Chapter 3. MATERIALS AND METHODOLOGY

3.1 Materials

3.1.1 DNA barcode sequence

3.1.1.1 Retrieved DNA barcode sequences

a) Indian freshwater fishes

The Public Data Portal of BOLD Ratnasingham S (2007) and Core Nucleotide database of GenBank were searched for *COI* barcode sequences of Indian freshwater fishes. The data were retrieved using Boolean operator 'AND' with two terms under a different context (taxonomic: Order and geographic: India) thereby extracting records that only matched both the terms. Sequences from both the databases were compiled together and duplicate records were removed, to finally get a set of 1413 barcode sequences for 179 species. Sequences of length >600 bp, with no missing nucleotides or gaps, were included, thereby reducing the possibility of NUMTs (nuclear DNA originating from mitochondrial DNA sequences) (Zhang and Hewitt 1996), and aligned using Clustal Omega McWilliam et al. (2013). Suspected erroneous sequences, with highly unlikely positions (species clustering with different family or order) or having extreme branch lengths were omitted, based on a Neighbor-Joining tree. The *COI* coding DNA sequence were translated using MEGA 5.1 and aligned with the available COI amino acid sequences to ensure the presence of an open reading frame (Tamura et al. 2011). The sequences were trimmed at either ends to exclude any gaps and a final set of 503 bp long 1383 consensus barcode sequences for 175 species were used for analysis and are detailed in Appendix 2. For calculating sequence information and selecting minibarcode segment, orders having good numbers of representative barcodes with sufficient species coverage were considered. Thus, 1307 sequences from three orders viz, Cypriniformes, Siluriformes and Perciformes were included in the analysis

b) Global Cypriniformes

All the species with representative *COI* barcode for the order Cypriniformes were retrieved from BOLD. Among them 24 species, had sequences submitted from India along with submissions from other geographical locations across the world. All the available *COI* barcode sequences for these species were sorted and aligned using MAFFT 7.015 (Katoh and Standley 2013). A final curated dataset of 540bp long 393 non-redundant sequences of global Cypriniformes species were obtained.

3.1.1.2 Generated barcode sequences from eastern and Northeastern India.

In this study, an attempt is made to authenticate species identity of market samples such as fillet and juvenile specimen, through DNA barcoding. Such samples are difficult to identify through traditional morphological method due to loss or absence of key identifying features. 15 freshwater fish specimens from market were collected; some were in the form of fillet. The specimens belonged to two main locations Barak river system and Hooghly river system. The samples were purchased from market based on their local names. The tissue samples from each of the specimens were collected aseptically and preserved in 90% ethanol. The specimens were then barcoded and the barcode sequences were checked against the reference barcode database to find their species identity. Taxonomic identification was resolved for the specimens for which morphological features were intact and the *COI* barcode sequence were deposited in GenBank.

Further, 172 sequences of 70 species from Northeast India were developed in our laboratory. Along with other group specific studies the sequences were used in this study for overall assessment of Indian freshwater fishes. The specimens have been vouchered and preserved which helped to confirm the identity of the species referred to in the study. This also helped to clarify misidentification and confusions arising from erroneous sequences present in the database. All the sequences developed in our laboratory and used in this study have been marked with asterisks in Appendix 2.

3.1.2 Chemicals

The chemicals were mostly of molecular biology grade and includes Tris hydroxymethyl) Aminomethane also called as Tris Base, Ethylenediamine Tetrachloro Acetic acid (EDTA), Sodium chloride (NaCl), Sodium Dodecyl Sulfate (SDS), Ethidium Bromide (EtBr), Proteinase-K (Promega, USA), Tris saturated phenol, Chloroform, Glacial Acetic acid, Isoamyl alcohol and Glycerol (SRL, India), Absolute Ethanol (Bengal Chemical and Pharmaceuticals Works Ltd. Kolkata), Taq DNA polymerase, dNTPs Master mix, Reagents kit, Magnesium chloride ($MgCl_2$) Buffer etc. (Fermentas, Germany; Bioline, U.K; Applied Biosystems USA).

3.1.2.1 Reagents

I. **Tris (hydroxymethyl) aminomethane-HCl or Tris-HCl- 10 ml (1M, pH-8.00).**

Trisbase 1.211 gm

Nuclease free water 10 ml

Tris base 1.211 gm was added to 7 ml of nuclease free water; pH was noted and adjusted to pH-8.00 by adding concentrated HCl. Once pH becomes stable at pH-8.00 remaining amount of nuclease free water was added to make the final volume 10 ml and then autoclaved at 15 lb for 10 minutes.

II. **Ethylenediamine tetrachloro acetic acid or EDTA-10 ml (0.5M, pH-8.00).**

EDTA 1.862 gm.

Nuclease free water 10 ml.

EDTA 1.862 gm was added in 7 ml of nuclease free water; initial pH was noted and adjusted to pH-8.00 by adding NaOH. Once pH becomes stable at pH-8.00 remaining amount of nuclease free water was added to make the final volume 10 ml and then autoclaved at 15 lb for 10 minutes and kept at 220C.

III. **Sodium chloride or NaCl- 10 ml (5M).**

NaCl 2.422 gm.

Nuclease free water 10 ml.

NaCl 2.422 gm was added in 10 ml of nuclease free water and dissolved properly with an autoclaved stirrer and then further autoclaved at 15 lb for 10 mins and kept at 220C.

IV. Sodium Dodecyl Sulfate or SDS- 10 ml (20%).

SDS 2 gm.

Nuclease free water 10 ml.

2 gm SDS powder was added with 10 ml of nuclease free water at room temperature and kept at 370C.

V. Ethidium Bromide or EtBr- (10 mg/ml).

Ethidium bromide 10 mg

Nuclease free water 1 ml

Mixed properly and wrapped the container in Aluminium foil and stored at room temperature.

VI. Proteinase-K- 200 µg/ml in nuclease free water.

VII. Tris saturated phenol (pH 8.00)

Solid phenol was liquefied at 68°C in water bath and equal volume of 1 M Tris HCl (pH 8.00) was added. Discarded the upper aqueous phase and repeated the extraction procedure until the pH of the aqueous phase became 8.00. After final extraction an equal volume of TE was added to it and stored in a bottle wrapped with aluminum foil at 4°C.

VIII. Chloroform-Isoamyl alcohol (24:1).

24 parts of chloroform was added with 1 part of Iso-amyl alcohol and mixed properly and stored at 4°C.

IX. Alcohol grades- 100%, 90%, 70%

3.1.2.2 Buffers

I. TES buffer (10 ml)

Tris-HCl 50 mM (added 500 µl from stock 1 M Tris-HCl)

EDTA 25 mM (added 500 µl from stock 0.5 M EDTA)

NaCl 150 mM (added 300 µl from stock 5 M NaCl)

Nuclease free water 8.7 ml

II. **Lysis buffer (10 ml)**

Tris-HCl 10 mM (added 100 µl from stock 1 M Tris-HCl)

EDTA 10 mM (added 200 µl from stock 0.5 M EDTA)

NaCl 50 mM (added 100 µl from stock 5 M NaCl)

SDS 20% (100 µl)

Nuclease free water 8.6 ml

III. **TAE buffer – 10 ml (50X)**

Tris-HCl 2.42 gm

Glacial acetic acid 570 µl

EDTA- 500mM 1000 µl

Nuclease free water 8.43 ml

IV. **TE buffer**

Tris-HCl 10 mM

EDTA 1 mM

V. **Loading dye (6X)**

Bromophenol Blue 25 mg

Glycerol 9 ml

Nuclease free water 7 ml

3.1.2.3 Primers

Primer pair proposed by (Ward et al. 2005) for fishes were used in this study and these primers have been widely used as standard primers for amplifying *COI* barcode region in many previous studies. The primer pair amplified the full-length barcode segment of mitochondrial *COI* gene sequence of all the studied fishes.

FishF1-5' TCAACCAACCACAAAGACATTGGCAC 3' and

FishR1-5' TAGACTTCTGGGTGGCCAAAGAATCA 3'

3.2 Methodology

3.2.1 Generating *COI* DNA barcodes from fish samples

Isolation of genomic DNA was done from tissue samples. Muscle tissue just above caudal fin was collected aseptically in alcohol. Further steps of isolation was done following phenol-chloroform-isoamyl method (Green 2012). The protocol is described below.

3.2.1.1 Isolation from tissue

1. The tissue sample was removed from alcohol and chopped with a sterilized blade into small pieces and then kept dry in -800C for 30 minutes.
2. The tissue was added to a pre-cooled (dry ice) mortar, and homogenized gently in 2 volume (w/v) cold TES buffer. The homogenizer was kept in ice previously. The volume (500µl) was adjusted with TES buffer.
3. To the above sample, 50 µl of 10% SDS was added followed by 5-10 µl of 20mg/ml of proteinase-K, incubated at 56°C for 1-18 hours until the tissue was totally dissolved.
4. Equal volume of phenol: chloroform: isoamylalcohol (25:24:1) was added and mixed thoroughly for few minutes.
5. The samples were centrifuged for 10 minutes with 12000 rpm.
6. The upper phase was transferred to new 1.5 ml tube and equal volume of chloroform: isoamylalcohol was added. It was then centrifuged with 12000 rpm for 10 minutes.
7. The upper aqueous layer was transferred in a fresh sterilized microcentrifuge tube and double the volume of chilled Absolute ethanol was added.
8. The above sample was kept overnight at -20° C for precipitation.
9. The sample was then centrifuged at 10,000 rpm for 10 minutes.
10. The supernatant was decanted and the pellet was retained.

11. 500 µl of 70% ethanol was added to the pellet and again centrifuged at 7000 rpm for 10 minutes and the supernatant was decanted.
12. The pellet was kept for air dry under laminar air flow.
13. The pellet was re-suspended in Nuclease free water and stored at -20° C for immediate use or at - 86° C for long preservation.

3.2.1.2 Purity check and yield of the extracted DNA

The DNA isolated to be applied for downstream process such as PCR should be pure i.e. free from most of the associated proteins that keeps DNA coiled and should also be in adequate quantity. Therefore, before undergoing PCR with the extracted DNA, the purity and yield of the DNA was checked by the processes as follows:

3.2.1.3 Spectrophotometric determination

Spectrophotometric determination of purity and yield of the extracted DNA may be explained stepwise as follows:

1. The isolated DNA stock solution was taken at different dilution rate viz. 200, 50 and 10 times by adding nuclease free water.

2. 50µl of nuclease free water was taken in a cuvette and the spectrophotometer was calibrated at 260 nm as well as 280nm.

3. 2 µl of each DNA sample was added to 48 µl of nuclease free water in a cuvette and mixed well.

4. Optical densities (OD) were measured at 260 (OD260) and 280 (OD280) in UV spectrophotometer (Biophotometer, Eppendrof) against nuclease free water as blank. The yield and purity of DNA samples were estimated as follows:

• Concentration of DNA stock solution (µg/ml) = OD260 X 100(dilution factor) X 50 µg/ml/1000

• Purity of DNA stock solution = OD260/OD280 (for pure DNA sample this ratio must be in the range of (1.75 – 1.80)

• From the concentration of DNA stock solution, the total yield of DNA was calculated and recorded.

3.2.1.4 Agarose gel electrophoresis for genomic DNA quantification and quality analysis

The stepwise procedure followed for agarose gel electrophoresis was as follows:

1. Agarose powder was poured in electrophoresis buffer (IX TAE) to the desired concentration and mixed properly and then heated in a microwave oven until completely melted.

2. Ethidium bromide (10 mg/ml) was added to the gel (final concentration 0.5 ug/ml) at this point to facilitate visualization of DNA after electrophoresis.

3. After cooling the solution to about 600C, it was poured into a casting tray containing a sample comb and allowed to solidify at room temperature

4. After the gel solidified, the comb was removed carefully without disturbing the wells formed.

5. The gel, still in its plastic tray, was inserted horizontally into the electrophoresis chamber and covered with buffer.

6. Samples containing DNA mixed with loading buffer were then pipetted into the sample wells, the lid and power leads are placed on the apparatus and current was applied and flow confirmed through observing bubbles coming off the electrodes.

7. The distance DNA had migrated in the gel was judged by visually monitoring migration of the tracking dyes, Bromophenol blue that migrate through agarose gels at roughly the same rate as double-stranded DNA fragments of 300 and 4000 bp.

8. When adequate migration occurred, DNA fragments were visualized by staining with Ethidium Bromide.

9. To visualize DNA, the gel was placed on an ultraviolet transilluminator and photograph was taken in Gel-DOC (BioRad).

3.2.1.5 PCR amplification of barcode segment of *COI* gene

As mentioned above published primers of Ward et al. 2005 was used to amplify the barcode segment of *COI* gene in a Veriti Mastercycler (Applied Biosystems Inc., CA, USA).

COI DNA PCR reaction setting:

Each 50 µl PCR reaction mixers contain:

10X PCR Buffer	:	5 µl
dNTPs Mastermix (10 mM)	:	5µl
Forward prim (20 pmole/µl)	:	1µl
Reverse primer (20 pmole/µl)	:	1µl
High fidelity DNA polymerase (5 Unit/µl)	:	1µl
Genomic DNA (100-200 ng)	:	variable
Nuclease free water	:	up to 50 µl

COI DNA-PCR cycling condition:

The PCR reaction was set with an initial denaturation temperature of 94ºC host start for 3 min and subsequently, 94ºC for 1 min for denaturation, 50ºC for 45 sec, 72ºC for 45 sec for extension primer annealing for 30 cycles followed by 72 ºC for 10 min for final extension using gradient thermal cycler (Applied Biosystem, Inc. USA). Aliquots for 10 µl of DNA products from PCR amplification were loaded in 0.8-1.2 % agarose gel for electrophoresis in 1X TAE. Gel was stained with ethidium bromide and observed under UV transilluminator and documented with Gel-DOC (BioRad).

3.2.1.6 Gel-electrophoresis and purification of PCR product

The PCR-amplified products were analysed in 1% low melting agarose gel containing Ethidium Bromide staining (10 mg/ml). The single uniform band was excised and

purified using QIAquickR Gel extraction kit (QIAGEN, USA), following manufacturer's instructions as given below.

1. The DNA fragment was excised from the agarose gel with a clean, sharp scalpel.

2. The gel slice was transferred to a 1.5 ml or 2.0ml tube and 650 µl Gel solubilizer was added.

3. The tube was incubated for 10 minutes at 50ºC in a water bath until the gel slice had completely dissolved (3-4 times during incubation).

4. 50µl Binding Optimizer was added to the sample. The solution was mixed by pipetting up and down, or by vortexing.

5. 750µl of the sample was transferred to Spin Column A placed in a 2ml collection tube. The mixture was centrifuged at 10,000 x g (12,000rpm) for 1 minute. The filtrate was discarded and the collection tube was reused by placing the Spin Column back in the collection Tube. The residual solution was loaded and the centrifugation step was repeated.

6. 700µl Wash Buffer A was added and centrifuged at 10,000 x g (12,000rpm) for 1 minute. The filtrate was discarded and the collection tube was reused by placing the spin column back in the collection Tube.

7. Step 6 repeated.

8. The mixture was centrifuged at maximum speed for 2 minutes to remove all traces of ethanol and the collection tube was discarded.

9. Spin column A was placed into a 1.5 ml Elution Tube and 30-50 µl Elution buffer was added directly to the spin column membrane. This was incubated at room temperature for 1 minute, centrifuged at 10,000 x g (12,000 rpm) for 1 minute to elute the DNA.

10. The purified DNA was stored at -20ºC.

3.2.1.7 Sequencing of PCR amplicons

The purified amplicons of the *COI* were bidirectionally sequenced in an automated DNA sequencer Genetic Analyzer 3500 by Applied Biosystem.

3.2.1.8 Sequence submission

All the analyzed sequences were deposited in GenBank through the Banklt sequence submission tool (http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank) of GenBank and received valid accession numbers.

3.2.2 *In silico* sequence analysis

3.2.2.1 Sequence formats used for sequence analysis

The sequence formats used in this study were as follows:

FASTA is the simplest text-based format for representing either nucleic acid sequence or protein sequence. The format begins with single line sequence description proceed by ">". On the next line, the actual sequence is represented in the standard IUB/IUPAC amino acid or nucleic acid code.

CLUSTAL is the standard output format of popular alignment software and is also widely supported as input format in phylogenetic programs. The format is recognized by the word CLUSTAL at the beginning of the file. The sequence alignment output from CLUSTAL software is usually given the default extension "aln".

In MEGA format, the "#Mega" keyword indicates that the data is prepared for analysis using MEGA. It must be present on the very fast line in the data file. On the second line, the word "Title" must be written, which can be followed by some description of data on the same line. Each taxon label must be written on a new line starting with #.

GenBank format consists of a mixture of compressed and uncompressed ASCII text files, containing sequence data and indices that cross reference author names, journal citations, gene names and keywords to individual GenBank records.

Nexus format of sequence was used as an input format of sequences for the character based barcoding process in MESQUITE. The #NEXUS is the keyword of the format and each block starts with BEGIN block name; and finishes with END.

3.2.2.2 Data analysis

Genetic distances were calculated using Kimura's two parameter (K2P) models (Kimura 1980), as implemented in MEGA 5.1 (Tamura et al. 2011) to quantify sequence divergences among individuals. K2P based method has been argued to be not the best suited method for species delimitation. Despite this the model was chosen as, studies have shown, differences in distance between best considered model and K2P model estimates were usually minimal, and identification success rates were largely unaffected by model choice (Collins et al., 2012). Interspecific K2P distances were calculated for species with at least 2 sequences, and intraspecific K2P distances were calculated between species in the entire data set. Genetic distances were analyzed at species, genus and family level in MEGA 5.1. Neighbor joining (NJ) analysis was performed for *COI* using the K2P distance model as recommended by (Hebert et al. 2003a) using MEGA 5.1. Node supports for the trees were evaluated with 1000 bootstrap pseudo-replicates. The details regarding the sequences generated and retrieved from the database is given in Appendix 2. Correctly delimited species based on cohesive clustering by the conspecifics in the NJ tree were considered as true species. The maximum conspecific and minimum congeneric divergences between these species were used to define the species boundaries as elaborated in the "Results" section. The divergence between the minimum congeneric and maximum conspecific divergence is the lowest divergence between congeners. This divergence value has been assumed as the threshold level of species delineation and thereby considered as a barcoding gap in this study. This process of species delimitation is preferred over other methods and has been used by many computational methods to partition a sequence alignment dataset into candidate species (Puillandre et al. 2012, Zhang et al. 2013). Based on the intra and interspecific distance variation of species from the threshold, the studied species were classified into 3groups as shown in Figure 3.1 and detailed in the "Results" section.

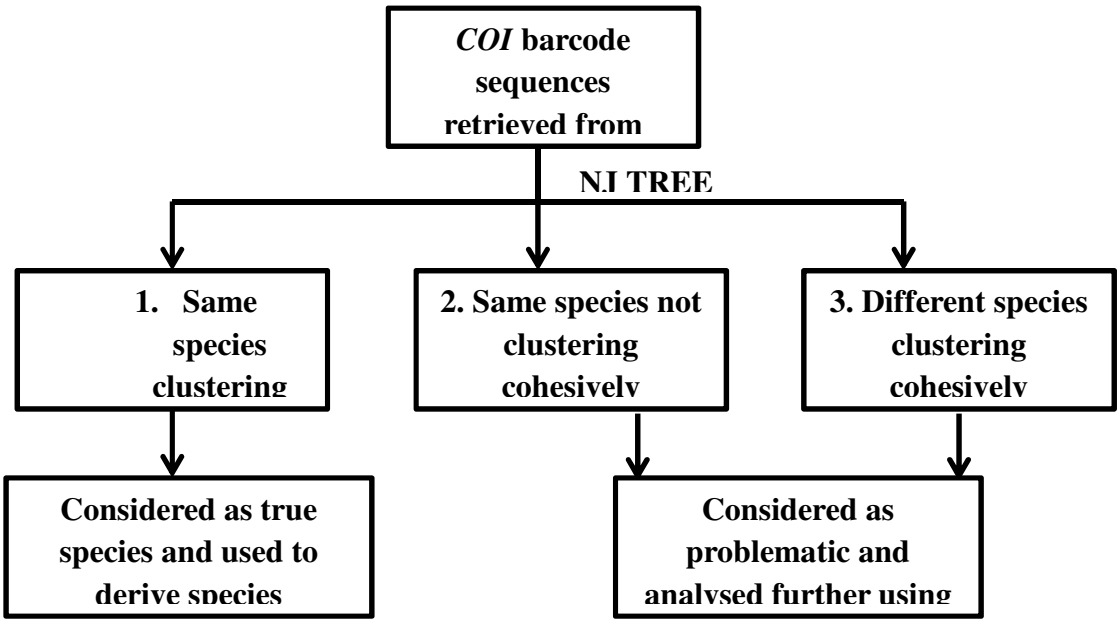

Figure 3.1 Flowchart describing the categorization of the species into three groups based on clustering pattern in NJ tree.

Sequences having same species name clustering cohesively and distinct from others are categorized as Group 1. Sequences having same species name but forming distinct clusters are categorized as Group 2. Sequences having different species name and forming cohesive cluster are categorized as Group 3. Species representing first group are considered as true species and used to derive threshold to define species boundaries while the remaining two groups are considered problematic.

3.2.2.3 R_{seq} calculation for sequence information analysis

The sequences were aligned using MAFFT and various sequence information viz: number of conserved sites, variable sites, singleton sites and parsimony informative sites were assessed in MEGA 5.1.

Information content at each nucleotide or amino acid position from the 1306 aligned sequences in an excel spreadsheet was calculated using Schneider and Stephens (1990). Here the degree of sequence conservation per site Rseq is defined as:

Rseq= log_2 N - (-Σplog_2 p)

where N is the number of options per site (four for DNA, 20 for proteins) and p is the observed frequency of each nucleotide base or amino acid at a particular position. The maximum degree of conservation possible per position is $\log_2 4$ for DNA (= 2) and $\log_2 20$ for protein (= 4.32).

3.2.2.4 Selecting high-informative segment within the *COI* barcode

Intraorder transition and transversion substitution pattern for each position in the full length *COI* barcode was calculated using Kimura 2 Parameter distance formula (employed in MS Excel) for 160 species of the three orders (Cypriniformes, Siluriformes and Perciformes) separately. To obtain the substitution pattern for maximum recoverable length of *COI* barcode, a consensus 648 bp alignment of representative sequences from 160 species was considered and few sequences that had shorter length were eliminated. A scatter plot was plotted, showing the pattern of the two types of substitution across 648bp in the three orders. From this distribution, a transversion dominant segment common to all the three orders was selected as a potential candidate for designing minibarcode.

3.2.2.5 *In silico* primer designing of minibarcode segment

The conspecific conserved sequences flanking the minibarcode segment were used to design primers for amplifying the target transversion hotspot region. The primers were designed with the help of *in silico* primer designing tool: Primer 3, abiding by the recommended thresholds of different parameters (Tm, specificity, length, hairpin formation, G-C clamp etc.). The specificity of the Primer in amplifying the target sequences was tested using representative sequences from the 160 studied species by the PCR Product program in Sequence Manipulation Suite (Stothard 2000). The program searched for perfectly matching primer annealing sites that could generate a PCR product.

3.2.2.6 Neighbor Joining tree construction

The transversion hotspot region was retrieved for all the species (160) of the three orders. Two neighbor-joining (NJ) trees were constructed, one using the full-length barcode and the other using the minibarcode region. The two trees were analyzed to compare the tree topologies derived from the two segments of different lengths. Based on the NJ trees, only those sequences, which had a confirmed species status, were considered for building the motifs.

3.2.2.7 Developing consensus barcode motifs using MOTIF-BUILD

Sequences that showed confirmed species status in the NJ tree were used for developing species-specific motif. These sequences were aligned along with a full-length template *COI* barcode, with known nucleotide positions. The sequences were trimmed on either end to get the desired minibarcode segment. The MOTIF-BUILD program was written in C++, to build consensus motifs for each species. The program identified sequences with same species name as unique entities. A motif was developed for each entity by assigning a consensus character for each position. Each position within an entity was marked by the corresponding conserved nucleotide while the variable positions were marked by degenerate nucleotides as represented by the standard International Union of Pure and Applied Chemistry (IUPAC) nomenclature (Cornish-Bowden 1985).

3.2.2.8 Specificity check of the barcode motifs using MEGA

The specificity of the motifs in identifying concerned species was first tested using the "Find motif" command in the Search menu. After the motif was entered in the search term, the Alignment Explorer found each occurrence of it and indicated it with yellow highlighting. Secondly, the program also checked whether the motifs aligned with the expected region or it showed alignment in an unexpected region.

3.2.2.9 Specificity check using MOTIF-MATCH

A straightforward pattern matching program, called MOTIF-MATCH was written in C++ to test specificity of the motifs. The motifs were expected to match exclusively with sequences of their respective species. Thus, the program used a stringent search criteria allowing only exact matching of the query motif sequence against the *COI* barcode data file. The program required users to explicitly enter a pattern to search for and is not meant for de novo pattern detection.

The main objective of this algorithm was to find dataset sequences that matched with the motif sequence. The fasta formatted full-length *COI* barcode data file and motif data file were used as input in two separate tab-delimited text files. The algorithm takes a sequence from the motif file and tries to find *COI* dataset sequences that match with the motif sequence. The *COI* dataset sequence consisted of a string of four nucleotide characters ('A','T','G','C') while the motif sequence consisted of a string of this four characters and 11 more (W, S, M, K, R, Y, B, D, H, V,N) degenerate characters. The two strings i.e. motif sequence and full-length *COI* sequence were tested for match. If a match was found then the dataset name was saved and the algorithm looked for the next dataset sequence. If the dataset sequence did not match with that of the motif sequence then the next dataset sequence was considered for the test. This process continued until the whole *COI* dataset file was traversed. The output file contains a list of sequence accession number and name against each motif's species-name. The algorithm has been described in a flowchart format in the Figure 3.2 and Figure 3.3.

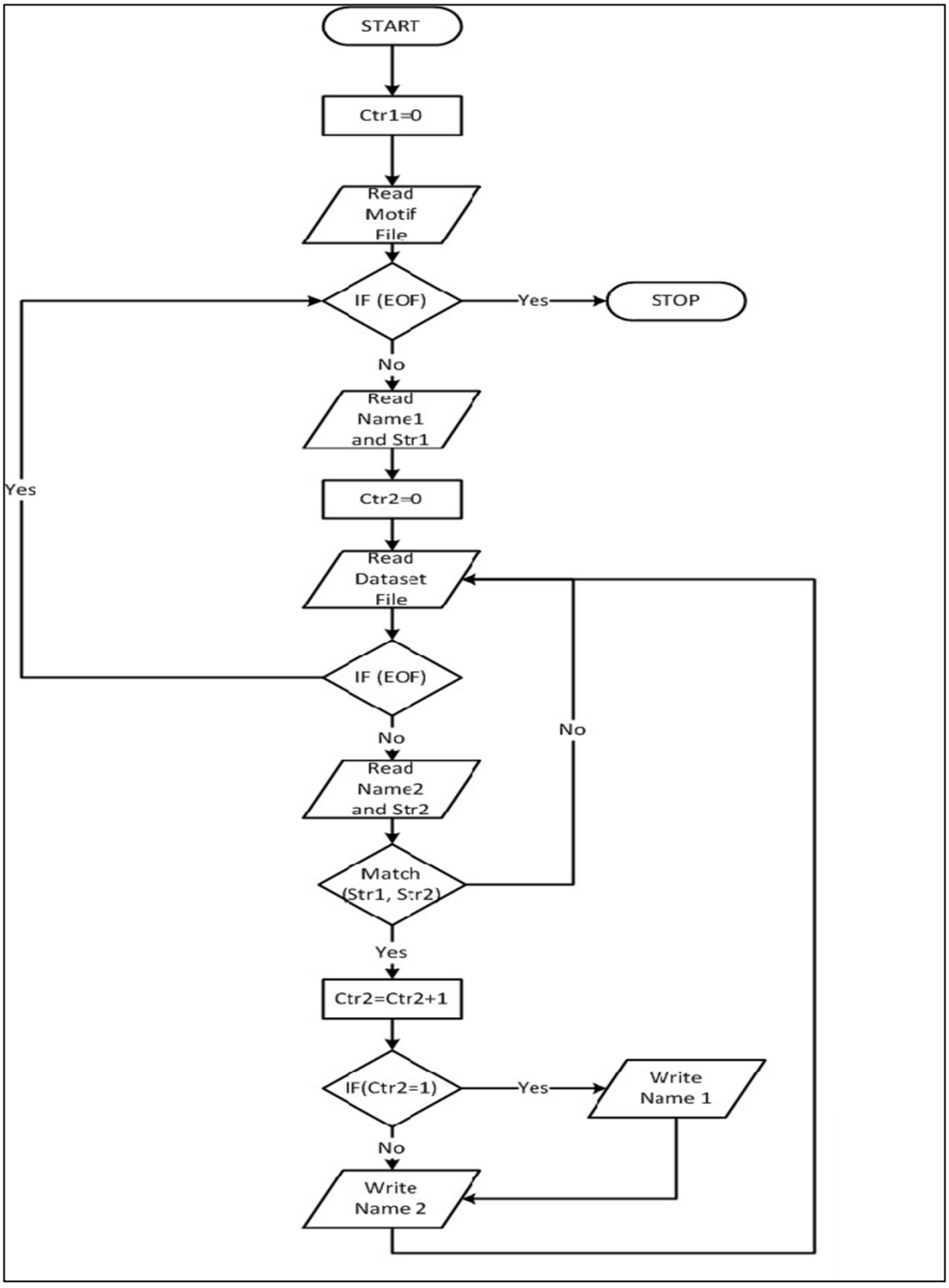

Figure 3.2 Flow chart describing the algorithm of MOTIF-MATCH Program.

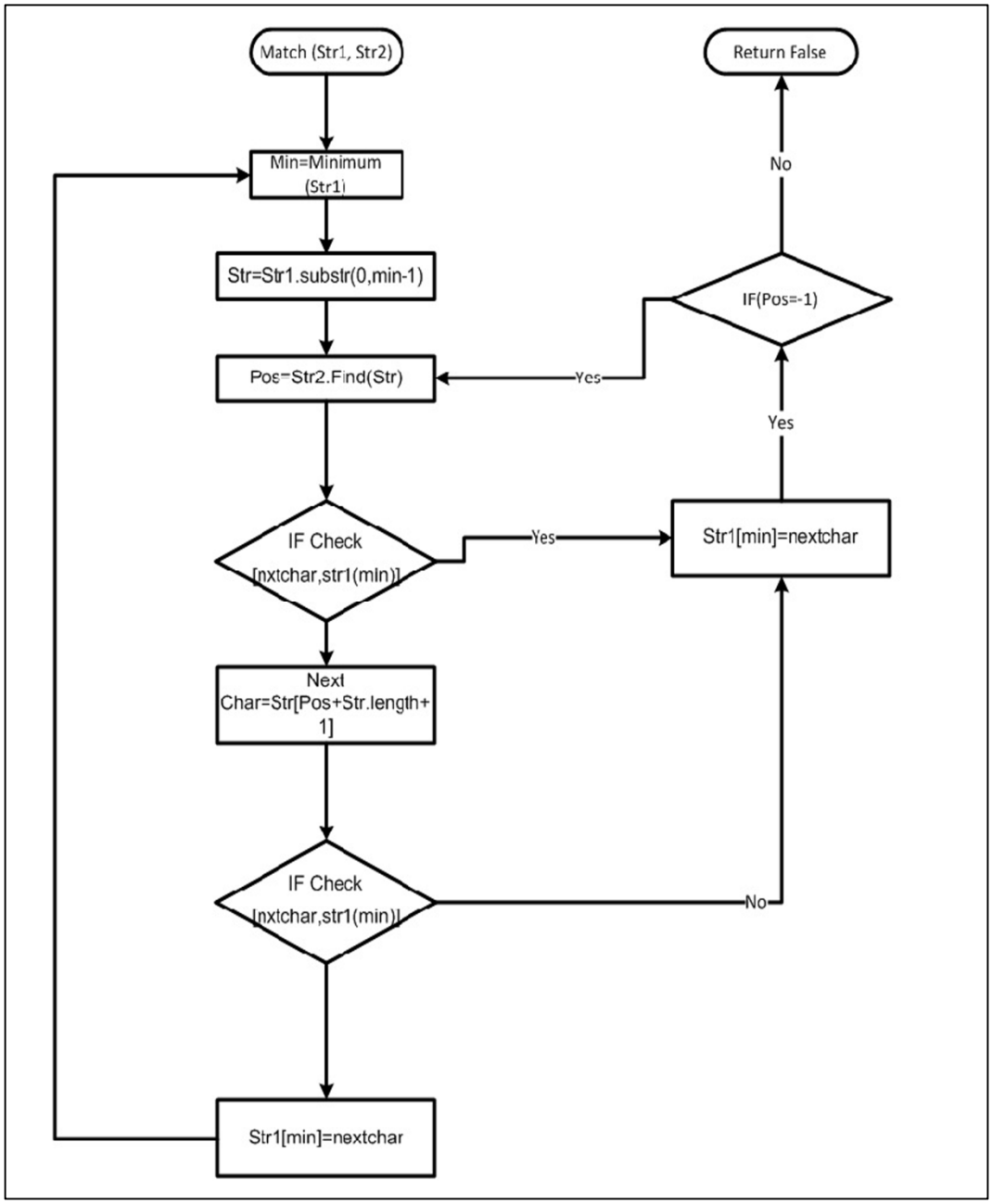

Figure 3.3 Flowchart describing the Match-String Module of the MOTIF-MATCH program