UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMÁTICA E EVOLUÇÃO

FILOGENIA MOLECULAR DAS ENZIMAS ISOCITRATO LIASE E MALATO SINTASE E SUA EVOLUÇÃO EM *Viridiplantae*

RICARDO VICTOR MACHADO DE ALMEIDA

Dissertação de Mestrado
Natal/RN, agosto de 2014

RICARDO VICTOR MACHADO DE ALMEIDA

Filogenia molecular das enzimas Isocitrato liase e Malato sintase e sua evolução em *Viridiplantae*

Dissertação apresentada ao programa de Pós-graduação em Sistemática e Evolução da Universidade Federal do Rio Grande do Norte, como requisito parcial para obtenção do título de Mestre.

Orientador: Prof. Dr. João Paulo Matos Santos Lima

Natal/RN
2014

Catalogação da Publicação na Fonte
Universidade Federal do Rio Grande do Norte - UFRN

Almeida, Ricardo Victor Machado de.
Filogenia molecular das enzimas isocitrato liase e malato sintase e sua evolução em Viridiplanta / Ricardo Victor Machado de Almeida. - Natal, 2014.
77f: il.

Orientador: Prof. Dr. João Paulo Matos Santos Lima.
Dissertação (Mestrado) - Universidade Federal do Rio Grande do Norte. Centro de Biociências. Programa de Pós-Graduação em Sistemática e Evolução.

1. Ciclo do Glioxilato - Dissertação. 2. Metabolismo - Dissertação. 3. Seleção Positiva - Dissertação. I. Lima, João Paulo Matos Santos. II. Universidade Federal do Rio Grande do Norte. III. Título.

RN/UF/BSE01 CDU 577.121

RICARDO VICTOR MACHADO DE ALMEIDA

FILOGENIA MOLECULAR DAS ENZIMAS ISOCITRATO LIASE E MALATO SINTASE E SUA EVOLUÇÃO EM *VIRIDIPLANTAE*

Dissertação apresentada ao Programa de Pós-graduação em Sistemática e Evolução da Universidade Federal do Rio Grande do Norte como requisito parcial para obtenção do título de Mestre em Sistemática e Evolução.

Área de concentração: Sistemática e Evolução.

Aprovada em 28/08/2014.

BANCA EXAMINADORA:

_____

Dr. João Paulo Matos Santos Lima
Universidade Federal do Rio Grande do Norte
(Orientador)

_____

Dr. Daniel Carlos Ferreira Lanza
Universidade Federal do Rio Grande do Norte

_____

Dr. Rodrigo Maranguape Silva da Cunha
Universidade Estadual Vale do Acaraú - CE

“The hardest thing to see is what is in front of your eyes.”

(Goethe)

A todos aqueles que nunca desistiram.

Dedico

AGRADECIMENTOS

Aos meus pais, pela sua excedente dedicação em fazerem aquilo que acham o mais correto para minha formação enquanto pessoa.

Ao meu orientador Prof. João Paulo Matos Santos Lima, por todos os ensinamentos, conselhos, experiências, paciência, citações, momentos de descontração, mas principalmente de dedicação e amizade, como exemplo de cientista. Sob sua orientação tenho certeza que nenhum uma conversa foi desperdiçada.

Aos coordenadores do Programa de Pós-Graduação em Sistemática e Evolução da Universidade Federal do Rio Grande do Norte, Profs. Dr. Iuri Baseia e Dr. Bruno Goto pela imensurável paciência.

Ao Prof. Dr. João Felipe de Souza Filho, por abrigar e partilhar pacientemente, mesmo as vezes tirando seu sossego (risos), o espaço que foi definitivamente essencial para a realização desta pesquisa.

A Profª. Dra. Juliana Espada Lichston, pelo voto de confiança em receber em seu laboratório um aluno do 2º semestre de biologia e confiar ser sua primeira orientadora dentro da instituição.

A Profª. Dra. Maria Iracema Bezerra Loiola, ex-curadora do herbário da UFRN, por demonstrar dentro da universidade o que é ser eficiente na rotina de ensino, pesquisa e extensão, sem nunca perder o bom humor.

Aos professores do Departamento de Bioquímica, pela disposição em ajudar e contribuindo de alguma forma com sugestões, especialmente o Prof. Dr. Daniel Lanza.

As minhas “quase irmãs”, Ana Paula C. de Pontes e Rafaella A. S. do Nascimento por serem as pessoas tão especiais e essenciais na minha vida, em que numa simples oração não há como descrever sua importância.

A meus primos Elisa Paiva e Guilherme Paiva por serem os irmãos que nunca tive e fazerem o possível e impossível para me apoiarem nesta etapa da minha vida.

Aos meus amigos e colegas de faculdade, que entre tantos e tantos merecem destaque os “Batatistas”: Daniel Chaves, João Felipe, Emanuel Sousa (Manel), Diego Fernandes, Rogério, Helaine (Cris), Mariana Dias, Priscilla França, Raysa (ysa), Rafael e Anthonieta. Obrigado por me ensinarem o significado da celebre frase “*O apresso não tem preço*”.

Aos meus companheiros de laboratório e amigos, que estão ou estiveram ao meu lado sempre que preciso: Ivanice Bezerra, João Paulo de Freitas, Érika Pinheiro, Taffarel Torres, Diego Gomes, Juliana (Jú), Delanno (agregado), Diego (rosa), Flávio, Maria Gabriela, Rômulo, Diego (Bob), Heverton, Dáfiny, Clara Dantas, Júlio e Suellen (Rebeca).

Aos demais membros da minha família, principalmente representada na pessoa da minha Tia, Maria Júlia de Paiva, graças a qual me ensinou desde pequeno o valor e importância na aquisição do conhecimento cultural e científico.

A CAPES / CNPq que concederam o apoio e permitiram a realização deste trabalho.

E, a todos que não mencionei aqui mas que de alguma forma participaram da realização deste trabalho.

RESUMO

O metabolismo vegetal é composto por uma complexa rede de eventos físicos e químicos que resultam na fotossíntese, respiração, e na síntese e degradação de compostos orgânicos. Isto só é possível graças aos diferentes tipos de respostas a inúmeras variações ambientais que um vegetal pode estar sujeito, adquiridas ao longo da evolução, levando também a conquistas de novos ambientes. O ciclo do glioxilato é uma via metabólica localizada nos glioxissomos de plantas, que possui papel único no estabelecimento das plântulas. Considerado como uma variação do ciclo do ácido cítrico esta via utiliza uma molécula de acetil-Coenzima A, oriunda da beta-oxidação de lipídios para sintetizar compostos que são utilizados na síntese de carboidratos. As enzimas Malato sintase (MLS) e Isocitrato liase (ICL) são exclusivas deste ciclo e essenciais na regulação da biossíntese de carboidratos. Devido à ausência das etapas de descarboxilação, como fatores limitantes da velocidade, estudos mais detalhados da filogenia e evolução molecular dessas proteínas permite o esclarecimento dos efeitos da presença desta rota nos processos evolutivos envolvidos em espécies vegetais. Portanto, o objetivo deste trabalho foi estudar a relação entre a evolução molecular das enzimas Isocitrato liase e Malato sintase e sua filogenia, nas plantas verdes (*Viridiplantae*). Para isso, foram utilizadas sequências de aminoácidos e nucleotídeos dos genes, a partir de repositórios online como o *Genbank* e *Uniprot*. As sequências foram alinhadas e, em seguida, submetidos à análise estatística dos modelos de melhor ajuste de substituição. A filogenia foi reconstruída por métodos de distância (*Neighbor-joining*) e métodos discretos (Máxima Verossimilhança, Máxima Parcimônia e Análise Bayesiana). O reconhecimento de padrões estruturais na evolução das enzimas foi feito por predição e modelagem por homologia das estruturas das sequências das proteínas obtidas. Com base nas análises comparativas entre modelos *in silico*, das enzimas, e partir dos resultados de inferência filogenética, ambas as enzimas apresentam um padrão de conservação relativamente elevado em sua estrutura e geram topologias condizentes com dois processos de seleção e especialização dos seus respectivos genes. Deste modo, confirmando a relevância em se realizar novos estudos para se elucidar o metabolismo vegetal sob uma perspectiva evolutiva das relações entre os genes e a expressão de suas enzimas.

Palavras-chave: Ciclo do Glioxilato; *IcL; MLS*; Seleção Positiva; Metabolismo.

ABSTRACT

The plant metabolism consists of a complex network of physical and chemical events resulting in photosynthesis, respiration, synthesis and degradation of organic compounds. This is only possible due to the different kinds of responses to many environmental variations that a plant could be subject through evolution, leading also to conquering new surroundings. The glyoxylate cycle is a metabolic pathway found in glyoxysomes plant, which has unique role in the seedling establishment. Considered as a variation of the citric acid cycle, it uses an acetyl coenzyme A molecule, derived from lipids beta-oxidation to synthesize compounds which are used in carbohydrate synthesis. The Malate synthase (MLS) and Isocitrate lyase (ICL) enzyme of this cycle are unique and essential in regulating the biosynthesis of carbohydrates. Because of the absence of decarboxylation steps as rate-limiting steps, detailed studies of molecular phylogeny and evolution of these proteins enables the elucidation of the effects of this route presence in the evolutionary processes involved in their distribution across the genome from different plant species. Therefore, the aim of this study was to establish a relationship between the molecular evolution of the characteristics of enzymes from the glyoxylate cycle (isocitrate lyase and malate synthase) and their molecular phylogeny, among green plants (*Viridiplantae*). For this, amino acid and nucleotide sequences were used, from online repositories as UniProt and Genbank. Sequences were aligned and then subjected to an analysis of the best-fit substitution models. The phylogeny was rebuilt by distance methods (neighbor-joining) and discrete methods (maximum likelihood, maximum parsimony and Bayesian analysis). The identification of structural patterns in the evolution of the enzymes was made through homology modeling and structure prediction from protein sequences. Based on comparative analyzes of *in silico* models and from the results of phylogenetic inferences, both enzymes show significant structure conservation and their topologies in agreement with two processes of selection and specialization of the genes. Thus, confirming the relevance of new studies to elucidate the plant metabolism from an evolutionary perspective.

Keywords: Glyoxilate cyclo; *IcL; MLS;* Positive selection; metabolism.

ÍNDICE DE FIGURAS

INDICE DE TABELAS

LISTA DE ABREVIAÇÕES / SIGLAS

Acetil-CoA ..... Acetil-Coenzima A
ACO ..... Aconitase
APG III ..... *Angiosperm Phylogeny Group III*
ASN.1 ..... *Abstract Syntax Notation One*
BLAST ..... *Basic Local Alignment Search Tool*
BLOSUM ..... *BLOcks of Amino Acid SUbstitution Matrix*
CAT .....Catalase
CSY ..... Citrato sintase
DAMBE ..... *Data Analysis and Molecular Biology and Evolution*
GTR ..... *General time-reversible*
HTTP ..... *Hypertext Transfer Protocol*
ICL ..... Isocitrato liase
MAFFT ..... *Multiple Alignment using Fast Fourier Transform*
MDH ..... Malato Desidrogenase
ME ..... *Minimum Evolution*
MEGA .....*Molecular Evolutionary Genetics Analysis*
ML ..... Máxima Verossimilhança
MLS ..... Malato Sintase
MP ..... Máxima Parcimônia
NCBI..... *National Center for Biotechnology Information*
NAD .....*Nicotinamide Adenine Dinucleotide*
NADH .....*Nicotinamide Adenine Dinucleótideo Reduced*
NJ ..... *Neighbor-joining*
PAM ..... *Point Accepted Mutation Matrix*
PEP .....Fosfoenolpiruvato
PP ..... Probabilidade *a posteriori*
TCA .....*Tricarboxylic Acid Cycle*
WGD .....*Whole Genome Shotgun*

SUMÁRIO

1. INTRODUÇÃO

1.1 Padrões e Processos Evolutivos em Plantas

As plantas que atualmente existem na Terra são o resultado de um longo e contínuo processo de seleção natural. Para todo e qualquer ser vivo, a seleção natural é o elemento regulador no estabelecimento e sucesso reprodutivo de uma espécie no meio ambiente (Darwin, 1859). O que não é diferente para os vegetais.

Sua origem e evolução remonta as diversas transições críticas de ambiente a que este grupo passou ao longo de gerações, tendo início a colonização do meio terrestre com as embriófitas, até alcançarem o auge da independência reprodutiva da água com as angiospermas (Lewis; McCourt, 2004; Judd *et al*., 2009; Niklas; Kutschera, 2009). Assim, de forma gradual, as plantas que originalmente habitavam os ambientes úmidos, se estabeleceram em *"terra firme"*, desenvolvendo inúmeras estruturas e mecanismos fisiológicos para sobreviver nesse meio (McCourt, 2004).

As angiospermas, ou plantas com flores, são tratadas como o maior grupo dentre as eudicotiledôneas, com mais de 300.000 espécies vivas (Niklas, 1997; Soltis & Soltis, 2004) distribuídas em cerca de 450 famílias (Soltis; Soltis, 2004) com representantes na maioria dos habitats, e cerca de 90 milhões de anos (Qiu *et al*, 1999). Diferentes hipóteses têm sido elaboradas para justificar o grande número de espécies de angiospermas em comparação com as demais plantas terrestres ou aquáticas, além de explicar o porquê são tão comuns ecologicamente (Doyle, 1978; Stebbins, 1981, 1992; Crepet, 2008; Grimaldi, 1999; Magallón *et al*, 1999; Verdú, 2002; Fenster *et al*, 2004; Grimaldi; Engel, 2005; Friis *et al*, 2006; Bomblies; Weigel, 2007). A maioria delas estão associadas a um dos seguintes campos: (1) atributos vegetativos, ou seja, sua morfologia organográfica e anatomia diversificada, taxas de crescimento rápido, juntamente com a alta condutividade dos vasos e sua capacidade de extensa plasticidade fenotípica, (2) importância dos caracteres reprodutivos, ou seja, tipo de floração e morfologia, síndromes de polinização, dupla fecundação, endosperma e embriogênese para difusão das sementes por meio de frutos comestíveis, e (3) uma perspectiva mais pluralista, explicando que o sucesso das plantas com flores está na união de todos os fatores mencionados acima (Crepet; Niklas, 2008).

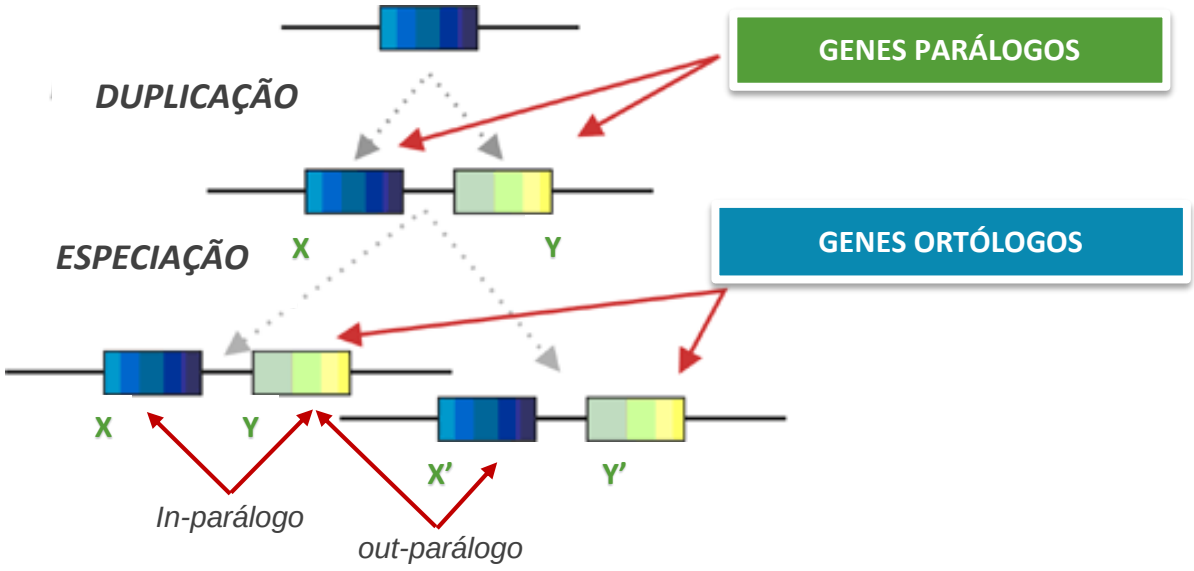
As principais fontes de diversificação nas Angiospermas estão na recombinação gênica, mutações e eventos de completa duplicação do genoma (*Whole-genome duplication* – WGDs), essenciais na seleção natural e deriva gênica. A recombinação gênica envolve a segregação de genes que ocorrem principalmente durante a meiose, na formação dos esporos durante o processo de *crossing-over*, onde há a troca de segmentos cromossômicos entres os pares de cromossomos homólogos alinhados. Já as mutações envolvem alterações na molécula de DNA, sejam elas pontuais (troca de nucleotídeos), inserções, duplicações, deleções e inversões de partes de um cromossomo, ganho ou perda de cromossomos inteiros ou mudanças no genoma por inteiro (Waters, 2003; Judd *et al*., 2009). Os efeitos dessas mudanças podem ser letais, neutros ou seletivamente vantajosos, caso novos arranjos gênicos promovam benefícios ao metabolismo ou a fisiologia das plantas.

Embora os processos de duplicações de genes individuais, segmentos cromossômicos, ou genomas inteiros sejam tratados como principal fonte de material para a origem de novidades evolutivas, não está claro em todos os vegetais como estes adotam com sucesso uma trajetória evolutiva de um estado inicial de redundância completa, em que um exemplar é dispensável, para uma situação estável em que ambas as cópias são mantidas (Lynch; Conery, 2000). Porém, sabe-se que após a duplicação do genoma em plantas há a subsequente ação da seleção purificadora (ou negativa) com remoção das mutações prejudiciais; neutra por deriva genética com variação aleatória nas frequências alélicas; ou positiva, fixando as mutações que tragam alguma vantagem ao indivíduo (Kimura, 1989; Futuyma, 2009; Nei *et al*, 2010). Esses fenômenos podem resultar em grandes modificações no padrão da expressão gênica e proporcionar uma importante fonte de variação adaptativa (Dickinson, 2012). Durante os processos de duplicação e seleção, os novos pares de genes podem gerar ortólogos, conjunto de genes que divergiram por especiação, a partir de um ancestral comum, ou se manterem como réplicas distintas dentro da mesma espécie, acarretando na formação dos parálogos (Ridley, 2006; Futuyma, 2009) (Fig.1). Sendo apenas esse último capaz de possuir ou não a mesma função de seu ancestral, ao sofrer pressão de seleção diferenciada.

Logo, existem três desfechos para evolução dos genes duplicados: (1) uma cópia pode ser silenciada por mutações degenerativas (não funcionalização ou pseudogenização), (2) uma cópia pode adquirir uma nova função e ser preservada pela seleção natural, retendo a outra cópia com a função original (neofuncionalização);

ou (3) ambas as cópias podem se tornar parcialmente comprometidas pelo acumulo de mutações, ao ponto em que a capacidade de ambas é reduzida para o nível da única cópia do gene ancestral (subfuncionalização) (Futuyma, 2009; Dickinson, 2012).

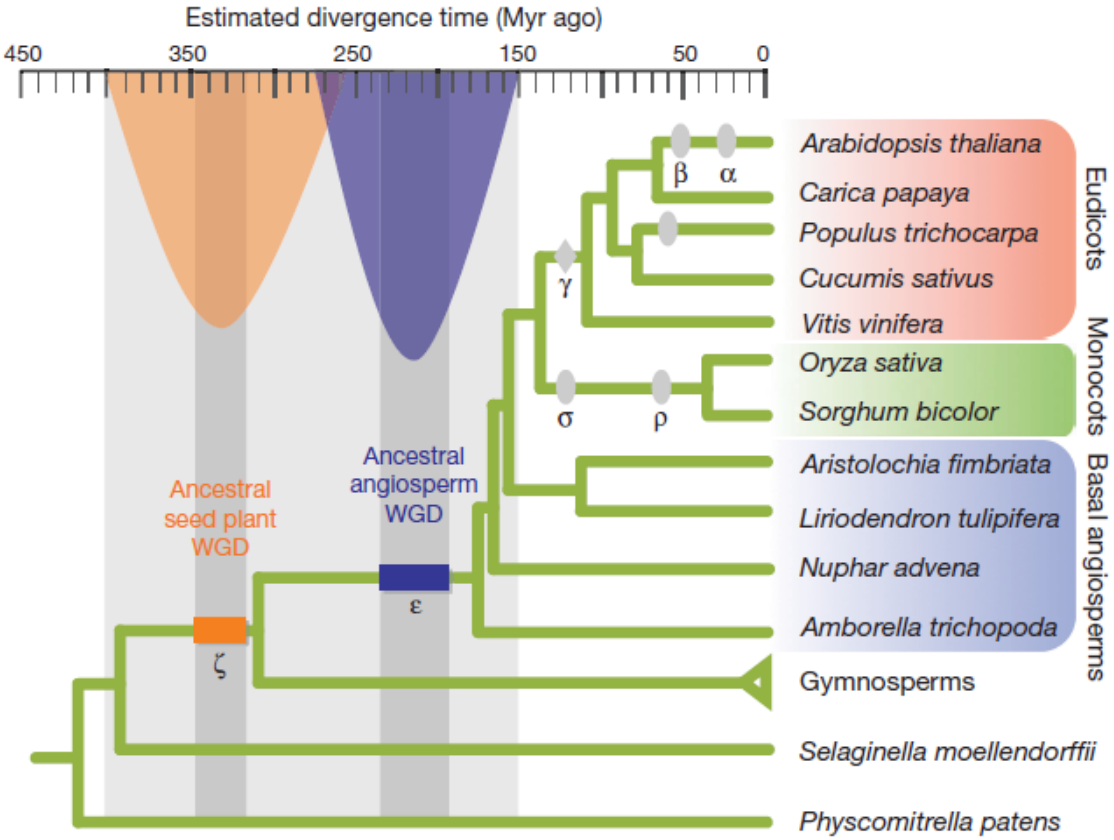
Figura 1: Processos de duplicação do gene.



Legenda: Duas possibilidades de duplicação dos genes identificados na evolução das plantas terrestres. Cópias em azul estão diretamente relacionadas ao ancestral comum.

Eventos de duplicação de genes foram intensamente concentrados em torno de 319 e 192 milhões de anos atrás, implicando em dois eventos de completa duplicação do genoma (*Whole-genome duplication* – WGDs) em linhagens ancestrais, pouco após a diversificação das plantas com sementes e angiospermas atuais, respectivamente (Lawton-Rauh, 2003; Jiao *et al*., 2011) (Fig.2). A duplicação desse genoma forneceu material genético bruto para o surgimento de novidades funcionais e metabólicas (Maira *et al.* 2013). Estas completas duplicações dos genomas ancestrais resultaram na diversificação de genes regulatórios importantes para o estabelecimento de modificações nos mecanismos enzimáticos e desenvolvimento de sementes e flores, sugerindo que eles estavam envolvidos em grandes inovações que, em última análise contribuiu para o aumento na dispersão e dominação das plantas com sementes (Magallón; Castillo, 2009, Jiao *et al*., 2011).

Figura 2: Eventos poliploides ancestrais nas plantas com sementes e angiospermas.



Legenda: Em destaque, dois eventos de duplicação (WGD) ancestrais identificados na evolução das plantas terrestres, e sua respectivas distribuições normais (parábolas). Elipses: duplicações aceitas e identificados em genomas já sequenciados. Losango: evento de triplicação (3ɳ) provavelmente compartilhada pelas eudicotiledôneas núcleo. Retângulos: indicam regiões de confiança para distribuição das estimativas médias das idades de seus ancestrais. Adaptado de Jiao *et al.*, 2011.

Devido a esses fatores, muitas linhagens das plantas verdes (*Viridiplantae*), por meio destes e de outros processos de evolução molecular, têm acumulado uma extraordinária diversidade de espécies que compreende uma vasta versatilidade morfológica, funcional e metabólica, constituindo a base energética da grande maioria dos ecossistemas terrestres (Lewis; McCourt, 2004; Niklas; Kutschera, 2009). No caso das angiospermas a sua marcante diversidade e abundância permitiu não apenas uma capacidade de aclimatação, como também o desenvolvimento de interações complexas dentre e entre os níveis tróficos, e o surgimento de novos mecanismos metabólicos para sua sobrevivência (Futuyma, 2009; Niklas; Kutschera, 2009).

O metabolismo vegetal é composto por uma complexa rede de eventos físicos e químicos resultantes na fotossíntese, respiração, e na síntese e degradação de compostos orgânicos. As vias metabólicas associadas a essa rede são unidades com funções bioquímicas que englobam uma série de conversões de substrato à produtos, conduzidos por meio de um intermediário químico (Schnarrenberger; Martin, 2002). Isto só é possível graças ao grande número de interações provenientes dos diferentes mecanismos químicos, enzimáticos e regulatórios oriundos da expressão gênica diferenciada, em resposta as inúmeras modificações a que um vegetal pode estar sujeito, provenientes da plasticidade na capacidade de resposta adquirida ao longo da evolução molecular e conquistas de novos ambientes (Futuyma, 2009; Niklas; Kutschera, 2009; Magallón; Castillo, 2009; Finet *et al.* 2010).

1.2 Ciclo do Glioxilato e suas Origens

Dentre as principais rotas bioquímicas dos eucariontes, o ciclo do ácido cítrico ou do ácido tricarboxílico (TCA) constitui um elemento central no metabolismo de carbono das plantas superiores, que fornecem, entre outras coisas, os elétrons para a fosforilação oxidativa na membrana mitocondrial, intermediários para a biossíntese de aminoácidos, e oxaloacetato para a gliconeogênese (Schnarrenberger; Martin, 2002). Embora o ciclo do ácido tricarboxílico seja uma via mitocondrial típica em eucariontes superiores, com a maioria de suas enzimas codificadas no núcleo, sua origem remonta às α-proteobactérias, reforçando a hipótese que os genes nucleares eucarióticos foram adquiridos a partir do genoma mitocondrial, após aquisição de uma α-proteobactéria de maneira endosimbiótica durante o curso de evolução (Huynen *et al.*, 1999; Gray *et al.*, 1999, 2001; Adams; Palmer, 2003).

Ao se estudar as enzimas do ciclo do TCA necessariamente também se envolve o estudo das várias enzimas envolvidas no Ciclo do Glioxilato em plantas, pois três etapas enzimáticas são comuns a ambos os ciclos, estas catalisadas por isoenzimas diferencialmente compartimentadas, processos análogos às isoenzimas do cloroplasto envolvidas no ciclo de Calvin e Glicólise nos vegetais (Schnarrenberger; Martin, 2002). Caracterizado nas Plantas por Kornberg & Beevers (1957), o Ciclo do Glioxilato é descrito como uma variante do ciclo do TCA capaz de realizar a conversão líquida de duas moléculas de acetil-CoA à succinato, evitando as duas etapas de descarboxilação do Ciclo do TCA, catalisados pela Isocitrato desidrogenase e α-

cetoglutarato desidrogenase (Cooper & Beevers, 1969; Eastmond & Graham, 2001; Kornberg & Beevers, 1957; Schnarrenberger & Martin, 2002). Esse apresenta sua atividade mais evidente nos tecidos de reserva, não sendo ausente sua expressão relacionada a outras funções no vegetal (Nelson; Cox, 2011).

Contudo, é amplamente aceito que o ciclo do glioxilato opere não apenas nas plantas verdes mas também em bactérias (Nelson; Cox, 2011), fungos (Lowenstein, 1967), protistas (Levy; Scherbaum, 1965; Nakazawa *et al.*, 2005), além de alguns invertebrados (ex. *C. elegans*), onde identificaram e caracterizaram uma enzima bifuncional com atividades de ambas, ICL e MLS, que aparentemente evoluiu por um processo de fusão dos respectivos genes (Liu *et al.*, 1995; Nakazawa *et al*. 2011).

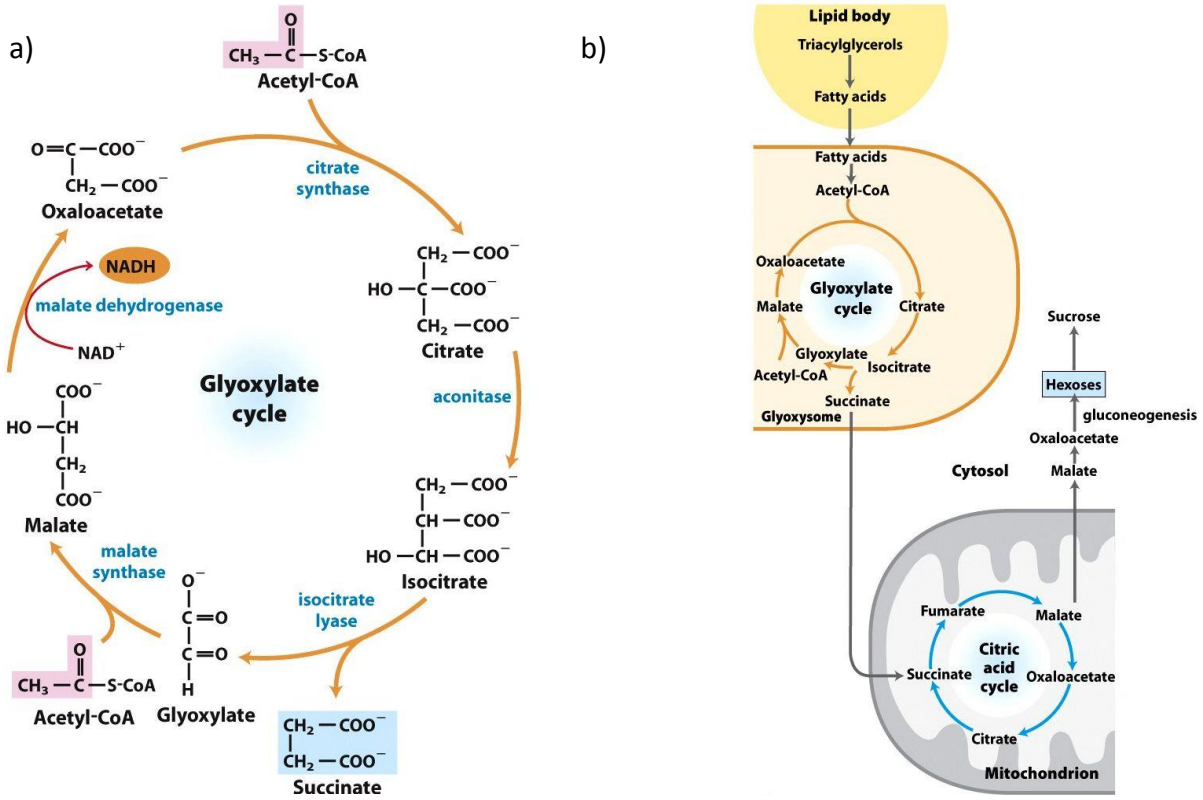
Em meio a maioria das sementes de plantas oleaginosas o endosperma é o tecido especializado na reserva de nutrientes e metabólitos, utilizado durante a germinação e desenvolvimento pós-germinativo da plântula (Eastmond; Graham, 2001). Neste tecido, os lipídios são a principal fonte energética das sementes, onde são armazenados durante a sua maturação (Graham, 2008). Quando necessário são rapidamente convertidos a metabólitos solúveis, que podem ser transportados para a plântula e utilizados para suprir o crescimento, permitindo que o sistema fotossintético se estabeleça antes do esgotamento das reservas (Quettier; Eastmond, 2009).

A mobilização dos óleos durante a germinação envolve uma regulação coordenada de uma complexa rede metabólica de genes, enzimas e proteínas associadas, integrada com vários outros aspectos do metabolismo celular (Graham, 2008). Após a mobilização proveniente da β-oxidação dos ácidos graxos, há a formação de moléculas de acetil-CoA, que serão posteriormente metabolizadas via ciclo do glioxilato (Fig.03; a), situado no Glioxissomo, formas especializadas do peroxissomas em plantas (Graham, 2008; Nelson; Cox, 2011). Logo, este ciclo é mobilizado essencialmente durante o início da germinação até plântula jovem ser capaz de emitir sua radícula e estabelecer seu sistema fotossintético.

Em um determinado espaço de tempo a semente germinada está sujeita a uma gama de variações ambientais que podem acarretar em respostas indesejadas no desenvolvimento do vegetal. Portanto, a ausência dos passos de descarboxilação (Eastmond; Graham, 2001; Graham, 2008), como fatores limitantes da velocidade de reação, permite o aproveitamento dos carbonos provenientes de moléculas lipídicas e molécula de acetil-CoA para a síntese de carboidratos, já que no ciclo do ácido cítrico os dois carbonos do resíduo acetil são liberados na forma de CO_2, impedindo

o seu aproveitamento para reações de síntese, após a entrada no ciclo. O que não é executado pelos representantes do subreino *Metazoa*, com exceção dos nematódeos, visto que não ocorre a síntese de carboidratos a partir da utilização de moléculas de carbono na forma de Acetil-CoA (Liu *et al.*, 1995; Nelson; Cox, 2011).

Figura 3: Ciclo do Glioxilato e aproveitamento do Succinato para síntese de carboidratos por rota alternativa do metabolismo de Acetil-CoA.



Legenda: (a) Ciclo do Glioxilato nos glioxissomos de sementes oleaginosas; (b) Relação entre o ciclo do glioxilato e TCA, conduzindo à síntese de carboidratos. **Fonte:** Nelson & Cox, 2011.

As enzimas Malato sintase (MLS; EC 4.1.3.2) e Isocitrato liase (ICL; EC 4.1.2.1) são únicas ao ciclo do glioxilato, e as outras 3 enzimas, Citrato sintase (CSY), Aconitase (ACO), e Malato desidrogenase (MDH) são compartilhadas com o Ciclo de Krebs. As proteínas MLS, ICL e CSY estão localizadas nos peroxissomos vegetais, enquanto as reações enzimáticas da ACO e MDH para o ciclo do glioxilato estão localizadas fora do glioxissomo (Fig.3 e 4). Ambas MLS e CSY utilizam acetil-CoA como substrato para produzir malato e citrato, respectivamente (Graham, 2008).

Assim, duas moléculas de acetil-CoA são introduzidos em cada volta do ciclo, resultando na síntese líquida de um mol do composto de quatro carbonos succinato,

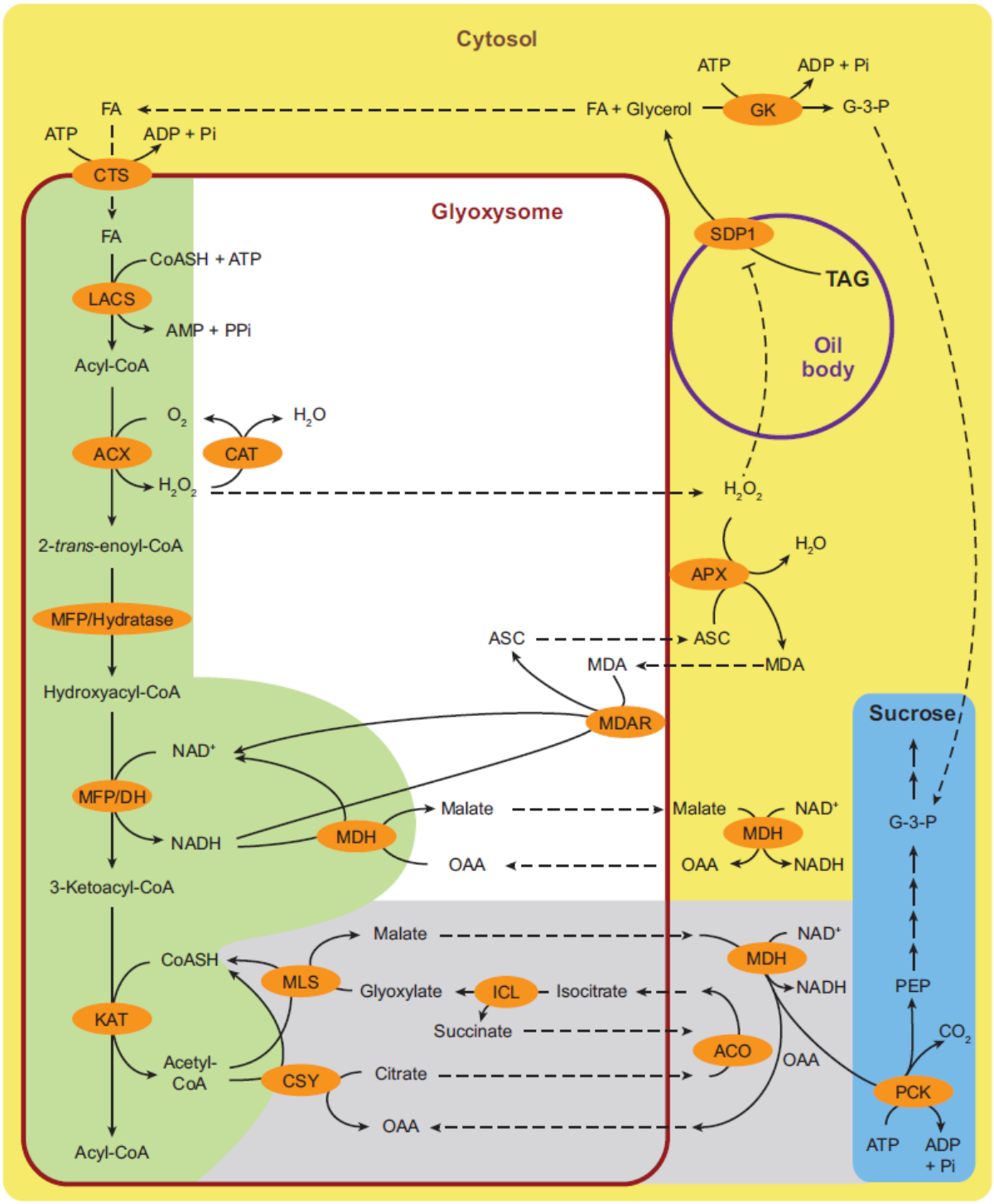
pela enzima ICL (Eastmond; Graham, 2001; Kunze *et al.,* 2006). O succinato passa do glioxissomo para a mitocôndria e entra no ciclo de Krebs (Fig.3; b), onde é convertido a malato. Este é exportado para o citosol em trocado de succinato e convertido a oxaloacetato pela isoforma citosólica da enzima Malato desidrogenase.

Por fim, a enzima Fosfoenolpiruvato (PEP) carboxiquinase, catalisa a conversão do oxaloacetato a PEP promovendo a síntese de carboidratos solúveis para a nutrição e desenvolvimento do embrião, via o processo de gliconeogênese (Fig.03; b) (Eastmond; Graham, 2001; Walli; Brows, 2010).

Devido a serem exclusivas do ciclo do glioxilato (Graham, 2008), as enzimas MLS e ICL são comumente utilizadas como marcadores para a mensuração do fluxo nesta rota metabólica (Fig.4). Em muitas sementes oleaginosas, existe uma forte correlação entre a quebra e mobilização dos lipídeos e a expressão da ICL e da MLS durante o crescimento pós-germinativo (Eastmond; Graham, 2001). Contudo, existem modelos vegetais em que a atuação desse é dispensável, em algumas condições. Plântulas do mutante da ICL de *Arabidopsis* têm dificuldade de crescimento, pois são incapazes de converter o acetato da β-oxidação em açúcares (Eastmond *et al.*, 2000), e tendem a utilizar o Citrato ou o Acetato dos peroxissomos nas reações oxidativas da mitocôndria (Eastmond; Graham, 2001; Cornah *et al*., 2004). A ausência da MLS em *Arabidopsis* resulta em um fenótipo menos severo (Kunze *et al*., 2006), pois as plântulas são capazes de realizar um mecanismo de gliconeogênese alternativo, empregando as enzimas da fotorrespiração em conjunto com as do ciclo do glioxilato, pois foi constatado que ambas as rotas coexistem nos mesmos peroxissomos durante o crescimento da plântula em outras espécies de oleaginosas (Cornah *et al*., 2004).

Embora esta via aparente não ser um requerimento essencial a germinação e ao crescimento inicial (Eastmond *et al*., 2000), a conversão líquida de lipídeos a carboidratos é extremamente necessária para um melhor desenvolvimento destas plântulas (Graham, 2008), propiciando uma vantagem evolutiva e por consequência levando a uma maior taxa de sucesso reprodutivo.

Figura 4: Regulação e processos envolvidos na mobilização dos ácidos graxos das reservas de sementes para conversão em sacarose, via ciclo do glioxilato em *Arabidopsis*.



Legenda: Em amarelo: a localização citosólica dos oleossomos; em verde: significa um envolvimento direto na β-oxidação glioxissomal; em cinza: denota o ciclo do glioxilato; e em azul: a gliconeogênese. **Fonte:** Graham, 2008.

Portanto, tendo em vista essas enzimas desempenharem um papel central nas etapas iniciais de utilização das reservas lipídicas das sementes e estabelecimento do estágio de plântula, a escassez de conhecimento sobre suas relações e processos evolutivos, que podem ter determinado a sua distribuição entre as plantas verdes (*Viridiplantae*), denota uma clara necessidade de estudos com suporte molecular filogenético para melhor caracterização de seus genes, origem e relações evolutivas do ciclo entre diferentes espécies vegetais.

1.3 As enzimas ICL e MLS

Conforme descrito anteriormente, o Ciclo do Glioxilato envolve cinco atividades enzimáticas que estão compartimentalizadas nos glioxissomos de plantas, exceto a Aconitase localizada no citosol (Graham, 2008). Duas entre essas enzimas são responsáveis por fornecer uma via intermediária do metabolismo, permitindo certo organismos atenderem suas demandas de carbono por meio de compostos de dois carbono: (1) a Isocitrato liase, com a clivagem de isocitrato a succinato e glioxilato (enquanto no ciclo Krebs o isocitrato é convertido a succinato e duas moléculas de dióxido de carbono); e (2) a Malato sintase, ao condensar glioxilato de um grupo acetila originado partir de uma molécula de acetil-CoA para a produção de malato, afim de repor o conjunto de intermediários do ciclo do TCA (Howard *et al.*, 2000; Eastmond; Graham, 2001; Nelson; Cox, 2011). De maneira sucinta, as equações das reações descritas, para cada enzima, consistem em:

(i) Isocitrato → succinato + glioxilato

Ao se representar a ação catalítica da ICL (EC 4.1.3.1) na clivagem do isocitrato para formação de succinato e glioxilato.

(ii) Glioxilato + H_2O + acetil-CoA → L-malato + CoASH

A MLS (EC 4.1.3.2) ao transferir o grupamento acetila do acetil-CoA para o glioxilato resultando em L-malato. Ou seja, a síntese de carboidratos oriunda do consumo das reservas lipídicas em plantas pode ser resumida por:

(iii) Ácido graxo → acetil-CoA → [(i)+(ii)] → fosfoenolpiruvato → [glicólise]

Embora as enzimas sejam consideradas unidade únicas sua ação catalítica está na verdade associada a um conjunto de unidades menores. No caso do ciclo do TCA, e por conseguinte sua variante, muitas enzimas consistem de múltiplas subunidades (Schnarrenberger; Martin, 2002). Estruturalmente ambas as enzimas têm sido designadas com formas monoméricas e oligoméricas (Howard *et al.*, 2000).

A MLS não segue um padrão estrutural conservado em todos os seres vivos, sendo isolada na forma de octâmero de subunidades idênticas de ≈ 60-kDa dentre algumas espécies vegetais, como um homotetrâmero em fungos, e como um homodímero em eubactérias. A ICL é caracteristicamente um homotetrâmero com subunidades de ≈ 64 kDa. As eucarióticas se enquadram em dois grupos: (a) um que contém as sequências eucarióticas de *Caenorhabditis* e *Chlamydomonas* e apresentam elevada similaridade a genomas homólogos em γ-proteobacterial, e (b) aquele que codifica as enzimas glioxissomais de plantas e fungos (Schnarrenberger; Martin, 2002; Kondrashov *et al*., 2006).

1.4 Bases Moleculares da Evolução em Proteínas

Para que se possa compreender as mudanças a que estão sujeitas as macromoléculas de DNA, RNA e/ou Aminoácidos nas vias metabólicas, sejam em plantas ou qualquer outro tipo de ser vivo, é imprescindível entender algumas das causas e origens destes processos evolutivos rumo a adaptação.

Como mencionado anteriormente, a evolução nos ácidos nucleicos ocorre continuamente por processos de mutação, os quais provocam alterações nos aminoácidos codificados. Estes resultam em subsequentes modificações na forma e funcionamento das proteínas (Futuyma, 2009; Nelson; Cox, 2011). Porém, as origens da evolução molecular não estão associadas apenas as falhas nos mecanismos de ligação dos pares de bases durante a replicação do DNA, aonde entram as mutações classificadas de acordo com a alteração estrutural que produzem (Mutações pontuais, Inserções e Deleção), mas também aqueles ligados a outros processos já citados para as plantas (Nei; Kumar, 2000). Dentre esses estão a Recombinação, Duplicação gênica e transferência horizontal de Genes.

Mutações por modificações estruturais são todas tipos de mutações em pequena escala mas que podem afetar diretamente um ou poucos nucleotídeos de um gene (Ridley, 2009). A Seleção Natural irá exercer seu papel no momento em que essa resposta resultar em alguma variação adaptativa de estado vantajosa ou prejudicial para o organismo (Futuyma, 2009). Podendo ser mantida nas gerações seguintes se a mutação conferir algum tipo de vantagem ou eliminada, pela ausência de sobrevivência ou sucesso reprodutivo do portador. Porém, ao ocorrer mutações neutras, ausência de vantagens adaptativas, o seu estabelecimento nas populações será pelo mecanismo de Deriva Genética, em que por terem frequências alélicas aleatórias ao longo das gerações não é possível convergir o sentido das mudanças evolutivas (Nei; Kumar, 2000; Ridley, 2009).

A deriva genética e a Seleção natural raramente ocorrem isoladas em populações naturais; ambas estão sempre agindo sobre uma população (Ridley, 2009). Entretanto, o grau de influência destes dois fenômenos pode variar em função das circunstâncias. Um exemplo de tal fenômeno está presente nas mutações sinônimas, pois não irão interferir no processo de adaptação do indivíduo já que não apresentam um impacto funcional significativo na codificação dos aminoácidos, mas irão gerar um efeito cumulativo que dará origem a diversidade biológica nas gerações seguintes (Nei; Kumar, 2000; Futuyma, 2009).

A molécula de DNA está sujeita a dois diferentes processos de substituição nucleotídica em suas sequências, que acarretarão em mutações (Jukes; Collis, 1994). Mudanças por “transição”, no que se refere as trocas envolvendo nucleotídeos de tipos químicos iguais (purina por outra purina ou de uma pirimidina por outra pirimidina) e/ou “transversão”, definida pelas substituições envolvendo os diferentes tipos químicos (pirimidinas por purinas ou vice-versa). Contudo, a probabilidade de ocorrência entre estas duas formas tende a favorecer as transições, mais comuns e frequentes se comparadas às transversões (Futuyma, 2009).

A taxa de substituição entre sequencias apresenta uma relação direta com o tempo de divergência entre elas. Devido a múltiplas mudanças que podem ocorrer no mesmo sitio a relação entre a diferença e o tempo de divergência das sequencias tende a torna-se repetitivas, levando por consequência a perda com o tempo da informação ali contida por saturação (Nei; Kumar, 2000; Ridley, 2009).

Portanto, a casualidade nas mutações está associada com a natureza dos processos que as originam, contudo ainda é necessária a ação dos processos

secundários de deriva genética ou seleção natural para se definir o papel que esta mudança trará ao indivíduo e sua espécie.

Ao se abordar as implicações diretas destas modificações para o metabolismo do indivíduo, podemos estabelecer uma relação com o produto proteico originado e sua influência sobre a conformação estrutural de uma proteína, pois como resultado dessas mutações pontuais a leitura dos códons durante a síntese proteica é alterada e apenas uma única mudança pode ocasionar transformações a longo prazo (Nelson; Cox, 2011; Meier *et al.,* 2007).

Durante a evolução a seleção pode agir não só ao nível de sequências, mas também ao nível de organização estrutural das proteínas (Meier *et al.,* 2007). Parâmetros subjacentes associados com os processos de mutação e fixação também são importantes. Estes incluem: a taxa de mutação, a taxa de recombinação e o tamanho da população. Compreender a coevolução de resíduos dentro das estruturas protéicas é fundamental tanto para a análise dos mecanismos de funcionamento enzimáticos das proteínas (Liberles *et al*., 2012), bem como para estudos evolutivos em redes de interatomas.

Modelos probabilísticos de mudança em sequências têm um papel central no estudo da evolução molecular. Suas vantagens são a exploração em simulações qualitativas na evolução das proteínas, como por admitirem a estimativa de parâmetros e avaliação de hipóteses através de testes estatísticas baseadas em verossimilhança. Para evitar implicações da presença de sítios independentes, as sequências de proteínas são tipicamente preditas para apresentarem propriedades tais como a estabilidade termodinâmica e dobramento, sendo a taxa de substituição expressa em função de alterações destas propriedades (Liberles *et al*., 2012).

Dentre as proteínas envolvidas nos sistemas biológicos, aqueles presentes no metabolismo apresentam um elevado grau de conservação no que diz respeito aos seus domínios funcionais. Logo, regiões estruturais que apresentem um domínio funcional nas suas sequencias tentem a ser mais conservada evolutivamente ao longo do tempo, uma vez que a pressão de seleção irá atuar diretamente sobre a porção envolvida na ação da proteína (Nei; Kumar, 2000; Futuyma, 2009). Deste modo, demonstrando a necessidade em se elucidar as relações das rotas/vias do metabolismo vegetal, sob uma perspectiva evolutiva molecular estrutural.

1.5 Métodos em Filogenia Molecular

Ao se realizar estudos de evolução molecular, a necessidade intrínseca do uso de ferramentas que se apoiem na utilização de caracteres moleculares como base para as análises da história evolutiva, seja de um grupo de organismos ou gene, se faz presente por meio da filogenia. O termo “filogenia molecular” se refere à filogenia macromolecular, ou seja, o estudo das relações evolutivas entre os organismos e seus genes pelo uso de sequências de ácidos nucléicos e aminoácidos, elementos transponíveis, ou outros marcadores (Graur; Li, 2000).

Logo, a utilização dos dados moleculares para o esclarecimento e proposição de teorias sobre processos evolutivos envolvidos na reconstrução filogenética de organismos e suas rotas metabólicas, bem como sua origem, tem se tornado o principal recurso dos pesquisadores nos últimos anos, a exemplo dos estudos de Gray (1999), Adams (2003), Schlüter *et al.* (2006), Hügler & Sievert (2011), Soltis *et al.* (2011) e Maira *et al.* (2013).

Incialmente, para se estabelecer as relações filogenéticas derivadas a partir do conjunto de sequências moleculares que se está estudando, o correto alinhamento é uma etapa indispensável (Nei; Kumar, 2000; Schneider, 2007). Na filogenia molecular a forma mais comum de alinhamento se trata do “alinhamento múltiplo”, utilizado na inferência das relações evolutivas entre sequências, e determinação dos padrões de funcionamento entre grupos de genes. Os métodos de alinhamento entre sequências são baseados num sistema de escores para cada alinhamento obtido entre cada pareamento (Schneider, 2007), no qual sua pontuação é calculada por algoritmos que determinam se a presença de trocas dos pares e eventos de inserção ou remoção (*indels*) serão necessários para a correta disposição do alinhamento.

Os métodos de inferência filogenética conhecidos, por meio de alinhamentos de sequências, podem ser baseados nas distâncias entre os pares de sequências, ou na presença ou ausência de variabilidade nos sítios analisados entre cada sequência. Sendo assim, os métodos de inferência filogenética podem ser categorizados em dois tipos básicos: os métodos de distância e os métodos discretos (Page; Holmes, 2001).

Dentre esses métodos, as técnicas de reconstrução mais comum aplicadas são: a *Neighbor-joining* (NJ) e Evolução Mínima (ME), como métodos de distância; Máxima parcimônia (MP), Máxima verossimilhança (ML) e Inferência Bayesiana (IB), como métodos discretos (Nei; Kumar, 2000; Lemey *et al*., 2009). Segundo Schneider

(2007), é necessário estar claro quais são as implicações na escolha dos métodos, pois cada um apresenta algoritmos diferentes, uma vez que as conclusões relevantes na evolução, em termos de topologias, podem acarretar em interpretações errôneas. Os métodos de distância como o *Neighbor-joining* ao converterem o alinhamento em uma matriz de distâncias entre sequências reflete a informação contida nelas sobre distância evolutiva atual entre seus representantes na forma de topologias (Nei; Kumar, 2000). Como métodos discretos mais utilizados na reconstrução filogenética estão o método da Máxima Parcimônia e o da Máxima Verossimilhança.

A Máxima Parcimônia é baseada na concepção de que a maior probabilidade de se reconstruir a filogenia está na escolha da hipótese que pressupõe o menor número de mudanças ou etapas para se obter o resultado final. Ao se trabalhar com esse método apenas parte da informação contida nas sequências pode ser recuperada, uma vez que nem toda variabilidade será utilizada como sítios informativos. Com isso, o número de substituições em cada sítio informativo é inferido e a totalidade em cada árvore é calculada, para assim a árvore final com menor número de modificações ser selecionada (Nei; Kumar, 2000).

Na Máxima Verossimilhança sua inferência considera todos os sítios indistintamente (Schneider, 2007) com base no princípio da verossimilhança, em que dentre todas as topologias possíveis apenas a com maior probabilidade será considerada a mais correta. Para isso, dentre as sequências analisadas um modelo é proposto com sua topologia e comprimento dos ramos levados em consideração. Em seguida, variações desse modelo são propostas e a filogenia que apresentar maior probabilidade é considerada a mais verossímil (Nei; Kumar, 2000).

Para ambos, afim de garantir a confiabilidade dos resultados obtidos nas topologias pelos métodos citados acima, é realizado o chamado teste de *bootstrap*, em que é feita uma reamostragem estatística com reposição pseudoaleatória dos dados a cada corrida (Nei; Kumar, 2000). Durante essa reamostragem uma nova árvore é construída com base em variações do conjunto de dados. A análise de *bootstrap* é uma técnica simples e eficaz para testar a estabilidade dos grupos dentro de uma árvore filogenética. A principal vantagem está em poder ser aplicado, essencialmente, em qualquer método de construção de árvores, embora deva ser lembrado que a aplicação do método de *bootstrap* multiplica o tempo computacional necessário para obtenção do número de amostras solicitadas (Lemey *et al*., 2009).

Por fim, a Inferência Bayesiana, na qual antes de se começar a análise é preciso especificar quais as *priori* sobre as relações entre sequências. Pois, na ausência de informações sobre suas relações, uma solução simples seria atribuir uma probabilidade equivalente para todos. Contudo, nesse método durante a construção da árvore filogenética o que é levado em consideração é a probabilidade a *posteriori* de um determinado conjunto de dados, sem esquecer as taxas de verossimilhança presente e suas probabilidades *a priori* (Ronquist *et al*., 2009). Seguindo o teorema de Bayes, para estimar a distribuição de probabilidade *a posteriori* o método se utiliza do algoritmo de Metropolis (Metropolis *et al., 1953*), através de amostragem obtida com a construção de uma cadeia de Markov Monte Carlo, cadeia essa que tem a propriedade de convergir para um estado de equilíbrio, independente do ponto de partida. E para que isso ocorra, ao longo de suas gerações pequenas mudanças são aceitas ou rejeitadas de acordo com os valores de verossimilhança do estágio seguinte na cadeia (Lemey *et al*., 2009).

Além disso, na análise e inferência filogenética para garantir a veracidade da reconstrução e influência dos processos evolutivos são levados em consideração no cálculo outros parâmetros, como: os modelos probabilísticos evolutivos, distribuições *Gamma*, proporção de sítios invariáveis, etc (Nei; Kumar, 2000; Schneider, 2007).

Todavia, independentemente do método utilizado para se realizar a reconstrução filogenética é necessário levar em consideração a escolha de uma característica distinta ao se enraizar as topologias que definam os estados primitivos e derivados de um estado de caráter, que ao se tratar de uma análise molecular sua raiz (*outgroup*) pode ser considerada a sequência de apenas um ancestral comum ao aparecimento do caráter observado (Schneider, 2007; Lemey *et al*., 2009).

Estudos evolutivos de rotas metabólicas têm sido baseados em análises filogenéticas convencionais dos participantes de suas vias individuais, sendo feita a partir destas apenas a comparação entre as árvores obtidas das diferentes enzimas, na busca por padrões de similaridade ou divergência filogenéticas (Schnarrenberger; Martin, 2002). Isso foi realizado para o ciclo de Calvin (a via de fixação de CO_2), na via glicolítica/gliconeogênica (Martin; Schnarrenberger, 1997; Henze *et al.*, 2001), e para as duas diferentes vias de biossíntese de isoprenóides (Lange *et al.*, 2000). O que segundo os autores revelaram um elevado grau de mosaicismo dentre as vias estudadas em ambos os tipos de organismo (procariotas e eucariotas), indicando que essas tendem a evoluir como entidades coerentes de atividade enzimática.

Sendo o ciclo do glioxilato um exemplo de via metabólica com genes de enzimas presentes no genoma de todos os vegetais, mesmo não tendo sua expressão designada na utilização de reservas das sementes durante a germinação, como um mecanismo compartilhado, trata-se de um possível modelo de rota metabólica que, por estar sujeitas a processos evolutivos de dispersão e colonização de novas áreas, podem ter determinado sua distribuição entre as plantas. Ao nosso conhecimento, apenas dois estudos anteriores abordaram o Ciclo do Glioxilato sob um enfoque evolutivo, que são os trabalhos de Kondrashov *et al.* (2006) e Schnarrenberger & Martin (2002). No entanto, estes autores não foram capazes de realizar uma correlação entre a evolução molecular das enzimas do ciclo com a evolução das espécies oleaginosas estudadas. Adicionalmente, nos últimos anos houve um intenso aumento no número de dados destas sequências disponíveis nos bancos de dados biológicos, devido ao crescimento do número de genomas vegetais sequenciados. Demostrando a clara necessidade de estudos evolutivos e filogenéticos para um melhor entendimento e conhecimento da evolução dos genes das enzimas MLS e ICL e do próprio ciclo do glioxilato nas plantas verdes (*Viridiplantae*).

2. OBJETIVOS

2.1 Geral

Realizar estudos de evolução molecular das enzimas Isocitrato liase e Malato sintase, no táxon *Viridiplantae.*

2.2 Específicos

- Obter sequencias homologas (ortólogas e parálogas) dos genes das enzimas Isocitrato liase e Malato sintase de diferentes representantes no táxon *Viridiplantae*;
- Realizar estudos de filogenia molecular dos genes das enzimas *icl* e *mls*;
- Comparar a filogenia do grupo das *Viridiplantae* com as análises de agrupamentos resultantes das sequências dos genes *icl* e *mls*;
- Elaborar modelos teóricos estruturais de ambas as enzimas para representantes do táxon *Viridiplantae*;
- Identificar os tipos de processos evolutivos envolvidos, a partir da observação de mudanças nas sequencias biológicas e nas estruturas tridimensionais das enzimas Icl e Mls;
- Inferir possíveis processos evolutivos envolvidos na diferenciação e especialização do ciclo do glioxilato entre os grupos do táxon *Viridiplantae*.

3. MATERIAL E MÉTODOS

3.1 Busca por similaridade e Obtenção das sequências

A Busca pelas sequências de interesse foi feita através das ferramentas BLAST (*Basic Local Alignment Search* Tool), com o uso do BLASTp, PHI-BLAST (*Pattern Hit Initiated* BLAST) e tBLASTn no NCBI *(National Center for Biotechnology Information* - www.ncbi.nlm.nih.gov) (Altschul *et al*., 1997) e UNIPROT (*Universal Protein Resource* - www.uniprot.org), optando-se apenas pela aquisição de RefSeq's de diferentes espécies em *Viridiplantae* que apresentam-se completas, com anotação e função devidamente caracterizada para ambos os bancos de dados. Com a intenção de se obter uma melhor qualidade dos resultados as buscas foram feitas utilizando as informações contidas nos domínios da proteína, famílias e sítios funcionais, bem como padrões e perfis de identificação associados no Pfam (*Protein families databases* - http://www.pfam.xfam.org) e PROSITE (*Database of protein domains, families and functional sites* - http://prosite.expasy.org), utilizando suas matrizes de escores de posição especifica de resíduos (PSSM - *Position-Specific Scoring Matrix*) e padrões de assinatura proteica. Todas as sequências foram obtidas com base nos seus genes reconhecidos para diferentes representantes do genoma vegetal e armazenadas em formato .fasta (Pearson, 1990), enquanto as informações sobre anotação em formato ASN.1 (NCBI).

3.2 Montagem de um banco de dados local

Um banco de dados local em servidor Apache 2.2.27 (*HTTP Server*) foi construído com as sequências obtidas, em uma estação de trabalho Linux, a partir da utilização do pacote wwwblast-2.2.26-x64 (*ftp.ncbi.nih.gov/blast /executables/release/*) fornecido pelo NCBI, para uso local da ferramenta BLAST, através de um conjunto de programas independentes que realizam buscas por semelhança utilizando o mesmo algoritmo heurístico dos servidores do NCBI (*http://www.ncbi.nlm.nih.gov/blast/*).

3.3 Múltiplo alinhamento, Análise de Saturação e Teste do Modelo

As sequências foram submetidas inicialmente a alinhamento múltiplo com o uso da ferramenta MAFFT (v.7.149; Katoh; Asimenos; Toh, 2009; Katoh; Toh, 2010), método de refinamento interativo L-NS-i e penalidades de *Gap opening* 1,53 e *Offset value* nulo. Adicionalmente, em proteínas foram utilizados os alinhamentos estruturais dos domínios conservados em sequências proteicas (Fig.5), retiradas do CDD (*Conserved Domains Database* - http://www.ncbi.nlm.nih.gov/Structure/cdd) (ref) das subfamílias de cada umas das enzimas (Tabela 1).

Em seguida, ambos os tipos de sequências (nucleotídeos e aminoácidos) foram realinhadas com auxílio do programa Geneious R7 (v.7.1.5; Biomatters©) e seu próprio algoritmo de alinhamento múltiplo padrão. Todas foram alinhadas de acordo com matrizes de substituição de aminoácidos e nucleotídeos adequadas às distâncias evolutivas (Matriz BLOSUM62 ou PAM 200) (Henikoff & Henikoff, 1992; Altschul, 1991). Durante a análise da qualidade dos dados obtidos por meio dos alinhamentos finais, por se tratar de uma etapa crucial na verificação da informação filogenética neles contida, precauções adicionais foram tomadas no tratamento de possíveis *gaps* indesejados oriundos de sequencias conservadas globalmente, mas possuindo seguimentos não segmentos independentes não condizentes com os reais eventos de inserção e remoção (*indels*).

As análises de saturação nas substituições nucleotídicas foram realizadas a partir de cada alinhamento múltiplo final com o auxílio do programa DAMBE (v.5.2.31; Xia; Xie, 2001). Para determinar qual modelo evolutivo de substituição mais adequado a cada conjunto de dados, foram utilizados os programas jModelTest 2 (v.2.1.4; Darriba *et al.*, 2012) e ProtTest 3 (v.3.2; Darriba *et al.,* 2011) na construção da matriz de distância, levando em consideração a heterogeneidade nas taxas de substituição ao longo dos sítios pelo uso da distribuição Gamma (+G), proporção de sítios invariáveis (+I), ambos (+I+G) e em alguns casos a frequência (+F) (de bases nucleotídicas e ou de aminoácidos). Totalizando cerca de 120 variações de modelos possíveis para proteínas e 88 para nucleotídeos. Ainda nesta etapa, com o auxílio dos programas Geneious R7 e MEGA6 (v.6.05; Tamura *et al*., 2013) foram calculadas as matrizes de substituição, determinada a composição de aminoácidos e nucleotídeos, e o conteúdo G+C do conjunto de sequências.

Tabela 1 – Identificação dos domínios conservados das enzimas malato sintase (MLS) e isocitrato liase (ICL), utilizados nos múltiplos alinhamentos.

Família	Subfamília	Identificado	Nº Seqs.
Malate_synt		cd00480	
	malate_synt_A	cd00727	28
	malate_synt_G	cd00728	10
ICL_KPHMT		cd06556	
	ICL_PEPM	cd00377	40
	KPHMT-like	cd06557	100

Fonte: *CDD - conserved domains and protein three-dimensional structure, 2014.*

3.4 Análises filogenéticas e Teste de Seleção

A reconstrução filogenética prévia envolveu métodos de distância (NJ) e caracteres discretos (MP) que foram feitas com o auxílio do programa MEGA6 (*Molecular Evolutionary Genetic Analysis*), utilizando o teste de *Bootstrap* com 500 repetições como medida de suporte. Isto foi realizado para verificar a consistência e confiança nas topologias obtidas, mas levando em consideração os parâmetros exclusivos de cada conjunto de dados.

Para construção das árvores de Máxima Parcimônia optou-se pelo método de busca heurístico TBR (*Tree Bisection and Reconnection*) capaz de reduzir o número de topologias procuradas sem que haja necessidade de uma busca exaustiva, e para o *Neighbor-joining* foi levado em consideração o modelo de substituição adequando a cada conjunto de dados. Nas duas análises optou-se por remover do cálculo de reconstrução todos os sítios que continham lacunas de alinhamentos e falta de informação (*Gaps/Missing datas*).

Em seguida, análises com o método de Máxima Verossimilhança foram realizadas duas análises para cada alinhamento de aminoácidos, utilizando modelos de substituição evolutivos distintos na reconstrução das topologias, utilizando os programas FastTree (v.2.1.5; Price *et al*., 2010) e PhyML 3 (v.3.1; Guindon *et al*.,

2010), ambos executados por meio do software Geneious R7. Nessa análise os parâmentos de construção das árvores de ambas as enzimas com o FastTree foram mantidos, por padrão a opção Fastest e otimização dos valores de verossimilhança com correção *Gamma*. No PhyML 3, por se tratar de um programa com algoritmo mais robusto, pôde-se calcular os valores de suporte aos ramos por *Bootstrap*, com 500 repetições, otimização no tamanho dos ramos das topologias e método de busca heurístico por NNI (*Nearest Neighbor Interchange*). A escolha dos modelos de substituição, valores de +G e proporção de sítios invariáveis foram efetuadas de acordo com os resultados obtidos na etapa anterior de testes dos modelos.

A Inferência Bayesiana foi conduzida a partir do pacote de programas BEAST (*Bayesian Evolutionary Analysis Sampling Trees*) (v.1.8.0; Drummond & Rambaut, 2007), utilizando duas corridas independentes, contendo quatro cadeias simultâneas de 10^6 gerações e amostragem a cada 100 gerações, levando em consideração nos modelos a distribuição *Gamma* (+G) na correção dos valores de alpha *a priori*. Os arquivos obtidos contendo os parâmetros utilizados ao longo da execução do programa foram utilizados como referência pela ferramenta TRACER (v.1.6; Drummond; Rambaut, 2007), para verificar a credibilidade e estabilidade das cadeias.

Em seguida, as árvores consenso foram calculadas aplicando um *burnin*[1] de dez por cento. Os arquivos com dendogramas gerados foram visualizados e editados pelo programa FigTree (v.1.3.1; Drummond; Rambaut, 2007). Adicionalmente foram implementados testde seleção para cada gene foi feito no MEGA6, pelo teste Z empregando o modelo de proporção de Nei-Gojobori (Nei; Gojobori, 1986), no qual é calculado o número de substituições sinônimas e não-sinônimas e nº de sítios potencialmente sinônimos e não-sinônimos.

[1] Burnin: Prática de descartar porção inicial da amostra da cadeia de Markov, para que os efeitos dos valores iniciais da reconstrução não interfiram na inferência *a posteriori.*

3.5 Modelagem de Proteínas

As sequências de aminoácidos obtidas e resultantes do múltiplo alinhamento de cada enzima foram submetidas a anotação e predição no Geneious R7 seguida da modelagem estrutural por homologia, por meio do servidor online do Phyre2 (*Protein Homology/analogY Recognition Engine* - http://www.sbg.bio.ic.ac.uk/phyre2) (v.2; Lawrence *et al.*, 2011). Para cada sequência de ICL e MLS, 10 modelos foram gerados a partir de 8 estruturas conhecidas (3CUX, 3CUZ, 3OYZ, 3CV1, 3CV2, 3ERB, 3EOL, 3POX) depositados no PDB (*Protein Data Bank* - http://www.rcsb.org/pdb), porém retornando apenas o seu melhor modelo teórico. Em seguida, a pontuação DOPE (Shen; Sali 2006), gráficos de *Ramachandran*, sobreposições estéricos, parâmetros de desvio do carbono beta *(*Cβ), qualidade dos rotâmeros e de interações fracas, para cada modelo foram estimados utilizando o MolProbity Server (Chen *et al*. 2010) e seus escore Z, escores QMEAN6 utilizando o servidor online do SWISS-MODEL (Arnold *et al*. 2006). Os modelos obtidos foram gerados em formato PDB e visualizados utilizando o UCSF Chimera (Pettersen *et al*. 2004). Após avaliação da qualidade dos resultados obtidos na predição, cinco modelos de cada enzima foram selecionados entre as espécies de cada representante das ordens/táxons das plantas verdes.

De posse das dez estruturas preditas, seus PDBs foram submetidos ao Servidor ConSurf (http://consurf.tau.ac.il/) (Ashkenazy *et al*., 2010) comparadas aos bancos de sequência UniRef (Suzek *et al*., 2007) para se identificar regiões funcionais por mapeamento de superfície de informação filogenética das proteínas. Este utiliza como critérios de investigação métodos de reconstrução filogenética por Máxima verossimilhança e Inferência bayesiana. Como resultado desta busca, os sítios conservados no múltiplo alinhamento entre as estruturas foram visualizados com ajuda do programa UCSF Chimera (v.1.9; Pettersen *et al.*, 2004).

Os modelos com os melhores escores foram sujeitos a uma etapa adicional para minimizar a energia da estrutura, onde foi utilizada a dinâmica molecular. Esta fase foi realizada utilizando Campo de Força Amber-99SB* (Hornaket *al.*, 2006) e o workflow preestabelecido “Gromacs FULL MD Setup” do Servidor MDWeb (Hospital *et al.,* 2012), neste servidor será utilizada a versão 4.0.2 do programa GROMACS (Hesset *al.*, 2008).

4. RESULTADOS

4.1 Sequências obtidas e selecionadas

Para primeira parte deste estudo, como resultado da busca feita nos bancos de dados biológicos do *NCBI* e *UniProt*, foram encontradas entre hipotéticas, *putative* (possíveis) e previamente identificas um total de quarenta e seis sequências de espécies de *Viridiplantae* do gene da enzima Malato Sintase e setenta e cinco sequências diferentes para Isocitrato Liase (disponíveis em 22/05/2014). Entre o conjunto completo de sequências encontradas, foram selecionadas apenas aquelas com representantes dos genomas vegetais completamente sequenciados. Porém, para se evitar a presença de sequências de genes parálogos, apenas vinte sequências diferentes para MLS e vinte e três para ICL (Tabelas 2 e 3) foram utilizadas como referências evidenciadas experimentalmente.

4.2 Alinhamento Múltiplo entre sequências e Teste do Modelo

O múltiplo alinhamento do conjunto de sequências demostrou a presença de inserções, remoções e mutações pontuais por substituições, sedo todos os tipos nas sequências nucleotídicas (transição e/ou transversão). Entre alguns indivíduos houve regiões de deleção, com abertura e extensão de *gaps*, o que gerou uma variação no tamanho das sequências. Embora isso tenha ocorrido, conforme o esperado os alinhamentos entre sequências de aminoácidos se mostraram relativamente mais conservados do que as sequencias nucleotídicas, contendo em ICL uma identidade entre pares de 68,4 à 73,9% e de seus sitos de até 17%, e em MLS de 68 à 69,3% com até 27,4%. Nessas enzimas, considerando os possíveis pareamentos positivos da matriz (BLOSUM62) seus percentuais chegam a 75,8% e 78%, respectivamente.

Ao se realizar a predição de estruturas associadas aos alinhamentos de organismos elencados como represente dos diferentes *taxa* das plantas verdes (Fig. 5 e 6), pôde-se constatar que regiões das proteínas tendem a apresentar conservação na conformação de locais específicos da estrutura, ao longo de sua evolução. Há regiões de beta folha e rotação do alinhamento associadas a montagem da estrutura do TIM Barrel, como sítio catalítico das enzimas. Sendo já perceptível a propensão

dos agrupamentos formados pertencerem ao mesmo táxon num determinado nível filogenético. A conservação de regiões das sequências de ambas as enzimas pode ser ainda verificada pela tendência nas taxas estimadas na matriz de substituição (Tabela 4 e 5), onde a frequência do número de transversões entre os sítios é baixa. O que também pode ser observado pela estimativa da razão (*R*) transição/transversão entre sequências nucleotídicas, obtidas com auxílio do programa MEGA 6, foi de 1,37 e 1,39 para ICL e MLS, respectivamente.

Tabela 2 – Relação das sequências do gene *mls* obtidos nos bancos de dados biológicos, mostrando seus respectivos conteúdos GC (Guanina + Citosina).

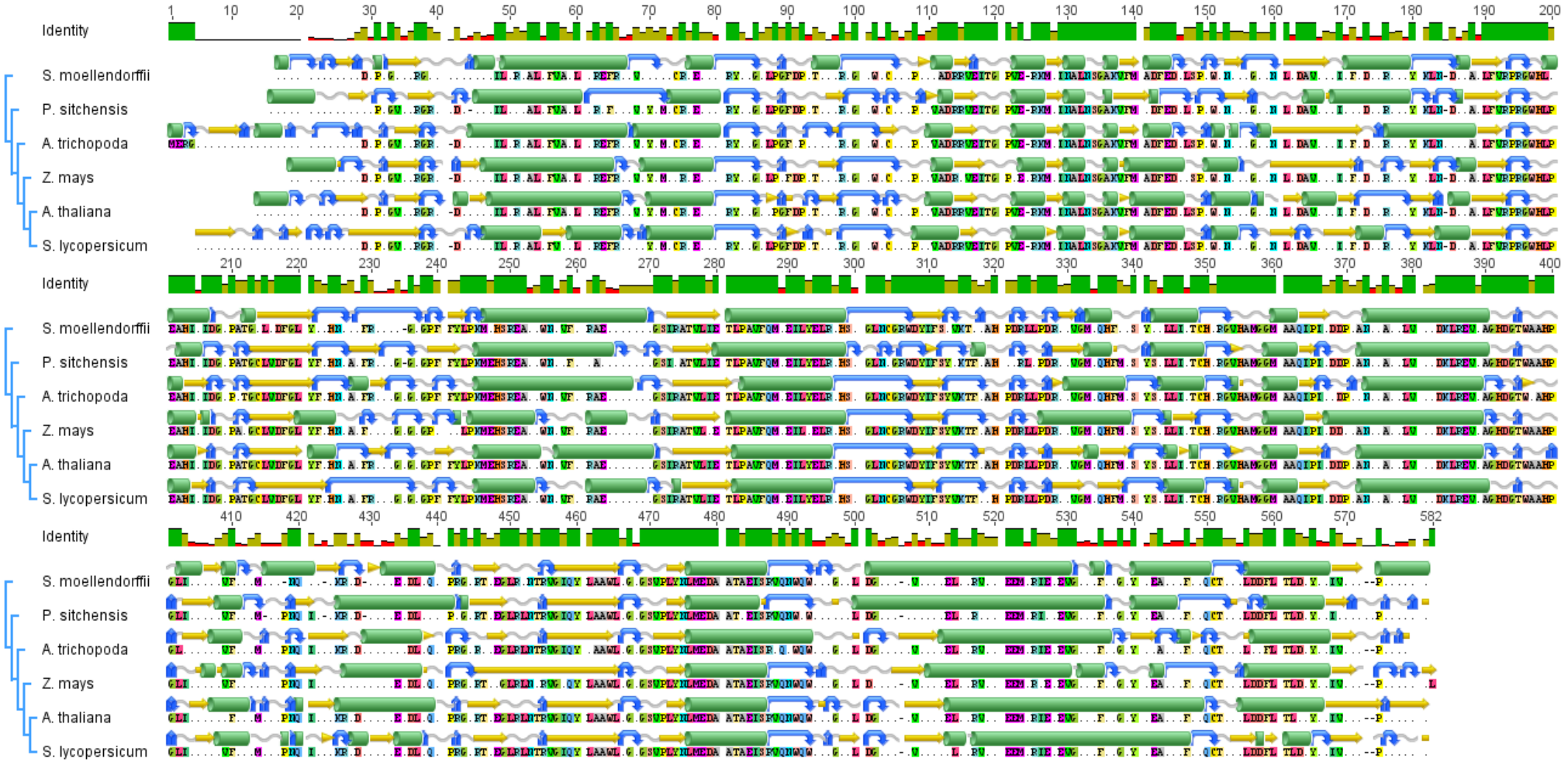
Táxon/Ordem	**Nome da Espécies**	***Acesso (GI)***	***Conteúdo GC (%)***
Amborellales	*Amborella trichopoda*	586752138	50,8
Brassicales	*Arabidopsis thaliana*	15237551	46,1
Brassicales	*Brassica napus*	126766	46,4
Brassicales	*Raphanus sativus*	6225656	46,3
Chlamydomonadales	*Chlamydomonas reinhardti*	75130149	64,1
Chlamydomonadales	*Volvox carteri*	302853993	61,0
Cucurbitales	*Cucumis sativus*	126768	47,3
Cucurbitales	*Curcubita maxima*	126767	47,1
Fabales	*Glycine max*	1170878	46,0
Fabales	*Glycine max*	356563182	46,0
Malpighiales	*Ricinus communis*	126773	45,3
Malvales	*Gossypium hirsutum*	126770	42,9
Pinales	*Picea sitchensis*	148909186	45,3
Poales	*Oryza sativa*	21741699	72,6
Poales	*Sorghum bicolor*	242076294	67,2
Poales	*Zea mays*	1346487	67,2
Selaginellales	*Selaginella moellendorffii*	302767316	62,7
Solanales	*Physcomitrella patens*	168014816	55,8
Solanales	*Solanum lycopersicum*	460381210	43,9
Solanales	*Solanum tuberosum*	565369726	44,4

Tabela 3 – Relação das sequências do gene *icl* obtidos nos bancos de dados biológicos, mostrando seus respectivos conteúdos GC (Guanina + Citosina).

Táxon/Ordem	Nome da Espécies	*Acesso (GI)*	*Conteúdo GC (%)*
Amborellales	*Amborella trichopoda*	586642755	50,5
Asparagales	*Dendrobium crumenatum*	11131348	53,3
Asterales	*Helianthus annuus*	113030	47,7
Brassicales	*Arabidopsis thaliana*	30686361	49,4
Brassicales	*Brassica napus*	113026	48,6
Chlamydomonadales	*Chlamydomonas reinhardti*	159474436	62,9
Chlamydomonadales	*Volvox carteri f. nagariensis*	302854455	55,2
Cucurbitales	*Cucumis sativus*	449490272	45,5
Cucurbitales	*Curcubita máxima*	8134299	49,2
Fabales	*Glycine max*	358248362	47,1
Fabales	*Glycine max*	356542840	45,3
Malpighiales	*Ricinus communis*	113032	43,5
Poales	*Aegilops tauschii*	475510959	61,0
Malvales	*Gossypium hirsutum*	113029	46,7
Pinales	*Pinus taeda*	3831487	46,7
Poales	*Oryza sativa Japonica Group*	115472481	61,4
Poales	*Oryza sativa Indica Group*	218199746	60,3
Poales	*Sorghum bicolor*	242050412	64,8
Poales	*Zea mays*	194702636	65,7
Selaginellales	*Selaginella moellendorffii*	302798645	60,1
Funariales	*Physcomitrella patens*	168037648	65,5
Solanales	*Solanum lycopersicum*	350539375	45,1
Solanales	*Solanum tuberosum*	565403667	42,8

Contudo, ao se examinar as taxas de divergência evolutiva dos aminoácidos (Tabela 6 e 7) pertencentes a cada representante de *taxa* distintos, ao contrário do que era esperado, há a ausência na correlação das distâncias seguirem um padrão decrescente de valores, de acordo com os agrupamentos do sistema de classificação atual dos vegetais.

Figura 5: Alinhamento e predição das estruturas tridimensionais das sequências de Malato sintase de cinco representantes dos genomas vegetais utilizados na análise filogenética.



Legenda: No alinhamento se encontram visíveis apenas os resíduos de aminoácidos conservados em todas as sequências; *Identidade*: Escala cores para similaridade entre pares. Resíduos 100% conservados (Verde), resíduos de 30 a 100% conservados (esverdeado a marrom), e abaixo de 10% (vermelho); *Estruturas*: [ ] *alpha helix*; [ ] espiral; [ ] beta folha; e [ ] rotação.

Figura 6: Alinhamento e predição das estruturas tridimensionais das sequências de Isocitrato liase de cinco representantes dos genomas vegetais utilizados na análise filogenética.

Legenda: No alinhamento se encontram visíveis apenas os resíduos de aminoácidos conservados em todas as sequências; *Identidade*: Escala cores para similaridade entre pares. Resíduos 100% conservados (Verde), resíduos de 30 a 100% conservados (esverdeado a marrom), e abaixo de 10% (vermelho); *Estruturas*: [ ] *alpha helix*; [ ] espiral; [ ] beta folha; e [ ] rotação.

Tabela 4 – Matriz de substituição das frequências de nucleotídeos entre as 23 sequências de Isocitrato liase, obtidas de espécies de plantas verdes

	A	T/U	C	G
A	-	*5.85*	*5.92*	**12.05**
T/U	*5.76*	-	**16.58**	*3.57*
C	*6.81*	**19.39**	-	*6.36*
G	**10.06**	*3.03*	*4.62*	-

Tabela 5 – Matriz de substituição das frequências de nucleotídeos entre as 20 sequências de Malato sintase, obtidas de espécies de plantas verdes.

	A	T/U	C	G
A	-	*4.54*	*6.58*	**12.57**
T/U	*5.18*	-	**15.76**	*6.72*
C	*6.58*	**13.80**	-	*7.03*
G	**10.49**	*4.90*	*5.86*	-

Na etapa seguinte de avaliação da qualidade dos dados e sinal filogenético, o teste de saturação das substituições demonstrou a presença de uma tendência gradativa à saturação entre as distâncias e as taxas de transição x transversão nos alinhamentos, descrito melhor no item 4.3 a seguir. Posteriomente, as análises do melhor modelo de ajuste à substituição nucleotídica utilizando o jModelTest, resultou no modelo *General Time Reversible* (GTR) (Tavaré, 1986) para ambos os genes. A proporção de sítios invariáveis e parâmetro de distribuição *Gamma* foram calculados para *icl* e *mls*, porém apenas os valores de (+G) foram incluídos nas análises posteriores, sendo eles 0,864 e 0,515, respectivamente.

Tabela 6 – Divergência evolutiva (*p-distance*) entre sequências de Isocitrato liase de representantes dos genomas vegetais.

	Amborella trichopoda	*Arabidopsis thaliana*	*Pinus taeda*	*Selaginella moellendorffii*	*Solanum lycopersicum*	*Zea mays*
Amborella trichopoda	—					
Arabidopsis thaliana	**0,205**	—				
Pinus taeda	0,223	0,228	—			
Selaginella moellendorffii	0,324	0,317	0,297	—		
Solanum lycopersicum	0,203	0,187	**0,210**	0,326	—	
Zea mays	0,239	**0,214**	0,252	0,326	**0,214**	—

Nota: Os valores correspondem ao número de diferenças de aminoácidos por sítio entre sequências no alinhamento. Em destaque, valores que indicam alta similaridade entre espécies, embora estejam separados por cerca de 100 à 300 milhões de anos.

Na estimativa do melhor modelo de substituição para proteinas, pelo uso do ProtTest 3, obteve-se o modelo LG (Le; Gascue, 2008) para o conjunto de sequências de Isocitrato Liase e *Jones-Taylor-Thornton* (JTT) (Jones *et al.*, 1992) para Malato sintase. Novamente, optou-se pelo uso apenas da correção *Gamma*, sendo seus valores fixados nas análises em 0,564 e 0,752, respectivamente.

Tabela 7 – Divergência evolutiva (*p-distance*) entre sequências de Malato sintase de representantes dos genomas vegetais.

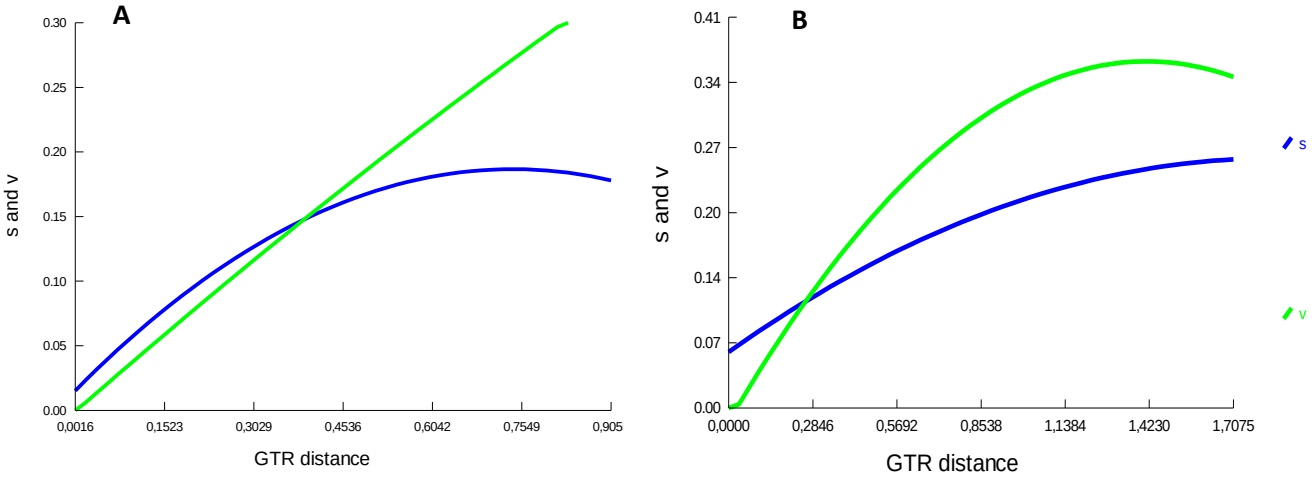
	Amborella trichopoda	*Arabidopsis thaliana*	*Picea sitchensis*	*Selaginella moellendorffii*	*Solanum lycopersicum*	*Zea mays*
Amborella trichopoda	—					
Arabidopsis thaliana	**0,257**	—				
Picea sitchensis	0,313	0,302	—			
Selaginella moellendorffii	0,313	0,304	0,293	—		
Solanum lycopersicum	0,268	0,175	0,304	**0,288**	—	
Zea mays	0,302	**0,253**	0,361	0,344	**0,260**	—

Nota: Os valores correspondem ao número de diferenças de aminoácidos por sítio entre sequências no alinhamento. Em destaque, valores que indicam alta similaridade entre espécies, embora estejam separados por cerca de 100 à 300 milhões de anos.

4.3 Análise de Saturação das Substituições

A análise de saturação de substituições, realizadas com auxílio do programa DAMBE (*Data Analysis in Molecular Biology and Evolution*) demonstrou que para ambos os conjuntos de dados de alinhamentos de sequências nucleotídicas houve saturação nas transições (S), conforme observado anteriormente nas tabelas de matrizes de frequência de substituição. A representação gráfica das distâncias entre sequências, pelo modelo GTR, contra as taxas de substituição do tipo transversões (V) tende a ser uma reta até dos valores utilizados de (Fig. 7), ratificando a presença do sinal filogenético e possível uso destas sequências para estudos de filogenia e evolução molecular de seus produtos protéicos ao nível taxonômico das *Viridiplantae*.

Figura 7: Representação gráfica das taxas de transições e transversões *versus* divergência evolutiva entre as sequências obtidas dos genes das enzimas, utilizando modelo GTR.



Legenda: (A) gráfico dos valores de MLS; (B) gráfico dos valores de ICL. Em azul, transições (s) e em verde transversões (v).

4.4 Inferências Filogenéticas

Para a análise filogenética com o programa MEGA6 foram utilizadas as vinte sequências das enzimas MLS e vinte e três de ICL, que apresentassem espécies vegetais comuns, sendo ao final da reconstrução definido como *outgroup* a espécie de alga verde *Volvox carteri* F. Stein (1873) (Gi 302854455). Os dendogramas consenso foram calculados com base na distância (*Neighbor-Joining*) e métodos discretos (Máxima Parcimônia), utilizando seus modelos de substituição adequados, respectivos valores de distribuição inferidos e parâmetros citados anteriormente. Os

testes de *Bootstrap* foram realizados para verificar a consistência topológica e confiança dos dados. Como resultado desta análise prévia, os dendogramas consenso obtidos para ambas as enzimas, apresentaram permutas no agrupamento de táxons em suas topologias, tendo seus índices de *Bootstrap* acima de 88 e 92 para o nó correspondente ao núcleo das eudicotiledôneas (*Eudicots Core*), porém com algumas incongruências taxonômicas nos ramos mais basais desse grupo, quando comparada com a classificação taxonômica mais atual de APG III (*Angiosperm Philogeny Group*, 2003) (Anexo).

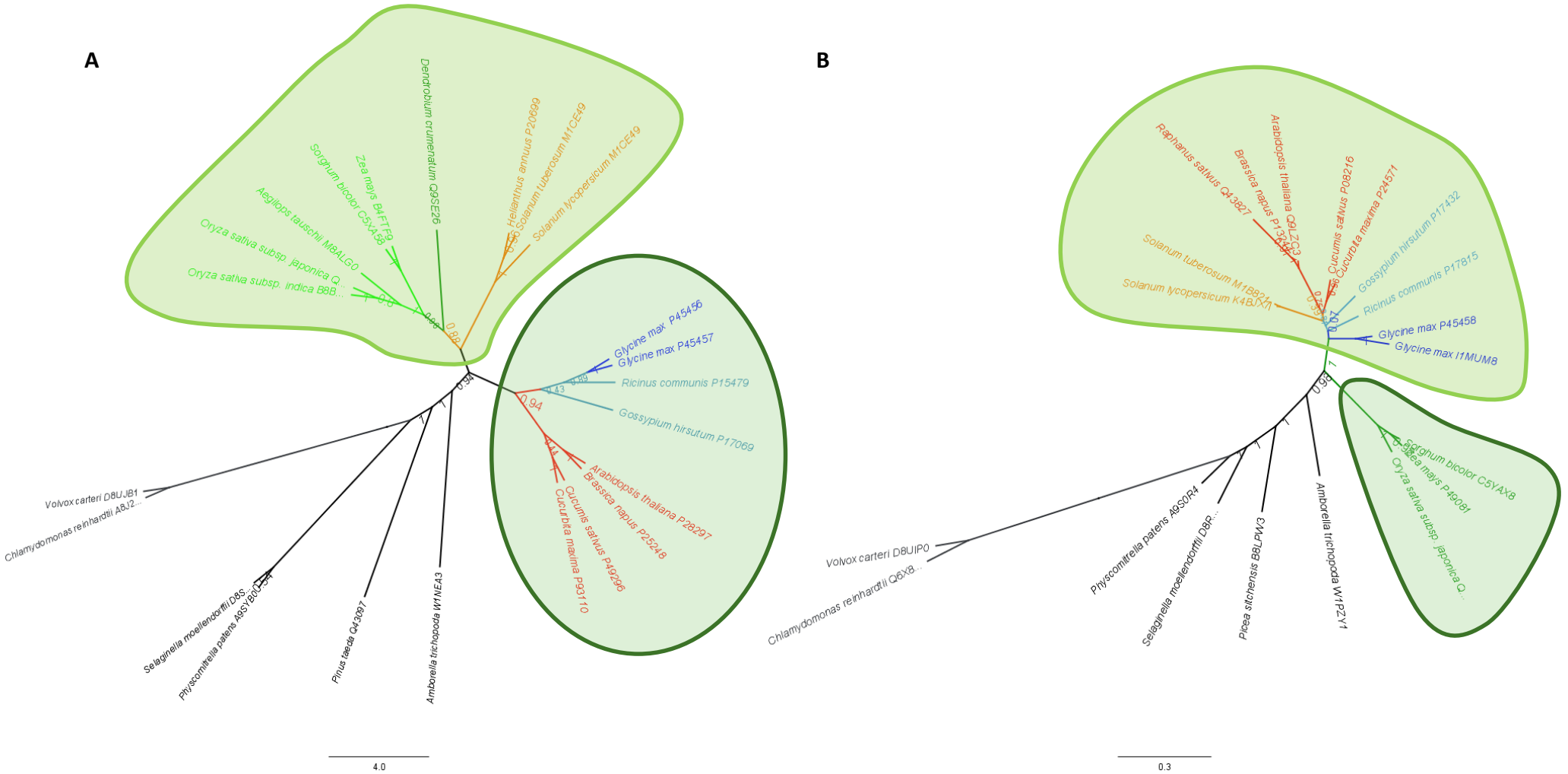
Embora os dendogramas obtidos por esses métodos apresentem valor filogenético, por se tratar de um estudo molecular de evolução das estruturas dessas proteinas seus critérios de reconstrução não levam em consideração todos os fatores e eventos presentes ao longo da diferenciação entre genomas e surgimento das espécies pertencente aos vegetais.

Sendo assim, para inferir as relações evolutivos entre as enzimas optou-se pelo uso apenas das topologias resultantes dos métodos baseados em caracteres discretos, que apresentem correção pela utilização do modelo evolutivo mais adequado aos tipos de mudanças na conservação de seus sítios e manutenção de sua estrutura para atividade catalítica. Para tanto, as árvores obtidas por meio das sequências de aminoácidos, foram inferidas pelos métodos de Máxima verossimilhança e Inferência Bayesiana.

As topologias acançadas com a ML se revelam condizentes com a realidade evolutiva dos genomas vegetais entre as plantas verdes, o que é suportado pelos seus valores de *Bootstrap* e comprimento dos ramos (Fig. 8) (Anexos). Além disso, o agrupamento entre um maior número de espércies em MLS demonstra uma tendência em terem uma maior taxa de conservação dos sitios.

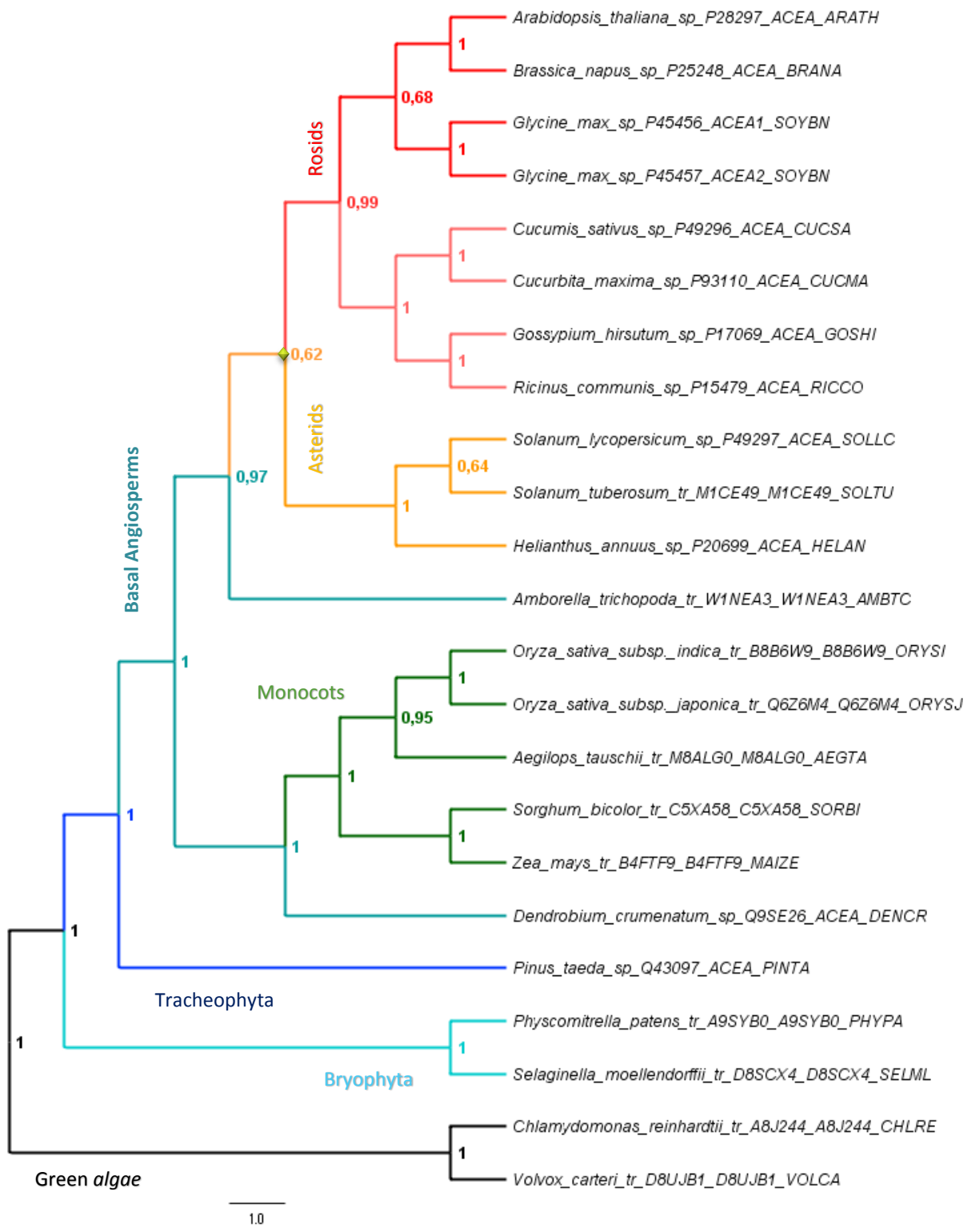
Para inferência bayesiana foram utilizadas os dois conjunto de sequências de aminoácidos de cada enzima para construção das árvores consensos. As árvores apresentaram valores máximo (1.0) de PP (probabilidade *a posteriori*) em quase todos os ramos em ambos os conjuntos de dados (Fig. 9, 10, 11 e 12), exceto pelos valores abaixo de 0.65 em ICL e 0.52 dos nós de separação entre as ordens das angiospermas, dentro do núcleo das Eudicotiledôneas.

Figura 8: Árvores filogenéticas das enzimas *Isocitrato liase* e *Malato sintase* por Máxima Verossimilhança.



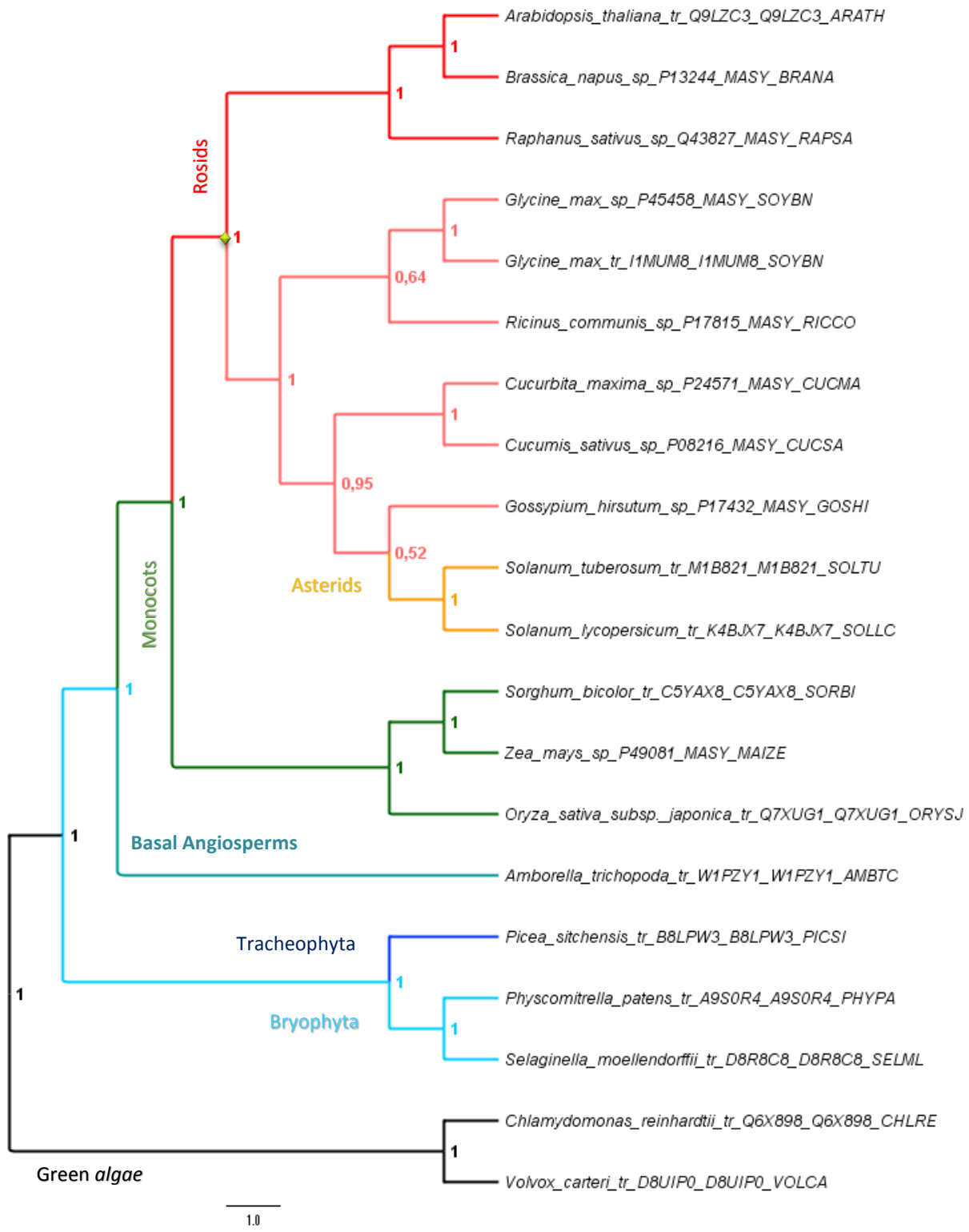
Legenda: Topologias obtidas utilizando os modelos de substituição LG+G para Isocitrato liase (a) e JTT+G para Malato sintase (b), no programa FastTree. Em destaque, agrupamentos associados a conservação na estrutura das enzimas. Nº: Valores de suporte dos ramos. **Esquema de cores dos ramos:** Organização das espécies em Classes correspondentes a classificação atual de APG III.

Figura 9: Árvore filogenética obtida por Inferência Bayesiana de sequências selecionadas da enzima *Isocitrato liase*.



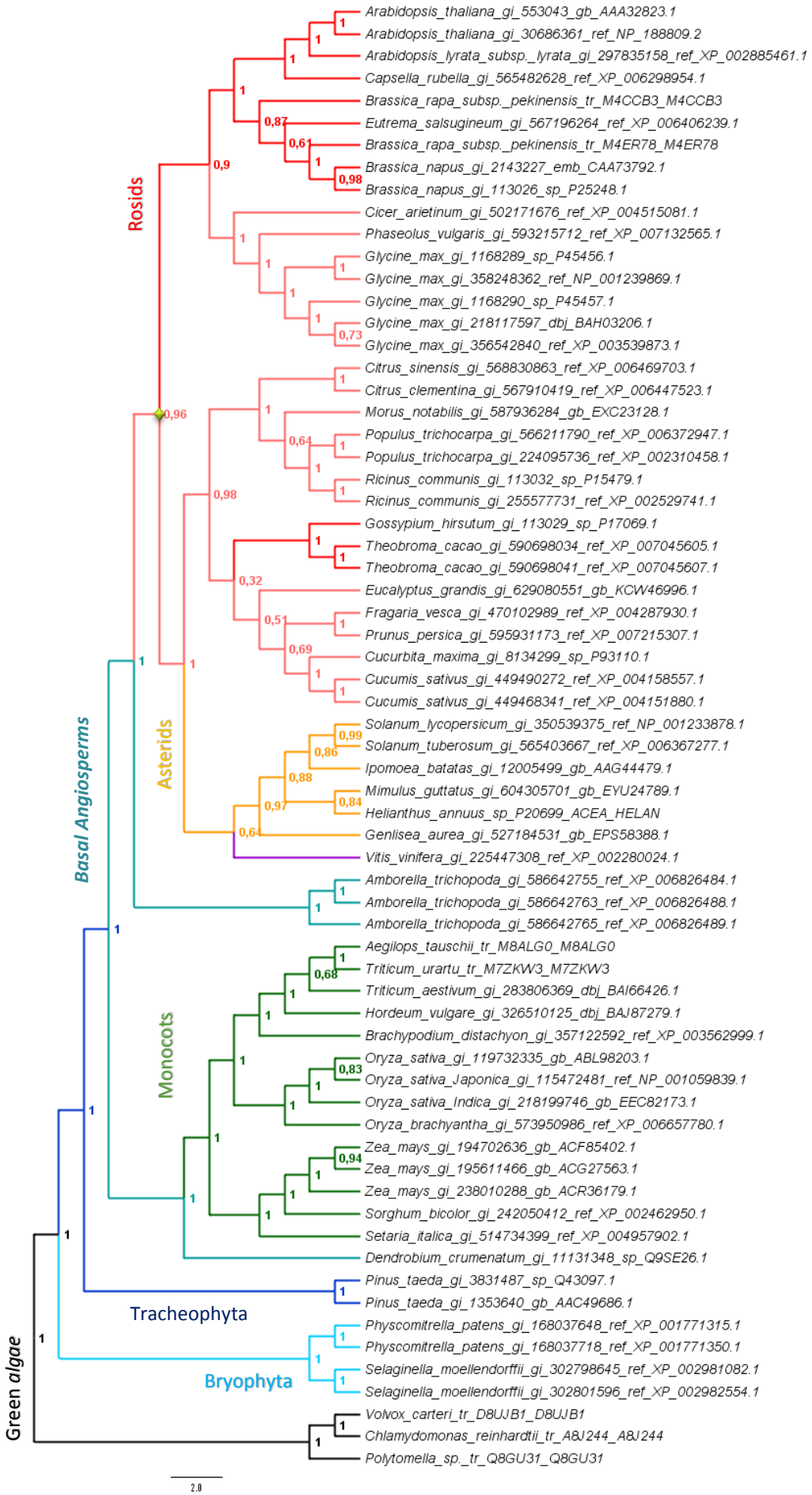
Legenda: Filogenia molecular obtida utilizando o modelo de substituição LG+G para conjunto de 23 sequências de aminoácido da Isocitrato liase. Valores de PP dos ramos exibidos. **Esquema de cores:** corresponde a separação de Divisão/Classe na classificação atual de APG III.

Figura 10: Árvore filogenética obtida por Inferência Bayesiana de sequências selecionadas da enzima *Malato sintase*.



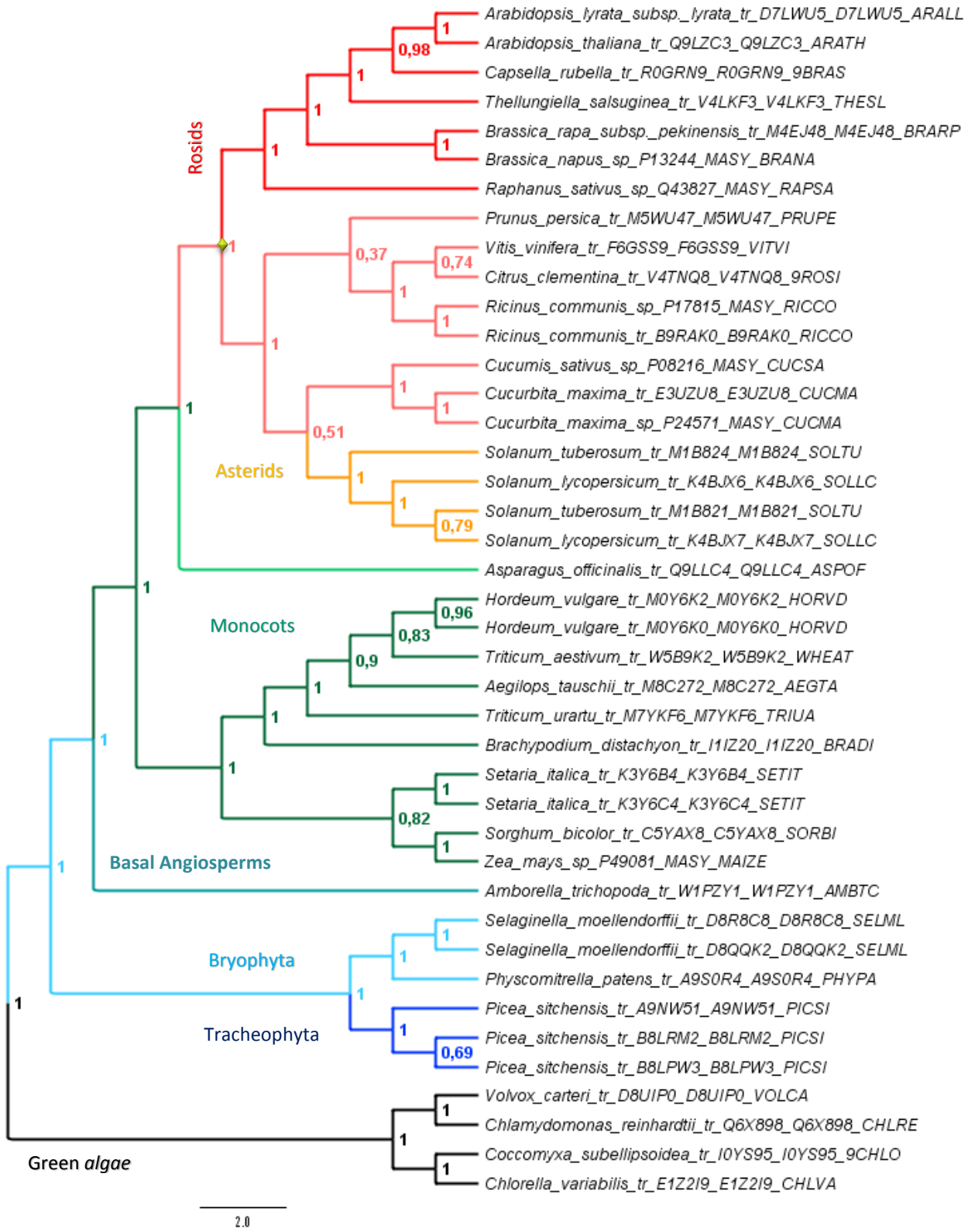
Legenda: Filogenia molecular obtida utilizando o modelo de substituição JTT+G para conjunto de 20 sequências de aminoácido da Isocitrato liase. Valores de PP dos ramos exibidos. **Esquema de cores:** corresponde a separação de Divisão/Classe na classificação atual de APG III.

Figura 11: Árvore filogenética obtida por Inferência Bayesiana de todas as sequências da enzima *Isocitrato liase*.



Legenda: Filogenia molecular obtida utilizando o modelo de substituição LG+G para o conjunto de 66 sequências de aminoácido da Isocitrato liase. Nº exibidos: Valores de PP. **Esquema de cores:** corresponde a separação de Divisão/Classe na classificação atual de APG III.

Figura 12: Árvore filogenética obtida por Inferência Bayesiana de todas as sequências da enzima *Malato sintase*.



Legenda: Filogenia molecular obtida utilizando o modelo de substituição JTT+G para o conjunto de 41 sequências de aminoácido da Malato sintase. Nº exibidos: Valores de PP. **Esquema de cores:** corresponde a separação de Divisão/Classe na classificação atual de APG III.

4.5 Teste de Seleção e Evolução

O teste de hipótese nula para os três tipos de seleção (positiva, neutra e purificadora) foram aplicados sobre o conjunto de sequências nucleotídicas dos genes de ambas as enzimas. Na tabela abaixo, a probabilidade de rejeição da hipótese nula de estrita neutralidade (dN = dS) em favor da hipótese alternativa é mostrada para cada gene na primeira coluna, seguidas dos resultados da diferença entre o número de substituições sinônimas e não-sinônimas (dN – dS), por sítio do alinhamento.

Tabela 8 – Resultados do teste de seleção (teste Z) na análise das médias entre todos os pares de sequências dos dois genes.

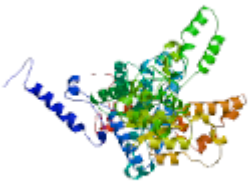
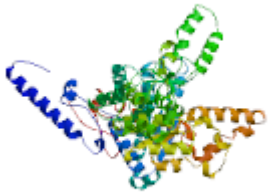
Hipótese alternativa	**Gene *icl***		**Genes de *mls***	
	P	(dN – dS)	*P*	(dN – dS)
Seleção Positiva (dN > dS)	**0,002**	2,980	1,000	-18,067
Seleção Neutra (dN = / = dS)	0,004	2,897	0,000	-19,053
Seleção Purificadora (dN < dS)	1,000	-2,892	**0,000**	18,057

Nota: Os valores de P inferior a 0,05 são considerados significativos ao nível de 5%. A análise envolveu 20 sequências de nucleotídeos. Todas as posições que continham lacunas e dados ausentes foram eliminados. Análises evolutivas foram realizadas em MEGA6 com o método de Nei-Gojobori.

4.6 Predição dos Modelos de Proteínas

Todos os modelos gerados pelo Phyre2 e correspondentes valores obtidos no MolProbity e Swiss-Model, utilizados na avaliação da qualidade de predição suas estruturas tridimensionais teóricas, foram sumarizados nas tabelas a seguir. Por apresentar uma certa conservação filogenética entres sequências (observada em análises prévias) seus modelos indicaram uma tendência a exibirem a conservação também em sua conformação. Dentre os escores obtidos para ambas as enzimas, pelo MolProbity, o *Clashscore* revelou valores abaixo de 0,75 em todos. No Swiss-Model, os modelos avaliados de Malato sintase se mostraram muito bons, por apresentarem em todos o Z-score e QMEANscore6 altos, que ao resultarem em valores próximos a 1 indicam uma melhor qualidade nas estruturas obtidas.

Tabela 9 – Modelos teóricos de *Isocitrato liase* de espécies representantes das *Viridiplantae* e respectivos valores de escores, avaliados pelo MolProbity e QMEAN6.

Modelo	Escores	
		Amborella trichopoda
	Clashscore	66.77
	Rotâmeros das cadeias laterais desfavoráveis	2.89%
	Ramachandran desfavorável	3.16%
	Ramachandran favorável	92.11%
	Desvio do carbono beta	0.76%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.43%
	MolProbity Score	3.13
	Z-Score	-2.075
	QMEANscore6	0.576
		Arabidopsis thaliana
	Clashscore	72.72
	Rotâmeros das cadeias laterais desfavoráveis	2.54%
	Ramachandran desfavorável	2.26%
	Ramachandran favorável	92.16%
	Desvio do carbono beta	1.50%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.47%
	MolProbity Score	3.12
	Z-Score	-2.357
	QMEANscore6	0.55
		Chlamydomonas reinhardtii
	Clashscore	50.36
	Rotâmeros das cadeias laterais desfavoráveis	3.99%
	Ramachandran desfavorável	1.20%
	Ramachandran favorável	94.94%
	Desvio do carbono beta	1.31%
	Ligações com comprimentos ruins	0.03%
	Ligações com ângulos ruins	1.52%
	MolProbity Score	2.98
	Z-Score	-1.909
	QMEANscore6	0.604

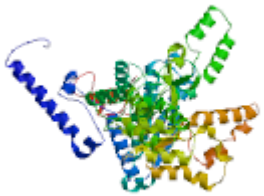
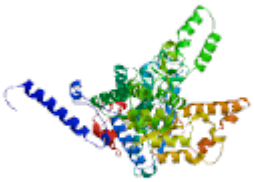
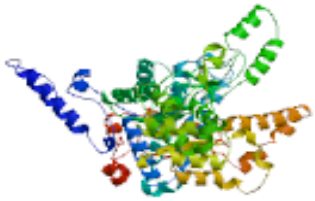
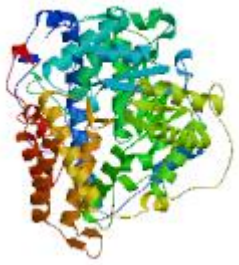
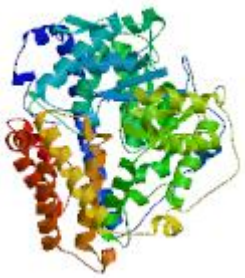
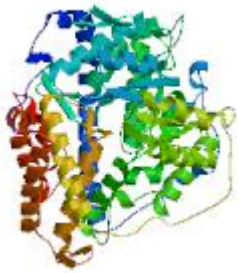
Modelo	Escores	
		Pinus taeda
	Clashscore	58.32
	Rotâmeros das cadeias laterais desfavoráveis	2.32%
	Ramachandran desfavorável	3.29%
	Ramachandran favorável	91.35%
	Desvio do carbono beta	0.37%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.44%
	MolProbity Score	3.02
	Z-Score	-2.434
	QMEANscore6	0.548
		Selaginella moellendorffii
	Clashscore	63.58
	Rotâmeros das cadeias laterais desfavoráveis	1.97%
	Ramachandran desfavorável	2.48%
	Ramachandran favorável	90.43%
	Desvio do carbono beta	0.56%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.47%
	MolProbity Score	3.04
	Z-Score	-1.699
	QMEANscore6	0.608
		Solanum lycopersicum
	Clashscore	67.44
	Rotâmeros das cadeias laterais desfavoráveis	1.89%
	Ramachandran desfavorável	2.62%
	Ramachandran favorável	91.97%
	Desvio do carbono beta	0.56%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.47%
	MolProbity Score	3
	Z-Score	-1.77
	QMEANscore6	0.602

Tabela 10 – Modelos teóricos de *Malato sintase* de espécies representantes das *Viridiplantae* e respectivos valores de escores, avaliados pelo MolProbity e QMEAN6.

Modelo	Escores	
		Amborella trichopoda
	Clashscore	55.16
	Rotâmeros das cadeias laterais desfavoráveis	3,81%
	Ramachandran desfavorável	1.57%
	Ramachandran favorável	93.19%
	Desvio do carbono beta	1.33%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.46%
	MolProbity Score	3.10
	Z-Score	-0.38
	QMEANscore6	0.725
		Arabidopsis thaliana
	Clashscore	60.82
	Rotâmeros das cadeias laterais desfavoráveis	1.48%
	Ramachandran desfavorável	0.54%
	Ramachandran favorável	94.82%
	Desvio do carbono beta	0.58%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.37%
	MolProbity Score	2.74
	Z-Score	-0.237
	QMEANscore6	0.738
		Chlamydomonas reinhardtii
	Clashscore	61.51
	Rotâmeros das cadeias laterais desfavoráveis	2.01
	Ramachandran desfavorável	0.18%
	Ramachandran favorável	94.67%
	Desvio do carbono beta	1.57%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.53%
	MolProbity Score	2.86
	Z-Score	-0.004
	QMEANscore6	0.763

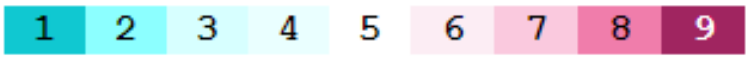
Modelo	Escores	
		Picea sitchensis
	Clashscore	59.41
	Rotâmeros das cadeias laterais desfavoráveis	3.58%
	Ramachandran desfavorável	0.36%
	Ramachandran favorável	94.96%
	Desvio do carbono beta	1.15%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.43%
	MolProbity Score	3.02
	Z-Score	-0.164
	QMEANscore6	0.746
		Selaginella moellendorffii
	Clashscore	67.12
	Rotâmeros das cadeias laterais desfavoráveis	1.93%
	Ramachandran desfavorável	0.36%
	Ramachandran favorável	95.85%
	Desvio do carbono beta	0.77%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.52%
	MolProbity Score	2.80
	Z-Score	-0.211
	QMEANscore6	0.741
		Solanum lycopersicum
	Clashscore	70.18
	Rotâmeros das cadeias laterais desfavoráveis	2.71%
	Ramachandran desfavorável	0.70%
	Ramachandran favorável	94.38%
	Desvio do carbono beta	1.33%
	Ligações com comprimentos ruins	0%
	Ligações com ângulos ruins	1.74%
	MolProbity Score	3.03
	Z-Score	-0.485
	QMEANscore6	0.717

4.7 Identificação de Regiões Funcionais Conservadas

Inicialmente, ao se realizar um alinhamento comparativo entre os modelos teóricos obtidos constatou-se a conservação na estrutura (Fig. 13 e 14) para ambas as enzimas, ocorrendo nas regiões do aminoácido 61 ao 239 e 440 a 610 em ICL e em quase toda sua totalidade nos modelos de MLS, aminoácido 81 ao 640.

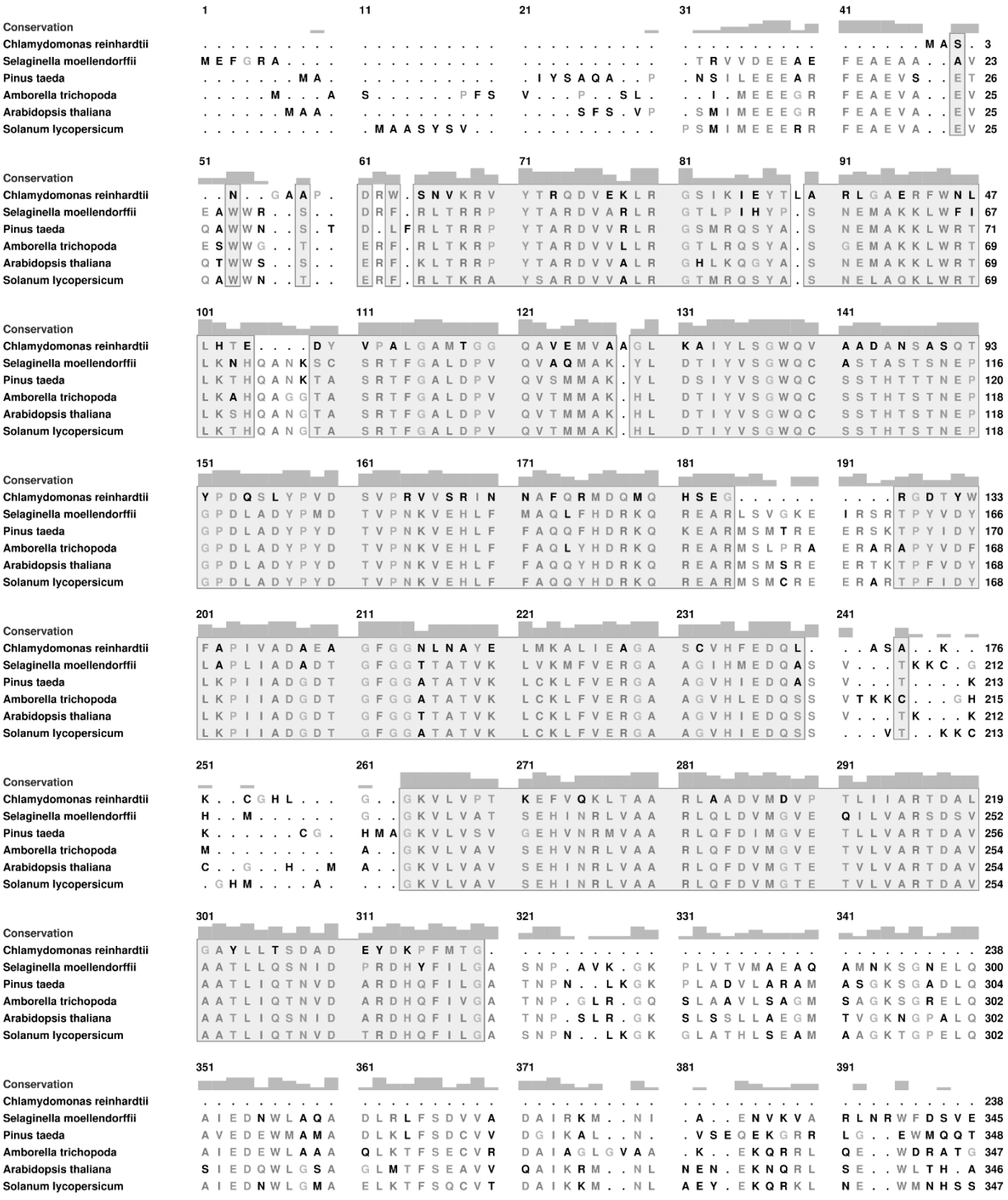
A relação evolutiva presente entre os aminoácidos foi analisada de acordo com o escore fornecido pelo ConSurf (Fig.15), no qual é levado em consideração o grau de conservação de acordo com sua importância e função para a estrutura. O programa estima o grau de conservação de cada posição pelo inverso da taxa evolutiva do site, ou seja, posições de rápida evolução se mostram variáveis em relação as de evolução lenta com maior conservação (Celniker *et al.*, 2013). Adicionalmente, tal escore leva em consideração parâmetros evolutivos de reconstrução filogenética, por métodos de distância, após realizar busca e obtenção de sequências homólogas na construção do múltiplo alinhamento utilizado como base pelo programa (Ashkenazy *et al.*, 2010)

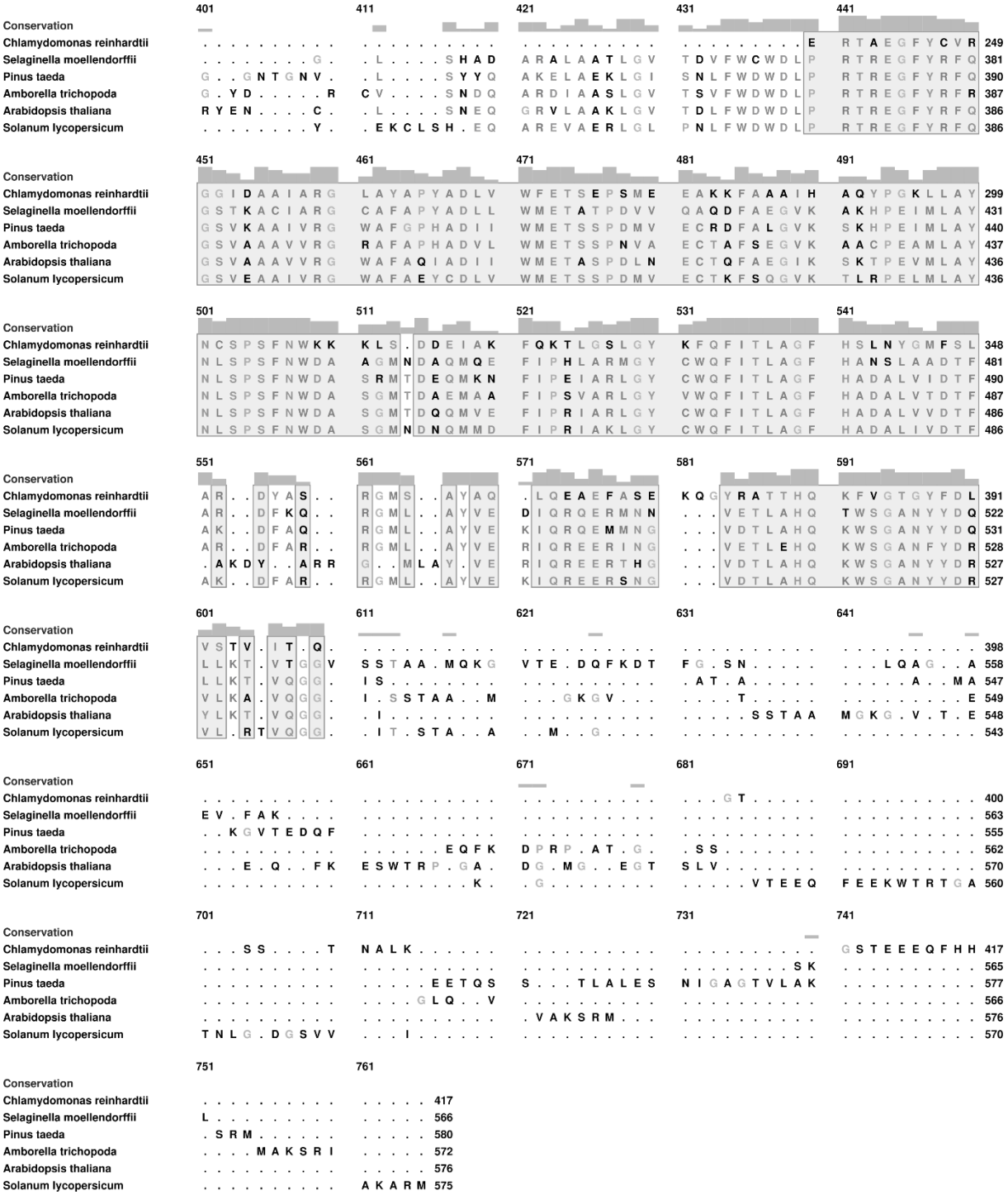
Figura 13: Representação do escore do programa ConSurf do grau de conservação entre os sítios na estrutura proteica.



De acordo a análise, os resíduos identificados como mais conservados entre os modelos descritos foram os sítios: Gly58, Ala65, Ser79, Gly80, Ala85, Thr93, Asp96, Ile112, Asp140, Glu142, Arg168, Glu169, Arg215, Asp217, Arg258 e Gly327, para Isocitrato liase (Fig.16 e 18). Enquanto que, para a Malato sintase, foram Arg89, Gly95, Asp116, Glu118, Asp119, Gly276, Asp279, Gly362, His370, Pro371, Leu373, Asn427, Ala455, Met451, Asp453, Ser460 e Thr531 (Fig.17 e 19). Nas figuras a seguir estão representados os resultados obtidos para o grau de conservação e importância dos resíduos de aminoácidos presentes em todos os modelos obtidos, bem como o padrão de conservação referente a posição dos resíduos em cada sequência, para ambas as enzimas ICL e MLS.

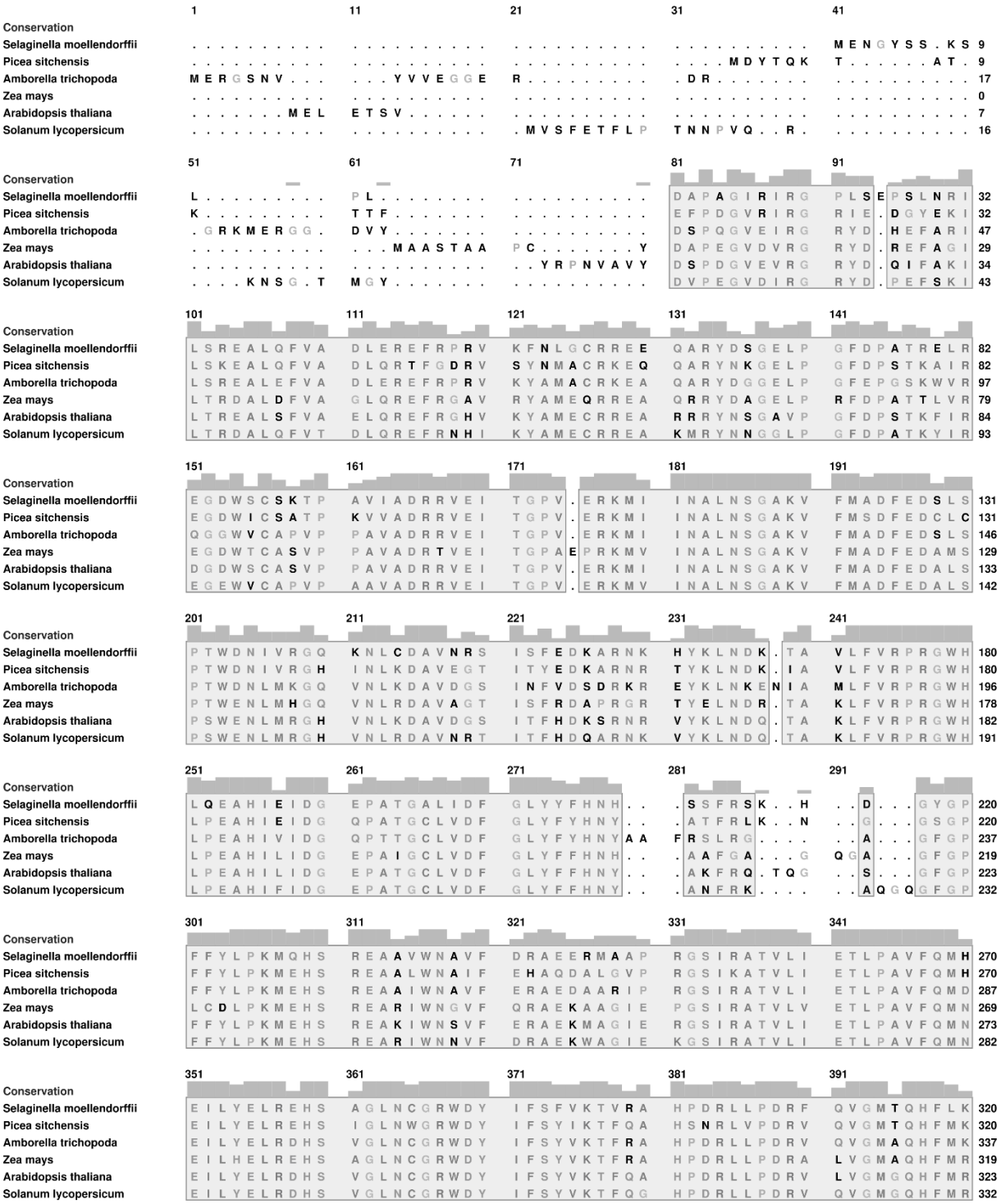
Figura 14: Alinhamento das estruturas tridimensionais das sequências dos modelos teóricos obtidos para *Isocitrato liase*.

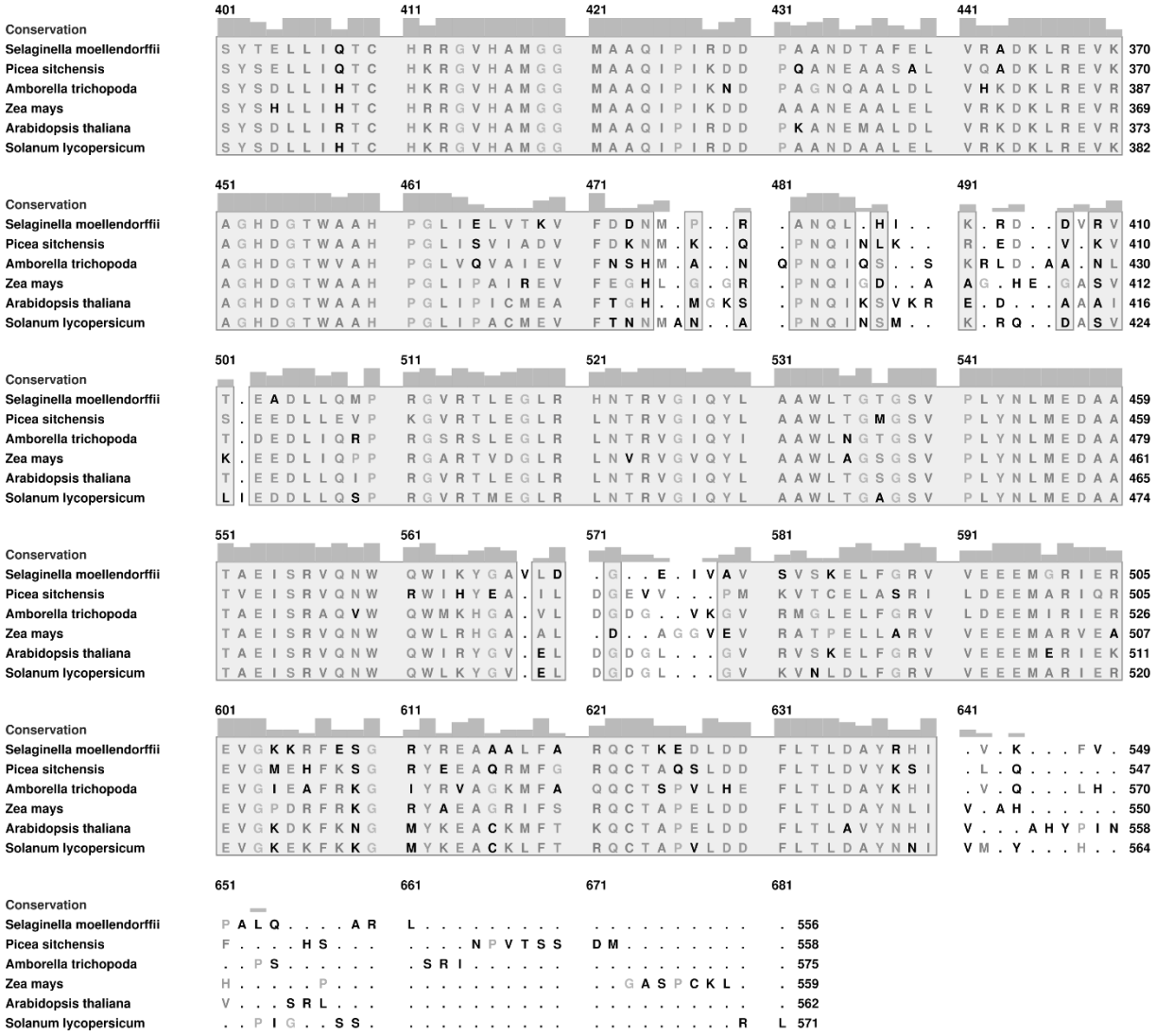




Legenda: Início do múltiplo alinhamento entre sequências de aminoácidos dos modelos obtidos na análise. Em destaque, regiões que se mostram conservadas na conformação estrutural presente entre os representantes dos diferentes *taxa*. Coluna: taxa de conservação dos sitio no alinhamento. Nº a direita: posição do resíduo em cada sequência.

Figura 15: Alinhamento das estruturas tridimensionais das sequências dos modelos teóricos obtidos para *Malato sintase.*





Legenda: Início do múltiplo alinhamento entre sequências de aminoácidos dos modelos obtidos na análise. Em destaque, regiões que se mostram conservadas na conformação estrutural presente entre os representantes dos diferentes *taxa*. Coluna: taxa de conservação dos sitio no alinhamento. Nº a direita: posição do resíduo em cada sequência.

Figura 16: Análise da conservação entre os resíduos de aminoácidos no modelo da *Isocitrato liase* proposto para *Chlamydomonas reinhardtii.*

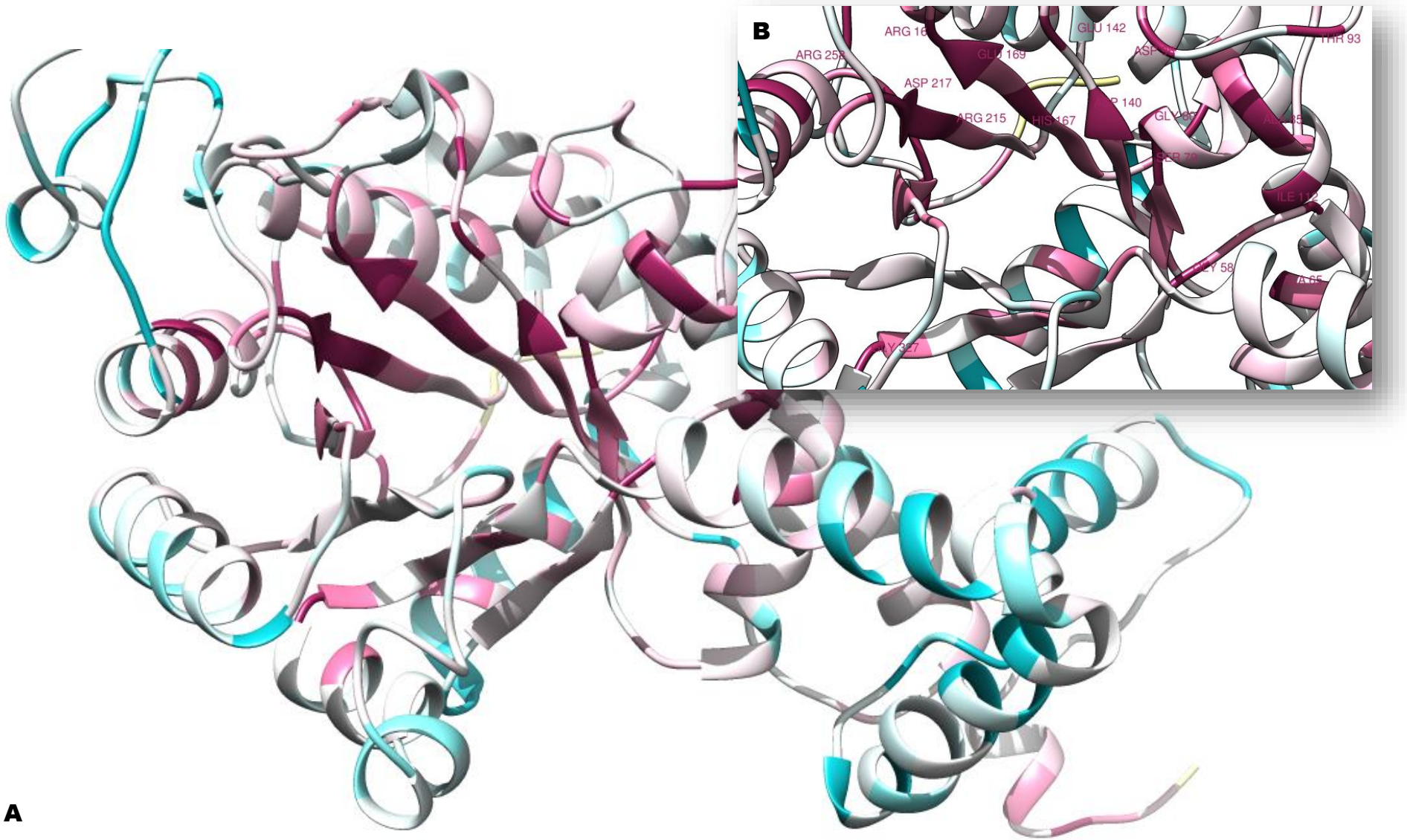
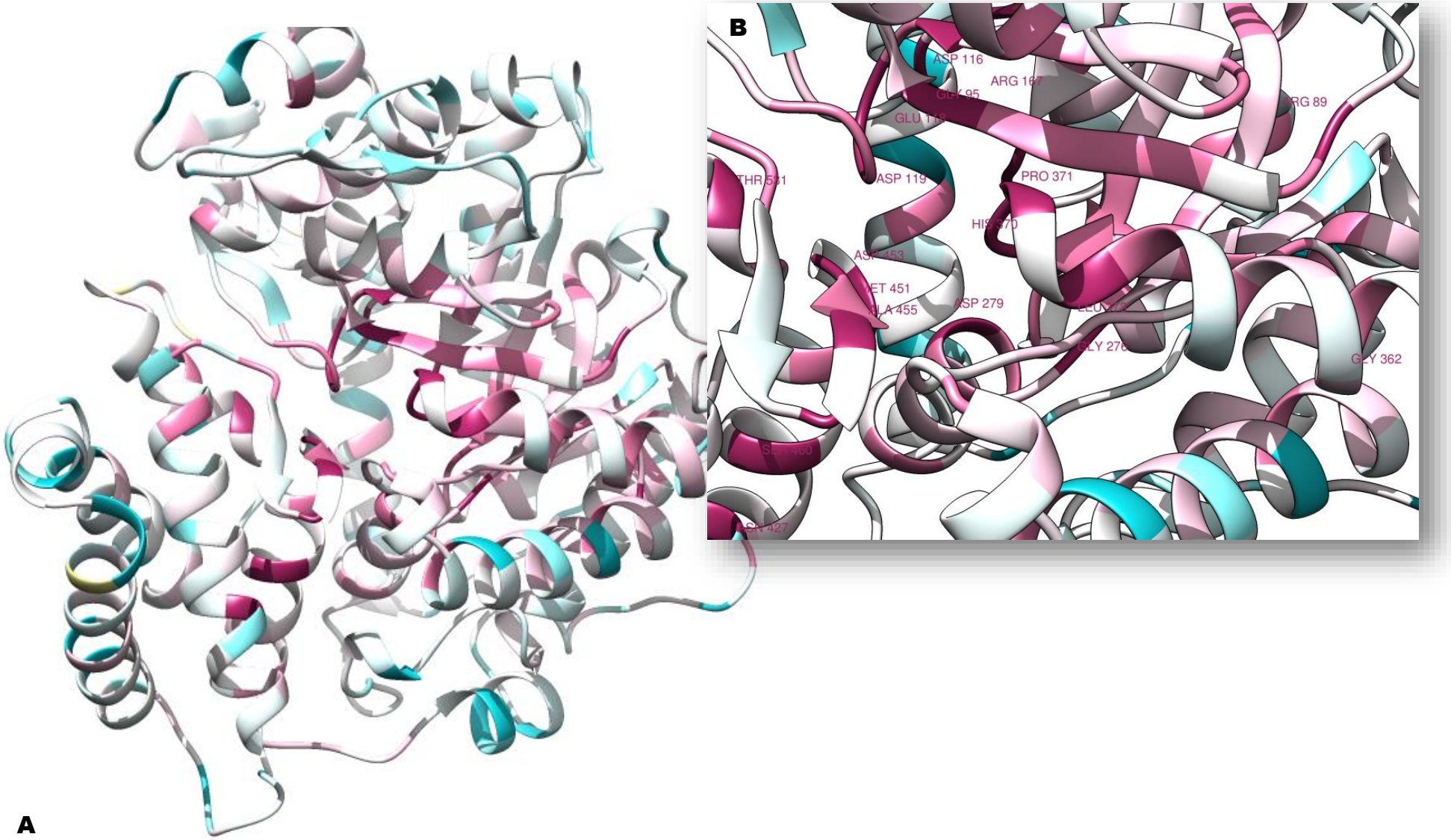
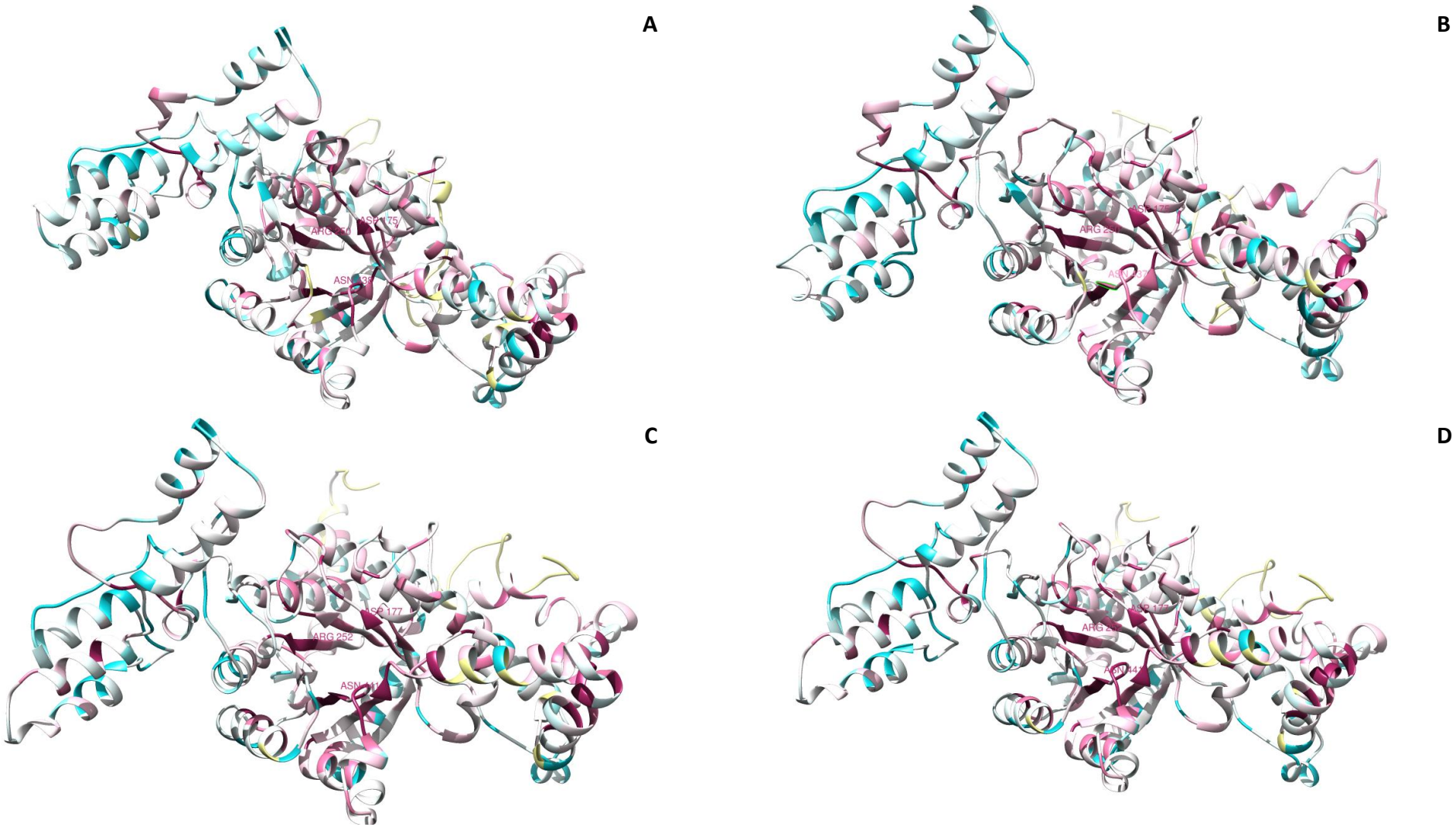

B
ARG 16
GLU 142
THR 93
ARG 25
GLU 169
ASP 217
ARG 215
HIS 167
A
Legenda: Modelo teórico da enzima representado em estilo cartoon, com suas estruturas secundárias (A); resíduos conservados nomeados e em destaque (B). A coloração na estrutura de acordo com o estabelecido pelo escore do ConSurf (Fig.13). Abrev. e nº referentes ao nome e posição do aminoácido, respectivamente.

Figura 17: Análise da conservação entre os resíduos de aminoácidos no modelo da *Malato sintase* proposto para *Chlamydomonas reinhardtii.*



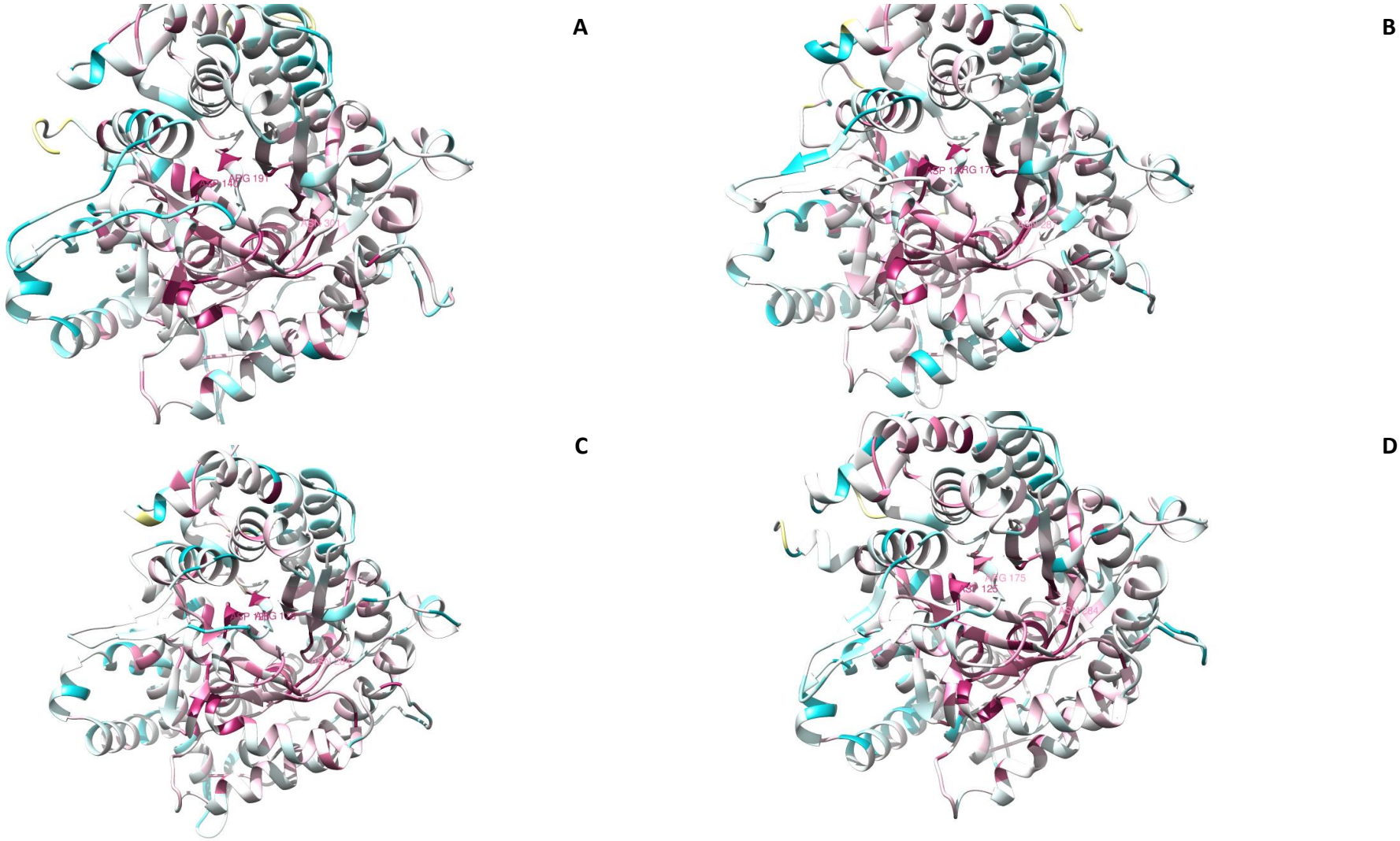
Legenda: Modelo teórico da enzima representado em estilo cartoon, com suas estruturas secundárias (A); resíduos conservados nomeados e em destaque (B). A coloração na estrutura de acordo com o estabelecido pelo escore do ConSurf (Fig.13). Abrev. e nº referentes ao nome e posição do aminoácido, respectivamente.

Figura 18: Análise do grau de conservação dos sítios da *Isocitrato liase* nos demais modelos obtidos para os representantes de *Viridiplantae*.



Legenda: Modelos teóricos de estrutura secundária de espécies dos táxons analisados. (A) *Amborella trichopoda*; (B) *Arabidopsis thaliana*; (C) *Pinus taeda*; e (D) *Selaginella moellendorffii*. A coloração na estrutura de acordo com o estabelecido pelo escore do ConSurf (Fig.13). Abrev. e nº referentes ao nome e posição do aminoácido, respectivamente.

Figura 19: Análise do grau de conservação dos sítios da *Malato Sintase* nos demais modelos obtidos para os representantes de *Viridiplantae.*



Legenda: Modelos teóricos de estrutura secundária de espécies dos táxons analisados. (A) *Amborella trichopoda*; (B) *Arabidopsis thaliana*; (C) *Picea sitchensis*; e (D) *Selaginella moellendorffii*. A coloração na estrutura de acordo com o estabelecido pelo escore do ConSurf (Fig.13. Abrev. e nº referentes ao nome e posição do aminoácido, respectivamente.

5. DISCUSSÃO

Durante a realização de estudos de filogenia molecular, para que as regiões homólogas possam ser inferidas e identificadas, o correto alinhamento múltiplo entre sequências de nucleotídeos ou aminoácidos de diferentes indivíduos ou espécies é uma etapa indispensável (Page; Holmes, 2001; Schneider, 2007). O alinhamento múltiplo entre as enzimas nas espécies estudadas mostrou variação entre suas sequências. Além das substituições nucleotídicas por mutações pontuais (transições e transversões), observou-se também um grande número de mudanças por remoção e inserção nessas sequências. Conforme o observado nas matrizes de substituição (Tabelas 4 e 5), os valores obtidos na taxa de transições/transversões (R) indicam uma menor quantidade de homoplasias entre as substituições, o que segundo Schneider (2007) numa análise filogenética para pares de espécies próximas está razão deve se manter elevada, já que o número de transversões é mais baixo que o de transições. Conforme as sequências do genoma vegetal passam por processos de divergência evolutiva, suas taxas de transversão tende a aumentar em relação às transições (Tabela 6; Fig. 7b).

Os alinhamentos obtidos entre as sequencias nucleotídicas, a partir do menor conjunto de dados de ICL e MLS, são influenciados pelo elevado número de mutações por inserção e/ou remoção, acarretando o aumento da necessidade de aberturas e extensão de *gaps* para preservar as homologias posicionais. Contudo, nesta mesma análise é notório que em MLS um menor número de *gaps* é necessário para corrigir a homologia posicional entre sítios.

A informação filogenética contida num conjunto de sequências está sujeita a processos evolutivos específicos de deriva genética e seleção natural que ocorreram ao longo das gerações (Magallón; Castillo, 2009; Finet *et al.*, 2010; Jiao *et al.*, 2011; Dickinson *et al.*, 2012; Bowman, 2013). A análise de saturação pela divergência *versus* transições e transversões é imprescindível para qualquer estudo com sequências de macromoléculas, pois permite uma maior confiabilidade da qualidade da informação filogenética nelas contida (Lemey *et al*., 2009). Os resultados mostraram que nas espécies analisadas, para ambos os conjuntos de sequências, está presente a saturação de pelo menos um tipo de substituição (Fig. 7a e b).

Ao se utilizar o modelo *General Time Reversible* (GTR) (Rodriguez *et al.,* 1990) para se representar a heterogeneidade do processo de substituição a que as enzimas estão submetidas, as condições de reversibilidade são corrigidas e incluem oito parâmetros independentes (sendo três referentes às frequências e cinco referentes às taxas de mudança nos nucleotídeos), pois no conjunto de dados nem todos os sítios evoluem numa mesma taxa (Tabelas 4 e 5). Ao lidar com a informação contida nas sequências proteicas, o uso dos modelos LG e JTT, para as respectivas enzimas Isocitrato liase e Malato sintase, possibilita que a reconstrução filogenética entre alinhamentos leve em consideração a probabilidade na substituição baseada na similaridade química estrutural e funcional dos resíduos de aminoácidos presente numa matriz. Para ambos os modelos, matrizes de substituição são construídas incorporando variabilidade nas taxas evolutivas entre sítios por valores de verossimilhança, porém com suporte em base de dados diferentes entre si (Jones *et al.*, 1992; Le; Gascue, 2008; Lemey *et al*., 2009).

A filogenia das duas enzimas resultou numa clara separação de grupos entre os ramos basais das *Viridiplantae* (Fig. 8), apoiadas pelos valores de suporte estatístico obtidos pelos testes referentes a cada método. Ao verificar a divergência dos ramos internos das Eudicotiledôneas podem ser observadas trocas de espécies entre Ordens (Fig. 8 - 12), porém sem deixarem de manter seus agrupamentos em Subclasses de acordo com o sistema de classificação vigente de APG III (Soltis; Soltis, 2004; Lewis; McCourt, 2004; Endress; Doyle, 2009). O motivo dessas trocas pode ser o reflexo da saturação no sinal filogenético de seus genes nas proteínas dos clados mais derivados das angiospermas, para ambas as enzimas (Tabela 6; Fig. 7). Além disso, embora tenham sido utilizados modelos evolutivos adequados a cada conjunto de sequência a inferência de suas topologias está atrelada a escolha do *Outgroup,* espécie mais distante filogeneticamente relacionada aos ramos mais derivados de uma árvore filogenética (Roquist *et al.*, 2009). Logo, caso o conjunto de dados utilizados não contenham uma amostragem com sítios homólogos suficientes (Lemey *et al*., 2009; Futuyma, 2009), que representem todos os estados de caráter presentes na evolução do genoma, sua baixa resolução entre táxons mais derivados é esperada.

De acordo com os resultados de reconstrução filogenética obtidos pela máxima verossimilhança (Fig. 8a) incoerências de posicionamento no tempo evolutivo entre grupos irmãos estão presentes entre membros ao nível de ordem, nos *taxa* pertencentes a *Asterids* e *Rosids,* dentro do núcleo das Eudicotiledôneas.

A Inferência Bayesiana para os dois conjuntos de sequências de tamanhos diferentes se mostrou capaz de apresentar melhor resolução dos agrupamentos na reconstrução de suas topologias (Fig. 9, 10, 11 e 12). Contudo, ainda assim foram encontradas diferenças entre as monofilias estabelecidas de sequencias proteicas nos dendogramas obtidos e atual sistema de classificação de espécies de APG III. Sendo assim, é possível que tais agrupamentos de enzimas estejam relacionados a evolução destas não apenas conforme o táxon a que pertence, mas que também esteja associado a pressões evolutivas que agiram sobre as sequências na adaptação do funcionamento de suas vias metabólica as condições ambientais expostas (Futuyma, 2009). Entretanto, apenas por meio de análise filogenética desse conjunto de dados nenhum padrão metabólico associado aos agrupamentos foi diagnosticado.

Conforme ilustrado pelas médias gerais obtidas com teste de seleção (Tabela 8), o conjunto de sequências analisadas passou por processos de seleção distintos que acarretaram em diferenciação na funcionalização da enzima Isocitrato liase, por seleção positiva, e manutenção da função na Malato sintase, por seleção purificadora, anterior ao surgimento das plantas verdes. O que pode ser explicado pela necessidade do ciclo do glioxilato ou destas enzimas no metabolismo estarem presentes no genoma de organismo mais primitivos (Roucourt *et al*., 2009; Kondrashov *et al*., 2006; Nakazawa *et al*., 2011; Hügler; Sievert, 2011). Tais informações também podem ser inferidas por meio da conservação relativa entre as estruturas preditas para cada enzima (Figs. 14 e 15). A tendência nas sequências da Isocitrato liase passarem por processo de seleção positiva pode estar relacionada ao grau de similaridade do seu sítio catalítico com outros membros de sua superfamília (ICL / PEP mutase) a caracterizam como produtos da evolução dirigidas pela especiação, uma vez que a sua homóloga a 2-Metilisocitratoliase (MICL), responsável pela quebra do 2-metil-isocitrato, irá atuar em organismos que evolutivamente apresentam vias metabólicas alternativas derivadas do ciclo do ácido cítrico, porém não utilizam o ciclo do glioxilato, como as plantas (Liu *et al*., 2004; 2005).

A qualidade dos modelos obtidos com a predição por homologia de proteínas de representantes dos diferentes táxons de *Viridiplantae* (Tabelas 9 e 10) são reflexos indiretos da diferença nas taxas de conservação dos genes de cada enzima ao longo da evolução dos genomas em plantas. Sendo os melhores modelos, para ambas, aqueles referentes ao grupo das algas verdes (*Chlamydomonas reinhardtii*).

Os modelos gerados para cada uma das proteínas mostraram a presença de estruturas e resíduos de sítios ativos conservados. Em ambas, a ocorrência do barril TIM como regiões de domínio críticos para atividade catalítica enzimática é relatada pela literatura (Howard *et al*., 2000; Liu *et al*., 2005) o que pode ser visto em todos modelos resultantes da análise. Nos modelos de *C. reinhardtii* da Isocitrato liase os sítios de interação à succinato (His167, Asn306, Ser312, Ser325 e Thr378) estão presentes (Fig. 16), bem como em Malato sintase os sítios de ligação a Mg+ e interação à Glioxilato (Glu251, Arg323, Glu421, Leu488 e Phe505) (Fig. 16), o que demonstra além da conservação ao longo da evolução e presença de atividade funcional em todos os modelos teóricos propostos. Porém, além desses resíduos já identificados na literatura, os resultados obtidos com a ferramenta ConSurf exibiram outros resíduos conservados presentes em todos os modelos teóricos propostos para espécies estudadas (Figs. 16b, 17b, 18 e 19). Entre eles destaque para os resíduos associados a estrutura do Barril TIM, presente em ambas as enzimas como principal domínio de interação ao substrato.

Segundo Torres (2012), ainda não há teorias comprovadas na literatura científica sobre sua origem, nem se ela é de origem única na evolução, contudo novos modelos com o objetivo de inferir sua filogenia não pela conservação de sequência, mas pela conservação nas interações presentes entre os aminoácidos. (Huang *et al.,* 2012; Vijayabaskar; Vishveshwara, 2012). A evolução molecular pode ser movida pela capacidade das biomoléculas em adotarem conformações múltiplas, pois enorme quantidade de informação presente nos genomas de organismos mais complexos gera um número limitado de dobras de proteínas, onde uma estrutura de domínio específico, pode assumir inúmeras funções (Meier *et al.,* 2007).

6. CONCLUSÃO

Com base nas análises comparativas entre modelos *in silico* das enzimas e partir dos resultados de inferência filogenética, por máxima Verossimilhança e inferência Bayesiana, ambas as enzimas apresentam um padrão de conservação relativamente elevado entre alguns sítios de suas estruturas gerando topologias condizentes com os processo de seleção positiva para as sequências de Isocitrato liase com neofuncionalização*;* e seleção purificadora para Malato sintase, com a manutenção de sua conformação estrutural ao longo da evolução do gene, ambas em gerações anteriores as *Viridiplantae*. Adicionalmente, as análises das relações evolutivas presentes entre os aminoácidos de acordo com a importância de sua conservação para a estrutura, feitas pelo programa ConSurf, geraram novos questionamentos referentes a possível associação de seus resíduos de aminoácidos com a identificação de novos domínios regulatórios dentre da estrutura proteica dessas enzimas na via metabólicas associadas as adaptações de um determinado grupo vegetal ao ambiente que vive. Corroborando com a importância em se realizar novos estudos para se elucidar o metabolismo vegetal sob uma perspectiva evolutiva das relações entre os genes e a expressão de suas enzimas.

REFERÊNCIAS

ABASCAL, F.; ZARDOYA, R.; POSADA, D. ProtTest: selection of best-fit models of protein evolution. **Bioinformatics** (Oxford, England), v. 21, n. 9, p. 2104–5, 1 maio 2005.

ADAMS, K. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. **Molecular Phylogenetics and Evolution**, v. 29, n. 3, p. 380–395, dez. 2003.

ALTSCHUL, S., MADDEN, T., SCHÄFFER, A., ZHANG, J., ZHANG, Z., MILLER, W., LIPMAN, D.Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v. 25, p. 3389–3402, 1997.

ANDERSON, J. T.; WILLIS, J. H.; MITCHELL-OLDS, T. Evolutionary genetics of plant adaptation. **Trends in genetics**: TIG, v. 27, n. 7, p. 258–66, jul. 2011.

ANDREANI, J.; GUEROIS, R. Evolution of protein interactions: From interactomes to interfaces. **Archives of biochemistry and biophysics**, v. 554C, p. 65–75, 20 maio 2014.

ASHKENAZY, H. et al. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. **Nucleic acids research**, v. 38, n. Web Server issue, p. W529–33, jul. 2010.

BOWMAN, J. L. Walkabout on the long branches of plant evolution. **Current opinion in plant biology**, v. 16, n. 1, p. 70–7, fev. 2013.

CASSIMJEE, K. E. et al. *Chromobacterium violaceum* ω-transaminase variant Trp60Cys shows increased specificity for (S)-1-phenylethylamine and 4'-substituted acetophenones, and follows Swain-Lupton parameterisation. **Organic & biomolecular chemistry**, v. 10, n. 28, p. 5466–70, 28 jul. 2012.

CELNIKER, G. et al. ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. **Israel Journal of Chemistry**, v. 53, n. 3-4, p. 199–206, 15 abr. 2013.

COLLINS, D. W.; JUKES, T. H. Rates of transition and transversion in coding sequences since the human-rodent divergence. **Genomics**, v. 20, n. 3, p. 386–96, abr. 1994.

COOPER, T. G.; BEEVERS, H. Beta oxidation in glyoxysomes from castor bean endosperm. **The Journal of biological chemistry**, v. 244, n. 13, p. 3514–20, 10 jul. 1969.

CORNAH, J. E., GERMAIN, V., WARD, J. L., BEALE, M. H., & SMITH, S. M. Lipid utilization, gluconeogenesis, and seedling growth in Arabidopsis mutants lacking the glyoxylate cycle enzyme malate synthase. **Journal of Biological Chemistry**, v. 279, n. 41, p. 42916–42923, 2004.

CREPET, W. L.; NIKLAS, K. J. Darwin's second'abominable mystery': Why are there so many angiosperm species? **American Journal of Botany**, v. 96, n. 1, p. 366, jan. 2009.

DARRIBA D, TABOADA GL, DOALLO R, POSADA D. jModelTest 2: more models, new heuristics and parallel computing. **Nature methods**, v. 9, n. 8, p. 772, ago. 2012.

_____. ProtTest 3: fast selection of best-fit models of protein evolution. **Bioinformatics** (Oxford, England), v. 27, n. 8, p. 1164–5, 15 abr. 2011.

DARWIN, Charles - **The origin of species by means of natural selection**. London : J. M. Dent, [s.d.]. XXIV, 488 p

DICKINSON, H.; COSTA, L.; GUTIERREZ-MARCOS, J. Epigenetic neofunctionalisation and regulatory gene evolution in grasses. **Trends in plant science**, v. 17, n. 7, p. 389–94, jul. 2012.

DRUMMOND, A. J.; RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC evolutionary biology**, v. 7, p. 214, jan. 2007.

EASTMOND, P. J.; GRAHAM, I. A. Re-examining the role of the glyoxylate cycle in oilseeds. **Trends in Plant Science**, v. 6, n. 2, p. 72–77, 2001.

ENDRESS, P. K., DOYLE, J. A., JAN. Reconstructing the ancestral angiosperm flower and its initial specializations. **Am. J. Bot.** 96 (1), 2009. pp. 22–66.

EWING, B., HILLER, L., WENDL, M., GREEN, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Research**, v. 8, p. 175–185, 1998.

FINET, C. et al. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. **Current biology** : CB, v. 20, n. 24, p. 2217–22, 21 dez. 2010.

FRIEDMAN, W.E. The meaning of Darwin's "abominable mystery". **Am. J. Bot.**, 96 (1), 2009. pp. 5-21.

FURNESS C. A.; RUDALL P. J. Pollen aperture evolution – a crucial factor for eudicot success? **Trends in Plant Science** 9, 2004. pp. 1360-1385

FUTUYMA, D*.* **Evolution,** 2nd ed. Sinauer Associates Inc., 2009.

GRAHAM, I. A. Seed storage oil mobilization. **Annual review of plant biology**, v. 59, p. 115–42, jan. 2008.

GRAUR, D. & LI, W. H. **Fundamentals of Molecular Evolution**. 2a edição. Sinauer Press, Sunderland, Massachusetts. 2000. 481p.

GROSDIDIER, A.; ZOETE, V.; MICHIELIN, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. **Nucleic acids research**, v. 39, n. Web Server issue, p. W270–7, jul. 2011.

HELED, J.; DRUMMOND, A. J. Bayesian inference of species trees from multilocus data. **Molecular biology and evolution**, v. 27, n. 3, p. 570–80, mar. 2010.

HENIKOFF, S.; HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 22, p. 10915–9, 15 nov. 1992.

HOFFMAN, G. E.; PUERTA, M. V. S.; DELWICHE, C. F. Evolution of light-harvesting complex proteins from Chl c-containing algae. **Bmc Evolutionary Biology**, v. 11, p. 101, 2011.

HOWARD, B. R.; ENDRIZZI, J. A; REMINGTON, S. J. Crystal structure of Escherichia coli malate synthase G complexed with magnesium and glyoxylate at 2.0 A resolution: mechanistic implications. **Biochemistry**, v. 39, n. 11, p. 3156–68, 21 mar. 2000.

HÜGLER, M.; SIEVERT, S. M. Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. **Annual review of marine science**, v. 3, p. 261–89, jan. 2011.

JIAO, Y., WICKETT, N.J., AYYAMPALAYAM, S., CHANDERBALI, A.S., LANDHERR, L., RALPH, P.E., TOMSHO, L.P., HU, Y., LIANG, H., SOLTIS, P.S., SOLTIS, D.E., CLIFTON, S.W., SCHLARBAUM, S.E., SCHUSTER, S.C., MA, H., LEEBENS-MACK, J., DE PAMPHILIS, C.W. Ancestral polyploidy in seed plants and Angiosperms. **Nature,** 473, 2011. pp. 97–100.

JONES DT, TAYLOR WR, THORNTON JM (1992). The rapid generation of mutation data matrices from protein sequences. **Comput Applic Biosci** 8: 275–282.

JUDD, W.S.; CAMPBELL, C.S.; KELLOG, E.A.; STEVENS, P.F. **Plant Systematics: A phylogenetic approach**. Sinauer Associates, Sunderland, 2009.

KATHRIARACHCHI, H., HOFFMANN, P., SAMUEL, R., WURDACK, K.J., CHASE M.W. Molecular phylogenetics of Phyllanthaceae inferred from five genes (plastid atpB, matK, 3′ndhF, rbcL, and nuclear PHYC). **Molecular Phylogenetics and Evolution**, v. 36, n. 1, p. 112–134, jul. 2005.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular biology and evolution**, 16 jan. 2013.

KATOH, K.; TOH, H. Parallelization of the MAFFT multiple sequence alignment program. **Bioinformatics** (Oxford, England), v. 26, n. 15, p. 1899–900, 1 ago. 2010.

KIMURA, M. The neutral theory of molecular evolution and the world view of the neutralists. **Genome,** v. 31, n. 1, p. 24–31, 1989.

KLIEBENSTEIN, D. J. Making new molecules--evolution of structures for novel metabolites in plants. **Current opinion in plant biology**, v. 16, n. 1, p. 112–7, mar. 2013.

KONDRASHOV, F. A, KOONIN, E. V, MORGUNOV, I. G., FINOGENOVA, T. V, & KONDRASHOVA, M. N. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. **Biology direct**, v. 1, p. 31, jan. 2006.

KORNBERG, H.; BEEVERS, H. The glyoxylate cycle as a stage in the conversion of fat to carbohydrate in castor beans. **Biochimica et biophysica acta**, v. 26, n. 3, p. 531–537, 1957.

KUNZE, M., PRACHAROENWATTANA, I., SMITH, S. A central role for the peroxisomal membrane in glyoxylate cycle function. **Biochimica et biophysica acta**, v. 1763, n. 12, p. 1441–52, dez. 2006.

LANG, B. F.; GRAY, M. W.; BURGER, G. Mitochondrial genome evolution and the origin of eukaryotes. **Annual Review of Genetics**, v. 33, n. 1, p. 351–397, 1999.

LARKIN, M., BLACKSHIELDS, G., BROWN, N., CHENNA, R., MCGETTIGAN, P., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J., HIGGINS, D. G. Clustal W and Clustal X version 2.0. **Bioinformatics** (Oxford, England), v. 23, n. 21, p. 2947–8, 1 nov. 2007.

LAWTON-RAUH, A. Evolutionary dynamics of duplicated genes in plants. **Molecular Phylogenetics and Evolution**, v. 29, n. 3, p. 396–409, dez. 2003.

LE, S. Q.; GASCUEL, O. An improved general amino acid replacement matrix. **Molecular biology and evolution**, v. 25, n. 7, p. 1307–20, jul. 2008.

LEMEY, P. et al. **The phylogenetic handbook:** a practical approach to phylogenetic analysis and hypothesis testing. 2. ed. Cambridge, UK: Cambridge University Press, 2009. p. 750

LEWIS, L. A; MCCOURT, R. M. Green algae and the origin of land plants. **American journal of botany**, v. 91, n. 10, p. 1535–56, out. 2004.

LIBERLES, D. A et al. The interface of protein structure, protein biophysics, and molecular evolution. **Protein science**: a publication of the Protein Society, v. 21, n. 6, p. 769–85, jun. 2012.

LIU, S. et al. Conformational flexibility of PEP mutase. **Biochemistry**, v. 43, n. 15, p. 4447–53, 20 abr. 2004.

_____. Crystal structures of 2-methylisocitrate lyase in complex with product and with isocitrate inhibitor provide insight into lyase substrate specificity, catalysis and evolution. **Biochemistry**, v. 44, n. 8, p. 2949–62, 1 mar. 2005.

LYNCH, M. The Evolutionary Fate and Consequences of Duplicate Genes. **Science**, v. 290, n. 5494, p. 1151–1155, 10 nov. 2000.

MAGALLON, S.; CASTILLO, A. Angiosperm diversification through time. **Am. J. Bot**., v. 96, n. 1, p. 349–365, jan. 2009.

MAIRA, N., TORRES, T. M., DE OLIVEIRA, A L., DE MEDEIROS, S. R. B., AGNEZ-LIMA, L. F., LIMA, J. P. M. S., & SCORTECCI, K. C. Identification, characterisation and molecular modelling of two AP endonucleases from base excision repair pathway in sugarcane provide insights on the early evolution of green plants. **Plant biology** (Stuttgart, Germany), p. 1–10, 19 ago. 2013.

MEIER, S. et al. Continuous molecular evolution of protein-domain structures by single amino acid changes. **Current biology**: CB, v. 17, n. 2, p. 173–8, 23 jan. 2007.

MORJAN, C. L., AND RIESEBERG, L. H. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. **Molecular Ecology** 13, 2004. pp. 1341-1356.

NAKAZAWA, M., MINAMI, T., TERAMURA, K., KUMAMOTO, S., HANATO, S., TAKENAKA, S., MIYATAKE, K Molecular characterization of a bifunctional glyoxylate cycle enzyme, malate synthase/isocitrate lyase, in Euglena gracilis.Comparative biochemistry and physiology. Part B, **Biochemistry & molecular biology**, ago. 2005.

NAKAZAWA, M., NISHIMURA, M., INOUE, K., UEDA, M., INUI, H., NAKANO, Y., & MIYATAKE, K. Characterization of a bifunctional glyoxylate cycle enzyme, malate synthase/isocitrate lyase, of Euglena gracilis. **The Journal of Eukaryotic Microbiology**, *58*(2), 2011. pp. 128–133.

NEI, M.; KUMAR, S**. Molecular evolution and phylogenetics**. [S.l.]: Oxford University Press, 2000. p. 352

NEI, M.; SUZUKI, Y.; NOZAWA, M. The neutral theory of molecular evolution in the genomic era. **Annual Review of Genomics and Human Genetics**, v. 11, n. June, p. 265–289, 2010.

NELSON, DAVID L.; COX, MICHAL M. **Princípios de bioquímica de Lehninger**. 5. ed. Porto Alegre : Artmed, 2011.

NIKLAS, K.; KUTSCHERA, U. The evolution of the land plant life cycle. **New Phytologist**, n. 2009, p. 27–41, 2010.

NIKLAS, K. **The evolutionary biology of plants**. The University of Chicago Press, Chicago, 1997. 449 pp.

NOTREDAME, C; HIGGINS, D. G.; HERINGA, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. **Journal of molecular biology**, v. 302, n. 1, p. 205-17, 8 set 2000.

QIU Y., LEE J., BERNASCONI-QUADRONI F., SOLTIS D. E., SOLTIS P. S., ZANIS M., ZIMMER E. A., CHEN Z., SAVOLAINEN V., CHASE M. W. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. **Nature**, v. 402, n. 6760, p. 404–407, 25 nov. 1999.

QUETTIER, A.-L.; EASTMOND, P. J. Storage oil hydrolysis during early seedling growth. **Plant physiology and biochemistry**: PPB / Société française de physiologie végétale, v. 47, n. 6, p. 485–90, jun. 2009.

RIDLEY, M. (2004). **Evolution** (3rd ed., pp. 1–786). Malden, USA: Blackwell Publishing.

RIESEBERG, L. H.; WILLIS, J. H. Plant speciation. **Science** (New York, N.Y.), v. 317, n. 5840, p. 910–4, 17 ago. 2007.

RIESEBERG, L.H., WOOD, T.E. & BAACK, E. The nature of plant species. **Nature,** 440, 2006. pp. 524 –527.

ROMANO, E.; BRASILEIRO, A.; CARNEIRO, V. **Extração de DNA de tecidos Vegetais**. [s.l: s.n.]. p. 163–177

RONQUIST, F.; HUELSENBECK, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics**, v. 19, n. 12, p. 1572–1574, 11 ago. 2003.

RONQUIST, F.; VAN DER MARK, P.; HUELSENBECK, J. P. 7. Bayesian phylogenetic analysis using MRBAYES *In:* LEMEY, P.; SALEMI, M.; VANDAMME, A.-M.; (EDS). **The phylogenetic handbook:** a practical approach to phylogenetic analysis and hypothesis testing. 2. ed. Cambridge, UK: Cambridge University Press, 2009. Pp. 210-237

ROUCOURT, B. et al. Biochemical characterization of malate synthase G of P. aeruginosa. **BMC biochemistry**, v. 10, p. 20, jan. 2009.

SAMBROOK, J.; FRITSCH, E.; MANIATIS, T. **Molecular Cloning. A laboratory manual.** [s.l: s.n.]. v. 3

SCHLÜTER, A., FOURCADE, S., RIPP, R., MANDEL, J. L., POCH, O., & PUJOL, A. The evolutionary origin of peroxisomes: an ER-peroxisome connection. **Molecular biology and evolution**, v. 23, n. 4, p. 838–45, abr. 2006.

SCHNARRENBERGER, C.; MARTIN, W. Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants. A case study of endosymbiotic gene transfer. **European journal of biochemistry** / FEBS, v. 269, n. 3, p. 868–83, fev. 2002.

SCHNEIDER, H. **Métodos de Análise Filogenética** - Um guia prático. 3. ed. Ribeirão Preto: SBG & Hollos, 2007. v. 1. 200 p.

SI QUANG LE AND; OLIVIER GASCUEL. An Improved General Amino Acid Replacement Matrix **Mol Biol Evol** (2008) 25 (7): 1307-1320.

SIMPSON, M.G. **Plant systematics**. Elsevier Academic Press, Amsterdam, 2006.

SOLTIS, D. E. et al. ANGIOSPERM PHYLOGENY: 17 GENES, 640 TAXA. **American Journal of Botany**, v. 98, n. 4, p. 704–730, 2011.

SOLTIS, P. S.; SOLTIS, D. E. The origin and diversification of angiosperms. **American Journal of Botany**, v. 91, n. 10, p. 1614–1626, 2004.

SOUZA, V. C., LORENZI, H. **Botânica Sistemática: Guia ilustrado para identificação das famílias de Angiospermas da flora brasileira, baseado em APG II**, Nova Odessa, Instituto Plantarum. (1), 2004. pp. 1614–1626.

TAMURA, K., PETERSON, D., PETERSON, N., STECHER, G., NEI, M., KUMAR, S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. **Molecular biology and evolution**, v. 28, n. 10, p. 2731-2739, 4 maio 2011.

TAVARÉ S. (1986) Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. **Lectures on Mathematics in the Life Sciences** (American Mathematical Society) 17: 57–86.

TORRES, D. C., LIMA, J. P. M. S., FERNANDES, A. G., NUNES, E. P., GRANGEIRO, T. B. Phylogenetic relationships within Chamaecrista sect. Xerocalyx (Leguminosae, Caesalpinioideae) inferred from the cpDNA trnE-trnT intergenic spacer and nrDNA ITS sequences. **Genetics and Molecular Biology**, v. 34, n. 2, p. 244–251, 2011.

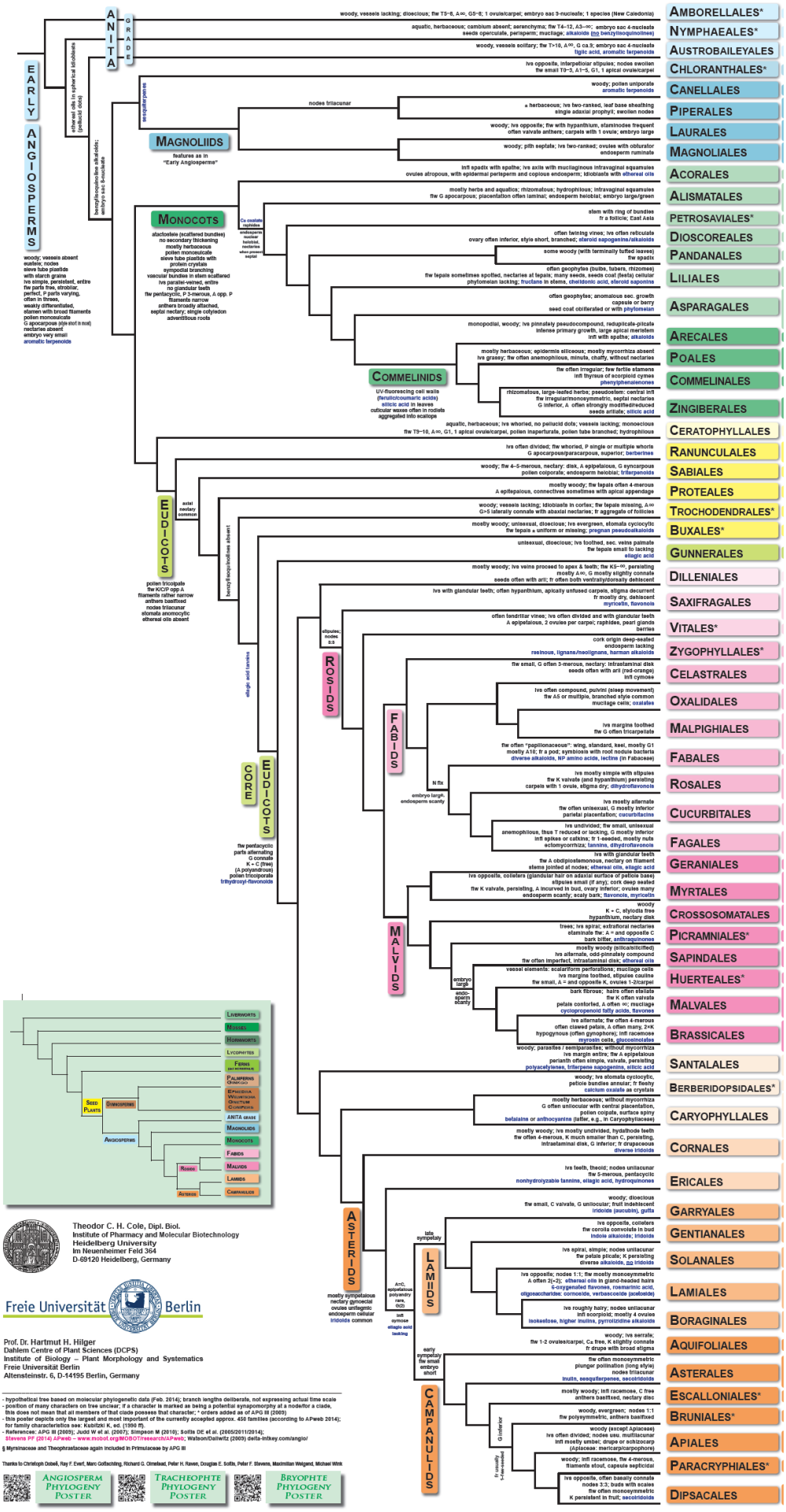
WALLI, J.; BROWS, J. Lipid biochemists salute the genome. **Plant Journal**, v. 61, p. 1092–1106, 2010.

WODNIOK, S. et al. Origin of land plants: do conjugating green algae hold the key? **BMC evolutionary biology**, v. 11, n. 1, p. 104, jan. 2011.

XIA, X., XIE. Z. (2001). DAMBE: Data analysis in molecular biology and evolution. **Journal of Heredity**, 92, 371-373.

ANEXOS

ANEXO 01 – Representação esquemática do Sistema de classificação APG III.



Theodor C. H. Cole, Dipl. Biol.
Institute of Pharmacy and Molecular Biotechnology
Heidelberg University
Im Neuenheimer Feld 364
D-69120 Heidelberg, Germany

Freie Universität Berlin

Prof. Dr. Hartmut H. Hilger
Dahlem Centre of Plant Sciences (DCPS)
Institute of Biology – Plant Morphology and Systematics
Freie Universität Berlin
Altensteinstr. 6, D-14195 Berlin, Germany

- hypothetical tree based on molecular phylogenetic data (Feb. 2014); branch lengths deliberate, not expressing actual time scale
- position of many characters on tree unclear; if a character is marked as being a potential synapomorphy at a node/for a clade, this does not mean that all members of that clade possess that character; * orders added as of APG III (2009)
- this poster depicts only the largest and most important of the currently accepted approx. 450 families (according to APweb 2014); for family characteristics see: Kubitzki K, ed. (1990 ff).
- References: APG III (2009); Judd W et al. (2007), Simpson M (2010), Soltis DE et al. (2005/2011/2014); Stevens PF (2014) APweb – www.mobot.org/MOBOT/research/APweb; Watson/Dallwitz (2003) delta-intkey.com/angio/

§ Myrsinaceae and Theophrastaceae again included in Primulaceae by APG III

Thanks to Christoph Dobeš, Ray F. Evert, Marc Gottschling, Richard G. Olmstead, Peter H. Raven, Douglas E. Soltis, Peter F. Stevens, Maximilian Weigend, Michael Wink

Angiosperm Phylogeny Poster
Tracheophyte Phylogeny Poster
Bryophyte Phylogeny Poster