

# MEGA-CC (Compute Core) and MEGA-Proto

Quick Start Tutorial

# MEGA-CC Input Files

- MEGA Analysis Options file
  - Specifies the calculation and desired settings.
  - Created using MEGA-Proto.
  - Has a *.mao* file extension.
- Data file (one of the following)
  - Multiple sequence alignment in MEGA or Fasta format.
  - Distance matrix in MEGA format.
  - Unaligned sequences in Fasta format (for alignment only).
- Tree file (required for some analyses)
  - Newick file format.

# MEGA-CC Output Files

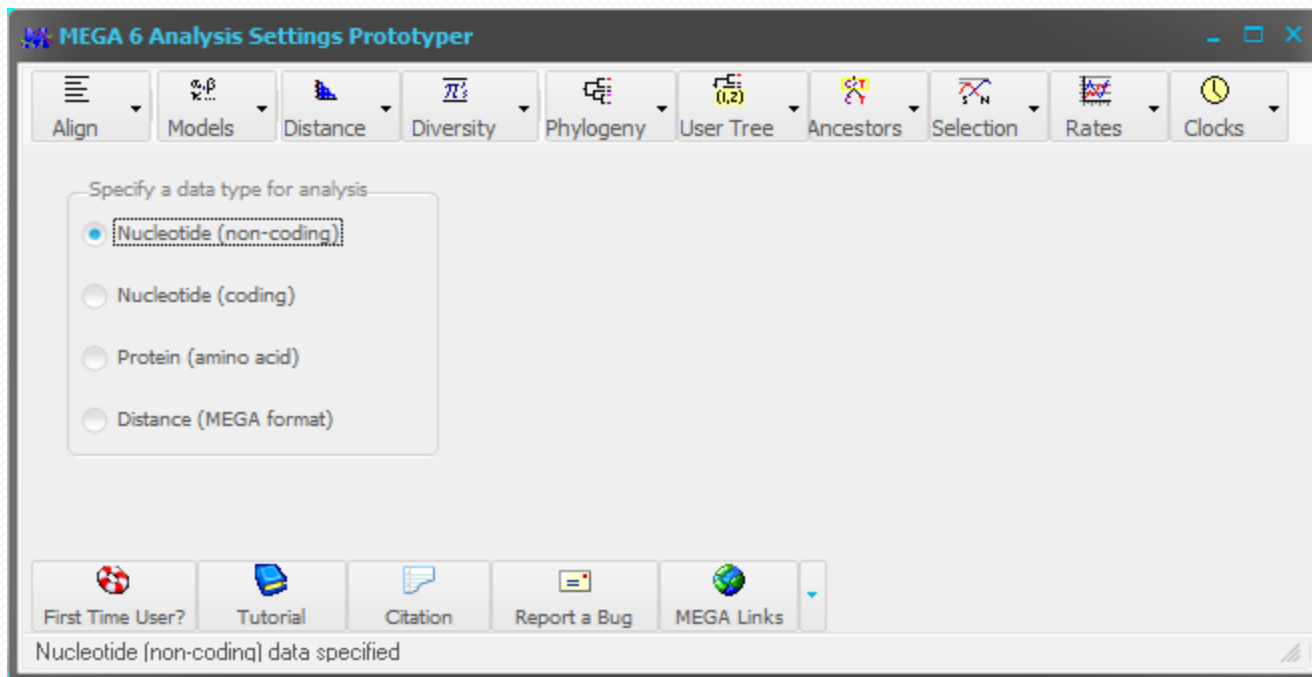
- In general, two output files are produced
  1. Calculation-specific results file (Newick file, distance matrix,...).
  2. A summary file with additional info (likelihood, SBL,...).
- Some analyses produce additional output (bootstrap consensus tree).
- Output directory/filename
  - Default is the same location as the input data file.
  - Specify an output directory and/or file name using `-o` option.
  - If no output filename is specified, MEGA-CC will assign a unique name.
- Errors/warnings
  - If MEGA-CC produces any errors or warnings, they will be logged in the the summary file.

# Running MEGA-CC

- Easiest to run using command-line or batch scripts:
  - `megacc -a settings.mao -d alignment.meg -o outFile`
- Can also be run using custom scripts (Perl, Python, ...):
  - `exec('megacc -a options.mao -d alignment.meg -o outFile');`
- Integrated *File Iterator* system can process multiple files without the need for using scripts (see Demo2 below)
- In addition, other applications can launch MEGA-CC:
  - `status = CreateProcess("path/to/megacc.exe...");`
- To see a list of available command options, call `megacc` from a command-line prompt with the `-h` flag.

# MEGA-Proto (analysis prototyper)

- Has the same look and feel as the GUI edition of MEGA.
- Produces MEGA Analysis Options files.
- Has no computational capabilities.

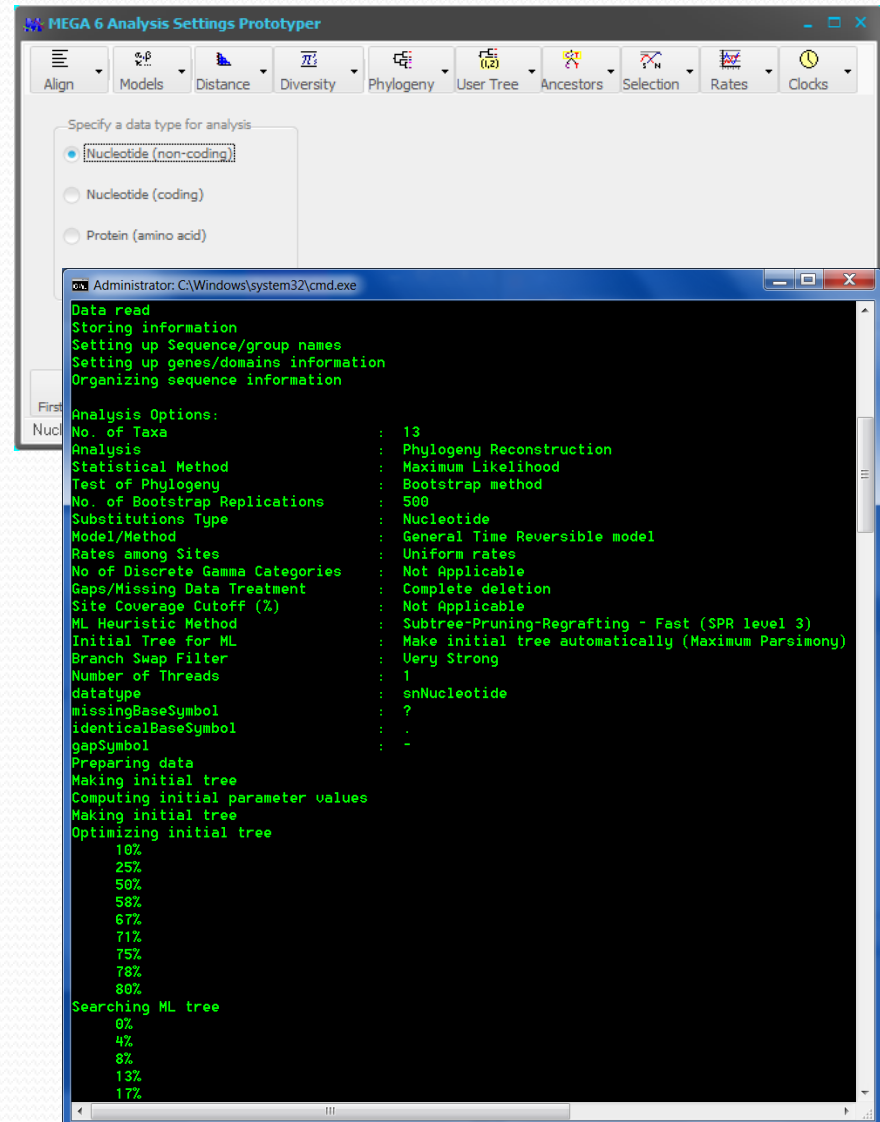


# Using MEGA-Proto

1. Select input data type.
  - Nucleotide (non-coding)
  - Nucleotide (coding)
  - Protein (amino-acid)
  - Distance matrix (MEGA format)
2. Select analysis from menu.
3. Adjust analysis settings.
4. Save the MEGA Analysis Options file.

# Demo1

- The following example demonstrates how to create a timetree using MEGA-Proto and MEGA-CC



The image shows two overlapping windows. The top window is the 'MEGA 6 Analysis Settings Prototyper' interface. It has a menu bar with options: Align, Models, Distance, Diversity, Phylogeny, User Tree, Ancestors, Selection, Rates, and Clocks. Below the menu bar, there is a section titled 'Specify a data type for analysis' with three radio buttons: 'Nucleotide (non-coding)' (selected), 'Nucleotide (coding)', and 'Protein (amino acid)'. The bottom window is a command prompt titled 'Administrator: C:\Windows\system32\cmd.exe'. It displays the following output:

```
Data read
Storing information
Setting up Sequence/group names
Setting up genes/domains information
Organizing sequence information

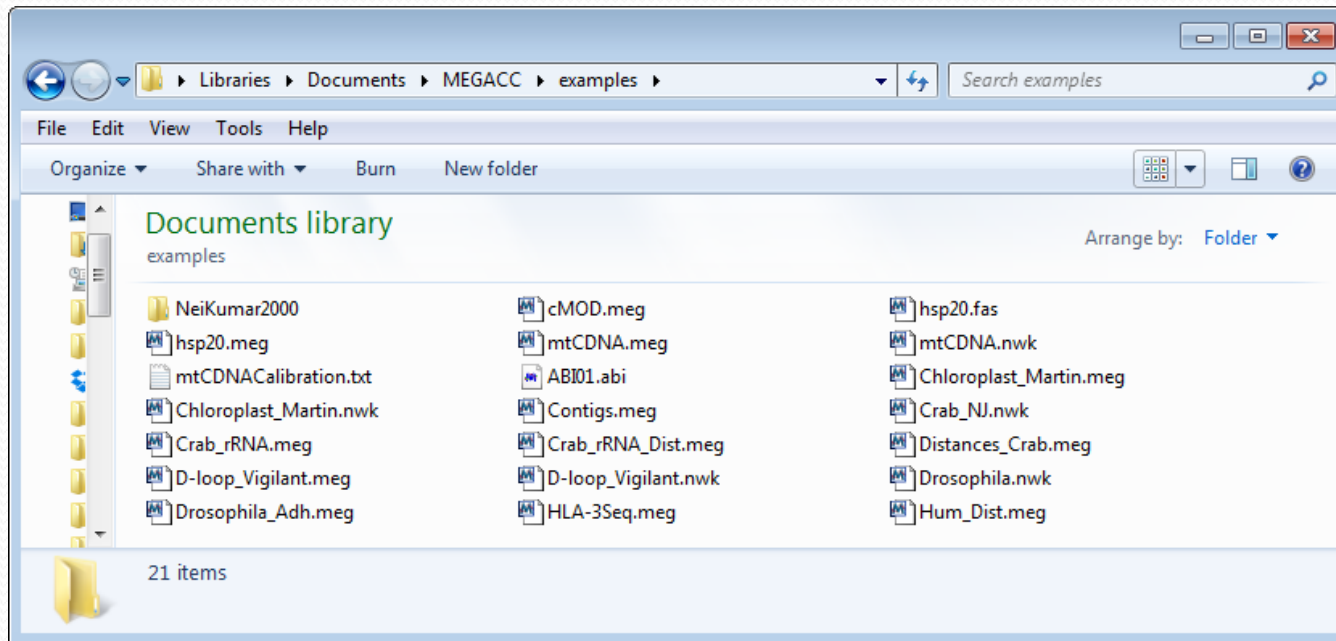
Analysis Options:
No. of Taxa           : 13
Analysis              : Phylogeny Reconstruction
Statistical Method    : Maximum Likelihood
Test of Phylogeny     : Bootstrap method
No. of Bootstrap Replications : 500
Substitutions Type    : Nucleotide
Model/Method          : General Time Reversible model
Rates among Sites     : Uniform rates
No of Discrete Gamma Categories : Not Applicable
Gaps/Missing Data Treatment : Complete deletion
Site Coverage Cutoff (%) : Not Applicable
ML Heuristic Method   : Subtree-Pruning-Regrafting - Fast (SPR level 3)
Initial Tree for ML  : Make initial tree automatically (Maximum Parsimony)
Branch Swap Filter    : Very Strong
Number of Threads     : 1
datatype              : snNucleotide
missingBaseSymbol     : ?
identicalBaseSymbol   : -
gapSymbol             : -

Preparing data
Making initial tree
Computing initial parameter values
Making initial tree
Optimizing initial tree
10%
25%
50%
58%
67%
71%
75%
78%
80%

Searching ML tree
0%
4%
8%
13%
17%
```

# Demo1 Data Files

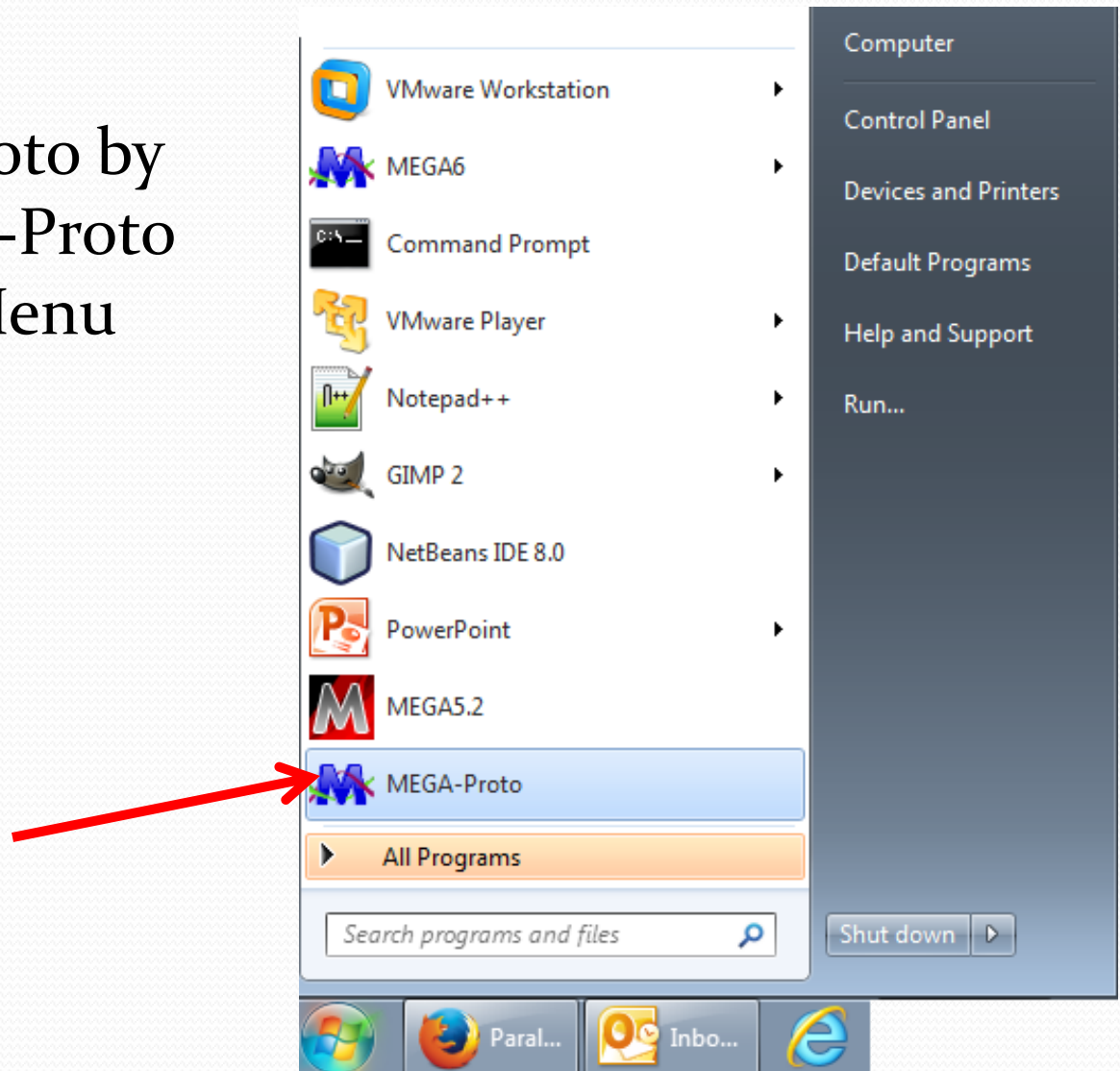
- For this demo, we will be using some of the example data files that were copied to your documents directory by the installer (*i.e.* Documents\MEGACC\examples).





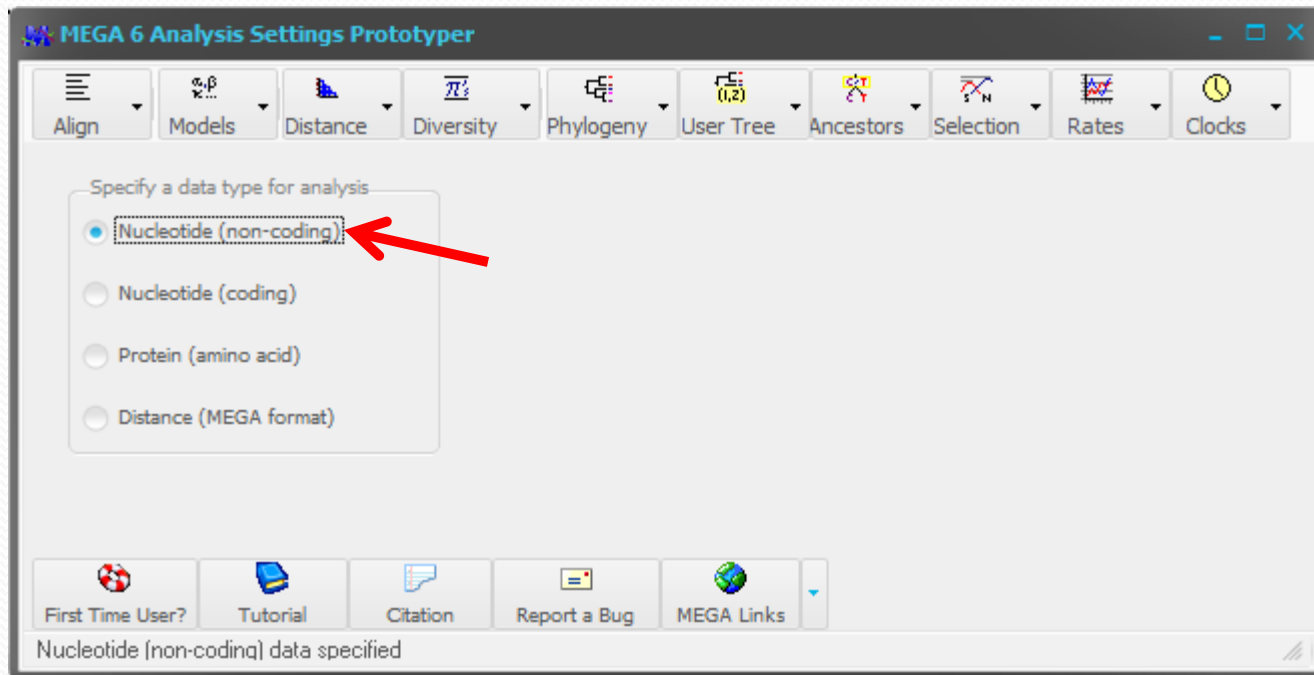
# Step 1

- Open MEGA-Proto by selecting MEGA-Proto from the Start Menu



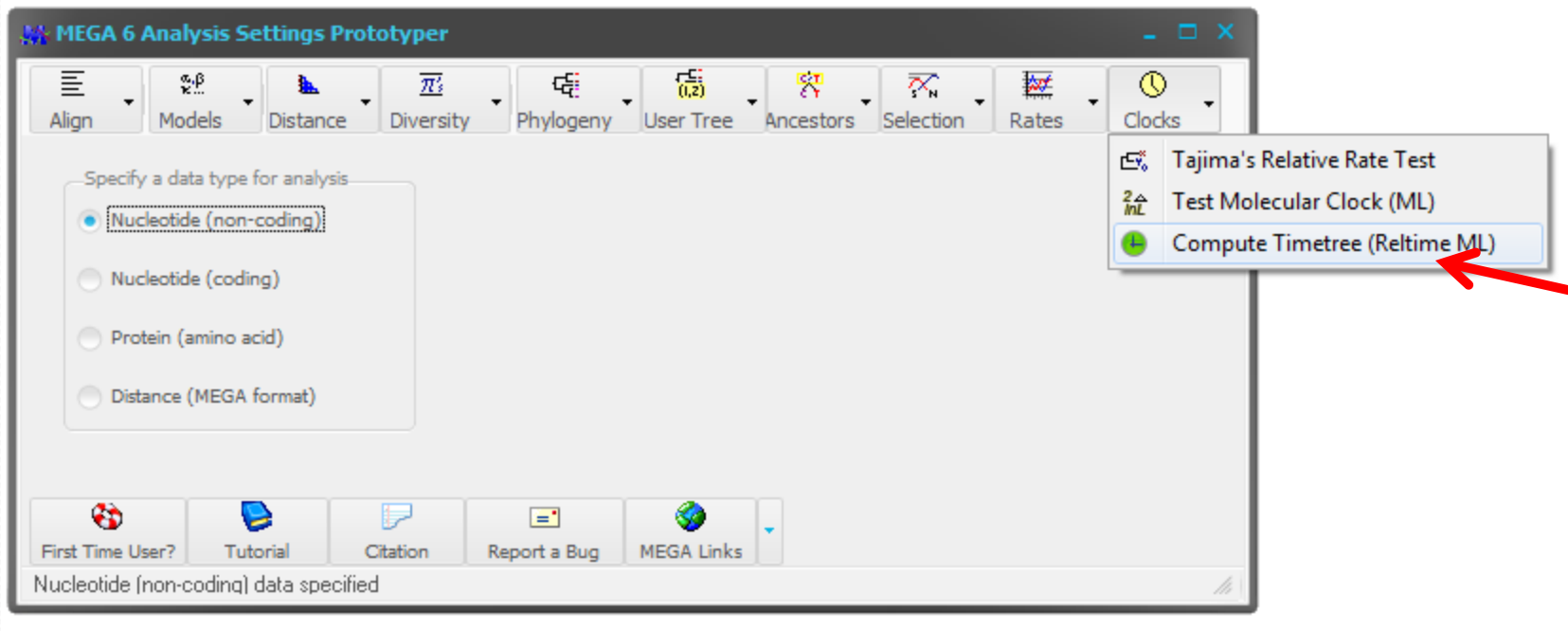
# Step 2

- Select the data type of the input data file to be analyzed. For this demo, we will accept the default setting - Nucleotide (non-coding).



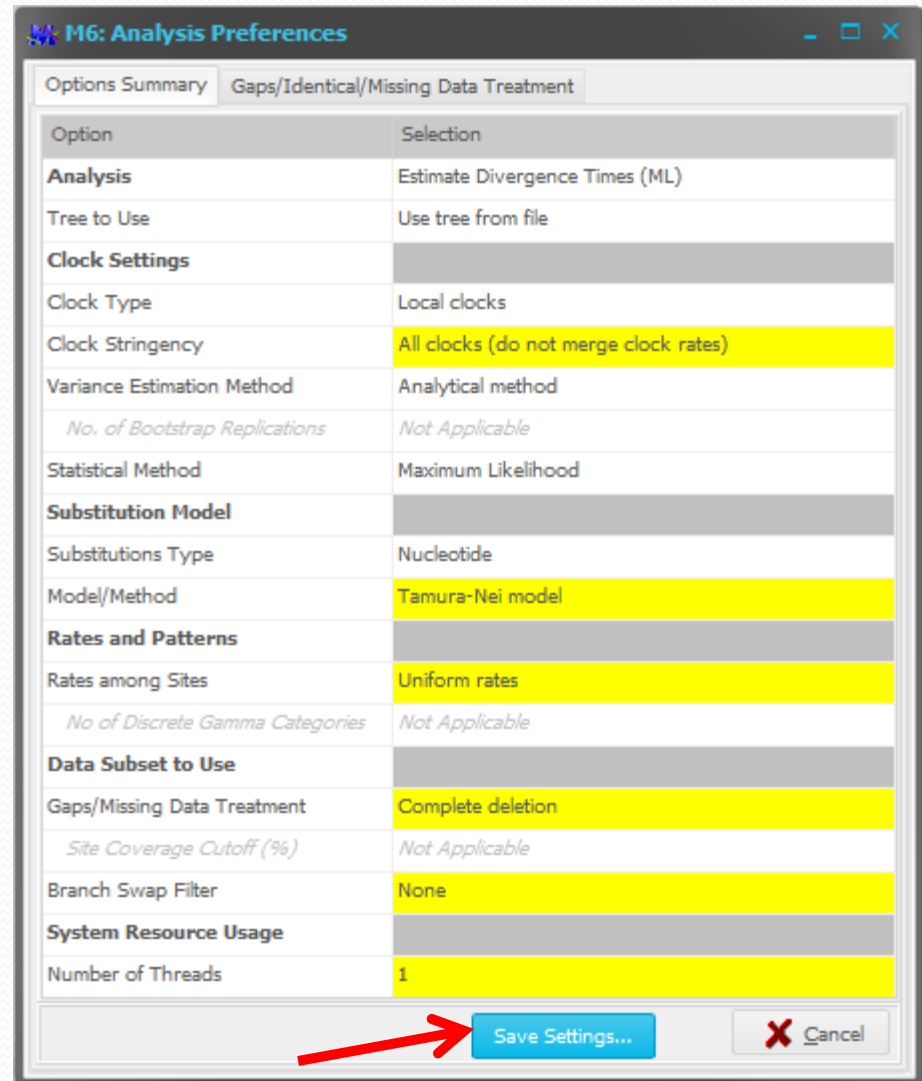
# Step 3

- Select *Compute Timetree (Reltime ML)* from the *Clocks* menu.



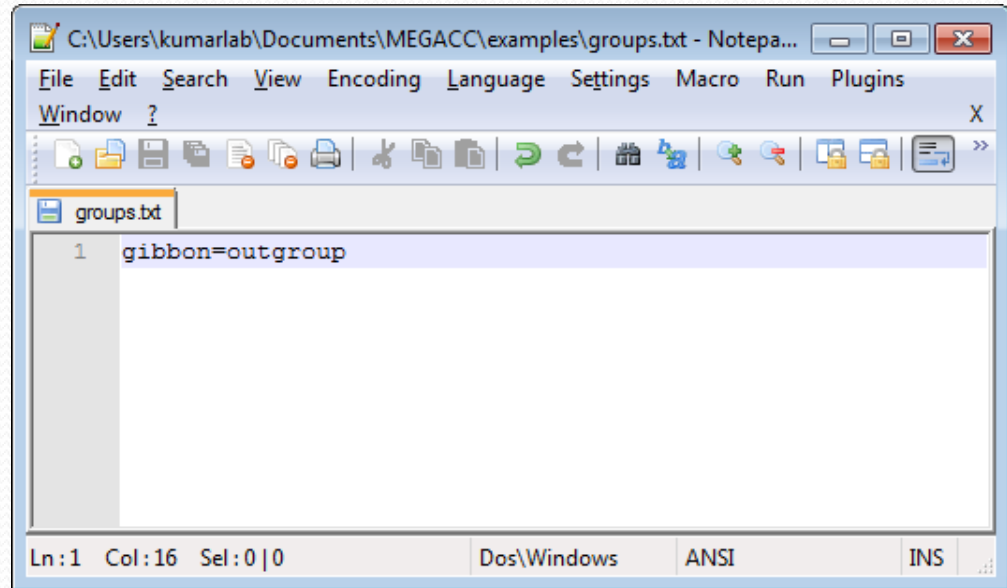
# Step 4

- Adjust the analysis preferences to match those shown.
- Click the *Save Settings...* button and save the analysis options file as *demoSettings.mao* in the MEGACC\examples directory.



# Step 5

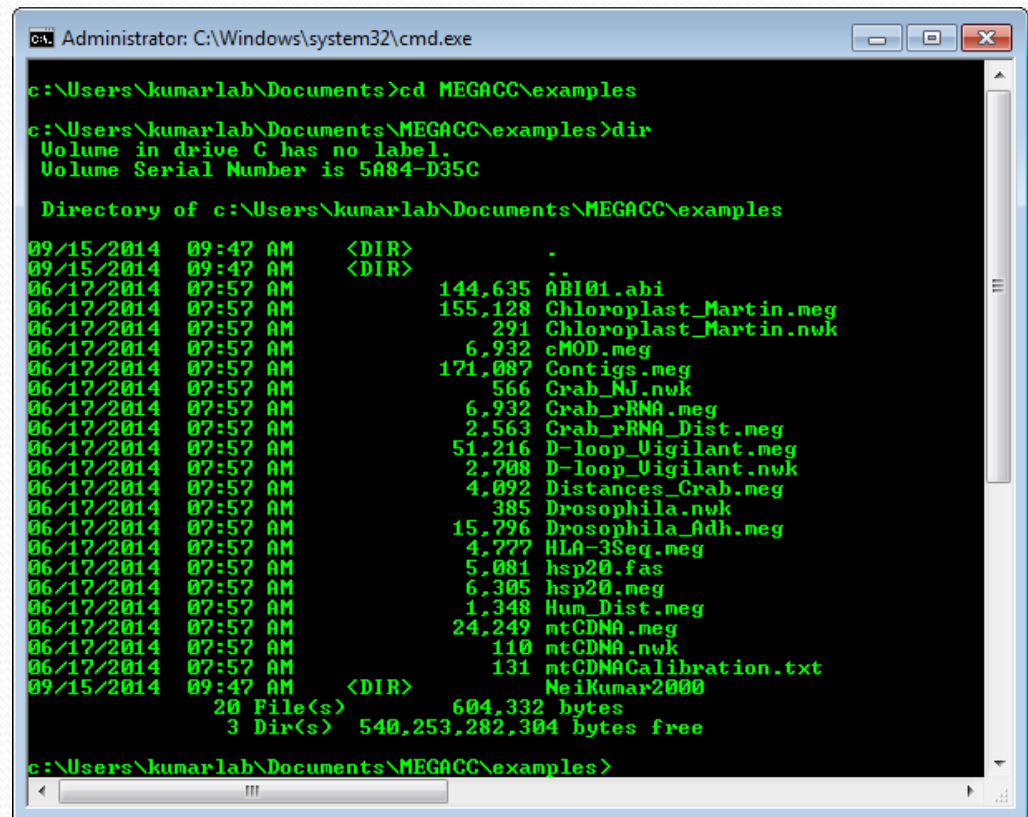
- The timetree analysis requires that we specify an outgroup. To do so, create a text file and add the line 'gibbon=outgroup'. Save this file as groups.txt in the MEGACC\examples directory.



A screenshot of a Notepad window titled 'C:\Users\kumarlab\Documents\MEGACC\examples\groups.txt - Notepa...'. The window displays a single line of text: '1 gibbon=outgroup'. The status bar at the bottom indicates 'Ln:1 Col:16 Sel:0|0', 'Dos\Windows', 'ANSI', and 'INS'.

# Step 6

- Open a command prompt.
- Navigate to the MEGACC\examples directory using the `cd` command



```
Administrator: C:\Windows\system32\cmd.exe
c:\Users\kumarlab\Documents>cd MEGACC\examples
c:\Users\kumarlab\Documents\MEGACC\examples>dir
Volume in drive C has no label.
Volume Serial Number is 5A84-D35C

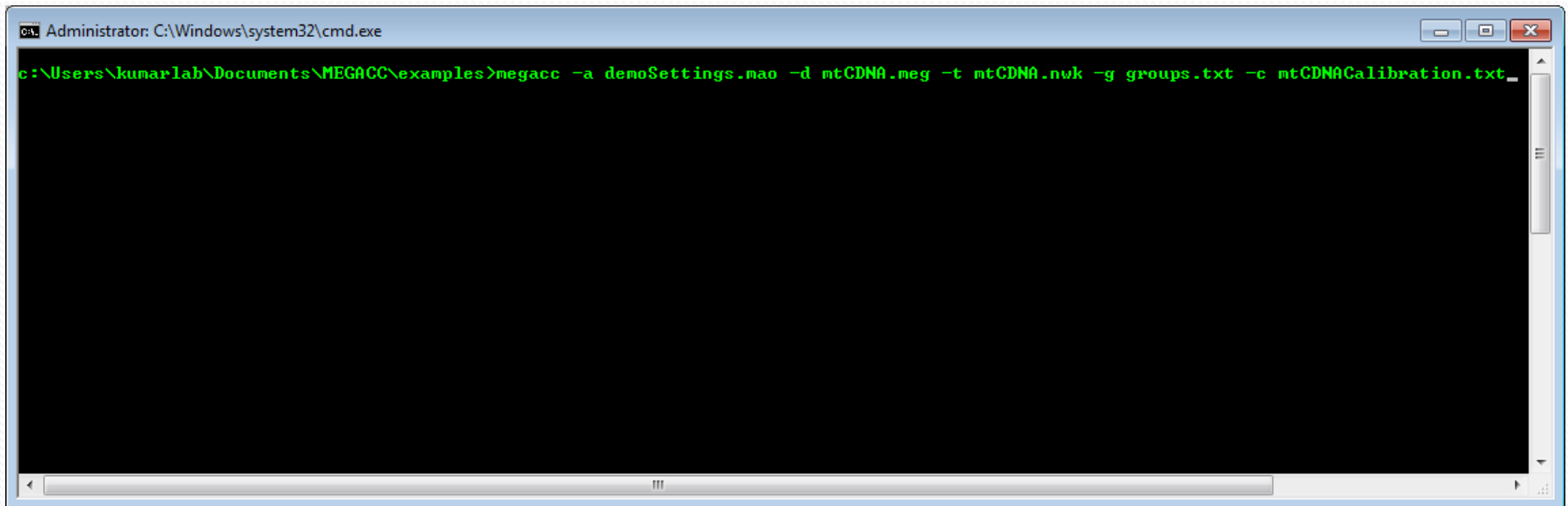
Directory of c:\Users\kumarlab\Documents\MEGACC\examples

09/15/2014  09:47 AM  <DIR>          .
09/15/2014  09:47 AM  <DIR>          ..
06/17/2014  07:57 AM             144,635  ABI01.abi
06/17/2014  07:57 AM             155,128  Chloroplast_Martin.meg
06/17/2014  07:57 AM              291  Chloroplast_Martin.nwk
06/17/2014  07:57 AM              6,932  cMOD.meg
06/17/2014  07:57 AM             171,087  Contigs.meg
06/17/2014  07:57 AM              566  Crab_MJ.nwk
06/17/2014  07:57 AM              6,932  Crab_rRNA.meg
06/17/2014  07:57 AM             2,563  Crab_rRNA_Dist.meg
06/17/2014  07:57 AM             51,216  D-loop_Vigilant.meg
06/17/2014  07:57 AM             2,708  D-loop_Vigilant.nwk
06/17/2014  07:57 AM             4,092  Distances_Crab.meg
06/17/2014  07:57 AM              385  Drosophila.nwk
06/17/2014  07:57 AM             15,796  Drosophila_Adh.meg
06/17/2014  07:57 AM             4,777  HLA-3Seq.meg
06/17/2014  07:57 AM             5,081  hsp20.fas
06/17/2014  07:57 AM             6,305  hsp20.meg
06/17/2014  07:57 AM             1,348  Hum_Dist.meg
06/17/2014  07:57 AM             24,249  mtCDNA.meg
06/17/2014  07:57 AM              110  mtCDNA.nwk
06/17/2014  07:57 AM              131  mtCDNACalibration.txt
09/15/2014  09:47 AM  <DIR>          NeiKumar2000
                20 File(s)          604,332 bytes
                3 Dir(s)  540,253,282,304 bytes free

c:\Users\kumarlab\Documents\MEGACC\examples>
```

# Step 7

- Execute the analysis by calling megacc from the command prompt as follows:
- `megacc -a demoSettings.mao -d mtCDNA.meg -t mtCDNA.nwk -g groups.txt -c mtCDNACalibration.txt`



A screenshot of a Windows command prompt window. The title bar reads "Administrator: C:\Windows\system32\cmd.exe". The command prompt shows the following command being executed: `c:\Users\kumar\lab\Documents\MEGACC\examples>megacc -a demoSettings.mao -d mtCDNA.meg -t mtCDNA.nwk -g groups.txt -c mtCDNACalibration.txt`. The command is displayed in green text on a black background. The window has standard Windows window controls (minimize, maximize, close) in the top right corner and a scrollbar on the right side.

# Step 8

- The analysis will be launched and progress updates will be displayed in the command prompt window.

```
Administrator: C:\Windows\system32\cmd.exe - megacc -a demoSettings.mao -d mtCDNA.meg -t mtCDN...
Branch Swap Filter      None
Number of Threads      1
datatype                snNucleotide
containsCodingNuc      False
MissingBaseSymbol      ?
IdenticalBaseSymbol    -
GapSymbol               -
Start time: 9/15/2014 10:21:08
Executing analysis:

    100% Analysis Complete

c:\Users\kumarlab\Documents\MEGACC\examples>megacc -a demoSettings.mao -d mtCDNA.meg
MEGA-CC.10 Molecular Evolutionary Genetics Analysis
Build#: 6140910
 0% Organizing sequence information
 0% 9/15/2014 10:21:16
Using the following analysis options:
No. of Taxa             7
No. of Groups          1
Analysis               Estimate Divergence Times (ML)
Tree to Use            Use tree from file
Clock Type             Local clocks
Clock Stringency       All clocks (do not merge clock rates)
Variance Estimation Method Analytical method
No. of Bootstrap Replications Not Applicable
Statistical Method     Maximum Likelihood
Substitutions Type     Nucleotide
Model/Method           Tamura-Nei model
Rates among Sites      Uniform rates
No of Discrete Gamma Categories Not Applicable
Gaps/Missing Data Treatment Complete deletion
Site Coverage Cutoff (%) Not Applicable
Branch Swap Filter     None
Number of Threads      1
datatype                snNucleotide
containsCodingNuc      False
MissingBaseSymbol      ?
IdenticalBaseSymbol    -
GapSymbol               -
Start time: 9/15/2014 10:21:16
Executing analysis:

    75% Optimizing user tree
```

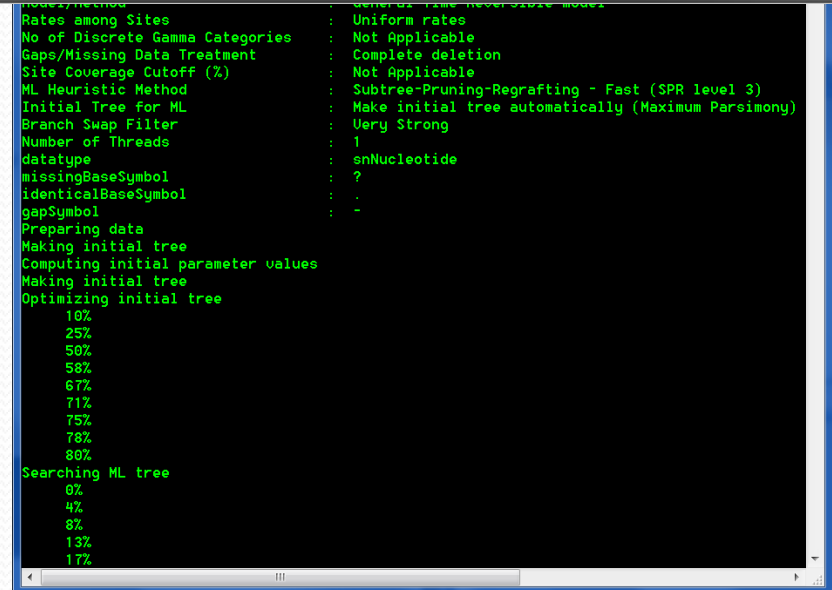
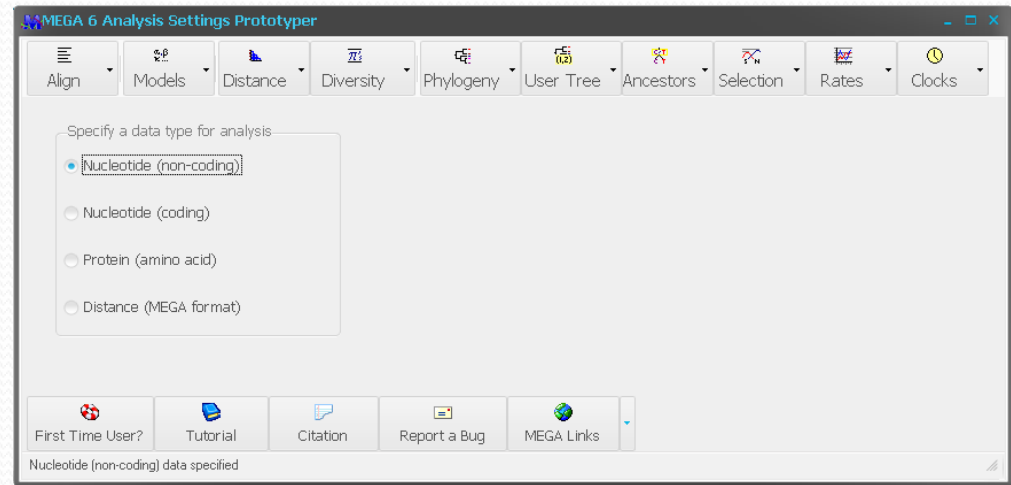


# Finally

- The analysis will produce several output files in the directory MEGACC\examples\M6CC\_Out
  - mtCDNA-xxxx\_exactTimes.nwk
    - This Newick file gives the timetree scaled according to the estimated divergence times.
  - mtCDNA-xxxx\_relTimes.nwk
    - This Newick file gives the timetree scaled according to the estimated relative divergence times.
  - mtCDNA-xxxx.txt
    - This text file gives a more detailed representation of the timetree, including relative times, exact times, evolutionary rates, and divergence time std errors.
  - mtCDNA-xxxx\_summary.txt
    - This file gives analysis information such as the log likelihood value of the Maximum Likelihood tree, ts/tv ratio, etc...

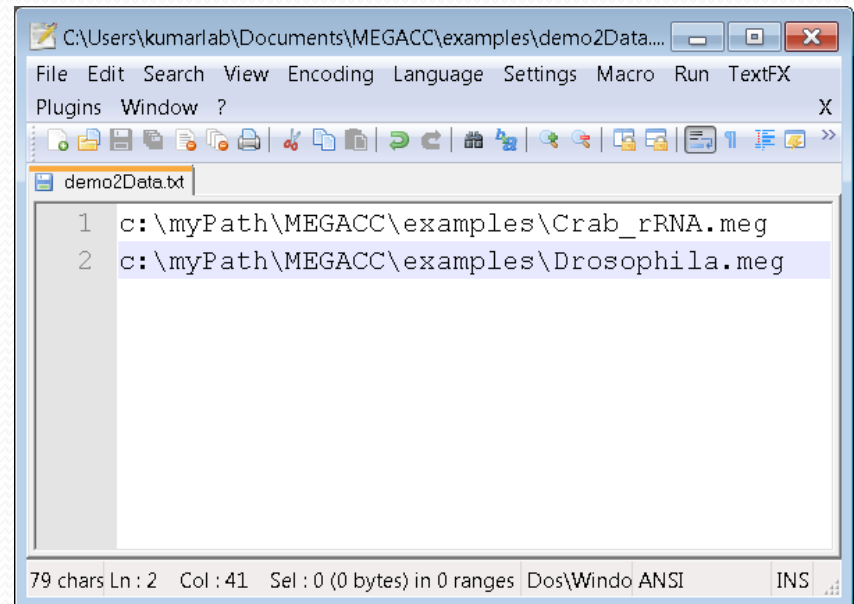
# Demo2

- The following example demonstrates how to use the File Iterator system in MEGA-CC to process multiple input data files using a single analysis options file.



# Step 1

- Create a text file named demo2Data.txt which we will use to specify multiple alignment files for ML phylogeny inference.
- In this file, add the full paths to the Crab\_rRNA.meg and Drosophila\_Adh.meg example files.



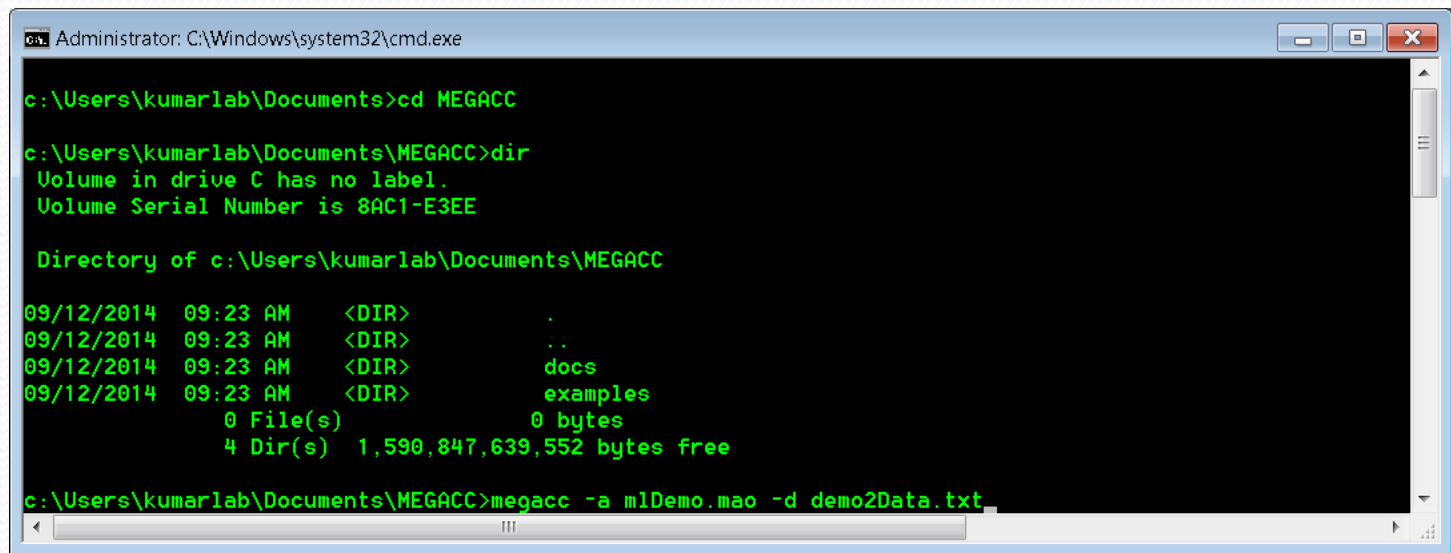
The screenshot shows a text editor window titled "demo2Data.txt" with the following content:

```
1 c:\myPath\MEGACC\examples\Crab_rRNA.meg
2 c:\myPath\MEGACC\examples\Drosophila.meg
```

The status bar at the bottom indicates "79 chars Ln : 2 Col : 41 Sel : 0 (0 bytes) in 0 ranges Dos\Windo ANSI INS".

# Step 2

- From a command-line prompt, call MEGA-CC as follows:
  - `megacc -a mlDemo.mao -d demo2Data.txt`



```
Administrator: C:\Windows\system32\cmd.exe

c:\Users\kumarlab\Documents>cd MEGACC

c:\Users\kumarlab\Documents\MEGACC>dir
Volume in drive C has no label.
Volume Serial Number is 8AC1-E3EE

Directory of c:\Users\kumarlab\Documents\MEGACC

09/12/2014  09:23 AM    <DIR>          .
09/12/2014  09:23 AM    <DIR>          ..
09/12/2014  09:23 AM    <DIR>          docs
09/12/2014  09:23 AM    <DIR>          examples
               0 File(s)              0 bytes
               4 Dir(s)  1,590,847,639,552 bytes free

c:\Users\kumarlab\Documents\MEGACC>megacc -a mlDemo.mao -d demo2Data.txt
```

# Step 3

- The analyses will be launched sequentially and progress updates will be displayed in the command prompt window.

```
Administrator: C:\Windows\system32\cmd.exe
c:\yourWorkingDirectory> M51CC.exe -a mlDemo.mao -d Examples\Crab_rRNA.meg -o demoResults
MEGA 5.1 Molecular Evolutionary Genetics Analysis
Build#: 5120301
Data file           : Examples\Crab_rRNA.meg
Reading header
Reading data
Data read
Storing information
Setting up Sequence/group names
Setting up genes/domains information
Organizing sequence information

Analysis Options:
No. of Taxa           : 13
Analysis              : Phylogeny Reconstruction
Statistical Method    : Maximum Likelihood
Test of Phylogeny     : Bootstrap method
No. of Bootstrap Replications : 500
Substitutions Type    : Nucleotide
Model/Method          : General Time Reversible model
Rates among Sites     : Uniform rates
No of Discrete Gamma Categories : Not Applicable
Gaps/Missing Data Treatment : Complete deletion
Site Coverage Cutoff (%) : Not Applicable
ML Heuristic Method   : Subtree-Pruning-Regrafting - Fast (SPR level 3)
Initial Tree for ML   : Make initial tree automatically (Maximum Parsimony)
Branch Swap Filter    : Very Strong
Number of Threads     : 1
datatype              : snNucleotide
missingBaseSymbol     : ?
identicalBaseSymbol   : -
gapSymbol              : -
Preparing data
Making initial tree
Computing initial parameter values
Making initial tree
Optimizing initial tree
10%
25%
50%
58%
67%
71%
75%
```

# Finally

- The analysis will produce output files for each input data file
- In this example, the same analysis options were used for each alignment file
- Enjoy!