NBU Journal of
Plant Sciences

Vol. 5, No.1 March 2011 p. 1-5
ISSN No. 0974 6927

Studying molecular evolution using tools of bioinformatics: an example from maize starch biosynthetic pathway

V Ghatak and PD Ghosh*
Cytogenetics and Plant Breeding Section, Plant Biotechnology Research Unit,
Department of Botany, University of Kalyani

Abstract
Various tools and techniques of bioinformatics are used to reveal the evolutionary pattern and chronology of different phylogenetic events in the evolutionary history of an organism. In maize, *shrunken, brittle, waxy* and *sugary* mutants all results from anomalies in the starch biosynthetic pathway.The genes responsible are *sh1,bt2,wx1 and su1* respectively. The present article explores and depicts an outline of computer application based methods adopted in a standard molecular phylogenetic analysis using sequence data of these gene products controlling different steps of maize starch biosynthetic pathway. Several such methods namely MSA,PSI-BLAST, Maximum Likelihood analysis and UPGMA analysis reveals that *shrunken1, waxy1 and sugary1* have evolutionary rates ranging in between 0.001-0.003 substitutions per site. Non-synonymous substitution (as polymorphism is analysed using protein sequences) substitution rate is much slower in *brittle2* (0.0006).The divergence times between the ancestors of *Sorghum* and maize-teosinte lineage has been calculated from these non-synonymous substitution rates. MEME analyses and distribution of motifs in these sequences and their homologues suggest changes (duplication and rearrangements) in the genetic material after *Sorghum*-maize split.

Keywords: MSA,PSI-BLAST, Maximum Likelihood, Neighbour Joining,, *Tajima's D value*, MEME

Sequence polymorphism analysis reveals evolutionary aspects imprinted within the sequences of genes and gene products. Phylogenetic studies using sequence data enables us to know the timing of events and pattern of changes in the ancestors of an organism as we cannot do experiments on evolutionary history. Various tools and techniques of bioinformatics are being used to reveal the evolutionary pattern and chronology of different phylogenetic events in the evolutionary history of an organism (Baxevanis and Ouellette, 2002). The present article explores and depicts an outline of these computer application based methods adopted in a standard molecular phylogenetic analysis using the gene products controlling different steps of starch biosynthetic pathway of *Z. mays* (Badenhuizen, 1965) ssp *mays*. (Linn.) as examples. These genes (namely *sh1,bt2,wx1 and su1*), determining the quantity\quality of starch stored in corn kernels, mutates to produce various starch deficient phenotypes including *shrunken* (Chourey and Nelson, 1976) ,*brittle* (Hannah *et al.*, 2001),*waxy* (Feng *et al.*, 2008; Shure *et al.*, 1983) and *sugary* (James *et al.*,1995) mutants. The products of these genes have been analyzed to bring out the nature of evolution in maize starch biosynthetic pathway. The work inspects all the important arenas of bioinformatics which can be utilized by a plant biologist while deducing origin and evolution of genes and biochemical pathways including the tools for sequence analysis and alignment (Jeanmougin *et al.*, 1998; Nei and Gojobori, 1986; Nei *et al.*, 2001) (Clustal X, Mega, Bioedit and Dot matrix analysis.), analysis of homology and protein profile (BLAST (Altschul *et al.*, 1990), MEME (Bailey and Elkan, 1994) and Sequence Logo study) and phylogenetic study (Nei and Gojobori, 1986; Nei *et al.*, 2001; Studier and Keppler, 1988) (NJ-UPGMA method and MAXIMUM LIKELIHOOD method).

*Corresponding author:
E-mail: pdgbot@yahoo.co.in

Materials and Methods

The protein sequences of interest are chosen from NCBI database at *http://ncbi.nlm.nih.*The mostly used format in phylogenetic analysis is the FASTA format (Thompson *et al.*, 1997) which has been used here as well. While measuring similarities between sequences, the principles followed are that, replacement of an amino acid by another one having similar size or physicochemical property is more common than replacement by one having dissimilar size\property and amino acids are more probable to be deleted in groups rather than the individual deletion of same number of bases. According to standardized scoring matrices of amino acid replacements, scoring is done. Scoring matrices used in the present study are mainly PAM**250** (which corresponds to an overall sequence identity of **20%**) and BLOSUM**62** (in which the threshold value of sequence identity is 62%). 'Clustal X 2.0' and 'BioEdit' software are used for MSA (Jeanmougin *et al.*, 1998).

Homology has been searched primarily by using BLAST - the online tool available at NCBI (Altschul *et al.*, 1990).Using the PSI BLAST programme patterns of similarity in same position of the query is searched against protein sequence database and then the matches of that initial pattern are extended by iterated search and

gradual refinement at each cycle Each entry in the database is checked for short regions which are similar to a particular pattern .i.e. contiguous region in the query sequence. Once a match is found, the programme extends it as possible and carries on further search cycles to find more matches of that extended pattern. BLAST performs a gapped alignment between the query sequence and the database sequence. Bit score (S') is derived from the raw alignment score S in which the statistical properties of the scoring system used are taken into account. By normalizing a raw score, using the formula:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

a bit score S' is attained. E value (number of hits one can "expect" to see by chance when searching a database of a particular size) below 0.05 is considered to be significant. The E values are interpreted in the present study according to the following guideline: E < 0.02 implies homology of the sequence with the query ; E = 0.02 – 0.05 indicates possible homology and E > 0.05 suggests chance similarity. In the present work, every tree presented has been subjected to 500-1000 bootstrap verifications done by randomly stripping off whole columns from the original MSA prepared from the sequences under study. The number of changes between pairs in a group of sequences including the maize protein sequence of interest has been subjected to two different types of distance analysis ,namely, Neighbour Joining and UPGMA (Unweighted Pair Group Method with Arithmatic Mean) using MEGA 4 (Tamura *et al.*, 2007). Prior to phylogenetic tree building,the sequences under study has been subjected to Tajima's neutrality test and Tajima's relative rate test to know the nature of evolution (Tajima 1989; Tajima 1993). Using Tajima's test of Neutrality, the number of segregating sites (*S*) and the average pairwise differences (π) are compared to detect the nature of selection in sequences under study. The difference between these two estimates has been depicted as *Tajima's D value*. Under neutral expectations, both π and *S* gives similar estimates. So when Tajima's D is approximately zero, neutrality of the locus has been predicted. A positive D value suggested balancing selection and a negative D value depicted the operation of purifying selection in the loci under study.Tajima's relative rate test has been used to measure uniformity of evolutionary rates among maize loci (under study) and its homologues in related organisms (identified by BLAST)with respect to an outgroup. Three sequences, 1, 2 and 3 has been considered where sequence 3 is the outgroup. If n_{ijk} is the observed number of sites in which sequences 1, 2 and 3 have nucleotides i, j and k , under the molecular clock hypothesis, E (n_{ij0}) = E (n_{ji0}), that is , when the evolutionary rates are same, the expected number of sites showing unique residue in the first sequence (n_{ij0}) is same as the expected number of sites showing unique residue in the second sequence (n_{ji0}). The null hypothesis is that these two values are same. The equality is tested from the chi-square value at 1 degree of freedom utilizing the formula:

$$Chi\text{-}square = [E(n_{ij0}) - E(n_{ji0})]^2 / [E(n_{ij0}) + E(n_{ji0})]$$

When this hypothesis is rejected, the molecular clock hypothesis has also been rejected for the studied set of sequences.

Neighbor-joining trees are drawn using distances between pairs of taxa rather than individual characters. As this method allows different rate of evolution along the different branches of the tree, it has limitations when a molecular clock is operative in the sequences under study (as known from Tajima's relative rate test). NJ trees depicted only the pattern of evolution of the sequences under study. The evolutionary rates and divergence times are calculated using UPGMA method which does not consider differential rates of evolution along the different branches of the evolutionary tree. Distance from nodes are calculated by arithmetic mean enabling the study of evolutionary rates and divergence times. However, this method has been applied only if the operation of molecular clock is established in the sequences under study, that is, UPGMA analysis is done only when the gene shows similar evolutionary substitution rate in all organisms. This uniformity of evolutionary rates has been determined by Tajima's relative rate test. For the calibration of UPGMA trees in this work, a divergence time of 50 million years between *Oryza* and the *Sorghum* – maize lineage has been used as calculated by Wolfe (1989). This divergence time has been used customarily by many researchers while working out the phylogeny of *Zea mays*.

Common motifs in these related sequences has been identified by using the MEME algorithm (Bailey and Elkan, 1994). MEME stands for '*Multiple Expectation Maximization for Motif Elicitation*'. The programme locates ungapped patterns in a series of protein sequence and display the location and sequence logo of these motifs. In the present work the online meme server at *www.meme.nchr.net* has been used for MEME analysis. MEME displays the occurrences (sites) of the motif which are shown aligned with each other. The position of these motifs are also shown. The motifs are listed in order of increasing statistical significance (p-value) which gives the probability of a random string (generated from the background letter frequencies) having the same match score or higher. The occurrences of the motif in the training set sequences are shown with block diagrams. One diagram is printed for each sequence showing all the occurrences of the motif in that sequence. The sequences are sorted by the lowest p-value among all occurrences of the motif in a given sequence.

Results and Discussion

The shrunken mutants in maize show cofaplete or nearly complete absence of *sucrose synthase* activity in the storage tissue and mutants show 'shrunken' phenotype of the kernels. The *shrunken1* gene product is a sucrose synthase which converts sucrose to UDP-glucose. The protein consists of 357 residues (gi 270314578, gb ACZ74304.1). MSA and Distance matrix shows closeness of maize *sh1* with its *Saccharum* and *Sorghum*

orthologues to be 0.022 and 0.014 respectively suggesting its comparative closeness with *Sorghum*. BLAST p suggests possible homology of maize *sh1* with *sucrose synthases* of monocots like *Sorghum,Hordeum, Bambusa, Triticum* and *Oryza* Sequence polymorphism and Neighbour Joining analysis suggests that maize, *Saccharum* and *Sorghum* derived the *sh1* gene from a common ancestor from which maize diverged earlier. Tajima's relative rate test (Chi-square = 1.00 with 1 degree of freedom) suggests that evolutionary rate of maize sh1 follows a molecular clock and the value of Tajima's D is 0.25, suggesting 'near neutrality' (Ohta, 1992) of this locus. Based on the assumption of a molecular clock, UPGMA tree is drawn with 500 bootstrap replicates using MEGA 4 (Figure 1). *Arabidopsis sus5* gene product ($E = 3e^{-04}$) is used as an outgroup. A divergence time of 29.6 mya between the ancestors of *Triticum-Hordeum* and *Sorghum*-maize lineages is predicted by UPGMA analysis in an evolutionary rate of 0.003 substitutions per site. Three motifs having lengths of 50-60 residues are identified using MEME. The sequence under study (*sh1* gene product of Zea *mays*) show unique distribution of all these three motifs with an increase in length suggesting duplication and rearrangements in this locus in course of origin of maize.

Maize *brittle2* (*bt2*) gene encodes the small subunit of ADP-glucose pyrophosphorylase (*AGPase*), which converts ADP-glucose into glucose-1-phosphate (the substrate for starch synthesis). Mutants are characterized by hard, brittle kernels which is the result of an alteration in starch composition and property.The protein is composed of 474 residues (gi 23664297 ,gb|AAN39300.1).NJ tree suggests the closeness of maize *bt2* with the *AGPase* small subunit of *Sorghum* and teosinte (*Z.mays* ssp. *parviglumis*).The value of Tajima's D is (0.24) suggests that maize *bt2* and *sh1* genes have undergone evolution in similar pattern(substitution rate is uniform and minimal operation of selection). However, the evolutionary rate is much slower in *bt2* (0.0006 substitutions per site in contrast to 0.003 in *sh1*).At this uniform evolutionary rate at the *bt2* locus as per the result of Tajima's relative rate test (chi square = 2 at 1degree of freedom suggesting the null hypothesis of operation of molecular clock in this locus to be true), *Oryza –Hordeum* shows a divergence time of 32.8 mya and *Sorghum* diverged from the maize-teosinte ancestor nearly 18.8 mya.ago. This divergence time between the progenitors of *Sorghum* and maize is a bit later than that calculated from the evolutionary rates in *su1* loci (18.3 mya) which can be explained by the slower evolutionary rate of maize *bt2* locus. MEME analysis is done to find out three motifs in the sequences under study. All of these three motifs in maize and teosinte are larger in length compared to its homologues including *Sorghum* suggesting changes in *bt2* gene after *Sorghum*-maize split.

Maize *waxy1* (*wx1*) gene encodes granule bound starch synthase which is essential for appropriate amylose to amylopectin ratio in the maize kernels. In *waxy* mutants (*wx1*), the amylose synthesized by soluble starch synthase isoenzymes is transformed into amylopectin by starch branching enzyme. Only amylopectin accumulates making the mutant kernels waxy (sticky).The *wx1* gene product is composed of 271 residues (gi|15055000|) and shows homology with the granule bound starch synthases of *Sorghum* and teosinte (*Z.mays*.subsp.*mexicana*). Tajima's neutrality test shows D= -2.1.This negative value of Tajima's D suggests that maize *wx1* locus has been subjected to purifying selection. Tajima's relative rate test (Chi-square= 1.00 with 1 degree of freedom) suggests that evolutionary rate follows a molecular clock in maize *wx1*.This rate is 0.001 substitutions per site.At this uniform evolutionary rate, UPGMA analysis suggests that the ancestor of maize and teosinte diverged from the *Coix-Sorghum* line 18.38 million years ago. Divergence time of *Sorghum* and *Coix* is 12.01 million years. MEME analysis is done to find out motifs in the *granule bound starch synthases* under study.These motifs in maize and teosinte (*Z.mays*.subsp. *mexicana*) are also larger in length compared to the homologous proteins in other plants (this is exactly what found by MEME analysis of *sh1* and *bt2 proteins*) including *Sorghum*. However in *Coix*, these motifs are in parity with maize and teosinte in lieu of length and distribution.

Maize *sugary1* (*su1*) gene encodes starch *debranching enzyme* which transforms amylopectin into amylose in kernels. The protein is composed of 357 residues. BLASTp and Neighbour Joining analysis suggests its possible homology with the isoamylases of *Sorghum,Triticum* and *Hordeum*.Tajima's D is 0.39, suggesting 'near neutrality'(Ohta, 1992; Ohta 1996) of the locus. Tajima's relative rate test (Chi-square= 1.00 with 1 degree of freedom) suggests that evolutionary rate follows a molecular clock in maize *su1*.This rate is 0.002 substitutions per site. The divergence time of maize and *Sorghum* is once again suggested to be 18.3 mya by distance analysis at *su1* locus which evolves at a rate of 0.002 substitutions per site. In all the proteins of monocots studied (including maize *su1* gene product), 3 copies of the motif has been found in contrast to the 2 copies present in the *Arabidopsis* protein. Moreover the distribution of these copies in maize *su1* gene product is intermediate to that found in *Triticum* and *Sorghum*.

To summarize , the evolutionary pattern and protein sequence polymorphism of maize in the loci involved in starch biosynthesis has been analysed in the present work.Among the genes determining quality and quantity of maize kernel starch, *shrunken1, waxy1 and sugary1* shows evolutionary rates ranging in between 0.001-0.003 substitutions per site.Non-synonymous substitution (as polymorphism is analysed on protein sequences) substitution rate is much slower in *brittle2* (0.0006). Among them, *wx1* bears evidence of strong purifying selection.Phylogenetic analysis of all these sequences suggests that the maize-teosinte lineage diverged from *Sorghum* nearly 18.3 million years ago and divergence of *Coix* and *Sorghum* occurred later. Distribution of motifs (known from MEME analysis)in all these proteins suggest changes (duplication and

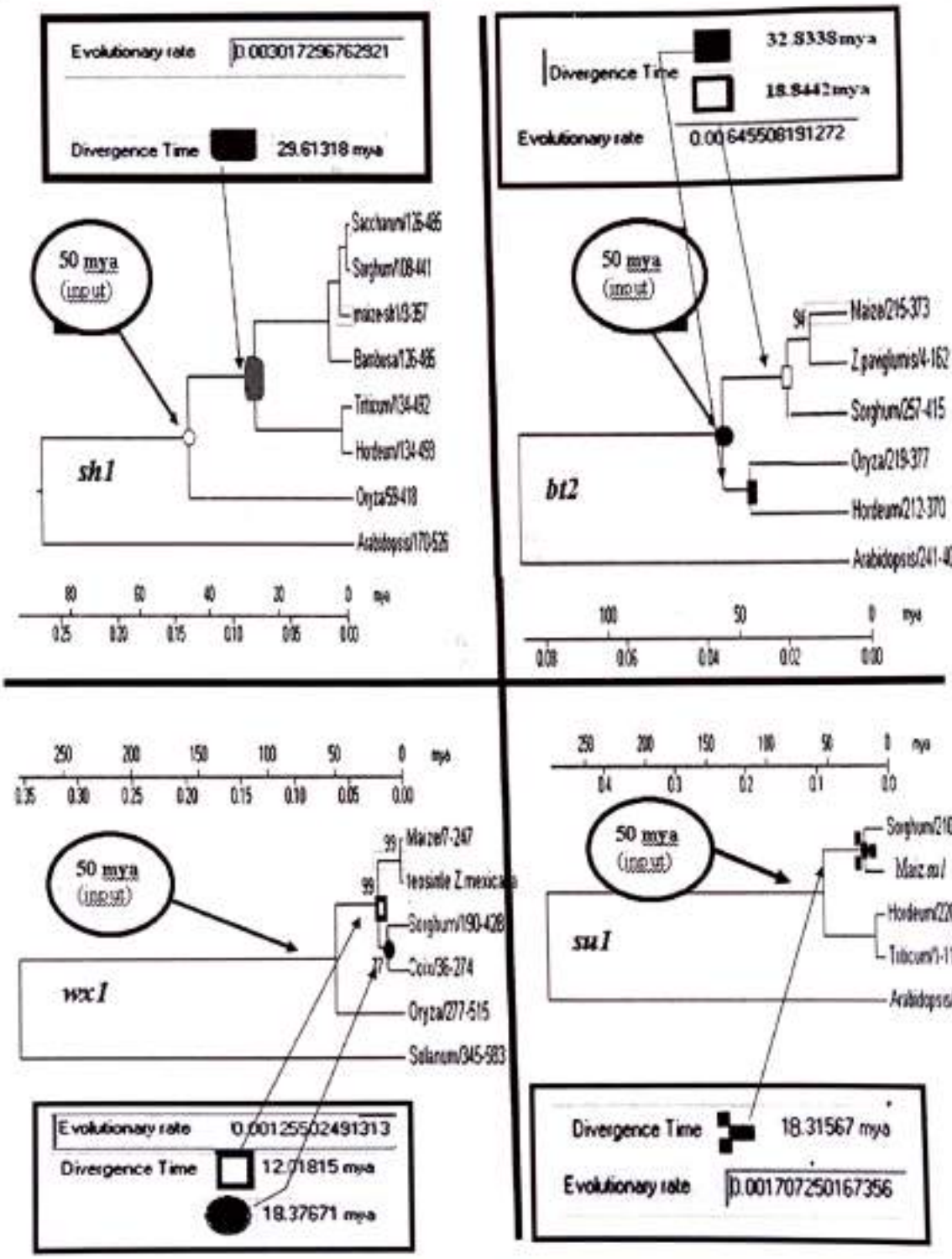

Figure 1: UPGMA trees showing evolutionary rates and divergence times of the maize loci involved in starch biosynthesis

rearrangements) in the genetic material after *Sorghum*-maize split.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol. Biol.* **215** (3): 403–410.

Badenhuizen NP. (1965). Occurrence and development of starch in plants.Starch: Chemistry and Technology, Vol. I. R.L. Whistler and E.F. Paschall eds.Academic Press, New York.

Bailey TL, Elkan C.(1994).Fitting a mixture model by expectation maximization to discover motifs in biopolymers.*Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*.28-36.

Baxevanis AD, Ouellette BFF (2002). Bioinformatics, A Practical Guide to the Analysis of Genes and proteins. 2nd edition; A Jhon Willy & Sons INC,NY.

Bromham L,Penny D.(2003).The Modern Molecular Clock.*Nature*.4:216-224.

Chourey PS., Nelson OE.(1976). Enzymatic deficiency conditioned by shrunken 1 mutations in maize. *Biochem. Genet.* **14**: 1041–1055.

Feng ZL, Liu J, Fu FL, Li WC.(2008). Molecular Mechanism of Sweet and Waxy in Maize. *Int. Jour. Pl Breed. Genet.***2 (2)**:93-100.

Hannah LC, Shaw JR, Giroux JR, Reyss A, Prioul JL, Bae JM,Lee JY.(2001). Maize genes encoding the small subunit of ADP-glucose pyrophosphorylase. *Pl. Physiol.* **127(1)**:173-83.

James, M.G., Robertson, D.S., and Myers, A.M.(1995). Characterization of the maize gene *Sugary1*, a determinant of starch composition in kernels. *Pl. Cell*. 7:417–429

Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J.(1998).Multiple sequence alignment with Clustal X. *Trends Biochem Sci*. **23**: 403-5.

Nei M & Gojobori T .(1986).Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. **3**:418-426.

Nei M,Ping Xu,Glazko G.(2001).Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms *PNAS* **98 (5)**: 2497-2502.

Ohta T .(1992). The nearly neutral theory of molecular evolution.Ann.Rev. EcoSyst.23: 263–286.

Ohta T. (1996).The neutralist-selectionist debate: The current significance and understanding of neutral and nearly neutral theories. *BioEssays* .**18(8)**.673–677

Shure M., Wessler S., Fedoroff N.(1983). Molecular identification and isolation of the waxy locus in maize. *Cell*.**35**:225–233.

Studier JA, Keppler KJ. (1988). A note on the Neighbor-Joining algorithm of Saitou and Nei. *Mol Biol Evol* .**5 (6)**: 729-731.

Tajima F.(1989).Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genet*. **123**: 585-595.

Tajima F.(1993).Simple methods for testing molecular evolutionary clock hypothesis.*Genet*.**135**:599-607.

Tamura K,Dudley J,Masatoshi Nei and Sudhir Kumar (2007).MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0.*Mol. Biol. Evol*.**24 (8)**:1596-1599.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997).The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acid Res*. **25**:4876-4882.

Wolfe K.(1989).Date of monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci*. **86**:6201-6205.