Universidade Federal do Rio Grande do Sul
Instituto de Biociências
Programa de Pós-Graduação em Genética e Biologia Molecular

EVOLUÇÃO MOLECULAR, DIVERGÊNCIA FUNCIONAL E ASPECTOS ESTRUTURAIS DA FAMÍLIA GÊNICA DA ÁLCOOL DESIDROGENASE

Claudia Elizabeth Thompson

Tese submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para a obtenção do grau de Doutor em Ciências.

Orientadora: Francisco Mauro Salzano
Co-orientadora: Loreta Brandão de Freitas

Porto Alegre.

Maio, 2009.

INSTITUIÇÕES E FONTES FINANCIADORAS

Essa Tese foi desenvolvida no Laboratório de Evolução Molecular (LEM) do Departamento de Genética da Universidade Federal do Rio Grande do Sul (UFRGS) e no Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas (LABIO) da Faculdade de Informática da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brasil.

As fontes financiadoras foram PRONEX (Programa de Apoio a Núcleos de Excelência), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul), PROPESQ-UFRGS (Pró-Reitoria de Pesquisa da Universidade Federal do Rio Grande do Sul) e PRPPG-PUCRS (Pró-Reitoria de pesquisa e Pós-Graduação da Pontifícia Universidade Católica do Rio Grande do Sul).

Dedicado a meus pais e irmãos,

a Ricardo e a meus orientadores

Salzano, Loreta e Osmar.

Sumário

RESUMO

A álcool desidrogenase é uma família gênica classicamente conhecida como pertencente à via glicolítica, dimérica em animais e plantas, mas tetramérica em fungos e alguns invertebrados. A proteína ADH (álcool desidrogenase) possui dois domínios principais: o domínio de ligação da coenzima, formado por um motivo estrutural conhecido como *Rossman fold* (seis fitas betas paralelas ligadas por alfa hélices); e o domínio catalítico. Análises filogenéticas mostraram que essa família estrutura-se formando três *clusters* principais, correspondentes a sequências de animais, plantas e fungos. As classes 1 e 2 de ADH de *Caenorhabditis elegans* agruparam-se próximas ao *cluster* monofilético das ADHs de fungos, muito provavelmente porque também são tetraméricas. Em animais e plantas, houve a formação de clados de acordo com o tipo de ADH, já em fungos os agrupamentos devem-se ao tipo de ADH e gênero do organismo. O padrão de evolução dessa família gênica pode ser explicado através do modelo por nascimento e morte. Estudos teóricos de divergência funcional conduzidos nos três grupos de organismos previamente citados indicaram os sítios que, provavelmente, estão submetidos a processos de surgimento de novidades funcionais após a duplicação gênica. As regiões onde foram encontrados os maiores números de aminoácidos divergentes incluem a região de ligação do segundo átomo de zinco, o segmento de interação entre os monômeros e o sítio ativo. Foram construídos dezessete modelos da estrutura tridimensional de ADH em plantas pertencentes a quatro famílias botânicas, a partir da modelagem molecular comparativa. Os resíduos funcionalmente divergentes foram localizados nas estruturas modeladas, tendo sido também encontradas diferenças no potencial eletrostático e no pI (ponto isoelétrico).

ABSTRACT

Alcohol dehydrogenase (ADH) is a gene family known to function in the glycolytic pathway, being dimeric in animals and plants, but tetrameric in fungi and some invertebrates. This protein presents two main domains: one which binds to the coenzyme, formed by a structural motif known as *Rossman fold* (six parallel beta sheets connected by alpha helices); and the catalytic domain. Phylogenetic analyses showed that this family is structured in three main clusters, corresponding to animal, plant, and fungi sequences. *Caenorhabditis elegans* ADHs 1 and 2 are placed near the fungi ADH monophyletic cluster, probably because they are also tetrameric. In mammals and plants clade formation occurs by ADH type, while in fungi it follows ADH type and organism genera. The evolutionary pattern of this gene family can be explained by the birth and death model. Theoretical functional divergence studies conducted in the three previously cited groups of organisms indicated the sites that probably are being submitted to processes involving the emergence of functional novelties after gene duplication. The largest numbers of sites of divergent amino acids were found in the second zinc binding region, the monomer interacting segment, and the active site. Seventeen models of ADH tridimensional structure in plants from four botanical families were built by Comparative Molecular Modeling. The functionally divergent residues were located in the modeled structures and electrostatic and pI (isoelectric point) differences found.

CAPÍTULO 1

INTRODUÇÃO

Evolução Molecular de Famílias Multigênicas

O modelo da estrutura do DNA proposto por Watson e Crick em 1953 propiciou a descoberta de como os genes são copiados e passados através das gerações. Nas cinco décadas seguintes, a biologia molecular foi utilizada com o propósito de elucidar os mecanismos envolvidos na duplicação, transcrição e tradução gênicas, levando a grandes avanços na biologia, genética e medicina (Clayton e Dennis, 2003).

Os genes e suas regiões regulatórias sofrem uma variedade de modificações, incluindo substituições nucleotídicas, duplicações, recombinações e eventos de reparo (Liberles, 2001). Recentemente, o sequenciamento de genomas em diferentes organismos modelo indicou claramente que as duplicações gênicas são mecanismos importantes para a criação de novos genes e sistemas genéticos, levando a uma rápida diversificação de reações catalisadas enzimaticamente, elementos regulatórios complexos e padrões de desenvolvimento, além de serem responsáveis pelo aumento no tamanho do genoma (Nei, 2005; Gogarten e Lorraine, 1999; Ohta, 1987).

Há dois mecanismos principais de produção de genes duplicados: (1) a duplicação do genoma e (2) a duplicação gênica *in tandem*. A duplicação do genoma não necessariamente duplica o número de genes funcionais, já que alguns são silenciados ou perdidos. Acreditava-se que era a forma mais eficiente de aumentar o número de genes no genoma, uma vez que tanto regiões regulatórias quanto codificadoras são duplicadas. Ao contrário do que se pensava, as duplicações *in tandem* não possuem nenhuma desvantagem relativamente às genômicas. Podem produzir centenas de genes se um longo tempo evolutivo for considerado, como nos casos dos genes que codificam os receptores olfativos (Nei, 2005; Niimura e Nei, 2003). As duplicações gênicas podem ser classificadas de acordo com a extensão da região genômica envolvida. A duplicação gênica completa produz duas cópias idênticas e estas sequências podem evoluir de diferentes formas: uma das cópias pode originar um pseudogene por apresentar mutações deletérias (pseudofuncionalização); ou ambas podem ser retidas. A preservação pode levar à persistência das duas cópias com similaridade de sequências, à subfuncionalização (onde cada cópia adota algumas das características do ancestral) ou à neofuncionalização (quando um gene mantém a função original, enquanto o outro adquire uma nova função) (Gonzàlez-Duarte e Albalat, 2005).

A redundância genética, característica presente em todos os genomas estudados até o momento, aumenta de acordo com o aumento do tamanho e da complexidade do genoma. Tanto regiões codificadoras e, particularmente, as regiões não-codificadoras são amplificadas, sendo que as últimas correspondem à maior porção em vertebrados, equivalendo a aproximadamente 95% de todo genoma humano (Santamaria *et al.*, 2004). O genoma de mamíferos contém cerca de 20.000 pseudogenes (Podlaha e Zhang, 2004), número semelhante ao de genes funcionais (cerca de 23.000). Há evidências de que pseudogenes evoluam para genes regulatórios, muito conservados, tal como ocorre com o transcrito do pseudogene *Makorin1-p1* de rato, que regula a estabilidade do transcrito parálogo *Makorin1*. O *Makorin1-p1* possui função biológica e evolui lentamente (Hirotsune *et al.*, 2003; Podlaha e Zhang, 2004; Nei, 2005).

Em regiões codificadoras, os genes podem ser duplicados ou retrotranspostos criando genes parálogos que se agrupam em famílias. Genes que possuem a mesma origem são chamados homólogos, podendo ser classificados em ortólogos, se sua divergência é devido à especiação, ou parálogos, se eles se originaram da duplicação gênica. Alguns parálogos podem manter a mesma função que o gene ancestral, enquanto outros podem adquirir novas funções (Doyle e Gaut, 2000). Assim, a redundância gênica é um importante mecanismo na evolução, criando inovações. Neste cenário, o entendimento das funções específicas de cada membro de uma família gênica, bem como a comparação das mesmas entre as várias espécies, constitui um problema altamente relevante (Santamaria *et al.*, 2004).

Nos genomas de eucariotos há numerosas famílias gênicas, de pequenas a grandes, com número uniforme ou variável de membros. Pesquisas sobre famílias multigênicas, famílias formadas por múltiplos genes de origem comum que codificam produtos de função idêntica ou similar derivadas uma da outra através de duplicação gênica e subsequente divergência, começaram na década de 1970, impulsionadas pela curiosidade de saber como a variabilidade genética gerada em um indivíduo ou existente em uma população está relacionada à redundância gênica. O número e tamanho dessas famílias variam enormemente de um organismo a outro (Rubin *et al.*, 2000). Na maior parte dos genomas sequenciados até o momento, as famílias contêm poucos membros (dois a cinco), mas há exemplos de famílias com mais de 50-100 membros. As maiores famílias envolvidas em funções celulares críticas, como genes ribossomais e RNA transportador mostram pouca heterogeneidade de sequências, contrastando com aquelas envolvidas em respostas versáteis do organismo a mudanças ambientais, como os genes transportadores em *Escherichia coli* (Blattner *et al.*, 1997), os antígenos de superfície de *Mycobacterium tuberculosis* (Cole *et al.*, 1998) e de *Plasmodium* (Su *et al.*, 1995; Mercereau-Puijalon *et al.*, 2002) e receptores olfativos em eucariotos (Li *et al.*, 2001). As famílias multigênicas constituem uma parte significativa do conteúdo do genoma de bactérias, arqueobactérias e eucariotos (Klenk *et al.*, 1997; Rubin *et al.*, 2000; Li *et al.*, 2001).

A homologia de sequência entre membros de uma família multigênica depende principalmente da taxa relativa de ocorrência de mutações e eventos ilegítimos de *crossing-over*, além de ser influenciada em alguma extensão pela seleção natural e deriva genética (Wagner, 2002; Ohta, 2003). Além disso, genes pertencentes a uma família multigênica estão, usualmente, sob o mesmo controle regulatório. Portanto, é interessante estudar como diferentes padrões de expressão são adquiridos.

Uma grande superfamília pode conter genes únicos ou várias famílias multigênicas. Praticamente todos os genes estudados pertencem a alguma família gênica. A busca por homologia tornou-se uma ferramenta importante para identificação das relações genéticas, assim como a análise filogenética, que é útil para o entendimento das relações entre os membros de uma mesma família, já que árvores de genes esclarecem a história dos eventos de duplicação e fornecem informação para a identificação de ortólogos em estudos comparativos de genes homólogos envolvendo várias espécies (Ohta, 2003). O número de cópias, a distribuição dentro do genoma, a diversidade de sequência, o perfil de expressão e abordagens comparativas, tais como comparação de posições, colinearidade dos genes e substituições nucleotídicas sob efeito de pressão seletiva em organismos relacionados são informações úteis para a caracterização de famílias multigênicas (Perovic *et al.*, 2007).

Os membros de uma família multigênica podem estar agrupados com repetições *in tandem*. Esse agrupamento surge como resultado de *croosing-over* desigual durante a meiose ou mitose de linhagens celulares germinativas. Exemplos são famílias codificadoras de histonas, RNA e superfamília das globinas. Existem superfamílias que apresentam membros que formam agrupamentos e outros que estão dispersos pelo genoma. Nesses casos, os genes agrupados, geralmente, formam famílias multigênicas com funções sobrepostas, enquanto as cópias dispersas possuem funções mais diversificadas. A superfamília das imunoglobulinas e a dos receptores olfativos são exemplos clássicos. Finalmente, há famílias gênicas que se apresentam completamente dispersas pelo genoma. Acredita-se que surjam através de transcrição reversa do RNA e subsequente integração ao genoma. Enquadram-se nesse caso, a argininosucinato sintetase, citocromo c e β-tubulina (Ohta e Dover, 1983).

Diferentes modelos de evolução tentam explicar a evolução de famílias multigênicas. Dentre eles, o modelo da evolução divergente implica a divergência gradual de proteínas filogeneticamente relacionadas, que posteriormente podem adquirir novas funções gênicas. Um segundo modelo, chamado evolução em concerto, surgiu como explicação para a maior similaridade das regiões intergênicas codificadoras do RNA ribossomal dentro de uma mesma espécie que entre duas espécies relacionadas (Ohta, 1985; Brown *et al.*, 1972). Assim, nesse modelo, todos os membros de uma família multigênica evoluem conjuntamente, de forma que uma mutação espalha-se por todos os membros da família através de eventos de permuta ou conversão gênica (Nagylaki, 1984; Ohta, 1984). A variação genética dentro da família pode "migrar" entre diferentes cópias e, eventualmente, um alelo pode ser fixado na espécie (Mano e Innan, 2008). Esse modelo foi utilizado para explicar a evolução de várias famílias. Entretanto, sua aplicabilidade a algumas famílias gênicas começou a ser questionada; consequentemente, um novo modelo chamado evolução por nascimento e morte foi proposto (Nei e Hughes, 1992; Nei *et al.*, 1997). Nesse modelo, novos genes são criados por eventos de duplicação gênica, sendo que alguns são mantidos no genoma durante um longo período, enquanto outros são deletados ou tornam-se não funcionais. Um dos mais bem estudados casos é o sistema imune adaptativo (SIA) em vertebrados, que inclui as famílias do complexo de histocompatibilidade (MHC), imunoglobulinas (Ig) e receptores de célula T (TCR). A evolução do SIA inicialmente ocorreu através de mudanças em genes não relacionados à resposta imune e, posteriormente, esses genes reuniram-se para formar o sistema imune adaptativo. Outro exemplo largamente estudado é dos genes dos receptores olfativos. Análises filogenéticas demonstraram que formam vários grupos altamente divergentes originados por duplicação gênica. Um importante sistema genético para o desenvolvimento de plantas e animais é a superfamília dos genes homeobox. É uma família muito antiga e compartilhada por animais, plantas e fungos. Em animais, essa superfamília compreende pelo menos 49 famílias com diferentes papéis no desenvolvimento, como HOX (responsável pelo padrão corporal) e PAX6 (envolvida na formação do olho) (Gehring e Ikeo, 1999; Burglin *et al.*, 1997). Em plantas, genes que codificam as proteínas regulatórias R e MADS-box, as proteínas de choque térmico e as proteínas de ligação à clorofila a e b, entre outras, foram extensivamente estudados nos últimos anos. A maioria dessas famílias possui diversos *loci* e apresenta uma grande variação no número de cópias entre as espécies. As famílias gênicas em plantas variam de pequenas famílias com poucos *loci*, como muitas enzimas metabólicas das famílias *Adh* e *rbcS*, a grandes famílias com centenas de *loci*, como as proteínas de choque térmico. Análises filogenéticas indicam que muito dessa variação pode ser atribuída a duplicações recentes. Portanto, a dinâmica evolutiva dessas famílias, como detalhado acima, apresenta flutuações no número de cópias através de eventos múltiplos de duplicação e deleção (Small e Wendel, 2000a; Morton *et al.*, 1996).

As famílias gênicas que apresentam produtos variados são usualmente sujeitas à evolução por nascimento e morte, já que esse modelo explica a variação genética. No entanto, é interessante observar que algumas famílias multigênicas, tais como a de proteínas de choque térmico, estão sujeitas a processos mistos de evolução, podendo evoluir através da evolução em concerto e da evolução por nascimento e morte. Além disso, algumas famílias evoluem segundo o modelo de nascimento e morte e estão submetidas a uma forte seleção purificadora, ao passo que outras evoluem de acordo com o mesmo modelo submetidas à seleção positiva (Nei e Rooney, 2005).

Se a evolução em concerto é o fator principal na evolução de uma determinada família, tanto o número de substituições sinônimas por sítio sinônimo (dS) quanto o número de substituições não-sinônimas por sítio não-sinônimo (dN) precisa ser nulo, uma vez que a evolução em concerto afeta os sítios sinônimos e os não-sinônimos de forma similar. No entanto, se a seleção purificadora estiver atuando verifica-se um valor de ω (dN/dS)<1,0; enquanto que um valor de ω (dN/dS)>1,0 indica a atuação da seleção positiva. No primeiro caso, as mutações não-sinônimas são removidas por seleção purificadora por causarem um efeito detrimental na função da proteína codificada; já no segundo caso, a proteína está sob efeito da seleção diversificadora favorável ao aumento da diversidade de aminoácidos. Assim, o valor de ω pode dar importantes evidências sobre a função de um gene, além de indicar quais resíduos de aminoácidos específicos são funcionalmente importantes.

Divergência Funcional

O estudo da função dos genes usando abordagens evolutivas é o objetivo da Genômica Funcional Evolutiva (Golding e Dean, 1998). Sob o ponto de vista da genética de populações, o termo função pode ser quantificado como a intensidade de seleção (S) aplicada a qualquer nível (resíduos de aminoácidos, genes, etc.). Assim, a "função evolutiva" é definida como um parâmetro da genética de populações que contribui para o *fitness* do organismo e está relacionada à função bioquímica, fenotípica ou estrutural de um gene. Isso envolve interações entre moléculas e diferentes níveis de organização biológica, incluindo complexos moleculares, rotas metabólicas e, eventualmente, engloba indivíduos, populações e espécies. Neste contexto, a maior contribuição da Genômica Funcional Evolutiva é a geração de hipóteses, a partir de sequências genômicas, que possam ser experimentalmente validadas (Gu, 2003).

O tópico central de interesse desta Tese foi o estudo da divergência funcional após duplicação gênica ou especiação da família gênica da álcool desidrogenase. Pesquisas objetivando compreender o processo de duplicação de genes, o seu significado e impacto na função dos produtos gênicos têm sido realizadas tendo como objeto de estudo as mais diversas famílias gênicas. O modelo clássico de duplicação gênica sugere que uma cópia manteria a função original, enquanto a outra estaria livre para acumular mudanças de aminoácidos. Muitos modelos surgiram posteriormente para explicar esse processo (Force *et al.*, 1999; Nei *et al.*, 1997; Fryxell, 1996; Clark, 1994; Hughes, 1994); porém, os detalhes da diversificação funcional entre genes duplicados permanecem desconhecidos (Gu *et al.*, 2002b).

Altos valores de intensidade de seleção indicam importância funcional (baixas taxas evolutivas, no caso de conservação da função); consequentemente, mudanças sítio-específicas nas taxas evolutivas (ou intensidade de seleção S) podem ser interpretadas como mudanças funcionalmente importantes. Adicionalmente, sabe-se que as taxas evolutivas podem variar durante a evolução para um resíduo de aminoácido específico.

Assim, tal resíduo pode mudar de muito conservado para altamente variável ou vice-versa (Gu, 2003). A detecção de variabilidade nas taxas evolutivas entre os genes é importante no estudo da divergência funcional de uma família protéica. Isto pode ser feito através da detecção de mudanças nas taxas evolutivas sítio-específicas, usando um alinhamento múltiplo de sequências de aminoácidos para uma determinada árvore filogenética. Se a função ou a estrutura da proteína está mudando, alguns resíduos de aminoácidos podem alterar as restrições funcionais durante a evolução. Isto implica que taxas evolutivas nesses sítios poderão variar em genes homólogos diferentes de uma mesma família gênica. Esta é a chamada divergência funcional do *Tipo I* (Gu e Vander Velden, 2002).

A divergência *Tipo I* é mais provável nos resíduos com diferentes taxas evolutivas entre os genes duplicados. Entretanto, em função da natureza estocástica da evolução molecular, a precisão de qualquer classificação *ad hoc* de aminoácidos será limitada. Além disso, existe uma grande sensibilidade a filogenias desbalanceadas entre dois agrupamentos de genes. Por exemplo, considerando dois grupos monofiléticos com o mesmo número de sequências, sendo que o grupo 1 possui sequências filogeneticamente próximas, enquanto o grupo 2 é composto por sequências distantes, um escore simples que ignore a filogenia pode induzir ao erro, já que muitos sítios serão invariáveis no grupo 1. Esse problema torna-se sério quando análise envolvendo um grande número de sequências é realizada, tendo em vista que nessa situação a inspeção visual torna-se quase impossível. Portanto, é necessária a utilização de modelagem estatística e predição baseada em uma árvore de genes (Gu e Gu, 2003; Gu *et al.*, 2002b).

Considerando-se uma filogenia com dois grupos monofiléticos gerados por duplicação gênica ou especiação, é proposto que um resíduo de aminoácido possa ter dois estados: em um estado (*S0*), o sítio tem a mesma taxa em ambos os grupos e, em outro (*S1*), as taxas nos dois grupos são tão diferentes que podem ser consideradas estatisticamente independentes. Em cada estado, as taxas evolutivas entre os sítios variam de acordo com uma distribuição gama. λ_A e λ_B identificam as taxas evolutivas para os sítios A e B. O coeficiente de divergência funcional (θ) entre os dois grupos é definido como a probabilidade de um sítio ser *S1*, isto é, $\theta = P\ (S1)$. A rejeição da hipótese nula, $\theta = 0$, sugere que as taxas evolutivas (ou as restrições funcionais) em alguns sítios têm variado entre os dois grupos significativamente, sendo tais sítios relevantes nas diferenças funcional-estruturais das proteínas (Gu e Vander Velden, 2002; Gu, 1999). A rejeição da hipótese nula H0: $\theta = 0$ fornece uma evidência estatística para mudanças nas taxas evolutivas ou nas restrições funcionais. Sendo $Q(k)=P_k(S_1|X)$ a probabilidade *a posteriori* de um sítio *k* ser S_1 (*status* relacionado à divergência funcional) quando uma configuração de aminoácido (*X*) é observada. Nesse caso, temos que $\lambda_A \neq \lambda_B$ (Gaucher *et al.*, 2002b). Se o *status* observado é o alternativo S_0 (não relacionado à divergência funcional), com probabilidade posterior $P_k(S_0|X)=1-P_k(S_1|X)$, isso significa que não há restrição funcional alterada e os resíduos preditos serão significativos somente quando $Q(k)>0.5$, no caso em que a taxa é $R(S_1|S_0)=P(S_1|X)/P(S_0|X)>1$. Um valor de corte mais rigoroso seria $Q(k)>0.67$, ou $R(S_1|S_0)\geq 2$ (Gu e Gu, 2003). Para o estado S_0 assume-se $\lambda_A = \lambda_B$.

Existem três tipos de divergência funcional adicionais à de *Tipo I*. *Tipo 0* representa configurações de aminoácidos que estão universalmente conservados em toda a família gênica, implicando que são resíduos importantes para a função compartilhada por todos os membros da família. A divergência do *Tipo II* refere-se a configurações de aminoácidos que são muito conservadas em ambas as cópias gênicas, mas possuem propriedades bioquímicas muito diferentes, como por exemplo: cargas positivas vs. cargas negativas;

portanto, podem ser responsáveis por especificação funcional. Gu classificou uma quarta classe, nomeando-a *Tipo U*, que enquadraria aquelas configurações de difícil padronização (Gu *et al.*, 2002a, 2002b). Dois modelos para evolução de sequências, um utilizando Cadeias de Markov e outro baseado em Poisson, foram desenvolvidos por Gu e colaboradores, cujas estruturas podem ser entendidas em parte através do fluxograma da *Figura 1*.

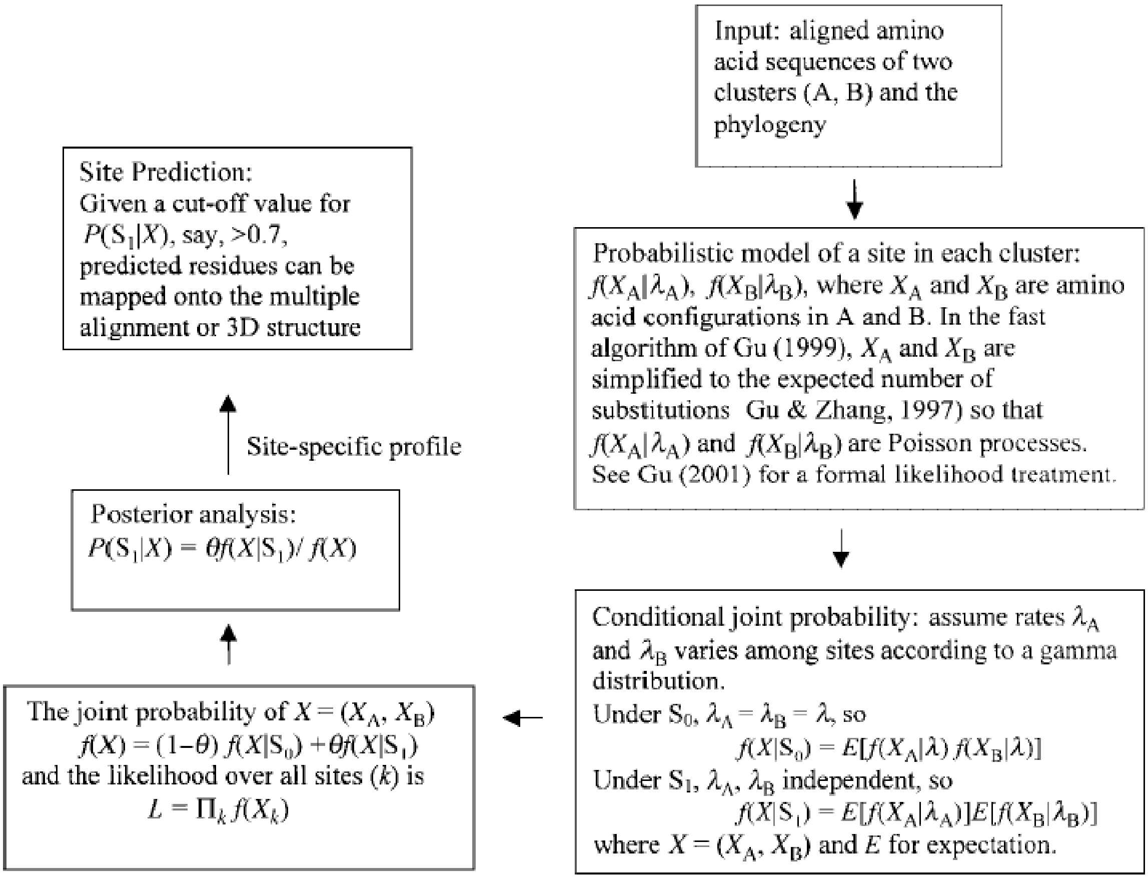

Input: aligned amino acid sequences of two clusters (A, B) and the phylogeny
Probabilistic model of a site in each cluster: $f(X_A|\lambda_A)$, $f(X_B|\lambda_B)$, where X_A and X_B are amino acid configurations in A and B. In the fast algorithm of Gu (1999), X_A and X_B are simplified to the expected number of substitutions Gu & Zhang, 1997) so that $f(X_A|\lambda_A)$ and $f(X_B|\lambda_B)$ are Poisson processes. See Gu (2001) for a formal likelihood treatment.
Conditional joint probability: assume rates λ_A and λ_B varies among sites according to a gamma distribution.
Under S_0, $\lambda_A = \lambda_B = \lambda$, so
$f(X|S_0) = E[f(X_A|\lambda)\,f(X_B|\lambda)]$
Under S_1, λ_A, λ_B independent, so
$f(X|S_1) = E[f(X_A|\lambda_A)]E[f(X_B|\lambda_B)]$
where $X = (X_A, X_B)$ and E for expectation.
The joint probability of $X = (X_A, X_B)$
$f(X) = (1-\theta)\,f(X|S_0) + \theta f(X|S_1)$
and the likelihood over all sites (k) is
$L = \Pi_k\, f(X_k)$
Posterior analysis:
$P(S_1|X) = \theta f(X|S_1)/\,f(X)$
Site-specific profile
Site Prediction:
Given a cut-off value for $P(S_1|X)$, say, >0.7, predicted residues can be mapped onto the multiple alignment or 3D structure

Figura 1 – Fluxograma ilustrando os métodos de Gu para inferência de divergência funcional. Figura adaptada de Gu (2003).

A seguir, um maior detalhamento matemático do método será dado.

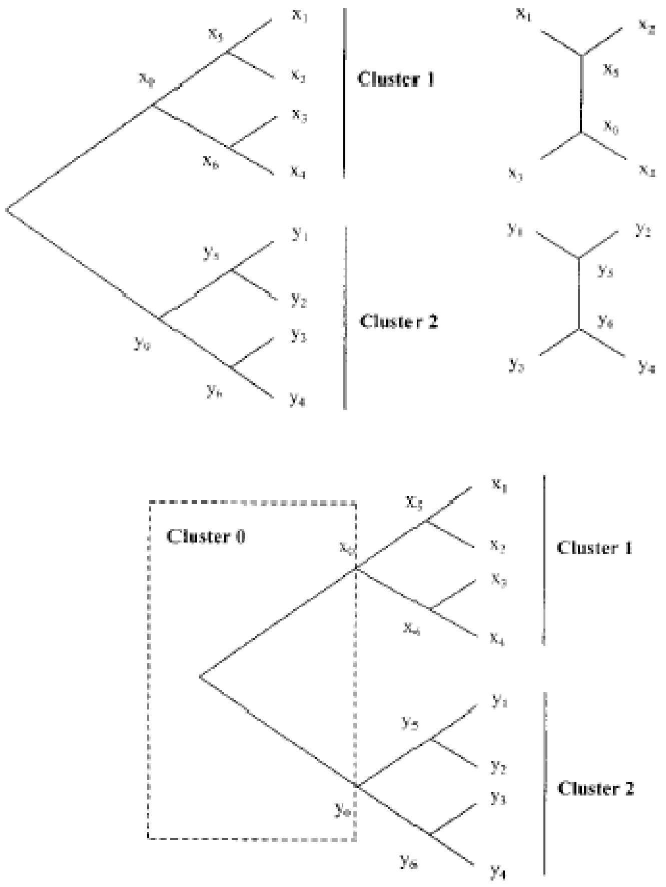


Figura 2 – Agrupamentos de genes para as subárvores (parte superior da figura) e a árvore completa (parte inferior da figura). O estado logo após a duplicação gênica é chamado de Cluster 0. Figura adaptada de Gu (2001a).

Considere a árvore dada na *Figura 2*, a máxima verossimilhança para a evolução das sequências pode ser derivada como segue. A matriz de probabilidade de transição para um dado período de tempo t computada como $\mathbf{P} = e^{\lambda \mathbf{R} t}$, onde **R** representa o padrão de substituições de aminoácidos, que pode ser determinado empiricamente através de modelos, tais como o de Dayhoff (Dayhoff *et al.*, 1978) e JTT (Jones *et al.*, 1992). A taxa evolutiva (λ) varia entre os sítios em função das diferentes restrições funcionais. Geralmente, λ é tratada como uma variável randômica que segue uma distribuição gama, ou seja,

$$\phi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \qquad (1)$$

onde a forma do parâmetro α descreve a intensidade da taxa de variação entre os sítios, isto é, um pequeno valor indica uma alta taxa de heterogeneidade entre os sítios e quando α = ∞ significa que não há taxa de variação entre eles. β é uma constante.

Novamente considere a árvore da *Figura 2*. Sejam X = (x_1, x_2, x_3, x_4) e Y = (y_1, y_2, y_3, y_4) as configurações de aminoácidos observadas em um sítio para os *clusters* (agrupamentos) 1 e 2, respectivamente. Para subárvores não enraizadas, *clusters* 1 ou 2, a probabilidade condicional de se observar X ou Y em um sítio pode ser escrita como

$$f(X|\lambda) = \sum_{x_5=1}^{20}\sum_{x_6=1}^{20} b_{x_5}P_{x_5x_1}P_{x_5x_2}P_{x_5x_6}P_{x_6x_3}P_{x_6x_4}$$
$$f(Y|\lambda) = \sum_{y_5=1}^{20}\sum_{y_6=1}^{20} b_{y_5}P_{y_5y_1}P_{y_5y_2}P_{y_5y_6}P_{y_6y_3}P_{y_6y_4}, \qquad (2)$$

onde $P_{ij} = P_{ij}\ (v_{ij})$ é a probabilidade de transição de um nodo *i* para um nodo *j*, *vij* é o comprimento do ramo entre eles a b_i é a frequência do aminoácido *i*. Integrando em função da variável λ, a probabilidade de observar X ou Y em um determinado sítio é dada por

$$p(X) = E[f(X|\lambda)] = \int_0^{\infty} f(X|\lambda)\phi(\lambda)\, d\lambda$$
$$p(Y) = E[f(Y|\lambda)] = \int_0^{\infty} f(Y|\lambda)\phi(\lambda)\, d\lambda, \qquad (3)$$

respectivamente, onde E é a esperança.

Utilizando um modelo de dois estados, há dois estados combinados não degenerados (configurações de divergência funcional), denotados por $S_0 = \{(F_0, F_0)\}$ e $S_1 = \{(F_0, F_1),(F_1, F_0),(F_1, F_1)\}$. A notação *F* descreve o *status* em um único *cluster*, enquanto *S* é usado para a configuração de divergência funcional de uma família gênica.

Uma vez que as taxas evolutivas (λ_1 e λ_2) em um sítio no estado S_1 são estatisticamente independentes entre dois *clusters*, enquanto que são completamente correlacionadas ($\lambda_1 = \lambda_2 = \lambda$) em um sítio no estado S_0, a probabilidade condicional total no estado S_0 ou S_1 é dada por

$$\begin{aligned} f^*(X, Y \mid S_0) &= \int_0^\infty f(X \mid \lambda) f(Y \mid \lambda) \phi(\lambda)\, d\lambda \\ &= E[f(X \mid \lambda) f(X \mid \lambda)] \\ f^*(X, Y \mid S_1) &= p(X)p(Y) \\ &= E[f(X \mid \lambda_1)] \times E[f(Y \mid \lambda_2)] \end{aligned} \quad (4)$$

onde $f\,(X \mid \lambda_1)$ ou $f\,(X \mid \lambda_1)$ é a verossimilhança de cada subárvore não enraizada dadas pelas equações (2). O asterisco serve para diferenciar a verossimilhança da subárvore da verossimilhança da árvore completa. Portanto, para um modelo de dois estados, a probabilidade total dadas duas subárvores pode ser escrita como

$$p^*(X, Y) = (1 - \theta_{12}) f^*(X, Y \mid S_0) + \theta_{12} f^*(X, Y \mid S_1). \quad (5)$$

Então, sob a suposição de sítios independentes, a função de verossimilhança considerando todos os sítios, excluindo os *gaps* (falhas), é dada por

$$L^*(\mathbf{x} \mid \text{data}) = \prod_k p^*(X^{(k)}, Y^{(k)}) \quad (6)$$

onde *k* indica o sítio e **x** é o conjunto de dados de parâmetros desconhecidos.

O algoritmo numérico desenvolvido por Gu e colaboradores utiliza algumas simplificações a fim de diminuir o tempo computacional, já que a computação da probabilidade total de duas subárvores ($p^*(X, Y)$) envolve a árvore filogenética, o comprimento dos ramos, parâmetro α de uma distribuição gama e o coeficiente de divergência funcional (Gu, 2001a; Gu, 2001b).

Para uma família multigênica com muitos membros (vários *clusters*), o padrão de configuração de aminoácidos é complicado. Mesmo considerando três *clusters*, a configuração de aminoácidos do *Tipo I* envolve vários subtipos: *cluster* 1 variável, *clusters* 2 e 3 conservados, etc. Assim, a verossimilhança da subárvore pode ser estendida para *n* *clusters*, mas o tempo computacional cresce enormemente com o aumento de *n*.

Considerando três agrupamentos (*n* = 3), as configurações funcionais equivalem a 2^3=8, sendo elas: (F_0, F_0, F_0), (F_0, F_0, F_1), (F_0, F_1, F_0), (F_1, F_0, F_0), (F_0, F_1, F_1), (F_1, F_0, F_1), (F_1, F_1, F_0) e (F_1, F_1, F_1). A primeira, segunda e terceira posição de cada conjunto indica o *status* dos *clusters* 1, 2 e 3, respectivamente. O segundo, terceiro e quarto conjuntos referem-se a *clusters* gênicos não degenerados. Os quatro últimos estados devem ser degenerados a um único estado S_4 porque λ_1, λ_2 e λ_3 são mutuamente independentes. Note que há m = 2^n – n estados combinados degenerados (configurações de divergência funcional) no caso de *n clusters*, que são denotados por S_j, com $j = 0, ..., m -1$. De fato, a relação entre as taxas evolutivas (λ_1, λ_2, λ_3) entre os três *clusters* é mostrada na *Tabela 1.*

Tabela 1 – Estados combinados (configurações de divergência funcional) para a verossimilhança de subárvore com três *clusters* gênicos. Adaptado de Gu (2001a).

Estado S_i	$F_0\,/F_1$	*P (Si)*	Independência das taxas evolutivas[1]	*Tipo I*[2]
S_0	(F_0, F_0, F_0)	π_0	$\lambda_1 = \lambda_2 = \lambda_3$	*No*
S_1	(F_1, F_0, F_0)	π_1	$\lambda_1, \lambda_2 = \lambda_3$	*Cluster 1*
S_2	(F_0, F_1, F_0)	π_2	$\lambda_1 = \lambda_3, \lambda_2$	*Cluster 2*
S_3	(F_0, F_0, F_1)	π_3	$\lambda_1 = \lambda_2,\ \lambda_3$	*Cluster 3*
S_4	(F_0, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	*Clusters 2 e 3*
	(F_1, F_0, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	*Clusters 1 e 3*
	(F_1, F_1, F_0)	π_4	$\lambda_1, \lambda_2, \lambda_3$	*Clusters 1 e 2*
	(F_1, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	*Clusters 1, 2 e 3*

[1] Quando duas taxas evolutivas são iguais significa que são não independentes. [2] Indica qual *cluster* está sob divergência funcional *Tipo I*.

Assim, a probabilidade total das três subárvores sob cada configuração funcional (S_j) é dada por

$$\begin{aligned}
f^*(X|S_0) &= E[f(X_1|\lambda)f(X_2|\lambda)f(X_3|\lambda)]\\
f^*(X|S_1) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda)f(X_3|\lambda)]\\
f^*(X|S_2) &= E[f(X_2|\lambda_2)] \times E[f(X_1|\lambda)f(X_3|\lambda)]\\
f^*(X|S_3) &= E[f(X_3|\lambda_3)] \times E[f(X_1|\lambda)f(X_2|\lambda)]\\
f^*(X|S_4) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda_2)]\\
&\quad \times E[f(X_3|\lambda_3)], \qquad (7)
\end{aligned}$$

onde $f\,(X_1 \mid \lambda), f\,(X_2 \mid \lambda)$ ou $f\,(X_3 \mid \lambda)$ é a verossimilhança para a subárvore não enraizada de cada *cluster* gênico, respectivamente. Seja π_j a probabilidade de um sítio no estado S_j, ou seja, $\pi_j = P\,(S_j)$; então, a probabilidade total das três subárvores para um determinado sítio é equivalente a

$$p^*(X) = \sum_{j=0}^{m-1} \pi_j f^*(X|S_j), \qquad (8)$$

onde m = 5. De maneira geral, π_j são denominados coeficientes de divergência funcional do *Tipo I* para configuração de divergência funcional S_j. Em particular,

$$\pi_f = 1 - \pi_0 = \sum_{j=1}^{m-1} \pi_j \qquad (9)$$

é o coeficiente de divergência funcional do *Tipo I* da família gênica. A equação (5) é um caso especial da equação (9) quando n = 2 (e m = 2), e $\pi_0 = 1 - \theta_{12}$ *e* $\pi_1 = \theta_{12}$.

Tabela 2 – Estados combinados (configurações de divergência funcional) para a verossimilhança da árvore com dois *clusters* gênicos. Adaptado de Gu (2001a).

Estado S_i	F_0 / F_1	*P (Si)*	Independência das taxas evolutivas	Divergência Funcional
S_0	(F_0, F_0, F_0)	π_0	$\lambda_1 = \lambda_2 = \lambda_3$	No
S_1	(F_1, F_0, F_0)	π_1	$\lambda_1, \lambda_2 = \lambda_3$	Tipo II
S_2	(F_0, F_1, F_0)	π_2	$\lambda_1 = \lambda_3, \lambda_2$	Tipo I
S_3	(F_0, F_0, F_1)	π_3	$\lambda_1 = \lambda_2, \lambda_3$	Tipo I
S_4	(F_0, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Tipo I
	(F_1, F_0, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Tipo I
	(F_1, F_1, F_0)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Tipo I
	(F_1, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Tipo I

[1] Quando duas taxas evolutivas são iguais significa que são não independentes. [2] Indica o tipo de divergência.

Para avaliar a divergência funcional do *Tipo II*, considera-se um ramo interno (*cluster 0*) entre os *clusters 1* e *2* (ver *Figura 2*, parte inferior). Cada *cluster* possui dois estados possíveis, exatamente como no caso do *Tipo I*; portanto, 2^3 = 8 estados combinados são possíveis, podendo ser degenerados a cinco configurações de divergência funcional. As relações entre os λs são mostradas na *Tabela 2*. Seja π_j (*j = 0, 1, ..., 4*) a probabilidade de um sítio no estado S_j, ou seja, $\pi_j = P\ (S_j)$; então, a probabilidade condicional de X e Y é dada por

$$f(X,\ Y|\lambda) = \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} P_{x_0 y_0}(\upsilon|\lambda_0) f(X|\lambda_1;\ x_0) f(Y|\lambda_2;\ y_0), \tag{10}$$

onde $f\ (X \mid \lambda_1;\ x_0),\ f\ (Y \mid \lambda_2;\ y_0)$ são as funções de verossimilhança para os *clusters 1* e *2*; x_0 *e* y_0 são as raízes e v é o comprimento do ramo interno. Assim, temos

$$\begin{aligned} f(X|\lambda; x_0) &= \sum_{x_5} \sum_{x_6} P_{x_0x_5} P_{x_5x_1} P_{x_5x_2} P_{x_0x_6} P_{x_6x_3} P_{x_6x_4} \\ f(Y|\lambda; y_0) &= \sum_{y_5} \sum_{y_6} P_{y_0y_5} P_{y_5y_1} P_{y_5y_2} P_{y_0y_6} P_{y_6y_3} P_{y_6y_4}. \end{aligned} \tag{11}$$

A probabilidade condicional de X e Y sob cada estado combinado é dada por

$$\begin{aligned} f(X,\ Y|S_0) &= \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} \\ &\quad \times E[P_{x_0y_0}(\upsilon|\lambda_0) f(X|\lambda;\ x_0) f(Y|\lambda;\ y_0)] \\ f(X,\ Y|S_1) &= \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0)] \\ &\quad \times E[f(X|\lambda;\ x_0) f(Y|\lambda;\ y_0)] \\ f(X,\ Y|S_2) &= \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} E[f(X|\lambda_1;\ x_0)] \\ &\quad \times E[P_{x_0y_0}(\upsilon|\lambda_0) f(Y|\lambda;\ y_0)] \\ f(X,\ Y|S_3) &= \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0) f(X|\lambda;\ x_0)] \\ &\quad \times E[f(Y|\lambda_2;\ y_0)] \\ f(X,\ Y|S_4) &= \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0)] \times E[f(X|\lambda_1; x_0)] \\ &\quad \times E[f(Y|\lambda_2;\ y_0)]. \end{aligned} \tag{12}$$

Portanto, a probabilidade total de X e Y pode ser genericamente descrita como

$$p(X,\ Y) = \sum_{j=0}^{m-1} \pi_j f(X,\ Y | S_j), \quad (13)$$

onde *m* = 5. Se não houver restrições funcionais alteradas entre os dois agrupamentos, isso significa que *cluster 0* está no estado F_1 e os *clusters 2* e *3* estão no estado F_0. Consequentemente, o coeficiente de divergência funcional pode ser definido como $\theta_{II} = P\ (S_1) = P\ (F_1, F_0, F_0) = \pi_1$. Por outro lado, divergência do *Tipo I* significa que pelo menos um agrupamento deve estar no estado F_1. De acordo com a *Tabela 2*, o coeficiente de *Tipo I* é dado por $\theta_I = \pi_2 + \pi_3 + \pi_4$. Adicionalmente, se o coeficiente total é definido como $\pi_f = 1 - P(S_0) = 1 - \pi_0$, tem-se $\theta_I + \theta_{II} = 1 - \pi_0$. Assim, π_0 pode ser chamado de coeficiente de restrição funcional da família gênica.

Utilizando a árvore completa pode-se identificar um perfil sítio-específico para divergências do *Tipo I* e *II*. No caso de dois agrupamentos, a probabilidade de cada estado combinado não degenerado S_i pode ser computado como

$$P(S_i | X,\ Y) = \frac{\pi_i f(X,\ Y | S_i)}{\sum_{j=0}^{m-1} \pi_j f(X,\ Y | S_j)}, \quad i = 0, 1, \ldots, 4, \quad (14)$$

onde $\pi_i = P\ (S_i)$ e $f\ (X\ |\ Y\ |\ S_j)$ é dada pela equação (12). Então, os perfis sítio-específicos para *Tipos I* e *II* são dados, respectivamente, por

$$\begin{aligned} P(\text{type I} \mid X,\ Y) &= P(S_2 \mid X,\ Y) + P(S_3 \mid X,\ Y) \\ &\quad + P(S_4 \mid X,\ Y) \\ P(\text{type II} \mid X,\ Y) &= P(S_1 \mid X,\ Y), \end{aligned} \qquad (15)$$

A metodologia acima descrita foi implementada no programa DIVERGE (Gu e Vander Velden, 2002; Gu 2001b). Tal software possui uma ferramenta para visualização da estrutura tridimensional da proteína. De forma que, aminoácidos preditos como sendo importantes na divergência estrutural e/ou funcional desta proteína, podem ser visualizados na estrutura tridimensional, permitindo inferências sobre as relações entre os resíduos.

O desempenho desse algoritmo já foi testado por vários estudos de caso (Gu, 1999; Wang e Gu, 2001), tendo sido verificado que, para obter a maior eficiência na detecção da divergência funcional entre resíduos, o uso da sequência de dados deve satisfazer as seguintes condições: (1) cada grupo precisa ter pelo menos quatro sequências; (2) exceto para um grande número de dados, deve-se ter cautela quanto aos resultados quando a identidade entre todas as sequências for superior a 90%, devido à perda do poder estatístico; e (3) devem ser utilizados alinhamentos múltiplos (Gu e Vander Velden, 2002). Todas as condições acima foram satisfeitas nesta Tese.

Vários algoritmos foram propostos para definir os tipos de divergência funcional (Landgraf *et al.*, 1999; Lichtarge *et al.*, 1996; Casari *et al.*, 1995). Casari e colaboradores utilizaram análise vetorial do perfil das sequências para identificar resíduos importantes. O grupo de Lichtarge desenvolveu um método chamado Rastreamento evolutivo, que foi melhorado posteriormente por Landgraf e colaboradores e recebeu o nome de Rastreamento evolutivo ponderado. Nesses métodos, o grau de conservação em cada posição recebe um valor para diferentes subfamílias e pode-se visualizar na proteína, através de colorações distintas, quando o resíduo é mais ou menos conservado. Adicionalmente, muitos métodos utilizando abordagens de evolução molecular estão disponíveis para predições semelhantes às funcionais (*functional-like*) (Pollock *et al.*, 1999; Golding e Dean, 1998). No entanto, o método desenvolvido por Gu e colaboradores foi o único a propor uma abordagem baseada em filogenias para predizer estatisticamente as configurações do Tipo *I* e *II*.

Tal metodologia já foi aplicada para o estudo da duplicação de diversas famílias gênicas, tais como as ciclooxigenases (Gu, 2001a), caspases (Wang e Gu, 2001), fatores de transcrição (Gaucher *et al.*, 2002a), globinas (Naylor e Gerstein, 2000), quinases (Gu *et al.*, 2002a), α-manosidades classe I (Jordan *et al.*, 2001), opsinas (Spaethe e Briscoe, 2004), proteína G (Zheng *et al.*, 2007), proteínas RAMP (moduladores de proteína G) (Benítez-Páez e Cárdenas-Brito, 2008), transportadores monocarboxilados (Liu *et al.*, 2008), bestrofinas (Milenkovic *et al.*, 2008), fosforilases (Georgelis *et al.*, 2008), receptores *toll-like* (Zhou *et al.*, 2007), proteínas de ligação ao RNA (TDP) (Wang *et al.*, 2004), tirosina quinases (Gu e Gu, 2003) e transferinas (Gu, 1999).

Modelagem Molecular no estudo da Evolução de Proteínas

Há vários métodos disponíveis que avaliam divergência funcional sob o ponto de vista da evolução molecular, mas não implementam facilidades computacionais que propiciem o estudo da evolução, função e estrutura de proteínas (Gu, 2003). A metodologia de detecção de divergência funcional utilizada nos estudos presentes nesta Tese foi a de Gu (2003, 2001a, 2001b, 1999). O programa desenvolvido pelo grupo de Gu fornece uma ferramenta gráfica através da qual é possível exibir a estrutura tridimensional de uma proteína, quando esta se encontra disponível nos bancos de dados, resultante de estudos de difração de raios-X ou ressonância magnética nuclear. No entanto, tal ferramenta não é suficiente para o estudo da relação entre evolução, função e estrutura protéica, necessitando muitos aprimoramentos até que possa ser utilizada como principal fonte geradora de resultados. Assim sendo, um dos desafios encontrados nesse trabalho foi a determinação da estrutura tridimensional das proteínas de estudo (álcool desidrogenases), já que as mesmas não estão disponíveis nos bancos de dados para diversos grupos de organismos, especialmente plantas. De forma que as sequências de interesse foram modeladas pelo método de Modelagem Molecular Comparativa por Homologia.

A Modelagem Comparativa por Homologia tem como principal objetivo a construção de um modelo tridimensional para uma proteína de estrutura desconhecida, com base na estrutura tridimensional (3D) conhecida de proteínas (moldes) cujas sequências primárias sejam semelhantes à da proteína a ser modelada. Esta modelagem só é possível porque pequenas mudanças na sequência protéica, geralmente resultam em pequenas mudanças na estrutura 3D (Sánchez e Sali, 1997). Este é o método mais apurado de predição de estruturas tridimensionais (Sánchez e Sali, 1997). Tais estruturas protéicas 3D são mais conservadas que suas sequências primárias. Assim, se é detectada similaridade entre sequências, usualmente pode-se supor similaridade estrutural. Além disso, proteínas com baixa similaridade na sequência podem ter estruturas similares (Sánchez e Sali, 2000). Os passos para a modelagem consistem em identificar estruturas relacionadas à proteína que se quer modelar; selecionar os moldes; fazer o alinhamento entre os moldes e a proteína a ser modelada, usando programas tais como o CLUSTAL; e, finalmente, construir um modelo usando a informação das estruturas dos moldes e validá-lo (*Figura 3*).

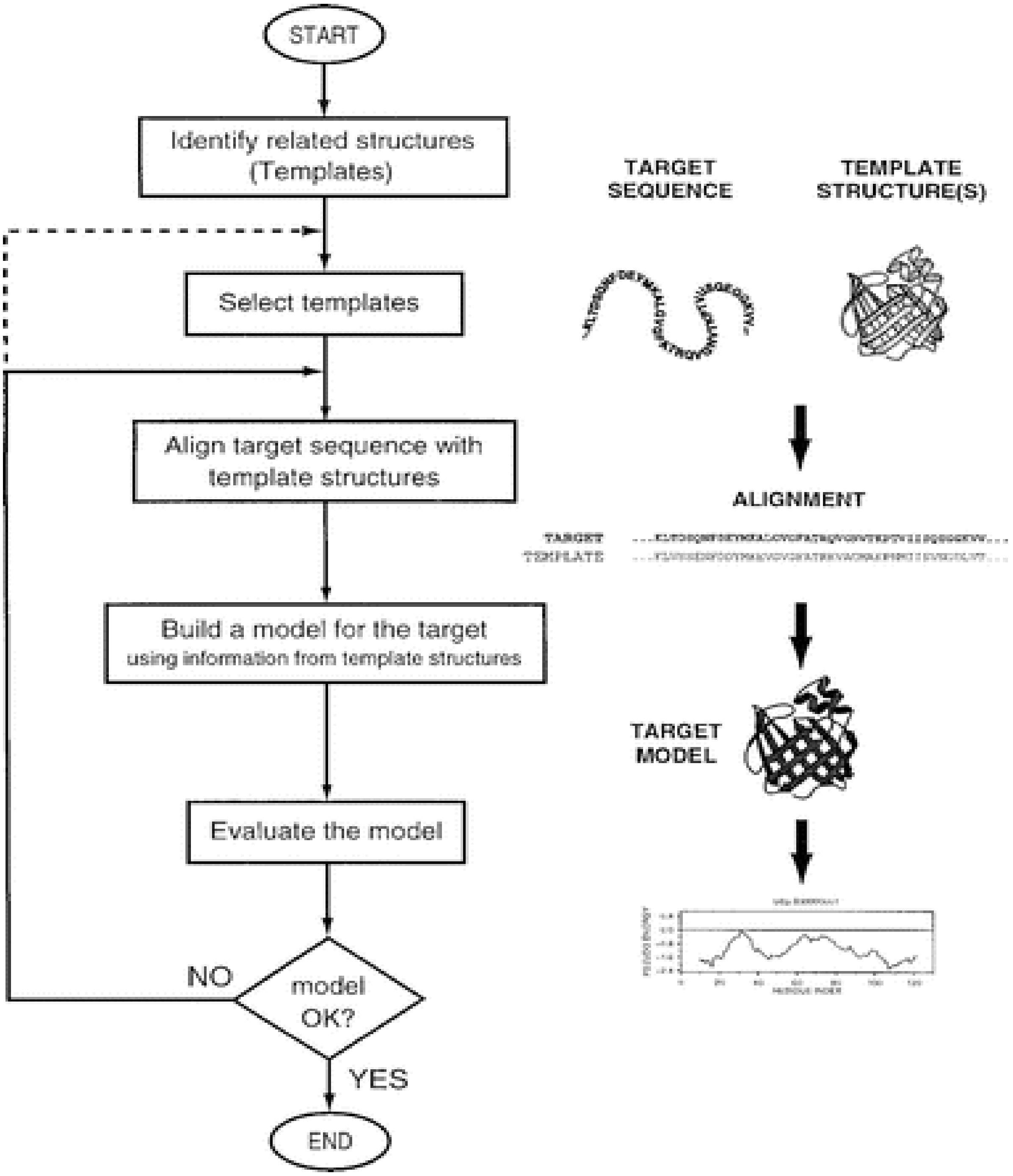

START
Identify related structures (Templates)
Select templates
Align target sequence with template structures
Build a model for the target using information from template structures
Evaluate the model
NO
model OK?
YES
END
TARGET SEQUENCE
TEMPLATE STRUCTURE(S)
ALIGNMENT
TARGET
TEMPLATE
TARGET MODEL

Figura 3 – Passos na Modelagem Comparativa por Homologia. Figura adaptada de Martí-Renom *et al.* (2000).

O modelo tridimensional é obtido através da otimização de uma função de densidade de probabilidade molecular (pdf), através do método da função alvo variável no espaço cartesiano que aplica métodos de gradientes conjugados e dinâmica molecular com *simulated annealing*. O programa implementa uma abordagem automatizada para modelar estrutura 3D satisfazendo as restrições espaciais obtidas empiricamente de um banco de dados de alinhamentos de estruturas de proteínas. As restrições angulares e de ligações são geradas através de uma análise estatística das relações entre muitos pares de estruturas homólogas, incluindo 416 proteínas de estrutura conhecida. Essas relações são expressas como funções de densidade de probabilidade condicionais e podem ser usadas diretamente como restrições espaciais. Posteriormente, as restrições espaciais e termos de energia que determinam a adequada estereoquímica são incluídas em uma função objetiva, que é otimizada no espaço cartesiano (Martí-Renom, 2000; Sánchez e Sali, 2000, 1997; Sali e Blundell, 1993).

A validação de um modelo produzido por modelagem molecular requer a verificação de características estereoquímicas, como comprimento de ligações, ângulos das ligações, planaridade das cadeias carbonadas, quiralidade e ângulos de torção das cadeias principal e laterais, tal como implementado no programa PROCHECK (Morris *et al.*, 1992; Laskowski *et al.*, 1993). Além disso, a avaliação do ambiente de cada resíduo de aminoácido no modelo em função do ambiente encontrado nas estruturas resolvidas que serviram como molde costuma ser feita. Um dos mais conhecidos programas utilizados para tal, trata-se do VERIFY-3D (Bowie *et al.*, 1991; Lüthy *et al.*, 1992).

Atualmente são conhecidas mais de 40.000 estruturas de proteínas, distribuídas em várias famílias contendo moléculas com o mesmo padrão básico de enovelamento.

Variações em famílias de proteínas homólogas que conservam a mesma função mostram como as estruturas acomodam mudanças na sequência de aminoácidos. Assim, resíduos localizados na superfície da proteína e não envolvidos em função estão livres para mutar, bem como alças que poderiam acomodar mudanças pelo rearranjo estrutural local. Mutações que alteram o volume de resíduos internos, em geral, não modificam a conformação estrutural de hélices e folhas beta; no entanto, causam distorções em seu arranjo espacial. À medida que uma sequência diverge, há um aumento progressivo na distorção da conformação da cadeia principal; portanto, existe uma correlação positiva entre divergência de sequências e alteração da estrutura tridimensional (Lesk, 2008).

Embora sequências muito similares possuam estruturas semelhantes, a relação entre estrutura e função é mais complexa, já que proteínas semelhantes podem desempenhar funções distintas. A contribuição da modelagem molecular para o entendimento da função de proteínas é imensa, inclusive com aplicações na indústria farmacêutica através da modelagem de ligantes e estudo da interação entre proteínas e proteínas-ligantes (Lesk, 2008; Coop e MacKerell, 2000). Tais análises quando associadas a abordagens evolutivas podem enriquecer enormemente nosso conhecimento sobre aspectos como divergência de funções e mudança de especificidade, podendo ser úteis para estudos de *docking*, desenho racional de fármacos, entre outras aplicações.

Família Gênica da Álcool Desidrogenase

A álcool desidrogenase (ADH) tem sido estudada em uma ampla variedade de grupos taxonômicos, incluindo plantas, animais e fungos. Os membros deste grupo pertencem à superfamília protéica das desidrogenases/redutases de cadeia média (MDR). Essa superfamília é formada por ADHs, redutases, leucotrieno B4 desidrogenases, entre diversas outras famílias gênicas. O grupo de MDRs cresceu consideravelmente nos últimos anos, sendo que no final de 2007 atingiu o número de 11.000 membros no banco de dados UniProt (Bairoch *et al.*, 2005). A maior parte das famílias do grupo das MDR possui centenas de membros, sendo que cerca de 1.000 famílias possuem 10 ou menos membros (Persson, *et al.*, 2008). A primeira MDR descrita foi uma álcool desidrogenase de mamífero, em 1970 (Jörnvall, 1970).

As proteínas MDR tipicamente possuem dois domínios: (1) domínio de ligação da coenzima (C-terminal), formado, frequentemente, por seis folhas betas paralelas, cada uma da qual seguida por uma hélice; e (2) domínio N-terminal de ligação ao substrato, formado por um núcleo de fitas β-antiparalelas e α-hélices posicionadas na superfície (Persson *et al.*, 2008).

Dois tipos de álcool desidrogenase (E.C. 1.1.1.1.) têm sido identificados: álcool desidrogenases clássicas, de cadeia longa (chamadas classe P em plantas), e álcool desidrogenases de cadeia curta (Charlesworth *et al.*, 1998), classificadas como SDRs (*short-chain dehydrogenases/reductases*). As proteínas clássicas são enzimas que requerem zinco como cofator, possuem aproximadamente 350 resíduos de aminoácidos, são todas diméricas e encontradas em um grande número de organismos, como mamíferos, plantas e leveduras. Já as de cadeia curta possuem cerca de 250 resíduos de aminoácidos, não necessitam de zinco como cofator e são encontradas em insetos, inclusive em *Drosophila* (Yokoyama *et al.*, 1990). Embora as sequências protéicas de álcool desidrogenase sejam altamente conservadas, sua função metabólica é bastante variável (Charlesworth *et al.*, 1998).

A ADH forma diferentes classes que compartilham cerca de 60% de identidade entre as sequências (Jörnvall, 2008). A classe mais conservada parece ser a ADH3 (classe III), que corresponde à formaldeído desidrogenase dependente de glutationa. Essa parece ser a forma ancestral da qual derivaram as outras classes de vertebrados, já que não possui representação em todos os invertebrados investigados.

Entre os vertebrados, os peixes apresentam uma forma mista de ADH, estruturalmente similar à classe III e funcionalmente semelhante à classe I (Dasmahapatra *et al.* 2005). Em mamíferos, tais enzimas são subdivididas em classes de I a VI. A classe I é uma enzima do fígado, contendo atividade de etanol desidrogenase; a classe III é idêntica à enzima formaldeído desidrogenase dependente de glutationa; a classe IV é uma forma preferencialmente expressa no estômago; enquanto que as classes II, V e VI, embora pouco estudadas, exibem propriedades distintas. Acredita-se que a origem das classes se deva a eventos de duplicação gênica no início da evolução dos vertebrados (duplicação I/III) ou durante a evolução dos mesmos (duplicação IV/I). A classe III corresponde a uma forma ancestral (Danielsson *et al.*, 1994), uma vez que apresenta alto grau de conservação na sequência de aminoácidos, e dela as outras classes de enzimas teriam derivado por duplicação e aquisição de novas especificidades aos substratos (Dolferus *et al.*, 1997). A classe II encontra-se presente nas linhagens de mamíferos e aves; enquanto as classes IV e VIII encontram-se presentes em anfíbios e a classe VII é encontrada em aves (Kedishvili *et* *al.*, 1997).

Em humanos, os sete genes que codificam ADH encontram-se em um agrupamento de cerca de 380 kb no braço longo do cromossomo 4 (4q21-23). A classe I é encontrada em uma região de 70 kb, formada por *Adh7* (*upstream*) e por *Adh6*, *Adh4* e *Adh5* (*downstream*). Membros da classe I estão envolvidos diretamente na oxidação do etanol em diferentes taxas, consequentemente, tornaram-se candidatos para o estudo do risco de desenvolvimento de alcoolismo (Osier *et al.*, 2002).

Há evidências de que a álcool desidrogenase desempenha funções distintas em mamíferos, tendo sido relacionada ao metabolismo da norepinefrina, dopamina, serotonina e ácido biliar. Adicionalmente, essa enzima pode catalisar a oxidação do retinol *in vitro* e *in vivo* (Gonzàlez-Duarte e Albalat, 2005). Assim, a expansão da família gênica da álcool desidrogenase em mamíferos exemplifica um processo de neofuncionalização com inúmeros eventos de duplicação levando a novas atividades.

Da mesma forma, foi demonstrado teoricamente que cópias em plantas provavelmente teriam sido retidas como uma consequência da substituição adaptativa de resíduos de aminoácidos, que confeririam uma mudança em função (Thompson *et al.*, 2007). Alterações sutis de restrição funcional nas sequências protéicas levam a diferenças nas taxas de substituições não-sinônimas entre as cópias gênicas, como encontrado por Gaut *et al.* (1999) no estudo da *Adh* de gramíneas. Dependendo da localização de tais substituições e da natureza da mudança, como por exemplo, a substituição de um resíduo carregado positivamente por um de carga negativa, pode-se avaliar o impacto que as mesmas podem ter na estrutura tridimensional da enzima.

A álcool desidrogenase apresenta função biológica variável dependendo da classe e do organismo a qual pertença. Nos animais a fermentação resulta em ácido láctico sob condições anaeróbicas, sendo que a ADH tem papel importante na detoxificação de alcoóis e aldeídos no fígado; portanto, trata-se de enzima glicolítica essencial no metabolismo anaeróbico. A atividade de formaldeído desidrogenase dependente de glutationa (GSH-FDH; formaldeído: NAD+ oxidoredutase, formil glutationa; E.C. 1.2.1.1.) tem sido demonstrada em plantas, com evidências de que estas enzimas de classe III poderiam ser o ancestral da classe P nas mesmas (enzimas ADH etanol-ativas) (Dolferus *et al.*, 1997). Nesse grupo a ADH juntamente com a piruvato decarboxilase (PDC, E.C. 4.1.1.1), faz parte da via de fermentação alcoólica, uma via de dois passos que converte piruvato via acetaldeído em etanol, diferentemente do que ocorre em animais. Esta via está presente em leveduras, gimnospermas, angiospermas e em algumas bactérias dá vantagem de crescimento a microrganismos (leveduras) na presença de altas concentrações de açúcar. Nas gimnospermas e angiospermas, a fermentação alcoólica é essencial para sua sobrevivência sob determinadas situações ambientais estressantes (Dolferus *et al.*, 1994, 1997), sendo fortemente induzidos em condições de baixa tensão de oxigênio. Adicionalmente, em A*rabidopsis* há indução por baixas temperaturas e estresse osmótico (desidratação). Em milho também ocorre indução pelo frio (Clegg *et al.*, 1997).

Em plantas, há uma pequena variação na quantidade de *loci Adh*. *Arabidopsis thaliana* tem um único *locus Adh* classe P, mas a maioria das plantas tem dois ou três *loci.* Em espécies diplóides de *Gossypium* há pelo menos sete *loci* (Small e Wendel, 2000b; Charlesworth *et al.*, 1998, Suiter, 1998). A posição dos íntrons nos *loci Adh* é, geralmente, conservada, tanto em angiospermas quanto em gimnospermas (Charlesworth *et al.*, 1998).

Adh é a família gênica nuclear de plantas melhor estudada, tanto em termos de biologia quanto de evolução molecular. Estudos evolutivos têm sido feitos em um número considerável de plantas, como gramíneas, especialmente o milho, *Arabidopsis* e algodão (Small e Wendel, 2000a). Apesar disso, não eram encontrados na literatura estudos sobre a estrutura da proteína álcool desidrogenase em plantas nem estrutura 3D disponibilizada nos bancos de dados. Esse último fato contrasta com o grande número de tais estruturas investigadas em outros organismos, como mamíferos, anfíbios, peixes, ascomicetos, bactérias e leveduras disponíveis no *Protein Data Bank* (Berman *et al.*, 2000). Em estudo publicado em 2007 (Thompson *et al.*), a primeira estrutura de ADH em plantas foi disponibilizada.

Estudos sobre a diversidade da sequência nucleotídica do *locus Adh1* em plantas têm mostrado um excesso substancial de sítios não-sinônimos em determinada região do *locus*, indicando que a distribuição dos polimorfismos é consistente com um balanço entre seleção fraca e mutação. Taxas de substituições sinônimas e não-sinônimas dentro de um gene são frequentemente relacionadas, mas não parece ser este o caso do gene *Adh*. Esse desacoplamento é mais óbvio entre os *loci Adh1* e *Adh2* de gramíneas. O *locus Adh2* tem uma taxa maior de substituições não-sinônimas que *Adh1*, sem que haja um aumento nas taxas de substituições sinônimas. A aceleração das taxas de substituições não-sinônimas em *Adh2* é consistente com estudos prévios que documentaram um aumento na taxa de substituição não-sinônima subsequente à duplicação gênica. Tanto a seleção positiva quanto o relaxamento da seleção purificadora podem causar um aumento nessa taxa após a duplicação gênica (Brandon *et al.*, 1996).

Em *Zea mays*, os dois genes *Adh* diferem em seus padrões de expressão tecido-específicos. O gene *Adh1* é expresso em tecidos da semente e pólen, enquanto *Adh1* e

Adh2 são expressos em tecidos de raiz sob condições de anóxia. Variação similar nos padrões de expressão é encontrada em outras gramíneas, girassóis, eucaliptos e pinheiros (Gaut *et al.*, 1999). Além disso, ensaios enzimáticos revelaram importantes diferenças entre os três alelos do *locus Adh1* em milho, entre elas, o fato de que produtos protéicos codificados por esses alelos diferem nas suas atividades específicas (Gaut e Clegg, 1993).

Estudos indicam que duplicações gênicas aumentam a diversidade de expressão gênica tanto dentro do genoma como entre genomas, tornando as cópias mais especializadas em diferentes tecidos ou estágios de desenvolvimento (Gu *et al.*, 2004). Há evidências de que a ADH também mostra padrões de expressão tecido-específicos em animais, tal como ocorre em ADH1 de *Oryzias latipes* (Hoffmann *et al.*, 2006; Dasmahapatra *et al.*, 2005), e ADH1, 2 e 4 em camundongos (Szalai *et al.*, 2002).

CAPÍTULO 2

OBJETIVOS

Geral

O objetivo geral desse trabalho é o estudo da evolução molecular, diversificação funcional e aspectos estruturais da família gênica da Álcool Desidrogenase.

Específicos

1. Analisar a evolução molecular da enzima ADH em plantas, animais e fungos.

2. Compreender a relação da formaldeído desidrogenase dependente de glutationa, ADH de classe III, com outras classes de ADH.

3. Estudar a diversificação funcional da família gênica da álcool desidrogenase.

4. Determinar os resíduos de aminoácidos importantes nas diferenças estrutural-funcionais dessas enzimas.

5. Deduzir a estrutura tridimensional de proteínas de interesse em plantas já que não há nenhuma estrutura resolvida desses organismos nos bancos de dados.

CAPÍTULO 3

Artigo 1

Publicado em Gene 396: 108-115, 2007 (ISSN: 0378-1119)

Available online at www.sciencedirect.com
ScienceDirect

Gene 396 (2007) 108–115

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

Sequence and structural aspects of the functional diversification of plant alcohol dehydrogenases

Claudia E. Thompson [a], Francisco M. Salzano [a], Osmar Norberto de Souza [b], Loreta B. Freitas [a,*]

[a] *Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970, Porto Alegre, RS, Brazil*
[b] *Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga 6681, 90610-001, Porto Alegre, RS, Brazil*

Received 16 October 2006; received in revised form 15 February 2007; accepted 27 February 2007
Available online 13 March 2007
Received by J.G. Zhang

Abstract

The glycolytic proteins in plants are coded by small multigene families, which provide an interesting contrast to the high copy number of gene families studied to date. The alcohol dehydrogenase (*Adh*) genes encode glycolytic enzymes that have been characterized in some plant families. Although the amino acid sequences of zinc-containing long-chain ADHs are highly conserved, the metabolic function of this enzyme is variable. They also have different patterns of expression and are submitted to differences in nonsynonymous substitution rates between gene copies. It is possible that the *Adh* copies have been retained as a consequence of adaptative amino acid replacements which have conferred subtle changes in function. Phylogenetic analysis indicates that there have been a number of separate duplication events within angiosperms, and that genes labeled *Adh1*, *Adh2* and *Adh3* in different groups may not be homologous. Nonsynonymous/synonymous ratios yielded no signs of positive selection. However, the coefficients of functional divergence (θ) estimated between the *Adh1* and *Adh2* gene groups indicate statistically significant site-specific shift of evolutionary rates between them, as well as between those of different botanical families, suggesting that altered functional constraints may have taken place at some amino acid residues after their diversification. The theoretical three-dimensional structure of the alcohol dehydrogenase from *Arabis blepharophylla* was constructed and verified to be stereochemically valid.

Keywords: Glycolytic proteins; Functional diversification; *Arabis blepharophylla*; Molecular modeling

1. Introduction

Many plant nuclear genes are members of multigene families. The formation of these families must have been fundamental in plant evolution. The recurrence of duplications and functional divergences has generated the present gene families. One copy may be silenced by deleterious mutations (pseudofunctionalization) or both copies may be preserved if such substitutions result in novel capacities. This preservation can lead to the persistence of both copies with sequence similarity, to subfunctionalization (where each copy adopts some of the tasks of the ancestor) or to neofunctionalization (when one gene maintains the original function, while the other acquires a new role) (Gonzàlez-Duarte and Albalat, 2005). Phylogenetic analyses are useful for understanding the relationships of member genes of a gene family, since gene trees may clarify the history of gene duplication events.

We studied patterns of molecular diversification among members of the *Adh* gene family in plants, which belong to the medium-chain dehydrogenase/reductase protein superfamily. Alcohol dehydrogenases (ADHs) are dimeric enzymes of the glycolytic pathway that constitute two evolutionary groups, one characterized by short protein chains (~250 residues) which do not require zinc as a cofactor, and another characterized by long protein chains (~370 residues) which require zinc as a cofactor. The former group is represented by *Drosophila* ADHs, and the

Abbreviations: ADH, alcohol dehydrogenase; LRSH, Shimodaira–Hasegawa test with RELL approximation; LRT, likelihood ratio test; MEGA, Molecular Evolutionary Genetics Analysis; ML, maximum likelihood; NAD$^+$, nicotinamide adenine dinucleotide; NCBI, National Center of Biotechnology Information; NJ, neighbor-joining; RELL, resample estimated log-likelihood; RMSD, root mean square deviation.

* Corresponding author. Tel.: +55 51 33086731; fax: +55 51 33089823. *E-mail address:* loreta.freitas@ufrgs.br (L.B. Freitas).

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.02.016

latter by ADHs from organisms as diverse as mammals, plants, and yeasts. Although the amino acid sequences of zinc-containing long-chain (LC) ADHs are highly conserved (Yokoyama et al., 1990; Charlesworth et al., 1998), the metabolic function of this enzyme is variable (Dolferus et al., 1997).

In vertebrates eight distinct classes have been defined based on sequence similarity, catalytic features and gene expression patterns (ADH1–8, classes I–VII, according to Duester et al., 1999, and class VIII following Peralba et al., 1999). In mammalian tissues, at least six classes of this enzyme occur. Class I is the well-known liver enzyme with ethanol dehydrogenase activity, class III is identical to the glutathione-dependent formaldehyde dehydrogenase, class IV is a form preferentially expressed in the stomach, while classes II, V and VI are known to exhibit diverse properties (Danielsson et al., 1994).

Transcription from *Adh* promoters increases under oxygen stress, as well as in response to stress due to low temperatures in maize. The *Arabidopsis Adh* gene is induced predominantly in roots by environmental stresses such as low oxygen levels, dehydration, low temperature and the phytohormone ABA (Dolferus et al., 1994). Two or three isozymes are observed in all flowering dicot or monocot plant species, with the exception of *Arabidopsis*, which has a single *Adh* locus. This gene family has been most intensively studied in the Poaceae (Morton et al., 1996). In *Zea mays*, for example, the two *Adh* genes differ in their pattern of tissue-specific expression. *Adh1* is expressed in dry seed and pollen tissues, while both *Adh1* and *Adh2* are expressed in roots under anoxic conditions. Similar variation in *Adh* expression is found in a wide variety of plant species, including other grasses, sunflower, eucalyptus, and pine (Gaut et al., 1999). Functional assays also reveal important differences among the three *Adh1* alleles from the locus in maize. The protein products encoded by these alleles differ in their specific activity, and the alleles vary in their ability to recombine intragenically. Allelic differences in protein function, different patterns of expression, gene conversion, and recombination make the *Adh* locus evolutionarily interesting (Gaut and Clegg, 1993).

Structurally the ADH zinc-containing and the nicotinamide adenine dinucleotide (NAD^+) dependent enzyme contains two domains, one (residues 177–322) links to the coenzyme, and the other (residues 1–176; 323–373) is the catalytic unit. The active site is located in the cavity between the two domains. Two conformations between the domains are distinguished, “open” in the apoenzyme and “closed” in the complex with the substrate. Three segments (V1, V2 and V3) can also be distinguished, that are responsible for the enzyme’s hypervariability. They correspond to a portion (V1, residues 49–61) adjacent to the active site; a loop near the zinc atom (V2, residues 100–130); and a region of monomer interaction (V3, residues 290–310) (Person et al., 1993; Danielsson et al., 1994).

The *Adh* copies may have been retained as a consequence of adaptative amino acid replacements which have conferred subtle changes in function. Slightly different constraints on the protein sequence may lead to subsequent differences in nonsynonymous substitution rates between gene copies, as found by Gaut et al. (1999).

Despite the large number of studies involving the *Adh* gene family, there does not exist a wide-ranging study correlating its molecular evolution and structural biology in plants. Here, we extend previous studies of this multigene family, with the goal of using molecular evolutionary and modeling tools to understand its process of diversification. This is the first study where a plant ADH three-dimensional structure is proposed.

2. Materials and methods

2.1. Sequence analysis

The gymnosperm and angiosperm amino acid and DNA sequences were obtained from the National Center of Biotechnology Information (NCBI) and are listed in the Supplementary material (Table 1A), together with information about the species from which they were obtained. Alignments were performed with the ClustalW program (Jeanmougin et al., 1998). They were inspected and manual changes were made when necessary using GeneDoc 2.6 (Multiple Sequence Alignment Editor & Shading Utility; Nicholas and Nicholas, 1997). Alignments are available upon request. Phylogenies were estimated by neighbor-joining (NJ) (Saitou and Nei, 1987), available in the MEGA (Molecular Evolutionary Genetics Analysis) program, version 3.1 (Kumar et al., 2000), and by maximum likelihood (ML) methods using PhyML (Guindon and Gascuel, 2003) and TreeFinder (Jobb et al., 2004). ADH sequences from *Pinus banksiana* were used as an outgroup. In the NJ method, the p distance and the Poisson-corrected amino acid distances were used to analyze the amino acid sequences. The Kimura two-parameter method was used as a substitution model in the DNA sequences. A total of 1500 repetitions were performed using the bootstrap method (Felsenstein, 1985) to determine the reliability of each node of the tree. The NJ method is known to be preferable to other commonly used methods when rates of evolution differ among the branches of a phylogenetic tree (Nei, 1991), as may often be true in the case of multigene families whose members have adaptations to different functions. Maximum likelihood topologies were generated using the HKY substitution model (Hasegawa et al., 1985) to build the DNA gene trees. The bootstrap analyses were conducted using the previously cited programs; however, in the TreeFinder analysis, an approximate bootstrap support was computed by applying the Shimodaira–Hasegawa test with RELL approximation (LRSH) to all local rearrangements of the tree topology around an edge. In this case, the reliability of the trees was tested using 1000 replications. The resultant tree topologies were used to calculate branch lengths using the M0 model available in the CODEML program of the PAML packet (Yang, 1997). Afterwards, to evaluate the presence of positive selection, 13 protein sequences were examined using the maximum likelihood models recommended by Yang (2004). The one-ratio model (M0) assumes one ω (d_N, nonsynonymous/d_S, synonymous) ratio for all sites. The nearly neutral model (M1a) presupposes a proportion p_0 of conserved sites with $\omega_0 < 1$ and $p_1 = 1 - p_0$ of neutral sites with $\omega = 1$. The positive selection model (M2a) adds an additional class of sites with frequency

$p_2 = 1 - p_0 - p_1$ and ω_2 is estimated from the data. In the discrete model (M3), the probabilities (p_0, p_1 and p_2) of each site being submitted to purifying, neutral and positive selection, respectively, and their corresponding ω ratios (ω_0, ω_1 and ω_2) are inferred from the data. The Beta model (M7) is a null test for positive selection, assuming a Beta distribution with ω between 0 and 1. Finally, the Beta&ω model (M8) adds one extra class with the same ratio ω_1 (Yang, 2004).

The likelihood ratio test (LRT) was used to verify whether ω was significantly different from 1 for each pairwise comparison: M1a *vs.* M2a, M0 *vs.* M3, and M7 *vs.* M8. LRT performs the comparison both with the constraint of $\omega = 1$ and without such constraint: $LR = 2(\ln_1 - \ln_2)$. These LRT statistics approximately follow a chi-square distribution and the number of degrees of freedom is equal to the number of additional parameters in the more complex model (Anisimova et al., 2001, 2002).

A statistical framework modeling the functional divergence was implemented by the DIVERGE program (Gu and Velden, 2002) using 61 sequences, to estimate the coefficient of functional divergence (θ). This coefficient is an indicator of the level of type I functional divergence caused by an evolutionary process resulting in either altered functional constraints or in a site-specific evolutionary rate shift between two duplicate genes (Gu, 1999; Gu and Velden, 2002). Rejection of the null hypothesis H0: $\theta = 0$ provides statistical evidence for shifts in evolutionary rates or in altered functional constraints. Let $Q(k) = P_k(S_1|X)$ be the *a posteriori* probability of a site k being S_1 (functional divergence related status) when the amino acid configuration (X) is observed. Since the alternative status S_0 (functional divergence unrelated status), with posterior probability $P_k(S_0|X) = 1 - P_k(S_1|X)$, means no altered functional constraint, the predicted residues are only meaningful when $Q(k) > 0.5$, in which case the ratio $R(S_1|S_0) = P(S_1|X)/P(S_0|X) > 1$. A more stringent cut-off may yield $Q(k) > 0.67$, or $R(S_1|S_0) \geq 2$ (Gu and Gu, 2003).

2.2. Modeling and structure analysis

We obtained results for the three-dimensional structure of the *Arabis blepharophylla* ADH (adhARABLE) using *Equus caballus* liver alcohol dehydrogenase (PDB code 1N8K), as a template, encountered using Blastp (Altschul et al., 1990; Altschul et al., 1997). Its structure has been solved to a 1.13 Å resolution (Rubach and Plapp, 2003). The amino acid sequence alignments were performed with the ClustalW program (Thompson et al., 1994), using the BLOSUM matrix (Henikoff and Henikoff, 1992) for scoring, and manually adjusting to obtain optimal alignment. Afterwards, the program MODELLER6v2 (Sali and Blundell, 1993) was used to build protein models according to the comparative protein modeling methodology. Finally, the best model was evaluated and selected on the basis of the results obtained by PROCHECK (Laskowski et al., 1993) and VERIFY-3D (Lüthy et al., 1992). The Swiss-PdbViewer (Guex and Peitsch, 1997) was used to calculate the root mean square deviation (RMSD) between the template and the model, as well as to draw all the figures and generate the molecular surface.

3. Results and discussion

3.1. Phylogenetic analysis

Comparative analysis presents an ideal opportunity to investigate the dynamics of an angiosperm gene family, and in particular, to expand our understanding of *Adh* evolution. In total, 1155 DNA sites from 176 sequences (94 monocot, 75 dicot, and seven gymnosperm) were used in the phylogenetic analysis. The NJ (Fig. 1) and ML (not shown) tree topologies did not differ significantly, especially when the major clades are considered. There was consensus in some important respects. First of all, recent gene duplications are evident in the Paeoniaceae and Cyperaceae, indicating that the *Adh1* and *Adh2* sequences of both families are more related to each other than to other dicot and monocot sequences. These duplications should have occurred after the diversification of these two botanical families. Other gene duplication responsible for the *Adh1a* and *Adh1b* split is observed at the base of the Paeoniaceae clade (Fig. 1).

The Arecaceae (palm) family always formed a basal clade within monocots in our data, as was also observed by Borsch et al. (2003) using the plastid *TrnT–TrnF* marker. It emerged ≈80 million years ago and radiated early when compared to the Poaceae and Cyperaceae families (Wilson et al., 1990; Duvall et al., 1993; Tamura et al., 2004). The Poaceae (grass) family emerged and diversified ≈60 million years ago and has been the most intensively studied monocot lineage due to its economic importance (Morton et al., 1996). Our results agree with those of Mathews and Sharrock (1996) using the phytochrome gene. Both

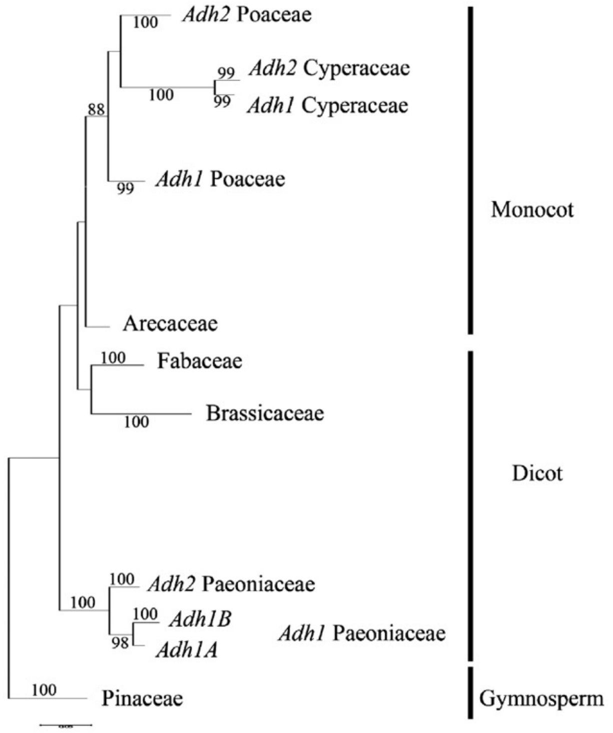


Fig. 1. Simplified phylogenetic tree inferred from DNA *Adh* sequences using the neighbor-joining method. Numbers represent bootstrap values; values higher than 84% are shown. Scale bar indicates levels of sequence divergence. The detailed tree with sequence identification can be obtained on request.

the DNA and protein trees produce two separated clades which correspond to the *Adh1* and *Adh2* sequences. A recent *Adh2* duplication has also originated a third (*Adh3*) locus in the genus *Hordeum* (Supplementary material). The Arecaceae *AdhA* and *AdhB* loci (Supplementary material) do not seem to correspond to the Poaceae loci, since they clearly separate in the trees. The Fabaceae adh1GLYMAX2, adh1PHAFOL, adh1SOPFLA, adh1LOTCOR, adh1PEA, adh1TRIREP sequences form a monophyletic group with high bootstrap support.

The Brassicaceae sequences cluster in four major groups of taxa (Fig. 2): (1) one (adhBRAOLE) with *Brassica oleraceae* as the most basal group; (2) a clade containing sequences from *Cardamine amara*, *Barbarea vulgaris* and taxa of the *Leavenworthia* genus (prefixes BAR, CAR and LEA); (3) one with those from *Arabidopsis thaliana* (located in the lower portion of the figure); and (4) one with *Arabis* and *Aubrieta* sequences. According to Koch et al. (2000) the *Arabidopsis*, *Brassica* and *Arabis stelleri* clades diverged early in the history of the Brassicaceae, at roughly 24 Mya. Yang et al. (1999) obtained very similar dates (14,000–20,000 Mya) for the *Brassica–Arabidopsis* divergence. Pollen from close relatives of *Cardamine* and *Barbarea* is common in Pliocene samples (2.5–5.0 Mya; Mai, 1995). Koch et al. (2000) estimated the *Cardamine* and *Barbarea* divergence at about 6.0 Mya. Since relatives of *Leavenworthia*, *C. cardamine* and *B. vulgaris* only have a single *Adh* copy, the extra copies in *Leavenworthia* (*Adhs* 1, 2 and 3) probably arose after the origin of this genus. In this taxon *Adh3* appears to be more closely related to *Adh2* than to *Adh1*. Since *Adh3* has no introns, it may be a product of a reverse transcription event involving a mRNA intermediate. Charlesworth et al. (1998) have shown that this locus is not closely linked to the other *Adh* loci. The species relationships within this family agree with those obtained by Johnston et al. (2005), using ITS sequences.

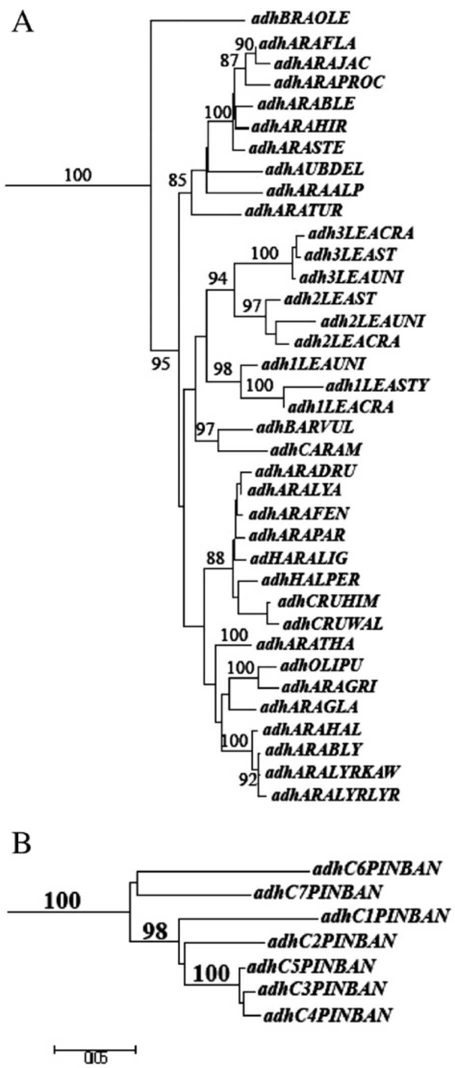


Fig. 2. Detailed representation of the relationships obtained with the DNA sequences of the Brassicaceae and Pinaceae families.

Table 1
Parameter estimates and log-likelihood values under models of variable ω ratios among sites

Model	Parameters	l	Sites indicating positive selection
M0	$\omega=0.07383$	−6025.028641	No
M1	$\omega_0=0.05640$, $\omega_1=1$, $p_0=0.94142$, $p_1=0.05858$	−5981.774704	Not allowed
M2	$\omega_0=0.05640$, $\omega_1=1$, $\omega_2=32.42120$, $p_0=0.94142$, $p_1=0.05858$, $p_2=0$	−5981.774710	No
M3	$\omega_0=0.00532$, $\omega_1=0.09775$, $\omega_2=0.38780$, $p_0=0.51672$, $p_1=0.37858$, $p_2=0.10470$	−5937.707770	No
M7	$p=0.33284$, $q=3.51763$	−5938.488181	Not allowed
M8	$p_0=1$, $p=0.33284$, $q=3.51763$, $p_1=0$, $\omega=2.88006$	−5938.488499	No

All these observations confirm the usefulness of *Adh* as a molecular phylogenetic marker and the evidence that there have been a number of separate duplication events within angiosperms (e.g., in the peony lineage, in the grasses, etc); therefore, genes labeled as *Adh1*, *Adh2* and *Adh3* in different groups may not be homologous.

The phylogenetic tree of Fig. 1 clearly shows three primary lineages corresponding to monocot, dicot, and gymnosperm

Table 2
Coefficients of functional divergence (θ) of pairwise comparisons in the alcohol dehydrogenase gene family

Comparison	Group 1	Group 2	$\theta\pm$SE[a]	LRT[b]
Between forms	Poaceae *Adh2*	Poaceae *Adh1*	0.729±0.200	13,303
	Poaceae *Adh2*	Fabaceae *Adh1*	0.552±0.118	21,964
	Poaceae *Adh1*	Fabaceae *Adh1*	0.541±0.137	15,467
Between taxonomic units	Poaceae	Pinaceae	0.549±0.122	20,330
	Poaceae	Fabaceae	0.442±0.094	22,198
	Poaceae	Brassicaceae	0.449±0.120	13,929
	Fabaceae	Pinaceae	0.436±0.092	22,136
	Brassicaceae	Pinaceae	0.505±0.113	19,914
	Fabaceae	Brassicaceae	0.666±0.097	47,055

All values are statistically significant at $P<0.001$ or less. Sequences of the Cyperaceae, Arecaceae, and Paeoniaceae families had incomplete information for this type of analysis.

[a] SE stands for standard error.

[b] LRT: Likelihood Ratio Test.

genes, the eudicot not forming a monophyletic block. As indicated above, gene duplications have taken place independently in each of these lineages. All members of the same linkage group, *P. banksiana AdhC1*, *AdhC2*, *AdhC3*, *AdhC4*, and *AdhC5* form a strongly supported lineage, while *AdhC6* and *AdhC7*, of other linkage group according to Perry and Furnier (1996), fall outside of it (Fig. 2). Duplication seems a recurrent theme within the *P. banksiana Adh* gene family. Not only are there more *Adh* genes than have been found in most angiosperms, but also repeated sequences within the genes are much more common. This pattern could represent one component contributing to the unusually large conifer genome sizes (Perry and Furnier, 1996). The *Adh* tissue specificity and regulated stress response may result in a unique pattern of evolution for this gene family.

3.2. Selection and functional diversification

Nonsynonymous rate differences appear to be common among members of plant multigene families. Gaut et al. (1999) found that *Adh2* sequences evolve significantly more rapidly than *Adh1* sequences at nonsynonymous sites, but evolve with similar synonymous rates. *Adh1* and *Adh2* differences in nonsynonymous rates could be fueled either by adaptative substitution events or by reduced selective constraints against duplicate gene copies.

Positive selection can be detected by the ratio of nonsynonymous to synonymous substitution rates ($\omega = d_N/d_S$). Table 1 lists the parameter estimates and log-likelihood values under models of variable ω between sites and those obtained with the M0 model (one-ratio), which assumes the same ratio for all sites.

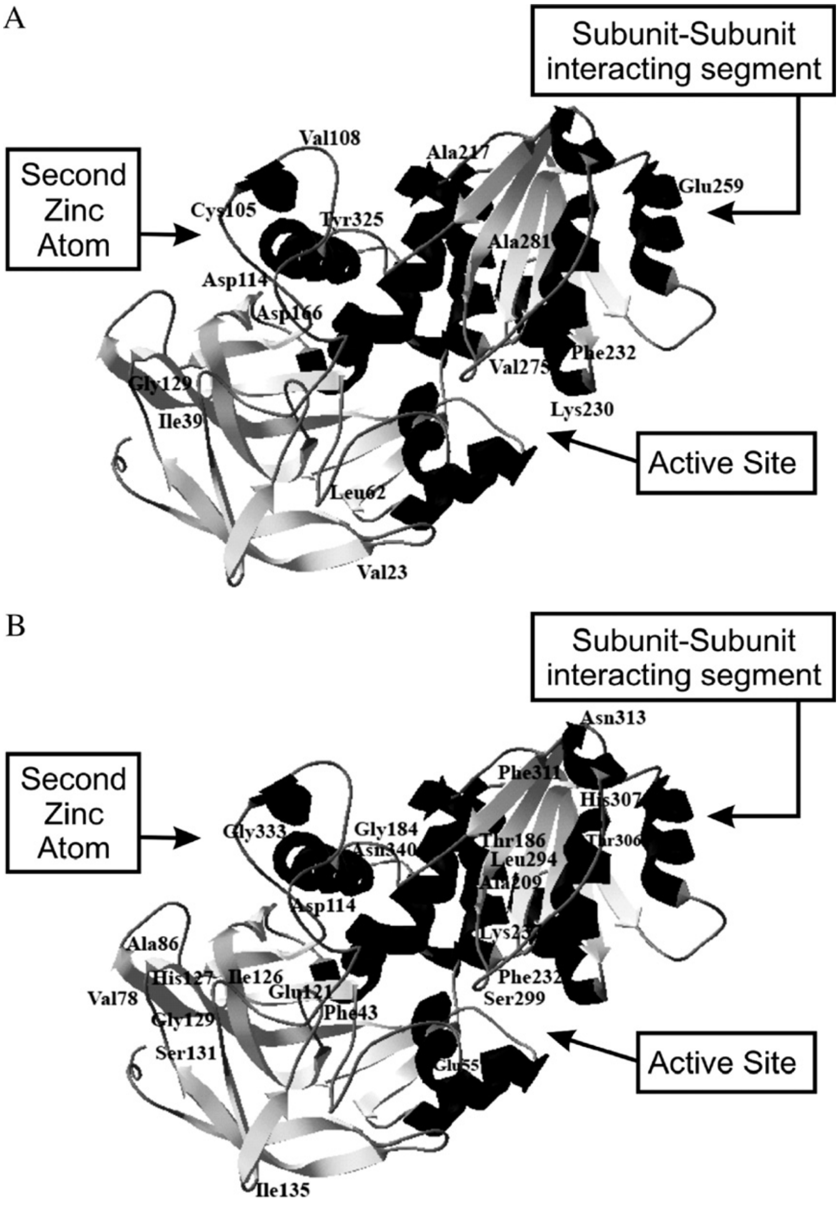


Fig. 3. Localization of the amino acid residues presumably submitted to altered functional constraints: A. after *Adh1*/*Adh2* gene duplication, cut-off value: $Q(k) \geq 0.85$, plus three (Cys105, Val108, Asp114) with $Q(k) \geq 0.8$ located in a strategical position of the molecule; B. after diversification among botanical families, cut-off value: $Q(k) \geq 0.85$.

Table 3
Amino acid residues important for the functional divergence between *Adh1* and *Adh2*

Amino acids residues	Secondary structure of *A. blepharophylla*
25	**s (Val23)**
41	**s (Ile39)**
45	s (Phe43)
62	l (Thr60)
64	**l (Leu62)**
79	s (Val75)
109	h (Cys105)
112	l (Val108)
118	l (Asp114)
133	**l (Gly129)**
170	**l (Asp166)**
178	h (Ile174)
183	h (Leu179)
185	h (Thr181)
190	h (Thr186)
200	s (Gln196)
204	s (Ile200)
221	**h (Ala217)**
229	s (Val225)
233	l (Ser229)
234	**h (Lys230)***
236	**h (Phe232)**
238	h (Glu234)
240	h (Lys236)
259	h (Glu255)
263	**h (Glu259)**
279	**h (Val275)**
285	**h (Ala281)***
329	**l (Tyr325)***
337	h (Gly333)

In bold amino acid residues with $Q(k) \geq 0.85$, and (*) indicates those found using $Q(k) \geq 0.9$ as a cut-off value; s: strand; l: loop; h: helix.

None of the estimates are significantly higher than 1.0; consequently, there is no indication that adaptative selection contributed to the *Adh1*/*Adh2* differences in nonsynonymous rates. However, the test detects only a limited set of adaptative selection events. To further investigate whether any amino acid replacement could have led to adaptative functional diversification, we examined the amino acid replacements that distinguish the *Adh* sequences by posterior analysis using the DIVERGE program. The coefficients of functional divergence θ estimated between the *Adh1* and *Adh2* gene groups reported in Table 2 indicate statistically significant site-specific shift of evolutionary rates between them, with θ varying markedly from 0.541 to 0.729. Moreover, the θ of the *Adh* genes considering different botanical families is significantly greater than zero (0.436–0.666), suggesting that altered functional constraints may have taken place at some amino acid residues after their diversification.

The amino acids responsible for the functional divergence after gene duplication or after speciation can be predicted based on a site-specific profile by choosing a suitable cut-off value. We used three cut-off values: $Q(k) \geq 0.8$, $Q(k) \geq 0.85$ and $Q(k) \geq 0.9$. These residues were mapped onto the three-dimensional structure of the *A. blepharophylla* ADH which we have modeled.

Of course, there is no present direct evidence of different functions between plant ADH1 and ADH2 proteins or between these proteins in diverse botanical families. But the diversification process which occurred in these substances along evolution in general indicates their variability. For instance, Höög et al. (2001) mentioned that in addition to ethanol oxidation, they may be involved, in different organisms, with norepinephrine, dopamine, serotonin, and bile acid metabolism.

3.3. Molecular modeling

We have modeled the alcohol dehydrogenase three-dimensional structure from *A. blepharophylla* (adhARABLE). The degree of identity between the sequence of the selected template and that of *A. blepharophylla* was around 48%, with 23% showing strong similarity.

Table 4
Amino acid residues important for the functional divergence among different botanical families

Amino acids residues	Secondary structure of *A. blepharophylla*
45	**s (Phe43)**
49	l (Cys47)
57	**h (Glu55)**
64	l (Leu62)
82	**s (Val78)**
90	**l (Ala86)**
112	l (Val108)
118	**l (Asp114)***
125	**l (Glu121)**
127	l (Gly123)
128	l (Gly124)
130	**l (Ile126)**
131	**l (His127)***
133	**l (Gly129)***
135	**l (Ser131)**
139	**s (Ile135)**
161	s (Ser157)
178	h (Ile174)
187	h (Leu183)
188	**h (Gly184)**
190	**l (Thr186)**
194	l (Ala190)
209	h (Ala205)
213	**h (Ala209)**
219	h (Arg215)
221	h (Ala217)
224	l (Ser220)
236	**h (Phe232)**
237	h (Asp233)
241	**h (Lys237)**
271	s (Arg267)
279	h (Val275)
295	**s (Leu294)**
303	**l (Ser299)***
310	**l (Thr306)***
311	**l (His307)***
313	h (Met309)
315	**l (Phe311)***
317	**l (Asn313)***
337	**h (Gly333)**
338	h (Val334)
344	**h (Asn340)**

In bold amino acid residues with $Q(k) \geq 0.85$, and (*) indicates those found using $Q(k) \geq 0.9$ as a cut-off values; s: strand; l: loop; h: helix.

Ten models were initially created, and they were analyzed using the PROCHECK and VERIFY-3D programs, together with the values of the root mean square deviations (RMSD). The Ramanchandran plots confirmed the excellent quality of the initial models, with the percentage of residues in most favoured and additional allowed regions being no lower than 99.7%. The RMSD between the backbone atoms of the template and the model was 0.2 Å, where the differences in the loop regions account for the largest differences. The stereochemical parameters and the VERIFY-3D results are presented as Supplementary material (Table 2A, Fig. 1A). Taken together, these data suggest that the models were stereochemically valid, and, therefore, suitable for further sequence–structural analysis.

The residues which are important for the functional divergence were mapped onto the alcohol dehydrogenase 3D structure obtained (Fig. 3A and B). The number of amino acid residues presumably submitted to altered functional constraints is smaller when the ADH1 and ADH2 groups are compared than in the comparison between different botanical families. The probably selected sites are located both inside and on the surface of the 3D structure. Thirty amino acid residues can be identified as being important for the functional divergence in the case of the ADH1/ADH2 comparison using $Q(k)\geq 0.8$ as a cut-off value, twelve using $Q(k)\geq 0.85$ and three using $Q(k)\geq 0.9$ (Table 3); while 42 amino acids residues are found comparing different botanical families using $Q(k)\geq 0.8$, 24 using $Q(k)\geq 0.85$ and eight using $Q(k)\geq 0.9$ as cut-off values (Table 4). Three residues identified as being significantly divergent between ADH1 and ADH2 (Cys105, Val108 and Asp114) are located in functionally important regions, near a part of the loop around the second zinc atom. Lys230, Phe232, and Val275 are near the active site (Fig. 3A). As for the divergent residues among different botanical families, residues Cys47 and Glu55 are found in a segment adjacent to the active site, residues Val108, Asp114, Glu121, Gly123, Gly124 and Ile126 are in the loop around the second zinc atom, and residues Leu294, Ser299 and Thr306 are in a subunit–subunit interacting segment.

Considering these findings as a whole, we obtained evidence for variation in three key areas of the enzyme: (a) the loop around the second zinc atom, an important cofactor for the enzyme's function; (b) the subunit–subunit interacting segment, responsible for the dimmer formation; and (c) the active site, that interacts with the substrate. All these results point to natural selection as an important factor in the evolution of this protein.

Acknowledgements

The authors would like to thank L.E. Dardenne and C.L. Fernandes, and two anonymous reviewers for helpful comments. This work was supported by the Institutos do Milênio and PRONEX Programs, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul), PROPESQ-UFRGS (Pró-Reitoria de Pesquisa da Universidade Federal do Rio Grande do Sul) and PRPPG-PUCRS (Pró-Reitoria de Pesquisa e Pós-Graduação da Pontíficia Universidade Católica do Rio Grande do Sul).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.02.016.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anisimova, M., Bielawski, J.P., Yang, Z., 2001. Accuracy and power of the likelihood ratio testing in detecting adaptative molecular evolution. Mol. Biol. Evol. 18, 1585–1592.

Anisimova, M., Bielawski, J.P., Yang, Z., 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19, 950–958.

Borsch, T., Hilu, K.W., Quandt, D., Wilde, V., Neinhuis, C., Barthlott, W., 2003. Noncoding plastid *trnT–trnF* sequences reveal a well resolved phylogeny of basal angiosperms. J. Evol. Biol. 16, 558–576.

Charlesworth, D., Liu, F.L., Zhang, L., 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol. Biol. Evol. 15, 552–559.

Danielsson, O., Atrian, S., Luque, T., Hjelmqvist, L., Gonzalez-Duarte, R., Jörnvall, H., 1994. Fundamental molecular differences between alcohol dehydrogenase classes. Proc. Natl. Acad. Sci. U. S. A. 91, 4980–4984.

Dolferus, R., Jacobs, M., Peacock, W.J., Dennis, E.S., 1994. Differential interactions of promoter elements in stress responses of the *Arabidopsis Adh* gene. Plant Physiol. 105, 1075–1087.

Dolferus, R., Osterman, J.C., Peacock, W.J., Dennis, E.S., 1997. Cloning of the *Arabidopsis* and rice formaldehyde dehydrogenase genes: implications for the origin of plant ADH enzymes. Genetics 146, 1131–1141.

Duester, G., et al., 1999. Recommended nomenclatures for the vertebrate alcohol dehydrogenase gene family. Biochem. Pharmacol. 58, 389–395.

Duvall, M.R., et al., 1993. Phylogenetic hypotheses for the monocotyledons constructed from *rbcL* sequence data. Ann. Mo. Bot. Gard. 80, 607–619.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Gaut, B.S., Clegg, M.T., 1993. Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. U. S. A. 90, 5095–5099.

Gaut, B.S., Peek, A.S., Morton, B.R., Clegg, M.T., 1999. Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. 16, 1086–1097.

Gonzàlez-Duarte, R., Albalat, R., 2005. Merging protein, gene and genomic data: the evolution of the MDR-ADH family. Heredity 95, 184–197.

Gu, X., 1999. Statistical method for testing functional divergence after gene duplication. Mol. Biol. Evol. 16, 1664–1674.

Gu, J., Gu, X., 2003. Natural history and functional divergence of protein tyrosine kinases. Gene 317, 49–57.

Gu, X., Velden, K.V., 2002. DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. Bioinform. Appl. Note 18, 500–501.

Guex, N., Peitsch, M.C., 1997. SWISS-MODEL and Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714–2733.

Guindon, S., Gascuel, O., 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89, 10915–10919.

Höög, J.O., Hedberg, J.J., Stromberg, P., Svesson, S., 2001. Mammalian alcohol dehydrogenase — functional and structural implications. J. Biomed. Sci. 8, 71–76.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. 23, 403–405.

Jobb, G., Von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 1–9.

Johnston, J.S., et al., 2005. Evolution of genome size in Brassicaceae. Ann. Bot. 95, 229–235.

Koch, M.A., Haubold, B., Mitchell-Olds, T., 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis* and related genera (Brassicaceae). Mol. Biol. Evol. 17, 1483–1498.

Kumar, S., Tamura, K., Jakobsen, I., Nei, M., 2000. MEGA2: molecular evolutionary genetics analysis software. Bioinform 17, 1244–1245.

Laskowski, R.A., McArthur, M.W., Moss, D.S., Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26, 283–291.

Lüthy, R., Bowie, J.U., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature 356, 83–85.

Mai, D.H., 1995. Tertiäre vegetationsgeschichte Europas. Fischer, Jena.

Mathews, S., Sharrock, R.A., 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the *loci* found in dicot angiosperms. Mol. Biol. Evol. 13, 1141–1150.

Morton, B.R., Gaut, B.S., Clegg, M.T., 1996. Evolution of alcohol dehydrogenase genes in the palm and grass families. Proc. Natl. Acad. Sci. U. S. A. 93, 11735–11739.

Nei, M., 1991. Relative efficiencies of different tree-making methods for molecular data. In: Miyamoto, M., Cracaft, J.L. (Eds.), Phylogenetic Analysis of DNA Sequences. Academic Press, New York, pp. 90–128.

Nicholas, K.B., Nicholas Jr., H.B., 1997. GeneDoc: a tool for editing and annotating multiple sequence alignment Distributed by the authors www.psc.edu/biomed/genedoc.

Peralba, J.M., et al., 1999. Structural and enzymatic properties of a gastric NADP(H)-dependent and retinal-active alcohol dehydrogenase. J. Biol. Chem. 274, 26021–26026.

Perry, D.J., Furnier, G.R., 1996. *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups. Proc. Natl. Acad. Sci. U. S. A. 93, 13020–13023.

Person, B., et al., 1993. Basic features of class-I alcohol dehydrogenase: variable and constant segments coordinated by inter-class and intra-class variability. Conclusions from characterization of the alligator enzyme. Eur. J. Biochem. 216, 49–56.

Rubach, J.K., Plapp, B.V., 2003. Amino acid residues in the nicotinamide binding site contribute to catalysis by horse liver alcohol dehydrogenase. Biochemistry 42, 2907–2915.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sali, A., Blundell, T.L., 1993. Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 243, 779–815.

Tamura, M.N., Yamashita, J., Fuse, S., Haraguchi, M., 2004. Molecular phylogeny of monocotyledons inferred from combined analysis of plastid *matK* and *rbcL* gene sequences. J. Plant Res. 117, 109–120.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalty and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Wilson, M.A., Gaut, B.S., Clegg, M.T., 1990. Chloroplast DNA evolves slowly in the palm family (Arecaceae). Mol. Biol. Evol. 7, 303–314.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comp. Appl. BioSci. 13, 555–556.

Yang, Z., 2004. PAML: Phylogenetic Analysis by Maximum Likelihood Version 3.14. University College London, London.

Yang, Y.W., Lai, K.N., Tai, P.Y., Li, W.H., 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequence and dates of divergence between *Brassica* and other angiosperm lineages. J. Mol. Evol. 48, 597–604.

Yokoyama, S., Yokoyama, R., Kinlaw, C.S., Harry, D.E., 1990. Molecular evolution of the zinc-containing long-chain alcohol dehydrogenase genes. Mol. Biol. Evol. 7, 143–154.

SUPPLEMENTARY MATERIAL

Table 1A

Alcohol dehydrogenase sequences considered, their NCBI accession numbers, and the species from which they were obtained

Botanical families	*Adh* type		Accession numbers	Species
Poaceae	*Adh1*	adh1ORYLAT	AAF37784	*Oryza latifolia*
		adh1ORYALT	AAF37781	*Oryza alta*
		adh1ORYGRA	AAF37403	*Oryza grandiglumis*
		adh1ORYEIC	AAO42684	*Oryza eichingeri*
		adh1ORYRHI	AAO42680	*Oryza rhizomatis*
		adh1ORYOFF	AAO42679	*Oryza officinalis*
		adh1ORYMIT	AAF37401	*Oryza minuta*
		adh1ORYNIV	AAF37389	*Oryza nivara*
		adh1ORYSCH	AAF37411	*Oryza schlechteri*
		adh1ORYMEY	AAF37415	*Oryza meyeriana*
		adh1ORYGNU	AAF37416	*Oryza granulata*
		adh1ORYRID	AAF37413	*Oryza ridleyi*
		adh1ORYLON	AAF37414	*Oryza longiglumis*
		adh1ORYCOA	AAF37412	*Oryza coarctata*
		adh1ORYBRA	AAF37417	*Oryza brachyantha*
		adh1ORYRUF	BAC87775	*Oryza rufipogon*
		adh1ORYGLU	BAC87778	*Oryza glumipatula*
		adh1ORYMER	BAC87779	*Oryza meridionalis*
		adh1ORYSAT	BAC87776	*Oryza sativa*
		adh1ORYPUN	AAF37396	*Oryza punctata*
		adh1ORYAUS	BAC87780	*Oryza australiensis*
		ORYBAR	BAC87777	*Oryza barthii*
		adh1HORVUL	P05336	*Hordeum vulgare*
		adh1HORVULVUL2	AF253472	*Hordeum vulgare* subsp. *vulgare*
		adh1ZEADIP	AAA74637	*Zea diploperennis*
		adh1ZEALUX	AAA74639	*Zea luxurians*
		adh1ZEAMAY	CAA27682	*Zea mays*
		adh1MISSINF	CAD56722	*Miscanthus sinensis* var. *Formorsanus*
		adh1MISFLO	CAD56717	*Miscanthus floridulus*
		adh1MISTRA	CAD56756	*Miscanthus transmorrisonensis*
		adhMISSINGL	CAD56726	*Miscanthus sinensis f. glaber*
		adh1BAMMUL	AAB71522	*Bambusa multiplex*
		adhPENGLA	AJ311047	*Pennisetum glaucum*
		adh1ZIZVIL	AAF37420	*Zizaniopsis villanensis*
		adh1LEEPER	AAF37419	*Leersia perrieri*
		adh1RHYSUB	AAF37418	*Rhynchoryza subulata*
	Adh2	adh2ORYGRA	AAF37783	*Oryza grandiglumis*
		adh2ORYRHI	AAF37778	*Oryza rhizomatis*
		adh2ORYALT	AAO42696	*Oryza alta*
		adh2ORYEIC	AAF37779	*Oryza eichingeri*
		adh2ORYMIT	AAF37780	*Oryza minuta*
		adh2ORYAUS	AAF37786	*Oryza australiensis*
		adh2ORYRID	AAF37787	*Oryza ridleyi*
		adh2ORYSCH	AAF37789	*Oryza schlechteri*
		adh2ORYLON	AAF37788	*Oryza longiglumis*
		adh2ORYNIV	AAF37768	*Oryza nivara*
		adh2ORYBAR	AAF37772	*Oryza barthii*
		adh2ORYOFF	AAO42689	*Oryza officinalis*
		adh2ORYBRA	AAF37795	*Oryza brachyantha*
		adh2ORYCOA	AAF37791	*Oryza coarctata*
		adh2ORYMEY	AAF37793	*Oryza meyeriana*
		adh2ORYGNU	AAF37794	*Oryza granulata*
		adh2ORYGLU	BAE00047	*Oryza glumipatula*
		adh2ORYRUF	BAE00043	*Oryza rufipogon*
		adh2ZEAMAY	P04707	*Zea mays*
		adh2ORYSAT	BAE00044	*Oryza sativa*
		adh2ORYPUN	AAF37775	*Oryza punctata*
		adh2HORVULVUL	P10847	*Hordeum vulgare* subsp. *vulgare*
		adh2HORVULSPO	AAO24260	*Hordeum vulgare* subsp. *spontaneum*
		adh2ORYMER	BAE00049	*Oryza meridionalis*
		adh2ZIZVIL	AAF37798	*Zizaniopsis villanensis*
		adh2LEEPER	AAF37797	*Leersia perrieri*
		adh2RHYSUB	AF37796	*Rhynchoryza subulata*
	Adh3	adh3HORVULVUL	CAA31231	*Hordeum vulgare* subsp. *vulgare*
		adh3HORVULSPO	AAG42522	*Hordeum vulgare* subsp. *spontaneum*

Table 1A (Cont.)

Botanical families	*Adh* type		Accession numbers	Species
Cyperaceae	*Adh1*	adh1CARLUC	AAV66036	*Carex lucorum* var. *lucorum*
		adh1CARTRUG	AAV66013	*Carex tonsa* var. *rugosperma*
		adh1CARCCO	AAV66016	*Carex communis* var. *communis*
		adh1CARDDE	AAV66019	*Carex deflexa* var. *deflexa*
		adh1CARPEC	AAV66024	*Carex peckii*
		adh1CARINO	AAV66009	*Carex inops* subsp. *inops*
		adh1CARFLO	AAV66015	*Carex floridana*
		adh1CARPEN	AAV66022	*Carex pensylvanica*
		adh1CARTON	AAV66025	*Carex tonsa* var. *tonsa*
		adh1CARSER	AAV66008	*Carex serpenticola*
		adh1CARROS	AAV66027	*Carex rossii*
		adh1CARBRE	AAV66005	*Carex brevicaulis*
		adh1CARGLO	AAV66010	*Carex globosa*
		adh1CARGEO	AAV66018	*Carex geophila*
	Adh2	adh2CARSER	AAV66044	*Carex serpenticola*
		adh2CARINO	AAV66055	*Carex inops* subsp. *inops*
		adh2CARROS	AAV66056	*Carex rossii*
		adh2CARGLO	AAV66046	*Carex globosa*
		adh2CARLUC	AAV66036	*Carex lucorum* var. *lucorum*
		adh2CARPEN	AAV66051	*Carex pensylvanica*
		adh2CARDDE	AAV66054	*Carex deflexa* var. *deflexa*
		adh2CARCCO	AAV66049	*Carex communis* var. *communis*
		adh2CARBRE	AAV66035	*Carex brevicaulis*
		adh2CARTRUG	AAV66048	*Carex tonsa* var. *rugosperma*
		adh2CARGEO	AAV66053	*Carex geophila*
Fabaceae	*Adh1*	adh1GLYMAX2	AAC62469	*Glycine max*
		adh1GLYMAX	AAN03476	*Glycine max*
		adh1LOTCOR	CAG30579	*Lotus corniculatus* var. *japonicus*
		adh1SOPFLA	BAD91183	*Sophora flavescens*
		adh1PHAFOL	Z23170	*Phaseolus acutifolius*
		adh1PEA	P12886	*Pisum sativum*
		adhTRIREP	X14826	*Trifolium repens*
	Adh2	adh2WISFLO	BAD91186	*Wisteria floribunda*
Paeoniaceae	*Adh1*	adh1BPAESUF	AAB63520	*Paeonia suffruticosa* subsp. *spontanea*
		adh1BPAEROC	AAB63519	*Paeonia rockii*
		adh1BPAELUT	AAB63518	*Paeonia lutea*
		adh1BPAEDEL	AAB70172	*Paeonia delavayi*
		adh1APAESUF	AAC12907	*Paeonia suffruticosa* subsp. *spontanea*
		adh1APAESZE	AAB63514	*Paeonia szechuanica*
		adh1APAEROC	AAB81208	*Paeonia rockii*
		adh1APAELUT	AAB81207	*Paeonia lutea*
		adh1APAEPAR	AAF04344	*Paeonia parnassica*
		adh1APAETEN	AAB63515	*Paeonia tenuifolia*
		adh1APAEVEI	AAB63516	*Paeonia veitchii*
		adh1APAEANO	AAB81247	*Paeonia anomala*
		adh1APAESIN	AAF04338	*Paeonia sinjiangensis*
		adh1PAEOFF	AAK50892	*Paeonia officinalis*
		adh1APAEARI	AAF04340	*Paeonia arietina*
		adh1APAECAL	AAB63512	*Paeonia californica*
	Adh2	adh2PAESUF	AAB70177	*Paeonia suffruticosa* subsp. *spontanea*
		adh2PAEOFF	AAK50895	*Paeonia officinalis*
		adh2PAELUT	AAB70175	*Paeonia lutea*
		adh2PAESZE	AAB70178	*Paeonia szechuanica*
		adh2PAEROC	AAB70180	*Paeonia rockii*
		adh2PAEHUM	AAF37598	*Paeonia humilis*
		adh2PAEARI	AAF37595	*Paeonia arietina*
		adh2PAETEN	AAB70182	*Paeonia tenuifolia*
		adh2PAEANO	AAB70181	*Paeonia anomala*
		adh2PAEPAR	AAF37600	*Paeonia parnassica*
		adh2PAEVEI	AAB70184	*Paeonia veitchii*
		adh2PAECAL	AAB70174	*Paeonia californica*
		adh2PAEDEL	AAB70176	*Paeonia delavayi*
		adh2PAESIN	AAF37594	*Paeonia sinjiangensis*

Table 1A (Cont.)

Botanical families	*Adh* type		Accession numbers	Species
Arecaceae	*AdhA*	adhAWASROB	AAB39598	*Washingtonia robusta*
		adhAPHOREC	AAB17256	*Phoenix reclinata*
	AdhB	adhBWASROB	AAB39597	*Washingtonia robusta*
	AdhC	adhCALUSI	AAB17255	*Calamus usitatus*
Pinaceae	*AdhC1*	adhC1PINBAN	AAC49539	*Pinus banksiana*
	AdhC2	adhC2PINBAN	AAC49540	*Pinus banksiana*
	AdhC3	adhC3PINBAN	AAC49541	*Pinus banksiana*
	AdhC4	adhC4PINBAN	AAC49542	*Pinus banksiana*
	AdhC5	adhC5PINBAN	AAC49543	*Pinus banksiana*
	AdhC6	adhC6PINBAN	AAC49544	*Pinus banksiana*
	AdhC7	adhC7PINBAN	AAC49545	*Pinus banksiana*
Brassicaceae		adhCRUHIM	BAA34681	*Crucihimalaya himalaica*
		adhARATHA	CAA54911	*Arabidopsis thaliana*
		adhARALYRKAW	BAA34679	*Arabidopsis lyrata* subsp. *kawasakiana*
		adhARAFLA	BAA34678	*Arabis flagellosa*
		adhBARVUL	AF110458	*Barbarea vulgaris*
		adhCARAM	AF110430	*Cardamine amara*
		adhBRAOLE	AB015508	*Brassica oleraceae*
		adhARAPROC	AAF23552	*Arabis procurrens*
		adhARAJAC	AF110446	*Arabis jacquinii*
		adhARABLE	AAF23531	*Arabis blepharophylla*
		adhARASTE	BAA34676	*Arabis stelleri*
		adhARAHIR	AF110443	*Arabis hirsuta*
		adhARABLY	AAF23551	*Arabidopsis lyrata* subsp. *petraea*
		adhARALYRLYR	AAF23547	*Arabidopsis lyrata* subsp. *lyrata*
		adhARAHAL	AAF23540	*Arabidopsis halleri*
		adhOLIPU	BAA34682	*Olimarabidopsis pumila*
		adhARAGRI	AF110440	*Arabidopsis griffithiana*
		adhCRUWAL	BAA34684	*Crucihimalaya wallichii*
		adhARAGLA	AAF23537	*Arabis glabra*
		adhARADRU	AAF23535	*Arabis drummondii*
		adhARAFEN	AAF23536	*Arabis fendleri*
		adhARALYA	AAF23546	*Arabis lyallii*
		adhARAPAR	AAF23548	*Arabis parishii*
		adhARALIG	AAF23545	*Arabis lignifera*
		adhHALPER	AAF23539	*Halimolobos perplexea* var. *lemhiensis*
		adhARAALP	AAF23527	*Arabis alpina*
		adhAUBDEL	AAF23523	*Aubrieta deltoidea*
		adhARATUR	AF110457	*Arabis turrita*
	Adh1	adh1LEASTY	AAC79422	*Leavenworthia stylosa*
		adh1LEACRA	AAC79420	*Leavenworthia crassa*
		adh1LEAUNI	AF037557	*Leavenworthia uniflora*
	Adh2	adh2LEAUNI	AAC79370	*Leavenworthia uniflora*
		adh2LEACRA	AAC79368	*Leavenworthia crassa*
		adh2LEAST	AAC79416	*Leavenworthia stylosa*
	Adh3	adh3LEAUNI	AAC79419	*Leavenworthia uniflora*
		adh3LEACRA	AAC79417	*Leavenworthia crassa*
		adh3LEAST	AAC79418	*Leavenworthia stylosa*

SUPPLEMENTARY MATERIAL

Table 2A

Quality of main-chain and side-chain parameters of the modeled *A. blepharophylla Adh*

Comparison values					
Stereochemical parameters	N° of data pts	Parameter value	Typical value	Band width	No de bandwidths from mean
Main-chain					
% residues in A, B, L	318	94	89.6	10	0.4
Omega angle SD	378	5.2	6	3	-0.3
Bad contacts / 100 residues	12	3.2	0.7	10	0.2
Zeta angle SD	339	1.7	3.1	1.6	-0.9
H-bond energy SD	228	0.7	0.6	0.2	0.4
Overall G-factor	379	-0.1	0	0.3	-0.2
Side-chain					
Chi-1 gauche minus SD	57	8	10.2	6.5	-0.3
Chi-1 trans SD	90	10.1	12.6	5.3	-0.5
Chi-1 gauche plus SD	145	8.1	11	4.9	-0.6
Chi-1 pooled SD	292	8.8	11.5	4.8	-0.6
Chi-2 trans SD	81	9.1	15.7	5	-1.3

SD, standard deviation; pts, points.

SUPPLEMENTARY MATERIAL

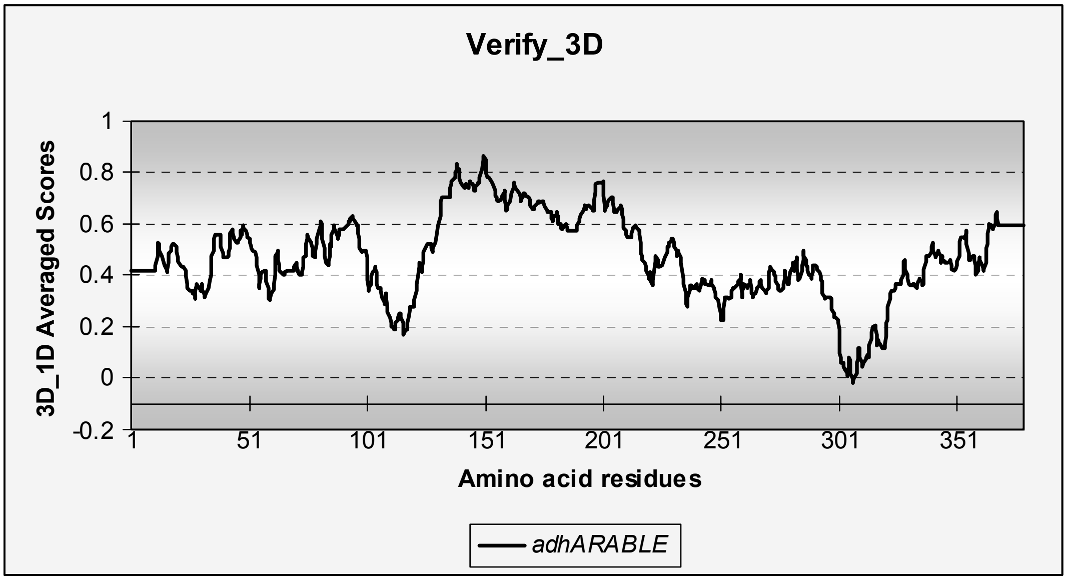


Fig. 1A. VERIFY-3D evaluation of the *Arabis blepharophylla* three-dimensional model.

CAPÍTULO 4

Artigo 2

Aceito no Journal of Molecular Modeling (ISSN: 1610-2940)

Evaluation of the impact of functional diversification on Poaceae, Brassicaceae, Fabaceae, and Pinaceae alcohol dehydrogenase enzymes

Claudia E. Thompson[a], Cláudia L. Fernandes[b], Osmar Norberto de Souza[b], Loreta B. Freitas[a], Francisco M. Salzano[a*]

[a]*Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970, Porto Alegre, RS, Brazil*

[b]*Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga 6681, 90610-001, Porto Alegre, RS, Brazil*

Keywords: Glycolytic proteins; Molecular modeling, Protein structure, Alcohol dehydrogenase, Functional Diversification, Divergence, Enzyme Function, ADH, Molecular evolution

***Corresponding author. Tel.: 55 51 3217-4182; Fax: 55 51 3308-9823; E-mail address: claudia.thompson@ufrgs.br**

Abstract

The plant alcohol dehydrogenases (ADHs) have been intensively studied in the last years in terms of phylogeny and they have been widely used as a molecular marker. However, almost no information about their three-dimensional structure is available. Several studies point to functional diversification of the ADH, with evidence of its importance, in different organisms, in the ethanol, norepinephrine, dopamine, serotonin, and bile acid metabolism. Computational results demonstrated that in plants these enzymes are submitted to a functional diversification process, which is reinforced by experimental studies indicating distinct enzymatic functions as well as recruitment of specific genes in different tissues. The main objective of this article is to establish a correlation between the functional diversification occurring in the plant alcohol dehydrogenase family and the three-dimensional structures predicted for 17 ADH belonging to Poaceae, Brassicaceae, Fabaceae, and Pinaceae botanical families. Volume, molecular weight and surface areas are not markedly different among them. Important electrostatic and pI differences were observed with the residues responsible for some of them identified, corroborating the function diversification hypothesis. These data furnish important background information for future specific structure-function and evolutionary investigations.

Introduction

The alcohol dehydrogenase (ADH) proteins belong to the medium-chain dehydrogenases/reductases (MDR) superfamily which has almost 1000 members spread in all types of organisms. MDR-ADH have been described in bacteria, archaea, yeast, plants and animals, and have been additionally implicated in ethanol oxidation, norepinephrine, dopamine, serotonin and bile acid metabolism (Höög et al. 2001), as well as in the *in vitro* and *in vivo* oxidation of retinol (Boleda et al. 1993; Martras et al. 2004).

Alcohol dehydrogenases (ADHs) are dimeric enzymes of the glycolytic pathway which encode two types of enzymes, one characterized by short protein chains (~250 residues), represented by *Drosophila* ADHs, which do not require zinc as a cofactor; and another characterized by long protein chains (~370 residues), represented by ADHs from organisms as diverse as mammals, plants and yeasts, which require zinc as a cofactor, and are called class P in dicot and monocot plants. The highest specificity of the ADHs among the latter is for ethanol, aldehyde, and acetaldehyde substrates, but they can also utilize other primary alcohols as well (Garabagi et al. 2005).

The catalysis, NAD interactions, evolution, and conformational changes of ADHs have been investigated (Eklund and Bränden 1979; Danielsson et al. 1994; Persson et al. 1994), using three-dimensional structures from the horse liver and focusing on the analysis of the differences among the enzymes of distinct species. Other studies have considered plant *Adh* evolution (Chang and Meyerowitz 1986; Gaut and Clegg 1993; Perry and Furnier 1996; Morton et al. 1996; Miyashita et al. 1998; Charlesworth et al. 1998; Gaut et al. 1999; Koch et al. 2000; Lin et al. 2001). Plant *Adh* transcription has been demonstrated to increase by environmental stresses such as low oxygen levels, dehydration, low temperatures, and in response to the ABA phytohormone (Dolferus et al. 1994). The activation of the fermentation pathway compensates the decrease of the Tricarboxylic Acid Cycle function and of oxidative phosphorilation, regenerating NAD$^+$ and producing energy. Phylogenetic studies indicated two or three isozymes, sometimes more than three, in all flowering dicot and monocot plant species, except in *Arabidopsis*, where a single *Adh* locus is found. Differences in *Adh1* alleles specific activity were detected in maize, while different patterns of tissue-specific expression were observed in the *Adh1* and *Adh2* loci (Gaut and Clegg 1993; Gaut et al. 1999). However, no consideration was given to the relationship between structure and evolution, since there was no three-dimensional model of the plant alcohol dehydrogenases available. Gaut et al. (1999) assumed that the horse and plant ADH structures were similar and mapped some amino acid replacements of plant onto the horse secondary structure. Actually, there is a powerful method to model protein three-dimensional (3D) structures, which makes easier to locate the amino acid residues important to the functional diversification of enzymes and predict substrate preferences. This method (comparative protein structure modeling) estimates the 3D structure of a given protein sequence based on its alignment to one or more templates (Martí-Renom et al. 2000).

Experimental studies have shown the ADH involvement in additional metabolic pathways in plants, indicating putative distinct enzymatic functions during tobacco's pollen tube growth (Bucher et al. 1995) and seed storage (Zhang et al. 1994, 1995a, 1995b, 1997), in potato's pollinated pistils (van Eldik et al. 1997) and in *Petunia*'s seed detoxification (Garabagi et al. 2005).

We recently proposed the first plant ADH three-dimensional model using *Arabis blepharophylla* data (Thompson et al. 2007), obtaining evidence for variation in the subunit-subunit interacting segment, active site and the loop around the second zinc atom. The present work provides 16 other 3D structures, which are considered together with the first described especially in relation to their electrostatic and pI properties. The amino residues theoretically important to the functional divergence among the Poaceae, Brassicaceae, Fabaceae, and Pinaceae modeled ADHs were indicated, as well as those between ADH subtypes, and their position in the 3D structure evaluated to contribute to the elucidation of their functional divergence and molecular evolution.

Materials and methods

Source of the data and sequence alignment

A total of 16 alcohol dehydrogenase sequences were retrieved from the National Center of Biotechnology Information (NCBI) and added to a previous one reported by Thompson et al. (2007). They are listed in Table 1. As indicated there, the 12 species from which they were isolated could be classified in four botanical families. Representatives from ADH1, ADH2 and ADH3 proteins were considered. The ClustalW program (Jeanmougin et al. 1998) was used to perform the alignments, which were inspected and manually changed when necessary using GeneDoc 2.6 (Multiple Sequence Alignment Editor & Shading Utility) (Nicholas and Nicholas 1997).

Modeling

Three-dimensional structures for the 17 ADH enzymes were built using the *Equus caballus* liver form (PDB code 1N8K) as a template, obtained through Blastp (Altschul et al. 1990, 1997). Its structure has been solved to a 1.13 Å resolution (Rubach and Plapp 2003). The ClustalW program (Thompson et al. 1994) was employed to perform the amino acid sequence alignments, using the BLOSUM62 matrix (Henikoff and Henikoff 1992) for scoring. The penalties for gap opening and gap extension were 10.0 and 0.2, respectively. The GeneDoc 2.6 program (Nicholas and Nicholas, 1997) was used to plot the percent identity of the sequences and manually adjust the alignment. The plot is created by sorting the data to be plotted into ascending order. For each data point the fraction of data points which have the same or a smaller value is computed. The data is then compressed to eliminate multiple points with the same value. The highest value is retained during the compression.

The protein models were obtained through MODELLER 8v2 (Sali and Blundell 1993), which implements an approach to comparative protein structure modeling by satisfaction of spatial restraints. The best model was selected using PROCHECK (Laskowski et al. 1993) and VERIFY-3D (Lüthy et al. 1992). The PROCHECK program calculates the stereochemical parameters of the main and side-chains, the residues in the most favored regions, bond lengths and the angle's standard deviation. VERIFY-3D evaluates the compatibility of a 3D model with the amino acid sequence considered using a 3D profile. Each residue position in the 3D model is characterized by its location and environment (alpha, beta, loop, polar, nonpolar, etc), and it is represented by 20 numbers in the profile. These numbers are called 3D_1D scores. The residue environments are defined by three parameters: the residue area that is buried, the fraction of side-chain area that is covered by polar atoms (O and N), and the local secondary structure. If the model is correct, the sum of the 3D profile scores is high, preferentially above zero. The protein signatures were obtained using a database of protein domains known as PROSITE (Gattiker et al. 2002). The protein volumes and surface areas were calculated according to the Richards' Rolling Probe Method (Richards 1974, 1977), using the 3V program (Voss Volume Voxelator) (Voss 2006), with a 1.5 Å probe radius and a high grid resolution (0.5 Å). The theoretical isoelectric point and molecular weight were obtained using the ExPASy Tools available at http://ca.expasy.org/tools/pi_tool.html (Gasteiger et al. 2005). Koch et al. obtained a value (5.81) not significantly different from our results (5.65) for the *Arabis blepharophylla* ADH isoelectric point.

The Swiss PDB Viewer (Guex and Peitsch 1997) was used to calculate the root mean square deviation (RMSD) between the template and the model and also to compute the electrostatic potential using the Coulomb method, as well as to draw all the figures and to generate the molecular surface. The nicotinamide-adenine-dinucleotide (acidic form) and two zinc atoms present in the PDB 1N8K code were located in the modeled three-dimensional structures using its fitting tool. The theoretical models are available upon request.

Functional divergence analysis

The amino acid residues responsible for the functional divergence of the plant ADHs were predicted based on site-specific profiles in combination with suitable cut-off values derived from the posterior probability of each comparison, using Gu's (2001) methodology, as in our previous analysis (Thompson et al. 2007). It is known that functional changes are highly correlated to variations in the evolutionary rates occurring during a certain period of time. Therefore, the identification of the residues submitted to this process in our material were evaluated by finding sites with very different patterns (e.g., very few changes in one cluster but many in the others).

The site-specific profile to identify responsible amino acid sites uses a *Qk* to be the posterior probability that site *k* is in state S1 ($0 \leq Qk \leq 1$). A large *Qk* indicates a high possibility that the functional constraint (or the evolutionary rate) of a site is different between two clusters. We used three cut-off values, equal to or above respectively 0.80, 0.85, and 0.90.

Results

Sequence alignment and modeling

The results obtained with the multiple alignments are presented in Fig. 1, and they show high similarity among the sequences. The degree of identity between the sequence of the selected template and the models was around 48%. In general, the number of gaps in the template's primary sequence is very low (Fig. 1), so it does not significantly affect the comparative molecular modeling. The inserted region of the alignment (Fig. 1 – positions 75 up to 83), which do not have an equivalent segment in the template, was modeled in the context of the whole molecule, using its primary sequence alone. The percent identity of the sequences is presented in Fig. 2. All target proteins have the signature of the zinc alcohol dehydrogenase family, which has a consensual pattern corresponding to G-H-E-X(2)-G-X(5)-[GA]-X(2)-[IVSAC]. Ten models were initially created, and they were considered using PROCHECK and VERIFY-3D, as well as the root mean square deviations (RMSD).

The stereochemical parameters used to verify the quality of the models are listed in Tables 1S-3S (Electronic Supplementary Information). In relation to the main-chain, of the six parameters considered, only the average of bad contacts per 100 residues show some differences among the families considered, the number in the Brassicaceae being higher (8.17) than in the Poaceae, Fabaceae, or Pinaceae (7.57, 6.00, and 7.00), respectively. The figures for the side-chains show better fit (smaller numbers) in relation to Chi1 gauche (-), Chi1-trans, and Chi-pooled SD for the Brassicaceae, the Poaceae model also comparing favorably in relation to the others for Chi-1 gauche (+), Chi1 pooled SD and Chi-2 trans. The mean of percentage of amino acid residues in most favored regions according to the

Ramachandran plot shows variation from 91.67% (Poaceae) to 93.2% (Fabaceae), which is not significant since all results above 90% are considered of good quality. No model value was lower than 90.8%, confirming the excellent quality of the initial models.

Taken together, these data suggest that the models were stereochemically valid. It is important to observe that the G factor measures how "normal" is a given stereochemical property, considering the torsion angles and the bond lengths in the main chain. Therefore, when applied to a specific residue, a low G factor indicates that the property corresponds to a low probability of conformation. A G factor value smaller than -1.0 could indicate geometry problems. In this work, all G factor results are near -0.1 (Tables 1S-3S, Supplementary Information). Observing the VERIFY_3D (Figures 1S-3S, Supplementary Information) results, we can see that the sum of 3D profile scores is high in all cases. The region near the 301 amino acid residue, however, shows the smaller 3D_1D average scores for all botanical families, which means that this is most likely the area with the higher number of structural problems. In a general way, the graphics show a similar pattern.

Number of residues, molecular weight, surface area, and volume

Information concerning these four variables is shown in columns 3-6 of Table 2. *Brassica oleraceae* (1BRAOLE) has a reduced number of amino acids (350), conditioning also lower values for the molecular weight and volume. The opposite occurs in 1ZEAMAY which presents the highest number of residues (388). No clear differences in relation to these variables were observed in the ADHs of different botanical families.

Electrostatic and pI differences

The molecular surface of this protein is electrostatically polarized (Figs. 3-5). The Brassicaceae have the most acid ADH proteins when compared to the other families (Table 2), with *Brassica oleraceae* (1BRAOLE) having the most negative pI (5.47) value, followed by *Arabis blepharophylla* (2ARABLE; 5.65), *Arabis griffthiana* (1ARAGRI; 5.69), and *Arabis parishii* (1ARAPAR; 5.88). The regions of the active site, the second zinc atom, and of the subunit-subunit interacting segment (middle portion, upper and lower right region of the figures, respectively) show the greatest differences (Fig. 3). These proteins have a pI value significantly different from those of the *Leavenworthia* proteins (2LEAST and 3LEAST), which show pI values equal to 6.37 and 6.40, respectively.

Considering now the Poaceae group (Table 2 and Fig. 4), it is seen that the ADH1 forms 1HORVUL and 1ORYSAT (pI 6.28 and 6.20; nos. 1 and 4 in the Figures) are more basic than the ADH2 forms of the same species (respectively 5.52 and 6.04; nos. 2 and 5 in the Figure), the same occurring in *Zea mays* (6.43 and 5.72, nos. 6 and 7 in the Figure). The most significant differences in electrostatic potential is in the region near the second zinc atom and in the subunit-subunit interacting segment (upper and lower right, Fig. 4), a smaller contrast being observed in the active site region.

The pI values for the Fabaceae are not much different (Table 2). However, there is a different concentration of negative charges between the models, the subunit-subunit segment of *Lotus corniculatus* (Fig. 5.1) showing a clear difference from the other two (Figs. 5.2 and 5.3). The *Pinus banksiana* model have a pI value of 5.91 (Table 2), and the protein has the negative charges concentrated in the active site region (Fig. 5.4).

Functional divergence analysis

Sites showing *Qk* values above 0.8 and therefore suggestive of being associated with functional divergences are listed in Table 3 for the comparisons involving different botanical families (60 sequences considered); while in Table 4 the comparisons are between the ADH1 and ADH2 forms. Data related to ADH3 could not be used for this analysis because the number of sequences available was less than those needed for statistical comparisons (Gu and Vander Velden 2002).

Concentrating our attention to the comparisons which yielded $Qk \geq 0.9$ only, we see that in the Brassicaceae vs. Fabaceae contrast, three residues which occur in loops (133, 303, 310) (Table 4S) show different amino acid conservation, the same being true for two others (315, 337) that are located in helices (Table 4S). Positions 133 (Gly), 303 (Ser), and 337 (Gly) are conserved within Brassicaceae, but highly divergent in Fabaceae (133: Asn, Gly, Ser; 303: Asn, Ser, Lys; 337: Asn, Leu, Ser, Gly) (Table 3). The three Fabaceae modeled show variability in position 303 only (Table 4S). In the Poaceae vs. Fabaceae comparative analysis, two amino acid residues of loop regions (118, 133) and one in the helix secondary structure (236) (Table 4S) should be considered; while in the Fabaceae vs. Pinaceae comparison the amino acids to be distinguished are 131 and 133 (loop) and 337 (helix) (Tables 3 and 4S). The Poaceae ADHs present conservation in residues 118 (Asp) and 133 (Gly), and the six Fabaceae in residue 236 (Phe) (Table 3). Considering the three Fabaceae ADH modeled, position 118 is variable in the Poaceae vs. Fabaceae, and position 131 in the Fabaceae vs. Pinaceae comparisons (Table 4S).

As presented in Tables 4 and 5S, both in the functional divergence analysis and in the models, amino acids that show different rates of change between Poaceae's ADH1 and

ADH2 are nos. 234 and 263 (in helices) and 329 (loop), ADH2 being conserved for all of them. In the Poaceae vs. Fabaceae comparison the Poaceae ADH1s exhibit differences in residues 263, located in helix and 329, in a loop.

A ribbon representation of one model of each botanical family showing sites identified as functional divergent ($Qk \geq 0.85$) is presented in Figs. 6 and 7. The subunit-subunit interaction segment seems to be the region with the highest number of functionally important residues in Brassicaceae and Fabaceae (Figs. 6.1 and 7.1, respectively). In Fabaceae the amino acids forming helices and loops around the second zinc atom region are variable (Fig. 7.1). Amino acid changes near the same region distinguish the Poaceae ADH forms, as well as substitutions in the dimer interaction zone (Fig. 6.2). The same regions are fundamental for the diversification of Pinaceae ADHs (Fig 7.2). There are also some differences among all ADHs near the coenzyme region (in green).

Discussion

Alcohol dehydrogenase is an essential enzyme in the anaerobic metabolism, and it has been widely used as a molecular marker in plants due to its convenient size (2-3kb in length with a ~1000 nucleotide coding sequence, 10 exons, 9 introns) and low copy number. The enzyme is important primarily for responses to hypoxic conditions, when its expression is highly induced. Moreover, it has an important role in fruit ripening, seedling and pollen development (Small and Wendel 2000). Despite the large number of phylogenetic investigations performed, no extensive work correlating its sequence and structure in plants exists.

Studies in *Zea mays* have revealed that different alloenzyme types of *Adh1* exhibit different specific activity, and distinct pattern of organ-specific gene expression (Schwartz and Laughner 1969; Freeling and Bennet 1985). An exchange of Tyr for Asp at residue 52, located in a helix structure in the *Adh1-C* allele, alters enzymatic properties by reducing the specific activity. Additionally, amino acid replacements changing the secondary structure were also reported (Gaut and Clegg 1993).

In humans, ADH is a cytosolic enzyme able to metabolize ethanol and a wide variety of substrates, including aliphatic alcohols, hydroxysteroids and lipid peroxidation products. Its catalytic properties are variable. The *Adh2* gene may be present as *Adh2*1*, *Adh2*2*, and *Adh2*3* encoding for β1, β2, and β3 subunits, respectively, which differ by a single nucleotide change. The enzyme containing the β1 subunit has high affinity and low capacity for ethanol, whereas the β2 and β3 forms show lower affinity and higher capacity. Additionally, the human tissues show measurable different *Adh* gene expressions (Gemma et al. 2006).

The proteins modeled in this work are composed by two domains and have a similar fold. The nucleotide binding domain is formed by a structural motif known as Rossmann fold (Lesk, 1995), consisting of parallel beta strands linked by alpha helices (Figs. 6 and 7, region of nicotinamide binding at lower right). The catalytic region containing residues involved in substrate binding has a zinc atom located deeply in the cleft formed between the two domains. There are divergent amino acid residues localized in three important regions (the loop around the zinc atom, an important cofactor for the enzyme's function; the subunit-subunit interacting segment, responsible for the dimer formation; and the active site) which are probably submitted to functional diversification.

Zinc seems to be important for the catalysis and geometry stabilization of the active site. These two processes could be achieved by moderating the electrostatic potential near the substrate or by zinc acting as ligand during the enzyme's catalysis (Baker et al. 2008). Thus the residues indicated as functionally divergent near the zinc atom region possibly have an impact on ADH function. Some residues located near the zinc atom region, such as 109 and 112, which were not previously discussed since they have $0.80 \leq Qk \leq 0.85$, may be also candidates for future investigations. The same can be said of 313 that is related to the subunit-subunit interaction, and residues nos. 49, 62 and 178, present near the active site. The first helix, located in residues 49 up to 55 using 1HORVUL as reference, can accommodate large movements associated with the loop near the active site (Baker et al. 2008); consequently, amino acid no. 64 (loop) has high probability to contribute to these movements.

Clearly the modeled proteins show electrostatic potential differences in the molecular surface. Comparing proteins of same species, ADH1 seems to be more basic than the ADH2 enzymes. *Arabis blepharophylla* ADH1, which was not model, has a theoretical pI equal to 5.74, greater than the 5.65 from the modeled ADH2, corroborating the pattern observed between the ADH forms.

Electrostatic interactions have an important role in the structure and function of biological molecules. Association of proteins in solution and in membranes, enzyme-substrate complexation, chemical reactions in enzyme active sites, charge transfer, are all drastically affected by the strength and distribution of the electrostatic field around regions in biological molecules. The protein-protein interactions are affected by several surfaces properties, such as cavities, hydrophobic residues, specific interaction residue pockets, and electrostatics. This latter has a high potential for functional protein classification (Valeyev, et al. 2008), since that it plays an important role in the specificity of protein-ligand or protein-protein interactions. Due to its attractive or repulsive forces, certain protein-protein interactions could be more or less favorable (Valeyev, et al. 2008). The electrostatic and pI differences described here most certainly lead to dissimilar functional efficiency, a subject that is now open for further investigation. Note that the number of plant proteomic papers is still quite reduced as compared to those of other organisms (only about 3% according to Jorrín et al. 2007).

It is well-known that variation in a specific DNA region not necessarily correlates with the evolutionary pattern of the organism as a whole. Our results summarized in Tables 3 and 4 add new information on this point. In a previous study (Thompson et al., 2007), based on 1155 sites from 176 sequences, we found a close relationship between the Brassicaceae and Fabaceae families. But it is between them that we find the largest number of site differences (a total of 33 with $Qk \geq 0.80$; 20 with $Qk \geq 0.85$; and five with $Qk \geq 0.90$). The other between-family comparisons show much less differences, despite the fact that they are placed far away in the phylogenetic three (Thompson et al., 2007).

The dissimilarities between the Poaceae ADH1 and ADH2 [26 sites with $Qk \geq 0.80$; six with $Qk \geq 0.85$; three with $Qk \geq 0.90$; Table 4] point to the functional differences which exist between these two forms. Our models clearly differentiate them structurally in *H. vulgare*, *O. sativa*, and *Z. mays* (Fig. 4). On the other hand, the ADH1 from the Poaceae and Fabaceae show three sites with clear functional differences. All these findings point to the subtle quantitative changes that occur at the molecular level as a result of the evolutionary process.

Acknowledgements

We would like to thank Dr. Ana Tereza R. Vasconcelos, Dr. Élgion L. S. Loreto, Dr. Laurent E. Dardenne, and Dr. Paulo M. Bisch for their helpful comments. This work was supported by the Institutos do Milênio and Programa de Apoio a Núcleos de Excelência Projects, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, Pró-Reitoria de Pesquisa da Universidade Federal do Rio Grande do Sul, and Pró-Reitoria de Pesquisa e Pós-Graduação da Pontifícia Universidade Católica do Rio Grande do Sul.

References

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

Altschul SF, Madden TL, Schaffer AA, Zhang JZ, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Baker PJ, Britton KL, Fisher M, Esclapez J, Pire C, Bonete MJ, Ferrer J, Rice DW (2008) Active site dynamics in the zinc-dependent medium chain alcohol dehydrogenase superfamily. PNAS 106: 779-784.

Boleda MD, Saubi N, Farrés J, Parés X (1993) Physiological substrates for rat alcohol dehydrogenase classes: aldehydes of lipid peroxidation, omega-hydroxyfatty acids, and retinoids. Arch. Biochem. Biophys. 307: 85-90.

Bucher M, Brander KA, Sbicego S, Mandel T, Kuhlemeier C (1995) Aerobic fermentation in tobacco pollen. Plant. Mol. Biol. 28: 739-750.

Chang C, Meyerowitz EM (1986) Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA 83: 1408-1412.

Charleswort D, Liu FL, Zhang L (1998) The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol. Biol. Evol. 15: 552-559.

Danielsson O, Atrian S, Luque T, Hjelmqvist L, Gonzalez-Duarte R, Jörnvall H (1994) Fundamental molecular differences between alcohol dehydrogenase classes. Proc. Natl. Acad. Sci. USA 91: 4980-4984.

Dolferus R, Jacobs M, Peacock WJ, Dennis ES (1994) Differential interactions of promoter elements in stress responses of the *Arabidopsis Adh* gene. Plant Physiol. 105: 1075-1087.

Eklund H, Bränden CI (1979) Structural differences between apo- and holoenzyme of horse liver alcohol dehydrogenae. J. Biol. Chem. 254: 3458-3461.

Freeling M, Bennet DC (1985) Maize *Adh 1*. Annu. Rev. Genet. 19: 297-323.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) *Protein identification and analysis tools on the ExPASy Server.* In: Walker JM (Ed) The Proteomics Protocols Handbook. Humana Press, Totowa, USA, pp 571-607.

Gattiker A, Gasteiger E, Bairoch A (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. Appl. Bioinformat. 1: 107-108.

Garabagi F, Duns G, Strommer J (2005) Selective recruitment of *Adh* genes for distinct enzymatic functions in Petunia hybrid. Plant Mol. Biol. 58: 283-294.

Gaut BS, Clegg MT (1993) Molecular evolution of the *Adh1* locus in the genus *Zea*. Proc. Natl. Acad. Sci. USA 90: 5095-5099.

Gaut BS, Peek AS, Morton BR, Clegg MT (1999) Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. 16: 1086-1097.

Gemma S, Vichi S, Testai E (2006) Individual susceptibility and alcohol effects: biochemical and genetic aspects. Ann. Ist. Super. Sanità 42: 8-16.

Gu X (2001) Mathematical modeling for functional divergence after gene duplication. J. Comp. Biol. 3: 221-234.

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18: 500-501.

Guex N, Peitsch MC (1997) SWISS-MODEL and Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18: 2714-2733.

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89: 10915-10919.

Höög JO, Hedberg JJ, Stromberg P, Svesson S (2001) Mammalian alcohol dehydrogenase – functional and structural implications. J. Biomed. Sci. 8: 71-76.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem. Sci. 23: 403– 405.

Jorrín JV, Maldonado AM, Castillejo MA (2007) Plant proteome analysis: a 2006 update. Proteomics 7: 2947-2962.

Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidospis*, *Arabis* and related genera (Brassicaceae). Mol. Biol. Evol. 17: 1483-1498.

Laskowski RA, McArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystalogr. 26: 283-291.

Lesk AM (1995) NAD-binding domains of dehydrogenases. Curr. Opin. Struct. Biol. 5: 775-783.

Lin J-Z, Brown AHD, Clegg MT (2001) Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). Proc. Natl. Acad. Sci. USA 98: 531-536.

Lüthy R, Bowie JU, Eisenberg, D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83-85.

Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29: 291-325.

Martras S, Alvarez R, Martinez SE, Torres D, Gallego O, Duester G, Farrés J, de Lera AR, Parés X (2004) The specificity of alcohol dehydrogenase with *cis*-retinoids. Activity with 11-*cis*-retinol and localization in retina. Eur. J. Biochem. 271: 1660-1670.

Miyashita NT, Kawabe A, Innan H, Terauchi R (1998) Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. Mol. Biol. Evol. 15: 1420-1429.

Morton BR, Gaut BS, Clegg MT (1996) Evolution of alcohol dehydrogenase genes in the palm and grass families. Proc. Natl. Acad. Sci. USA 93: 11735-11739.

Nicholas KB, Nicholas HB Jr (1997) GeneDoc: a tool for editing and annotating multiple sequence alignment. Distributed by the authors (www.psc.edu/biomed/genedoc).

Perry DJ, Furnier GR (1996) *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups. Proc. Natl. Acad. Sci. USA 93: 13020-13023.

Persson B, Bergman T, Keung WM, Waldenström U, Holmquist B, Vallee BL, Jörnvall H (1994) Structural and functional divergence of class II alcohol dehydrogenase – cloning and characterisation of rabbit liver isoforms of the enzyme. Eur. J. Biochem. 216: 49-56.

Richards FM (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol. 82: 1-14.

Richards FM (1977) Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6: 151-176.

Rubach JK, Plapp BV (2003) Amino acid residues in the nicotinamide binding site contribute to catalysis by horse liver alcohol dehydrogenase. Biochemistry 42: 2907-2915.

Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 243: 779-815.

Small RL, Wendel JF (2000) Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155: 1913-1926.

Schwartz D, Laughner WJ (1969) A molecular basis for heterosis. Science 166: 626-627.

Thompson CE, Salzano FM, Souza ON, Freitas LB (2007) Sequence and structural aspects of the functional diversification of plant alcohol dehydrogenases. Gene 396: 108-115.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalty and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Valeyev NV, Downing AK, Sondek J, Deane C (2008) Electrostatic and functional analysis of the seven-bladed WD β-Propellers. Evol. Bioinf. 4: 203-216.

Van Eldik GJ, Ruiter RK, Van Herpen MMA, Schrauwen JAM, Wullems GJ (1997) Induced ADH gene expression and enzyme activity in pollinated pistils of *Solanum tuberosum*. Sex Plant Reprod 10: 107–109.

Voss NR (2006) The geometry of the ribosomal polypeptide exit tunnel. J. Mol. Biol. 360: 893-906.

Zhang M, Maeda Y, Furihata Y, Nakamaru Y, Esashi Y (1994) A mechanism of seed deterioration in relation to the volatile compounds evolved by dry seeds themselves. Seed Sci. Res. 4: 49–56.

Zhang M, Yajima H, Umezawa Y, Nakagawa Y, Esashi Y (1995a). GC-MS identification of volatile compounds evolved by dry seeds in relation to storage conditions. Seed Sci. Tech. 23: 59–68.

Zhang M, Nakamaru Y, Tsuda S, Nagashima T, Esashi Y (1995b) Enzymatic conversion of volatile metabolites in dry seeds during storage. Plant Cell Physiol. 36: 157–164.

Zhang M, Nagata S, Miyazawa K, Kikuchi H, Esashi Y (1997) A competitive enzyme-linked immunosorbent assay to quantify acetaldehyde-protein adducts that accumulate in dry seed during aging. Plant Physiol. 113: 397–402.

Table 1

Alcohol dehydrogenase sequences considered, their NCBI accession numbers, and the species from which they were obtained[1]

Botanical family	Abbreviation	NCBI accession number	Species
Brassicaceae	1BRAOLE	BAA34686	*Brassica oleraceae*
	2ARABLE	AAF23531	*Arabis blepharophylla*
	1ARAGRI	AAF23538	*Arabidopsis griffithiana (Arabidopsis pumila* var. *griffithiana)*
	1ARAPAR	AAF23548	*Arabis parishii*
	2LEAST[2]	AAC79416	*Leavenworthia stylosa*
	3LEAST	AAC79418	*Leavenworthia stylosa*
Poaceae	1HORVUL	AAK49116	*Hordeum vulgare subsp. vulgare*
	2HORVUL	P10847	*Hordeum vulgare*
	3HORVUL	CAA31231	*Hordeum vulgare subsp. vulgare*
	1ORYSAT	BAC87776	*Oryza sativa subsp. indica*
	2ORYSAT	BAE00044	*Oryza sativa subsp. indica*
	1ZEAMAY	Q5GA23	*Zea mays*
	2ZEAMAY	P04707	*Zea mays*
Fabaceae	1LOTCOR	CAG30579	*Lotus corniculatus*
	1TRIREP	CAA32934	*Trifolium repens*
	1PISSAT	P12886	*Pisum sativum*
Pinaceae	1PINBAN	AAC49539	*Pinus banksiana*

[1] The number before the sequence identification indicates the ADH subtype (ADH1, ADH2, ADH3).

[2] Only a partial sequence of *Leavenworthia stylosa* ADH1 sequence was described, preventing its modeling.

Table 2

Theoretical values obtained for the ADH models and the template from *Equus caballus*[1]

			Theoretical values			
Botanical families	**Abbreviation**	**Number of residues**	**Molecular weight (D)**	**Surface area (Å^2)**	**Volume (Å^3)**	**Isoelectric point**
Brassicaceae	1BRAOLE	350	38001.58	14006.5	47976.37	5.47
	2ARABLE	379	40994.03	13889.1	51463.00	5.65
	1ARAGRI	379	41308.23	13923.9	51977.75	5.69
	1ARAPAR	379	41165.19	13729.7	52032.37	5.88
	2LEAST	379	41454.79	13613.9	52275.25	6.37
	3LEAST	380	41255.52	14081.2	52307.50	6.40
Poaceae	1HORVUL	379	40903.29	13873.0	51820.50	6.28
	2HORVUL	373	40511.62	13683.4	51113.75	5.52
	3HORVUL	379	41011.48	14243.1	52027.37	6.08
	1ORYSAT	379	40984.30	13994.8	52085.75	6.20
	2ORYSAT	379	41176.75	14089.5	52134.00	6.04
	1ZEAMAY	388	41975.50	14529.8	53186.50	6.43
	2ZEAMAY	379	41054.43	14467.6	52977.87	5.72
Fabaceae	1LOTCOR	380	41096.13	14156.2	51981.75	5.92
	1TRIREP	380	41172.33	14336.3	52204.00	6.08
	1PISSAT	380	41155.37	14198.7	52050.25	6.09
Pinaceae	1PINBAN	375	40465.59	13794.6	51078.00	5.91
Template	1N8K	374	39806.29	13187.4	51493.12	8.31

[1] The number before the sequence identification indicates the ADH subtype (ADH1, ADH2, ADH3).

Table 3

Amino acid residues changes associated with the functional divergence among the botanical families[1]

Comparison[2]	Amino acid residue position	Amino acid residue	
Brassicaceae (31) vs. Poaceae (16)		in Brassicaceae	in Poaceae
	236	F	F, Y, H
Brassicaceae (31) vs. Pinaceae (7)		in Brassicaceae	in Pinaceae
	271	R	Y, C
	310	T, S	T
	315	F, L	F
	317	N	N, C, T, S
Brassicaceae (31) vs. Fabaceae (6)		in Brassicaceae	in Fabaceae
	45	F	*Y, F*
	49	C, *S, W*	C
	57	E	E, D
	64	L, *W, R*	L
	82	V, I, A	V
	90	Q, A, K	K
	112	E, *V, G*	E
	125	E, D	D
	127	G, V, R	G
	128	G, V	V
	130	I	I, L
	133*	G	N, G, S
	135	S	S, T
	139	I	I, K
	178	I	I, V
	187	L	*F, L*
	188	G, E, R	G
	190	T, V, I, P	T
	194	A, V	A
	213	A, G	A
	219	R, K	R
	221	A, S	S
	224	S, G	S
	237	D, E	E
	241	K, E	K
	295	V	V, *L, T*
	303*	S	*N, S, K*
	310*	T, S	T
	311	H	H, *A, N*
	315*	F, L	F
	337*	G	N, *L, S, G*
	338	V, I, L	V
	344	N	*N, K* , *R, S*
Poaceae (16) vs. Fabaceae (6)		in Poaceae	in Fabaceae
	118*	D	D, E, N
	133*	G	N, G, S
	236*	F, Y, H	F
	279	I, V, A	I
Fabaceae (6) vs. Pinaceae (7)		in Fabaceae	in Pinaceae
	131*	*S, H, N*	S
	133*	N, G, S	G
	209	A	G, A, *T, S*
	271	R	Y, C
	337*	N, *L, S, G*	G
Poaceae (16) vs. Pinaceae (7)		in Poaceae	in Pinaceae
	161	V	V, A, S
	209	A	G, A, *T, S*
	271	R	Y, C
	313	M	V, L, I

[1] Only sequences which yielded $Qk \geq 0.80$ are listed; amino acid residues with $Q(k) \geq 0.85$ are in bold face, and those with $Qk \geq 0.90$ are distinguished by an asterisk (*). The amino acid residues are displayed by decreasing order of frequency. Residues in italics are those with the same frequency. Those in italics and gray have smaller frequencies than the residues placed before them.

[2] Numbers in parentheses indicate the number of sequences used in this analysis (data supplied on request).

Table 4

Amino acid residues changes associated with the functional divergence between ADH1 and ADH2[1]

Comparison[2]	Amino acid residue position	Amino acid residue	
Poaceae ADH1 (9) vs. Poaceae ADH2 (6)		Poaceae ADH1	Poaceae ADH2
	25	V, S, T	S
	41	V	V, *D, I*
	45	F, Y	Y
	62	T, I	T
	64	V, M	V
	79	V, I	V
	109	C, S	C
	112	A, P	E
	170	A, E, Q	E
	178	V	I, L
	183	I	F, I
	185	T, S	T
	190	T, S	T
	200	S	Q, *M, S*
	204	I, V	I
	221	A	S, A
	229	I, V	V
	233	A, P	P
	234*	N, S, V	A
	236	F	F, *H, Y*
	240	R, K	K
	259	Q, E	E
	263*	E, D	E
	285	A	C, A
	329*	Y, F	Y
	337	N	N, G
Poaceae ADH1 (9) vs. Fabaceae ADH1 (6)		Poaceae ADH1	Fabaceae ADH1
	64	V, M	L
	263*	E, D	E
	329*	Y, F	Y
Poaceae ADH2 (6) vs. Fabaceae ADH1 (6)		Poaceae ADH2	Fabaceae ADH1
	41	V, *D, I*	L
	118	D	D, E, N
	133	G	N, G, S
	221	S, A	S
	236	F, *Y, H*	F
	238	Q	*L, E, G, Q*
	279	I	I, *V, A*
	285	C, A	A

[1] Only sequences which yielded *Qk* ≥ 0.80 are listed; amino acid residues with Q(k) ≥ 0.85 are in bold face, and those with *Qk* ≥ 0.90 are distinguished by an asterisk (*). The amino acid residues are displayed by decreasing order of frequency. Residues in italics are those with the same frequency. Those in italics and gray have smaller frequencies than the residues placed before them.

[2] Numbers in parentheses indicate the number of sequences used in this analysis (data supplied on request).

Fig. 1 Multiple alignment of the protein sequences modeled and the template used in the modeling.

Fig. 2 Percent identity of the ADH sequences. The horizontal axis presents the data values being plotted. The vertical axis shows the fraction of data points with as small or smaller a data value.

Fig. 3 View of the surface topology of the Brassicaceae ADH models with the electrostatic potential represented as red (most negative), white (neutral) and blue (most positive). Numbers in black refer to the sites identified as showing functional divergence ($Qk \geq 0.90$) among botanical families. Those numbered 315 and 337 are placed on the other side of the figure and cannot be displayed. Since the molecules are shown in the same position, only the first was labeled.

Fig. 4 View of the surface topology of the Poaceae ADH models with the electrostatic potential represented as red (most negative), white (neutral) and blue (most positive). Numbers in black refer to the sites identified as showing functional divergence ($Qk \geq 0.90$) among botanical families. That numbered 236 is placed on the other side of the figure and cannot be displayed. The number in blue refers to the site showing functional divergence ($Qk \geq 0.90$) between ADH forms. Sites nos. 234 and 329 are placed on the other side of the figure. Since the molecules are shown in the same position, only the first was labeled.

Fig. 5 View of the surface topology of the Fabaceae and Pinaceae ADH models with the electrostatic potential represented as red (most negative), white (neutral) and blue (most positive). Numbers in black refer to the sites identified as showing functional divergence ($Qk \geq 0.90$) among botanical families. Those numbered 315 and 337 are

on the other side of the figure. The number in blue refers to the site showing functional divergence (*Qk* ≥ 0.90) between ADH forms. Site no. 329 is on the other side of the figure. Since the molecules are shown in the same position, only the first was labeled.

Fig. 6 Ribbon representation of the ADHs three-dimensional structures in the same orientation shown in Figs. 3-4: (1) 2ARABLE and (2) 1HORVUL. Numbers in black refer to the sites identified as showing functional divergence (*Qk* ≥ 0.85) among botanical families. Numbers in blue identified sites showing functional divergence (*Qk* ≥ 0.85) between ADH forms. Zinc atoms are displayed in blue, and the nicotinamide-adenine-dinucleotide (acidic form) is shown in green.

Fig. 7 Ribbon representation of the ADHs three-dimensional structures in the same orientation shown in Fig. 5: (1) 1LOTCOR and (2) 1PINBAN. Sites showing functional divergence (*Qk* ≥ 0.85) among botanical families are in black. Those showing functional divergence (*Qk* ≥ 0.85) between ADH forms are in blue. Residues 133 and 236 are distinguished by an asterisk (*) in 1LOTCOR, since they are important both to the divergence among botanical families and between ADH forms, as can be seen in Tables 3, 4, 4S, and 5S. Zinc atoms are displayed in blue, and the nicotinamide-adenine-dinucleotide (acidic form) is shown in green.

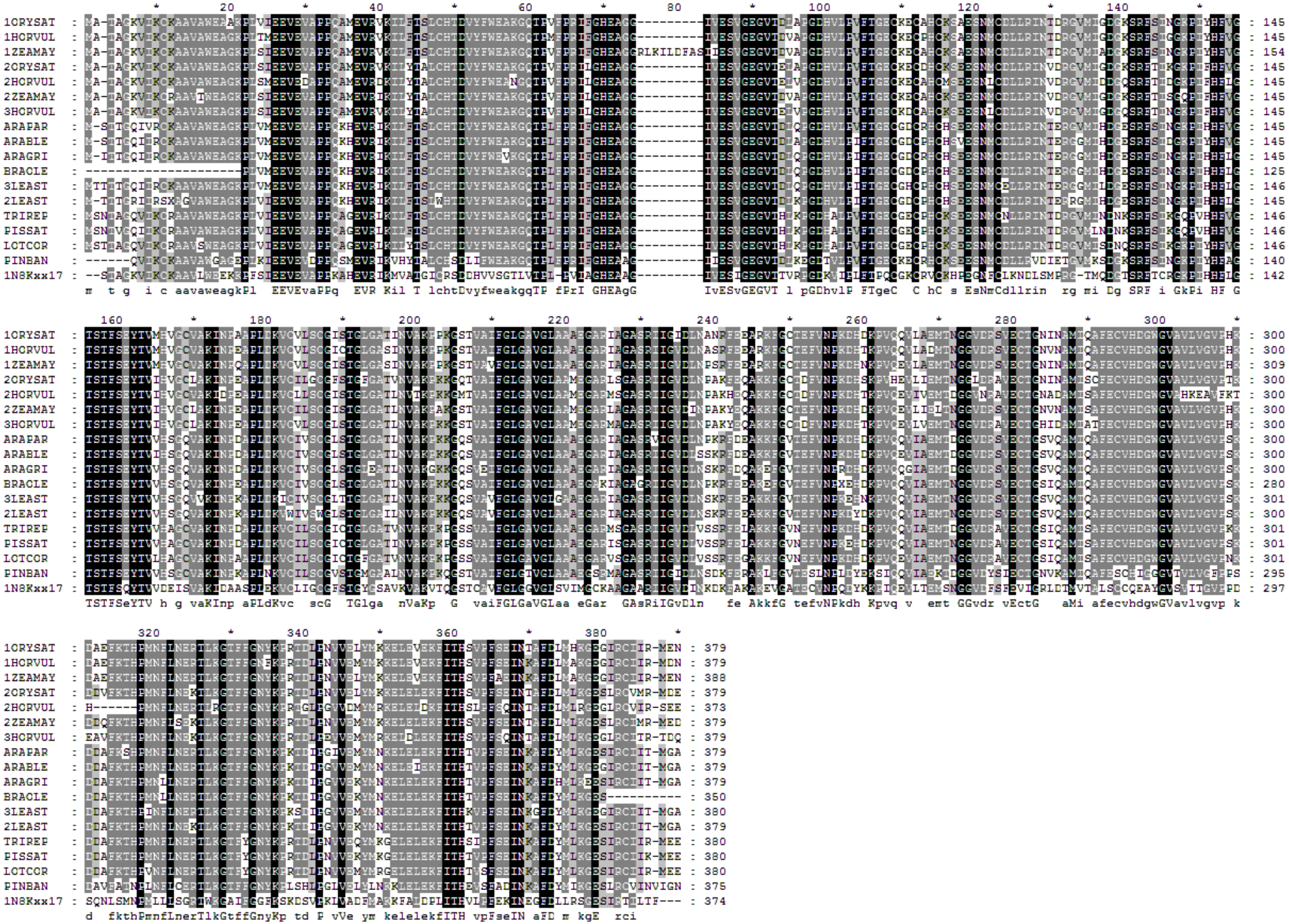

1ORYSAT
1HORVUL
1ZEAMAY
2ORYSAT
2HORVUL
2ZEAMAY
3HORVUL
ARAPAR
ARABLE
ARAGRI
BRAOLE
3LEAST
2LEAST
TRIREP
PISSAT
LOTCOR
PINBAN
1N8Kxx17

Fig. 1

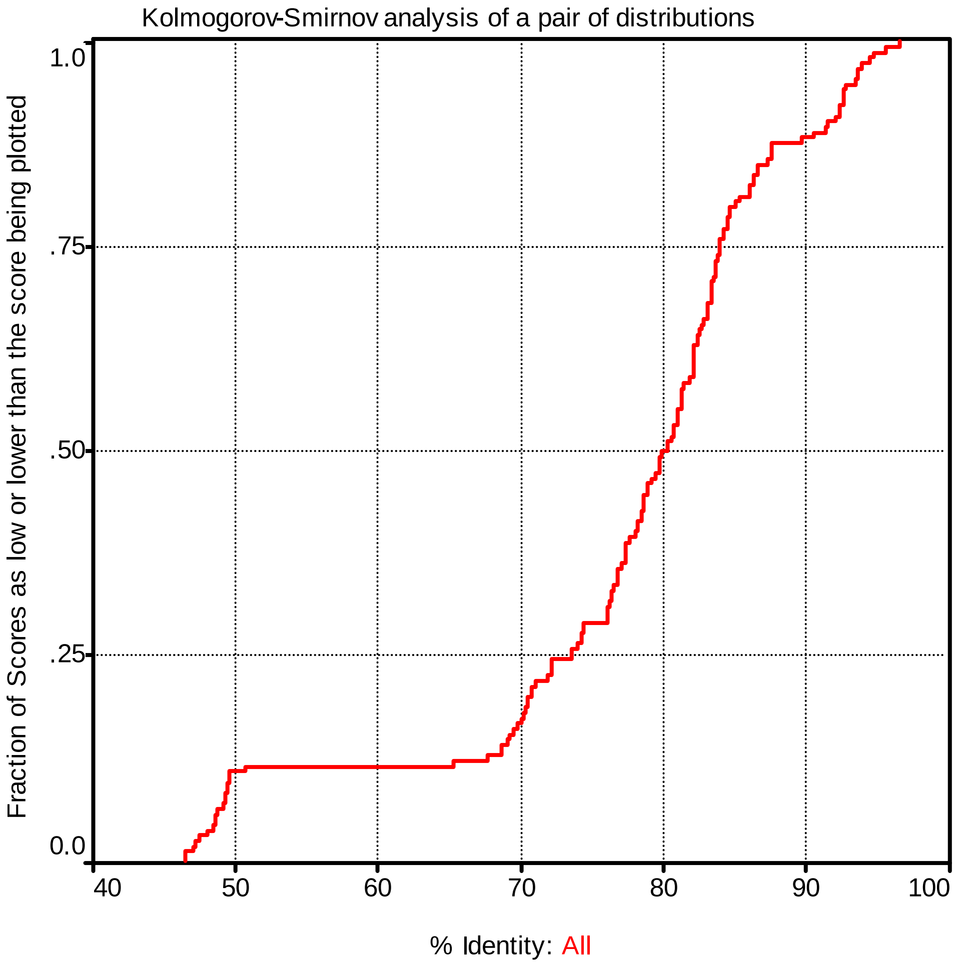
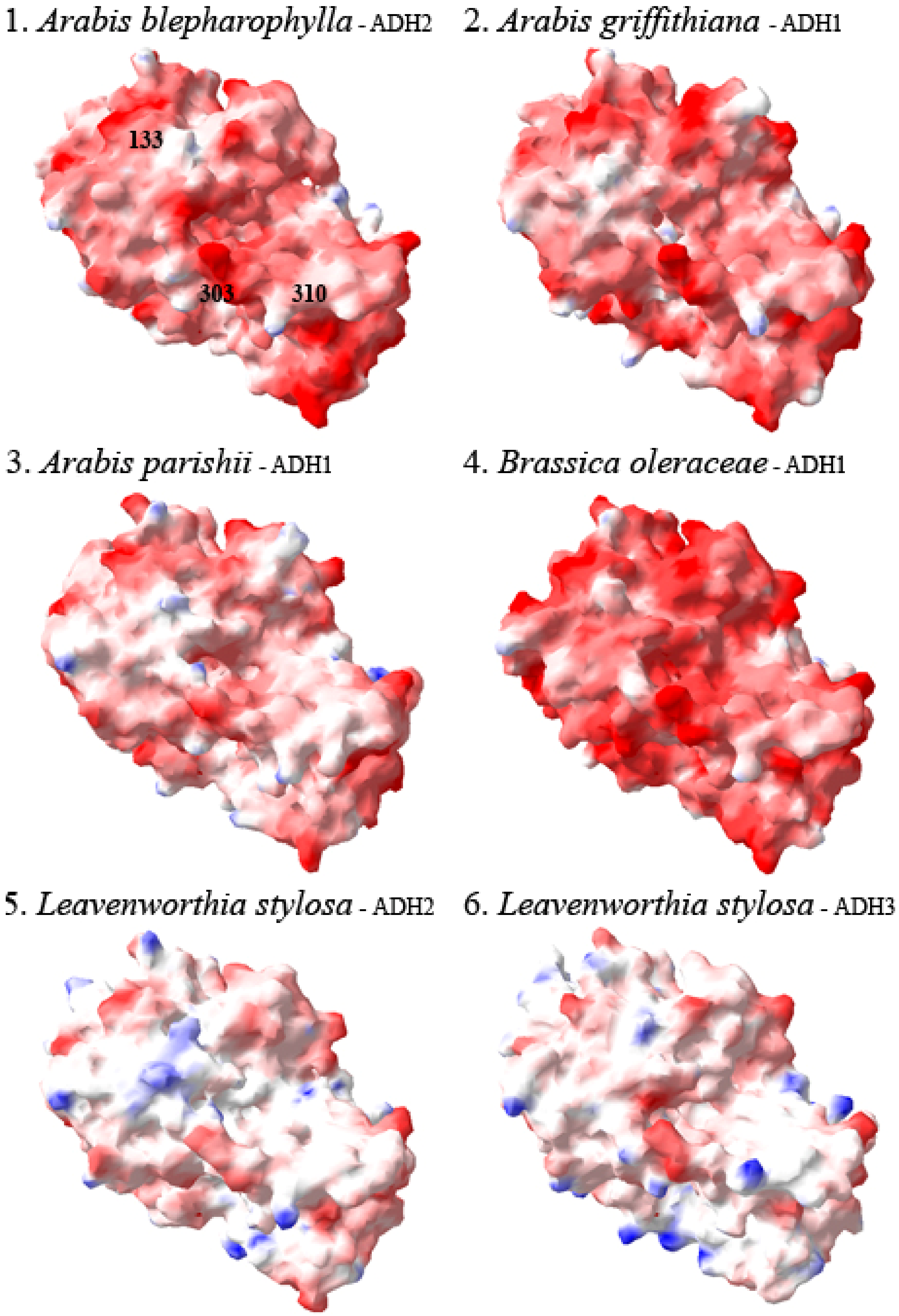

Kolmogorov-Smirnov analysis of a pair of distributions
Fraction of Scores as low or lower than the score being plotted
1.0
.75
.50
.25
0.0
40
50
60
70
80
90
100
% Identity: All

Fig. 2



1. *Arabis blepharophylla* - ADH2

2. *Arabis griffithiana* - ADH1

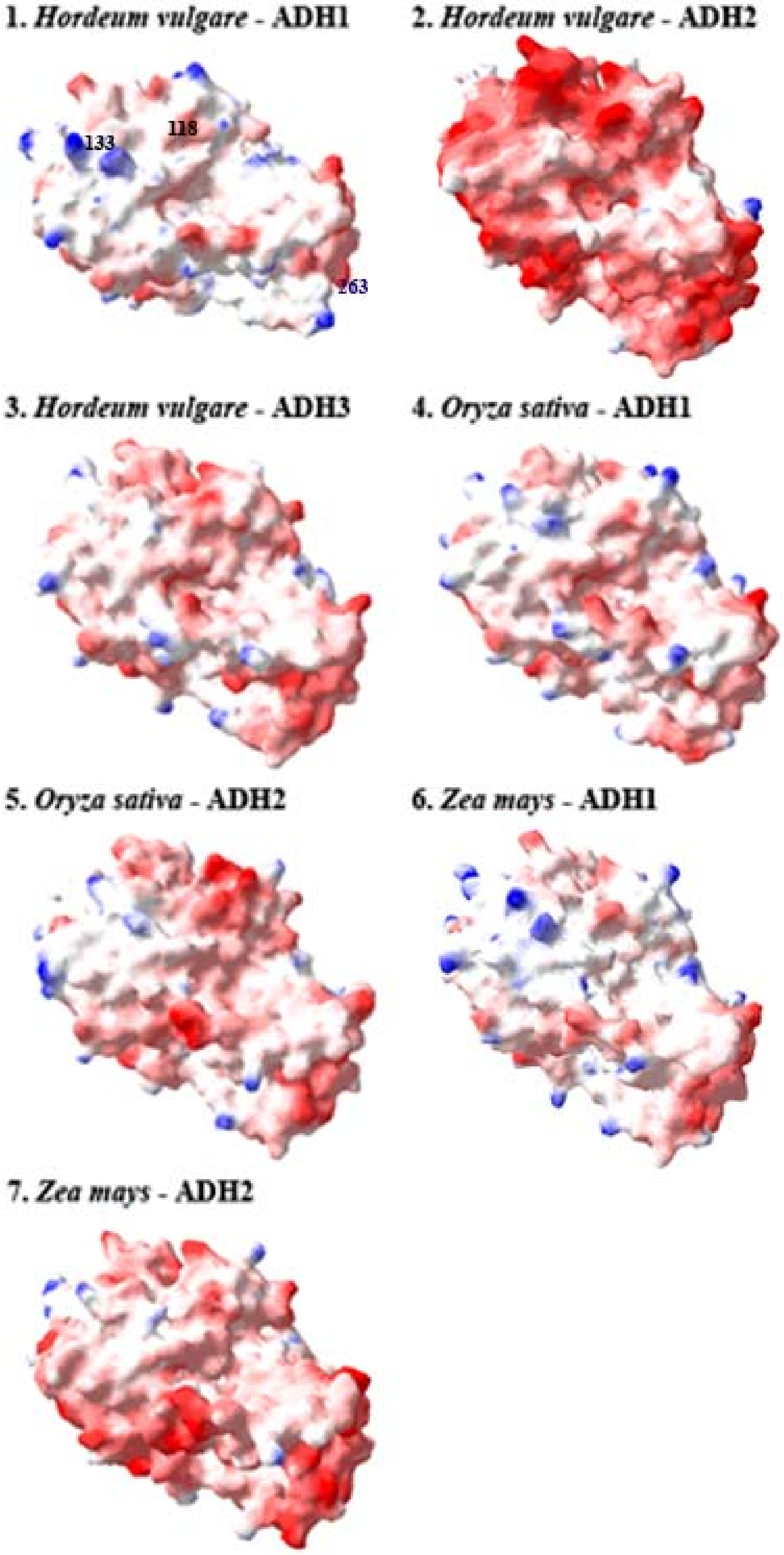
3. *Arabis parishii* - ADH1

4. *Brassica oleraceae* - ADH1

5. *Leavenworthia stylosa* - ADH2

6. *Leavenworthia stylosa* - ADH3

Fig. 3



1. *Hordeum vulgare* - ADH1

2. *Hordeum vulgare* - ADH2

3. *Hordeum vulgare* - ADH3

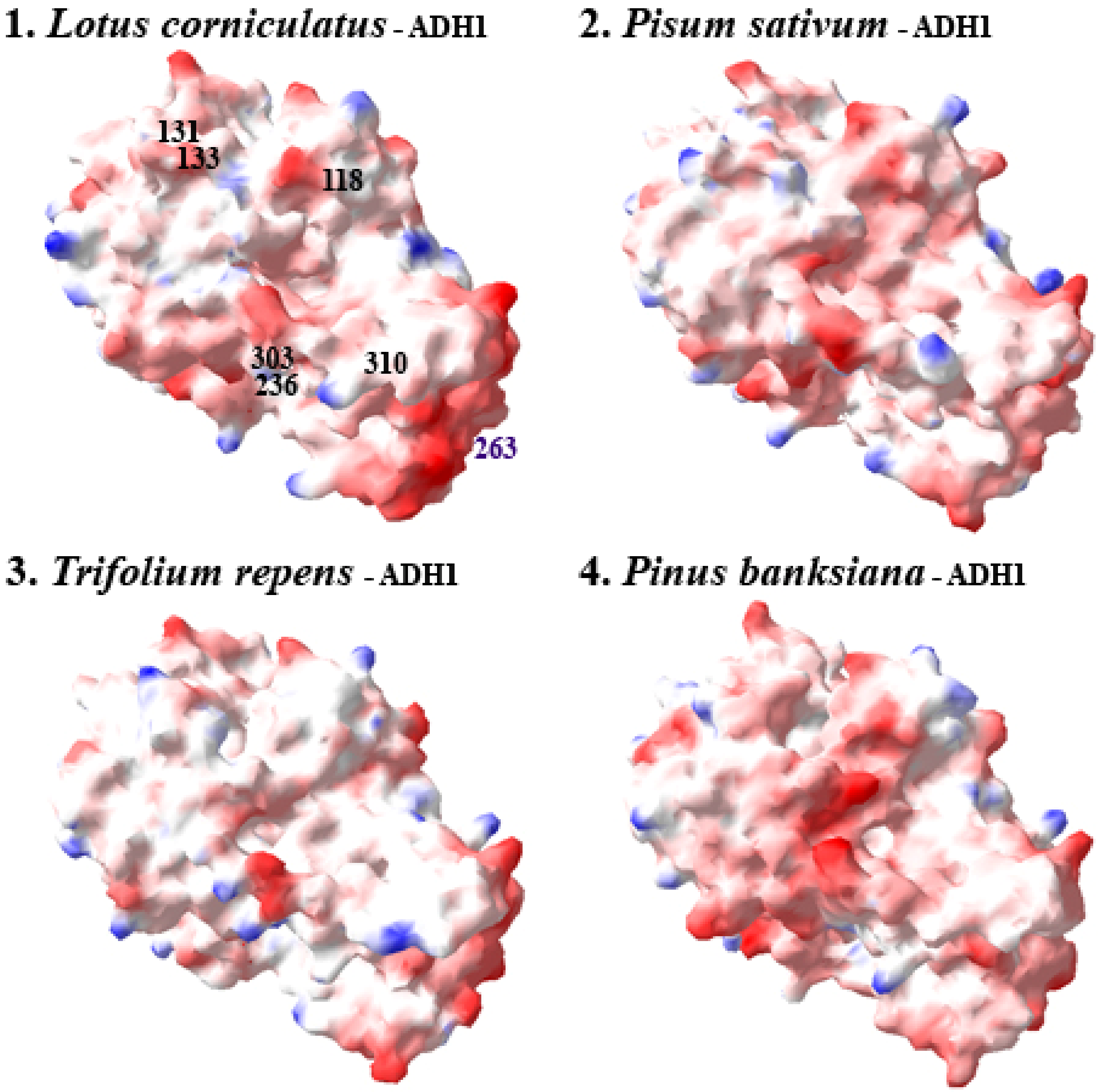
4. *Oryza sativa* - ADH1

5. *Oryza sativa* - ADH2

6. *Zea mays* - ADH1

7. *Zea mays* - ADH2

Fig. 4



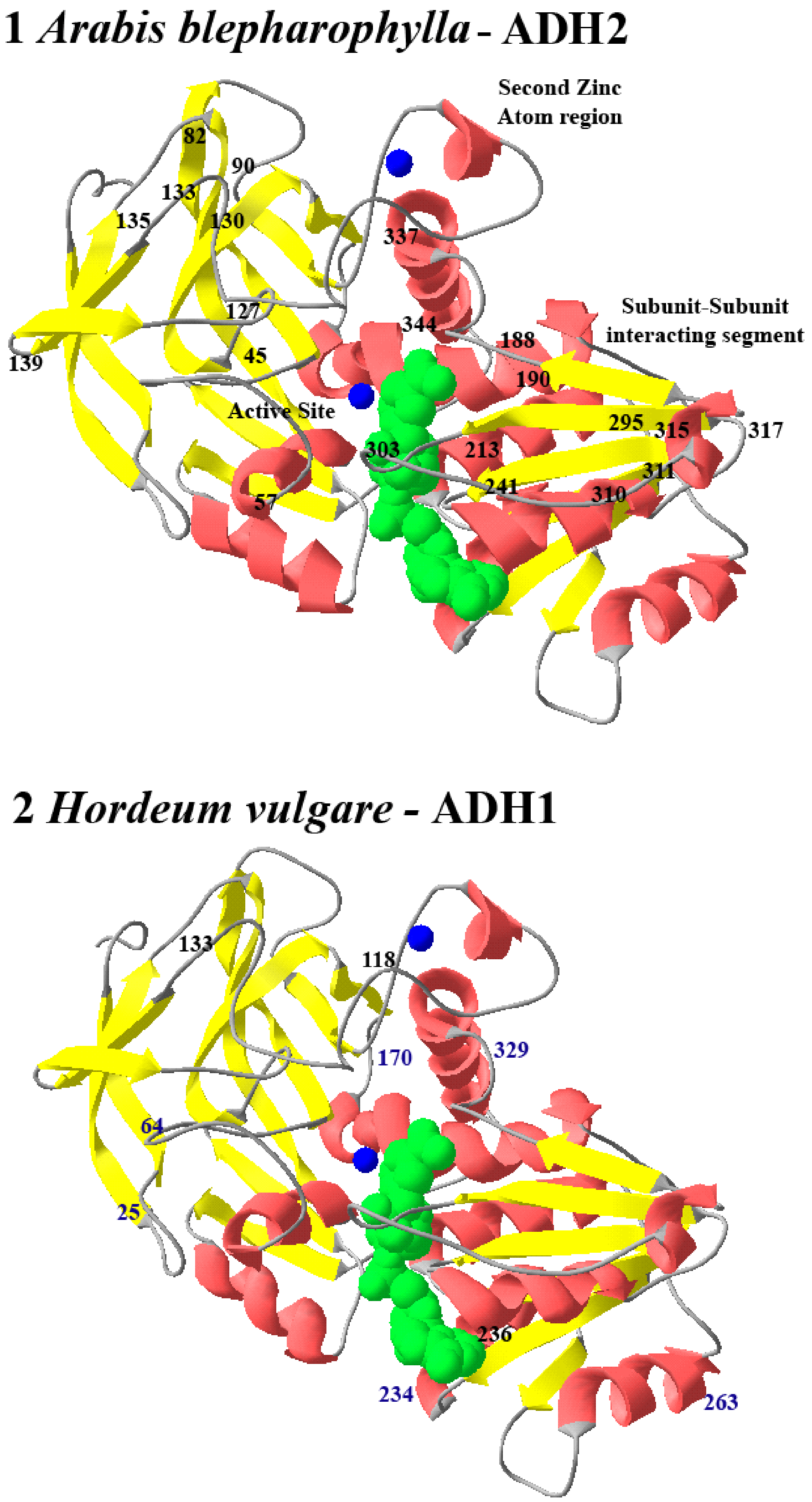
1. ***Lotus corniculatus*** **- ADH1**

2. ***Pisum sativum*** **- ADH1**

3. ***Trifolium repens*** **- ADH1**

4. ***Pinus banksiana*** **- ADH1**

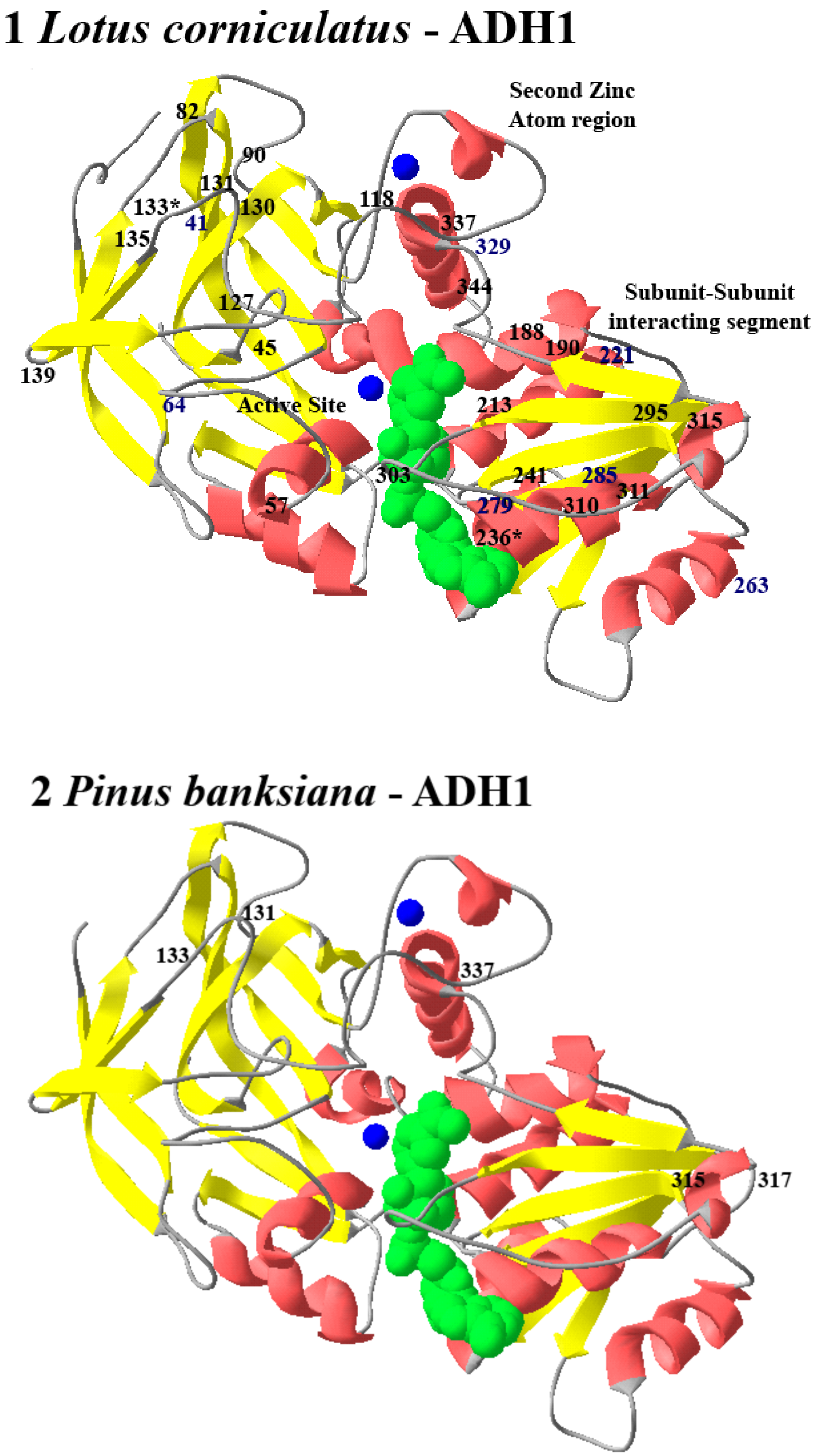
Fig. 5



1 *Arabis blepharophylla* - ADH2

2 *Hordeum vulgare* - ADH1

Fig. 6



1 *Lotus corniculatus* - ADH1

2 *Pinus banksiana* - ADH1

Fig. 7

Electronic Supplementary Information

Table 1S. Quality of main-chain and side-chain parameters of the modeled Brassicaceae ADH

		Comparative values - Brassicaceae					
Stereochemical parameters		ARABLE[1]	ARAGRI	ARAPAR	BRAOLE	2LEAST	3LEAST
Main-chain							
% residues in A, B, L	(1)	318	321	320	292	319	317
	(2)	94	92.8	92.5	92.8	92.5	93.1
	(3)	0.4	0.3	0.3	0.3	0.3	0.3
Omega angle SD	(1)	378	378	377	348	378	378
	(2)	5.2	5.3	4.6	4.3	5.1	4.4
	(3)	-0.3	-0.2	-0.5	-0.6	-0.3	-0.5
Bad contacts / 100 residues	(1)	12	5	8	6	11	7
	(2)	3.2	1.3	2.1	1.7	2.9	1.8
	(3)	0.2	0.1	0.1	0.1	0.2	0.1
Zeta angle SD	(1)	339	340	341	312	339	338
	(2)	1.7	1.3	1.5	1.4	1.6	1.4
	(3)	-0.9	-1.1	-1	-1.1	-1	-1.1
H-bond energy SD	(1)	228	235	231	202	230	230
	(2)	0.7	0.7	0.7	0.6	0.7	0.7
	(3)	0.4	0.5	0.5	0.4	0.5	0.5
Overall G-factor	(1)	379	379	379	350	379	380
	(2)	-0.1	-0.1	-0.1	-0.1	-0.1	0
	(3)	-0.2	-0.2	-0.3	-0.1	-0.3	-0.2
Side-chain							
Chi-1 gauche (-) SD	(1)	57	44	49	49	56	53
	(2)	8	6.7	6.9	6.2	6.5	5.7
	(3)	-0.3	-0.5	-0.5	-0.6	-0.6	-0.7
Chi-1 trans SD	(1)	90	108	109	98	99	101
	(2)	10.1	8.3	9	7.6	9.7	10.1
	(3)	-0.5	-0.8	-0.7	-0.9	-0.5	-0.5
Chi-1 gauche (+) SD	(1)	145	146	137	123	139	141
	(2)	8.1	6.3	7.1	7	6.6	6.1
	(3)	-0.6	-1	-0.8	-0.8	-0.9	-1
Chi-1 pooled SD	(1)	292	298	295	270	294	295
	(2)	8.8	7.2	7.8	7.1	7.8	7.5
	(3)	-0.6	-0.9	-0.8	-0.9	-0.8	-0.8
Chi-2 trans SD	(1)	81	73	90	74	82	82
	(2)	9.1	10.3	9.4	8.7	11.8	9.4
	(3)	-1.3	-1.1	-1.3	-1.4	-0.8	-1.3

Note: Numbers indicate the following attributes: (1) N° of data points; (2) Parameter value; (3) N° of bandwidths from mean. The stereochemical parameters that are not immediately obvious are as follows: % residues in ABL: % of protein's residues in most favored regions according to the Ramachandran plot; Bad contacts are defined as the number of non-bounded atoms in a distance ≥ 2.6 Å; Overall G-factor: measures how "normal" is the specified stereochemical property; Chi1 and Chi2 are the side chain torsion angles, and the different conformations are referred to as gauche (+), trans or gauche (-) according to their position in relation to the main chain carbonyl group and the torsion degree (+/-60 or 180 degrees). SD: Standard deviation. Typical values and bandwidth for refined structures at a similar resolution are as follows: (a) 89.6; 10; (b) 6; 3; (c) 0.7; 10.

[1] According to Thompson, C.E., Salzano, F.M., Souza, O.N., Freitas, L.B., 2007. Sequence and structural aspects of the functional diversification of plant alcohol dehydrogenases. Gene 396, 108-115.

Table 2S. Quality of main-chain and side-chain parameters of the modeled Poaceae ADH[1]

		Comparative values - Poaceae						
Stereochemical parameters		1HORVUL	2HORVUL	3HORVUL	1ORYSAT	2ORYSAT	1ZEAMAY	2ZEAMAY
Main-chain								
% residues in A, B, L	(1)	319	316	321	322	321	330	332
	(2)	91.5	90.8	92.2	91.9	92.2	90.9	92.2
	(3)	0.2	0.1	0.3	0.2	0.3	0.1	0.3
Omega angle SD	(1)	377	371	378	377	378	387	378
	(2)	4.3	4.7	5.2	4.3	5.4	5.5	5.3
	(3)	-0.6	-0.4	-0.3	-0.6	-0.2	-0.2	-0.2
Bad contacts / 100 residues	(1)	9	11	6	6	8	5	8
	(2)	2.4	2.9	1.6	1.6	2.1	1.3	2.1
	(3)	0.2	0.2	0.1	0.1	0.1	0.1	0.1
Zeta angle SD	(1)	341	336	342	343	342	352	343
	(2)	1.3	1.3	1.4	1.3	1.3	1.7	1.5
	(3)	-1.1	-1.1	-1.1	-1.1	-1.2	-0.9	-1.0
H-bond energy SD	(1)	225	224	227	228	229	232	228
	(2)	0.7	0.7	0.7	0.7	0.7	0.7	0.7
	(3)	0.3	0.4	0.4	0.3	0.4	0.5	0.4
Overall G-factor	(1)	379	373	379	379	379	388	379
	(2)	-0.1	-0.2	-0.1	-0.1	-0.1	-0.2	-0.1
	(3)	-0.2	-0.5	-0.2	-0.2	-0.2	-0.4	-0.2
Side-chain								
Chi-1 gauche (-) SD	(1)	45	49	38	52	49	56	49
	(2)	10.2	7.1	8.5	7.7	6.5	8.1	8.0
	(3)	0	-0.5	-0.3	-0.4	-0.6	-0.3	-0.3
Chi-1 trans SD	(1)	100	101	104	103	107	97	97
	(2)	8.6	9.2	8.7	9.2	9	11.2	9.6
	(3)	-0.7	-0.6	-0.7	-0.6	-0.7	-0.3	-0.6
Chi-1 gauche (+) SD	(1)	143	138	150	134	139	143	148
	(2)	5.9	6.6	6.9	5.7	6.3	6.8	7.4
	(3)	-1	-0.9	-0.8	-1.1	-0.9	-0.9	-0.7
Chi-1 pooled SD	(1)	288	288	292	289	295	296	294
	(2)	7.7	7.6	7.8	7.3	7.4	8.5	8.2
	(3)	-0.8	-0.8	-0.8	-0.9	-0.8	-0.6	-0.7
Chi-2 trans SD	(1)	75	77	74	81	79	84	79
	(2)	9.3	8.7	9.7	9.8	8.7	10.6	10.8
	(3)	-1.3	-1.4	-1.2	-1.2	-1.4	-1.0	-1.0

[1] Parameter explanations in Table 1S.

Table 3S. Quality of main-chain and side-chain parameters of the modeled Fabaceae and Pinaceae ADH[1]

		Comparative values			
Stereochemical parameters		Fabaceae			Pinaceae
		LOTCOR	PISSAT	TRIREP	PINBAN
Main-chain					
% residues in A, B, L	(1)	321	321	321	320
	(2)	93.1	93.1	93.5	92.5
	(3)	0.4	0.4	0.4	0.3
Omega angle SD	(1)	379	378	378	374
	(2)	5.3	4.4	4.3	5.3
	(3)	-0.2	-0.5	-0.6	-0.2
Bad contacts / 100 residues	(1)	4	8	6	7
	(2)	1.1	2.1	1.6	1.9
	(3)	0	0.1	0.1	0.1
Zeta angle SD	(1)	342	343	343	338
	(2)	1.6	1.4	1.5	1.4
	(3)	-1	-1	-1	-1.1
H-bond energy SD	(1)	228	228	229	228
	(2)	0.7	0.7	0.7	0.7
	(3)	0.3	0.4	0.4	0.4
Overall G-factor	(1)	380	380	380	375
	(2)	-0.1	-0.1	-0.1	0.1
	(3)	-0.3	-0.2	-0.3	-0.2
Side-chain					
Chi-1 gauche (-) SD	(1)	52	47	53	43
	(2)	8.7	8.4	10	7.5
	(3)	-0.2	-0.3	0	-0.4
Chi-1 trans SD	(1)	100	99	112	107
	(2)	9.5	9.7	12.1	9.2
	(3)	-0.6	-0.5	-0.1	-0.6
Chi-1 gauche (+) SD	(1)	141	148	128	144
	(2)	8.4	6.7	8.3	6.4
	(3)	-0.5	-0.9	-0.5	-0.9
Chi-1 pooled SD	(1)	293	294	293	294
	(2)	8.8	8.1	10.1	7.7
	(3)	-0.5	-0.7	-0.3	-0.8
Chi-2 trans SD	(1)	72	86	74	87
	(2)	10.7	9.3	12.7	8.8
	(3)	-1	-1.3	-0.6	-1.4

[1] Parameter explanations in Table 1S.

Table 4S. Secondary structure of the amino acid residues identified as functionally divergent among the botanical families according to *Qk* ≥ 0.85 cut off values (h: helix; l: loop; s: strand)[1]

Comparison	Amino acid residue position	Amino acid residue and Secondary structure	
Brassicaceae vs. Pinaceae		in Brassicaceae	in Pinaceae
	315	1BRAOLE: h (Leu291) 2ARABLE: h (Phe311) 1ARAGRI: h (Leu311) 1ARAPAR: h (Phe311) 2LEAST: h (Phe311) 3LEAST: h (Phe312)	1PINBAN: h (Phe306)
	317	1BRAOLE: l (Asn293) 2ARABLE: l (Asn313) 1ARAGRI: l (Asn313) 1ARAPAR: l (Asn313) 2LEAST: l (Asn313) 3LEAST: l (Asn314)	1PINBAN: l (Cys308)
Poaceae vs. Fabaceae		in Poaceae	in Fabaceae
	118*	1HORVUL: l (Asp114) 2HORVUL: l (Asp114) 3HORVUL: l (Asp114) 1ORYSAT: l (Asp114) 2ORYSAT: l (Asp114) 1ZEAMAY: l (Asp123) 2ZEAMAY: l (Asp114)	1LOTCOR: l (Asp115) 1TRIREP: l (Asn115) 1PISSAT: l (Asp115)
	133*	1HORVUL: l (Gly129) 2HORVUL: l (Gly129) 3HORVUL: l (Gly129) 1ORYSAT: l (Gly129) 2ORYSAT: l (Gly129) 1ZEAMAY: l (Gly138) 2ZEAMAY: l (Gly129)	1LOTCOR: l (Asn130) 1TRIREP: l (Asn130) 1PISSAT: l (Asn130)
	236*	1HORVUL: h (Phe232) 2HORVUL: h (His232) 3HORVUL: h (Tyr232) 1ORYSAT: h (Phe232) 2ORYSAT: h (Phe232) 1ZEAMAY: h (Phe241) 2ZEAMAY: h (Tyr232)	1LOTCOR: h (Phe233) 1TRIREP: h (Phe233) 1PISSAT: h (Phe233)
Fabaceae vs. Pinaceae		in Fabaceae	in Pinaceae
	131*	1LOTCOR: l (Ser128) 1TRIREP: l (Asn128) 1PISSAT: l (Asn128)	1PINBAN: l (Ser122)
	133*	1LOTCOR: l (Asn130) 1TRIREP: l (Asn130) 1PISSAT: l (Asn130)	1PINBAN: l (Gly124)
	337*	1LOTCOR: h (Asn334) 1TRIREP: h (Asn334) 1PISSAT: h (Asn334)	1PINBAN: h (Gly328)

Table 4S. (Cont.)

Comparison	Amino acid residue position	Amino acid residue and Secondary structure	
Brassicaceae vs. Fabaceae		in Brassicaceae	in Fabaceae
	45	1BRAOLE: s (Phe23)	1LOTCOR: s (Tyr44)
		2ARABLE: s (Phe43)	1TRIREP: s (Phe44)
		1ARAGRI: s (Phe43)	1PISSAT: s (Phe44)
		1ARAPAR: s (Phe43)	
		2LEAST: s (Phe43)	
		3LEAST: s (Phe44)	
	57	1BRAOLE: h (Glu35)	1LOTCOR: h (Glu56)
		2ARABLE: h (Glu55)	1TRIREP: h (Glu56)
		1ARAGRI: h (Glu55)	1PISSAT: h (Glu56)
		1ARAPAR: h (Glu55)	
		2LEAST: h (Glu55)	
		3LEAST: h (Glu56)	
	82	1BRAOLE: s (Val58)	1LOTCOR: s (Val79)
		2ARABLE: s (Val78)	1TRIREP: s (Val79)
		1ARAGRI: s (Val78)	1PISSAT: s (Val79)
		1ARAPAR: s (Val78)	
		2LEAST: s (Val78)	
		3LEAST: s (Val79)	
	90	1BRAOLE: l (Gln66)	1LOTCOR: l (Lys87)
		2ARABLE: l (Ala86)	1TRIREP: l (Lys87)
		1ARAGRI: l (Gln86)	1PISSAT: l (Lys87)
		1ARAPAR: l (Gln86)	
		2LEAST: l (Lys86)	
		3LEAST: l (Gln87)	
	127	1BRAOLE: l (Gly103)	1LOTCOR: l (Gly124)
		2ARABLE: l (Gly123)	1TRIREP: l (Gly124)
		1ARAGRI: l (Gly123)	1PISSAT: l (Gly124)
		1ARAPAR: l (Gly123)	
		2LEAST: l (Arg123))	
		3LEAST: l (Gly124)	
	130	1BRAOLE: l (Ile106)	1LOTCOR: l (Ile127)
		2ARABLE: l (Ile126)	1TRIREP: l (Ile127)
		1ARAGRI: l (Ile126)	1PISSAT: l (Leu127)
		1ARAPAR: l (Ile126)	
		2LEAST: l (Ile126)	
		3LEAST: l (Ile127)	
	133*	1BRAOLE: l (Gly109)	1LOTCOR: l (Asn130)
		2ARABLE: l (Gly129)	1TRIREP: l (Asn130)
		1ARAGRI: l (Gly129)	1PISSAT: l (Asn130)
		1ARAPAR: l (Gly129)	
		2LEAST: l (Gly129)	
		3LEAST: l (Gly130)	
	135	1BRAOLE: l (Ser111)	1LOTCOR: l (Ser132)
		2ARABLE: l (Ser131)	1TRIREP: l (Ser132)
		1ARAGRI: l (Ser131)	1PISSAT: l (Ser132)
		1ARAPAR: l (Ser131)	
		2LEAST: l (Ser131)	
		3LEAST: l (Ser132)	
	139	1BRAOLE: s (Ile115)	1LOTCOR: s (Ile136)
		2ARABLE: s (Ile135)	1TRIREP: s (Ile136)
		1ARAGRI: s (Ile135)	1PISSAT: s (Ile136)
		1ARAPAR: s (Ile135)	
		2LEAST: s (Ile135)	
		3LEAST: s (Ile136)	
	188	1BRAOLE: h (Gly164)	1LOTCOR: h (Gly185)
		2ARABLE: h (Gly184)	1TRIREP: h (Gly185)
		1ARAGRI: h (Glu184)	1PISSAT: h (Gly185)
		1ARAPAR: h (Gly184)	
		2LEAST: h (Gly184)	
		3LEAST: h (Gly185)	

Table 4S. (Cont.)

Comparison	Amino acid residue position	Amino acid residue and Secondary structure	
Brassicaceae vs. Fabaceae		in Brassicaceae	in Fabaceae
	190	1BRAOLE: h (Thr166)	1LOTCOR: h (Thr187)
		2ARABLE: h (Thr186)	1TRIREP: h (Thr187)
		1ARAGRI: h (Thr186)	1PISSAT: h (Thr187)
		1ARAPAR: h (Thr186)	
		2LEAST: h (Ile186)	
		3LEAST: h (Thr187)	
	213	1BRAOLE: h (Ala189)	1LOTCOR: h (Ala210)
		2ARABLE: h (Ala209)	1TRIREP: h (Ala210)
		1ARAGRI: h (Ala209)	1PISSAT: h (Ala210)
		1ARAPAR: h (Ala209)	
		2LEAST: h (Ala209)	
		3LEAST: h (Gly210)	
	241	1BRAOLE: h (Glu217)	1LOTCOR: h (Lys238)
		2ARABLE: h (Lys237)	1TRIREP: h (Lys238)
		1ARAGRI: h (Glu217)	1PISSAT: h (Lys238)
		1ARAPAR: h (Lys237)	
		2LEAST: h (Lys237)	
		3LEAST: h (Lys238)	
	295	1BRAOLE: s (Val271)	1LOTCOR: s (Val292)
		2ARABLE: s (Val291)	1TRIREP: s (Val292)
		1ARAGRI: s (Val291)	1PISSAT: s (Val292)
		1ARAPAR: s (Val291)	
		2LEAST: s (Val291)	
		3LEAST: s (Val292)	
	303*	1BRAOLE: l (Ser279)	1LOTCOR: l (Asn300)
		2ARABLE: l (Ser299)	1TRIREP: l (Lys300)
		1ARAGRI: l (Ser299)	1PISSAT: l (Ser300)
		1ARAPAR: l (Ser299)	
		2LEAST: l (Ser299)	
		3LEAST: l (Ser300)	
	310*	1BRAOLE: l (Thr286)	1LOTCOR: l (Thr307)
		2ARABLE: l (Thr306)	1TRIREP: l (Thr307)
		1ARAGRI: l (Thr306)	1PISSAT: l (Thr307)
		1ARAPAR: l (Ser306)	
		2LEAST: l (Thr306)	
		3LEAST: l (Thr307)	
	311	1BRAOLE: l (His287)	1LOTCOR: l (His308)
		2ARABLE: l (His307)	1TRIREP: l (His308)
		1ARAGRI: l (His307)	1PISSAT: l (His308)
		1ARAPAR: l (His307)	
		2LEAST: l (His307)	
		3LEAST: l (His307)	
	315*	1BRAOLE: h (Leu291)	1LOTCOR: h (Phe312)
		2ARABLE: h (Phe311)	1TRIREP: h (Phe312)
		1ARAGRI: h (Leu311)	1PISSAT: h (Phe312)
		1ARAPAR: h (Phe311)	
		2LEAST: h (Phe311)	
		3LEAST: h (Phe312)	
	337*	1BRAOLE: h (Gly313)	1LOTCOR: h (Asn334)
		2ARABLE: h (Gly333)	1TRIREP: h (Asn334)
		1ARAGRI: h (Gly333)	1PISSAT: h (Asn334)
		1ARAPAR: h (Gly333)	
		2LEAST: h (Gly333)	
		3LEAST: h (Gly334)	
	344	1BRAOLE: h (Asn320)	1LOTCOR: h (Arg341)
		2ARABLE: h (Asn340)	1TRIREP: h (Lys341)
		1ARAGRI: h (Asn340)	1PISSAT: h (Lys341)
		1ARAPAR: h (Asn340)	
		2LEAST: h (Asn340)	
		3LEAST: h (Asn341)	

[1] Amino acid residues with Q(k) ≥ 0.90 are distinguished by an asterisk (*)

Table 5S. Secondary structure of the amino acid residues identified as functionally divergent between ADH forms according to *Qk* ≥ 0.85 cut off values (h: helix; l: loop; s: strand)[1]

Comparison	Amino acid residue position	Amino acid residue and Secondary structure	
Poaceae ADH1 vs. Poaceae ADH2		in Poaceae ADH1	in Poaceae ADH2
	25	1HORVUL: s (Thr23)	2HORVUL: s (Ser23)
		1ORYSAT: s (Val23)	2ORYSAT: s (Ser23)
		1ZEAMAY: s (Ser23)	2ZEAMAY: s (Ser23)
	64	1HORVUL: l (Met62)	2HORVUL: l (Val62)
		1ORYSAT: l (Val62)	2ORYSAT: l (Val62)
		1ZEAMAY: l (Val62)	2ZEAMAY: l (Val62)
	170	1HORVUL: l (Glu166)	2HORVUL: l (Glu166)
		1ORYSAT: l (Ala166)	2ORYSAT: l (Glu166)
		1ZEAMAY: l (Gln175)	2ZEAMAY: l (Glu166)
	234*	1HORVUL: h (Ser230)	2HORVUL: h (Ala230)
		1ORYSAT: h (Asn230)	2ORYSAT: h (Ala230)
		1ZEAMAY: h (Ser239)	2ZEAMAY: h (Ala230)
	263*	1HORVUL: h (Asp259)	2HORVUL: h (Glu259)
		1ORYSAT: h (Glu259)	2ORYSAT: h (Glu259)
		1ZEAMAY: h (Glu268)	2ZEAMAY: h (Glu259)
	329*	1HORVUL: l (Phe325)	2HORVUL: l (Tyr319)
		1ORYSAT: l (Tyr325)	2ORYSAT: l (Tyr325)
		1ZEAMAY: l (Tyr334)	2ZEAMAY: l (Tyr325)
Poaceae ADH1 vs. Fabaceae ADH1		in Poaceae ADH1	in Fabaceae ADH1
	64	1HORVUL: l (Met62)	1LOTCOR: l (Leu63)
		1ORYSAT: l (Val62)	1TRIREP: l (Leu63)
		1ZEAMAY: l (Val62)	1PISSAT: l (Leu63)
	263*	1HORVUL: h (Asp259)	1LOTCOR: h (Glu260)
		1ORYSAT: h (Glu259)	1TRIREP: h (Glu260)
		1ZEAMAY: h (Glu268)	1PISSAT: h (Glu260)
	329*	1HORVUL: l (Phe325)	1LOTCOR: l (Tyr326)
		1ORYSAT: l (Tyr325)	1TRIREP: l (Tyr326)
		1ZEAMAY: l (Tyr334)	1PISSAT: l (Tyr326)
Poaceae ADH2 vs. Fabaceae ADH1		in Poaceae ADH2	in Fabaceae ADH1
	41	2HORVUL: s (Asp39)	1LOTCOR: s (Leu40)
		2ORYSAT: s (Val39)	1TRIREP: s (Leu40)
		2ZEAMAY: s (Ile39)	1PISSAT: s (Leu40)
	133	2HORVUL: l (Gly129)	1LOTCOR: l (Asn130)
		2ORYSAT: l (Gly129)	1TRIREP: l (Asn130)
		2ZEAMAY: l (Gly129)	1PISSAT: l (Asn130)
	221	2HORVUL: h (Ser217)	1LOTCOR: h (Ser218)
		2ORYSAT: h (Ser217)	1TRIREP: h (Ser218)
		2ZEAMAY: h (Ala217)	1PISSAT: h (Ser218)
	236	2HORVUL: h (His232)	1LOTCOR: h (Phe233)
		2ORYSAT: h (Phe232)	1TRIREP: h (Phe233)
		2ZEAMAY: h (Tyr232)	1PISSAT: h (Phe233)
	279	2HORVUL: h (Ala275)	1LOTCOR: h (Ile276)
		2ORYSAT: h (Ile275)	1TRIREP: h (Ile276)
		2ZEAMAY: h (Val275)	1PISSAT: h (Ile276)
	285	2HORVUL: h (Ala281)	1LOTCOR: h (Ala282)
		2ORYSAT: h (Cys281)	1TRIREP: h (Ala282)
		2ZEAMAY: h (Ala281)	1PISSAT: h (Ala282)

[1] Amino acid residues with Q(k) ≥ 0.90 are distinguished by an asterisk (*)

Fig. 1S. 3D_1D averaged scores, as determined by the VERIFY_3D program for the models of the Brassicaceae botanical family.

Fig. 2S. 3D_1D averaged scores, as determined by the VERIFY_3D program for the models of the Poaceae botanical family.

Fig. 3S. 3D_1D averaged scores, as determined by the VERIFY_3D program for the models of the Fabaceae and Pinaceae botanical families.

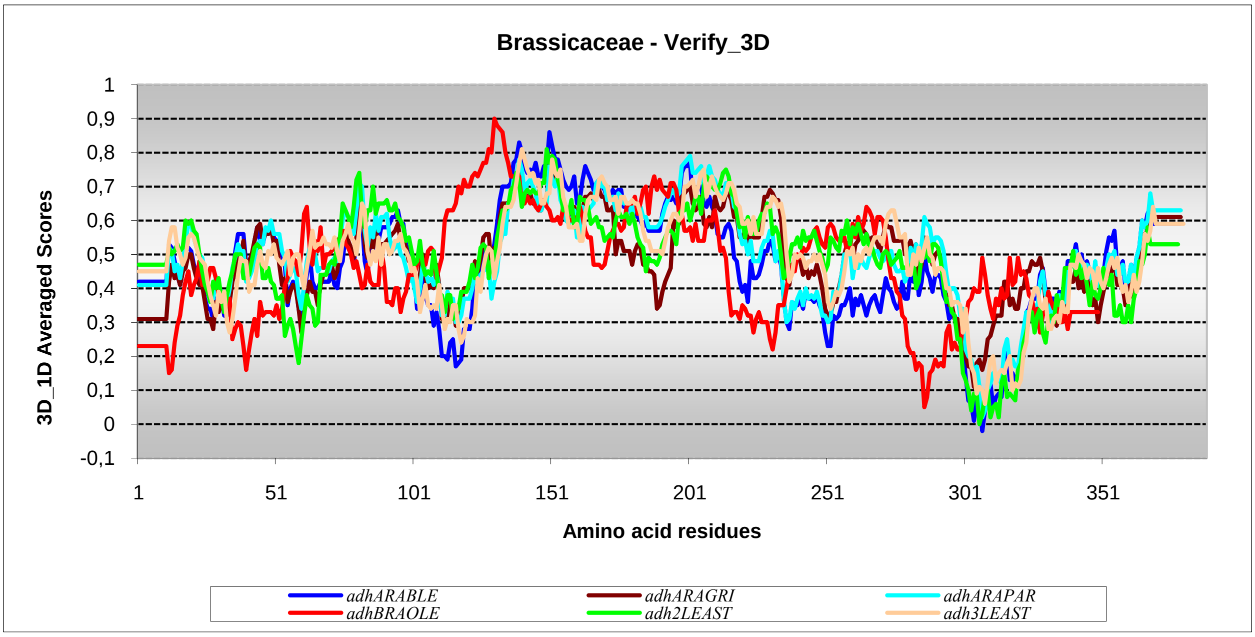

Brassicaceae - Verify_3D
3D_1D Averaged Scores
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
1
51
101
151
201
251
301
351
Amino acid residues
adhARABLE
adhARAGRI
adhARAPAR
adhBRAOLE
adh2LEAST
adh3LEAST

Fig. 1S

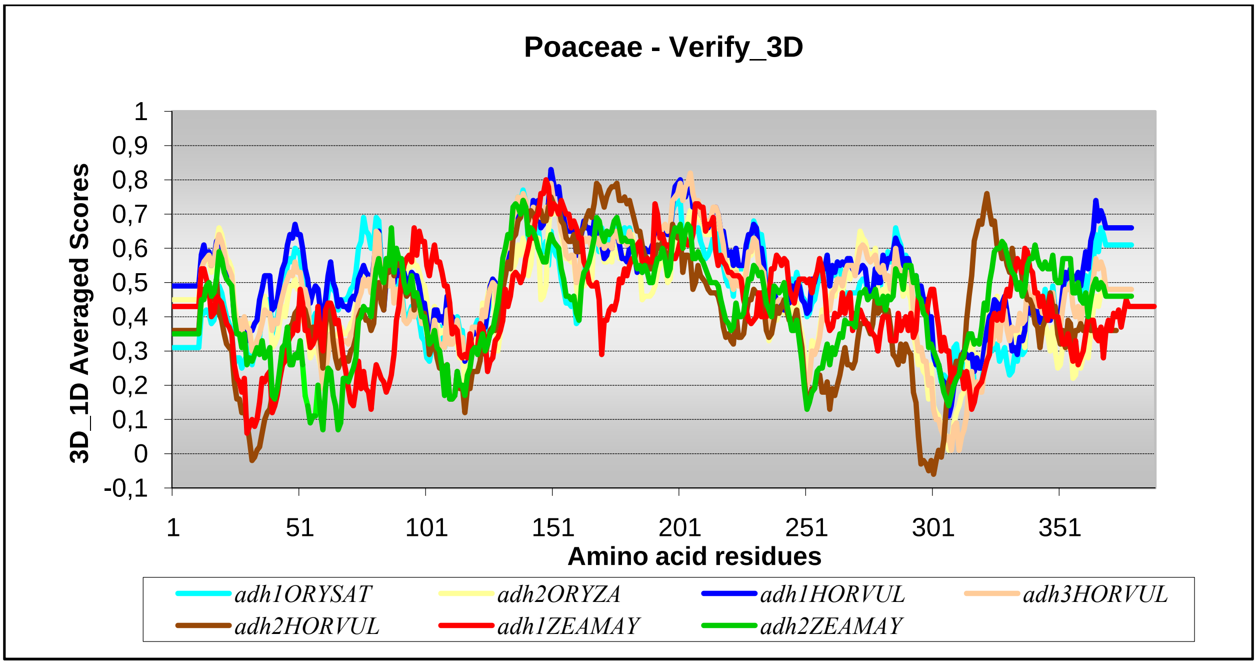

Poaceae - Verify_3D
3D_1D Averaged Scores
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
1
51
101
151
201
251
301
351
Amino acid residues
adh1ORYSAT
adh2ORYZA
adh1HORVUL
adh3HORVUL
adh2HORVUL
adh1ZEAMAY
adh2ZEAMAY

Fig. 2S

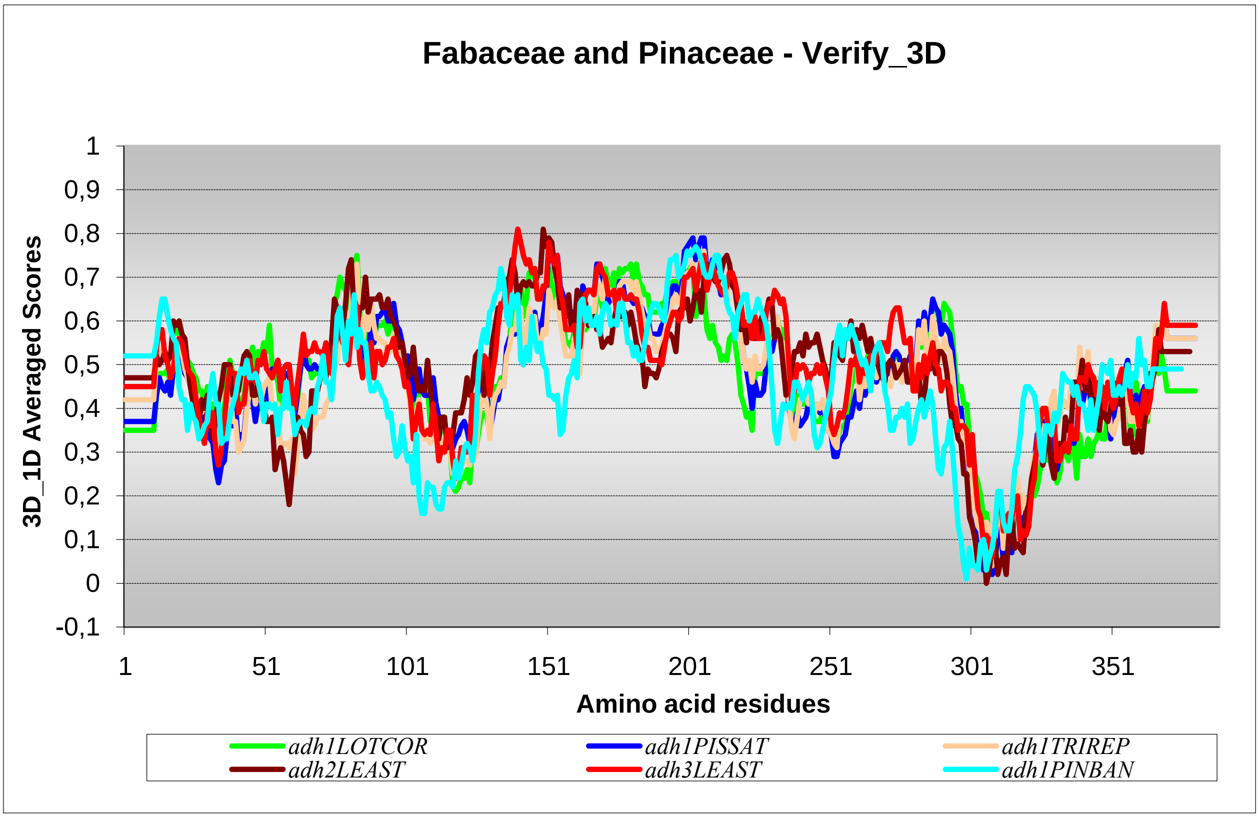

Fabaceae and Pinaceae - Verify_3D
3D_1D Averaged Scores
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
1
51
101
151
201
251
301
351
Amino acid residues
adh1LOTCOR
adh1PISSAT
adh1TRIREP
adh2LEAST
adh3LEAST
adh1PINBAN

Fig. 3S

CAPÍTULO 5

Artigo 3

A ser submetido a Journal of Molecular Evolution (ISSN: 0022-2844)

Molecular Evolution and Functional Divergence of Alcohol Dehydrogenases in Animals, Fungi and Plants

Claudia E. Thompson · Loreta B. Freitas · Francisco M. Salzano

C. E. Thompson (🖂) · L. B. Freitas · F. M. Salzano

Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970, Porto Alegre, RS, Brazil

🖂 Corresponding author. Tel.: 55 51 3217-4182; Fax: 55 51 3308-9823; *E-mail address:* claudia.thompson@ufrgs.br

Abstract

The alcohol dehydrogenase enzyme belongs to the large superfamily of medium-chain dehydrogenases/reductases, which have been characterized in animals, fungi, plants, protozoan, and bacteria. They are involved, in different organisms, with the ethanol, norepinephrine, dopamine, serotonin, bile acid metabolism, and other important metabolic routes. However, the diversification process which occurred in these substances in the course of evolution is not well known. In the present report we considered the phylogeny of 192 sequences of animals, fungi and plants. They formed distinct clusters. Non-class III *Caenorhabditis elegans* are closely related to the tetramer fungi ADH type. Within chordates, duplications lead to a multiplicity of ADHs forms, ADH3 being basal in Fishes, Amphibia, Reptilia, Aves, and Mammalia groups. In fishes, two main clusters are identified: ADH1 and ADH3. Amphibians present the highest ADH diversity. ADH2 is found in Mammalia and Aves. Additionally, ADH4 and ADH5 seem to result from ADH1 duplication. A more complex pattern was identified in Fungi, where ADH formed clusters based on types and genera. The plant results were published in Gene (v. 396, p. 108-115, 2007). Amino acid residues responsible for functional divergence between ADH types and among fungi, plant and animals were identified. For mammals sites 209 and 264, near the enzymes' active center, seem to be especially important. Functional site differences occur mainly between ADH1/ADH4 and ADH2/ADH3. Site no. 301 is important in birds ADH1/ADH3 and fungi ADH3/ADH5. Within fungi site 373 differs markedly between their ADH3/ADH5 and ADH4/ADH5 forms. Both sites occur in the enzyme's subunit-subunit interacting segment. The ADH family expansion exemplifies a neofunctionalization process with reiterative duplication events leading to new activities.

Keywords Glycolytic proteins; Molecular evolution; Alcohol dehydrogenase, Functional diversification; Molecular modeling

Introduction

The alcohol dehydrogenase (ADH, EC 1.1.1.1) enzyme belongs to the large superfamily of medium-chain dehydrogenases/reductases, which comprise different enzyme activities, such as alcohol, sorbitol, xylitol, threonine dehydrogenase, and quinone reductase (Persson et al. 1993). Its activity appears to be universal in all types of life forms, derived from enzymes of separate family assignments, and frequently of multiple occurrences in a complex fashion (Norin et al. 1997).

Class III ADH, with little or almost no ethanol activity, and similar to the glutathione-dependent formaldehyde dehydrogenase, seems to be an ancestral form and has been characterized in vertebrates (Jörnvall et al. 1995; Hjelmqvist et al. 1995b), invertebrates (Kaiser et al. 1993; Danielsson et al. 1994), plants (Martínez et al. 1996), fungi (Sasnaukas et al. 1992; Fernández et al. 1995), and prokaryotes (Gutheil et al. 1992; Ras et al. 1995). Class III ADH functions as a glutathione-dependent dehydrogenase in the oxidative elimination of formaldehyde, and do not function in ethanol or retinol oxidation, which has been assumed by other ADH classes (Duester et al. 1999). It is considered to be vertebrate ADH most ancient form due to the fact that it is the only one detected in invertebrates (Kaiser et al. 1993).

Vertebrate ADH is a cytosolic, dimeric, zinc-containing, NAD-dependent enzyme with a subunit molecular masss of 40 kDa. At least eight distinct classes have been identified in this group, based upon sequence alignment, phylogenetic analysis, catalytic properties, and gene expression patterns. Within the same organism, ADH classes share around 60% amino acid sequence identity, and multiple ADH isoenzymes within a single class share above 90% identity (Jörnvall 2008). They metabolize a wide variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products (Duester et al. 1999).

In humans, only ADH classes I (with three isoforms: A, B, and C, earlier called α, β, and γ, respectively), II, III, IV, and V have been identified, and in mouse classes I, II, III, and IV have been described (Boleda et al. 1993; Zheng et al. 1993; Höög and Brandt 1995; Höög et al. 2001). Class VI ADH has been observed in the rat and deer mouse (Zgombic-Knight et al. 1995), and class VII ADH has been found in chicken (Kedishvili et al. 1997), which may act as a steroid/retinoid dehydrogrenase. An amphibian ADH class VIII (class IV-like) has specificity towards NADP(H), with high catalytic efficiency specificity for retinoids and high K_m for ethanol (Rosell et al. 2003).

Some classes originally defined for separate species may not be true classes, but simply species variants in rapidly evolving classes. This seems to be the case in vertebrate ADH classes V and VI. Another problem is the different nomenclature given for genes and protein classes, such as ADH classes II-V, which have a very distinct gene number designation.

Several fungal and bacterial ADH enzymes are tetramers with two zinc atoms per monomer, while the animal and plant ADHs characterized to date are thought to be dimeric with also two Zn atoms. In *Saccharomyces cerevisiae* and *Kluyveromyces* five distinct ADHs were found. ADH classes I and II of *S. cerevisiae* are cytoplasmic enzymes expressed under fermentative and respiratory conditions. Class III corresponds to a mitochondrial protein. Class IV is distantly related to the other four ADHs and is probably originated from a bacterium (Williamson and Paquin 1987). Finally, class V was discovered during the *S. cerevisiae* genome sequencing. The function of fungi classes III, IV, and V is not completely understood (Wills and Jörnvall 1979; Young et al 2000; Ladrière et al. 2000; Thomson et al. 2005).

In plants, the alcohol dehydrogenase gene family has been intensively studied in order to understand its genetics and molecular evolution. Generally this family is characterized by a small number of copies and very diverse expression patterns. ADHs are involved in the energy production pathway, converting the acetaldehyde into ethanol via fermentation during episodes of low oxygen or low temperatures. Despite the large number of studies, there does not exist a clear correlation among their molecular evolution, function and structure. Thompson et al. (2007) proposed that functional diversification during evolution has been responsible for site-specific shifts after ADH gene duplication in plants, and furnished the first three-dimensional model of a plant ADH. Subsequently, they evaluated the impact of functional divergence on Poaceae, Brassicaceae, Fabaceae, and Pinaceae enzymes (unpublished data).

In this article, the relationship among the different classes of ADH belonging to animals, fungi and plants were investigated and the amino acid residues crucial for different types of functional divergence between duplicate genes identified using evolutionary and modeling tools to better understand the ADH diversification process.

Materials and Methods

Source of the Data and Sequence Alignment

ADH amino acid sequences from Phylum Chordata (Classes Myxini, Actinopterygii, Elasmobranchii, Sarcopterygii, Amphibia, Reptilia, Aves, and Mammalia), Phylum Mollusca (Class Cephalopoda), Phylum Nematoda (Class Chromadorea), Phylum Platyhelminthes (Class Turbellaria), and Phylum Ascomycota (Classes Saccharomycetes, Sordariomycetes, and Eurotiomycetes) were obtained from the National Center of Biotechnology Information (NCBI). Plant amino acid sequences used in our previous studies (Thompson et al. 2007) were incorporated to the analysis. Alcohol dehydrogenase DNA sequences were also downloaded from the NCBI site to evaluate the presence of positive selection. Protein alignments were performed using the ClustalW multiple sequence alignment program (Thompson et al. 1994; Jeanmougin et al. 1998) with the BLOSUM matrix (Henikoff and Henikoff 1992) for scoring. The penalties for gap opening and gap extension were 10.0 and 0.2. GeneDoc 2.6 (Multiple Sequence Alignment Editor & Shading Utility; Nicholas and Nicholas 1997) was employed to align DNA sequences based on their corresponding manually adjusted protein alignment. Alignments are available upon request.

Phylogenetic analysis

Phylogenies were estimated by neighbor-joining (NJ) (Saitou and Nei 1987), available in the MEGA program version 4.1 (Molecular Evolutionary Genetics Analysis; Kumar et al. 2000, 2008; Tamura et al. 2007), and by maximum likelihood (ML) methods using PhyML (Phylogenetic Maximum Likelihood; Guindon and Gascuel 2003) and TreeFinder (Jobb et al. 2004) programs. All the analyses were performed for (1) Actinopterygii + Elasmobranchii + Sarcopterygii + Myxini, Amphibia, Reptilia, Aves, Mammalia, and Fungi separately; (2) the combined vertebrate and invertebrate sequences; and (3) the combined animals, fungi and plants sequences.

The p-distance, the Poisson-corrected amino acid distances, and the complete and pairwise deletion of gaps/missing data were implemented to analyze the amino acid sequences in the neighbor joining (NJ) method, with 2,000 repetitions performed using the bootstrap test of phylogeny.

The selection of the best-fit model of amino acid substitution for the ML analysis was performed with ProtTest 1.4 program (Abascal et al. 2005) using the slow strategy (optimization of model, branches and topology of the tree) and without restricting the set of protein evolution candidate models for the uncombined data. For the combined data a fast strategy and a restriction of the set of candidate models was applied to decrease computer processing time. The protein empirical matrices which were the best candidates for the uncombined data formed the restricted candidates group applied to the combined data. A BIONJ tree was calculated, which is a distance based on a phylogeny reconstruction algorithm with better topological accuracy than NJ in all evolutionary conditions (Gascuel 1997). The BIONJ algorithm is an alternative version of NJ, where long genetic distances present a higher variance than short ones when distances from a newly defined node to all other nodes are estimated (Van der Peer 2003). All statistical frameworks (AIC, AICc or BIC) for the selection of the best model were comparatively evaluated considering the relative importance and the model-averaged estimate of parameters. AICc (Corrected Akaike Information Criterion) (Posada and Crandall 2001; Burnham and Anderson 2003) and BIC (Bayesian Information Criterion) (Schwarz 1978) include penalties for sample size.

The PhyML and TreeFinder programs performed the analyses using the trees and the best models of protein sequence evolution which resulted from ProtTest 1.4. In

TreeFinder, an approximate bootstrap support was computed by using the Expected Likelihood-Weights (Strimmer and Rambaut 2002) applied to all local rearrangements of the tree topology (LR-ELW). This edge support value, in percent, can be directly interpreted as confidence in the configuration of branches adjacent to a particular edge. An approximate likelihood-ratio test (aLRT) for branches were used in PhyML. This approach is based on the conventional LRT principle. However, it is a faster test since the log-likelihood value l_2 is computed by optimizing over the branch of interest and the four adjacent branches, whereas other parameters are fixed at their optimal values corresponding to the best ML tree (Anisimova and Gascuel 2006).

Selection and Functional Diversification Analysis

The resultant tree topologies were used to calculate branch lengths using the M0 model available in the CODEML program of the PAML packet (Yang 1997). The presence of positive selection was verified through the maximum likelihood models recommended by Yang (2004) using alcohol dehydrogenase DNA sequences. A series of likelihood ratio tests (LRTs) was carried out to investigate whether ω was significantly different from 1 for each pairwise comparison: M1a *vs*. M2a, M0 *vs*. M3, and M7 *vs*. M8. LRT performs the comparison both with the constraint of ω=1 and without such constraint: $LR=2(ln_1-ln_2)$. These LRT statistics approximately follow a chi-square distribution and the number of degrees of freedom is equal to the number of additional parameters in the more complex model (Anisimova et al. 2001, 2002).

Type-I and Type-II functional divergences were examined using a statistical framework model implemented by the DIVERGE2 program (Gu and Vander Velden 2002;

Gu 2006), which determines if the coefficients of divergence (θ_I and θ_{II}) are significantly greater than zero. Type-I functional divergence (site-specific rate shift) refers to the evolutionary process resulting in site-specific rate shifts after gene duplication. It identifies amino acid residues highly conserved in one gene copy and highly variable in the other. Type-II divergence results in a site-specific property shift. It considers if a radical shift of amino acid property (positively vs. negatively charged; hydrophobic vs. hydrophilic) between the proteins of the duplicate genes occurs. An amino acid substitution will be radical if it changes the residue from one of the previous cited groups to another. The probability of a residue being under Type-II divergence is denoted θ_{II}. Q_I *(k)* or Q_{II} *(k)* are the site *(k)*-specific scores corresponding to the posterior probability that site *k* is related to type-I or type-II functional divergences (Zheng et al. 2007).

Results and Discussion

Phylogenetic analysis

Gene duplication is an important precursor of evolutionary diversification. The majority of new genes originate through duplication, chromosomal rearrangement, and the subsequent divergence of pre-existing genes (Lawton-Rauth 2003). The existence of several multigene families is an indication of the importance of gene duplication in the origin of new function novelties (Wendel 2000). Phylogenetic analysis has been a powerful approach to investigate the role of gene duplications in evolution.

The alcohol dehydrogenase enzymes form a large and diverse family which has contributed to the understanding of protein evolution, enzymatic mechanisms, metabolic functions, and regulatory roles. They show chemically modified sub-forms, isoenzymes, classes, and separate enzymes, presenting a wide range of distinct functions as well as redundancy with overlaps in activity (Jörnvall 2008). Recently we have theoretically demonstrated that different plant ADH forms may be submitted to an evolutionary diversification process which occurred after gene duplication (Thompson et al. 2007). The next step was to evaluate the importance of this process in ADHs of other organisms, to obtain a comprehensive panorama for ADH molecular evolution.

In total, 192 ADH amino acid sequences belonging to animals, fungi, and plants were submitted to a comparative phylogenetic approach. The taxonomic classification, ADH types, accession numbers, and sequence sizes are shown in Table 1S (Electronic Supplementary Material).

The tree topologies resulting from the NJ and ML methods did not differ significantly, especially when major clades are considered. Figure 1 shows that obtained with the NJ approach. Three monophyletic groups corresponding to fungi, plants, and a larger group formed by animals, with fungi ADH closer to plant enzymes were obtained. A similar pattern was found by Glasner et al. (2005) who analyzed a smaller number of sequences (22). Two *Caenorhabditis elegans* sequences (ADH1 and ADH2) were placed closer to the tetrameric fungal ADHs (Fig. 1). This result agrees with that obtained by Glasner et al. (1995) who described for the first time fungal-like ADH sequences among metazoans. Both ADH *C. elegans* forms show ethanol activity, preferentially for longer alcohols. It may be possible that additional fungal-like sequences will be discovered in other animals or plants. This phenomenon could be explained by one or multiple deletions in lineages generating the modern plants and animals or it may be the result of convergent evolution (Glasner et al. 1995).

It is interesting to note that *C. elegans* ADH3 clusters with those of *Octopus vulgaris* and *Schmidtea mediterranea* (a freshwater planarian), within the large group of ADH3 from all animals. Godoy et al. (2007) also found a close relationship between the *S. mediterranea* and *C. elegans* ADH3s. Kaiser et al. (1993) described the *O. vulgaris* ADH3, which was the only ADH form detectable in this animal. They were the first detected group of animals that lack ethanol dehydrogenase activity. No other ADH class is present in planarians also, as suggested by *in silico* analysis which indicated that only one contig was sufficient to account for the cDNA and 40 trace sequences from the current planarian databases (Godoy et al. 2007). The invertebrate ADH3s formed a highly supported monophyletic group. In relation to chordate ADHs, all ADH3 also formed a monophyletic cluster (Fig. 1), as seen in other studies which considered a smaller number of species and sequences (Dasmahapatra et al. 2001; Reimers et al. 2004; Gonzàlez-Duarte and Albalat 2005). This enzyme is widely known as a glutathione-dependent formaldehyde dehydrogenase which can oxidize ethanol at high concentrations (Dasmahapatra et al. 2001), and preferentially metabolize longer aliphatic and aromatic alcohols (Reimers et al. 2004). ADH3 has been described as a ubiquitous enzyme in vertebrates (Funkenstein and Jakowlew 1996), with a spatio-temporal regulation in zebrafish development (Dasmahapatra et al. 2001; Cañestro et al. 2003). Additionally, ADH3 is found in the cell nucleus, with a probable DNA protection function (Iborra et al. 1992; Fernández et al. 2003), differently from the other ADHs, which commonly have a cytosolic location (Gonzàlez-Duarte and Albalat 2005). In invertebrates, its expression is mainly found in digestive tissues (Godoy et al. 2007).

Most of the ADH1s are located in a large set with high bootstrap support for the individual clusters within the considered group (Fig. 1). This form is the classical and highly variable liver enzyme responsible for ethanol metabolism. In fishes, only two ADH classes were detected: class III and a second mixed class (here named ADH1, but also called ADH8 in the literature), structurally similar to class III, but functionally similar to ADH1 (the classical alcohol-metabolzing enzyme) (Dasmahapatra et al. 2005). This hybrid characteristic may explain why the Actinopterygii ADH1 cluster separately from all other class I forms (Fig. 1). Mammal ADH4s are highly similar to ADH1 in terms of primary sequence, and are placed near them in the phylogenetic tree (Figs. 1 and 2). Our results corroborated Gonzàlez-Duarte and Albalat's (2005) hypothesis that ADH4 may be the result of mammalian-specific *Adh1* duplication, since this class has not been detected in avians or reptilians and the amphibian ADH4 does not seem to be orthologous to the mammalian form (Fig. 1). ADH4 functions in retinoid oxidation *in vitro* (Boleda et al. 1993). However, ADH4-null mutant mice showed weak phenotypic effects, which may indicate a contribution in specific routes, not involved in systemic retinol metabolism (Deltour et al. 1999).

Class II are found in mammalian and avian/reptilian lineages forming a sister group to tetrapod nonclass III proteins (Fig. 1), reinforcing the results of Hjelmqvist et al. (1995b). Based on the phylogenetic analysis, as well as biochemical and structural characteristics (Höög et al. 2001, Gonzàlez-Duarte and Albalat 2005), it is reasonable to suggest that ADH2 is derived from a tetrapod ADH3. ADH2 proteins have higher K_m values toward ethanol and preferentially metabolize larger aliphatic and aromatic alcohols/aldehydes (Reimers et al. 2004). Moreover, they are structurally more divergent than the ADH1 forms, classically known as variable (Hjelmqvist et al. 1995a).

Amphibian ADH8 is positioned as a distinct cluster in the phylogenetic analysis (Fig. 1), confirming its special characteristics such as (i) a large active site pocket; (ii) the probably different proton-relay pathway, (iii) very specific rearrangements in the phosphate-binding site cofactor; and (iv) the weak interactions of the adenine moiety (Rosell et al. 2003). This form has a unique NADP(H) specificity and was first described as ADH4-like. However, its characteristics led to classification in a new class.

Phylogenetic relationships among mammal ADH sequences are displayed in more detail in Fig. 2, where it can be seen that monophyletic groups were formed according to ADH type. ADH1 has sub-clusters (ADH1A, ADH1B, ADH1C), corresponding to different isoenzymes. Both in Figs. 1 and 2 ADH4 is placed close to ADH1, suggesting that it originated from a ADH1 duplication. Similar results were obtained by Estonius et al. (1994), Parés et al. (1994) and Strömberg and Höög (2000). In mammals ADH4 is specifically expressed in epithelial tissues, as stomach mucosa. ADH5 is located between ADH1-ADH4 and ADH2 in the tree of Fig. 2. Its function is not well-understood yet. ADH3 forms a basal clade.

Clusters corresponding to ADH1 and ADH3 are formed in the basal Chordate Classes Myxini, Actinopterygii, Elasmobranchii, and Sarcopterygii (Fig. 1S). The ethanol-active class is a mixed type, structurally similar to class III but functionally similar to class I. Fishes constitute the first vertebrate class with documented expression of more than one ADH class (Dasmahapatra et al. 2005).

The greatest ADH diversity is found in amphibians (Fig. 2S). As in other animals, ADH4 shows highest similarity with other amphibian ADH1. Additionally, a new form (ADH8) appears in these animals and ADH3 forms a basal clade. Few reptile ADH sequences were found, ADH3 forming a distinguishable group from ADH1 (Fig. 3S). In addition to ADH1 and ADH3, ADH2 appears in the mammalian (Fig. 2), and avian (Fig. 4S) lineages; it is assumed that it derived from ADH1. ADH2 is basal in relation to ADH1 in both mammals and birds, and ADH3 is basal to all sequences in the two groups.

A more complex duplication pattern is seen in fungi (Fig. 3), where the ADH sequences cluster according to ADH type and fungi genera. A larger cluster composed by Saccharomycetes sequences is distinguishable. Additionally, Sordariomycetes and Eurotiomycetes ADHs formed distinct monophyletic groups. Sequences from *Saccharomyces* were grouped by ADH type, with ADH2 closer to the ADH1. The ADH1-ADH2 duplication seems to have occurred before the divergence of the *Saccharomyces* species and after the divergence between *Saccharomyces* and *Kluyveromyces*, which has been estimated to have occurred 80 ± 15 million years ago (Thomson et al. 2005). *Saccharomyces* ADHs 1, 2, and 5 probably derived from a common ancestor, as also suggested by Ladrière et al. (2000). Moreover, ADH5 has the highest rate of sequence divergence. ADH1 and ADH2 forms from *Lachancea* grouped together, ADH4 from *Kluyveromyces* formed a different group, as well as *Saccharomyces* ADH3 and ADH5. The *Yarrowia* sequences also separate by type. *Pichia* and *Candida* ADHs did not form a monophyletic cluster, while ADH3 from *Kluyveromyces* and *Lachancea* cluster together.

Saccharomyces ADH1 and ADH2 are cytoplasmatic enzymes acting in the fermentation and gluconeogenesis processes respectively, while ADH3 is located in the mitochondria. *Kluyveromyces* ADH has two cytoplasmatic (ADHs 1 and 2) and two mitochondrial (ADH3 and ADH4) enzymes. A single gene is sufficient to allow *Kluyveromyces lactis* to grow in ethanol. Since *Saccharomyces* and *Kluyveromyces* genomes are similar but their ADH sequences are submitted to different rates of divergence (Ladrière et al. 2000), they may have a lower structural constraint or are being submitted to a functional divergence process that would lead to new enzyme functions, as seen to occur in animals (Höög et al. 2001) and was theoretically demonstrated in plants (Thompson et al. 2007).

Selection and Functional Diversification Analyses

Natural selection has been described as responsible for the evolution of many genes (Hey 1999). A widely used method to detect positive selection is through the ratio of nonsynonymous to synonymous rates (ω = d_N/d_S). It is assumed that synonymous substitutions are neutral, while the nonsynonymous are influenced by selection. Consequently, a ω statistically higher than one indicates the action of positive selection or a relaxed selective constraint and low values for d_N/d_S ratio mean conservation of the gene product due to selective constraints (Tennessen 2008). No indication of positive selection acting on *Adh* genes were found (data not shown, available upon request). However, a relationship between a statistically detectable selection ($\omega>1$) and functional divergence might not necessarily exist (Tennessen 2008). Thus, to further investigate if any amino acid replacement could have lead to adaptative functional diversification, type-I and type-II divergences were estimated by posterior analysis using DIVERGE2 which evaluates shifted evolutionary rates and altered amino acid properties after gene duplication.

Coefficients of functional divergence (θ) of pairwise comparisons between mammalian and fungi alcohol dehydrogenases are reported in Table 1. They show statistically significant site-specific shift of evolutionary rates, with θ varying markedly from 0.0370 to 0.9704. A site-specific profile based on the posterior probability (Q_k) was used to identify the amino acid residues responsible for functional divergences after gene duplication or speciation. To reduce false positives, a conservative cut-off value was empirically used: $Q_k \geq 0.80$. Amino acid residue positions identified as functionally important between the mammalian ADH forms with their respective Q_k values are shown in Table 2, while those important for the differentiation between fish and fungi forms are listed in Table 3.

For mammals (Table 1) two sites (nos. 209 and 264) seem to be especially important for the differences between ADH3/ADH2 and ADH3/ADH5. Both are near the enzymes' active sites and therefore should have a functional role. The number of important site functional differences occur mainly between ADH4 and ADH1 (19 sites) and ADH3/ADH2 (18 sites). As for the fishes and fungi results (Table 2), site no. 301 is important for the differences between fishes ADH3/ADH1 and fungi ADH3/ADH5. Within fungi site 373 differs in an important way between their ADH3/ADH5 and ADH4/ADH5 forms. Both sites occur in the subunit-subunit interacting segment. The largest number of site differences (14) occurs in the ADH3/ADH5 comparison.

Experimental structural-functional studies are mainly restricted to the ADH1 and ADH3 enzymes. In ADH1, the classical liver ethanol dehydrogenase binding of the coenzyme induces the catalytic domain to approach the coenzyme-binding domain and to narrow the active site cleft. The two domain conformations are thus described as 'open' in the apoenzyme, and 'closed' in the binary and ternary complexes. These conformations could account for the different substrate specificity and kinetic mechanisms of ADH1 and ADH3. They are consistent with the ordered kinetics of ADH1, while the random mechanism of ADH3 is coherent with its 'semiopen' domain conformation (Sanghani et al 2000).

The proton relay pathway is also significantly different in the two classes. In class I, the components are the Thr/Ser48 that hydrogen binds with the alcohol hydroxyl group, the hydroxyl groups of nicotinamide ribose, and His51, a general base in contact with the solvent. However, in ADH3 His51 is not found, suggesting that proton transfer proceeds directly to the solvent (Sanghani et al 2002). Besides Thr/Ser48, class I enzymes share three strictly conserved positions, His67, Glu68 and Phe140. This triad has been proposed as a signature for class assignment (Norin et al 1997), although preservation of these positions does not necessarily imply ethanol oxidizing activity (Reimers et al 2004). In ADH1 three segments stand out as variable. They lie near the substrate-binding pocket and participate in the subunit interactions. In contrast, these regions are among the most conserved ADH3 segments (Cañestro et al 2003). In addition to ethanol oxidation, ADH1 has been implicated in other physiological pathways, for example, norepinephrine, dopamine, serotonin and bile acid metabolism (Höög et al 2001), and it can catalyze the oxidation of retinol *in vitro* (Boleda et al 1993) and *in vivo* (Deltour et al 1999). However, analysis of *Adh1*-null mutant mice challenged a major role in retinol metabolism and, instead, suggested a protective function against vitamin A toxicity. The *Adh1* gene is expressed at a very high level in the liver and also at a significant degree in the small and large intestine, kidney, adrenal, testis, epididymis, and uterus. It has tissue-specific expression, in contrast to the widespread distribution of *Adh3* in vertebrates (Gonzàlez-Duarte e Albalat, 2005).

Acknowledgements

This work was supported by the Institutos do Milênio and Programa de Apoio a Núcleos de Excelência Projects, Conselho Nacional de Desenvolvimento Científico e Tecnológico, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul, and Pró-Reitoria de Pesquisa da Universidade Federal do Rio Grande do Sul.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinform. 21: 2104-2105.

Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55: 539-552.

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio testing in detecting adaptative molecular evolution. Mol. Biol. Evol. 18: 1585-1592.

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol. Biol. Evol. 19: 950-958.

Boleda MD, Saubi N, Farrés J, Parés X (1993) Physiological substrates for rat alcohol dehydrogenase classes: aldehydes of lipid peroxidation, omega-hydroxyfatty acids, and retinoids. Arch. Biochem. Biophys. 307: 85-90.

Burnham KP, Anderson DR (2003) Multimodel inference: understanding AIC and BIC in model selection. Sociol. Method. Res. 33: 261-304.

Cañestro C, Godoy L, Gonzàlez-Duarte R, Albalat R (2003) Comparative expression analysis of *Adh3* during arthropod, urochordate, cephalochordate and vertebrate development challenges its predicted housekeeping role. Evol. Dev. 5: 157-162.

Danielsson O, Atrian S, Luque T, Hjelmqvist L, Gonzalez-Duarte R, Jörnvall H (1994) Fundamental molecular differences between alcohol dehydrogenase classes. Proc. Natl. Acad. Sci. USA 91: 4980-4984.

Dasmahapatra AK, Doucet HL, Bhattacharyya C, Carvan MJ (2001) Developmental expression of alcohol dehydrogenase (ADH3) in zebrafish (*Danio rerio*). Biochem. Biophys. Res. Commun. 286: 1082-1086.

Dasmahapatra AK, Wang X, Haasch ML (2005) Expression of *Adh8* mRNA is developmentally regulated in japanese medaka (*Oryzias latipes*). Comp. Biochem. Physiol. 140: 657-664.

Deltour L, Foglio Mh, Duester G (1999) Impaired retinol utilization in *Adh4* alcohol dehydrogenase mutant mice. Dev. Genet. 25: 1-10.

Duester G, Farrés J, Felder, MR, Holmes RS, Höög JO, Parés X, Plapp BV, Yin SJ, Jörnvall H (1999) Recommended nomenclature for the vertebrate alcohol dehydrogenase gene family. Biochem. Pharmacology 58: 389 – 395.

Estonius M, Hjelmqvist L, Jörnvall H (1994) Diversity of vertebrate class I alcohol dehydrogenase: mammalian and non-mammalian enzyme functions correlated through the structure f a ratite enzyme. Eur. J. Biochem. 224: 373-378.

Fernández MR, Biosca JA, Norin A, Jörnvall H, Parés X (1995) Class III alcohol dehydrogenase from *Saccharomyces cerevisae*: structural and enzymatic features differ toward the human/mammalian forms in a manner consistent with functional needs in formaldehyde detoxication. FEBS Lett. 370: 23 – 26.

Fernández MR, Biosca JA, Parés X (2003) S-nitrosoglutathione reductase activity of human and yeast glutathione-dependent formaldehyde dehydrogenase and its nuclear and cytoplasmic localization. Cell. Mol. Life Sci. 60: 1013-1018.

Funkenstein B, Jakowlew SB (1996) Molecular cloning of fish alcohol dehydrogenase cDNA. Gene 174: 159-164.

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14: 685-695.

Glasner JD, Kocher TD, Collins JJ (1995) *Caenorhabditis elegans* contains genes enconding two new members of the Zn-containing alcohol dehydrogenase family. J. Mol. Evol. 41: 46-53.

Godoy L, Gonzàlez-Duarte R, Albalat R (2007) Analysis of planarian *Adh3* supports an intron-rich architecture and tissue-specific expression for the urbilaterian ancestral form. Comp. Biochem. Physiol. 146: 489-495.

Gonzàlez-Duarte R, Albalat R (2005) Merging protein, gene and genomic data: the evolution of the MDR-ADH family. Heredity 95: 184-197.

Gu X (2006) A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol. Biol. Evol. 23: 1937-1945.

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinform. 18: 500-501.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696-704.

Gutheil WG, Holmquist B, Vallee BL (1992) Purification, characterization, and partial sequence of the glutathione-dependent formaldehyde dehydrogenase from *Escherichia coli*: a class III alcohol dehydrogenase. Biochemistry 31: 475 – 481.

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89: 10915-10919.

Hey J (1999) The neutralist, the fly, and the selectionist. Trends Ecol. Evol. 14: 35-38.

Hjelmqvist L, Estonius M, Jörnvall H (1995a) The vertebrate alcohol dehydrogenase system: variable class II type form elucidates separate stages of enzymogenesis. Proc. Natl. Acad. Sci. USA 92: 10904-10908.

Hjelmqvist L, Shafqat J, Siddiqi AR, Jörnvall H (1995b) Alcohol dehydrogenase of class III: consistent patterns of structural and functional conservation in relation to class I and other proteins. FEBS Lett. 373: 212 – 216.

Höög JO, Brandt M (1995) Mammalian class VI alcohol dehydrogenase: novel types of the rodent enzymes. Adv Exp. Med. Biol. 372: 355–364.

Höög JO, Hedberg JJ, Stromberg P, Svesson S (2001) Mammalian alcohol dehydrogenase – functional and structural implications. J. Biomed. Sci. 8: 71-76.

Iborra FJ, Renau Piqueras J, Portoles M, Boleda MD, Guerri C, Parés X (1992) Immunocytochemical and biochemical demonstration of formaldehyde dehydrogenase (class III alcohol dehydrogenase) in the nucleus. J. Histochem. Cytochem. 40: 1865-1878.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem. Sci. 23: 403– 405.

Jobb G, Von Haesler A, Strimmer K (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4: 1-9.

Jörnvall H (2008) MDR and SDR gene and protein superfamilies. Cell. Mol. Life Sci. 65: 3875-3878.

Jörnvall H, Höög JO (1995) Nomenclature of alcohol dehydrogenases. Alcohol Alcoholism 30: 153 – 161.

Kaiser R, Férnandez MR, Parés X, Jörnvall H (1993) Origin of human alcohol dehydrogenase system: implications from the structure and properties of the octopus protein. Proc. Natl. Acad. Sci. USA 90: 11222 – 11226.

Kedishvili NY, Gough WH, Chernoff EAG, Hurley TD, Stone CL, Bowman KD, Popov KM, Bosron WF, Li TK (1997) cDNA sequence and catalytic properties of a chick embryo alcohol dehydrogenase that oxidizes retinol and 3b,5a-hydroxysteroids. J. Biol. Chem. 272**:** 7494–7500.

Kumar S, Tamura K, Jakobsen I, Nei M (2000) MEGA2: molecular evolutionary genetics analysis software. Bioinform. 17: 1244-1245.

Kumar S, Dudley J, Nei M, Tamura K (2008) MEGA: a biologist centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioninform. 9: 299-306.

Ladrière JM, Georis I, Guérineau M, Vandenhaute J (2000) *Kluyveromyces marxianus* exhibits an ancestral *Saccharomyces cerevisiae* genome organization downstream of ADH2. Gene 255: 83-91.

Lawton-Rauth A (2003) Evolutionary dynamics of duplicated genes in plants. Mol. Phylogenet. Evol. 29: 396-409.

Martínez MC, Achkor H, Persson B, Férnandez MR, Shafqat J, Farrés J, Jórnvall H, Parés X (1996) Arabidopsis formaldehyde dehydrogenase. Eur. J. Biochem. 241: 849 – 857.

Nicholas KB, Nicholas HB Jr (1997) GeneDoc: a tool for editing and annotating multiple sequence alignment. Distributed by the authors (www.psc.edu/biomed/genedoc).

Norin A, Van Ophen PW, Piersma SR, Persson B, Duine JA, Jörnvall H (1997) Mycothiol-dependent formaldehyde dehydrogenase, a prokaryotic medium-chain dehydrogenase/reductase, phylogenetically links different eukaryotic alcohol dehydrogenases. Eur. J. Biochem. 248: 282 – 289.

Parés X, Cederlund E, Moreno A, Hjelmqvist L, Farrés J, Jörnvall H (1994) Mammalian class IV alcohol dehydrogenase (stomach alcohol dehydrogenase): structure, origin, and correlation with enzymology. Proc Natl. Acad. Sci. 91: 1893-1897.

Persson B, Bergman T, Keung WM, Waldenström U, Holmquist B, Vallee BL, Jörnvall H (1993) Basic features of class-I alcohol dehydrogenase: variable and constant segments coordinated by inter-class and intra-class variability. Conclusions from characterization of the alligator enzyme. Eur. J. Biochem. 216: 49-56.

Posada D, Crandall KA (2001) Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50: 580-601.

Ras J, Van Ophem PW, Reijnders WNM, Van Spanning RJM, Duine JA, Stouthamer AH, Harms N (1995) Isolation, sequencing, and mutagenesis of the gene enconding NAD- and glutathione-dependent formaldehyde dehydrogenase (GD-FALDH) from *Paracoccus denitrificans*, in which GD-FALDH is essential for methylotrophic growth. J. Bacteriol. 177: 247 – 251.

Reimers MJ, Hahn ME, Tanguay RL (2004) Two zebrafish alcohol dehydrogenases share common ancestry with mammalian class I, II, IV, and V alcohol dehydrogenase genes but have distinct functional characteristics. J. Biol. Chem. 279: 38303-38312.

Rosell A, Valencia E, Parés X, Fita I, Farrés J, Ochoa WF (2003) Crystal structure of the vertebrate NADP(H)-dependent alcohol dehydrogenase (ADH8). J. Mol. Biol. 330: 75-85.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

Sanghani PC, Bosron WF, Hurley TD (2002) Human glutathione-dependent formaldehyde dehydrogenase. Biochem. 41: 15189-15194.

Sasnaukas K, Jomantiené R, Januska A, Lebediené E, Lebedys J, Janulaitis A (1992) Cloning and analysis of a Candida maltosa gene which confers resistance to formaldehyde in *Saccharomyces cerevisiae*. Gene 122: 207 – 211.

Schwarz G (1978) Estimating the dimension of a model. Ann. Statist. 6: 461-464.

Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene tress. Proc. Roy. Soc. B. 269: 137-142.

Strömberg P, Höög J (2000) Human class V alcohol dehydrogenase (ADH5): a complex transcription unit generates C-terminal multiplicity. Biochem. Biophys. Res. Commun. 278: 544-549.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24: 1596-1599.

Tennessen JA (2008) Positive selection drives a correlation between nonsynonymous / synonymous divergence and functional divergence. Bioinform. 24: 1421-1425.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalty and weight matrix choice. Nucleic Acids Res. 22: 4673-4680.

Thompson CE, Salzano FM, Norberto de Souza O, Freitas LB (2007) Sequence and structural aspects of functional diversification of plant alcohol dehydrogenases. Gene 396: 108-115.

Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA (2005) Resurrecting ancestral alcohol dehydrogenases from yeast. Nat. Genet. 37: 630-635.

Wendel JF (2000) Genome evolution in poliploids. Plant Mol. Biol. 42: 225-249.

Williamson VM, Paquin CE (1987) Homology of *Saccharomyces cerevisiae* ADH4 to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis*. Mol. Gen. Genet. 209: 374-381.

Wills C, Jörnvall H (1979) The two major isoenzymes of yeast alcohol dehydrogenase. Eur. J. Biochem. 99: 323-331.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comp. Appl. BioSci. 13: 555-556.

Yang Z (2004) PAML: Phylogenetic Analysis by Maximum Likelihood version 3.14. University College London, London.

Young ET, Sloan J, Miller B, Li N, van Riper K, Dombek KM (2000) Evolution of a glucose-regulated ADH gene in the genus *Saccharomyces*. Gene 245: 299-309.

Van de Peer Y (2003) Phylogeny inference based on distance methods. In: Salemi M, Vandamme A-M (eds) The phylogenetic handbook. A practical approach to DNA and protein phylogeny. Cambridge University Press, Cambridge, United Kingdom, pp. 101-119.

Zheng Y, Xu D, Gu X (2007) Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits. J. Exp. Zool. (Mol. Dev. Evol.) 308B: 85-96.

Zheng Y-W, Bey M, Liu H, Felder MR (1993) Molecular basis of the alcohol dehydrogenase-negative deer mouse. Evidence for deletion of the gene for class I enzyme and identification of a possible new enzyme class. J. Biol. Chem. 268: 24933–24939.

Zgombic-Knight M, Ang HL, Foglio MH, Duester G (1995) Cloning of the mouse class IV alcohol dehydrogenase (retinol dehydrogenase) cDNA and tissue-specific expression patterns of the murine ADH gene family. J. Biol. Chem. 270: 10868–10877.

Table 1

Coefficients of functional divergence (θ) of pairwise comparisons in the alcohol dehydrogenase gene family

Comparison	*Group 1*	*Group 2*	*θ±SE*[a]	*LRT*[b]
Between forms	Mammals ADH3	Mammals ADH2	*0.7136±0.222*	*10.330*
	Mammals ADH3	Mammals ADH5	*0.6712±0.204*	*10.824*
	Mammals ADH3	Mammals ADH4	*0.9704±0.257*	*14.281*
	Mammals ADH2	Mammals ADH5	*0.4032±0.159*	*6.414*
	Mammals ADH2	Mammals ADH1	*0.4296±0.105*	*16.585*
	Mammals ADH5	Mammals ADH4	*0.2512±0.289*	*0.758**
	Mammals ADH5	Mammals ADH1	*0.4264±0.113*	*14.348*
	Mammals ADH4	Mammals ADH1	*0.8094±0.188*	*18.431*
	Fishes ADH1	Fishes ADH3	*0.5112±0.080*	*40.475*
	Fungi ADH1	Fungi ADH3	*0.6040±0.252*	*5.745*
	Fungi ADH1	Fungi ADH4	*0.6496±0.179*	*13.186*
	Fungi ADH1	Fungi ADH5	*0.8824±0.115*	*58.618*
	Fungi ADH3	Fungi ADH4	*0.0370±0.240*	*0.023**
	Fungi ADH3	Fungi ADH5	*0.7704±0.130*	*35.361*
	Fungi ADH4	Fungi ADH5	*0.7968±0.091*	*76.845*

[a] SE stands for standard error. [b] LRT: Likelihood Ratio Test. All values are statistically significant at P<0.001 or less, except those labeled with (*). Sequences of bird, amphibia, and reptilia had incomplete information for this type of analysis.

Table 2

Amino acid residues important for the functional divergence between ADH mammalian forms

Amino acid residues	Mammals ADH3/ADH2	Mammals ADH3/ADH5	Mammals ADH4/ADH1	Mammals ADH2/ADH5	Mammals ADH2/ADH1
60		**0.889377**			
70	0.834021			0.837070	
78			**0.892779**		
91			0.802919		
103	0.811543				
105			0.802919		
128	0.807681		0.802919		
133			**0.892774**		
134			**0.892774**		0.802368
135			**0.892774**		
142	0.800877				
161			**0.865930**		
163	0.809431		0.802919		
165	0.806196				
177			0.802919		
194			0.802919		
201	0.807681				
209	0.903189*	**0.889513**			
212			0.802919		
214	0.805189				
215			**0.892534**		
223	0.807681				
233	0.811543				
239		**0.889513**			
243	0.809517				
250			0.802919		
256			0.802919		
257	**0.892105**	**0.876564**			
259			0.802919		
264	0.956807*	0.951870*			
268	0.809431				
272	0.802480				
275	0.806196				
282			0.802919		
285	0.817554				
289			0.802919		
297			0.802919		

[a] In bold amino acid residues with $Q(k)\geq 0.85$, while an asterisk indicates those with $Q(k)\geq 0.90$ as a cut off value.

Table 3

Amino acid residues important for the functional divergence between fungi and fishes ADH forms

Amino acid residues	Fishes ADH3/ADH1	Fungi ADH4/ADH1	Fungi ADH3/ADH5	Fungi ADH4/ADH5
36			0.938280*	
49			0.926019*	
58	0.914537*			
59			0.948280*	
62	0.945906*			
89			0.948280*	
129	**0.898560**			
175		0.933913*		
200			0.935322*	
209	**0.858626**			
232		0.920585*		
233	0.902046*			
246		**0.859203**		
259	0.851359*			
262			0.935322*	
279			0.938280*	
301	0.940363*		0.935322*	
313			0.948280*	
312				0.957991*
316				0.949383*
324			0.938280*	
327	0.940363*			
328				0.911417*
348			0.948280*	
354	0.940363*			
358				0.973481*
360				0.949383*
366			0.938280*	
373			0.948280*	0.976559*
382				0.949383*
383				0.949383*
385			0.948280*	0.975113*

[a]In bold amino acid residues with $Q(k)\geq0.85$, while an asterisk indicates those with $Q(k)\geq0.90$ as a cut off value.

Fig. 1 Linearized phylogenetic tree inferred from protein ADH sequences using the neighbor-joining method, employing the Poisson-corrected amino acid distance and the pairwise deletion of gaps/missing data.

Fig. 2 Detailed representation of the relationships obtained with mammal protein sequences. Numbers represent bootstrap values; values higher than 80% are shown. Scale bar indicates levels of sequence divergence. Clusters distinguishable by ADH type are highlighted.

Fig. 3 Fungi protein sequence relationships with labels indicating clusters distinguishable by ADH type and fungi genera. Numbers represent bootstrap values; values higher than 80% are shown. Scale bar indicates levels of sequence divergence.

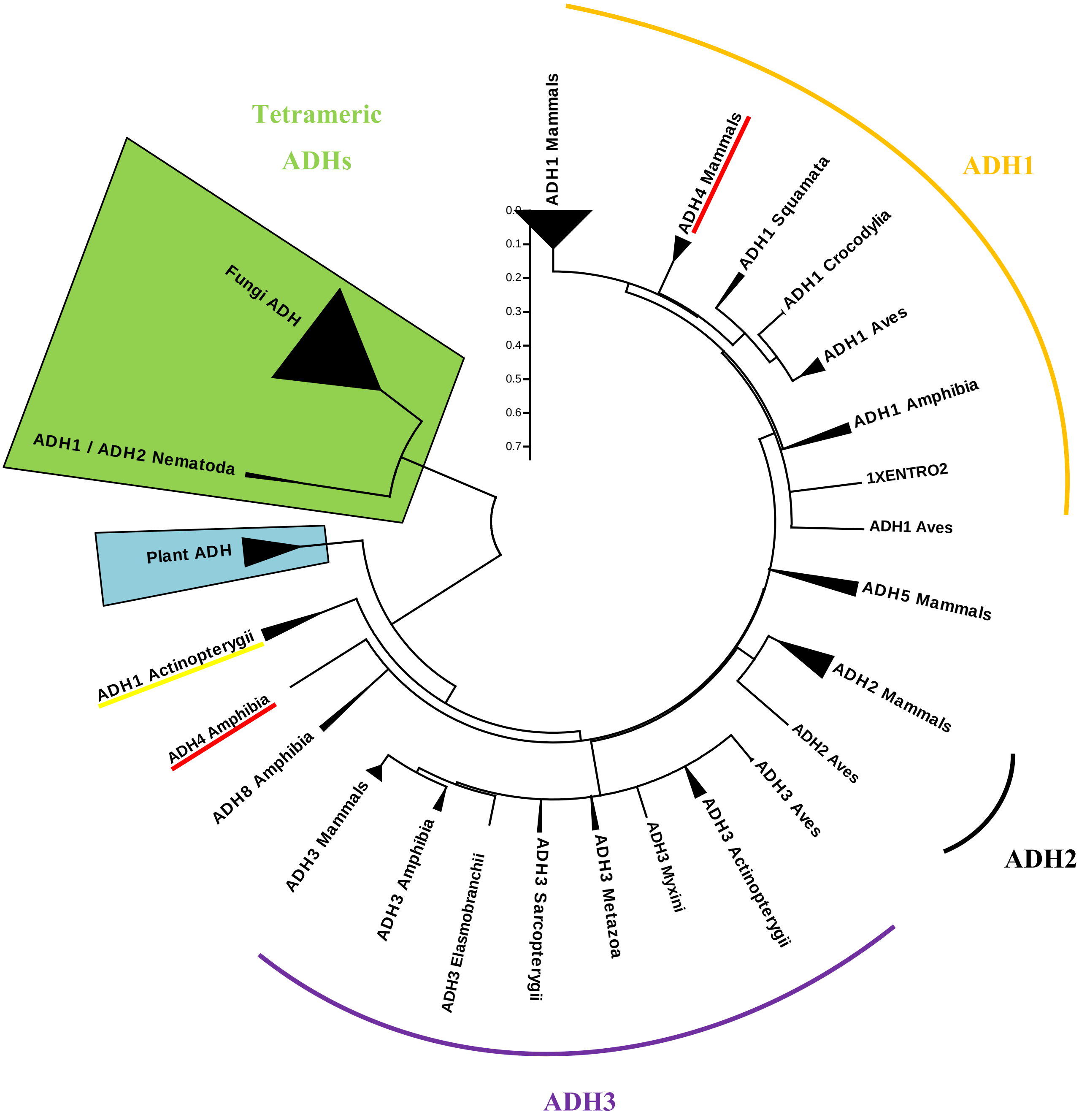

Tetrameric
ADHs
ADH1
ADH2
ADH3
Fungi ADH
ADH1 / ADH2 Nematoda
Plant ADH
ADH1 Actinopterygii
ADH4 Amphibia
ADH8 Amphibia
ADH3 Mammals
ADH3 Amphibia
ADH3 Elasmobranchii
ADH3 Sarcopterygii
ADH3 Metazoa
ADH3 Myxini
ADH3 Actinopterygii
ADH3 Aves
ADH2 Aves
ADH2 Mammals
ADH5 Mammals
ADH1 Aves
1XENTRO2
ADH1 Amphibia
ADH1 Aves
ADH1 Crocodylia
ADH1 Squamata
ADH4 Mammals
ADH1 Mammals
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Fig. 1

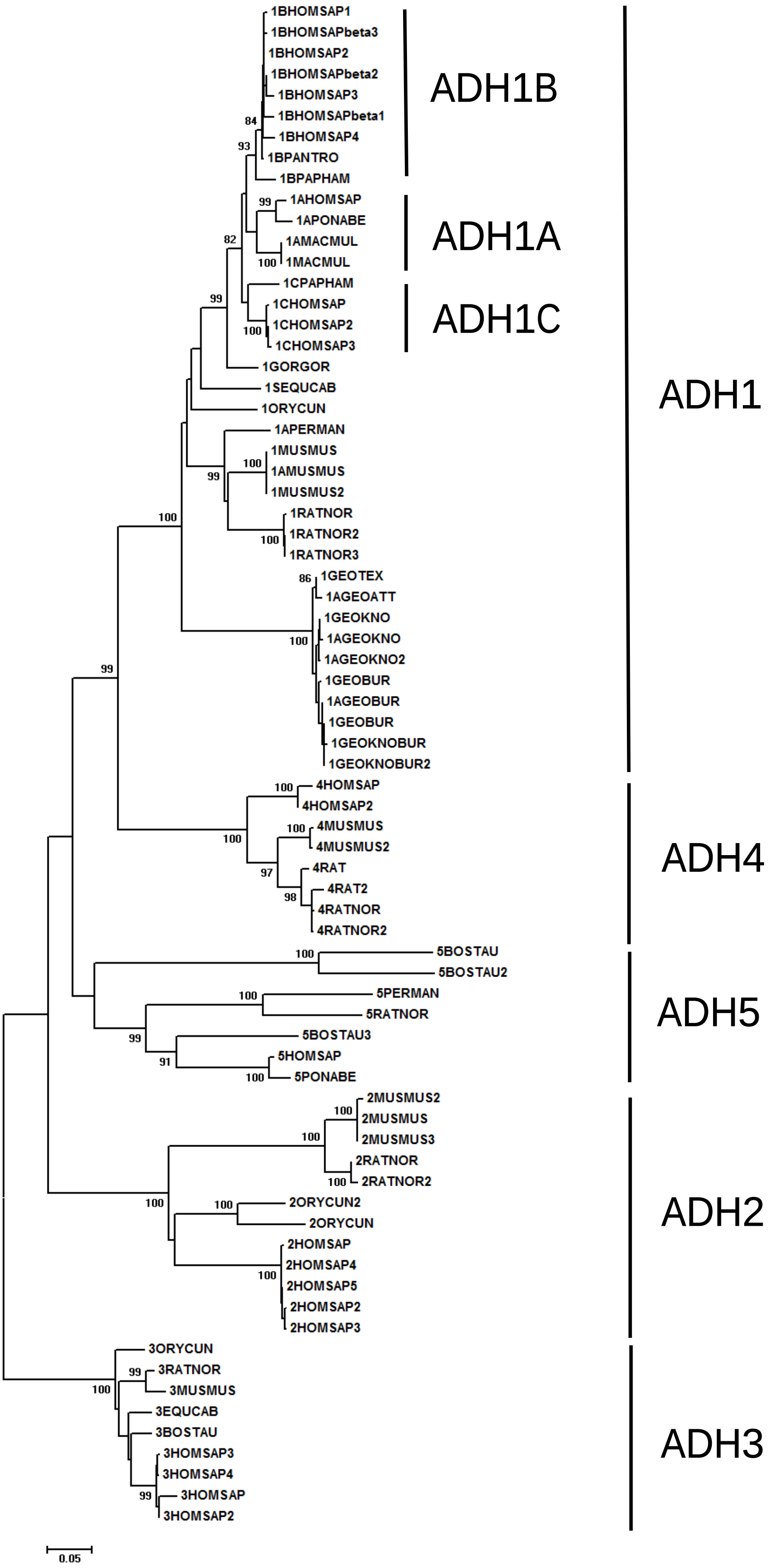

1BHOMSAP1
1BHOMSAPbeta3
1BHOMSAP2
1BHOMSAPbeta2
1BHOMSAP3
84
1BHOMSAPbeta1
1BHOMSAP4
93
1BPANTRO
1BPAPHAM
ADH1B
99
1AHOMSAP
1APONABE
82
1AMACMUL
100
1MACMUL
ADH1A
1CPAPHAM
99
1CHOMSAP
100
1CHOMSAP2
1CHOMSAP3
ADH1C
1GORGOR
1SEQUCAB
1ORYCUN
ADH1
1APERMAN
1MUSMUS
100
99
1AMUSMUS
1MUSMUS2
100
1RATNOR
1RATNOR2
100
1RATNOR3
86
1GEOTEX
1AGEOATT
1GEOKNO
100
1AGEOKNO
1AGEOKNO2
99
1GEOBUR
1AGEOBUR
1GEOBUR
1GEOKNOBUR
1GEOKNOBUR2
100
4HOMSAP
4HOMSAP2
100
4MUSMUS
100
4MUSMUS2
ADH4
4RAT
97
4RAT2
98
4RATNOR
4RATNOR2
100
5BOSTAU
5BOSTAU2
100
5PERMAN
5RATNOR
ADH5
5BOSTAU3
99
91
5HOMSAP
100
5PONABE
2MUSMUS2
100
2MUSMUS
100
2MUSMUS3
2RATNOR
100
2RATNOR2
100
100
2ORYCUN2
2ORYCUN
ADH2
2HOMSAP
2HOMSAP4
100
2HOMSAP5
2HOMSAP2
2HOMSAP3
3ORYCUN
99
3RATNOR
100
3MUSMUS
3EQUCAB
3BOSTAU
ADH3
3HOMSAP3
3HOMSAP4
99
3HOMSAP
3HOMSAP2
0.05

Fig. 2

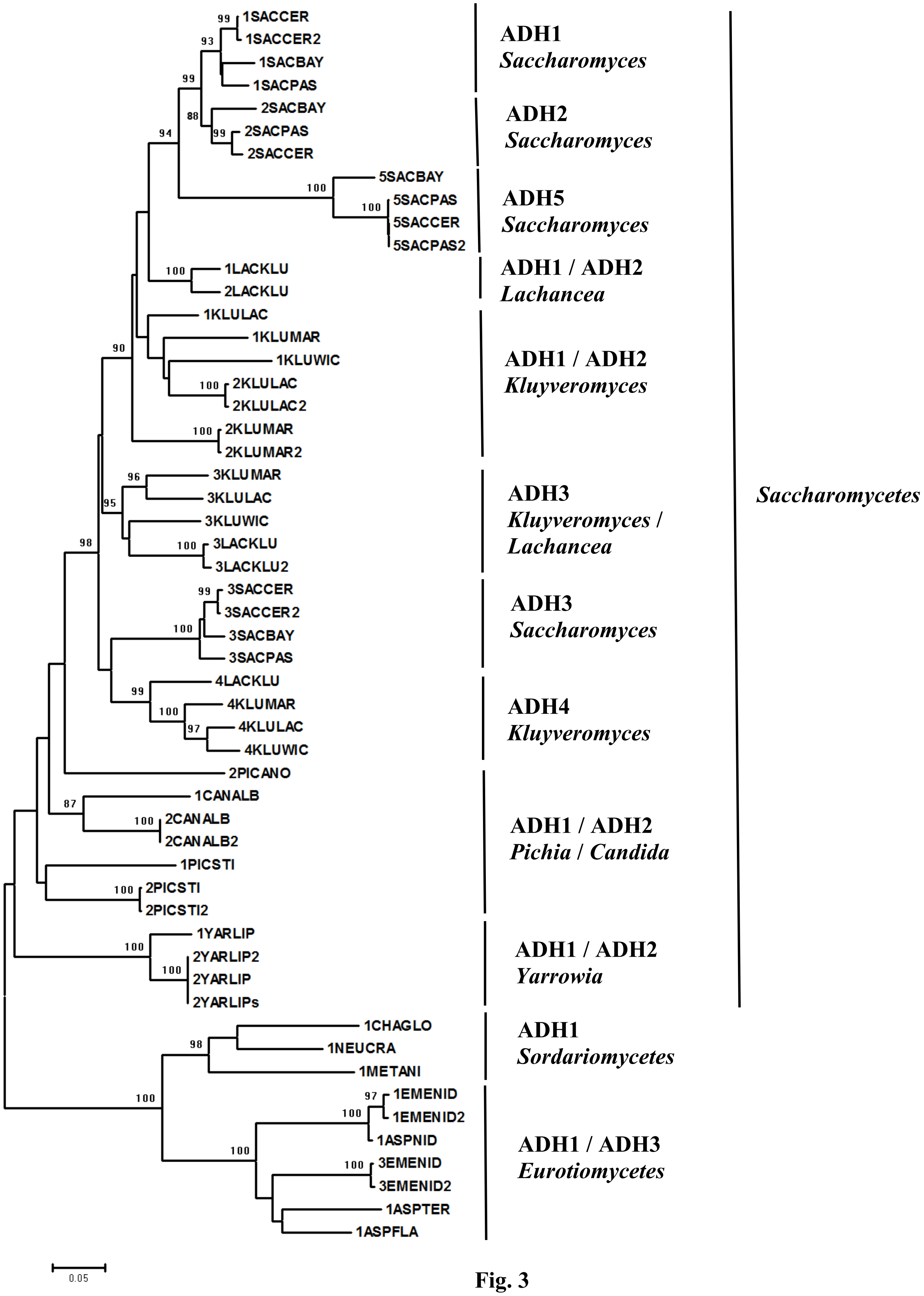

99
1SACCER
93
1SACCER2
1SACBAY
99
1SACPAS
2SACBAY
88
99
2SACPAS
94
2SACCER
5SACBAY
100
5SACPAS
100
5SACCER
5SACPAS2
100
1LACKLU
2LACKLU
1KLULAC
1KLUMAR
90
1KLUWIC
100
2KLULAC
2KLULAC2
100
2KLUMAR
2KLUMAR2
96
3KLUMAR
95
3KLULAC
3KLUWIC
98
100
3LACKLU
3LACKLU2
99
3SACCER
3SACCER2
100
3SACBAY
3SACPAS
4LACKLU
99
4KLUMAR
100
97
4KLULAC
4KLUWIC
2PICANO
1CANALB
87
100
2CANALB
2CANALB2
1PICSTI
100
2PICSTI
2PICSTI2
1YARLIP
100
2YARLIP2
100
2YARLIP
2YARLIPs
1CHAGLO
98
1NEUCRA
1METANI
100
97
1EMENID
100
1EMENID2
1ASPNID
100
100
3EMENID
3EMENID2
1ASPTER
1ASPFLA
0.05
ADH1 *Saccharomyces*
ADH2 *Saccharomyces*
ADH5 *Saccharomyces*
ADH1 / ADH2 *Lachancea*
ADH1 / ADH2 *Kluyveromyces*
ADH3 *Kluyveromyces / Lachancea*
ADH3 *Saccharomyces*
ADH4 *Kluyveromyces*
ADH1 / ADH2 *Pichia / Candida*
ADH1 / ADH2 *Yarrowia*
ADH1 *Sordariomycetes*
ADH1 / ADH3 *Eurotiomycetes*
Saccharomycetes

Fig. 3

Electronic Supplementary Material

Table 1S Alcohol dehydrogenase proteins to which the sequences considered belong, and respective taxonomic information and accession number.

Taxonomy					ADH type	Accesion number	Species	Sequence size
Kingdoms	Phylum	Class	Order	Families				
Animalia	Chordata	Actinopterygii (Osteichthyes)	Perciformes	Cypirinidae	1DANRER	AAK97853	*Danio rerio*	377
					3DANRER	AAL26325	*Danio rerio*	376
					3DANRER2	NP_571924	*Danio rerio*	376
					1B8DANRER	NP_982285	*Danio rerio*	376
					1A8DANRER	NP_001001946	*Danio rerio*	377
					1A8DANRER2	AAH96855	*Danio rerio*	377
				Sparidae	3SPAAUR	P79896	*Sparus aurata*	376
			Gadiformes	Gadidae	3GADMORL	P81601	*Gadus morhua*	375
					3GADMORH	P81600	*Gadus morhua*	375
					1GADCAL	P26325	*Gadus callaris*	375
			Beloniformes	Oryziinae	1ORYLAT	AAT81592	*Oryzias latipes*	378
					3ORYLAT	AAS15570	*Oryzias latipes*	379
		Myxini	Myxiniformes	Myxinidae	3MYXGLU	P80360	*Myxine glutinosa*	376
		Elasmobranchii (Chondrichthyes)	Carcharhiniformes	Scyliorhinidae	3SCYCAN	AAS49606	*Scyliorhinus canicula*	345
		Sarcopterygii	Coelacanthiformes	Coelacanthidae	3LATCHA	AAS49517	*Latimeria chalumnae*	344
			Lepidosireniformes	Protopteridae	3PRODOL	AAS49516	*Protopterus dolloi*	346
		Amphibia	Anura	Ranidae	1RANPER	P22797	*Rana perezi*	375
					8RANPER	O57380	*Rana perezi*	373
					8ARANPER	1P0F_A	*Rana perezi*	373
				Pipidae	1XENTRO2	NP_001011391	*Xenopus tropicalis*	376
					1BXENTRO	NP_001011431	*Xenopus tropicalis*	378
					5XENTRO	NP_001011423	*Xenopus tropicalis*	376
					3XENTRO	CAJ83953	*Xenopus tropicalis*	376
					3XENLAE	NP_001086427	*Xenopus laevis*	376
					8XENLAE	NP_001089094	*Xenopus laevis*	373
					5XENLAE	NP_001086903	*Xenopus laevis*	378
					3XENTRO2	AAS49608	*Xenopus laevis*	376
					4XENLAE	AAC33715	*Xenopus laevis*	271
					1XENLAE	AAH97612	*Xenopus laevis*	376
		Reptilia	Crocodylia	Crocodylidae	1ALLMIS	P80222	*Aligator mississippiensis*	374
					1AMEALL	S35669	*American alligator*	374
			Squamata	Agamidae	1AUROHAR	P25405	*Uromastyx hardwickii*	375
					1BUROHAR	P25406	*Uromastyx hardwickii*	375
					3UROHAR	P80467	*Uromastyx hardwickii*	373
				Elapidae	1NAJNAJ	P80512	*Naja naja*	375

Table 1S Cont.

Taxonomy					ADH type	Accesion number	Species	Sequence size
Kingdoms	Phylum	Class	Order	Families				
Animalia	Chordata	Aves	Galliformes	Phasianidae	1FGALGAL	AAB66676	*Gallus gallus*	376
					1CGALGAL	CAA38433	*Gallus gallus*	329
					3GALGAL	NP_001026323	*Gallus gallus*	374
					3GALGAL2	AAS49609	*Gallus gallus*	371
					1GALGAL	P23991	*Gallus gallus*	376
					1COTJAP	P19631	*Coturnix japonica*	375
			Struthioniformes	Struthionidae	1STRCAM2	AAB32020	*Struthio camelus*	374
					2STRCAM	P80468	*Struthio camelus*	379
					1STRCAM	P80338	*Struthio camelus*	374
			Anseriformes	Anatidae	1ANAPLA	P30350	*Anas platyrhynchos*	185
		Mammalia	Primates	Hominidae	1BHOMSAP1	AAA51884	*Homo sapiens*	375
					1BHOMSAP2	P00325	*Homo sapiens*	375
					1BHOMSAPbeta3	AAB48003	*Homo sapiens*	375
					1BHOMSAPbeta2	AAD37446	*Homo sapiens*	375
					1BHOMSAP3	CAA33487	*Homo sapiens*	376
					1BHOMSAPbeta1	AAA51592	*Homo sapiens*	375
					1BHOMSAP4	EAX06097	*Homo sapiens*	355
					1CHOMSAP3	AAH67421	*Homo sapiens*	375
					1CHOMSAP	P00326	*Homo sapiens*	375
					1AHOMSAP	P07327	*Homo sapiens*	375
					1CHOMSAP2	AAH62476	*Homo sapiens*	375
					4HOMSAP	AAB38424	*Homo sapiens*	374
					4HOMSAP2	P40394	*Homo sapiens*	374
					5HOMSAP	AAA35509	*Homo sapiens*	368
					2HOMSAP	P08319	*Homo sapiens*	380
					2HOMSAP2	AAA51595	*Homo sapiens*	392
					2HOMSAP3	NP_000661	*Homo sapiens*	380
					2HOMSAP4	EAX06088	*Homo sapiens*	380
					2HOMSAP5	EAX06089	*Homo sapiens*	380
					3HOMSAP	EAX06085	*Homo sapiens*	395
					3HOMSAP2	P11766	*Homo sapiens*	374
					3HOMSAP3	AAA51597	*Homo sapiens*	392
					3HOMSAP4	AAV38636	*Homo sapiens*	374
					1GORGOR	AAL56229	*Gorilla gorilla*	269
					5PONABE	Q5R7Z8	*Pongo abelii*	375
					1APONABE	Q5RBP7	*Pongo abelii*	375
					1BPANTRO	Q5R1W2	*Pan troglodytes (chimpanzee)*	375
				Cercopithecidae	1BPAPHAM	P14139	*Papio hamadryas*	375
					1AMACMUL	P28469	*Macaca mulatta (rhesus)*	375
					1MACMUL	AAA36830	*Macaca mulatta*	375
					1CPAPHAM	O97959	*Papio hamadryas*	375

Table 1S Cont.

Taxonomy					ADH type	Accesion number	Species	Sequence size
Kingdoms	Phylum	Class	Order	Families				
Animalia	Chordata	Mammalia	Perissodactyla	Equidae	1SEQUCAB	P00328	*Equus caballus (horse)*	375
					3EQUCAB	P19854	*Equus caballus (horse)*	374
			Lagomorpha	Leporidae	1ORYCUN	Q03505	*Oryctolagus cuniculus (rabbit)*	375
					2ORYCUN	O46650	*Oryctolagus cuniculus*	379
					2ORYCUN2	O46649	*Oryctolagus cuniculus*	379
					3ORYCUN	O19053	*Oryctolagus cuniculus*	374
			Rodentia	Cricetidae	1APERMAN	P41680	*Peromyscus maniculatus (deer mouse)*	375
					5PERMAN	P41681	*Peromyscus maniculatus*	375
				Murinae	1MUSMUS2	AAH13477	*Mus musculus (house mouse)*	375
					1AMUSMUS	P00329	*Mus musculus*	375
					1MUSMUS	1402250A	*Mus musculus*	375
					4MUSMUS	AAH47267	*Mus musculus*	374
					4MUSMUS2	Q64437	*Mus musculus*	374
					2MUSMUS3	CAB57455	*Mus musculus*	377
					2MUSMUS	Q9QYY9	*Mus musculus*	377
					2MUSMUS2	NP_036126	*Mus musculus*	377
					3MUSMUS	P28474	*Mus musculus*	374
					1RATNOR	P06757	*Rattus norvegicus (norway rat)*	376
					1RATNOR2	AAA40681	*Rattus norvegicus*	376
					1RATNOR3	gi|225439	*Rattus norvegicus*	376
					4RATNOR	CAA67297	*Rattus norvegicus*	374
					4RATNOR2	P41682	*Rattus norvegicus*	374
					2RATNOR	AAI27505	*Rattus norvegicus*	377
					2RATNOR2	Q64563	*Rattus norvegicus*	377
					5RATNOR	Q5XI95	*Rattus norvegicus*	376
					3RATNOR	P12711	*Rattus norvegicus*	374
					4RAT	AAB30153	*Rattus sp*	374
					4RAT2	A53142	*Rattus sp*	374
				Geomyidae	1GEOTEX	AAC98960	*Geomys texensis*	375
					1AGEOATT	Q9Z2M2	*Geomys attwateri*	375
					1GEOKNO	AAC98957	*Geomys knoxjonesi*	375
					1AGEOKNO	Q64415	*Geomys knoxjonesi*	375
					1GEOKNO2	AAA03600	*Geomys knoxjonesi*	375
					1GEOBUR	AAC98958	*Geomys bursarius major*	375
					1AGEOBUR	Q64413	*Geomys bursarius (plains pocket gopher)*	375
					1GEOKNOBUR	AAA03596	*Geomys knoxjonesi x Geomys bursarius major*	375
					1GEOBUR	AAA03595	*Geomys bursarius*	375
					1GEOKNOBUR2	AAA03597	*Geomys knoxjonesi x Geomys bursarius major*	375
			Artiodactyla	Bovidae	5BOSTAU	AAI20379	*Bos taurus*	375
					5BOSTAU2	AAI30017	*Bos taurus*	375
					5BOSTAU3	AAI12631	*Bos taurus*	375
					3BOSTAU	Q3ZC42	*Bos taurus*	375

Table 1S Cont.

					ADH type	Accesion number	Species	Sequence size
Kingdom	Phylum	Class	Order	Families				
Animalia	Mollusca	Cephalopoda	Octopoda	Octopodidae	3OCTVUL	P81431	*Octopus vulgaris*	378
	Nematoda	Chromadorea	Rhabditida	Rhabditidae	2CAEELE	O45687	*Caenorhabditis elegans*	351
					1CAEELE	Q17334	*Caenorhabditis elegans*	349
					3CAEELE	Q17335	*Caenorhabditis elegans*	384
	Platyhelminthes	Turbellaria	Tricladida	Dugesiidae	3SCHMED	ABG78601	*Schmidtea mediterranea*	379
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	Saccharomycetaceae	1SACCER	2HCY	*Saccharomyces cerevisiae (Yeast)*	347
					1SACCER2	P00330	*Saccharomyces cerevisiae*	348
					1SACBAY	AAP51042	*Saccharomyces bayanus*	348
					1SACPAS	AAP51050	*Saccharomyces pastorianus*	348
					2SACBAY	AAP51043	*Saccharomyces bayanus*	348
					2SACPAS	AAP51051	*Saccharomyces pastorianus*	348
					2SACCER	P00331	*Saccharomyces cerevisiae*	348
					5SACBAY	AAP51045	*Saccharomyces bayanus*	351
					5SACPAS	AAP51053	*Saccharomyces pastorianus*	351
					5SACCER	P38113	*Saccharomyces cerevisiae*	351
					5SACPAS2	Q6XQ67	*Saccharomyces pastorianus*	351
					2KLUMAR	AAF91235	*Kluyveromyces marxianus*	348
					2KLUMAR2	Q9P4C2	*Kluyveromyces marxianus*	348
					1LACKLU	AAP51046	*Lachancea kluyveri*	351
					2LACKLU	AAP51047	*Lachancea kluyveri*	351
					1KLULAC	P20369	*Kluyveromyces lactis*	350
					1KLUMAR	Q07288	*Kluyveromyces marxianus*	348
					1KLUWIC	AAP51039	*Kluyveromyces wickerhamii*	348
					2KLULAC	CAA45739	*Kluyveromyces lactis*	348
					2KLULAC2	P49383	*Kluyveromyces lactis*	348
					3SACCER	AAA34409	*Saccharomyces cerevisiae*	375
					3SACCER2	P07246	*Saccharomyces cerevisiae*	375
					3SACBAY	AAP51044	*Saccharomyces bayanus*	375
					3SACPAS	AAP51052	*Saccharomyces pastorianus*	375
					4LACKLU	AAP51049	*Lachancea kluyveri*	377
					4KLUMAR	BAF43529	*Kluyveromyces marxianus*	379
					4KLUWIC	AAP51041	*Kluyveromyces wickerhamii*	374
					4KLULAC	P49385	*Kluyveromyces lactis*	375
					3KLUWIC	AAP51040	*Kluyveromyces wickerhamii*	371
					3LACKLU	AAF43645	*Lachancea kluyveri*	373
					3LACKLU2	AAP51048	*Lachancea kluyveri*	373
					3KLUMAR	BAF43528	*Kluyveromyces marxianus*	375
					3KLULAC	P49384	*Kluyveromyces lactis*	374
					2PICANO	CAH56496	*Pichia anomala*	372

Table 1S Cont.

Kingdoms	Phylum	Class	Order	Families				
Fungi	Ascomycota	Saccharomycetes	Saccharomycetales	Saccharomycetaceae	2PICSTI	AAC49990	*Pichia stipitis*	348
					2PICSTI2	O13309	*Pichia stipitis*	348
					1PICSTI	O00097	*Pichia stipitis*	348
					1CANALB	P43067	*Candida albicans*	350
					2CANALB	CAA21988	*Candida albicans*	348
					2CANALB2	O94038	*Candida albicans*	348
				Dipodascaceae	1YARLIP	AAD51737	*Yarrowia lipolytica*	349
					2YARLIP	AAD51738	*Yarrowia lipolytica*	351
					2YARLIPs	CAG79670	*Yarrowia lipolytica (strain: CLIB122)*	351
					2YARLIP2	XP_504077	*Yarrowia lipolytica*	351
		Sordariomycetes	Hypocreales	Clavicipitaceae	1METANI	ABD49723	*Metarhizium anisopliae*	353
			Sordariales	Chaetomiaceae	1CHAGLO	EAQ83781	*Chaetomium globosum (strain: CBS148.51)*	356
				Sordariaceae	1NEUCRA	Q9P6C8	*Neurospora crassa*	353
		Eurotiomycetes	Eurotiales	Trichocomaceae	1ASPNID	EAA64311	*Aspergillus nidulans (strain: FGSC A4)*	350
					1EMENID	AAA33291	*Emericella nidulans (=Aspergillus nidulans)*	349
					1EMENID2	gP08843	*Emericella nidulans*	350
					1ASPTER	EAU30544	*Aspergillus terreus NIH2624*	350
					1ASPFLA	P41747	*Aspergillus flavus*	349
					3EMENID	CAA26541	*Emericella nidulans*	352
					3EMENID2	P07754	*Emericella nidulans*	352
Viridiplantae	Streptophyta	Coniferopsida	Coniferales	Pinaceae	1PINBAN	AAC49539	*Pinus banksiana*	375
		Liliopsida	Poales	Poaceae	1ORYRUF	BAC87775	*Oryza rufipogon*	379
					1ORYMER	BAC87779	*Oryza meridionalis*	379
					1ORYGLU	BAC87778	*Oryza glumipatula*	379
					1ORYSAT	BAC87776	*Oryza sativa*	379
					1HORVUL	P05336	*Hordeum vulgare*	379
					1ZEAMAY	CAA27682	*Zea mays*	379
					2HORVULVUL	P10847	*Hordeum vulgare subsp. vulgare*	373
					3HORVULVUL	CAA31231	*Hordeum vulgare subsp. vulgare*	379
					2ZEAMAY	P04707	*Zea mays*	379
					2ORYRUF	BAE00043	*Oryza rufipogon*	379
					2ORYSAT	BAE00044	*Oryza sativa*	379
					2ORYGLU	BAE00047	*Oryza glumipatula*	379
					2ORYMER	BAE00049	*Oryza meridionalis*	379

Fig. 1S Detailed representation of the relationships obtained with fish protein sequences using the neighbor-joining method, Poisson-corrected amino acid distances, and pairwise deletion of gaps/missing data. Numbers representing bootstrap values higher than 80% are shown. Scale bar indicates levels of sequence divergence.

Fig. 2S Representation of the relationships obtained with the amphibian protein sequences using the neighbor-joining method and Poisson-corrected amino acid distances. Bootstrap values higher than 80% are shown. Scale bar indicates the sequence divergence level.

Fig. 3S Detailed representation of the relationships obtained with the reptilian protein sequences using the neighbor-joining method and Poisson-corrected amino acid distances. Bootstrap values higher than 80% are shown. Scale bar indicates the sequence divergence level.

Fig. 4S Bird protein sequences relationships obtained using the neighbor-joining method and Poisson-corrected amino acid distances. Bootstrap values higher than 80% are shown. Scale bar indicates the sequence divergence level.

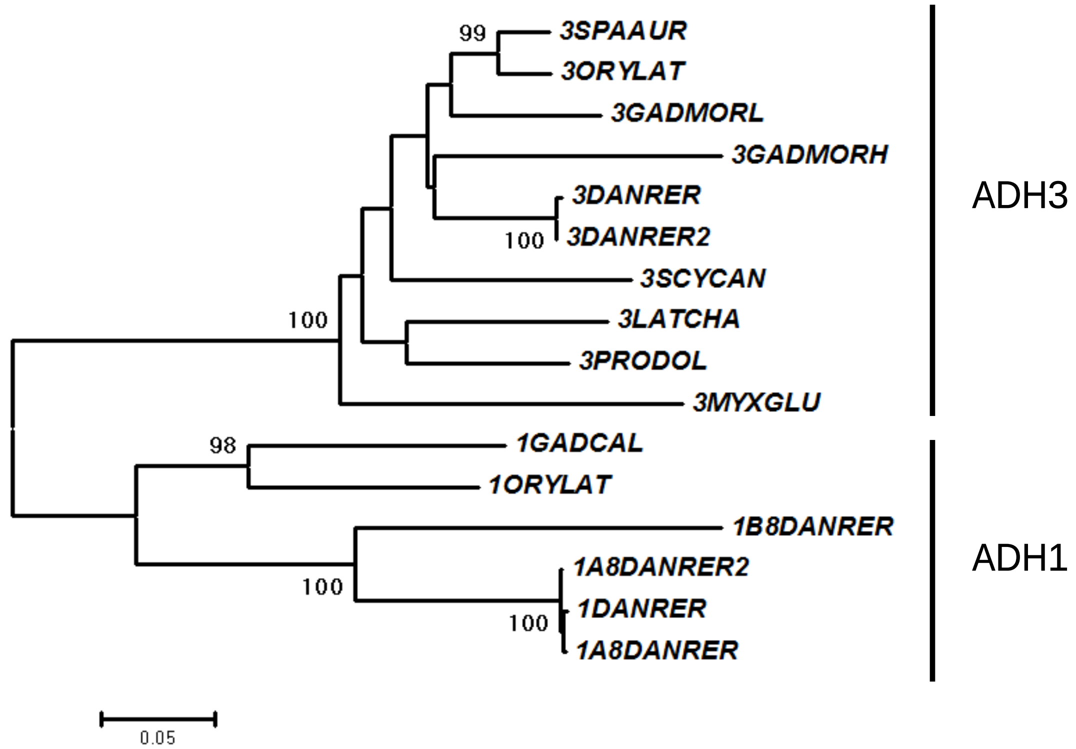

99
3SPAAUR
3ORYLAT
3GADMORL
3GADMORH
3DANRER
100
3DANRER2
3SCYCAN
100
3LATCHA
3PRODOL
3MYXGLU
98
1GADCAL
1ORYLAT
1B8DANRER
100
1A8DANRER2
1DANRER
100
1A8DANRER
ADH3
ADH1
0.05

Fig. 1S

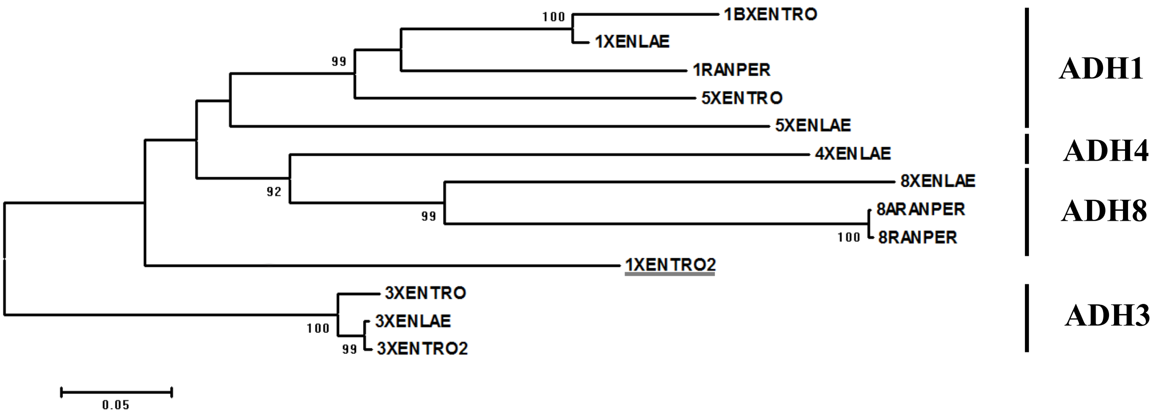

100
1BXENTRO
1XENLAE
99
1RANPER
5XENTRO
5XENLAE
4XENLAE
92
8XENLAE
99
8ARANPER
100
8RANPER
1XENTRO2
3XENTRO
100
3XENLAE
99
3XENTRO2
0.05
ADH1
ADH4
ADH8
ADH3

Fig. 2S

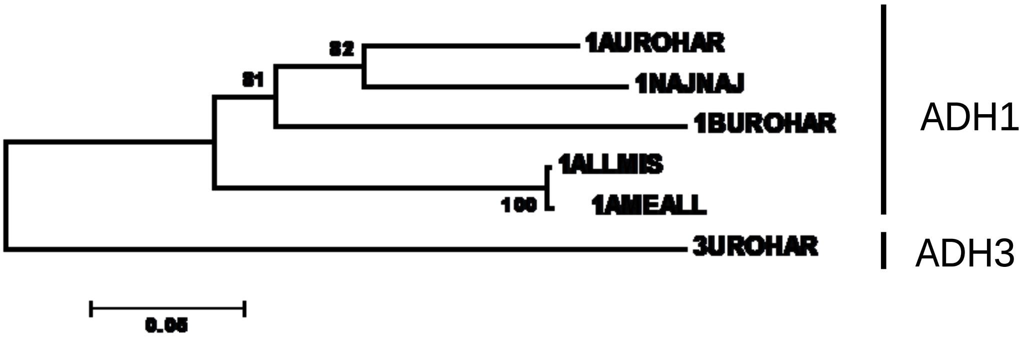

82
1AUROHAR
81
1NAJNAJ
1BUROHAR
1ALLMIS
100
1AMEALL
3UROHAR
ADH1
ADH3
0.05

Fig. 3S

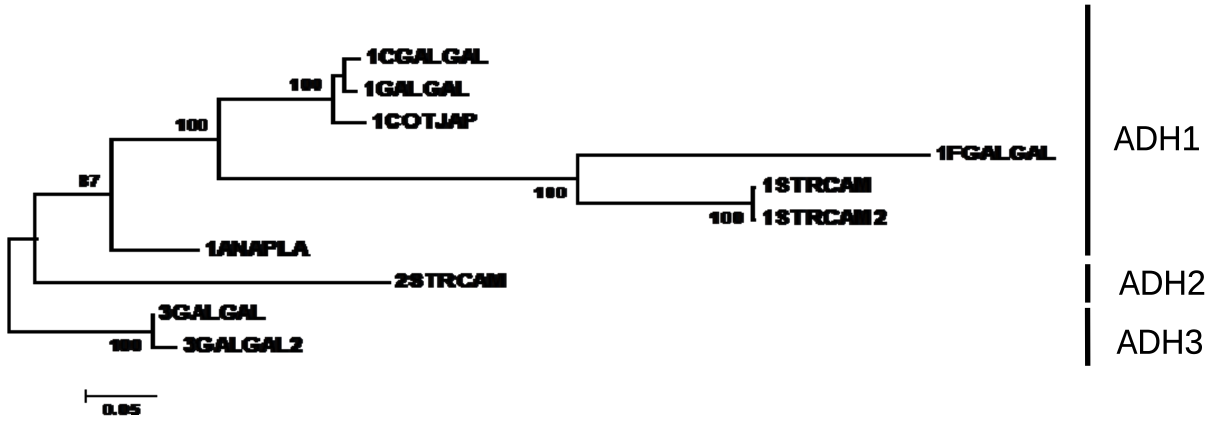

1CGALGAL
100
1GALGAL
100
1COTJAP
87
1FGALGAL
100
1STRCAM
100
1STRCAM2
1ANAPLA
2STRCAM
3GALGAL
100
3GALGAL2
0.05
ADH1
ADH2
ADH3

Fig. 4S

CAPÍTULO 6

DISCUSSÃO

O entendimento da diversidade funcional de famílias gênicas sempre foi um dos maiores tópicos de interesse da área de evolução molecular (Nei, 1987). Recentemente, sua importância para a Genômica Funcional tem sido reconhecida (Bork e Koonin, 1998; Henikoff *et al.*, 1997). O sucesso de projetos objetivando desvendar o genoma de diferentes organismos depende da habilidade em lidar com a informação resultante de forma a realizar a predição de funções. Somente uma pequena parcela dos genes sequenciados tem sido estudada experimentalmente. À medida que aumenta o número de sequências disponíveis em banco de dados, aumenta também a necessidade de caracterizá-las funcionalmente (Bork e Koonin, 1998). Assim, abordagens computacionais tornam-se essenciais em virtude de sua maior rapidez quando comparadas aos experimentos de bancada.

Atualmente, existem dois principais gargalos que necessitam ser superados para que predições funcionais eficientes possam ser implementadas na análise de sequências protéicas. O primeiro deles refere-se à falta de um sistema amplamente aceito, robusto e continuamente atualizado de métodos de análise de sequências integrados a um sistema de predição funcional eficiente. O outro ponto a ser considerado é o excessivo "ruído" na apresentação de dados experimentais, levando a descrições funcionais insuficientes ou errôneas nos bancos de dados de sequências.

A produção de genes duplicados ocorre através de processos de duplicação envolvendo grandes partes do genoma ou eventos de duplicação gênica *in tandem*. Adicionalmente, novos tipos de proteínas com múltiplos domínios podem ser gerados através de um mecanismo de rearranjo dos domínios (*domain-shuffling*). Assim, em consequência desses processos de duplicação de genes e genomas e rearranjo de domínios, muitos genes estão representados como vários parálogos no genoma, com funções relacionadas, mas não idênticas (Gu, 1999).

O surgimento de famílias gênicas propicia inovações funcionais, de forma que a identificação de sítios de aminoácidos responsáveis por essa diversificação é de fundamental importância e possui grande potencial para a Genômica Funcional porque é custo-efetivo e pode ser testada experimentalmente. Para os evolucionistas moleculares é desejável conhecer o nível de divergência funcional após a duplicação de genes ou genomas, bem como saber quantas substituições de aminoácidos estão realmente envolvidas em alterações funcionais. Uma vez que grande parte das mutações não possui impacto funcional, pois representam evolução neutra, é crucial desenvolver métodos estatísticos apropriados para distinguir entre essas duas possibilidades (Gu, 1999). De fato, algumas abordagens computacionais, tais como a busca por homologia e o alinhamento múltiplo, podem não ser suficientes para solucionar determinado problema. Isso ocorre especialmente quando as diferenças de aminoácidos entre cópias de genes duplicados são resultantes de uma duplicação gênica antiga ou de um evento rápido e recente de divergência (Golding e Dean, 1998).

O objeto de estudo dessa Tese foi a família gênica da álcool desidrogenase. A ADH é uma enzima classicamente caracterizada como glicolítica, mas que tem sido associada ao metabolismo da noradrenalina, dopamina, serotonina e ácido biliar (Höög *et al.*, 2001), bem como oxidação de retinol (Boleda *et al.*, 1993; Martras e tal., 2004). Em plantas, estudos experimentais indicam envolvimento no crescimento do tubo polínico (Bucher *et al.*, 1995), em pistilos polinizados (Van Eldik *et al.*, 1997) e em detoxificação de sementes (Garabagi *et al.*, 2005), sendo que sua expressão aumenta significativamente em situações de estresse, tais como desidratação, baixas concentrações de oxigênio e temperaturas (Dolferus *et al.*, 1994).

Assim como ocorre em outras famílias de proteínas, uma enorme quantidade de sequências de ADH tanto de DNA quanto de proteínas encontra-se disponível em bancos de dados. Essa região tem sido extensivamente utilizada como marcador molecular em estudos que objetivam estabelecer as relações filogenéticas entre organismos, especialmente em plantas.

Em humanos, variantes da classe I de ADH têm sido associadas a um efeito de proteção ao alcoolismo (Osier *et al.*, 2002). Elevada expressão de ADH classe III tem sido associada a diversos tipos de câncer. Em pacientes com câncer pancreático, credita-se essa expressão aumentada à liberação dessa enzima por células cancerosas (Jelski e Zalewski, 2008). Em pacientes homozigotos para ADH3 foi comprovado um risco aumentado de adenocarcinomas de esôfago (Terry *et al.*, 2007). É possível que a ADH contribua para o dano causado pelo *Helicobacter pylori* na mucosa gástrica. Esse organismo está fortemente associado à gastrite crônica, úlcera e adenocarcinoma. (Cornally *et al.*, 2008). A ação combinada de xantina oxidoredutase e ADH produz espécies reativas de O_2 que causam danos ao DNA contribuindo para a carcinogênese e câncer de mama (Wright *et al.*, 1998). Há evidências de que ADH também gera espécies reativas de oxigênio e óxido nítrico em neurônios humanos (Haorah *et al.*, 2008).

Acredita-se que essa enzima possa ser candidata, em estudos de quimioterapia, potencialmente como agente anti-giardíase (Dan e Wang, 2000) e anti-candidíase (Swoboda *et al.*, 1993). Há um considerável interesse em álcool desidrogenases estáveis para uma ampla gama de aplicações nas indústrias de alimentos, farmacêuticas e de química fina. Representam elas um importante grupo de biocatalisadores, já que possuem a capacidade de reduzir estereoquimicamente compostos carbonilados. Assim podem ser usadas eficientemente na síntese de alcoóis oticamente ativos (Chen *et al.*, 2008). A produção de dióis é particularmente desejada visto que são importantes blocos de construção de produtos químicos (Haberland *et al.*, 2002). Por exemplo, o (2S,5S)-hexanodiol é utilizado para compostos farmacêuticos. Para obtenção desses compostos, alguns grupos têm utilizado a engenharia de proteínas, alterando a composição de aminoácidos da ADH de microorganismos, tais como *Pyrococcus furiosus*, para aumentar sua atividade catalítica, fornecendo um ambiente eficiente para a catálise dos compostos de interesse (Machielsen *et al.*, 2008). Esses estudos mostram que várias partes da proteína, incluindo a região de ligação ao substrato, sítio de ligação do cofator e regiões perto do sítio ativo, podem resultar em adaptações que permitem o aumento da atividade em baixas temperaturas, provavelmente devido ao aumento na flexibilidade das alças da molécula (Liang *et al.*, 2004; Shiraki *et al.*, 2001; Vieille e Zeikus, 2001). ADH de levedura, fígado de cavalo e *Thermoanaerobium brockii* estão disponíveis comercialmente (Chen *et al.*, 2008). Em plantas, foi demonstrado que a introdução de haloalcano dehalogenase juntamente com ADH endógena em plantas de tabaco cria uma via completa para a degradação de 1,2-DCA (haloalcano 1,2-dicloroetano), usado para a produção de herbicidas e desinfetantes (Mena-Benitez *et al.*, 2008).

Acima foi descrito como estudos computacionais contribuem para o desenvolvimento da Genômica Funcional, com especial aplicação na análise de divergência funcional no entendimento da evolução e diversificação da função de genes duplicados de determinada família protéica. Foram mostrados vários exemplos do papel da ADH em diversos processos, evidenciando a importância de seu estudo tanto para finalidades médicas quanto industriais. A identificação de resíduos funcionalmente importantes através de estudos combinados de evolução molecular e biologia estrutural é uma poderosa ferramenta para fornecer a localização de resíduos de aminoácidos que podem servir como potenciais alvos de engenharia de proteínas e desenho racional de fármacos e compostos de interesse na indústria farmacêutica ou de química fina. A seguir os principais resultados obtidos nessa Tese serão avaliados.

No Capítulo 3 foram analisadas 176 sequências (94 monocotiledôneas, 75 dicotiledôneas e sete gimnospermas) de DNA e proteína de álcool desidrogenases, pertencentes a sete famílias botânicas (Poaceae, Cyperaceae, Arecaceae, Fabaceae, Brassicaceae, Paeoniaceae e Pinaceae). Os resultados da análise filogenética mostraram que houve duplicações recentes em Paeoniaceae e Cyperaceae, indicando que *Adh1* e *Adh2* são mais relacionadas uma a outra do que a outras sequências de dicotiledôneas ou monocotiledôneas. Portanto, tais duplicações devem ter ocorrido após a diversificação dessas famílias botânicas. A família Arecaceae formou um grupo basal dentro das monocotiledôneas, corroborando os dados obtidos por Borsch *et al.* (2003) usando o marcador *TrnT-TrnF*. As sequências de Poaceae formaram dois clados, correspondendo a *Adh1* e *Adh2*. Um terceiro clado (*Adh3*) parece ter surgido de um evento de duplicação de *Adh2*. Tais resultados concordaram com aqueles obtidos por Mathews e Sharrock (1996) usando o gene do fitocromo. Com relação à Brassicaceae, a distribuição dos ramos e, portanto, a relação entre as espécies, foi bastante similar à encontrada utilizando-se sequências de ITS (Johnston *et al.*, 2005). Como conclusão geral, observa-se que diversos eventos independentes de duplicação aconteceram em três linhagens primárias (monocotiledôneas, dicotiledôneas e gimnospermas), podendo ter sido seguidos por pseudogenização e eventuais deleções.

Apesar de ter sido detectada, em estudo realizado por Gaut e colaboradores (1999), diferença nas taxas de substituições não-sinônimas entre *Adh1* e *Adh2* de plantas, nenhuma evidência estatisticamente significante de atuação da seleção positiva foi encontrada. Tal fato não descarta a diversificação funcional de sítios específicos de aminoácidos, uma vez que os testes de seleção positiva detectam um conjunto limitado de eventos de seleção. Assim, os estudos de diversificação funcional, apresentados no Capítulo 3, identificaram doze resíduos de aminoácidos como funcionalmente importantes entre os genes duplicados *Adh1* e *Adh2* e vinte e quatro comparando esta mesma enzima em diferentes famílias botânicas. Esses resultados foram obtidos para um $Q(k) \geq 0,85$. Quando um valor mais restritivo ($Q(k) \geq 0,90$) foi aplicado encontraram-se três sítios divergentes entre as cópias gênicas e sete entre as famílias botânicas. Os sítios importantes foram mapeados na estrutura tridimensional de ADH de *Arabis blepharophylla* obtida utilizando-se modelagem comparativa a partir da estrutura resolvida de ADH de *Equus caballus*. Considerando-se os aminoácidos divergentes entre as cópias gênicas, observa-se que Cys105, Val108 e Asp114 foram localizados em uma alça próxima ao segundo átomo de zinco, e Lys 230, Phe232 e Val275 estão próximos ao sítio ativo. Quando consideramos a divergência entre as famílias botânicas verifica-se que Cys47 e Glu55 estão adjacentes ao sítio ativo; Val108, Asp114, Glu121, Gly123, Gly124 e Ile126 estão em alça próxima ao segundo átomo de zinco; e Leu294, Ser299 e Thr306 localizam-se no segmento de interação subunidade-subunidade. Portanto, tais resíduos estão localizados em regiões de reconhecida importância funcional da proteína. Esses resultados indicam que a seleção natural é um importante fator na evolução da álcool desidrogenase em plantas.

Objetivando realizar uma análise mais detalhada do impacto da diversificação funcional na estrutura tridimensional da álcool desidrogenase em plantas foram deduzidas

17 estruturas de ADH pertencentes às famílias botânicas Poaceae, Brassicaceae, Fabaceae e Pinaceae (Capítulo 4). Resultados indicaram que as Brassicaceae possuem as proteínas com maior carga negativa, sendo que as regiões do sítio ativo, segundo átomo de zinco e segmento de interação entre as subunidades mostram as maiores diferenças. As formas ADH1 das Poaceae são mais básicas que as ADH2 da mesma espécie, sendo que diferenças significativas no potencial eletrostático encontram-se próximas ao segundo átomo de zinco e no segmento de interação entre os monômeros. Com relação às Fabaceae, a ADH de *Lotus corniculatus* possui uma concentração diferenciada de cargas negativas na região do segmento de interação entre subunidades quando comparada com outras ADH da mesma família. A ADH de *Pinus banksiana* possui uma concentração de cargas negativas no sítio ativo. Considerando os resíduos com $Q(k) \geq 0{,}85$, a região de interação entre as subunidades é aquela que apresenta os resíduos mais importantes funcionalmente em Brassicaceae, Fabaceae e Poaceae, sendo que os sítios presentes em hélices e alças próximos ao átomo de zinco são variáveis em Fabaceae. A zona de interação dos dímeros é importante na diferenciação de Poaceae e Pinaceae. Alguns resíduos próximos ao sítio de ligação da coenzima também são fundamentais para a diversificação de ADH nas famílias.

A evolução da álcool desidrogenase em animais e fungos, juntamente com sequências de plantas que já haviam sido utilizadas nos manuscritos anteriores, foi avaliada no Capítulo 5. A análise filogenética de 192 sequências mostrou a ocorrência de três grupos monofiléticos, correspondendo a esses três tipos de organismos, confirmando os resultados obtidos por Glasner *et al.* (1995). Duas sequências de *C. elegans* localizaram-se próximas às de fungos, já que também são ADHs tetraméricas. ADH3, correspondente à formaldeído desidrogenase dependente de glutationa, formou um grupo monofilético com grande suporte de *bootstrap*, incluindo sequências de vertebrados e invertebrados. Um grande grupo de ADH1 foi formado. A ADH classe I de peixes apresenta características mistas, já que é estruturalmente similar à classe III e funcionalmente parecida com a classe I. Tal condição pode explicar porque sequências de ADH dos Actinopterygii formaram um *cluster* separado das outras formas de classe I. Essa ADH parece ter surgido cedo na evolução dos vertebrados. De fato, quando foi encontrada a primeira sequência de ADH de animais dessa classe mostrando características mistas, foi possível estabelecer que a duplicação responsável pela separação entre as ADH de classes I e III deve ter ocorrido há 500 milhões de anos atrás (Cañestro *et al.*, 2002). As ADH4 de mamíferos ficaram próximas as de ADH1, o que não é surpreendente, uma vez que são estruturalmente similares. ADHs de classe II são encontradas nas linhagens de mamíferos e aves formando um grupo irmão de ADH3, reforçando os resultados de Hjelmqvist *et al.* (1995). As ADH8 de anfíbios formam um grupo distinto, confirmando suas características especiais, tais como uma grande cavidade no sítio ativo e rearranjos específicos no sítio de ligação do cofator (Rosell *et al.*, 2003).

Verificou-se que, em mamíferos, as sequências formaram *clusters* diferentes de acordo com o tipo de ADH, enquanto que em fungos houve uma disposição de acordo com o gênero do organismo e o tipo de ADH. Situação semelhante à de mamíferos ocorreu em outros grupos de vertebrados. Assim como aconteceu em plantas, não foi identificado um $\omega > 1$, não tendo sido encontrada evidência de atuação da seleção positiva. Há a possibilidade de que o teste aqui aplicado não tenha sido sensível o suficiente para detectar eventos adaptativos nos genes *Adh*, mas é também possível que os *loci Adh* estejam submetidos a eventos pontuais de substituições não-sinônimas. Levando em consideração que uma única modificação de aminoácido em uma enzima pode ter grande impacto na estrutura e função protéica, foi conduzida uma análise de divergência funcional, que identificou diversos resíduos potencialmente importantes para a função da enzima.

Durante a evolução das sequências, um resíduo de aminoácido pode mudar de muito conservado para altamente variável e vice-versa. A divergência funcional depois da duplicação gênica pode causar mudanças nas taxas evolutivas de alguns sítios. Portanto, detectar alterações nas taxas sítio-específicas pode fornecer uma lista de resíduos de aminoácidos que seriam responsáveis pela divergência funcional entre os membros de uma família gênica.

A álcool desidrogenase é composta por dois domínios e todas as proteínas dessa família possuem similar enovelamento. O domínio de ligação do nucleotídeo é formado por um motivo conhecido como *Rossman fold* (Lesk, 1995) , consistindo de fitas beta paralelas e alfa hélices. O domínio catalítico contém resíduos envolvidos na ligação do substrato e um átomo de zinco localizado na cavidade entre os dois domínios. Em mamíferos e plantas, a ADH é uma enzima dimérica, já em fungos e nas ADH1 e 2 de C. *elegans* elas são tetraméricas. Os resíduos divergentes nesses organismos localizaram-se nas mesmas três regiões onde aminoácidos funcionalmente importantes foram identificados nas plantas: posição do segundo átomo de zinco, região próxima ao sítio ativo e segmento de interação entre monômeros. É importante ressaltar que o zinco parece ser importante para a catálise e estabilização da geometria do sítio ativo. Adicionalmente, estudos focados na classe I indicaram que a ligação da coenzima com um monômero de ADH induz a aproximação do domínio catalítico ao domínio de ligação da coenzima do outro monômero. Essa última conformação é descrita como "fechada", os dímeros mantendo-se abertos na ausência da coenzima promovem a ligação de pequenos alcóois, diminuindo a probabilidade de ligação de alcoóis maiores (Sanghani *et al.*, 2000) (Persson *et al.*, 1993, Danielsson *et al.*, 1994).

O impacto dos resíduos de aminoácidos identificados como potencialmente funcionais em fungos, plantas e animais precisa ser investigado em mais detalhe. Estudos futuros de docking e dinâmica molecular podem ajudar a elucidar o papel de substituições nas posições positivamente relacionadas à divergência. Tais procedimentos já foram conduzidos para a ADH de *Equus caballus* (código PDB: 1N8K), a qual foi utilizada como molde para a modelagem molecular comparativa no presente estudo. Tal estrutura e seus complexos com a coenzima e o substrato mostram dois grupos carboxila próximos, mas não ligados ao zinco do sítio ativo. A carboxila de Asp49 forma uma ligação de hidrogênio ao grupo imidazólico de His67 (um dos grupos que se liga ao zinco), enquanto que Glu68 está localizado atrás do íon, no sentido oposto ao sítio de ligação do substrato (Ganzhorn e Plapp, 1988). Alinhamentos de sequências de aminoácidos indicam que ambos os resíduos são conservados em todas as proteínas álcool desidrogenases de mamíferos, plantas e fungos (Dennis *et al.*, 1985; Chang e Meyerowitz, 1986). Estudos de mutagênese dirigida indicaram que a mutação de Asp49 para Asn49 diminuiu a eficiência catalítica para a oxidação do etanol e acarretou a redução do acetaldeído por um fator de 1000 vezes. Já a mutação do resíduo de aminoácido Glu68 para Gln68 diminuiu a eficiência catalítica por um fator de 100 vezes (Ganzhorn e Plapp, 1988). Aparentemente, a atividade catalítica diminuiu em função do ambiente eletrostático alterado.

Tão importante quanto identificar resíduos divergentes é avaliar quais posições são conservadas entre as proteínas de diferentes organismos. Os sítios Thr/Ser48, His67, Glu68 e Phe140 são compartilhados por todas as ADH, sendo altamente conservados; entretanto, a conservação nessas posições não necessariamente implica em equivalência na especificidade quanto à oxidação do etanol. Um alto grau de conservação é verificado nas ADH3; consequentemente, apresentam-se elas como um grupo monofilético quando são conduzidas análises filogenéticas. Com relação à função biológica, sabe-se que a ADH está envolvida em vias fisiológicas distintas em humanos, tais como no metabolismo da dopamina, serotonina, norepinefrina e ácido biliar. Além disso, ela pode catalisar a oxidação do retinol *in vitro* e *in vivo* e tem sido sugerido que possa exercer uma função de proteção contra a toxicidade da vitamina A. Em camundongos adultos o gene *Adh1* é expresso em altos níveis no fígado e também em níveis significativos no intestino grosso e delgado, rins, adrenal, epidídimo, útero e ovário (Gonzàlez-Duarte e Albalat, 2005). Tal diversidade de funções foi inferida computacionalmente através dos estudos aqui realizados de divergência funcional em vertebrados e invertebrados. De forma similar, demonstramos teoricamente que a álcool desidrogenase de plantas e fungos provavelmente deve estar submetida a processos de neofuncionalização, com os eventos de duplicação que lá ocorreram levando ao surgimento de novas atividades, além de identificar os resíduos de aminoácidos importantes para a diversificação dessas classes de enzimas.

Com relação à evolução molecular dessas enzimas, nossa hipótese inicial foi a de que a ADH está submetida a um processo de evolução por nascimento e morte (Nei e Rooney, 2005), onde novos genes são criados por eventos de duplicação gênica e alguns desses genes duplicados são mantidos no genoma por um longo período de tempo, enquanto outros são deletados ou se tornam não-funcionais. A família gênica da álcool desidrogenase apresenta vários eventos de duplicação, levando a um aumento no número de cópias. Mostrou-se nesse trabalho que tais cópias são fundamentais no processo de diversificação de funções. Alguns aspectos indicam claramente que a aplicação do modelo de nascimento e morte é adequada aos nossos resultados, como indicado por (Longhurst *et* *al.*, 1994): (1) vários eventos de duplicação gênica ocorreram ao longo do tempo; (2) os loci da mesma espécie são mais distantemente relacionados entre si do que a seus correspondentes em espécies diferentes; e (3) há evidências de que alguns genes após sua duplicação tornam-se pseudogenes.

O estudo da diversificação funcional das macromoléculas é de suma importância para uma melhor compreensão da diversidade biológica. A evolução de genomas, a origem de novos genes e os mecanismos responsáveis pela aquisição de novas funções são essenciais para o entendimento das adaptações bioquímicas. Esses mecanismos incluem os moleculares que geram a diversidade e os processos evolutivos que fixam as mutações e aperfeiçoam as funções. Como a redundância é mantida, e em que extensão afeta a viabilidade de um organismo são questões intrigantes. A avaliação de um maior número de famílias gênicas e o entendimento das relações entre as proteínas codificadas pelas mesmas são necessários a fim de começarmos a entender os processos que geram a diversidade biológica que conhecemos.

REFERÊNCIAS BIBLIOGRÁFICAS

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al.* (2005) The universal protein resource (UniProt). Nucleic. Acids Res. 33D: 154-159.

Benítez-Páez A e Cárdenas-Brito S (2008) Dissection of functional residues in receptor activity-modifying proteins through phylogenetic and statistical analysis. Evol. Bioinf. 4:153-169.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov HI e Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res. 28: 235–242.

Blattner FR, Plunkett GR, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B e Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–74.

Boleda MD, Saubi N, Farrés J, Parés X (1993) Physiological substrates for rat alcohol dehydrogenase classes: aldehydes of lipid peroxidation, omega-hydroxyfatty acids, and retinoids. Arch. Biochem. Biophys. 307: 85-90.

Bork P e Koonin EV (1998) Predicting functions from protein sequences – where are the bottlenecks? Nat. Genet. 18: 313-318.

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C e Barthlott W (2003) Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. J. Evol. Biol. 16: 558-576.

Bowie JU, Luthy R e Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253: 164-70.

Brandon SG, Morton BR, Mccaig BC e Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL.* Proc. Natl. Acad. Sci. USA 83: 10274–10279.

Brown DD, Wensink PC e Jordan E (1972) *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J. Mol. Biol. 63: 57-73.

Bucher M, Brander KA, Sbicego S, Mandel T, Kuhlemeier C (1995) Aerobic fermentation in tobacco pollen. Plant. Mol. Biol. 28: 739-750.

Burglin TR (1997) Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. Nucleic Acids Res. 25: 4173-4180.

Casari G, Sander C e Valencia A (1995) A method to predict functional residues in proteins. Nat. Struct. Biol. 2: 171-178.

Cañestro C, Albalat R, Hjelmqvist L, Godoy L, Jörnvall H e González-Duarte R (2002) Ascidian and amphioxus Adh genes correlate functional and molecular features of the ADH family expression during vertebrate evolution. J. Mol. Evol. 54: 81-89.

Chang C e Meyerowitz Em (1986) Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA 94: 7791-7798.

Charlesworth D, Liu FL e Zhang L (1998) The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol. Biol. Evol. 15: 552-559.

Chen Q, Hu Y, Zhao W, Zhu C e Zhu B (2008) Cloning, expression, and characterization of a novel (S)-specific alcohol dehydrogenase from *Lactobacillus kefir*. Appl. Biochem. Biotechnol. (versão online, DOI 10.1007/s12010-008-8442-6).

Clark AG (1994) Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA 91: 2950-2954.

Clayton J e Dennis C (2003) 50 years of DNA. Nature Publishing Group, London, UK, 146 pp.

Clegg MT, Cummings MP e Durbin ML (1997) The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. 94: 7791-7798.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CER., Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–44.

Coop A e MacKerell ADJr (2000) The future of opioid analgesics. Amer. J. Pharm. Educ. 66: 153-156.

Cornally D, Mee B, MacDonaill C, Tipton KF, Kelleher D, Windle HJ e Henehan GTM (2008) Aldo-keto reductase from Helicobacter pylori – role in adaptation to growth at acid pH. FEBS J. 275: 3041-3050.

Dan M e Wang CC (2000) Role of alcohol dehydrogenase E (ADHE) in the energy metabolism of *Giardia lamblia*. Mol. Biochem. Parasitol. 109: 25-36.

Dasmahapatra AK, Wang X e Haasch ML (2005) Expression of *Adh8* mRNA is developmentally regulated in japanese medaka (*Oryzias latipes*). Comp. Biochem. Physiol. 140: 657-664.

Danielsson O, Atrian S, Luque T, Hjelmqvist L, Gonzalez-Duarte R e Jörnvall H (1994) Fundamental molecular differences between alcohol dehydrogenase classes. Proc. Natl. Acad. Sci. USA 91: 4980-4984.

Denis ES, Sachs MM, Gerlach WL, Finnegan EJ e Peacock WJ (1985) Molecular analysis of the alcohol dehydrogenase 2 (*Adh2*) gene of maize. Nucleic Acids Res. 13: 727-743.

Dayhoff MO, Schwartz RM e Orcutt BC (1978) A model of evolutionary change in proteins. In: Atlas of protein sequence structure. National Biomedical Research Foundation, Washington, DC. pp 342-352.

Dolferus R, Osterman JC, Peacock WJ e Dennis ES (1997) Cloning of the *Arabidopsis* and rice formaldehyde dehydrogenase genes: implications for the origin of plant ADH enzymes. Genetics 146: 1131-1141.

Dolferus R, Jacobs M, Peacock WJ e Dennis ES (1994) Diferential interactions of promoter elements in stress responses of the Arabidopsis *Adh* gene. Plant Physiol. 105: 1075-1087.

Doyle JJ e Gaut BS (2000) Evolution of genes and taxa: a primer. Plant Mol. Biol. 42: 1-23.

Force A, Lynch M, Pickett FB, Amores A, Yan YL e Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-1545.

Fryxell KJ (1996) The coevolution of gene family trees. Trends Genet. 12: 364-369.

Ganzhorn AJ e Plapp BV (1988) Carboxyl groups near the active site zinc contribute to catalysis in yeast alcohol dehydrogenase. J. Biol. Chem. 263: 5446-5454.

Garabagi F, Duns G, Strommer J (2005) Selective recruitment of *Adh* genes for distinct enzymatic functions in *Petunia* hybrid. Plant Mol. Biol. 58: 283-294.

Gaucher EA, Das UK, Miyamoto MM e Benner SA (2002a) The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. Mol. Biol. Evol. 19: 569-573.

Gaucher EA, Gu X, Miyamoto MM e Benner SA (2002b) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. 27: 315-321.

Gaut BS e Clegg MT (1993) Molecular evolution of the *Adh1 locus* in the genus Zea. Proc. Natl. Acad. Sci. USA 90: 5095-5099.

Gaut BS, Peek AS, Morton BR e Clegg MT (1999) Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. 16(8): 1086-1097.

Gehring WJ e Ikeo K (1999) *Pax6*: mastering eye morpho-genesis and eye evolution. Trends Genet. 15: 371-377.

Georgelis N, Braun EL e Hannah LC (2008) Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants. BMC Evol. Biol. 8:1-10 (versão online).

Glasner JD, Kocher TD e Collins JJ (1995) Caenorhabditis elegans contains genes encoding two new members of the Zn-containing alcohol dehydrogenase family. J. Mol. Evol. 41: 46-53.

Gogarten JP e Lorraine O (1999) Orthologs, paralogs and genome comparisons. Curr. Opin. Genet. Dev. 9: 630-636.

Golding GB e Dean AM (1998) The structural basis of molecular adaptation. Mol. Biol. Evol. 15: 355-369.

Gonzàlez-Duarte R e Albalat R. (2005) Merging protein, gene and genomic data: the evolution of the MDR-ADH family. Heredity 95: 184-97.

Gu J e Gu X (2003) Natural history and functional divergence of protein tyrosine kinases. Gene 317: 49-57.

Gu J, Wang Y e Gu X (2002a) Evolutionary analysis of functional divergence of Jak protein kinase domains and tissue-specific genes. J. Mol. Biol. 54: 725-733.

Gu X (2003) Functional divergence in protein (family) sequence evolution. Genetica 118: 133-141.

Gu X (2001a) Maximum-likelihoood approach for gene family under functional divergence. Mol. Biol. Evol. 18: 453-464.

Gu X (2001b) A site-specific measure for rate difference after gene duplication or speciation. Mol. Biol. Evol. 18: 2327-2330.

Gu X (1999) Statistical method for testing functional divergence after gene duplication. Mol. Biol. Evol. 16: 1664–1674.

Gu X e Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinform. Applic. Note 18: 500–501.

Gu X, Wang Y, Gu J e Vander Velden K (2002b) Evolutionary perspective for functional divergence of gene family and applications in functional genomics. Curr. Genomics 3: 201-211.

Gu Z, Rifkin SA, White KP e Li W-H (2004) Duplicate genes increases gene expression diversity within and between species. Nat. Genet. 36: 577-579.

Haberland J, Kriegesmann A, Wolfram E, Hummel W e Liese A (2002) Diastereoselective synthesis of optically active (2R,5R)-hexanediol. Appl. Microbiol. Biotechnol. 58:595–599.

Haorah J, Ramirez SH, FLoreani N, Gorantla S, Morsey B e Persidsky Y (2008) Mechanism of alcohol-induced oxidative stress and neuronal injury. Free Radic. Biol. Med. 45: 1542-1550.

Henikoff S, Greene EA, Pietrokovski S, Bork P, Atwood TK e Hood L (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278: 609-614.

Hirotsune S, Yoshida N, Chen A, Garret L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A e Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423: 91-96.

Hjelmqvist L, Estonius M e Jörnvall H (1995) The vertebrate alcohol dehydrogenase of class III: consistent patterns of structural and functional conservation in relation to class I and other proteins. FEBS Lett. 373: 212-216.

Hoffmann JL, Torontali SP, Thomason RG, Lee DM, Brill JL, Price BB, Carr GJ e Versteeg DJ (2006) Hepatic gene expression profiling using genechips in zebrafish exposed to 17α-ethynylestradiol. Aquat. Toxicol. 79: 233-246.

Höög JO, Hedberg JJ, Stromberg P, Svesson S (2001) Mammalian alcohol dehydrogenase – functional and structural implications. J. Biomed. Sci. 8: 71-76.

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B. Biol. Sci. 256: 119-124.

Jelski W e Zalewski B (2008) Alcohol dehydrogenase (ADH) isoenzymes and aldehyde dehydrogenase (ALDH) activity in the sera of patients with pancreatic cancer. Dig. Dis. Sci. 53; 2276-2280.

Johnston JS *et al.* (2005) Evolution of the genome size in Brassicaceae. Ann. Bot. 95: 229-235.

Jordan IK, Bishop GR e Gonzalez DS (2001) Sequence and structural aspects of functional diversification in class I α-mannosidase evolution. Bioinform. 17: 965-976.

Jones DT, Taylor WR e Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8: 275-282.

Jörnvall H (1970) Horse liver alcohol dehydrogenase: the primary structure of the protein chain of the ethanol-active isoenzyme. Eur. J. Biochem. 16:25-40.

Kedishvili NY, Gough WH, Chernoff EA, Hurley TD, Stone CL, Bowman KD, Popov KM, Bosron WF e Li TK (1997) cDNA sequence and a catalytic properties of a chick embryo alcohol dehydrogenase that oxidizes retinol and 3beta,5alpha-hydroxysteroids. J. Biol. Chem. 272: 7494-7500.

Klenk HP, Clayton RA, Tomb EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B e Venter JC (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390: 364–70.

Landgraf R, Fischer D e Eisenberg D (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng. 12: 943-951.

Laskowski RA, McArthur MW, Moss DS e Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Crystalogr. 26: 283-291.

Lesk A (2008) Estrutura de proteínas e descoberta de fármacos. In: Introdução à bioinformática. Artes Médicas, Porto Alegre. 384 p.

Lesk AM (1995) NAD-binding domains of dehydrogenases. Curr. Opin. Struct. Biol. 5: 775-783.

Li WH, Gu Z, Wang H e Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409: 847–849.

Liberles DA (2001) Evaluation of methods for determination of a reconstructed history of gene sequence evolution. Mol. Biol. Evol. 18: 2040-2047.

Lichtarge O, Bourne HR e Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257: 342-358.

Liang ZX, Tsigos I, Bouriotis V e Klinman JP (2004) Impact of protein felexibility on hybride-transfer parameters in thermophilic and psychrophilic alcohol dehydrogenases. J. Am. Chem. Soc. 126: 9500-9501.

Liu Q, Dou S, Wang G, Li Z e Feng Y (2008) Evolution and functional divergence of monocarboxylate transporter genes in vertebrates. Gene 423: 14-22.

Longhurst T, Lee E, Hinde R, Brady C e Speirs J (1994) Structure of the tomato *Adh2* gene and *Adh2* pseudogene, and a study of *Adh2* gene expression in fruit. Plant Mol. Biol. 26: 1073-1084.

Lüthy R, Bowie JU e Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356: 83-85.

Machielsen R, Leferink NGH, Hendriks A, Brouns SJJ, Hennemann H, Dauβmann T e van der Oost J (2008) Laboratory evolution of *Pyrococcus furiosus* alcohol dehydrogenase to improve the production of (2S,5S)-hexanediol at moderate temperatures. Extremophiles 12: 587-594.

Mano S e Innan H (2008) The evolutionary rate of duplicated genes under concerted evolution. Genetics 180: 493-505.

Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F e Sali A (2000) Comparative protein structure modelling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29: 291-325.

Martras S, Alvarez R, Martinez SE, Torres D, Gallego O, Duester G, Farrés J, de Lera AR, Parés X (2004) The specificity of alcohol dehydrogenase with *cis*-retinoids. Activity with 11-*cis*-retinol and localization in retina. Eur. J. Biochem. 271: 1660-1670.

Mathews S e Sharrock RA (1996) The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. Mol. Biol. Evol. 13: 1141-1150.

Mena-Benitez GL, Gandia-Herrero F, Graham S, Larson TR, McQueen-Mason SJ, French CE, Rylott EL e Bruce NC (2008) Engineering a catabolic pathway in plants for the degradation of 1,2-dichloroethane. Plant Physiol. 147: 1192-1198.

Mercereau-Puijalon O, Barale JC e Bischoff E (2002) Three multigene families in *Plasmodium* parasites: facts and questions. Int. J. Parasitol. 32: 1323-1344.

Milenkovic VM, Langman T, Schreiber R, Kunzelmann K e Weber BHF (2008) Molecular evolution and functional divergence of the bestrophin protein family. BMC Evol. Biol. 72: 1-10 (versão online).

Morris AL, Macarthur MW, Hutchinson EG e Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins 12: 345-364.

Morton BR, Gaut BS e Clegg MT (1996) Evolution of alcohol dehydrogenase genes in the Palm and Grass families. Proc. Natl. Acad. Sci. USA 93: 11735-11739.

Nagylaki T (1984) The evolution of multigene families under intrachromosomal gene conversion. Genetics 106: 529-548.

Naylor GJP e Gerstein M (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J. Mol. Evol. 51: 223-233.

Nei M (2005) Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22: 2318-2342.

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York. 521 pp.

Nei M, Gu X e Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA 94: 7799-7806.

Nei M e Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: Tsuji K, Aizawa M, Sasazuki T (eds) 11th Histocompatibility workshop and conference. Oxford Univ. Press, Oxford, UK, pp 27-38.

Nei M e Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39: 121-152.

Niimura Y e Nei M (2003) Evolution of olfactory receptor genes in the human genome. Proc. Natl. Acad. Sci. USA 100: 12235-12240.

Ohta T (2003) Gene families: multigene families and superfamilies. Encyclopedia of the human genome (on-line).

Ohta T (1987) Simulating evolution by gene duplication. Genetics 115: 207-213.

Ohta T (1985) Variances and covariances of identity coefficients of a multigene family. Proc. Natl. Acad. Sci. USA 82: 829-833.

Ohta T (1984) Some models of gene conversion for treating the evolution of multigene families. Genetics 106: 517-528.

Ohta T e Dover GA (1983) Population genetics of multigene families that are dispersed into two or more chromosomes. Proc. Natl. Acad. Sci. USA 80: 4079-4083.

Osier MV, Pakstis AJ, Soodyall H, Comas D, Goldman D, Odunsi A, Okonofua F, Parnas J, Schulz LO, Bertranpetit J, Bonne-Tamir B, Lu RB, Kidd JR e Kidd KK (2002) A global perspective on genetic variation at the *Adh* genes reveals unusual patterns of linkage disequilibrium and diversity. Am. J. Hum. Genet. 71: 84-99.

Perovic D, Tiffin P, Douchkov D, Bäumlein H e Graner A (2007) An integrated approach for the comparative analysis of a multigene family: the nicotianamine synthase genes of barley. Funct. Integr. Genomics 7: 169-179.

Person B, Bergman T, Keung WM, Waldenström U, Holmquist B, Vallee BL e Jörnvall H (1993) Structural and functional divergence of class II alcohol dehydrogenase – cloning and characterization of rabbit liver isoforms of the enzyme. Eur. J. Biochem. 216: 49-56.

Persson B, Hedlund J e Hörnvall H (2008) The MDR superfamily. Cell. Mol. Life Sci. 65: 3879-3894.

Podlaha O e Zhang J (2004) Nonneutral evolution of the transcribed pseudogene *Makorin1-p1* in mice. Mol. Biol. Evol. 21: 2202-2209.

Pollock D, Taylor WR e Goldman N (1999) Coevolving protein residues: maximum likelihood identification and relationships to structure. J. Mol. Biol. 287: 187-198.

Rosell A, Valencia E, Páres X, Fita I, Farrés J e Ochoa WF (2003) Crystal structure of the vertebrate NADP(H)-dependent alcohol dehydrogenase (ADH8). J. Mol. Biol. 330: 75-85.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini, ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski

MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH e Lewis S (2000) Comparative genomics of the eukaryotes. Science 287: 2204–15.

Sali A e Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. 243: 779-815.

Sánchez R e Sali A (2000) Comparative protein structure modeling: introduction and practical examples with MODELLER. In: Protein structure prediction: methods and protocols. Humana Press, New Jersey, 415 p.

Sánchez R e Sali A (1997) Advances in comparative protein-structure modelling. Curr. Opin. Struct. Biol. 7: 206–214.

Sanghani PC, Stone CL, Ray BD, Pindel EV, Hurley TD e Bosron WF (2000) Kinetic mechanism of human glutathione-dependent formaldehyde dehydrogenase. Biochem. 39: 10720-10729.

Santamaria M, Lanave C e Saccone C (2004) The evolution of the adenine nucleotide translocase family. Gene 333: 51–59.

Shiraki K, Nishikori S, Fujiwara S, Hashimoto H, Kai Y, Takagi M e Imanaka T (2001) Comparative analyses of the conformational stability of a hyperthermophilic protein and its mesophilic counterpart. Eur. J. Biochem. 268: 4144-4150.

Small RL e Wendel JF (2000a) Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155: 1913-1926.

Small RL e Wendel JF (2000b) Phylogeny, duplicaton and intraspecific variation of *Adh* sequences in New World diploid cottons (*Gossypium L.*, Malvaceae). Mol. Phylogenet. Evol. 16: 73-84.

Spaethe J e Briscoe AD (2004) Early duplication and functional diversification of the opsin gene family in insects. Mol. Biol. Evol. 21: 1583-1594.

Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA e Wellems TE (1995) The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell 82: 89–100.

Suiter KA (1998) Genetics of allozyme variation in *Gossypium arboretum* L. and *Gossypium herbaceum* L. (Malvaceae). Theor. Appl. Genet. 75: 259 – 271.

Szalai G, Duester G, Friedman R, Jia H, Lin S, Roe BA e Felder MR (2002) Organization of six functional mouse alcohol dehydrogenase genes on two overlapping bacterial artificial chromosomes. Eur. J. Biochem. 269: 224-232.

Swoboda RK, Bertram G, Hollander H, Greenspan D, Greenspan JS, Gow NAR, Gooday GW e Brown AJP. Glycolytic enzymes of *Candida albicans* are nonubiquitous immunogens during candidiasis. Infect. Immun. 61: 4263-4271.

Terry MB, Gammon MD, Zhang FF, Vaughan TL, Chow W-H, Risch HA, Schoenberg JB, Mayne ST, Stanford JL, West AB, Rotterdam H, Blot WJ, Fraumeni Jr, JF e Santella RM (2007) *Alcohol dehydrogenase* 3 and risk of esophageal and gastric adenocarcinomas. Cancer Causes Control 18: 1039-1046.

Thompson CE, Salzano FM, Norberto de Souza O e Freitas LB (2007) Sequence and structural aspects of the functional diversification of plant alcohol dehydrogenases. Gene 396:108-115.

Van Eldik GJ, Ruiter RK, Van Herpen MMA, Schrauwen JAM, Wullems GJ (1997) Induced ADH gene expression and enzyme activity in pollinated pistils of *Solanum tuberosum*. Sex Plant Reprod 10: 107–109.

Vieille C e Zeikus GJ (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65: 1-43.

Wagner A (2002) Selection and gene duplication: a view from the genome. Genome Biol. 3: 1012.1-1012.3.

Wang YF e Gu X (2001) Functional divergence in caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158: 1311-1320.

Wang HY, Wang IF, Bose J e Shen CKJ (2004) Structural diversity and functional implications of the eukaryotic TDP gene family. Genomics 83: 130-139.

Wright RM, McManaman JL e Repine JE (1998) Alcohol-induced breast cancer: a proposed mechanism. Free Radic. Biol. Med. 26: 348-354.

Yokoyama S, Yokoyama R, Kinlaw CS e Harry DE (1990) Molecular evolution of the zinc-containing long-chain alcohol dehydrogenase genes. Mol. Bio. Evol. 7: 143-154.

Zheng Y, Xu D e Gu X (2007) Functional divergence after gene duplication and sequence-structural relationship: a case study of G-protein alpha subunits. J. Exp. Zool. (Mol. Dev. Evol.) 308B: 85-96.

Zhou H, Gu J, Lamont SJ e Gu X (2007) Evolutionary analysis for funtional divergence of the toll-like receptor gene family and altered functional constraints. J. Mol. Evol. 65: 119-123.

What limit can be put to this power, acting during long ages and rigidly scrutinising the whole constitution, structure, and habits of each creature, - favouring the good and rejecting the bad? I can see no limit to this power, in slowly and beautiful adapting each form to the most complex relations of life"

(Charles Darwin)