

PROTOCOL

DNA BARCODING OF INSECTS FOR NON EXPERTS

LABORATORY OF ENTOMOLOGY

TÂNIA NOBRE
2021





PREFACE

This protocol is developed to aid students in their first steps in DNA barcoding. It assumes that lab safety rules are known, as well as procedures to work sterile.

It is intended to provide basic knowledge to allow a non-expert (with some skills) to extract, amplify and analyse the sequence of a fragment of the COI in an insect.

DNA BARCODING

DNA barcoding refers to the technique of sequencing a short fragment of the mitochondrial cytochrome c oxidase subunit I (COI) gene, the "DNA barcode," from a taxonomically unknown specimen and performing comparisons with a reference library of barcodes of known species origin to attempt establishing a species-level identification.

WHY?

Often, morphological characterization is not enough to identify an insect at species level, and certainly it is not a task for a non-expert!

REFERENCE BARCODE LIBRARY

The Barcode of Life Datasystem (BOLD), carefully curated by experts all over the world.

CONTENT

EXTRACTION

Steps

Protocol with extraction kit

CTAB extraction Protocol

AMPLIFICATION

Components

Chain reaction

Protocol for COI

Visualization

Protocol for gel electrophoresis

PCR product purification

SEQUENCING AND ANALYSES

Sequencing

Visualisation of the chromatogram

Homology search

BOLD - Barcode of Life

Genbank



EXTRACTION

STEPS

1. Grinding/blending separates the cells.

Each cell is surrounded by the cell membrane and the DNA is found inside a second membrane, the nuclear membrane, within each cell. To free the DNA, we have to break open these two membranes.

2. A detergent is added

The lipids and proteins in the cell membrane and nuclear membrane get "captured" by the detergent, breaking open the cell membrane and the nuclear membrane. This frees the DNA.

3. Proteases are added

These enzymes, called proteases, cut proteins apart just like a pair of scissors.

4. And then, the purification step

Now the DNA is in a "soup" of different materials (mainly proteins and grease). DNA precipitates when in the presence of alcohol. This allows for separation of the DNA from the rest of the material.

SIMPLY PUT, DNA EXTRACTION IS THE REMOVAL OF DEOXYRIBONUCLEIC ACID (DNA) FROM THE CELLS OR VIRUSES IN WHICH IT NORMALLY RESIDES.



EXTRACTION KITS

EASY TO USE, ERROR-RESISTANT

There are several extraction kits available from different companies. They are optimized for their aim and downstream use.

Our purpose is one of the simplest: DNA for amplicon sanger sequencing; so, any good conventional kit should do the job provided that the instructions are followed.

As we are dealing with insects, and depending on the development stage, care needs to be taken in the grinding step. Chitin is a major component of the insect cuticle and chitinous material may clog columns. A good grinding with pestle, and eventually aided by liquid nitrogen, might be useful when dealing with highly coriaceous species.

CTAB EXTRACTION

1. Allow specimens to dry on a filter paper
2. Using sterile forceps (flame over burner to sterilize) place each pupae in a 1.5 ml eppendorf tube
3. Sterilize forceps between samples
4. Crush the pupae with a pestle
5. Wash pestle in sodium hypochlorite between samples, and dry on paper
6. Add 500 μ L CTAB (2%) and 2 μ L Proteinase K (15 -20 mg/ml)
7. Put the eppendorf tubes in the shaker at 55°C and let them incubate for at least 1 h (but can also stay overnight)

----- Interruption of the protocol -----

8. Add 500 μ L Chloroform : Isoamylalcohol (24:1), mix it gently
9. Centrifuge at 10.000 rpm for 15 minutes at room temperature
10. Put the water phase (about 300 - 400 μ L) in a new tube, this is on top and contains the DNA. The lower phase can be wasted
11. Add 1 volume of cold Isopropanol to the solution
12. Put it in the fridge at -20 °C for a minimum of 30 minutes
13. Cool the centrifuge to 4°C
14. Centrifuge at 10.000 rpm for 15 minutes at 4°C
15. Remove the fluid from the pellet
16. Wash the pellet with 300 μ l 100% EtOH
17. Centrifuge at 14.000 rpm for 5 minutes at room temperature
18. Remove the fluid from the pellet
19. Wash the pellet with 300 μ l 70% EtOH
20. Centrifuge at 14.000 rpm for 5 minutes at room temperature
21. Remove the fluid from the pellet
22. Dry the pellet
23. Solute the pellet in 50 μ l MQ water and store it at -20°C



AMPLIFICATION

COMPONENTS

DNA polymerase: it is an enzyme that synthesizes DNA molecules by copying one double-stranded DNA molecule. There are several common DNA polymerases that are used for PCRs. DNA polymerases can only add new nucleotides to an existing strand of DNA.

Two primers (Forward and Reverse): are complementary to the 3' (three prime) ends of the sense and anti-sense strand of the DNA target. They are based on a previously known sequence and they serve as initiators for the DNA polymerase.

Deoxynucleoside triphosphates (dNTPs, nucleotides containing triphosphate groups): a deoxyribonucleotide is the monomer or single unit of the DNA, which is polymerized by the DNA polymerase to form DNA during the PCR reaction.

Buffer solution: it provides a suitable chemical environment for optimum activity and stability of the DNA polymerase.

Divalent cations: generally Mg^{2+} ($MgCl_2$) is used.

MiliQ water: an ultrapure deionized water.

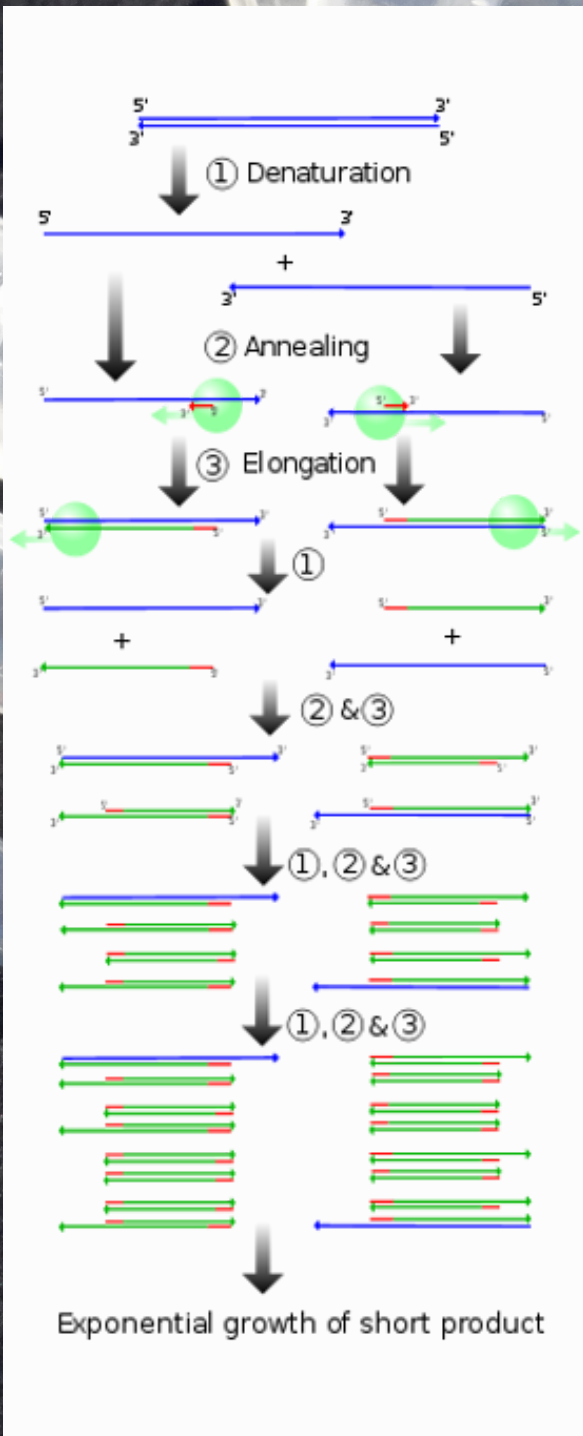
DNA template: contains the DNA region (target) to be amplified.

POLYMERASE CHAIN REACTION - PCR!

...TO MAKE A HUGE NUMBER OF COPIES OF A GENE FRAGMENT

PCR

HOW DOES THE CHAIN REACTION WORK?



Typically, PCR consists of a series of 20-40 repeated temperature changes, called cycles, with each cycle commonly consisting of usually three discrete temperature steps.

① Denaturation step:

the first regular cycling event and consists of heating the reaction to 94°C for 20-30 seconds. It causes DNA melting by disrupting the hydrogen bonds between complementary bases, yielding single-stranded DNA molecules.

② Annealing step:

temperature is lowered to 50-65 °C for 20-40 seconds allowing annealing of the primers to the template.

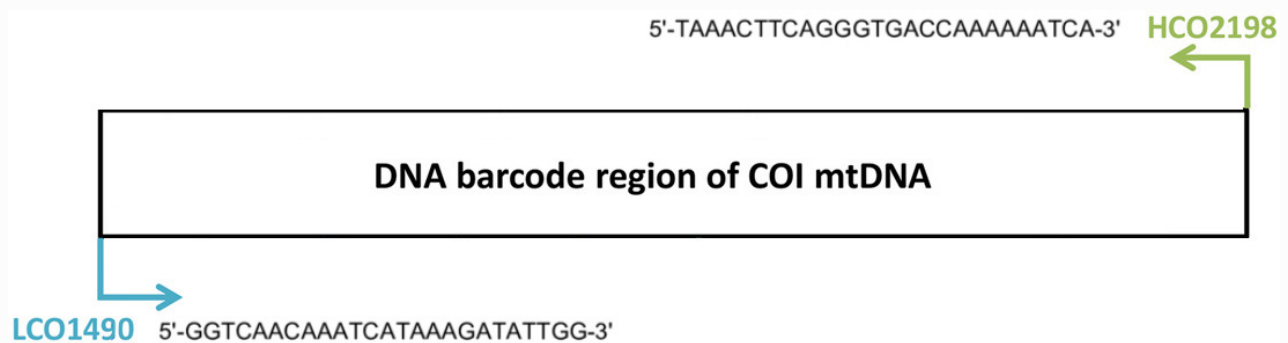
③ Elongation step:

DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs that are complementary to the template.

Several cycles of this steps result in an exponential growth of the product.

PROTOCOL FOR COI

The DNA primers designed by Folmer and co-workers, LCO1490 and HCO2198, amplified a 710-bp region of the mitochondrial cytochrome oxidase subunit I gene (COI) from a broad range of invertebrates.



They are called Universal Primers, however, sometimes their use resulted in a relatively low amplification success rate or even failure to amplify.

These problems are due either to the degradation of DNA or the target gene fails to be amplified by one or both primers. This led to the development and design of several taxa-specific primers.

TRY THIS COMBINATION FIRST. IF IT DOES NOT WORK... WELL THERE IS PLENTY OF LITERATURE TO INSPIRE THE NEXT STEP.

Think on which group/family your sample belongs to, and look for taxa specific primers, for instance. If it also fails (which is not likely) you can always design new primers, but that is a next level.

PROTOCOL FOR COI

Depending on the taq polymerase enzyme chosen, there will be the need to make adjustments to the protocol. However, as "rule of thumb" the following works well for most of the standard enzymes available.

In a traditional PCR protocol (final volume of 25 μL) a master mix is prepared according to the following table (where n is the number of DNA extractions to amplify, considering a positive and a negative control, plus one reaction for pipetting errors):

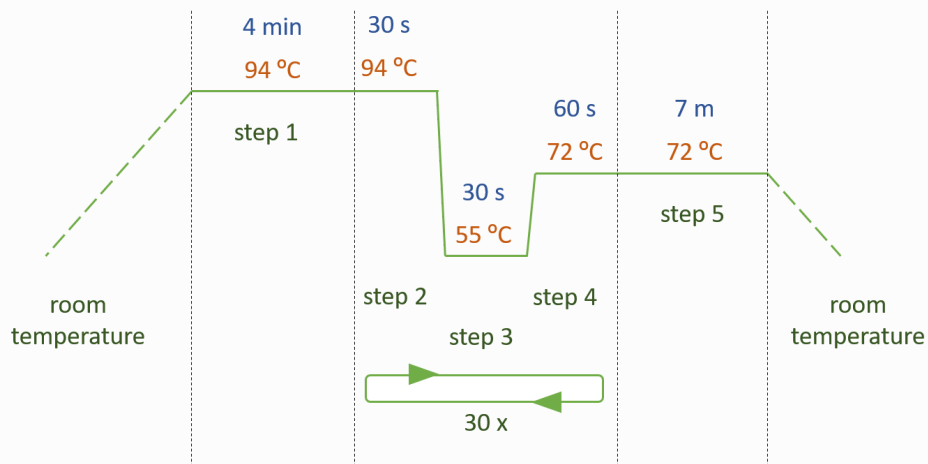
	Volume 1X	Master mix for samples (n+1)
Water	15.75	
dNTP	1.0	
Buffer 10X	2.5	
Mg ²⁺ solution (25 mM)	1.5	
Primer Rv (10 μM)	1.0	
Primer Fw (10 μM)	1.0	
Taq polymerase	0.2	
Total	24	

The master mix should be assembled as follows:

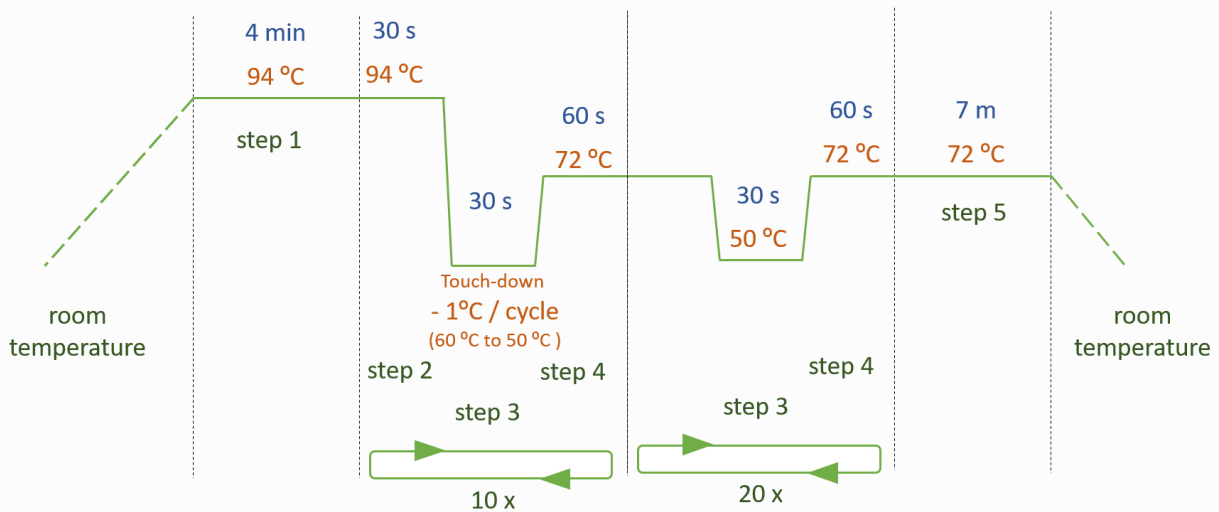
- Thaw all reagents and keep them on ice.
- Assemble reaction mix into a sterilized 1,5 or 2.0 mL eppendorf tubes.
- Add reagents in following order: water, buffer, dNTPs, MgCl₂, primers, and Taq polymerase.
- Gently mix by tapping the tube. Briefly centrifuge to settle tube contents.
- Aliquot 24 μL of the master mix into sterile 0.2 mL PCR tubes.
- Prepare negative control reaction without template DNA.
- Prepare positive control reaction with 1 μL template of known size.
- Prepare barcode reactions adding 1 μL of each DNA extraction.

PROTOCOL FOR COI

Usually, you can program your thermocycler for your PCR reaction using the following protocol:



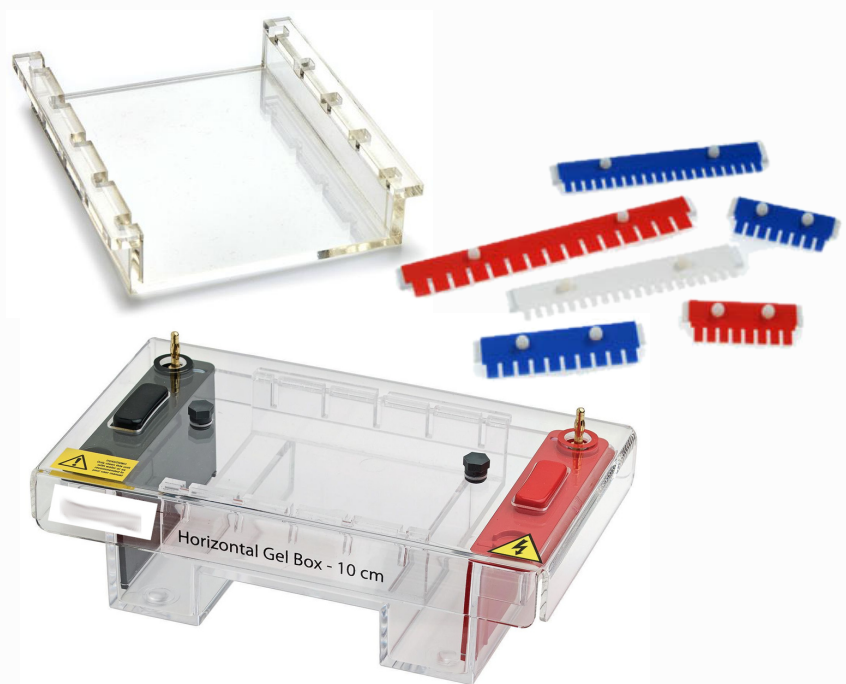
Or you might want to try a more robust, albeit less specific, protocol that is less sensitive to not having the right annealing temperature for the reaction mix (remember that the annealing temperature relates to the melting temperature of the primers):



VISUALIZATION

AGAROSE GEL ELECTROPHORESIS

To check whether the PCR generated the anticipated DNA fragment, agarose gel electrophoresis is employed for size separation of the PCR products. The size(s) of PCR products is determined by comparison with a DNA ladder (a molecular weight marker), which contains DNA fragments of known size, run on the gel alongside the PCR products.



PROTOCOL GEL ELECTROPHORESIS

The aim is to separate by size the DNA fragments obtained in the PCR.

Material:

- Agarose powder
- 1X Tris-acetate (TAE buffer)
- Gel casting tray and combs
- DNA ladder
- 6X Loading buffer
- Electrophoresis chamber with power supply
- Staining reagent (Greensafe or equivalent)

PROCEDURE

1-Prepare the gel

- Dilute 20 mL of 50X TAE buffer on 980 ml of water.
- Dissolve 1g agarose powder on 100 mL TAE 1X buffer.
- Heat it on the microwave until the agarose is completely melt.
- Add 2 μ L of the staining reagent if greensafe (or otherwise follow the manufacturer instructions)
- When the agarose is at 55 °C (when you can hold the bottle with your hands without scalding) pour the agarose on the casting tray with the comb(s).
- Wait until it is solid (it should appear milky white).
- Carefully pull out the comb(s) and place the gel in the electrophoresis chamber.
- Add enough TAE Buffer so that there is about 2-3 mm of buffer over the gel.

PROTOCOL GEL ELECTROPHORESIS

The aim is to separate by size the DNA fragments obtained in the PCR.

2-Load the gel

- Add 5 mL of 1X Sample Loading Buffer to each 5 mL PCR reaction
- Carefully pipette 10 mL of each sample/Sample Loading Buffer mixture into separate wells in the gel.
- Pipette 5 mL of the DNA ladder standard into at least one well of each row on the gel.

3-Run the gel

- Place the lid on the gel box, connecting the electrodes.
- Connect the electrode wires to the power supply.
- Turn on the power supply to about 100 V.
- Check to make sure the current is running through the buffer by looking for bubbles forming on each electrode.
- Check to make sure that the current is running in the correct direction by observing the movement of the loading dye - this will take a couple of minutes (it will run in the same direction as the DNA).
- Let the power run until the dye approaches the end of the gel.
- Turn off the power and disconnect the wires from the power supply.
- Remove the lid of the electrophoresis chamber.
- Using gloves, carefully remove the tray and gel.

4-Visualize the gel

- Place the gel on the UV light viewing tray. Close the tray with the cone and switch on the UV light.
- Take a picture and save it on a computer.



DNA PURIFICATION FOR SEQUENCING

“PREPARING ‘DNA FOR SEQUENCING’ IS A CRUCIAL AND IMPORTANT STEP IN THE DNA SEQUENCING PROCESS. THE QUALITY AND QUANTITY OF THE TEMPLATE DNA ARE TWO KEY FACTORS FOR GETTING BETTER SEQUENCING RESULTS.”

The quality and the quantity of the amplified DNA fragment intended to sequence, plays an important role in the process. A good quality template performs well during sequencing without any background noisy sounds, gaps or non-amplifications.

The PCR products are selected only when one single prominent DNA band of our interest obtained. Even though the amplicons or PCR products are the purest forms of DNA fragments, the unused primers, dNTPs other chemicals can block sequencing reaction. Therefore, they need to be removed!

Alcohol purification is usually not recommended for PCR product purification and using ready to use DNA purification kits is advisable.

SEQUENCING AND ANALYSES

SEQUENCING

Sequencing includes any method or technology that is used to determine the order of the four bases: adenine, guanine, cytosine, and thymine. There are several technologies and equipments, and the speed of innovation is astonishing. For the purpose of sequencing individual fragments of DNA, traditional Sanger sequencing is used.

This is the classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphates (dNTPs), and modified di-deoxynucleotide triphosphates (ddNTPs), the latter of which terminate DNA strand elongation. No worries, from the practical point of view you only need to send the purified DNA PCR product to the selected commercial sequence services provider and follow their instructions on template preparation. Then you only have to wait for the results.

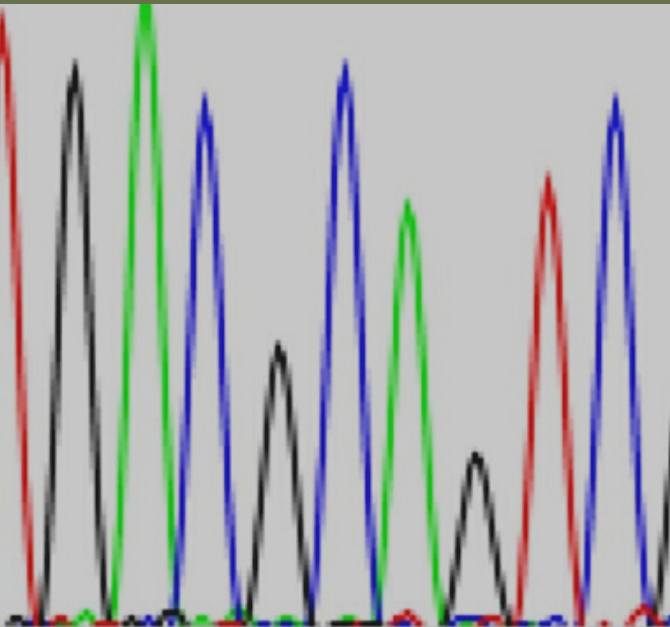
**SEQUENCING, IT IS
SIMPLY READING
THE ORDER OF DNA
NUCLEOTIDES**



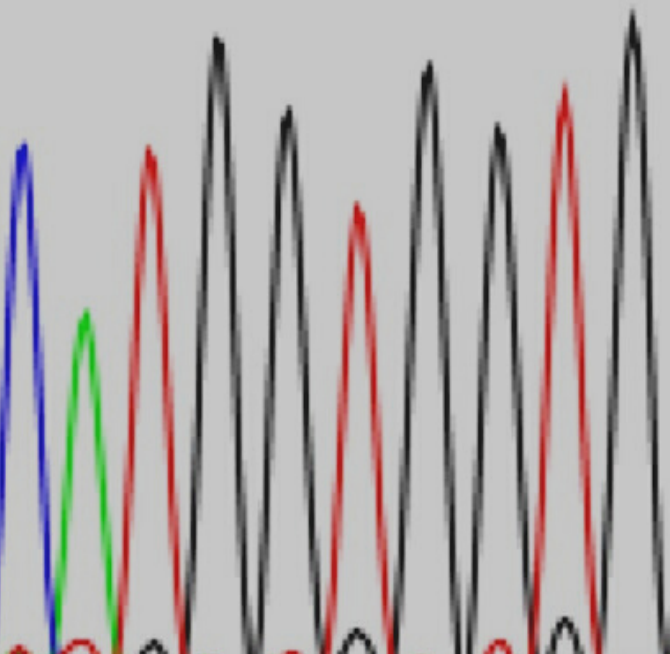
**Sequences are available,
and then what?**

T G A C G C A G T C C
200

VISUALIZATION OF CHROMATOGRAMS



C A T G G T G G T G
190



WHAT FILES TO LOOK AT?

In order to obtain good sequencing results, you **MUST** download and examine your sequencing chromatogram. If you are using just the text data, you could be getting conclusions based on data that is completely invalid!

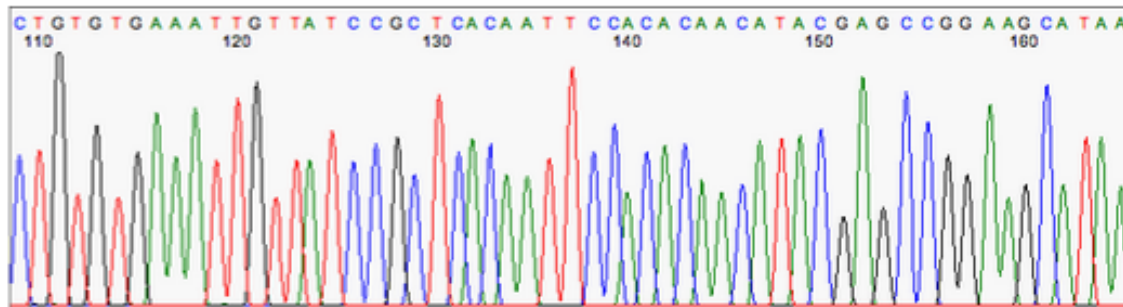
There are several free softwares available, and tools are often being added. Therefore, it makes no sense to show a list of available softwares as it probably be outdated in no time.

For the examples that follow, I will use Chromas, but feel free to choose the one that you feel more comfortable with.

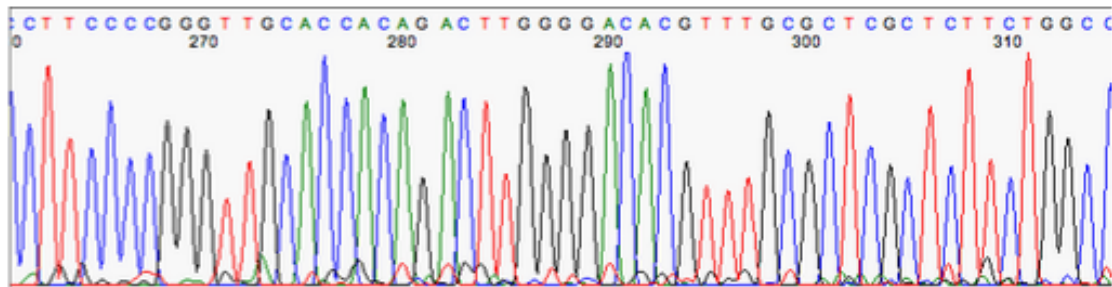
THE CHROMATOGRAM

1. Get a general sense of how clean the sequence is

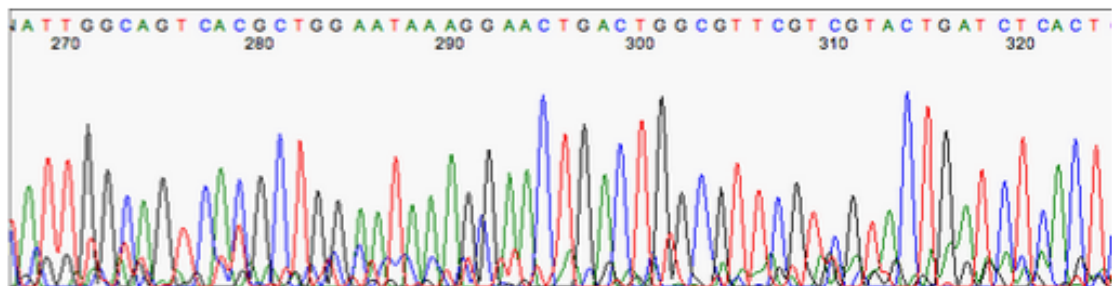
Here's an example of excellent sequence. Note the evenly-spaced peaks and the lack of baseline 'noise' (see further down for examples of higher baseline noise):



The next example has a little baseline noise, but the 'real' peaks are still easy to call, so there's no problem with this sample:



Now we get to an example that has a bit too much noise. Note the multicolored peaks at 271, 273 and 279, the oddly-spaced interstitial peaks near 291 and 301, and it is impossible to determine the real nucleotide is at 310.

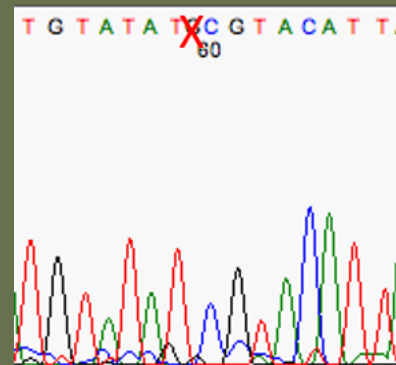
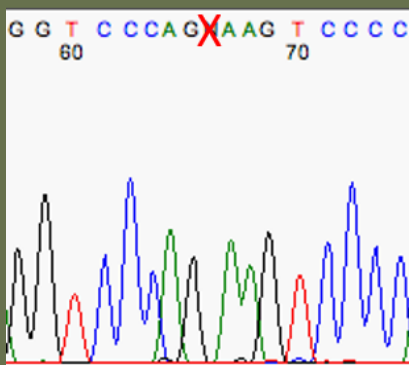


THE CHROMATOGRAM

2. Check for mis-called nucleotides

Sometimes the computer will mis-call a nucleotide when a human could do better. Most often, this occurs when the basecaller calls a specific nucleotide, when the peak really was ambiguous and should have been called as 'N'. Occasionally, the computer will call an 'N' when a human would be confident in making a more specific basecall.

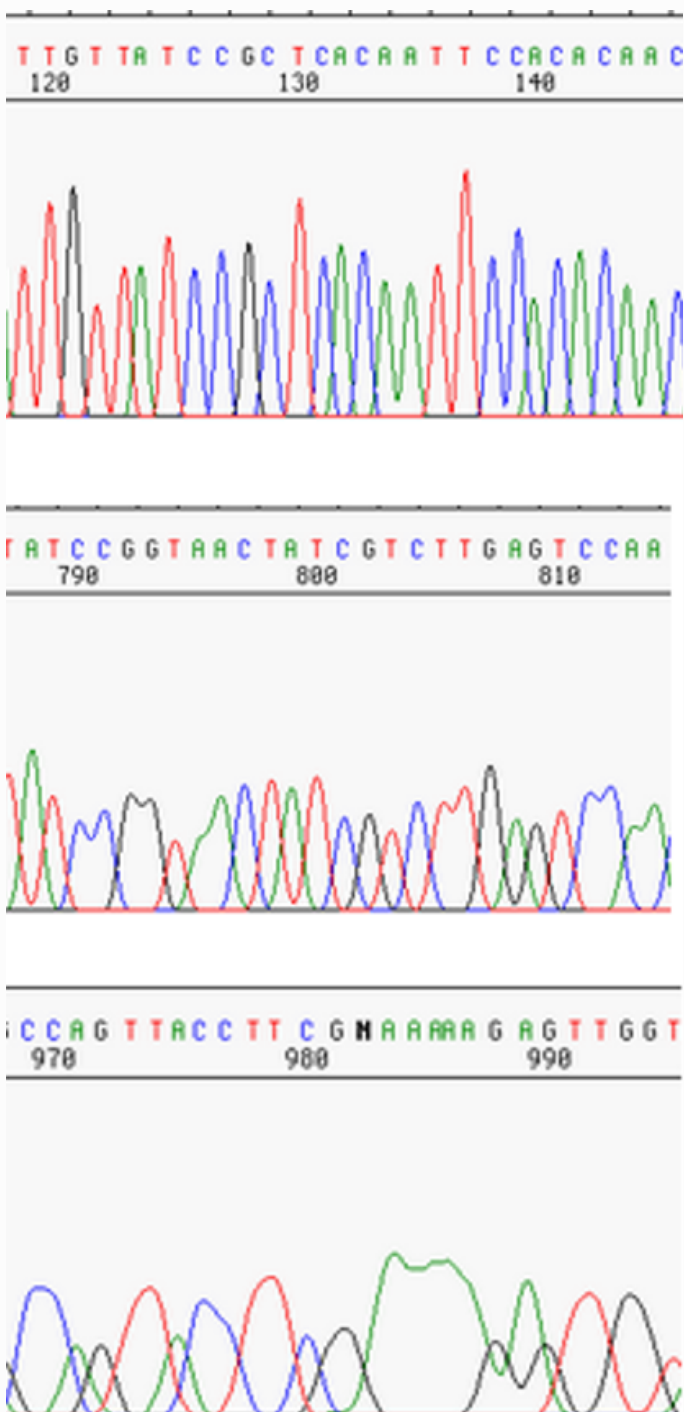
Quickly scan the chromatogram for extremely small peaks, 'N' calls, and any mis-spaced peaks or nucleotides.



This is a great example of why a weak sample, with its consequent noisy chromatogram, is untrustworthy.

THE CHROMATOGRAM

3. Loss of resolution at the ends of the chromatogram



As the sequencing progresses, it loses resolution. This is normal; peaks broaden and shift, making it harder to make them out and call the bases accurately.

Late in the chromatogram, watch for multiple bases of any one nucleotide where there really should be only one. Watch, too, for wide peaks miscounted by the program as two nucleotides, when it should have been just one.

Truncate the sequence when problems become too frequent for the purpose.

HOMOLOGY SEARCH

CTGGAG

CTGGGG

CTGGAG

CTGGTGGA

~~CTGGAG~~

CTAG

~~The fat cat s~~

hef atc ats a

WHICH DATABASES TO LOOK AT?

Once your sequence is clean the next step is to see its identity. Did you get the expected DNA fragment or you have a contamination? What is the similarity with other already published sequences?

Sequences are stored in databases, and they exchange information on a daily basis so that they are up-to-date and are synchronized.

Several databases are available and the choice depends mainly on aim. For insect DNA barcoding, two should be checked by this order:

1. BOLD
2. Genbank

BOLD

Boldsystems.org

BARCODE OF LIFE DATA SYSTEM ^{v4}

Advancing biodiversity science through DNA-based species identification.

EXPLORE THE DATA

DESIGNED TO SUPPORT THE GENERATION & APPLICATION OF DNA BARCODE DATA

BOLD is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada. It consists of four main modules, a data portal, an educational portal, a registry of BINs (putative species), and a data collection and analysis workbench.

The Barcode of Life Data System (BOLD) is an online workbench and database that supports the assembly and use of DNA barcode data. It is a collaborative hub for the scientific community and a public resource for citizens at large.

BOLD is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada. It consists of four main modules, a data portal, an educational portal, a registry of BINs (putative species), and a data collection and analysis workbench.

BOLD

BOLDSYSTEMS

DATABASES IDENTIFICATION TAXONOMY WORKBENCH RESOURCES LOGIN

IDENTIFICATION ENGINE

ANIMAL IDENTIFICATION [COI] FUNGAL IDENTIFICATION [ITS] PLANT IDENTIFICATION [RBCL & MATK]

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial Cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

Historical Databases: **Current** Jul-2019 Jul-2018 Jul-2017 Jul-2016 Jul-2015 Jul-2014 Jul-2013 Jul-2012 Jul-2011 Jul-2010 Jul-2009

Search Databases:

- All Barcode Records on BOLD (8,454,525 Sequences)**
Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.
- Species Level Barcode Records (4,212,554 Sequences/228,141 Species/106,497 Interim Species)**
Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.
- Public Record Barcode Database (2,210,094 Sequences/141,796 Species/55,059 Interim Species)**
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.
- Full Length Record Barcode Database (2,715,381 Sequences/203,712 Species/85,089 Interim Species)**
Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Enter fasta formatted sequences in the forward orientation:

```
>AY954425.1 Reticulitermes grassei isolate P22 cytochrome oxidase subunit II gene, partial
cds; mitochondrial
TTATATTAATAATCATTACAACCGTAATATACATAATAACCACCTACTTTGAAATAAACATACTAGACG
ATTCATACTAGAAGGACAATTAATTGAAACACGTGAACAATTGCTCCGGCAATCATCCTAGTATTCATT
GCAATACCATCCTTACGACTTCTATATTTAATAGACGAAATCCACAACCAACAATAACCCAAAAAGCAG
TAGGACACCAATGATACTGAAGATATGAATATTCAGACTTTACAAAATAAGAATTCGACTCCTACATAAT
TCCACAGGAGGAAAATCAAACAAGAACCCTCCGACTATTAGATACAGATAACCGAATCGTGCTACCTATA
AACTCCCAATCCGACTAGTAGTAACAGCAGCAGACGTATTACACTCATGAACAATCCAAGATTGGGGG
TGAAAACAGACGCCACACCAGGACGATTAATCAAACAAGATTCTCAATCAATCACCCCTGGTATCCTATA
CGGTCAGTGCTCAGAAATCTGTGGGCAAAATCACAGA
```

SUBMIT

Just copy and paste
your curated sequence here,
and press submit

GENBANK



U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite

blastn

blastp

blastx

tblastn

tblastx


Enter Query Sequence

Basic Local Alignment Search Tool - BLAST

Finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

GENBANK

 U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST[®] » blastn suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

Just copy and paste your curated sequence here, and press BLAST

Query subrange [?](#)
From
To

Or, upload file Não foi escolhido nenhum ficheiro [?](#)

Job Title
Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases
 [?](#)

Organism Optional
 exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional
 Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional
 Sequences from type material

Entrez Query Optional

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for
 Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm [?](#)



This work is supported by National Funds through FCT - Foundation for Science and Technology under the research project PTDC/ASP-PLA/30650/2017.



**"IF ALL MANKIND WERE TO
DISAPPEAR, THE WORLD WOULD
REGENERATE BACK TO THE RICH
STATE OF EQUILIBRIUM THAT
EXISTED TEN THOUSAND YEARS
AGO. IF INSECTS WERE TO
VANISH, THE ENVIRONMENT
WOULD COLLAPSE INTO CHAOS."
- E.O. WILSON**

LABORATORY OF ENTOMOLOGY

TÂNIA NOBRE
2021

