CSE 397-497:
Computational Issues in Molecular Biology

Lecture 24

Spring 2004

LEHIGH UNIVERSITY

Important points to remember

- Final paper / project is due by 5:00 pm on Friday, April 30.
- If you have questions about this, just ask.
- If you still owe me a scribe report, get it to me ASAP.
- Interested in giving an optional 5-minute presentation on your final project on last day of class? Let me know.
- For those who are interested, workshop on graduate program in bio-engineering to be held on May 19 – details to follow.

Today we step back in time

"The Invention of the Genetic Code," Brian Hayes, *American Scientist*, vol. 86, no. 1, January-February 1998, pp. 8-14.

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

1865 Gregor Mendel, working alone in Austrian monastery, discovers that some characteristics are inherited in 'units'.

1870 Friedrich Miescher isolates chemicals from cell nucleus, including 'nucleic acids'. However, most people are more interested in proteins in nucleus.

1879 Walter Flemming describes behavior of chromosomes during cell division, implicating these nuclear structures in inheritance.

1900 Hugo DeVries and others rediscover Mendel's work and establish first laws of inheritance.

1909 Wilhelm Johannsen coins term 'gene'.

1911 Thomas Hunt Morgan is first to show that genes are arranged in linear fashion along chromosomes.

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

Genetic Code timeline #2

1928 Frederick Griffith uses chemical extract to convert harmless pneumonia bacteria into pathogenic forms, but nature of 'inheritance factor' is unknown.

1929 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.

1941 George Beadle & Edward Tatum show genes direct production of proteins.

1943 William Astbury makes first X-ray diffraction images of DNA.

1944 Building on Griffith's work, Oswald Avery & colleagues show that DNA can 'transform' cells, cementing link between DNA and genes.

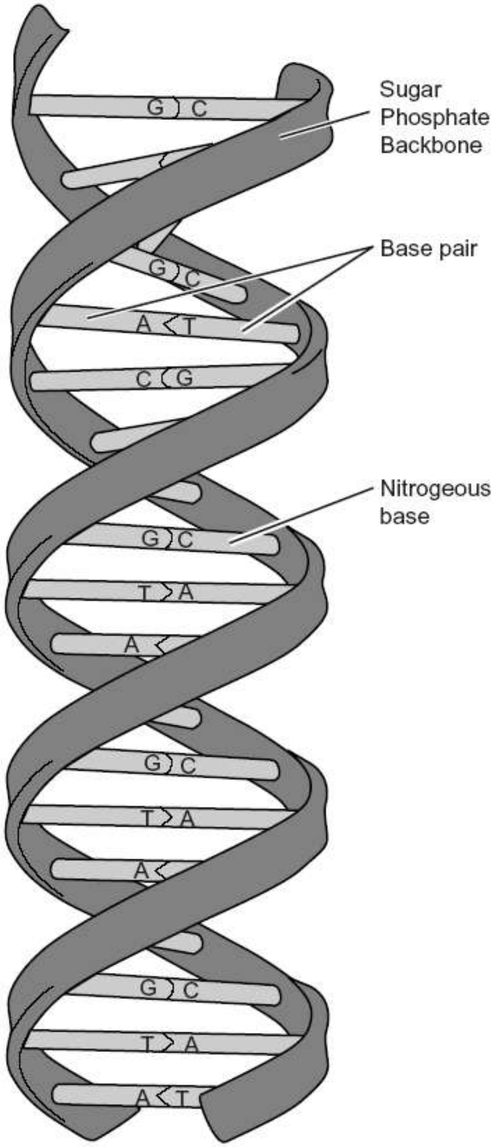
1950 Edwin Chargaff discovers patterns in amounts of four bases in DNA: amounts of G and C, and of A and T, are always same.

1951 Rosalind Franklin takes her first X-ray diffraction pictures.

1953 James Watson & Francis Crick publish first paper proposing double helix structure for DNA.

http://www.wellcome.ac.uk/en/fourplus/DNA_timeline.html

What was known in 1953?



http://www.accessexcellence.org/AB/GG/nhgri_PDFs/dna.pdf

- 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 replicates itself during cell division (transcription).

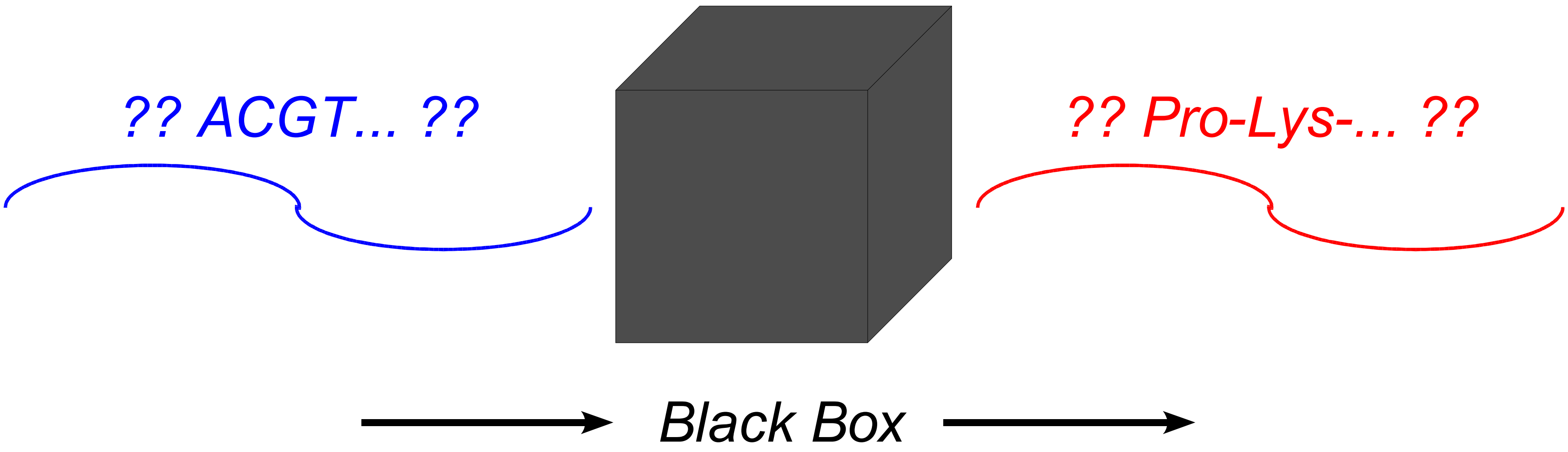
... and ...

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

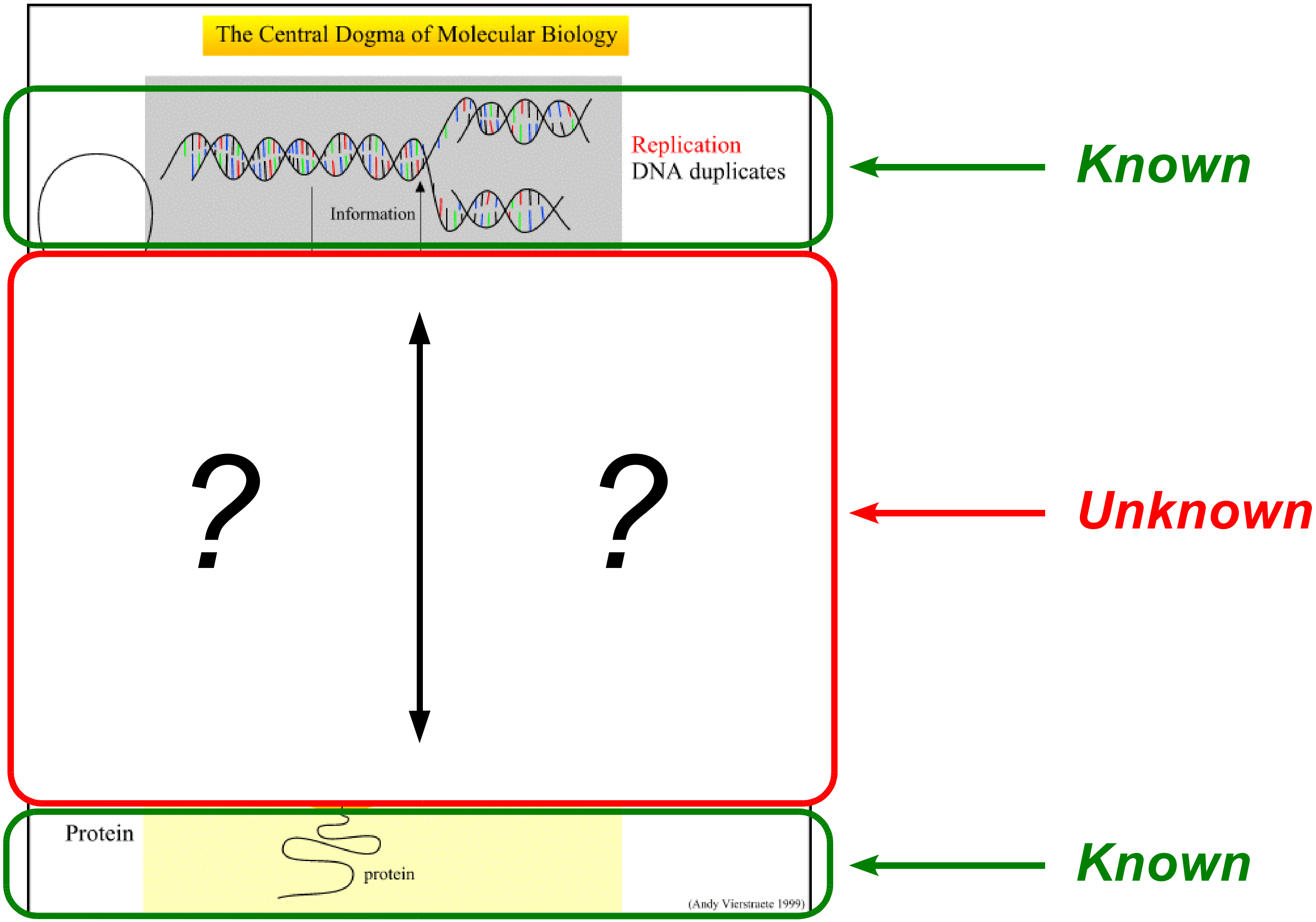
But what is the connection???

What wasn't known in 1953?

- ✗ No DNA sequences (had not been sequenced yet).
- ✗ Fragmentary information about protein sequences (insulin).
- ✗ Concept of RNA (including mRNA and tRNA).
- ✗ The Genetic Code – mapping from a four symbol alphabet to a 20 symbol alphabet – and how it is implemented.



Back in 1953



http://allserv.rug.ac.be/~avierstr/principles/centraldogma.html

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:

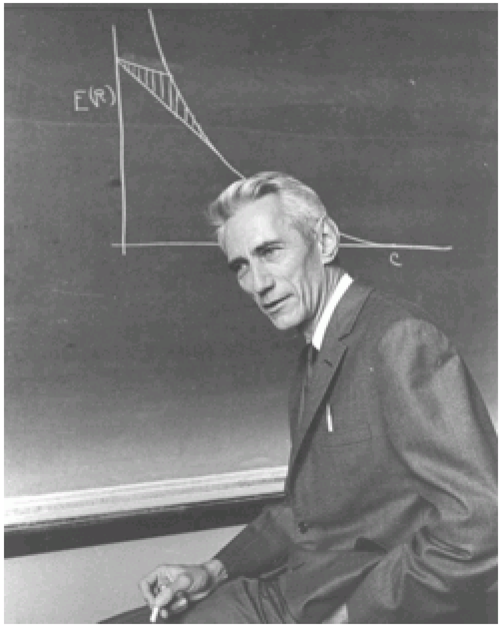
- You don't know any sequence for a real DNA molecule. Sequences for a few proteins are just becoming available.

What does this imply?

- Anything you propose will be an abstract theory awaiting later experimental validation. But that's okay ...

Some additional context

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:

$4^1 = 4 < 20$ ... *nope, not enough*

$4^2 = 16 < 20$ ... *nope, not enough*

$4^3 = 64 \geq 20$ ... *looks good!*

- 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.

First theory: George Gamow's diamond code

George Gamow was an extremely famous physicist, one of the early proponents of the “Big Bang” theory in astrophysics.

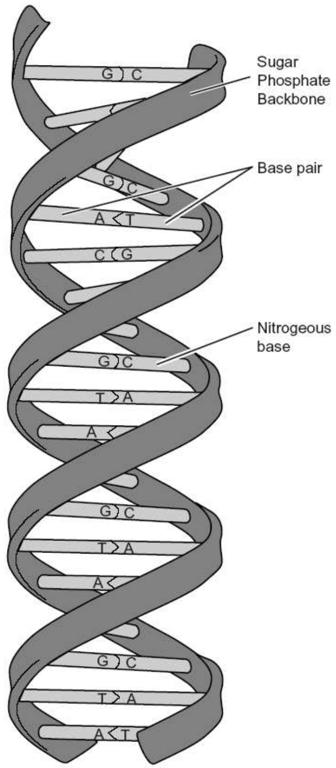
Recall that the concept of RNA as a mediator between DNA and proteins was completely unknown.

In the absence of RNA, Gamow made the reasonable assumption that proteins form directly on a template created by the DNA double helix.

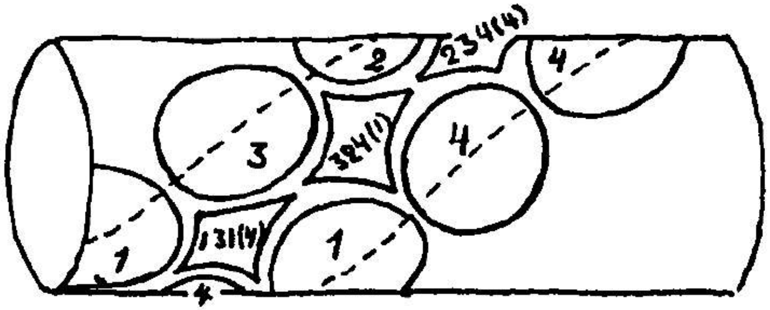
The various combinations of nucleotides along the grooves create disinctive cavities to attract a specific amino acid.

http://www.gwu.edu/~physics/gwmageh.html

First theory: George Gamow's diamond code



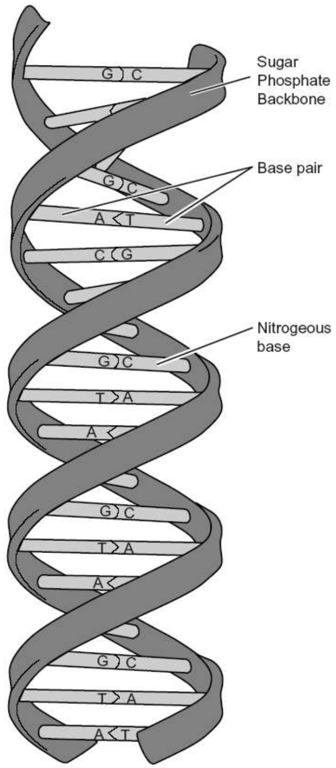
Turn double helix on its side and picture it like this:

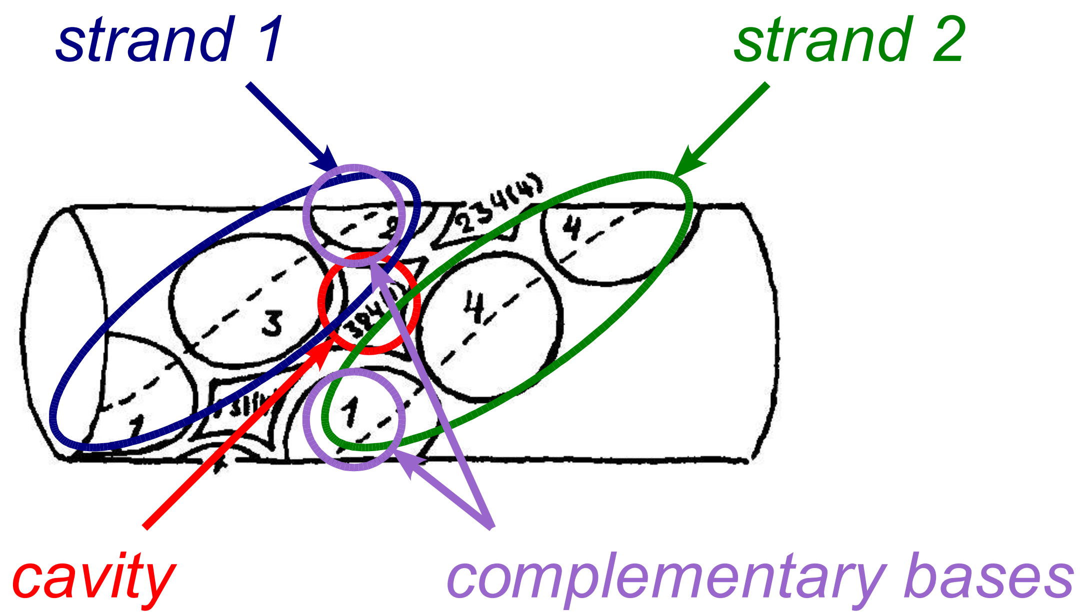


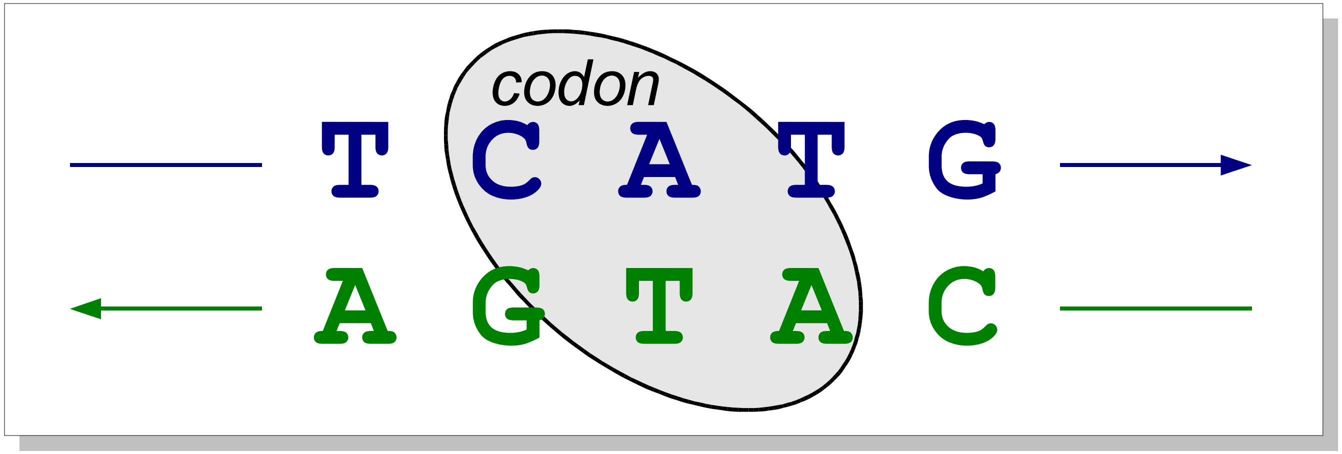
Note: I think he got the "twist" backwards.

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

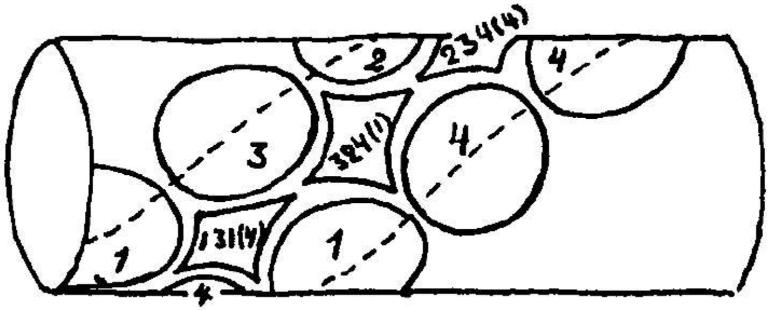
First theory: George Gamow's diamond code







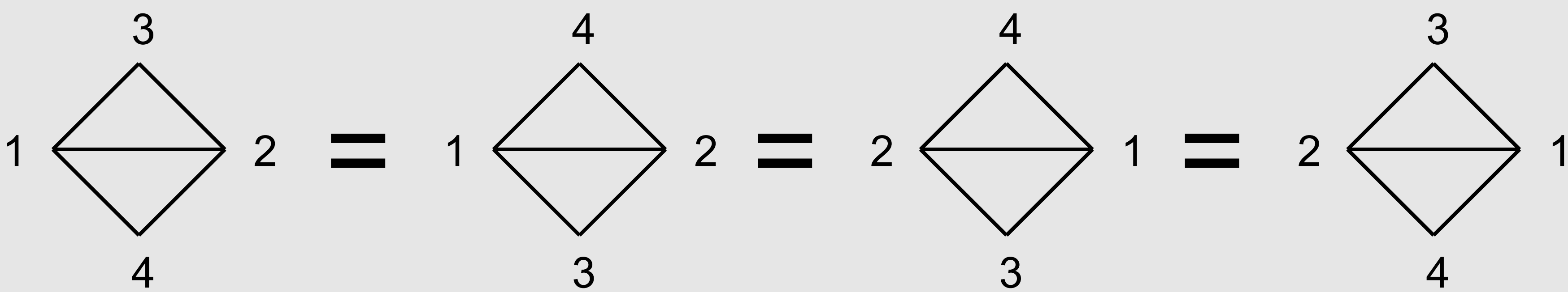
First theory: George Gamow's diamond code



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?

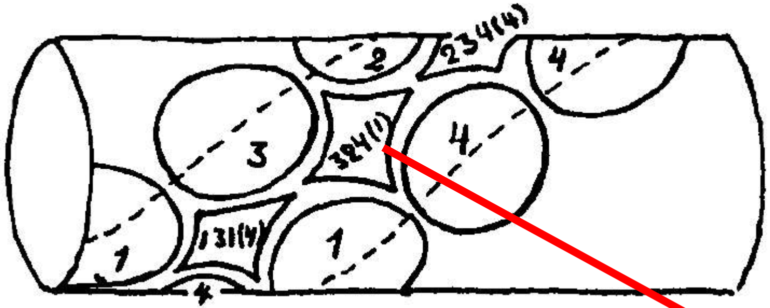
He exploited symmetries:

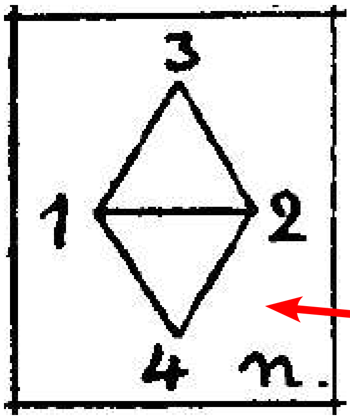


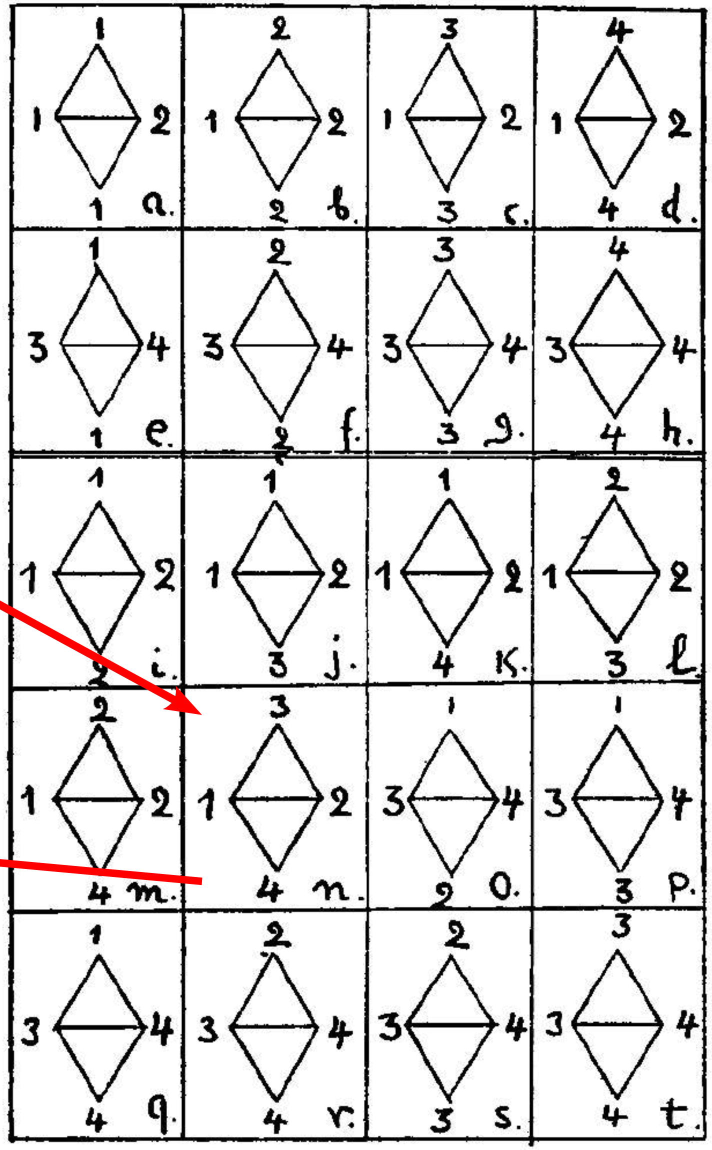
Say 3 = *C*, 1 = *A*, 4 = *G*. Then *CAG* = *GAC* = *GTC* = *CTG*.

First theory: George Gamow's diamond code

Fortuitously, this yields 20 codons!







n = some amino acid (we don't know which without experiments).

Symmetries of the diamond code

AAA	AUA	ACA	AGA
CAC	CUC	CCC	CGC
GAG	GUG	GCG	GGG
UAU	UUU	UCU	UGU

AAC	CAA	AUC	CUA
AAG	GAA	AUG	GUA
AAU	UAA	AUU	UUA
ACC	CCA	AGC	CGA
ACG	GCA	AGG	GGA
ACU	UCA	AGU	UGA
CAG	GAC	CUG	GUC
CAU	UAC	CUU	UUC
CCG	GCC	CGG	GGC
CCU	UCC	CGU	UGC
GAU	UAG	GUU	UUG
GCU	UCG	GGU	UGG

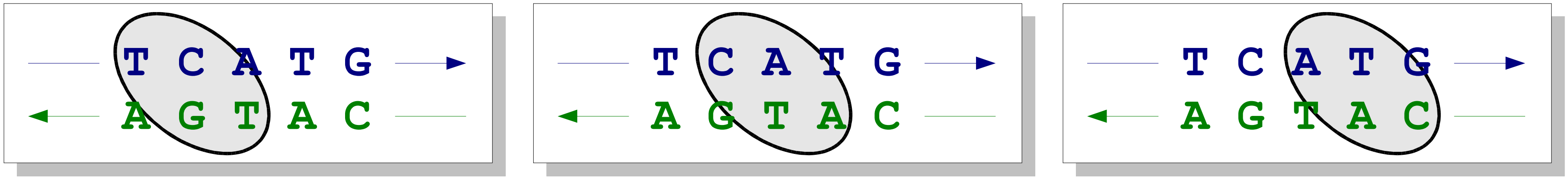
case we considered earlier

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

All codons in each class specify same amino acid.

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.

Persistence ...

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.

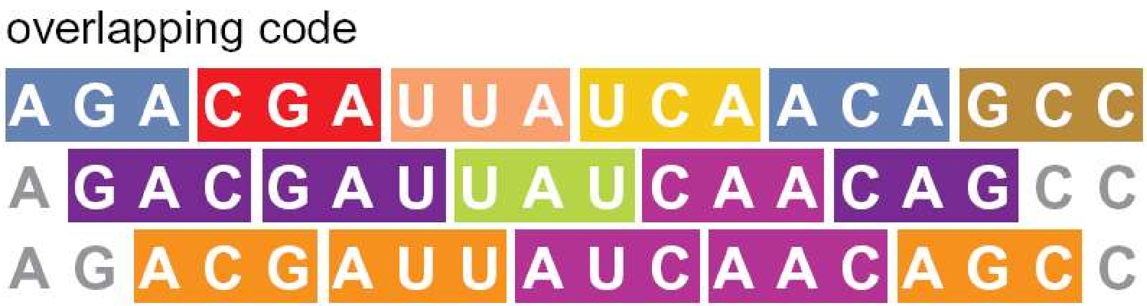
Recall in diamond code we had *CAG* = *GAC* = *GTC* = *CTG*.

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*.

Overlapping codes



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):

$20^2 = 400$ possible amino acid sequences

$4^4 = 256$ possible codons

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

Overlapping codes

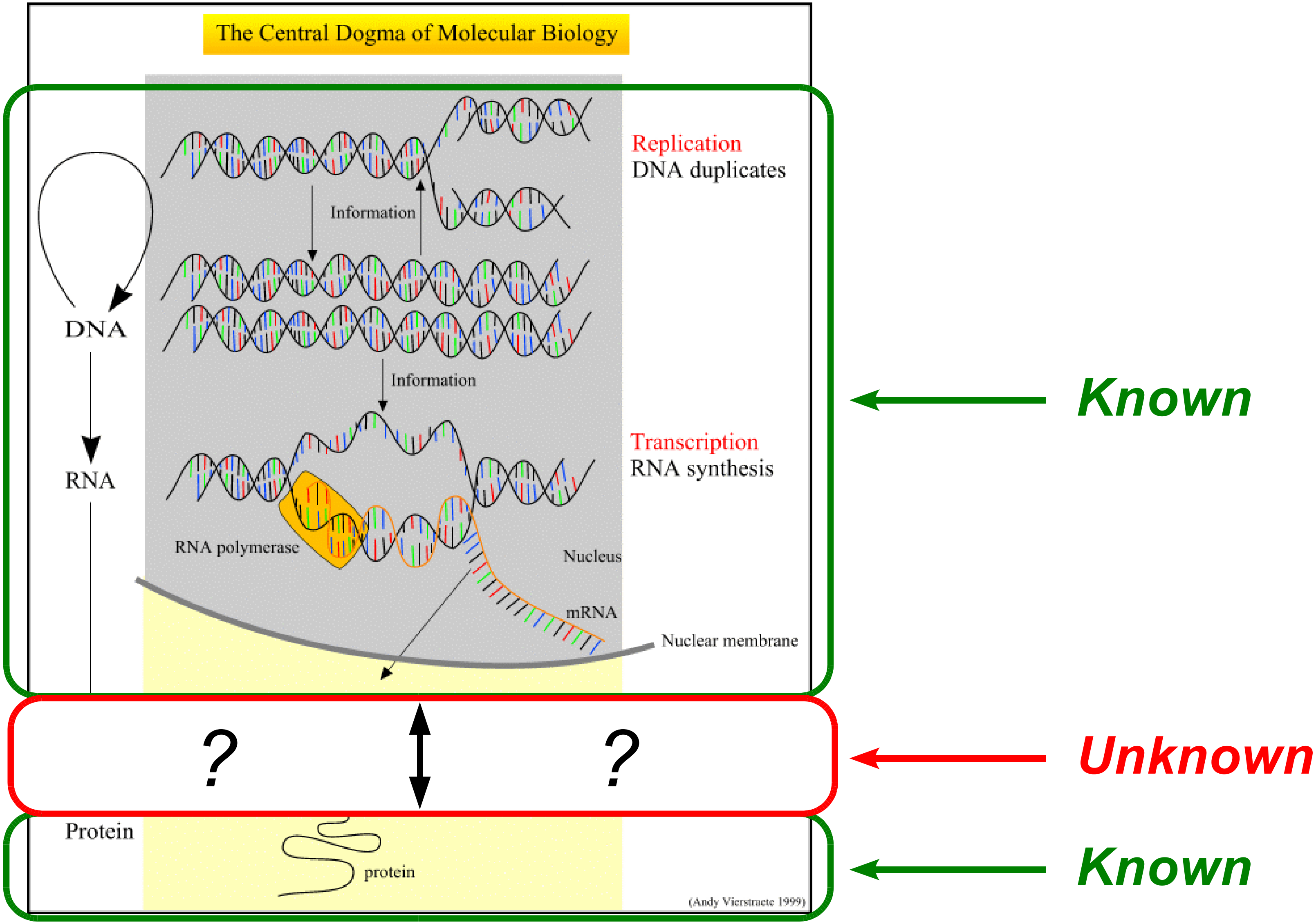
While overlapping codes were eventually eliminated from consideration, it was obvious from the start that they had one undesirable property: a single nucleotide mutation could affect up to three adjacent amino acids. This seems a bit dangerous.

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 ...

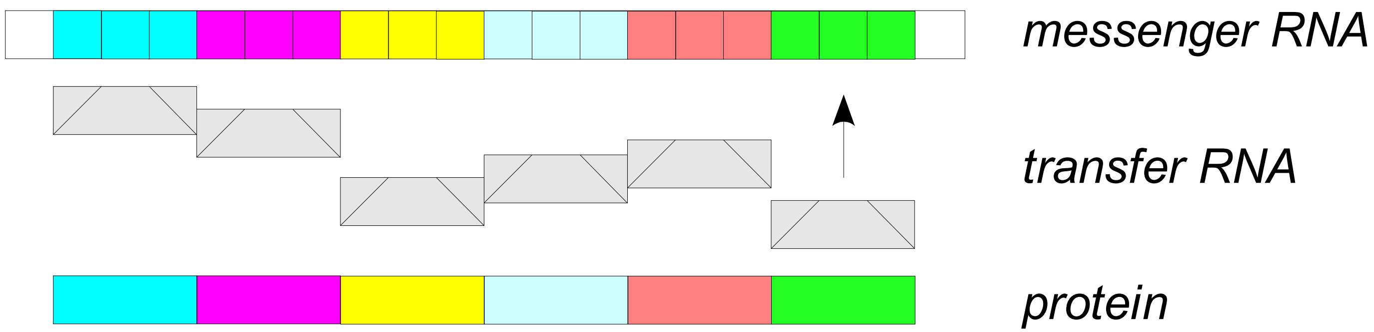


http://allserv.rug.ac.be/~avierstr/principles/centraldogma.html

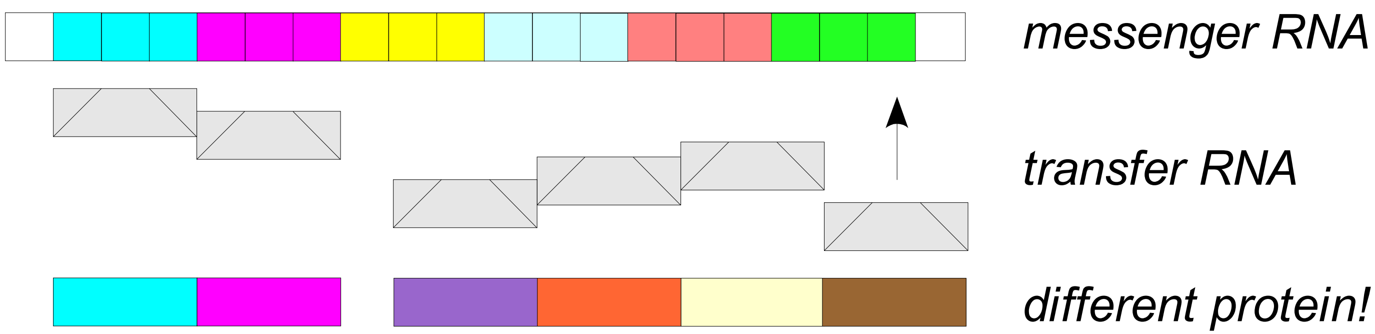
RNA enters the picture ...

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:



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



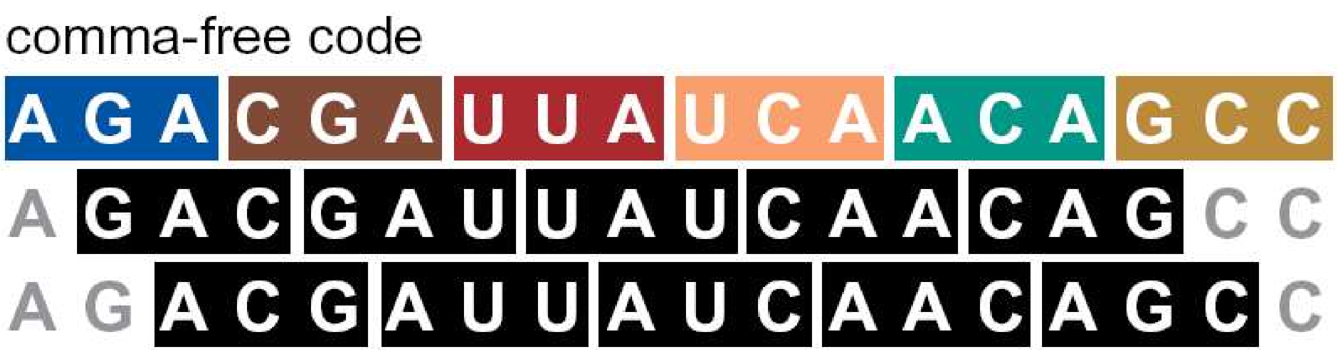
Comma-free codes

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).



LEHIGH UNIVERSITY

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 <u>20</u> !

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:

AAC	ACA	CAA	AUG	UGA	GAU
AAG	AGA	GAA	AUU	UUA	UAU
AAU	AUA	UAA	CCG	CGC	GCC
ACC	CCA	CAC	CCU	CUC	UCC
ACG	CGA	GAC	CGG	GGC	GCG
ACU	CUA	UAC	CGU	GUC	UCG
AGC	GCA	CAG	CUG	UGC	GCU
AGG	GGA	GAG	CUU	UUC	UCU
AGU	GUA	UAG	GGU	GUG	UGG
AUC	UCA	CAU	GUU	UUG	UGU

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 perspetive, 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.

Diamond code versus nature's code

diamond code

AAA	AAC	ACA	ACC	CAA	CAC	CCA	CCC
AAG	AAU	ACG	ACU	CAG	CAU	CCG	CCU
AGA	AGC	AUA	AUC	CGA	CGC	CUA	CUC
AGG	AGU	AUG	AUU	CGG	CGU	CUG	CUU
GAA	GAC	GCA	GCC	UAA	UAC	UCA	UCC
GAG	GAU	GCG	GCU	UAG	UAU	UCG	UCU
GGA	GGC	GUA	GUC	UGA	UGC	UUA	UUC
GGG	GGU	GUG	GUU	UGG	UGU	UUG	UUU

nature's code

AAA	AAC	ACA	ACC	CAA	CAC	CCA	CCC
AAG	AAU	ACG	ACU	CAG	CAU	CCG	CCU
AGA	AGC	AUA	AUC	CGA	CGC	CUA	CUC
AGG	AGU	AUG	AUU	CGG	CGU	CUG	CUU
GAA	GAC	GCA	GCC	UAA	UAC	UCA	UCC
GAG	GAU	GCG	GCU	UAG	UAU	UCG	UCU
GGA	GGC	GUA	GUC	UGA	UGC	UUA	UUC
GGG	GGU	GUG	GUU	UGG	UGU	UUG	UUU

Composition code versus nature's code

composition code

nature's code

Comma-free code versus nature's code

comma-free code

nature's code