UNIVERSIDADE ESTADUAL DE CAMPINAS

José Pedro Fonseca

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE GENES POTENCIALMENTE TRANSFERIDOS HORIZONTALMENTE NO GENOMA DO FITOPATÓGENO *C. PERNICIOSA*, CAUSADOR DA DOENÇA "VASSOURA DE BRUXA" NO CACAUEIRO.

Tese apresentada ao Instituto de Biologia para obtenção do Título de Mestre em Genética e Biologia Molecular na área de Genética e Microorganismos.

Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira

Campinas
2005

Fonseca, Jose Pedro

L628a Identificação e caracterização de genes potencialmente transferidos horizontalmente no genoma do fitopatógeno *C. perniciosa*, causador da doença "vassoura de bruxa" no cacaueiro./ Jose Pedro Fonseca. Campinas, SP : [s.n.], 2005.

65f: ilus

Orientador : Goncalo Amarante Guimarães Pereira

Dissertação (Mestrado) Universidade Estadual de Campinas. Instituto de Biologia.

1. Transferência Horizontal de Genes (HGT) 2. Bioinformática
3. Vassoura de Bruxa.
I- Universidade Estadual de Campinas. Instituto de Biologia.

Campinas, 7 de Março de 2005.

Banca Examinadora

Prof. Dr. Gonçalo Amarante Guimarães Pereira (Orientador)

_____

Assinatura

Prof. Dr. Michel Georges Albert Vincentz _____

Assinatura

Prof. Dr. Sérgio Furtado dos Reis _____

Assinatura

Prof. Dr. Marcelo Brocchi _____

Assinatura

Agradecimentos

Gostaria de agradecer primeiramente ao meu orientador, Gonçalo, pela chance de trabalhar em seu laboratório, pelos conselhos, paciência e pela confiança depositada em mim ao longo dos dois últimos anos. Espero que esta seja somente a primeira de muitas parcerias vindouras.

Gostaria de agradecer ao Prof. Michel Vincentz pelo apoio no inicio e durante meu mestrado na UNICAMP, apoio este que perdurou durante todo o mestrado e pelas conversas, sempre esclarecedoras sobre o trabalho.

Ao Professor de economia José Maria Silveira pela ajuda com análises de estatística.

Agradeço também aos colegas bioinformatas, especialmente o Marcelo pelas parcerias e debates ao longo do mestrado.

Agradeço aos diversos colegas de labroratorio que estiveram presentes ao longo de dois anos de trabalho, em especial a Odalys, minha parceira na bancada, quem me introduziu na prática diversas técnicas de biologia molecular, Carol e Ana Deckman que sempre estavam lá para ajudar e todos os outros membros do LGE.

Por fim, gostaria imensamente de agradecer minha família e minha namorada por tudo, esta tese é para vocês; minha mãe Maria, meu Pai Mário, meu irmão João e minha querida Carolina. Voces sempre torceram por mim e me apoiaram em todos os momentos tornando esta empreitada possível.

"- Esta tese é dedicada á memória de meu querido avô
Bertoldo Derengowski, um grande homem."

Campinas, 09/03/2005

Índice

I. RESUMO

Transferência horizontal de genes (HGT) pode ser definida como a transmissão de genes entre diferentes grupos taxonômicos no qual o gene incorporado ao genoma do organismo receptor pode ser mantido ao longo de sucessivas gerações pelo sistema reprodutivo tradicional do receptor. HGT é reconhecida como uma das principais forças atuantes na evolução de genomas de procariotos, que têm significantes percentagens adquiridas por esse processo (1.5% a 14.5%).

Em eucariotos, entretanto, o papel de HGT começou apenas recentemente a ser avaliado e geralmente é relacionado com algum cenário evolutivo específico como, por exemplo, a relação hospedeiro-patógeno. Espécies de fungos são consideradas especialmente suscetíveis a HGT por seu modo íntimo de associação com hospedeiros. Em vista disso, no presente projeto foi elaborado um protocolo para identificação de potenciais candidatos a HGT no genoma de *Crinipellis perniciosa*, o fungo causador da vassoura de bruxa nos cacauais.

Primeiramente foi criado um banco de dados contendo todas as seqüências putativas de proteínas (ORFs) de fungos disponíveis no banco de dados do NCBI nr (03/2004). Uma montagem-rascunho de seqüências genômicas "shotgun" de *C. perniciosa*, num total de 17.000 contigs, foi então comparada, através de BLASTX, com esse banco de ORFs de fungos. Cerca de 5000 contigs de *C. perniciosa* que apresentaram similaridade com o banco de ORFs de fungos (e-value < e^{-5}) foram eliminados, sendo considerados como contendo genes verticalmente transferidos. As demais 12.000 seqüências foram então comparadas por blastX contra nr, gerando um total de 357 "hits" com alta similaridade (e-value < e^{-5}) com seqüências não encontradas dentro das seqüências de fungos até então depositadas.

Estes 357 contigs foram então analisados pelo algoritmo BLASTP contra o NCBI-nr e destes 43 apresentaram alta similaridade com proteínas ou domínios conservados de organismos evolutivamente distantes a fungos: 16 apresentaram alta similaridade com plantas, 19 com bactérias e outros 09 com vertebrados/invertebrados.

Entre os genes putativos mais interessantes destacam-se o gene "Thaumatin" (THN) e Transposase; cujas ORFs, à exceção de THN estão completas, não tem introns e apresentam sinais de endereçamento celular. Esses genes foram amplificados a partir do DNA genômico e cDNA da fase necrotrófica de *C. perniciosa*, confirmando assim sua presença no genoma do fungo, sendo inclusive validados através de sequenciamento. Entre os genes candidatos amplificados, Pollen e DAD não apresentaram uma amplificação consistente e análises através de BLASTN contra banco de dados de EST e do genoma do eucalipto indicaram uma possível contaminação. Um gene identificado neste trabalho cujo produto é biologicamente interessante é o gene "thaumatin" (THN), da família de genes PR-5 ("pathogenesis-related"), descrito como uma proteína antifúngica relacionada à patogênese e à resposta sistêmica a patógenos em plantas.

Assim, o presente trabalho gerou uma lista de genes candidatos a terem a sua origem pela transferência horizontal e sua análise poderá permitir a compreensão de processos de interação patógeno hospedeiro.

II. Abstract

Horizontal gene transfer (HGT) can be defined as the mobilization and transmission of DNA across species barriers other than through sexual transmission to the offspring (vertical descent). HGT has long been recognized as a major force affecting the evolution of prokaryotes as significant proportions of such genomes have been subjected to horizontal transfer (1.5% to 14.5%).

The role of HGT in eukaryotes is still poorly demonstrated and only recently some individual cases have been reported, normally linked to some evolutive scenario such as host-pathogen interaction. It is said that fungal species are especially susceptible to HGT for their close associations with their hosts. This is especially the case for fungi with saprophytic mode of feeding that could somehow facilitate HGT. This is also the case in pathogens that share an intracellular environment with its host. We investigate here the possible role of HGT in the basidiomycete, saprophytic fungi *C. perniciosa* and its possible consequences to plant-pathogen interactions in the context of putative acquired HGT genes.

Given the large number of sequences to examine and the expectation that only a few were horizontally transferred we created a stepwise approach based on genomic comparisons in order to reduce the number of HGT candidates through elimination of spurious HGT candidates.

We first created a fungal database containing all fungal sequences available at NCBI at the time of the analysis and all fungal genomes available through the FTP site (NCBI). This fungal dataset was then compared to a draft assembly of *C. perniciosa* shotgun genomic sequences using BLASTX with an e-value cutoff higher than E>-5. This served to eliminate all sequences with high similarities to fungal sequences that are not considered as being subjected to HGT.

Secondly, we made a second round of genomic comparison, this time against the NCBI-nr, containing all protein coding sequences and putative ORFs, through BLASTX with a cutoff BLAST e-value of e<-5. We managed to reduce our sequences from around 17.000 contigs to around 7.000 in the first round of screening and to only 357 contigs after the second round of genomic screening. Those 357 contigs were then examined for putative homologs using BLASTP against NCBI-nr and we selected 43 candidates of HGT showing high similarity to sequences of distantly related organisms, mainly plants and bacteria. We then moved on to perform Phylogenetic and Parametric analysis on those candidates to see if they confirmed HGT. Out of those 43 candidates we found a strong bias between plants and fungi, with 15 HGT candidates being proposed to have been transferred from plants to fungi. Out of those 15 plant candidates we found 6 transposons with high similarities to plants class 1 transposons, suggesting that transposons indeed might play a role in HGT. Two HGT candidates in our list (Transposase and THN) were successfully amplified through PCR and were sequenced. Further analysis suggested the presence of 6 contaminant sequences in the *C. perniciosa* genome from Eucalyptus through BLASTN scores against EST and eucalyptus databases including the putative genes for Pollen and DAD.

III Introdução

A doença da "Vassoura de Bruxa" do cacaueiro

O agente causador da vassoura de bruxa, *C. perniciosa,* é um fungo patógeno Himenomiceto pertencente ao filo Basidiomycota, ordem Agaricales e família Tricholomataceae. A espécie foi primeiramente identificada em 1895 (Stahel 1915) como um patógeno do cacau (*Theobroma cacao*) que causa a doença vassoura de bruxa. Para obter carbono para sua nutrição, Himenomicetos decompõe matéria orgânica morta ou entram em diversos tipos de associação (ambas antagonísticas e benignas) com plantas, animais e outros fungos.

A espécie *C. perniciosa* é subdividida em diferentes biótipos que variam de acordo com o hospedeiro que ataca. O biótipo-C ataca especificamente o cacaueiro, apresentando um ciclo de vida hemibiotrófico, ou seja, durante uma fase do seu ciclo de vida se hospeda em tecido vivo (fase biotrófica), e posteriormente em tecido morto (fase necrotrófica ou saprofítica). *C. perniciosa* ataca obrigatoriamente tecidos jovens, em crescimento, como flores, almofada floral e outros tecidos meristemáticos, causando hipertrofia com a formação de ramificações vegetais (vassouras verdes) após infecção dos "buds" auxiliares ou terminais. Uma espécie de colchão de ramos, similar a uma inflorescência, que se assemelha a estas ramificações vegetais surge quando a almofada floral do cacaueiro é infectada podendo produzir frutos partenocarpicos popularmente conhecidos como "moranguinhos".

Figura 1. Fruto paternocárpico originado a partir da infecção da flor do cacaueiro ou da almofada floral por *C. perniciosa*.

Seu ciclo de vida começa com os basidiocarpos desenvolvendo-se a partir de micélios dicarióticos e saprofíticos em vassouras mortas. Ainda no basidiocarpo, quatro basidiósporos mononucleares são formados através de divisões meióticas resultantes da fusão de núcleos de células dicarióticas. As quatro células resultantes da 2ª divisão meiótica desenvolvem-se em basidiósporos, que irão espalhar-se através do vento e/ou chuva, infectando assim novos tecidos meristemáticos. Basidiospóros germinam uma vez em contato direto com a superfície da planta

saudável, produzindo tubos germinativos monocarióticos que penetram o hospedeiro iniciando assim uma nova infecção (Purdy and Schmidt 1996). Logo após a penetração no hospedeiro o micélio biotrófico ramifica o tecido hospedeiro intercelularmente, e após apenas 4h a 6h está bem distribuído pelo hospedeiro. Hifas não foram visualizadas dentro de células vivas no período biotrófico (Stahel 1915; Calle, Cook et al. 1982). Entretanto, hifas binucleadas com conexões de grampo foram encontradas intracelularmente logo após o início da fase necrotrófica (Calle, Cook et al. 1982).

Após diversas semanas as folhas da até então vassoura verde começam a necrosar, seguido de necrose do caule até a ponta da vassoura. Depois de 3-9 semanas, dicariotização tem inicio e o micélio biotrófico muda para a fase necrotrófica tendo como característica hifas dicarióticas com grampo de conexão. O mecanismo desta mudança da condição nuclear do micélio ainda é desconhecido. Durante esta transição da vassoura verde para a vassoura necrosada, o micélio de ambas as fases pode ser encontrado nos tecidos do hospedeiro.

A doença da Vassoura de Bruxa já causou perdas econômicas expressivas em alguns estados produtores de cacau do Brasil, sendo o caso mais drástico o sul da Bahia, que viu recentemente a exportação de cacau cair em 60%, passando de US$ 700 milhões/ano para cerca de US$ 260 milhões/ano (Pereira 1996).

Projeto genoma de *C. perniciosa*

Apesar da importância econômica de *C. perniciosa* pouco se sabe sobre sua genética e foi por isso que esforços foram feitos no sentido de se estabelecer um projeto genoma (**http://www.lge.ibi.unicamp.br/vassoura**). Ferramentas de bioinformática foram desenvolvidas e vêm sendo utilizadas para gerar análises quantitativas e qualitativas dos dados provenientes do projeto genomas. Até o dia 19/11/2004 já haviam sido registrados 151.077 reads seqüenciados através do método de “Shotgun” aleatório total ou “Whole genome Shotgun” (Venter, Adams et al. 1998), perfazendo cerca de 2,5 vezes o genoma inteiro de *C. Perniciosa* estimado em 30 Mb através de Pulse Field Gel Electrophoresis (PFGE) (Rincones, Meinhardt et al. 2003).

Um exemplo de ferramenta criado para anotar os reads submetidos pelos pesquisadores é o programa Gene Projects que também pode ser usado na análise de dados provenientes de EST (“Expressed Sequence Tags”). O programa Gene Projects está ligado a um sistema de submissão que incorpora reads recém seqüenciados em um banco de dados no formato FASTA através de algoritmos específicos do programa PHRED/PHRAP (Ewing and Green 1998; Ewing, Hillier et al. 1998). Este sistema além de ler os cromatogramas em formato FASTA também executa operações conhecidas como trimagens nestas seqüências de modo a excluir vetores e regiões de baixa complexidade (repetições e vetores). Este sistema de submissão é ainda responsável para executar automaticamente o programa BLAST (Basic Local Alignment Search Tool) (Altschul, Gish et al. 1990) em todas as seqüências ou reads submetidos e armazenar o resultado em um banco de dados que será posteriormente usado para consulta pelo programa Gene Projects.

De maneira resumida, o Gene Projects funciona como uma interface de anotação e montagem para que os pesquisadores procurem genes ou regiões de interesse entre os reads já submetidos. Cabe aos anotadores a tarefa de analisar a informação biológica gerada a partir dos esforços de sequenciamento de um projeto genoma de modo a identificar e organizar as seqüências nucleotídicas resultantes que compõe o código genético do organismo. Mas exatamente do que se trata esta tarefa de sistematização e organização? O sequenciamento de um genoma eucariótico através de sequenciamento de "shotgun" total aleatório gera tanto seqüências codantes como introns, portanto uma dos esforços iniciais do anotador é a identificação de seqüências codantes através da busca de ORFs ("open reading frames") de modo a identificar genes codantes de proteínas. Como um número considerável de genomas já foi seqüenciado até hoje é possível identificar estes genes através de comparação genômica usando o programa BLAST. Isto funciona de modo a comparar a seqüência de uma ORF nova e ainda desconhecida contra o banco de dados de seqüências biológicas já anotadas conhecido por NCBI de modo a inferir sua função por similaridade e descobrir possíveis ortólogos deste gene.

Devido a esta disponibilidade de seqüências de diversos projetos genoma nos bancos de dados biológicos hoje é mais rápido e mais barato achar seqüências através de similaridade com seqüências já anotadas através comparação genômica usando uma das diversas possibilidades de BLAST como será visto a seguir.

BLAST (Basic Local Alignment Search Tool) e bancos de dados biológicos

Busca de similaridade entre seqüências através de banco de dados biológicos tornou-se uma atividade indispensável em laboratórios tradicionais ao redor do mundo. A ferramenta mais utilizada de busca por similaridade entre seqüências tanto de proteína como de nucleotídeo é o BLAST (Altschul, Gish et al. 1990). Os resultados de busca do BLAST podem ser utilizados para inferir função de cDNAs recém descobertos, encontrar novos membros de uma família de genes ou explorar relações evolutivas entre seqüências (Altschul, Boguski et al. 1994). O programa BLAST funciona, de maneira sucinta, através da busca de uma seqüência de entrada que pode tanto ser de nucleotídeo ou aminoácido conhecido como "query", seguido de busca desta seqüência "query" contra um banco de dados tanto de aminoácido quanto de nucleotídeo. O programa retorna como resultado desta busca uma série de seqüências "hits", os alinhamentos das seqüências dos bancos de dados com a seqüência de busca e valores estatísticos suportando estes alinhamentos. Existem diferentes implementações do programa BLAST para realizar diferentes tarefas e para lidar com diferentes tipos de seqüência e estes podem ser visualizados na tabela 1 abaixo.

Um erro freqüente ao realizarmos buscas por similaridade de seqüência é a falha de procurarmos uma seqüência contra um banco de dados desatualizado. O NCBI produz e atualiza o Gen Bank diariamente além de dividir dados em bases diárias com o banco de DNA do Japão (DDJB) e o laboratório de biologia molecular europeu (EMBL). Uma busca do banco de dados do NCBI através da página Web do BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) garante acesso as

mais recentes atualizações. O banco de dados nr (não-redundante) é talvez o mais "compreensivo" dentre todos os bancos. Apesar de nr não ser propriamente não redundante de modo a conter múltiplas cópias de seqüências similares, as seqüências idênticas são postas em uma só entrada.

Tabela 1. Programas de busca do BLAST

Programa de busca por similaridade	Sequência de busca ou "query"	Banco de dados (sequências-alvo)	Comentarios
BLASTP	Proteína	Proteína	Pode ser usado em seu modo convencional ou sensitivo (PSI-BLAST)
BLASTN	Nucleotídeo (dupla fita)	Nucleotídeo	Possui os parâmetros otimizados para velocidade e não para sensitividade; não aconselhável para achar seqüências codantes pouco relacionadas.
BLASTX	Nucleotídeo (tradução dos seis quadros de leitura)	Proteína	Útil para análise de dados preliminares
TBLASTN	Proteína	Nicleotídeo (tradução dos seis quadros de leitura)	Essencial para procurar "queries" de proteínas contra banco de dados de EST. Útil para achar ORFs não documentadas ou erros em quadros de leitura em bancos de dados de seqüências.
TBLASTX	Nucleotídeo (tradução dos seis quadros de leitura)	Nucleotídeo (tradução dos seis quadros de leitura)	Somente deve ser usado quando os demais BLASTs não funcionarem. Restrito para buscas contra banco de dados de EST e Alu.

*Tabela adaptada de (Ausubel, Brent et al. 1999)

O NCBI nr considera que duas seqüências são idênticas quando possuem tamanhos iguais e possuem resíduos idênticos em todas as posições (Ausubel, Brent et al. 1999). O nr é composto somente por seqüências apresentando anotações de boa qualidade e não inclui aqui "expressed sequence tags" (ESTs), "sequence tagged sites" (STSs) ou "high-throughput genomic" (HTG). Existem bancos nr tanto para seqüências de nucleotídeos como para aminoácidos. GenBank, DDJB e EMBL são bancos de dados de nucleotídeos. Caso qualquer nucleotídeo em qualquer um destes bancos seja anotado com uma CDS ("coding sequence"), este aparece em nr. NCBI-nr

também contém seqüências de proteínas do banco Swiss Prot (um banco de dados de seqüências de proteína altamente anotado e curado; Bairoch and Apweiler, 1998), PIR (“protein identification resort) e PRF (“ protein research foundation”).

Uma típica lista de “hits” produzida usando-se, por exemplo, o BLASTP retorna quatro campos de informação referentes a uma busca em particular (Altschul, Boguski et al. 1994). No primeiro campo temos a designação do banco de dados com seu número de acesso juntamente com o nome do locus da seqüência alinhada. O segundo campo, muito importante para a anotação da função do gene, consiste em uma breve descrição textual da seqüência e seu conteúdo varia de acordo com diferentes bancos de dados. O mais importante, principalmente para detecção de transferência horizontal de genes é o nome do organismo cuja seqüência é proveniente e o tipo de seqüência (ex: mRNA ou DNA), além de alguma informação sobre função e fenótipo. O terceiro campo é composto pela nota do alinhamento entre a seqüência de busca e a seqüência alvo em bits. Quanto melhor for o alinhamento, ou seja, mais identidades e similaridades, maior será sua nota em bits. As maiores notas sempre ficam no topo da lista de “hits” ou alinhamento do BLAST.

Em termos gerais esta nota é calculada de uma fórmula que leva em consideração o alinhamento de resíduos similares ou idênticos, assim como “espaços” que devem ser introduzidos de modo a alinhar as seqüências. A matriz padrão usada por BLAST para o cálculo da nota de alinhamento em bits, ou a matriz de substituição é a BLOSUM-62 e é usada como padrão na maioria das buscas com BLAST.

O quarto campo contém o “valor esperado”, mais conhecido como “e-value”, já embutido no jargão da maioria dos biólogos atuais e cuja função é fornecer uma estimativa de significância estatística.

O valor de e-value reflete quantas vezes podemos esperar que um resultado (um alinhamento) ocorra aleatoriamente. Em estatística, podemos considerar um resultado de e-value de $e<0.05$ como significante. Este valor é o indicador mais confiável para calcular a importância de um alinhamento de BLAST uma vez que tanto as notas dos alinhamentos em bits como a nota absoluta do alinhamento não podem ser julgadas isoladamente da informação usada para calcular o e-value. O valor de e-value leva em consideração tanto o tamanho da seqüência de busca como o tamanho do banco de dados sendo procurado, bem como o sistema de pontuação (BLOSUM 62). Devido ao fato de notas de e-value serem sensitivas ao tamanho do banco de dados, podendo variar de um banco para o outro de acordo com seu tamanho (um “hit” significante estatisticamente em um banco relativamente pequeno pode não ser significante em uma busca contra nr). De maneira mais abrangente, quanto maior for o valor em bits do alinhamento e menor for seu e-value, maior será a significância estatística do alinhamento.

É comum hoje usarmos um “valor de corte” de modo a selecionar alinhamentos estatisticamente e biologicamente significantes, normalmente o valor de ($e<-05$) entre seqüências

de busca e seqüências alvos em bancos de dados é o valor de corte padrão. Apesar destas medidas estatísticas, sempre é recomendável uma inspeção dos alinhamentos de modo a determinar sua significância biológica o que pode ser feito através do exame do alinhamento dos pares de seqüências na logo abaixo das notas de cada alinhamento. Estes alinhamentos sempre apresentam uma linha intermediária ou consenso entre as seqüências indicando as similaridades entre elas; caso os resíduos sejam idênticos entre a seqüência de busca e a seqüência alvo, o resido é repetido na linha consenso; as substituições conservativas, julgadas pela matriz de substituição são indicadas pelo símbolo de (+); barras entre as seqüências (-) representam inserções ou deleções. Outra opção padrão do algoritmo BLAST é a opção de filtro que elimina regiões da seqüência de baixa complexidade, indicando-as com um (X), normalmente repetições de baixa qualidade, vetores ou plasmídios.

Uma vez que o produto do sequenciamento é obtido para um fragmento gênico cuja ORF não é conhecida, ou sua região codante é desconhecida torna-se necessário o uso de BLASTX. Uma rápida busca com BLASTX pode revelar rapidamente qualquer similaridade entre todas as ORFs possíveis em uma seqüência de nucleotídeo com todas as proteínas caracterizadas (Ausubel, Brent et al. 1999).

O algoritmo de BLAST mais usado para buscas envolvendo seqüências de “expressed sequence tags” (EST) é o TBLASTN. Buscas com TBLASTN envolvem seqüências de proteína contra a tradução em seis quadros de um banco de dados de nucleotídeos (Ausubel, Brent et al. 1999). Na prática podem ser usados para descobrirmos novas similaridades de uma proteína através da comparação desta com as ORFs de seqüências não caracterizadas ou de baixa qualidade, como aquelas comumente encontradas em ESTs e outros tipos de bancos de dados. As traduções destas seqüências geralmente não estão presentes no nr porquê as proteínas não estão anotadas. O uso mais comum de TBLASTN, portanto é sua busca contra ESTs que sejam similares a proteínas de interesse.

Panorama da genômica de fungos parasitas e simbióticos

Dez anos após o sequenciamento do primeiro genoma completo, da bactéria *Haemophilus influenzae* em 1995, temos hoje, segundo o banco de dados KEGG (www.genome.jp/kegg/catalog/org_list.html) 41 genomas de eucariotos sendo seqüenciados, sendo 10 completos; 202 genomas completos de bactérias seqüenciados e 21 genomas completos de Archae. Deste total de 264 genomas, 18 são de fungos, sendo 17 de fungos Ascomicetos e apenas 1 basidiomiceto (*Ustilago maydis*). Neste contexto o projeto genoma de *C. perniciosa*, um fitopatógeno basidiomiceto será de importância para o propósito de comparação genômica, ao passo que uma vez anotados, estes genes estarão à disposição para reconstrução filogenética ou busca por genes ortólogos em bancos de dados biológicos.

Com o aumento constante do número se seqüências genômicas de eucariotos sendo produzidas a cada ano, as oportunidades para comparação genômica tornaram-se sem precedentes. Existe uma demanda crescente para o desenvolvimento de novas ferramentas e métodos de modo a aproveitar todas as informações que possam advir destas comparações (Galagan 2004). Genomas de fungos são ideais para este propósito devido ao tamanho relativamente pequeno de seus genomas tornando as tarefas de múltiplas comparações genômicas inteiras com outros organismos mais acessíveis. Outro fator que faz dos fungos um bom organismo modelo para comparações inteiras de genoma é que eles possuem diversas características genômicas em comum com outros eucariotos e estas comparações podem nos informar sobre a biologia de todos os eucariotos.

Comparado com o tamanho de outros genomas eucariotos como animais e plantas, o genoma de fungos é relativamente pequeno; *S. cerevisiae* e *S. pombe* possuem respectivamente 13.7Mb e 13.8Mb (Goffeau, Barrell et al. 1996; Wood, Gwilliam et al. 2002). Basiciomicetos e Ascomicetos possuem geralmente entre 13-42Mb (Kupfer, Reece et al. 1997; Yoder and Turgeon 2001), de acordo com o tamanho predito do genoma de *C. perniciosa* (30Mb) determinado através de "Pulse Field Gel Electrophoresis" (PFGE) (Rincones, Meinhardt et al. 2003). Portanto os genomas de fungo possuem apenas um terço do tamanho do genoma de *C. elegans* e *Arabidopsis thaliana*. Outra característica dos genomas de fungos é sua alta densidade gênica e uma baixa proporção de seqüências repetitivas. *S. cerevisiae* contém aproximadamente um gene a cada 2Kb (Goffeau, Barrell et al. 1996) enquanto *N. crassa* contém um gene a cada 4Kb. A densidade genômica de *M. grisae* foi estimado em 4.2 genes por Kb (Hamer, Pan et al. 2001).

Mais de 100.000 espécies de fungos foram descritas até hoje, sendo que 10% destes obtém nutrientes através de associações íntimas (simbiose e modo de alimentação saprofítico) com outros organismos como plantas e animais, incluindo humanos (Tunlid and Talbot 2002). Até agora nosso conhecimento sobre os processos de patogenicidade como mecanismos de reconhecimento de hospedeiros, desenvolvimento de estruturas de infecção e penetração e colonização de tecidos são limitados. Entretanto espera-se que esta situação mude rapidamente nos próximos anos à medida que um grande volume de informações do sequenciamento de diversos patógenos fúngicos como *C. perniciosa* e simbiontes tornem-se disponíveis.

Além da finalização da anotação de *Schizosaccaromyces pombe* em 2002 e *Neurospora crassa* em 2003, o sequenciamento de diversas outras espécies de fungo estão quase chegando ao fim como, por exemplo, o dos patógenos fúngicos humanos *Candida albicans* e *Cryptococcus neoformans.* Outros organismos modelo importantes para o estudo da interação patógeno-hospedeiro são os patógenos de planta *Magnaporthe grisae*, o agente causal da doença brusone no arroz e o basidiomiceto *Ustilago maydis.* A disponibilidade destas seqüências para comparação somada ao advento de análises de genomas inteiros poderá possibilitar, por exemplo, a definição de todos os componentes genéticos que são expressos durante os processos ligados a patogenicidade. A concretização destes avanços no conhecimento de patógenos fúngicos, no

entanto trará uma série de desafios sendo o primeiro deles a construção de recursos de bioinformática para armazenarmos e investigarmos a gigantesca quantidade de informação gerada a partir dos projetos genoma de fungos patogênicos. Esforços no presente momento para a coleta de seqüências genomicas e "expressed sequence tag" (EST) de diversos outros patógenos fúngicos em plantas além de *C. perniciosa* estão tornando-se mais comuns (tabela 1). Estes projetos estão diretamente relacionados a anotação de *C. perniciosa* uma vez que servirão como medida de comparação tanto para seqüências genômicas como para ESTs. Recursos genômicos disponíveis para pesquisadores estudando fungos fitopatogênicos ainda são limitados. Uma das exceções é o banco de dados de seqüências de EST de fungos fitopatógênicos e Oomycetos altamente curado, o COGEME (http://cogeme.ex.ac.uk.), contendo significante anotação para cada EST depositado, que contém até o presente momento (02/2005) 57.727 seqüências de ESTs de 13 fungos fitopatógenos além de três fungos não patogênicos com o propósito de comparação (tabela 1) (Soanes, Skinner et al. 2002).

Tabela 1. Projetos de sequenciamento de patógenos fúngicos de plantas e Oomycetos.

Patógenos de plantas	Endereço URL	Atualização (Fevereiro 2005)
Blumeria graminis f.sp *hordei*	http://cogeme.ex.ac.uk/*	3253 unisequences (ESTs)
Sclerotinia sclerotiorum	http://cogeme.ex.ac.uk/*	738 unisequences (ESTs)
Botrytis cinerae	http://www.genoscope.cns.fr http://cogeme.ex.ac.uk	2901 unisequences (ESTs)
Cladosporum fulvum (bolor da folha do tomateiro)	http://www.ncbi.nlm.nih.gov/dbEST	513 unisequences (ESTs)
Fusarium graminearium	http://cogeme.ex.ac.uk	4112 unisequences (ESTs)
Fusarium sporotrichioides	http://www.genome.ou.edu/fsporo.html	3448 unisequences (ESTs)
Ustilago maydis	http://cogeme.ex.ac.uk	4276 unisequences (ESTs)
Magnaporthe grisae	http://cogeme.ex.ac.uk http://www.ncbi.nlm.nih.gov/dbEST http://www.fungalgenomics.ncsu.edu***	12465 unisequences (ESTs)
Cryphonectria parasitica	http://cogeme.ex.ac.uk	2185 unisequences (ESTs)
Mycosphaerella graminicola	http://cogeme.ex.ac.uk	2926 unisequences (ESTs)
Phytophthora infestans	http://www.ncbi.nlm.nih.gov/dbEST*	1414 unisequences (ESTs)
Colletotrichum trifolii	http://cogeme.ex.ac.uk	550 unisequences (ESTs)
Leptosphaeria maculans	http://cogeme.ex.ac.uk	118 unisequences (ESTs)
Verticillium dahliae	http://cogeme.ex.ac.uk	1455 unisequences (ESTs)
Phytophthora sojae	http://www.ncbi.nlm.nih.gov/dbEST*	5849 unisequences (ESTs)

Um recurso oferecido pelo COGEME é a confirmação dos produtos de genes putativos para cada EST através de buscas por similaridade contra o banco de dados NCBI nr ("non-reduntant") (ftp://ftp.ncbi.nih.gov/blast/db/) usando o algoritmo BLASTX. Os primeiros 20 "hits", obedecendo a um valor de corte de 1 x 10-5 foram inseridos no banco de dados. Este recurso possibilita tanto estudos de inferência de função como relações evolutivas entre os genes como foi neste trabalho (ver resultados).

Para entendermos melhor a importância do desenvolvimento destes recursos genômicos em fungos atualmente basta observarmos nos últimos oito anos um fluxo semelhante de informação genômica em procariotos que culminou com grandes impactos na pesquisa de patogênese em bactérias e simbiose (Wren 2000; Ochman and Moran 2001). A análise genômica comparativa de espécies de bactéria proporcionou novas descobertas na evolução da virulência e adaptação de hospedeiros (Tunlid and Talbot 2002).

Uma observação surpreendente surgida de comparações de seqüências genômicas disponíveis de fungos e outros organismos é que uma significante proporção de seqüências não exibem similaridade a seqüências de proteína ou DNA presentes em bancos de dados. Um exemplo disto é que de 40 a 60% das seqüências de EST de patógenos fúngicos exibem nenhuma ou pouca similaridade a proteínas de função conhecida (Skinner, Keon et al. 2001). Estes genes são conhecidos por órfãos (ORFs que não exibem funções), ou como são conhecidos na linguagem do BLAST "no hits". Órfãos são comumente achados nos genomas de modelos organismos eucarióticos. Um exemplo disto é que foi estimado que 1/3 de todas as proteínas codantes preditas em *S. cerevisiae* são órfãs (Oliver 1996). As funções destes órfãos precisam ser comprovadas através de métodos genéticos ou bioquímicos e permanece um dos grandes desafios da genômica funcional (Tunlid and Talbot 2002).

Um estudo comparativo de larga escala envolvendo os genomas de *N. crassa* e *S. cerevisiae* apresenta evidência que existe uma maior proporção de genes órfãos em *N.crassa* do que *S. cerevisiae*. Alguns destes órfãos que não estão presentes em *S. cerevisiae* podem refletir a aquisição ou manutenção de novos genes e seria consistente com o genoma maior e complexidade morfológica de *N. crassa* (Tunlid and Talbot 2002). Filogenias moleculares de fungos parasíticos e simbiônticos nos mostram que estes organismos podem ser encontrados em vários grupos taxonômicos, sugerindo que estes estilos de vida evoluíram repetidamente no reino dos fungos.

Ainda, de acordo com Tunlid e Talbot (2002); existem três mecanismos ao "nível genômico", que podem explicar a múltipla emergência e adaptação de fungos. O primeiro é que parasitismo e simbiose estariam diretamente envolvidos com a aquisição de genes novos. Estes genes poderiam ter um papel durante a infecção do hospedeiro e poderiam ser adquiridos através de duplicação gênica e transferência horizontal de genes (HGT), descrito na próxima sessão deste

trabalho. A segunda opção é a de que adaptações aos hábitos parasíticos e simbióticos poderiam resultar de diferenças na regulação da expressão gênica. Sendo a última; parasitismo e simbiose são associados a perda gênica e deleção. Vários estudos demonstram que fungos parasíticos possuem fatores patogênicos únicos. Dentre os exemplos mais famosos estão as conhecidas toxinas hospedeiro-específicas produzida por várias espécies de patógenos de plantas (Tunlid and Talbot 2002).

Análise de seqüências genômicas de bactérias demonstraram que muitos genes necessários para virulência são restritos a organismos patogênicos e estes foram introduzidos no genoma através de HGT (Ochman and Moran 2001; Gamieldien, Ptitsyn et al. 2002). Existem evidencias de HGT em fungos (Rosewich and Kistler 2000). Começamos aqui a introduzir o tema da transferência horizontal de genes de maneira cronológica; de fato, as primeiras evidências robustas de HGT começaram a surgir a medida que os dados dos primeiros projetos genoma passaram a estar disponíveis para comparações genômicas. Sabemos hoje que os casos mais robustos de HGT geralmente estão ligados a algum contexto evolutivo-ambiental, como por exemplo, a relação hospedeiro-patógeno, porém, antes de abordarmos transferência horizontal propriamente alguns parágrafos abaixo, daremos ainda alguns exemplos provenientes dos primeiros estudos resultantes da comparação de genomas bacterianos. Em bactéria HGT geralmente envolve a transferência de um cassete inteiro de genes de tamanho de 5 a 100Kb. Caso venha a ser comprovado que estes cassetes contribuam para virulência de fato, eles já foram chamados de “ilhas de patogenicicade” em um alusão ao fato destas regiões serem intrinsecamente diferentes das outras regiões do genoma como é visto a seguir (Ochman and Moran 2001). Estas Ilhas de patogenicidade são conhecidas por sua instabilidade genética, distribuição filogenética restrita, proximidade a elementos genéticos móveis (transposons, retrotransposons), conteúdo G+C atípico relativo a media do genoma. Vários genes de patogenicidade são conhecidamente “clusterizados” em fungos patogênicos em plantas, incluindo genes codificadores de reguladores do crescimento de plantas, toxinas e metabolitos secundários (Rosewich and Kistler 2000). Alguns destes clusters como os genes de patogenicidade da ervilha PEP (“Pea pathogenicity genes”) no patógeno fúngico *Nectria haematococca* (Han, Liu et al. 2001) e o locus TOX1 em *Cochliobolus heterostrophus* (Yang, Rose et al. 1996), outro fungo patogênico, possuem características semelhantes as ilhas de patogenicidade em procariotos, incluindo diferenças em “códon usage” e conteúdo G+C da média do genoma. A Linhagen altamente patogênica Race-T de *C. heterostrophus* difere da linhagem Race-O no locus Tox1, responsável pela produção da T-toxina por um gene PKS1. O patógeno da linhagem Race-O não possui homólogos de PKS1 sugerindo que Race-T possa ter adquirido PKS1 através de transferência horizontal de genes (Yang, Rose et al. 1996).

Os mecanismos potenciais de como a transferência destas ilhas de patogenicidade possa ocorrer ainda é desconhecido. O cluster PEP é localizado em um cromossomo supernumerario e estes cromossomos demonstram a capacidade de transferência entre indivíduos geneticamente

isolados de um fungo patógeno de plantas (He, Rusu et al. 1988). Outros casos envolvendo HGT serão descritos mais adiante (ver logo abaixo e mecanismos de HGT).

Transferência horizontal de genes (HGT).

Transferência horizontal de genes pode ser definido como a aquisição de material genético (genes, introns, transposons) entre diferentes grupos taxonômicos reprodutivamente isolados, na qual o material genético é integrado ao genoma recipiente de modo que seja herdável. Este processo é conhecido como Transferência Horizontal de Genes (HGT em inglês: Horizontal Gene Tranfer) e contrasta com o modo tradicional de aquisição de material genético (transferência vertical), na qual o material genético é transmitido verticalmente ao longo de sucessivas gerações em um grupo taxonômico.

Um dos pilares da teoria da evolução de Charles Darwin é a herança vertical de características de pais para filhos ao longo de sucessivas gerações. Entretanto, recentemente biólogos moleculares evolucionistas mostraram que extensivo intercâmbio de genes pode ocorrer entre espécies distantes ou diversos grupos taxonômicos, mudando o padrão de evolução linear desses organismos (Brown 2003). De acordo com Brown (Brown 2003), HGT pode ter sido o principal agente na evolução de vida celular e na emergência dos três reinos da vida: Archae, Bactéria e Eukarya. O motivo para tanto é simples: genes são achados em lugares que eles não deviam estar, ou seja, alguns genes em espécies de eucariotos são mais similares a homólogos em procariotos do que em outras espécies de eucariotos (Doolittle, Feng et al. 1990), assim como alguns genes de bactérias são mais semelhantes com versões similares em Archae.

Do ponto de vista da evolução molecular hoje em dia, baseada em árvores filogenéticas de seqüências moleculares, toda a vida é dividida em três domínios: Bactéria, Eucarya e Archae. O grupo Bactéria é composto por todos os procariotos (organismos sem um núcleo definido) exceto o grupo Archae. Incluído no grupo das bactérias encontramos as bactérias gram-positivas e gram-negativas, “mycoplasmas”, “cyanobactérias” e alguns outros. Archae (ou Archaebactéria) também são procariotos, que, no entanto diferem de bactéria por apresentarem seqüências de RNA ribossômico (rRNA), lipídeos de membrana, introns além de outras características. Os eucariotos consistem de organismos que apresentam núcleo e diversas organelas únicas a este domínio. Os eucariotos incluem fungos, plantas e animais e o que vêm sendo proposto é que as barreiras entre os grupos taxonômicos vêm sendo crescentemente erodida através do reconhecimento de extensa transferência horizontal de genes (Bushman 2002). Portanto a monofilia de cada grupo taxonômico, em alguns casos, dependendo da magnitude de transferência horizontal pode estar em cheque.

HGT pode, portanto, significar uma mudança de paradigma em relação ao nosso conhecimento de como a vida evoluiu até agora. De fato, se considerarmos que as características mais salientes da árvore universal da vida baseada em seqüências de SSU rRNA são a singularidade ou monofilia de cada um dos três domínios da vida; similaridade maior entre archae e eucariotos em relação as bactérias, e a evolução precoce de bactérias termófilas no domínio

bactéria. HGT violaria a hipótese da árvore universal da vida ao misturar seqüências de diferentes taxas em um clado polifilético (Brown 2003). Um gene é dito ser transferido horizontalmente caso a topologia da árvore filogenética, seja ela construída com seqüências moleculares, ou através do padrão de distribuição dos ortólogos daquele gene entre diversas espécies, seja incongruente com a árvore universal da vida.

Existem dois tipos básicos de HGT; aquele no qual a transferência de um dado gene ocorreu há muito tempo em uma escala evolutiva e que, portanto sua composição nucleotídica adaptou-se àquela da espécie recipiente. Este tipo de evento só pode ser detectado com inferência filogenética. O segundo tipo de HGT envolve casos de aquisição recente de genes. Nestes casos a transferência pode ser detectada através da diferença na composição nucleotídica do gene recém transferido e a composição nucleotídica do genoma hospedeiro. Existem várias técnicas disponíveis para este tipo de análise paramétrica (análise de “codon usage” e conteúdo G+C são as mais usadas).

Entretanto, casos de transferência horizontal em eucariotos permanecem um desafio para geneticistas. Desde que HGT foi especulado pela primeira vez por Syvaneh (Syvanen 1994) que a idéia vêm sofrendo resistências dentro das escolas mais tradicionais da biologia, como é o caso da escola neo-Darwinista. Esta resistência pode ser considerada natural uma vez que, superficialmente, o conceito de HGT causaria contradições com o poder explicativo de árvores filogenéticas em taxonomia e o papel importante do isolamento reprodutivo como mecanismo de especiação (Syvanen 1994). Ao mesmo tempo um progresso na área de genética molecular nas últimas décadas aliada a novas ferramentas de bioinformática possibilitaram várias observações diferentes a respeito de HGT. Hoje, não somente existem dezenas de exemplos de prováveis casos de HGT, mas também os mecanismos que poderiam clarificar esta transferência e a integração de um novo DNA no genoma, como as diferentes formas de transformação. Alguns exemplos podem ser encontrados listados na tabela 2.

A medida em que os dados de diversos projetos genoma se tornam disponíveis para comparação genômica, diversos casos de transferência horizontal vem sendo amplamente demonstrados entre espécies pouco relacionadas taxonomicamente. Estudos recentes em um grande número de espécies (17 bactérias e 7 archae) mostram que uma porção significativa do genoma em procariotos (1.5%-14.5%) pertencem a aquisições horizontais recentes (Garcia-Vallvé 2000). Os métodos usados para esta estimativa foram baseados no cálculo do conteúdo G+C e “códon usage” dos genes presentes em um dado genoma e comparados ao seu valor médio no genoma total. Genes que apresentem um desvio padrão significativo desta média genômica são considerados como sendo horizontalmente adquiridos.

Tabela 2. Algumas transferências horizontais envolvendo eucariotos (Syvanen, 1994).

	Proteína	Caso	Referência
1.	NodL	similaridade a proteínas da bactéria *Rhizobium l* . de proteínas e genes do nematódeo *Meloidogyne s*. Confirmados filogenéticamente.	(Scholl, Thorne et al. 2003)
2.	Cutinase	Ausência de ortólogos deste gene de *Mycobacterium tuberculosis (Gamieldien, Ptitsyn et al. 2002)* em bactéria.	
3.	Adenylate cyclase	Somente presente em *tubercle bacilli* e eucariotos.	(Gamieldien, Ptitsyn et al. 2002)
4.	Fe superoxide dismutase	Fe SOD do protista *Entamoeba histolítica* possui origens procarióticas.	(Smith, Feng et al. 1992)
5.	Aldolase	Aldolase classe 11 de levedura mostra afinidade com aldolase de *E.coli*.	(Smith, Feng et al. 1992)
6.	Citocromo C	a proteína de *Arabidopsis thaliana* possui afinidade com fungos.	(Kemmerer, Lei et al. 1991)
7.	Xylanase	gene encontrado em "rumen fungi" é similar a *Rumoncoccus*.	(Gilbert, Hazlewood et al. 1992)
8.	Thioredoxin	Thiorredoxina-m em plantas possui origem bacterial	(Hartman, Syvanen et al. 1990)
9.	Glyceraldehyde-3-phosphate dehydrogenase	*gapdhA* de *E.coli* e *Anabaena* possuem uma afinidade com eucarioto *gapC*	(Doolittle, Feng et al. 1990),
10.	Elongation facto Tu	*rufA* em *Arabidopsis* é do seu endosimbionte	(Baldauf and Palmer 1990)
11.	Proteína ribossômica L21 e L22	L21 e L22 em algumas plantas são de seus endosimbiontes	(Gantt, Baldauf et al. 1991)

*Tabela modificada a partir de (Syvanen, 1994).

A magnitude da aparente transferência horizontal de genes em procariotos talvez não devesse ser recebida com tanta surpresa, como apresentado por Koonin e equipe (Koonin, Makarova et al. 2001), nem com ceticismo exagerado. De fato, a habilidade dos microorganismos absorverem DNA do ambiente que os cercam e integrá-lo ao seu genoma foi demonstrado já pelo

clássico experimento de Avery-McLeod-McCarthy em 1943, que provou o papel do DNA como sendo o principio transformante.

Portanto os primeiros estudos comparativos entre genomas completamente seqüenciados sugerem a importância da contribuição de genes horizontalmente transferidos. Talvez mais até do que previamente pensado (Wolf, Aravind et al. 1999). Já podemos afirmar sem dúvida alguma que há um fluxo de genes de bactérias para eucariotos, assim como é predominante entre eventos de HGT, sendo talvez a teoria endosimbiontica da mitocôndria e plastídio a mais evidente (Koonin, Mushegian et al. 1997).

Estudos recentes indicam que relações evolutivas entre organismos, como a relação parasita-hospedeiro, geram condições propícias para um fluxo de DNA entre estes. HGT poderia favorecer a adaptação do patógeno ao ambiente intracelular (Wolf, Aravind et al. 1999). Análises mostram que mais de 30 genes do genoma da bactéria *Chlamydia trachomatis* agrupam com homólogos em eucariotos e não foram achados em genomas bacterianos. Outros agruparam claramente com homólogos eucariotos em arvores filogenéticas (Wolf, Aravind et al. 1999).
Outros dois estudos recentes (Scholl, Thorne et al. 2003; Nitz, Gomes et al. 2004) um utilizando nematóides parasitas de plantas da espécie *Meloydogyne* e outro com o protozoário *Trypanossoma cruzi,* demonstram como HGT pode ser um mecanismo importante na evolução de parasitismo e na interação patógeno-hospedeiro. Em *T. cruzi* ficou provado experimentalmente a existência de integração do DNA do cinetoplasto (kDNA) em células do hospedeiro durante a infecção. Esse processo foi realizado artificialmente em células germinativas de coelhos e aves, tendo sido obtida a transmissão para as gerações F1, F2 e subseqüentes de cruzamentos, indicando que quando processos dessa natureza acontecem em células germinativas eles podem dar origem a HGT.

Genes sujeitos a HGT entre nematódeos e bactéria foram recentemente identificados [36] através de um método de comparação de bancos de dados de EST de três espécies de nematóides *Meloidogyne* contra bancos de dados formados a partir de seqüências dos invertebrados *C. elegans e Drosophila* e outro de proteínas bacterianas. Como similaridade a uma seqüência de bactéria é o critério mais simples para se considerar uma proteína deste nematódeo e conseqüentemente o gene que a codifica, como sendo um possível candidato a HGT, o banco de proteínas de bactéria foi gerado com a finalidade de identificar este tipo de similaridade. Os outros dois bancos de dados de invertebrados foram gerados de modo a eliminar candidatos em *Meloidogyne* que apresentassem similaridade com estes invertebrados, pois estes não seriam adquiridos através de HGT. Um gene similar a bactéria presente em *Meloidogyne e Drosophila*, mas ausente em *C. elegans* provavelmente não foi adquirido através de HGT, tendo sido resultado de perda gênica na linhagem de *C. elegans*, após a divergência destas espécies de seu último ancestral em comum, o que não seria identificado caso o banco de *Drosophila* não tivesse sido gerado.

O método identificou seis novos candidatos a HGT, cuja análise filogenética indicou a origem rhizobial (bactéria fixadora de nitrogênio no solo e que povoam raízes de plantas) de quatro deles. Curiosamente, *Meloidogyne* e *Rhizobia* são simpátricos, isto é, dividem um nicho ecológico no solo, e provavelmente na planta também, satisfazendo o requisito de proximidade física para que HGT ocorra.

Existem diversos mecanismos afetando evolução de parasitas. Entre estes podemos incluir a adaptação de genes pré-existentes para codificar novas funções, duplicação gênica, perda gênica e aquisição de genes de outras espécies (HGT). Estudos recentes mostram a relevância do papel de HGT na relação patógeno-hospedeiro (Wolf, Aravind et al. 1999; Scholl, Thorne et al. 2003; Nitz, Gomes et al. 2004) indicando inclusive a direção da transferência sendo proposta nos casos específicos. Parece provável que no caso de parasitas intracelulares, a transferência do hospedeiro para o patógeno poderia estar sendo facilitado se comparado com parasitas extracelulares ou bactérias de vida livre ("free-living").

Recentemente Gamieldien e colegas reportaram a presença de 19 genes de origem eucariótica no genoma de *Mycobacterium tuberculosis* e a sua retenção no mesmo através de vantagem seletiva (Gamieldien, Ptitsyn et al. 2002). Sem dúvida, muitos patógenos bacterianos evoluíram sua capacidade de produzir fatores de virulência adaptados ao ambiente do hospedeiro (Gamieldien, Ptitsyn et al. 2002). Os autores chegaram a estes resultados baseados na comparação de valores de e-value de BLASTP para cada proteína predita em *M. tuberculosis* contra conjuntos de seqüências de proteínas de eucariotos e procariotos do GenBank como um "screening" preliminar de candidatos a HGT. Proteínas que apresentarem similaridade contra eucariotos ao invés de bactéria, utilizando um valor de corte de e<-10, foram selecionados como possíveis candidatos. Na outra etapa eles fizeram uma nova comparação genômica, desta vez contra o conjunto de seqüências completo do banco de dados nr utilizando para tal PSI-BLAST (**http://www.ncbi.nlm.nih.gov/blast**). Nesta etapa foram identificados 25 candidatos. Destas, 11 obtiveram similaridade somente com eucariotos e foram selecionados como candidatos putativos a HGT. Para confirmar transferência horizontal nos casos em que seqüências de bactéria foram identificados por PSI-BLAST, as proteínas representativas dos três reinos (bactéria, eucaria e archae) foram alinhadas por ClustalW (Thompson, Higgins et al. 1994) e sujeitas a análise filogenética através do software PHYLIP *(Felsenstein 1989)*. Dos 14 candidatos remanescentes, 8 apresentaram filogenias não-congruentes; agruparam com eucariotos à exclusão de procariotos com suporte de bootstrap ≥70%.

Evidência recente de genes mitocondriais codantes em plantas sujeitos a HGT entre espécies não relacionadas de plantas (Bergthorsson, Adams et al. 2003) através de métodos filogenéticos sugere a existência de um mecanismo para troca de material genético entre plantas de diferentes espécies. Árvores filogenéticas inferidas apenas com genes mitocondriais de espécies diferentes de plantas foram analisadas levando a conclusão de que a topologia da árvore só poderia ser suportada se levada em conta HGT. Como houve agrupamentos de genes em

clados de espécies diferentes criando-se uma filogenia incongruente (ex: monocotiledôneas com dicotiledôneas), as hipóteses alternativas a HGT foram refutadas, como transmissão vertical de gene do núcleo da mesma planta para a mitocôndria.

Critérios para identificação de HGT

Existem diversos métodos para detecção ou confirmação de HGT: (1) Métodos filogenéticos, (2) padrões anômalos de conteúdo G+C e "códon usage" em genomas, (3) padrões filéticos (Rosewich and Kistler 2000; Katz 2002) e (4) Homologia. O primeiro destes métodos é aquele que fornece a mais direta evidência de HGT, pois linhagens doadoras e recipientes podem ser analisadas na filogenia observando-se a topologia discordante (Philippe and Douady 2003).

O ponto de partida para obtenção de uma lista de genes candidatos a HGT em um genoma é naturalmente a comparação genômica. Dado o grande número de seqüências para analisar e a expectativa de que a maioria não são transferidos horizontalmente pode ser desenvolvido um método de comparação genômica através de BLAST para eliminação de genes herdados de modo tradicional (Wolf, Aravind et al. 1999; Gamieldien, Ptitsyn et al. 2002; Scholl, Thorne et al. 2003), através de transferência vertical. Este método seria uma espécie de filtro filogenético de modo a separar as seqüências que apresentassem alta similaridade com genes do mesmo taxa do genoma estudado.

O método mais rápido e simples para determinarmos homologia significante entre dois genes evolutivamente distantes, (diferentes grupos taxonômicos) é através de similaridade por BLAST. Entretanto valores de similaridade do BLAST nem sempre indicam relações evolutivas com precisão além do fato do tamanho e o tipo do banco de dados poder afetar o resultado obtido por este método.

Um exemplo de predição inflacionada de genes candidatos a HGT através de similaridade por BLAST foi descrito recentemente (Stanhope, Lupas et al. 2001). O consórcio internacional para o seqüenciamento do genoma humano relatou 113 casos de HGT em 2001 entre bactéria e vertebrados sem nenhuma aparente ocorrência em intermediários evolutivos (não-vertebrados). Estudos posteriores (Stanhope, Lupas et al. 2001) usando buscas mais rigorosas em banco de dados e análises evolutivas-filogenéticas em 28 destes genes propostos pelo consórcio, que haviam sido comprovados por PCR, indicam que estes genes podem ser explicados em termos de descendência desde o ancestral em comum e portanto não são exemplos de HGT. Uma das conclusões desta experiência é que relações evolutivas entre proteínas não podem ser concluídas somente através de similaridade por BLAST resultantes de buscas com bancos de dados. No caso do consórcio humano, muitos resultados de BLAST eram sem dúvida de espécies de bactéria, no entanto outras seqüências apresentando similaridade a humanos e que não foram cuidadosamente analisadas consistiam de organismos eucariotos não vertebrados, o que rejeitaria a hipótese de HGT e uma análise filogenética acurada teria mostrado esta relação. Portanto a análise filogenética deve ser um componente central de qualquer esforço de busca de genes candidatos a HGT, anotação de famílias de proteína ou anotação genômica. Segundo (Stanhope, Lupas et al.

2001): "*Reconstrução filogenética é fundamental para a síntese, a partir da crescente quantidade de dados de sequenciamento, da nossa visão da evolução de genomas*".

Métodos Filogenéticos e Paramétricos

Análise da topológica de árvores filogenéticas inferidas para famílias de genes sempre foi tradicionalmente um dos modos principais de se decifrar cenários evolutivos, incluindo HGT (Syvanen 1994). Este tipo de análise depende de congruência de arvores filogenéticas com a árvore canônica universal baseada em SSUrRNA. Um exemplo de HGT em uma árvore filogenética seria o agrupamento de uma proteína bacteriana com o seu homólogo eucarionte, excluindo o agrupamento desta mesma proteína bacteriana com homólogos de outras bactérias. Esta mistura de taxas contradiz a monofilia da árvore canônica universal, agrupando seqüências de eucariotos com seqüências de procariotos em um clado polifilético.

Conflitos entre árvores sob comparação não são necessariamente devido a HGT, mas podem surgir devido a questões específicas como perda de genes e convergência (Xie, Bonner et al. 2003). De fato, sendo convergência gênica a evolução independente de caracteres similares (genes por exemplos), em linhagens evolutivas distintas, uma árvore construída com seqüências gênicas que sofreram convergência formaria uma incongruência similar a HGT, gerando grupos polifiléticos de reinos diferentes (Bacteria e eucariotos por exemplo) (Brown 2003).

O fato de seqüências de proteínas se modificarem ao longo da evolução mais lentamente do que suas seqüências codantes de DNA tornam as arvores filogenéticas de proteínas a melhor opção para detecção de HGT ocorrido há muito tempo na escala evolutiva ("ancient HGT") . Árvores filogenéticas baseadas em nucleotídeos são, portanto, melhores para o estudo de espécies semelhantes, que possuem poucas diferenças em suas seqüências nucleotídicas (Brown 2003).

Métodos paramétricos incluem análise da composição de nucleotídeos (G+C) e "codon usage". Ambos os métodos baseiam-se na chamada 'hipótese Genômica' que diz que "codon usage" e composição G + C são assinaturas distintas de cada genoma (Grantham, Gautier et al. 1980). Segundo (Campbell 2000), a maioria dos eucariotos e procariotos possui DNA de composição homogênea. Portanto, aqueles genes cujo nucleotídeo ou composição de codons são significativamente diferentes da média esperada de um dado genoma, provavelmente foram adquiridos através de HGT.

Vários genes transferidos horizontalmente revelados por este critério são transposons, profagos e outros elementos genéticos cuja 'mobilidade evolutiva' não representa surpresa devido a suas características de mobilidade (retrotransposição, transposição) de um ponto para outro em genomas.

O consenso que se tem chegado sobre qual o melhor método de análise para se reconstruir eventos evolutivos como HGT envolve tanto a combinação de métodos filogenéticos como métodos paramétricos (Brown 2003).

Padrão Filético

Podemos definir padrão filético como simplesmente o "padrão de espécies", ou seja, o padrão de presença ou ausência em um dado cluster de genes ortólogos de espécies. Quando um grupo de genes ortólogos mostra, por exemplo, a presença de uma família de proteínas tipicamente encontrada em eucariontes em uma única linhagem de bactéria, a probabilidade de HGT é alta.

Com a disponibilidade de vários genomas completos seqüenciados, métodos novos, relativamente simples e altamente eficientes tornaram-se possíveis. Com a sistemática delineação de famílias de genes ortólogos a noção de padrão filético (filogenético) foi introduzida inclusive com a criação de bancos de dados de genes ortólogos (COGs e KOGs; "cluster of orthologous genes", em eucariotos –KOG e em procariotos -COG) (Tatusov, Fedorova et al. 2003; Koonin, Fedorova et al. 2004) através da clusterização de seqüências por similaridade de seqüência. Isto irá ajudar muito a tarefa de inferência filogenética, já que para recuperar a árvore filogenética de um determinado organismo é necessário a utilização de genes ortólogos ou todos os parálogos, o que é muito difícil (relativamente poucos genomas seqüenciados, especialmente eucariotos).

Para caracterização de um cenário de HGT ou perda gênica ("gene loss") nós recorreremos aos padrões filéticos da proteína em questão. Caso um gene esteja presente em um genoma mas ausente em vários outros genomas do mesmo taxa ou semelhantes evolutivamente (distribuição filogenética anômala), para este cenário não ser uma transferência horizontal e sim transferência vertical, teria de ter ocorrido múltiplas perdas gênicas ("gene loss") independentes o que evolutivamente é um tanto improvável.

Estudos recentes (Koonin, Fedorova et al. 2004) demonstram que padrões filéticos inesperados refletem em sua maioria eventos de perda gênica e HGT. O banco de dados atualizado de KOGs continha até 01/2005 4.852 clusteres de genes ortólogos, que são formados a partir de 60.759 proteínas ou ~54% dos ~111.000 produtos de genes eucarióticos.

Evolução molecular e filogenética

Estudos filogenéticos do sub-reino dos metazoários vêm sendo feitos há mais de um século com dados tanto da embriologia como morfológicos, contudo não foi possível inferir uma filogenia bem suportada dos metazoários através da utilização destes métodos. Existe um crescente interesse em resolver a filogenia deste período devido ao fato de que a maioria dos filos conhecidos hoje terem surgido na metade do período cambriano, 530 milhões de anos atrás (Bowring, Grotzinger et al. 1993). É predominantemente aceita a idéia de que todos os filos vivos hoje em dia possam ter tido a sua origem até o fim da explosão cambriana (Valentine, Jablonski et al. 1999). O período cambriano foi caracterizado após a evidência geológica que mostra uma súbita aparição de inúmeros fósseis com características morfológicas deste período.

Para entender os padrões de radiação das espécies que surgiram durante o período cambriano, as relações entre os filos devem ser entendidas para tanto surgiu a evolução molecular. Métodos filogenéticos também são necessários para entendermos o padrão da

evolução dos genes, assim como os processos que levaram a diferenciação no desenvolvimento morfológico que caracterizam os diversos filos. Observa-se a exaustão da maioria das características morfológicas informativas para fins de estudos filogenéticos, o que tem levado a controvérsias em relação a contribuição efetiva dessas características (Raff, MArshall et al. 1994).

O aumento da importância do uso de computadores em 1950 levou eventualmente ao desenvolvimento de métodos de "clusterização", e com isso os taxonomistas gradualmente tornaram-se mais receptivos a procedimentos numéricos e algorítmicos para inferir filogenias, culminando com o uso de filogenias moleculares para construção de árvores filogenéticas de 1960 até hoje. Zuckerkandl e Pauling foram os primeiros a perceber que as seqüências primárias de ácidos nucléicos e proteínas continham uma rica fonte de informações sobre a história evolutiva, que poderia ser recuperada através de alinhamentos de seqüências e comparação (Zuckerkandl and Pauling 1965). Essas análises levaram ao início da era da filogenia molecular a partir de 1965.

Filogenias Moleculares

O poder revolucionário destes métodos tornou-se óbvio treze anos depois quando Woese e Fox publicaram análises de um número significante de organismos baseados nas seqüências conservadas e de baixas taxas evolutivas do gene 16S rRNA (RNA ribossômico, moléculas conservadas que compõe o ribossomo de procariotos e eucariotos, também descritas como SSUrRNA), culminando com a definição de uma visão da vida na terra representada em uma filogenia composta por três domínios: Bactéria, Eucaria e Archae (Woese and Fox 1977), sendo que cada domínio possui ramos monofiléticos, ou seja, derivados de um único ancestral em comum. Alternativamente, um taxa polifilético possuiria descendentes de diferentes ancestrais. Esta representação das relação evolutivas dos principais domínios taxonômicos e três grandes taxa monofiléticos ficou conhecida como "árvore canônica universal" ou simplesmente "árvore da vida". Adicionalmente, filogenias de proteínas independentes mostram que tanto Archae quanto Eucaria são grupos irmãos, portanto a árvore deve possuir sua raiz em bactéria.

Com este estudo os rRNAs tornaram-se os marcadores referenciais para estudos filogenéticos. Usando comparações baseadas em seqüências de rRNA, as relações filogenéticas dos principais taxas de procariotos e eucariotos foram obtidos (Woese 1987; Sogin 1991). Este sucesso das SSU rRNAs como marcadores filogenéticos em parte é explicado pelo fato destas serem moléculas de evolução lenta (baixas taxas de substituição), conterem um número significativo de pares de base e não ter sido sujeita a transferência horizontal de genes. De acordo com Raff, o fato desta ser uma molécula de evolução lenta a faz possuir regiões conservadas que sofreram poucas modificações desde os seus ancestrais até os organismos de hoje em dia (Raff, MArshall et al. 1994). Por esta característica esta molécula é muito usada para estudos de homologia entre taxas distantes.

É importante ressaltar agora a diferença entre árvores de genes e árvores de organismos. A primeira se propõe a estudar a evolução de determinado gene ou família de genes ou proteínas

enquanto que a segunda se propõe a entender as relações evolutivas de organismos. A expectativa ingênua da sistemática molecular (árvores filogenéticas baseadas em seqüências de ácidos nucléicos) é a de que a filogenia de genes são compatíveis com filogenias de organismos, ou seja, através da obtenção da primeira invariavelmente descobriríamos a segunda. Na verdade existem inúmeras razões para que isto não seja verdade: a principal delas é o fato que duplicações gênicas podem resultar em genes diferentes porém relacionados em uma espécie. Estes genes são conhecidos por parálogos e caso não tenhamos todos os genes parálogos ao inferirmos uma filogenia de organismos não seria possível inferir a filogenia correta, como será exemplificado logo abaixo. Portanto é importante distinguirmos dois tipos básicos de homologia entre genes: Paralogia e Ortologia

Genes parálogos descendem de um ancestral que sofreu uma ou mais duplicações gênicas. Portanto genes paralogos se diferenciaram ao longo da evolução (através de duplicação) em linhagens diferentes de genes, muitas vezes formando famílias gênicas. No caso de dois genes homólogos que não sofreram duplicação gênica, nos referimos a eles como ortólogos

O paradoxo dos genes paralogos é bem exemplificado no caso do gene da globina que em determinado momento da evolução sofreu duplicação gênica, o que gerou duas cópias distintas do mesmo gene em suas espécies; α-globina e β-globina. Os genes de β-globina são ortólogos entre si em diferentes espécies mas são parálogos em relação ao gene da α-globina. Esse fenômeno pode prejudicar a inferência da filogenia organismal caso na análise não sejam considerados todos os parálogos de uma mesma espécie (Page and Holmes 1998). Portanto este exemplo ilustra bem a fato de que para inferirmos a filogenia organismal correta é necessário a utilização de um grupo completo de genes parálogos. Alternativamente, é possível a utilização estrita de genes ortólogos; que não sofreram duplicação gênica.

Um terceiro tipo de relação evolutiva seria a Xenologia; quando um gene deve a sua presença em um organismo à transferência horizontal.

Alternativamente a relações homólogas entre genes, uma similaridade entre duas seqüências pode ter ocorrido independentemente, caso pela qual são determinadas de homoplásicas.

Homoplasia é o termo dado a dois ou mais caracteres similares que evoluíram independentemente. É oposta a idéia de homologia na qual dois ou mais caracteres são similares devido a herança destes diretamente de seu ancestral. Convergência gênica seria um exemplo de homoplasia na qual a seqüência de dois genes são similares devido a evolução independente desta similaridade em duas seqüências não relacionadas por ancestralidade (Page and Holmes 1998). Foi sugerido que HGT também seja um fator causador de homoplasia entre seqüências (Syvanen 1994).

No caso dos basidiomicetos, filo a que pertence o fungo *C. perniciosa*, ele aparece ramificando com outro fungo filamentoso (Ascomycota) em um clade irmão em uma filogenia

construídas com SSU rRNA (“small subunit”) como pode ser observado na figura 2 (Van de Peer and De Wachter 1997).

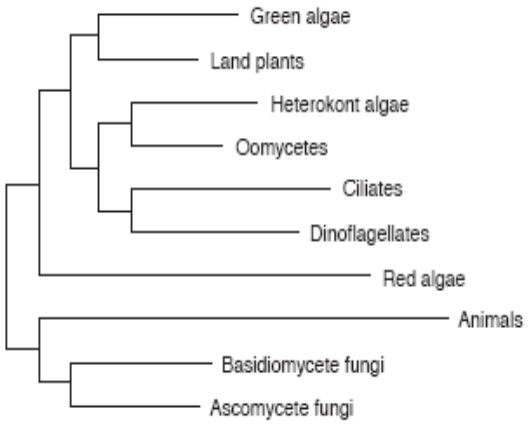


Figura 2. Àrvore filogenética mostrando as relações evolutivas entre os maiores grupos eucariotos. Oomicetos não agrupam com outros patógenos de plantas (fungos filamentosos), compostos por ascomicetos e basidiomicetos, este último do qual faz parte *Crinipellis perniciosa*. Árvore gerada a partir de seqüências da pequena unidade de rRNA ou SSU rRNA (Van de Peer, Y. & Wachter, R, 1997).

Como conseqüência deste sucesso, filogenias moleculares substituíram quase que completamente abordagens taxonômicas clássicas baseadas em dados morfológicos, e um grande banco de dados contendo seqüências de rRNA foi gerado. Hoje em dia, portanto, filogenias moleculares baseadas em seqüências de nucleotídeos ou aminoácidos tornaram-se ferramentas indispensáveis para o estudo da evolução molecular.

Os métodos atuais para construção de árvores filogenéticas dividem-se em dois tipos: Métodos baseados em matrizes de distância (Fenéticos) ou métodos probabilísticos (Cladística), sendo distância genética definido como a medida da evolução que a molécula sofreu (substituição, transição ou transversão).

Dentre o primeiro estão os métodos da evolução mínima e agrupamento de vizinhos (NJ, neighbour-joining). No segundo estão os métodos da Máxima parsimônia, da Máxima verossimilhança e inferência Bayesiana.

Métodos de distância primeiramente convertem as seqüências alinhadas em uma matriz de distância que representa uma estimativa da distância evolutiva entre duas seqüências (número de mudanças que ocorreram) enquanto métodos probabilísticos analisam cada sítio de nucleotídeo ou resíduo de aminoácido.

Um dos métodos mais simples e mais populares para construção de árvores filogenéticas é o de agrupamento de vizinhos (NJ). Ele foi desenvolvido por Saitou e Nei em 1987, sendo uma simplificação do princípio da evolução mínima com a diferença de não examinar todas as topologias possíveis mas sim agrupar vizinhos estreitamente relacionados (Saitou and Nei 1987).

Visando aumentar a robustez da topologia de árvores filogenéticas, análises estatísticas foram desenvolvidas e associadas aos métodos evolutivos para inferência filogenética. O método mais popular destes é o "bootstrap", que consiste na reorganização da amostragem aleatoriamente dos dados para determinar erros de amostragem nas topologias das árvores. Na prática os caracteres são reorganizados aleatoriamente com substituições de modo a criar vários conjuntos de réplicas bootstrap de uma amostra original. Cada replicata é então analisada através de um dos métodos evolutivos usados em programas de filogenia (Neighbour Joining - NJ, Parcimônia) e a concordância das árvores resultantes são resumidas em uma "majority-rule consensus tree" ou simplesmente árvore consenso. As freqüências de ocorrência de grupos concordantes ou BP ("bootstrap proportions") são a medida de suporte final para estes grupos e conseqüentemente para a topologia da árvore.

É importante ressaltar que "bootstrapping" não deve ser tratado da mesma forma como intervalos de confiança. Não existe unanimidade sobre o que seria um bom valor de bootstrap (>70%, >80%, >85%?). Uma árvore pode ter valores de bootstrap muito altos mas estar completamente errada, sendo este mais uma medida de precisão do que correção.

Incongruência em árvores filogenéticas

Dentre os critérios para comprovação de HGT encontra-se a incongruência filogenética. Portanto os fatores responsáveis pelas incongruências das árvores filogenéticas podem ser classificados em biológicos e metodológicos. Entre os biológicos destacam-se os fenômenos da transferência horizontal de genes (Horizontal Gene Transfer – HGT), tratado posteriormente (Syvanen 1994), e a duplicação de genes seguida de perdas gênicas randômicas e independentes (Wolfe and Shields 1997). Por exemplo, quando um número de seqüências de eucariotos cai em um clado bem suportado (bootstrap) de seqüências de procariotos, fala-se de filogenia incongruente. De maneira geral quando observamos uma incongruência entre a "árvore do gene" e a "árvore da espécie", um caso para transferência horizontal de genes pode ser feito.

Entre os fatores de natureza metodológica observa-se o uso de modelos evolutivos incorretos para a reconstrução de árvores filogenéticas. Talvez o problema mais conhecido seja o "long branch attraction" (LBA), ou atração dos galhos longos, em árvores filogenéticas *(Moreira and Philippe 2000)* que pode ocorrer quando o método parsimônia é utilizado para inferência da filogenia. Este problema foi primeiramente identificado por Felsenstein quando observou que as taxas evolutivas eram diferentes entre espécies diferentes (Felsenstein 1978). No caso específico de comparação de um conjunto de seqüências que evoluem a taxas distintas, é freqüente que aquelas que evoluem a taxas mais rápidas sejam desenhadas juntas erroneamente. Isto ocorre pois seqüências de rápida evolução (altas taxas de substituição) estão mais sujeitas a dividirem caracteres idênticos aleatoriamente (falsas sinapomorfias) do que seqüências que evoluem vagarosamente (pequenas taxas de substituição). Este fenômeno é relevante no caso de HGT, pois este tipo de evento usualmente é acompanhado de "evolução acelerada" (Koonin, Makarova

et al. 2001), portanto sujeito a “long branch attraction”. Portanto na ausência de um modelo adequado para medir a evolução das seqüências este problema ocorre e as seqüências que evoluem a taxas mais rápidas são agrupadas juntas erroneamente, independentemente de suas verdadeiras relações. Como atualmente os modelos estão tendendo a um excesso de simplificação, LBA ocorre freqüentemente sempre que as diferenças das taxas evolutivas das seqüências existir. Um problema adicional de LBA é o fato de a seqüência de “outgroup” possuir freqüentemente ramos longos. De modo que as próprias seqüências evoluindo rapidamente no grupo interno não somente atraem umas as outras mas também o “outgroup”, levando ao seu posicionamento errôneo na base da árvore.

Foi concluído que devido a este relevante problema de atração dos galhos longos nossa visão da evolução de certos grupos foi infelizmente influenciada através de filogenias incorretas resultantes de LBA. Um exemplo disto ocorreu no filo *Microsporidia*. Estes protistas são parasitas altamente divergentes e que não possuem mitocôndria e ramificam na base da árvore eucariótica de SSU rRNA (Leipe, Gunderson et al. 1993). Devido a esta posição eles eram considerados réplicas vivas de uma era pre-mitocondrial da evolução dos eucariotos. Os problemas com este modelo começaram a surgir quando genes de origem mitocondrial foram encontrados em seu genoma além do fato de filogenias construídas com várias outras proteínas que não SSU rRNA indicaram seu posicionamento próximo a fungos. Este posicionamento foi ainda suportado pela presença de quitina na parede de seus esporos além da similaridade de seu ciclo de vida com o de outros fungos (Muller 1998). Portanto estes protistas são na verdade um grupo de fungos que experimentou um forte aumento da taxa evolutiva de alguns de seus genes como o gene de SSU rRNA que foi usado para reconstrução da árvore errônea. Este aumento da taxa evolutiva do gene de SSU rRNA ocorre provavelmente devido a seu estilo de vida parasítico, caracterizado pela redução da atividade metabólica e “gargalos” seletivos. Portanto sua rápida taxa evolutiva explica seu posicionamento filogenético basal como resultado de um artefato causado por LBA.

De um modo geral filogenias são trabalhosas e consomem muito tempo além de dependerem em última instância de seqüências corretas de alinhamento, portanto são muito difíceis de serem automatizadas sem comprometer sua qualidade.

Genomas de organelas e HGT: a teoria endosimbiôntica

O conceito de transferência horizontal aplicado a teoria endosimbiôntica possui uma longa história, tendo sido descrito pela primeira vez em 1905 por Mereschkowvski que argumentou que cloroplastos foram endosimbiontes bacterianos (Mereschkowsky 1905). A mais convincente demonstração da teoria foi realizada por Woese & Fox usando rRNA de mitocôndria e cloroplasto, os quais eram mais similares ao de bactéria do que seus próprios homólogos nucleares (Woese 1977; Woese and Fox 1977). Este artigo, aliás, é um marco em estudos de transferência horizontal pois é o primeiro a apresentar testes de incongruência filogenética para provar HGT.

Foi sugerido ainda que adicionalmente ao evento endosimbiôntico original que formou a mitocôndria e os plastídeos, outros eventos adicionais de HGT ocorreram. Outros estudos

envolvendo filogenias incongruentes com diferentes marcadores filogenéticos (rubisco e 16SrRNA) demonstraram que de fato pode ter ocorrido mais de um evento simbiôntico (Shivji, Li et al. 1992). Além disso, existem inúmeras descrições de DNA organelar que moveu para o núcleo e estes eventos são importantes pois definem parte da via para HGT mediada por endosimbiosis. Estes tipos de transferência tanto podem ter ocorrido há muito tempo em uma escala evolutiva, como podem ter sido adquiridos recentemente, como exemplos com o gene da tierredoxina F e GAPDH (Erwin and Valentine 1984), (tabela 1).

Mecanismos moleculares de HGT

De acordo com Bushman (Bushman 2002) são três os mecanismos abrangentes que mediam um movimento eficiente de DNA entre as células: transdução, conjugação e transformação.

Transdução refere-se a transferência de uma seqüência de DNA entre células através de vírus. Quando uma seqüência de DNA é introduzida em uma nova célula hospedeira, ela pode ser integrada no genoma através de recombinação homóloga.

Conjugação é a transferência direta de uma seqüência de DNA de uma célula para outra. Normalmente conjugação refere-se a transferência de DNAs entre células bacterianas, entretanto *Agrobacterium tumefaciens* utiliza-se de um mecanismo essencialmente idêntico para transferir DNA para células eucarióticas. O mecanismo de conjugação envolve um aparato especial chamado de pilus para juntar fisicamente duas células fazendo que entrem em contato direto. Uma fita de DNA da célula doadora adentra a célula recipiente. Caso o DNA transferido seja integrado ao genoma da célula recipiente, esta pode adquirir novas características codificadas pelo DNA transferido.

Transformação normalmente envolve simplesmente a absorção por uma célula de DNA presente no ambiente que a cerca e expressão deste DNA na célula recipiente. Algumas células bacterianas são naturalmente competentes para absorção de DNA.

Outro mecanismo relacionado a HGT é a transposição e pode ser definido como a transferência de segmentos de DNA de uma região do genoma para outra. Estes segmentos móveis de DNA são chamados de transposons após sua descoberta em 1950 por Barbara McClintock. Transposição não é um mecanismo de transferência horizontal *per se*, pois não ocorre necessariamente saída de material genético de uma célula e entrada em outra ou mesmo transferência entre genomas mas em um mesmo genoma. Entretanto, alguns casos de transposição são diretamente associados com transferência horizontal porque podem “ligar” regiões vizinhas de transposons a veículos de transferência horizontal como plasmídeos conjugativos ou vírus. Transposons representam um dos mais convincentes exemplos de HGT em eucariotos como no caso do transposon *mariner* descoberto primeiramente em *Drosophila* (Jacobson, Medhora et al. 1986) e posteriormente descritos em outras espécies, incluindo dois casos em humanos (Robertson, Zumpano et al. 1996).

Existem dois modos alternativos de transposição: um que usa um RNA intermediário para transposição, denominado de retrotransposição e um outro que não utiliza RNA como intermediário.

Retrotransposons replicam-se através de transcrição reversa, através da proteína integrase, de modo a integrar o cDNA linear no genoma. Retrotransposons do tipo LTR, com longas repetições terminais, são similares a retrovírus, com o único diferencial de que estes não possuem uma fase extracelular em seu ciclo de vida (Bushman 2002). Ambos possuem estruturas e proteínas em comum como *gag-pol* (ver figura 3), sendo que o retrovírus possui ainda um gene *env* que codifica a cápsula protéica, ausente em retrotransposons.

Diversos estudos documentam transferência horizontal envolvendo retrotransposons do tipo LTR, sendo talvez o caso mais conhecido a transferência entre *D. melanogaster* e *D. willistoni* (Flavell 1999). Outros estudos demonstram HGT de retrotransposons, desta vez entre plantas através da descoberta de retrotransposons similares em espécies evolutivamente distantes de plantas hospedeiras, como é o caso da família de retrotransposons tipo Ty1-*copia.*

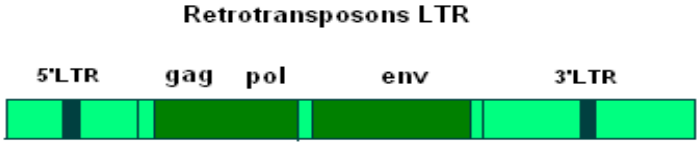


Fig 5. Estrutura de transposons eucarióticos de tipo LTR.

Este retrotransposon foi seqüenciado em oito espécies diferentes de plantas evolutivamente distantes. A filogenia dos transposons mostraram-se incongruentes em relação a filogenia do hospedeiro (planta) sugerindo HGT (Flavell 1992). De um modo geral o fato de que a família de retrotransposons do tipo Ty1-copia é distinta em fungos, plantas e animais é consistente com a hipótese de que extensiva transferência tenha mantido o caráter específico de cada família de retrotransposons (Flavell 1992; Bushman 2002).

Existem dois tipos de mecanismo para transposição envolvendo transposons que não necessitam de um RNA intermediário; transposição replicativa, que gera uma segunda cópia do transposon e o reintegra em um novo sítio e transposição conservativa, que envolve corte e reintegração do elemento em outro sito no cromossomo. O processo de integração do transposon em um novo sítio é catalisado pela enzima Transposase, codificada por um dos genes do transposon.

Além do transposon *mariner* descrito anteriormente, transposons tipo PIF ("P instability factor") foram identificados em *Zea mays* (Zhang, Feschotte et al. 2001), sendo estes associados a uma nova família de transposons recém identificados denominados de "tourist like MITEs" ("Miniature inverted repeat transposable elements") (Zhang, Jiang et al. 2004). Resumidamente MITEs são pequenos trechos de DNA, não autônomos que são abundantemente espalhados em genomas de plantas. Podem ser divididos em dois grupos: "Tourist like" e "Stowaway like". Como

MITEs não codificam proteínas eles dependem de transposases codificadas por outros elementos: “P Instability factor” no milho (PIF) e outro no arroz –(PONG) foram identificados como transposases usadas por MITEs. Estes transposons (PIFs e PONGs) são comuns em todos os eucariotos, tanto em genomas de plantas como fungos e animais.

Foi demonstrado ainda que elementos de transposição do tipo P são regulados de modo a somente “pularem” eficientemente em células de linhagem germinativa e não na linhagem somática e isto é feito através da regulação do “splicing” do gene da transposase (Bushman 2002). Existem estudos que indicam HGT de transposons do tipo P em populações naturais de Drosophila através do seu parasita *Proctolaelaps regalis* (Bushman 2002).

Um exemplo de HGT envolvendo transposons em fungos ficou evidente após um estudo da distribuição das seqüências homólogas do transposon Fot1, que é encontrado muito bem representado em linhagens de *F. oxysporum,* mas ausente em espécies filogenéticamente relacionadas de Fusarium sendo encontrado entretanto em outras espécies mais distantes (Daboussi 1997).

De maneira mais abrangente, transferência horizontal ocorre tanto em procariotos como em eucariotos. Muitos dos princípios descrevendo HGT em procariotos são os mesmos em eucariotos, de maneira modificada. Segundo Bushman (Bushman 2002), transferência horizontal foi uma força central na construção dos cromossomos de ambos procariotos e eucariotos. Segundo o autor elementos móveis (transposons) entraram no genoma de ambos, proliferaram e criaram mudanças herdáveis no DNA. Mudanças genômicas devido à integração de DNA, multiplicação de transposons ou integração de vírus “geram novas regiões de homologia que servem como substrato para recombinação homóloga celular”, favorecendo troca de material genético. O autor conclui que em ambos, eucariotos e procariotos, grupos de genes foram adquiridos de outros organismos e incorporados em seu genoma. As condições para ocorrência de transferência horizontal diferem entre eucariotos e procariotos; em eucariotos qualquer DNA entrando em uma célula deve atravessar não somente a membrana citoplasmática como a membrana nuclear também, ao passo que procariotos não apresentam membrana nuclear. Conseqüentemente, eucariotos exigem sinais direcionando distribuição dentro da célula. Outra diferença é a ausência de conjugação em células eucarióticas. Nada que impeça, entretanto a presença de mecanismos parecidos com conjugação que possam transferir DNA de procariotos para eucariotos.

Portanto em eucariotos, para uma seqüência de DNA mover-se entre células, deve sair da membrana do doador e entrar pela membrana da célula receptora. O DNA transferido deve então adquirir um estado na célula do recipiente que permita sua persistência no genoma. Um exemplo de como isto poderia acontecer foi demonstrado recentemente (Nitz, Gomes et al. 2004), sendo a integração de kDNA do patógeno nematóide *T. cruzi* em células tronco embrionárias de hospedeiros mamíferos e aves comprovada. Células tronco embrionárias são pluripotentes e podem se desenvolver em qualquer tipo de célula, inclusive células de linhagem germinativa.

Foi visto que diversos casos em que existe a suspeita de HGT em eucariotos envolve elementos móveis “egoístas”, como por exemplo, introns, transposons e plasmídeos. Este quadro não é diferente para fungos com algumas exceções em que há algum tipo de relação evolutiva envolvida (Rosewich and Kistler 2000).

Além de fagos existem diversos outros tipos de vírus que transferem DNA entre células como os retrovírus, porque RNAs celulares podem ser encapsulados em vírions, transportados entre células e integrados ao novo hospedeiro. Integração pode ocorrer, ás vezes, em células germinativas, fazendo com que a alteração seja herdada pela próxima geração da célula modificada. Uma vez que o ciclo de vida da infecção é iniciado com a formação de cDNA através de transcrição reversa a partir das fitas molde de RNA, este mesmo cDNA é integrado no genoma do hospedeiro através da ação da proteína Integrase, codificada pelo vírus e que liga o cDNA ao DNA alvo no hospedeiro. Talvez o exemplo mais extremo da conseqüência da infecção de um organismo por um retrovírus é o caso da AIDS. O vírus da imunodeficiência humana, assim como tantos outros vírus apresentando virulência, originou-se através de um “pulo recente” de um vírus animal em humanos (origem zoonótica). O retrovírus endógeno humano (HERV) por sua vez fez o caminho contrário, infectando um progenitor dos macacos modernos há muito tempo atrás, tendo sido este elemento integrado nas células germinativas do primata, tornando-se um provírus endógeno. HIV difere de HERV no sentido de que não há um provírus endógeno de HIV, portanto ele não está envolvido em transferência horizontal de genes, no sentido de que não houve uma modificação herdável das células germinativas, pelo menos por enquanto (Bushman 2002).

IV. Objetivo.

O objetivo deste projeto é a identificação de genes candidatos a transferência horizontal no genoma de *Crinipellis perniciosa* através de comparação genômica, padrão filético, métodos paramétricos e árvores filogenéticas.

Uma vez identificadas os genes (proteínas) serão caracterizados através de anotação de modo a analisar sua provável contribuição nos diversos processos biológicos realizados pelo fungo, principalmente aqueles envolvidos na patogenicidade.

Os candidatos mais interessantes segundo este critério serão selecionados para validação através do desenho de primers, PCR, amplificação a partir de cDNA e sequenciamento.

V. Materiais e métodos

Comparação Genômica

Uma montagem do genoma de *C. perniciosa* composta de mais de 17.000 contigs gerados a partir dos reads obtidos do seqüenciamento do genoma de *C. Perniciosa* foi usada para uma comparação de genoma inteiro através do algoritmo BLASTX contra um banco de dados contendo apenas seqüências de proteínas de fungos.

Este banco de seqüências de proteínas de fungos continha todas as seqüências de proteínas de fungos disponíveis no GenBank NR (77.821 seqüências) na época desta comparação (03/2004).

Este banco de fungos foi gerado através de um script PERL que selecionou todas as seqüências de proteína de fungos do banco de dados mundial através de um script por palavras-chave que selecionava apenas as seqüências do NCBI que possuíam nomes de fungos, através de uma comparação por palavra-chave contra uma lista contendo apenas nomes fungos, inserida no programa.

Identificação de "open reading frames"(ORFs)

Orfs foram identificadas com o programa ORF finder (www.ncbi.nlm.nih.gov/gorf/gorf.html).

Alinhamento Múltiplo

Alinhamentos múltiplos das proteínas obtidas através do GenBank e as proteínas candidatas a HGT de *C. perniciosa* foram construídos com as opções padrões do programa ClustalX versão 1.8 (Thompson, Higgins et al. 1994) e editadas para remoção de regiões com baixa qualidade de alinhamento com o programa GeneDoc (Nicholas, Jr. et al. 1997).

Análise filogenética de genes candidatos a HGT

Todas as árvores filogenéticas foram construídas com o programa PHYLIP versão 3.2 (Felsenstein 1989). Matrizes de distância foram geradas com o programa PROTDIST usando a opção "Dayhoff" para calcular a probabilidade de mudança de um amino ácido para outro com a matriz PAM-001(1%). O programa NEIGHBOR foi usado como método de clusterização das árvores com a opção Neighbour-Joining (N-J). Os programas SEQBOOT e CONSENSE foram usados para estabelecer limites de confiança dos pontos de ramificação de 1000 replicatas de BOOTSTRAP. As árvores filogenéticas construídas através do PHYLIP foram visualizadas através do programa TREEVIEW (Page 1996).

Todas as filogenias foram construídas através da seleção de seqüências que apresentassem alta similaridade (alinhamentos) de BLASTP entre a seqüência da proteína candidata a HGT ("query") e seus alinhamentos mais significantes estatisticamente em bancos de dados. Portanto aquelas seqüências apresentando baixos valores de e-value com um "valor de corte" de e<-05 e valores altos de alinhamento em bits foram incluídos na filogenia.

Os critérios usados para definir quais candidatos teriam a filogenia inferida foram; 1) Blastn dos candidatos para checar possíveis contaminantes; 2) Candidatos a HGT que representavam preferencialmente o gene inteiro. Caso o candidato seja apenas um fragmento do gene, seu alinhamento múltiplo foi checado de modo a conferir possibilidade de recuperação de seu sinal filogenético. 3) Proteínas interessantes de um ponto de vista bioquímico ou relacionadas a patogenicidade do fungo. 4) Candidatos que obtivessem alta similaridade ou homologia com proteínas de plantas (e-value), levando-se em consideração a relação saprofítica de *C. perniciosa* com suas plantas hospedeiras.

Desenho de Primers

Primers específicos foram desenhados com o programa PRIMER3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) a partir das ORFs preditas dos contigs candidatos com a finalidade de amplificá-lo através de PCR e posterior clonagem. Foram adicionados sítios para enzimas de restrição palíndrômicos nas extremidades 5´ (Nde1;CATATG) e 3' (BamH1;GGATCC) do oligo para posterior corte dos sítios de restrição para isolar o gene (seqüências em azul, tabela 2).

No total foram desenhados 12 oligos, sendo três primers para cada gene (dois "forward" e um reverso) para um total de quatro genes candidatos. Isto foi feito de modo que um dos primers "forward" para cada gene pulasse o peptídeo sinal de cada gene (For2). Estes primers foram então usados para realização de PCR com DNA genômico total de *C. perniciosa* como molde.

Tabela 3. Lista de Oligonucleotídeos usados para PCR neste trabalho. Trechos marcados em azul representam sítios de corte para enzimas de restrição NDE1 e BamH1 para clonagem direcional.

Oligonucleotídeo	Seqüência
Pólen	
PoFor1	CATATGAAGCTTGAATCTTTTTCG
Pofor2	CATATGACGGGGAAGGTCTACTG
PoRev1	GGATCCCCAAGTAATGCTATTATG
DAD	
DaFor1	CATATGAAGAGGACGACGACGA
DaFor2	CATATGGATCTGTACGTGGCCTT
DaRev1	GGATCCCTCCTCAAGCTATGTA
Transposase	
TrFor1	CATATGGTATCACCAGAGGAGCAA
TrFor2	CATATGCAATTGTTGGACTCCCT
TrRev1	GGATCCTTGAAAAGCTCCTTGTAA
THN	
ThFor1	CATATGACCGGCGGTACTGTT
ThFor2	CATATGAGGTGCAGGAAACCT
ThRev1	GGATCCAATGATATTGTCGAAGG

Extração de DNA genômico de *C. perniciosa*

Extração de DNA genômico do micélio de *C. perniciosa* na fase necrotrófica foi feito segundo o método tradicional de extração de DNA de fungos e plantas (Ausubel, Brent et al. 1999) com modificações:

O micélio foi macerado em nitrogênio líquido, distribuído em 09 tubos de 1,5mL e acrescentado 500 uL de tampão de extração CTAB em cada tubo [CTAB 2% (p/v), NACL 1,4M,

EDTA 20mM, TRIS-HCL 100mM pH 8.0, PVP 1% (p/v) e β-mercaptoetanol 0,2% (v/v)]. Esses tubos foram incubados a 65°C por 30 minutos. Após este período de incubação foi adicionado em cada tubo 500uL de clorofórmio: álcool isoamílico na proporção de 24:1 respectivamente, misturados e centrifugados a 12.000rpm em temperatura ambiente por 15 minutos para a separação das fases orgânica e aquosa. A fase aquosa foi transferida para um novo tubo e a ela foi adicionado 0,6 volume de isopropanol na temperatura de aproximadamente -20 °C e centrifugado a 12.000rpm por 15 minutos a 4 °C para a precipitação do DNA. O precipitado foi lavado com 500uL de etanol 70% (v/v) centrifugado a 12.000 rpm por 5minutos. Após a secagem do precipitado, esse foi dissolvido em 30uL de tampão TE (Tris-HCl 10 mM pH8.0 e EDTA 1 mM) acrescido de 1uL de RNAse (10mg/mL) e incubado a 37°C por 1hora. O DNA foi armazenado a –20°C para posterior utilização.

Amplificação dos fragmentos gênicos por PCR ("Polymerase Chain Reaction")

Todas as reações de PCR (Mullis 1990) foram feitas em um volume final de 20uL, sendo 2uL de DNA, 0.8uL de enzima Taq polimerase (Promega), 2uL de tampão da enzima, 4uL de dNTPs (1,5pmol/uL), 2uL cada de primers específicos sense e antisense na concentração de 15pmol/uL e (7uL) de água milli-Q. Essas reações foram realizadas em termo ciclador (Peltier tetrad PTC-225 (MJ Research), seguindo o programa de amplificação de: denaturação inicial (94°C por 4 minutos) seguido de 30 ciclos compostos por denaturação (94°C por 40 segundos), anelamento do primer (58°C por 40 segundos) e extensão (72°C por 1 minunto); finalizando a 4°C por 4 minutos.

Purificação do produto de PCR

Após a amplificação, os fragmentos foram purificados usando o kit S.N.A.P (Invitrogen).

Clonagem direcional de um homólogo de Póllen Ole E 1 em *C. perniciosa.*

Uma banda de aproximadamente 700pb contendo as enzimas de restrição Nde1 e BamH1 amplificada através de PCR do DNA genômico foi clonado no vetor pGEMT-Easy (Promega) através de ligação (T4 DNA Ligase) da banda isolada em gel "low melting". Posteriormente o plasmídio foi transformado em E.coli DH10b (Invitrogen) e selecionado por X-gal/IPTG. Foi realizado um PCR das colônias brancas para a seleção dos clones recombinantes.

RT-PCR

A transcrição reversa seguida de PCR foi utilizada para a síntese de cDNA fita simples, a partir do RNA total extraído por Trizol (*Invitrogen*) e tratado com DNase 1 (50 U/μl, Invitrogen), e subsequente amplificação do cDNA por PCR com primers específicos.

Nesta reação os oligos (dT) se ligam a cauda poli-A na extremidade 3´ do mRNA e sintetizam cDNAs. A síntese de cDNA seguiu a seguinte reação: 4μl de RNA total, 1μl 10mMdNTPs, 5μl tampão1st5X, 1μl transcriptase reversa [SuperScript II-GibcoBRL (200U/μl)], 3μl 15pM iniciador (Oligo dT), 1μl DTT e 10μl de água milliQ para completar volume final de 25μl. Primeiramente a mistura de água, oligonucleotídeos, dNTPs e RNA foi colocada à 70°C por 10 minutos e resfriada

em gelo. Acrescentaram-se os demais componentes e as reações ocorreram à 42°C por 1 hora. Em seguida, a enzima foi inativada por 15 minutos à 70°C.

Seqüenciamento

O seqüenciamento dos produtos de PCR foi feito segundo o método de Sanger (Sanger 1977) . As reações de seqüenciamento foram feitas para um volume final de 10uL usando 2uL do mix "DYEnamic ET Terminator Cycle Sequencing Kit – Amersham" que contém a enzima Taq DNA polimerase, dNTPs, Dye terminators , TrisHCl pH9.0 e $MgCl_2$; 2uL de primer específico (1,5pmol/uL), 2uL de DNA (10ng/uL), 2uL de tampão Tris-Cl- $MgCl_2$ pH9.0 e 2uL de água deionizada. Essas reações foram realizadas em termo ciclador (Peltier tetrad PTC-225 (MJ Research), seguindo o programa de amplificação de: denaturação inicial (95°C por 1 minutos) seguido de 25 ciclos compostos por denaturação (95°C por 20 segundos), anelamento do primer (58°C por 15 segundos) e extensão (60°C por 60 minutos); finalizando a 4°C por 4 minutos.
As reações foram precipitadas com 8uL de água deionizada, 32 uL de etanol 95%, centrifugadas a 12.000rpm por 30 minutos, seguido por lavagem com 100uL de etanol 70% centrifugadas a 12.000rpm por 10 minutos e secagem em temperatura ambiente no escuro.

Estas reações foram seqüenciadas em seqüenciador automático plataforma AbiPrism 377 (Perkin Elmer) adquirido pelo Instituto de Biologia, departamento de genética e evolução da UNICAMP.

Análise de "codon usage"

Genes selecionados para uma análise de "Códon Usage" foram computados com o programa CodonW (Peden) e representados gráficamente através de uma análise estatística bivariada com o programa SPADN (Bélgica).

O conjunto de genes candidatos a HGT selecionados para esta análise obedeceu a um critério de presença de códons de terminação dentro de ORFs (Peden) (15 candidatos apresentando múltiplos códons de terminação dentro de suas ORFs foram excluídos).

Análise do conteúdo G+C

Um script PERL foi gerado para contar as bases G e C das seqüências nucleotídicas de cada candidato a HGT e gerar uma porcentagem deste valor relativo ao total de bases da seqüência.

VI) Resultados

Comparação Genômica

Os 17.906 contigs resultantes de uma montagem rascunho do genoma de *C. perniciosa* foram comparados (blastX) contra um banco de dados composto por seqüências de proteínas de fungos (77.821 seqüências) e todos os contigs "no-hits" (e>-5) selecionados. Como resultado desta

primeira comparação foi obtido um primeiro banco com 7.824 seqüências “no hits” (órfãs), ou 43.64% das seqüências da montagem inicial. Este resultado serviu para eliminarmos seqüências gênicas sujeitas a transferência vertical de modo a selecionar resultados que não apresentassem “hits” com seqüências de fungos. Uma nova comparação foi feita usando as seqüências órfãs do primeiro banco, desta vez contra o banco de dados NCBI-NR de modo a detectar altas similaridades com seqüências de organismos distantes evolutivamente a fungos através de BLASTX, usando um e-value de e<-5 como valor de corte, gerando 359 “hits” ou 2% da montagem rascunho original.

317 destas 359 seqüências foram eliminadas por não apresentarem ORFs com similaridade contra seqüências anotadas em NCBI-nr (BLASTP) e uma pequena parte destas seqüências foi excluída por ser consideradas contaminações bacterianas (*E. coli*). Somente ORFs ou genes apresentando similaridade a seqüências codantes (CDSs) anotadas no NCBI-nr foram selecionados.

Portanto os 43 contigs de *C. perniciosa* foram selecionados, suas ORFs encontradas e anotadas através de similaridade por BLASTP contra o banco de seqüências de proteínas do NCBI-NR (figura 4).

Baseado no alinhamento destas 43 ORFs preditas com seus homólogos anotados no NCBI-NR apenas sete são prováveis genes completos enquanto que os outros 36 candidatos consistem de fragmentos gênicos. Deste total, 16 são similares a genes de plantas, sendo 06 destes elementos móveis. 19 apresentavam maior similaridade com genes provenientes de bactérias enquanto outros 09 genes candidatos são mais similares a genes de origem animal.

Uma vez identificados os genes candidatos a transferência horizontal no genoma de *C. Perniciosa* o próximo passo seria o de validar estes genes, confirmando sua presença no genoma de *C. Perniciosa* através de PCR e sequenciamento, além de eliminar a hipótese de contaminação genômica. Para tanto seriam escolhidos quatro genes da lista de 43 candidatos que obedecessem aos seguintes critérios: 1) genes que apresentem uma alta similaridade com homólogos em plantas; 2) Genes cujos produtos sejam relacionados a patogenicidade ou desenvolvimento de estruturas relacionadas a infecção 3) Genes relacionados a mecanismos de HGT. Baseado nestes critérios de função os genes escolhidos foram: *Cp*PLP (*Cp* para *Crinipellis perniciosa* e PLP para “pollen like protein”), *Cp*TLP (*ibidem* e TLP para “thaumatin like protein”), *Cp*DAD (gene de apoptosis em *C. Perniciosa*) e *Cp*Transposase (transposon relacionado a família de transposons do tipo “tourist like”).

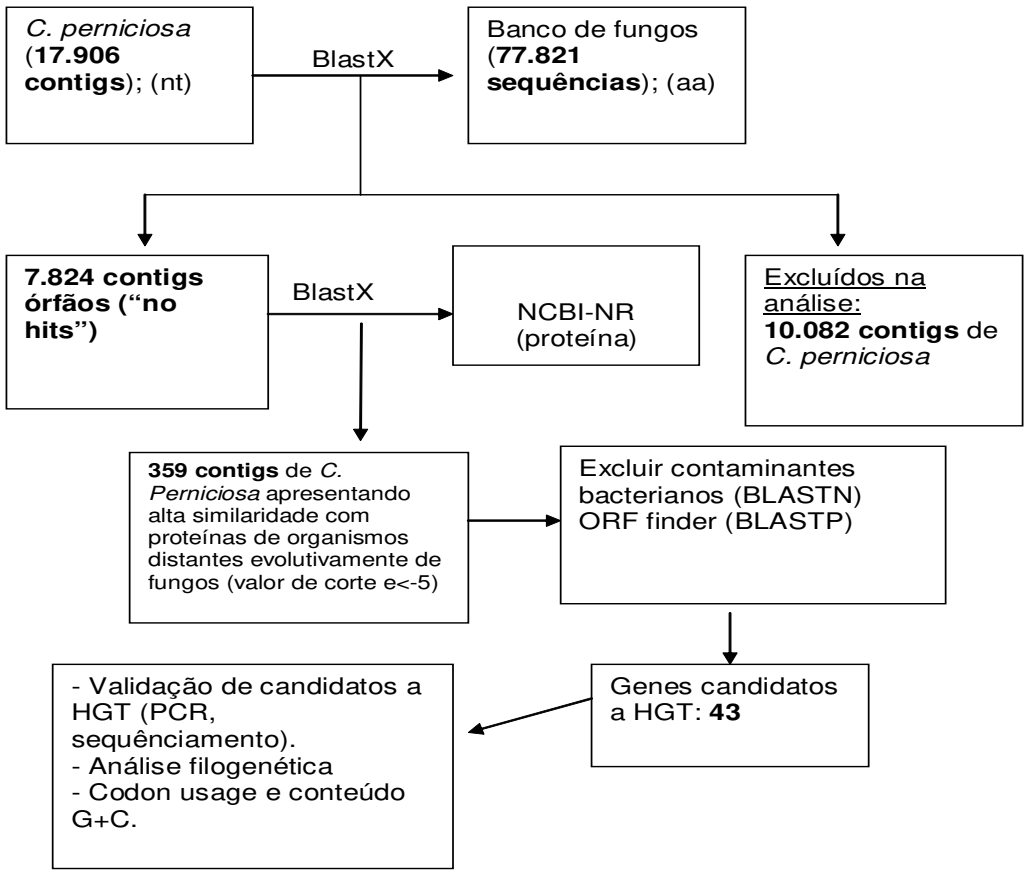


Figura 4. "Pipeline" de comparação genômica funciona como um filtro filogenético para eliminar candidatos espúrios a HGT através de comparação com o algoritmo BLASTX.

Estes quatro genes nunca foram descritos antes em fungos, a exceção do gene THN, e são muitos bem distribuídos em plantas. Podemos visualizar suas anotações bem como suas respectivas funções em negrito na tabela 4.

A proteína antifungo THN (“thaumatin-like”) é parte da família de proteínas PR-5 (“Pathogenesis-related"), muito bem distribuída em diversas espécies de plantas. Esta proteína é relacionada a patogênese, sendo estruturalmente diversa e amplamente distribuída em plantas. A acumulação de proteínas tipo PR, localmente e sistêmicamente, é parte integral do sistema de resposta da planta iniciado por plantas hospedeiras sob stress biótico ou abiótico de modo a impedir o ingresso do patógeno.

Outra proteína importante selecionada foi a proteína DAD (“Defender against cell death”). Membros desta família fazem parte da membrana celular e causam apoptosis (morte celular programada) caso sejam mutadas. É importante ressaltar que o gene *Cp*DAD foi identificado e selecionado com base em uma busca feita através do algoritmo BLASTN com o contig do gene de Pollen contra o banco de dados de *C. perniciosa*. Um dos “hits” surpreendentemente foi um read com alta similaridade a proteína DAD em *Citrus u.* Este resultado indica que estes dois genes estão posicionados em um mesmo contig em cluster no genoma de *C. Perniciosa*, o que configuraria uma ilha de patogenicidade caso suas funções sejam comprovadas.

A proteína de Pollen além de apresentar alta similaridade a seqüências de plantas (figura 5), é muito bem distribuída neste reino. Pollen não possui ortólogos em outros fungos o que caracterizaria talvez o cenário mais robusto de HGT dos candidatos aqui apresentados.

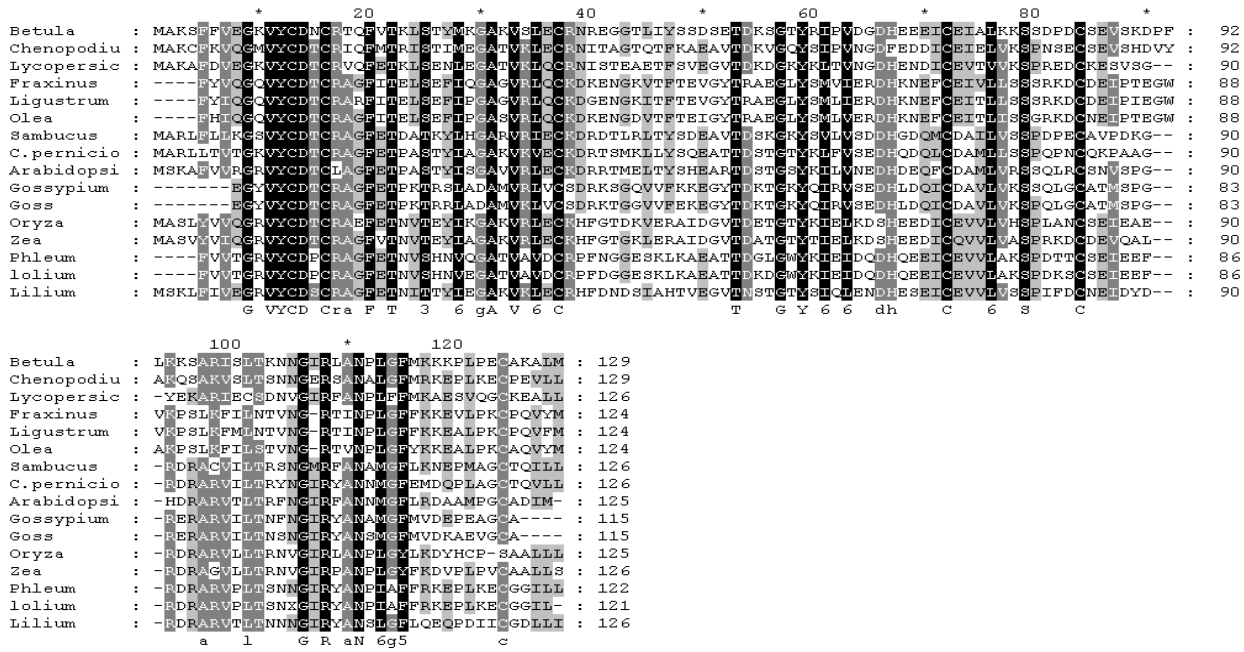


Figura 5. Alinhamento de seqüências de aminoácidos da família de proteínas de Pollen. Aminoácidos marcados em negrito indicam resíduos idênticos em todas as seqüências do alinhamento, enquanto aqueles marcados em cinza indicam substituições bem conservadas. Alinhamento ilustra alto grau de Identidade (>55%) da provável proteína de Póllen em *C. perniciosa* (Pollen_olle_e_1) e homólogos em plantas.

A proteína transposase identificada em *C. perniciosa* apresenta alta similaridade através de BLASTP (59%) ao transposon “P instability factor” (PIF) em *Zea mays*.

Outro motivo que nos levou a optar pela escolha destes quatro candidatos para validação experimental foi o fato destes serem os únicos a apresentarem um peptídeo sinal (sinal localizador) na região N-terminal do gene, um indicativo de que esta proteína seria secretada.

Portanto as seqüências de aminoácidos dos 43 candidatos acima foram examinadas com o programa iPSORT (Bannai, 2002) para detecção do peptídeo sinal. Esta pequena seqüência de resíduos de aminoácidos localiza-se na região N-terminal de proteínas recém sintetizadas e contém a informação que determinará sua localização sub-celular. Como a maioria das proteínas são sintetizadas no citosol elas necessitam destes sinais localisadores para levá-las a organelas específicas da célula. IPSORT é um programa de predição destes sítios N-terminais de localização através de métodos baseados em redes neurais. Quatro candidatos (Pollen_Ole_E_1, DAD, THN e transposase) apresentaram este sinal confirmando a suspeita de que são genes codantes de proteína. Ainda segundo as predições feitas pelo programa, a proteína putativa de Pólen apresenta um peptídeo sinal, DAD e transposase apresentam um peptídeo sinal direcionador para a mitocôndria e THN apresenta um peptídeo direcionador para o cloroplasto.

Podemos dividir os seis elementos móveis encontrados em nossa análise em dois grandes grupos de elementos móveis: retrotransposons e transposons. Os dois retrotransposons encontrados em *C. perniciosa* apresentaram similaridade a retrotransposons com longas repetições terminais (LTR), de tipo cópia enquanto que as quatro transposases, que codificam enzimas que catalizam a transposição apresentaram similaridade a transposons do tipo “P Instability Factor” ou PIF, abundante em eucariotos em geral e em plantas especificamente.

Estes elementos móveis foram demonstrados anteriormente como sendo sujeitos a HGT, como foi descrito anteriormente (mecanismos de HGT) no dramático caso dos *mariners* em eucariotos e também no caso dos retrotransposons LTR de tipo cópia e portanto análises subseqüentes irão indicar HGT destes elementos no caso específico de *C. perniciosa.*

Tabela 4. Anotação de 43 genes e suas proteínas codantes candidatos a HGT em *C.perniciosa* demonstrando homologia inesperada (BLASTP) com organismos evolutivamente distantes. Proteínas em negrito foram amplificadas através de PCR cDNA da fase necrotrófica do fungo.

Gene/Proteína (Contigs)	Estrutura/ função	Organismo mais similar (BLASTP)	E-value (BLASTP)	Tamanho (nt)
Pólen_Ole_e_1 (pfam01190)	**Família de proteínas de pólen. Função biológica desconhecida.**	**Planta; *Arabidopsis***	**e-51**	**528**
HYPK	Proteína Huntingtin	Planta;*Oryza sativa*	e-39	336
Retrotransposon de tipo LTR-cópia	Domínio catalítico do retrotranspson *gag-pol*. Media retrotransposição no hospedeiro.através de transcrição reversa.	Planta;*Oryza sativa*	e-56	1776
Transposase	**Envolvida em recombinação, replicação e reparo.**	**Planta;*Oryza sativa***	**e-28**	**564**
DUF614	Proteínas de função desconhecida. Proteína eucariótica não caracterizada.	Plantas; Petúnia hybrida	e-04	420
KOG0769	Proteína mitocondrial predita	Planta; *Arabidopsis*	e-30	291
Thaumatin (THN)	**Família de proteínas tipo Thaumatin. Reúne proteínas relacionadas a patogênese em plantas. Vários membros desta família demonstram significante atividade in-vitro para inibir o crescimento da hifa ou germinação do esporo de vários fungos.**	**Planta;*Arabidopsis***	**e-07**	**342**
Transposase	Cataliza a transposição de um transposon	Planta; *Oryza sativa*	e-27	615
UDPGT	UDP-glucoronosyl e UDP-glucosyl transferase. Envolvida em transporte e metabolismo do	Planta; *Arabidopsis*	e-11	345
Ps2 (pfam00421)	Proteína de fotossíntese	Planta;*Ascarina*	e-66	363
DUF674	Proteína também encontrada em *Arabidopsis* não caracterizada e de função desconhecida.	Planta;*Oryza sativa*	e-19	279
Transposase predita	Envolvida em replicação, recombinação e reparo.	Planta;*Oryza sativa*	e-23	444
Retrotransposon putativo do tipo LTR-cópia	Contém os genes *gag-pol*	Planta;*Oryza sativa*	e-41	1176
Transposase (pfam 02992)	Transposase_21; família de transposases tnp2.	Planta;*Arabidopsis*	e-09	2124
DAD (pfam 02109)	**Família de proteínas de defesa contra morte celular programada. Causam apoptosis quando mutadas.**	**Planta, *Citrus u.***	**e-55**	**348**

NinG	Proteína do fago bacteriano Lambda	Bactéria; *Yersinia*	e-64	576
Faa1	Envolvido no metabolismo do lipídeo.	Bactéria; *Yersínia*	e-25	351
SapC (pfam07277)	Família de proteína bacteriana de aproximadamente ~250 resíduos. Em Campylobacter f. SapC faz parte de uma "superfície em S", a qual confere resistência a Serum.	Bactéria;*Caulobacter*	e-17	704
COG3956	Proteína contendo domínio de metiltransferase.	Bactéria; *Yersínia*	e-30	345
FKBP_N		Bactéria; *Buchnera aphidicola.*	e-09	741
HLyD	Família de proteínas secretoras	Bactéria; *Serratia*	e-32	444
DKSA	Supressor do Dnak. Envolvido em mecanismos de sinais de transdução.	Bactéria; env. Seq.	e-20	171
SFSA (pfam 03749)		Bactéria; *Yersínia*	e-40	444
SFSA		Bactéria;*Yersínia*	e-35	384
TDO2 e TRP	Triptofano 2,3-dioxygenase (vermílion). Transporte de amino acido e metabolismo.	Bactéria; *Burkholderia fungorum*	e-29	186
Kinase	Kinase	Bactéria; env. Seq	e-16	432
Peptidase M32	Peptidase, Caroxypeptidase	Bactéria; *Yersínia*	e-66	492
COG2911	Proteína não caracterizada conservada em bactérias.	Bactéria; *Yersínia*	e-29	447
MHPC	Hydrolase predita	Bactéria;*Bradyrhizobium japonicum*	e-16	267
Metil-transferase	O-Metiltransferase; Podem estar envolvidas em produção de antibiótico.	Bactéria;*Mycobacterium*	e-08	255
Permease	Permease do tipo ABC do sistema de transporte.	Bactéria; *Burkholderia fungorum*	e-06	288
RnfD	NADH predito, oxireductase; Produção de energia e conservação. Regula negativamente a expressão de fatores de virulência em *Vibrio cholerae* através da inibição do ativador de transcrição ToxT.	Bactéria; *Yersínia*	e-38	510
KOG0804	Proteína tipo dedo de zinco citoplasmático BRAP2	Bactéria; *Microbulbifer*	e-12	999
MtlA	Proteína envolvida no sistema de fosfotransferase, transporte de carboidrato e metabolismo.	Bactéria; *Bacillus s.*	e-08	597
COG3236	Proteína desconhecida conservada em bactéria.	*Yersínia*	e-14	532
ManB e PGM	Phosphomannomutase. Proteína envolvida em transporte de carboidratos e metabolismo; PGM: Phosphoglucomutase/Phosphomannomutase.	Bactéria; *Shigella sonnei*	e-30	478

JmjC (KOG1356)	Este domínio está envolvido na organização da cromatina através da modulação da heterocromatização.	Animais; *Drosófila*	e-17	444
KOG4182	Proteína conservada não caracterizada.	Animais; *Rattus n*	e-15	666
P53 (pfam00870)	Proteína reguladora do ciclo celular.	Animais; *Homo s*.	e-42	456
DAP2 (COG 1506 e KOG2100)	COG: Dipeptidil aminopeptidase. Envolvido no transporte de amino ácido e metabolismo; KOG: Dipeptidil aminopeptidase. Função: Modificação pós traducional, Turn-over de proteína e chaperones.	Animais; *Danio r.*	e-16	525
Hydrolase	Hidrolase secretada predita.	Animais; *Leishmania*	e-13	1095
Heme Oxygenase (pfam01126)	Catalisa conversão da heme para biliverdin. Essencial para reciclar ferro da heme. Também participa da homeostasis da heme, resposta a estresse oxidativo, formação de pigmentos fotossintéticos e aquisição bacterial de ferro da heme do hospedeiro.	Animais; *Gallus g.*	e-13	660
KOG0965	Proteína liga-se ao RNA. Contém SWAP e domínios G-patch que são encontrados em certos retrovírus (pfam01585).	Animais; *C. elegans*	e-05	402

Em uma nova análise com cada um dos 43 genes candidatas a HGT foram submetidas a buscas por similaridade através do algoritmo TBLASTN contra o banco de dados de ESTs de patógenos fúngicos COGEME de forma a encontrar possíveis homologias. Das 43 seqüências apenas cinco apresentaram “hits” significantes contra seqüências de EST (tabela 5).

A proteína DAD em *C. perniciosa* apresentou alta similaridade com ESTs de fungos como, por exemplo, o basidiomiceto *Ustilago maydis* que por sua vez apresentaram resultados de BLASTX contra o NCBI-nr com seqüências de organismos distantes evolutivamente como *X. laevis.*

Dois retrotransposons previamente preditos com base em similaridade por BLASTP em *C. perniciosa* também apresentaram “hits” contra ESTs de patógenos fúngicos em plantas como *M. grisea.* Estes resultados reforçam sua similaridade a retrotransposons de tipo “long terminal repeat” (LTR) da família cópia também identificados pelo algoritmo BLASTP. Estes ESTs por sua vez apresentaram “hits” contra retrotransposons do tipo cópia em plantas.

Outra proteína predita em *C. perniciosa* apresentando alta similaridade a ESTs anotados como proteínas da membrana do peróxissomo de *Ustilago maydis* foi a proteína mitocondrial predita.

Curiosamente a proteína candidata mtla apresentou similaridade alta a um EST considerado como sendo um contaminante de vetor pelo banco COGEME em *C. parasítica,* que por sua vez apresentou alta similaridade a seqüências bacterianas de *E. coli,* indicando a presença de um possível contaminante.

Tabela 5. Busca por similaridade com o algoritmo TBLASTN dos 43 candidatos a HGT contra o banco de dados de ESTs de patógenos fúngicos em plantas COGEME indica homologia apenas para cinco candidatos.

Candidato a HGT (*C. Perniciosa)*	Organismos "hit" (ESTs-COGEME)	e-value	Função predita	1 organismo "hit" EST vs. NCBI-NR	Função
DAD	*U. maydis*	4e-20	defesa contra morte celular programada (DAD)	*Xenopus laevis*	DAD
	C. parasitica	2e-15	defesa contra morte celular programada(DAD)	*Xenopus laevis**	DAD
	M. grisea	8e-14	defesa contra morte celular programada(DAD)	*Xenopus laevis**	DAD
rve1	*M. graminicola*	8e-27	Retrotransposon poliproteína-retroelemento LTR	*Zea mays*	Gag-pol
	B. graminis	2e-12	retrotransposon tipo cópia-LTR	*Arabidopsis*	retrotransposon de tipo cópia
	M. grisea	3e-6	Retrotransposon poliproteína-LTR	*Zea mays*	Gag-pol
rve2	*M. grisea*	5e-4	Retrotransposon poliproteína-LTR	*Zea mays*	Gag-pol
	M. graminicola	8e-4	Retrotransposon poliproteína-LTR	*Arabidopsis*	Gag-pol
Proteína transportadora mitocondrial predita	*U. maydis*	2e-4	Proteína da membrana do peroxissomo 47.	Proteína da membrana do peroxissomo 47	*Candida boidinii*
mtla	*C. parasitica*	1e-9	Enzima de sistema Mannitol-específica PTS/ Contaminante bacteriano ou vetor de clonagem	*E. coli Shigella f.*	Enzima de sistema Mannitol-específica PTS/ contaminante bacteriano

Buscas através dos algoritmos de BLAST com a proteína de pollen não apresentaram similaridade com a seqüência de EST depositadas no banco de dados COGEME. Portanto uma nova busca foi feita através do algoritmo BLASTN usando a seqüência de nucleotídeos da proteína de Pollen como "query" contra o banco de dados EST_Others (ftp://ftp.ncbi.nlm.nih.gov/blast/db/est_others), composto de seqüências de EST de todos os organismos exceto humano e camundongo. Os 5 primeiros hits resultantes podem ser visualizados na figura 6. Estes ESTs poderiam representar ortólogos desta proteína em espécies diferentes ou membros adicionais de uma família de genes.

Todos os “hits” mostrados na figura 6 são de bibliotecas de cDNA de órgãos de plantas e apresentam alta similaridade ao gene de pólen de *C. Perniciosa*. Pode-se observar que o primeiro “hit” é contra um EST de eucalipto, com um valor máximo de e-value de 0 suportando este alinhamento. Inspeção do alinhamento demonstrou 100% de similaridades entre as duas seqüências indicando uma possível contaminação. Análises posteriores através de BLASTN contra o banco de dados do eucalipto confirmaram suspeitas de contaminação apresentando alinhamentos suportados pela nota máxima de e-value (0) e similaridade de 100% com estas seqüencias.

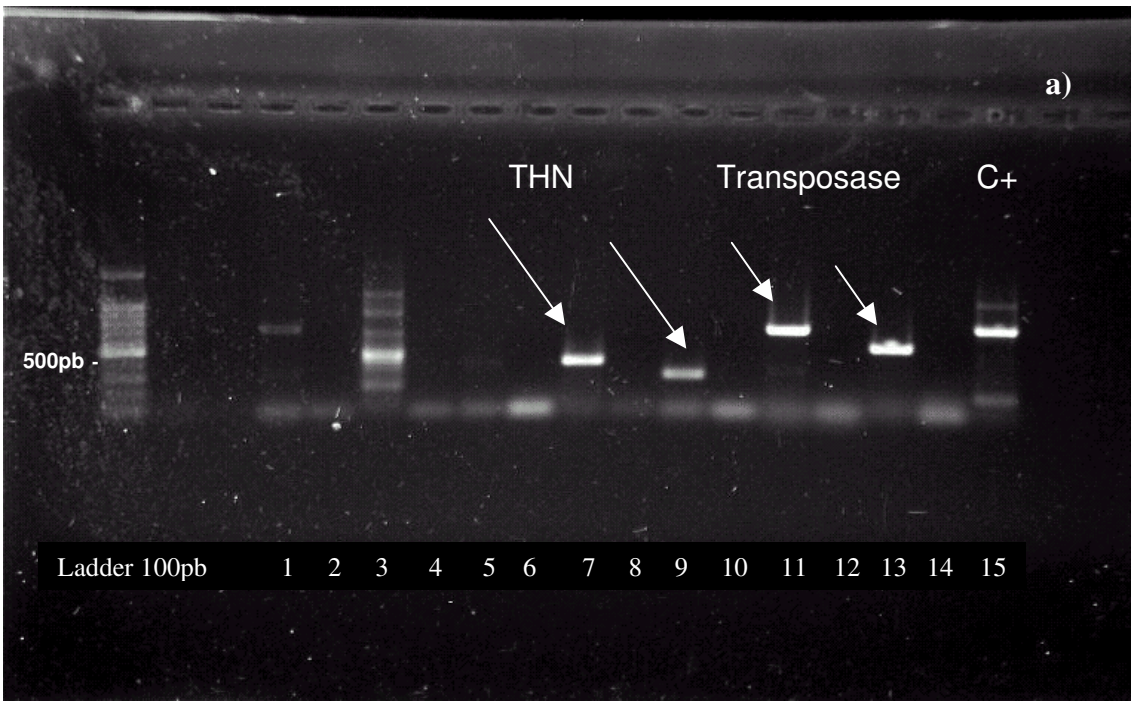
Figura 6. Os cinco primeiros “hits” da seqüência gene de Póllen contra o banco de dados EST_OTHERS através do algoritmo BLASTN demonstram homologia a cDNAs de órgãos de plantas como o Eucalipto e genes relacionados a inflorescência na uva.

```
Query= Contig12357
       (964 letters)

Database: EST: other organisms
        11,409,048 sequences; 6,107,871,864 total letters

Searching.....done

                                                  Score    E
Sequences producing significant alignments:       (bits) Value

gb|CB967989.1|CB967989 egx28a01_F Differentiating xylem Eucalypt...   664   0.0
gb|CF610077.1|CF610077 INFIO01_000915 Grape Inflorescence pSPORT...   135   1e-28
gb|CF609172.1|CF609172 GERMO01_000906 Grape Shoot pSPORT1 Librar...   135   1e-28
gb|CF608663.1|CF608663 GERMO01_000348 Grape Shoot pSPORT1 Librar...   135   1e-28
gb|CF607712.1|CF607712 GEMMA01_001238 Grape Bud pSPORT1 Library ...   135   1e-28
```

Pólen, THN, DAD e Transposase: Caracterização de quatro novos genes candidatos a HGT.

Amplificação por PCR dos genes de Póllen, “Thaumatin” (THN), DAD (“defender against cell death”) e Transposase foi feito a partir do DNA genômico da fase necrotrófica de *C. perniciosa.* Foram usados primers específicos desenhados para flanquearem as extremidades de cada um dos genes a serem amplificados de modo a validar a presença destes genes no genoma de *C. perniciosa.* As bandas resultantes do PCR para estes quatro genes podem ser visualizadas na figura 7 a confirmando a presença destes no DNA genômico de *C. perniciosa* e portanto anulando a possibilidade destes genes serem contaminação genômica ou artefatos da análise de bioinformática. Todos as bandas amplificadas a partir do DNA genômico são compatíveis com os tamanhos dos genes preditos a partir de suas ORFs através do algoritmo BLASTP. Note que estas ORFs não apresentam introns uma vez que estão sendo comparados com seqüências de aminoácidos depositadas a partir de cDNAs (mRNAs dupla fita) e portanto sem íntrons. Com isto

qualquer pequena variação de tamanho entre a seqüência predita e o tamanho amplificado pela banda a parti de DNA genômico deve-se a presença de introns. A banda para o gene de pólen da figura 7 a apresenta ~700 pb enquanto a seqüência predita pela bioinformática para o mesmo gene possui 528 pb. Esta diferença de aproximadamente ~200 pb pode ser devido a presença de introns da banda amplificada do DNA genômico, enquanto que a seqüência predita pela bioinformatica é a proteína codificada do mRNA, portanto sem íntrons. Experimentos subseqüentes (figura 8 b) confirmam o tamanho do gene de Pollen exatamente como previsto pela bioinformática.



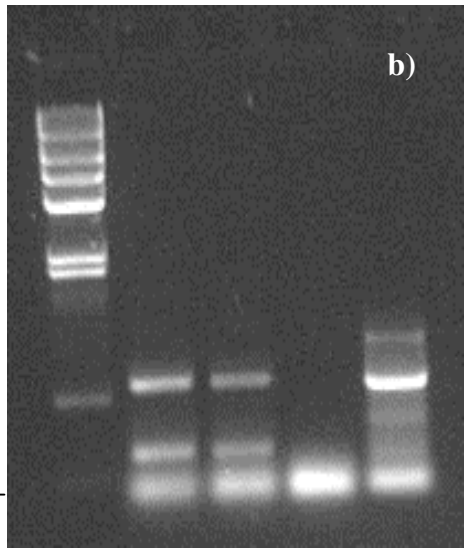


Figura 7.(a) PCR para os genes de Póllen, DAD, THN (setas) e Transposase (setas) confirmando presença destes genes no genoma de *C. perniciosa.* Bandas foram amplificadas com oligos específicos a partir do DNA genômico extraído da fase necrotrófica do fungo. Bandas amplificadas correspondem a números para facilitar visualização de cada gene: 1) Pólen (~700pb); 3) DAD-1 (~350pb); 5 DAD-2 (~342pb); 7) THN-1 (~342pb); 9) THN-2 (~300pb); 11) Transposase-1 (~700pb); 13) Transposase-2 (600pb). Números pares de 1 a 14 são controles negativos deste experimento. Banda número 15 é um controle positivo da proteína de necrose (~700pb, cordialmente cedida).

Gel agarose concentrado a 1.2%, corado com azul. Marcador usado foi a Ladder 100pb (PROMEGA). **(b)** PCR mostrando duas bandas amplificadas com primers específicos para o gene de pólen. Banda maior de aproximadamente 700pb e a menor de aproximadamente 300pb . DNA ladder usado foi λ-hind (1Kb). Gel agarasoe foi concentrado a 1%.

Os óligos usados para amplificar o gene de Póllen (figura 7 a, banda número 1) foram especificamente desenhados de modo a amplificar o gene somente depois do sinal peptídico (PoFor2 e PoREV), composto por 30 resíduos de aminoácidos, de modo a facilitar uma purificação ou expressão desta proteína futuramente uma vez que os sinais peptídicos são caracterizados por serem regiões de alta hidrofobicidade. PCR do gene de Póllen com óligos específicos (PoFor1 e PoREV), desenhados para amplificar o gene inteiro incluindo o sinal peptídico, amplificou bandas não-específicas (resultado não mostrado aqui).

Nos casos específicos de DAD, THN e Transposase, assim como no caso do gene de póllen, óligos foram desenhados de modo a amplificar bandas de dois tamanhos; um par de óligos desenhado especificamente para amplificar o gene inteiro e outro par de modo a pular o sinal peptídico, gerando uma banda menor do que o par de óligos que amplificam o gene inteiro. Estas diferenças de tamanho entre os produtos amplificados para cada gene podem ser visualmente inspecionados na figura 7 a, sendo: bandas 3 e 5 (DAD); bandas 7 e 9 (THN); bandas 11 e 13 (Transposase).

O tamanho das bandas amplificadas neste experimento é similar ao tamanho dos genes preditos pela bioninformática (tabela 4). O tamanho do gene putativo "Thaumatin" (THN) predita pela bioinformática de 342 pb é compatível com o tamanho da banda amplificada neste experimento (bandas número 7 e 9, ~350 e ~300 pb respectivamente).

Da mesma forma, os genes DAD e Transposase com tamanhos preditos pela bioinformática de 348 pb e 564 pb respectivamente são compatíveis com as bandas amplificadas de DNA genômico de *C. perniciosa* que podem ser observadas na figura 7 a.

Amplificação a partir de cDNA

O cDNA foi obtido através da transcrição reversa do RNA total extraído da fase necrotrófica e biotrófica de *C. perniciosa* previamente tratado com DNAse1 de modo a evitar contaminação. O cDNA foi então amplificado por PCR com os oligos específicos dos genes de Póllen, DAD, THN e Transposase como pode ser observado na figura 8 a (setas indicando bandas de número 2, 7, 12 e 17 respectivamente). O cDNA necrotrófico amplificou bandas com tamanhos similares a bandas amplificadas do DNA genômico: Póllen (banda número 4); THN (bandas número 8 e 9); DAD (13 e 14) e Transposase (18 e 19) de *C. perniciosa* para cada um dos quatro genes. Nenhuma banda foi amplificada do cDNA biotrófico de *C. perniciosa*.

Outro PCR feito com os óligos específicos flanqueando o gene de Pollen amplificaram uma banda de aproximadamente 500 pb do cDNA necrotrófico de *C. perniciosa* (setas, figura 8 b), confirmando a ORF identificada previamente de 528 pb de um contig de *C. perniciosa*. Estes mesmos oligos foram usados para amplificação através de PCR do DNA genômico de *C. perniciosa* neste experimento resultando em uma banda de 700 pb (~200 pb maior do que o fragmento amplificado do cDNA) como pode ser visualizado na figura 8b. Esta diferença de

tamanho se deve ao fato da seqüência amplificada do DNA genômico possuir introns ao contrario da seqüência de cDNA (mRNA dupla-fita) já processada portanto sem introns.

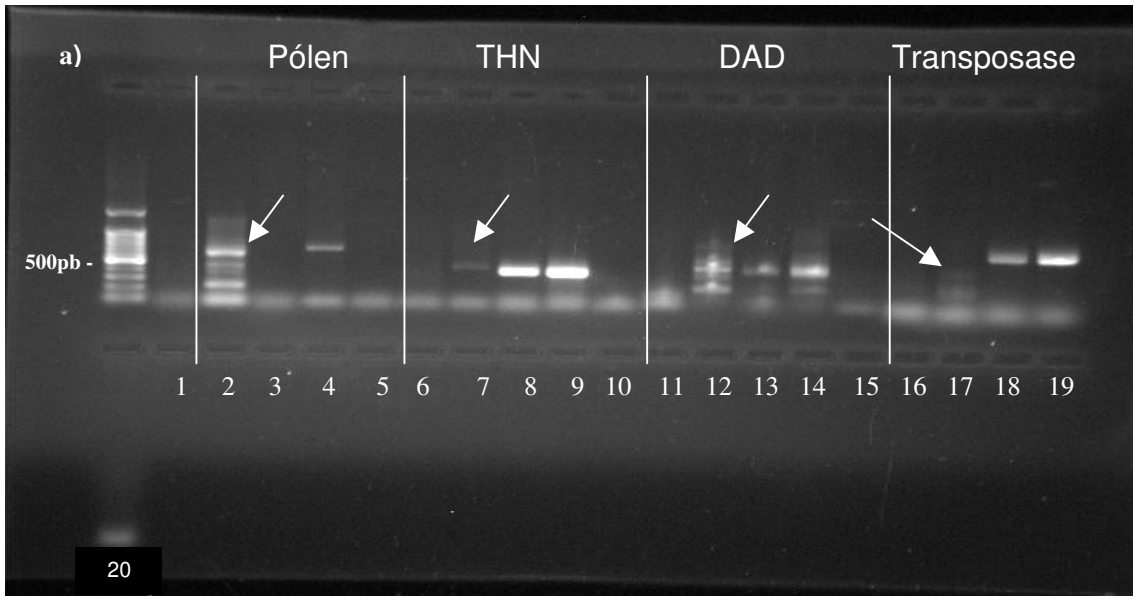
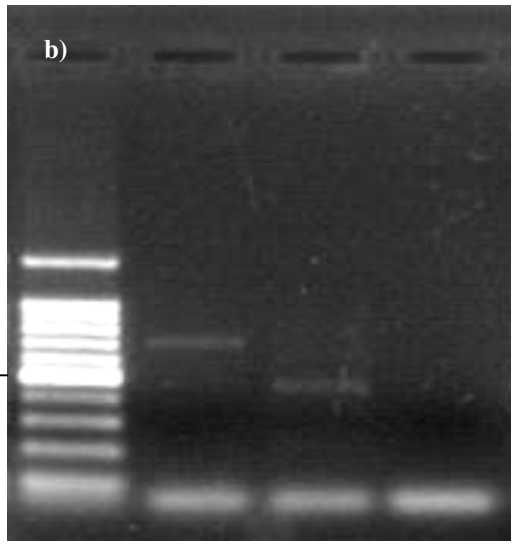


Figura 8 (a). RT-PCR mostrando os produtos (setas) resultantes da amplificação dos genes de Póllen (banda número 2), THN (banda número 7), DAD (12) e Transposase (17) a partir de cDNA necrotrófico de *C. perniciosa*. O controle positivo consiste em bandas amplificadas a partir de DNA genômico de duas cepas de *C. perniciosa* (BP10 e CP02), ambas extraídas da fase necrotrófica do fungo: Pollen (banda n° 4); THN (bandas 8 e 9); DAD (bandas n° 13 e 14) e Transposase (18 e 19). Os poços de número 5, 10, 15 e 20 são controles negativos para o experimento.



8 (b). RT-PCR mostrando a amplificação de uma banda a partir de cDNA para o gene de Pólen (~500pb, seta). DNA genômico foi utilizado como controle positivo. Podemos observar que a banda para o gene de pólen é ~200pb menor do que a banda amplificada a partir de DNA genômico devido a ausência de introns do mRNA dupla fita (cDNA).

Entre os candidatos a HGT cujas bandas foram amplificadas, Pollen e DAD não apresentaram um padrão de bandas consistente em todas as amplificações. Este dado somado ao resultado de BLASTN com o gene de Pollen descrito anteriormente (ver BLASTN, figura 6) confirmou a suspeita de contaminação do banco de dados de *C. perniciosa* com seqüencias de eucalipto. Estas suspeitas foram confirmadas por um BLASTN de todos os 43 candidatos a HGT contra o banco de dados do eucalipto.

Análise filogenética de candidatos a HGT.

Além dos quatro genes descritos anteriormente através de PCR e RT-PCR, outros três genes foram escolhidos da lista de 43 prováveis candidatos a HGT, com base em sua anotação biológica para terem as suas filogenias inferidas; o gene da Heme Oxigenase que apresenta similaridade com seqüência de bactéria, o gene JMJC, envolvido na organização da cromatina e apresentando alta similaridade com *Drosóphila* além do provável retrotransposon LTR do tipo cópia em *C. perniciosa*.

A seguir faremos uma análise detalhada de cada uma destas sete árvores filogenéticas, levando em consideração suas topologias e ramificações, bem como seus valores de Bootstrap, de modo a detectar árvores de genes incongruentes com a árvore canônica universal baseada em seqüências de SSU rRNA, que considera o agrupamento de todos os organismos em três domínios (bactéria, eucaria e archae), sendo que cada um destes é monofilético.

Uma eventual mistura de seqüências de grupos taxonômicos distintos (não necessariamente entre bactéria, archae e eucária) em um clado polifilético também pode ser um indicativo de HGT (plantas e fungos por exemplo) apesar de ambos serem eucariotos, pois HGT poderia ter ocorrido depois da divergência destas espécies.

Na árvore filogenética construída para a família de proteínas de Pollen na figura 9, podemos observar a ramificação de uma seqüência de um provável contaminante de eucalipto detectado no genoma de *C. perniciosa* com uma seqüência de *Arabidopsis,* apresentando um valor de 68% de bootstrap suportando este agrupamento. Caso este gene não fosse uma contaminação de eucalipto a filogenia mostrada na figura 9 implicaria HGT, uma vez que agrupa uma seqüência de fungo em um clado de plantas. No caso deste hipotético evento de HGT ser verdadeiro ele teria ocorrido depois da divergência de plantas e fungos desde seu último ancestral em comum, uma vez que não foi encontrado homólogos do gene de Pollen em nenhum dos diversos fungos até agora seqüenciados. Este fato exclui também a possibilidade de perda gênica, pois seria no mínimo improvável evolutivamente a ocorrência de múltiplos eventos de perda gênica em todos os fungos a exceção de *C. perniciosa*.

Entretanto, como foi identificado uma origem vegetal para os genes de Pollen e DAD através de contaminação do banco de dados de *C. perniciosa* por seqüências de eucalipto, as filogenias construídas logo abaixo para estes genes ramificaram estas seqüências com homólogos de plantas e portanto não caracterizam um evento de HGT. Os nomes do organismo de origem destes genes contaminantes nas filogenias de Pollen e DAD foram mantidos como sendo de *C. perniciosa* ao invés de eucalipto de modo a indicar a detecção da contaminação genômica através das filogenias e também para ilustrar um hipotético caso de HGT através de incongruência filogenética.

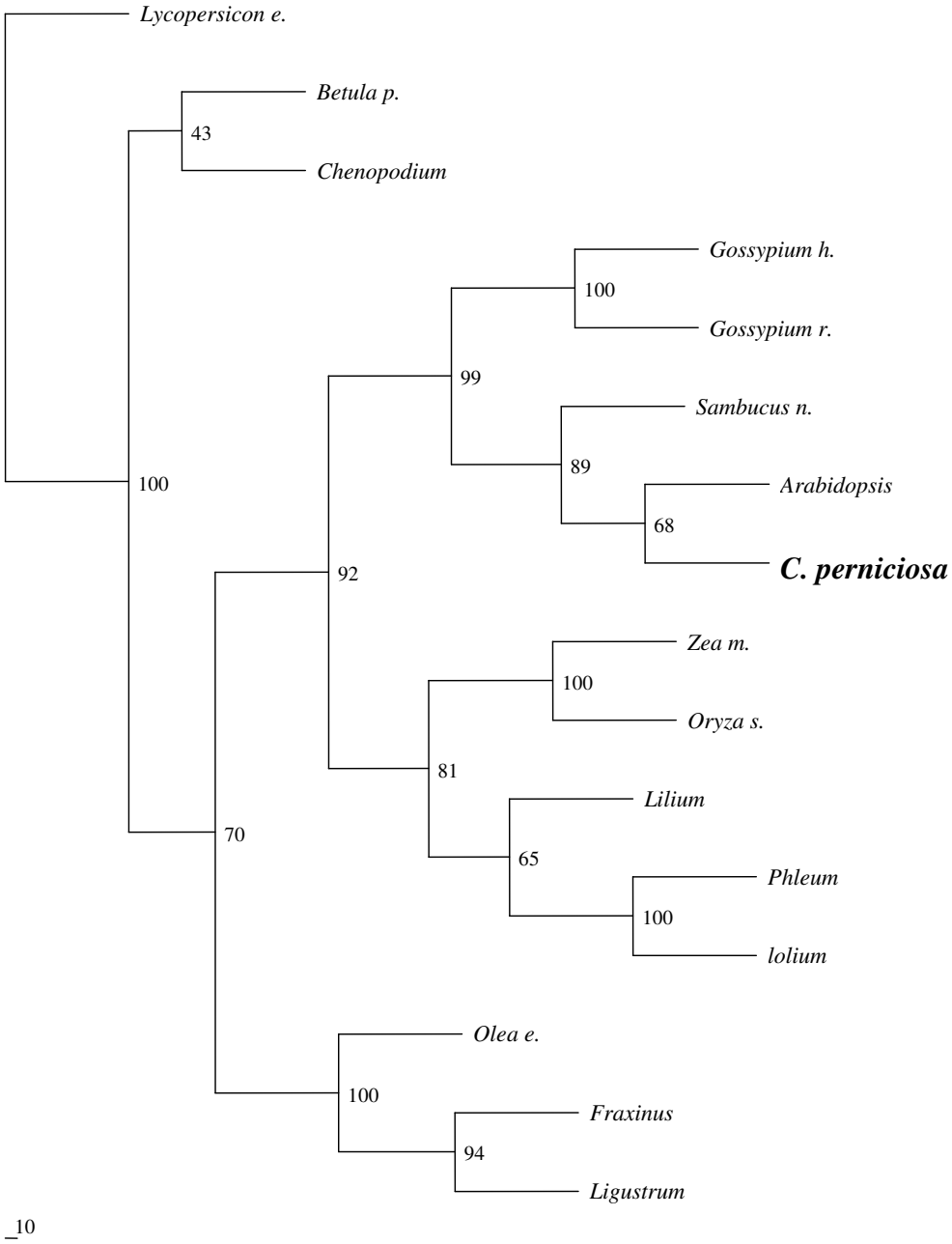


Figura 9. Árvore filogenética da família de proteínas de Pollen, um provável contaminante de eucalipto no genoma de *C. perniciosa* agrupa *com* com *Arabidopsis* sendo suportada por um valor robusto de bootstrap (68%). Árvore foi construída com o método "neighbour-joining" (N-J). Os programas utilizados para tal foram NEIGHBOR e PROTDIST do PHYLIP versão 3.6. A opção de programa "Dayhoff" foi utilizada para computar a matriz de distância. Os programas SEQBOOT e CONSENSE foram usados para estabelecer limites de confiança dos pontos de ramificação de 1000 replicatas de bootstrap. As barras do lado esquerdo embaixo da filogenia correspondem a 10 substituições de aminoácido por sítio ao longo dos ramos.

Na árvore filogenética construída na figura 10 para a família de proteínas DAD ("defender against cell death") podemos observar a ramificação da seqüência de *C. perniciosa* misturada com seqüências de plantas caracterizando uma árvore não monofilética. A sequência de DAD encontrada em *Crinipellis perniciosa* representa um possível contaminante do banco de dados do eucalipto e aparece agrupando especificamente com *Citrus u.,* apresentando um baixo valor de bootstrap suportando esta ramificação (40%).

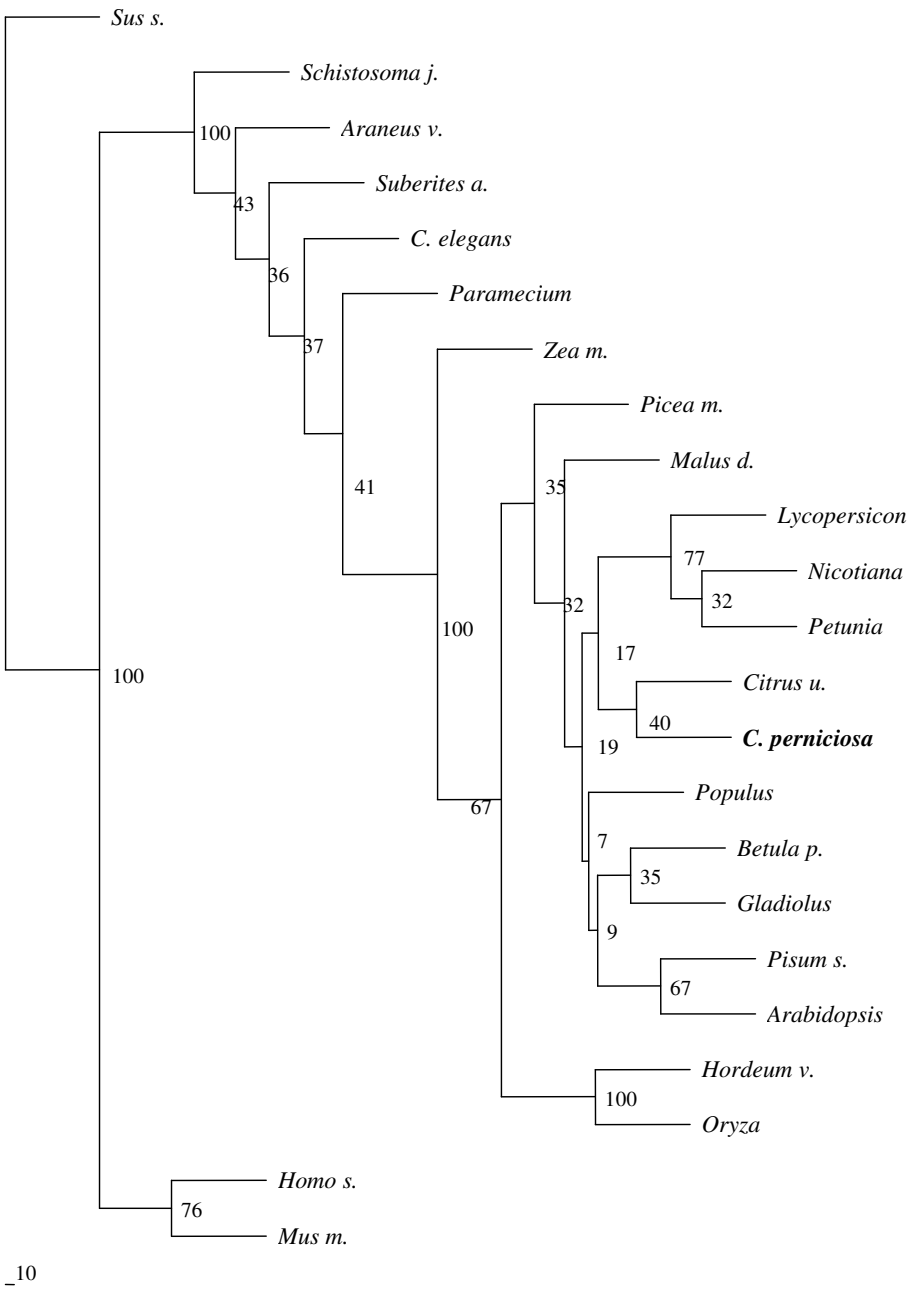


Figura 10. Árvore filogenética da proteína DAD. Método usado para reconstrução da árvore filogenética foi o Neighbor-Joining usando as mesmas opções selecionadas na figura 9. A seqüência de DAD em *C. perniciosa* é um provável

contaminante de eucalipto e aparece ramificando com seqüências de plantas. Pode-se observar um baixo valor de bootstrap suportando a ramificação de *C. perniciosa* e *Citrus u.* de 40% portanto a filogenia é inconclusiva para definir um evento de HGT.

Outro candidato a HGT a ter sua filogenia inferida foi a proteína "Thaumatin", tipo PR ("Patogênese related"). Podemos observar que *C. perniciosa* aparece como "outgroup" de um clado de proteínas de THN em plantas, mantendo a monofilia da árvore.

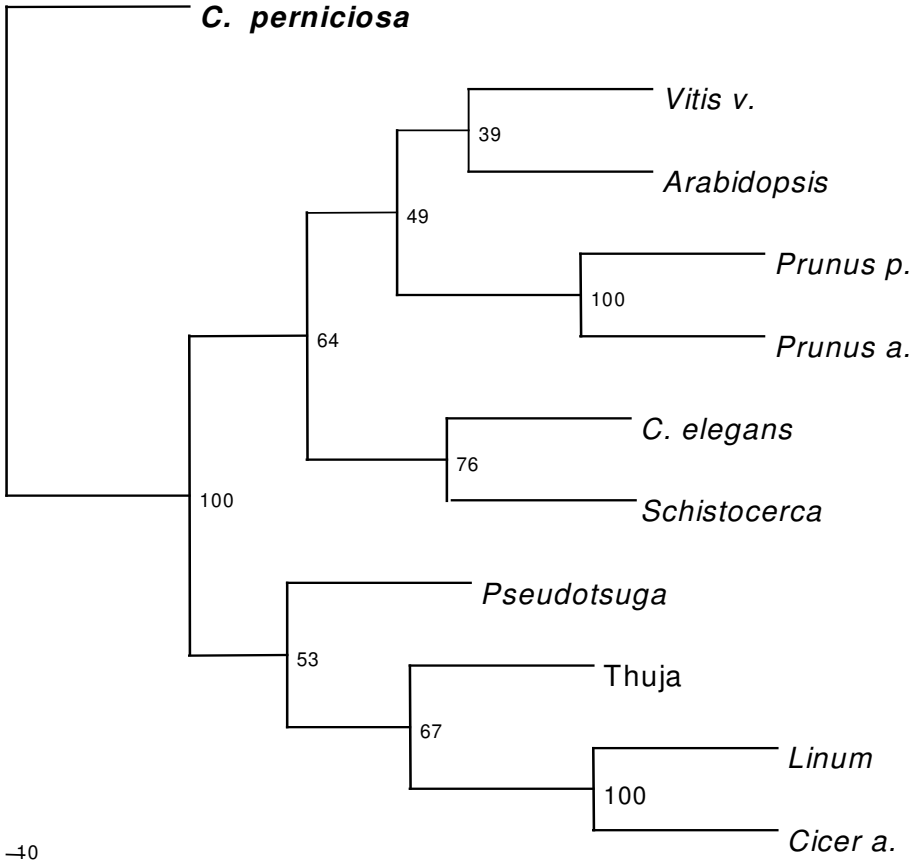


Figura 11. Reconstrução da Árvore filogenética para a proteína THN. Novamente o método usado foi o Neighbor Joining e as opções escolhidas as mesmas que na figura 9. Podemos observar *C. perniciosa* ramificando como "outgroup" do clado monofilético de plantas portanto não apresentando uma ramificação incongruente com sequências de plantas o que caracterizaria um evento de HGT.

Na árvore filogenética para a proteína Transposase (figura 12), podemos observar o agrupamento de *C. perniciosa* com seqüências hipotéticas do fungo basidiomiceto *cryptococcus* sendo suportado por valores robustos de bootstrap (91%).

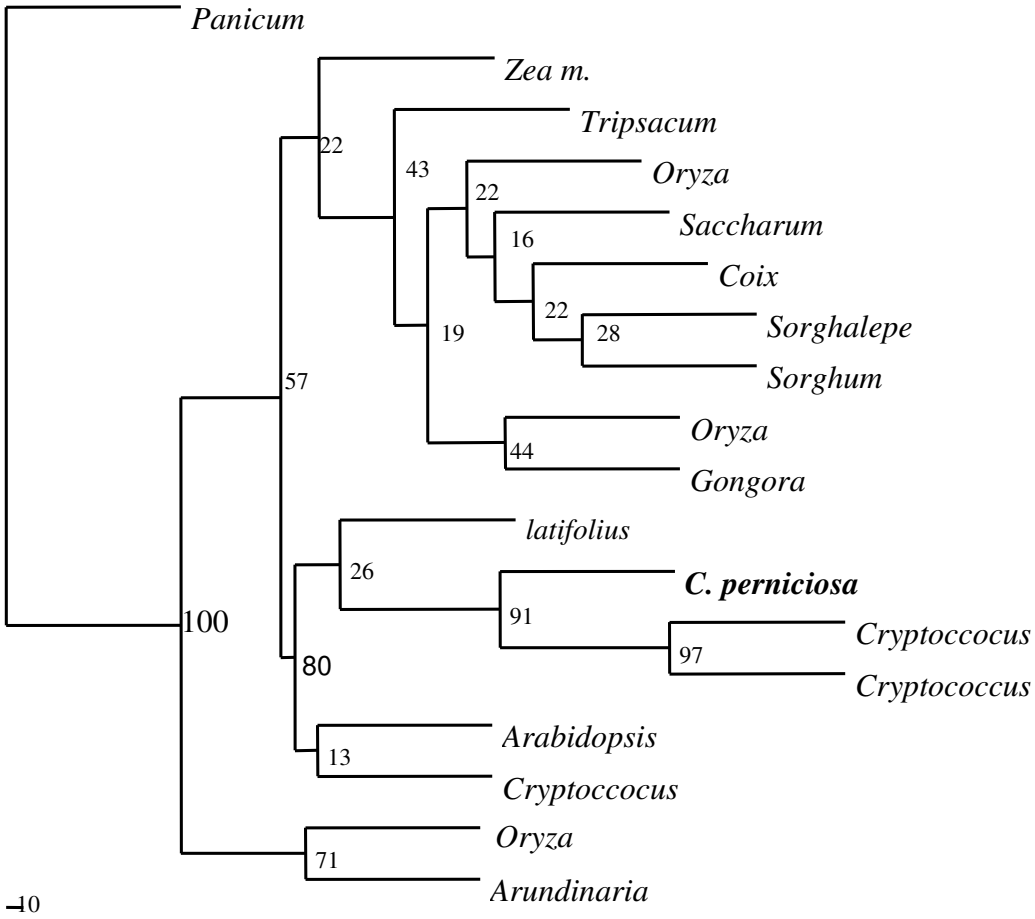


Figura 12. Arvore filogenética da proteína transposase. Método usado foi o Neighbour Joining de acordo com as especificações da figura 9. *C. perniciosa* ramifica com duas sequências hipotéticas de Cryptoccocus com valores robustos de bootstrap suportando esta ramificação (91%).

Podemos observar a filogenia da proteína Integrase (rve) na figura 13 ramificando *C. perniciosa* com a alga verde *Chlorophyta* (77%). *Volvox c.* e *Crinipellis peniciosa* aparecem ramificando com um clado irmão formado de duas proteínas do tipo *gag-pol* indicando uma homologia entre estas seqüências.

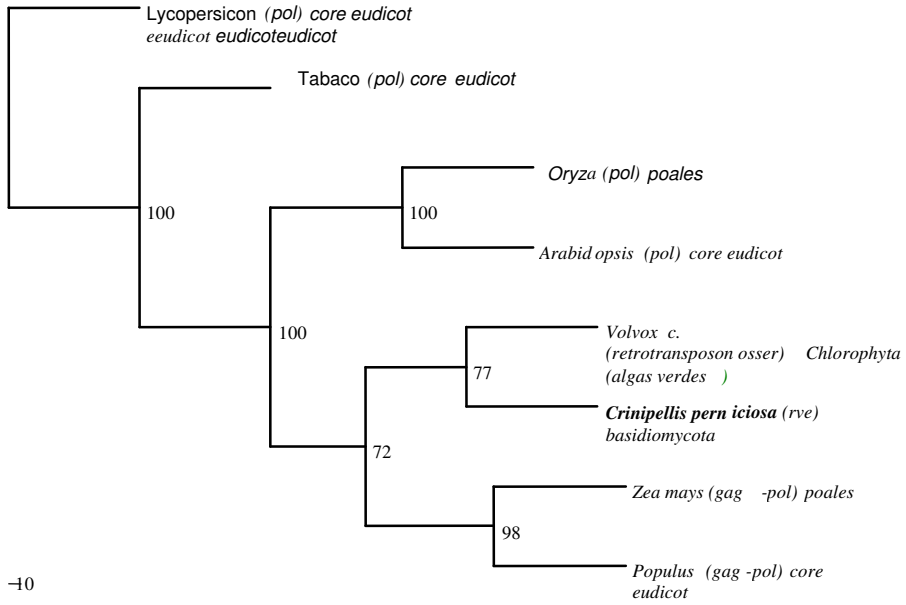


Figura 13. Árvore filogenética da proteína Integrase ramifica *C. perniciosa* com *Volvox c.* e portanto não é monofilética, indicando HGT. Esta árvore de proteína foi construída usando o método Neighbour Joining, baseado nas estimativas das distâncias par-a-par do numero esperado de subsituições de aminoácidos por sítio (10 na barra de escala).

A árvore filogenética da proteína "Heme Oxygenase" pode ser observada na figura 14. Enquanto as espécies de eucariotos se mantiveram monofiléticas,no topo da árvore, de acordo com a árvore universal, *C. perniciosa* aparece agrupando em um clado de bactérias, suportada por um valor de bootstrap de 46%, indicando que os dois grupos são polifiléticos, indicando HGT.

A filogenia da proteína JMJC pode ser visualizada na figura 15. A filogenia consiste de dois clados irmãos eucariotos; um monofilético com seqüências de proteínas JMJC presentes em animais enquanto o outro clado é polifilético com seqüências de plantas e *C. perniciosa*, com altos valores de bootstrap suportando esta ramificação (100%).

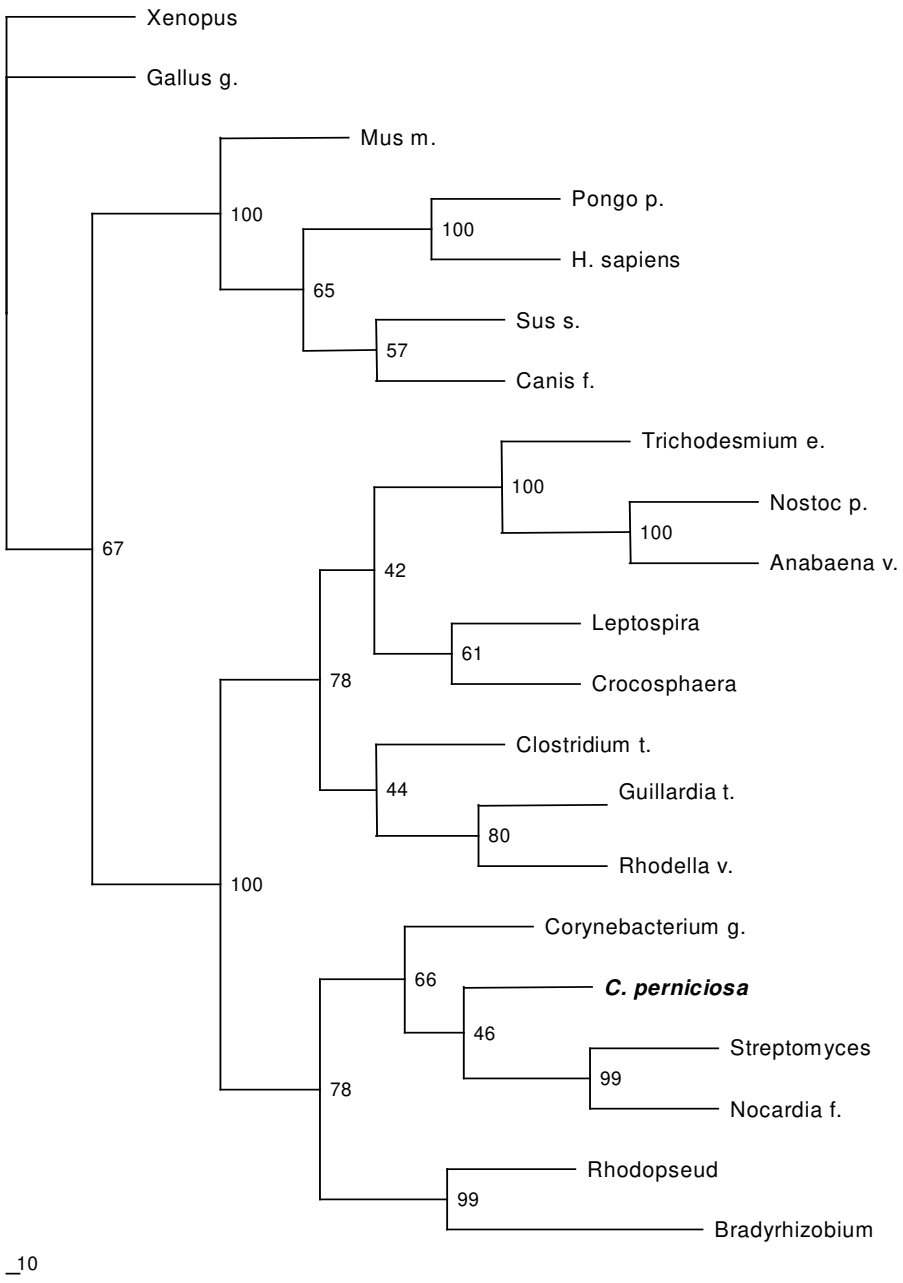


Figura 14. Árvore filogenética da proteína heme oxigenase indica HGT com o agrupamento de C. perniciosa em um clado polifilético com seqüências de bactéria com um valor de 46% de bootstrap suportando esta ramificação. Método empregado para reconstrução da árvore filogenética foi o Neighbour Joining com as mesmas opções descritas na figura 9.

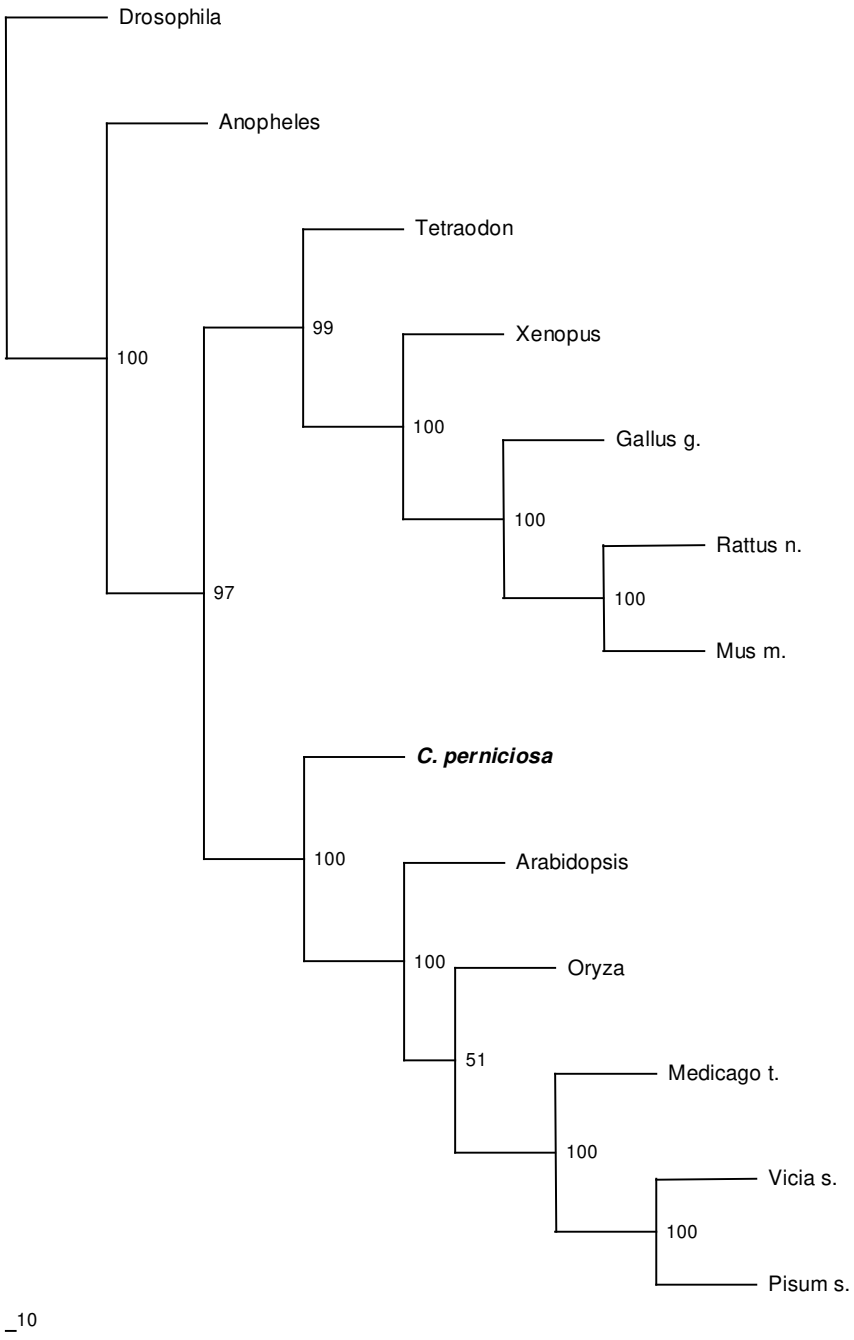


Figura 15. Árvore filogenética da proteína JMJC apresenta uma filogenia incongruente com a árvore universal pois agrupa C. perniciosa em um grupo polifilético com plantas sendo suportado por um valor robusto de bootstrap (100%).

Análise de "Códon Usage"

A distribuição dos códons de 57 genes (27 candidatos a HGT, sendo 05 transposons e 25 genes de EST similares a fungos) foram calculados a partir de freqüências de genes individuais computados pelo programa codonW. Um programa de estatística (SPADN) foi usado para interpretar os dados gerados graficamente, mostrando a variância destas freqüências de codons para cada gene ao longo de três eixos (ou classes) através da análise bivariada. Genes foram agrupados em três classes de acordo com a variância de suas freqüências (ver abaixo os genes pertencentes a cada uma das três classes). Genes estão representados através de notação, por exemplo; genes de EST possuem a letra "e" na frente; genes candidatos a HGT possuem um "h" na sua frente e os transposons um "t" na frente.

CLASSE 1 / 3 (verde no gráfico)

(e1 e10 e11 e12 e13 e15 e16 e17 e18 e19 e2 e20 e21 e22 e23
e24 e25 e3 e4 e5 e6 e7 e8 e9 hcog hduf hfa1 hidr hjm hkog
hp53 hps2 hTHN hudp hzn Tra tra2 traK trve tr17)

CLASSE 2 / 3 (em vermelho no gráfico)

(e14 hdad hduf heme hk17 hk42 hmit hmtl hPo hsfs hypk)

CLASSE 3 / 3 (amarela)

(h122 h661 hfkb Hlyd hpep hsap)

Podemos observar na figura 12 que os genes pertencentes a classe dois e três demonstram maior variância e agruparam apenas genes candidatos a HGT à exceção de um gene de EST (e14). Já os genes de EST foram todos agrupados na classe 1 mostrando uma similaridade em suas freqüências e inclusive os transposons. Portanto os genes pertencentes aa classes dois e três possuem um "códon usage" diferentes dos da classe 1, basicamente compostos por genes de EST. Isto confirmaria, a princípio, a teoria da Hipótese genômica na qual os genes de um organismo tendem a ter composições homogêneas em seu "códon usage", e aqueles que diferem em freqüência para cada codon do total do genoma são provavelmente genes "álien", ou seja,.são provenientes de outros organismos e portanto transferidos horizontalmente.

É importante ressaltar que estes dados confirmam HGT para dois dos genes amplificados através de PCR; DAD e Pólen, que podem ser visualizados em vermelho no gráfico e compõe a classe 3 (hpo e hdad). Outros genes candidatos também demonstram uma freqüência distinta de códons como por exemplo o gene da heme oxidase (heme) e o gene HYPK, também em vermelho e apresentando uma variância de códons em relação aos genes da classe 1 (verde), majoritáriamente compostos por genes de EST. Os genes em amarelo (classe 2) também contribuem significantemente para a variância destes dados.

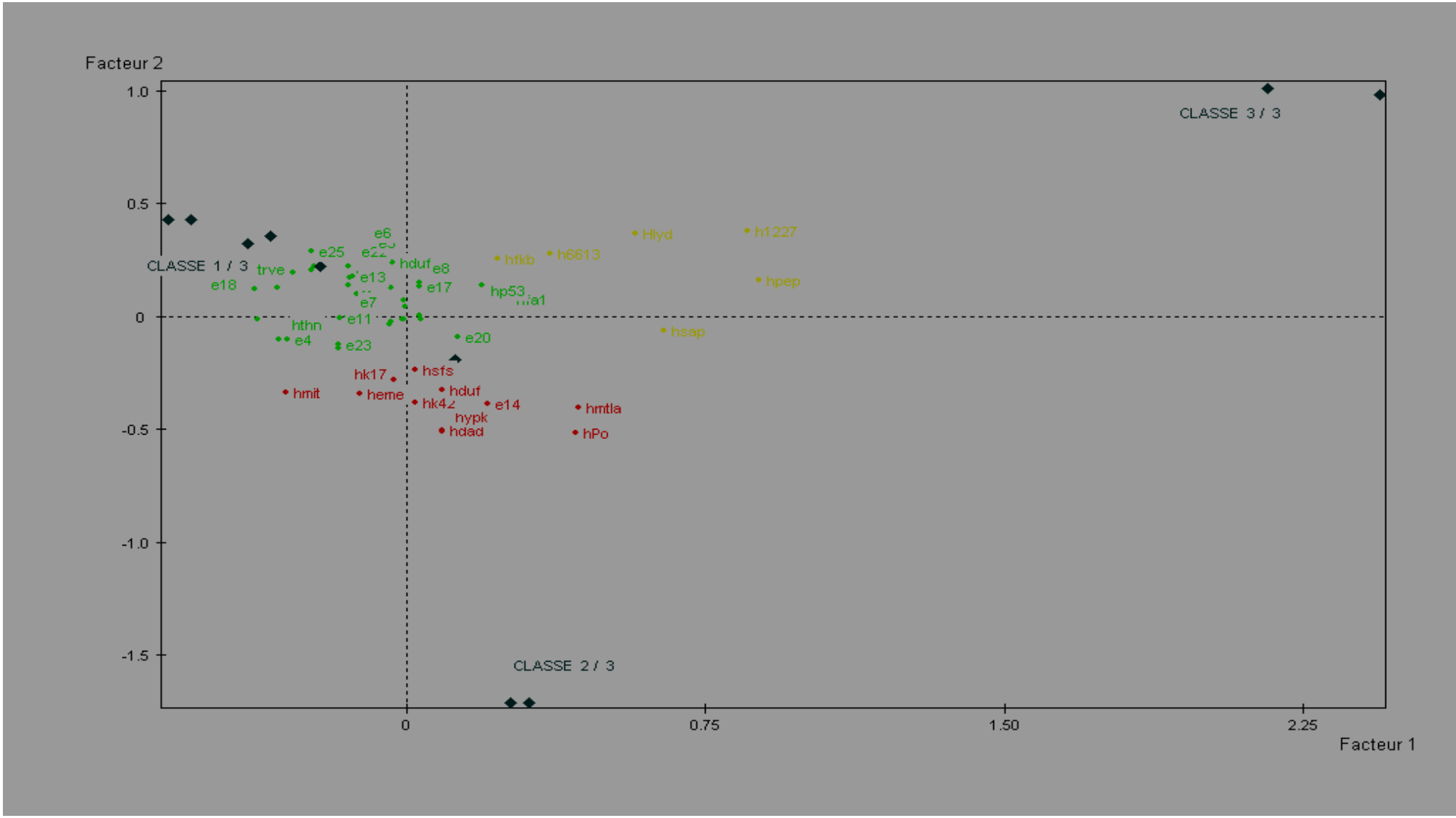


Figura 16. Análise bivariada mostrando a variância da freqüência dos códons de 57 genes em *Crinipellis perniciosa*. Genes foram agrupados em 3 classes de acordo com a freqüência do “códon usage” para cada gene..

Análise de conteúdo GC dos 43 genes candidatos

O conteúdo G+C do genoma de *C.perniciosa* foi estimado em 46.1% através da utilização de um script PERL na montagem-rascunho do genoma obtida com o programa Pherd/Phrap e comparada a média individual de genes candidatos a HGT e genes convencionais obtidos da biblioteca de EST. A média G+C destes genes candidatos variou entre 43% a 67%. Portanto alguns genes possuem valores diferentes da média de *C.perniciosa* como é o caso do gene putativo de pólen que possui um valor de 60% (figura 13). É interessante notar que os valores relativos a composição média G+C em diversos genomas tendem a possuir valores iguais tanto para as montagens "draft" como para regiões codantes.

Para chegar a tal conclusão nós calculamos a média G+C em *Aspergillus nidulans* com 50,4% de média G+C em sua montagem "Draft" e 53% de média apenas para genes codantes e *Ustilago maydis* com valores de 54,2% para o "draft" e 56,3% para regiões codantes apenas. Portanto o resultado preliminar de genes candidatos a HGT cujos valores de conteúdo G+C contrastam com a média de 46.1% em *C. perniciosa* deve ser consistente. Apesar de não termos ainda o genoma completamente finalizado podemos dizer, com base na análise acima, que a média G+C das regiões codantes em *C. perniciosa* deve ser similar ao resultado da média do "draft", levando-se em consideração o fato de que regiões codantes geralmente possuem uma maior percentagem de bases G+C do que regiões não codantes.

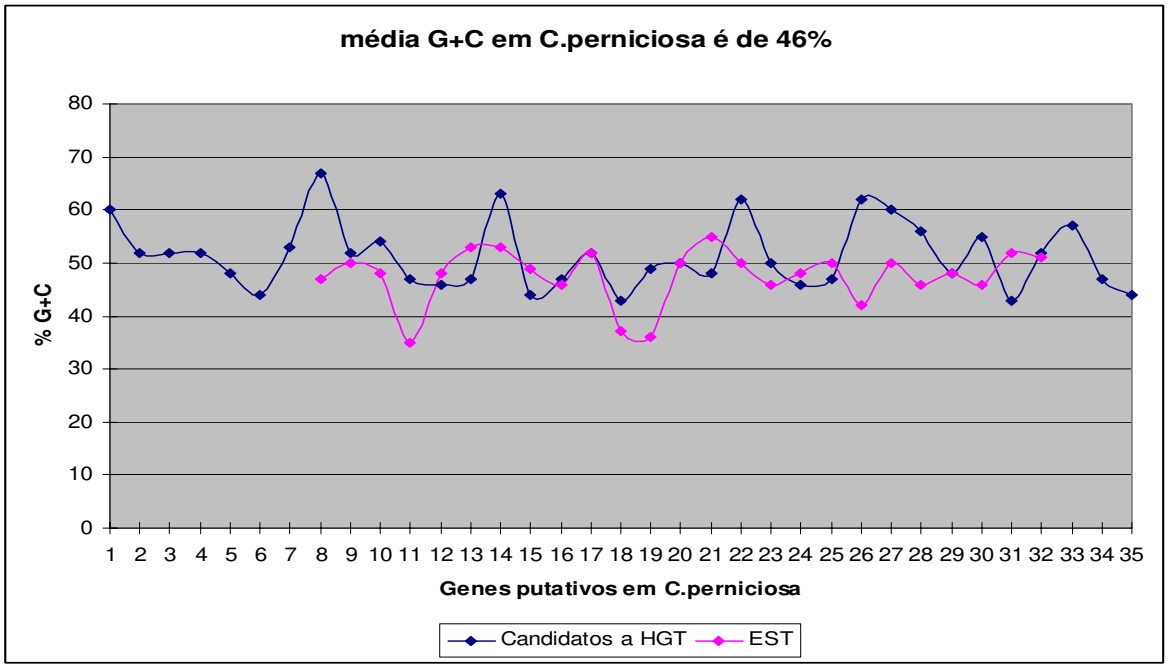


Figura 17. Histograma comparativo mostrando a média de conteúdo G+C de 35 genes candidatos a HGT (em azul) com 25 ESTs adquiridos através de transferência vertical (linha rosa).

VII. Discussão

O que é a transferência horizontal.

O fato de HGT incorporar tão bem a idéia da migração de informação genética entre organismos parece oferecer uma simples resposta a diversas questões biológicas atuais. Ao invés de ser criar conflito com noções úteis de árvores filogenéticas e de especiação, HGT pode, em realidade vir a agregar valor a teoria da evolução e problemas filogenéticos até então não resolvidos como no caso de aparentes conflitos da filogenia de angiospermas. Foi sugerido neste caso que HGT pode ter sido tão freqüente que anularia critérios baseados no conhecimento da filogenia da espécie como isolamento reprodutivo (Syvanen 1994). Foi demonstrado que a árvore filogenética para o gene citocromo c em angiospermas é internamente homoplásica se comparado com a filogenia deste mesmo gene em vertebrados e que esta seja resultado de HGT (Syvanen 1994).Como vimos anteriormente, homoplasia é o termo dado a dois ou mais caracteres similares mas que evoluíram independentemente. Foi sugerido que HGT também seja um fator causador de homoplasia entre seqüências (Syvanen 1994).

Estudos com o gene de citocromo c em plantas demonstram que o gene é de fato bem conservado e que é possível recuperar filogenias corretas com ele como no caso do citocromo c em vertebrados (Syvanen 1994). Ambos apresentam quantidade de divergência entre as seqüências uniforme e taxas de substituição comparáveis ao longo do tempo (relógio molecular), portanto erros artefatuais, como o fenômeno do “long branch attract” não poderiam explicar a filogenia incongruente em plantas. Entretanto enquanto que a filogenia do gene de citocromo c em vertebrados é razoavelmente congruente com a filogenia aceita, a filogenia do citocromo c em plantas gera quatro grupos de árvores através do método da evolução mínima completamente incongruentes com a árvore canônica universal baseada em seqüências de SSU rRNA. Segundo Syvanen isto seria o equivalente a dizer em filogenética que mais homoplasia ocorre na evolução do citocromo c em plantas do que na evolução dele em vertebrados de modo que caso a homoplasia interna do gene de citocromo c seja devido a HGT poderíamos explicar estas árvores incongruentes.

HGT poderia, portanto resolver uma crise clássica da teoria neo-Darwiniana da sistemática de angiospermas pois se HGT for muito freqüente neste caso não caberia a aplicação de critérios que dependam do conhecimento de filogenia de espécies.Tal crise poderia ser resolvida através da incorporação de HGT na teoria da evolução. O exemplo descrito acima descreve a importância de considerarmos HGT como um fenômeno evolutivo de poder explicativo.

Outra questão importante que diz respeito a HGT são os possíveis mecanismos para penetração de DNA entre barreiras de espécies. Um trabalho que ilustra bem esse mecanismo foi realizado com *T. cruzi*. Foi observado que o DNA do cinetoplasto é integrado em células do músculo cardíaco no coração dos pacientes infectados cronicamente, levando a fusão de genes de forma inadequada o que em alguns casos leva a formação de proteínas quiméricas antigênicas, que acabam sendo combatidas pelo sistema imune, levando a lesões histopatológicas do órgão, caracterizando, portanto a natureza autoimune da doença. Para comprovar esse processo, células tronco embrionário de aves e coelhos foram infectadas

com o parasita *T. cruzi*, levando a um organismo com DNA de cinetoplastos integrados em praticamente todas as células, inclusive as germinativas. As progênies desses organismos apresentaram então cinetoplastos integrados e, como conseqüência, lesões histopatológicas relacionadas a autoimunidade provocada pelas integrações. Assim sendo, esse é um exemplo de um intercâmbio de genes entre espécies diferentes através da interação parasita-hospedeiro, com a integração de DNA intracelularmente nos hospedeiros que foi comprovado experimentalmente em coelhos e aves.

Este é um exemplo da possibilidade de genes transferidos horizontalmente serem integrados em um dado genoma, tendo o DNA introduzido se adaptado completamente ao genoma do hospedeiro, inclusive tendo sido transferido verticalmente durante processos de especiação.

Discussão dos resultados

Com base nas sete filogenias de proteínas apresentadas aqui, somente uma (THN), apresenta uma filogenia congruente e condizente com a árvore canônica universal baseada em SSU rRNA. As outras seis filogenias apresentam uma topologia incongruente e, *a priori,* indicariam HGT; ora ramificando seqüências de fungos em clados de bactérias, ou ramificando seqüências de fungos em clados de plantas; em ambos os casos formando grupos polifiléticos.

Esta polifilia dos clados não condiz com a monofilia dos principais reinos (Bacteria, eucaria e archae) ou dos domínios da vida(Plantas, animais, Fungos, Bactérias) presentes na árvore canônica universal. Hoje em dia sabemos que estas divergências podem ser o resultado de transferência horizontal de genes ou menos provavelmente de convergência gênica.

Outro indicador de HGT tanto para o gene de Pólen quanto para o gene DAD em *C. perniciosa* é o fato destes não apresentarem genes ortólogos em nenhum fungo seqüenciado até hoje. Entretanto estes genes apresentaram similaridade de 100% através de BLASTN com o seqüências do banco de dados de eucalipto, indicando uma contaminação.

A árvore filogenética do gene de Pólen o agrupa com valores razoáveis de bootstrap com seqüências de plantas indicando, *a priori*, HGT. Alternativamente, este agrupamento de Pollen com seqüências provenientes de plantas explicaria uma possível contaminação genômica uma vez que a seqüência contaminante pertence a outra planta; o eucalipto. Este gene mostrou tambem um “códon usage” que apresenta uma variância em relação ao “códon usage” médio de *C. perniciosa* (representado através de genes transferidos verticalmente - ESTs) e possui um conteúdo G+C bem acima da média da montagem de *C. perniciosa* (60% do Pollen contra 46% de *C. perniciosa*).

Com esses resultados o método empregado neste trabalho indicou com sucesso uma origem “alien” para este gene no genoma de *C. perniciosa*, sendo ele transferido horizontalmente ou apenas uma contaminação genômica, de acordo com os resultados descritos acima.

Existe uma vasta literatura de potenciais casos de HGT em fungos, incluindo genes nucleares, e elementos genéticos como Transposons, plasmídeos e íntrons. A maioria dos casos envolvendo HGT de genes nucleares em fungos como, por exemplo, a família de genes das Gliycosyl hydrolases como a endoglucanase B e β-glucanase foram inferidas através da presença/ausência de genes e similaridade

ao invés de filogenias e trabalhos posteriores serão necessários para confirmar estas transferências (Rosewich and Kistler 2000).

Filogenias incongruentes para genes de endoglucanases e xylanases demonstram que estes genes sejam mais homólogos a seqüências bacterianas do que seqüências de fungos pois aparecem ramificando no ramo das bactérias a exclusão do ramo dos fungos (Garcia-Vallve 2000).

Os exemplos acima são apenas alguns dos já descritos sobre HGT em fungos que vem reforçar a idéia que fungos são organismos extremamente suscetíveis a HGT devido a seu modo de vida de íntima de associação com outros organismos. Especialmente no caso de fungos com modo saprofítico de alimentação como é o caso dos himenomicetos de um modo geral e *C. perniciosa* especificamente.

Os 43 genes candidatos a HGT em *C. perniciosa* identificados como sendo horizontalmente transferidos apresentaram alta similaridade com seqüências de organismos distantes evolutivamente, principalmente plantas e bactéria e nenhuma similaridade com seqüência de fungos, salvo algumas exceções e nunca como o primeiro "hit" de BLASTP.

Outro gene importante descrito caracterizado em *C. perniciosa* pela primeira vez é o gene THN ("Thaumatin"), descrito como sendo uma proteína antifungo e só descrita em plantas até o momento.

Esta proteína só havia sido descrita em plantas mas foi recentemente identificada em alguns insetos provavelmente com alguma função de proteção a patogênese por fungos já que THN vem sendo descrita como uma proteína antifungo. Isto pode indicar que esta proteína é mais bem conservada do que se pensava previamente sendo encontrada em diversos taxa.

Elementos móveis

Uma parte significante do DNA em eucariotos é derivada de retrotransposons, elementos móveis que utilizam transcrição reversa para se replicarem. No DNA genômico da seqüência "draft" do projeto genoma humano foram identificados 29% como sendo retrotransposons ou reminiscências de retrovírus . Normalmente retrotransposons vêm sendo associados com HGT. Mas como esta retrotransposição poderia ocorrer? Existem alguns mecanismos plausíveis atualmente. Um deles seria a possibilidade de "pegar carona" em algum tipo de elemento móvel, como vírus por exemplo (Burke, Malik et al. 1998). Caso o retrotransposon seja integrado ao genoma de um vírus ele poderia, por exemplo, ser transmitido entre as células através de infecção viral. Caso a célula sobreviva a infecção o elemento poderia ser transcrito do genoma viral e, através de retrotransposição se integrar ao cromossomo do hospedeiro. Diversos vírus infectam múltiplos hospedeiros, providenciando um vetor extremamente favorável a HGT entre taxa. Existem vários exemplos de retrovírus integrados ao genoma de elementos móveis: O baculovírus em insetos é um exemplo bem documentado de vírus que transportaria transposons. Existem vários exemplos bem documentados de retrotransposons bem integrados ao genoma de vírus (Bushman 2002), que possibilitariam sua movimentação entre taxa. Existe ainda uma terceira classe de transposons, os introns móveis, que não são abordados neste trabalho.

Com relação aos elementos móveis encontrados pela nossa metodologia de identificação de candidatos transferidos horizontalmente em eucariotos, de fato, eles não deveriam constituir uma surpresa uma vez que suspeitas de casos de HGT normalmente envolvem "elementos egoístas", não

infectivos (transposons), contrastando com genes adquiridos possivelmente de vírus como proteínas de prófagos e retrotovírus.

Distinções entre retrotransposons do tipo LTR e retrovírus pode ser uma tarefa difícil pois estes possuem estruturas similares como é o caso das proteínas *gag* e *pol.* Duas proteínas candidatas apresentaram similaridade tanto a seqüências de retrovírus como seqüências de retrotransposons exatamente devido a esta similaridade genética. Uma das maneiras de distinguirmos entre estes dois elementos é por meio de buscas por similaridade contra diversos bancos de dados, incluindo o banco nr e bancos de ESTs. As duas proteínas identificadas neste trabalho foram comparadas contra ambos os bancos, sendo os "hits" mais similares encontrados em ambos com retrotransposons LTR do tipo cópia.

VIII. Conclusão e perspectivas

O método empregado neste trabalho identificou 43 genes (proteínas) provavelmente transferidos horizontalmente entre *C. perniciosa* e três taxas: Plantas, Bactérias e Animais, com base na alta similaridade entre estas seqüências através de altos valores de e-value (BLASTP). Estes 43 genes foram preliminarmente anotados e tiveram sua função inferida com base em seu maior "hit" de BLASTP contra o ncbi-nr ou contra o banco de ESTs de fungos patogênicos COGEME.

Amplificação por PCR com primers específicos a partir de DNA genômico da fase necrotrófica de *C. perniciosa* de quatro candidatos (Pólen, THN, DAD e Transposase) amplificou bandas com tamanhos similares aos mesmos genes preditos pela bioinformática.

Um segundo experimento consistindo na amplificação dos mesmos quatro genes citados acima, desta vez a partir de cDNA confirmou a amplificação destes genes a partir de mRNAs dupla fita (cDNA) de modo que estes genes possivelmente codificam proteínas. Este experimento amplificou ainda bandas do tamanho predito conforme a seqüência codante (CDS) preditas pela bioinformática.

Os resultados de amplificação para os genes de Pollen e DAD não apresentaram um padrão de bandas consistente em todas as amplificações e outros experimentos serão necessários de modo a comprovar a presença destes dois genes no genoma de *C. perniciosa*.

Dois dos quatro genes amplificados através de PCR foram seqüenciados (THN e Transposase) e um BLAST2seq confirmou similaridade do sequenciamento com as seqüências destes dois genes preditas pela bioinformática.

Devido ao fato dos genes de Pollen e DAD não terem sido sequenciados e os resultados de amplificação das bandas destes genes não ter sido consistente, foi feito um BLASTN do gene de Pollen contra o banco de dados EST_others de modo a procurar genes ortólogos. O resultado de BLASTN apresentou alinhamentos com a nota máxima de e-value (0) e similaridade de 100% com seqüências de EST de eucalipto. Com isso foi feito um novo BLASTN com a seqüência do gene de Pollen contra o banco de dados do eucalipto apresentando alinhamentos com a nota máxima de e-value (0) e similaridade de 100% com seqüências nucleotídicas de eucalipto confirmando a suspeita de contaminação.

A mesma metodologia foi aplicada para os 43 genes candidatos a HGT e cinco novas contaminações provenientes de eucalipto foram identificadas além de Pollen (DAD, HYPK, 2 DUFs e uma proteína mitocondrial), apresentaram alinhamentos com a nota máxima de e-value (0) e similaridade de 100%.

Análises adicionais identificaram três placas contaminantes provenientes do sequenciamento do genoma do eucalipto, representando 276 reads no banco de dados do genoma da vassoura de bruxa que foram identificadas como prováveis contaminantes e tiveram seus respectivos reads removidos de modo a evitar futuras contaminações.

Com esses resultados o método empregado neste trabalho indicou com sucesso uma origem "alien" para estes genes no genoma de *C. perniciosa*, sendo eles transferidos horizontalmente ou apenas uma contaminação genômica do eucalipto, através de métodos como conteúdo G + C, "Códon Usage" e inferência filogenética.

Somente uma das sete árvores filogenéticas inferidas para os genes candidatos a HGT, a árvore da proteína "thaumatin" (THN) da família PR-5 ou ("pathogen related proteins") apresentou uma topologia congruente com a árvore canônica universal baseada em seqüências de SSU rRNA, não indicando portanto HGT. Todas as outras filogenias inferidas apresentaram árvores incongruentes com a árvore universal indicando HGT. Uma destas árvores incongruentes foi a da proteína de Pollen que agrupou a seqüência de *C. perniciosa* em um clado contendo somente seqüências de proteínas de Pollen em plantas, com um robusto valor de bootstrap suportando esta ramificação (68%). Esta mistura de seqüências de organismos filogenéticamente distantes, neste caso de planta e fungos, caracterizando um grupo polifilético onde deveríamos observar clados monofiléticos é um indicador de HGT.

Posteriormente verificou-se que esta ramificação de Pollen com homólogos de plantas poderia ser devido a uma contaminação genômica do banco de dados do eucalipto e não um evento de HGT, caracterizado pela ramificação de proteínas de grupos taxonômicos distintos (ex. fungos e plantas).

A filogenia da família de proteínas da Heme Oxygenase apresentou um agrupamento polifilético entre seqüências de bactéria e a seqüência de *C. perniciosa.* Este agrupamento, no entanto foi suportado por um baixo valor de bootstrap (46%).

Como as notas de BLAST não são um indicador confiável de relações evolutivas (Brown 2003), o posicionamento anômalo de uma espécie em particular, ou várias espécies em uma árvore filogenética continua sendo o melhor indicador de eventos de HGT ocorridos há muito tempo em uma escala evolutiva.

Duas classes gerais de transposons previamente relacionados a HGT foram identificados nesta lista de candidatos a HGT em *C. perniciosa*: transposons tipo "PIF" e retrotransposons tipo cópia. O gene putativo da Transposase, que foi inclusive seqüenciado, demonstrou similaridade a família de transposons do tipo "P instability factor" (PIF), amplamente distribuídos e abundantes nos genomas de eucariotos (plantas, fungos e animais), além do gene Integrase, similar ao retrotransposon LTR do tipo cópia.

IX. Referências Bibliográficas.

Altschul, S. F., M. S. Boguski, et al. (1994). "Issues in searching molecular sequence databases." Nat Genet **6**(2): 119-29.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Ausubel, F. M., R. Brent, et al. (1999). "Current protocols in molecular biology." **3**: 19.3.1-19.3.29.

Baldauf, S. L. and J. D. Palmer (1990). "Evolutionary transfer of the chloroplast tufA gene to the nucleus." Nature **344**(6263): 262-5.

Bergthorsson, U., K. L. Adams, et al. (2003). "Widespread horizontal transfer of mitochondrial genes in flowering plants." Nature **424**(6945): 197-201.

Bowring, S. A., J. P. Grotzinger, et al. (1993). "Calibrating rates of early Cambrian evolution." Science **261**: 1293-8.

Brown, J. R. (2003). "Ancient horizontal gene transfer." Nat Rev Genet **4**(2): 121-32.

Burke, W. D., H. S. Malik, et al. (1998). "Are retrotransposons long-term hitchhikers?" Nature **392**(6672): 141-2.

Bushman, F. (2002). Lateral DNA Transfer: Mechanisms and consequences, CSHL PRESS.

Calle, H., C. H. Cook, et al. (1982). "*Histology of Witches' -Broom Caused in Cacao by Crinipellis perniciosa*." Cytology and Histology **72**(11): 1479-1481.

Campbell, A. M. (2000). "Lateral Gene Transfer in Prokaryotes." Theoretical Pop Bio **57**: 71-77.

Daboussi, M. J. (1997). "Fungal transposable elements and genome evolution." Genetica **100**(1-3): 253-60.

Doolittle, R. F., D. F. Feng, et al. (1990). "A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote." J Mol Evol **31**(5): 383-8.

Erwin, D. H. and J. W. Valentine (1984). ""Hopeful Monsters", transposons, and metazoan radiation." Procl. Natl. Acad. Sci. USA **81**: 5482-83.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-94.

Ewing, B., L. Hillier, et al. (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-85.

Felsenstein, J. (1978). "Cases in which parsimony or compatibillity methods will be positively misleading." Syst Zool **27**: 401-410.

Felsenstein, J. (1989). "PHYLIP - Phylogeny inference package (version 3.2)." Cladistics **5**: 164-166.

Flavell, A. J. (1992). "Ty1-copia group retrotransposons and the evolution of retroelements in the eukaryotes." Genetica **86**(1-3): 203-14.

Flavell, A. J. (1999). "Long terminal repeat retrotransposons jump between species." Proc Natl Acad Sci U S A **96**(22): 12211-2.

Galagan, J. (2004). Academy colloquium "fungal phylogenomics". Comparative fungal genomics, Netherlands.

Gamieldien, J., A. Ptitsyn, et al. (2002). "Eukaryotic genes in Mycobacterium tuberculosis could have a role in pathogenesis and immunomodulation." Trends Genet **18**(1): 5-8.

Gantt, J. S., S. L. Baldauf, et al. (1991). "Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron." Embo J **10**(10): 3073-8.

Garcia-Vallve, S., A. Romeu and J. Palau, (2000). "Horizontal gene transfer of glycosyl hydrolases of the rumen fungi." Mol Biol Evol **17**: 352-61.

Garcia-Vallvé, S., A. Romeu and J. Palau. (2000). "Horizontal gene transfer in bacterial and archaea complete genomes." Genome Res **10**: 1719-1725.

Gilbert, H. J., G. P. Hazlewood, et al. (1992). "Homologous catalytic domains in a rumen fungal xylanase: evidence for gene duplication and prokaryotic origin." Mol Microbiol **6**(15): 2065-72.

Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.

Grantham, R., C. Gautier, et al. (1980). "Codon catalog usage and the genome hypothesis." Nucleic Acids Res **8**(1): r49-r62.

Hamer, L., H. Pan, et al. (2001). "Regions of microsynteny in Magnaporthe grisea and Neurospora crassa." Fungal Genet Biol **33**(2): 137-43.

Han, Y., X. Liu, et al. (2001). "Genes determining pathogenicity to pea are clustered on a supernumerary chromosome in the fungal plant pathogen Nectria haematococca." Plant J **25**(3): 305-14.

Hartman, H., M. Syvanen, et al. (1990). "Contrasting evolutionary histories of chloroplast thioredoxins f and m." Mol Biol Evol **7**(3): 247-54.

He, C., A. Rusu, et al. (1988). "Transfer of supernumerary chromossomes between vegetatively incompatible biotypes of the fungus *Colletotrichum gloeosporioides.*" Genetics(150): 1459-1466.

Jacobson, J. W., M. M. Medhora, et al. (1986). "Molecular structure of a somatically unstable transposable element in Drosophila." Proc Natl Acad Sci U S A **83**(22): 8684-8.

Katz, L. A. (2002). "Lateral gene transfers and the evolution of eukaryotes: theories and data." Int J Syst Evol Microbiol **52**(Pt 5): 1893-900.

Kemmerer, E. C., M. Lei, et al. (1991). "Structure and molecular evolutionary analysis of a plant cytochrome c gene: surprising implications for Arabidopsis thaliana." J Mol Evol **32**(3): 227-37.

Koonin, E. V., N. D. Fedorova, et al. (2004). "A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes." Genome Biol **5**(2): R7.

Koonin, E. V., K. S. Makarova, et al. (2001). "Horizontal gene transfer in prokaryotes: quantification and classification." Annu Rev Microbiol **55**: 709-42.

Koonin, E. V., A. R. Mushegian, et al. (1997). "Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea." Mol Microbiol **25**(4): 619-37.

Kupfer, D. M., C. A. Reece, et al. (1997). "Multicellular ascomycetous fungal genomes contain more than 8000 genes." Fungal Genet Biol **21**(3): 364-72.

Leipe, D. D., J. H. Gunderson, et al. (1993). "Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree." Mol Biochem Parasitol **59**(1): 41-8.

Mereschkowsky, C. (1905). "Uber Natur Und Ursprung der Chromatophoren in Pflanzenteilein." Biol. Zentrabl. **25**(593-).

Moreira, D. and H. Philippe (2000). "Molecular phylogeny: pitfalls and progress." Int Microbiol **3**(1): 9-16.

Muller, M. (1998). "What are the Microsporidia?" Parasitol Today **13**: 455-456.

Mullis, K. B. (1990). "Target amplification for DNA analysis by the polymerase chain reaction." Ann Biol Clin (Paris) **48**(8): 579-82.

Nicholas, K. B., N. H. B. Jr., et al. (1997). "GeneDoc: Analysis and Visualization of Genetic Variation." EMBNEW.NEWS **4**(14).

Nitz, N., C. Gomes, et al. (2004). "Heritable integration of kDNA minicircle sequences from Trypanosoma cruzi into the avian genome: insights into human Chagas disease." Cell **118**(2): 175-86.

Ochman, H. and N. A. Moran (2001). "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis." Science **292**(5519): 1096-9.

Oliver, S. G. (1996). "From DNA sequence to biological function." Nature **379**(6566): 597-600.

Page, R. D. (1996). "TreeView: an application to display phylogenetic trees on personal computers." Comput Appl Biosci **12**(4): 357-8.

Page, R. D. M. and E. C. Holmes (1998). Molecular Evolution: A Phylogenetic Approach.

Peden, J. CodonW as a freeware release for codon usage analysis.

Pereira, J. L., Almeida LCC and SM Santos. (1996). "Witche's broom disease of cocoa in Bahia: attempts at eradication and containment." Crop Protection **15**: 743-752.

Philippe, H. and C. J. Douady (2003). "Horizontal gene transfer and phylogenetics." Curr Opin Microbiol **6**(5): 498-505.

Purdy, L. H. and R. A. Schmidt (1996). "STATUS OF CACAO WITCHES' BROOM: biology, epidemiology, and management." Annu Rev Phytopathol **34**: 573-94.

Raff, R. A., C. R. MArshall, et al. (1994). "Using DNA sequences to unravel the Cambrian radiation of the animal phyla." Annu. Rev. Ecol. Syst **25**: 351-375.

Rincones, J., L. W. Meinhardt, et al. (2003). "Electrophoretic karyotype analysis of Crinipellis perniciosa, the causal agent of witches' broom disease of Theobroma cacao." Mycol Res **107**(Pt 4): 452-8.

Robertson, H. M., K. L. Zumpano, et al. (1996). "Reconstructing the ancient mariners of humans." Nat Genet **12**(4): 360-1.

Rosewich, U. L. and H. C. Kistler (2000). "Role of Horizontal Gene Transfer in the Evolution of Fungi." Annu Rev Phytopathol **38**: 325-363.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.

Scholl, E. H., J. L. Thorne, et al. (2003). "Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach." Genome Biol **4**(6): R39.

Shivji, M. S., N. Li, et al. (1992). "Structure and organization of rhodophyte and chromophyte plastid genomes: implications for the ancestry of plastids." Mol Gen Genet **232**(1): 65-73.

Skinner, W., J. Keon, et al. (2001). "Gene information for fungal plant pathogens from expressed sequences." Curr Opin Microbiol **4**(4): 381-6.

Smith, M. W., D. F. Feng, et al. (1992). "Evolution by acquisition: the case for horizontal gene transfers." Trends Biochem Sci **17**(12): 489-93.

Soanes, D. M., W. Skinner, et al. (2002). "Genomics of phytopathogenic fungi and the development of bioinformatic resources." Mol Plant Microbe Interact **15**(5): 421-7.

Sogin, M. L. (1991). "Early evolution and the origin of eukaryotes." Curr Opin Genet Dev **1**(4): 457-63.

Stahel, G. (1915). "Marasmius perniciosus, the cause of the krulloten disease of cacao in suriname." AMW ter laag **25**: 25.

Stanhope, M. J., A. Lupas, et al. (2001). "Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates." Nature **411**(6840): 940-4.

Syvanen, M. (1994). "Horizontal gene transfer: evidence and possible consequences." Annu Rev Genet **28**: 237-61.

Tatusov, R. L., N. D. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**(1): 41.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Tunlid, A. and N. J. Talbot (2002). "Genomics of parasitic and symbiotic fungi." Curr Opin Microbiol **5**(5): 513-9.

Valentine, J. W., D. Jablonski, et al. (1999). "Fossils, molecules and embryos: new perspectives on the Cambrian explosion." Development **126**(5): 851-9.

Van de Peer, Y. and R. De Wachter (1997). "Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA." J Mol Evol **45**(6): 619-30.

Venter, J. C., M. D. Adams, et al. (1998). "Shotgun sequencing of the human genome." Science **280**(5369): 1540-2.

Woese, C. R. (1977). "Endosymbionts and mitochondrial origins." J Mol Evol **10**(2): 93-6.

Woese, C. R. (1987). "Bacterial evolution." Microbiol Rev **51**(2): 221-71.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proc Natl Acad Sci U S A **74**(11): 5088-90.

Wolf, Y. I., L. Aravind, et al. (1999). "Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange." Trends Genet **15**(5): 173-5.

Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-13.

Wood, V., R. Gwilliam, et al. (2002). "The genome sequence of Schizosaccharomyces pombe." Nature **415**(6874): 871-80.

Wren, B. W. (2000). "Microbial genome analysis: insights into virulence, host adaptation and evolution." Nat Rev Genet **1**(1): 30-9.

Xie, G., C. A. Bonner, et al. (2003). "Lateral gene transfer and ancient paralogy of operons containing redundant copies of tryptophan-pathway genes in Xylella species and in heterocystous cyanobacteria." Genome Biol **4**(2): R14.

Yang, G., M. S. Rose, et al. (1996). "A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin." Plant Cell **8**(11): 2139-50.

Yoder, O. C. and B. G. Turgeon (2001). "Fungal genomics and pathogenicity." Curr Opin Plant Biol **4**(4): 315-21.

Zhang, X., C. Feschotte, et al. (2001). "P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases." Proc Natl Acad Sci U S A **98**(22): 12572-7.

Zhang, X., N. Jiang, et al. (2004). "PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements." Genetics **166**(2): 971-86.

Zuckerkandl, E. and L. Pauling (1965). "Molecules as documents of evolutionary history." J Theor Biol **8**(2): 357-66.