Edward N. Trifonov
University of Haifa
and Masaryk University, Brno

Early Molecular Evolution

Edward N. Trifonov

(kakhol ve lavan)
(blue and white)

Contents:

Introduction

Molecular evolution is commonly known as the discipline initiated by seminal study of E. Zuckerkandl and L. Pauling on evolutionary distances between similar protein sequences. It deals with events of last 2-3 billion years, when the Life already operated with long sequences.

Zuckerkandl, E., and Pauling, L. (**1962**) Molecular disease, evolution and genetic heterogeneity. In: Kasha, M., and Pullman, B., (eds.) **Horizons in Biochemistry**. Academic Press, New York, pp. 189-225.

Early Molecular Evolution is a new discipline. It is reconstruction of the earliest molecular events and structures, starting with origin of the triplet code and continuing to the very first small nucleic acids and short protein chains. The first steps of the reconstruction have been made by W. Loeb, S. Miller, M. Eigen and P. Schuster.

Löb W (1913) Über das Verhalten des Formamids unter der Wirkung der stillen Entladung: Ein Betrag zur Frage der Stickstoff-Assimilation. Ber 46:684-697

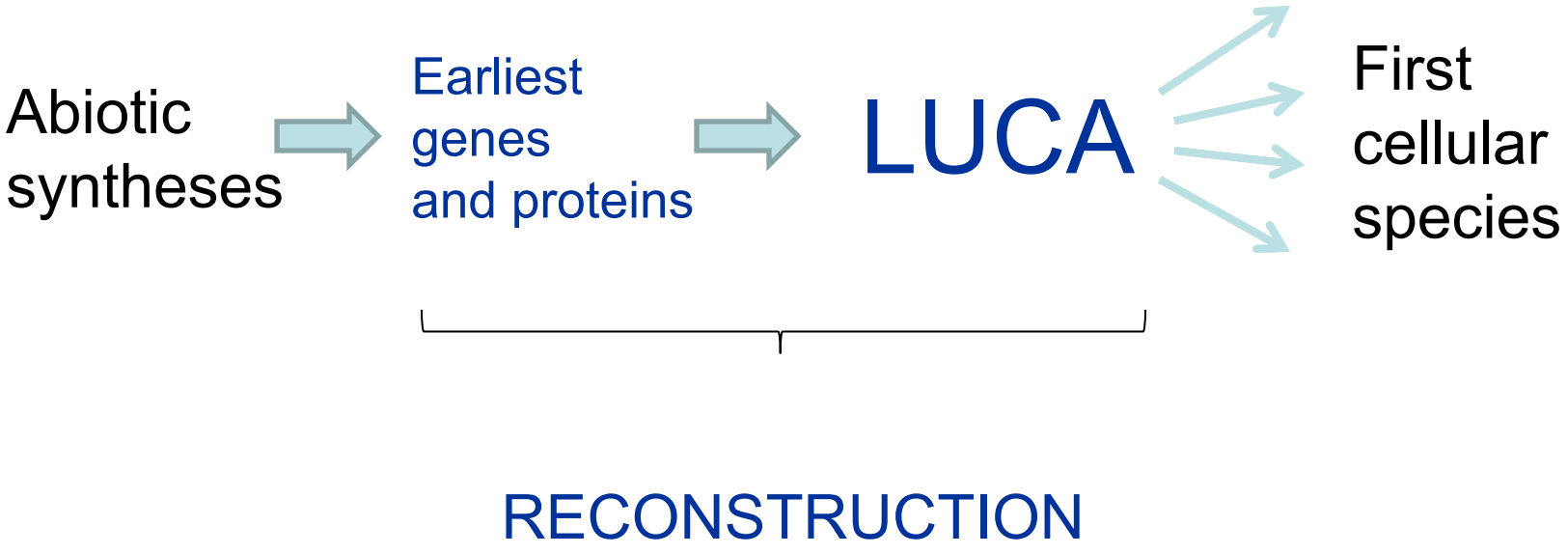
Yockey, H.P., 1997. Walther Löb, Stanley L. Miller and prebiotic "building blocks" in the silent electrical discharge. Persp. Biol. Med. 41, 125-131.

Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117:528-529

Miller SL, Urey HC, 1959, Organic compound synthesis on the primitive Earth, Science 130, 245-251

Miller SL (1987) Which organic compounds could have occurred on the prebiotic Earth? Cold Spr Harb Symp Quant Biol 52:17-27

Eigen M, Schuster P (1978) The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65:341-369


Abiotic syntheses
Earliest genes and proteins
LUCA
First cellular species
RECONSTRUCTION

Life on Earth, landmarks

NOW- Homo sapiens

Homo erectus

1-

earliest eukaryotic fossils

2-

3-

earliest prokaryotic fossils

oldest rocks

4-

origin of Earth

5-

billion
years
back

adapted from L. Margulis,
K. V. Schwartz. Five kingdoms

“millions of years, in pain, labors and fight
this shining beauty has been created
from primordial slime, and here it is:
just a rooster walking on the grass.
And it occurs to nobody what a Life cost
has been paid…

…in a thousand year long blink,
in a tremendous effort
dead particles fused together -
and the Life, selfconfident,
joyfully runs across the road,
disregarding those incredible sufferings
that have been sacrificed to its fate”.

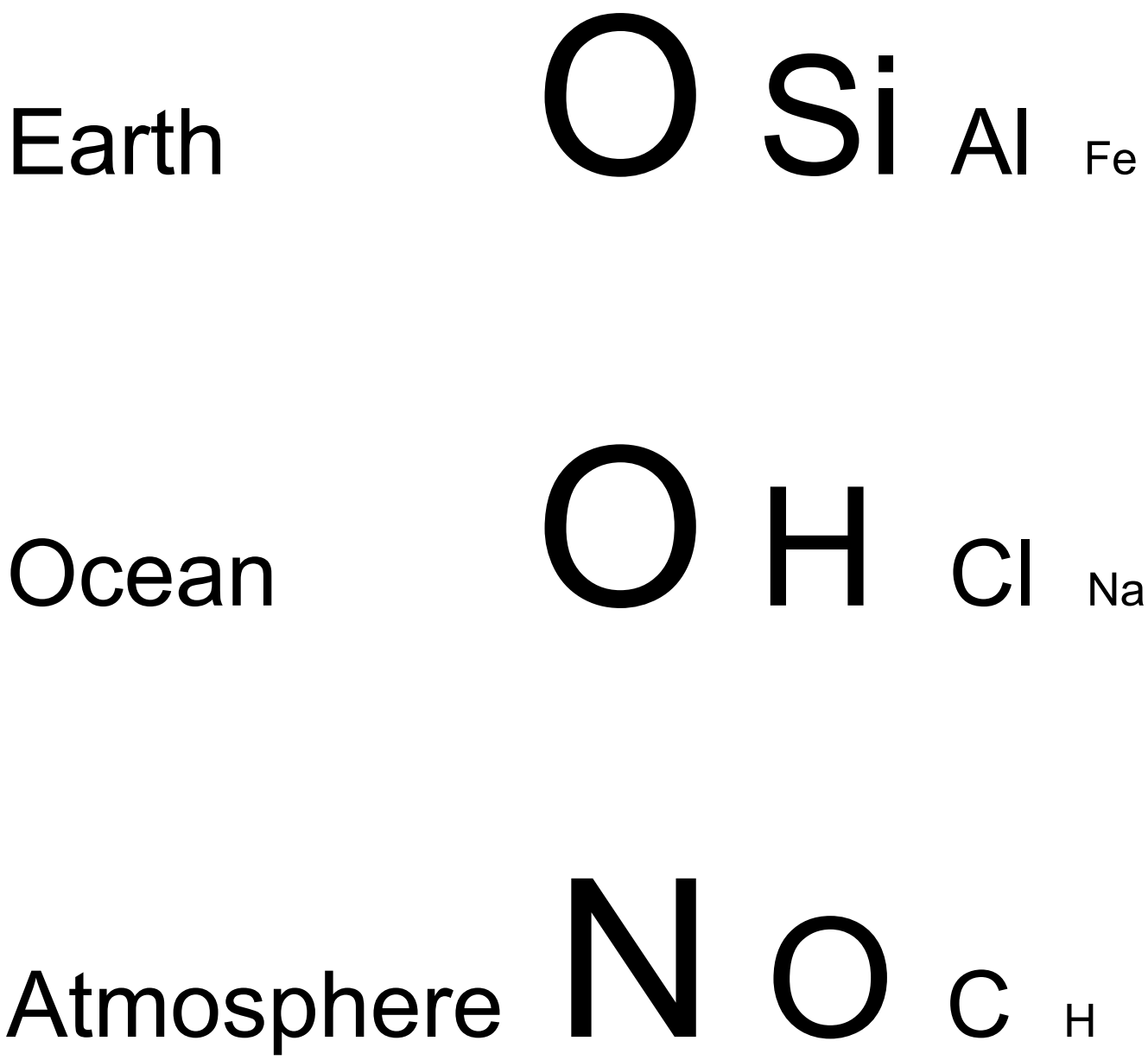
(Veresaev. Dead end. Translation by ENT

Put in original Russian text

1a	2a	3b	4b	5b	6b	7b	8			1b	2b	3a	4a	5a	6a	7a	0
H 1																	He 2
Li 3	Be 4											B 5	C 6	N 7	O 8	F 9	Ne 10
Na 11	Mg 12											Al 13	Si 14	P 15	S 16	Cl 17	Ar 18
K 19	Ca 20	Sc 21	Ti 22	V 23	Cr 24	Mn 25	Fe 26	Co 27	Ni 28	Cu 29	Zn 30	Ga 31	Ge 32	As 33	Se 34	Br 35	Kr 36
Rb 37	Sr 38	Y 39	Zr 40	Nb 41	Mo 42	Tc 43	Ru 44	Rh 45	Pd 46	Ag 47	Cd 48	In 49	Sn 50	Sb 51	Te 52	I 53	Xe 54
Cs 55	Ba 56	La 57	Hf 72	Ta 73	W 74	Re 75	Os 76	Ir 77	Pt 78	Au 79	Hg 80	Tl 81	Pb 82	Bi 83	Po 84	At 85	Rn 86
Fr 87	Ra 88	Ac 89	Rf 104	Ha 105	?? 106												
Lanthinide Series		Ce 58	Pr 59	Nd 60	Pm 61	Sm 62	Eu 63	Gd 64	Tb 65	Dy 66	Ho 67	Er 68	Tm 69	Yb 70	Lu 71		
Actinide Series		Th 90	Pa 91	U 92	Np 93	Pu 94	Am 95	Cm 96	Bk 97	Cf 98	Es 99	Fm 100	Md 101	No 102	Lr 103		

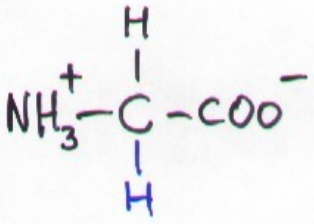
Living matter

O C H N

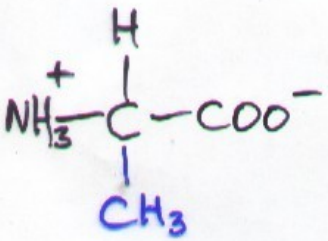

Earth
O Si Al Fe
Ocean
O H Cl Na
Atmosphere
N O C H

Atmosphere N O C H

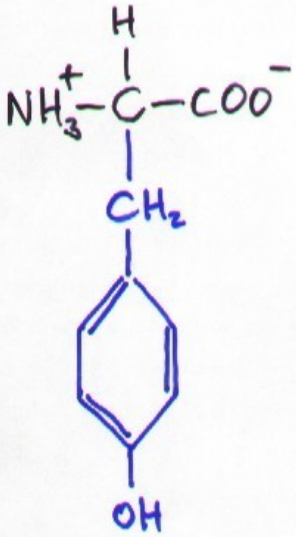
Life O C H N



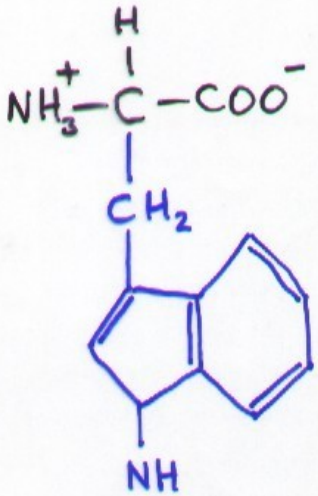
GLYCINE



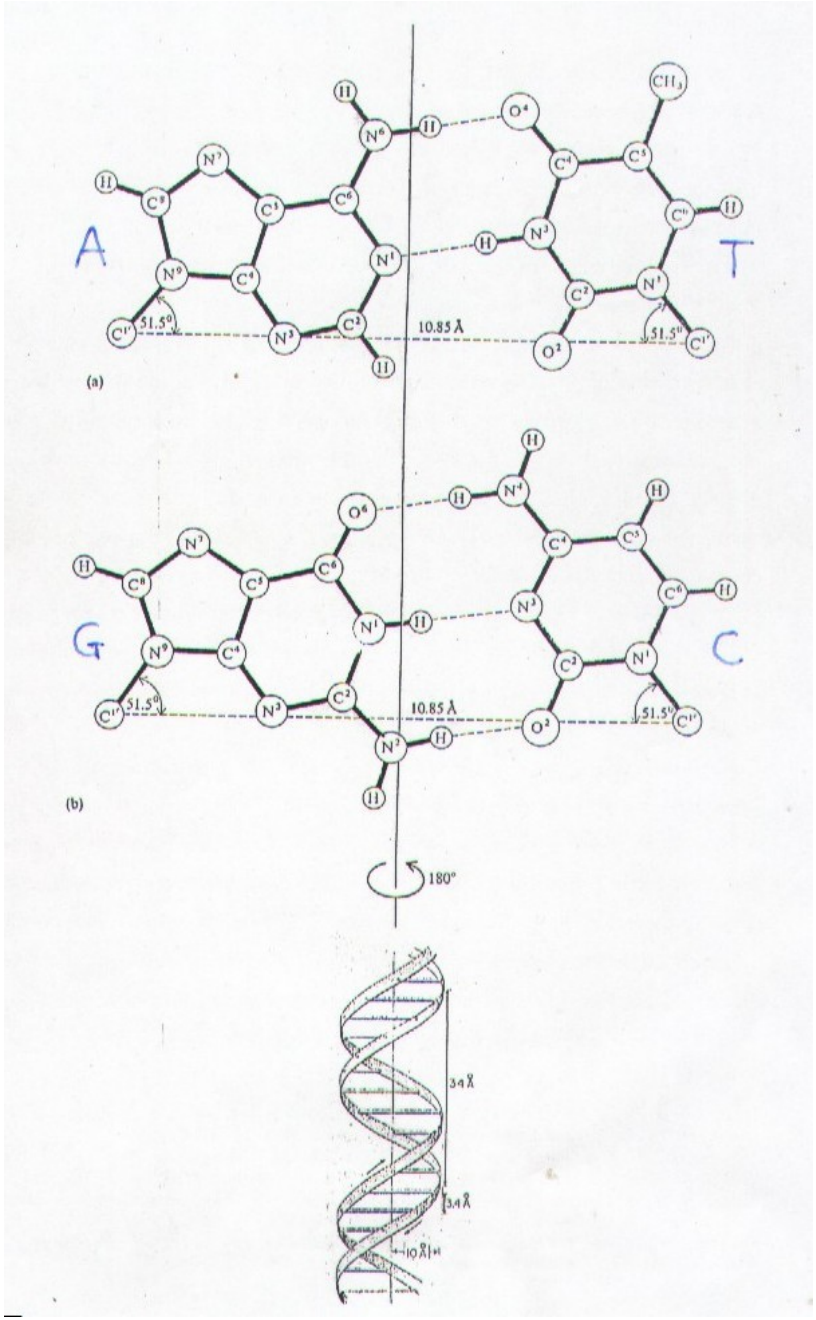
ALANINE



TYROSINE



TRYPTOPHAN


A
T
10.85 Å
51.5°
(a)
G
C
10.85 Å
51.5°
(b)
180°
34 Å
3.4 Å
10 Å

Steps of reconstruction of the earliest Life:

1953-1983 Stanley Miller imitation experiments yielded
A, G, V, D, S, E, P, L, T, I – 10 natural amino acids

1976 Manfred Eigen and Peter Schuster noted that
Alanine and Glycine are encoded today by the most stable and complementary codons GCC/GGC

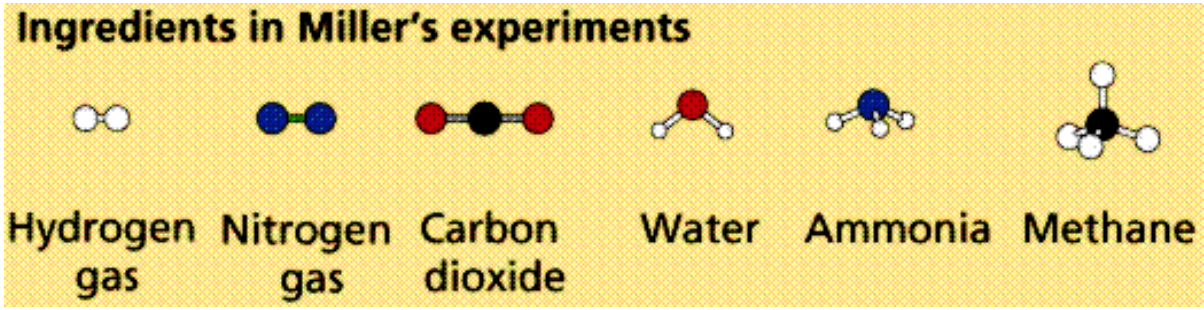
1987-92 Jaime Lagunez-Otero and ENT discovered that
consensus of mRNA is $(GCU)_n$

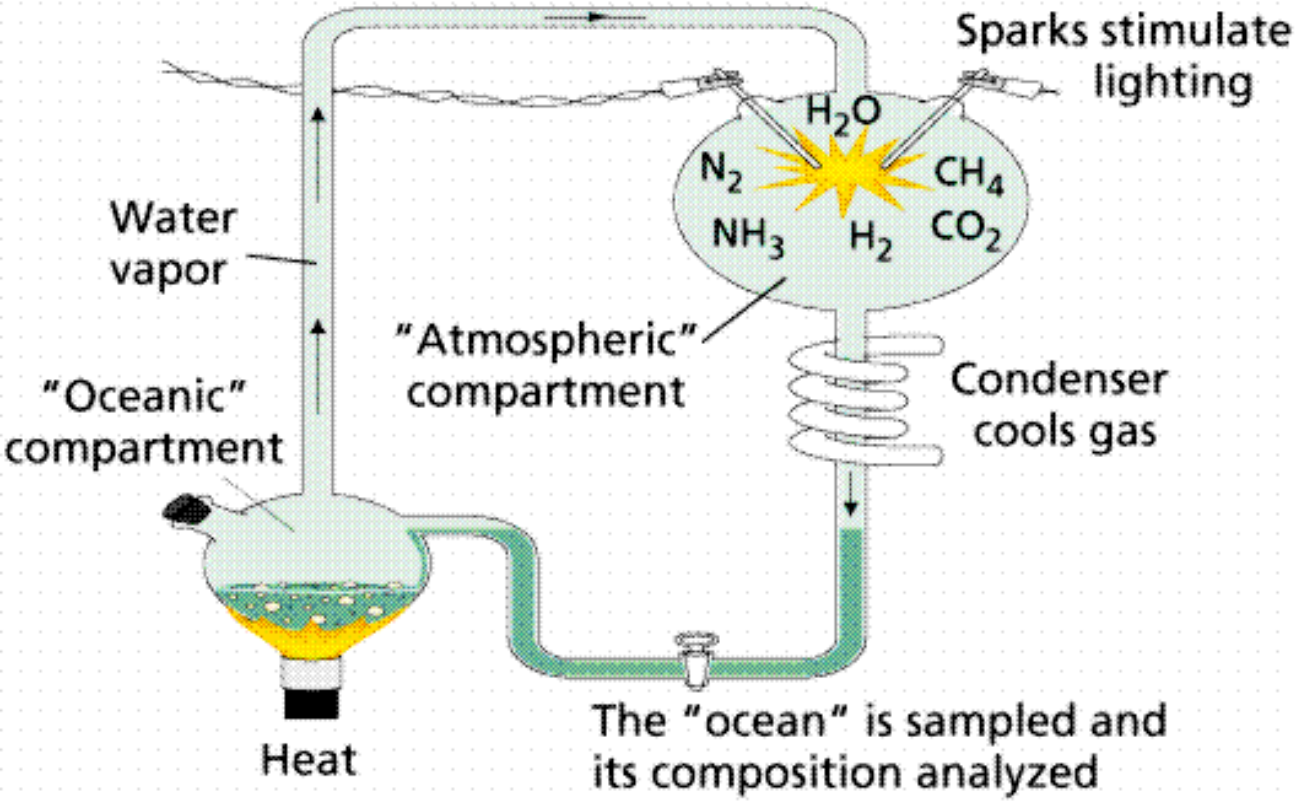
1997 Thomas Bettecken and ENT speculated that
$(GCC)_n/(GGC)_n$ could be the first duplex gene.
This duplex is the most expandable still today.

2000 Evolutionary Chart of Codons is derived

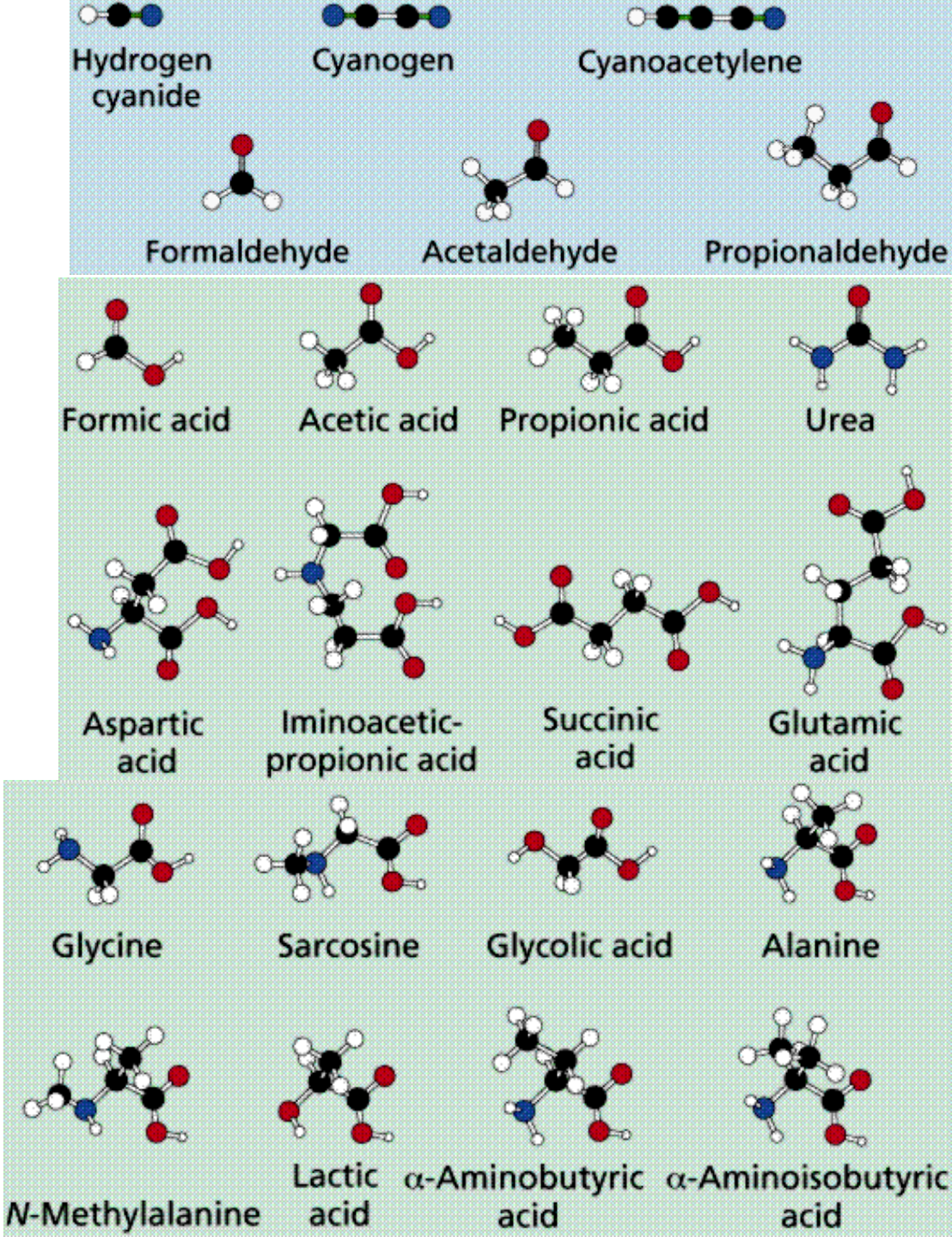
Origin of Life

- Miller's Soup





- Miller's products



aa composition of modern proteins	aa's of Miller mix
L	**L**
A	**A**
G	**G**
S	**S**
V	**V**
E	**E**
I	**I**
T	**T**
K	
D	**D**
R	
P	**P**
N	
Q	
F	
Y	
M	
H	
C	
W	

The imitation experiments of Miller, then Ph. D. student of Harold Urey, have been conducted as side-project, with permission of the supervisor.

Walther Loeb (1913)

first synthesized glycine
in experiments imitating
primordial conditions.

this was recognized only in 1995,
when translation mistake was noticed
(German to English).

“Kohlenoxyd”, carbon monooxide CO,
Instead of “Kohlensaure”, carbonic acid
H_2CO_3 (carbon dioxide CO_2)

Raffaele Saladino
Umberto Ciambecchini,
Claudia Crestini,
Giovanna Costanzo,
Rodolfo Negri,
Ernesto Di Mauro, 2003

first synthesized in primordial conditions
in presence of catalyzers,(TiO_2),

all four nucleobases

in appreciable amounts

J. Biol. Chem. 2007

What are the simplest Living organisms?

Bacteria?
Viruses?

The simplest are viroids.

They consist of just infectious RNA molecules, about 300 bases.

They attack plants (avocado, citruses, potato).

Is that life? But what is life?

“The evolution of life is a trick
of nature to ensure a faster and
better reproduction of the nucleic acids”.

Sol Spiegelman

MASTER t-RNA SEQUENCE

(Eigen and Winkler-Ostwatitsch, Naturwissenschaften 68, 217, 1981)

GCC GGG GUA GCU CAG UUG GUA GAG

anticodon

CGC CGG ACU XXX AAU CCG GAG GUC

GCG GGU UCG AAU CCC GUC CCC GGC ACC A

Consensus sequence of ancient RNA:

($\mathbf{R}$N$\mathbf{Y}$)$_n$ Eigen, Schuster, 1976

MASTER t-RNA:

	I	II	III
A+G	**16**	10	11
C+U	8	13	**13**

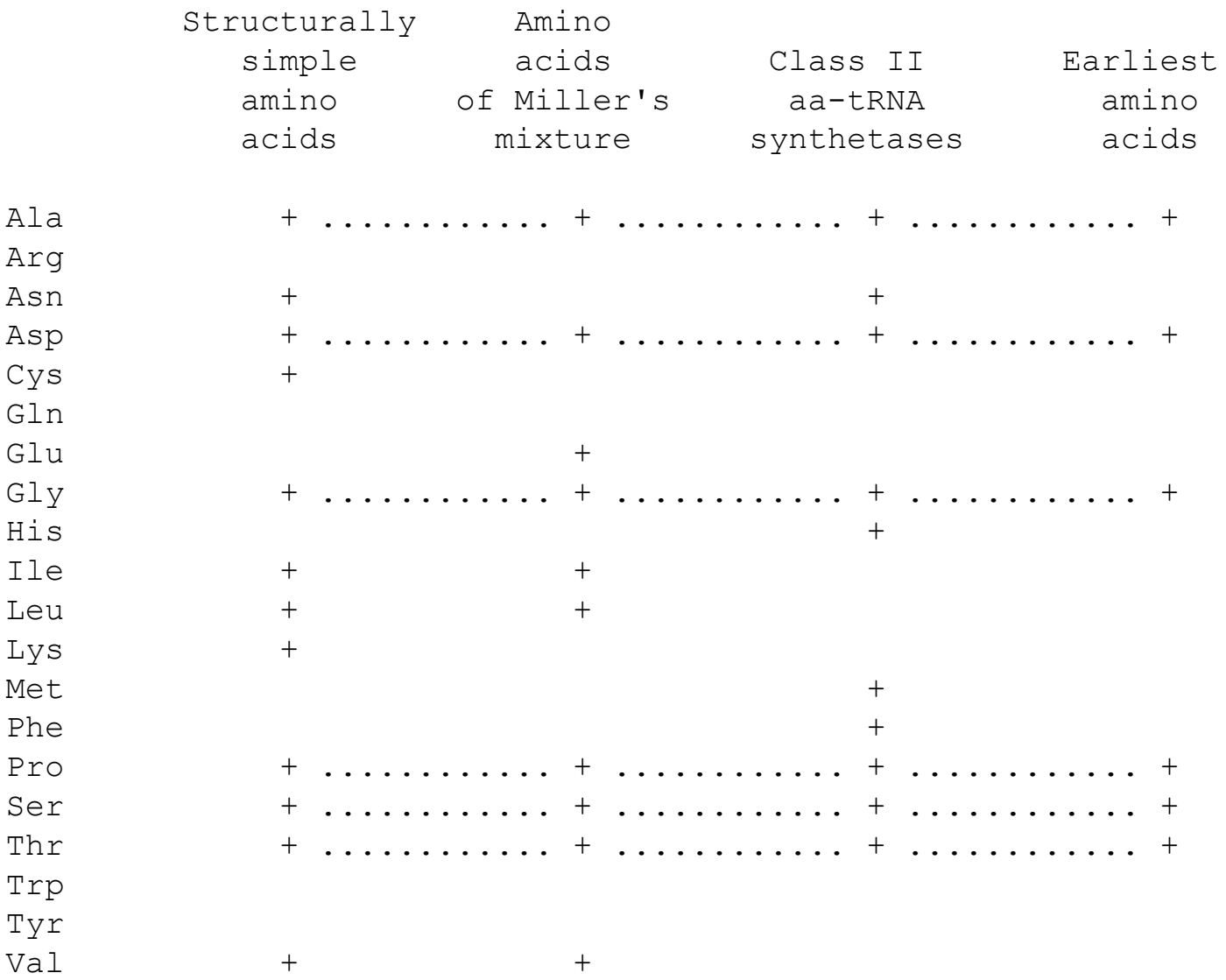
BUT, ACTUALLY:

	I	II	III	
A	4	5	2	
C	6	8	8	($\mathbf{G}$NN)$_n$
G	**12**	5	9	
U	2	5	5	

“We must admit that we had expected more noise accumulation during later stages of evolution, so that the memory of a triplet pattern - which has no foundation in tRNA present adaptor function – came out as a true surprise”

Eigen, Winkler-Ostwatitsch,
Naturwissenschaften 68, 282-292, 1981

-the headacke surprise since 1979 (Braunlage) until 2006 (Les Treilles).


Structurally simple amino acids
Amino acids of Miller's mixture
Class II aa-tRNA synthetases
Earliest amino acids
Ala
Arg
Asn
Asp
Cys
Gln
Glu
Gly
His
Ile
Leu
Lys
Met
Phe
Pro
Ser
Thr
Trp
Tyr
Val

Triplet code and its early form

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

Evolutionary chart of codons

39 criteria for amino-acid chronology (2000)

1. Simplicity (number of non-hydrogen atoms)
2. Involvement with more ancient synthetases of class II
3. Yield in the Miller's experiments
4. Amino-acid composition of extant proteins
5. Chemical inertness
6. Stability of codon-anticodon interactions
7. Molecular clock sequence analysis of synthetases
8. Stability of ("older") assignments in the table of the code
9. Jukes' theory of the origin of the code
10. Coevolution theory of Wong
11. GCU-based theory of Trifonov and Bettecken
12. RRY hypothesis of Crick
13. RNY hypothesis, Eigen and Schuster
14. Hypothesis of Hartman
15. Hypothesis of Ferreira
16. Prebiotic physicochemical code of Altshtein-Efimov
17. Early copolymerization code of Nelsestuen
18. Composition of proteinoids of Fox
19. Coevolution theory of Dillon
20. Yield in imitation experiments of Fox and Windsor
21. Yield in experiments of Harada and Fox, high temperatures.
22. Yield in shock wave experiments of Bar-Nun
23. Coevolution theory of Wächtershäuser
24. Remnants of primordial code in tRNA (Möller and Janssen)
25. Evolutionary distances between isoacceptor tRNAs
26. Hypothesis of O. Ivanov
27. Match scores of BLOSUM matrix
28. A/U start, Jimenez-Sanchez
29. N-fixing amino acids first, Davis
30. GNN codons first, Taylor and Coates
31. Algebraic model of Hornos and Hornos
32. Composition of translated Urgen
33. Murchison meteorite
34. Minimal graph complexity, amino acids
35. Minimal graph complexity, amino-acid residues
36. Hypothesis of Jimenez-Montano
37. "Size/complexity" score, Dufton
38. Minimal alphabet for folding
39. DNA stability 40. RNA duplex stability

(ENT, 2000)

```
    1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1. G  A  -CS-  - PTV -  -  -DILMN-  -  - EKQ -  H  -FR-  Y  W
 2. -AG-  -  -  -DFHKNPST-  -  -  -  -  -  CEILMQRVWY  -  -  -
 3. A  G  D  V  L  E  I  S  P  T  M  K  -  -  -CFHNQRWY-  -  -
 4. L  A  G  S -VE- -IT-  K  D  R  P  N  Q  F  Y -HM- -CW-
 5. -  -  AFGILPV  -  -  -  NQST  -  -  -  - CDEHKMRWY -  -  -
 6. A -GP-  - DES -  -TV-  R  L  -  CHQW  -  -IM-  Y  - FKN -
 7. -  -  -  -  -  - ACDEFGHIKLMNPQRSTV -  -  -  -  -  -  -WY-
 8. -  -  ADEFGHP  -  -  V  I  -  KNSY  -  - MTW -  L  R  -CQ-
 9. -  -  -  ADEGHLPQRV  -  -  -  -  -  -CFIKNSTY-  -  -  -MW-
10. -  -ADEGS-  -  V  -PT-  -IL-  F  C  Y  -KR-  -NQ-  H  -MW-
11. A  -  - DGPSTV -  -  E  -  -  -  - CFHIKLMNQRWY -  -  -  -
12. -  DGNS  -  -  -  -  -  -  ACEFHIKLMPQRTVWY  -  -  -  -  -
13. -AG-  -  - DINSTV -  -  -  -  -  - CEFHKLMPQRWY -  -  -  -
14. G  P  A  R  -  - DENQST -  -  -HK-  C  -  -FILVY-  -  -MW-
15. -  - FGKLNP -  -  -  -  -  CDEHQRSTVW  -  -  -  -  AIMY  -
16. -  -  - ADEGKRSTV -  -  -  -  -  -  -CFHILMNPQWY-  -  -  -
17. -  -  -  - DEFHIKLMSTVY -  -  -  -  -  -  -ACGNPQRW-  -  -
18. A  E  V -GK-  M  L  C  Y -NQ-  I -DF-  R  H  P  W  T  S
19. G  A  D  V  E  Q  -  HLPR  -  N  T -IS-  -KM-  F -CY-  W
20. G  I -AP-  S  E  D  F  L  V  -  -  -  CHKMNQRTWY  -  -  -
21. G  A  E  D  L -PV-  S  I  T -FY-  -  -  -CHKMNQRW-  -  -
22. G  A  V  L  -  -  -  -  -  CDEFHIKMNPQRSTWY  -  -  -  -  -
23. -DE-  -  -  -ACGNPQST-  -  -  -  ILMV  -  -  - FHKRWY -  -
24. -  ADGV  -  -  -  -  -  -  CEFHIKLMNPQRSTWY  -  -  -  -  -
25. Q  H  P -LS-  G  C  W  R  V -DE-  A  Y  T -IM-  F -KN-
26. -  -  -  ADEGLPRSTV  -  -  -  -  -  -  CFHIKMNQWY  -  -  -
27. -  -AILSV-  -  -  - EKMQRT -  -  -  DFGN  -  -PY-  H  C  W
28. -  -  FIKLMNY  -  -  -  -  -  CDEHQRSTVW  -  -  -  - AGP -
29. -  DENQ  -  -  APSV  -  -CG-  T  - ILM -  R  K -FY-  H  W
30. -  -ADEGV-  -  -  -  -  -  - CFHIKLMNPQRSTWY -  -  -  -  -
31. -  -CDFSV-  -  -  -EKLRY-  -  -HP-  -  -  -AGIMNQTW-  -  -
32. V  - AGP -  -  ENRT  -  - LQS -  -  -  - CDFHIKMYW -  -  -
33. -AG-  -  DEPV  -  -  -  -  -  -CFHIKLMNQRSTWY-  -  -  -  -
34. G  A  D  P -CS-  N  E  V  K  Q  T  L  M  I  R  H  F  Y  W
35. G  A -CS-  P  V  K  M  T  L -DI-  N  E  Q  H  F  R  Y  W
36. -  ADGV  -  - LPR -  -  - CIKQST -  -  -  -  EFHMNWY  -  -
37. G  A  V -IL-  S  T  K  P  D  N  E  Q  F  R  Y  C  H  M  W
38. -  -AGEIK-  -  -  -  -  -  - CDFHLMNPQRSTVWY -  -  -  -  -
39. A  G  S  R  C  T  D  V  P  E  W -HN-  F  L  I  Y  M -KQ-
40. G  A  P  W -RS-  C  D  T  E  H  V -LM-  Q  I  Y  N  F  K
41. -  ADGS  -  -  CPQV  -  -  - EFIKNT -  -  -  - HLMRWY -  -
42. G  A  C  S  D  V  -  -  -  -  -EFHIKLMNPQRTWY-  -  -  -  -
43. -  -AGPTV-  -  L  R  S  I  -  -  - CDEFHKNQY -  -  -  -MW-
44. G  A  S  P  D  C  N  T  E  V  Q  H  M -LI-  K  R  F  Y  W
45. G  L  A  V  D  E  P  I  T  R  F  K  S  Y  N  H  Q  M  W  C
46. -  -  -  -  ADEFGIKLNQTVY  -  -  -  -  -  -  CHMPRSW  -  -
47. -  -  - ADGHINSTV -  -  -  - MPR -  -  -  -CEFKLQWY-  -  -
48. -  -  -  -  -  -ADEGHIKLMNPQRSTVW-  -  -  -  -  -  - CFY -
49. -  AGPR  -  -  -  -  CDEHLQSTVW  -  -  -  -  - FIKMNY  -  -
50. -  ADGV  -  -  - EHLPQR -  -  -  -  -  CFIKMNSTWY  -  -  -
51. D  N  T  E  Q  K  P  I  M  S  G  A  R  V  L  C  H  Y  F  W
52. -  -  -  -  ADEFGHLPQRSTV  -  -  -  -  -  -  CIKMNWY  -  -
53. -  -AILPV-  -  -  -DEGST-  -  -  -CFHNY-  -  -  -KMQRW-  -
54. -  - ADEGSV -  -  -  -KLPRT-  -  -  -  - CFHIMNQWY -  -  -
```

Table 2. Thermostability of the codons (complementary pairs, kcal/M)

A	GCC	28.3	K	AAG	17.3	R	AGG	23.9
	GCG	25.5		AAA	13.6		AGA	22.9
	GCU	25.4	L	CUC	22.9	S	UCC	25.8
	GCA	25.3		CUG	20.9		UCG	23.1
C	UGC	25.3		CUA	18.2		UCU	22.9
	UGU	21.8		CUU	17.3		UCA	22.9
D	GAC	23.8	L	UUG	17.3	S	AGC	25.4
	GAU	21.8		UUA	14.5		AGU	21.9
E	GAG	22.9	M	AUG	19.8	T	ACC	24.8
	GAA	19.3	N	AAC	18.2		ACG	22.0
F	UUC	19.3		AAU	16.3		ACU	21.9
	UUU	13.6	P	CCC	26.8		ACA	21.8
G	GGC	28.3		CCG	24.0	V	GUC	23.8
	GGG	26.8		CCU	23.9		GUG	21.8
	GGA	25.8		CCA	23.8		GUA	19.1
	GGU	24.8	Q	CAG	20.9		GUU	18.2
H	CAC	21.8		CAA	17.3	W	UGG	23.8
	CAU	19.8	R	CGC	25.5	Y	UAC	19.1
I	AUC	21.8		CGG	24.0		UAU	17.1
	AUA	17.1		CGA	23.1			
	AUU	16.3		CGU	22.0			

(Xia et al., 1998)

Consensus temporal order of amino acids (single-factor criteria)

amino acids of Miller		average rank (0.7)	order	codon capture cases
+	G	2.8	1	
+	A	3.9	2	
+	V	6.5	3	
+	S	7.1	4	
+	P	7.4	5	
+	D	7.7	6	
+	T	9.0	7	
+	E	9.9	8	
+	L	10.3	9	(+)
+	I	10.9	10	(+)
	N	11.2	11	
	R	11.7	12	
	H	12.7	13	+
	Q	12.8	14	+
	K	13.2	15	
	F	13.2	16	+
	C	13.9	17	+
	M	15.0	18	+
	W	15.3	19	+
	Y	15.3	20	+

Consensus temporal order of amino acids (multi-factor criteria)

amino acids of Miller		average rank (0.7)	order	codon capture cases
+	A	4.1	1	
+	G	4.2	2	
+	D	4.2	3	
+	V	6.1	4	
+	E	6.3	5	
+	P	7.2	6	
+	S	8.0	7	
+	L	9.5	8	(+)
+	T	9.8	9	
	Q	9.9	10	(+)
	R	10.2	11	
	N	11.4	12	
+	I	11.9	13	(+)
	H	13.2	14	+
	K	13.4	15	
	C	13.8	16	+
	F	15.1	17	+
	Y	15.2	18	+
	M	15.9	19	+
	W	17.7	20	+

Consensus temporal order of amino acids (final)

amino acids of Miller		average rank (0.7)	order	codon capture cases
+	G	3.5	1	
+	A	4.0	2	
+	D	6.0	3	
+	V	6.3	4	
+	P	7.3	5	
+	S	7.6	6	
+	E	8.1	7	
+	T	9.4	8	
+	L	9.9	9	(+)
	R	11.0	10	
	N	11.3	11	
+	I	11.4	12	(+)
	Q	11.4	13	(+)
	H	13.0	14	+
	K	13.3	15	
	C	13.8	16	+
	F	14.2	17	+
	Y	15.2	18	+
	M	15.4	19	+
	W	16.5	20	+

Persistence of the ranking

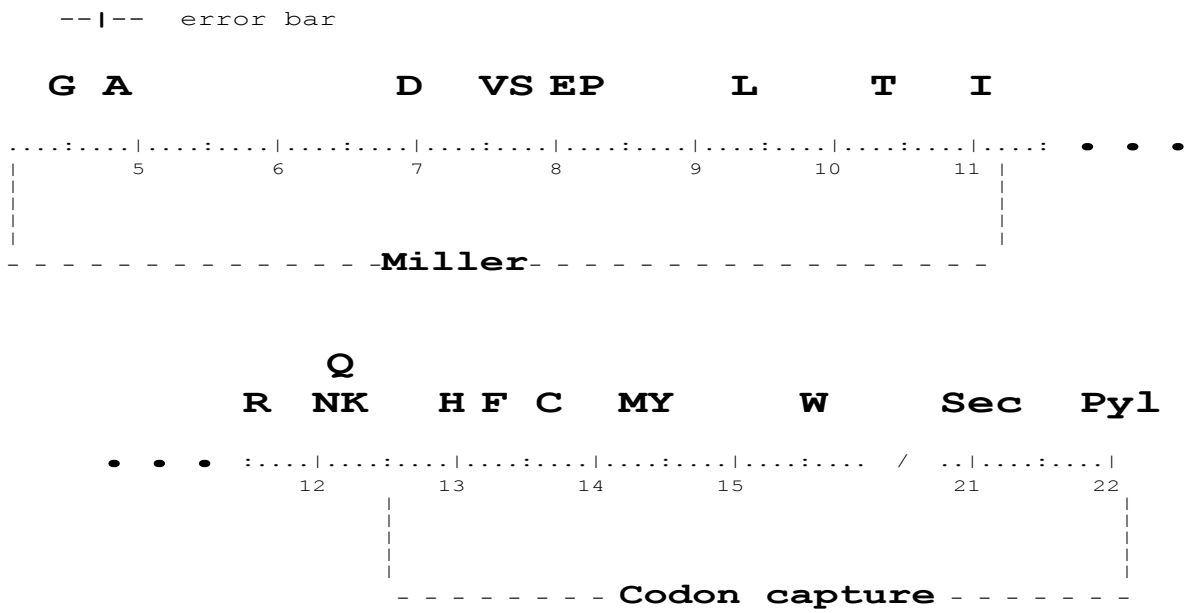
	Number of criteria (simple averaging)					Filtered	
	3	**7**	**25**	**28**	**40**	**one**	**two**
rank							
1.	G	A	G	G	G	G	G
2.	A	G	A	A	A	A	A
3.	S	S	D	V	V	V	V
4.	D	P	V	D	D	D	D
5.	P	V	P	S	S	S	E
6.	T	T	S	P	E	E	P
7.	V	L	E	E	P	P	S
8.	L	D	L	L	L	L	L
9.	I	I	T	T	T	T	T
10.	K	E	I	I	I	N	R
11.	N	N	N	N	N	R	N
12.	E	F	F	R	R	K	K
13.	C	K	H	F	K	I	Q
14.	M	R	K	K	Q	Q	I
15.	H	Q	R	Q	C	H	C
16.	F	C	Q	H	F	C	H
17.	Q	H	C	C	H	F	F
18.	R	M	M	M	M	M	M
19.	Y	W	Y	Y	Y	Y	Y
20.	W	Y	W	W	W	W	W

Consensus chronology of amino acids (2000)

Raw data				Filtered data				Miller
G	4.4	0.7	1	G	2.9	0.3	1	G
A	4.9	0.8	2	A	2.9	0.3	2	A
V	6.9	0.6	3	V	6.6	0.6	3	V
D	7.2	0.7	4	D	7.0	0.7	4	D
S	7.9	0.7	5	E	7.2	0.6	5	E
E	8.2	0.7	6	P	7.5	0.6	6	P
P	8.3	0.7	7	S	7.7	0.7	7	S
L	9.4	0.7	8	L	9.5	0.7	8	L
T	10.1	0.6	9	T	9.8	0.6	9	T
I	11.2	0.7	10	R	11.5	0.7	10	
N	11.8	0.7	11	N	12.2	0.7	11	
R	12.0	0.7	12	K	12.3	0.5	12	
K	12.0	0.7	13	Q	13.0	0.4	13	
Q	12.4	0.7	14	I	13.0	0.5	14	I
C	12.4	0.7	15	C	14.3	0.6	15	
F	13.0	0.7	16	H	14.9	0.5	16	
H	13.3	0.6	17	F	15.1	0.4	17	
M	14.0	0.6	18	M	15.4	0.4	18	
Y	14.7	0.5	19	Y	15.6	0.4	19	
W	15.8	0.6	20	W	16.7	0.5	20	

CONSENSUS TEMPORAL ORDER OF AMINO ACIDS

(101 VECTORS AVERAGED)



GCC – codon for alanine (A),

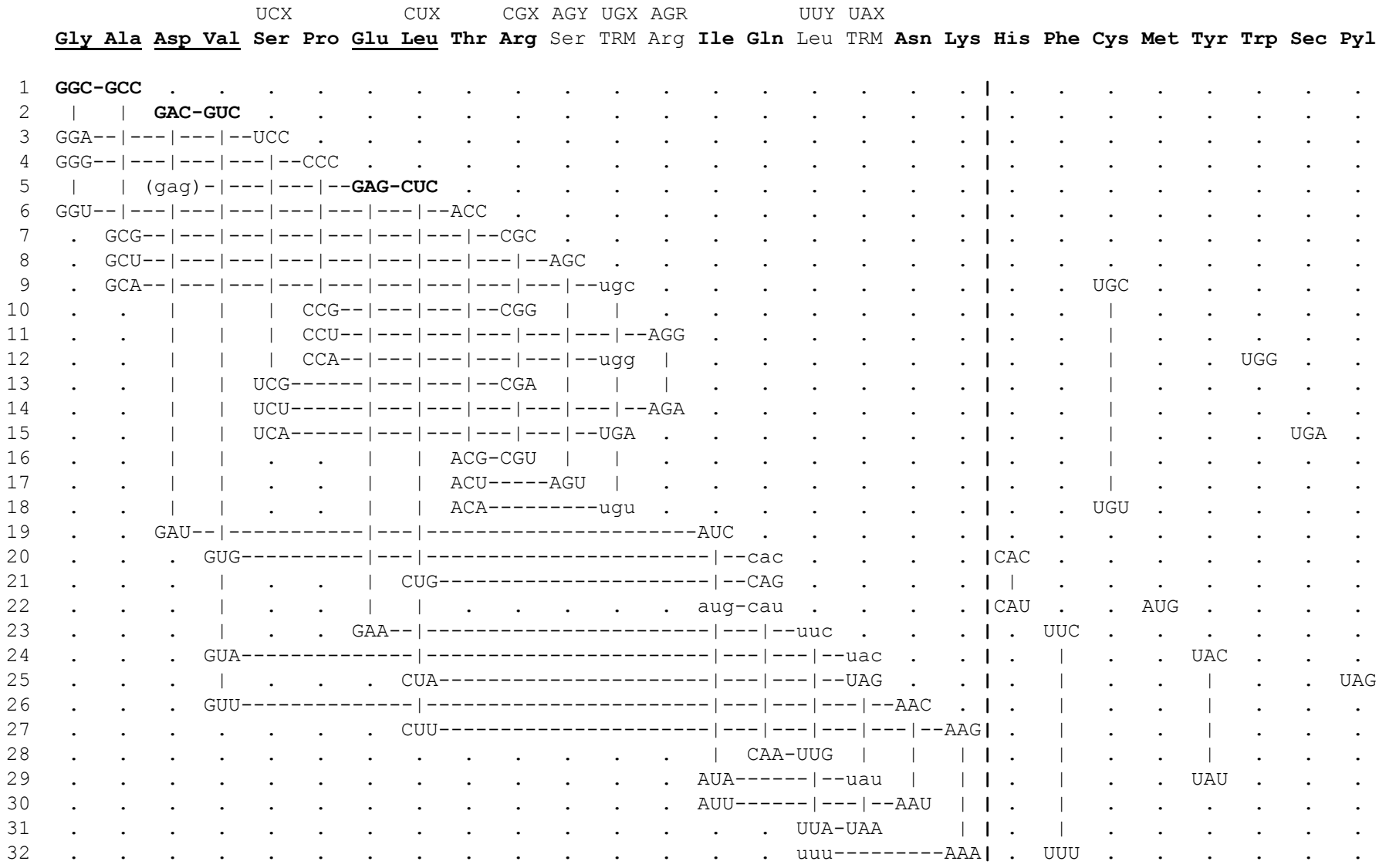
GGC – codon for glycine (G).

Both are of the highest yield
in imitation experiments of Stanley Miller

EVOLUTION OF THE TRIPLET CODE

E. N. Trifonov, December 2007, Chart 101

Consensus temporal order of amino acids:



CONSECUTIVE ASSIGNMENT OF 64 TRIPLETS

CODON CAPTURE

aa "age":

17 17 16 16 15 14 13 13 12 11 10 9 8 7 6 5 4 3 2 1

THE OLD NEW RULES IN EVOLUTION OF THE TRIPLET CODE

1. **ABIOTIC START (Miller, 1953)**
   Initial set of amino acids is
   of purely chemical origin

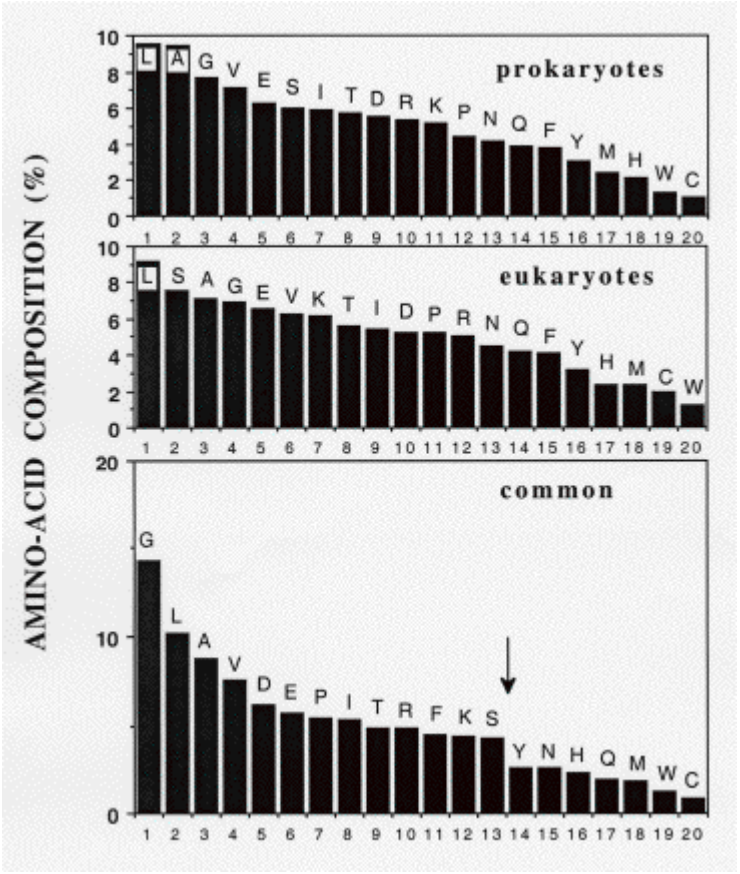
2. **COMPLEMENTARITY (Eigen and Schuster, 1978)**
   New codons are introduced as
   complementary pairs

3. **THERMOSTABILITY (Eigen and Schuster, 1978)**
   The codons that make the most
   stable pairs with their
   anticodons are engaged first

4. **PROCESSIVITY**
   New codons are derived from
   the earlier ones by mutations
   in redundant third positions
   and complementary copying

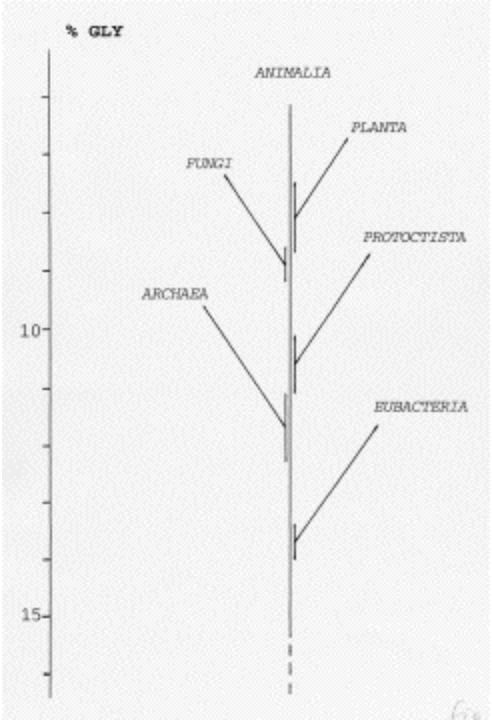
GLYCINE CLOCK

Set	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
1.	52	10	51	31	30	**111**	17	39	26	67	19	20	32	16	41	28	44	54	9	21	
	7.2	1.4	7.1	4.3	4.2	**15.5**	2.4	5.4	3.6	9.3	2.6	2.8	4.5	2.2	5.7	3.9	6.1	7.5	1.3	2.9	%
2.	58	6	37	35	18	**80**	17	34	31	65	11	18	37	6	32	21	29	42	6	16	
	9.7	1.0	6.2	5.8	3.0	**13.4**	2.8	5.7	5.2	10.9	1.8	3.0	6.2	1.0	5.3	3.5	4.8	7.0	1.0	2.7	%
3.	65	8	34	39	32	**108**	12	46	27	79	11	18	30	18	27	40	35	54	8	15	
	9.0	1.1	4.7	5.4	4.4	**14.9**	1.7	6.4	3.7	10.9	1.5	2.5	4.1	2.5	3.7	5.5	4.8	7.5	1.1	2.1	%
4.	72	7	67	55	51	**133**	26	45	40	72	16	33	51	13	36	46	49	85	10	25	
	7.7	0.8	7.2	5.9	5.5	**14.3**	2.8	4.8	4.3	7.7	1.7	3.5	5.5	1.4	3.9	4.9	5.3	9.1	1.1	2.1	%
5.	135	10	78	83	49	**188**	21	68	78	126	19	26	74	18	61	54	72	117	6	29	
	10.3	0.8	5.9	6.3	3.7	**14.3**	1.6	5.2	5.9	9.6	1.4	2.0	5.6	1.4	4.6	4.1	5.5	8.9	0.5	2.2	%
6.	54	4	44	33	29	**91**	20	40	23	86	6	11	32	15	31	20	22	34	8	18	
	8.7	0.6	7.2	5.4	4.7	**14.9**	3.3	6.5	3.8	14.1	1.0	1.8	5.2	2.5	5.1	3.3	3.6	5.6	1.3	2.9	%
7.	55	7	35	39	41	**82**	12	23	21	71	16	16	44	18	42	28	20	38	18	19	
	8.5	1.1	5.4	6.0	6.4	**12.7**	1.9	3.6	3.3	11.0	2.5	2.5	6.8	2.8	6.5	4.3	3.1	5.9	2.8	2.9	%
Tot.	491	52	346	315	250	**793**	125	295	246	566	98	142	300	104	270	237	271	424	65	143	
	8.8	0.9	6.2	5.7	4.5	**14.3**	2.3	5.3	4.4	10.2	1.8	2.6	5.4	1.9	4.9	4.3	4.9	7.6	1.2	2.6	%

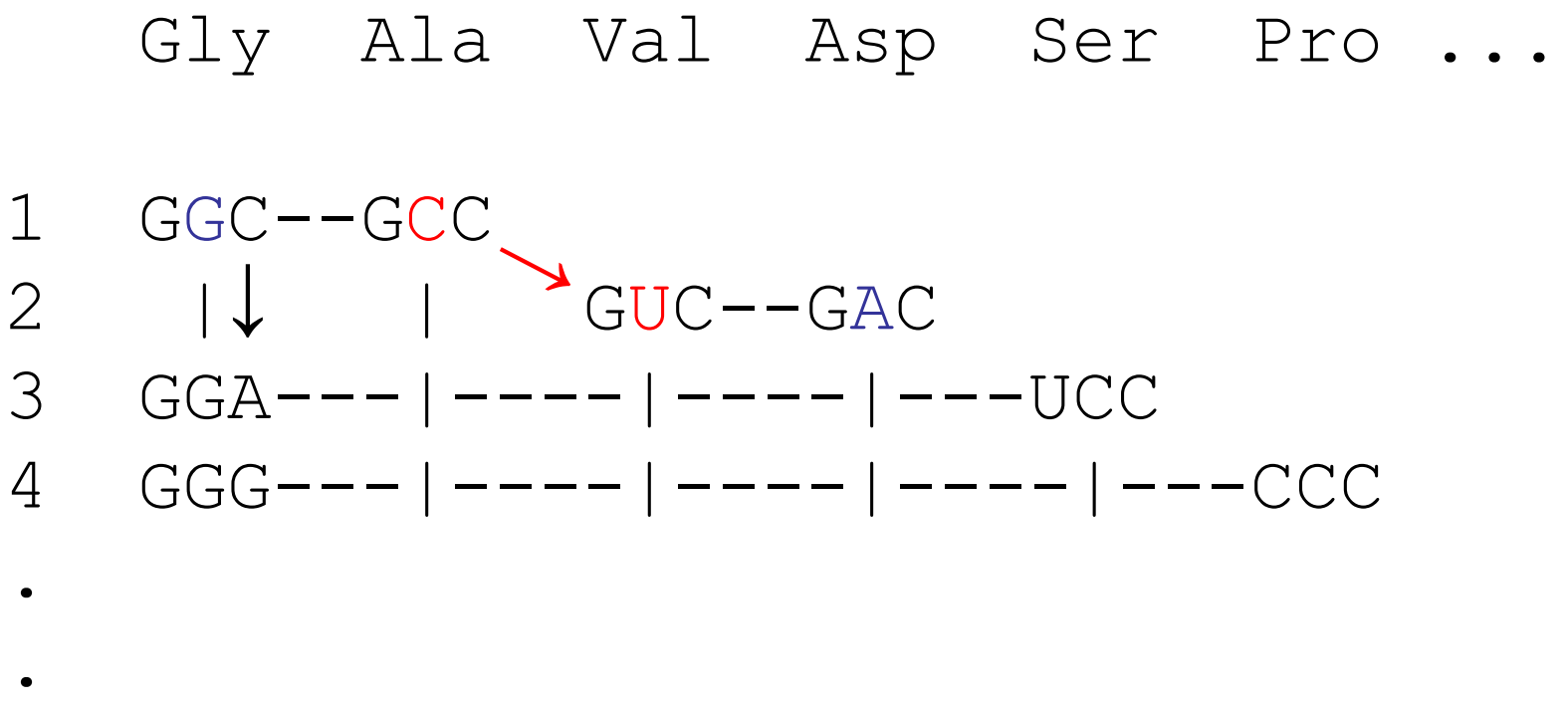

AMINO-ACID COMPOSITION (%)
prokaryotes
L A G V E S I T D R K P N Q F Y M H W C
eukaryotes
L S A G E V K T I D P R N Q F Y H M C W
common
G L A V D E P I T R F K S Y N H Q M W C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Contents of shared glycine (%) in kingdom-to-kingdom protein sequence alignments

	ANIMALIA	*PLANTA*	*FUNGI*	*PROTOCTISTA*	*ARCHAEA*	Branching level
PLANTA	8.8 ± 0.4 (*51*)					**8.8 ± 0.4** (426/4862, *51*)
FUNGI	8.8 ± 0.4 (573/6479, *70*)	8.8 ± 0.4 (391/4427, *50*)				**8.8 ± 0.3** (964/10906, *120*)
PROTOCTISTA	9.6 ± 0.6 (300/3127, *28*)	9.9 ± 0.6 (324/3283, *27*)	9.8 ± 0.5 (321/3262, *27*)			**9.8 ± 0.3** (945/9672, *82*)
ARCHAEA	11.1 ± 0.7 (222/1994, *30*)	12.9 ± 0.9 (215/1669, *26*)	12.5 ± 0.8 (245/1961, *31*)	13.9 ± 1.3 (109/787, *13*)		**12.3 ± 0.4** (791/6411, *100*)
EUBACTERIA	14.9 ± 0.6 (685/4590, *70*)	13.5 ± 0.6 (546/4041, *44*)	13.4 ± 0.5 (667/4966, *70*)	11.4 ± 0.7 (304/2656, *28*)	13.3 ± 0.8 (304/2288, *35*)	**13.5 ± 0.3** (2506/18541, *247*)


% GLY
ANIMALIA
PLANTA
FUNGI
PROTOCTISTA
ARCHAEA
EUBACTERIA
10
15

Ancient binary alphabet

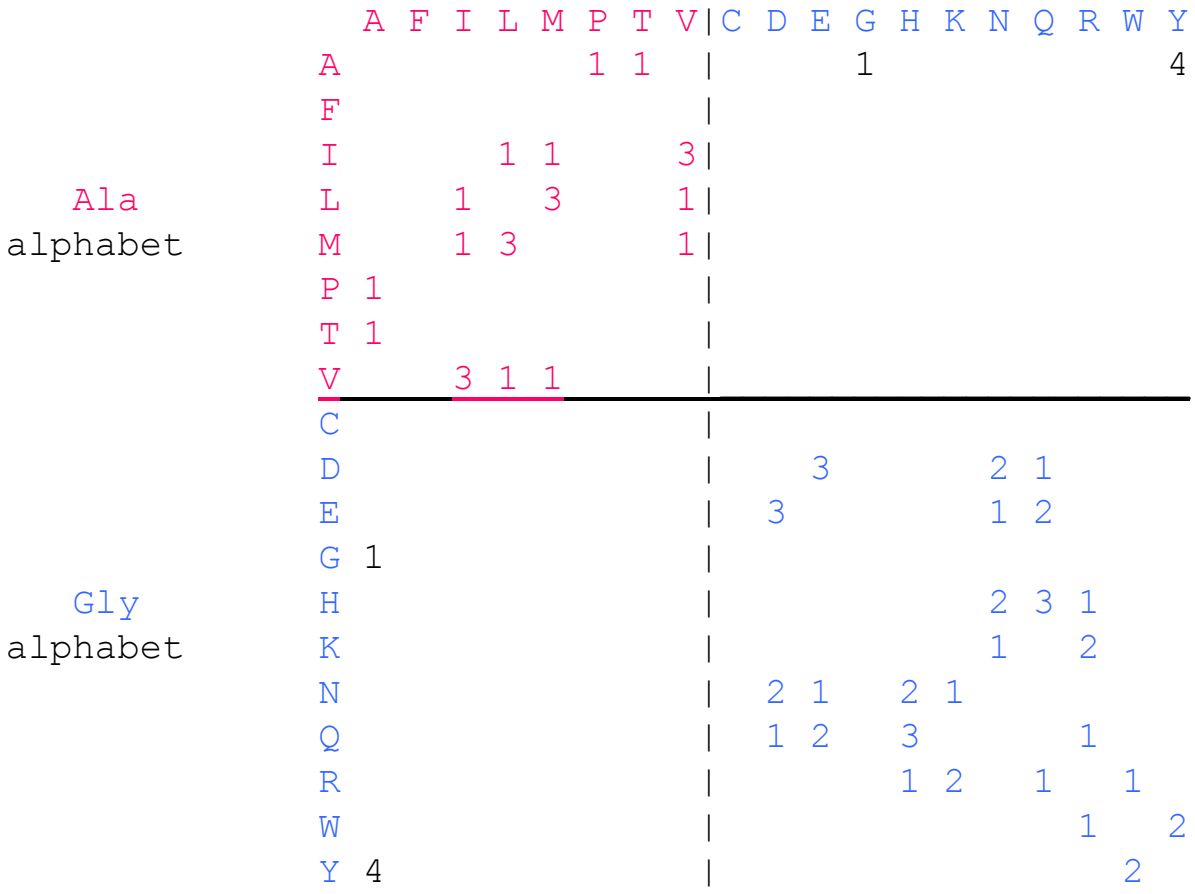


At every step of the evolution of the codons
middle purines remain purines (R→R),
middle pyrimidines remain pyrimidines (Y→Y).

Reconstruction of evolutionary history
of the triplet code suggests that
the earliest protein sequences could be presented
in the binary alphabet of two types of amino acids –

those encoded by xYx triplets (Ala family, *A*) and
those encoded by xRx triplets (Gly family, *G*).

The conclusion about two alphabets is strongly supported by respective rearrangements of substitution matrices:



Rearranged PAM120 substitution matrix

(original matrix in Altschul SF, JMB 219, 555, 1991)

Rearranged BLOSUM substitution matrix

(original matrix in Henikoff S, Henikoff JG, PNAS 89, 10915,1992)

Using the two-letter alphabet one can rewrite modern sequences in their (presumed) ancient version

AFLIIMVRKREDQNFFVTAMAQQNEDGR

AFLIIMVRKREDQNFFVTAMAQQNEDGR

AAAAAAAGGGGGGGAAAAAAAGGGGGGG

"I assume that the earliest proteins were smal
of **about ten amino acids**, and specified by sma
primitive genes, probably made of RNA"

"In the next stage, I postulate that the genes
joined together at random and a primitive
splicing mechanism concatenates the peptides
into longer molecules"

Sidney Brenner,
Nature 334, 528-530, 1988

Rewriting modern amino acid sequence in the binary form

would suggest
what was the **ancestral form** of that sequence,

all the way to original Alanines and Glycines only

The *G* to *A* and *G* to *G* distance analysis of modern protein sequences suggests that the very first miniproteins had the structure

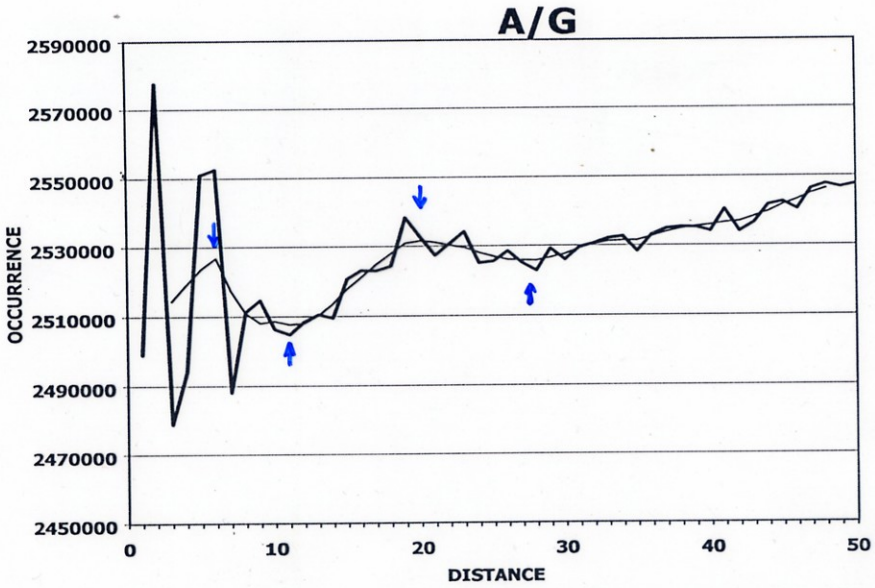
GGGGGGG and AAAAAAA

encoded by the duplex

xRx xRx xRx xRx xRx xRx xRx

xɅx xɅx xɅx xɅx xɅx xɅx xɅx

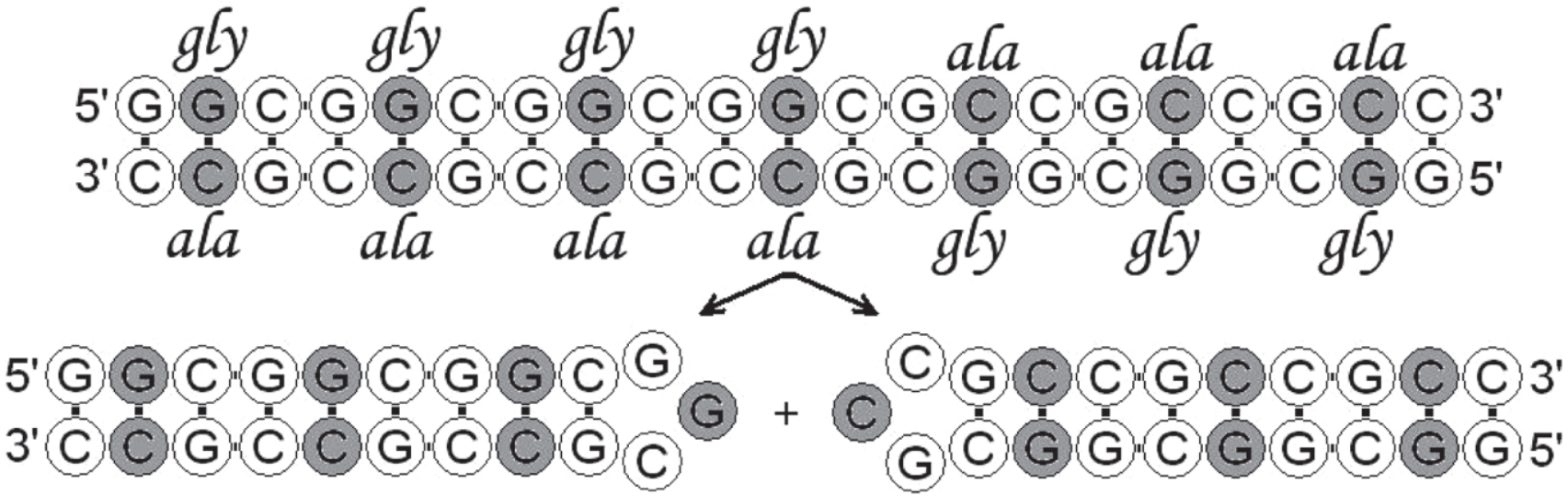
The size of the original miniproteins is estimated from modern sequences written in binary form to be 7 amino acid residues (J. Mol. Evol. 53, 394-401, 2001).The same estimate is provided by sequence fossils of ancient hairpins in mRNA(J Biomol Str Dyn 24, 163-170, 2006)

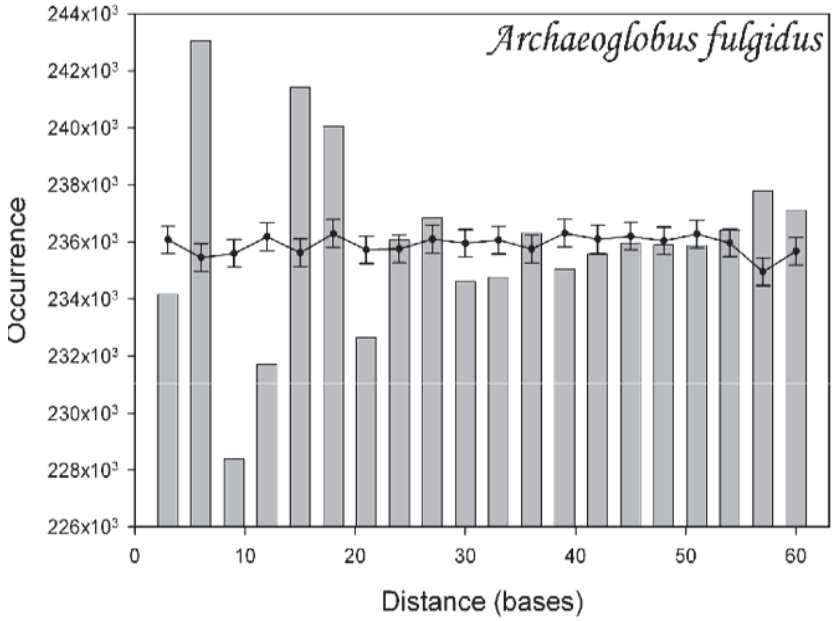


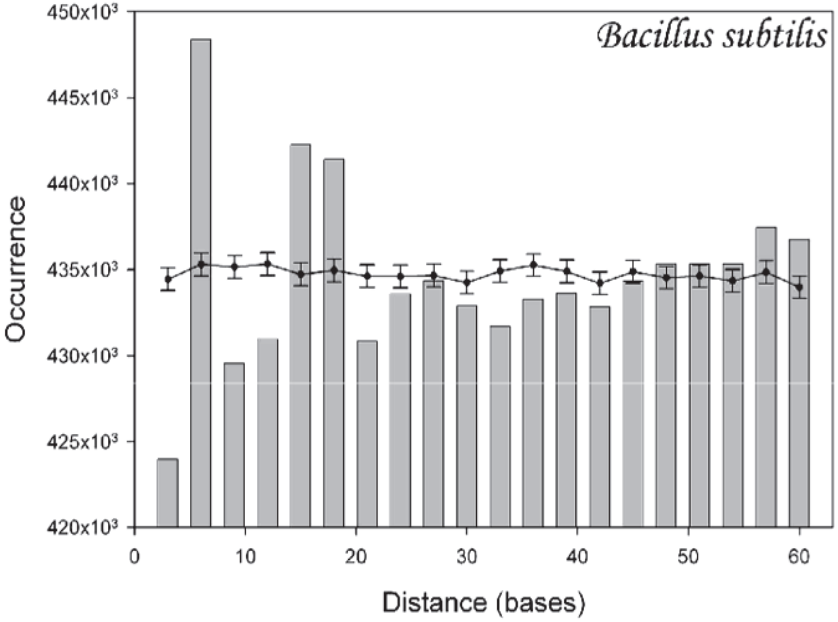
THE SIZE n OF A_n AND G_n UNITS
IS 6 TO 7 RESIDUES

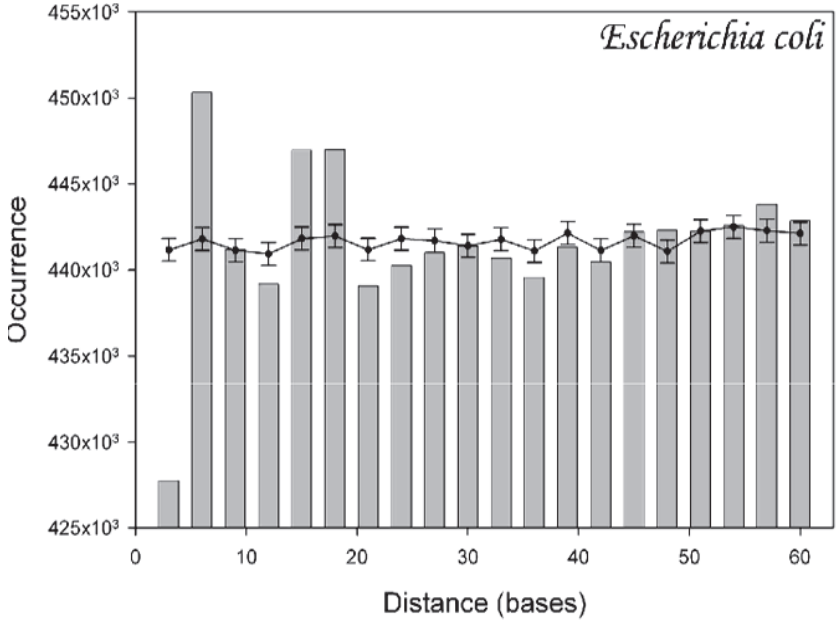
2001
Kizhner V
Kizhner A
Berezovsky I

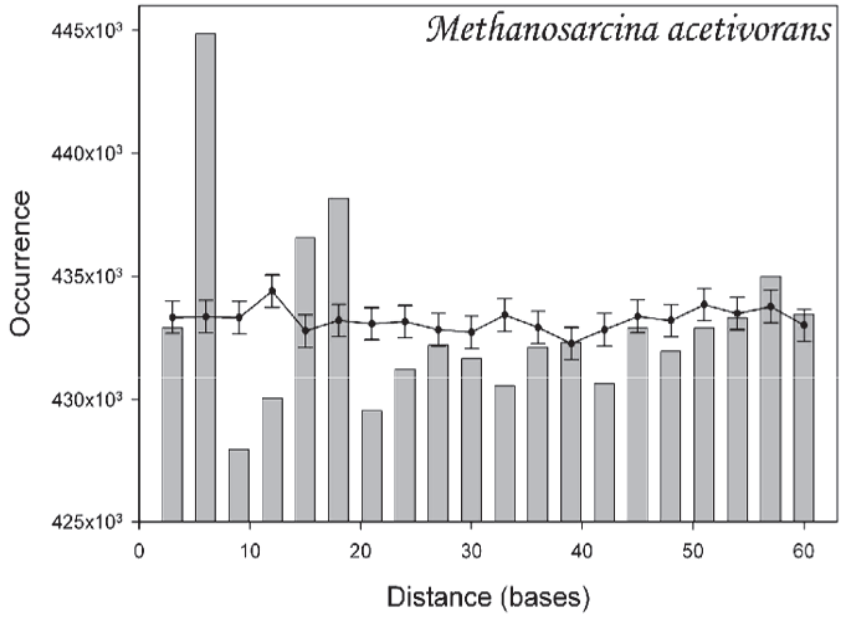
One possible early hairpin

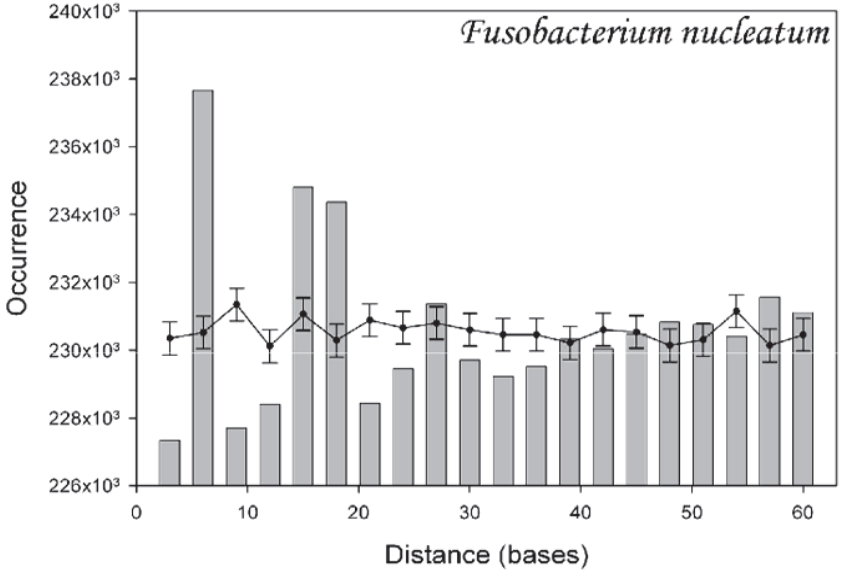


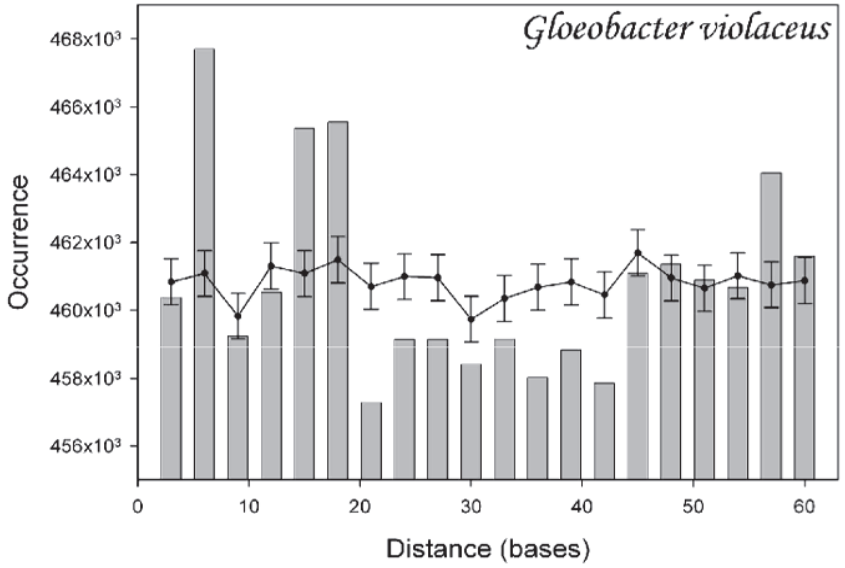

Archaeoglobus fulgidus
Occurrence
244x10³
242x10³
240x10³
238x10³
236x10³
234x10³
232x10³
230x10³
228x10³
226x10³
0
10
20
30
40
50
60
Distance (bases)

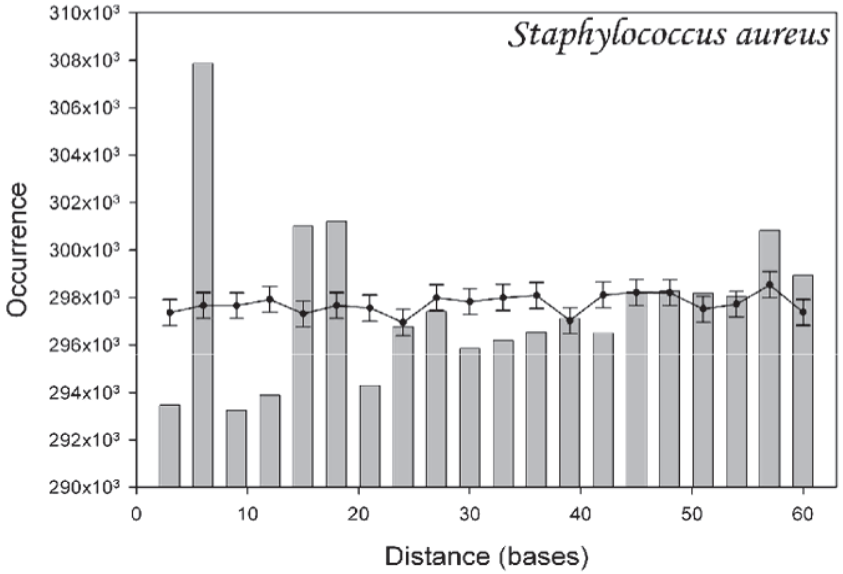

Bacillus subtilis
Occurrence
450x10³
445x10³
440x10³
435x10³
430x10³
425x10³
420x10³
0
10
20
30
40
50
60
Distance (bases)

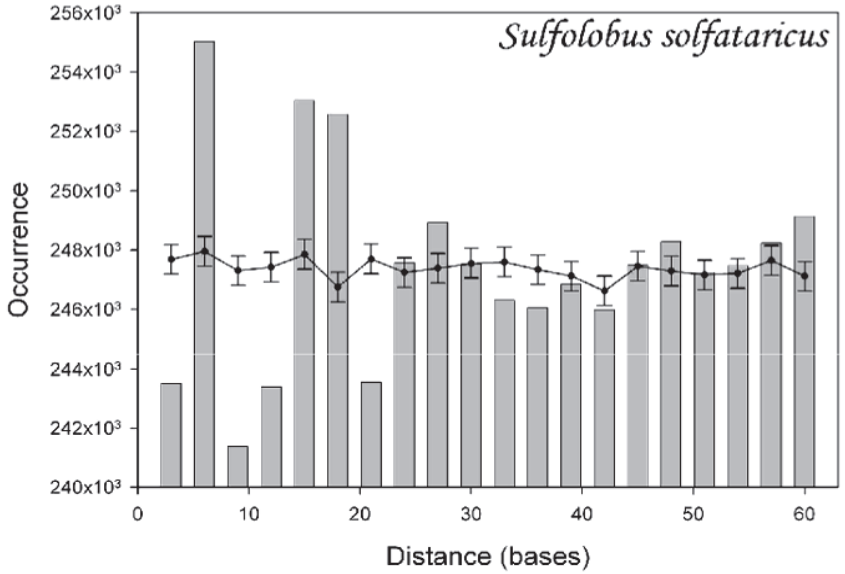

Escherichia coli
Occurrence
455x10³
450x10³
445x10³
440x10³
435x10³
430x10³
425x10³
0
10
20
30
40
50
60
Distance (bases)

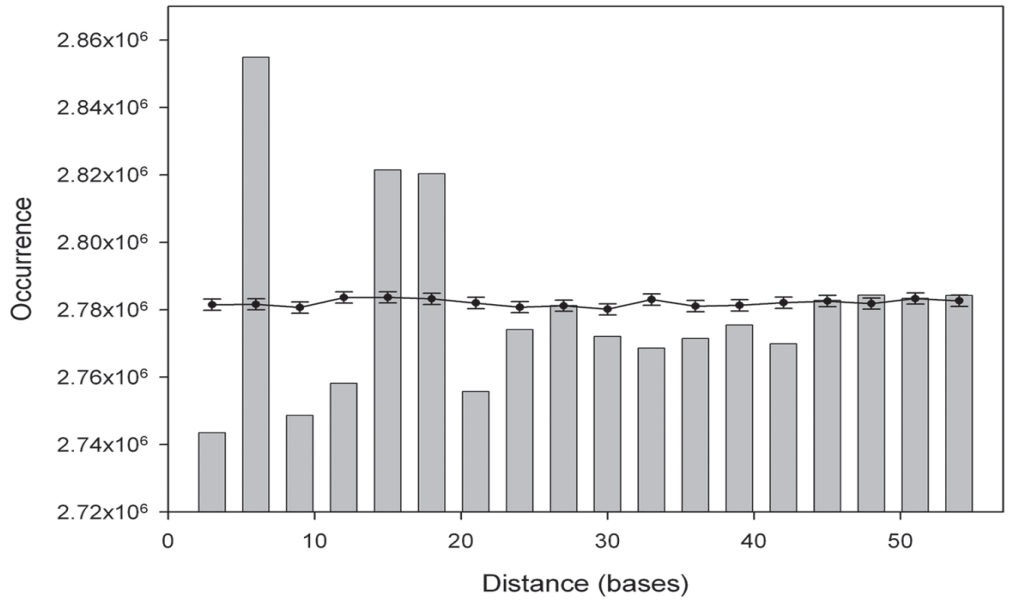

Methanosarcina acetivorans
Occurrence
445x10³
440x10³
435x10³
430x10³
425x10³
0
10
20
30
40
50
60
Distance (bases)

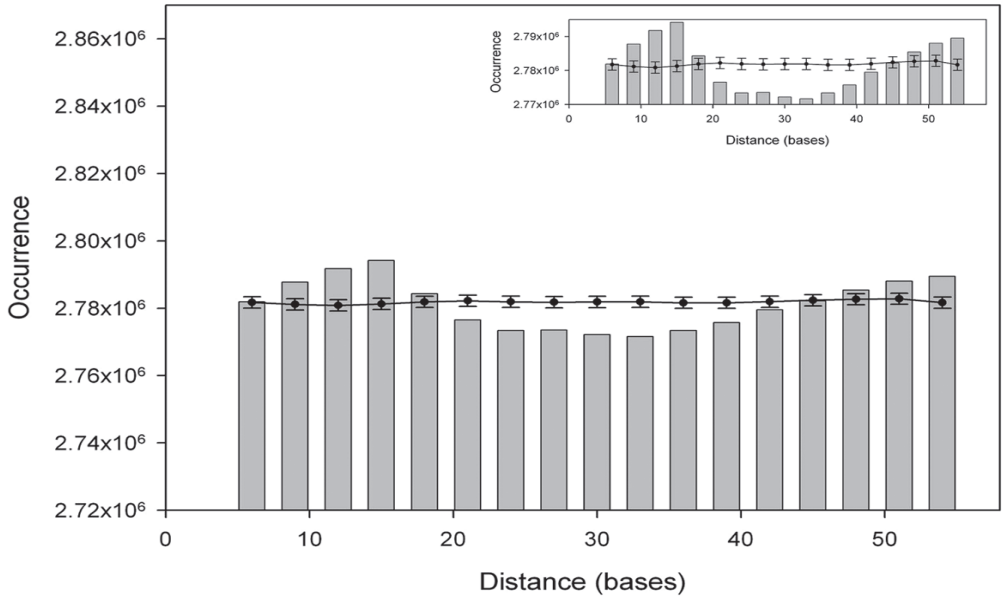

Fusobacterium nucleatum
Occurrence
240x10³
238x10³
236x10³
234x10³
232x10³
230x10³
228x10³
226x10³
0
10
20
30
40
50
60
Distance (bases)


Gloeobacter violaceus
Occurrence
468x10³
466x10³
464x10³
462x10³
460x10³
458x10³
456x10³
0
10
20
30
40
50
60
Distance (bases)


Staphylococcus aureus
Occurrence
310x10³
308x10³
306x10³
304x10³
302x10³
300x10³
298x10³
296x10³
294x10³
292x10³
290x10³
0
10
20
30
40
50
60
Distance (bases)


Sulfolobus solfataricus
Occurrence
256x10³
254x10³
252x10³
250x10³
248x10³
246x10³
244x10³
242x10³
240x10³
0
10
20
30
40
50
60
Distance (bases)


Occurrence
2.86x10⁶
2.84x10⁶
2.82x10⁶
2.80x10⁶
2.78x10⁶
2.76x10⁶
2.74x10⁶
2.72x10⁶
0
10
20
30
40
50
Distance (bases)


Occurrence
2.86x10⁶
2.84x10⁶
2.82x10⁶
2.80x10⁶
2.78x10⁶
2.76x10⁶
2.74x10⁶
2.72x10⁶
0
10
20
30
40
50
Distance (bases)
Occurrence
2.79x10⁶
2.78x10⁶
2.77x10⁶
0
10
20
30
40
50
Distance (bases)

Codon evolution chart as basis of new theory of early evolution:

predictions and confirmations

1. Oldest proteins were glycine-rich. Glycine
2. Alanine- and Glycine-family amino acids. Binary code. Substitutions keep the code.
3. The earliest mini-proteins had the size of 6-7 amino acids.
4. The earliest mini-genes had the size of 18-21 bases.
5. The earliest mRNA were duplexes, coding in both strands.
6. The most conserved protein sequence motifs consist of early amino acids.

Protein modules (closed loops)

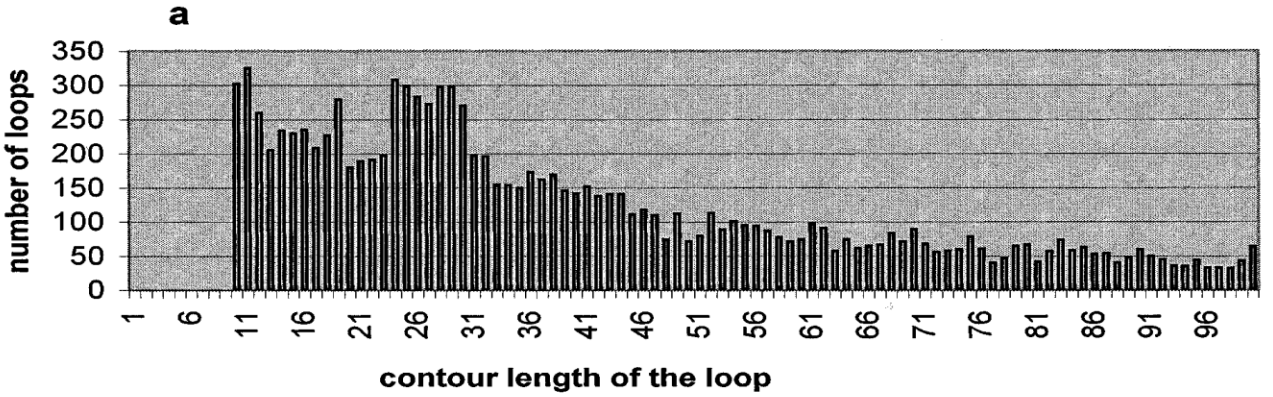
Polymer statistics of polypeptide chains

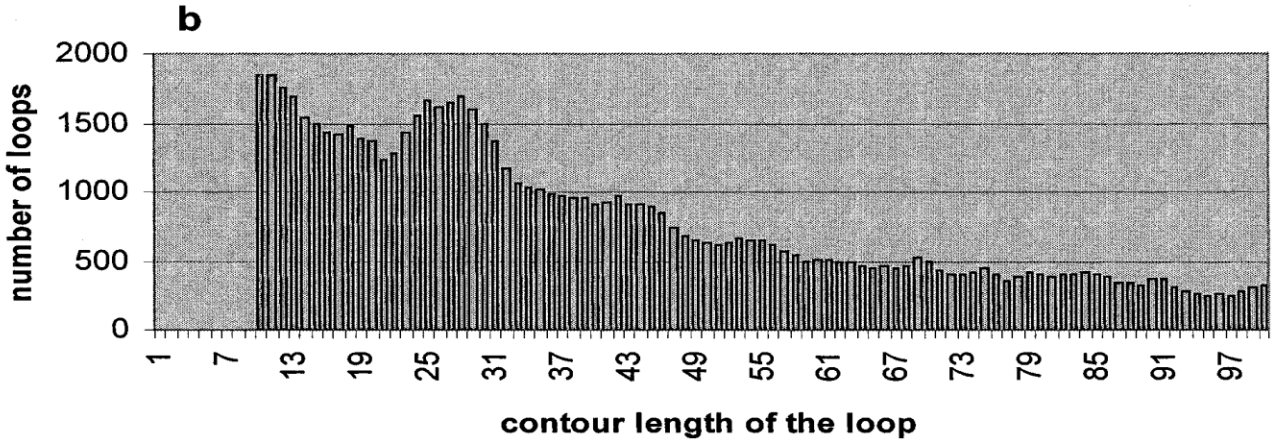
The chain returns to itself
with optimal loop closure size
of 3-4 persistence lengths (Shimada and Yamaka

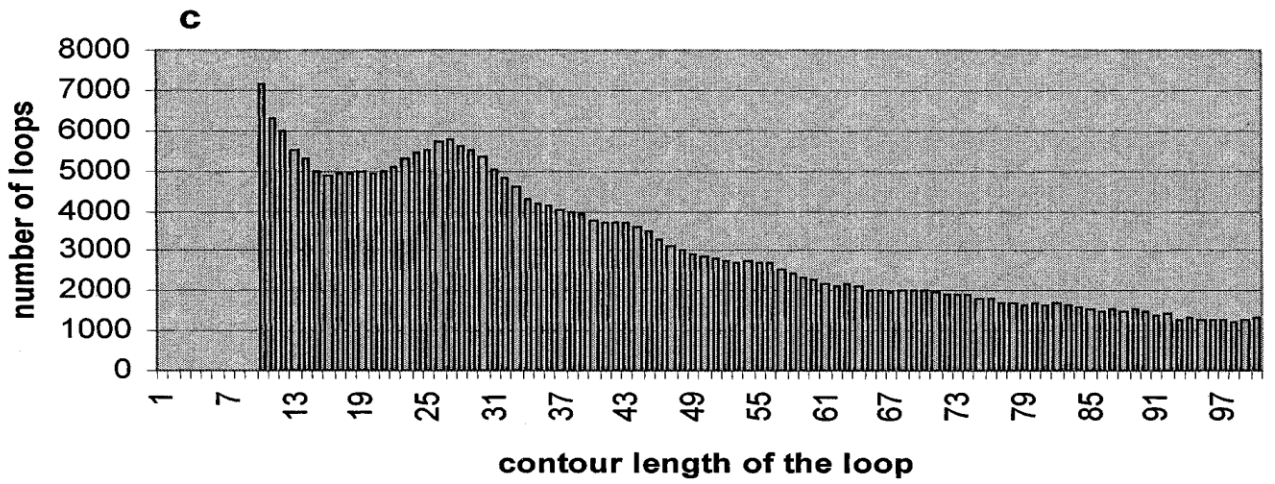
Persistence length of mixed sequence polypepti
is ~5 amino acid residues (Flory).

Natural closed loops are expected to be
15-20 residues (non-structured)

and 25-35 residues long (α-helix containing lo


a
number of loops
350
300
250
200
150
100
50
0
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
contour length of the loop


b
number of loops
2000
1500
1000
500
0
1
7
13
19
25
31
37
43
49
55
61
67
73
79
85
91
97
contour length of the loop


c
number of loops
8000
7000
6000
5000
4000
3000
2000
1000
0
1
7
13
19
25
31
37
43
49
55
61
67
73
79
85
91
97
contour length of the loop

OUT-OF-CONTEXT SEQUENCES I, II and III

```
original seq.  ACC GCU AUA CAG AUG UGU CAU ACC GCC CAU GAC GGC ACU UGC AAU GCA CGU UUA
     I         A   G   A   C   A   U   C   A   G   C   G   G   A   U   A   G   C   U
     II          C   C   U   A   U   G   A   C   C   A   A   G   C   G   A   C   G   U
     III           C   U   A   G   G   U   U   C   C   U   C   C   U   C   U   A   U   A
```

original seq. ACCGCUAUACAGAUGUGUCAUACCG**CCC**AUGACGGCA**CUU**GCAAUGCACG**UUU**A

I AGACAUCAGCGGAUAGCU

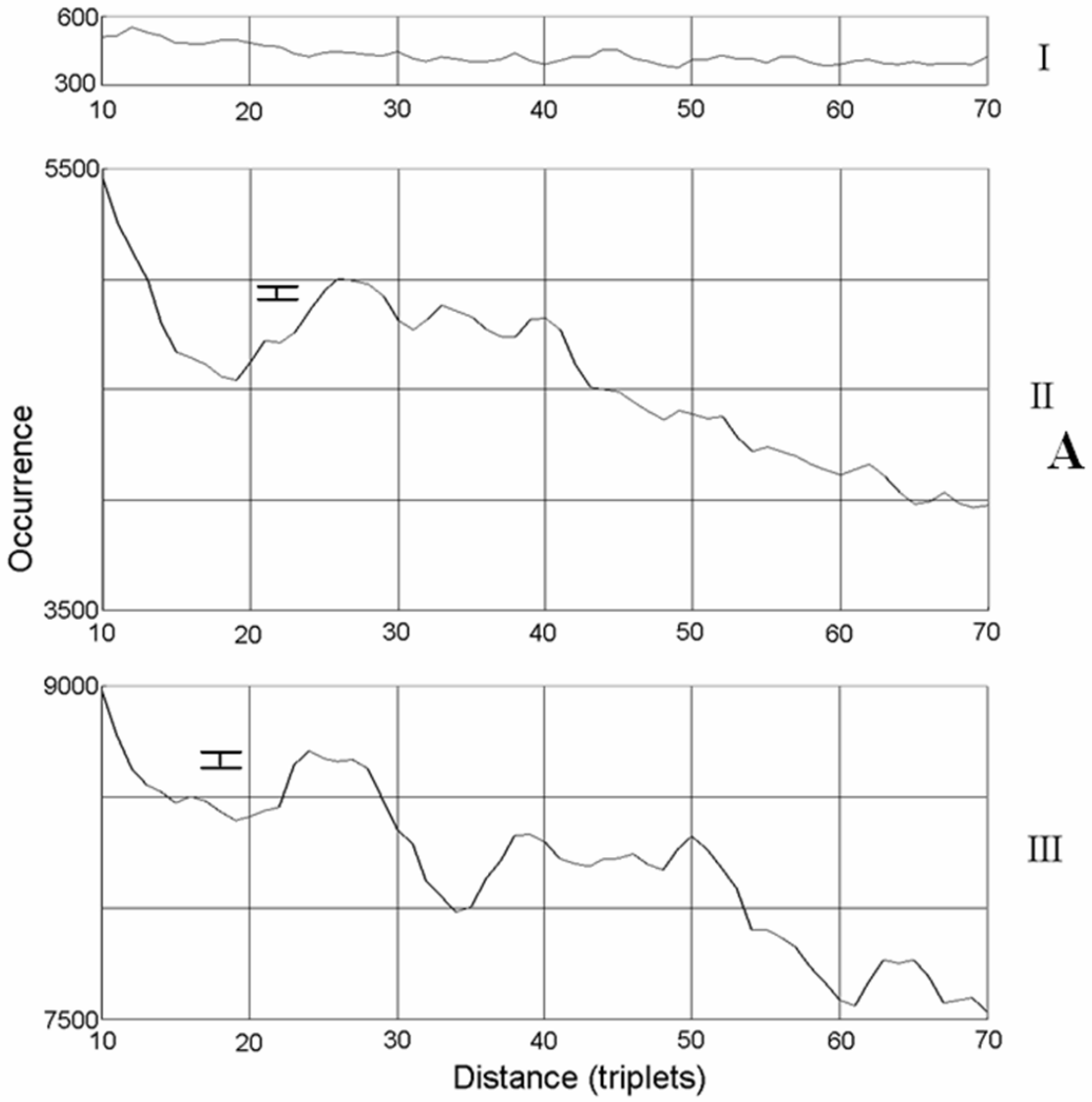
II **CCU**AUGACCAAGCGACGU

III CUAGG**UUCCUCCUCU**AUA

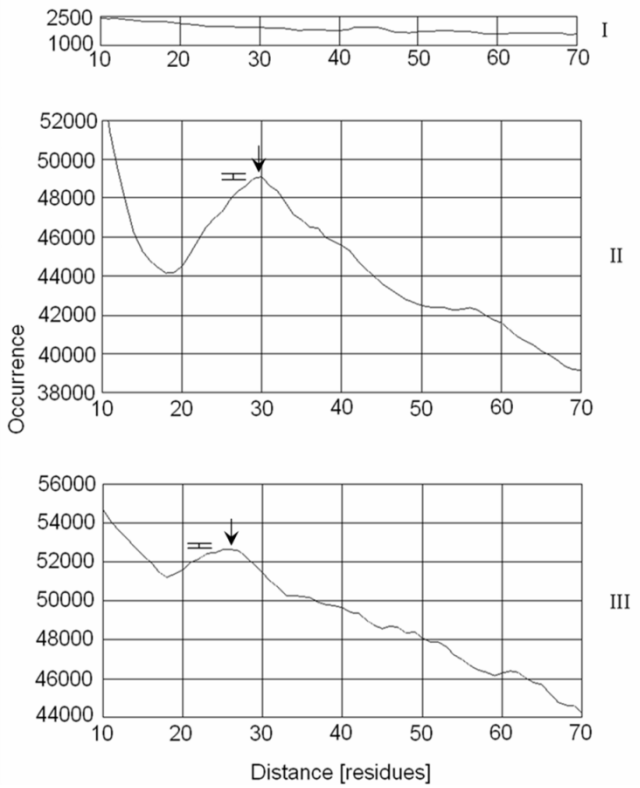
A. Rapoport, 2008

Pyrimidine clusters in different codon positions. The highest

	Position I			Position II			Position III		
	Natural	Random	Ratio	Natural	Random	Ratio	Natural	Random	Ratio
	Bradyrhizobium japonicum								
Y_5	29757	26041	1.14	157363	146121	1.08	214525	150012	1.43
Y_6	12846	10460	1.23	95764	83157	1.15	135458	84731	1.6
Y_7	5616	4213	1.33	60556	47624	1.27	85807	47918	1.79
Y_8	2499	1700	1.47	39758	27455	1.45	54740	27139	**2.02**
Y_9	1166	687	1.7	26915	15938	1.69	35100	15396	**2.28**
	Chromobacterium violaceum								
Y_5	22413	18361	1.22	70680	62766	1.13	104311	60872	1.71
Y_6	10443	7910	1.32	41858	34333	1.22	65390	33047	**1.98**
Y_7	4894	3431	1.43	25831	18923	1.37	41265	18046	**2.29**
Y_8	2358	1498	1.57	16602	10514	1.58	26237	9918	**2.65**
Y_9	1207	658	1.84	10904	5891	1.85	16775	5488	**3.06**
	Thermotoga maritima								
Y_5	3285	2783	1.18	26752	23210	1.15	20941	15676	1.34
Y_6	1246	992	1.26	16412	12540	1.31	10960	7656	1.43
Y_7	470	358	1.31	10659	6862	1.55	5755	3751	1.53
Y_8	177	131	1.35	7329	3806	1.93	3105	1843	1.68
Y_9	61	48	1.27	5216	2139	**2.44**	1688	909	1.86
	Methanosarcina acetivorans								
Y_5	9255	8316	1.11	61310	54328	1.13	60914	56666	1.07
Y_6	3780	3143	1.2	36752	29118	1.26	33395	30070	1.11
Y_7	1676	1221	1.37	23284	15797	1.47	18493	16031	1.15
Y_8	846	490	1.72	15559	8682	1.79	10343	8592	1.2
Y_9	444	204	**2.18**	10759	4837	**2.22**	5806	4634	1.25
	Sulfolobus sulfataricus								
Y_5	6380	4193	1.52	43090	36761	1.17	21356	18400	1.16
Y_6	2783	1529	1.82	26790	20511	1.31	10867	8693	1.25
Y_7	1220	568	**2.15**	17416	11632	1.5	5553	4130	1.34
Y_8	556	214	**2.6**	11810	6704	1.76	2834	1974	1.44
Y_9	250	81	**3.1**	8212	3922	**2.09**	1457	949	1.53


600
300
10
20
30
40
50
60
70
I
5500
3500
H
Occurrence
II
A
9000
7500
H
III
Distance (triplets)

pyrimidines of 2-nd and 3-rd codon positions cluster at distance 25-30 triplets



Levinthal paradox:

$t = n^L \cdot \tau = 3^{150} \cdot 10^{-12}\ s = 10^{48}\ yrs$

(L = 150 residues)

Solution:

$t = n^L \cdot \tau = 3^{23\ \text{to}\ 31} \cdot 10^{-12}\ s = 0.1$ to 1000 sec

(L = 23 to 31 residues)

Berezovsky, ENT, 2002

Hullabaloo around Levinthal

Berezovsky, I. N., Trifonov, E. N., Loop fold structure of proteins: Resolution of Levinthal's paradox, J. Biomolec. Str. Dyn. 20, 5-6 (2002)

Finkelstein A. V., Cunning simplicity of a hierarchical folding,
J. Biomolec. Str. Dyn. 20, **311-313** (2002)

Berezovsky, I. N., Trifonov, E. N., Back to units of protein folding, J. Biomolec. Str. Dyn. 20, **31**

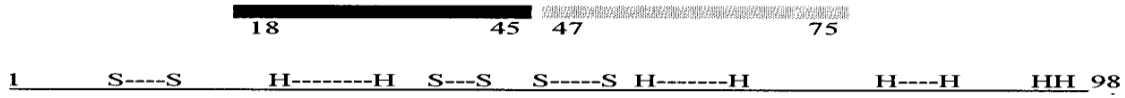
Grosberg, A., A few disconnected notes related to Levinthal paradox, J. Biomolec. Str. Dyn. 20, **317**

Kloczkowski, A., Jernigan, R. L., Loop folds in proteins and evolutionary conservation of folding n
J. Biomolec. Str. Dyn. 20, **323-325** (2002)

Rooman M., Dehouck, Y., Kwasigroch, J. M., Biot, C., Gilis, D.,
What is paradoxical about Levinthal paradox?
J. Biomolec. Str. Dyn. 20, **327-329** (2002)

Fernandez, A., Belinky, A., de las Mercedes Boland, M., Protein folding: where is the paradox?
J. Biomolec. Str. Dyn. 20, **331-332** (2002)

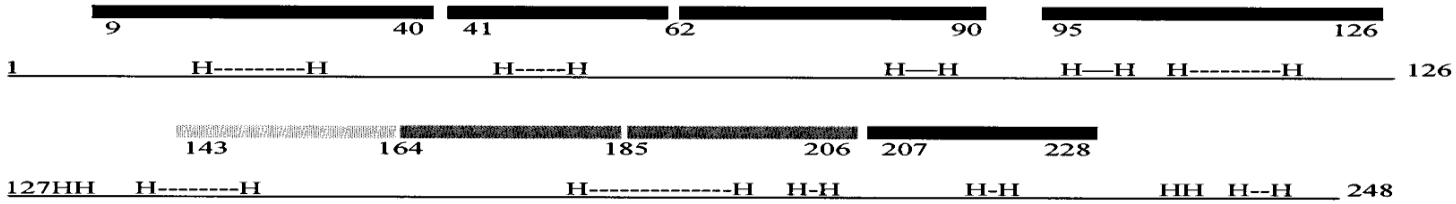
α/β Sandwich (1aps):



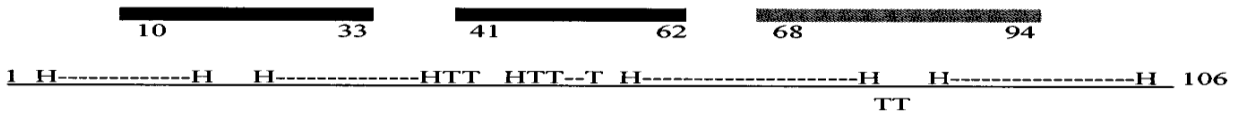
Trefoll (1i1b):

Globin (1thb):

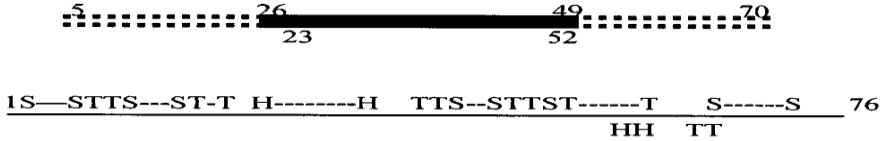
TIM barrell (7tim):



Up-down (256b):

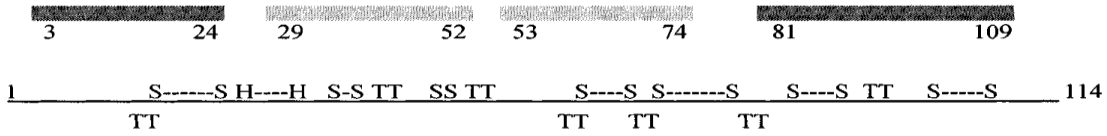


UB α/β roll (1ubq):

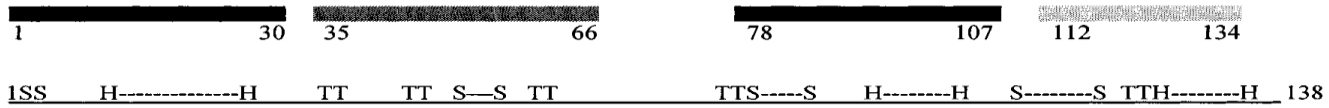


Jelly roll (2stv):

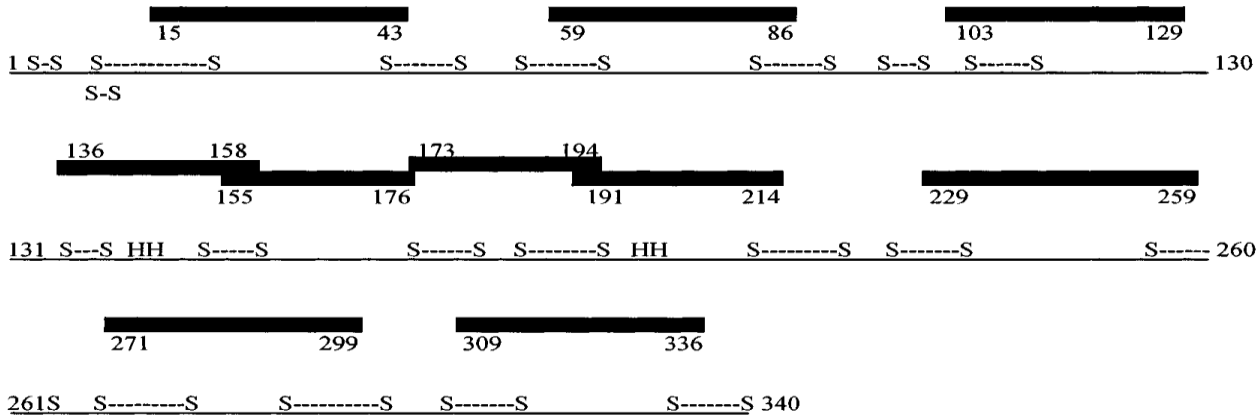
Immunoglobulin folds (2rhe):



Doubly Wound (4fxn):



Matrix Porin Outer Membrane Protein F (2omf):



A

4tim, 9-40 (32 residues)

B

1bnh, 26-53 (28 residues)

C

2omf, 156-175 (20 residues)

D

1kap, 352-370 (19 residues)

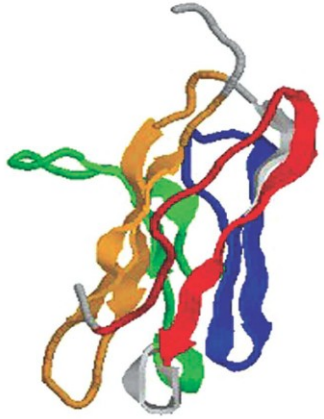
E

1tsp, 167-190 (24 residues)

F

1thj, 128-146(19 residues)

A
Immunoglobulin fold (2rhe)



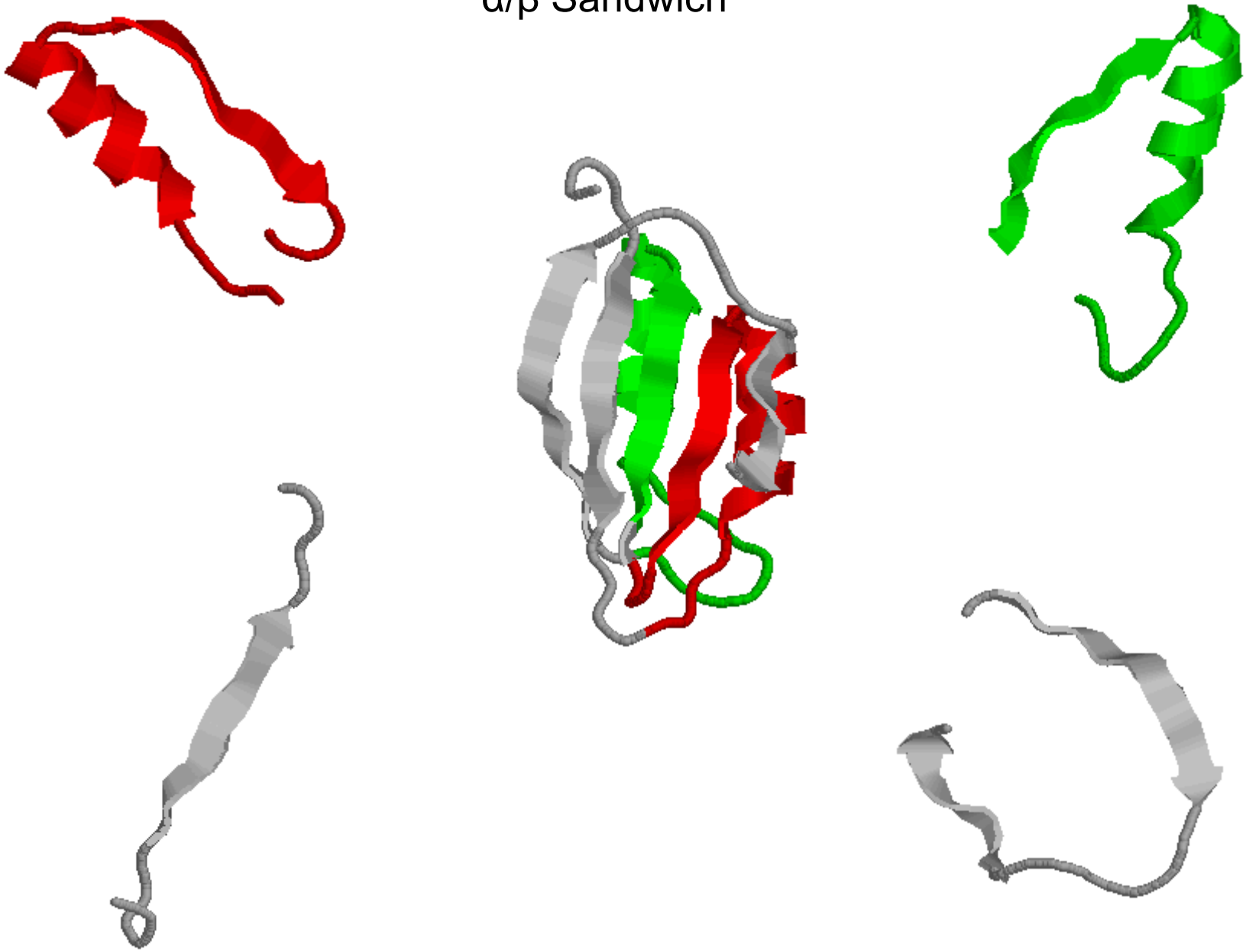
B
loop 3-24

C
loop 29-52

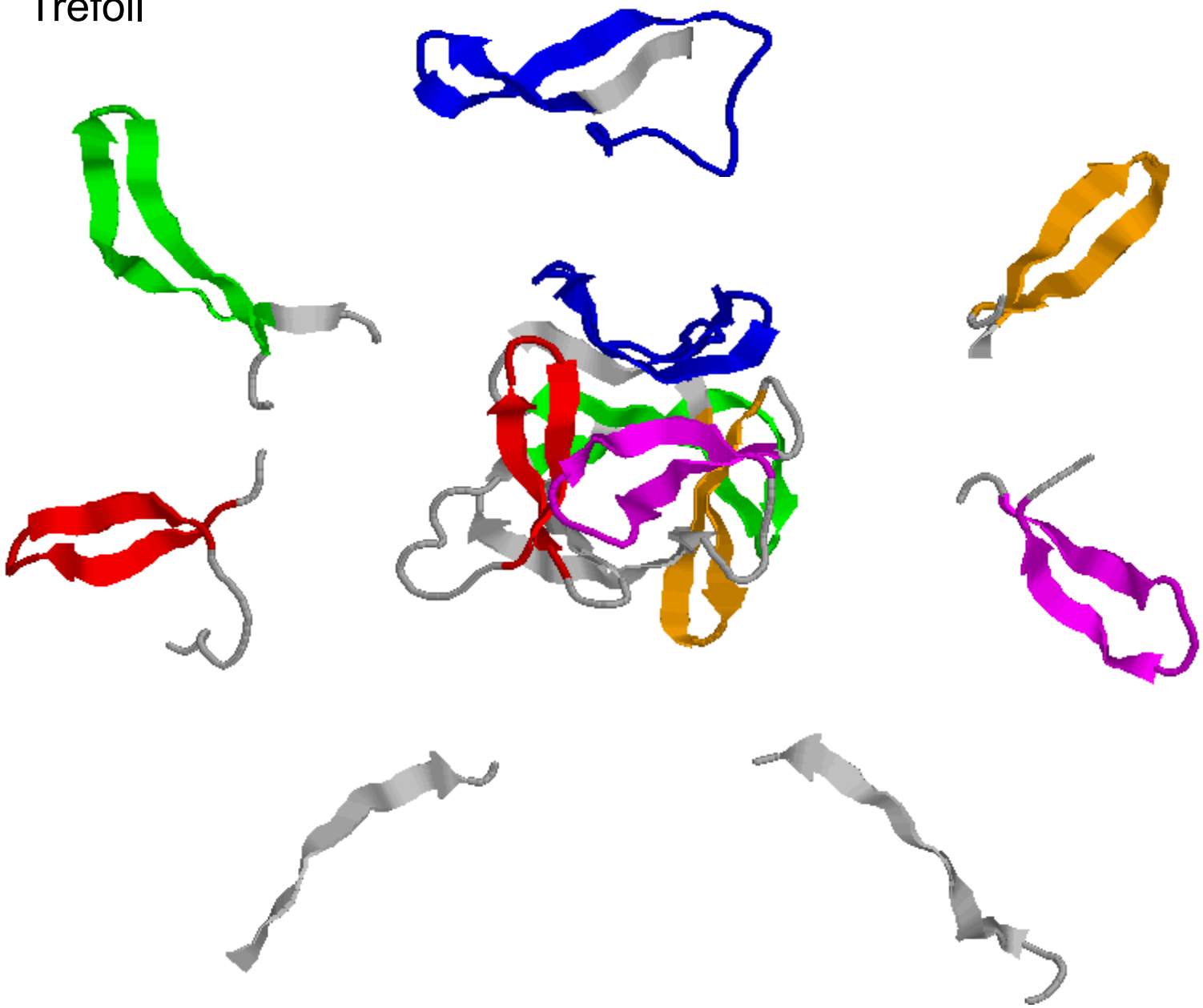
D
loop 53-74

E
loop 81-109

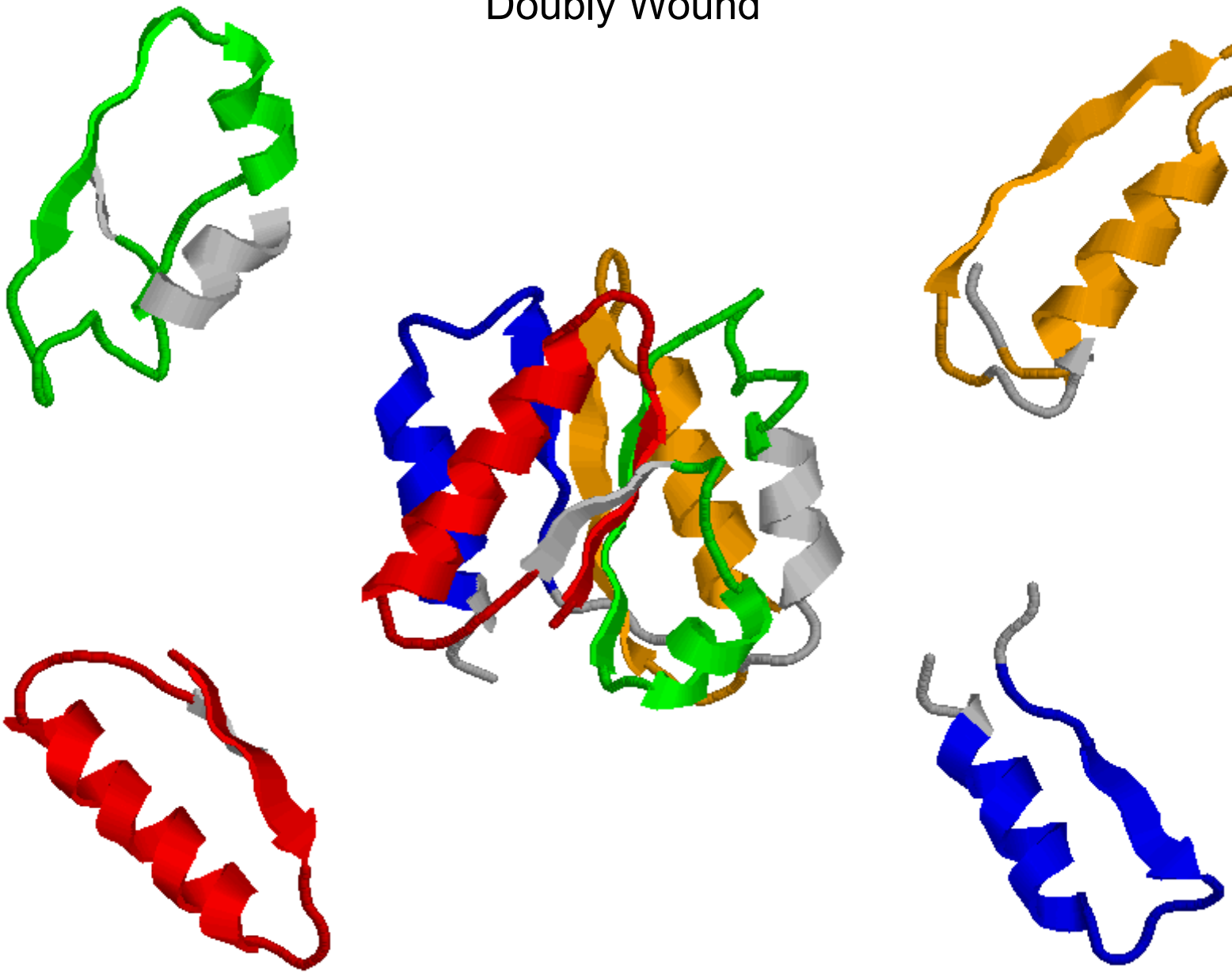
α/β Sandwich



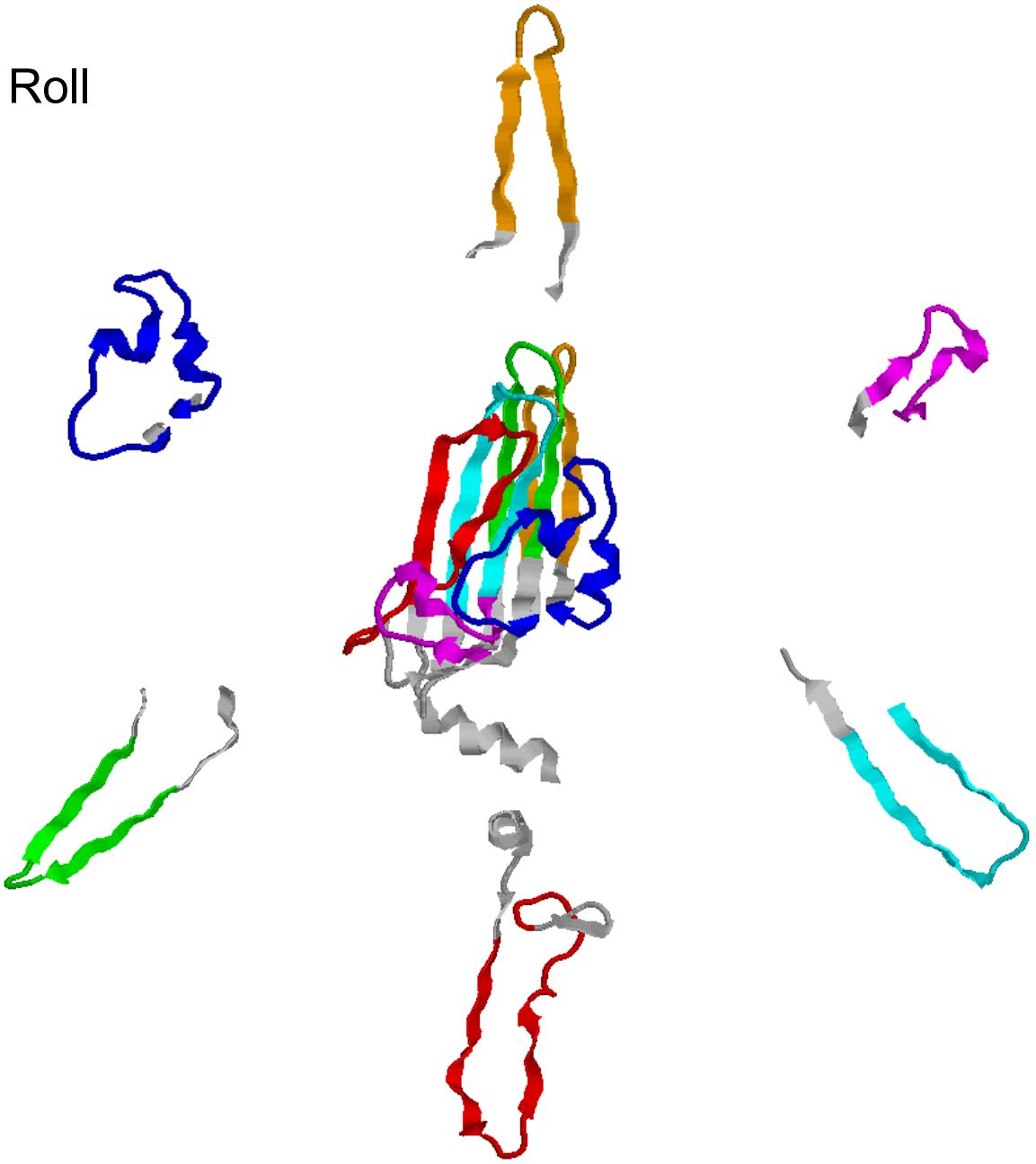
Trefoil



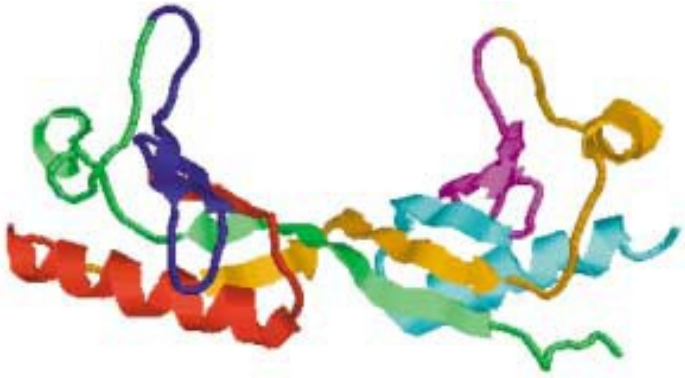
Doubly Wound

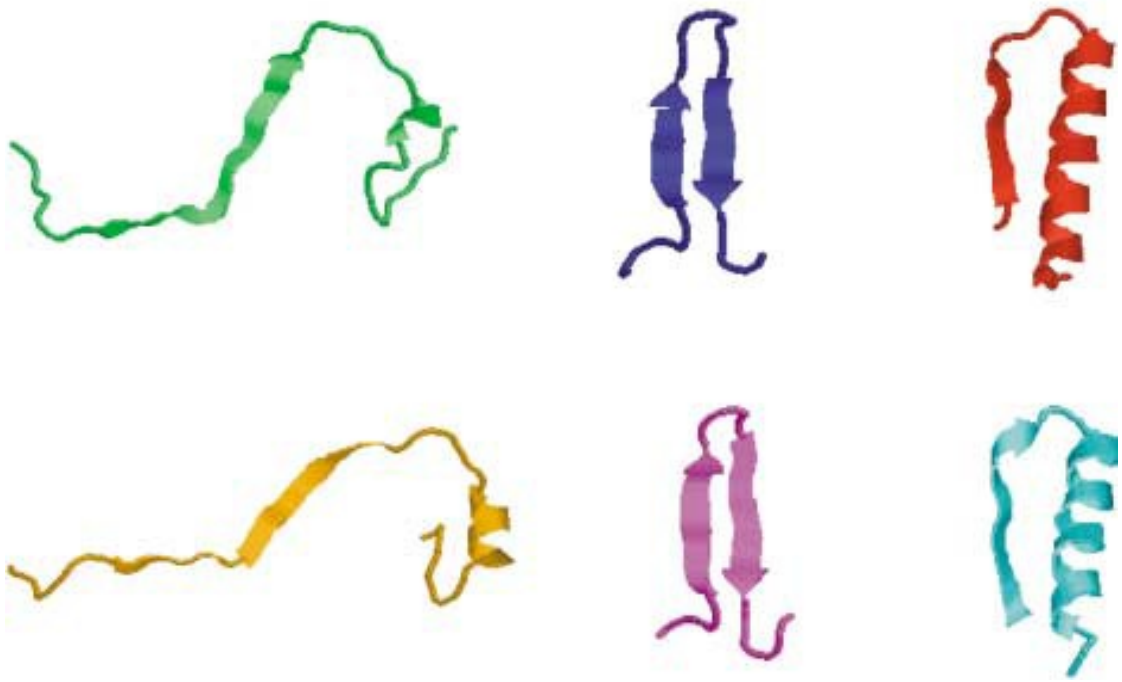


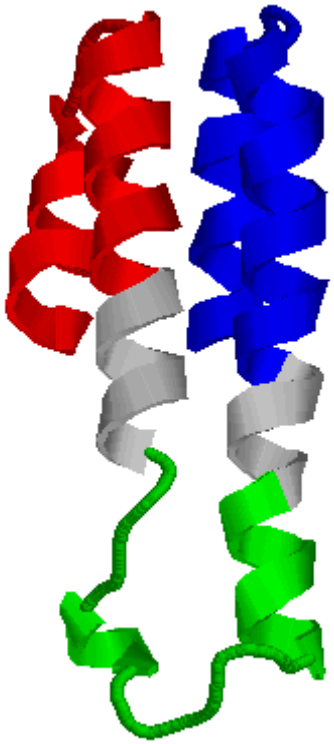
Jelly Roll



TATA binding protein





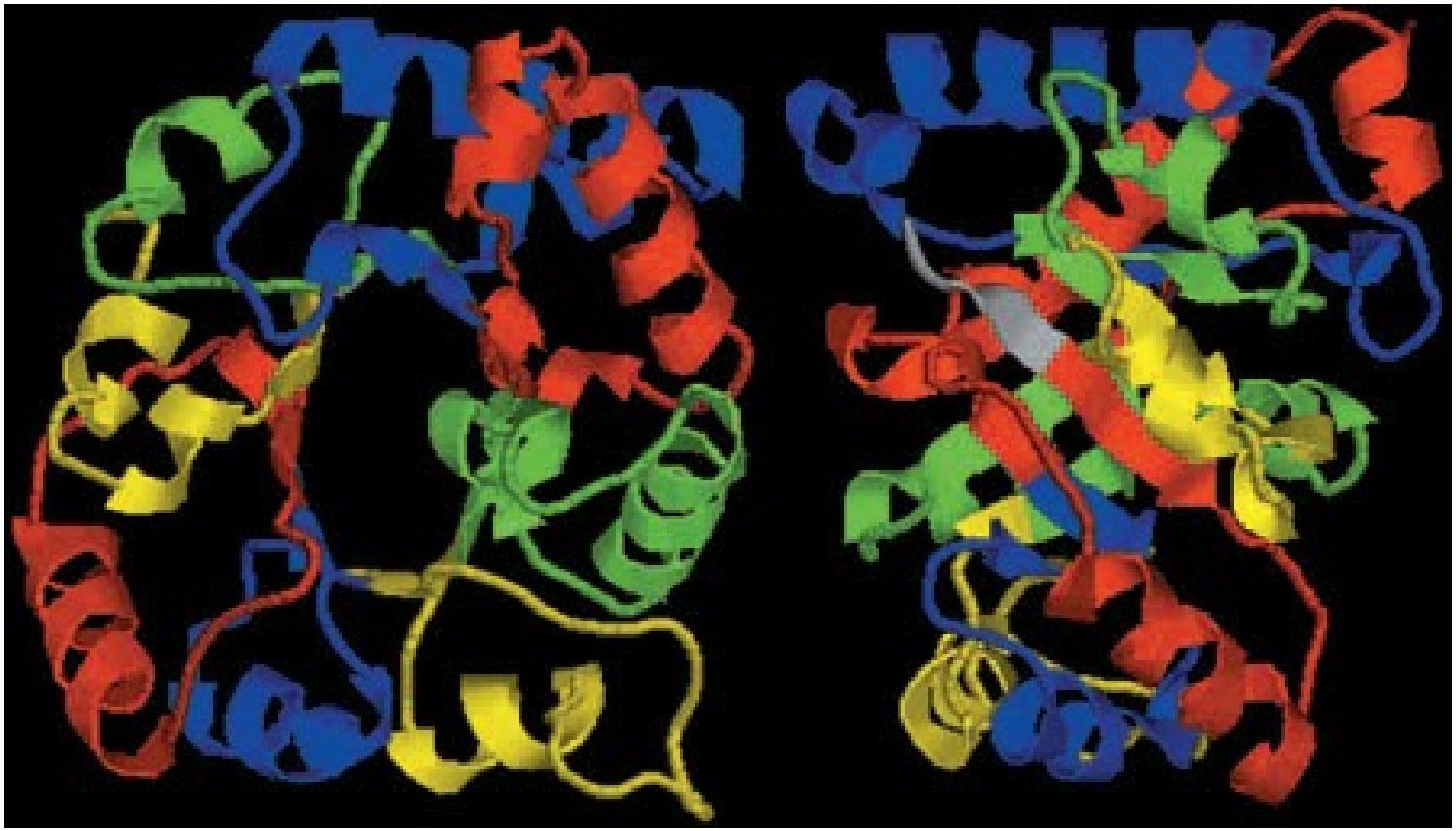


Cytochrome C

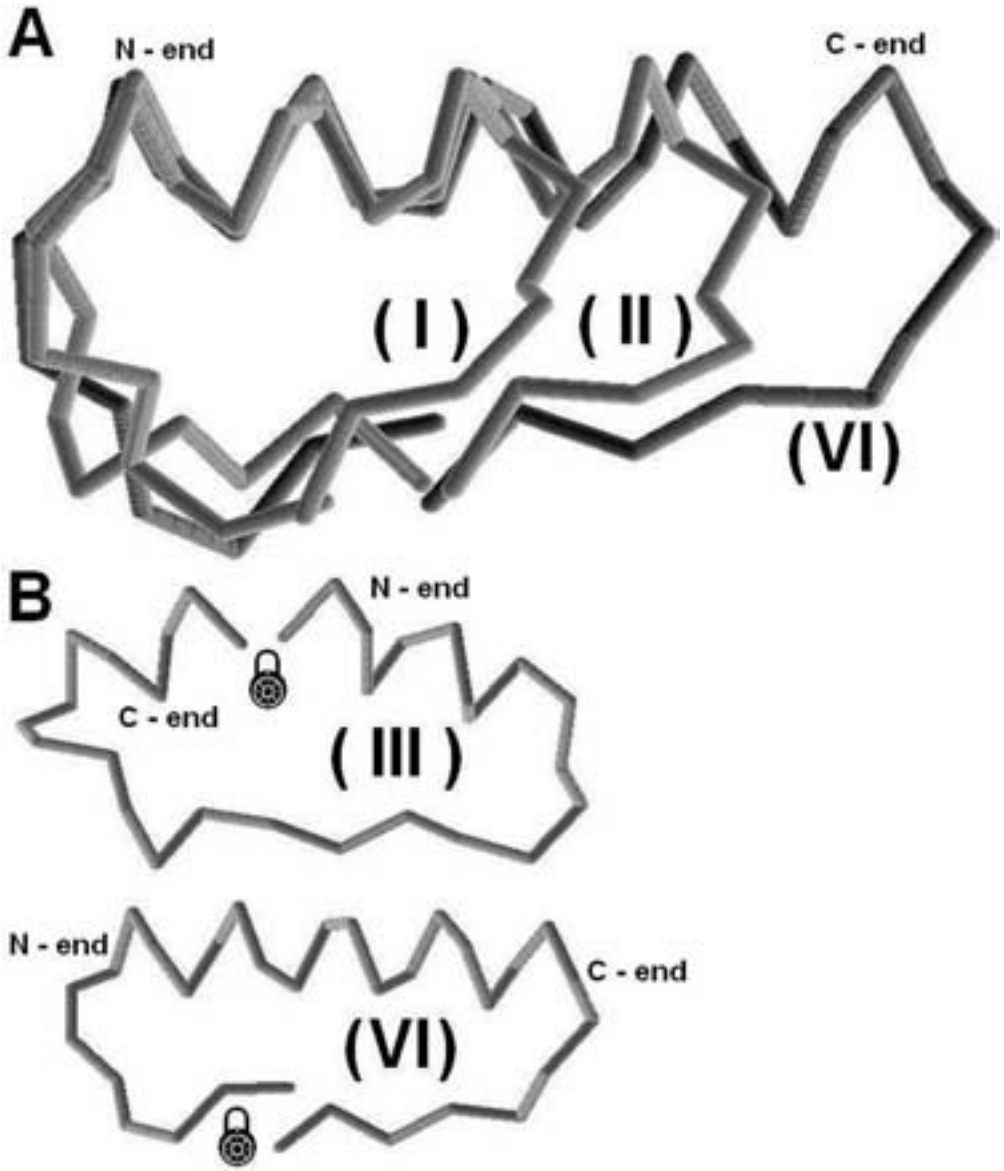
Cytochrome 256b

Cytochrome C

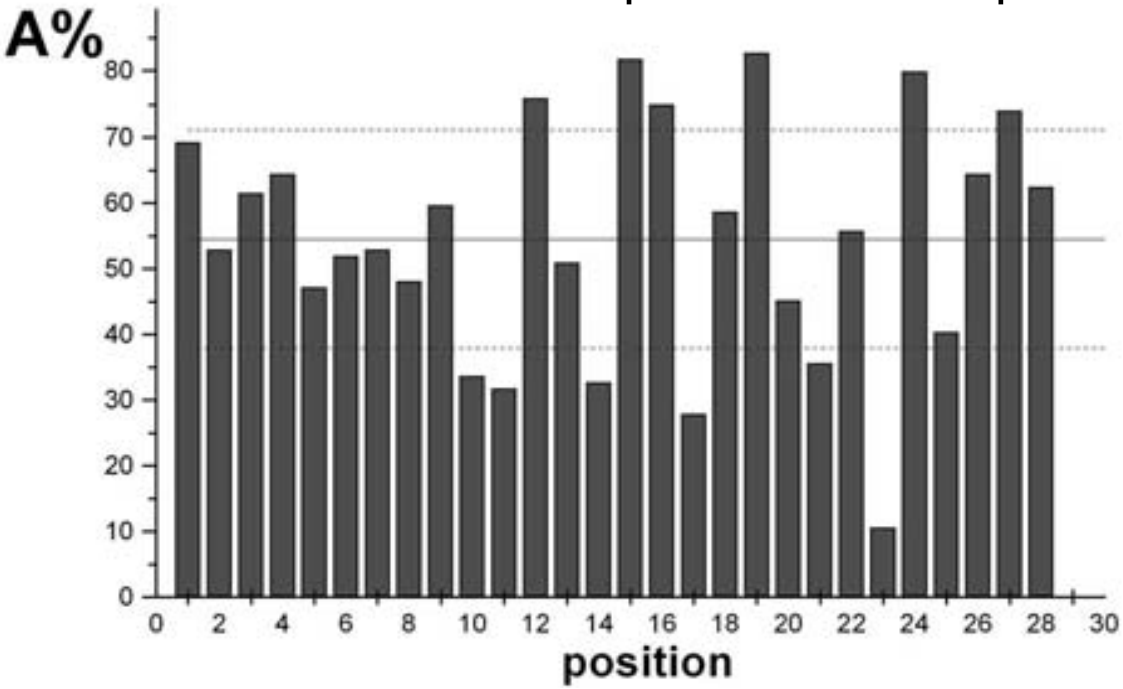
Cytochrome 256b



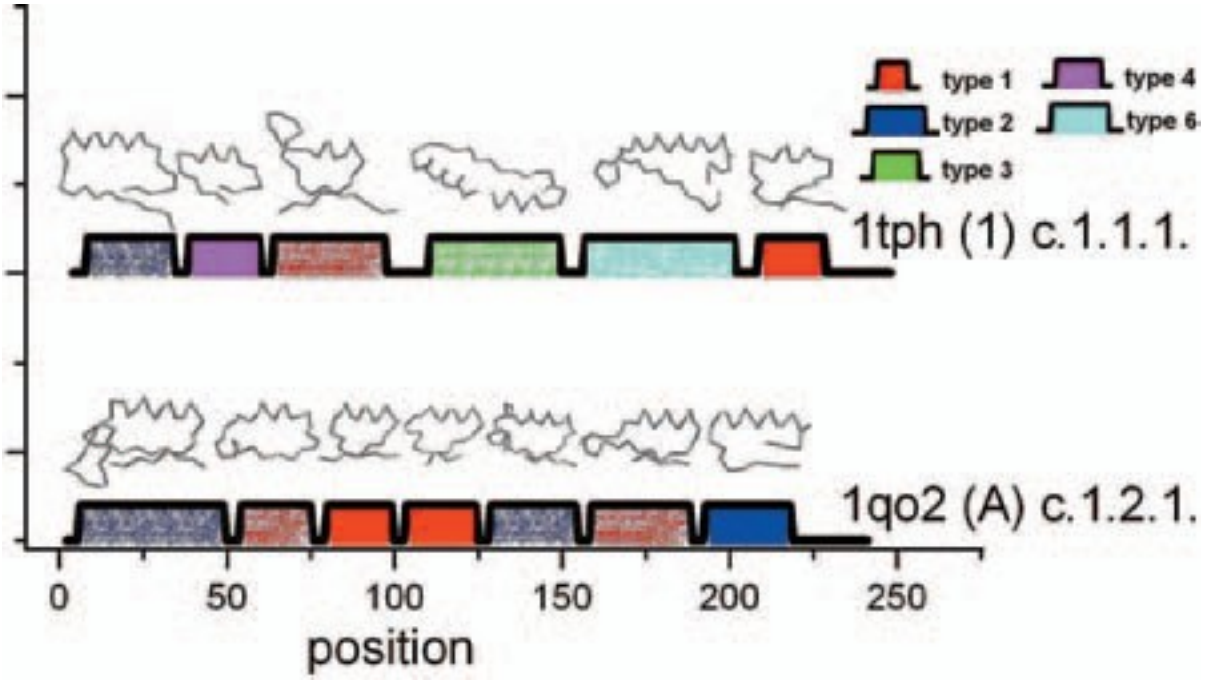
TIM barrell protein

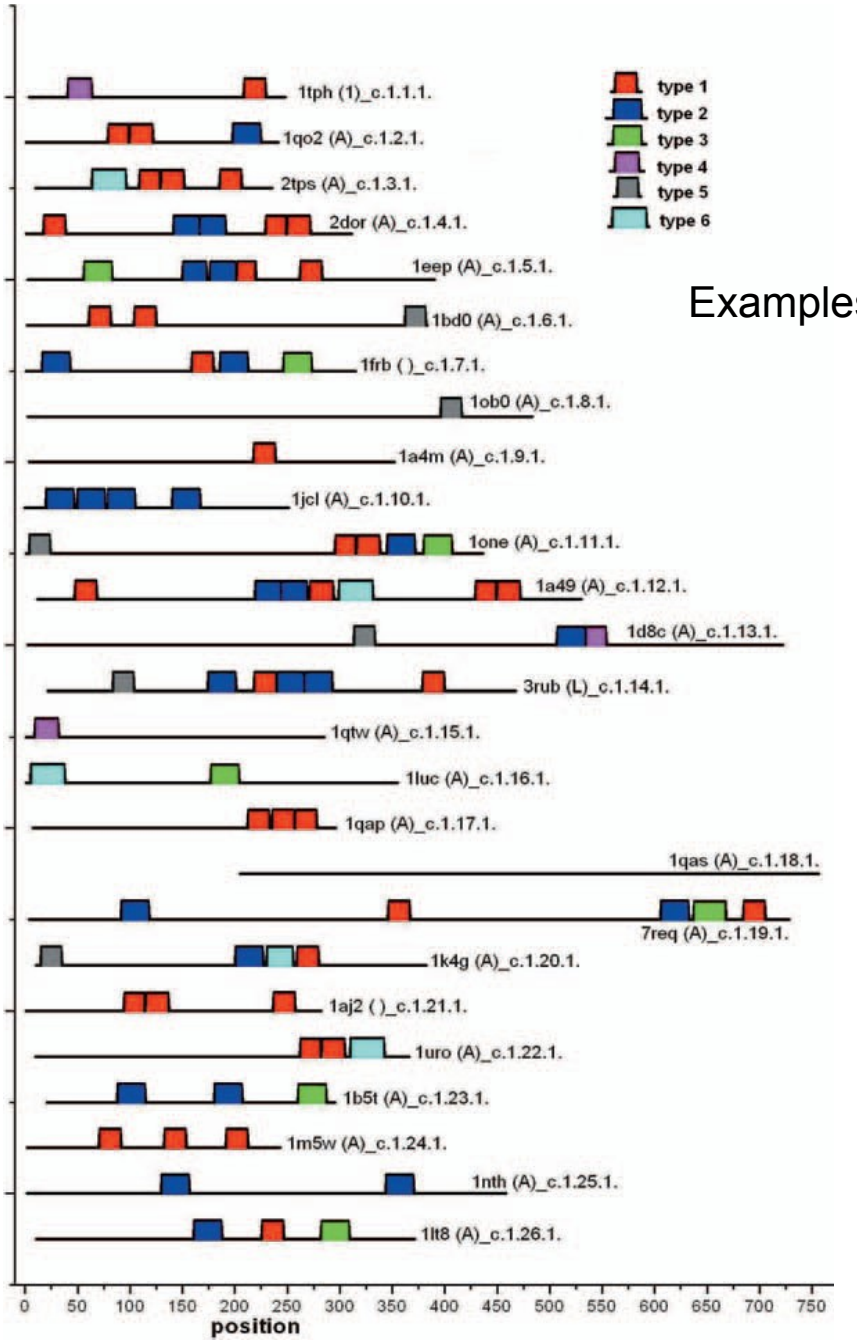

A
N - end
C - end
(I)
(II)
(VI)
B
N - end
C - end
(III)
N - end
C - end
(VI)

Generic closed loop of TIM barrell proteins

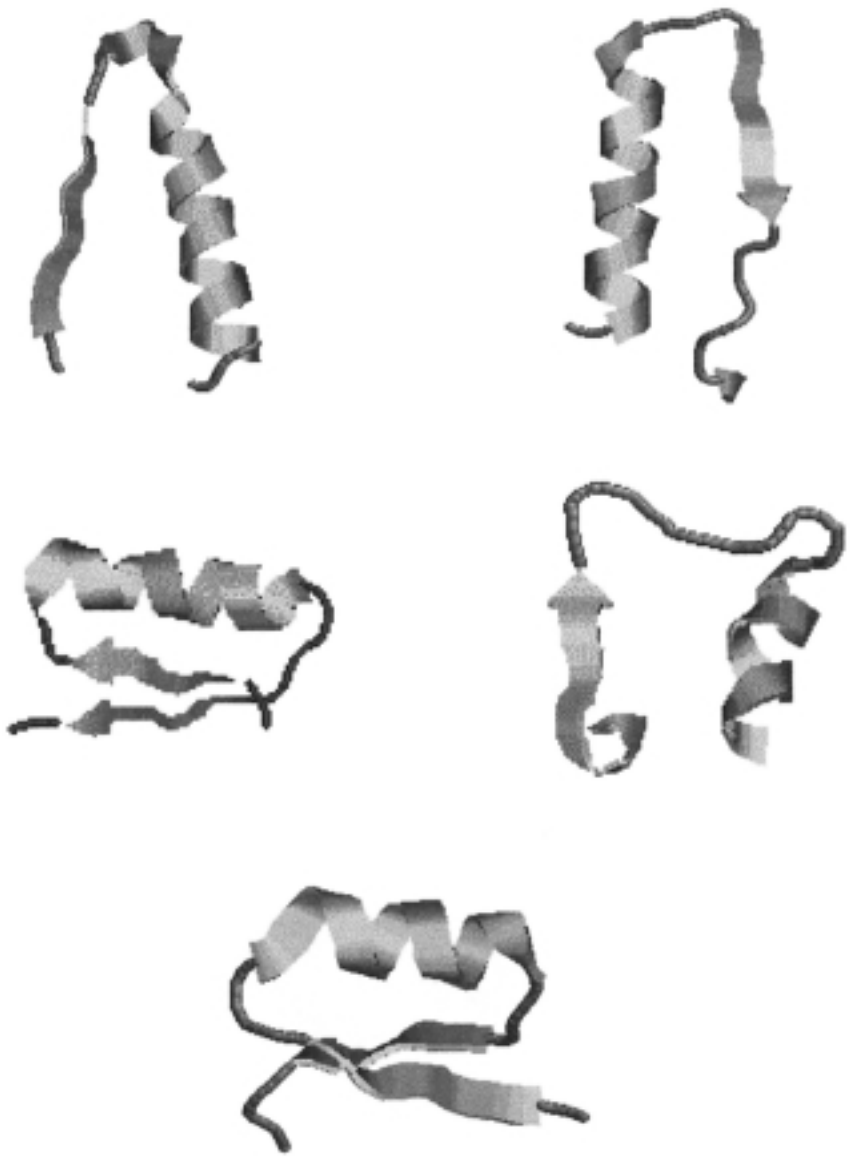


ILLLGIGSPEEVRELARAAKEAGADALI


type 1
type 2
type 3
type 4
type 6
1tph (1) c.1.1.1.
1qo2 (A) c.1.2.1.
0
50
100
150
200
250
position



Examples of TIM barrell proteins



First five presumably ancient sequence prototypes identified

(previous Figure)

```
Aleph   GEIVALVGPSGSGKSTLLRALAGLLKPTSG
Beth    LSGGQRQRVAIARALALEPKLLLLDEPTSALD
Gimel   DVIVVGAGPAGLAAALVLARAGAKVLVIE
Dalet   RRGIGMVFQNYALFPHLTVLENVALGL
Heh     PVIILTARDDEEDRVEGLELGADDYLTKPF
```

Histidine permease

Aleph

Dalet

Beth

Vav

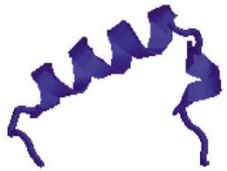
Zayin

Prototype IV

Dalet

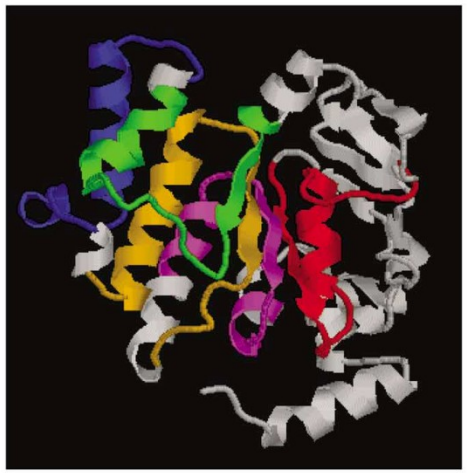
Prototype VI

Vav



Prototype I

Aleph



Prototype II

Beth

Prototype VII

Zayin

1b0u (16)

1d8y_A (12)

1f3o_A (11)

1g29_1 (11)

1pii (10)

1hqc_A (10)

1eld_A (9)

Vav in PDB crystals

Zayin in PDB crystals

1g29_1 (14)

1f3o_A (13)

1ion_A (10)

1b0u_A (8)

Seven prototypes

Aleph	**GEIVALVGPSGSGKSTLLRALAGLLKPDGG**
Beth	**LSGGQRQRVAIARALALEPKLLLLDEPTSALD**
Gimel	**DVIVVGAGPAGLAAALVLARAGAKVLVIE**
Dalet	**RRRIGMVFQNYALFPHLTVLENVALGL**
Heh	**PVIILTARDDEEDRVEGLELGADDYLTKPF**
Vav	**VLGLSKEEARERALKLLAKVGLDERADGKP**
Zayin	**LLKKLQKELGLTILLVTHDLGEA**

1 Aleph  30	2 Beth  32	3 Gimel  29
4 Dalet  27	5 Heh  30	6 Vav  30
7 Zayin  23	8 	9 
10 	11 	12 

ALEPH

BETH

GIMEL

DALET

HEH

VAV

ZAYIN

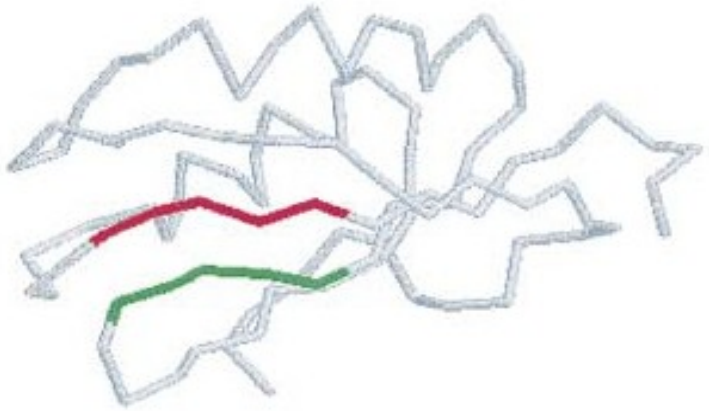
THE EARLIEST STEPS OF LIFE

0. Heptapeptides GGGGGGG and AAAAAAA encoded in RNA duplexes of 21 bp.

1. "Complementary" heptapeptides of Gly- and Ala- alphabets. Some encoded by hairpins.

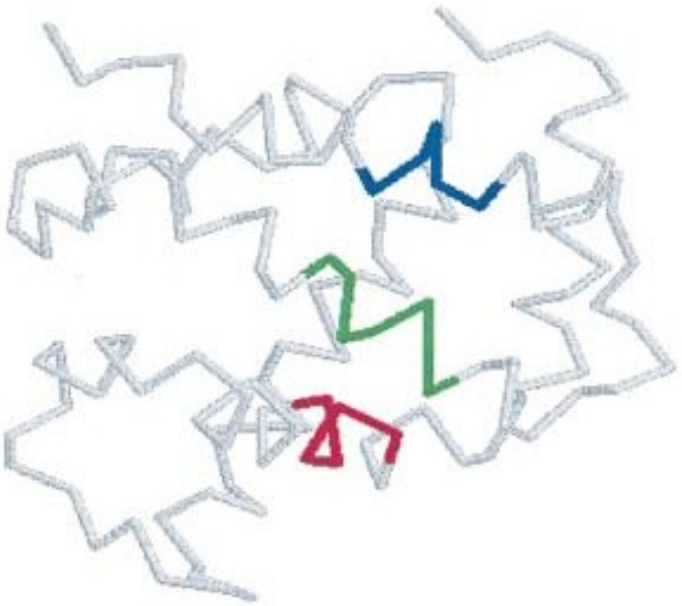
2. The peptides fuse in closed loops of ~28 aa, by end-ligation of the alternating minigenes for all-Gly- and all-Ala-fragments.

3. The closed loops develop in standard sequence/structure/function prototype modules.

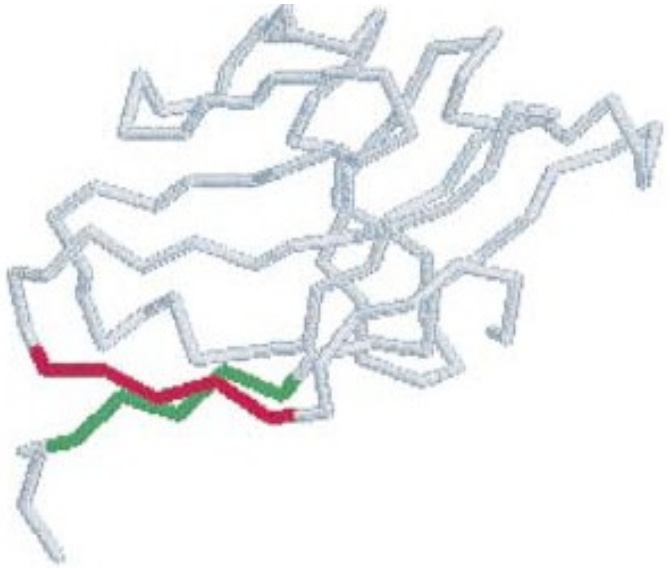


(a)

(b)

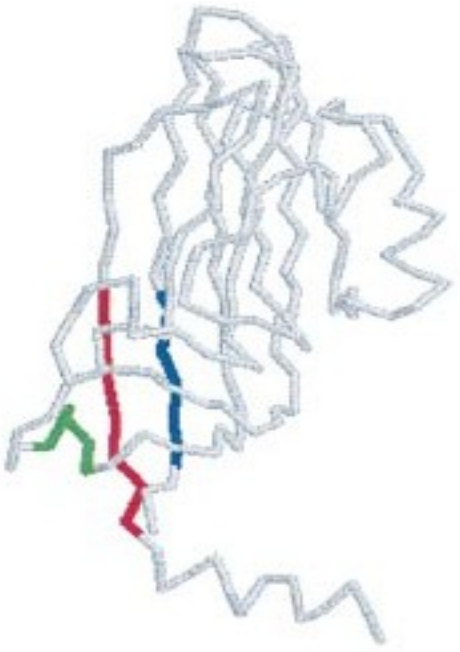


(c)



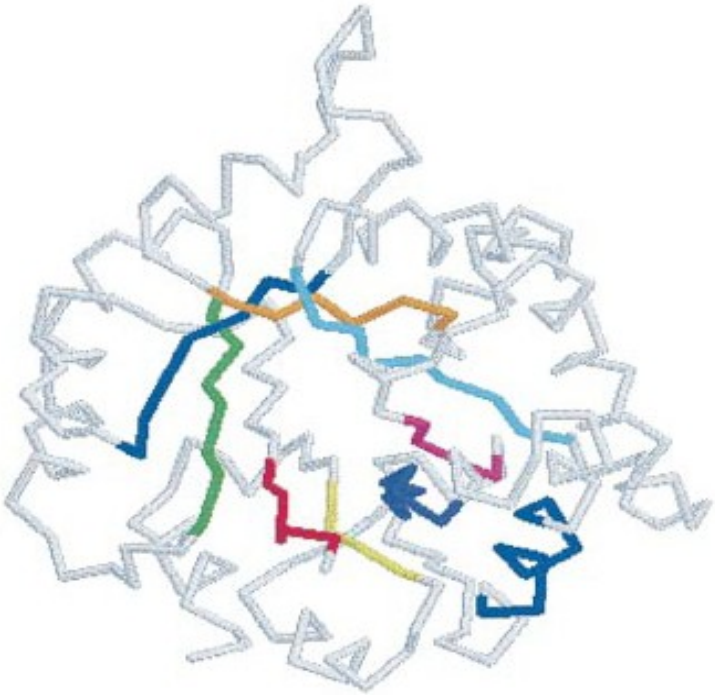
(d)

(e)

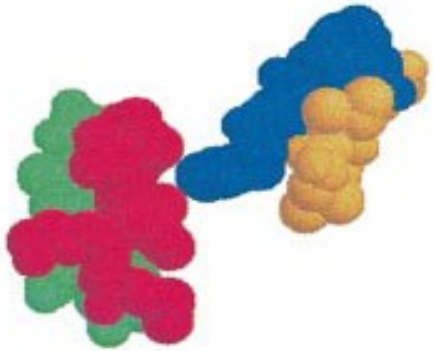


(f)

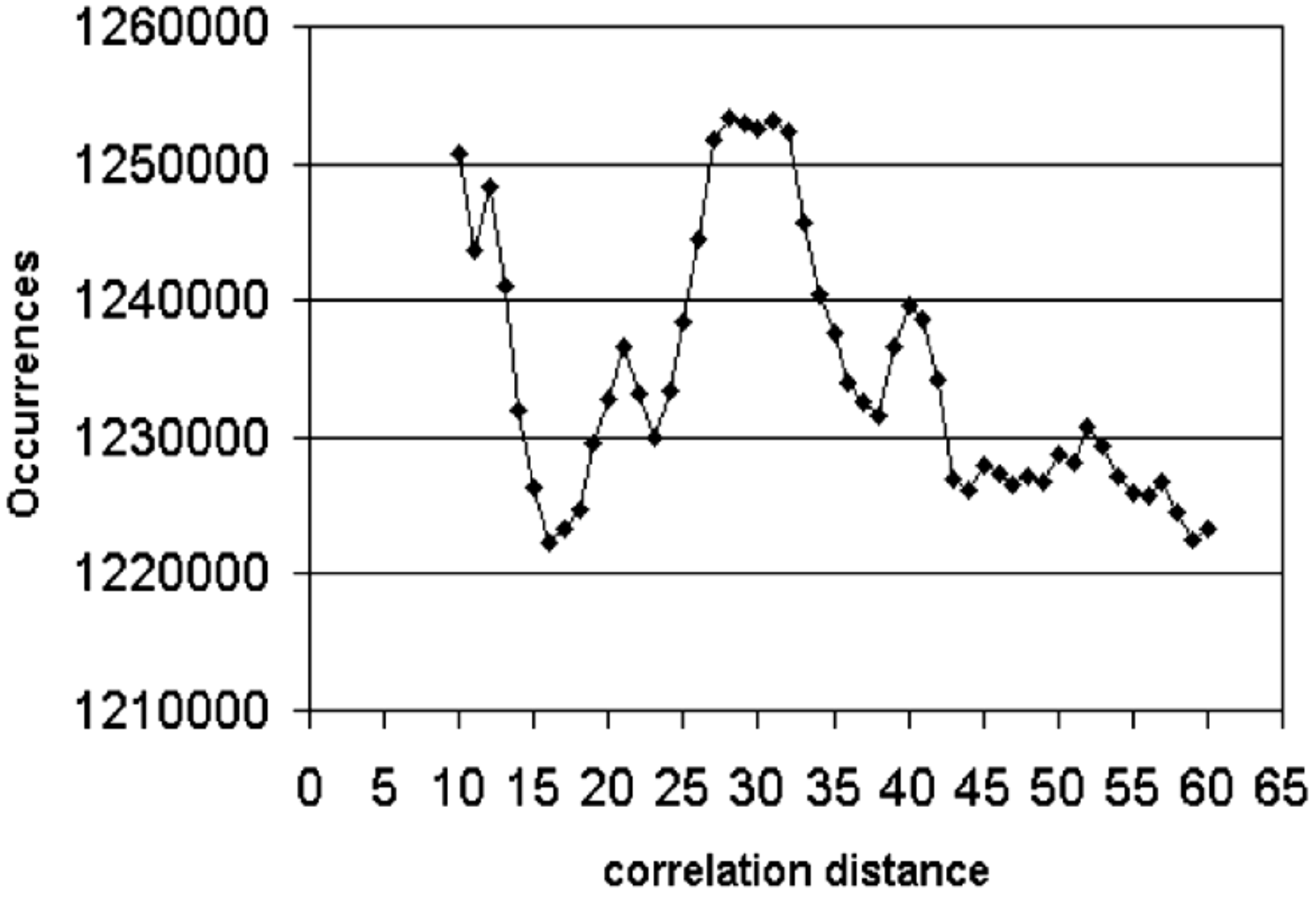
(g)

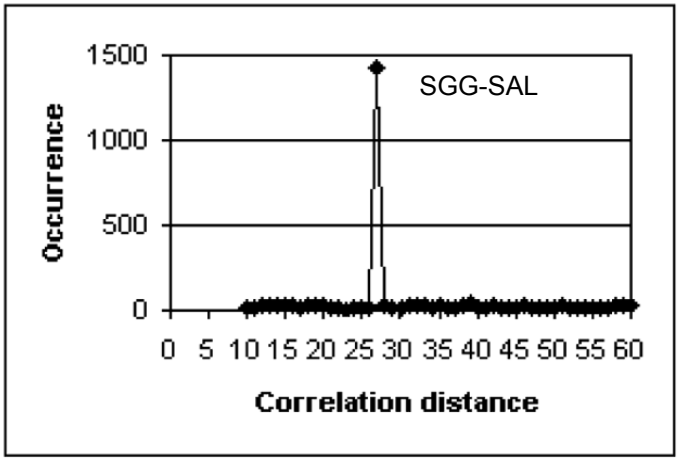


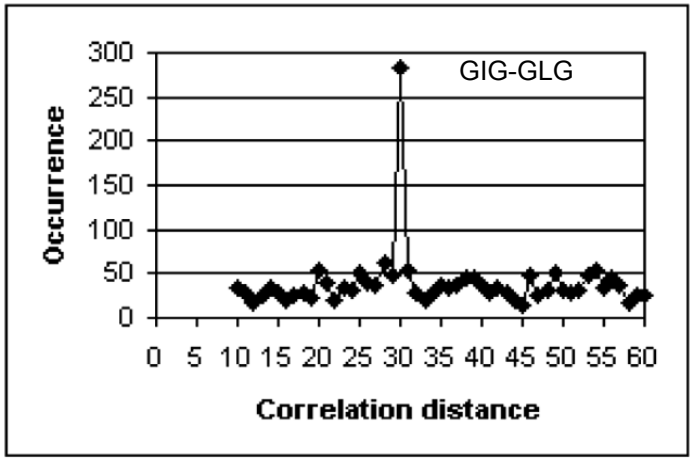
(h)

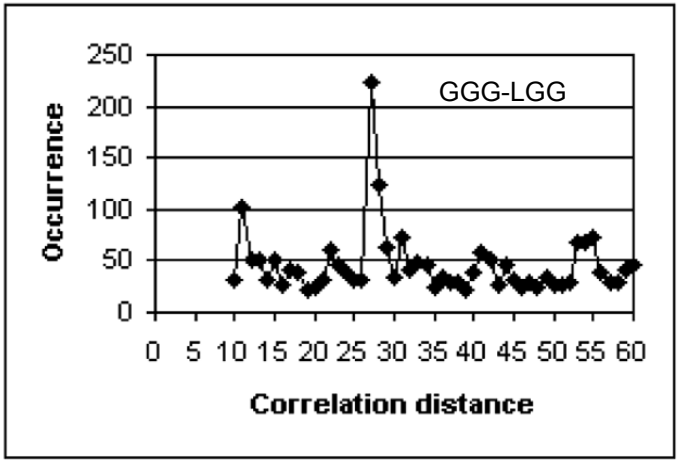


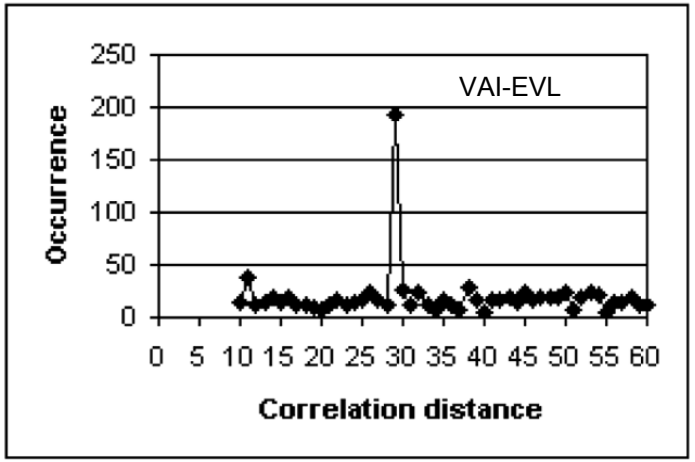
Preferred distance between hydrophobic triplets

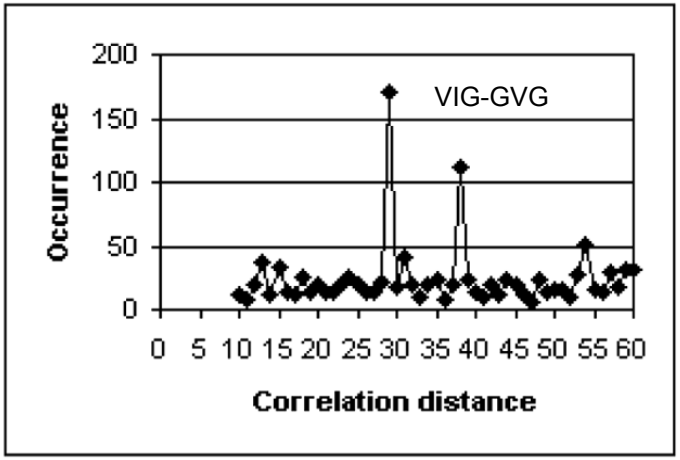


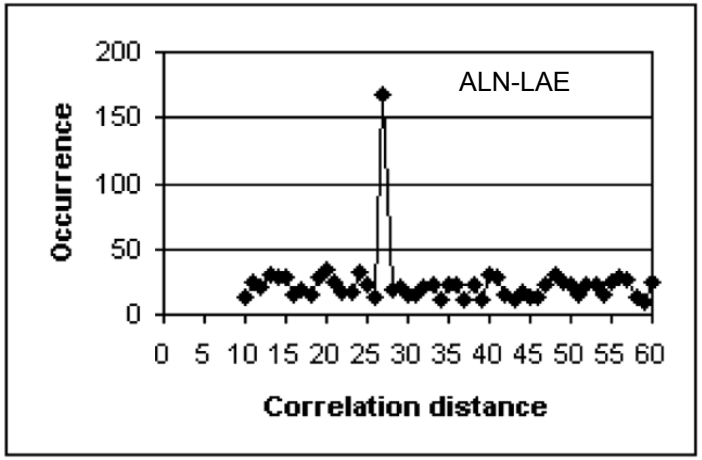

SGG-SAL
Occurrence
1500
1000
500
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance

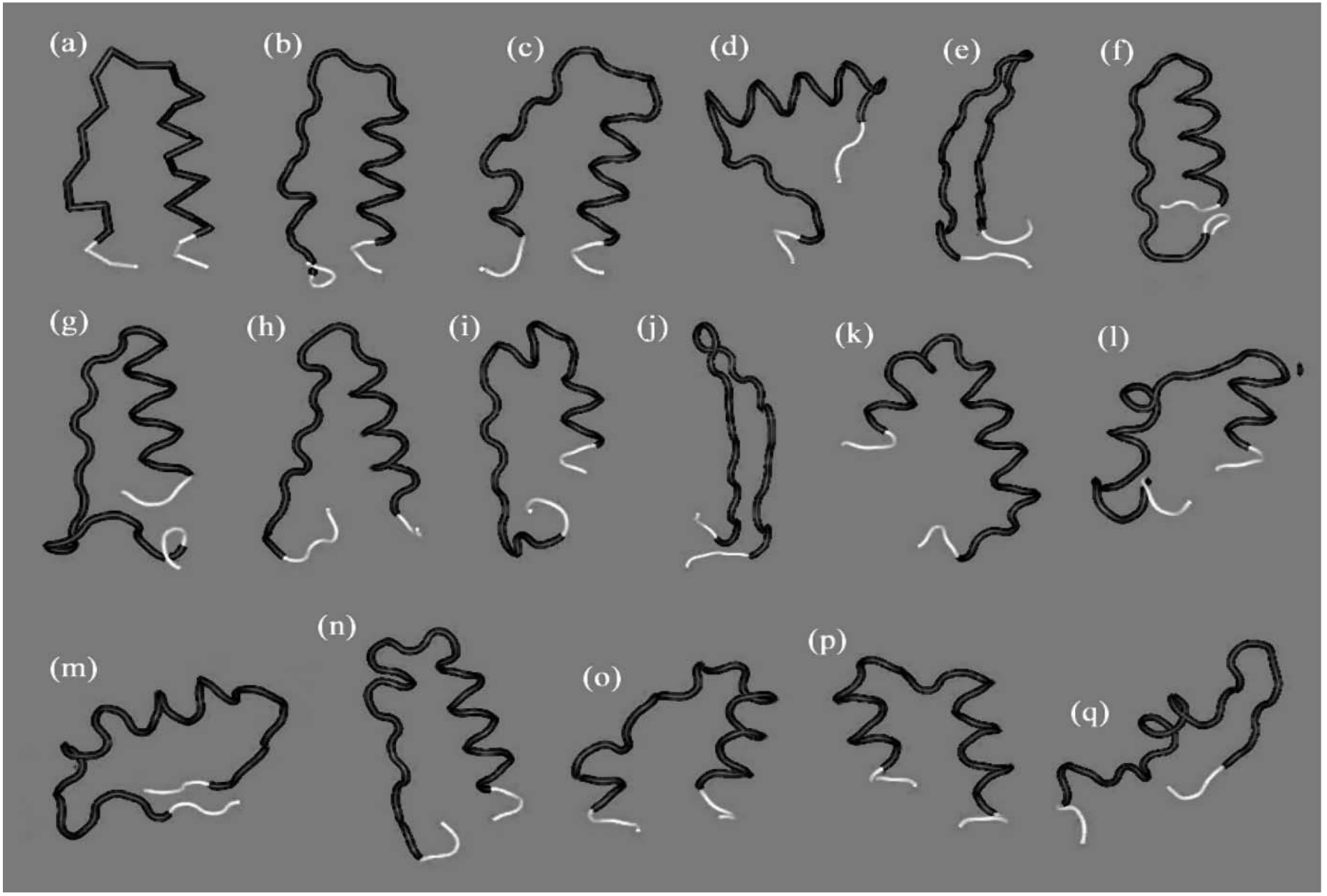

GIG-GLG
Occurrence
300
250
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance

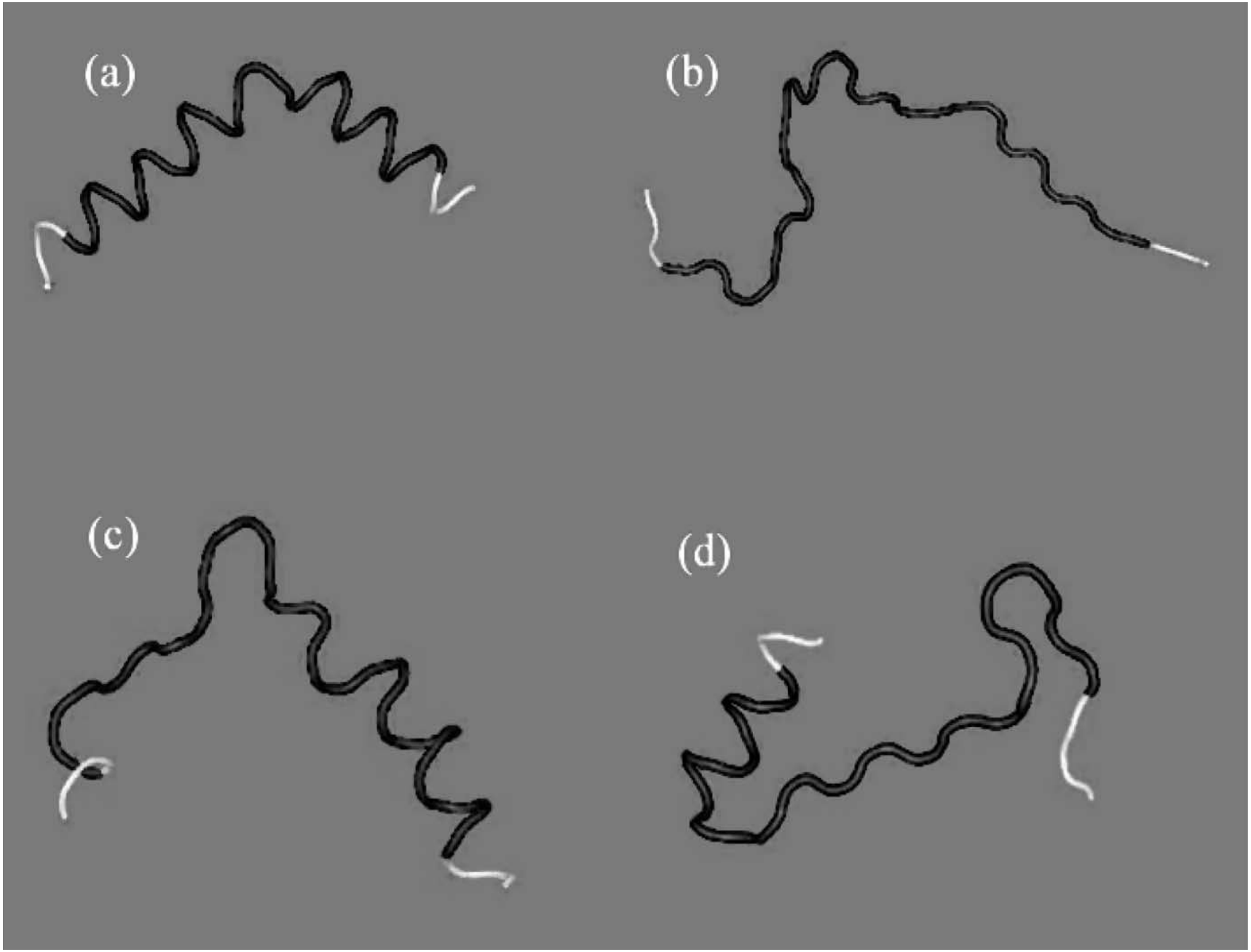

GGG-LGG
Occurrence
250
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance


VAI-EVL
Occurrence
250
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance


VIG-GVG
Occurrence
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance


ALN-LAE
Occurrence
200
150
100
50
0
0 5 10 15 20 25 30 35 40 45 50 55 60
Correlation distance


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)
(q)


(a)
(b)
(c)
(d)

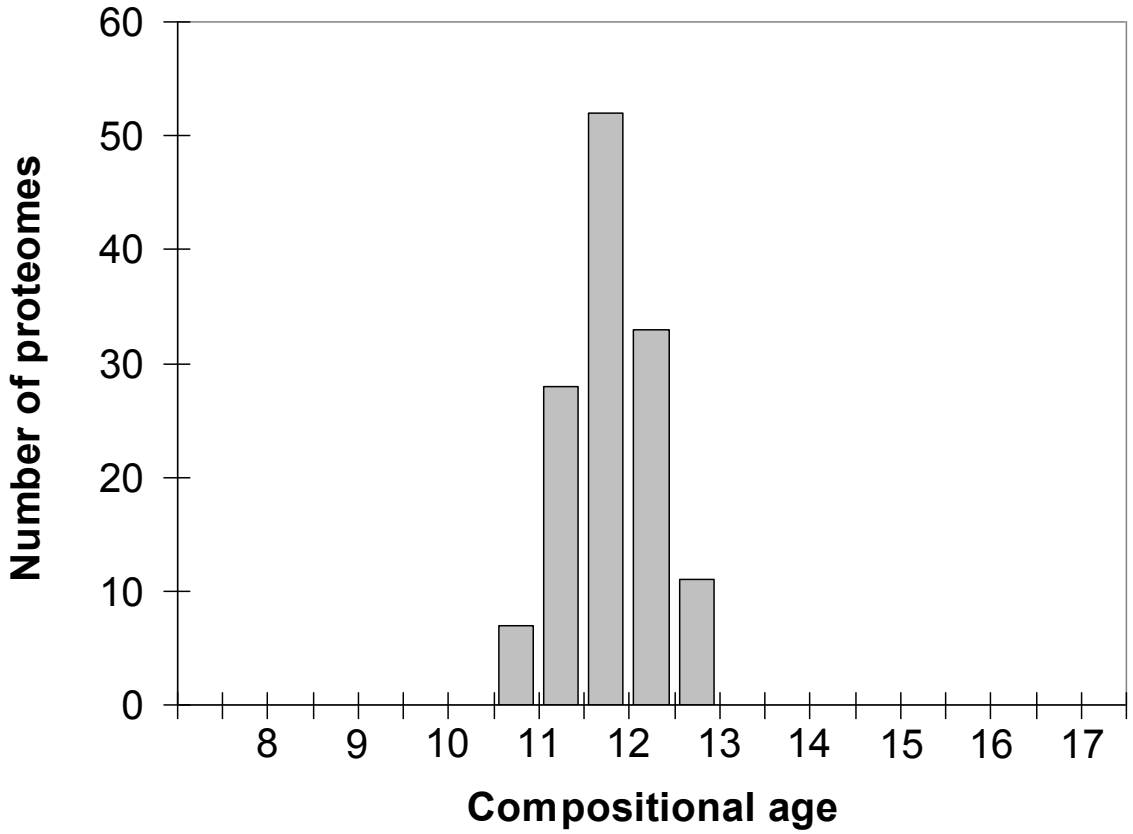
Omnipresent oligopeptides

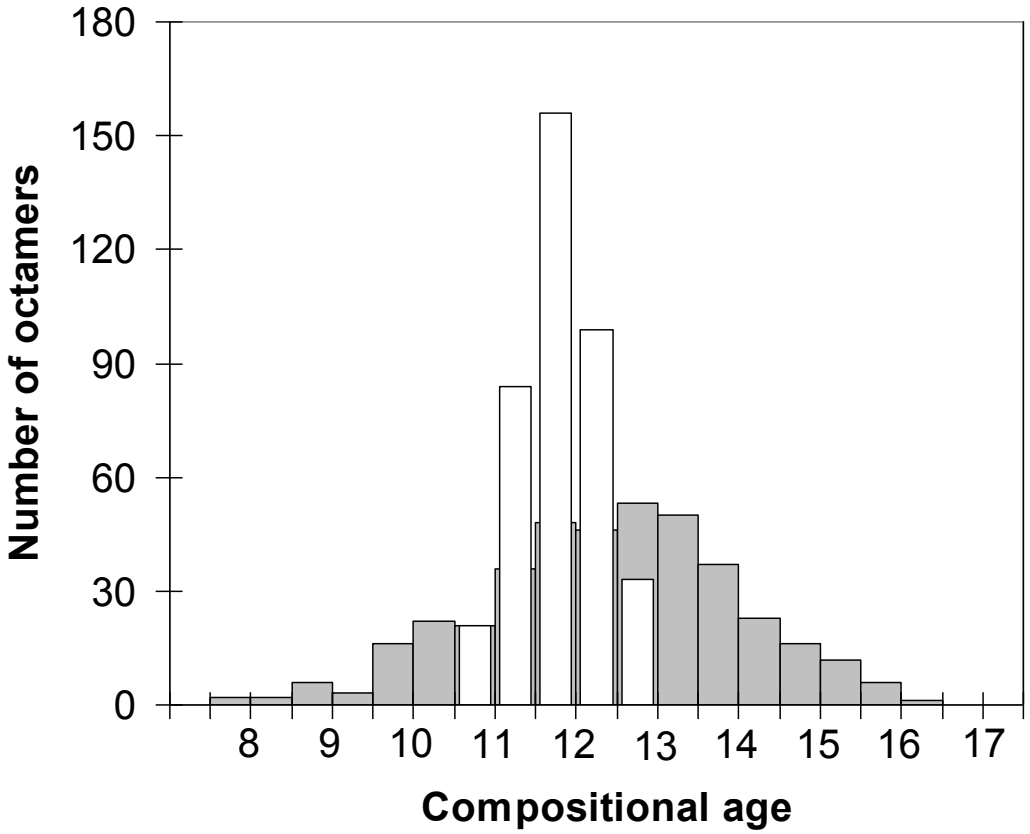
Omnipresent and frequent motifs

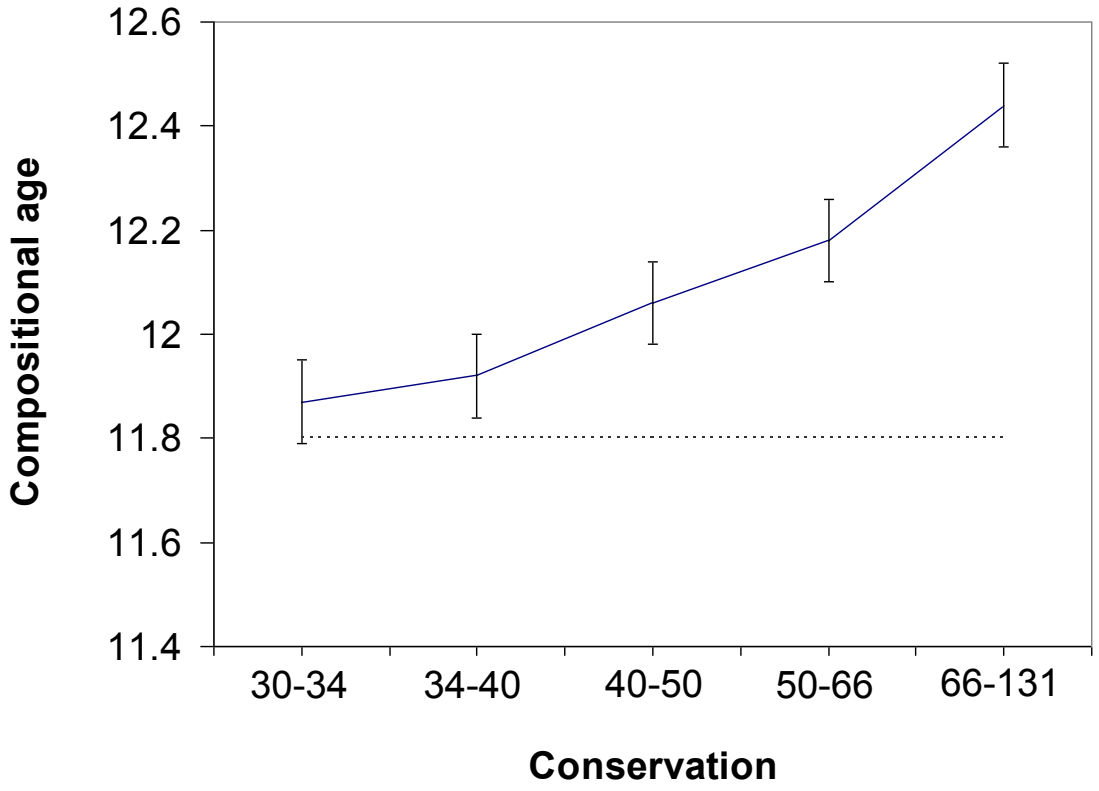
GHVDHGKT	131	AGRHGNKG	104
SGSGKSTL	125	PRSNPATY	104
LSGGQQQR	125	MTDADVDG	104
GPPGTGKT	122	LTEAGYVG	104
KMSKSLGN	121	INGFGRIG	104
LRPGRFDR	119	TQQPLGGK	104
QRVAIARA	119	PIGRTPRS	104
DEPTSALD	119	LPGKLADC	104
SIGEPGTQ	117	GDEGGFAP	104
SGGLHGVG	117	ERHRHRYE	103
VEGDSAGG	116	RYKGLGEM	103
GLPNVGKS	116	ATPIPRTL	103
DEPSIGLH	115	AVKAPGFG	103
DLGGGTFD	115	ATWWIRQA	103
GPNGAGKS	114	GTQLTMRT	102
GIDLGTTN	113	EPTAAALA	102
VITVPAYF	113	TLHRLGIQ	102
LNRAPTLH	113	NIIDTPGH	102
NADFDGDQ	113	SYYDYYQP	101
NLLGKRVD	113	EMFVGVGA	101
AGDGTTTA	112	LFGGAGVG	101
GPTGVGKT	112	TGRTHQIR	101
GIAVGMAT	112	PESSGKTT	101
GFDYLRDN	112	KPETINYR	101
ERERGITI	111	RERIRQIE	101
KPNSALRK	111	GQRFGEME	100
NMITGAAQ	111	GVQQALLK	100
SHRSGETE	110	PSAVGYQP	100
MAGRGTDI	110	EPTTALDV	99
IIFIDEID	110	QLSQFMDQ	99
GGTVGDIE	110	SRQLWWGH	99
KFSTYATW	109	DVLDTWFS	99
DEARTPLI	108	ADKEGFLR	99
HHNVGGLP	108	AHIDAGKT	99
GHNLQEHS	107	VRKRPGMY	99
GGRVKDLP	107	GYLTRRLV	98
LPDKAIDL	107	AAQMDGAI	98
NPRSTVGT	107	GVGERTRE	98
NEKRMLQE	106	NVISITDG	98
CPIETPEG	106	GGITQHIG	98
NPETVSTD	106	NMQRQAVP	97
LEYRGYDS	106	RIDNQLRG	97
SRSSALAS	106	DCPGHADY	97
HTRWATHG	106	EMEVWALE	97
DEREQTLN	105	GPGSICTT	97
DVSGEGVQ	105	GLTGRKII	97
GPSGCGKS	105	VDYSGRSV	96
KTKPTQHS	105	NPLGVPSR	96
DHPHGGGE	105	SAASFQET	96
GRFRQNLL	105	VPSGASTG	96

Less frequent motifs

SSDSQAMG	30	TGNTVIDA	30
LRQDPDII	30	GAIEQDAD	30
TGGEPLLR	30	VNAQQARR	30
SGVSGAGR	30	HDVKAVEY	30
PAMREGSG	30	LTDSTVLR	30
QASRISGV	30	NVVMMGMG	30
TSMGFTPL	30	VQIPCIER	30
GHRELPIR	30	WREPGCSM	30
LNVFPVPD	30	GHEQYTRN	30
AFANAFLG	30	TGYITEGQ	30
LLKILEGT	30	KATKVDGV	30
AYLFSGPR	30	TESFISAA	30
LLTFFYRY	30	RRLPKRGF	30
MLLRGQNL	30	AYSARNRS	30
DTALKTAD	30	SHEIRTPM	30
GQLTEKVR	30	GKSPNIFF	30
ASDMSGWL	30	EIWNLVFM	30
DNHYVPNL	30	NVNDSVTK	30
FPFIFRGA	30	GTAAGPHP	30
PVGFKNGT	30	SVKVPDPK	30
EDWGRRQL	30	FWAEWCGP	30
DASAERSA	30	GLPGNPVS	30
IGHTQPRR	30	CRNVLIYF	30
AINAPMQG	30	FLTGITEP	30
ETDSPYLA	30	GIEYGDMQ	30
KQFDVTRE	30	GAIGTGLF	30
GREQILKV	30	AVMGCVVN	30
DVAGCDEA	30	RRLLWPIK	30
AGANSIFY	30	DAANILKP	30
MAGLQGAG	30	RISLGIKQ	30
KGPAVRAT	30	DYVGSWGP	30
ATHYFELT	30	LVKTMRAS	30
GSKVSTKL	30	GDVSAFVP	30
RALWRATG	30	KPIVVINK	30
GMPESFNV	30	FPDLNTGN	30
KISVDSAT	30	GPVKDYEC	30
GGVQPQSE	30	DPHNLGAC	30
GYMYMLKL	30	LEEVGKQF	30
GRIVEIYG	30	EADESDAS	30
ALTPKAEI	30	GGGIANTF	30
GDLKYGRT	30	ALIIDSWF	30
TNGDTHLG	30	NAGSFFKN	30
ASSSSVYG	30	IATDHAPH	30
QTIISGMG	30	RAGTKAGN	30
ILHVSAKD	30	IAGNWKMN	30
AYIRFASV	30	NAGMNQFK	30
GYNFEDSI	30	HGTGCTLS	30
RTTDVTGV	30	GTSHGAYK	30
WDDPRMPT	30	TEETTTGV	30
AYLKISEG	30	LGIFLPLI	30

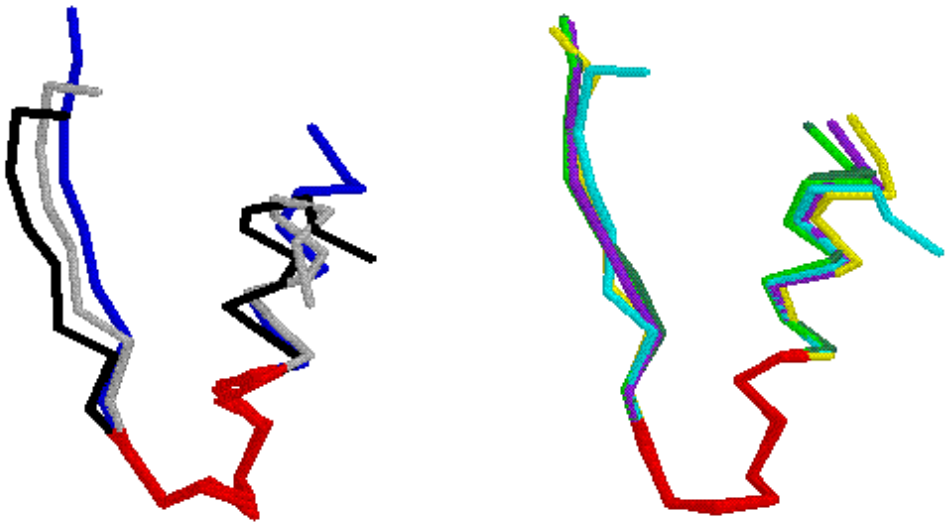

Number of proteomes
60
50
40
30
20
10
0
8
9
10
11
12
13
14
15
16
17
Compositional age


180
150
120
90
60
30
0
Number of octamers
8
9
10
11
12
13
14
15
16
17
Compositional age


12.6
12.4
12.2
12
11.8
11.6
11.4
Compositional age
30-34
34-40
40-50
50-66
66-131
Conservation

A

B

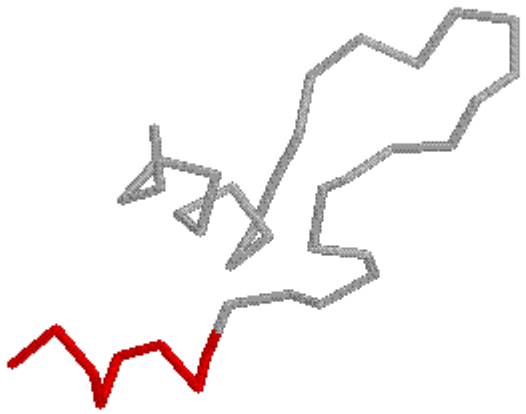


GHVDHGKT

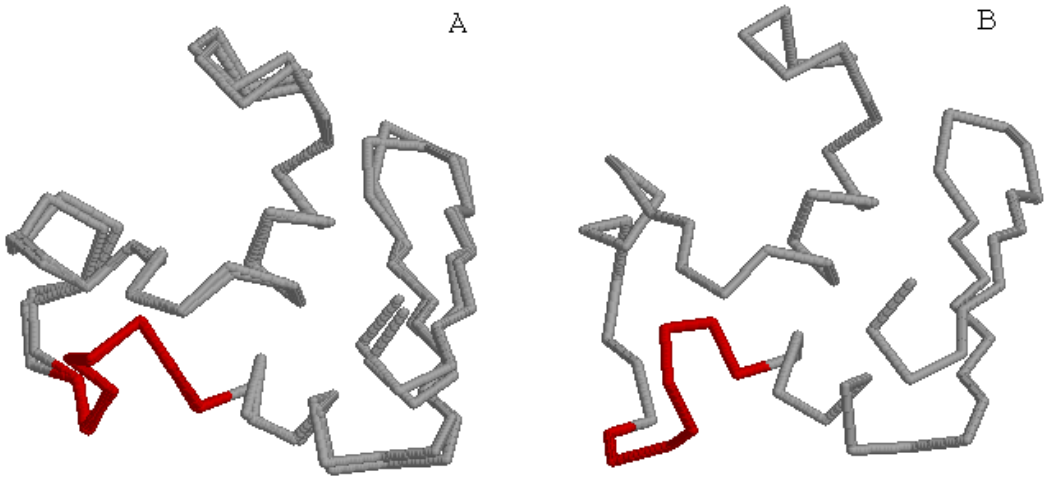
LSGGQQQR

KMSKSLGN

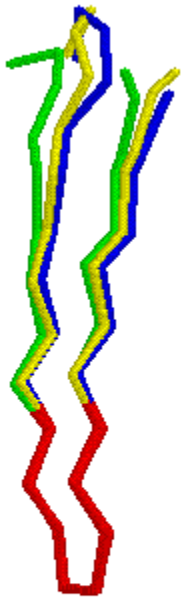

LRPGRFDR



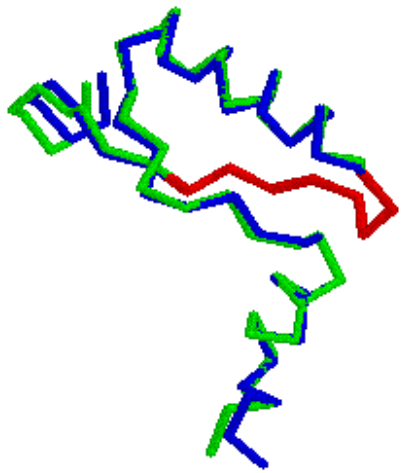
SIGEPGTQ



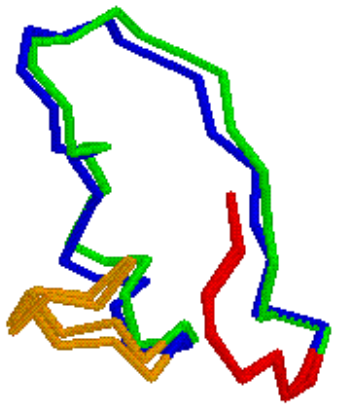
SGGLHGVG



DLGGGTFD



VITVPAYF



NADFDGDQ LNRAPTLH

AGDGTTTA

MOST COMMON
PROTEIN SEQUENCE MODULES (PROTOTYPES)

Aleph **GEIVLLVGPSGSGKTTLLRALAGLLGPDGG**

Beth **LSGGQRQRVAIARALALEPKLLLLDEPTSALD**

Gimel DVVVIGAGGAGLAAALALARAGAKVVVVE

Dalet RRGIGMVFQEYALFPHLTVLENVALGL

Heh PVIMLTARGDEEDRVEALLEAGADDYLTKPF

Vav LLGLSKKEARERALELLELVGLEEKADRYP

Zayin LLLKLLKELGLTVLLVTHDLEEA

Berezovsky et al. 2000-2003

The underlined motifs are omnipresent

KVALVGRSGSGKTTVTSLLM

FIAVEGIDGAGKTTLAKSLS

GxxxxGKT - Walker A motif (NTP binding)

Phylogenetically diverse prokaryotes
used for calculation of the omnipresent motifs

Bradyrhizobium japonicum
Streptomyces coelicolor
Rhodopirellula baltica
Bacillus cereus
Bacteroides thetaiotaomicron
Gloeobacter violaceus
Treponema denticola
Thermus thermophilus
Fusobacterium nucleatum
Thermotoga maritime
Aquifex aeolicus
Chlamydophila pneumoniae

Methanosarcina acetivorans
Nanoarchaeum equitans
Sulfolobus solfataricus

	sequences			
	NATURAL	SHUFFLE1	SHUFFLE2	SHUFFLE3
Tetramers	36593	40553	40485	40652
Pentamers	2326	1554	1442	1527
Hexamers	46	0	0	0
Heptamers	21	0	0	0
Octamers	9	0	0	0
Nonamers	3	0	0	0

Omnipresent 6-9 mers of 15 prokaryotes from different phyla

ALEPH ATP/GTP binding

```
 1      HVDHGKTTL
 2     GPPGTGKT
 3     GHVDHGKT
 4       GSGKTTLL
 5 IDTPGHV
 6     GPSGSGK
 7      PTGSGKT
 8       NGSGKTT
 9          GKSTLLN
10       SGSGKT
11       TGSGKS
12       PGVGKT
13       PNVGKS
14        GVGKTT
15        GTGKTT
16        DHGKST
17          GKTTLA
18          GKTTLV
19           KSTLLK
```

BETH ATPases of ABC transporters

```
20          QRVAIARAL
21   LSGGQQQRV
22                          LADEPT
23 TLSGGE
```

Other omni:

```
24 FIDEID
25 KMSKSL
26 WTTTPWT
27 NADFDGD
```

Omnipresence is a new measure of sequence conservation. These elements are the most conserved ones, coming, presumably from last common ancestor

EVOLUTIONARY ELITE
(OMNIPRESENT 6- to 9-MERS)

HVDHGKTTL	Aleph		
LSGGQQQRV		Beth	
QRVAIARAL		Beth	
GHVDHGKT	Aleph		
GPPGTGKT	Aleph		
GSGKTTLL	Aleph		
GKSTLLN	Aleph		
GPPGTGK	Aleph		
GPSGSGK	Aleph		
IDTPGHV			Dalet
NADFDGD			
NGSGKTT	Aleph		
PTGSGKT	Aleph		
WTTTPWT			
DHGKST	Aleph		
FIDEID			
GKTTLA	Aleph		
GKTTLV	Aleph		
GTGKTT	Aleph		
GVGKTT	Aleph		
KMSKSL			
KSTLLK	Aleph		
LADEPT		Beth	
PGVGKT	Aleph		
PNVGKS	Aleph		
SGSGKT	Aleph		
TGSGKS	Aleph		
TLSGGE		Beth	

Functional involvement of the most conserved octamers present in all (131) or almost all (125 and less) prokaryotic proteomes.

		number of genomes	protein function
1.	GHVDHGKT	131	initiation and elongation factors
2.	SGSGKSTL	125	ABC transporter family proteins
3.	LSGGQQQR	125	ABC cassettes, transporters
4.	GPPGTGKT	122	cell division proteins
5.	KMSKSLGN	121	aa-tRNA synthetases class I
6.	QRVAIARA	119	ABC cassettes, transporters
7.	DEPTSALD	119	ABC cassettes, transporters
8.	LRPGRFDR	119	cell division proteins
9.	SIGEPGTQ	117	DNA-directed RNA polymerases
10.	SGGLHGVG	117	topoisomerases
11.	VEGDSAGG	116	topoisomerases
12.	GLPNVGKS	116	GTP/ATP binding proteins
13.	DEPSIGLH	115	exinuclease ABC (UvrA)
14.	DLGGGTFD	115	chaperones (heat shock) proteins
15.	GPNGAGKS	114	ABC transporters
16.	GIDLGTTN	113	chaperones
17.	VITVPAYF	113	ATPase of heat shock protein 70
18.	LNRAPTLH	113	RNA polymerase beta' subunit
19.	NADFDGDQ	113	RNA polymerase beta' subunit
20.	NLLGKRVD	113	RNA polymerase beta' subunit
21.	AGDGTTTA	112	chaperonin GroEL
22.	GPTGVGKT	112	chaperone ClpB
23.	GIAVGMAT	112	DNA gyrase subunit A
24.	GFDYLRDN	112	preprotein translocase secA subunit
25.	ERERGITI	111	GTP-binding protein lepA
26.	KPNSALRK	111	30S ribosomal protein S12
27.	NMITGAAQ	111	elongation factor TU
28.	SHRSGETE	110	**enolase (phosphopyruvate hydratase)**
29.	MAGRGTDI	110	preprotein translocase secA subunit
30.	IIFIDEID	110	cell division protein FtsH
31.	GGTVGDIE	110	**CTP synthase**
32.	KFSTYATW	109	RNA polymerase sigma factor rpoD
33.	DEARTPLI	108	preprotein translocase secA subunit
34.	HHNVGGLP	108	**GMP synthase**
35.	GHNLQEHS	107	30S ribosomal protein S12
36.	GGRVKDLP	107	30S ribosomal protein S12
37.	LPDKAIDL	107	chaperone ClpB
38.	NPRSTVGT	107	excinuclease ABC subunit A
39.	NEKRMLQE	106	DNA-directed RNA polymerase beta' chain
40.	CPIETPEG	106	DNA-directed RNA polymerase beta chain
41.	NPETVSTD	106	**carbamoyl-phosphate synthase large chain**
42.	LEYRGYDS	106	**glucosamine-fructose-6-phosphate aminotransferase**
43.	SRSSALAS	106	**carbamoyl-phosphate synthase large chain**
44.	HTRWATHG	106	**glucosamine-fructose-6-phosphate aminotransferase**
45.	DEREQTLN	105	cell division protein FtsH
46.	DVSGEGVQ	105	Clp protease ATP-binding subunit clpX
47.	GPSGCGKS	105	phosphate import ATP-binding protein pstB
48.	KTKPTQHS	105	**CTP synthase**

Motifs involved in elementary syntheses appear late

Many of the 27 omnipresent elements do not match to one another

(e. g. WTTTPWT and QRVAIARAL)

yet, they turn out to belong to the same network.

Major nuclei in sequence space
(10% Monster)

LSGGQRQRVAIARALALDPD	3753	60%
++++++++++++++++++---		
LSGGQRQRVAIARALALEPKLLLLDEPTSALD	Beth	
GEFVAIV**G**PSGC**GKS**TLLRL	3043	60%
++-+--++++++-++-++++-		
GEIVLLV**G**PSGS**GKT**TLLRALAGLLGPDGG	Aleph	

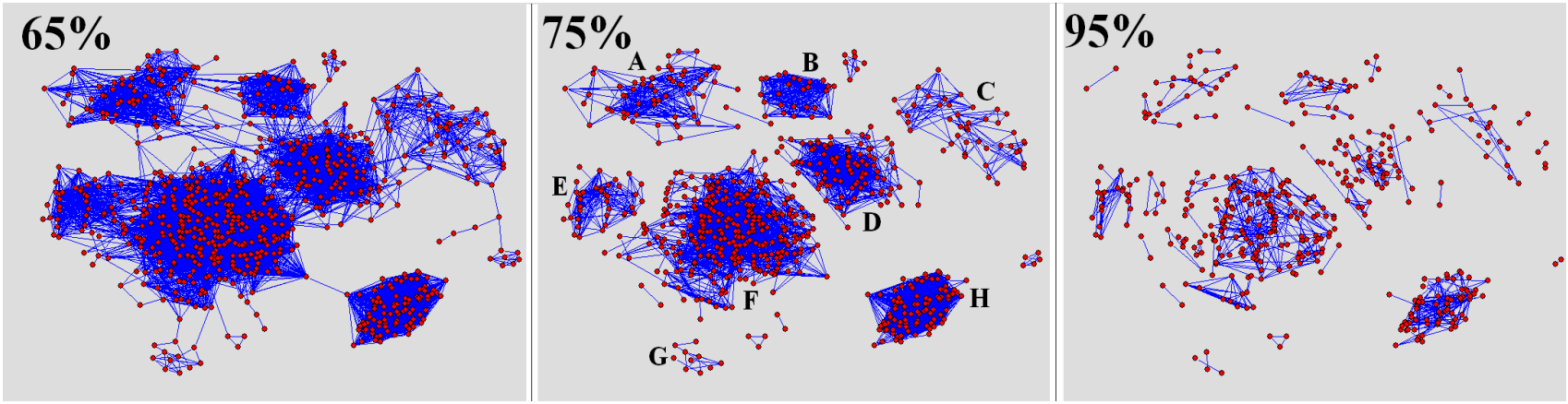
All 20 aa fragments of all proteins of prokaryotes make a

sequence space

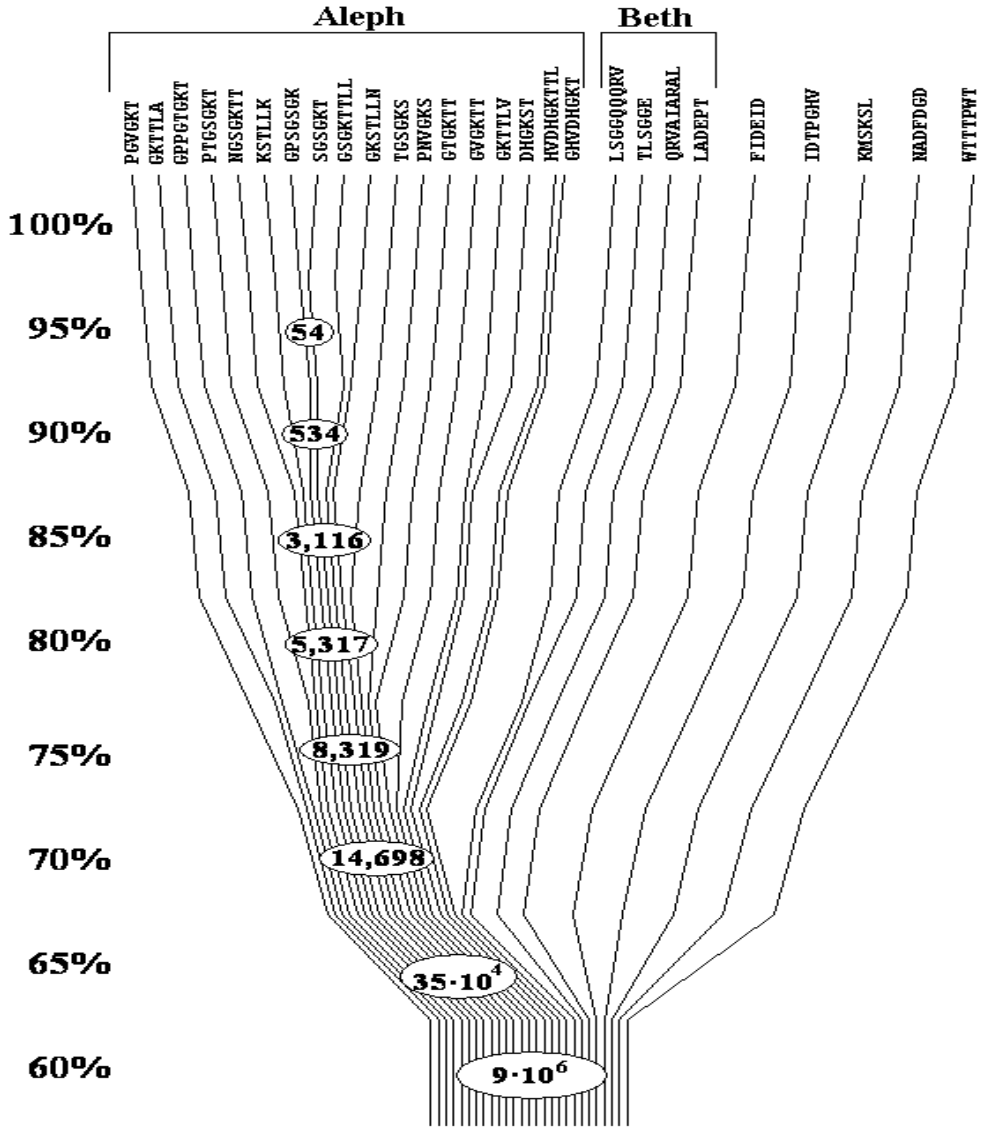
Those fragments that are close relatives (matching >60%) are pair-wise connected. This makes

networks

that allow tracing evolutionary relatedness of protein sequence motifs



A tyr trp B met C arg trp D cys
E leu F met leu ile val G ile H lepA

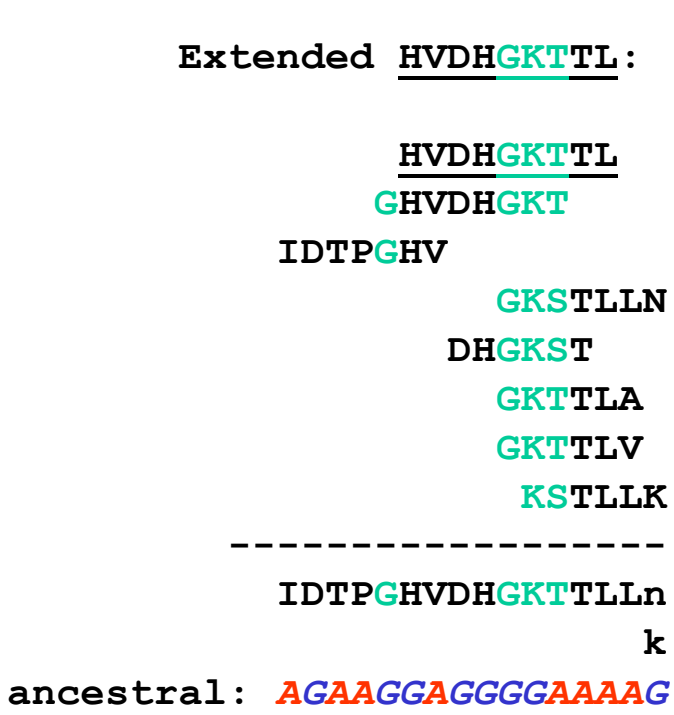


Sequence space based
evolutionary tree of omnipresent elements

All omnipresent elements are relatives!

They belong to the same 60% match network

RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 1.



RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 2.

Extended QRVAIARAL and LSGGOOORV:

```
                    QRVAIARAL
                LSGGQQQRV
               TLSGGE
             ------------------
               TLSGGqQQRVAIARAL
                    e
ancestral:     AASGGGGGGAAAAGAA
```

RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 3.

Remaining Aleph motifs:

```
          GPPGTGKT
             GSGKTTLL
          GPSGSGK
           PTGSGKT
            NGSGKTT
            SGSGKT
            TGSGKS
            PGVGKT
            PNVGKS
             GVGKTT
             GTGKTT
         ------------
consensus:  GPPGSGKTTLL

   binary:  GAAGSGGAAAA
```

RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 4.

Other omni:

WTTTPWT	*GAAAAGA*
NADFDGD	*GAGAGGG*
LADEPT	*AAGGAA*
FIDEID	*AAGGAG*
KMSKSL	*GASGSA*
	----------
consensus:	*GAAAGGAA*

RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. ALIGNMENT OF FOUR GROUPS.

```
AGAAGGAGGGGAAAAG      Aleph
  AASGGGGGGAAAAGAA    Beth
    GAAGSGGAAAA       rest of Aleph
          GAAAGGAA    rest of omni
--------------------
AGAAGGAGGGGAAAAGAA    common prototype
```

The above mentioned example of no match:

```
          GAAAAGA     WTTTPWT
        GGAAAAGAA     QRVAIARAL
```

This is, apparently,
why the omnipresent elements
belong to one common network of relatives


A G AA GG A GGGG AAAA G AA prototype
I D TP GH V DHGK TTLL N Aleph
TL SG G QQQR VAIA R AL Beth

In binary form ALEPH and BETH are rather similar

AGAAGGAGGGGAAAAG

++-+-++++++++++

*AA*S*GGGGGGAAAAGAA*

Compare to

IDTPGHVDHGKTTLLN

+

TLSGGQQQRVAIARAL

Symmetry properties of common prototype

AGAAGGAGGGGAAAAGAA

AAAAGAA GGAGGGG AAAAGAA

GGAGGGG AAAAGAA GGAGGGG

This is blunt end fusion of the same element

GGAGGGG

AAAAGAA

TWO RECONSTRUCTIONS MEET

OMNIPRESENT ELEMENTS

RECONSTRUCTION OF ALEPH AND BETH

ALEPH: IDTPGHVDHGKTTLLnk

BETH: TLSGGqQQRVAIARALe

COMMON BINARY PROTOTYPE OF ALEPH AND BETH

AGAAGGAGGGGAAAAGAA

AAAAAAA | *GGGGGGG* | *AAAAAAA*

AGAA | *GGAGGGG* | *AAAAGAA*

AAAAAAAGGGGGGGAAAAAAA

BINARY MOSAIC

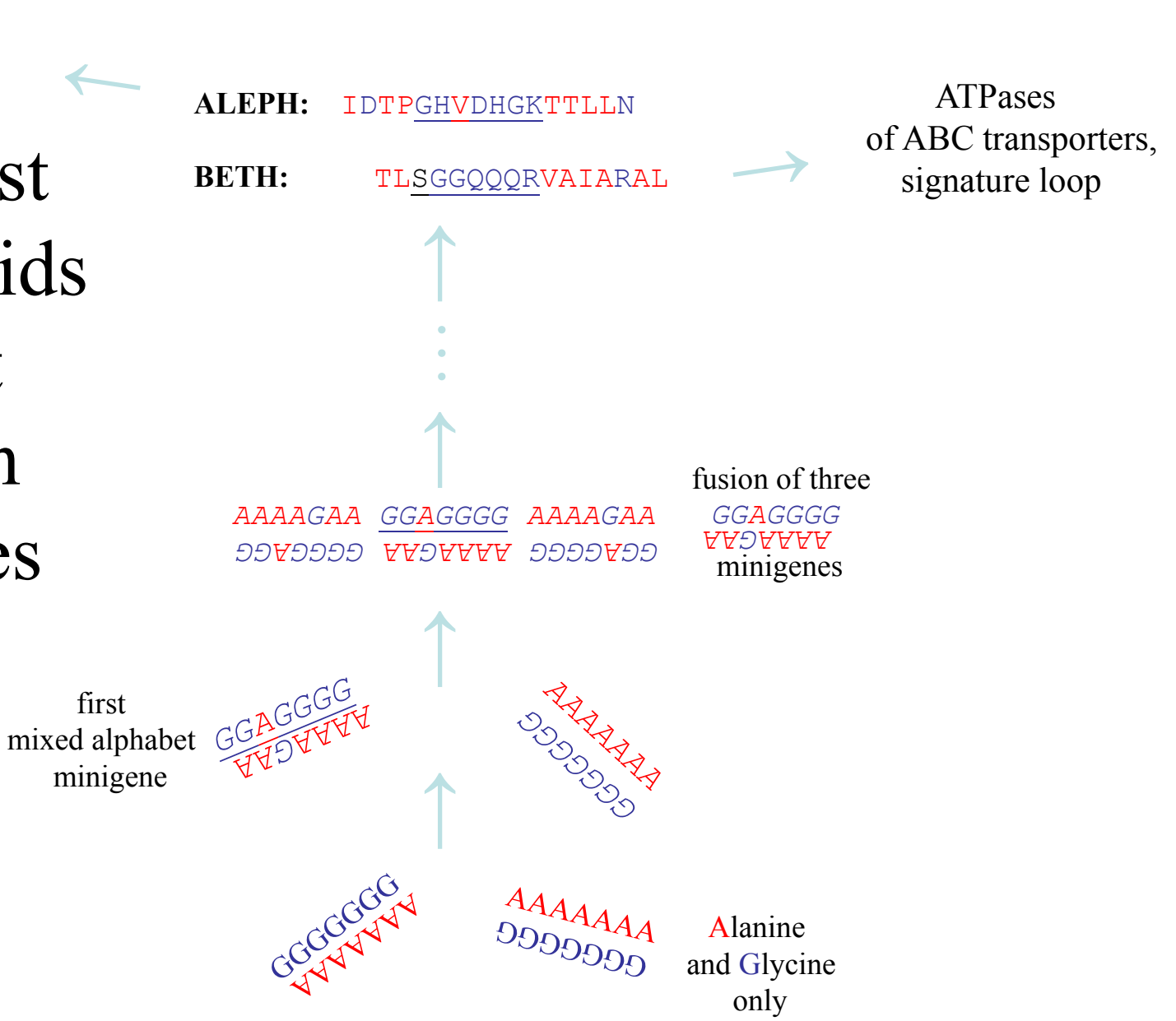
GGGGGGG **&** *AAAAAAA*

FIRST PEPTIDES

BINARY ALPHABET

EVOLUTIONARY CHART OF CODONS

from first amino acids to first protein modules



According to the same theory
(reconstruction of evolutionary history of the triplet code)

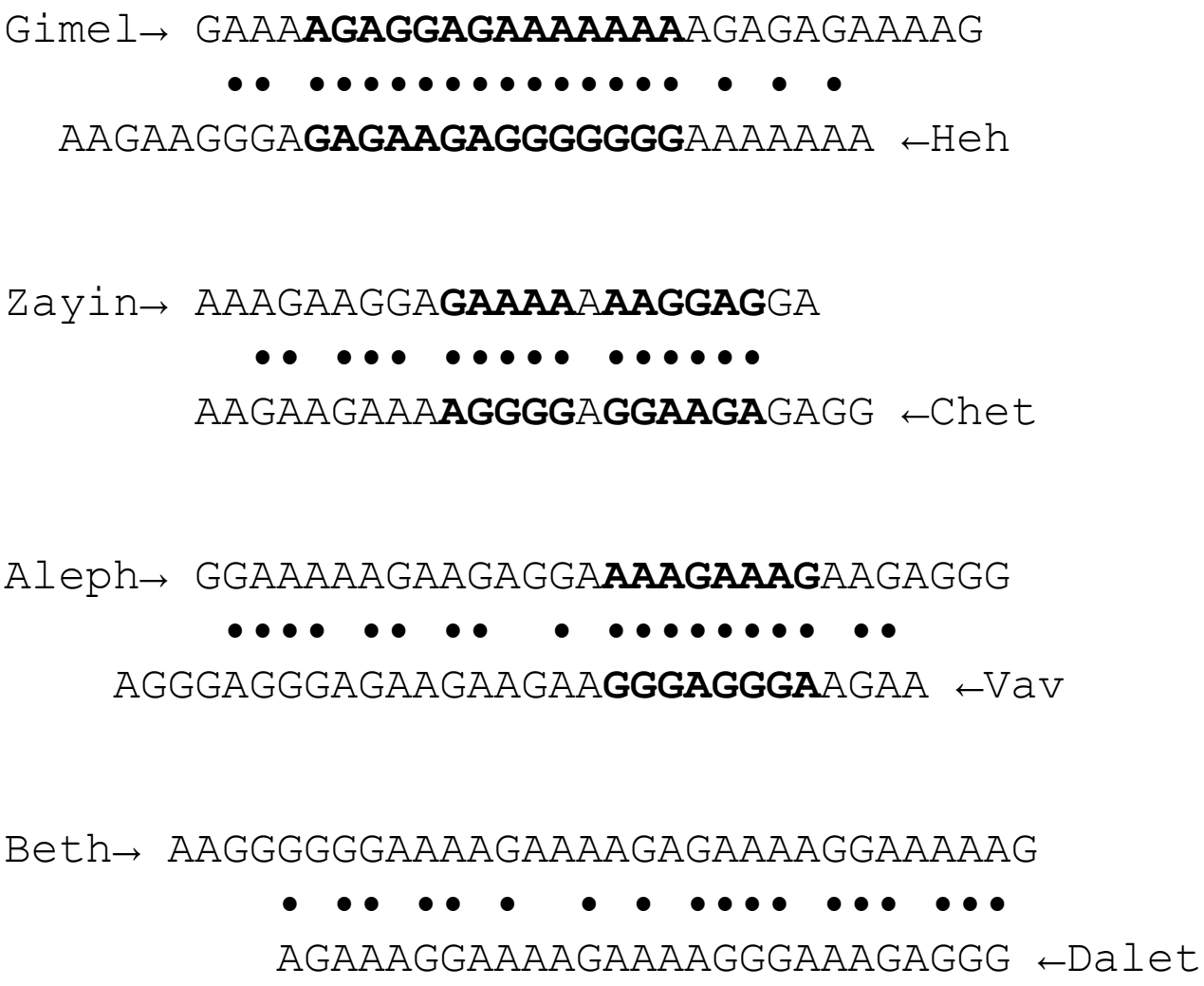
the earliest proteins have been encoded in both strands of the genes-duplexes,
so that the xYx codons of one strand
would be complementary to xRx codons of another strand.

Remarkably, the above ALEPH and BETH are, indeed, complementary:

ALEPH *AGAAGGAGGGGAAAAG*

| | | | | | | | | | | | | | | | – |

AA **S** *GGGGGGAAAAGAA* **BETH**


Gimel→ GAAAAGAGGAGAAAAAAAAGAGAGAAAAG
AAGAAGGGAGAGAAGAGGGGGGGAAAAAAA ←Heh
Zayin→ AAAGAAGGAGAAAAAAAGGAGGA
AAGAAGAAAAGGGGGAGGAAGAGAGG ←Chet
Aleph→ GGAAAAAGAAGAGGAAAAGAAAGAAGAGGG
AGGGAGGGAGAAGAAGAAGGGAGGGAAGAA ←Vav
Beth→ AAGGGGGGAAAAGAAAAGAGAAAAGGAAAAAG
AGAAAGGAAAAGAAAAGGGAAAGAGGG ←Dalet

All 27 omnipresent LUCA motifs originate from one prototype sequence, which is:

Ala Ala Ala Ala Gly Ala Ala Gly Gly Ala Gly Gly Gly Gly

encoded in

GCC GCC GCC GCC GGC GCC GCC GGC GGC GCC GGC GGC GGC GGC

which is self-complementary:

GCC GCC GCC GCC GGC GCC GCC GGC GGC GCC GGC GGC GGC GGC

CGG CGG CGG CGG CCG CGG CGG CCG CCG CGG CCG CCG CCG CCG

The very first gene

was a short duplex,

encoding the same thing in both strands

ENZYMATIC REPERTOIRE OF LUCA

Omnipresent cassette of

ABC transporters

(32-72)**GPSGSGKTTLL**(29-41)MVFQNYALFPHLTALENV(31-42)Q**LSGGQQQRVAIARAL** (6) L**LADEPT**SALD(21-22)IYVTHDQ(28-263) consensus

Bacteria

(35)	GPSGcGKTTmL	(36)	MVFQsYAvwPHmnvfdNi	(36)	eLSGGQQQRVAlgRAL	(6)	LLlDEPlSnLD	(22)	IYVTHDQ	(158)	Q8RGI3	-	Fnu
(38)	GPSGSGKsTLm	(38)	fVFQqfnLmarsdALENV	(36)	QLSGGQQQRVAvARAL	(6)	LLADEPTgALD	(21)	lviTHDQ	(28)	Q7NNB9	-	Gvi
(32)	GPSGSGKTTfL	(39)	MVFQhhnLFPHLTALqNV	(38)	QLSGGQQQRVgIARAL	(6)	LLfDEPTSALD	(21)	viVTHem	(44)	Q81HE0	-	Bce
(33)	GknGSGKTTLL	(29)	yVFQNpssqiigatvEed	(37)	nLSGGQkQRlAIAsmL	(6)	LalDEPvSmLD	(21)	IlVTHel	(68)	Q9x1z1	-	Tma
(37)	GPSGcGKTTLL	(32)	fVFQdYALFPHLTALgNV	(31)	eLSGGQQQRVAlARAL	(6)	vLlDEPfSsLD	(22)	llVTHDQ	(158)	AAS81608	-	Tth
(35)	GeSGSGKssiL	(41)	MVFQepsLyldplftvgs	(42)	QLSGGlkQRVcIAnAi	(6)	vLADEPTtALD	(21)	IliTHDf	(43)	O67913	-	Aae
(45)	GPSGSGKTTtL	(32)	MVFQNYALFPHLTiaENi	(36)	QLSGGQQQRVAlARAL	(6)	vLmDEPlgALD	(22)	vYVTHDQ	(165)	Q89FQ5	-	Bja
(41)	GPSGcGKTTLL	(32)	tVFQkYALFPHLnvydNi	(36)	sLSGGQQQRVAIARAi	(6)	LLlDEPlaALD	(22)	vYVTHDQ	(263)	Q8A883	-	Bth
(52)	GeSGSGKsTLa	(37)	lVFQNpqaslnprktild	(40)	QLSGGQQQRVsIARAL	(6)	iicDEivSALD	(22)	lfisHDl	(104)	Q9Z7M1	-	Cpn
(72)	GPSGSGKsTLL	(38)	fVFQsYnLiqqLsvvENi	(36)	QLSGGQQQRVAIARsL	(6)	iLADEPTgnLD	(21)	IlVTHed	(50)	Q7UPF2	-	Rba
(49)	GPSGSGKsTLc	(36)	MVFQsfnLFaHkTvLENV	(37)	QLSGGQQQRVAIARAL	(6)	mLfDEPTSALD	(21)	IvVTHem	(46)	O50495	-	Sco
(34)	GPSGSGKTTLm	(38)	lVFQqfhLvnyLTALENV	(33)	QLSGGeQQRVcIARAL	(6)	LLADEPTglnD	(21)	IvVTHDp	(34)	AAS12033	-	Tde

Archaea

(41)	GPSGSGKsTmm	(38)	fVFQqYnLiPgmTALENV	(36)	QLSGGQQQRVsIARAL	(6)	vLADEPTgALD	(22)	vmVTHDm	(31)	Q8TNL0	-	Mac
(35)	GPSGSGKTTLL	(39)	fVFQhsyLiPvLTALENV	(33)	QLSGGQQQRVAIARAL	(6)	iLADEPTasLD	(21)	vmVTHDp	(33)	AAR39266	-	Neq
(40)	GPSGeGKTTiL	(32)	MVpQNYAiyPfmsvydNi	(36)	QLSGGQmQRVAIARAL	(6)	iLmDEPlSnLD	(22)	IYVTHDQ	(169)	Q97YY4	-	Sso

Omnipresent cassette of

Proteases (cell division protein FtsH, zinc-dependent metalloprotease)

```
(146-463)LLVGPPGTGKTLLARAVAGEA (7) SGSDFVEMFVGVGASRVRD (9) PCIIFIDEIDAVGR(7-11)DEREQTLNQLLVEMDGF  consensus (cont.)

  (191)  LLyGePGvGKTLLAkAiAGEA (7) SGSDFVEMFVGVGAaRVRD (9) PCIIFIDEIDAVGR (10) DEREQTLNQLLVEMDGF  O67077   - Aae
  (198)  LLVGPPGTGKTLLARAVAGEA (7) SGSDFVEMFVGVGASRVRD (9) PCIIFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q81J82   - Bce
  (192)  LLVGPPGTGKTLiARAVAGEA (7) SGSDFVEMFVGVGASRVRD (9) PCIIFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q9XBG5   - Bja
  (213)  LLVGPPGTGKTLLAkAVAGEA (7) aGSDFVEMFVGVGASRVRD (9) PCIvFIDEIDAVGR (10) DEREnTLNQLLtEMDGF  Q8A0L4   - Bth
  (463)  LLiGPPGTGKTLiAkAVsGEA (7) aGSDFVEMFVGVGASRiRD (9) PCIIFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q9Z6R1   - Cpn
  (309)  LLlGePGTGKTLLAkAVAGEA (7) SGSeFVEMFVGVGASRVRD (9) PCIvFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q8R6D4   - Fnu
  (210)  LLVGPPGTGKTLLAkAiAGEA (7) SGSeFVEMFVGVGASRVRD (9) PCIvFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q7NHF9   - Gvi
  (233)  LLnGPPGTGKTLLARAVAGEA (7) nGSeFiqMFVGVGASRVRD (9) PsIIFIDEIDAVGR (11) DEREQTLNQILgEMDGF  Q7UUZ7   - Rba
  (239)  LLtGPPGTGKTLLARAVAGEA (7) SaSeFiEMiVGVGASRVRe (9) PsIIFIDEIDtiGR (10) DEREQTLNQILtEMDGF  O69875   - Sco
  (241)  LLVGPPGTGKTLLARAVAGEA (7) SGSDFVEMFVGVGASRVRD (9) PCIIFIDElDAiGk (11) DEREQTLNQLLVEMDGF  AAS10965 - Tde
  (197)  LLVGPPGTGKTLLARAVAGEA (7) SGSDFVElFVGVGAaRVRD (9) PCIvFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  Q9WZ49   - Tma
  (192)  LLVGPPGvGKThLARAVAGEA (7) SGSDFVEMFVGVGAaRVRD (9) PCIvFIDEIDAVGR (11) DEREQTLNQLLVEMDGF  AAS81470 - Tth

  (213)  LLhGPPGTGKTmiAkAVAsEt (7) SGpeiVskyyGeseqklRe (9) PsIIFIDEIDsiap (11) emerrvvaQLLslMDGl  Q8THE2   - Mac
  (146)  LLyGPPGTGKTLigkAlAksA (7) vGSelVqkyiGeGAklVke (9) PaIvFIDEIDAiaa (11) rEvqrTfmQLLaEiDGF  AAR39040 - Neq
  (238)  LLyGPPGvGKTLLARAlAnEi (7) nGpeimskFyGeseqRlRe (9) PaIIFIDEIDAiap  (7) evekrvvaQLLtlMDGi  Q97ZZ9   - Sso

(8) IAATNRPDxLDPALLRPGRFDRQ (95-415)  consensus

(8) IAATNRPDILDPALLRPGRFDRQ   (314)   O67077   - Aae
(8) vAATNRPDILDPALLRPGRFDRQ   (307)   Q81J82   - Bce
(8) IAATNRPDvLDPALLRPGRFDRQ   (320)   Q9XBG5   - Bja
(8) lAATNRvDvLDkALLRaGRFDRQ   (354)   Q8A0L4   - Bth
(8) mAATNRPDvLDkALLRPGRFDRr   (319)   Q9Z6R1   - Cpn
(8) lAATNRaDvLDkALrRPGRFDRQ   (277)   Q8R6D4   - Fnu
(8) IAATNRPDvLDaAiLRPGRFDRQ   (292)   Q7NHF9   - Gvi
(8) IAATNRPDvLDPALLRPGRFDRh   (311)   Q7UUZ7   - Rba
(8) IAATNRaDILDaALtRPGRFDRv   (280)   O69875   - Sco
(8) lAATNRPDvLDPALLRPGRFDRQ   (290)   AAS10965 - Tde
(8) mAATNRPDILDPALLRPGRFDkk   (285)   Q9WZ49   - Tma
(8) mAATNRPDILDPALLRPGRFDRQ   (304)   AAS81470 - Tth

(8) IAATNRPnsiDeALrRgGRFDRe   (415)   Q8THE2   - Mac
(8) IgATNRlDILDPAiLRPGRFDRi    (95)   AAR39040 - Neq
(8) IgATNRPDavDPALrRPGRFDRe   (406)   Q97ZZ9   - Sso
```

Omnipresent cassette of **Initiation factor 2**

(10-546)MGHVDHGKTTLL (11) EAGGITQHIGA(11-29)F**IDTPGH**EAFT (14) LVVAADDGV (21) INKIDLP(381-458)consensus

(313)	MGHVDHGKTTLL	(11)	EkGGITQHIGA	(12)	FlDTPGHEAFT	(14)	LVVAADDGV	(21)	vNKIDKP	(384)	O67825	-	Aae
(195)	MGHVDHGKTTLL	(11)	EAGGITQHIGA	(11)	FlDTPGHaAFT	(14)	LVVAADDGV	(21)	vNKmDKP	(384)	Q812X7	-	Bce
(345)	MGHVDHGKTsLL	(11)	EAGGITQHIGA	(13)	FIDTPGHaAFT	(14)	LVVAADDGV	(21)	INKIDKP	(388)	Q89WA9	-	Bja
(546)	MGHVDHGKTsLL	(11)	EAGGITQHIGA	(12)	FlDTPGHEAFT	(14)	iiVAADDnV	(21)	INKvDKP	(386)	Q8A2A1	-	Bth
(342)	MGHVDHGKTTLI	(11)	EAGaITQHmGA	(11)	ilDTPGHEAFs	(14)	LVVAgDeGi	(21)	INKcDKP	(381)	Q9Z8M1	-	Cpn
(244)	MGHVDHGKTsLL	(11)	EAGGITQkIGA	(11)	FIDTPGHEAFT	(14)	LVVAADDGV	(21)	vNKIDKP	(386)	Q8R5Z1	-	Fnu
(424)	MGHVDHGKTsLL	(11)	EAGGITQHIGA	(15)	FlDTPGHEAFT	(14)	LVVAADDGV	(21)	INKvDKP	(390)	Q7NH85	-	Gvi
(536)	lGHVDHGKTsLL	(11)	EAGGITQHIrA	(11)	FvDTPGHEAFT	(14)	LVVAADDGi	(21)	lNKIDle	(395)	Q7URR0	-	Rba
(533)	MGHVDHGKTrLL	(11)	EAGGITQHIGA	(15)	FIDTPGHEAFT	(14)	LVVAAnDGV	(21)	vNKIDve	(389)	Q8CJQ8	-	Sco
(322)	MGHVDHGKTKTL	(11)	EfGGITQHIGA	(11)	FlDTPGHEAFT	(14)	LVVAADDGV	(21)	vNKvDKP	(407)	AAS11595	-	Tde
(185)	MGHVDHGKTTLL	(11)	EeGGITQsIGA	(11)	FIDTPGHElFT	(14)	LVVAADDGV	(21)	INKIDKP	(398)	Q9WZN3	-	Tma
(78)	MGHVDHGKTTLL	(11)	EAGGITQHvGA	(11)	FIDTPGHEAFT	(14)	iViAADDGi	(21)	INKIDlP	(386)	AAS80695	-	Tth
(20)	MGHVDHGKTTLL	(11)	EAGAITQHIGA	(27)	FIDTPGHhAFT	(14)	vVVdineGf	(21)	aNKIDri	(454)	Q8TQL5	-	Mac
(10)	lGHVDHGKTTLL	(11)	EAGGITQHIGA	(29)	FIDTPGHEAFs	(14)	vVidineGi	(21)	aNKIDKi	(439)	AAR39338	-	Neq
(17)	lGHVDHGKTTLL	(11)	EpGemTQevGA	(29)	FIDTPGHEyFs	(14)	LVVditeGl	(21)	aNKIDKi	(458)	Q980Q8	-	Sso

Omnipresent cassette of

Aminoacyl-tRNA synthases (class I)

(495-671) DQTRGWF(29-84)GR**KMSKSL**GN(318-467)consensus

(585)	DQhRGWF	(29)	GRKMSKSLGN	(325)	O66651	- Aae
(554)	DQyRGWF	(29)	GRKMSKSiGN	(321)	Q819R4	- Bce
(632)	DQhRGWF	(29)	GRKMSKSLGN	(324)	Q89DF8	- Bja
(671)	DQTRGWF	(29)	GnKMSKrLnN	(445)	Q8A9K9	- Bth
(552)	DQTRGWF	(29)	GnKMSKrLnN	(445)	Q9Z972	- Cpn
(568)	DQhRGWF	(29)	GkKMSKSLGN	(320)	Q8RH47	- Fnu
(606)	DQhRGWF	(29)	GRKMSKSLGN	(327)	Q7NF75	- Gvi
(648)	DQTRGWF	(84)	tgKMSKSLrN	(464)	Q7UNZ2	- Rba
(562)	DQTRGWF	(29)	GRKMSKhLGN	(440)	Q9S2X5	- Sco
(587)	DQTRGWF	(29)	GkKMSKSLrN	(467)	AAS13180	- Tde
(555)	DQhRGWF	(29)	GRKMSKSLGN	(318)	P46213	- Tma
(576)	DQTRGWF	(29)	GqKMSKSkGN	(445)	AAS81050	- Tth
(556)	DQTRGWF	(29)	GkKMSKSLGN	(455)	Q8TN62	- Mac
(622)	DQiRGWF	(29)	GRKMSKSLGN	(348)	AAR39083	- Neq
(495)	DQlRGWF	(29)	GReMhKSLGN	(445)	Q9UXB1	- Sso

Omnipresent cassettes

(1) ABC transporters

(32-72)**GPSGSGKTTLL**(29-41)MVFQNYALFPHLTALENV(31-42)Q**LSGGQQQRVAIARAL** (6) L**LADEPT**SALD(21-22)IYVTHDQ(28-263)

(2) Proteases (cell division protein FtsH, zinc-dependent metalloprotease)

(146-463)LLV**GPPGTGKTLLA**RAVAGEA (7) SGSDFVEMFVGVGASRVRD (9) PCII**FIDEID**AVGR(7-11)DEREQTLNQLLVEMDGF

(3) RNA polymerase beta′ (gamma) subunit

LDGGRFATSDLNDLYRRVINRNNRLK 12 RNEKRMLQEAVDAL 25-33 GKQGRFRQNLLGKRVDYSGRSVIVVGP
59-84 HPVLLNRAPTLHRLGIQAF 18 AF**NADFDGD**QMAVH

(4) Initiation factor 2

M**GHVDHGKTTLV** 11 EAGGITQHIGA 12-29 F**IDTPGH**EAFT 14 LVVAADDGV 21 INKIDLP

(5) Elongation factor G

GIMAHIDA**GKTT**TTERIL 22-26 ERERGITIT 12-27 INI**IDTPGHV**DFTxEVERSLRVLDGAV 13 ETVWRQA

(6) tRNA synthase (isoleucine synthases and class I synthases)

(495-671) DQTRGWF(29-84)GR**KMSKSL**GN(318-467)consensus

Two most widespread modules ALEPH and BETH, apparently, represent the earliest duplex gene
that encoded in the earliest past two vitally important activities involved in energy supply (ATP binding and ATP-ase).

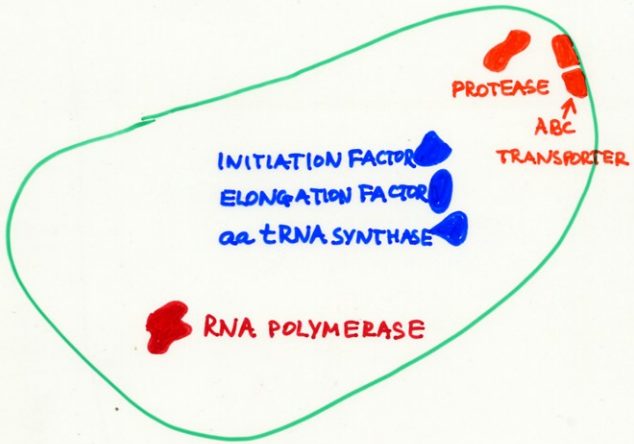
Today the module ALEPH is located in a variety of enzymes that require ATP, including the most ancient ones:

1. ABC cassettes of transporters,
2. cell division proteins (proteases),
3. initiation and
4. elongation translation factors.

Other most ancient enzymes are

5. RNA polymerase and
6. Amino acyl tRNA synthetase

THE OLDEST COMBINATIONS
OF THE OMNIPRESENT MOTIFS
ARE FOUND IN 6 PROTEIN TYPES



THE EARLIEST CELLS
WERE ELEMENTARY SUPPLY DEPENDENT.
(transport and digestion of external peptides
and amino acids)

Functional definition of LUCA:

Early organism that contained
functionally unique
omnipresent cassettes
and functionally unique
omnipresent singular modules

HVDHGKTTL Elongation factor EF-TU
GHVDHGKT Elongation factor EF-TU
GSGKTTLL ABC transporters (UraD)
GKSTLLN ABC transporters
SGSGKT Amino acid ABC transporters
GPSGSGK Amino acid (glutamine) ABC transporter
NGSGKTT ABC transporters
KSTLLK ABC transporters
GPPGTGKT Cell division control protein
GVGKTT ParA (chromosome partitioning) family protein
PGVGKT Clp protease, ATP binding
GKTTLA Holiday junction DNA helicase RuvB
PTGSGKT General secretion pathway protein
TGSGKS Twitching motility protein
PNVGKS GTP-binding protein era
GKTTLV GTP-binding protein TypA
DHGKST GTP-binding protein LepA
GTGKTT Signal recognition particle receptor protein

LSGGQQQRV ABC transporters, ATPases
QRVAIARAL ABC transporters, ATPases
TLSGGE ABC transporters, ATPases
LADEPT ABC transporters, ATPases

IDTPGHV Elongation factors G
NADFDGD DNA-directed RNA polymerases
WTTTPWT Isoleucyl-tRNA synthetases
KMSKSL Amino acyl tRNA synthetases, class I
FIDEID Cell division proteins

None of the omnipresent motifs
is involved in elementary syntheses.

ATP binding and breaking up,
peptide digestion,
membrane transport
and template functions only

Most of the singular omnipresent modules
are involved in many different multimodular activities.

For complete functional characterization of LUCA
one has to determine
what are specific functions of the
omnipresent modules themselves

GENOME SEGMENTATION

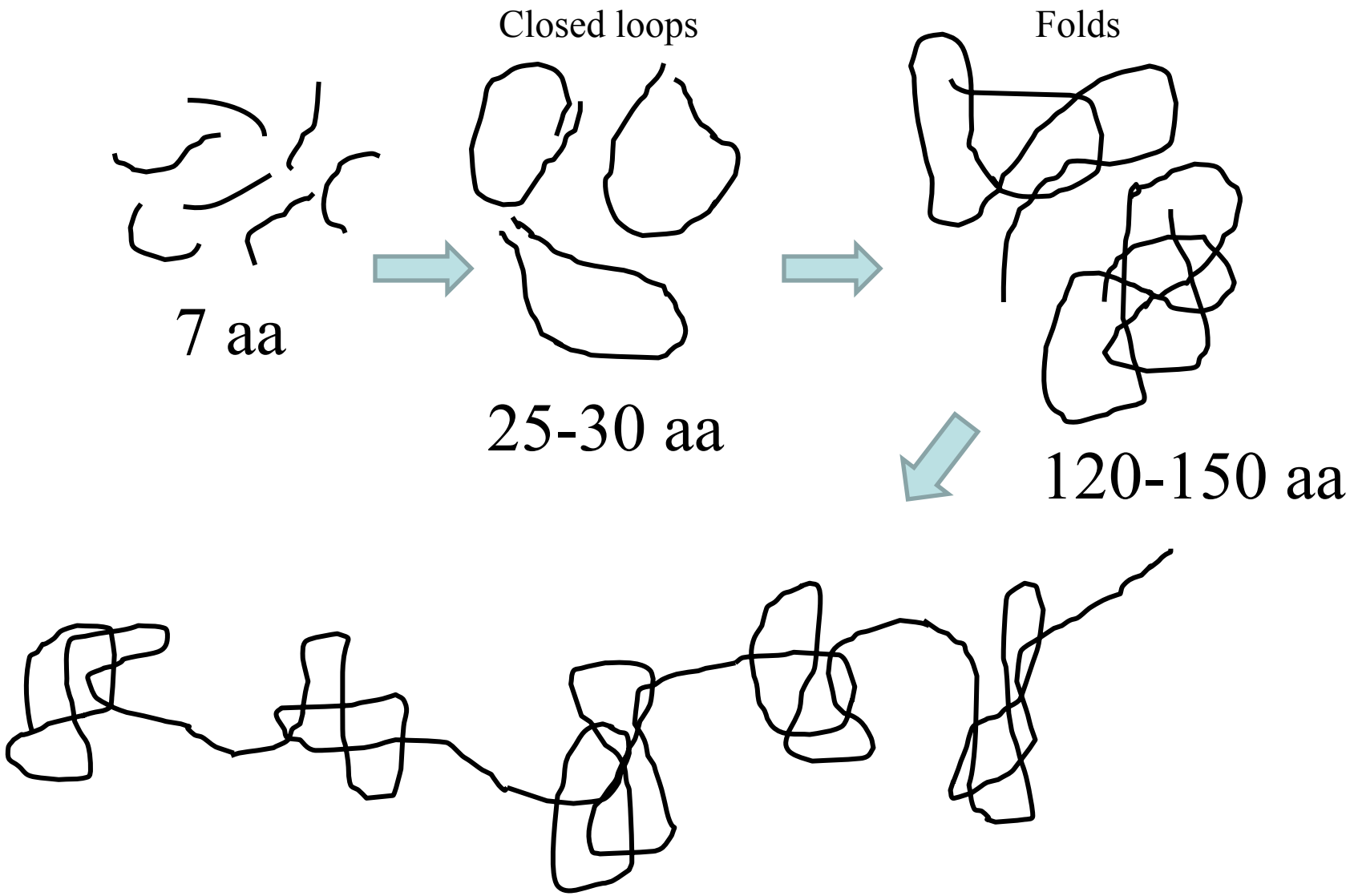
TABLE 2. Genome Unit Sizes Estimated by Various Techniques

	Prokaryotes		Eukaryotes		
	aa	bp	aa	bp	Ref.
Protein chain lengths	148–156	444–468	120–126	360–378	22
Periodical methionines	135–155	405–465	114–126	342–378	23
Curved DNA				350–370	25
Mobile DNA		400–440		335–355	24
Translation pausing	150–160	450–480	120–130	360–390	26,27
Mean	148 ± 3	444 ± 9	121 ± 2	363 ± 6	

“Evolution may have proceeded largely, rather than periferally, through extrachromosomal elements”

D. Reanney

Bact. Rev. 40, 552, 1976



Multifold proteins

Does complexity go together

with evolution of species?

YES Genome changes open
new opportunities, new niches

NO Loss of functions/structures
in parasites and symbionts

with evolution of biosphere?

YES speciation

NO extinction

Active **PATH SELECTION**

by life

(marching to all permissive niches and subniches)

VERSUS

Passive **NATURAL SELECTION**

by environment

(condemning unfortunate individuals and whole species in underpermissive conditions)

DEFINITIONS OF LIFE

"... if **variations** useful to any organic being ever do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life;
and from the strong principle of inheritance, these will tend to **produce offspring similarly characterized**“

Charles Darwin, Origin of Species (1859)

Rephrasing (ET):

Individuals with useful **variations** will **self-reproduce**

The essential criteria of life are twofold:
(1)the ability to direct chemical change by catalysis;
(2) the ability to reproduce by autocatalysis.
The ability to undergo heritable catalysis changes is general, and is essential where there is competition between different types of living things, as has been the case in the evolution of plants and animals (Alexander 1948).

Any system capable of replication and mutation is alive (Oparin 1961).

The criteria of living systems are: metabolism, self-reproduction and spatial proliferation.
The more complicated kinds also have the ability to mutate and evolve (G anti 1974).

We regard as alive any population of entities which has the properties of multiplication, heredity and variation (Maynard-Smith 1975).

Life is synonymous with the possession of genetic properties.
Any system with the capacity to mutate freely
and to reproduce its mutation must almost inevitably evolve
in directions that will ensure its preservation.
Given sufficient time, the system will acquire the complexity,
variety and purposefulness that we recognize as being alive
(Horowitz 1986)

To biologists, life is an outcome of ancient events that led to the assembly of nonliving materials into the first organized, living cells. ‘Life’ is a way of capturing and using energy and materials. ‘Life’ is a way of seeing and responding to specific changes in the environment. ‘Life’ is a capacity to reproduce; it is a capacity to follow programs of growth and development. And ‘life’ evolves, meaning that details in the body plan and functions of each kind of organism can change through successive generations (Starr and Taggart 1992).

Life is a self-sustained chemical system capable of undergoing Darwinian Evolution
(NASA working definition of life, Joyce 1994, 2002)

A living entity is defined as a system which, owing to its internal process of component production and coupled to the medium via adaptative changes, persists during the time history of the system (Luisi 1998).

Life on the Earth [*. . .*] seems to possess three properties (strongly related to each other and in fact being different aspects of the same thing) which are absent in inanimate systems. Namely, life is (1) composed of particular individuals, that (2) reproduce (which involves transferring their identity to progeny) and (3) evolve (their identity can change from generation to generation). A living individual is defined as a network of inferior negative Feedbacks (regulatory mechanisms) subordinated to (being at the service of) a superior positive feedback (potential of expansion of life) (Korzeniewski 2001).

Life is the process of existence of open non-equilibrium complete systems, which are composed of carbon-based polymers and are able to selfreproduce and evolve on the basis of template synthesis of their polymer components (Altstein 2002).

Life is defined as a system capable of 1. self-organization; 2. selfreplication; 3. evolution through mutation; 4. metabolism and 5. concentrative encapsulation (Arrhenius 2002).

Life is defined as a self-sustained molecular system transforming energy and matter, thus realizing its capacity of replication with mutations and anastrophic evolution (Baltcheffsky 2002).

Life is a chemical system capable of transferring its molecular information independently (self-reproduction) and also capable of making some accidental errors to allow the system to evolve (evolution) (Brack 2002).

Life is synonymous with the possession of genetic properties, i.e., the capacities for self-replication and mutation (Horowitz 2002).

A living entity is an ensemble of molecules which exhibit spatial organization and molecular-informational feedback loops in utilization of materials and energy from the environment for its growth, reproduction and evolution (Lahav and Nir 2002).

Any definition of life that is useful must be measurable.We must define life in terms that can be turned into measurables, and then turn these into a strategy that can be used to search for life. So what are these? a. structures, b. chemistry, c. replication with fidelity and d. evolution (Nealson 2002).

Life is a population of functionally connected, local, non-linear, informationally-controlled chemical systems that are able to self-reproduce, to adapt, and to coevolve to higher levels of global functional complexity (Von Kiedrowski 2002).

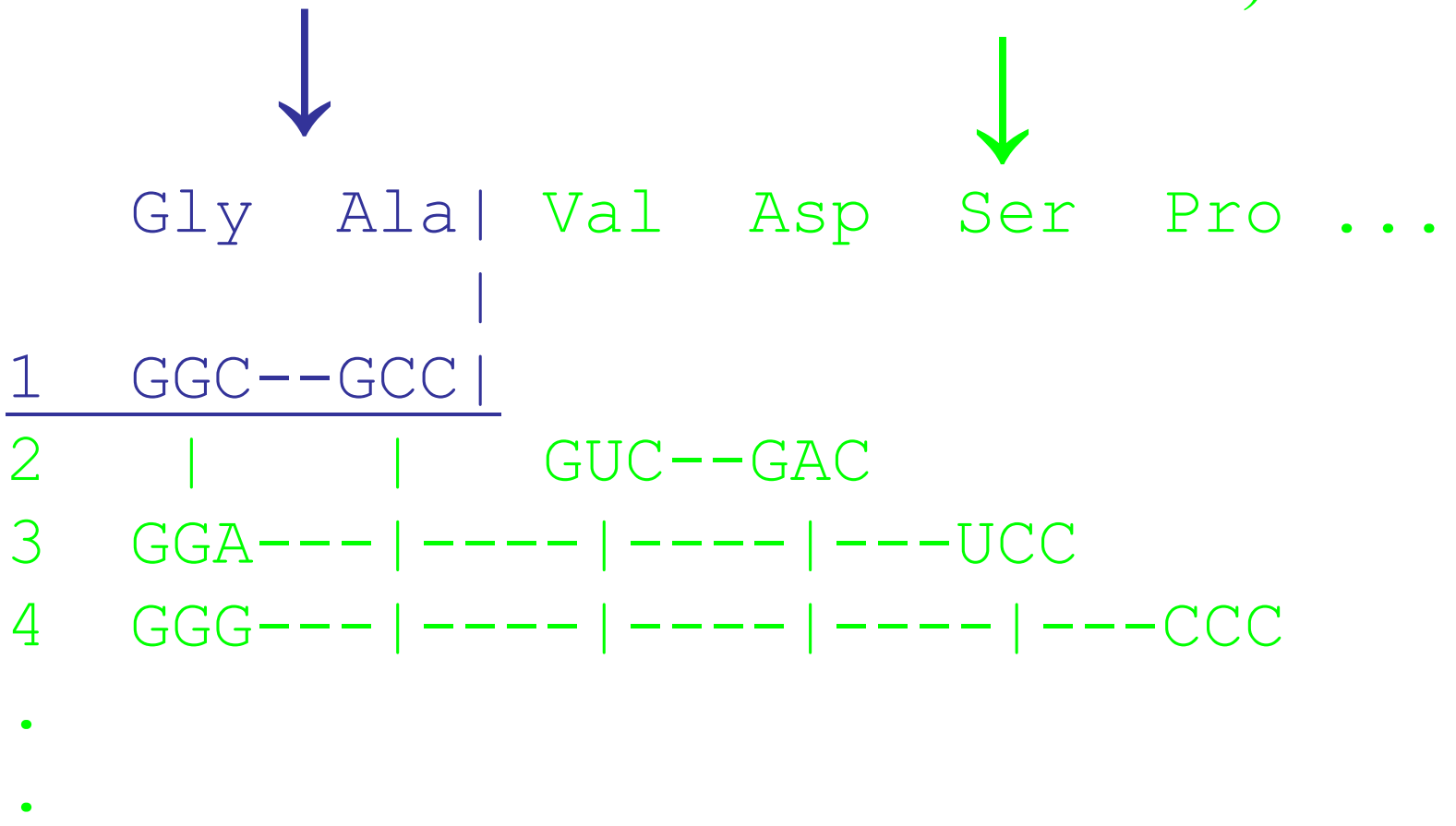
A living system is one capable of reproduction and evolution, with a fundamental logic that demands an incessant search for performance with respect to its building blocks and arrangement of these building blocks. The search will end only when perfection or near perfection is reached. Without this built-in search, living systems could not have achieved the level of complexity and excellence to deserve the designation of life (Wong 2002).

Rephrasing Darwin and all above:

Life is self-reproduction with variations

not Life yet
(self-reproduction only)

Life
(self-reproduction and variations)



WANTED

Self-reproducing composite replicon

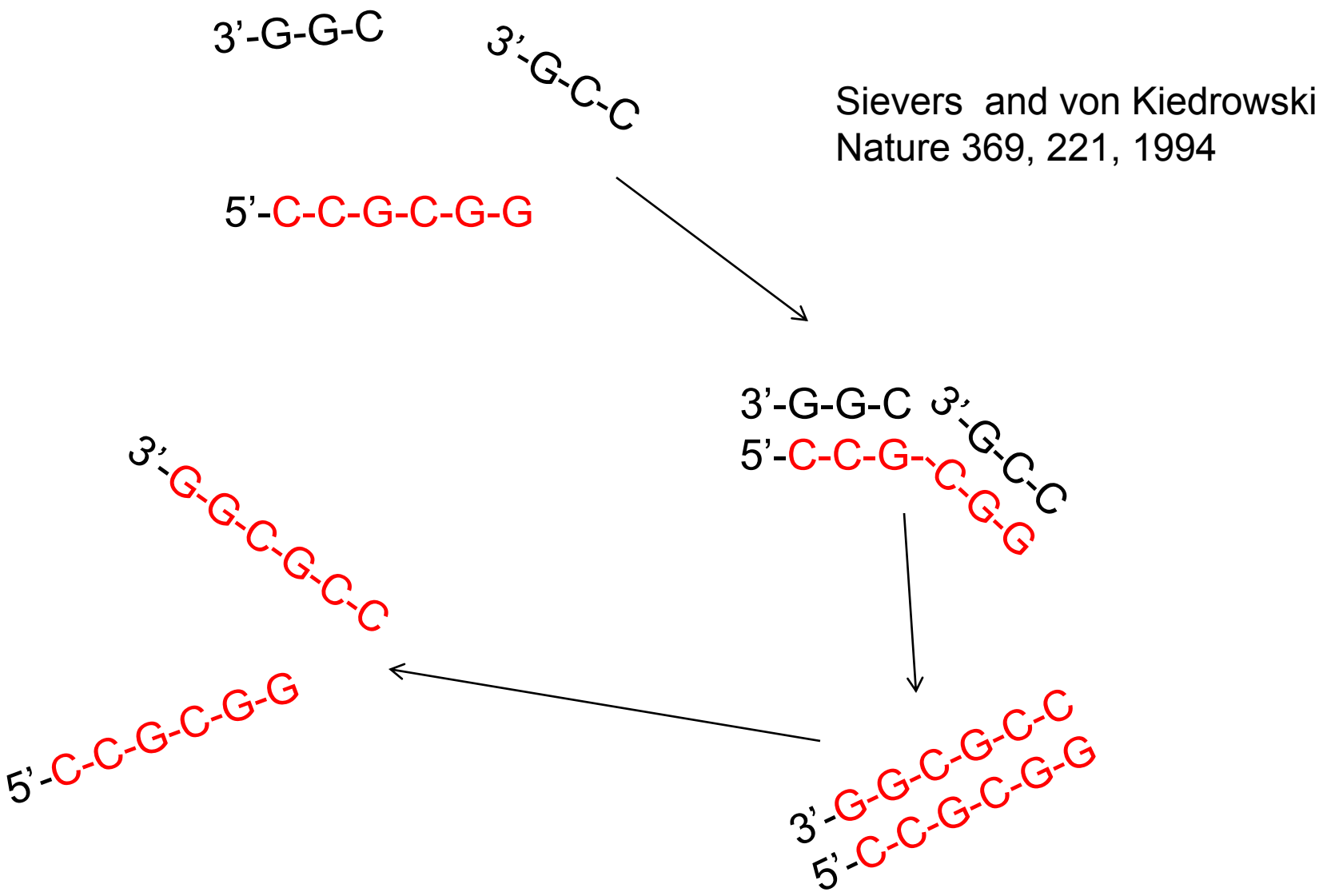
duplex of 5′-GCCGCCGCCGCCGCCGCCGCC-3′ **1**

and 3′-CGGCGGCGGCGGCGGCGGCGG-5′ **2**

and heptapeptides

ala ala ala ala ala ala ala **3**

gly gly gly gly gly gly gly **4**


3’-G-G-C
3’-G-C-C
5’-C-C-G-C-G-G
Sievers and von Kiedrowski
Nature 369, 221, 1994
3’-G-G-C
5’-C-C-G-C-G-G
3’-G-C-C
3’-G-G-C-G-C-C
5’-C-C-G-C-G-G
3’-G-G-C-G-C-C
5’-C-C-G-C-G-G

Another life before triplets

Well organized sequences GCC GCC GCC GCC….
and GGC GGC GGC GGC….

could not appear from nowhere.

Obviously, some other (simpler?) RNA molecules
had to come before.

This suggests that
the early biomolecular life, actually, started earlier,
before the triplet stage.

Moreover, one could speculate that
there were two lifes, one after another

The abiotic synthesis

of RNA (homopolyribonucleotides) in water

is experimentally established fact

(Di Mauro, 2009, 2010)

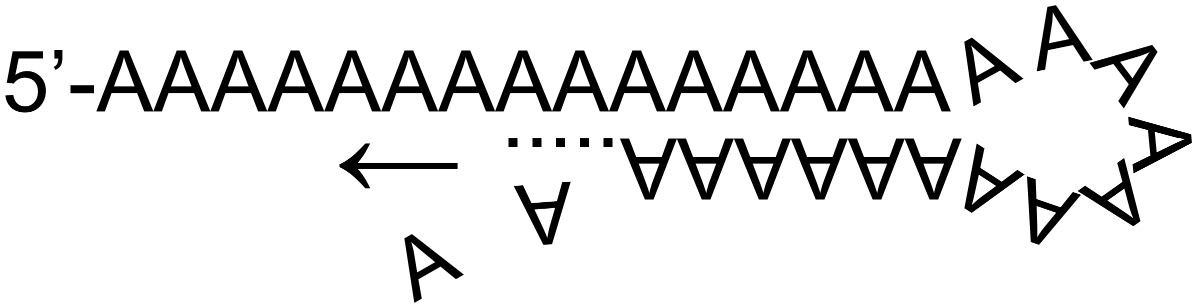
The abiotic synthesis of
5’-AAAAA….
stops at 5-mers, because
the degradation starts to dominate
over condensation

If, however, one starts with

hexamers or longer oligonucleotides

a magic thing happens:

the synthesis resumes
and continues to over hundred steps.



A•A complementary pairs are formed, first discovered by J. Brahms in 70s

Nature, thus, discovered the

complementary template synthesis,

although not Watson-Crick complementarity yet

In the above AAAAAAAAAAA… system
erroneous incorporation of
bases other than A
has lead to formation of a spectrum of
mixed sequence RNAs

The Watson-Crick pairing entered the scene

The competition started between the
replicating molecules

The simple repeating sequences took over

due to their ability to form slippage structures and expand

The champions of the slippage and expansion
GCC GCC GCC GCC ….
and GGC GGC GGC GGC …. appeared

This first pre-triplet life started with
primitive elongating homooligonucleotides
(self-reproduction),
went through the heterooligonucleotide stage
(self-reproduction and variation – LIFE),
and ended with, again, primitive simple repeats
(self-reproduction)
This was beginning of second life,
now with triplets and encoded amino acids

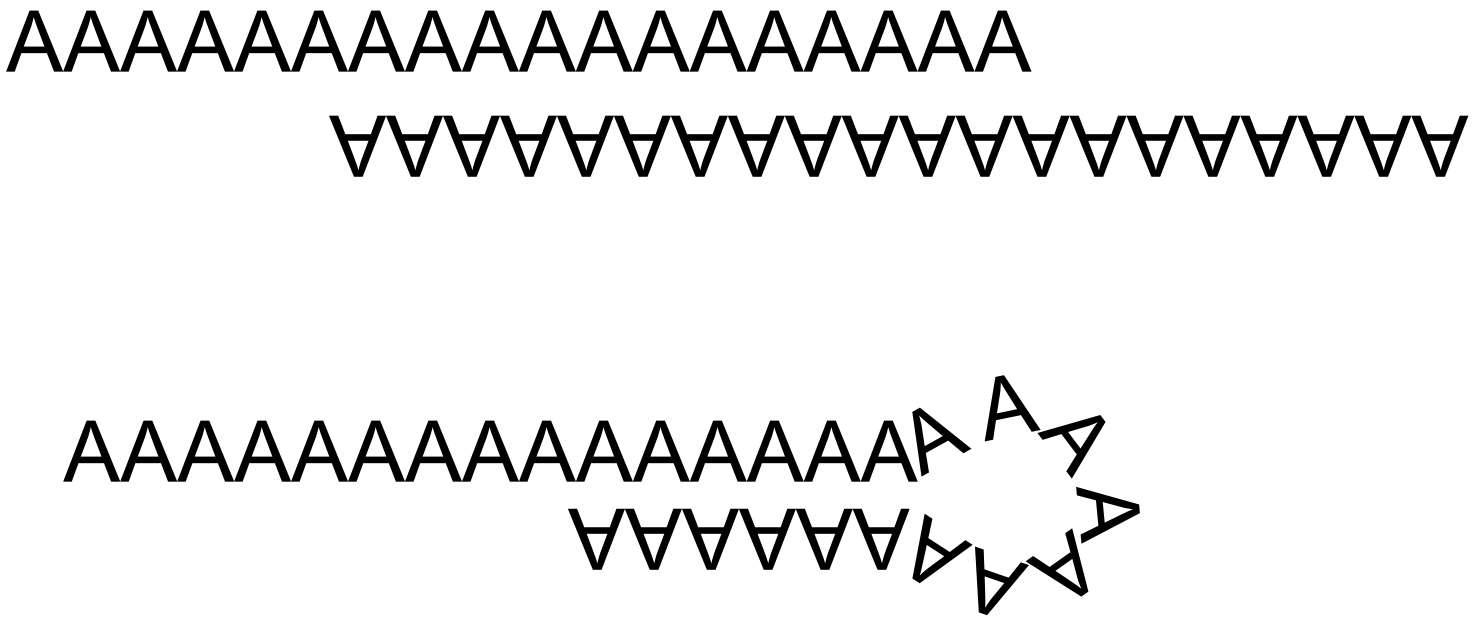
Major steps of early molecular evolution

I. Life before triplet code

1. Abiotic syntheses of monomers
2. Oligomerization, mixed sequence peptides, RNA oligonucleotides
3. Homooligonucleotides (polyA) take over, due to A•A complementarity
4. Inclusion of non-A bases, mixed sequences
5. Appearance of Watson-Crick pairs and takeover
6. Competition between RNA replicons, and appearance of simple repeats
7. GCCn•GGCn take over – first stage of the triplet code life

ACC
CCGG
UAG
CUUGGG
AAAA
AUAUCGC
AUGG
GAU
.....
CCUUGAG
GUCUU
UUU

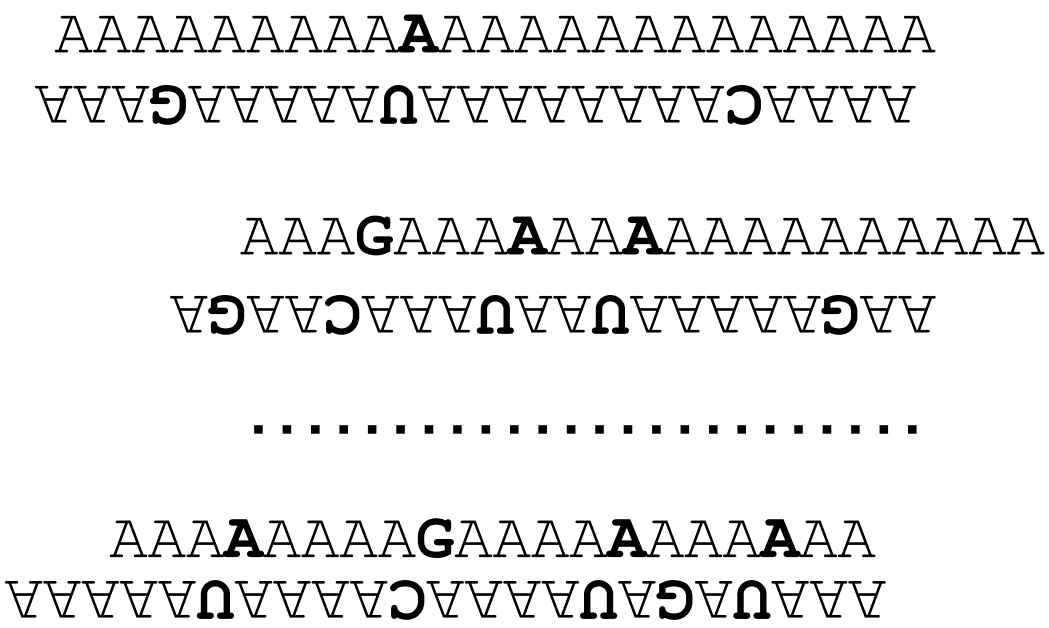
short
mixed
sequences



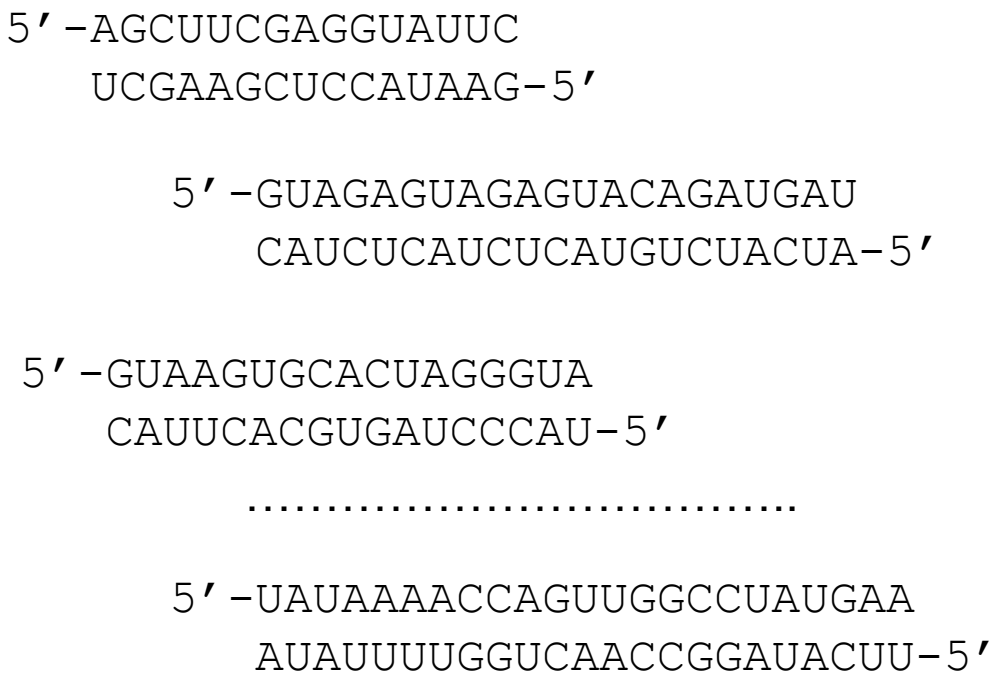
Hairpins and duplexes of oligoA.

Degradation barrier by-passed

Birth of complementarity



Development of Watson-Crick complementarity



Variety of mixed sequence complementary duplexes

(GAU)n•(AUC)n
(GU)n•(AC)n
(UAU)n•(AUA)n
(AAG)n•(CUU)n
(UUCC)n•(GGAA)n
(UC)n•(GA)n
....
(CUC)n•(GAG)n
(AUCG)n•(CGAU)n

variety of repetitive duplexes

5′-…GGCGGCGGCGGCGGCGGC…
CCGCCGCCGCCGCCGCCG…-5′

GGC•GCC duplexes.
Triplet life started.

II. Triplet code life

1. Appearance of first codons, in addition to GCC and GGC

2. First complementary mini-genes encoding peptides of 7 Ala-family residues and of 7 Gly-family residues

3. Fusion of minigenes, alternation of Ala-family and Gly-family units

4. Completion of the assignment of 64 codons to 17 amino acids and terminators

5. Codon capture stage, completion of modern codon table

6. Formation of closed polypeptide loops, first protein modules

7. Fusion of the early modules, formation of LUCA protein repertoire

8. Fusion of the genes encoding fold-size proteins, appearance of multi-fold proteins