JONAS WEISSMANN GAIARSA

**HISTÓRIA EVOLUTIVA DE CARBO-HIDROLASES LIGNO-CELULÓSICAS DA
FAMÍLIA *XANTHOMONADACEAE***

Tese apresentada ao Programa de
Pós-Graduação Interunidades em
Biotecnologia USP/Instituto
Butantan/IPT, para obtenção do Título
de Doutor em Biotecnologia.

São Paulo
2013

JONAS WEISSMANN GAIARSA

**HISTÓRIA EVOLUTIVA DE CARBO-HIDROLASES LIGNO-CELULÓSICAS DA
FAMÍLIA *XANTHOMONADACEAE***

Tese apresentada ao Programa de
Pós-Graduação Interunidades em
Biotecnologia USP/Instituto
Butantan/IPT, para obtenção do Título
de Doutor em Biotecnologia.

Área de concentração: Biotecnologia

Orientadora: Profa. Dra. Marie-Anne Van
Sluys

Versão original

São Paulo
2013

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Gaiarsa, Jonas Weissmann.

História evolutiva de carbo-hidrolases ligno-celulósicas da família Xanthomonadaceae / Jonas Weissmann Gaiarsa. -- São Paulo, 2013.

Orientador: Profa. Dra. Marie-Anne Van Sluys.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan. Área de concentração: Biotecnologia. Linha de pesquisa: Fitopatologia.

Versão do título para o inglês: Evolutionary history of lignocellulosic carbo-hydrolases of the Xanthomonadaceae family.

1. Xanthomonas 2. Parede celular vegetal 3. Fitopatógenos
4. Polissacarídeos 5. Evolução molecular I. Sluys, Profa. Dra. Marie-Anne Van II. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan III. Título.

ICB/SBIB090/2013

UNIVERSIDADE DE SÃO PAULO
Programa de Pós-Graduação Interunidades em Biotecnologia
Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

---

Candidato(a): Jonas Weissmann Gaiarsa.

Título da Tese: História evolutiva de carbo-hidrolases ligno-celulósicas da família Xanthomonadaceae.

Orientador(a): Profa. Dra. Marie-Anne Van Sluys.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a ...../...../....., considerou

(   ) Aprovado(a)

(   ) Reprovado(a)

Examinador(a):   Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a):   Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a):   Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a):   Assinatura: .....
Nome: .....
Instituição: .....

Presidente: Assinatura: .....
Nome: .....
Instituição: .....

À minha família, aos meus colegas e professores,
pelo apoio, incentivo e inspiração.

AGRADECIMENTOS

Agradeço a Marie pela parcela que me foi confiada na formação do laboratório e pela formação que recebi durante todos esses anos de trabalho.

Ao Robson por me inspirar a sempre ir atrás da melhor maneira de fazer uma análise e achar a biologia por trás dos bits.

A todos que pelo GaTE passaram e que lá ainda estão. A todos que lá me ensinaram, me apoiaram, me amolaram e em mim confiaram.

A professora Jan Leach pelo material cedido.

A Simone Vidigal Alves e o professor Paulo Lee Ho pelo apoio na parte de proteômica.

Ao professor Marcos e todos do Lafieco por me ajudarem e abrirem espaço para os ensaios enzimáticos.

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro.

“Imagina ainda que esse homem volta à caverna e vai sentar-se no seu antigo lugar:
Não ficará com os olhos cegos pelas trevas ao se afastar bruscamente da luz do
Sol?”

Platão. A República. Livro VII

RESUMO

Gaiarsa JW. História evolutiva de carbo-hidrolases ligno-celulósicas da família *Xanthomonadaceae*. [tese (Doutorado em Biotecnologia)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2013.

O presente trabalho visa compreender o processo de degradação da parede celular vegetal de hospedeiros de fitopatógenos da família *Xanthomonadaceae*. Criamos e aperfeiçoamos uma técnica de enumeração dos genes relacionados ao metabolismo de polissacarídeos, com enfoque na distinção entre aqueles que agem sobre os componentes da parede celular vegetal e sobre outros polissacarídeos. A história evolutiva desse conjunto de enzimas foi delineada através de inferências sobre as relações de homologia entre os genes enumerados, sua presença ou ausência nos diversos genomas abordados e comparação das taxas de mutação entre grupos de homólogos. Além disso, procuramos também, com essa etapa de bioinformática e a etapa seguinte, incrementar a anotação desses genes, muitos descritos como hipotéticos ou com vaga definição de sua função. Na segunda parte do desenvolvimento do projeto foram feitos experimentos de expressão heteróloga e verificação da atividade enzimática para validação da anotação de alguns dos genes identificados. Foram exploradas diferentes técnicas *in silico* para procurar diferenças de potencial de maceração da parede celular vegetal. Quatro técnicas de enumeração de genes codificantes para enzimas de degradação da parede celular vegetal (PCWDE) foram construídas em sucessão, para melhor distinguir a atuação sobre polissacarídeos da parede celular vegetal e outros polissacarídeos. A primeira técnica procurou termos relativos à PCWDE sobre a anotação dos genomas; a segunda comparou esses genes com mais genomas; a terceira baseou-se na escolha de modelos de Markov (HMM) de conservação de aminoácidos do projeto Pfam para a seleção de genes e a quarta estratégia no uso do banco dbCAN de HMMs das sequências do projeto CAZy (*Carbohydrate Active enZymes*) em conjunto com um levantamento das atividades enzimática (ECs) relevantes, de forma a correlacionar polissacarídeos, ECs e domínios proteicos. Foi possível, então, a descoberta de novas putativas PCWDE, além da recuperação das já descritas. Os resultados aqui apresentados revelam que o repertório genômico PCWDE varia entre os genomas de *Xanthomonadaceae*. Procuramos aperfeiçoar a técnica de inferência automática de homologia de sequências de aminoácidos baseando-se em alinhamentos múltiplos globais. Essa inferência foi usada para traçar-se um perfil de ação do aparato de PCWDE de cada genoma sobre os diferentes polissacarídeos. Árvores filogenéticas de homólogos foram comparadas para suas taxas de mutação no intuito de verificar movimentos adaptativos semelhantes entre cada grupo de homólogos. Além disso, um levantamento dos genes de PCWDE pertencentes a regiões de transferência horizontal gênica (HGT) foi feito caracterizando suas contribuições aos genomas. As análises indicam que alguns genomas, em especial algumas espécies do gênero *Xanthomonas*, são detentores da maior parte da diversidade de PCWDE, tendo outros genomas apenas um subconjunto. No intuito de validar o trabalho de bioinformática fizemos um processo de seleção de genes de PCWDE candidatos para uma caracterização molecular. Os genes envolvidos no início da cascata de degradação, com sinal para secreção e que ainda não tivessem nenhum relato específico na literatura foram selecionados.

Dois dos genes submetidos a esse processo chegaram ao final com retificação de suas anotações pela atividade enzimática verificada.

Palavras-chave: *Xanthomonas*. Parede celular vegetal. Fitopatógenos. Polissacarídeos. Evolução molecular.

ABSTRACT

Gaiarsa JW. Evolutionary history of lignocellulosic carbo-hydrolases of the *Xanthomonadaceae* family. [Ph. D. thesis (Biotechnology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2013.

This study aims to understand the process of degradation of host plant cell walls by plant pathogens of the *Xanthomonadaceae* family. We created and perfected a technique for enumeration of genes related to the metabolism of polysaccharides, focusing on the distinction between those who act on components of plant cell wall and on other polysaccharides. The evolutionary history of this group of enzymes has been outlined through inferences about the relations of homology between the genes listed, their presence or absence in different genomes and comparison of mutation rates between groups of homologues. Moreover, we also intended with this bioinformatics step and the next step, to enhance the annotation of these genes, many described as hypothetical or vague in the determination of its function. In the second part of the project development heterologous expression and enzymatic activity assays were made to validate the annotation of some of the genes identified. We explored different *in silico* techniques to assess differences in plant cell wall maceration potential. Four techniques for enumeration of genes coding for plant cell wall degrading enzymes (PCWDE) were built in succession to better distinguish between activities on plant cell wall polysaccharides and other polysaccharides. The first technique searched for keywords relating to PCWDE in genome annotations, the second comparing these genes with a bigger set of genomes, the third was based on the selection of amino acid conservation Markov models (HMM) from the Pfam project used for gene selection and the fourth in the use of dbCAN HMMs constructed from sequences of the CAZy project (Carbohydrate Active enZymes) together with a survey of relevant enzymatic activities (ECs), in order to correlate polysaccharides, ECs and protein domains. It was then possible to discover new putative PCWDE, besides the recovery of already described ones. The results presented here show that the genomic PCWDE repertoire varies among *Xanthomonadaceae*. We seek also to improve the technique of automatic inference of homologous amino acid sequences through the use of multiple global alignments. This inference was used to outline the potential of action of the apparatus of PCWDE of each genome on the different polysaccharides. Phylogenetic trees of homologous groups were compared through their mutation rates in order to verify adaptive movements similar between each group. In addition, a survey of PCWDE genes belonging to horizontal gene transfer (HGT) regions was made assessing their contributions to the genomes. The analysis indicates that some genomes, especially certain species of the genus *Xanthomonas*, harbor most of the PCWDE diversity, with other genomes having only a subset. In order to validate the bioinformatics work done, a process of selection of candidate PCWDE genes was made for molecular characterization. The genes involved in the early steps of the degradation cascade, with secretion signal and not having yet had any specific reports in the literature were selected. Two of the genes undergoing this process have had their annotation corrected based on the enzymatic activity observed.

Keywords: *Xanthomonas*. Plant cell wall. Phytopathogens. Polysaccharides. Molecular evolution.

LISTA DE ILUSTRAÇÕES

Figura 1 - Representação esquemática dos componentes da parede celular vegetal primária.	27
Figura 2 – Vetores de clonagem.....	53
Figura 3 - 4-nitrofenil- β -D-glicopiranosídeo. A seta indica a ligação clivada por uma endoglucanase.	59
Figura 4 – Cluster do blastCORE mais favoravelmente resolvido pelos seus correspondentes do OrthoMCL. Cada caixa indica o ramo da árvore que contém as sequências dos agrupamentos do OrthoMCL.	80
Figura 5 - Alinhamento de um cluster exemplo desfavorável ao critério do OrthoMCL. Alinhamento global das sequências do cluster de dois genes. Baixa similaridade entre as duas sequências com muitos gaps e mismatches.	80
Figura 6 - Alinhamento de um cluster favorável ao critério do blastCORE. A mesma sequência da base da Figura 5 está no topo dessa figura. Essa proteína está dividida nesse genoma em relação aos outros genomas do mesmo gênero.	81
Figura 7 - Árvore do concatenado de óperons. As flechas segmentadas indicam a presença ou ausência de genes dos óperons indicados.	82
Figura 8 - Árvore do cluster e4o0025 com a mesma escala da árvore do cluster e4o0018, porém com uma disposição diferente.....	83
Figura 9 – Árvore do cluster e4o0028 com com disposição semelhante à árvore do cluster e4o0025, porém com uma escala 10 vezes menor	84
Figura 10 – Árvore do cluster e4o0018 com uma disposição diferente das árvores dos clusters e4o0025 e e4o0028, porém com a mesma escala da árvore de e4o0025.	84
Figura 11 – Relação de clusters, HGTs e distâncias entre árvores filogenéticas.	86
Figura 12 - Trecho do alinhamento de aminoácidos do cluster 106. Grifados os resíduos com polimorfismo ligados à mudança da classificação do CAZy.	91
Figura 13 – (p. anterior e esta) Diagrama do processo de clonagem e expressão de genes candidatos. Ao lado de cada bloco correspondente a um vetor de clonagem encontra-se um exemplo de teste ou etapa bem sucedido. Um painel de imagens dos géis encontra-se no APÊNDICE A -.	100
Figura 14 - Painel de absorbâncias de PNP. Cada bloco se refere ao teste com um dos 11 PNP-sacarídeos para os oito clones com proteína produzida. Foram feitas 9	

replicas e os dados submetidos ao teste de Tukey para distinção entre clones para cada PNP. BP e BZ indicam clones dos genes xop de *Xo oryzae* PXO99a e xoz de *Xo oryzicola* BLS256. 103

Figura 15 - Curva de carga por tempo de retenção em HPAEC-PAD de xilano de faia tratado com lisado dos 8 clones por 72 h a 37 °C. À esquerda a curva para o tratamento e a direita o controle. O pico a 42 min corresponde ao polissacarídeo intacto, o pico de 7 min corresponde a um dissacarídeo não identificado, a 4 min o pico de um monossacarídeo não identificado e a 2 min um contaminante não identificado do processo de lise. O pico de 2 min está presente em todas as outras curvas não mostradas. 104

Figura 16 – (Próxima página) Géis de agarose do processo de subclonagem do vetor pGEM-T Easy para o vetor pET-44b. Referência de peso molecular (bp) no terceiro gel na base da figura. Géis de agarose com concentração entre 0,5 e 0,7% em TAE 0,5X. 113

LISTA DE TABELAS

Tabela 1 - Espécies e patovares do gênero <i>Xanthomonas</i> (Rademaker et al., 2005).	22
Tabela 2 - Fenótipos de maceração.	24
Tabela 3 – Monômeros e ligações de hemiceluloses e pectinas.	26
Tabela 4 - Coleção de polissacarídeos da parede celular vegetal, suas ocorrências em monocotiledôneas e dicotiledôneas.	66
Tabela 5 - Coleção de polissacarídeos da parede celular vegetal, suas subunidades e ligações componentes.....	67
Tabela 6 - Coleção de polissacarídeos da parede celular vegetal, números EC das atividades enzimáticas associadas e domínios correspondentes das classes do CAZy. (continua)	68
Tabela 7 – Frequências absolutas de ocorrência de domínios CAZy para cada tipo de substrato.....	70
Tabela 8 - Cepas e abreviações correspondentes aos 36 genomas analisados.	71
Tabela 9 – Frequências de genes nas estratégias de enumeração de PCWDE.	72
Tabela 10 – Características de diversidade e frequências relativas de genes enumerados.	72
Tabela 11 – Frequência de genes enumerados para cada genoma em cada uma das estratégias.....	74
Tabela 12 - ECs e substratos.....	76
Tabela 13 – Médias de frequências de genes e substratos.	77
Tabela 14 - Tabela de frequências de genes e substratos.	78
Tabela 15 - Comparação entre blastCORE e OrthoMCL.....	81
Tabela 16 - Relação dos grupos de árvores de clusters, frequência de clusters e cores representativas.	87
Tabela 17 - Clusters com variação de CAZy. ....(continua)	89
Tabela 18 - Processo de seleção de genes candidatos de Xoc e Xop.	91
Tabela 19 - Tabela de predição da localização subcelular dos candidatos para clonagem de Xop e Xoz.	94

Tabela 20 - Oligonucleotídeos gerados com o auxílio dos softwares Primer-BLAST e Pribnon sobre as sequências genômicas.	95
Tabela 21 - Relação de oligonucleotídeos desenhados para subclonagem dos genes candidatos do vetor pGEM-T Easy para o vetor pET-44b.	96
Tabela 22 - Relação dos 24 genes candidatos de <i>Xanthomonas oryzae</i> pv. <i>oryzae</i> e <i>X.o.pvoryzicola</i>	98
Tabela 23 - Substratos sintéticos com um grupamento 4-Nitrophenyl (Sigma Chem. Co. St. Louis, USA) que pode identificar uma ação enzimática específica dependendo do açúcar e suas respectivas concentrações utilizadas no ensaio.....	102

LISTA DE ABREVIATURAS E SIGLAS

PCWDE – *Plant Cell Wall Degrading Enzymes*

PCWPS – *Plant Cell Wall Polysaccharides*

HMM – *Hidden Markov Model*

CAZy – *Projeto Carbohydrate Active enZymes*

cazy – enzima ativa sobre carboidrato

dbCAN - *DataBase for automated Carbohydrate-active enzyme Annotation*

KEGG - *Kyoto Encyclopedia of Genes and Genomes*

IUBMB – *International Union of Biochemistry and Molecular Biology*

HGT – *Horizontal Gene Transfer*

BLAST – *Basic Local alignment and Search Tool*

ORF – *Open Reading Frame*

EC – *Enzyme Commission*

Xo – *Xanthomonas oryzae*

Xc – *Xanthomonas campestris*

Xa – *Xanthomonas axonopodis*

PF – *Protein Family* (PFam)

GTR – *General Time Reversible*

JTT – *James-Taylor-Thornton*

RAxML - *Randomized Axelerated Maximum Likelihood*

BBH – (BLAST) *Bidirectional Best Hit*

MCL – *Markov Clustering Algorithm*

NCBI – *National Center for Biotechnology Information*

PCR – *Polymerase Chain Reaction*

X-gal – 5-bromo-4-cloro-3-indolil- β -D-galactopiranosídeo

EDTA – *Ethylenediamine tetraacetic acid*

IPTG – isopropil- β -D-1-tiogalactopiranosídeo

SDS-PAGE – *Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis*

TEMED – *Tetrametiletlenodiamina*

APS – *Ammonium persulfate*

PMSF – *Phenylmethanesulfonylfluoride*

PVPP – Polyvinylpolypyrrolidone

PNP - 4-nitrofenil- β -D-sacarídeo

HPAEC-PAD - cromatografia de troca iônica de alto desempenho com detecção amperométrica pulsada

JCVI – *John Craig Venter Institute*

CBM – *Carbohydrate-Binding Modules*

GH – *Glycoside Hydrolases*

GT – *Glycosyl Transferases*

CE – *Carbohydrate Esterases*

PL – *Polysaccharide Lyases*

IS – *Insertion Sequences*

OMCL, oxc, e4o – grupo do OrthoMCL

16S – RNA da subunidade maior do ribossomo de bactérias

BRENDA - BRAunschweig ENzyme DAtabase

T2SS, T3SS – Sistemas bacterianos de secreção do tipo 2 e 3.

pb, bp – pares de bases (DNA).

Kb, Mb – quilo-, megapares de bases.

SUMÁRIO

1 INTRODUÇÃO	21
1.1 <i>Xanthomonadaceae</i>	21
1.2 Mecanismos de instalação do patógeno	21
1.3 Enzimas de degradação da parede celular vegetal (PCWDE)	25
1.4 Polissacarídeos da parede celular vegetal (PCWPS)	25
1.5 PCWDE de <i>Xanthomonadaceae</i>	28
2 OBJETIVO	30
3 MATERIAL E MÉTODOS	31
3.1 Formatos e manipulação de dados	31
3.2 Enumeração dos genes de PCWDE	32
3.2.1 PCWDE nos clusters do projeto de genômica comparativa de <i>Xylella</i>	32
3.2.2 Genes PCWDE recuperados por gene isca	32
3.2.3 Seleção de PCWDE por text mining e dobramento proteico	33
3.2.4 Uma relação de composição dos polissacarídeos e modelos HMM curados do dbCAN	35
3.3 Designação de substratos às enzimas	35
3.3.1 EC atribuído na anotação	35
3.3.2 Pfam e InterPro	36
3.3.3 IUBMB e sequências moleculares de referência do KEGG	36
3.3.4 Substratos e conteúdo de acordo com EC, CAZy e dbCAN	36
3.4 Clusterização	37
3.4.1 Desenvolvimento de uma técnica de clusterização: BlastCORE	37
3.4.2 Comparação de clusterizações para inferências sobre relações de homologia	37
3.4.3 OrthoMCL	38
3.5 Árvores	38
3.5.1 Árvores de máxima verossimilhança	38
3.5.2 Distância de árvores e clusterização	39
3.6 Comparação de perfis de genes homólogos (PHROG)	39
3.7 PCWDE e HGTs	41
3.7.1 Mapeamento de ilhas de transferência horizontal gênica	41
3.8 Seleção de genes candidatos para trabalho de biologia molecular	41
3.8.1 Dobramento de glicosil hidrolase	41

3.8.2 Início da cascata de degradação.....	42
3.8.3 Sinal para periplasma, membrana externa ou secreção	42
3.8.4 Genes não abordados na literatura	43
3.8.5 Software de desenho de primers de alta especificidade para amplificação sobre um genoma previamente conhecido (PribNon)	43
3.9 Biologia Molecular: Clonagem, expressão e testes enzimáticos.....	44
3.9.1 Reações de PCR.....	44
3.9.2 Células competentes.....	45
3.9.2.1 Preparo de células DH10B competentes por eletroporação.....	45
3.9.2.2 Preparo de células BL21(DE3) quimio competentes por choque térmico	45
3.9.3 Transformações	46
3.9.3.1 Transformação de células DH10B competentes por eletroporação	46
3.9.3.2 Transformação de células BL21(DE3) quimio competentes por choque térmico	46
3.9.4 Recuperação de transformantes	47
3.9.5 Preparação de DNA plasmidial em pequena escala	47
3.9.6 Eletroforese em gel de agarose	49
3.9.7 Purificações.....	50
3.9.7.1 Purificação de reação de PCR	50
3.9.7.2 Purificação de DNA por precipitação.....	51
3.9.7.3 Purificação de banda excisada de gel agarose	51
3.9.8 Clonagem de produto de PCR com enzima Pfu em vetor pGEM-T Easy	51
3.9.9 Testes de confirmação de clonagem.....	52
3.9.9.1 Teste de confirmação para pJET1.2.....	52
3.9.9.2 Teste de confirmação para pET-44b	52
3.9.10 Procedimento de sequenciamento pela técnica de Sanger automatizada	53
3.9.11 Subclonagem sítio dirigida em pET-44b.....	54
3.9.12 Desenho de primers com adaptadores para subclonagem sítio dirigida em pET-44b	54
3.9.13 PCR sobre material de clones pGEM.....	55
3.9.14 Clonagem ponta cega no vetor pJET1.2	55
3.9.15 Digestão de DNA de clones no vetor pJET1.2 e ligação no pET-44b	55
3.9.16 Indução de expressão de genes clonados na cepa de E.coli BL21 (DE3)	56
3.9.17 SDS-PAGE.....	56

3.9.17.1 Lise bacteriana para ensaios enzimáticos	58
3.9.18 Testes enzimáticos	58
3.9.18.1 Atividade endoglunásica sobre 4-nitrofenil- β -D-sacarídeos	58
3.9.18.2 Atividade sinérgica sobre hemiceluloses e extratos de parede celular	59
4 RESULTADOS E DISCUSSÃO	61
4.1 Enumeração dos genes cazy e das PCWDE	61
4.1.1 Genes não-PCWDE e número restrito de genomas	61
4.1.2 Heterogeneidade entre anotações, hidrolases não-PCWDE e baixa diversidade de PCWDE	62
4.1.3 Dicionários, PFs e aumento da confiabilidade	63
4.1.3.1 Domínios RHS	64
4.1.4 Listagem de cazys ligados a PCWDE	64
4.1.5 Resumo das estratégias	70
4.1.6 Diferenças entre <i>Xanthomonas</i> , <i>Xylella</i> e <i>Stenotrophomonas</i>	72
4.2 Predição de atividades e substratos associados às enzimas	73
4.2.1 EC atribuído na anotação	75
4.2.2 Pfam e InterPro	75
4.2.3 Substratos de acordo com ECs da IUBMB e sequências do KEGG	75
4.2.4 Substratos atribuídos através da relação entre CAZys e ECs	76
4.2.5 Conteúdo de PCWDE relativo às atividades sobre cada PCWPS	76
4.3 Relação de homologia entre os genes enumerados	79
4.3.1 Clusterizações: BLASTclust com genes isca e o desenvolvimento do blastCORE	79
4.3.2 Comparação entre blastCORE e OrthoMCL	79
4.3.3 União dos resultados do blastCORE e do OrthoMCL	81
4.4 Árvores filogenéticas de genes homólogos	82
4.4.1 Comparação do perfil gênico de PCWDE homólogos entre genomas com agrupamento por distância entre árvores filogenéticas	84
4.4.2 Variação na estrutura de genes homólogos	88
4.5 Seleção de candidatos para validação da anotação por biologia molecular	91
4.5.1 Genes na literatura	92
4.6 Clonagem	92

4.6.1 Amplificações de material genômico, ligações em pGEM e transformação em DH10B.....	92
4.6.2 Iniciadores de PCR.....	95
4.6.3 Subclonagem em pET-44b e BL21(DE3)	97
4.7 Indução de expressão heteróloga e SDS-PAGE	101
4.8 Testes enzimáticos	101
4.8.1 Atividade endoglunásica sobre 4-nitrofenil- β -D-sacarídeos	101
4.8.2 Atividade sinérgica sobre hemiceluloses e extratos de parede celular	104
5 CONCLUSÕES	105
REFERÊNCIAS*	108
APÊNDICE A - Géis de eletroforese	113
APÊNDICE B - Código de computador gerado	115
APÊNDICE C - Artigo de periódico	126

1 INTRODUÇÃO

1.1 *Xanthomonadaceae*

A ordem *Xanthomonadales*, à qual pertence sua maior família *Xanthomonadaceae*, é um grupo que contém dois gêneros de fitopatógenos com diversos genomas sequenciados desde 2000 (Simpson et al., 2000). Dentre os gêneros, *Xanthomonas* se destaca por possuir uma ampla gama de hospedeiros, sendo o gênero composto por 20 espécies classificadas em 80 patovares (Rademaker et al., 2005) (Tabela 1). Perdas mundiais de safra causada por patógenos somam tipicamente 10% da produção (Strange, Scott, 2005). Três das espécies de destaque dentre as que causam os maiores prejuízos são *Xanthomonas campestris*, *Xanthomonas axonopodis* e *Xanthomonas oryzae*. No entanto, já há quatro décadas vem se empregando a espécie *Xanthomonas campestris* de forma industrial na produção de goma xantana, um importante biopolímero empregado nas indústrias de alimentos, cosméticos e petróleo. O gênero *Stenotrophomonas*, embora tenha dois genomas sequenciados de cepas causadoras de infecções oportunistas em pacientes imunodeprimidos, também pode ser encontrada associada a plantas apresentando atividade antifúngica (*Stenotrophomonas rhizophila*) (Wolf et al., 2002) e na forma de espécie endofítica fixadora de nitrogênio (*Stenotrophomonas pavanii*) (Ramos et al., 2011).

1.2 Mecanismos de instalação do patógeno

Existem duas fases sequenciais distintas no início da patogênese no que tange a instalação da infecção. A primeira é à entrada do agente patogênico em um órgão da planta, vencendo quaisquer barreiras físicas que se interponham, e a segunda, constituída em uma de duas possibilidades, a detecção da invasão pelos sistemas de resistência da planta e a resposta de hipersensibilidade ou a de fato colonização, fruto da não detecção do patógeno, nesse caso considerando-se o hospedeiro susceptível a infecção. Existe um período de incubação antes da manifestação dos sintomas patogênicos, mais longo em algumas variedades do que em outras, tempo onde acontece a colonização. Esse tempo pode ser necessário, dependendo da variedade, para aumento do número de células no biofilme, ativação do metabolismo

derivado da sinalização de *quorum sensing* e atuação dos mecanismos de evasão da detecção como os atuadores secretados pelo sistema de secreção do tipo 3, que modificam o metabolismo da célula vegetal impedindo uma resposta de hipersensibilidade.

Tabela 1 - Espécies e patovares do gênero *Xanthomonas* (Rademaker et al., 2005).

Espécie	Patovar	Espécie	Patovar	Espécie	Patovar
<i>fragariae</i>		<i>vasicola</i>	<i>vasculorum</i>	<i>axonopodis</i>	<i>Sesbaniae</i>
<i>populi</i>		<i>axonopodis</i>	<i>begoniae</i>	<i>axonopodis</i>	<i>Vignaeradiatae</i>
<i>arboricola</i>		<i>axonopodis</i>	<i>alfalfae</i>	<i>axonopodis</i>	<i>Vignicola</i>
<i>cassavae</i>		<i>axonopodis</i>	<i>cassavae</i>	Sublinhados táxons com genomas sequenciados.	
<i>vodiaei</i>		<i>axonopodis</i>	<i>cassiae</i>		
<i>bromi</i>		<i>axonopodis</i>	<i>citrumelo</i>		
<i>cucurbitae</i>		<i>axonopodis</i>	<i>coracanae</i>		
<i>pisi</i>		<i>axonopodis</i>	<i>cyamopsidis</i>		
<i>melonis</i>		<i>axonopodis</i>	<i>desmodii</i>		
<i>vesicatoria</i>		<i>axonopodis</i>	<i>desmodiigangetici</i>		
<i>hyacinthi</i>		<i>axonopodis</i>	<i>desmodiitundifolii</i>		
<i>theicola</i>		<i>axonopodis</i>	<i>erythrinae</i>		
<i>sacchari</i>		<i>axonopodis</i>	<i>lespedezae</i>		
<i>albilineans</i>		<i>axonopodis</i>	<i>patelii</i>		
<i>arboricola</i>	<i>corylina</i>	<i>axonopodis</i>	<i>phyllanthi</i>		
<i>arboricola</i>	<i>juglandis</i>	<i>axonopodis</i>	<i>poinsettiicola</i>		
<i>arboricola</i>	<i>poinsettiicola</i>	<i>axonopodis</i>	<i>ricini</i>		
<i>arboricola</i>	<i>populi</i>	<i>axonopodis</i>	<i>tamarindi</i>		
<i>arboricola</i>	<i>pruni</i>	<i>axonopodis</i>	<i>vesicatoria</i>		
<i>campestris</i>	<i>aberrans</i>	<i>axonopodis</i>	<i>axonopodis</i>		
<i>campestris</i>	<i>armoraciae</i>	<i>axonopodis</i>	<i>vasculorum</i>		
<i>campestris</i>	<i>barbareae</i>	<i>axonopodis</i>	<i>manihotis</i>		
<i>campestris</i>	<i>campestris</i>	<i>axonopodis</i>	<i>phaseoli</i>		
<i>campestris</i>	<i>incanae</i>	<i>axonopodis</i>	<i>nova citri E</i>		
<i>campestris</i>	<i>raphani</i>	<i>axonopodis</i>	<i>dieffenbachiae</i>		
<i>campestris</i>	<i>musacearum</i>	<i>axonopodis</i>	<i>bauhiniae</i>		
<i>hortorum</i>	<i>hederae</i>	<i>axonopodis</i>	<i>cajani</i>		
<i>hortorum</i>	<i>pelargonii</i>	<i>axonopodis</i>	<i>citri</i>		
<i>hortorum</i>	<i>vitians</i>	<i>axonopodis</i>	<i>clitoriae</i>		
<i>oryzae</i>	<i>oryzae</i>	<i>axonopodis</i>	<i>desmodiilaxiflori</i>		
<i>oryzae</i>	<i>oryzicola</i>	<i>axonopodis</i>	<i>glycines</i>		
<i>translucens</i>	<i>arrhenatheri</i>	<i>axonopodis</i>	<i>malvacearum</i>		
<i>translucens</i>	<i>cerealis</i>	<i>axonopodis</i>	<i>vitians</i>		
<i>translucens</i>	<i>graminis</i>	<i>axonopodis</i>	<i>nova vitians</i>		
<i>translucens</i>	<i>bromi</i>	<i>axonopodis</i>	<i>nova phaseoli</i>		
<i>translucens</i>	<i>hordei</i>	<i>axonopodis</i>	<i>cajani</i>		
<i>translucens</i>	<i>phlei</i>	<i>axonopodis</i>	<i>aurantifolii</i>		
<i>translucens</i>	<i>phleipratensis</i>	<i>axonopodis</i>	<i>aurantifolii</i>		
<i>translucens</i>	<i>poae</i>	<i>axonopodis</i>	<i>aurantifolii</i>		
<i>translucens</i>	<i>secalis</i>	<i>axonopodis</i>	<i>dieffenbachiae</i>		
<i>translucens</i>	<i>translucens</i>	<i>axonopodis</i>	<i>phaseoli</i>	<i>var.</i>	
<i>translucens</i>	<i>undulosa</i>		<i>fuscans</i>		
<i>vasicola</i>	<i>holcicola</i>	<i>axonopodis</i>	<i>rhynchosiae</i>		

Na fase de entrada, a atuação das PCWDE (Tabela 2) pode estar presente no rompimento da barreira física imposta pela parede celular, permitindo um contato mais direto entre o patógeno e as células do hospedeiro ou o acesso ao sistema vascular da planta hospedeira. Já na fase de colonização da planta essas enzimas podem atuar na hidrólise da parede proporcionando carboidratos nutrientes para o patógeno. Os dois gêneros fitopatogênicos tratados nesse trabalho, *Xanthomonas* e *Xylella*, apesar de compartilharem boa parte de seus genes, possuem diferentes estratégias de dispersão, invasão e colonização de seus hospedeiros. *Xanthomonas* com genomas sequenciados possuem três estratégias principais de dispersão e invasão. Trazidos através de agentes físicos ambientais, como chuva e vento, e entrando por estômatos foliares, se estabelecendo nas câmaras estomatares ou no mesófilo e depois atingindo o sistema vascular; ou entrando por hidatódios, entrando diretamente no sistema vascular e lá se estabelecendo e depois se dispersando pelos órgãos da planta. Também pode invadir pela injúria mecânica, em geral fruto da manipulação no cultivo, estabelecendo-se possivelmente diretamente junto a células e também sendo disperso pelo sistema vascular. *Xylella* possui um mecanismo que envolve a atuação de um inseto vetor sugador, que ao se alimentar da planta injeta a bactéria, que carrega no seu trato digestivo, diretamente no sistema vascular da planta.

Dentre as espécies do gênero *Xanthomonas* existe também uma variação de estratégias, sendo a maior distinção encontrada entre patovares que colonizam sistemicamente o hospedeiro e patovares que se restringem ao sistema vascular.

Tabela 2 - Fenótipos de maceração.

Nome	Hospedeiros	Tecidos	Maceração
<i>Xanthomonas campestris</i> patovar <i>vesicatoria</i>	Tomate e <i>Capsicum</i>	Mesófilo e parênquima (EPPO XANTVE)	Manchas superficiais granulosas ou crostas, com margens encharcadas de água, oval ou irregular, 2-10 mm de diâmetro.
<i>Xanthomonas axonopodis</i> patovar <i>citri</i>	Citros	Mesófilo e parênquima (EPPO XANTCI)	Lesões nas folhas, caules e frutos com margens levantadas, marrom, banhadas de água, geralmente com um halo amarelo ou anel em torno da lesão.
<i>Xanthomonas campestris</i> patovar <i>armoraciae</i>	<i>Brassica oleracea</i>	Mesófilo e parênquima (Kamoun; Kado, 1990).	Numerosas pequenas manchas pretas, circulares, encharcadas de água, eventualmente, com halos amarelos na parte inferior das folhas, dando uma cor amarelada a folhagem e aparência irregular. Manchas foram também visíveis na superfície superior das folhas e em pecíolos
<i>Xanthomonas campestris</i> patovar <i>campestris</i>	Crucíferas e Brassicáceas	Vascular (Kamoun; Kado, 1990).	Interno (enegrecimento vascular)
<i>Xanthomonas campestris</i> patovar <i>musacearum</i>	Ensete, banana e milho.	Mesófilo e vascular	Folhas infectadas tendem a quebrar ao longo do limbo foliar (http://www.cabi.org/isc/?compid=5&dsid=56917&loadmodule=datasheet&page=481&site=144)(http://www.promusa.org/tiki-index.php?page=Xanthomonas+wilt)
<i>Xanthomonas oryzae</i> patovar <i>oryzae</i>	Arroz, gramíneas e sedges.	Mesófilo, parênquima e vascular (EPPO XANTOR).	Folhas de plantas jovens, como verde-claro a verde-cinza, estrias encharcadas de água perto da ponta da folha e margens. Essas lesões se aglutinam e se tornar branco-amarelada, com bordas onduladas.
<i>Xanthomonas oryzae</i> patovar <i>oryzicola</i>	Todas as espécies selvagens do gênero <i>oryza</i>	Coloniza o parênquima e se limita a regiões vasculares da planta (EPPO XANTOR)	A doença aparece inicialmente como leves estrias intervenais encharcadas de água. As estrias são de início de cor verde escuro, mas depois se tornam translúcidas. As estrias ampliam e aglutinam-se, eventualmente tornando-se castanho claro.
<i>Xylella fastidiosa</i>	Citros, parreira, espiroleira, amêndoa.	Vascular (EPPO XYLEFA)	Interno

Hospedeiros, tecidos afetados e características da maceração de algumas espécies e patovares com genomas sequenciados. EPPO se refere aos códigos de patógenos da Organização Europeia e Mediterrânea para a proteção de plantas (<http://www.eppo.int/QUARANTINE/quarantine.htm>) e suas listas de descrições.

1.3 Enzimas de degradação da parede celular vegetal (PCWDE)

Genes codificadores para enzimas com atividade sobre carboidratos ou mais especificamente, enzimas de degradação de polissacarídeos da parede celular de plantas são conhecidos por ser uma característica particular de genomas de microrganismos associados a plantas incluindo a família *Xanthomonadaceae*. Em parasitas, PCWDE têm um papel fundamental no processo de infecção do hospedeiro, de modo que já foram implicados como fatores de virulência. Essas enzimas também desempenham um papel muito importante na indústria nascente de etanol lignocelulósico também conhecido como o bioetanol de segunda geração. Dentro da família *Xanthomonadaceae*, apenas algumas das cepas de *Xanthomonas* foram analisadas quanto à expressão, atividade enzimática de alguns dos genes PCWDE e sobre o papel que os sistemas de secreção têm na atuação das enzimas (Hsiao et al., 2007; Hu et al., 2007; Rajeshwari et al., 2005; Ray et al., 2000; Schröter et al., 2001; Sun et al., 2005; Xiao et al., 2008).

Uma das restrições da atividade de uma enzima sobre múltiplos polissacarídeos é o tipo de ligação, sendo que enzimas que agem sobre ligações betas não conseguem agir sobre ligações alfa e vice-versa. Portanto PCWDE em geral não agem sobre glicose alfa ligada, preponderante em amilose, amilopectina, glicogênio e dextrana, mas não em PCWPS.

1.4 Polissacarídeos da parede celular vegetal (PCWPS)

A parede celular vegetal (Figura 1) é composta por três ambientes: a membrana celular, a parede e o apoplasto. São enumerados três grandes grupos de polissacarídeos que compõe a parede celular vegetal, celulose, hemicelulose e pectina, sendo celulose um polissacarídeo de composição invariável de glicopirranose β -(1 $\rightarrow$ 4) ligada e as outras duas classes compostas por diversas variações de estrutura (Tabela 3). Dentre os polissacarídeos hemicelulósicos a literatura enumera 4 tipos com ainda outras variações de composição: mananos (manano linear, glucomanano, galactoglucomanano, galactomanano), xilanos (xilano, glucuronoxilano, glucuronoarabinoxilano), xiloglucano (xiloglucano, fucogalactoxiloglucano) e glucanos β -(1 $\rightarrow$ 3,1 $\rightarrow$ 4) (liquenano). Já os polissacarídeos pécticos podem ser classificados em até 6 tipos: homogalacturonanos,

ramnogalacturonano-I, galacturonanos substituídos (ramnogalacturonano-II, xilogalacturonano, apiogalacturonano), arabinano, arabinogalactano-I, arabinogalactano-II.

Na distribuição desses polissacarídeos entre as espécies vegetais celulose e mananos ocorrem ambos em dicotiledôneas e monocotiledôneas. Xilanos ocorrem principalmente na parede de monocotiledôneas, mas em pequenas quantidades em dicotiledôneas. Xiloglucanos, ao contrário, ocorrem principalmente em dicotiledôneas, mas em pequenas quantidades na parede primária de monocotiledôneas. Já glucanos $\beta(1\rightarrow3,1\rightarrow4)$ ocorrem única e exclusivamente na parede primária de monocotiledôneas. Pectinas em geral ocorrem principalmente em dicotiledôneas, mas em pequena quantidade na parede celular primária de monocotiledôneas.

Tabela 3 – Monômeros e ligações de hemiceluloses e pectinas.

			Hemicelulose	Pectina
Ligação principal	da	cadeia	$\beta(1\rightarrow4)$	$\alpha(1\rightarrow4)$
Açúcares principal	da	cadeia	D-Glcp, D-Xylp, D-Galp, D-Manp	D-GalpA, L-Rhap, D-GlcpA.
Ligações secundárias	nas	cadeias	$\alpha(1\rightarrow2)$ $\beta(1\rightarrow2)$ $\beta(1\rightarrow6)$	$\alpha(1\rightarrow2)$ $\alpha(1\rightarrow3)$ $\alpha(1\rightarrow5)$ $\alpha(3\rightarrow5)$ $\alpha(2\rightarrow3)$ $\beta(1\rightarrow4)$ $\beta(1\rightarrow2)$ $\beta(1\rightarrow3)$ $\beta(1\rightarrow5)$ $\beta(2\rightarrow3)$ [$\beta(1\rightarrow3)$ -D-Apif]
Outros açúcares nas cadeias secundárias			L-Fucp, L-Araf.	L-Araf, L-Arap, L-Fucp, D-Dha, D-Kdo, L-AcefA, D-Xylp.
Modificações de açúcares de cadeia lateral			O-Acetil	4-O-Metoxi

. As composições de hemicelulose e pectina podem variar de acordo com tecido, estágio de desenvolvimento, contexto ambiental e também com a espécie.

Figura 1 - Representação esquemática dos componentes da parede celular vegetal primária.

1.5 PCWDE de *Xanthomonadaceae*.

Dois dos artigos se voltaram para o sistema de secreção do tipo 2 a influência que ele pode ter no processo de infecção. Ray et al. (2000) e Sun et al. (2005) demonstraram a atividade de uma xilanase e uma celulase em isolados de *Xanthomonas oryzae* pv. *oryzae* (Xoo) de Beijing e das Filipinas (cepa sequenciada PXO99), respectivamente. Principalmente, que mutantes para genes do sistema de secreção de proteínas de *Xanthomonas* (xps; sistema tipo dois) apresentavam virulência reduzida e ausência de bandas de 42 kDa e 50 kDa (xilanase e celulase, respectivamente) num gel de proteína da porção extracelular. Não houve, no entanto, alteração na resposta de hipersensibilidade em espécies não suscetíveis e ausência da banda de xilanase para o selvagem da cepa 8004 de Xcc.

Com foco no melhoramento da produção de goma xantana o grupo de Schröter et al. (2001) produziu um mutante da cepa FC2 (NRLLB-1459 derivada) de *Xanthomonas campestris* pv. *campestris* (Xcc) para os genes engXC A e B. O duplo mutante possuía redução de 80% da atividade endoglucanásica sobre celulose.

No trabalho de Rajeshwari et al. (2005) foi exposto o efeito de diminuição de virulência da cepa BXO43 na mutagênese da xilanase xynB e da lipase/esterase lipA, conjuntamente e individualmente. Foi também verificada a atividade de ambas sobre xilano e tween80/tributirina, respectivamente.

Hsiao et al. (2007) exploraram o papel regulatório do fator sinalizador difusível (DSF) sintetizado pelo gene *rpfF* e da proteína fator de transcrição análoga à proteína receptora de cAMP (*clp*). Mostraram a existência de sítio promotor *clp* à montante de uma série de genes codificantes para pectatolases e poligalacturonases de diversas *Xanthomonas*. Além disso, fizeram uma caracterização enzimática da poligalacturonase *pehA* da cepa Xc17 de Xcc e averiguaram a indução dessa atividade na presença em meio de cultura de ácido poligalacturônico e a inibição na presença de catabólitos da reação (glicose), na limitação de oxigênio e na falta de nitrogênio.

Já Hu et al. (2007) fizeram ensaios sobre Xoo e expressão heteróloga em *E. coli* da celulase *eglXoB* da cepa PXO99. O mutante para esse gene tinha virulência reduzida, a expressão do gene foi comprovada *in planta*, porém a verificação de atividade em placa de cultura teve retorno negativo, assim como a atividade só foi verificada nas porções intracelulares no heterólogo em *E. coli*.

No trabalho de Xiao et al. (2008) fez-se um trabalho biotecnológico de mineração dos genomas de *Xcc* e *Streptomyces coelicolor* para pectinases. Verificou-se a atividade em expressão heteróloga em *E. coli* de três pectatolases e duas poligalacturonases de *Xcc*. Uma possível pectatolase não apresentou atividade, fato atribuído a uma especificidade da enzima que não estava apta a degradar a pectina de citros utilizada. Dentre os genes contemplados nesse trabalho temos os mesmos de Hsiao et al. (2007).

Por fim, Aparna et al. (2009) divulgaram uma análise cristalográfica, enzimática e de virulência do gene *lipA* de *Xoo* BXO43 mostrando que este difere do dobramento canônico de outras lipases. Sua atividade se restringe a triacilglicerídeos de cadeia curta, não agindo sobre cadeias mais longas que C6 e possui características que indicam a uma ação sobre alquil ésteres de ligação cruzada entre polissacarídeos da parede vegetal. A mutação de alguns resíduos particulares dessa proteína inibem respostas de imunidade inata na planta de arroz (deposição de calose e morte celular programada). Esse gene não foi incluído na análise por se tratar de um dobramento proteico pouco usual para PCWDE e novo na literatura. Porém, foram encontrados outros dois grupos homólogos: um com um gene de cada um dos genomas de *Stenotrophomonas* e o outro em todas as *Xanthomonadaceae*. Esses genes possuem o domínio Pfam LIP, pertencente ao clã AB_hydrolase, mas ainda não é contemplado no CAZy.

2 OBJETIVO

Como *Xanthomonas* faz a hidrólise da parede celular de seu hospedeiro? O objetivo desse trabalho é traçar o perfil de PCWDE nos genomas desses fitopatógenos.

Qual o repertório gênico de PCWDE da família *Xanthomonadaceae* e qual sua história nos genomas da família? O trabalho propõe estabelecer uma técnica de bioinformática para enumeração desses genes, comparar o conteúdo entre os genomas de *Xanthomonadaceae* e sondar o potencial de ação sobre a planta hospedeira. Algumas das perguntas que surgiram ao longo do desenvolvimento foram: Como distinguir entre enzimas que ajam sobre PCWPS e não sobre outros polissacarídeos? Como melhor determinar grupos de genes homólogos? Como melhor prever os substratos de uma PCWDE?

Qual é a confiabilidade do processo de anotação eletrônica? Na segunda parte do trabalho elencam-se candidatos relevantes para ensaios moleculares com o objetivo de validar as atividades enzimáticas anotadas para as proteínas.

O conhecimento gerado deve dar subsídio a novas abordagens de tratamento de cultivares para as doenças causadas por *Xanthomonadaceae* ou auxiliar na seleção de variedades resistentes aos patógenos. Ela também estabelece um framework para a descoberta de genes PCWDE em outros organismos, contribuindo para pesquisas de hidrólise de biomassa.

3 MATERIAL E MÉTODOS

3.1 Formatos e manipulação de dados

A manipulação de informações de anotação foi feita em sua maioria na forma de tabelas de texto com campos separados por caracteres de tabulação (TSV) para o uso com as ferramentas de linha de comando do sistema operacional Linux. Por exemplo, as informações de anotação obtidas dos bancos de genomas foram obtidas ou formatadas de acordo com a convenção de TSV *Protein Table* (PTT) do NCBI. Outras manipulações não necessariamente foram usaram formatos convencionais. Algumas manipulações mais pontuais, ou em menor volume de operações, foram feitas sobre os TSV, mas com software de planilha como Microsoft Excel, OpenOffice Calc e Gnumeric, por vezes sendo mantido no formato nativo desses softwares.

Sequências de DNA e proteína foram obtidas e manipuladas de acordo com a convenção Pearson/IUPAC (FASTA) e sua derivação de múltiplas entradas, multifasta. Convenções NCBI de extensões de nomenclatura de multifasta: faa, multifasta do proteoma de um genoma; ffn, multifasta do DNA das ORFs de um genoma; fna, fasta do DNA do cromossomo do genoma.

Os arquivos de saída do software de procura e alinhamento local (*Basic Local Alignment and Search Tool*; BLAST) foram usados principalmente no formato extenso com alinhamentos (-m 0; mais comumente usado) e no formato de tabela resumida sem cabeçalho (-m 8) de acordo com a utilização a ser feita.

Os softwares de alinhamento múltiplo global, principalmente o T-Coffee, mas por vezes Clustal e Muscle, foram usados com formato de saída FASTA alinhados ou, no caso da análise CORE particular do T-Coffee, em formato próprio. Para análise do RaxML foi usado o formato PHYLIP intercalado e formato de manipulação de árvores filogenéticas foi o Newick.

Os locustags de genes contidos nas informações de anotação foram abreviados para compatibilidade com vários programas de análise, como os softwares de alinhamento múltiplo. Isto é, vários softwares de análise não permitem que os identificadores possuam mais que 12 caracteres.

Os programas gerados para automatização da sequência de análises (*pipeline*) são compostos por programas escritos nas linguagens Perl e Shell Script.

3.2 Enumeração dos genes de PCWDE

As quatro estratégias descritas a seguir foram desenvolvidas em série, uma como consequência da outra.

3.2.1 PCWDE nos clusters do projeto de genômica comparativa de *Xylella*

O trabalho iniciou-se com um levantamento de clusters da web site do projeto de genômica comparativa de *Xylella fastidiosa* (adiante referido como Comparativo; www.xylella.incc.br). Esse projeto é constituído da reanotação dos genomas das cepas 9a5c e Temecula1 e da anotação final de Dixon e Ann1, isoladas das plantas de amêndoa e oleandro (espírradeira ou flor de são josé), além do estabelecimento de agrupamentos (clusters) gênicos incluindo outros genomas além de *Xylella*.

Os clusters foram recuperados valendo-se de palavras chaves retiradas dos artigos que compõem parte das referências bibliográficas. As palavras usadas foram: *Cellulase*, *Cellobiosidase*, *Xylan*, *Xylosidase*, *Beta-glucosidase*, *Arabinosidase*, *Pectatylase*, *Pectin*, *Polygalacturonase* e *Rhamnogalacturon*.

3.2.2 Genes PCWDE recuperados por gene isca

O diretório de genomas de bactéria do NCBI foi baixado como constava em 19/08/2008, de 767 genomas. Desse diretório foi gerada uma lista das espécies de *Xanthomonadaceae* lá presentes e essa lista usada para gerar um faa contendo todas as sequências dessas espécies. Foram também incorporados alguns genomas encontrados disponíveis no projeto Comprehensive Microbial Resources (CMR) do J. Craig Venter Institute (JCVI) (tabela 2).

Foram selecionados 19 genomas de *Xanthomonadaceae* para execução dessa etapa. O genoma de *Xanthomonas vasicola* pv. *musacearum* utilizado estava no estado de *draft* à época dessa análise.

Esse faa foi agregado aos faa dos clusters do Comparativo e submetido a uma clusterização por vínculo simples de BLAST, com o programa BLASTclust do NCBI. Esse algoritmo se baseia na identificação de uma série de sequências como sendo de um mesmo cluster se estiverem dentro de um parâmetro mínimo de

cobertura do alinhamento de um dado gene em relação aos outros e dentro de um parâmetro mínimo de corte de identidade de aminoácidos/nucleotídeos ou de pontuação por posição alinhada. Através dessa técnica procurou-se recuperar os genes correspondentes nos genomas não contemplados no Comparativo. Isto é, usaram-se os genes recuperados do Comparativo como molde, ou isca, para procurar genes homólogos em outros genomas.

Ao se listar os diferentes identificadores (*tags*) dos *locustags* de anotação verificou-se que o Comparativo contava com uma nova versão da anotação de alguns dos genomas de *Xylella* e era então necessária a obtenção de uma correspondência entre as duas versões com o mantenedor do site para comparação entre todas as *Xylellas*. A compilação final dos genes foi realizada através de um programa para recuperar as sequências das proteínas com um faa por cluster em 19 genomas constantes da lista unificada.

3.2.3 Seleção de PCWDE por text mining e dobramento proteico

O objetivo é selecionar um conjunto de genes que pertençam a algum contexto funcional metabólico delineado por um conjunto de termos (dicionário). No contexto desse trabalho o dicionário foi composto com palavras retiradas das páginas web do CCRC (*Complex Carbohydrate Research Center*) da Universidade da Geórgia, que descrevem processos metabólicos envolvendo componentes da parede celular vegetal, somando ao conjunto de termos usados na abordagem inicial.

Esse dicionário é utilizado para a identificação de dobramentos proteicos relevantes, em primeira instância, e na filtragem de linhas de definição dos genes, ou “anotações” em segunda instância, ou seja, na ausência de dobramentos identificáveis. Escolhemos a ferramenta HMMER em conjunto com os modelos probabilísticos de sequências de aminoácidos para domínios proteicos disponibilizados através do Pfam. Procuramos nos valer também do vínculo parcial entre Pfam, CAZy e Interpro para recuperar mais domínios e mais informações a respeito dos domínios. No Interpro constam descrições adicionais sobre os domínios do Pfam (PFs).

O método de procura do HMMER 2 utilizado foi o hmmpfam, com base no banco Pfam_ls contra todos os multifasta de proteínas dos genomas desse trabalho

e com corte de E-value de e^{-1} . Foi usado esse E-value, pois empiricamente demonstrou ser um valor estanque para a detecção, já que houve pequena detecção de domínios até um valor de corte maior em pelo menos duas ordens de grandeza (e^1).

Em seguida é feita a seleção de sequências que não apresentaram identificação com nenhum PF.

Para tentar encontrar PCWDE dentre essas sequências foi feita então outra seleção elencando-se uma lista de termos e expressões regulares a ser usada para filtragem de genes pela sua linha de definição fornecida na forma de uma tabela de descrição de proteínas (ptt). Para tal o dicionário usado na seleção de PFs é reconstruído e novamente manualmente curado incluindo-se dessa vez o vocabulário encontrado na descrição dos genes selecionados na etapa anterior.

Faz-se um dicionário de expressões inicial a partir de arquivos texto ou web sites, que são então manualmente reduzidas a expressões regulares compatíveis com Perl e selecionadas para conter apenas termos ligados aos processos metabólicos de interesse. São baixadas todas as descrições dos PFs constantes das páginas web do próprio Pfam e também do InterPro e são selecionados modelos Pfam (PFs) de acordo com a presença de palavras do dicionário. Uma primeira leva de PFs é filtrada com o dicionário e em seguida os PFs restantes do Clã correspondente também podem ser incluídos. Os PFs selecionados foram mais reduzidos, baseando-se na descrição do modelo, averiguando-se indícios de ação sobre monômeros e ligações químicas pertencentes aos polissacarídeos da parede celular vegetal.

Adicionalmente foi escrito um pipeline, chamado `glipfammer.sh`, que faz predição de ORFs de um genoma e tradução das sequências nucleotídicas para aminoácidos antes da procura `hmmpfam`.

Para tornar o pipeline mais eficiente também foi escrito o programa `pfam_process_threads.2.0.sh` que é um paralelizador do interpretador de resultados do `hmmpfam`, `process_hmmpfam.pl` (Copyright Robson Francisco de Souza 2008), usado na análise. Ao final é produzido um multifasta com os genes selecionados.

3.2.4 Uma relação de composição dos polissacarídeos e modelos HMM curados do dbCAN

O dicionário foi substituído por um levantamento mais detalhado das relações entre todos os polissacarídeos conhecidos por ocorrer na parede celular vegetal, a maneira como ocorrem em espécies de plantas eudicotiledôneas e monocotiledôneas (onde são classificados todos os hospedeiros das *Xanthomonadaceae*), as atividades enzimáticas conhecidas por ocorrerem na hidrólise desses polissacarídeos e as famílias do CAZy correspondentes a essas atividades.

O trabalho desenvolvido por Yiu et al. (2011) teve o objetivo de tornar a identificação de cazys em novos genomas e metagenomas uma atividade mais rápida e precisa, contando com o conhecimento gerado e depositado nos banco de dados do projeto CAZy. Para tal, foi feita a composição de HMMs específicos a cada família de classificação do projeto, utilizando-se das sequências lá referenciadas, e disponibilizadas conjuntamente com scripts Perl para filtragem e reformatação da saída do HMMER.

Dessa forma, foi feita uma procura dos modelos CAZy relevantes da tabela de polissacarídeos da parede celular vegetal sobre os genomas das *Xanthomonadaceae*.

3.3 Designação de substratos às enzimas

O processo de designação de substratos às enzimas levantadas evoluiu ao longo de uma série de 4 diferentes estratégias.

3.3.1 EC atribuído na anotação

Inicialmente partimos da descrição do número EC associado à proteína para fazer uma, ou mais, atribuições de substrato sobre o qual a enzima poderia estar agindo. Isso depende da existência de um número EC já associado à proteína avinda da sua anotação, o que não é sempre o caso.

3.3.2 Pfam e InterPro

Em seguida procuramos na descrição InterPro dos PFs atribuídos a um cluster evidências sobre o tipo de monômeros e ligações químicas, além de números EC, presentes nos polissacarídeos da parede celular vegetal compatíveis com aquela enzima.

3.3.3 IUBMB e sequências moleculares de referência do KEGG

Uma da literatura direcionada permitiu a composição de uma listagem dos polissacarídeos presentes na parede celular vegetal, sua ocorrência em monocotiledôneas e dicotiledôneas, atividades enzimáticas de degradação associadas, famílias do projeto CAZy associadas e a subunidades componentes dos polissacarídeos. A listagem de nomes dos polissacarídeos foi usada na seleção das atividades enzimáticas e números EC listados no site web da União Internacional de Bioquímica e Biologia Molecular (<http://www.chem.qmul.ac.uk/iubmb/enzyme>).

A atribuição de atividade e levantamento de novos genes foi realizada em banco local constituído pelas sequências de aminoácidos referência do banco KEGG. Uma procura foi efetuada com o programa tBLASTn e o banco de nucleotídeos dos genomas. Foi usado o corte de E-value de e^{-30} e coberturas de 75%.

Números EC órfãos de sequências moleculares foram submetidos à técnica de Yamada e Bork (2012). Nesse trabalho foram analisadas vizinhanças genômicas, e metagenômicas montadas, de regiões com genes de atividade conhecida. Através de correlações entre a conexão de reações em um metabolismo e a proximidade entre genes codificantes para as enzimas em questão no genoma, foi possível fazer associações de números EC às sequências. Para fazer a associação com os genes de *Xanthomonadaceae* foi feito um novo tBLASTn das proteínas elencadas naquele trabalho contra os genomas.

3.3.4 Substratos e conteúdo de acordo com EC, CAZy e dbCAN

Por fim com a geração da tabela que relaciona polissacarídeos, ECs e CAZys pudemos mais plenamente fazer atribuições de substratos às enzimas apenas

cruzando a informação levantada com o dbCAN e os polissacarídeos correspondentes na tabela. Dessa forma unimos o processo de enumeração e de atribuição de substratos.

3.4 Clusterização

3.4.1 Desenvolvimento de uma técnica de clusterização: BlastCORE

Na segunda etapa do processo de enumeração de genes foi utilizado o programa BLASTclust para a formação de conjuntos por similaridade de sequência de aminoácidos. A dificuldade em se obter parâmetros de cobertura e similaridade adequados para esse programa motivou o desenvolvimento de um procedimento de escolha objetiva dos parâmetros, de forma a fortalecer o resultado da clusterização para uma análise de homologia gênica.

O blastCORE parte da formação de clusters de vínculo não recíproco (em contraste ao BBH) formados pela ferramenta BLASTclust e da verificação iterativa da qualidade desses clusters pelo alinhador múltiplo global T-Coffee. O T-Coffee possui um método de alinhamento que permite a geração de uma pontuação de qualidade do alinhamento chamada pontuação CORE. Esta é usada internamente pelo T-Coffee para qualificar os alinhamentos gerados durante o processo iterativo de refinamento do alinhamento global. Esse critério é avaliado para cada posição alinhada, mas também é usado para o cálculo de uma estatística para cada par de sequências e também para o alinhamento múltiplo. O critério de alinhamento satisfatório é uma pontuação CORE de no mínimo 70. No blastCORE o processo de construção e verificação da qualidade de clusters é feito iterativamente por secção dos intervalos de parâmetros de percentual de identidade e cobertura de sequências do BLASTclust.

3.4.2 Comparação de clusterizações para inferências sobre relações de homologia

Com o intuito de verificar a qualidade da clusterização do blastCORE recorremos ao OrthoMCL, um software usado amplamente para clusterização e recomendado no trabalho de comparação de Chen et al. (2007).

Foi feita a execução dos dois algoritmos distintos sobre o mesmo conjunto de sequências e os clusters dos dois processos são então escolhidos de forma a aumentar o CORE médio dos clusters.

3.4.3 OrthoMCL

Usamos o pacote OrthoMCL (Li et al., 2003), uma variante da técnica já estabelecida de melhores resultados BLAST recíprocos (Bidirectional Best Hit; BBH) analisada nesse caso pelo software de clusterização de grafos MCL (Van Dongen, 2008).

O pacote é composto dos seguintes passos de análise: filtragem de sequências por critérios qualidade (tamanho mínimo de 10 aminoácidos e no máximo 20% de stop códons), BLAST de todos os proteômas (tradução das ORFs) contra todos os proteômas, cálculo de percentual de acerto para cada par de BLAST, procura de possíveis pares de parálogos recentes, ortólogos e co-ortólogos e agrupamento de pares pelo programa MCL.

Eventos de paralogia recente se tratam duplicações recentes de um gene dentro de um genoma, ortologia são eventos de descendência recente entre genomas e co-ortologia é uma relação composta da duplicação em um genoma de uma relação de descendência recente entre genomas.

O pacote de análise estatística MCL é usado para agrupamento através de uma matriz de distâncias geradas a partir da esperança das melhores parêlas de resultados recíprocos de BLAST (*BBH*) de sequências de aminoácidos.

Os dados processados no antepenúltimo e penúltimo passo são feitos num banco de dados mySQL para manipulação pelo código do projeto OrthoMCL.

Os parâmetros utilizados foram padrão, *percentMatchCutoff* 50 e *evalueExponentCutoff* -5.

3.5 Árvores

3.5.1 Árvores de máxima verossimilhança

A partir dos alinhamentos múltiplos globais de aminoácidos e nucleotídeos de cada cluster, foram gerados 100 bootstraps de máxima verossimilhança com o

auxílio do software RAxML (Stamatakis, 2006), usando-se o modelo de substituição de nucleotídeos GTR para nucleotídeos e JTT para aminoácidos, ambos com estimação de distribuição γ para até quatro categorias de taxas de mutação. Ao final é produzida a árvore de máxima verossimilhança consenso dos bootstraps, com comprimentos de ramos.

3.5.2 Distância de árvores e clusterização

A distância entre árvores foi comparada com o software treedist do pacote Phylip (Felsenstein, 2005) usando-se o método de pontuação de ramos para árvores com comprimento de ramos e comparados todos os pares de árvores. Para a análise das sequências de aminoácidos foi escrito um software para se achar táxons comuns não repetidos entre clusters, uma restrição para análise do *treedist*. Também fizemos uma nova fragmentação dos clusters, separando grupos de ramos próximos entre genomas, porém repetidos. Os pares de árvores foram comparados duas a duas gerando-se um novo alinhamento com esses táxons, dessa vez com o software MUSCLE (parâmetros padrão), por ser mais rápido que o T-Coffee. Fez-se então nova análise do RAxML para então se submeter o par de árvores ao *treedist*. As distâncias produzidas foram reformatadas para uma tabela com três colunas de pares únicos usando-se o colog da distância normalizado para valores positivos. Essa tabela então submetida ao software de clusterização de grafos MCL, o mesmo usado no OrthoMCL, e fazendo-se uma varredura do parâmetro de inflação do MCL escolheu-se aquele produzisse o maior número de clusters. No caso das sequências de nucleotídeos foi feita uma comparação similar, porém apenas com a árvore da subunidade 16S ribossomal, com táxons comuns não repetidos. Como essa comparação foi feita sempre em relação à árvore de 16S, as distâncias das árvores foram agrupadas por picos de frequência de distâncias e não pelo MCL.

3.6 Comparação de perfis de genes homólogos (PHROG)

Uma análise sobre o número máximo de ocorrências de arquiteturas de modelos PFam dentre todos os genomas serviu de inspiração para a criação de um

programa escrito em Perl que foi chamado de Perfil de Homólogos no Repertório Gênico (PHROG).

Um perfil de homólogos nessa forma, como caracteres dos genomas estudados, é uma forma de verificação de diferenças entre genomas e no auxílio de inferências sobre diferenças de fenótipos em organismos com conteúdo gênico similar; em particular no caso desse trabalho diferenças qualitativas sobre a capacidade do aparato de catabólise de polissacarídeos da parede celular vegetal.

O programa recebe como entrada a lista de clusters e é gerada uma tabela com uma linha para cada genoma e colunas para cada um dos clusters. O primeiro passo é transformar a tabela de caracteres de cada táxon em uma tabela, com um eixo de táxons e outro de caracteres, e criar células vazias de caracteres inexistentes de um táxon. O passo mais trabalhoso desse processo, e por isso feito pelo programa `phrog_plot.pl`, é inserir as células nulas de forma a respeitar a organização dos caracteres em uma ordem geral comum. Durante o processo podemos lidar com caracteres repetitivos de duas maneiras: somar ou não células de caracteres equivalentes e, se somadas, transformadas em caracteres binários, presença ou ausência do caractere. Podemos também, ao invés de imprimir nas células quaisquer símbolos para representar a presença, ausência ou repetitividade dos caracteres, imprimir quaisquer dados que se julgar relevante a respeito dos genes representados. No caso das análises aqui realizadas foram impressas as coordenadas de início de cada ORF. Podemos então reordenar os dados no eixo dos caracteres de acordo com um genoma de referência. Essa tabela pode então ser submetida a um software de plotagem, no caso desse trabalho o Excel (Microsoft Corp., Redmond, WA, EUA). As duas maneiras encontradas para melhor representação visual através do Excel foram o gráfico do tipo 3D, achatado e reformatado para se assemelhar a uma tabela, ou fazer uso do recurso de formatação condicional da própria tabela para colorir as células. Em ambos os casos ajustes dos recursos do Excel podem ser feitos para melhor representar quantidades binárias, absolutas ou outras informações. No caso das coordenadas das ORFs, como feito nesse trabalho, foi usada uma formatação condicional degrade para denotar as diferentes regiões do cromossomo bacteriano.

3.7 PCWDE e HGTs

Procuramos mapear a influência de fenômenos de transferência horizontal gênica (HGT) com um levantamento em regiões que compõem os genomas de *Xanthomonadaceae*.

3.7.1 Mapeamento de ilhas de transferência horizontal gênica

Foi gerado um mapeamento das ilhas de *X. campestris* pv. *campestris* e *X. axonopodis* pv. *citri* a partir do que foi estabelecido como eventos de HGT em Lima et al. (2005) propagando-se esses resultados sobre os genomas agregados a esse trabalho. As regiões HGT do trabalho anterior serviram como referência para uma busca por BLASTn com corte de E-value de 0 sobre os genomas acrescentados no atual trabalho.

3.8 Seleção de genes candidatos para trabalho de biologia molecular

Dentro dos putativos genes de PCWDE elencados procuramos refinar genes que seriam de interesse para um trabalho de caracterização molecular. Para tal procuramos determinadas características disponíveis através da anotação feita sobre os genomas, a informação disponível na literatura e preditores *in silico*.

3.8.1 Dobramento de glicosil hidrolase

Dentre os domínios Pfam encontrados, na 3ª enumeração, parte deles foi atribuída a outros tipos de enzimas do tipo hidrolase. Particularmente, AB_hydrolase, Lipase_GDSL e LIP são dobramentos encontrados em hidrolases com substratos proteicos, lipídicos, além de glicosídicos. Com o intuito de obter um conjunto com um sinal de predição o mais confiável possível decidimos por excluir proteínas com dobramentos como estes, que não fossem imediatamente associados à PCWDE.

Em seguida, foram selecionados apenas genes com função enzimática, em detrimento daqueles apenas com domínios para função de associação ao substrato

(Pfam: RHS_repeat) e também com domínio Pfam que não tivessem associação mais imediata com polissacarídeos da parede celular vegetal (Pfam: Alpha-amylase).

3.8.2 Início da cascata de degradação

Da mesma forma, alguns dos dobramentos encontrados são em geral associados com a degradação de oligossacarídeos produzidos após a hidrólise inicial dos componentes da parede celular vegetal. Em especial, os dobramentos Glyco_hydro_2 e Glyco_hydro_3 aparentemente são ligados à decomposição intracelular de oligossacarídeos.

3.8.3 Sinal para periplasma, membrana externa ou secreção

De modo a fortalecer a seleção de um gene candidato, e de forma que também a focar em genes de interesse biotecnológico, restringimos o conjunto de genes candidatos para clonagem procurando por proteínas com indícios de localização no periplasma, na membrana externa ou direcionada para secreção. Essas enzimas também são mais prováveis de atuar no início da cascata de degradação.

Para tal fizemos uso de quatro preditores de localização subcelular consensuados na forma de uma média ponderada dos resultados (probabilidades percentuais). As ferramentas Psortb (Gardy et al., 2005) e PA-SUB (Lu et al., 2004) contaram cada uma com 40% da pontuação, enquanto as ferramentas Cello (Yu et al., 2004) e SOSUIgramn (Imai et al., 2008), cada uma pesou 10% na pontuação. O percentual possível não totaliza 100%, pois algumas das ferramentas consideram alguns pares de compartimentos em outras ferramentas como um só. Essa ponderação se deve ao fato de que os dois últimos preditores sempre forçam a localização em um ou outro compartimento enquanto os dois primeiros possuem um critério de localização incerta que não tenta fazer uma predição que pode ser potencialmente errada.

Na primeira etapa de seleção observou-se a pontuação relativa à localização subcelular e foram procurados genes que tivessem obtido pelo menos 40% em qualquer um dos compartimentos periplasmático, membrana externa e meio extracelular e incerteza até 40%. Além disso, foram descartados genes que

obtiveram sinal positivo para secreção pelo sistema do tipo três (T3SS) através da ferramenta T3SS_prediction (Löwer, Schneider, 2009) e da presença de PIP boxes (Lu et al., 2008).

3.8.4 Genes não abordados na literatura

Fizemos uma última filtragem excluindo genes, ou seus homólogos através da identificação por clusterização, que já tivessem sido abordados na literatura em algum tipo de ensaio de expressão heteróloga, *knock-out*, silenciamento, caracterização na virulência ou outros ensaios moleculares.

3.8.5 Software de desenho de primers de alta especificidade para amplificação sobre um genoma previamente conhecido (*PribNon*)

O objetivo do programa desenvolvido é o de reproduzir o comportamento da ferramenta Primer-BLAST (Ye et al., 2012) disponível no site do NCBI, que no entanto não permite a execução sobre um genoma que não esteja contemplado no banco de dados da instituição. Essa ferramenta permite o desenho de iniciadores para uma reação de PCR com poucas inespecificidades em termos de anelamentos secundários, ou fora de região alvo para a amplificação.

De forma geral o software procura a região genômica a ser amplificada, agregando uma janela de pares de bases em redor da região do gene, e repassa essas informações para o software Primer3 (Untergasser et al., 2012) que gera uma série de oligonucleotídeos que obedeçam aos parâmetros de região de anelamento, temperatura de melting, tamanho do oligonucleotídeo, entre outros. Os parâmetros usados no software Primer3 foram: 70 a 500 bp de região para geração do oligonucleotídeo, geração de 3000 pares, T_m com mínimo de 57 °C, ótimo em 55 °C e máximo de 63 °C, janela de 1 °C entre os iniciadores e algoritmo Santa Lucia/1998 de cálculo e correção de sal da T_m.

Esses oligonucleotídeos são então submetidos a uma procura por BLAST com parâmetros permissivos: word size pequeno, abertura e extensão de gap pouco custosos e E-value alto para resgatar hits pequenos ($W=8-9$, E-value: $10e^4$) sobre o genoma conhecido. São então considerados bons iniciadores aqueles que apresentarem baixa correspondência, ou hits inespecíficos de BLAST, com regiões

genômicas que não a região alvo em pares dos oligonucleotídeos com até 5000 pb de distância.

3.9 Biologia Molecular: Clonagem, expressão e testes enzimáticos

Os procedimentos de biologia molecular foram feitos usando-se como referência “*Molecular Cloning: A laboratory manual*” (Sambrook et al., 2001).

3.9.1 Reações de PCR

As amplificações da primeira etapa foram feitas usando-se como molde DNA genômico das duas cepas aqui abordadas, cedido pela Profa. Jan Leach da Colorado State University.

Foram feitas variações nas condições da reação de PCR de concentração de magnésio (Mg^{2+}), temperatura de dissociação da dupla fita de DNA e volume de reação até se chegar às condições de 2 mM de $MgCl$ e temperatura de annealing de 55 °C e feito em 3 reações de 15 µL. Foi usada a polimerase com verificação de leitura (Pfu) contida no produto Elongase da empresa Life Technologies.

Foi usado um programa de termociclador para amplificação de longos trechos de DNA de acordo com parâmetros empíricos testados em trabalhos anteriores do GaTE Lab. A rotina usada foi: 30 s a 95 °C; 10 repetições compostas por 30 s a 95 °C, 30 s a 55 °C e 3 min a 68 °C; 25 repetições compostas por 30 s a 95 °C, 30 s a 55 °C e 3 min a 68 °C com acréscimo de 20 s por repetição; 4 min a 95 °C; 7 min a 68 °C; término a 12 °C.

Na PCR para teste de confirmação de clonagem foi usada a enzima DreamTaq da empresa Fermentas com a condição padrão de concentração magnésio já incluso no tampão de reação que acompanha a enzima. O programa de termociclador utilizado foi: 30 s a 95 °C; 35 repetições compostas por 30 s a 95 °C, 30 s a 55 °C e 2 min a 72 °C; 2 min a 95 °C; 5 min a 72 °C; término a 12 °C.

3.9.2 Células competentes

3.9.2.1 Preparo de células DH10B competentes por eletroporação

Células de *Escherichia coli* cepa DH10B foram submetidas a um protocolo de competência para transformação por eletroporação. O processo é iniciado estriando uma colônia ou glicerol da cepa DH10B em uma placa de LB-Agar com estreptomicina 12,5 µg/mL e incubando-se por 16 a 18 horas a 37 °C. Também se prepara de antemão 3 a 4 l de água deionizada autoclavada. É então feito o préinóculo com uma colônia isolada da estria recém-feita em 5 mL de meio LB contendo estreptomicina 12,5 µg/mL (em um tubo de polipropileno de 50 mL: 5 mL LB + 1,25 µL de estreptomicina 20 mg/mL). Incuba-se por 18 horas a 37 °C sob agitação de 200 rpm. Após as 18 horas, colocar um volume de 400 mL de meio SOB em estufa a 37 °C. Retirar 4 mL do préinóculo e inocular o meio SOB a 37 °C sem o antibiótico. Agitar a 200 rpm por 2 h a 37 °C e medir a OD₆₀₀ em intervalos de 10 min até atingir aproximadamente 0,65. Ao atingir a OD desejada, dividir os 400 mL em 8 tubos de polipropileno de 50 mL e incubar em gelo por 20 min. Centrifugar por 5 min a 7000 G e descartar o sobrenadante. Ressuspender cada um dos pellets em 50 mL de H₂O deionizada autoclavada gelada. Neste passo o volume total será 400 mL. Centrifugar por 5 min a 7000 G e descartar o sobrenadante. Ressuspender cada um dos pellets em 25 mL de H₂O deionizada autoclavada gelada. Neste passo o volume total será 200 mL. Juntar dois falcons em 1, totalizando 4 tubos. Centrifugar por 5 min a 7000 G e descartar o sobrenadante. Ressuspender um dos 4 pellets em 8 mL de glicerol 10% autoclavado e gelado. Enquanto isso os outros 3 pellets devem ficar no gelo. Ressuspender os demais pellets na mesma solução, um de cada vez, juntando todas as células no final. Centrifugar por 5 min a 7000 G e descartar o sobrenadante. Ressuspender em 1,2 a 1,6 mL de glicerol 10% (v/v) autoclavado e gelado e fazer alíquotas de 50 µL e congelar imediatamente em nitrogênio líquido. Estocar em freezer -80 °C (rende cerca de 50 alíquotas).

3.9.2.2 Preparo de células BL21(DE3) quimio competentes por choque térmico

Células de *Escherichia coli* cepa BL21(DE3) foram submetidas a um protocolo de competência para transformação por choque térmico. Faz-se um préinóculo em 5

mL de meio LB crescido a 37 °C por 16 h. O volume de 1 mL do préinóculo é inoculado em um novo volume de 50 mL de meio LB. Incuba-se a 37 °C por 2 h ou até atingir uma OD₆₀₀ de 0,3. As células são coletadas por centrifugação a 1000 G por 15 min a temperatura ambiente e então ressuspensas em 20 mL de cloreto de magnésio 0,1M gelado. A ressuspensão é incubada no gelo por 30 min e então centrifugada a 1000 G por 15 min a 4 °C. Finalmente o pellet de células é ressuspendido em 2 mL de cloreto de magnésio 0,1M gelado e aliquoteado e congelado a -80 °C em tubos de microcentrífuga de 1,5 mL, para uso posterior.

3.9.3 Transformações

3.9.3.1 Transformação de células DH10B competentes por eletroporação

A reação de ligação do kit pGEM-T Easy requereu a dessalinização através de uma membrana de nitrocelulose 0,22 µM Millipore sobre uma solução de polietileno glicol a 10% por 30 min. Do contrário a condutância da solução causaria formação de arco voltaico na cubeta de transformação, sem no entanto proporcionar a abertura de poros na membrana celular das células competentes. 5 µL recuperados desse processo foi usado para transformação de células competentes por eletroporação. A concentração salina da reação de ligação do kit CloneJet não foi considerada impeditiva à transformação, porém sendo usado um volume de 2 µL da reação. Foram usados os parâmetros 2,5 kV, 25 µF, 200 Ω. Em se produzindo um pulso com duração em torno de 5 ms o processo é considerado bem sucedido do ponto de vista da condutância da solução. Foi usado um eletroporador Bio-Rad Gene Pulser II dotado dos módulos Capacitance Extender e Pulse Controller.

3.9.3.2 Transformação de células BL21(DE3) quimio competentes por choque térmico

A transformação de células quimio competentes de *E.coli* BL21(DE3) para transformação por choque térmico foi feita com DNA extraído para o processo de confirmação após a ligação no vetor pET-44b. O procedimento foi feito de acordo com o que se segue. São feitas placas de LB-ágar mais o antibiótico marcador. As alíquotas de células competentes são descongeladas no gelo por 5 min. Adiciona-se

2 μ L do DNA a um tubo de polipropileno de 15 mL e a alíquota de células competentes. O tubo é mantido no gelo por 20 min e então a 42 °C por 20 a 30 s em banho de circulação. Rapidamente remove-se o tubo para o gelo por 2 min adicionando-se 300 μ L de SOC. Incuba-se a 37 °C sob agitação por 30 min, plaqueia-se e então se incuba a 37 °C por 16 a 18 h.

3.9.4 Recuperação de transformantes

Os transformantes foram recuperados em 300 μ L de meio de cultura SOC por 30 min e em seguida espalhados com alça de Drigalski sobre placas de Petri contendo meio de cultura Luria-Bertani (LB), antibiótico Ampicilina (Amp) na concentração 100 μ g/ml e X-Gal 40 μ g/ml. A presença de X-Gal só é necessária no caso de ser possível a triagem de colônias azuis e brancas, como na cepa DH10B com o vetor pGEM mas não em BL21(DE3) e também não com os vetores pJET1.2 ou pET-44b, já que DH10B não possui a cadeia alfa da enzima β -galactosidase contida em redor do sítio de clonagem do vetor pGEM. As placas foram então incubadas em estufa a 37 °C por 16 a 18 h. No caso da cepa DH10B contendo o vetor pGEM essa etapa rendeu em média 20 colônias sendo que 75% eram azuis e 25% eram brancas.

3.9.5 Preparação de DNA plasmidial em pequena escala

Colônias com menos de 7 dias foram transferidas com palito de madeira para 2 ml de meio de cultura LB contendo ampicilina 100 mM e crescidas sob agitação de 200 rpm e 37 °C por 16 h em uma incubadora (New Brunswick Innova 44). Esse inóculo foi então submetido a um protocolo de preparação em pequena escala de DNA plasmidial.

Transfira 1,5 ml da cultura para um tubo de microcentrífuga e centrifugue a velocidade máxima por 30 s a 4 °C. Guarde a parte não utilizada da cultura original, a 4 °C. Remover o meio por aspiração, deixando o sedimento bacteriano tão seco quanto possível.

Ressuspender em 300 μ l de solução STE (10 mM Tris-Cl (pH 8,0), cloreto de sódio 0,1 M, 1 mM de EDTA (pH 8,0)) gelada por vortex vigoroso, centrifugar a

velocidade máxima por 30 s a 4 °C e remover novamente o sobrenadante. Ressuspender em 100 µl de solução de lise alcalina gelada I (50 mM de glicose, 25 mM Tris-Cl (pH 8,0), 10 mM de EDTA (pH 8,0)) por vortex vigoroso. Adicionar 200 µl da solução de lise alcalina solução II (NaOH 0,2 N recentemente diluído a partir de um estoque de 10 N, 1% (p/v) de SDS. Prepare a solução na hora e use à temperatura ambiente.), fechar o tubo firmemente, e misturar o conteúdo invertendo o tubo rapidamente cinco vezes. Não use vortex. Armazenar o tubo em gelo e adicionar 150 µl de solução de lise alcalina III (60 mL acetato de potássio 5 M, 11,5mL ácido acético glacial, 28,5mL H₂O) gelada. Fechar o tubo de lise alcalina e inverter o tubo várias vezes. Armazenar o tubo no gelo por 3 a 5 min e centrifugar o lisado bacteriano à velocidade máxima durante 5 min a 4 °C. Transferir o sobrenadante para um novo tubo e adicionar um volume igual de fenol: clorofórmio. Misturar as fases orgânica e aquosa por vortex e depois centrifugar a emulsão a uma velocidade máxima durante 2 min a 4 °C. Transferir a fase aquosa superior para um novo tubo. Precipitar os ácidos nucleicos por adição de 2 volumes de etanol à temperatura ambiente. Misturar a solução por vórtex e, em seguida, deixar a mistura em repouso durante 2 min à temperatura ambiente. Sedimentar os ácidos nucléicos precipitados por centrifugação à velocidade máxima durante 5 min a 4 °C. Remover o sobrenadante por aspiração suave. Coloque o tubo em posição invertida sobre uma toalha de papel para permitir que todo o líquido seja retirado. Usar um papel toalha ou uma ponta da pipeta para remover quaisquer gotas de fluido que aderem às paredes do tubo. Adicionar 1 ml de etanol a 70% e inverter o tubo fechado várias vezes. Recuperar o DNA pela centrifugação à velocidade máxima durante 2 min a 4 °C. Remova todo o sobrenadante por aspiração tomando cuidado com o pellet que por vezes não adere firmemente ao tubo. Remover quaisquer gotas de etanol, que restarem nas paredes do tubo. Deixe o tubo aberto à temperatura ambiente até que o etanol evapore e nenhum fluido seja visível no tubo (5 a 10 min). Dissolvem-se os ácidos nucleicos em 50 µl de TE (pH 8,0) ou água deionizada contendo 20 µg/ml de DNase livre de RNase A (RNase pancreática). Agite em vortex suavemente durante alguns segundos. Armazene a solução à temperatura de -20 °C.

3.9.6 Eletroforese em gel de agarose

Géis de agarose foram preparados com concentrações de 0,5% a 1% em peso por volume de TAE 0,5X em tamanhos de 8,4 cm x 7,1 cm, 15,2 cm x 12,1 cm e 25 cm x 24 cm. A mistura foi aquecida em forno de micro-ondas até a fervura e então resfriada até uma temperatura tolerável ao toque e despejada sobre o molde. O gel solidificado era então colocado dentro da cuba de eletroforese e coberto com TAE 0,5X. As amostras eram misturadas a tampão de carregamento (10 mM Tris-HCl (pH 7.6), 0.03% azul de bromofenol, 0.03% xileno cianol FF, 60% glicerol, 60 mM EDTA) na proporção 6:1 e carregadas nos poços do gel com o auxílio de uma micropipeta. A cuba era então ligada a uma fonte de corrente contínua regulada para 5 a 30 V/cm, dependente do tamanho do gel, por 20 a 60 min, tempo necessário para o azul de bromofenol atingir a região do gel próxima ao anodo. O gel foi então submetido a um banho de brometo de etídeo 5 µg/ml por 5 min e então lavado por 15 min em água. Alternativamente foi usado o corante *Blue green Loading dye* (LGC biotecnologia, Cotia, SP, Brasil) na forma de um tampão de carregamento misturado às amostras. Foi usada a referência de peso molecular GeneRuler 1 kb DNA Ladder (Thermo Fisher Scientific Inc., Waltham, MA, EUA) carregada em paralelo às amostras nos géis.

3.9.7 Purificações

3.9.7.1 Purificação de reação de PCR

Foi feita a purificação por coluna de afinidade através do kit NucleoSpin III (MACHEREY-NAGEL GmbH & Co. KG, Düren, Alemanha). O protocolo seguinte é apropriado para remoção de sais e enzimas da reação de PCR. Antes de iniciar a preparação verifique se o tampão NT3 foi preparado da forma correta. Para os volumes de amostra muito pequenos, menores que 30 µL, ajustar o volume a mistura de reação para 50 a 100 µl com água deionizada. Mistura-se 1 volume da amostra com 2 volumes do tampão NTI. Para a remoção de pequenos fragmentos, como dímeros de iniciadores, diluições de tampão NTI podem ser usadas ao invés de 100%. Coloque uma Coluna NucleoSpin Gel e PCR Clean-up em um tubo de coleta (2 mL) e carregar até 700 µL da amostra. Centrifugar por 30 s a 11.000 G. Descarte o líquido e coloque a coluna de volta no tubo de coleta. Carregar amostra restante, se necessário, e repetir o passo de centrifugação. Adicionar 700 µl de tampão de NT3 à coluna e centrifugar por 30 s a 11.000 G. Descartar o líquido e colocar a coluna de volta para no tubo de coleta. É recomendado repetir o passo anterior para minimizar a presença de sal e melhorar a razão A260/A230 do purificado. Centrifugar durante 1 min a 11.000 G para remover o tampão TE3 completamente. Certifique-se que a coluna de spin não entre em contato com o líquido ao removê-lo do tubo de coleta e descartar o líquido. Note, o etanol residual de tampão TE3 pode inibir reações enzimáticas. A remoção total do etanol pode ser obtida por incubação das colunas por 2 a 5 min a 70 °C antes da eluição. Coloque a coluna em um novo tubo de microcentrifuga de 1,5 mL (não fornecido). Adicionar 15 a 30 µL de tampão NE e incubar a temperatura ambiente (18 a 25 °C) durante 1 min Centrifugar por 1 min a 11.000 G. A recuperação de fragmentos maiores (> 1000 bp) pode ser aumentada em etapas múltiplas de eluição com mais tampão aquecido a 70 °C e incubação durante 5 min.

3.9.7.2 Purificação de DNA por precipitação

O produto de PCR foi purificado por co-precipitação com glicogênio usando-se um protocolo do kit Topo Shotgun (Thermo Fisher Scientific Inc., Waltham, MA, EUA).

Para 300 μ L de amostra adicione 36 μ L de acetato de sódio 3 M pH 5,2, 1,8 μ L de glicogênio 20 mg/mL e 600 μ L de etanol 100% gelado. Misture e incube a -20 °C por 15 min. Centrifugue a 12.000 G por 15 min a 4 °C e descarte o sobrenadante. Adicione 300 μ L de etanol 80% e centrifugue a 12.000 G por 5 min a 4 °C. Descarte o sobrenadante, seque a temperatura ambiente e elua em água ou TE.

3.9.7.3 Purificação de banda excisada de gel agarose

As bandas de interesse foram cortadas com uma lâmina de bisturi e depois submetidas à purificação usando-se um protocolo desenvolvido no laboratório. Os materiais necessários são 2 tubos de microcentrifuga de 1,5 ml, lâã acrílica (perlon) estéril e agulha hipodérmica. A preparação da coluna de centrifugação de lâã acrílica é feita utilizando uma agulha hipodérmica limpa. Fazer um furo na extremidade de um dos tubos a partir do interior e destacar a tampa cortando a alça. Pegue um pequeno pedaço de lâã acrílica e enrole-o em uma bola entre as mãos. Coloque a bola de lâã acrílica no fundo do tubo perfurado. Pode ser necessário utilizar uma ponta de pipeta de 1 ml para empurrar a lâã acrílica até o fundo do tubo. Não corte a lâã acrílica, pois pequenas fibras podem contaminar o material. Engaste o tubo com a lâã acrílica sobre o outro tubo de microcentrifuga. Coloque o fragmento de agarose na coluna de centrifugação. Centrifugar o tubo a aproximadamente 2900 G durante não mais do que 45 s para eluir o DNA. Centrifugações mais longas coeluem substâncias que são inibidoras de reações enzimáticas que podem se seguir. Em seguida o eluído é submetido ao protocolo de coprecipitação com glicogênio.

3.9.8 Clonagem de produto de PCR com enzima Pfu em vetor pGEM-T Easy

O purificado da banda extraída do gel de agarose foi então submetido à adição de cauda de adenosina e novamente purificado. Esse passo foi feito para possibilitar a entrada no protocolo de clonagem do kit pGEM-T Easy (Promega

Biotecnologia do Brasil, São Paulo, SP, Brasil) que se vale de uma adenosina não pareada na porção 3' do amplificado para ligação a uma timina não pareada no vetor linearizado incluso no kit de clonagem. A reação de ligação foi feita por 18 h a 4 °C.

3.9.9 Testes de confirmação de clonagem

As colônias da transformação recuperadas de cor branca, no caso de DH10B transformada com pGEM, eram então submetidas a uma preparação de DNA plasmidial em pequena escala. Esse DNA foi então submetido à eletroforese em gel de agarose com e sem o tratamento com a endonuclease EcoRI, que excisa o inserto com cortes em sítios que cercam o sítio de clonagem, para confirmação da presença do inserto. O inóculo de clones confirmados para a presença do inserto era então preservado com a adição de um volume de glicerol 50% e armazenado em tubo de microcentrífuga com tampa de rosca em freezer a -80 °C. O DNA era então submetido a sequenciamento pela técnica de Sanger para confirmação da sequência do inserto clonado. Em casos de confirmação por tamanho esperado de banda em gel de agarose, uma parte reservada do inóculo bacteriano foi misturada na proporção 1:1 com uma solução de glicerol 50% e armazenada em um freezer a -80 °C.

3.9.9.1 Teste de confirmação para pJET1.2

A confirmação da clonagem em pJET1.2 (Thermo Fisher Scientific Inc., Waltham, MA, EUA) foi feita com uma PCR utilizando-se dos mesmos iniciadores desenhados no início desse processo sobre DNA preparado a partir das colônias geradas na transformação de *E.coli* DH10B com o produto da ligação.

3.9.9.2 Teste de confirmação para pET-44b

A confirmação da clonagem foi feita da mesma forma que para a clonagem em pJET1.2.

Na ordem da figura: Vetor pGEM-T Easy utilizado na clonagem inicial a partir da amplificação a partir do material genômico; Vetor pJET1.2 usado para clonagem em ponta cega do produto de PCR a partir dos clones em pGEM-T Easy para posterior subclonagem; Vetor de expressão pET-44b com sítios *PacI* e *NdeI* para onde a clonagem foi dirigida e o trecho do polipeptídio *Nus Tag* retirado.

O DNA produzido na reação de preparação em pequena escala foi usado para uma reação de PCR usando o tampão e o preparado BigDye (Thermo Fisher Scientific Inc., Waltham, MA, EUA). As reações foram feitas usando-se 2 µL de água deionizada, 2 µL de iniciadores, 2 µL do BigDye, 2 µL do tampão e 2 µL de DNA. O programa de termociclador utilizado foi 95 °C por 90 s e 35 ciclos de 95 °C por 18 s, 55 °C por 12 s e 60 °C por 4 min. Essa reação foi precipitada com o protocolo de coprecipitação com glicogênio e ressuspensa em formamida para então ser submetida aos sequenciadores automatizados ABI3130xl ou ABI3730 (Thermo Fisher Scientific Inc., Waltham, MA, EUA). Esses sequenciadores recebem placas de 96 poços e fazem a eletroforese das reações em microcapilar com a leitura por CCD

da emissão sob excitação de laser dos quatro diferentes dideoxinucleosídeos fluorescentes, correspondentes à adenina, timina, guanina e citosina. Essa leitura produz um cromatograma com a variação temporal da captação dos quatro comprimentos de onda dos nucleosídeos. Esse cromatograma é então submetido a um script desenvolvido no GaTE Lab (run_GaTE_phred_crossmatch) para distinção e escolha de bases pelo software phred e em seguida a filtragem de sequências do vetor de clonagem com auxílio do software Crossmatch.

3.9.11 Subclonagem sítio dirigida em pET-44b

As etapas de todo o processo de subclonagem e expressão estão apresentadas na Figura 13. De modo breve: PCR dos alvos clonados inicialmente em pGEM-T Easy, purificação, clonagem ponta cega no vetor pJET1.2, teste de confirmação para pJET1.2 por PCR, digestão de DNA de clones no vetor pJET1.2, purificação, ligação com T4 ligase do fragmento purificado de pJET1.2 para clonagem por ligação coesiva em pET-44b e posterior transformação em BL21(DE3).

3.9.12 Desenho de primers com adaptadores para subclonagem sítio dirigida em pET-44b

Os iniciadores 5' para subclonagem do vetor pGEM-T Easy para o vetor pET-44b foram desenhados de forma a ter a mesma fase de leitura do trecho ATG do sítio de restrição NdeI que está à distância correta do sítio de ligação a ribossomo incluso no vetor. Além disso, os iniciadores 3' contêm um sítio de restrição PacI de forma a excluir a etiqueta de nusA inclusa no vetor. Foi permitida uma variação de 14 a 25 bp de qualquer um dos dois oligonucleotídeos iniciadores, no trecho específico de amplificação da fita molde de forma a manter compatível a temperatura de melting entre os oligonucleotídeos e foi incluído um trecho de 6 nucleotídeos (TATATA) antes do sítio de restrição de acordo com indicação do fabricante da enzima de restrição para digestão do produto de PCR. A relação dos oligonucleotídeos encontra-se na Tabela 21.

O vetor pET-44 foi usado no intuito de otimizar o processo de expressão, pelo que é reportado na literatura (Hu et al., 2007) e também porque não houve evidência

da produção da proteína em pGEM-T Easy. Há diferenças entre as clonagens no vetor pGEM-T Easy e pET-44b que propiciam um processo mais eficiente e controlável de transcrição e tradução do gene clonado. Elas são: a presença de um repressor lacI, um terminador de transcrição, um sítio de ligação de ribossomo à distância apropriada do início de tradução do gene clonado e a expressão sem a formação de fusão quimérica, já que no pGEM o gene é clonado internamente ao gene lacZ.

3.9.13 PCR sobre material de clones pGEM

Foi feita uma preparação de pequena escala do DNA dos clones em pGEM-T Easy gerados anteriormente. Esse DNA foi usado como molde para a reação de PCR com os iniciadores desenhados para clonagem no vetor pET-44b.

3.9.14 Clonagem ponta cega no vetor pJET1.2

Procedimento de clonagem executado de acordo com o indicado pelo fabricante do kit. De acordo com os parâmetros do protocolo foi usado entre 50 e 100 ng de DNA produto de PCR, incubação a 22 °C por 15 a 30 min em termociclador e extração da reação de ligação com clorofórmio para transformação de células competentes de *E.coli* DH10B para eletroporação.

3.9.15 Digestão de DNA de clones no vetor pJET1.2 e ligação no pET-44b

A digestão do DNA extraído dos clones positivos para subclonagem em pET-44b foi feita com enzimas (New England Biolabs, Ipswich, MA, EUA). Foram utilizadas 1U de PaeI e 10U de NdeI em tampão NEB 4 com todo DNA de uma alíquota da preparação em pequena escala. As reações foram incubadas a 37 °C por 18 a 21 h.

As digestões foram submetidas à eletroforese em gel de agarose 0,5%, as bandas de interesse cortadas com uma lâmina de bisturi e depois submetidas à purificação. Foram cortadas as bandas com tamanho previsto para o vetor pET-44b, sem o Nus-Tag, e aos genes a serem subclonados.

Os purificados foram então ligados com a enzima T4 ligase (Thermo Fisher Scientific Inc., Waltham, MA, EUA). O procedimento foi feito de acordo com as instruções do fabricante. Foi usado 50 ng de DNA de pET-44b e uma proporção de 3:1 entre gene e pET-44b. Foi usado tempo de reação de 1 h a 22 °C e feita inativação da reação a 65 °C por 10 min. Foi utilizado 4 µl da reação para transformação e células eletrocompetentes de E.coli DH10B.

3.9.16 Indução de expressão de genes clonados na cepa de E.coli BL21 (DE3)

O préinóculo do meio de indução foi feito crescido em 2 ml nas mesmas condições descritas anteriormente, porém por não mais que 6 a 8 h. A indução foi feita por 6 h com diferentes concentrações de IPTG em volume final de 10 ml de meio LB.

Os plasmídeos com sequência clonada confirmada foram então usados para transformação de células de E. coli cepa BL21 (DE3) quimio competentes. Colônias confirmadas por visualização de miniprep em gel de agarose foram armazenadas em glicerol a -80 °C. Novas placas de LB ágar Amp foram estriadas com o armazenado em glicerol e dentro de no máximo de 20 h uma colônia formada era transferida para meio líquido para crescimento por 16 h a ser usado como préinóculo. 5 mL de meio líquido LB Amp foi então inoculado com 50 µL do préinóculo colocado em estufa agitada para crescimento após 4 h adicionado IPTG para concentração final de 0,1 mM e mais 2 h de crescimento sob agitação a 37 °C. O inóculo foi então centrifugado a 5000 G por 1 min e separados o sobrenadante e o pellet bacteriano.

3.9.17 SDS-PAGE

Os ensaios de eletroforese em gel denaturante de poliacrilamida e SDS foram feitos de acordo com o protocolo que se segue.

Lavar e secar as placas de vidro. Montar o aparato de corrida com os espaçadores de espessura adequada. Colocar água para verificar se há vazamento. Verter a água na pia e secar com papel filtro. Fazer o gel de separação em um tubo de polipropileno pequeno. Colocar o TEMED e APS por último. Com uma pipeta, colocar o gel entre as placas e em seguida colocar uma finíssima camada de etanol

e aguarda pelo menos 15 min. Verter e colocar papel filtro para retirar a água e etanol e lavar com tampão lower. Colocar o pente de mesma espessura dos espaçadores, preenchendo se houver retração. Lavar os poços com tampão de corrida para retirar a acrilamida excedente. Mergulhar o gel em tampão de corrida 1x. Separar 5 µL de marcador de peso. Submeter o pellet bacteriano obtido no processo de indução de produção da proteína heteróloga a lise a 100 °C em tampão de carregamento de SDS-PAGE por 3 min e imediatamente carregar no gel. Correr a 50 V por 3 a 3,5 h. Parar a corrida imediatamente antes ou depois que o bromofenol blue comece a sair do gel. Coloque o gel de poliacrilamida num recipiente contendo solução de Coomassie Blue suficiente para cobri-lo. Deixe agitando por 30 min a temperatura ambiente. Retire a solução de Coomassie Blue e adicione a solução descorante A por 1 h com um pedaço de papel absorvente. Troque pela solução B até que as bandas do gel fiquem bem nítidas. Retire a solução descorante e adicione água para hidratar o gel.

Soluções:

- Upper buffer pH 8,8 4x: 91 g de Tris base, q.s.p 500 mL de água deionizada, filtrar em filtro de 0,45 µm, 2 g de SDS e estocar a 4 °C.
- Lower Buffer pH 6,8 4x: 30 g de Tris base, q.s.p 500 mL de água deionizada, filtrar em filtro de 0,45 µm, 2 g de SDS e estocar a 4 °C.
- Tampão de corrida 10x: 30,3 g Tris base, 144 g de glicina, q.s.p 1 l de água deionizada, ajustar pH para 8,5, 10 g de SDS, agitar vagarosamente e estocar a 4 °C. Para uso diluir 10 vezes em água deionizada.
- Tampão da amostra 6x: 7 mL de upper buffer, 1 g de SDS, 3 mL de glicerol, 0,4 mL de 2-mercaptoetanol, 1,2 mg de azul de bromofenol e q.s.p 10 mL de água deionizada.
- Solução de Coomassie Blue: 50% de metanol, 10% de ácido acético glacial, 40% de água destilada, 0,05% (p/v) de Coomassie Blue. Dissolva o corante em metanol antes de adicionar o ácido acético e água. A solução pode ser estocada por até 6 meses à temperatura ambiente. Se for observada a formação de precipitado, filtre com papel filtro para obter uma solução homogênea.
- Solução descorante A: 227 mL de metanol, 37,5 mL de ácido acético, 235,5 mL de água.

- Solução descorante B: 25 mL de metanol, 36,5 mL de ácido acético, 437,5 mL de água.
- Bis/Acrilamida: 150 g de acrilamida, 4 g de bis-acrilamida, q.s.p 500 mL de água deionizada, filtrar com papel filtro e estocar a 4 °C no escuro.
- APS 10%: Dissolver em água a 10% (p/v) e fazer alíquotas de 35 uL. Manter a 20 °C por 3 meses.
- Gel de empilhamento de 5% (2 mL para 1,5 mm): 1140 µL de água, 500 µL de Upper Buffer 4x, 340 µL de Bis/Acrilamida, 20 µL de APS 10%, 2,7 µL de TEMED.
- Gel de separação de 12% (10 mL para 1,5 mm): 3,5 mL de água, 2,5 mL de Lower Buffer 4x, 4 mL de Bis/Acrilamida, 33 µL de APS 10%, 17 µL de TEMED.

3.9.17.1 Lise bacteriana para ensaios enzimáticos

Os pellets de cultura bacteriana produzidos na indução com IPTG foram lisados para os testes enzimáticos que se seguiram. Para isso foi usado o produto BugBuster (Merck S.A., Rio de Janeiro, R.J., Brasil). Foi usada uma mistura de 2,5 mL de BugBuster, 1 µL da DNase Benzonase, 0,6 mg de lisozima, 0,5 mg do inibidor de serina protease fluoreto de fenilmetilsulfonila (PMSF) e 1 mg do antioxidante polivinilpolipirrolidona (PVPP) para cada 50 mL de cultura líquida bacteriana. A mistura é adicionada ao pellet e mantida à temperatura ambiente em agitador orbital por 30 min.

3.9.18 Testes enzimáticos

3.9.18.1 Atividade endoglunásica sobre 4-nitrofenil-β-D-sacarídeos

O ensaio para verificar atividades de hidrólise sobre ligações internas à cadeia, distantes das pontas com potencial redutor, foi feita através de reações sobre mono ou dissacarídeos ligados a uma nitrofenila (PNP) (Figura 3) pelo

carbono oxidativo do açúcar. Com o rompimento da ligação a nitrofenila se torna amarela e passa a ter uma forte absorbância em 405 nm.

Figura 3 - 4-nitrofenil- β -D-glicopiranosídeo. A seta indica a ligação clivada por uma endoglucanase.

O ensaio foi feito em placas de 96 poços de fundo chato transparente. Foi feita a mistura de 30 μ L de cada PNP com 2,4 mL de buffer Tris-HCl pH 6,8 e dispensado em uma coluna de duas placas para cada PNP resultando em 150 μ L para cada poço. 10 μ L de lisado de cada clone foi adicionado a cada poço de uma linha de cada placa e incubado a 37 °C por 2 h. Foi então adicionado 300 μ L de uma solução de carbonato de sódio 1 M e feita a medição em espectrofotômetro de leitura de placa. Foram feitas 9 replicas e os dados foram submetidos a um teste de Tukey para distinguir grupos de ensaios com absorbâncias estatisticamente distintas.

3.9.18.2 Atividade sinérgica sobre hemiceluloses e extratos de parede celular

Para verificar a presença de atividade exoglucanásica nos lisados bacterianos foi feito um ensaio em cromatografia de troca iônica de alto desempenho com detecção amperométrica pulsada (HPAEC-PAD) do perfil de oligossacarídeos produzidos na digestão de diferentes polissacarídeos. O protocolo que se segue foi modificado da dissertação de mestrado de Encarnação (2012).

Digestão: solução de polissacarídeo a 1% ou 0,1% (p/v) em tampão acetato de amônio 50 mM pH 5.0 acrescido de 20% da mistura de lisados dos clones bacterianos. Incubar a mistura a 30 °C por 24 h e em seguida manter a 100 °C por 5 min. Centrifugar à 2700 G, por 5 min à temperatura ambiente e coletar o sobrenadante. Ao sobrenadante adicionar 3 volumes de etanol absoluto, para a precipitação dos polissacarídeos não digeridos e centrifugar à 4000 G por 3 min.

Coletar o sobrenadante e secar em concentrador de amostras sem aquecimento. Ressuspender em 1 mL de água deionizada para análise por HPAEC-PAD.

Cromatografia: performedo no modelo ICS-3000 (Dionex®), utilizando coluna CarboPac PA-100, eluente de base utilizado 88 mM de NaOH, eluente de arraste acetato de sódio 200 mM em hidróxido de sódio 88 mM. Um gradiente da solução de acetato de sódio 200 mM em hidróxido de sódio 88 mM foi aplicado à amostra visando separar os oligossacarídeos antes de sua detecção pelo detector de pulso amperométrico.

Foram feitos ensaios com os seguintes polissacarídeos e extratos:

1. Arabinoxilano de trigo
2. Fração Oxalato de Cana-De-Açúcar
3. Arabinogalactano 15:85
4. Extrato em 4 M hidróxido de sódio de raiz de Cana-De-Açúcar
5. Xiloglucano de Jatobá
6. Raiz de Cana-De-Açúcar
7. Resíduo da extração de raiz de Cana-De-Açúcar
8. Xilano de Bétula
9. Arabinoxilano de Centeio
10. Xilano de Faia
11. Extrato em 1 M hidróxido de sódio de raiz de Cana-De-Açúcar
12. Liqueenano de musgo da Islândia
13. Arabinoxilano insolúvel de Trigo
14. Manano de Café
15. Beta-glucano de baixa viscosidade de cevada
16. Arabinoxilano desramificado
17. Arabinoxilano de Trigo
18. Galactomanano de *Sesbania virgata*

4 RESULTADOS E DISCUSSÃO

4.1 Enumeração dos genes cazy e das PCWDE

Ao longo do trabalho procuramos refinar a série de técnicas de enumeração, sempre tentando corrigir uma questão encontrada na etapa anterior. A 1ª técnica se restringiu ao Comparativo de *Xylella*, mas selecionou genes não-PCWDE e contava com um número restrito de genomas. A 2ª técnica encontrou mais genes porém encontrou heterogeneidade entre as anotações, causando incerteza sobre os genes levantados; recuperou mais hidrolases porém ainda contendo não-PCWDE e subestimou a diversidade de PCWDE. A incerteza causada pelas anotações e baixa diversidade encontrada foi a motivação para a 3ª estratégia. Essa estratégia buscou uma abordagem mais semântica, formando dicionários de termos ligados a PCWDE e PCWPS, usados para selecionar modelos HMM (PFams). Esses modelos permitiram uma maior abrangência dos genes levantados com maior confiabilidade de distinção entre ação sobre PCWPS e outros polissacarídeos. No entanto, muitos dos PFams continham informação desatualizada ou permissiva demais, deixando de recuperar genes relevantes e ainda trazendo enzimas não-PCWDE. A 4ª estratégia, final, uniu o processo de definição de substratos com o de enumeração de genes. Com a geração pelo projeto dbCAN de modelos HMM baseados nas sequências curadas pelo CAZy adquiriu-se uma precisão maior no processo de enumeração. A catalogação no projeto das classes de enzimas cazy, permitiu a recuperação de números EC que foram confrontados com os EC recuperados do banco da IUBMB. Esses EC foram selecionados baseando-se na informação retirada da literatura relativa à PCWPS. Essa estratégia conseguiu reduzir o número de genes fracamente relacionados à PCWDE levantados na estratégia anterior e contemplou uma série de genes que haviam sido subestimados em todas as outras estratégias.

4.1.1 Genes não-PCWDE e número restrito de genomas

Ao iniciar o processo de seleção de genes de PCWDE partimos do conjunto de clusters do projeto de genômica comparativa de *Xylella fastidiosa*. Nele procuramos sobre a descrição da anotação de genes de acordo com a coleção de termos relativos à PCWDE, encontrados na bibliografia levantada à época

(*Arabinosidase*, *Beta-glucosidase*, *Cellobiosidase*, *Cellulase*, *Pectate lyase*, *Pectin*, *Polygalacturonase*, *Rhamnogalacturon*, *Xylan*, *Xylosidase*). Essa procura, no entanto, não contemplou alguns genes envolvidos no processo de degradação por estarem anotados como proteínas hipotéticas ou com anotação incompleta ou defasada. Também incluiu genes codificantes para elementos regulatórios e outros que evidentemente não eram ligados a PCWDE devido a um artefato da técnica BBH usada, ligados a elementos de transposição fusionados aos genes de PCWDE. A seleção inicial do Comparativo deixou também de incluir outros genomas de *Xanthomonadaceae*, não disponíveis à época do fechamento do projeto.

4.1.2 Heterogeneidade entre anotações, hidrolases não-PCWDE e baixo diversidade de PCWDE

Uma grande heterogeneidade entre anotações e termos utilizados para descrição das anotações dos genes levantados na etapa anterior foi averiguada, sendo que anotações de genes em um mesmo cluster do Comparativo não eram uniformes. Isso é explicado pela diferença de até 7 anos entre os projetos de anotação e, por consequência, da informação disponível nos bancos de dados à época de cada um. Sem ter de recorrer a uma uniformização das anotações, procuramos recuperar mais genes PCWDE através de clusterização BLASTclust.

Procuramos também enriquecer a diversidade de genomas de *Xanthomonadaceae*, ao todo 19, não contemplados no banco do projeto Comparativo de *Xylella* recuperando do NCBI, banco onde a maior parte dos genomas parciais e completos são depositados, e JCVI, onde alguns dos genomas são depositados antes do NCBI.

Usando-se as sequências de PCWDE levantadas na estratégia inicial em conjunto com os proteômas de todos os genomas de *Xanthomonadaceae* encontrados à época foi possível enriquecer a enumeração. No entanto, algumas das proteínas recuperadas exibiam indícios de ação sobre polissacarídeos não-PCWPS como amido, quitina e polissacarídeos da parede celular bacteriana. Em oposição, algumas das PCWDE que se esperaria encontrar devido à complexidade da parede celular vegetal estavam ausentes dessa listagem, como algumas hidrolases de cadeia lateral de mananos, xilanos, xiloglucanos (hemiceluloses), arabinanos e galacturonanos em geral (pectinas).

4.1.3 Dicionários, PFs e aumento da confiabilidade

A capacidade de busca dessa classe de enzimas exclusivamente com base na descrição da anotação gênica se mostrou insuficiente em elencar todos os genes do aparato de PCWDE. Frente uma grande variedade de combinações de monômeros e ligações possíveis dentre os PCWPS e suas PCWDE correspondentes (Tabela 3) e a capacidade de maceração de tecido já averiguada (Tabela 2) para alguns desses patógenos, desenvolvemos uma abordagem mais geral. Desenvolvemos um método de procura baseando-se nos termos encontrados na literatura e em modelos de evolução molecular da conservação da estrutura terciária das enzimas que hidrolisam carboidratos.

A construção de uma lista de 544 palavras e posterior filtragem de termos produziram uma lista de 15 palavras-chave e expressões regulares (da linguagem de programação Perl) para seleção de HMMs/PFs do projeto Pfam e de genes sem PFs detectáveis. A lista inicial usada sobre as descrições dos genes no banco do Comparativo já contava com 10 palavras-chave, também incluídas. O processo de seleção automática de PFs produziu uma lista de 92 clãs e 693 PFs e sua filtragem manual resultou numa lista curada de 104 PFs usados para selecionar ORFs. Da lista de vocabulário encontrado na descrição dos genes selecionados com os PFs resultaram 78 termos e expressões regulares utilizados na seleção de genes sem PFs detectáveis.

Foram encontrados 1057 genes, dentre 72101, contendo 88 arquiteturas de PFs. Além desses, 27 genes foram selecionados que não acusaram nenhum PF dentro dos parâmetros de procura, mas foram recuperados com o auxílio do dicionário de termos usado sobre as descrições de anotação. No total foram selecionados 1084 genes.

Apesar da grande riqueza de informação disponibilizada pelo Pfam alguns dos domínios não estavam lá vinculados, apesar de constarem no banco do CAZy. Ademais, as bases de sequências para construção dos PFs de PCWDE encontravam-se bastante desatualizadas frente às listas de genes do CAZy. Consequentemente, o uso do dbCAN se seguiu como consequência natural ao se tomar conhecimento dessa base de HMMs.

4.1.3.1 Domínios RHS

Nessa análise foram também encontrados genes considerados ligados ao metabolismo de carboidratos porém não a PCWDE. Em particular numerosos genes compostos de motivos repetitivos RHS (*recombination hot spots*) (Lin et al., 1984) que na base do Pfam eram indicados como possíveis motivos de ligação a carboidrato. Foram encontrados na literatura resultados indicando o papel das proteínas RHS_repeat na adesão e formação de biofilme (Pradhan et al., 2012) ou como toxina (Zhang et al., 2012), reduzindo a probabilidade de essas proteínas serem ligadas à hidrólise de PCWPS.

4.1.4 Listagem de cazys ligados a PCWDE

O projeto CAZy se propõe a listar genes e classificar domínios proteicos de enzimas que agem sobre carboidratos (cazy) em 6 classes contendo subclasses numeradas. As classes são: módulos de ligação a carboidrato (CBM), glicosil hidrolases (GH), glicosil transferases (GT), carboxil esterases (CE), pectato liases (PL) e domínios de atividades redox auxiliares (AA). O banco dbCAN (Yin et al., 2011), usado nessa etapa em lugar do Pfam, foi gerado a partir da coleta das sequências de referência do CAZy e geração de modelos HMM com o software HMMER3. O CAZy/dbCAN possui maior especificidade e é mais rico em informação do que os domínios depositados no Pfam, pois neste faltam vários domínios e sequências de referência de várias das classes do banco do CAZy. Além disso, a informação atrelada aos modelos na página do projeto é muito mais curada e atualizada do que aquilo foi transposto para o banco de descrições de domínios do Pfam e do Interpro. Essa riqueza permitiu correlacionar os números EC atrelados a um membro de uma classe com os números levantados do banco da IUBMB. A quarta estratégia trouxe um maior número de genes elencados e o uso da informação que correlaciona polissacarídeos, ECs e cazys, levantada da literatura, permitiu uma maior sensibilidade para diferenciar PCWDE de outras cazys.

O banco de atividades enzimáticas da IUBMB foi submetido à procura de termos encontrados na literatura de PCWPS (tabelas Tabela 4, Tabela 5 e Tabela 6), o que gerou uma lista de 86 números EC. A intersecção entre o conjunto dos números EC levantados e o daqueles descritos nos domínios do CAZy resultou em

120 domínios prováveis de estarem envolvidos com PCWDE (Tabela 7), sendo que 90 ocorrem em mais de um tipo de substrato, em concordância com a inespecificidade ou promiscuidade observada para essas enzimas, e 30 em apenas um tipo de substrato. Com o incremento de mais 17 novos genomas, totalizando 36 genomas (Tabela 8), foram analisadas um total de 137.123 proteínas, das quais foram levantadas cerca de 5620 cazys e das quais 3047 são possíveis PCWDE (tabelas Tabela 9 e Tabela 10).

Tabela 4 - Coleção de polissacarídeos da parede celular vegetal, suas ocorrências em monocotiledôneas e dicotiledôneas.

Nome	Classe	táxons
Celulose	Celulose	Dicot. e monocot.
Manano, galactoglucomanano, galactomanano, GGM, manano linear, glucomanano	Hemicelulose	Dicot. e monocot.
Glucuronoarabinoxilano, xilano, GAX, Glucuronoxilano	Hemicelulose	Principalmente monocot., pouco em dicot.
Fucogalactoxiloglucano, xiloglucano, XG	Hemicelulose	Principalmente dicot., paredes primárias de monocot.
β-(1→3,1→4)-glucanos	Hemicelulose	Paredes primárias de monocot.
homogalacturonano, HG	Pectina	
ramnogalacturonano-I	Pectina	
Galacturonanos substituídos, SG, ramnogalacturonano-II, xilogalacturonano, XGA, apiogalacturonano	Pectina	Principalmente dicot., pouco em paredes primárias de monocot
Arabinano	Pectina	
Arabinogalactano I	Pectina	
Arabinogalactano II	Pectina	

Tabela 5 - Coleção de polissacarídeos da parede celular vegetal, suas subunidades e ligações componentes.

Nome	Ligações e monômeros nas respectivas cadeias
Celulose	Principal: $\beta(1-4)$ -glicose
Manano, galactoglucomanano, galactomanano, GGM, manano linear, glucomanano	Principal: $\beta(1,4)$ -manose/glicose; Secundárias: $\alpha(1,6)$ -galactose e $\alpha(1,2)/\alpha(1,3)$ acetil manose
Glucuronoarabinoxilano, xilano, GAX, Glucuronoxilano	Principal: $\beta(1,4)$ -xilose. Secundárias: $\alpha(1,2)$ -glucurônico, $\alpha(1,2)/\alpha(1,3)$ -arabinose, arabinose esterificada com ferrúlico e p-cumárico, $\alpha(1,4)$ -metilglucurônico, acetilações em O-3 e O-2 de xilose.
Fucogalactoxiloglucano, xiloglucano, XG	Principal: $\beta(1,4)$ -glicose. Secundárias: $\alpha(1,6)$ -xilose, xilose derivada com galactose, fucose e arabinose
$\beta(1\rightarrow3,1\rightarrow4)$-glucanos	Principal: $\beta(1-3)/\beta(1-4)$ -glicose
homogalacturonano, HG	Principal: $\alpha(1,4)$ -galacturônico. Secundárias: 4 diferentes cadeias (hetero-oligoméricas) com resíduos peculiares como apiose, acérico, dihidroxiacetona e 3-deoximanoctulosônico, carboxilas metilesterificadas e acetilações em O-3 e O-2.
ramnogalacturonano-I	Principal: repetições do dissacarídeo $\alpha(1,4)$ - galactouronil- $\alpha(1,2)$ -L-ramnosil. Secundárias: ramnose substituídas em C-4 com cadeias de $\alpha(1,4)$ -galactosil, arabinogalactano I, arabinano (50 resíduos ou mais), α -arabinosil, β -galactosil, β -glycosil, α -fucosil, $\beta(1,4)$ -glucuronosil, ferrulato, cumarato, 4-O-metil- β -glucuronosil, acetilação em C-2 e C-3 de galA da principal
Galacturonanos substituídos, SG, ramnogalacturonano-II, xilogalacturonano, XGA, apiogalacturonano	Principal: $\alpha(1,4)$ -galactosilurônico. Secundárias: 4 cadeias diferentes ligadas a C-2 e C-3 da principal contendo 12 resíduos diferentes, 2-O-metil-l-fucose, 2-O-metil-xilose, apiose, ácido acérico, 3-deoximanoctulosônico e dihidroxiactenoa; xilogalacturonanos tem resíduos β -xil ligados a C-3 da principal; resíduos da principal de XGA podem ser metil-esterificados como HG; apiogalacturonanos com cadeias laterais substituídas com β -apiosil e β -apiosil-(1,3)- β -apiosil em C-2 e C-3 da principal; resíduos ramnosil estão presentes nas cadeias laterais de ramnogalacturonano-II, com ligações cruzadas de éster de borato
Arabinano	Principal: $\alpha(1,5)$ -arabinosil. Secundárias: substituições com $\alpha(1,2)/\alpha(1,3)$ -arabinosil e/ou $\alpha(1,3)$ -arabinosil- $\alpha(1,3)$ -arabinosil
Arabinogalactano I	Principal: $\beta(1,4)$ -galactil. Secundárias: $\alpha(1,3)$ -arabinosil
Arabinogalactano II	Principal: $\beta(1,3)$ -galactil. Secundárias: curtas de $\beta(1,6)$ -arabinosil-[(1,6)-galactil-]n e $\alpha(1,3)$ -arabinosil.

Tabela 6 - Coleção de polissacarídeos da parede celular vegetal, números EC das atividades enzimáticas associadas e domínios correspondentes das classes do CAZy. (continua)

Nome	ECs	CAZy
Celulose	3.1.6.7, 3.2.1.176, 3.2.1.21, 3.2.1.4, 3.2.1.74, 3.2.1.86, 3.2.1.91	CBM1, CBM10, CBM11, CBM12, CBM17, CBM18, CBM2, CBM28, CBM3, CBM30, CBM32, CBM35, CBM4, CBM44, CBM46, CBM49, CBM5, CBM6, CBM63, CBM65, CBM8, CBMNC, CE2, GH1, GH10, GH116, GH12, GH124, GH18, GH19, GH26, GH3, GH30, GH4, GH44, GH45, GH48, GH5, GH51, GH6, GH7, GH74, GH8, GH9, GHNC
Manano, galactoglucomanano, galactomanano, GGM, manano linear, glucomanano	3.2.1.100, 3.2.1.101, 3.2.1.163, 3.2.1.21, 3.2.1.22, 3.2.1.24, 3.2.1.25, 3.2.1.4, 3.2.1.74, 3.2.1.78, 3.2.1.86, 4.2.2.3	CBM1, CBM10, CBM11, CBM12, CBM13, CBM16, CBM17, CBM18, CBM2, CBM23, CBM27, CBM28, CBM3, CBM30, CBM32, CBM35, CBM4, CBM44, CBM46, CBM49, CBM5, CBM59, CBM6, CBM63, CBM65, CBM8, CBMNC, CE2, GH1, GH10, GH110, GH113, GH116, GH12, GH124, GH18, GH19, GH2, GH26, GH27, GH3, GH30, GH32, GH36, GH38, GH4, GH44, GH45, GH48, GH5, GH51, GH57, GH6, GH7, GH74, GH76, GH8, GH9, GH92, GH97, GHNC, PL14, PL15, PL17, PL18, PL5, PL6, PL7, PLNC
Glucuronoarabinoxilano, xilano, GAX, Glucuronoxilano	3.1.1.73, 3.2.1.131, 3.2.1.136, 3.2.1.156, 3.2.1.177, 3.2.1.31, 3.2.1.32, 3.2.1.37, 3.2.1.55, 3.2.1.8	CBM1, CBM10, CBM13, CBM15, CBM2, CBM22, CBM3, CBM31, CBM35, CBM36, CBM37, CBM4, CBM42, CBM48, CBM49, CBM5, CBM59, CBM6, CBM60, CBM64, CBM9, CBMNC, CE1, CE10, CE3, CE4, CENC, GH1, GH10, GH11, GH115, GH116, GH12, GH120, GH16, GH2, GH26, GH3, GH30, GH31, GH39, GH43, GH44, GH5, GH51, GH52, GH54, GH62, GH67, GH78, GH79, GH8, GH9, GHNC
Fucogalactoxiloglucono, xiloglucano, XG	3.2.1.21, 3.2.1.22, 3.2.1.4, 3.2.1.55, 3.2.1.74, 3.2.1.86, 2.4.1.207, 3.2.1.120, 3.2.1.121, 3.2.1.150, 3.2.1.151, 3.2.1.155, 3.2.1.72	CBM1, CBM10, CBM11, CBM12, CBM13, CBM17, CBM18, CBM2, CBM22, CBM28, CBM3, CBM30, CBM32, CBM35, CBM36, CBM4, CBM42, CBM44, CBM46, CBM49, CBM5, CBM6, CBM63, CBM65, CBM8, CBMNC, CE2, GH1, GH10, GH110, GH116, GH12, GH124, GH16, GH18, GH19, GH26, GH27, GH3, GH30, GH32, GH36, GH4, GH43, GH44, GH45, GH48, GH5, GH51, GH54, GH57, GH6, GH62, GH7, GH74, GH8, GH9, GH97, GHNC
β-(1$\rightarrow$3,1$\rightarrow$4)-glucanos	3.2.1.4, 3.2.1.6	CBM1, CBM10, CBM11, CBM12, CBM13, CBM17, CBM18, CBM2, CBM28, CBM3, CBM30, CBM32, CBM35, CBM4, CBM44, CBM46, CBM49, CBM5, CBM54, CBM6, CBM63, CBM65, CBM8, CBMNC, CE2, GH10, GH12, GH124, GH16, GH18, GH19, GH26, GH44, GH45, GH48, GH5, GH51, GH6, GH7, GH74, GH8, GH9, GHNC

Tabela 6 - Coleção de polissacarídeos da parede celular vegetal, números EC das atividades enzimáticas associadas e domínios correspondentes das classes do CAZy. (conclusão)

Nome	ECs	CAZy
homogalacturonano, HG	3.1.1.11, 3.1.1.72, 3.2.1.146, 3.2.1.15, 3.2.1.55, 3.2.1.67, 3.2.1.82, 4.2.2.10, 4.2.2.2, 4.2.2.22, 4.2.2.6, 4.2.2.9	CBM1, CBM13, CBM2, CBM22, CBM3, CBM35, CBM36, CBM37, CBM42, CBM6, CBMNC, CE1, CE12, CE15, CE2, CE3, CE4, CE5, CE6, CE7, CE8, GH10, GH11, GH28, GH3, GH4, GH43, GH5, GH51, GH54, GH62, PL1, PL10, PL2, PL22, PL3, PL9
ramnogalacturonano-I	3.1.1.11, 3.1.1.72, 3.1.1.73, 3.1.1.86, 3.2.1.146, 3.2.1.15, 3.2.1.171, 3.2.1.172, 3.2.1.173, 3.2.1.174, 3.2.1.40, 3.2.1.43, 3.2.1.55, 3.2.1.67, 3.2.1.82, 4.2.2.10, 4.2.2.2, 4.2.2.22, 4.2.2.23, 4.2.2.24, 4.2.2.6, 4.2.2.9	CBM1, CBM13, CBM2, CBM22, CBM3, CBM35, CBM36, CBM37, CBM42, CBM6, CBM67, CBMNC, CE1, CE12, CE15, CE2, CE3, CE4, CE5, CE6, CE7, CE8, CENC, GH10, GH105, GH106, GH11, GH13, GH28, GH3, GH4, GH43, GH5, GH51, GH54, GH62, GH78, GHNC, PL1, PL10, PL2, PL22, PL3, PL9
Galacturonanos substituídos, SG, ramnogalacturonano-II, xilogalacturonano, XGA, apiogalacturonano	3.1.1.11, 3.1.1.72, 3.2.1.146, 3.2.1.15, 3.2.1.171, 3.2.1.172, 3.2.1.173, 3.2.1.174, 3.2.1.40, 3.2.1.43, 3.2.1.55, 3.2.1.67, 3.2.1.82, 4.2.2.10, 4.2.2.2, 4.2.2.22, 4.2.2.6, 4.2.2.9	CBM1, CBM13, CBM2, CBM22, CBM3, CBM35, CBM36, CBM37, CBM42, CBM6, CBM67, CBMNC, CE1, CE12, CE15, CE2, CE3, CE4, CE5, CE6, CE7, CE8, GH10, GH105, GH106, GH11, GH13, GH28, GH3, GH4, GH43, GH5, GH51, GH54, GH62, GH78, GHNC, PL1, PL10, PL2, PL22, PL3, PL9
Arabinano	3.2.1.55, 3.2.1.99	CBM1, CBM13, CBM2, CBM22, CBM3, CBM35, CBM36, CBM42, CBM6, CBMNC, GH10, GH3, GH43, GH51, GH54, GH62
Arabinogalactano I	3.2.1.145, 3.2.1.164, 3.2.1.23, 3.2.1.55, 3.2.1.89, 3.2.1.99	CBM1, CBM13, CBM2, CBM22, CBM3, CBM32, CBM35, CBM36, CBM42, CBM6, CBM61, CBMNC, GH1, GH10, GH2, GH3, GH30, GH35, GH42, GH43, GH5, GH50, GH51, GH53, GH54, GH62, GHNC
Arabinogalactano II	3.2.1.145, 3.2.1.164, 3.2.1.23, 3.2.1.55, 3.2.1.89, 3.2.1.99	CBM1, CBM13, CBM2, CBM22, CBM3, CBM32, CBM35, CBM36, CBM42, CBM6, CBM61, CBMNC, GH1, GH10, GH2, GH3, GH30, GH35, GH42, GH43, GH5, GH50, GH51, GH53, GH54, GH62, GHNC

Tabela 7 – Frequências absolutas de ocorrência de domínios CAZy para cada tipo de substrato.

Arabinano	arabinogalactano-I	arabinogalactano-II
16	27	27
homogalacturonanos	β -1-3,1-4 glucanos	galacturonanos
37	43	43
ramnogalacturonanos-I	celulose	xilanos
44	45	54
xiloglucanos	mananos	
59	69	

4.1.5 Resumo das estratégias

Ao longo do desenvolvimento foi possível adquirir-se maior sensibilidade de distinção sobre aquilo que é o conjunto das PCWDE e notar-se que a distinção entre a ação sobre PCWPS e outros polissacarídeos não é uma tarefa trivial. Há várias sobreposições de atividade entre polissacarídeos, já que há também muitas sobreposições de monômeros e ligações, e os genes levantados na 4ª estratégia, apesar de ser considerada a mais confiável, ainda provavelmente contém essas sobreposições. Outra característica que ainda não é passível de distinção através dessas estratégias é a da sensibilidade de ação das enzimas em diferentes contextos de cadeias secundárias de polissacarídeos ramificados. Ainda sim, foi possível enriquecer a anotação dos genomas tratados, designando função a várias proteínas anotadas como hipotéticas, entre outras com anotações de baixa especificidade.

Tabela 8 - Cepas e abreviações correspondentes aos 36 genomas analisados.

Nome extenso	Nome abreviado	4 ^a	3 ^a
<i>Stenotrophomonas maltophilia</i> D457	Sm D457	Sm	
<i>Stenotrophomonas maltophilia</i> JV3	Sm JV3	Sm1	
<i>Stenotrophomonas maltophilia</i> K279a	Sm K279a	Sm2	sml
<i>Stenotrophomonas maltophilia</i> R551 3	Sm R551 3	Sm3	smt
<i>Stenotrophomonas</i> SKA14	<i>Stenotrophomonas</i> SKA14	SS	
<i>Xanthomonas albilineans</i> GPE PC73	<i>X albilineans</i> GPE PC73	Xa	
<i>Xanthomonas fuscans aurantifolii</i> ICPB 10535	<i>X fuscans auranti</i> ICPB 10535	Xf	
<i>Xanthomonas fuscans aurantifolii</i> ICPB 11122	<i>X fuscans auranti</i> ICPB 11122	Xf1	
<i>Xanthomonas gardneri</i> ATCC 19865	<i>X gardneri</i> ATCC 19865	Xg	
<i>Xanthomonas perforans</i> 91 118	<i>X perforans</i> 91 118	Xp	
<i>Xanthomonas sacchari</i> NCPPB 4393	<i>X sacchari</i> NCPPB 4393	Xs	
<i>Xanthomonas vesicatoria</i> ATCC 35937	<i>X vesicatoria</i> ATCC 35937	Xv	
<i>Xanthomonas axonopodis citri</i> 306	Xa citri 306	Xa1	xac
<i>Xanthomonas axonopodis citrumelo</i> F1	<i>Xa citrumelo</i> F1	Xa2	
<i>Xanthomonas axonopodis manihotis</i>	<i>Xa manihotis</i>	Xa3	
<i>Xanthomonas campestris</i> 8004	Xc 8004	Xc2	xcb
<i>Xanthomonas campestris</i> ATCC 33913	Xc ATCC 33913	Xc3	xcc
<i>Xanthomonas campestris</i> B100	Xc B100	Xc4	xca
<i>Xanthomonas campestris musacearum</i> NCPPB4381	Xc musacearum NCPPB4381	Xc	xvm
<i>Xanthomonas campestris raphani</i> 756C	Xc raphani 756C	Xc5	xar
<i>Xanthomonas campestris vasculorum</i> NCPPB702	<i>Xc vasculorum</i> NCPPB702	Xc1	
<i>Xanthomonas campestris vesicatoria</i> 85 10	Xc vesicatoria 85 10	Xc6	xcv
<i>Xylella fastidiosa</i> 9a5c	Xf 9a5c	Xf4	xfa
<i>Xylella fastidiosa</i> Dixon	Xf Dixon	Xf2	xfi
<i>Xylella fastidiosa</i> GB514	<i>Xf GB514</i>	Xf5	
<i>Xylella fastidiosa</i> M12	Xf M12	Xf6	xfm
<i>Xylella fastidiosa</i> M23	Xf M23	Xf7	xfn
<i>Xylella fastidiosa</i> sandyi Ann 1	Xf sandyi Ann 1	Xf3	xfo
<i>Xylella fastidiosa</i> Temecula1	Xf Temecula1	Xf8	xfp
<i>Xanthomonas oryzae</i> KACC10331	Xo KACC10331	Xo	xoo
<i>Xanthomonas oryzae</i> MAFF 311018	Xo MAFF 311018	Xo1	xom
<i>Xanthomonas oryzae</i> NA8	<i>Xo NA8</i>	Xo4	
<i>Xanthomonas oryzae</i> X8-1A	<i>Xo X8-1A</i>	Xo5	
<i>Xanthomonas oryzae oryzicola</i> BLS256	Xo oryzicola BLS256	Xo2	xoz
<i>Xanthomonas oryzae oryzicola</i> MAI10	<i>Xo oryzicola</i> MAI10	Xo6	
<i>Xanthomonas oryzae</i> PXO99A	Xo PXO99A	Xo3	xop

Em negrito os 19 da 3^a enumeração. 3^a e 4^a referem-se aos locustags usados em cada etapa de enumeração.

Tabela 9 – Frequências de genes nas estratégias de enumeração de PCWDE.

Estratégia	Genomas	Enumerados	Hipotéticos	Concordantes	Discordantes	Subestimados
1 ^a	9	190	14,21%	162	28	496
2 ^a	19	498	0,80%	415	83	1063
3 ^a	19	1084	6,09%	829	255	649
4 ^a	36	3047	16,02%	3047	0	0

Os genomas estão listados na tabela 3. Enumerados – número absoluto de PCWDE enumerados nos genomas da estratégia. Hipotéticos – frequência relativa de genes enumerados com anotação como hipotéticos ou hipotéticos conservados. Concordantes – Genes em comum com a 4^a estratégia. Discordantes – genes ausentes na 4^a estratégia nos genomas usados. Subestimados – genes presentes na 4^a estratégia mas não presentes nas outras. Comparações feitas apenas com genomas em comum entre estratégias. Grifos – azul para números menores, passando por amarelo e vermelho para números maiores, de acordo com cada coluna.

Tabela 10 – Características de diversidade e frequências relativas de genes enumerados.

Estratégia	PCWDE CAZys	Pfams	Genes por genoma	Concordantes	Discordantes	Subestimados
1 ^a	25	35	21	85,26%	14,74%	261,05%
2 ^a	25	36	23	83,33%	16,67%	213,45%
3 ^a	45	82	48	76,48%	23,52%	59,87%
4 ^a	74	94	75	100,00%	0,00%	0,00%

PCWDE CAZys - ocorrências apenas dos 120 HMMs ligados a PCWDE, excluindo-se outros CAZys. Pfams – número de ocorrências de PFs como usado na 3^a estratégia. Genes por genoma – média de genes entre os 9 genomas em comum nas 4 estratégias (Sm R551 3, Xa citri 306, Xc ATCC 33913, Xc vesicatoria 85 10, Xf 9a5c, Xf Dixon, Xf sandyi Ann 1, Xf Temecula1 e Xo KACC10331). Concordantes, Discordantes e Subestimados – como na tabela 1, mas na forma de frequências relativas.

4.1.6 Diferenças entre *Xanthomonas*, *Xylella* e *Stenotrophomonas*

A característica mais marcante da enumeração (Tabela 11) é a diferença de genes PCWDE entre *Xanthomonas* e *Xylella*. Além da diferença de cerca de 2 kb entre os cromossomos dos dois gêneros, *Xylella* também possui um estilo de vida mais restrito nos ambientes que habita e na maneira como invade e coloniza o hospedeiro. Essa diferença, portanto, se deve provavelmente ao número mais restrito de polissacarídeos que encontra ao longo de sua vida.

Há outra diferença, no entanto bem menos marcante, entre *Xanthomonas oryzae* e outras *Xanthomonas*. Os cromossomos correspondentes quase não diferem em tamanho e nem o modo de vida das espécies. Porém, a influência da presença massiva de elementos de transposição IS no caso de *Xo* teve

provavelmente um papel fundamental na perda de genes por truncamento, recombinação ou inativação.

Notável sim, é a presença de vários genes de PCWDE no gênero *Stenotrophomonas*, ao qual pertence o patógeno humano. No entanto, os isolados submetidos a sequenciamento são pertencentes a apenas uma espécie, *Stenotrophomonas maltophilia*, desse gênero que também conta com outras duas espécies ligadas plantas, *Stenotrophomonas rizophila* e *Stenotrophomonas pavani*. Os possíveis ancestrais mais recentes de *Stenotrophomonas maltophilia* provavelmente contam com espécies que tiveram contato próximo com plantas, senão fitopatógenos.

4.2 Predição de atividades e substratos associados às enzimas

Da mesma forma que na enumeração, desenvolvemos 4 estratégias em série para a definição de substratos para as enzimas enumeradas (Tabela 12). Na 1ª procuramos números EC junto da descrição dos genes levantados, com um baixo número de ECs recuperados. Na 2ª usamos a informação levantada na 3ª estratégia de enumeração para tentar deduzir um substrato provável a partir de ECs atrelados aos PFs ou da descrição da anotação dos PFs no banco do Pfam ou do Interpro; obtivemos uma maior diversidade de ECs porém com pouca confiabilidade nas atribuições. Já na 3ª foi onde fizemos o levantamento de ECs associados à PCWPS do banco da IUBMB, mas em conjunto com as sequências de referência para os ECs presentes no banco do KEGG. Como o banco KEGG conta com um número restrito dessas sequências, apelamos para técnica de Bork et al. (2012) mas ainda sem muito sucesso no número de designações. Por fim, na 4ª estratégia usamos os ECs levantados sobre a anotação das classes do CAZy, com todos os genes tendo uma designação de substrato porém com uma grande multiplicidade de substratos atribuídos a um mesmo gene.

Tabela 11 – Frequência de genes enumerados para cada genoma em cada uma das estratégias.

Genoma	Estratégia			
	1 ^a	2 ^a	3 ^a	4 ^a
Sm D457				71
Sm JV3				61
Sm K279a		4	39	69
Sm R551 3	4	4	32	65
<i>Stenotrophomonas</i> SKA14				63
<i>X albilineans</i> GPE PC73				71
<i>X fuscans auranti</i> ICPB 10535				117
<i>X fuscans auranti</i> ICPB 11122				120
<i>X gardneri</i> ATCC 19865				140
<i>X perforans</i> 91 118				131
<i>X sacchari</i> NCPPB 4393				125
<i>X vesicatoria</i> ATCC 35937				137
<i>Xa citri</i> 306	39	40	86	131
<i>Xa citrumelo</i> F1				114
<i>Xa manihotis</i>				101
<i>Xc</i> 8004		44	90	128
<i>Xc</i> ATCC 33913	36	45	95	128
<i>Xc</i> B100		44	91	127
<i>Xc musacearum</i> NCPPB4381		39	94	98
<i>Xc raphani</i> 756C		45	92	126
<i>Xc vasculorum</i> NCPPB702				99
<i>Xc vesicatoria</i> 85 10	40	40	89	121
<i>Xf</i> 9a5c	11	11	17	30
<i>Xf</i> Dixon	10	12	18	30
<i>Xf</i> GB514				27
<i>Xf</i> M12		9	15	28
<i>Xf</i> M23		10	16	30
<i>Xf sandyi</i> Ann 1	11	12	16	57
<i>Xf Temecula</i> 1	10	10	14	30
<i>Xo oryzae</i> KACC10331	29	32	68	85
<i>Xo oryzae</i> MAFF 311018		34	65	85
<i>Xo oryzae</i> NA8				47
<i>Xo oryzae</i> X8-1A				45
<i>Xo oryzicola</i> BLS256		29	71	81
<i>Xo oryzicola</i> MAI10				45
<i>Xo oryzae</i> PXO99A		34	76	84

4.2.1 EC atribuído na anotação

Foi possível recuperar poucos genes com atribuição de EC vindas diretamente da anotação. Alguns poucos projetos genoma consistentemente colocaram números EC junto da descrição dos genes anotados.

4.2.2 Pfam e InterPro

Apesar do baixo número de ECs encontrados associados aos PFs essa estratégia foi mais bem sucedida que a estratégia anterior em achar substratos para a maioria dos genes. Isso porque os indícios coletados foram mais indiretos, baseados principalmente na descrição da anotação de proteínas cujas sequências foram usadas para construir o HMM. Isso no entanto causou uma distorção na designação de substratos, favorecendo apenas a designação de apenas um substrato, o que se opõe ao achado de que existe uma inespecificidade inerente às PCWDE. Da mesma maneira, as designações foram mais gerais, ficando no nível das classes de PCWPS celulose, hemicelulose e pectina.

4.2.3 Substratos de acordo com ECs da IUBMB e sequencias do KEGG

Através da relação de números EC, mantida pela União internacional de Bioquímica e Biologia Molecular (IUBMB), e usando a informação obtida na literatura pudemos obter 86 números ECs ligados a PCWDE. Todos esses constam do banco do KEGG apesar de alguns não ocorrerem nos mapas metabólicos, porém 39 ECs são órfãos de sequências moleculares representativas no banco do KEGG. Ao se fazer um tBLASTn das sequências proteicas obtidas no KEGG contra os genes enumerados puderam se verificar 20 ECs que não tem resultado de BLAST e 27 ECs com resultado de BLAST e, usando-se a técnica de Bork et al. (2012) recuperou-se mais um EC com resultado de BLAST.

4.2.4 Substratos atribuídos através da relação entre CAZys e ECs

Como mencionado anteriormente, nos valem das informações disponíveis tanto na literatura quanto nos bancos da IUBMB e do CAZy/dbCAN para atribuir substratos a cada domínio. Dessa forma o levantamento dos ECs dos CAZys encontrados através do dbCAN permitiram atribuições de substratos aos genes, unindo o processo de enumeração de PCWDE e de designação de substratos. Essas atribuições foram na maioria múltiplas para cada gene, característica da promiscuidade das cazys frente a polissacarídeos com monômeros e ligações semelhantes. Em contrapartida essa análise não incorpora informação para distinguir enzimas que tenham restrições de ação sob diferentes contextos de posicionamento e composição da vizinhança de cadeias laterais em polissacarídeos ramificados.

Tabela 12 - ECs e substratos.

		ECs			Genes		
		Substratos	na estratégia	PCWDE/IUBMB	sem substratos	com substratos	1 substrato
Estratégia							2 ou mais substratos
Anotação	Todos		27	22	2989	58	0
Pfam e Interpro	Todos		60	23	582	2465	2381
IUBMB, KEGG e Bork et al	Todos		28	28	627	2420	31
dbCAN, CAZy e IUBMB	Todos		152	86	0	3047	620
							2427

Frequências de ECs e genes com substrato encontrados através das diferentes estratégias. Todos os substratos foram contemplados em todas as estratégias. ECs na estratégia – total de números EC encontrados na estratégia. ECs PCWDE/IUBMB – quantos dos ECs totais pertenciam à coleção de 86 levantados do banco da IUBMB. Genes com substratos é a soma de genes com 1 substrato e 2 ou mais substratos.

4.2.5 Conteúdo de PCWDE relativo às atividades sobre cada PCWPS

A parede da célula de espécies da família *Poales* (gramíneas; liliopsida/monocotiledôneas), do tipo II, possui maior conteúdo de xilanos, enquanto outras plantas, com paredes do tipo I, possuem mais pectinas (Carpita, Gibeaut, 1993). Procuramos investigar se existe uma correlação da composição da parede do

hospedeiro com o conteúdo relativo de PCWDE para cada tipo de PCWPS de cada genoma. Dentre os hospedeiros das espécies com genomas aqui tratados, arroz é o único hospedeiro da família *Poales*, com parede do tipo II. Seu patógeno, *Xanthomonas oryzae*, no entanto não apresenta nenhuma discrepância notável no seu conteúdo relativo de xilanases.

As diferenças mais notáveis, no entanto, são notadas nos genomas de *Xylella* e *Stenotrophomonas*. *Xylella* tem um alto conteúdo de enzimas que agem sobre celulose, mananos, xiloglucanos e β 1-3,1-4 glucano, enquanto *Stenotrophomonas* tem um baixo conteúdo, e a relação oposta é verdadeira para enzimas que agem sobre xilanos, galacturonanos, homogalacturonanos e ramnogalacturonano-I.

De forma geral as proporções, entre os diferentes substratos, encontradas (Tabela 14 e Tabela 13) eram esperadas de acordo com a complexidade conhecida de cada polímero, com a exceção de xiloglucanos e mananos. Estes dois polímeros contam com uma maior complexidade em relação à celulose, no entanto as proporções desses três polímeros foram encontradas bastante próximas. Possivelmente esse fato pode estar relacionado com a sensibilidade dependente do contexto de cadeias secundárias ou até mesmo com o modo de invasão ou colonização do tecido do hospedeiro.

Tabela 13 – Médias de frequências de genes e substratos.

Tabla 10. Medias de frecuencias de genes e substratos.														
Grupo de genomas	Genes		Gene x Substratos	Celulose	Mananos	Xiloglucanos	β 1-3,1-4 glucanos	Xilanos	Galacturonanos	Homogalacturonanos	Ramnogalacturonano-I	Arabinogalactano-I	Arabinogalactano-II	Arabinano
<i>Xanthomonas</i>	121	569	7%	9%	8%	5%	15%	14%	11%	14%	7%	7%	3%	
<i>X. albilineans</i>	71	377	7%	8%	8%	5%	15%	14%	12%	14%	6%	6%	5%	
<i>X. oryzae</i>	67	334	7%	8%	8%	5%	15%	14%	11%	14%	7%	7%	4%	
<i>Stenotrophomonas</i>	66	258	6%	9%	7%	5%	18%	15%	13%	15%	4%	4%	2%	
<i>Xylella</i>	33	181	9%	11%	9%	6%	13%	12%	11%	12%	7%	7%	3%	

Como na Tabela 14 porém fazendo-se a média para *Xylella*, *Stenotrophomonas*, *Xanthomonas oryzae*, *Xanthomonas albilineans* e o restante das *Xanthomonas*.

Tabela 14 - Tabela de frequências de genes e substratos.

Genoma	Genes	Gene x Substratos	Celulose	Mananos	Xiloglucanos	β 1-3,1-4 glucanos	Xilanos	Galacturonanos	Homogalacturonanos	Ramnogalacturonano-I	Arabinogalactano-I	Arabinogalactano-II	Arabinano
<i>Xa citri</i> 306	131	605	7%	9%	8%	5%	15%	14%	11%	14%	7%	7%	3%
<i>Xa citrumelo</i> F1	114	540	7%	9%	8%	5%	15%	13%	11%	13%	8%	8%	4%
<i>Xa manihotis</i>	101	506	8%	9%	9%	6%	14%	13%	10%	13%	7%	7%	4%
<i>Xc musacearum</i> NCPPB4381	98	486	7%	9%	8%	6%	15%	13%	10%	13%	8%	8%	3%
<i>Xc vasculorum</i> NCPPB702	99	487	7%	9%	8%	6%	15%	13%	11%	13%	7%	7%	3%
<i>Xc</i> 8004	128	600	6%	8%	8%	5%	15%	15%	12%	15%	7%	7%	3%
<i>Xc</i> ATCC 33913	128	593	7%	9%	8%	5%	15%	14%	12%	14%	7%	7%	3%
<i>Xc</i> B100	127	590	7%	9%	8%	5%	15%	14%	12%	14%	7%	7%	3%
<i>Xc raphani</i> 756C	126	575	7%	9%	8%	5%	15%	14%	11%	14%	7%	7%	3%
<i>Xc vesicatoria</i> 85 10	121	563	7%	9%	8%	5%	15%	14%	12%	14%	7%	7%	3%
<i>X.fus. auranti.</i> ICPB 10535	117	549	7%	9%	8%	5%	15%	13%	11%	13%	8%	8%	4%
<i>X.fus. auranti.</i> ICPB 11122	120	566	7%	9%	8%	5%	15%	13%	11%	13%	7%	7%	3%
<i>X.gardneri</i> ATCC 19865	140	646	7%	9%	8%	5%	15%	14%	12%	14%	7%	7%	3%
<i>X.perforans</i> 91 118	131	597	7%	9%	8%	5%	15%	14%	11%	14%	7%	7%	3%
<i>X.sacchari</i> NCPPB 4393	125	607	7%	9%	8%	5%	14%	14%	11%	14%	8%	8%	3%
<i>X.vesicatoria</i> ATCC 35937	137	591	6%	8%	8%	5%	16%	14%	12%	14%	6%	6%	3%
<i>Xo</i> KACC10331	85	429	7%	8%	8%	5%	15%	14%	12%	14%	7%	7%	4%
<i>Xo</i> MAFF 311018	85	424	6%	7%	7%	5%	16%	14%	12%	14%	7%	7%	4%
<i>Xo oryzicola</i> BLS256	81	400	6%	8%	7%	5%	16%	15%	12%	15%	7%	7%	4%
<i>Xo</i> PXO99A	84	430	7%	7%	8%	5%	15%	14%	12%	14%	7%	7%	4%
<i>Xo oryzae</i> NA8	47	246	7%	8%	7%	4%	15%	14%	11%	14%	8%	8%	4%
<i>Xo oryzae</i> X8-1A	45	210	7%	9%	8%	5%	16%	14%	10%	14%	7%	7%	4%
<i>Xo oryzicola</i> MAI10	45	198	7%	8%	8%	6%	15%	15%	11%	15%	7%	7%	3%
<i>X.albilineans</i> GPE PC73	71	377	7%	8%	8%	5%	15%	14%	12%	14%	6%	6%	5%
<i>Stenotrophomonas</i> SKA14	63	252	7%	10%	8%	5%	18%	14%	13%	14%	4%	4%	2%
<i>Sm</i> D457	71	282	6%	9%	7%	4%	18%	16%	13%	16%	4%	4%	2%
<i>Sm</i> JV3	61	242	6%	10%	7%	5%	19%	14%	13%	14%	4%	4%	2%
<i>Sm</i> K279a	69	259	6%	9%	7%	4%	19%	17%	14%	17%	3%	3%	2%
<i>Sm</i> R551 3	65	256	6%	9%	7%	5%	18%	15%	14%	15%	4%	4%	2%
<i>Xf</i> Dixon	30	175	9%	11%	9%	6%	13%	12%	11%	12%	7%	7%	3%
<i>Xf sandyi</i> Ann 1	57	311	8%	11%	8%	6%	13%	12%	12%	12%	7%	7%	3%
<i>Xf</i> 9a5c	30	164	9%	11%	9%	6%	13%	12%	12%	12%	7%	7%	3%
<i>Xf</i> GB514	27	138	8%	11%	8%	7%	14%	12%	12%	12%	7%	7%	2%
<i>Xf</i> M12	28	150	9%	11%	9%	5%	13%	12%	11%	12%	7%	7%	3%
<i>Xf</i> M23	30	163	9%	11%	9%	6%	13%	12%	11%	12%	7%	7%	3%
<i>Xf Temecula</i> 1	30	163	9%	11%	9%	6%	13%	12%	11%	12%	7%	7%	3%

Frequências absolutas de genes e gene x substratos e relativa dos substratos. Gene x substratos indica a soma de múltiplos substratos atribuíveis a cada gene do genoma. As frequências relativas de substratos reflete o potencial de ação relativo à frequência de gene x substratos. Média de 4,8 substratos por gene.

4.3 Relação de homologia entre os genes enumerados

4.3.1 Clusterizações: *BLASTclust* com genes isca e o desenvolvimento do *blastCORE*

Durante o desenvolvimento da 2ª técnica de enumeração de PCWDE o levantamento por BLAST com gene isca motivou o delineamento da técnica *blastCORE*. Durante esse desenvolvimento procurou-se controlar os parâmetros do *BLASTclust* de forma objetiva, gerando clusters que refletissem possíveis relações de homologia gênica derivada da identidade ou similaridade de sequências. O desenvolvimento do *blastCORE* se deu antes que se tomasse conhecimento da existência prévia de técnicas de inferência de homologia de sequências como o *OrthoMCL* ou o *InParanoid*. A técnica desenvolvida, nomeada de *BlastCORE*, apoia-se no critério CORE de confiabilidade do alinhamento global. Este é uma medida usada internamente pelo software de alinhamento múltiplo T-Coffee para quantificar a confiabilidade dos alinhamentos gerados durante o processo iterativo de refinamento do alinhamento global (Notredame et al., 2000). Esse critério é avaliado para cada posição alinhada, mas também é usado para o cálculo de uma estatística para cada par de sequências e também para o alinhamento múltiplo global. O critério de alinhamento satisfatório do T-Coffee é de uma pontuação CORE de no mínimo 70. No *blastCORE* procuramos maximizar os COREs dos clusters procurando aperfeiçoar os parâmetros do *BLASTclust*.

4.3.2 Comparação entre *blastCORE* e *OrthoMCL*

Muitos dos *clusters* gerados inicialmente pelo *blastCORE* quando submetidos a uma análise de filogenia molecular (árvores de máxima verossimilhança) apresentavam subgrupos de sequências que aparentavam ser parálogos. Durante o processo de avaliação das técnicas de clusterização descobriu-se que o *OrthoMCL* era capaz de repartir em clusters menores esses clusters de parálogos do *blastCORE*, como ilustrado na Figura 4.

Por outro lado vários genes singletons pelo *OrthoMCL* foram incorporados em clusters pelo *blastCORE*, em especial em casos de genes divididos ou

interrompidos, como é mais frequente em *Xanthomonas oryzae*. Na Figura 5 podemos observar um cluster do OrthoMCL formado por apenas duas seqüências que possuem baixa similaridade. O critério de alinhamento local usado pelo OrthoMCL considerou que essas seqüências de aminoácido teriam maior similaridade entre si que com o cluster similar ao gerado com critério de alinhamento global (CORE) do blastCORE (Figura 6). A seqüência mal agrupada pelo OrthoMCL faz parte de um gene dividido em duas partes em relação a outros genomas, sendo que sua outra metade já fazia naturalmente parte do cluster com seus ortólogos íntegros.

Figura 4 – Cluster do blastCORE mais favoravelmente resolvido pelos seus correspondentes do OrthoMCL. Cada caixa indica o ramo da árvore que contém as seqüências dos agrupamentos do OrthoMCL.

Figura 5 - Alinhamento de um cluster exemplo desfavorável ao critério do OrthoMCL. Alinhamento global das seqüências do cluster de dois genes. Baixa similaridade entre as duas seqüências com muitos gaps e mismatches.

Figura 6 - Alinhamento de um cluster favorável ao critério do blastCORE. A mesma sequência da base da Figura 5 está no topo dessa figura. Essa proteína está dividida nesse genoma em relação aos outros genomas do mesmo gênero.

4.3.3 União dos resultados do blastCORE e do OrthoMCL

Tabela 15 - Comparação entre blastCORE e OrthoMCL.

Característica	3ª Técnica	4ª Técnica
blastCORE clusters	111	187
blastCORE singletons	59	67
OrthoMCL clusters	124	174
OrthoMCL singletons	30	101
Iguais (clusters)	85	49
Iguais (singletons)	22	31
semelhantes (clusters+singletons)		
blastCORE	26	138
OrthoMCL	39	125
Total	110	369
Final		
blastCORE semelhantes	15	85
OrthoMCL semelhantes	19	91
Clusters	119	225
Singletons	44	53
CORE médio		
blastCORE	86,20	94,80
OrthoMCL	97,20	91,20
Final	98,46	95,16

Analizados os genes da 3ª técnica e 4ª técnica de enumeração. Iguais - aqueles os quais tem exatamente os mesmos genes membros. Semelhantes - mesmos genes mas divididos de forma diferente, possivelmente num número diferente de clusters. Final – semelhantes escolhidos na união entre OrthoMCL e blastCORE e número de clusters e singletons após a união. CORE médio calculado sem singletons (CORE=0).

Ao unirmos os resultados das duas técnicas (Tabela 15) procuramos encontrar fragmentações ou uniões de clusters de forma também a maximizar o CORE dos clusters finais. Comparando esse processo para os genes da 3ª técnica e 4ª técnica de enumeração encontramos duas situações, na primeira em que o

blastCORE tinha uma média de COREs menor que o OrthoMCL antes da união e na segunda em que ocorreu o inverso. No caso da 4ª técnica houve também uma redução de singletons de ambos blastCORE e OrthoMCL.

O blastCORE ainda tem um custo computacional relativamente alto, mas que pode se assemelhar ao OrthoMCL com aplicação de uma heurística de procura de parâmetros do BLASTclust. No entanto, seu ponto forte é um menor tempo necessário para preparação dos dados de entrada e simplicidade de uso em relação ao OrthoMCL.

4.4 Árvores filogenéticas de genes homólogos

A árvore que foi usada como referência foi a publicada no artigo de Lu et al. (2008). Essa árvore (Figura 7) foi construída usando-se o concatenado do óperon de RNAs ribossomais com alguns genes de óperons relevantes à filogenia das *Xanthomonadaceae* como abordado no artigo.

Figura 7 - Árvore do concatenado de óperons. As flechas segmentadas indicam a presença ou ausência de genes dos óperons indicados.
Fonte: Lu et al. (2008).

De forma geral, as árvores geradas a partir dos conjuntos de sequências dos clusters têm ramos com suporte de pelo menos 70% das árvores de *bootstrap*. Há exceções onde uma bissecção entre genomas próximos, como diferentes cepas de um mesmo patovar, tem uma queda de até 40% do suporte. Isso indica que é mais provável uma politomia ao invés da bissecção, como observado na árvore de .óperons.

No entanto, a comparação manual apenas de valores de suporte de ramo entre árvores deixa a desejar quando pretendemos analisar algumas centenas de árvores, como as geradas para os clusters. Se observarmos a disposição e as escalas das árvores geradas (exemplos nas figuras Figura 8, Figura 9 e Figura 10), podemos notar que existem variações da taxa de mudança de aminoácidos entre os clusters e também entre genes de um mesmo genoma. Para automatizar esse processo e impormos um critério mais objetivo de comparação de árvores recorreremos ao algoritmo de comparação de ramos do software *treedist* e clusterização do grafo gerado pelo software MCL.

Figura 8 - Árvore do cluster e4o0025 com a mesma escala da árvore do cluster e4o0018, porém com uma disposição diferente.

Figura 9 – Árvore do cluster e4o0028 com com disposição semelhante à árvore do cluster e4o0025, porém com uma escala 10 vezes menor

Figura 10 – Árvore do cluster e4o0018 com uma disposição diferente das árvores dos clusters e4o0025 e e4o0028, porém com a mesma escala da árvore de e4o0025.

4.4.1 Comparação do perfil gênico de PCWDE homólogos entre genomas com agrupamento por distância entre árvores filogenéticas

Para comparar o perfil de genes homólogos, determinados pela clusterização, construímos uma tabela (

Figura 11) com um cluster por linha, um genoma por coluna e as células sombreadas de acordo com a presença ou ausência de genes. Podemos dizer que cada linha no perfil de clusters corresponde a um conjunto de genes ortólogos. A conservação de um gene entre genomas tão próximos quanto os de *Xanthomonadaceae* torna mais verossímil a hipótese de ortologia. Os clusters fragmentados por repetição, como descrito, são prováveis parálogos por duplicações recentes. A essa tabela também foi integrada a informação dos grupos de árvores (Tabela 16), tanto as de aminoácidos quanto as em relação ao 16S e também um levantamento de clusters pertencentes às regiões HGT do trabalho de Lima et al. (2005). O ordenamento da tabela foi feito primeiro em relação aos grupos de árvores de aminoácidos, depois em relação aos cromossomos de *Xa. citri* e *Xc. campestris*.

Figura 11 – Relação de clusters, HGTs e distâncias entre árvores filogenéticas.

(p. anterior) A cor do sombreado nas células da terceira coluna indica um grupo de árvores de máxima verossimilhança próximas de acordo com o algoritmo *treedist/MCL*, a da segunda coluna indica o grupo de distância em relação à árvore de 16S e a primeira coluna se os genes fazem parte de HGT. As caixas vermelhas indicam genomas com maior número de genes dentro do grupo de árvores. Grupos de genomas por espécie: Xf-*Xylella fastidiosa*, Sm-*Stenotrophomonas maltophilia*, Xo-*Xanthomonas oryzae*, Xal-*Xanthomonas albilineans*, Xa-*Xanthomonas axonopodis* patovares citri, citrumelo e manihotis, Xc-*Xanthomonas campestris* patovares campestris, vesicatoria, raphani, musacearum e vasculorum, X-*Xanthomonas fuscans aurantifolii*, *Xanthomonas gardneri*, *Xanthomonas perforans*, *Xanthomonas sacchari* e *Xanthomonas vesicatoria*.

Tabela 16 - Relação dos grupos de árvores de clusters, frequência de clusters e cores representativas.

Grupo	Clusters	Cor
1	118	
2	15	
3	11	
4	8	
5	8	
6	6	
7	6	
8	2	
9	2	
-	77	

Na tabela é possível ver um grande número de genes conservados pertencentes principalmente ao gênero *Xanthomonas* como já observado no processo de enumeração. O grupo 1 de árvores é o mais extenso e portanto considerado a base de PCWDE dos genomas. Esses ortólogos estão provavelmente sujeitos a um mesmo tipo de evolução por estarem sujeitos a taxas de mudança próximas e estarem em diferentes genomas. Também devem fazer parte a mais tempo do genoma e é onde estão os HGT de Lima et al. (2005) Trouxemos as informações das regiões HGT, às propagamos para todos os genomas em estudo e pudemos observar que uma parte das carboidrazes está contida nas regiões HGT, mas que não são em número tão grande como poderia se esperar para um caractere de virulência do patógeno. Ainda sim, não podemos encontrar nenhuma dessas regiões nos genomas de *Xylella*. Esses HGT, como já observado, provavelmente ocorreram já há bastante tempo, no ancestral comum entre *Xanthomonas* e *Stenotrophomonas*, mas após a divergência de *Xylella*. Árvores sem agrupamento têm taxas mais distantes do restante dos genes, fazem parte principalmente de *X. sacchari*, *X. vesicatoria*, *X. glycines* e *Stenotrophomonas*. Sua

distância 16S não pode ser calculada pois não são compartilhados com quase nenhum outro genoma. Portanto, são prováveis HGTs ou genes com problemas de anotação.

As distâncias em relação à árvore de 16S não revelaram nenhum padrão que se correlacionasse com HGTs, algum grupo de genomas ou de árvores. Aparentemente 16S não constitui um bom parâmetro para comparação filogenética com genes PCWDE.

Xa. citri é uma das espécies mais basais do gênero *Xanthomonas* como mostrado na árvore de óperons. *Xa citri* é representada na tabela pela maior linha contígua de clusters nos grupos 1, 3, 5 e 8. Assim está mais próxima do evento de consolidação da maior parte do aparato de PCWDE, isto é, tem uma cópia da maior parte dos genes PCWDE. *X. campestris* conta com vários eventos de duplicação exclusivos no grupo 1 de árvores, ou seja, a maior parte dos genes parálogos dos clusters fragmentados. Como estes genes estão sujeitos às mesmas taxas de genes basais do genoma, devem ser duplicações recentes. Essas diferenças de números de cópia se correlacionam com fenótipos de maceração mais agressiva de tecido do hospedeiro e, portanto, com uma maior capacidade de degradação da parede celular vegetal. São três as justificativas possíveis para tal: um maior número de cópias pode abastecer melhor a necessidade de geração de RNA mensageiro; ter mais cópias permite uma mudança de função dependente de regulação, permitindo a adaptação das diferentes cópias a diferentes contextos do metabolismo da célula, variáveis ambientais, estágio da infecção, tecido do hospedeiro, entre outros; e, por fim, uma maior quantidade de membros específicos de famílias multigênicas de PCWDE podem conferir uma melhor habilidade de hidrólise de determinados substratos.

4.4.2 Variação na estrutura de genes homólogos

Outra evidência interessante surgiu ao fazermos um levantamento da variação de domínios CAZy nos genes dos diferentes clusters. Encontramos polimorfismos entre genes ortólogos que ocasionam a classificação de um domínio CAZy de maneira diferente entre genes de um mesmo cluster. Essa mudança na classificação provavelmente tem uma ligação com a adaptação à arquitetura e composição dos diferentes PCWPS de cada hospedeiro.

Essas pequenas diferenças são exemplificadas com grifos no trecho do alinhamento do cluster 106 (Figura 12) e listado na tabela (Tabela 17).

Tabela 17 - Clusters com variação de CAZy. ....(continua)

Cluster	CAZys	Genomas
106	CE10	Xa1,Xf1,Xf
106	CE5	Xa2,Xc6,Xo1,Xo3,Xo,Xp
10	CE10	SS,Sm1,Sm2,Sm3,Sm,Xa1,Xa2,Xa3,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf2,Xf3,Xf3,Xf5,Xf6,Xf7,Xf8,Xf,Xg,Xo1,Xo2,Xo3,Xo4,Xo5,Xo6,Xo,Xp,Xs,Xv
10	PL22-CE10	Xa,Xf4
13	CBM23-CBM32	Xa1,Xa,Xc4,Xc5,Xc,Xf1,Xf,Xg,Xo1,Xo2,Xo3,Xo5,Xo,Xs
13	CBM23-CBM32-GH15	SS,Sm1,Sm3,Sm,Xf3,Xf4
13	CBM23-CBM32-GH78	Xf2,Xf3,Xf5,Xf6,Xf7,Xf8
13	CBM32	Xa2,Xc2,Xc3,Xc6,Xo4,Xo6,Xp,Xv
13	CBM32-GH15	Sm2
159	CE1	SS,Sm2
159	CE10	Sm1
162	CE1	Xc2,Xc3
162	CE7	Xg
25	GH12	Xa1,Xa2,Xa3,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf,Xg,Xg,Xg,Xo1,Xo3,Xo,Xp,Xs,Xv
25	GH12-CBM2	Xa
31	CE10	SS,Sm1,Sm3,Sm,Xa1,Xa2,Xa3,Xa,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf,Xg,Xo1,Xo2,Xo3,Xo5,Xo6,Xo,Xp,Xs,Xv
31	PL22-CE10	Sm2
36	GH6	Xc5,Xf3,Xo2,Xo6
36	GH6-CBM2	Xa,Xf2,Xf3,Xf4,Xf6,Xf7,Xf8
36	GH6-CBM37	Xa3,Xc1,Xc2,Xc2,Xc3,Xc3,Xc4,Xc4,Xc,Xo1,Xo3,Xo4,Xo5,Xo,Xv
e4o0001	CE1	Xc1
e4o0001	CE10	Sm2,Xs,Xs
e4o0011	CBM48-CBM48-GH13	Sm,Xa1,Xa1,Xa2,Xa2,Xa3,Xa3,Xa,Xc1,Xc1,Xc2,Xc2,Xc3,Xc3,Xc4,Xc4,Xc5,Xc5,Xc6,Xc6,Xc,Xc,Xf1,Xf1,Xf,Xf,Xg,Xg,Xo1,Xo1,Xo2,Xo2,Xo3,Xo3,Xo4,Xo4,Xo5,Xo5,Xo6,Xo,Xo,Xp,Xp,Xs,Xs,Xv,Xv
e4o0011	CBM48-GH13	Sm2
e4o0011	GT5-CBM48-CBM48-GH13	Xo6

Alguns dos clusters possuem até 4 variações de CAZys que compõe seus genes.

Tabela 17 - Clusters com variação de CAZy. ....(conclusão)

Cluster	CAZys	Genomas
e4o0018	CE10	SS,Sm1,Sm2,Sm3,Sm,Xa3,Xa,Xc1,Xc,Xf1,Xf2,Xf3,Xf3,Xf4,Xf5,Xf6,Xf7,Xf8,Xf,Xo1,Xo2,Xo3,Xo5,Xo,Xv
e4o0018	CE7	Xa1,Xa2,Xc2,Xc3,Xc4,Xc5,Xc6,Xp,Xp,Xs
e4o0027	GH43	Xa1,Xa1,Xa2,Xa2,Xa3,Xa3,Xc1,Xc2,Xc2,Xc3,Xc3,Xc4,Xc4,Xc5,Xc5,Xc6,Xc6,Xc,Xf1,Xf1,Xf,Xf,Xg,Xp,Xp,Xv,Xv
e4o0027	GH43-CBM32	Xs
e4o0030	GH5	Xv
e4o0030	GH5-CBM63	Xa3,Xc1,Xc2,Xc3,Xc4,Xc5,Xc,Xf2,Xf3,Xf3,Xf4,Xf5,Xf6,Xf7,Xf8,Xg,Xo1,Xo2,Xo3,Xo4,Xo5,Xo6,Xo
e4o0044	GH91-PL6	Xa
e4o0044	PL1	Xa1,Xa2,Xa3,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf,Xg,Xo1,Xo2,Xo3,Xo4,Xo6,Xo,Xp,Xv
e4o0052	GH2	Xa1,Xa2,Xa3,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf,Xg,Xp,Xs,Xv
e4o0052	GH53	Xo1,Xo2,Xo6,Xo
e4o0064	CE1	Xa1,Xa2,Xc1,Xc2,Xc3,Xc4,Xc5,Xc6,Xc,Xf1,Xf,Xo1,Xo2,Xo3,Xo,Xp
e4o0064	CE1-CE10	Xg,Xv
e4o0086	CBM35	Xc1
e4o0086	GH98-CBM35	Xa1,Xa2,Xa3,Xa,Xc6,Xf1,Xf,Xo1,Xo2,Xo3,Xo4,Xo,Xp
e4o0091	GH78	Sm1,Sm2,Sm,Xa1,Xa3,Xc2,Xc3,Xc4,Xc5,Xp,Xs,Xv
e4o0091	GH78-CBM35	Sm3
e4o0105	CE1	SS
e4o0105	CE7	Xf2,Xf3,Xf3,Xf4,Xf5,Xf6,Xf7,Xf8
e4o0115	GH16	Xa1,Xc4,Xf1,Xg,Xv
e4o0115	GH16-CBM23	Xc2,Xc3,Xc5
e4o0116	CBM12-GH18	SS,Sm1,Sm2,Sm3,Sm,Xs
e4o0116	GH18	Xa3,Xa
e4o0121	CBM32	Sm1,Sm3
e4o0121	CBM32-GH15	SS,Sm2,Sm,Xs

Alguns dos clusters possuem até 4 variações de CAZys que compõe seus genes.

```

Xa1|NP_640728.1  LPPFALLGYSGARVALLLAQQQPQEISALVLFGFDDVDTPPAS
Xf1|ZP_06706162  VPPFALLGYSGARVALLLAQQQPQEISALVLFGFDDVDTPPAS
Xf|ZP_06732442.  VPPFALLGYSGARVALLLAQQQPQEISALVLFGFDDVDTPPAS
Xa2|YP_00484996  LPPFALLGYSGARVALLLAQQQPQAMSALVLFGFDDVDAAPAP
Xc6|YP_362117.1  LPPFALLGYSGARVALLLAQQQPQAMSALVLFGFDDVDAAPAP
Xo1|YP_449483.1  LPPFALLGYSGARVALLLAQQQPQAMSAFVLFGFDDVDAPPAP
Xo3|YP_00191583  LPPFALLGYSGARVALLLAQQQPQAMSAFVLFGFDDVDAPPAP
Xo|YP_199126.1   LPPFALLGYSGARVALLLAQQQPQAMSAFVLFGFDDVDAPPAP
Xp|ZP_08190530.  LPPFALLGYSGARVALLLAQQQPQAMSALVLFGFDDVDAAPAP
cons             :*****:****:*****:***.
Xa1|NP_640728.1  AVPADRPLRTPTTAAAAGEDFITANAAPPVRAAYVAQALAADPV
Xf1|ZP_06706162  AVPADRPLRTPTTAAAAGEDFITANAAPPSVRAAYVAQALAADPV
Xf|ZP_06732442.  AVPADRPLRTPTTAAAAGEDFITANAAPPSVRAAYVAQALAADPV
Xa2|YP_00484996  AVPADRPLRAPTTAAAAGEDFITANAAPPGVRAAYVAQALAADPV
Xc6|YP_362117.1  AVPADRPLRAPTTAAAAGEDFITANAAPPGVRAAYVAQALAADPV
Xo1|YP_449483.1  AVAADRPLRAPTTAAAAGEDFITANAAPPAVRAAYVAQALASDPV
Xo3|YP_00191583  AVAADRPLRAPTTAAAAGEDFITANAAPPAVRAAYVAQALASDPV
Xo|YP_199126.1   AVAADRPLRAPTTAAAAGEDFITANAAPPAVRAAYVAQALASDPV
Xp|ZP_08190530.  AVPADRPLRAPTTAAAAGEDFITANAAPPGVRAAYVAQALAADPV
cons             **.*****:***** *****:***

```

Figura 12 - Trecho do alinhamento de aminoácidos do cluster 106. Grifados os resíduos com polimorfismo ligados à mudança da classificação do CAZy.

4.5 Seleção de candidatos para validação da anotação por biologia molecular

O enfoque do trabalho molecular se voltou para duas cepas de *Xanthomonas oryzae*: Xo pv. *oryzae* PXO99A (Xop) e Xo pv. *oryzicola* BLS256 (Xoc). Isso foi motivado por um grande número de trabalhos direcionados à fitopatologia da espécie com enfoque nos genes de interesse, o grande impacto causado na economia da planta de arroz, e principalmente à diferença de fenótipo de infecção entre dois patovares com conteúdo gênico muito similar que infestam uma mesma espécie de planta. Dos 149 genes de PCWDE levantados na 3ª enumeração para esses dois genomas, 24 genes foram considerados candidatos para o desenvolvimento de um trabalho de caracterização molecular (

Tabela 18).

Tabela 18 - Processo de seleção de genes candidatos de Xoc e Xop.

Genomas	Inicial	CAZy	Início da cascata	Loc. Sub. Cel	Sem literatura
Xoc	72	66	58	20	11

Xop	78	73	67	22	13
Xo	150	140	126	42	24
3ª enum.	1084	971	866	243	129
PFs	57	50	48	21	16

Cada coluna corresponde a uma etapa de exclusão de genes.

Foi estabelecida uma série de critérios de seleção dos genes (tabela 4): genes com modelos CAZy, genes do início da cascata de degradação, genes com localização (Tabela 19) no periplasma, membrana externa ou secreção e não secretado por um sistema do tipo três, e, por fim, genes não abordados na literatura. Os pares de ortólogos foram ambos incluídos nos candidatos por conterem diferenças de sequência que podem levar a atividades enzimáticas diferentes.

4.5.1 Genes na literatura

Ao todo 8 trabalhos publicados nos últimos 9 anos por diversos grupos da Índia, China, Alemanha e Canadá foram incluídos, os quais tratavam mais diretamente de atividades carboidrolíticas de espécies do gênero *Xanthomonas* e dos quais 5 continham indicação dos genes aqui elencados. 24 foram os genes abordados nesses trabalhos sendo que 8 são os clusters correspondentes. Isso corresponde a 10,5% dos 1084 genes da 3ª enumeração e a uma redução final para 129 genes no processo de filtragem para o posterior trabalho de caracterização molecular.

4.6 Clonagem

Um diagrama do processo de clonagem e teste enzimático encontra-se na Figura 13 e o estado da progressão de todos os genes candidato está na Tabela 22

4.6.1 Amplificações de material genômico, ligações em pGEM e transformação em DH10B

O processo de amplificação (Tabela 20) dos genes candidatos teve uma série de tentativas até se chegar numa boa condição de reação. Ao final foi possível

completar a ligação dos amplificados no vetor pGEM-T Easy e transformar a cepa DH10B com 22 genes candidatos do total de 24. Confirmações foram feitas por PCR de colônia com iniciadores ancorados no plasmídeo carreador do inserto e por sequenciamento.

Tabela 19 - Tabela de predição da localização subcelular dos candidatos para clonagem de Xop e Xoz.

Locustag	Anotação	CAZy	Extracel.	Mmb. Ext.	Perip.	Mmb. Int.	Citop.	Incert.	T3SS
xop04176	exoglucanase A	GH6~CBM37	0,5	0,1	0,4	0	0	0,4	Não
xoz00640		GH6	0,5	0,1	0,4	0	0	0,4	Não
xop02432	cellulase S	GH12	0,6	0	0,4	0	0	0,4	Não
xop02431		GH12	0,6	0	0,4	0	0	0,4	Não
xop03852	alpha-glucuronidase	GH67	0,4	0	0,5	0	0,1	0,4	Não
xoz04417		GH67	0,4	0	0,5	0	0,1	0,4	Não
xop03869	beta-xylosidase	GH43	0,4	0	0,4	0	0,2	0,4	Não
xoz04521		GH43	0,4	0	0,5	0	0,2	0,4	Não
xop01459	xylosidase; arabinosidase	GH43	0,4	0	0,6	0	0	0,4	Não
xoz01906		GH43	0,4	0	0,6	0	0	0,4	Não
xop02430	rhamnogalacturonase B	GH28	0,5	0,1	0	0	0	0,4	Não
xoz03726		GH28	0,6	0,1	0	0	0	0,4	Não
xop04177	Celulase	GH5~CBM63	0,4	0,1	0,5	0	0	0,4	Não
xoz00638		GH5~CBM63	0,5	0	0,5	0	0	0,4	Não
xop04327	biofilm PGA synthesis lipoprotein PgaB	CE4	0,4	0	0	0	0,2	0,4	Não
xoz00762		CE4	0,4	0	0	0	0,2	0,4	Não
xop00373	Pectinesterase	CE8	1	0,4	0,4	0	0	0	Não
xop04558	Xylanase	GH30	0,6	0,4	0,4	0	0	0,4	Não
xoz02153	mannan endo-1,4-beta-mannosidase	GH5	0,5	0,1	0,4	0,1	0	0,4	Não
xop04681	putative secreted protein	GH98~CBM35	0,2	0,5	0	0	0	0,4	Não
xoz00878	carbohydrate binding module (family 6)	GH98~CBM35	0,6	0,1	0	0	0	0,4	Não
xop00565	pectate lyase L	PL9	1	0	0	0	0	0	Não
xoz02478		-	1	0	0	0	0,1	0	Não
xoz02477	pectate lyase	PL9	0,9	0	0	0	0,1	0	Não

Extracel., Mmb.Ext., Perip., Mmb.Int., Citop. – Média ponderada da pontuação percentual para presença no meio extracelular, na membrana externa, no periplasma, na membrana interna ou no citoplasma. Incert. – média ponderada da pontuação percentual para incerteza de localização. T3SS – Sinal para secreção pelo sistema do tipo 3.

4.6.2 Iniciadores de PCR

Tabela 20 - Oligonucleotídeos gerados com o auxílio dos softwares Primer-BLAST e Pribnon sobre as sequências genômicas.

Locustag	Software	Forward (5'→3')	Reverse (5'→3')	Tamanho (bp)
xop03852	Primer-BLAST	GCCGACAGCGTGCCGATCAAT	CTGCATGCACGGGTGACGCT	2282
xoz04417	pribnonW9N70	CCGACAGCGTGCCGATCAGTAT	CGACTGCATGCACGGGTGAC	2283
xop04327	Primer-BLAST	GCCGTACAGCCATGCGGAC	GCGAGTTGGTCACGTCCGGT	2009
xoz00762	pribnonW8N500	TGCCGTGCCGCCAGTATGTC	CGGCTGGGTGGCAGTCAGTT	2661
xop04177	Primer-BLAST	TACCGGCCGTTCTGGGGTCG	GGCCGCTTCGCACGCATTTC	1853
xoz00638	pribnonW9N70	TACCGGCCGTTCTGGGGTCG	TCCACCCGCACAGACCGCTT	1864
xop02430	Primer-BLAST	CGATGCCCCCATCAGGCACA	CGTGCCTTACCACCGCAAGGT	1777
xoz03726	pribnonW8N500	CGACGCGTGCGCTATCCCTG	CCCCTGTTGCGGGGTCTCAC	2242
xop01459	Primer-BLAST	CGATGCCTACGCGCCCTCAG	GCACCCGCCGGGAAAGTCTC	1703
xoz01906	pribnonW8N500	CGTGCACCCACCGGGAAAGTC	AAGCGCGTGGCGGGGATATG	2001
xop00373	Primer-BLAST	GCCTGGCTTGTTGCGCAGTC	ACCTGCCCCACACGTCCCTT	1257
xop00565	Primer-BLAST	CGCCCCCTTAATCCTGCCGGG	TGCTGGGCTGTCCGTCCATAC	761
xop02431	Primer-BLAST	TGGCGCGACAATAGCCGTGT	GCAGGTGGGCTTGTTGGTCC	860
xop02432	Primer-BLAST	AGCTGCGGCTGCGCTTACTT	CGCCACTGATAGGGGGCCAACC	919
xop03869	Primer-BLAST	CGCGATGGATCTCGGCAGGC	ACCCTCTGACCTCCCCTCCCA	1111
xop04176	Primer-BLAST	CCGCTCTGTGGCAACGCAGT	ACACACCGTTTCAGCACTTAACCA	1769
xop04558	Primer-BLAST	CCGGCGGCATCCCCTTTCC	CCACGCCCACCTGACGACTC	1293
xop04681	Primer-BLAST	GCACCGCCGCTGTGGTATCG	GCACCAACCACTACGCAGGCAT	2211
xoz00640	pribnonW9N70	GACGAACTGCGCGTGGAACC	TCCTGCCACTGACGCAAGCC	921
xoz00878	pribnonW9N70	GGCACCGCCGCTGTGGTAT	TCAAGCTGACCAACACGTACGGC	1712
xoz02153	pribnonW9N70	ATGTGCGTTCGACTCAAATGGCG	CCGCTCAACCACCGACGCTTA	1125
xoz02477	pribnonW8N500	ACCGCCGTAGCAGAACCGAG	CCCGCACTGACCACAAGGGG	1288
xoz02478	pribnonW9N70	GGCAATCGATCCAACGGCCTCT	CCGCCGCCCTTCATCCTTC	504
xoz04521	pribnonW9N70	ACGCGATGGATCTCGCCAGG	GCCACCCTCTGACCTCCCCT	1114

Tabela 21 - Relação de oligonucleotídeos desenhados para subclonagem dos genes candidatos do vetor pGEM-T Easy para o vetor pET-44b.

Gene	Produto (bp)	Sequência do oligonucleotídeo <i>forward</i> (5'→3')	Sequência do oligonucleotídeo <i>reverse</i> (5'→3')
xop04558	1221	TATATACATATGAAGAAATCCTTGTCGATGTTGC	TATATATTAATTAATCAGTTGCCGACCAGTGTGG
xoz04417	2196	TATATACATATGAAGCGGCATGGCGC	TATATATTAATTAATCACGTCGGATTCCCCGG
xop03852	2196	TATATACATATGAAGCGGCATGGCGCCGAC	TATATTAATTAATCACGTCGGATTCCCCGGTGCAT
xoz01906	1587	TATATACATATGATGCGTCTGTTCTCCCGC	TATATATTAATTAATCACGGCGCATCGCAACG
xoz02478	345	TATATACATATGCAGTCCACCCTGAGCGAC	TATATATTAATTAATCAACGGGCCTCGAATGCCC
xoz03726	1602	TATATACATATGCCGCGTCATCGCCTG	TATATATTAATTAACTACAACGCCACCAGTTGGATG
xop02430	1602	TATATACATATGCCGCGTCATCGCCTGC	TAATATTAATTAACTACAACGCCACCAGTTGGATGC
xop01459	1584	TATATACATATGCGTCTGTTCTCCCGCGT	TATATATTAATTAATCACGGCGCATCGCAACG
xop04177	1773	TATATACATATGCTGTACACATCCATTTCT	TATATATTAATTAATCAGGGAAACTGTACGTGGC
xoz00638	1773	TATATACATATGCTGTACACATCCATTTCT	TATATATTAATTAATCAGGGAAACTGTACGTGGC
xoz00878	1656	TATATACATATGGTGAGCTTCACCGGCGGCGTCT	TATATATTAATTAATTACTGCGCGACGGCCAGAGTGTCC
xop03869	1035	TATATACATATGTCCGATGAGCTGCACACCGCT	TATATATTAATTAATCACCACGGCGACACCGCATCT
xoz04521	1035	TATATACATATGTCCGATGAGCTGCACACCGCT	TATATATTAATTAATCACCACGGCGACACCGCATCT
xop00373	1182	TATATACATATGTCTGAATCGCACATCCGTACG	TATATATTAATTAACTAATTCCCAGCGCCGGCG
xop02431	909	TATATACATATGGTGACGCACAGGCGTCGTCT	TATATATTAATTAATTACTTGGCTGTAATCGACCCGGTC
xop02432	591	TATATACATATGTGAACAATTTCAACGATCCCAACA	TATATATTAATTAATCAAAGCCCGCCTCGGC
xoz02477	795	TATATACATATGTGATGGCCGCGCAGG	TATATTAATTAACTACTCCTGCCACCTATGTGGTG
xop00565	654	TATATACATATGTGTTGCACTTGACTCTTGAGACCG	TATATATTAATTAATCAACGGGCCTCGATTGCCC
xop04681	2103	TATATACATATGTTTCGTGATCCATGCCGACACC	TATATATTAATTAATTACTGCGCGACGGCCAGAT
xop04327	1896	TATATACATATGCGCGTACTTGTGTTG	TATATATTAATTAACCTTGTCGGGTAAGGG
xoz00762	1896	TATATACATATGCGCGTACTTGTGTTG	TATATATTAATTAACCTTGTCGGGTAAGGG
xoz02153	987	TATATACATATGTTGAACGCACTGCTGCCTTT	TATATATTAATTAATCAACGCGGGCCGAACAC
xop04176	1660	TATATACATATGTTGCTGACAGGCGCGCTGC	TATATATTAATTAAGGCAACGGGAACGAGCAG
xoz00640	1660	TATATACATATGTTGCTGACAGGCGCGCTGC	TATATATTAATTAAGGCAACGGGAACGAGCAG

4.6.3 Subclonagem em pET-44b e BL21(DE3)

Depois de repetidas tentativas de clonagem direta do produto de PCR (Tabela 21) no vetor pET-44b optou-se pela clonagem em vetor pJET1.2, com o auxílio do kit CloneJet de clonagem de ponta cega da empresa Fermentas, e posterior subclonagem em pET-44b.

Outra dificuldade encontrada nesse processo foi a obtenção de digestões completas com as enzimas *PacI* e *NdeI*. Usualmente a quantidade de material gerado numa digestão de preparação de pequena escala do vetor de origem em uma subclonagem é suficiente para obter-se uma reação de ligação que gere colônias transformadas. No entanto, descobriu-se que a enzima *NdeI* tem particular sensibilidade a impurezas advindas dos meios de cultura utilizados usualmente para o crescimento de *E.coli* (LB, SOB, CircleGrow) e do processo de lise celular durante o processo de preparação. Foram tomadas medidas que melhoraram a digestão, de acordo com a recomendação da literatura. Foram elas: a de se executar uma pré-lavagem do precipitado de cultura com o tampão STE, manutenção das reações em gelo na maior parte possível do tempo antes da precipitação do lisado e uma etapa de purificação com uma mistura de fenol, clorofórmio e álcool isoamílico. Apesar das condições de reação utilizadas inicialmente fossem aquelas definidas pelo fabricante, foi necessário também alterá-las para 5 vezes mais enzima e até 20 h mais.

As purificações das bandas em gel de agarose da digestão também tiveram seu rendimento melhorado com a troca de um protocolo de kit, por um procedimento desenvolvido internamente no laboratório. Trata-se da montagem de uma coluna de lã acrílica para separação por centrifugação da fase líquida da banda cortada do gel, seguida de precipitação do DNA.

Tabela 22 - Relação dos 24 genes candidatos de *Xanthomonas oryzae* pv. *oryzae* e *X.o.pvoryzicola*.

Locustag	Status	DNA (bp)	Prt. (kDa)	Descrição da anotação	CAZy
xop04176	pGEM	1670	57,8	exoglucanase A	GH6~CBM37
xoz00640	pGEM	866	30,4		GH6
xop02431	Proteína	909	32,6	cellulase S	GH12
xop02432	pGEM	591	21,8		GH12
xop03852	pGEM	2196	80,4	alpha-glucuronidase	GH67
xoz04417	Genômico	2196	80,5		GH67
xop03869	pJET	1035	37,8	beta-xylosidase	GH43
xoz04521	Proteína	1035	37,7		GH43
xop01459	Proteína	1584	57,6	xylosidase; arabinosidase	GH43
xoz01906	pGEM	1587	58		GH43
xop02430	pJET	1602	57,2	rhamnogalacturonase B	GH28
xoz03726	pJET	1602	57,1		GH28
xop04177	Enzima	1773	61,9	cellulase	GH5~CBM63
xoz00638	Enzima	1773	62		GH5~CBM63
xop04327	pGEM	1898	71,3	biofilm PGA synthesis lipoprotein PgaB	CE4
xoz00762	pGEM	1922	72,2		CE4
xop00373	Proteína	1182	40,5	pectinesterase	CE8
xop04558	pGEM	1221	43,5	xylanase	GH30
xoz02153	Proteína	987	35,3	mannan endo-1,4-beta-mannosidase	GH5
xop04681	Genômico	2103	76,5	putative secreted protein	GH98~CBM35
xoz00878	Genômico	1656	59,5		GH98~CBM35
xop00565	pET-44b	654	23,2	pectate lyase L	PL9
xoz02478	pET-44b	345	12		-
xoz02477	pET-44b	795	28,3	pectate lyase	PL9

A tabela descreve os genes candidatos, suas características e na coluna dois em que etapa se encontra. Na ordem de progressão: Genômico, pGEM, pJET, pET+BL21, Proteína e Enzima. O estado de progressão de um gene denota que aquele candidato não progrediu para etapa seguinte do processo.

Figura 13 – (p. anterior e esta) Diagrama do processo de clonagem e expressão de genes candidatos. Ao lado de cada bloco correspondente a um vetor de clonagem encontra-se um exemplo de teste ou etapa bem sucedido. Um painel de imagens dos géis encontra-se no APÊNDICE A -.

4.7 Indução de expressão heteróloga e SDS-PAGE

As melhores condições de expressão encontradas permitiram a verificação da produção heteróloga de 7 dos 10 clones induzidos, porém com a provável formação exclusiva de corpos de inclusão para 5 deles. Isso porque não foi possível verificar nenhuma atividade enzimática para esses clones. O tamanho verificado em SDS-PAGE para os outros dois clones, xop04177 e xoz00638, foi o dobro do predito suscitando a possibilidade de que essas proteínas são capazes de formar dímeros estáveis sob a condição denaturante de SDS.

4.8 Testes enzimáticos

4.8.1 Atividade endoglunásica sobre 4-nitrofenil- β -D-sacarídeos

O teste de comparação de médias de Tukey (Figura 14) sobre os dados das 9 réplicas do experimento mostrou que apenas no caso das duas celulasas homólogas xop04177 e xoz00638 sobre β -glucoronano (Tabela 23) houve uma diferença significativa da média desses experimentos em relação ao restante. O controle desse experimento foi feito usando-se o clone xop00565 pois esse não exibiu sinal da presença de nenhuma proteína heteróloga no SDS-PAGE. Houve outros casos com diferenças significativas de médias em relação ao controle e alguns outros clones, em particular no caso dos mesmos clones já mencionados e do clone xoz04521 sobre α -arabinano, porém nenhum deles houve uma distinção tão clara, isto é, em dois grupos não sobrepostos.

Dessa forma, podemos dizer que a anotação original desses dois clones como celulase talvez seja mais bem descrita como endo- β -glucoronidase ou, se levado em conta o experimento sobre α -arabinano, endo-glucoronoarabinoxilanase, uma xilanase. O domínio catalítico desses dois genes homólogos se enquadrou na categoria de glicosil hidrolases do tipo 5 do CAZy o que ainda não permite determinar o substrato com grande certeza já que esse tipo de domínio GH pode ser associado a todos as categorias de PCWPS descritas.

No entanto, o domínio CBM do tipo 63 presente na proteína aponta para uma direção diferente. Esse CBM está em geral associado com celulose, mananos e

xiloglucanos, o que seria condizente com a clivagem de resíduos de arabinose na cadeia lateral de xiloglucanos, mas excluiria a ação sobre xilanos.

Tabela 23 - Substratos sintéticos com um grupamento 4-Nitrophenyl (Sigma Chem. Co. St. Louis, USA) que pode identificar uma ação enzimática específica dependendo do açúcar e suas respectivas concentrações utilizadas no ensaio.

Nomenclatura do substrato	Sigla	Concentração do substrato no ensaio	Descrição
4-Nitrophenyl α -D-galactopiranoside	a-gal	25 mM	hidrólise de mananos
4-Nitrophenyl β -D-glucopiranoside	b-glc	25 mM	hidrólise de celulose
4-Nitrophenyl α -D-xilopiranoside	a-xil	25 mM	hidrólise de xiloglucano
4-Nitrophenyl α -L-arabinofuranoside	a-araf	10 mM	hidrólise de pectinas (arabinano)
4-Nitrophenyl α -L-arabinopiranoside	a-arapi	10 mM	hidrólise de pectinas (arabinano)
4-Nitrophenyl α -L-fucopiranoside	a-fuc	25 mM	hidrólise de xiloglucano
4-Nitrophenyl β -D-cellobioside	b-cel	25 mM	hidrólise de celulose
4-Nitrophenyl β -D-galactopiranoside	b-gal	25 mM	hidrólise de pectinas (galactanos)
4-Nitrophenyl β -D-glucuronide	b-glr	25 mM	hidrólise de arabinoxilanos e pectinas
4-Nitrophenyl β -D-xilopiranoside	b-xil	25 mM	hidrólise de arabinoxilanos e pectinas
4-Nitrophenyl β -D-mannopiranoside	b-man	10 mM	hidrólise de mananos

Cada substrato possui um nome em relação à enzima que pode estar agindo na hidrólise deste e a atividade da enzima é quantificada pela liberação do produto durante a reação. a=alfa e b=beta.

Figura 14 - Painele de absorvâncias de PNP. Cada bloco se refere ao teste com um dos 11 PNP-sacarídeos para os oito clones com proteína produzida. Foram feitas 9 replicas e os dados submetidos ao teste de Tukey para distinção entre clones para cada PNP. BP e BZ indicam clones dos genes xop de *Xo oryzae* PXO99a e xoz de *Xo oryzae* BLS256.

4.8.2 Atividade sinérgica sobre hemiceluloses e extratos de parede celular

O ensaio de digestão e detecção de oligossacarídeos por HPAEC-PAD foi feito usando-se uma mistura do lisado dos 8 clones do ensaio anterior sobre 18 extratos e polissacarídeos de paredes celulares vegetais. Na maioria dos experimentos houve problemas para obter concentrações iguais de polissacarídeo no tratamento e no controle. No caso do experimento com xilano de faia pudemos obter curvas comparáveis (Figura 15). As diferenças dos picos de monossacarídeo e de dissacarídeo entre tratamento e controle não foram consideradas relevantes frente aos parâmetros determinados de amostragem para o ensaio, sendo necessárias repetições do experimento. Se no entanto o aparecimento do monossacarídeo se tornar relevante, podemos dizer que se trata de resíduos de arabinose ou glucoronato retirados das cadeias laterais pelas glucoronoarabinoxilanases presentes na mistura. Não há no entanto o aparecimento de oligômeros da cadeia principal, indicando que a xilosidase xoz04521 não foi capaz de hidrolisar a cadeia principal desse xilano.

Figura 15 - Curva de carga por tempo de retenção em HPAEC-PAD de xilano de faia tratado com lisado dos 8 clones por 72 h a 37 °C. À esquerda a curva para o tratamento e a direita o controle. O pico a 42 min corresponde ao polissacarídeo intacto, o pico de 7 min corresponde a um dissacarídeo não identificado, a 4 min o pico de um monossacarídeo não identificado e a 2 min um contaminante não identificado do processo de lise. O pico de 2 min está presente em todas as outras curvas não mostradas.

5 CONCLUSÕES

Nesse projeto procuramos estabelecer um processo de validação de anotação de genes PCWDE enumerando-os e prevendo seus substratos *in silico* e testando suas atividades enzimáticas *in vitro*. Procuramos também delinear uma provável história evolutiva das enzimas de *Xanthomonadaceae* pertencentes a esse aparato de catabolismo.

Cerca de 20% dos genes enumerados estavam anotados como proteínas hipotéticas. No restante, alguns genes tinham uma anotação ainda superficial ou defasada e havia grande descompasso no vocabulário utilizado para descrição das anotações dos genes. Algumas dificuldades no início da enumeração foram encontradas em casos em que encontramos elementos de transposição fusionados aos genes de PCWDE, especialmente em *X. oryzae*.

Com a otimização do processo de enumeração eliminamos genes com indícios de ação sobre não-PCWPS, especialmente genes relacionados aos processos da parede celular bacteriana. Apesar disso, dada à complexidade dos PCWPS, procuramos abranger ao máximo as PCWDE correspondentes, talvez até abarcando algumas das cazys não-PCWDE, sendo que há partes compartilhadas da composição de PCWPS com outros polissacarídeos.

Desenvolvemos um método que liga os termos encontrados na literatura a modelos estatísticos que refletem a conservação da estrutura terciária de proteínas PCWDE. Para se adquirir a sensibilidade alcançada, os bancos de dados do projeto CAZy foram de grande valia, especialmente na forma de HMMs montados no projeto paralelo dbCAN. CAZy e dbCAN são em conjunto, em seu estado atual, um grande exemplo do potencial que os HMM tem a oferecer.

Apesar da riqueza de números EC recuperados do banco da IUBMB encontramos uma limitação no número de sequências moleculares de referência no banco do KEGG. Isso provavelmente não reflete o estado atual do KEGG, pois se descobriu posteriormente que o projeto foi fechado para acesso pago por falta de verbas de pesquisa, sendo disponibilizada publicamente uma versão que se torna mais defasada a cada dia.

Uma característica das PCWDE que se tornou mais clara ao longo do desenvolvimento é a inespecificidade de ação perante diferentes substratos, que também pode ser averiguada nos testes enzimológicos. Uma consulta ao site

BRENDA a respeito de PCWDE já caracterizadas, mostra que de fato podem existir diversos substratos para uma mesma enzima, porém talvez com diferentes níveis de afinidade. Por outro lado, há restrições de ação das enzimas envolvendo o micro ambiente próximo contendo ramificações e outras modificações existentes nos PCWPS.

A diferença no número de genes PCWDE levantados para *Xanthomonas* e *Xylella* está provavelmente correlacionada com a diferença de estilo de vida observada para esses dois patógenos. Enquanto uma pode ser encontrada na forma de vida livre no ambiente, a outra fica restrita ao xilema de seu hospedeiro e ao trato digestivo do inseto vetor. Já *Stenotrophomonas* ainda tem laços fortes com seus ancestrais que provavelmente tinham um contato bastante próximo com plantas. O espectro de ação de *Xylella* parece ser voltado para celulose, mananos, xiloglucanos e β 1-3,1-4 glucanos enquanto o ancestral de *Stenotrophomonas* aparenta ter tido maior contato com xilanos, galacturonanos, homogalacturonanos e ramnogalacturonano-I. *Xanthomonas* não aparenta ter uma tendência no seu potencial de hidrólise. Talvez por isso, e pelo grande número de genes, seja tão mais capaz de macerar o tecido de seus hospedeiros.

O grupo mais abrangente de genes que compõem o aparato de PCWDE deve ter ocorrido em concerto próximo da derivação de *Xanthomonas axonopodis citri*, contando com alguns eventos de HGT que ocorreram antes da divergência entre *Xanthomonas* e *Stenotrophomonas* e após divergência de *Xylella*. Alguns eventos isolados de HGT possivelmente também ocorreram nas *Xanthomonas* mais basais (*X. sachari*, *X. vesicatoria* e *X. glycines*) que não são compartilhados com quase nenhum outro genoma. *X. campestris* teve duplicações recentes que devem contribuir para o fenótipo de maceração mais agressiva. Algumas diferenças mais sutis de polimorfismos foram implicadas na adaptação à arquitetura e composição dos diferentes PCWPS de cada hospedeiro.

Estabelecemos também critérios de seleção de genes candidatos para testes enzimáticos com o intuito de contribuir para a anotação e para o levantamento com fins tecnológicos de uso dessas enzimas. Foram selecionados 24 genes candidatos dos quais 3 permaneceram no início do processo, 11 foram clonados a partir do material genômico, 3 puderam ser subclonados no vetor de expressão. 5 produziram proteína e 2 demonstraram atividade enzimática. Pudemos ao final fazer a refino da anotação de dois genes ortólogos de *Xanthomonas oryzae* pv. *oryzae* e

Xanthomonas oryzae pv. *oryzicola* (xop04177 e xoz00638). Originalmente eram descritos como celulase, uma anotação pouco específica, mas após os ensaios verificamos que se tratava de uma endo-glucoronoarabinoxilanase. Os processos de enumeração e predição de substrato foram validados, permitindo também um refino. Apesar de esses genes terem sido recuperados por serem glicosil hidrolases do tipo 5, com um amplo espectro de atuação sobre os PCWPS, o resultado dos ensaios indicou a existência de um subtipo de GH5 com atuação específica sobre arabinoxilano.

Com o advento das tecnologias de sequenciamento massivo de DNA ganhando momento à mesma época do início do projeto, foi natural a incorporação de vários genomas ao longo do projeto. Ao iniciarmos as primeiras análises contávamos com cerca de 10 genomas disponíveis publicamente em diferentes níveis de fechamento e anotação. Já a última análise incorporou 36 genomas com vistas a dezenas ainda por virem a público em um período de poucos meses. Tornando-se o completo sequenciamento de genomas um evento cada vez mais trivial muda-se a restrição do processo experimental da leitura do DNA para sua interpretação e para as derivações dessa interpretação. Isto é, para a bioinformática e para a biologia molecular, aonde esse trabalho vem, portanto, a contribuir.

REFERÊNCIAS*

- Aparna G, et al. A cell wall-degrading esterase of *Xanthomonas oryzae* requires a unique substrate recognition module for pathogenesis on Rice. *The Plant Cell*. 2009;21:1860-73.
- Carpita NC, Gibeaut DM. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J*. 1993;3(1):1-30.
- Da Silva ACR, et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*. 2002;417:459-63.
- Darvill JE, et al. Structure of Plant Cell Walls - XI. Glucuronoarabinoxylan, a second hemicellulose in the primary cell walls of suspension-cultured sycamore cells. *Plant Physiol*. 1980;66:1135-9.
- De Vries RP. Synergy between enzymes from *Aspergillus* involved in the degradation of plant cell wall polysaccharides. *Carbohydrate Research*. 2000;327:401–10.
- Dongen SV. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*. 2008;30(1):121-41.
- Encarnação TBC. Modulação da degradação enzimática de galactomanano por sua própria estrutura fina. [dissertação (Mestrado em Ciências Biológicas)]. São Paulo: Instituto de Biociências, Universidade de São Paulo; 2012.
- Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle; 2005. [cited from 2008 Jun 11]. Available from: evolution.genetics.washington.edu/phylip.html.
- Gardy JL. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*. 2005;21(5):617-23.
- Gilbert HJ, Stalbrand H, Brumer H. How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation. *Current Opinion in Plant Biology*. 2008;11:338–48.
- Hsiao YM, et al. Regulation of the *pehA* gene encoding the major polygalacturonase of *Xanthomonas campestris* by Clp abnd RpfF. *Microbiology*. 2008;154:705-13.
- Hsieh YSY, Harris PJ. Xyloglucans of Monocotyledons Have Diverse Structures. *Molecular Plant*. 2009;2(5):943–65.

*De acordo com: International Committee of Medical Journal Editors. [Internet]. Uniform requirements for manuscripts submitted to Biomedical Journal: sample references. [updated 2011 Jul 15]. Available from: <http://www.icmje.org>

Hu J, Qian W, He C. The *Xanthomonas oryzae* pv. *oryzae* eglXoB endoglucanase gene is required for virulence to Rice. FEMS Microbiol Lett. 2007;269:273-9.

Imai K, et al. SOSUI-GramN: high performance prediction for subcellular localization of proteins in Gram-negative bacteria. Bioinformatics. 2008;2(9):417-21.

Kamoun S, Kado CI. Phenotypic switching affecting chemotaxis, xanthan production, and virulence in *xanthomonas campestris*. Appl Environ Microbiol. 1990;56(12):3855-60

Kirchberg J, Büttner D, Thiemer B, Sawers RG. Aconitase B is required for optimal growth of *Xanthomonas campestris* pv. *vesicatoria* in pepper plants. PLoS One. 2012;7(4):e34941.

Lagaert S, Beliën T, Volckaert G. Plant cell walls: Protecting the barrier from degradation by microbial enzymes. Seminars in Cell & Developmental Biology. 2009;20:1064–73.

Lee BM, et al. The genome sequence of *Xanthomonas oryzae* pathovar *oryzae* KACC10331, the bacterial blight pathogen of rice. Nucleic Acids Research. 2005;33(2):577–86.

Lee ST, Lee JJ. Insoluble dye substrate for screening and assay of xylan-degrading enzymes. Journal of Microbiological Methods. 1997;29:1-5.

Li L, Stoeckert Jr. CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003;13:2178-89.

Lima WC, et al. Non-gamma-proteobacteria gene islands contribute to the *Xanthomonas* genome. OMICS. 2005;9(2):160-72.

Lin RJ, Capage M, Hill CW. A repetitive DNA sequence, rhs, responsible for duplications within the *Escherichia coli* K-12 chromosome. Journal of Molecular Biology. 1984;177(1):1–18.

Löwer M, Schneider G. Prediction of type III secretion signals in genomes of gram-negative bacteria. PLoS One. 2009;4(6):e5917.

Lu Z, et al. Predicting subcellular localization of proteins using machine learned classifiers. Bioinformatics. 2004;20(4):547–56.

Lu H, et al. Acquisition and evolution of plant pathogenesis-associated gene clusters and candidate determinants of tissue-specificity in *Xanthomonas*. PLoS One. 2008;3(11):e3828

Moreira LRS, Filho EXF. An overview of mannan structure and mannan-degrading enzyme systems. Appl Microbial Biotechnology. 2008;79:165–78.

Moreira LM, et al. Comparative analyses of *Xanthomonas* and *Xylella* complete genomes. OMICS. 2005;9(1):43-76.

Moreira LM, et al. Comparative genomics analyses of citrus-associated bacteria. *Annu Rev Phytopathol.* 2004;42:163–84.

Ochiai H, et al. Genome sequence of *Xanthomonas oryzae* pv. *oryzae* suggests contribution of large numbers of effector genes and insertion sequences to its race diversity. *JARQ.* 2005;39(4):275–87.

Pradhan BB, Ranjan M, Chatterjee S. XadM, a novel adhesin of *Xanthomonas oryzae* pv. *oryzae*, exhibits similarity to Rhs family proteins and is required for optimum attachment, biofilm formation and virulence. *Mol Plant Microbe Interact.* 2012;25(9):1157-70.

Rademaker JLW, et al. A comprehensive species to strain taxonomic framework for *Xanthomonas*. *Phytopathology.* 2005;95:1098-111.

Rajeshwari R, Jha G, Sonti RV. Role of an in planta-expressed xylanase of *Xanthomonas oryzae* pv. *oryzae* in promoting virulence on rice. *Molecular Plant-Microbe Interactions.* 2005;18(8):830-7.

Ramos PL, et al. Screening for endophytic nitrogen-fixing bacteria in Brazilian sugar cane varieties used in organic farming and description of *Stenotrophomonas pavanii* sp. nov. *International Journal of Systematic and Evolutionary Microbiology.* 2011;61:926–31.

Ray SK, Rajeshwari RR, Sonti RV. Mutants of *Xanthomonas oryzae* pv. *oryzae* deficient in General Secretory Pathway are virulence deficient and unable to secrete xylanase. *Mol Plant Microbe Interact.* 2000;13(4):394-401.

Ridley BL, O'Neill MA, Mohnen D. Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. *Phytochemistry.* 2001;57:929–67.

Sambrook J, MacCallum P, Russell D. *Molecular cloning: a laboratory manual.* 3rd ed. New York: CSH Press; 2001.

Scheller HV, Ulvskov P. Hemicelluloses. *Annu Rev Plant Biol.* 2010;61:263–89.

Schröter K, et al. *Xanthomonas campestris* pv. *campestris* secretes the endoglucanases ENGXCA and ENGXCB: construction of an endoglucanase-deficient mutant for industrial xanthan production. *Appl Microbiol Biotechnol.* 2001;55:727-33.

Simpson AJ, et al. The genome sequence of the plant pathogen *Xylella fastidiosa*. *Nature.* 2000;406:151-60.

Stamatakis A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 2006;22(21):2688-90.

Strange RN, Scott PR. Plant Disease: A Threat to Global Food Security. *Annu Rev Phytopathol.* 2005;43:83–116.

Sun QH, et al. Type-II secretion pathway structural gene xpsE, xylanase and cellulose secretion and virulence in *Xanthomonas oryzae* pv. *oryzae*. Plant Pathology. 2005;54:15-21.

Suzek BE, et al. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics. 2001;17:1123–30.

Ten LN, et al. Development of a plate technique for screening of polysaccharide-degrading microorganisms by using a mixture of insoluble chromogenic substrates. J. Microbiol. Methods. 2004;56(3):375-82.

Thieme F, et al. Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium *Xanthomonas campestris* pv. *vesicatoria* revealed by the complete genome sequence. Journal of bacteriology. 2005;187(21):7254–66.

Toth IK, Pritchard L, Birch PRJ. Comparative genomics reveals what makes an enterobacterial plant pathogen. Annual Review of Phytopathology. 2006;44:305-36.

Untergasser A, et al. Primer3 - new capabilities and interfaces. Nucleic. Acids. Res. 2012;40(15):e115.

Van Sluys MA, et al. Comparative analyses of the complete genome sequences of Pierce's Disease and Citrus Variegated Chlorosis strains of *Xylella fastidiosa*. Journal of Bacteriology. 2003;185(3):1018-26.

Van Sluys MA, et al. Comparative genomic analysis of plant-associated bacteria. Annu. Rev. Phytopathol. 2002;40:169–89.

Vincken JP, et al. If homogalacturonan were a side chain of rhamnogalacturonan i. implications for cell wall architecture. Plant Physiology. 2003;132:1781–9.

Vorhölter FJ, et al. Comparison of two *Xanthomonas campestris* pathovar *campestris* genomes revealed differences in their gene composition. Journal of Biotechnology. 2003;106:193–202.

Wolf A, et al. *Stenotrophomonas rhizophila* sp. nov., a novel plant-associated bacterium with antifungal properties. International Journal of Systematic and Evolutionary Microbiology. 2002;52:1937–44.

Xiao Z, et al. Mining *Xanthomonas* and *Streptomyces* genomes for new pectinase-encoding sequences and their heterologous expression in *Escherichia coli*. Appl. Microbiol. Biotechnol. 2008;78:973-981.

Yamada T, et al. Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. Molecular Systems Biology. 2012;8:581.

Ye J, et al. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;18:13-134.

Yin Y, et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*. 2012;40:1–7.

Yu C, et al. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci*. 2004;13:1402–06

Zhang D, et al. Polymorphic toxin systems: comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biol Direct*. 2012;7(1):18.

APÊNDICE A - Géis de eletroforese

Figura 16 – (Próxima página) Géis de agarose do processo de subclonagem do vetor pGEM-T Easy para o vetor pET-44b. Referência de peso molecular (bp) no terceiro gel na base da figura. Géis de agarose com concentração entre 0,5 e 0,7% em TAE 0,5X.

Dois primeiros géis no topo: Produtos de PCR com iniciadores para clonagem sítio dirigida em vetor pET. PCR com bandas de tamanho esperado para todos os candidatos com exceção de xoz01906. Candidatos com bandas positivas (17): xop03852, xoz04417, xop03869, xoz04521, xop01459, xop02430, xoz03726, xop04177, xoz00638, xop00373, xoz02153, xop00565, xoz02477, xoz02478, xop04558, xop02432, xop02431.

Terceiro e quarto gel do topo: teste sobre DNA de clones no vetor pET-44b. Terceiro: digestão dupla com enzimas *PacI* e *NdeI*. Quarto: PCR com os iniciadores desenhados. Marcados apenas os clones positivos nos testes (10): xoz04521, xop01459, xop04177, xoz00638, xop00373, xoz02153, xop00565, xoz02477, xoz02478, xop02431.

Restante dos géis: Testes de clones de candidatos em pJET por PCR com os iniciadores desenhados. Marcados apenas os clones positivos nos testes (15): xoz04417, xop03869, xoz04521, xop01459, xop02430, xoz03726, xop04177, xoz00638, xop00373, xoz02153, xop00565, xoz02477, xoz02478, xop02432, xop0243.

APÊNDICE B - Código de computador gerado

blastCORE

```
#!/bin/bash

FASTA=$1

if ! [[ -r $FASTA ]];
then
    echo "File $FASTA does not exist or is not readable.";
    exit 1;
fi

if [[ -z $THREADS ]]
then
    echo "Number of threads set to 4. Could not find \ $THREADS value.";
    THREADS=4;
fi

NAME=`echo $FASTA | perl -pe "s/\.\w+$///;"`

echo -n "Started ${NAME}_L${COVERAGE}_S${SCORE}, "
echo -n `date`
echo -n ", on "
echo "${HOSTNAME}."

#create and access the work directory
mkdir ${NAME}
cd ${NAME}
#mkdir ${NAME}_L${COVERAGE}_S${SCORE}
#cd ${NAME}_L${COVERAGE}_S${SCORE}

#Should iteratively run blastclust depending on CORE statistics
feedback
# Write cluster number and locustag into separate lines,
echo "Started blastclust run."
COVERAGE=5;
SCORE=5;
while [[ ! $DONE ]];
do
    echo
    echo "COVERAGE 0.$COVERAGE SCORE $SCORE"
    blastclust -L 0.$COVERAGE -S $SCORE -a $THREADS -v
    blclust_L0.${COVERAGE}_S${SCORE}.log -i ../$FASTA -e F\
    2>> blclust_L0.${COVERAGE}_S${SCORE}.error.log\
    | cat -n -\
    | perl -ne 'chomp; split; $n=shift @_; foreach (@_){print
"$n\t$_\n";}
' > Genelist_L0.${COVERAGE}_S${SCORE}.tsv
    echo "Finished blastclust run."

#Retrieve fasta for each cluster
perl -e '
    open LIST, "$ARGV[0]" or die;
    while(<LIST>){
        chomp;
```

```

        split /\t/;
        @{$list{$_[1]}}=($_[0]);
    }

    open FAA, "$ARGV[1]" or die ("FAA $!\n");
    $/=">";
    <FAA>;
    while(<FAA>){
        chomp;
        $_=">".$_.$_;
        />(\S+)/;
        if (exists $list{$1}){
            open FILE, ">>", "${list{$1}}[0].faa" or die
("$${list{$1}}[0].faa : $!");
            print FILE;
        }
    }
    ' Genelist_L0.${COVERAGE}_S${SCORE}.tsv ../$FASTA

# Run T-Coffee
echo "Running t_coffee for all fastas."
for f in `ls *faa | sort -g`;
do
    bn=`basename $f .faa`;
#     for p in `ps hocmd -C t_coffee | cut -s -d ' ' -f 4 | sort |
uniq`;
#     do
        while [[ ((`ps hocmd -C t_coffee | sort | uniq | wc -
l`>=$THREADS)) ]];
        do
            sleep 1;
            echo "While wait.";
        done
#         if [[ $p =~ "$f" ]];
#         then
#             echo "$f already running!";
#         else
#             if [[ -e $bn.aln ]];
#             then
#                 echo "$f already ran!";
#             else
#                 ../../t_coffee -quiet $bn.tclog -infile=$f -
output=clustalw,score_html,score_ascii 2>>align.log &
#                 break;
#                 fi
#             fi
#         done
##         if [[ ((`ps hocmd -C t_coffee | sort | uniq | wc -
l`<$THREADS)) ]];
##         then
##             mkdir $f.dir
##             mv $f $f.dir
##             cd $f.dir
##             ../../t_coffee -mode quickaln -quiet -infile=$f -
output=score_ascii &
##             ../../t_coffee -mode quickaln -quiet -infile=$f -
output=score_ascii;
##             cd ..;

```

```

##         fi
done 2>align.log

##   while [[ (`ps hocmd -C t_coffee | sort | uniq | wc -l`>0)) ]];
##   do
##       echo "Final wait.";
##       sleep 1;
##   done
##   echo "Last faas done.";

##   for f in [0-9]*.dir;
##   do
##       mv $f/* .;
##       rmdir $f;
##   done

#Statistics
CLUSTERS=`ls *.faa | wc -l`;
CORE=`grep -h SCORE *score_ascii | perl -ne '/\d+//; $s++ if $& >=
70; print 100*$s/'$CLUSTERS' if eof;`
echo "Overall CORE: $CORE" >> Genelist_L0.${COVERAGE}_S${SCORE}.tsv
if [[ (($CORE>95)) ]]
then
    DONE=1;
else
    rm align.log
    rm [1-9]*.{faa,score_ascii}
    rm blclust*log
    rm error.log
    COVERAGE=$((COVERAGE+5))
    SCORE=$((SCORE+5))
#       if [[ (($SCORE>100)) ]]
#       then
#           exit 0;
#       fi
    fi
done

```

Pruner – compatibiliza análise par-a-par com o treedist

```

#!/usr/bin/perl

use IO::Handle;
use File::Copy;

while(<>){
    chomp;
    split /\t/;
    $query=shift @_;
    $target=shift @_;

    $/=">";
#   open FASTA, "-|", <cat fasta/$query.fasta> or die($?);
    open FASTA, "-|", <hadoop dfs -cat
s3n://gatelab.pruner/fasta/$query.fasta> or die($?);
    <FASTA>;
    while(<FASTA>){
        chomp;

```

```

    $_=">".$_;
    s/>([A-Za-z]+[0-9]*)\\|([\\w.]+)\\n/>$1\\n/;
    $index{$query}{$1}=$_ unless exists $index{$query}{$1};
}
close FASTA;
#   open FASTA,,"-|",<cat fasta/$target.fasta> or die($?);
open FASTA,,"-|",<hadoop dfs -cat
s3n://gatelab.pruner.fasta/$target.fasta> or die($?);
<FASTA>;
while(<FASTA>){
    chomp;
    $_=">".$_;
    s/>([A-Za-z]+[0-9]*)\\|([\\w.]+)\\n/>$1\\n/;
    $index{$target}{$1}=$_ unless exists $index{$target}{$1};
}
close FASTA;
$/"\\n";

%common=();
foreach $one (keys %{$index{$query}}){
    undef $common{$one} if exists $index{$target}{$one};
}
#   print STDERR join "\\t",keys %common;
#   print STDERR "\\n";

#####retrieving#####
#####
    foreach $species (keys %common){
        open QUERY, ">>", "${query}_${target}.match" or die
("${query}_${target}.match:$!");
        print QUERY $index{$query}{$species};
        open TARGET, ">>", "${target}_${query}.match" or die
("${target}_${query}.match:$!");
        print TARGET $index{$target}{$species};
    }
#####aligning#####
#####
    system "muscle -in ${target}_${query}.match -out
${target}_${query}.aln 1>&2" and die
("${target}_${query}.clustalw:$!");
#   system "t_coffee -infile=${target}_${query}.match 1>&2" and die
("${target}_${query}.clustalw:$!");
    system "clustalw -infile=${target}_${query}.aln -convert -
output=PHYMLIP -outfile=${target}_${query}.phy 1>&2" and die
("${target}_${query}.phy:$!");
    system "muscle -in ${query}_${target}.match -out
${query}_${target}.aln 1>&2" and die
("${query}_${target}.clustalw:$!");
#   system "t_coffee -infile=${query}_${target}.match 1>&2" and die
("${query}_${target}.clustalw:$!");
    system "clustalw -infile=${query}_${target}.aln -convert -
output=PHYMLIP -outfile=${query}_${target}.phy 1>&2" and die
("${query}_${target}.phy:$!");
#####treeing#####
#####
#   system "raxml -f a -x 12345 -p 12345 -# 100 -m PROTGAMMAJTT -s
${target}_${query}.phy -n ${target}_${query} -T 2 1>&2" and die
("RAXML_info.${target}_${query}:$!");

```

```

#      system "raxml -f a -x 12345 -p 12345 -# 100 -m PROTGAMMAJTT -s
${query}_${target}.phy -n ${query}_${target} -T 2 1>&2" and die
("RAxML_info.${query}_${target}:$!");
      system "raxmlHPC -f a -x 12345 -p 12345 -# 100 -m PROTGAMMAJTT -s
${target}_${query}.phy -n ${target}_${query} 1>&2" and die
("RAxML_info.${target}_${query}:$!");
      system "raxmlHPC -f a -x 12345 -p 12345 -# 100 -m PROTGAMMAJTT -s
${query}_${target}.phy -n ${query}_${target} 1>&2" and die
("RAxML_info.${query}_${target}:$!");
#####disting#####
#####
      mkdir "${query}_${target}.dir",0770 or die
("${query}_${target}",$!);
      copy
"RAxML_bipartitions.${query}_${target}", "${query}_${target}.dir/intree"
or die ("intree:$!");
      copy
"RAxML_bipartitions.${target}_${query}", "${query}_${target}.dir/intree2
" or die ("intree2:$!");
      chdir "${query}_${target}.dir" or die;
      system("/usr/lib/phylib/bin/treedist 1>&2 <<word
2
c
s
y
word") and die;
      open SPLIT,"outfile" or die;
      $distance=<SPLIT>;
      $distance=~s/^1 //;
      close SPLIT;
      print "${query}\\t${target}\\t$distance";
      chdir ".." or die;
}

```

Glipfammer – 3ª estratégia de enumeração

```

#!/bin/bash

if [[ -z $1 && -z $2 ]]
then
    echo "
Usage ./glipfammer.sh FASTA TAG
Where:
FASTA = Nucleotide fasta or multifasta of sequence assembly to be ORF
predicted.
TAG    = Tag to be used for gene locus tag. Should be a three letter
acronym.
"
    exit 1
fi

echo -n "Started Glimmer at"
echo -n `date`
echo "."

```

```

fastaexplode $1
for f in *fa; do tigr-glimmer long-orfs $f $f.long; done
for f in *fa; do tigr-glimmer extract $f $f.long > $f.seq; done
for f in *seq; do tigr-glimmer build-icm $f.icm < $f; done
for f in *fa; do tigr-glimmer glimmer3 $f $f.seq.icm $f; done
for f in *fa; do tigr-glimmer extract $f $f.predict > $f.ffn; done
for f in *ffn; do bn=`basename $f .ffn`; ./translate.pl -g 11 $f >
$bn.faa; done
for f in *faa; do if [[ ! -s $f ]]; then rm $f; fi; done
for f in *faa; do perl -pe 'BEGIN{$ARGV[0]=~/([0-9]+)\D*$/; $cont=$1;
$/=">"; <>;} chomp; $_=">".$_; s/>orf([0-9]+).+>/'$2'$cont$1/;' $f;
done > $2.aa
mv $2.aa $2.faa
echo -n "Finished Glimmer at"
echo -n `date`
echo "."

```

```

echo -n "Started hmmpfam at"
echo -n `date`
echo "."
hmmpfam --compat -E 1e-01 Pfam_ls $2.faa > $2.hmm.out
echo -n "Finished hmmpfam at"
echo -n `date`
echo "."

```

```

echo -n "Started pfam result parsing at"
echo -n `date`
echo "."
./pfam_process_threads.2.0.sh $2.hmm.out $2.faa
echo -n "Finished pfam result parsing at"
echo -n `date`
echo "."

```

```

echo -n "Started candidate selection at"
echo -n `date`
echo "."
perl -e '
open FAM,"pfam_filtered.fam" or die ("Could not open pfam_filtered.fam:
$!");
while(<FAM>){
    chomp;
    undef $fams{$_};
}

open ARCH,"'$2'.hmm.out.arch.tsv";
while(<ARCH>){
    chomp;
    split /\t/;
    @arch=split /\~/,$_[1];
    foreach $dom (@arch){
        if(exists $fams{$dom}){
            undef $candidates{$_[0]};
            last;
        }
    }
}

```



```

}

open FAA,"'$2'.faa";
$/=">";
<FAA>;
while(<FAA>){
    chomp;
    $_=">".$_._;
    />(.)\n/;
    print if exists $candidates{$1};
}
' > $2.candidates.faa

grep \> $2.candidates.faa \
| perl -pe 's/\d//g; s/\>//;' \
| sort \
| uniq \
| perl -e '
    while(<>){
        chomp;
        undef $star{$_};
    }
    open CAN,"'$2'.candidates.faa" or die;
    $/=">";
    <CAN>;
    while(<CAN>){
        chomp;
        $_=">".$_._;
        />([A-Za-z]+)\n/;
        print if not exists $star{$1};
    }
'

echo -n "Finished candidate selection at"
echo -n `date`
echo "."

```

Reformat.single.entry.to.matrix – PHROG ou transformando um TSV numa matriz

```

#!/usr/bin/perl

while(<>){
    chomp;
    split /\t/;
    $index{$_[2]}{$_[1]}=$_[0];
    undef $tags{$_[1]};
}

print "geno/subst\t";
print join "\t", (sort keys %tags);
print "\n";
foreach $subst (sort keys %index){
    print "$subst";
    foreach $geno (sort keys %tags){

```

```

        print "\t$index{$subst}{$geno}";
    }
    print "\n";
}

```

Pfam_process_threads – paralelizador do *parser* da saída do HMMER

```

#!/bin/bash

if [[ ! -e $1 ]]
then
echo "File $1 does not exist!";
exit 1;
fi

if [[ ! -e $2 ]]
then
echo "File $2 does not exist!";
exit 1;
fi

if [[ ! -e split*.out ]]
then
perl -e '
    open OUT, <'$1'>;
    $c=0;
    open (SPLIT, ">>", "split$c.out");
    print SPLIT "hmmpfam - search one or more sequences against HMM
database\nHMMER 2.3.2 (Oct 2003)\nCopyright (C) 1992-2003
HHMI/Washington University School of Medicine\nFreely distributed under
the GNU General Public License (GPL)\n- - - - -\nHMM file:
- - - - -\nSequence file:      '$2'\n- - - - -\n";
    while(<OUT>){
        if(/hmmpfam - search one or more sequences against HMM
database/){
            $h++;
        }
        if($h>0){
            if($h<8){
                $h++;
                next;
            }
            else{
                print SPLIT "//\n";
                $h=0;
                next;
            }
        }
        if($i<1000){
            if(/\n\n\n/){
                print SPLIT; $i++;
            }
        }
    }
'

```

```

        else{
            print SPLIT;
        }
    }
    else{
        $c++;
        $i=0;
        open (SPLIT, ">>", "split$c.out");
        print SPLIT "hmmpfam - search one or more sequences against
HMM database\nHMMER 2.3.2 (Oct 2003)\nCopyright (C) 1992-2003
HHMI/Washington University School of Medicine\nFreely distributed under
the GNU General Public License (GPL)\n- - - - -\nHMM file:
Pfam_ls\nSequence file:          '$2'\n- - - - -\n";
        unless(/\//){
            print SPLIT;
        }
    }
}

';
else
echo "File splited.";
fi

for f in `ls split*.out | sort -n`;
do
    bn=`basename $f .out`;
    for p in `ps hocmd -C process_hmmpfam.pl | cut -s -d ' ' -f 4 |
sort | uniq`;
    do
        while [[ `ps hocmd -C process_hmmpfam.pl | sort | uniq | wc -l` >
$THREADS ]];
        do
            sleep 1;
        done
        if [[ $p =~ "$f" ]];
        then
            echo "$f already running!";
        else
            if [[ -e $bn.log ]];
            then
                echo "$f already ran!";
            else
                ./process_hmmpfam.pl -s $2 $f &
                break;
            fi
        fi
    done
    if [[ `ps hocmd -C process_hmmpfam.pl | sort | uniq | wc -l` == 0
]];
    then
        if [[ -e $bn.log ]];
        then
            echo "$f already ran!";
        else
            ./process_hmmpfam.pl -s $2 $f &

```

```

        fi
    fi
done

echo -n "Waiting for remaining threads to finish..."

while [[ `ps hocmd -C process_hmmpfam.pl | sort | uniq | wc -l` != 0
]];
do
    sleep 1;
done

cat split*.tsv | sort -r | uniq > $1.tsv

cut -f 2,9 $1.tsv | uniq | sort | grep -v sequence > $1.arch.tsv

for f in split*.log;
do
    tail -n +6 $f | perl -pe 's/\n$/\t-\n/' >> $1.arch.tsv;
done

echo -e "Architecture\tNofProteins" > $1.arch;
tail -qn +2 split*.arch | sort | perl -ne '
chomp;
split /\t/;
$a{$_[0]}+=$_[1];
if eof){
    foreach (keys %a){
        print "$_\t$a{$_}\n";
    }
}' | sort >> $1.arch

echo -e "Architecture\tNofProteins\tNofMatches" > $1.model;
tail -qn +2 split*.model | sort | perl -ne '
chomp;
split /\t/;
$a{$_[0]}+=$_[1];
$b{$_[0]}+=$_[2];
if eof){
    foreach (keys %a){
        print "$_\t$a{$_}\t$b{$_}\n";
    }
}' | sort >> $1.model

cat split*.log | sort | uniq | perl -ne '
next if /Command:/;
next if /File:/;
if(/Sequences without known domains \((\d+)\):/){
    $w+=$1;
}
if(/Total number of non-repeated queries: (\d+)/){
    $nr+=$1;
}
else{
    unless(/Sequences without known domains/){
        push @a,$_;
    }
}
}

```

```
if (eof) {
    print "Command: ./process_hmmpfam.pl -s '$2' '$1'\n\nFile: '$1'\nTotal
number of non-repeated queries: $nr\nSequences without known domains
($w):\n", @a;
}' > $1.log

rm split*

echo -e "Finished."

exit 0;
```

APÊNDICE C - Artigo de periódico

Gaiarsa, JW, De Souza RF, Van Sluys MA. Plant cell wall degrading gene discovery and validation in *Xanthomonas*. Em fase de elaboração

Introduction

Xanthomonadaceae is a group rich in phytopathogens with one of the broadest host range among the proteobacteria. Worldwide crop losses caused by *Xanthomonadaceae* amounts to millions of dollars annually (REF). *Xanthomonas campestris* is used in the production of xanthan gum, an important biopolymer employed in the food, cosmetics and petroleum industries. One genera of this group, *Stenotrophomonas*, although known to be associated with plants, is mainly a human pathogen causing opportunistic infections in HIV patients.

Carbohydrases coding genes, or more specifically, plant cell wall degrading enzymes (PCWDE) are known to be a particular feature of plant associated bacteria genomes, including the *Xanthomonadaceae* group (REFS). PCWDE have a key role in the infection process of the host as they have been implicated in bacteria virulence (Cooper RM 1984, +REFS). They also play a very important role in the production of ligno-cellulosic ethanol also known as second generation bioethanol (REFS). Within the *Xanthomonadaceae* group, only particular strains of *Xanthomonas* have been analyzed regarding the expression and enzymatic activities of the PCWDE genes (Ray, 2000; Schröter, 2001; Sun, 2005; Rajeshwari, 2005; Hsiao, 2007; Hu, 2007; Xiao, 2008).

Although a great volume of genomic data has been made available for *Xanthomadaceae* in the last decade (genomas REFS), knowledge regarding the genomic repertoire coding for these enzymes is limited. Indeed, some of this group genomes have been submitted to a comparative genomics approach (comparativos Xantho/Xyl REFS, Toth et al., 2006), but the published results did not shed light into PCWDE genomic repertoire diversity among this group. We used in silico techniques to search for these potential differences. The results presented herein reveal for the first time that PCWDE genomic repertoire varies among the *Xanthomonadaceae* genomes. Moreover we found that insertion sequences (IS) are likely to have had

impacted genome remodeling of the group and that HGT regions may be an important contribution to *Xanthomonas* virulence.

This study aims to improve the definition of a PCWDE gene repertoire in the *Xanthomadacea* family. The results may shed light into the role of PCWDE into the mechanisms by which bacteria invade and colonize their host. This knowledge can lead to novel ways of treating crops for *Xanthomonadaceae* caused diseases or assist in selecting pathogen resistant varieties. It also establishes a discovery framework for PCWDE genes in other organisms, useful in bioethanol production research.

Materials and methods

19 *Xanthomonadaceae* genomes were selected for performing this work in October 2008 as shown in table 1.

Gene selection through PFam HMM model and text mining (glipfammer)

This pipeline is composed of programs written in Perl and Shell Script programming languages that should be used in sequence. The goal is to select a set of genes that belong to some functional context delineated by a set of expressions, based on identifying relevant protein folds in the first step and in second instance, namely in the absence of identifiable folds, in the filtering of definition lines of genes, or "annotation".

The dictionary.init.sh program makes an initial expressions dictionary (which may be Perl regular expressions) to be used for filtering. In the context of this work the dictionary was populated with words taken from the web pages of the Carbohydrate Active Enzymes database and the University of Georgia's Complex Carbohydrate Research Center.

The program that follows, hmmpfam_selection.pl, filters Pfam models (PFs) according to the presence of dictionary words. First the InterPro descriptions of PFs are filtered with the dictionary and then the remaining PFs in the pertaining clans are

also included. In the context of this work the PFs filtered (pfam_filtered.fam) were filtered one more time manually based on the description of the PF, looking for evidence of action on monomers and links belonging to the plant cell wall polysaccharides.

Both the initial dictionary and the pfam_filtered.fam file are just starting points. Manual intervention is required for the selection process to be specific.

Finally the program glipfam.sh does a search with hmmpfam [HMMER software package] with E-value 1e-01 against the set of selected PFs on the multifasta of the genome ORFs. The genes without PFs detected are then selected by their definition line. Optionally ORF predictions are made on the multifasta and/or translation of the sequences.

In addition to these programs, others have also been written to assist the process of analysis: pfam_process_threads.2.0.sh is a multithread hmmpfam output parser based on previous code written for other projects.

At the end a multifasta with selected genes is produced.

Clustering and Blastcore

The clustering method developed involves executing two different algorithms on the same sequences and later manual selection of clusters. One of the methods is done with the OrthoMCL package, in which MCL clustering is done on Blast bidirectional best hits [REFERENCES] and the other one which was developed for this work, called blastcore.

Blastcore works based on one-way clusters formation by the blastclust tool [REF] and iterative scanning of quality of such clusters by guidance of the multiple global alignment program T-Coffee. T-Coffee alignment has a method that provides a quality score called CORE [CORE description literature] in which the criterion for a satisfactory alignment is a CORE of 70. In blastcore the construction process and quality verification of the clusters is done iteratively by bisection of the range of identity and coverage parameters of sequences in blastclust until the process converges. In the case of clustering made in this work the blastcore produced the

best parameters with identity of 55% and coverage of 40% with 95% of the clusters having a CORE greater than 70.

At the end of the process a table similar to the PTT format is generated, with the addition of columns for PFam architecture and cluster id which is then subjected to analysis by the RAxML and Phrog programs.

Phrogs

The character profile plotted against the genomes studied (phrog) is a straightforward way to check for differences between genomes. The first step is to transform the character table of each taxon in a common table, with an axis for taxa and the other for characters, and create empty cells for nonexistent characters of each taxon. The more labor-intensive step of this process, and so made by the `phrog_plot.pl` program, is to insert empty cells respecting the organization of the characters in a common general order. During the process we can deal with repetitive characters in two ways: sum equivalent character cells or not and, if added, whether they should or should not be normalized to presence and absence, or binary characters. If the table has all repetitive characters explicit, we can also rather than writing cells with characters to represent the presence or absence, replace the cells by other information such as the coordinates of genes. We can then reorder the data on the axis of the characters according to a reference genome. This table can then be subjected to a plotting software, as Microsoft Excel was used for this work. The two ways found to better represent the data through Excel were the 3D plot flattened and reformatted to resemble a table or the conditional formatting of the table itself with painted cells. In both cases adjustments can be made to better represent binary quantities or not.

Islands

The data of the work of Lima et al 2003 on *Xanthomonas campestris* pv. *campestris* and *Xanthomonas axonopodis* pv. *citri* were used to perform BlastN searches on the chromosome of additional genomes analyzed in this work. Hits with E-value equal to 0.0 were regarded as the sintenic region of the island. From this step the program

island_pipeline.sh compiles a list of locustags of genes belonging to those regions, along with a list of genes of carbohydrases and transposition elements (TE). Also the locustags are ordered according to the position of the genes in chromosome. This is then submitted to the island_render.pl script that writes a HTML document where each genome has a series of rows of small pictures corresponding to each ORF. Each ORF may have a certain image depending on the information associated with it (HGT, carbohydrase, TE), in addition a pop-up balloon may be displayed over the ORF, containing related information, when the cursor is positioned over the image.

Prediction of the active site and plotting it in the multiple alignment

The prediction of the active site of selected genes was done using the active site prediction feature of the HMMER software. This information was then used to generate rows within the files of multiple alignments of clusters that contain an underscore in a region without the active site and an asterisk in the residue of the active site. Opening these files in the software Jalview and coloring the residues by percentage of identity between the sequences was possible to identify areas of change in or around the active site.

ECs

The genomes were reannotated with the aid of the Blast2Go4Pipe software which allowed the assignment of EC numbers and generation of Kegg metabolism maps with the ORFs of the 19 genomes as input.

Regions without PFam

The coordinates of the regions of selected genes that lacked correspondence with PFs were used to extract the sequence of these regions that were then submitted to blastcore for clustering and subsequent verification of the CORE of the alignment of each cluster.

Results

Survey of ECs

Our goal was to list the set of ORFs coding for enzymes involved in plant cell wall degradation. In a survey using the Kegg database through the blast2go4pipe program we were able to ascertain the designating EC number for 530 genes, or roughly half of the genes found by our pipeline (glipfammer) of analysis. The coverage of these 530 genes on the map of starch and sucrose metabolism of Kegg is not complete.

Comparison of surveys for carbohydrases through CAZy, KEGG, and glipfammer

A survey through the Kegg ontology for the name of the three substrates (pectin, cellulose and hemicellulose) was able to retrieve 80 genes also found through our pipeline and CAZy database. Nevertheless 2 genes were covered only by the ontology and by our method. Only the taxons present in all the databases were compared.

Glipfammer selection

The selection process of HMMs with a dictionary of 15 expressions has produced a list of 104 PFs from the 693 initially found, belonging to 92 clans, and a list of 78 keywords and regular expressions for selection of the definition lines of genes without detectable PFs.

The selected genes of 19 the genomes were altogether 1084 with 88 PF architectures, are more numerous in *Xanthomonas*, *Stenotrophomonas* has about one half to one third of what *Xanthomonas* has and almost an order of magnitude less is present in *Xylella*. The number of genes correlates with the decreasing size of chromosomes for each genus: *Xanthomonas* 5 MB, *Stenotrophomonas* 4 MB and *Xylella* 3 MB.

Clustering

Selecting clusters for the set of selected genes generated 163 clusters from both our pipeline (BlastCore) and OrthoMCL.

During the process of evaluation of clustering techniques it was discovered that OrthoMCL was able to divide large blastcore clusters that contain very similar genes into smaller clusters, that is genes with high sequence similarity and physically close, that probably constitute inparalogy events. On the other hand several singletons left aside by OrthoMCL were incorporated into clusters by blastcore, in particular in cases of broken or interrupted genes, as is frequently found in *Xanthomonas oryzae*. For this reason clusters were manually chosen in order to best solve these issues.

Phrogs: different ECs and clusters

In the profile of EC numbers for the genomes much of the metabolism is common to all organisms with the exception in the case of reactions linked to carbohydrates of 16 ECs. They are present in most *Xanthomonas*, but are largely absent in *Stenotrophomonas* and *Xylella*. These reactions are associated with bonds and monomers of polysaccharides that can be found both in the plant cell wall as in other types of cells and tissues.

If however we look at the profile of clusters present in the 19 genomes we can observe a larger number of genes (than the profile of ECs) possibly involved with the degradation of carbohydrates. If we look at this same profile now using the gene symbols a horizontal flattening of the phrog plot occurs since multiple clusters of a few genes or singletons were condensed with larger clusters that have the same symbols.

This way we can say that each column in the cluster profile corresponds to a set of orthologous genes and that sequences in adjacent columns with the same symbol in the symbol profile are paralogous (the same applies to the pfam_arch profile). In addition, between genes with the same symbol, those who belong to the same cluster are inparalogous and different cluster are outparalogous.

The columns of the plot were organized according to a reference genome (XCC) and were then painted with a heat scale in which Red corresponds to the start of the chromosome, yellow the middle of the chromosome and green the end of the chromosome. This way we can identify reorganizations of chromosomes. The position of a gene conserved between syntenic regions of genomes reinforces the idea of orthology, although this is only observable within genomes of the same genus. Although in the *Xanthomonas oryzae* genomes the large number of small stretches of chromosome rearranged hinders the determination of homology relationships, in some cases being impossible, cases in which the gene was interrupted, disrupted or deleted, forming a separate cluster, singleton or missing.

In these distributions we can observe that most of the empty cells are in the genera *Stenotrophomonas* and *Xylella*, as one would expect by the number of genes observed. However, these missing characters are largely genes that are present in multiple copies (e.g. *egl*, *bgl*, *pel*, *xyl*, ...) in *Xanthomonas* genus genomes, in particular in pathovars of the *Xanthomonas campestris* species. In addition, some are missing multigene family members, or possibly operons, which are found only in the larger genomes (e.g. *xynA*, *xynB*, *xynC*; *by*, *manA*, *manB* *pelB*; ...).

Many of these duplications occur in an exogenous way, i.e. a copy of the gene belongs to a region of vertical gene transfer and other one belongs to a region of horizontal gene transfer.

Substrates and relative content according to EC and InterPro

From the description of the EC number an assignment of the substrate on which the enzyme acts was made. When classifying the genes in this way we observe that most genes are assigned to a substrate not belonging to the plant cell wall or a nonspecific assignment, to more than one substrate. Those who are directed to the plant cell wall belong in the vast majority to pectinases.

The percentage of genes assigned to each of the three plant cell wall carbohydrates that we observe is greater for pectinases in all cases.

On the other hand, if we look at the InterPro description of PFs assigned to a cluster an inference can be made about what the kind of polysaccharide monomers and chemical bonds are compatible with that enzyme. In this case, the relative percentage of each plant cell wall carbohydrate displays a pattern that correlates well with the composition of the wall of the host plant in some of the *Xanthomonas*. Plant cell wall type I hosts, infected by *Xanthomonas campestris*, tend to have more pectin in their composition than type II hosts, that have more hemicellulose and are infected by *Xanthomonas oryzae*.

In the plot of predicted active sites over the multiple alignments of sequences of each cluster, some particular mutation events were: changes in the vicinity of 50 amino acids of the site, a change in the residue of the active site, a deletion of the active site and no change. In cases where it was possible to predict the position of the residues of the active sites of proteins, vicinity changes were found in almost all cases and some active site residue changes and deletions were also largely accompanied by changes in the vicinity.

Islands and ballplot

In the plot of ORFs for each chromosome of *Xanthomonas* chromosomal rearrangement phenomena due to TEs were apparent and also HGT events of carbohidrazes. It is clear the ubiquitous presence of TEs in *Xanthomonas oryzae* genomes is more pervasive than in other genomes, including various elements which flank genes of carbohidrazes. In this plot the HGT Islands information from the work of Lima et al 2004 was propagated to all genomes in this study revealing that part (# of genes HGT/genes VGT) of the carbohidrazes are contained in HGT regions.

Regions without PFam

Of 824 clusters of regions without PFam only two exhibited CORE less than 70. A gene of multiple Rhs repetitive motifs and two repeated stretches of a cellulose synthase in *Xanthomonas campestris* pv. *armoraciae*. The first case constitutes a recombination hotspot with many breaks and great variation in the number of

repetitions with many genes present specially in *Xanthomonas oryzae*. The second is a peculiar case of a repetition within the same gene found only once in the gene *Xanthomonas axonopodis* orthólogo of pv. *citri*. Also there is the occurrence of similar regions without PFam between different clusters of genes that mainly occur in the Rhs motifs, also because they are hotspots of recombination, and in genes of *Xanthomonas oryzae*, genomes with high level of recent reorganization due to TEs.

Discussion

The collection of EC numbers, although contemplate various activities of degradation of the three plant cell wall components, is not able to represent of the all activities necessary for the decomposition of variants of the polymers. In the work of Preston et al 2001 it is demonstrated the degree of specificity that these enzymes may have, but that is revealed only by its structure.

The fact that two genes were found in common between the KEGG ontology and our selection, but not in the CAZy database, and also the work of Sonti et al 2009 indicates that there is still unknown genes involved in the degradation of polysaccharides or plant cell wall polymer cross-linking.

The existence of genes selected that do not have PFs detected (# Gs without PF) can also indicate the possibility of a diversity of carbohydrases, and its protein folds, still unknown. The regions that have no Pfam detectable are quite conserved between the 19 genomes. This occurs probably due to very similar genomes and/or recent divergence. We can assume that many of these regions are motifs of protein folding with functional relevance that are not yet covered in the PFam database. This is reinforced by the fact that these regions are distributed along more than three-quarters of the set of genes, clusters and architectures of Pfams found.

The numbers of enzymes found in each genome also correlates with the proficiency of plant tissue maceration, *Xanthomonas* being more destructive and *Xylella* less destructive. This way we can assume that a greater number of copies and the presence of more complete multigene families foster greater capacity of plant cell wall degradation. There are three possible reasons for this: a greater number of copies can better supply the need for generation of messenger RNA, it can also allow

the adaptation of the different copies to different environmental contexts, stage of infection, among other circumstances, and finally, the presence of more members of a multigene family can provide better ability to hydrolyze certain substrates.

The second reason presented has evidence at the level of amino acid sequence of orthologous gene, in which we can observe changes, in particular the active site and its surroundings, in the multiple alignments of the clusters. The variation in the vicinity of active sites can be attributed to subtle changes in the construction of host carbohydrate and/or the context in which the gene is expressed. There are cases in which the variation is more radical, with changes or deletions of active site itself, occurring between inparalogous clusters as is the case of clusters 68 and oxc2507.

Many of these TEs may be interacting with these genes modulating their expressions, modifying their structures or inactivating them.

The genes belonging to HGT regions are not as much as could be expected for a virulence character of the pathogen. Still, we cannot find any of these regions in the genomes of *Stenotrophomonas* or *Xylella*, and can be a feature that distinguishes the phenotype of intense host tissue maceration of *Xanthomonas* and *Xylella* fastidiousness.

There is still much to be revealed about how the decomposition of plant cell walls occurs (REF CCRC). The search capabilities for this class of enzymes on the basis of the classification of metabolic activities appear to be incomplete given the wide variety of combinations of monomers and chemical bonds possible among the plant cell wall polysaccharides and tissue maceration capacity displayed by some of these pathogens. In order to reveal this diversity we developed a search method based on the conservation of the structure of enzymes that hydrolyze carbohydrates.

Also, the likely parallelism of functions exist between the metabolism of the bacterial and the plant cell walls as there are common chemical bonds and similar among the two monomers (figure X). To better define the full PCWDE set future studies would require a larger set of genomes, including plants, non-pathogenic bacteria, fungi and even animals known to digest biomass directly (REFS). Still the bacterial cell wall degrading enzyme genes revealed here may be the source of genetic material is novel PCWD.

Taken together our results indicate that the repertoire of PCWDE genes may represent an adaptive *Xanthomonadaceae* movement of the family of pathogens towards their particular niche. The data presented herein may help to advance the knowledge on how these bacteria invade and make sure you colonize on their host and in particular, the role of PCWDE in this process. This knowledge can lead to novel ways of *Xanthomonadaceae* treating crops for diseases caused or assist in selecting pathogen resistant varieties. The data may also be of biotechnological application as it may assist the discovery of novel PCWDEs involved in the production of cellulosic ethanol. Plant biomass co-fermentation through discovery of new enzymes may have other applications than simpler carbohydrates and celulosic ethanol because ethanol itself is a raw material for many other industries (bioplastics, Amiris). Also, production of xanthan gum, an important industrial biopolymer, can also benefit from knowledge of PWCDE (Schröter, 2001).

Our analyses towards the PCWDE apparatus may have excluded genes that encode for enzymes of biotechnological interest and which are part of the initial steps of plant cell wall hydrolysis process. To better understand this step, future work on the new PCWDE proteins reported here should be targeted towards the subset of enzymes predicted to be secreted by the bacteria.

Acknowledgements

This work was supported by Fundação de Amparo a Pesquisa do Estado de São Paulo [grant number 2010/50735-9].

Tables and figures

Table 1. Genomes used in this work with their abridged tag representation, the source database and their accession.

Tag	Genome	Acession	Database
XO	<i>Xylella fastidiosa</i> str. Dixon	XO	Xylella fastidiosa comparative genome project (www.xylella.lncc.br)
XP	<i>Xylella fastidiosa</i> str. Temecula1	XP	
XA	<i>Xylella fastidiosa</i> str. Ann1	XA	
XF	<i>Xylella fastidiosa</i> str. 9a5c	XF	
xfm	<i>Xylella fastidiosa</i> str. M12	NC_010513	NCBI (ftp.ncbi.nih.gov)
		NC_010579	
xfn	<i>Xylella fastidiosa</i> str. M23	NC_010577	
sml	<i>Stenotrophomonas maltophilia</i> str. K279a	NC_010943	
smt	<i>Stenotrophomonas maltophilia</i> str. R551-3	NC_011071	
		NC_003919	
xac	<i>Xanthomonas axonopodis</i> pv. citri str. 306	NC_003921	
		NC_003922	
		NC_007508	
		NC_007507	
xcv	<i>Xanthomonas campestris</i> pv. vesicatoria str. 85-10	NC_007505	
		NC_007504	
		NC_007506	
xoo	<i>Xanthomonas oryzae</i> pv. oryzae str. KACC10331	NC_006834	JCVI Comprehensive Microbial Resource (cmr.jcvi.org)
xom	<i>Xanthomonas oryzae</i> pv. oryzae str. MAFF 311018	NC_007705	
xop	<i>Xanthomonas oryzae</i> pv. oryzae str. PXO99A	NC_010717	
xcc	<i>Xanthomonas campestris</i> pv. campestris str. ATCC 33913	NC_003902	
xca	<i>Xanthomonas campestris</i> pv. campestris str. 8004	NC_007086	
xcb	<i>Xanthomonas campestris</i> pv. campestris str. B100	NC_010688	
xoz	<i>Xanthomonas oryzae</i> pv. oryzicola str. bls256	Xoc	
xvm	<i>Xanthomonas vasciola</i> pv. musacearum str.	ntxv01	
xar	<i>Xanthomonas campestris</i> pv. armoraciae str. 756c	Xca	

Figure 1. Diagram of presence and absence of EC number for the 1084 genes, their clusters and the PFs architectures found.

Figure 2. Diagram and table of the number of carbohydrases found by three different routes.

Figure 3. Chart with number of genes found for each one of the 19 genomes.

Figure 4. The top part of the picture is the phrog plot of clusters of the 19 genomes sorted with chromosome xcc as the reference. The bottom part is the same done but for the gene symbols from each cluster.

Figure 5. Relative content of cellulases, hemicelulases and pectinases in each genome according to the description of PFs in InterPro.

Active sites

Figure 6. Modification events on active sites found in the cluster sequences.

Figure 7. Representation of the output of the drawing program of chromosomes.
<https://gate.ib.usp.br/~tmbogus/ballplot2.3.html>.