

SSR MARKER DEVELOPMENT IN MEDICAGO
&
PHYLOGENETIC STUDIES IN LEGUMES.

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

MASTER OF TECHNOLOGY
in
BIOTECHNOLOGY

By

ADDAGADA BALAKRISHNA



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

20002

TO
MY
LOVING PARENTS



ICRISAT

International Crops Research Institute for the Semi-Arid Tropics

Patancheru 502 324
Andhra Pradesh
India



CGIAR

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 MEDICAGO & PHYLOGENETIC STUDIES IN LEGUMES**” Submitted by **A.Balakrishna** 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

Visit our worldwide web site at <http://www.icrisat.org>

ICRISAT is part of the global agricultural research network called the Consultative Group on International Agricultural Research (CGIAR)



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 MEDICAGO & PHYLOGENETIC STUDIES IN LEGUMES**" 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 **Mr.Balakrishna.Addagada** under the guidance of **Dr.V.Mahalakshmi**, Senior scientist, International Crops Research Institute For Semi-Arid Tropic.(ICRISAT)

[Dr.M.Lakshmi Narasu]

DECLARATION

I Balakrishna.addagada, a bonafied student of IPGSR, JNTU, Hyderabad here by declare that the dissertation entitled "**SSR Marker Development In Medicago And Phylogenetic Studies In Legumes**" 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.

(Balakrishna addagada)

ACKNOWLEDGEMENTS

This project work is carried out with valuable suggestions and guidance under the supervision of Dr.V. Mahalakshmi.Principalscientist,GT1,ICRISAT,patancheru.

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 ApprenticeAnd 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. hemabindu, Ms.Manjula, Mr.PVNS Prasad, Ms. Rekha, Ms. Leela, Mr. Kumar, Ms. Sasikala, Ms. shanthi, and other colleagues in ICRISAT for their encouragement and cooperation during my course of project work.

I show my gratitude to my beloved parents, brother's and my Roommates, for their constant encouragement and good support throughout this course of work.

Abstract:

Bioinformatics is 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 finding of tandem repeats for medicago truncatula. For this purpose bioinformatics tools like Tandem Repeats Finder, Primer3, Windows software and MS Access were used. The complete database of medicago truncatula plant was created and placed in the INTRANET of ICRISAT, which could be accessed by the scholars and scientists for their requirements. The database of medicago truncatula gives complete information regarding the different tandemrepeats, their complete sequence, the accession number of that sequence which was given by NCBI, left primer, right primer, left temperature, right temperature, and total sequence size.

Chapter 2 gives an introduction about the molecular phylogenetic studies and describes completely about the phylogeny of selected legumes 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 success. 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 by using bio-informatics tools.

CONTENTS

1.	SSR Marker Development For Medicago Data Base.....	1
1.1	Introduction.....	1
1.1.1	DNA repeat – finding tools.....	2
1.1.2	Tandem Repeats Finder.....	2
1.1.2.1	Levels of Tandem Repeat Finder.....	3
1.1.2.2	Advanced Tandem Repeat Finder Program Parameters.....	4
1.1.2.3	Options.....	5
1.1.2.4	Procedure for finding Tandem repeats.....	6
1.1.2.5	Alignment Explanation.....	11
1.2	Primer Design.....	12
1.2.1	Introduction.....	12
1.2.1.1	Primer Design Programs.....	12
1.2.1.2	Primer Design Considerations.....	13
1.2.1.3	Features of Primer Design.....	15
1.2.1.4	Limits of Primer Design.....	15
1.2.2	Primer3.....	16
1.2.2.1	Primer3 input parameters.....	19
1.2.2.2	Procedure for Primer Design by using Primer3.....	28
1.2.2.3	<i>Output of Primer3</i>	28
1.2.2.4	Results.....	31
1.2.2.5	Data base of medicago.....	32
1.2.2.6	Discussion.....	35

2.	Phylogenetic studies in legumes.....	37
2.1	Introduction	37
2.1.2	Phylogenetic Terms.....	38
2.2	Molecular phylogenetics.....	43
2.3	Multiple alignment.....	44
2.3.1	Phylogenetic tree construction methods.....	45
2.3.2	Phylogenetic tree.....	46
2.3.3	Tree building.....	49
2.4	Method Of Phylogenetic Studies Of Legumes.....	51
2.4.1	Reasons For Taking Conserved Common Enzymes In Phylogenetic Studies.....	51
2.4.1.1	Nuclear Enzyme.....	51
2.4.1.2	Mitochondrial Enzyme.....	52
2.4.1.3	Chloroplast Enzyme.....	53
2.4.2	Multiple Alignment Method.....	53
2.4.2.1	Steps involved in multiple alignment method.....	54
2.5	Design Primers For The Sequences Of Medicago.....	80
2.5.1	Selection of first set of primers.....	81
2.5.2	Calculation of product size.....	89
2.6	Selection of second set of primers.....	90
2.7	Results.....	90
2.8	Discussion.....	93

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/cluster/>
- 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>

1. SSR Marker Development For Medicago Database:

1.1 Introduction:

DNA simple sequence repeats (SSRs) are also called as microsatellites. These micro satellites are becoming used as DNA markers in marker assisted breeding. 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)). SSR are becoming the standard DNA markers for plant genome analysis. A wide variety of methods for construction of libraries enriched for micro-satellite sequences have been reported, the most popular among these being the ones based on vectorette PCR using anchored primers (Lench et al. 1996).

The standard procedure for developing SSRs as genetic markers is to isolate and sequence SSR –containing DNA clones from size-fractionated and (or) enriched genomic DNA libraries, and to design, produce, and test PCR primer sets for SSRs contained in the sequenced clones. The most rapidly reassociating DNA is simple sequence DNA, which is composed of short [5-to 10-base] oligonucleotides that are tandemly repeated. The tandem repetition of a short sequence often creates a fraction with distinctive physical properties that can be used to isolate it. The term satellite DNA is essentially synonymous with simple sequence DNA. 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.

1.1.1 DNA repeat – finding tools:

DST- locate human repeats:

Two search modes are available, a search for human repeats using the file BR3X and a self-homology search that will find repeats more than 2kb apart.

Tandem Repeats Finder:

A Tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides.

Large Dot Plots:

This page accesses a very fast dot plot algorithm designed for large DNA sequences.

REPuter- Fast Computation of Maximal Repeats in complete genome:

REPuter computes all maximal duplications and reverse, complemented and reverse complemented repeats in a DNA input sequence.

Repeat Masker- mask out repeat sequences:

Repeat Masker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches as well as a table annotating the masked regions.

1.1.2 Tandem Repeats Finder

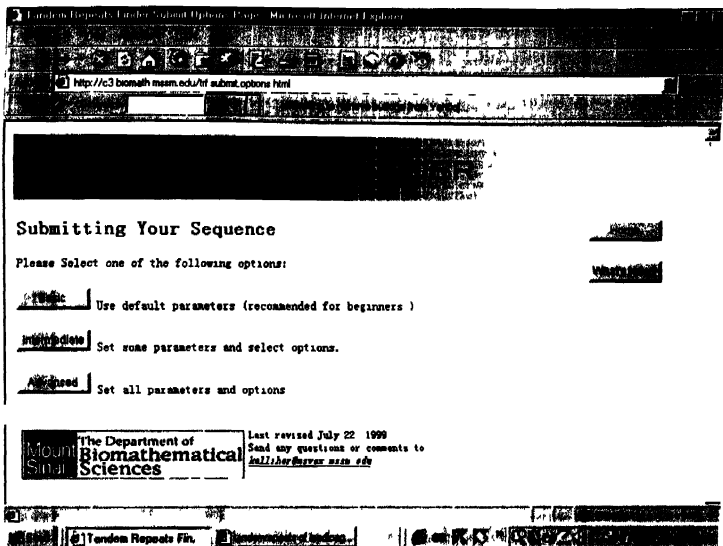
Tandem Repeats Finder is a program to locate and display Tandem Repeats in DNA sequences. In order to use this program, we have to submit the

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 entry 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.

1.1.2.1 Levels of Tandem Repeat Finder:

There are 3 levels of tandem repeat finders.

- 1. Basic:** It uses default parameters (recommended for beginners.)
- 2. Intermediate:** It provides the parameter Maximum period size, and options Flanking sequence and Masked Sequence File.
- 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.



1.1.2.2 *Advanced Tandem Repeat Finder Program Parameters:*

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

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.

2. Minimum Alignment Score: The alignment score must meet or exceed this value for the repeat to be reported.

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.

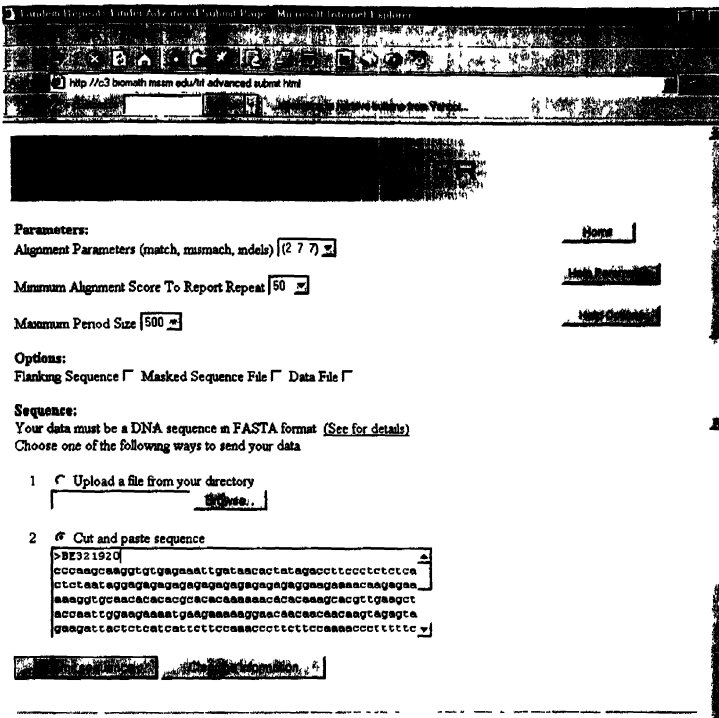
4. Detection parameters: Matching probability P_m and indeal probability P_i , $P_m = .80$ and $P_i = .10$ by default and it cannot be modified in this version of the program.

1.1.2.3 Options:

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.

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.

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

1.1.2.4 Procedure for finding Tandem repeats:

Web tools like <http://c3.biomath.mssm.edu/trf/advanced.submit.html>

can be used to find out the tandem repeats of a particular sequence (source sequence). This web tool is provided by The Department of Biomathematical sciences, Mount Sinai School of Medicine.

Download the source sequence from the www.ncbi.nlm.nih.gov/entrez in FASTA format.

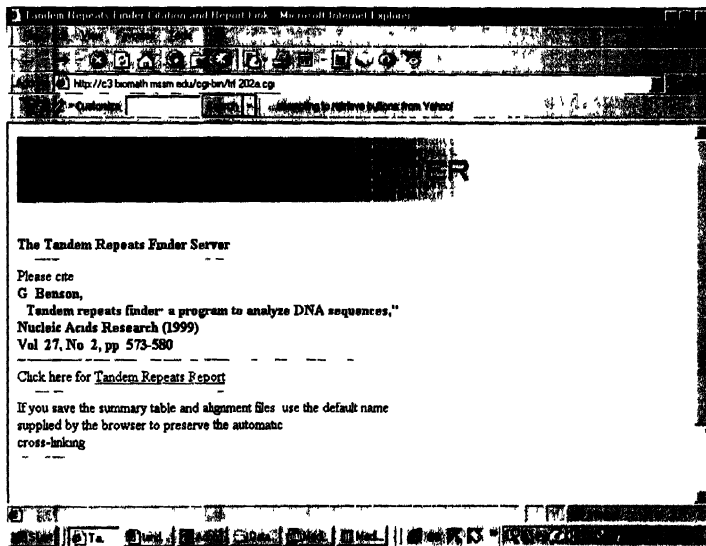
Open the page http://c3.biomath.mssm.edu/tf/advanced_submit.html

Enter the source sequence in **cut and paste** sequence blank.

Adjust the parameters according to our practical requirement

Click the **submit sequence** button.

- By clicking on the submit sequence the tandem repeats finder server will display on screen



- By clicking on the **tandem repeats report** displays the program analysis.
- It shows the results in a table format.
- Results of search:

The program's analysis is sent back to the user's web browser as two files, a summary table file and an alignment file.

http://c3.biomath.mssm.edu/10c/P/W/0/Wo 2 7 7 80 10 50 500 1.htm

Tandem Repeats Finder Program written by

Gary Benson
 Department of Biomathematical Sciences
 Mount Sinai School of Medicine
 Version 2.02

Please cite
 G. Benson,
 "Tandem repeats finder: a program to analyze DNA sequences"
 Nucleic Acid Research(1999)
 Vol 27, No 2, pp 573-580

Sequence BE321920
 Parameters 2 7 7 80 10 50 500
 Length 555

Tables 1

This is table 1 of 1

Click on indices to view alignment

Table Explanation

Indices	Period Size	Copy Number	Consensus Size	Percent Matches	Percent Indels	Score	A	C	G	T	Entropy (0-2)
<u>60-84</u>	2	125	2	100	0	50	48	0	52	0	1.00
<u>218-246</u>	14	21	14	93	6	51	24	41	0	34	1.55

The End!

• **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. The links to the other table appears at the top of each table.
- Clicking on the location indices of the table opens a second web browser which shows an alignment of the copies against a consensus pattern.

Tandem Repeats Finder Program written by:

Gary Benson
Department of Biomathematical Sciences
Mount Sinai School of Medicine

Version 2.02

Sequence: BE321920

Parameters: 2 7 7 80 10 50 500

Fmatch=0.80, Findel=0.10
tuple sizes 0,4,5,7
tuple distances 0, 29, 159, MAXDISTANCE

Length: 555
ACGTCcount: A:0.36, C:0.23, G:0.19, T:0.23

Found at 1.65 original size*2 final size 2

Alignment explanation

Indices: 60 -84 Score: 50
Period size: 2 Copynumber: 12.5 Consensus size: 2

50 ACTCTAATAG

60 GA GA GA GA GA GA GA GA GA GA GA G
1 GA GA GA GA GA GA GA GA GA GA GA G

85 GAAGAAAACA

Statistics

Matches: 23, Mismatches: 0, Indels: 0
1.00 0.00 0.00

Matches are distributed among these distances:
2 23 1.00

ACGTCcount: A:0.48, C:0.00, G:0.52, T:0.00

Consensus pattern (2 bp):
GA

Found at 1:236 original size:13 final size:14

Alignment explanation

Indices: 218--246 Score: 51
Period size: 14 Copynumber: 2.1 Consensus size: 14

208 ACTCTCATCA

218 TTCTTCC-AAACCC
1 TTCTTCCAAAACCC

231 TTCTTCCAAAACCC
1 TTCTTCCAAAACCC

245 TT
1 TT

247 TTCTTCACT

Statistics

Matches: 15, Mismatches: 0, Indels: 1
0.94 0.00 0.06

Matches are distributed among these distances:
13 7 0.47
14 0 0.53

ACGTCcount: A:0.24, C:0.41, G:0.00, T:0.34

1.1.2.5 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 are 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. The Symbol (*) indicates a mismatch.
5. The 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.

1.2 Primer Design:

1.2.1 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.

1.2.1.1 *Primer Design Programs:*

These are all primer design tools, which are generally used for the primer prediction and analysis programs.

1. The PCR Jump Station: The ultimate Web page for information and links on all aspects of the Polymerase Chain Reaction (PCR).

2. Gene Fisher: Gene fisher processes aligned or unaligned sequences.

3. Gene Walker: It allows working with two primer sequences

4.Cyber Gene: Cyber Gene is a company that provides commercial oligonucleotide synthesis, DNA sequencing, genotyping and bioinformatics services.

5.Web Primer: An application that designs primers for PCR or sequencing purpose.

6.Primer Design: A free primer design utility, from the EMBL.

7.Primer3: Primer3 picks primers from PCR reactions, according to the conditions specified by us. Primer considers things like melting temperature, concentrations of various solutions in PCR reactions, primer bending and folding, and many other conditions when attempting to choose the optimal pair of primers for reaction.

8.Poland – melting profiles of double stranded DNA: The Poland server will calculate the thermal denaturation profile of double stranded RNA or DNA based on sequence input and parameter settings in this form.

9.Net Primer: Net primer combines the latest primer design algorithms with a Web-based interface allowing the user to analyze primers over the Internet.

10.Gene Primer: It gives computational support of gene experiments. This software implements an algorithm for experimental gene identification by multiple PCR amplification.

1.2.1.2 *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, that 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.

1.2.1.3 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**.

1.2.1.4 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 explored up to 6000 pairs.

1.2.2 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.

Sequence Quality

Min Sequence Quality Min End Site Min Start Site Min Site Min Site Min Site Min Site Min Site

Objective Function Penalty Weights for Primers

Pr L1 L1 Pr

we L1 L1 Pr

Pr L1 L1 Pr

Self Complementarity

Self Complementarity

HTs

Mispriming

Secondary Structure

End Site Quality

Start Site Quality

End Site Quality

Objective Function Penalty Weights for Primer Pairs

Pr L1 L1 Pr

Product In L1 L1 Pr

HT Differ

Any Complementarity

Self Complementarity

End Site Quality

Start Site Quality

Hyb Oligo Feasibility Weight

Primer Primer

Hyb Oligo (Internal Oligo) Per Sequence Inputs

Hyb Oligo

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size Min Max

Hyb Oligo L1 Min Max

Hyb Oligo L2 Min Max

Hyb Oligo Self Complementarity Hyb Oligo Max Self Complementarity

Max HTs Hyb Oligo Max HTs

Hyb Oligo Hybridization Hyb Oligo Hybridization

Hyb Oligo Min Sequence Quality

Hyb Oligo Self Complementarity Hyb Oligo Self Complementarity

Primer Primer

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo L1 L1 L1

Hyb Oligo L2 L1 L1

Hyb Oligo L1 L1 L1

Hyb Oligo Self Complementarity

Hyb Oligo HTs

Hyb Oligo Hybridization

Hyb Oligo Sequence Quality

Primer Primer

Copyright Notice and Disclaimer

1.2.2.1 Primer3 input parameters:

Source Sequence:

The sequence from which to select primers.

Sequence Id:

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

Targets:

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.

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.

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,

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.

Max 3' Stability:

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

Max Mispriming:

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

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

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.

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.

Maximum T_m Difference:

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

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.

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

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.

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

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.

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.

Max Poly-X:

The maximum allowable length of a mononucleotide repeats for example AAAAAA.

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.

Start Codon Position:

This parameter should be considered EXPERIMENTAL at this point. Some erroneous input 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.

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.

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.)

Salt Concentration:

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

Annealing Oligo Concentration:

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

Max Ns Accepted:

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

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).

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.

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.

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.

Sequence Quality**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.

Min Sequence Quality:

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

Min 3' Sequence Quality:

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

Sequence Quality Range Min:

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

Sequence Quality Range Max:

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

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

Hyb Oligos (Internal Oligos):

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

1.2.2.2 Procedure for Primer Design by using Primer3:

- Select the query sequence and Id, which has the Tandem repeats in it
- 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.

1.2.2.3 Output of Primer3

In primer 3 input after adjusting the parameters like temperature, (the default temperatures of minimum, optimum and maximum are 57,60,63 respectively. But adjust them to 59,60,61 respectively for more reliable results, which are more acceptable for practical purpose.) By clicking on the PICK PRIMERS gives the results as PRIMER3 OUTPUT. It will display like as shown below.

The top of the output displays the sequence id. The next part of the output displays the best left and right primers, and their characteristics (starting position, length, melting temperatures, and so forth). Then the output displays information specific to the input sequence and the selected pair.

The next information is a quasi-graphical representation of the location of the left (>>>>>>...) and right (<<<<<<....) primers in the source sequence. The position of the target is marked by asterisks (*****).

Primer3 Output

--
 --
 No mismatching library specified
 Using I-based sequence junctions
 JLIGOS

	LEN	INT	IN	MC	MIN	S	JL
LEFT PRIMER	228	20	59 43	45 00	3 00	3 00	ccctctttccaaaacccttt
RIGHT PRIMER	434	20	59 97	55 00	6 00	0 00	tagggtagcaggtggagaa

 SEQUENCE SIZE 555
 IN-LUDEF PF01 R 17E 555

PRODUCT SIZE: 107 PAIR ANY COMPL: 5 00 PAIR 3 JNPL 0 00

```

1 cccnagcagggtgtgagaactgataaacctatagacctccctctctccctctaatgg
61 egeggagagaauegagagagagagaggaacccagagaaagagagaaagagagaaagaa
121 cccanaaaacacacacacag Agcttgaagrrarraatggagaaatgaaagaaaagjaa
181 ccaacacacagagragagtagaagarractctcatcatctctccaaacttctctccaa
241 ac tctctctctt cctctctcttcttcccaaaaactgggaatgtgaaluyagagaga
301 agatrraatgggaacarratcatagctfractcaaacctatragaggtcagtgatccat
361 gctctacccctgatgggtctctgaaactctctccctctctgagtraatgggtcctatccc
421 aactctagrcctagaalantcaatctctctcttcaagttcaactf Anjct cgtutui
481 tccactacaaaucctgatgagatgaaagttccagcccttctctgatgcaaacctaatc
541 aggtctatcagatcat
  
```

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

ADDITIONAL JLIGOS

	LEN	INT	IN	MC	MIN	S	JL
1 LEFT PRIMER	220	20	59 43	45 00	3 00	3 00	ccctctttccaaaacccttt
RIGHT PRIMER	437	20	59 97	55 00	6 30	0 00	ttcttaggagragcaggtgga
PRODUCT SIZE	110	PAIR ANY	COMPL: 4 00	PAIR 1	COMPL	1 1	
2 LEFT PRIMER	220	20	59 43	45 00	3 00	3 00	ccctctttccaaaacccttt
RIGHT PRIMER	433	20	60 43	55 00	6 00	2 00	agggtagcaggtggagaa
PRODUCT SIZE	706	PAIR ANY	COMPL: 5 00	PAIR 3	COMPL	1 00	
3 LEFT PRIMER	220	20	59 43	45 00	3 00	3 00	ccctctttccaaaacccttt
RIGHT PRIMER	431	20	60 36	55 00	6 00	1 00	gggtagcaggtggagaa
PRODUCT SIZE	204	PAIR ANY	COMPL: 5 00	PAIR 3	COMPL	1 00	
4 LEFT PRIMER	222	20	59 52	40 30	2 00	2 00	cccaaacctcttccaaaa
RIGHT PRIMER	434	20	59 97	55 00	6 00	0 00	tagggtagcaggtggagaa
PRODUCT SIZE	213	PAIR ANY	COMPL: 5 00	PAIR 3	COMPL	2 00	

Statistics

con	too	in	in	no	tm	tm	high	high	poly	high			
sid	many	tax	excl	basl	GC	too	too	any	3	end			
Left	ered	Na	gat	reg	G% clamp	low	high	compl	compl	X stab	ok		
Left	4140	0	0	0	1	0	2544	1117	0	12	45	17	404
Right	4059	0	0	0	1	0	2108	1470	0	2	44	29	405

Pair Stats
 considered 220 unacceptable product size 194 high end compl 13 ok 113
 primer3 release 0 9

1.2.2.4 Results:

For our study, we selected medicago, which is considered the nodal crop for all legumes. All available (approximately 1,56,000) nucleotide sequences were analyzed for the presence of tandemrepeats upto 50bp (maximum) length of repeat motif and no penalty gaps or indels were allowed. All the genomic sequences of medicago from public domain database were searched and analysed of Di, Tri, and tetra nucleotide repeats. Of the total about 1,56,000 sequences, which were searched, 7325 sequences were found to contain repeat motif and may yield SSRs, which would yield product sizes of around 200 bp. Of these mostly abundantly found repeats were the Tri-nucleotide group.

Out of 7325 tandemrepeats the Di-nucleotides are 1290,tri-nucleotides are 5210,and tetra –nucleotides are 925.

Summary by nucleotide unit length

Unit length	SSR Count
2	1290
3	5210
4	925

1.2.2.5 Data base of medicago:

As described in the above procedure the nucleotide sequences are allowed to find tandem repeats by submitting them to Tandem repeat finder. Then the primers are found to these repeats by using Primer3. The results of Primer3 output are collected and created as database. This database contains eight columns, id, accession number, sequence, left primer sequence, right primer sequence, left primer temperature, right primer temperature and total sequence.

This database link to gene annotation database at TIGR (www.tigr.org)

To facilitate further exploration of this resource, a dynamic database with options to search and link to other resources is available at (<http://www.icrisat.org/text/research/grep/homepage/genomics/mcdssrs1.asp>) and on CDs from V.Mahalakshmi@cgjar.org.

Group	SSR Count	SSR Units in Group
AC	89	AC CA GT TG
AG	828	AG CT GA TC
AT	371	AT TA
CG	2	CG GC
AAC	721	AAC ACA CAA GTT TGT TTG
AAG	2067	AAG AGA CTT GAA TCT TTC
AAT	364	AAT ATA ATT TAA TAT TTA
ACC	462	ACC CAC CCA GGT GTG TGG
ACG	508	ACG AGC CAG CGA CGT CTG GAC GCA GCT GTC TCG TGC
ACT	809	ACT AGT ATC ATG CAT CTA GAT GTA TAC TAG TCA TGA
AGG	196	AGG CCT CTC GAG GGA TCC
CCG	83	CCG CGC CGG GCC GCG GGC
AAAC	73	AAAC AACA ACAA CAAA GTTT TGTT TTGT TTTG
AAAG	141	AAAG AAGA AGAA CTTT GAAA TCIT TTCT TTTC
AAAT	172	AAAT AATA ATAA ATTT TAAA TATT TTAT TTTA
AACC	1	AACC ACCA CAAC CCAA GGTT GTTG TGGT TTGG
AACG	5	AACG AAGC ACGA AGCA CAAG CGAA CGTT CTTG GAAC GCAA GCTT GTTC TCGT TGCT TTCG TTGC
AACT	44	AACT AATC ACTA AGTT ATCA ATTG CAAT CTAA GATT GTTA TAAC TAGT TCAA TGAT TTAG TTGA
AAGG	39	AAGG AGGA CCTT CTTC GAAG GGAA TCCT TTCC
AAGT	80	AAGT AATG ACTT AGTA ATGA ATTC CATT CTTA GAAT GTAA TAAG TACT TCAT TGAA TTAC TTCA
AATT	123	AATT ATTA TAAT TTAA
ACAG	44	ACAG AGAC CAGA CTGT GACA GTCT TCTG TGTC
ACAT	56	ACAT ATAC ATGT CATA GTAT TACA TATG TGTA
ACCC	2	ACCC CACC CCAC CCCA GGGT GGTG GTGG TGGG
ACCT	3	ACCT AAGT ATCC ATGG CATC CCAT CCTA CTAC GATG GGAT GGTA GTAG TACC TAGG TCCA TGGA
ACGC	1	ACGC CACG CGCA CGTG GCAC GCGT GTGC TGCC

ACGG	2	ACGG AGGC CAGG CCGT CCTG CCGA CGTC CTGC GACG GCAG GCCT GGAC GGCA GTCC TCCG TGCC
ACGT	24	ACGT ATGC CATG CGTA GCAT GTAC TACG TGCA
ACTC	37	ACTC AGTG CACT CTCA GAGT GTGA TCAC TGAG
AGAT	47	AGAT ATAG ATCT CTAT GATA TAGA TATC TCTA
AGCG	1	AGCG CGAG CGCT CTCG GAGC GCGA GCTC TCGC
AGCT	9	AGCT ATCG CGAT CTAG GATC GCTA TAGC TCGA
AGGG	20	AGGG CCCT CCTC CTCC GAGG GGAG GGGG TCCC
CCCG	1	CCCG CCGC CGCC CGGG GCCC GCGG GGCG GGGC
CGGCG	2	CGGCG GGCCC
ACCGG	9	ACCGG ACGCC CAGGC CCACG CGCCA
AAAAAT	79	AAAAAT AAATA AATAA AATTA ATAAA ATATT ATTAT ATTTT TAAAA TATAA TATTT TTAATA TTATA TTATT TTAA TTTAT TTTTA
AACCC	54	AACCC AAGGG ACACC AGGAG AGGTG ATCCC CAACC CAGAT CTCTC GAAGC GGAGA GGGAT GGTTC GTCCT GTGGA GTGTC TCACC TCCAC TCCGA TCGAC TCGCA TCTCC TCGCA TTCCG TTGGC
AAAGG	237	AAAGG AACAC AACAG AACCA AAGAG AAGCA AATCC AATCG ACAAC ACACA ACAGA ACTCA ACTTC AGACA AGAGA AGCAT AGTAC ATCCA ATCTG ATGAG ATGGA ATGTC CAAAC CAACA CAACT CACAA CACAT CACTA CAGAT CATCT CCAAA CCAAT CTCAA CTCAT CTCTT CTTC TCCAA TCCTT TCGAT TCTCA TCTCT TCTGT TCTTC TGAAG TGACA TGCA TGTC TTCTT TTCTC TTCTG TTGGA TTGTG TTTCC TTTGG
AAAAAC	236	AAAAAC AAAAG AAACA AAAGA AAATC AACAA AACAT AAGAA AAGAT AATAG AATCA AATGA AATTIC AATTG AAAAA ACAAT ACATA AGAAA ATCAA ATCTA ATCTT ATGAA ATGAT ATGTA ATTAG ATTCA ATTCT ATTGA ATTTC ATTTG CAAAA CAATT CATAA CATAT CATTI CTA CTAAT CTATT CTAT CTITT GAAAA GAATA GTTTT TAGTT TATCA TATTC TCAAA TCAAT TCTAT TCTTA TGTTT TTCAA TTCAT TTCTA TTCTT TTGAA TTGAT TTGTA TTGTT TTCA TTTCT TTTGT TTTTC TTTTG

1.2.2.6 Discussion:

Data mining encompasses the use of pattern recognition technologies and statistical techniques to examine large amounts of data. De Novo generation of microsatellite markers through laboratory based screening of SSR enriched genomic libraries is highly time consuming and expensive. An alternative is to screen the public databases of related model species where abundant sequence data is already available. Beyond the cost savings, this approach also offers the possibility of identity 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.

Since the advent of recombinant DNA technology in population genetics in the mid-1980s, the repertoire of genetic markers available for population studies and for crop improvement has increased enormously. Plant breeding has changed with the introduction of these molecular techniques. Molecular markers allow for the extension of traditional breeding methods with one important difference-to transfer greater variety of genetic information in a more precise and controlled manner. Since the advent of molecular markers 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 action being co-dominance with the exception of null alleles at some loci, show large variation within population s and are generally found in non-coding regions, which may contribute to the genome stability.

Our project work suggests that we can take advantage of DNA marker technology, a core set of at least 1000 sequence tagged sites (STSs) that are universal among all legume species could be developed. Though these approaches are proving to be useful as SSR markers for the same species or closely related species within the same genus, their utility in other related species is yet to be tested.

Phylogenetic studies in legumes

2.1 Introduction:

Phylogenetics, the science of phylogeny, is one part of the larger field of **systematics**, which also includes **taxonomy**. Taxonomy is the science of naming and classifying the diversity of organisms.

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. It is a hypothesis based on 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 ancestral relationships of species or groups of species and on how to incorporate these relationships into classification systems.

Phylogeny is used to classify organisms on the basis of their inferred evolutionary relationships (the phylogenetic approach to classification). 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 plants and herbivores.

Problems: If the evolutionary clock is not constant, the procedure generates results, which can be misleading.

1. Within practical computational limits, this often leads in the generation of tens or more "equally most parsimonious trees" which make it difficult to justify the choice of a particular tree.
2. Long computation time to construct a tree.

2.1.2 Phylogenetic Terms

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.

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.

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.
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.
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.
Lineage	Any continuous line of descent; any series of organisms connected by reproduction by parent of offspring.
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.
Stasis	A period of little or no discernible change in a lineage.

Extinction	when all the members of a clade or taxon die, the group is said to be extinct.
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.
Sister group	The two clades resulting from the splitting of a single lineage.
Cladogenesis	The development of a new clade; the splitting of a single lineage into two distinct lineages; speciation.
Stem group	All the taxa in a clade preceding a major cladogenesis event. They are often difficult to recognize because they may not possess synapomorphies found in the crown group.

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.
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.
Homology	Two structures are considered homologous when they are inherited from a common ancestor who possessed the structure. This may be difficult to determine when the structure has been modified through descent.
Relatedness	Two clades are more closely related when they share a more recent common ancestor between them than they do with any other clade.
Diversity	Term used to describe numbers of taxa, or variation in morphology.

Convergence

Similarities that have arisen independently in two or more organisms that are not closely related. Contrast with homology.

2.2 Molecular phylogenetics:

Molecular phylogenetics attempts to determine the rates and patterns of change occurring in DNA and proteins and to reconstruct the evolutionary history of genes and organisms. Two general approaches may be taken to obtain this information. In the first approach, scientists use DNA to study the evolution of an organism. In the second approach, different organisms are used to study the evolution of DNA. Whatever the approach, the general goal is to **infer process from pattern**: the processes of organismal evolution deduced from patterns of DNA variation and processes of molecular evolution inferred from the patterns of variations in the DNA itself.

Molecular Phylogenetic Analysis: Fundamental Elements

Nucleic acid and protein sequences can also be used to generate trees. DNA, RNA and protein sequences can be considered as phenotypic traits. The sequences depict the relationship of genes and usually of the organism in which the genes are found.

As we just discussed, macromolecules, especially gene and protein sequences have surpassed morphological and other organismal characters as the most popular forms of data for phylogenetic analyses.

First, it is important to point out that a single, all-purpose recipe does not exist for phylogenetic analysis of this type of data. Although numerous algorithms, procedures, and computer programs have been developed, their

reliability and practicality are, in all cases, dependent upon the size and structure of the data set under analysis. Phylogenetic tree-building models presume particular evolutionary models. For any given set of data, these models may be violated because of various occurrences, such as the transfer of genetic material between organisms. Therefore, when interpreting a given analysis, a person should always consider the models used and entertain possible explanations for the results obtained. For example, models used in molecular phylogenetic analysis methods make "default" assumptions, including:

1. The sequence is correct and originates from the specified source;
2. The sequences are homologous--are all descended in some way from a shared ancestral sequence;
3. Each position in a sequence alignment is homologous with every other in that alignment;
4. Each of the multiple sequences included in a common analysis has a common phylogenetic history with the other sequences;
5. The sampling of taxa is adequate to resolve the problem under study;
6. Sequence variation among the samples is representative of the broader group; and
7. The sequence variability in the sample contains phylogenetic signal adequate to resolve the problem under study.

2.3 Multiple alignment:

The most practical and widely used method is multiple alignment method. This method is the hierarchical extensions of pairwise alignment methods. Multiple alignments are built by successive application of pairwise methods:

- It Compares all the sequences pairwise; (for N sequences there are $N.(N-2)/2$ pairs or scores)
- It performs cluster analysis on pairwise scores to generate a hierarchy for alignments;
- It builds the multiple alignment by aligning the most similar pair of sequences first, 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.

- Phylogenetic tree will strongly depends on the alignments obtained.
- In simple cases, the quality of the alignments is good, in more difficult cases, the alignments give good starting points for further automatic or manual refinements ;
- The multiple alignment is dependent of the score calculation model (gap, transversions, weights...)

2.3.1 Phylogenetic tree construction methods:

Its topology (form) and its length (sum of its branch lengths) characterize a phylogenetic tree.

Each node of a tree is an estimation of the ancestor of the elements included in this node.

The phylogenetic methods for constructing phylogenies from sequence data:

1.Methods directly based on sequences:

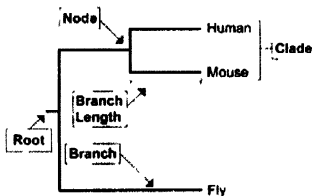
- Parsimony ;
- Maximum likelihood ;

2.Methods indirectly based on sequences:

- Distance matrices (UPGMA, NJ,) ;

2.3.2 PHYLOGENETIC TREE: In phylogenetic studies, the most convenient way of visually presenting evolutionary relationships among a group of organisms is through illustrations called **phylogenetic trees**.

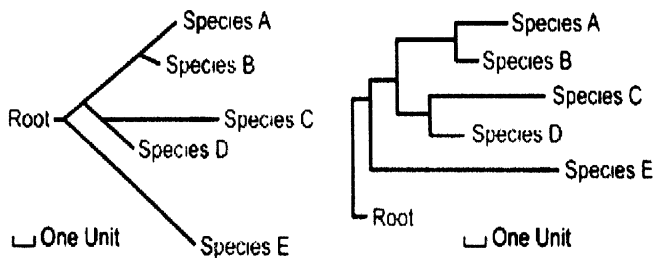
A phylogenetic tree is a graph composed of *nodes* and *branches*, in which only one branch connects any two adjacent nodes. The nodes represent the *taxonomic units* and the branches define the *relationships* among the units in terms of descent and ancestry. The branching pattern of a tree is called the *topology*. The branch length usually represents the number of changes that have occurred in that branch. The taxonomic units represented by the nodes can be species, populations, individuals or genes.



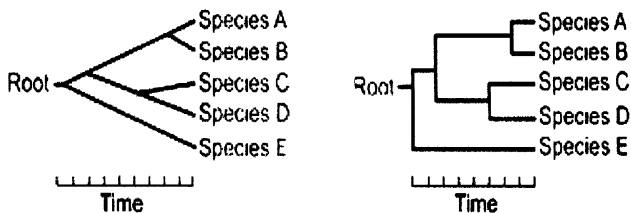
- **Node:** represents a taxonomic unit. This can be either an existing species or an ancestor.
- **Branch:** Defines the relationship between the taxa in terms of descent and ancestry.
- **Branch length:** Represents the number of changes that have occurred in the branch.
- **Root:** The common ancestor of all taxa.
- **Clade:** a group of two or more taxa or DNA sequences that includes both their common ancestor and all their descendents.

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.

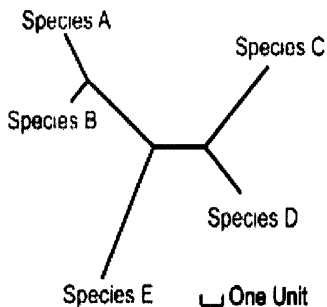
Scaled Branches



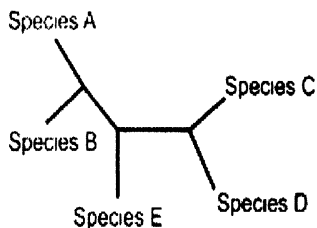
Unscaled Branches



Unrooted Tree with Scaled Branches



Unrooted Tree with Unscaled Branches



2.3.3 TREE BUILDING:

The other type of phylogenetic analysis we'll discuss is *tree building*. This form of analysis is more work than signature analysis, but is quantitative and more reliable. There are several methods for building trees, including distance matrix methods and parsimony methods. We'll discuss the 'least-squares distance matrix' method.

Tree building starts with a sequence alignment. Here is an example alignment of 5 sequences with 25 positions in the alignment:

Seq. A AGAUUCGUCUGUAGGUUCCACCAA

Seq. B ACAUUCGUGUAUAGGUUCCACUAA

Seq. C ACAUUCGUGUAGAGGUUCCACUAA

Seq. D AAGUUCGCUUGGAGGUUCCACGAA

Seq. E AUCGUGAGAUCAGGUAUCCACAAU

The first step toward building a tree is to generate a *similarity matrix*: Just count the fraction of identical bases in every pair of sequences in the alignment.

Seq. A AGAUUCGUCUGUAGGUUCCACCAA

|X|||||X|X|||||||X|| 21/25 = 0.84

Seq. B ACAUUCGUGUAUAGGUUCCACUAA

	A	B	C	D	E
A	----	----	----	----	----
B	0.84	----	----	----	----
C	0.80	0.96	----	----	----
D	0.76	0.72	0.76	----	----
E	0.52	0.52	0.52	0.52	----

In this example, sequences A and B are 0.84 (= 84%) similar, A and C are 0.80 similar, B and C are 0.96 similar, etc, etc.

With all of the similarities converted to evolutionary distances (whether or not they are corrected, or how they are corrected), you have a *distance matrix*:

Evolutionary distance					
A	B	C	D	E	
A	-----	-----	-----	-----	-----
B	0.18	-----	-----	-----	-----
C	0.23	0.04	-----	-----	-----
D	0.29	0.35	0.29	-----	-----
E	0.77	0.77	0.77	0.77	-----

These distances can then be used to construct a tree that best fits these evolutionary distances. This done by starting with two of the sequences separated by a line equal in length to the evolutionary distance between the sequences.

Then the next sequence is added to the tree such that the distances between A, B and C are approximately equal to the evolutionary distances.

Notice that the fit isn't perfect. If we could determine the evolutionary distances exactly, they would fit the tree exactly, but since we have to estimate these distances, the numbers are fit to the tree as closely as possible using a least-squares best fit.

The next step is to add the next sequence, again re-adjusting the tree to fit the distances as well as possible.

And at last we can add the final sequence and readjust the branch lengths one last time using least squares.

Notice that the distance between any two sequences is (approximately) equal to the sum of the length of the line segments joining those two sequences - in other words, the tree is additive.

This type of tree is called a **dendrogram**. The nodes connecting different sets of branches represent common ancestors of those branches. This tree is unrooted - the single common ancestor of *all* of the sequences cannot be determined in this tree. Some people prefer dendrogram because evolutionary distance is easily visualized. In this example, sequence B and C are the most closely related. Each of these are somewhat less similar to A (a little closer in the case of seq. B; that's why the branch to B is shorter than to C). A, B, and C are less similar to D, and E is only distantly related to the rest.

2.4 Method Of Phylogenetic Studies Of Legumes

Take the legumes i.e., *Pisum sativum*, *Medicago*, *Glycine max*, *Arabidopsis*, *Lotus* and co – related the phylogenetic relationship between these Legumes by taking a nuclear enzyme, a mitochondrial enzyme and a chloroplast enzyme.

2.4.1 Reasons For Taking Conserved Common Enzymes In Phylogenetic Studies.

2.4.1.1 Nuclear Enzyme

- 1) Nuclear enzymes have BI-parental inheritance.
- 2) Nuclear enzymes are in abundance in a Genome.
- 3) Amplification and sequencing 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 Intron that necessitates reverse transcriptase PCR.
- 7) The nuclear enzyme which we are taken are Chitinase due to its role in metabolism a key functional pathway.

2.4.1.2 Mitochondrial Enzyme

- 1) Mitochondrial enzymes are maternally inherited.
- 2) These enzymes are used to construct a Phylogenetic Tree 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 Aspartate amino transferase a key enzyme in the respiratory pathway.

2.4.1.3 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 reconstruction's using Parsimony analysis may be produced.
- 5) Restriction-site variation is used to estimate total sequence divergence between taxa. Such distance measures may be used to reconstruct phylogenies.
- 6) Chloroplast 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 Glutamine synthase the key enzyme of the photosynthetic pathway.

2.4.2 Multiple Alignment Method

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

2.4.2.1 Steps involved in multiple alignment method

1. Select the most conserved enzymes:

First we selected the most conserved mitochondrial, chloroplast and nuclear enzymes of legumes. In legumes the most conserved functional enzymes are

Mitochondrial enzymes:

1. Aspartate amino transferase	2.6.1.1
2. NADH dehydrogenase	1 6 5 3
3. ATP synthase	3 6 3 14
4. Succinate dehydrogenase	1 3 99 1
5. Malate dehydrogenase	1 1 1 82
6. Citrate synthase	4 1 3 28

Chloroplast enzymes:

1. Glutamine synthase	1.4.1.13
2. Fructose 1,6 bis phosphatase	3 1 3 11
3. Phospho enol pyruvate carboxylase	4 1 1 31
4. Glyceraldehyde 3-phosphate dehydrogenase	1 2 1 9

Nuclear enzyme:

1. Chitinase	3.2.1.14
2. Methyl transferase	2 1 1 37
3. Alcohol dehydrogenase	1 1 1 1
4. Cysteine synthase	4 2 99 8

2. From the all above-mentioned enzymes the following enzymes which were selected for the practical purpose. i.e. sequence available from GenBank

- Aspartate amino transferase (mitochondrial enzyme)
- Glutamine synthase (chloroplast enzyme)
- Chitinase (nuclear enzyme)

3. Search The Sequences From NCBI:

Search the sequences of the Aspartate amino transferase (mitochondrial enzyme), Glutamine synthase (chloroplast enzyme), and Chitinase (nuclear enzyme)

From NCBI.

Then accession numbers and their sequence of the above enzymes of interest of legumes were downloaded on to a local database .

Mitochondrial Enzyme (Aspartate amino transferase)

Legume	Accession number
Medicago	L25335
Glycine max	L40579
Arabidopsis	X91865
Lotus	AF029898

4. Paste accession numbers of these legumes for a Mitochondrial Enzyme (Aspartate amino transferase) in the NCBI. We will get the sequences for that accession number of the Legumes for ex:- (Medicago, Arabidopsis, Glycine, and Lotus).



NCBI

National Center for Biotechnology Information

National Library of Medicine

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search [input type="text" value=""] [button type="submit" value="Go"]

SITE MAP

Guide to NCBI resources

About NCBI

The science behind our resources. An introduction for researchers, educators and the public.

GenBank

Sequence submission support and software

Molecular databases

Sequences, structures and taxonomy

Literature databases

PubMed, OMIM, Books and PubMed Central

Genomic biology

The human genome, whole genomes and related resources

Tools

For data mining

Research at NCBI

People, projects and programs

Software and computing tools

RAD and databases

Education

Teaching resources and on-line tutorials

FTP site

Download data and software

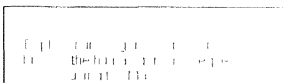
Contact information

How to reach us

What does NCBI do?

- [What does NCBI do?](#)
- [GenBank](#)
- [Molecular databases](#)
- [Literature databases](#)
- [Genomic biology](#)
- [Tools](#)
- [Research at NCBI](#)
- [Software and computing tools](#)
- [Education](#)
- [FTP site](#)
- [Contact information](#)

Hot Spots



RefSeqs for viral genomes¹



► [RefSeqs for viral genomes¹](#)

NCBI in the News

- [NCBI in the News](#)
- [GenBank](#)
- [Molecular databases](#)
- [Literature databases](#)
- [Genomic biology](#)
- [Tools](#)
- [Research at NCBI](#)
- [Software and computing tools](#)
- [Education](#)
- [FTP site](#)
- [Contact information](#)

Entrez-Nucleotide - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide&term=Medicago

Search: attempting to retrieve buttons from layout

NCBI Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books

Search: Nucleotide for Medicago salvia arparata aramo transcribe rRNA 5S gene complete cDNA
 [18729696] [2507914] [FAAATLAA111111]

Display: Summary Save Text Clip/Save

Show: 10 Items 1-5 of 5 One page

1: Medicago salvia arparata aramo transcribe rRNA 5S gene complete cDNA
 [18729696] [2507914] [FAAATLAA111111]

2: Medicago salvia arparata aramo transcribe rRNA 5S gene complete cDNA
 [18729696] [2507914] [FAAATLAA111111]

3: Glycine max clone p2727a nuclear encoded nuclear ribosomal large subunit rRNA, complete cDNA
 [216999] [240579] [A037Max] [116999]

4: Arabidopsis thaliana gene
 [1014410] [201195] [ATAG11981] [1014410]

5: Lotus corniculatus aramo transcribe rRNA 5S gene complete cDNA
 [260593] [244767] [2696] [A0099] [260593]

Display: Summary Save Text Clip/Save

Show: 10 Items 1-5 of 5 One page

Search for Genes, Links, protein data, similar information for human, rat, mouse, rat, and species.

Entrez Nucleotide Help | FAQ

Both Entrez and EMBL file of GI or accession numbers to retrieve sequences.

Check sequence region history.

NCBI maps to WWW | NCBI Entrez

Links

Guppy

Feedback | Contact Us

- 5) First click on accession number of Medicago then Glycine followed by Arabidopsis and Lotus. We will get sequence of Medicago, Glycine, Arabidopsis, and lotus.

NCBI Sequence Viewer - Microsoft Internet Explorer

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=5711837&dopt=GenBank

NCBI Nucleotide

Search Nucleotide for: [] Limits [] Preview/Index [] History [] Clipboard [] Display: Default [] Save Text [] Add to Clipboard []

1: L25335. Medicago sativa a...[gi:413726]

LOCUS ALFAAT2A 5788 bp DNA linear PLN 10-FEB-1997

DEFINITION Medicago sativa aspartate aminotransferase (AAT2) gene, complete cds.

ACCESSION L25335

VERSION L25335.1 GI:413726

KEYWORDS aspartate aminotransferase.

SOURCE Medicago sativa.

ORGANISM
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae; eurosids 1; Fabales; Fabaceae; Papilionoideae; Trifolieae; Medicago

REFERENCE 1 (bases 1 to 5788)

AUTHORS Gregerson,R.G., Miller,S.S., Petrowski,M., Gannt,J.S. and Vance,C.P.

TITLE Genomic structure, expression and evolution of the alfalfa aspartate aminotransferase genes

JOURNAL Plant Mol. Biol 25 (3), 387-399 (1994)

MEDLINE PUBMED

FEATURES

source Location/Qualifiers

1..5788

/organism="Medicago sativa"

/cultivar="Saranac"

/sub_species="sativa"

/db_xref="taxon:3879"

1525..5247

/gene="AAT2"

join(1525..1599,1680..1721,2537..2731,2873..2911,3000..3089,3205..3313,3405..3514,3767..3910,4107..4273,4700..4964,5116..5247)

/gene="AAT2"

/codon_start=1

/product="aspartate aminotransferase"

/protein_id="AA14c_----"

/db_xref="GI:777387"

/translation="MASSSLLSSVPSHSASLSILDNTIKGKLLKGINNFSNLRSSGRI
CMAVATNVSREFGIPMAPDPDILGVSEAFKADTNDVKNLNLGVGAYRTEELQPVVNLV
KKAENLMLERGENKEYLPTEGLAENKATAELLGADNPAIKQQRVATVQGLSGTGL
RLGAALIERYPFGAKVLISNPTWGNHKNIFNDARVPWSEYRYDPKTVGLDFEGMIED
IKSAPEGTFVLLHGCAHNPTGIDPTPEQWEKIADVIOQKNHFFPFVDVAYQGFASGLD
EDAASVRLFESRGMVELVAQSYSKNLGLYAERVGAINVISSPESATRVKSQLKRLAR
PMSYNPPVHGARIIVANIVGTALPFDENKAEEMMAGRIKTVRQALYDSISKDKSGKD
WSFILKQIGMFSPTGLNKSQSDNMTNKHVIYMTKDGRIISLAGLSLAKCEYLADAIIDS
YHNVS"

BASE COUNT 1791 a 934 c 963 g 2100 t

ORIGIN

1 aactgtatga tgaagttaa gtgaccaat. agtcttttaq gqatctgaat tctcattaca
61 aagttataag ttatatctgg taatcaatt. tqatcaratf acaaatftac taactcaatt
121 ttaaaatact tctaccatta atttaaalft atttagatct actaatcgta
181 tgtattctct ctttcgagga agtggatgca. ttcaactctf gqatattttt tttaatggac
241 aatgttaata tghtaaftqt tqtatgttta ttttttttca ctttqttaa qcttqaatca
301 tgaccttcaa gtccttaacc ttttaggtca aaccaattga gctacctaat ccaccctcc
361 aqtagtagat gtgtgttttt cctcctctta tttcaacttt tctccttttt talactalaaa
421 actagactct aaaactttta gatagacaaa tcaatctccc ctttcaaac actctcttt
481 ttttaaaact cttaactaatt taaaagtatg agccttttql aqgtqaccac gactccttt
541 caactctaat tgacttcaac tctantatca ctctcattca ttttactgta ccaccggtaa
601 gtaaaaaagt cttacatttg aggaaaaaaa atgagatgac acalatagtt aatatcttta
661 aggttttagg tgggaalgta atgtctcttt catttqtatg acqtatctcq actcattgtg
721 aatggctctc atagagactt cctctatgat ccaaacactt catcactqaa qatatttttt
781 ggtgttaacc tctgtttctt aggagaaaqt aqctatagta atcqaqaatt cqtcaaaaa
841 ataagtaaaa cctgacaaaa aactcgtact. tcaactatua aqatataate ttctactgta
901 ctcarattta atgaaatttg tactaagaaf atatttttga aaaaaalqt ccacacaaa
961 aaacttgtt ttcattttcc cctctttaal. cttttatcaa aqaaaalaaa aaactatttt
1021 aaacccaaatc atcagattttt tttttgtqr aaccogaaat qaacctctgaq taactatgaa
1081 tcaactcgtg aagtaattat atctcgacca aaaaattalc caactcaaaa ttgaactttg
1141 ttactctgaa cgaattctcc tttagtttqa atctctaac actttgaact caattggttg
1201 gttctatgaa cttaaaaaaa tgaatttttt. cctctatnat ttgaagaaaa tatattatt
1261 attattattt ltttttgaca aqaattalga laattttttt tttatbaag ttlaaaatgaa
1321 aagtgataaa acaaaaactt gtctgttttg ataqaalqqa gqacattgta gttgtctgca
1381 gccactagtt tcaatttata tcaactctac tcaactcttg tgaactctac taactttta
1441 ttacaaaaca acacaaaana atcaaacca. tttcgaact ctttctgttg atgtctttg
1501 gtttttctct ctcaacaaca aaatatggca tnatcttcat tactctctc tglactctca
1561 cactctqct. cactttgat cctcgaacc. acaactcaagg tttctqtat gaactcaact
1621 ttactcahrt atgaaatgatt taactcaartg attaatttat ttttqtftc attcaacagg
1681 gaaagctlaa gcttggaaatc aacaactct. ccaatttng ggtgtgata tataatgta
1741 ggaactctca tttatctcat tccaattttt atlatttta tatcattta ctctatgatc
1801 aatttaaggv tgtgtataa laattcaatt. catcttact. actaatgta ttcattaact
1861 actttttttt tttttglac ttttttaaa. ttatcattt. tcaactctac tttagattaa
1921 aaatgtaatt. tcaagctoga ggtattagtc ttgtcaqitg ccgqcgaga tttagaccgt
1981 gttgtlaate qatgttcca atttttgtg. tcaataalg attgnaaac agttttaaaa
2041 ttgcaattca aagtggttta tattattaci. aacacacatg tcaqtact taatagtlac
2101 tgalgtlctt caacaactac ttttaattgt. talttgtta atttactaat qttacttata
2161 caactctact tttagattaga atgtl.aaaac. caagtgttg gatcaagcaa ttaacaaggt
2221 ttagttaaaa ttgaaatgat. tatgaaatf. gagtttqaag cctqactgaa acaattgttg
2281 atcagagtca gactttalac. acctccaqac. atcagat tac caggggtcc. ttctctgga
2341 aaccggaggy ttaagcaaaa. aataagaaa. taqaalagta attcaattcc tgaatgtaaac
2401 tctgttctta tttttgtttc tataatgga. caaagaatg ttttaaaatt gcaattgaaa
2461 ttggttaatt aggatttagg aggtqgtct. aclacctga. tcaattctat ttlyctgtt
2521 gattttgtt attcagtcac ctggllcggal. ctqcatgct. gttcgacta atgtttctcg
2581 gtttgaggtt atacogattg ctctcctgta. tccaaltct. gtagtftct. aactatttaa
2641 agcagacag atgatgtca agctcaact. ttgagttggg gcclacagaa ccggaagaact
2701 acaaccatatt qtgcttaatg ttgtgaaaaa ggtatltgat gctgtcttt ttttttaca
2761 gtttgcacat tctcgtgccc atgtattgat tcttatttt actcttgtt gaaaaactg
2821 aggcattttt ttcaagttgag tcaatattga ttgatgaaat latlgttttt aggcagaaaa
2881 tcttatqctg gagagagggg aaaaacaaagv ggttaattttc tttaggtcgt atgcaattta
2941 tttgtgcata aaattgttga gctttagatt. attgactgtg atttcaactg gaattcgact
3001 actccccat tgagggtttg gctgcattta. acaaggcaac tgcagagttg ttctcggag
3061 cagacaatcc agcaactcaa cagcaaaagv tatgtagtgg atttagaaaa tataaatatt
3121 tgaagtata taacatttat tgactgtgag atcaaaaaa caaaccaagaa ataaatcaac
3181 tttttctctc tcttttttcc tcaagttgcc. actgtccaag gttcttcagg aactggtct
3241 ctgcagactg ctgactctct gatagaacg. tattttctg. gacgaaaaat ttgtatca
3301 aatcctacgt ggggtgtgaa ttttatctct. ttctcggttt aacttggtca tgaatttgt
3361 ttcgagggaa tctgtacac gtaggtaach tttttgtct gcaggtaac caagaattg
3421 tttcaacgat gctcgtacac catggctgta gtaccgatac tatgaccca agacagttgg
3481 cttggatttt gagggcactga tagaagat. aaaggttaca tctcatcca ttaattgaaa
3541 aatgaaatata atcatacttt ttatattgac ttgatattt. tttcaactct tttcttttt
3601 tgcattagaa agtaagatg atactgattt tgatttataa gtatgtttta atctacttg

3661 tagtgaaggc ctacagaaga acacgaatc acatttaatt agaatacca tgcaattcta
3721 aagaaaatgg aacttcaact aattgcaact gtttgcatt ttatagtcgg ctcgggaagg
3781 aactttttgg ctacttcaat gatgtgcaca faacctact ggtattgac caaccaccga
3841 acagtgggaa aaaatagctg atgtaatca acaaaagac cactttccat ttttgatgt
3901 tgctaccag qcacacaac tttggtaaac gaactactt rtttgraaat ttgtggacag
3961 cgtgttcaaa tgcgtlaac aaaatlgtaa ttaagacaaa taacacttac acaaaatgaa
4021 gtggatacaa catgaaaccg gtgggtatcg tgcagtaacc gtgaatatat ttcacgatal
4081 actaaactgc tcattaaaaa tttcagggtl ttgctagtgg aagccttgat gaagatgagg
4141 ctctctgtgag attgtttgag tcacgtggca tggaaagtct tgtagctcag tcatacagta
4201 aaaaacctcg cctttatgct gaaagggttg gagctattaa tgtcatttcc tcataccagg
4261 aatctgcaac aagggtgtgt aaattacgta tgccttgcct tctaataatt ttgaaattca
4321 atgcacttgt ctaaaatatt gaattctgal tctttictag gtagggtgaa gtttcatatt
4381 atcaatcaag agtcataggg aataaactta acgataggca ttttqttaga atagaaaagt
4441 tgaanaatc ttgtat tact gctacagtc ttttgcataa tcacltctcq aatgaqttga
4501 aatgaacttt cacacalaa cttctttgal agctctctca ttttctacat gagcacctat
4561 agtatatat caacttaatc aattgttttg ttttaagaagg aagacatctt gtattcaact
4621 tgnccaatgc aatgtatgct alagtaaac tctgtcttg gtcnaactc acaatttga
4681 tqattactat gtatttcagg gtaaaagacc aatlgaaaag gcttgcctga ccaatgtact
4741 ctlaatccacc agttcaaggq gctaggatlg ttgctaatat tgttqggact ccagctctct
4801 ttqatgaatg yaaagcagaa atggaaatga tggctggaag gataaaaact gtlaggcagg
4861 cgcctgatga lagtatctt tcaaaagaca aaagtggaaa ggatlggtca ttcatactca
4921 aqcatatag catgtttctca ttcacagqcl tgaacaaagag ccagggtttt gactcccccl
4981 ttgacttate lattttatac qttatctlna qtectalile tttgacteta tcaacgtaat
5041 atgagratgt gatttqtatt taatgtattq catlattgar ttgtgttctc tttgttatat
5101 cttclgtttg gtcagagtga caatatgaca aataagtgcc atatacatat gacaaggat
5161 ggaagqattt cctgtggcagg attgtcctq qccaaatqlq aatacctggc agatgctatt
5221 atcgattcat uLcataatgt cagctgaaac qcadtgaaac alqcttttga agcaagcata
5281 tatgtgtgagt attataccaa atcatagttu ttgacacatl acaataattl tatcatgtat
5341 qcatttqttg tcatttttca taigtaccca aagtcctctt qqaanaatg tttgtaacct
5401 gaaLaagttg aatcaaatg Ltgatgtaac aaacgaqtct cttttgcaga cttgaaacaa
5461 gtlgaaccta attattgaat tttagagtat tttcttalgt ttgnaatct actatattaa
5521 atcatttttt gcagacttca agcaaagttga acctaatlat tgatttggag tatgtttctt
5581 attgtttgtaa tactactata ttaaactatf ttacgttal f agttctcaa aattggtaac
5641 accaatgtca atagaggaca ttgggtgaaa cagttaacgg tttaagtcac cataaatgat
5701 tcaagcaact tgaaatggat tcaaacctq ccttcaagga aaacatgcat tcatqcagtc
5761 aatgtatcat aatcgttaga caaaqatc

NCBI Nucleotide
 Search Nucleotide for
 Limits Preview/Index History Clipboard
 Display default Save Text Add to Clipboard

1: L40579. Glycine max (clon. .[gi:710595])
 LOCUS SOYMAA 1717 bp mRNA linear PLN 30-APR-1996
 DEFINITION Glycine max (clone pSAT2) nuclear-encoded mitochondrial aspartate aminotransferase mRNA, complete cds.
 ACCESSION L40579
 VERSION L40579.1 GI:710595
 KEYWORDS aspartate aminotransferase; nuclear-encoded mitochondrial protein.
 SOURCE soybean.
 ORGANISM Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae; eurosids I; Fabales; Fabaceae; Papilionoideae; Phaseoleae; Glycine.
 REFERENCE 1 (bases 1 to 1717)
 AUTHORS Wadsworth, G.J., Gebhardt, J.S. and Matthews, B.F.
 TITLE Characterization of a soybean cDNA clone encoding the mitochondrial isozyme of aspartate aminotransferase, AAT4
 JOURNAL Plant Mol. Biol. 27 (6), 1085-1096 (1995)
 MEDLINE
 PUBMED
 FEATURES
 source
 1..1717
 /location/Qualifiers
 1..1717
 /organism="Glycine max"
 /strain="Century"
 /db_xref="taxon:3847"
 /clone="pSAT2"
 /tissue_type="green leaf"
 /dev_stage="14 day old light grown seedlings"
 1..53
 54..1337
 /EC_number="____"
 /note="nuclear-encoded protein"
 /codon_start=1
 /evidence=experimental
 /product="mitochondrial aspartate aminotransferase"
 /protein_id="____"
 /db_xref="GI:710596"
 /translation="MAIRNSLTGQFLRRSSVAGARLMSSSSSWFRSIEPAKPDILGV
 TEAFLADQSPNKVNVGVGAYRDDQRKPVVLECVREARRVAGSQFMEYLPMMGSSIKMI
 EESLKLAFGDNSEFIKDKRIA AVQALSGTGACRLFAAFQQRFRHPTQIYIYIPVPTWANH
 HNIWRDAGVPMKTRFRYHYPESRGLDFSGLMDDIKNAPDGSFFLLVLTAHNPTGVDPSE
 EQWREISSQIKAKGHFFFFDMAYQGFASGDPERDAKAIKIFLEDGHLI GLAQSYAKNM
 GLYQRRAGLSVLCEDEKQAVKSQLQLIARPMYSNPPLHGALIVSTVLGDPDLKLL
 WLKEVKVMADRIIGMRTTLRENLEKKGSTLPQWHITNQIGMFCYSGLTPEQVDRWNTN
 FHYIMTRNGRISMAGLNTGNVGYVLDLAIHEVTKSF"
 54..128
 /function="targets protein to mitochondria"


```

                                /note="putative"
                                129..1334
                                /product="mature mitochondrial aspartate aminotransferase"
                                /EC number="___ _"
                                1335. 1717
                                1717
BASE COUNT: 477 a 361 c 382 g 49 t
ORIGIN
1  ctctccctct ctgttgcac tetgtcttc cctctttc gcctactga gtcatggca
61  ttgcraactc gctcaacggc caattctcc gccgcagctc cglcgcgga gcaaggctca
121 tctctctctc qctctcatgg ttccggagca tccgaqccc tcccaaggat cctatcctg
181 gadtcaactg agctttrct gccgataga gtccaaacaa agtcaacgtc ggaagtgggtg
241 cgtalccgga tgaccacgg aaacctglgq ttttggantg tgttagagaa gcagagagga
301 gggttgccqg aagtcattc atggagtatc tcccatggq tggaaqcata aaaatgataq
361 aagaatcctc gaagctggca tttggaqaca actctgaqt cactaaaggat aaaaqaatag
421 ctgcagtgca ggctttatc tgggctggtg cctgtccact ttttgcctca tttcaacaga
481 gatttcctc taataccc aa atctatata caqtgcctac ctgggccaa caccataaca
541 tttggagaga tgcctggagt cctalgaaga ctctccgtta ctatcaccct gagtctagag
601 gattggallt ttcagatctg atggatqaca taanaaatgc tccagatggt tctctcttc
661 tgcctgtctc tactgctat aatctatctg gqtagatc ttcagaaga cnatggagag
721 agactctctc ccagataaag gctaaaggtc atttccctt ctctgacatg qcatatcaag
781 gttttgctag tggatgacca gaggagatg ccaagccat aaagatttt cttagagatg
841 qtcatttaat aggaactcct cagtcatact caaaaatct yggactat ggcacgcag
901 caggaagcct gagtgtgctc tglgaagatg aaaaaaaqc tglggctgta aaaagtcagt
961 tccagctgat tctagacc cttacagta accacctc ccatggaqca cttatagtt
1021 clactgtcct tggatcaca gatttgaagc agttatggtc taagaagtc aaggttatgg
1081 cagacgcgat ccttggaaag aqgactaca tccgagaaac cttagaaaag aaggggtcta
1141 ctttgcctag ccagcaata actauctaga ttggtatgtt ctgtcactg ggatgacac
1201 clgaacaggt tgcctgatg acaaacgagt tccataatc catgaccct aacggtcgt
1261 tccagtatgg tggctctaat accggcaacg ttggtatgct cttgagcct atccatgagq
1321 ttacaaaatc attcctaat attcaatca ccaacaggtc accaaatctt tgtctttgga
1381 qacaaaatc tcagaaqtaq taccctatc cagggagaca aanaaattt aatnaagct
1441 gacaaaatct caactgttt cgggtaagc tagtgatct cttcaggggt taccttgcac
1501 attgaagacc ctattgcaa tacgtatcct atttatata ttrattgac atagattata
1561 tccattttta cagacatct gtttagcctc ttaagagca taaccctgta tcatcatttt
1621 tttataaaa taatttgaqt agtagattga atnaaggtt falctatggt qaacaggaag
1681 tacgtagctt cactctctac taccattta tctgtt

```



1: X91865. A.thaliana asp5
g...[gi:1017410]
LOCUS ATASP5GEN 4836 bp DNA linear PLN 14-NOV-1996
DEFINITION A.thaliana asp5 gene.
ACCESSION X91865
VERSION X91865.1 GI:1017410
KEYWORDS asp5 gene; aspartate aminotransferase.
SOURCE thale cress.
ORGANISM

Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis.

REFERENCE
1 (bases 1 to 4836)
AUTHORS Wilkie,S.E., Lambert,R. and Warren,M.J.
TITLE Chloroplastic aspartate aminotransferase from Arabidopsis thaliana:
an examination of the relationship between the structure of the
gene and the spatial structure of the protein
JOURNAL Biochem. J. 319 (Pt 3), 969-976 (1996)
MEDLINE 9706001
REFERENCE
2 (bases 1 to 4836)
AUTHORS Wilkie,S.
TITLE Direct Submission
JOURNAL Submitted (28-SEP-1995) S. Wilkie, Institute of Ophthalmology,
University College London, Dept.Molecular Genetics, Bath St.,
London, EC1V 9EL, UK

FEATURES
source Location/Qualifiers
1..4836
/organism="Arabidopsis thaliana"
/cultivar="Landsberg erecta"
/db_xref="taxon:3702"
/tissue_type="leaf"
/clone_lib="lambda DASHII"
/germline
1821..1826
join(2237..2296,2389..2439,2563..2757,2832..2870,
2955..3044,3142..3250,3335..3444,3652..3795,3888..4054,
4193..4457,4552..4683)
/gene="asp5"
join(2237..2296,2389..2439,2563..2757,2832..2870,
2955..3044,3142..3250,3335..3444,3652..3795,3888..4054,
4193..4457,4552..4683)
/gene="asp5"
/EC_number="1.3.1.11"
/note="homodimer"
/codon_start=1
/product="aspartate aminotransferase"
/protein_id="At0g10174.1"

```

/db_xref="GI:1017411"
/db_xref="SWISS-PROT:P46248"
/translation="MASLMLSLGSTSLLPREINKDNVKGTSASNPFLKAKSFSRVTH
TVAVKPSRFEGITMAPDPILGVSEAFKADTNGMKLNLGVGAYRTEELQPYVLNVVKK
AENLMLEKRGDNKEYLPIEGLAAFNKATAELLFAGHPVVIKEQRVATIQQLSGSGSLRL
AAALIERYFFGAKVVISPTWGNHKNIFNDAKVPWSEYRYDPKTI GLDFEGMIADIK
EAPGGSFILLHGCAHNPTGIDPTPEQWVKIADVIQEKNHIPFFDVAYQGFASGSLDEI
AASVRLFAERGMEFFVAQSYSKNLGLYAERIGAINVVCSSADAATRVKSQLKRIARPM
YSNPVPHGARIVANVVGVDVTFSEWKAEMEMMAGRIKTVRQELYDSLVSKDKSGKDWSS
FILKIQGMFSFTGLNKAQSDNMTDKWHVYMTKDGRISLAGLSLAKCELYADAIDSYH
NVS"

```

```

2237..2296
/gene="asp5"
/number=1
2297..2388
/gene="asp5"
/number=1
2389..2439
/gene="asp5"
/number=2
2440..2562
/gene="asp5"
/number=2
2563..2757
/gene="asp5"
/number=3
2758..2831
/gene="asp5"
/number=3
2832..2870
/gene="asp5"
/number=4
2871..2954
/gene="asp5"
/number=4
2955..3044
/gene="asp5"
/number=5
3045..3141
/gene="asp5"
/number=5
3142..3250
/gene="asp5"
/number=6
3251..3334
/gene="asp5"
/number=6
3335..3444
/gene="asp5"
/number=7
3445..3651
/gene="asp5"
/number=7
3652..3795
/gene="asp5"
/number=8
3796..3887
/gene="asp5"
/number=8
3888..4054
/gene="asp5"
/number=9
4055..4192

```

```

/gene="asp5"
/number=9
4193 4457
---
/gene="asp5"
/number=10
4458 4551
1
/gene="asp5"
/number=10
4552 4683
---
/gene="asp5"
/number=11
BASE COUNT      1377 a      886 c      959 g      1614 t
ORIGIN
1  qgtaccacaaa caaacacaaa gtacilaalca tttttarttt ttaccgggat tcttgatct
61  tqtattttate atctaaatct cagaatgaaq ttgatttgtt gagcaagatt caccaccoga
121 arataratrc atgtgttggg tatggaaatq aactcagttc gaggtttatc gtcctcagcg
181 tqatggaagg cggatcattg gatcacaggt tacacggtaa aaatagctca aaacatcttc
241 tcatcatcgt gtaatagatc atgagttaaa ttaatarctat agtataacag agataaaca
301 aatcagatgt cccacatgaa ataaataat1 tttgttttqa taaaatttaa tgtttgcctc
361 attaatggtt aaggaccttc tcggggatcq gctt taacat qgcacatgcy gatgaagatt
421 gctccttgata cagcaaggya ctacactcaa gtgcccacat ttaacaatcc tctctaccgg
481 ttgattaatc tcggttiagg taaaccggtt trttatgttg gtttggaatt agagctgttg
541 aqatctccca cgagcgttgt cgtctcccgq tttctccacq agatrttaa tcgtcaaaaa
601 tttctcttca tttctctctc aacgc caag taaacaactt gaatcattaa tgatgattct
661 gttctgagct aaccgaaaca gatttgtttt ttgtattcaa gaatcalttc attcatttgt
721 ctttagttcc atctctaacg atagatcaar attc agattt cggattttgg tcttgcggta
781 atggtggggg ctcaacggaa aaacaacat1 aacactatrag caacatcttg tctgttgcct
841 ccagaatafc tcttagatgg taaaatatta acatctccct ctatattgat cgactttgac
901 trgtgtcatc atcactagag ttatagcaca tagtttgata gtcataatca gttgaccaca
961 taacattaag tttcttctat tacgtctaaa attgttgrtc taaatcaaat ctttttgttt
1021 tgtatgatat tatttcaact aatcacaact1 aaaaatcatt1 atttgattag agatgtagt1
1081 tgtfgggtcg ctcaaacggg ttatigagt1 tgtctcattg tcaactcttt attgtccctca
1141 ggaataatga cggataagag tgalgtttat cggctttggg tggttttact tgaactcttg
1201 ttaggaagac ggccgggtga gaaatl gagi tcggttcagt gtcaattctt tgcactctgg
1261 gtaactcctg tttcgttga cgttatgtrc atcaatacat gaatcagata gctttgggta
1321 atgagtgtae cgtttc gaaa aatttgcag1 aatqccccaa ctacgggata gatc aaagt1
1381 tccqaaatc cgtggatccgg ttatcaaaq1 ta aatggat1 caacatcttg taccaggt1
1441 tlgcgtcatc ttttctctt1 tlggttagc1 ftaaaaatcc gattgggtac atatttttg
1501 tqttyaaaac taaataaaac cggaaataaa ttgttcaggt ggcaqccgtg ccaglgctt1
1561 gttcaaacac agaaccgagt tarcgac1 gl tyataac1 rya tgtctctcac tca1 tagttc
1621 c atiggttaa ggtagagcta ggagggact1 tccggttat1 accatctcgt tcttgatt1
1681 aqaaaataat atttttttgc tctctttttg aattt1 acag1 ttgatctct1 gatlgcttt1
1741 agtaatt1taa tttgggggtg aaaaatgtga ggacagaagg gttatcact1 gtcattatt1
1801 lat1tgtaaa ttgattaaata taaaattg1 tttgtrc1a agtgtaaaa1 tcaaat1
1861 tgttgcataa accactaatc acaatccact agtarccatc catccactct tcaactttt
1921 aaacataatc ctaaatgact atcttaccct1 fatcaagaca agccfggaaa gtgagaac1
1981 ttcgtagaca tctcgtgttc tctctctc1g tgaatcct1g ttaactcttt tcttcaact1
2041 ctctagctga tegtattcga aaccgcgc1a tcttaacc1a gtrcaggtga caaatcag1
2101 ctg1 taaatt atctcaaaag tctcgtgat1 atattttg1g gatcacatca tttgtctatt
2161 tttgqaattg cttatctctg cactagtcca ctgatttga1 atrtgtgcag ttaattttg
2221 gr tccatagc gattccatgg cttctttaa1 tctatctctc ggttccact1 cctctgtacc
2281 gcrcgagatt aacaaggtae ttttgc1g1 tctatcagat atatggtaga aatcagct1
2341 ragltttatc1 aaaaatctac tttttgtt1 tcgat1aaac tfgaacagga taactgtaa1
2401 cttggaactt ctgctcgaa1 cccgttcta1 aaagcaagg1 tactcttt1 tctgtttag
2461 tgcgcgaact tgaaccat1a1 gaacaatttt1 gaatgacgt1 ttagagtggt1 attagtcaa
2521 taagtgtagt ttgtatctga aatttgtgga1 ttttggttc1 agtcttttag1 cagagtgact
2581 atgacggtg1 cagtgaaacc1 tctctgtt1g1 gagggtataa1 ctatggctcc1 accagacct1
2641 attcttggag1 tcagtgaagc1 attcaaaag1c1 gacactaac1g1 ggatgaaact1 caatctgtt1
2701 gttggtgct1 atcgtactga1 ggaactcca1g1 ccttatgtgc1 ttaagtgtg1 taaaaggtt1
2761 ggaactgct1 accttagtra1 atctcgtat1g1 agagaggaga1 taacaagaag1 gtaactgtg1
2821 tgcgtttcta1 gccggagaat1 ttgat1ttg1g1 agagaggaga1 taacaagaag1 gtaactgtg1
2881 ttatttgaat1 ttgtcaagc1 gattacatca1 tcagaat1a1 actaaat1a1 accctcacgt1
2941 gatggaactt1 gcagtatctt1 ccaattgagg1 ggttggcagc1 attcaacaag1 gctactgct1

```

3001	iqttgctatt	tggagctggg	catcctqtt	ttaaggaica	aaqagtaatt	ctgcaccttt
3061	tgttcatcat	gttatataaa	tgttttctca	tgatgttcc	gtttctgagc	taatccattt
3121	accttggctc	taaatcacca	ggtagcaaca	atccagggct	tltcgggaac	aggttcactg
3181	cgattagcag	cggctcttat	agagcgttat	ttccctggag	caaaaagttg	gatctcatca
3241	craacctggg	gtacattgtc	tggaccaraa	acatttttga	gtgatttgg	ttcatatfta
3301	accgccttta	tcaactgaca	aaataccaca	tcaggtaatc	acaagaatat	cttcaatgat
3361	gccaaaagttc	cgtygtccga	ataccgctac	tatgatccaa	aaacaattgg	tttggatttt
3421	qagggaaatga	tagcagatat	aaaggtttgt	ttgtccataga	aggtctatgt	ttagttgaca
3481	ttgclttata	tatgcatcaa	caaacaggat	atcatttttg	tgatgccaaa	gtgagtattt
3541	gtttgcagtt	atactgtctg	tgtgataaaa	cgaqtacttg	tttaaaqaca	acgtgaaacc
3601	ctgatggatt	ataatggtta	atcttgaatl	tctqtatgt	aaaatdtaca	ggaagctcca
3661	gaaggatcct	tcactcttgc	tcacggatgl	gctcacacc	caactggaat	tgacccaaca
3721	ccaqaacagt	gggtaaaaat	tgctgatgtc	atrcaggaaa	agaaccatat	cccatttttc
3781	gatgttgcat	accaggatc	cccctattcc	taratttctg	aatatcgtgt	ttccgagaga
3841	agtaaqcgaa	accatattat	gtaacgctac	atgcacctat	tttccagggc	tttctagtgt
3901	gaagccttga	tgaagatgca	gcatctgtga	gattatttgc	tgagcgggga	atggagtttt
3961	ttgttgctca	gtcalatagt	aaaaatttaq	gtttgtatqc	agaaaagaat	ggggcaatca
4021	atgtctgtgt	ctcctcagct	gatgctgtca	caaggtaaaa	cttactctaa	attttcttat
4081	catgagcctg	ttagcaagtga	tcctggagtr	tcagtttcta	ctttaccgr	cacgaaccat
4141	atctagctac	aggttttatg	gaaccttla	cagtaaacct	accttcttcc	aggttcaaga
4201	ccr agttgaa	aaggatlgct	cggcctatyl	actcqaatcc	accagttcat	ggggcgagaa
4261	lcgtggccaa	tgtagtgggl	gatglaacta	tgttcagtga	atggaagca	qagatgaaa
4321	tgatggcagg	aagaataaag	acggitaga	aaagctgla	tgaagcctc	qtttcaaaag
4381	acaagagcgg	gaaggactgq	tccttcaatc	tgaagcaat	tggcatgttc	tctttcacccg
4441	qctaaaiaa	agctcaggta	tqccaacaa	tttaatacta	cacagatccc	ctttggcctt
4501	qllttagact	gtggtttgaa	gtttgtgctg	tttcttttq	gtaacgagca	gagcgataac
4561	atgacqqaca	aatggcatgt	gtatnlgact	aaagacggga	ggatalcatt	gqccggatfa
4621	tctctgqcca	aatcgagila	tcttqctgai	qcgatcatc	actctacca	taacgtaagc
4681	tgagcttcca	tctcagtaga	tqacaataag	aaacagtttl	atataccctt	tttaqcttct
4741	lcttgatctt	tgattgaaac	cagatggcta	aat taatag	ctagatcac	actatgattg
4801	ctagacaait	tccaaggaaa	cttttgaiit	gaaitc		

NIH Sequence Viewer - Microsoft Internet Explorer

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=5711837&top=GenBank

NCBI Nucleotide

Search Nucleotide for Limits Preview/Index History Clipboard Details

1 AF029898 Lotus corniculatus [gi 2605931]

LOCUS AF029898 1685 bp mRNA linear PLN 06-APR-1998

DEFINITION Lotus corniculatus aspartate aminotransferase mRNA, complete cds

ACCESSION AF029898

VERSION AF029898.1 GI 2605931

KFYWORDS

SOURCE Lotus corniculatus

ORGANISM

Fukaryota, Viridiplantae, Streptophyta, Embryophyta, Tracheophyta, Spermatophyta, Magnoliophyta, eudicotyledons, core eudicots, Rosidae, eurosids I, Fabales, Fabaceae, Papilionoideae, Loteae, Lotus

REFERENCE

1 (bases 1 to 1685)

AUTHORS Morlier, J M and Gregerson, R G

TITLE Isolation and DNA sequence analysis of an aspartate aminotransferase cDNA clone (Accession No AF029898) from Lotus corniculatus (PGR98-034)

JOURNAL Plant Physiol 116 (3), 1191 (1998)

REFERENCE

2 (bases 1 to 1685)

AUTHORS Morlier, J M and Gregerson, R G

TITLE Direct Submission

JOURNAL Submitted (13-OCT-1997) Biology, Lyon College, 2300 Highland Road, Batesville, AR 72501, USA

FEATURES

Location/Qualifier

source 1 1685

/organism="Lotus corniculatus"

/db_xref="taxon:47247"

/tissue_type="nodule"

119 1492

/EC_number=" "

/codon_start=1

/product="aspartate aminotransferase"

/protein_id="/^ " "

/db_xref="GI 2605932"

/translation="MAASSVFSVASHSVSPSNHHAHKGKTKIGGSGLRLANSRFSGG
RISMAYVNASRFEIGIPMAPDDPLGVSEAFKADCKDLKLNGLVGVAYRTEELQPYVLN
VVKKAENMLNNGENKEYLPIEGWAAFNKATAELLGADNPALKEQQRVATVQGLSGTG
SLRHAAALIERYPFGAKVLI^SPTWGNHKNIFNDARVFWSEYRYYPKTVGLDFEGLM
EDIKSAPEGSFVLLHGCAHNPTGIDPTPEQWVKIADLIQQKNHIPPFDVAYQGFASGS
LDEDAASVRLFVSRGMEVLVAQSYSKNGLYAERIGAINVSSSPESAARVKSQLKRI
ARPMYSNPVHGARIIVADIVGNPDLFNEWKAEMEMMAGRIKNVRQKLYDSISSKDKSG
KDWSPILKQIGMFSFTGLNKNQSDNMTNKHVYMTKDGRIISLAGLSLAKCEYLADAI
DSYHNVS"

BASE COUNT 501 a 316 c 380 g 488 t

ORIGIN

1 ttcgattcat ccggtacgcc attaaaccac tctgcaacta gatccatcac tctttcactc
61 gagtttggtt gtaaccggtt ccgttatttc tgttgtgtgc ggttgttttc cgttgagat

```

121 ggcggcgtct tcagtggtct ctgtagcttr acactctggt tcgccttcga atccatgc
181 tcacaagggg aaaaccaaga ttggaggtag cggtttgaga ttggcaaatt caaggtcttt
241 tggtagtggc cggatctcta tggctgtgct tgttaatgct tctcgatttg agggatacc
301 gatggctcca cctgatccaa ttctcggagt ttctgaaqca tttaaagcgg acaaatgcga
361 tctcaagctc aatcttggag tcggggccta cagaactqaa gaattacagr catatgtct
421 taatgttgtt aagaaggcag agaactctat gctgaataga ggggaaaaa aagagtatct
481 acctattgag ggttggqctg catttaataa ggcaactgca gagtgtttac tcggagctga
541 caaccagca atcaaaagac aaagagtgr cactgtccaa ggtctttctg gaactggtc
601 tctgcgacat gctgctgctc tgatagagcy atattttcca ggggaaaaa gtttgcatac
661 atccccacc tgggtaatc acaagaata tttcaatgat gctagagtc catggtcaga
721 gtaccgatat tatgat.cta agacagtgtg attggatttt gagggcatgt fagaagatat
781 aaagtcagct cctgaaggat ctttcgtgct acttcatgga tgtgctcata accctaccg
841 tattgatccc acaccagaac agtgggtgaa aatagctgat ctaattcac aaaagaacca
901 cattccattt tttgatgtcg cttaccaggg gtttgcctagt ggaagcctgg atgaagatgc
961 tgcttctgtg cgattgtttg tgracgtgq catggaggtt ctatagctc agtcatacag
1021 taaaaatctt ggtctctatg ctgaaaggal tggagcaatc aatgtcattl cctcatcacc
1081 agaatctgct qcaaggqtaa agagccaatl gaaaagqatt qcaaggccaa tgtactctaa
1141 tccacgggtt lacgggqcta gगतtgttg: tgatatagtt qgaaatccaq atcttctcaa
1201 tgaatggaaa qcagagatqg aaatgatggr aggaaqqata aagaatgta gacagaagct
1261 atatgatagt atttctcaa aagacaagay tggaaaaggat tggtrattca tacttaagca
1321 gataggcatg tttctattca caggrttgaa caagaatcaq agtgalaata tgacaaataa
1381 gtggcatgta tacatgacaa aagatggaaq gatttccctg gcaggattgt cactggccaa
1441 atgtgaatc ctgtcagatg ctattattga ctccatcat aatgtcagct gaaactagat
1501 gaaatattct ttatcaccda gcttatattt tttgggtgagt attgtatcca atcatagtg
1561 tggcacaca tgacaataat tcatgtaaa attaatgtca tacatgtact ttttaattccc
1621 taggaattt gtaaccttaa aataagttga atcaaaactat tgatgcaaaa aaaaaaaaaa
1681 aaaaa

```


- 6) Note the Exon regions of Medicago, Arabidopsis, Glycine, and lotus.
- 7) Make a Multiple alignment by putting the sequences (only Exons) of Medicago, Arabidopsis, Glycine, and lotus in Clustal W.

Job running: <http://www.ebi.ac.uk/services/mp/341166.716022-11800.html> - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ebi.ac.uk/services/mp/341166.716022-11800.html> Go

Customize Search attempting to retrieve buttons from Yahoo!



Your job is currently running...
please be patient

The results of your job will appear in this browser window.

Your job output will be available again in 15 minutes.

Please Note the Following:

- You may press Shift+Refresh or Reload on your browser at any time to check if results are ready. Should this window go blank please press the Shift+Refresh or Reload button on your browser.
- You may bookmark this page to view your results later if you wish.
Netscape users: Use Bookmark - Add Bookmark or (F11.5) Alt+ to bookmark this page.
IE users: Click on [Bookmark](#) to bookmark this page.
- Results are stored for 24 hours. Some big files will be deleted after 15 minutes.

Done Internet

Start [Navigation icons] [Address bar] [Status bar] 11:53AM

Help Index **ClustalW** Your results
General Help
Formats [Seq](#) [JalView](#) [JalView](#) [SUBMIT ANOTHER JOB](#)
Gaps
Matrix
References [Pairwise Branch](#)
ClustalW Help



glycine 7246327GGCATTTGGAGCAAACTCTGGACTTCATCAAAGATATAGAAATATTCATTC 7247

medicago
Lotus
arabidopsis
glycine 7246328AACTTTCTCAGGAACTGGTTCTCTCCGACTAGGTCGACACTTTCATATAAAGAAATAATTTT 7249
7246329AAGTCTTTCTGGAGCTGCTCTCTCCGACATGCTGCTGCTCTCTATATAAAGAAATAATTTT 7252
7246330AATGCTTTCTGGAGCTGCTCTCTCCGACTAGGTCGACACTTTCATATAAAGAAATAATTTT 7255
7246331AGGTTTATCTTGGGACTTGTGCATCTGCACCTTTTGGGGATATTCAGAAAGAAATTTTAA 7258

medicago
Lotus
arabidopsis
glycine 7246332TTGGATGAAAGTCTTGATATCAAATCTGACTGGGGTAAATTCAGAGAAATTTTAA 7261
7246333CARGGGCAAAAGTCTTGATATCATACTGCTGACTGGGGTAAATTCAGAAAGAAATTTTAA 7264
7246334TTGGATGAAAGTCTTGATATCATACTGCTGACTGGGGTAAATTCAGAAAGAAATTTTAA 7267
7246335TAAATCCCAAAATCTATATACCAGTCTGCTACTGGGGTAAATTCAGAAAGAAATTTTAA 7270

medicago
Lotus
arabidopsis
glycine 7246336ATCTTTAGATACCATGCTTTGAGTAATGATACTATGAACTTAAATAATCTTTTATAT 7273
7246337ATCTTAGAGTCCCATGCTTTAGAGTAATGATACTATGATGCTTAAACAAATTCAGAAAGAA 7276
7246338ATCTTCAAAATCTGCTGCTGCAAGATACTGCTACTATGATGCTTAAACAAATTCAGAAAG 7279
7246339ATCTTTAGATACCATGCTTTGAGGACTTTAGGACTATGCTTAAATAATTCAGAAAGAA 7282

medicago
Lotus
arabidopsis
glycine 7246340TTGAGGGCATGATAGAAGATATAAAATGGCTCCGGAAAGAAATTTTGTCTATTTTATG 7285
7246341TTGAGGGCATGTTAGAGATATAAATCCAGCTCCTGAAGGATCTTCTTCTACTCTTTATG 7288
7246342TTGAGGAAATGATAGACAGATATAAATGAAGCTCCGAAAGGATCTTTATCTTTCTTCCG 7291
7246343TTTCAGACTGATGATGACATATAAATGATCTCCAGATGGTCTCTCTCTCTTTCTTTGG 7294

medicago
Lotus
arabidopsis
glycine 7246344GATGTCACATAACCTATCTGGTATTGATCCACACCGAGAACTCTGGAAATAATACTG 7297
7246345GATGTCCTCATAAATCTATCTGGTATTGATCCACACCGAGAACTCTGGAAATAATACTG 7299
7246346TTGATTTCTCAACCTGATCTGGATTTGACCCACCGAGAACTCTGGAAATAATACTG 7302
7246347TTGATTTCTCATAAATCTATCTGGGATGATCTCTCAGAAGAAATTCAGAAAGATTTT 7305

medicago
Lotus
arabidopsis
glycine 7246348ATTTAAATTCAGAAAAGAACCACTTTTCATTTTTGATTTTCTTACAGATCTTTTGT 7308
7246349ATTTAAATTCAGCAAAAAGAACCACTTCATTTTTGGATTTCTTACAGATCTTTTGT 7311
7246350ATCTTCATCCAGAAAAGAACCACTATCCATTTTGGATTTCTTACAGATCTTTTGT 7314
7246351CCGAGATAAAGCTTAGGCTGATTTTTCTTTCTTGTGAATGGCTATCAGATCTTTTGT 7317

medicago
Lotus
arabidopsis
glycine 7246352GTGGAAGCTCTGATTAAGATGCCGCTCTCTGGAGATTTTTTAAATGACTGGATTAAG 7320
7246353GTGGAAGCTCTGATTAAGATGCCGCTCTCTGGAGATTTTTTAAATGACTGGATTAAG 7323
7246354GTGGAAGCTCTGATTAAGATGCCGCTCTCTGGAGATTTTTTAAATGACTGGATTAAG 7326
7246355GTGCTATCCAGAGAGATGCAAAAAGCATAAAGATTTTCTTGGAGATCTTTATTA 7329

medicago
Lotus
arabidopsis
glycine 7246356TTTTTGTAKCTCAGTATACAGTAAAAAACTGAGCCCTTTATCTTAAAGATTTTGAAGTA 7332
7246357TTCTTATAGCTCAGTATACAGTAAAAAACTGAGCCCTTTATCTTAAAGATTTTGAAGTA 7335
7246358TTTTTTGTAKCTCAGTATACAGTAAAAAACTTAKGCTTTTATCTTAAAGATTTTGAAGTA 7338
7246359TAGGATTTGCTCATCTATATGCAAAAAATATGGGACTGTATCTTGGAGGATCTGAT 7341

medicago
Lotus
arabidopsis
glycine 7246360TTAAATGTCAATTTCTCATCACCGAATCTGCAACAAGGCTAAAGAGCTAAATTTAAAGG 7344
7246361TCAATGTCAATTTCTCATCACCGAATCTGCTCAAGGCTAAAGAGCTAAATTTAAAGG 7347
7246362TCAATGTCTGTGCTCTCAGCTGATCTGCTACAAAGGCTCAAGAGCTTTGAAGAAAG 7350
7246363TGATGTTGCTTTTGGAGATGAGAAACAGCTGTGGCTGTAAMATTCATCTTGAATTA 7353

medicago
Lotus
arabidopsis
glycine 7246364TTGCTCGAATAATGTACTAATGCCACAGTTACGCGGGCTAGCATTTCTTCAATATAT 7356
7246365TTGCACGGCCAAATGTACTAATGCCACAGTTACGCGGGCTAGCATTTCTTCAATATAT 7359
7246366TTGCTCGGCTTGTACTTCAATGCCACAGTTCTATGGGGCGAGATCTTCTTCAATATAT 7362
7246367TTCTTAGACCCATGTACATTAACCAACTCTTCCATGGAGGACTTATATATTTCTATCT 7365

medicago
Lotus
arabidopsis
glycine 7246368TTGGACTCCAGCTCTCTTTGATGAATGGAAGACGAAATGAAATGATGCTTTGAAGTA 7368
7246369TTGGAATCCAGACTCTCTTCAAGATGGAAGACGAGATGGAAATGATGCTTGAAGTA 7371
7246370TCTGTGATTTAACTATGTTTCAGTGAATGAAAGACGAGATGAAATGATGCTTGAAGTA 7374
7246371TTGGACTCCAGACTCTCTTGAAGGATTTAGCTTAAAGAGTCAAGGCTATGCTGAGCCCT 7377

medicago
Lotus
arabidopsis
glycine 7246372TAAAACTGTTAGGACGGCTGTATGATAGTATTTCTTCAAAAGCAAAAGTGGAAAG 7380
7246373TAAAGAAATGTTAGACGAAAGCTATATGATAGTATTTCTTCAAAAGCAAAAGTGGAAAG 7383
7246374TAAAGACGGTTAGACGAAAGCTGTATGATAGCTCTTCTTCAAAAGCAAAAGTGGAAAG 7386
7246375TCTATGAAATGAGGACTACTACAGAAAACCTAGA---AAAGAGGCTTACTTTG 7389

medicago
Lotus
arabidopsis
glycine 7246376ATTGGTCATTCACTCAAGCAGATAGGCATGTTCTCATTTCACTGGCTTTAAACAAGGC 7392
7246377ATTGGTCATTCACTTAAGCAGATAGGCATGTTCTCATTTCACTGGCTTTAAACAAGAA 7395
7246378ACTGGCTGCTTCACTTAAGCAAAATGGCATGTTCTCTTCACTGGCTTTAAACAAGCT 7398
7246379CATGGGACGATAAATCAATCAGATTTCTGATGATGTTCTGTCAGAGGATGAGCTGAC 7401

```

      TTA      A      A      A
      AA  AA  AA  AA  AA  AA
      AT  AT  AT  CAAAT  ATCTATAA  AA  AA  A
      ATCTATA  A  AA  CGA  T  ATCTATA  AA  A
      *
      CAA  T  A  TCC  A  AT  AA
      AA  AA  AA  CAA  TCA  TT  A  AT  AA
      ATTA  T  C  CAAAT  A  TA  AT  A  A
      AT  AT  AA  AA  AA  AA  AA  AA
      *

```

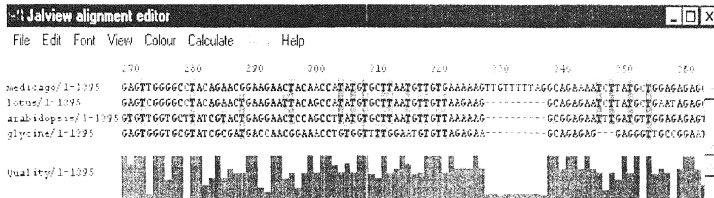
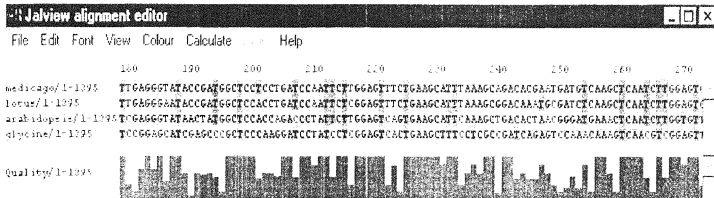
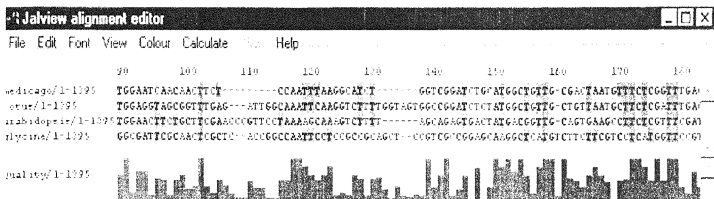
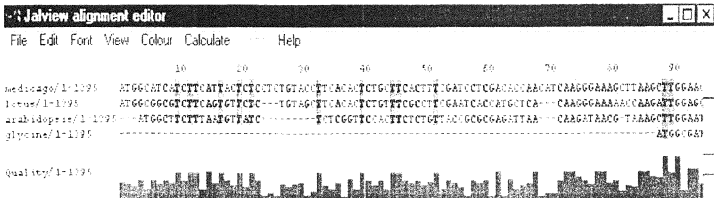
941166 716022-11800 dnd

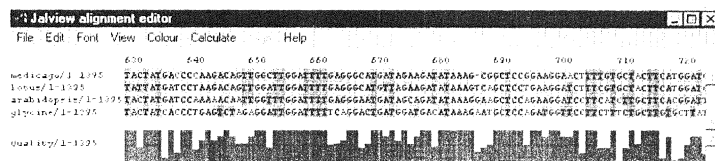
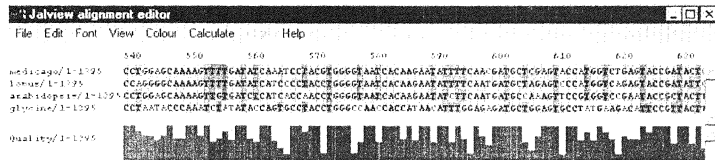
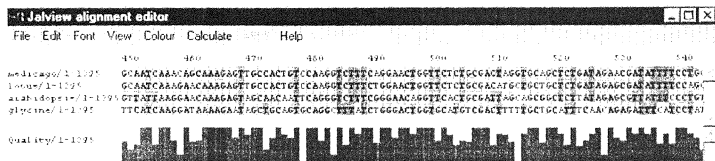
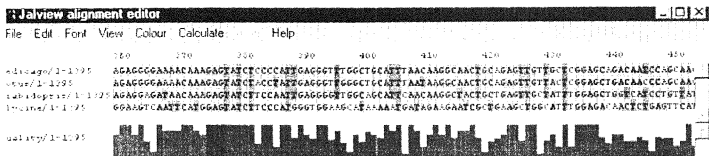
```

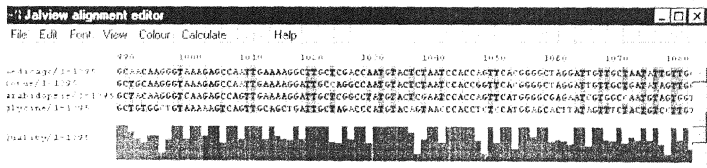
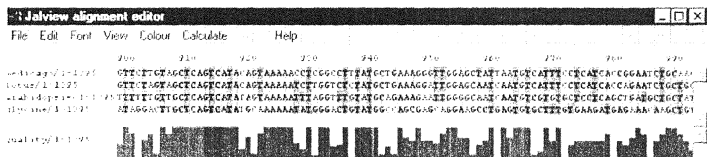
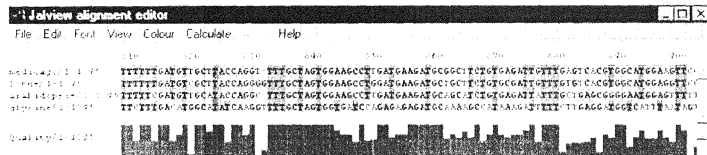
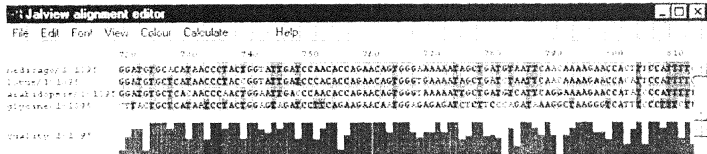
      1  0  H06
      /  788
      1  0  11
      (  )
      (  7)

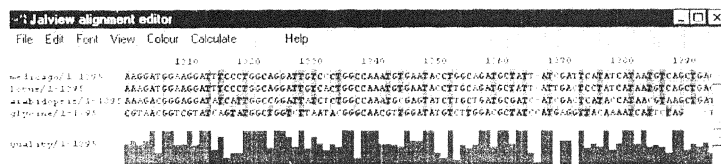
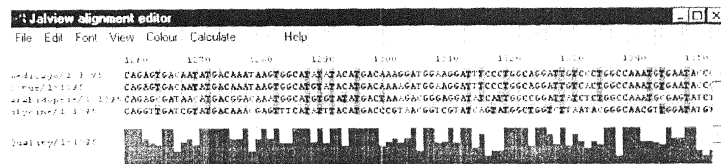
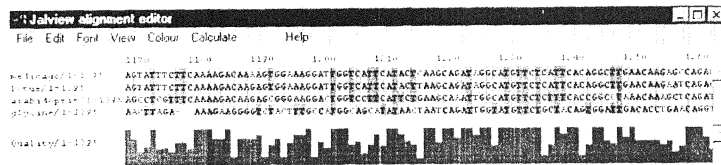
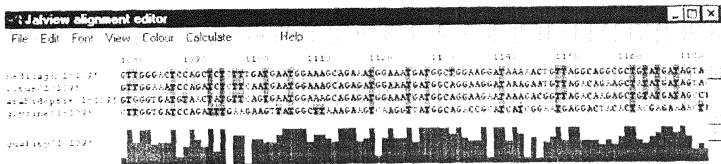
```

8) An alignment graph among Medicago, Arabidopsis, lotus and Glycine is obtained.

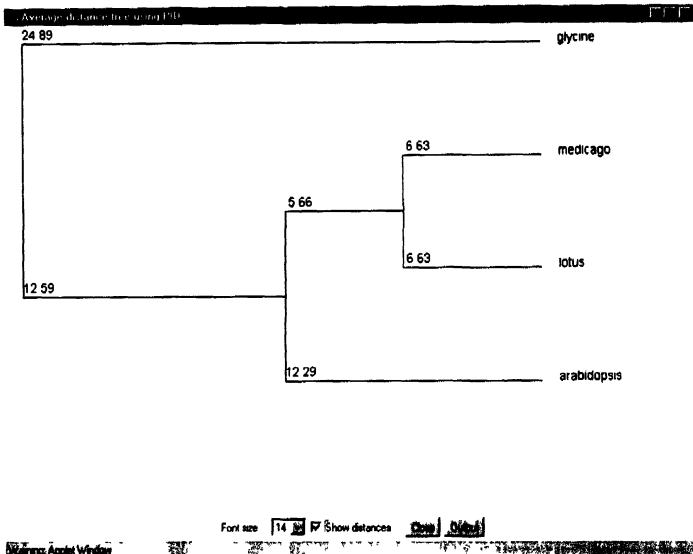








Construct the phylogenetic tree among *Medicago*, *Arabidopsis*, *Lotus* and *Glycine* with distances.



- Check the two species which having maximum alignment By tree Analysis.
- Take the species, which have maximum alignment.
- By Tree Analysis we have got that there is maximum alignment between Medicago and Lotus.

2.5 Design Primers For The Sequences Of Medicago:

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

2.5.1 SELECTION OF FIRST SET OF PRIMERS

- 1 Design left primer from one Exon of the sequence and right primer from another Exon, which is adjacent right position of the left exon 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 Medicago sequence has both Exons and introns. But Glycine sequence for this enzyme has only coding regions i.e.Exons.so we can't design primer from Glycine for sequencing of DNA but can be used as a marker.**
- 2 Take one Exon region sequence of the Medicago and then paste it in PRIMER3 and pick primers. From this primer output left primer only has to consider. Then design right primer from another Exon, which is adjacent right position of the left exon sequence of the Medicago. The parameters for both the primers should be same.**

Sequence Quality

Min Sequence Quality Min End Sequence Quality Sequence Quality Range Min Sequence Quality Range Max

Objective Function Penalty Weights for Primers

Tm Ls Gs
 Size Ls Gs
 GC% Ls Gs
 Self-Complementarity
 3'-Self-Complementarity
 #Ns
 Mismatches
 Sequence Quality
 End Sequence Quality
 Position Penalty
 End Stability

Objective Function Penalty Weights for Primer Pairs

Product Size Ls Gs
 Product Tm Ls Gs
 Tm Differences
 Any-Complementarity
 3'-Complementarity
 Hair Mismatches
 Primer Penalty Weight
 Hyb Oligo Penalty Weight

~~Hyb Oligo Penalty Weight~~

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo Excluded Region:

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size Min Opt Max
 Hyb Oligo Tm Min Opt Max
 Hyb Oligo GC% Min Opt Max
 Hyb Oligo Self-Complementarity Hyb Oligo Max 3'-Self-Complementarity
 Max #Ns Hyb Oligo Max Poly-X
 Hyb Oligo Mismatch Library Hyb Oligo Max Mismatch
 Hyb Oligo Min Sequence Quality
 Hyb Oligo Salt Concentration Hyb Oligo DNA Concentration

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo Tm Ls Gs
 Hyb Oligo Size Ls Gs
 Hyb Oligo GC% Ls Gs
 Hyb Oligo Self-Complementarity
 Hyb Oligo #Ns
 Hyb Oligo Mismatches
 Hyb Oligo Sequence Quality

~~Hyb Oligo Penalty Weight~~

Copyright Notice and Disclaimer

- **Select another exon region of sequence which is right adjacent to the left primer from the Medicago sequence to select the right primer.**

Primer3

enter one or more pick primers from a BLAST response

download
cautions

Note: Some sequences of interest may contain low-complexity regions which may be filtered out by the software. For more information, click on the "cautions" link. To use this tool, click on "pick primers" in the left navigation pane. [Help](#)

Sequence ID:

Start: End:

Strand:

pick	pick	pick
left	left/strand	right
primer	primer	primer
or	or	or
use	use	use
left	or	right
primer	use	primer
follow	edge	below

Accepted by the following:

Figure 2 requires primer3 primer set file. Use of primer3 with BLAST or other BLAST-like programs may require that primer3 user than the normal user.

Fig 2. Pick primer3 primer set file. Input primer set file. Output primer3 primer set file. Input primer set file. Output primer3 primer set file.

pick: Max: Min:

pick: Max: Min:

pick: Max: Min:

General Primer Picking Conditions

Target Size: Min: Max:

Primer Size: Min: Max:

Primer GC: Min: Max:

Primer GC2: Min: Max:

Primer GC3: Min: Max:

Primer GC4: Min: Max:

Primer GC5: Min: Max:

Primer GC6: Min: Max:

Primer GC7: Min: Max:

Primer GC8: Min: Max:

Primer GC9: Min: Max:

Primer GC10: Min: Max:

Primer GC11: Min: Max:

Primer GC12: Min: Max:

Primer GC13: Min: Max:

Primer GC14: Min: Max:

Primer GC15: Min: Max:

Primer GC16: Min: Max:

Primer GC17: Min: Max:

Primer GC18: Min: Max:

Primer GC19: Min: Max:

Primer GC20: Min: Max:

Primer GC21: Min: Max:

Primer GC22: Min: Max:

Primer GC23: Min: Max:

Primer GC24: Min: Max:

Primer GC25: Min: Max:

Primer GC26: Min: Max:

Primer GC27: Min: Max:

Primer GC28: Min: Max:

Primer GC29: Min: Max:

Primer GC30: Min: Max:

Internal Primer Check for overlap with

Other Per Sequence Inputs

Primer 1:

Primer 2:

Primer 3:

Primer 4:

Sequence Quality

Quality:

Start: End:

Sequence Quality

Min Sequence Quality Min End Sequence Quality Sequence Quality Range Min Sequence Quality Range Max

Objective Function Penalty Weights for Primers

In. Lt Ct
 Size Lt Ct
 GC% Lt Ct
 Self Complementary
 3' Self Complementary
 #N's
 Mismatches
 Sequence Quality
 End Sequence Quality
 Position Penalty
 End Stability

Objective Function Penalty Weights for Primer Pairs

Product Size Lt Ct
 Product Tm Lt Ct
 Tm Difference
 Any Complementarity
 3' Complementarity
 Pair Mismatches
 Primer Penalty Weight
 Hyb Oligo Penalty Weight
~~Hyb Oligo Penalty Weight~~

Hyb Oligo (Internal Oligo) Per-Sequence Inputs

Hyb Oligo Excluded Reason

Hyb Oligo (Internal Oligo) General Conditions

Hyb Oligo Size Min Opt Max
 Hyb Oligo Tm Min Opt Max
 Hyb Oligo GC% Min Opt Max
 Hyb Oligo Self Complementary Hyb Oligo Max 3' Self Complementary
 Max #N's Hyb Oligo Max Poly-X
 Hyb Oligo Mismatch Library Hyb Oligo Max Mismatch
 Hyb Oligo Min Sequence Quality
 Hyb Oligo Salt Concentration Hyb Oligo DNA Concentration
~~Hyb Oligo Salt Concentration~~

Objective Function Penalty Weights for Hyb Oligos (Internal Oligos)

Hyb Oligo Tm Lt Ct
 Hyb Oligo Size Lt Ct
 Hyb Oligo GC% Lt Ct
 Hyb Oligo Self Complementary
 Hyb Oligo #N's
 Hyb Oligo Mismatches
 Hyb Oligo Sequence Quality
~~Hyb Oligo Sequence Quality~~

Copyright Notice and Disclaimer

Note down the right primer from the above PRIMER3 output.

2.5.2 CALCULATION OF PRODUCT SIZE

a) Put the sequence of the Exon of the Medicago for the left primer in the primer output and by clicking pick primer we will get the left primer and right primer both. But we have to consider the left primer only.

For example:

Left primer start point is 6

End of the left primer is 109

So the bases from which left primer starts are 103.

b) Put another sequence of Exon of the Medicago for the right primer in the primer output and by clicking pick primer we will get the right primer and left primer both. but we have to consider right primer only.

Right primer end point is 105 bases.

Sequence size of the right primer is 110.

c) Introns between these two Exons are 92 bases.

Calculation the product sizes are as follows:

Left primer sequence size is 109

Left primer starting point is 6

Left primer total length is 103

Introns length is 92

Right primer sequence size is 110

Right primer ending point is 105

Total right primer length is 105

Product size is = Left primer total length (103)+ Introns length (92)+

Total right primer length (105)= 300

2.6 SELECTION OF SECOND SET OF PRIMERS

Repeat the same process as explained in the first set of primers. Adjust the parameters for getting the primers for PCR.

2.7 Results: For our study, we selected the different conserved nuclear, mitochondrial and chloroplast enzymes, which has the abundant sequence data, which is available from public databases (NCBI) for the different legumes. Out of these different enzymes for our project work we have taken one enzyme from each nuclear, chloroplast and mitochondrial areas from each legume. Then we find out the genetic relationship by constructing the phylogenetic tree among different legumes for all enzymes. It was observed that in case of mitochondrial enzyme *Medicago* and *Lotus* are phylogenetically closely related, In case of chloroplast enzyme *Glycine* and *Medicago* are closely related, In case of nuclear enzyme *Medicago* and *Pisum* are closely related. The results are summarized in the following tables.

id	enzymes	Crop name	accession no	exon region
1	Aspartate aminotransferase	medicago	L25335	1525-1599 1680-1721 2537-2731 2873-2911 3000-3089 3205-3313 3405-3514 3767-3910 4107-4273 4700-4964 5116-5247
		Glycine max	L40579	54-1337
		Arabidopsis	X91865	2237-2296 2389-2439 2563-2757 2832-2870 2955-3044 3142-3250 3335-3444 3652-3795 3888-4054 4193-4457 4552-4683
		Lotus	AF029898	119-1492
		medicago	X03931	741-814 1529-1568 1682-1785 1910-1958 2404-2510 2604-2691 2978 3106 3221-3295 3424-3477 3613-3650 3741-3900 4078-4230
		Arabidopsis	AB015045	1791-2078 2160-2263 2340-2388 2472-2578 2664-2751 2834-2962 3044-3118 3192-3443 3530-3590 3671-3810
		Pisum	U28925	1141-1302 1642-1681 1791-1894 2013-2061 21298-2304 2617-2704 2835-2963 3211-3285 3412-3465 3595-3631 3859-4018 4145-4561
		lotus	Y12859	10123-10409 10919-10958 11055-11158 11286-11334 11727-11833 11932-12019 12721-12849 13137-13211 13307-13361 13593-13629 13714-13873 13979-14131
		Glycine	AF091456	3013-3120 3579-3707
		3	Chitinase	medicago
Pisum	L37876			269-1243
trifolium	AJ011940			25-921
Glycine	AF335589			3338-3796 4265-4412 5215-5719
Arabidopsis	AF422179			884-1271 1490-1649 1751-2168

id	enzyme	legume name	accession no	exon region	left primer	left temp	right primer	right temp	product size
1	aspartate aminotransferase second set	medicago	L25335	3205-3313 3405-3514 3767-3910 4107-	cactgtccaaggtcttc	50.9 3	tatctctatcatgccctca	53.28	300
2	glutamine synthase second set	glycine	AF091456	3013-3120 3579-4557-4716 5036-5193	gtgtgatgcttacactcctg ct	59.7 1 atg	attccggatgaggaa	59.75	494
					gtgtgatgcttacactcctg ct	59.4 3 tt	ccaggcatcactctcca	60.07	694
					agtcctatgagaaacgatg gtg	59.9 8 ga	gctgcttatggtttccaaa	59.35	634
3	chitinase is	arabidops is	AF422179	884-1271 1490-1649	atcatttctggccttgggtg	59.9 3 ag	accgcgctccgtagtattc	60.15	693
					ggctatggagttgcaaca gg	60.6 6 gt	agctgagctcatcgtttg	60.02	621

2.8 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 legumes, 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 is 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 bhatramakki,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 opinion in Structural biology 8:321-330

Yuan, q.; Quackenbush,J.;Sultana ,R.;Pertea,m.;Salzberg,S.L and Buell, C.R.(2010). 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 microsatellite 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. STEMLER. 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.

KOLUKISAAGLU, H.U., S. MARX., C. WIEGMANN, S. HANELT and H. A. SCHNEIDER-POETSCH.1995. Divergence of the pytochrome 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.