences over a 20 symbol alphabet.
- You believe that DNA is the genetic material.
- What's the connection between a DNA molecule and the proteins it is purported to produce?

Just to make it interesting:
- You don't know any sequence for a real DNA molecule. Sequences for a few proteins are just becoming available.

What does this imply?
- Anything you propose will be an abstract theory awaiting later experimental validation. But that's okay ...
At about the same time, information theory was coming into vogue. Claude Shannon joined Bell Labs in 1941 and soon started working on a fundamental approach for expressing information in a quantitative way. The goal was to make information a measurable quantity, like density or mass.

The repercussions were felt throughout science. Now we could talk, in a formal way, about coding theory, i.e., efficient schemes for storing and transmitting information.

Surely nature is just as efficient as anything we could invent?

http://www.nyu.edu/pages/linguistics/courses/v610003/shan.html
Some preliminaries (should look familiar)

- DNA is sequence over a four symbol alphabet.
- Protein is sequence over a 20 symbol alphabet.

Already obvious, even without support of experimental data:

\[
\begin{align*}
4^1 &= 4 < 20 \quad \text{... nope, not enough} \\
4^2 &= 16 < 20 \quad \text{... nope, not enough} \\
4^3 &= 64 \geq 20 \quad \text{... looks good!}
\end{align*}
\]

- This means that codon length must be at least three nucleotides, assuming all codons same length.
- It doesn't mean codons can't be longer or shorter.
- It doesn't mean all codons must be same length.
George Gamow was an extremely famous physicist, one of the early proponents of the “Big Bang” theory in astrophysics.

Recall that the concept of RNA as a mediator between DNA and proteins was completely unknown.

In the absence of RNA, Gamow made the reasonable assumption that proteins form directly on a template created by the DNA double helix. The various combinations of nucleotides along the grooves create distinctive cavities to attract a specific amino acid.

http://www.gwu.edu/~physics/gwimageh.html
First theory: George Gamow's diamond code

Nucleotide bases are designated by numbers and the 20 codons by letters (remember, no experimental evidence yet).

Note: I think he got the “twist” backwards.
First theory: George Gamow's diamond code

strand 1

strand 2

cavity

complementary bases

codon

T C A T G

A G T A C
First theory: George Gamow's diamond code

While each codon has four bases, only three of these are independent – one pair must be complements (1 & 2 here).

Hence, this is a *triplet code*. As we saw before, that seems to be exactly what we need. But such a code defines 64 codons. What did Gamow do?

He exploited symmetries:

Say 3 = C, 1 = A, 4 = G. Then CAG = GAC = GTC = CTG.
First theory: George Gamow's diamond code

Fortuitously, this yields **20** codons!

\[ n = \text{some amino acid (we don't know which without experiments).} \]
Symmetries of the diamond code

Symmetries of diamond code sort 64 codons into 20 classes, indicated here by 20 colors.

All codons in each class specify same amino acid.

case we considered earlier
Overlapping codes

Consider successive codons:

Note that each nucleotide is used in three successive codons. Hence, not only is diamond code a triplet code, it's an *overlapping triplet code*.

At time, this seemed like a good idea:
• inter-nucleotide and inter-amino acid spacings similar,
• maximizes information storage density (recall Shannon),
• imposes constraints on possible protein sequences.

But eventually this last point used to rule out diamond code.
Not one to be easily disuaded (an attribute that is vital in a successful scientist), Gamow proposed another overlapping triplet code with an even simpler interpretation.

Each triple of nucleotides maps to the same amino acid regardless of the order in which the nucleotides appear.

Recall in diamond code we had $CAG = GAC = GTC = CTG$. Here we have $CAG = CGA = GAC = GCA = ACG = AGC$.

How many codons does this give us? **20** codons!

This is known as Gamow's *composition code*. 
An overlapping code packs 16 codons into 18 base-pairs by exploiting triplets in all three phases, or *reading frames*.

But, as noted earlier, this prohibits some protein sequences. Consider dipeptides (sequences two amino acids in length, which require four nucleotides to code for):

\[
20^2 = 400 \quad \text{possible amino acid sequences}
\]

\[
4^4 = 256 \quad \text{possible codons}
\]

So we should see at most 256 different dipeptides in nature. This was used to rule out all overlapping codes experimentally.
Overlapping codes

While overlapping codes were eventually eliminated from consideration, it was obvious from the start that they had one undesirable property: a single nucleotide mutation could affect up to three adjacent amino acids. This seems a bit dangerous.

Moreover, earlier experimental evidence showed signs that single amino acid mutations occurred in nature, providing an initial clue that overlapping codes weren't the right model.

The belief that nature would somehow try to optimize coding efficiency is, as we now know, a bit humorous, given the vast quantities of “junk” DNA that appear in our genome.

After overlapping codes had been conclusively ruled out, another important development took place ...
RNA enters the picture ...

The Central Dogma of Molecular Biology

Replication
DNA duplicates

Information

Transcription
RNA synthesis

RNA polymerase
Nucleus
mRNA
Nuclear membrane

Protein

Unknown

Known

Known

RNA enters the picture ...

Still didn't know how DNA and RNA used for making proteins.

It was clear the code had to be a non-overlapping triplet code:

- messenger RNA
- transfer RNA
- protein

A big concern arose, however – the frame-shift problem:

- messenger RNA
- transfer RNA
- different protein!
Comma-free codes

How is poor transfer RNA molecular supposed to know where it is supposed to bind to mRNA? It has no global context.

Overlapping code didn't have the problem, but that's ruled out.

Solution is *comma-free code*.

A comma-free code is constructed so that only the codons in one reading frame are meaningful; the overlap triplets are nonsense (indicated in black below).

```
AGA|CGA|AUU|AUA|UCA|ACA|AGC|CCC
AGA|CGA|AUU|UAU|CA|AAC|CAG|CCC
AGA|CGA|AUU|UAU|CA|ACA|CAG|CCC
```
Comma-free codes

In 1957, Crick suggested that “adaptor molecules” (i.e., tRNA) might only exist for a subset of the 64 codons that corresponded to a comma-free code. This completely solves the frame shift problem.

Example:

If CGU and AAG are sense codons, then:
(a) GUA and UAA are ruled out (because of CGUAAG),
(b) AGC and GCG are ruled out (because of AAGCGU).
Comma-free codes

How many codons could a comma-free triplet code include?

Must immediately exclude AAA, CCC, GGG, and UUU. Why?

Now consider codon like AGU. Say we have AGUAGU. There would be a frame-shift problem if we allowed GUA or UAG. So we can't use more than one codon related by a cyclic shift.

Hence, we can partition the remaining 60 codons into groups of three, each group related by a cyclic shift. Then we can choose at most one representative codon from each group.

The "magic" number of groups this yields is 20!
Comma-free codes

Exclude AAA, CCC, GGG, UUU:  AAA  CCC  GGG  UUU

Divide remaining 60 triplets into groups of three based on cyclic permutation. Can use no more than one from each group:
Comma-free codes

Comma-free codes have one very desirable property: translation process can take place without requiring global coordination of “adapter” molecules (i.e., tRNA).

These can bind in any order

These can't exist

Concept that translation might be controlled on molecular level in sequential way, like computer reading paper tape, was thought implausible.

No way!
Comma-free codes

Francis Crick, et al. noted:

“This scheme ... allows the intermediates to accumulate at the correct positions on the template without ever blocking the process by settling ... in the wrong place.”

... but also with the caveat ...

“The ... assumptions which we have had to employ to deduce this code are too precarious for us to feel much confidence in it on purely theoretical grounds ... We put it forward because it gives the magic number – 20 – in a neat manner and from reasonable physical postulates.”

Still, the comma-free code idea received almost universal acceptance and was pursued heavily for about 5 years.
Comma-free codes

Solomon Golomb took comma-free codes even further.

Formula for upper bound on size of comma-free code:

\[ n = \text{size of alphabet} \quad (4 \text{ in case of DNA}) \]
\[ k = \text{length of “words”} \quad (3 \text{ in case of codons}) \]

When \( k \) is prime, then upper bound is \( \frac{n^k - n}{k} \).

In case of DNA, upper bound is of course \( \frac{4^3 - 4}{3} = 20 \).

Golomb, et al. found 408 maximal comma-free codes for DNA.

Also developed:

• Transposable comma-free code where both strands of DNA are comma-free, but requires codons with 4 nucleotides.
• Error-correcting code for codons with 6 nucleotides.
Real life intrudes ...

In 1961 this coding craze came to an end – experimental science finally caught up. Nirenberg and Matthaei of the National Institutes of Health announced that artificial RNA's could stimulate protein synthesis in a cell-free system.

The first RNA they tried was poly-U, a long chain of repeating uracil units. In comma-free codes, UUU has to be a nonsense codon, but Nirenberg and Matthaei’s result implied that it codes for the amino acid phenylalanine.

By 1965 the genetic code was mostly solved.

Viewed from nature's perspective, the "magic" number 20 held no magic after all. All the clever attempts for getting 20 amino acids out of 64 codons turned out to be "figments of the human urge to find a pattern." (Brian Hayes)
Diamond code versus nature's code
Composition code versus nature's code

**Composition code**

<table>
<thead>
<tr>
<th>AAA</th>
<th>AAC</th>
<th>ACA</th>
<th>ACC</th>
<th>CAA</th>
<th>CAC</th>
<th>CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAG</td>
<td>AAU</td>
<td>ACG</td>
<td>ACU</td>
<td>CAG</td>
<td>CAU</td>
<td>CCG</td>
</tr>
<tr>
<td>AGA</td>
<td>AGC</td>
<td>AUA</td>
<td>AUC</td>
<td>CGA</td>
<td>GGC</td>
<td>CUA</td>
</tr>
<tr>
<td>AGG</td>
<td>AGU</td>
<td>AUG</td>
<td>AUA</td>
<td>CGG</td>
<td>GGU</td>
<td>CUG</td>
</tr>
<tr>
<td>GAA</td>
<td>GAC</td>
<td>GCA</td>
<td>GCC</td>
<td>UAA</td>
<td>UAC</td>
<td>UCC</td>
</tr>
<tr>
<td>GAG</td>
<td>GAU</td>
<td>GCG</td>
<td>GCU</td>
<td>UAG</td>
<td>UAU</td>
<td>UCU</td>
</tr>
<tr>
<td>GGA</td>
<td>GCC</td>
<td>GUA</td>
<td>GUC</td>
<td>UGA</td>
<td>UGC</td>
<td>UUA</td>
</tr>
<tr>
<td>GGG</td>
<td>GGU</td>
<td>GUG</td>
<td>GUU</td>
<td>UGG</td>
<td>UGU</td>
<td>UUU</td>
</tr>
</tbody>
</table>

**Nature's code**

<table>
<thead>
<tr>
<th>AAA</th>
<th>AAC</th>
<th>ACC</th>
<th>CAA</th>
<th>CAC</th>
<th>CCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAG</td>
<td>AAU</td>
<td>ACG</td>
<td>ACU</td>
<td>CAG</td>
<td>CAU</td>
</tr>
<tr>
<td>AGA</td>
<td>AGC</td>
<td>AUA</td>
<td>AUC</td>
<td>CGA</td>
<td>GGC</td>
</tr>
<tr>
<td>AGG</td>
<td>AGU</td>
<td>AUG</td>
<td>AUA</td>
<td>CGG</td>
<td>GGU</td>
</tr>
<tr>
<td>GAA</td>
<td>GAC</td>
<td>GCA</td>
<td>GCC</td>
<td>UAA</td>
<td>UAC</td>
</tr>
<tr>
<td>GAG</td>
<td>GAU</td>
<td>GCG</td>
<td>GCU</td>
<td>UAG</td>
<td>UAU</td>
</tr>
<tr>
<td>GGA</td>
<td>GCC</td>
<td>GUA</td>
<td>GUC</td>
<td>UGA</td>
<td>UGC</td>
</tr>
<tr>
<td>GGG</td>
<td>GGU</td>
<td>GUG</td>
<td>GUU</td>
<td>UGG</td>
<td>UGU</td>
</tr>
</tbody>
</table>
Comma-free code versus nature's code

<table>
<thead>
<tr>
<th>comma-free code</th>
<th>nature's code</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>AAA</td>
</tr>
<tr>
<td>AAC</td>
<td>AAC</td>
</tr>
<tr>
<td>ACA</td>
<td>ACA</td>
</tr>
<tr>
<td>ACC</td>
<td>ACC</td>
</tr>
<tr>
<td>CAA</td>
<td>CAA</td>
</tr>
<tr>
<td>CAC</td>
<td>CAC</td>
</tr>
<tr>
<td>CCA</td>
<td>CCA</td>
</tr>
<tr>
<td>CCC</td>
<td>CCC</td>
</tr>
<tr>
<td>AAG</td>
<td>AAG</td>
</tr>
<tr>
<td>AAU</td>
<td>AAU</td>
</tr>
<tr>
<td>ACG</td>
<td>ACG</td>
</tr>
<tr>
<td>ACU</td>
<td>ACU</td>
</tr>
<tr>
<td>CAG</td>
<td>CAG</td>
</tr>
<tr>
<td>CAU</td>
<td>CAU</td>
</tr>
<tr>
<td>CCG</td>
<td>CCG</td>
</tr>
<tr>
<td>CCU</td>
<td>CCU</td>
</tr>
<tr>
<td>AGA</td>
<td>AGA</td>
</tr>
<tr>
<td>AGC</td>
<td>AGC</td>
</tr>
<tr>
<td>AUA</td>
<td>AUA</td>
</tr>
<tr>
<td>AUC</td>
<td>AUC</td>
</tr>
<tr>
<td>CGA</td>
<td>CGA</td>
</tr>
<tr>
<td>CGC</td>
<td>CGC</td>
</tr>
<tr>
<td>CUA</td>
<td>CUA</td>
</tr>
<tr>
<td>CUC</td>
<td>CUC</td>
</tr>
<tr>
<td>AGG</td>
<td>AGG</td>
</tr>
<tr>
<td>AGU</td>
<td>AGU</td>
</tr>
<tr>
<td>AUG</td>
<td>AUG</td>
</tr>
<tr>
<td>AUA</td>
<td>AUA</td>
</tr>
<tr>
<td>CGG</td>
<td>CGG</td>
</tr>
<tr>
<td>CGU</td>
<td>CGU</td>
</tr>
<tr>
<td>CUG</td>
<td>CUG</td>
</tr>
<tr>
<td>CUU</td>
<td>CUU</td>
</tr>
<tr>
<td>GAA</td>
<td>GAA</td>
</tr>
<tr>
<td>GAC</td>
<td>GAC</td>
</tr>
<tr>
<td>GCA</td>
<td>GCA</td>
</tr>
<tr>
<td>GCC</td>
<td>GCC</td>
</tr>
<tr>
<td>UAA</td>
<td>UAA</td>
</tr>
<tr>
<td>UAC</td>
<td>UAC</td>
</tr>
<tr>
<td>UCA</td>
<td>UCA</td>
</tr>
<tr>
<td>UCC</td>
<td>UCC</td>
</tr>
<tr>
<td>GAG</td>
<td>GAG</td>
</tr>
<tr>
<td>GAU</td>
<td>GAU</td>
</tr>
<tr>
<td>GCG</td>
<td>GCG</td>
</tr>
<tr>
<td>GCU</td>
<td>GCU</td>
</tr>
<tr>
<td>UAG</td>
<td>UAG</td>
</tr>
<tr>
<td>UAU</td>
<td>UAU</td>
</tr>
<tr>
<td>UCG</td>
<td>UCG</td>
</tr>
<tr>
<td>UCU</td>
<td>UCU</td>
</tr>
<tr>
<td>GGA</td>
<td>GGA</td>
</tr>
<tr>
<td>GGC</td>
<td>GGC</td>
</tr>
<tr>
<td>GUA</td>
<td>GUA</td>
</tr>
<tr>
<td>GUC</td>
<td>GUC</td>
</tr>
<tr>
<td>UGA</td>
<td>UGA</td>
</tr>
<tr>
<td>UGC</td>
<td>UGC</td>
</tr>
<tr>
<td>UUA</td>
<td>UUA</td>
</tr>
<tr>
<td>UUC</td>
<td>UUC</td>
</tr>
<tr>
<td>GGG</td>
<td>GGG</td>
</tr>
<tr>
<td>GGU</td>
<td>GGU</td>
</tr>
<tr>
<td>GUG</td>
<td>GUG</td>
</tr>
<tr>
<td>GUU</td>
<td>GUU</td>
</tr>
<tr>
<td>UGG</td>
<td>UGG</td>
</tr>
<tr>
<td>UGU</td>
<td>UGU</td>
</tr>
<tr>
<td>UUG</td>
<td>UUG</td>
</tr>
<tr>
<td>UUU</td>
<td>UUU</td>
</tr>
</tbody>
</table>

After the true genetic code was determined experimentally in the 1960's, that put an end to this sort of speculation. Right?

Wrong! Even to this day, researchers continue to study the genetic code (we have innate curiosity about our origins).

Obvious question: why this code over all others? Is there some compelling reason we can discern? Or is it just a "frozen accident" (according to Francis Crick)?
Why this genetic code over all others?

Recall that early attempts at devining the genetic code were aimed at maximizing encoding density (motived by Shannon's study of information theory).

Later realized this is not important; nature doesn't care about efficiency. Error tolerance far more critical (natural selection).

Two basic kinds of errors we might worry about:

• Translation errors (wrong protein produced from DNA). Tend to be one-of-a-kind events, probably less important.
• Mutations (DNA itself gets changed). Much more important as they alter the genome.
Fault tolerance already evident in genetic code. Often, altering one nucleotide ("point mutation") does not change amino acid. Even when amino acid is different, single mutation often leads to similar chemical properties. All codons with middle \( U \) are hydrophobic, so 2/3 of point mutations in such cases have same property.

http://www.people.virginia.edu/~rjh9u/code.html
Coding transitions due to point mutations

Changes in amino acids induced by point mutations in codons:

Thickness of edge or node border indicates number of transitions. As colors go from blue to red, amino acids go from hydrophilic to hydrophobic.
Trying to design a better genetic code

It may sound a bit egotistical, but we might ask:
Is nature's genetic code the best possible?

The answer to this question will depend on:
• Our criteria. E.g., what does “best” mean?
• Universe of possible genetic codes we're able to examine.
Trying to design a better code

In 1969, Alff-Steinberger began trying to quantify the code’s resilience to error by means of computer simulation:

• Shuffle codon table, but maintain certain statistical properties of real code (number of codons for each amino acid).
• Measure error resistance by determining whether point mutations yield “similar” amino acids.
• Tested 200 random codes – found that real code works best.

Later, Wong approached same question from another angle:

• Try hand-crafted solution instead of random codes.
• Look for best substitution for each amino acid.
• Found that substitutions generated by natural code are less than half as close, on average, as the best possible.
Trying to design a better genetic code

Returning to studies of random codes, Haig and Hurst of the University of Oxford generated 10,000 random permutations in 1991. Natural code was better than all but two of them (measured in terms of polar requirement, i.e., hydrophobicity vs. hydrophilicity).

Later study of 1,000,000 random codes by Hurst and Freeland, also accounting for nature's biases in mutations and mistranslations, led to conclusion that real genetic code is more error tolerant than all but one other code.
Trying to design a better genetic code

Nature's code (left) vs. "optimal" genetic code (right) determined experimentally by Freeland and Hurst. Critereon is amino acid substitutions similarly hydrophilic or hydrophobic.
Some questions to ponder ...

• What evidence do we have that the genetic code is evolving? (If it were evolving, wouldn't we see some evidence of this?)

• What sorts of experiments come to mind (either computer simulations or real biological experiments)?

• Is the genetic code really just a “frozen accident”?
Wrap-up

Readings for next time:

• Papers on Bioethics (see "Readings for Final Papers and Projects" folder on Blackboard).

Remember:

• Come to class having done the readings.
• Check Blackboard regularly for updates.