International Symposium on Ciliate Biology 2025 & Interdisciplinary Research

Pre-Symposium Workshops

on

1) Ciliate Identification, Taxonomy and Phylogeny
2) Genomics, Metagenomics and Bioinformatics (Eukaryotes)

organized by

Acharya Narendra Dev College

University of Delhi

under the aegis of DBT STAR STATUS SCHEME/IQAC

in association with

Indian Society of Ciliate Biology (ISoCB)

affiliated society of

Indian Network of Soil Contamination Research (INSCR) &
International Society of Protistologists (ISOP)

दिल्ली विश्वविद्यालय UNIVERSITY OF DELHI निष्ठा धृतिः सत्यम्

ACHARYA NARENDRA DEV COLLEGE असतो मा सद्गमय UNIVERSITY OF DELHI

CILIATE BIOLOGY LABORATORY ACHARYA NARENDRA DEV COLLEGE

INSCR Indian Network for Soil Contamination Research

partnered by

Ministry of Environment, Forest and Climate Change Government of India

ZOOLOGICAL SURVEY OF INDIA 1916

Department of Science & Technology Govt. of India

Department of Biotechnology Govt. of India

105 years IUBS

CSIR

Ministry Of Earth Sciences Government of India

Pre-Symposium Workshop: February 03, 2025

Acharya Narendra Dev College, New Delhi, INDIA

विकसित भारत अभियान

1947 TO 2047

Schedule

Time	**Activity (Venue)**	**Resource Person /Affiliation**
9:00am - 09:15am	Collection of ciliates from freshwater and soil (Zoology Lab II and III)	Prof Seema Makhija Ms Jyoti Dagar Ms Swati Maurya
9:15am - 10:00am	Physico-Chemical Analysis of water and soil sample	Ms Swati Maurya Mr Sandeep Antil Ms Jyoti Dagar Ms VijayaLaxmi PA Ms Arpita Sharma
10:00am - 10:15am	**Tea**	
10:15am - 10:45am	Ciliate Culturing	Ms Swati Maurya Mr Sandeep Antil Ms Anshika Tyagi Ms Urvashi Arora
10:45am - 1:00pm	Ciliate Identification	Prof. Seema Makhija Ms Swati Maurya Mr Sandeep Antil Ms Jyoti Dagar Mr Hritik Kadian
1:00pm - 1:30pm	**Lunch**	
1:30pm - 2:00pm	DNA Isolation from ciliates	Mr Sandeep Antil Ms Jyoti Dagar Mr Hritik Kadian
2:00pm- 2:30pm	Primer Designing	Prof. Ravi Toteja Mr Sandeep Antil
2:30 pm - 3:00 pm	PCR and Sequencing	Prof Ravi Toteja Dr S. Sripoorna Mr Sandeep Antil Ms Vinita Kumari Ms Rimjhim Shukla
3:00 pm – 4:00 pm	Phylogenetic Analysis using Bioinformatics Tools	Prof. Ravi Toteja Prof Seema Makhija Dr S. Sripoorna
4:00 pm	**Valedictory and Feedback**	
4:15 pm	**Tea**	

Workshop Organizers:

Patron	Prof. Ravi Toteja
Conveners	Dr Komal Kamra, Prof. Seema Makhija
Co-Conveners	Dr S. Sripoorna, Ms Swati Maurya, Mr Sandeep Antil, Ms Jyoti Dagar
Resource Person	Prof. Ravi Toteja, Prof. Seema Makhija, Dr S. Sripoorna, Ms Swati Maurya, Mr Sandeep Antil, Ms Jyoti Dagar, Mr Hritik Kadian, Ms Urvashi Arora, Ms Anshika Tyagi, Ms Vinita Kumari, Ms Rimjhim Shukla, Ms VijayaLaxmi PA, Ms Arpita Sharma
Support Staff (ANDC)	Mr Sachidanand Mishra, Mr Tara Dutt, Mr Promod Bhat, Mr Vikas Sharma, Mr Mahesh Kandpal, Mr Sanjeev Kumar, Mr Sagar Laishram, Mr Harshal, Ms Sumitra, Mr Dharmender, Mr Sachin

Organizing Institute

Acharya Narendra Dev College

Acharya Narendra Dev College is 32 years old, and through all these years, the college has helped to unfold the enormous potentialities of the students and empower them to meet the challenges of the future. Within this short span of the journey, the college has carved a niche of its own, and ANDC is considered one of the top colleges in India. This is evident from the NAAC score of 3.31 and 18 th NIRF-2022 ranking. Besides, India Today and Week rankings are also testimony to this. The College has one of the best infrastructure and research facilities for students and faculty at the University of Delhi. Our pedagogy is student centric. The College provide ample opportunities for students to excel in both academics and co-curricular activities like theatre, sports and NSS. Our college is Wi-Fi enabled, boasts of a well-equipped library, provides RFID enabled ID cards to students and issues laptops to them for their use. Other facilities in the college include a women's development cell, excursions (EXPLORE) etc. Several flagship schemes like *Paramarsh*, *Unnat Bharat Abhiyan and* DBT STAR COLLEGE SCHEME of Government of India have been sanctioned to the College. College has been granted DBT STAR STATUS by Department of Biotechnology. Collaborations with Auburn University of Montgomery, Alabama, USA, IIT, Delhi, School of open Learning, THISTI, NII, SPIE (USA) ANDC Chapter and NPTEL-SWAYAM chapter continue to report robust progress at national and international level. The 'Skill Hubs Pilot' is implemented in the college under central component of Pradhan Mantri Kaushal Vikas Yojana 3.0 (PMKVY 3.0). The existence of this collective erudite atmosphere in the college is seen to promote higher standards wherein innovative teaching methods become harbingers of quality education.

Workshop outline

Ciliate Identification, Taxonomy and Phylogeny

1) Sample Collection and Physico-Chemical Analysis
2) Culturing of Ciliates
3) Ciliate studies by classical and Molecular Tools

- **Classical Tools:**
  a. Live Cell Observation
  b. Protargol and Silverline Staining
  c. Feulgen Staining
- **Molecular Tools**
  a. DNA Isolation
  b. Primer Designing
  c. PCR
  d. Sequencing
  e. Phylogenetic analysis using Bioinformatics Tools

Experiment 1: Collection of freshwater and soil samples.
Materials required: Beakers, bottles, gloves, thermometer, pH strips, tissue paper.
Procedure:

I. Water Samples

1. Collect the samples at a depth of one foot from various corners of the lakes mixed with some vegetation.
2. Use Nytex nets of decreasing mesh sizes to filter out large crustaceans, debris and other unwanted materials
3. Transfer the concentrated ciliate fauna to large troughs in the laboratory.
4. Mixed planktonic cultures are initially grown at room temperature with the addition of freshly boiled cabbage pieces to promote the growth of bacteria which serve as their food organism.
5. Observe the water samples under the microscope for about 20 days.

II. Soil samples

1. Collect soil samples from various locations of the soil-water interface of water bodies along with partially decomposed leaves and small roots.
2. Record the temperature and pH of soils on the sites for each collection.
3. Store samples in sterile plastic bags, bring to the laboratory and process them immediately to record the presence of ciliate fauna.
4. Analyze soil samples with the non-flooded Petri Dish method (Foissner, 1997).
5. Draw about 3 ml of run-offs from the Petri dishes after 24, 48 and 72 hr and observe under microscope.
6. Record and isolate the ciliates that excyst and use for subculturing.

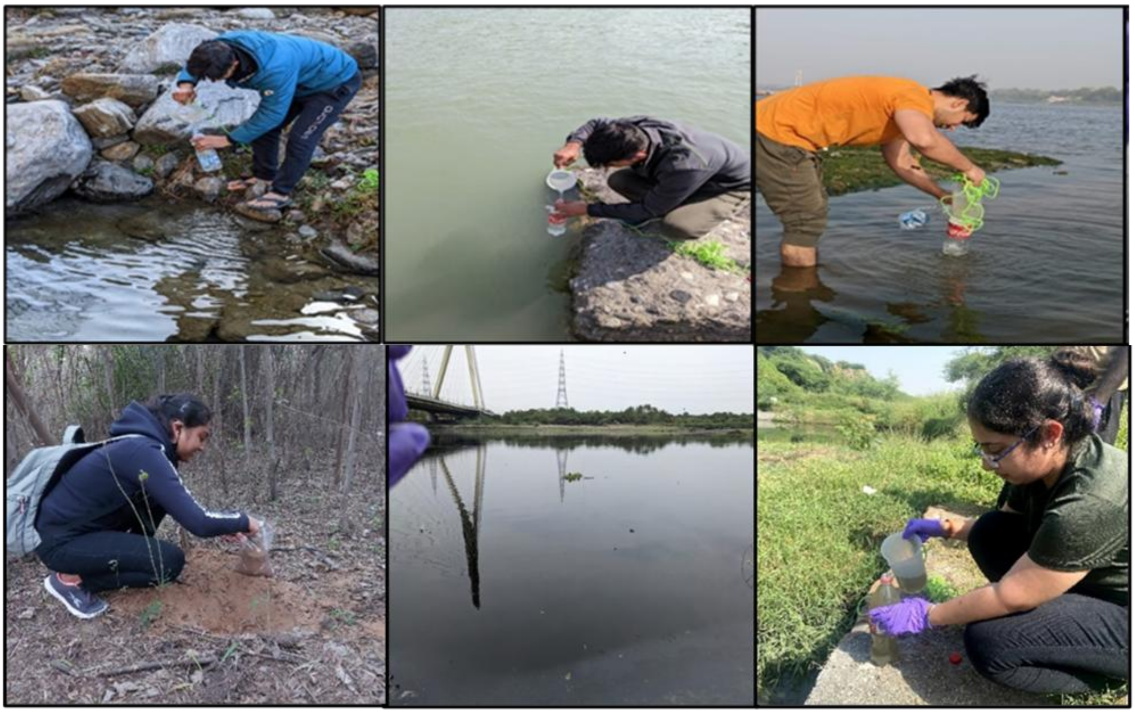


Fig. 1. Images showing sample collection.

Experiment 2: Physico-Chemical Analysis of Freshwater and Soil samples

I. Electrical conductivity (EC), total dissolved solids (TDS), oxidation-reduction potential (ORP), resistivity, and salinity will be measured using a water multiparameter analysis apparatus.

Procedure

1. Prepare the Sample:

   Take a 20 mL water sample in a clean container.

2. Insert the Electrodes:

   Place the electrodes of the multiparameter analysis apparatus into the water sample. Ensure they are submerged adequately for accurate readings.

3. Select the Measurement Parameter:

   Choose the parameter you wish to measure (Electrical Conductivity (EC), Total Dissolved Solids (TDS), Oxidation-Reduction Potential (ORP), Resistivity, or Salinity).

4. Start the Measurement:

   Press the "Measure" button on the apparatus to initiate the measurement for the selected parameter.

5. Record the Readings:

   Once the measurement stabilizes, record the reading displayed on the device.

6. Repeat for Other Parameters:

   If measuring multiple parameters, clean the electrodes between measurements, and repeat steps 3–5 for each parameter.

7. Clean Up:

   After measurements, clean the electrodes properly to prevent contamination for future samples.

Notes:

- Ensure that the electrodes are calibrated correctly for accurate results.
- Be aware of any specific instructions or settings recommended by the device manufacturer for each parameter

II. Ammonia (NH_3), Nitrite (NO_2^-), and Nitrate (NO_3^-) concentrations will be estimated by API-test kit following the manufacturer's protocol.

Ammonia

This salicylate-based ammonia test kit reads the total ammonia level in parts per milion (ppm) which are equivalent to mg/L from 0-8.0 ppm (mg/L).

Components

Ammonia Test Solution #1, Test Solution #2, Test tube, Ammonia Color Chart**.**

Procedure

1. Fill a clean test tube with 5 ml of water to be tested.
2. Add 8 drops from Ammonia Test Solution #1, holding the dropper bottle upside down in a completely vertical position to assure uniformity of drops.
3. Add 8 drops from Ammonia Test Solution #2, holding the dropper bottle upside down in a completely vertical position to assure uniformity of drops.
4. Cap the test tube & shake vigorously for 5 seconds
5. Wait 5 minutes for the color to develop.
6. Read the test results by comparing the color of the solution to the Ammonia Color Chart. The tube should be viewed in a well-lit area against the white area of the chart. The closest match indicates the ppm (mg/l) of ammonia in the water sample, Rinse the test tube with clean water after use.

Nitrite (NO_2^-)

This test kit reads total nitrite level in parts per million (ppm) which are equivalent to milligrams per liter (mg/L) from 0 - 5.0 ppm (mg/L).

Components

Nitrite test solution, Test tube, Nitrite color chart.

Procedure

1. Fill a clean test tube with 5 ml of water to be tested.
2. Add 5 drops of Nitrite Test Solution, holding dropper bottle upside down in a completely vertical position to assure uniformity of drops.
3. Cap the test tube and shake for 5 seconds.
4. Wait 5 minutes for the color to develop.
5. Read the test results by comparing the color of the solution to the Nitrite Color Chart. The tube should be viewed in a well-lit area against the white area of the chart. The closest match indicates the ppm (mg/L) of nitrite in the water sample. Rinse the test tube with clean water after use**.**

Nitrate (NO_3^-)

This test kit reads total nitrate level in parts per millon (ppm) which are equivalent to mg/L from 0-160 ppm.

Components

Nitrate test solution #1, Nitrate test solution #2, Test tube, Nitrite color chart.

Procedure

1. Fill a clean test tube with 5 ml of water to be tested.
2. Add 10 drops from Nitrate Test Solution #1, holding dropper bottle upside down in a completely vertical position to assure uniformity of drops.
3. Cap the test tube & invert tube several times to mix solution.

4. Vigorously shake the Nitrate Test Solution #2 for at least 30 seconds. This step is extremely important to insure accuracy of test results.
5. Now add 10 drops from Nitrate Test Solution #2. Holding dropper bottle upside down in a completely vertical position to insure uniformity of drops.
6. Cap the test tube and shake vigorously for 1 minute. This step is extremely important to insure accuracy of test results.
7. Wait 5 minutes for the color to develop
8. Read the test results by comparing the color of the solution to the Nitrate Color Chart. The tube should be viewed in a well-lit area against the white area of the card. The closest match indicates the ppm (mg/L) of nitrate in the water sample. Rinse the test tube with clean water after use.

Experiment 3: Culturing of ciliates

Materials required: Petri dishes, Pringsheim’s medium (PM), micropipettes, cavity blocks, tissue paper, filter sheet, BOD incubator.

Pringsheim’s medium (PM)

Stock solution:

- 0.85 mM Calcium nitrate tetrahydrate $[Ca(NO_3)_2.4H_2O]$
- 0.35 mM Potassium chloride (KCl)
- 0.08 mM Magnesium sulfate ($MgSO_4.7H_2O$)
- 0.11 mM Sodium hydrogen phosphate ($Na_2HPO_4.2H_2O$)

Working solution:

Add 5 ml of each stock solutions to 1 liter of distilled water and boil before use to sterilize the medium. After cooling, the medium is used for culturing ciliates.

Procedure

1. Raise clonal cultures of ciliates by isolating single cells in cavity blocks from the freshwater samples by using micropipettes.
2. Transfer the clonal cultures in Pringsheim's medium (Chapman-Andersen, 1958) and keep it at 22º–23ºC in BOD incubator.
3. Add small pieces of boiled cabbage to the medium to promote the growth of bacteria which serve as the primary food for ciliates (Gupta et al., 2001).

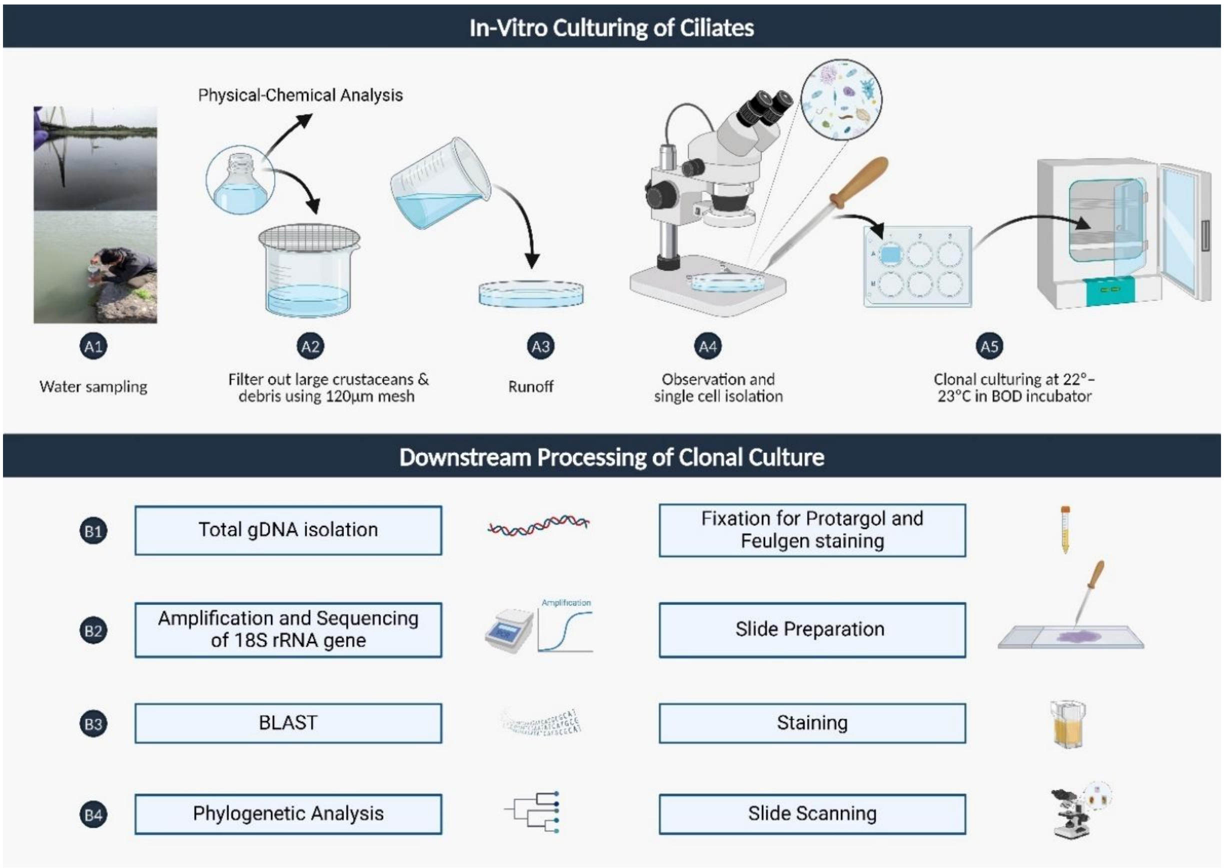

In-Vitro Culturing of Ciliates
Physical-Chemical Analysis
A1
Water sampling
A2
Filter out large crustaceans & debris using 120µm mesh
A3
Runoff
A4
Observation and single cell isolation
A5
Clonal culturing at 22°–23°C in BOD incubator
Downstream Processing of Clonal Culture
B1
Total gDNA isolation
Fixation for Protargol and Feulgen staining
B2
Amplification and Sequencing of 18S rRNA gene
Amplification
Slide Preparation
B3
BLAST
Staining
B4
Phylogenetic Analysis
Slide Scanning

Fig. 2. Images showing ciliate culturing in the laboratory.

Experiment 4: Identification of ciliates by live cell observation and staining.
Materials required: Ciliate samples, stereoscopic microscope, Slides, coverslips etc.

<u>Live cell observation</u>

Procedure:

1. Identify freshwater ciliates by stereoscopic, phase contrast and differential interference contrast (DIC) microscopy.
2. Picking up cells from cultures with a micropipette and place them on a clean and non-greasy slide.
3. Apply petroleum jelly at the corners of the coverslip. Petroleum jelly creates a seal that keeps the cells alive for up to a few hours and helps in taking live cell images.
4. Cells are viewed under the microscope at 4X, 10X, 40X and 100X.

Results:

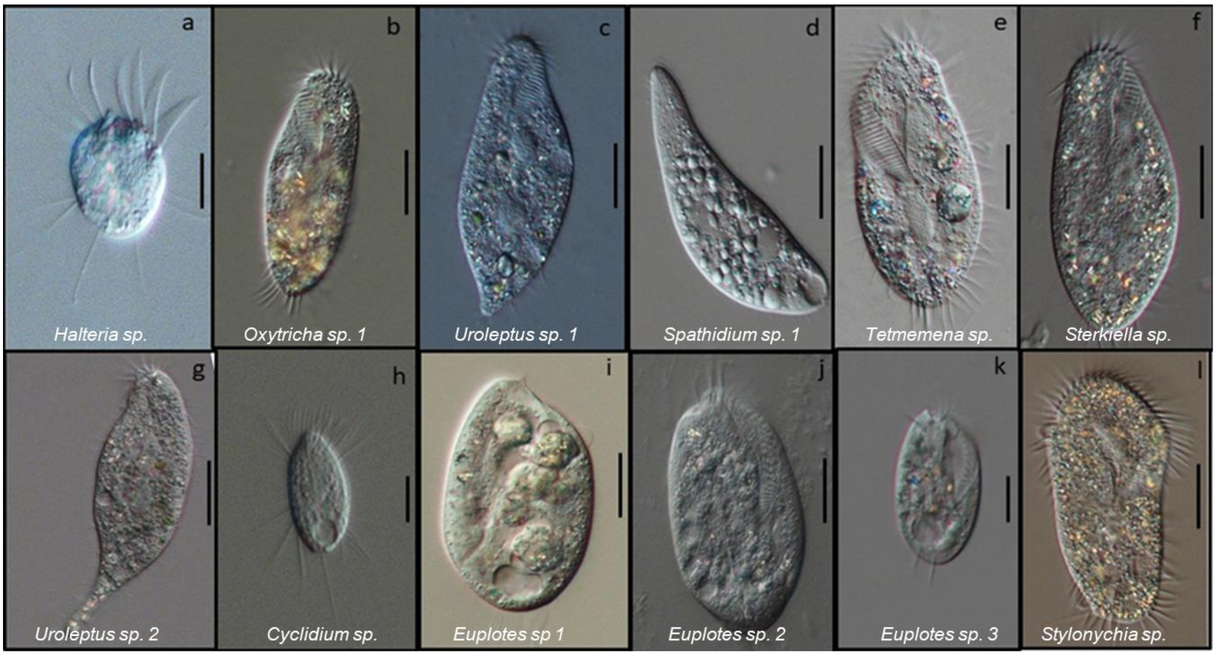


Fig. 3. Live cell images of ciliates.

Staining techniques

Protargol impregnation is used for visualizing surface ciliature (Kamra and Sapra, 1990). Chatton and Lwoff wet silver nitrate staining is used for identifying silver-line system (Chatton and Lwoff 1930, 1936; Foissner 2014). Nuclear observations are done by using the Feulgen stain (Feulgen, 1914; Chieco and Derenzini, 1999).

a. Protargol impregnation

Materials required:

(i) Bouin's fixative (prepared fresh)

- 4.5 ml Picric acid
- 0.5 ml Formalin
- 5 ml Mercuric chloride
- 9 drops Glacial acetic acid

(ii) Meyer's albumin (prepared fresh)

Egg white mixed with glycerol in a 1:1 ratio.

(iii) Protargol stain (prepared fresh)

0.4 gm of Protargol stain is added to 33 ml of distilled water and incubate at 60°C for proper mixing of stain with water.

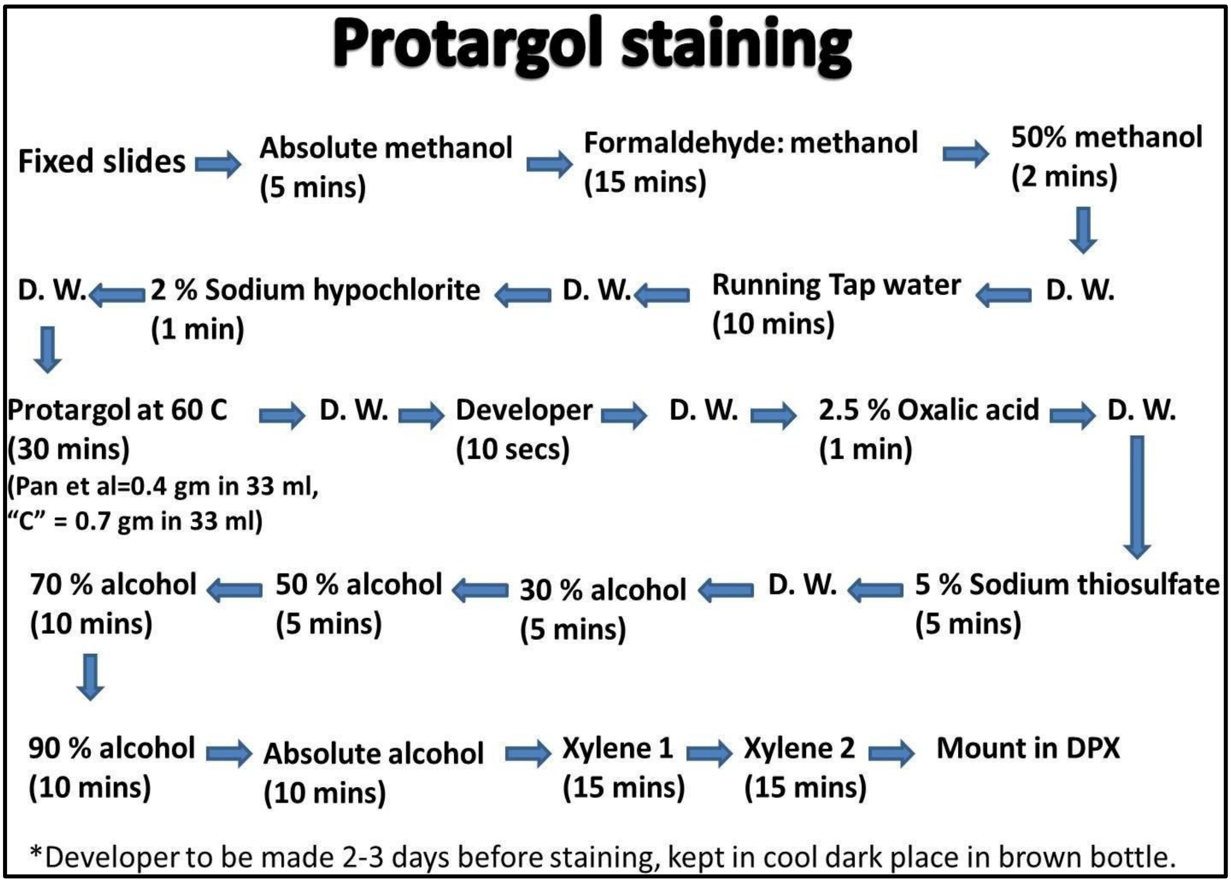
(iv) Developer (prepared one day before the experiment)

- 2 gm Sodium sulfate
- 0.3 gm Hydroquinone
- 1.4 gm Boric acid
- 15 ml Acetone

Make up the total volume to 100 ml by adding distilled water.

Procedure

1. Pellet the cells and fix in Bouin's fixative for half an hour.
2. After half an hour, wash the cells in distilled water three to four times to remove the excess fixative.
3. Enrobe the fixed cells on clean slides with Meyer's albumin and keep for 24 h.
4. Dip the slides in absolute methanol (5 min), 1:1 formaldehyde: methanol (15 min) and 50% methanol (2 min) to remove the albumin-film covering the cells.
5. After a rinse in distilled water, wash the slides thoroughly in tap water to remove the excess of formalin.
6. Bleach the slides in 2% sodium hypochlorite for 1–3 min depending on the thickness of the albumin layer and wash in distilled water.
7. Stain the slides in Protargol stain for 30–40 min.
8. Develop the slides by dipping into the developer.
9. After 1–2 min, wash the slides in distilled water and dip in 2.5% oxalic acid for a minute to stop the developing process.
10. Dip the slides in 5% sodium thiosulfate for 5 min for fixation of silver staining and wash again.
11. Finally, dehydrate the slides by passing through different grades of alcohol (30%, 50%, 70%, 90%, and 100%) for 5–10 min each, dip in xylene for 15 min and mount in Distyrene Plasticizer Xylene (DPX).



*Developer to be made 2-3 days before staining, kept in cool dark place in brown bottle.

Results:

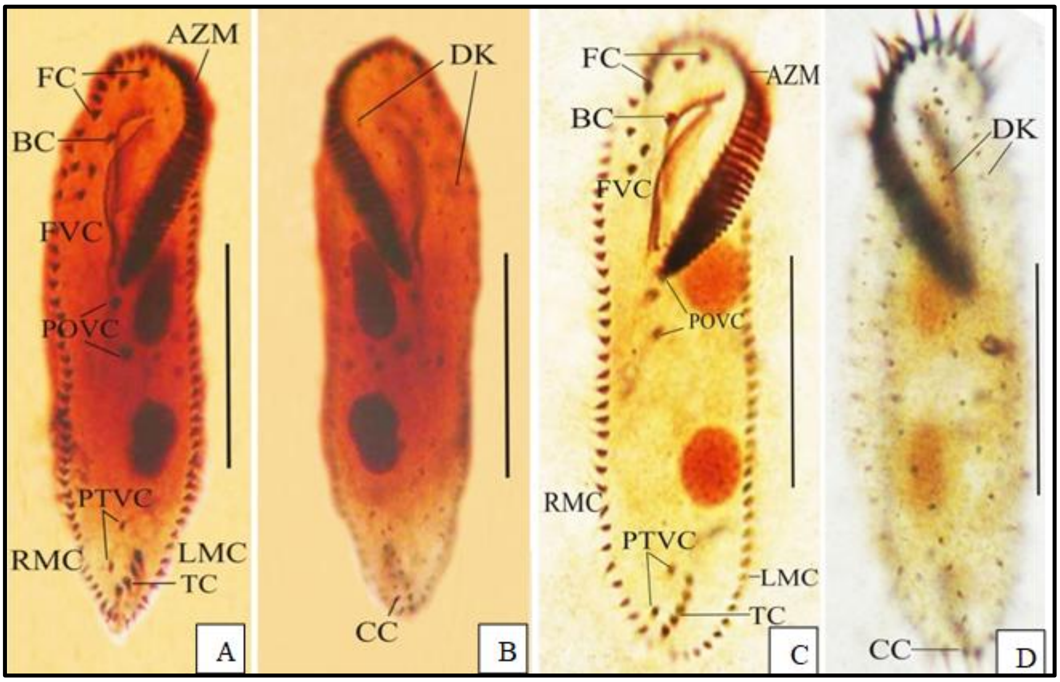


Fig. 4. Protargol stained images of ciliates.

b. Silver staining

Materials required:

(i) Champy fluid (prepared fresh)

- 3% Potassium dichromate (7 drops)
- 1% Chromate (7 drops)
- 2% Osmium tetroxide (8 drops)

(ii) Da Fano's fluid (prepared fresh)

- 1 gm Cobalt nitrate
- 1 gm Sodium chloride
- 10 ml Formalin

Total volume made up to 100 ml using distilled water.

(iii) Gelatin (prepared a week before use)

- 10 gm powdered Gelatin
- 0.1 gm NaCl

Mix the above-mentioned contents in distilled water and make the total volume up to 100 ml. Slightly heat this mixture to dissolve gelatin in water and store in the refrigerator.

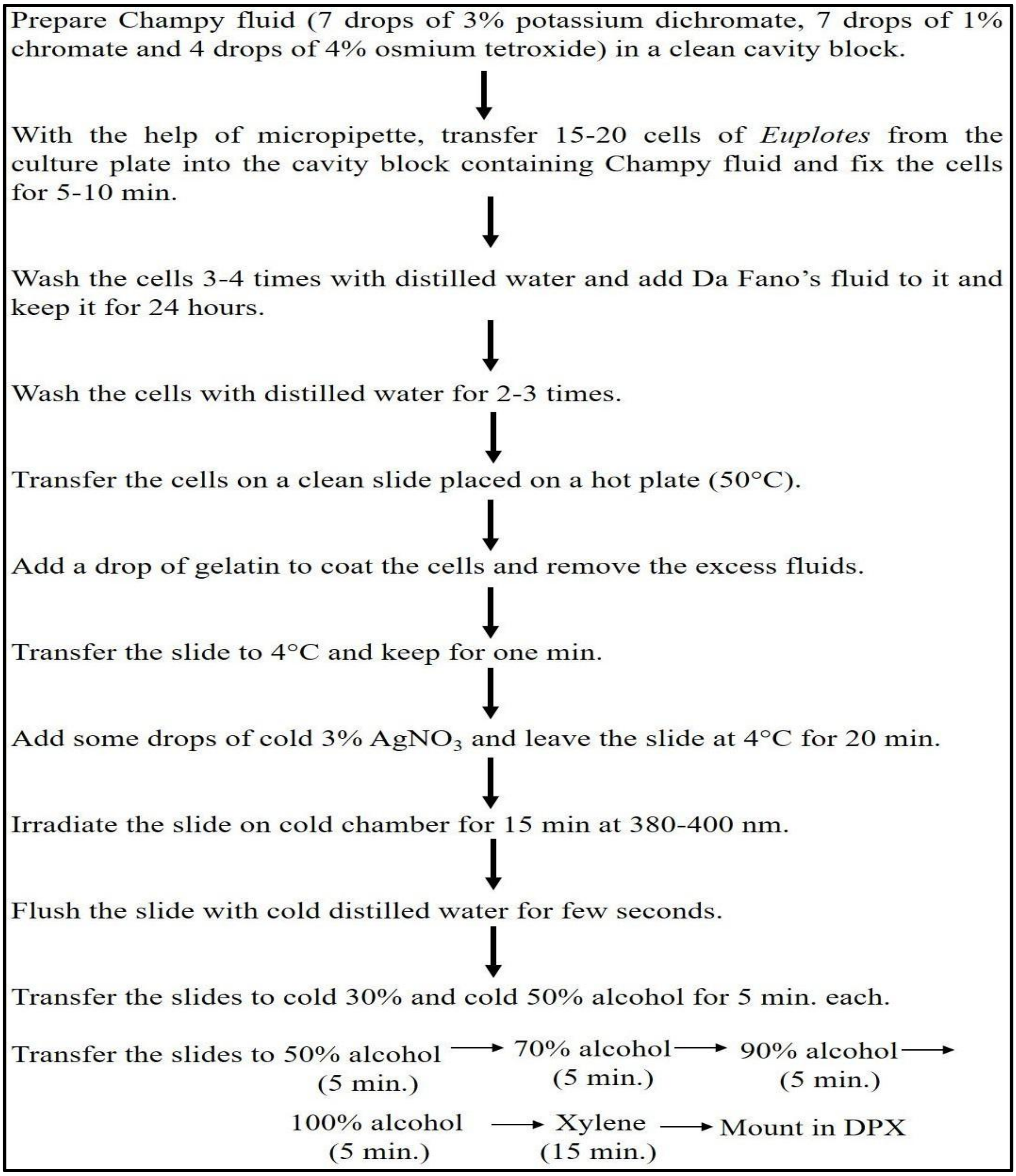
(iv) 3% Silver nitrate ($AgNO_3$) (prepared fresh)

- 1.5 gm $AgNO_3$ dissolve in 50 ml of distilled water.

Procedure

1. Fix the ciliates in Champy's fluid for 10 min.
2. After fixation, wash the ciliates 3–4 times with distilled water to remove Champy's fluid.
3. To this, add Da Fano's fluid and keep it for 24 hours.
4. After 24 h of incubation, wash the cells twice with distilled water.
5. Pick the fixed cells and transfer them to pre-heated, clean and non-greasy slides.
6. To the slides, add a drop of gelatin quickly and mix it with the organisms thoroughly.
7. Remove the excess fluid with a micropipette to have a thin layer of gelatin.
8. Immediately, transfer the slides to a cold chamber to solidify gelatin.

9. To the solidified gelatin, add 3–5 drops of cold 3% $AgNO_3$ and keep in the refrigerator for 20 min.
10. Irradiate the slides with ultraviolet light on a cold chamber for 15 min at 380–400 nm till $AgNO_3$ turns light brown in color.
11. Flush the slides with cold distilled water, dip in cold 30% and 50% alcohol grade for 5 min each.
12. Dehydrated the slides by passing through different grades of alcohol (50%, 70%, 90%, and 100%) for 10 min each, dip in xylene for 15 min and mount in DPX.



Prepare Champy fluid (7 drops of 3% potassium dichromate, 7 drops of 1% chromate and 4 drops of 4% osmium tetroxide) in a clean cavity block.

With the help of micropipette, transfer 15-20 cells of *Euplotes* from the culture plate into the cavity block containing Champy fluid and fix the cells for 5-10 min.

Wash the cells 3-4 times with distilled water and add Da Fano's fluid to it and keep it for 24 hours.

Wash the cells with distilled water for 2-3 times.

Transfer the cells on a clean slide placed on a hot plate (50°C).

Add a drop of gelatin to coat the cells and remove the excess fluids.

Transfer the slide to 4°C and keep for one min.

Add some drops of cold 3% $AgNO_3$ and leave the slide at 4°C for 20 min.

Irradiate the slide on cold chamber for 15 min at 380-400 nm.

Flush the slide with cold distilled water for few seconds.

Transfer the slides to cold 30% and cold 50% alcohol for 5 min. each.

Transfer the slides to 50% alcohol (5 min.) → 70% alcohol (5 min.) → 90% alcohol (5 min.) → 100% alcohol (5 min.) → Xylene (15 min.) → Mount in DPX

Results:

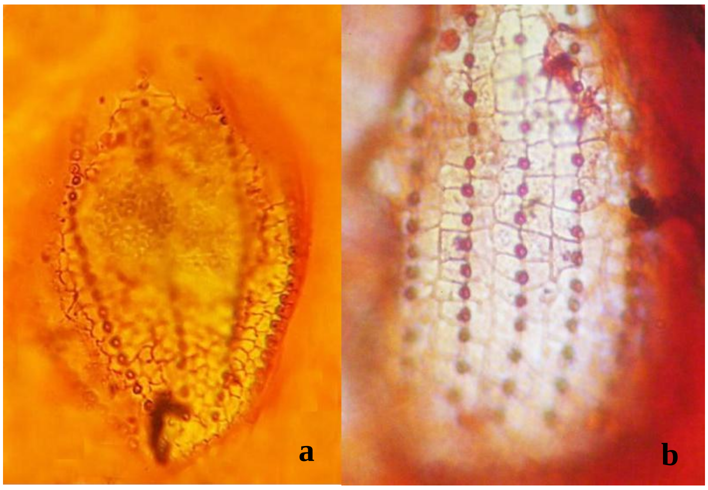


Fig. 5. Silver line network of Euplotid ciliates.

c. Feulgen staining

Materials

(i) Carnoy's fixative (prepared fresh)

- 4 ml Methanol
- 1 ml Glacial acetic acid

(ii) 1N Hydrochloric acid (HCl) (prepared fresh)

To make 100 ml of 1N HCl, add 12 ml of HCl to 88 ml of distilled water.

(iii) Schiff's reagent

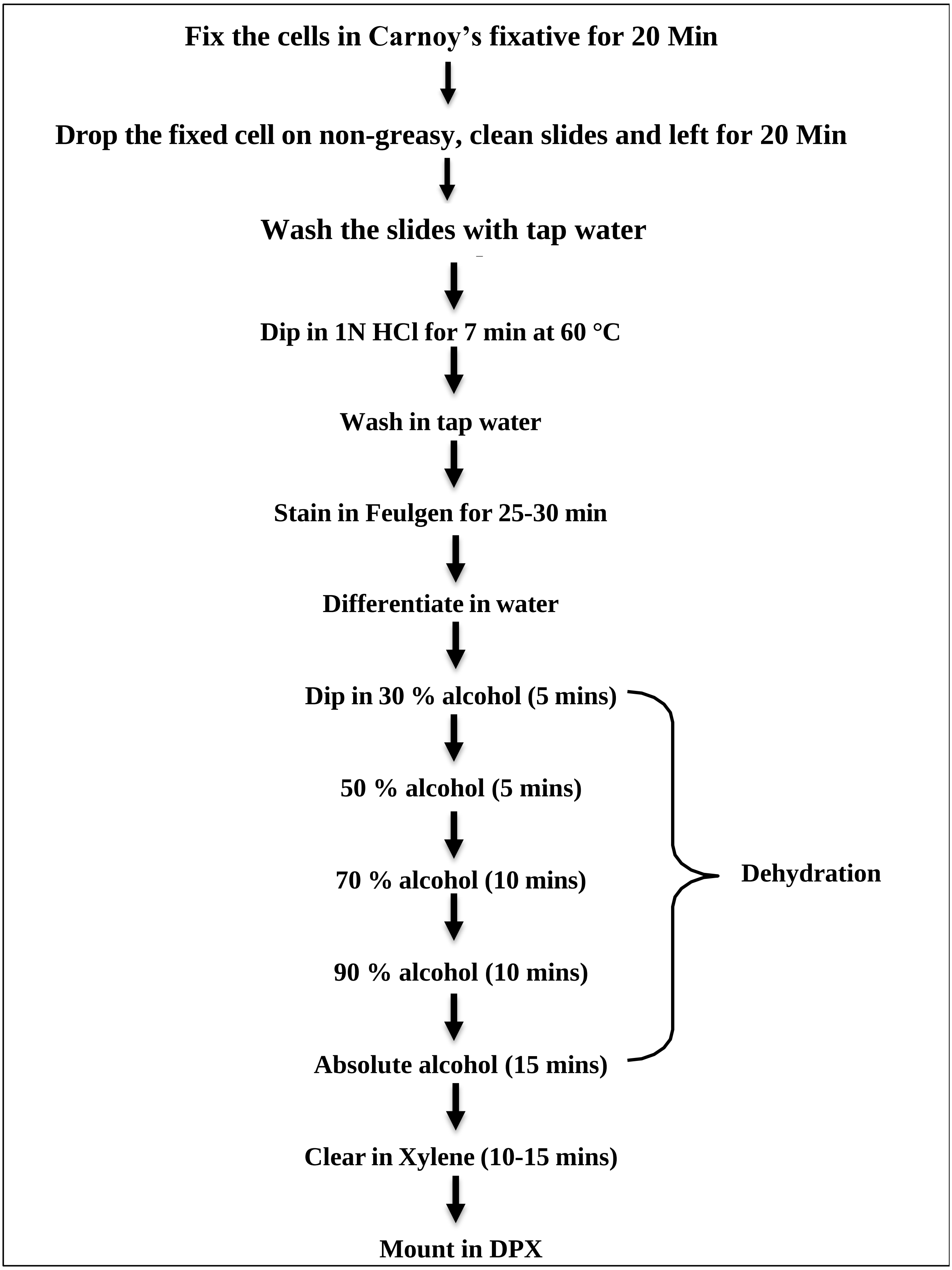
- 0.5 gm Basic Fuschin
- 2 gm Potassium metabisulfate
- 12 ml 1N HCl

Dissolve the above-mentioned contents in 100 ml of boiled water and store in dark to avoid light contact for 24 h. After 24 h, add 1 gm of charcoal, mix thoroughly and keep for half an hour. Transfer this mixture to another dark bottle by filtering it through 1 mm of Whatman filter paper to obtain colorless and transparent Schiff's reagent.

Procedure

1. Pellet the cells and fix in Carnoy's fixative for 20 min.
2. After 20 min of fixation, drop the cells on non-greasy and clean slides from a distance so that the cells can properly disperse onto the slide for analysis.
3. After the slides get dried, hydrolyze in 1N HCl maintained at 60°C for 7 min and wash the slides with distilled water.
4. Stain the slides in Schiff's reagent for 30 min.
5. After 30 min of staining, wash the slides in running tap water and dehydrate by passing through different grades of alcohol (30%, 50%, 70%, 90%, and 100%) for 5– 10 min each, dip in clear xylene for 15 min and mount in DPX.

Feulgen Staining



Fix the cells in Carnoy’s fixative for 20 Min

Drop the fixed cell on non-greasy, clean slides and left for 20 Min

Wash the slides with tap water

Dip in 1N HCl for 7 min at 60 °C

Wash in tap water

Stain in Feulgen for 25-30 min

Differentiate in water

Dip in 30 % alcohol (5 mins)

50 % alcohol (5 mins)

70 % alcohol (10 mins)

90 % alcohol (10 mins)

Absolute alcohol (15 mins)

Dehydration

Clear in Xylene (10-15 mins)

Mount in DPX

Results:

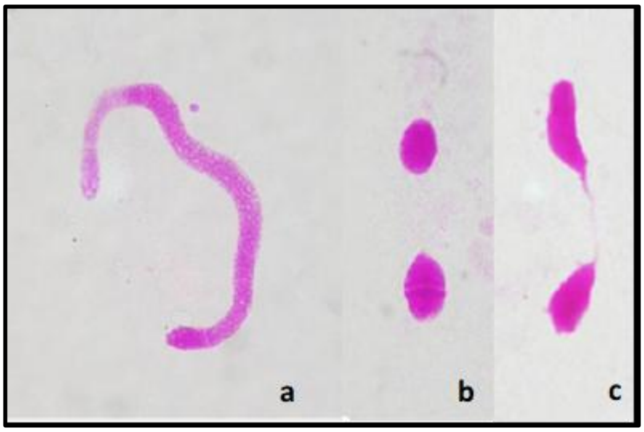


Fig. 6. Feulgen- stained nuclei of ciliates (a) *Euplotes aediculatus*, (b) *Aponotohymena isoaustralis* and (c) *Blepharisma* sp.

Significance of collection, culturing and staining of ciliates

- **Ciliate**s are present in freshwater bodies (ponds, lakes, rivers etc.) and in marine environments. They are also present in soil and terrestrial ecosystems in the form of cysts. Their role is immense as they are part of the microbial loop, function as recyclers, remineralizers of organic material in terrestrial and aquatic systems, they prey upon the bacteria and smaller protists and maintain the ecosystem balance. They are extremely useful models in studies of cell biology, molecular biology, genetics, ecology and evolution. They have extensively been used as model system to assess heavy metal toxicity in freshwater ecosystems and also been used to detect stress induced morphological anomalies and in the expression of genes such as hsp70, superoxide dismutase and metallothionein.
- **Ciliate culturing** by raising clonal population are useful because it can minimize the error in experimentation and can produce reproducible results.
- Ciliates can be identified using microscopic techniques involving observation of live cells or by staining using Protargol (for ciliature), wet silver nitrate staining or Feulgen stain (for nucleus).
- **Live cell observations** can be done using stereoscopic, phase contrast and differential interference contrast microscopes. Live cell observation reveals the cell morphology in living cells, the presence or absence of granules, the pigmentation and the shape and placement of contractile vacuole etc.
- **Protargol staining** reveals the ciliate species' distinct ventral and dorsal morphology.
- **Wet silver nitrate staining** reveals the silver line network in ciliates, which is an important characteristic for species identification (especially the Euplotids).
- **Feulgen staining** reveals the nuclear morphology of the ciliates. Ciliates have two morphologically and functionally distinct nuclei, the macronuclei and the micronuclei. The shape and size of the nuclear apparatus also varies from species to species.

Molecular Tools

Experiment 5: Isolation of Genomic DNA from Eukaryotic cells (Ciliate cells)

Introduction:

DNA is found in both prokaryotes and eukaryotes. In prokaryotes, DNA is double-stranded and circular and is found throughout the cytoplasm and is called nucleoid. In eukaryotes, DNA is located in the nucleus and mitochondria or chloroplasts. The DNA in the nucleus is double-stranded and linear, whereas the DNA in mitochondria and chloroplasts is like prokaryotic DNA, double-stranded and circular. The DNA in prokaryotes is relatively free of associated proteins, but the DNA in the nucleus of eukaryotes is associated with basic proteins, called histones. Now that the structure of DNA has been studied for over 100 years and has been accepted, procedures have been devised to isolate almost pure DNA from its other components.

The isolation of DNA is a fundamental step in molecular biology studies. The DNeasy Blood and Tissue Kit (Qiagen) is widely used for purifying high-quality DNA from animal tissues, blood, and other biological samples. The kit uses a silica membrane-based technology that allows for efficient extraction and purification of DNA, yielding material suitable for various downstream applications such as PCR, sequencing, and cloning. The process is rapid and does not involve toxic chemicals like phenol or chloroform, making it safe and reproducible.

Principle:

The DNeasy Blood and Tissue Kit operates on the principle of selective binding of DNA to a silica-based membrane in the presence of chaotropic salts. These salts disrupt hydrogen bonds and denature proteins, allowing the DNA to bind to the silica surface while contaminants are washed away. The key steps include:

1. **Cell Lysis:** The sample is lysed using proteinase K (digests proteins) and lysis buffer, which releases nucleic acids into the solution.
2. **Binding:** The lysate is mixed with a buffer containing chaotropic salts that help DNA bind to the silica membrane in the spin column.
3. **Washing:** Contaminants like proteins, lipids, salts and other cellular debris are removed by passing wash buffers through the spin column.
4. **Elution:** Pure DNA is eluted from the silica membrane (Mini Spin column) with a low-salt buffer or Milli-Q water.

Role of Components:

1. **Proteinase K**: Degrades proteins and helps in cell lysis, freeing the DNA.
2. **Buffer ATL/Buffer AL**: Helps in cell lysis and solubilization of the sample.
3. **Ethanol**: Facilitates precipitation of DNA by reducing the solubility of DNA.
4. **DNeasy Mini Spin Column**: Contains the silica membrane that binds the DNA during the purification process.
5. **Wash Buffers (AW1 and AW2)**: Used to wash away impurities like proteins, salts, and other contaminants without disturbing DNA bound to the column.
6. **Buffer AE/Elution Buffer**: A low-salt buffer that releases the DNA from the silica membrane for collection.

Requirements:

- DNeasy blood and tissue kit (Qiagen): {Components- Proteinase K, Buffer AL, AW1 buffer, AW2 buffer, Elution buffer, Spin columns}
- Clonal cultures of the ciliates (1000 cells/ml)
- Ethanol
- Milli-Q water
- Eppendorf
- Micro tips

- Pipettes
- Centrifuge
- Hand-Centrifuge
- Water Bath
- Vortex

Procedure:

1. Clonal culture of the ciliates (1000 cells/ml) was taken and filtered using 120-micron Nytex mesh to remove the cabbage from the medium.
2. Cells were then centrifuged to obtain a cell pellet with a hand-centrifuge machine.
3. Discard the supernatant and resuspend the pellet in approximately 1 ml media.
4. Transferred it in an Eppendorf and added 20 µL proteinase K to it.
5. Added 200 µL Buffer AL. Mixed thoroughly by vortexing.
6. Incubated the sample at 56°C in the water bath for 10 mins. So that the proteinase K enzyme could work optimally.
7. Added 200 µL ethanol (96-100%) and mix thoroughly by vortexing.
8. Pipetted the mixture into a DNeasy Mini spin column placed in a 2 ml collection tube. Centrifuged at 8000 rpm for 1 min. Discarded the flow through.
9. Placed the spin column in a new 2 ml collection tube, added 500 microliter Buffer AW1 (wash buffer) and centrifuged at 8000 rpm for 1 min. Discarded the flow through.
10. Placed the spin column in a new 2 ml collection tube, added 500 microliter Buffer AW2 and centrifuged for 3 min at 14,000 rpm. Discarded the flow through.
11. Transferred the spin column to a new 1.5 ml or 2 ml microcentrifuge tube.
12. Eluted the DNA by adding 100 µL Milli-Q water/Elution buffer and incubated it overnight at room temperature, then centrifuged for 1 min. at 8000 rpm.
13. Eluted DNA was subjected to Gel electrophoresis and the gel was visualized under UV-transilluminator.

Results: The quality of DNA was assessed using a spectrophotometer (A_{260}/A_{280} ratio) or gel electrophoresis. On electrophoresis, one band of micronuclear DNA near the well which indicated that the band was of high molecular weight, and a smear of macronuclear DNA was observed.

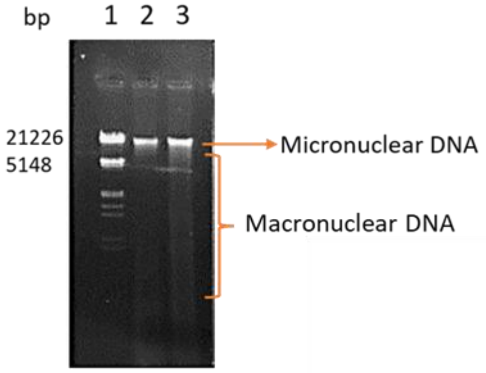


Figure 2: EtBr stained 0.8% agarose gel showing total genomic DNA

Precautions:

1. Handle all samples with care to avoid contamination. Use clean, sterile tools and workspaces.
2. Ensure the complete digestion of samples by incubating them with proteinase K for the recommended time. Incomplete digestion can lower yield and quality.

3. Prepare all buffers according to the instructions.
4. Avoid overloading the spin column with too much starting material, as this may reduce the efficiency of the DNA isolation.
5. Eluted DNA should be stored at -20°C to prevent degradation.

Experiment 5: Primer designing for the DNA template

Introduction

A primer is a short synthetic oligonucleotide which is used in many molecular techniques from PCR to DNA sequencing. These primers are designed to have a sequence which is the reverse complement of a region of template or target DNA to which we wish the primer to anneal.

Types of primers

- Universal
- Degenerate
- Sequence-specific primers

Criteria for designing the Primers

1. Primers should be 17-28 bases in length;

- Too short primer: uniqueness not ensured;
- Long length: more chances of primer being unique;
- Too long length: high melting and annealing temperature;
- But length may vary according to the requirement to attain specificity

2. Base composition should be 50-60% (G+C)

- affects hybridization specificity and melting/annealing temperature.

3. Melting Temperature (Tm) between 55-80^{o}C are preferred;

- **Tm**– the temperature at which half the DNA strands are single-stranded and half are double-stranded.
- Tm is characteristic of the DNA composition; Higher G+C content DNA has a higher Tm due to more H bonds.
- **Calculation**
  Tm= (A+T) X 2 + (G+C) X 4

4. **Annealing Temperature**
   Tanneal = Tm_primer – 4^{0}C
5. 3'-ends of primers should not be complementary (ie. base pair), as otherwise primer dimers will be synthesised preferentially to any other product;
6. Primer self-complementarity (ability to form 2^{o} structures such as hairpins) should be avoided.
7. Runs of three or more Cs or Gs at the 3'-ends of primers may promote mispriming at G or C-rich sequences (because of the stability of annealing) and should be avoided.

Software used for Designing primer

There are several online software for designing primers like

- Primer3
- Oligo
- GC
- Biotools
- Primer Blast (NCBI)

Primer3

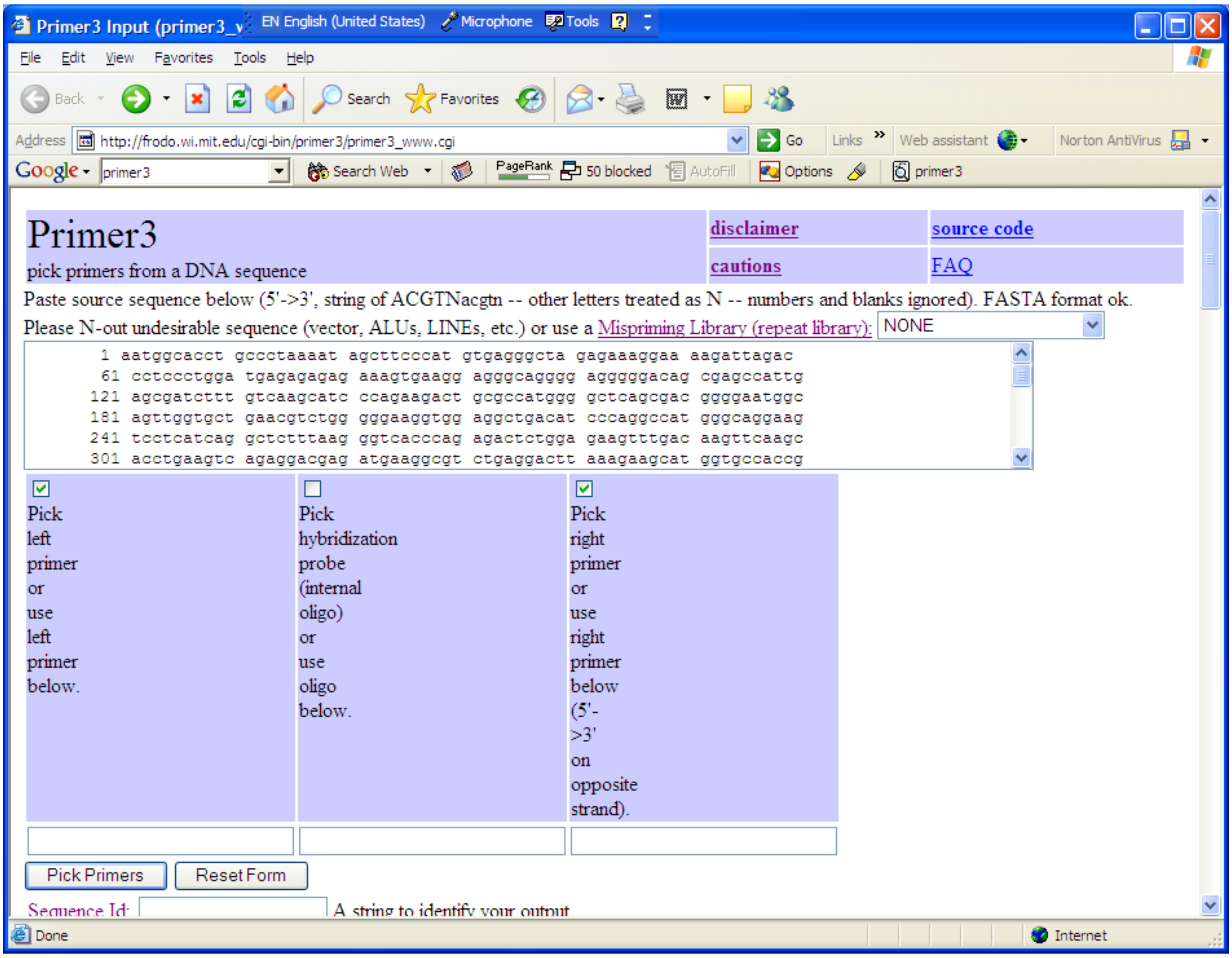
Primer3 is a free online tool for designing and analyzing primers for PCR and real-time PCR experiments. Primer3 can also select single primers for sequencing reactions and can design oligonucleotide hybridization probes. The online tool constitutes some important features like primer detection, cloning, sequencing and Primer listing

Parameters setting

- Start and end of the region to prime (in bases) **[MANDATORY]**
- **Product Size [MANDATORY]**
- Primer melting temperature (min/max/optimal) [Default]
- Primer content in GC (%, min/max) [Default]
- DNA concentration (mM) [Default]
- Salt concentration (mM) [Default]
- Number of 3' GC clamps [Default]

Compute

Once the parameters have been chosen, by pressing the "**Pick Primers**" button, then the primers are designed and the combinations found are displayed on a clickable chart.



Address http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

Google primer3

Primer3
pick primers from a DNA sequence

disclaimer | source code | cautions | FAQ

Paste source sequence below (5'->3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a Mispriming Library (repeat library): NONE

```
      1 aatggcacct gccctaaaat agcttcccat gtgagggcta gagaaaggaa aagattagac
     61 cctccctgga tgagagagag aaagtgaagg agggcagggg agggggacag cgagccattg
    121 agcgatcttt gtcaagcatc ccagaagact gcgccatggg gctcagcgac ggggaatggc
    181 agttggtgct gaacgtctgg gggaaggtgg aggctgacat cccaggccat gggcaggaag
    241 tcctcatcag gctctttaag ggtcacccag agactctgga gaagtttgac aagttcaagc
    301 acctgaagtc agaggacgag atgaaggcgt ctgaggactt aaagaagcat ggtgccaccg
```

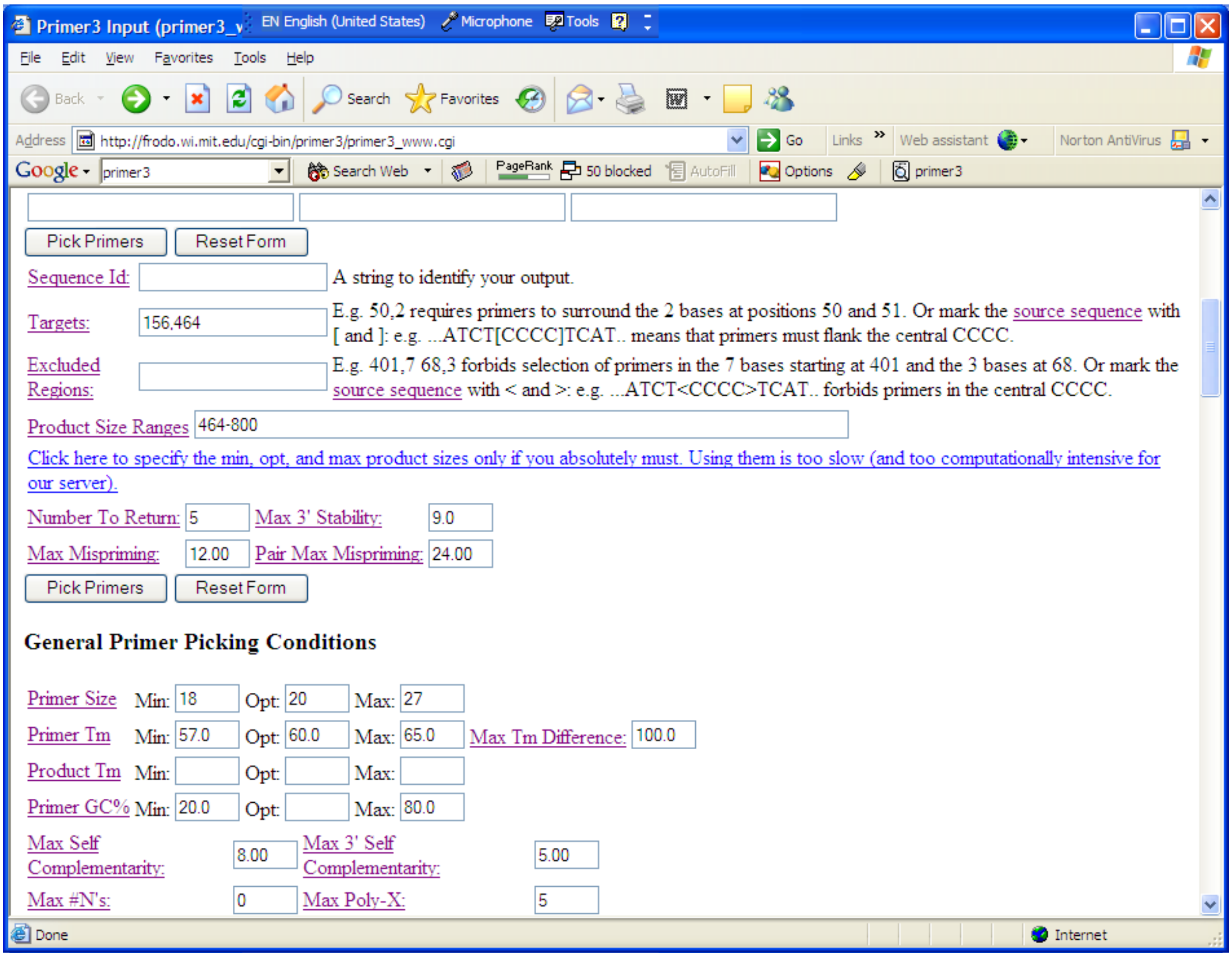
☑ Pick left primer or use left primer below.

□ Pick hybridization probe (internal oligo) or use oligo below.

☑ Pick right primer or use right primer below (5'->3' on opposite strand).

Sequence Id: A string to identify your output

http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi



Address: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi

Google: primer3

Pick Primers | Reset Form

Sequence Id: A string to identify your output.

Targets: 156,464 — E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the source sequence with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

Excluded Regions: E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the source sequence with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

Product Size Ranges: 464-800

Click here to specify the min, opt, and max product sizes only if you absolutely must. Using them is too slow (and too computationally intensive for our server).

Number To Return: 5 — Max 3' Stability: 9.0

Max Mispriming: 12.00 — Pair Max Mispriming: 24.00

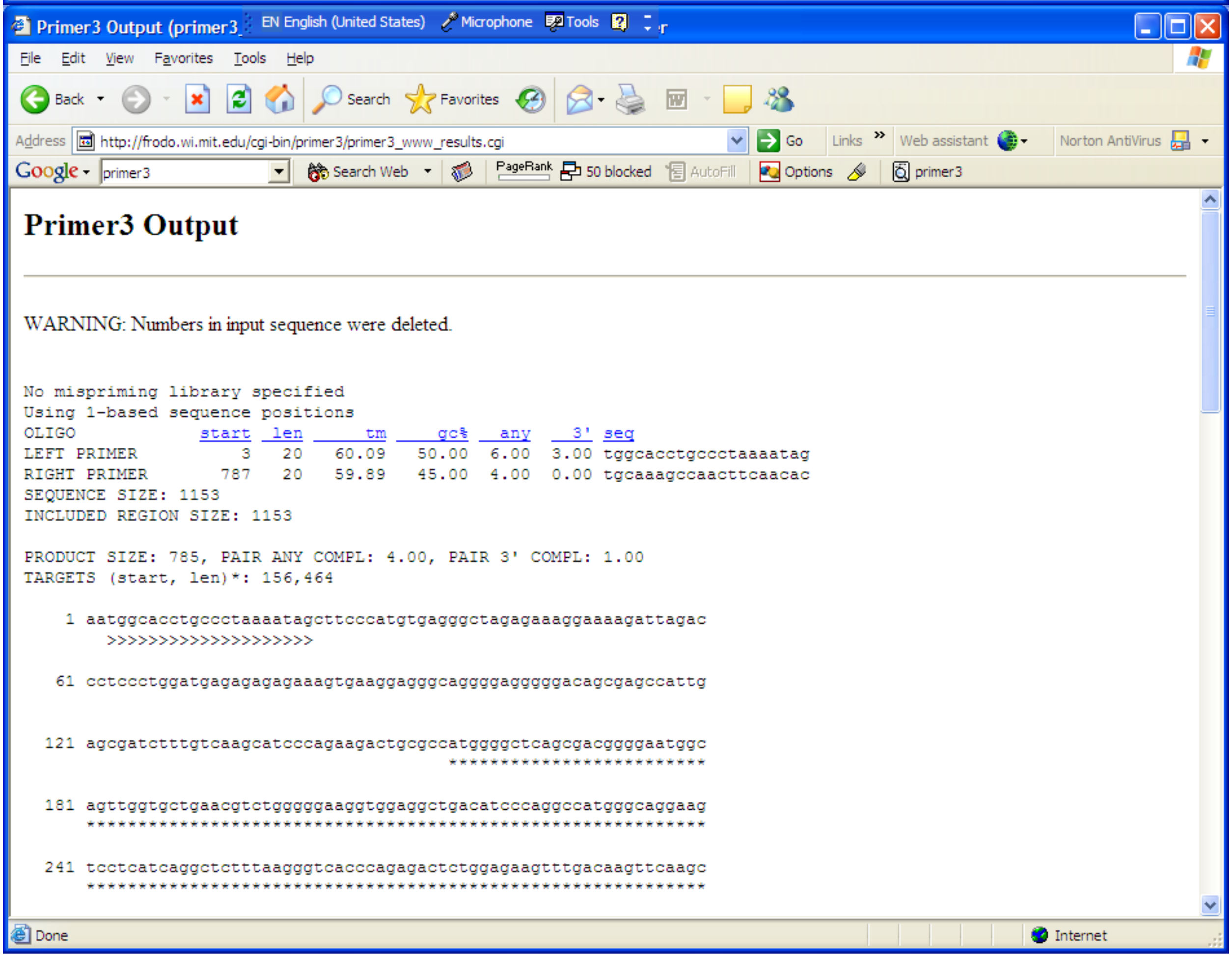
Pick Primers | Reset Form

General Primer Picking Conditions

Primer Size	Min: 18	Opt: 20	Max: 27				
Primer Tm	Min: 57.0	Opt: 60.0	Max: 65.0	Max Tm Difference: 100.0			
Product Tm	Min:	Opt:	Max:				
Primer GC%	Min: 20.0	Opt:	Max: 80.0				

Max Self Complementarity: 8.00 — Max 3' Self Complementarity: 5.00

Max #N's: 0 — Max Poly-X: 5



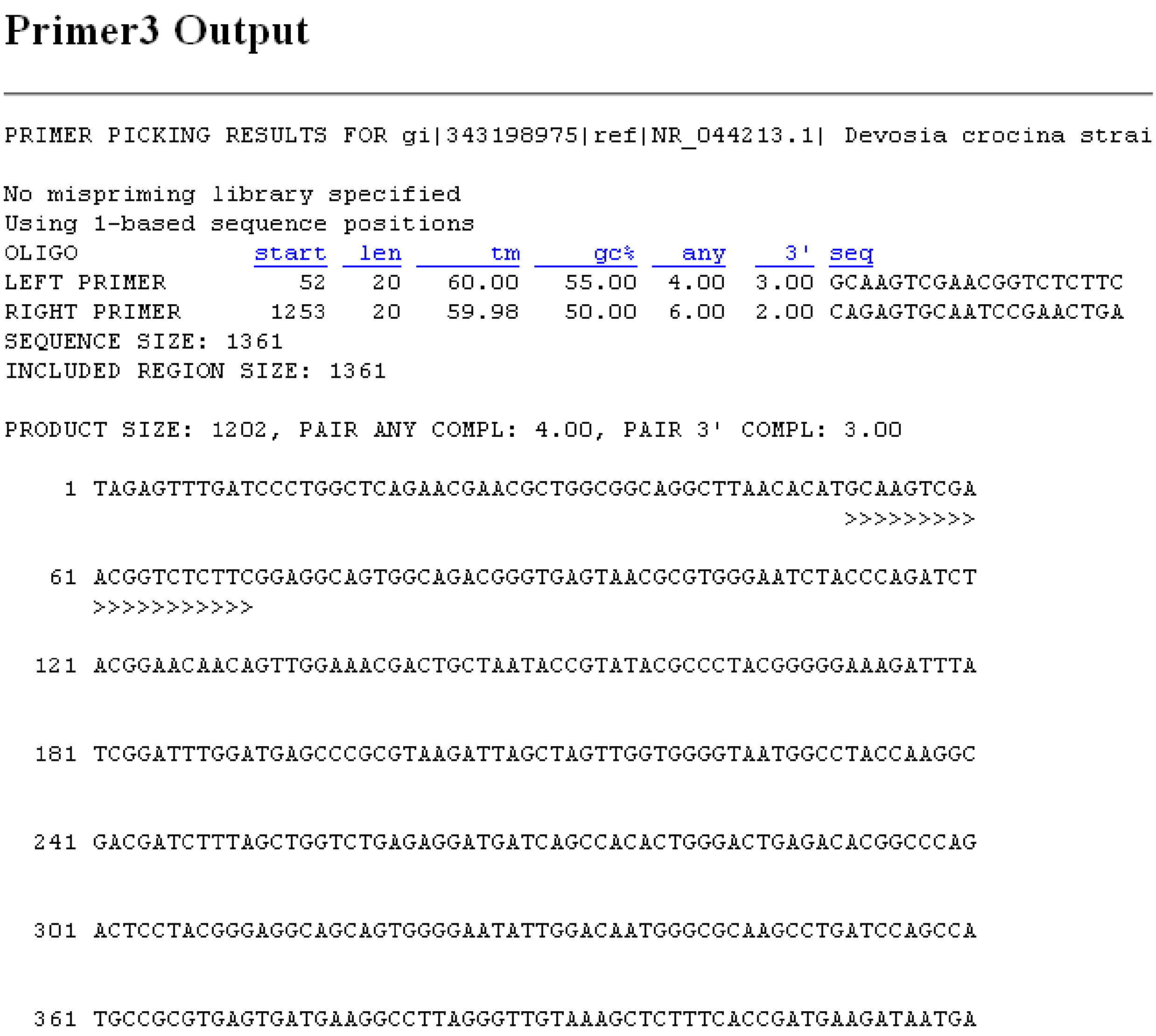
Address: http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www_results.cgi

Google: primer3

Primer3 Output

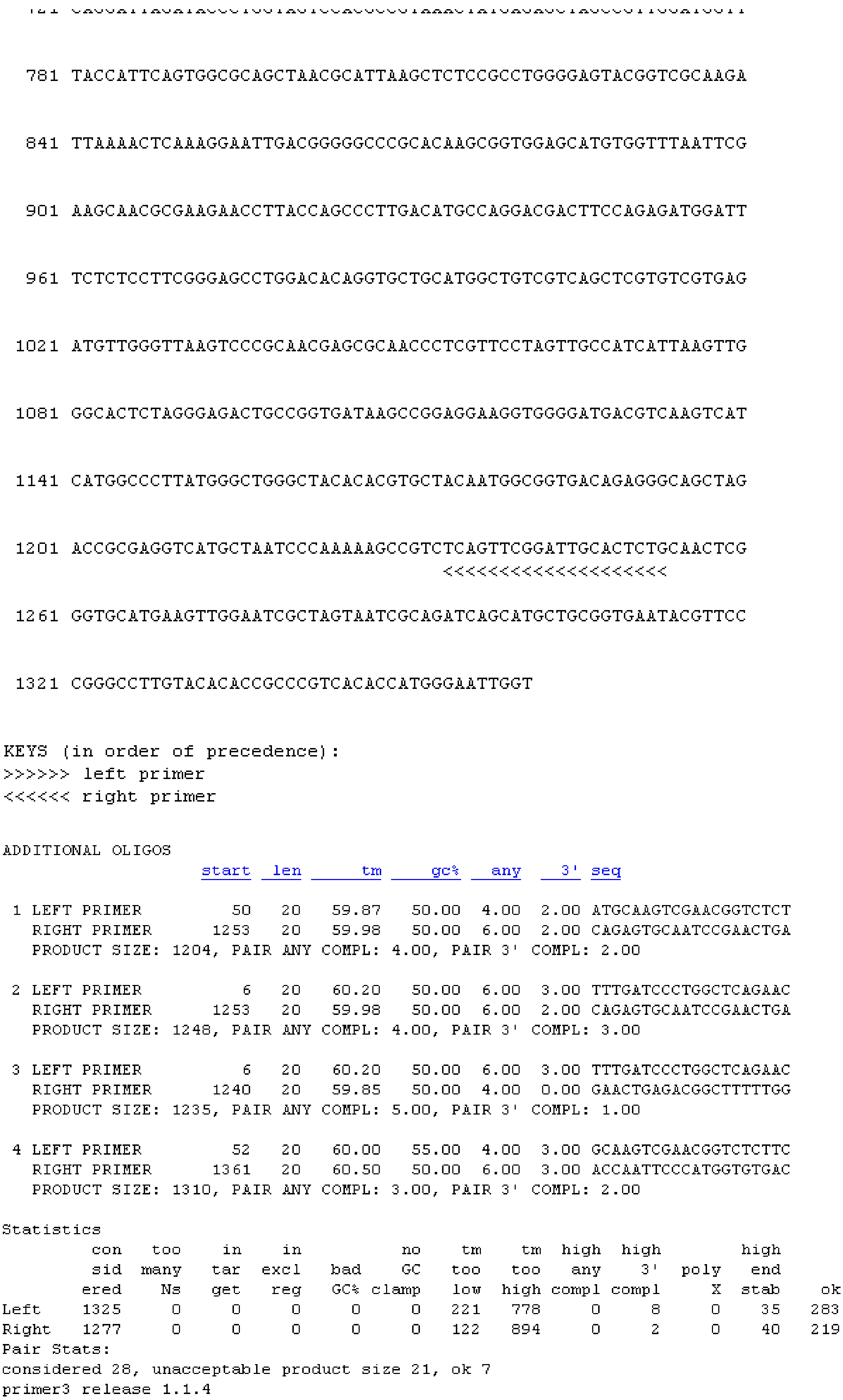
WARNING: Numbers in input sequence were deleted.

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3' seq
LEFT PRIMER          3   20   60.09   50.00  6.00  3.00 tggcacctgccctaaaatag
RIGHT PRIMER       787   20   59.89   45.00  4.00  0.00 tgcaaagccaacttcaacac
SEQUENCE SIZE: 1153
INCLUDED REGION SIZE: 1153

PRODUCT SIZE: 785, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 1.00
TARGETS (start, len)*: 156,464

    1 aatggcacctgccctaaaatagcttcccatgtgagggctagagaaaggaaaagattagac
        >>>>>>>>>>>>>>>>>>>>

   61 cctccctggatgagagagagaaagtgaaggagggcaggggagggggacagcgagccattg


  121 agcgatctttgtcaagcatcccagaagactgcgccatggggctcagcgacggggaatggc
                                         *************************

  181 agttggtgctgaacgtctgggggaaggtggaggctgacatcccaggccatgggcaggaag
      ************************************************************

  241 tcctcatcaggctctttaagggtcacccagagactctggagaagtttgacaagttcaagc
      ************************************************************
```



Primer3 Output

```
PRIMER PICKING RESULTS FOR gi|343198975|ref|NR_044213.1| Devosia crocina strai

No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm     gc%   any    3' seq
LEFT PRIMER         52   20   60.00   55.00  4.00  3.00 GCAAGTCGAACGGTCTCTTC
RIGHT PRIMER      1253   20   59.98   50.00  6.00  2.00 CAGAGTGCAATCCGAACTGA
SEQUENCE SIZE: 1361
INCLUDED REGION SIZE: 1361

PRODUCT SIZE: 1202, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 3.00

     1 TAGAGTTTGATCCCTGGCTCAGAACGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGA
                                                          >>>>>>>>>

    61 ACGGTCTCTTCGGAGGCAGTGGCAGACGGGTGAGTAACGCGTGGGAATCTACCCAGATCT
       >>>>>>>>>>>

   121 ACGGAACAACAGTTGGAAACGACTGCTAATACCGTATACGCCCTACGGGGGAAAGATTTA


   181 TCGGATTTGGATGAGCCCGCGTAAGATTAGCTAGTTGGTGGGGTAATGGCCTACCAAGGC


   241 GACGATCTTTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAG


   301 ACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCA


   361 TGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTCACCGATGAAGATAATGA
```



```
  781 TACCATTCAGTGGCGCAGCTAACGCATTAAGCTCTCCGCCTGGGGAGTACGGTCGCAAGA


  841 TTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCG


  901 AAGCAACGCGAAGAACCTTACCAGCCCTTGACATGCCAGGACGACTTCCAGAGATGGATT


  961 TCTCTCCTTCGGGAGCCTGGACACAGGTGCTGCATGGCTGTCGTCAGCTCGTGTCGTGAG


 1021 ATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTCGTTCCTAGTTGCCATCATTAAGTTG


 1081 GGCACTCTAGGGAGACTGCCGGTGATAAGCCGGAGGAAGGTGGGGATGACGTCAAGTCAT


 1141 CATGGCCCTTATGGGCTGGGCTACACACGTGCTACAATGGCGGTGACAGAGGGCAGCTAG


 1201 ACCGCGAGGTCATGCTAATCCCAAAAAGCCGTCTCAGTTCGGATTGCACTCTGCAACTCG
                                       <<<<<<<<<<<<<<<<<<<<

 1261 GGTGCATGAAGTTGGAATCGCTAGTAATCGCAGATCAGCATGCTGCGGTGAATACGTTCC


 1321 CGGGCCTTGTACACACCGCCCGTCACACCATGGGAATTGGT


KEYS (in order of precedence):
>>>>>> left primer
<<<<<< right primer

ADDITIONAL OLIGOS
                    start  len      tm     gc%    any    3' seq

 1 LEFT PRIMER         50   20   59.87   50.00   4.00   2.00 ATGCAAGTCGAACGGTCTCT
   RIGHT PRIMER      1253   20   59.98   50.00   6.00   2.00 CAGAGTGCAATCCGAACTGA
   PRODUCT SIZE: 1204, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00

 2 LEFT PRIMER          6   20   60.20   50.00   6.00   3.00 TTTGATCCCTGGCTCAGAAC
   RIGHT PRIMER      1253   20   59.98   50.00   6.00   2.00 CAGAGTGCAATCCGAACTGA
   PRODUCT SIZE: 1248, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 3.00

 3 LEFT PRIMER          6   20   60.20   50.00   6.00   3.00 TTTGATCCCTGGCTCAGAAC
   RIGHT PRIMER      1240   20   59.85   50.00   4.00   0.00 GAACTGAGACGGCTTTTTGG
   PRODUCT SIZE: 1235, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 1.00

 4 LEFT PRIMER         52   20   60.00   55.00   4.00   3.00 GCAAGTCGAACGGTCTCTTC
   RIGHT PRIMER      1361   20   60.50   50.00   6.00   3.00 ACCAATTCCCATGGTGTGAC
   PRODUCT SIZE: 1310, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00

Statistics
         con   too    in    in          no    tm    tm  high  high        high
         sid  many   tar  excl   bad    GC   too   too   any    3'  poly   end
        ered    Ns   get   reg   GC% clamp   low  high compl compl     X  stab     ok
Left    1325     0     0     0     0     0   221   778     0     8     0    35    283
Right   1277     0     0     0     0     0   122   894     0     2     0    40    219
Pair Stats:
considered 28, unacceptable product size 21, ok 7
primer3 release 1.1.4
```

Experiment 6: To amplify the 18S ribosomal gene using gene-specific primers

Principle

The polymerase chain reaction (PCR) is a technique for exponentially amplifying a specific fragment of DNA, via enzymatic replication, without using a living organism. PCR can be used for amplification of a single or few copies of a piece of DNA across several orders of magnitude, generating millions or more copies of the DNA piece. Developed in 1983 by Kary Mullis, who won the Nobel Prize for his work, PCR is now a common technique with wide applications in medical and biological research labs, such as in gene manipulation, prenatal diagnosis, DNA profiling, archaeology, paleontology, diagnosis of hereditary and infectious diseases etc.

Two primers anneal to denatured DNA template at opposite sides of the target region, and are extended by DNA polymerase to give new strands. Denaturation of DNA template, annealing of primers to the template and extension of the annealed primer constitute one cycle of reaction. After one cycle, new DNA strands of variable length are produced. In cycle 2, the original template strands and the new strands from cycle 1 are separated, yielding a total of four primer sites with which primers anneal. The primers that are hybridized to the new strands from cycle 1 are extended by the polymerase as far as the end of the template, leading to a precise copy of the target region. In cycle 3, double stranded DNA molecules are synthesized, that are precisely identical to the target region. Further cycles lead to exponential doubling of the target region.

PCR primers are short fragments of single stranded DNA (17-30 nucleotides in length) that are complementary to DNA sequences that flank the target region of interest. Primer provides 3’OH group to which the Taq DNA polymerase and synthesize a new strand, complimentary to the template strand.
Forward & Reverse primers
For a PCR reaction one set of primers is designed- forward & reverse, which are illustrated as:

```
Reverse Primer
                                      3'<-------GGAA----5'
                             Plus Strand         ||||
5'-----ATCG-------=========------------CCTT---- 3'
      ||||        Target          ||||
3'-----TAGC-------=========------------GGAA---- 5'
      ||||      minus Strand
5'--ATCG----> 3'
Forward Primer
```

Rules for designing primers

- In designing primers for PCR, the following rules are followed:
- LENGTH of individual primers should be between 17-30 nucleotides. There is a ¼ chance of finding an A, G, C or T in any given DNA sequence; 1/16 of finding any dinucleotide sequence (eg. AG); 1/256 of finding a given 4-base sequence. A sixteen base sequence will statistically be present only once in every 4,29,4967,296 bases (4 billion). Thus, the association of a greater-than-16-base oligonucleotide with its target sequence is an extremely sequence-specific process.
- MELTING TEMPERATURE (Tm). It is the temperature at which the duplex DNA is denatured to 50%. The Tm of the forward & reverse primers should be nearly the same and should be in the range of 55-80ºC. A simple formula for the calculation of Tm is:
  Tm = 4(G + C) + 2(A + T) ºC
- ANNEALING TEMPERATURE (Ta). The temperature at which the primer anneals to the template. It is required to be standardized and usually it is 5-10ºC less than the Tm of the primers.
- GC CONTENT of the primers should between 40-60%.

- The sequence of forward and reverse primers should not be complimentary to each other, as it could lead to primer-dimer formation.
- The sequence of the primers should be such that it does not lead to the formation of secondary structures (eg. hairpin loop) as it would prevent the binding of primer to the template DNA.

Components of PCR

- DNA TEMPLATE that contains the region of the DNA fragment to be amplified.
- PRIMERS (Forward and Reverse).
- DNA POLYMERASE (Taq polymerase or any DNA polymerase with a temperature optimum at around 70°C), used to synthesize a DNA copy of the region to be amplified. Taq polymerase is a thermo stable DNA polymerase widely used in current PCR practice. It is isolated from the thermophillic bacterium *Thermus aquaticus*.
- DEOXYNUCLEOTIDE TRIPHOSPHATES (dNTPs) from which the DNA polymerase builds the new strands of DNA.
- Enzyme buffer, which provides a suitable chemical environment for optimum activity and stability of the DNA polymerase.
- The PCR is carried out in small reaction tubes (0.2-0.5 ml volumes), containing a reaction volume typically of 20-50 μl that are inserted into a THERMAL CYCLER. This is a machine that heats and cools the reaction tubes within it to the precise temperature required for each step of the reaction.

STEPS OF PCR

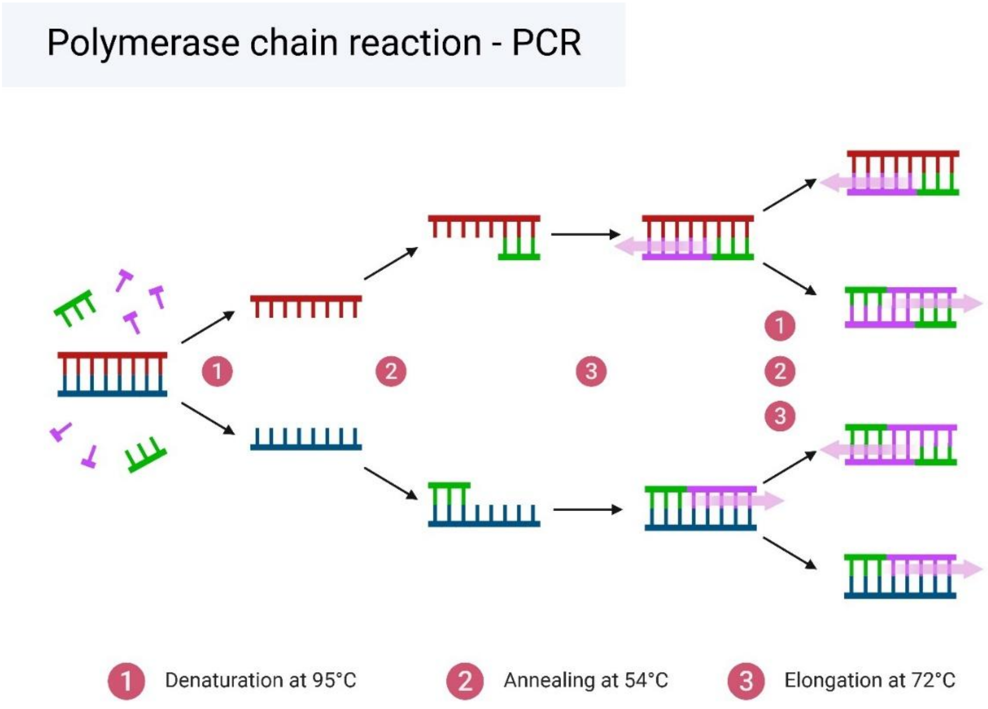
The PCR usually consists of a series of 25 to 35 cycles. Most commonly, PCR is carried out in three steps, often preceded by one temperature hold at the start and followed by one hold at the end.

1) **Primary Denaturation:**
   Prior to the first cycle, during an initialization step, the PCR reaction is often heated to a temperature of 94°C, and this temperature is then held for 5 minutes. This first hold is employed to ensure that the DNA template is completely denatured.
   Secondary Denaturation:
   Following this first hold of primary denaturation, cycles of amplification begin, with one step at 94°C for one minute.
2) **Annealing:**
   The denaturation is followed by the annealing step. In this step the reaction temperature is lowered so that the primers can anneal to the single-stranded DNA template. Brownian motion causes the primers to move around, and hydrogen bonds are constantly formed and broken between primer and template. Stable bonds are only formed when the primer sequence very closely matches the template sequence, and to this short section of double-stranded DNA the polymerase attaches. The annealing temperature (Ta) is calculated as discussed above.
3) **Extension:**
   The annealing step is followed by an extension/elongation step during which the DNA polymerase synthesizes new DNA strands complementary to the DNA template strands. The temperature at this step depends on the DNA polymerase used. Taq polymerase has a temperature optimum of 70-74°C; thus, in most cases a temperature of 72°C is used. The hydrogen bonds between the extended primer and the DNA template are now strong enough to withstand the higher temperature. Primers that have annealed to DNA regions with mismatching bases dissociate from the template and are not extended.
   Final Extension:

A final elongation step of 5-15 minutes after the last cycle may be used to ensure that any remaining single-stranded DNA is fully extended.

Final Hold:

A final hold at 4°C for an indefinite time may be employed for storage of the reaction, e.g., if reactions are run overnight.



Requirements

Chemicals

- Template DNA (25ng/µl):
- Primers: Forward primer (10 µM) & Reverse primer (10 µM)
- Mix of deoxyribonucleotides (dNTPs): 10 mM each of the four deoxyribonucleotide- dATP, dCTP, dGTP, dTTP
- Enzyme: Taq Polymerase (5 units/µl)
- Enzyme buffer (10X)
- Autoclaved distilled water

Equipments

- PCR machine
- Micropipettes

Plasticware

- Microfuge tubes
- Microtips

Procedure

SSU rRNA

Primers	Sequences	Length	Tm	GC%
Forward primer	5' CGGTAATTCCAGCTCCAATAG 3'	20 bases	59.53	50
Reverse primer	5' AACTAAGAACGGCCATGCAC 3'	20 bases	60.42	55

Collect all the reagents & materials required for setting up a PCR and place them in ice. Composition of the PCR mix for a standard PCR (50 µL) reaction is:

PCR Reaction Mix:

	Amount	**Final Concentration**
Sterile Distilled Water	51 µl	---
10XReaction buffer	10 µl	1X
dNTPs Mix	04 µl	100 µM of each Dntp
Forward primer	04 µl	20 µM
Reverse primer	04 µl	20 µM
Template DNA	25 µl	300 ng

Taq DNA polymerase	02 µl	6U
TOTAL	**100 µl**	

As described above, prepare one more PCR mix that shall serve as a negative control. The DNA template is not added to the mix; instead, an equivalent volume of autoclaved water is added to keep the final volume of the mix constant.

After preparation of the cocktail, place PCR tubes in a thermal cycler. Programme the machine as follows:

Cycle	**Denaturation**	**Annealing**	**Extension**
First	95°C for 5 min.	54°C for 1 min.	72°C for 1 min.
Subsequent	95°C for 45 sec.	54°C for 45 sec.	72°C for 45 secs.
Last	95°C for 45 sec.	54 C for 45 sec.	72°C for 10 min.

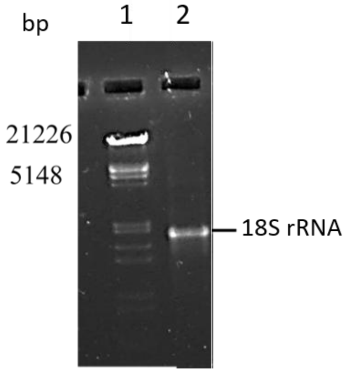
Once the machine's temperature comes to 4°C, take out the tubes and place them in ice. Take an aliquot (3-5-µL) of the PCR product from the tube and analyze it on agarose gel (1%). Alongside load a 100 bp DNA marker.

Observation

A single DNA band of 1700 bp (base pair) was observed on agarose gel [size was estimated by comparing the position of the DNA band with that of the corresponding band in the DNA marker (100 bp ladder)].

Result

PCR amplification of SSU rRNA gene was successfully done.



Discussion

PCR product of the expected size was observed. Presence of a single sharp DNA band and absence of smear indicates that the non-specific annealing of the primers did not take place.

Precautions

- Keep all the reagents and PCR tubes in ice at the time of preparation of PCR mix.
- Frequent freeze-thaw of the reagents should be avoided. Store all the stock solutions as aliquots at -20°C deep freezers.
- All the plasticware (micropipettes tips, PCR tubes) should be autoclaved.
- Before putting up the PCR reaction, test the quality of the template DNA by agarose gel electrophoresis.
- Ensure that all the components of the PCR mix are added to the tube.
- Every PCR reaction should include a negative control to ensure that reagents are not contaminated with any DNA template, which could result in non-specific amplification.

Flow Chart for PCR Reaction

Add the following components of PCR Reaction Mix to the PCR Tube

Milli Q	9µl
Buffer	2.5µl
dNTPs	1µl
Forward Primer	1µl
Reverse Primer	1µl
Template DNA	10µl
Taq Polymerase	0.5µl
Total	**25µl**

↓

Vortex mix the PCR Reaction Mix properly

↓

Switch on the PCR Machine

↓

Select “Palm Cycler”

↓

Set the following conditions in the PCR Machine

Cycle	**Denaturation**	**Annealing**	**Extension**
First	95°C for 5 min.	54°C for 1 min.	72°C for 1 min.
Subsequent	95°C for 45 sec.	54°C for 45 sec.	72°C for 45 sec.
Last	95°C for 45 sec.	54°C for 45 sec.	72°C for 10 min.

↓

Select “Start” option from ‘‘Control”

↓

Select “Exit” from “File” once your reaction is complete

↓

Run your PCR Product on 0.8% Agarose gel

Experiment 7: Construction of phylogenetic trees with the help of bioinformatics tools (Clustal X, MEGA, NJ) and its interpretation.

Introduction:

Phylogenetic Systematics

Phylogenetic systematics is a field of biology that deals with identifying and understanding the evolutionary relationships among the many different kinds of life on earth, both living (extant) and dead (extinct). Evolutionary theory states that similarity among individuals or species is attributable to common descent or inheritance from a common ancestor. Thus, the relationships established by phylogenetic systematics often describe a species' evolutionary history and, hence, its phylogeny, the historical relationships among lineages or organisms or their parts, such as their genes.

Traits used in Phylogenetics

- Morphological and developmental
  - –the importance of Fossils
- Molecular
  - –Protein sequences
  - –Genetic markers
  - –DNA sequences

Molecular Phylogenetics

Phylogenetics can be studied in various ways. It is often studied using fossil records, which contain morphological information about ancestors of current species and the timeline of divergence. However, fossil records have many limitations; they may be available only for certain species. Existing fossil data can be fragmentary and their collection is often limited by abundance, habitat, geographic range, and other factors. The descriptions of morphological traits are often ambiguous, which are due to multiple genetic factors. Thus, using fossil records to determine phylogenetic relationships can often be biased. For microorganisms, fossils are essentially nonexistent, which makes it impossible to study phylogeny with this approach. Fortunately, molecular data that are in the form of DNA or protein sequences can also provide very useful evolutionary perspectives of existing organisms because, as organisms evolve, the genetic materials accumulate mutations over time causing phenotypic changes. Because genes are the medium for recording the accumulated mutations, they can serve as *molecular fossils.* Through comparative analysis of the molecular fossils from a number of related organisms, the evolutionary history of the genes and even the organisms can be revealed.

The advantage of using molecular data is obvious.

(1) Molecular data are more numerous than fossil records and easier to obtain.
(2) There is no sampling bias involved, which helps to mend the gaps in real fossil records.
(3) More clear-cut and robust phylogenetic trees can be constructed with the molecular data.

Therefore, they have become favorite and sometimes the only information available for researchers to reconstruct evolutionary history. The advent of the genomic era with tremendous amounts of molecular sequence data has led to the rapid development of molecular phylogenetics. The field of molecular phylogenetics can be defined as the study of evolutionary relationships of genes and other biological macromolecules by analyzing mutations at various positions in their sequences and developing hypotheses about the evolutionary relatedness of the biomolecules. Based on the sequence similarity of the molecules, evolutionary relationships between the organisms can often be inferred.

PROCEDURE

Molecular phylogenetic tree construction can be divided into five steps:

Choosing molecular markers; Retrival, Notepad File, BLAST

↓

Performing multiple sequence alignment; Software Clustal X

↓

Choosing a model of evolution; Jukes Canter Model and Kimura Model

↓

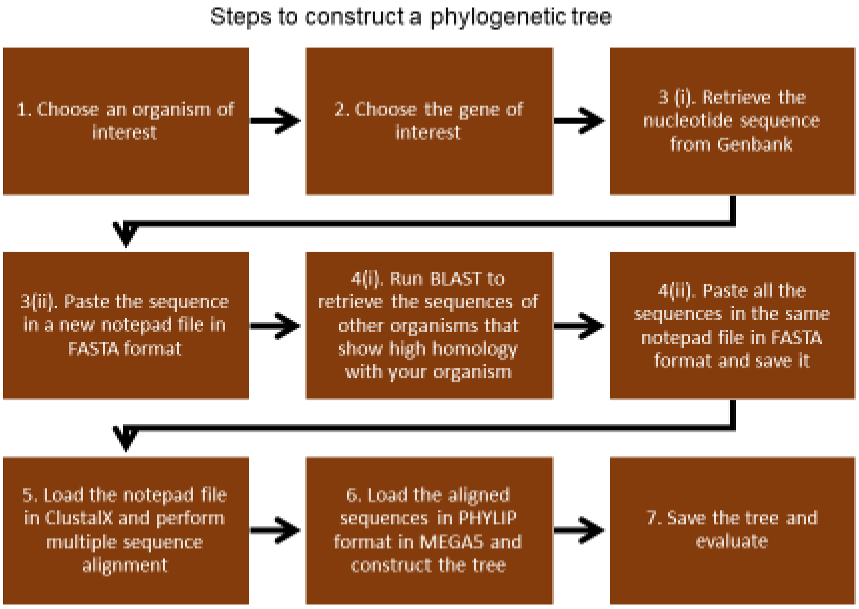
Determining a tree building method; Neighbor Joining Method, Maximum Likelihood Method, Parsimony Method

↓

Assessing tree reliability: Bootstrap

↓

Tree evaluation


Steps to construct a phylogenetic tree
1. Choose an organism of interest
2. Choose the gene of interest
3 (i). Retrieve the nucleotide sequence from Genbank
3(ii). Paste the sequence in a new notepad file in FASTA format
4(i). Run BLAST to retrieve the sequences of other organisms that show high homology with your organism
4(ii). Paste all the sequences in the same notepad file in FASTA format and save it
5. Load the notepad file in ClustalX and perform multiple sequence alignment
6. Load the aligned sequences in PHYLIP format in MEGA5 and construct the tree
7. Save the tree and evaluate

Step 1: Choice of Molecular Markers (Query Sequence)

For constructing molecular phylogenetic trees, one can use either nucleotide or protein sequence data. The choice of molecular markers is an important matter because it can make a major difference in obtaining a correct tree. The decision to use nucleotide or protein sequences depends on the properties of the sequences and the purposes of the study. For studying very closely related organisms, nucleotide sequences, which evolve more rapidly than proteins, can be used. For example, for evolutionary analysis of different individuals within a population, non-coding regions of mitochondrial DNA are often used. For studying the evolution of more widely divergent groups of organisms, one may choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences. If the phylogenetic relationships to be delineated are at the deepest level, such as between bacteria and eukaryotes, using conserved protein sequences makes more sense than using nucleotide sequences. The reason is explained in more detail next.

Retrieval of molecular markers from Biological Databases (bioinformatics tools)

Bioinformatics is an interdisciplinary research area at the interface between computer science and biological science. A variety of definitions exist in the literature and on the world wide web (www); some are more inclusive than others. Here, we adopt the definition proposed by Luscombe et al. in defining bioinformatics as a union of biology and informatics: bioinformatics involves the technology that uses computers for storage, retrieval, manipulation, and distribution of information related to biological macromolecules such as DNA, RNA, and proteins. The emphasis here is on the use of computers because most of the tasks in genomic data analysis are highly repetitive or mathematically complex. The use of computers is absolutely indispensable in mining genomes for information gathering and knowledge building.

DATABASES

Databases are fundamental to modern biological research, especially to genomic studies. The goal of a biological database is two-fold: information retrieval and knowledge discovery.

WHAT IS A DATABASE? A database is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria. Databases are composed of computer hardware and software for data management. The chief objective of the development of a database is to organize data in a set of structured records to enable easy retrieval of information. Each record, also called an **entry**, should contain a number of fields that hold the actual data items, for example, fields for names, phone numbers, addresses, dates. To retrieve a particular record from the database, a user can specify a particular piece of information, called value, to be found in a particular field and expect the computer to retrieve the whole data record. This process is called making a **query**.

Biological databases can be roughly divided into three categories: **primary databases, secondary databases, and specialized databases**. Primary databases contain original biological data. They are archives of raw sequence or structural data submitted by the scientific community. GenBank and Protein Data Bank (PDB) are examples of primary databases. Secondary databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR) (successor of Margaret Dayhoff's Atlas of Protein Sequence and Structure). Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

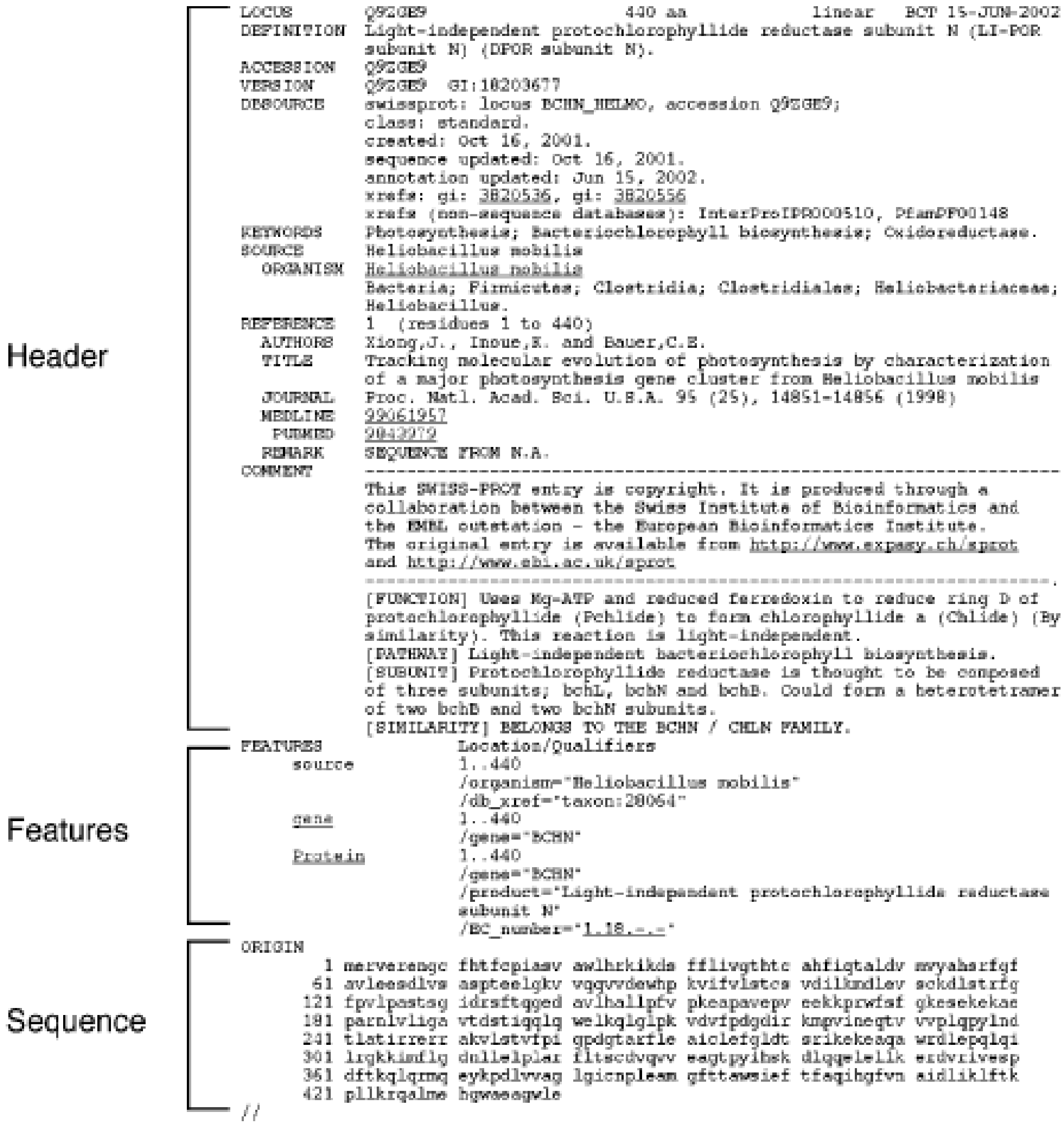
Table. Major Biological Databases Available via the World Wide Web.

Databases and Retrieval Systems	**Brief summary of Content**	**URL**
DDBJ	Primary nucleotide sequence database in Japan	www.ddbj.nig.ac.jp
EMBL	Primary nucleotide sequence database in Europe	www.ebi.ac.uk/embl/index.html
Entrez	NCBI portal for a variety of biological databases	www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
GenBank	Primary nucleotide sequence database in NCBI	www.ncbi.nlm.nih.gov/Genbank
OMIM	Genetic information of human diseases	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
PubMed	Biomedical literature	www.ncbi.nlm.nih.gov/PubMed

	information	
Ribosomal database project	Ribosomal RNA sequences and phylogenetic trees derived from the sequences	http://rdp.cme.msu.edu/html
SWISS-Prot	Curated protein sequence database	www.ebi.ac.uk/swissprot/access.html

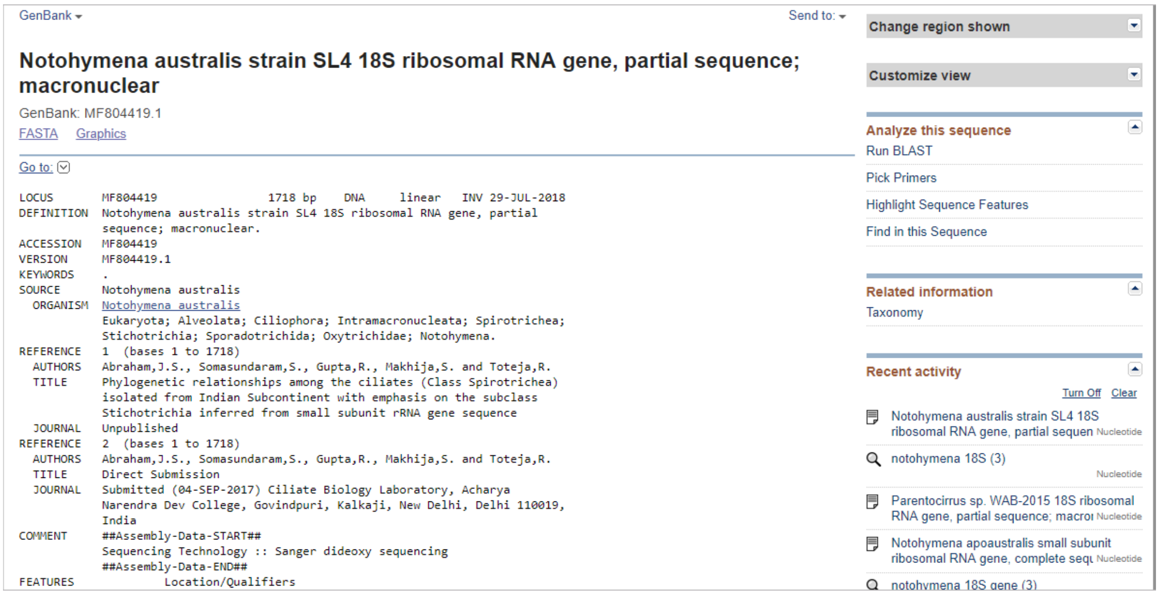
GenBank Sequence Format

To search GenBank effectively using the text-based method requires an understanding of the GenBank sequence format. GenBank is a relational database. However, the search output for sequence files is produced as flat files for easy reading. The resulting flat files contain three sections – Header, Features, and Sequence entry. There are many fields in the Header and Features sections. Each field has a unique identifier for easy indexing by computer software. Understanding the structure of the GenBank files helps in designing effective search strategies.



```
LOCUS       Q9ZGE9                  440 aa            linear   BCT 15-JUN-2002
DEFINITION  Light-independent protochlorophyllide reductase subunit N (LI-POR
            subunit N) (DPOR subunit N).
ACCESSION   Q9ZGE9
VERSION     Q9ZGE9  GI:10203677
DBSOURCE    swissprot: locus BCHN_HELMO, accession Q9ZGE9;
            class: standard.
            created: Oct 16, 2001.
            sequence updated: Oct 16, 2001.
            annotation updated: Jun 15, 2002.
            xrefs: gi: 3820536, gi: 3820556
            xrefs (non-sequence databases): InterProIPR000510, PfamPF00148
KEYWORDS    Photosynthesis; Bacteriochlorophyll biosynthesis; Oxidoreductase.
SOURCE      Heliobacillus mobilis
  ORGANISM  Heliobacillus mobilis
            Bacteria; Firmicutes; Clostridia; Clostridiales; Heliobacteriaceae;
            Heliobacillus.
REFERENCE   1  (residues 1 to 440)
  AUTHORS   Xiong,J., Inoue,K. and Bauer,C.E.
  TITLE     Tracking molecular evolution of photosynthesis by characterization
            of a major photosynthesis gene cluster from Heliobacillus mobilis
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 95 (25), 14851-14856 (1998)
  MEDLINE   99061957
   PUBMED   9843979
  REMARK    SEQUENCE FROM N.A.
COMMENT     -------------------------------------------------------------------
            This SWISS-PROT entry is copyright. It is produced through a
            collaboration between the Swiss Institute of Bioinformatics and
            the EMBL outstation - the European Bioinformatics Institute.
            The original entry is available from http://www.expasy.ch/sprot
            and http://www.ebi.ac.uk/sprot
            -------------------------------------------------------------------.
            [FUNCTION] Uses Mg-ATP and reduced ferredoxin to reduce ring D of
            protochlorophyllide (Pchlide) to form chlorophyllide a (Chlide) (By
            similarity). This reaction is light-independent.
            [PATHWAY] Light-independent bacteriochlorophyll biosynthesis.
            [SUBUNIT] Protochlorophyllide reductase is thought to be composed
            of three subunits; bchL, bchN and bchB. Could form a heterotetramer
            of two bchB and two bchN subunits.
            [SIMILARITY] BELONGS TO THE BCHN / CHLN FAMILY.
FEATURES             Location/Qualifiers
     source          1..440
                     /organism="Heliobacillus mobilis"
                     /db_xref="taxon:28064"
     gene            1..440
                     /gene="BCHN"
     Protein         1..440
                     /gene="BCHN"
                     /product="Light-independent protochlorophyllide reductase
                     subunit N"
                     /EC_number="1.18.-.-"
ORIGIN
        1 merverengc fhtfcpiasv awlhrkikda fflivgthtc ahfiqtaldv mvyahsrfgf
       61 avleesdlvs aspteelgkv vqgvvdewhp kvifvlstcs vdilkmdlev sckdlstrfg
      121 fpvlpastsg idrsftqged avlhallpfv pkeapavepv eekkprwfsf gkesekekae
      181 parnlvliga vtdstiqqlq welkqlglpk vdvfpdgdir kmpvineqtv vvplqpylnd
      241 tlatirrerr akvlstvfpi gpdgtarfle aiclefgldt srikekeaqa wrdlepqlqi
      301 lrgkkimflg dullelplar fltscdvqvv eagtpyihsk dlqqelellk erdvrivesp
      361 dftkqlqrmq eykpdlvvag lgicnpleam gfttawsief tfaqihgfvn aidliklftk
      421 pllkrqalme hgwaeagwle
//
```

Figure 2.3: NCBI GenBank/GenPept format showing the three major components of a sequence file.


GenBank
Send to:
Change region shown
Customize view
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence
Related information
Taxonomy
Recent activity
Turn Off
Clear
Notohymena australis strain SL4 18S ribosomal RNA gene, partial sequen Nucleotide
notohymena 18S (3) Nucleotide
Parentocirrus sp. WAB-2015 18S ribosomal RNA gene, partial sequence; macroı Nucleotide
Notohymena apoaustralis small subunit ribosomal RNA gene, complete sequ Nucleotide
notohymena 18S gene (3)

Notohymena australis strain SL4 18S ribosomal RNA gene, partial sequence; macronuclear

GenBank: MF804419.1

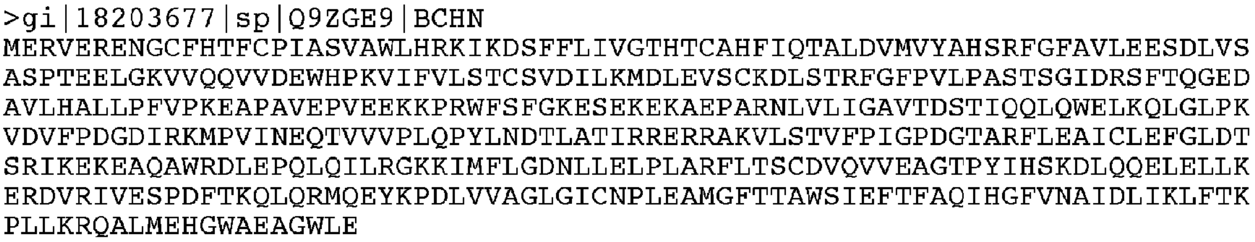
FASTA Graphics

Go to:

```
LOCUS       MF804419                1718 bp    DNA     linear   INV 29-JUL-2018
DEFINITION  Notohymena australis strain SL4 18S ribosomal RNA gene, partial
            sequence; macronuclear.
ACCESSION   MF804419
VERSION     MF804419.1
KEYWORDS    .
SOURCE      Notohymena australis
  ORGANISM  Notohymena australis
            Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Spirotrichea;
            Stichotrichia; Sporadotrichida; Oxytrichidae; Notohymena.
REFERENCE   1  (bases 1 to 1718)
  AUTHORS   Abraham,J.S., Somasundaram,S., Gupta,R., Makhija,S. and Toteja,R.
  TITLE     Phylogenetic relationships among the ciliates (Class Spirotrichea)
            isolated from Indian Subcontinent with emphasis on the subclass
            Stichotrichia inferred from small subunit rRNA gene sequence
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1718)
  AUTHORS   Abraham,J.S., Somasundaram,S., Gupta,R., Makhija,S. and Toteja,R.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-SEP-2017) Ciliate Biology Laboratory, Acharya
            Narendra Dev College, Govindpuri, Kalkaji, New Delhi, Delhi 110019,
            India
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
FEATURES             Location/Qualifiers
```

Alternative Sequence Formats

FASTA. In addition to the GenBank format, there are many other sequence formats.
FASTA is one of the simplest and the most popular sequence formats because it contains plain sequence information that is readable by many bioinformatics analysis programs. It has a single definition line that begins with a right angle bracket (>) followed by a sequence name (Fig. 2.4). Sometimes, extra information such as gi number or comments can be given, which are separated from the sequence name by a "|" symbol. The extra information is considered optional and is ignored by sequence analysis programs. The plain sequence in standard one-letter symbols starts
in the second line. Each line of sequence data is limited to sixty to eighty characters in width. The drawback of this format is that much annotation information is lost.



```
>gi|18203677|sp|Q9ZGE9|BCHN
MERVERENGCFHTFCPIASVAWLHRKIKDSFFLIVGTHTCAHFIQTALDVMVYAHSRFGFAVLEESDLVS
ASPTEELGKVVQQVVDEWHPKVIFVLSTCSVDILKMDLEVSCKDLSTRFGFPVLPASTSGIDRSFTQGED
AVLHALLPFVPKEAPAVEPVEEKKPRWFSFGKESEKEKAEPARNLVLIGAVTDSTIQQLQWELKQLGLPK
VDVFPDGDIRKMPVINEQTVVVPLQPYLNDTLATIRRERRAKVLSTVFPIGPDGTARFLEAICLEFGLDT
SRIKEKEAQAWRDLEPQLQILRGKKIMFLGDNLLELPLARFLTSCDVQVVEAGTPYIHSKDLQQELELLK
ERDVRIVESPDFTKQLQRMQEYKPDLVVAGLGICNPLEAMGFTTAWSIEFTFAQIHGFVNAIDLIKLFTK
PLLKRQALMEHGWAEAGWLE
```

Step 2: Sequence Alignment

The second step in phylogenetic analysis is to construct sequence alignment. This is probably the most critical step in the procedure because it establishes positional correspondence in evolution. Only the correct alignment produces correct phylogenetic inference because aligned positions are assumed to be genealogically related. Incorrect alignment leads to systematic errors in the final tree or even a completely wrong tree. For that reason, it is essential that the sequences are correctly aligned.

Manual editing is often critical in ensuring alignment quality. However, there is no firm rule on how to modify a sequence alignment. As a general guideline, a correct alignment should ensure the matching of key cofactor residues and residues of similar physicochemical properties. If secondary structure elements are known or can be predicted, they can serve to guide the alignment.

Alignment Methods

Pairwise Alignment

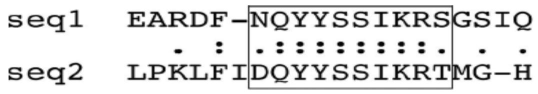
Pairwise sequence alignment is the fundamental component of many bioinformatics applications. It is extremely useful in structural, functional, and evolutionary analyses of sequences. Pairwise sequence alignment provides inference for the relatedness of two sequences. Strongly similar sequences are often homologous. However, a distinction needs to be made between **sequence homology and similarity**. The former is the inference drawn from sequence comparison, whereas the latter relates to actual observation after sequence alignment.

There are two sequence alignment strategies, **local alignment and global alignment**, and three types of algorithms that perform both local and global alignments. They are the **dot matrix method, dynamic programming method, and word method.**

Global Alignment and Local Alignment

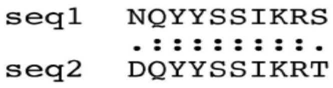
In *global alignment*, two sequences to be aligned are assumed to be generally similar over their entire length. Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences. This method is more applicable for aligning two closely related sequences of roughly the same length. For divergent sequences and sequences of variable lengths, this method may not be able to generate optimal results because it fails to recognize highly similar local regions between the two sequences.

Local alignment, on the other hand, does not assume that the two sequences in question have similarity over the entire length. It only finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions. This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences. The two sequences to be aligned can be of different lengths. This approach is more appropriate for aligning divergent biological sequences containing only modules that are similar, which are referred to as *domains*
or *motifs*.



```
seq1   EARDF-NQYYSSIKRSGSIQ
         . : .::::::::. . .
seq2   LPKLFIDQYYSSIKRTMG-H
```

global sequence alignment



```
seq1   NQYYSSIKRS
       .::::::::.
seq2   DQYYSSIKRT
```

local sequence alignment

1. Query: MRDPYNKLIS

2. Scan every three residues to be used in searching BLAST word database.

3. Assuming one of the words finds matches in the database.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...

4. Calculate sums of match scores based on BLOSUM62 matrix.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...
Sum of score	20	16	10	10	...

5. Find the database sequence corresponding to the best word match and extend alignment in both directions.

Query	M	R	D	PYN	K	L	I	S
Database	M	H	E	PYN	D	V	P	W

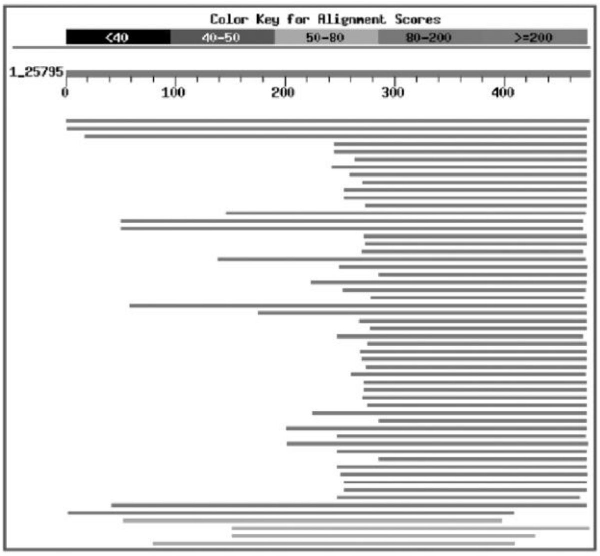
extension to left ← → extension to right

6. Determine high scored segment above threshold (22).

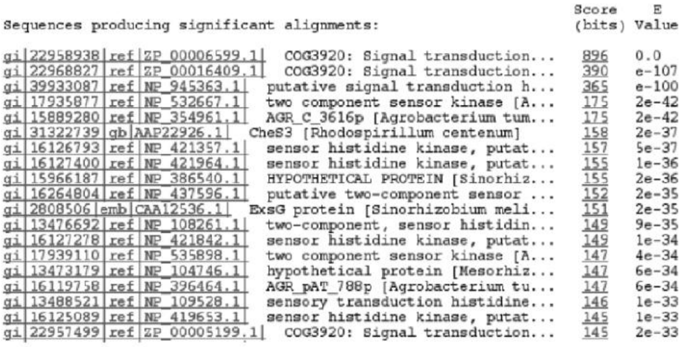
Query	M	R	D	PYN	K	L	I	S
Database	M	H	E	PYN	D	V	P	W
	5	0	2	20	−1	1	−3	−3

HSP, total score 24

Graphical overview



Matching list



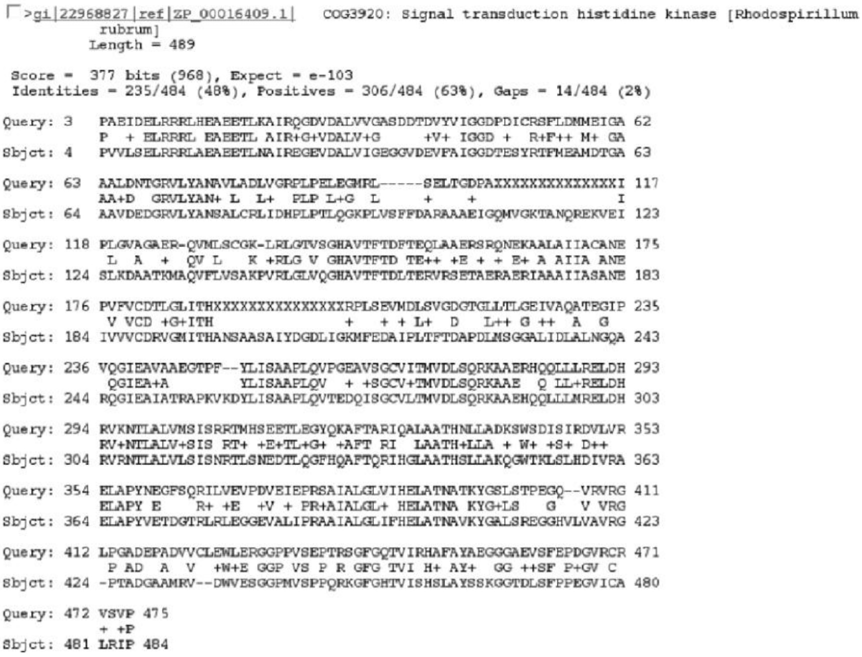
```
                                                                  Score    E
Sequences producing significant alignments:                      (bits) Value

gi|22958938|ref|ZP_00006599.1|  COG3920: Signal transduction...   896   0.0
gi|22968827|ref|ZP_00016409.1|  COG3920: Signal transduction...   390   e-107
gi|39933087|ref|NP_945363.1|  putative signal transduction h...   365   e-100
gi|17935877|ref|NP_532667.1|  two component sensor kinase [A...   175   2e-42
gi|15889280|ref|NP_354961.1|  AGR_C_3616p [Agrobacterium tum...   175   2e-42
gi|31322739|gb|AAP22926.1|  CheS3 [Rhodospirillum centenum]       158   2e-37
gi|16126793|ref|NP_421357.1|  sensor histidine kinase, putat...   157   5e-37
gi|16127400|ref|NP_421964.1|  sensor histidine kinase, putat...   155   1e-36
gi|15966187|ref|NP_386540.1|  HYPOTHETICAL PROTEIN [Sinorhiz...   155   2e-36
gi|16264804|ref|NP_437596.1|  putative two-component sensor ...   152   2e-35
gi|2808506|emb|CAA12536.1|  ExsG protein [Sinorhizobium meli...   151   2e-35
gi|13476692|ref|NP_108261.1|  two-component, sensor histidin...   149   9e-35
gi|16127278|ref|NP_421842.1|  sensor histidine kinase, putat...   149   1e-34
gi|17939110|ref|NP_535898.1|  two component sensor kinase [A...   147   4e-34
gi|13473179|ref|NP_104746.1|  hypothetical protein [Mesorhiz...   147   6e-34
gi|16119758|ref|NP_396464.1|  AGR_pAT_788p [Agrobacterium tu...   147   6e-34
gi|13488521|ref|NP_109528.1|  sensory transduction histidine...   146   1e-33
gi|16125089|ref|NP_419653.1|  sensor histidine kinase, putat...   145   1e-33
gi|22957499|ref|ZP_00005199.1|  COG3920: Signal transduction...   145   2e-33
```

Alignment output

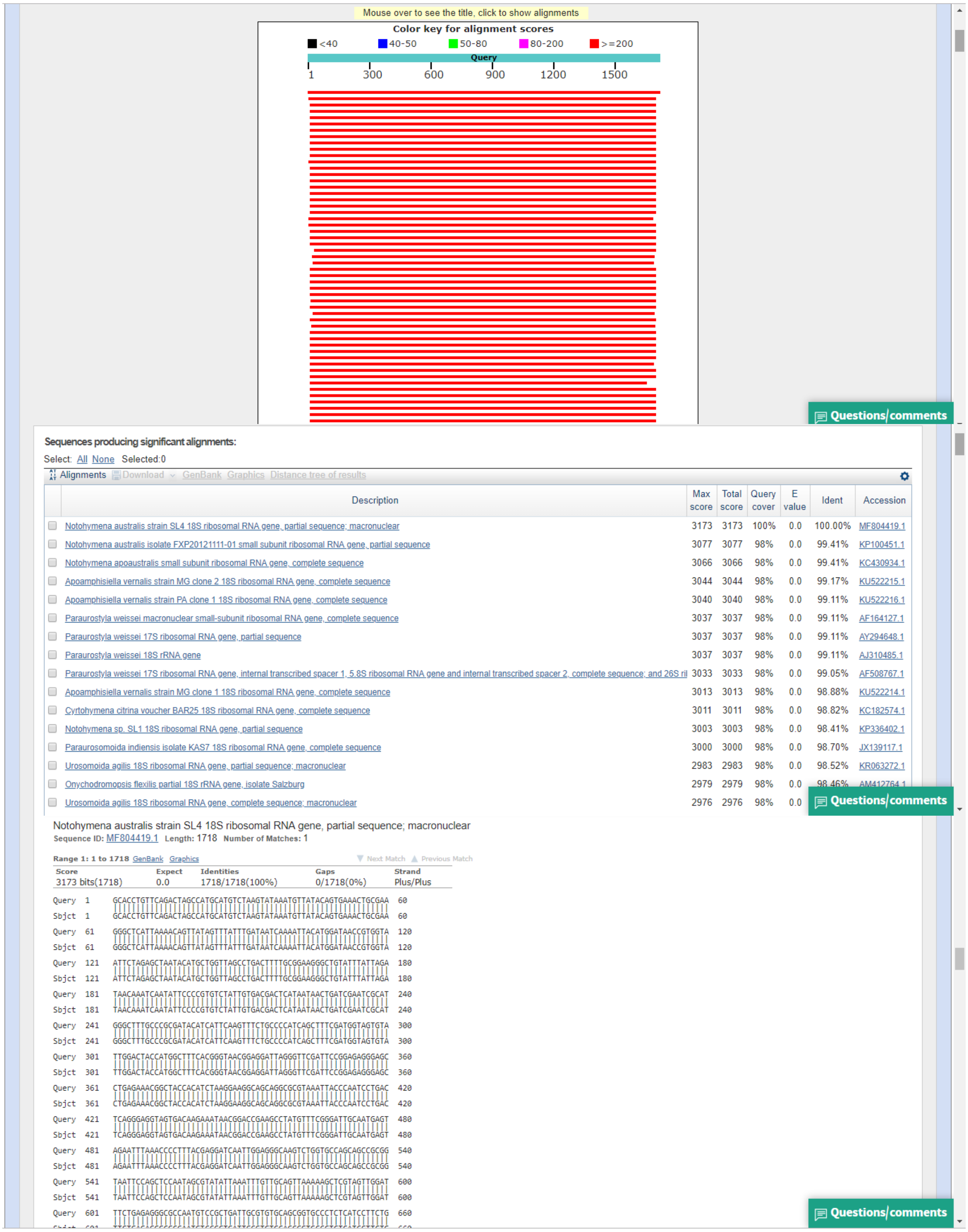
header

statistics

alignment



```
>gi|22968827|ref|ZP_00016409.1|   COG3920: Signal transduction histidine kinase [Rhodospirillum
          rubrum]
         Length = 489

 Score =  377 bits (968), Expect = e-103
 Identities = 235/484 (48%), Positives = 306/484 (63%), Gaps = 14/484 (2%)

Query: 3   PAEIDELRRRLHEAEETLKAIRQGDVDALVVGASDDTDVYVIGGDPDICRSFLDMMEIGA 62
           P  + ELRRRL EAEETL AIR+G+VDALV+G     +V+ IGGD +  R+F++ M+ GA
Sbjct: 4   PVVLSELRRRLAEAEETLNAIREGEVDALVIGEGGVDEVFAIGGDTESYRTFMEAMDTGA 63

Query: 63  AALDNTGRVLYANAVLADLVGRPLPELEGMRL-----SELTGDPAXXXXXXXXXXXXXXI 117
           AA+D  GRVLYAN+ L  L+  PLP L+G  L     +    +                I
Sbjct: 64  AAVDEDGRVLYANSALCRLIDHPLPTLQGKPLVSFFDARAAAEIGQMVGKTANQREKVEI 123

Query: 118 PLGVAGAER-QVMLSCGK-LRLGTVSGHAVTFTDFTEQLAAERSRQNEKAALAIIACANE 175
            L  A  +  QV L   K +RLG V GHAVTFTD TE++ +E + + E+ A AIIA ANE
Sbjct: 124 SLKDAATKMAQVFLVSAKPVRLGLVQGHAVTFTDLTERVRSETAERAERIAAAIIASANE 183

Query: 176 PVFVCDTLGLITHXXXXXXXXXXXXXXXRPLSEVMDLSVGDGTGLLTLGEIVAQATEGIP 235
            V VCD +G+ITH               +   + + L+  D   L++ G ++  A  G
Sbjct: 184 IVVVCDRVGMITHANSAASAIYDGDLIGKMFEDAIPLTFTDAPDLMSGGALIDLALNGQA 243

Query: 236 VQGIEAVAAEGTPF--YLISAAPLQVPGEAVSGCVITMVDLSQRKAAERHQQLLLRELDH 293
            QGIEA+A        YLISAAPLQV  + +SGCV+TMVDLSQRKAAE  Q LL+RELDH
Sbjct: 244 RQGIEAIATRAPKVKDYLISAAPLQVTEDQISGCVLTMVDLSQRKAAEHQQLLLMRELDH 303

Query: 294 RVKNTLALVMSISRRTMHSEETLEGYQKAFTARIQALAATHNLLADKSWSDISIRDVLVR 353
           RV+NTLALV+SIS RT+ +E+TL+G+ +AFT RI  LAATH+LLA + W+ +S+ D++
Sbjct: 304 RVRNTLALVLSISNRTLSNEDTLQGFHQAFTQRIHGLAATHSLLAKQGWTKLSLHDIVRA 363

Query: 354 ELAPYNEGFSQRILVEVPDVEIEPRSAIALGLVIHELATNATKYGSLSTPEGQ--VRVRG 411
           ELAPY E    R+ +E  +V + PR+AIALGL+ HELATNA KYG+LS   G   V VRG
Sbjct: 364 ELAPYVETDGTRLRLEGGEVALIPRAAIALGLIFHELATNAVKYGALSREGGHVLVAVRG 423

Query: 412 LPGADEPADVVCLEWLERGGPPVSEPTRSGFGQTVIRHAFAYAEGGGAEVSFEPDGVRCR 471
            P AD  A  V  +W+E GGP VS P R GFG TVI H+ AY+  GG ++SF P+GV C
Sbjct: 424 -PTADGAAMRV--DWVESGGPMVSPPQRKGFGHTVISHSLAYSSKGGTDLSFPPEGVICA 480

Query: 472 VSVP 475
           + +P
Sbjct: 481 LRIP 484
```


Mouse over to see the title, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
300
600
900
1200
1500
Questions/comments

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Notohymena australis strain SL4 18S ribosomal RNA gene, partial sequence; macronuclear	3173	3173	100%	0.0	100.00%	MF804419.1
□	Notohymena australis isolate FXP20121111-01 small subunit ribosomal RNA gene, partial sequence	3077	3077	98%	0.0	99.41%	KP100451.1
□	Notohymena apoaustralis small subunit ribosomal RNA gene, complete sequence	3066	3066	98%	0.0	99.41%	KC430934.1
□	Apoamphisiella vernalis strain MG clone 2 18S ribosomal RNA gene, complete sequence	3044	3044	98%	0.0	99.17%	KU522215.1
□	Apoamphisiella vernalis strain PA clone 1 18S ribosomal RNA gene, complete sequence	3040	3040	98%	0.0	99.11%	KU522216.1
□	Paraurostyla weissei macronuclear small-subunit ribosomal RNA gene, complete sequence	3037	3037	98%	0.0	99.11%	AF164127.1
□	Paraurostyla weissei 17S ribosomal RNA gene, partial sequence	3037	3037	98%	0.0	99.11%	AY294648.1
□	Paraurostyla weissei 18S rRNA gene	3037	3037	98%	0.0	99.11%	AJ310485.1
□	Paraurostyla weissei 17S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 26S ril	3033	3033	98%	0.0	99.05%	AF508767.1
□	Apoamphisiella vernalis strain MG clone 1 18S ribosomal RNA gene, complete sequence	3013	3013	98%	0.0	98.88%	KU522214.1
□	Cyrtohymena citrina voucher BAR25 18S ribosomal RNA gene, complete sequence	3011	3011	98%	0.0	98.82%	KC182574.1
□	Notohymena sp. SL1 18S ribosomal RNA gene, partial sequence	3003	3003	98%	0.0	98.41%	KP336402.1
□	Paraurosomoida indiensis isolate KAS7 18S ribosomal RNA gene, complete sequence	3000	3000	98%	0.0	98.70%	JX139117.1
□	Urosomoida agilis 18S ribosomal RNA gene, partial sequence; macronuclear	2983	2983	98%	0.0	98.52%	KR063272.1
□	Onychodromopsis flexilis partial 18S rRNA gene, isolate Salzburg	2979	2979	98%	0.0	98.46%	AM412764.1
□	Urosomoida agilis 18S ribosomal RNA gene, complete sequence; macronuclear	2976	2976	98%	0.0		

Notohymena australis strain SL4 18S ribosomal RNA gene, partial sequence; macronuclear

Sequence ID: MF804419.1 Length: 1718 Number of Matches: 1

Range 1: 1 to 1718 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
3173 bits(1718)	0.0	1718/1718(100%)	0/1718(0%)	Plus/Plus

```
Query  1    GCACCTGTTCAGACTAGCCATGCATGTCTAAGTATAAATGTTATACAGTGAAACTGCGAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GCACCTGTTCAGACTAGCCATGCATGTCTAAGTATAAATGTTATACAGTGAAACTGCGAA  60

Query  61   GGGCTCATTAAAACAGTTATAGTTTATTTGATAATCAAAATTACATGGATAACCGTGGTA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GGGCTCATTAAAACAGTTATAGTTTATTTGATAATCAAAATTACATGGATAACCGTGGTA  120

Query  121  ATTCTAGAGCTAATACATGCTGGTTAGCCTGACTTTTGCGGAAGGGCTGTATTTATTAGA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  ATTCTAGAGCTAATACATGCTGGTTAGCCTGACTTTTGCGGAAGGGCTGTATTTATTAGA  180

Query  181  TAACAAATCAATATTCCCCGTGTCTATTGTGACGACTCATAATAACTGATCGAATCGCAT  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  TAACAAATCAATATTCCCCGTGTCTATTGTGACGACTCATAATAACTGATCGAATCGCAT  240

Query  241  GGGCTTTGCCCGCGATACATCATTCAAGTTTCTGCCCCATCAGCTTTCGATGGTAGTGTA  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GGGCTTTGCCCGCGATACATCATTCAAGTTTCTGCCCCATCAGCTTTCGATGGTAGTGTA  300

Query  301  TTGGACTACCATGGCTTTCACGGGTAACGGAGGATTAGGGTTCGATTCCGGAGAGGGAGC  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  TTGGACTACCATGGCTTTCACGGGTAACGGAGGATTAGGGTTCGATTCCGGAGAGGGAGC  360

Query  361  CTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCAATCCTGAC  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  CTGAGAAACGGCTACCACATCTAAGGAAGGCAGCAGGCGCGTAAATTACCCAATCCTGAC  420

Query  421  TCAGGGAGGTAGTGACAAGAAATAACGGACCGAAGCCTATGTTTCGGGATTGCAATGAGT  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421  TCAGGGAGGTAGTGACAAGAAATAACGGACCGAAGCCTATGTTTCGGGATTGCAATGAGT  480

Query  481  AGAATTTAAACCCCTTTACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGG  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481  AGAATTTAAACCCCTTTACGAGGATCAATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGG  540

Query  541  TAATTCCAGCTCCAATAGCGTATATTAAATTTGTTGCAGTTAAAAAGCTCGTAGTTGGAT  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541  TAATTCCAGCTCCAATAGCGTATATTAAATTTGTTGCAGTTAAAAAGCTCGTAGTTGGAT  600

Query  601  TTCTGAGAGGGCGCCAATGTCCGCTGATTGCGTGTGCAGCGGTGCCCTCTCATCCTTCTG  660
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

BLAST Output Format

The BLAST output includes a graphical overview box. The graphical overview box contains colored horizontal bars that allow quick identification of the number of database hits and the degrees of similarity of the hits. The color coding of the horizontal bars corresponds to the ranking of similarities of the sequence hits (red: most related; green and blue: moderately related; black: unrelated).

The length of the bars represents the spans of sequence alignments relative to the query sequence. Each bar is hyperlinked to the actual pairwise alignment in the text portion of the report. Below the graphical box is a list of matching hits ranked by the *E*-values in ascending order. Each hit includes the accession number, title (usually partial) of the database record, bit score, and *E*-value.

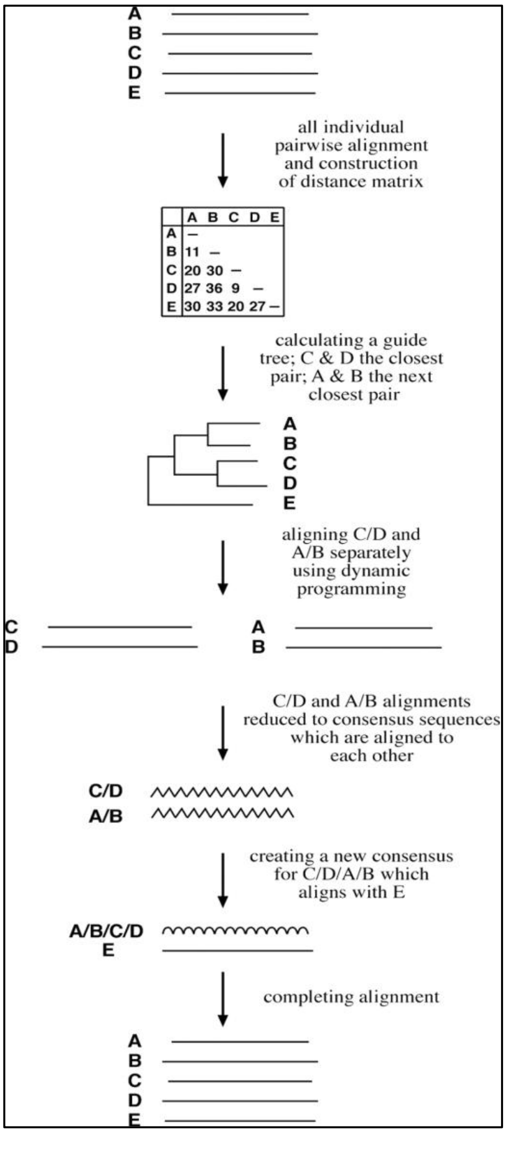
Multiple Sequence Alignment

Multiple sequence alignment is an essential technique in many bioinformatics applications.

Many algorithms have been developed to achieve optimal alignment. Some programs are exhaustive in nature; some are heuristic. Because exhaustive programs are not feasible in most cases, heuristic programs are commonly used. These include progressive, iterative, and block-based approaches. The progressive method is a stepwise assembly of multiple alignment according to pairwise similarity. A prominent example is Clustal, which is characterized by adjustable scoring matrices and gap penalties as well as by the application of weighting schemes.

Clustal (www.ebi.ac.uk/clustalw/) is a progressive multiple alignment program available either as a stand-alone or on-line program. The stand-alone program, which runs on UNIX and Macintosh, has two variants, ClustalW and ClustalX. The W version provides a simple text-based interface and the X version provides a more user-friendly graphical interface. One of the most important features of this program is the flexibility of using substitution matrices. Clustal does not rely on a single substitution matrix. Instead, it applies different scoring matrices when aligning sequences, depending on degrees of similarity. The choice of a matrix depends on the evolutionary distances measured from the guide tree. For example, for closely related sequences that are aligned in the initial steps, Clustal automatically uses the BLOSUM62 or PAM120 matrix.

Progressive Alignment methods (e.g. Clustal)



A
B
C
D
E
all individual pairwise alignment and construction of distance matrix
calculating a guide tree; C & D the closest pair; A & B the next closest pair
A
B
C
D
E
aligning C/D and A/B separately using dynamic programming
C
D
A
B
C/D and A/B alignments reduced to consensus sequences which are aligned to each other
C/D
A/B
creating a new consensus for C/D/A/B which aligns with E
A/B/C/D
E
completing alignment
A
B
C
D
E

	A	B	C	D	E
A	–				
B	11	–			
C	20	30	–		
D	27	36	9	–	
E	30	33	20	27	–

Multiple Substitutions

A simple measure of the divergence between two sequences is to count the number of substitutions in an alignment. The proportion of substitutions defines the observed distance between the two sequences. However, the observed number of substitutions may not represent the true evolutionary events that actually occurred. When a mutation is observed as A replaced by C, the nucleotide may have actually undergone a number of intermediate steps to become C, such as A→T→G→C. Similarly, a back mutation could have occurred when a mutated nucleotide reverted back to the original nucleotide. This means that when the same nucleotide is observed, mutations like G→C→G may have actually occurred. Moreover, an identical nucleotide observed in the alignment could be due to parallel mutations when both sequences mutate into T, for instance. Such multiple substitutions and convergence at individual positions obscure the estimation of the true evolutionary distances between sequences. This effect is known as *homoplasy*, which, if not corrected, can lead to the generation of incorrect trees. To correct homoplasy, statistical models are needed to infer the true evolutionary distances between sequences.

Step 3: Choosing Substitution Models

The statistical models used to correct homoplasy are called *substitution models* or *evolutionary models*. For constructing DNA phylogenies, there are a number of nucleotide substitution models available. These models differ in how multiple substitutions of each nucleotide are treated. The caveat of using these models is that if there are too many multiple substitutions at a particular position, which is often true for very divergent sequences, the position may become saturated. This means that the evolutionary divergence is beyond the ability of the statistical models to correct. In this case, true evolutionary distances cannot be derived. Therefore, only reasonably similar sequences are to be used in phylogenetic comparisons.

Jukes–Cantor Model

The simplest nucleotide substitution model is the Jukes–Cantor model, which assumes that all nucleotides are substituted with equal probability. A formula for deriving evolutionary distances that include hidden changes is introduced by using alogarithmic function.

$d\text{AB} = -(3/4)\ \ln[1 - (4/3)p\text{AB}]$

where d AB is the evolutionary distance between sequences A and B and p AB is the observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.

For example, if an alignment of sequences A and B is twenty nucleotides long and six pairs are found to be different, the sequences differ by 30%, or have an observed distance 0.3. To correct for multiple substitutions using the Jukes–Cantor model, the corrected evolutionary distance based on Equation is:

$d\text{AB} = -3/4\ \ln[1 - (4/3 \times 0.3)] = 0.38$

The Jukes–Cantor model can only handle reasonably closely related sequences. According to the given Equation, the normalized distance increases as the actual observed distance increases. For distantly related sequences, the correction can become too large to be reliable. If two DNA sequences have 25% similarity, p AB is 0.75. This leads the log value to be infinitely large.

Kimura Model

Another model to correct evolutionary distances is called the Kimura two-parameter model. This is a more sophisticated model in which mutation rates for transitions and transversion are assumed to be different, which is more realistic. According to this model, transitions occur more frequently than transversions, which, therefore, provides a more realistic estimate of evolutionary distances. The Kimura model uses the following formula:

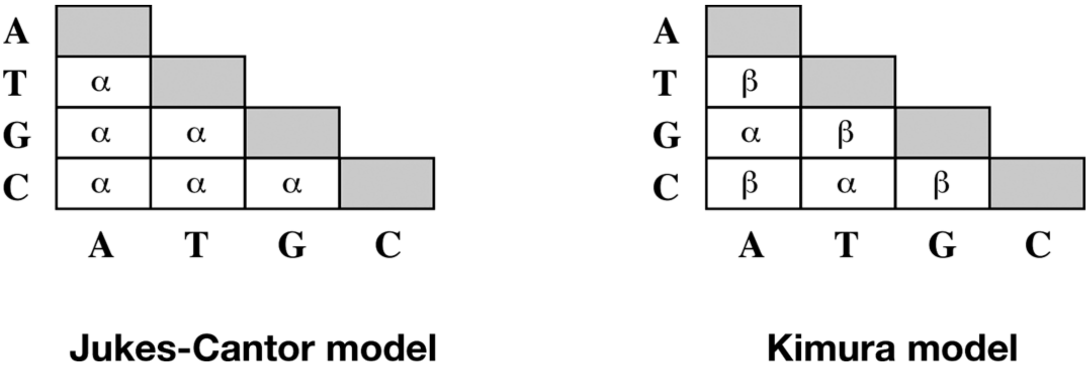
$$d\text{AB} = -(1/2)\ \ln(1 - 2p\text{ti} - p\text{tv}) - (1/4)\ \ln(1 - 2p\text{tv})$$

where d AB is the evolutionary distance between sequences Aand B, p ti is the observed frequency for transition, and p tv the frequency of transversion.

.

An example of using the Kimura model can be illustrated by the comparison of sequences A and B that differ by 30%. If 20% of changes are a result of transitions and 10% of changes are a result of transversions, the evolutionary distance can be calculated using Equation:

$$d\text{AB} = -1/2\ \ln(1 - 2 \times 0.2 - 0.1) - 1/4\ \ln(1 - 2 \times 0.1) = 0.40$$



	A	T	G	C
A				
T	α			
G	α	α		
C	α	α	α	

Jukes-Cantor model

	A	T	G	C
A				
T	β			
G	α	β		
C	β	α	β	

Kimura model

Step 4: Phylogenetic Tree Construction Methods and Programs:

There are currently two main categories of tree-building methods, each having advantages and limitations. The first category is based on discrete characters, which are molecular sequences from individual taxa. The basic assumption is that characters at corresponding positions in a multiple sequence alignment are homologous among the sequences involved. Therefore, the character states of the common ancestor can be traced from this dataset. Another assumption is that each character evolves independently and is therefore treated as an individual evolutionary unit. The second category of phylogenetic methods is based on distance, which is the amount of dissimilarity between pairs of sequences, computed on the basis of sequence alignment. The distance-based methods assume that all sequences involved are homologous and that tree branches are additive, meaning that the distance between two taxa equals the sum of all branch lengths connecting them. More details on procedures and assumptions for each type of phylogenetic method are described.

DISTANCE-BASED METHODS Clustering-Based Methods

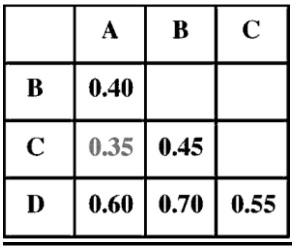
Unweighted Pair Group Method Using Arithmetic Average (UPGMA)

The simplest clustering method is UPGMA, which builds a tree by a sequential clustering method. Originally proposed in the early 1960s to help with the evolutionary analysis of morphological characters, the unweighted pair group method with arithmetic averages (UPGMA) is largely statistically based and requires data that can be condensed to a measure of genetic distance between all the pairs of taxa being considered. To illustrate the construction of a phylogenetic tree using the UPGMA method, consider a group of four taxa called A, B, C, and D. Assume that the pairwise distances between each of the taxa are given in the following matrix: In this matrix, dAB represents the distance (perhaps as calculated by the Jukes–Cantor model) between taxa A and B, dAC is the distance between taxa A and C, and so on. UPGMA begins by clustering the two taxa with the smallest distance separating them into a single, composite taxon. In this case, assume that the smallest value in the distance matrix corresponds to dAB , in which case taxa A and B are the first to be grouped together (AB). After the first clustering, a new distance matrix is computed, with the distance between the new taxon (AB) and taxa C and D being calculated as d(AB)C=1/2(dAC+dBC) and d(AB)D=1/2(dAD+dBD . The taxa separated by the smallest distance in the new matrix are then clustered together to make another new composite taxon. The process is repeated until all taxa have been grouped together. If scaled branch lengths are to be used on the tree to represent the evolutionary distance between taxa, then branch points are positioned at a distance halfway between the taxa being grouped (i.e., at for the first clustering).

A strength of distance matrix approaches in general is that they work equally well with morphological and molecular data as well as combinations of the two. They, like maximum likelihood analyses, also take into consideration all the data available for a particular analysis. In contrast, the alternative parsimony approaches discard many “non informative” sites (described later).

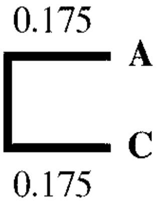
A weakness of the UPGMA approach in particular is that it presumes a constant rate of evolution across all lineages, something that is known to not always be the case. Several distance matrix-based alternatives to UPGMA such as the transformed distance method and the **neighbor-joining method** are more complex but capable of incorporating different rates of evolution within different lineages.

An Example of Phylogenetic Tree Construction Using the UPGMA Method

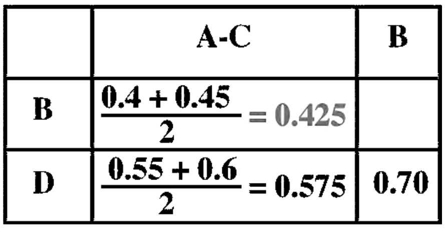


	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

1. Using a distance matrix involving four taxa, A, B, C and D, the UPGMA method first joins two closest taxa together which are A and C (0.35 in grey). Because all taxa are equidistant from the node, the branch length for A to the node is AC/2 = 0.35/2 = 0.175.

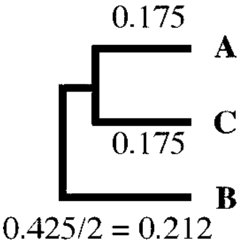


2. Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A-C cluster to every other taxa is one half of a taxon to A and C, respectively. That means that the distance of B to A-C is (AB + BC)/2; and that of D to A-C is (AD + CD)/ 2.



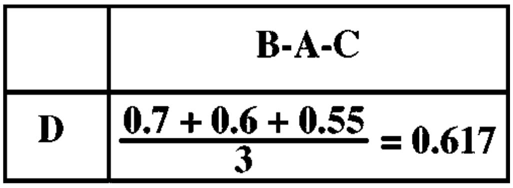
	A-C	B
B	$\frac{0.4+0.45}{2} = 0.425$	
D	$\frac{0.55+0.6}{2} = 0.575$	0.70

3. In the newly reduced-distance matrix, the smallest distance is between B and A-C (in grey), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.



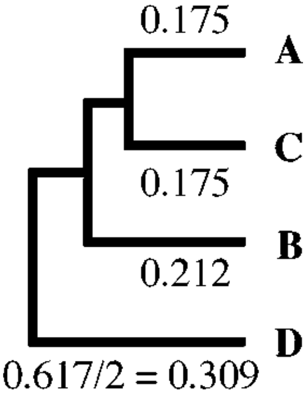
4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to

every single component which is (BD + AD + CD)/3.

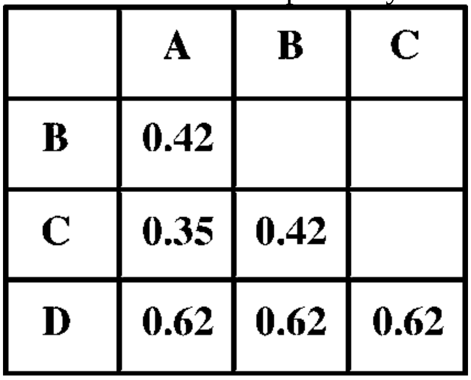


	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

5. D is the last branch to add to the tree, whose branch length is one half of D to B-A-C.



6. Because distance trees allow branches to be additive, the resulting distances between taxa from the tree path can be used to create a distance matrix. Obviously, the estimated distance matrix. Obviously, the estimated distances do not match the actual evolutionary distances shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.



	A	B	C
B	0.42		
C	0.35	0.42	
D	0.62	0.62	0.62

Neighbor Joining

The UPGMA method uses unweighted distances and assumes that all taxa have constant evolutionary rates. Since this molecular clock assumption is often not met in biological sequences, to build a more accurate phylogenetic trees, the neighborjoining (NJ) method can be used, which is somewhat similar to UPGMA in that it builds a tree by using stepwise reduced distance matrices. However, the NJ method does not assume the taxa to be equidistant from the root. It corrects for unequal evolutionary rates between sequences by using a conversion step.

CHARACTER-BASED METHODS

Character-based methods (also called *discrete methods*) are based directly on the sequence characters rather than on pairwise distances. They count mutational events accumulated on the sequences and may therefore avoid the loss of information when characters are converted to distances. This preservation of character information means that evolutionary dynamics of each character can be studied. Ancestral sequences can also be inferred. The two most popular character-based approaches are the maximum parsimony (MP) and maximum likelihood (ML) methods.

Maximum Parsimony

The parsimony method chooses a tree that has the fewest evolutionary changes or shortest overall branch lengths. It is based on a principle related to a medieval philosophy called *Occam's razor*. The theory was formulated by William of Occam in the thirteenth century and states that the simplest explanation is probably the correct one. This is because the simplest explanation requires the fewest assumptions and the fewest leaps of logic. In dealing with problems that may have an infinite number of possible solutions, choosing the simplest model may help to "shave off" those variables that are not really necessary to explain the phenomenon. By doing this, model development may become easier, and there may be less chance of introducing inconsistencies, ambiguities, and redundancies, hence, the name Occam's razor.

For phylogenetic analysis, parsimony seems a good assumption. By this principle, a tree with the least number of substitutions is probably the best to explain the differences among the taxa under study. This view is justified by the fact that evolutionary changes are relatively rare within a reasonably short time frame. This implies that a tree with minimal changes is likely to be a good estimate of the true tree. By minimizing the changes, the method minimizes the phylogenetic noise owing to homoplasy and independent evolution.

Maximum Likelihood Method

Another character-based approach is ML, which uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data. It finds a tree that most likely reflects the actual evolutionary process. ML is an exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites. By employing a particular substitution model that has probability values of residue substitutions, ML calculates the total likelihood of ancestral sequences evolving to internal nodes and eventually to existing sequences. It sometimes also incorporates parameters that account for rate variations across sites.

With this approach, probabilities are considered for every individual nucleotide substitution in a set of sequence alignments. For instance, we know that transitions are observed roughly three times as often as transversions. In a three-way alignment where a single column is found to have a **C**, a **T**, and an **A**, it can be reasonably argued that a greater likelihood exists that the sequences with the **C** and the **T** are more closely related to each other than they are to the sequences with an **A** (because the **C** to **T** change represents a transition, while the **C** or **T** to **A** change represents a transversion). Calculation of probabilities is complicated because the sequence of the common ancestor to the sequences being considered is generally not known. Determining the most likely evolutionary history is further complicated by the fact that multiple substitutions may have occurred at one or more of the sites being considered, and all sites are not necessarily independent or equivalent. Still, objective criteria can be applied to calculating the probability for every site *and* for every possible tree that describes the relationship of the sequences in

a multiple alignment. The number of possible trees for even a modest number of sequences makes this a very computationally intensive proposition, yet the one tree with the single highest aggregate probability is, by definition, the most likely to reflect the true phylogenetic tree under the proposed model of nucleotide substitution.

The dramatic increase in the raw power of computers has made maximum likelihood approaches feasible and trees inferred in this way are becoming increasingly common in the literature.

Conclusion: The number of possible trees that describe the relationship between even a small number of taxa can be very large. Distance matrix and maximum likelihood methods rely on statistical relationships between taxa to group them. Parsimony approaches assume that the tree that invokes the fewest number of mutations is most likely to be the best. No method can guarantee that it will yield the true phylogenetic tree, but when multiple substitutions are not likely to have occurred and evolutionary rates within all lineages are fairly equal, all three methods have been demonstrated to work well.

Software used for Phylogenetic tree construction are:

MEGA5: Molecular evolutionary genetics analysis version 5 is a user-friendly software for mining online databases, building sequence alignments and phylogenetic trees and using methods of evolutionary bioinformatics in basic biology, biomedicine and evolution. MEGA software is an integrated tool for statistical analyses of DNA and protein sequence data from an evolutionary standpoint.

PHYLIP (the Phylogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Step 5: Assessing tree reliability: Bootstrap value

It is also possible for portions of inferred trees to be determined with varying degrees of confidence. Bootstrap procedures allow a rough quantification of those confidence levels by randomly changing the weighting of each site. The basic approach of the bootstrap procedure is

Straight forward: a subset of the original data is drawn (with replacement) from the original data set, and a tree is inferred from the new data set. This whole process is repeated to create hundreds or thousands of resampled data sets, and portions of the inferred tree that have the same groupings in many of the repetitions are those that are especially well supported by the entire original data set. Numbers that correspond to the fraction of bootstrapped trees yielding the same grouping are often placed next to the corresponding nodes in phylogenetic trees to convey the relative confidence in each part of the tree. Bootstrapping has become very popular in phylogenetic analyses even though some methods of tree inference can make it very time-consuming to perform.

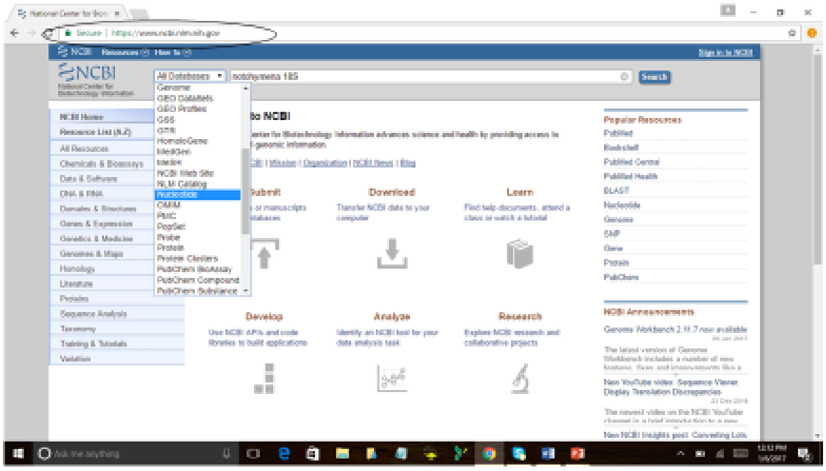
<u>Steps for Construction of Phylogenetic tree of ciliate species:</u>

USE OF DATABASE (NCBI) TO RETREIVE THE SEQUENCES

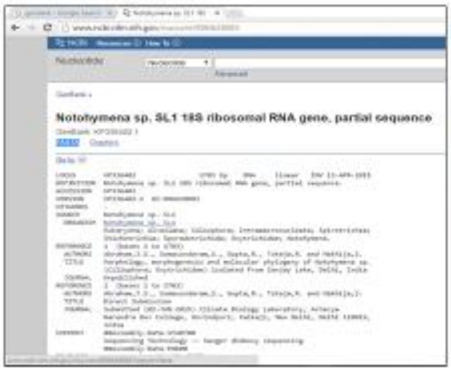
STEP I

Go to site www.ncbi.nlm.gov

Select "Nucleotide" option and enter the organism name and gene of interest and click "Search"

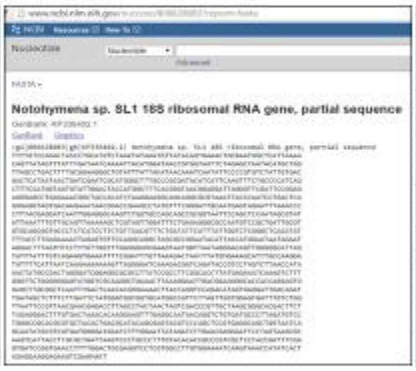


STEP II As soon as flat file opens, click on FASTA to retrieve the sequence in FASTA format. Then copy the sequence in FASTA in a new notepad file.



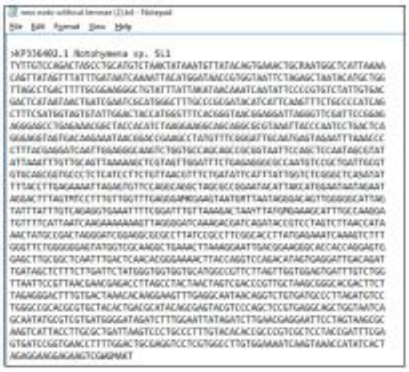
Notohymena sp. SL1 18S ribosomal RNA gene, partial sequence

GenBank: KP336402.1



Notohymena sp. SL1 18S ribosomal RNA gene, partial sequence

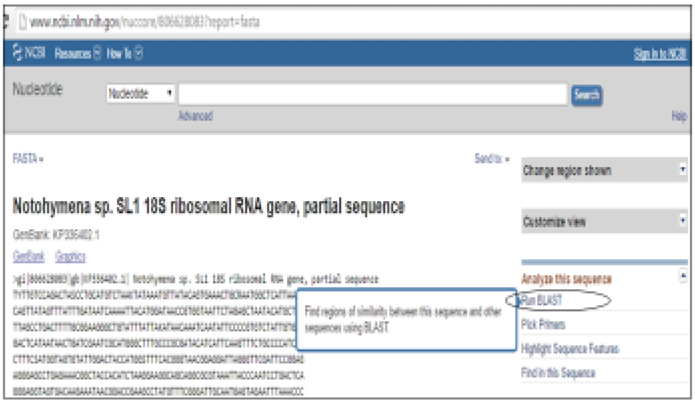
GenBank: KP336402.1



>KP336402.1 Notohymena sp. SL1

[illegible]

STEP III Run BLAST to get the homology sequences

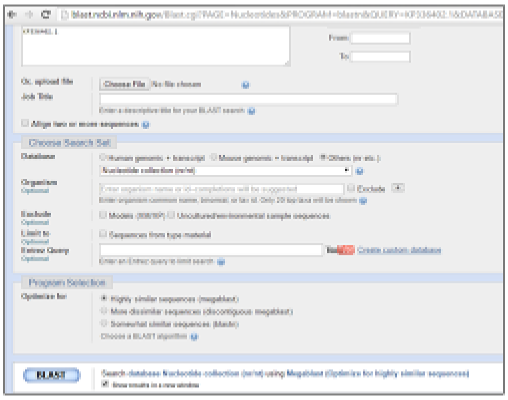


Notohymena sp. SL1 18S ribosomal RNA gene, partial sequence

GenBank: KP336402.1

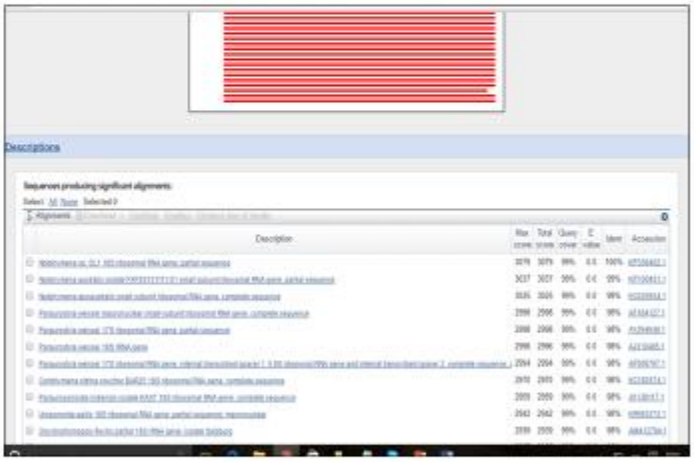
>gi|806628083|gb|KP336402.1| Notohymena sp. SL1 18S ribosomal RNA gene, partial sequence

Find regions of similarity between this sequence and other sequences using BLAST.


blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&QUERY=KP336402.1&DATABASE
From
To
Or, upload file
Choose File
No file chosen
Job Title
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.)
Nucleotide collection (nr/nt)
Organism
Exclude
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Sequences from type material
Entrez Query
Create custom database
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window

STEP IV

Retrieve the sequences of all the organisms showing maximum similarity with the organism of interest and paste in the same notepad in FASTA format. Save the file.

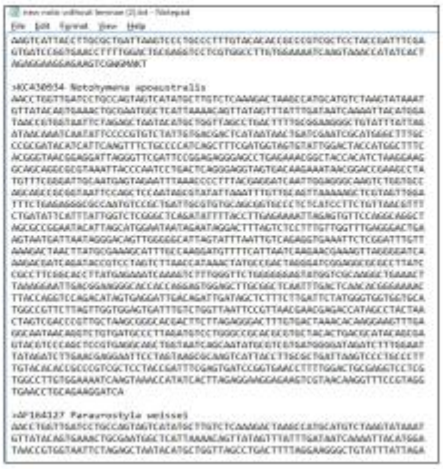


Descriptions

Sequences producing significant alignments:

Select: All None Selected 0

Description	Max score	Total score	Query cover	E value	Ident	Accession
☐ Notohymena sp. [illegible] 18S ribosomal RNA gene, partial sequence	3079	3079	99%	0.0	100%	[illegible]
☐ Notohymena [illegible] small subunit ribosomal RNA gene, partial sequence	3037	3037	99%	0.0	99%	[illegible]
☐ Notohymena apoaustralis small subunit ribosomal RNA gene, complete sequence	3035	3035	99%	0.0	99%	KC430934.1
☐ Paraurostyla weissei macronuclear small subunit ribosomal RNA gene, complete sequence	2998	2998	99%	0.0	98%	AF164127.1
☐ Paraurostyla weissei 17S ribosomal RNA gene, partial sequence	2998	2998	99%	0.0	98%	[illegible]
☐ Paraurostyla weissei 18S rRNA gene	2998	2998	99%	0.0	98%	[illegible]
☐ Paraurostyla weissei 17S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence	2994	2994	99%	0.0	98%	[illegible]
☐ Cyrtohymena citrina [illegible] 18S ribosomal RNA gene, complete sequence	2970	2970	99%	0.0	98%	[illegible]
☐ [illegible] 18S ribosomal RNA gene, complete sequence	2959	2959	99%	0.0	98%	[illegible]
☐ Urosomoida agilis 18S ribosomal RNA gene, partial sequence; [illegible]	2942	2942	99%	0.0	98%	[illegible]
☐ [illegible] partial 18S rRNA gene, [illegible]	2939	2939	99%	0.0	98%	[illegible]



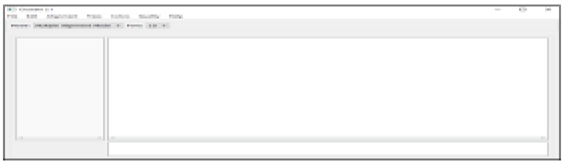
File Edit Format View Help

```
AAGTCATTACCTTGCGCTGATTAAGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTTCGA
GTGATCCGGTGAACCTTTTGGACTGCGAGGTCCTCGTGGCCTTGTGGAAAATCAAGTAAACCATATCACT
AGAGGAAGGAGAAGTCGAGAAKT

>KC430934 Notohymena apoaustralis
AACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGACTAAGCCATGCATGTCTAAGTATAAAT
GTTATACAGTGAAACTGCGAATGGCTCATTAAAACAGTTATAGTTTATTTGATAATCAAAATTACATGGA
TAACCGTGGTAATTCTAGAGCTAATACATGCTGGTTAGCCTGACTTTTGCGGAAGGGCTGTATTTATTAG
ATAACAAATCAATATTCCCCGTGTCTATTGTGACGACTCATAATAACTGATCGAATCGCATGGGCTTTGC
CCGCGATACATCATTCAAGTTTCTGCCCCATCAGCTTTCGATGGTAGTGTATTGGACTACCATGGCTTTC
ACGGGTAACGGAGGATTAGGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCTAAGGAAG
GCAGCAGGCGCGTAAATTACCCAATCCTGACTCAGGGAGGTAGTGACAAGAAATAACGGACCGAAGCCTA
TGTTTCGGGATTGCAATGAGTAGAATTTAAACCCCTTTACGAGGATCAATTGGAGGGCAAGTCTGGTGCC
AGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAATTTGTTGCAGTTAAAAAGCTCGTAGTTGGA
TTTCTGAGAGGGCGCCAATGTCCGCTGATTGCGTGTGCAGCGGTGCCCTCTCATCCTTCTGTTAACGTTT
CTGATATTCATTTATTGGTCTCGGGCTCAGATATTTTACCTTGAGAAAATTAGAGTGTTCCAGGCAGGCT
AGCGCCGGAATACATTAGCATGGAATAATAGAATAGGACTTTAGTCTCCTTTGTTGGTTTGAGGGACTGA
AGTAATGATTAATAGGGACAGTTGGGGGCATTAGTATTTAATTGTCAGAGGTGAAATTCTCGGATTTGTT
AAAGACTAACTTATGCGAAAGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTAGGGGATCA
AAGACGATCAGATACCGTCCTAGTCTTAACCATAAACTATGCCGACTAGGGATCGGAGGCGCGCCTTATC
CGCCTTCGGCACCTTATGAGAAATCAAAGTCTTTGGGTTCTGGGGGGAGTATGGTCGCAAGGCTGAAACT
TAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCTTGCGGCTCAATTTGACTCAACACGGGAAAAC
TTACCAGGTCCAGACATAGTGAGGATTGACAGATTGATAGCTCTTTCTTGATTCTATGGGTGGTGGTGCA
TGGCCGTTCTTAGTTGGTGGAGTGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCATAGCCTACTAA
CTAGTCGACCCGTTGCTAAGCGGGCACGACTTCTTAGAGGGACTTTGTGACTAAACACAAGGAAGTTTGA
GGCAATAACAGGTCTGTGATGCCCTTAGATGTCCTGGGCCGCACGCGTGCTACACTGACGCATACAGCGA
GTACGTCCCAGCTCCGTGAGGCAGCTGGTAATCAGCAATATGCGTCGTGATGGGGATAGATCTTTGGAAT
TATAGATCTTGAACGAGGAATTCCTAGTAAGCGCAAGTCATTACCTTGCGCTGATTAAGTCCCTGCCCTT
TGTACACACCGCCCGTCGCTCCTACCGATTTCGAGTGATCCGGTGAACCTTTTGGACTGCGAGGTCCTCG
TGGCCTTGTGGAAAATCAAGTAAACCATATCACTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCCGTAGG
TGAACCTGCAGAAGGATCA

>AF164127 Paraurostyla weissei
AACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGACTAAGCCATGCATGTCTAAGTATAAAT
GTTATACAGTGAAACTGCGAATGGCTCATTAAAACAGTTATAGTTTATTTGATAATCAAAATTACATGGA
TAACCGTGGTAATTCTAGAGCTAATACATGCTGGTTAGCCTGACTTTTAGGAAGGGCTGTATTTATTAGA
```

5. Multiple Sequence Alignment by ClustalX

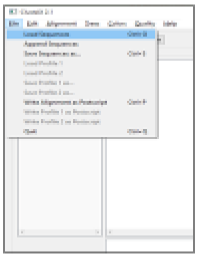
STEP I

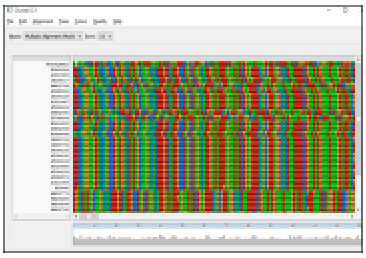
Open ClustalX 2.1 and set the Mode in “Multiple Alignment Mode”.



STEP II

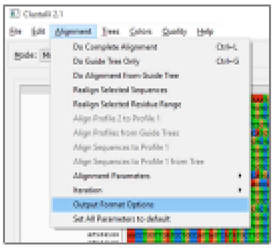
Open File, click on “Load Sequences” option and open the file saved in .txt format

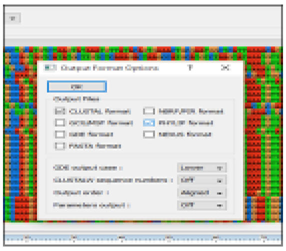

ClustalX 2.1
File Edit Alignment Trees Colors Quality Help
Load Sequences Ctrl+O
Append Sequences
Save Sequences as... Ctrl+S
Load Profile 1
Load Profile 2
Save Profile 1 as...
Save Profile 2 as...
Write Alignment as Postscript Ctrl+P
Write Profile 1 as Postscript
Write Profile 2 as Postscript
Quit Ctrl+Q



STEP III

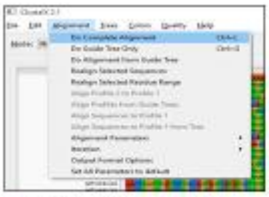
Go to “Alignment”. Select “Output Format Options” from the dropdown menu and select “PHYLIP format.” Then click on “OK”

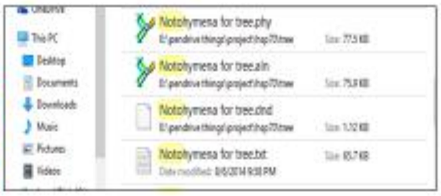

ClustalX 2.1
File Edit Alignment Trees Colors Quality Help
Mode: M
Do Complete Alignment Ctrl+L
Do Guide Tree Only Ctrl+G
Do Alignment from Guide Tree
Realign Selected Sequences
Realign Selected Residue Range
Align Profile 2 to Profile 1
Align Profiles from Guide Trees
Align Sequences to Profile 1
Align Sequences to Profile 1 from Tree
Alignment Parameters
Iteration
Output Format Options
Set All Parameters to default



STEP IV

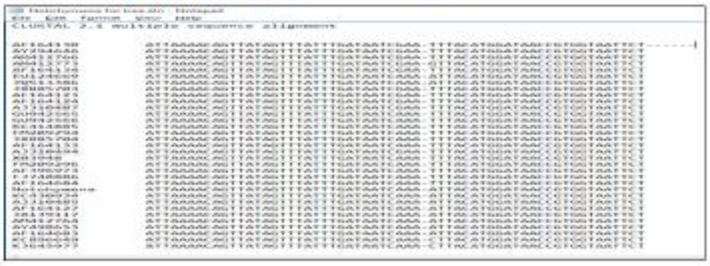
Again go to Alignment. Click on “Do Complete Alignment”. After alignment, 3 files will be created: in .aln format, in .phy format and .dnd format


ClustalX 2.1
File Edit Alignment Trees Colors Quality Help
Do Complete Alignment Ctrl+L
Do Guide Tree Only Ctrl+G
Output Format Options
Set All Parameters to default



STEP V

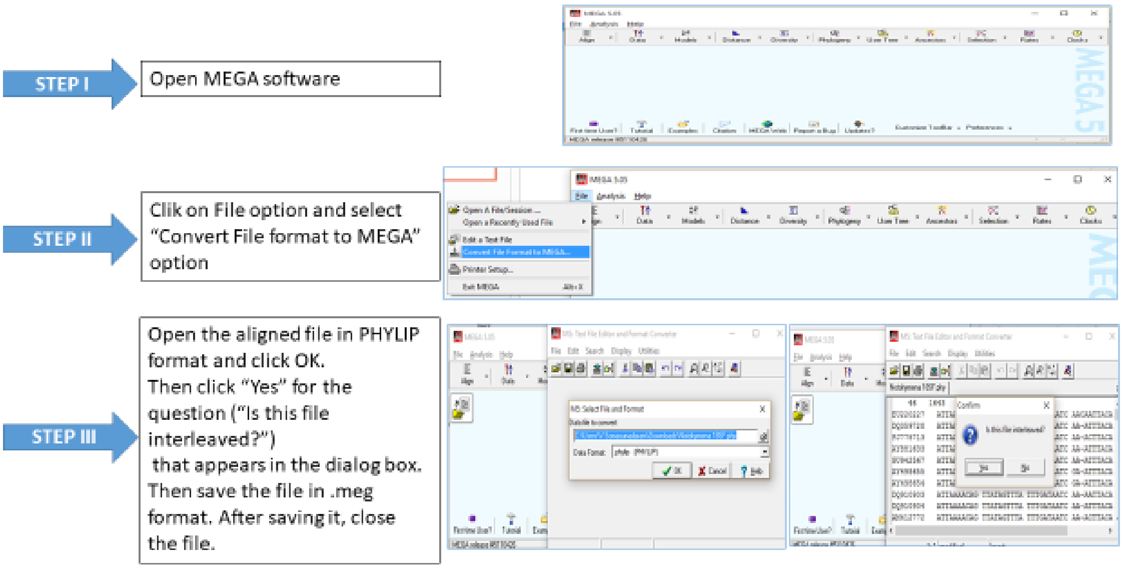
Open .aln file and do the alignment by removing the blank spaces. Then again go to file, select load sequences and open the modified .aln file. Go to “Alignment” and “Do Complete Alignment”.


Notohymena for tree.aln - Notepad
File Edit Format View Help
CLUSTAL 2.1 multiple sequence alignment

STEP VI

Again 3 files in .aln, .phy and .dnd format will be formed. Now save the .phy file for constructing tree by MEGA software

6. STEPS FOR CONSTRUCTING PHYLOGENETIC TREE USING MEGA 5.05 SOFTWARE:



STEP I

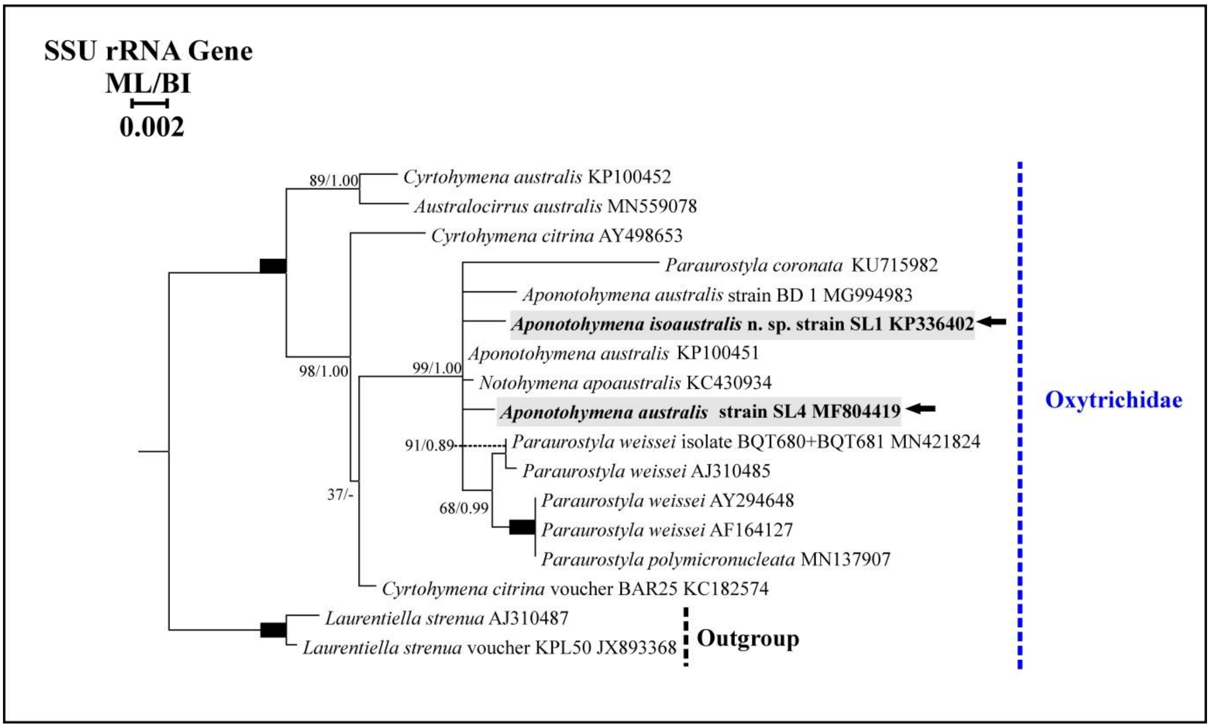
Open MEGA software

STEP II

Clik on File option and select "Convert File format to MEGA" option

STEP III

Open the aligned file in PHYLIP format and click OK. Then click "Yes" for the question ("Is this file interleaved?") that appears in the dialog box. Then save the file in .meg format. After saving it, close the file.

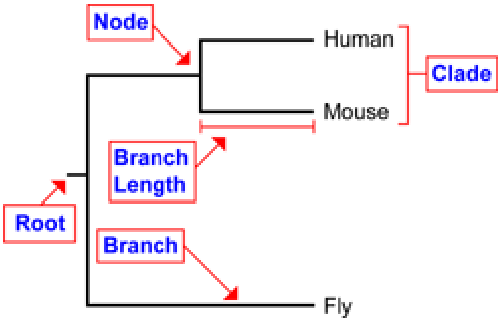


SSU rRNA Gene
ML/BI
0.002
89/1.00
Cyrtohymena australis KP100452
Australocirrus australis MN559078
Cyrtohymena citrina AY498653
Paraurostyla coronata KU715982
Aponotohymena australis strain BD 1 MG994983
Aponotohymena isoaustralis n. sp. strain SL1 KP336402
98/1.00
99/1.00
Aponotohymena australis KP100451
Notohymena apoaustralis KC430934
Aponotohymena australis strain SL4 MF804419
91/0.89
Paraurostyla weissei isolate BQT680+BQT681 MN421824
Paraurostyla weissei AJ310485
37/-
68/0.99
Paraurostyla weissei AY294648
Paraurostyla weissei AF164127
Paraurostyla polymicronucleata MN137907
Cyrtohymena citrina voucher BAR25 KC182574
Laurentiella strenua AJ310487
Laurentiella strenua voucher KPL50 JX893368
Outgroup
Oxytrichidae

Fig : Maximum likelihood (ML) phylogenetic tree inferred from small subunit (SSU) rRNA gene sequences of the new species *Apontohymena isoaustralis* and *Aponotohymena australis*.

Step 6: Tree evaluation

Phylogenetic Trees are a representation of evolutionary relationships.



The lines in the tree are called *branches*. At the tips of the branches are present-day species or sequences known as *taxa* (the singular form is *taxon*) or operational taxonomic units (OTUs).

The connecting point where two adjacent branches join is called a *node*, which represents an inferred ancestor of extant taxa.

The bifurcating point at the very bottom of the tree is the *root node*, which represents the common ancestor of all members of the tree.

A group of taxa descended from a single common ancestor is defined as a *clade* or *monophyletic group*. In a monophyletic group, two taxa share a unique common ancestor not shared by any other taxa. They are also referred to as *sister taxa* to each other. The branch path depicting an ancestor–descendant relationship on a tree is called a *lineage*, which is often synonymous with a tree branch leading to a defined monophyletic group.

When a number of taxa share more than one closest common ancestors, they do not fit the definition of a clade. In this case, they are referred to as *paraphyletic* (e.g., taxa B, C, and D).

The branching pattern in a tree is called *tree topology*. When all branches bifurcate on a phylogenetic tree, it is referred to as *dichotomy*. In this case, each ancestor divides and gives rise to two descendants. Sometimes, a branch point on a phylogenetic tree may have more than two descendents, resulting in a *multifurcating node*. The phylogeny with multifurcating branches is called *polytomy*. A polytomy can be a result of either an ancestral taxon giving rise to more than two immediate descendants simultaneously during evolution, a process known as *radiation*, or an unresolved phylogeny in which the exact order of bifurcations cannot be determined precisely.

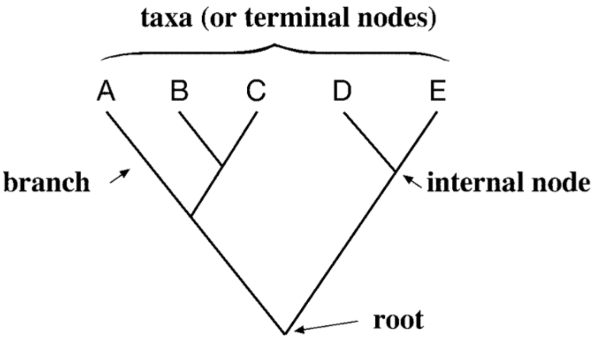
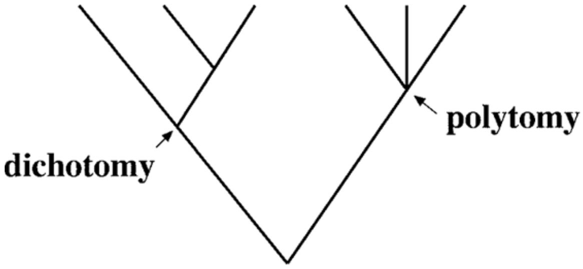


Fig. A typical bifurcating phylogenetic tree showing root, internal node, terminal nodes and branches.

A phylogenetic tree can be either rooted or unrooted. An *unrooted phylogenetic tree* does not assume knowledge of a common ancestor, but only positions the taxa to show their relative relationships. Because there is no indication of which node represents an ancestor, there is no direction of an evolutionary path in an unrooted tree. To define the direction of an evolution path, a tree must be rooted.

Fig. A phylogenetic tree showing an example of bifurcation and multifurcation. Multifurcation is normally a result of insufficient evidence to fully resolve the tree or a result of evolutionary process known as radiation



In a *rooted tree*, all the sequences under study have a common ancestor or root node from which a unique evolutionary path leads to all other nodes. Obviously, a rooted tree is more informative than an unrooted one. To convert an unrooted tree to a rooted tree, one needs to first determine where the root is. Strictly speaking, the root of the tree is not known; the common ancestor is already extinct. In practice, however, it is often desirable to define the root of a tree. There are two ways to define the root of a tree. One is to use an *outgroup*, which is a sequence that is homologous to the sequences under consideration, but separated from those sequences at an early evolutionary time. Outgroups are generally determined from independent sources of information. For example, a bird sequence can be used as a root for the phylogenetic analysis of mammals based on multiple lines of evidence that indicate that birds branched off prior to all mammalian taxa in the ingroup. Outgroups are required to be distinct from the in group sequences, but not too distant from the ingroup. Using too divergent sequences as an outgroup can lead to errors in tree construction.

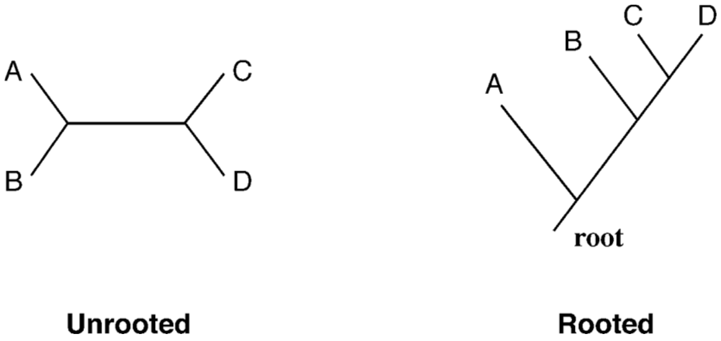


Fig. An illustration of rooted versus unrooted trees. A phylogenetic tree without definition of a root is unrooted (left). The tree with a root is rooted (right).

FORMS OF TREE REPRESENTATION

The topology of branches in a tree defines the relationships between the taxa. Thetrees can be drawn in different ways, such as a cladogram or a phylogram In each of these tree representations, the branches of a tree can freely rotate without changing the relationships among the taxa.

In a *phylogram*, the branch lengths represent the amount of evolutionary divergence.Such trees are said to be scaled. The scaled trees have the advantage of showingboth the evolutionary relationships and information about the relative divergence time of the branches.

In a *cladogram*, however, the external taxa line up neatly in a row or column. Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning. In such an unscaled tree, only the topology of the tree matters, which shows the relative ordering of the taxa.

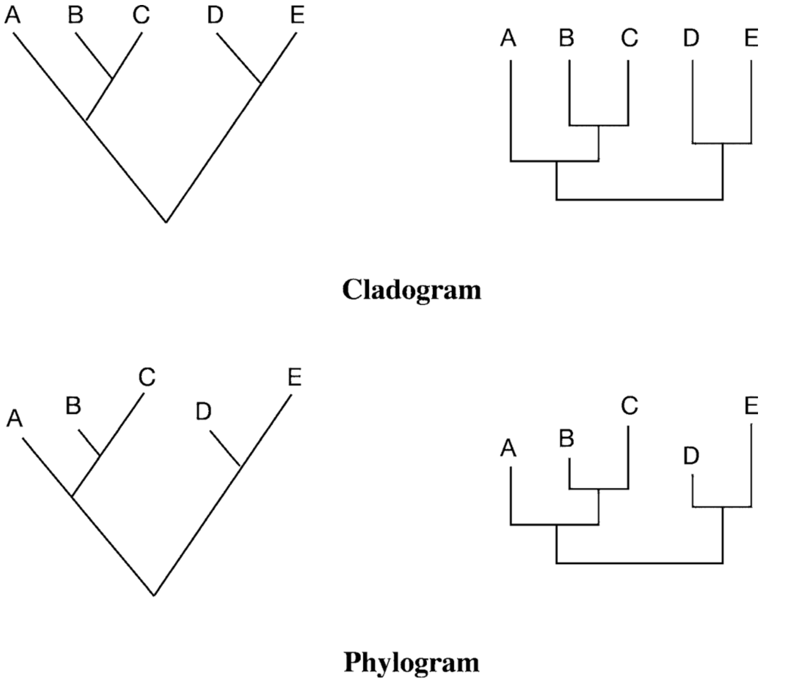


Fig. Phylogenetic trees drawn as cladograms (top) and phylograms (bottom). The branch lengths are unscaled in the cladograms and scaled in the phylograms. The trees can be drawn as angled form (left) or squared form (right).

To provide information of tree topology to computer programs without having to draw the tree itself, a special text format known as the *Newick format* is developed. In this format, trees are represented by taxa included in nested parentheses. In this linear representation, each internal node is represented by a pair of parentheses that enclose all member of a monophyletic group separated by a comma.

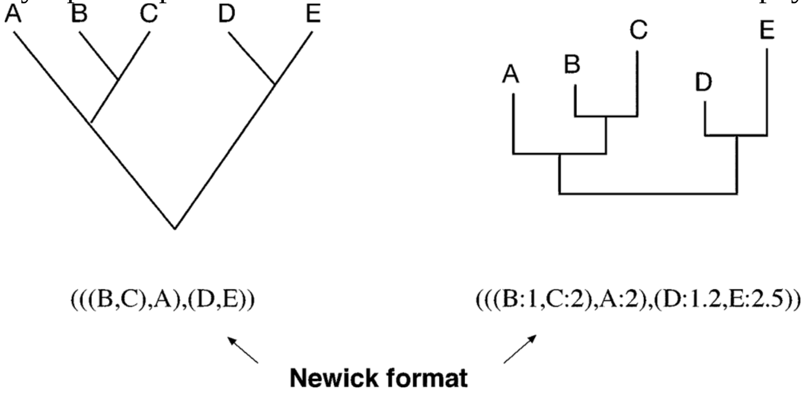


Fig. Newick format of tree representation that employs a linear form of nested parentheses within which taxa are separated by commas. If the tree is scaled, branch length are indicated immediately after the taxon name. The numbers are relative units that represent divergent times.

ISCB2025