Evolution and multilevel optimization of the genetic code

T. Bollenbach, K. Vetsigian, R. Kishony, Genome Research March 2007

or

What information is hidden in the DNA?

Marion Reuter
October 30th, 2007
Journal Club

Content

- History of DNA: Watson, Crick & Co
- The Code and it´s decryption
- Evolution of the genetic code
- Optimization: frameshift and double coding
- A current work: Itzkovitz, Alon
- Conclusions

When all began

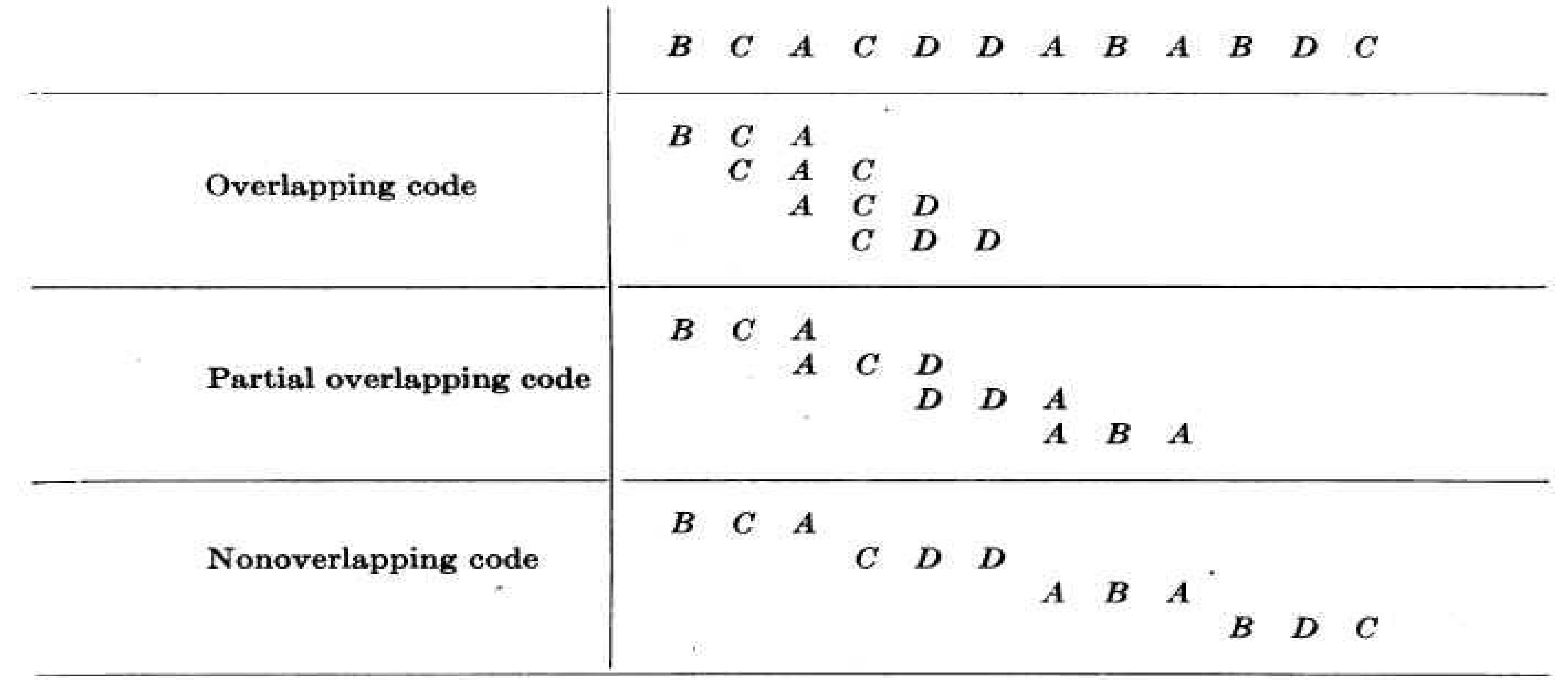
- **1868** first nucleid acids found (name because found in nucleus; F. Miescher)
- 1919 compounds found (sugar, base and phosphate; P. Levene)
- 1937 hints for repetitive structure of DNA (W. Astbury)
- 1952 proposal that order of nucleotides determines the order of amino acids (L.A. Dounce)
- **1953** : discovery of the DNA structure of the double helix (Watson & Crick)

The encryption race started

- 1954 Gamow suggested a „key to lock“ mechanism for binding the aa at special „holes“
- *quartet of nucleotides of nucleotides code for each aa but two are complementary,* so a triplet codes for each aa
  - 20 different aa
- *codons are overlapping: 2nd position of the first codon is the 1st position of the second codon,* ruled out in 1957 (Brenner)

Theories at 1957

(Crick proposed *„code without commas“*)



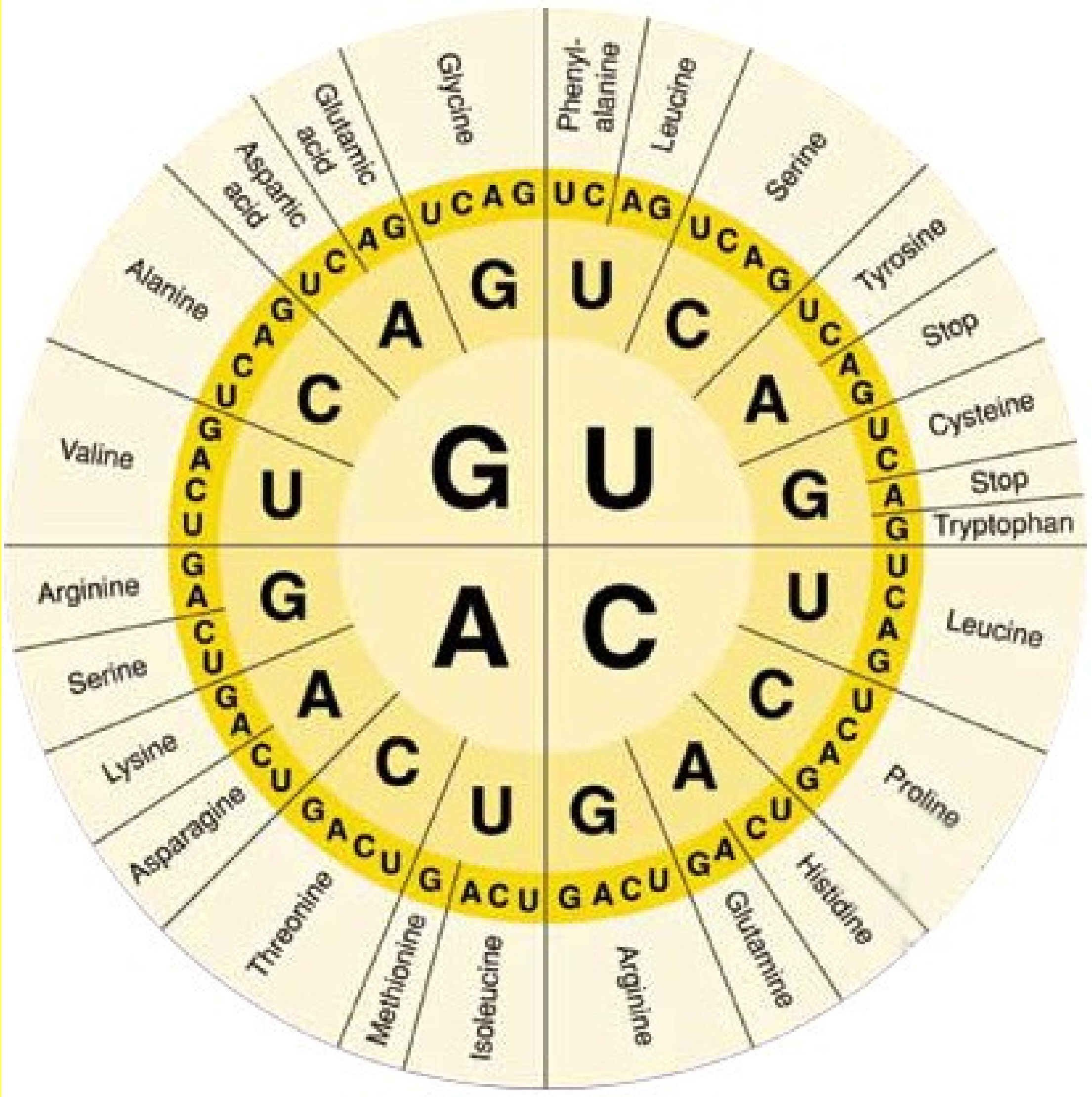
	B	*C*	*A*	*C*	*D*	*D*	*A*	*B*	*A*	*B*	*D*	*C*
Overlapping code	*B*	*C*	*A*									
		C	*A*	*C*								
			A	*C*	*D*							
				C	*D*	*D*						
Partial overlapping code	*B*	*C*	*A*									
			A	*C*	*D*							
					D	*D*	*A*					
							A	*B*	*A*			
Nonoverlapping code	*B*	*C*	*A*									
				C	*D*	*D*						
							A	*B*	*A*			
										B	*D*	*C*

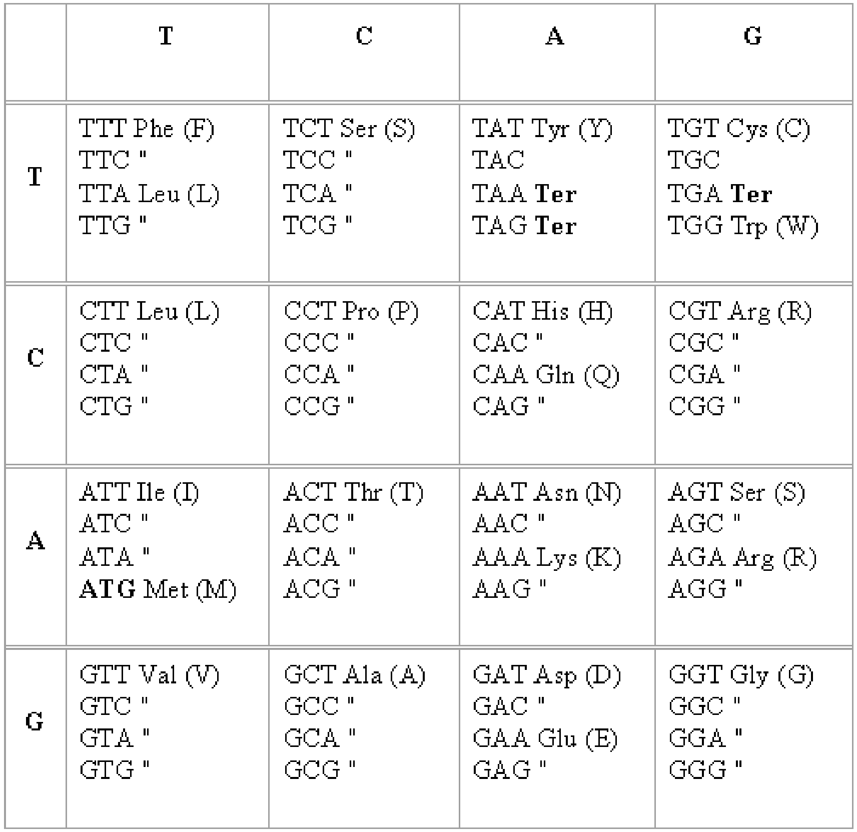
FIG. 1.—The letters *A*, *B*, *C*, and *D* stand for the four bases of the four common nucleotides. The top row of letters represents an imaginary sequence of them. In the codes illustrated here each set of three letters represents an amino acid. The diagram shows how the first four amino acids of a sequence are coded in the three classes of codes.

Lots of guessing until...

- *comma free*? which reading frame to choose? one is ok, the out-of-frame sequences are nonsense?
- several theories like sixtuplet, two letter or the one letter code + additional informtion from outside the DNA
- 1961 Matthaei found the first Codon (UUU to phenylalanine)
- 1966 all 64 codons were found

- family boxes and wobble rules (Crick) for system of same aa to several similar codons





	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

- similar codons assigned to aa´s with similar chemical properties (Woese)

Evolutionary optimization

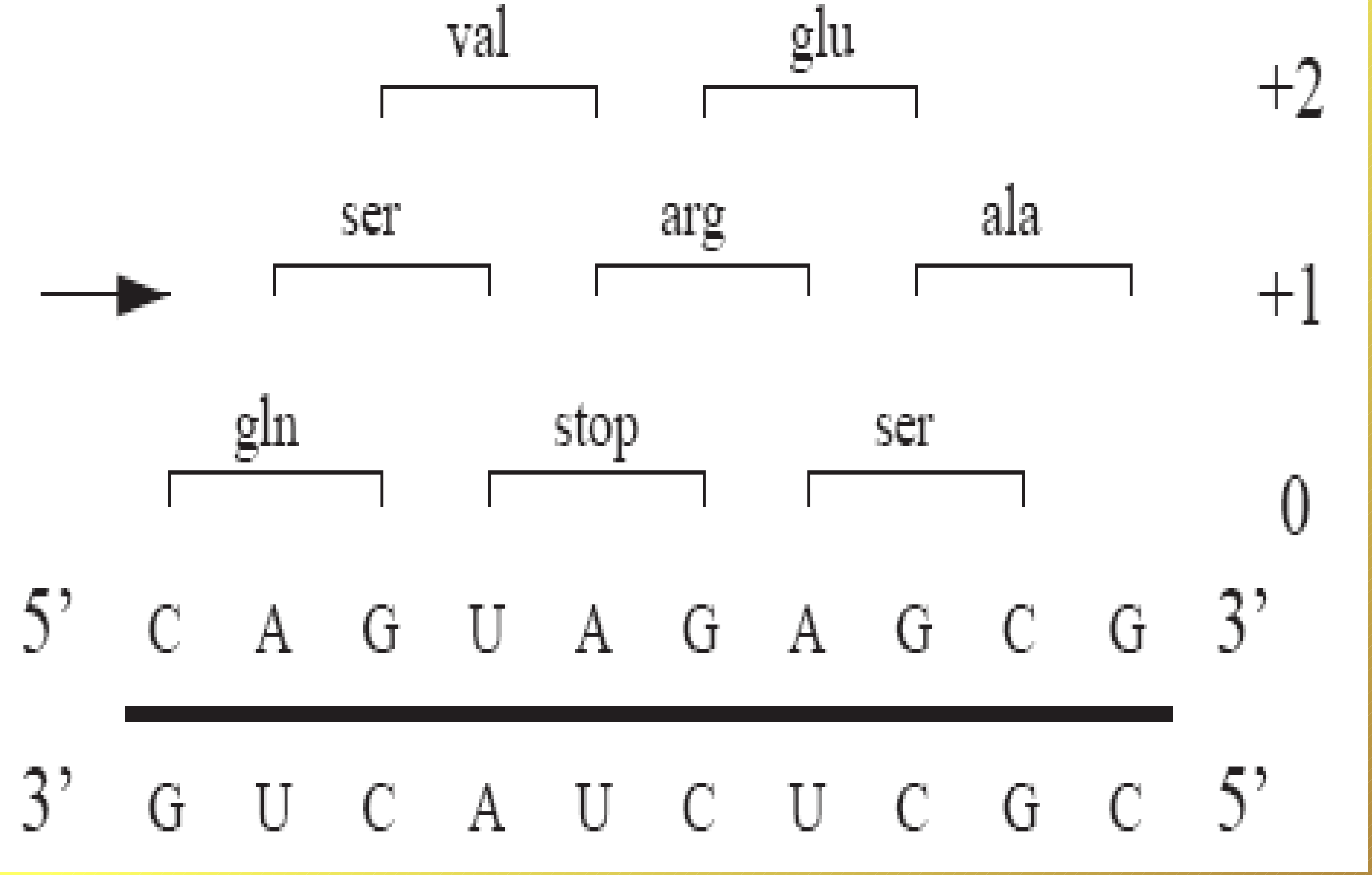
- Minimize the impact of translational errors
- Minimize the impact of mutations
- To deal with the increasing number of aa during evolutionary time
- etc.

variant codes, but all with minor differences (why? proposal of extensive horizontal gene transfer during early evolution, would lead to optimality and universality)

Frameshift mutation

- leads to <u>nonfunctional proteins, waste of ressources and could be toxic</u>
- minimizing with termination of elongation as quickly as possible after frameshift
- in many organisms bioinformaticians found tendencies to stop codons if read off-frame
- not always because then every point mutation would result to nonsense codons
- **Antagonistic goals: low price with low error rate**

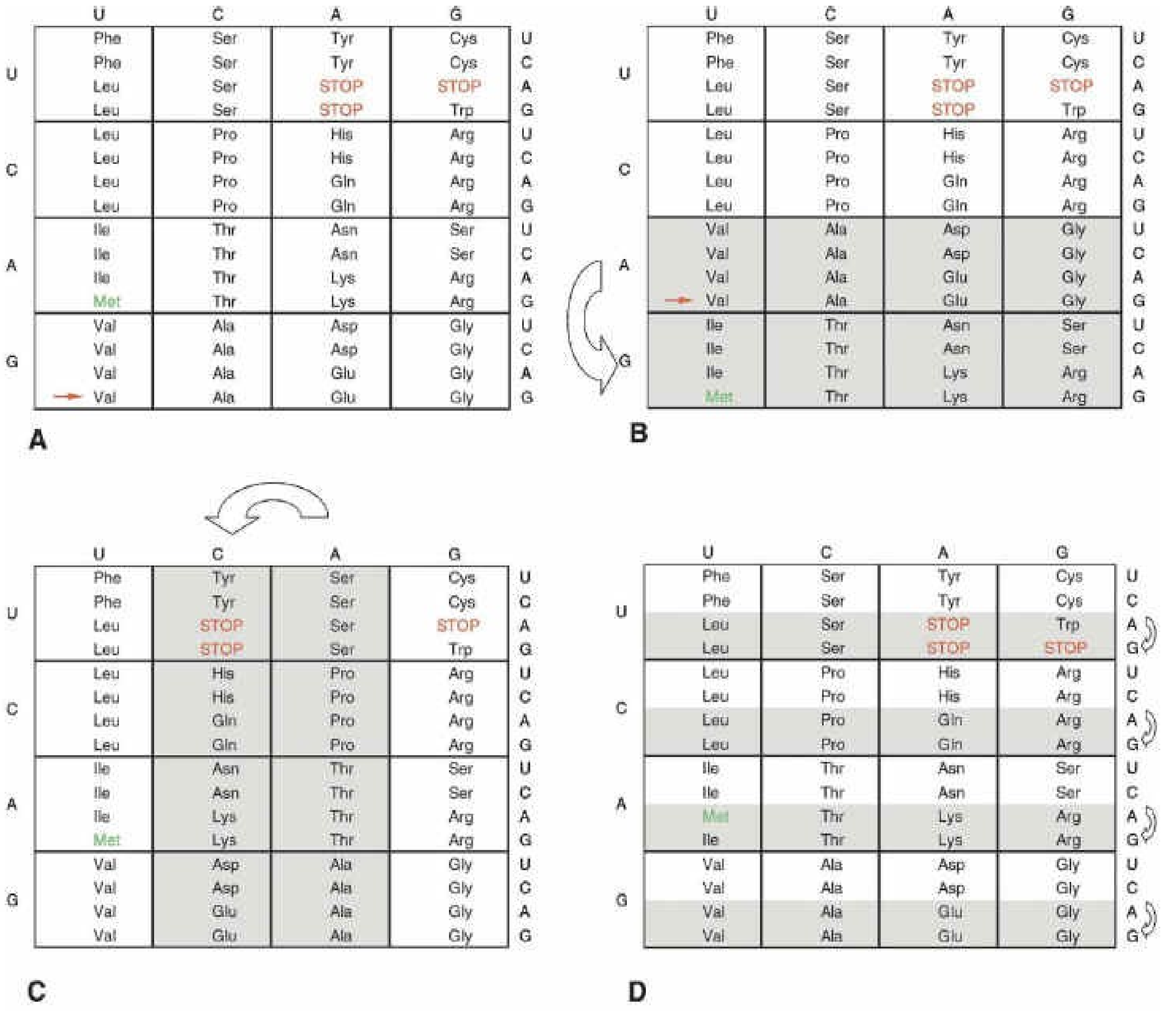
Frame shift


val
glu
+2
ser
arg
ala
+1
gln
stop
ser
0
5' C A G U A G A G C G 3'
3' G U C A U C U C G C 5'

Itzkovitz and Alon (Genome Research March2007):

The genetic code is nearly optimal for allowing additional information within protein-coding sequences

- First new property of optimization
  - compared genetic code with others that are equally optimized (with respect to mistranslation or mutation)
  - assume the usage frequency of the aa is fixed while codon assignments vary for other models
  - **Result: actual genetic code ist far better in minimizing the aa chain length after frameshift error**



	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	→ Val	Ala	Glu	Gly	G

A

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	→ Val	Ala	Glu	Gly	G
G	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G

B

	U	C	A	G	
U	Phe	Tyr	Ser	Cys	U
	Phe	Tyr	Ser	Cys	C
	Leu	STOP	Ser	STOP	A
	Leu	STOP	Ser	Trp	G
C	Leu	His	Pro	Arg	U
	Leu	His	Pro	Arg	C
	Leu	Gln	Pro	Arg	A
	Leu	Gln	Pro	Arg	G
A	Ile	Asn	Thr	Ser	U
	Ile	Asn	Thr	Ser	C
	Ile	Lys	Thr	Arg	A
	Met	Lys	Thr	Arg	G
G	Val	Asp	Ala	Gly	U
	Val	Asp	Ala	Gly	C
	Val	Glu	Ala	Gly	A
	Val	Glu	Ala	Gly	G

C

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	Trp	A
	Leu	Ser	STOP	STOP	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Met	Thr	Lys	Arg	A
	Ile	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

D

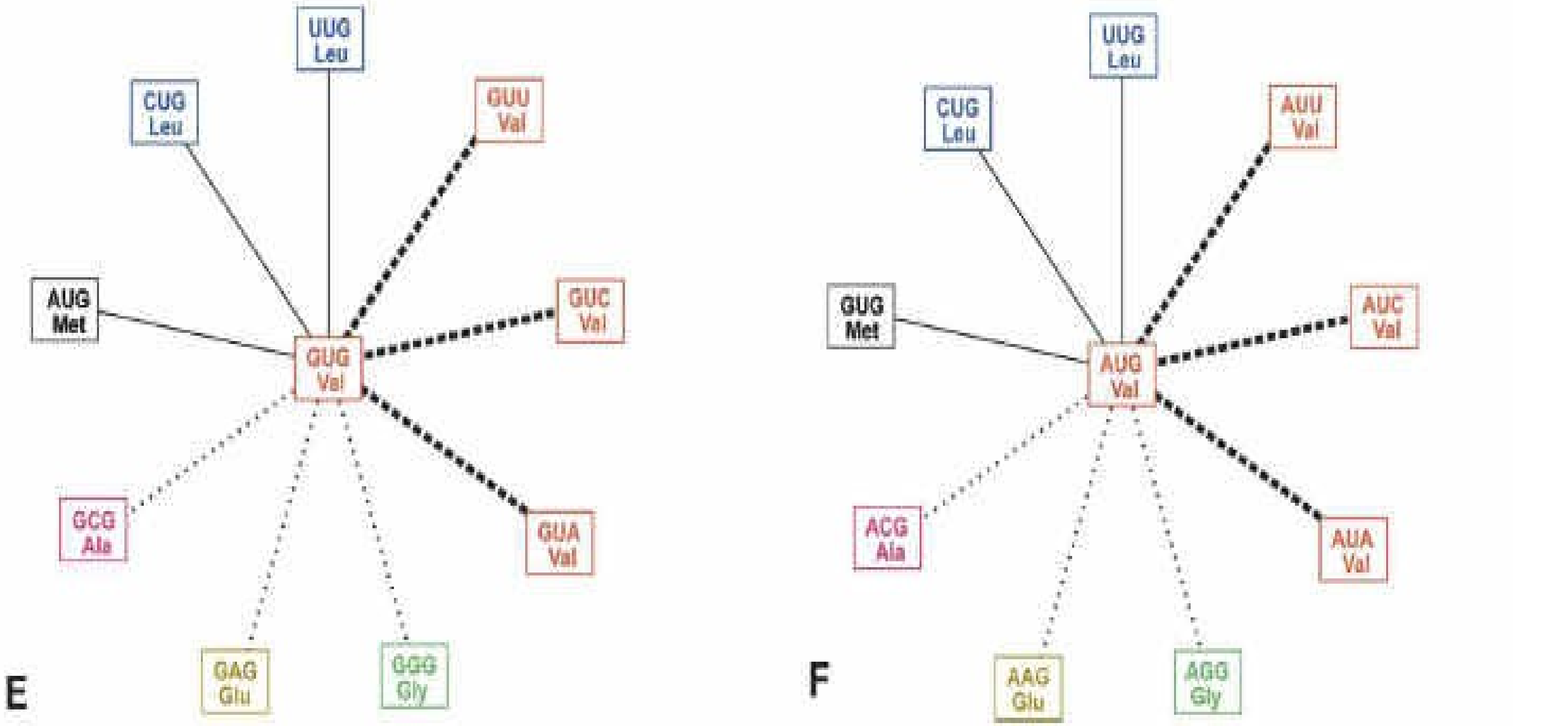


Figure 1. Alternative genetic codes. (*A*) The real code. (*B*) An alternative code obtained by an A↔G permutation in the first position. (*C*) An alternative code obtained by an A↔C permutation in the second position, and (*D*) A↔G permutation in the third position. Stop codons are marked in red, start (Met) codons in green. Codons that are changed relative to the real code are in gray. There are 4! × 4! × 2 = 1152 alternative codes obtained by independent permutations of the nucleotides in each of the three codon positions. (*E,F*) Structural equivalence of real and alternative genetic codes. For example, (*E*) the nine neighboring codons of the Valine codon marked with a red arrow in the real code (shown in *A*) are the same as (*F*) the nine neighboring codons of the Valine codon marked with a red arrow in the alternative code shown in *B*. Solid lines connect codons differing in the first letter, dotted lines connect codons differing in the second letter, and dashed lines connect codons differing in the third letter. Different amino acids are displayed in different colors. This equivalence applies to all codons.

Itzkovitz and Alon (2)

- Second new property/proposal
  - genetic code is highly optimal for encoding arbitrary additonal information, i.e., information other than aa code
  - like RNA splicing signals
  - signals recognized by the translation apparatus (e.g. usally stop codons can in special combinations be translated as rare aa)
  - nucleosome positioning
  - RNA secondary structure
  - additional genes (common in viruses; double coding)

Additional information as hidden messages

- example from „Sherlock Holmes“ (Conan Doyle 1893)

„The supply of game for London is going steadily up. Head keeper Hudson, we believe, has been now told to receive all orders for fly-paper and for preservation of your hen pheasant´s life.“

There is more information in the sentence than it seams. Read only every third word...

Additional information as hidden messages

- example from „Sherlock Holmes“ (Conan Doyle 1893)

„The supply of game for London is going steadily up. Headkeeper Hudson, we believe, has been now told to receive all orders for fly-paper and for preservation of your hen pheasant´s life.“

Hidden message/additional information: „The game is up. Hudson has told all. Fly for your life.“

Conclusions

- The degeneracy of the genetic code optimizes a combination of several different functions simultaneously. Low cost with high quality.
- Looking deeper into the structure of the genetic code, the more possibilities seems to occur.

Thanks for your attention! Questions?