Reconstruction of (bio)chemical life by interpolation from modern sequences to zero time

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

Sant Feliu 2009

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
A and G are encoded today by the most stable
and complementary codons GCC/GGC

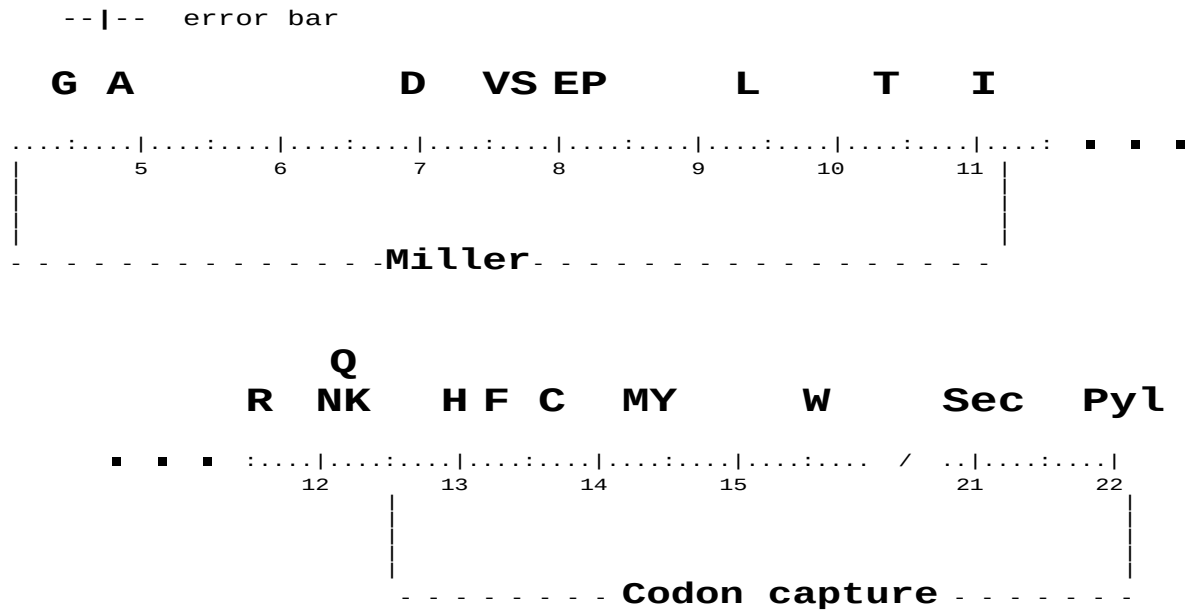
1987-92 Jaime Lagunez-Otero and ENT discovered that
consensus of mRNA is $(GCU)_n$

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

CONSENSUS TEMPORAL ORDER OF AMINO ACIDS

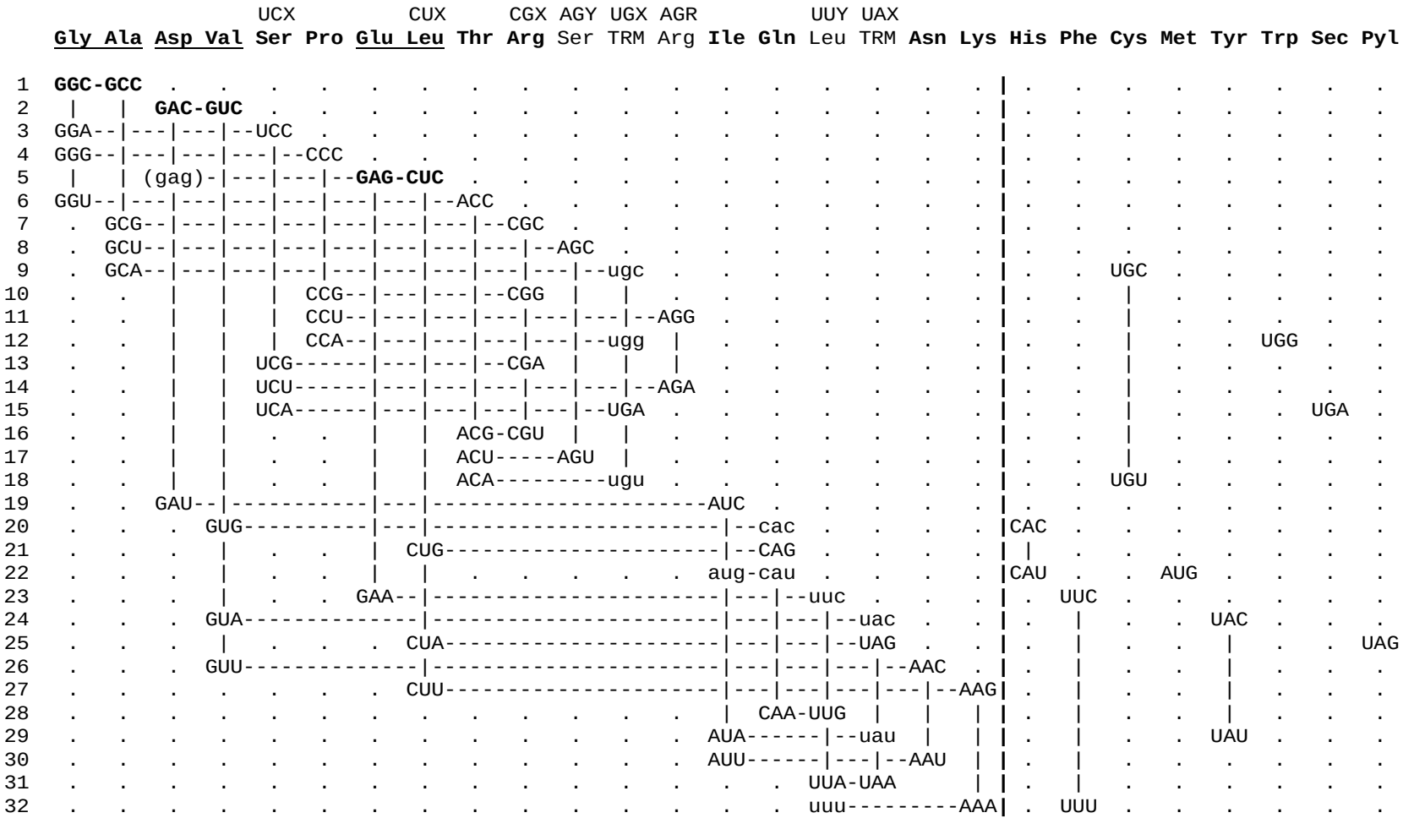
(101 VECTORS AVERAGED)



EVOLUTION OF THE TRIPLET CODE

E. N. Trifonov, December 2007, Chart 101

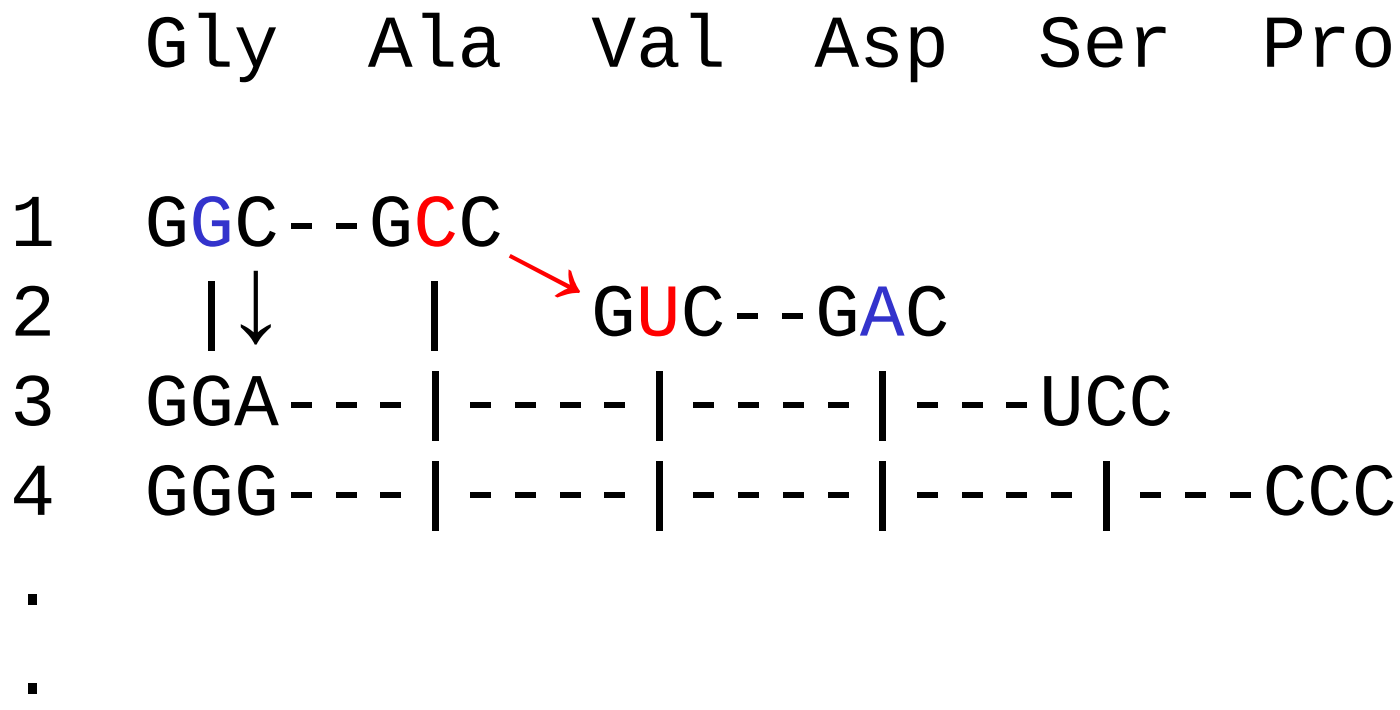
Consensus temporal order of amino acids:



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

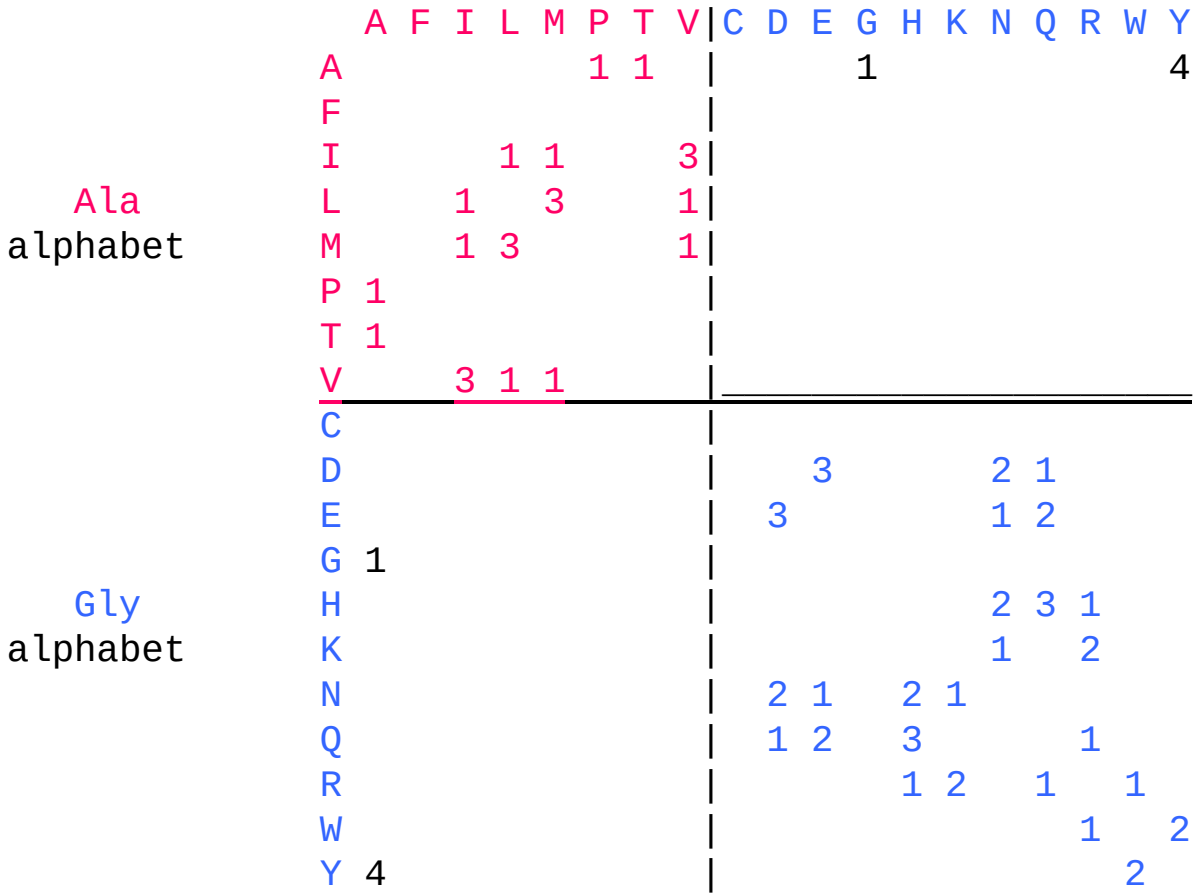


At every step of the evolution of the codons

middle purines remain purines (R→R),

middle pyrimidines remain pyrimidines (Y→Y).

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)

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
xʎx xʎx xʎx xʎx xʎx xʎx xʎx

At a later stage they fused in mosaics:

…*GGGGGGGAAAAAAAGGGGGGGAAAAAAA*…

The unit size is estimated to be 7 amino acid residues

(J. Mol. Evol. 53, 394-401, 2001; J Biomol Str Dyn 24, 163-170, 2006)

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

AFLIIMVRKREDQNFFVTAMAQQNEDGR

AFLIIMVRKREDQNFFVTAMAQQNEDGR

AAAAAAAGGGGGGGAAAAAAAGGGGGGG

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

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

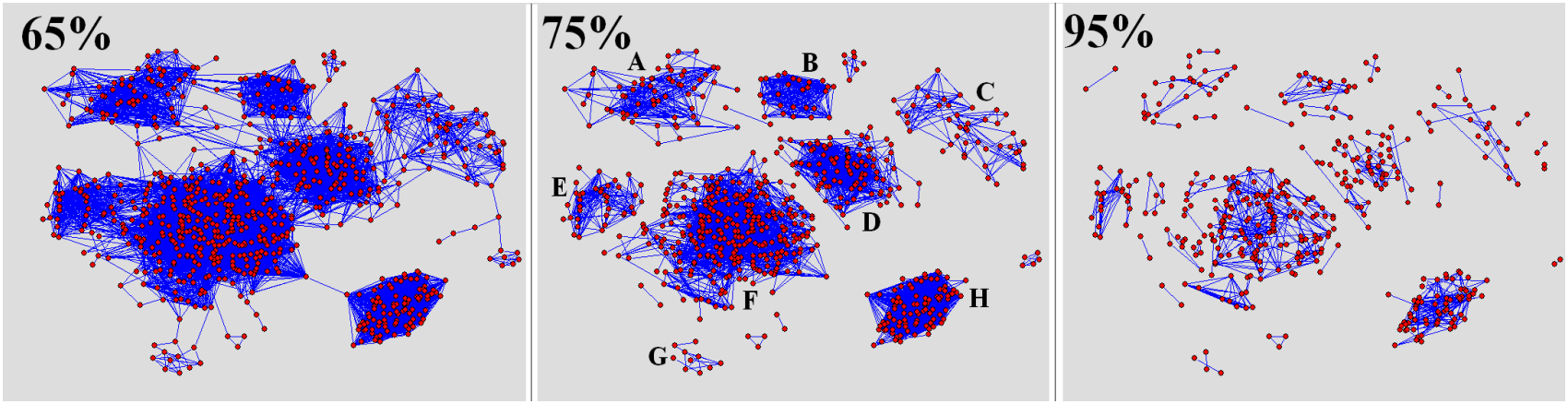
All 20 aa fragments of all proteins of prokaryotes make a

sequence space

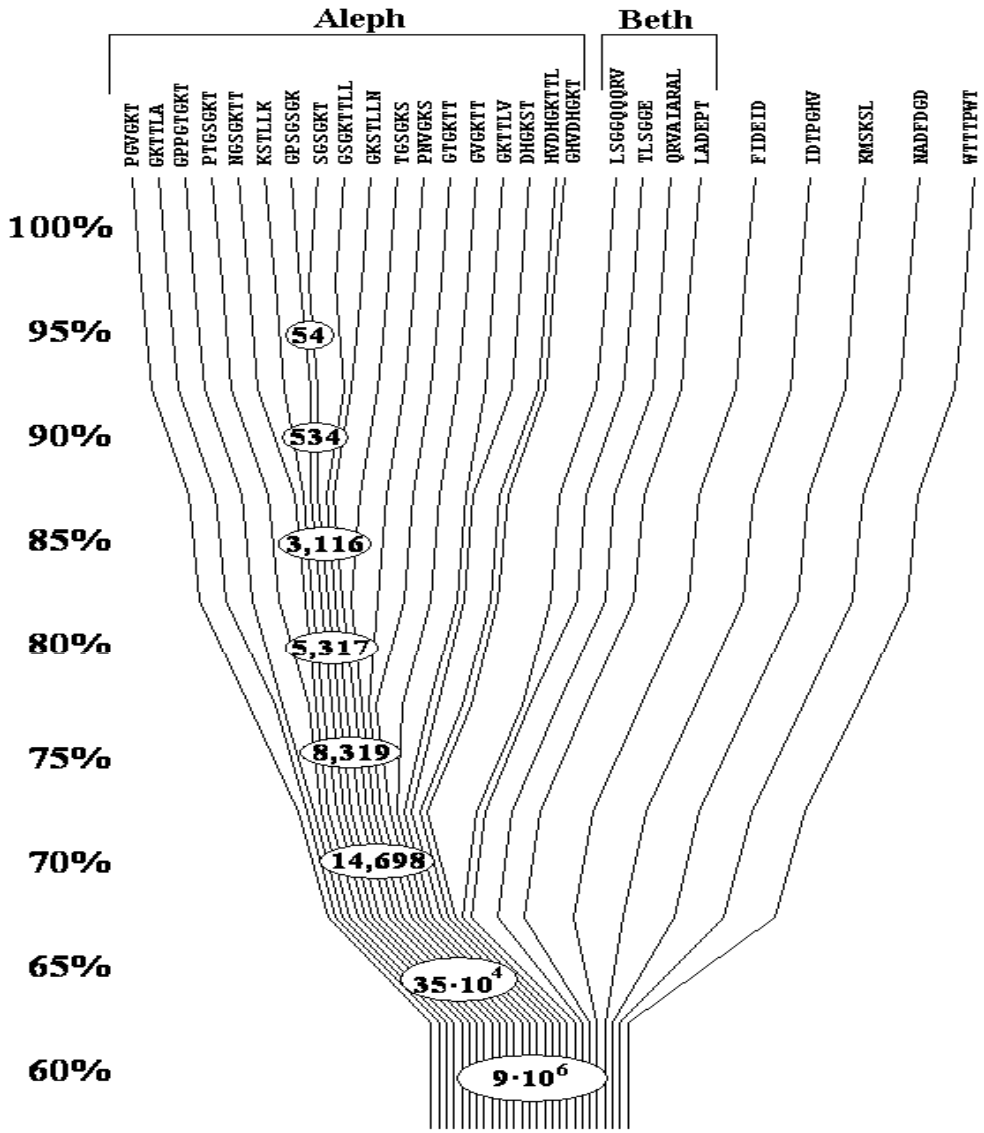
Those fragments that are close relatives (matching >60%)
are pair-wise connected. This makes

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

In binary form all 27 omnipresent motifs have common consensus prototype

AGAAGGAGGGGAAAAGAA

including extensions of Aleph and Beth:

AGAAGGAGGGGAAAAG	***Aleph***
AA*S*GGGGGGAAAAGAA	***Beth***

This explains the common tree for all omnipresent elements

TWO RECONSTRUCTIONS MEET

OMNIPRESENT ELEMENTS

RECONSTRUCTION OF ALEPH AND BETH

ALEPH: IDTPGHVDHGKTTLLnk

BETH: TLSGGqQQRVAIARALe

COMMON BINARY PROTOTYPE OF ALEPH AND BETH

(AAA)AGAAGGAGGGGAAAAGAA

AAAAAAA*|*GGGGGGG*|*AAAAAAA

AAAAGAA*|*GGAGGGG*|*AAAAGAA

AAAAAAAGGGGGGGAAAAAAA

BINARY MOSAIC

GGGGGGG* & *AAAAAAA

FIRST PEPTIDES

BINARY ALPHABET

EVOLUTIONARY CHART OF CODONS

Remarkably, the ALEPH and BETH are complementary:

ALEPH *AGAAGGAGGGGAAAAG*

| | | | | | | | | | | | | | | | -

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

Same for the reconstructed common prototype of Aleph and Beth :

AGAAGGAGGGGAAAAGAA

| | | | | | | | | | | | | | | | | | |

AGAAGGAGGGGAAAAGAA

Two most widespread modules ALEPH and BETH, apparently, represent the earliest duplex gene

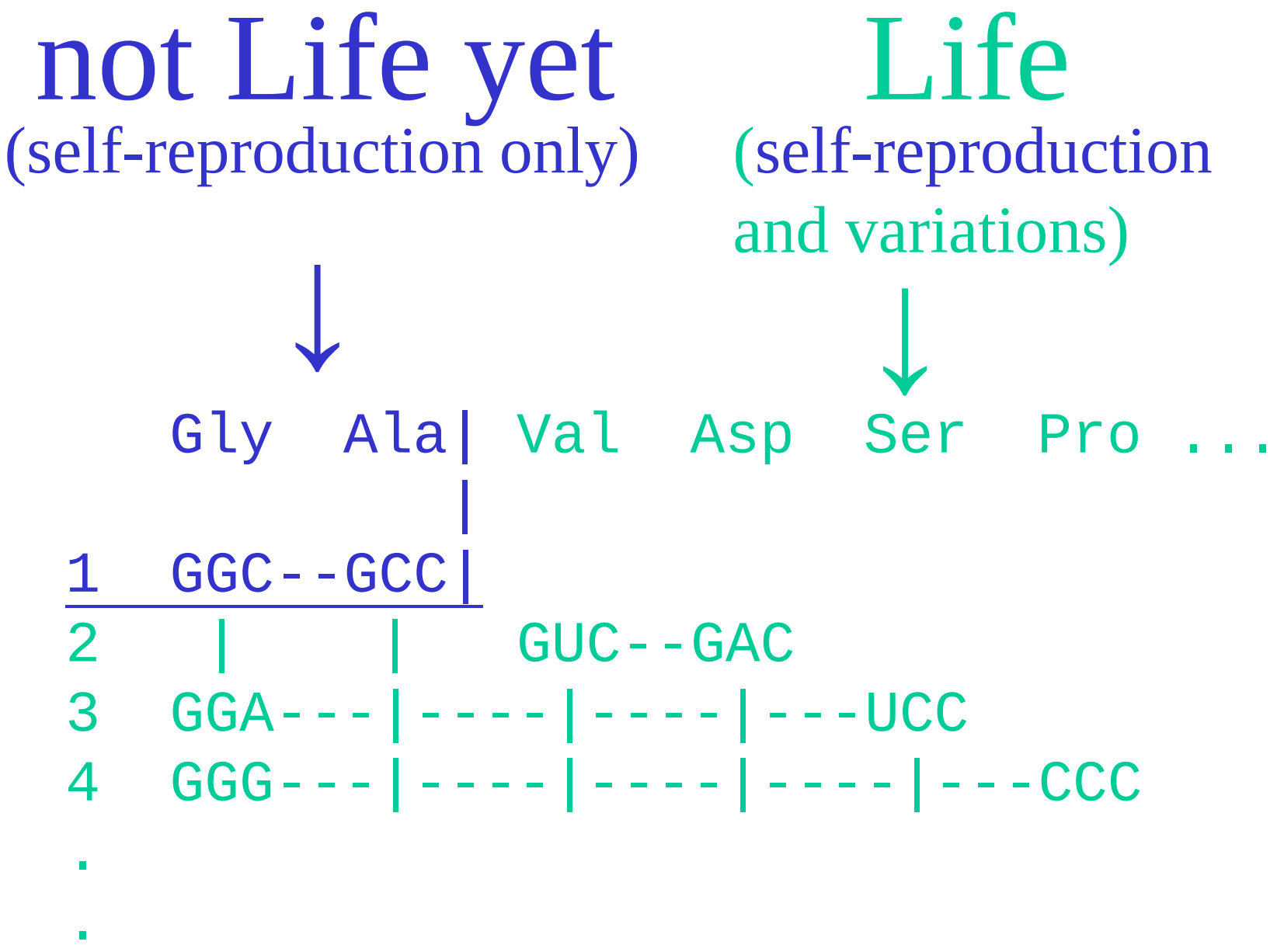
that encoded in the earliest past two vitally important activities involved in energy supply (ATP binding and ATP-ase).

"... 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 Darwin:

Life is self-reproduction with variations


not Life yet
(self-reproduction only)
Life
(self-reproduction
and variations)
Gly Ala| Val Asp Ser Pro ...
1 GGC--GCC|
2 | | GUC--GAC
3 GGA---|----|----|---UCC
4 GGG---|----|----|----|---CCC

WANTED

Self-reproducing four-component replicon

duplex of

5′-GCC GCC GCC GCC GCC GCC GCC-3′ **1**

and 3′-CGG CGG CGG CGG CGG CGG CGG-5′ **2**

and heptapeptides:

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

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

THANKS TO

Networks - Zacharia M. **Frenkel**
University of Haifa

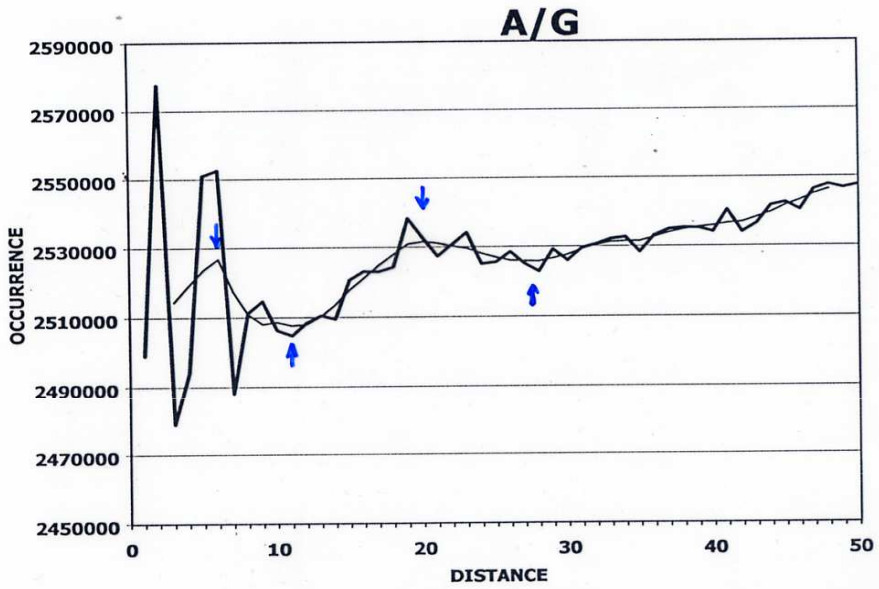
Omnipresent motifs - Yehoshua **Sobolevsky**
University Minas Gerais, Brazil

Modules – closed loops – Igor N. **Berezovsky**
University of Bergen, Norway

Ancestral complementarity Idan **Gabdank**
Danny **Barash**
Ben Gurion University,
Beer Sheva

AND TO THE AUDIENCE

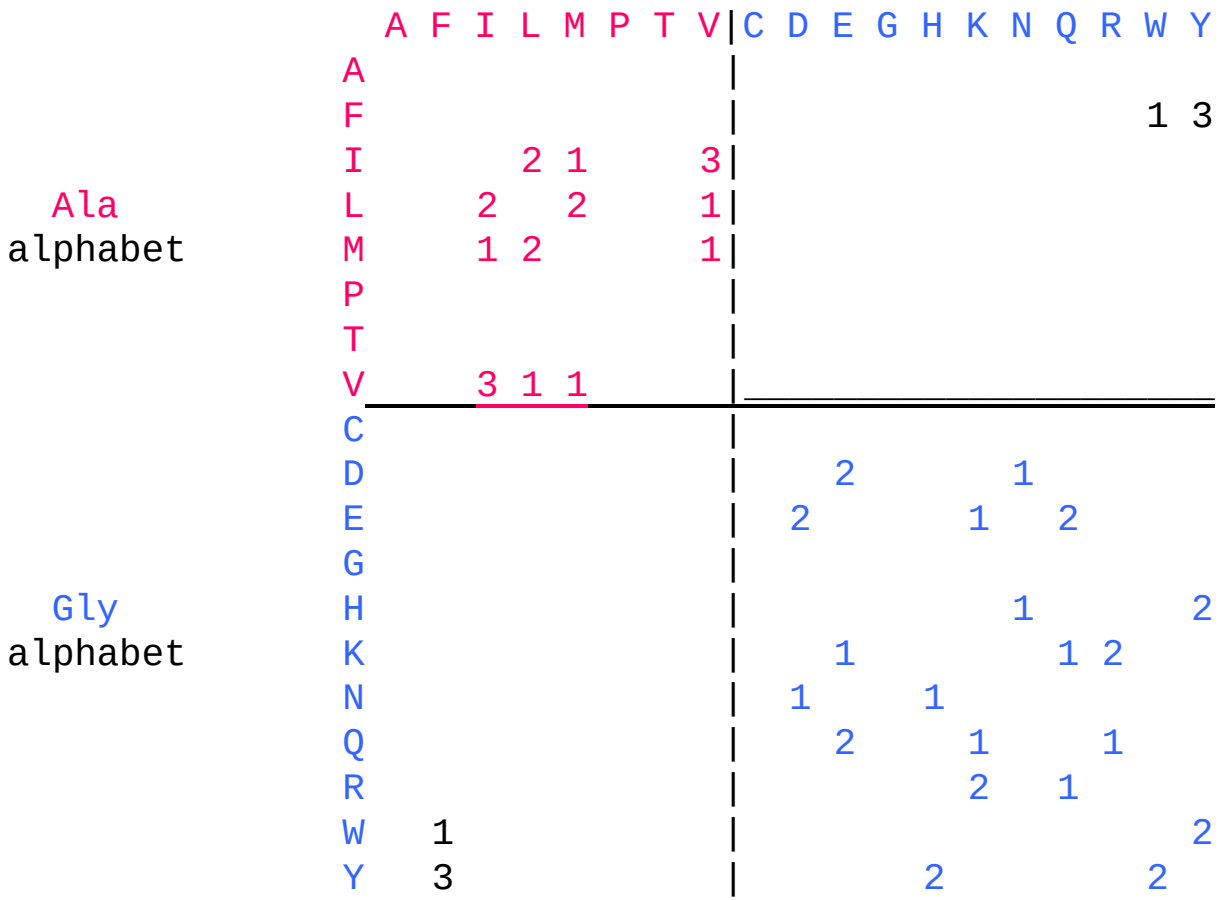
Support by:
Israel Science Foundation,
Center of Complexity Science, and
Masaryk University, Brno



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

2001

Kizzhner V
Kizzhner A
Berezovsky I



Rearranged BLOSUM substitution matrix

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

⟶ temporal order of amino acids ⟶

Gly Ala Asp Val Ser Pro Glu Leu Thr Arg TRM Ile Gln TRM Asn Lys

GGC GCC GAC GUC UCC CCC GAG CUC ACC CGC ugc AUC cac uac AAC AAG

⟶ descending thermostability of triplet pairs ⟶

Newcomers (codon capturers)

His Phe Cys Met Tyr Trp Sec Pyl

CAC UUC UGC AUG UAC UGG UGA UAG

Complementary symmetry properties of common prototype

AGAAGGAGGGGAAAAGAA

AAAAGAA GGAGGGG AAAAGAA

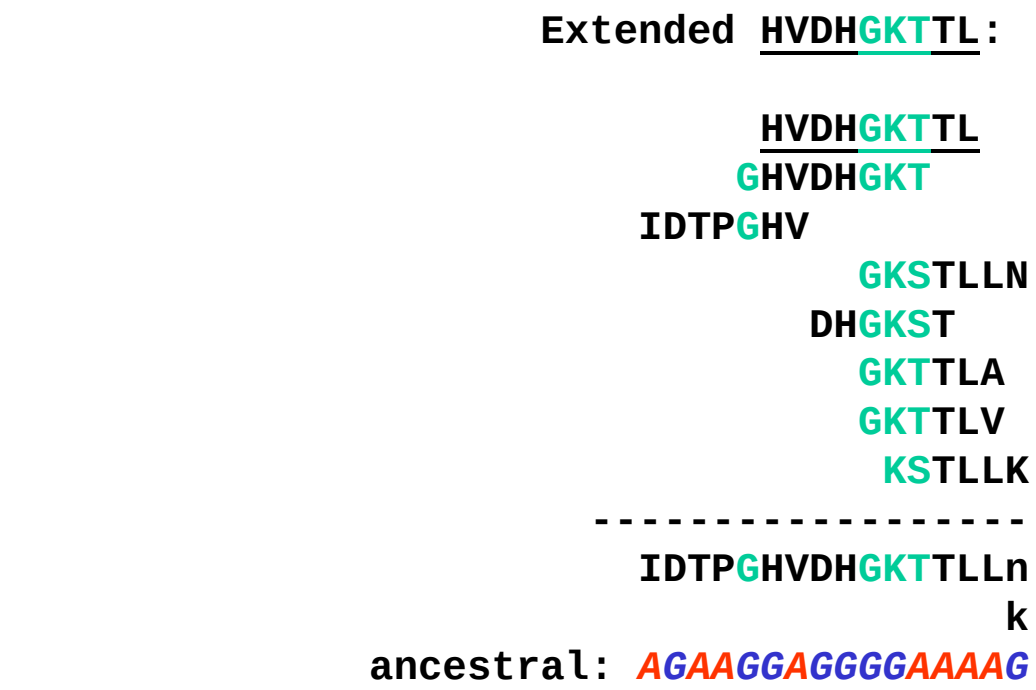
GGAGGGG AAAAGAA GGAGGGG

This is blunt end fusion of the same element

AAAAGAA

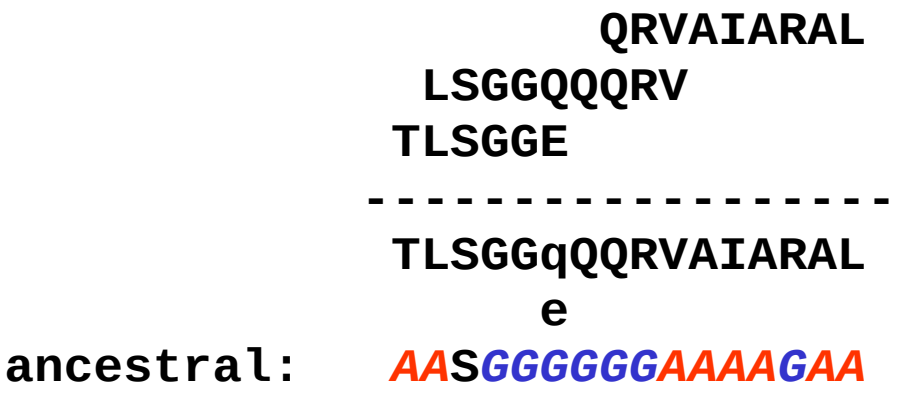
GGAGGGG

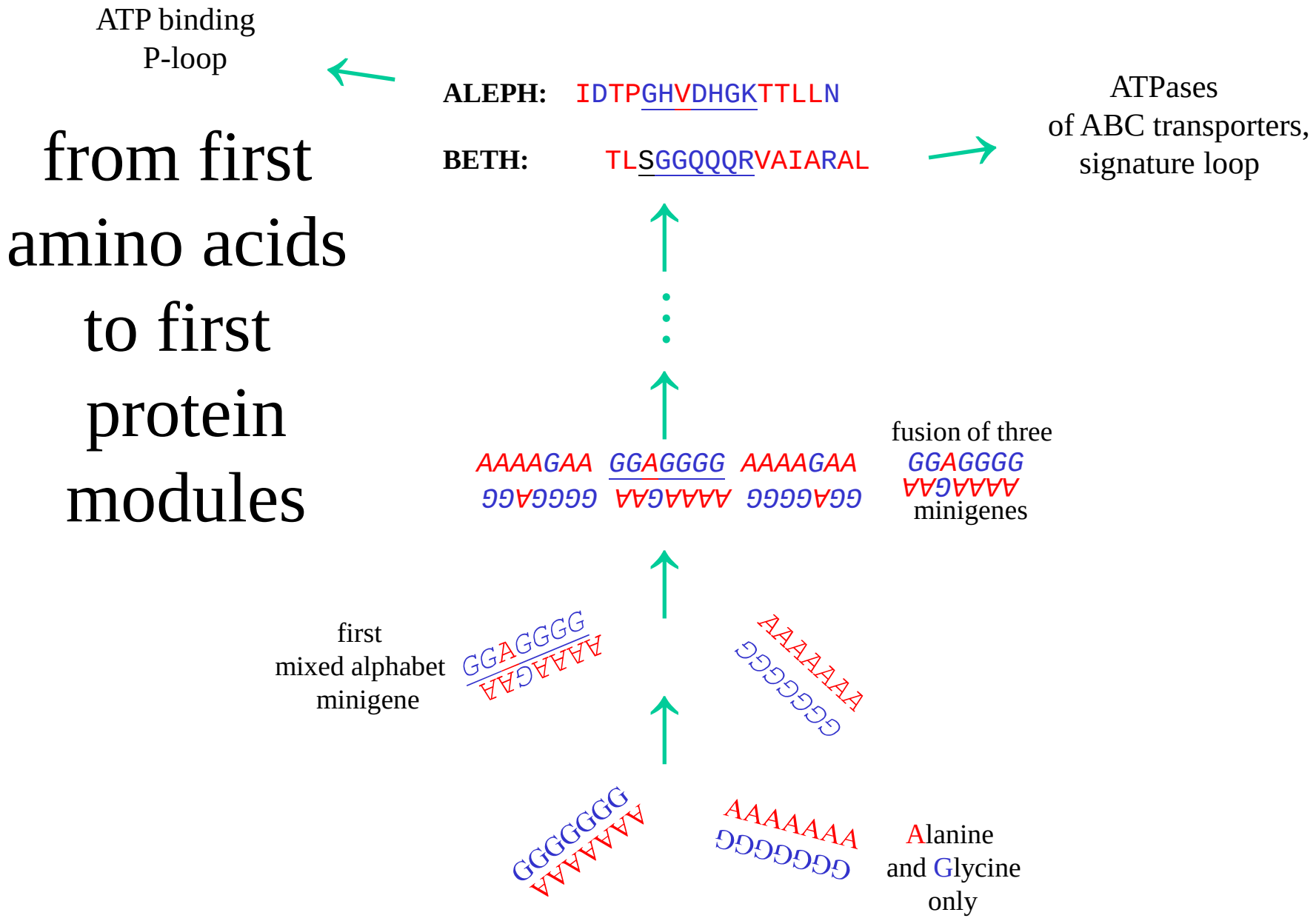
RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 1.



RECONSTRUCTION OF COMMON PROTOTYPE OF OMNIPRESENT ELEMENTS. Step 2.

Extended QRVAIARAL and LSGGOOORV:




ATP binding
P-loop
ALEPH: IDTPGHVDHGKTTLLN
BETH: TLSGGQQQRVAIARAL
ATPases
of ABC transporters,
signature loop
from first
amino acids
to first
protein
modules
fusion of three
minigenes
first
mixed alphabet
minigene
Alanine
and Glycine
only

GCC – codon for alanine (A),

GGC – codon for glycine (G).

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

Среди болезней триплетной экспансии

самые распространенные
вызываются повторами

…GCCGCCGCCGCCGCCGCCGCCGCC…

!

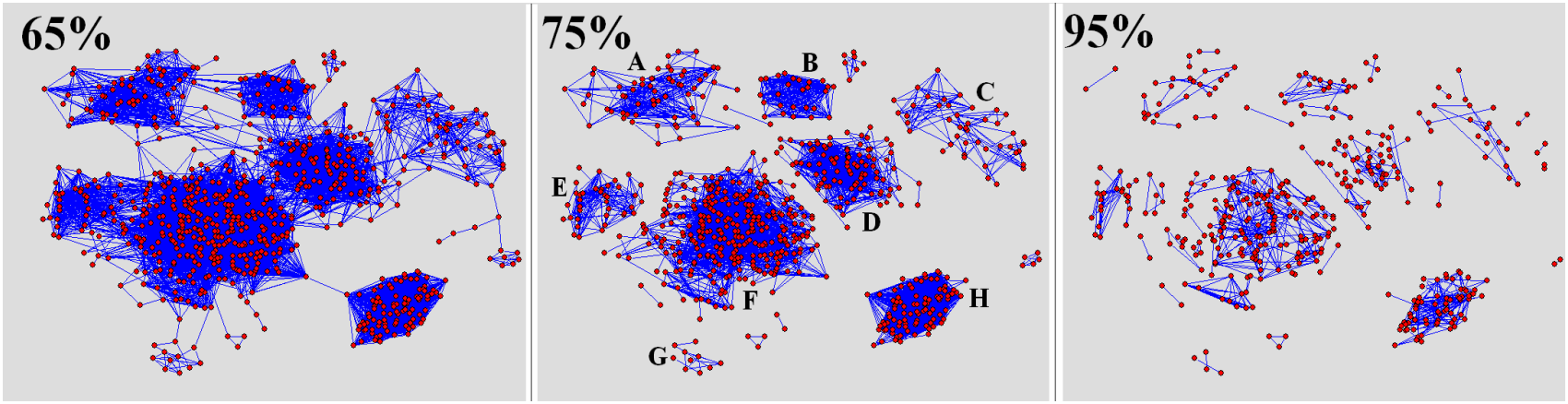
All 20 aa fragments of all proteins of prokaryotes make a

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

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*


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

Proteases (cell division proteins FtsH)

GPP (Aleph) **FVE** **FID**

(197) LLVGPPGTGKTLLARAVAGEA(7)SGSDFVELFVGVGAARVRD(9)PCIVFIDEIDAVGR (10) 2CEA

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

DER **RPG**

DEREQTLNQLLVEMDGF(8)MAATNRPDILDPALLRPGRFDKK (297) 2CEA

DEREQTLNQLLVEMDGF(8)IAATNRPDxLDPALLRPGRFDRQ (95-415) consensus

- another example of the omnipresent cassette

Omnipresent cassette of RNA polymerases

FAT **NEK** **NLL**

(529) VDGGRFATSDLNDLYRRLINRNNRLK (12) RNEKRMLQEAVDAL (27) GKQGRFRQNLLGKRVDYSGRSVIVVGP 2A6E

(224-518)LDGGRFATSDLNDLYRRVINRNNRLK (12) RNEKRMLQEAVDAL(25-27)GKQGRFRQNLLGKRVDYSGRSVIVVGP consensus

VLL **NAD**

(62) KVVLLNRAPTLHRLGIQAF (18) AFNADFDGDQMAVH (776) 2A6E

(59-84)HPVLLNRAPTLHRLGIQAF (18) AFNADFDGDQMAVH (131-961) consensus