DNA barcoding: theoretical aspects and practical applications

Maurizio Casiraghi, Massimo Labra, Emanuele Ferri,
Andrea Galimberti, Fabrizio De Mattia

Abstract — DNA barcoding is a molecular-based identification system, recently introduced in the scientific community. The method is not completely new to science, but the real innovation is not in the discrimination system itself: DNA barcoding can be considered as the core of an integrated taxonomic system, where bioinformatics plays a key role. Time is now ripe for a real collaboration of all the different forces working in taxonomy, towards a "next generation systematics".

Index Terms — DNA barcoding, DNA taxonomy, molecular identification, species identification.

◆

1 Introduction

The classification and monitoring of biodiversity are playing a key role in different contexts (e.g.: biological, social, economical), even if several aspects linked to these topics are far to be completely understood. A common assumption is that the central unit of taxonomy is the species, and the unequivocal association of a scientific name to a biological entity is an essential step to build a reliable reference system of biological information [1].

In the last 250 years, since Carl Linnaeus' classification system, about 1.7 million species have been formally described by taxonomists, but it is largely accepted that this number probably represents only a small fraction of the real biodiversity present on the planet (presently estimated in tens of millions of species) [2]. To help discovering this hidden biodiversity and in order to provide a useful and standardized tool for species identification, a molecular and bioinformatical tool called DNA barcoding has been proposed in 2003 [3].

The basic idea of this approach is quite simple (and not completely new

The authors are with the Department of Biotechnologies and Biosciences, University of Milan-Bicocca, Milan, Italy. E-mail:maurizio.casiraghi@unimib.it, tgalimba@gmail.com.

to science): through the analysis of the variability in a single or in a few standard molecular marker(s), it is possible to discriminate biological entities (hopefully belonging to the taxonomic rank of species). This method relies on the assumption that the genetic variation between species exceeds that within species. Consequently, the ideal DNA barcoding analysis mirrors the distributions of intra- and interspecific variabilities separated by a distance called 'DNA barcoding gap' [4], [5]. The original idea was to apply DNA barcoding systematically to all metazoans, by the use of one or a few (mitochondrial) markers (e.g. coxI, [1]). Rapidly, but with less coherent results, the idea was extended to flowering plants [6], [7] and fungi [8], and now the DNA barcoding initiative can be considered as a tool suitable for all of the tree kingdoms. Efforts in DNA barcoding development and management are coordinated by the Consortium for the Barcode of Life (CBoL; http://barcoding.si.edu/).

One of the major properties of a DNA barcode is the possibility to easily associate all life history stages and genders, to identify organisms from part/ pieces, or to discriminate a matrix containing a mixture of biological species. Quite soon it became clear that DNA barcoding was suitable for two different purposes: (1) the molecular identification of already described species [1], and (2) the discovery of undescribed species [9].

A lot of rumours raised around this approach, but what is the revolution introduced by DNA barcoding? In our opinion, the big leap forward is not only the discrimination power itself, but the joint use of three innovations of modern taxonomy: (1) molecularization (i.e. the use of variability in a molecular marker as a discriminator); (2) computerization (i.e. the non redundant transposition of data using informatic supports) and (3) standardization (i.e. the extension of the approach to vast groups of not deeply related organisms). For the first time, by DNA barcoding, it is possible to introduce in taxonomy a generalization, allowing researchers specialized in different fields to work on a shared framework.

In the space of few years, DNA barcoding has moved from fantasy to reality. In some of the first enthusiastic reports, DNA barcoding was even claimed as the way to make true the dreams of Gene Roddenberry, the creator of the science fiction drama Star Trek: the creation of a tool for organism identification, the DNA barcoder, as a homologous to the fictional Tricorder [10]. A few years later we are not yet in the spaceship Enterprise, but DNA barcoding has deeply impacted the scientific community, becoming a widely used approach.

Presently, the most relevant DNA barcoding tool, The Barcode of Life Data Systems, BOLD (http://www.barcodinglife.org/, [11]) is still in constant evolution and update.

2 THE 'DNA BARCODING MOLECULAR ENTITY' VERSUS 'SPECIES' DEBATE

Most of the questions raised by the use of DNA barcoding are directly linked to the essence of an identification method. In a strict sense, to identify simply means to differentiate, but the choice of the discriminator is essential, because the difficulty is in giving a biological meaning to what it has been discriminated [12].

Even if not always fully acknowledged, DNA barcoding implies two different

approaches to discrimination. DNA barcoding sensu stricto is a simple sorting method that could differentiate biological entities. It is not significantly different from a dichotomic key in the traditional taxonomical framework. On the other hand, DNA barcoding sensu lato represents a system that reflects the true sense of taxonomy. The discrimination method itself can be considered as an epiphenomenon - and the subject of major criticisms (DNA barcoding sensu stricto) - but it also becomes a system implementing all the aspects of taxonomy towards the representation of the living world as a whole (DNA barcoding sensu lato). It should be clear to users which kind of DNA barcoding philosophy they are going to adopt.

3 THE 'DNA BARCODING MOLECULAR ENTITY' VERSUS 'SPECIES' DEBATE

It is well known that no identification method (morphological, biochemical, genetic, etc.) can truly identify species, because species are entities in continuous evolution and it is theoretically impossible to define statically such a dynamic matter. DNA barcoding, in its original generalization, follows the typological species approach, a concept that theoretically fails because it freezes the evolutionary continuum of species. To cope with these limitations, some development of DNA barcoding has shifted towards other species concepts [13].

The entities identified by molecular approaches have been named in several ways: 'Genospecies'; 'Phylospecies', 'Recognizable Taxonomic Units', RTUs, 'Phylotypes' sensu, 'Molecular Operational Taxonomic Units', 'MOTUs' [12]. A general naïve assumption considers 'molecular entities' and 'species' as synonyms. This is the (almost) insurmountable problem for DNA barcoding sensu stricto: the biological meaning of the identified ranks cannot be directly derived, unless we have clearly and unequivocally linked a species to the variability pattern of a single DNA barcoding marker. In all other cases, we need DNA barcoding sensu lato [12].

The identification and then the interpretation of molecular entities is the main goal of DNA barcoding. This can be reached only by users with a sound theoretical background on what this technique is able to identify.

4 THE CHOICE OF THE BARCODE MARKER

DNA barcoding is not *coxI* only. A precise portion in the 5' end region of this mitochondrial gene has been proposed as a standard for metazoans [1]. Even if *coxI* has proven to be useful to discriminate species in most of the tested groups, its limits in some animal taxa are already evident (e.g. [14]). The choice of regions usable for DNA barcoding has been little investigated in many other eukaryotes. For instance, a marker was already available in fungi: the nuclear ITS region, which has been now confirmed as the main DNA barcode for this group [15]. In terrestrial plants, compared to animals, mitochondrial DNA has slower substitution rates and shows intramolecular recombination [16]. The search for an analogous to *coxI* or ITS in plants, that matches with the DNA barcoding criteria, has focused attention on the plastid genome. Several plastid genes have been proposed, such as the most conserved *rpoB, rpoC1*

and *rbcL* or a section of *matK* showing a rapid rate of evolution, but in some plant families these genes showed amplification problems. At the same time, the intergenic spacers such as *trnH-psbA, atpF-atpH* and *psbK-psbI* were also tested for their rapid evolution [17. Recently, the CBoL Plant Working Group [18] provided a recommendation on a standard plant barcode suggesting the 2-locus combination of *rbcL* and *matK*.

5 BIOLOGICAL AND BIOINFORMATICAL REPOSITORIES FOR DNA BARCODING DATA

DNA barcoding data are meant to be easily and widely accessed. To reach this aim, a proper sequence submission procedure is available for GenBank (http://www.ncbi.nlm.nih.gov/WebSub/?tool=barcode). This procedure slightly modifies the standard sequence submission procedure, introducing a DNA barcoding label to the sequence in order to simplify database querying and searching. Moreover, additional data are requested to link barcode sequence data to its voucher specimen. This standardization is mirrored by the establishment of the Registry of Biological Repositories initiative (http://www.biorepositories.org/), an on-line registry of organisms linked to DNA sequences. DNA barcoding sequences can also be deposited as projects in BOLD databases, characterized by an automatic submission tool to publish sequences to GenBank. By December 2009, BOLD database encompassed more than 760,000 sequences, corresponding to more than 65,300 formally described 'species'. The amount of data managed by the BOLD database is impressive: it collects, for a large amount of deposited barcode sequences, specimen details such as morphology, photographs, geographical distribution, collection points and others [11].

6 CONCLUSION

DNA barcoding is not a "perfect" method, but it has deeply impacted the scientific community, becoming a widely used approach, characterized by many relevant aspects of uniformity and generalization. A critical knowledge of the method is essential for a proper use of it.

REFERENCES

[1]   Q. D. Wheeler, "Taxonomic triage and the poverty of phylogeny". *Phil. Trans. R. Soc. Lond.* B, vol. 359, pp. 571-583, 2004.
[2]   R. Vernooy, E. Haribabu, M. R. Muller, J. H. Vogel, P. D. N. Hebert, et al., "Barcoding Life to Conserve Biological Diversity: Beyond the Taxonomic Imperative". *PLoS Biol.*, vol. 8(7): e1000417. doi:10.1371/journal.pbio.1000417, 2010.
[3]   P. D. N. Hebert, A. Cywinska, S. L. Ball, et al., "Biological identifications through DNA barcodes". *Proc. R. Soc. London, Biol. Sci. Series* B, vol. 270, pp. 313–321, 2003.

[4] C. P. Meyer and G. Paulay, "DNA Barcoding: error rates based on comprehensive sampling". *PLoS Biol.*, vol. 3, e422, 2005.

[5] M. Wiemers and K. Fiedler, "Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae)". *Frontiers in Zoology*, vol. 4, p. 8, 2005.

[6] W. J. Kress, K. J. Wurdack K., E. A. Zimmer, et al., "Use of DNA barcodes to identify flowering plants". *PNAS*, vol. 102, pp. 8369-8374, 2005.

[7] M. L. Hollingsworth, A. Clark, L. L. Forrest, et al., "Selecting barcoding loci for plants: evaluation of seven candidate loci with species level sampling in three divergent groups of land plants". *Mol. Ecol. Res.*, vol. 9, pp. 439-457, 2009.

[8] X. J. Min and D. A. Hickey, "Assessing the effect of varying sequence length on DNA barcoding of fungi". *Mol. Ecol. Notes.,* vol. 7, pp. 365–373, 2007.

[9] P. D. N. Hebert, E. H. Penton, J. M. Burns, et al., "Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*." *PNAS*, vol. 101, pp. 14812-14817, 2004.

[10] K. J. Gaston and M. A. O'Neill, "Automated species identification: why not?" *Phil. Trans. R. Soc. Lond.* B, vol. 359, pp. 655-667, 2004.

[11] S. Ratnasingham and P. D. N. Hebert, "BOLD: The Barcode of Life Datasystem (www. barcodinglife.org)". *Mol. Ecol. Notes*, vol. 7, pp. 355-364, 2007.

[12] M. Casiraghi, M. Labra, E. Ferri, A. Galimberti and F. De Mattia, "DNA barcoding: a six-question tour to improve users' awareness about the method." *Brief Bioinform.*, vol. 11(4), pp. 440-453. Epub 2010 Feb 15, 2010.

[13] J. M. Padial, A. Miralles, I. De la Riva and M. Vences, "The integrative future of taxonomy." *Front Zool.*, vol. 7, p. 16, 2010.

[14] T. L. Shearer and M. A. Coffroth, "Barcoding corals: limited by interspecific divergence, not intraspecific variation." *Mol. Ecol. Resour.*, vol. 8, pp. 247-255, 2008.

[15] R. H. Nilsson, M. Ryberg, E. Kristiansson, et al., "Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective." *PLoS One*, vol. 1, e59, 2006.

[16] G. D. D. Hurst and F. M. Jiggins, "Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts." *P. Roy. Soc. Lond.* B Bio, vol. 272, pp. 1525-1534, 2005.

[17] A. J. Fazekas, K.S. Burgess, P. R. Kesanakurti, et al. "Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well." *PLoS One*, vol. 3, e2802, 2008.

[18] CBoL Plant Working Group. "A DNA barcode for land plants." *PNAS*, vol.106, pp. 12794-12797, 2009.