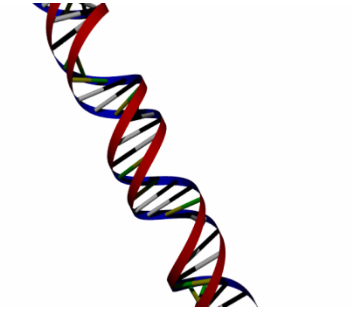
Evolution of the Genetic Code



Supratim Sengupta

Centre for Computational Biology & Bioinformatics
Jawaharlal Nehru University, New Delhi, India.

KITP, Santa Barbara – Feb. 8, 2010

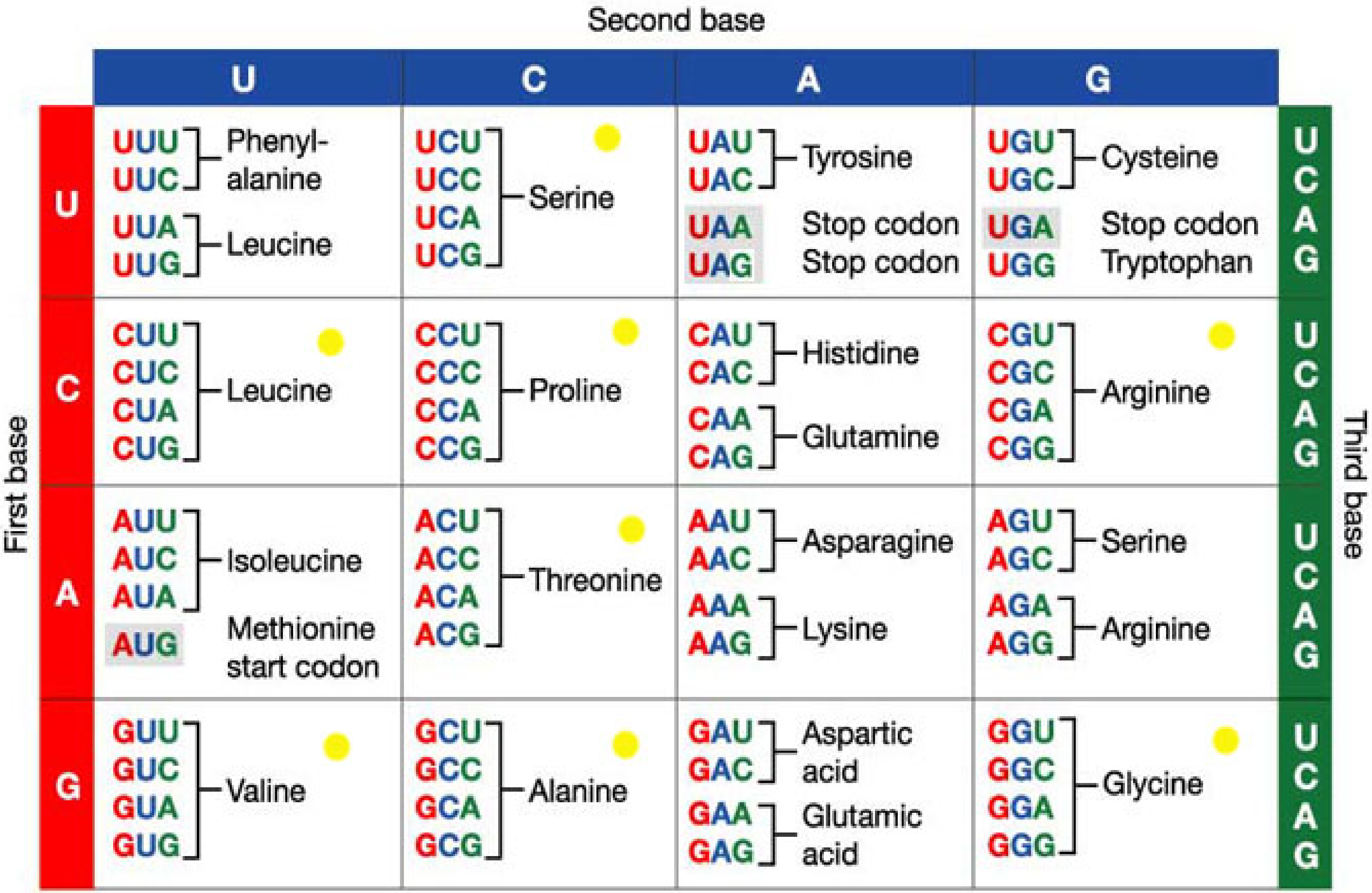
Major Transitions in Evolution

(Maynard-Smith & Szathmary)

- Replicating molecules → Populations of molecules in compartments.
- Independent replicators → Chromosomes.
- RNA as genes and enzymes → DNA & proteins – Development of the Genetic Code and its subsequent evolution.
- Prokaryotes → Eukaryotes (chromosomes enclosed within nucleus)
- Asexual clones → Sexual populations
- Single celled eukaryotes → Multicellular organisms (fungi, plants, animals)
- Solitary individuals → colonies (Division of labour and origin of cooperation)
- Primate societies → Human societies (evolution of the nervous system & origin of language) .

Genetic Code Evolution

- Pre-LUCA Phase
- Post-LUCA Phase



	Second base								
First base	U		C		A		G		Third base
U	UUU	Phenylalanine	UCU	Serine ●	UAU	Tyrosine	UGU	Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	Leucine	UCA		UAA	Stop codon	UGA	Stop codon	A
	UUG		UCG		UAG	Stop codon	UGG	Tryptophan	G
C	CUU	Leucine ●	CCU	Proline ●	CAU	Histidine	CGU	Arginine ●	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Isoleucine	ACU	Threonine ●	AAU	Asparagine	AGU	Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lysine	AGA	Arginine	A
	AUG	Methionine start codon	ACG		AAG		AGG		G
G	GUU	Valine ●	GCU	Alanine ●	GAU	Aspartic acid	GGU	Glycine ●	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

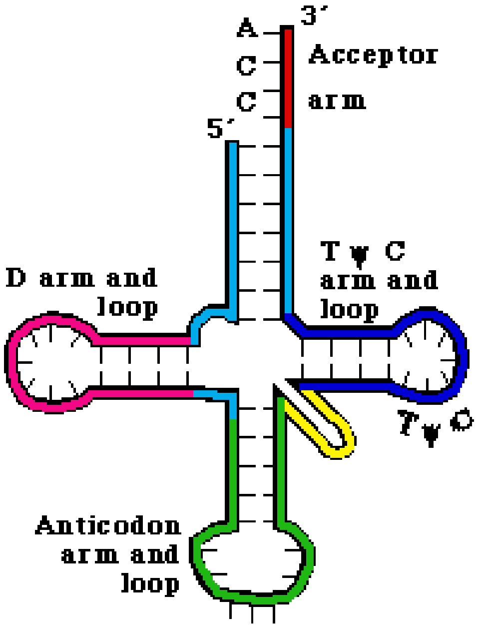
Khorana, Holley, Nirenberg (**Nobel 1968**), Matthei, Crick

Number of Synonymous changes = 137

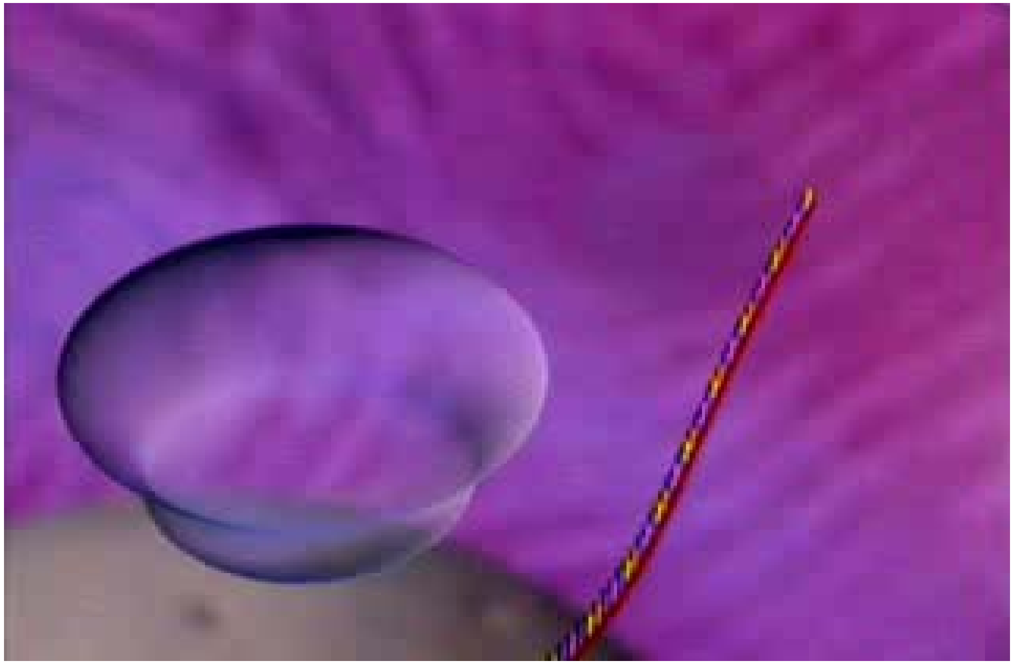
Number of Non-Synonymous Changes = 392

- The Genetic Code provides a recipe for synthesis of proteins from DNA sequences.

DNA sequence $\xrightarrow{\text{Transcription}}$ RNA $\xrightarrow{\text{Translation}}$ Protein



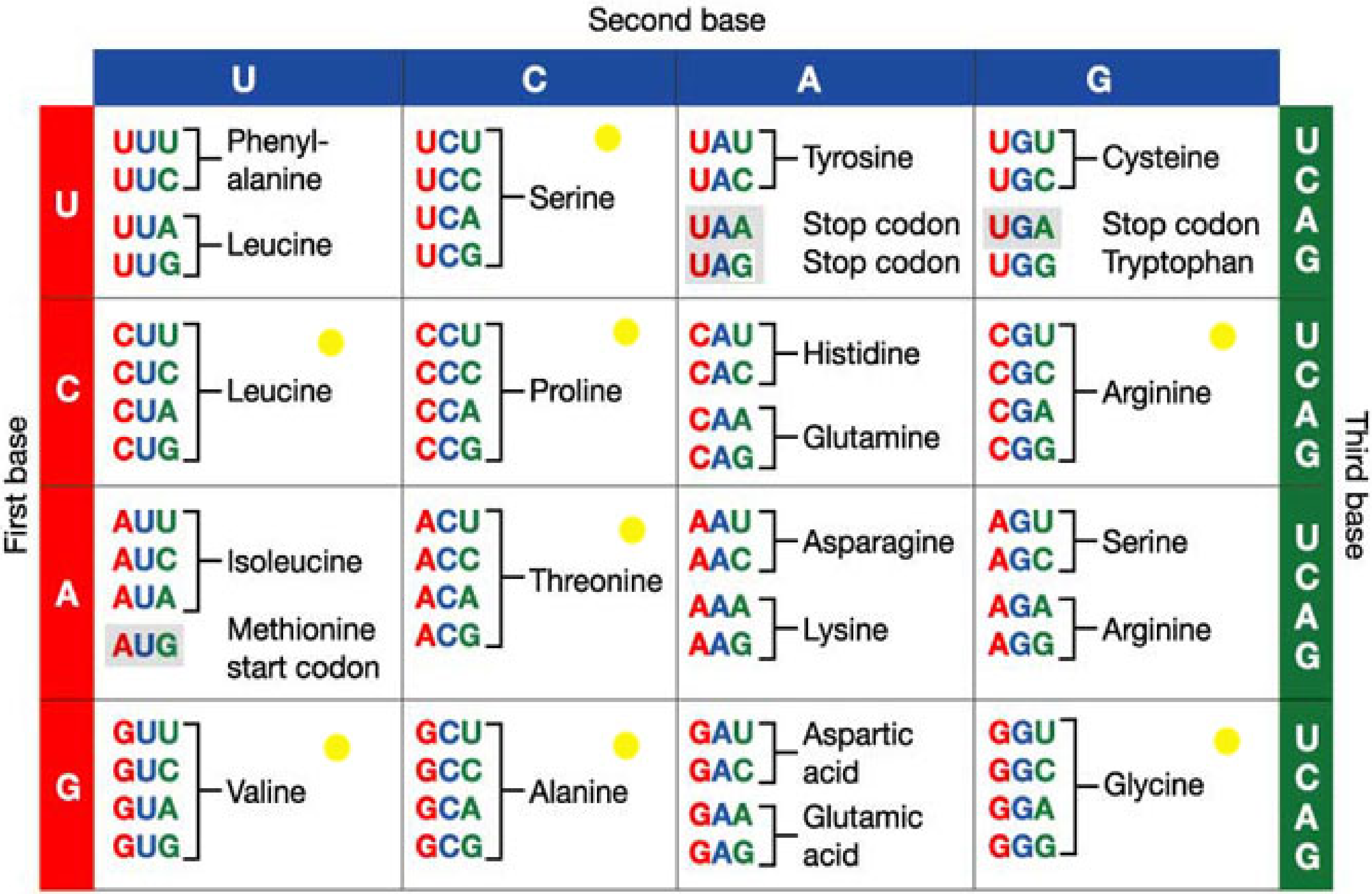
$$_1\text{CAU}_3$$
$$|\;\;|\;\;|$$
$$3' \text{———}\ {}_3\text{GUA}_1\ \text{———} 5'$$



Source : *National Human Genome Research Institute*

Wobble Pairing : Pairing between 1’st anti-codon position and 3’rd codon position.

→ Allows greater pairing flexibility compared to pairing at other codon/anti-codon positions which strictly follow Watson-crick rules.



	Second base				
First base	U	C	A	G	Third base
U	UUU UUC Phenylalanine UUA UUG Leucine	● UCU UCC UCA UCG Serine	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	● CUU CUC CUA CUG Leucine	● CCU CCC CCA CCG Proline	CAU CAC Histidine CAA CAG Glutamine	● CGU CGC CGA CGG Arginine	U C A G
A	AUU AUC AUA Isoleucine AUG Methionine start codon	● ACU ACC ACA ACG Threonine	AAU AAC Asparagine AAA AAG Lysine	AGU AGC Serine AGA AGG Arginine	U C A G
G	● GUU GUC GUA GUG Valine	● GCU GCC GCA GCG Alanine	GAU GAC Aspartic acid GAA GAG Glutamic acid	● GGU GGC GGA GGG Glycine	U C A G

Khorana, Holley, Nirenberg (**Nobel 1968**), Matthei, Crick

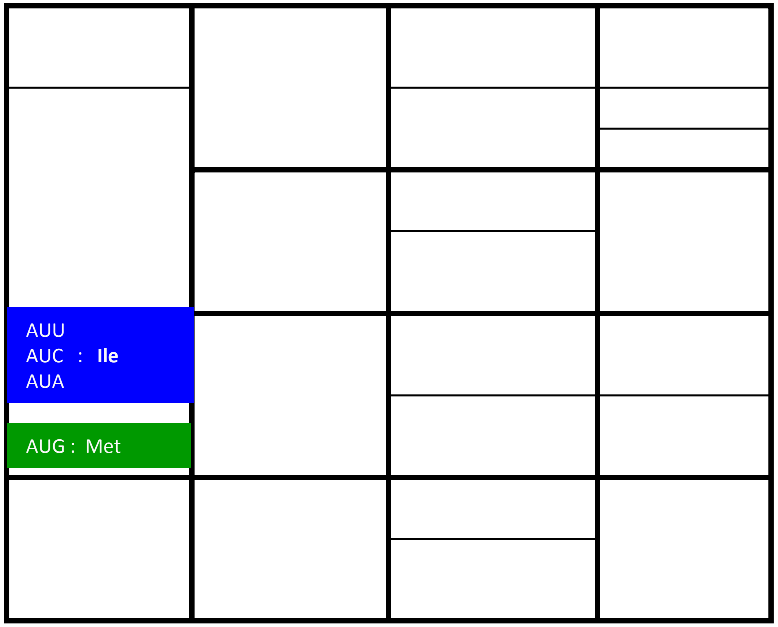
Number of Synonymous changes = 137

Number of Non-Synonymous Changes = 392

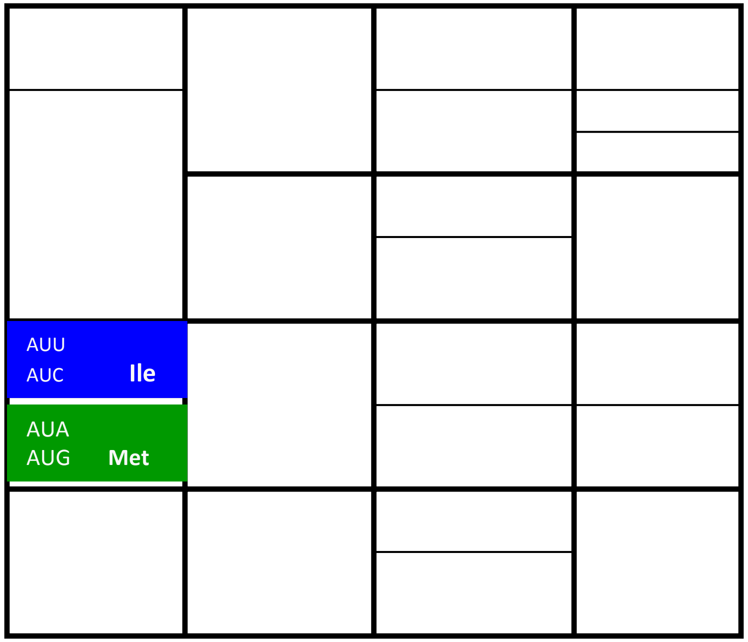
The code is ***not*** universal.

Reassignments of one or more codons to a different amino acid have occurred in many separate groups of organisms.

Codon reassignments ➔ Alternative Genetic Codes



AUA: Ile ➔ Met



Total number of observed reassignments = **44** (mitochondrial + nuclear + bacterial)

11 *distinct* codon reassignments - mitochondrial genome.

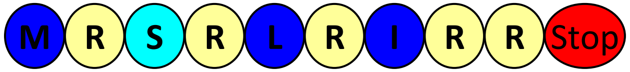
4 *distinct* codon reassignments - nuclear genome.

1 *distinct* codon reassignment - bacteria

Implications of codon reassignment

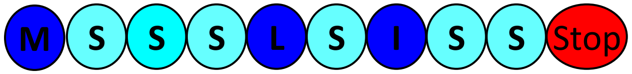
- mRNA sequence: **A U G AGGU C A A G G C U UAGG A U CAGGAGG U G**
  A

Protein *before* reassignment :



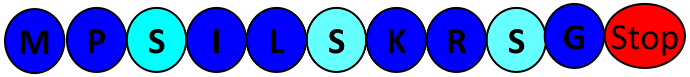
AGG : Arg ➔ Serine – observed in vertebrates

Protein *after* reassignment :



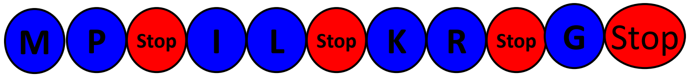
- mRNA sequence: **AUG CCA UCA AUU CUU UCA AAA AGG UCA GGG UGA**

Protein *before* reassignment :

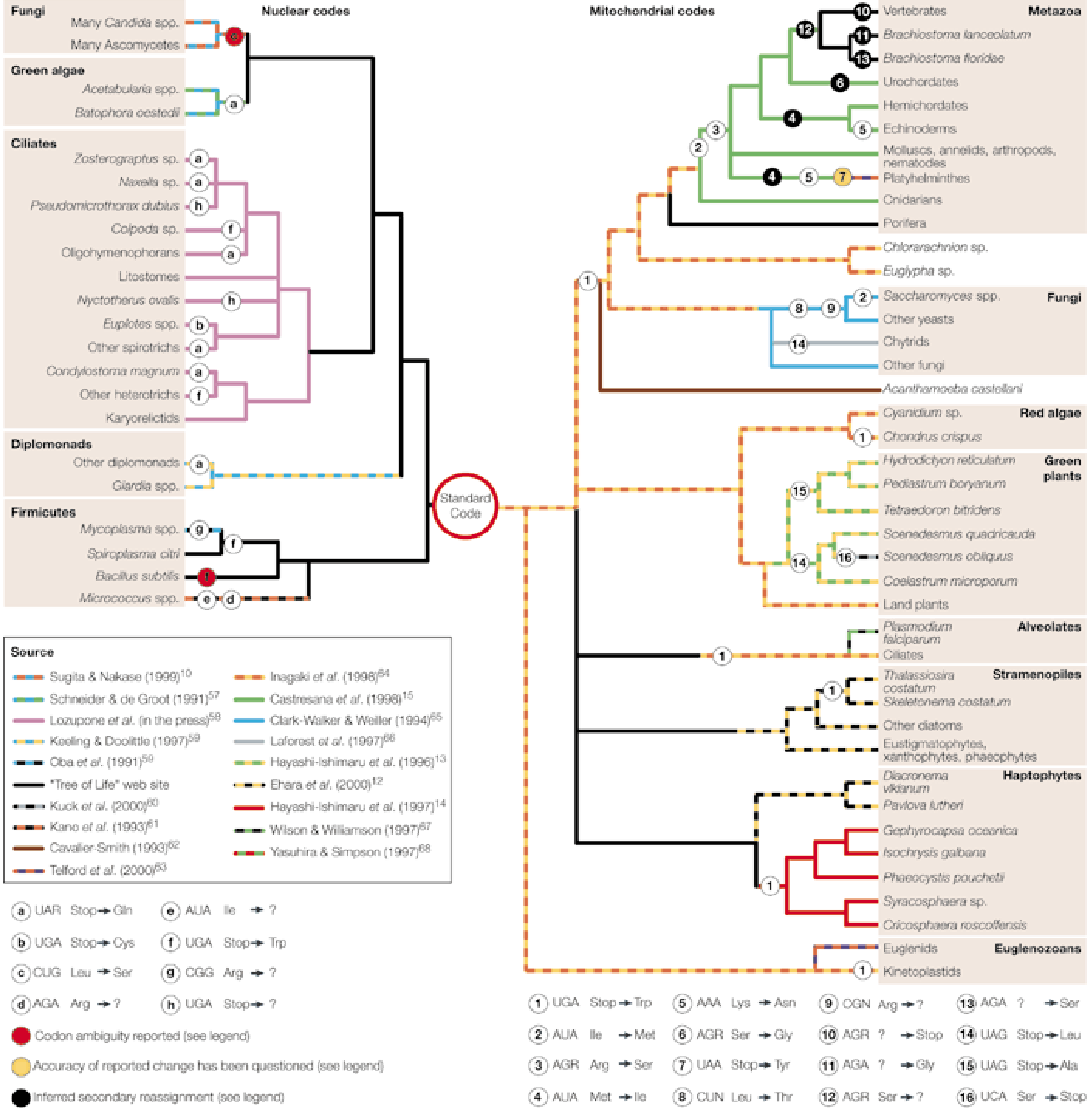


UCA : Ser ➔ Stop – observed in some green algae

Protein *after* reassignment :



Codon reassignment changes the amino acid sequence of ***every*** protein which uses the codon.



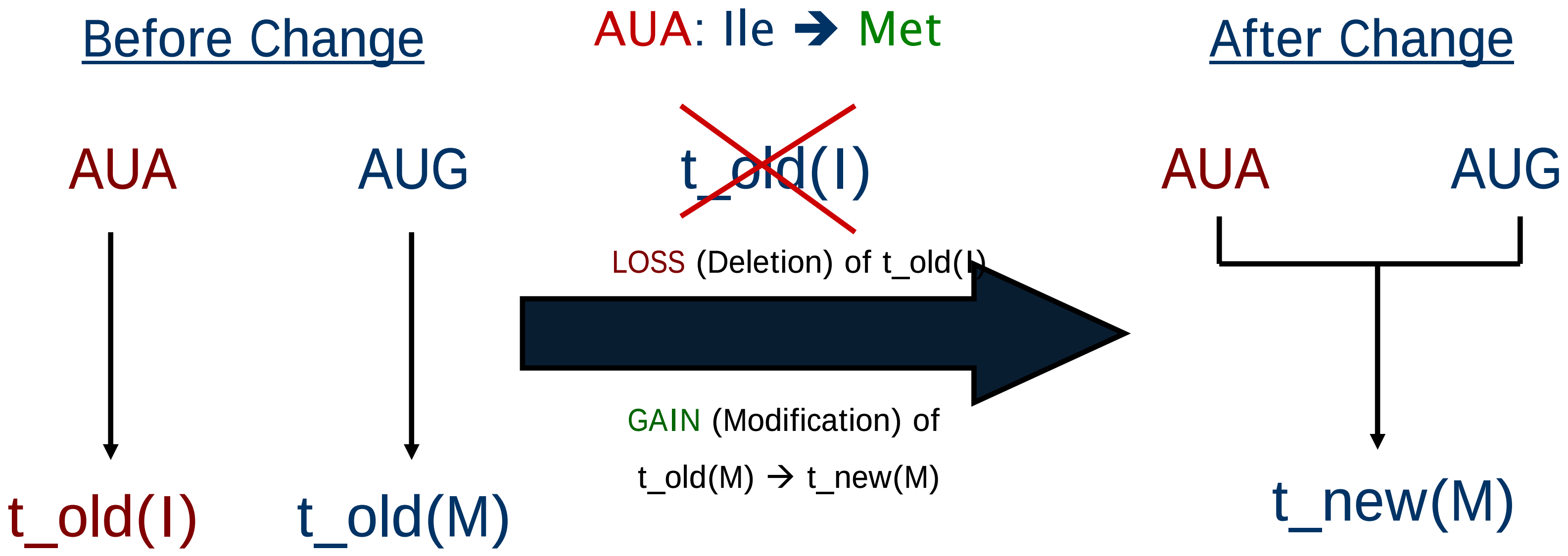
Nature Reviews | Genetics

Knight *et al.* (2001)
Nature Reviews Genetics

Molecular Basis of Codon Reassignment

Example : AUA was reassigned from Ile to Met during the early evolution of the mitochondrial genome

	Before Change				After Change			
Amino Acid	**Ile**			**Met**	**Ile**		**Met**	
Codon	$_1$AUU$_3$	AUC	AUA	AUG	$_1$AUU$_3$	AUC	AUA	AUG
Anti-codon	$_3$UAG$_1$	UAG	UA^{k2}C	UAC	$_3$UAG$_1$	UAG	UA^{f5}**C,U**	UA^{f5}**C,U**
Effect	Proteobacteria and some protist mitochondria possess an **extra tRNA-Ile** (^{k2}CAU) with a modified base that translates **AUA only**.				**Loss** of the ***special*** tRNA-Ile(^{k2}CAU) and **Gain** of function of the tRNA-Met so that it translates **both AUA** and **AUG**. UAU (Sea Squirts), ^{f5}CAU (insects).			



Before Change

AUA → t_old(I)

AUG → t_old(M)

AUA: Ile → Met

~~t_old(I)~~

LOSS (Deletion) of t_old(I)

GAIN (Modification) of

t_old(M) → t_new(M)

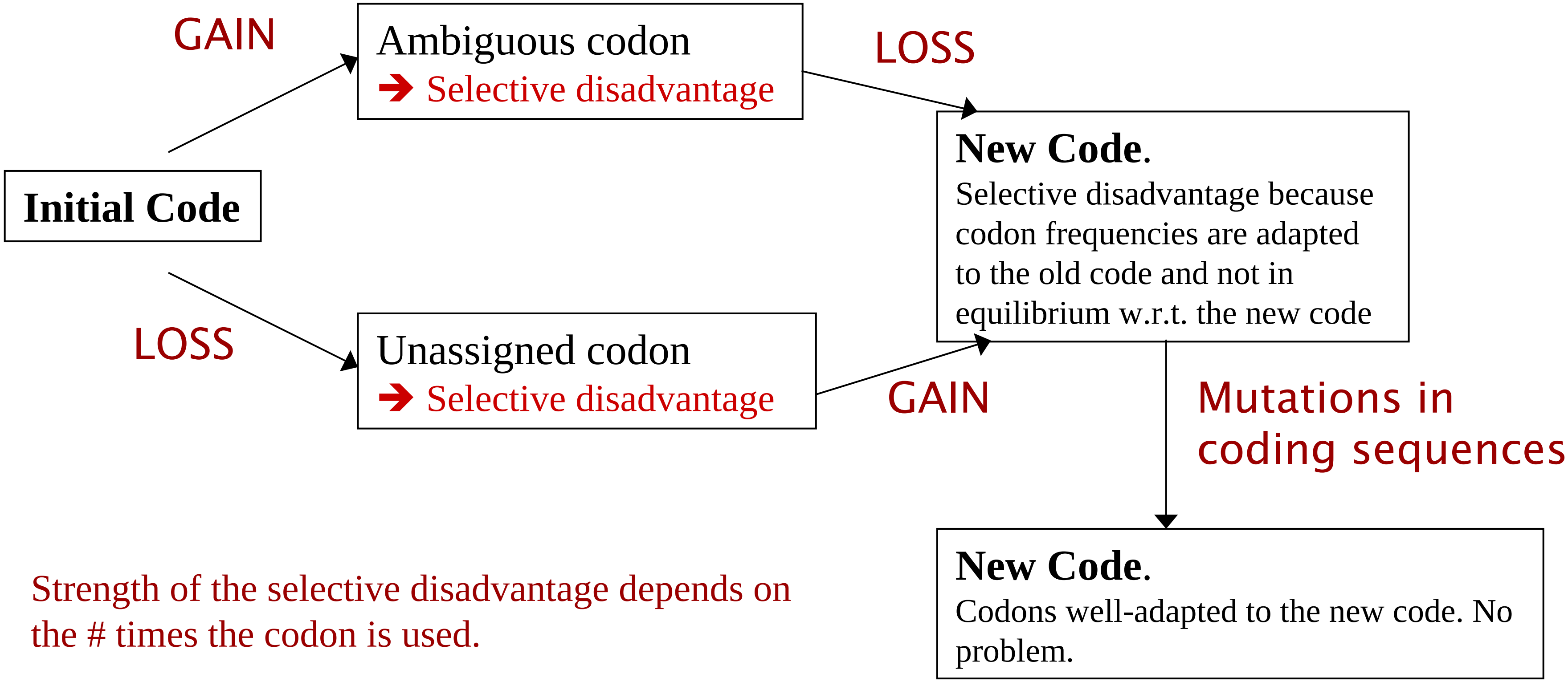
After Change

AUA, AUG → t_new(M)

Temporal Ordering of occurrence of GAIN/LOSS events determine mechanism of change

LOSS before GAIN ≠ GAIN before LOSS

Proposition: Codon reassignments involve a LOSS (i.e. deletion or loss of function of a tRNA or RF) and a GAIN (i.e. gain of a new tRNA or a gain of function of an existing one).



GAIN

Ambiguous codon
➔ Selective disadvantage

LOSS

Initial Code

LOSS

Unassigned codon
➔ Selective disadvantage

GAIN

New Code.
Selective disadvantage because codon frequencies are adapted to the old code and not in equilibrium w.r.t. the new code

Mutations in coding sequences

New Code.
Codons well-adapted to the new code. No problem.

Strength of the selective disadvantage depends on the # times the codon is used.

NO selective disadvantage if the codon disappears.

SS & P. Higgs, Genetics **170:**831-840(2005)

The Gain-Loss framework allows for **4** *distinct* mechanisms of codon reassignment.

Codon Disappearance (CD) - Osawa & Jukes

Ambiguous Intermediate (AI) - Schultz & Yarus

Unassigned Intermediate (UI)

Compensatory Change (CC) – Code change occurs *without fixation* of the AI or UI state

Objectives of Modeling

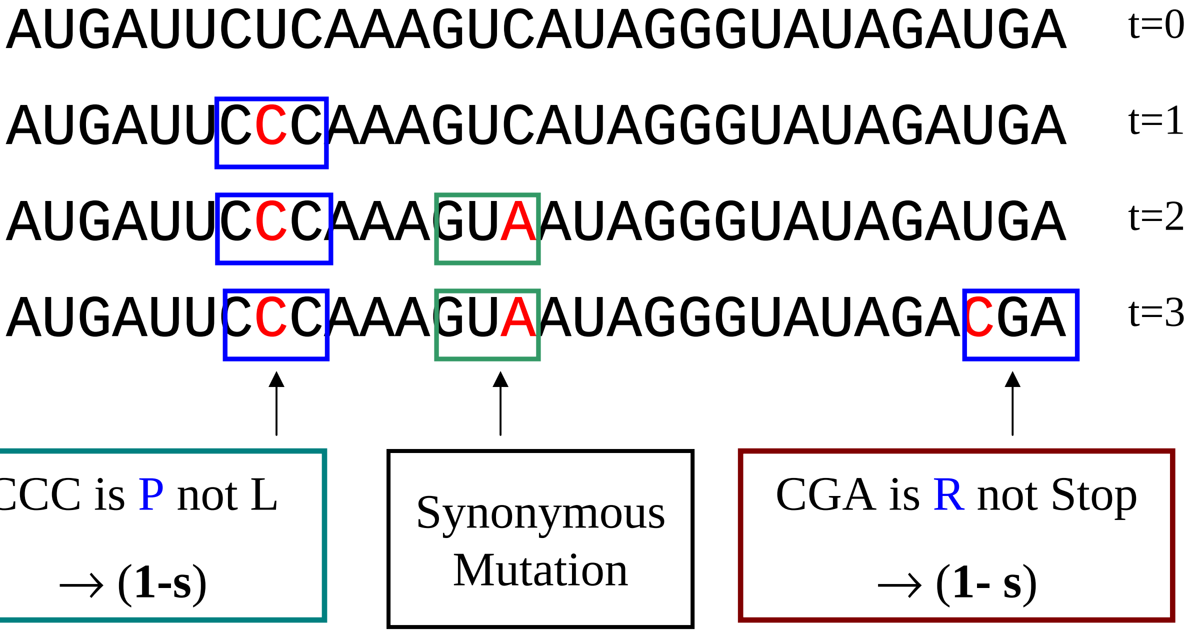
- Does the codon disappear from the genome during the intermediate stages?
- In the ***absence*** of codon disappearance, does the population pass through the ambiguous state or the unassigned state?
- Does the intermediate ambiguous or unassigned state get ***fixed*** in the population ?

Computational Model of a Population Undergoing Codon Reassignment

Part 1 – Coding sequence evolution

A population of ***N*** individuals.

Each individual has a genome that evolves with mutation rate ***u*** per site per generation.


AUGAUUCUCAAAGUCAUAGGUAUAGAUGA t=0

AUGAUUCCCAAAGUCAUAGGUAUAGAUGA t=1

AUGAUUCCCAAAGUAAUAGGUAUAGAUGA t=2

AUGAUUCCCAAAGUAAUAGGUAUAGACGA t=3

CCC is P not L

→ **(1-s)**

Synonymous Mutation

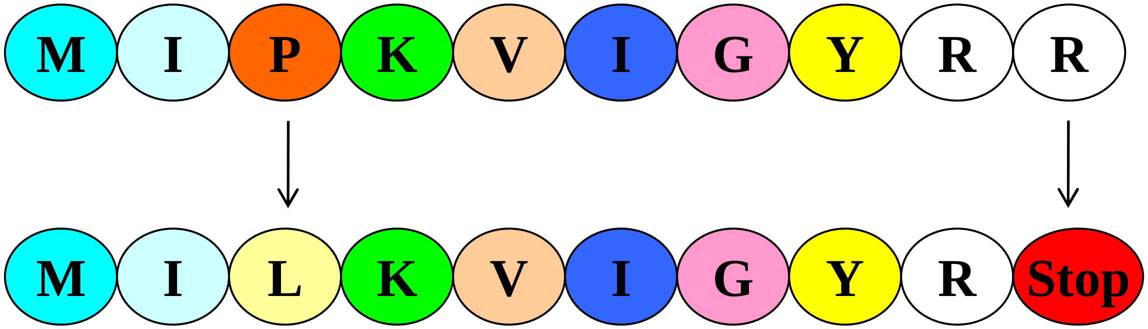
CGA is R not Stop

→ **(1- s)**

Base sequence at t=3

TRANSLATE

Evolved amino acid sequence



Optimum amino acid sequence

Calculating the fitness of an individual sequence in the population

- Ascertain the genetic code used by the individual sequence.
- Compare the translated mRNA sequence to the optimal protein sequence.
- Assign factor (1-**s**) whenever there is a ‘wrong’ amino acid.

Eg.: AUGAUUCCCAAAGUAAUAGGGUAUAGACGA

Fitness = (**1**-**s**)(**1**- **s**)

Individuals reproduce in proportion to their fitness

Computer Model of a Population Undergoing Codon Reassignment

Part 2 – tRNA genes controlling translation

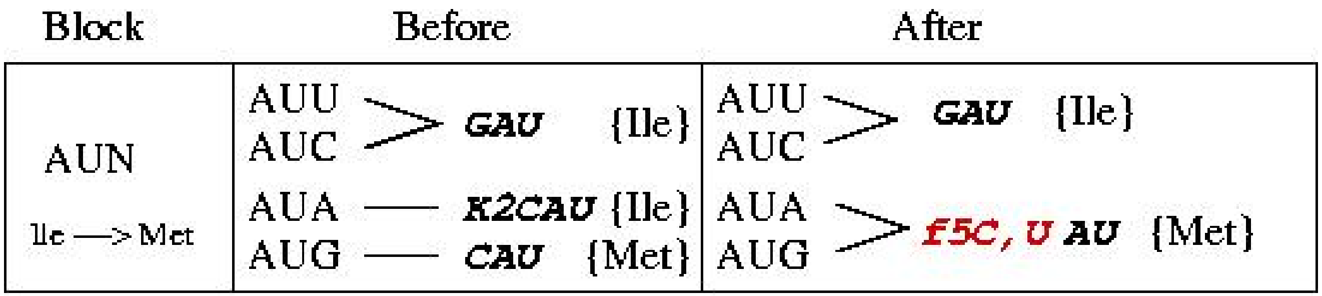
- Each individual has a LOSS gene and a GAIN gene with two alleles, 0 or 1. These genes represent tRNAs associated with one codon (e.g. AUA in the Ile → Met case).

LOSS = 0 - the original tRNA is still present.

LOSS = 1 - the original tRNA has been deleted or is non-functional.

GAIN = 0 - the new tRNA is not present.

GAIN = 1 - the new tRNA has been acquired.



Block	Before			After		
AUN Ile —> Met	AUU AUC	GAU	{Ile}	AUU AUC	GAU	{Ile}
	AUA	K2CAU	{Ile}	AUA AUG	f5C,U AU	{Met}
	AUG	CAU	{Met}			

There are 4 types of individuals defined by the GAIN-LOSS combination.

- Type 00 (old code) - fitness's calculated as in Part 1.
- Type 11 (new code) - translation follows the modified genetic code.
- Type 01 (unassigned codon) - Specify fitness factor $(1-\mathbf{s}_{\mathbf{unas}})$ wherever this codon appears.
- Type 10 (ambiguous codon) - This will be wrongly translated half the time. Specify fitness $\mathbf{(1-s)^{1/2}}$ when it appears at either the old or the new amino acid position.

Possible distribution of tRNA-types and individual sequences in a generation

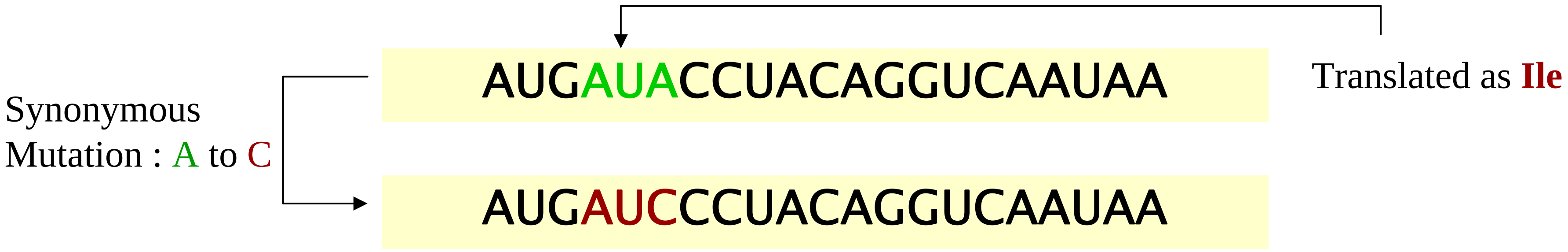
Sequence	Gain–Loss	Code Used	Fitness Assignment for AUA
Seq-1	Type 00	Old Code	1 (Ile); (1-s) - otherwise
Seq-2	Type 01	Old Code	(1-s_unas)
Seq-3	Type 01	Old Code	(1-s_unas)
Seq-4	Type 10	Old Code	$(1-s)^{1/2}$
Seq-5	Type 11	New Code	1 (Met); (1-s) - otherwise
Seq-6	Type 01	Old Code	(1-s_unas)

- Individuals begin in state 00. Alleles mutate from state 0 to 1 with a probability $\mathbf{U}_{\mathbf{gl}}$ per individual per generation.

- Simulations begin with all type 00 and end with all type 11. Type 01 and/or type 10 are present at intermediate stages.

Codon Disappearance (CD) [Osawa & Jukes]

- Codon undergoing reassignment vanishes during a transient intermediate stage.



AUGAUACCUACAGGUCAAUAA

Translated as **Ile**

Synonymous
Mutation : A to C

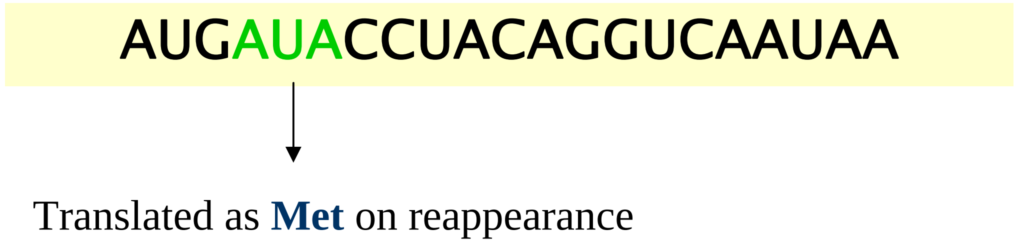
AUGAUCCCUACAGGUCAAUAA

- A new tRNA acquires the ability to decode the codon

  E.g: Gain of function of tRNA-Met(UAU or ^{f5}CAU)

  Loss (deletion) of tRNA-Ile (^{k2}CAU)

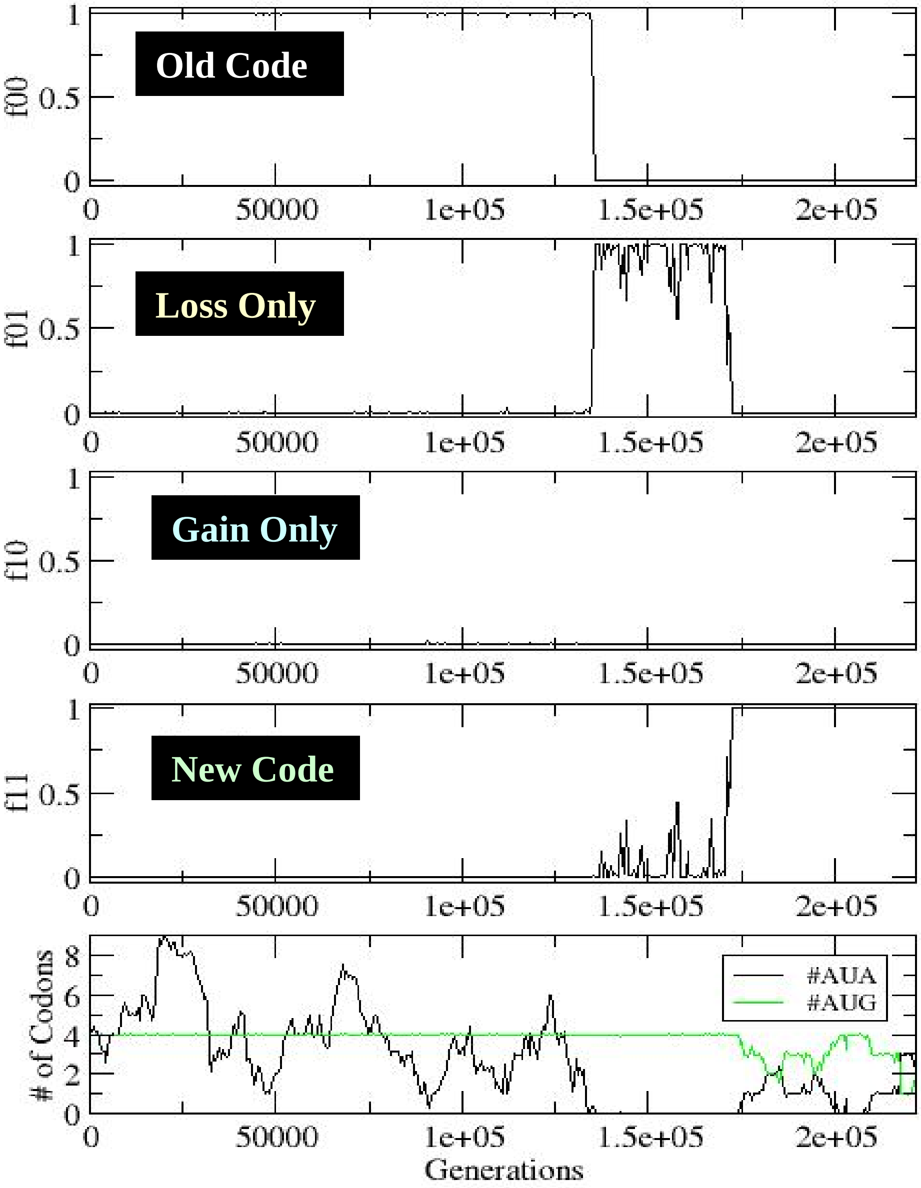
- On subsequent reappearance ,the AUA codon has a new meaning (amino acid) associated with it.



AUGAUACCUACAGGUCAAUAA

Translated as **Met** on reappearance

***S.S & P.Higgs, Genetics* 170:**831-840(2005)



Evidence of codon reassignment via the ***Codon Disappearance*** mechanism.

fij is the *frequency* of individuals with tRNA **type ij**.

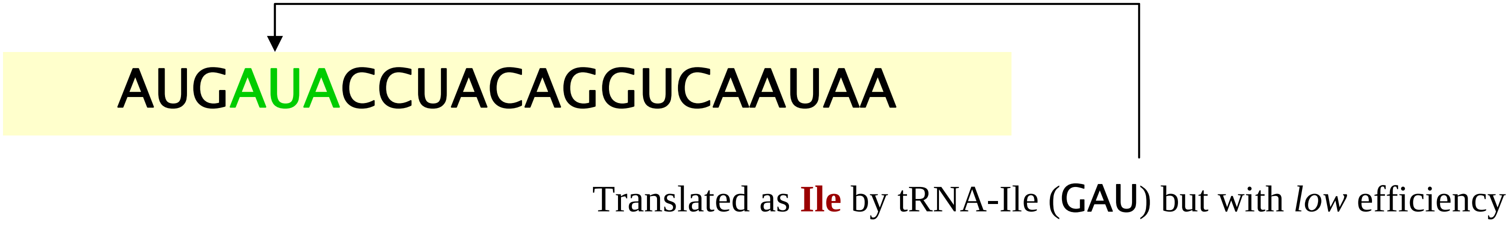
Parameters

N=1000; u=0.0001;
s=s_stop=s_unas=0.05; #AUA=4.
#AUG=4

Unassigned Intermediate Mechanism (UI)

- tRNA decoding a codon is lost (Loss of function), codon becomes unassigned during a transient intermediate stage.

  Eg.: Loss (deletion) of tRNA-Ile (^{k2}CAU) occurs first; AUA - unassigned



AUGAUACCUACAGGUCAAUAA

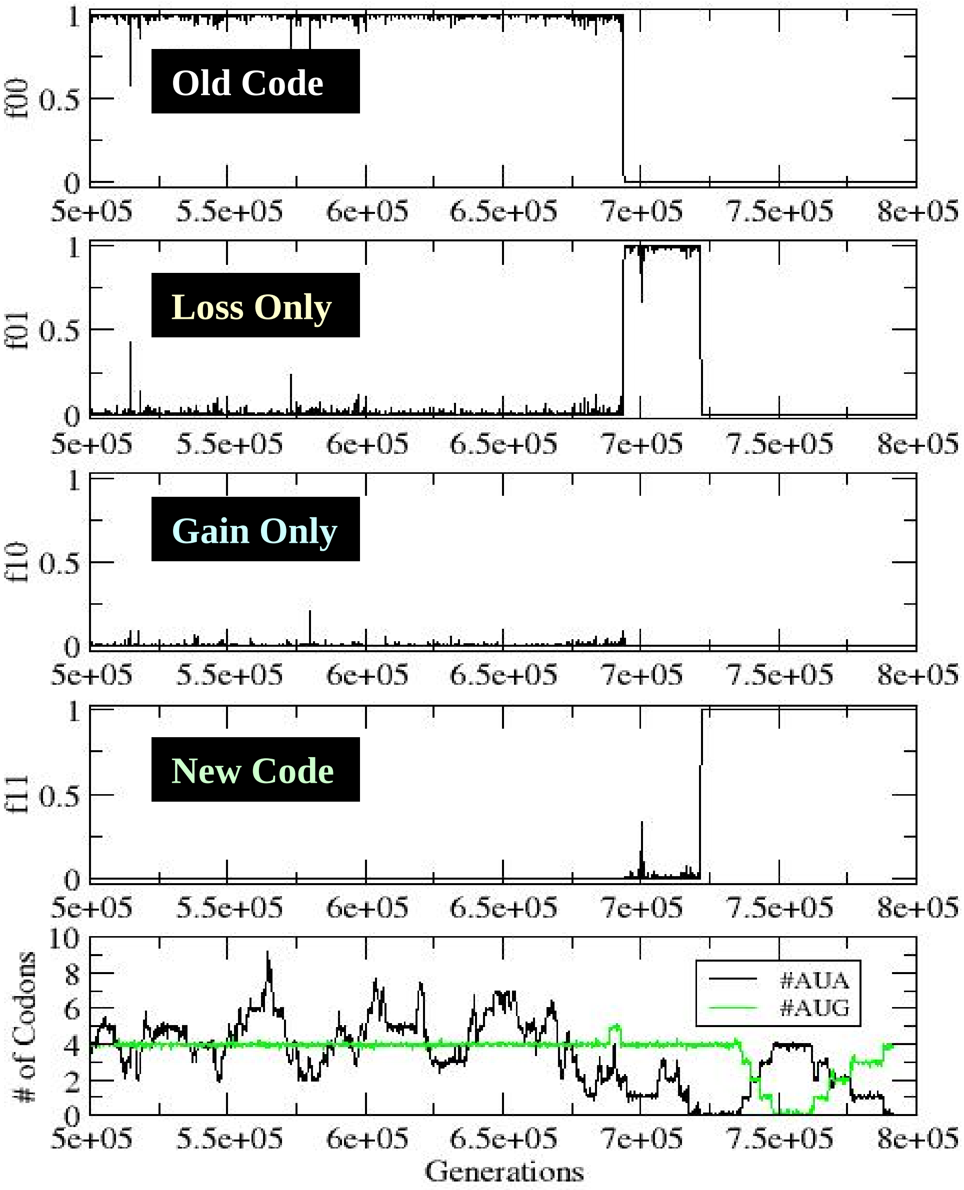
Translated as **Ile** by tRNA-Ile (GAU) but with *low* efficiency

- A new tRNA acquires the ability to efficiently decode the codon.

  Eg.: Gain of function of tRNA-Met(UAU or ^{f5}CAU)

- Codon gets new meaning ; new genetic code results.

  Eg.: AUA translated as **Met**



Evidence of Codon Reassignment initiated by the ***Loss*** of tRNA.

Parameters

N=1000; u=0.0001;
s_unas=0.007;
s=s_stop=0.02; #AUA=4.

Ambiguous Intermediate (AI) [Schultz & Yarus]

- New tRNA acquires the ability to decode the codon.

  Eg.: Gain of function of tRNA-Met(UAU or ^{f5}CAU)

- Codon undergoing reassignment is translated as two distinct amino acids during a transient intermediate stage.

AUGAUACCUAUAGGUCAAUAA

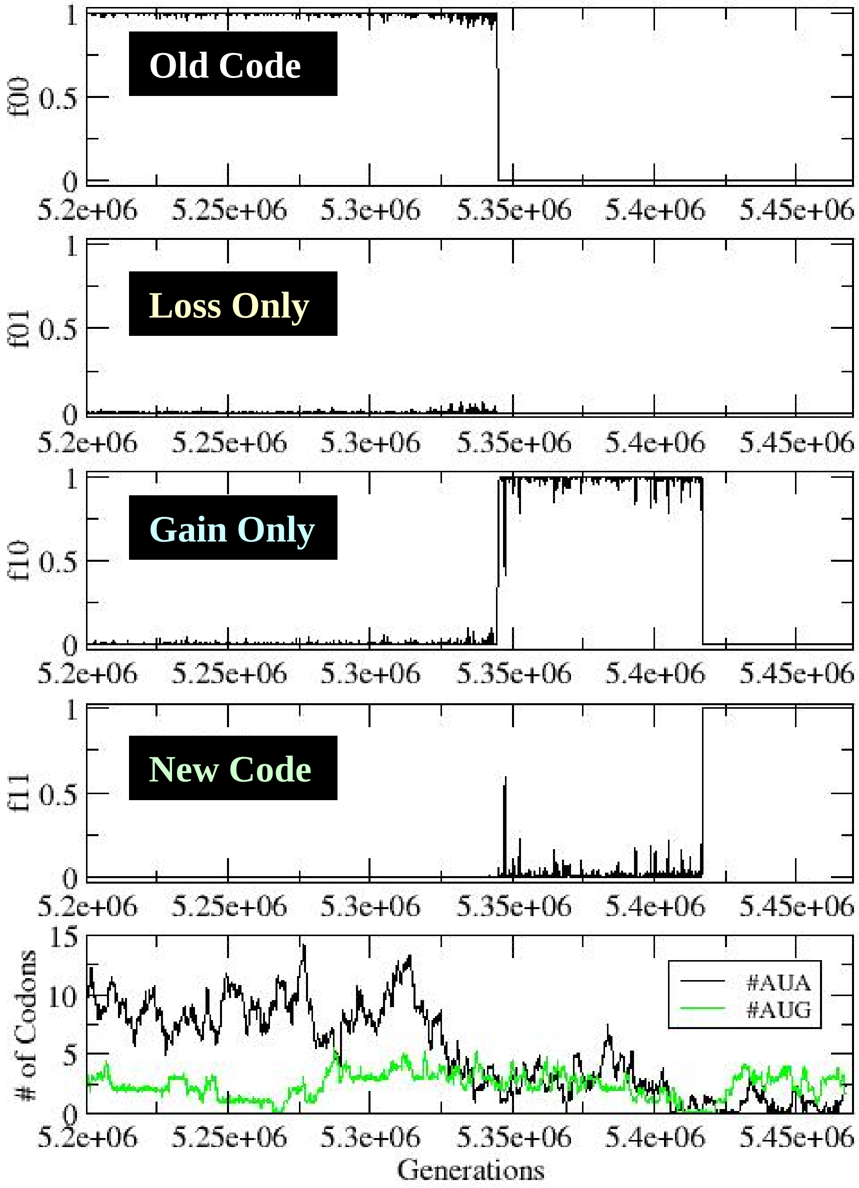
Translated as **Ile** by tRNA-**Ile**(^{k2}CAU)

Translated as **Met** by tRNA-**Met**(^{f5}CAU)

- New code is established when one meaning is selected over the other.

  Eg.: Loss (deletion) of tRNA-Ile (^{k2}CAU) ; **ambiguity removed**.

  AUA uniquely translated as **Met**.



Evidence of codon reassignment occurring via the **Ambiguous Intermediate** mechanism.

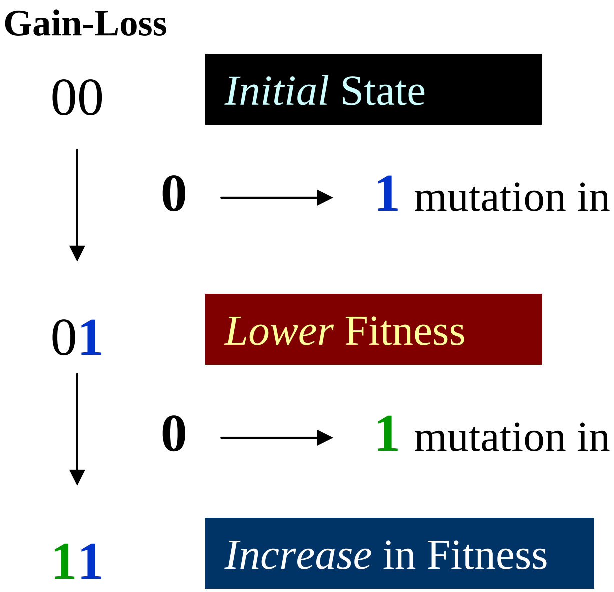
36 **Ile** codons (including 12 AUA's) and 4 **Met's** were present in the original sequence.

Parameters

N=1000; u=0.0001;
s_sense=s_stop=s_unas=0.01;#AUA=12.

Compensatory Change (Kimura)

Changes in the Gain-Loss gene used to represent the state of the tRNA



Gain-Loss

00 *Initial* State

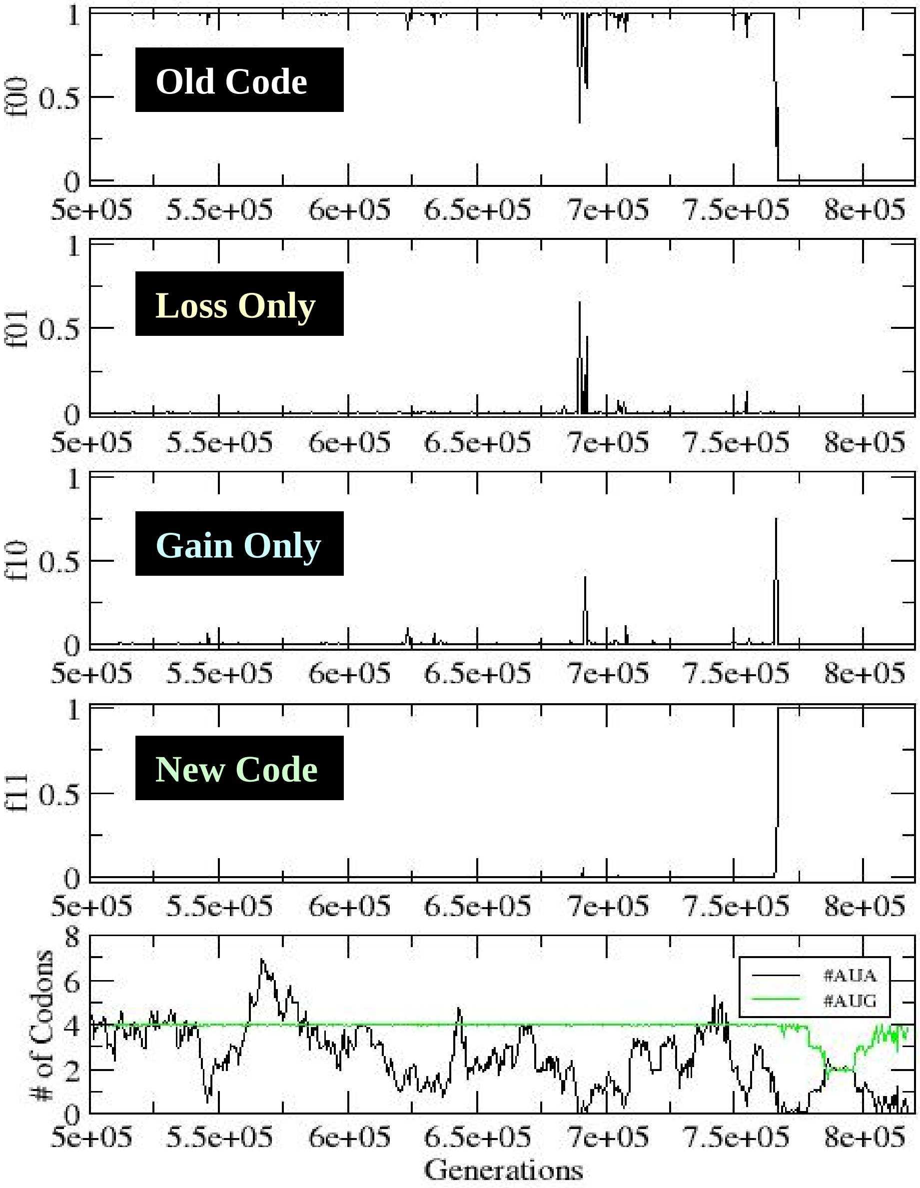
0 → **1** mutation in the Loss gene

01 *Lower* Fitness

0 → **1** mutation in the Gain gene

11 *Increase* in Fitness

The selective disadvantage of the intermediate stage is compensated for by a subsequent mutation.



Evidence of codon reassignment occurring via the **Compensatory Change** mechanism.

Parameters

N=1000; u=0.0001;
s_sense=s_stop=s_unas=0.05;
#AUA=4.

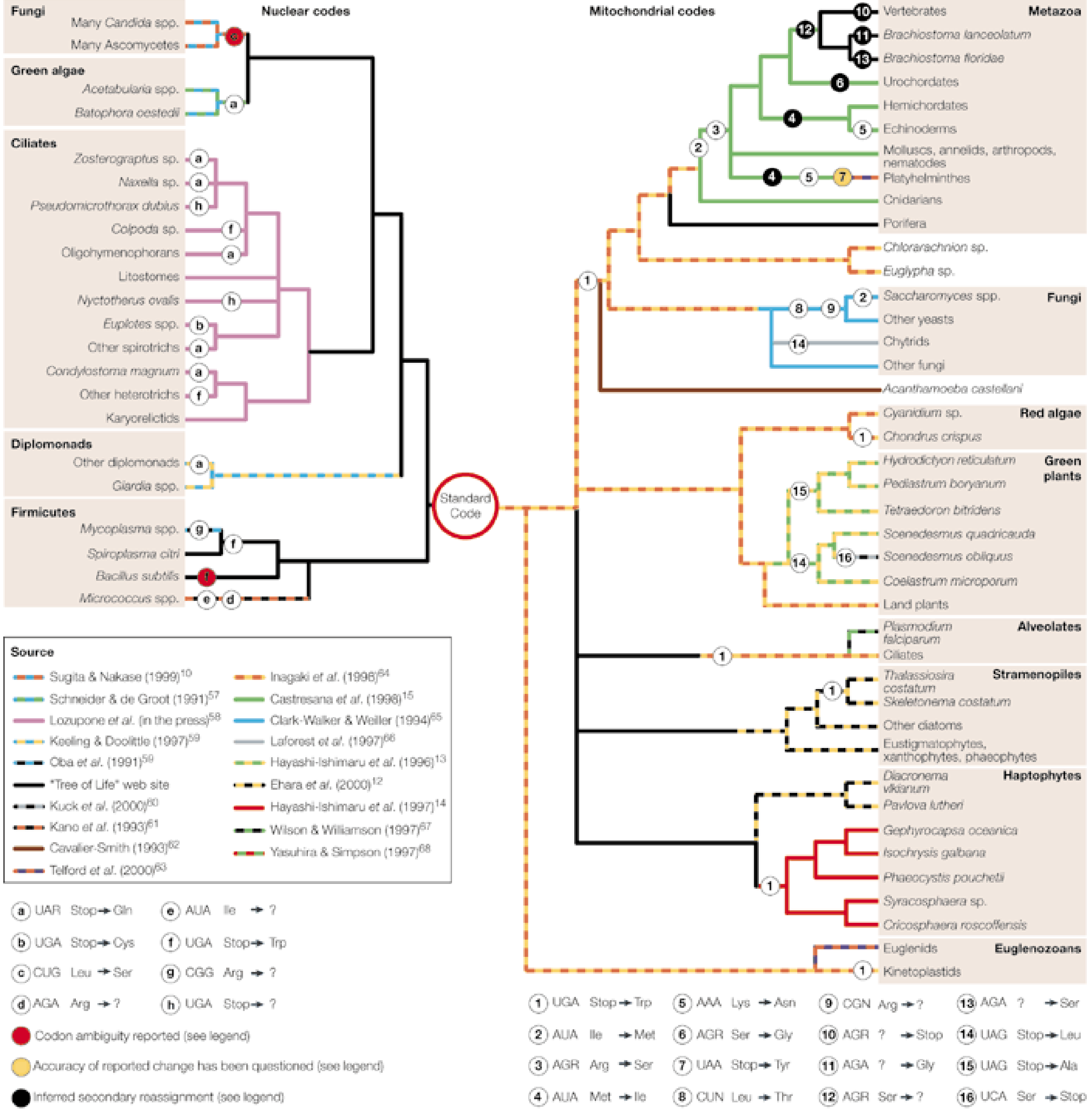
Model Predictions

Ambiguous Intermediate mechanism seems more likely for reassignment involving sense codons.

Codon Disappearance mechanism could be more relevant for reassignment involving stop codons.

Reassignment initiated by **tRNA Loss** is likely in scenarios in which a codon can be translated by a near-cognate tRNA without substantial fitness cost.

Compensatory Change mechanism is more likely when penalty against ambiguous & unassigned states are large.



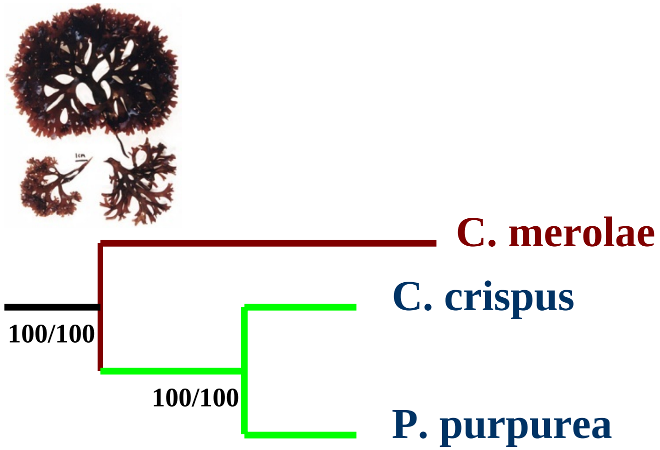
Nature Reviews | Genetics

Knight *et al.* (2001)
Nature Reviews Genetics

Genomic Analysis of Changes in the Mitochondrial Genetic Code

SS, X. Yang & P. Higgs (J.Mol.Evol. **64**:662-688 (2007))

- Determine location of each codon reassignment on a phylogenetic tree.
- Correlate location of codon reassignment with codon usage and tRNA types available for the specific codon(s) undergoing reassignment.



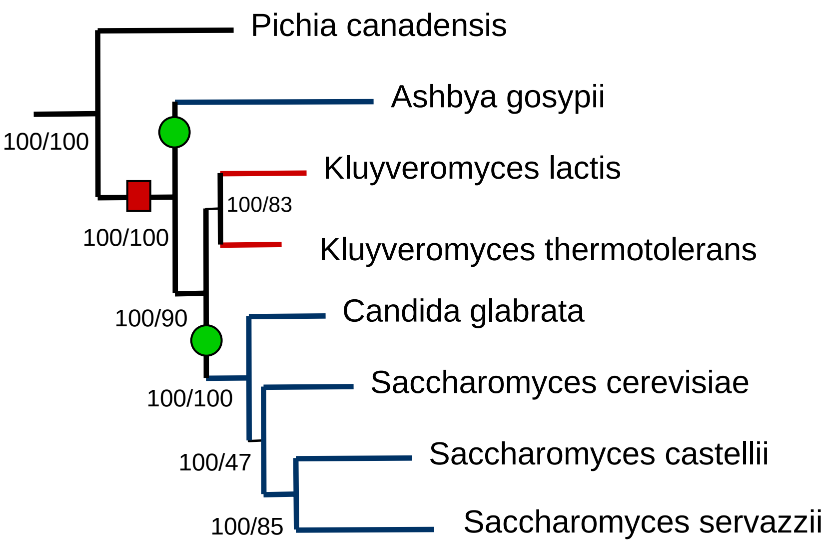
UGA: Stop to Trp

Species	**# UGA**	**tRNA-Trp** anticodon	$P_{dis} = (1-f_{UGA})^N$
C.merolae	2 (stop)	CCA	
C.crispus	101 (Trp)	UCA	0.036
P.purpurea	117 (Trp)	UCA	0.0014

- UGA disappeared in the common ancestor of *C.crispus* and *P.purpurea*.
- Anticodon of tRNA-Trp mutated from CCA to UCA

➡ UGA on reappearance is translated as Trp.

AUA: **Ile** to **Met** in Fungi



Species	**#AUA**	tRNA Ile(^{k2}CAU)	Gain of tRNA Met(CAU)
P.canadensis	562	Present	No
A.gosypii	95	Absent	Yes
K.lactis	7	Absent	No
K.thermotolerans	2	Absent	No
C.glabrata	16	Absent	Yes
S.cerevisiae	60	Absent	Yes
S.castellii	101	Absent	Yes
S.servazzii	95	Absent	Yes

🟥 **Location of disappearance of tRNA-Ile**(^{k2}CAU)

🟢 **AUA: Ile to Met**

Reassignment initiated by deletion of **tRNA-Ile**(^{k2}CAU)

In K.Lactis & K. Thermotolerans AUA translated as **Ile** by tRNA-**Ile**(**GAU**)

Completed after gain of function of **tRNA-Met**(CAU) → AUA translated as **Met**

Predictions of Codon Reassignment Mechanisms in Mitochondria

Codon Reassignment	**# of times**	**Mechanism**
UAG: Stop → Leu	2	CD
UAG: Stop → Ala	1	CD
UGA: Stop → Trp	12	CD
CUU/CUC/CUA/CUG: Leu → Thr	1	CD
CGU/CGC/CGA/CGC: Arg → Unassigned	5	CD
AUA: Ile → Met	5	UI
AAA: Lys → Asn or Unassigned	3	AI
AGA/AGG: Arg → Ser	1	UI
AGA/AGG: Ser → Stop	1	AI
AGA/AGG: Ser → Gly	1	AI
UUA: Leu → Stop	1	UI/AI
UCA: Ser → Stop	1	UI/AI

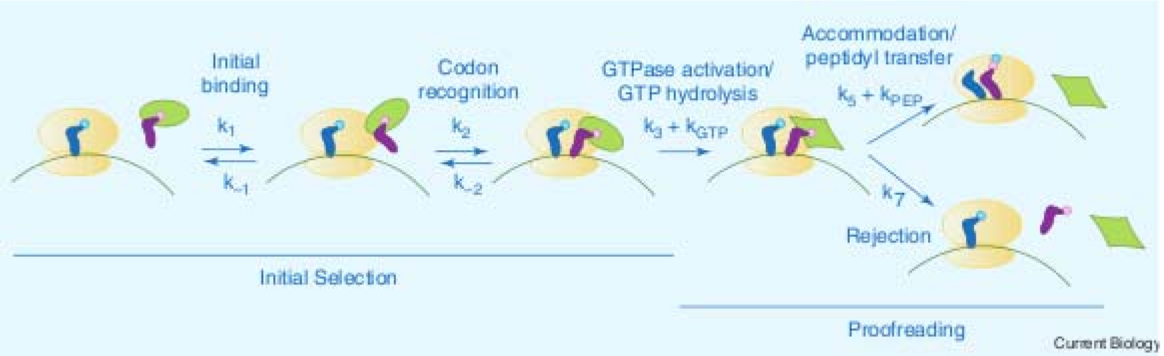
- AI and UI mechanisms can only be *indirectly* inferred after ruling out the CD mechanism

- For some reassignments, its *not* possible to determine whether the AI/UI mechanism was responsible.

Testing the viability of the AI & UI mechanisms *in vitro*

Translational Error correction *before* Peptide bond formation

(W. Wintermeyer, M. Rodnina, R. Green)



Source : Cochella & Green; Current Biology : 15 (2005); R536

Aminoacyl–tRNA's arrive at the ribosome in a ternary complex with GTPase EF–Tu and GTP.

- A cognate ternary complex hydrolyzes with greater rate ($k_3 > k_{-2}$) than a non-cognate ternary complex

A non-cognate ternary complex is more likely to dissociate ($k_{-2} > k_3$) than hydrolyze.

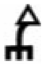
- A cognate ternary complex undergoes the peptidyl transfer reaction with greater rate ($k_5 > k_7$) than a non-cognate ternary complex.

- Conformational changes at the ribosome *and* in the cognate / non-cognate tRNA is induced by the pairing between the codon and anticodon and affects the peptidyl transfer reaction rate.

Such conformational changes increases the rejection rate ($k_7 > k_5$) of a non-cognate ternary complex from the ribosome.

Testing the UI mechanism : Effect on Translation Rates

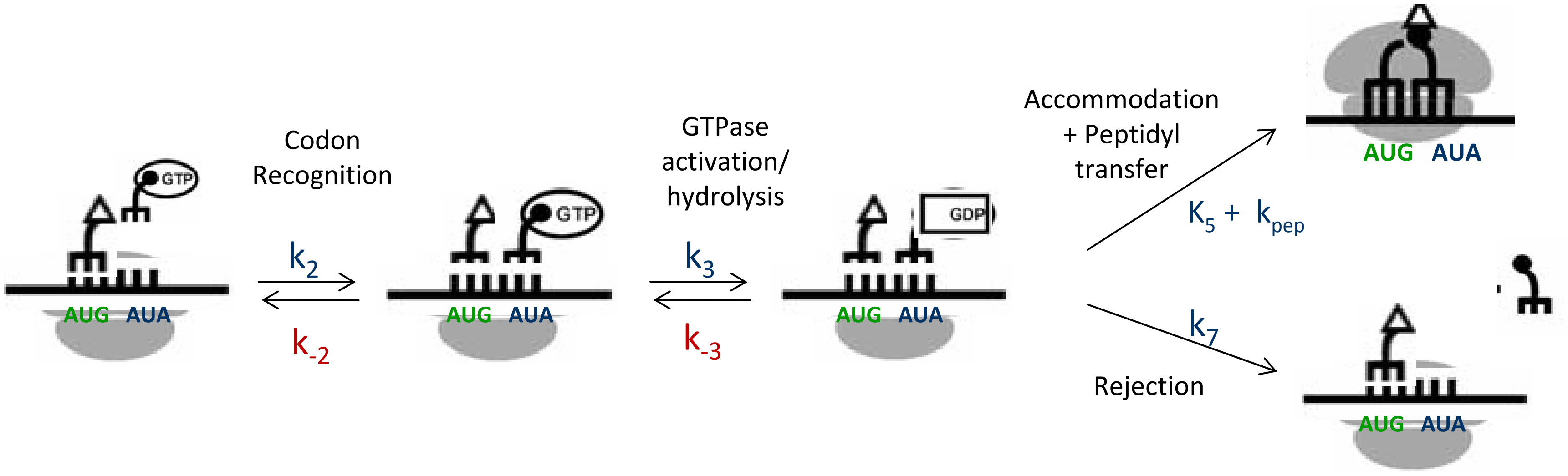
- Produce a ribosomal nascent chain (RNC) complex



tRNA-fMet

- Produce a ternary complex of EF-Tu-GTP-Ile-tRNA-Ile(GAU) for translating the AUA codon

- Determine the rates (k_{-2}, $k_{3,}$ $k_{7,}$ k_{pep}) of AUA codon recognition by the tRNA.



Codon Recognition

k_2

k_{-2}

GTPase activation/ hydrolysis

k_3

k_{-3}

Accommodation + Peptidyl transfer

$K_5 + k_{pep}$

k_7

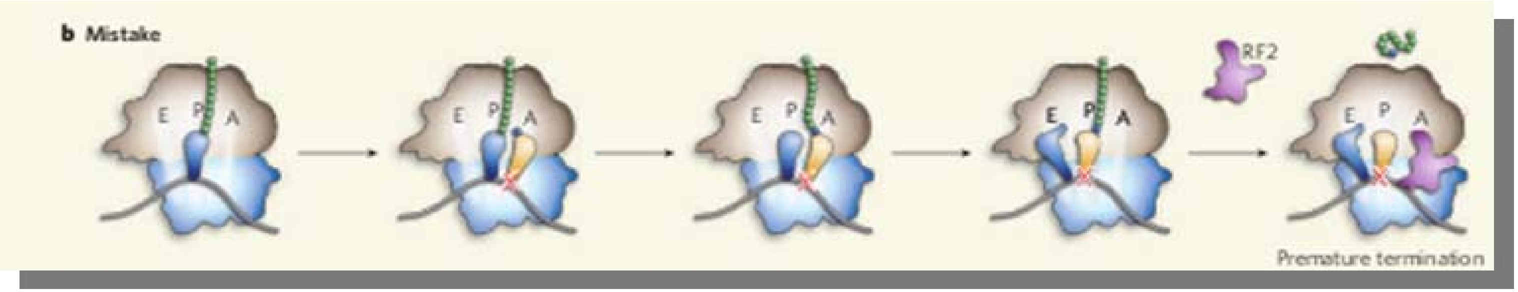
Rejection

Ile-tRNA-Ile(GAU)

Di-Peptidyl tRNA with amino acids Met-Ile when tRNA-Ile(GAU) is used to decode AUA

Translational Error Correction *after* Peptide bond formation

Zaher & Green; Nature 457 (2009) 161



Source : Fredrick & Ibba; Nature : 457 (2009); 157-158

The ribosomal complex becomes more susceptible to premature termination induced by RF2

➔ The partial formed erroneous sequence is prematurely released and eventually degraded.

During *ambiguous* translation of AAA by tRNA-Asn(GUU) a “wrong” amino acid Asn can be inserted into the peptide chain ?

How viable is *ambiguous* translation of AAA by tRNA-Asn(GUU) and tRNA-Lys(UUU) ?

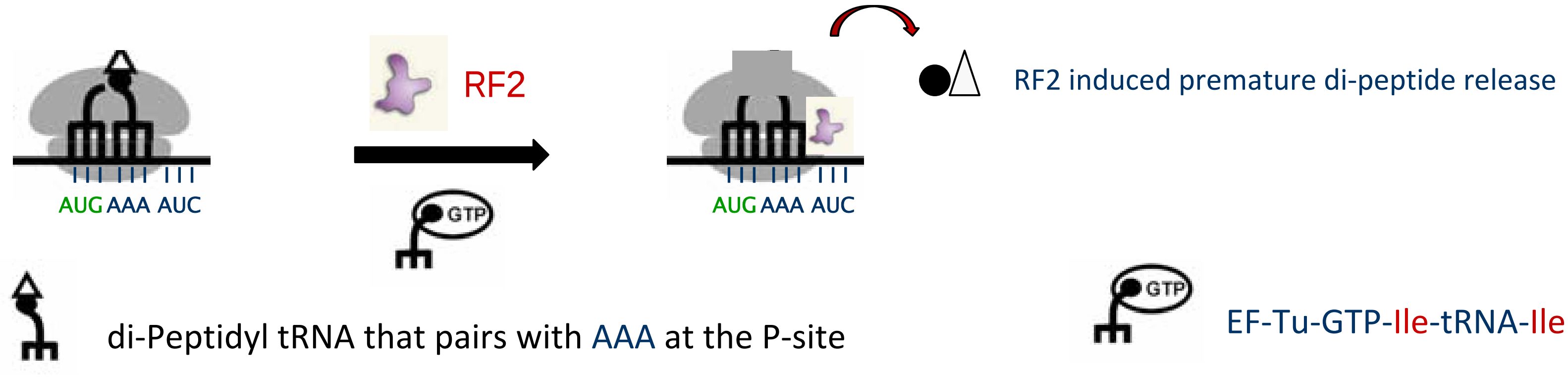
Retrospective Error Correction during Ambiguous Translation of AAA

- Create two di-peptidyl-tRNA-ribosome complexes both of which carry a sense codon (AUC) at the A-site.

- “Wrong” di-Peptidyl tRNA with amino acids Met-Asn when tRNA-Asn(GUU) is used to decode AAA

  Correct di-Peptidyl tRNA has amino acids Met-Lys when tRNA-Lys(UUU) is used to decode AAA

- Determine if recognition of AAA by tRNA-Asn(GUU) leads to RF2 induced peptide release.



RF2 induced premature di-peptide release

di-Peptidyl tRNA that pairs with AAA at the P-site

EF-Tu-GTP-Ile-tRNA-Ile(GAU)

- A comparison of the **differential** rates of translation by the tRNA-Asn(GUU) and tRNA-Lys(UUU) will determine the extent of ambiguous decoding AAA.

Conclusions

- The Genetic Code is *still* evolving

- The evolutionary pathways leading to codon reassignment can be deciphered

    Codon Disappearance mechanism is responsible for **all** reassignments involving **loss** of **stop** codons.

    Unassigned Intermediate and **Ambiguous Intermediate** mechanisms are mostly responsible for reassignments involving sense codons.

- *in vitro* experiments can be used to test the viability of AI & UI mechanisms.

- Open Question : tRNA import into mitochondria and its connection to codon reassignment

    - Understanding the tRNA import machinery and how it varies across organisms ?
    - When did the tRNA import machinery evolve and did it evolve independently multiple times ?
    - Why are some tRNA’s imported and not others ?

Acknowledgements

Paul Higgs – Genetic Code evolution model, genomic analysis of codon reassignments.

Xiaoguang Yang – Genomic analysis of codon reassignments.

Ashutosh Vishwabandhu – codon reassignment in bacterial and nuclear genomes

Howsun Jow
Cendrine Hudelot
Vivek Gowrishankar
Magnus Rataray

PHASE – PHylogenetics And Sequence Evolution package

Daniel Jameson
Bin Tang
Wei Xu

OGRe. - Database for mitochondrial genomes
drake.physics.mcmaster.ca/ogre