GB→TNT

by P. Goloboff & S. Catalano 2011

What for ?

GB→TNT (GenBank to TNT) is a Windows program for easily creating data sets for TNT (Goloboff et al., 2003, 2008) from GenBank files. It is a merge of the C scripts (gb2tnt, fas2fas) used by Goloboff et al. (2009) in their analysis of eukaryotes, with a graphic interface and several functionalities added. The TNT matrices contain (by default) all the taxonomic information for each terminal taxon, so that you can easily diagnose the results in TNT (automatically labeling tree branches with the taxonomic groups they represent, or coloring different taxonomic groups).

GB→TNT is designed to extract defined genomic region(s) from a bulk of sequences included in a GenBank file (alternatively Fasta and TNT files are also allowed). Several filters (genome, length of the sequence, taxonomy, etc.) can be defined in order to generate the desired dataset. Each genomic region to be parsed is included in a different block of a project that will subsequently constitute a different block in the final TNT matrix.

The first step in the pipeline (Figure 1) is the creation of Fasta files from the GenBank files, retaining only one sequence (longest) when multiple copies for the same species exist. After that an alignment program is called (Mafft, Muscle or any alternative defined by the user). Finally all the aligned files are merged into a single TNT matrix. All these steps are automatically effected in the proper sequence by **GB→TNT**.

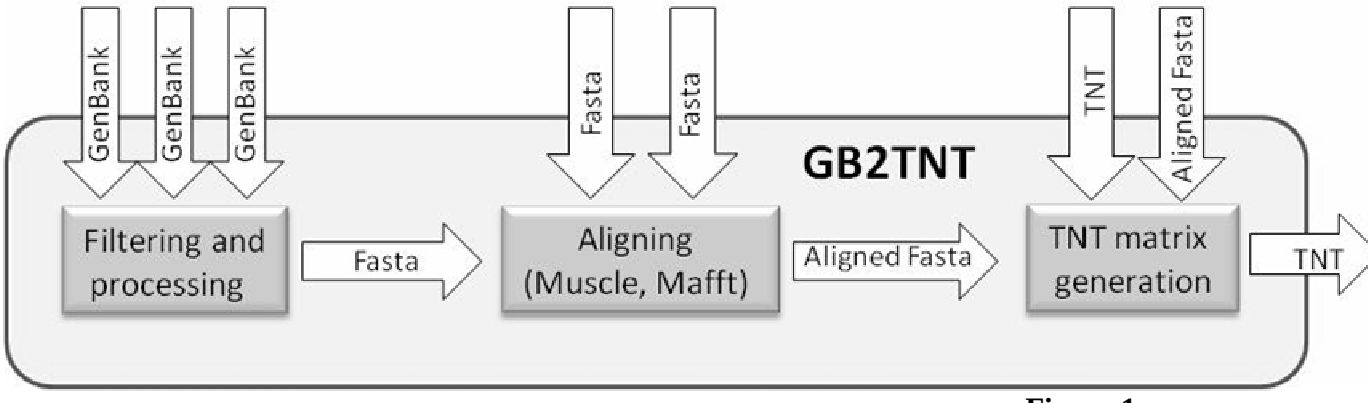


Figure 1

What do you need ?

The input for the program is sequence data in GenBank format files (Fasta and TNT files can also be processed). You also need to install Mafft and/or Muscle, or other alignment program of your preference. You can get Mafft from http://mafft.cbrc.jp/alignment/software/ (you need to install the Windows version

which runs with a batch file, without installation of Cygwin) and Muscle from http://www.drive5.com/muscle/.

How ?

You define a "project" (with the menu option **File/New Project**) or modify an existing one (**File/Edit Project**). A project is composed of several blocks (=genomic regions). A **GB→TNT** project consist of the definitions of all blocks as well as the parsing settings. Each block corresponds to a different genomic region to be parsed. Prealigned matrices in aligned Fasta format and TNT matrices can also be given as input. In this case each matrix corresponds to a project block. The "Edit project" window present two boxes (Figure 2). In the top one the user defines the location of the files to be parsed. In the bottom one the user defines the name(s) of the genomic region to be extracted. The settings of each block are defined by clicking on the "Options" button below the boxes.

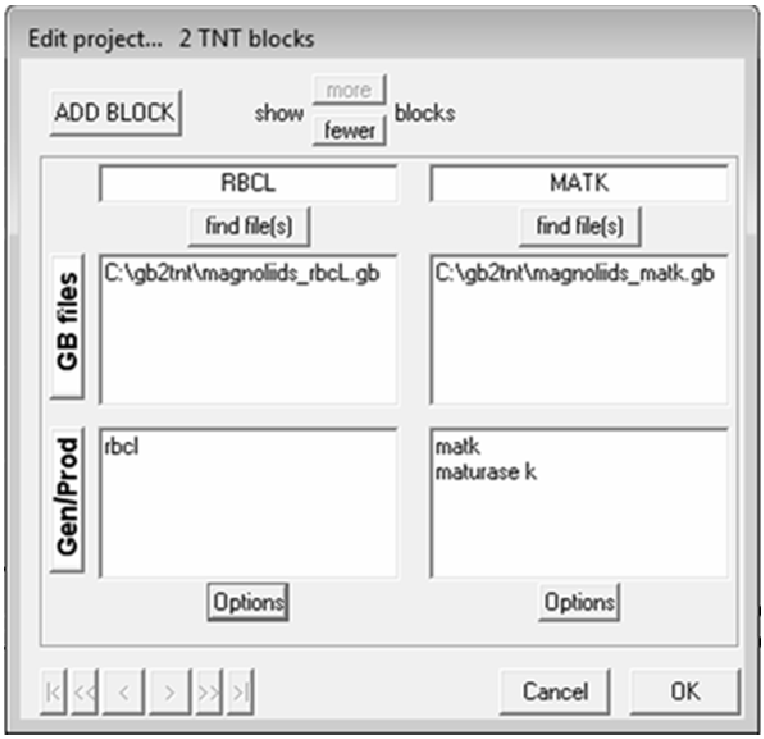


Figure 2

To build the matrix, **GB→TNT** parses the GenBank input files, creating temporary files for each block of TNT data. For a given block, you can use several input GenBank files. Each GenBank file is first extracted into a file named *project_name_bkN_fileN.tmp*. Then, **GB→TNT** combines these files into a single file for alignment, *project_name_bkN.fas*, retaining only one sequence (longest) when multiple copies exist for the same species. The alignment program is then called to align each of the blocks, outputting for block number N a file called *project_name_bkN.aln*. The aligned files for

each block are then merged into a single TNT matrix. At this point, the program checks for similar names (e.g. *Drosophila_melanogaster* and *Drosophila_melanogaxter*) so that typing errors can be identified, and creates a unique TNT matrix that is ready to run. **GB→TNT** writes the matrix for TNT in a file called project_name.tnt.

Block Options Window

The settings for each block are defined by clicking the OPTIONS button that is below the bottom (GEN/PROD) box (Figure 3).

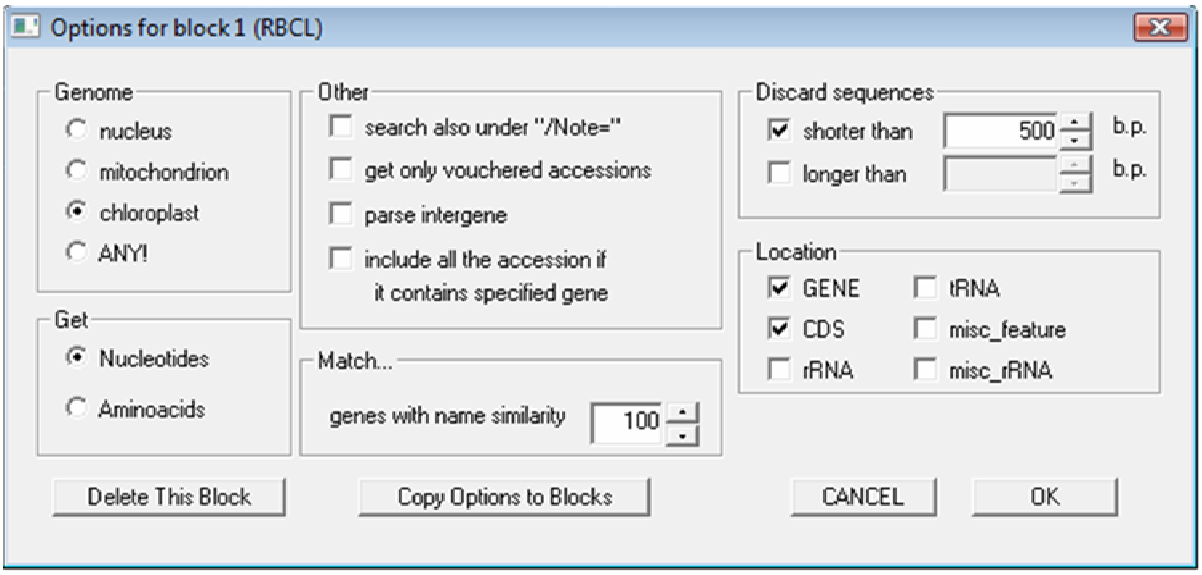


Figure 3

"Genome " box

GB→TNT will process only those sequences that match the defined genome. If the "Any" option is chosen the accessions with any genome or no genome definition will be extracted.

"Location" box

The user defines in this box where in the GB file the program will look for the names defined in the Gene/Product windows. The different qualifiers present in GB format can be chosen (gene, CDS, tRNA, rRNA, misc_feature, misc_rRNA). Within these qualifiers **GB→TNT** will look for the names given in the GEN/PROD box in the line that starts with "/gene=" and , if it corresponds, in the lines that starts with "/product=". Optionally **GB→TNT** can also look for the name in the line that starts with "/note=". More than one qualifier can be chosen at a time. If gene option is chosen, the part of the sequence

to be kept will be that defined in the "gene" headerline (e.g.: "gene <1..2521"). If the option CDS is chosen the part of sequence to be kept will be the one defined in the "CDS" headerline (e.g.: "CDS 1..1225,1501..2521"). This possibility allows the used to keep only the coding sequence of a gene. Alternatively the user can choose to extract the coding sequence as aminoacids instead of nucleotides. This is defined in the "get" box.

If more than one qualifier is chosen at the same time the program will keep the part of the sequence that is found first in the GB file (given that there is a match with the name defined).

"Other" box

"*Get only vouchered accesions* ": keeps only those sequences with a voucher of reference defined in "Source" as "/specimen_voucher"
"Include all the accesion if it contains the specified gene": Keeps not only the chosen gene but the complete accession that presents the genomic region defined in the GEN/PROD box.

"*Parse intergen* ": In this case two different gene/product name should be defined and the region between both will be kept. The sequences of the genes determined as limits can be keep optionally. The order of the genes should be coincident with that present in the GB file.

Alternatively, the genes that will function as limits can be defined in the gene/product window. The syntax is:
Name1_gene1
Name2_gene1
<> (less and greater than symbols)
Name1_gene2

The option to include the genes that act as limits in the region to be kept is available only in the options menu.

"Discard sequences" box:

Quite obvious

"Match" box

Useful for instance for those cases where an accession can be rejected given a misspelled name. Use it with caution given that sometimes two different genes differs only in a single character (e.g. rps12 vs. rps15)

“Copy options to blocks”:

All options (except those of the intergene parsing) are copied to the selected blocks.

BUILD MATRIX

Once all the options are set, the next step is the processing of the GenBank files. There are three steps in this process The first step is to extract the sequences from GenBank files. To do so **GB→TNT** considers the settings defined for each block At the end of this step **GB→TNT** creates a single Fasta file for each block. The second step is the alignment of the Fasta files. The program does not produce the alignments by itself but calls an alignment program. You need to install Mafft and/or Muscle, or other alignment program of your preference. It is very common that once the alignments are created the user may want to visually inspect the alignments and, if there is some misaligned sequences, exclude them to generate subsequently a new alignment. This is done in **GB→TNT** by calling BioEdit (**Project/Check with BioEdit**). If the user modifies some alignments the program will ask whether to mark this file for realignment or to keep it as it stands.

The final step is the generation of a TNT matrix. Each block in TNT file will correspond to one block of the project. Besides the data itself, the TNT file includes different settings and data: memory settings, names of the blocks, etc.

One of the main advantages of using **GB→TNT** to generate TNT matrices is the possibility to easily add all the GenBank taxonomy in the name of the species. TNT has a series of tools to make use of this information: it is possible for instance to evaluate the fit of the groups found in a tree search in the taxonomy, to label and colour branches, etc. It is possible to define the number of categories to be kept in the name. It is also possible to filter the sequences by the taxonomic identity for instance keeping only the sequences belonging to Primates or excluding the sequences belonging to Carnivora.

Details ?

GB→TNT keeps track of the modification dates and times of all the files it uses in the process, only re-aligning or re-extracting those files that need to be processed. This behavior can be overridden, as you can also specify one by one the files that are to be extracted/processed. If in doubt, choose "all" (under the menu option **Project/ Build**), for both extraction and alignment.

By default, **GB→TNT** will call the alignment program with their default options (i.e. just specification of input/output files). If you want, you can specify your own parameters for alignment (in this case, you will probably want to choose yourself the blocks that are to be aligned or re-aligned), by typing those options in the corresponding box; note that the parameters to be specified do not include the intput/output specification, which is always added by **GB→TNT** to the list of parameters. However, if you are using an alignment program of your choice for the alignment, you must specify the input and output files (using the convention names above for naming input/output files for each block). **GB→TNT** keeps track for you of where each of the sequences came from. The TNT matrix contains, in each block, a list of the taxa included for that block/gene. The name of the species is identical in each block (in case **GB→TNT** established some equivalenecies, based on highly similar names, the "valid" name –i.e. the one first encountered– is used throughout). The GenBank accession number is (normally) different in each block; for establishing taxon identity, TNT only uses the name of the species, not the accession number or the taxonomic information. This is so because **GB→TNT** adds to the name of the species a quadruple underscore, and TNT is designed to use only the string preceding

that quadruple underscore for name-matching. In this way, you can reconstruct the steps taken by **GB→TNT** to create the matrix, or manually change some doubtful accession (e.g. if it is sequenced by your mortal enemy).
Note that TNT stores in memory the name of the first occurrence of each taxon, so that if the first block contains:

*Drosophila_buzzatti*___AJK564_@Eukaryota_ …

And the second block contains:

*Drosophila_buzzatti*___XYZ321_@Eukaryota_ …

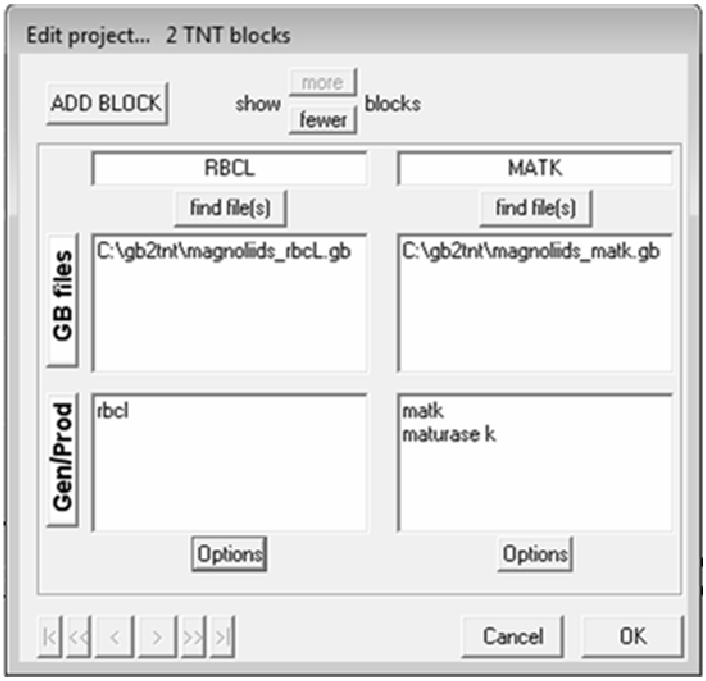
TNT will store the name for that terminal as the first string (i.e. accession AJK564). For this species, to know what accession the sequence in the second block came from, you need to look at the matrix, in the file project_name.tnt.

Example

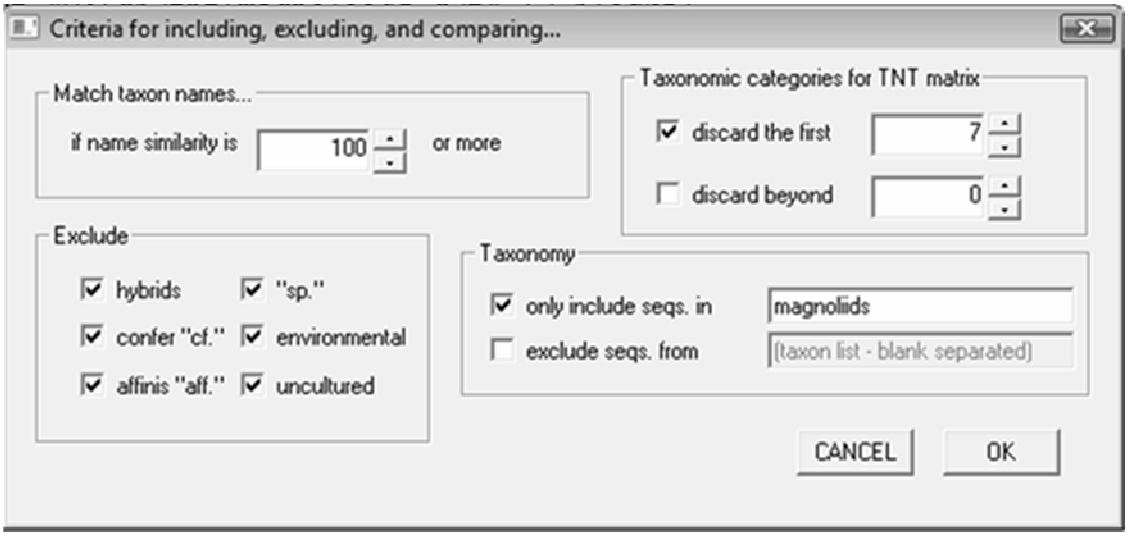
This sections shows how to create, with **GB→TNT**, a **TNT** matrix including rbcL and matK for all the species of magnoliids available in GenBank, with at least 500 bp sequenced. The GB files, *magnoliids_rblc.gb* and *magnoliids_matk.gb*, are included together with the **GB→TNT** package; these two files were generated and downloaded directly from the NCBI homepage.

We will start first by generating a new project (**File/New Project**, which opens the "**Edit project**" dialog). We will rename the first block as *rbc*L, and indicate to the program where the GB file to be parsed is (by clicking on the "**find file(s)**" button). After that we will define the name of the gene we want to retrieve (rbcl in this case) in the "**Gen/Prod**" box. The options for this block are set with the "**Options**" button (below the "**Gen/Prod**" box); since *rbc*L is a gene from chloroplast we will click the "**Chloroplast**" option in the "**Genome**" box; the default option of the "**Get**" is "**Nucleotide**", so no modification is needed there.

To generate the second block we press the "**Add block**" button of the "**Edit Project**" window. As in the case of the rbcl block we will rename the block, define where the GB file to be parsed is, and write the name of the gene in the "**Gen/Prod**" box. Here, we will include two ways of naming the same gene, *mat*k and maturase k. Since the second is a product name, we must also choose the **CDS** option of the "**Location**" box (this is there where the product names are defined in GB files).



After that, the "**Edit project**" dialog can be closed by pressing "**OK**" and the project can be saved (with **File/Save**). We will then define the taxonomy options (**Taxonomy/Inclusion Exclusion criteria)**. Here we will indicate to the program that we want to exclude the first seven categories from the taxonomy that will be added to each terminal in the **TNT** file (since those seven categories are common to all the sequences) and that we just want species belonging to magnoliids (just in case species for other groups were infiltrated in the GB file).



The alignment program to use can be defined from the menu "**Options/Align with...**"; in this case, we will select **Muscle**. Finally, we will proceed to build the matrix (**Project/Build Matrix**). For alignment ("**STEP 2**" box), we will set "-maxiters 1" (a faster **Muscle** option, since we do not need a lot of accuracy for this example). To generate the **TNT** matrix we will leave the "**STEP 3**" options as default (i.e. "**Create matrix for TNT**" and "**Launch TNT**" selected). After pressing "**OK**", **GB→TNT** parses the GB files, calls **Muscle** to perform the alignment, and generates the **TNT** matrix, without the need for human intervention. With the settings shown, the matrix consists of 1376 taxa and 4247 characters. All the work needed to assemble this large data set, from the downloading of the GB files to the last step in the creation of the matrix (including approximate alignments, with default program options), could be completed in a couple of hours.