Origin of Life Part 3:
The emergence of hierarchy

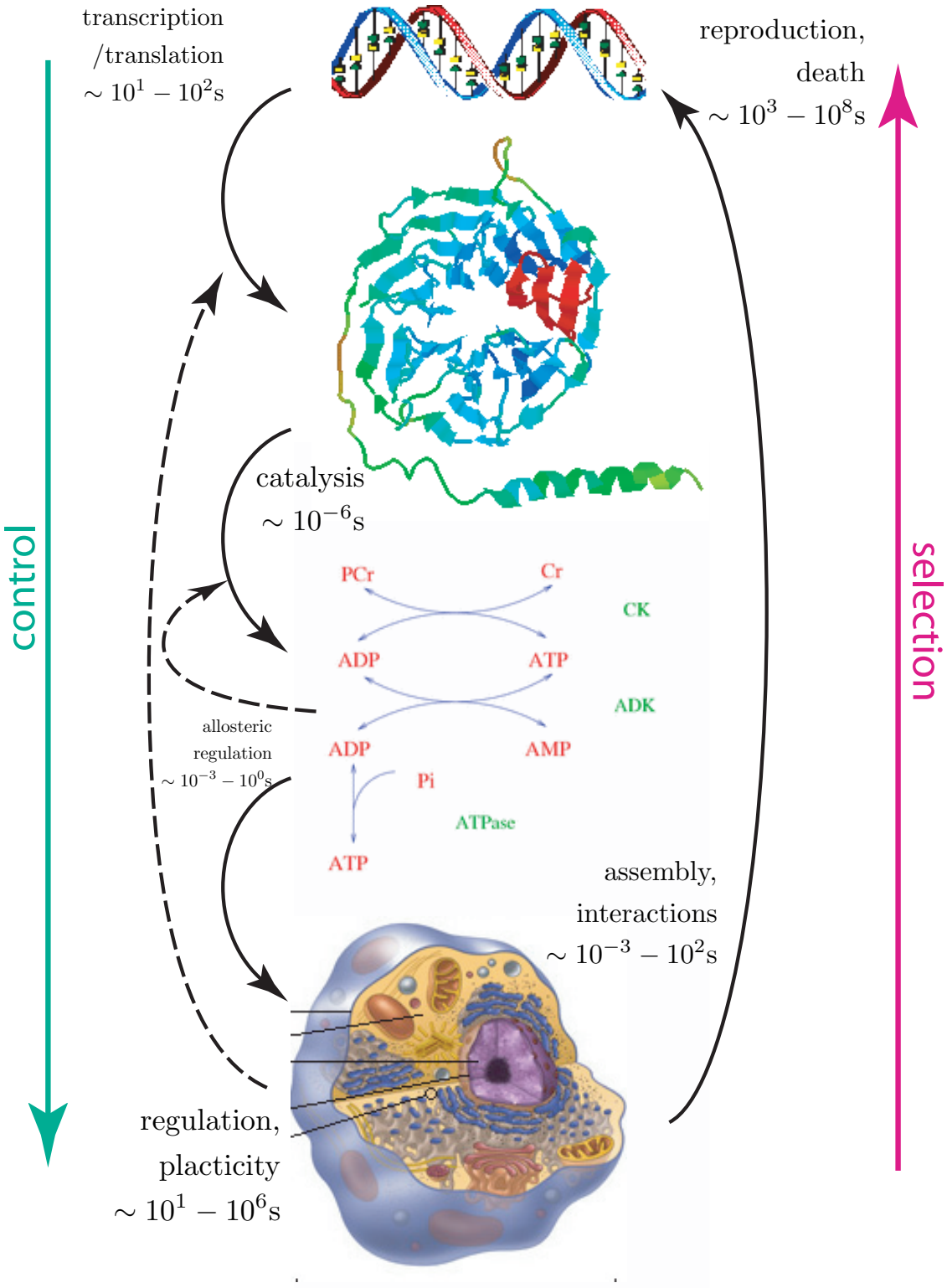
Who controls and who is controlled? Evidence from the genetic code and modularity

Outline

- Hierarchy, control, and the complex notion of information flow in biology
- The genetic code within the control-flow system, and surprising links to metabolism
- Modularity and its significance for origin
- How should we think of directions of causation or information flow, in origins and today?

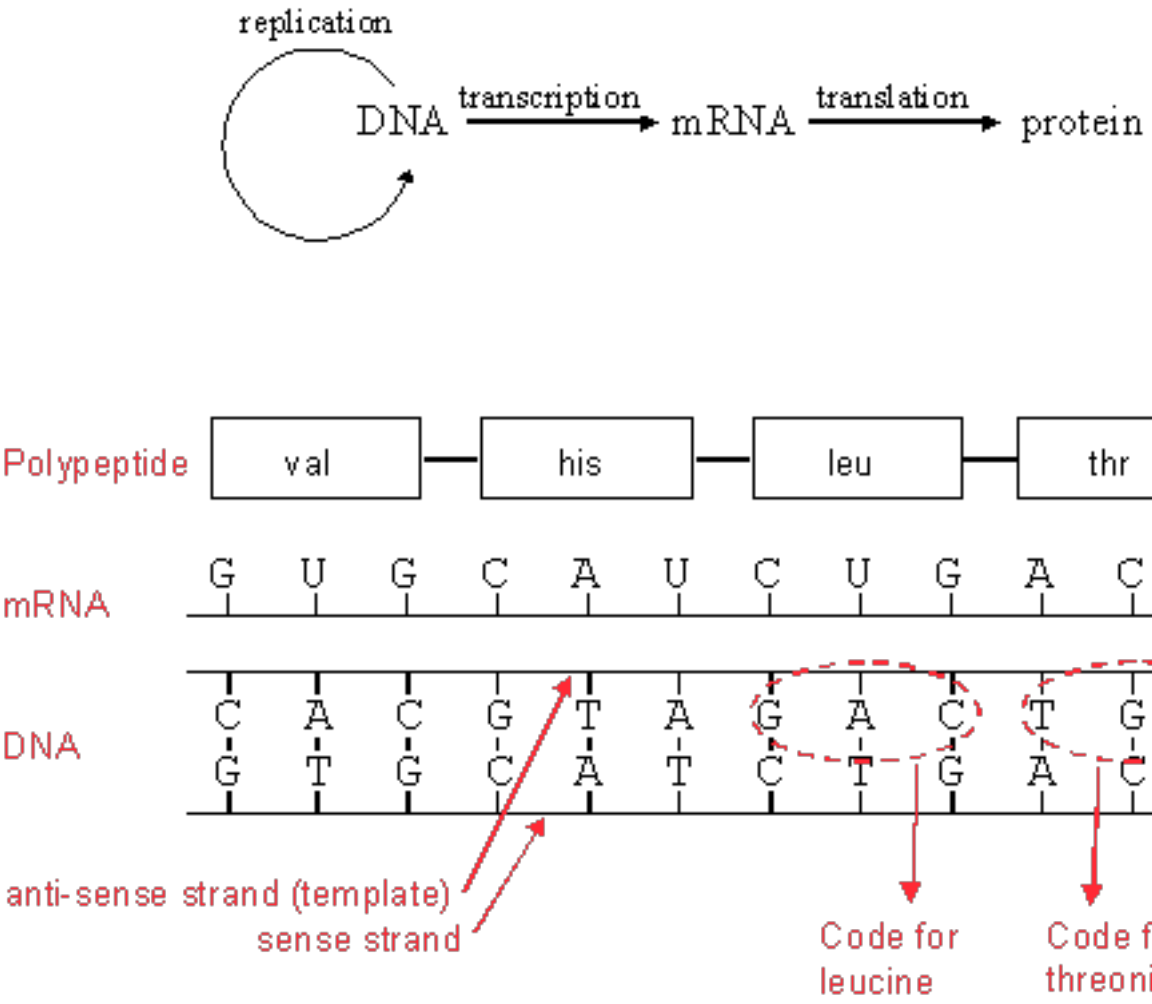
Control flows and error correction

- Long-lived states "control" faster processes
- "Errors" removed by both control and selection
- References are contained in both system and environment



Central Dogma: the place of DNA and RNA in evolutionary and developmental control

- Information flows from DNA genes, to RNA messengers, to proteins during development and physiology
- Correction by natural selection acts on the genome
- Translation uses a complex *ribosome* with both protein and RNA catalysts



http://faculty.clintoncc.suny.edu/faculty/Michael.Gregory/files/Bio%20101/Bio%20101%20Lectures/Protein%20Synthesis/protein.htm

Paradoxes of the emergence of coding and the formation of a code

- The function of coding presumes the most complex apparatus in the cell
- This apparatus currently depends on cellular organization -- and to some extent on protein -- made possible only by coding
- What intermediate forms could have stabilized and guided the emergence of such complexity?

Look for clues to the structure of the code itself

The genetic code is a great biological universal

- Almost all organisms use exactly the same map from 64 NNN to 20 AA and start and stop signals
- Mitochondria and some bacteria use a slightly modified code
- The modifications are interesting in themselves

First Codon \ Second Codon	G	C	A	U	Third Codon
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine* *Threonine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine *(Serine / Stop)*	Threonine	Lysine	Isoleucine (Methionine)	A
	Arginine *(Serine / Stop)*	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
	(Tryptophan)	Serine		Leucine	A
	Tryptophan	Serine		Leucine	G

The standard genetic code

Start
Stop

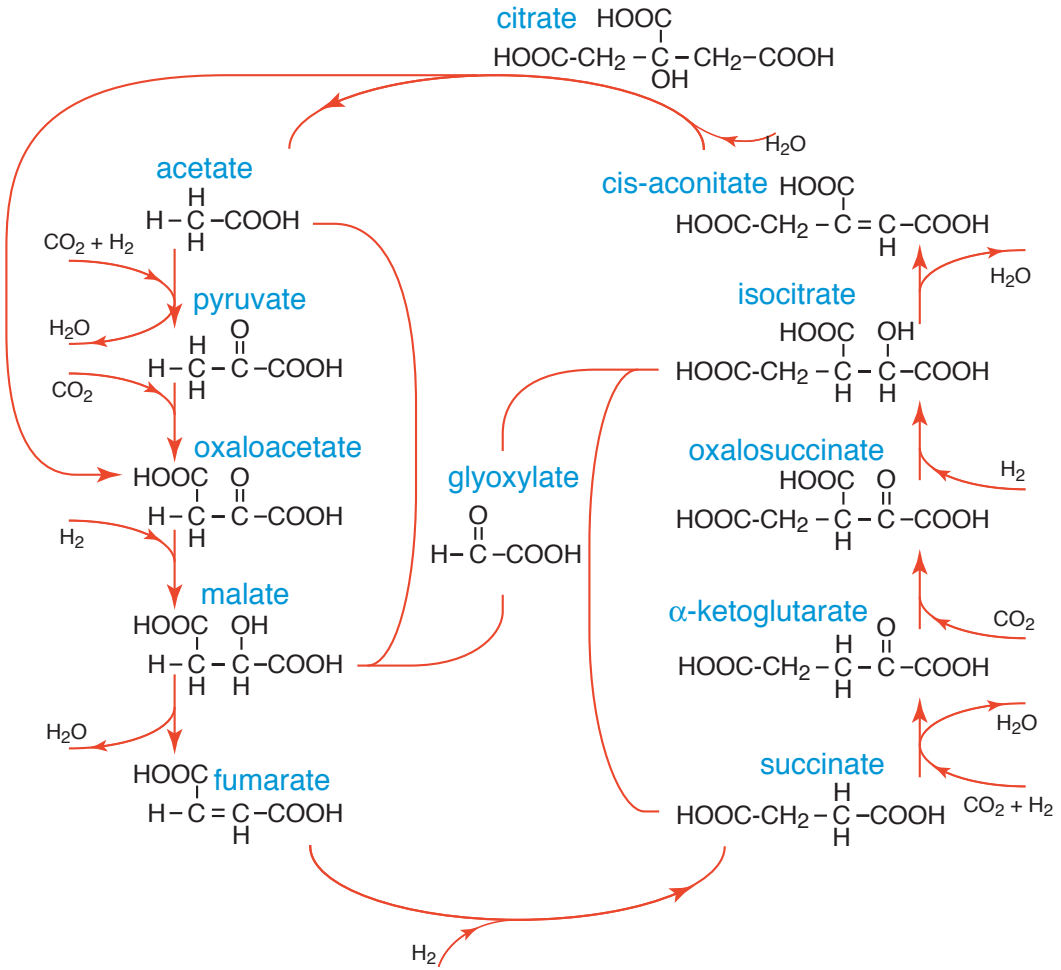
The standard code is special: optimization and/or redundancy?

- For error correction (exact and approximate), not all permutations are equally good
- For both exact correction, and substitution of amino acids with similar properties, the standard code is extremely good

First Codon \ Second Codon	G	C	A	U	Third Codon
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine	Threonine	Lysine	Isoleucine	A
	Arginine	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
		Serine		Leucine	A
	Tryptophan	Serine		Leucine	G

Start
Stop

We will find the organizational structure of metabolism echoed in the genetic code

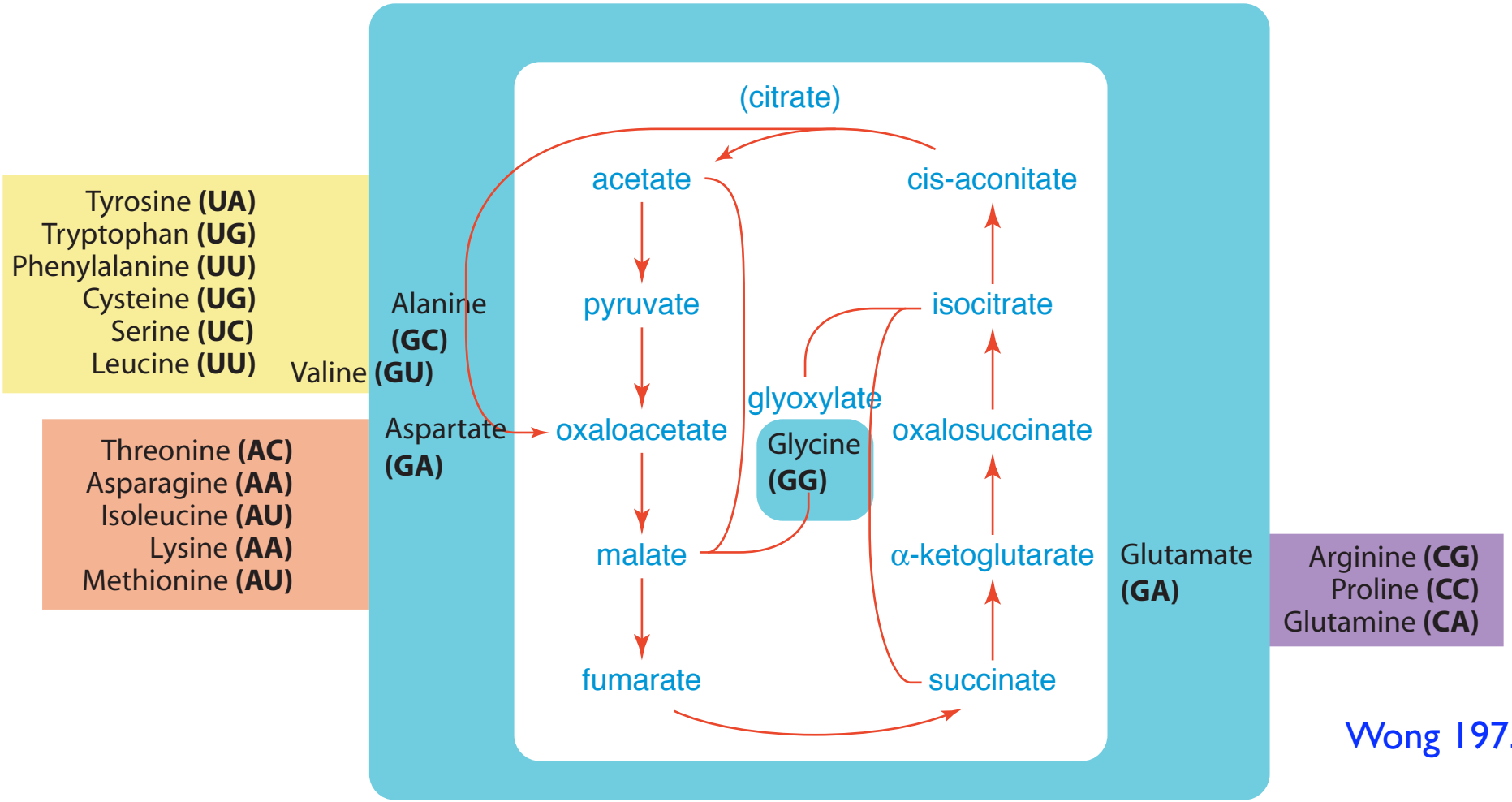


First Codon \ Second Codon	G	C	A	U	Third Codon
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine	Threonine	Lysine	Isoleucine	A
	Arginine	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
		Serine		Leucine	A
	Tryptophan	Serine		Leucine	G

Start
Stop

The first base of the code tells about backbones from rTCA

- Carbon backbone is specified if U, A, or C is first
- A single kind of reaction is specified if G is first



Wong 1975

Second codon gives physical properties

- A second indicates amino acids that dissolve readily in water
- U second indicates acids that dissolve better in oils
- G and C are intermediate

	G	C	A	U	
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine	Threonine	Lysine	Isoleucine	A
	Arginine	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
		Serine		Leucine	A
	Tryptophan	Serine		Leucine	G
			Hydrophilic	**Hydrophobic**	

Italic: Capture (?)
Font size ~ age

Start
Stop

Third base correlates with biosynthetic complexity; (?) with the emergence of coding

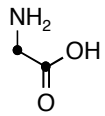
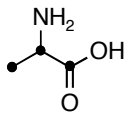
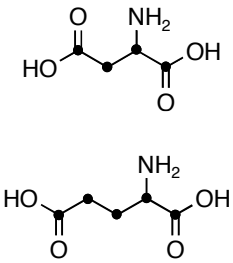
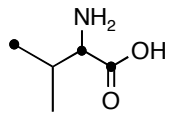
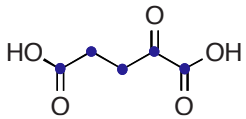
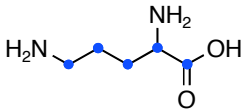
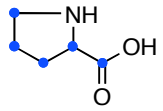
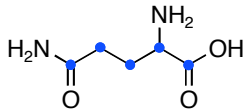
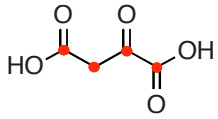
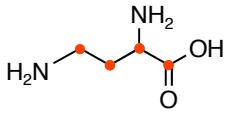
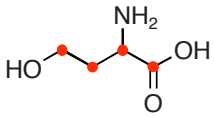
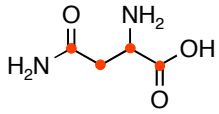
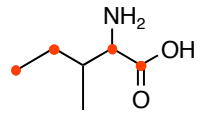
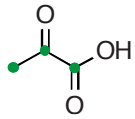
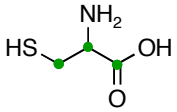
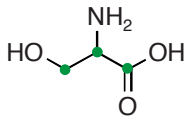
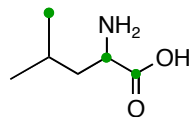
- 15 amino acids are structurally simple, and draw from small regions of the metabolic chart
- Simple acids are either totally redundant in the third base, or specified only at purine/pyrimidine level
- Complex acids, like start/ stop, are specified at third-codon position, often as minorities

First Codon \ Second Codon	G	C	A	U	Third Codon
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine	Threonine	Lysine	Isoleucine	A
	Arginine	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
		Serine		Leucine	A
	Tryptophan	Serine		Leucine	G

Start
Stop

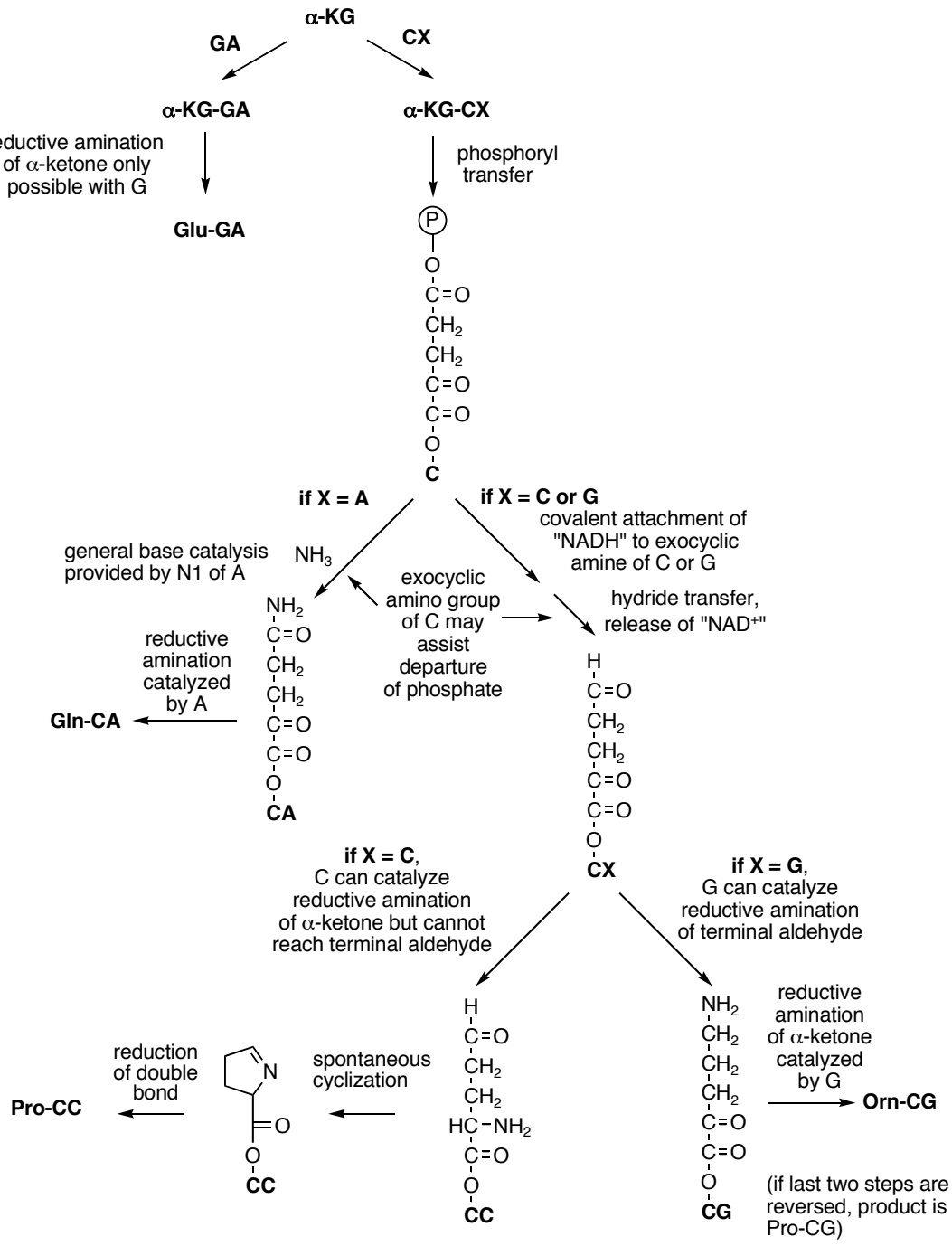
So was there something here before coding?

A two-base "code" is more regular than today's

first position	second position: G	C	A	U
G	**Gly**	**Ala**	**Asp/Glu**	**Val**
	Gly	**Ala**	**Asp/Glu**	**Val**
				
C (α-ketoglutarate)	**Arg**	**Pro**	**Gln**	**Leu**
	Orn	**Pro**	**Gln**	**?**
				**?**
A (oxaloacetate)	**Ser / Arg**	**Thr**	**Asn**	**Ile**
	Dab	**Hsr**	**Asn**	**Ile**
				
U (Pyruvate)	**Cys**	**Ser**	**Tyr/stop**	**Leu**
	Cys	**Ser**	**X**	**Leu**
			**X**	

Capture logic of the code as an "instruction set" in a decision tree

- Start with all pairs of dinucleotides and alpha-keto acids from rTCA
- For each dinucleotide, assign a systematic reaction type
- Dead ends predict no association
- This system compactly represents the associations found in the first two bases of the modern code!



Deviations occur in *captured* positions

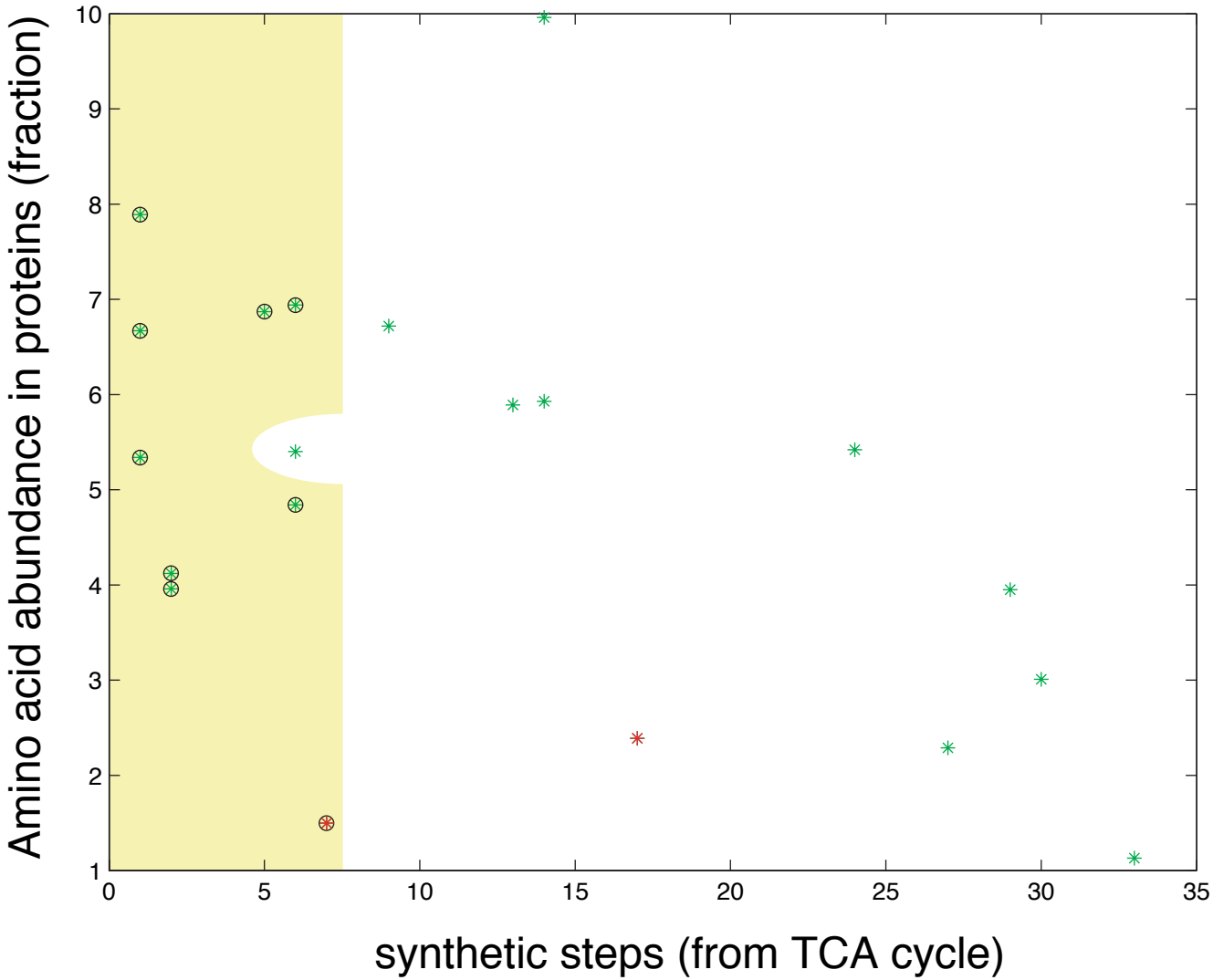
- There is no way to make a pyruvate homologue at UA, leaving it open
- AG merely duplicates either CG or UC
- UA and AU are used for start/stop today

First Codon \ Second Codon	G	C	A	U	Third Codon
G	Glycine	Alanine	Aspartate	Valine	U
	Glycine	Alanine	Aspartate	Valine	C
	Glycine	Alanine	Glutamate	Valine	A
	Glycine	Alanine	Glutamate	Valine	G
C	Arginine	Proline	Histidine	*Leucine*	U
	Arginine	Proline	Histidine	*Leucine*	C
	Arginine	Proline	Glutamine	*Leucine* *Threonine*	A
	Arginine	Proline	Glutamine	*Leucine*	G
A	*Serine*	Threonine	Asparagine	Isoleucine	U
	Serine	Threonine	Asparagine	Isoleucine	C
	Arginine (Serine / Stop)	Threonine	Lysine	Isoleucine (Methionine)	A
	Arginine (Serine / Stop)	Threonine	Lysine	Methionine	G
U	Cysteine	Serine	Tyrosine	Phenylalanine	U
	Cysteine	Serine	Tyrosine	Phenylalanine	C
	(Tryptophan)	Serine		Leucine	A
	Tryptophan	Serine		Leucine	G

Start
Stop

Biosynthetic simplicity and essentiality structures evolutionary specialization

- Biosynthetic cost reflected in trophic ecology

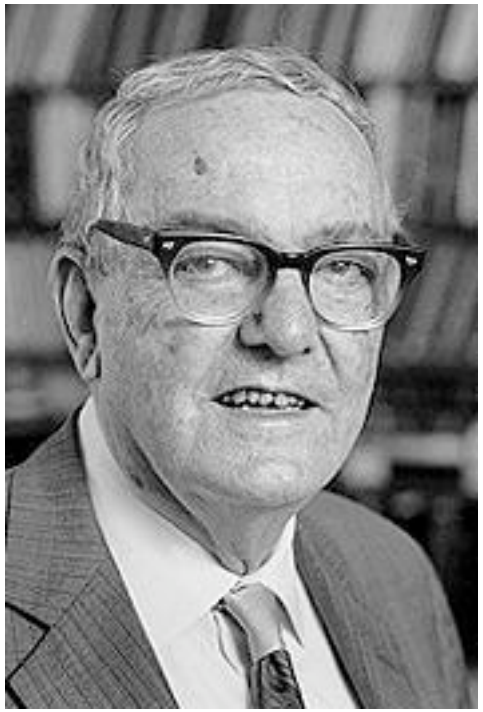


Summary comments concerning the genetic code

- Main function of coding would work with an arbitrary code, were it not for errors
- The actual non-randomness of the observed code is correlated with biosynthesis
- The same non-randomness is compatible with prior constraint, co-evolution, or optimization

Modularity as the key to complexity?

- Herb Simon (on theory of organization); parable of the watchmakers
- Without intermediate stability of modules, complexity is improbable
- Given observed complexity, mechanisms with intermediate modularization have higher (Bayesian) posterior probability

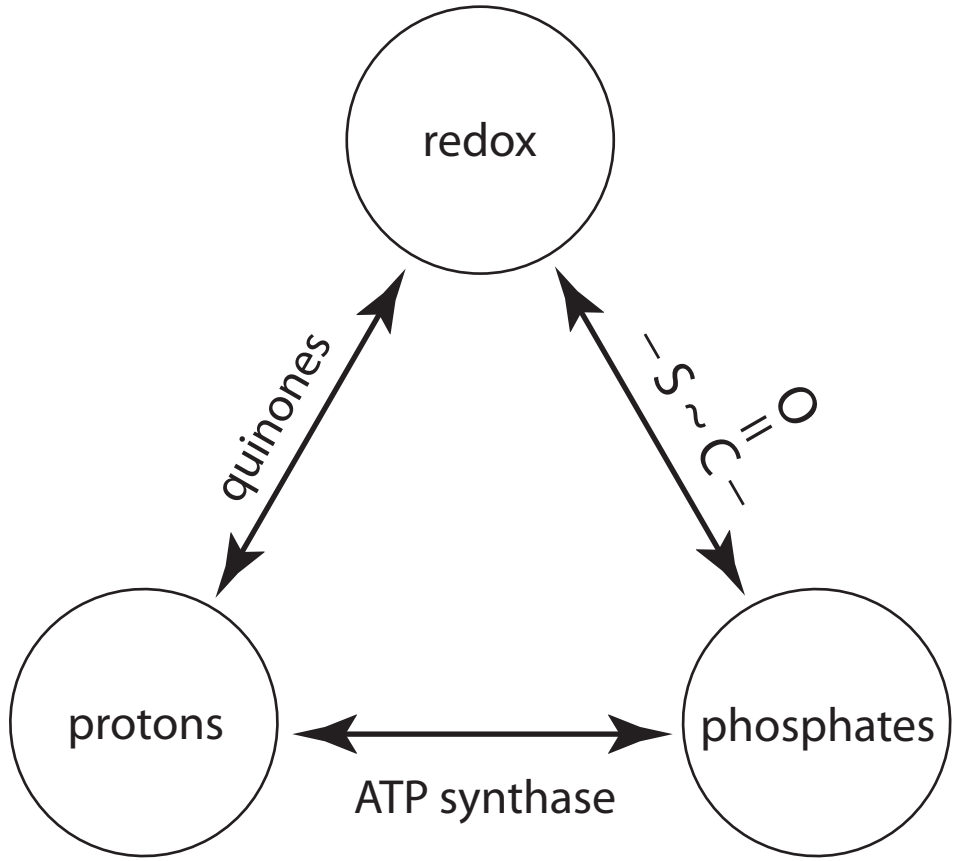


"... if one would be Alexander, one should be born into a world where large stable political systems already exist"

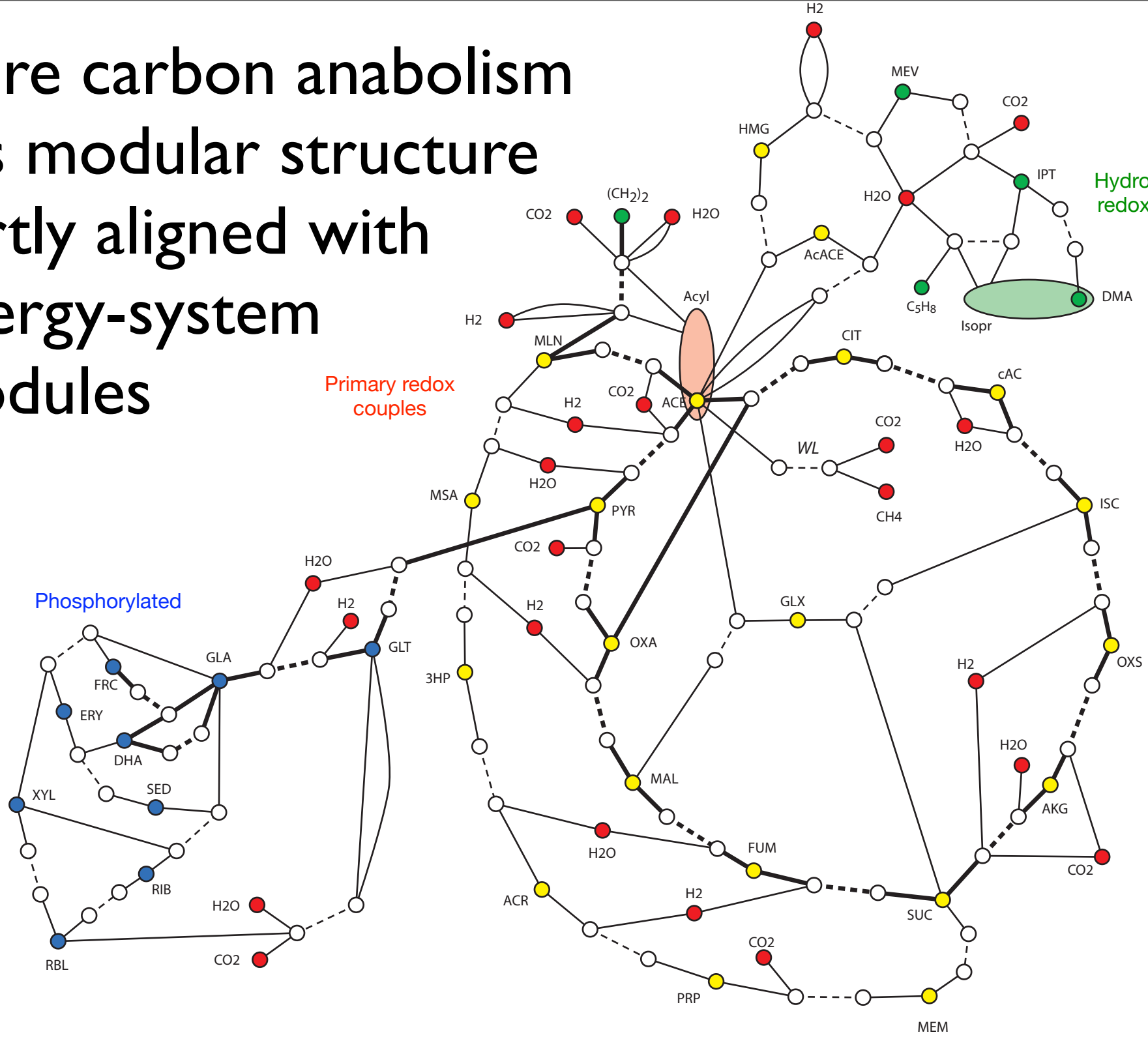
-- H.A. Simon (1962)

Modularity and microenvironments

- Three energy systems do different things; have different requirements
  - e^-: makes high-energy bonds; requires quantum structure
  - P_i: dehydrates to make polymers; needs fnal groups
  - p^+: no QM and powers motors; needs compartments
- Each system could run in a suitable geochemical environment
- Cell processes couple these: balance and buffer them

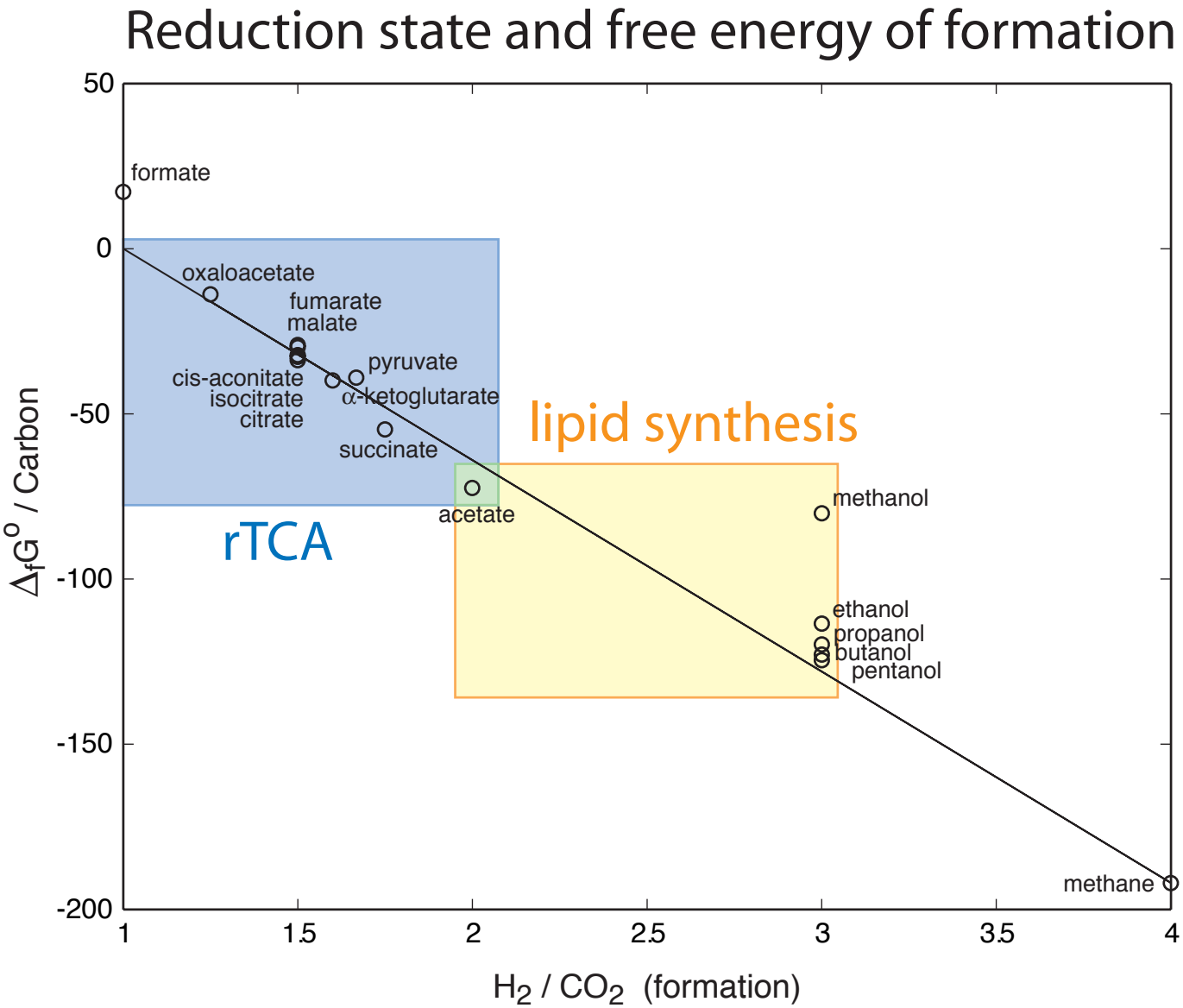


Core carbon anabolism has modular structure partly aligned with energy-system modules



Modularity in redox-driven network links

energy flow, chemical redundancy, and molecule classes



The earliest RNA controllers: emergent individuality or canalization?

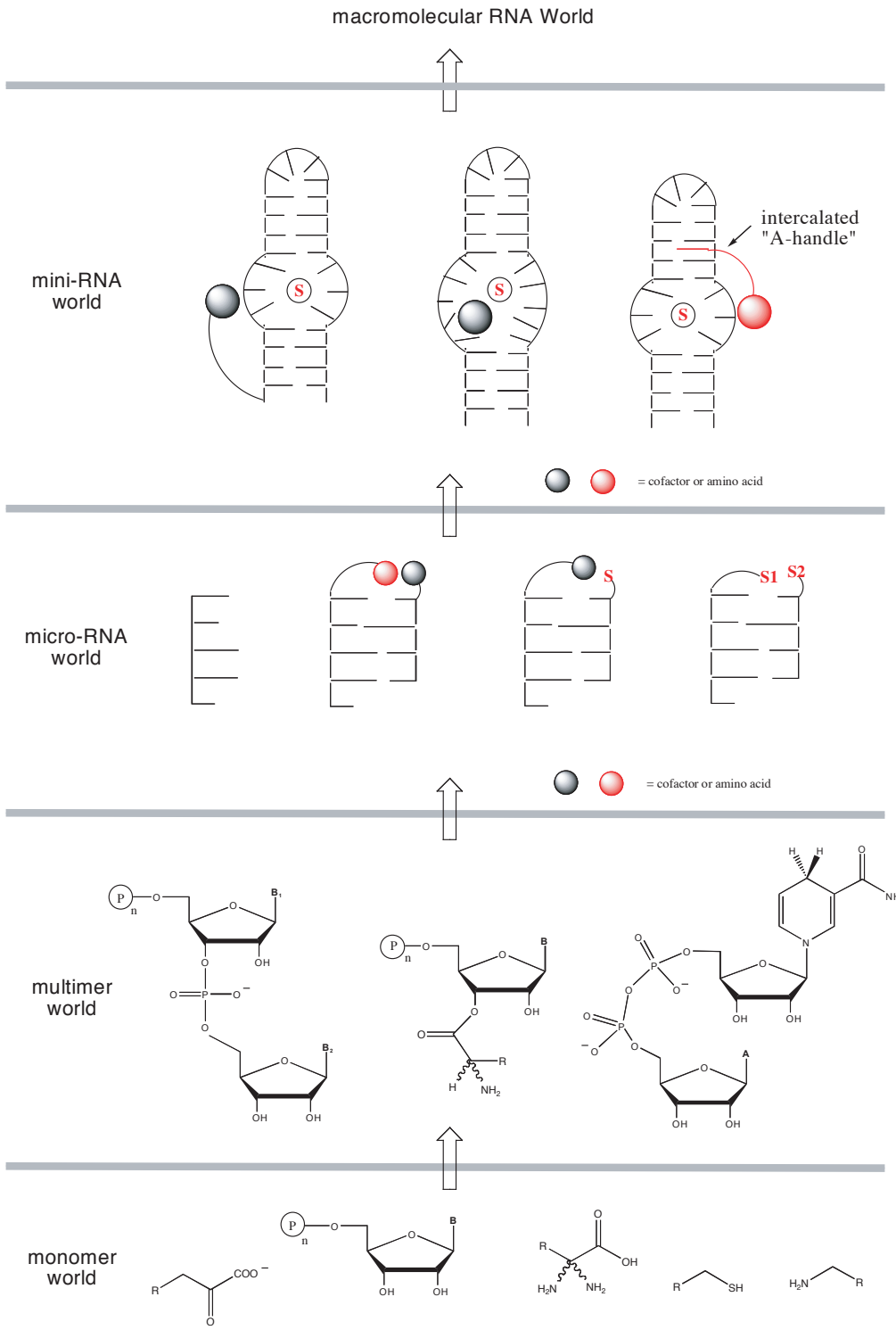
- Darwinian dynamics based on heritable variation and selection at individual level
- Competing RNAs or RNA hypercycles in primordial soup are like molecular individuals
- But weaker control would be needed merely to couple systems that already existed
- Canalization a better biological concept here than individuality?

Canalization: a measure of the ability of a population to produce the same phenotype regardless of variability of its environment or genotype.

The ladder of catalysis

hierarchy recapitulates biosynthesis

- Reconsider the emergence of RNA control from a perspective of canalization
- Between core metabolites and macromolecules lie a ladder of intermediate forms
- Much organization of metabolism is governed by such forms today
- Suppose they were originally selected for service of metabolism, not for reproductive competition



Implications of the directions of information flow for origin-of-life inference

Control-first

Abiotically-generated organics (primordial soup, not-metabolism)

Self-replicating RNA

Ribozymes for metabolic reactions

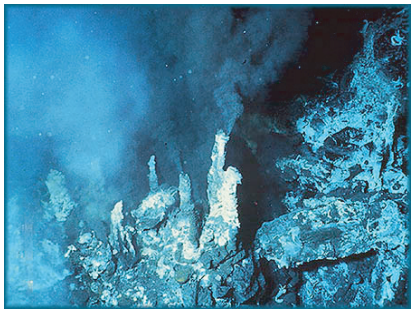
Metabolism contingent on control

Darwinian selection

Complexity/ relevance

Time

Metabolism-first



Restricted set of C, e^-, P, etc. sources

Self-organized organosynthetic network

Molecular replication through template-directed ligation

Metabolism recapitulates biogenesis

Chemical Self-organization

Darwinian selection

Complexity/ relevance

Further reading

- Smith, Eric and Morowitz, Harold J **Universality in intermediary metabolism** Proc. Nat. Acad. USA **101:** 13168, 2004
- Wong Tse-Fei **A co-evolution theory of the genetic code** Proc. Nat. Acad. USA **72:**1909, 1975
- Copley, Shelley D, Smith, Eric and Morowitz, Harold J **A mechanism for the association of amino acids with their codons and the origin of the genetic code** Proc. Nat. Acad. USA **102:**4442, 2005
- Copley, Shelley D, Smith, Eric and Morowitz, Harold J **The origin of the RNA world: co-evolution of genes and metabolism** Bioorganic Chemistry **35:**430, 2007
- Sinanoglu, Oktay and Lee, Lih-Syng **Finding the possible mechanisms for a given type of overall reaction. The case of the (A+B to C+D) overall reaction types** Theoretica Chimica Acta **51:**1, 1979; **On the algebraic construction of chemistry from quantum mechanics. A fundamental valency vector field defined on the euclidean 3-space and its relation to the Hilbert space** Theoretica Chimica Acta **65:**249, 1984
- C. H. Waddington **Canalization of development and the inheritance of acquired characters** Nature **14**:563, 1942
- Simon, Herbert A. **The architecture of complexity** Proceedings of the American Philosophical Society **106**:467, 1962