UNIVERSIDADE ESTADUAL DE CAMPINAS

doutorado

INSTITUTO DE BIOLOGIA

Universidade Estadual de Campinas

Instituto de Biologia

Departamento de Genética e Evolução

Montagem e anotação do genoma de *Xylella fastidiosa*: identificação dos genes responsáveis pela síntese da goma fastidiana

Felipe Rodrigues da Silva

Orientador: Prof. Dr. Adilson Leite

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas (UNICAMP) para a obtenção do título de Doutor em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento.

2001

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

Si38m

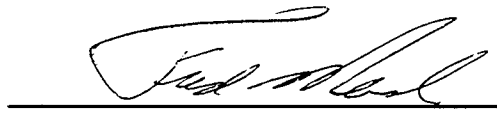
Silva, Felipe Rodrigues da
Montagem e anotação do genoma de *Xylella fastidiosa*: identificação dos genes responsáveis pela síntese da goma fastidiana/Felipe Rodrigues da Silva. --
Campinas, S.P:[s.n.], 2001.

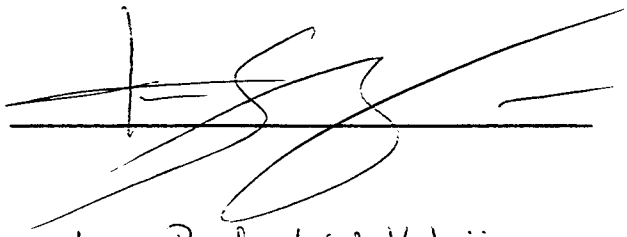
Orientador: Adilson Leite
Tese (doutorado) – Universidade Estadual de Campinas. Instituto de Biologia.

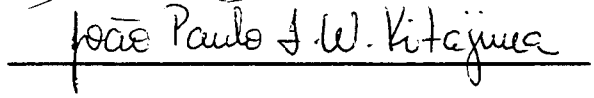
1. Genoma. 2. Biopolímeros. 3. Engenharia genética. I.Leite, Adilson. II. Universidade Estadual de Campinas. Instituto de Biologia. III.Título.

Banca examinadora

Prof. Dr. Adilson Leite 

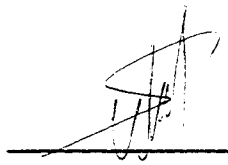
Prof. Dr. Carlos Frederico Martins Menck 

Prof. Dr. Francisco Gorgonio da Nóbrega 

Prof. Dr. João Paulo Kitajima 

Prof. Dr. Paulo Arruda

Prof. Dr. Aline Maria da Silva

Prof. Dr. Michel G. A. Vincentz 

Acreditar
Na existência dourada do sol
mesmo que em plena boca
nos bata o açoite contínuo da noite.
Arrebentar
a corrente que envolve o amanhã,
despertar as espadas,
varrer as esfinges das encruzilhadas.
Todo esse tempo
foi igual a dormir num navio:
sem fazer movimento,
mas tecendo o fio da água e do vento.
Eu, baderneiro,
me tornei cavaleiro,
malandramente,
pelos caminhos.
Meu companheiro
tá armado até os dentes:
já não há mais moinhos
como os de antigamente.

João Bosco & Aldir Blanc , *O Cavaleiro e os Moinhos*

Dedicatória

Ao Adilson, por ser mais que um amigo, mais que um chefe, muito mais que um orientador.

Por ser um exemplo.

Agradecimentos

Ao Adilson Leite, orientador desde a Iniciação Científica, pelo óbvio, pelas partidas de pebolim na cantina da Agrícola, por ter me convidado a estagiar com ele em 1992 (e jurar até hoje que não se arrependeu), pelas noites agradabilíssimas preparando placas de lise de bibliotecas de fago λ, pela parceria no curso de Biologia Computacional, pelos bolos do Dia do Eppendorf, pelas discussões de resultados para as duas teses e, principalmente, pelo bom-humor formidável (quebrado apenas pelas brincadeiras do FAS ou da FARRA) .

Ao Paulo Arruda, por ter criado, e ajudado muito a manter, o espaço onde esta grande aventura de nove anos aconteceu, pelo convite, no final de 1997, para fazer parte do Grupo Genoma do CBMEG, por todas as oportunidades de contato com a comunidade científica do Brasil e do mundo, por ter me emprestado ao LBI e pelas histórias, principalmente a dos peões correndo atrás do tatu.

Ao André L. Vettore, por ter sido o meu irmão no laboratório desde o início, pela capacidade de ponderar, por ter me ensinado a jamais subestimar a besta, pela companhia em tantas viagens, pela incrível compreensão e respeito às nossas (muitas) diferenças. Ao Edson L. Kemper por ser a prova viva de que criação é 10% de inspiração e 90% de transpiração. Ao André e ao Edson por terem compartilhado os melhores anos de minha vida, nossa brincadeira de Genoma no CBMEG. A eles e a Kely, Fabiana, Adriana, Marco Antônio, Silvana e, mais tarde, Almir, Ana, Danilo, Rodrigo, Juliana, Elaine, Susan, Renato e Fábio, por terem feito o grupo Genoma. À Thaís e à Natalia, além do muito obrigado, desejo o maior sucesso na aventura delas, que está apenas começando.

Ao Andrés, ao Germano, ao Gonçalo, ao Freitas, ao Chico, à Mayrimar, ao Dante e a Marília, que assim com Edson e André, foram alçar vôos em outras paragens, mas que fizeram, juntamente com a Silvia e o Márcio, o Laboratório de Biologia Molecular de Plantas e muito do CBMEG.

À Sandra, à Tânia e mais tarde à Fabiana e à Patrícia por serem muito mais que metade do motivo de sucesso do CBMEG. Ao seu Chico, por tanto coix e milho plantado e colhido. E ao Maurício, por tanto vidro lavado.

Ao Setúbal e ao Meidanis, por terem me recebido tão bem no LBI. Ao Zanoni, à Lin, ao Renato, à Marília, ao Werneck, ao César, ao Guilherme, ao José Augusto e ao Frank, por terem tornado a estadia lá tão mais agradável. Ao Katsumi, por isso também, mas principalmente pelos almoços.

Ao Menck, ao Nóbrega, ao Kitajima e ao Michel pelo carinho especial na banca. Fui, definitivamente, privilegiado por ter partilhado os projetos genoma com vocês e com os demais participantes (lista completa em http://aeg.lbi.ic.unicamp.br/xf/project/participants.html e http://sucest.lad.dcc.unicamp.br/en/Teams_people/people.html).

Finalmente, aos meus pais, pelo apoio irrestrito e pela transmissão de um gene não cariado, o da academia.

Índice

BANCA EXAMINADORA ..... **III**

DEDICATÓRIA ..... **V**

AGRADECIMENTOS ..... **VI**

ÍNDICE ..... **VII**

ÍNDICE DE FIGURAS A TABELAS ..... **IX**

ABREVIAÇÕES E TERMOS EM INGLÊS ..... **X**

RESUMO ..... **XI**

ABSTRACT ..... **XII**

INTRODUÇÃO **1**

OBJETIVOS **3**

METODOLOGIA **4**

GERAÇÃO DE SEQÜÊNCIAS 4
Construção da biblioteca shotgun *5*
Seqüenciamento dos clones *5*
Seqüenciamento com primers específicos *5*
Construção de bibliotecas de bacteriófago lambda. *5*
MONTAGEM DOS CONTIGS 6
Programas utilizados *7*
teste de co-linearidade *7*
ANÁLISE DA INFORMAÇÃO CONTIDA NAS SEQÜÊNCIAS 7
Localização de ORFs *7*
Identificação de ORFs *8*

RESULTADOS E DISCUSSÃO **9**

AUXÍLIO NA CONCLUSÃO DO PROJETO 9
Montagem de extremidades de cosmídeos e desenho de primers *9*
SEQÜENCIAMENTO, MONTAGEM E FINALIZAÇÃO DE CLONES 10
Cosmídeos *10*
FECHAMENTO DA SEQÜÊNCIA GENÔMICA 15
Construção e análise da biblioteca de fagos lambda *15*
Seqüenciamento de fagos lambda *15*
Seqüenciamento de insertos de bibliotecas shotgun totais *16*
ANOTAÇÃO DO GENOMA 17
Anotação manual *18*

Anotação automática *19*
ANOTAÇÃO DE SEQÜÊNCIAS CONCLUÍDAS 21
O GENOMA COMPLETO DE XYLELLA FASTIDIOSA 21
GOMA FASTIDIANA 23
FASTIDIAN GUM: THE *XYLELLA FASTIDIOSA* EXOPOLISACCHARIDE 25
APÊNDICE A 32
ORGANISMOS DE VIDA LIVRE COM O GENOMA SEQÜENCIADO 32
APÊNDICE B 33
LISE ALCALINA - MICROPREPS 33
BIBLIOTECAS SHOTGUN 36
PURIFICAÇÃO DE DNA DE COSMÍDEOS 39
EXTRAÇÃO DE DNA DE GEL LOW MELTING 43
MIDIPREP E SEQÜENCIAMENTO DE EXTREMIDADES DE FAGO LAMBDA 45
Preparing the Bacteria *45*
Preparing liquid phage lysate *45*
Macroplates preparation *45*
Purification of phage particles and DNA extraction *46*
APÊNDICE C 48
BASIC PRIMER ON ANNOTATION 48
Working with sequin *48*
Finding ORFs *49*
Marking the ORFs *49*
Finding similarities *51*
Annotating, finally! *51*
Another features *52*
APÊNDICE D 53
TODAS AS CATEGORIAS EMPREGADAS NA ANOTAÇÃO DE XYLELLA FASTIDIOSA 53
APÊNDICE E 57
PRIMERS PARA FECHAMENTO DO GENOMA 57
BIBLIOGRAFIA 61

Índice de figuras a tabelas

FIGURA 1: GENOMAS COMPLETOS DE ORGANISMOS 1
FIGURA 2. MAPA DO COSMÍDEO 07A01. 11
FIGURA 3. MAPA DO COSMÍDEO 07A02. 12
FIGURA 4. MAPA DO COSMÍDEO 10H05. 13
FIGURA 5. MAPA DO COSMÍDEO 11A02. 14
FIGURA 6. FLUXO DE INFORMAÇÕES DA ANOTAÇÃO AUTOMÁTICA. 20

TABELA 1. COSMÍDEOS CONCLUÍDOS PELO GENOMA - CBMEG 10
TABELA 2. CLONES SEQÜENCIADOS PARA A RESOLUÇÃO DE *GAPS*. 17
TABELA 3. OCORRÊNCIA DE SEQÜÊNCIAS CODIFICADORAS EM GENOMAS BACTERIANOS. 18
TABELA 4. COSMÍDEOS COM ANOTAÇÃO COMPLETA. 23

Abreviações e termos em inglês

aa	aminoácido
browser	navegador. Programa para visualizar páginas na Internet
contig	contíguo. Seqüência de DNA formada pela sobreposição de duas ou mais seqüências
CVC	Clorose Variegada dos Citrus
DNA	Ácido Desoxiribonucléico
DO	absorbância ótica
EPS	exopolissacarídeo
gap	buraco. Pedaço ainda não seqüenciado de um clone
HMM	Hiden Markov Model. Modelos Ocultos de Markov
Kb	milhares de pares de base
Mb	milhões de pares de base
nm	nanômetros
ORFs	Open Reading Frames. Quadros abertos de leitura
pb	pares de base
PCR	Polimerase Chain Reaction. Reação da Polimerase em Cadeia
primers	seqüências iniciadoras da síntese de ácidos nucléicos
RNA	Ácido Ribonucléico
rpm	rotações por minuto
shotgun	metralhadora. Técnica de obtenção de fragmentos aleatórios de um inserto
template	molde. Fita de DNA usada para síntese de sua complementar reversa

Resumo

A bactéria Gram-negativa *Xylella fastidiosa* foi o primeiro fitopatógeno a ser completamente seqüenciado. Esta bactéria causa um série de doenças em plantas de importância econômica, incluindo a Clorose Variegada dos Cítrus (CVC), também conhecida como amarelinho. Esta tese descreve a participação do autor na montagem e anotação deste genoma, com especial ênfase à descrição do sistema de produção de um composto provavelmente ligado à patogenicidade da bactéria: a goma fastidiana.

A análise da seqüência genômica de *Xylella fastidiosa* revelou um fragmento de 12 Kb contendo um operon muito semelhante ao operon *gum* de *Xanthomonas campestris*. A presença de genes ligados à síntese de precursores de açúcares, a existência de reguladores da produção de exopolissacarídeos (EPS) em seu genoma e a ausência de 3 dos genes *gum* de *Xanthomonas campestris* sugerem *que Xylella fastidiosa* é capaz de sintetizar um EPS diferente da goma xantana. Este novo EPS seria fruto da polimerização de unidades repetidas de um tetrassacarídeo formado pela adição seqüencial de glicose-1-fosfato, glicose, manose e ácido glicurônico a um poliprenol carreador.

Abstract

The Gram-negative bacterium *Xylella fastidiosa* was the first plant pathogen to be completely sequenced. This species causes several economically important plant diseases, including citrus variegated chlorosis (CVC). Analysis of the genomic sequence of *X. fastidiosa* revealed a 12 kb DNA fragment containing an operon closely related to the *gum* operon of *Xanthomonas campestris.* The presence of all genes involved in the synthesis of sugar precursors, existence of exopolysaccharide (EPS) production regulators in the genome, and the absence of three of the *X. campestris* gum genes suggested that *X. fastidiosa* is able to synthesize an EPS different from that of xanthan gum. This novel EPS consists of polymerized tetrasaccharide repeating units assembled by the sequential addition of glucose-1-phosphate, glucose, mannose and glucuronic acid on a polyprenol phosphate carrier.

Introdução

O desenvolvimento da tecnologia de seqüenciamento automático de DNA em larga escala levou a uma mudança dramática no ritmo de identificação de genes. Há pouco tempo alguns anos de trabalho eram necessários para identificar um único gene ou proteína de interesse (por exemplo, da Silva, F.R., 1996: Gama-coixina: isolamento do clone genômico e caracterização da região promotora, tese de mestrado do autor). Hoje, pequenos genomas (~5 Mb) podem ser completamente seqüenciados em menos de um ano.

Até 1995, apenas genomas virais e de organelas haviam sido descritos. Em julho de 1995 foi publicada a seqüência completa do primeiro genoma de um organismo de vida livre, o *Haemophilus influenzae* (Fleischmann *et al.*, 1995). Até dezembro de 2000, as seqüências completas de 43 genomas foram disponibilizadas publicamente. A relação completa destes genomas se encontra no apêndice A. O crescimento na geração de informação com projetos genoma pode ser percebido na Figura 1.

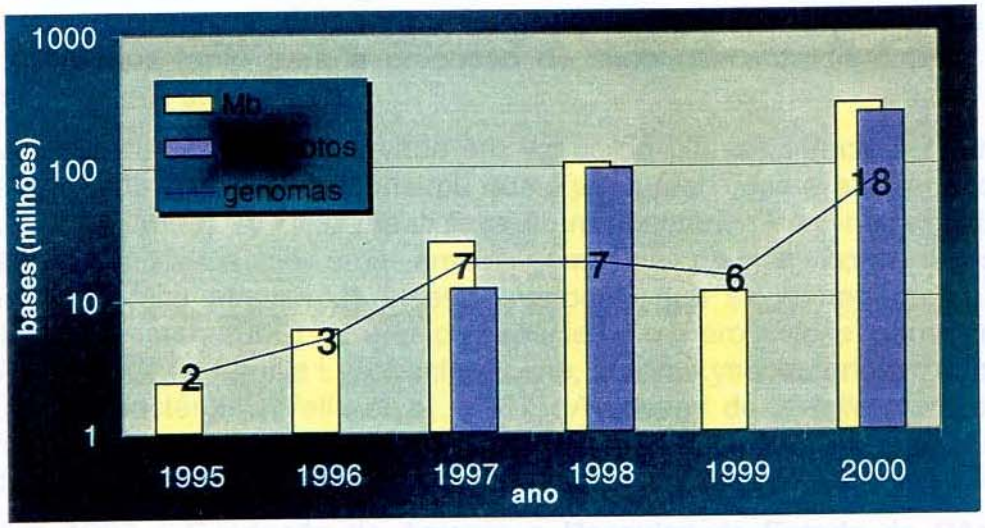


Figura 1: Genomas completos de organismos

As barras amarelas indicam o total de bases compreendidas pelos genomas, as barras azuis, o total de bases de genomas eucariotos e a linha azul, o número de genomas concluídos em um determinado ano. Note que os dados ***não*** *são cumulativos,* i.e., *os números apresentados para determinado ano não incluem os anteriores. Note também que a ordenada está apresentada em escala logarítmica e que a escala não está apresentada para a linha azul.*

Esta mudança de escala demanda novas técnicas para o tratamento dos dados gerados. O trabalho com um único gene, ou pequeno conjunto deles, permite um trabalho quase artesanal em sua exploração. Algumas ferramentas relativamente simples permitem que toda a informação contida nas seqüências seja explorada. Quando se trabalha com todo um genoma, o volume de informação é grande demais para ser explorado artesanalmente. Além disso, uma

enormidade de dados, inexistentes quando se trabalha com genes isoladamente, passa a estar disponível. Estes novos dados incluem a verificação de vias metabólicas (se estão completas e, quando não, que alternativas este determinado organismo possui para tal via de síntese, *e.g.*) e as vias completas de regulação de expressão gênica. Por fim, a possibilidade de se conhecer de antemão todos os genes de um organismo permite a aplicação de métodos de detecção em massa, como *microarrays,* em estudos de indução gênica.

A Clorose Variegada dos Citros (CVC), popularmente conhecida como Amarelinho, foi identificada pela primeira vez em 1987 nos estados de São Paulo e Minas Gerais. Inicialmente encontrada em pomares de Colina, município do centro-oeste do estado de São Paulo, a doença foi posteriormente detectada nas regiões norte e nordeste do estado de São Paulo e no estado de Minas Gerais. A partir do início da década de 1990 a CVC teve uma rápida disseminação. Em 6 anos já estava presente em 83% da principal área de citricultura do estado de São Paulo e era encontrada também outros estados (Rossetti *et al.*, 1998).

Dentre todos os sintomas da doença, que inclui clorose foliar, o mais importante e de maior conseqüência econômica é sem dúvida a diminuição do tamanho dos frutos. Estes chegam a ser 3 vezes menores que um fruto normal, apresentam sabor muito ácido, amadurecimento precoce e constituição rígida, sendo impróprios tanto para a produção de suco concentrado quanto para o consumo (Rossetti *et al.*, 1998).

Apesar de ser sido observada em vasos de plantas infectadas em 1989, apenas em 1993 é que se demonstrou que *Xylella fastidiosa* era a causadora da CVC (Li *et al.*, 1999). A *Xylella fastidiosa* é uma bactéria Gram-negativa limitada aos vasos do xilema de uma ampla gama de plantas hospedeiras. Sua transmissão é feita através de insetos suctoriais que se alimentam de xilema (Derrick and Timmer, 2000) . É a única espécie do gênero *Xylella*, pertencente ao grupo *Xanthomonas*, família Lysobacteriaceae, ordem Lysobacteriales, subdivisão Gama das Eubactérias (Wells *et al.*, 1987). As cepas de *Xylella fastidiosa* que atacam citros e café no Brasil são bastante semelhantes entre si, diferindo genética e biologicamente das demais cepas já descritas (Beretta *et al.*, 1996).

Em 1997 a Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP), com auxílio do Fundo de Defesa da Citricultura de São Paulo (FUNDECITRUS), iniciou o projeto de seqüenciamento do genoma do isolado 8.1.b clone 9.a.5.c, do isolado original *Xylella fastidiosa* que cumpriu os postulados de Koch (Li *et al.*, 1999).

Para tal foi definida uma rede virtual de laboratórios denominada de rede ONSA (Organization for Nucleotide Sequence and Analysis). Durante o projeto genoma de *Xylella fastidiosa* esta rede foi composta por um laboratório de coordenação do projeto, dois laboratórios de seqüenciamento centrais, 29 laboratórios de seqüenciamento e um centro de bioinformática. Embora haja 208 nomes listados na página que descreve o projeto (http://aeg.lbi.ic.unicamp.br/xf/project/participants.html) estima-se que houve o envolvimento direto de cerca de 400 pessoas ao longo do projeto.

Objetivos

Esta tese descreve a participação do autor no projeto genoma de *Xylella fastidiosa*. Sua participação se deu em duas frentes distintas: o grupo genoma do CBMEG e o Laboratório de Bioinformática (LBI) do Instituto de Computação, ambos na UNICAMP. Desta forma, enumeram-se os objetivos específicos para cada um destes lugares.

No CBMEG:

1. Descrição da seqüência exata dos clones atribuídos ao grupo Genoma do CBMEG pela coordenação do projeto.
2. Atribuição de significado biológico às seqüências contidas nestes clones, *i.e.*, anotação dos clones.

No LBI:

3. Auxílio na finalização do projeto.
4. Auxílio na anotação do genoma.

Além disso, o contato com a informação genômica de *Xylella fastidiosa* levou à percepção de uma série de fenômenos biológicos inéditos. A descrição de um destes fenômenos constitui também objetivo desta tese:

5. Descrição da via metabólica responsável pela produção de uma nova substância potencialmente produzida por *Xylella fastidiosa*: a goma fastidiana.

Geração de seqüências

A estratégia central prevista para o seqüenciamento do genoma da *Xyllela fastidiosa* era dividida em cinco fases (FAPESP, 1998):

i. construção de um mapa físico do genoma bacteriano baseado em digestões com enzimas de restrição.

ii. preparação de uma biblioteca de plasmídeos contendo insertos de aproximadamente 4 Kb, correspondendo a uma cobertura de 10 vezes o genoma total.

iii. seqüenciamento das extremidades dos clones da biblioteca de plasmídeos.

iv. construção de uma biblioteca de cosmídeos, contendo insertos de aproximadamente 40 Kb.

v. mapeamento da biblioteca de cosmídeos e ordenação dos clones a serem seqüenciados.

O rápido sucesso na construção, validação e mapeamento da biblioteca de cosmídeos (Frohme *et al.*, 2000), aliado à baixa qualidade das bibliotecas de plasmídeos fizeram com que os cosmídeos se transformassem na estratégia central do projeto.

Assim, cada um dos laboratórios centrais assumiu o compromisso inicial de concluir o seqüenciamento de dois cosmídeos, enquanto os demais laboratórios deveriam concluir a seqüência de um.

A abordagem *shotgun* foi empregada para elucidação das seqüências dos cosmídeos. A estratégia de seqüenciamento com uso de bibliotecas *shotgun*, na forma como foi proposta originalmente (Anderson *et al.*, 1982), compreende duas fases.

A primeira, a fase *shotgun*, consiste na geração de leituras de seqüenciamento de subclones aleatórios. Nenhuma evidência existe sobre que região da seqüência do cosmídeo será coberta por um determinado subclone. A cada lote de leituras, tentativas de "agrupamento" (*i.e.*, montagem) das seqüências obtidas até o momento são feitas. Duas ou mais leituras que se juntem em seqüências contíguas recebem o nome de *contig*. O aumento do número de leituras de uma biblioteca *shotgun* converge para a formação de um único *contig*, que representa a seqüência do cosmídeo original.

Na fase seguinte, de *finalização*, a correção da montagem é inspecionada. Verifica-se ainda diversas anomalias de dados, tais como leituras contaminantes, presença de seqüência de vetores e leituras quiméricas. Nesta fase, a obtenção de novos dados passa a ser dirigida para regiões que ainda não seguem os padrões de finalização.

Construção da biblioteca *shotgun*

O DNA de cosmídeos foi preparado e purificado utilizando-se o Mega-Kit (Qiagen). O DNA de plasmídeos foi preparado e purificado com o Plasmid Midi Kit (Qiagen). DNA de bacteriófagos lambda (λ) para a construção de sub-bibliotecas de *shotgun* foi preparado pela amplificação do inserto por XL-PCR (Perkin-Elmer) ou, alternativamente, utilizando-se o Lambda Kit (QIAGEN).

As bibliotecas *shotgun* foram preparadas através da sonicação de 20-30μg de DNA, preenchimento das extremidades com T4 Polimerase e Klenow e clonagem dos fragmentos contendo entre 500 pb e 3 Kb no plasmídio pUC 18. A faixa de tamanho dos fragmentos dependeu do tamanho do inserto do clone original, entre 2 Kb e 3 Kb para os provenientes da biblioteca de cosmídeo ou bacteriófago lambda e entre 500 pb e 1,5 Kb para os provenientes das bibliotecas de *shotgun* total do projeto. O protocolo detalhado desta técnica pode ser encontrada no Apêndice B. Para cada biblioteca foram isolados cerca de 1000 clones recombinantes.

Seqüenciamento dos clones

O DNA dos clones das bibliotecas de *shotgun* foi preparado em micro-placas de 96 poços através de lise alcalina (Sambrook *et al.*, 1989) com algumas modificações (a técnica se encontra descrita detalhadamente no Apêndice B) e seqüenciados através de reações com *dye-terminator* (Big-Dye, Applied) em um seqüenciador automático ABI 377-XL, usando o *primer* T3 ou o T7.

Seqüenciamento com *primers* específicos

Quando necessário, *primers* específicos foram desenhados a partir de seqüências já determinadas dos cosmídeos. Estes *primers* foram empregados na geração de seqüências da porção central dos insertos dos clones *shotgun* (*i.e.*, a mais de 400 pb das extremidades dos insertos) ou usando o próprio DNA do cosmídeo como *template* na reação de seqüenciamento.

Construção de bibliotecas de bacteriófago lambda.

Xylella fastidiosa foi crescida conforme descrito por Schafer *et al.* (1981) Dois litros de meio PW foram inoculados com 20 mL de uma cultura de *Xylella fastidiosa* com $DO_{600nm} = 1,0$. O inóculo foi mantido a 37°C por 15 dias sob agitação (300 rpm). Após este período, as células foram coletadas por centrifugação (15 minutos, 3.000 rpm) e lavadas com água. O DNA genômico de *Xylella fastidiosa* foi extraído com CTAB (Sambrook *et al.*, 1989). Após a extração o DNA foi quantificado em espectrofotômetro.

A metodologia da confecção da biblioteca segue fundamentalmente a descrita por Ottoboni *et al.* (1993). Cem microgramas de DNA genômico de *Xylella fastidiosa* foram parcialmente digeridos com 2 unidades de *Sau3AI* por 3 minutos a 37°C. A enzima de restrição foi inativada termicamente (15 minutos,

70°C) e o produto de digestão foi defosforilado pela ação de 5 unidades de CIAP (*Calf Intestinal Alkaline Phosphatese*) em uma incubação de 60 minutos a 37°C seguida de outra incubação de 30 minutos a 56°C. A enzima CIAP foi inativada termicamente (10 minutos a 70°C) e removida pelo tratamento com 40ng/μL de Proteínase K, 0,8% de SDS e 20mM de EDTA seguido por extração com Fenol:Clorofórmio:Álcool Isoamílico (24:23:1). Após precipitação com 0,1 volumes de acetato de sódio e etanol, o DNA foi ressuspenso em 200 μL de água e submetido a ultracentrifugação (16 horas, 30.000 rpm) em gradiente de sacarose de 40% a 5%. Fragmentos com tamanho entre 12 e 20 Kb foram recolhidos. Após nova precipitação com acetato de sódio e etanol, estes fragmentos foram ressuspensos em 20 μL de água e quantificados em espectrofotômetro.

As reações de ligação e empacotamento foram realizadas conforme especificações do *Lambda DASH II/BamH I Cloning Kit* (Stratagene). Foram realizadas reações de ligação com 4 diferentes relações molares entre o DNA do vetor lambda DASH, previamente digerido com BamH I, e os fragmentos de *Xylella fastidiosa* (1:2; 1:1; 2:1; 3:1). Dois microlitros de cada uma destas reações foram misturados e utilizados na reação de empacotamento *in vitro* utilizando-se o extrato de empacotamento *Gigapack III Gold* (Stratagene). Após o empacotamento, células da bactéria DL538 foram infectadas com os clones recombinantes (Sambrook *et al.*, 1989).

O DNA de fago lambda foi preparado segundo a metodologia de Helms *et al.* (1987) descrita mais detalhadamente no Apêndice B.

Montagem dos contigs

Após alguns ensaios preliminares, optou-se por uma abordagem *shotgun* ligeiramente modificada para obtenção dos *contigs*.

Todos os subclones da biblioteca *shotgun* são seqüenciados a partir de uma única extremidade (apenas o *primer* T7, *e.g.*). Como a inserção dos fragmentos nos subclones não segue qualquer orientação, há uma distribuição razoavelmente eqüitativa de leituras nos dois sentidos em relação à seqüência final do cosmídeo. No instante em que a soma do comprimento dos *contigs* obtidos na tentativa de montagem se aproxima do valor esperado para o cosmídeo inteiro, verificam-se os clones presentes nas extremidades dos *contigs* e, daqueles cuja leitura tem orientação voltada para fora do *contig*, é gerada a seqüência da outra extremidade (usando o *primer* T3, neste exemplo).

O uso desta estratégia permitiu que o dobro de clones fosse amostrado com o mesmo volume de trabalho da estratégia original. Um maior número de clones amostrados aumenta a chance de se obter leituras que cubram todo o cosmídeo. O seqüenciamento oposto (*i.e.*, usando o *primer* oposto ao usado para todas as leituras) apenas dos clones de região de interesse diminui a redundância improdutiva de leituras.

Além disso, a amostragem de um número maior de subclones facilitou consideravelmente o trabalho de finalização das regiões de fita simples (aquelas cuja seqüência do *contig* é gerada por leituras em um único sentido) e de

discrepâncias (quando as leituras que descrevem uma determinada base do *contig* são discordantes). Com efeito, os cosmídeos concluídos pelo grupo GENOMA – CBMEG com esta abordagem modificada contém menos *primers* específicos para resolução de problemas que os cosmídeos concluídos por grupos que empregam a estratégia original.

Programas utilizados

Dois pacotes distintos de softwares foram empregados na montagem e finalização dos *contigs*, o *Staden sequence analysis package* (Staden, 1996; Bonfield *et al.*, 1995), pioneiro nesta área (Dear and Staden, 1991) e o *Phil Green package*, nome corriqueiramente empregado para descrever o conjunto dos programas *phred* (Ewing *et al.*, 1998), *phrap* (Green, 1999) e *consed* (Gordon *et al.*, 1998).

Após uma comparação inicial, optou-se pelo trabalho com o Staden, devido a sua grande capacidade de padronização / automação, principalmente no tocante à estratégia de *seqüenciamento oposto* descrita acima.

Mais tarde, com a decisão do projeto de associar os critérios de finalização com o *Phil Green package*, as montagens passaram a ser feitas nos dois programas simultaneamente.

teste de co-linearidade

O teste de co-linearidade dos clones montados foi realizado de duas formas. Primeiramente através da comparação entre o padrão obtido com a digestão do DNA total do *Xylella* clone com determinada enzima de restrição e o previsto pela digestão eletrônica da seqüência montada. Além disso, foi realizada a verificação das posições que as duas pontas de um sub-clone contido na região representada pelo clone sendo montado ocupam em relação à seqüência gerada pela montagem. Espera-se que estas pontas estejam em orientações opostas e a uma distância compatível com o vetor (entre 9 Kb e 12 Kb no caso de bacteriófagos lambda, *e.g.*).

Análise da informação contida nas seqüências

Localização de ORFs

A performance e a acuidade de três programas localizadores de ORFs (*open reading frames*), ou regiões codificadoras, foi testada. Conquanto não foi possível observar diferenças relevantes entre eles, optou-se por trabalhar com o Glimmer (Salzberg *et al.*, 1998). O Genemark.hmm (Lukashin and Borodovsky, 1998) foi descartado por não ter uma versão que pudesse ser executada localmente (era necessário enviar, via internet, as seqüências dos cosmídeos) e o SelfId (Audic and Claverie, 1998) por sua menor operacionalidade e base instalada.

Os três programas se baseiam em modelos ocultos de Markov (HMM) (Rabiner, 1989). Glimmer implementa HMM interpolado, o que permite a análise de contextos de comprimentos e ordens variáveis.

Identificação de ORFs

Uma vez delimitadas as seqüências das ORFs, estas foram traduzidas utilizando-se o código genético bacteriano (Elzanowski and Ostell, 1999). O programa BLASTP (Altschul *et al.*, 1997) foi usado na busca de similaridades entre as seqüências de *Xylella fastidiosa* traduzidas e proteínas já descritas depositadas no GenBank, no banco nr (não-redundante) (Benson *et al.*, 2000). Optou-se pelo emprego da matriz de substituição BLOSUM62 na procura por ser uma das matrizes mais apropriadas na detecção de similaridades fracas (como as existentes entre genes homólogos que divergiram há muito tempo), o que pode ser bastante importante na inferência da função de novos genes (Henikoff and Henikoff, 1992).

O valor de corte do *e-value* (Karlin and Altschul, 1990) para considerar um alinhamento entre uma ORF e uma proteína presente no banco de dados significativo foi fixado em 10^{-5}, *i.e.*, apenas quando o *e-value* era igual ou menor que este valor a função da ORF foi inferida por homologia com a proteína a ela alinhada.

Quando o *e-value* se encontrava entre 10^{-3} e 10^{-5}, ao menos três ciclos interativos no PSI-BLAST (Altschul *et al.*, 1997) foram realizados para verificar a significância dos alinhamentos encontrados.

Auxílio na conclusão do projeto

O laboratório genoma do CBMEG foi, todo o tempo, extremamente comprometido com o projeto, realizando não só as tarefas que lhe cabiam como também auxiliando outros laboratórios. Desde um curso introdutório de análise genômica, ministrado no final de março de 1998 para 29 integrantes de 18 laboratórios da rede ONSA, passando pelo auxílio na preparação de 11 bibliotecas de *shotgun* para 9 laboratórios distintos, até a ajuda na montagem dos cosmídeos de 5 grupos.

Entre novembro de 1998 e janeiro de 2000 o autor passou boa parte do seu tempo junto ao LBI, o Laboratório de Bioinformática do Instituto de Computação da UNICAMP, que centralizava os trabalhos computacionais do projeto.

Montagem de extremidades de cosmídeos e desenho de *primers*

No início de dezembro de 1998, ficou claro que os cosmídeos distribuídos não representavam todo o genoma de *Xylella fastidiosa*. Para estimar a dimensão dos *gaps* entre os cosmídeos, foi necessário o desenho de *primers* específicos a partir da seqüência das extremidades dos cosmídeos. Como a absoluta maioria destes cosmídeos estava ainda em uma fase inicial de acabamento, foi necessária a montagem cuidadosa de suas extremidades, seguida da avaliação de sua fidedignidade para que os *primers* fossem desenhados baseados em uma seqüência real.

No total, 23 pontas foram montadas e foram desenhados 35 *primers*. Além disso, uma discrepância entre a sobreposição de dois cosmídeos exigiu o desenho de mais 6 primers para verificar as possibilidades erro na montagem. Os arquivos originais dos *primers* podem ser vistos no Apêndice E.

Os *primers* desenhados para estimar o tamanho dos *gaps* entre os cosmídeos através de PCR combinatório acabaram não sendo muito úteis. Sabemos hoje que a dimensão destes *gaps* era maior que o esperado. O erro decorreu de uma sub-estimativa do tamanho total do genoma de *Xylella*. No início do projeto, calculou-se que o genoma desta bactéria teria pouco mais que 2 Mb, ou seja, bem menor que os 2,7 Mb hora descritos. Os poucos produtos obtidos através das 506 reações de PCR eram, em sua maioria, ou artefatos, ou amplificavam (devido à presença de repetições no genoma) regiões que não a esperada.

Seqüenciamento, montagem e finalização de clones

Cosmídeos

Embora o acordo inicial da rede ONSA previsse que um laboratório Central, como era o caso do CBMEG, devesse seqüenciar 2 cosmídeos, o desenrolar do projeto tornou clara a necessidade do seqüenciamento de mais cosmídeos. O CBMEG finalizou quatro cosmídeos, descrevendo 157295 bases inéditas do genoma de *Xylella fastidiosa* contendo cerca de 131 regiões codificadoras.

Estes cosmídeos excedem o rigoroso critério determinado pelo *steering committee* do projeto genoma – FAPESP: apresentam taxa esperada de erro de 0.00 / 10 Kb (exige-se abaixo de 1 / 10 Kb), não possuem nenhuma base do consenso com *phred quality* (Ewing *et al.*, 1998) inferior a 20, não apresentam, nas leituras, bases com qualidade acima de 40 que discordem da seqüência consenso e todas as bases do consenso foram confirmadas em ambas as fitas.

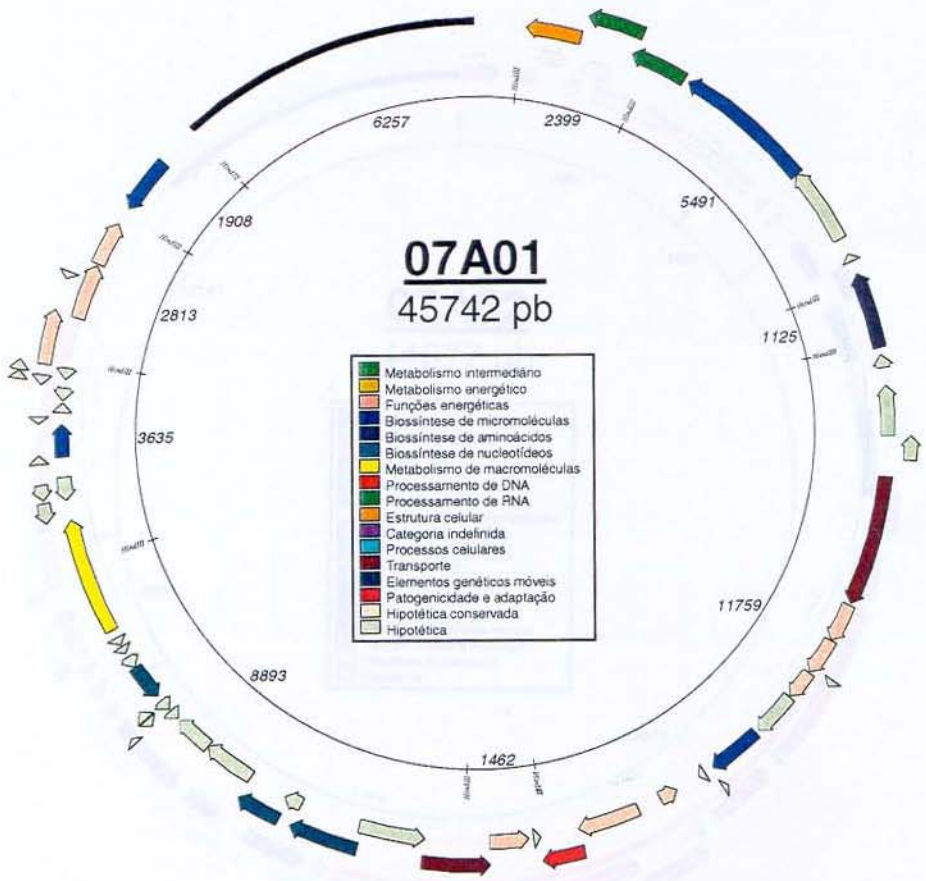
Tabela 1. Cosmídeos concluídos pelo GENOMA - CBMEG

Cosmídeo	Inserto	Leituras	phred	ORFs
07A01	40347	832/579 (632)	86,5 (30)	36
07A02	38685	572/495 (728)	84,7 (33)	33
10H05	38186	1649/1228 (1293)	88,4 (28)	27
11A02	40077	952/863 (1037)	88,0 (34)	35

A coluna Leituras indica o total de leituras realizadas na biblioteca shotgun *do cosmídeo, seguido do número destas empregadas na montagem. O número entre parênteses indica o total de leituras (*i.e., *incluindo- se as não geradas pelo CBMEG) usadas na montagem. A coluna phred indica o valor médio da* phred quality *para as bases da seqüência consenso do inserto. O número entre parênteses indica a qualidade da base de valor mais baixo encontrado. O número de ORFs ignora ORFs pequenas (menores que 100 aa) para as quais não foi possível encontrar homólogos já descritos.*

Para se chegar a esta qualidade final, 4005 leituras de bibliotecas *shotgun* dos cosmídeos foram geradas no CBMEG. Destas, apenas 3165 (79%) foram efetivamente usadas na montagem dos cosmídeos. Os cerca de 20% de perda de leituras devem-se ao seqüenciamento de clones sem inserto, leituras de qualidade muito baixa e contaminação por seqüências de *E.coli* (hospedeiro em todos os passos de clonagem). A montagem empregou também 179 leituras de cosmídeos seqüenciados por outros laboratórios que apresentavam sobreposição com o cosmídeo. Foram utilizadas, ainda, mais 346 leituras dos clones da biblioteca de plasmídeos (fase ii da estratégia central original).

Um mapa detalhado e o teste de co-linearidade destes cosmídeos pode ser visto nas figuras 2-5.



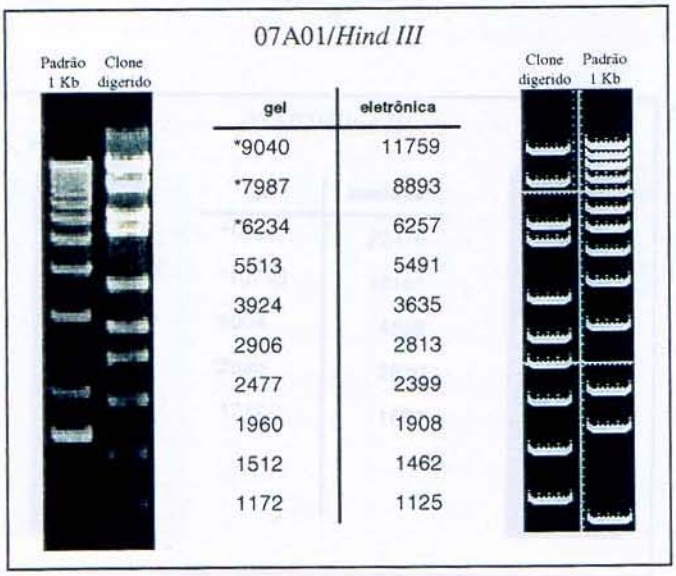
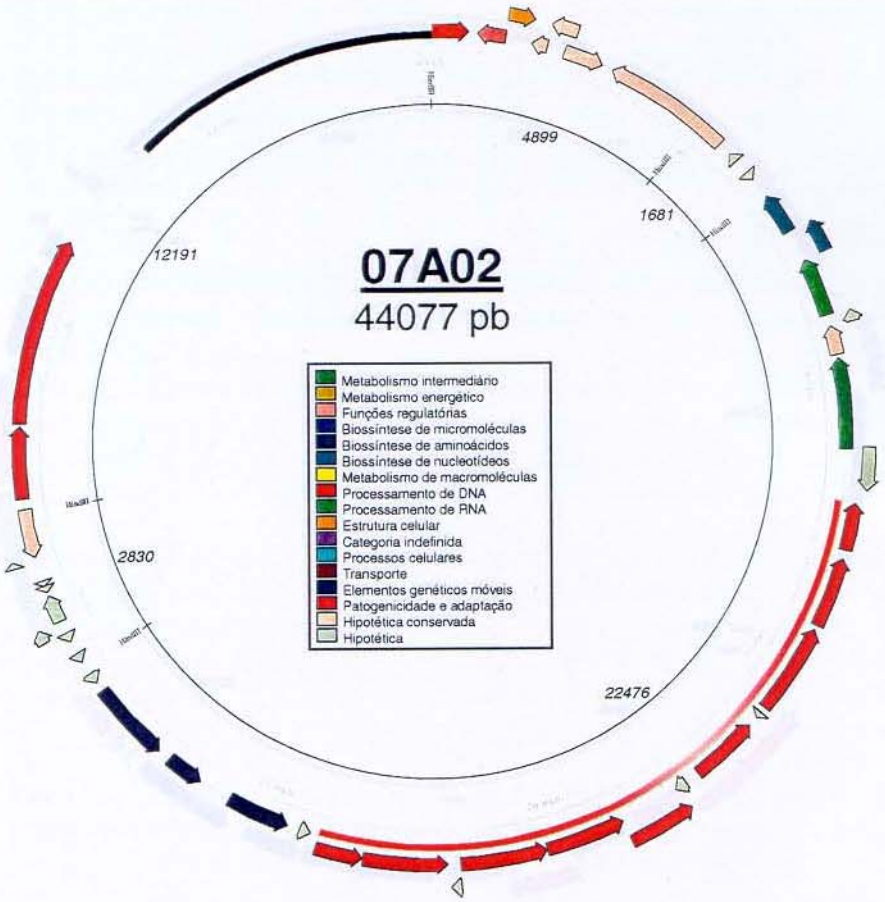


Figura 2. Mapa do cosmídeo 07A01.

O círculo representa a seqüência completa do cosmídeo. A barra preta, na parte de cima à esquerda, representa o vetor, Lawrist. As setas coloridas indicam a orientação de transcrição dos genes encontrados no inserto e sua cor, a classe a que as proteínas codificadas pertencem, segundo a legenda. Os números no círculo mais interno representam os tamanhos dos fragmentos obtidos por digestão com Hind III. *O quadrado na parte inferior da figura mostra, à esquerda, a diestão do DNA do clone e, à direita, o resultado da digestão eletrônica do clone montado. Os números entre as figuras representam os tamanhos dos fragmentos. Os da esquerda foram calculados a partir da migração observada. Aqueles contendo um asterisco (*) estão apontados como fora da faixa de precisão do programa de cálculo.*



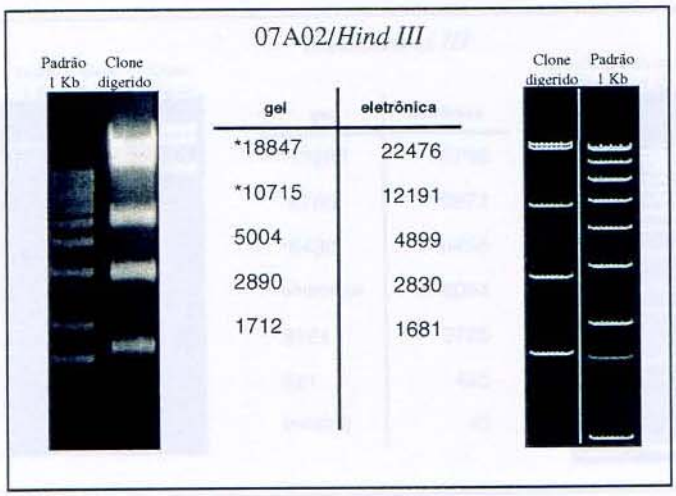
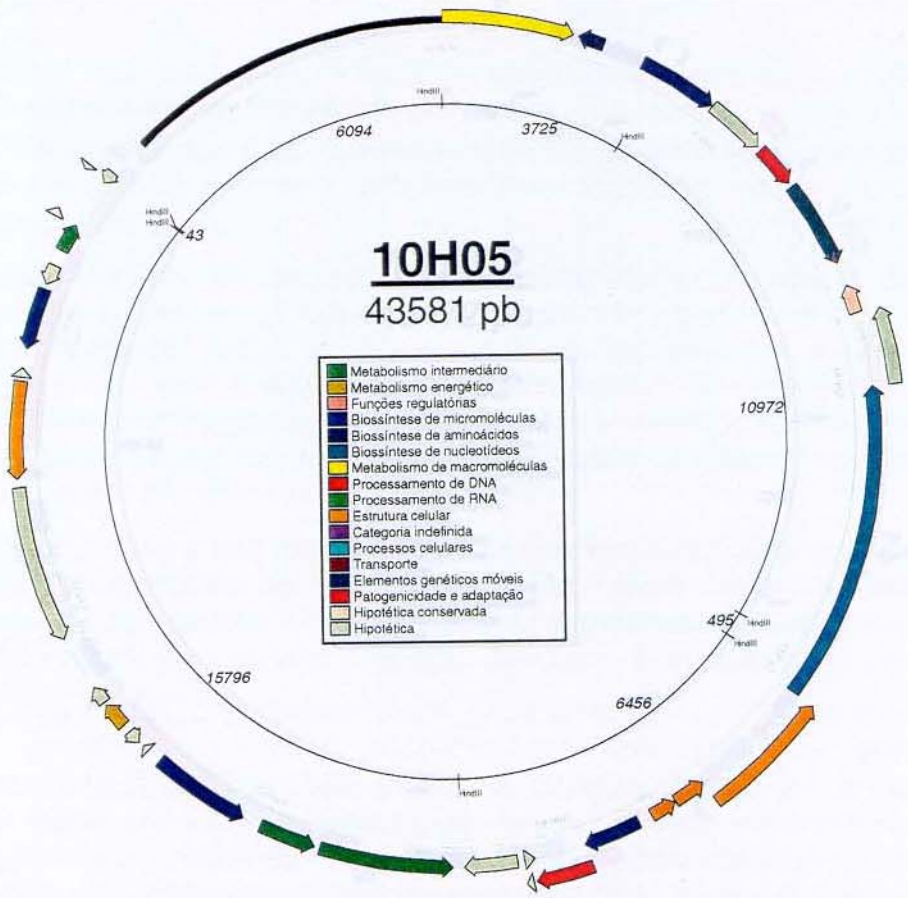


Figura 3. Mapa do cosmídeo 07A02.

O círculo representa a seqüência completa do cosmídeo. A barra preta, na parte de cima à esquerda, representa o vetor, Lawrist. As setas coloridas indicam a orientação de transcrição dos genes encontrados no inserto e sua cor, a classe a que as proteínas codificadas pertencem, segundo a legenda. Note a grande região destacada em rosa, referente ao operon gum. *Os números no círculo mais interno representam os tamanhos dos fragmentos obtidos por digestão com* Hind III. *O quadrado na parte inferrior da figura mostra, à esquerda, a diestão do DNA do clone e, à direita, o resultado da digestão eletrônica do clone montado. Os números entre as figuras representam os tamanhos dos fragmentos. Os da esquerda foram calculados a partir da migração observada. Aqueles contendo um asterisco (*) estão apontados como fora da faixa de precisão do programa de cálculo.*



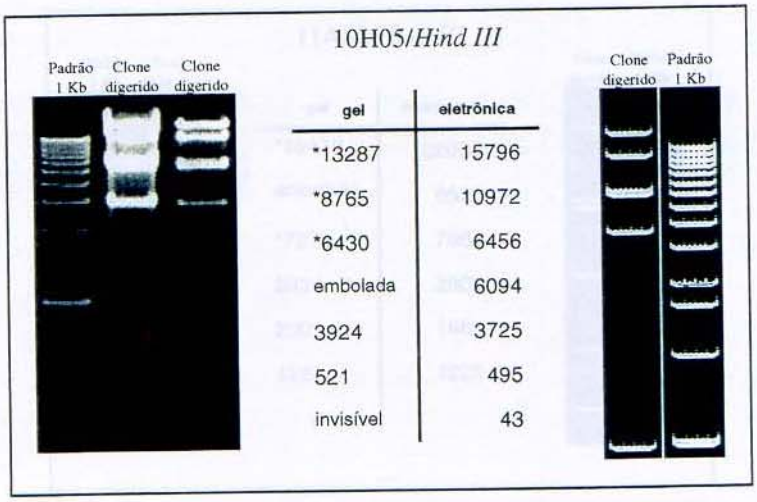
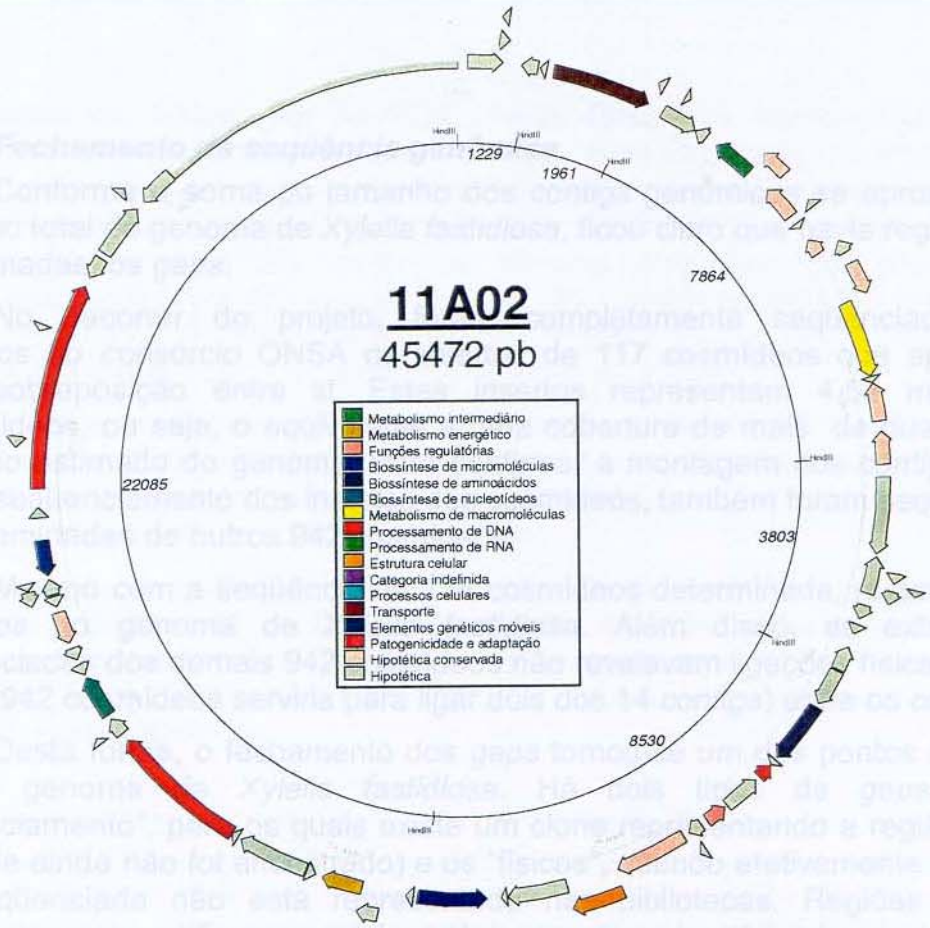


Figura 4. Mapa do cosmídeo 10H05.

O círculo representa a seqüência completa do cosmídeo. A barra preta, na parte de cima à esquerda, representa o vetor, Lawrist. As setas coloridas indicam a orientação de transcrição dos genes encontrados no inserto e sua cor, a classe a que as proteínas codificadas pertencem, segundo a legenda. Os números no círculo mais interno representam os tamanhos dos fragmentos obtidos por digestão com Hind III. *O quadrado na parte inferrior da figura mostra, à esquerda, a diestão do DNA do clone e, à direita, o resultado da digestão eletrônica do clone montado. Os números entre as figuras representam os tamanhos dos fragmentos. Os da esquerda foram calculados a partir da migração observada. Aqueles contendo um asterisco (*) estão apontados como fora da faixa de precisão do programa de cálculo.*



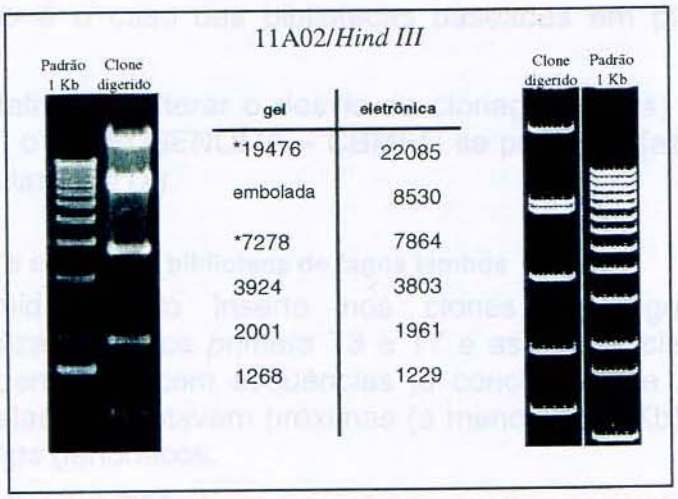


Figura 5. Mapa do cosmídeo 11A02.

O círculo representa a seqüência completa do cosmídeo. A barra preta, na parte de cima à esquerda, representa o vetor, Lawrist. As setas coloridas indicam a orientação de transcrição dos genes encontrados no inserto e sua cor, a classe a que as proteínas codificadas pertencem, segundo a legenda. Os números no círculo mais interno representam os tamanhos dos fragmentos obtidos por digestão com Hind III. *O quadrado na parte inferrior da figura mostra, à esquerda, a diestão do DNA do clone e, à direita, o resultado da digestão eletrônica do clone montado. Os números entre as figuras representam os tamanhos dos fragmentos. Os da esquerda foram calculados a partir da migração observada. Aqueles contendo um asterisco (*) estão apontados como fora da faixa de precisão do programa de cálculo.*

Fechamento da seqüência genômica

Conforme a soma do tamanho dos *contigs* genômicos se aproximava do tamanho total do genoma de *Xylella fastidiosa*, ficou claro que havia regiões ainda não clonadas, os *gaps*.

No decorrer do projeto, foram completamente seqüenciados pelos membros do consórcio ONSA os insertos de 117 cosmídeos que apresentam certa sobreposição entre si. Estes insertos representam 4,58 milhões de nucleotídeos, ou seja, o equivalente a uma cobertura de mais de duas vezes o tamanho estimado do genoma. Para confirmar a montagem dos *contigs* obtidos com o seqüenciamento dos insertos dos cosmídeos, também foram seqüenciadas as extremidades de outros 942 cosmídeos.

Mesmo com a seqüência de 114 cosmídeos determinada, restavam ainda 14 *gaps* no genoma de *Xylella fastidiosa*. Além disso, as extremidades seqüenciadas dos demais 942 cosmídeos não revelavam ligações físicas (*i.e.*, um destes 942 cosmídeos serviria para ligar dois dos 14 *contigs*) entre os *contigs*.

Desta forma, o fechamento dos *gaps* tornou-se um dos pontos críticos no projeto genoma da *Xylella fastidiosa.* Há dois tipos de *gaps*, os "de seqüenciamento", para os quais existe um clone representando a região faltante (mas ele ainda não foi encontrado) e os "físicos", quando efetivamente a região a ser seqüenciada não está representada nas bibliotecas. Regiões contendo seqüências que codifiquem proteínas tóxicas para as bactérias hospedeiras, *e.g.*, podem ser impossíveis de clonar em sistemas quem empreguem elementos replicativos, como é o caso das bibliotecas baseadas em plasmídeos ou em cosmídeos.

Numa tentativa de alterar o desvio de clonagem (*bias*) representado por estas estratégias, o grupo GENOMA – CBMEG se propôs a fazer uma biblioteca em bacteriófagos lambda (λ).

Construção e análise da biblioteca de fagos lambda

As extremidades do inserto nos clones de fago lambda foram seqüenciadas utilizando-se os *primers* T3 e T7 e as seqüências obtidas usadas em buscas de identidade com seqüências já concluídas de *Xylella fastidiosa*. Procurou-se aquelas que estavam próximas (a menos de 5 Kb) às extremidades dos grandes *contigs* genômicos.

Foram analisados 759 clones através do seqüenciamento de, pelo menos, uma das extremidades de seus insertos. Localizando-se as seqüências obtidas nos *contigs* então montados do genoma de *Xylella fastidiosa*, foi possível identificar sete fagos cujos insertos aparentemente cobririam regiões ainda não clonadas (Tabela 2).

Seqüenciamento de fagos lambda

A amplificação do inserto do clone de fago utilizando a metodologia descrita por Helms *et al.* (1987) para preparação de mini-lisados seguida de

amplificação do inserto por XL-PCR (Perkin-Elmer) foi empregada nos dois primeiros clones identificados (04L02 e 04L31). A conclusão da montagem destes insertos, contudo, revelou que o sistema introduzia erros com freqüência muito maior que a obtida com bibliotecas *shotgun*. Estes erros se evidenciaram pela presença, em cerca de 15% das leituras, de uma única base, com alta qualidade (*phred value* acima de 40) que discordava do consenso (também formado por leituras com bases de altíssima qualidade). Imagina-se que estas discordâncias teriam sido introduzidas na fase de amplificação do inserto.

Para se estar mais certo da correção da seqüência final, o número de leituras destas duas bibliotecas foi triplicado (passando de uma cobertura de cerca de seis vezes para mais de dezoito), de sorte a sempre haver mais de duas leituras concordantes em cada uma das fitas.

Para os demais fagos, o DNA para a sub-biblioteca de *shotgun* foi preparado utilizando-se o QIAGEN Lambda Kit (QIAGEN), seguindo-se as instruções do fabricante. A presença de discrepâncias entre leituras nestas bibliotecas foi comparável à obtida nas bibliotecas de *shotgn* de clones de cosmídeos ou plasmídeos.

Seqüenciamento de insertos de bibliotecas *shotgun* totais

Paralelamente ao esforço realizado no CBMEG com a construção das bibliotecas de bacteriófago lambda, a coordenação do projeto estava produzindo novas bibliotecas *shotgun* de plasmídeos, preparadas a partir de DNA genômico total de *Xylella*. Assim como no caso dos insertos dos fagos, alguns clones plasmidiais continham insertos que pareciam cobrir regiões inéditas do genoma. Coube ao grupo GENOMA do CBMEG realizar o seqüenciamento completo e montagem do inserto de 3 destes clones (Tabela 2).

Assim como já havia sido observado para a biblioteca de cosmídeo e as bibliotecas de *shotgun* total em plasmídeo, a biblioteca genômica em bacteriófago lambda também apresentou um desvio de amostragem. A distribuição dos fragmentos clonados não foi aleatória. Determinadas regiões do genoma estão representadas em vários clones, enquanto outras regiões não foram cobertas por clone algum. A grande vantagem da utilização de diferentes vetores para construção de bibliotecas está no fato de que, embora todas apresentem algum tipo de desvio de amostragem, as regiões pouco representadas, ou ausentes, em uma biblioteca podem estar bem representadas em uma outra biblioteca construída em um vetor diferente. Pode-se dizer que há uma espécie de complementação entre bibliotecas construídas com vetores distintos. Este fato pode ser claramente observado com as bibliotecas construídas para o seqüenciamento do genoma de *Xylella fastidiosa*.

Tabela 2. Clones seqüenciados para a resolução de *gaps.*

clone	Inserto (pb)	leituras
00J05	8338	189 (*125*)
00J11	4840	68 (*59*)
00J66	2233	96 (*89*)
04L02	13663	397 (*302*)
04L31	13918	394 (*323*)
08L88	11764	146 (*119*)
09L11	15536	164 (*137*)
20L02	12456	307 (*95*)
21L40	13333	143 (*124*)
21L77	16662	197 (*171*)

Os clones destacados em azul são provenientes de bibliotecas shotgun *de plasmídeo e os demais, da biblioteca de fagos lambda. A coluna Inserto indica o tamanho em pares de base do inserto. A coluna Leituras indica o total de leituras empregadas na montagem do clone. O número entre parênteses indica quantas das leituras eram oriundas apenas da sub-biblioteca deste determinado clone*

A determinação da seqüência completa do genoma de *Xylella fastidiosa* (2,7 Mb) aconteceu no início de janeiro de 2000. Como já era esperado, embora tenhamos conseguido concluir mais de 95% do seqüenciamento em pouco mais de 1 ano, a obtenção dos 5% restantes levou cerca de seis meses.

Anotação do genoma

À seqüência de DNA gerada por um projeto genoma é fundamental acrescentar informação biológica. Conquanto a comprovação de função biológica necessite de experimentos genéticos e/ou bioquímicos, a grande quantidade de seqüências existentes nos bancos de dados permite que a similaridade entre uma seqüência recém descrita e uma já caracterizada seja empregada na inferência da função da primeira.

A anotação de uma seqüência genômica é um trabalho contínuo, no qual novas informações são adicionadas ao longo dos anos. Em um primeiro instante é necessário localizar as seqüências que possivelmente codificam proteínas ou contém os genes de RNA ribossômico e transportador. O passo seguinte consiste na atribuição de função a estas moléculas putativas.

No caso dos RNAs, as estruturas secundária e terciária podem ser determinadas a partir da seqüência. Desta forma, a determinação da função destas moléculas é relativamente simples. Além disso, seu número no genoma é relativamente pequeno.

Em proteínas, entretanto, exceto para pequenos trechos de estrutura secundária, a estrutura primária não permite a determinação da forma (e, portanto, função) da molécula. Além disso, o número de seqüências que codificam estas moléculas no genoma é bem maior que o de seqüências codificadoras de RNAs (transportadores e ribossômicos). Nos genomas já concluídos, são descritas em média 35 vezes mais seqüências codificadoras de proteínas do que de RNAs (Tabela 3).

Tabela 3. Ocorrência de seqüências codificadoras em genomas bacterianos.

organismo	tamanho (pb)	RNAs	pb/RNA	proteínas	pb/proteína	proteína/RNA
Escherichia coli	4.639.221	**115**	40.341,1	**4.289**	1.081,7	**37,3**
Bacillus subtilis	4.214.814	**121**	34.833,2	**4.100**	1.028,0	**33,9**
Aquifex aeolicus	1.551.335	**51**	30.418,3	**1.522**	1.019,3	**29,8**
Archaeoglobus fulgidus	2.178.400	**49**	44.457,1	**2.407**	905,0	**49,1**
Helicobacter pylori	1.667.867	**43**	38.787,6	**1.553**	1.074,0	**36,1**

A coluna RNAs *inclui todas as moléculas putativas descritas como tais no genoma (excluindo-se mRNAs), mesmo as de função desconhecida. O mesmo ocorre para proteínas. As colunas* pb/RNA *e* pb/proteína *descrevem o a freqüência média (em pb) de ocorrência de RNAs e proteínas, respectivamente. Os cálculos de média estão arredondados para uma casa decimal. Interessante notar que a razão menos variável é* pb/proteína *(20%).* pb/RNA *(47%) e* proteína/RNA *(65%) são bem mais variáveis.*

A parte de anotação relacionada a proteínas pode ser dividida nos seguintes passos:

I. localização das ORFs na seqüência de DNA;

II. tradução das ORFs;

III. busca de similaridade entre cada uma das ORFs e seqüências já descritas;

IV. inferência de função a partir das similaridades encontradas.

Anotação manual

A anotação dos dois primeiros cosmídeos foi realizada antes da implementação do serviço de anotação automática (descrito mais adiante). Desta forma, procedeu-se a anotação destes clones de forma "manual". Conquanto bastante trabalhosa, a experiência adquirida neste processo foi empregada na elaboração do procedimento automático.

Fundamentalmente, empregou-se o Genemark.hmm (Lukashin and Borodovsky, 1998) na procura de ORFs na seqüência finalizada dos cosmídeos 07A01 e 07A02. Estas ORFs foram traduzidas utilizando-se o software GeneRunner (Hastings Software) e a seqüência de aminoácidos foi utilizada na procura de similaridades com seqüências já descritas presentes no GenBank (Benson *et al.*, 2000), com o uso do BLAST (Altschul *et al.*, 1997). A anotação foi

feita no software sequin (Benson *et al.*, 2000). Instruções detalhadas do processo ainda podem ser encontradas na antiga página do projeto *Xylella fastidiosa* (http://www.lbi.ic.unicamp.br/xf-old/Annotation_primer.html) e estão reproduzidas no Apêndice C.

Estas instruções estavam disponibilizadas devido a mais uma idiossincrasia deste projeto. Normalmente inicia-se a anotação de um genoma após a conclusão de seu seqüenciamento. O projeto genoma de *Xylella fastidiosa* estava baseado em cosmídeos contendo, em média, insertos de 38,9 Kb. Cada laboratório era responsável por gerar a seqüência do cosmídeo com qualidade excepcional. Desta forma, cada laboratório tinha, no final do processo, um trecho de cerca de 40 kb do genoma de *Xylella fastidiosa*. Já que um dos objetivos do projeto era justamente capacitar os laboratórios em análise genômica (daí a insistência para que cada um deles montasse seu próprio cosmídeo), nada mais natural que responsabilizá-los pela anotação dos insertos de seus cosmídeos.

A necessidade de se homogeneizar a anotação do genoma, contudo, acabou inviabilizando a anotação dos cosmídeos por quem os seqüenciava. Um grupo de anotação, que se reunia periodicamente para ficar mais coeso, foi formado. A estratégia de anotação baseada em cosmídeos, entretanto, permitiu que finalizasse a anotação paralelamente à finalização do seqüenciamento.

Anotação automática

O volume de dados a serem analisados em um projeto genoma torna necessária alguma automatização no processo. As primeiras anotações, realizadas "manualmente", revelaram passos do processo que poderiam ser automatizados, tornando a tarefa de anotação simultaneamente mais organizada e menos tediosa.

Desta forma foi desenvolvido um serviço de "anotação automática", em parceria com os membros do LBI. Este serviço gerou uma primeira anotação (chamada de "preliminar-automática") para cada cosmídeo que atingia a situação de *close to finished* (nesta situação, admite-se 0,1% de tolerância nos critérios de fechamento de clones já descritos) e simultaneamente disponibilizava uma série de ferramentas para o anotador responsável por ele.

O processo de criação da anotação "preliminar-automática" , ilustrado na Figura 6, envolve o fluxo de informações que se inicia com uma seqüência FASTA (Pearson, 1990) e termina com um arquivo no formato ASN.1. (Abstract Syntax Notation One) que é o formato empregado pelo programa de submissão de seqüências do NCBI, o sequin (Benson *et al.*, 2000).

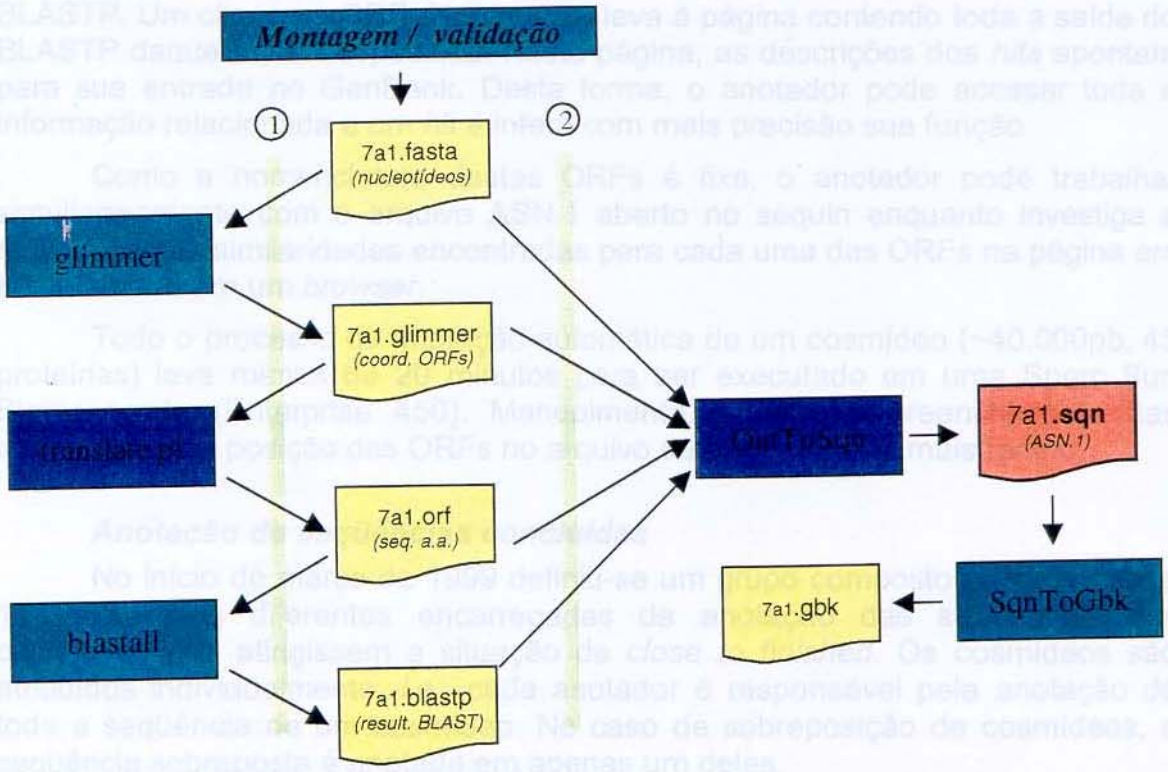


Figura 6. Fluxo de informações da Anotação Automática.

Um *script* em perl (Wall *et al.*, 1996), desenvolvido juntamente com os membros do LBI (laboratório de bioinformática, DCC – UNICAMP), **makesqn.pl**, se encarrega de fazer a chamada a cada um dos programas descritos nos quadrados da Figura 6.

Ao ser disparado, ele executa o glimmer contra o arquivo fasta. Utiliza, então, o arquivo de saída do glimmer como entrada para um outro *script* perl, **translate.pl**, encarregado de fazer a tradução das ORFs.

O arquivo de saída do translate.pl é empregado em um BLASTP contra o banco *nr* do GenBank (Benson *et al.*, 2000). Informações dos quatro arquivos são empregadas por um terceiro *script*, **OutToSqn**, na construção do arquivo ASN.1 compatível com o sequin. Este arquivo ASN.1 contém a seqüência nucleotídica original, genes relativos a cada uma das ORFs encontradas, com coordenadas relativas a esta seqüência, um nome apropriado e único para cada um dos genes, a seqüência de aminoácidos resultante da tradução de cada um dos genes e, associada a esta seqüência de aminoácidos, a descrição do *hit* mais alto encontrado no GenBank.

Para facilitar ainda mais o processo de anotação mais dois *scripts* são executados: o **SqnToGbk** transforma a informação presente no arquivo ASN.1 em um formato semelhante ao encontrado nas páginas de descrição de seqüências do GenBank, mais agradável para a leitura humana; o **parseador.pl** cria um arquivo HTML contendo a lista das ORFs que foram empregadas no

BLASTP. Um clique na ORF de interesse leva à página contendo toda a saída do BLASTP daquela ORF específica. Nesta página, as descrições dos *hits* apontam para sua entrada no GenBank. Desta forma, o anotador pode acessar toda a informação relacionada a um *hit* e inferir com mais precisão sua função.

Como a nomenclatura destas ORFs é fixa, o anotador pode trabalhar simultaneamente com o arquivo ASN.1 aberto no sequin enquanto investiga a relevância das similaridades encontradas para cada uma das ORFs na página em HTML aberta em um *browser*.

Todo o processo de anotação automática de um cosmídeo (~40.000pb, 45 proteínas) leva menos de 20 minutos para ser executado em uma Sparc Sun Biprocessada (Enterprise 450). Manualmente, apenas o preenchimento das informações de posição das ORFs no arquivo sequin consome mais tempo.

Anotação de seqüências concluídas

No início de março de 1999 definiu-se um grupo composto de pessoas de 15 laboratórios diferentes encarregadas da anotação das seqüências dos cosmídeos que atingissem a situação de *close to finished*. Os cosmídeos são atribuídos individualmente, *i.e.*, cada anotador é responsável pela anotação de toda a seqüência de um cosmídeo. No caso de sobreposição de cosmídeos, a seqüência sobreposta é anotada em apenas um deles.

O autor realizou a anotação de 12 cosmídeos, 4 deles apenas da parte não sobreposta a cosmídeos anotados por outras pessoas. No total foram analisadas 466 ORFs putativas distribuídas ao longo de 366.080 pb, o que corresponde a aproximadamente 1/7 do tamanho total do genoma de *Xylella fastidiosa*.

O genoma completo de Xylella fastidiosa

A partir da classificação dos genes de *E.coli* proposta por Riley (1998), uma lista de categorias foi definida pelo grupo de anotação para *Xylella fastidiosa*. A lista completa pode ser vista no Apêndice D. A seguir, descreve-se o número de genes encontrados nas principais categorias:

I. Intermediary metabolism (**232**)
   - A. Degradation (**27**)
   - B. Central intermediary metabolism (**44**)
   - C. Energy metabolism, carbon (**84**)
   - D. Fermentation (**2**)
   - E. General regulatory functions (**39**)

II. Biosynthesis of small molecules (**218**)
   - A. Amino acids (**77**)
   - B. Nucleotides (**41**)
   - C. Sugars and sugar nucleotides (**2**)
   - D. Cofactors, prosthetic groups, carriers (**74**)
   - E. Fatty acid and phosphatidic acid biosynthesis (**21**)
   - F. Polyamines (**3**)

III. Macromolecule metabolism (**321**)
   A. DNA (**111**)
   B. RNA (**137**)
   C. Protein (**70**)
   D. Other macromolecules (**9**)

IV. Cell structure (**119**)
   A. Membrane components (**40**)
   B. Murein sacculus, peptidoglycan (**25**)
   C. Surface polysaccharides and antigens (**26**)
   D. Surface structures (**28**)

V. Cellular processes (**129**)
   A. Transport (**107**)
   B. Cell division (**21**)
   C. Chemotaxis and mobility (**1**)
   D. Osmotic adaptation (**0**)
   E. Cell killing (**0**)

VI. Mobile genetic elements (**91**)
   A. Phage-related functions and prophages (**60**)
   B. Plasmid-related functions (**24**)
   C. Transposon-related functions (**7**)

VII. Pathogenicity, virulence, and adaptation (**131**)
   A. Avirulence (**5**)
   B. Hypersensitive response and pathogenicity (**3**)
   C. Toxin production and detoxification (**36**)
   D. Host cell wall degradation (**9**)
   E. Exopolysaccharides (**10**)
   F. Surface proteins (**12**)
   G. Adaptation, atypical conditions (**30**)
   H. Other (**26**)

VIII. Hypothetical, unknown, dubious (**1534**)
   A. Conserved proteins with unknown functions (**318**)
   B. No hits/only low score hits (**1216**)

Nesta lista é possível notar que, dos 2853 possíveis genes codificadores de proteínas encontrados no genoma, cerca de 54% não puderam ser classificados e 43% simplesmente não possuíam um possível homólogo já depositado em bancos de dados. Estes números parecem excessivamente grandes, e realmente são quando comparados a dados obtidos por genomas como o de *Mycoplasma pneumonia* (Himmelreich *et al.*, 1996) onde apenas 9.9% dos genes descritos não apresentam similares nos bancos de dados. Alguns genomas, contudo, apresentam números ainda maiores, como é o caso de *Pyrococcus horikoshii* (Kawarabayasi *et al.*, 1998) onde apenas 19.7% dos genes preditos têm similares já descritos de função conhecida, 22% são similares a

proteínas hipotéticas e 58.3% carecem de qualquer similaridade significativa com seqüências já descritas.

Tabela 4. Cosmídeos com anotação completa.

cosmídeo	ORFs	bases anotadas
01G04	43	37493
01H09	47	37445
03H12	32	16480
05F05	49	39596
07A01	59	40347
07A02	44	38685
07A11	13	12702
07B07	46	39733
07B08	13	11662
07G02	40	36500
10H05	10	15360
11A02	70	40077
TOTAL	**466**	**366080**

A coluna ORFs indica o total de quadros de leitura encontrados pelo programa glimmer. A coluna bases anotadas *mostra o tamanho total do cosmídeo ou (destacado em azul) a porção efetivamente anotada em caso de sobreposição com outro cosmídeo. Os cosmídeos destacados em vermelho foram seqüenciados pelo CBMEG.*

Goma fastidiana

A despeito de não ser possível, neste instante, inferir a função de cerca de metade das proteínas descritas no genoma, as cerca de 1300 outras proteínas revelaram uma série de fenômenos biológicos bastante interessantes. Um deles, particularmente, nos chamou a atenção.

O trabalho de anotação manual do segundo cosmídeo cujo seqüenciamento foi concluído pelo grupo genoma do CBMEG (07A02 – Figura 3) revelou a presença de um agrupamento de genes similares aos responsáveis pela síntese da goma xantana.

A goma xantana é um exopolissacarídeo (EPS) diretamente envolvido com a virulência da bactéria fitopatogênica *Xanthomonas campestris* pv. *Campestris* (Chou *et al.*, 1997). Está bastante claro que, em ambientes naturais, agrícolas, industriais e médicos, a maior parte das bactérias não cresce individualmente em suspensão, mas colonizam superfícies organizadas em comunidades usando biofilmes (Stickler, 1999). A importância de biofilmes na patogenicidade de diversos organismos foi comprovada (Costerton *et al.*, 1999). Já foi mesmo demonstrada a relação direta entra a presença de goma, descrita apenas morfologicamente, e os sintomas causados por *Xylella fastidiosa* em cafeeiros (Queiroz -Voltan *et al.*, 1998).

O trabalho reproduzido a seguir, publicado na edição de setembro de 2001 da FEMS Microbiology Letters, descreve detalhadamente as evidências que levam a crer que *Xylella fastidiosa* produz um exopolissacarídeo semelhante à goma xantana. Esta nova goma, batizada de fastidiana, já foi comentada por outros autores em periódicos científicos de grande renome, como *Science* (Hagmann, 2000), *Nature* (Bevan, 2000) e *Plant Cell* (Kamoun and Hogenhout, 2001) e foi patenteada pela FAPESP (Arruda *et al.*, 2000).

ELSEVIER

FEMS Microbiology Letters 203 (2001) 165–171

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Fastidian gum: the *Xylella fastidiosa* exopolysaccharide possibly involved in bacterial pathogenicity

Felipe Rodrigues da Silva [a], André Luiz Vettore [a], Edson Luis Kemper [a], Adilson Leite [a], Paulo Arruda [a,b,*]

[a] *Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6010, CEP 13083-970, Campinas, SP, Brazil*

[b] *Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Caixa Postal 6109, CEP 13083-970, Campinas, SP, Brazil*

Received 14 June 2001; accepted 10 July 2001

First published online 29 August 2001

Abstract

The Gram-negative bacterium *Xylella fastidiosa* was the first plant pathogen to be completely sequenced. This species causes several economically important plant diseases, including citrus variegated chlorosis (CVC). Analysis of the genomic sequence of *X. fastidiosa* revealed a 12 kb DNA fragment containing an operon closely related to the *gum* operon of *Xanthomonas campestris*. The presence of all genes involved in the synthesis of sugar precursors, existence of exopolysaccharide (EPS) production regulators in the genome, and the absence of three of the *X. campestris gum* genes suggested that *X. fastidiosa* is able to synthesize an EPS different from that of xanthan gum. This novel EPS probably consists of polymerized tetrasaccharide repeating units assembled by the sequential addition of glucose-1-phosphate, glucose, mannose and glucuronic acid on a polyprenol phosphate carrier.

Keywords: Exopolysaccharide; Citrus variegated chlorosis; Plant pathogen; Biofilm; Gum; Genome analysis

1. Introduction

Polysaccharides are important constituents of the surface of bacterial cells, and play a critical role in the interaction of bacteria with the environment. Many bacteria produce complex exopolysaccharides (EPSs), which can remain attached to the cell surface in a capsular form or be released as slime. Rather than growing as individuals cells, in natural habitats most bacteria colonize surfaces in organized biofilm communities [1]. These biofilms consist of microorganisms immobilized on a variety of polymeric compounds, with EPSs as the major constituent, along with proteins, nucleic acids, lipids and humic substances [2].

The formation of biofilms is an important factor in the pathogenicity of several organisms [3]. Examples include the biofilm produced by *Pseudomonas aeruginosa*, an infectious bacterium in humans [4], and those produced by *Ralstonia solanacearum*, the ESP of which is important for the rapid systemic colonization of tomato plants and the subsequent symptoms caused by bacterial infection [5]. The EPS produced by *Xanthomonas campestris* may be required for this species' virulence [6].

Xylella fastidiosa is the causal agent of economically important plant diseases including Pierce's disease of grapevine, alfalfa dwarf, phony peach disease, periwinkle wilt, leaf scorch in several plant species, and citrus variegated chlorosis (CVC) [7]. CVC, which has so far been found only in Brazil and Argentina, is a major concern to the citrus industry, and is considered to be potentially the most devastating citrus disease. *X. fastidiosa* was the first phytopathogenic bacterium to have its genome completely sequenced [8], which makes this species an interesting model for understanding pathogenicity of plant bacteria. The *X. fastidiosa* genome comprises two plasmids and one circular chromosome encoding genes important for metabolic functions and virulence [8]. One of the interesting features of the *X. fastidiosa* genome, which perhaps

* Corresponding author. Tel.: +55 (19) 37881137; Fax: +55 (19) 37881089.

E-mail address: parruda@unicamp.br (P. Arruda).

0378-1097/01/$20.00
PII: S0378-1097(01)00348-2

represents a major factor in this species' pathogenicity, is the presence of genes related to the biosynthesis of an EPS. *X. fastidiosa* has 22 genes encoding regulatory proteins and enzymes involved in the synthesis of an EPS similar to xanthan gum, the EPS produced by *X. campestris*. In this work, we discuss the characteristics of these genes and propose a putative pathway for the synthesis of *X. fastidiosa* EPS, as well as its possible chemical nature.

2. Materials and methods

Gene identification in the completed *X. fastidiosa* sequence was done using GLIMMER [9] and the GLIMMER post-processor RBSfinder (S.L. Salzberg, unpublished, http://www.tigr.org/softlab). Similarity searches between the translated open reading frame (ORF) sequences and previously described proteins were done with gapped BLAST [10] using the NCBI nr protein database.

The statistical significance of sequence similarities between putative homologues was established using the RDF2 [11] and Gap [12] programs with 1000 random shuffles. Binary comparison scores were expressed in standard deviations [13]. A value of 10 standard deviations was considered sufficient to establish homology [14]. Homologous sequences were aligned using CLUSTAL X [15]. Identity between homologues was calculated as the percentage of identical residues in an alignment. Mean hydropathy analyses were done as described by Kyte and Doolittle [16] with a sliding window of 20 residues. PSORT [17] was used to tentatively predict the cellular location and transmembrane domains of the proteins.

3. Results

The sequencing of the *X. fastidiosa* genome revealed the presence of a 12 kb DNA fragment containing nine ORFs, in a typical operon structure, which showed homology to genes related to EPS production in several bacterial species. These ORFs shared the highest similarity with genes in the *X. campestris* xanthan gum operon (Table 1). The xanthan gum operon of *X. campestris* consists of a tandem array of 12 genes, *gumB*, *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J*, *K*, *L* and *M*, which encode the enzymes responsible for EPS synthesis [6]. Of the 12 genes in the xanthan gum operon, the 12 kb DNA fragment of *X. fastidiosa* had nine ORFs homologous to genes *gumB*, *C*, *D*, *E*, *F*, *H*, *J*, *K* and *M* (Fig. 1A). The genes *gumG*, *I* and *L* were missing from this DNA fragment and did not occur at any other site in the *X. fastidiosa* genome, including the plasmids. Because of the high similarity to the *X. campestris* operon, we named this cluster of ORFs 'the fastidian gum operon' and its constituent genes were given the same names as the corresponding genes in the xanthan gum operon. The genes in both gum operons were arranged in an identical genomic structure (Fig. 1A).

The biosynthesis of xanthan gum requires sugar precursors and regulatory proteins [18]. A list of already identified genes involved in the regulation of xanthan biosynthesis and the biosynthesis of sugar precursors is shown in

Table 1
Putative genes related to fastidian gum production in *X. fastidiosa*

Class[a]	*Xylella* gene ID	Homologous gene[b]	% Identity	Function	Reference
Regulator	XF0287	*rpfB*	72.3	regulatory protein (DSF)	[32]
Regulator	XF0290	*rpfA*	80.0	aconitase	[31]
Regulator	XF1109	*rpfE*	65.2	regulatory protein	[18]
Regulator	XF1113	*rpfG*	77.0	two-component system, regulatory protein	[28]
Regulator	XF1114	*rpfC*	60.0	fused two-component sensor-regulator protein	[28]
Regulator	XF1115	*rpfF*	65.7	regulatory protein (DSF)	[32]
Precursor	XF0232	*pgi*	79.1	glucose-6-phosphate isomerase	[34]
Precursor	XF0259	*xanB*	84.5	phosphomannose isomerase-GDP-mannose pyrophosphorylase	[20]
Precursor	XF0260	*xanA*	84.8	phosphoglucomutase/phosphomannomutase	[20]
Precursor	XF1064	*glk*	41.4	glucose kinase	[35]
Precursor	XF1460	*glk*	32.7	glucose kinase	[36]
Precursor	XF1606	*algD*	66.1	UDP-glucose dehydrogenase	[37]
Precursor	XF2432	*gtaB*	81.8	UTP-glucose-1-phosphate uridylyltransferase	[38]
EPS-synt	XF2360	*gumM*	63.1	GumM protein	[6]
EPS-synt	XF2361	*gumK*	68.7	GumK protein	[6]
EPS-synt	XF2364	*gumH*	64.7	GumH protein	[6]
EPS-synt	XF2365	*gumF*	41.9	GumF protein	[6]
EPS-synt	XF2367	*gumD*	73.6	GumD protein	[6]
EPS-exp	XF2362	*gumJ*	62.7	GumJ protein	[6]
EPS-exp	XF2366	*gumE*	59.9	GumE protein	[6]
EPS-exp	XF2369	*gumC*	61.2	GumC protein	[6]
EPS-exp	XF2370	*gumB*	67.1	GumB protein	[6]

[a]EPS-synt: genes related to the synthesis of the repeating tetramer of the EPS. EPS-exp: genes related to the export/polymerization of the EPS.
[b]The homologous genes for *glk* are from *E. coli* (XF1064) and *Zymomonas mobilis* (XF1460). All others are from *X. campestris*.

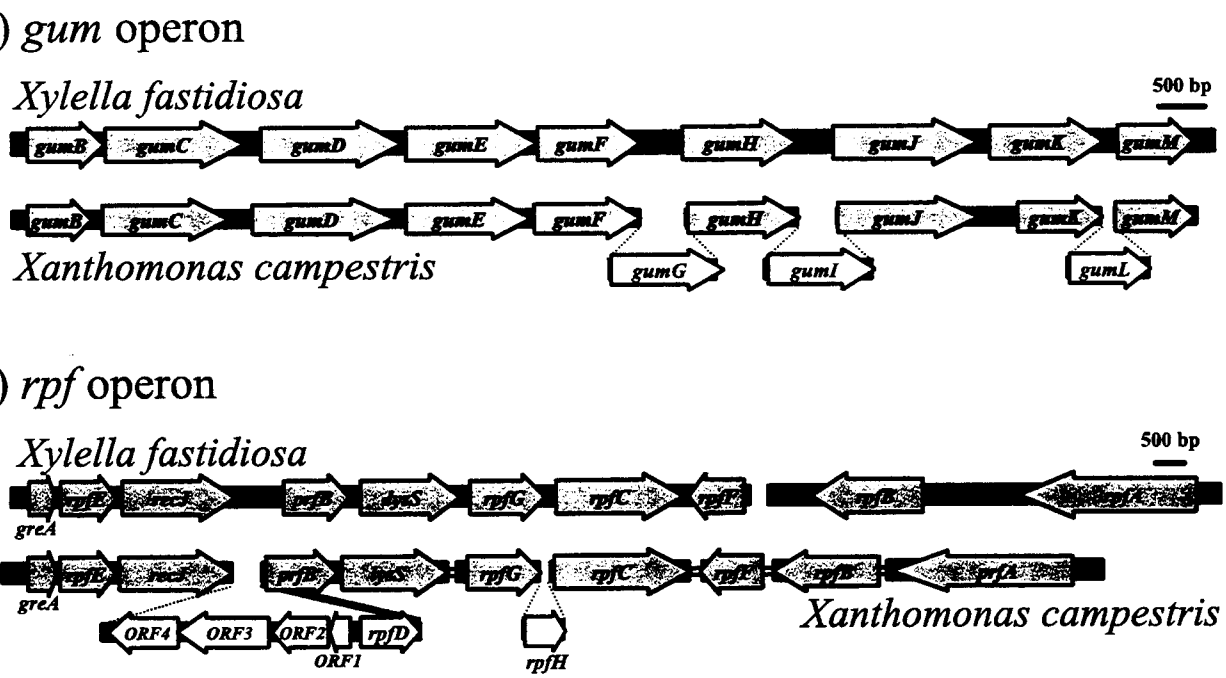


Fig. 1. Genetic map of the *X. campestris gum* (A) and *rpf* (B) operons compared to the region of the homologous genes in the *X. fastidiosa* genome. The black bar represents the DNA sequence and the overlying arrows represent the genes. The 'white' genes in the *X. campestris* operons are not found in *X. fastidiosa* and are drawn shifted out of the sequence in order to maintain the alignment. Regions described as contiguous in the *X. campestris rpf* operon, but not described as a single sequence (or two sequences that overlap) are joined by two thin black lines. The thin black bar above each figure represents the length of 500 nucleotides. Note that the scale is not the same for both figures.

Table 1. All the genes related to the synthesis of the precursor molecules needed as substrates for the activity of the enzymes encoded by the *gum* operon were found in the *X. fastidiosa* genome (Table 1). Most of these genes encode enzymes involved in the basal carbohydrate metabolism common to almost all living organisms. In microorganisms, these genes are regulated by processes directed to specific requirements related to the habitat or to environmental changes, and are therefore arranged in clusters or operons. In *Escherichia coli* and other bacteria [19], the genes encoding enzymes responsible for the production of some of the sugar precursors depicted in Fig. 2 are arranged in operons. Curiously, this is not the case for *X. fastidiosa*, where these genes are scattered throughout the genome (http://www.lbi.ic.unicamp.br/xf/), except for the genes *xanA* and *xanB* (Fig. 2), which are contiguous in both *X. fastidiosa* and *X. campestris* [20].

Several genes involved in the regulation of xanthan biosynthesis in *X. campestris* have been identified by mutagenesis [18], which in many cases affects the ability of bacteria to produce EPS and also decreases bacterial virulence. For this reason these genes are known as *rpf* (regulation of pathogenicity factors) genes. Six sequences homologous to the *rpf* genes of *X. campestris* were found in the *X. fastidiosa* genome (Table 1). Whereas the *gum* operon structure was considerably conserved in both species, the genes in the *rpf* operon of *X. campestris* were split into two segments in the *X. fastidiosa* genome (Fig. 1B). No other putative EPS regulatory gene was found in the *X. fastidiosa* genome.

4. Discussion

The EPS-related genes found in *X. fastidiosa* showed the highest similarity to *X. campestris* genes. *X. campestris* produces an EPS known as xanthan gum, which is of great industrial importance and is involved in the pathogenicity of *X. campestris* [21]. Xanthan is a polymer consisting of repeating pentasaccharide units with the structure mannose-1,4-β-glucuronic acid-1,2-β-mannose-1,3-α-cellobiose [22] (Fig. 3). Acetyl groups can be present to varying degrees in the two mannoses as 6-*O* substituents, and a pyruvic acid moiety joined by ketal linkage may also occur in some of the terminal mannoses [23].

The biosynthetic pathway of xanthan involves at least three stages: (i) conversion of simple sugars to nucleotidyl derivative precursors, (ii) assembly of pentasaccharide subunits attached to an inner-membrane polyprenol phosphate carrier, with addition of acetyl and pyruvate groups, and (iii) polymerization of the pentasaccharide repeating units and secretion of the polymer [24]. The genes encoding all the enzymes involved in the last two stages are located in a cluster of 12 genes (Fig. 1A) encompassing a 16 kb region of the *X. campestris* genome known as *xpsI* or *gum* [6]. Conversely, only two (*xanA* and *xanB*, Fig. 2)

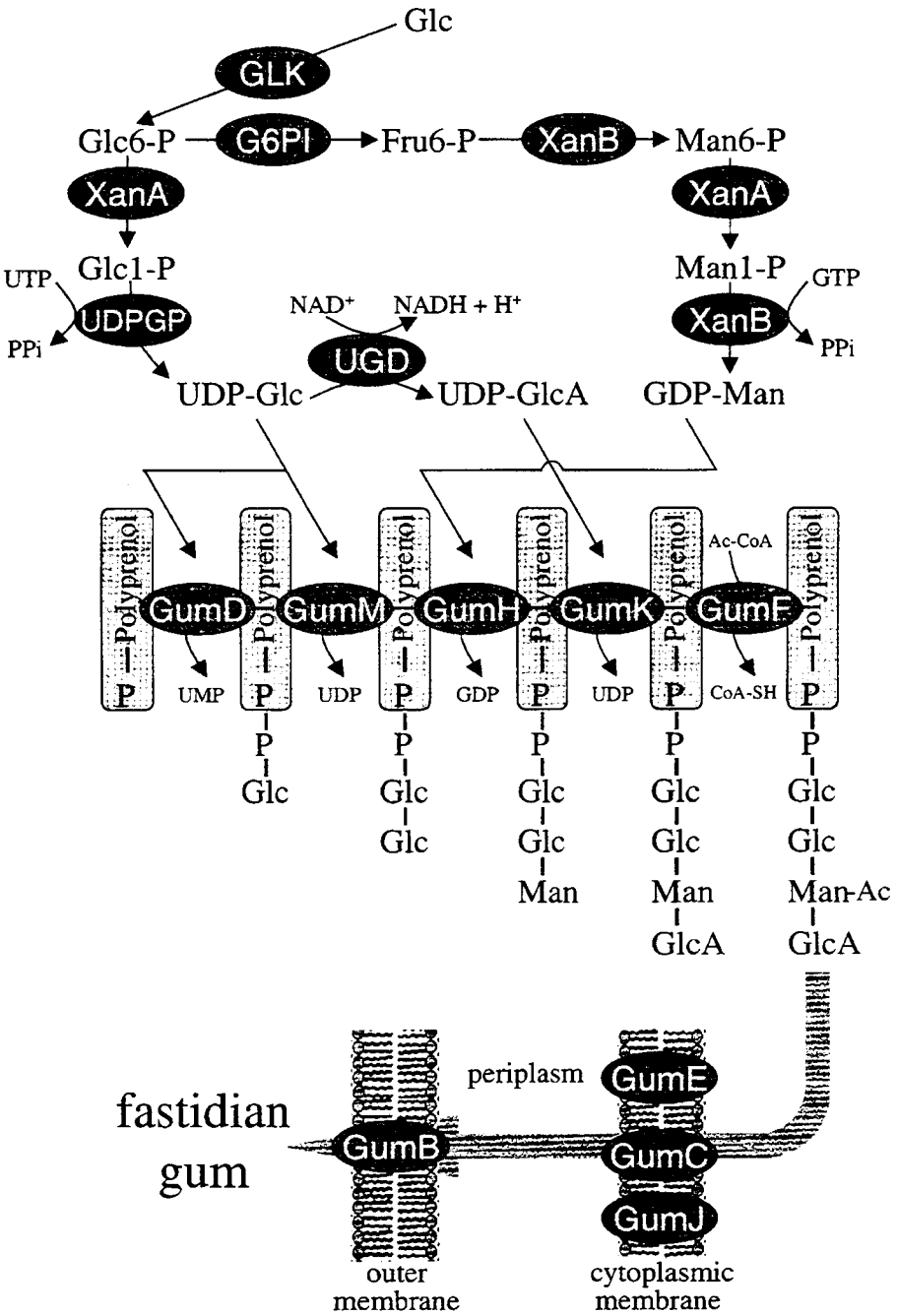


Fig. 2. Proposed synthesis pathway for fastidian gum. Putative proteins are shown in the gray ovals. The names of the enzymes correspond to those shown in Table 1. The arrows for the sequential reactions which add sugars to polyprenol are not shown. Glc, D-glucose; GlcA, D-glucuronic acid; Man, D-mannose; Ac, acetyl ester; PPi, pyrophosphate.

of the five genes already associated with the first stage are genetically linked [25]. Transcriptional analysis has shown that the *gum* genes are expressed mainly as an operon from a promoter upstream of the first gene, *gumB* [26]. Secondary (weak) promoters may also exist upstream of *gumK* [26] and *gumD* [27]. The synthesis of xanthan gum is coordinately regulated by the cluster of *rpf* genes [28].

The function of some of the *X. campestris gum* gene products has been established biochemically [6]. GumD, GumM, GumH, GumK and GumI are responsible for the assembly of the pentasaccharide lipid intermediate. GumL is the ketal pyruvate transferase and GumF and GumG are the acetyltransferases. GumB, GumC and GumE are related to the polymerization and translocation of xanthan gum. GumJ is speculated to also have a role in the polymerization/translocation process [29].

Since the nine *X. fastidiosa gum* genes share high similarity with *X. campestris* genes, and both bacteria have

these genes in similarly structured operons, we hypothesize that *X. fastidiosa* should be able to synthesize an EPS with a structure different from that of xanthan gum. The three *X. campestris* genes not found in the *X. fastidiosa* gum operon (*gumG*, *I* and *L*) are all related to the assembling and/or modification of the last mannose of the pentamer repeating unit of xanthan gum. *X. campestris* GumI adds a mannose to the tetrasaccharide lipid intermediate. GumL and GumG decorate this mannose with pyruvyl and acetyl groups, respectively [6]. As these genes are not found elsewhere in the *X. fastidiosa* genome, the repeating unit of the *X. fastidiosa* EPS is most probably a tetramer (Fig. 3B). This tetrasaccharide would be assembled by the sequential addition of UDP-glucose, UDP-glucose, GDP-mannose and UDP-glucuronic acid on a polyprenol phosphate carrier, by the glycosyltransferase I, II, III and IV homologues encoded by *gumD*, *gumM*, *gumH* and *gumK* genes. The absence of *gumL* and *gumI* in *X. fastidiosa* would not prevent the EPS assembling since in *X. campestris* these genes do not appear to interfere with the rate of polysaccharide production [6].

The *gumF* gene encodes an acetyltransferase I homologue which would catalyze the acetylation of the mannose residue (Fig. 2). For xanthan gum, the acetylation occurs at the prenyl-phospho-sugar stage and can also take place at the trisaccharide-diphosphate-prenol level [30]. These findings corroborate the hypothesis of an acetylated EPS being produced by *X. fastidiosa*.

While the roles of the genes responsible for the assembling, acetylation and pyruvylation of the pentasaccharide repeating units in *X. campestris* have been very well characterized by biochemical analysis of gum mutants [6], the roles of the remaining genes are unknown, but may be associated with the assembling and export of the EPS. The GumE protein has so far been described only in *X. fastidiosa* and *X. campestris*. In both species, this protein is predicted to be an inner-membrane protein [17], with 11 transmembrane domains in *X. fastidiosa* and nine transmembrane domains in *X. campestris* (Fig. 4). Since *X.*

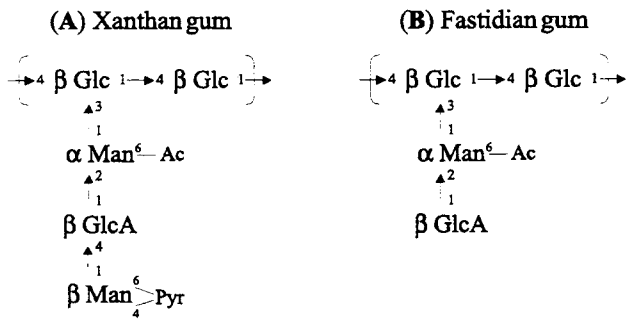


Fig. 3. Structure of the repeating unit of xanthan gum (A) from *X. campestris*, and the proposed structure of fastidian gum (B) from *X. fastidiosa*. The arrows point towards the reducing end of each repeat. Glc, D-glucose; GlcA, D-glucuronic acid; Man, D-mannose; Ac, acetyl ester; Pyr, acetyl-linked pyruvic acid. Some external mannoses of xanthan gum may contain an acetyl instead of a pyruvyl substituent [23]. The structure of the xanthan gum is based on Jansson et al. [22].

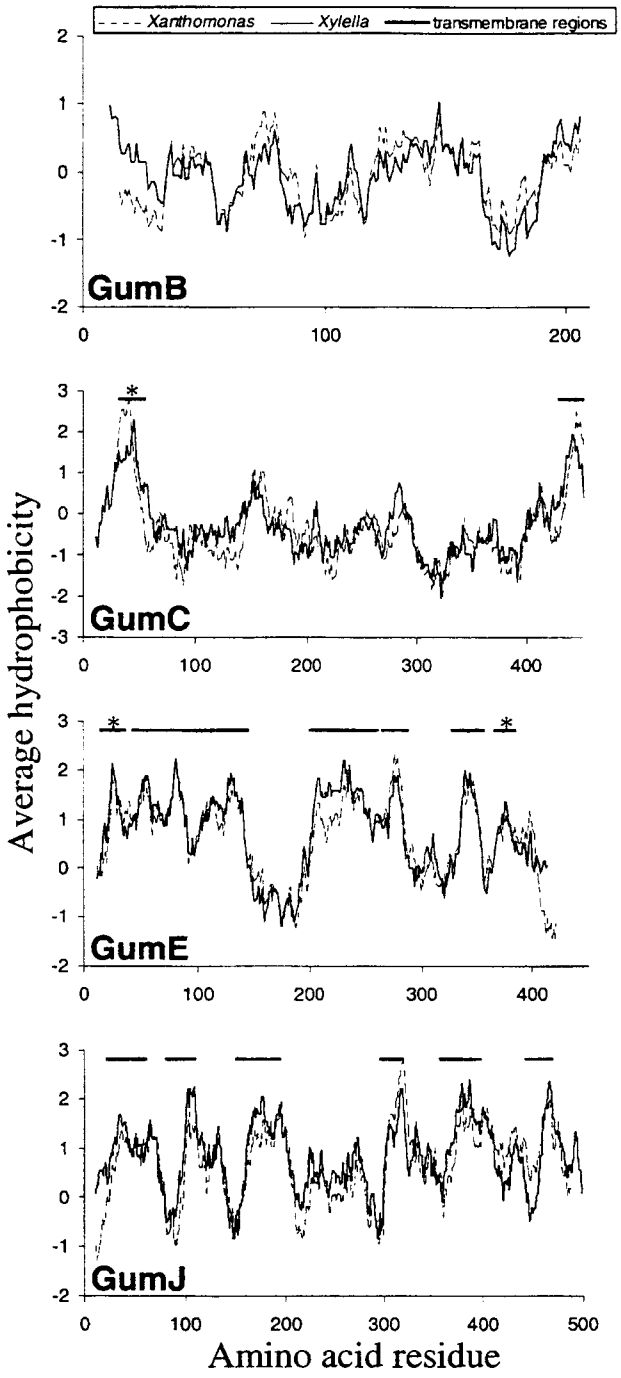


Fig. 4. Mean hydropathy plots for GumB, GumC, GumE and GumJ from *X. fastidiosa* (continuous line) and *X. campestris* (dashed line). The thick lines above the graphs show the segments of the *X. fastidiosa* protein predicted to be transmembrane domains by PSORT. All transmembrane regions identified in *X. fastidiosa* were also predicted in the homologous *X. campestris* protein, except those indicated by an asterisk (*). Position refers to the amino acid position in the *X. fastidiosa* protein. For the best alignment of the profiles, the *X. campestris* sequences have been shifted to the right by four amino acids for GumB, and by 18 amino acids for the GumC graph.

campestris strains lacking *gumE* gene appear to accumulate complete xanthan pentasaccharide subunits in vitro but are unable to synthesize the polymer, the gene is believed to be directly involved in the polymerization of xanthan gum [6].

Paulsen et al. [29] analyzed the families of homologous proteins that mediate the excretion of complex carbohydrates across the bacterial cell membrane. These authors identified a novel family of putative transporters, which they call PST (polysaccharide-specific transport) involved in the translocation of EPS substrates across the cytoplasmic membrane of Gram-negative bacteria. In this model, the PST system requires the presence of a cytoplasmic-membrane-periplasmic auxiliary protein (MPA2) as well as an outer-membrane auxiliary protein (OMA). Based on the criteria used by Paulsen et al. [29], the *X. fastidiosa* proteins GumJ, GumC and GumB belong to the PST, MPA2 and OMA families, respectively. The hydrophobicity profiles of those proteins are very similar between *X. fastidiosa* and *X. campestris*, so their localization in membrane is very plausible. Since a lipid-linked tetrasaccharide fulfills the substrate requirements for polymerization in *X. campestris* [6], we propose that the *X. fastidiosa* GumE, GumJ, GumC and GumB proteins would polymerize and export the fastidian EPS. The conservation and uniqueness of the GumE protein in both species suggest a particular class of PST system.

As shown in Fig. 1B, *X. fastidiosa* contains homologues to all 16 genes of the *X. campestris rpf* cluster [28], except for *rpfD*, *rpfH*, ORF1, ORF2, ORF3 and ORF4. Of these six missing genes, only ORF4 was associated with the modulation of EPS production levels in *X. campestris* [18]. The seven genes related to the control of EPS production found in *X. fastidiosa* (Table 1) are homologous to positive regulators. *rpfA* encodes the major aconitase of *X. campestris*, which regulates the production of xanthan gum, possibly by responding to changes in intracellular iron levels [31]. *X. campestris* mutants in *rpfE*, although having no detectable effects on the interaction with plants, show a reduced level of xanthan [18]. RpfC and RpfG are members of a two-component sensory transduction system usually involved in regulation of gene expression in response to environmental stimuli; mutations in both genes reduce xanthan levels [28]. Finally, *rpfB* and *rpfF* exert regulation via a low molecular mass diffusible substance [32].

Based on the following points, we propose that *X. fastidiosa* produces a novel EPS. First, the three genes absent in the *X. fastidiosa gum* operon are all related to the same process. If the *X. fastidiosa* genes were simply missing because of the lack of selective pressure, as would be the case if the operon was not functional, one would expect the absent genes to be contiguous (a single deletion event), or to be the first in the pathway (to save precursor for other pathways). Second, *X. fastidiosa* has all the genes needed to produce the intermediate compounds required for the synthesis of this novel EPS, including *glk*, which has not yet been described in *X. campestris*. Third, all of the proteins are very similar to those described in *X. campestris*. Except for GumF, which showed only 41.9% identity, all pairs of homologues have identities ranging from 59.9% to 84.8% (Table 1). The relatively low identity between the GumF proteins of *X. fastidiosa* and *X. campestris* may reflect a conformational difference between the EPSs of these bacteria. Alternatively, the *X. fastidiosa* GumF may have a different activity, and the fastidian gum could be acetylated to a different degree, or not acetylated at all. A non-acetylated polytetramer EPS has already been described in a mutated *X. campestris* and its viscosity is almost double that of the wild xanthan gum [33].

The fastidian gum may be linked directly to the pathogenicity of this bacterium. The EPS could be involved in the formation of a biofilm required for the attachment and survival of the bacteria in the two hydrodynamically turbulent environments it is found: the xylem vessels and the sucking pumps of insect vectors. A lack of the EPS would therefore prevent the plant symptoms caused by vessel occlusion (and/or embolism) and also affect the spread of the disease. Most of the genes in the *rpf* cluster of *X. campestris* which are absent in *X. fastidiosa* are unrelated to EPS production. The ORF4 product, although important for xanthan production, did not affect the pathogenicity of *X. campestris* and has homologues only among animal pathogens [18].

References

[1] Stickler, D. (1999) Biofilms. Curr. Opin. Microbiol. 2, 270–275.
[2] Mayer, C., Moritz, R., Kirschner, C., Borchard, W., Maibaum, R., Wingender, J. and Flemming, H.C. (1999) The role of intermolecular interactions: studies on model systems for bacterial biofilms. Int. J. Biol. Macromol. 26, 3–16.
[3] Costerton, J.W., Stewart, P.S. and Greenberg, E.P. (1999) Bacterial biofilms: a common cause of persistent infections. Science 284, 1318–1322.
[4] Costerton, J.W., Lewandowski, Z., Caldwell, D.E., Korber, D.R. and Lappin-Scott, H.M. (1995) Microbial biofilms. Annu. Rev. Microbiol. 49, 711–745.
[5] Kang, Y., Saile, E., Schell, M.A. and Denny, T.P. (1999) Quantitative immunofluorescence of regulated eps gene expression in single cells of *Ralstonia solanacearum*. Appl. Environ. Microbiol. 65, 2356–2362.
[6] Katzen, F., Ferreiro, D.U., Oddo, C.G., Ielmini, M.V., Becker, A., Phler, A. and Ielpi, L. (1998) *Xanthomonas campestris* pv. campestris *gum* mutants: effects on xanthan biosynthesis and plant virulence. J. Bacteriol. 180, 1607–1617.
[7] Pooler, M.R. and Hartung, J.S. (1995) Genetic relationships among strains of *Xylella fastidiosa* from RAPD-PCR data. Curr. Microbiol. 31, 134–137.
[8] Simpson, A.J.G., Reinach, F.C., Arruda, P., Abreu, F.A., Acencio, M., Alvarenga, R., Alves, L.M.C., Araya, J.E., Baia, G.S., Baptista, C.S., Barros, M.H., Bonaccorsi, E.D., Bordin, S., Bove, J.M., Briones, M.R.S., Bueno, M.R.P., Camargo, A.A., Camargo, L.E.A., Carraro, D.M., Carrer, H., Colauto, N.B., Colombo, C., Costa, F.F., Costa, M.C.R., Costa-Neto, C.M., Coutinho, L.L., Cris-

tofani, M., Dias-Neto, E., Docena, C., El Dorry, H., Facincani, A.P., Ferreira, A.J.S., Ferreira, V.C.A., Ferro, J.A., Fraga, J.S., Franca, S.C., Franco, M.C., Frohme, M., Furlan, L.R., Garnier, M., Goldman, G.H., Goldman, M.H.S., Gomes, S.L., Gruber, A., Ho, P.L., Hoheisel, J.D., Junqueira, M.L., Kemper, E.L., Kitajima, J.P., Krieger, J.E., Kuramae, E.E., Laigret, F., Lambais, M.R., Leite, L.C.C., Lemos, E.G.M., Lemos, M.V.F., Lopes, S.A., Lopes, C.R., Machado, J.A., Machado, M.A., Madeira, A.M.B.N., Madeira, H.M.F., Marino, C.L., Marques, M.V., Martins, E.A.L., Martins, E.M.F., Matsukuma, A.Y., Menck, C.F.M., Miracca, E.C., Miyaki, C.Y., Monteiro-Vitorello, C.B., Moon, D.H., Nagai, M.A., Nascimento, A.L.T.O., Netto, L.E.S., Nhani, J., Nobrega, F.G., Nunes, L.R., Oliveira, M.A., de Oliveira, M.C., de Oliveira, R.C., Palmieri, D.A., Paris, A., Peixoto, B.R., Pereira, G.A.G., Pereira, J., Pesquero, J.B., Quaggio, R.B., Roberto, P.G., Rodrigues, V., Rosa, A.J.D., de Rosa, V.E., de Sa, R.G., Santelli, R.V., Sawasaki, H.E., da Silva, A.C.R., da Silva, F.R., da Silva, A.M., Silva, J., da Silveira, J.F., Silvestri, M.L.Z., Siqueira, W.J., de Souza, A.A., de Souza, A.P., Terenzi, M.F., Truffi, D., Tsai, S.M., Tsuhako, M.H., Vallada, H., Van Sluys, M.A., Verjovski-Almeida, S., Vettore, A.L., Zago, M.A., Zatz, M., Meidanis, J. and Setubal, J.C. (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406, 151–157.

[9] Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26, 544–548.

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

[11] Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

[12] Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12, 387–395.

[13] Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) Establishing homologies in protein sequences. Methods Enzymol. 91, 524–545.

[14] Saier Jr., M.H. (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol. Rev. 58, 71–93.

[15] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.

[16] Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

[17] Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. 24, 34–36.

[18] Dow, J.M., Feng, J.X., Barber, C.E., Tang, J.L. and Daniels, M.J. (2000) Novel genes involved in the regulation of pathogenicity factor production within the *rpf* gene cluster of *Xanthomonas campestris*. Microbiology 146, 885–891.

[19] Stevenson, G., Andrianopoulos, K., Hobbs, M. and Reeves, P.R. (1996) Organization of the *Escherichia coli* K-12 gene cluster responsible for production of the extracellular polysaccharide colanic acid. J. Bacteriol. 178, 4885–4893.

[20] Koplin, R., Arnold, W., Hotte, B., Simon, R., Wang, G. and Puhler, A. (1992) Genetics of xanthan production in *Xanthomonas campestris*: the *xanA* and *xanB* genes are involved in UDP-glucose and GDP-mannose biosynthesis. J. Bacteriol. 174, 191–199.

[21] Chou, F.L., Chou, H.C., Lin, Y.S., Yang, B.Y., Lin, N.T., Weng, S.F. and Tseng, Y.H. (1997) The *Xanthomonas campestris gumD* gene required for synthesis of xanthan gum is involved in normal pigmentation and virulence in causing black rot. Biochem. Biophys. Res. Commun. 233, 265–269.

[22] Jansson, P.E., Kenne, L. and Lindberg, B. (1975) Structure of extracellular polysaccharide from *Xanthomonas campestris*. Carbohydr. Res. 45, 275–282.

[23] Stankowski, J.D., Mueller, B.E. and Zeller, S.G. (1993) Location of a second O-acetyl group in xanthan gum by the reductive-cleavage method. Carbohydr. Res. 241, 321–326.

[24] Ielpi, L., Couso, R.O. and Dankert, M.A. (1993) Sequential assembly and polymerization of the polyprenol-linked pentasaccharide repeating unit of the xanthan polysaccharide in *Xanthomonas campestris*. J. Bacteriol. 175, 2490–2500.

[25] Harding, N.E., Raffo, S., Raimond, A., Cleary, J.M. and Ielpi, L. (1993) Identification, genetic and biochemical analysis of genes involved in synthesis of sugar nucleotide precursors of xanthan gum. J. Gen. Microbiol. 139, 447–457.

[26] Katzen, F., Becker, A., Zorreguieta, A., Pühler, A. and Ielpi, L. (1996) Promoter analysis of the *Xanthomonas campestris* pv. campestris *gum* operon directing biosynthesis of the xanthan polysaccharide. J. Bacteriol. 178, 4313–4318.

[27] Pollock, T.J., Thorne, L., Yamazaki, M., Mikolajczak, M.J. and Armentrout, R.W. (1994) Mechanism of bacitracin resistance in Gram-negative bacteria that synthesize exopolysaccharides. J. Bacteriol. 176, 6229–6237.

[28] Tang, J.L., Liu, Y.N., Barber, C.E., Dow, J.M., Wootton, J.C. and Daniels, M.J. (1991) Genetic and molecular analysis of a cluster of *rpf* genes involved in positive regulation of synthesis of extracellular enzymes and polysaccharide in *Xanthomonas campestris* pathovar campestris. Mol. Gen. Genet. 226, 409–417.

[29] Paulsen, I.T., Beness, A.M. and Saier Jr., M.H. (1997) Computer-based analyses of the protein constituents of transport systems catalysing export of complex carbohydrates in bacteria. Microbiology 143, 2685–2699.

[30] Ielpi, L., Couso, R.O. and Dankert, M.A. (1983) Xanthan gum biosynthesis: acetylation occurs at the prenyl-phospho-sugar stage. Biochem. Int. 6, 323–333.

[31] Wilson, T.J., Bertrand, N., Tang, J.L., Feng, J.X., Pan, M.Q., Barber, C.E., Dow, J.M. and Daniels, M.J. (1998) The *rpfA* gene of *Xanthomonas campestris* pathovar campestris, which is involved in the regulation of pathogenicity factor production, encodes an aconitase. Mol. Microbiol. 28, 961–970.

[32] Barber, C.E., Tang, J.L., Feng, J.X., Pan, M.Q., Wilson, T.J., Slater, H., Dow, J.M., Williams, P. and Daniels, M.J. (1997) A novel regulatory system required for pathogenicity of *Xanthomonas campestris* is mediated by a small diffusible signal molecule. Mol. Microbiol. 24, 555–566.

[33] Hassler, R.A. and Doherty, D.H. (1990) Genetic engineering of polysaccharide structure: production of variants of xanthan gum in *Xanthomonas campestris*. Biotechnol. Prog. 6, 182–187.

[34] Tung, S.Y. and Kuo, T.T. (1999) Requirement for phosphoglucose isomerase of *Xanthomonas campestris* in pathogenesis of citrus canker. Appl. Environ. Microbiol. 65, 5564–5570.

[35] Meyer, D., Schneider-Fresenius, C., Horlacher, R., Peist, R. and Boos, W. (1997) Molecular characterization of glucokinase from *Escherichia coli* K-12. J. Bacteriol. 179, 1298–1306.

[36] Barnell, W.O., Yi, K.C. and Conway, T. (1990) Sequence and genetic organization of a *Zymomonas mobilis* gene cluster that encodes several enzymes of glucose metabolism. J. Bacteriol. 172, 7227–7240.

[37] Lin, C.S., Lin, N.T., Yang, B.Y., Weng, S.F. and Tseng, Y.H. (1995) Nucleotide sequence and expression of UDP-glucose dehydrogenase gene required for the synthesis of xanthan gum in *Xanthomonas campestris*. Biochem. Biophys. Res. Commun. 207, 223–230.

[38] Wei, C.L., Lin, N.T., Weng, S.F. and Tseng, Y.H. (1996) The gene encoding UDP-glucose pyrophosphorylase is required for the synthesis of xanthan gum in *Xanthomonas campestris*. Biochem. Biophys. Res. Commun. 226, 607–612.

Apêndice A

Organismos de vida livre com o genoma seqüenciado

(até dezembro de 2000)

organismo	ano	super-reino	tamanho (Kb)	orfs	referência
Haemophilus influenzae KW20	1995	bactérias	1830	1850	Science 269,496-512
Mycoplasma genitalium G-37	1995	bactérias	580	468	Science 270,397-403
Methanococcus jannaschii DSM 2661	1996	arqueobactérias	1664	1750	Science 273,1058-1073
Mycoplasma pneumoniae M129	1996	bactérias	816	677	NAR 24,4420-4449
Synechocystis sp. PCC6803	1996	bactérias	3573	3168	DNA Res 3,109-136
Archaeoglobus fulgidus DSM4304	1997	arqueobactérias	2178	2493	Nature 390,364-370
Bacillus subtilis 168	1997	bactérias	4214	4099	Nature 390,249-256
Borrelia burgdorferi B31	1997	bactérias	1230	1256	Nature 390,580-586
Escherichia coli K12- MG1655	1997	bactérias	4639	4289	Science 277,1453-1474
Helicobacter pylori 26695	1997	bactérias	1667	1590	Nature 388,539-547
Methanobacterium thermoautotrophicum delta H	1997	arqueobactérias	1751	1918	J.Bacteriology 179,7135-7155
Saccharomyces cerevisiae S288C	1997	eucariotos	12069	6294	Nature 387,5-105
Aquifex aeolicus VF5	1998	bactérias	1551	1544	Nature 392,353-358
Caenorhabditis elegans	1998	eucariotos	97000	19099	Science 282,2012-2018
Chlamydia trachomatis serovar D	1998	bactérias	1042	896	Science 282,754-759
Mycobacterium tuberculosis H37Rv	1998	bactérias	4411	4402	Nature 393,537-544
Pyrococcus horikoshii(shinkaj) OT3	1998	arqueobactérias	1738	1979	DNA Research 5,55-76
Rickettsia prowazekii Madrid E	1998	bactérias	1111	834	Nature 396,133-140
Treponema pallidum Nichols	1998	bactérias	1138	1041	Science 281,375-388
Aeropyrum pernix K1	1999	arqueobactérias	1669	2620	DNA Research 6,83-101
Chlamydia pneumoniae CWL029	1999	bactérias	1230	1052	Nat Genet 21,385-389
Deinococcus radiodurans R1	1999	bactérias	3284	3187	Science 286,1571-1577
Helicobacter pylori J99	1999	bactérias	1643	1495	Nature 397,176-180
Pyrococcus abyssi GE5	1999	arqueobactérias	1765	1768	**não publicado**
Thermotoga maritima MSB8	1999	bactérias	1860	1877	Nature 399,323-329
Arabidopsis thaliana	2000	eucariotos	115428	25598	Nature 408, 796-815
Bacillus halodurans C-125	2000	bactérias	4202	4066	Extremophiles 4, 99-108
Buchnera sp. APS	2000	bactérias	640	564	Nature 407, 81-86
Campylobacter jejuni NCTC 11168	2000	bactérias	1641	1654	Nature 403,665-668
Chlamydia pneumoniae AR39	2000	bactérias	1229	1052	NAR 28,1397-1406
Chlamydia pneumoniae J138	2000	bactérias	1228	1070	NAR 28,2311-2314
Chlamydia muridarum	2000	bactérias	1069	924	NAR 28,1397-1406
Drosophila melanogaster	2000	eucariotos	137000	14100	Science 287,2185-95
Halobacterium sp. NRC-1	2000	arqueobactérias	2014	2058	PNAS 97, 12176-12181
Mesorhizobium loti MAFF303099	2000	bactérias	7596	6752	DNA Research 7, 331-338
Neisseria meningitidis MC58 (serogroup B)	2000	bactérias	2272	2158	Science 287,1809-1815
Neisseria meningitidis Z2491 (serogroup A)	2000	bactérias	2184	2121	Nature 404,502-506
Pseudomonas aeruginosa PAO1	2000	bactérias	6264	5570	Nature 406,959-964
Thermoplasma acidophilum	2000	arqueobactérias	1564	1478	Nature 407, 508-513
Thermoplasma volcanium GSS1	2000	arqueobactérias	1584	1524	PNAS 97, 14257-14262
Ureaplasma urealyticum serovar 3	2000	bactérias	751	650	Nature 407, 757-762
Vibrio cholerae O1, Biotype El Tor, strain N16961	2000	bactérias	4033	3885	Nature 406,477-483
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	2000	bactérias	2679	2904	Nature 406,151-157

Apêndice B

Protocolos empregados no laboratório de genoma do CBMEG durante o projeto genoma de *Xylella fastidiosa*.

Lise Alcalina - Micropreps

1- Fill the wells of a microplate with 1ml of Circle Grow medium containing ampicilin at 100µg/ml.

2- Pick colonies with a toothpick into the medium. Seal the plate with an adhesive and pierce each well with a needle to allow aeration during growth.

3- Place the box into a shaker at 37°C at 300rpm for 22 hours.

4- Spin at 4000rpm for 6 minutes to pellet the cells.

5- Remove the supernatant and leave dry on a tissue for 1 minute.

6- Add to each well 240µl of GTE. Seal the plate with an adhesive and vortex the cells to resuspend.

7- Spin the plate at 4000rpm (Jouan Br4I centrifuge) for 6 minutes to pellet the cells. Remove the supernatant.

8- Add to each well 80µl of GTE. Seal the plate with an adhesive and vortex the cells to resuspend.

9- Add 5µl of RNase (10mg/ml) to each well of a 96-wells microplate.

10-Transfer 60µl of the cells to the 96-wells microplate. Add to each well 60µl of NaOH/SDS solution. Seal the plate with an adhesive and mix by inversion 10 times.

11-Leave on bench for 10 minutes. Spin for few seconds.

12-Add to each well 60µl of 3M KOAc (stored at 4°C). Seal the plate with an adhesive and mix by inversion 10 times.

13-Leave on bench for 10 minutes. Spin for few seconds.

14-Remove the adhesive and place the plate in a oven at 90°C for EXACTLY 30 minutes.

15-Cool the microplate by placing it on ice for 10 minutes. Spin for 4 minutes at 4000rpm at 20°C

16-Sellotape a Millipore filter plate (MAGV N22) to the top of a new 250µl 96-wells microplate ensuring the filters and receiving wells line up.

17-Transfer the full volume of the 96-wells microplate to the filter plate and spin for 4 minutes at 4000rpm at 20°C.

18-Remove and discard the filter plate and add 110µl of isopropanol to the filtrate.

19-Seal the plate with an adhesive and mix by inversion 10-20 times.

20-Spin for 45 minutes at 4000rpm at 20°C

21-Remove the supernatant and add 200µl of ice cold ethanol 70%.

22-Spin for 5 minutes at 4000rpm at 20°C. Remove the supernatant.

23-Invert the plate over absorbent paper and spin for 3 minutes at 900rpm at 20°C. (Optional)

24-Leave the plates to dry for 60 minutes at room temperature.

25-Dissolve the DNA in 40µl of water (overnight).

Solutions

Circle Grow Medium

Circle Grow	40g
H2O to	1000ml
Autoclave	
Add 1µl of Amp. stock per ml of medium	

Ampicillin Stock, 100mg/ml

Ampicillin	1g
H2O to	10ml
Store at -20°C	

Glucose 20%

Glucose	20g
H2O to	100ml
Autoclave	

EDTA, 0.5M pH8.0

EDTA . 2 H2O	186.1g
H2O	700ml
Adjust pH to 8.0 with 10M NaOH (~50ml)	
H2O to	1000ml

Tris-HCl, 1M pH7.4 or 8.5

Tris base	121g
H2O	800ml
Adjust pH to 7.4 or 8.5 with Conc. HCl	
H2O to	1000ml

Tris-HCl, 1mM pH8.5

1M Tris-HCl pH8.5	100µl
H2O to	100ml

GTE

20% Glucose	23ml
0.5M EDTA pH8.0	10ml
1M Tris-HCl pH7.4	13ml
H2O to	500ml
Autoclave	

Sodium Hydroxide, 4M

NaOH	16g
H2O to	100ml

SDS, 10%

SDS	10g
H2O to	100ml

NaOH/SDS solution

0.2M NaOH, 1% SDS

4M NaOH	25ml
10% SDS	50ml
H2O to	500ml

Potassium Acetate, 3M

KOAc	147.2g
Glacial Acetic Acid	7.5ml
H2O to	500ml
..Autoclave	
Stock at 4°C	

RNase A Stock, 10mg/ml

RNase A	100mg
H2O to	1ml
boil for 15 minutes	
Cool at Room Temperature	
Stock at -20°C	

Ethanol 70%

Ethanol Abs.	350ml
H2O to	150ml

REAGENTS

Reagent	Supplier	Cat #
Circle Grow Medium	BIO 101	3000-142
Ampicilim Sodium Salt	SIGMA	A-9518
EDTA . 2 H2O	SIGMA	E-6635
Ethanol Absolute	MERCK	21604
Glacial Acetic Acid	MERCK	21606
Glucose	MERCK	8347
Isopropanol	MERCK	21504
Potassium Acetate	SIGMA	P-3542.
RNase A	PHARMACIA	27-0323
SDS	SIGMA	L-4390
Sodium Hydroxide	MERK	6468.
Tris base	SIGMA	T-8524

***Bibliotecas* shotgun**

Sonication:

1. Put 10µg of DNA in a final volume of 20µl of water. Divide in two microtubes of 200µl (10µl each). Put on ice.
2. Using a cuphorn, sonicate for three different times (ex.: 15" 20" and 25").
3. Load 5µl in an agarose gel to check the DNA fragmentation.
4. Mix the content of the tubes.
5. If necessary, you can precipitate the DNA by the addition of 0.1 volumes of 3M NaOAc pH5.2 and 2 volumes of Ethanol. Incubate for 10 minutes at -80°C. Spin for 20 minutos at 14.000rpm at 4°C. Wash with Ethanol 70%. Dry in SpeedVac.
6. Resuspend in 30µl of water.

Blunt-ends:

1. Reaction with *T4 DNA Polymerase* and *Klenow:*

   30µl DNA

   2µl *T4 DNA Polymerase* (5U/µl)

   10µl *T4 DNA Pol.* Buffer (5X)

   2,5µl dNTPs Mix (1mM)

   7.5µl water

   - incubate for 30 minutes at 37°C
   - add 2µl of *Klenow* (5U/µl)
   - incubate for 15 minutes, at room temperature.
2. Run a 0.8% agarose Low-Melting Point gel (40-50V, TAE 1X, ~3 hours).
3. Remove the gel band containing de DNA fragments with size between 1 and 2Kb. Extract the DNA from the gel.
4. After extraction, run an 1% agarose gel to confirm the size of the extracted DNA

Ligation:

(Kit: Ready-to-go pUC18 SmaI/BAP + Ligase – Pharmacia cat# 27-5266-01)

1. Dissolve 2μl and 5μl in 20μl of water. Usually, this amount of DNA is enough to obtain a good number of clones.
2. Add 20μl of the DNA solution to the tube Ready-to-go pUC18 SmaI/BAP + Ligase.
3. Incubate at room temperature for 3-5 minutes. Then mix by gentlly pipetting up and down several times.
4. Centrifuge briefly to collect the contents at the botton of the tube. The centrifugation will also remove any bubbles that were created.
5. Incubate at 16°C for 45-90 minutes.
6. Store at 4°C

Transformation:

1. Thaw competent cells on ice
2. Mix ligation (about 5μl) with 20μl of Transformation Buffer and place on ice.
3. Add 100μl of thawed cells. Let sit on ice for 15-30 minutes
4. Move to room temperature for 10 minutes.
5. Add 1ml of LB medium and incubate at 37°C for 50 minutes.
6. Spread 100μl of the cells on a X-Gal LB_{amp} plate. Incubate overnight at 37°C.

Solutions

EDTA, 0.5M pH8.0

EDTA . 2 H2O	186.1g
H2O	700ml
Adjust pH to 8.0 with 10M NaOH (~50ml)	
H2O to	1000ml

Tris-HCl, 1M pH8.0

Tris base	121g
H2O	800ml
Adjust pH to 8.0 with Conc. HCl (~42ml)	
H2O to	1000ml

TE

10mM Tris-HCl, 1mM EDTA, pH8.0

1M Tris-HCl pH8.0	5ml
0.5M EDTA pH8.0	1ml
H2O to	500ml

SDS, 10%

SDS	10g
H2O to	100ml

Sodium Acetate, 3M pH5.2

NaOAc	408g
H2O	800ml
Adjust pH to 5.2 with Acetic Acid	
H2O to	1000ml

Ethanol 70%

Ethanol Abs.	350ml
H2O to	150ml

dNTPs Mix, 1mM each one

100mM dATP	10µl
100mM dCTP	10µl
100mM dTTP	10µl
100mM dGTP	10µl
H2O to	1ml

TAE 50X

Tris base	242g
0.5M EDTA pH8.0	100ml
Glacial Acetic Acid	57.1ml
H2O to	1000ml

X-Gal Stock, 20mg/ml

X-Gal	0.2g
Dimetylformamide	10ml
Store at -20°C	

10X KCM

KCl	0.745g
CaCl2	0.441g
MgCl2 1M	5ml
H2O to	10ml

10% PEG (6000 or 4000)

PEG	1g
H2O to	10ml

Transformation Buffer

10X KCM	1ml
10% PEG	1.5ml
H2O	7.5ml
Store at –20°C	

Purificação de DNA de cosmídeos

1-Inoculate 500ml of LB medium containing kanamicin at 30µg/ml. Incubate at 37°C at 300rpm for 12-16 hours.

2- Spin for 5 minutes at 6.000rpm at 4°C.

NucleoBond Purification

3- Resuspend the bacterial cell pellet in 12ml of **Sol. S1** (50mM Tris-HCl pH8.0; 10mM EDTA; 100µg/ml RNase A).

4- Transfer the mixture to a SS-34 tube. Add 12ml of **Sol. S2** (200mM NaOH; 1% SDS). Incubate at room temperature for 5 minutes.

5- Add 12ml of **Sol. S3** (2.8M KOAc, pH5.1). Incubate on ice for 5 minutes.

6- Spin for 45 minutes at 13.000rpm at 4°C.

7- Filter the supernatant with filter paper and load it on a Nucleobond AX cartridge, preequilibrated with 5ml of **Sol. N2** (100mM Tris; 15% Ethanol; 900mM KCl; pH 6.3 adjusted with H_3PO_4). Collect the flow-through and load it a second time on the cartridge.

8- Wash the cartridge with 12ml of **Sol. N3** (100mM Tris; 15% Ethanol; 1150mM KCl; pH 6.3 adjusted with H_3PO_4). Wash again with more 12ml of **Sol.N3.**

9- Elute the DNA with 9ml of **Sol. N5** (100mM Tris; 15% Ethanol; 1000mM KCl; pH 8.5 adjusted with H_3PO_4).

10- Transfer the DNA solution to eppendorf tubes (800µl). Add 0.7 volumes of isopropanol (600µl). Mix by inversion.

11- Centrifuge at 14.000rpm at 4°C for 20 min.

12- Wash with Ethanol 70%.

13- Centrifuge at 10.000rpm at R.T. for 7 minutes.

14- Dry in SpeedVac.

15- Dissolve the DNA pellets in 20µl of water. Wash the tubes with more 50µl of water. Final Volume: 250µl.

16- Digest 5µl with the appropriated restriction enzyme to check the DNA quality.

17- Cesium Chloride purification: 4.5g CsCl

4.3ml TE

100µl BrEt 10mg/ml

200µl DNA

Refraction = 1.3860 (1.55g CsCl/ml)

Ultracentrifuge at 58.000rpm at 20°C for 12-16 hours

18- Remove the DNA band with a syringe.

19- Remove the BrEt by extraction with 1 volume of isoamyl alcohol/H_2O*. Repeat more 4 times.

20- Remove the isoamyl alcohol by extraction with ether/H_2O*. Repeat more 2 times.

**saturated with water*

21- Transfer 250µl to an eppendorf tube. Add 250µl of water.

22- Add 2 volumes of ethanol (1ml). Incubate for 10 minutes on ice

23- Centrifuge at 14.000rpm at 4°C for 20 minutes.

24- Wash with Ethanol 70%. Dry in SpeedVac.

25- Dissolve the DNA pellets in a total volume of 200-300µl of water.

26- Digest 2µl with the appropriated restriction enzyme to check the DNA quality.

SOLUTIONS

LB Medium

Yeast Extract	10g
Tryptone	10g
NaCl	5g
H2O to	1000ml
Autoclave	
Add 1µl of Kanam. stock per ml of medium	

Kanamicin Stock, 30mg/ml

Kanamicin Monosulfate	0.3g
H2O to	10ml
Store at -20°C	

Ethanol 70%

Ethanol Abs.	350ml
H2O to	150ml

Ethidium Bromide, 10mg/ml

Ethidium Bromide	0.2g
H2O to	20ml
Store in dark	

EDTA, 0.5M pH8.0

EDTA . 2 H2O	186.1g
H2O	700ml
Adjust pH to 8.0 with 10M NaOH (~50ml)	
H2O to	1000ml

Tris-HCl, 1M pH8.0

Tris base	121g
H2O	800ml
Adjust pH to 8.0 with Conc. HCl (~42ml)	
H2O to	1000ml

TE

10mM Tris-HCl, 1mM EDTA, pH8.0

1M Tris-HCl pH8.0	5ml
0.5M EDTA pH8.0	1ml
H2O to	500ml

Isoamyl Alcohol saturated with H2O

Isoamyl Alcohol	100ml
H2O	50ml

Diethyl Ether saturated with H2O

Diethyl Ether	100ml
H2O	50ml

REAGENTS

Reagent	Supplier	Cat #
Cesium Chloride	BOEHRINGER	757-306
Diethyl Ether	MERCK	1.00921.
Ethanol Absolute	VEL	1115
Ethidium Bromide	BIORAD	161-0430
EDTA . 2 H2O	ACROS	14785-0010
Isoamyl Alcohol	MERCK	1.00979.
Isopropanol	BAKER	8067
Kanamicin Monosulfate	SIGMA	K-4378
Nucleobond AX	MACHEREY-NAGEL	709-241
Sodium Chloride	MERCK	1.06.404.
Tris base	SIGMA	T-8524
Tryptone	DFICO	0123-17-3
Yeast Extract	DIFCO	0127-17-9

Extração de DNA de gel low melting

1- Cut the appropriate DNA band from the gel. Transfer it into an eppendof tube.

2- Spin for few seconds to pellet the agarose pieces.

3- Heat for 5 minutes at 65°C with periodically vortexing.

4- Estimate the volume and add 0.1 volumes of 4M NaCl.

5- Heat for 3 minutes at 65°C with periodically vortexing.

6- Add 0.1 volumes of phenol pH8.0. Mix well.

7- Spin for 10 minutes at 14.000 rpm at room temperature. Take the upper phase.

8- Repeat the steps 6 and 7 for more 2 times.

9- Add 2 volumes of diethyl ether/H_2O. Mix well.

10- Spin for 1 minute at 14.000 rpm at room temperature and remove the upper phase.

11- Repeat the steps 9 and 10 for more 2 times.

12- Add 0.1 volumes of 3M NaAc pH5.2 and 2 volumes of Ethanol Abs.

13- Incubate for 10 minutes at -80°C.

14- Spin for 15 minutes at 14.000 rpm at 4°C. Remove de supernatant.

15- Wash 2 times with Ethanol 70%. Dry the pellet in SpeedVac.

16- Dissolve the DNA pellet in 50µl of water.

SOLUTIONS

Sodium Chloride, 4M

NaCl	23.4g
H2O to	100ml

Diethyl Ether saturated with H2O

Diethyl Ether	100ml
H2O	50ml

Sodium Acetate, 3M pH5.2

NaOAc . 3 H2O	408g
H2O	800ml
Adjust pH to 5.2 with Acetic Acid	
H2O to	1000ml

Ethanol 70%

Ethanol Abs.	350ml
H2O to	150ml

REAGENTS

Reagent	Supplier	Cat #
Diethyl Ether	MERCK	1.00921.
Ethanol Absolute	VEL	1115
Glacial Acetic Acid	VEL	1005
Low Melting Point Agarose	FMC	50-102
Phenol	ALDRICH	32-811-1
Sodium Acetate . 3 H2O	MERCK	1.06267.
Sodium Chloride	MERCK	1.06.404.

Midiprep e seqüenciamento de extremidades de fago lambda.

Preparing the Bacteria

1. Inoculate a single DL538 colony in 5ml of NZY medium supplied with 0,2% of Maltose and 10mM of $MgSO_4$. Incubate overnight at 37°C at 300 RPM.
2. Inoculate 50ml of fresh NZY medium supplied with 0,2% of Maltose and 10mM of $MgSO_4$ with 500µl of the overnight culture. Incubate at 37°C at 300 RPM until OD_{600} = 0,5.
3. Spin for 10 minutes at 3.000 RPM at 4°C.
4. Ressuspend the cells in 50ml of $MgSO_4$ 10mM. Stock at 4°C.

Preparing liquid phage lysate

Microplate preparation

1. Infect 500µl of DL538 with a desirable dilution of the library stock (to produce a ~100 plaques per 150 mm dish). Incubate for 20 minutes at 37°C.
2. Add 8ml of NZY-top agar to the infection and plate it in a NZY plate (150mm). Incubate at 37°C for 24 hours.
3. Store the plate at 4°C

Macroplates preparation

1. Overlay an LBM plate containing 10mM $MgSO_4$ with 0.1 ml of a fresh overnight cell culture DL538 in 3 ml LBM top agar. Let solidify at room temperature, then store at 4°C for an hour to firm the surface.
2. Transfer phage from a single well-isolated plaque to its grid position using a sterile toothpick or wood applicator stick. Inoculate an area about the size of the wide end of a pasteur pipet by lightly touching the agar surface several times.
3. Incubate at 37°C. The grid areas should clear in 6-7 hours creating macroplaques.
4. Chill plates at least 30 minutes.
5. Excise macroplaques using a 1ml pipette tip cut at middle.
6. Put each plug into a 1,5ml microtube containing 250µl of SM. Phage extraction of the agar plug can be done at 4°C overnight or by gently shaking for 1-2 hours at room temperature. At this point phage plugs can be stored at 4°C for several weeks provided they are tightly capped.
7. Put each plug into a 15 ml disposable tube containing 250µl of SM. Incubate overnight at 4°C. At this point, the phage can be stored at 4°C for several weeks provided tubes are tightly capped.

Purification of phage particles and DNA extraction

DEAE-cellulose method

1. Remove an aliquot of 50-100µl, add 2 drops of chloroform and store at 4°C.
2. Add 4ml of DE52 suspension and gently mix by inverting the tube 20-30 times.
3. Centrifuge at 10,000 rpm for 10 minutes and transfer supernatant to a new tube. Spin again for 10 minutes to remove any remaining DE52.
4. Add NaCl to a final concentration of 0.55 M and 0.6 volume of cold IsPrOH. Mix and chill for 50 minutes at –40°C.
5. Centrifuge at 10,000 rpm for 10 minutes, discard supernatant, and wash the pellet with 3ml of 70% cold ethanol. Centrifuge at 10,000 rpm for 5 minutes.
6. Re-collect the washed pellet by centrifugation. Dry for 30 minutes at room temperature.
7. Dissolve the pellet in 200µl TE (GENTLELY).
8. Extract the DNA twice with an equal volume of phenol pH 8.0.
9. Extract once with an equal volume of a 24:1 mixture of chloroform:isoamyl alcohol.
10. Precipitate the DNA by adding 15µl 5 M NaCl and 2 volumes of cold 95% ethanol. Chill at -70°C for 30 minutes.
11. Spin for 15 minutes at 13,000 RPM at 4°C, wash once with 1 ml 70% ethanol and dry the pellet.
12. Dissolve the pellet in 30µl of H_2O.
13. Check the quality of the DNA by electrophoresis with 5µl of each DNA.

Preparation of DE52 suspension

1. Place 100g of DE52 in a backer. Slowly add 200ml of 0.05 N HCl.
2. Mix the solution by gentle stirring for 10 minutes.
3. Allow the resin to settle, decant the supernatant.
4. When the solution is separated in two phases, remove the upper phase
5. Repeat these last steps until the pH goes below 4.5.
6. Then, with constant gentle stirring, add NaOH until pH goes to 6.8.
7. Wash the resin by ressuspend it in 2 volumes of LB medium (pH 6,8).
8. Allow the resin to settle and decant the excess medium.
9. Repeat the steps 7 and 8 for more 2 times.
10. After the last wash, make a final slurry of approximately 75% resin and 25% LB medium.

11. Add 0.01% sodium azide
12. Store at 4°C

End-Sequencing

Template	Primer	BigDye	H_2O	Final Volume
0.5 - 1.0 µg (10 µl)	7.5 pmoles (1.5 µl)	8 µl	0.5 µl	20 µl

PCR conditions

1 x	95°C 5 minutes
70x	95°C 30 sec
	55°C 20 sec
	60°C 4 minutes

Basic primer on annotation

This primer intends to provide a very basic guideline in tools and strategies for annotating a cosmid. Different approaches could be attempted. However, the standard in the format of the final file (Sequin) and, most important, the fields in which data is entered should be the ones indicated here. Suggestions are very welcome.

Working with sequin

Each cosmid is going to be annotated in an individual sequin file. The program sequin is available for all platforms used by the ONSA labs. There is extensive documentation for the use of this program (including an context sensitive online help open all the time the program is running) so only the very basic steps are going to be described here.

Initiating the sequin file

You will need a FASTA format file of the consensus sequence of the cosmid insert (this means ONLY the insert sequence, with no Lawrist). Edit your FASTA file in order to have only your cosmid number (7h3, *e.g.*) after the ">" character. The beginning of your file should look like this:

```
>7h3
ATGCGACGTATTGAGCAGACATAGACAGATATAGAGAC
TTAGGCCAGAGACATAGCGAGATAGACAGAGATAGACA
TGGGGTTTGGGGGACACGCGGCATAGACTTTGACACAC...
```

Initiate sequin and press **Start New Submission**.

Fill the **Tentative title for manuscript** with the cosmid number and press the **Next Page** button. The next three pages are not important, fill them as you like. When you press the **Next Page** button in the **Affiliation** page, the **Sequence Format** window pops-up. The default values (*Single sequence / FASTA*) are the right ones.

In the next window (**Organism and Sequences / Organism**) type *Xylella fastidiosa*, let *Genomic* as the **Location of the Sequence** and choose *Bacterial* as the **Genetic Code for**

Translation. In the **Nucleotide** panel choose *Genomic DNA* as the **Molecule**, let the **Topology** *Linear*, tick *Incomplete* for both ends, let the **Fasta definition line...** ticked and click the **Import Nucleotide FASTA** button. Chose your FASTA file in the **file_selection** window and verify the values (1 segment, n nucleotides and the Sequence ID, with the ID=your cosmid number and n=the right size of the insert). Press the **Next Page** and them the **Next Form** button. Answer *OK* to the "**You have not entered proteins. Is that correct?**"

In the new window (the one with your cosmid name) choose **File / Save**. You may now quit the program.

From now on, when working with this cosmid, choose **Read Existing Record** after starting Sequin.

Note: depending on the way you exit the program or save your work, a window may pop-up saying "**Submission is now written. Please e-mail '*your_cosmid_file*' to gb-sub@ncbi.nlm.nih.gov**". **DO NOT** do this. When the annotation is complete, this file should be sent **ONLY** to the Bioinformatics Lab of the ONSA.

Finding ORFs

Although it may miss some genes, the initial searches should be with the ORFs. Several methods exist for locating them in the cosmid sequence. We suggest the use of GLIMMER or GeneMark.HMM. Both lead to similar results. GeneMark can be run directly in the Web (changing the **Species** in the **Running Options** from *E.coli* to *otherprokariotes* before starting the search) while GLIMMER has to be downloaded and installed in a UNIX/LYNUX machine. Don't forget that the sequence used to search for the ORFs has to be identical to the one used in the sequin file.

Marking the ORFs

Open the sequin file and choose **Annotate / Coding Regions and Transcripts / CdRgn**. Choose the **Location** panel and impute the position of the beginning and the end of the first ORF in the **From** and **To** fields. If the ORF is in the reverse orientation (*i.e.*, if the beginning of the ORF is in the 3') change the **Strand** field to *Minus*, otherwise let it *Plus*. Now choose the **Coding Region** panel

and the **Product** sub panel. Click the **Translate Product** button and inspect what you see.

The most important thing to note is that there should be only one asterisk (*) in the data shown (schrool down the window, if necessary, to see the whole sequence). The asterisk represents the stop codon and must appear in the end of the sequence. If there's none in the translated sequence, return to the **Location** panel and increase the **To** field by three positions. It may be that the program used to locate the ORFs does not consider the stop codon as part of them. If there are some stop codons inside the translated sequence it may be that you forget to choose the right orientation in the **Strand** field. To be sure that you locate correctly the ORF, there should be only one stop codon at the end.

Repeat this procedure to all ORFs.

It is obviously a tedious task. There are more practical ways of doing this. Lets suppose that you manually annotated the position of some genes and is sure now of what are the right numbers to put in the **From** and **To** fields. In this case, you could generate a text file to tell sequin where are all the (other) ORFs. This text file should be in the following format:

- the first line contains ">Features" followed by a space and then the cosmid name

  `>Features 7h3` (*for example*)

- the following lines have three columns separated by tab marks. The first column contains the start of the ORF, the second the end and the third the text "CDS"

The file should look like this:

```
>Features                                    7h3
871                     17                  CDS
1186                    1668                CDS
1598                    1765                CDS
3478    1991    CDS
```

Note that the **Strand** field is automatically generated depending on the orientation a CD appears in the text file. Lines 2 and 5 in this example would generate *Minus* entries and the others *Plus*. Moreover, despite the fact you don't see a translation where you did in the previous method, you can still press the translate button, but IF THE ORFs ARE CORRECTLY LOCATED this is not necessary.

Finding similarities

Like in the location of ORFs, the search for similarities in ORFs can be done in a multitude of ways. There are (commercial) programs that, once a ORF is found, permits a immediate BLAST search. The method depicted bellow is intended to be used by those that have only the sequin and a browser. Still, it works very well.

The sequin file, once opened, can be seen in a series of formats (choose in the **Display Format**). If it is in the *GenBank* format, the **File** menu will show a option called "**Export GenBank...**". Export the sequence in this format, with a name different from the one you use for the sequin file. Then, with any text editor, open the resulting file. Every translated ORF can be found in the field "/translation" after the CDS feature. Copy the sequence of the ORF (the text that is between double quotes only), don't worry about the spaces in the beginning of each line.

Go to the BLAST web server, chose *blastp* as the **Program**, let *nr* (non-redundant) as the **Database** and paste the translated ORF in the square field. Don't worry about the FASTA format, it's not necessary, really. Click the **Submit Query** button.

The output will consist of a color graphic, lines describing the hits found and the best alignments. The most important result to consider is the E-value (the last column in the hits brief description and the second paragraph (field "Expect = ") in the alignments) found for the hits.

The value depicted there is in Scientific Notation, *i.e.*, the "e" letter is the same as "x 10^{-n} " where n is the number after the dash. So an e-value of 4e-59 means $4x10^{-59}$. This number represents the probability of this specific hit be produced be chance (*i.e.*, be of no significance). So, the most significative hits are the ones with small e-value. Values bellow 10^{-5} (1e-5) can already be considered significant, but it is usual to see values well bellow that.

Deciding which of the hits is the one that best describes the ORF can be very confusing. If there is an E.coli protein among the best hits, we would suggest this to be chosen, specially if the hits above it describe the same (kind of) protein.

Annotating, finally!

In the sequin file (opened in the sequin) choose de *Graphic* **Display Format** and adjust the **Scale** to a value where you can see well each CD. Double-click in the CD whose sequence you have just sent to blastp. In the **Comment** sub panel of the **Properties** panel, type what you think the ORF is. Pay attention to not mistake the ORF sent to blastp with another one in your sequin file. Just check the numbers in the first line of the CD field in the GenBank file in your

text editor (the file where you took the sequence to send to blastp) with the position of the ORF (CDS) you are clicking in the sequin file.

Another features

For a more complete annotation, you may choose to find another interesting features. One that should be done is search for tRNAs. We suggest you doing this using the tRNAscan-SE. Remember to use the same FASTA file used to generate the sequin file originally.

In the **tRNAscan-SE** page, let **Search Mode** *Default*, choose *Prokaryotic* as **Source**, check *Other* as **Format** and **Browse** for your FASTA file. Click the **Run tRNAscan-SE** button.

To annotate the tRNAs you might find, choose **Annotate / Structural RNAs / tRNA** in the sequin file. Enter the tRNAscan-SE **tRNA Bounds Begin** and **End** values in the **From** and **To** fields of the **Location** panel in sequin. In the **tRNA** panel, select the corresponding **Amino Acid** (the **tRNA Type** in the tRNAscan output) in the pull down menu in the **Amino Acid** sub panel. In the **Codons** sub panel, type the Recognized Codons (the **Anti Codon** column in the tRNAscan output). Click the **Accept** button.

Once you have covered all ORFs and any other features, save the file and send it to us.

Comments, suggestions, complains? Send us a note! ;-)

Este documento pode ser encontrado na antiga página do projeto Xylella fastidiosa em http://www.lbi.ic.unicamp.br/xf-old/Annotation_primer.html

Apêndice D

***Todas as categorias empregadas na anotação de* Xylella fastidiosa.**

Intermediary metabolism (**195**)

A. Degradation (**24**)
    1. Degradation of polysaccharides
    2. Degradation of small molecules
        - Amines
        - Amino acids
        - Carbon compounds
        - Fatty acids

B. Central intermediary metabolism (**46**)
    1. Amino sugars
    2. Entner-Douderoff
    3. Gluconeogenesis
    4. Glyoxylate bypass
    5. Miscellaneous glucose metabolism
    6. Non-oxidative branch, pentose pathway
    7. Nucleotide hydrolysis
    8. Nucleotide interconversions
    9. Phosphorus compounds
    10. Pool, multipurpose conversions
    11. Sugar-nucleotide biosynthesis, conversions
    12. Sulfur metabolism

C. Energy metabolism, carbon (**84**)
    1. Aerobic respiration
    2. Anaerobic respiration
    3. Electron transport
    4. Glycolysis
    5. Oxidative branch, pentose pathway
    6. Pyruvate dehydrogenase
    7. TCA cycle
    8. ATP-proton motive force interconversion

D. Fermentation (**2**)

E. General regulatory functions (**39**)

II. Biosynthesis of small molecules (**218**)

A. Amino acids (**77**)
    1. Glutamate family|nitrogen assimilation
        - Arginine
        - Glutamate
        - Glutamine

- Proline

2. Aspartate family, pyruvate family
    - Alanine
    - Asparagine
    - Aspartate
    - Isoleucine, Valine
    - Leucine
    - Lysine
    - Methionine
    - Threonine
3. Glycine-serine family|sulfur metabolism
    - Cysteine
    - Glycine
    - Serine
4. Aromatic amino acid family
    - Chorismate
    - Phenylalanine
    - Tryptophan
    - Tyrosine
5. Histidine

B. Nucleotides (**41**)
    1. Purine ribonucleotides
    2. Pyrimidine ribonucleotides
    3. 2'-Deoxyribonucleotides
    4. Salvage of nucleosides and nucleotides

C. Sugars and sugar nucleotides (**2**)

D. Cofactors, prosthetic groups, carriers (**74**)
    1. Biotin
    2. Folic acid
    3. Lipoate
    4. Molybdopterin
    5. Pantothenate
    6. Pyridoxine
    7. Pyridine nucleotides
    8. Thiamin
    9. Riboflavin
    10. Thioredoxin, glutaredoxin, glutathione
    11. Menaquinone, ubiquinone
    12. Heme, porphyrin
    13. Biotin carboxyl carrier protein (BCCP)
    14. Cobalamin
    15. Enterochelin

E. Fatty acid and phosphatidic acid biosynthesis (**21**)

F. Polyamines (**3**)

III. Macromolecule metabolism (**321**)

- A. DNA (**111**)
  1. Replication
  2. Structural DNA binding proteins
  3. Recombination
  4. Repair
  5. Restriction, modification
- B. RNA (**137**)
  1. Ribosomal and stable RNAs
  2. Ribosomal proteins
  3. Ribosomes - maturation and modification
  4. Aminoacyl tRNA synthetases, tRNA modification
  5. RNA synthesis, modification, DNA transcription
  6. RNA degradation
- C. Protein (**70**)
  1. Translation and modification
  2. Chaperones
  3. Protein degradation
- D. Other macromolecules (**9**)
  1. Polysaccharides
  2. Phospholipids
  3. Lipopolysaccharides

IV. Cell structure (**119**)

- A. Membrane components (**40**)
  1. Inner membrane
  2. Outer membrane constituents
- B. Murein sacculus, peptidoglycan (**25**)
- C. Surface polysaccharides and antigens (**26**)
- D. Surface structures (**28**)

V. Cellular processes (**129**)

- A. Transport (**107**)
  1. Amino acids, amines
  2. Anions
  3. Carbohydrates, organic acids, alcohols
  4. Cations
  5. Nucleosides, purines, pyrimidines
  6. Protein, peptide secretion
  7. Other
- B. Cell division (**21**)
- C. Chemotaxis and mobility (**1**)
- D. Osmotic adaptation (**0**)
- E. Cell killing (**0**)

VI. Mobile genetic elements (**91**)

- A. Phage-related functions and prophages (**60**)

B. Plasmid-related functions (**24**)
C. Transposon-related functions (**7**)

VII. Pathogenicity, virulence, and adaptation (**131**)

A. Avirulence (**5**)
B. Hypersensitive response and pathogenicity (**3**)
C. Toxin production and detoxification (**36**)
D. Host cell wall degradation (**9**)
E. Exopolysaccharides (**10**)
F. Surface proteins (**12**)
G. Adaptation, atypical conditions (**30**)
H. Other (**26**)

VIII. Hypothetical, unknown, dubious (**1518**)

A. Conserved proteins with unknown functions (**326**)
B. No hits/only low score hits (**1188**)

Apêndice E

Primers para fechamento do genoma

Os cinco lotes de *primers* desenhados na tentativa de fechamento de *gaps* do projeto genoma de *Xylella fastidiosa*.

======= lote 1 (28 primers) ==========

>06G11-05A04 3'+ (32, 50.3oC)
GTCTTATCAGGCGTCCTC
>06G11-05A04 3'- (1077, 48.7oC)
CCCTATACGGTTAAGCTC

>06G11-05A04 5'+ (10g6, 49.8oC, 43647-43665)
CAGGTTTGGTAATAGCTTG
>06G11-05A04 5'- (10g6, 50.8oC, 44814-44797, 202pb ext.)
AACACAGCATCCCTTGAC

>07C01-03H11 3'+ (3h11, 48.6oC, 188-205)
ATATTGCCTGACACTTGG
>07C01-03H11 3'- (3h11, 49.9oC, 1523-1502, hairpin a 35.3oC!!!)
GGTGTCTCTTTGTTATCTAGAG

>07C01-03H11 5'- (7c1, 45.4 oC, 34732-34749)
AAGACTACACTGACTGCG
>07C01-03H11 5'+ (7c1, 44.8oC, 35710-35693 82pb ext.)
CTACACCCTCAGGTAAAC

>05B08-07H03 3'+ (7h3, 46.0oC, 46219-46237
CAAAGCACTCTAGTTAACC
>05B08-07H03 3'- (7h3, 47.6oC, 47812-47795, 530pb ext.)
ATCCTGTGTTACTCGAGC

>05B08-07H03 5'+ (5b8, 55oC, 81-98)
CCTTGGCACCGTACTAGC
>05B08-07H03 5'- (5b8, 50.3, 1115-1098)
TCAGCATCAGGAGACTCC

>02A11-07C09 3'- (7c9, 48.0oC, 172-189)
GACACAATCATGGACCTG
>02A11-07C09 3'+ (7c9, 45.7oC, 1320-1301)
GTAACGAACATTAGTGAGAG

>02A11-07C09 5'- (2a11, 52.4oC, 37621-37638)
ATGGTTCCGTCCTCCTAC
>02A11-07C09 5'+ (2a11, 53.8oC, 38654-38637, 202pb ext.)
TGCGGTGTGTGGTTACAG

>07F02-11H05 3'- (11h5, 54.1oC, 137-154)
CTGGCTCGTCGGATAGAC
>07F02-11H05 3'+ (11h5, 50.2oC, 1910-1892)
GACCTGTGTTTCCTGAAAC

>07F02-11H05 5'- (7f2, 52.2oC, 37207-37224)
CAGTCTGCGGTGATTAGG
>07F02-11H05 5'+ (7f2, 54.8oC, 38156-38139, 68pb ext.)
GCTGCACCTTTACCGTTC

>01E01-09E10 3'+ (9d10, 49.1oC, 37130-37147)
AGCTTCGCTCTGTTAAAG
>01E01-09E10 3'- (9d10, 45.9oC, 38159-38142)
ATCACACCTTGAAATACG

>01G04-09D11 3'+ (9d11, 50.7oC, 35362-35380)
CGACTTCTTTGATCTTTCC
>01G04-09D11 3'- (9d11, 55.4oC, 36427-36406, 53pb ext.)
TCGGTGTATAACAGCAAATTAG

>01G04-09D11 5'+ (1g4, 50.9oC, 340-357)
GATGTCAGGTGTCATGGG
>01G04-09D11 5'- (1g4, 49.7oC, 1197-1180)
TGGCCAGACACATATAGC

>10F11-03A12 3'- (7a8, 53.8oC, 50-67)
CGTTTGCGATGGTGTTAG
>10F11-03A12 3'+ (7a8, 54.9, 890-973)
GTACGCCTTCCACTGTGC

======= lote 2 (6 primers) ==========

>7a9NAO7c10
sequence=GGTGGTAGGTGGGGTTGG
PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
300 317 18 4.0 66.7 61.9 4.0 0.0 26.0 16.0
(94pb da cicatriz de BamHI)

>11a7NAO11a8
sequence=AGGAACGGCGGCAAGAGG

PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
595 612 18 16.0 66.7 61.9 8.0 4.0 14.0 14.0
(139pb da cicatriz de BamHI)

>7c10NAO7a9
sequence=CTATCAACTGAGGCGATTCCG

PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
534 554 21 24.0 52.4 60.0 8.0 8.0 14.0 8.0
(194pb da cicatriz de BamHI)

>11a2E
sequence=GTCAGCATTTCCAGCAGCG

PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
40923 40941 19 24.0 57.9 59.7 8.0 8.0 16.0 8.0
(140pb da cicatriz BamHI)

>11a2D
sequence=GCAAGTGTCAATCACATGAGCG

PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
1028 1049 22 28.0 50.0 60.1 12.0 8.0 16.0 8.0

>7a2NAO7a11
sequence=GTCGTGGTTGTTGAAAGTGCC

PRIMER-SELF PRIMER-OTHER

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
561 581 21 24.0 52.4 60.0 8.0 8.0 14.0 14.0
(78pb da cicatriz BamHI)

======= lote 3 (8 primers) ==========

serial#=1a1NAO3c12
template=X0PM-01A01-A028
sequence=GTGTGTGTTTGATGGAGTGTGG
flags=OK
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
526 547 22 4.0 50.0 60.1 4.0 0.0 16.0 14.0

serial#=2d3d
template=X0IU-02D03-B110
sequence=CAGTGGTATCGGCAACAGG
flags=OK

5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
112 130 19 16.0 57.9 59.7 8.0 4.0 16.0 14.0

serial#=2d3e
template=X0UI0501A04
sequence=GCTGGTGTTGTAGGCAAGG
flags=OK (hairpin loop 0oC, loop interno -84oC)
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
494 512 19 16.0 57.9 59.7 8.0 4.0 14.0 10.0

serial#=2g4e (atencao!!! esta regiao e' muito complicada)
template=X0EZ-02G04-C165
sequence=TGCAGTTGGGTGAGCTGGG
flags=OK (hairpin 5.6oC, dimero -92oC)
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
1709 1727 19 20.0 63.2 61.9 12.0 4.0 20.0 14.0

serial#=2g4d
template=
sequence=GGATGGTGAGTCGGTTAAGG
flags=OK (dimero -98oC)
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
437 456 20 16.0 55.0 59.9 8.0 4.0 20.0 10.0

serial#=7b8e
template=
sequence=CGATACTTTGCATCGGCAGG
flags=
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
582 601 20 20.0 55.0 59.9 12.0 4.0 18.0 10.0

serial#=7b8d
template=X0PM-07B08-A115
sequence=CGTGCAGCGATCAATGCG
flags=
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
753 770 18 28.0 61.1 59.6 12.0 8.0 18.0 14.0

serial#=9e9d
template=X0JJ-09E09-B078
sequence=ACTCCAAGGCACTGCACC
flags=
5' end 3' end length Score G+C(%) Tm 3' Internal 3' Internal
754 771 18 28.0 61.1 59.6 12.0 8.0 16.0 10.0

======= lote 4 (3 primers) ==========

serial#=7a8e

template=X0UV0504A10
sequence=GGCTAACACCATCGCAAACG
flags=OK

5' end	3' end	length	Score	G+C(%)	Tm	3' Internal	3' Internal		
353	372	20	24.0	55.0	59.9	8.0	8.0	16.0	10.0

serial#=7a8d
template=
sequence=GTGATGCGTTTGGTGCGG
flags=OK

5' end	3' end	length	Score	G+C(%)	Tm	3' Internal	3' Internal		
879	896	18	16.0	61.1	59.6	8.0	4.0	18.0	18.0

serial#=7a1Nao7c4 (7a1e)
template=
sequence=TCGACGACAGTGCCAACC
flags=OK (estruturas sub-zero)

5' end	3' end	length	Score	G+C(%)	Tm	3' Internal	3' Internal		
541	558	18	20.0	61.1	59.6	12.0	4.0	18.0	10.0

======= lote 5 (2 primers) ==========

>7c6d(Lau2, 246pb)
CGCTCAAGGTTACGCACTCCAGACTCACG

>7c6e (Lau1, 1486pb)
GCAGCACAACAAAGCTCAGGCGCTACAGG

Bibliografia

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J. (1997). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res 25* (17): 3389-3402.

Anderson,S., de Bruijn,M.H., Coulson,A.R., Eperon,I.C., Sanger,F., and Young,I.G. (1982). **Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome**. *J Mol Biol 156* (4): 683-717.

Arruda,P., da Silva,F.R., Kemper,E.L., Vettore,A.L., Leite,A., and Silva,M.J. (2000). **Isolated gum operon from *Xylella fastidiosa*, isolated nucleic acid molecules therefrom, and uses thereof. 09/567,465**. USA. 9-5-2001.

Audic,S. and Claverie,J.M. (1998). **Self-identification of protein-coding regions in microbial genomes**. *Proc Natl Acad Sci U S A 95* (17): 10026-10031.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., and Wheeler,D.L. (2000). **GenBank**. *Nucleic Acids Res 28* (1): 15-18.

Beretta,M.J.G., Harakava,R., Chagas,C., Derrick,K.S., and Barthe,G.A. (1996). **First report of *Xylella fastidiosa* in coffe**. *Plant Dis 80*: 821-826.

Bevan,M. (2000). **Plant pathology. The bugs from Brazil**. *Nature 406* (6792): 140-141.

Bonfield,J.K., Smith,K., and Staden,R. (1995). **A new DNA sequence assembly program**. *Nucleic Acids Res 23* (24): 4992-4999.

Dear,S. and Staden,R. (1991). **A sequence assembly and editing program for efficient management of large projects**. *Nucleic Acids Res 19* (14): 3907-3911.

Derrick,K.S. and Timmer,L.W. (2000). **Citrus blight and other diseases of recalcitrant etiology**. *Annu.Rev.Phytopathol. 38*: 181-205.

Elzanowski, A. and Ostell, J. (1999). **The Genetic Codes**: *http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=t#SG11*

Ewing,B., Hillier,L., Wendl,M.C., and Green,P. (1998). **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res 8* (3): 175-185.

FAPESP (1998). **The Sequencing of the *Xyllela fastidiosa* genome**: *http://www.lbi.ic.unicamp.br/xf-old/strategy.html*

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., and et,a. (1995). **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science 269* (5223): 496-512.

Frohme,M., Camargo,A.A., Heber,S., Czink,C., Simpson,A.J., Hoheisel,J.D., and de,S. (2000). **Mapping analysis of the Xylella fastidiosa genome**. *Nucleic Acids Res 28* (16): 3100-3104.

Gordon,D., Abajian,C., and Green,P. (1998). **Consed: a graphical tool for sequence finishing**. *Genome Res 8* (3): 195-202.

Green, P. (1999). **phrap.doc**: *http://bozeman.genome.washington.edu/phrap.docs/phrap.html*

Hagmann,M. (2000). **Genomes 2000. Intimate portraits of bacterial nemeses.** *Science 288* (5467): 800-801.

Helms,C., Dutchik,J.E., and Olson,M.V. (1987). **A lambda DNA protocol based on purification of phage on DEAE-cellulose**. *Methods Enzymol 153*: 69-82.

Henikoff,S. and Henikoff,J.G. (1992). **Amino acid substitution matrices from protein blocks**. *Proc Natl Acad Sci U S A 89* (22): 10915-10919.

Kamoun,S. and Hogenhout,S.A. (2001). **Agricultural Microbes Genome 2. First glimpses into the genomes of plant-associated microbes**. *Plant Cell 13* (3): 451-458.

Karlin,S. and Altschul,S.F. (1990). **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes**. *Proc Natl Acad Sci U S A 87* (6): 2264-2268.

Li ,W.B., Zreik,L., Fernandes,N., Miranda,V.S., Teixeira,D.C., Ayres,A.J., Garnier,M., and Bove,J.M. (1999). **A triply cloned strain of *Xylella fastidiosa* multiplies and induces symptoms of citrus variegated chlorosis in sweet orange**. *Curr Microbiol* **39** (2): 106-108.

Lukashin,A.V. and Borodovsky,M. (1998). **GeneMark.hmm: new solutions for gene finding**. *Nucleic Acids Res 26* (4): 1107-1115.

Ottoboni,L.M., Leite,A., Yunes,J.A., Targon,M.L., de Souza,F., and Arruda,P. (1993). **Sequence analysis of 22 kDa-like alpha-coixin genes and their comparison with homologous zein and kafirin genes reveals highly conserved protein structure and regulatory elements**. *Plant Mol Biol 21* (5): 765-778.

Pearson,W.R. (1990). **Rapid and sensitive sequence comparison with FASTP and FASTA**. *Methods Enzymol 183*: 63-98.

Queiroz -Voltan,R.B., Paradela,O., Carelli,M.L.C., and Fahl,J.I. (1998). **Aspectos estruturais de cafeeiro infectado com *Xylella fastidiosa***. *Bragantia 57* (1): 23-33.

Rabiner,L.R. (1989). **A tutorial on hidden markov models and selected applications in speech recognition**. *Proceedings of the IEEE* **77**: 257-285.

Riley,M. (1998). **Systems for categorizing functions of gene products**. *Curr Opin.Struct.Biol 8* (3): 388-392.

Rossetti,V., Gonzales,M.A., and Donadio,L.C. (1998). **History**. In Citrus Variegated Chlorosis. L.C.Donadio and C.S.Moreira, eds. (Bebedouro:

Salzberg,S.L., Delcher,A.L., Kasif,S., and White,O. (1998). **Microbial gene identification using interpolated Markov models**. *Nucleic Acids Res 26* (2): 544-548.

Sambrook,J., Fritsch,E.F., and Maniatis,T. (1989). **Molecular Cloning: a laboratory manual** (New York: Cold Spring Harbor Laboratory).

Schafer,B.W., French,W.J., and Schaad,M.W. (1981). **Axenic Culture of the bacteria-assiciated with phony disease of peach and plum leaf scald**. *Curr Microbiol 6*: 309-314.

Staden,R. (1996). **The Staden sequence analysis package**. *Mol Biotechnol 5* (3): 233-241.

Wall,L., Christiansen,T., and Schwartz,R. (1996). **Programming Perl** (Sebastopol - CA - USA: O'Reilly & Associates).

Wells,J., Raju,B., Jung,H., Weisburg,W., Mandelco-Paul,L., and Brenner,D. (1987). ***Xylella fastidiosa* gen nov, sp nov gramnegative, xylem limited fastidious plant bacteria related to *Xanthomonas* ssp.** *International Jouranl of Systematic Bacteriology 37*: 136-143.