

Analysis of Protein Evolution Group



Head:

Dr. Vincenzo E. A. Russo

Phone: +49 (0)30-8413 1264

Fax: +49 (0)30-8413 1394

Email: russo@molgen.mpg.de

Scientist:

Yean-Su Lee (until 1/98)

Technician:

Uta Marchfelder (until 6/01)

About the tree of life

My work until June 2001 was mainly concerned with the moss *Physcomitrella patens* where we found, for the first time in a land plant, a way to obtain high homologous recombination. Since July 2001 I have been working alone on a fascinating problem of theoretical biology, the Tree of Life.

Charles Darwin was the first to suggest that all living organisms are descended from one common ancestor (“The origin of the Species”, 1859). The exponential growth of the number of completely sequenced genomes, today 140 Bacteria/Archaea and 13 Eukarya (and these numbers probably will double in the next year), provided a great hope to realize the dream of Darwin, namely to identify LUCA (Last Common Cellular Ancestor). Until now, however, there is little consensus regarding LUCA except that it was living circa 3.4-3.8 billions years ago.

1) *Russell Doolittle et al.* (Determining divergence times of the major Kingdoms of living organisms with a protein clock (1996), *Science* 271:470-477) suggested that LUCA was a Eubacterium.

2) *William Martin* (Mosaic bacterial chromosomes: a challenge en route to a tree of genomes (1999), *BioAssays* 21:99-104) draws a Tree of Life which has two roots: the Eubacteria and the Archaeobacteria. The Eukaryotes are then a complicated mixture of the two Kingdoms.

3) *Forterre and Philippe* (Where is the root of the universal tree of life? (1999), *BioAssays* 21:871-879) argue that the very first cell was a Eukaryote.

4) *Woese* (On the evolution of cells (2002), *PNAS* 99:8742-8747) states that “Extant life on Earth is descendent not from one, but from three distinctly different cell types. However, the designs of the three have developed and matured in a communal fashion”.

Despite these models and a plethora of phylogenetic trees and bioinformatic analysis published to date, there is no detailed information on the evolution of proteins in well known biosynthetic pathways. Without this information I believe that it will be not possible to fully understand evolution.

Making the bold assumption that *Homo sapiens* is at the top of the evolutionary tree, I asked if the human proteins of important cellular pathways (transcription, translation,



DNA synthesis, lipid biosynthesis, glycolysis, biosynthesis of amino acid, purines, pyrimidines) are more similar to the equivalent of Eubacteria, or of Archaea, or equal similar to both, or have no counterpart in Prokaryotes.

The technique I have employed is simple: I blasted each of the 281 human proteins involved in these pathways, on one hand against the proteins data base and on the other hand against the genomic sequences of the completely sequenced Eubacteria and Archaea genomes (both kind of databases at NCBI in Washington).

Preliminary data are summarized in table 1. It is immediately apparent that different proteins have different origins, however it is not a random process, but seems to follow a pattern, suggesting a logical choice in evolution.

The eukaryotic genome was shown already to have a mosaic structure (Horiike T. et al. (2001), *Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis*, Nature Cell Biology 3:210-214). In this paper there is also a table where thousands of eukaryotic genes (*S. cerevisiae*) in 43 pathways were classified as of Archaea or of Eubacteria origin. However this table is often misleading, as I will show below with one example:

Under *amino-acid metabolism* all the 201 proteins considered to be in this pathway are reported to be of Eubacteria origin. In contrast, my table shows that only 7 out of 16 of the enzymes analyzed are of Eubacteria origin while 7 are of Archaea AND Eubacteria origin (common origin), and for two is difficult to make a decision. Two examples out of those 7 proteins of common origin:

a) The human **3-phosphoglycerate dehydrogenase**, an enzyme of the serine biosynthetic pathway, has 533 amino acids; the sequence of this protein blasted against all Archaea proteins show 43% identity (ID) with a *M. jannashii* protein, 524 amino acids (aa) long, over 447 aa, 45% ID with a *M. acetivorans* protein (523 aa) over 405 amino acids and 41% ID with a *A. fulgidus* protein (527 aa) over 401 amino acids; blasted against all Eubacteria proteins show a 43% ID with a *M. loti* protein (533 aa) over 399 aa, 42% ID with a *B. subtilis* protein (525 aa) over 416 aa and 42% ID with a *B. melitensis* protein (538 aa) over 400 aa.

	N° of proteins				Total
	of Archaea origin	of Eubacteria origin	of Archaea & Eubacteria origin	"new"	
Transcription					
RNA pol I	6	0	0	0	6
Transcription Factors	1	1	1	13	16
RNA pol II	9	1	0	2	12
Transcription Factors & mediators	1	0	0	19	20
RNA pol III	6	0	0	5	11
Translation					
Ribosomal proteins large subunit	38	0	1	4	43
Ribosomal proteins small subunit	28	0	0	3	31
Initiation factors	7	0	6	5	18
Elongation factors	2	0	0	3	5
Release factors	2	0	0	0	2
aa-tRNA synthetases	6	5	8	0	19
Protein folding machine: Chaperonin CCT	8	0	0	0	8
DNA polymerases	3	1	3	16	23
Biosynthesis of lipids					
Fatty acid synthetase	0	5	0	0	5
Triacyl-glycerols	0	3	0	2	5
Biosynthesis of small molecules					
glycolysis	0	8	1	0	9
Biosynthesis of nonessential aminoacids	0	7	7	0	16
Biosynthesis of purine	0	6	3	0	13
		(1)	(3)		
Biosynthesis of pyrimidine	0	(1)	6	0	7
TOTAL	117	39	41	72	269
Mitochondrial proteins as control					
Krebs cycle	0	9	2	1	12

Table 1: Origin of nuclear-coded cytoplasmic proteins from selected biochemical pathways of *H. sapiens*

b) The human **serine hydroxymethyltransferase**, an enzyme that catalyzes glycine from serine, has 483 aa; the sequence of this protein blasted against all Archaea proteins show 46% ID with a *M. mazei* protein (419 aa) over 402 aa, 43% ID with a *Halobacterium* protein (424 aa) over 415 aa and a 32% ID with a *A. fulgidus* protein (451 aa) over 444 aa; blasted against all Eubacteria proteins show a 47% ID with an *A. tumefaciens* protein over 440 aa, 49% ID with a *T. maritima* protein (427 aa) over 389 aa, and 47% ID with a *C. acetobutylicum* protein (411 aa) over 406 aa.

In each of these cases it would be incorrect to conclude that the human genes evolved from either the Eubacteria or the Archaea for several reasons: 1) the length of the Eubacteria or Archaea proteins most homologous to the human pro-

tein is very similar to each other; 2) The percentage of identity is very similar over a long stretch of amino acids of comparable length in both cases; 3) The best hits with Eubacteria are with bacteria coming from very divergent families like *bacillus* (*B. subtilis*), *clostridium* (*C. acetobutylicum*), *rhizobiaceae* group (*A. tumefaciens*, *M. loti*, *B. melitensis*), *thermotogales* (*T. maritima*). The best explanation for these results to me is to assume that these particular proteins are very ancient and have developed to a form of “perfection” and have remained so in the three different kingdoms.

A similar analysis reported in my table 1 show, for example, that 6 enzymes out of 7 enzymes of the biosynthetic pathway of pyrimidines are of Archaea & Eubacteria origin contrary to the results published by Horiike et al. who states that nucleotide metabolism enzymes are all of Eubacteria origin; 8 out of 19 aa-tRNA synthetases were found to be of Archaea & Eubacteria origin, 5 are of Eubacteria origin, and only 5 are of Archaea origin, while the just quoted authors state that all the enzymes for protein biosynthesis are of Archaea origin. However, there are human proteins that have much higher identity to Archaea than to Eubacteria proteins, such as the great majority of ribosomal proteins.

Similarly there are clearly human proteins that have much higher identity to Eubacteria proteins than to Archaea proteins, such as the 8 enzymes involved in glycolysis.

My studies to date are a warning against quick bioinformatic analysis. I believe that each protein and enzyme must be studied carefully, and comprehensive information about the origin of human protein must be collected and discussed, for all biochemical pathways known, before we can make any reasonable model about the Tree of Life. It is a long way to go. But the real challenge will be then to understand why Nature decided that the archaea proteins of transcription and translation were the best ones to select for the Eukarya, while for glycolysis, the biosynthesis of lipids and of small molecules Nature selected the eubacteria proteins for the Eukarya.

General information

Selected Publications 1998-2003

Ayora S, Piruat JI, Luna R, Reiss B, **Russo VEA**, Aguilera A & Alonso JC (2002). *Characterization of two highly similar Rad 51 homologs of Physcomitrella patens*. J Mol Biol 316:35-49

Markmann-Mulisch U, Hadi MZ, Koepchen K, Alonso JC, **Russo VEA**, Schell J & Reiss B (2002). *The organization of Physcomitrella patens RAD51 genes is unique among the eukaryotic organisms*. PNAS 99: 2959-2964

Musa A, Lehrach H & Russo VEA (2001). *Distinct expression patterns of two zebrafish homologues of the human APP gene during embryonic development*. Dev Genes Evol 211: 563-567

Schulz P, **Hofmann AH, Russo VEA**, Hartmann E, Lalouche M & von Schwartzberg K (2001). *Cytokinin overproducing over mutants of Physcomitrella patens show increased riboside to base conversion*. Plant Physiology 126:1-8

Hofmann AH, Codón AC, Knight C, Cove D, Schaefer DG, Chakparonian M, Zryd J-P & **Russo VEA** (1999). *A specific member of the CAB multigene family is efficiently targeted and disrupted in the moss Physcomitrella patens*. MGG 261:92-99

Russo VEA (Editor-in-chief), Cove D, Edgar L, Jaenisch R & Salamini F (1999). *Development - Genetics, Epigenetics and Environmental Regulation*. Monography, Springer-Verlag, Berlin Heidelberg

Yarden O & **Russo VEA** (1999). *Genetic and Environmental Influence on the Development of the Filamentous Fungus Neurospora crassa*. In *Development - Genetics, epigenetics and environmental regulation*, Russo VEA, Cove D, Edgar L, Jaenisch R & Salamini F, eds., Springer-Verlag, Berlin Heidelberg

Lauter F-R, Marchfelder U, Russo VEA, Yashamiro C, Yatzkan E & Yarden O (1998). *Photoregulation of cot-1, a kinase-encoding gene involved in hyphal growth in Neurospora crassa*. Fungal Gen Biol 23:300-310