**Supplementary figures and tables captions:**

**Supplementary figure 1**. Length distribution of the pDNCs in Odonata and *Conus* species.

**Supplementary figure 2**. RAxML phylogenetic trees (COI) of *Crocothemis erythraea* and *C*. *servilia* (**A**), *Pseudagrion massaicum* and *P. tanganyicum* (**B**) and *Trithemis stictica* (**C**). COI sequences included in the main Odonata\_COI data set highlighted with violet. Sequences, precluding compilation of a sDNC-based diagnosis valid for the expanded data set shown with red arrow.

**Supplementary table 1**. Metrics of the DNC based diagnoses for Odonata species (ND1, 316 positions).

**Supplementary table 2**. Metrics of the DNC based diagnoses for Odonata species (COI, 541 positions).

**Supplementary table 3**. Metrics of the DNC based diagnoses for Odonata families (data sets COI and ND1).

**Supplementary table4**. Metrics of the DNC based diagnoses for *Conus* species (COI, 658 positions).

**Supplementary table 5**. Metrics of the DNC based diagnoses for *Conus* subgenera (CO1, 658 positions).

**Supplementary table 6**. Metrics of the DNC based diagnoses for Conidae genera (CO1, 658 positions).

**Supplementary table 7**. Taxon-shared VS informative nucleotide positions in three codon positions.

**Supplementary table 8**. Validity of the pDNCs recovered for Odonata and Conidae species from main COI data sets on the respective expanded data sets.

**Supplementary table 9**. Inter-run variation of the sDNCs for taxa in analysed data sets.

**The folder ‘Scripts’ contains Python scripts used in the present paper:**

**MolD\_Clade\_weighted\_resampling\_pDNC\_PVS\_dLengths.py** – performs taxonomic resampling, picking up an increasing number of species from a dataset (all sequences from each picked species), provides pDNC based diagnosis for each sampled number of species in 10 iterations, and calculates proportion of the EDV-pDNCs in the output.

**MolD\_Clade\_weighted\_resampling\_pDNC\_PVS\_dLengths\_down.py** - performs taxonomic resampling, picking up an increasing number of species from a dataset (only one sequence from each picked species), provides pDNC based diagnosis for each sampled number of species in 10 iterations, and calculates proportion of the EDV-pDNCs in the output.

**MolD\_Clade\_weighted\_resampling\_sDNC.py** - performs taxonomic resampling, picking up an increasing number of species from a dataset (all sequences from each picked species), and provides sDNC based diagnosis for each sampled number of species in 10 iterations and, calculates proportion of the EDV-sDNCs in the output.

**MolD\_pDNC\_x5\_command\_line.py**, **MolD\_pDNC\_x5\_input\_in\_script.py**, **MolD\_sDNC\_x5\_command\_line.py**, **MolD\_sDNC\_x5\_input\_in\_script.py** – ready-to-use MolD versions. The input (location and name of the file with analysis parameters – default MolD\_parameters.txt) either specified in the body of the file (input\_in\_script), or provided via command line. The ‘input in script’ files are designed to be used with Spyder environment for Python, the command line files work by typing in command line input:  
python <PATH-TO>MolD\_pDNC\_x5\_command\_line.py –i <PATH-TO>MolD\_parameters.txt

**MolD\_parameters.txt** – a file, providing all parameters of the user MolD analysis, and brief instructions on their use.

**Read-Plot\_RESAMPLING\_lenghts\_Pvalid\_double2.py** – reads resampling output (can be found in the folder with output files) and plots the results.

**The folder ‘Input\_alignments’ contains input data used in the present paper:**

**Conidae\_COI\_spp.txt** – main Conidae dataset with second column reflecting allocation to species.

**Conidae\_COI\_genera.txt** - main Conidae dataset with second column reflecting allocation to genera

**Conidae\_COI\_5sg.txt** - main Conidae dataset with second column reflecting allocation to genera, plus (for four selected subgenera of *Conus*) – to subgenera.

**Furnariidae\_sf\_ND2.txt** - main Furnariidae dataset with second column reflecting allocation to subfamilies.

**Furnariidae\_genera\_ND2.txt** - main Furnariidae dataset with second column reflecting allocation to genera.

**Furnariidae\_genera\_ND2\_NCBI.txt** - extended Furnariidae dataset with second column reflecting allocation to genera.

**Furnariidae\_species\_ND2-NCBI.txt** - extended Furnariidae dataset with second column reflecting allocation to species.

**Odonata\_spp\_ND1.txt** – main Odonata ND1 dataset with second column reflecting allocation to species.

**Odonata\_spp\_COI.txt** – main Odonata COI dataset with second column reflecting allocation to species.

**Odonata\_spp\_COI\_stictica.txt** – main Odonata COI dataset with second column reflecting allocation to species; the divergent sequence of *Trithemis stictica* INS508317 added.

**Odonata\_spp\_4kCOI.txt** – extended Odonata COI dataset with second column reflecting allocation to species.

**Odonata\_Fam\_ND1.txt** - main Odonata ND1 dataset with second column reflecting allocation to families.

**Odonata\_Fam\_COI.txt** - main Odonata COI dataset with second column reflecting allocation to families.

**The folder ‘Output\_diagnoses’ contains pDNC and sDNC based diagnoses:**

Each file name contains information on what dataset was analysed and what type of diagnosis it provides.

The pDNC based diagnoses include

1. Information of the number of species and number of sequences in the analysed taxon;
2. Number of independent pDNCs sampled, and the pDNCs in the form: [position1: ‘nucleotide1’, position2: ‘nucleotide2’, ...]
3. Number of key positions
4. Number of pDNCs comprising 1, 2, 3, 4, 5 positions sampled (Lcount).
5. Information on the codon position usage in pDNCs.

The sDNC based diagnoses include:

1. Information of the number of species and number of sequences in the analysed taxon;
2. The sDNC scores obtained based on their validity in 100 artificial datasets;
3. Resulting sDNC in the form : [position1: ‘nucleotide1’, position2: ‘nucleotide2’, ...]

**The folder ‘Output\_resampling’ contains resampling results**

The txt files output by the scripts MolD\_Clade\_weighted\_resampling\_pDNC\_PVS\_dLengths.py (all sequences of picked species included in the partial dataset) and MolD\_Clade\_weighted\_resampling\_pDNC\_PVS\_dLengths\_down.py (only one sequence of each picked species included in the partial dataset). The filename reflects dataset analysed and focus taxon.

These files are read and plotted by pairs by the Read-Plot\_RESAMPLING\_lenghts\_Pvalid\_double2.py script.