*Available online at http://www.journalcra.com*

*International Journal of Current Research*
*Vol. 7, Issue, 05, pp.15667-15673, May, 2015*

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

# TOWARDS A NEW PROKARYOTIC BIODIVERSITY ASSESSMENT: THE *PASTA* INDEX

*Jacques Thierie

Belgium

ARTICLE INFO

*Article History:*
Received 20th February, 2015
Received in revised form
06th March, 2015
Accepted 04th April, 2015
Published online 25th May, 2015

*Key words:*

Bacterial consortium,
Statistical functions,
Partition at random,
Functional and spatial diversity,

## ABSTRACT

We assume that the distribution of "biological objects" (species, family, phyla,) inside a "set" (community, assemblage, etc.) is at random. Although similar assumptions have been put forward by others (see text), our treatment is basically different. Based on a simple algorithm describing the random partition of a set, we obtain the resulting statistical laws characterizing the partition (that is to say: probability density, repartition function and moments). The theoretical results are supported by mathematical simulations of thousands random drawings. Assuming that the reduced variable characterizing a biological object is analogous to a probability, we adjust curves of rank / relative abundance type (like non-log plots Witthaker type, Fisher plots,) with probabilities calculated through our statistical functions. The results are compelling both for eukaryotes (not shown) and for prokaryotes. The sole fitting parameter is the partition order, *M* (number of parts of the set). As a rule, this number is an integer. Nevertheless, we are discussing the possibility that this variable can be considered as a simple index and may have non-integer values. This index, which we call the *PASTA index* (from the French *Partition STAtistiques*) is defined as a biodiversity estimate and characterizes what we call the bioclusters number of the assemblage. We take as true that this index can be applied to characterize structural, genetic biodiversities and thus applies to various levels of description of an assemblage.

## INTRODUCTION

For many decades, microbiologists have used a typological methodology to study bacteria. Despite the limitation of this method partly due to the difficulty of cultivating many strains (Rodriguez-Valera, 2002), the typological approach persists (rightly) today. Meanwhile, the central paradigm evolved up to consider a bacterial population as a multicellular organism (Shapiro, 1998). Perhaps, in between, we could talk about consortia to describe coherent associations of different bacteria species endowed with emergent properties (complementary species developed enhanced metabolic properties compared with a single species). The consortium concept applies to many areas including physiology (rumen, dental plaque ...), food (sauerkraut, various cheeses, kefir ...), the environment (soil, compost, biological treatment, roots ...). All the above concepts are based in one way or another on the notion of bacterial species (or groups of species). For prokaryotes, the species concept is problematic for many reasons (Mora *et al*., 2011; Magurran, 2004). This of course puts the typological idea less acceptable, but poses a problem for environmentalists. While measuring biodiversity seems to apply successfully to eukaryotes, the concept becomes more controversial in relation to prokaryotes. Mora and co-workers (Mora *et al*., 2001) found that prokaryotes increase for only 0.1% the number of known species. They attribute this not to a real lack of bacterial diversity (10,000 species only), but to the definition of the species concept in bacteria. There is not only the difficulty of clearly defining the species concept for prokaryotes, but the concept of biodiversity in itself is often problematic. The definition adopted by Magurran (Magurran, 2004) clearly shows that biodiversity goes far beyond the single notion of varieties of species, but also includes the number of individuals, ecological considerations, genetic, space, etc. Our approach is based on two concepts which are often challenging: the characterization of prokaryotes ("species") and their biodiversity. We obviously don’t aim to definitely clear up problems associated with these two concepts, but rather to construct a method of interpretation measurements linked to these two concepts. Our approach is based on the mathematical construction of an index that can integrate the fundamentals of biodiversity (variety, number of individuals, etc.) and, if possible, elements of metabolic variability or spatial distribution. Colwell (Colwell, 2009) defines an index of diversity as “a mathematical expression that combines species richness and evenness as measure of diversity." The *PASTA* index that we look for should be an expanded version of this definition. Work on prokaryotes we conducted in 2011 (Thierie, 2011) allowed us to show that allelic partition permitted to diversify (or make more

***Corresponding author: Jacques Thierie**
*Belgium*

inhomogeneous) population uniform at the start. In what way these various classes could they be distributed? At first, we assumed a random distribution and successfully tested this hypothesis on various eukaryotes (publication in preparation). The idea came up to apply our results to certain situations with eukaryotes, given the theoretical and practical interest of this problem. After giving a very condensed summary of how we built statistical functions applied to a random distribution, we will examine an example of complex bacterial community. In the field of prokaryotic communities, examples are few (less than eukaryotes) ... and we leave the responsibility for the validity of the measures to the authors who made them. We believe that the results are encouraging. Other examples will be treated and some mathematical enhancements (particularly in the area of the curves adjustments) are developed.

Derivation of statistical theories. –The PASTA theory

The rigorous derivation of statistical laws of random partitions (*PASTA* from French «Partition STAtistique") is relatively long and is not (in our opinion), useful or inevitable in an applied short communication like this. This demonstration will be given later elsewhere. The quantities we are considering are in the form of reduced variables (or frequencies) of the type

$$f_i = \frac{q_i}{\max(\{q_k\})} > 0 \qquad (1)$$

forming a partition (no empty subset) of a given set.
The consequence of the partition is that

$$\sum_{i=1}^{M} f_i = 1 \qquad (2)$$

(where *M* is the number of parts of the partition ("order" of the partition).

We then seek the statistical laws governing the objects distribution within the set if the objects frequencies are distributed at random and in an equiprobable way. The simplified algorithm for obtaining the distribution of statistical laws is described below.

```
BEGIN

• Let be a uniform distribution defined on the interval [min, max] (typically, min = 0, max = 1; working in frequency).
• Define the order $M$ of the partition ($M \geq 2$).
• $i = 1$
• Sum = 0
i The frequency $f_i$is obtained by a random draw on the uniform interval
[min, max].
ii max = max - $f_i$
iii Sum = Sum + $f_i$
iv$i = i + 1$
v If $i <$ M – 1 go to $i$.
vi$f_M = 1 - f_{M-1}$
END
```

(Despite a possible formal analogy, this approach of the problem is not comparable to other statistical theories, like those of Tokeshi (1990, 1996), MacArthur (1957) – the "broken stick" model ...).

In this way, we obtain a random sequence of *M* frequencies, satisfying (2). The procedure generates a result that could be called "chronological". Obtaining truly random suite values needs to repeat this algorithm a lot of times (theoretically, at infinity) and to "mix" the resulting frequencies using a stochastic process, while respecting the constraint (2). Proceeding in this way, it can be shown by recurrence that, for each value of index *i* in a series of mixed sizes, the probability density function is given by

$$g_i(f) = (-1)^{i-1} . \frac{(\ln f)^{i-1}}{(i-1)!} > 0; \forall i \qquad (3.a)$$

and the repartition function by

$$G_i(f) = f \sum_{r=1}^{i} (-1)^{r-1} \frac{(\ln f)^{r-1}}{(r-1)!} \qquad (3.b)$$

The statistical functions of the unique variable *f* are then

$$g(f) = \sum_{i=1}^{M} g_i(f) \qquad (4.a)$$

$$G(f) = \sum_{i=1}^{M} G_i(f) \qquad (4.b)$$

bearing in mind that

$$g_M(f) = g_{M-1}(f) \, ; G_M(f) = G_{M-1}(f)$$

It is easy to verify that

$$G(f = 0) = 0 \, ; G(f = 1) = 1$$

and that

$$\frac{dG(f)}{df} = g(f)$$

Furthermore

$$g(f = 0) = \infty \, ; g(f = 1) = \frac{1}{M}$$

Moments.
The moment of order *m* is defined by

$$E(f_i^m) = \int_0^1 f_i^m g(f_i).df_i ; i = 1, \ldots, M-1$$

Using (3.a):

$$E(f_i^m)=\frac{(-1)^{i-1}}{(i-1)!}\int_0^1 f_1^m(\ln f_i)^{i-1}.df_i$$

Leading to

$$E(f_i^m)=\frac{1}{(m+1)^i} \tag{5}$$

Insofar as the integral of a sum equals the sum of the integrals, the result regarding the times is immediate.

Let designate by

$$E_i(f^m)=\frac{1}{(m+1)^i} \tag{6}$$

Where *m* is the moment order. In the equiprobable case

$$E(f^m)=\frac{1}{M}\sum_{i=1}^{M-1}E_i(f^m) \tag{7}$$

with $E_M(f^m)=E_{M-1}(f^m)$ .

Mean.

For *M* = 1, using (6) and (7), it follows that

$$E(f^1)=\langle f\rangle=\frac{1}{M}.\left(\frac{1}{2}+\frac{1}{2^2}+\ldots+\frac{1}{2^{M-1}}\right)$$

Furthermore, the relationship (Gradshteyn *et al.*, 1980)

$$\sum_{i=1}^{\infty}\frac{1}{a^i}=\frac{1}{a-1}\ ;\ a \text{ integer} >0 \tag{8}$$

permits to show that the mathematical expectation takes the remarkable form

$$\langle f\rangle=\frac{1}{M} \tag{9}$$

Variance.

The moment of order two is

$$E(f^2)=\frac{1}{M}.\left(\frac{1}{3}+\frac{1}{3^2}+\ldots+\frac{1}{3^{M-1}}\right) \tag{10}$$

and thus the variance is given by

$$\sigma_f^2=\frac{1}{M}.\left(\frac{1}{3}+\frac{1}{3^2}+\ldots+\frac{1}{3^{M-1}}\right)-\frac{1}{M^2} \tag{11}$$

Using (II.8), leads to

$$\sigma_f^2=\frac{1}{2M}.\left(1+\frac{1}{3^{M-1}}\right)-\frac{1}{M^2} \tag{12}$$

Note: for $M \geq 4$ , the following approximation can be used

$$\sigma_f^2\approx\frac{1}{2M}-\frac{1}{M^2} \tag{13}$$

More compact forms.

We show, using the explicit forms that the functions (4) can be put in more practical forms, such as

$$g(f,M)=\frac{1}{M}\left(2\frac{(-\ln f)^{M-2}}{(M-2)!}+\sum_{r=0}^{M-3}\frac{(-\ln f)^r}{r!}\right)$$

$$G(f,M)=\frac{f}{M}\sum_{r=0}^{M-2}(M-r).\frac{(-\ln f)^r}{r!} \tag{14.b}$$

where the $Z(f,M)$ notation indicates that the statistics functions change with the order *M* of the partition.

Asymptotic approximations.

Taking into account that (Dwight, 1961; Gradshteyn *et al*, 1980)

$$e^x=\sum_{i=0}^{\infty}\frac{x^i}{i!};x^2<\infty$$

one shows that, using x=-ln*f*

$$\sum_{i=0}^{\infty}\frac{(-\ln f)^i}{i!}=\frac{1}{f};f>0$$

Passing to the limit, (4.b) becomes

$$\lim_{M\to\infty}g(f,M)=\frac{1}{M}.\frac{1}{f}=0;\forall f>0 \tag{15}$$

Relation (15) is suitable for all *f*, excepted for the singular point *f*=0.
Therefore, for *M* big, we may assume that

$$g(f,M)\,\square\,\frac{1}{M}.\frac{1}{f};f>\varepsilon>0 \tag{16}$$

where ε is an infinitesimal value.

By integration, and considering that $G(f=1,M)=1$ it follows that

$$G(f,M)\square\frac{1}{M}.\ln f+1;\forall f>\varepsilon;\varepsilon>0 \qquad (17)$$

This relationship deserves a closer thorough review. Indeed, it is an approximation for *M* large, which is not always the case. For, $M\approx 3$ the error using (17) can be roughly estimated in the range of 4-5%, as the case might be (data not shown). However, if we consider that *M* is not only a number of parts of a partition but a simple index characterizing a partition, then *M* doesn't necessarily be an integer. This may seem paradoxical, but obtaining statistical functions *PASTA* by the described method here can be considered as an analogy and not as strictly formal derivation of the results ("it is as if we were doing a partition ..."). The big advantage of (17) is to avoid calculation of factorials and summations on non-integers, as they appear in (14). This comment will probably appear more evident in the next section, when working on practical examples.

**Note:** Many numerical simulations have shown the accuracy of the theoretical results obtained above (data not shown).

Fundamental assumption.

The basic assumption underlying our statistical theory is that the distribution of the reduced variables in a set (1) is uniformly at random, as derived here. In practice, this means that a reduced variable *f* can be equated to a calculable probability of (7).

Based on the calculation of the probability of *f* (Kaufmann, 1965; Ventsel, 1987)

$$pr\{x_1\le f<x_2\}=pr\{f<x_2\}-pr\{f<x_1\}$$

$$pr\{x_1\le f<x_2\}=G(f_2,M)-G(f_1,M)$$

we assume that

$$f\equiv G(f_2,M)-G(f_1,M) \qquad (18)$$

with $f=pr\{x_1\le f<x_2\}$.

The examples which follow have been calculated in this manner. In addition, numerous other examples (not shown here) proved the validity of this hypothesis for both prokaryotes and eukaryotes.

For example, Figure 1 shows the type of adjustment we can make based on our fundamental hypothesis and using (19) (courtesy Colwell, Colwell 2009).

**Figure 1**

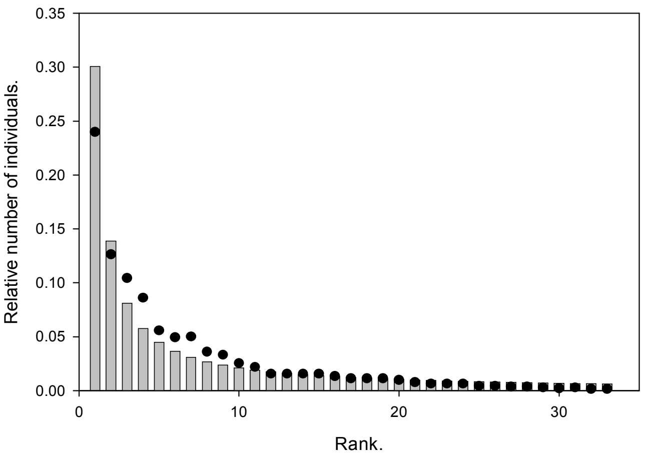

**Figure 1. Relative variable (relative number of individuals) versus partition set (rank) fitting (after Colewell, 2009)**

The gray bars in the histogram represent the probability (Prob) calculated according to (II.18); black points are "experimental data" (reduced variables (RedVar) without specification here) corresponding to the ranks. Relative mean squares LSq (∑ (RedVar - Prob) ² / data number) is calculated to roughly estimate the adjustment rightness (which consists in minimizingLSq by varying *M*). The adjustment produces a *PASTA* index M = 5 (LSQ = 1.8 $10^{-4}$), which corresponds to a good biodiversity. A more detailed interpretation of this index is given in the Discussion. (Softwares used for adjustments were written in single precision by the author)

(This example is given for purely illustrative purposes, conclusions or biological speculations can obviously not be given about this graph without units or explanation)

## RESULTS

This section is intended to illustrate the possibilities offered by our "PASTA theory" in the field of prokaryotes. We consider a case of an already high degree of sophistication. Observations on a Crabtree effect in a bacterial consortium grown in a chemostat (Thierie *et al.*, 2004). In a 2000 thesis (Bensaid, 2000), we highlighted a Crabtree effect in a bacterial consortium stably grown in a chemostat over a long period (see Figure 2). (PCR analysis has demonstrated that the consortium had only bacterial genetic material)

An additional fermentation product, the butyrate, is not shown; this metabolite was produced at a constant rate over the whole range of dilution rates (see Thierie and Penninckx, 2004). It is noted that at low dilution rates (D <0.2 $h^{-1}$; D = Q / V ratio of the volumetric flow (Q) of the substrate inflowing in the reactor and the working volume (V) of the reactor), no lactate appears in the bulk: the bacterial respiration is purely oxidative. On the contrary, beyond D = 0.2 $h^{-1}$, lactate suddenly appears in the medium, subsequent to a respiratory shift of "purely oxidative" to "respirofermentative" mode. This transition is therefore an abrupt and deep metabolic change within the consortium. Figure 5 shows the composition of the main species that make up the consortium before (D = 0.1 $h^{-1}$) and after (D = 0.4 $h^{-1}$) respiratory shift.

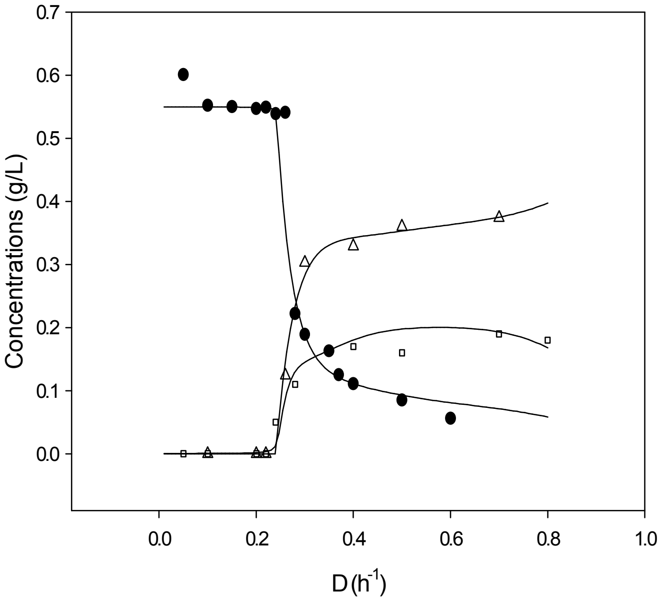

**Figure 2. Steady states of some components in the chemostat. The figure shows the mass concentrations (g / L) of the major metabolites produced by the consortium. ● bacterial biomass (dry weight); Δ substrate (SSF: synthetic sewage feed (OECD, 1981) + 0.5 g / L glucose in final concentration); □ lactate**

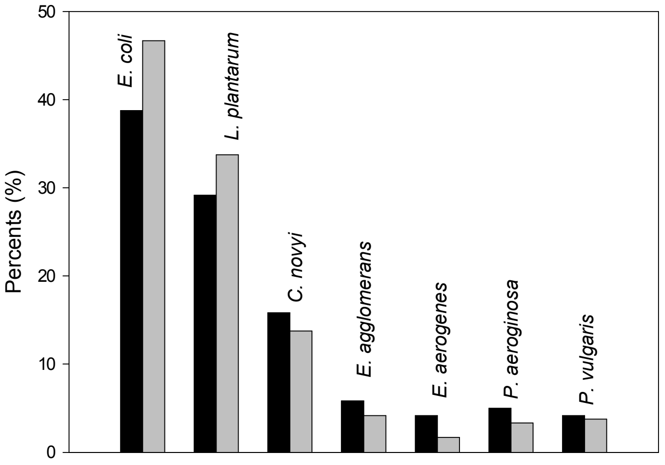

**Figure 3. Relative specific composition of the consortium**

The figure shows the relative specific composition of the consortium at a dilution rate below (black bars: D = 0.1 h$^{-1}$) and above (gray bars: D = 0.4 h$^{-1}$) the critical value. Seven species accounted for nearly 100% of the total species detected. PCR analysis showed that only bacterial genetic material was present in the consortium.

Rather unexpectedly, at first approximation, the consortium's composition remains unchanged. We wanted to characterize this observation in a more qualitative manner. Figures 4a and 4b show the results of the adjustments for the two dilution rates values.

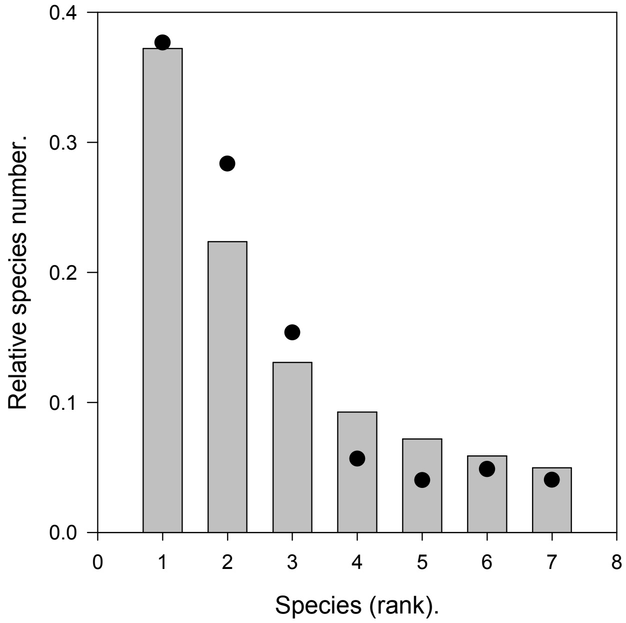

**Figure 4a**

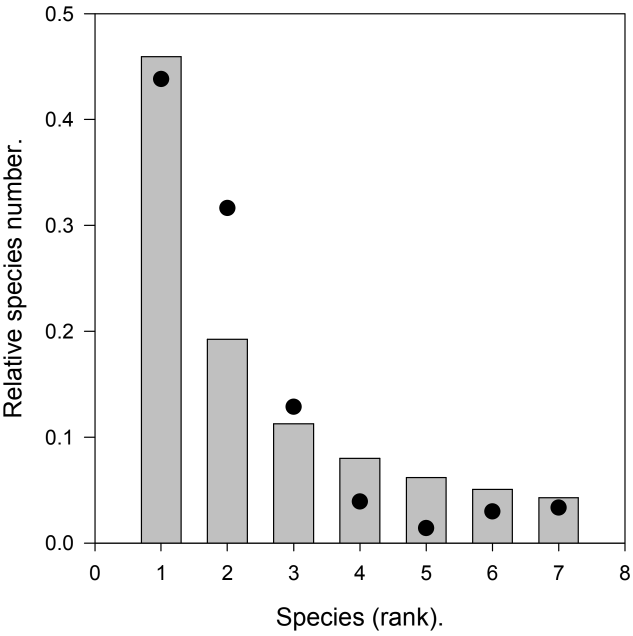

**Figure 4b.**

**Figure 4. Fitting of the relative number of bacterial species in the consortium versus rank. Special attention has been taken in determining the *PASTA* index: a relative least squares curve (see legend to Figure 1) LSq = f (*M*) was performed and the minimum of this quadratic curve was analytically calculated. Minimization process of least squares was thus optimal in this example**

a. D = 0.1 h$^{-1}$. The adjustment gave a value of $M$ = 3.1 with LSq = 9.4 10$^{-4}$. Unless a larger deviation for *L. plantarum* (rank = 2), the fit is very good.
b. D = 0.4 h$^{-1}$. Here we obtained $M$ = 3.6 with LSq = 2.9 10$^{-3}$. The difference of rank = 2 being more important, the fit is not as good as for D=0.1 h-1, but remains significant.

The results obtained gave $M$ = 3.1 and $M$ = 3.6 and are difficult to interpret. However, we can conclude that, in fact, the consortium is not catastrophically distressed BUT there is still a significant difference between the two values of the *PASTA* index, sufficient to prevent us from decide wether $M$ = 3 or $M$ = 4. The full discussion of a non-integer PASTA index will be in the following section.

## DISCUSSION

We believe that the obtained results regarding prokaryotes are convincing enough to be validated. This remark is also justified concerning eukaryotes studies (more numerous) and will be published later. However, many questions remain unanswered regarding the in depth interpretation of the approach and the *PASTA* index. Concerning the Crabtree effect in a consortium two strategies are possible. First, the *PASTA* index is simply a number of parts in a partition, and then it is necessarily an integer. In this case, we must choose: either $M$ = 3 or $M$ = 4, which is almost undecidable. Assume, however, that an integer value is chosen: what then is its meaning? The consortium consists of seven species (but probably many more). However, the *PASTA* index suggests a partition in much fewer parts (four maximum). Given the very coherent structure of a floc (Thierie *et al*., 1999) and the uniform spatial flocs distribution in the bioreactor, it seems unavoidable to adopt a functional rather than a structural interpretation of the consortium.

In other words, the *PASTA* index here represents a number of metabolic (or physiological ...) properties belonging to the entire consortium. We propose to call "biocluster" each group of such properties. A biocluster would then be a set of functional and / or structural specific biological properties, belonging to part of a spatially or not spatially distributed coherent whole. Thus, in our case, at small dilution rates (<2 $h^{-1}$), the consortium is made up of only 3 bioclusters essential (among others) to a purely oxidative respiration; at highest dilution rates (> 2 $h^{-1}$), four bioclusters would be required for an oxidative function and an additional fermentative mode. There would be an increase in functional biodiversity of the consortium. The other strategy is to recognize that the *PASTA* index is nothing but an analogy with a Partitionsensustricto. This analogy would only allow deriving the corresponding statistical functions. Once statistical laws obtained, $M$ would not be a number of divisions (an integer) but a simple index characterizing a particular situation and therefore no longer necessarily an integer.

We admit, for sake of argumentation, that the $M$ values are exact, which is certainly excessively optimistic. We only know, in reality, that they are just significantly different. In oxidative respiration, $M$ = 3.1. We know that only a part of the consortium exhibits oxidative respiration because 1) oxygen does not penetrate to the core of the floc (there is therefore an anaerobic fraction); 2) butyrate is produced at a constant rate over the entire range dilution rates (data not shown; cf. Thierie *et al*, 2004). Greatly simplifying (and just for the sake of argumentation) we could say that $M$ = 3 + 0.1 or 3 "oxidative" bioclusters plus a fraction representing 0.1 equivalent "fermentative" bioclusters. The Crabtree effect would be characterized by three oxidative bioclusters + 0.6 fermentative equivalents clusters. (We insist that these numbers are only used for illustration of an explanation of the non-integer character of $M$ and are not a serious attempt to assess the functional properties of the current consortium). We assume that $M$ is a not necessarily an integer index, characterizing the number of bioclustersin a coherent whole. In the consortium Crabtree effect example, the entire system is spatially homogeneous. We must then admit that bioclusters are completely functional (and in that case, greatly involved in bacteria respiration mode).

## REFERENCES

Bensaïd, A. 2000. Comportement complexe d'un consortium bactérien dans un chémostat. Ph. D. Thesis, Faculté des Sciences, Université Libre de Bruxelles, Brussels.

Coe R. 2008. Designing ecological and biodiversity sampling strategies. Working paper no. 66. World Agroforesty Center, Nairobi, Kenia.

Colwell, R.K. 2009. Biodiversity: concepts, patterns and measurement. Ch. III.1., pp. 257-263. In S.A. Levin, Editor. The Princeton Guide to Ecology. Princeton Univ. Press, Princeton, NJ.

Dwight, H.B. 1961. Table of integrals and other mathematical data. Fourth edition. Macmillan Publishing C°., Inc., New York.

Gradshteyn, I.S., Ryzhyk, I.M. and Jeffrey, A. 1980. Table of integrals, series, and products. Fourth edition. Academic Press. Inc., Orlando, Florida.

Kaufmann, A. 1965. Coursmoderne de calcul des probabilités. Ed. Albin Michel, Paris.

Hortal, J., Lbob, J.M. and Jiménez-Valverde, A. 2006. Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology,* 21:853-863.

MacArthur, R.H. 1957. On the relative abundance of birds species. Proc. Nat. Acad. Sci. USA, 43:293-295

Magurran, A.E. 2004. Measuring biological diversity. Blackwell Publishing, UK.

Mora, C., Tittensor, D.P., Adl, S. *et al*. 2011. How many species are there on earth and in the ocean? *Plos Biology,* 9:1-8

OECD, 1981. Guidelines for testing of chemicals. Ed OECD, Paris

Rodriguez-Valera, F. 2002. Approaches to prokaryotic biodiversity: a population genetics perspective. *Environmental Microbiology*, 4:628-633.

Shapiro, J.A. 1998. Thinking about bacterial population as multicellular organisms. Annu. Rev. Microbiol. 52:81-104.

Thierie, J., Bensaid, A. and Penninckx, M. 1999. Robustness, coherence and complex behaviors of a bacterial consortium from an activated sludge cultivated in a chemostat. Med. Fac. Landbouw, Univ. Gent, 64/5a:205-210 / Thirteenth Forum for Applied Biotechnology, Gent, 1999.

Thierie, J. 2011. The Partition Method applied to biological evolution and geographical distribution. *Editeur J. Thierie. Brussels.*

Thierie, J. and Penninckx, M. 2004. Possible occurrence of a Crabtree effect in the production of lactic and butyric acids by a floc forming bacterial consortium. *Current Microbiology,* 48(3):224-229 (march 2004)

Tokeshi, M. 1990. Niche apportionment or random assortment: species abundance patterns revisited. *J. Animal. Ecol.,* 59.1129-1146.

Tokeshi, M. 1996. Power fraction: a new explanation for species abundance patterns in species-rich assemblage. Oikos. 75: 543-550.

Ventsel, M. 1987. Théorie des probabilités. Eds. MIR, Moscou.

*******