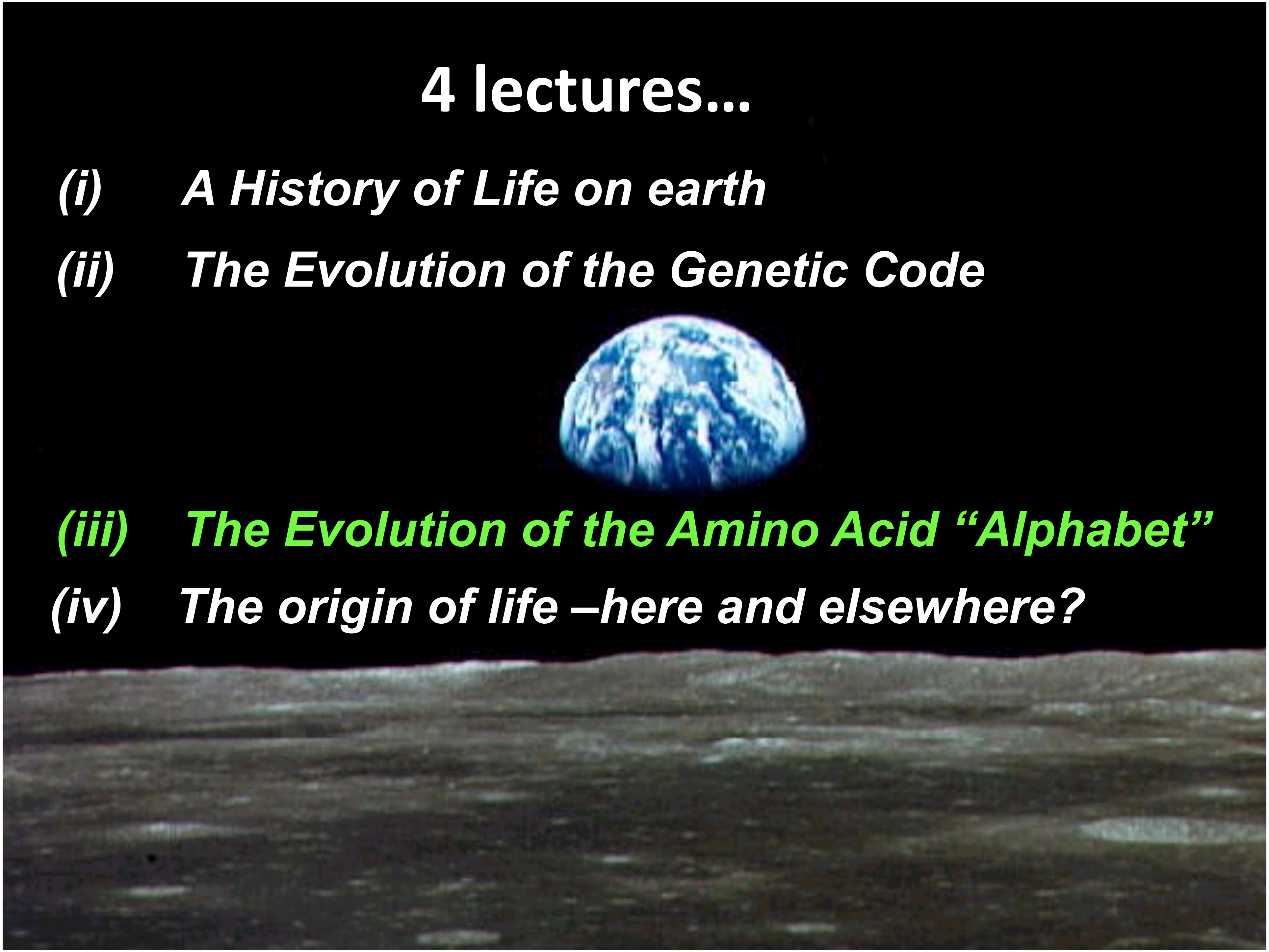
The evolution of the amino acid alphabet

Stephen Freeland

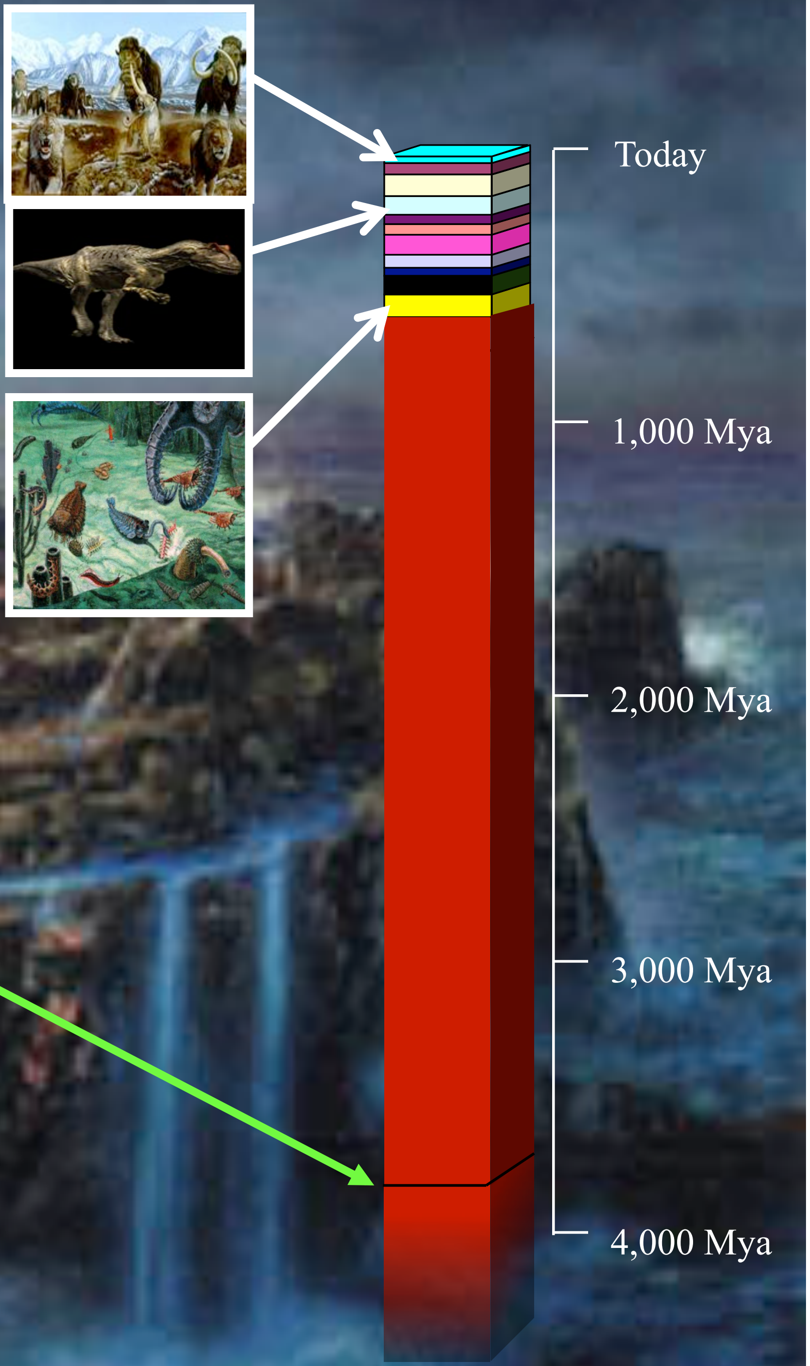
UMBC
AN HONORS
UNIVERSITY
IN MARYLAND



4 lectures…

- *(i) A History of Life on earth*
- *(ii) The Evolution of the Genetic Code*
- *(iii) The Evolution of the Amino Acid “Alphabet”*
- *(iv) The origin of life –here and elsewhere?*

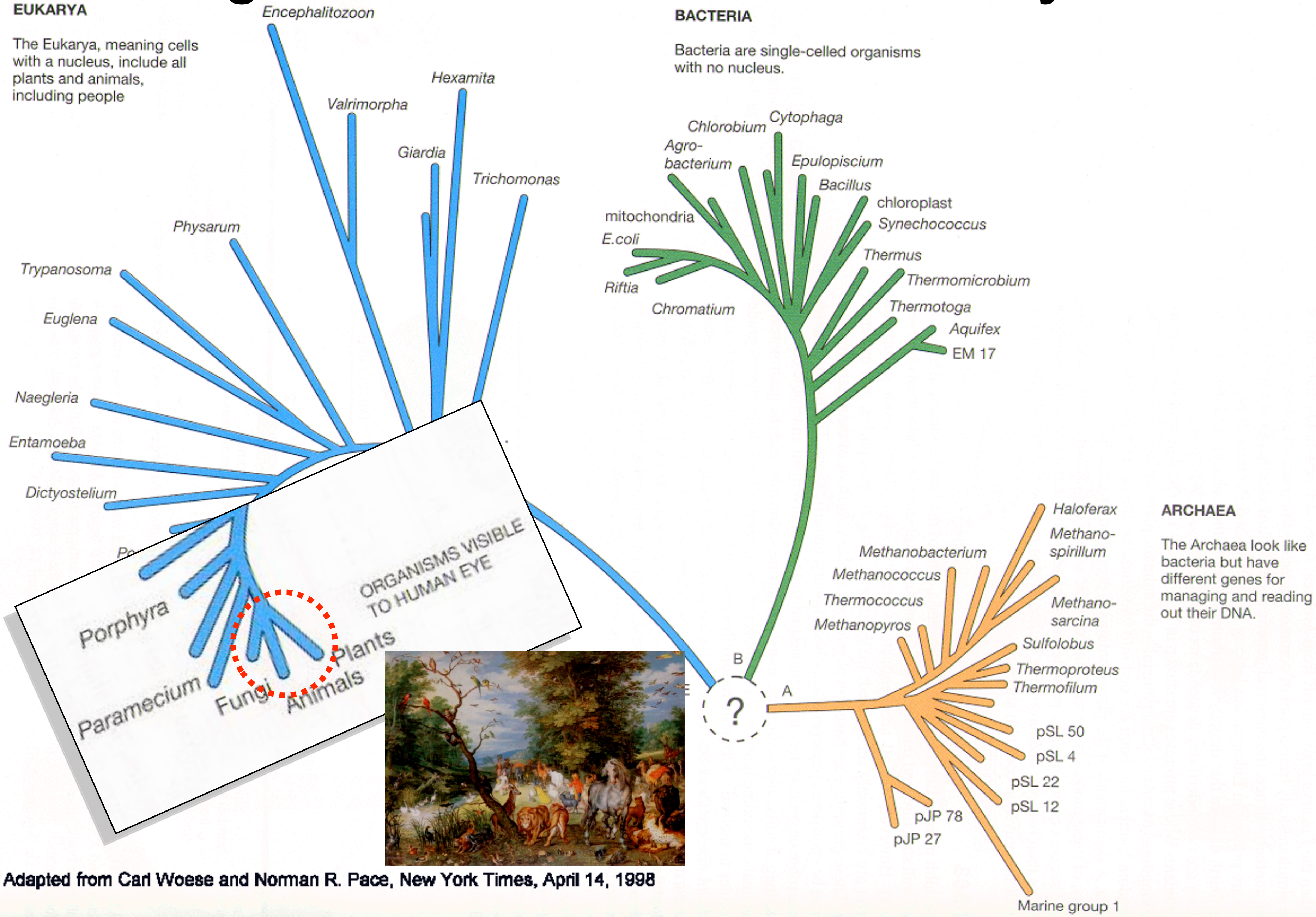
Life on earth to scale...



Nature (1996) 384:55-9. “**Evidence for life on Earth before 3,800 million years ago**”, *Eiler JM, Mojzsis SJ, Arrhenius G.*

“It is unknown when life first appeared on Earth…Here we … provide evidence for the emergence of life on Earth by at least 3,800 Myr before present.”

◆ A genetic tree of life's diversity...



EUKARYA

The Eukarya, meaning cells with a nucleus, include all plants and animals, including people

BACTERIA

Bacteria are single-celled organisms with no nucleus.

ARCHAEA

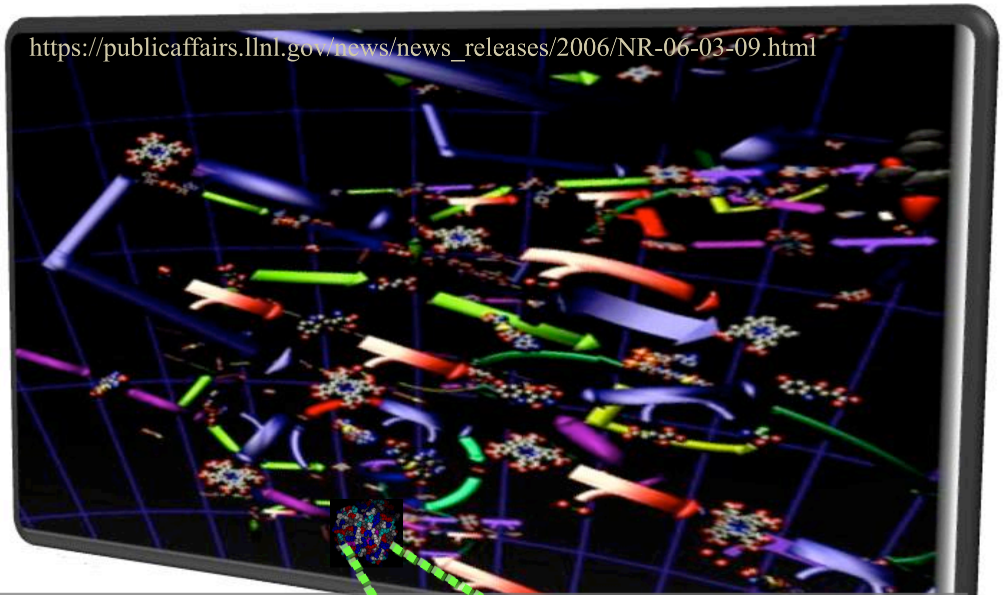
The Archaea look like bacteria but have different genes for managing and reading out their DNA.

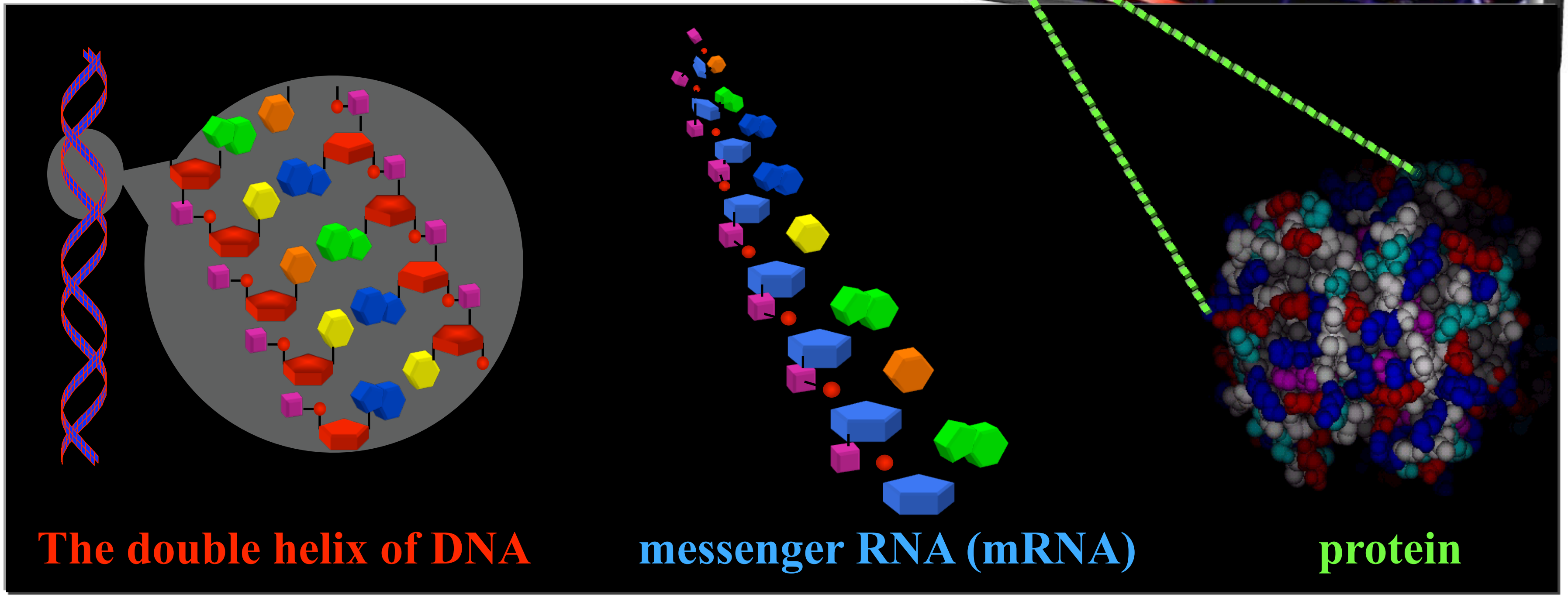
Adapted from Carl Woese and Norman R. Pace, New York Times, April 14, 1998

The Central Dogma of Molecular Biology …

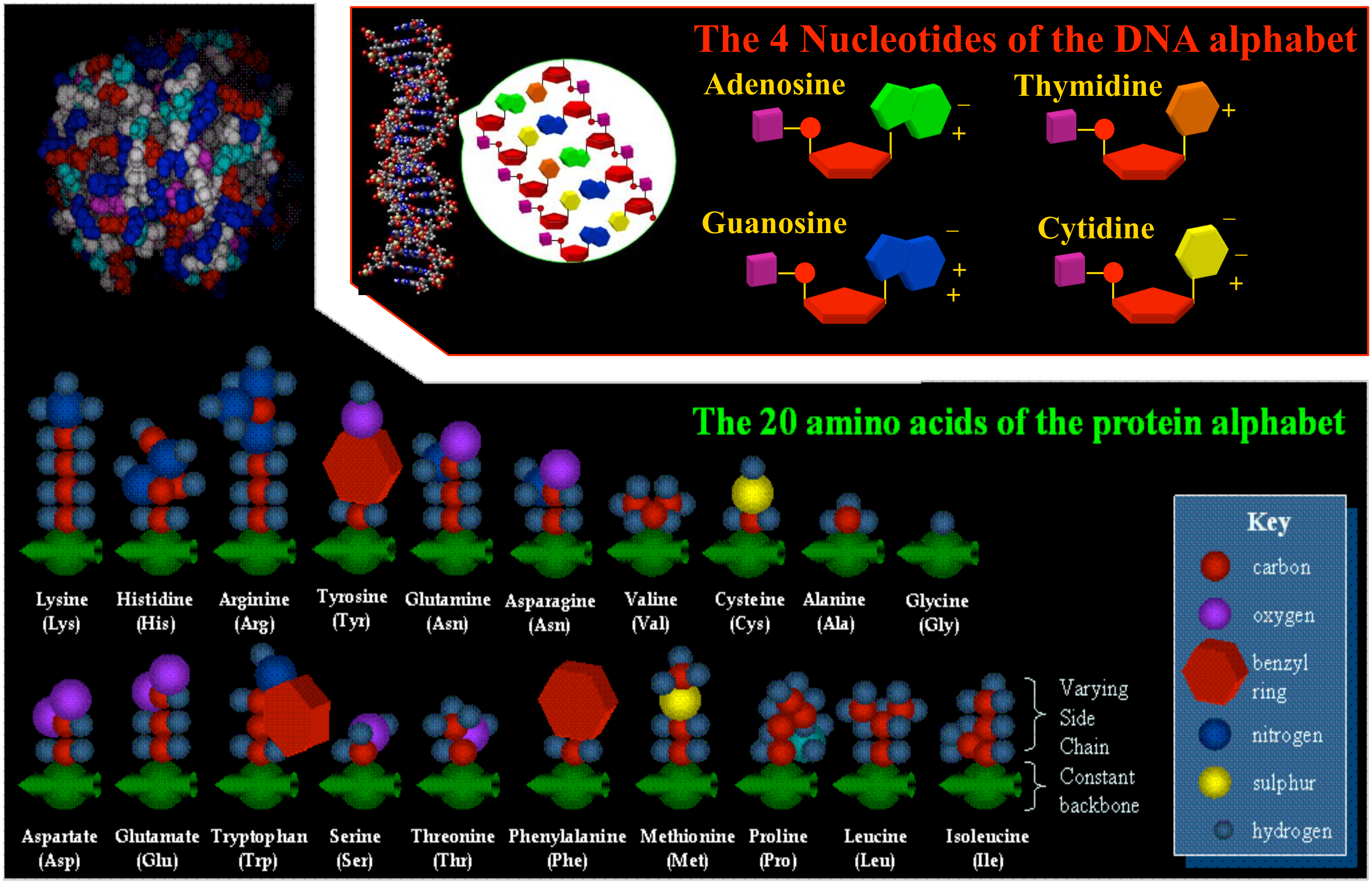
Today, in every cell of every organism of every species...

genetic information stored in nucleic acids (**DNA**→**RNA**) is being translated into the structural and catalytic **proteins** that produce **metabolism**

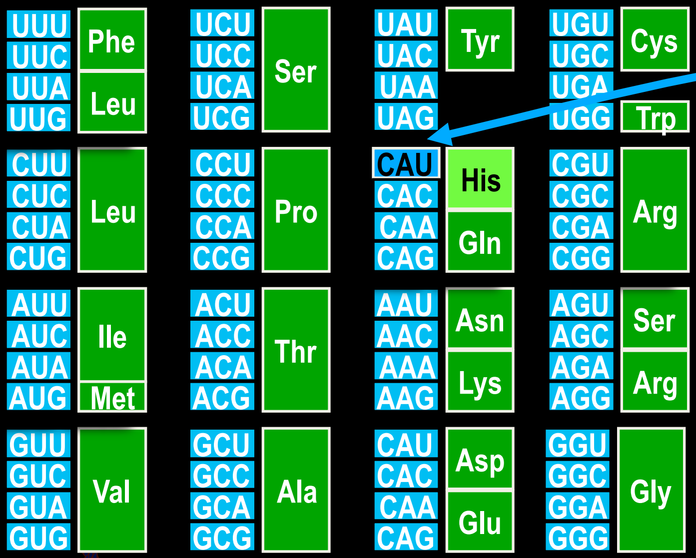

https://publicaffairs.llnl.gov/news/news_releases/2006/NR-06-03-09.html



The “central dogma” of molecular biology – involves two *utterly different* chemical languages

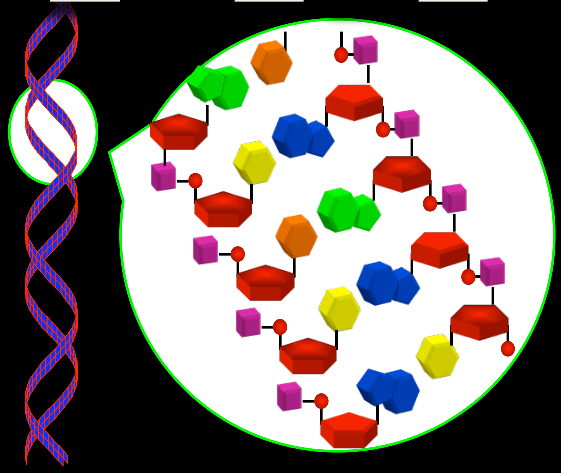


Translation between chemical languages requires decoding…

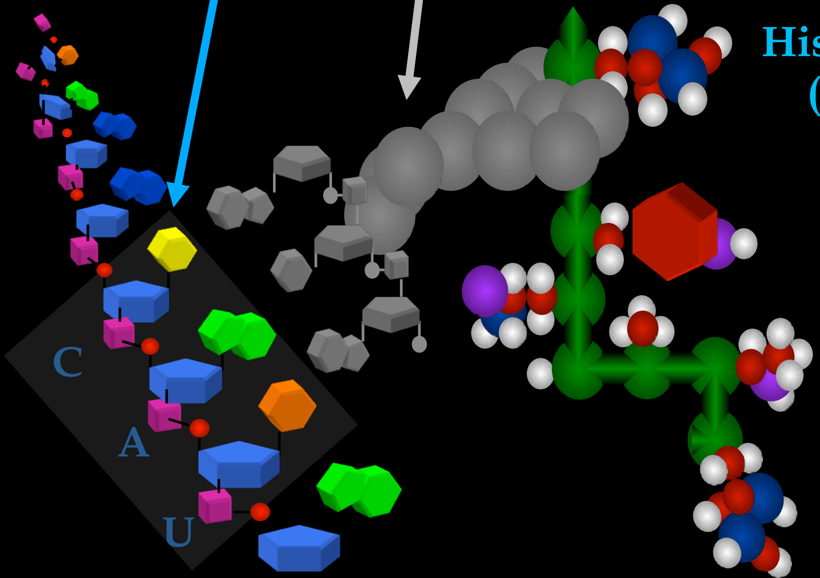


UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC		UCC		UAC		UGC	
UUA	Leu	UCA		UAA		UGA	
UUG		UCG		UAG		UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Gln	CGA	
CUG		CCG		CAG		CGG	
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lys	AGA	Arg
AUG	Met	ACG		AAG		AGG	
GUU	Val	GCU	Ala	CAU	Asp	GGU	Gly
GUC		GCC		CAC		GGC	
GUA		GCA		CAA	Glu	GGA	
GUG		GCG		CAG		GGG	

A triplet ‘*Codon*’ of mRNA



The double helix of DNA



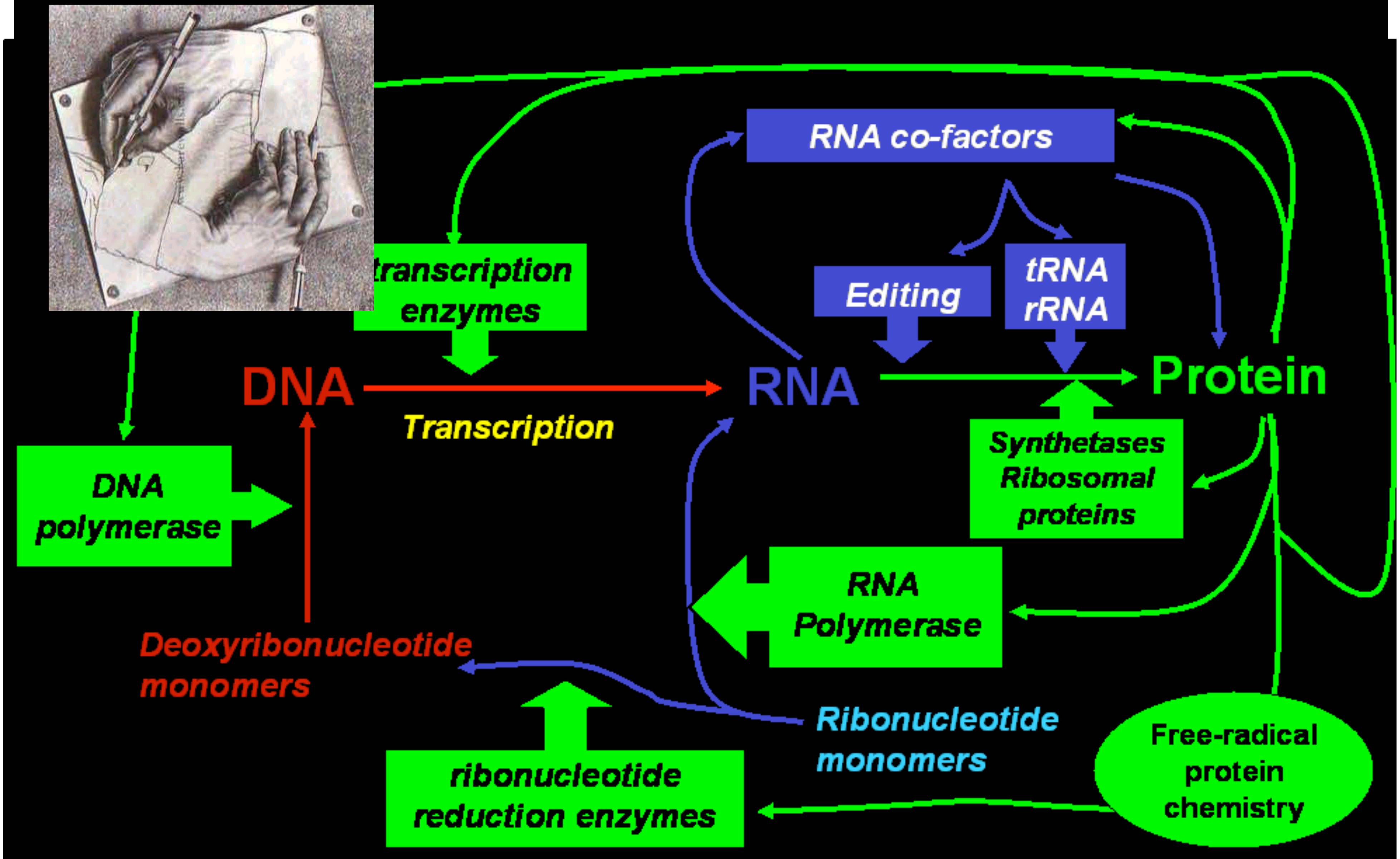
tRNA – the ‘adaptor’ molecule

Histidine (His)

Messenger RNA (mRNA)

Protein

Solving the Puzzle of Primordial Self-Replication



RNA co-factors

transcription enzymes

Editing

tRNA
rRNA

DNA

Transcription

RNA

Protein

DNA polymerase

Synthetases
Ribosomal
proteins

RNA
Polymerase

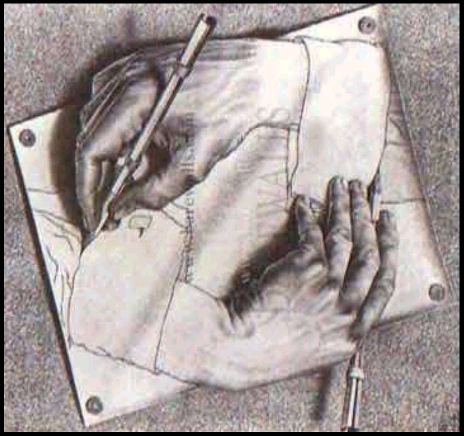
Deoxyribonucleotide monomers

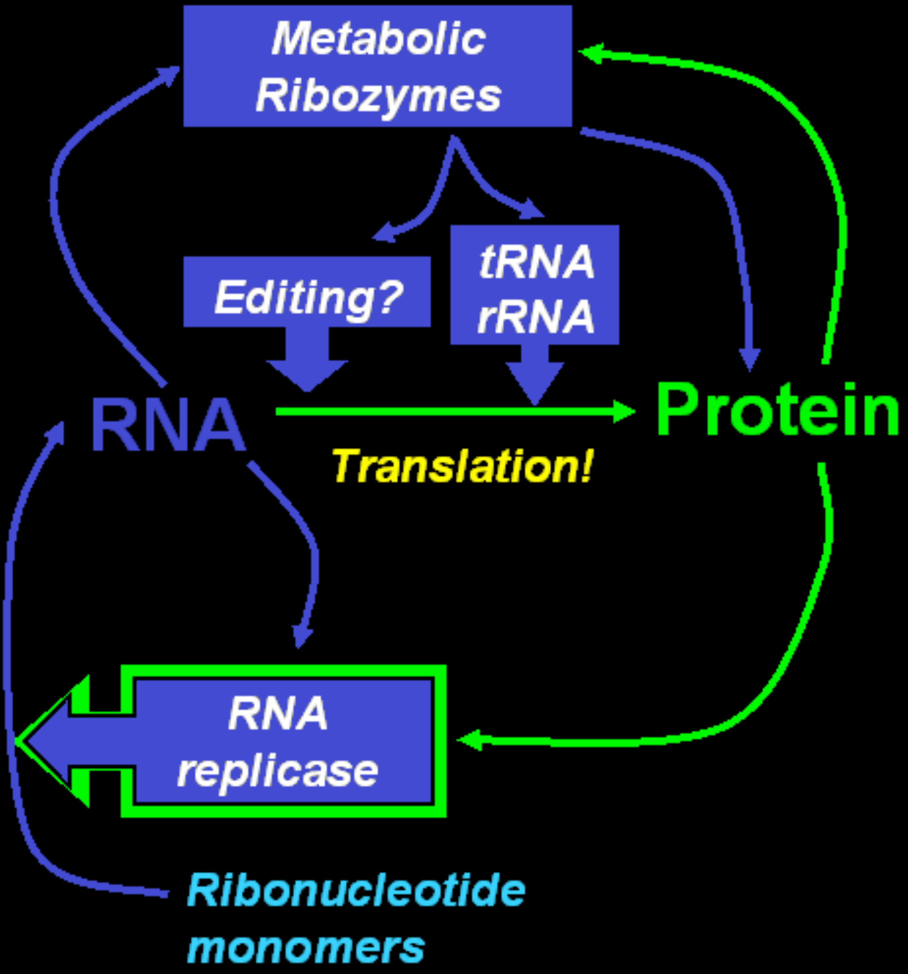
Ribonucleotide monomers

ribonucleotide reduction enzymes

Free-radical
protein
chemistry

Solving the Puzzle of Primordial Self-Replication





Metabolic Ribozymes

Editing?

tRNA
rRNA

RNA

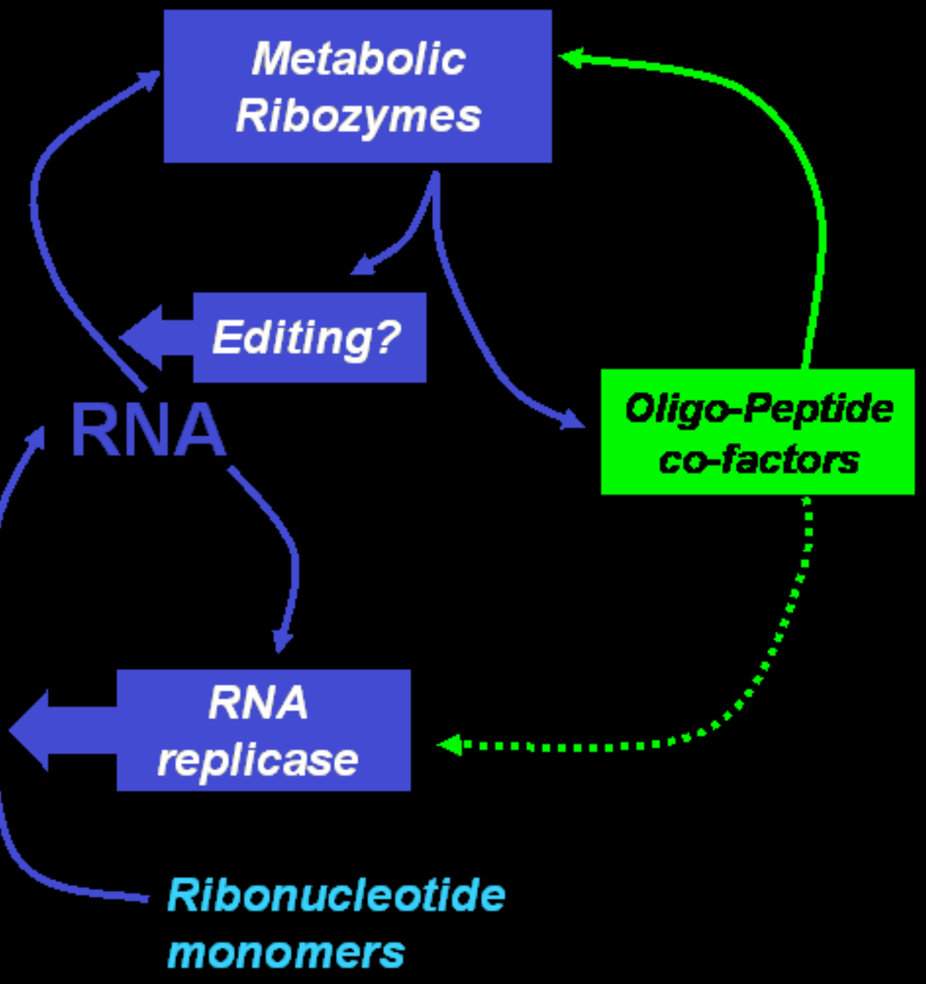
Protein

Translation!

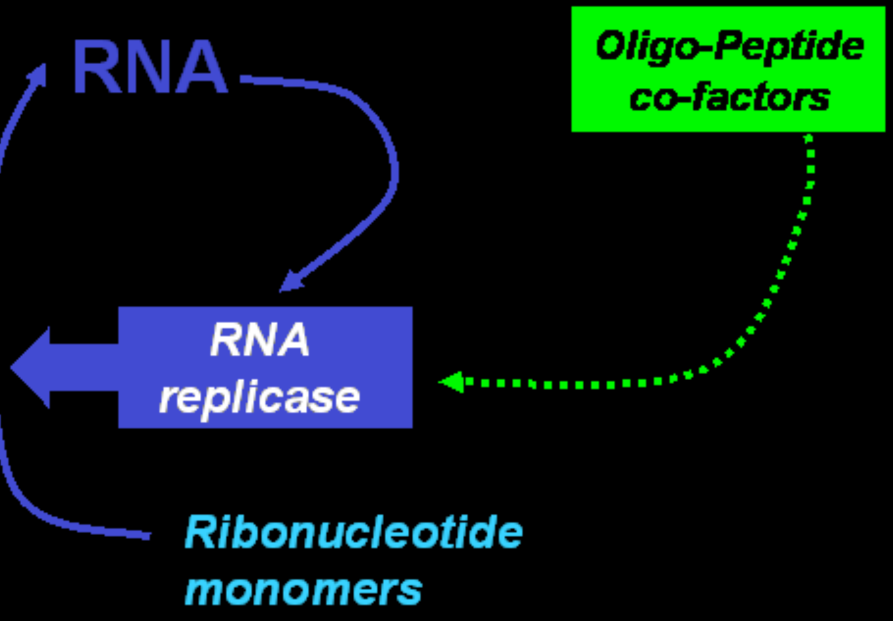
RNA
replicase

Ribonucleotide
monomers

Solving the Puzzle of Primordial Self-Replication



Solving the Puzzle of Primordial Self-Replication



RNA

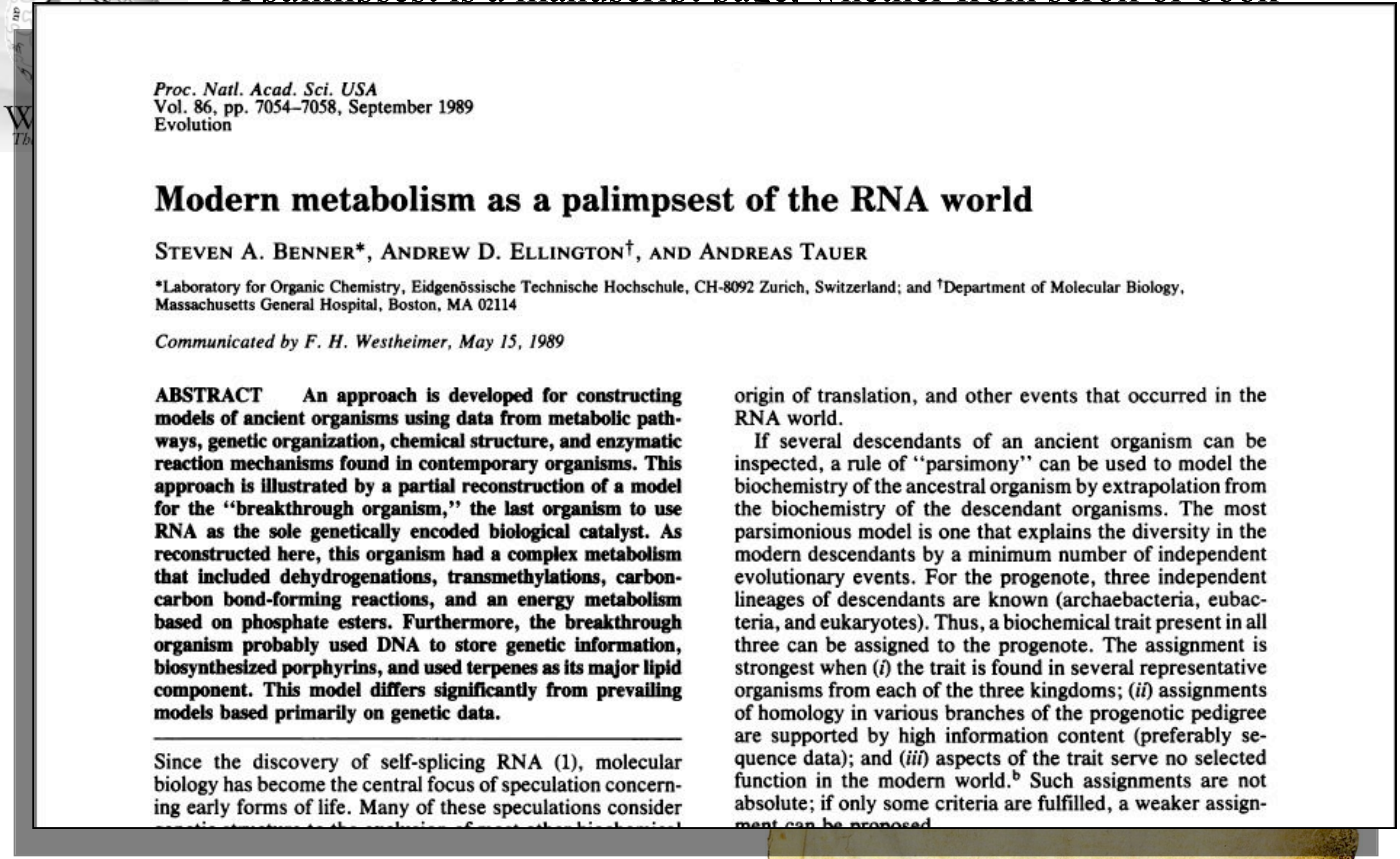
RNA replicase

Oligo-Peptide co-factors

Ribonucleotide monomers

the *palimpsest* view of modern metabolism

“A palimpsest is a manuscript page, whether from scroll or book



Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 7054–7058, September 1989
Evolution

Modern metabolism as a palimpsest of the RNA world

STEVEN A. BENNER*, ANDREW D. ELLINGTON†, AND ANDREAS TAUER

*Laboratory for Organic Chemistry, Eidgenössische Technische Hochschule, CH-8092 Zurich, Switzerland; and †Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114

Communicated by F. H. Westheimer, May 15, 1989

ABSTRACT An approach is developed for constructing models of ancient organisms using data from metabolic pathways, genetic organization, chemical structure, and enzymatic reaction mechanisms found in contemporary organisms. This approach is illustrated by a partial reconstruction of a model for the "breakthrough organism," the last organism to use RNA as the sole genetically encoded biological catalyst. As reconstructed here, this organism had a complex metabolism that included dehydrogenations, transmethylations, carbon-carbon bond-forming reactions, and an energy metabolism based on phosphate esters. Furthermore, the breakthrough organism probably used DNA to store genetic information, biosynthesized porphyrins, and used terpenes as its major lipid component. This model differs significantly from prevailing models based primarily on genetic data.

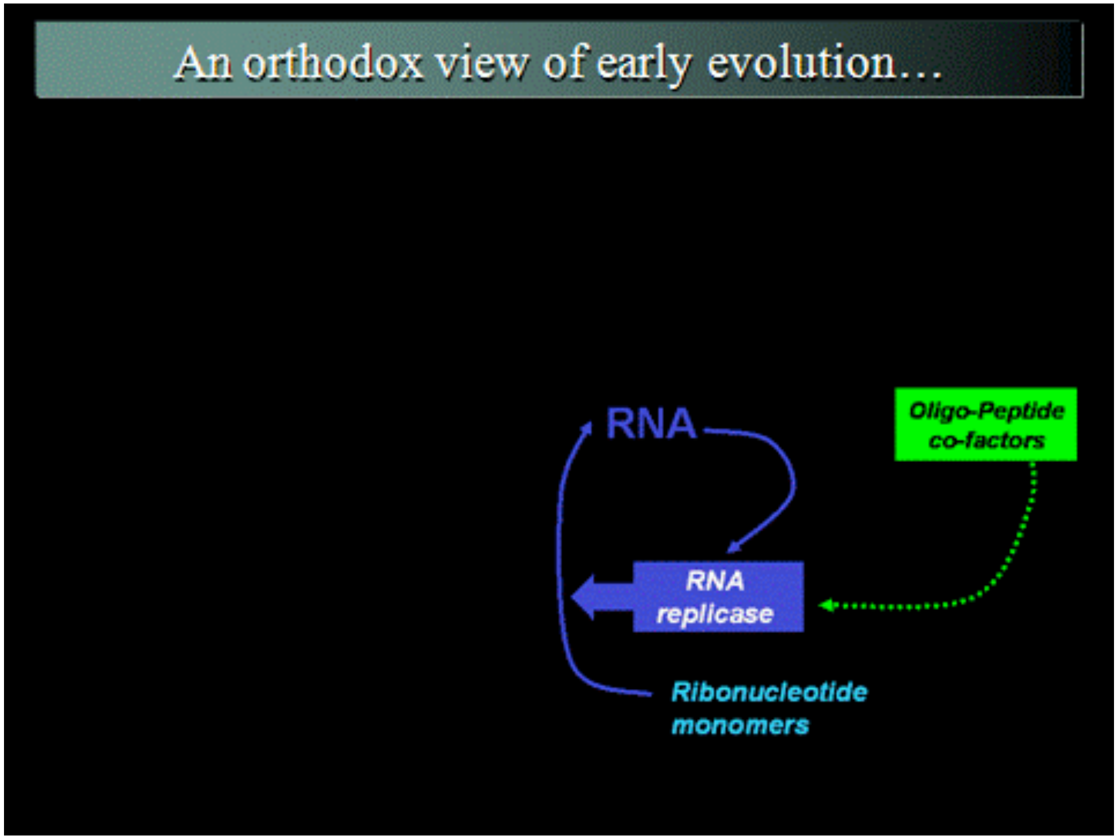
Since the discovery of self-splicing RNA (1), molecular biology has become the central focus of speculation concerning early forms of life. Many of these speculations consider

origin of translation, and other events that occurred in the RNA world.

If several descendants of an ancient organism can be inspected, a rule of "parsimony" can be used to model the biochemistry of the ancestral organism by extrapolation from the biochemistry of the descendant organisms. The most parsimonious model is one that explains the diversity in the modern descendants by a minimum number of independent evolutionary events. For the progenote, three independent lineages of descendants are known (archaebacteria, eubacteria, and eukaryotes). Thus, a biochemical trait present in all three can be assigned to the progenote. The assignment is strongest when (*i*) the trait is found in several representative organisms from each of the three kingdoms; (*ii*) assignments of homology in various branches of the progenotic pedigree are supported by high information content (preferably sequence data); and (*iii*) aspects of the trait serve no selected function in the modern world.[b] Such assignments are not absolute; if only some criteria are fulfilled, a weaker assignment can be proposed.

From the Schoyen Collection, Section 5 (Patristic literature)
http://www.schoyencollection.com/patristic.htm

Where do amino acids come from?



An orthodox view of early evolution…

(why do we think that amino acids were around from the beginning?)

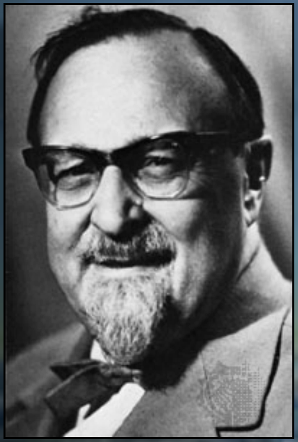
Amino acids predate life...

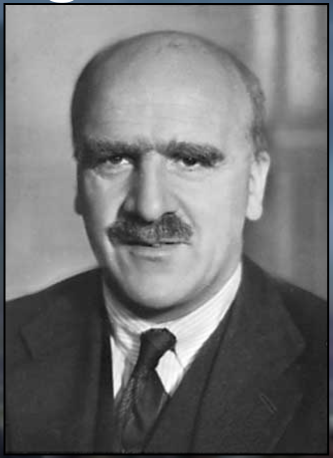
Nature (1996) 384:55-9. "**Evidence for life on Earth before 3,800 million years ago**", *Eiler JM, Mojzsis SJ, Arrhenius G.*

"It is unknown when life first appeared on Earth…Here we … provide evidence for the emergence of life on Earth by at least 3,800 Myr before present."

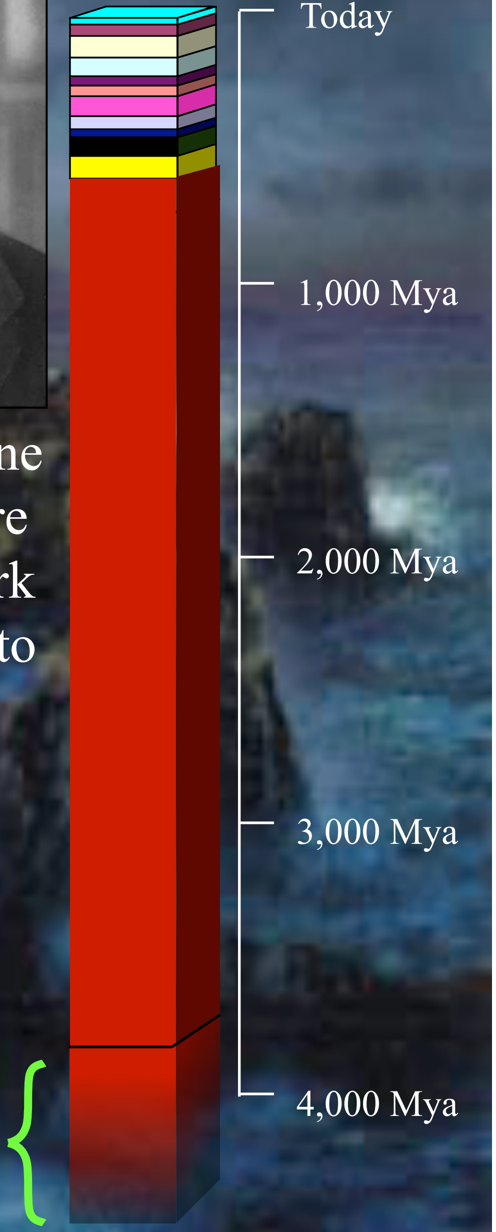
Half-ish of the 20 'modern' amino acids existed <u>before</u> life evolved

“Prebiotic Soup” and the Origin of Life...

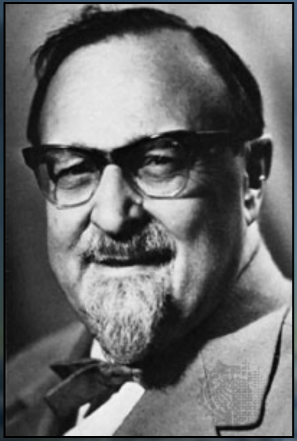

Aleksandr Oparin
(in 1924)


J.B.S. Haldane
(1929, before
Oparin's work
translated into
English)

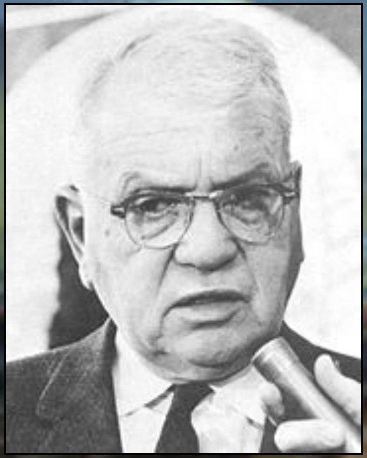
…proposed that the primordial sea served as a vast chemical laboratory powered by solar energy (ultra violet radiation) acting on an oxygen free atmosphere of carbon dioxide and ammonia: became a prebiotic soup containing large populations of organic monomers and polymers



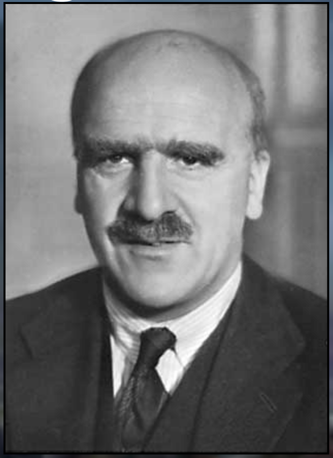
"Prebiotic Soup" and the Origin of Life...



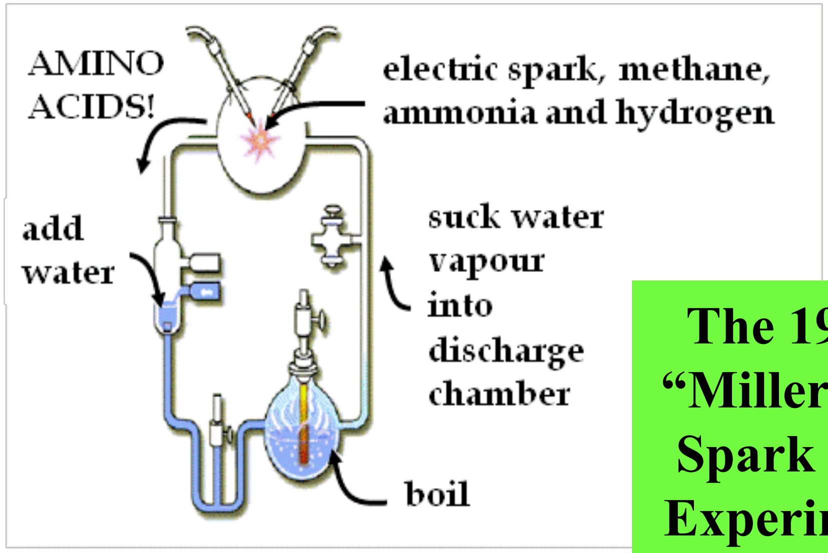
Aleksandr Oparin



Harold Urey



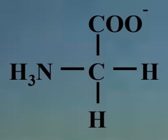
J.B.S. Haldane



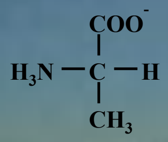
The 1950's "Miller/Urey Spark Tube Experiments

…spawning a huge literature of "prebiotic syntheses"

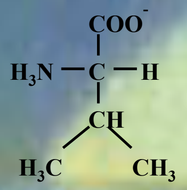
Glycine (Gly)



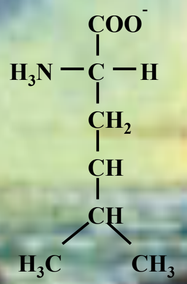
Alanine (Ala)



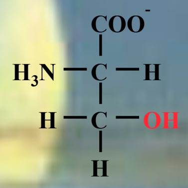
Valine (Val)



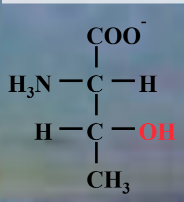
Leucine (Leu)



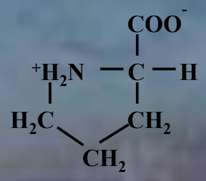
Serine (Ser)



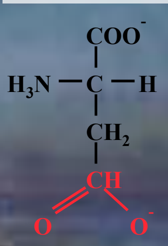
Threonine (Thr)



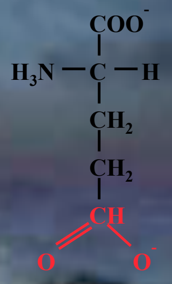
Proline (Pro)



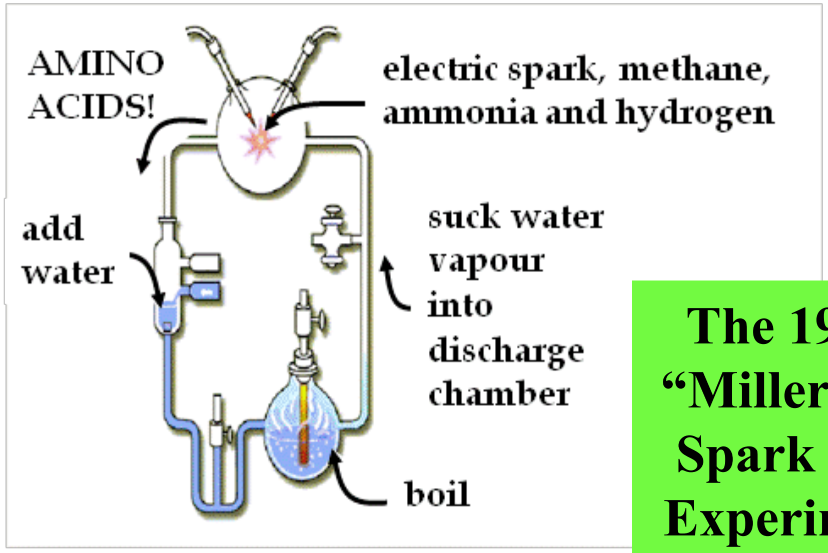
Aspartate (Asp)



Glutamate (Glu)



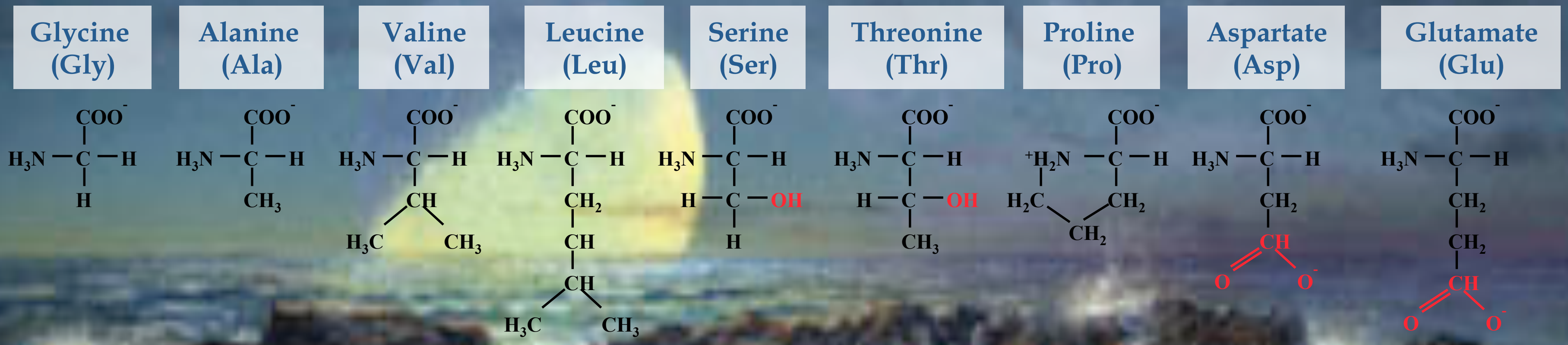
"endogenous" sources of abiotic amino acids: atmosphere, oceans, terrestrial ponds, evaporating shores, hot streams, cold streams etc.

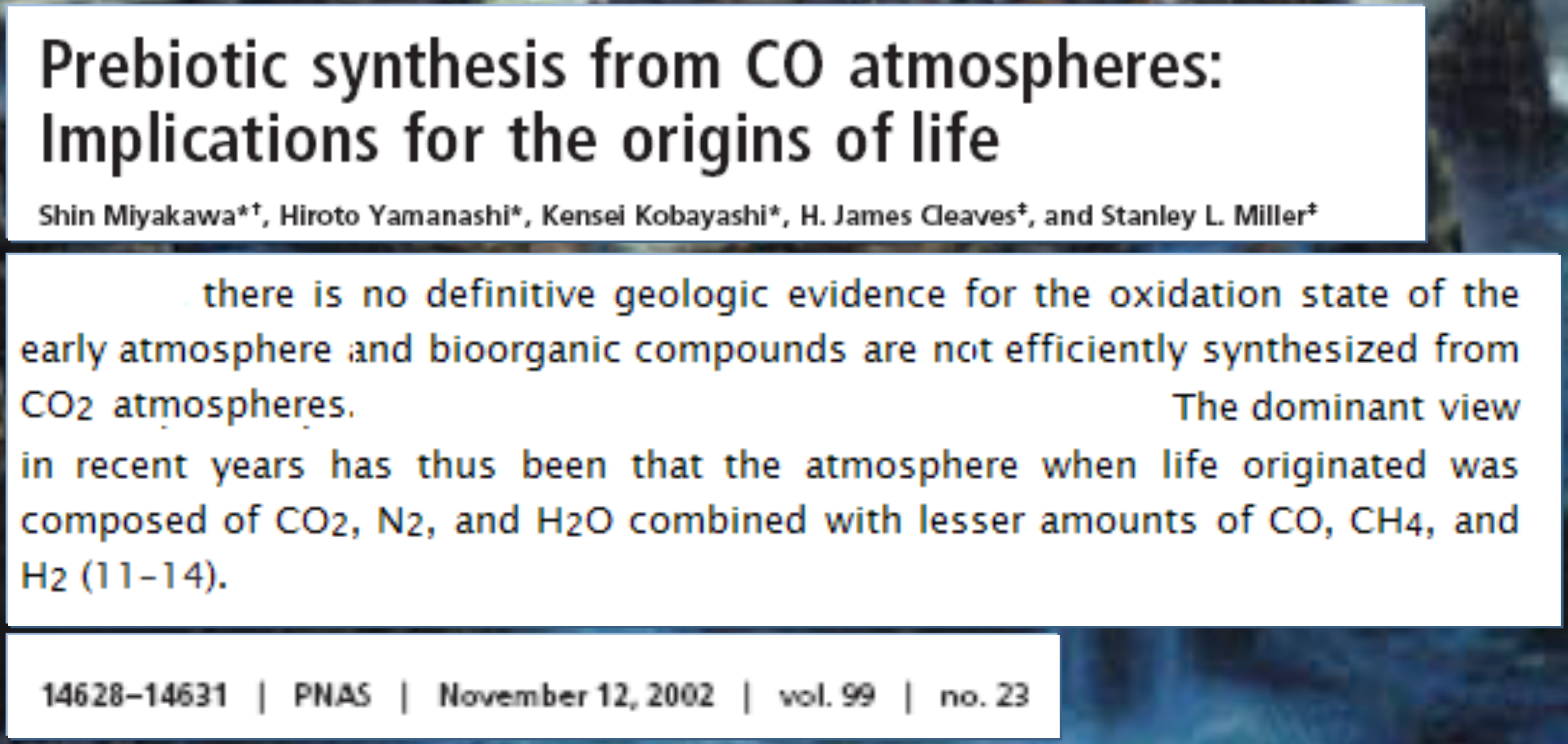


The 1950's "Miller/Urey Spark Tube Experiments

(e.g. Miller, 1953; Harada and Fox, 1964; Sagan and Khare, 1971; Wolman, et al., 1972; Schlesinger and Miller, 1983; Kobayashi and Ponnamperuma, 1985; Weber, 1995; Nelson et al., 2001; Plankensteiner et al., 2004; Tian et al., 2005; Russell, 2007)

A "strongly reducing" early atmosphere now seems unlikely





Prebiotic synthesis from CO atmospheres: Implications for the origins of life

Shin Miyakawa*†, Hiroto Yamanashi*, Kensei Kobayashi*, H. James Cleaves‡, and Stanley L. Miller‡

there is no definitive geologic evidence for the oxidation state of the early atmosphere and bioorganic compounds are not efficiently synthesized from CO_2 atmospheres. The dominant view in recent years has thus been that the atmosphere when life originated was composed of CO_2, N_2, and H_2O combined with lesser amounts of CO, CH_4, and H_2 (11–14).

14628–14631 | PNAS | November 12, 2002 | vol. 99 | no. 23

"we are stardust; we are golden. we are billion year old carbon"

Joni Mitchell , "Woodstock"

Inside a star, gravitational forces condense hydrogen and nuclear fusion begins: hydrogen is fused into helium, which in turn is converted into carbon and onwards into oxygen, neon, magnesium, silicon, sulfur and iron …When the star exhausts its "fuel" the core can collapse resulting in a spectacular explosion, producing a gas-cloud that expands at a rate of about 10,000 km/s. The cloud slows and dissipates, seeding the inter-stellar neighborhood with heavy elements and providing the shock waves that initiate new star formation.

www.novaspace.com/GICLEE/Poor/Supernova.html

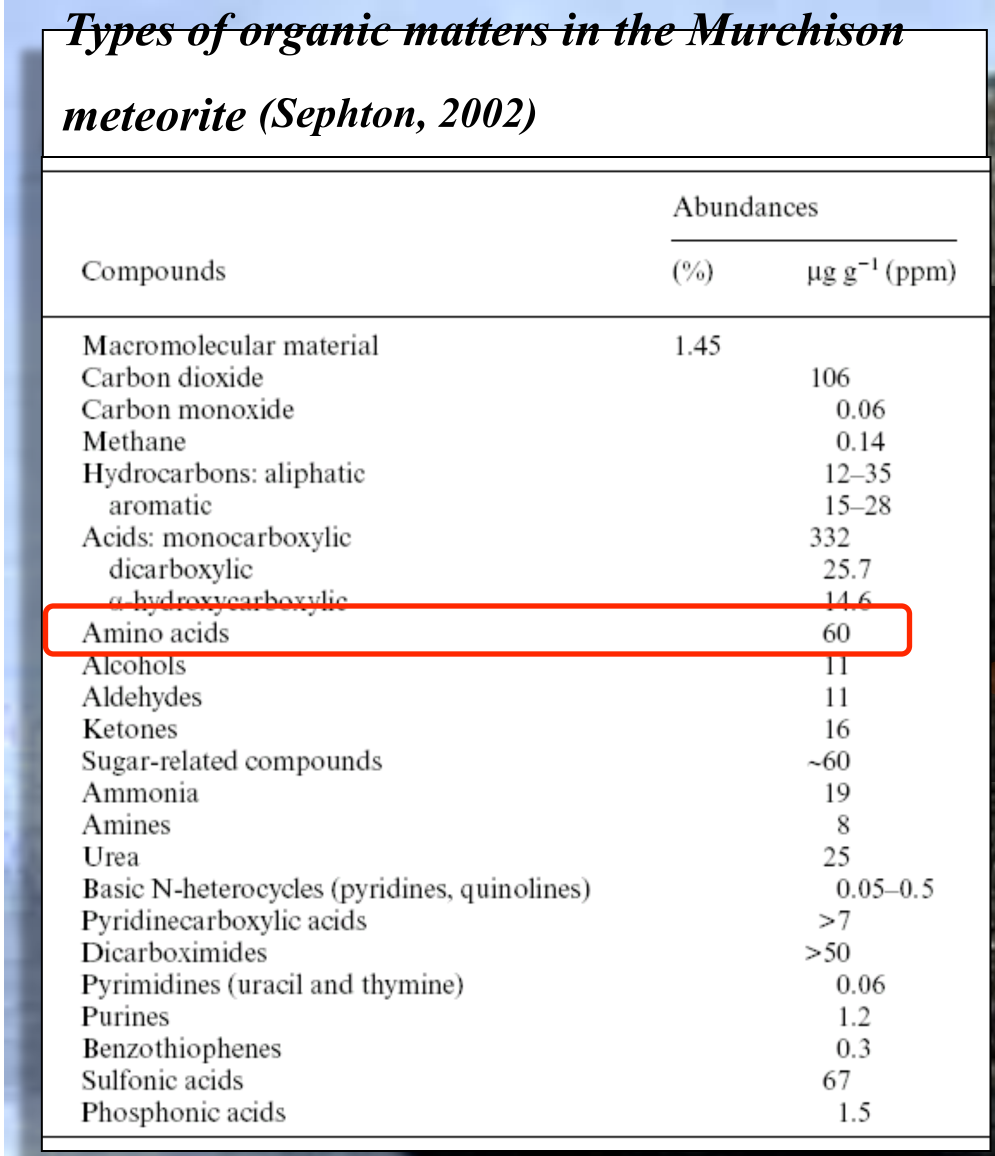
…amino acid delivery from space seems *very* likely!

The Earth acquires 100,000 to 1,000,000 kg of extraterrestrial material each day…

Ehrenfreund P, Sephton MA, "Carbon molecules in space: from astrochemistry to astrobiology." Faraday Discuss. 2006;133:277-88; discussion 347-74, 449-52

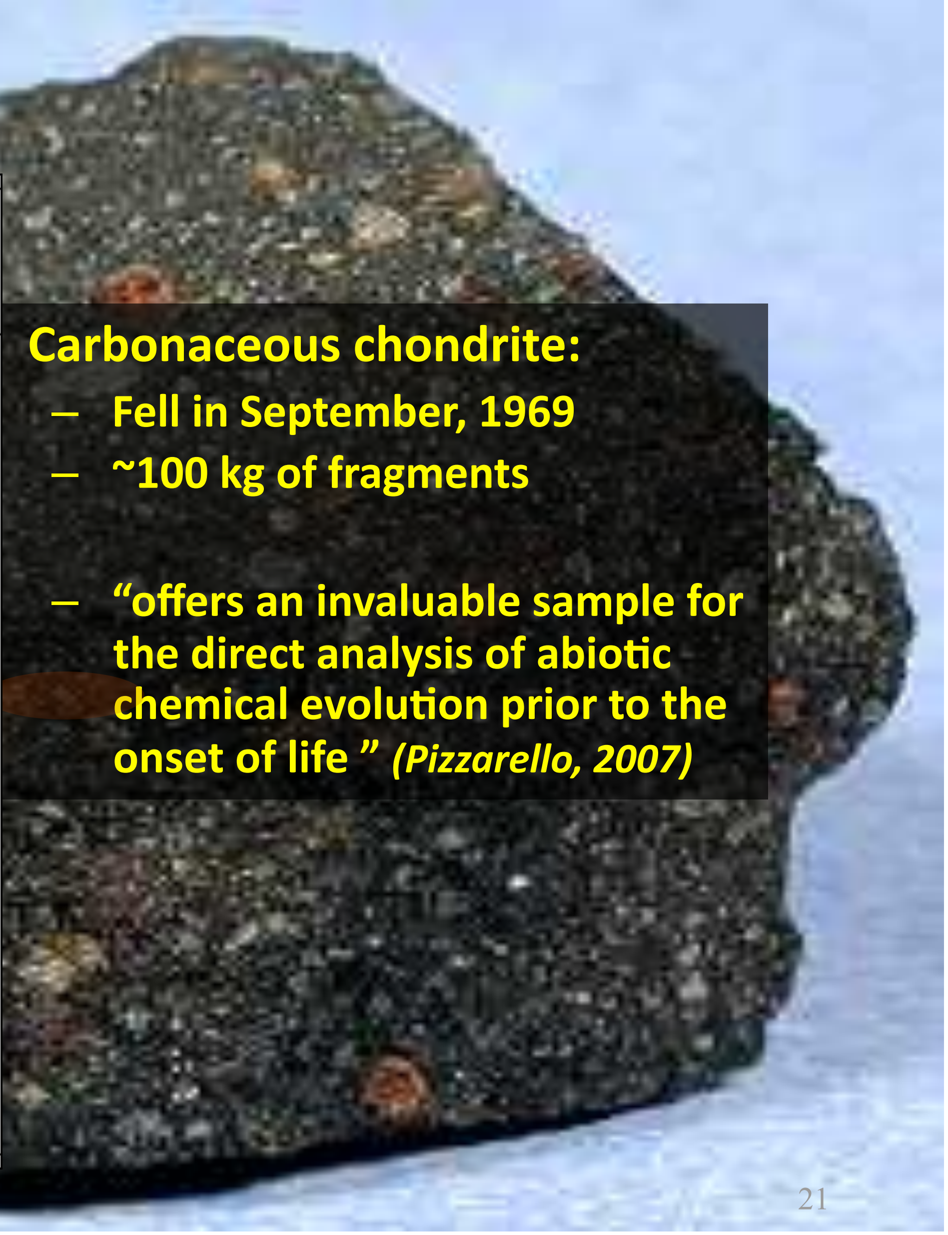
http://www.nightskyinfo.com/maps_images/meteor_tezel.jpg

The Murchison Meteorite



Types of organic matters in the Murchison meteorite (Sephton, 2002)

Compounds	Abundances (%)	Abundances µg g^{-1} (ppm)
Macromolecular material	1.45	
Carbon dioxide		106
Carbon monoxide		0.06
Methane		0.14
Hydrocarbons: aliphatic		12–35
aromatic		15–28
Acids: monocarboxylic		332
dicarboxylic		25.7
α-hydroxycarboxylic		14.6
Amino acids		60
Alcohols		11
Aldehydes		11
Ketones		16
Sugar-related compounds		~60
Ammonia		19
Amines		8
Urea		25
Basic N-heterocycles (pyridines, quinolines)		0.05–0.5
Pyridinecarboxylic acids		>7
Dicarboximides		>50
Pyrimidines (uracil and thymine)		0.06
Purines		1.2
Benzothiophenes		0.3
Sulfonic acids		67
Phosphonic acids		1.5

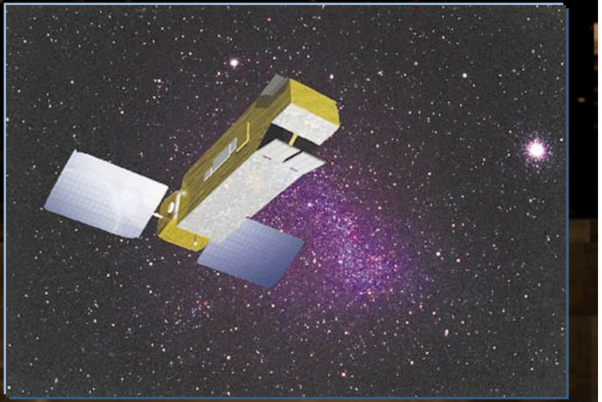


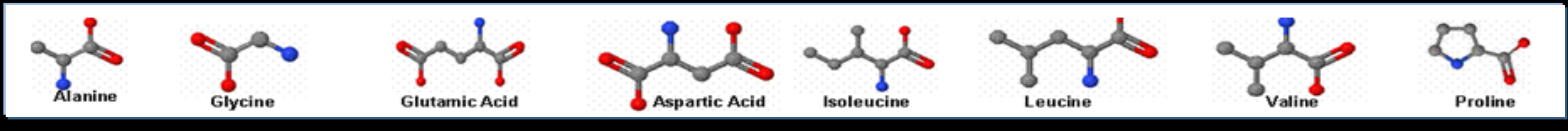
Carbonaceous chondrite:

- **Fell in September, 1969**
- **~100 kg of fragments**

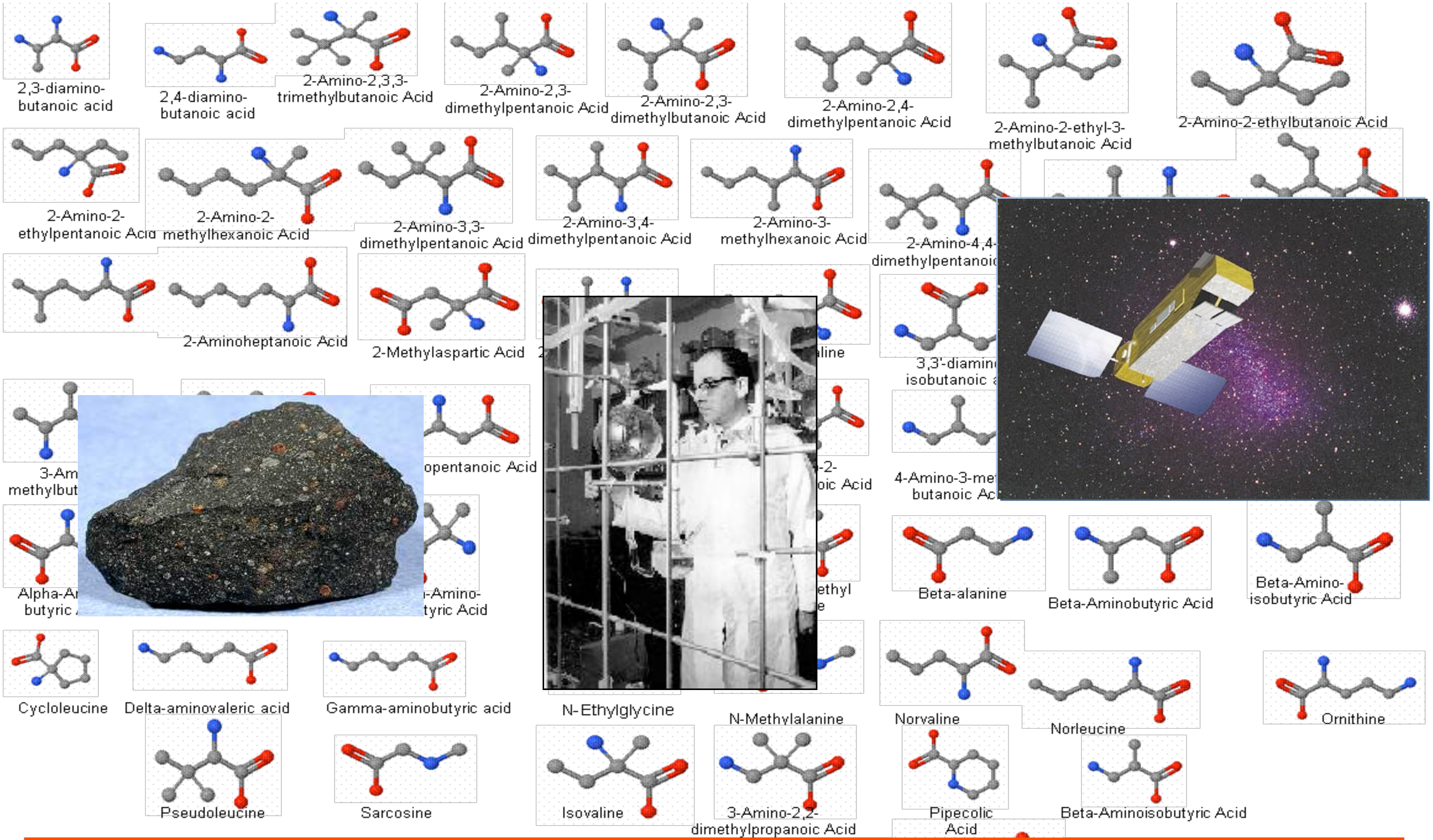
- **"offers an invaluable sample for the direct analysis of abiotic chemical evolution prior to the onset of life "** ***(Pizzarello, 2007)***

Some amino acids are prebiotically plausible!





…but this includes MANY "ignored" by biology



2,3-diamino-butanoic acid

2,4-diamino-butanoic acid

2-Amino-2,3,3-trimethylbutanoic Acid

2-Amino-2,3-dimethylpentanoic Acid

2-Amino-2,3-dimethylbutanoic Acid

2-Amino-2,4-dimethylpentanoic Acid

2-Amino-2-ethyl-3-methylbutanoic Acid

2-Amino-2-ethylbutanoic Acid

2-Amino-2-ethylpentanoic Acid

2-Amino-2-methylhexanoic Acid

2-Amino-3,3-dimethylpentanoic Acid

2-Amino-3,4-dimethylpentanoic Acid

2-Amino-3-methylhexanoic Acid

2-Amino-4,4-dimethylpentanoi

2-Aminoheptanoic Acid

2-Methylaspartic Acid

2

aline

3,3'-diamin
isobutanoic a

3-Am
methylbut

opentanoic Acid

-2-
oic Acid

4-Amino-3-me
butanoic Ac

Alpha-Ar
butyric /

-Amino-
tyric Acid

ethyl
e

Beta-alanine

Beta-Aminobutyric Acid

Beta-Amino-isobutyric Acid

Cycloleucine

Delta-aminovaleric acid

Gamma-aminobutyric acid

N-Ethylglycine

N-Methylalanine

Norvaline

Norleucine

Ornithine

Pseudoleucine

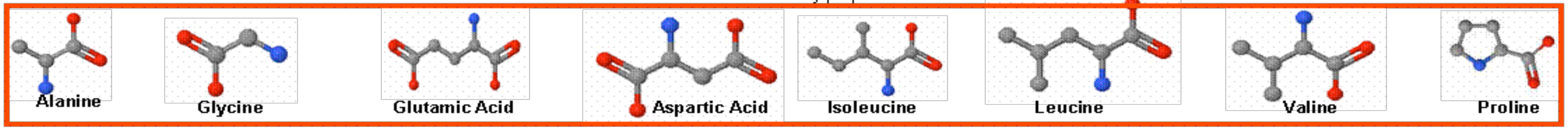
Sarcosine

Isovaline

3-Amino-2,2-dimethylpropanoic Acid

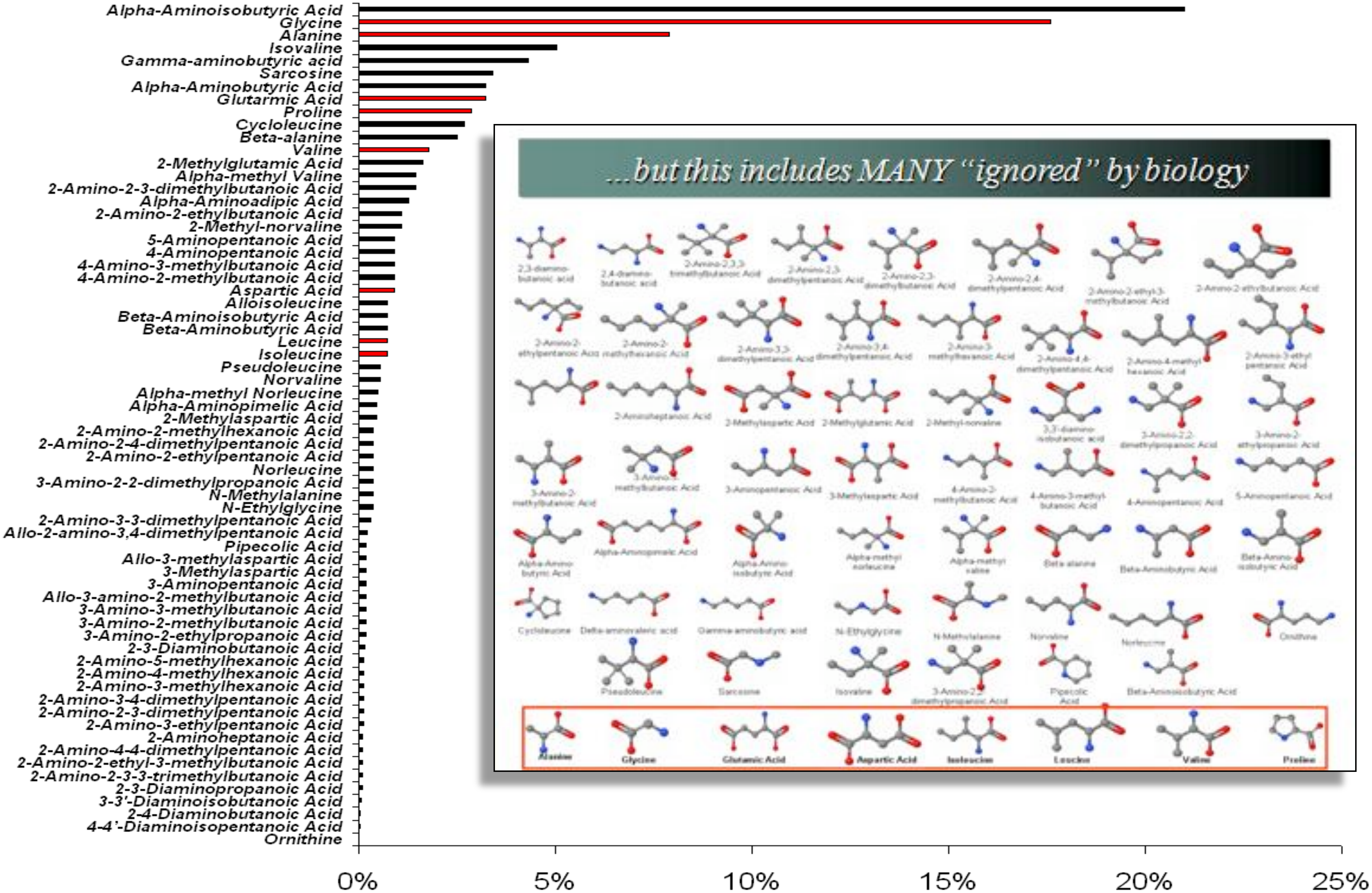
Pipecolic Acid

Beta-Aminoisobutyric Acid



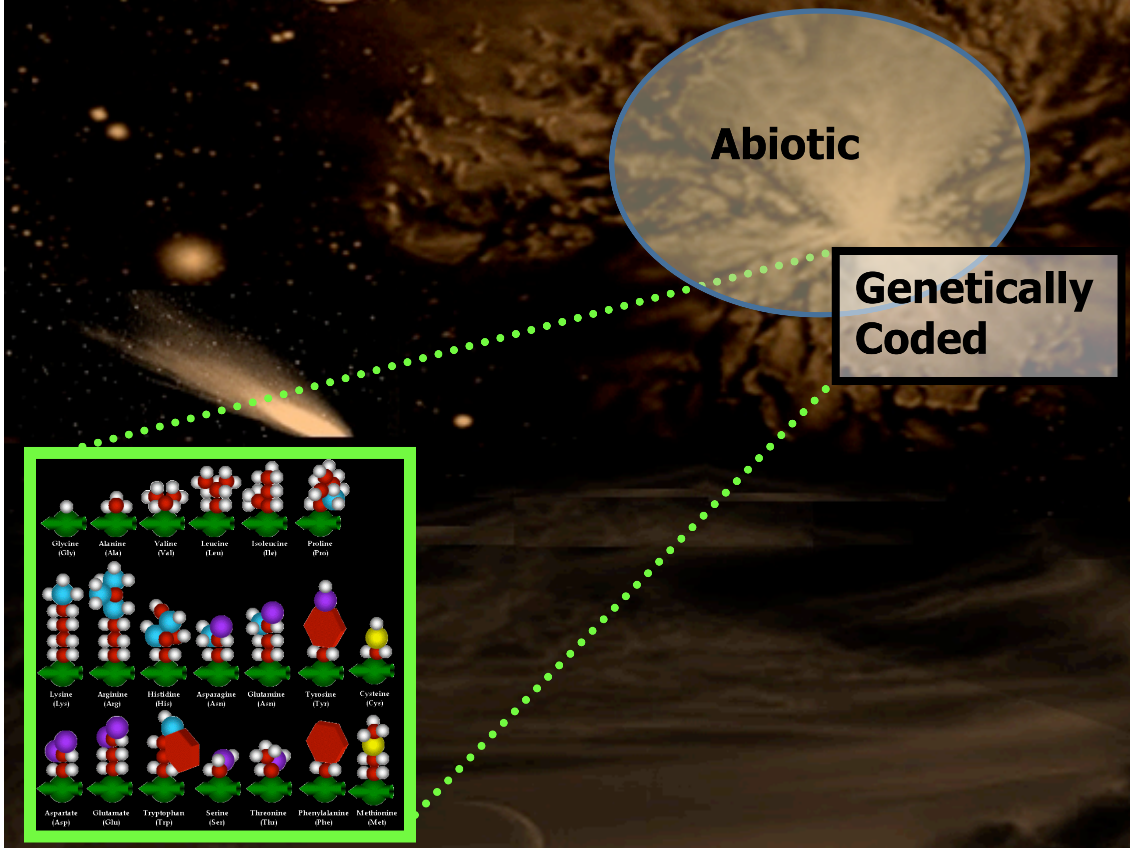
Alanine **Glycine** **Glutamic Acid** **Aspartic Acid** **Isoleucine** **Leucine** **Valine** **Proline**

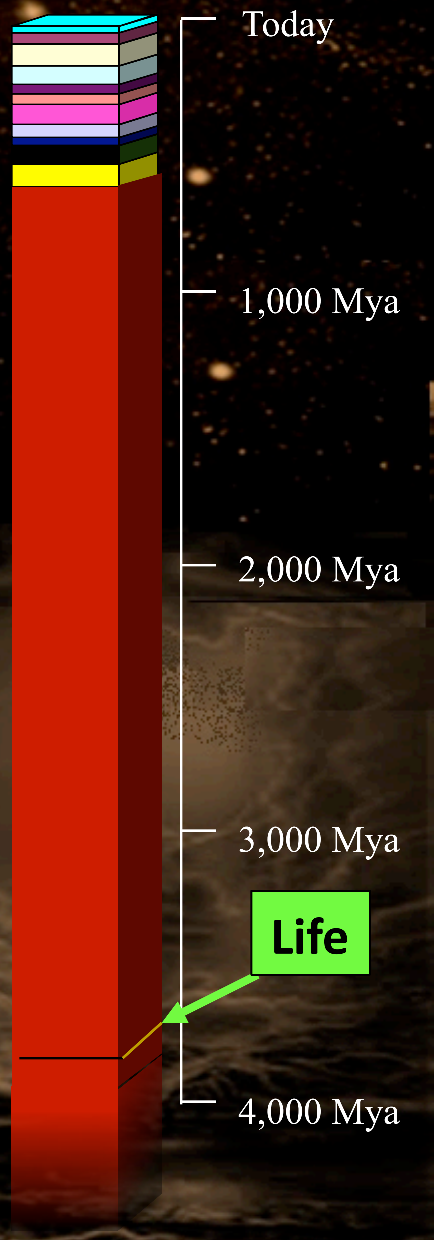
◆ The biological amino acids are not obvious choices



Abundance of each amino acid within the Murchison meteorite, as a proportion of the total amount of the 66 amino acids found there. "Biological" amino acids are shown in red *(courtesy Pizarello)*

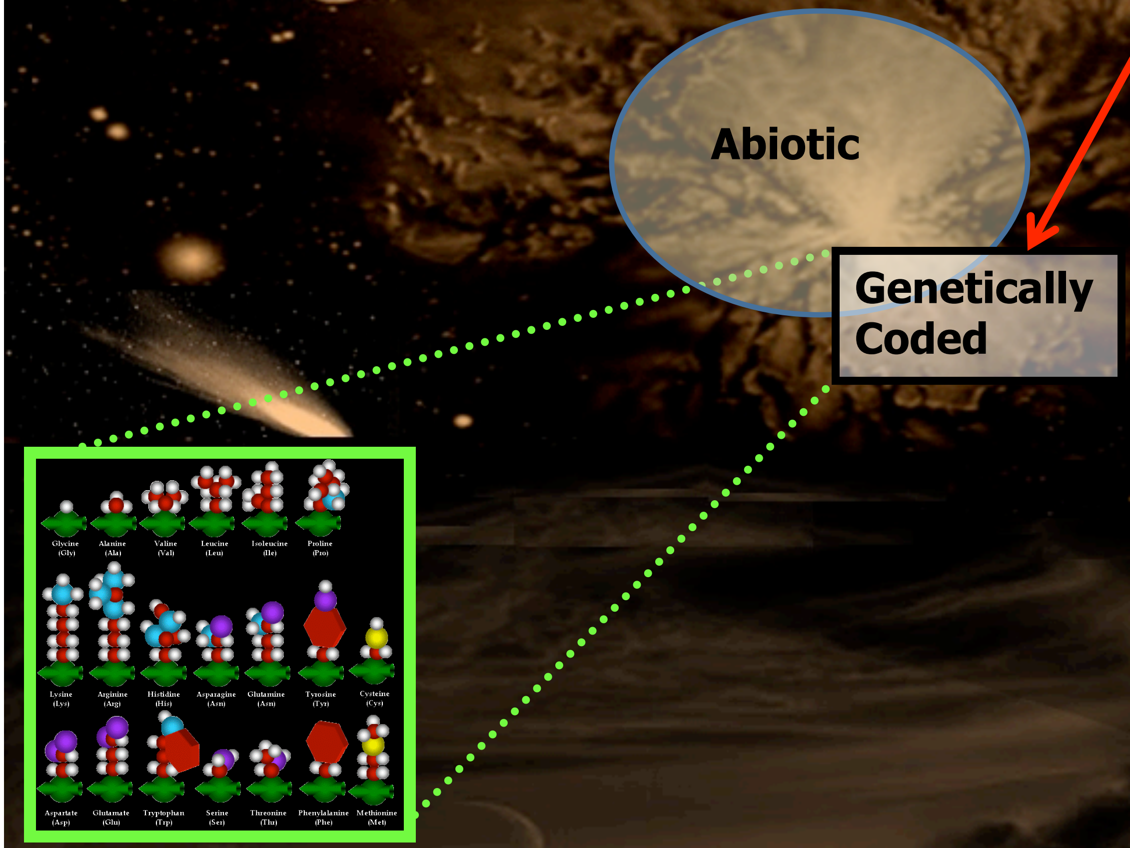
The genetic code uses a small fraction of prebiotically plausible amino acids

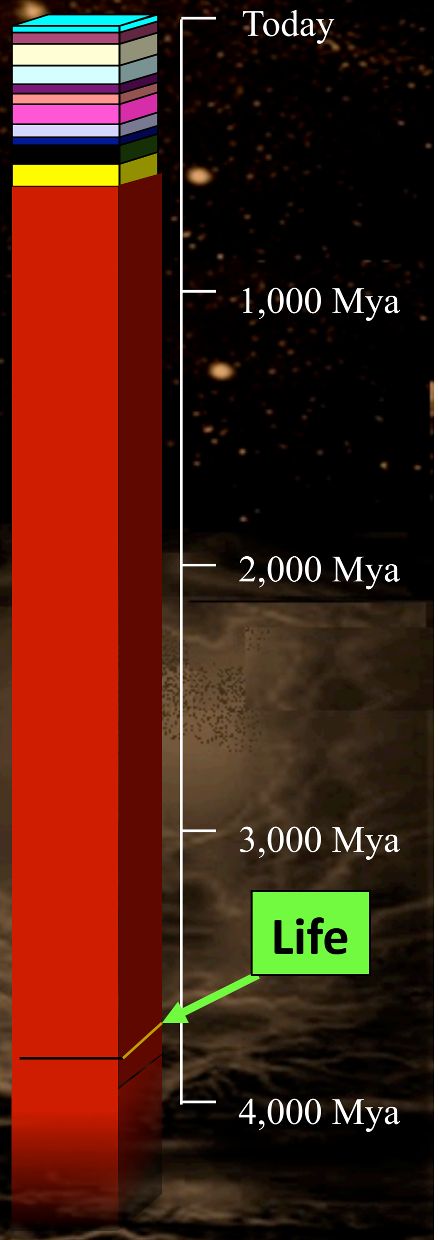




Half-ish of the 20 ‘modern’ amino acids existed __before__ life evolved

Meanwhile, many genetically encoded amino acids are not prebiotically plausible

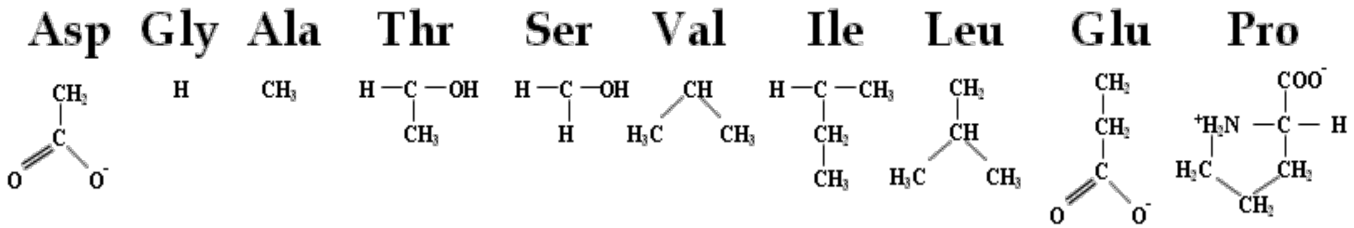




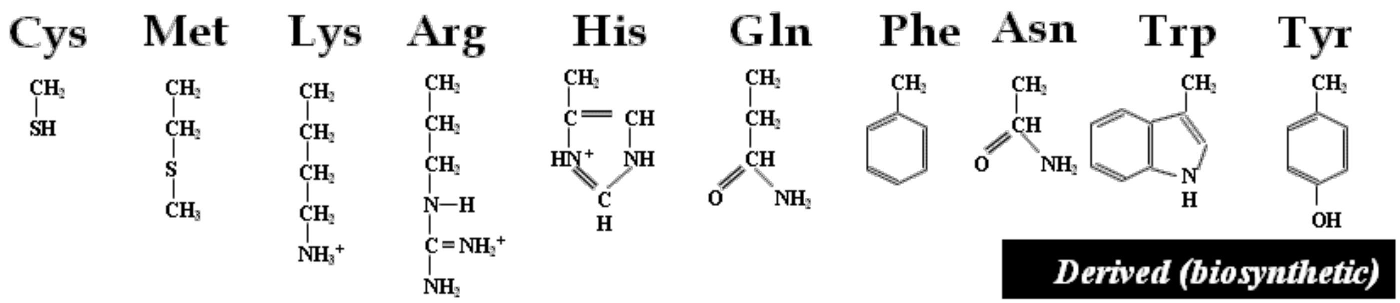
Half-ish of the 20 'modern' amino acids existed before life evolved

“early” versus “late” amino acids

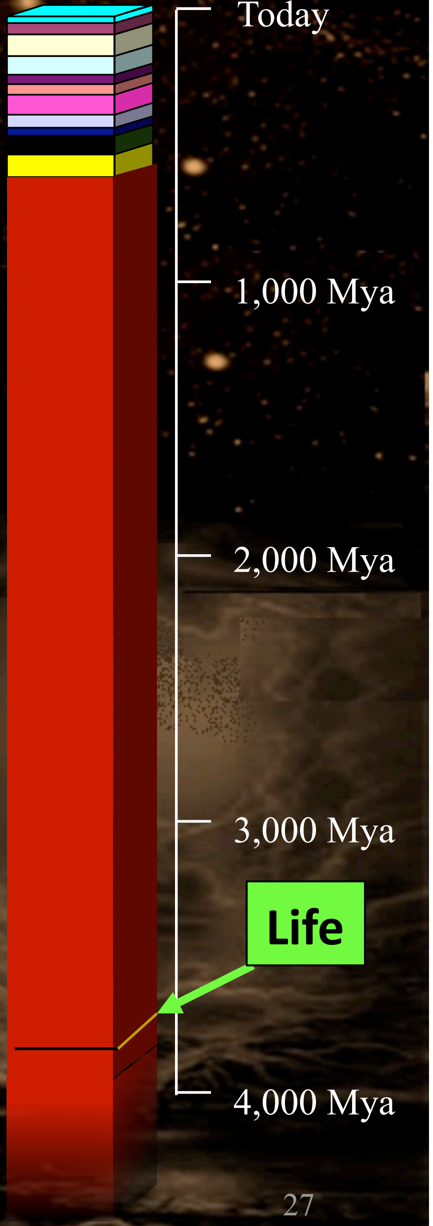
Primordial (abiotic)



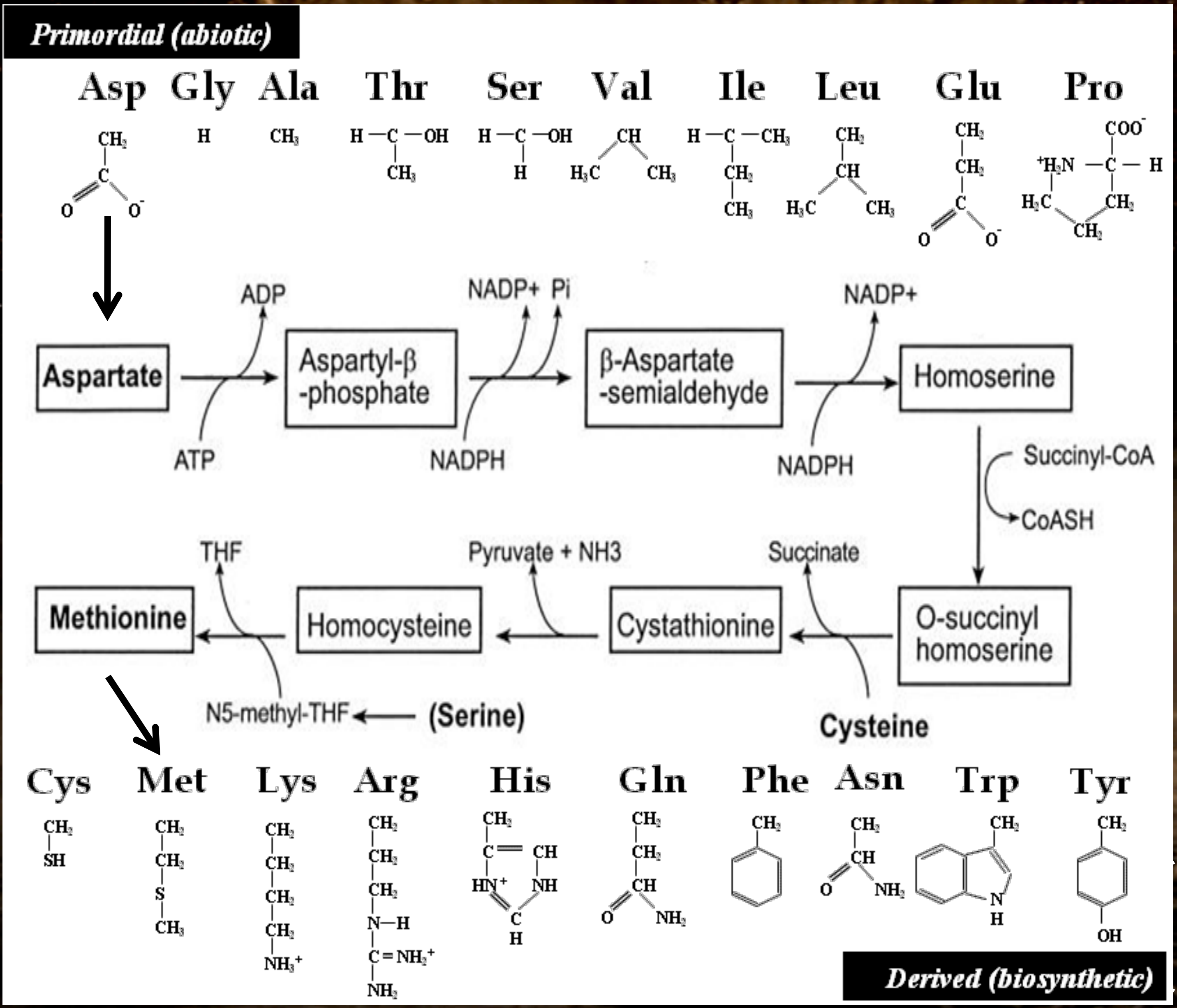
Genetically Coded



Derived (biosynthetic)



“early” versus “late” amino acids…connected by biosynthesis



Primordial (abiotic)

Asp Gly Ala Thr Ser Val Ile Leu Glu Pro

Aspartate

ATP

ADP

Aspartyl-β -phosphate

NADPH

NADP+ Pi

β-Aspartate -semialdehyde

NADPH

NADP+

Homoserine

Succinyl-CoA

CoASH

O-succinyl homoserine

Cysteine

Succinate

Cystathionine

Pyruvate + NH3

Homocysteine

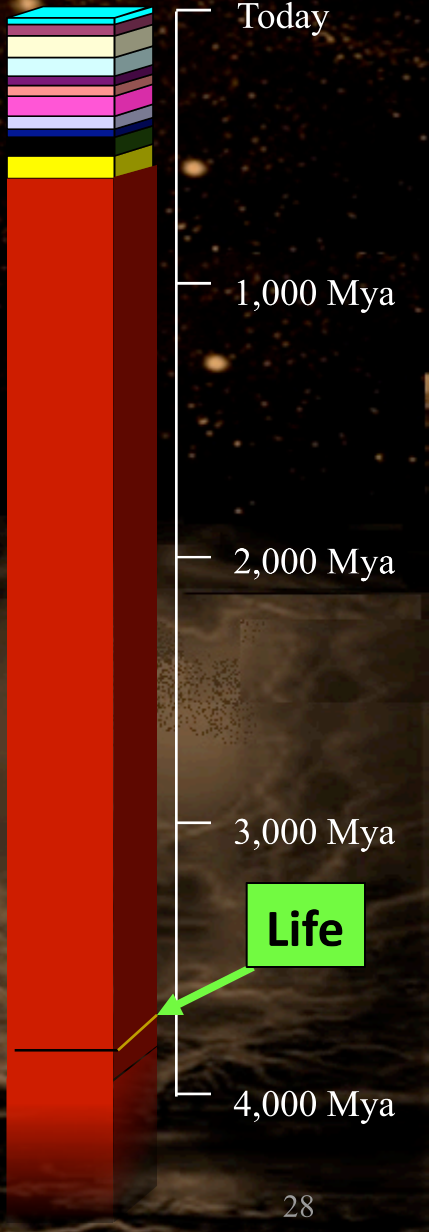
N5-methyl-THF ← (Serine)

THF

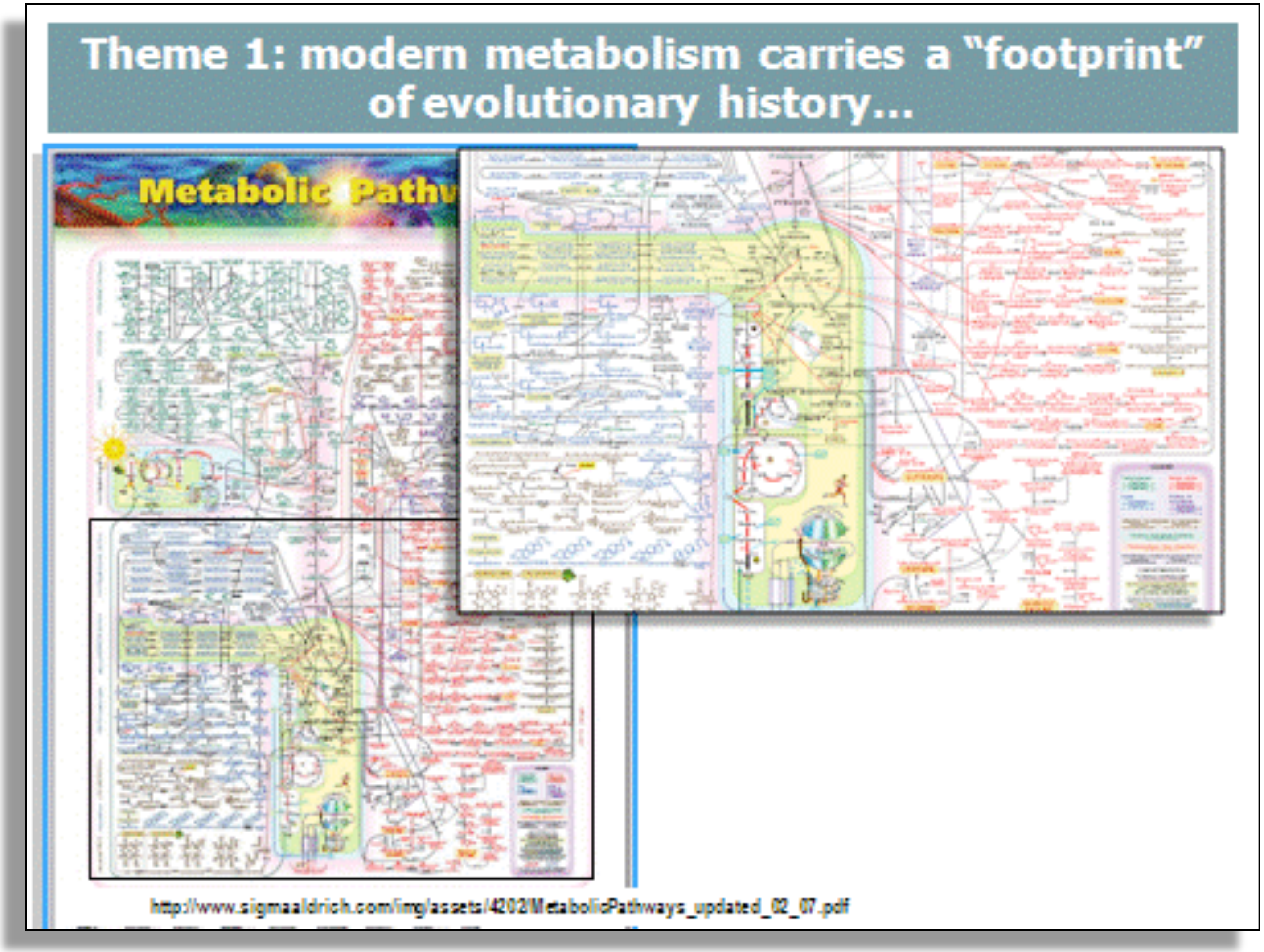
Methionine

Cys Met Lys Arg His Gln Phe Asn Trp Tyr

Derived (biosynthetic)



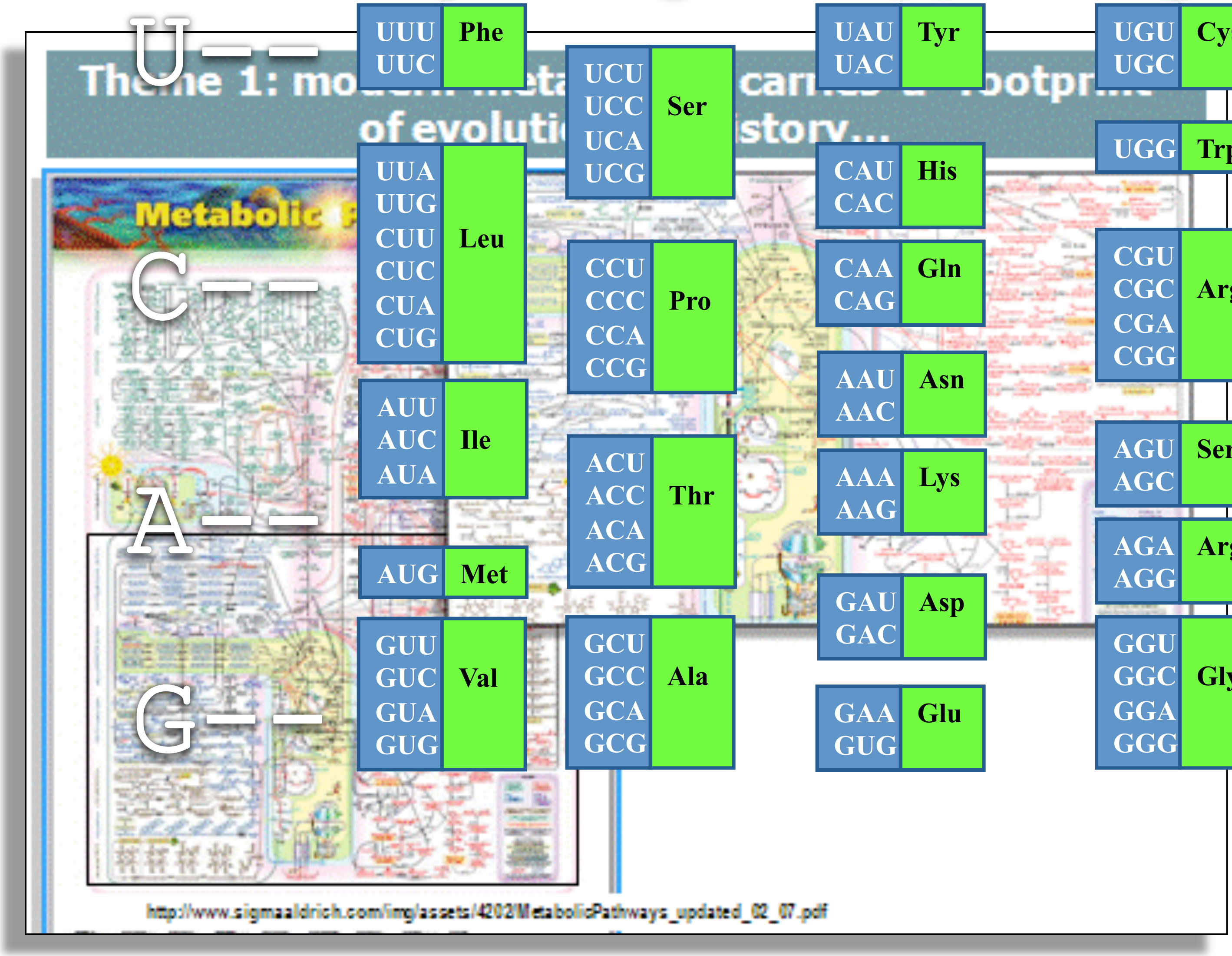
Amino acid metabolism & codon assignments



Theme 1: modern metabolism carries a "footprint" of evolutionary history...

http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

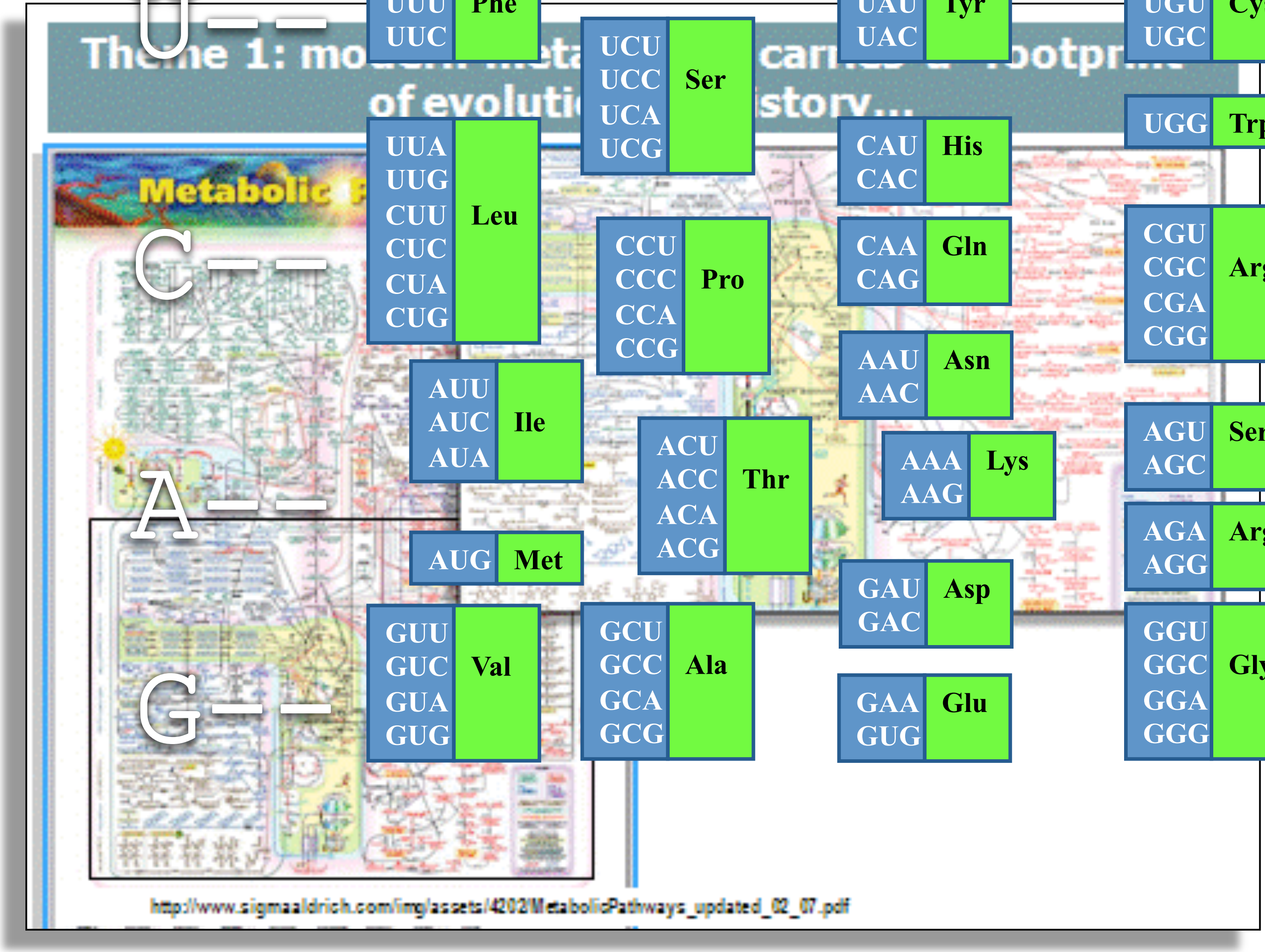
Amino acid metabolism & codon assignments



http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

	–U–	–C–	–A–	–G–	
U– –	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	– –U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	– –C
	UUA Leu	UCA Ser			– –A
	UUG Leu	UCG Ser		UGG Trp	– –G
C– –	CUU Leu	CCU Pro	CAU His	CGU Arg	– –U
	CUC Leu	CCC Pro	CAC His	CGC Arg	– –C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	– –A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	– –G
A– –	AUU Ile	ACU Thr	AAU Asn	AGU Ser	– –U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	– –C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	– –A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	– –G
G– –	GUU Val	GCU Ala	GAU Asp	GGU Gly	– –U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	– –C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	– –A
	GUG Val	GCG Ala	GUG Glu	GGG Gly	– –G

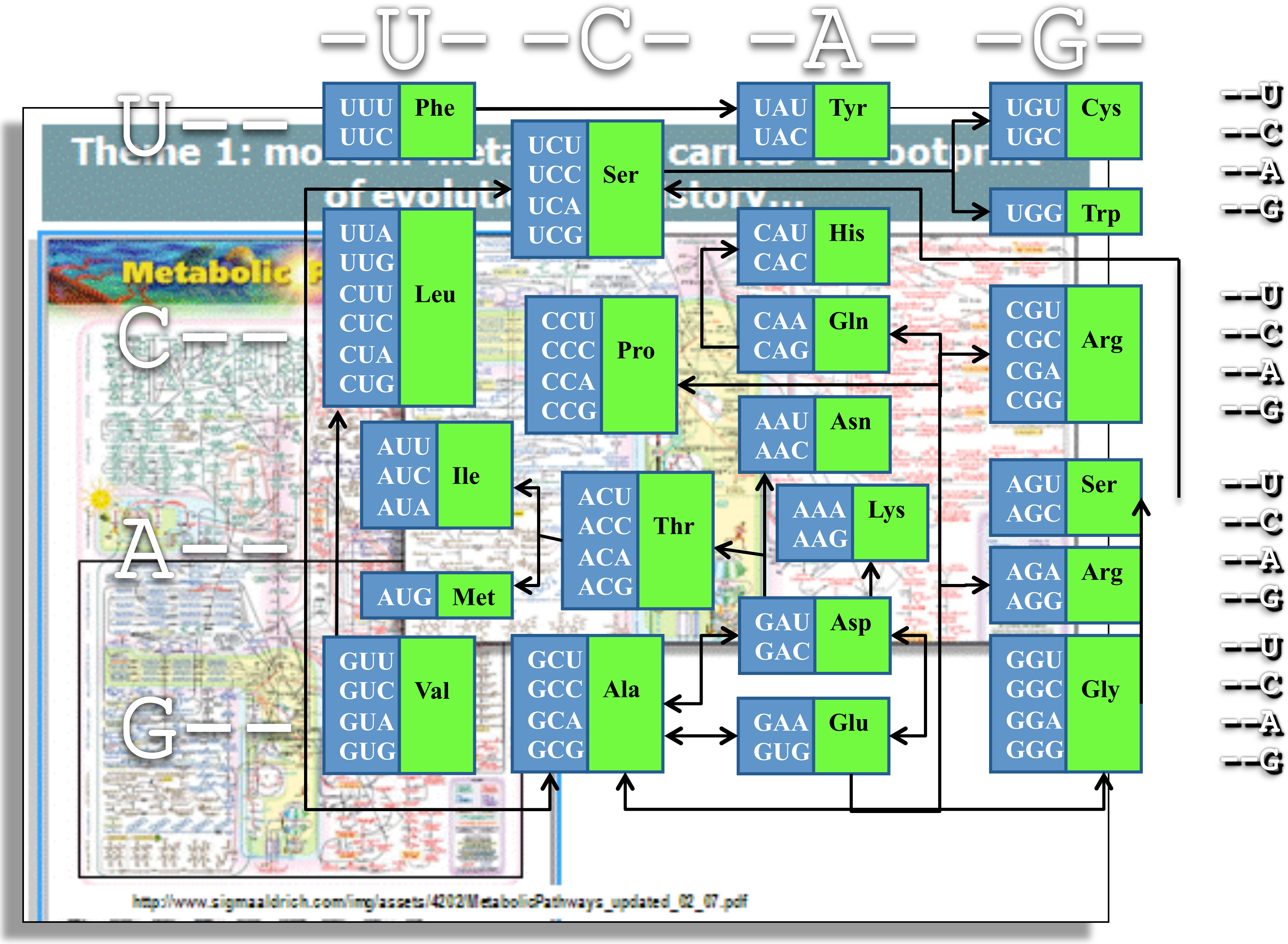
Amino acid metabolism & codon assignments



	–U–		–C–		–A–		–G–		
U––	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	––U
	UUC		UCC		UAC		UGC		––C
	UUA	Leu	UCA						––A
	UUG		UCG				UGG	Trp	––G
C––	CUU		CCU	Pro	CAU	His	CGU	Arg	––U
	CUC		CCC		CAC		CGC		––C
	CUA		CCA		CAA	Gln	CGA		––A
	CUG		CCG		CAG		CGG		––G
A––	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	––U
	AUC		ACC		AAC		AGC		––C
	AUA		ACA		AAA	Lys	AGA	Arg	––A
	AUG	Met	ACG		AAG		AGG		––G
G––	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	––U
	GUC		GCC		GAC		GGC		––C
	GUA		GCA		GAA	Glu	GGA		––A
	GUG		GCG		GUG		GGG		––G

http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

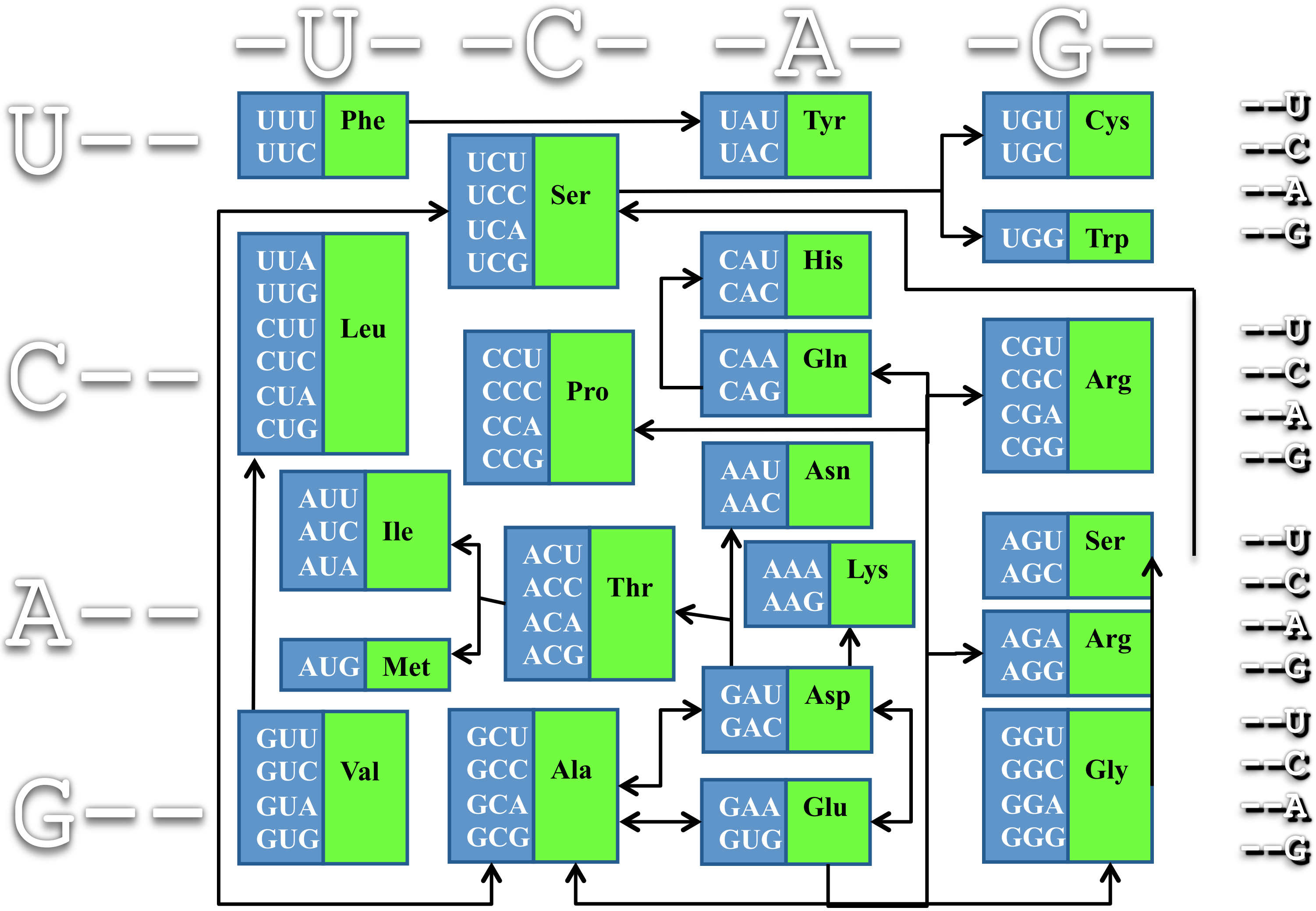
Amino acid metabolism & codon assignments



	–U–	–C–	–A–	–G–	
U– –	UUU UUC Phe UUA UUG CUU CUC CUA CUG Leu	UCU UCC UCA UCG Ser	UAU UAC Tyr	UGU UGC Cys UGG Trp	––U ––C ––A ––G
C– –		CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	––U ––C ––A ––G
A– –	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg	––U ––C ––A ––G
G– –	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU GAC Asp GAA GUG Glu	GGU GGC GGA GGG Gly	––U ––C ––A ––G

http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

Amino acid metabolism & codon assignments



–U– –C– –A– –G–

U– –

UUU UUC **Phe**

UCU UCC UCA UCG **Ser**

UAU UAC **Tyr**

UGU UGC **Cys**

UGG **Trp**

C– –

UUA UUG CUU CUC CUA CUG **Leu**

CCU CCC CCA CCG **Pro**

CAU CAC **His**

CAA CAG **Gln**

CGU CGC CGA CGG **Arg**

A– –

AUU AUC AUA **Ile**

AUG **Met**

ACU ACC ACA ACG **Thr**

AAU AAC **Asn**

AAA AAG **Lys**

AGU AGC **Ser**

AGA AGG **Arg**

G– –

GUU GUC GUA GUG **Val**

GCU GCC GCA GCG **Ala**

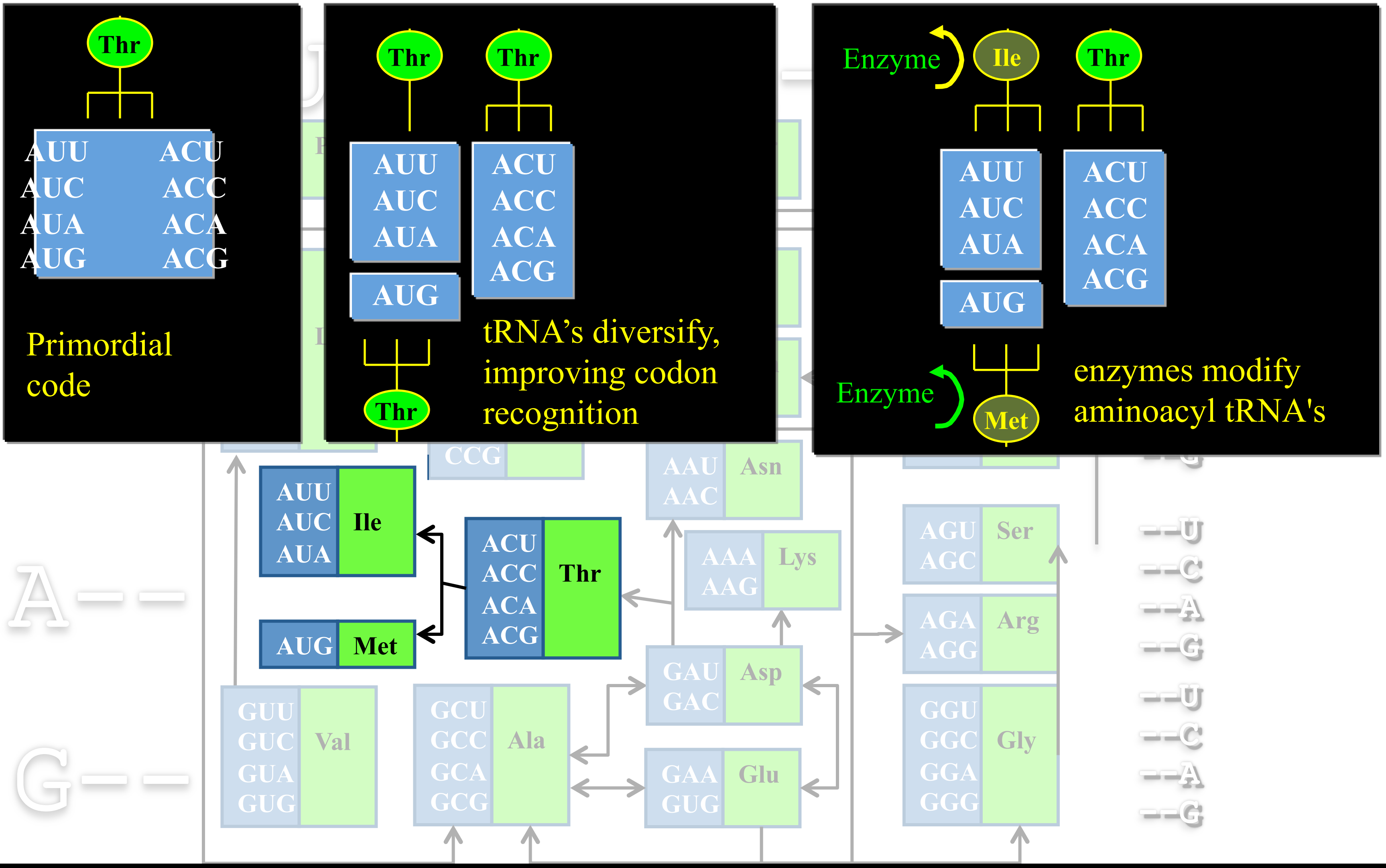
GAU GAC **Asp**

GAA GUG **Glu**

GGU GGC GGA GGG **Gly**

– –U
– –C
– –A
– –G

Amino acid metabolism & codon assignments



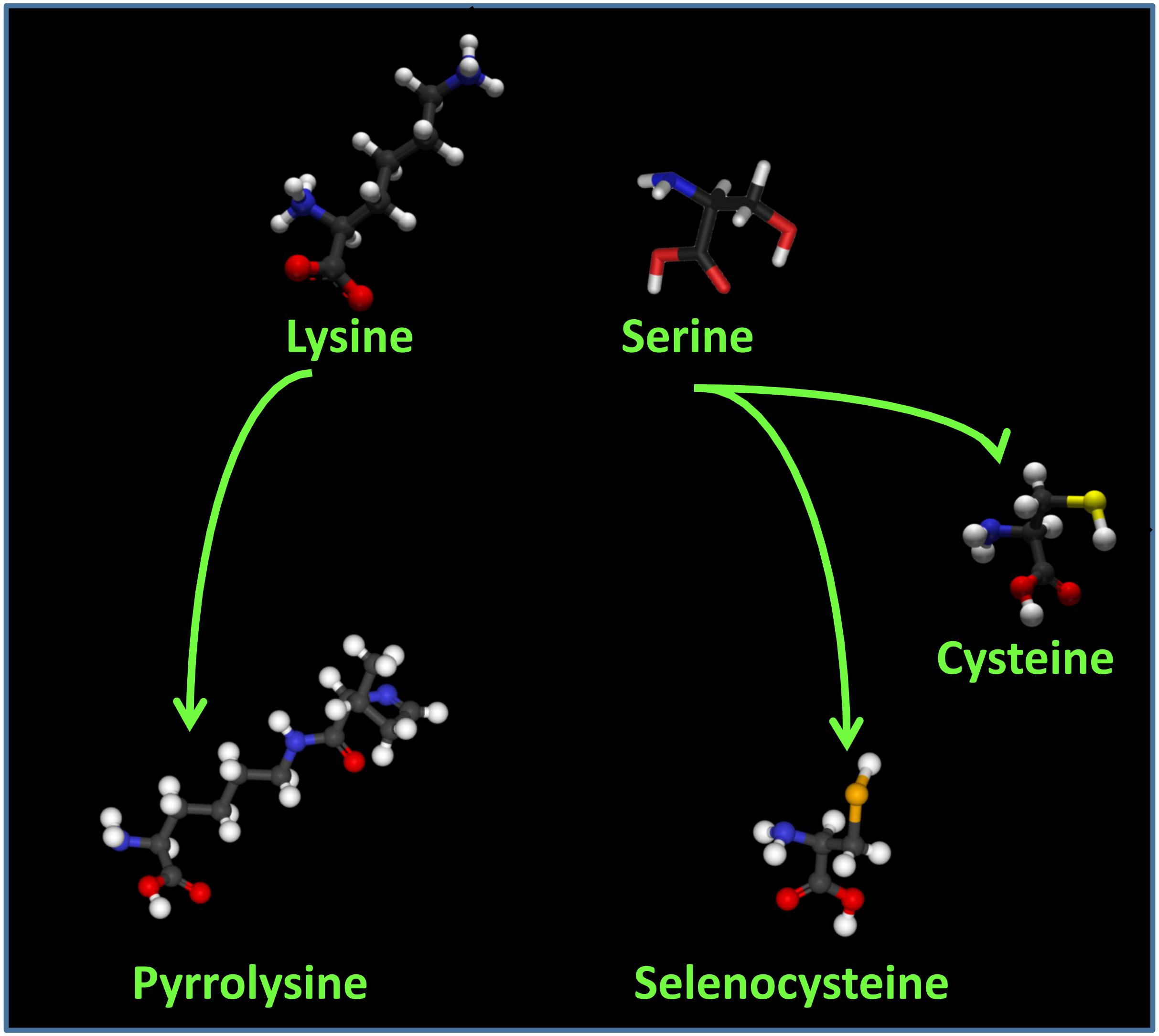
Primordial code

tRNA's diversify, improving codon recognition

enzymes modify aminoacyl tRNA's

amino acid biosynthetic derivatives may have become incorporated into the code by modifying precursor amino acids on the tRNA ~'novel' amino acids would thus capture a sub-set of the codons belonging to their pre-cursors

The "21st and 22nd amino acids":



We see something very much like this hypothesized *genetic code expansion* in lineages that have "recently" evolved to encode pyrrolysine and selenocysteine

Growth of the amino acid alphabet...

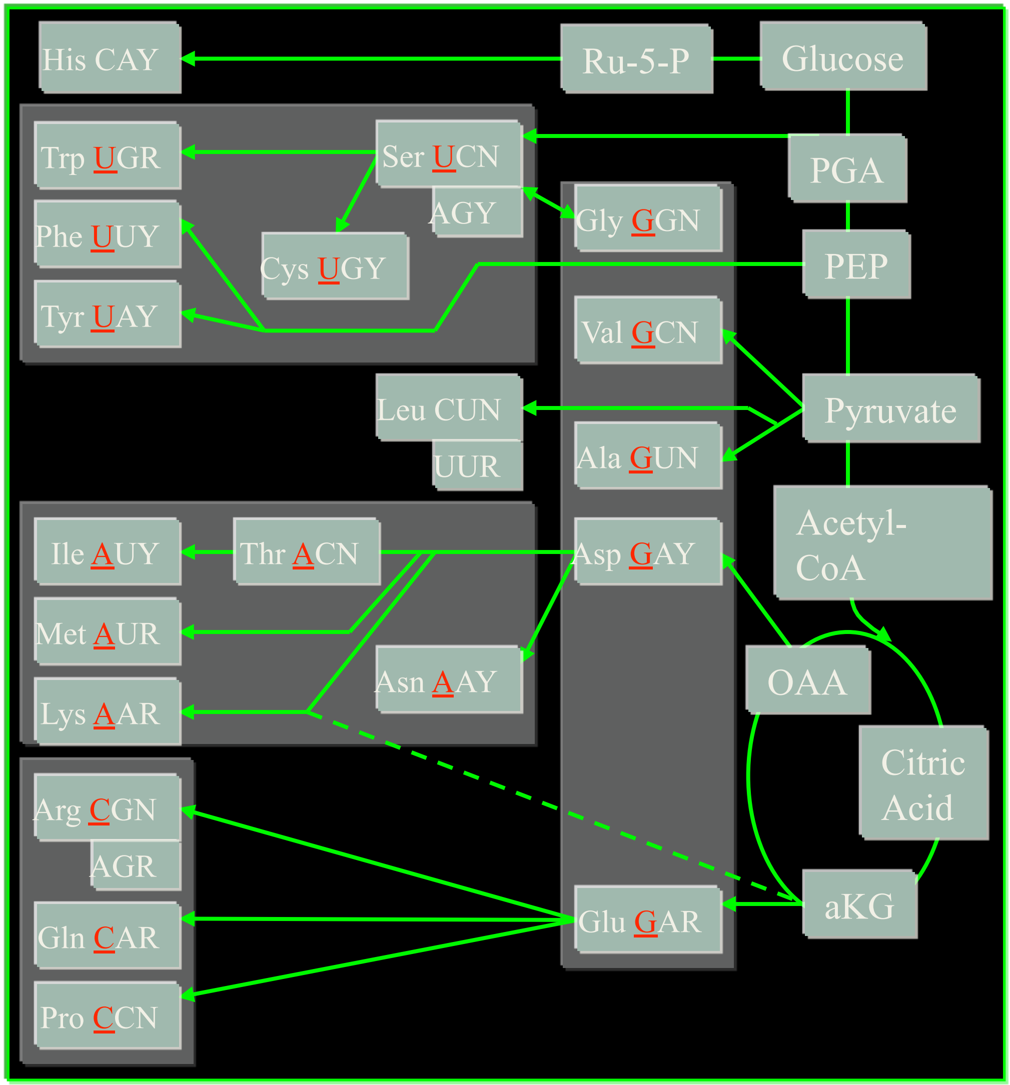
- One study has compiled more than 60 incorporation orders from various genetic code studies and derived several “consensus” orders *(Trivonov, 2000, 2004).*

- However they were derived, these various ideas largely reflect an underlying (intuitive) theme that small, chemically simple amino acids were encoded before big, complex amino acids

- According to Trifonov, the following one “has a merit of a best guess” based on computer analysis *(Table V in Trifonov, 2004)*:

Gly, Ala, Asp, Val, Pro, Ser, Glu, Thr, Leu, Arg, Asn, Ile, Gln, His, Lys, Cys, Phe, Tyr, Met, Trp

Simple, general patterns of biosynthetic relatedness



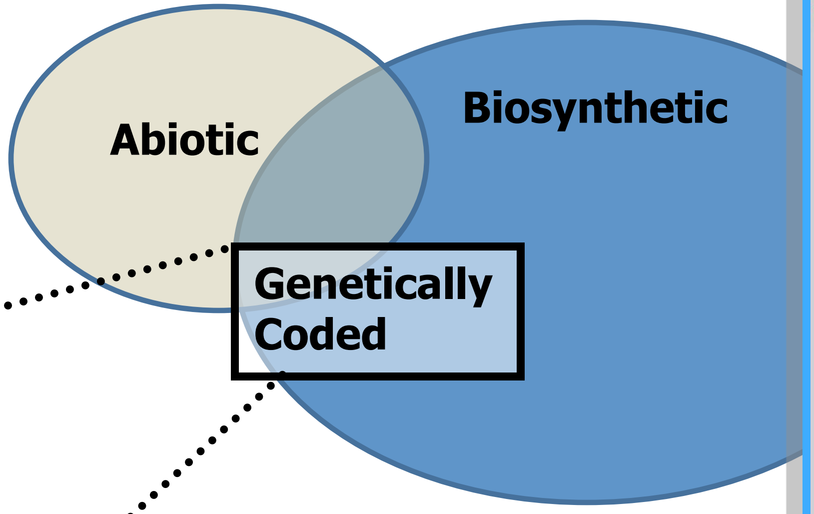
~ Amino acids from the same biosynthetic pathway tend to be assigned codons with the same first base.

~ The most likely prebiotic amino acids are all assigned to GNN codons

Taylor & Coates (1989) *BioSystems* 22, 177 - 187.

Miseta, A. (1989) *Physiol. Chem. Phys. Med. NMR* 21, 237 - 242.

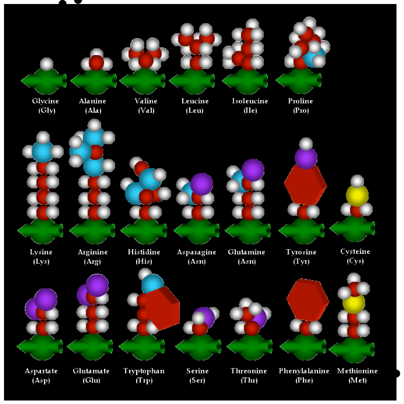
Genetically encoded, biosynthesized amino acids



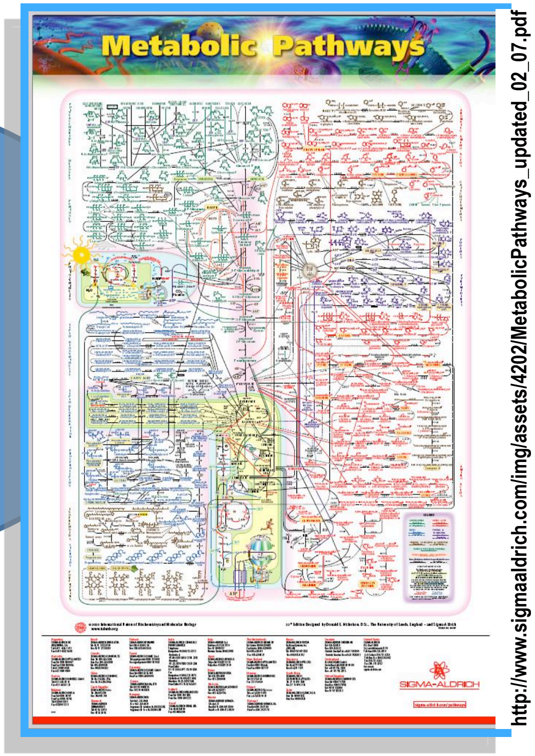
Abiotic

Biosynthetic

Genetically Coded



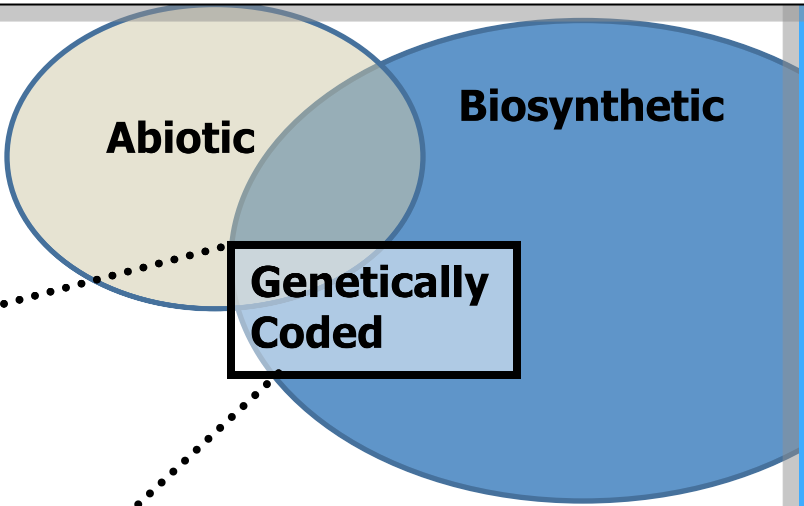
20 amino acids

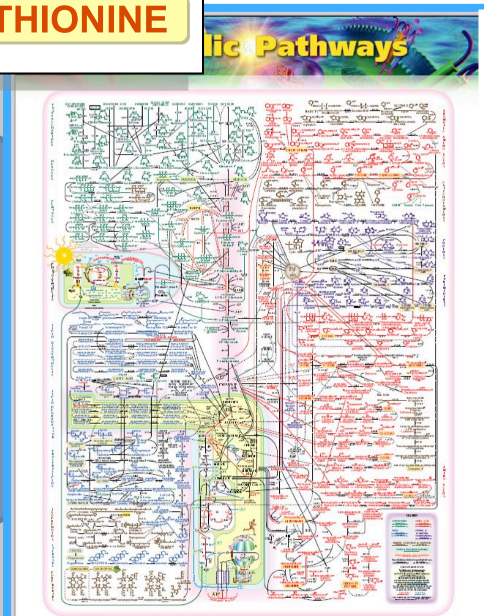


http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

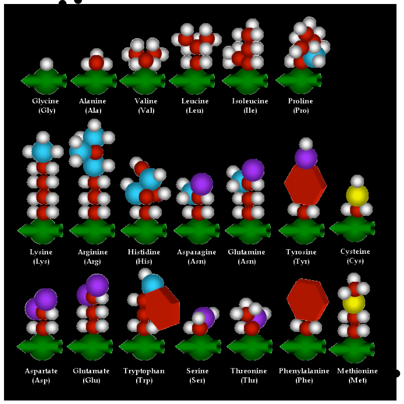
Genetically encoded, biosynthesized amino acids

Cystine → CYSTEINE → Cystathionine → Homocysteine → METHIONINE

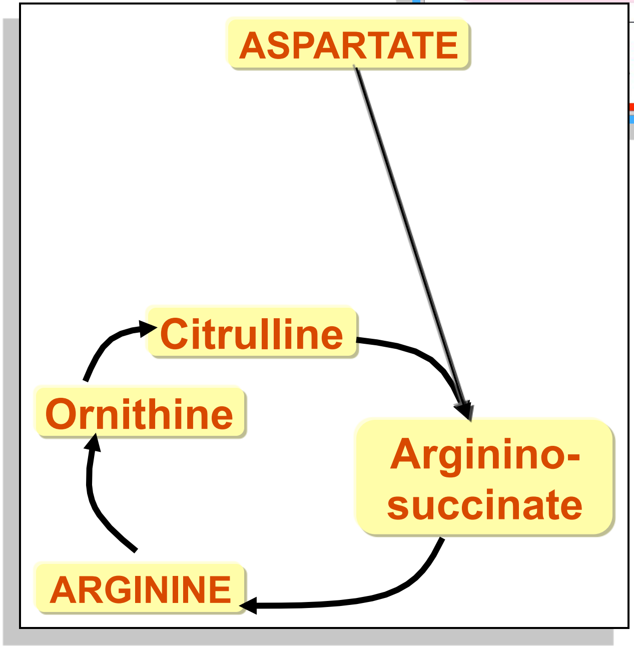




http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

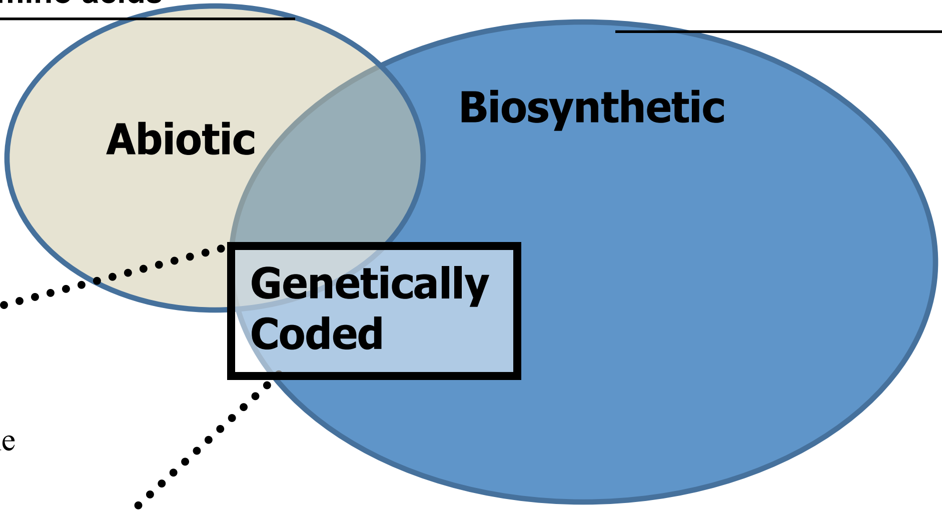


20 amino acids

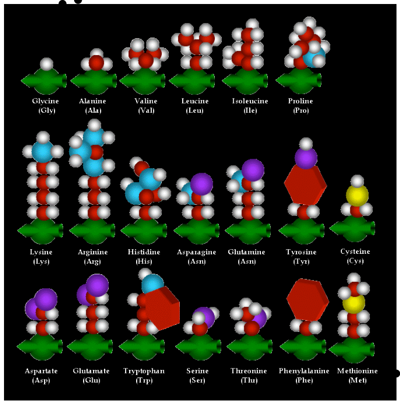


Genetically encoded, biosynthesized amino acids

12 biosynthetic standard amino acids
Arginine
Asparagine
Cysteine
Glutamine
Histidine
Lysine
Methionine
Phenylalanine
Serine
Threonine
Tryptophan
Tyrosine



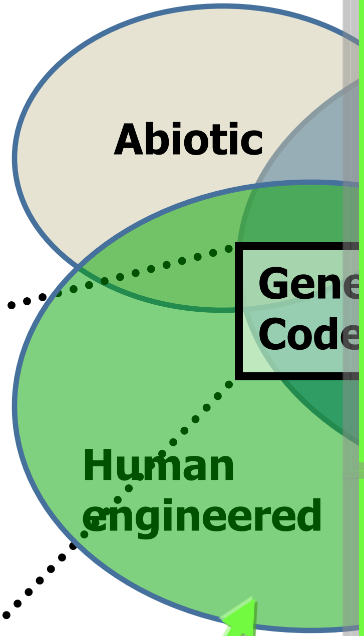
14 biosynthetic amino acid intermediates
Acetylserine
Anthranilate
Carnosine
Citrulline
Cystathione
Diaminopimelate
Formylkynurenine
Homocysteine
Homoserine
Kynurenine
2-amino-3-carboxy muconate semialdehyde
3-hydroxy-anthranilate
3-hydroxy-kynurenine
5-hydoxy-tryptophan

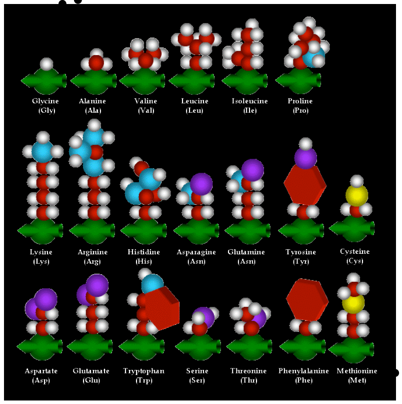


20 amino acids

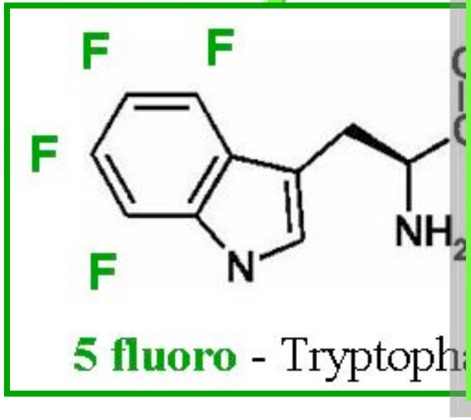
…are a small subset of what is available

A vast “chemical space” is possible

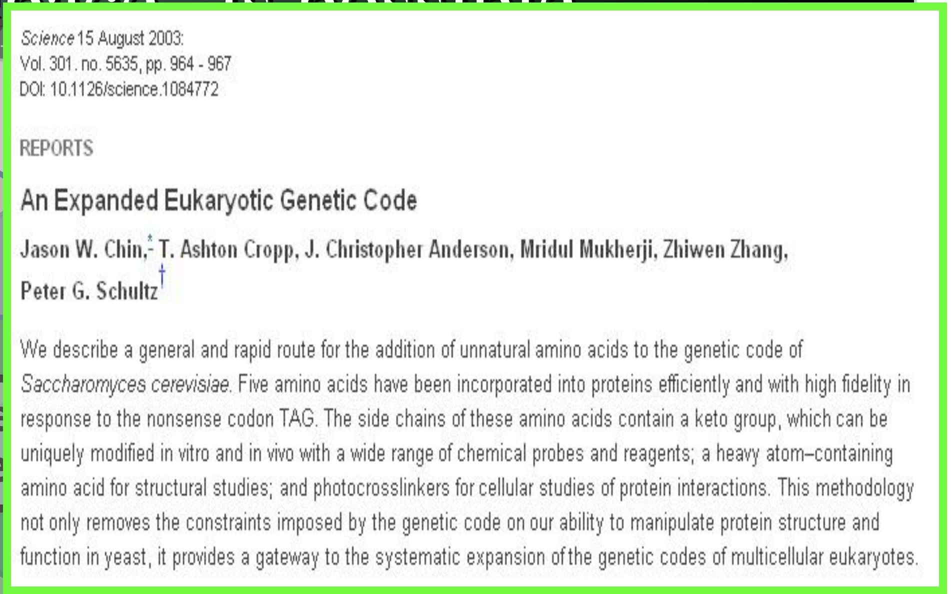




20 amino acids



5 fluoro - Tryptoph



Science 15 August 2003:
Vol. 301. no. 5635, pp. 964 - 967
DOI: 10.1126/science.1084772

REPORTS

An Expanded Eukaryotic Genetic Code

Jason W. Chin,* T. Ashton Cropp, J. Christopher Anderson, Mridul Mukherji, Zhiwen Zhang,
Peter G. Schultz†

We describe a general and rapid route for the addition of unnatural amino acids to the genetic code of *Saccharomyces cerevisiae*. Five amino acids have been incorporated into proteins efficiently and with high fidelity in response to the nonsense codon TAG. The side chains of these amino acids contain a keto group, which can be uniquely modified in vitro and in vivo with a wide range of chemical probes and reagents; a heavy atom–containing amino acid for structural studies; and photocrosslinkers for cellular studies of protein interactions. This methodology not only removes the constraints imposed by the genetic code on our ability to manipulate protein structure and function in yeast, it provides a gateway to the systematic expansion of the genetic codes of multicellular eukaryotes.

Expanding the Genetic Code

Summary (May 20, 2004): Life as we know it depends on the action of 20 protein building blocks, or amino acids. By adding a few new candidates, scientists hope to find out why certain organisms may be able to depart ever so slightly from the great Tree of Life.

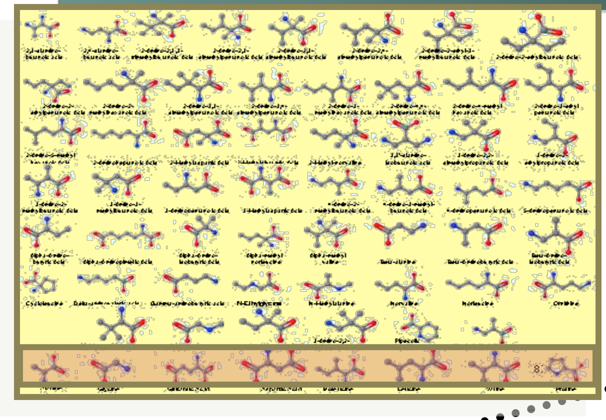
Expanding the Genetic Code

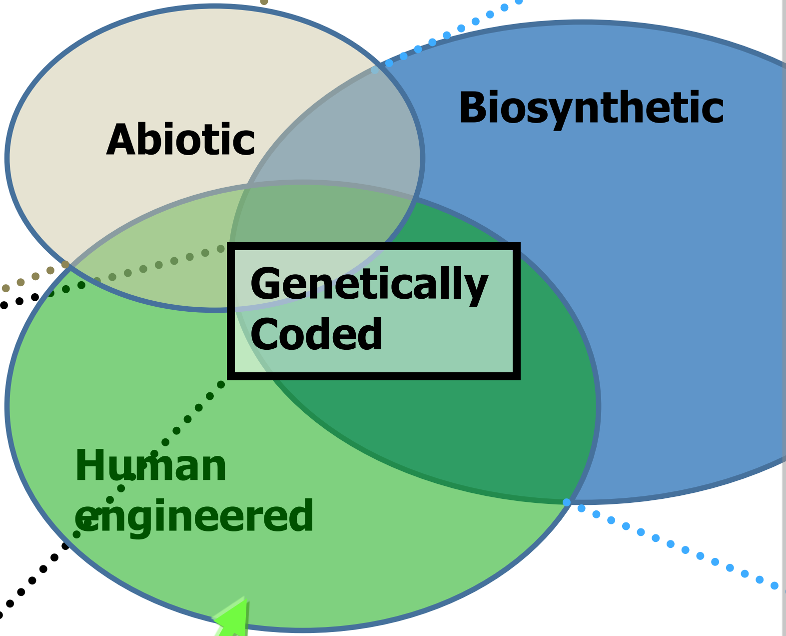
based on Scripps Research Institute report

A team of investigators at The Scripps Research Institute and its Skaggs Institute for Chemical Biology in La Jolla, California has modified a form of the bacterium Escherichia coli to use a 22-amino acid genetic code.

"We have demonstrated the simultaneous incorporation of two unnatural amino acids into the same polypeptide," says Professor Peter G. Schultz, Ph.D., who holds the Scripps Family Chair in Chemistry at Scripps Research. "Now that we know the genetic code is amenable to expansion to 22 amino acids, the next question is, how far can we take it?"

A vast “chemical space” is possible



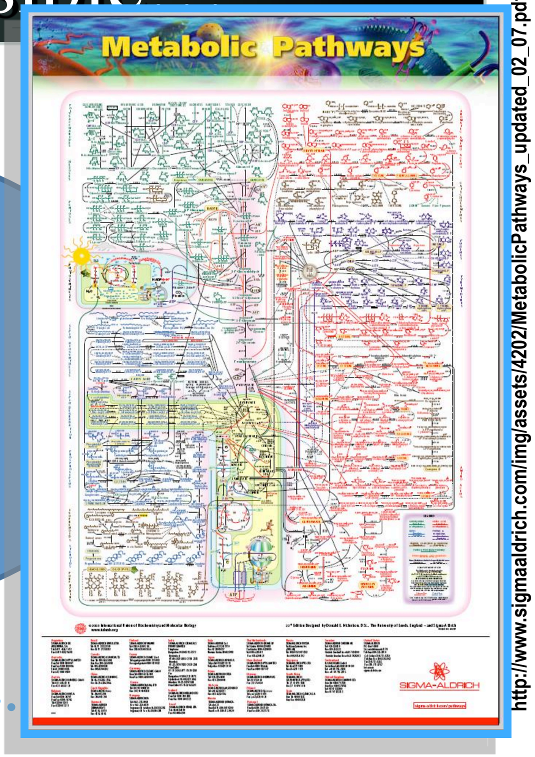


Abiotic

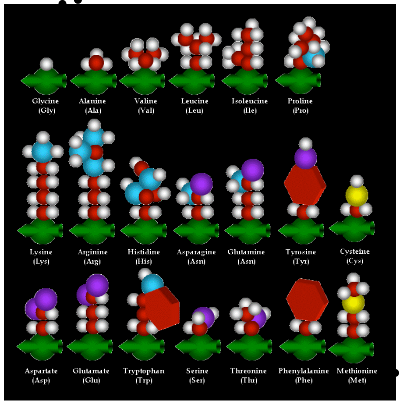
Biosynthetic

Genetically Coded

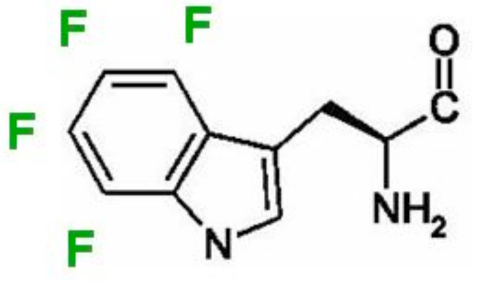
Human engineered



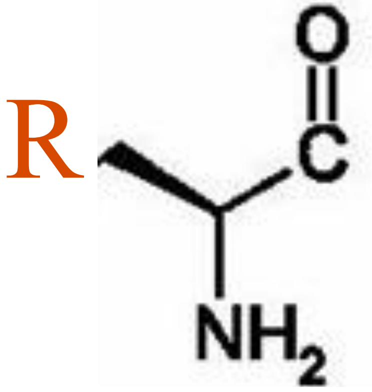
http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf



20 amino acids

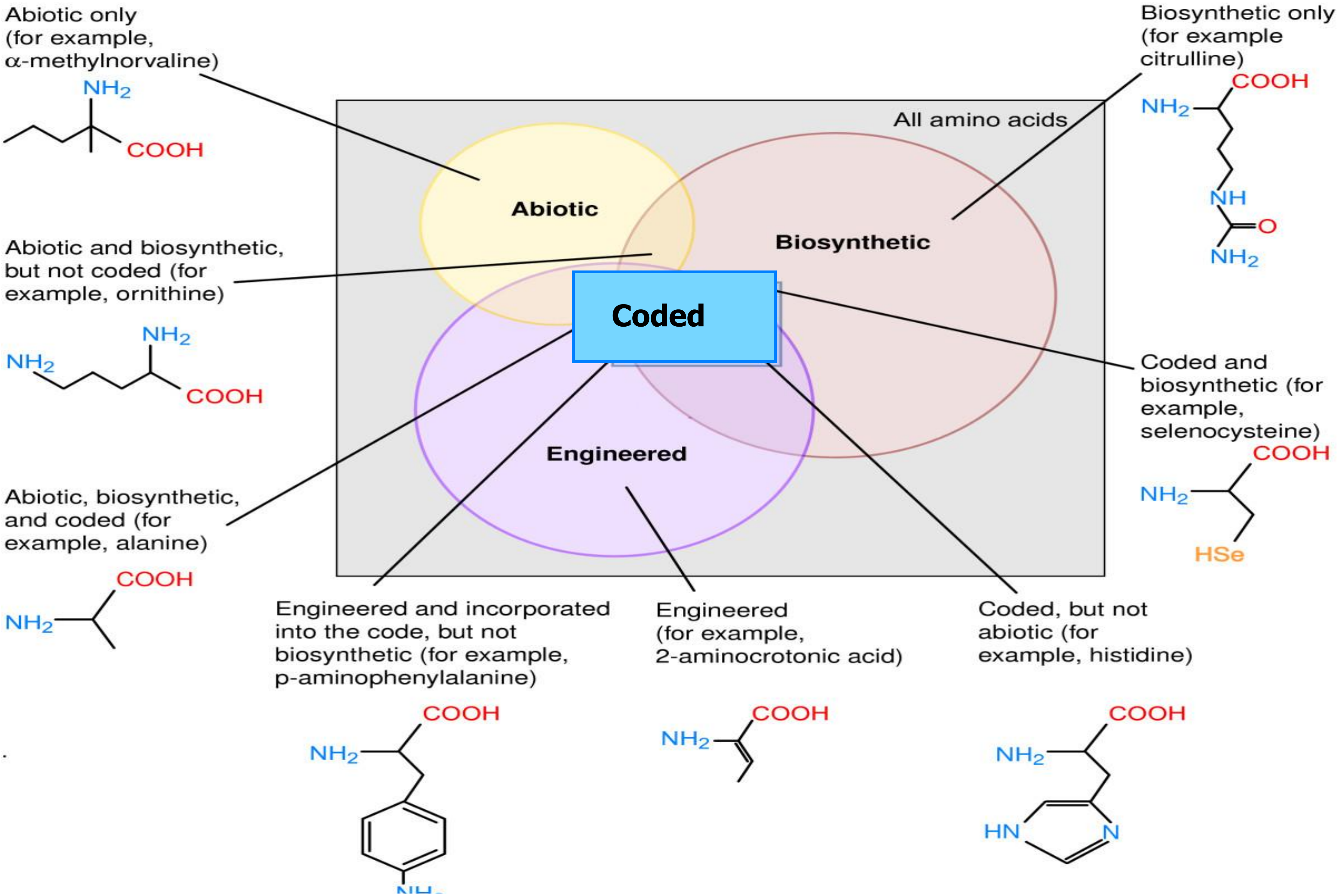


5 fluoro - Tryptophan



...and R = anything!

Genetically encoded amino acids are a small subset of chemical possibilities



Abiotic only (for example, α-methylnorvaline)

Biosynthetic only (for example citrulline)

Abiotic and biosynthetic, but not coded (for example, ornithine)

Coded and biosynthetic (for example, selenocysteine)

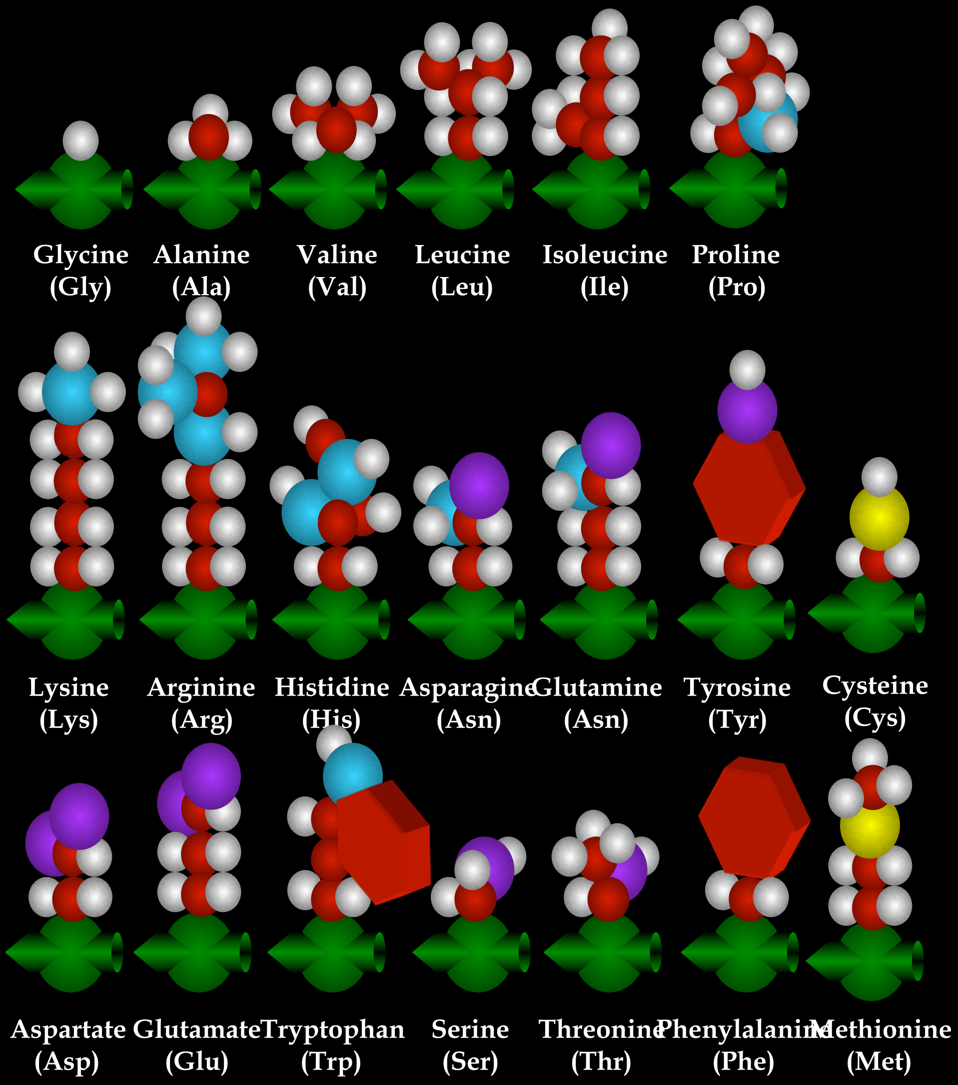
Abiotic, biosynthetic, and coded (for example, alanine)

Engineered and incorporated into the code, but not biosynthetic (for example, p-aminophenylalanine)

Engineered (for example, 2-aminocrotonic acid)

Coded, but not abiotic (for example, histidine)

"On the evolution of the standard amino-acid alphabet" Lu & Freeland (2006) **Genome Biol.** 7:102



So why did nature "choose" these 20 amino acids?

The scientific literature on the topic…

"Proteins provided a greater catalytic versatility than nucleic acids"

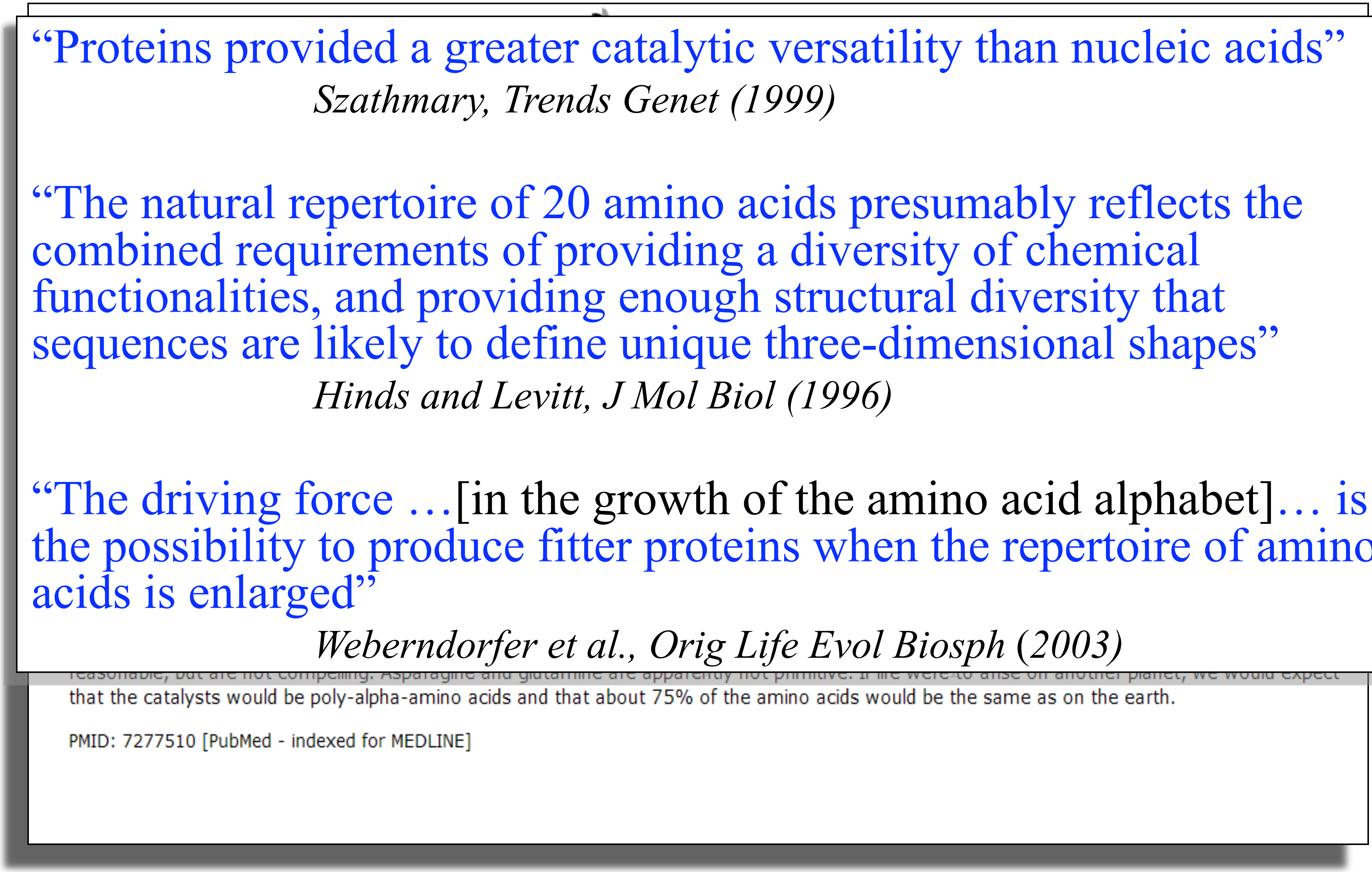
Szathmary, Trends Genet (1999)

"The natural repertoire of 20 amino acids presumably reflects the combined requirements of providing a diversity of chemical functionalities, and providing enough structural diversity that sequences are likely to define unique three-dimensional shapes"

Hinds and Levitt, J Mol Biol (1996)

"The driving force …[in the growth of the amino acid alphabet]… is the possibility to produce fitter proteins when the repertoire of amino acids is enlarged"

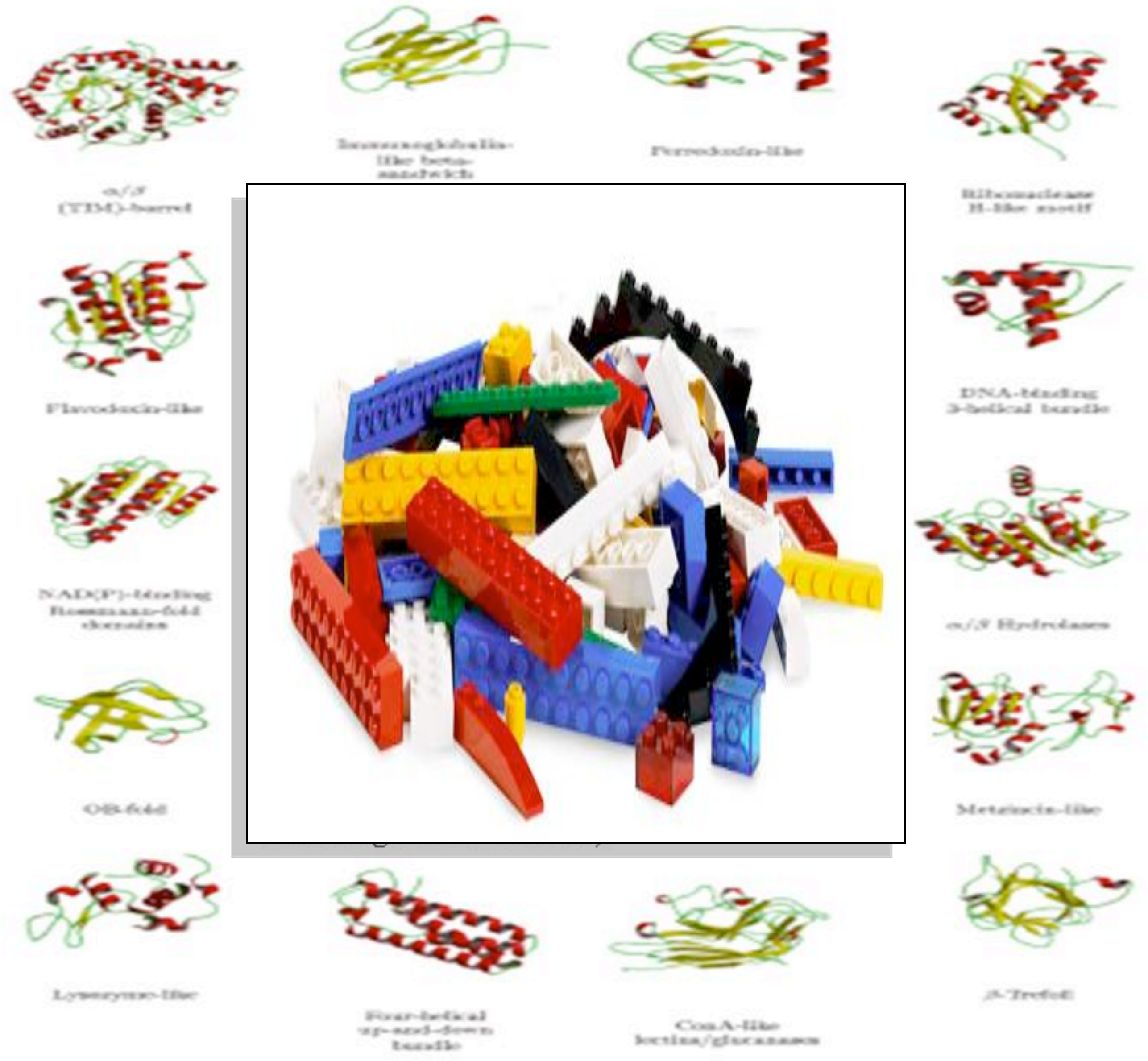
Weberndorfer et al., Orig Life Evol Biosph (2003)

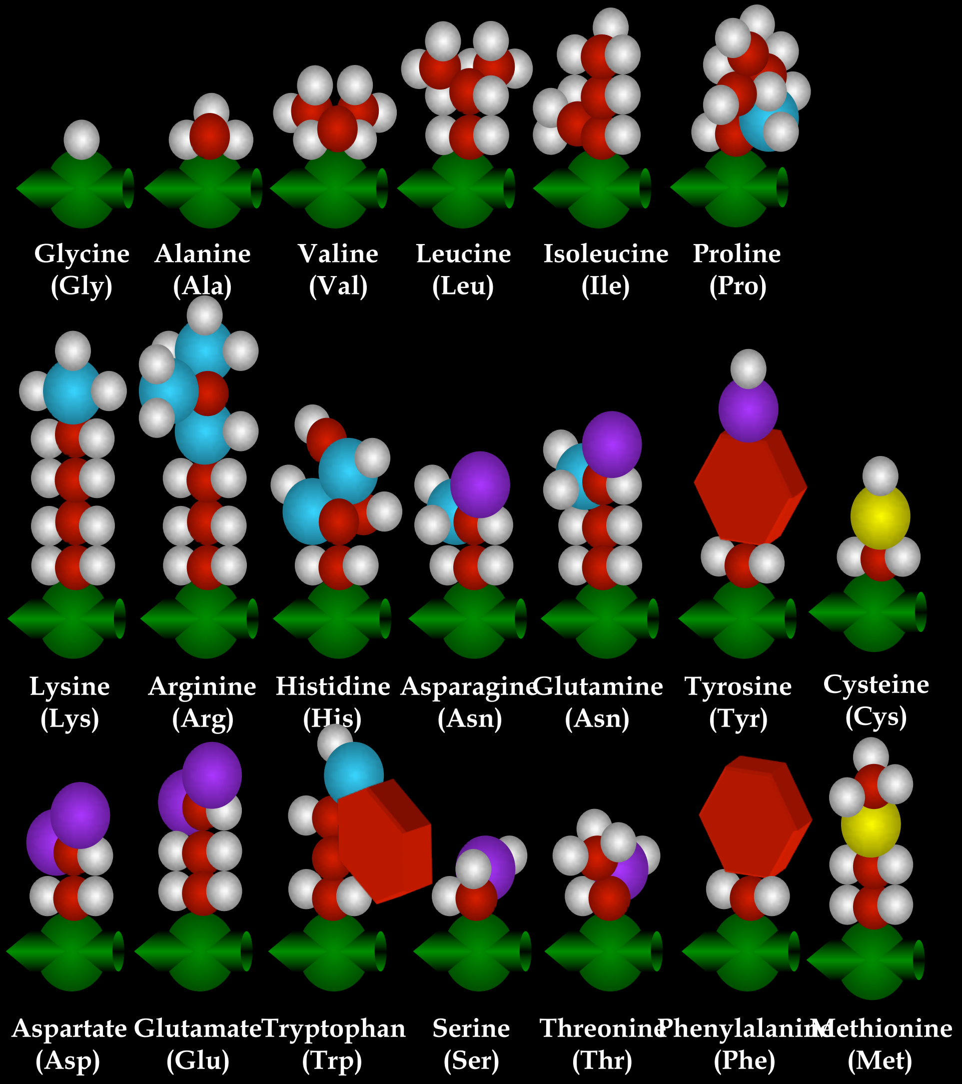


reasonable, but are not compelling. Asparagine and glutamine are apparently not primitive. If life were to arise on another planet, we would expect that the catalysts would be poly-alpha-amino acids and that about 75% of the amino acids would be the same as on the earth.

PMID: 7277510 [PubMed - indexed for MEDLINE]

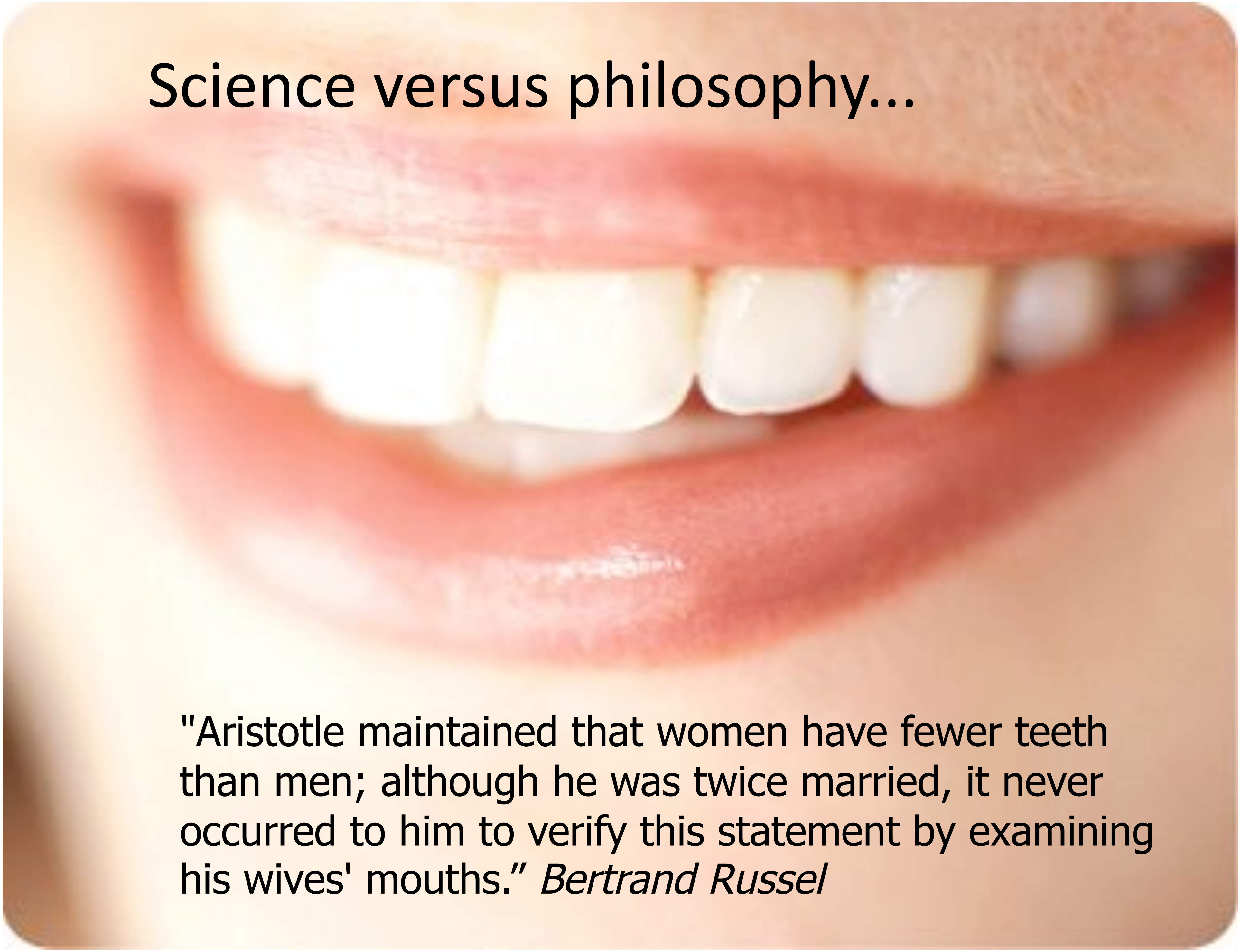
- **More building blocks = more shapes + functions**





The amino acid alphabet was “chosen” to allow the construction of varied proteins

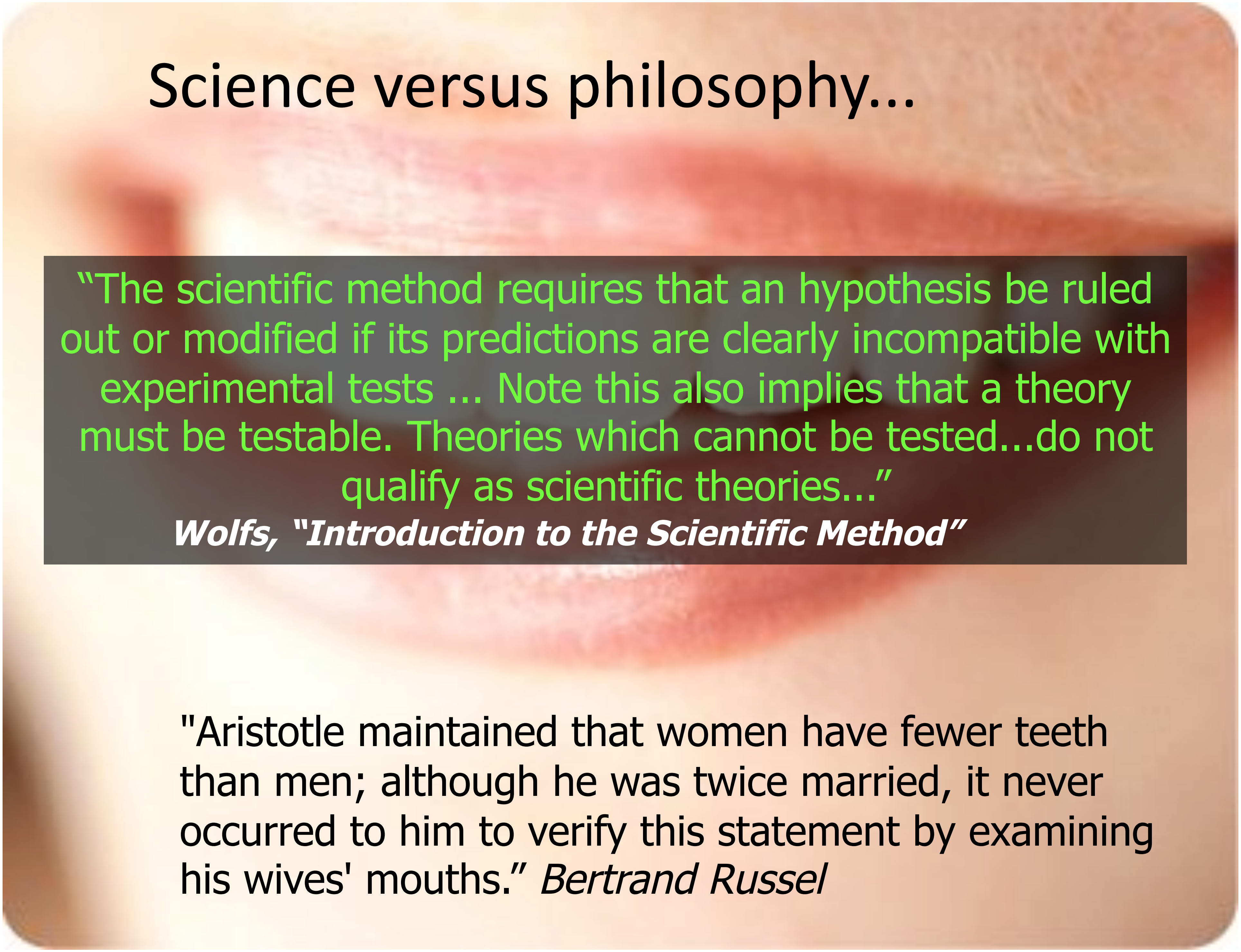
How can we test hypotheses here?



Science versus philosophy...

"Aristotle maintained that women have fewer teeth than men; although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths." *Bertrand Russel*

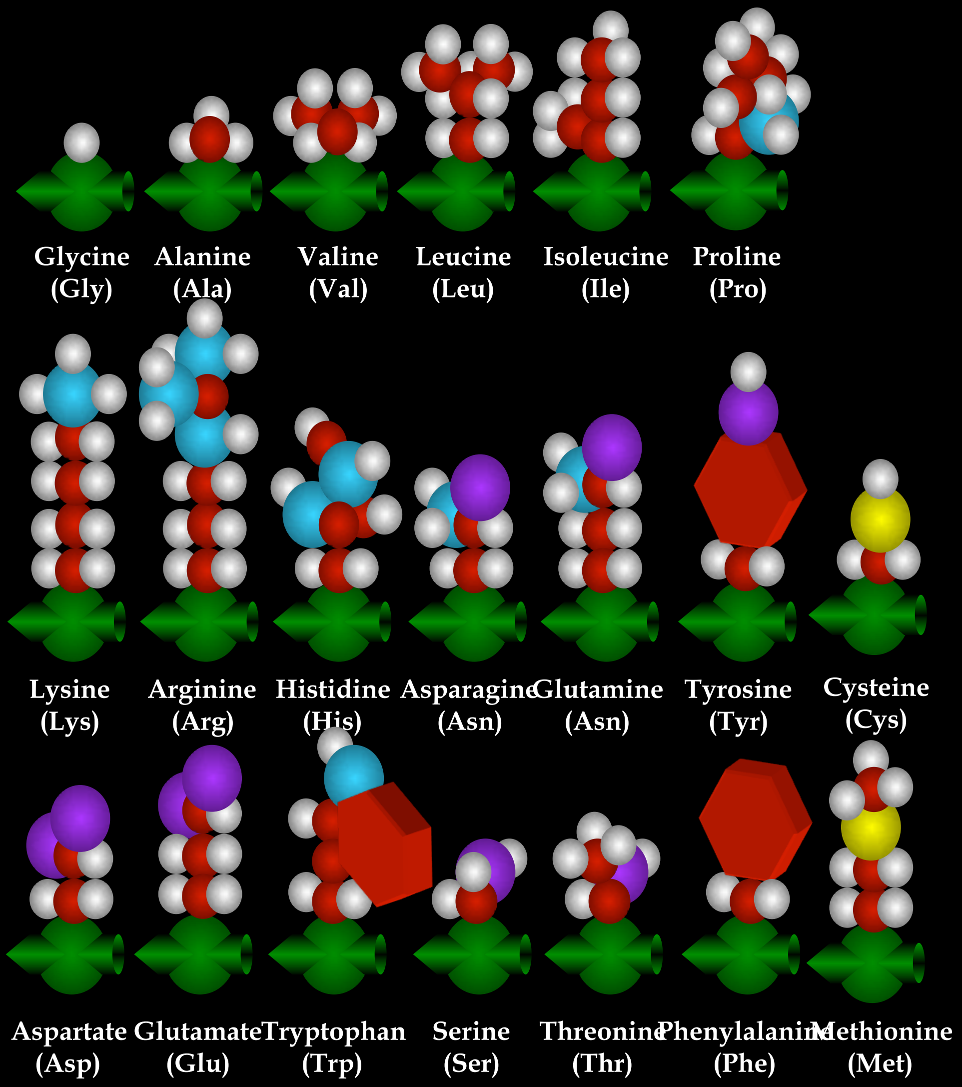
Science versus philosophy…



"The scientific method requires that an hypothesis be ruled out or modified if its predictions are clearly incompatible with experimental tests … Note this also implies that a theory must be testable. Theories which cannot be tested…do not qualify as scientific theories…"

Wolfs, "Introduction to the Scientific Method"

"Aristotle maintained that women have fewer teeth than men; although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths." *Bertrand Russel*



H_1: these 20 amino acids are more biochemically diverse than a random collection

H_0: no they aren’t

◆ the challenge:

1: what is biochemical diversity & how do you measure it?

2: what does a random collection of amino acids look like?

the challenge:

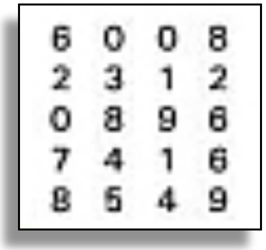
1: what is biochemical diversity & how do you measure it?

2: what does a random collection of amino acids look like?

- **the problem:** ***(why has no one asked even simple questions about "biochemical diversity" for the non-proteinaceous amino acids?)***

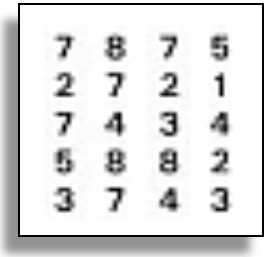
1: what is biochemical diversity & how do you measure it?

Estimating “functional diversity” of an alphabet



6	0	0	8
2	3	1	2
0	8	9	6
7	4	1	6
8	5	4	9

measure the corresponding values for your collection of objects



7	8	7	5
2	7	2	1
7	4	3	4
5	8	8	2
3	7	4	3

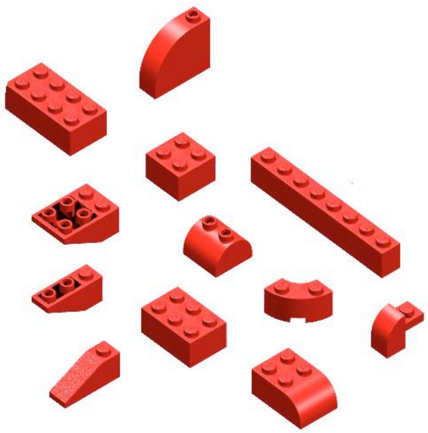
σ^2

Variance$(x_1, x_2 \ldots x_n)$

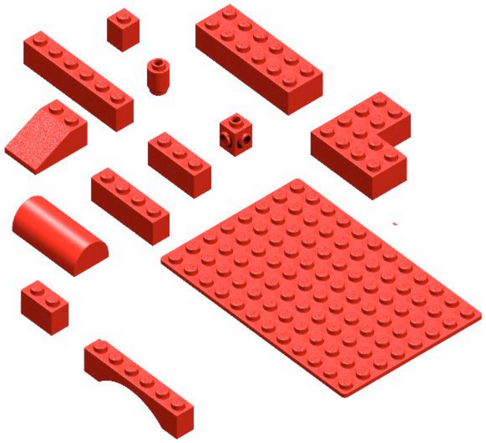
=

squared deviation around the mean

σ^2



Choose a relevant, quantifiable property



From chemical properties to *molecular descriptors*

Each of these is a different
“index”: something that has
be
pr

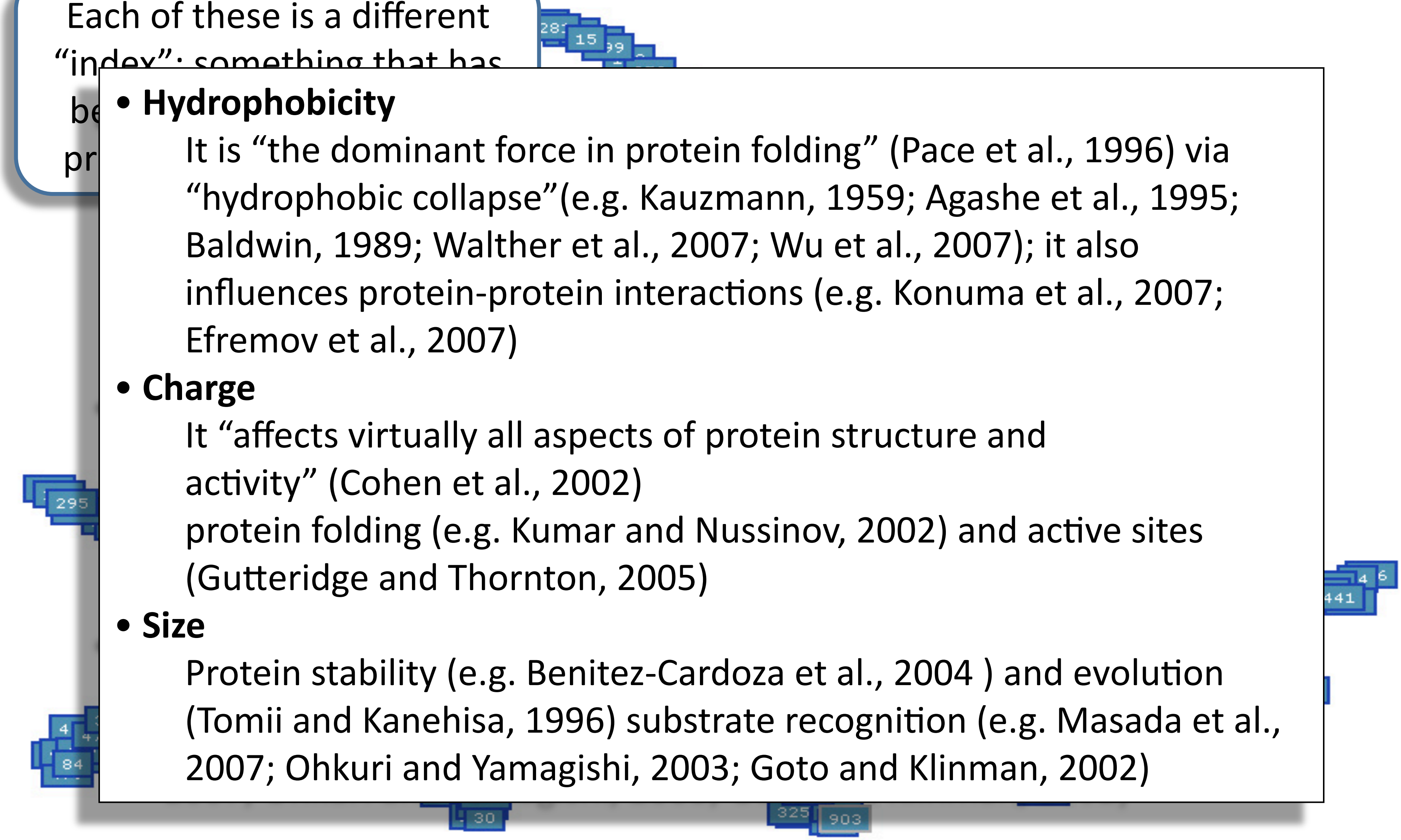
- **Hydrophobicity**

  It is “the dominant force in protein folding” (Pace et al., 1996) via “hydrophobic collapse”(e.g. Kauzmann, 1959; Agashe et al., 1995; Baldwin, 1989; Walther et al., 2007; Wu et al., 2007); it also influences protein-protein interactions (e.g. Konuma et al., 2007; Efremov et al., 2007)
- **Charge**

  It “affects virtually all aspects of protein structure and activity” (Cohen et al., 2002)

  protein folding (e.g. Kumar and Nussinov, 2002) and active sites (Gutteridge and Thornton, 2005)
- **Size**

  Protein stability (e.g. Benitez-Cardoza et al., 2004) and evolution (Tomii and Kanehisa, 1996) substrate recognition (e.g. Masada et al., 2007; Ohkuri and Yamagishi, 2003; Goto and Klinman, 2002)



Minimum Spanning Tree of 496 Amino Acid Indices

(The AAIndex database)

(Nature, 2004)

insight introduction

Chem

Christopher M. D

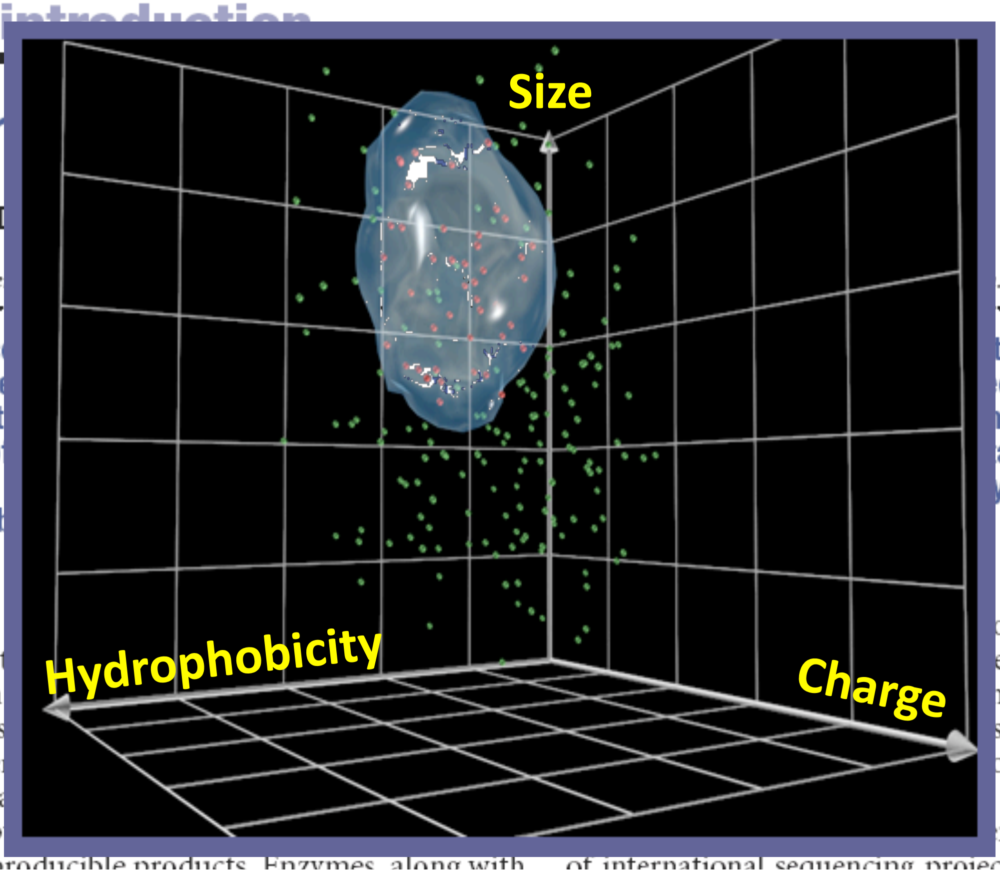
Department of Che *.ac.uk)*

Chemical spac
biological syst
Nevertheless, t
development o
understanding
knowledge of b

those present in
en explored.
nd have led to the
ated by a deeper
y, will advance our

L iving
years t
in an
almos
Unde
the chemical rea
occur at percep
specific and reproducible products. Enzymes, along with

olecules, and that biolog-
erse despite the relatively
n to have been exploited
stems (see Box 2 for a
of chemistry might have
nt triumphs of a variety
of international sequencing projects, the genomes of the



Molecules can be represented by points in a “multi-dimensional molecular descriptor space”

(Molecular descriptors are the specific measures of properties)

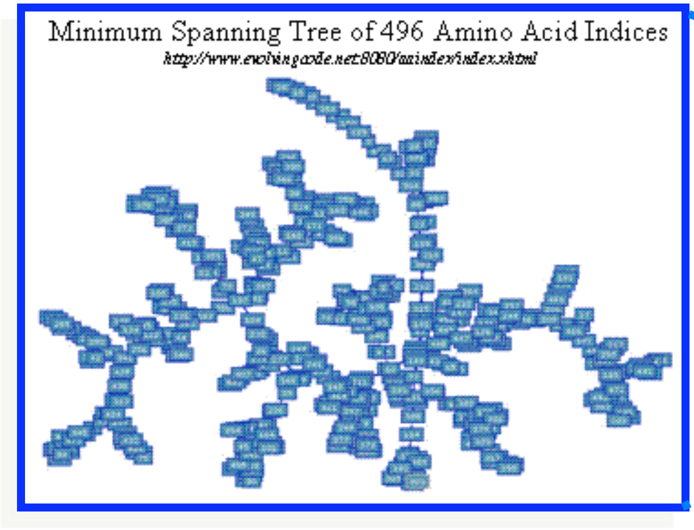
possible number of small carbon-based compounds with molecular masses in the same range as those of living systems (that is, less than about 500 daltons). Some estimates of this number are in excess of 10^{60} (ref. 2). The simplest living organisms can function with just a few hundred different types of such molecule, and fewer than 100 account for nearly the entire molecular pool[3,4]. Moreover, it seems that the total number of different small molecules within our own bodies

single molecule of each of these polypeptides were to be produced, their combined mass would vastly exceed that of the known universe. Natural proteins are therefore also a very select group of molecules.

The characteristics of this select group of natural proteins are linked to those of the small molecules that are used in living systems, and to those of the relatively small number of synthetic small molecules that we have developed into

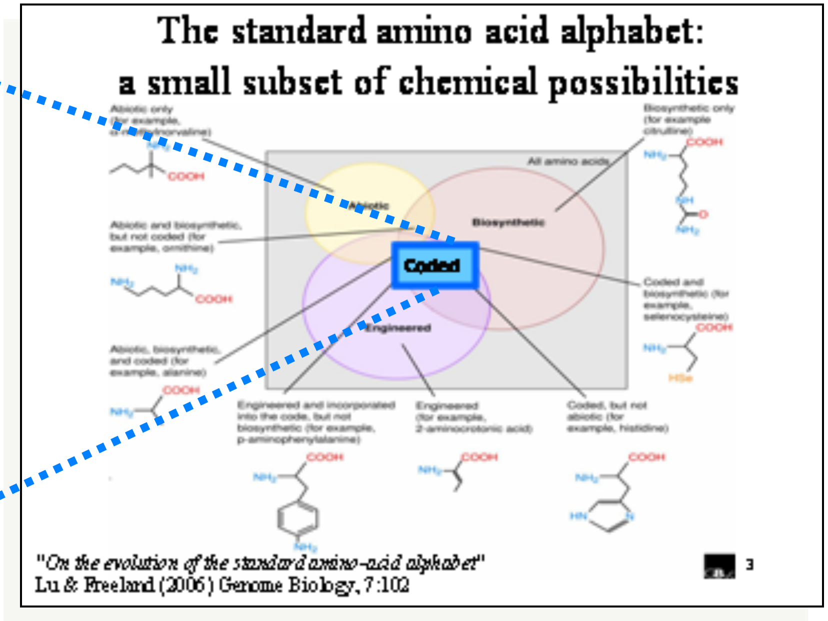
the problem: ***(why has no one asked even simple questions of this kind?)***

Most molecular descriptors are only available for the 20 standard amino acids…



Minimum Spanning Tree of 496 Amino Acid Indices

http://www.evolvingcode.net:8080/aaindex/index.xhtml



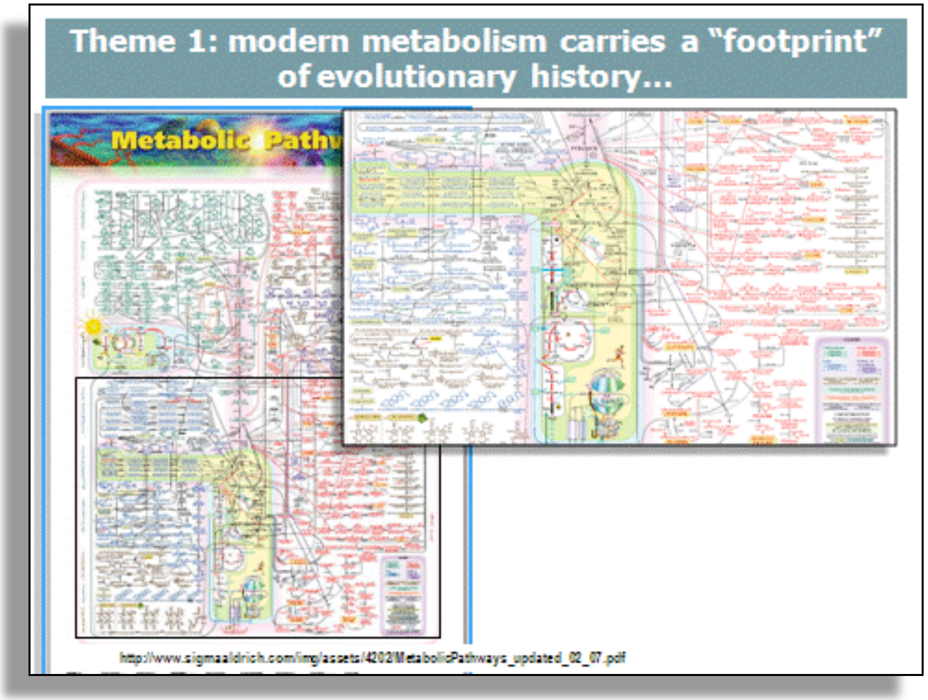
The standard amino acid alphabet:

a small subset of chemical possibilities

"On the evolution of the standard amino-acid alphabet"
Lu & Freeland (2006) Genome Biology, 7:102

How can we measure size, charge and hydrophobicity for all the other plausible choices of amino acid?

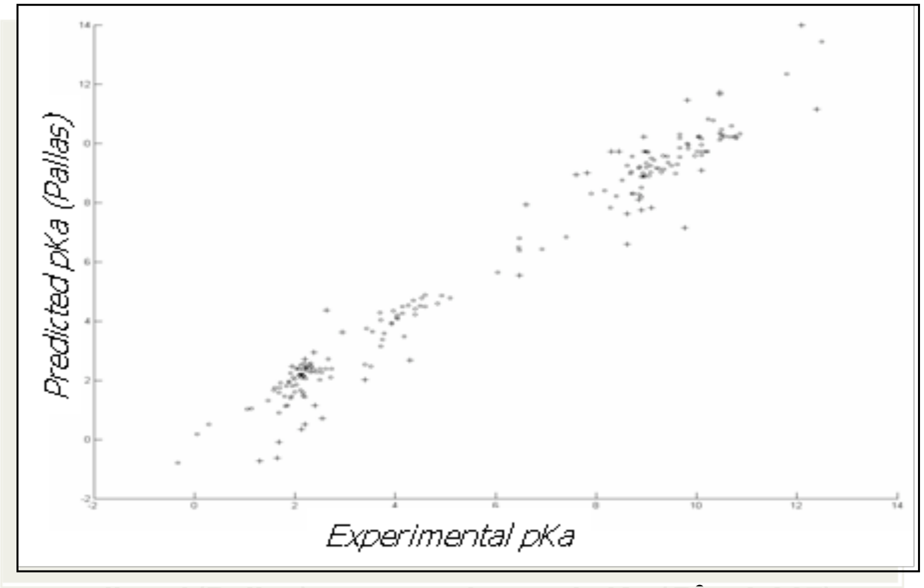
Theme II: Computers can quickly and cheaply "sketch" possibilities beyond life-as-we-know-it



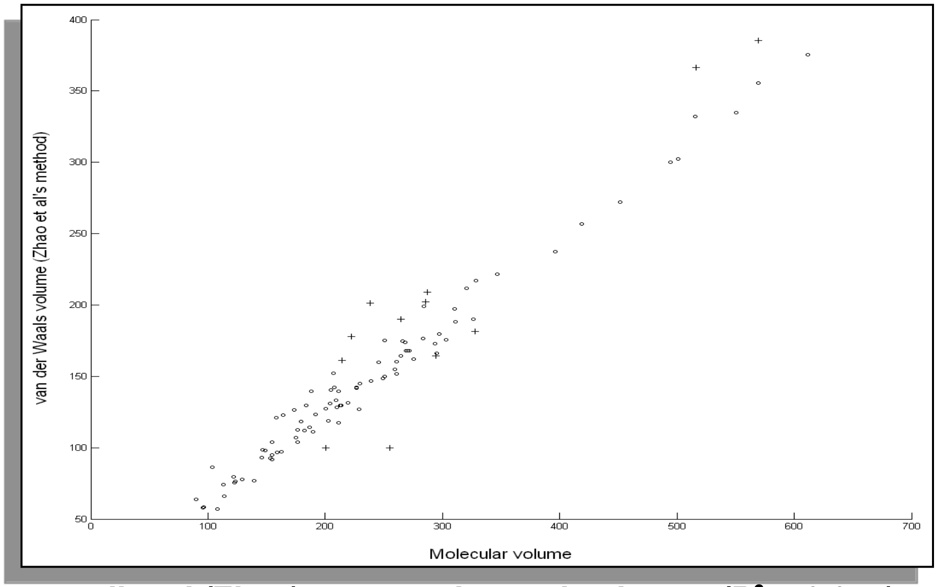
Theme 1: modern metabolism carries a "footprint" of evolutionary history…

http://www.sigmaaldrich.com/img/assets/4202/MetabolicPathways_updated_02_07.pdf

computers can sketch amino acid chemical space for us

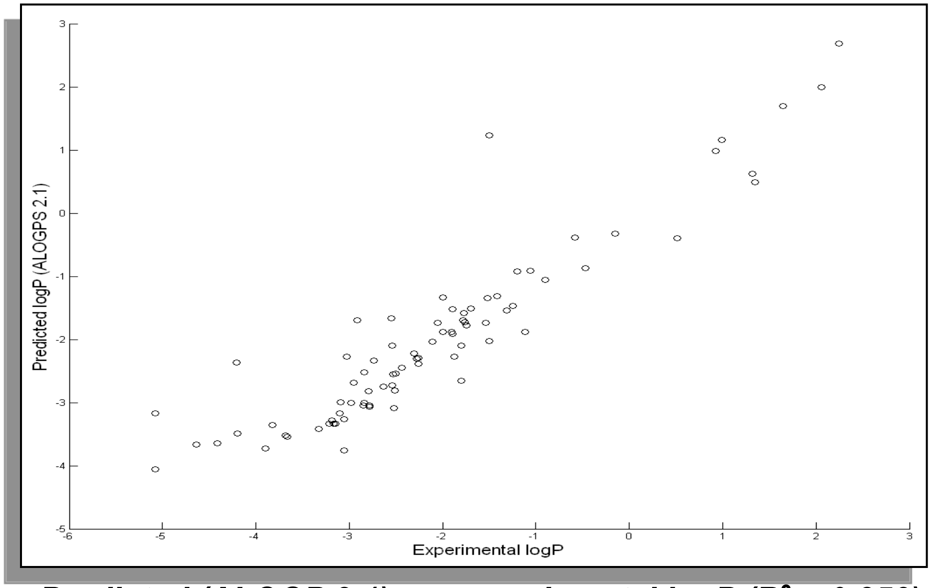


predicted **(Pallas) vs.** ***experimental*** **pKa (*R^2* = 0.967)**



predicted (Zhao) vs. experimental volumes (R^2 = 0.955)

"*Testing the Potential for Computational Chemistry to Quantify Physiochemical Properties of the Non-Proteinaceous Amino Acids*" ~ (2006), Lu Y and Freeland SJ., Astrobiology, 6(4): 606-624



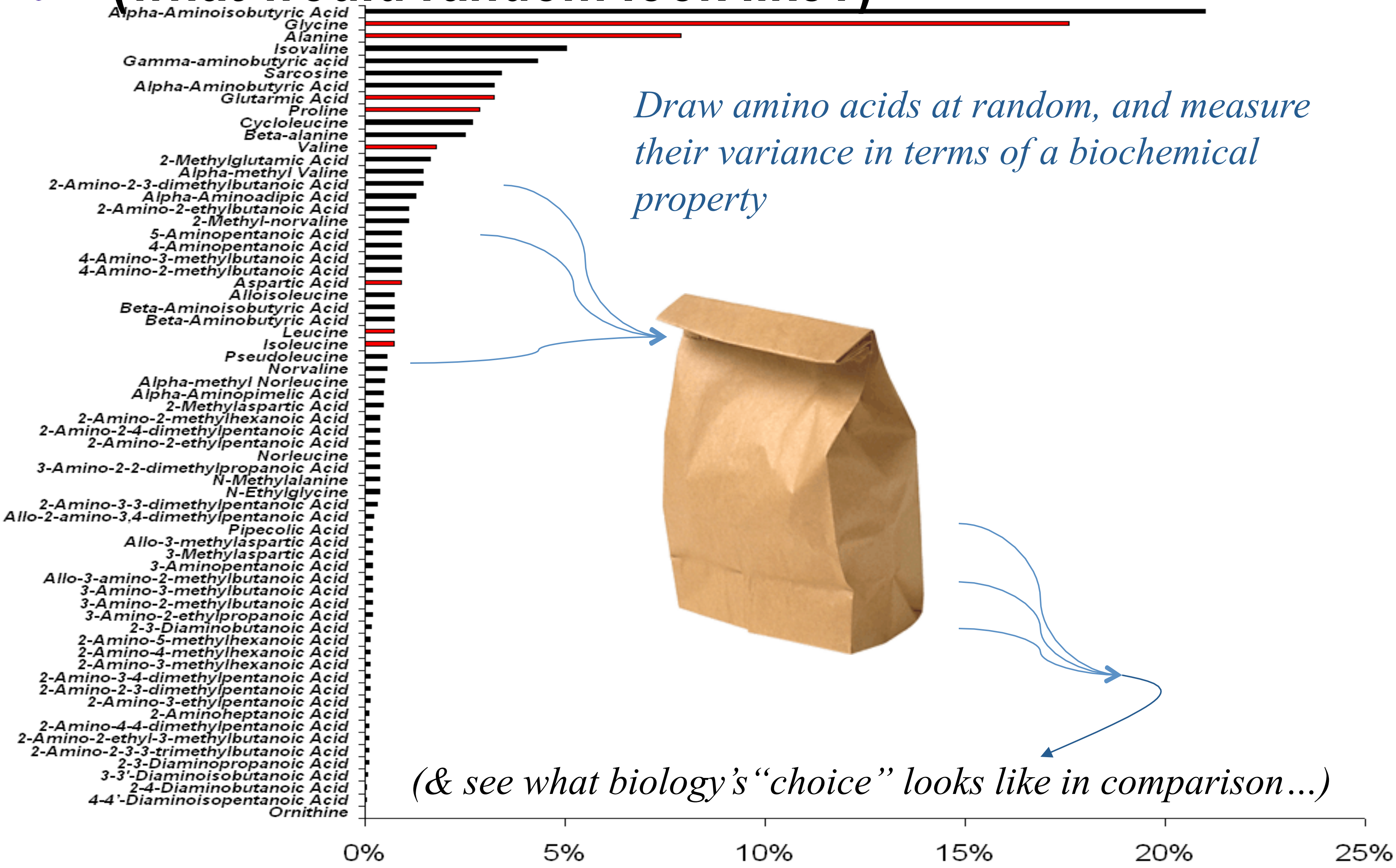
Predicted (ALOGP 2.1) vs. experimental logP (R^2 = 0.853)

◆ the challenge:

1: what is biochemical diversity & how do you measure it?

2: what does a random collection of amino acids look like?

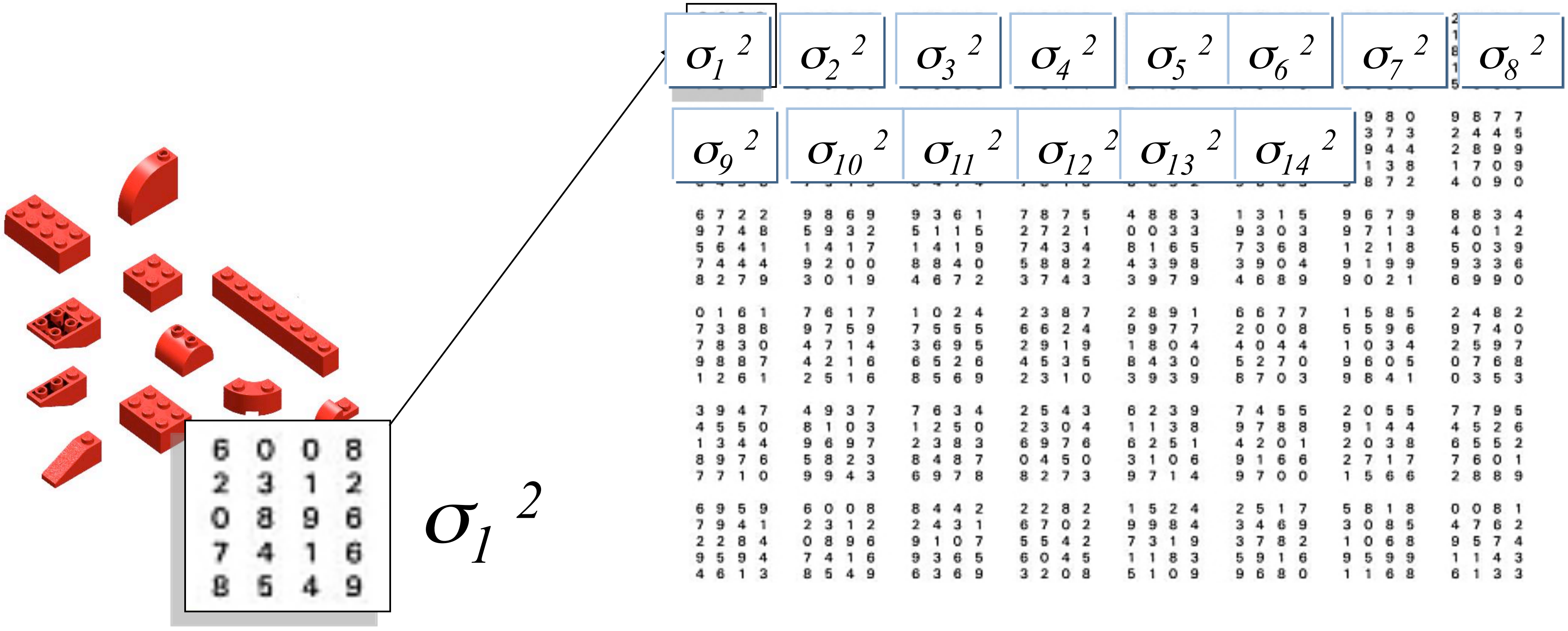
(what would random look like?)

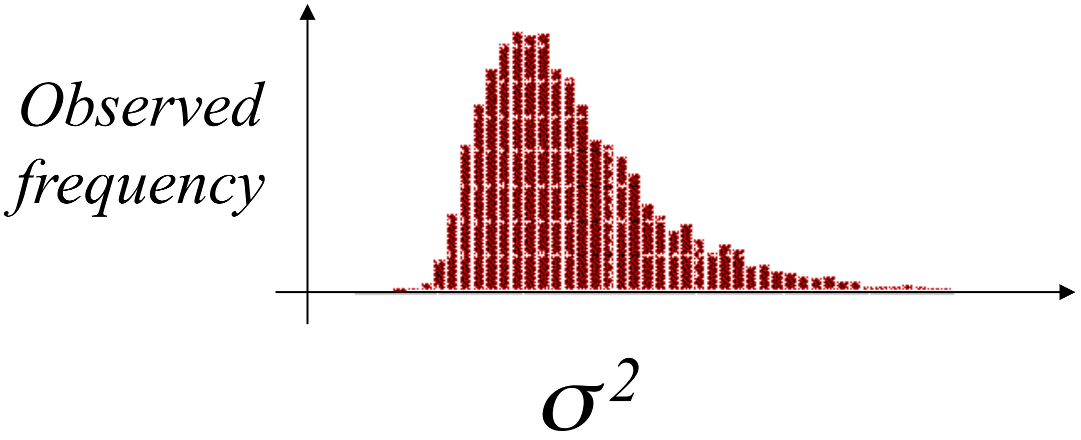


Draw amino acids at random, and measure their variance in terms of a biochemical property

(& see what biology's "choice" looks like in comparison…)

Comparing to expectation of “randomness”





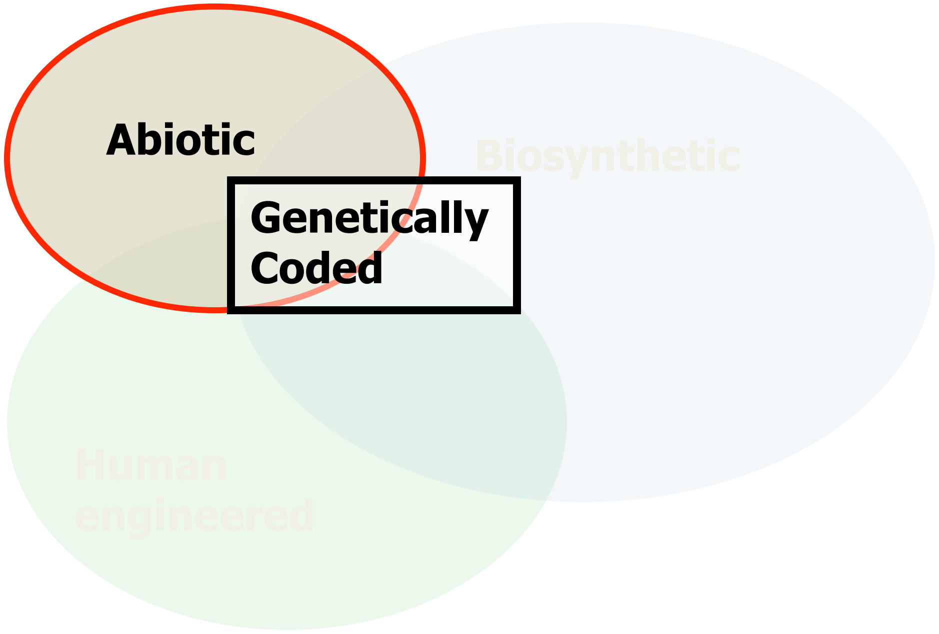
◆ 3 variants of our hypothesis:

A) Does the diversity of the standard amino acid alphabet exceed the diversity of an alphabet, size 20, drawn from pool of pre-biotic candidates?

B) Does the diversity of the 8 prebiotically plausible standard amino acids exceed the diversity of an alphabet, size 8, drawn from pool of pre-biotic candidates?

C) Does the diversity of the standard amino acid alphabet exceed the diversity of an alphabet, size 20, drawn from pool of pre-biotic + biosynthetic candidates?

A) Are the 20 standard amino acids more diverse than a set of Random 20 Drawn from abiotic amino acid Chemical Space?



Pool of available amino acids = 66 amino acids from the Murchison meteorite

Draw samples of size 20

Measure diversity and plot distribution of “random”

Compare diversity of 20 standard amino acids

A) Is *diversity (std amino acid alphabet) > diversity(random sample of 20 drawn from pre-biotic candidates)*?

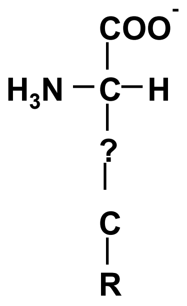
Percentage of Random Sets of 20 PREBIOTIC Amino Acids that show Lower σ^2 Values than those of the 20 STANDARD Amino Acids

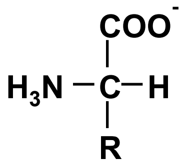
	logP (%)	**van der Waals (%)**	**pI (%)**
	38.2 ±0.03	100.0 ±0.001	97.9 ±0.01

A) Is *diversity (std amino acid alphabet) >* *diversity(random sample of 20 drawn from pre-biotic candidates)*?

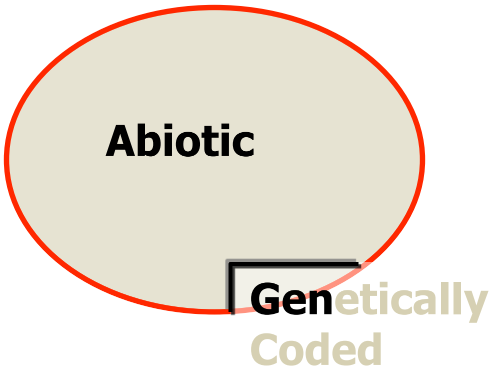
Percentage of Random Sets of 20 PREBIOTIC Amino Acids that show Lower σ^2 Values than those of the 20 STANDARD Amino Acids

	logP (%)	van der Waals (%)	pI (%)
A) 66 abiotic amino acids			
	38.2 ±0.03	100.0 ±0.001	97.9 ±0.01
B) 50 abiotic alpha- amino acids			
	59.9 ±0.03	100.0 ±0.001	99.7 ±0.003





2. Did the amino acid alphabet *start out* unusually diverse?



Pool of available amino acids = 66 from the Murchison meteorite (inc. 8 std alphabet)

Draw samples of size 8

Measure diversity and plot distribution of “random”

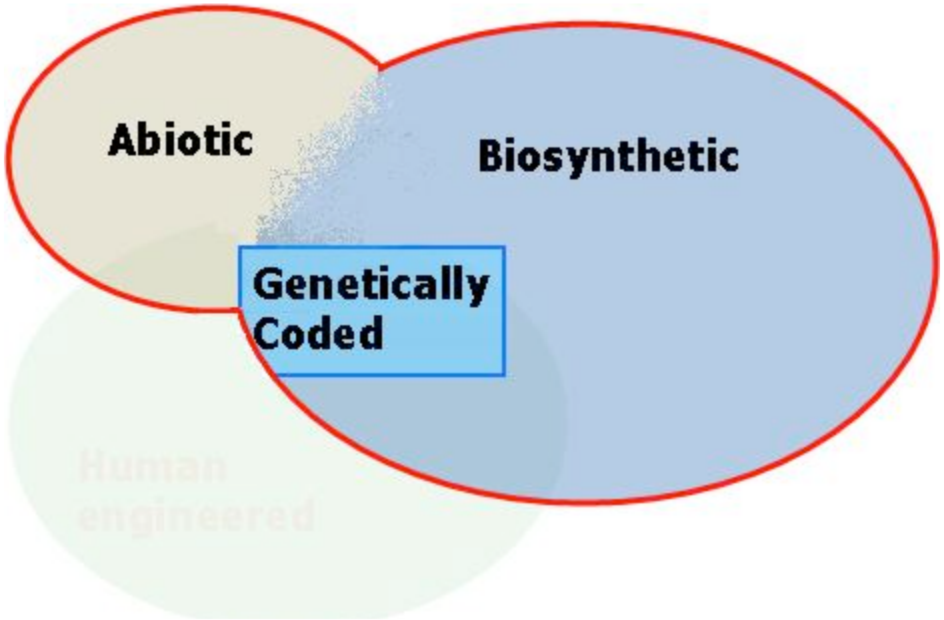
Compare diversity of 8 standard amino acids found in Murchison Meteorite

(i) Draw at random from the contents of the Murchison meteorite 10,000 sets of 8 prebiotically plausible amino acids

(ii) Find the percentage that show lower variance (less diversity) than the set of 8 genetically encoded amino acids found there…

	LogP	Van der Waals	pI
Un-weighted	33.7%	81.1%	57.3%
[illegible]	[illegible]	[illegible]	[illegible]

A) Are the 20 Standard Amino Acids more Diverse than a set of Random 20 Drawn from an Abiotic & Biosynthetic Amino Acid Chemical Space?



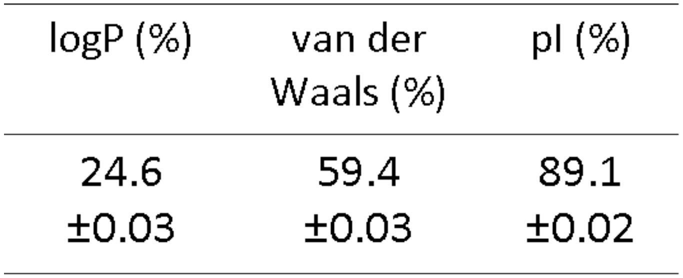
Pool of available amino acids = 66 from the Murchison meteorite (inc. 8 std alphabet)
\+ 12 additional standard amino acids
\+ 14 additional biosynthetic amino acids

Draw samples of size 20

Measure diversity and plot distribution of “random”

Compare diversity of 20 standard amino acids

The percentage of 10,000 sets of 20 biosynthetically plausible amino acids, drawn at random from the appropriate chemical space*, that show lower variance than the set of 20 genetically encoded amino acids found there…



logP (%)	van der Waals (%)	pI (%)
24.6 ±0.03	59.4 ±0.03	89.1 ±0.02

** combined chemical space: 66 abiotic amino acids (including 8 standard amino acids),*
12 standard amino acids and 14 biosynthetic intermediates.

◆ Conclusions…

- *The 20 amino acids of the standard alphabet are more diverse than a random sample of 20 taken from abiotic chemical space*
  (though for hydrophobicity, this is only true if we draw random samples weighted by the abundances observed in the Murchison meteorite)
  Life appears to be using more diverse building blocks than "whatever was lying around"
- *The 20 amino acids of the standard alphabet are not more diverse than a random sample taken from abiotic + biosynthetic chemical space*

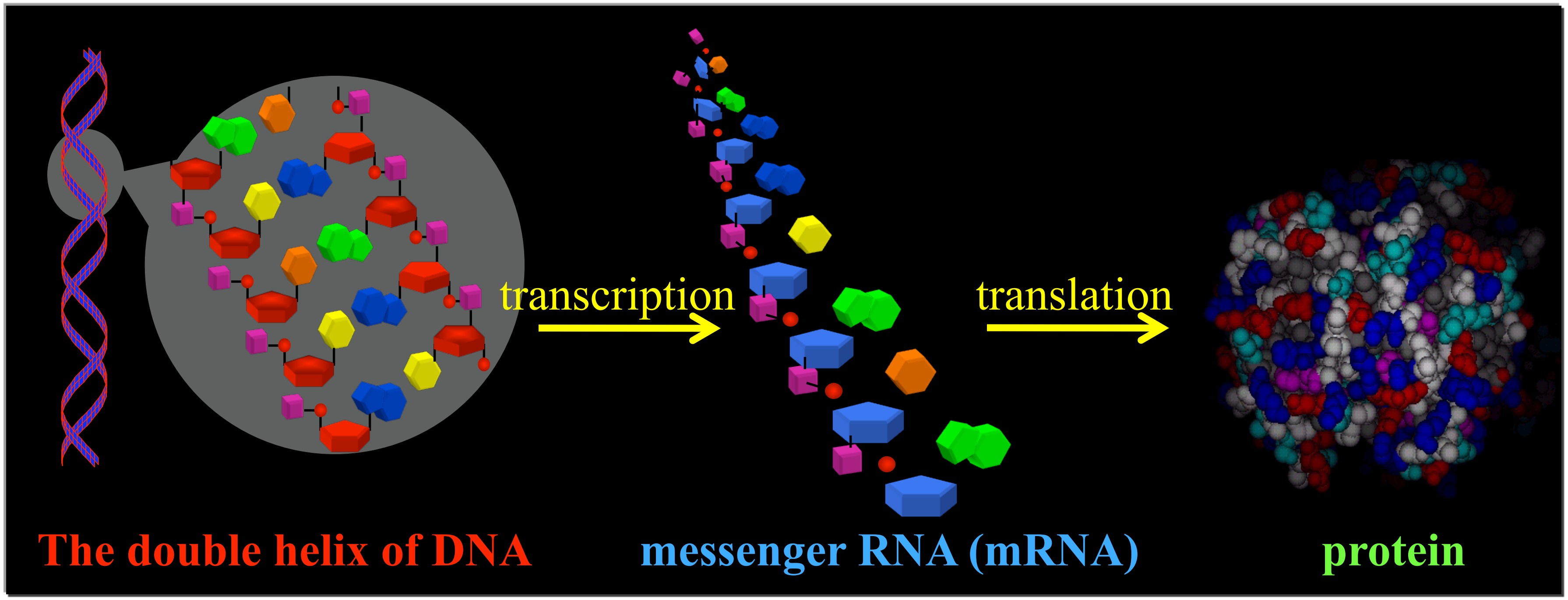
  Life appears NOT to be using more diverse building blocks than "whatever it could invent"
- *The 8 prebiotically plausible amino acids of the standard alphabet are not more diverse than a random sample of 8 taken from taken from abiotic chemical space*

  Life appears NOT to have started with the most diverse set of building blocks available

~ This is, at best, a start point to thinking about the presence or absence of amino acids within the standard genetic code

Different scenarios predict different sets of possible amino acids…

Back to basics…



transcription

translation

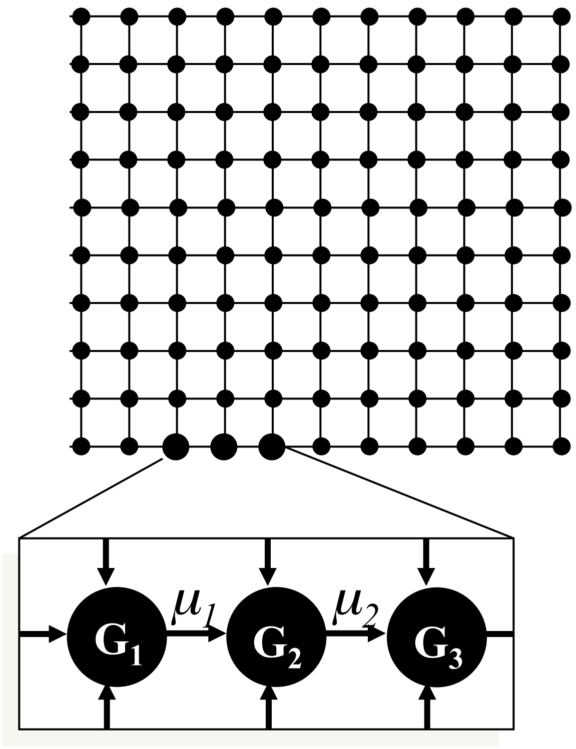
The double helix of DNA

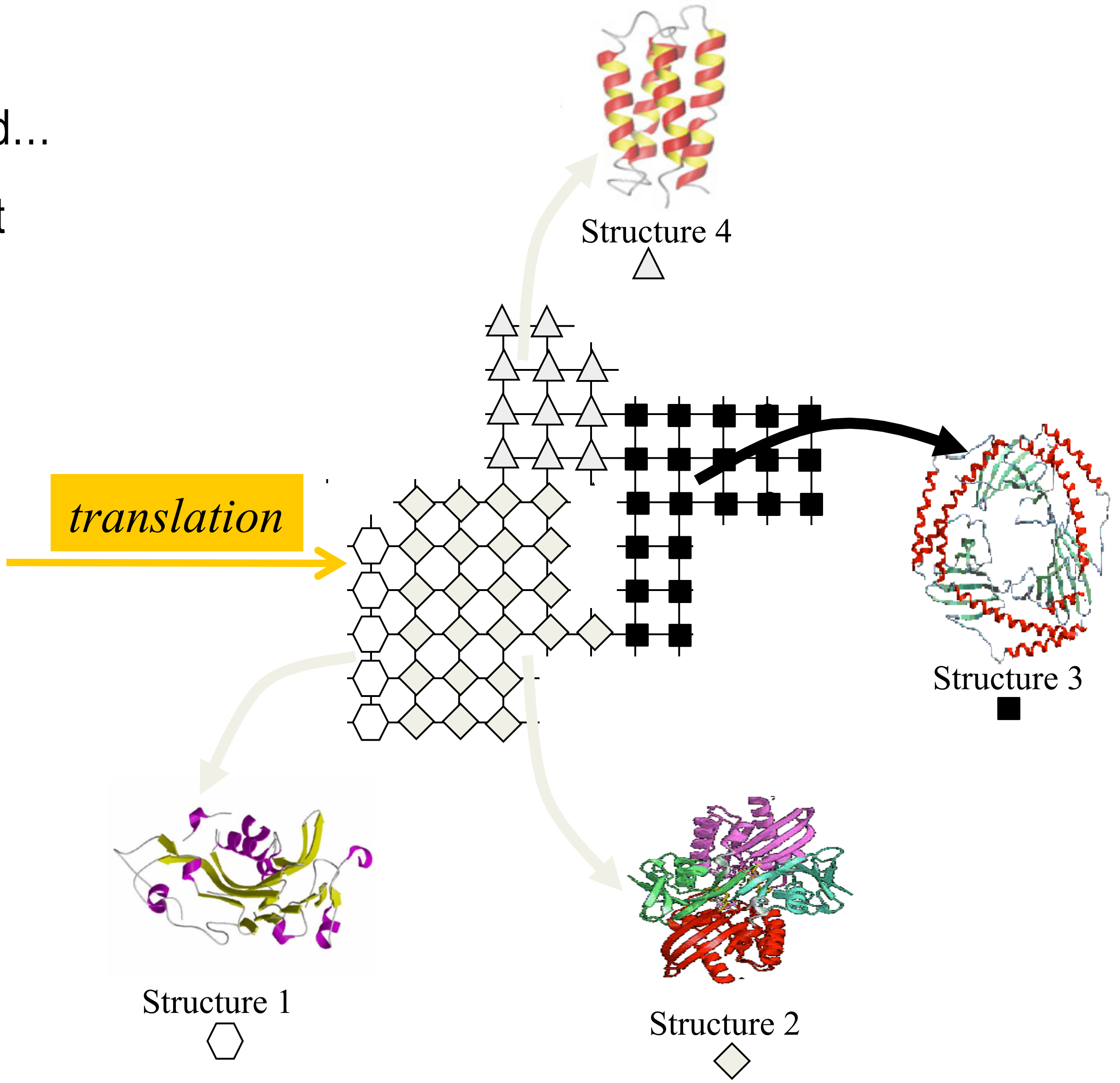
messenger RNA (mRNA)

protein

Imagine each possible gene-sequence as a point on a grid…

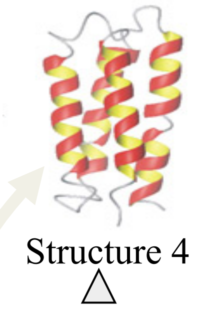
…possible mutations connect these different “genotypes”

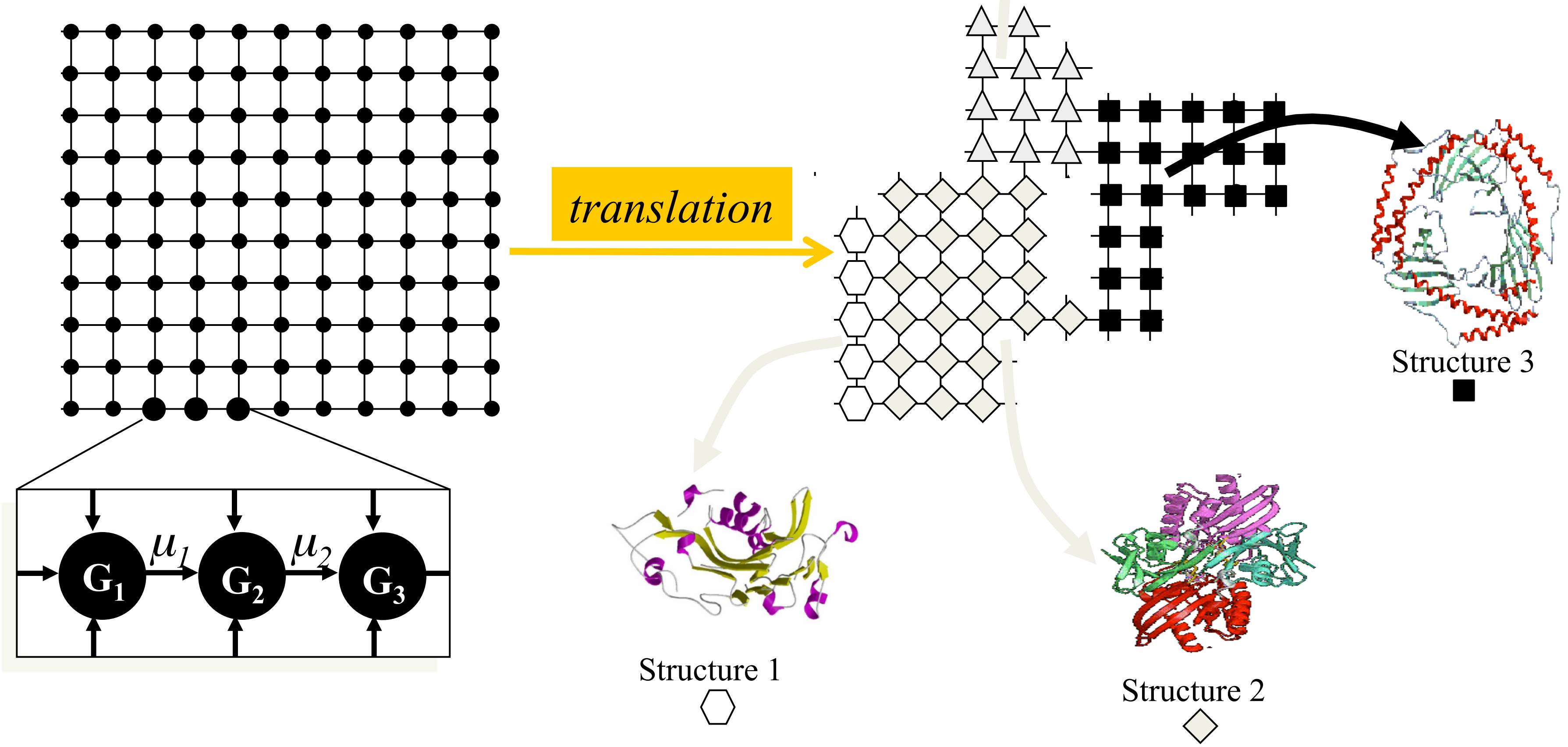




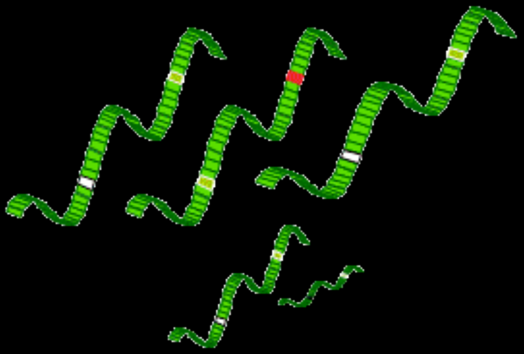
Translation converts these genotypes into corresponding protein structures: 4 stable structures or “no structure” (represented by different shapes & “whitespace” respectively)

Each viable phenotype is encoded by a ‘neutral network’ of multiple genotypes that all encode this same structure: variability in the size of this neutral network corresponds to the robustness of this structure to mutation (structure 2 is more robust than structure 1).





Differences in the proportion of genotypes that allow interconversion can lead to asymmetric mutational distances between structures (for example, every genotype for structure 1 lies within a point mutation of structure 2, but the reverse is not true)



Mutation causes a population of genes to vary from one another

...the organisms we see today represent lineages that have successfully adapted under continually changing demands of natural selection!