SSR MARKER DEVELOPMENT IN SORGHUM

&

PHYLOGENETIC STUDIES IN CEREALS.

Dissertation Submitted In Partial Fulfillment Of Requirement For The Award Of Degree Of

MASTER OF TECHNOLOGY
in
BIOTECHNOLOGY

By

HEMABINDU CHINTALA

CENTRE FOR BIOTECHNOLOGY
Institute of postgraduate studies &research
Jawaharlal Nehru Technology University
Hyderabad-500028

20002

TO

LORD

SHRI SHIRIDHI SAI BABA

ICRISAT

CGIAR

International Crops Research Institute for the Semi-Arid Tropics

atancheru 502 324
Andhra Pradesh
India

Tel +91 40 3296161 (19 lines)
Fax +91 40 241239
+91 40 3296182
Email ICRISAT@CGIAR.ORG

CERTIFICATE

This is to certified that the work reported in the dissertation entitled "**SSR MARKER DEVELOPMENT IN SORGHUM & PHYLOGENETIC STUDIES IN CEREALS**" Submitted by **CH.Hemabindu** have been carried out under my supervision. This work is towards the partial fulfillment of her **M.Tech** Degree from **Jawaharlal Nehru Technological University**, Hyderabad. This work is original and has not been submitted in part or full for any other degree or diploma of any university

[Dr. V. MAHALAKSHMI]
Principal Scientist,
GT-1,
ICRISAT.

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY
CENTRE FOR BIOTECHNOLOGY
Institute of Post Graduate Studies and Research
Mahaveer Marg, Hyderabad - 500 028, A.P., India
Phone No. : 040-3373020

Dr. M. LAKSHMI NARASU
Associate Professor & Head

CERTIFICATE

This is certified that the work reported in the dissertation entitled "**SSR MARKER DEVELOPMENT IN SORGHUM & PHYLOGENETIC STUDIES IN CEREALS**" Submitted in partial fulfillment for the award of **M.Tech** in biotechnology from **Jawaharlal Nehru Technological University**, Hyderabad, is a bonafied work carried out by **Ms CH. HEMEBINDU** under the guidance of **Dr.V.Mahalakshmi**, Senior scientist, International Crops Research Institute For Semi-Arid Tropic.**(ICRISAT)**

[Dr.M.Lakshmi Narasu]

DECLARATION

I Hemabindu.Ch, a bonafied student of IPGSR, JNTU, Hyderabad here by declare that the dissertation entitled "**SSR Marker Development In Sorghum And Phylogenetic Studies In Cereals**" is solely done by me under the expertise guidance of Dr.V.Mahalakshmi at International Crops Research Institute For Semi-Arid Tropics **(ICRISAT)**, Hyderabad.

The facts and figures enumerated in this project work are in accordance with the results of the modeling done in computer. This project work has not been submitted to any university or institution for the award of any degree or diploma.

ACKNOWLEDGEMETS

This project work is carried out with valuable suggestions and guidance under the supervision ofDr.V.Mahalakshmi,seniorscientist,GREP,ICRISAT,patancheru,Hyderabad. Her unstilted encouragement and deep concern helped me to complete my project work in time. I am highly grateful and indebted to her.

It gives me immense pleasure in expressing my deep sense of gratitude to Dr. Lakshmi Narasu, Head, Centre for biotechnology, Jawaharlal Nehru University, Hyderabad.I take this opportunity to thank her for suggesting me to join in the ICRISAT as an Apprentice And helped me in finishing my project work.

My sincere thanks to Dr.Prameela Devi, Dr.Archana giri associate professors and Mr.kiran, Ms.Anuradha and Ms. Radhika academic associates, Centre for biotechnology, JNTU, for their valuable guidance.

I would like to thanks to my friends Ms.Manjula, Ms.Leela, Mr.Kumar, Mr.PVNS Prasad, Ms Rekha and Mr. A.Balakrishna and other colleagues in ICRISAT for their encouragement and cooperation during my course of project work.

I show my gratitude to my beloved mother, father, brother's and sister& my Husband for their constant encouragement and good support throughout this course of work.

HEMABINDU.CH

ABSTRACT

Bioinformatics, the application of computational techniques to analyze the information associated with bio-molecules on a large scale and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies.

Chapter 1 gives an introduction and overview of the SSR marker development in sorghum . For this purpose bioinformatics tools like Tandem Repeats Finder, Primer3, Windows software and MS Access were used. More than 50,000 records were collected and placed in the INTRANET of ICRISAT, which could be accessed by the scholars and scientists for their requirements.

A tandem repeat in DNA is two or more contiguous approximate copies of a pattern of nucleotides. Extensive knowledge about pattern size, copy number, mutational history, etc. for tandem repeats has been limited by the inability to easily detect them in genomic sequence data. These sequences range in size from 3kb up to 700kb. A World Wide Web server interface at http://c3.biomath.mssm.edu/trf_advanced_submit.html has been established for automated use of the program.

Primer design is crucial for the success of PCR. Inappropriate primers cause low yield, equivocal results and misinterpretation. An ideal primer should only bind with its 3' end to a unique sequence.

Chapter 2 is dealt with an introduction about Phylogenetic studies and describes completely about the phylogeny of selected cereals for conserved enzymes. . For this purpose we utilized the software tools like CLUSTAL W for multiple alignment, JALVIEW for alignment analysis and phylogenetic tree construction, AND PRIMER 3 for primer designing which is crucial for PCR.Phylogeny is about evolution and is used to reconstruct evolutionary events. It is now possible to construct phylogenetic evolution at a molecular level through analysis of molecular sequences. namely proteins & nucleic acids.To construct phylogenetic tree among grass family, the sequences of conserved enzymes from mitochondria, chloroplast and nucleus are probed using bio-informatics tools.

CONTENTS

WEBSITES USED IN THE PROJECT WORK

1) http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
2) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
3) http://www.ebi.ac.uk/clustalw/
4) http://www.ncbi.nlm.nih.gov/
5) http://c3.biomath.mssm.edu/trf.html
6) http://c3.biomath.mssm.edu/example.html
7) http://c3.biomath.mssm.edu/trf.definitions.html#fasta
8) http://c3.biomath.mssm.edu/trf.submit.options.html
9) http://c3.biomath.mssm.edu/trf.advanced.submit.html
10) http://c3.biomath.mssm.edu/trf.upload.form.html
11) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
12) http://www2.ebi.ac.uk/~michele/jalview/contents.html
13) http://www.expasy.ch/enzyme/
14) http://www.genome.ad.jp/
15) http://www.mssm.edu/school.html

LIST OF FIGURES

BIOINFORMATICS

Introduction:

Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied math, Computer Science, and statistics) to understand and organize the associated with these molecules, on a large scale.

In the last few decades, advances in molecular biology and the equipment available for research in this field have allowed the increasingly rapid sequencing of large portions of the genomes of several species. In fact, to date, several bacterial genomes, as well as those of some simple eukaryotes (e.g., *Saccharomyces cerevisiae*, or baker's yeast) have been sequenced in full. The Human Genome Project, designed to sequence all 24 of the human chromosomes, is also progressing. Popular sequence databases, such as Genbank and EMBL, have been growing at exponential rates. This deluge of information has necessitated the careful storage, organization and indexing of sequence information. Information science has been applied to biology to produce the field called **Bioinformatics**.

The most pressing tasks in bioinformatics involve the analysis of sequence information. **Computational Biology** is the name given to this process, and it involves the following:

1. Finding the genes in the DNA sequences of various organisms
2. Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.
3. Clustering protein sequences into families of related sequences and the development of protein models.
4. Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

The Need For Bioinformatics

1. Whole Genome Analyses and Sequences
2. Experimental Analyses involving Thousands of Genes simultaneously
3. DNA Chips and Array Analyses
4. Expression Arrays
5. Comparative Analyses between Species and Strains
6. Proteomics: 'Proteome' of an Organism ... 2D gels, Mass Spec
7. Medical applications: Genetic Disease ... SNPs
8. Pharmaceutical and Biotech Industry
9. Forensic applications
10. Agricultural applications

Evolution Of Bioinformatics

After years of research in structure-function relationships of genes and proteins, the last decade proved to be extremely important and immensely satisfying due to its technical advances in genome sequences of several species and protein structure and identification. To handle this ever-increasing voluminous data, computer processing power and disk storage has been instrumental. Besides gathering all these data, it is necessary to compare these nucleotide and amino acid sequences to find similarities and differences. Since it is not convenient to compare the sequences, that are several hundreds of nucleotides by hand, several computational techniques were developed to approach this problem. In addition these are less error-prone than the manual approach. So bioinformatics has taken its place to cater the needs of biological community.

Divisions Of Bioinformatics

Bioinformatics is a multi disciplinary subject. Though only about a decade old, it has become very important for the growth of biosciences, biotechnology, and the economic prosperity of nations. Three well-identified subdivisions of Bioinformatics are:

a) Molecular Bioinformatics
b) Cellular and sub-cellular Bioinformatics and
c) Orgasmic and community Bioinformatics.

Out of these three, most Bioinformatics scientists and workers practice molecular Bioinformatics. The other two areas are more recent and are at different stages of development. In the next 5-10 years, cellular and sub-cellular Bioinformatics that will include metabolic pathways, epigenetic, and neuro Bioinformatics on one hand and Bioinformatics of Species Diversity, behavior, evolution and the effect of pollutants on higher as well lower species, on the other will occupy the main stage.

Global Importance Of Bioinformatics

Bioinformatics has acquired great importance due to its recent application in vast amount of data generated in the Genome sequence projects. The target of decoding the three billion base pairs of the human DNA has become achievable only through the use of various innovative techniques and methods evolved by the Bioinformatics scientists. Bioinformatics has become an essential component of biotechnology based product and process development. The process of drug design and development is expensive and time consuming. The application of the tools and techniques of Bioinformatics has resulted in the reduction in cost and the development cycle of the drugs. This aspect has a tremendous impact on the society. If a newly discovered drug is a life-saving one, then the resulting gains are not only in terms of financial savings but also in saving the lives of several million people. Major pharmaceutical and biotechnology companies have set-up large R&D groups in Bioinformatics.

Current research has identified biotechnology the fastest growing sector of production technology. Further advances in this sector will depend quite a lot upon the progress of Bioinformatics and hence there is a great emphasis on Bioinformatics world over. The following are a few important Websites for international networks/ institutions/ groups on Bioinformatics.

The Role Of Bioinformatics In Research

As traditional way of research is very time consuming and laborious, in these days there is introduction of computers and even biological sciences are no exception to this. The long sequences and enormous flow data in molecular biology it becomes difficult to manage data by persons or by organizations. Further with the help of computers through Internet and Intranet, sharing of data with others in the organization or organizations through out the world is becoming routine. This aids in the research and makes it easier for scientists to download, analyze, compare and exchange information. This also saves a lot of time and finances.

Uses Of Bioinformatics

1. The main role of bioinformatics for today's world is to increase productivity and quality of all plants and animals.
2. For understanding evolution, comparative genomics will add hither to unimaginable comprehensiveness to the study of the relationships between species.
3. In cell biology. the components of cellular activity, how these components interact and how they are influenced by environmental states can be identified comprehensively.
4. In medicine, sequences will provide a basis for the study of susceptibility to disease, it's pathogenesis and the development of new preventive and therapeutic approaches.
5. Bioinformatics is extremely useful to mankind as it can help in increasing crop yields and animal produces, to cater the needs of ever increasing human

population. It also helps to develop more effective drugs to protect public health as well as to control pests and diseases on the crop.

1. SSR Marker Development In Sorghum

1.1 Introduction

DNA molecules are subject to a variety of mutational events. One of the less well understood is **TANDEM DUPLICATION** in which a stretch of **DNA**, which we call the pattern, is converted into two or more copies, each following the preceding one in a contiguous fashion.for example we could have

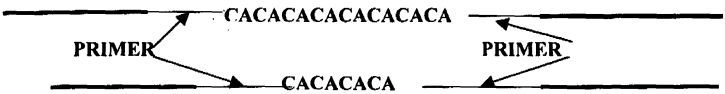
...TCGGA... ⍰ ... TCGGCGGCGGA...

In which the single occurrence of triplet **CGG** has been transformed into three identical, adjacent copies. The result of a tandem duplication event is termed as **TANDEM REPEAT.** Over time, individual copies within a tandem repeat may undergo additional, uncoordinated mutations so that so typically, only approximate tandem copies are present. Tandem repeats are ubiquitous sequence features in both prokaryotic and eukaryotic genomes. These are highly useful as genetic markers. They are codominant, occur in high frequency, and appear to be distributed through out the genomes of most, not all the higher plants and animals. They also display a high level of polymorphism, even among closely related accessions, and are amenable to simple and inexpensive Polymerase Chain Reaction **(PCR)** assays (Brown et. al. 1996).

Tandem repeats are usually classified among satellites (spanning megabases of DNA, associated with heterochromatin), minisatellites (repeat units in the range 6-100 bp, spanning hundreds of base pairs) and microsatellites (repeat units in the range 1-5 bp, spanning a few tens of nucleotides). The minisatellites are also called "various number tandem repeats" or **VNTRs**. The microsatellites are also called "short tandem repeats" or **STRs**. (Short tandem repeats, **STRs**) contain 2-5 bp repeats **VNTR 'S** are scattered at various locations in the Genome are regions that are highly variable. These regions contain a type of **DNA** sequence called variable number Tandem Repeat. Tandem repeats are multiple copies of sequence of base pairs arranged in head to tail fashion. For example, a frequently found Tandem Repeat is **CA,** and one strand containing this type of repeat reads **CACACA...** notated as **$(CA)_n$** the other strand would read

GTGTGT.... In this example, the number of repeating base pairs is two, but it can be more. When the repeating unit is less than four, the **VNTR** is called a **MICROSATELLITE** and when the repeating unit is longer it is a **MINISATELLITE.**

MICROSATELLITE are DNA regions with variable numbers of short tandem repeats flanked by a unique sequence. Microsatellites make good genetic Markers because they each have many different 'alleles'-i. e. There can be many different lengths of the repeat region. The number of repeats there are at the same location defines an allele. With many alleles, most individuals are heterozygous, giving power to note association between marker allele and performance in progeny inheriting a favorable linked QTL allele.



Tandemly repeated sequences are especially liable to undergo misalignments during chromosome pairing, and the size of tandem clusters tends to be highly polymorphic. The smaller clusters of this simple sequence can be used to characterize individual genomes in the technique of **"DNA** finger printing". Comparisons of corresponding regions of simple sequence DNA with in and between species are informative about the mechanisms involved in manipulating sequences

1.2 Different Tools For Repeat Finder

1.2.1 censor_xxx_humrep [censor]

This tool searches for genome wide repeats and classifies its findings. Each incarnation of this tool produces two tables containing ALU (xxx=ALU) and non ALU (xxx=NONALU) repeats. In this incarnation the ``Humrep" repeat database [humrep] is used and the search sensitivity is set to ``moderate". If the input sequence was derived from other mammalians than human the ``MamRep" repeat database - which comes along with CENSOR too - is used.

1.2.2 repeatmasker_xxx primate [repmask]

This tool searches for both genome wide and simple repeat and classifies its findings. Each incarnation of this tool produces three tables containing ALU (xxx=ALU), non ALU (xxx=NONALU) and simple (xxx=REPEAT) repeats. In this incarnation the ``Primate'' repeat database is used. The search sensitivity is set to ``high'' and tagging of simple repeats is turned on. If the input sequence was derived from other mammalians than human the ``Rodent'' or ``Vertebrate'' repeat database - which comes along with RepeatMasker too - is used.

1.2.3 sst_xxx_humrep [sst]

This tool searches for both genome wide and simple repeats and classifies its findings. Each incarnation of this tool produces three tables containing ALU (xxx=ALU), non ALU (xxx=NONALU) and simple (xxx=REPEAT) repeats. The ``HumRep'' repeat database [humrep] is used.

1.2.4 xnun_repeat_default [xnun]

This tool searches for short-period repeats (micro-satellites) and masks its findings. In order to reduce the rate of false positives the default probability cut is lowered from 1 % to 0.1 %.

1.2.5 tandem_xxx_default [tandem]

This tool searches for short-period repeats (micro-satellites) and reports the consensus repeat unit. Each incarnation of this tool produces two tables. The first table (xxx=STRAND) contains all repeats found together with the consensus repeat unit and the strand information. The second table (xxx=REGION) contains just the repeat regions and is cross-linked with the first table.

1.2.6 inVerted_xxx_default [inverted]

This tool searches for inverted repeats within a window of 2 *kbp*. The window size is an internally fixed parameter of the analysis tool itself. Each incarnation of this tool produces two tables. The first table (xxx=STRAND) contains two entries for each

inverted repeats. One for the forward and one for the reverse strand repeat unit. The second table (xxx=REGION) contains just the repeat regions and is cross-linked with the first table.

1.3 *Tandem Repeats Finder*

A Tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides. Tandem Repeats Finder is a program to locate and display Tandem Repeats in DNA sequences. In order to use this program, we have to submit the sequence in **FASTA** format. There is no need to specify the pattern, the size of the pattern or any other parameter. The program's analysis is sent back as two files, a summary table file and an alignment file. The summary table contains information about each repeat, including its location, size, number of copies and nucleotide content. Clicking on the location indices for one of the table entries opens a second web browser that shows an alignment of the copies against a consensus pattern. The program is very fast, analyzing sequences on the order of .5Mb in just a few seconds. Submitted sequences may be up to 5Mb in length. Repeats with pattern size in the range from 1 to 500 bases are detected.

1.4 Levels Of Tandem Repeat Finder:

There are 3 **LEVELS** of tandem repeat finders.

1.4.1 Basic

It uses default parameters (recommended for beginners.)

1.4.2 Intermediate

It provides the parameter Maximum period size, and options Flanking sequence and Masked Sequence File.

1.4.3 Advanced

It provides the parameters like Alignment parameters (match, mismatch and indels), Minimum Alignment Score To Report Repeat, Maximum Period Size and options like Flanking sequence, Masked sequence file, Data file.

File Edit View Favorites Tools Help

TANDEM REPEATS FINDER

Submitting Your Sequence

Please Select one of the following options:

Use default parameters (recommended for beginners)

Set some parameters and select options

Set all parameters and options

Last revised July 22, 1999
Send any questions or comments to

TANDEM REPEAT FINDER OPTION PAGE

1.5 Advanced Tandem Repeat Findeer Program Parameters:

Input to the program consists of a sequence file and the following parameters:

1.5.1 Alignment Parameters

Weights for match, mismatch and indels. Lower weights allow alignments with more mismatches and indels. Match weight is +2 in all options here. Mismatch and indels weights [interpreted as negative numbers] are 3, 5, or 7. A 3 is more permissive and a 7 is less permissive of these types of alignment choices.

1.5.2 Minimum Alignment Score

The alignment score must meet or exceed this value for the repeat to be reported.

1.5.3 Maximum Period Size

The period size must be no larger than this value for the repeat to be reported. The program will find all repeats with period size between 1 and 500, but the output table can be limited to some other range.

1.5.4. Detection parameters

Matching probability P_m and indeal probability P_i. P_m =.80 and P_i =.10 by default and it cannot be modified in this version of the program.

1.5.5 Options

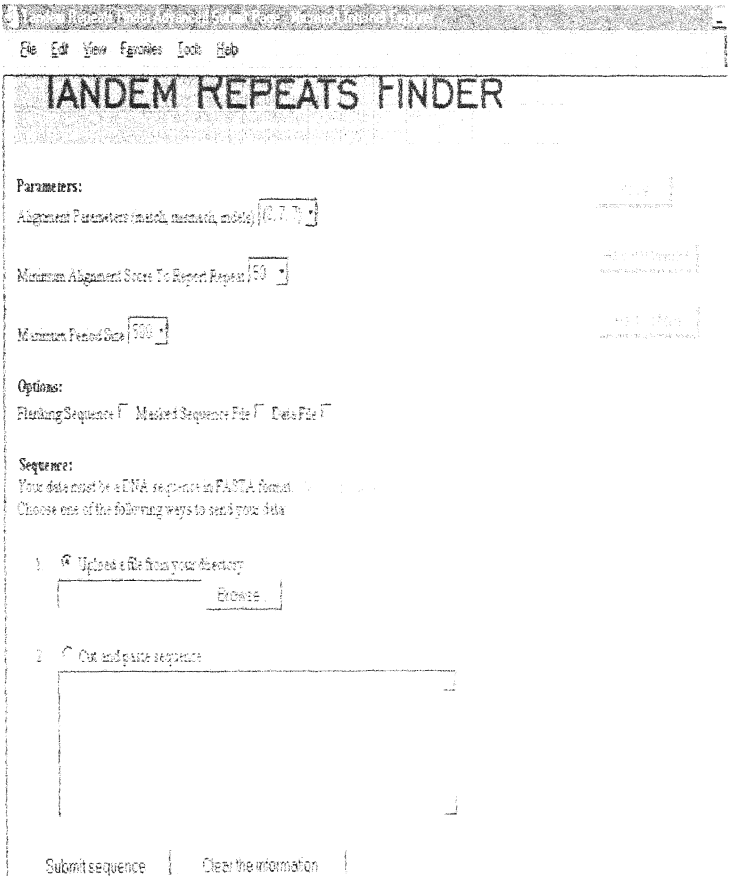
1.5.5.1 Flanking sequence

Flanking sequence consists of the 200 nucleotides on each side of a repeat. Flanking sequence is recorded in the alignment file. This may be useful for PCR primer determination.

1.5.4.2 Masked sequence File

The masked sequence file is a FASTA format file containing a copy of the sequence with every character that occurred in a tandem repeat changed to the letter 'N'. The word "masked" is added to the sequence description line just after the '>' character.

1.5.4.3 Data File: The data file is a text file, which contains the same information, in the same order, as the summary table file, plus consensus sequences. This file contains no labeling and is suitable for additional processing, for example with a perl script, outside of the program.



TANDEM REPEAT FINDER ADVANCED SUBMIT PAGE.

1.6 Procedure For Finding Tandem Repeats:

Web sites like http://c3.biomath.mssm.edu/trf.advanced.submit.html are used to find out the tandem repeats of particular sequence (source sequence). This web site is provided by The Department of Biomathematical sciences, Mount sinai School of Medicine.

- Down load the source sequence from the www.ncbi.nlm.nih/gov/entrez
- Open the page http://c3. Biomath.mssm.edu/trf.advanced.submit.html
- Enter the source sequence in **cut and paste sequence** blank.
- Click the **submit sequence** button.

1.7 Entering The Sequence For Finding Tandem Repeats:

1.7.1 Sequence:

Data must be a DNA sequence in FASTA format.

1.7.2 FASTA format:

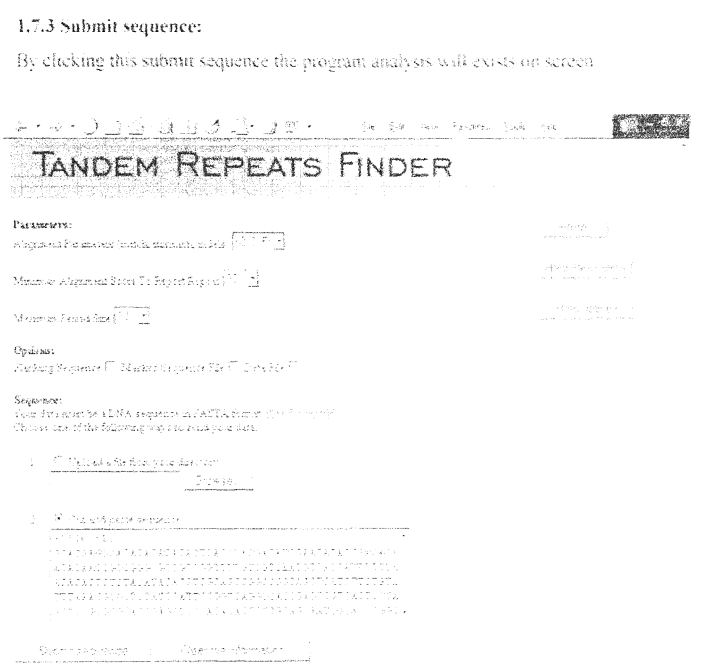
The FASTA format looks like this,

gi|18063613|gb|BM324719.1|BM324719 PIC1_34_G05.b1_A002 Pathogen-infected compatible 1 (PIC1) Sorghum bicolor cDNA, mRNA sequence
GCACGAGGCACACACACACACTGACGTACGATATCGAACACACTGACACAA
CACAACCGCCGGGCGCCGCCGGCTTTGTCGTCAACGTCACCATTCCTTAACA
CACTCTCTACATACACGTCGCAGCCGGCCGGCAGCTCATCTTCTGTATTTAA
ACGCGCTCCATCCATTCCGGTCAGGGCACGGAGTCGTCACTCCTAGCTCCGC
GCGCACCGAGCTCCCACACACTCTCGCAGCAATGGCACGCGGCGGCTCTACC
GCGGCTCCTCCACCCCTCCCGCCGCTGCTGCTGCTGCTGGTGGTAGTAGTGG
TGTCCTCGGCCTTTTTCGCTCGCTGCTGCCACGCCGCGCCGCCGGGTGCACTG
GTGACGAGCGTCCCGGGGTACACGGCGGGCAGCGGCGCGCTGCCGTCGAAG
CACTACGCCGGGTACGTGACGGTGGACGAGGCGCACGGCAGGAGGCTCTTC
TACTACCTGGTGGAGCCGAGCGTGACCCCGCCAAGGACCCCGTCGTGCTGTG
GCTCAACGGCGGGCCTGGCTGCTCCAGCTTCGACGGC

The first line starts with a greater than sign ">" and contains a name or other identifier for the sequence. The remaining lines contain the sequence data. The sequence can be in upper or lower case letters. Anything other than letters (numbers for example) is ignored.

1.7.3 Submit sequence:

By clicking this submit sequence the program analysis will exists on screen

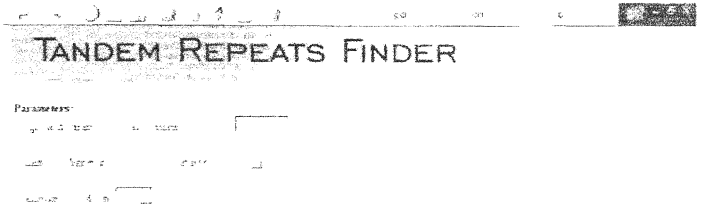


SUBMISSION OF SEQUENCE TO ADVANCED TANDEM REPEAT FINDER

The first line starts with a '>' (greater than) sign and contains a name or other identifier for the sequence. The remaining lines contain the sequence data. The sequence can be in upper or lower case letters. Anything other than letters (numbers or example) is ignored.

1.3 Submit sequence

By clicking this submit sequence the program then tests with various options

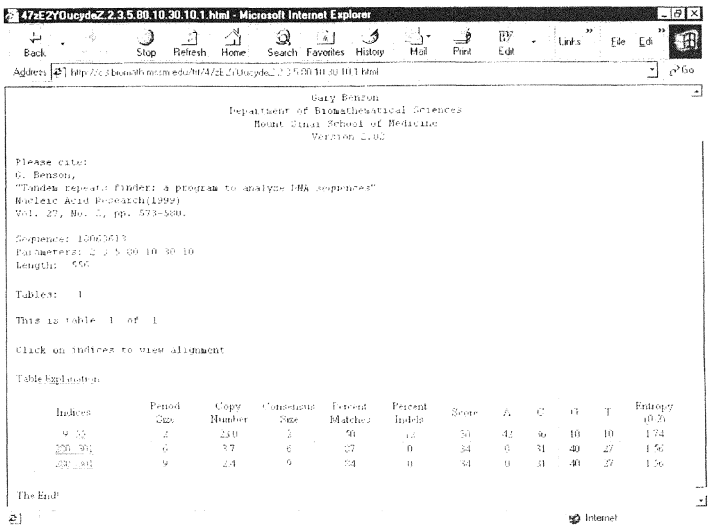


SUBMISSION OF SEQUENCE TO ADVANCED TANDEM REPEAT FINDER

1.8 Table Explanation:

The summary table includes the following information:

1. Indices of the repeat relative to the start of the sequence.
2. Period size of the repeat.
3. Number of copies aligned with the consensus pattern.
4. Size of consensus pattern (may differ slightly from the period size).
5. Percent of matches between adjacent copies overall.
6. Percent of indels between adjacent copies overall.
7. Alignment score.
8. Percent composition for each of the four nucleotides.
9. Entropy measure based on percent composition.
10. If the output contains more than 140 repeats, multiple linked tables are produced.



SUMMARY TABLE

1.9 Alignment Explanation:

The alignment is presented as follows:

1. In each pair of lines, the actual sequence on top and a consensus sequence for all the copies is on the bottom.
2. Each pair of lines is one period except very small patterns.
3. The 10 sequence characters before and after a repeat is shown.
4. Symbol * indicates a mismatch.
5. Symbol - indicates an insertion or deletion.
6. Statistics refers to the matches, mismatches and indels overall between adjacent copies in the sequence, not between the sequence and consensus pattern.

7. Distances between matching characters at corresponding positions are listed as distance, number at that distance, percentage of all matches.
8. A, C, T, G count is percentage of each nucleotide in the repeat sequence.

1.10 Primer Design

1.10.1 Introduction:

Designing **PCR** and sequencing primers are essential activities for molecular biologists around the world. Primer design was developed to find suitable primers for PCR or oligo nucleotides for probes and **DNA** sequencing. Primer design is crucial for the success of PCR. Inappropriate primers cause low yield and misinterpretation. Primers that bind to multiple DNA loci can synthesize side products and render sequencing illegible, especially with high amplification of small amounts of DNA and with impure DNA. They are generally the result of short DNA sequence repeats. An ideal primer should only bind to a unique sequence. To ensure this the given sequence must be compared with itself to identify repeats.

Primer Design is a DOS-program to choose primer for PCR or oligonucleotide probes. Napiwotzki, J. and Becker, A. wrote this program in 1995. It is tailored to check known sequences for repeats and unique sequences and subsequently to create proper primers according to this data.

1.11 Primer Design Programs:

- **PrimerGen** searches strings of amino acid residues in order to reverse-translate oligonucletide primers of a desired range of lengths and maximum number of degeneracies.
- **Primer** (Stanford) Sun Sparcstations only
- **Primer** (Whitehead) Unix, Vms (and DOS and Mac if you can compile it)

- **Amplify**, Bill Engels (Macintosh only), is for use in designing, analyzing, and simulating experiments involving the polymerase chainreaction (PCR). You can obtain your copy of Amplify here
- **PrimerDesign 1.04**, a DOS-program to choose primer for PCR or oligonucleotide probes. See also the PrimerDesign Welcome Page.
- **PC-Rare**, a new software by R. Griffais, which uses a rare octamer at the 3' termini of the primer. This powerful (but user friendly) software is available for Macintosh and Windows environment.
- **Primer Design**, a Java applet by Luca Ida Giovanni TOLDO.
- **CODEHOP**, PCR primers designed from protein multiple sequence alignments (Local mirror site at WIS).
- **Primer3**, an online service to pick PCR primers from nucleotide sequence. (Local mirror site at WIS).
- **NetPrimer** (PREMIER Biosoft International). NetPrimer combines the latest primer design algorithms with a web-based interface allowing the user to analyze primers over the internet

1.12 Primer Design Considerations:

One of the single most important factors in successful automated **DNA** sequencing is proper primer design. It is important that a primer has the following characteristics:

1. Primers should be at least **18-20 nucleotides** in length to minimize the chance of encountering problems with a secondary hybridization site on vector or insert.
2. Primers with long runs of a single base should generally be avoided. It is especially to avoid **3** or more G's or C's in a row.
3. For cycle sequencing, primers with melting temperatures above **55°C** are generally produce better results than primers with lower melting temperatures.

4. Primers should have a **G/C** content between **40** and **60** percent. For primers **GC** content of less than **50%,** it may be necessary to extend the primer sequence beyond 18 bases to keep the melting temperature above the recommended lower limit of **55°C**.
5. Primers should be "stickier" on their **5'** ends than on their **3'** ends. A "sticky" 3' end as indicated by a high **G/C** content could potentially anneal at multiple sites on the template.
6. **"G"** or **"C"** is desirable at the **3'** end.
7. Primers should not contain complementary (palindromes) within themselves, which is they should not form hairpins. If this state exists a primer will fold back on itself and result in an unproductive priming event, which decreases the overall signal obtained.
8. Primers should not contain sequences of nucleotides that would allow one primer molecule to anneal to it self or to the other primer used in a **PCR** reaction (primer dimer formation).
9. If possible, run a computer search against the vector and insert **DNA** sequences to verify that the primers, and especially the 8-10 bases of its 3' end, are unique.
10. Do not design degenerate primers. Do not request inosine in sequencing primers. They either do not work or give poor cycle sequencing results.

1.13 Features Of Primer Design:

- Creating of new **primer pairs.**
- Creating one suitable primer to a given primer.
- Finding of **repeats** within a sequence.
- Finding of **unique** sequences within a sequence.
- Handling of sequences up to **32,000bp.**

1.14 Limits Of Primer Design:

- The sequence length, which can be used for primer design, repeat and unique search is limited to **32,000bp.**

- Maximal **16000** repeats can be found and sorted.
- Primer combinations can be explorated up to **6000 pairs.**

1.15 Primer Design Parameters:

The most critical parameter for successful PCR is the design of primer. The primer sequence determines several things such as the length of the product, it's melting temperature and ultimately the yield. A poorly designed primer can result in little or no product due to non-specific amplification or primer dimer formation, which can be competitive enough to suppress to product formation. The main parameters for primer design are:

1.15.1 primer length

Specificity, the temperature and the time of annealing are at least partly dependent on primer length. Length of primer is critical parameter in primer designing. It is also proportional to annealing efficiency. For general studies primers of typically 17-34 nucleotides in length are the best primers, <16 nucleotides generally anneal non-target DNA sequences. Longer primers are required when a perfect complementary sequence to the entire template is not found.

1.15.2 primer sequence

a) **3' self/cross complementarity** – 3 or more bases of complementarity at the 3' end can give a strong primer dimer and reduce PCR efficiency.
b) **Runs of single bases** – avoid long runs of single bases (i.e.5 or more).
c) **Internal self-complementarity** - We should check and avoid sequences of internal self-complementarity otherwise these regions can form hairpin loops.
d) **Internal sequence structure** – Primer stability is affected by nearest neighbour interactions between bases. Consequently it is better to have Gs and Cs together in pairs rather than dispersed within As and Ts.

1.15.3 GC content:

Melting temperature and annealing temperature are strictly dependent on each other. GC is an important characteristic of DNA and provides information about the strength of annealing (GC have hydrogen bonds between them). The base composition of primer should be between 45% and 55% of GC. The primer sequence must be chosen such that there is no poly G or poly C stretches that can promote non-specific annealing. Poly A, poly T stretches also to be avoided as these will breath and open up stretches of primer template complex.

1.15.4 Melting temperature:

a) **Matching the primer for melting temperature** – the stability of a primer-template DNA duplex can be measured by its melting **temperature(T_m).** The T_m of any duplex of DNA is defined as the temperature at which **50%** is dissociated as **ssDNA.** T_m is affected by sequence, length, p^H, salt concentration etc. The T_m of any primer pair (or multiplexed set) should be reasonably well matched otherwise amplification efficiency may be compromised.

There are many formulas used for calculating T_m the most accurate account of nearest neighbour interactions and are best left to computer programs.

$T_m (^0 C) = 2(A+T) + 4(G+C)$

This formula becomes inaccurate for longer primers.

Alternatively another more versatile formula is as follows: -

$T_m (^0 C) = 69.3 + 0.41(\%G+C) - (650/L)$

%G + C= %age GC content of the primer

L = length of primer (bp)

After calculating the T_ms of a primer pair they differ by too much it is best to add further bases to the less stable of the pair from the 5' end.

b) **Choosing an annealing temperature for a primer set** – the first time a pair of primers is used it is best to attempt an annealing temperature 5^0 C below the average T_m for the primer pair. Annealing temperature is then adjusted depending on the preliminary results or alternatively a Mg^{2+} titration can be useful especially where the product size is large or the GC content is high.

1.15.5 Secondary structure formation:

An important factor to consider when designing a primer is the presence of secondary structures. Presence of secondary structures greatly reduces the number of primer molecules available for the reaction. The presence of hairpin loops reduces the efficiency by limiting the ability to bind to the target site. The single stranded nucleic acid sequences may have secondary structure due to the presence of complementary sequences in the primer sequence. Secondary structure formation occurs mainly because of repetitive sequence.

1.15.6 Specificity:

Primer specificity is partly dependent on primer length. Primers must be chosen so that they have a unique sequence with in the template DNA that is to be amplified. A primer designed with a high repetitive sequence will result a smear when amplifying genomic DNA. How ever the same primer may give a single band if clone from a genomic library is amplified. Because Taq DNA Polymerase is active over a broad range of temperature, primer extension will occur at the lower temperatures of annealing. If the temperature is too low, non-specific priming may occur which can be extended by the Polymerase if there is a short homology at the 3' end. In general, a melting temperature of 55^0 C to 72^0 C gives the best results.

1.15.7 Primer ends:

The 3' terminal position in PCR primers is essential for the control of mis-priming. Another variable to look at is the inclusion of a G or C residue at the 3' end of primer. This

will helps to improve the efficiency of the reaction by minimizing local unwind of DNA double helix. This occurs because of AT rich sequence.

1.16 PRIMER3

To design primers for a region of interest, Genotator i.e.*Primer3* is used. The development of *Primer3* and the Primer3 WWW interface were funded by Howard Hughes Medical institute and by the National Institutes of Health, National Human Genome Research Institute, under grants ROI – HG00257 and P50-HG0098.

Primer 3 started as a reimplementation of *Primer .5* as software component; the design of *Primer 3* draws heavily on the design of *Primer .5* and Primer v2 and WWW interface designed by Richard Resnick for *Primer v2.*

Primer 3 is a computer program that suggests PCR primers for a variety of applications.

a) To create STS (sequence tagged sites).
b) To amplify sequences for single nucleotide polymorphism discovery.
c) To select single primers for sequencing reactions.
d) Do design oligo nucleotide hybridization probes.

1.17 Procedure For Primer Design By Using Primer3:

- Select the query sequence and Id, which has the Tandem repeats in i
- Paste source sequence in **FASTA** format.
- Paste the sequence Id number in the sequence Id blank.
- Then put the tandem repeats in brackets [ ], which are present in the source sequence.
- Then adjust parameters according to our requirement.
- Then click the **Pick Primers** option.
- Then it shows the results as output.

1.18 Primer3 Input Parameters:

1.18.1 Source Sequence

The sequence from which to select primers or hybridization oligos.

1.18.2 Sequence Id

An identifier that is reproduced in the output to enable the user to identify the chosen primers.

1.18.3Targets

If one or more Targets are specified then a legal primer pair must flank at least one of them. A Target might be a simple sequence repeat site (for example a CA repeat) or a single-base-pair polymorphism.

1.18.4 Excluded Regions

Primer oligos may not overlap any region specified in this tag. The associated value must be a space-separated list of

Start, length

Pairs where *start* is the index of the first base of the excluded region, and *length* is its length.

E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. e.g. ...ATCT<CCCC>TCAT. Forbids primers in the central CCCC.

1.18.5 Product Size

Minimum, Optimum, and Maximum lengths (in bases) of the PCR product. Primer3 will not generate primers with products shorter than Min or longer than Max, and with default arguments Primer3 will attempt to pick primers producing products close to the Optimum length,

1.18.6 Number To Return

The maximum number of primer pairs to return. Primer pairs returned are sorted by their "quality", in other words by the value of the objective function (where a lower number indicates a better primer pair). Setting this parameter to a large value will increase running time.

1.18.7 Max 3' Stability

The maximum stability for the five 3' bases of a left or right primer. Bigger numbers mean more stable 3' ends.

1.18.8 Max Mispriming

The maximum allowed weighted similarity with any sequence in Mispriming Library. Default is 12.

1.18.9 Pair Max Mispriming

The maximum allowed sum of similarities of a primer pair (one similarity for each primer) with any single sequence in Mispriming Library. Default is 24

1.18.10 Primer Size

Minimum, Optimum, and Maximum lengths (in bases) of a primer oligo. Primer3 will not pick primers shorter than Min or longer than Max, and with default arguments will attempt to pick primers close with size close to Opt. Min cannot be smaller than 1. Max cannot be larger than 36. (This limit is governed by maximum oligo size for which melting-temperature calculations are valid.) Min cannot be greater than Max.

1.18.11 Primer T_m

Minimum, Optimum, and Maximum melting temperatures (Celsius) for a primer oligo. Primer3 will not pick oligos with temperatures smaller than Min or larger than Max, and with default conditions will try to pick primers with melting temperatures close to Opt.

1.18.12 Maximum T_m Difference

Maximum acceptable (unsigned) difference between the melting temperatures of the left and right primers.

1.18.13 Product T_m

The minimum, optimum, and maximum melting temperature of the amplicon. Primer3 will not pick a product with melting temperature less than min or greater than max.

$$T_m = 81.5 + 16.6(\log_{10}([Na+])) + .41*(\%GC) - 600/length,$$

where [Na+] is the molar sodium concentration, (%GC) is the percent of Gs and Cs in the sequence, and length is the length of the sequence.

1.18.14 Primer GC%

Minimum, Optimum, and Maximum percentage of Gs and Cs in any primer.

1.18.15 Max Complementarity:

The maximum allowable local alignment score when testing a single primer for (local) self-complementarity and the maximum allowable local alignment score when testing for complementarity between left and right primers. For example, the alignment

```
5' ATCGNA 3'
   ||| |
3' TA-CGT 5'
```

is allowed (and yields a score of 1.75), but the alignment

5' ATCCGNA 3'

|| ||

3' TA--CGT 5'

is not considered. Scores are non-negative, and a score of 0.00 indicates that there is no reasonable local alignment between two oligos.

1.18.16 Max 3' Complementarity:

The maximum allowable 3'-anchored global alignment score when testing a single primer for self-complementarity, and the maximum allowable 3'-anchored global alignment score when testing for complementarity between left and right primers. The 3'-anchored global alignment score is taken to predict the likelihood of PCR-priming primer-dimers, for example

5' ATGCCCTAGCTTCCGGATG 3'

||| |||||

3' AAGTCCTACATTTAGCCTAGT 5'

or

5' AGGCTATGGGCCTCGCGA 3'

||||||

3' AGCGCTCCGGGTATCGGA 5'

The scoring system is as for the Max Complementarity argument. In the examples above the scores are 7.00 and 6.00 respectively. Scores are non-negative, and a score of 0.00 indicates that there is no reasonable 3'-anchored global alignment between two oligos. In order to estimate 3'-anchored global alignments for candidate primers and primer pairs, Primer assumes that the sequence from which to choose primers is presented 5'->3'. It is nonsensical to provide a larger value for this parameter than for the Maximum (local) Complementarity parameter because the score of a local alignment will always be at least as great as the score of a global alignment.

1.18.17 Max Poly-X:

The maximum allowable length of a mononucleotide repeat, for example AAAAAA.

1.18.18 Included Region:

A sub-region of the given sequence in which to pick primers. For example, often the first dozen or so bases of a sequence are vector, and should be excluded from consideration. The value for this parameter has the form

start,length

where *start* is the index of the first base to consider, and *length* is the number of subsequent bases in the primer-picking region.

1.18.19 Start Codon Position:

This parameter should be considered EXPERIMENTAL at this point. some erroneous inputs might cause an error in Primer3. Index of the first base of a start codon. This parameter allows Primer3 to select primer pairs to create in-frame amplicons.

1.18.20 Mispriming Library:

This selection indicates what mispriming library (if any) Primer3 should use to screen for interspersed repeats or for other sequence to avoid as a location for primers.

1.18.21 CG Clamp:

Require the specified number of consecutive Gs and Cs at the 3' end of both the left and right primer. (This parameter has no effect on the hybridization oligo if one is requested.)

1.18.22 Salt Concentration:

The millimolar concentration of salt (usually KCl) in the PCR. Primer3 uses this argument to calculate oligo melting temperatures.

1.18.23 Annealing Oligo Concentration:

The nanomolar concentration of annealing oligos in the PCR. Primer3 uses this argument to calculate oligo melting temperatures.

1.18.24 Max Ns Accepted:

Maximum number of unknown bases (N) allowable in any primer.

1.18.25 Liberal Base:

This parameter provides a quick way to get Primer3 to accept IUB / IUPAC codes for ambiguous bases (i.e. by changing all unrecognized bases to N.

1.18.26 First Base Index:

The index of the first base in the input sequence. For input and output using 1-based indexing (such as that used in GenBank and to which many users are accustomed) set this parameter to 1. For input and output using 0-based indexing set this parameter to 0. (This parameter also affects the indexes in the contents of the files produced when the primer file flag is set.) In the WWW interface this parameter defaults to 1.

1.18.27 Inside Target Penalty:

Non-default values valid only for sequences with 0 or 1 target regions. If the primer is part of a pair that spans a target and overlaps the target, then multiply this value times the number of nucleotide positions by which the primer overlaps the (unique) target to get the 'position penalty'. The effect of this parameter is to allow Primer3 to include overlap with the target as a term in the objective function.

1.18.28 Outside Target Penalty:

Non-default values valid only for sequences with 0 or 1 target regions. If the primer is part of a pair that spans a target and does not overlap the target, then multiply this value times the number of nucleotide positions from the 3' end to the (unique) target to get the

'position penalty'. The effect of this parameter is to allow Primer3 to include nearness to the target as a term in the objective function.

1.18.29 Sequence Quality

A list of space separated integers. There must be exactly one integer for each base in the Source Sequence if this argument is non-empty. High numbers indicate high confidence in the base call at that position and low numbers indicate low confidence in the base call at that position.

1.18.30 Min Sequence Quality:

The minimum sequence quality (as specified by Sequence Quality) allowed within a primer.

1.18.31 Min 3' Sequence Quality:

The minimum sequence quality (as specified by Sequence Quality) allowed within the 3' pentamer of a primer.

1.18.32 Sequence Quality Range Min:

The minimum legal sequence quality (used for interpreting Min Sequence Quality and Min 3' Sequence Quality).

1.18.33 Sequence Quality Range Max:

The maximum legal sequence quality (used for interpreting Min Sequence Quality and Min 3' Sequence Quality).

1.18.34 Penalty Weights:

This section describes "penalty weights", which allow the user to modify the criteria that Primer3 uses to select the "best" primers. There are two classes of weights: for some parameters there is a 'Lt' (less than) and a 'Gt' (greater than) weight. These are the weights

that Primer3 uses when the value is less or greater than (respectively) the specified optimum. The following parameters have both 'Lt' and 'Gt' weights:

- Product Size
- Primer Size
- Primer T_m
- Product T_m
- Primer GC%
- Hyb Oligo Size
- Hyb Oligo T_m
- Hyb Oligo GC%

For the remaining parameters the optimum is understood and the actual value can only vary in one direction from the optimum:

- Primer Self Complementarity
- Primer 3' Self Complementarity
- Primer #N's
- Primer Mispriming Similarity
- Primer Sequence Quality
- Primer 3' Sequence Quality
- Primer 3' Stability
- Hyb Oligo Self Complementarity
- Hyb Oligo 3' Self Complementarity
- Hyb Oligo Mispriming Similarity
- Hyb Oligo Sequence Quality
- Hyb Oligo 3' Sequence Quality

The following are weights are treated specially:

- Position Penalty Weight

- Determines the overall weight of the position penalty in calculating the penalty for a primer.
- Primer Weight
- Determines the weight of the 2 primer penalties in calculating the primer pair penalty.
- Hyb Oligo Weight

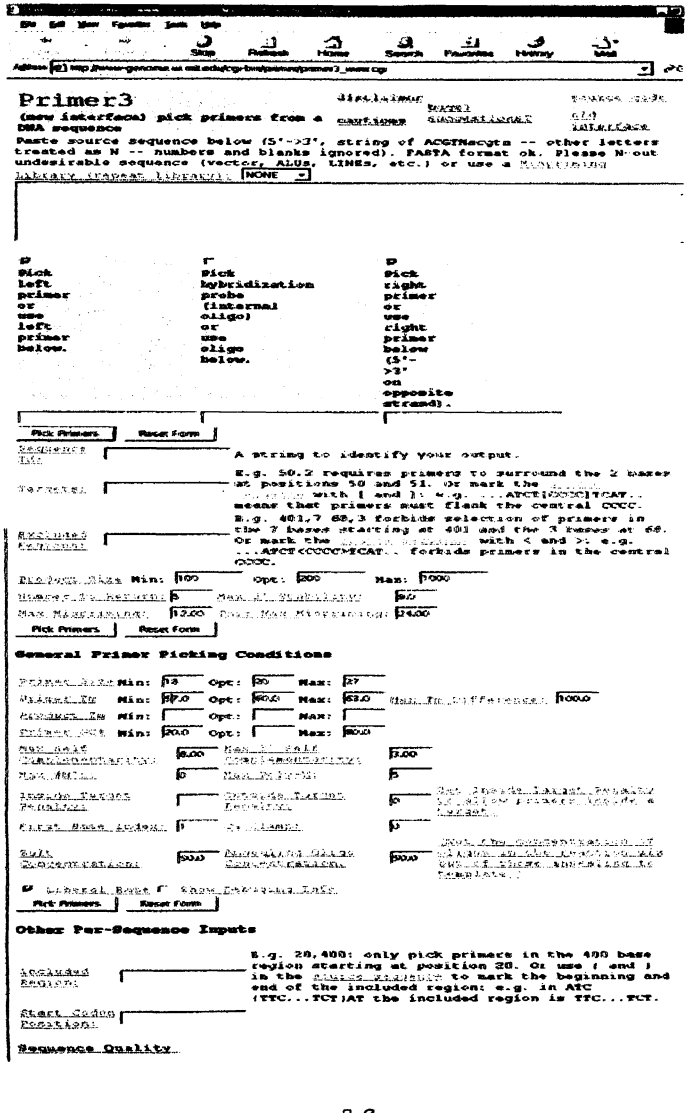
Determines the weight of the hyb oligo penalty in calculating the penalty of a primer pair plus hyb oligo.

The following govern the weight given to various parameters of primer pairs (or primer pairs plus hyb oligo).

- T_m difference
- Primer-Primer Complementarity
- Primer-Primer 3' Complementarity
- Primer Pair Mispriming Similarity

1.18.35 Hyb Oligos (Internal Oligos):

Parameters governing choice of internal oligos are analogous to the parameters governing choice of primer pairs.


Primer3
(new interface) pick primers from a DNA sequence
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
NONE
Pick Primers
Reset Form
A string to identify your output.
E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the
with [and]: e.g.
means that primers must flank the central CCCC.
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
with < and >: e.g.
...AFCTCCCCCMTCAT.. forbids primers in the central CCCC.
Min:
Opt:
Max:
Pick Primers
Reset Form
General Primer Picking Conditions
Min:
Opt:
Max:
Pick Primers
Reset Form
Other Per-Sequence Inputs
E.g. 20,400: only pick primers in the 400 base region starting at position 20. Or use < and > is the
to mark the beginning and end of the included region: e.g. in ATC (TTC...TCT)AT the included region is TTC...TCT.
Sequence Quality

Sequence Quality

Min Sequence Quality: 0 | Min First Sequence Quality: 0 | Sequence Quality Range: 0 | Sequence Quality Range Max: 100

Objective Function Penalty Weights for Primers

Tm: Lt 1.0 Gt 1.0
Size: Lt 1.0 Gt 1.0
GC%: Lt 0.0 Gt 0.0
Self Complementarity: 0.0
3' Self Complementarity: 0.0
#N's: 0.0
Mismatches: 0.0
Sequence Quality: 0.0
Hyb Sequence Quality: 0.0
Position Penalty: 0.0
End Stability: 0.0

Objective Function Penalty Weights for Primer Pairs

Product Size: Lt 0.05 Gt 0.05
Product Tm: Lt 0.0 Gt 0.0
Tm Difference: 0.0
Any Complementarity: 0.0
3' Complementarity: 0.0
Pair Mismatches: 0.0
Primer Penalty Weight: 1.0
Hyb Oligo Penalty Weight: 0.0

Pick Primers | Reset Form

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo Excluded Region:

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size: Min 18 Opt 20 Max 27
Hyb Oligo Tm: Min 57.0 Opt 60.0 Max 63.0
Hyb Oligo GC%: Min 20.0 Opt Max 80.0
Hyb Oligo Self Complementarity: 12.00 | Hyb Oligo Max 3' Self Complementarity: 12.00
Max #N's: 0 | Hyb Oligo Max Poly-X: 5
Hyb Oligo Mishyb Library: NONE | Hyb Oligo Max Mishyb: 12.00
Hyb Oligo Min Sequence Quality: 0
Hyb Oligo Salt Concentration: 50.0 | Hyb Oligo DNA Concentration: 50.0

Pick Primers | Reset Form

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo Tm: Lt 1.0 Gt 1.0
Hyb Oligo Size: Lt 1.0 Gt 1.0
Hyb Oligo GC%: Lt 0.0 Gt 0.0
Hyb Oligo Self Complementarity: 0.0
Hyb Oligo #N's: 0.0
Hyb Oligo Mishybing: 0.0
Hyb Oligo Sequence Quality: 0.0

Pick Primers | Reset Form

1.19 Methods For Finding Tandem Repeats

All available nucleotide sequences of sorghum from the NCBI (http://www.ncbi.nlm.nih/gov/entrez) were downloaded in FASTA format on to a local database.

1.19.1 Steps Involved After Downloading The Sequence Of Sorghum From NCBI.

1.19.1.1 Fasta Format

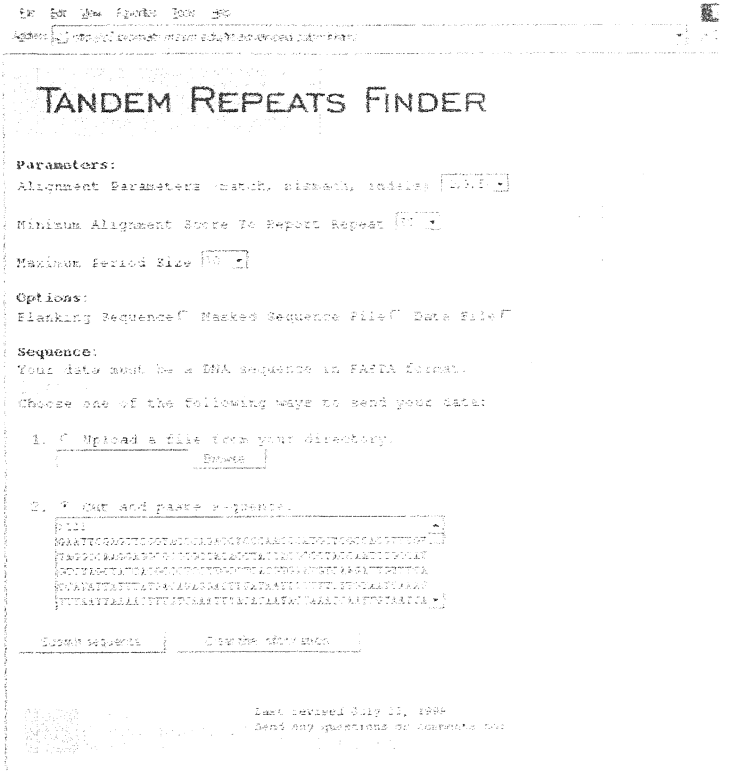
The FASTA format looks like this,

gi|18063613|gb|BM324719.1|BM324719 PIC1_34_G05.b1_A002 Pathogen-infected compatible 1 (PIC1) Sorghum bicolor cDNA, mRNA sequence
GCACGAGGCACACACACACACTGACGTACGATATCGAACACACTGACACAA
CACAACCGCCGGGCGCCGCCGGCTTTGTCGTCAACGTCACCATTCCTTAACAC
ACTCTCTACATACACGTCGCAGCCGGCCGGCAGCTCATCTTCTGTATTTAAAC
GCGCTCCATCCATTCCGGTCAGGGCACGGAGTCGTCACTCCTAGCTCCGCGC
GCACCGAGCTCCCACACACTCTCGCAGCAATGGCACGCGGCGGCTCTACCGC
GGCTCCTCCACCCCTCCCGCCGCTGCTGCTGCTGGTGGTAGTAGTGGTGT
CCTCGGCCTTTTTCGCTCGCTGCTGCCACGCCGCGCCGCCGGGTGCACTGGTG
ACGAGCGTCCCGGGGTACACGGCGGGCAGCGGCGCGCTGCCGTCGAAGCAC
TACGCCGGGTACGTGACGGTGGACGAGGCGCACGGCAGGAGGCTCTTCTACT
ACCTGGTGGAGCCGAGCGTGACCCCGCCAAGGACCCCGTCGTGCTGTGGCTC
AACGGCGGGCCTGGCTGCTCCAGCTTCGACGGC

1.19.1.2 Enter The Sequence For Finding Tandem Repeats.

Each of the sequence was searched for tandem repeats region and motif using the Tandem Repeat finder at http://c3.biomath.mssm.edu/trf.advanced.submit.html. This program was developed by the department of biomathematical sciences, Mount Sinai School of Medicine (http://www.mssm.edu/school.html). Advanced options were used to be find the tandem repeat region.

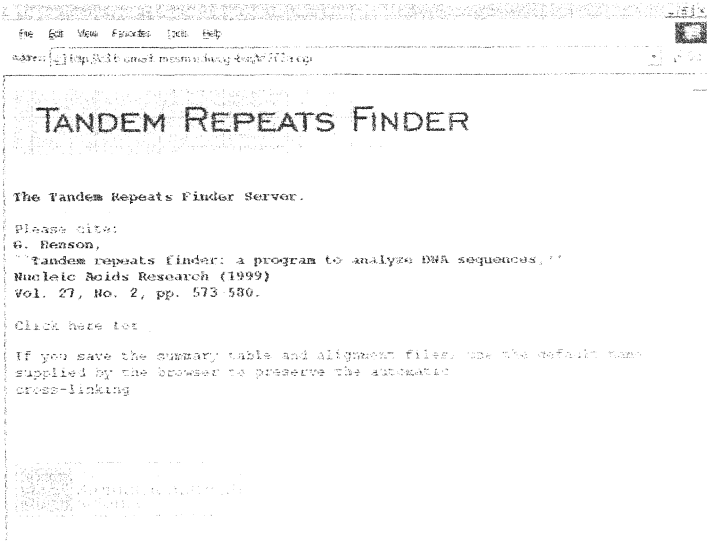
Parameters are set according the requirement. Alignment parameters (match, mismatch, indels) are set to (2,3,5). Minimum alignment score to report repeat is set to 30 and option period size to 10.



SUBMISSION OF SEQUENCE TO ADVANCE TANDEM REPEAT FINDER

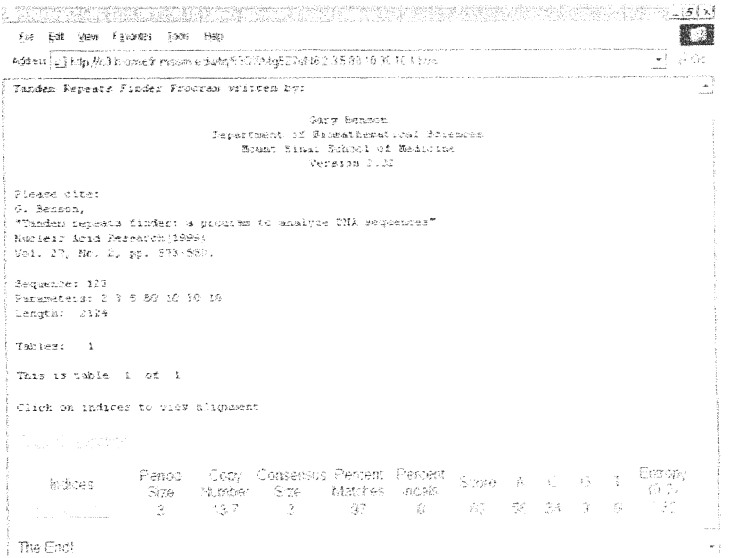
1.19.1.3 Submit Sequence

By clicking the submit sequence button which is on the advanced tandem repeat finder we will get program analysis of DNA sequence on screen



1.19.1.4 Tandem Repeats Report

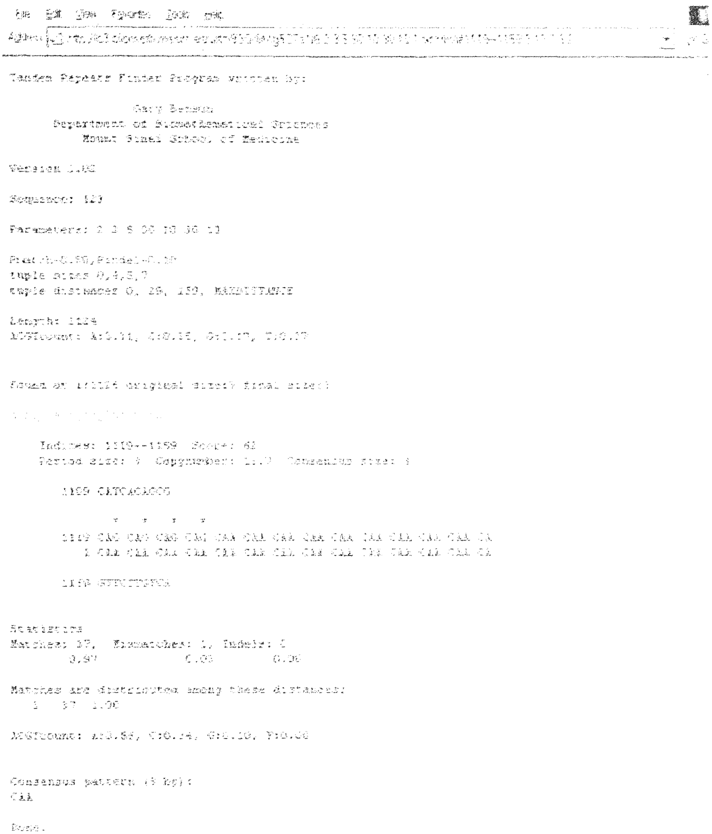
Click on the tandem repeats reports button on the program to analyze DNA sequence. We will get the summary table.



SUMMARY TABLE

1.19.1.5 Summary Table

By clicking the column of indices of table explanation, which is present in summary table, we will get the alignment explanation.



```
Tandem Repeats Finder Program written by:

                  Gary Benson
      Department of Biomathematical Sciences
          Mount Sinai School of Medicine

Version 3.02

Sequence: 123

Parameters: 2 3 5 80 10 50 10

Pmatch=0.80,Pindel=0.10
tuple sizes 0,4,5,7
tuple distances 0, 29, 159, MAXDISTANCE

Length: 1124
ACGTcount: A:0.31, C:0.15, G:0.17, T:0.37


Found at i:1124 original size:9 final size:3

    Indices: 1119--1159  Score: 62
    Period size: 3  Copynumber: 13.7  Consensus size: 3

      1109 CATCACAGCG

      1159 GTTCTTGTCA


Statistics
Matches: 37,  Mismatches: 1, Indels: 2
        0.97              0.03          0.05

Matches are distributed among these distances:
   3   37  1.00

ACGTcount: A:0.66, C:0.24, G:0.10, T:0.00


Consensus pattern (3 bp):
CAA

Done.
```

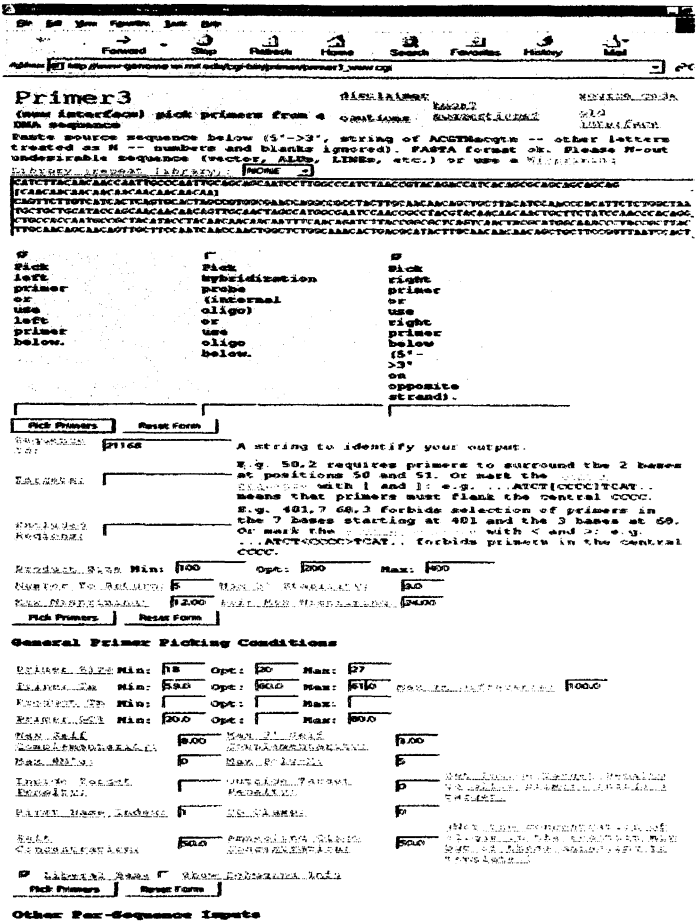
ALIGNMENT EXPLANATION TABLE

1.19.1.6 Alignment Explanation Table

Alignment explanation table explains the number of times the repeat is repeated. for Ex - In the above case the repeat CAA is repeated 9 times

1.19.1.7 Primer3

PRIMER3 program to used design the primers in sequences where repeat region were found. The sequence and its Id. Were pasted in the sequence text box in FASTA format. Tandem repeats regions were masked by placing them in brackets [ ]. The parameters were set according to the requirement. The selected options for product size Min imum100,Optimum 200 ,maximum 400 . the primer anealing temperature Minimum 59.0,optimum 60.O,Maximum 61.0 degree celcius. Primer GC% should be Minimum 20.0 and Maximum 80.0.primer size should be minimum 18bp,Optimum 20bp maximum27bp.Then click the **pick primer** option.



Primer3

(new interface) pick primers from a DNA sequence

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 Sequence Quality file.

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).

Pick Primers | Reset Form

A string to identify your output.

E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the ... with [and]: e.g. ...ATCT[CCCCTTCAT... means that primers must flank the central CCCC.

E.g. 401,7 60,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 60. Or mark the ... with < and >: e.g. ...ATCT<CCCC>TCAT... forbids primers in the central CCCC.

Product Size Min: 100 Opt: 200 Max: 400

Number To Return: 5 Max 3' Pr Tm Diff: 3.0

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: 59.0 Opt: 60.0 Max: 61.0 Max Tm Difference: 100.0

Primer GC% Min: 20.0 Opt: Max: 80.0

Max Self Complementarity: 8.00 Max 3' Self Complementarity: 3.00

Max Poly-X: 4 Max Pr Inside: 5

Excluded Region: Oligo Excluded Region: 0

Primer Num Ns Accepted: 1 Gc Clamp: 0

Salt Concentration: 50.0 Annealing Oligo Concentration: 50.0

Lowercase Masking Pick Primers | Reset Form

Other Per-Sequence Inputs

Included Region: E.g. 20,400: only pick primers in the 400 base region starting at position 20. Or use [and] in the Source Sequence to mark the beginning and end of the included region: e.g. in ATC[TTC...TCT]AT the included region is TTC...TCT.

Start Codon Position:

Sequence Quality

Sequence Quality

Min Sequence Quality:	0	Min End Sequence Quality:	0	Sequence Quality Range Min:	0	Sequence Quality Range Max:	100

Objective Function Penalty Weights for Primers

Tm Lt: 1.0 Gt: 1.0
Size Lt: 1.0 Gt: 1.0
GC% Lt: 0.0 Gt: 0.0

Self Complementarity	0.0
3' Self Complementarity	0.0
#N's	0.0
Mispriming	0.0
Sequence Quality	0.0
End Sequence Quality	0.0
Position Penalty	0.0
End Stability	0.0

Objective Function Penalty Weights for Primer Pairs

Product Size Lt: 0.05 Gt: 0.05
Product Tm Lt: 0.0 Gt: 0.0

Tm Difference	0.0
Any Complementarity	0.0
3' Complementarity	0.0
Pair Mispriming	0.0
Primer Penalty Weight	1.0
Hyb Oligo Penalty Weight	0.0

Pick Primers | Reset Form

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo Excluded Region:

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size: Min 18 Opt 20 Max 27
Hyb Oligo Tm: Min 59.0 Opt 60.0 Max 61.0
Hyb Oligo GC% Min: 20.0 Opt: Max: 80.0

Hyb Oligo Self Complementarity:	12.00	Hyb Oligo Max 3' Self Complementarity:	12.00
Max #Ns:	0	Hyb Oligo Max Poly-X:	5
Hyb Oligo Mishyb Library:	NONE	Hyb Oligo Max Mishyb:	12.00
Hyb Oligo Min Sequence Quality:	0		
Hyb Oligo Salt Concentration:	50.0	Hyb Oligo DNA Concentration:	50.0

Pick Primers | Reset Form

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo Tm Lt: 1.0 Gt: 1.0
Hyb Oligo Size Lt: 1.0 Gt: 1.0
Hyb Oligo GC% Lt: 0.0 Gt: 0.0

Hyb Oligo Self Complementarity	0.0
Hyb Oligo #N's	0.0
Hyb Oligo Mishybing	0.0
Hyb Oligo Sequence Quality	0.0

Pick Primers | Reset Form

1.19.1.7 0ut Put Of Primer3

Primer3 Output

```
No mispriming library specified
Using 1-based sequence positions
OLIGO                start  len       tm      gc%    any   3' GC_clamp seq
LEFT PRIMER           1088   20    59.96    55.00   4.00  3.00 GGCCCATCTAACCGTACAGA
RIGHT PRIMER          1285   20    59.50    45.00   7.00  1.00 TGCAACTGTTGTTGTTGCTG
SEQUENCE SIZE: 2124
INCLUDED REGION SIZE: 2124

PRODUCT SIZE: 198, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00
TARGETS (start, len)*: 1131,27

    1 GAATTCGAGCTCGGTACCCAGACCTCCCAACCCATGCTCGCCACGTTTGTTAGGCCAAGG

   61 AGGCGCCCGCCACACCTACCACGCCCTACCAATCCGCCATGTCTAGCTATCACGCCCTCC

  121 TTGCCTCACGTGAATGTCAAGATTGTTTCACCATATTATTTATGACAGAGGACTTGATAA

  181 TTTTTTTCTTGTAATCAAAGTTTAATTAAAACTTTGTCAAATTTACACAATACTAAAGCA

  241 ATTGTAATCAAGAACACAGGGAGTGCGTGTGATAATAGCTATAAGCATTCTAATTTGTAC

  301 ATTCTATTTGTGTGCATACATCTGTACATACTGGGATTTCAATTTGTCTTGATCTTGTAG

  361 CATTTTTCAATCATTGATGAACAACTTCATCTAACTACGTTGCAAGACAAATAGTACAGT

  421 AGTACAACAAAGTCCTTTGATAAAGGCTTTGATATACATGAGCAAGTCATAACTTTACTT

  481 GCACATCATGTCTGTAAAGTGGAACATTTCTGATGTGGCTAAGGCTATAACATGTGTAAA

  541 GGTGAAGTGATGTCACTCCTCATTTATCGAAAAGTTCCAATAGAAAATGACAACTTTTCC

  601 TTGTAAGTAGTGGAAATTGTCTTTCCTACACAGACCATATAATCCAATAAAATTGATAAC

  661 TAAATGTCAAAATCGACTAGGTGCCATGTCATCTATAGTTTATCTGTTGTTCGCAAAAGC

  721 CAAAATCTAAACAGATATCTATGAGCTCTCACTCATATAAATAGGCCCCAAATCAGTAGT

  781 TAAACCATCGCCCATAACATTGAGAGGAATTAGAAAAATACCAAGTGAAACGAACTAGCA
```

841 ACGTCCTTCCAACAATGGCTACCAAGATATTCGCCCTCCTTGCGCTCCATGCTCTTTTAG

901 TGAGCGGTACAACTGCGGCCATTATTCCACAGTGCTCACTTGCTCCTAATGCTATTATTC

961 CACAGTTCATCCCACCTGTTACTGCTTTAGGGAATGAACACCTAGCTGTGCAGGCCTATC

1021 CTGGACAGCAGGTGCTTTCGGCGAGCATCTTACAACAACCAATTGCCCAATTGCAGCAGC

1081 AATCCTTGGCCCATCTAACCGTACAGACCATCACAGCGCAGCAGCAGCAGCAACAACAAC
>>>>>>>>>>>>>>>>>>> **********

1141 AACAACAACAACAACAACAGTTCTTGTCATCACTCAGTGCACTAGCCGTGGCGAACCAGG

1201 CCGCCTACTTGCAACAACAGCTGCTTACATCCAACCCACATTCTCTGGCTAATGCTGCTG

1261 CATACCAGCAACAACAACAGTTGCAACTAGCCATGGCGAATCCAACCGCCTACGTACAAC
<<<<<<<<<<<<<<<<<<<

1321 AACAACTGCTTCTATCCAACCCACAGGCTGCCACCAATGCCGCTACATACCTACAACAAC

1381 AACAATTTCAACAGATCTTACCGGCGCTCAGTCAACTACGCATGGCAAACCCTACCGCTT

1441 ACTTGCAACAGCAACAGTTGCTTCCAATCAACCAACTGGCTCTGGCAAACACTGACGCAT

1501 ACTTGCAACAACAACAGCTGCTTCCGGTTAATCCACTGGTAGTTGCCAATCCATTAGTTG

1561 CCGCCTTCCTACAGCAACAACAATTGTCGTCATTCAACCAGATATCTTTGGTTAACCCTG

1621 CCTTGTCGTGGCAGCAACCCATCATTGGTGGTGCCATCTTCTAGATTACATATGAGTTGT

1681 TGTACTTGATAACAAAGCTCTCATACCGGCATGGGCAACTTTCCTAAAATAATCAATATA

1741 TTGGTTGAGATTTATTTGTCTATTTCGTAATTATGTTCTTCATATATGCGATTGAAAACA

1801 TCACATCATAATTAAAGACACATGCTTGGTTAATTTGTGGACAATAACATAATACTTCAT

1861 CAATCTAAAGATGTGCCTGCCCGACCTGAATATTCTAACTTGGGTGTGTGTGTTTTTATA

1921 CGATAGATGCACTATTGGAATCGCGCGCTTTGCCTAGTGTCTAGGGCACTCGGCAAACGG

1981 TAAAAAAAACACTCGGCAAACCGTTTGCTGGGTGCCACACTCGGAAAGCGGAGACAGCAA

2041 CAAAAATAACGGCAAAGCAAAACATAGTCAAATCAGATGCATGTATGGGGATCCTCTAGA

2101 GTCGACCTGCAGGCATGCAAGCTT

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

ADDITIONAL OLIGOS

1	LEFT PRIMER	1081	20	60.15	50.00	5.00	0.00	AATCCTTGGCCCATCTAACC
	RIGHT PRIMER	1285	20	59.50	45.00	7.00	1.00	TGCAACTGTTGTTGTTGCTG
	PRODUCT SIZE: 205, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00							
2	LEFT PRIMER	1080	20	59.39	50.00	5.00	0.00	CAATCCTTGGCCCATCTAAC
	RIGHT PRIMER	1285	20	59.50	45.00	7.00	1.00	TGCAACTGTTGTTGTTGCTG
	PRODUCT SIZE: 206, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 3.00							
3	LEFT PRIMER	1001	20	60.00	60.00	6.00	2.00	CCTAGCTGTGCAGGCCTATC
	RIGHT PRIMER	1188	20	59.06	55.00	8.00	3.00	CGGCTAGTGCACTGAGTGAT
	PRODUCT SIZE: 188, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00							
4	LEFT PRIMER	1002	20	60.00	60.00	6.00	2.00	CTAGCTGTGCAGGCCTATCC
	RIGHT PRIMER	1188	20	59.06	55.00	8.00	3.00	CGGCTAGTGCACTGAGTGAT
	PRODUCT SIZE: 187, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00							

Statistics

	con sid ered	too many Ns	in tar get	in excl reg	bad	no GC clamp	tm too low	tm too high	high any compl	high 3' compl	poly X	high end stab ok
Left	10645	0	0	0	448	0	5809	3518	5	22	14	58
771												
Right	8773	0	0	0	142	0	3498	4419	11	24	24	78
577												

Pair Stats:
considered 11174, unacceptable product size 10518, high end compl 117, ok 539
primer3 release 0.9

(primer3_www results.cgi v 0.2)

1.20 Results

All available (approximately 110055) nucleotide sequences were analyzed for the presence of tandem repeats up to 80bp (maximum) length of repeat motif and no penalty gaps or indels were allowed. Although such methods may not detect some repeats containing insertions or deletions, it was considered that such a conservative approach might not be affecting the results very dramatically. The summary of the total number of potential SSRs in each of the nucleotide length group is presented in table 1. As descried in the above procedure the nucleotide sequences are allowed to find tandem repeats by submitting them to Tandem repeat finder. Possible primers were found to these sequences which had repeat regions using Primer3. All the sequences with tandem repeats do not yield primers either due the sequence being short and less than minimum product size that is prescribed in the parameter or the tandem repeat regions could be skewed to the sides resulting in the loss as a flanking region. The results of Primer3 output were collated and entered into a database table, with right primer sequence, left primer temperature, right primer temperature and total sequence as shown in Table 2.

Of the total 9232 a little over (6232) were from tri-nucleotide groups.the tetra-nucleotide groups that potentially form the polypeptide repeats were the next largest groups followed by di-nucleotide and penta-nulceotide. The SSRs summary by various nucleotide groups for a minimum repeat size length of 10bp and the units that are part of the group, are presented in Table3. The largest group is the tri-nucleotide group 'CCG'consisting of 3312 accessions. Of these, the largest sub-group were from the 'GCG' and 'GCC' group which both code for alanine. In the di-nucleotide repeats the most common repeat was the 'AG'group which translated to di-oligopeptide of Arginine and Glutamine.

Table 1 . Summary of the repeat motif by nucleotide unit length of more than 10bp repeat length

Nucleotide unit length	Number of SSRs
Di	962
Tri	6232
Tetra	1468
Penta	569
Total	**9232**

Table 2

Id	Accessio	Details	Sequence	Repeat	Leftprimer	Rightprimer	Lefttem	Rightte	Totalsequence
1	6674175	AW284211	AAGAGGAGCCAAAGAA AAAGAAGAAGAAGGAG AAGAAGGCCAAAAAGG CCAAGAAGAAGAAGAA GTCCAAAAAGGATTGA GGTGGTCAGGCAAAG GCGGGGCAATAAAATC C	AAG(5)	GAGGAGCCA AAGAAAAAG AAGA	ATTTTATTGC CCCGCCTTT	59.13	60.63	112
2	7554456	AW680658	GCACGAGGATCGATCA GCGTGCATCCCTGACA AAAAAAAGCACACACA CACATACATACACCTA CTAGTCTTCGTGGCTG CCAAGCTTAAGCTCGA GAAGGATTAACAATGG CGGTAGTCTCGCGG	CA(6)	GGATCGATC AGCGTGCAT	CCGCCATTG TTAATCCTTC T	60.78	59.04	126
3	8087874	AW922049	TCAATGGGTGGACCAA ACACAATTGCAAACCC TCAAGCTGCTGCTGCT GCTCAACTGACCCTCT TACAGCAGCAACGGAT CCTGCCTATGCAACAA CAGCATCAG	TGC(4)	ATGGGTGGA CCAAACACA AT	CTGATGCTGT TGTTGCATAG G	59.95	59.36	105
4	8089381	AW923556	CGCTTTAATGTCGCGA GATTGATAGCAAGGTT CCTGCTGCTGCTGCTG CTGCTCTGGAACAGTG GGAAACACTGCTGTCT TGCTCCGTTGTGTCTA GGGCCGGGGTGTGCA TCGCTGGATTT	CTG(6)	CGCGAGATT GATAGCAAG G	AATCCAGCG ATGCACACC	59.53	60.66	122

5	9847911	BE592838	GCACGAGGCTCGGCC AAAAATCTGCGGCGAG CTCCGCTCCGATCCGA TCCGATCCGACTGCTC CCCACGCCGTTGAGG CCGAGCGGGGGTCCA GGAGAGGAGAGGAGA GGCGCGGC	TCCGA(4)	AGGCTCGGC CAAAAATCT	GCCTCTCCT CTCCTCTCCT G	59.77	60.63	116
6	11680518	BF588194	GCACGAGGGGATTGCT TTCCCCACCCCACCTC CGTCTCCGTCTCTTGC CGACCGAAGCGACTCC ACCTGAGGAGAGGAG AGGAGAGGACAGAGG AGAGGACAGCAAGCTA CCGCCCCCGTCCCCG	GAGGA(4)	CACGAGGGG ATTGCTTTC	CGGTAGCTT GCTGTCCTCT C	59.17	60.16	125
7	11920692	BF655560	GCACGAGGCTCTCTCT CTCTCTCTCTCTCTCTC TCTTCTCTCGAGAATG GTTGTGTAGGCGGCG GGCCCAGGAGGGAGG GAGGGAGGAGAGAGA GAGGGGAGCTAGGGT TTT	GGAG(4)	GCACGAGGC TCTCTCTCTC T	CCCTAGCTC CCCTCTCTCT C	59.03	59.53	112
8	12500452	BG049068	TACTACTAGTAGGACA GGAGTGCTCCATTATT AATTAATTATTCTATTA TAAAGTGCTTGCTTGC TTGCTGCTAATCCATG CACCATTAGGAAGAGA ACCCTCTCTCCGTCGT GCATGCAGCTTGTAC	TGCT(4)	AGGACAGGA GTGCTCCAT TATT	CACGACGGA GAGAGGGTT C	59.13	60.81	128

9	13588582	BG559584	GAACTTTGGTTTCGTTT CCCTGAATGATCAACC ATCATCATCATCATCAT CATGTGTCGAACTCTG CTGTCCCAAGGGATTT CCGTCTCCTCCTTTCC TCTCCTGCTCTTCT	ATC(6)	CTTTGGTTTC GTTTCCCTG A	AGCAGGAGA GGAAAGGAG GA	60.08	60.47	112
10	14570584	BI099002	GCACGAGGCTCTAGC GCCCAGCACGCTCTCG CAGTAGCAGCGAGCG AGCGTGCCGAGGGAG TGAGAGGTGCGGAGC AGTCGCCGGCGTCGA CCTTCGTTGTCGGAGA TGAAGGAGAGACG	AGCG(3)	CAGCACGCT CTCGCAGTA	TCTCTCCTTC ATCTCCGACA A	60.03	59.93	120

Table 3. Summary by nucleotide group and members of the group

Click on the group or the inidividual unit to retrieve data

Group	SSR Count	SSR Units in Group
AC	255	AC CA GT TG
AG	357	AG CT GA TC
AT	260	AT TA
CG	91	CG GC
AAC	39	AAC ACA CAA GTT TGT TTG
AAG	294	AAG AGA CTT GAA TCT TTC
AAT	29	AAT ATA ATT TAA TAT TTA
ACC	281	ACC CAC CCA GGT GTG TGG
ACG	1378	ACG AGC CAG CGA CGT CTG GAC GCA GCT GTC TCG TGC
ACT	278	ACT AGT ATC ATG CAT CTA GAT GTA TAC TAG TCA TGA
AGG	621	AGG CCT CTC GAG GGA TCC
CCG	3312	CCG CGC CGG GCC GCG GGC
AAAC	31	AAAC AACA ACAA CAAA GTTT TGTT TTGT TTTG
AAAG	29	AAAG AAGA AGAA CTTT GAAA TCTT TTCT TTTC
AAAT	20	AAAT AATA ATAA ATTT TAAA TATT TTAT TTTA
AACC	27	AACC ACCA CAAC CCAA GGTT GTTG TGGT TTGG
AACG	46	AACG AAGC ACGA AGCA CAAG CGAA CGTT CTTG GAAC GCAA GCTT GTTC TCGT TGCT TTCG TTGC

AACT	25	AACT AATC ACTA AGTT ATCA ATTG CAAT CTAA GATT GTTA TAAC TAGT TCAA TGAT TTAG TTGA
AAGG	29	AAGG AGGA CCTT CTTC GAAG GGAA TCCT TTCC
AAGT	15	AAGT AATG ACTT AGTA ATGA ATTC CATT CTTA GAAT GTAA TAAG TACT TCAT TGAA TTAC TTCA
AATT	7	AATT ATTA TAAT TTAA
ACAG	7	ACAG AGAC CAGA CTGT GACA GTCT TCTG TGTC
ACAT	77	ACAT ATAC ATGT CATA GTAT TACA TATG TGTA
ACCC	12	ACCC CACC CCAC CCCA GGGT GGTG GTGG TGGG
ACCT	108	ACCT AGGT ATCC ATGG CATC CCAT CCTA CTAC GATG GGAT GGTA GTAG TACC TAGG TCCA TGGA
ACGC	97	ACGC CACG CGCA CGTG GCAC GCGT GTGC TGCG GCCA GCTG GGTC TGGC ACCG AGCC CAGC CCAG CCGA CGGT GACC
ACGG	84	ACGG AGGC CAGG CCGT CCTG CGGA CGTC CTGC GACG GCAG GCCT GGAC GGCA GTCC TCCG TGCC
ACGT	181	ACGT ATGC CATG CGTA GCAT GTAC TACG TGCA AGTC CAGT GTCA TCAG TGAC
ACTC	71	ACTC AGTG CACT CTCA GAGT GTGA TCAC TGAG
AGAT	194	AGAT ATAG ATCT CTAT GATA TAGA TATC TCTA
AGCG	54	AGCG CGAG CGCT CTCG GAGC GCGA GCTC TCGC
AGCT	228	AGCT ATCG CGAT CTAG GATC GCTA TAGC TCGA
AGGG	52	AGGG CCCT CCTC CTCC GAGG GGAG GGGA TCCC
CCCG	74	CCCG CCGC CGCC CGGG GCCC GCGG GGCG GGGC CCGG CGGC GCCG GGCC
AAAAG	69	AAAAG AAACA AAATG AAGAA ATACA ATTAC CATAT GTTTT TAATG TCATT TGTAA TGTTA TGTTT TTCTT TTTCT TTTGT TTTTG TTTTC TCTTT

AAAAT	12	AAAAT ATAAA TAAAA TTATT
AAAGC	155	AAAGC AAGCA AAGCT ACATC ACTGT AGAGA AGCTA AGGAA AGTAC AGTGT ATGGA ATTCG CAGAT CATCT CATTG CCAAT CTAGA CTCAT CTGAT CTTCT GAATC GAGAA GAGTA GATCA GATCT GATTG GCAAA GGTTT GTGTA GTGTT GTTGT GTTTC TACTG TAGCA TAGGA TAGTC TCAGA TCCAT TCCTA TCGAA TCTTC TCTCT TGATC TGCAA TGCTT TGTAC TGTTG TTCAG TTCTC TTGCA TTGCT TTGTG TTTCG TGATT
AACGG	227	AACGG ACAGG ACCGA AGAGC AGAGG AGCAG AGCGA AGCTG AGGGA AGGTG AGTCG AGTGG ATCCC ATCGG ATGCG CACCT CACGC CAGCT CAGTG CCACA CCCTT CCTAC CCTGA CCTTC CGATC CGATG CGGTT CGTCT CGTGT CTAGC CTCCT CTCGT CTCTC CTCTG CTGAG CTGCA CTGCT GAGAG GAGCA GAGCT GAGGA GAGGT GATCG GCACA GCACT GCGTT GCTCA GCTCT GCTTG GGAGA GGCTA GGGAA GGGAT GGTGA GTGCA GTGTG TCCGA TCCGT TCGGA TCGGT TCTCC TGCAC TGCGT TGCTC TGGGT TGGTG TTCCC
ACCGC	91	ACCGC AGGCC CACCC CAGCG CAGGC CCCCT CCCGT CCCTG CCGTC CCTCC CCTGC CGCCT CGCGA CGGAG CGGCA CGGCT CTCCC CTCCG CTCGC CTGCC CTGCG GAGGG GCCGT GCGGA GCTCC GCTGC GGGAC GGGGA GTGCG TCCCC TCCGC TCCGG TCGCC TGCCC TGCGG
CCCCG	15	CCCCG CCCGC CCGCC CGCCG CGCGC CGGCG GCCCC GCGGC GGCGC

1.21 Discussion:

De novo generation of micro-satellite markers through laboratory-based screening of SSR- enriched genomic libraries is highly time consuming and expensive.an alternative is to screen the public database of related model species where abundant sequence data is already available.beyond cost savings, this approach also offers the possibility of identifying rare micro-satellite motifs, which would be uneconomical to identify through laboratory protocols.the availability of massive amounts of nucleotide sequence data has led to the development of innovative ways to examine these data as reflected in their functions. Various types of DNA markers have been used in plant breeding and of these the most extensively used are the micro-satellite markers. The reasons for their extensive use are due to their mode of transmission, which is bi-parental-nuclear with few loci and many alleles per locus. Mode of gene action being co- dominance with exception of null allele at some loci, show large variation within populations and are generally found in non-coding regions, which may contribute to the genome stability. Genome sequence and protein sequence information is publicly available for large –scale analysis from Gene bank at (NCBI) and European Molecular biology Laboratory (EMBL). Today, the search for a gene of interest starts with sequence

Information, including expressed sequence tags (EST). Genome related public database are an invaluable part of the scientific community and most notably the model organism database, have two major consumers: the focused scientific community actively studying that system, and the large scientific community interested in relating this specialized information to and from other systems. The thrust of any high-throughput facility is the creation of large, well-organized, rigorous datasets. The model system database can be mined by other related crop specialist to design markers for marker-assisted selection and -aided introgression methods. such an approach can save valuable resources both in terms of time and funds.

2. Phylogenetic studies in cereals

2.1 Introduction

Phylogeny is really about Evolution and is used to reconstruct evolutionary events at a molecular level through analysis of molecular sequences, namely proteins & nucleic acids. Phylogeny is a diagram (a phylogenetic tree or cladogram) that depicts the evolutionary relationships among organisms. Comparative morphological, anatomical, embryological, molecular, behavioral, physiological, chemical, geographical, and fossil data can all be used, together or separately, to construct the phylogeny. Phylogeny provides the historical perspective from which to interpret the evolution of characters, patterns and processes of diversification, rates of evolution, historical biogeography, and co-evolutionary phenomena, such as the relationships between hosts and parasites or plants and herbivores. Phylogeny is used to classify organisms on the basis of their inferred evolutionary relationships (the phylogenetic approach to classification). Phylogeny is a hypothesis based on the best interpretation of the data at hand and subject to further evaluation (and possibly change) as new data become available.Phylogenetic focuses on the construction of descent relationships of species or groups of species and on how to incorporate these relationships into classification systems. Probably the first steps on establishing relationships and classifications were by Carolus Linnaeus in the 18^{th} century. He is the founder of taxonomy, the biological branch of study dedicated to naming and classifying life forms. Linnaeus first established the binomial naming system, which continues today. In fact, many of the names he developed are still being used as scientific names. He also established a filing hierarchy, but avoided making his classifications based on evolutionary relationships. He was a natural theologian and believed that species were permanent creations that existed according to God's plan.

2.2 Phylogenetic terms

Adaptation	Change in a organism resulting from natural selection: a structure which is the result of such selection.
Anagensis	Evolutionary change along an unbranching lineage: change without speciation.
Ancestor	Any organism, population, or species from which some other organism, population, or species is descended by reproduction.
Basal group	The earliest diverging group within a clade; for instance, to hypothesize that sponges are basal animals is to suggest that the lineage(s) leading to sponges diverged from the lineage that gave rise to all other animals.
Character	Heritable trait possessed by an organism; characters are usually described in terms of their states, for example: "hair present" vs. "hair absent," where "hair" is the character, and "present" and "absent" are its states.
Clade	A monophyletic taxon; a group of organisms which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. From the Greek word "klados", meaning branch or twig.
Cladogenesis	The development of a new clade; the splitting of a single lineage into two distinct lineages; speciation.
Cladogram	A diagram, resulting from a cladistic analysis, which depicts a hypothetical branching sequence of lineages leading to the taxa under consideration. The points of branching within a cladogram are called nodes. All taxa occur at the endpoints of the cladogram.

Convergence	Similarities, which have arisen independently in two or more organisms that are not closely related. Contrast with homology.
Crown group	All the taxa descended from a major cladogenesis event, recognized by possessing the clade's synapomorphy. See: stem group.
Derived	Describes a character state that is present in one or more subclasses, but not all, of a clade under consideration. A derived character state is inferred to be a modified version of the primitive condition of that character, and to have arisen later in the evolution of the clade. For example, "presence of hair" is a primitive character state for all mammals, whereas the "hairlessness" of whales is a derived state for one subclade within the Mammalia.
Diversity	Term used to describe numbers of taxa, or variation in morphology.
Endosymbiosis	When one organism takes up permanent residence within another, such that the two become a single functional organism. Mitochondria and plastids are believed to have resulted from endosymbiosis.
Evolution	Darwin's definition: descent with modification. The term has been variously used and abused since Darwin to include everything from the origin of man to the origin of life.
Evolutionary tree	A diagram, which depicts the hypothetical phylogeny of the taxa under consideration. The points at which lineages split represent ancestor taxa to the descendant taxa appearing at the terminal points of the cladogram.
Extinction	When all the members of a clade or taxon die, the group is said to be extinct.

Gradualism	A model of evolution that assumes slow, steady rates of change. Charles Darwin's original concept of evolution by natural selection assumed gradualism. Contrast with punctuated equilibrium.
Homology	Two structures are considered homologous when they are inherited from a common ancestor, which possessed the structure. This may be difficult to determine when the structure has been modified through descent.
Hypothesis	A concept or idea that can be falsified by various scientific methods.
In-group	In a cladistic analysis, the set of taxa which are hypothesized to be more closely related to each other than any are to the outgroup.
Lineage	Any continuous line of descent; any series of organisms connected by reproduction by parent of offspring.
Monophyletic	Term applied to a group of organisms, which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. A monophyletic group is called a clade.
Outgroup	In a cladistic analysis, any taxon used to help resolve the polarity of characters, and which is hypothesized to be less closely related to each of the taxa under consideration than any are to each other.
Paraphyletic	Term applied to a group of organisms, which includes the most recent common ancestor of all of its members, but not all of the descendants of that most recent common ancestor.
Parsimony	Refers to a rule used to choose among possible cladogram, which states that the cladogram implying the least number of changes in character states is the best.

Phylogenetic	Field of biology that deals with the relationships between organisms. It includes the discovery of these relationships, and the study of the causes behind this pattern.
Phylogeny	The evolutionary relationships among organisms; the patterns of lineage branching produced by the true evolutionary history of the organisms being considered.
Plesiomorphy	A primitive character state for the taxa under consideration.
Polarity of characters	The states of characters used in a cladistic analysis, either original or derived. Original characters are those acquired by an ancestor deeper in the phylogeny than the most recent common ancestor of the taxa under consideration. Derived characters are those acquired by the most recent common ancestor of the taxa under consideration.
Polyphyletic	Term applied to a group of organisms, which does not include the most recent common ancestor of those organisms; the ancestor does not possess the character shared by members of the group.
Primitive	Describes a character state that is present in the common ancestor of a clade. A primitive character state is inferred to be the original condition of that character within the clade under consideration. For example, "presence of hair" is a primitive character state for all mammals, whereas the "hairlessness" of whales is a derived state for one subclade within the Mammalia.
Pseudoextinction	The apparent disappearance of a taxon. In cases of pseudoextinction, this disappearance is not due to the death of all members, but the evolution of novel features in one or more lineages, so that the new clades are not recognized as belonging to the paraphyletic ancestral

	group, whose members have ceased to exist. The Dinosauria, if defined so as to exclude the birds, is an example of a group that has undergone pseudoextinction.
Punctuated equilibrium	A model of evolution in which change occurs in relatively rapid bursts, followed by longer periods of stasis.
Radiation	Event of rapid cladogenesis, believed to occur under conditions where a new feature permits a lineage to move into a new niche or new habitat, and is then called an adaptive radiation.
Rank	In traditional taxonomy, taxa are ranked according to their level of inclusiveness. Thus a genus contains one or more species; a family includes one or more genera, and so on.
Relatedness	Two clades are more closely related when they share a more recent common ancestor between them than they do with any other clade.
Reticulation	Joining of separate lineages on a phylogenetic tree, generally through hybridization or through lateral gene transfer. Fairly common in certain land plant clades; reticulation is thought to be rare among metazoans.
Selection	Process which favors one feature of organisms in a population over another feature found in the population. This occurs through differential reproduction -- those with the favored feature produce more offspring than those with the other feature, such that they become a greater percentage of the population in the next generation.
Sister group	The two clades resulting from the splitting of a single lineage.
Stasis	A period of little or no discernible change in a lineage.

Stem group	All the taxa in a clade preceding a major cladogenesis event. They are often difficult to recognize because they may not possess synapomorpies found in the crown group.
Synapomorphy	A character which is derived, and because it is shared by the taxa under consideration, is used to infer common ancestry.
Systematics	Field of biology that deals with the diversity of life. Systematics is usually divided into the two areas of phylogenetics and taxonomy.
Taxon	Any named group of organisms, not necessarily a clade.
Taxonomy	The science of naming and classifying organisms.
Vicariance	Speciation, which occurs as a result of the separation and subsequent isolation of portions of an original population.

2.3 Phylogenetic classifications

Integrating the results of a phylogenetic analysis into a classification is an extremely contentious issue in systematics. Basically, taxa are recognized on the basis of monophyly.

1. Linnaean system
   a) Named groups are monophyletic
   b) Not all groups are necessarily named
   c) Ranks are arbitrary

2. Phylogenetic or "rankless" systems
   a) Groups given unranked names
   b) Groups defined by ancestry - i.e. phylogenetic tree
   c) Groups described/diagnosed by a character(s) on the branch of the monophyletic

2.4 Methods of phylogenic analysis

There are three methods

1) Cladistic
2) Phenetic
3) Multiple alignment

2.4.1 Cladistic Method

Cladistics is a particular method of hypothesizing relationships among organisms. Like other methods, it has its own set of assumptions, procedures, and limitations. Cladistics is now accepted as the best method available for phylogenetic analysis, for it provides an explicit and testable hypothesis of organismal relationships. The basic idea behind cladistics is that members of a group share a common evolutionary history, and are "closely related", more so to members of the same group than to other organisms. These groups are recognized by sharing unique features, which were not present in distant ancestors. These *shared derived* characteristics are called synapomorphies.The Cladistic approach takes into account both ancestral relationships between organisms and current data stored in their sequences. Here, morphological similarities/differences are also considered as part of the evolutionary history of organisms being compared. The software employed in this approach includes parsimony and maximum likelihood methods.

2.4.2 phenetic Method

Phenetics, also known as numerical taxonomy, involves the use of various measures of overall similarity for the ranking of species. There is no restriction on the number or type of characters (data) that can be used, although all data must be first converted to a numerical value, without any character "weighting." Each organism is then compared with every other for all characters measured, and the number of similarities (or differences) is calculated. The organisms are then clustered in such a way that the most similar are grouped close together and the more different ones are linked more distantly. The taxonomic clusters, called phenograms, that result from such an analysis do not necessarily reflect genetic similarity or evolutionary relatedness. The lack of evolutionary

significance in Phenetics has meant that this system has had little impact on animal classification, and as a consequence, interest in and use of Phenetics has been declining in recent years.

2.4.3 Multiple Alignment Method

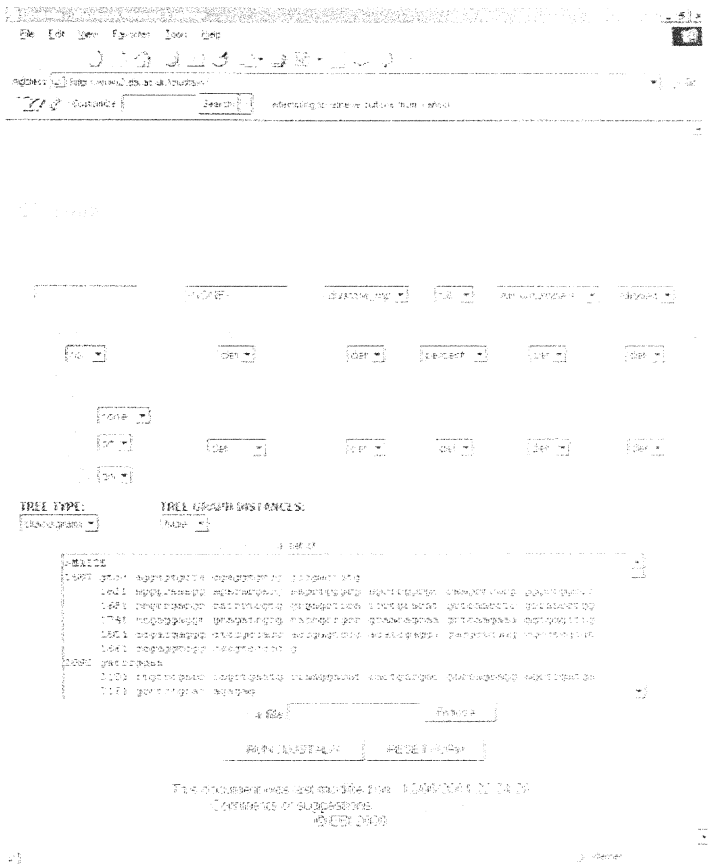
The most practical and widely used method in multiple sequence alignment is the hierarchical extentions of pair wise alignment methods.

- Multiple alignments are built by successive application of pair wise methods:
- Compare all sequences pair wise; (for N sequences there are N.(N-2)/2 pairs or scores)
- Perform cluster analysis on pair wise scores to generate a hierarchy for alignments;
- Build the multiple alignment by first aligning the most similar pair of sequences, then the next most similar pair and so on;

Once an alignment of 2 sequences has been made, then this is fixed. Thus for a set of sequences A, B, C, D having aligned A with C and B with D the alignment of A, B, C, D is obtained by comparing the alignments of A and C with that of B and D using averaged scores at each aligned position.

2.5 Clustalw

ClustalW is a sequence alignment tool, which is used to produce Multiple alignments of a protein sequence./nucleotide sequence. This web tool available at **http://www.ebi.ac.uk/clustalw/**



2.5.1 upload a file

Upload a file from the computer which containing a valid protein nucleotide sequence in any format (GCG, FASTA, PIR, etc.) by using this option. This option only works with Netscape Browsers'

2.5.2 Sequences

The sequences have different names, as the first 30 characters of the name are significant. If clustalw finds two or more sequences with the same name it will fail!

The following are the points to follow not to get the fatal error.

A set of sequences in different formats has been used as input. For example:

>Seq1
MRVLLAALGLLFLGALRAFPQDRPFEDTCHGNPSHYYDKAVRRCCYRCP
MGLFPTQQCPQRPTDCRKQCEPDYYLDEADRCTACVTCSRDD
>Seq2
PVAEERGLMSQPLMETCHSVGAAYLESLPLQDASPAGGPSSPRDLPEPRV
STEHTNNKIEKIYIMKADTVIVGTVKAELPEGRGLAGPAEP
ELEEELEADHTPHYPEQETEPPLGSCSDVMLSVEEEGKEDPLPTAASGK
>Seq3
MSALLTAAGL LFLGMLQAFP TDRPLKTTCA GDLSHYPGEA
ARNCCYQCPS GLSPTQPCPR
GPAHCRKQCA PDYYVNEDGK CTACVTCLPG LVEKAPCSGN
SPRICECQPG MHCCTPAVNS
CARCKLHCSG EEVVKSPGTA KKDTICELPS SGSGPNCSNP
GDRKTLTSIIA TPQAMPTLES

All input sequence must be in the same format. The sequence separators every 10 aa in seq3 are not allowed in the f asta format specification. The program does not allow this kind of input.

- Sequences with spaces (deletions) will have the same affect as above.
- Sequences should not have illegal characters as for example.

  >Badsequence
  *SNERNVCN*WFVYWADQL*WIELKCFA*HICNSG*LQRYNKPKNNQPI
  WGGIKRRCSS*

- Need at least two sequences to generate an alignment.

Illegal characters in the input should be removed. This could be due to cuting & pasting from a word processing program..

Remove any empty space or empty lines from the beginning of the input. put format for the alignment.

Example:

In FASTA format:

>gi|21326110|gb|AF503433.1| Sorghum bicolor BAC 170F8, complete sequence

AAGCTTTTGCTAATCGGAAGCGAGACTTTGCTCCCATCAGAGCAAAGT

ATGATCTTGAGAAGTACAAGTGGGAAAAGTCAAACCGCAAGTGCAAG

ATTGTCATCAAGAAAACCATCACAGAGGGCCTAAGAGGTGCCATCCAT

GTGAGACAGCAAAAGAATACCTTGAGAAAGTGAAGAATCAGTTTACT

GGTTCAACCAAAGCCCATGCTTCCACCCTGATCCAGAAACTCACTAAC

ATGAGGTTCACAGGGGGGAGTGTGAGAGAGCACATTCT

2.5.3 search title

To identify the database search results.

2.5.4 cpu mode

This option allows the choice the version of clustalw based on cpu mode(Central processing unit). There are two options:

Clustalw_mp this is a parallelised version of clustalw that runs on multiprocessor SGI systems.

Clustalw this is the generic (single CPU) version of the program. Use this version if we are having prolems with the MP version of the program.

2.5.5 alignment

choose a full alignment or using a stringent algorithm for generating the tree guide or a fast algorithm.

2.5.6 Output

Output format for multiple alignment results display. The options are ALN, GCG, PHYLIP, PIR and GDE.

2.5.7 Jalview

A new experimental option has been added to the results page, which involved using a Java Applet called JalView. This is a fully featured MSA editor which allows not only to edit the alignment but also, to exchange the alignment formats. For documentation click on the JalView Hyperlink.

2.5.8 Outorder

Decide, which order the sequences, should be printed in the alignment.

2.5.9 color

To get the alignment in color this option should be selected. This option only works when we choose ALN or GCG as the output format.

2.5.10 Fast paiwise alignment options

Ktup	This option allows you to choose which 'word-length' to use when calulating fast paiwise alignments.(note: make sure you have choosen 'fast' in the ALIGNMENT.
Window	Use this option to set the window length when calculating fast pairwise alignments.(Note: make sure you have choosen 'fast' in the ALIGNMENT.
Score	This option allows you to decide which score to take into account when calculating a fast pairwise alignment. (Note: make sure you have choosen 'fast' in the ALIGNMENT.
Topdiag	Select here how many top diagonals should be integrated when calculating a fast pairwise alignment.(Note: make sure you have choosen 'fast' in the ALIGNMENT.
Pairgap	Select here to set the gap penalty when generating fast pairwise alignments.

2.5.11 multiple sequence alignment options

matrix	This option allows you to choose which matrix series to use when generating the mulitple sequence alignment. The program goes through the choosen matrix series, spanning the full range of amino acid distances.

blosum (henikoff) These matrices appear to be the best available for carrying out data base similarity (homology searches). The matrices used are: Blosum80, 62, 40 and 30.

pam (dayhoff) These have been extremely widely used since the late '70s.We use the PAM 120, 160, 250 and 350 matrices.

gonnet These matrices were derived using almost the same procedure as the Dayhoff one (above) but are much more up to date and are based on a far larger data set. They appear to be more sensitive than the Dayhoff series. We use the GONNET 40, 80, 120, 160, 250 and 350 matrices. We also supply an identity matrix, which gives a score of 10 to two identical amino acids and a score of zero otherwise.

2.5.12 Gapopen

Set the penalty for opening a gap.

2.5.13 Endgap

Set the penalty for closing a gap.

2.5.14 Gapext

Set the penalty for extending a gap.

2.5.15 Gapdist

Set the gap separation penalty.

2.5.16 Phylogenetic Tree

To use this option need to input a sequence alignment, if this alignment is in PIR or PHYLIP format. ALN and GCG MSF files are not supported so this has to convert MSF files to PIR format.

This option allows choosing the following output formats for the tree:

Neighbor-Joining

Phylip

Distance

2.5.17 kimura correction of distances

This option allows setting on distances correction (correction for multiple subsitutions). This is because, as sequences diverge, more than one substitution will happen at many

Node: represents a taxonomic unit. This can be either an existing species or an ancestor.

Branch: Defines the relationship between the taxa in terms of descent and ancestry.

Topology: The branching patterns of the tree.

Branch length: Represents the number of changes that have occurred in the branch.

Root: The common ancestor of all taxa.

Distance scale: scale that represents the number of differences between organisms or sequences.

Clade: a group of two or more taxa or DNA sequences that includes both their common ancestor and all their descendents.

Operational Taxonomic Unit (OTU): Taxonomic level of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains.

A phylogenetic tree is composed of nodes--each representing a taxonomic unit (species, populations, and individuals)--and branches, which define the relationship between the taxonomic units in terms of descent and ancestry. Only one branch can connect any two adjacent nodes. The branching pattern of the tree is called the topology and the branch length usually represents the number of changes that have occurred in the branch. This is called a scaled branch. Scaled trees are often calibrated to represent the passage of time. Such trees have a theoretical basis in the particular gene or genes under analysis. Branches can also be unscaled, which means that the branch length is not proportional to the number of changes that has occurred, although the actual number may be indicated numerically somewhere on the branch. Phylogenetic trees may also be either rooted or unrooted. In rooted trees, there is a particular node, called the root--representing a common ancestor--from which a unique path leads to any other node. An unrooted tree only specifies the relationship among species, without identifying a common ancestor, or evolutionary path.

2.7 Methods For The Phylogenetic Studies Of Cereals

Cereals i.e., sorghum bicolor, Hordeum vulgare, Zea mays, Triticum aestivum, orizya sativum are taken and co – related the phylogenetic relationship between these cereals by taking a nuclear enzyme, a mitochondrial enzyme and a chloroplast enzyme.

2.8 Reasons For Taking Enzymes In Phylogenetic Studies

2.8.1 Nuclear Enzyme

1) Nuclear enzymes have bi-parental inheritance.
2) Nuclear enzymes are in abundance in a Genome.
3) Amplification and sequencings are easy.
4) They have mosaics of highly conserved variable region. The conserved regions have been informative for resolving relationship at higher taxonomic levels. Alignment of the variable region is often problematic.
5) They exhibit a wide range of evolutionary rate in phylogenetic utility.
6) Many nuclear genes may contain large Introns that necessitate reverse transcriptase PCR.
7) The nuclear enzyme which we are taken are alcohol-dehydrogenase due to its role in metabolism a key functional pathway

2.8.2 Mitochondrial Enzyme

1) Mitochondrial enzymes are maternally inherited.
2) These enzymes are used to construct a PhylogeneticTree to display the Evolutionary relationships between Individual sequences.
3) The structure of this gene tree contains information which in conjunction with a calibrated mutation rate for the DNA sequence under study, can be used to estimate a time-scale for events in evolutionary prehistory.
4) Sites that have frequently undergone mutations are less conserved among species compared to those where the consensus is more the sequence is highly

conserved. Evolutionary changes are found the non-conserved regions of the sequence.

5) These will provide the phylogenetic evolution of a given mitochondrial gene.
6) The mitochondrial enzyme investigated in this study is Malate-dehydrogenase a key enzyme in the respiratory pathway.

2.8.3 Chloroplast Enzyme

1) Chloroplast gene Restriction-Site Variation has been shown to be well suited for studies of genetic relationships at or below the family level.
2) The chloroplast genome consists of a large and a small region of Single-copy DNA separated by a pair of identical but inverted repeat sequences.
3) Restriction-pattern differences between taxa may be interpreted as site changes caused by single base substitutions or single insertion deletion events.
4) By relating variation in chloroplast DNA restriction-fragment patterns to specific mutations, either base substitution or indels data sets suitable for phylogenetic reconstructions using Parsimony analysis may be produced.
5) Restriction-site variation is used to estimate total sequence divergence between taxa. Such distance measures may used to reconstruct phylogenies.
6) Chloplast Enzymes found in plants only. These enzymes are mostly related to C3 and C4 pathways of photosynthesis.
7) These enzymes are maternally inherited.
8) The chloroplast enzyme, which we taken is Phosphoenol pyruvate the key enzyme of the C4 photosynthetic pathway.

2.9 Multiple Alignment Method

The most practical and widely used method in multiple sequence alignment is the hierarchical extensions of pair wise alignment methods.

2.10 Steps Involved In Multiple Alignment Method

2.10.1Select The Most Conserved Enzymes:

Select the most conserved and functional mitochondrial, chloroplast and nuclear enzymes of cereals. In cereals the most conserved functional enzymes are as follows: -

Mitochondrial Enzymes:

1. NADH dehydrogenase	1.6.5.3
2. ATP synthase	3.6.3.14
3. Succinate dehydrogenase	1.3.99.1
4. Malate dehydrogenase	1.1.1.82
5. Citrate synthase	4.1.3.28

Chloroplast Enzymes:

1. Fructose 1,6 bis phosphatase	3.1.3.11
2. Phosphoenolpyruvate carboxylase	4.1.1.31
3. Glyceraldehyde- 3-phosphate dehydrogenase	1.2.1.9

Nuclear Enzyme:

1.Methyl transferase	2.1.1.37
2.Alcohol dehydrogenase	1.1.1.1
3.Cysteine synthase	4.2.99.8

From the all above-mentioned enzymes we selected the following enzymes, which are more convenient for our practical purpose.i.e

Malate dehydrogenase (mitochondrial enzyme)

Phosphoenol pyruvate carboxylase (chloroplast enzyme)

Methyl transferase (nuclear enzyme)

2.10.2 Search The Sequences From NCBI:

Search the sequences of the Malate dehydrogenase (mitochondrial enzyme), Phosphoenol pyruvate carboxylase (chloroplast enzyme), and Methyl transferase (nuclear enzyme) From NCBI.(National Centre For Biotechnology Information)

Then note the accession number of the above enzymes of interest of cereals.

Chloroplast Enzyme (Phosphoenol Pyruvate Carboxylase)

Cereal	Accession number
Maize	E17154
Rice	AF271995
Sorghum	AF399915

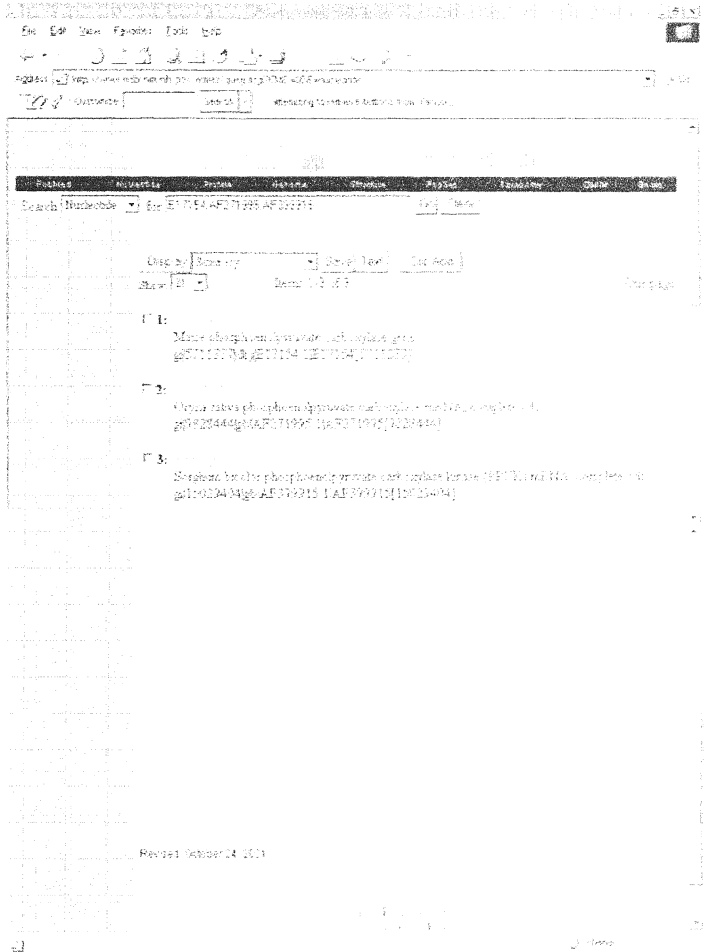
Paste accession numbers of the cereals for a Chloroplast enzyme (Phosphoenol pyruvate carboxylase) in the NCBI. The sequences for the accession number of the cereals for ex:- (maize, oryza sativa and sorghum bicolor) will display.

National Center for Biotechnology Information

What does NCBI do?

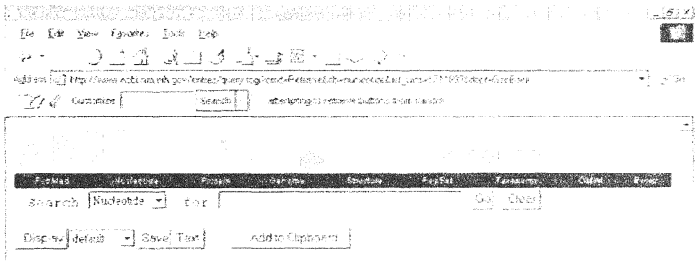
RefSeqs for viral genomes!

NCBI in the News



2.10.3 Accession Number

First click on accession number of maize then oryza sativa followed by sorghum bicolor Sequences of Maize, oryza sativa and Sorghum bicolor phosphenol pyruvate gene will obtain.



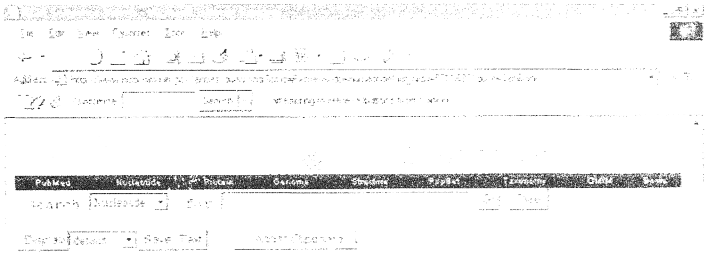
1: E17154, Maize phosphoenol [gi:5711837]
LOCUS E17154 6781 bp DNA linear PAT 28-JUL-1999
DEFINITION Maize phosphoenolpyruvate carboxylase gene.
ACCESSION E17154
VERSION E17154.1 GI:5711837
KEYWORDS JP 19982484l9-A 2.
SOURCE unidentified.
ORGANISM
unclassified.
REFERENCE 1 (bases 1 to 6781)
AUTHORS Matsuoka,M., Tokutomi,M., Toki,S. and Morris,S.K.
TITLE C3 PLANT BODY EXPRESSING PHOTOSYNTHETIC ENZYME OF C4 PLANT
JOURNAL Patent: JP 19982484l9-A 2 22-SEP-1998;
NORIN SUISANSYO NOGYO SEIBUTSU SHIGEN KENKYU SHO
COMMENT OS Zea mays (maize)
PN JP 19982484l9-A 2
PD 22-SEP-1998
PF 11-MAR-1997 JP 1997056742
PI MATSUOKA MAKOTO, TOKUTOMI MITSUE, TOKI SEIICHI, PI MORIS
SUN-BEN KU
PC A01H5 00;A01H1 00;C07H21 04;C12N5 10;C12N9 00;C12N15 09; PC (C12N5 10;
PC C12R1 91);(C12N15 09;C12R1 91);
CC strandedness: Double;
CC topology: Linear;

```
FH   Key             Location/Qualifiers
FH
FT   source          1..6781
FT                   /organism='Zea mays'
FT                   /cultivar='Golden Cross Bantam' FT   intron
     1477..1586
FT   exon            1587..1981
FT   intron          1982..2091
FT   exon            2092..2176
FT   intron          2177..3073
FT   exon            3074..3296
FT   intron          3297..3745
FT   exon            3746..3849
FT   intron          3850..3974
FT   exon            3975..4065
FT   intron          4066..4202
FT   exon            4203..4357
FT   intron          4358..4453
FT   exon            4454..5452
FT   intron          5453..5549
FT   exon            5550..5936
FT   intron          5937..6024
FT   CDS
FT                   join(1297..1476,1587..1981,2092..2176, FT
          3074..3296,
FT                   3746..3849,3975..4065,4203..4357, FT
        4454..5452,5550..5936,
FT                   6025..6318)
FT                   /product='phosphoenolpyruvate coaboxylase'.
FEATURES             Location/Qualifiers
  source        1..6781
              /organism="unidentified"
              /db_xref="taxon:32644"
BASE COUNT     1687 a   1854 c   1548 g   1692 t
ORIGIN
      1 tctagagatg taatggtgtt aggacacgtg gttagctact aatataaatg taaggtcaaa
     61 attcgatggt ttatttttcta ttttcactta cctagcatta tctcatttct aattgtgtga
    121 taacaaatgc attagaccat aattctgtaa atacgtacat ttaagcacac agtctatatt
    181 ttaaaattct tctttttgtg tggatatccc aaccccaaatc cacctctctc ctcaatccgt
    241 gtatcttcac cgctgccaag tgccaacaac acatcgcatc gtgcaaatct ttgttggttt
    301 gtgcacggtc ggcgccaatg gaggagacac ctgtacggtg cccttggtag aacaacatcc
    361 ttatccctat atgtatggtc ccttcgtaga tggcaccctt atccctacaa tagccatgta
    421 ttgcatacca agaatttaaat atacttttttc ttgaaccaca ataatttatt atagcggcac
    481 ttcttgttct ggttgaacac ttatttggaa caataaaatc ccgagttcct aaccacaggt
    541 tcacttttti tccttatcct cctaggaaac taaattttaa attcataaat ttaattgaaa
    601 tgttaatgaa aacaaaaaaa ttatctacaa agacgactct tagccacagc cgcctcactg
```

661 cacccteaac cacatcctgc aaacagacac cctcgccaca tcccteeaga ttcttcccte
721 cgatgcagcc tacttgctaa cagacgccct ctccacatcc tgcaaagcat tcctccaaat
781 tcttgcgatc cccegaatcc agcattaact gctaagggac gccctctcca catcctgcta
841 cccaattagc caacggaata acacaagaag gcaggtgagc agtgacaaag cacgtcaaca
901 gcaccgagcc aagccaaaag gagcaaggag gagcaagccc aagccgcagc cgcagctctc
961 caggtcccct tgcgattgcc gccagcagta gcagacaccc ctctccacat cccctccggc
1021 cgctaacagc agcaagccaa gccaaaaaga agcctcagcc acagccggtt ccgttgcggt
1081 taccgccgat cacatgccca aggccgcgcc tttccaaacg ccgagggccg cccgttcccg
1141 tgcacagcca cacacacacc cgcccgccaa cgactcccca tcoctatttg aacccacccg
1201 cgcactgcat tgatcaccaa tcgcatcgca gcagcacgag cagcacgccg tgccgctcca
1261 accgtctcgc ttcoctgctt agcttcccgc cgcgccatgg cgtcgaccaa ggcaccccggc
1321 cctggcgaga agcaccactc catcgacgcg cagctccgtc agctggtccc aggcaaggtc
1381 tccgaggacg acaagctcat cgagtacgat gcgctgctcg tcgaccgctt cctcaacatc
1441 ctccaggacc tccacgggcc cagccttcgc gaatttgtaa ctaaccaccg ccgccgccca
1501 tttcttcttc gaccggttgc cgcctgcgcg cggcactgct cgtgtcgtgt gcttagtctt
1561 agtgcttact actgtaatgc atgcaggtcc aggagtgcta cgaggtgtcg gccgactatg
1621 agggcaaagg agacacgacg aagctgggcg agctcggcgc caagctcacg gggctggccc
1681 ccgccgacgc catcctcgtg gcgagctcca tcctgcacat gctcaacctc gccaacctgg
1741 ccgaggaggt gcagatcgcg caccgccgcc gcaacagcaa gctcaagaaa ggtgggttcg
1801 ccgacgaggg ctccgccacc accgagtccg acatcgagga gacgctcaag cgcctcgtgt
1861 ccgaggtcgg caagtccccc gaggaggtgt tcgaggcgct caagaaccag accgtcgacc
1921 tcgtcttcac cgcgcatccc acgcagtccg cccgccgctc gctcctgcag aaaaacgcca
1981 ggtatatatt tctcaatggc ttgatcgata tgctactcac gttatatacc cttaagtctt
2041 aaccattatt attatttttg ataaaataaaa atgtcggtct tgtcgctgca ggatccggaa
2101 ttgtctgacc cagctgaatg ccaaggacat cactgacgac gacaagcagg agctcgatga
2161 gcctctgcac agagaggtac gtacatatta catttcacac cagggaatgc aagaacttta
2221 tcaagagaca ttcattcttt gatagagata gaatagaaca catgcacagt acacgtggac
2281 tcatgagctt gcaagacatc gagcacgaca cgtgtaagtt agtgcgccag agaaatcttc
2341 aatttatatg tcaagtcagg tcaggttctc ccattaaaac acatataaat aaatattcat
2401 tattatcaag ctaaggtaat aaacaaccaa acttttccac tatttaaact gtctttgcaa
2461 actccaaagt agaaactaac ctaatcagga aagaactaga ctgcacattt atgttttaac
2521 aatgcaatga gagaactgct acatgtataa cagaattatt tatatgaggc cgacttgact
2581 taagattcaa tgttgaagac cacttgatga aaactacact gaattattta tatgctattc
2641 tccagctgtg ctcaaagcat tttcctttta cttaaaaaag atcatttgt acaaagatct
2701 cttactcata tagagccatt tgagtagaac ttcggtacca cagatgcatt aatggtttag
2761 ttgtaatcaa gttgttgtac tcatcattat attttcctaa caagtagggc atccagtttc
2821 tccttgatga ggaatcaaac ctagatagcc ttaactccac accctcaatt agctaggcta
2881 tgctcaagtt cctagtgtta caaatttcag acgagtcata atgtcatcac tgagcactcg
2941 gtaaagagcg tctctctcat ggtgcatata tatgatgcag accacctgag aagtttactg
3001 cttcaagcca ccaaagtggt atttttgttg tttgggttgt ttagttctaa ttcctttttt
3061 tgggtgttca cagatccaag cagccttcag aaccgatgaa atcaggaggg cacaacccac
3121 cccccaggac gaaatgcgct atgggatgag ctacatccat gagactgtat ggaagggcgt
3181 gcctaagttc ttgcgccgtg tggatacagc cctgaagaat atcggcatca atgagcgcct
3241 tccctacaat gtttctctca ttcggttctc ttcttggatg ggtggtgacc gcgatggtac
3301 atttctgcct acccttttca ataaagtggc aggagctctc tgtctttcag cttgagagaa
3361 accttcctgc tttactctga ctgcaataga tgttcagaaa aactagtcta tcatttcgag

3421 ctctcaggag ctagaatttt aaaattgaaa ttatttagta cacctcacta ataaaaattt
3481 atcatccata catgctagca caacatataa gcataattta atcaaatctt tatattgcaa
3541 cctggaacct aacttgttga atttttttata tcacagaatt atacgtgtag tattatttta
3601 tatatcaaag agtgcttata ttatatcagt acttgtcctg tcaatattca aggctaacgt
3661 ttttctttte tcgccagaaa attatatata cagaattata tgtttttttct aagcctgtat
3721 atctttgcaa tctatcgcta tataggaaat ccaagagtta cccggaggct gacaagagat
3781 gtatgcttgc tggccagaat gatggctgca aacttgtaca tcgatcagat tgaagagctg
3841 atgtttgagg tactgtacat ccatactgca gatttgtttg attgaatgct ctatgatttt
3901 ttgcttgccc tgttttttgc tgtctccggt ccataccaga actctcatgc atgcatcgtc
3961 tgatatatct gtagctctct atgtggcgct gcaacgatga gcttcgtgtt cgtgccgaag
4021 agctccacag ttcgtctggt tccaaagtta ccaagtatta cataggtaac cacaaacaga
4081 agcatttatg tttgcttaat ttttgcctgc cgtacaggct tttgcaaaag tctccactag
4141 tgtttttcaaa ttaatttgag ggctcttttg gcatcttttc tgaagtgtat ttgctggcgc
4201 agaattctgg aagcaaattc ctccaaacga gccctaccgg gtgatactag gccatgtaag
4261 ggacaagctg tacaacacac gcgagcgtgc tcgccatctg ctggcttctg gagtttctga
4321 aatttcagcg gaatcgtcat ttaccagtat cgaagaggta aatatcgtca tgtatatatt
4381 atatatattc atagtatgac atcagcactg caactaacaa aaaaaaaatc actactgtcg
4441 tgcatgcatg cagttccttg agccacttga gctgtgctac aaatcactgt gtgactgcgg
4501 cgacaaggcc atcgcggacg ggagcctcct ggacctcctg cgccaggttt tcacgttcgg
4561 gctctccctg gtgaagctgg acatccggca ggagtcggag cggcacaccg acgtgatcga
4621 cgccatcacc acgcacctcg gcatcgggtc gtaccgcgag tggtccgagg acaagcggca
4681 ggagtggctg ctgtcggagc tgcgaggcaa gcgcccgctg ctgccccccgg accttcccca
4741 gaccgaggag atcgccgacg tcatcggcgc gttccacgtc ctcgcggagc tcccgcccga
4801 cagcttcggc ccctacatca tctccatggc gacggccccc tcggacgtgc tcgccgtgga
4861 gctcctgcag cgcgagtgcg gcgtgcggcc agccgtgccc gtggtgccgc tgttcgaaag
4921 gctggccagc ctgcagtcgg cgcccgcgtc cgtggagcgc ctcttctcgg tggactggta
4981 catggaccgg atcaagggca agcagcaggt catggtcggc tactccgact ccggcaagga
5041 cgccggccgc ctgtccgcgg cgtggcagct gtacagggcg caggaggaga tggcgcaggt
5101 ggccaagcgc tacggcgtca agctcacctt gttccacggc cgcggaggca ccgtgggcag
5161 gggtggcggg cccacgcacc ttgccatcct gtcccagccg ccggacacca tcaacgggtc
5221 catccgtgtg acggtgcagg gcgaggtcat cgagttctgc ttcggggagg agcacctgtg
5281 cttccagact ctgcagcgct tcacggccgc cacgctggag cacggcatgc acccgccggt
5341 ctctcccaag cccgagtggc gcaagctcat ggacgagatg gcggtcgtgg ccacggagga
5401 gtaccgctcc gtcgtcgtca aggaggcccg cttcgtcgag tacttcagat cggtatgctg
5461 ccattgccca ttgctttgtg acgatcgaat tcatccatgt cgatcgttct tttcattcat
5521 tcgagcgttt gtgcgtcact cactatcagg ctacaccgga gaccgagtac gggaggatga
5581 acatcggcag ccggccagcc aagaggaggc ccggcggcgg catcacgacc ctgcgcgcca
5641 tcccctggat cttctcgtgg acccagacca ggttccacct ccccgtgtgg ctgggagtcg
5701 gcgccgcatt caagttcgcc atcgacaagg acgtcaggaa cttccaggtc ctcaaagaga
5761 tgtacaacga gtggccattc ttcagggtca ccctggacct gctggagatg gttttcgcca
5821 agggagaccc cggcattgcc ggcttgtatg acgagctgct tgtggcagaa gaactcaagc
5881 cctttgggaa gcagctcagg gacaaatacg tggagacaca gcagcttctc ctccaggtac
5941 caaaaccagc actgcactgt acgatatgaa taaaagtctg ttgtctggct cctgatcgat
6001 gactgactac tccatttttgt gcagatcgct gggcacaagg atattcttga aggcgatcca
6061 ttcctgaagc agggactggt gctgcgcaac ccctacatca ccaccctgaa cgtgttccag
6121 gcctacacgc tgaagcggat aagggacccc aacttcaagg tgacgccccа gccgccgctg

6181 tccaaggagt tcgccgacga gaacaagccc gccggactgg tcaagctgaa cccggcgagc
6241 gagtacccgc ccggcctgga agacacgctc atcctcacca tgaagggcat cgccgccggc
6301 atgcagaaca ctggctaggc ggcttctctt cactcacctg cagagtgcac cgcaataatc
6361 agcttccgga tggtggcgtt ttgtcagttt tggatggaaa tgccgaactg gcagcgtctg
6421 tttccctat gcatatgtaa tttcctgcct cttatattc actcttgttg tcaagtccaa
6481 gtggaaaatc ttggcatatt atacatattg taataataaa catcgtacaa tctgcatgct
6541 gtttgtaat aattaattaa tatcccagcc cattggatgg actgtttac catggtgtta
6601 cttcagccac cctctcttag ttgtgctaaa catttctga ttggtatttt tttttaga
6661 gtaacctagt gcatttactt aagagagatg atatctagtg gcactagtga ttagtttgca
6721 agattgagaa cttgttactc gctcctagag gttaacacta gcaagtgatt ggagcttagg
6781 g



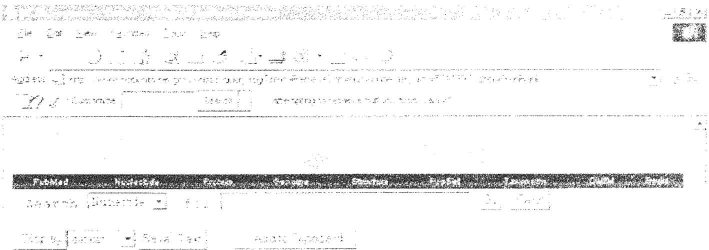
1: AF271995 Oryza sativa phos...[gi:9828444]
LOCUS AF271995 3307 bp mRNA linear PLN 16-AUG-2000
DEFINITION Oryza sativa phosphoenolpyruvate carboxylase mRNA, complete cds
ACCESSION AF271995
VERSION AF271995.1 GI:9828444
KEYWORDS .
SOURCE Oryza sativa (japonica cultivar-group)
ORGANISM
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae;
Ehrhartoideae; Oryzeae; Oryza.
REFERENCE 1 (bases 1 to 3307)
AUTHORS Yamamoto,N., Kurita,A., Masumura,T., Sugimoto,T., Morita,S.,
Shiraishi,N., Oji,Y. and Tanaka,K.
TITLE Root type of phosphoenolpyruvate carboxylase in developing rice
seeds
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 3307)
AUTHORS Yamamoto,N., Kurita,A., Masumura,T., Sugimoto,T., Morita,S.,
Tanaka,K., Shiraishi,N. and Oji,Y.

```
  TITLE     Direct Submission
  JOURNAL   Submitted (24-MAY-2000) Lab. Plant Nutrition, Faculty of
            Agriculture, Kobe University, Nada, Kobe 657-8501, Japan
FEATURES             Location/Qualifiers
     source          1..3307
                     /organism="Oryza sativa (japonica cultivar-group)"
                     /cultivar="Nipponbare"
                     /db_xref="taxon:39947"
                     /dev_stage="developing seeds"
     CDS             188..3067
                     /codon_start=1
                     /product="phosphoenolpyruvate carboxylase"
                     /protein_id="AAG00180.1"
                     /db_xref="GI:9828445"

/translation="MERHQSIDAQLRLLAPGKVSEDDKLVEYDALLVDRFLDILQDLH

GPHLREFVQECYELSAEYENDRDEARLDELGRKLTSLPPGDSIVVSSSFSHMLNLANL

AEEVQIAHRRRIKLKRGDFADEASAPTESDIEETLKRLVTQLGKSREEVFDALKNQTV

DLVFTAHPTQSVRRSLLQKHGRIRNCLRQLYAKDITADDKQELDEALQREIQAAFRTD

EIRRTPPTPQDEMRAGMSYFHETIWKGVPKFLRRIDTALKNIGINERLPYNAPLIQFS

SWMGGDRDGNPRVTPEVTRDVCLLARMMAANLYFSQIEDLMFELSMWRCSDELRIRAD

DLHCSSRKAAKHYIEFWKQIPPNEPYRVILGGVRDKLYYTRERTRHLLTTGVSEIPEE

ATFTNVEEFLEPLELCYRSLCACGDKPIADGSLLDFLRQVSTFGLALVKLDIRQESDR

HTDVLDAITTYLGIGSYAEWSEEKRQDWLLSELRGKRPLFGPDLPQTEEIADVLGTFH

VLAELPADCFGAYIISMATAPSDVLAVELLQRECHIKQPLRVVPLFEKLADLEAAPAA

VARLFSIDWYMNRINGKQEVMIGYSDSGKDAGRLSAAWQMYKAQEELVKVAKHYGVK
L

TMFHGRGGTVGRGGGPSHLAILSQPPDTIHGSLRVTVQGEVIEHSFGEEHLCFRTLQR

FTAATLEHGMHPPISPKPEWRALMDEMAVVATKEYRSIVFKEPRFVEYFRSATPETEY

GRMNIGSRPSKRKPSGGIESLRAIPWIFAWTQTRFHLPVWLGFGGAFKHIMQKDIRNI

HTLKEMYNEWPFFRVTLDLLEMVFAKGDPGIAALYDKLLVAGDLQSFGEQLRNNFEET

KQLLLQVAGHKDILEGDPYLKQRLRLRESYITTLNVCQAYTLKRIRDPSFEVMSQPAL
          SKEFVDSNQPAELVQLNAASEYAPGLEDTLILTMKGIAAGMQNTG"
BASE COUNT    816 a   814 c   880 g   797 t
```

ORIGIN

1 cccacactct ctacccccaa acttgccctc caatctatct ccgccgccgc cgccgcctcc
61 gccgcgagag ggcgtcgccg agctgcgcgg gaggctcgcc tcccgcagca tctggccgga
121 ttcgtcgtag tcgaggcagc tgtaagcgag agagagggag atggccgcgg ccgccgggaa
181 ggcggcgatg gagcggcacc agtcgatcga cgccgcagctg cggctgctcg cgccggggaa
241 ggtatccgag gacgacaagc tcgtcgagta cgacgccctc ctcgtcgacc gcttcctcga
301 catcctccag gacctgcacg gccctcacct ccgcgaattc gtgcaggagt gctacgagct
361 gtcggccgag tacgagaacg acagggacga ggcgccggct gacgagctgg ggaggaagct
421 caccagcctg ccgccggggg actccatcgt cgtctccagc tccttctcgc acatgctcaa
481 cctcgccaac ctcgccgagg aggtgcagat cgccgcaccgc cgccggatca agctcaagcg
541 cggggacttc gccgacgagg cctccgcgcc caccgagtcc gacatcgagg agacgctcaa
601 gcgcctcgtc acccagctcg gcaagtcgag ggaggagctg ttcgatgcgc tcaagaacca
661 gaccgttgac ctcgtgttca ccgcgcatcc aacgcagtcc gtcaggaggt cctgctctca
721 gaagcacggg aggatccgga attgcctgag gcagttatat gcgaaagaca tcactgctga
781 tgacaagcag gagcttgatg aggccctgca gagggagatt caagcagctt tcagaactga
841 tgaaatccgc aggactcctc ccactcctca ggatgaaatg cgcgccggga tgagttattt
901 tcatgaaact atatggaagg gtgtaccaaa gttcttgcgc cgcattgata ctgctctgaa
961 aaatattggg attaatgagc gtcttcctta caatgctcct ctcatccagt tctcatcctg
1021 gatgggtggt gaccgtgatg gaaacccaag agttacacca gaggttacaa gagatgtatg
1081 cttgttggca agaatgatgg ctgctaacct gtacttctct cagatagaag atctgatgtt
1141 tgagctctct atgtggcgct gcagtgatga acttcgaatt cgtgcagatg atttgcattg
1201 ctcctccaga aaagctgcaa agcactatat agaattctgg aagcaaattc ctccaaatga
1261 gccttatcgt gtcatacttg gtggtgtcag ggataaattg tactatacgc gtgaacgtac
1321 tcgccatcta ttgacaactg gagtttctga aattccagag gaggcaacgt ttactaatgt
1381 tgaagagttt ctggagccgc ttgagctgtg ctacagatca ttatgtgctt gtggtgacaa
1441 acctatagct gatggaagcc ttcttgattt cttgcgtcaa gtatcaactt ttggggctggc
1501 tcttgtaaaa cttgacataa ggcaggagtc tgatcgacac actgatgtcc ttgatgccat
1561 aactacatat cttgggattg gatcttacgc tgaatggtct gaggagaaac gccaggattg
1621 gcttttgtcc gaactaaggg gcaaacgtcc tttgttttggt cctgatcttc ctcagactga
1681 agaaattgct gatgttttag gaacatttca cgtccttgct gagctccccgg cagattgttt
1741 tggtgcttac atcatctcaa tggcaactgc accatctgat gtgcttgctg tggagcttct
1801 acagcgggag tgccatataa aacagcctct gcgagttgtt ccctatttg agaaacttgc
1861 agatcttgaa gcagctccag cagcggtggc acgactattt tcaatagact ggtacatgaa
1921 taggattaat ggcaagcagg aggtcatgat tggatactca gactctggca aggatgctgg
1981 tcgtctttct gcagcatggc aaatgtacaa agcacaagag gagcttgtca aggtggcaaa
2041 gcattatggt gtaaagttga caatgttcca tggaagaggt ggaacagttg gcagaggagg
2101 tggtcccagt catcttgcca tattatctca accaccagac acgatacatg gatcacttcg
2161 tgtaacagta cagggcgagg ttattgaaca ctcgtttgga gaggaacact tgtgctttag
2221 aactctgcag cgctttactg cagctactct tgagcatgga atgcatcctc caatttcccc
2281 caaaccagaa tggcgtgctc taatggatga gatggctgtt gtggcaacaa aggaatatcg
2341 atcaatcgtc ttcaaggagc cacggtttgt tgaatacttc cgatcggcaa cacctgagac
2401 agaatatggc aggatgaaca ttggtagccg accatcaaag agaaagccta gtggtggcat
2461 tgaatcgctc cgtgcaattc cttggatttt tgcttggaca caaactaggt ttcatcttcc
2521 tgtatggctt ggatttggtg gagcgttcaa acatataatg cagaaggata tcaggaacat
2581 ccatacactg aaagaaatgt acaatgagtg gccattcttc agggtcacct tggacttgct
2641 tgagatggtt ttcgctaagg gagatccagg aattgctgct ttatatgaca aattgcttgt

2701 ggcaggtgat ctgcaatcct tggagagca gctgagaaac aacttgaag agacaaaaca
2761 gctactcctt caggttgccg gtcacaagga tattctggaa gggatcctt acctgaagca
2821 gcgtctgcgg ctgcgtgagt cgtacatcac taccttgaac gttgccaag cctacaccct
2881 gaagcggata agggacccta gcttcgaggt gatgtcgcag ccggccttgt cgaaggagtt
2941 cgtcgacagc aaccagccctg ccgagctggt ccaactgaac gctgcgagcg agtacgcacc
3001 tggcctggag gacaccctca tcctgacgat gaagggcatt gctgccggca tgcagaacac
3061 aggctagatt gtccacaaga atgcttccag caggtgcaat caatcaatca tcaataactc
3121 cttccatgat gagatggtga atatgttaa ctagtacgtt cgccgatgct aggaattct
3181 ggagacttt ttgctcgc acgtttctt tggttgtgt actgatgta aatgtcaagg
3241 gggacttat tggcttg cagcccgatc caaaataaaa tatggaccaa tcattctacg
3301 aaaaaa



1: AF399915. Sorghum bicolor p...[gi:15029404]
LOCUS AF399915 1389 bp mRNA linear PLN 30-JUL-2001
DEFINITION Sorghum bicolor phosphoenolpyruvate carboxylase kinase (PPCK) mRNA,
complete cds.
ACCESSION AF399915
VERSION AF399915.1 GI:15029404
KEYWORDS .
SOURCE sorghum
ORGANISM Sorghum bicolor
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACC
clade; Panicoideae; Andropogoneae; Sorghum
REFERENCE 1 (bases 1 to 1389)
AUTHORS Hartwell,J., Gill,A., Nimmo,G.A., Wilkins,M.B., Jenkins,G.I. and
Nimmo,H.G.
TITLE Phosphoenolpyruvate carboxylase kinase is a novel protein kinase
regulated at the level of expression
JOURNAL Plant J. 20 (3), 333-342 (1999)
MEDLINE 20046076
PUBMED 10571893

REFERENCE 2 (bases 1 to 1389)
AUTHORS Fontaine,V., Hartwell,J., Jenkins,G.I. and Nimmo,H.G.
TITLE Direct Submission
JOURNAL Submitted (12-JUL-2001) IBLS, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK
FEATURES Location/Qualifiers
source 1..1389
/organism="Sorghum bicolor"
/db_xref="taxon:4558"
gene 1..1389
/gene="PPCK"
CDS 21..944
/gene="PPCK"
/note="protein kinase"
/codon_start=1
/product="phosphoenolpyruvate carboxylase kinase"
/protein_id="AAK81871.1"
/db_xref="GI:15029405"

/translation="MSAELKRDYEIGAEIGRGRFGVVHRCTSRATGEAFAVKSVDRSQ

LADDLDRELAELEPKLAQLAGAGNPGVVQTHAVYEDETWTHTVMDLCAGPDLLEWVR
L

RRGAPVPEPLAAAIVAQVAQALALCHRRGVAHRDVKPDNILIDTAPAGSDDGEDEEED

SGEAEEAETAPRARLADFGSAAWVGAGGLGRAEGLVGTPHYVAPEVVGGGEYGAKAD
V

WSAGVVMYALLSGGALPFGGETAAEVLAAVLRGSVRFPPRLFSGVSPAAKDLLRRMIC
RDEWRRFTAEQVLAHPWIVSGDGARAIERPT"
BASE COUNT 263 a 393 c 507 g 226 t
ORIGIN
1 gcacgaggaa ttactgaagt atgagtgcgg agctgaagag ggactacgag ataggcgcgg
61 agatcggtcg cggccgcttc ggggtggtcc accgctgcac gtcccgcgcc accggcgagg
121 cgttcgccgt caagtccgtg gaccggtcgc agctggccga cgacctggac cgcgagctcg
181 cggagctgga gcccaagctg gcgcagctcg ccggcgcggg caacccccgc gtggtgcaga
241 cgcacgcggt gtacgaggac gagacgtgga cccacacggt gatggacctg tgcgcggccc
301 cggacctgct cgagtgggtg cgcctccgcc gcggcgcgcc tgtgccggag ccctggcgg
361 ccgccatcgt cgcgcaggtc gcccaggcgc tcgcgctctg ccaccgccgc ggcgtcgccc
421 accgcgacgt aaagccccgac aacatcctca tcgacaccgc cccgcgaggg agcgacgacg
481 gcgaggacga ggaagaggac agcggcgagg ccgaggaggc cgagacgggcg ccgcgcgcgc
541 ggctggcgga cttcgggtcg gcggcgtggg tgggcgccgg tgggctgggc cgcgcggaag
601 gcctggtggg gacgccccac tacgtggcgc ccgaggtggt gggcggcggc gagtacggcg
661 cgaaggcgga cgtgtggagc gccggcgtgg tgatgtacgc gctgctctcc ggcggcgcgc
721 tccccgttcgg cggcgagacc gcggcggagg tgctaggcgc cgtgctgcgg ggcagcgtcc
781 ggttcccgcc caggctgttc agcggggtgt ccccgcggc caaggacctg ttgaggcgca

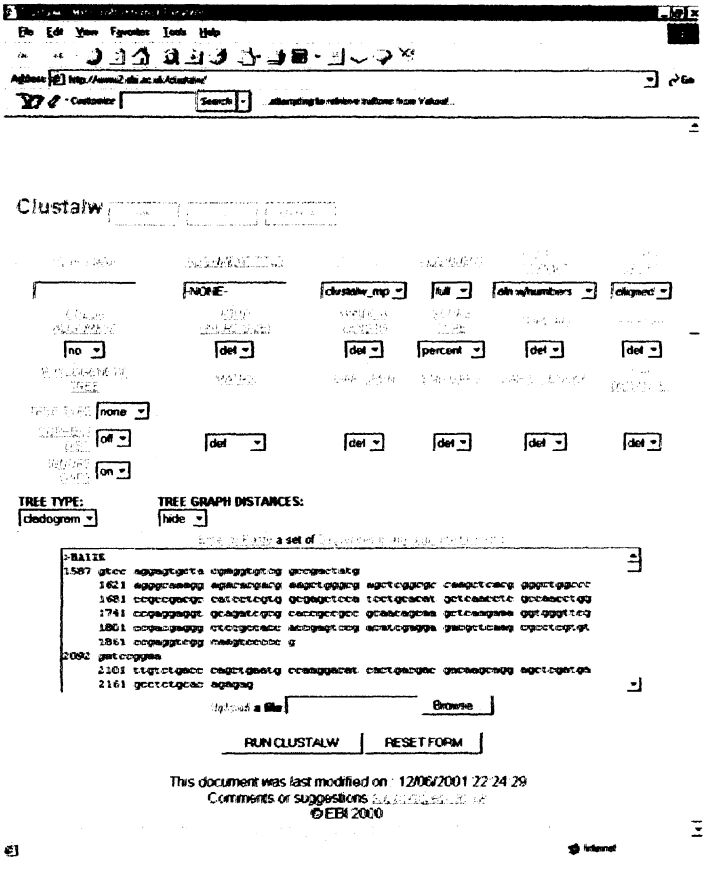
841 tgatctgccg cgacgagtgg cggaggttca ccgccgagca agtcctcgct cacccgtgga
901 tcgtgagcgg cggaggagca cgggcaatcg agcggccaac ctgaggatga ggagggcata
961 ccataccagc agaggaggag accatatggc gggcggacgt gtattgccgt agagacacag
1021 acacggcaag gaatcccgac gtcgttatgt atacggtttt gtatgtatgt atgtatgtac
1081 gtacggacca gagagtagta gagtacggta gctgctgccg aatgatgagt agcatcatcg
1141 attccgcgct gtgtaggcgc agcagagcag atcgatcggc aaagaagctt gtataggctt
1201 ctggtttttcc ttacttccgg gtttttggga tgctctttac ctagcagtgc agagcacect
1261 ttttttggat gtgcctggta agtggtagct ggatgtatat acactgaccg tcttgtacaa
1321 atggagtgga gttggataaa taaataaaat cgagcctttg tacttccaaa aaaaaaaaaa
1381 aaaaaaaaa

2.10.4 Exon Regions

Note the Exon regions of maize, oryza sativa and sorghum bicolor pyruvate carboxylase from the NCBI, which is given above.

2.10.5 Multiple Alignment

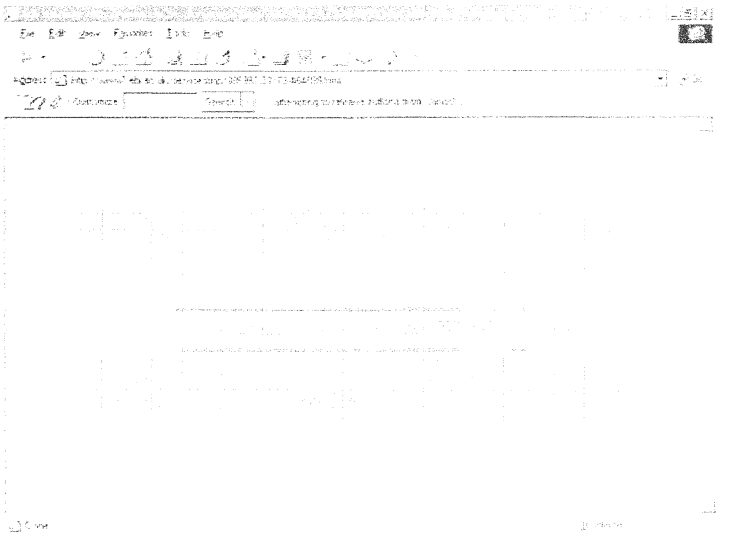
Make a Multiple alignment by putting the sequences (only Exons) of Maize, oryza sativa and Sorghum bicolor phosphenol pyruvate carboxylase in Clustal W.



ClustalW Submission Form

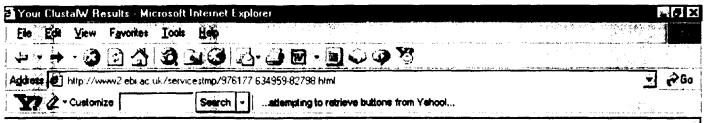
2.10.6 Run Clustalw

Click RUNCLUSTALW button, which is present on CLUSTALW. Job output will be displayed which is given below.



2.10.7 Jalview

Jalview will obtain after the job output by clicking on RESULT SAVE button



Your ClustalW Results:

View JalView: JalView SUBMIT ANOTHER JOB

Pairwise Scores:

CLUSTAL W (1.81) Multiple Sequence Alignments

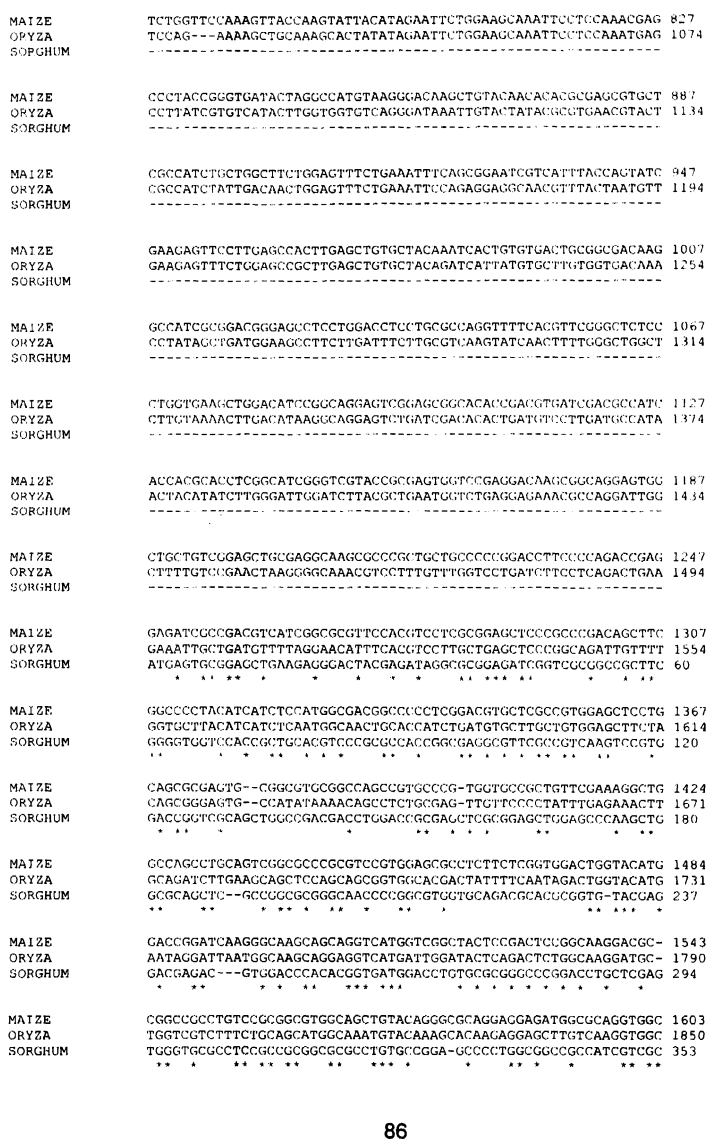
CLUSTAL W (1.81) Multiple Sequence Alignments

```
Sequence format is Pearson
Sequence 1: MAIZE          2339 bp
Sequence 2: ORYZA          2878 bp
Sequence 3: SORGHUM        9?4 bp
Start of Pairwise alignments
Aligning...
Sequences (2:3) Aligned. Score:  11
Sequences (1:2) Aligned. Score:  74
Sequences (1:3) Aligned. Score:  ?
Guide tree        file created:   [/net/nfs0/vol1/production/w3nobody/tmp/976177.634959-
82798.dnd]
Start of Multiple Alignment
There are 2 groups
Aligning...
Group 1: Sequences:    2      Score:35531
Group 2:                      Delayed
Sequence:3     Score:869
Alignment Score 16642
CLUSTAL-Alignment file created  [/net/nfs0/vol1/production/w3nobody/tmp/976177.634959-
82798.aln]
```

Your Multiple Sequence Alignment:

```
CLUSTAL W (1.81) multiple sequence alignment


MAIZE          ------------------------------------------------------------
ORYZA          ATGGAGCGGCACCAGTCGATCGACGCGCAGCTGCGGCTGCTCGGGCCGGGGAAGGTATCC  60
SORGHUM        ------------------------------------------------------------


MAIZE          ------------------------------------------------------------
ORYZA          GAGGACGACAAGCTCGTCGAGTACGACGCCCTCCTCGTCGACCGCTTCGTCGACATCCTC  120
SORGHUM        ------------------------------------------------------------


MAIZE          ----------------------------GTCCAGGAGTGCTACGAGGTGTCGGCC  27
ORYZA          CAGGACCTGCACGGCCCTCACCTCCGCGAATTCGTGCAGGAGTGCTACGAGCTGTCGGCG  180
```


SORGHUM
MAIZE GACTATGAGGGCAAAGGAGACACGACGAAGCTGGGCGAGCTCGGCGCCAAGCTCACGGGG 87
ORYZA GAGTACGAGAACGACAGGGACGAGGCGCGGCTCGACGAGCTCGGGAGGAAGCTCACCAGC 240
SORGHUM
MAIZE CTGGCCCCCGCCGACGCCATCCTCGTGGCGAGCTCCATCCTGCACATGCTCAACCTCGCC 147
ORYZA CTGCCGCCCGGGGGACTCCATCGTCGTCTCCAGCTCCTTCTCGCACATGCTCAACCTCGCC 300
SORGHUM
MAIZE AACCTGGCCGAGGAGGTGCAGATCGCGCACCGCCGCCGCAACAGCAAGCTCAAGAAAGGT 207
ORYZA AACCTCGCCGAGGAGGTGCAGATCGCGCACCGCCGCCGGAT---CAAGCTCAAGCGCGGG 357
SORGHUM
MAIZE GGGTTCGCCGACGAGGGCTCCGCCACCACCGAGTCCGACATCGAGGAGACGCTCAAGCGC 267
ORYZA GACTTCGCCGACGAGGCCTCCGCGCCCACCGAGTCCGACATCGAGGAGACGCTCAAGCGC 417
SORGHUM
MAIZE CTCGTGTCCGAGGTCGGCAAGTC---------- 290
ORYZA CTCGTCACCCAGCTCGGCAAGTCGAGGGAGGAGGTGTTCGATGCGCTCAAGAACCAGACC 477
SORGHUM
MAIZE -----CCCCG---------- 295
ORYZA GTTGACCTCGTGTTCACCGCGCATCCAACGCAGTCCGTCAGGAGGTCCCTGCTCCAGAAG 537
SORGHUM
MAIZE --------GATCCGGAATTGTCTGACCCAGCTGAATGCCAAGGACATCACTGACGACGAC 347
ORYZA CACGGGAGGATCCGGAATTGCCTGAGGCAGTTATATGCGAAAGACATCACTGCTGATGAC 597
SORGHUM
MAIZE AAGCAGGAGCTCGATGAGCCTCTGCACAGACGAGATCCAAGCAGCCTTCAGAACCGATGAA 407
ORYZA AAGCAGGAGCTTGATGAGGCCCTGCAGAGGCAGATTCAAGCAGCTTTCAGAACTGATGAA 657
SORGHUM
MAIZE ATCAGGAGGGCACAACCCACCCCCCAGGACGAAATGCGCTATGGGATGAGCTACATCCAT 467
ORYZA ATCCGCAGGACTCCTCCCACTCCTCAGGATCAAATGCGCGCCGGGATGAGTTATTTTCAT 717
SORGHUM
MAIZE GAGACTGTATGGAAGGGCGTGCCTAAGTTCTTGCGCCGTGTGGATACAGCCCTGAAGAAT 527
ORYZA GAAACTATATGGAAGGGTGTACCAAAGTTCTTGCGCCGCATTGATACTGCTCTGAAAAAT 777
SORGHUM
MAIZE ATCGGCATCAATGAGCGCCTTCCCTACAATGTTTCTCTCATTCGGTTCTCTTCTTGGATG 587
ORYZA ATTGGGATTAATGAGCGTCTTCCTTACAATGCTCCTCTCATCCAGTTCTCATCCTGGATG 837
SORGHUM
MAIZE GGTGGTGACCGCGATGGAAATCCAAGAGTTACCCCGGAGGTGACAAGAGATGTATGCTTG 647
ORYZA CGTGGTGACCGTGATGGAAACCCAAGAGTTACACCAGAGGTTACAAGAGATGTATGCTTG 897
SORGHUM
MAIZE CTGGCCAGAATGATGGCTGCAAACTTGTACATCGATCAGATTGAAGAGCTGATGTTTGAG 707
ORYZA TTGGCAAGAATGATGGCTGCTAACCTGTACTTCTCTCAGATAGAAGATCTGATGTTTGAG 957
SORGHUM
MAIZE CTCTCTATGTGGCGCTGCAACGATGAGCTTCGTGTTCGTGCCGAAGAGCTCCACAGTTCG 767
ORYZA CTCTCTATGTGGCGCTGCAGTGATGAACTTCGAATTCGTGCAGATGATTTGCATTGCTCC 1017
SORGHUM


MAIZE TCTGGTTCCAAAGTTACCAAGTATTACATAGAATTCTGGAAGCAAATTCCTCCAAACGAG 827
ORYZA TCCAG---AAAAGCTGCAAAGCACTATATAGAATTCTGGAAGCAAATTCCTCCAAATGAG 1074
SORGHUM
MAIZE CCCTACCGGGTGATACTAGGCCATGTAAGGGACAAGCTGTACAACACACGCGAGCGTGCT 887
ORYZA CCTTATCGTGTCATACTTGGTGGTGTCAGGGATAAATTGTACTATACGCGTGAACGTACT 1134
SORGHUM
MAIZE CGCCATCTGCTGGCTTCTGGAGTTTCTGAAATTTCAGCGGAATCGTCATTTACCAGTATC 947
ORYZA CGCCATCTATTGACAACTGGAGTTTCTGAAATTCCAGAGGAGGCAACGTTTACTAATGTT 1194
SORGHUM
MAIZE GAAGAGTTCCTTGAGCCACTTGAGCTGTGCTACAAATCACTGTGTGACTGCGGCGACAAG 1007
ORYZA GAAGAGTTTCTGGAGCCGCTTGAGCTGTGCTACAGATCATTATGTGCTTGTGGTGACAAA 1254
SORGHUM
MAIZE GCCATCGCGGACGGGAGCCTCCTGGACCTCCTGCGCCAGGTTTTCACGTTCGGGCTCTCC 1067
ORYZA CCTATAGCTGATGGAAGCCTTCTTGATTTCTTGCGTCAAGTATCAACTTTTGGGCTGGCT 1314
SORGHUM
MAIZE CTGGTGAAGCTGGACATCCGGCAGGAGTCGGAGCGGCACACCGACGTGATCGACGCCATC 1127
ORYZA CTTGTAAAACTTGACATAAGGCAGGAGTCTGATCGACACACTGATGTCCTTGATGCCATA 1374
SORGHUM
MAIZE ACCACGCACCTCGGCATCGGGTCGTACCGCGAGTGGTCCGAGGACAAGCGGCAGGAGTGG 1187
ORYZA ACTACATATCTTGGGATTGGATCTTACGCTGAATGGTCTGAGGAGAAACGCCAGGATTGG 1434
SORGHUM
MAIZE CTGCTGTCGGAGCTGCGAGGCAAGCGCCCGCTGCTGCCCCCGGACCTTCCCCAGACCGAG 1247
ORYZA CTTTTGTCGGAACTAAGGGCAAACGTCCTTTGTTTGGTCCTGATCTTCCTCAGACTGAA 1494
SORGHUM
MAIZE GAGATCGCCGACGTCATCGGCGCGTTCCACGTCCTCGCGGAGCTCCCGCCCGACAGCTTC 1307
ORYZA GAAATTGCTGATGTTTTAGGAACATTTCACGTCCTTGCTGAGCTCCCGGCAGATTGTTTT 1554
SORGHUM ATGAGTGCGGAGCTGAAGAGGGACTACGAGATAGGCGCGGAGATCGGTCGCGGCCGCTTC 60
MAIZE GGCCCCTACATCATCTCCATGGCGACGGCCCCCTCGGACGTGCTCGCCGTGGAGCTCCTG 1367
ORYZA GGTGCTTACATCATCTCAATGGCAACTGCACCATCTGATGTGCTTGCTGTGGAGCTTCTA 1614
SORGHUM GGGGTGGTCCACCGCTGCACGTCCCGCGCCACCGGCGAGGCGTTCCGCCGTCAAGTCCGTG 120
MAIZE CAGCGCGAGTG--CGGCGTGCGGCCAGCCGTGCCCG-TGGTGCCGCTGTTCGAAAGGCTG 1424
ORYZA CAGCGGGAGTG--CCATATAAAACAGCCTCTGCGAG-TTGTTCCCCTATTTGAGAAACTT 1671
SORGHUM GACCGGTCGCAGCTGGCCGACGACCTGGACCGCGAGCTCGCGGAGCTGGAGCCCAAGCTG 180
MAIZE GCCAGCCTGCAGTCGGCGCCCGCGTCCGTGGAGCGCCTCTTCTCGGTGGACTGGTACATG 1484
ORYZA GCAGATCTTGAAGCAGCTCCAGCAGCGGTGGCACGACTATTTTCAATAGACTGGTACATG 1731
SORGHUM GCGCAGCTC--GCCGGCGCGGGCAACCCCGGCGTGGTGCAGACGCACGCGGTG-TACGAG 237
MAIZE GACCGGATCAAGGGCAAGCAGCAGGTCATGGTCGGCTACTCCGACTCCGGCAAGGACGC- 1543
ORYZA AATAGGATTAATGGCAAGCAGGAGGTCATGATTGGATACTCAGACTCTGGCAAGGATGC- 1790
SORGHUM GACGAGAC---GTGGACCCACACGGTGATGGACCTGTGCGCGGGCCCGGACCTGCTCGAG 294
MAIZE CGGCCGCCTGTCCGCGGCGTGGCAGCTGTACAGGGCGCAGGAGGAGATGGCGCAGGTGGC 1603
ORYZA TGGTCGTCTTTCTGCAGCATGGCAAATGTACAAAGCACAAGAGGAGCTTGTCAAGGTGGC 1850
SORGHUM TGGGTGCGCCTCCGCCGCGGCGCGCCTGTGCCGGA-GCCCCTGGCGGCCGCCATCGTCGC 353


MAIZE CAAGCGCTACGGCGTCAAGCTCACCTTGTTCCACGGCCGCGGAGGCACCGTGGGCAGGG- 1662
ORYZA AAAGCATTATGGTGTAAAGTTGACAATGTTCCATGGAAGAGGTGGAACAGTTGGCAGAG- 1909
SORGHUM ---GCAGGTCGCCCAGGCGCTCGGCGCTCTGCCACCGCCGCGGCGTCGCCCACCGCGACGT 410
MAIZE GTGGCGGGCCCACGCACCTTGCCATCCTGTCCCAGCCGCCGGACACCATCAACGGGTCCA 1722
ORYZA GAGGTGGTCCCAGTCATCTTGCCATATTATCTCAACCACCAGACACGATACATGGATCAC 1969
SORGHUM AAAGCCCGACAACATCCTCATCGACACCGCCCCCGCAGGGAGCGACGGCGAGGACGA 470
MAIZE TCCGTGTGACGGTGC--AGGGCGAGGTCATCGAGTTCTGCTTCGGGGAGGAGCACCTGTG 1780
ORYZA TTCGTGTAACAGTAC--AGGGCGAGGTTATTGAACACTCGTTTGGAGAGGAACACTTGTG 2027
SORGHUM GCAAGAGGACAGCGGCGAGGCCGAGGAGGCCGACACGGCGCCGCGCGCGGCTGGCGGA 530
MAIZE CTTCCAGACTCTGCAGCGCTTCACGGCCGCCACGCTGGAGCACGGCATGCACCCGCCGGT 1840
ORYZA CTTTAGAACTCTGCAGCGCTTTACTGCAGCTACTCTTGAGCATGGAATGCATCCTCCAAT 2087
SORGHUM CTTCGGGTCGGCGGCGTGGGTGGGCGCCGGTGGGCTGGGCCGCCGCGGAAGGCCTGGTGGG 590
MAIZE CTCTCCCAAGCCCGAGTGGCGCAAGCTCATGGACGAGATGGCGGTCGTGGCCACGGAGGA 1900
ORYZA TTCCCCCAAACCAGAATGGCGTGCTCTAATGGATGAGATGGCTGTTGTGGCAACAAAGGA 2147
SORGHUM GACGCCCCACTACGTGGCGCCCGAGGTGGTCGGCGG--CGGCGAGTACGGC-GCGAAGGC 647
MAIZE GTACCGCTCCGTCGTCGTCAAGGAGGCCCGCTTCGTCGAGTACTTCAGATCGGCTACACC 1960
ORYZA ATATCGATCAATCGTCTTCAAGGAGCCACGCTTTGTTGAATACTTCCGATCGGCAACACC 2207
SORGHUM GGACGTGTGGAGCGCCGGCGTGGTGATGTACGCGCTGCTCTCCGGCGGCGCGCTCCCGTT 707
MAIZE GGAGACCGAGTACGGGAGGATGAACATCGGCAGCCGGCCAGCCAAGAGGAGGCCCGGCGG 2020
ORYZA TGAGACAGAATATGGCAGGATGAACATTGGTAGCCGACCATCAAAGAGAAAGCCTAGTGG 2267
SORGHUM CGGCGGCGAG-ACCGCGGCGGAGGTGCTAGCGGCCGTGCTGCGGGGCAGCGTCCGGTTCC 766
MAIZE CGGCATCACGACCCTGCCGCCATCCCCTGGATCTTCTCGTGGACCCAGAC-CAGGTTCC 2079
ORYZA TGGCATTGAATCGCTCCGTGCAATTCCTTGGATTTTTGCTTGGACACAAAC-TAGGTTTC 2326
SORGHUM CGCCCAGGCTGTTCAGCGGGGTGTCCCCCGCGGCCAAGGACCTGTTGAGGCCATGATCT 826
MAIZE ACCTCCCCGTGTGGCTGGGAGTCGGCGCCGCATTCAAGTTCGCCATCGACAAGGACGTCA 2139
ORYZA ATCTTCCTGTATGGCTTGGATTTGGTGGAGCGTTCAAACATATAATGCAGAAGGATATCA 2386
SORGHUM GCGGCGACGAGTGGCGGAGGTTCACCGCCGAG--CAAGTCC-TCGCTCACCCGTGGATCG 883
MAIZE GGAACTTCCAGGTCCTCAAAGAGATGTACAACGAGTGGCCATTCTTCAGGGTCACCCTGG 2199
ORYZA GGAACATCCATACACTGAAAGAAATGTACAATGAGTGGCCATTCTTCAGGGTCACCTTGG 2446
SORGHUM TGAGCGGCGGAGGAGCACGGGCAAT------CGAGCGGCCAACCTGA---------- 924
MAIZE ACCTGCTGGAGATGGTTTTCGCCAAGGGAGACCCCGGCATTGCCGGCTTGTATGACGAGC 2259
ORYZA ACTTGCTTGAGATGGTTTTCGCTAAGGGAGATCCAGGAATTGCTGCTTTATATGACAAAT 2506
SORGHUM ----------
MAIZE TGCTTGTGGCAGAAGAACTCAAGCCCTTTGGGAAGCAGCTCAGGGACAAATACGTGGAGA 2319
ORYZA TGCTTGTGGCAGGTGATCTGCAATCCTTTGGAGAGCAGCTGAGAAACAACTTTGAAGAGA 2566
SORGHUM ----------
MAIZE CACAGCAGCTTCTCCTCCAG---------- 2339
ORYZA CAAAACAGCTACTCCTTCAGGTTGCCGGTCACAAGGATATTCTGGAAGGGGATCCTTACC 2626
SORGHUM ----------
MAIZE ----------
ORYZA TGAAGCAGCGTCTGCGGCTGCGTGAGTCGTACATCACTACCTTGAACGTTTGCCAAGCCT 2686
SORGHUM ----------

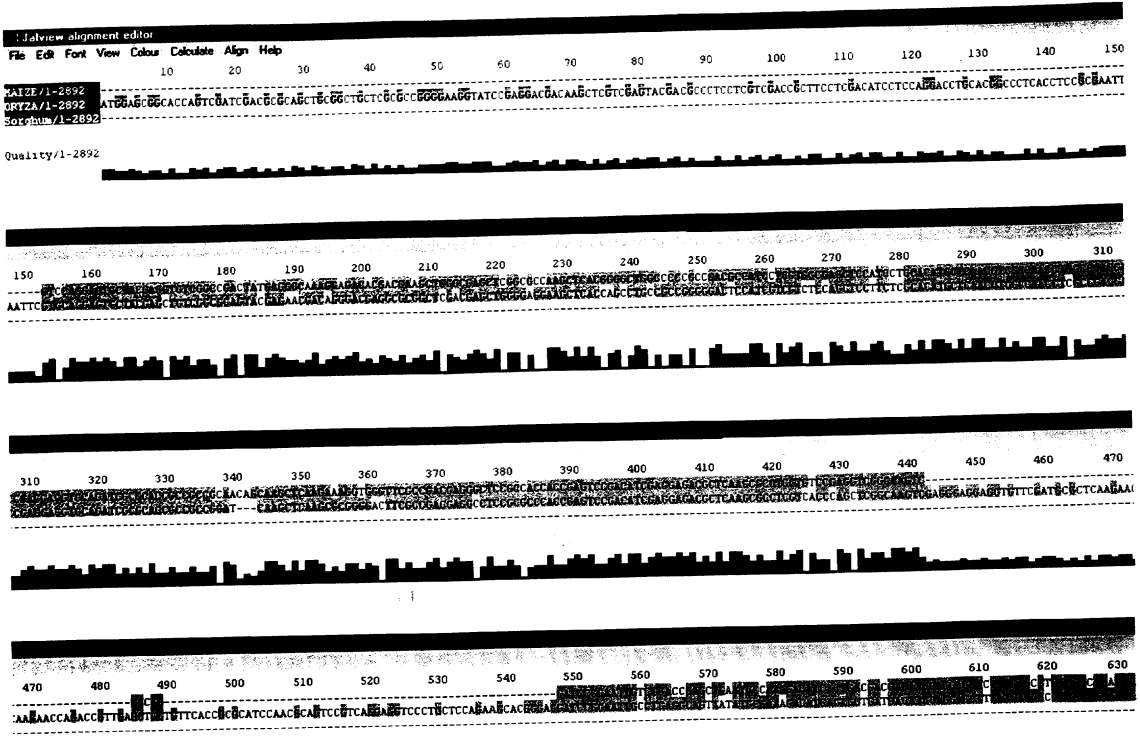
```
MAIZE           ------------------------------------------------------------
ORYZA           ACACCCTGAAGCGGATAAGGGACCCTAGCTTCGAGGTGATGTCGCAGCCGGCCTTGTCGA 2746
SORGHUM         ------------------------------------------------------------

MAIZE           ------------------------------------------------------------
ORYZA           AGGAGTTCGTCGACAGCAACCAGCCTGCCGAGCTGGTCCAACTGAACGCTGCGAGCGAGT 2806
SORGHUM         ------------------------------------------------------------

MAIZE           ------------------------------------------------------------
ORYZA           ACGCACCTGGCCTGGAGGACACCCTCATCCTGACGATGAAGGGCATTGCTGCCGGCATGC 2866
SORGHUM         ------------------------------------------------------------

MAIZE           ------------
ORYZA           AGAACACAGGCT 2878
SORGHUM         ------------
```

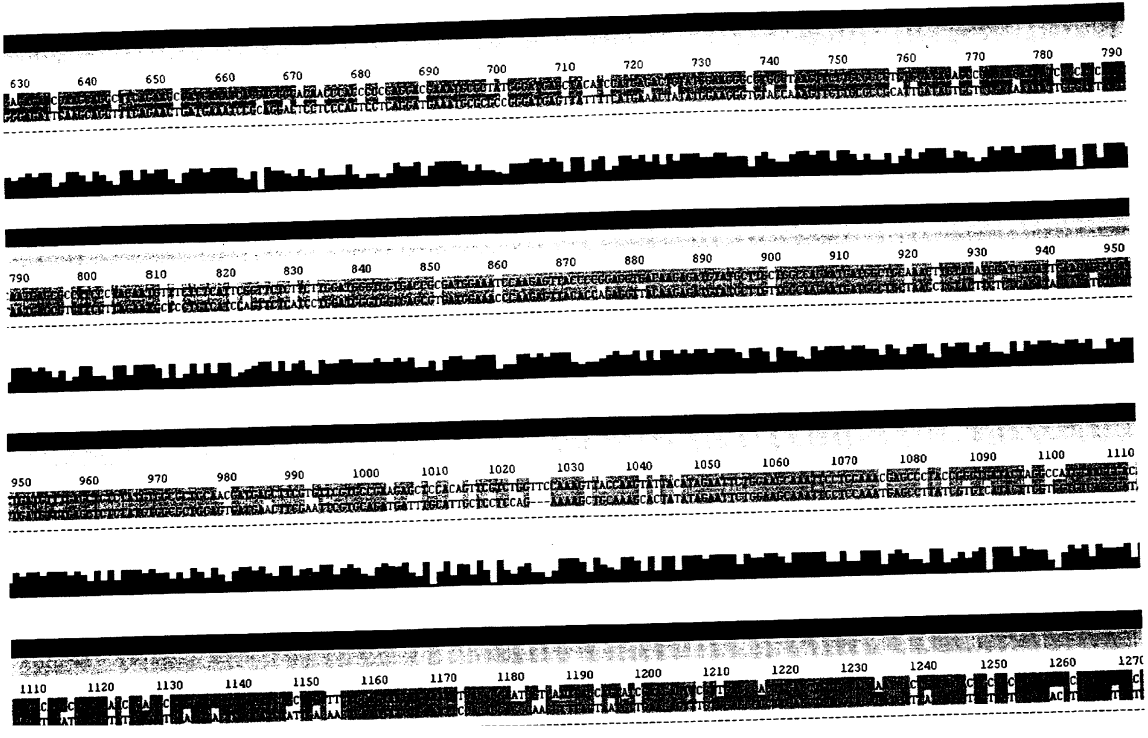
Your guide tree:

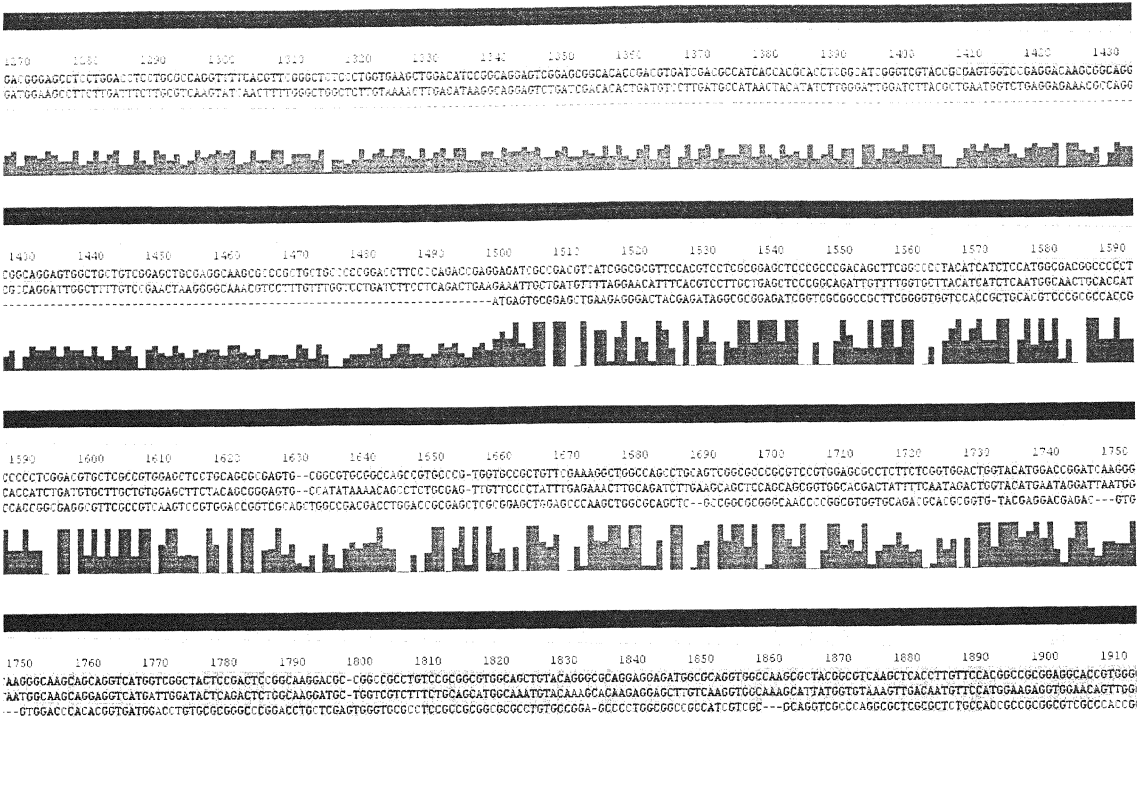
```
(
MAIZE:0.03716,
ORYZA:0.21465,
SORGHUM:0.669551);
```

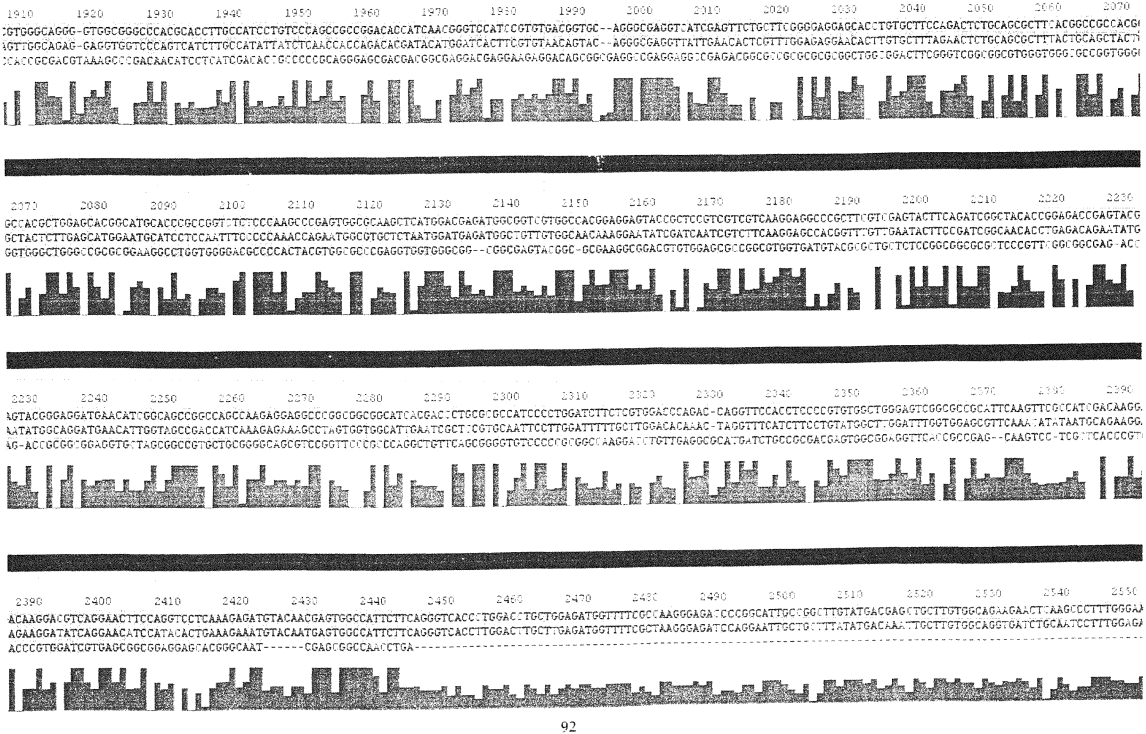
2.10.8 Alignment Graph

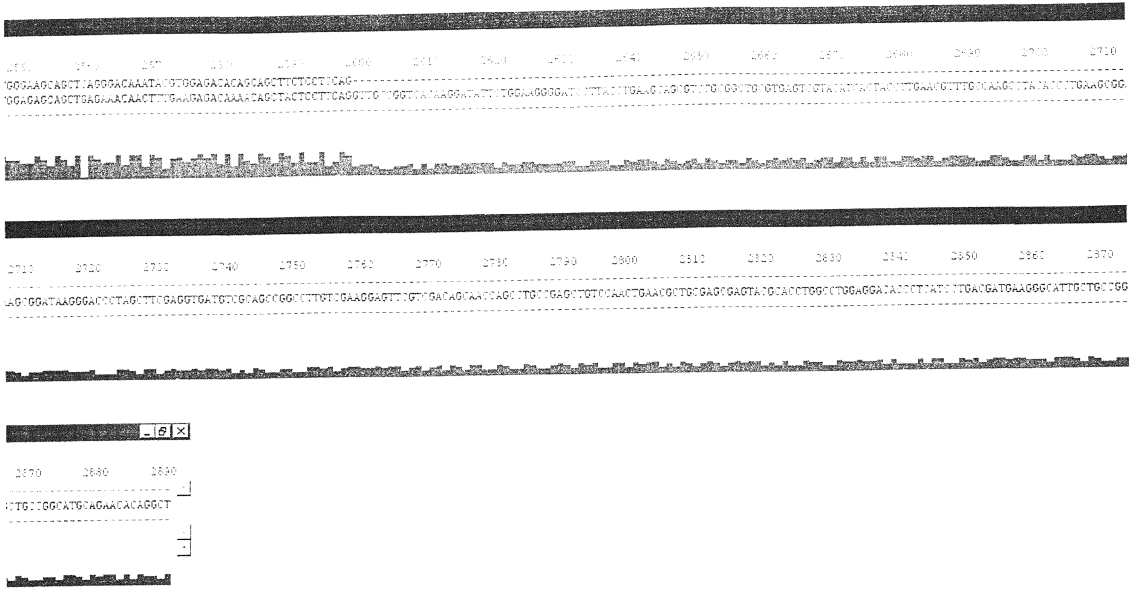
An alignment graph among Maize, oryza sativa and Sorghum bicolor phosphenol carboxylase gene will obtain.


Jalview alignment editor
File Edit Font View Colour Calculate Align Help
MAIZE/1-2892
ORYZA/1-2892
Sorghum/1-2892
Quality/1-2892


630 640 650 660 670 680 690 700 710 720 730 740 750 760 770 780 790
790 800 810 820 830 840 850 860 870 880 890 900 910 920 930 940 950
950 960 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110
1110 1120 1130 1140 1150 1160 1170 1180 1190 1200 1210 1220 1230 1240 1250 1260 127

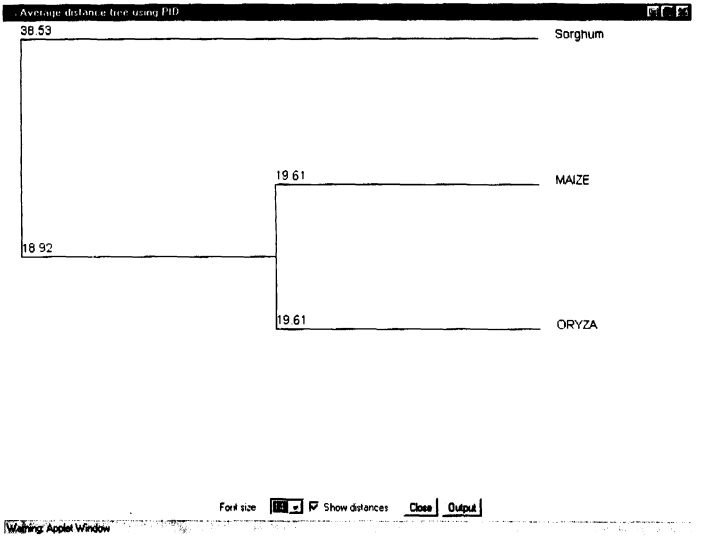






2.10.9 Phylogenetic Tree

Construct the phylogenetic tree among Maize, oryza sativa & Sorghum bicolor phosphenol pyruvate carboxylase.



2.10.10 Tree Analysis

Check the two species which having maximum alignment By tree Analysis. Take the species, which have maximum alignment. By Tree Analysis it has been found that there is maximum alignment between Maize and oryza sativa phosphenol pyruvate carboxylase gene.

2.10.11 Design Primers For The Sequences Of Maize:

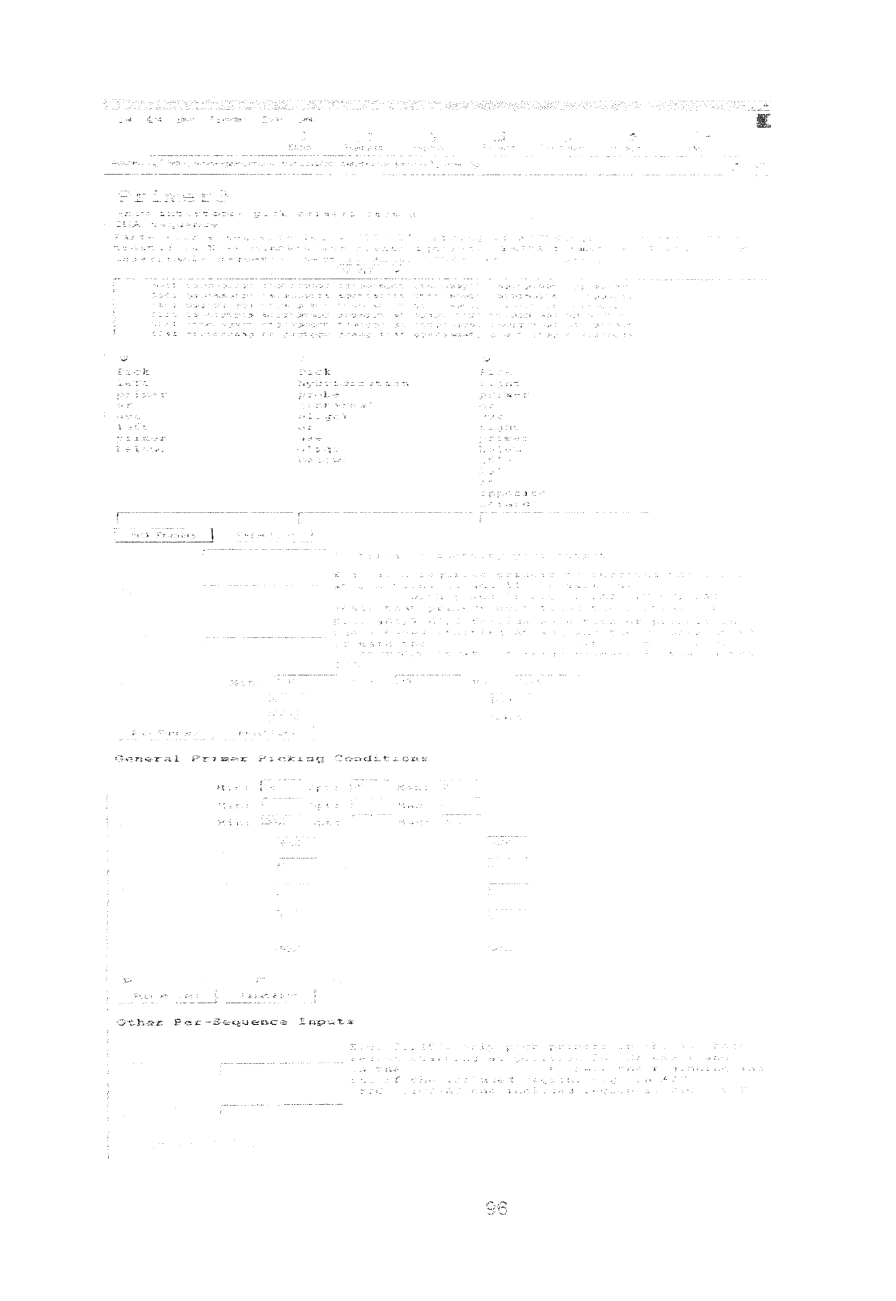
Design two sets of Left and Right Primers for the exon regions of maize sequences by using PRIMER3. One set of primers are sufficient for PCR, but the second set is taken as additional primer in the event the first fails hybridize.

2.10.12 Selection Of First Set Of Primers

Design left primer from one Exon of the sequence and right primer from another Exon of the sequence. The distance between the two primers should be about 500 to 1000 base pairs. Take the sequence from the species, which has both Exons and introns.i.e we have taken Maize sequence because it has both Exons and introns. But rice sequence for this enzyme has only coding regions i.e.Exons.so we can't design primer from rice for sequencing of DNA but can be used as a marker.

2.10.12.1 Primer3 For Left Primer

Take one Exon sequence of the maize and then paste it in PRIMER3 and pick primer. The region of the exon from which the left primer is designedare 4454-5452. The left primer is picked from the Exons, which is right of the exon from which the right primer is picked. The parameters for both the primers should be same.the product size should be Min: 100,Opt: 200,Max: 1000.Primer size should be Min: 18bp,Opt: 20bp,Max: 27bp.Primer annealing temperature should be Min: 57,Opt: 60,Max: 63 degree celcius. The GC% should be Min: 20and Max: 80.



General Primer Picking Conditions

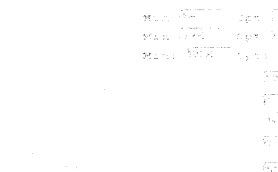
Other Per-Sequence Inputs

Objective Function Penalty Weights for Primers

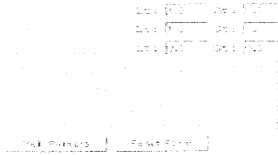
Objective Function Penalty Weights for Primer Pairs

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo (Internal Oligo) General Conditions



Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)



2.10.12.2 Primer3 Output For Left Primer.

Put the sequence of Exon of the maize for the left primer in the primer3 output and by clicking pick primer the left primer and right primer both will be obtained. But, left primer only has to be taken into cosideration. Primer3 output for left primer are given below.

Primer3 Output

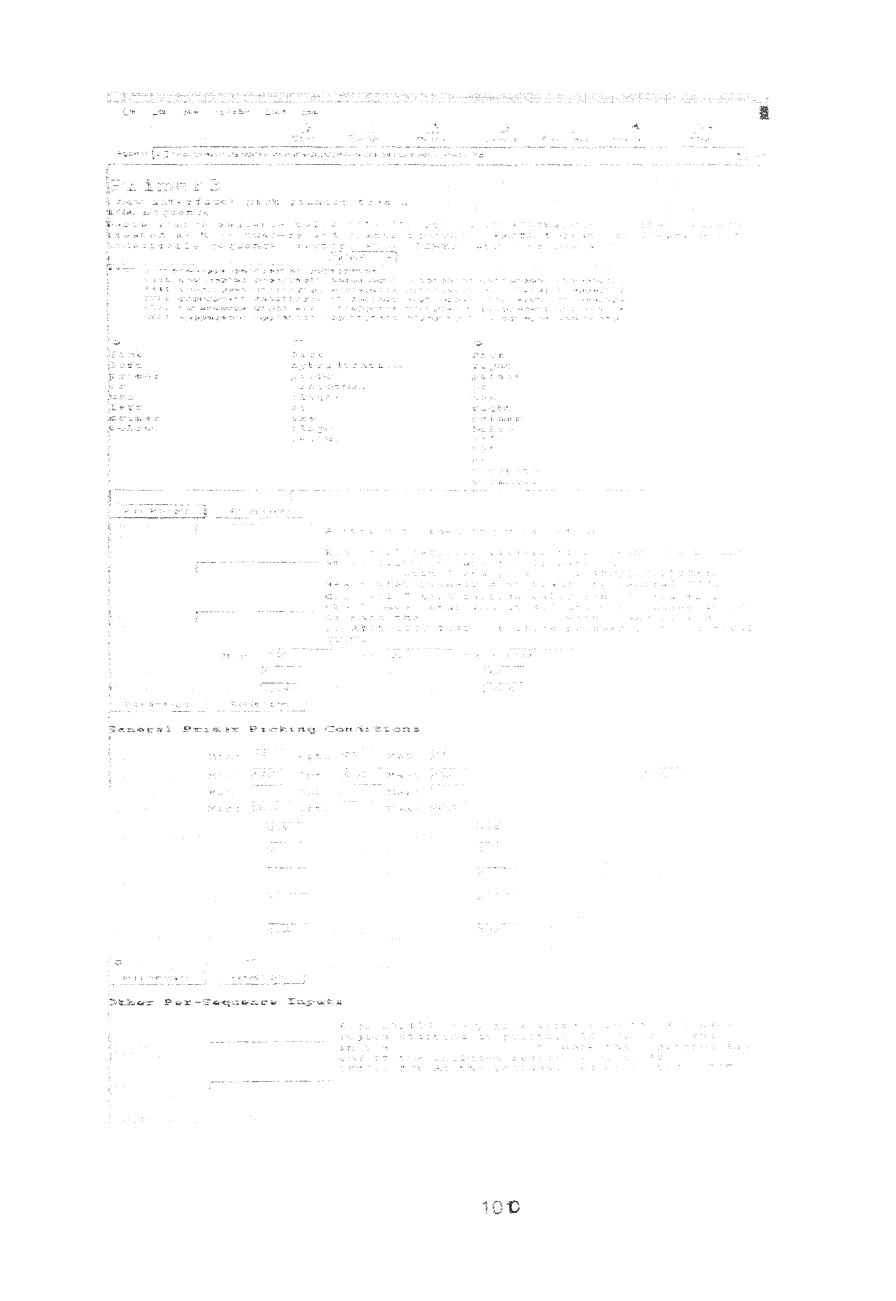
```
WARNING: Numbers in input sequence were deleted.

No mispriming library specified
Using 1-based sequence positions
OLIGO                 start  len      tm    gc%   any   3' seq
LEFT PRIMER             663   20   60.83   55.00  7.00  1.00 cgtcaagctcaccttgttcc
RIGHT PRIMER            843   20   60.33   55.00  7.00  0.00 ctgcagagtctggaagcaca
SEQUENCE SIZE: 999
INCLUDED REGION SIZE: 999

PRODUCT SIZE: 181, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 3.00

     1 ttccttgagccacttgagctgtgctacaaatcactgtgtgactgcggcgacaaggccatc

    61 gcggacgggagcctcctggacctcctgcgccaggttttcacgttcgggctctccctggtg

   121 aagctggacatccggcaggagtcggagcggcacaccgacgtgatcgacgccatcaccacg

   181 caccteggcatcgggtcgtaccgcgagtggtccgaggacaagcggcaggaglggctgctg

   241 tcggagctgcgaggcaagcgcccgctgctgcccccggaccttcccccagaccgaggagatc

   301 gccgacgtcatcggcgcgttccacgtcctcgcggagctcccgcccgacagcttcggcccc

   361 tacatcatctccatggcgacggccccctcggacgtgctcgccgtggagctcctgcagcgc

   421 gagtgcggcgtgcggccagccgtgcccgtggtgccgctgttcgaaaggctggccagcctg

   481 cagtcggcgcccgcgtccgtggagcgcctcttctcggtggactggtacatggaccggatc

   541 aagggcaagcagcaggtcatggtcggctactccgactccggcaaggacgccggccgcctg

   601 tccgcggcgtggcagctgtacagggcgcaggaggagatggcgcaggtggccaagcgctac

   661 ggcgtcaagctcaccttgttccacggccgcggaggcaccgtgggcaggggtggcgggccc
         >>>>>>>>>>>>>>>>>>>

   721 acgcaccttgccatcctgtcccagccgccggacaccatcaacgggtccatccgtgtgacg
```

```
  781 gtgcagggcgaggtcatcgagttctgcttcggggaggagcacctgtgcttccagactctg
                                                <<<<<<<<<<<<<<<<<

  841 cagcgcttcacggccgccacgctggagcacggcatgcacccgccggtctctcccaagccc
      <<<

  901 gagtggcgcaagctcatggacgagatggcggtcgtggccacggaggagtaccgctccgtc

  961 gtcgtcaaggaggcccgcttcgtcgagtacttcagatcg

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

ADDITIONAL OLIGOS

 1 LEFT PRIMER         662   20   60.99   55.00  7.00  3.00 gcgtcaagctcaccttgttc
   RIGHT PRIMER        843   20   60.33   55.00  7.00  0.00 ctgcagagtctggaagcaca
   PRODUCT SIZE: 182, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

 2 LEFT PRIMER         662   20   60.99   55.00  7.00  3.00 gcgtcaagctcaccttgttc
   RIGHT PRIMER        842   20   60.33   55.00  7.00  1.00 tgcagagtctggaagcacag
   PRODUCT SIZE: 181, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

 3 LEFT PRIMER         665   20   58.85   50.00  7.00  1.00 tcaagctcaccttgttccac
   RIGHT PRIMER        843   20   60.33   55.00  7.00  0.00 ctgcagagtctggaagcaca
   PRODUCT SIZE: 179, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

 4 LEFT PRIMER         664   21   59.76   52.38  7.00  1.00 gtcaagctcaccttgttccac
   RIGHT PRIMER        843   20   60.33   55.00  7.00  0.00 ctgcagagtctggaagcaca
   PRODUCT SIZE: 180, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 3.00

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   8334     0     0     0   770     0   147  6981     0    85     0    75   276
Right  8279     0     0     0   770     0    85  7041     0    32     0    82   269
Pair Stats:
considered 1342, unacceptable product size 828, high end compl 82, ok 432
primer3 release 0.9

(primer3_www_results.cgi v 0.2)
```

note down the left primer from the above PRIMER3 output.

2.10.12.3 Primer3 For Right Primer

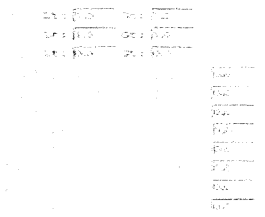
Pick the next Exon region sequence after the left exon region from which the left primer was designed for the right primer. The right primer from this region is designed as descbried about for the left primer in Primer 3 but only the right primer is to be considered from the output for right primer are given below..The parameters are similar to that of PRIMER3 for left primer.the exon region, which are taken to design the right primer are5550-5936.



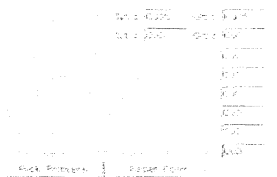
Several Primary Picking Conditions

Other Pre-Sequence Inputs

Objective Function Penalty Weights for Primers

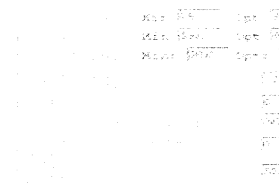


Objective Function Penalty Weights for Primer Pairs

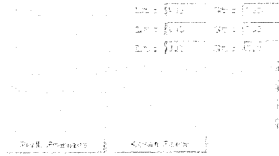


Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo (Internal Oligo) General Conditions



Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)



2.10.12.4 Primer3 Output For Right Primer

By clicking pick primer we will get the right primer and left primer both. but right primer only has to be consider.primer3 output for right primer are given below.

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              1   20   60.28   60.00   4.00   3.00 gctacaccggagaccgagta
RIGHT PRIMER           197   20   59.70   50.00   7.00   3.00 tggaagttcctgacgtcctt
SEQUENCE SIZE: 387
INCLUDED REGION SIZE: 387

PRODUCT SIZE: 197, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00

    1 gctacaccggagaccgagtacgggaggatgaacatcggcagccggccagccaagaggaag
      >>>>>>>>>>>>>>>>>>>

   61 cccggcggcggcatcacgaccctgcgcgccatcccctggatcttctcgtggaccccagacc

  121 aggttccacctccccgtgtggctgggagtcggcgccgcattcaagttcgccatcgacaag
                                                                ...

  181 gacgtcaggaacttccaggtcctcaaagaggatgtacaacgagtggccattcttcagggtc
      <<<<<<<<<<<<<<<<<

  241 accctggaccctgctggagatggttttcgccaagggagaccccggcattgccggcttgtat

  301 gacgagctgcttgtggcagaagaactcaagccctttgggaagcagctcagggacaaatac

  361 gtggaggcacagcaggcttctcctccag


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

ADDITIONAL OLIGOS
                     start  len      tm     gc%    any    3' seq

 1 LEFT PRIMER           1   20   60.28   60.00   4.00   3.00 gctacaccggagaccgagta
   RIGHT PRIMER        196   20   59.70   55.00   7.00   1.00 ggaagttcctgacgtccttg
   PRODUCT SIZE: 196, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00

 2 LEFT PRIMER         174   20   60.30   55.00   6.00   1.00 cgacaaggacgtcaggaact
   RIGHT PRIMER        377   20   59.65   55.00   5.00   2.00 agctgctgtgtctccacgta
   PRODUCT SIZE: 204, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00

 3 LEFT PRIMER           1   20   60.28   60.00   4.00   3.00 gctacaccggagaccgagta
   RIGHT PRIMER        193   20   60.30   55.00   6.00   2.00 agttcctgacgtccttgtcg
   PRODUCT SIZE: 193, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00

 4 LEFT PRIMER         186   20   60.23   55.00   7.00   3.00 caggaacttccaggtcctca
   RIGHT PRIMER        377   20   59.65   55.00   5.00   2.00 agctgctgtgtctccacgta
   PRODUCT SIZE: 192, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00
```

```
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   req   GC  clamp   low  high compl compl    X  stab    ok
Left    2541     0     0     0    86     0   144  1972     0    10    0    41   288
Right   2438     0     0     0     7     0   224  1738     1     8    0    52   418
Pair Stats:
considered 224, unacceptable product size 159, high end compl 13, ok 52
primer3 release 0.9

(primer3_www_results.cgi v 0.2)
```

Note down the right primer from the above PRIMER3 output.

2.10.12.5 Calculation Of Product Size

Left primer sequence size is	999
Left primer starting point is	663
Left primer total length is	336
Introns length is	98
Right primer sequence size is	387
Right primer ending point is	197
Total right primer length is	190

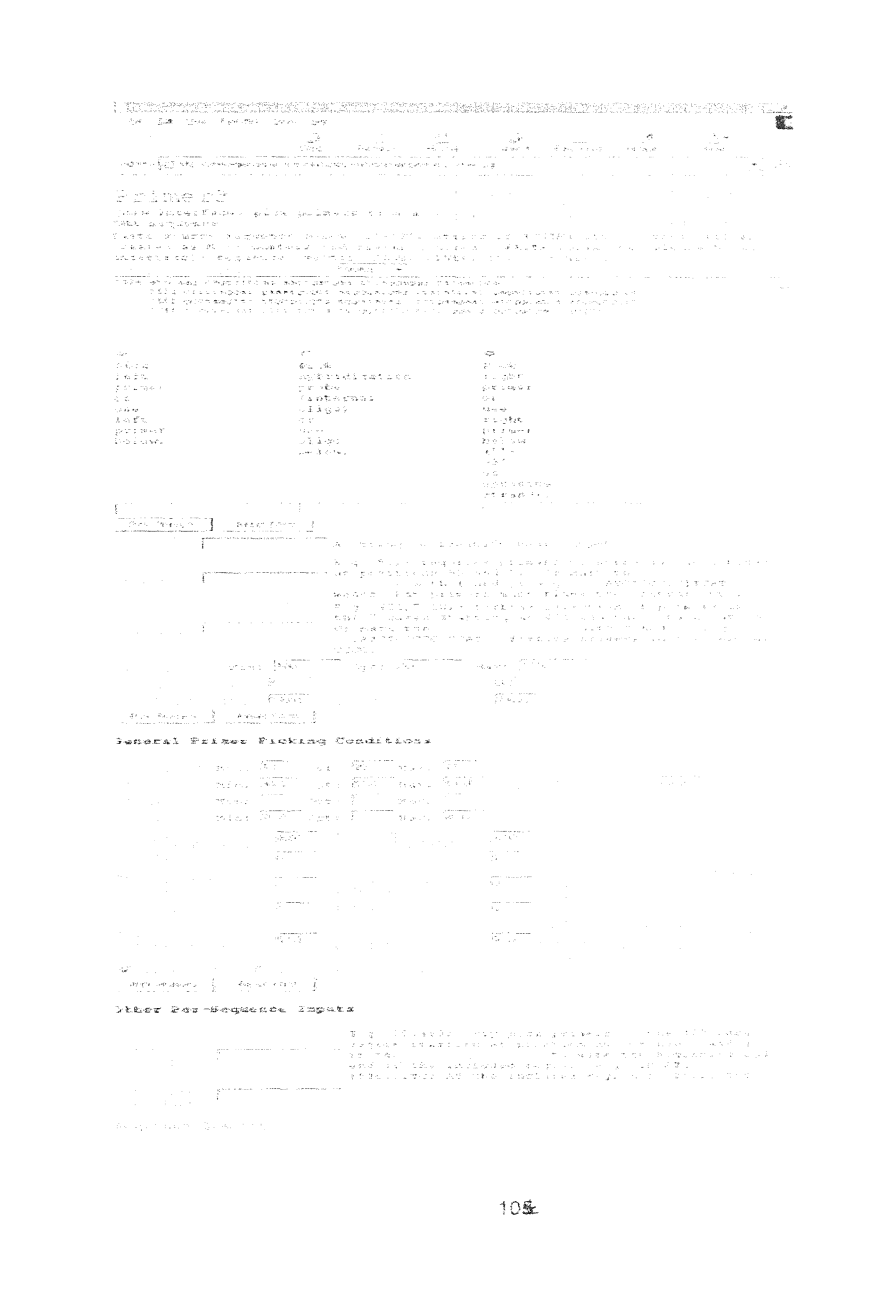
Product size = Left primer total length is (336) +introns(98) +Right primer ending point is (197) = 631

2.10.13 Selection Of Second Set Of Primers

Repeat the same process as explained in the first set of primers.Set the parameters for designing the primers for PCR.

2.10.13 .1 Primer3 For Left Primer

The parameters set to design the left primer for second set of primers are: the product size should be Min: 100,Opt: 200,Max: 1000.the primer annealing temperature is Min: 49, Opt:51, Max: 54.the GC% isMin : 20 and Max: 80.the primer size is Min: 18bp,Opt: 20bp,Max: 27bp. The exon regions which are taken to design the left primer are 3074-3296.



General Primer Picking Conditions

Other Per-Sequence Inputs

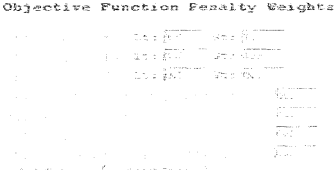
Objective Function Penalty Weights for Primers

Objective Function Penalty Weights for Primer Pairs

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo (Internal Oligo) General Conditions

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)



2.10.13 .2 Primer3 Output For Left Primer.

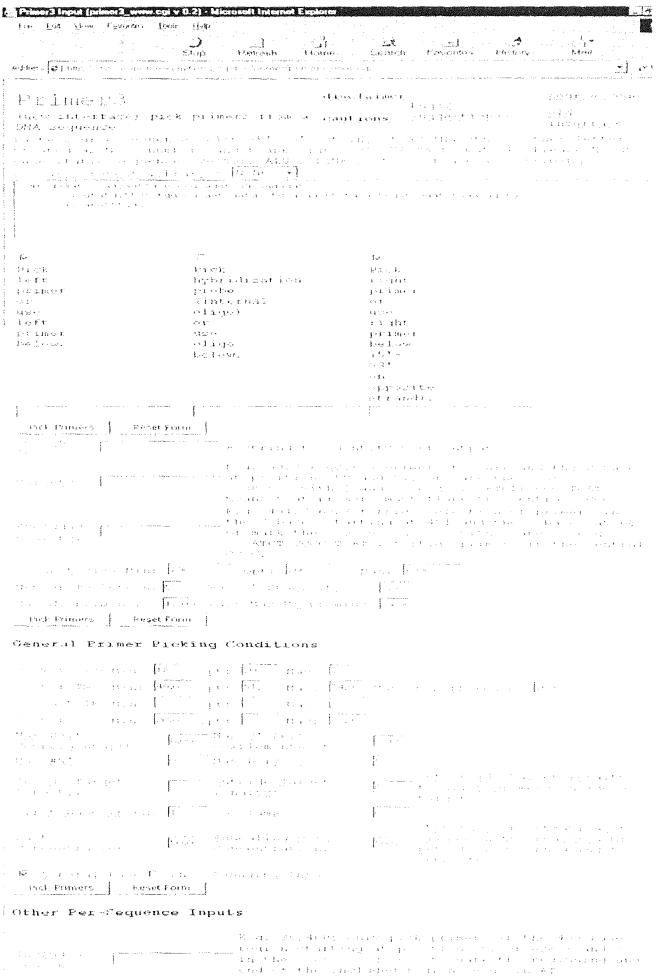
Primer3 Output

```
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           12   18   50.69   38.89  5.00  2.00 cttcagaaccgatgaaat
RIGHT PRIMER         191   20   51.23   35.00  5.00  3.00 cgaatgagagaaacattgta
SEQUENCE SIZE: 223
INCLUDED REGION SIZE: 223

PRODUCT SIZE: 180, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 2.00

    1 atccaagcagccttcagaaccgatgaaatcaggagggcacaaccacccccagcagacgaa
                >>>>>>>>>>>>>>>>>>

   61 atgcgctatgggatgagctacatccatgagactgtatggaagggcatgcctaagttcttg


  121 cgccgtgtggatacagccctgaagaatatcggcatcaatgagcgccttcctacaatgtt


  181 tctctcattcggttctgctttqqatgggtqgtgaccqcgatq
      <<<<<<<<<<

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

ADDITIONAL OLIGOS
                start  len      tm    gc%   any   3' seq

 1 LEFT PRIMER          12   18   50.69   38.89  5.00  2.00 cttcagaaccgatgaaat
   RIGHT PRIMER        185   20   51.05   40.00  5.00  0.00 agagaaacattgtagggaag
   PRODUCT SIZE: 174, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 0.00

 2 LEFT PRIMER          13   18   51.55   38.89  5.00  3.00 ttcagaaccgatgaaatc
   RIGHT PRIMER        191   20   51.23   35.00  5.00  3.00 cgaatgagagaaacattgta
   PRODUCT SIZE: 179, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00

 3 LEFT PRIMER          13   18   51.55   38.89  5.00  3.00 ttcagaaccgatgaaatc
   RIGHT PRIMER        185   20   51.05   40.00  5.00  0.00 agagaaacattgtagggaag
   PRODUCT SIZE: 173, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00

 4 LEFT PRIMER          12   18   50.69   38.89  5.00  2.00 cttcagaaccgatgaaat
   RIGHT PRIMER        188   20   51.69   40.00  5.00  0.00 atgagagaaacattgtaggg
   PRODUCT SIZE: 177, PAIR ANY COMPL: 3.00, PAIR 3' COMPL: 1.00

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    %  stab   ok
Left    1069     0     0     0     0     0    18   983     4    27    22     2   13
Right   1206     0     0     0     0     0    17  1107     0     1     0     4   77
Pair Stats:
considered 303, unacceptable product size 149, high end compl 4, ok 150
primer3 release 0.9
```

(primer3_www_results.cgi v 0.2)
note the left primer from the above PRIMER3 output.

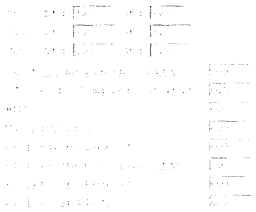
2.10.13 .3 Primer3 For Right Primer

The exon region which are taken to design the rigth primer are 3746-3849.the parameters are similar to that of left primer.

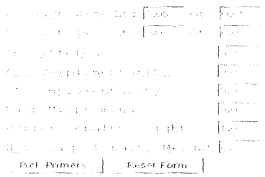

Primer3
Pick Primers
Reset Form
General Primer Picking Conditions
Pick Primers
Reset Form
Other Per-Sequence Inputs
Sequence Quality

Sequence Quality

Objective Function Penalty Weights for Primers

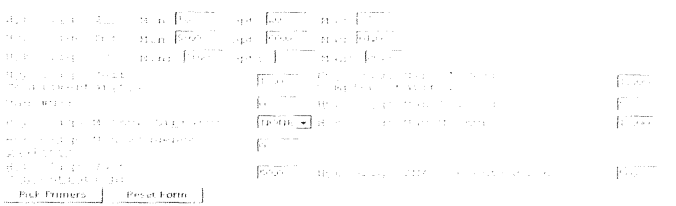


Objective Function Penalty Weights for Primer Pairs

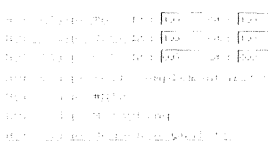


Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo (Internal Oligo) General Conditions

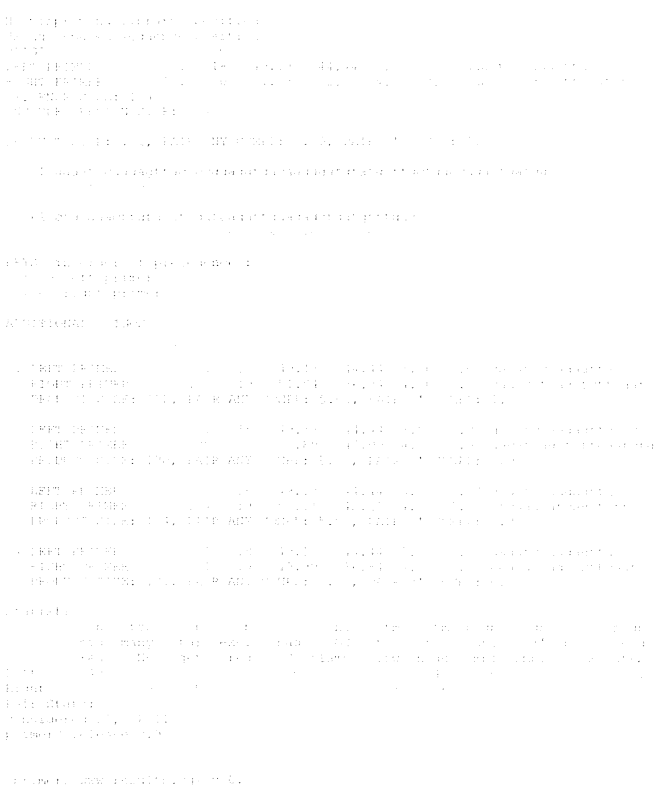


Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)



2.10.13 .4 Primer3 Output For Right Primer

Primer3 Output



note down the right primer from the above PRIMER3 output.

2.10.13 .5 Calculation Of The Product Size

Left primer sequence size is	223
Left primer starting point is	12
Left primer total length is	211
Introns length is	450
Right primer sequence size is	104
Right primer ending point is	101
Total right primer length is	101

Product size = Left primer total length is (211) +introns(101) +Right primer ending point is (450) = 762

2.10.14 Mitochondrial And Nuclear Enzyme

Mitochondrial and nuclear enzyme selected are Malate dehydrogenase and methyl transferase. The procedures of multiple alignments for these two enzymes are similar to that of chloroplast enzyme, which are explained above.

The summary sheet of the three enzymes is given in the table below

	Enzyme	Accessionnumber	Cerealname	Exonregions
1	malatedehydrogenase	M31965	soghumvulgare	603-783,873-963, 1461-1547,1643-1727, 1984-2165,2420-2490, 2570-2660,2749-2823, 2912-3013,3100-3222, 3303-3371,3456-3524, 3619-3672,3765-4638
	malatedehydrogenase	AF353203	Oryza sativa	90-1088
	malatedehydrogenase	AF007581	zea mays	78-1076
2	methyl transferase	AF063403	zea mays	1854-3316,3481-4960, 5107-5271,5361-5553, 5634-5745,5819-5966, 6069-6266,6364-6642, 6723-6884,6960-7178, 7259-7417
	methyl transferase	AF063403	zea mays	5107-5271,5361-5553
	methyl transferase	AB028870	hordeum vulgare	83-3349
	methyl transferase	AF387790	sorghum bicolor	110-1198
	methyl transferase	U76384	Triticum aestivum	52-1173
	methyl transferase	AF042332	Oryza sativa	187-1227
3	Phosphenolpyruvate carboxylase	EI7154	maize	1587-1981,2092-2176,3074-3296,3746-3849,3975-4065,4203-4357,4454-5452,5550-5936
	Phosphenolpyruvate carboxylase	AF399915	sorghum bicolor	21-944
	Phosphenolpyruvate carboxylase	AF271995	oryza sativa	188-3065

2.11 Results

For our study we have taken Phosphoenol pyruvate carboxylase a chloroplast enzyme, Malate dehydrogenase a mitochondrial enzyme in sorghum bicolor, maize and oryza sativa, and methyl transferase a nuclear enzyme in zea mays, Hordeum vulgare, sorghum bicolor, Triticum aestivum and oryza sativa. In case of Phosphoenol pyruvate carboxylase, Malate dehydrogenase and methyl transferase, there is a maximum alignment is between oryza sativa and maize, maize and sorghum bicolor, zea mays and Hordeum vulgare.the maximum alignment can be studied by alignment graph. The distance between the two species can be studied by phylogenetic tree. The exons of the above mentioned enzymes are used for designing primers for PCR.The summary sheet of the primers for the above enzymes are given in the table1.

Table1:

	Gene name	Accession number	Crop	Exons region	Leftprimer	Left temp	Rightprimer	Right temp	Size
1	Phosphenolpyruvate carboxylase	EI7154	zea mays	4454-5452, 5550-5936	cgtcaagctcaccttgttcc	60.83	tggaagttcctgacgtcctt	59.7	631
	Phosphenolpyruvate carboxylase	EI7154	zea mays	3074-3296, 3746-3849	cttcagaaccgatgaaat	50.69	aaacatcagctcttcaatct	51.53	762
2	malatedehydrogenase	M31965	soghumvulgare	2912-3013, 3100-3222	gttcctgatttcttgaatgc	54.82	ccctgtggagaaccagtc	56.33	312
3	methyl transferase	AF063403	zea mays	5107-5271, 5361-5553	ttgatgtgcctgtggatatga	59.94	tgcagccctttctgataaacc	60.21	417
	methyl transferase	AF063403	zea mays	5634-5745, 5819-5966	tatcttttacaaaatgggcgattg	61.30	acacggaggaccaccattta	60.23	317

2.12 Discussion:

Phylogeny is about evolution and is used to reconstruct evolutionary events. It is now possible to construct phylogenetic evolution at a molecular level through analysis of molecular sequences, namely proteins & nucleic acids.

To construct phylogenetic tree among grass family, the sequences of conserved enzymes from mitochondria, chloroplast and nucleus are probed using bio-informatics tools. The scheme for such study is the following

- Identify exon regions for the enzyme to be investigated.
- An exon region of the particular enzyme is used to design the primers.
- Confirm the presence the particular sequence of the enzyme (exon) in the species of interest using wet lab techniques.

  - ✓ Isolation of chloroplast,mitochondrial and nuclear DNA
  - ✓ Amplification of DNA by using PCR
  - ✓ Hybridization techniques(southern blotting)
  - ✓ DNA sequencing by chemical and enzymatic methods.
  - ✓ Analysis of sequence based on mitochondrial and chloroplast to determine maternal inheritance.
  - ✓ Analysis based on nucleus to determine paternal inheritance.
  - ✓ Comparison of sequences using multiple alignment tools

Determine the relationship among the species under study

References

Benson,G.(1999)Tandem repeats finder program to analyze DNA sequences Nucleic acids Research 27:573-580.

GerardF.B.(2001).The use of the Monsanto draft rice genome sequence research. Plant physiology 125:1164-1165.

Katti, V.M.;Sami-Subbu,R.;Ranjekar,P.K. and gupta, V.S.(2000). Amonoacid repeat patterns in protein sequences:Their diversity and structural and functional implications. Protein science 9:1203-1209.

Lench,N.J.;Norris,A;Bailey,a.;Booth,A. and Markham,A.F.(1996).Vectoreete PCR isolation of microsatellite repeat sequences using anchored dinucleotide primers.Nucleic acids research 24:2190-2191.

Macas,J.;Meszaros.T. and Nouzova,M.(2002).Plantsat:a specialized database for plant satellite repeats .Bioinformatics 18:28-35.

Mahalakshmi,V. and Ortiz, R.(2001). Plant genomics and agriculture :from model organisms to crops, the role of data mining for gene discovery. Electronic journal of biotechnology.

Dinakar bhattramakki,jianmindong,ashok K.chhabra, and Gary E.Hart. An integrated SSR and RFLP linkage map of Sorghum bicolor(L.) Moench

Pearson,C.E,nad Sinden,R,R.(1998). Trinucleotide repeat DNA structures:Dynamic mutations from dynamic DNA. Current openion in Structural biology 8:321-330

Yuan, q.; Quackenbush,J.;Sultana ,R.;Pertea,m.;Salzberg,S.L and Buell, C.R.(20010. Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species. Plant physiology,125:1166-1174.

References

BASU,D.,K.DEHESH,H.J.SCHNEIDER-

POETSCH,S.E.HARRINGTON,S.R.McCOUCH and P.H.QUAL.2000.rice PHYC gene: structure,expression,map position and evolution. plant mol bio44:27-42.

DJE, Y., M. HEUERTZ, C. LEFEBVRE and X.VENKEMANS.2000.Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatelite markers. Theoretical and applied genetics.100:918-925.

DOEBLEY, J. 1989. Isozymic evidence and the evolution of crop plants. Pp. 165-191 in D. E. soltis and P. S. soltis, eds. Isozymes in plant biology.Dioscorides press, Portland, Oregon.

DGGETT,H. 1976.Pp. 112-117.Evolution of crop plants. Longman, Essex, UK.

GAUT,B.S. and M. T. CLEGG. 1993a.Molecular evolution of the Adh1 locus in the genus Zea.proc Natl Acad Sci U S A 90: 5095-5099.

HARLAN, J. R. and A. B. L. STEMILER. 1976. The races of Sorghum in Africa. Pp. 465-478 in J. R. Harlan, J. M. de wet, and A. B. L. STEMLER, eds.origins of African plant Domestication. Mouton Press, The Hague.

HILTON,H. and B. S. GAUT. 1998.Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene. Genetics 150:863-872.

HUDSON, R. R., M. KREITMAN and M.AGUADE.1987.Atest of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences Mol Evol 16: 111-20.

KOLUKISAOGLU, H.U., S. MARX., C. WIEGMANN, S. HANELT and H. A. SCHNEIDER-POETSCH.1995. Divergence of the pyhtochrome gene family predates angiosperm evolution and suggests that selaginella and Equisetum arose prior to psilotum. J Mol Evol 41: 329-337.

KUMAR, S., K. TAMURA, I. B. JAKOBSEN and M. NEI, 2001 MEGA2: Molecular Evolutionary Genetics Analysis software, Pp., Arizona state University, temp, Arizona, USA.

Nei,m.1987. molecular Evolutionary Genetics.Columbia University Press,New York.