R-Syst::diatom: a barcode database for diatoms and freshwater biomonitoring

data sources and curation procedure

Frédéric RIMET [1 - 2], Philippe CHAUMEIL [3 - 4], François KECK [1 - 2], Lenaïg KERMARREC [5], Valentin VASSELON [1 - 2] , Maria KAHLERT [6], Alain FRANC [3 - 4], Agnès BOUCHEZ [1 - 2]

1 INRA - UMR Carrtel, 75 av. de Corzent - BP 511, FR-74203 Thonon les Bains cedex, France
2 University of Savoie, UMR CARRTEL, FR-73370 Le Bourget du Lac, France
3 INRA, UMR BioGeCo, 69 route d'Arcachon, FR-33612 Cestas cedex, France
4 University of Bordeaux 1, UMR BioGeCo, FR-33400 Talence, France
5 ASCONIT Consultants, Naturopôle – Bât. C, 3, Bd de Clairfont, FR-66350 Toulouges, France
6 Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE- 750 07 UPPSALA, Sweden

Citation

Rimet F., Chaumeil P., Keck F., Kermarrec L., Vasselon V., Kahlert M., Franc A., Bouchez A., 2015. R-Syst::diatom: a barcode database for diatoms and freshwater biomonitoring - data sources and curation procedure. INRA Report, 14 pages

frederic.rimet@thonon.inra.fr

Introduction

We present the data sources and the curation procedure of an open-access and curated reference barcoding database for diatoms, called R-Syst::diatom, developed in the framework of R-Syst, the barcoding network of INRA (French National Institute for Agricultural Research). R-Syst::diatom links DNA-barcodes to their taxonomical identifications, and is dedicated to identify barcodes from natural samples. The data come from two sources, a culture collection of freshwater algae maintained in INRA which is regularly barcoded for new strains and from the NCBI (National Center for Biotechnology Information) nucleotide database. Two kinds of barcodes were chosen to support the database: 18S and *rbc*L, because of their efficiency. Data are curated using innovative (Declic) and classical bioinformatic tools (Blast, classical phylogenies) and up-to-date taxonomy (Catalogue of Diatom Names and peer reviewed papers). Every 6 months R-Syst::diatom is updated. The database is available through the R-Syst website (http://www.rsyst.inra.fr/). In addition to these information, morphological features (e.g. biovolumes, chloroplasts...), life-forms (mobility, colony-type) or ecological features (taxa preferenda to pollution) are indicated in R-Syst::diatom. This database should get a foot in the door of biomonitoring 2.0.

Data sources

Two data sources are used to fill R-Syst::diatom: the barcoded strains of the Thonon Culture Collection (TCC) and the nucleotide database of NCBI.

Barcoded strains of the TCC:

The UMR-CARRTEL is a research unit of the French National Institute for Agricultural Research (INRA) working on aquatic ecosystems. It has maintained the TCC since 1968, which is registered to the World Data Centre for Microorganisms (# 1030) and to the Global Registry Biorepository (http://grbio.org/institution/thonon-culture-collection-umr-carrtel-inra). A total of 858 monoclonal strains of freshwater microalgae are registered, among which 505 are diatoms. For each culture, DNA extracts and raw material are kept in the UMR-CARRTEL. Moreover, for diatoms, at least one permanent slide (Naphrax) of clean frustules as well as nitric acid treated material (in a vial). This material is accessible for subsequent studies. Two hundred nineteen diatom strains are maintained as live cultures as of Feb 2015, the oldest was isolated in 1985 and the most recent in 2015. These strains are available on request through a website dedicated to the collection (http://www6.inra.fr/carrtel-collection_eng/). Each strain is sequenced for at least two barcodes: 18S and *rbc*L. Several research programs financed the isolations and sequencings (see acknowledgments). All information about these strains, the sampling site location (georeferenced on a google map), the isolator, the barcode (including type of barcode, amplified region, primer used, protocols), the phenotypic data, the photos, the associated research programs (for sampling and sequencing) and its taxonomic affiliation are available on the R-syst website (http://www.rsyst.inra.fr/). The strains are identified using updated literature such as the entire collection of Diatoms of Europe, Iconographia Diatomologica, Bibliotheca Diatomologica and peer reviewed papers.

The TCC is regularly enriched with new isolated strains, which are sequenced for at least the two barcodes (18S, *rbc*L). Their entry in R-Syst::diatom is submitted to the curation process described in the section "Data curation" here below.

Nucleotide database of NCBI:

NCBI (National Center for Biotechnology Information) maintains a webserver that collects and provides molecular data and software. In particular, NCBI allows access to all public DNA sequence data via the GenBank database ([1], http://www.ncbi.nlm.nih.gov/genbank). We recovered all the nucleotide sequences of diatoms (freshwater and marine) available on GenBank main collection (CoreNucleotide) for the 18s (including V4-region of 18s) and *rbc*L whatever their length and their quality. We limited ourselves to these markers because they showed good abilities for species identification (e.g. [2] [3] [4] [5]), they provide access to the largest taxonomic diversity and showed the best results for metabarcoding ([6], [7], Zimmermann et al. 2014b). The sequences for the 28S, the ITS and *cox*1 were not gathered in the database.

These sequences are retrieved regularly (every 6 months) using the following keywords on the Nucleotide Advanced Search Builder: "(18s OR rbcL) and (diatom OR Bacillariophyta)". In addition to these keywords, a publication interval in NCBI is indicated in the Advanced Search Builder: the oldest is corresponding to the last R-Syst::diatom update and the most recent to the current date. R-Syst::diatom is thus updated every 6 months. As well as the barcodes coming from the TCC, their entry in R-Syst::diatom is submitted to the curation process described in the next paragraphs.

Phenotypic data

For most species, three kind of phenotypic data are given: -1- morphological -2- life-form and -3- ecological information.

-1- Morphological data are gathering information about chloroplast shape and number. Bibliographical references are given for each taxon, most of the time, the publication of [8] was used. When possible, photos of the strains were look at to get such information. Cell-dimensions (length, width, thickness), biovolume and size-class are given. These information are most of the time coming from [9] which is a database gathering morphological and ecological information about freshwater diatoms. Omnidia [10] database which is gathering information about cell-biovolumes and sizes was also used. Original references where such information can be found are given.

-2- Even if diatoms are basically unicellular algae, they exhibit an important diversity of life-forms, and many of them can form colonies. Taxa can even present several successive life-forms during their life-cycle (e.g. *Cymbella* can be unicellular and free moving at a time and attached to a peduncle and then immobile at another time). Different kinds of life-forms information are documented in R-Syst::diatom [9], such as motility, kind of colony, type of attachment (pad, stalk, adnate, pedunculate).

-3- Several kind of ecological information are given. Nutrients, organic matter and moisture preferences of the species according to [11] are given. Habitat preferences (benthic, planktonic, epipsamic, epipelic) are given mostly according to [12]. Ecological guilds belonging (high-profile, low-profile, motile, euplanktonic) are given according to [9]. Finally, pollution sensitivity values and ecological weight of several diatom indices are given, such as the TDI - Trophic Diatom Index- [13], the TDI-Sweeden, [14], the IPS -Pollution Sensitivity Index- [15] or the Phylogenetic-IPS [16].

Data curation

Diatom identifications and sequencing which potentially feed R-Syst::diatom were carried out by different operators, and they may suffer significant heterogeneity. There are three important drawbacks to take into account when gathering new sequences in R-Syst::diatom:

- First, in NCBI, data were deposited by different authors at different times: the first data were deposited in 1998. From this date to the present, taxonomy has evolved.

- Second, the identifications and taxonomic skills of the different authors who deposited their data on NCBI can be shifting. The same problem is also visible for TCC.

- Third, the length or the quality of the sequences cannot be adapted for correct taxonomic affiliation.

These three drawbacks are necessary to take into account and underline the necessity to curate the taxonomic names of the strains and their corresponding sequences in order to have homogeneous taxonomic names in R-Syst::diatom. The aim is to achieve for similar sequences a similarity in their taxonomic names that are as taxonomically correct as possible according to the most recent taxonomy literature. However, as diatom taxonomy is under active development, there will be cases where only a consensus for practical use can be made and solutions regarding the correct name will have to await further scientific studies. In any case, if the original taxonomic name given by the authors of the sequence was changed during the curation procedure, the traceability of the original name is kept in the database and is visible on R-Syst web portal.

This data curation is carried out in three steps (the first two steps are mandatory):

- The 1st step is pre-curation. The objective of this 1st step is to check if each newly retrieved sequence from NCBI or the TCC have comparable taxonomic names with similar sequences formerly deposited in NCBI and if the quality and length of these sequences is correct. For this purpose, the newly retrieved sequences are compared to the entire NCBI database using Blast.

- The 2nd step is detailed curation. The objective of this 2nd step is to compare the new sequences which met criteria of the 1st step to the sequences that are already in R-Syst::diatom, based on a local alignment methodology, called "Declic algorithm" (for detail see "Second curation step" section of this "Data curation" part). If these sequences have taxonomic names similar to comparable sequences then their taxonomic name and the sequences is kept for the 3rd curation step. If the taxonomic names are different from comparable sequences, the taxonomic name is checked through a taxonomic curation procedure.

- The 3rd step, is an optional curation (for reasons see "Third curation step" of the "Data curation" part). The objective is to compare the new sequences, which met the criteria of the 2nd step to those of the R-Syst::diatom database, based on a global alignment and phylogenetic analyses. From these analyses taxonomic names of similar sequences are homogenized if necessary using the same taxonomic curation procedure.

Figure 1 gives an overview of the general workflow of the curation procedures. Figure 2 gives details on the taxonomic curation procedure which is used several times in the general workflow of Figure 1.

First curation step: pre-curation using NCBI and Blastn of each sequence

For each sequence (whatever its quality or length), newly gathered from NCBI or coming from new strains of the TCC, a Blastn is run on the entire NCBI database. The 20 sequences showing the best pairwise identity matching to this newly sequence are consulted.

If the taxonomic affiliation of the new sequence is close to those of the 20 other sequences then the taxonomic affiliation is kept and this new sequence is kept for the second curation step. Taxonomic affiliation designated here is not necessarily the species level: it can be the genus or family level in the case of newly isolated genera or families never isolated before.

If there is a discordance between the taxonomic affiliation of this new sequence and those of the 20 other sequences, then the taxonomic curation procedure is applied (fig. 2) :

(1) First, we check if a peer-reviewed publication is associated with this new sequence. In this case this new sequence and its taxonomic name are kept for the second curation step. If it is not the case, then point (2) of the taxonomic curation procedure is considered.

(2) If no peer-reviewed publication is available, taxonomic synonymies are checked using Algaebase website (http://www.algaebase.org/) [17], the catalogue of diatom names of E. Fourtanier & P. Kociolek (http://researcharchive.calacademy.org/research/diatoms/names/index.asp) or Omnidia software [10]. If this enables the homogenization of its taxonomic name, then the new sequence and its new taxonomic name are kept for the second curation step. If it is not the case, then point (3) of the taxonomic curation procedure is considered.

(3) If no peer-reviewed publication exists, and if the taxonomic synonymies check was not successful, we check if some photos or slides associated to the sequence are available (e.g. in TCC or Algaterra databases -[18]- or other websites). If the re-examination of this material (photos/slides) shows that the strain was wrongly identified then a correct taxonomic name is given. If this new taxonomic name is similar to those of the 20 most similar NCBI sequences, this sequence is kept for the 2nd curation step. If it is not the case, -no photos/slides are available or the new taxonomic name still different from those of the 20 most similar sequences-, then this new sequence is not accepted in the database.

After gathering all the new sequences from the first curation step, additional curation steps are done by bringing them face to face to the sequences already in the R-Syst database. Two different and complementary tools are used. The first tool is Declic analyses (second curation step). This analysis is based on local alignments which are useful when sequences of dissimilar sizes have to be compared, which. If this is the case of the data usually gathered: depending on the authors, only parts of 18s/rbcl are sequenced. The second tool is phylogenetic trees (third optional curation step) based on global alignments. Global alignments are more useful when sequences of similar sizes are compared and are carried out on a sub-set of sequences of homogeneous size and similar regions.

Second curation step: use of Declic on the entire database:

For each marker (18S and *rbc*L) a Declic analysis is carried out on all sequences (new sequences and R-Syst database sequences). It can be run with an R-package [19] or under a galaxy platform with in the Virtual BiodiversityL@b [20] (https://galaxy-pgtp.pierroton.inra.fr/). Briefly, Declic analysis is run after computing pairwise local alignment score [21] which are then transformed into a distance. We then have a full pairwise distance matrix. Pairwise distances can be visualized by running MDS on the distance matrix. Second, a graph is attached to the matrix, where the nodes are the sequences, and there is an edge between two nodes if the distance between the two sequences is lower than a given threshold (one graph per threshold). The threshold is selected by the user [generally a threshold around 1% for species level is selected according to [22] and [3]], and a graph can be built for any threshold value. The graph is projected onto the plane using the Fruchterman–Reingold layout [23]. If edit distances were evolutionary distances, and if a threshold exists for separating taxa, then a taxon would be a clique (i.e. in this case a subset of sequences which are all connected with each other by an edge, see fig. 3a). As we have edits distances from best local alignment, we built the connex components (i.e. in this case a subset of sequences which are connected by at least one edge, see fig 4b) of such a graph, expecting they are close to cliques, and related to taxa. Such a threshold may play the role of a barcoding gap, although some sequences within a connex component can be at a distance larger than the gap. Colors are given to sequences belonging to the same taxon. The taxonomic levels which are selected for data curation are genus level or species level.

It is expected that a new sequence has a homogeneous taxonomic name with the other sequences of the clique it belongs to. If it is the case, the sequences of this clique and their taxonomic names are kept for the 3rd curation step.

If the new sequence and the other sequences have heterogeneous taxonomic names inside the same clique, then the taxonomic curation procedure is applied (fig 2) in a similar way as in the first curation step:

(1) We check if a peer-reviewed publication is associated with the new sequence. If it is the case, based on the results of the publication, the taxonomic names of the sequences are homogenized in the clique and the sequences are kept for the 3rd curation step. If no peer-reviewed publication is available, then point (2) is considered.

(2) In the case that there is no publication available for the new sequence, then the synonymies of the taxonomic names of the sequences inside this clique are checked (using i.e. Algaebase, Catalogue of diatom names, Omnidia). If this enables us to homogenize the taxonomic names, the sequences and their new taxonomic names are kept for the 3rd curation step. If this is not the case, point (3) is considered.

(3) If no publication is associated to this new sequence and if the synonymies check did not enable us to homogenize taxonomic names, we check if photos/slides associated to it (collections as TCC, Algaterra, Bold) are available. If the re-examination of the photos/slides enable us to change the taxonomic names and make it similar to the those of the other sequences in the clique, the sequence and its new taxonomic name is kept for the 3rd curation step. If the taxonomic name is still different after checking photos/slide, the sequence is rejected. The taxonomic synonymies are checked. If this enable us to homogenize the taxonomic names, the sequences and their new taxonomic names are kept for the 3rd curation step.

Connex component which were not cliques are also checked. Sequences belonging to the same connex component should show homogeneous taxonomic names. If not, the same procedure (described here above) for cliques was adopted (check of literature, synonymies, photos ...).

3rd curation step (optional): Phylogenetic analyses:

Similarly as the Declic analyses, phylogenetic analyses are carried out for each marker (18S and *rbc*L) on all sequences (new sequences and R-Syst::diatom). A general alignment is carried out on all the sequences with Muscle in Seaview [24]. The best compromise between removing the shortest sequences and trimming the alignment is found in order to keep an alignment long enough to get phylogenetic analyses robust enough. Usually, for 18S and *rbc*L the alignment is carried out on 1000-1500 bp and so shorter sequences are not taken into account in this curation step. From this general alignment, a neighbor joining tree is run with Seaview [24] or Mega5 [25]. The same verifications as those carried out in the 2nd curation step are done: if taxonomic names in a given clade are heterogeneous, the taxonomic curation procedure is applied (fig. 2).

These phylogenetic analyses are done to confirm the curation completed with Declic analyses. Nevertheless, phylogenetic analyses are carried out on a sub-set of the database, since short sequences are not integrated in this analysis. The shortest sequences, which were not integrated in the phylogenetic analyses, are only curated with the Declic analyses and if they meet all the criteria of taxonomic homogeneity in the 2nd curation step they are directly integrated in R-Syst::diatom.

Data storage and open access

All the curated data are stored in a PostgresSQL database built in the frame of the R-Syst network. R-Syst is a French INRA network of several tens of research teams including technicians, researchers and engineers in the fields of molecular biology, genetics and bioinformatics who are involved in the

molecular and morphological characterization of organisms. Among those, micro-algae are represented and a dedicated web interface is available from the R-Syst web portal (http://www.rsyst.inra.fr/en) to browse the stored data of the diatom barcoding database.

On this website, the algae section of the database gathers information about diatom strains (but also of several Chlorophyta and Cyanophyta strains of the TCC) which were characterized for three kinds of criteria: taxonomic, phenotypic and genetic.

For each strain, the following information are given when available: (i) sampling site origin (name and location on Google map), (ii) type of habitat, (iii) strain code given by the laboratory, (iv) name of the project which funded for field sampling, sequencing, (v) laboratory responsible for field sampling, (vi) DNA extraction, (vii) PCR, sequencing, and (viii) the dates of the different steps. A species name is given to each strain, except in a few case where only genus level is given. Moreover the taxonomic affiliation is given until the regnum following [12], [26] and [17]. For molecular criteria, the database gives the type of marker (18S or *rbc*L), the primers used for sequencing and PCR. Protocols for DNA extraction and PCR are also given. The laboratory responsible of the sequence is given. For phenotypic information, photos (living material and empty frustules) of the TCC strains are given.

Reference List

1. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Research 41: 36-42.

2. Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). Protist 158: 349-364.

3. Hamsher SE, Evans KM, Mann DG, Poulickova A, Saunders GW (2011) **Barcoding Diatoms: Exploring Alternatives to COI-5P** . Protist 1-18.

4. Kermarrec L, Bouchez A, Rimet F, Humbert JF (2012) Using a polyphasic approach to explore the diversity and geographical distribution of the *Gomphonema parvulum* (Kützing) Kützing complex (Bacillariophyta). Protist in review.

5. Zimmermann J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org Divers Evol 1-20.

6. Kermarrec L, Franc A, Rimet F, Chaumeil P, Humbert JF, Bouchez A (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Molecular Ecology Resources 13: 607-619.

7. Kermarrec L, Franc A, Rimet F, Chaumeil P, Frigerio JM, Humbert JF, Bouchez A (2014) A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshwater Science 33: 349-363.

8. Cox EJ (1981) The use of chloroplast and other features of living cell in the taxonomy of naviculoid diatoms. 6τη: 115-133.

9. Rimet F, Bouchez A (2012) Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. Knowledge and Management of Aquatic Ecosystems 406: 1-14.

10. Lecointe C, Coste M, Prygiel J (1993) "Omnidia": software for taxonomy, calculation of diatom indices and inventories managment. Hydrobiologia 269/270: 509-513.

11. Van Dam H, Mertens A, Sinkeldam J (1994) A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology 28: 117-133.

12. Round, F., Crawford, C. G., and Mann, D. G (1990) The diatoms. Biology and morphology of the genera. Cambridge University Press. 747 p.

13. Kelly MG, Whitton BA (1995) The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. Journal of Applied Phycology 7: 433-444.

14. Kahlert M (2015) TDI-Sweeden, Trophic Diatom Index adapted to Swedish rivers.

15. Cemagref (1982) Etude des méthodes biologiques quantitative d'appréciation de la qualité des eaux. Rapport Q.E.Lyon-A.F.Bassin Rhône-Méditerranée-Corse.: 1-218.

16. Keck F, Rimet F, Franc A, Bouchez A (2015) **Developing phylogenetically based biomonitoring methods: a test with diatoms**. In prep .

17. Guiry MD, Guiry GM (2014) *AlgaeBase*. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 24 november 2014.

18. Kusber WH, Abarca N, Skibbe O, Zimmermann J, Jahn R (2012) Reference library of DNA-barcoded diatoms - A use case for publishing data via the GBIF database AlgaTerra. - S. 65.

19. Franc A (2014) A primer for MIAB. 1-31.

20. Laizet YC, Chaumeil P, Frigerio JM, Bouchez A, Kermarrec L, Rimet F, Franc A (2014) **Virtual_BiodiversityL@b and "*Declic*": A userfriendly galaxy platform with tools for species delineation and taxonomic annotation with genomics and metagenomics**. In prep .

21. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.

22. Evans KM, Mann DG (2009) A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. Phycologia 48: 70-74.

23. Fruchterman TMJ, Reingold EM (1991) Graph Drawing by Force-Directed Placement. Software - Practice & Experience 21: 1129-1164.

24. Gouy M, Guindon S, Gascuel O (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Molecular Biology and Evolution 27: 221-224.

25. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 28: 2731-2739.

26. Monnier O, Coste M, Rosebery J (2009) Une classification des taxons de l'Indice Biologique Diatomées (IBD, norme AFNOR NF T90-354, décembre 2007). Diatomania 13: 17-47.

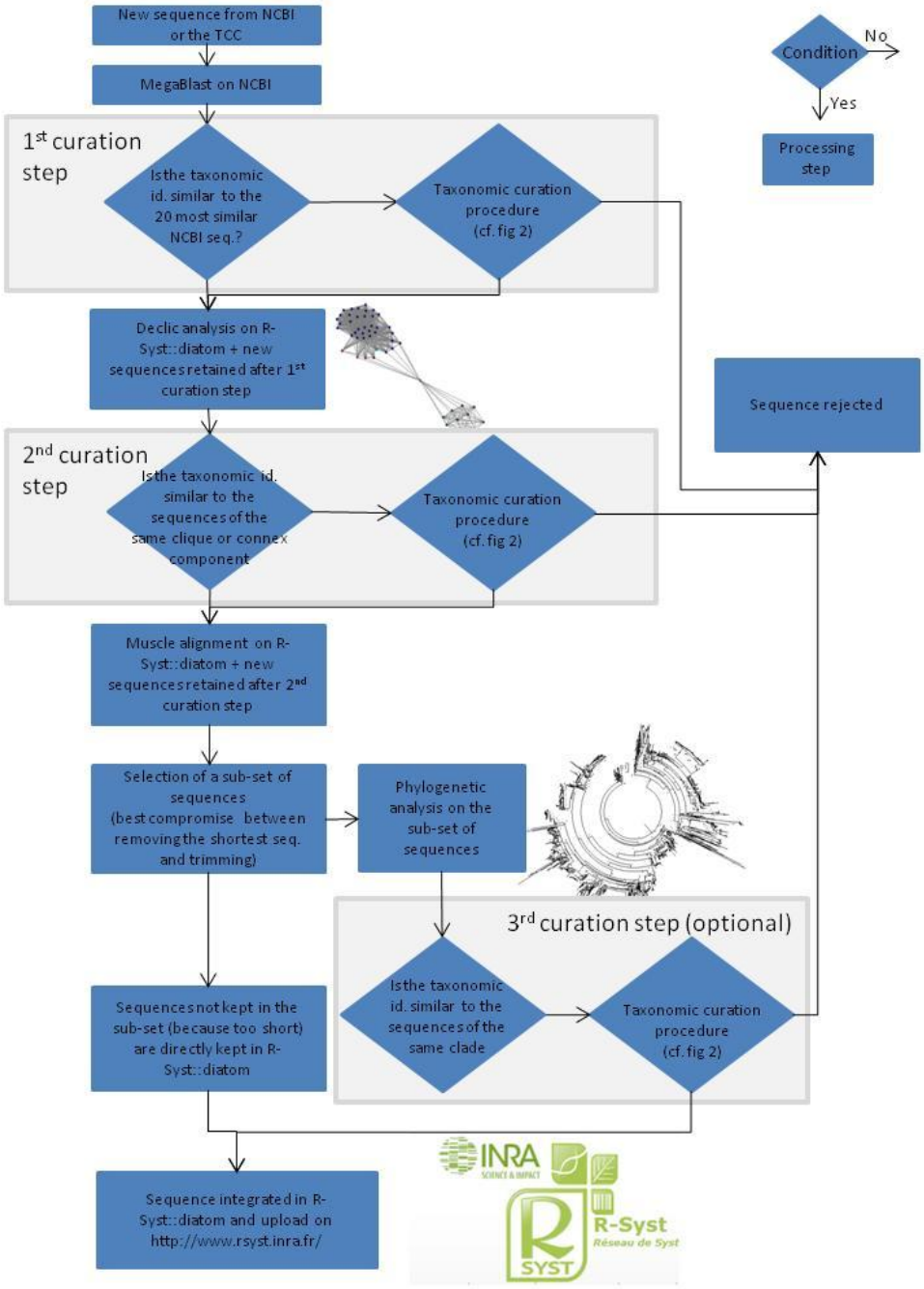

Figure 1: General flowchart of the curation and integration of new sequences in the R-Syst::diatom. Taxonomic curation procedure is detailed in a flowchart (fig 2). Diamonds are conditions, the arrow from the bottom point of the diamond corresponds to « Yes », the arrow from the right point of the diamond corresponds to « No ». Rectangles are processing steps.

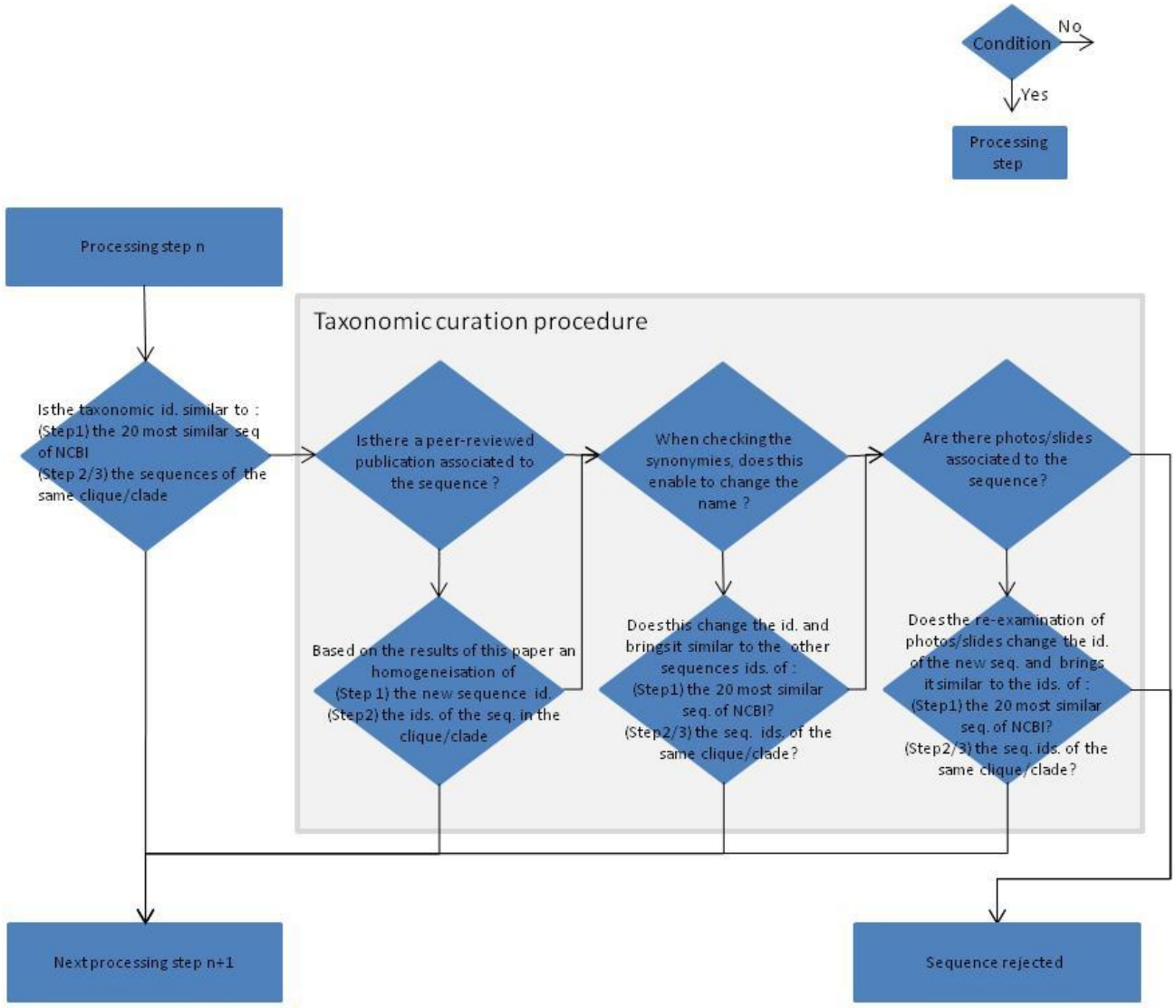

Figure 2: Flowchart of the taxonomic curation procedure. Diamonds are conditions, the arrow from the bottom point of the diamond corresponds to « Yes », the arrow from the right point of the diamond corresponds to « No ». Rectangles are processing steps.

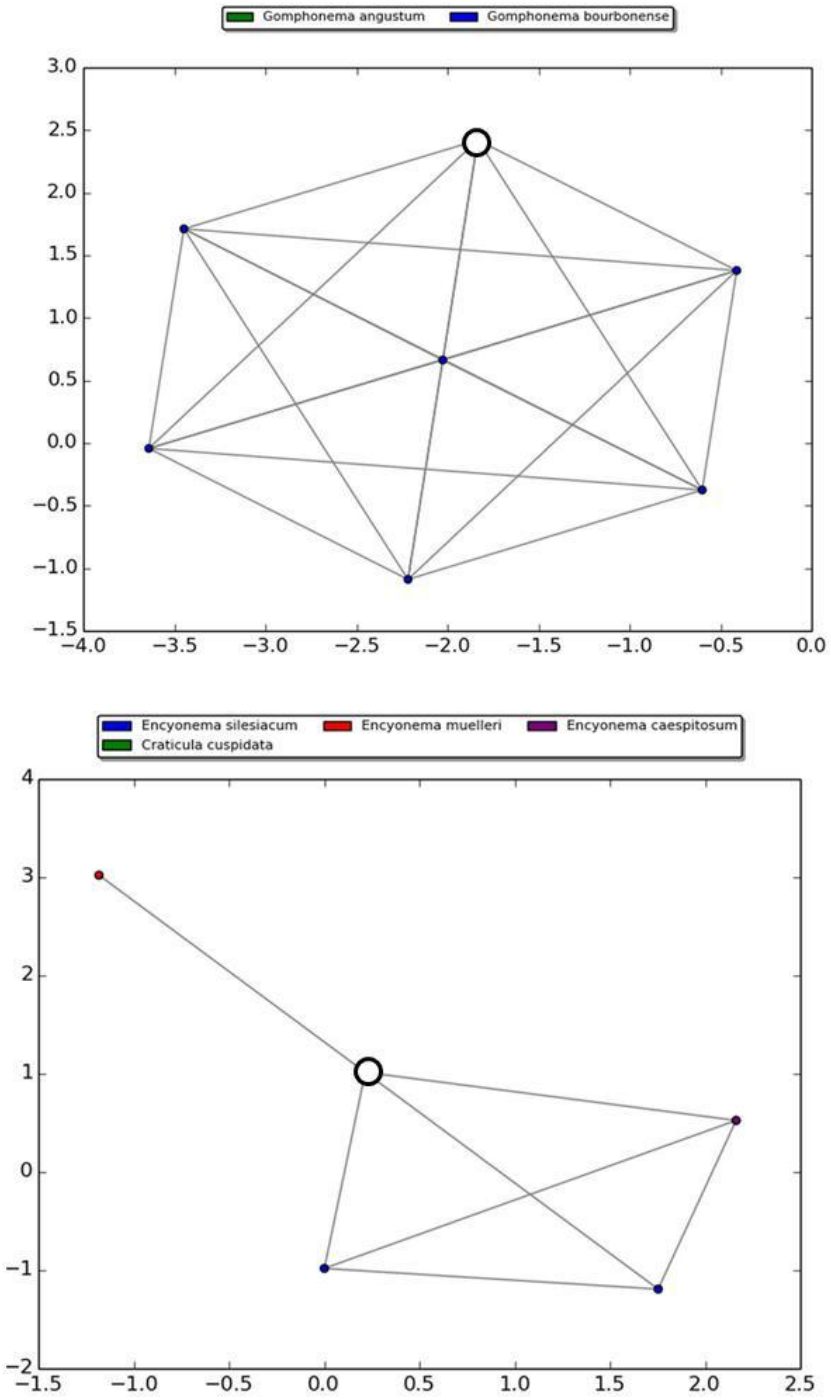

Figure 3: Use of Declic analyses to curate the database: case of taxonomically a heterogeneous clique (a) and connex component (b). (a) *Gomphonema bourbonense* clique (18s, gap 8) with one *G. angustum* sequence (TCC460) -white circle- and (b) *Encyonema* spp. connex component (18s, gap 8) with one *Craticula cuspidata* sequence (KM084917) -white circle-. TCC460 strain identification was changed into *G. bourbonense* after checking photos. KM084917 was rejected since there is an obvious mistake of identification.

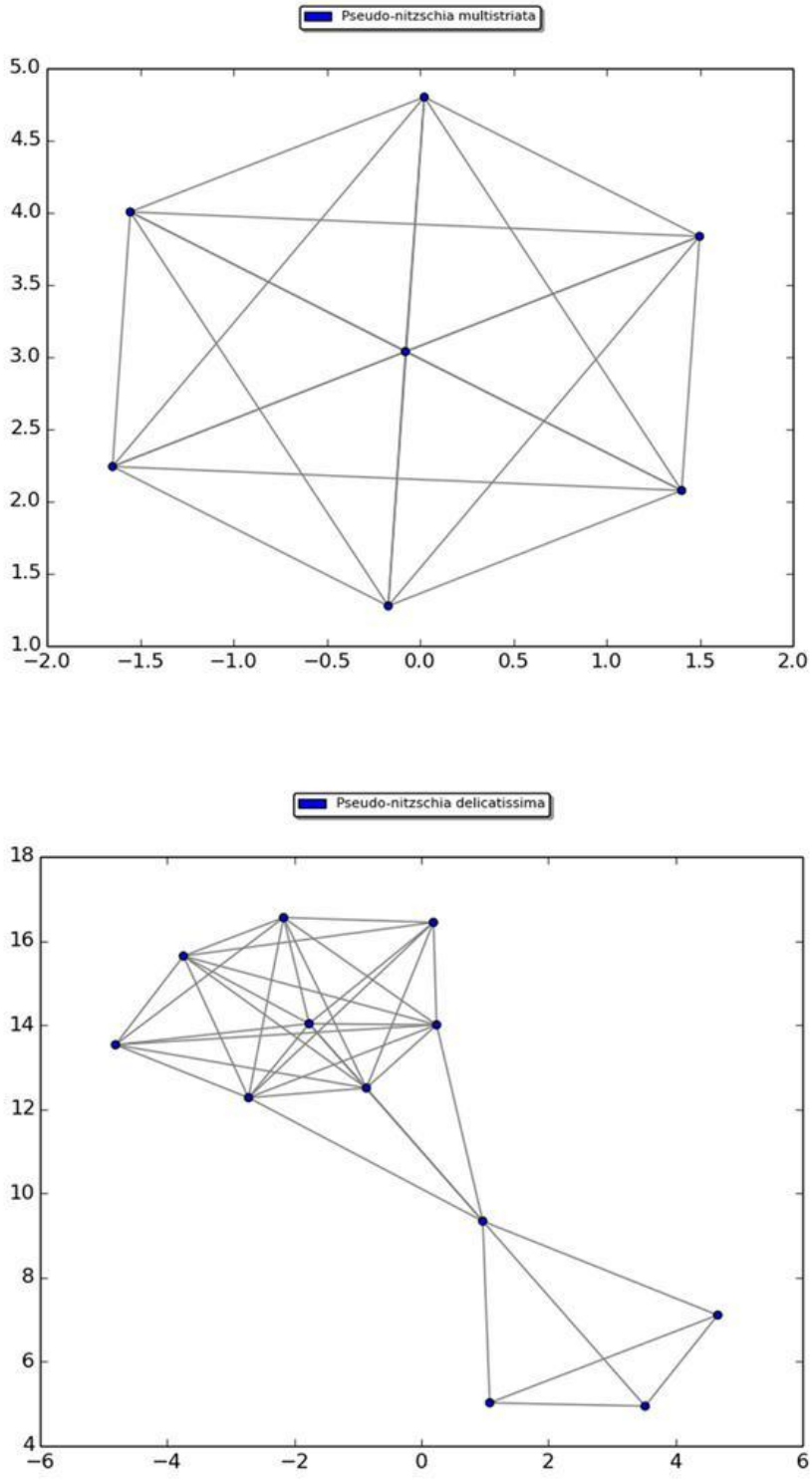

Figure 4: Use of Declic analyses to curate the database: case of taxonomically homogeneous clique (a) and connex component (b). (a) *Pseudonitzschia multistriata* (b) *Pseudo-nizschia delicatissima*. No changes were done in these cases.