Open Access

# A Multidisciplinary Approach to the Microbial Species Concept: The Role of Bioinformatics in the Search of Detectable Discontinuities

Livio Antonielli[1], Laura Corte[1], Luca Roscini[1], Vincent Robert [2], Ambra Bagnetti[1], Fabrizio Fatichenti[1] and Gianluigi Cardinali[*,1]

[1]*Department of Applied Biology, Microbiology Division, University of Perugia, Italy*

[2]*CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands*

**Abstract:** The problem of species and in particular microbial species is central in biology. An active collaboration of various specialists such as taxonomists, epistemologists, mathematicians and bioinformatics experts is desirable for its solution.

This article intended to show the possibilities and perspectives of bioinformatics research in understanding the ontological problem of species, i.e. the problem of the existence of the microbial species. The approach undertaken of this paper is based on the concept that if microbial species exist, then there should be detectable discontinuities or disconnections between microorganisms assigned to different species. A yeast model has been used to show how the distances from a reference organism raise with the increase of the number of different strains included in the study.

**Keywords:** Species, microbial species, discontinuity, nominalistic category, realistic category.

## INTRODUCTION

### A Multidisciplinary Approach for Species Definition Through the Search for Discontinuities

The rationale of this article is that finding discontinuities among organisms is instrumental to species definition. Since the approach undertaken involves quite different aspects including epistemology, biology and history of the thought around the species problem, this introduction aims to furnish a basic background for non-experts of the field. Paragraph 2 deals with the general problem on species definition, paragraph 3 concentrates on the Biological Species Concept and on the difficulty to apply it universally in biology. The various problems concatenated in the species definition, with a particular emphasis on microbial species concept, is presented in paragraph 4. Finally, some conceptual aspects on the treatment and on the meaning of the discontinuities in multivariate objects sets are treated in the paragraph 5.

The reader already acquainted with the general aspects treated in the first four paragraphs can move directly to the fifth paragraph to understand the essential background and the rationale of this contribution.

### Why the Species Problem should be Approached with Using Several Disciplines Including Informatics and Statistics?

The problem around the understanding and definition of the species concept is the field of an intense debate from both the theoretical and practical viewpoint. Theoretically, the species concept is part of the wider problem on the existence of universal categories and sees the contraposition between "nominalists" and "realists" [1]. The former believes that the species is a mere category of the thought, necessary to indicate groups of organisms, but deprived of any possible biological meaning. Among the most famous nominalists are enumerated Buffon, Lamarck and Darwin [2]. In a famous letter on the Christmas eve of 1856 to Joseph D. Hooker, Darwin stated: - *It is really laughable to see what different ideas are prominent in various naturalists minds, when they speak of "species". It all comes, I believe, from trying to define the undefinable* [3]. On the contrary, a prominent modern synthesis scientist as Ernst Mayr regarded the species as a real entity, universally known even by the tribal cultures [2].

These different positions raise the ontological problem: *does the species exist*? This question is even more complex in microbiology due to the total or partial lack of sexuality, as outlined briefly in the following chapter.

Another challenge is the semantic problem: *how can we define the species*? This is apparently a merely technical aspect focused on the choice of the descriptors to define the species, but it is instead a complex problem based on the designation of the "types" of descriptors. In biology, basically there are two different types of descriptors: the functional and the morphological. Functional descriptors include all the activities exerted by the living organisms. By definition, no functionality can be tested in dead organisms. On the other hand, morphological descriptors are roughly similar in living and dead organisms. Non exhaustive examples of morphological descriptors in microbiology are the cell shape, the cell dimension, and the way of aggregation (e.g. *Staphylococcus vs streptococci* can be

*Address correspondence to this author at the Department of Applied Biology, Microbiology Division, University of Perugia, Borgo XX Giugno, 74, I - 06121 Perugia, Italy; Tel: +39 075 585 6478; Fax: +39 075 585 6470; E-mail: gianlu@unipg.it

**1874-1363/11** **2011 Bentham Open**

discriminated because the round-shaped cells form bunch of grape-like and necklace-like structures, respectively). According to this view, and to the fact that the greek word "*morphè*" indicates the inner form, the design permeating each single object, and molecular descriptors based on DNA analysis share some basic features with molecular characters. For this reason, the morphological and the DNA descriptors will be referred hereinafter with the term "morpho-molecular".

The advantage of morpho-molecular descriptors is the insensitivity to the environment, their stability and their experimental robustness. Finally, morpho-molecular character is normally expressed as discontinuous or categorical data, e.g. the presence or absence of a given structure is normally indicated with a binary notation (1/0), molecular data are reported with a well known quaternary notation (A, C, G, T). The fact that these descriptors are discontinuous simplifies several analyses and remove a large degree of uncertainty typical of the continuous data, which is treated with the statistical analysis to assess the significance.

Conversely, functional data indicate not what an organism *is*, but rather what it *does*, making functional data key factors for the understanding of the complex machinery represented by cells, tissues and organisms, and for the applicative exploitation of microbes, plant and animals in biotechnology. These data are sensitive to the environment and need a careful experimental standardization. As for the understanding of "how can we describe species" the two types of descriptors have been differently evaluated during the short history of microbial taxonomy. In the last 150 years, since the publication of the Darwin's "The origin of the species" it has become increasingly clear that the species definition cannot be disjoined from the evolutionary problem, "*do the microbial species evolve?"*. This latter question can be easily misunderstood and it is important to stress that the evolution problem is not on whether the evolution exists (since we can measure it), but rather if the unit of evolution is the species or the individuals.

## The Species Problem: A General Overview for Informatics Experts - Biology Outsiders

The species concept is one of the most debated aspects in Biology and is particularly complex when applied to microbes. Biologists studying animal and plants have a strong consensus toward the biological species concept based on the inter-fertility among members of the same species [4]. This approach has the advantage of binding the definition of the species to an unambiguous test, which is insensitive to experimental techniques and assumptions and can therefore be considered rather universal among plants and animals. Most unfortunately, the biological species concept requires sexuality, a character not shared by all organisms and largely absent in the microbial world. All prokaryotes (*Eubacteria* and *Archea*) lack sexuality and what is sometimes denominated sexuality, the bacterial conjugation, is in fact a very sophisticated mode of horizontal gene transfer, i.e. a transfer of genetic materials between cells not involved in the reproduction process. It is possible, on the other hand, to discuss the sexuality of all microorganisms provided with the nucleus and organelles: the eukaryotic microorganisms including *Fungi*, *Algae* and *Protozoa*. Without entering too much into single subdivisions of these large taxonomic groups (kingdoms), for the purposes of this article, we will consider the well known case of *Fungi*. These microorganisms can be discriminated according to the sexuality as telomorph and anamorph. The former has a complete life cycle including the conjugation (two haploid cells form a diploid cell) and the meiosis (a diploid cell forms four haploid spores), the latter are not known to have sexuality and reproduce always in a vegetative way (either haploid or diploid). Interestingly, the telomorph is not compelled to reproduce sexually all the time, but alternate sexual and asexual reproduction, normally responding to environmental conditions. The largely accepted idea is that these microbes tend to reproduce asexually when in optimal conditions and tend to reproduce sexually when somehow stressed. These general observations lead to the conclusion that sexuality, if present, is not the only means of reproduction among eukaryotic microorganisms. This situation limits the application of the Biological Species Concept in microbiology. In fact, if the sexuality is not the only why to reproduce, then the species is not anymore confined by the reproductive barrier, in that every single cell can reproduce asexually without any limitation. Nonetheless, intensive studies have been undertaken to define some telomorphic microbial species on the basis of their interfertility

## An Overview of the Different Problems Inherent to the Species Concept in Microbiology

The specie's problem is primarily a problem on the existence of a structure similar to the species in the broad sense used in the biology of macroorganisms. This is normally called the ontological problem and is obviously much more challenging in microbiology than in the rest of biology, because there is little consensus on the general definition of species. Once the ontological problem is, partial, then one should consider the evolution problem (Fig. **1**). The species is considered as the unit of evolution, the set of organisms descending from one speciation event [5]. This problem is mainly a theoretical one: if the species were considered not to exist among microbes, then no speciation event can be assumed and all phylogenetic studies should be reconsidered in microbiology. From a more biological viewpoint, the problem occurs, when heritable changes occur within clones and bring to the splitting of two clones (Fig. **2**), whether this can be assumed as a real speciation or not as a part of the more complex problem of the microbial species. The third theoretical step is the semantic problem, i.e. "*how can we define species?*". This is not a mere problem of putting forward a good definition, but rather of choosing a biological criterion to define the microbial species. Here the challenge is also to find a criterion applicable to the rest of the living words if one wants to avoid the risk of definition pluralism, i.e. the situation in which each major group of organisms has its own species concept. The Biological Species Concept itself is based on a criterion that cannot be applied to all organisms, as demonstrated above and as recognized by its two major extensors [6, 7]. As this article is not the appropriate forum to propose and compare different criteria for defining species, we only consider that any criterion should be based on some form of discontinuity that separates one species from another.

After these three major theoretical problems, there are three more practical aspects to deal with: the classification, the technological and the analytical problem (Fig. **1**). The classification problem deals with several points, among which the choice between phenetics (classification based on the overall similarity) and phylogenetics (classification based on the evolutionary relationships) and the application of hierarchical or numerical systems. The numerical system was proposed by Michel Adanson in the 18$^{th}$ century assumes that all characters have the same weight and that each species can be defined by a unique string reporting the state of the considered descriptors. It is usually considered that failure of the initial method of Adanson was caused mainly by the lack of computers to compare the strings of all species. As a matter of fact, the numerical taxonomy has gained popularity after the introduction of the personal computers in the last decades.

The technological problem is treated with much attention by taxonomists. It deals with techniques to describe the microorganisms and thus the species. Presently, the techniques can be morphological, physiological and molecular. From the mathematical point of view, the descriptors can be either continuous, discontinuous or classified. Continuous data are seldom used in species description, although are rather important to characterize strains. Continuous descriptors can be the cell size, the physiological performances etc. Discontinuous and categorical descriptors are the most used in species description, examples are the DNA sequences and the binary descriptors indicating the presence (1 or +) or the absence (0 or -) of a given character. Species description is often based on some characters which should be considered continuous, but are classified for practical reasons. Cases are intermediate level of growth (classified as “slow” or “weak”), colony colors etc.

Last but not the least, the analytical problem deals with the statistical treatments to synthesize data in a easily perceivable way.

The six steps should not be considered as a fixed routine, but rather as a repetitive circular scheme to meliorate our understanding on the real nature of the species and on its practical applications. In fact, one could start with a hypothetical species concept, fitting with our general and evolutionary knowledge, that yields a definition and is applied in a classificatory criterion. Then the most convenient biological tools will be defined and analyzed appropriately with the best possible statistical approaches. Once the first round has been carried out, data will be available to tune more finely the species concept and to start the routine over and over again, in order to gain the best possible knowledge on the microbial species.

A practical aspect to underline the necessity of an approach based on successive approximations is the fact that without a larger knowledge on the *taxa* effectively present in nature, we currently lack a basilar aspects to deal with the problem of discontinuity. In fact, we cannot currently be sure whether some discontinuities among *taxa* (especially higher rank *taxa*) are due to real evolutionary jumps or simply to a lack of sampling. Microbiologists estimate to know a minimal part of the microbial diversity, maybe a figure ranging from 1 to 10%. These low values indicate an absolute need to work on the theoretical framework, but at the same time to extend the sampling and description of those *taxa* (maybe species) that are still unknown.

**Role and Importance of Discontinuities**

Seeking discontinuities is probably the main problem in the microbial species definition. In fact, the ontological, evolutionary and semantic problems could be summarized

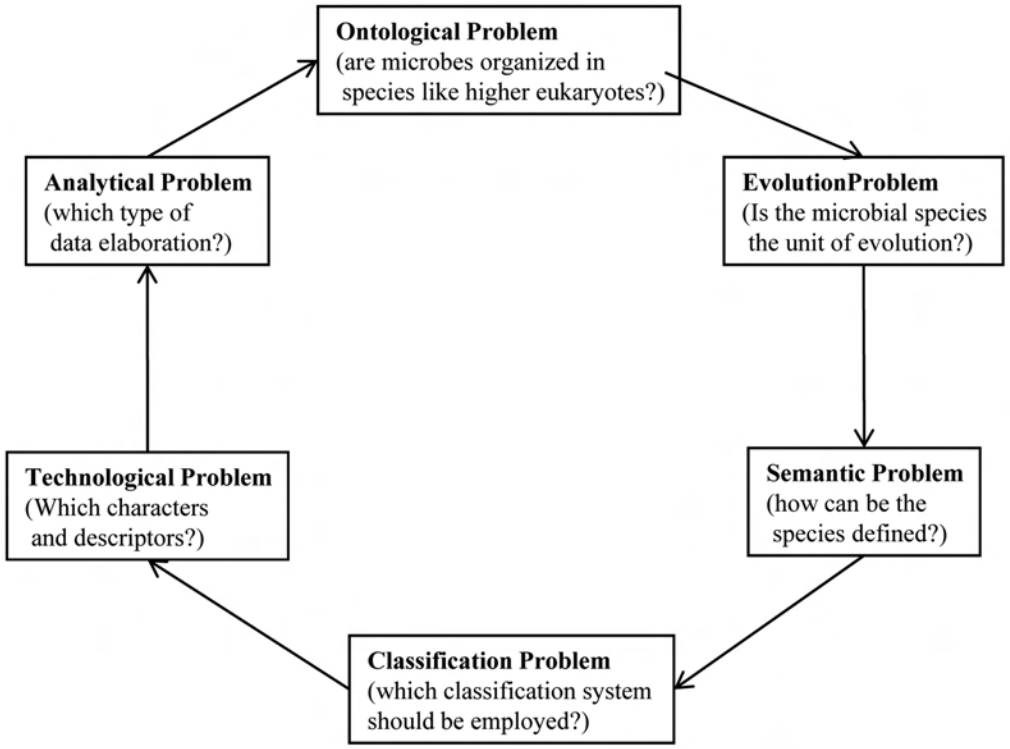


**Fig. (1).** Schematic flow of the problems involved in the species definition.

with the questions: "*do discontinuities exist? Do evolution changes occur at the species level? Which discontinuities can be used to define microbial species?*" Under a multi-disciplinary perspective, these questions should be addressed by the biologists along with experts in statistics, bioinformatics and epistemologists.

The major issue is probably demonstrating that indeed discontinuities exist and maybe to define clearly what a discontinuity at the species level should look like. Using categorical or discontinuous data, it is tempting to state that any change is a discontinuity. In other terms, two groups are different species if at least one descriptor has two different states as, for instance, the OTU1 and the OTU2 of Table **1**, differing only in the fourth descriptor. Such a definition is a strong one and probably the most objective criterion that can be used, in fact no difference exists with less than one mismatch, but any number of mismatches higher than one is likely due to some kind of convention and might not be shared. The argument against this criterion is that any strain would be a species. Using DNA sequences, instead of binary characters, would produce no conceptual difference. This simple argumentation shows that finding clear and shared definitions is a challenging and not so immediate task. On the other hand, the lack of shared definitions clearly hampers the scientific development.

**Table 1. Binomial Description of four OTUs with Four Descriptors**

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| OTU1 | 1 | 1 | 1 | 1 |
| OTU2 | 1 | 1 | 1 | 0 |
| OTU3 | 1 | 0 | 1 | 0 |
| OTU4 | 1 | 1 | 0 | 1 |

**Legend**: OTU stands for Operational Taxonomic Unit.
In many instances a OTU is a species.
D1, D2, D3 and D4 represent four binary descriptors.
The states 0 and 1 indicate lack or presence of the character.

A possible solution to this impasse could be found by switching the focus from the minimum distance to the minimum biologically relevant distance for two strain groups to be considered distinct species. Taxonomists deal with this issue since Linnaeus, in fact many classificatory systems are based on a hierarchy of characters starting with the most important ones (e.g. the reproductive organs in plants) and continuing with a long series of decreasingly important traits. The advantage of the system is that along the hierarchy of characters, a given level is designed to discriminate at the species level. This approach works reasonably well with higher organisms displaying a wealth of morphological rather stable characters. The situation in microbiology is complicated by the extremely low number of morphological traits, especially if limited to those observable with an ordinary light microscope. A solution to this limitation was the introduction of increasing numbers of physiological traits, represented mainly by the ability to assimilate or to ferment different carbon sources [8-11]. Defining which carbon or nitrogen source is more important than others is a difficult task, which will unlikely produce shared results. Invariant descriptors in state "1", as D1 in Table **1**, could be regarded as the most important because shared by all organisms considered and then probably absolutely necessary. Unfortunately, these indicators are useless just because they are invariant.

Another statistical and biological problem to consider is; if two OTUs sharing a descriptor in state "0" should be considered identical with respect to that specific trait (e.g. D4 of OUT2 and OTU3 in Table **1**). Statistics have provided different distance algorithms, some of which consider the "0" "0" a match, some other keeping this situation out of the calculation [12]. From a biological point of view, the absence of a character is likely due to many possible causes, suggesting that the "0" "0" should not possibly be considered in the computation of the similarity level, although it is obvious that the two organisms will behave identically regarding this trait.

This example introduces another problem: should taxonomy consider the functionality of the organisms or rather the genetic information that encodes for the various characters? Two organisms could be both lacking one function, say the ability to assimilate galactose, and therefore be considered somehow similar. However, there are several genes encoding the Leloir pathway, whose mutation can lead to the inability to grow on galactose. In other words, identical phenotypes could be caused by totally different genetic histories. This is a further argument against the concept that a single difference (especially if at the phenotypic level) can be considered enough for two OTUs to be considered different species.

The rest of this article will present a case study on the continuity within a group of yeast species, as a practical example of a possible approach to the ontological problem, i.e. to the question on whether discontinuities exist among related microbial strains.

## MATERIALS AND METHODS

### Sequences

Sequences were obtained from GenBank (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the BLAST algorithm within the Geneious software version 4.8.5 (http://www.geneious.com/) [13].

### Statistical Analyses

Statistical analyses were carried out in the open source programmable "R" environment (http://cran.r-project.org/) [14] with the addition of the packages VEGAN (http://cran.r-project.org/web/packages/vegan/index.html) [15], ADE4 (http://cran.r-project.org/web/packages/ade4/index.html) [16], APE (http://cran.r-project.org/web/packages/ape/index.html) [17] and ESTHER [18].

Plots and other graphics were prepared in the R environment.

The analyses with the *distconnected* function were carried out with the VEGAN package.

## RESULTS

### Sequence Dataset

The current system of identification in yeast biology is based on the D1/D2 domain sequence encoding the 26S ribosomal DNA (hereinafter referred to as 26S rDNA or

D1/D2 domain). GenBank contains a vast repository of such sequences and allows to retrieve them with the BLAST algorithm on the basis of the overall similarity to a "query" sequence.

As query sequence the D1/D2 domain sequence of the *Debaryomyces hansenii* type strain (CBS 767) was used, obtaining a hit list of 500 sequences then reduced to 486 by eliminating 14 too short entries. The Dataset was aligned and trimmed at the extremities in order to obtain a set of sequences of the same length (Database A). Sequences were listed according to the decreasing identity to the query. For each group of sequences with the same level of identity only one representative was left, producing a list of 52 sequences (Table **2**). The sequences were aligned and trimmed at the two extremities. Three sequences were removed because of containing ambiguities, yielding the final dataset of 49 sequences spanning from 100 to 93.3% identity (Database B). It is useful to highlight that, according to the current yeast taxonomy, members of the same species might share no less than 99% identities [19]. Consequently, this sequence alignment spans on a relative wide interval containing several yeast species.

### Seeking Discontinuities Among Sequences of the *D. hansenii* Group

The rationale of this investigation was to verify the pattern of genetic distances from a reference strain, in order to determine if hints of discontinuity existed and, more in general, to verify, whether such an approach can be taken for further extensive studies.

The aligned sequences of dataset B were imported in the R environment and the pairwise distances were calculated with the APE package using the *dist.dna* function (http://ape.mpl.ird.fr/) according to the "raw" method. The "square" distance matrix (i.e. the distance matrix with both upper and lower triangles plus the diagonal) was transformed in an object of class "matrix" and the column relative to the *D. hansenii* type strain CBS 767 was extracted. This vector included all distances of the 49 Dataset B members from the CBS 767 strain. The data of this vector were sorted in ascending order and plotted (Fig. **3**). The distances showed an evenly increasing trend with a few areas in which a subset of strains had essentially the same distance from the CBS 767. Only two regions presented gaps, namely after the first three strains and around the 25$^{th}$ strain. The seven strains identified as *D. hansenii* were obviously plotted in the lower left part of the graph with a maximum distance of five substitutions from the type strain (≥99%) and presented one such gap within their distribution. The distances of the strains of the closest species increased smoothly with a trend not different from that visible among the strains of the *D. hansenii* species. These observations suggested that, at least in this case, there was no evidence of discontinuities among the distances from one single reference strain. Another approach was taken in order to better investigate the pattern of distances present in Database B. All distances of this database (i.e. all the pairwise 49 *x* 49 distances) were sorted and plotted as dots in Fig. (**4**). The horizontal segments of Fig. (**4**) represent cluster of dots, i.e. of distances of the same value. Distances were calculated as mismatches (i.e. not identities in the alignment) and therefore increased by discrete steps of 1 mismatch. All together, the sorted distances showed a pattern without evidence of discontinuities and produced an even smoother plotting, as expected due to the larger number of data. The same analysis carried out with the distances of Database A produced analogous results (Data not shown).

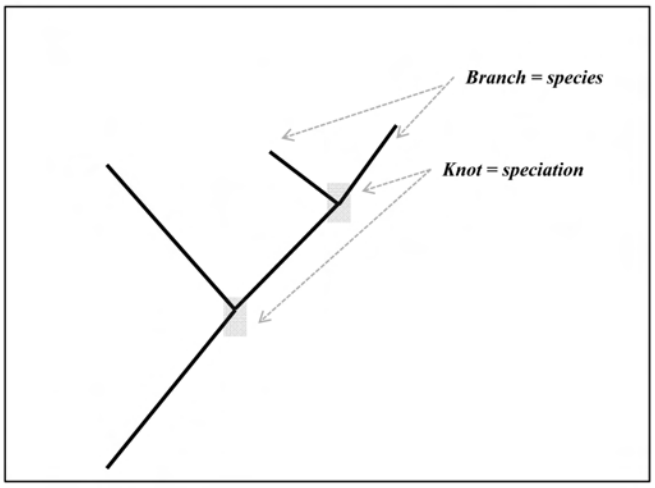


**Fig. (2).** Schematic differences between anagenesis and cladogenesis.

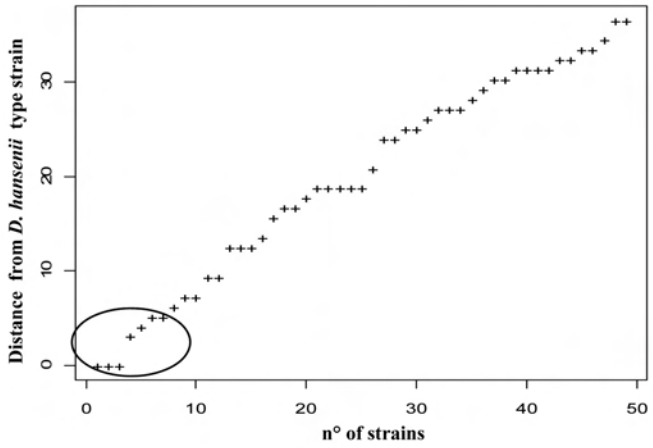


**Fig. (3).** Distribution in ascending order of the distances of 49 related strains with the *D. hansenii* type strain. The circle surrounds the strains identified as *D. hansenii* with no less than 99% identity.

### The *distconnected* Approach

The R package VEGAN includes an interesting function called *distconnected*, which seeks groups connected, according to the reciprocal pairwise distances. The algorithm disregards dissimilarities equal to zero at or above a threshold (*toolong*: an argument in the R function) chosen by the analyst. This algorithm simulates in some way the "classification" carried out in taxonomy when a given distance from the reference (type) strain is a criterion to include or exclude an isolate from a known species.

**Table 2. Results of the BLAST Search Using the *D. hansenii* TS (CBS 767) Sequence as Query**

| Seq. ID | Organism | %Pairwise Identity | Sequence Length |
|---|---|---|---|
| GU225766 | *Debaryomyces hansenii* isolate G346 | 100.00% | 534 |
| AJ508560 | *Debaryomyces hansenii*var. fabryi | 99.08% | 570 |
| AJ716109 | *Debaryomyces hansenii* | 99.6% | 528 |
| AF485978 | *Debaryomyces hansenii* A50 | 99.5% | 572 |
| AB385600 | *Debaryomyces hansenii* | 99.4% | 482 |
| FJ432605 | *Debaryomyces hansenii* strain 12-1 | 99.3% | 571 |
| AB438126 | *Debaryomyces nepalensis* | 99.2% | 503 |
| FJ475230 | *Debaryomyces hansenii* strain S08-1.2 | 99.01% | 571 |
| FJ986612 | *Debaryomyces nepalensis* | 99.00% | 572 |
| AY040651 | *Candida psychrophila* | 98.9% | 571 |
| U48844 | *Debaryomyces udenii* | 98.8% | 570 |
| AJ716116 | *Debaryomyces maramus* | 98.4% | 572 |
| FJ527166 | *Debaryomyces maramus* | 98.3% | 545 |
| FJ527170 | *Debaryomyces* sp.GY12S01 | 98.2% | 544 |
| FJ527175 | *Debaryomyces* sp.GJ14S01 | 98.0% | 545 |
| AF440014 | *Debaryomyces mycophilus* | 97.4% | 570 |
| AY520399 | *Candida* sp.BG02-5-27-1-2-C | 97.3% | 560 |
| AF440016 | *Debaryomyces mycophilus* | 97.2% | 570 |
| AY520396 | *Candida* sp.BG02-3-29-2-1 | 97.1% | 558 |
| DQ377632 | *Candida anglica* VTT C-04517 | 96.8% | 569 |
| GU213452 | *Saccharomycetes*sp.HZ10 | 96.7% | 568 |
| AB054994 | *Debaryomyces polymorphus* var. african | 96.6% | 564 |
| AJ539356 | *Candida beechii* | 96.4% | 562 |
| EU285537 | *Candida zeylanoides* strain SY6X-2 | 96.3% | 570 |
| FJ480853 | *Candida zeylanoides* strain 10C | 96.1% | 545 |
| EU359821 | *Candida zeylanoides* strain TJY7a | 96.0% | 554 |
| AF178052 | *Candida oleophila* strain CBS 8177 | 95.9% | 555 |
| EU131536 | *Candida zeylanoides* isolate W35 | 95.8% | 527 |
| EU131535 | *Candida zeylanoides* isolate W34 | 95.7% | 553 |
| AF257274 | *Candida railenensis* KCTC 7835 | 95.6% | 569 |
| AY242318 | *Candida* sp. BG01-8-20-001A-2-1 | 95.5% | 560 |
| EF452234 | *Candida oleophila* | 95.4% | 563 |
| DQ404496 | *Debaryomyces* sp. ST-310 | 95.4% | 570 |
| U45761 | *Candida shehatae*var. shehatae | 95.2% | 568 |
| AY731813 | *Candida oleophila* strain SDY 4.5.4 | 95.1% | 573 |
| FN667840 | *Candida oleophila* | 95.0% | 538 |
| DQ377636 | *Candida natalensis*strain VTT C-04521 | 94.9% | 572 |
| AB436395 | *Candida natalensis* | 94.8% | 572 |
| AB513345 | *Pichia*sp. MT-LUC0016 | 94.7% | 568 |
| DQ409151 | *Pichiasego biensis* strain CECT 10210 | 94.6% | 569 |
| AY332082 | Uncultured eukaryote clone | 94.5% | 568 |
| U39474 | *Cephaloascus albidus* | 94.4% | 575 |
| AY529522 | *Candida quercitrusa* isolate 129 | 94.2% | 569 |
| AY520393 | *Candida* sp. BG02-7-18-022A-1-1 | 94.1% | 564 |
| AM410999 | *Candida* sp. YS155 | 94.0% | 566 |
| FJ914895 | *Pichia spartinae* strain ATCC MYA-3201 | 93.9% | 570 |
| AB513344 | *Candida* sp. MT-LU0013 | 93.8% | 569 |
| AB361594 | *Candida palmioleophila* | 93.7% | 567 |
| AY518529 | *Candida athensensis* | 93.6% | 563 |
| AY845350 | *Candida lignicola* | 93.5% | 573 |
| FJ614693 | *Candida ascalaphidarum* | 93.4% | 572 |
| FJ196739 | *Candida athensensis* strain ATCC MYA-4479 | 93.3% | 571 |

**Legend.** Sequence length reported are those of the original sequences downloaded from the database before alignment and trimming.

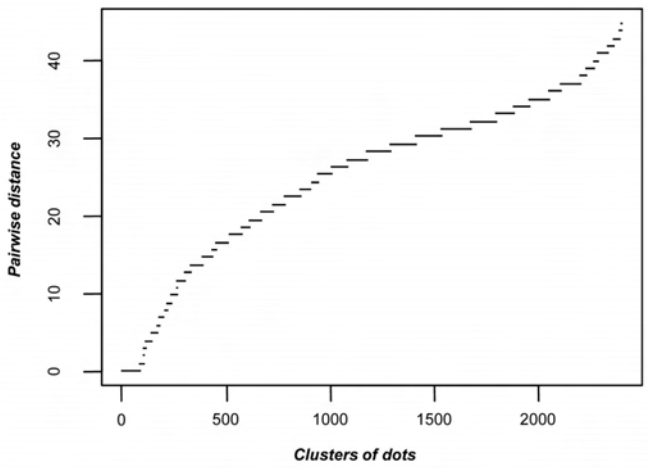


**Fig. (4).** Plotting of all distances of Database B in ascending order.

The analysis of the distance matrix obtained from the database B with *disconnected* produced different groupings according to the threshold distance (argument *toolong*) used as input (Fig. **5**). Threshold values were tested in the range from 1 to 12 mismatches (equivalent to approximately 0.1 to 2.5% distance). With 0.1% threshold value, the 49 strains were clustered in 36 groups, most of which included one single element, one included five strains and four two strains each (Fig. **5a**). Increasing the threshold value up to 2% the number of groups deceased from 36 to 15 (Fig. **5b**-**e**). Finally with *toolong* = 2.5% there were 12 groups some of which with more than ten strains (Fig. **5f**). This analysis showed as the choice of the threshold is critical to produce clusters of strains and that, as expected, increasing this value larger and fewer groups are obtained, although not necessarily in agreement with the biological delimitation of the species.

## DISCUSSION

The main aim of this article was to show that a complex problem such as that of the microbial species definition requires multidisciplinary answers and that bioinformatics and statistics could give important contributions. We could demonstrate that without a universally applicable criterion such as the biological species concept, only two approaches remain: the nominalistic definition of species based on some criteria, widely accepted but not necessarily related to the biological situation, and the search for the presence of discontinuities.

For practical reasons, the nominalistic approach has the advantage of simplicity and to provide a widely shared identification approach which can indeed help in the primary description of the microbial biodiversity. It is obviously not optimal from a general and theoretical viewpoint, but could be positively employed as a sort of first approximation to reach in future a better knowledge of the taxonomic structures of microorganisms.

On the other hand, discontinuities (or disconnections) could be a universally accepted to discriminate among microorganisms, but there are a series of challenges mostly for bioinformatics, statistics and epistemology. Firstly, one should define what a discontinuity (or a disconnection) should be and describe its properties in detail. It is largely possible that different types of disconnections are possible. Let's imagine, for instance, the people sitting in a room, if a

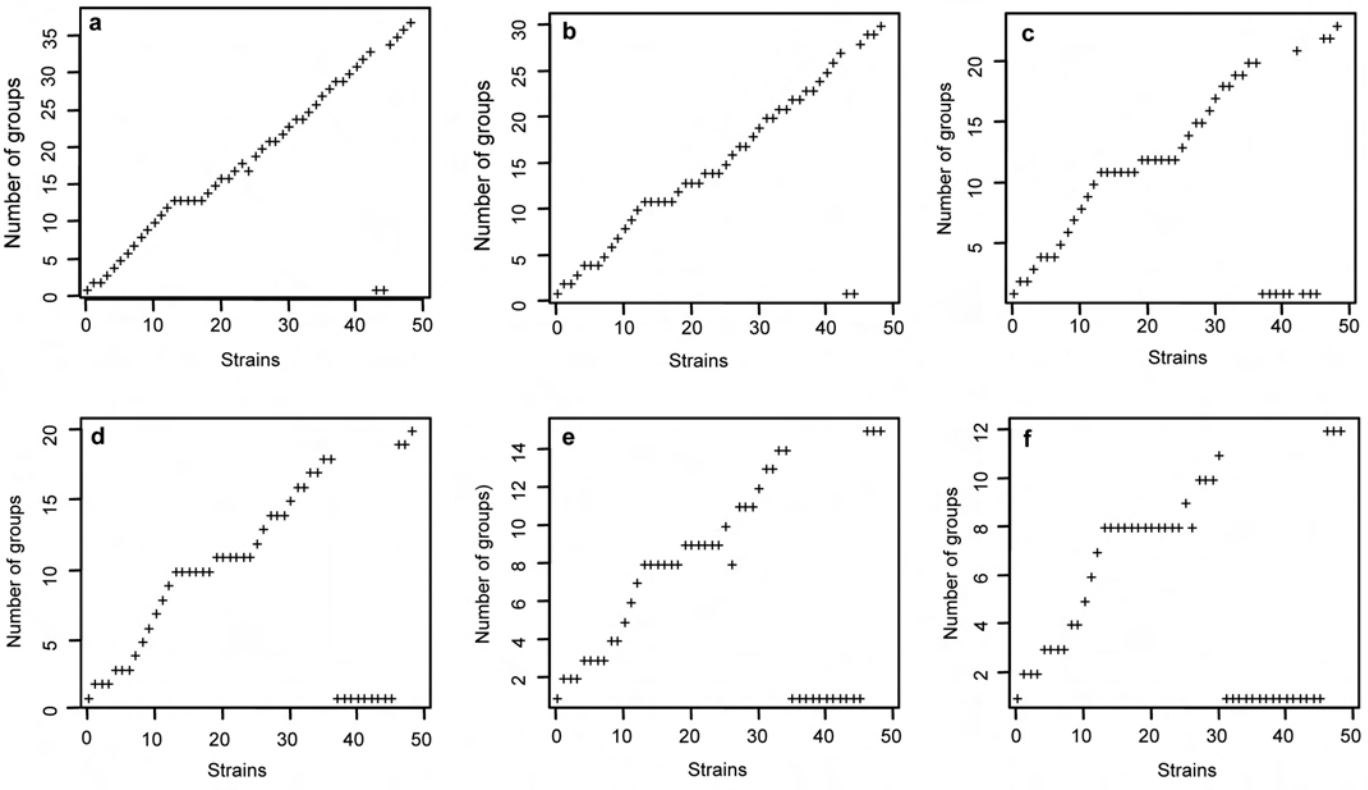


**Fig. (5).** Ordination of the 49 strains of the *D. hansenii* group according to the *distconnected* function. Panels show 6 different ordinations according to the toolong parameter which was set at 0.1, 0.5, 1.0, 1.5, 2.0 and 2.5% from panel **A** to panel **F**.

aisle is present, then at least two disconnected groups would be identifiable, but can we say that they are really separated? What do we mean by "separation" in biology? Another question is: can we really detect this disconnection? Again let's imagine the room example: the observer's position can allow to see or not the aisle, leading to different conclusions. We have suggested that the analysis of the reciprocal pairwise distances could be postulated as criterion and tested this hypothesis in a group of yeast species similar to *D. hansenii*. The results presented indicated that the distances from a reference strain increase almost smoothly and that disconnected elements in this situation are largely based on the distance threshold chosen, which is obviously a nominalistic approach. In another article of this Special Issue, we have dealt with the further problem of choosing a significant reference strain (the type) in order to avoid major problems of multivariate object grouping.

One could tentatively conclude that the taxonomic structure of yeast is rather continuous and that the nominalistic approach proposed is amply justified [20]. However, it is obvious that such a critical aspect of biology requires many more case studies and a whole set of statistical and bio-informatics to argue on the nature of the microbial species. Our intention was to show the complexity of the problem and to suggest some possible approaches of novel research lines for microbiologists, but also for non biologist's experts in biology, epistemology, statistics and informatics, interested to this question.

## ACKNOWLEDGEMENTS

L.A. and L.R. were supported by a Grant of the Italian Ministry of University and Research.

## CONFLICT OF INTEREST

None declared.

## REFERENCES

[1] M. Ereshefsky, "Darwin's solution to the species problem," *Synthese*, Vol. 175, pp. 405-425, 2009.

[2] J. Kupiec and P. Sonigo, *Ni Dieu ni gène. Pour une autre théorie de l'hérédité*, Seuil, Collection Science Ouverte, Paris, Seuil, 2000.

[3] C. Darwin, *The life and letters of Charles Darwin*, Basic Books, 1959.

[4] E. Mayr, *La biologie de l'évolution*, Hermann, Paris, 1981.

[5] M. Ridley, "The cladistic solution to the species problem," *Biology and Phylosophy*, vol. 4, pp. 1-6, 1989.

[6] T.G. Dobzhansky, *Genetics and the Origin of Species*, Columbia University Press, New York, 1951.

[7] E. Mayr, J. Hey, W.M. Fitch, and F.J. Ayala, *Systematics and the Origin of Species*, National Academies Press, 2005.

[8] N. Kreger-van Rij, "*The Yeasts, A Taxonomic Study"*, Amsterdam: Elsevier, 1984.

[9] C.P. Kurtzman, and C.J. Robnett, "Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences," *Antonie van Leeuwenhoek*, vol. 73, pp. 331-371, 1998.

[10] J. Lodder, and N. Kreger-van Rij, "*The Yeasts, A Taxonomic Study"*, Amsterdam: North-Holland, 1952.

[11] J. Lodder, "*The Yeasts, A Taxonomic Study"*, Amsterdam: North Holland, 1970.

[12] P. Legendre and L. Legendre, "*Numerical Ecology"*, Amsterdam: Elsevier Science *B.V*, 1998.

[13] A.J. Drummond, M. Kearse, J. Heled, R. Moir, T. Thierer, B. Ashton, A. Wilson, and S. Stones-Havas, "Geneious v2. 5," *Biomatters, Ltd., Auckland, New Zealand*, 2006.

[14] R Development Core Team, "R: A language and environment for statistical computing," *R: A Language and Environment for Statistical Computing*, 2010, Version 2.11.1. Available at: http://cran.r-project.org/

[15] J. Oksanen, R. Kindt, and B. O'Hara, "Vegan: R functions for vegetation ecologists," *Version*, vol. 1, pp. 8-3, 2006.

[16] D. Charif, J. Thioulouse, J.R. Lobry, and G. Perriere, "Online synonymous codon usage analyses with the ade4 and seqinR packages," *Bioinformatics*, vol. 21, pp. 545-547, 2005.

[17] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in r language," *Bioinformatics*, vol. 20, pp. 289-290, 2004.

[18] G. Cardinali, L. Antonielli, P. Rellini, and F. Fatichenti, "ESTHER: A "R Package" implementing a novel approach to bidimen-sional display of multidimensional binary data," *Open Appl. Info. J.*, vol. 1, pp. 20-27, 2007.

[19] C. Kurtzman, and C. Robnett, "Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences," *Antonie van Leeuwenhoek*, vol. 73, pp. 331-371, 1998.

[20] E. Stackebrandt, *Molecular Identification, Systematics, and Population Structure of Prokaryotes*, Springer-Verlag Berlin Heidelberg, Germany, 2006.

Received: April 29, 2010 Revised: October 10, 2010 Accepted: March 27, 2011

© Antonielli *et al.*; Licensee *Bentham Open*.

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.