Journal of Computational Interdisciplinary Sciences (2011) 2(2): 123-129

Printed version ISSN 1983-8409 / Online version ISSN 2177-8833
http://epacis.net
doi: 10.6062/jcis.2011.02.02.0039

In silico restriction analysis for identifying microbial communities in T-RFLP fingerprints

César A. Caretta* and Elcia M.S. Brito**

Manuscript received on October 21, 2010 / accepted on August 16, 2011

ABSTRACT

We present here a technique for correlating T-RFLP fingerprints with genomic clone libraries, using a *in silico* restriction analysis, in order to identify the microbial (Bacteria and Archaea) communities present in an environmental sample. This technique is very useful for independent of culture metagenomic studies. We also show some results on the application of this technique to 3 environmental samples from an anthropogenically extreme site. We confirm that T-RFLP results are quite reproducible both in peak's location and area. We found that the peak position (Terminal Restriction Fragment, TRF) can be identified with an uncertainty of only 0.3 base pair, while the percentage o fluorescence (frequency of population) has about 4% of relative uncertainty. Using the TRFs obtained from the *in silico* restriction as reference, we found that deletion and insertion during electrophoresis step of the T-RFLP and cloning must be taken into account. They typically produce a shift in the range $[-2$ to $+1]$ in PCR/cloning/sequencing and $[-4$ to $+1]$ in PCR/T-RFLP, independent of fragment size. This also means that deletion is more usual than insertion. The *in silico* restriction analysis allowed us to recognize 100% of the T-RFLP peaks of abundant populations, 60% of intermediate and 50% of poorly represented ones. Also, almost all populations recovered in the clone library could be associated to T-RFLP peaks, but not *vice-versa*, confirming that the T-RFLP is an efficient technique for detecting the less dominant populations of a microbial community.

Keywords: T-RFLP, clone libraries, microbial communities, Computational Biology and Bioinformatics.

1 INTRODUCTION

The taxonomy of living beings is traditionally accomplished on the basis of their *phenotypic* (external aspects, such as morphology, color, etc) and *physiologic* (concerning internal functions and biochemistry) characteristics. When the microbiological world was unveiled, the same rules were applied, although with limited success (e.g. [11]). In the second half of twentieth century, the idea of using *genotypic* information (based on the molecular sequencing of genes and proteins) for improving the taxonomical and phylogenetical[1] studies was introduced (e.g. [15]). Specifically, a group of microbiologists [7] proposed the use of ribosomal RNA (rRNA) genes, which are common to all living cells, were sufficiently preserved in structure along the life evolution on planet Earth and have an adequate length for being completely sequenced. From these innovative works arose the proposal for classifying the organisms in three Domains based on their genome. Then, the previously called *Monera* Kingdom was divided in two Domains: *Archaea* and *Bacteria*, while the other four Kingdoms (*Protista*, *Fungi*, *Plantae* and *Animalia*)

Correspondence to: César A. Caretta – E-mail: caretta@astro.ugto.mx
*Departamento de Astronomía, DCNyE, Universidad de Guanajuato, Guanajuato, Gto., Mexico
**Grupo de Ingeniería Ambiental, Departamento de Ingeniería Civil, DI, Universidad de Guanajuato, Guanajuato, Gto., Mexico.
E-mail: emsbrito@gmail.com

[1] More than classifying and labeling the organisms, the real concern currently is to disentangle the evolutionary relationships among them.

were regrouped in the *Eukarya* Domain [14]. This phylogeny comes from a measure of the level of divergence between their rRNA sequences, using the 16S small subunit gene for prokaryotes (Bacteria and Archaea) and the 18S small subunit gene for eukaryotes (Eukarya)[2].

For describing the genome (or some specific genes) of an organism it is necessary first to isolate the organism. Although relatively simple for macroscopic organisms, this may constitute an impossible task for some microorganisms. The isolation of a microorganism depends on its detection and cultivation in laboratory. For cultivation, the inoculation of the microorganism in a culture media with adequate nutrients (especially a carbon source and an electrons donor) is required. The problem is that these adequate nutrients are not usually known *a priori* when the organism was not isolated and studied yet. Also, the techniques for counting colony formation units (CFUs), in which one sees the microorganisms by their ability to grow in colonies till the size of being visually detected in a culture plate, have revealed results spread on up to 3 orders of magnitude [13]. The direct conclusion of this finding is that only a small fraction of the microorganisms is apt to be cultivated with the current techniques (probably less than 1%).

During the 90's, new molecular tools[3], called independent of culture techniques, were proposed. These tools are based on extracting information from an environmental sample without the necessity of first isolating the distinct organisms that are present there. The total gene content of this environmental sample is thus called *metagenome*. The characterization of this metagenome is then carried out based on the variation of the genetic patterns, which is also known as "fingerprinting" (e.g. [6]).

In this paper we will describe and discuss a technique for helping the identification and analysis of metagenome fingerprints. More specifically, we present some results on the *in silico* version of the enzymatic restriction of rRNA segments (after a PCR amplification, see details below) used for separating and identifying the distinct sequences present on a metagenome. As we show afterwards, this *in silico* restriction can help the analysis and correlation of combined T-RFLP fingerprints and clone libraries. In section 2 we discuss these molecular techniques, while in section 3 we present the *in silico* restriction. Finally, in section 4 we discuss the results of the application of these techniques to an environmental sample from an anthropogenically extreme site. The conclusions are presented in section 5.

2 MOLECULAR TECHNIQUES, T-RFLP AND CLONE LIBRARIES

The molecular techniques used to study microorganisms independently of their culture in laboratory begin with the extraction of pooled DNA from an environmental sample. This total DNA is firstly amplified by the *Polymerase Chain Reaction* (PCR) method, using specific oligonucleotides (e.g. [9]). In PCR, not the whole DNA molecule of each existing population in the sample is amplified, but only the stretch delimited by the specific oligonucleotide primers used. The complete rRNA 16S gene is a segment of about 1500 base pairs (bp). In the present study, for instance, we used the bacterial universal primers 8F and 907R[4], resulting in rRNA 16S sequences of 899 bp. The distribution of the nitrogenous bases along the obtained sequence is characteristic of an specific *Operational Taxonomic Unit* (OTU), which represents a distinct population.

There are many methods for separating the different sequences amplified by PCR (see, e.g., [2]). Here, we will concentrate ourselves in the *Terminal Restriction Fragment Length Polymorphism* (T-RFLP) and clone libraries. The T-RFLP consists on running the PCR using fluorescently labeled primers and making a new cleavage (enzymatic digestion) of the amplified DNA segments. The restriction enzyme (also known as restriction endonuclease) cuts the segment at an specific site every time it recognizes a certain short sequence of 4 to 6 nucleotides (enzyme gene), producing a set of fragments. Each sequence will give a distinct set of fragments, in such a way that the sizes of the fragments can be used to separate the different populations. When the primers are not fluorescently labeled, as are the cases of *Restriction Fragment Length Polymorphism* (RFLP) and *Amplified Ribosomal DNA Restriction Analysis* (ARDRA), all the restriction fragments can be visualized in an agarose gel electrophoresis run. On the other hand, in T-RFLP, only the fluorescently end-labeled PCR products (terminal fragments) are laser detected in a capillary electrophoresis (e.g. [8]). The capillary electrophoresis produces an *electropherogram*, where the position of the peaks mark the size of the fragments, while their intensity is proportional to the relative frequency of the populations they represent (Fig. 1).

[2] S, abbreviation for "svedberg", is a non S.I. unit for measuring sedimentation coefficients (for instance during centrifugation), very useful for distinguishing ribosomes. The sedimentation coefficient is defined as the rate between the sedimentation velocity and the applied acceleration, resulting in a quantity with dimension of time. Thus, one svedberg is defined as exactly 10^{-13} seconds (100 fs).

[3] methodologies based on extraction and manipulation of genomical material.

[4] F and R refer to "forward" and "reversed" primers respectively.

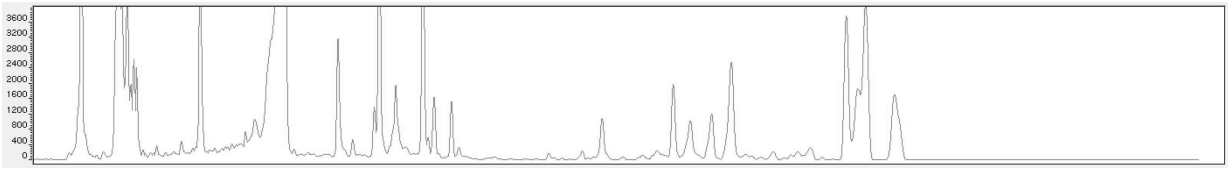

Figure 1 – Example of an electropherogram from PCR/T-RFLP. Each peak was produced by fragments of a distinct size (a specif population in the sample). The position of the peak corresponds to the size of the TRF in base pair units, while its area is proportional to the relative abundance of the population in the community.

No two peaks are associated with the same population present in the sample, but one peak may be associated to more than one population. This last confusion may happen more frequently in environmental samples with high biodiversity. To minimize this effect it is recommended to use more than one restriction enzyme (usually two or three), since the fluorescently labeled terminal fragments (TRFs) of a population can be different for each distinct enzyme used.

For a possible identification of the TRFs, it is necessary to create a genomic clone library from the same DNA sample. For that, the PCR amplified genes (amplicons), using unlabeled primers, are first inserted into a vector. For cloning rRNA 16S genes the usual vectors are the plasmids, which are small mobile segments of bacteria DNA, very stable and easily manipulated. The vector is, then, incorporated into a bacteria (*E. coli*, for instance) by chemically treating and modifying the sensibility of its plasmatic membrane. Since the targeted rRNA 16S genes are now inside new organisms (the clones), they can be separated by spread plate culture, in standard laboratory media containing X-Gal and ampiciline in order to inhibit the growth of the *E. coli* that did not receive the insert, followed by the isolation of their colonies. A new PCR amplification is done, using unlabeled primers that flank the rRNA 16S gene inserted in the plasmids, and the resulted amplicons are taken to be sequenced.

Since we are dealing with a metagenome, a certain number of clones is produced. The relatively more frequent populations will produce more clones, while some rare populations may be missed. In fact, the number of clones expected for adequately represent an environmental sample depends on the diversity of populations (species richness[5]) existing in such sample. For a richer sample, a bigger clone library is necessary for catching even the less representative OTUs.

The obtained base sequences are then compared in order to identify their similarity to other known sequences deposited in a Gene Bank (such as the NCBI[6]). This is known as the BLAST technique [1]. Using the sample sequences and the most similar ones found in the Gene Bank, a phylogenetic tree can be constructed.

3 THE *IN SILICO* RESTRICTION ANALYSIS

Once the community T-RFLP fingerprint and its respective genomic library have been obtained, the *in silico* restriction analysis can be applied. The entries are the library sequences (usually in FASTA format) and the restriction site. It is recommended to check if the sequences are all in the same direction and begin by the same primer. The algorithm for the restriction is very simple: after reading a sequence, the program inspects each group of # nucleotides (where # is the number of nucleotides of the used enzyme) for the first time a coincidence with the enzyme gene appears. When it happens, a counter for the checked nucleotides, plus the restriction site, gives the size of the excised fragment. The output is a list of the sequences for which the enzyme gene was found and the longitude of their respective restricted fragment.

4 APPLICATION TO ENVIRONMENTAL SAMPLES

We have applied the methodology described in sections 2 and 3 to three environmental samples taken from an anthropogenically extreme (hyper-alkaline and contaminated with metals) site, as described in [3]. The samples were provided by a chromite processing industry located in the state of Guanajuato, Mexico. The three samples are referred hereafter as Re, LW and LD, respectively for superficial soil from the industrial residue, surface sediment from a lixiviate channel collected during the wet season and surface sediment from the same lixiviate channel collected during the dry season. Each sample was processed with three distinct restriction enzymes: *Hae*III, *Hin*P1I and *Hpa*II, in triplicate or, in some cases, quadruplicates.

The populations (represented by TRFs) found by T-RFLP analysis were separated in the most abundant ones (more than 10%

[5]There are some indices for estimating the richness of species in a sample based on the observed number of species, such as the Chao Estimator [4, 5].
[6]http://www.ncbi.nlm.nih.gov/

Table 1 – Number of TRFs found in the studied samples (taking the means of triplicates/quadruplicates for each enzyme and sample), separated by the abundance of respective populations (A: abundant, I: intermediate, P: poor, T: Total). Only peaks with more than 1% of total fluorescence were taken into account.

Sample	*Hae*III				*Hin*P1I				*Hpa*II			
	A	I	P	T	A	I	P	T	A	I	P	T
Re	2	8	8	18	1	8	7	16	0	10	15	25
LW	2	10	7	19	1	9	8	18	2	7	9	18
LD	3	3	4	10	3	3	9	15	3	5	5	13

of total fluorescence), the intermediate ones (between 2 and 10% of total fluorescence) and the poorly represented ones (between 1 and 2%). The distribution of these populations according to their abundance is shown in Table 1.

Comparing the results from distinct replicates we found that they are very consistent, with differences in the TRF value lower than 0.3 bp. This mean that the T-RFLP is quite reproducible. They also match in intensity: the average spread in percentage of total fluorescence is below 4% for all the used enzymes. When no match is seen between the position of the TRFs, as was the case of one of LW samples, we feel secure to discard such sample as an outlier. We present the list of the found TRFs, for each enzyme, with the respective match between samples, in Table 2.

It is worth to note that the same total DNA samples used for obtaining the T-RFLP fingerprints were also used for preparing the clone libraries. A total of 99 clones were obtained, distributed in 25 OTUs (for the phylogenetic tree of the found clones, see Figure 3 of [3]). The Chao-1 estimator gives, for the 3 samples together, an estimate for species richness of 81 ± 30 OTUs. Thus, our sampling effort was able to detect about 23 to 49% of the populations probably present in the studied microbial community. Although not ideal for representing the community, this library may be enough for the following verification. Thus, the obtained sequences of these clones were used for the *in silico* restriction analysis. The results of this analysis are presented in Table 3.

We found that the sequences of clones that are similar ($>97\%$) to a particular OTU do not always produce exactly the same fragment longitudes in the *in silico* restriction analysis, with differences in the range [-2 to $+1$] base pairs around the most frequent value (the one quoted in Table 3). This is the case on about 20% of the time, with a similar rate for the three enzymes. Such uncertainty may be a consequence of both incorrect identification in the BLAST step (since it is usual to accept similarities between 97 and 99%) or instrumental errors during sequencing (capillary electrophoresis) or PCR (formation of heteroduplexes during *Taq* polymerase transcription). This suggests that the deletion is more frequent than the insertion in the process of PCR/cloning/sequencing.

The probable matches between the TRFs and the *in silico* restriction fragments are presented in Table 4. Since the T-RFLP has a limited detection range (between 35 and 500 bp), due to the internal standard used[7], there are some populations present in the community that will be missed. These are the cases of our clone sequences nearest to the genera *Thiobacillus*, *Paracoccus*, *Dietzia* and *Alkalibacterium* with *Hin*P1I enzyme, and *Stenotrophomonas* with *Hpa*II. Furthermore, the populations with fragments between 35 and 46 bp (*Lysobacter*, *Stenotrophomonas*, *Algoriphagus* and *Deinococci* with *Hae*III, *Fulvimarina* with *Hin*P1I and *Deinococci* with *Hpa*II) seem to be detected with peaks between 41 and 46 bp, probably because there is a small delay in the beginning of capillary electrophoresis run due to the presence of the fluorochrome.

The sizes of the TRFs (Table 2) are systematically smaller than the longitudes given by the *in silico* restriction analysis (Table 3). The range of this difference is between -4 and $+1$. This is probably an artifact of the T-RFLP (interference of fluorochrome during capillary electrophoresis). Note that this difference is independent of the size of the fragment, contrary to what is usually supposed. On the other hand, almost all (above 85%) populations revealed in the clone library had a matched peak in the T-RFLP fingerprint, while many peaks in the electropherogram were not identified in the bank. This means that T-RFLP is able to detect rare populations in the community that can not be cloned in the metagenomic library.

Therefore, the *in silico* analysis showed to be effective in helping the TRFs identification in T-RFLP fingerprints, specially the ones associated to the abundant and intermediate popula-

[7] Gene Scan™-500 TAMRA™ Size Standard, Applied Biosystems.

tions. For the abundant populations the match was of 100%, for intermediate of 64% and for poorly represented about 52%.

Table 2 – TRFs found with each enzyme and separated by their presence in the 3 distinct samples. The numbers in boldface are the TRFs that could be positively identified by the match with the ones revealed by the *in silico* restriction analysis (as presented in Table 4).

Found in	*Hpa*II	*Hin*P1I	*Hae*III
Re, LW, LD	**42**	**42**	**64**
	44	**44**	**198**
	431	**56**	**203**
		203	221
		369	**229**
Re, LW	52	52	**42**
	63	75	**44**
	137	**90**	**206**
	138	434	**226**
	146		
	434		
	489		
LW, LD	95	**95**	**41**
	126	**208**	**192**
			227
Re	121	200	**68**
	123	**213**	**187**
	125	329	**191**
	129	336	**214**
	134	340	218
	135	354	219
	158	**442**	**223**
	274		**228**
	279		290
	291		
	394		
	398		
	465		
	469		
	485		
LW	**43**	**41**	**43**
	86	86	51
	144	89	52
	436	99	59
	458	141	**65**
	484	181	99
		204	253
LD	**45**	**43**	180
	46	**45**	251
	53	**46**	
	100	54	
	104	100	
	161	126	
	162	**233**	
	457	**235**	

5 CONCLUSIONS

The usual way of studying microbiological community samples is by the use of two or more techniques, which work complementarily. In the case of T-RFLP and clone banking, the first one is much cheaper than the second, and can be applied to follow the changes in the community with time or changes of environmental conditions. The clone library may be done, in this case, at least once, in one representative time, to identify the T-RFLP peaks (specially the ones that correspond to the most abundant populations).

The results of the present work can be summarized as follows:

– We confirmed that the T-RFLP technique is robust and gives quite reproducible results (different replicates give TRFs spread in a range of only 0.3 bp and with percentages of total fluorescence uncertain to about 4% on average). We also confirmed that it allows the detection of less frequent (rare) populations that may be lost in constructing a clone library.

– Errors of deletion or insertion during the capillary electrophoresis step must be taken into account, both in clone libraries and T-RFLP (specially this last one). The *in silico* fragments of the clone sequences vary in longitude from -2 to $+1$ around their most probable value, while the T-RFLP peak sizes range from -4 to $+1$ bp respect to the longitude of the TRFs found by the *in silico* restriction analysis. These results suggest that deletion is more usual than insertion. For the *in silico* analysis, the above, associated with the uncertainty in the identification of species by BLAST/phylogenetic trees and possible PCR errors, may also lead to losing the restriction site in 20% of the cases.

– For more than 85% of the populations revealed in the clone library we could find a matched peak in the T-RFLP fingerprints. We were able to identify 100% of the abundant populations, 60% of the intermediate and 50% of the less representative ones. This shows that the combination of T-RFLP fingerprints, clone libraries and *in silico* restriction analysis is effective.

ACKNOWLEDGMENTS

We are grateful to the industry Quimica Central de México, located in Léon (Guanajuato, México) for providing the environmental samples. The T-RFLP and clone libraries were processed at Laboratory of Environnement et Microbiologie (UMR

Table 3 – TRFs provided by the *in silico* restriction analysis of the clone libraries. For each population given in column 1, the numbers in boldface are the fragment longitudes produced with each enzyme, while the numbers in square brackets are the number of sequences that produced such fragments.

OTU	*Hae*III	*Hin*P1I	*Hpa*II	Library[1]
Lysobacter enzymogenes	**39** [31]	**371** [29]	**436** [29]	[37]
Stenotrophomonas maltophilia	**39** [3]	**211** [3]	**496** [3]	[3]
Herbaspirillum sp.	**132** [1]	**294** [1]	**185** [1]	[1]
Malikia spinosa	**212-213** [2]	**146-147** [2]	**483-484** [2]	[2]
Thiobacillus thioparus	**202** [5]	**568** [5]	**492** [5]	[9]
Nitrosomonas sp.	**71** [1]	**233** [1]	**157** [1]	[1]
Hoeflea sp.	**225** [3]	**59** [2]	**128-130** [2]	[3]
Sinorhizobium sp.	**190** [1]	**58** [1]	**436** [1]	[1]
Fulvimarina sp.	**75** [1]	**36** [1]	**136** [1]	[1]
Paracoccus sp.	**193** [3]	**511** [3]	**130** [3]	[3]
Dietzia sp.	**67** [1]	**664** [1]	**149** [1]	[1]
Nocardioides jensenii	**225** [1]	**439** [1]	**140** [1]	[1]
Bacillus akibai	**207** [5], **231** [13]	**238** [19]	**165** [19]	[19]
Alkalibacterium pelagium	**232** [1]	**580** [1]	**148** [1]	[1]
Algoriphagus aquatilis	**39** [1]	**94** [1]	**144** [1]	[1]
Deinococci bacterium	**39** [1]	**280** [1]	**36** [1]	[1]

Notes: (1) "Library" is the total number of sequences in the clone library.

Table 4 – Probable identification of TRFs based on the comparison with the *in silico* restriction analysis[1].

Most similar to	*Hae*III	*Hin*P1I	*Hpa*II
Lysobacter enzymogenes	beginning	369	431?, 434, 436
Stenotrophomonas maltophilia	beginning	208, 213	out
Herbaspirillum sp.	–	–	–
Malikia spinosa	214	141?	484, 485
Thiobacillus thioparus	198, 203	out	489
Nitrosomonas sp.	68	233	158
Hoeflea sp.	223, 225, 227	56	125, 126, 129
Sinorhizobium sp.	187, 191	56	431?, 434, 436
Fulvimarina sp.	–	beginning	134, 135, 136
Paracoccus sp.	192	out	126, 129
Dietzia sp.	64, 65	out	146
Nocardioides jensenii	226	434?, 442?	137, 138
Bacillus akibai	206, 228, 229	233, 235	161, 162
Alkalibacterium pelagium	229	out	146
Algoriphagus aquatilis	beginning	90, 95	144
Deinococci bacterium	beginning	–	beginning

Notes: (1) "beginning": the fragment was detected at the starting point of the electropherogram (between 41 and 46 bp); "out": the fragment is out of the detection range of T-RFLP with the internal standard used.

IPREM5254), Université de Pau et des Pays de l'Adour (Pau Cedex, France), for whom we are greatly indebted. We thank the two referees for very useful comments that helped us to improve this paper. E.M.S.B also acknowledges financial support from FONCICyT (Ref. 95887) and PROMEP (*Apoyo a la Incorporación de Nuevos PTC*).

REFERENCES

[1] ALTSCHUL SF, GISH W, MILLER W, MYERS EW & LIPMAN DJ. 1990. J. Mol. Biol., 215: 403–410.

[2] BRITO EMS, ANDRADE LH, CARETTA CA & DURAN R. 2007. Microorganisms Bioprospection: A New Tendency in Microbial Ecology, in Leading-Edge Environmental Biodegradation Research, pp. 199–222, ed. Lyman E. Pawley, Nova Science Publ.

[3] BRITO EMS, PIÑÓN-CASTILLO HA, GUYONEAUD R, CARETTA CA, GUTIÉRREZ-CORONA JF, DURAN R, NEVAREZ-MOORILLON GV, REYNA-LÓPEZ GE & GOÑI-URRIZA MS. 2011. Appl. Microbiol. and Biotechn., submitted.

[4] CHAO A. 1984. Scandinavian J. Stat., 11: 265–270.

[5] CHAO A. 1987. Biometrics, 43: 783–791.

[6] DAHLLÖF I. 2002. Curr. Opinion Biotech., 13: 213–217.

[7] FOX GE, STACKEBRANDT E, HESPELL RB, GIBSON J, MANILOFF J, DYER TA, WOLFE RS, BALCH WE, TANNER RS, MAGRUM LJ, ZABLEN LB, BLAKEMORE R, GUPTA R, BONEN L, LEWIS BJ, STAHL DA, LUEHRSEN KR, CHEN KN & WOESE CR. 1980. Science, 290: 457–463.

[8] OSBORN AM, MOORE ERB & TIMMIS KN. 2000. Environmental Microbiology, 2: 39–50.

[9] SAIKI RK, GELFAND DH, STOFFEL S, SCHARF SJ, HIGUCHI R, HORN GT, MULLIS KB & ERLICH HA. 1988. Science, 239: 487–491.

[10] MUYZER G, DE WAAL EC & UITTERLINDEN AG. 1993. Appl. Env. Microbiol., 59: 695–700.

[11] VAN NEIL CB. 1946. Cold Spring Harbor Sympo. Quant. Biol., 11: 285–301.

[12] STACH JEM, BATHE S, CLAPP JP & BURNS RG. 2001. FEMS Microbiol. Ecol., 36: 139–151.

[13] TORSVIK V, GOKSOYR J & DAAE FL. 1990. Appl. Environ. Microbiol. 56:782-787.

[14] WOESE C. 1987. Microbiological Reviews, 51: 221–271.

[15] ZUCKERKANDL E & PAULING L. 1965. J. Theor. Biol., 8: 357–366.