UNIVERSIDADE ESTADUAL PAULISTA "JULIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CAMPUS DE JABOTICABAL

PREDIÇÃO COMPUTACIONAL DE PROMOTORES EM *Xanthomonas axonopodis* pv. *citri*

Renata Izabel Dozzi Tezza
Tecnóloga em Processamento de Dados

JABOTICABAL – SÃO PAULO – BRASIL
2008

UNIVERSIDADE ESTADUAL PAULISTA "JULIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CAMPUS DE JABOTICABAL

PREDIÇÃO COMPUTACIONAL DE PROMOTORES EM *Xanthomonas axonopodis* pv. *citri*

Renata Izabel Dozzi Tezza

Orientadora: Profª Drª Maria Inês Tiraboschi Ferro

Co-orientador: Dr Marcelo Luiz de Laia

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Campus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Agronomia (Genética e Melhoramento de Plantas).

JABOTICABAL – SÃO PAULO – BRASIL

Agosto de 2008

Tezza, Renata Izabel Dozzi
T356p Predição computacional de promotores em ***Xanthomonas axonopodis*** pv. ***citri*** / Renata Izabel Dozzi Tezza. -- Jaboticabal,2008
xiv, 79 f. : il. ; 28 cm

Dissertação (mestrado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2008
Orientadora: Maria Inês Tiraboschi Ferro
Banca examinadora: Marcelo Luiz de Laia, Manoel Victor Franco Lemos, Poliana Fernanda Giachetto
Bibliografia

1. ***Xanthomonas axonopodis*** pv. ***citri***. 2. Promotor. 3. Modelo Oculto de Markov. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 632.23:631.52

DADOS CURRICULARES DO AUTOR

RENATA IZABEL DOZZI TEZZA – natural de Santo André – SP, nascida aos 8 de novembro de 1973. Em novembro de 1990, formou-se Técnica em Programação de Microcomputadores pelo SENAC – Unidade de Santo André, SP. Em fevereiro de 1993, ingressou no curso graduação de Tecnologia em Processamento de Dados da FATEC - Faculdade de Tecnologia (vinculada ao Centro Paula Souza e à UNESP - Universidade Estadual Paulista "Julio de Mesquita Filho"), na cidade de Taquaritinga, estado de São Paulo, graduando-se em dezembro de 1997. Trabalha na área de Tecnologia da Informação desde fevereiro de 1991, tendo atuado nas sub-áreas de treinamento em informática, suporte técnico, programação, consultoria e análise de sistemas em algumas empresas e micro-empresas. Desde 2000, ocupa função de Analista de Sistemas em Bioinformática no LBM – Laboratório de Bioquímica e Biologia Molecular do Departamento de Tecnologia do campus de Jaboticabal (FCAV) da UNESP, contratada inicialmente através do FUNDECITRUS (Fundo de Defesa da Citricultura do Estado de São Paulo). Atualmente, é aluna do curso de Mestrado, do programa de Agronomia (Genética e Melhoramento de Plantas) da FCAV/UNESP, iniciado em março de 2006.

UMA ORAÇÃO

Recuse-se a cair.

Se não puder se recusar a cair, recuse-se a ficar no chão.

Se não puder se recusar a ficar no chão,

eleve o coração aos céus e, como um mendigo faminto,

peça que o encham, e ele será cheio.

Podem empurrá-lo para baixo.

Podem impedí-lo de se levantar.

Mas ninguém pode impedí-lo de elevar seu coração aos céus - só você.

É no meio da aflição que tantas coisas ficam claras.

Quem diz que nada de bom resultou disso,

ainda não está escutando.

Clarissa Pinkola Estés (poeta e psicanalista americana)

AGRADECIMENTOS

A Deus, que me fortalece a cada dia, me dando a oportunidade de novos desafios e energia para vencê-los!

À minha orientadora Profª Drª Maria Inês Tiraboschi Ferro e ao Prof Dr Jesus Ap. Ferro, pelo incentivo e a amizade, pelos valiosos conselhos, pela oportunidade e pela disponibilização da infraestrutra de bioinformática do Laboratório de Bioquímica e Biologia Molecular (LBM) do Departamento de Tecnologia para a realização deste projeto.

Ao Dr Marcelo Luiz de Laia, pela co-orientação, idéias e disponibilidade para ensinamentos, correções e constante presença neste trabalho.

À Drª Agda Paula Facincani, por ceder os dados e resultados experimentais de sua tese de doutorado, que complementou de forma significativa este projeto. E pela constante disponibilidade em discutir detalhes sobre o mesmo. Pelo carinho e amizade.

Ao Prof Dr Nalvo Franco de Almeida, Professor Adjunto do Departamento de Computação e Estatística da Universidade Federal de Mato Grosso do Sul, pela dispobilibilização de um modelo matemático, imprescindível para a realização deste trabalho, e por toda a pronta disponibilidade em colaborar explicando e discutindo este modelo.

Aos membros da Banca Examinadora do Exame Geral de Qualificação: Profª Drª Lucia Maria Carareto Alves, Profª Drª Janete Apparecida Desidério Sena e Dr Marcelo Luiz de Laia (presidente da banca), pela presença e valiosas sugestões.

A todos os Professores das disciplinas que cursei antes e durante o curso, por todos os ensinamentos!

Aos membros da Comissão Examinadora do Exame de Dissertação: Drª Poliana Fernanda Giachetto, Prof Dr Manoel Victor Franco Lemos e Dr Marcelo Luiz de Laia (presidente da banca), pela presença e valiosas sugestões.

A todos os funcionários do Setor de Pós-Graduação, que sempre nos atendem com exemplar profissionalismo e competência.

A todos os funcionários, pesquisadores, professores e alunos do LBM, BCCCenter e Departamento de Tecnologia da FCAV UNESP, que estiveram presente no meu dia-a-dia.

Aos meus “anjos da guarda”, disfarçados de amigas, que me ofereceram não só carinho, amizade, compreensão e um ombro amigo sempre disponível, como também auxílio profissional com valiosas explicações, em muitos momentos de dúvidas tanto durante as disciplinas como também durante o desenvolvimento deste trabalho. Impossível expressar todo o meu sentimento de gratidão a essas “irmãs de coração”.

A todo o pessoal do Reiki pela compreensão da minha ausência, pela amizade e pelas orações.

Em especial, agradeço à minha família que caminhou junto a mim em todos os momentos desta importante etapa de minha existência. Em especial, agradeço por me lembrarem mais uma vez que o limite de um ser humano vai até o limite da sua fé. “Nada é impossível !” Amo vocês!

À multidão de pessoas que cruzou o meu caminho durante este curso: tenham sempre em mim alguém profundamente grata!

SUMÁRIO

Página

LISTA DE TABELAS

Página

LISTA DE FIGURAS

Página

PREDIÇÃO COMPUTACIONAL DE PROMOTORES EM *Xanthomonas axonopodis* pv. *citri*

RESUMO - Com o seqüenciamento completo do genoma do fitopatógeno ***Xanthomonas axonopodis*** pv. ***citri*** (Xac), em 2002, inúmeras possibilidades de estudo foram viabilizadas, dando margem à busca de novas formas de controle do cancro cítrico, baseadas em alvos moleculares. Estudos dessa natureza têm mostrado a existência de genes que somente são expressos quando a bactéria está se desenvolvendo ***in planta***. Sabe-se que essa regulação é dependente da região promotora e sua identificação pode possibilitar avanços significativos na busca do controle dessa doença. Apesar do crescente avanço das técnicas experimentais ***in vitro*** em biologia molecular, identificar um número significante de promotores ainda é uma tarefa difícil e dispendiosa. Os métodos computacionais existentes enfrentam a falta de um número adequado de promotores conhecidos para identificar padrões conservados entre as espécies. Logo, um método para predizê-los em qualquer organismo procariótico ainda é um desafio. O Modelo Oculto de Markov é um modelo estatístico aplicável a muitas tarefas em biologia molecular. Entre elas, predição e mapeamento de seqüências promotoras no genoma de um procarioto. Neste trabalho, estudou-se o mapeamento ***in silico*** de promotores gênicos de todo o genoma da Xac e em 68% dos genes a localização de um promotor foi indicada. A análise comparativa com dados experimentais de proteômica mostrou que 72% das proteínas expressas identificaram-se com elementos desta lista de promotores, o que corresponde a 27% do total de genes do genoma. À partir destes dados foi possível levantar um rol de informações sobre estes candidatos a promotores incitando a novos estudos moleculares.

Palavras-chave: Promotor, Modelo Oculto de Markov, Cancro cítrico, ***Xanthomonas axonopodis*** pv. ***citri***, Proteoma, Genoma.

Xanthomonas axonopodis pv. *citri* PROMOTERS COMPUTATIONAL PREDICTION

SUMMARY - With the complete genome sequencing of the phytopathogen ***Xanthomonas axonopodis*** pv. ***Citri*** (Xac), in 2002, several study possibilities were made practical and then creating the search of new citrus canker control ways, based in molecular aims. This kind of studies has shown the genes existences that are only expressed when the bacteria are developing itself ***in plant.*** It has been known that this regulation is promoter region dependent and its identification can allow significant advances in the search of this disease control. Although increasing advance of ***in vitro*** experimental techniques in molecular biology, identifying a meaningful number of promoters is still a hard and expensive task. The existents computer science methods face the need of a proper number of known promoters to identify conserved patterns among the species. Therefore, a method to predict them in any prokaryote organism is still a challenge. The ***Hidden Markov Model*** (HMM) is a statistic model applicable in many tasks in molecular biology. Among them, prediction and mapping of the promoters sequences in prokaryotic genome. In this work, which has studied the genic promoters ***in silico*** mapping of the whole Xac genome, in 68% of the genes the promoter localization was indicated. The proteomic experimental data comparative analysis showed that 72% of the expressed proteins identified with elements from the promoters list, which corresponds 27% of the genome genes total. According to these data it was possible to generate an information roll about these promoters candidates inciting new molecular studies.

Keywords: Promoter, Hidden Markov Model, Citrus canker, ***Xanthomonas axonopodis*** pv. ***citri***, Proteome, Genome.

I. INTRODUÇÃO

O sistema agroindustrial da laranja é sem dúvida um caso de sucesso no agronegócio brasileiro. Afinal, é um produto que atende cerca de 81% das transações internacionais, fato que coloca o Brasil na primeira posição do ranking dos maiores exportadores de suco de laranja do mundo correspondendo a 1,8% da pauta brasileira de exportação e a um terço de toda a produção mundial, com seus quase 9 milhões de toneladas por ano (EXAME, 2005; EXAME, 2006). O setor emprega diretamente cerca de 400 mil pessoas e é a atividade econômica essencial de 322 municípios paulistas e 11 mineiros. A maior citricultura do mundo, em resumo (ABECITRUS, 2008). No entanto, a manutenção dessa posição é constantemente ameaçada por pragas e doenças, tais como Tristeza (CTV-Citrus Tristeza Virus), “Greening”, cigarrinhas, dentre outros, sendo o cancro cítrico uma das mais importantes (FUNDECITRUS, 2007).

O fitopatógeno ***Xanthomonas axonopodis*** pv. ***citri*** (Xac), causador do cancro cítrico, é capaz de provocar intensa desfolha, depreciação e queda de frutos, o que causa enormes prejuízos (FUNDECITRUS, 2007). Não existe um método de combate ao cancro cítrico definitivo, porém, os métodos possíveis são: sanitização, controle químico (cobre) ou erradicação através de queimada. No caso do estado de São Paulo e Minas Gerais, onde o cancro cítrico é controlado desde o seu surgimento, a erradicação, apesar de ser uma metodologia drástica, ainda é a melhor opção. Se as plantas infectadas somam menos de 0,5% do talhão, todas as plantas a um raio de 30 metros da infecção são erradicadas (aproximadamente 100 plantas). Se mais de 0,5% do talhão estiver infectado, todo o talhão é queimado. Após a erradicação, fica proibido o cultivo de citros no local pelos próximos dois anos. No entanto, a erradicação das plantas contaminadas não garante a eliminação da bactéria causadora do cancro cítrico (FUNDECITRUS, 2007; FUNDECITRUS, 2008).

Segundo levantamento amostral do FUNDECITRUS (Fundo de Defesa da Citricultura do Estado de São Paulo), o índice de contaminação do parque citrícola de São Paulo e Minas Gerais, em 2007, foi de 0,10%. Mesmo a doença estando

controlada, 7 milhões de plantas foram erradicadas de 1999 até 2007. Em recente pesquisa da Dra Margarida Figueiredo – ESALQ/USP, verificou-se que a erradicação desde o surgimento da doença no estado de São Paulo, apesar de ser uma metodologia totalmente anti-econômica, impediu a redução da produção citrícola de 20% a 24%. Ou seja, se a campanha de erradicação tivesse sido interrompida em 1970, a citricultura teria perdido R$ 2 bilhões (FUNDECITRUS, 2008).

A partir de 1996, uma intensificação no número de focos de cancro cítrico inspirou a necessidade de se atuar contra esta doença, através de medidas baseadas numa maior compreensão dos mecanismos de interação entre planta hospedeira e patógeno. Isso pôde ser feito através de estudos dos mecanismos de patogenicidade e virulência utilizados pela bactéria ***Xanthomonas axonopodis*** pv. ***citri*** em sua estratégia de infecção (RUDOLPH, 1993). Um estudo de retorno de investimentos em pesquisa e tecnologia empregadas na citricultura, referente aos dados do período de 1970 a 2004, mostrou que para cada R$1,00 utilizado em pesquisa na citricultura paulista, houve um retorno médio de R$13,67 (FUNDECITRUS, 2008).

Dada toda a sua importância, a bactéria ***Xanthomonas axonopodis*** pv. ***citri*** teve seu genoma totalmente seqüenciado por DA SILVA e colaboradores (2002), com investimento da FAPESP (Fundação de Apoio à Pesquisa do Estado de São Paulo) em parceria com o FUNDECITRUS. Desde então, estudos na área de biologia molecular tiveram início na tentativa de descobrir formas de controle da doença baseadas no gene ou no seu produto.

Um gene, em bactérias, é composto basicamente por uma região codificadora (CDS), por uma região promotora (“upstream” ao gene) e por outras regiões, também importantes, como o operador que inibe ou ativa a transcrição em diferentes taxas e o terminador, seqüência “downstream” ao final do gene que sinaliza à RNA polimerase a parada ou pausa na transcrição (GRIFFITHS, 1998). Formas alternativas de um mesmo gene são chamados de alelos.

Dada a co-evolução entre planta e patógeno, barreiras que impedem o desenvolvimento do patógeno no interior do hospedeiro são freqüentemente superadas

por meio, muitas vezes, do aparecimento de um novo alelo de virulência no patógeno ocasionado por mutações gênicas. Por outro lado, a ocorrência de mutações em promotores gênicos é muito rara, fato este que os torna excelentes alvos no controle de doenças.

Estudos baseados em biologia molecular têm mostrado que um promotor é constituído de 3 regiões características: uma seqüência de 6 nucleotídeos (hexâmero) centrada por volta de -35 bases (b) do ponto inicial de transcrição (+1), outra centrada por volta de -10 b (conhecido como TATA-Box) e a região de intervalo que separa esses hexâmeros (distância ou “spacer”). Esses hexâmeros apresentam um certo grau de conservação em sua constituição quando comparados a promotores de um mesmo organismo. Além disso, a distância do intervalo entre eles, em média de 17 nucleotídeos, parece ser relevante, apesar do tamanho variável e da baixa conservação nas bases que a compõe, pois pode ser crítica na interação dos hexâmeros com a geometria da RNA polimerase (REIS, 2005).

No entanto, mesmo em procariotos, é complexo caracterizar os promotores estruturalmente. Para ***Escherichia coli***, modelo biológico padrão, o promotor é reconhecido por conter a seqüência ***TTGACA...N***$_{17}$...***TATAAT***, onde ***N*** corresponde a qualquer um dos 4 nucleotídeos. Entretanto, esse padrão exato não ocorre em nenhuma região promotora real, mas sim variações do mesmo (REIS, 2005).

Apesar do crescente avanço das técnicas experimentais em biologia molecular, caracterizar e identificar um número considerável de promotores, presentes em um dado genoma, continua sendo uma tarefa demorada, cara e nada trivial. O avanço de técnicas de seqüenciamento de genomas gerou um acúmulo de dados nos últimos anos, mostrando que os meios convencionais de análise ***in vitro*** são restritos tanto pelo custo quanto pela dificuldade em capturar informações subjacentes à regulação gênica (REIS, 2005).

Neste sentido, ferramentas de Aprendizado de Máquina[1] ganham aplicabilidade ao problema, por serem capazes de aprender de forma automatizada, a partir de dados

[1] É uma área da Inteligência Artificial (IA) que estuda métodos computacionais para adquirir novos conhecimentos bem como meios de organizar o conhecimento já existente.

disponíveis, e levantar hipóteses relevantes sobre mecanismos biológicos ocultos (BALDI & BRUNAK, 2001). Abordagens ***in silico***[1] são muito utilizadas para reconhecer essas regiões em procariotos. Porém, além do alto número de falsos positivos obtidos, um consenso entre a maioria dos trabalhos encontrados na literatura é sobre a falta de informação comprovada experimentalmente sobre promotores para se ter certeza de que os modelos computacionais aplicados são realmente funcionais (ALMEIDA & SETUBAL, 1998; MONTEIRO, 2005; REIS, 2005).

Como objetivo, no presente trabalho, buscou-se levantar o máximo possível de informações sobre os potenciais candidatos a promotores gênicos de Xac utilizando-se técnicas ***in silico***. Desse modo, as regiões que não codificam genes, ou seja, regiões intergênicas, do cromossomo e dos dois plasmídeos da Xac foram mapeadas e categorizadas segundo o seu tamanho em número de bases de nucleotídeos, indicando aquelas que possuíam características ideais para ter um promotor mapeado com o modelo ***in silico*** utilizado. Afim de se ter indícios de dados comprobatórios, os genes que puderam ter o seu candidato a promotor predito foram comparados com dados experimentais de FACINCANI (2007), trabalho este sobre o perfil proteômico da Xac, onde foram investigadas proteínas expressas sob uma condição não infectante (***in vitro***) e 3 condições infectantes (***in vitro*** e ***in vivo***).

[1] em ou através de uma simulação computacional.

II. REVISÃO DE LITERATURA

A. Pesquisas com a bactéria *Xanthomonas axonopodis* pv. *citri*

1. Cancro cítrico

As bactérias pertencentes ao gênero ***Xanthomonas*** constituem um dos grupos de fitopatógenos mais dispersos na natureza, com capacidade de infectar aproximadamente 120 diferentes tipos de plantas monocotiledôneas e 270 dicotiledôneas. A capacidade de infectar uma grande variedade de plantas torna o estudo das ***Xanthomonas*** de grande interesse para a pesquisa, pois a produtividade de diversas culturas de interesse agronômico pode ser seriamente afetada por doenças causadas por bactérias deste gênero (CARVALHO, 2006).

Existe uma diversidade de tipos de cancro cítrico, porém, o cancro cítrico asiático (cancrose A), causado por um grupo de isolados oriundos da Ásia, é a forma mais propagada e agressiva da doença e é causado pela bactéria Gram-negativa ***Xanthomonas axonopodis*** pv. ***citri***.

Em 1957, o cancro cítrico (cancrose A) foi constatado pela primeira vez no Brasil, inicialmente encontrada em pomares na região de Presidente Prudente, SP, trazida em mudas importadas por imigrantes japoneses. A doença foi posteriormente detectada em regiões dos estados de São Paulo, Mato Grosso do Sul, Paraná, Santa Catarina e Rio Grande do Sul (NAMEKATA ***et al.,*** 1996). O cancro cítrico se alastrou rapidamente pela região produtora e, atualmente, é citado como um dos fatores responsáveis pela retração na produtividade de citrus do país, indicando o potencial destrutivo desta doença altamente contagiosa, causadora de grandes prejuízos todos os anos ao país (FUNDECITRUS, 2007).

Os sintomas do cancro cítrico são verificados em toda a parte aérea da planta cítrica, nas folhas, ramos e frutos, sendo que, nas folhas os sintomas incluem lesões iniciais constituídas de pequenas pústulas salientes, circundadas ou não por um halo

amarelo e/ou halo aquoso, nos dois lados da folha. Com o passar do tempo, as lesões aumentam de diâmetro e tornam-se pardacentas e escuras, corticosas e duras. Em frutos e ramos, os sintomas são semelhantes aos observados em folhas, porém, a parte corticosa mais pronunciada e as lesões podem ou não estar envolvidas pelo halo amarelo (KOLLER ***et al.***, 1993). A doença em estágio avançado em folhas e frutos é capaz de provocar intensa desfolha, depreciação e queda de frutos. (WHITESIDE ***et al.***,1988). Um sintoma característico e essencial para o diagnóstico do cancro cítrico é a indução da formação de tecido hiperplástico (divisões mitóticas excessivas – calogênicas), que resulta em lesões do tipo cancro. Os sintomas do cancro em si não constituem o principal problema ocasionado pela doença, pois raramente leva a planta à morte. Entretanto, em resposta ao estresse biótico, a planta responde com a produção excessiva de etileno provocando um desequilíbrio hormonal, levando à queda prematura do fruto ainda não suficientemente maduro. Tal característica torna os frutos impróprios tanto para o mercado de frutas frescas como para a produção de suco concentrado (RUDOLPH, 1993; BROWN, 2001).

A formação de cancros, com ruptura da epiderme, funciona como uma “porta de saída” para grandes quantidades de bactérias na superfície foliar ou do fruto infectado, constituindo-se em importante fonte de inóculo para a dispersão do patógeno. À curta e média distâncias, a partir de lesões, as bactérias são disseminadas principalmente através da água de chuva e vento, enquanto que à longa distância, a bactéria também pode ser levada pelo homem, propriamente ou indiretamente, por meio de materiais vegetais infectados, como frutos e mudas, assim como através de equipamentos agrícolas provenientes de áreas endêmicas (LEITE, 1990).

Como em muitas outras doenças bacterianas de plantas, o patógeno entra pelos estômatos e hidatódios do hospedeiro, por meio de ferimentos mecânicos ou pela atividade dos insetos. A infecção sempre se dá por penetração da bactéria nos tecidos, através de aberturas naturais ou ferimentos. Para folhas e caules jovens, os períodos de predisposição à doença coincidem com os fluxos de crescimento, permanecendo por seis semanas após cada brotação. Nos frutos, o período crítico vai até 90 dias após a queda das pétalas. No entando, uma vez no interior dos tecidos de citros, a colonização

pelo patógeno restringe-se ao sítio de infecção, onde se utiliza de sofisticadas estratégias para adaptar-se e multiplicar-se em plantas hospedeiras (BROWN, 2001).

2. Genoma

O seqüenciamento do genoma de bactérias fitopatogênicas é um importante passo para o entendimento da especificidade dessas bactérias aos respectivos hospedeiros e das diferenças em seus processos de patogenicidade e virulência (CARVALHO, 2006).

Concluído em maio de 2002, por mais de 10 laboratórios da Rede ONSA-FAPESP (sigla em inglês para Organização para Seqüenciamento e Análise de Nucleotídeos) em parceria com o FUNDECITRUS, o seqüenciamento completo do genoma da ***Xanthomonas axonopodis*** pv. ***citri*** (estirpe 306) possibilitou sua caracterização estrutural e foram identificadas 3 estruturas distintas: um cromossomo com 5175554 pares de bases (pb), onde 4374 ORFs[1] foram identificadas e anotadas; um plasmídeo (B) com 64920 pb e um outro plasmídeo (A) com 33699 pb, onde 73 e 42 ORFs, respectivamente, foram identificadas e anotadas. A procura, descoberta e identificação de regiões codificadoras na anotação do genoma revelaram um total de 4489 ORFs (Tabela 1). Deste total, 2770 (61,71%) ORFs puderam ter suas respectivas funções associadas a proteínas com funções conhecidas descritas na literatura e 1658 (36,93%) correspondem a proteínas hipotéticas sem função descrita até então.

Diversos genes envolvidos em patogenicidade e virulência foram descritos neste genoma. Porém, o arsenal utilizado pelo patógeno na colonização e ataque está longe de ser totalmente caracterizado e compreendido em suas particularidades individuais e inter-relacionais, ainda mais se considerar que 37,4% das ORFs contidas no genoma correspondem a proteínas sem função deduzível, e que, dentre esses, muitos podem codificar proteínas com funções importantes de patogenicidade e virulência ainda

[1] ORF: - sigla em inglês para “Open Reading Frame”, ou Quadro Aberto de Leitura; - região que se inicia por um códon iniciador e termina com um códon de parada; - é determinado por programas de computador a partir de seqüência de DNA; - é potencialmente uma região codificadora para alguma proteína.

desconhecidas (DA SILVA ***et al.***, 2002).

Tabela 1: Distribuição das ORFs anotadas do cromossomo (4374 ORFs) e plasmídeos (73 e 42 ORFs) de Xac em nove categorias funcionais primárias (DA SILVA ***et al.***, 2002).

	Categorias	**ORFs**
I	Metabolismo Intermediário	727
II	Biossíntese de pequenas moléculas	352
III	Metabolismo de macromoléculas	558
IV	Estrutura celular	202
V	Processos celulares	391
VI	Elementos genéticos móveis	190
VII	Patogenicidade, virulência e adaptação	304
VIII	Hipotéticas	1658
IX	ORFs com Categoria Indefinida	107
	Total	4489

3. Proteoma

Com as informações geradas através do seqüenciamento de DNA de um organismo pode-se inferir o perfil de proteínas produzidas pelo mesmo. Agregando informações oriundas do seqüenciamento completo do DNA de uma bactéria às informações obtidas a partir de um projeto proteoma, aumenta-se a amplitude de conhecimentos em relação às proteínas.

O interesse no estudo de proteomas tem aumentado recentemente com o aumento de seqüências disponíveis resultantes de análises de seqüenciamento de genomas. Esta foi a motivação de FACINCANI (2007), que fez o primeiro mapa de referência proteômica do fitopatógeno mais agressivo de todas as variedades de citros: ***Xanthomonas axonopodis*** pv. ***citri***.

Extratos protéicos obtidos de Xac, cultivada em uma condição não infectante (***in vitro***) e em 3 condições infectantes (***in vitro*** e ***in vivo***), foram digeridos com tripsina e os

peptídeos resultantes foram analisados através da tecnologia MudPIT (Tecnologia Multidimensional na Identificação de Proteínas), que possibilita resultados qualitativos, apontando a presença ou ausência das proteínas detectadas nos extratos em estudo. A condição não infectante, Xac cultivada em meio de cultura mínimo, representou o tempo zero, Xac cultivada em meio de cultura infectante XAM1[1] por 24 horas representou o tempo 1, Xac recuperada de folhas de laranjeira inoculadas por 3 dias representou o tempo 2, e recuperadas após 5 dias de infecção representou o tempo 3, simulando uma cinética de infecção.

FACINCANI (2007) observou a presença de 1162 proteínas expressas ***in vitro*** no tempo zero, 1167 proteínas expressas ***in vitro*** no tempo 1, 1157 proteínas expressas ***in planta*** no tempo 2 e 1072 proteínas expressas de Xac no tempo 3.

A estratégia metodológica utilizada no proteoma de Xac, por FACINCANI (2007), permitiu uma alta cobertura das proteínas codificadas pelo genoma, tanto em condições mínimas quanto em condições infectivas. Um total de 1661 proteínas distintas de Xac foram identificadas por MudPIT e estão distribuídas por todas as categorias funcionais da anotação do genoma (Tabela 2). Em comparação com a previsão teórica, cerca de 37% das proteínas foram identificadas. Segundo FACINCANI (2007), este valor é equivalente ao proteoma mais abrangente, projeto proteoma de ***E. coli,*** de TAOKA (2004), onde aproximadamente 35% das proteínas previstas no genoma foram identificadas. Assim, os resultados proteômicos apresentam-se como uma maneira real de validação de possíveis promotores.

[1] Meio de cultura XVM2 (WENGELNIK ***et al***., 1996b) com modificações, segundo comunicação pessoal do Dr Frank F. White.

Tabela 2: Distribuição das proteínas codificadas pelo cromossomo e pelos plasmídeos de Xac nas nove categorias funcionais primárias do genoma, quando expressas por MudPIT (FACINCANI, 2007).

	Categorias	Proteínas Expressas
I	Metabolismo Intermediário	333
II	Biossíntese de pequenas moléculas	218
III	Metabolismo de macromoléculas	282
IV	Estrutura celular	96
V	Processos celulares	172
VI	Elementos genéticos móveis	28
VII	Patogenicidade, virulência e adaptação	124
VIII	Hipotéticas	359
IX	ORFs com Categoria Indefinida	49
	Total	**1661**

B. Papel dos Promotores na Expressão Gênica de Organismos Procarióticos

1. Organismo Procarioto

As células compõem os organismos, que podem ser procarióticos, estruturas mais simples e sempre unicelulares (como as bactérias), ou eucarióticas, que incluem plantas multicelulares, animais e fungos, assim como organismos unicelulares, como leveduras e algas verdes. A principal diferença entre ambos os organismos está na ausência do envoltório nuclear nas células dos organismos procarióticos (ZAHA, 2003; POSTLETHWAIT & HOPSON, 2006).

No presente trabalho, o foco será em duas espécies de bactérias: ***Escherichia coli*** e ***Xanthomonas axonopodis*** pv. ***citri***.

A biologia molecular tem focalizado a sua atenção, há muito tempo, na espécie ***E. coli***. Pois, além de viver em ambientes como o intestino de humanos e de outros vertebrados, resíduos e dejetos recentes, ela pode ser cultivada facilmente em um meio de cultura simples, como em placas de Petri. A evolução tem otimizado a ***E. coli*** a

sobreviver em condições químicas variáveis e reproduzir-se rapidamente. Atualmente, temos mais conhecimento da ***E. coli,*** em termos moleculares, do que qualquer outro organismo. Na ***E. coli*** são analisadas a síntese de proteínas ou os mecanismos genéticos que foram conservados ao longo da evolução e são, essencialmente, os mesmos mecanismos que existem em nossas próprias células (MONTEIRO, 2005).

2. Promotores em Procariotos

As regiões do DNA que indicam o início da transcrição em procariontes são chamados de promotores (GRIFFITHS ***et al.***, 1998).

A transcrição dos genes em procariotos acontece de forma diferenciada dos organismos eucariotos, que possuem três tipos de RNA polimerase (não estudados neste trabalho). Em bactérias, só uma RNA polimerase sintetiza todos os tipos de RNAs. A enzima RNA polimerase é formada por quatro diferentes tipos de subunidades: beta (β), beta linha (β'), alfa (α) e sigma (σ). Esta última também conhecida como fator σ. A RNA polimerase completa é formada pelas quatro subunidades e esse conjunto é denominado de holoenzima da RNA Polimerase. O fator σ pode se dissociar do resto do complexo deixando o cerne da enzima. O fator σ está presente no momento da interação da RNA polimerase com o DNA, e assim que inicia a transcrição ele é liberado, permanecendo somente a enzima principal para a realização do alongamento da molécula (LEHNINGER ***et al***., 2004; LEWIN, 2000; GRIFFITHS ***et al.***, 1998).

O fator de transcrição σ reconhece promotores, em bactérias. E diferentes fatores σ reconhecem diferentes promotores. Em ***E. coli***, por exemplo, pelo menos seis diferentes fatores σ são sintetizados, sendo o mais importante o sigma 70 (σ^{70}), que está envolvido no metabolismo geral da célula. Outras bactérias possuem fator σ^{70} homólogo ao de ***E. coli***. Uma região promotora contendo em cada uma de suas extremidades um hexâmero, chamados de -35 e de –10, com um certo grau de conservação em sua constituição, são regiões de ligação ao DNA pelo fator σ^{70} de ***E. coli***. Tais hexâmeros foram chamados de "–35" e "-10" devido às suas localizações

relativas ao ponto de início da transcrição. Esta localização não ocorre de forma exata, mas o centro dos hexâmeros situa-se em torno de 35 e de 10 bases “upstream” ao sítio de início da transcrição, com a região central ou distância entre esses hexâmeros totalmente variável em seu tamanho e praticamente sem nenhuma conservação entre os diversos promotores deste mesmo tipo, encontrados em um mesmo organismo (LEWIN, 2000; GRIFFITHS ***et al.***, 1998).

Observa-se que estas regiões também estão presentes em outras bactérias. Apesar destas seqüências não serem perfeitamente idênticas entre os diferentes organismos, alguns nucleotídeos são encontrados com maior freqüência do que outros, numa determinada posição, constituindo o que se denomina seqüência consenso. Dois consensos, com 6 nucleotídeos cada (hexâmero), foram determinados a partir desse tipo de análise. A seqüência consenso -10 é TATAAT e a seqüência consenso -35 é TTGACA. Os consensos estão separados entre si por cerca de 15 a 18 nucleotídeos na maioria dos casos analisados em ***E. coli*** (LEHNINGER et al., 2004; LEWIN, 2000; GRIFFITHS ***et al.***, 1998).

Em especial, na ***E. coli***, a transcrição envolve três estágios distintos: iniciação, alongamento e término.

Na etapa de iniciação, a subunidade dissociativa da RNA polimerase, o fator σ, permite que a RNA polimerase reconheça e se ligue especificamente ao DNA na região promotora. Primeiro a holoenzima procura um promotor e inicialmente se liga frouxamente a ele reconhecendo as regiões –35 e –10. A estrutura resultante é chamada de complexo promotor fechado. Então, a enzima liga-se mais fortemente deselicoidizando bases próximas ao hexâmero –10. Quando a RNA polimerase ligada causa esta desnaturação local do duplex de DNA, diz-se que se formou um complexo promotor aberto. Nesta etapa de iniciação, a formação de um complexo aberto, requer o fator σ (LEHNINGER et al., 2004; GRIFFITHS ***et al.***, 1998).

Em bactérias a transcrição e a tradução ocorrem simultaneamente. Tão logo o RNA esteja sendo sintetizado pela RNA polimerase, ribossomos já estão se ligando para iniciarem a tradução. O mesmo transcrito poderá ser retranscrito várias vezes durante o ciclo celular. Uma outra RNA polimerase pode reiniciar a transcrição a partir

do promotor, assim que esse transcrito for liberado do DNA (LEWIN, 2000).

O processo de alongamento da cadeia continua até que um sinal de terminação seja atingido, ou pode ocorrer espontaneamente.

A RNA polimerase também reconhece sinais para término da cadeia, o que envolve a liberação do RNA nascente e da enzima do molde. Existem dois tipos de mecanismos para terminação, em ***E. coli***. No primeiro o término é direto. O sinal de terminação pode ser constituído por uma estrutura secundária em forma de “grampo de cabelo ou haste-alça”, em inglês denominado “hairpin”, seguida por uma seqüência de aproximadamente 8 resíduos “U”, que ocorre após o códon de parada***.*** A estrutura de grampo é formada por uma região rica em G e C, de aproximadamente 7 a 20 bases, que exibe simetria dupla pela presença de um palíndromo. A formação do grampo leva a RNA polimerase a diminuir a sua velocidade ou até a interromper a síntese. A interação U-A é fraca e, portanto, a seqüência de Us permite à RNA polimerase se dissociar do DNA molde. Esse tipo de terminação é denominado terminação independente de ρ (ro) (GRIFFITHS ***et al.***, 1998; LEHNINGER ***et al.***, 2004).

No segundo tipo de terminação, o auxílio de um fator protéico adicional, chamado de fator ρ, é necessário para que a RNA polimerase reconheça os sinais de término. Esta terminação é chamada de terminação dependente de ρ e nessas também é encontrada uma região de simetria dupla, rica em C e fraca em G e portanto, apresenta baixa complementaridade entre as bases, formando um grampo "fraco" no RNA. Além disto, não existe uma seqüência rica em Us após o grampo. Neste caso, torna-se necessária a participação da proteína auxiliar (fator ρ) para a liberação da RNA polimerase (LEHNINGER ***et al.***, 2000; GRIFFITHS ***et al.***, 1998). A primeira etapa do término dependente de ρ, é a ligação de ρ a um ponto específico no RNA, chamado de “rut”. Após a ligação, ρ retira o RNA da RNA polimerase, provavelmente translocando-se ao longo do mRNA[1]. Os sítios “rut” estão situados “upstream” de seqüências nas quais a RNA tende a fazer uma pausa. A eficiência de ambos os mecanismos de término é influenciado pelas seqüências vizinhas e outros fatores protéicos também

[1] mRNA: RNA Mensageiro. Classe de RNA que copia a mensagem de um gene no DNA e se dirige para os ribossomos, onde a informação é utilizada para sintetizar uma proteína.

(GRIFFITHS ***et al.***, 1998; LEHNINGER ***et al.***, 2004).

Assim como acontece na transcrição, a tradução também envolve três etapas: iniciação, alongamento e terminação. A iniciação inclui a reunião dos componentes envolvidos na tradução: as duas subunidades ribossômicas, o mRNA a ser traduzido, e o tRNA[1] com o aminoácido específico para o primeiro códon[2]. Envolve também GTP, que fornece energia para o processo, e fatores de iniciação, facilitadores do processo. O ribossomo se liga à seqüência de ligação do ribossomo RBS[3], localizada de 6 a 10 bases à montante do códon de início. Em seguida, um tRNA iniciador especial reconhece o códon de início e o alongamento da cadeia polipeptídica promove a adição de aminoácidos na extremidade carboxila da cadeia. Ao formar a ligação peptídica, o ribossomo avança três nucleotídeos em direção ao terminal 3' do mRNA. A terminação da tradução ocorre quando o códon de parada é reconhecido (LEHNINGER ***et al.***, 2004; GRIFFITHS ***et al.***, 1998).

3. Estruturas de *Operons*

Genomas bacterianos estão organizados em unidades de expressão envolvidas por sítios onde a transcrição do DNA em RNA inicia e termina. Uma unidade de expressão, também chamada de unidade de transcrição, pode conter mais de um gene com um único promotor realizando a regulação de sua transcrição, e dando origem a um transcrito único contendo todos os genes a serem traduzidos. Neste caso trata-se de um ***operon***. Um ***operon*** é um grupamento de genes adjacentes sob controle do mesmo promotor sendo, portanto, uma unidade genética de expressão coordenada. É uma forma de otimizar o processo, colocando próximos os genes que executam funções relacionadas (LEHNINGER ***et al.***, 2004; GRIFFITHS ***et al.***, 1998).

Quando um ***operon*** é induzido, todos os genes que dele fazem parte são transcritos numa molécula única de mRNA. Tal molécula é denominada RNA

---

[1] tRNA: RNA transportador. Classe de RNA que carrega os aminoácidos que serão ligados na cadeia polipeptídica em formação nos ribossomos.

[2] Seqüência de três bases de DNA ou RNA que codifica um aminoácido ou proteína.

[3] RBS: sigla em inglês para Sítio de Ligação ao Ribossomo ("Ribossome Binding Site").

policistrônico. A tradução ocorre de forma independente. O ribossomo se liga ao RBS e cada gene é traduzido. A tradução de cada gene, também pode ocorrer várias vezes (LEWIN, 2000).

A estrutura de um ***operon*** é uma importante característica organizacional de genomas bacterianos. Muitos conjuntos de genes ocorrem na mesma ordem em múltiplos genomas. Estes grupos de genes conservados representam candidatos a ***operons***. As características de um grupo de genes candidatos a um ***operon*** incluem: compartilhamento de determinados elementos regulatórios; hexâmeros –35 e –10 localizados na região “upstream” do ***operon***; genes arranjados em seqüência na mesma fita; genes separados por distâncias curtas, em geral no máximo 150 bases; genes conservados em dois ou mais genomas filogeneticamente relacionados e suas funções estão geralmente relacionadas (CHEN ***et al***., 2004).

C. Biologia Molecular Computacional

Vive-se atualmente, um dos momentos mais privilegiados na história da Ciência. Tem-se, como um dos grandes benefícios da Biologia Molecular, uma impressionante quantidade de informações genômicas acumuladas nas bases de dados mundiais. Em nenhuma outra época houve tantos dados à disposição dos pesquisadores como agora.

No entanto, a Biologia Molecular, por si só, não seria capaz de disponibilizar esses dados genômicos sem o auxílio da informática. Deste modo, nasceu a Biologia Molecular Computacional. Pode-se, então, definí-la como o estudo e a aplicação de modelos e técnicas computacionais aos problemas de biologia molecular. Uma de suas sub-áreas, a Bioinformática, consiste no desenvolvimento de ferramentas para a manipulação e análise de biosseqüências (DNA ou proteína) (GIBAS & JAMBECK, 2001).

Apesar dos grandes avanços ocorridos dentro da Bioinformática, tais como programas para a montagem de seqüências pequenas de DNA até ferramentas para a montagem de genomas complexos, muitos desafios ainda estão por serem

suplantados. Um desses desafios, de suma importância quanto ao estudo da regulação gênica, é o conhecimento da região promotora de um dado gene.

Um gene, em bactérias, é composto basicamente por uma região codificadora, por uma região promotora e por outras regiões, também importantes. A região codificadora é bem conhecida, mas sobre a região promotora não é trivial indicar com precisão a sua localização. Em contraste com regiões expressas de seqüências de DNA, cuja função se torna aparente quando são traduzidas para proteínas ou para outras seqüências de DNA, a função de um promotor é dada diretamente pela sua seqüência de nucleotídeos. Além disso, as seqüências que definem os vários promotores de um organismo podem apresentar variação relativamente pequena entre si, mesmo entre organismos diferentes mas pertencentes a um mesmo grupo. Quando esse fenômeno ocorre diz-se que ele se conserva (ALMEIDA & SETUBAL, 1998).

Com base em promotores procarióticos conhecidos, basicamente, é possível prever três regiões características:

- uma seqüência de 6 bases de nucletídeos (hexâmero) centrada em –35 do sítio inicial de transcrição +1;
- um hexâmero centrado a –10;
- e a região que separa os dois hexâmeros (distância).

Nos casos conhecidos, a principal característica do promotor consiste no fato de que cada posição de ambos os hexâmeros obedece a algum tipo de conservação. Por outro lado, o tamanho (número de bases) da região entre eles, que é variável, é relevante, enquanto que as suas bases parecem não ter qualquer conservação (LEWIN, 2000).

Ainda que os promotores sejam estruturas de importância indiscutível, a habilidade em identificá-los é menos desenvolvida comparada à de regiões codificantes de um gene. Isso acontece porque os promotores são muitos divergentes e, até os seus padrões mais característicos, como os hexâmeros centrados nos sítios –35 e –10, nem sempre são conservados (QIU, 2003).

Apesar do crescente avanço das técnicas experimentais em biologia molecular, caracterizar e identificar um número considerável de promotores, presentes em um dado genoma, continua sendo uma tarefa demorada e cara.

Abordagens ***in silico*** são bastante utilizadas para reconhecer essas regiões em procariotos. Entretanto, além do alto número de falsos positivos obtidos, elas enfrentam a inexistência de um número adequado de promotores conhecidos para identificar ***in silico*** padrões conservados entre as espécies. Logo, um método criterioso e confiável para predizê-los em qualquer organismo procariótico ainda é um desafio (ALMEIDA & SETUBAL, 1998; MONTEIRO, 2005; REIS, 2005).

Um consenso entre a maioria dos trabalhos encontrados na literatura versa sobre a falta de informação comprovada experimentalmente sobre promotores para se ter certeza de que os modelos computacionais aplicados são realmente funcionais (ALMEIDA & SETUBAL,1998; MONTEIRO, 2005; REIS, 2005).

Mesmo assim, o uso de ferramentas computacionais tem se mostrado importante para inferir funções e estruturas de seqüências e proteínas regulatórias (REIS, 2005). Especificamente para promotores de procariotos, os algoritmos[1] encontrados na literatura se enquadram em três abordagens:

- Baseada em sinal: que opera no reconhecimento de sinais relativamente conservados, assim como de distâncias entre esses elementos. Como exemplo, tem-se o conhecimento de que alguns genes de bactérias fitopatogênicas possuem uma seqüência consenso de nucleotídeos (TTCGC...N15...TTCGC) denominada “PIP-Box” (**P**lant-**I**nducible-**P**romoter-**B**ox”), localizada na região promotora e responsável pela ativação da expressão de fatores de patogenicidade e virulência quando o patógeno entra em contato com a planta hospedeira (FENSELAU & BONAS, 1995).
- Baseada em conteúdo: que utiliza as diferenças do conteúdo das seqüências para classificá-las em promotor ou não-promotor. Por

[1] Uma seqüência finita de passos ou operações bem-definidas que objetiva resolver determinado problema.

exemplo, preferência de códons para codificação de aminoácidos na região próxima ao sítio +1;

- Aprendizado de Máquina (AM): que usam um conjunto de informações estruturais e funcionais disponíveis sobre os promotores para "aprender" automaticamente a reconhecê-los, e produzir hipóteses relevantes sobre os mesmos (BALDI & BRUNAK, 2001).

O uso de técnicas de AM em biologia molecular computacional é bem difundido. Dentre estas técnicas, os Modelos Ocultos de Markov ***(***"Hidden Markov Models" **–** HMM***)*** constituem a mais utilizada atualmente e que tem dado melhores resultados no estudo de promotores. Essa preferência está firmada na hipótese de que regiões características de promotores, relevantes para que a RNA polimerase se direcione corretamente ao sítio +1, devem se apresentar conservadas entre os promotores de um genoma ou até mesmo entre promotores de genomas de organismos próximos evolutivamente (OPPON, 2000). Um HMM descreve uma série de observações através de um processo estocástico[1] oculto, tendo uma seqüência de DNA como uma série de símbolos observados e as posições do promotor como a componente oculta do modelo. Com HMM é possível tratar de forma probabilística a variação estrutural dos elementos de uma mesma classe biológica, apontando regiões conservadas em seqüências envolvidas na regulação da expressão gênica de organismos procarióticos. Como os promotores podem ser tratados como seqüências de nucleotídeos que apresentam bases com diferentes graus de conservação em cada posição, o HMM, o qual pode ser definido como um autômato estocástico de estados finitos, se mostra adequado frente à sua capacidade de capturar regularidades em seqüências de caracteres, considerando a variação nos símbolos observados em cada estado (REIS, 2005).

REIS (2005) propôs um protocolo para predizer e reconhecer promotores utilizando um HMM e o testou em quatro organismos procarióticos: ***E. coli, Bacillus subtilis, Helicobacter pylori e Helicobacter hepaticus***. Para ***E. coli*** houve redução de quase 50% da taxa de erros com relação a trabalhos anteriores; para ***B. subtilis***, o

[1] Estocástico: padrões que surgem através de eventos aleatórios. Ex.: lançar dados tem um resultado estocástico.

protocolo se mostrou muito satisfatório atingindo taxas de reconhecimento e predição altos (95% e 78%, respectivamente). Já com as espécies do gênero ***Helicobacter***, o protocolo teve uma baixa capacidade de predizer promotores, gerando um alto número de falsos positivos.

ALMEIDA & SETUBAL (1998) desenvolveram uma solução para predição de promotores gênicos em procariotos, utilizando a combinação de um HMM com o algorítmo "Expectation Maximization" (EM), para o reconhecimento de regiões promotoras do organismo modelo ***E. coli***, comparando-os com um banco de promotores desta bactéria previamente identificados e estudados por outros grupos de pesquisa. Em sua pesquisa, obtiveram aproximadamente 78% a 83,3% de desempenho, em termos do número de segmentos conservados (hexâmeros) encontrados corretamente. Esses mesmos autores também citam o uso de HMMs em outros trabalhos científicos como, por exemplo, na construção de alinhamentos múltiplos de seqüências, identificação de ***introns***, ***exons***, sítios de junção e sítios de ligação.

Como os métodos de análise e reconhecimento de padrões ainda são objeto de pesquisa ativa, existem muitas metodologias diferentes para a descoberta de padrões e a construção de perfis adequados para as análises.

Os métodos computacionais são uma poderosa ferramenta para a resolução de problemas biológicos. Como visto, pode-se determinar a seqüência correta das bases em um dado gene, como também pode-se determinar a localização exata em um dado genoma. Por outro lado, na maioria dos casos, as soluções apontadas por meio da Bioinformática carece de uma comprovação biológica para validar futuras predições.

III. MATERIAL E MÉTODOS

O presente trabalho foi desenvolvido no Laboratório de Bioquímica e Biologia Molecular (LBM) do Departamento de Tecnologia da Faculdade de Ciências Agrárias e Veterinárias - UNESP - Campus de Jaboticabal - SP.

Os experimentos foram realizados com biologia molecular computacional, ou bioinformática, utilizando-se os resultados experimentais gerados tanto pelo genoma (DA SILVA ***et al.,*** 2002) quanto pelo perfil proteômico da Xac (FACINCANI, 2007).

A. Experimentos com Biologia Computacional

Todos os códigos de programas foram desenvolvidos em linguagem de programação "Perl" ("Practical Extraction and Reporting Language": http://www.perl.org/), em ambiente de sistema operacional Linux ("RedHat Enterprise Linux": http://www.redhat.com/), que se apresenta como a melhor linguagem de programação para detectar padrões em grande quantidade de dados, por uma questão de eficiência do ponto de vista do programador (GIBAS & JAMBECK, 2001). Todos os programas desenvolvidos no presente trabalho, encontram-se no Apêndice.

1. Mapeamento das Regiões Intergênicas "Upstream"

Primeiramente, foi desenvolvido um programa (upstream_intergenic_mapping.pl) para gerar uma tabela com os dados gerais das regiões intergênicas (regiões que não codificam genes) na porção "upstream[1]" às ORFs. Este programa, assim como todos os que se seguem, foram aplicados às três estruturas presentes em Xac, separadamente: cromossomo e plasmídeos A e B. Os dados informados a este programa foram:

[1] A montante.

- Arquivo texto contendo dados do genoma da ***Xac*** (http://genoma4.fcav.unesp.br/xanthomonas/). Foram recuperadas informações das duas fitas de DNA (“forward” e “reverse”), referentes à identificação das ORFs e a localização exata destas no genoma.
- Arquivo FASTA[1] completo da estrutura.
- Presença de circularidade da estrutura (as três estruturas de Xac são circulares).

O programa foi desenvolvido com o objetivo de gerar como saída um arquivo formato texto com as informações: identificação da ORF, direção (a que fita pertence: “forward” ou “reverse”), posição inicial e final da ORF em relação ao FASTA completo da estrutura, grupo de tamanho da Região Intergênica “Upstream” (RIU), tamanho em quantidade de bases de nucleotídeo e respectiva seqüência FASTA da RIU.

A seqüência FASTA da RIU e seu tamanho em bases foram identificados à partir do cálculo da distância entre o final de uma dada ORF e o início da próxima, ambas de uma mesma fita. Como há circularidade nas três estruturas de Xac, o cálculo da distância da RIU da primeira ORF do FASTA genoma foi definida como sendo a soma do tamanho da porção entre o final da última ORF do FASTA do genoma até o final deste FASTA, mais a porção da primeira base do FASTA do genoma até o início da primeira ORF.

Um tipo de categorização de tamanhos de RIU foi aplicada, dividindo-se em grupos de tamanho de RIU. Este processo foi necessário, pois é importante para se decidir qual porção da RIU será a seqüência de observação para o mapeamento de promotores. Todas as RIU foram categorizadas em 5 tamanhos diferentes: “<=30” para as RIU de tamanho menor ou igual a 30 bases (b) ; “~50” para RIU com tamanho entre 31 e 75 b; “~100” para RIU entre 75 e 150 b; “~200” para intervalos entre 151 e 250 b; e “~300” para RIU maiores que 251 b.

Segundo XIONG (2006), as regiões promotoras em procariotos, em geral, ficam no máximo a aproximadamente –200 bases “upstream” à região codificadora da

[1] Arquivo em formato texto que contém seqüências de bases de nucleotídeos ou aminácidos.

proteína. Como não se sabe exatamente onde está situado o início do sítio de transcrição do gene, definiu-se neste trabalho que as seqüências de observação das regiões intergênicas a se trabalhar seriam com até -350 b “upstream” da posição +1 da ORF, descartando-se da análise a porção excedente mais à esquerda da RIU, acima de 350 b (Figura 1).

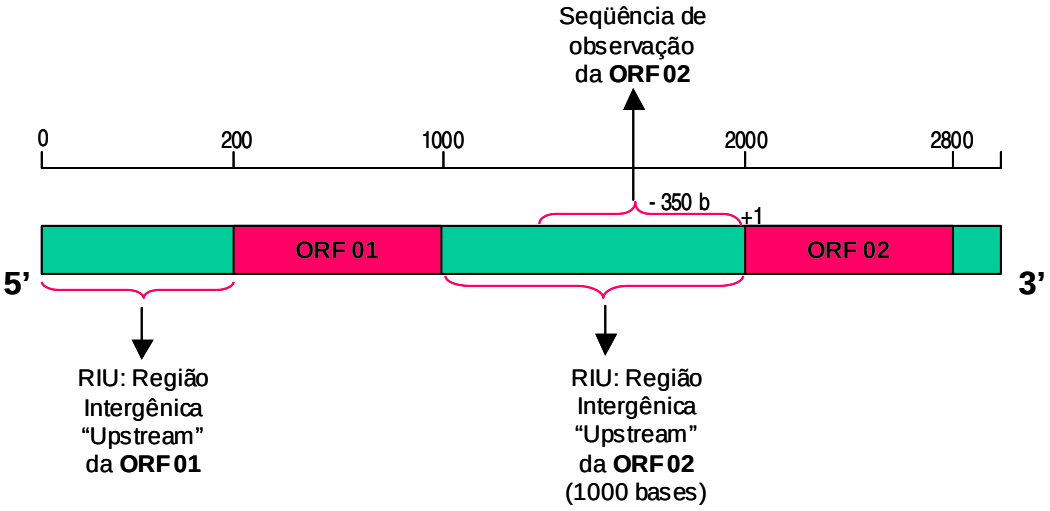

Figura 1: Esquema geral sobre a localização da seqüência de observação em uma região intergênica “upstream” de uma ORF qualquer.

Um dos aspectos para se predizer um promotor procariótico é a determinação das estruturas de ***operon.*** Os genes que se agrupam nestas estruturas compartilham um mesmo promotor, localizado na região “upstream” do primeiro gene da estrutura deste ***operon***. Porém, essas estruturas não são abordadas neste trabalho. Apenas foi necessário apontar as ORFs com potencial a fazer parte de um ***operon*** e descartá-las do processo de mapeamento dos promotores. Isso foi feito considerando-se apenas o tamanho da RIU encontrada para dada ORF. Utilizando-se o conceito de critérios de WANG e colaboradores (2004), a ORF onde a sua RIU tiver 30 b ou menos será uma forte candidata a fazer parte de um ***operon***. Um gene pode fazer parte desta estrutura quando tem até 300 b na sua RIU. Porém, descartou-se do processo de mapeamento apenas as ORFs com RIU<=30 b, pois, além de serem candidatas a fazerem parte de

operons, tamanhos de RIU próximos a este valor sugerem que a existência de promotores é altamente improvável, pois sabe-se que antes da posição +1 do gene ainda se tem o início do sítio de transcrição. Para todas as regiões maiores do que 30 b houve a tentativa de mapear os promotores, já que quanto maior esta distância, maior será a probabilidade deste gene não fazer parte de uma estrutura de ***operon***.

2. Ferramenta para predição dos promotores

A predição ***in silico*** dos promotores da Xac, foi baseada em um Modelo Oculto de Markov (HMM - “Hidden Markov Models”) e um algoritmo de “Expectation-Maximization” (EM), desenvolvido em linguagem de programação “Perl” por ALMEIDA & SETUBAL (1998). Este modelo procura pelo “melhor” candidato a promotor dentro dos possíveis promotores encontrados em seu processamento para cada seqüência de observação, à partir de um arquivo de entrada com parâmetros de exemplos de observação sobre promotores gênicos. Dentre esses parâmetros, destacam-se a própria seqüência FASTA da porção de interesse da RIU (seqüência de observação), o tamanho em bases, as probabilidades de cada sítio dentro dos hexâmeros e a freqüência dos nucleotídeos do código genético alvo, fora dos hexâmeros. Uma vez que o programa foi desenvolvido para trabalhar individualmente com algumas seqüências, a sua utilização no presente trabalho demandou uma série de adaptações para mapeamento de todo o genoma da Xac. Mais detalhes sobre o modelo propriamente dito pode ser visto em ALMEIDA & SETUBAL (1998).

a) Definição dos parâmetros de observação

Como não se tem todas as informações necessárias da própria Xac para os parâmetros de observação, convencionou-se utilizar os da ***E. coli*** e sempre baseado-se nos valores utilizados no modelo de ALMEIDA & SETUBAL (1998).

Para o tamanho médio da distância entre os hexâmeros, utilizou-se sugerir a média da distância da ***E. coli***, que é de 17 b (ALMEIDA & SETUBAL, 1998). Segundo OLEKHNOVICH & KADNER (1999), 90% dos promotores conhecidos de ***E. coli*** têm distância entre hexâmeros de 16 a 18 b.

Quanto à provável localização do promotor dentro da RIU, decidiu-se adotar que ele pode estar localizado mais ao meio da seqüência de observação, assim como na ***E. coli*** (ALMEIDA & SETUBAL, 1998).

O padrão de conservação dos hexâmeros também foi mantido como sendo o de ***E. coli***: $T_{82}T_{84}G_{78}A_{65}C_{54}A_{45}$ e $T_{80}A_{95}T_{45}A_{60}A_{50}T_{96}$, onde o índice denota o percentual de ocorrência da base encontrada com mais freqüência em cada sítio (ALMEIDA & SETUBAL, 1998).

O tamanho das seqüências de observação e, conseqüentemente, o próprio FASTA da seqüência, foram definidas especialmente para o presente trabalho em função de ter-se trabalhado com o genoma inteiro. Ficou como sendo a própria seqüência RIU nos casos em que estas apresentavam tamanho inferior a 350 b de nucleotídeos. Para as RIU maiores, apenas 350 b mais à direita foram selecionadas. As RIU menores do que 30 b foram descartadas, por representarem regiões de sobreposição de genes (inexistência de RIU) ou genes que podem fazer parte de ***operons*** (RIU quase que inexistente, entre zero e 30 b).

Em seguida, foi necessário calcular os valores da freqüência de nucleotídeos do genoma de Xac, para o modelo de ALMEIDA & SETUBAL (1998). Para o genoma de ***E. coli***, os valores de freqüência são praticamente iguais para as quatro bases de nucleotídeos (~25% cada). Porém, para a Xac, foi desenvolvido um programa (background_fasta.pl), o qual, de acordo com a seqüência FASTA completa do genoma em estudo calcula-se a porcentagem da freqüência de cada base na fita indicada. Esse programa foi aplicado em 6 versões: para o cromossomo e plasmídeos A e B, todos nas duas fitas.

A estimativa desses valores se faz necessária para “calibrar” os dados de entrada para o modelo aplicado. Esta “calibragem” se dá através do preenchimento de um arquivo de parâmetros de probabilidades, onde é informada a freqüência de

nucleotídeos do código genético do organismo, o tamanho médio da distância entre os hexâmeros e os valores correspondes às probabilidades iniciais dos estados à partir de onde o modelo pode começar a procurar o hexâmero –35. O que melhor se adaptou ao organismo foi a probabilidade de 90% do algorítmo continuar no estado chamado de ***before,*** antes de sugerir o primeiro promotor, o que o faz procurar por mais tempo por um padrão ótimo até o mais próximo possível do gene, mesmo em seqüências de observação maiores, como as de 350 b.

À partir destes parâmetros, foi necessário desenvolver um programa (make_input_probabilities_FILES.pl) cujo objetivo foi gerar um arquivo independente para cada ORF com RIU passível de estudo, automaticamente, com sua própria seqüência de observação e com os parâmetros das probabilidades do programa do HMM, de acordo com a direção do gene em relação ao genoma. A Figura 2 exemplifica um arquivo de parâmetros de observação de uma ORF, o qual é lido pelo programa durante a busca do promotor.

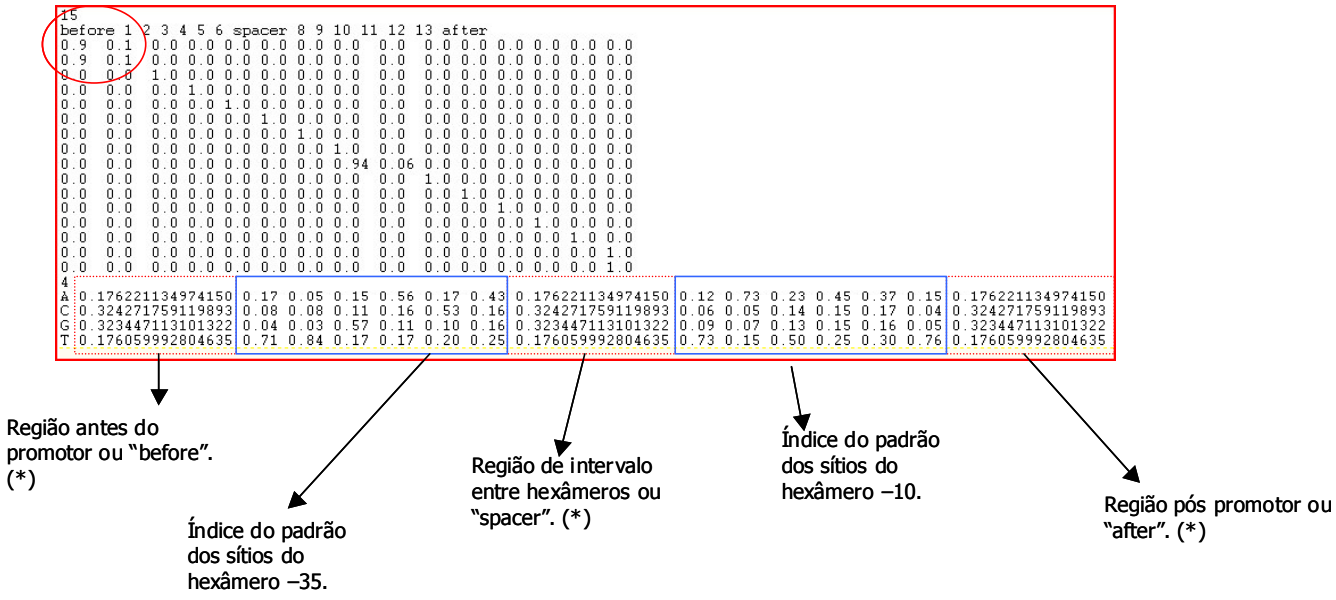

Figura 2: Exemplo de arquivo texto de parâmetros de observação. Logo abaixo destes valores e ainda no mesmo arquivo, é incluída a seqüência de observação FASTA. (*) Os índices do estados "before", "spacer" e "after" são da freqüência de todo o genoma.

b) Aplicação do modelo de predição de promotores

O programa do modelo de ALMEIDA & SETUBAL (1998) foi executado em lote (tata2HMM_for_all.pl), ou seja, o modelo foi aplicado automaticamente a todas as ORFs cuja RIU foi considerada passível de estudo quanto a promotores. E os resultados (arquivos gerados) foram gravados individualmente para cada uma delas (com mapeamento do provável promotor incluído), no mesmo formato do programa original dos autores. A Figura 3 exemplifica um arquivo de saída com o melhor promotor encontrado e mapeado.

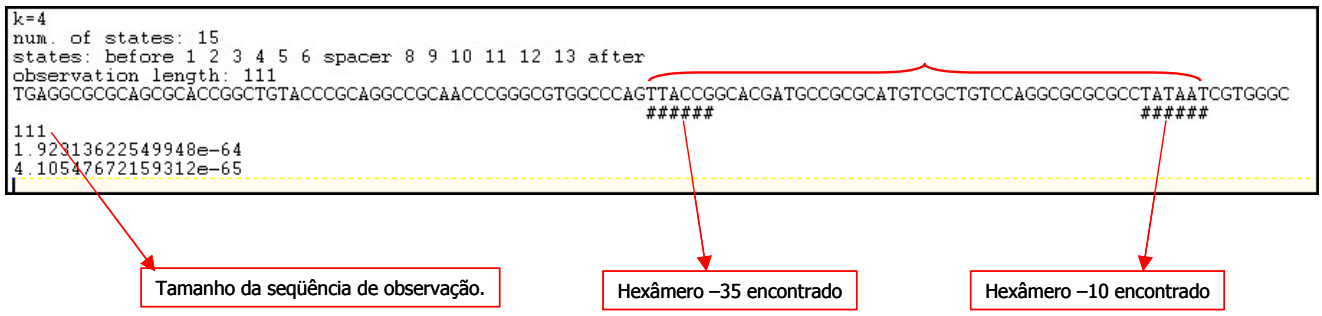

Figura 3: Exemplo de arquivo texto de saída, resultando no melhor promotor encontrado. Os hexâmeros encontrados delimitam a região do promotor.

c) Organização dos dados gerados dos promotores mapeados

Finalmente, à partir de todos os arquivos de saída gerados para o genoma, um único relatório para cada fita analisada de cada estrutura do código genético foi gerado, à partir de dois outros programas desenvolvidos neste trabalho (make_final_report.pl e make_mix_with_data_annotation.pl), contendo:

- todas as informações finais sobre o mapeamento dos promotores identificados: tamanho do promotor, tamanho da distância entre hexâmeros, posição do hexâmero -35 e do -10 quanto à ORF, seqüência FASTA completa do promotor encontrado.
- todas as informações das RIU mapeadas.

- todas as informações sobre as ORFs que não entraram nestas análises.
- informações da própria ORF, como a sua localização no genoma, o seu produto/proteína e a categoria a qual pertence.

Ao todo, foram 6 relatórios: cromossomo, plasmídeo A e plasmídeo B, todos os três nas fitas “forward” e “reverse” correpondentes.

3. Comparação com dados experimentais

a) Identificação de prováveis promotores

No trabalho de FACINCANI (2007), proteínas expressas diferencialmente em quatro tempos diferentes da Xac foi apresentada: uma sob condição não infectante (***in vitro***) e 3 em condições infectantes (***in vitro*** e ***in vivo).*** FACINCANI (2007) observou a presença de 1162 proteínas com expressão presente no tempo 0 ou condição não infectante; 1167 proteínas com expressão presente no tempo 1, cultivada em condição infectante (meio de cultura XAM1) por 24 horas; 1157 proteínas com expressão presente no tempo 2, 3 dias pós-inoculação (d.p.i.) ***in planta***; 1072 proteínas com expressão presente no tempo 3, 5 dias pós-inoculação ***in planta***. Um total de 1661 proteínas distintas de Xac foram identificadas por MudPIT. Em comparação com a previsão teórica das proteínas codificadas pelo genoma da Xac, cerca de 37% das proteínas foram identificadas. Esses dados foram comparados com todos os promotores preditos ***in silico*** neste trabalho.

Tabelas foram produzidas com os dados finais do seqüenciamento das proteínas dos genes expressos, oriundos do experimento de FACINCANI (2007), dividindo-se nos quatro tempos, como tempo 0, 1 (1 d.p.i.), 2 (3 d.p.i) e 3 (5 d.p.i.). Um programa (cross_experimental_data.pl) foi desenvolvido para que os dados finais das análises dos promotores fossem mesclados com estes resultados de expressão. As ORFs que tiveram o promotor mapeado ***in silico*** corretamente e apresentaram expressão de

proteína presente ao menos em um dos tempos estudados por FACINCANI (2007), foram aqui chamadas de “ORFs com prováveis promotores identificados”. Os dados foram importados para uma tabela em uma planilha eletrônica BrCalc, que faz parte do pacote de software BrOffice.org 2.4 criado por Sun Microsystems Inc, com base em OpenOffice.org (http://www.openoffice.org/welcome/credits.html).

b) Análise estrutural dos prováveis promotores de Xac

(1) Distância entre hexâmeros

As ORFs foram divididas nos 4 grupos de tamanho de RIU estudadas (~50, ~100, ~200 e ~300 bases). Medidas estatísticas foram calculadas para cada grupo, com relação ao tamanho da distância entre os hexâmeros***:*** média, mediana, desvio-padrão e coeficiente de variação (“Pearson”), utilizando-se as funções estatísticas disponíveis na planilha eletrônica BrCalc.

(2) Padrão de conservação dos hexâmeros

Foi calculado o padrão de conservação dos hexâmeros -35 e -10 dos prováveis promotores identificados, apontando o percentual de ocorrência da base de cada sítio dos hexâmeros. Para isso foi utilizado o software TextPad 4.7.3 criado por Helios Software Solutions (http://www.textpad.com/). Os dados foram organizados em uma tabela.

Gráficos de Logo foram criados para obter-se representações qualitativas e quantitativas da conservação dos nucleotídeos contidos nas seqüências dos hexâmeros, através da técnica de logos de seqüências, que resume estas informações graficamente. Os gráficos foram gerados através da ferramenta disponível no endereço de internet http://weblogo.berkeley.edu/ (CROOKS ***et al.,*** 2004). A representação é feita através de letras (A, T, C e G), com a maior localizada no topo da pilha. A altura de cada letra é proporcional à freqüência do nucleotídeo que ela representa, tendo a

visualização da quantidade de informação contida em dado sítio.

c) Outras análises

O programa ClustalX (THOMPSON ***et al.***, 1994) foi utilizado para fazer alinhamento de seqüências candidatas a promotoras. Buscou-se identificar os sítios conservados entre algumas espécies de ***Xanthomonas***.

Pesquisas no banco de dados do BLAST do NCBI (“National Center for Biotechnology Information” - http://www.ncbi.nlm.nih.gov/) foram feitas em julho de 2008, utilizando-se a ferramenta BLAST (“Basic Local Alignment Search Tool”), com os algorítmos TblastX e BlastN (ALTSCHUL ***et al.***, 1997), para se buscar seqüências do presente trabalho em bancos de dados de proteínas e de nucleotídeos.

IV. RESULTADOS E DISCUSSÃO

A. Estudos das Regiões Intergênicas “Upstream” às ORFs de Xac

O mapeamento das regiões intergênicas mostrou coerência no genoma do cromossomo da Xac, pois existe um equilíbrio tanto na quantidade de ORFs mapeadas em cada fita bem como na distribuição de tamanhos em bases de nucleotídeos da região intergênica “upstream” (RIU) em relação ao número de ORFs. Já o mesmo não foi observado nos plasmídeos A e B, onde há uma enorme variação nestes valores, fato esperado em função de haver diversos genes de patogenicidade, muito provavelmente adquiridos por meio de transferência lateral (Tabela 3).

Tabela 3: Quantidade de ORFs de Xac, segundo o tamanho em bases de nucleotídeos das Regiões Intergênicas “Upstream” (RIU), presentes no cromossomo e nos plasmídeos.

Tamanho da RIU em bases de nucleotídeos	Número de ORFs por Fita					
	Cromossomo		Plasmídeo A		Plasmídeo B	
	“Forward”	“Reverse”	“Forward”	“Reverse”	“Forward”	“Reverse”
RIU < = 30	579	576	7	5	4	27
31 <= RIU < = 50	93	94	1	1	1	0
51 <= RIU < = 100	214	204	1	3	1	3
101 <= RIU < = 200	288	260	1	2	0	6
RIU >= 201	1034	1032	10	11	11	20
Total de ORFs	2208	2166	20	22	17	56

Observando as RIU do cromossomo, das 4374 ORFs conhecidas, 3219 apresentaram RIU passíveis de serem estudadas, pois todas estas regiões são maiores que 30 bases de nucleotídeos. De todo o cromossomo, 1629 ORFs são da fita de direção “forward” e 1590 da fita de direção “reverse”.

Já nos plasmídeos, 42 das 73 ORFs do plasmídeo B e 30 das 42 ORFs do plasmídeo A, tiveram RIU consideradas passíveis de estudo.

B. Freqüência de nucleotídeos quanto ao genoma de Xac

Uma das principais características na predição ***in silico*** de promotores é a busca por padrões nos sítios do hexâmeros. Porém, é importante identificar o que acontece fora dos hexâmeros, ou seja, antes do promotor ou na região entre os hexâmetros ou depois do promotor, pois durante a aplicação do modelo HMM é necessário diferenciar o que seria um promotor do restante do genoma do organismo em estudo. A Tabela 4 mostra a freqüência com que cada um dos quatro nucleotídeos aparece em relação ao total de nucleotídeos presente no genoma de todo o organismo de Xac.

O cálculo das freqüências de nucleotídeos de Xac, mostrou que o genoma é mais rico em G e C e que os valores são muito próximos quando se observa cada nucleotídeo independente entre as três estruturas (cromossomo e plasmídeos A e B).

Tabela 4: Porcentagem de freqüência de cada nucleotídeo no genoma de Xac.

Fita	Nucleotídeo	Cromossomo	Plasmídeo A	Plasmídeo B
"Forward"	A	17,62%	19,13%	18,81%
	C	32,43%	30,87%	30,54%
	G	32,34%	30,99%	30,85%
	T	17,61%	19,01%	19,80%
"Reverse"	A	17,61%	19,01%	19,80%
	C	32,34%	30,99%	30,85%
	G	32,43%	30,87%	30,54%
	T	17,62%	19,13%	18,81%

C. Mapeamento dos promotores

Das 3219 ORFs com RIU passíveis de estudo, 552 têm RIU entre 31 e 92 bases e somente nesse intervalo é que observou-se que, em algumas seqüências de observação, não foi possível o mapeamento de promotor ou foi mapeado somente um hexâmero, o que o invalida. Um total de 229 das 552 ORFs deste intervalo não tiveram um promotor mapeado, ou seja, cerca de 41%. Todas as seqüências de observação maiores do que 93 b tiveram um promotor mapeado. Do total de promotores mapeados, 1516 são da fita de direção “forward” e 1474 da fita de direção “reverse”.

Já nos plasmídeos, 9 das 72 ORFs apontadas como contendo RIU passíveis de estudo, não tiveram um promotor mapeado. As seqüências de observação onde não foram encontrados promotores nos plasmídeos são menores do que 128 b. Um total de 24 ORFs do plasmídeo A e 39 ORFs do plasmídeo B, tiveram promotores mapeados corretamente.

D. Análise estrutural dos prováveis promotores

Os candidatos a promotores das ORFs do cromossomo e dos plasmídeos, com expressão de proteína presente (MudPIT) e com um promotor mapeado ***in silico*** corretamente (1192), foram analisados quanto à sua estrutura.

As ORFs foram divididas nos 4 grupos de tamanho de RIU estudadas (~50, ~100, ~200 e ~300 bases). Os dados foram organizados na Tabela 5.

1. Distância entre hexâmeros

Em relação à distância entre os hexâmeros, as ORFs com RIU aproximada de 50 b apresentaram, em média, 14 b no sentido “forward” e 12 b no sentivo “reverse”, para o cromossomo. Para as ORFs com RIU aproximada de 100 b, a média no sentido “forward” foi de 23 b e no “reverse” foi de 20 b, no cromossomo. Nos plasmídeos A e B, as ORFs desse grupo apresentaram média de 16 b. Para as ORFs com RIU

aproximada de 200 b, a média no sentido “forward” foi de 22 b e no “reverse” foi de 17 b, para o cromossomo. Nos plasmídeos esta média ficou em 13 b. Já no grupo de aproximadamente 300 b, a média da distância entre hexâmeros ficou em 18 b na direção “forward” e 20 b na direção “reverse”, para o cromossomo. E 7 b nos plasmídeos.

No cromossomo, tanto a média como a mediana de 3 dos 4 grupos (~100, ~200 e ~300 b), nas duas fitas, apontaram para uma distância entre hexâmeros dentro dos padrões de promotores conhecidos de ***E. coli***, ou seja, entre 17 e 23 bases, apesar do desvio-padrão e do coeficiente de variação mostrarem que há uma grande variação dentro dessa lista de valores. Para os plasmídeos, a quantidade de prováveis promotores foi pequena para se deduzir algo à partir de medidas estatísticas. Porém, tanto a média das distâncias entre os hexâmeros dos grupos de ~100 e ~200 b quanto a mediana do grupo de ~200 b são comparáveis aos padrões estabelecidos para ***E. coli***.

Tabela 5: Análise da distância entre os hexâmeros dos 1176 prováveis promotores identificados do cromossomo e 16 dos plasmídeos.

Análise da distância entre os hexâmeros –35 e –10				
Tamanho da RIU*	Medidas Estatísticas	Cromossomo		Plasmideos A e B
		"Forward"	"Reverse"	"Forward / Reverse"
~50 b	Total de Genes	46	36	0
	Média dos Tamanhos	14	12	-
	Mediana	11	10	-
	Desvio-padrão	11,57	9	-
	Coeficiente de Variação ("Pearson")	82,78%	77,62%	-
~100 b	Total de Genes	99	94	4
	Média dos Tamanhos	23	20	16
	Mediana	19	14	8
	Desvio-padrão	19,25	17	20
	Coeficiente de Variação ("Pearson")	83,30%	87,48%	128,22%
~200 b	Total de Genes	106	77	2
	Média dos Tamanhos	22	17	13
	Mediana	17	12	13
	Desvio-padrão	18,21	18	1
	Coeficiente de Variação ("Pearson")	84,64%	104,77%	10,88%
~300 b	Total de Genes	355	363	10
	Média dos Tamanhos	18	20	7
	Mediana	12	14	7
	Desvio-padrão	16,67	17	6
	Coeficiente de Variação ("Pearson")	95,20%	87,48%	90,10%

*RIU: Região Intergênica "Upstream"

2. Padrão de conservação dos hexâmeros

Quanto aos hexâmeros desses prováveis promotores de Xac, o padrão de conservação das bases neles encontradas foi $\boldsymbol{T_{64}T_{76}G_{72}A_{39}C_{58}A_{32}}$ e $\boldsymbol{T_{53}A_{65}T_{42}A_{36}A_{38}T_{80}}$, onde o índice denota o percentual de ocorrência da base encontrada com mais freqüência. As bases encontradas em cada posição dos hexâmeros desse padrão mostraram-se iguais às de ***E. coli***, porém diferem nos índices. Todas as freqüências podem ser observadas na Tabela 6.

Tabela 6: Freqüência de nucleotídeos encontrada nos hexâmeros dos prováveis promotores de Xac (1176 do cromossomo e 16 dos plasmídeos).

	Hexâmero -35						Hexâmero –10					
Quantidade de nucleotídeos encontradas em cada sítio												
Nucleotídeo	1°	2°	3°	4°	5°	6°	1°	2°	3°	4°	5°	6°
A	260	78	114	464	188	380	183	780	234	435	454	156
C	114	162	124	324	693	291	140	67	266	292	218	37
G	55	48	859	248	135	336	233	151	188	264	223	48
T	763	904	95	156	176	185	636	194	504	200	297	951
Freqüência dos nucleotídeos em cada sítio												
A	22%	7%	10%	39%	16%	32%	15%	65%	20%	36%	38%	13%
C	10%	14%	10%	27%	58%	24%	12%	6%	22%	24%	18%	3%
G	5%	4%	72%	21%	11%	28%	20%	13%	16%	22%	19%	4%
T	64%	76%	8%	13%	15%	16%	53%	16%	42%	17%	25%	80%

Este padrão de conservação dos hexâmeros, assim como nos promotores estudados de ***E. coli***, não ocorre exatamente igual em nenhum provável promotor identificado para Xac, mas sim variações do mesmo.

As Figuras de 4 a 7, são representações qualitativas e quantitativas da conservação dos nucleotídeos contidos nas seqüências dos hexâmeros centrados em torno de –35 e –10 bases do início do sítio de transcrição, através da técnica de logos de seqüências, que resume estas informações graficamente. A altura de cada letra é proporcional à freqüência do nucleotídeo que ela representa, tendo a visualização da quantidade de informação contida em dado sítio. Quanto ao padrão dos hexâmeros, observa-se que há uma maior freqüência de 4 das 6 bases aparentemente conservadas no hexâmero –35 (posições 1, 2, 3 e 5) e de 3 bases no hexâmero –10 (posições 1, 2 e 6), tanto na fita “forward” como na “reverse”. No hexâmero –10, também conhecido como TATA-Box, o consenso da seqüência modelo é muito difundido como “TATAAT”. Porém, em ***E. coli***, segundo SCHNEIDER (2001), as bases conservadas no TATA-Box são segregadas em duas partes: primeira e segunda posições (“TA”), na extremidade 5´, e posição 6 (“T”), na extremidade 3´. Este fato também foi observado em Xac, conforme Figuras 6 e 7. Também observou-se em Xac a base “T” altamente

conservada na extremidade 3´ do TATA-Box (Figuras 6 e 7), o que também é observado por SCHNEIDER (2001) em ***E. coli***. Isso é justificado por estudos experimentais que apontam este sítio como pertencente ao primeiro pareamento rompido na dupla fita de DNA da ***E. coli*** durante o início da transcrição (SCHNEIDER, 2001).

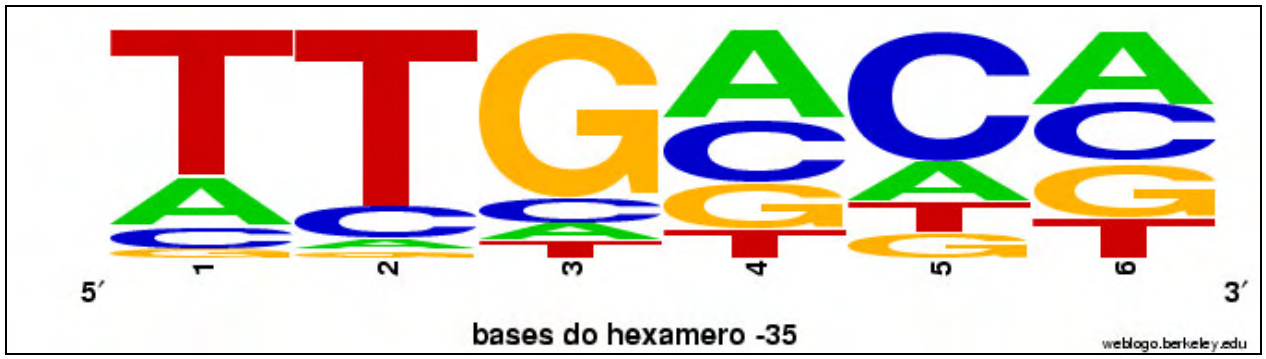

Figura 4: Logo com alinhamento das seqüências dos hexâmeros –35 da fita “forward” dos prováveis promotores.

Figura 5: Logo com alinhamento das seqüências dos hexâmeros –35 da fita “reverse” dos prováveis promotores.

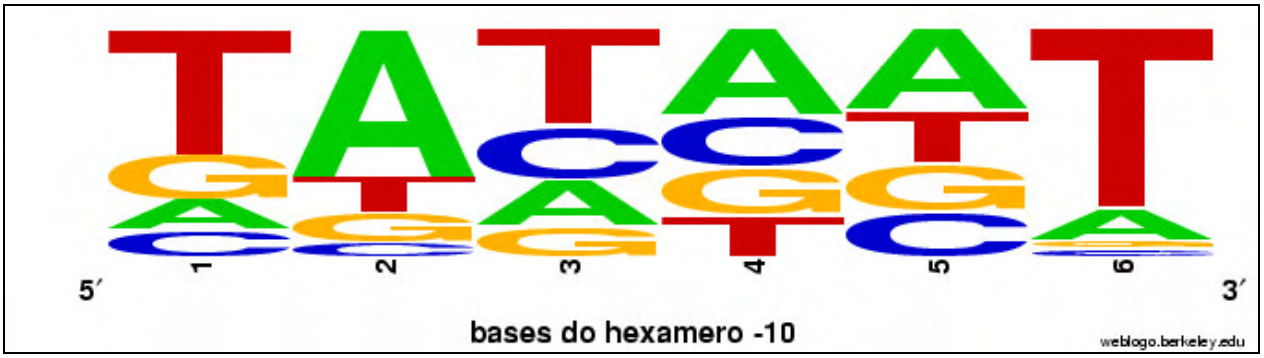

Figura 6: Logo com alinhamento das seqüências dos hexâmeros –10 da fita “forward”, dos prováveis promotores.

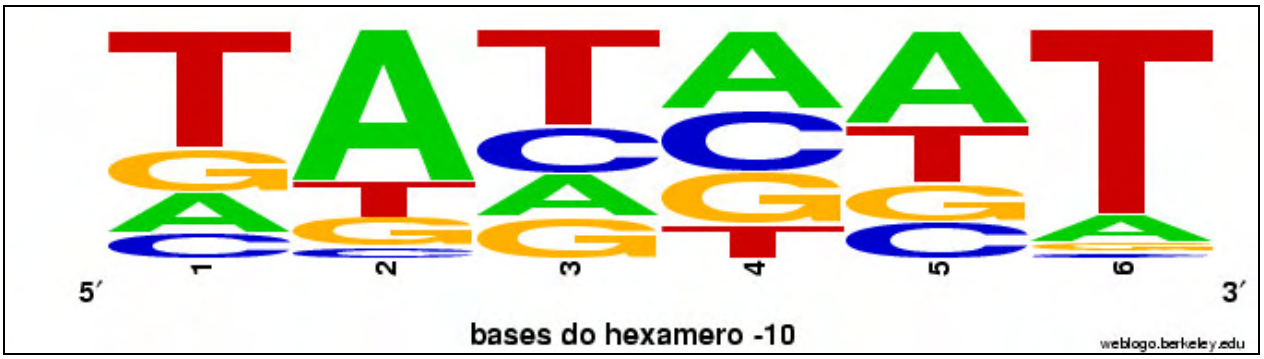

Figura 7: Logo com alinhamento das seqüências dos hexâmeros –10 da fita “reverse”, dos prováveis promotores.

E. Análise comparativa dos prováveis promotores identificados com os dados experimentais do perfil proteômico de Xac.

As ORFs que tiveram em sua RIU uma seqüência completa de promotor mapeada ***in silico*** foram comparadas com os resultados do experimento do perfil proteômico de Xac, de FACINCANI (2007). Aqueles que também deram resultado positivo quanto à expressão de sua proteína por MudPIT foram classificados como “ORFs com um provável promotor identificado” (Tabela 7).

Tabela 7: Comparação das ORFs com promotor mapeado na sua RIU* ***versus*** genes com expressão de proteína presente através do experimento de MudPIT de FACINCANI (2007). A proteína expressa mais o promotor mapeado para dada ORF, classifica-a como "ORF com Provável Promotor Identificado".

	ORFs no genoma de Xac	Cromossomo		Plasmídeo A		Plasmídeo B	
		"Forward"	"Reverse"	"Forward"	"Reverse"	"Forward"	"Reverse"
Proteínas Expressas	37%	833	799	3	8	4	14
ORFs com promotor mapeado	68%	1516	1474	12	12	13	26
Proteína expressa + promotor mapeado	27%	606	570	1	5	4	6

*RIU: Região Intergênica "Upstream"

A Tabela 8 mostra a distribuição das categorias primárias, segundo a anotação do genoma de Xac (DA SILVA ***et al***., 2002), às quais pertencem as ORFs cuja região promotora pôde ser provavelmente identificada.

Tabela 8: Distribuição das ORFs com provável promotor identificado, nas categorias primárias da anotação do genoma de Xac (DA SILVA ***et al***., 2002).

Categorias do Genoma da Xac	N° de ORFs com Provável Promotor Identificado					
	Cromossomo	% *	Plasmídeos	% *	Total	% *
I. Metabolismo Intermediário	242	33%	0	0	242	33%
II. Biosíntese de Pequenas Moléculas	148	42%	0	0	148	42%
III. Metabolismo de Macromolécula	215	39%	0	0	215	39%
IV. Estrutura Celular	58	29%	1	100%	59	29%
V. Processos Celulares	124	32%	0	0	124	32%
VI. Elementos Genéticos Móveis	10	7%	8	19%	18	9%
VII. Patogenicidade, Virulência e Adaptação	86	30%	4	24%	90	30%
VIII. Hipotéticas	264	16%	3	5%	267	16%
IX. ORFs com Categoria Indefinida	29	27%	0	0	29	27%

* % em relação ao número de ORFs da anotação do genoma de Xac.

Na Figura 8, observa-se a comparação de ORFs pertencentes às 9 categorias funcionais nos projetos genoma, proteoma e na indicação de ORFs com prováveis

promotores identificados do presente trabalho, utilizando-se das categorias pré-definidas no genoma da bactéria por DA SILVA e colaboradores (2002). Pode-se observar que tanto as proteínas expressas do proteoma quanto as ORFs com provável promotor mapeado estão distribuídas por todas as categorias funcionais. Das 1658 ORFs anotadas como pertencentes à categoria de proteínas hipotéticas por DA SILVA e colaboradores (2002) para o genoma de Xac, FACINCANI (2007) observou que apenas 22% apresentaram expressão de proteína por MudPIT e, no presente trabalho, 16% dessas ORFs, além de expressão da proteína por MudPIT, apresentaram também um provável promotor identificado. Isso mostra que a maioria das ORFs anotadas como hipotéticas para o genoma parecem realmente não codificar genes funcionais.

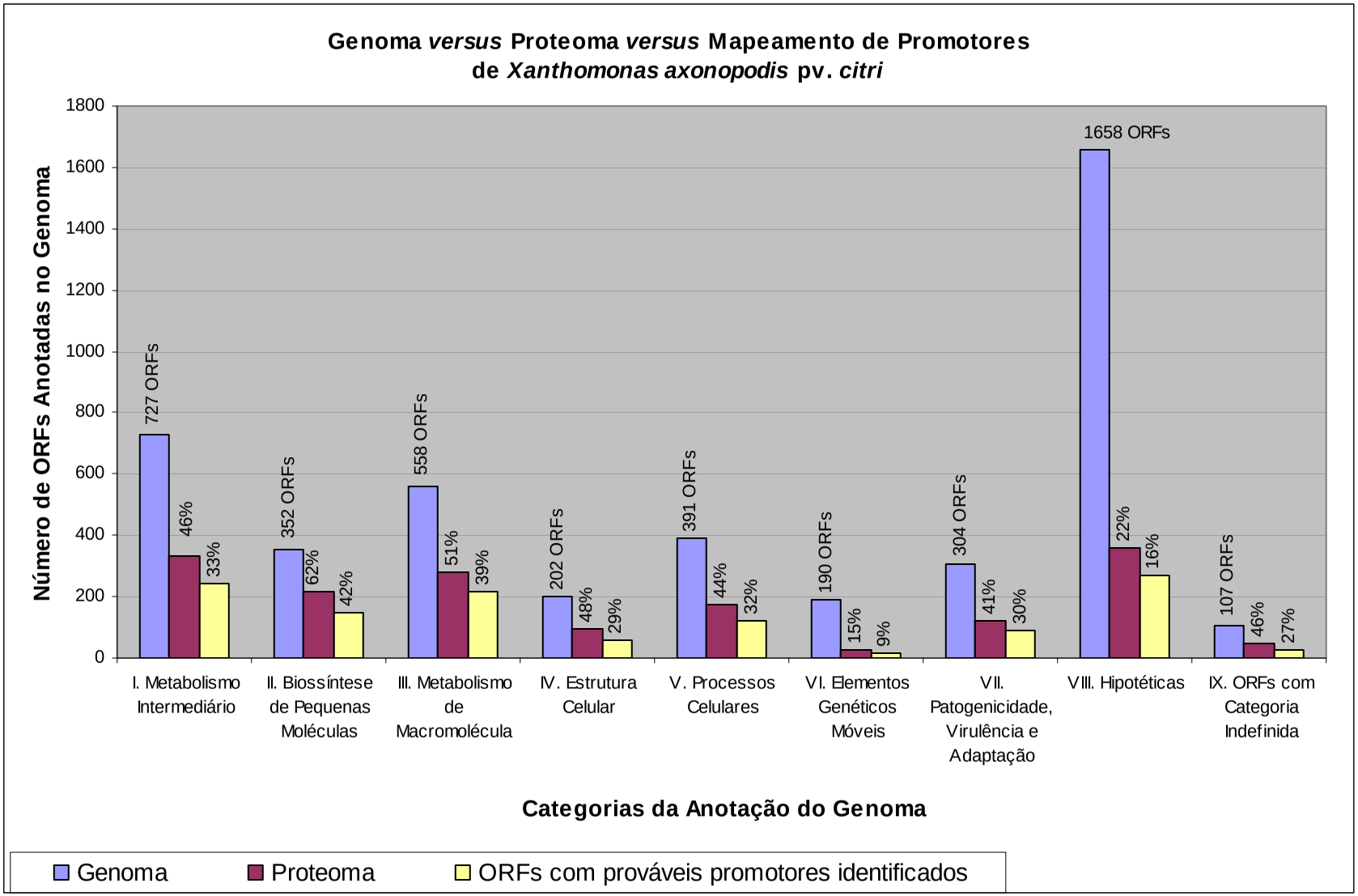

Figura 8: Distribuição de categorias comparando-se resultados do genoma, do proteoma e das ORFs com provável promotor identificado. No topo da barra do “Genoma” está indicada a quantidade de ORFs pertencentes àquela categoria. No topo das barras do “Proteoma” e das “ORFs com provável promotor identificado”, está indicada a porcentagem das ORFs identificadas, relacionadas àquela categoria.

O não mapeamento dos promotores em RIU de tamanhos menores, pode caracterizar a possibilidade destas ORFs fazerem parte de algum ***operon*** ou até serem regulados por outro tipo de promotor (e não os que contém TATA-Box). Pesquisas direcionadas a estudos de ***operons*** em procariotos, inferem que as distâncias intergênicas entre genes que fazem parte de um mesmo ***operon*** tendem a ser muito menores do que entre genes que estão fora dessas estruturas (WANG ***et al***., 2004 ; SALGADO ***et al***., 2000; YADA ***et al***., 1999). Porém, não é somente a distância intergênica que prediz a existência de um ***operon***. Além deste, uma combinação de outros critérios são utilizados, como por exemplo, o mapeamento de “clusters” de genes conservados, a relação funcional dos genes envolvidos, a análise dos elementos da seqüência, evidências experimentais, entre outros (BROUWER ***et al***., 2008).

Comparando-se as ORFs com provável promotor identificado com as ORFs anotadas do genoma (DA SILVA ***et al***., 2002) e expressas no proteoma (FACINCANI, 2007), os resultados permitiram que fosse feito uma ordenação dos dados de expressão por MudPIT em seus quatro tempos proporcionando a visualização da ordenação de alguns conjuntos de genes expressos em razão do tempo (0, 1, 3 e 5 dias, ou seja, tempo 0, 1, 2 e 3, respectivamente), e fazer uma correlação entre esta ordem e o mapeamento ***in silico*** do promotor, sugerindo uma temporalidade na expressão desses genes e indícios de presença de alguns ***operons***.

Na Tabela 9, alguns exemplos desses possíveis ***operons*** foram escolhidos nos dados e organizados, separados por grupos identificados por letras de “A” até “M”.

Por exemplo, no grupo “A”, as ORFs XAC0009, XAC0010 e XAC0011 estão numa seqüência com intervalos não caracterizados como ORFs com RIU passível de estudo, com exceção da primeira (XAC0009) que possui uma RIU com 88 b. Entre a XAC0009 e a XAC0010 (segunda ORF) têm-se 49 b, porém o modelo não encontrou promotor completo. Entre a XAC0010 e a XAC0011 (terceira ORF), têm-se uma RIU de 6 b, ou seja, menor que 30 b. Pode-se dizer que, provavelmente, este grupo compreende um ***operon***, já que pode-se observar também que todas as três ORFs apresentaram expressão presente em duas ou mais delas. Fato que amplia essa possibilidade é o mapeamento de um provável promotor à frente da ORF XAC0009.

Um indício de temporalidade da expressão nestes exemplos da Tabela 9, pode ser observado nos grupos “A”,”H”, “K” e “L”. Todos apresentaram a expressão no tempo zero na primeira ORF e a última proteína expressa da lista apresentou a expressão no tempo 3 (5 d.p.i.).

Os grupos “C”, “D” e “E” apresentaram expressão presente em todos os tempos de todas as ORFs. Estes grupos são formados por duas ORFs cada, sempre com provável promotor indicado apenas na primeira delas.

Os experimentos com proteômica utilizados (FACINCANI, 2007) mostraram que várias ORFs hipotéticas ou hipotéticas conservadas foram expressas ao menos em um dos quatro tempos, o que prova que codificam algum gene funcional. Alguns desses casos pode ser observado nestas estruturas de candidatos a ***operons***, como por exemplo, as ORFs: XAC3808 do grupo “H”, XAC1727 do grupo “K”, XAC1131 do grupo “L” e XAC0241 do grupo “M” (Tabela 9).

Outras ORFs hipotéticas ou hipotéticas conservadas, que fazem parte destas estruturas de possíveis ***operons***, não apresentaram nenhuma expressão, mesmo em grupos onde as outras ORFs foram expressas em praticamente todos os tempos. Fato que pode ser observado nos grupos “F”, “G”, “I” e “M” (Tabela 9).

Ainda sobre as hipotéticas e hipotéticas conservadas que não apresentaram expressão, um fato observado em vários grupos foi o caso de uma hipotética ser a primeira ORF destas estruturas de possíveis ***operons*** e, portanto, ser a ORF que teve um provável promotor identificado. Como por exemplo, os grupos “J”, “L” e “M”. Essas ORFs podem fazer parte de uma região não codificadora de genes.

O grupo “B”, da Tabela 9, apresentou o maior grupo contínuo de expressão presente de todos os dados, totalizando 39 ORFs. Nesse grupo observou-se vários possíveis ***operons*** iniciados pelas ORFs: XAC0959, XAC0961, XAC0967, XAC0971 (representa o maior deles, com 15 ORFs), XAC0986, XAC0989 e XAC0993. Entre essas estruturas observou-se alguns que apresentam expressão e têm seu próprio promotor indicado.

Tabela 9: Alguns exemplos de grupos de ORFs que representam possibilidades de ***operons***, em função da organização dos dados (com prováveis promotores identificados ou não) e a expressão da proteína dos dados de FACINCANI (2007). (Continua)

	PROTEOMA – Tempos				ORF	ANÁLISE DOS PROMOTORES MAPEADOS						ANOTAÇÃO NO GENOMA	
	0	1	2	3	ORF	GRUPO	TAMANHO SEQ OBSERV	TAMANHO PROMOTOR	NR HEX	SEQUENCIA DO PROMOTOR	3' do PROMOTOR	PRODUTO / PROTEÍNA	CATEGORIA
A	XAC0009	XAC0009	XAC0009	XAC0009	XAC0009	~100	88	40	2	ATCCCACGGGCCGGCCACGGTCGGCCCAGGATC CTGTCTT	-47	biopolymer transport ExbB protein	VII.C
	no	XAC0010	XAC0010	no	XAC0010	~50	49	0	1	* * * * * unique hexamer	0	biopolymer transport ExbD1 protein	VII.C
	no	no	XAC0011	XAC0011	XAC0011		6					biopolymer transport ExbD2 protein	VII.C
B	no	no	no	no	XAC0959	~50	52	36	2	TTTCCGGCCCAGCGACTCGGGTCAGATCGGTAACT	-8	preprotein translocase subunit	V.A.6
	XAC0960	XAC0960	XAC0960	XAC0960	XAC0960		16					transcription antitermination factor	III.B.5
	XAC0961	XAC0961	XAC0961	XAC0961	XAC0961	~200	200	13	2	ATGCTGCTATAGT	-153	50S ribosomal protein L11	III.B.2
	XAC0962	XAC0962	XAC0962	XAC0962	XAC0962		6					50S ribosomal protein L1	III.B.2
	XAC0963	XAC0963	XAC0963	XAC0963	XAC0963	~300	350	41	2	CTGGCAAGGAGTCGCCGTCGATGCAGCGGGATCG CTTAATT	-302	50S ribosomal protein L10	III.B.2
	XAC0964	XAC0964	XAC0964	XAC0964	XAC0964	~50	57	37	2	ATGTCCTAGCGTTCGCTAGAACCCCAATCCAGAAAA	-19	50S ribosomal protein L7/L12	III.B.2
	XAC0965	XAC0965	XAC0965	XAC0965	XAC0965	~300	282	20	2	TTGCATCGCCAGCGTTTCAT	-252	RNA polymerase beta subunit	III.B.5
	XAC0966	XAC0966	XAC0966	XAC0966	XAC0966	~200	153	53	2	CTGCAACCCGGAAGTGGCGAGCGCTCAGCGCGTT CGCTCTTCCCGGGTACGGT	-77	RNA polymerase beta' subunit	III.B.5
	XAC0967	XAC0967	XAC0967	XAC0967	XAC0967	~200	221	27	2	TAAGCGCTCCGACGGCGGTCGCAAGAT	-195	30S ribosomal protein S12	III.B.2
	XAC0968	XAC0968	XAC0968	XAC0968	XAC0968		15					30S ribosomal protein S7	III.B.2
	XAC0969	XAC0969	XAC0969	XAC0969	XAC0969	~100	140	31	2	TTGGTAACGGGGCCCTTCAGCACCATATGAG	-96	elongation factor G	III.C.1
	XAC0970	XAC0970	XAC0970	XAC0970	XAC0970	~50	51	14	2	TTCTCACAAATTTT	-21	elongation factor Tu	III.C.1
	XAC0971	XAC0971	XAC0971	XAC0971	XAC0971	~300	350	29	2	TTGCGCGCCCAAATCATGTTCGCTACAAT	-312	30S ribosomal protein S10	III.B.2
	XAC0972	XAC0972	XAC0972	XAC0972	XAC0972		14					50S ribosomal protein L3	III.B.2
	XAC0973	XAC0973	XAC0973	XAC0973	XAC0973		15					50S ribosomal protein L4	III.B.2
	XAC0974	XAC0974	XAC0974	XAC0974	XAC0974		-1					50S ribosomal protein L23	III.B.2
	XAC0975	XAC0975	XAC0975	XAC0975	XAC0975		13					50S ribosomal protein L2	III.B.2
	XAC0976	XAC0976	XAC0976	XAC0976	XAC0976		9					30S ribosomal protein S19	III.B.2
	XAC0977	XAC0977	XAC0977	XAC0977	XAC0977		9					50S ribosomal protein L22	III.B.2
	XAC0978	XAC0978	XAC0978	XAC0978	XAC0978		20					30S ribosomal protein S3	III.B.2
	XAC0979	XAC0979	XAC0979	XAC0979	XAC0979		8					50S ribosomal protein L16	III.B.2
	XAC0980	XAC0980	XAC0980	XAC0980	XAC0980		2					50S ribosomal protein L29	III.B.2
	XAC0981	XAC0981	XAC0981	no	XAC0981		14					30S ribosomal protein S17	III.B.2
	XAC0982	XAC0982	XAC0982	XAC0982	XAC0982		17					50S ribosomal protein L14	III.B.2
	XAC0983	XAC0983	XAC0983	XAC0983	XAC0983		18					50S ribosomal protein L24	III.B.2
	XAC0984	XAC0984	XAC0984	XAC0984	XAC0984		14					50S ribosomal protein L5	III.B.2
	no	XAC0985	XAC0985	XAC0985	XAC0985		21					30S ribosomal protein S14	III.B.2
	XAC0986	XAC0986	XAC0986	XAC0986	XAC0986	~200	207	64	2	TTGAACAGGATCGTTCTGTTCAACAAGGGAGTCCGA CGCAAGTCGGGCTCTTTTGTCGTATACT	-115	30S ribosomal protein S8	III.B.2
	XAC0987	XAC0987	XAC0987	XAC0987	XAC0987		18					50S ribosomal protein L6	III.B.2
	XAC0988	XAC0988	XAC0988	XAC0988	XAC0988	~100	91	43	2	TTGGTAGGAGCACGCTTTCCCTTCAGCTTTCGTAG GATAAATA	-3	50S ribosomal protein L18	III.B.2
	XAC0989	XAC0989	XAC0989	XAC0989	XAC0989	~200	161	55	2	TCGACCGCCTCCGCAATACCGCCGGCAGGGAAAT GCCGGACCGCCCCGTTCTTTT	-83	30S ribosomal protein S5	III.B.2
	XAC0990	XAC0990	XAC0990	XAC0990	XAC0990		25					50S ribosomal protein L30	III.B.2
	XAC0991	XAC0991	XAC0991	XAC0991	XAC0991		7					50S ribosomal protein L15	III.B.2
	no	XAC0992	no	no	XAC0992		10					preprotein translocase SecY subunit	V.A.6
	XAC0993	XAC0993	XAC0993	XAC0993	XAC0993	~300	335	13	2	TACACTAGATCTT	-316	30S ribosomal protein S13	III.B.2
	XAC0994	XAC0994	XAC0994	XAC0994	XAC0994		13					30S ribosomal protein S11	III.B.2
	XAC0995	XAC0995	XAC0995	XAC0995	XAC0995		18					30S ribosomal protein S4	III.B.2
	XAC0996	XAC0996	XAC0996	XAC0996	XAC0996	~50	56	17	2	TTCACATTGGAGAACCT	-37	RNA polymerase alpha subunit	III.B.5
	XAC0997	XAC0997	XAC0997	XAC0997	XAC0997	~200	183	45	2	ATGCGTCACCGCATGGGCGTCTTCGGGCGCCCAT GCTGTTTTTCT	-137	50S ribosomal protein L17	III.B.2

Tabela 9: Continuação.

	PROTEOMA – Tempos				ORF	ANÁLISE DOS PROMOTORES MAPEADOS						ANOTAÇÃO NO GENOMA	
	0	1	2	3	ORF	GRUPO	TAMANHO SEQ OBSERV	TAMANHO PROMOTOR	NR HEX	SEQUENCIA DO PROMOTOR	3' do PROMOTOR	PRODUTO / PROTEÍNA	CATEGORIA
C	XAC0487	XAC0487	XAC0487	XAC0487	XAC0487	~300	350	13	2	TGGAGACGCTGCT	-332	50S ribosomal protein L13	III.B.2
	XAC0488	XAC0488	XAC0488	XAC0488	XAC0488		5					30S ribosomal protein S9	III.B.2
D	XAC3579	XAC3579	XAC3579	XAC3579	XAC3579	~300	350	36	2	TTGACCAAGCCCGAATCCGTCGACGCCAGCGTTGT	-312	phosphoglucomutase/phosphomannomutase	VII.E
	XAC3580	XAC3580	XAC3580	XAC3580	XAC3580	~50	50	0	1	* * * * * unique hexamer	0	phosphomannose isomerase/GDP-mannose	VII.E
E	XAC3586	XAC3586	XAC3586	XAC3586	XAC3586	~300	350	35	2	TTCAGATTTCCCGCATAGGTCAGCGCGTCAAGATT	-316	electron transfer flavoprotein beta subunit	I.C.3
	XAC3587	XAC3587	XAC3587	XAC3587	XAC3587		2					electron transfer flavoprotein alpha subunit	I.C.3
F	XAC4339	XAC4339	XAC4339	XAC4339	XAC4339	~300	350	14	2	TCGTCACGTTTTGT	-337	toluene tolerance protein	VII.G
	XAC4340	no	no	XAC4340	XAC4340		2					toluene tolerance protein	VII.G
	XAC4341	XAC4341	XAC4341	XAC4341	XAC4341		20					toluene tolerance protein	VII.G
	XAC4342	XAC4342	XAC4342	XAC4342	XAC4342		-1					toluene tolerance protein	VII.G
	no	no	no	no	XAC4343		-8					conserved hypothetical protein	VIII.A
	XAC4344	XAC4344	XAC4344	XAC4344	XAC4344		13					lipoprotein	III.D.3
G	XAC3916	XAC3916	XAC3916	XAC3916	XAC3916	~50	69	35	2	TTGCTCGGACCCTGCGGCACGCGGTCGGGTAAAA	-8	arginyl-tRNA synthetase	III.B.4
	no	no	no	no	XAC3917	~200	182	17	2	TAAACAGGCGGTAGCTC	-166	conserved hypothetical protein	VIII.A
	no	no	no	no	XAC3918		9					Xanthomonas conserved hypothetical protein	VIII.C
	XAC3919	XAC3919	XAC3919	XAC3919	XAC3919	~50	40	0	1	* * * * * unique hexamer	0	oxidoreductase	I.C.3
H	XAC3807	XAC3807	XAC3807	XAC3807	XAC3807	~200	185	36	2	TTGACACGGCGACGCCCGCCGTGATCTTTCTAATA	-127	DNA topoisomerase I	III.A.1
	no	no	XAC3808	XAC3808	XAC3808		13					conserved hypothetical protein	VIII.A
	no	no	no	no	XAC3809		-71					Xanthomonas conserved hypothetical protein	VIII.C
	no	no	no	no	XAC3810		-1					Xanthomonas conserved hypothetical protein	VIII.C
I	XAC2466	XAC2466	XAC2466	XAC2466	XAC2466	~100	97	32	2	TCCACAGCGCCGTGCCGGCTTCCTGCTAGACT	-42	polar amino acid transporter	V.A.4
	no	no	no	no	XAC2467		-1					Xanthomonas conserved hypothetical protein	VIII.C
	no	XAC2468	XAC2468	no	XAC2468		3					magnesium and cobalt transport protein	V.A.4
J	no	no	no	no	XAC1827	~200	218	17	2	CTGTCAATTTGAAGATT	-197	conserved hypothetical protein	VIII.A
	XAC1828	no	XAC1828	no	XAC1828		12					ATP phosphoribosyltransferase	II.A.5
	no	no	XAC1829	XAC1829	XAC1829		-1					histidinol dehydrogenase	II.A.5
	no	XAC1830	no	no	XAC1830		-1					histidinol-phosphate aminotransferase	II.A.5
	no	XAC1831	XAC1831	XAC1831	XAC1831		-1					imidazoleglycerolphosphate dehydratase/histidinol-phosphate phosphatase bifunctional enzyme	II.A.5
	no	no	no	no	XAC1832		-1					amidotransferase	II.A.5
	no	no	no	XAC1833	XAC1833		-1					phosphoribosylformimino-5- aminoimidazole carboxam	II.A.5
	no	no	no	no	XAC1834		-4					cyclase	II.A.5
	XAC1835	XAC1835	XAC1835	no	XAC1835		-8					phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase bifunctional enzyme	II.A.5
K	XAC1725	no	no	no	XAC1725	~300	350	24	2	TTGGAACACGCGGGTGGGAAGGTT	-325	survival protein	VII.G
	no	no	no	no	XAC1726		-1					L-isoaspartate protein carboxylmethyltransferase type II	III.C.1
	XAC1727	XAC1727	XAC1727	no	XAC1727		21					conserved hypothetical protein	VIII.A
	no	no	no	XAC1728	XAC1728		-1					lipoprotein	III.D.3
L	XAC1131	XAC1131	XAC1131	no	XAC1131	~50	63	32	2	TTCAACGGGACATCACCGGCGCGCGGTATCCT	-25	conserved hypothetical protein	VIII.A
	XAC1132	no	XAC1132	no	XAC1132		-1					DNA polymerase III, delta' subunit	III.A.1
	XAC1133	XAC1133	XAC1133	XAC1133	XAC1133		-1					type IV fimbriae assembly protein	IV.D
M	XAC0241	XAC0241	XAC0241	XAC0241	XAC0241	~300	350	27	2	ATACCGATCAGCACCGGCTCGAAGCGT	-323	conserved hypothetical protein	VIII.A
	XAC0242	XAC0242	XAC0242	XAC0242	XAC0242		-1					ubiquinone biosynthesis protein	II.D.11
	no	no	no	no	XAC0243	~50	33	0	1	* * * * * unique hexamer	0	conserved hypothetical protein	VIII.A

Um outro caso observado diz respeito à ORF XAC3788, cujo produto é o Fator σ^{70} da RNA Polimerase, que foi expressa em todos os quatro tempos estudados por FACINCANI (2007). A seqüência completa do promotor identificada ***in silico*** permitiu verificar que o padrão encontrado difere apenas nos sítios 5 e 6 do hexâmero –35, onde é provável encontrar um "C" e um "A" é encontrado um "A" e um "T", respectivamente, quando comparado ao padrão encontrado para a ***E. coli*** (Figura 9).

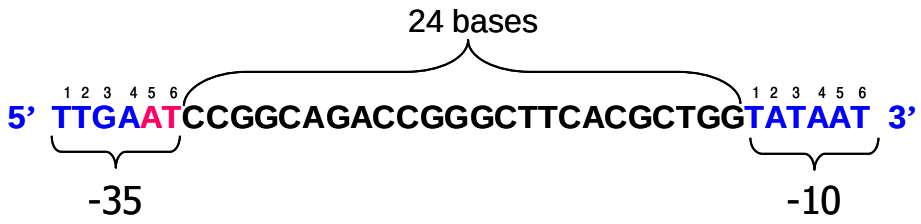

Figura 9: Provável promotor identificado da ORF XAC3788 (Fator σ^{70} da RNA Polimerase), com 36 bases de nucleotídeos.

Uma análise dessa seqüência no NCBI ("National Center for Biotechnology Information" - http://www.ncbi.nlm.nih.gov/blast/), por meio da ferramenta BLAST ("Basic Local Alignment Search Tool") e do algorítmo TblastX (ALTSCHUL ***et al.***, 1997) com parâmetros padrão, retornou similaridade exclusivamente de espécies de ***Xanthomonas***, tais como, ***Xanthomonas oryzae*** pv. ***oryzae*** str. KACC10331, ***Xanthomonas oryzae*** pv. ***oryzae*** str. MAFF 311018, ***Xanthomonas oryzae*** pv. ***oryzae*** str. PXO99A ***e Xanthomonas campestris*** pv. ***vesicatoria.*** Em todos os casos houve alinhamento perfeito com regiões "upstream" ao início da mesma ORF do Fator σ^{70} da RNA Polimerase***.*** Além disso, a distância da extremidade 3' do provável promotor mapeado está a exatas 45 bases da extremidade 5' da ORF que codifica a proteína Fator σ^{70} da RNA Polimerase em três dessas quatro ***Xanthomonas***, assim como ocorre em Xac (Figura 10).

```
>gb|CP000967.1|  Xanthomonas oryzae pv. oryzae PXO99A, complete genome
Length=5240075
 Features flanking this part of subject sequence:
   29 bp at 5' side: D-tyrosyl-tRNA(Tyr) deacylase
   45 bp at 3' side: RNA polymerase sigma-70 factor

 Score = 34.5 bits (69),  Expect = 5.3
 Identities = 12/12 (100%), Positives = 12/12 (100%), Gaps = 0/12 (0%)
 Frame = +1/+2

Query  1       LNPADRASRWYN  36
               LNPADRASRWYN
Sbjct  715526  LNPADRASRWYN  715561

>dbj|AP008229.1|  Xanthomonas oryzae pv. oryzae MAFF 311018 DNA, complete genome
Length=4940217
 Features flanking this part of subject sequence:
   29 bp at 5' side: conserved hypothetical protein
   45 bp at 3' side: RNA polymerase sigma-70 factor

 Score = 34.5 bits (69),  Expect = 5.3
 Identities = 12/12 (100%), Positives = 12/12 (100%), Gaps = 0/12 (0%)
 Frame = +1/+3

Query  1       LNPADRASRWYN  36
               LNPADRASRWYN
Sbjct  610857  LNPADRASRWYN  610892

>emb|AM039952.1|  Xanthomonas campestris pv. vesicatoria complete genome
Length=5178466
 Features flanking this part of subject sequence:
   45 bp at 5' side: RNA polymerase sigma-70 factor
   29 bp at 3' side: D-tyrosyl-tRNA(Tyr) deacylase

 Score = 34.5 bits (69),  Expect = 5.3
 Identities = 12/12 (100%), Positives = 12/12 (100%), Gaps = 0/12 (0%)
 Frame = +1/-2

Query  1        LNPADRASRWYN  36
                LNPADRASRWYN
Sbjct  4492854  LNPADRASRWYN  4492819

>gb|AE012027.1|  Xanthomonas axonopodis pv. citri str. 306,  section 405 of 469 of the complete
genome
Length=11067
 Score = 34.5 bits (69),  Expect = 5.3
 Identities = 12/12 (100%), Positives = 12/12 (100%), Gaps = 0/12 (0%)
 Frame = +1/-1

Query  1     LNPADRASRWYN  36
             LNPADRASRWYN
Sbjct  7416  LNPADRASRWYN  7381

>gb|AE013598.1|  Xanthomonas oryzae pv. oryzae KACC10331, complete genome
Length=4941439
 Features flanking this part of subject sequence:
   29 bp at 5' side: conserved hypothetical protein
   30 bp at 3' side: RNA polymerase sigma-70 factor

 Score = 34.5 bits (69),  Expect = 5.3
 Identities = 12/12 (100%), Positives = 12/12 (100%), Gaps = 0/12 (0%)
 Frame = +1/+3

Query  1       LNPADRASRWYN  36
               LNPADRASRWYN
Sbjct  622878  LNPADRASRWYN  622913
```

Figura 10: Resultados de alinhamentos do algorítmo TblastX (ALTSCHUL ***et al.***, 1997) com a seqüência promotora e a própria ORF do Fator σ^{70} da RNA Polimerase de Xac, comparando-se com o banco de seqüências do NCBI.

Quando se utilizou a ferramenta BlastN, com parâmetros padrão, contra o banco de nucleotídeos completo, além das ***Xanthomonas*** mencionadas no resultado do TblastX, também apareceu nesta nova análise um resultado referente a ***Xanthomonas campestris*** pv. ***campestris.*** No entanto, apesar de haver diferença em uma base (Figura 11), a distância de 45 b entre a extremidade 3' do promotor e a extremidade 5' da ORF persiste. Outros resultados de alinhamento com outros organismos foram apresentados, porém nenhum alinhamento relevante. Somente espécies de ***Xanthomonas*** apresentaram resultados de alta conservação neste provável promotor indicado.

```
>emb|AM920689.1|  Xanthomonas campestris pv. campestris complete genome, strain B100
Length=5079002

 Features flanking this part of subject sequence:
   45 bp at 5' side: RNA polymerase sigma-70 factor
   28 bp at 3' side: D-tyrosyl-tRNA(Tyr) deacylase

 Score = 63.9 bits (32),  Expect = 2e-08
 Identities = 35/36 (97%), Gaps = 0/36 (0%)
 Strand=Plus/Minus

Query  1        TTGAATCCGGCAGACCGGGCTTCACGCTGGTATAAT  36
                |||| ||||||||||||||||||||||||||||||
Sbjct  4433812  TTGATTCCGGCAGACCGGGCTTCACGCTGGTATAAT  4433777


>gb|AE012494.1|  Xanthomonas campestris pv. campestris str. ATCC 33913,  section 402 of 460
of the complete genome
Length=10873

 Score = 63.9 bits (32),  Expect = 2e-08
 Identities = 35/36 (97%), Gaps = 0/36 (0%)
 Strand=Plus/Minus

Query  1     TTGAATCCGGCAGACCGGGCTTCACGCTGGTATAAT  36
              |||| ||||||||||||||||||||||||||||||
Sbjct  7252  TTGATTCCGGCAGACCGGGCTTCACGCTGGTATAAT  7217


>gb|CP000050.1|  Xanthomonas campestris pv. campestris str. 8004, complete genome
Length=5148708

 Features flanking this part of subject sequence:
   45 bp at 5' side: RNA polymerase sigma-70 factor
   28 bp at 3' side: conserved hypothetical protein

 Score = 63.9 bits (32),  Expect = 2e-08
 Identities = 35/36 (97%), Gaps = 0/36 (0%)
 Strand=Plus/Minus

Query  1        TTGAATCCGGCAGACCGGGCTTCACGCTGGTATAAT  36
                |||| ||||||||||||||||||||||||||||||
Sbjct  4501552  TTGATTCCGGCAGACCGGGCTTCACGCTGGTATAAT  4501517


>gb|U82763.1|XCU82763  Xanthomonas campestris pv. campestris primary sigma factor RpoD (rpoD)
gene, complete cds
Length=2503

 Score = 63.9 bits (32),  Expect = 2e-08
 Identities = 35/36 (97%), Gaps = 0/36 (0%)
 Strand=Plus/Plus

Query  1    TTGAATCCGGCAGACCGGGCTTCACGCTGGTATAAT  36
             |||| ||||||||||||||||||||||||||||||
Sbjct  395  TTGATTCCGGCAGACCGGGCTTCACGCTGGTATAAT  430
```

Figura 11: Resultados de alinhamentos do algorítmo BlastN (ALTSCHUL ***et al.***, 1997) com a seqüência promotora e a própria ORF do Fator σ^{70} da RNA Polimerase de Xac, comparando-se com o banco de seqüências de nucleotídeos do NCBI.

Além disso, o alinhamento das seqüências de nucleotídeos dessa região nessas ***Xanthomonas,*** com o provável promotor identificado para Xac mostra que, da região promotora até o sítio +1, há uma alta conservação nas bases principalmente no próprio promotor. Há uma exceção na seqüência de ***Xanthomonas campestris*** pv. ***campestris***, onde se observa uma base "T" no lugar de uma base "A" no quinto sítio do hexâmero –35, sendo que as demais são totalmente conservadas (Figura 12). Por outro lado, pode-se constatar que a região codificante da ORF do Fator σ^{70} da RNA Polimerase não é tão conservada quanto a região promotora (Figura 12).

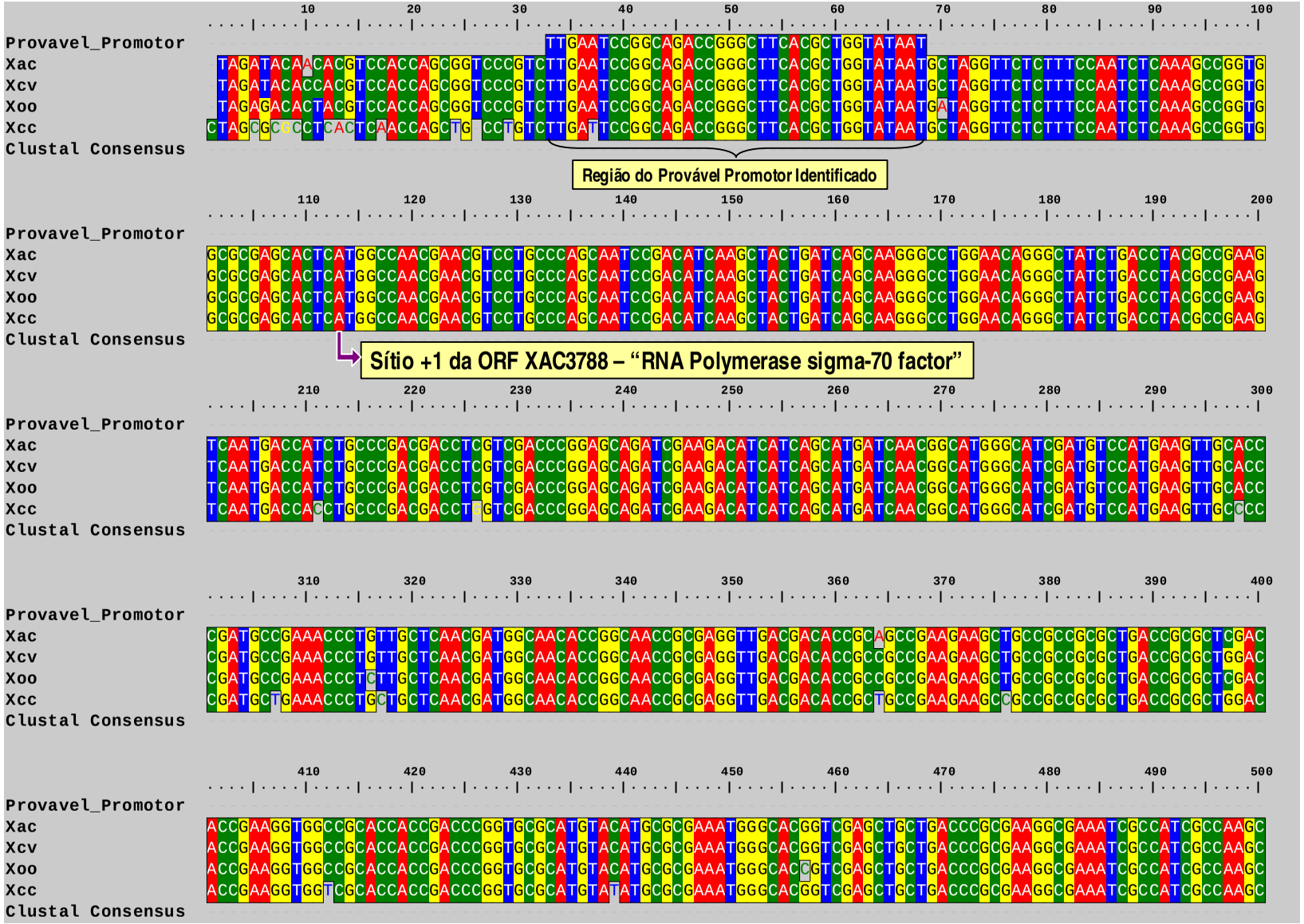

Figura 12: Alinhamento da seqüência da ORF XAC3788 de Xac precedida pela seqüência de sua RIU, com a mesma região da Xcv (***Xanthomonas campestris*** pv. ***vesicatoria***), da Xoo (***Xanthomonas oryzae*** pv. ***oryzae***) e da Xcc (***Xanthomonas campestris*** pv. ***campestris***). (Continua...)

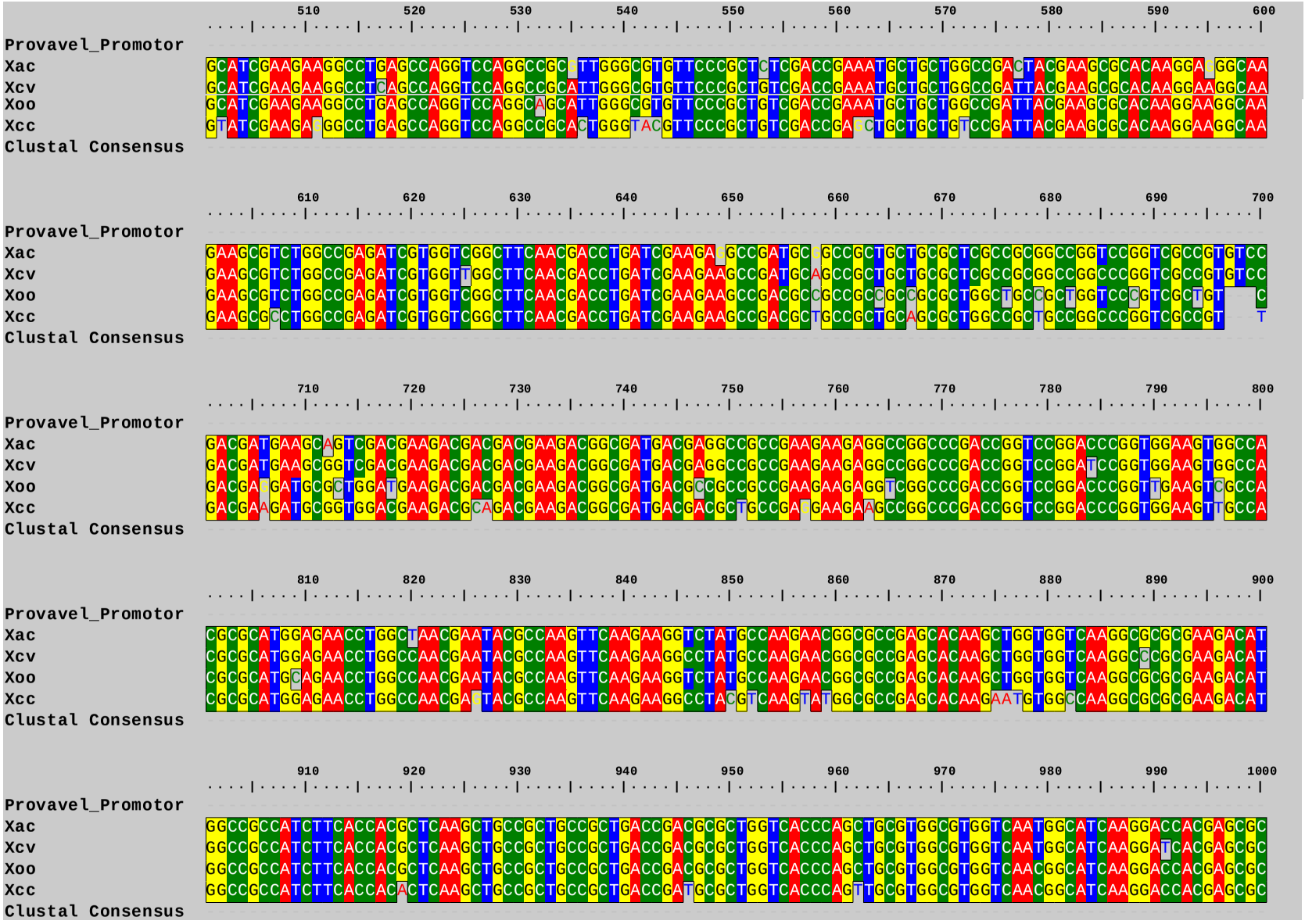

Figura 12: (Continua)

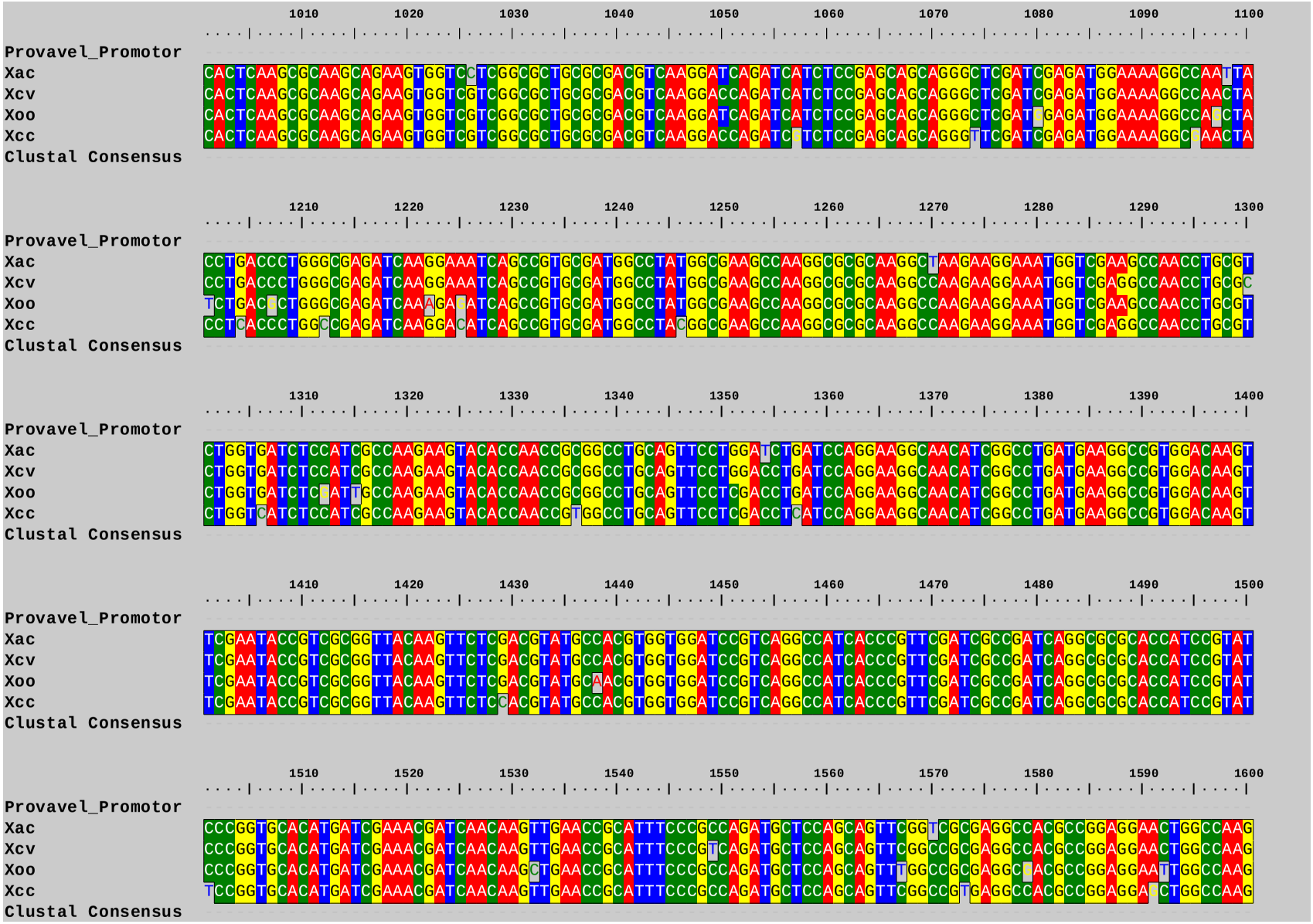

Figura 12: (Continua)

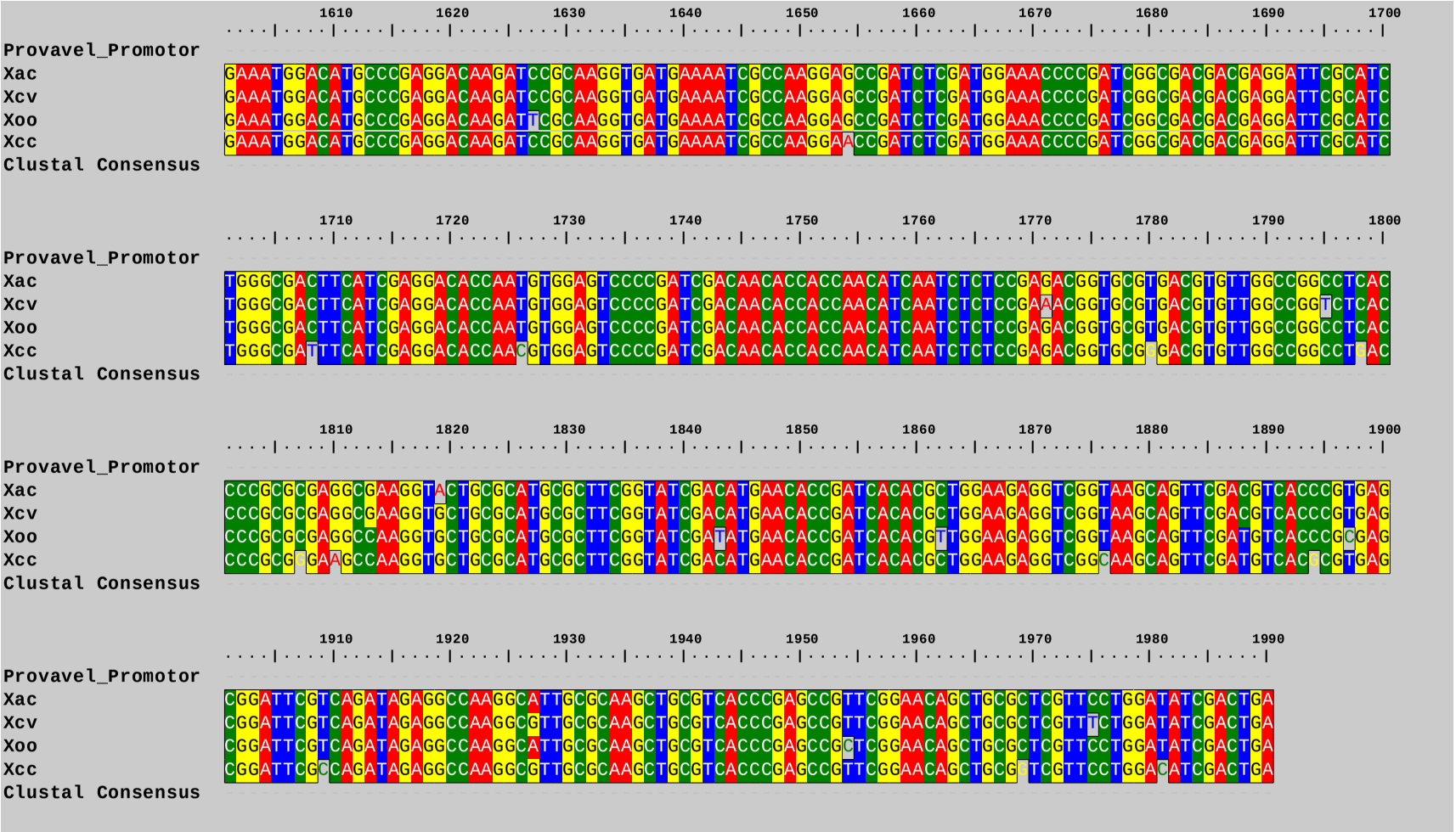

Figura 12: Término da Figura.

V. CONCLUSÕES

O modelo proposto mostrou-se eficaz para indicar as regiões promotoras da bactéria ***Xanthomonas axonopodis*** pv. ***citri*** (Xac), em razão dos resultados das análises estruturais dos prováveis promotores identificados e da comparação com o perfil proteômico de Xac.

Uma análise comparativa dos dados obtidos com os dados experimentais de proteômica mostrou que para aproximadamente 72% das proteínas expressas foi mapeado um promotor, o que corresponde a 27% do total de ORFs do genoma.

O seqüenciamento completo do genoma e a ordenação dos genes em tabelas, em fitas separadas, mostrou que pode-se levantar novos estudos sobre estruturas de ***operon*** à partir destes dados comparando-se com o estudo da temporalidade da expressão das proteínas.

A. Dados suplementares

No endereço https://genoma4.fcav.unesp.br/xacpromoters/ são disponibilizados todos os dados obtidos neste estudo.

VI. REFERÊNCIAS

ABECITRUS - Associação Brasileira dos Exportadores de Cítricos do Estado de São Paulo. **A História da Laranja - Um caso de sucesso**. Disponível em: http://www.abecitrus.com.br/historia_br.html . Acesso em: 01 de fevereiro de 2008.

ALMEIDA N. F. de; SETUBAL J. C. **Um Modelo Oculto de Markov para encontrar promotores em seqüências de DNA.** Campinas, SP: UNICAMP, 1998. 12 p. (Relatório Técnico, IC-98-37).

ALTSCHUL, S. F.; MADDEN, T. L.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleid Acids Research**, Inglaterra, v. 25, n. 17, p. 3389-3402. 1997.

BALDI, P.; BRUNAK, S. **Bioinformatics:** the machine learning approach. 2ª ed. Massachusetts: MIT Press, 2001. 351 p.

BROUWER, R. W. W.; KUIPERS, O. P.; HIJUM, S. A. F. T. The relative value of operon predictions. **Briefings in Bioinformatics**, Abril. 2008.

BROWN, K. Florida fights to stop citrus canker. **Science**, Estados Unidos, n. 292, p. 2275-2278, 2001.

CARVALHO, F. M. de S. **Expressão gênica em *Xanthomonas axonopodis* pv. *citri* controlada por promotores induzidos pela planta hospedeira.** 2006. 177 f. Tese (Doutorado em Ciências, área de concentração em Genética). Faculdade de Medicina

de Ribeirão Preto - Universidade de São Paulo. Ribeirão Preto – SP, 2006.

CHEN, X.; SU, Z.; DAM, P.; PALENIK, B.; XU, Y.; JIANG, T. Operon prediction by comparative genomics: an application to the ***Synechococcus*** sp WH8102 genome. **Nucleic Acids Research.** Inglaterra, v. 32, n. 7, p. 2147-2157. Abril. 2004.

CROOKS G. E.; HON G.; CHANDONIA J. M.; BRENNER S. E. WebLogo: A sequence logo generator. **Genome Research**, Estados Unidos, v. 14, n. 6, p. 1188-1190. Junho. 2004.

DA SILVA, A.C.R.; FERRO, J. A.; REINACH, F. C.; FARAH, C. S.; FURLAN, L. R.; QUAGGIO, R. B.; MONTEIRO-VITORELLO, C. B.; VAN SLUYS, M. A.; ALMEIDA, N. F.; ALVES L. M. C.; DO AMARAL, A. M.; BERTOLINI, M. C.; CAMARGO, L. E. A.; CAMAROTTE, G.; CANNAVAN, F.; CARDOSO, J.; CHAMBERGO, F.; CIAPINA, L. P.; CICARELLI, R. M. B.; COUTINHO, L. L.; CURSINO-SANTOS, J. R.; EL-DORRY, H.; FARIA, J. B.; FERREIRA, A. J. S.; FERREIRA, R. C. C.; FERRO, M. I. T.; FORMIGHIERI, E. F.; FRANCO, M. C.; GREGGIO, C. C.; GRUBER, A.; KATSUYAMA, A. M.; KISHI, L. T.; LEITE, R. P.; LEMOS, E. G. M.; LEMOS, M. V. F.; LOCALI, E. C.; MACHADO, M. A.; MADEIRA, A. M. B. N.; MARTINEZ-ROSSI, N. M.; MARTINS, E. C.; MEIDANIS, J.; MENCK, C. F. M.; MIYAKI, C. Y.; MOON, D. H.; MOREIRA, L. M.; NOVO, M. T. M.; OKURA, V. K.; OLIVEIRA, M. C.; OLIVEIRA, V. R.; PEREIRA, H. A.; ROSSI, A.; SENA, J. A. D.; SILVA, C.; DE SOUZA, R. F.; SPINOLA, L. A. F.; TAKITA, M. A.; TAMURA, R. E.; TEIXEIRA, E. C.; TEZZA, R. I. D.; TRINDADE DOS SANTOS, M.; TRUFFI, D.; TSAI, S. M.; WHITE, F. F.; SETUBAL, J. C.; KITAJIMA, J. P. Comparison of the genomes of two ***Xanthomonas*** pathogens with differing host specificities. **Nature**, Inglaterra, v.417, p.459-463, 2002.

EXAME. Anuário do Agronegócio. São Paulo - SP: Editora Abril. Agosto. 2005. Suplemento.

EXAME. Anuário do Agronegócio. São Paulo - SP: Editora Abril. Junho. 2006. Suplemento.

FACINCANI, A. P. **Análise Proteômica do Fitopatógeno *Xanthomonas axonopodis* pv. *citri*.** 2007. 100 f. Tese (Doutorado em Agronomia, Genética e Melhoramento de Plantas). Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista – FCAV UNESP. Jaboticabal – SP, 2007.

FENSELAU, S., BONAS, U. Sequence and expression analysis of the hrpB pathogenicity operon of ***Xanthomonas campestris*** pv. vesicatoria which encodes eight proteins with similarity to components of the Hrp, Ysc, Spa, and Fli secretion systems. **Molecular Plant-Microbe Interact**, Estados Unidos, v.8, p. 845 - 854. Novembro - Dezembro. 1995.

FUNDECITRUS – Fundo de Defesa da Citriculltura do Estado de São Paulo. Disponível em: http://www.fundecitrus.com.br/doencas/cancro.html. **Cancro Cítrico.** Acesso em: 01 de dezembro 2007.

FUNDECITRUS. Revista do Fundo de Defesa da Citricultura. Araraquara – SP: São Francisco Gráfica e Editora Ltda, Ano XXIV, n. 144, março/abril. 2008.

GIBAS C.; JAMBECK P. **Desenvolvendo Bioinformática:** Ferramentas de software para aplicações em biologia. Tradução: Cristina de Amorim Machado. Rio de Janeiro – RJ: Campus - O'Reilly, 2001. 440 p.

GRIFFITHS, A. J. F.; MILLER, J. H.; SUZUKI, D. T.; LEWONTIN, R. C.; GELBART, W. M. **Introdução à Genética.** 6ª ed. Tradução: Paulo Armando Motta. Rio de Janeiro – RJ: Ed. Guanabara Koogan S.A., 1998. 856 p.

POSTLETHWAIT, J. H.; HOPSON, J. L. **Modern Biology**. Estados Unidos: Holt, Rinehart ans Winston – A Harcourt Education Company. 2006. 1130 p.

KOLLER, O. L.; SOPRANO, E.; BONAS, U. **Normas técnicas para a cultura de citros em Santa Catarina.** Sistemas de Produção n° 14. Empresa de Pesquisa Agropecuária e Difusão de Tecnologia de Santa Catarina S.A (EPAGRI), 1993.

LEITE JR, R. P. **Cancro Cítrico:** Prevenção e Controle no Paraná. IAPAR, Londrina, PR, Brasil. 51p. (IAPAR. Circular, 61), 1990.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Principles of Biochemistry.** Estados Unidos: W. H. Freeman. 2004. 1119 p.,

LEWIN, B**. Genes VII**. Oxford: Oxford University Press, 2000. 990 p.

MONTEIRO, M. I. **Predição de Promotores de *Bacillus subtilis* usando técnicas de Aprendizado de Máquina.** 2005. 75 f. Dissertação (Mestrado em Ciências no Programa de Pós-graduação em Engenharia Elétrica) - Universidade Federal do Rio Grande do Norte - UFRN. Natal - RN, 2005.

NAMEKATA, T.; ROSSI, A. C.; CERÁVOLO, L. C. In: **Laranja**. Avaliação de novos métodos de erradicação de CC. Cordeirópolis-SP, v. 17, n. 1, p. 67-78. 1996.

OLEKHNOVICH, I. N.; KADNER, R. J. RNA Polymerase alpha and sigma 70 subunits participate in transcription of the ***Escherichia coli uhpT*** promoter. **Journal of Bacteriology**, Estados Unidos, v.181, n.23, p.7266-7227. Dezembro. 1999.

OPPON, E. C. **Synergistic use of promoter prediction algorithms**: a choice for small

training dataset ? 2000. 238 f. Tese (Doutorado em Ciência da Computação) – Department of Computer Science – University of the Western Cape – África do Sul, 2000.

QIU, P. Recent advantages in computational promoter analysis in understanding the transcriptional regulatory network. **Biochemical and Biophysical Research Communications**, Estados Unidos, vol.309, p.495-501. 2003.

REIS, A. N**. Reconhecimento e Predição de Promotores Procarióticos**: investigação de uma metodologia ***in silico*** baseada em HMMs. 2005. 116 f. Dissertação (Mestrado em Computação Aplicada) – Universidade do Vale do Rio dos Sinos – UNISINOS, São Leopoldo - RS, 2005.

RUDOLPH, K. In: **Xanthomonas**. Infection of the plant by ***Xanthomonas***. Ed. Swings JG and Giverolo EL. p. 193-264. 1993.

SALGADO, H.; MORENO-HAGELSIEB, G.; SMITH, T. F.; COLLADO-VIDES, J. Operons in Escherichia coli: genomic analyses and predictions. **Proceedings of the National Academy of Sciences of the United States of America**, Estados Unidos, 97, p.6652-6657. 2000.

SCHNEIDER, T. D. Strong minor groove base conservation in sequence logos implies DNA distortion or base flipping during replication and transcription initiation. **Nucleid Acids Research**, Inglaterra, v. 29, n. 23, p. 4881-4891. Dezembro. 2001.

TAOKA, M.; YAMAUCHI, Y.; SHINKAWA, T.; KAJI, H.; MOTOHASHI, W.; NAKAYAMA, H.; TAKAHASHI, N.; ISOBE T. Only a small subset of the horizontally transferred

chromosomal genes in ***Escherichia coli*** are translated into proteins. **Molecular & Cellular Proteomics**, Estados Unidos, v.3, p.780-787. Agosto. 2004.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Res**, v. 22, n. 22, p. 4673-4680, 1994.

WANG, L.; TRAWICK, J. D.; YAMAMOTO, R.; ZAMUDIO, C. Genome-wide operon prediction in ***Staphylococcus aureus***. **Nucleid Acids Research**, Inglaterra, v. 32, n. 12, p. 3689-3702. Julho. 2004.

WENGELNIK, K.; MARIE, C.; RUSSEL, M.; BONAS, U. Expression and localization of HrpA1, a protein of ***Xanthomonas campestris*** pv. ***vesicatoria*** essencial for pathogenicity and induction of the hypersensitive reation. **Journal of Bacteriology**, Estados Unidos, v.178, p.1061-1069. Dezembro. 1996b.

WHITESIDE, J. O.; GARNSEY, S. M.; TIMMER, L. W. **Compendium of citrus diseases.** Saint Paul: APS Press, 1988. 80p.

XIONG J. **Essential Bioinformatics.**New York:Cambridge University Press,2006. 339 p.

YADA, T.; NAKAO, M.; TOTOKI, Y.; NAKAI, K. Modeling and predicting transcriptional units of ***Escherichia coli*** genes using hidden Markov models. **Bioinformatics**, Inglaterra, v.15, n.12, p.987-993. Dezembro. 1999.

ZAHA, A. **Biologia Molecular Básica**. Mercado Aberto, 2003. 421 p.

APÊNDICES

Apêndice A: Programa upstream_intergenic_mapping.pl

```
#!/usr/local/bin/perl -w
##############################################################################################################
#
# File: upstream_intergenic_mapping.pl
##############################################################################################################
#
# PARAMETER TESTING
if ($#ARGV != 2) {
  die "Usage: $0 <FASTA_FILE> <GENE_LIST_FILE> <[C]IRCULAR/[N]OT CIRCULAR>\n";
  exit;
}

# VARIABLES INITIALIZATION
my ($a,$circular,$gene,$reverse_fasta) = "";
my (@forward_fasta,@intergenic_fasta) = ();
my ($pos1,$pos2,$i) = 0;
my (@genes_F,@start_pos_F,@stop_pos_F,@size_F,@size_group_F,@intergenic_upstream_fasta_F) = ();
my (@genes_R,@start_pos_R,@stop_pos_R,@size_R,@size_group_R,@intergenic_upstream_fasta_R) = ();

# OPEN FASTA FILE
open (FASTA,"<$ARGV[0]") or die "Impossible open $ARGV[0] file: $! \n";
while (<FASTA>)
{
   chomp;
   $a .= $_;
}
# FORWARD FASTA
$a =~ s/\s//g;
@forward_fasta=split(//,$a);
close (FASTA);


# HEADING OUT FILE
print "Program: $0 \n";
print
"=============================================================================================\n";
print "User parameters:\n";
print "Fasta file: $ARGV[0]\n";
print "Complete gene list text file: $ARGV[1]\n";
if ($ARGV[2] eq "C")
{
   $circular = $ARGV[2];
   print "CIRCULAR fasta.\n";
}
else
{
   $circular = "N";
   print "NO CIRCULAR fasta.\n";
}
print "Obs.: The upstream intergenic sequences observe gene's direction with regard to complete
fasta.\n";
print
"=============================================================================================\n";
print ">>>\n";
print
"GENE"."\t"."GENE_DIRECTION"."\t"."GENE_START_POSITION"."\t"."GENE_END_POSITION"."\t"."UPSTREAM
```

```
_INTERGENIC_SIZE_GROUP"."\t"."UPSTREAM_INTERGENIC_SIZE"."\t"."UPSTREAM_INTERGENIC_FASTA_SEQUENC
E"."\n";


# READ FILE OF COMPLETE GENE LIST AND SPLIT FOR DIRECTION
open (GENES_LIST,"<$ARGV[1]") or die "Impossible open file $ARGV[1]: $!\n";
while(<GENES_LIST>)
{
   if ($_ =~ /^(\S+)\t(\d+)\t(\d+)/)
   {
      $gene=$1;
      $pos1=$2;
      $pos2=$3;
   # Gene Forward
      if ($pos1<$pos2)
      {
         push @genes_F,$gene ;
         push @start_pos_F,$pos1 ;
         push @stop_pos_F,$pos2;
      }
   # Gene Reverse
      if ($pos1>$pos2)
      {
         push @genes_R,$gene ;
         push @start_pos_R,$pos1 ;
         push @stop_pos_R,$pos2;
      }
   }
}
close(GENES_LIST);

##############################################################################################
# INTERGENIC UPSTREAM SEQUENCE SIZE CALCULATION FOR FORWARD GENES
$i=0;
while (<@genes_F>)
{
   $intergenic_upstream_fasta_F[$i] = "";
   if($circular eq "C")
   {
      ### formula geral
      if ($start_pos_F[$i]>$stop_pos_F[$i-1])
      {
         $size_F[$i] = ($start_pos_F[$i]-$stop_pos_F[$i-1])-1;
         if ($size_F[$i] != 0)
         {
            # GUARDA O FASTA DA RI -
            foreach (($stop_pos_F[$i-1]+1)..($start_pos_F[$i]-1))
            { $intergenic_upstream_fasta_F[$i] .= "$forward_fasta[$_-1]";}
         }
         else
         {$intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";}

      }
      ### formula para negativos ou zero
      if (($start_pos_F[$i]<=$stop_pos_F[$i-1]) && $i>0)
      {
         $size_F[$i] = ($start_pos_F[$i]-$stop_pos_F[$i-1])-1;
          $intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * * *";
      }
```

```
     if (($start_pos_F[$i]<$stop_pos_F[$i-1]) && $i==0)
      {
         $size_F[$i] = $start_pos_F[$i] + ((($#forward_fasta+1)-$stop_pos_F[$#stop_pos_F])-1);
         if ($size_F[$i] != 0)
         {
            foreach (($stop_pos_F[$i-1]+1)..($#forward_fasta+1))
            { $intergenic_upstream_fasta_F[$i] .= "$forward_fasta[$_-1]";}
            foreach (1..($start_pos_F[$i]-1))
            { $intergenic_upstream_fasta_F[$i] .= "$forward_fasta[$_-1]";}
         }
         else{$intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";}
      }
   }
   else
   {
      ### formula geral
      if ($start_pos_F[$i]>$stop_pos_F[$i-1])
      {
         $size_F[$i] = ($start_pos_F[$i]-$stop_pos_F[$i-1])-1;
         if ($size_F[$i] > 0)
         {
            foreach (($stop_pos_F[$i-1]+1)..($start_pos_F[$i]-1))
            { $intergenic_upstream_fasta_F[$i] .= "$forward_fasta[$_-1]";}
         }
         else
         {
            $intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
         }
      }
      if (($start_pos_F[$i]<=$stop_pos_F[$i-1]) && $i>0)
      {
         $size_F[$i] = ($start_pos_F[$i]-$stop_pos_F[$i-1])-1;
         ### precisa deixar o -1 para simbolizar em qtas bases os genes se sobrepoe.
         $intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * * *";
      }
      if (($start_pos_F[$i]<$stop_pos_F[$i-1]) && $i==0)
      {
         $size_F[$i] = ($start_pos_F[$i])-1;
         if ($size_F[$i] > 0)
         {
            foreach (1..($start_pos_F[$i]-1))
            { $intergenic_upstream_fasta_F[$i] .= "$forward_fasta[$_-1]";}
         }
         else
         {
            $intergenic_upstream_fasta_F[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
         }
      }
   }
    $i++;
}
# INTERGENIC UPSTREAM SEQUENCE SIZE CALCULATION FOR REVERSE GENES
$i=0;
while (<@genes_R>)
{
   $intergenic_upstream_fasta_R[$i] = "";
   if($circular eq "C")
```

```
   {
      if ($i<$#genes_R)
      {
         ### formula geral
         if ($start_pos_R[$i]<$stop_pos_R[$i+1])
         {
            $size_R[$i] = ($stop_pos_R[$i+1]-$start_pos_R[$i])-1;
            if ($size_R[$i] != 0)
            {
               foreach (($start_pos_R[$i]+1)..($stop_pos_R[$i+1]-1))
               { $intergenic_upstream_fasta_R[$i] .= $forward_fasta[$_-1];}
            }
            else
            {
            $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
            }

         }
         ### formula para negativos ou zero
         if ($stop_pos_R[$i+1]<=$start_pos_R[$i])
         {
            $size_R[$i] = ($stop_pos_R[$i+1]-$start_pos_R[$i])-1;
            $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
         }
      }
      elsif ($i == $#genes_R)
      {
         ### formula para final da seq com circularidade
         if (((((($#forward_fasta+1)-$start_pos_R[$i])+$stop_pos_R[0])-1) <= 0)
         {
            $size_R[$i] = ((((($#forward_fasta+1)-$start_pos_R[$i])+$stop_pos_R[0])-1) ;
            $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
         }
         if (((((($#forward_fasta+1)-$start_pos_R[$i])+$stop_pos_R[0])-1) > 0)
         {
            $size_R[$i] = ((((($#forward_fasta+1)-$start_pos_R[$i])+$stop_pos_R[0])-1);
            # GUARDA O FASTA do final do fastao
            if (($start_pos_R[$i]+1) <= ($#forward_fasta+1))
            {
               foreach (($start_pos_R[$i]+1)..($#forward_fasta+1))
               { $intergenic_upstream_fasta_R[$i] .= "$forward_fasta[$_-1]";}
            }
            # GUARDA O FASTA do inicio do fastao
            if (($stop_pos_R[0]-1) >= 1)
            {
               foreach (1..($stop_pos_R[0]-1))
               { $intergenic_upstream_fasta_R[$i] .= "$forward_fasta[$_-1]";}
            }
         }
      }
   }
   else
   {
      if ($i<$#genes_R)
      {
         ### formula geral
         if ($start_pos_R[$i]<$stop_pos_R[$i+1])
```

```
        {
           $size_R[$i] = ($stop_pos_R[$i+1]-$start_pos_R[$i])-1;
           # GUARDA O FASTA DA RI e não for zero
           if ($size_R[$i] > 0)
           {
              foreach (($start_pos_R[$i]+1)..($stop_pos_R[$i+1]-1))
              { $intergenic_upstream_fasta_R[$i] .= $forward_fasta[$_-1];}
           }
           else
           {
           $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
           }
        }
        if ($stop_pos_R[$i+1]<=$start_pos_R[$i])
        {
           $size_R[$i] = ($stop_pos_R[$i+1]-$start_pos_R[$i])-1;
           $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
        }
     }
     elsif ($i == $#genes_R)
     {
        ### formula para final da seq SEM circularidade
        if ((($#forward_fasta+1)-$start_pos_R[$i]) == 0)
        {
           $size_R[$i] = (($#forward_fasta+1)-$start_pos_R[$i]) ;
           $intergenic_upstream_fasta_R[$i] = "* * * NOT FOUND INTERGENIC REGION UPSTREAM * *
*";
        }
        if ((($#forward_fasta+1)-$start_pos_R[$i]) > 0)
        {
           $size_R[$i] = (($#forward_fasta+1)-$start_pos_R[$i]);
           foreach (($start_pos_R[$i]+1)..($#forward_fasta+1))
           { $intergenic_upstream_fasta_R[$i] .= "$forward_fasta[$_-1]";}
        }
     }
   }
  $i++;
}
#############################################################################################
$i=0;
while (<@genes_F>)
{
   $size_group_F[$i] = ('<=0') if ($size_F[$i] <= 0);
   $size_group_F[$i] = ('1..30') if ($size_F[$i]>0 && $size_F[$i] <= 30);
   $size_group_F[$i] = ('~50') if ($size_F[$i]>30 && $size_F[$i] <= 75);
   $size_group_F[$i] = ('~100') if ($size_F[$i]>75 && $size_F[$i] <= 150);
   $size_group_F[$i] = ('~200') if ($size_F[$i]>150 && $size_F[$i] <= 250);
   $size_group_F[$i] = ('~300') if ($size_F[$i]>250);
#  print $size_group_F[$i]."\n";
   $i++;
}
$i=0;
while (<@genes_R>)
{
   $size_group_R[$i] = ('<=0') if ($size_R[$i] <= 0);
   $size_group_R[$i] = ('1..30') if ($size_R[$i]>0 && $size_R[$i] <= 30);
   $size_group_R[$i] = ('~50') if ($size_R[$i]>30 && $size_R[$i] <= 75);
   $size_group_R[$i] = ('~100') if ($size_R[$i]>75 && $size_R[$i] <= 150);
```

```
   $size_group_R[$i] = ('~200') if ($size_R[$i]>150 && $size_R[$i] <= 250);
   $size_group_R[$i] = ('~300') if ($size_R[$i]>250);
#  print $size_group_R[$i]."\n";
   $i++;
}

# PRINT THE OUTPUT FILE
foreach (0..$#genes_F)
{
   print
"$genes_F[$_]"."\t"."FORWARD"."\t"."$start_pos_F[$_]"."\t"."$stop_pos_F[$_]"."\t"."$size_group_
F[$_]"."\t"."$size_F[$_]"."\t"."$intergenic_upstream_fasta_F[$_]"."\n";
}


# SELECT INTERGENIC UPSTREAM SEQUENCE OF REVERSE GENES
foreach (0..$#genes_R)
{
   # gera o reverso complementar
   if ($intergenic_upstream_fasta_R[$_] !~ /^\*/)
   {
     $reverse_fasta = uc (reverse($intergenic_upstream_fasta_R[$_]));
     $reverse_fasta =~ s/G/c/g;
     $reverse_fasta =~ s/C/g/g;
     $reverse_fasta =~ s/T/a/g;
     $reverse_fasta =~ s/A/t/g;
     $intergenic_upstream_fasta_R[$_] = uc ($reverse_fasta);

   }
   print
"$genes_R[$_]"."\t"."REVERSE"."\t"."$start_pos_R[$_]"."\t"."$stop_pos_R[$_]"."\t"."$size_group_
R[$_]"."\t"."$size_R[$_]"."\t"."$intergenic_upstream_fasta_R[$_]"."\n";
}

exit;
```

Apêndice B: Programa background_fasta.pl

```
#!/usr/local/bin/perl
##############################################################################################
# File: background_fasta.pl
##############################################################################################
# PARAMETER TESTING
if ($#ARGV != 1) {
  die "Usage: $0 <FASTA_FILE> <BASES>\n";
  exit;
}

# VARIABLES INITIALIZATION
my ($a,$rev_fasta) = "";
my (@forward_fasta,@reverse_fasta,@bases_observation) = ();
my ($i,$total) = 0;

# OPEN FASTA FILE
open (FASTA,"<$ARGV[0]") or die "Impossible open $ARGV[0] file: $! \n";
while (<FASTA>)
{
   chomp;
   $a .= $_;
}
# FORWARD FASTA
$a =~ s/\s//g;
@forward_fasta=split(//,uc($a));
close (FASTA);

# REVERSE FASTA
$rev_fasta = uc (reverse($a));
$rev_fasta =~ s/G/c/g;
$rev_fasta =~ s/C/g/g;
$rev_fasta =~ s/T/a/g;
$rev_fasta =~ s/A/t/g;
$a = uc ($rev_fasta);
@reverse_fasta=split(//,$a);


# BASES FOR OBSERVATION

$a = uc($ARGV[1]);

@bases_observation=split(//,$a);
print 'BACKGROUND (%) of '."\'@bases_observation\' in total of ".($#forward_fasta+1)."
bases in fasta file $ARGV[0]"."\n\n";
print ">FORWARD FASTA:\n";

foreach $base (@bases_observation)
{
   $i=0;
   for ($j=0;$j<=$#forward_fasta;$j++)
   {
      if ($forward_fasta[$j] eq $base) {
      $i++; }
   }

   $total= $i / ($#forward_fasta+1);
   print "$base"."\t"."$total"."\n";
}
```

```
print "\n";
print ">REVERSE FASTA:\n";

foreach $base (@bases_observation)
{
   $i=0;
   for ($j=0;$j<=$#reverse_fasta;$j++)
   {
      if ($reverse_fasta[$j] eq $base) {
      $i++; }
   }

   $total= $i / ($#reverse_fasta+1);
   print "$base"."\t"."$total"."\n";
}

exit;
```

Apêndice C: Programa make_input_probabilities_FILES.pl

```
#!/usr/local/bin/perl -w
##############################################################################
# File: make_input_probabilities_FILES.pl
##############################################################################

# PARAMETER TESTING
if ($#ARGV != 1) {
  die "Usage: $0  <GENE_LIST_WITH_UPSTREAM_INTERGENIC_FASTA_FILE>\n";
  exit;
}

# VARIABLES INITIALIZATION
my
($file_50_F,$file_50_R,$file_100_F,$file_100_R,$file_200_F,$file_200_R,$file_300_F,$fil
e_300_R) = "";
my ($file_NEG_F,$file_NEG_R,$file_30_F,$file_30_R) = "";
my ($gene,$direction,$size_group,$seq_observ,$seq_intergenic) = "";
my ($size_intergenic_region) = 0;


$file_NEG_F = "NEG_WITHOUT_INTERGENIC_REGION_F.txt";
$file_NEG_R = "NEG_WITHOUT_INTERGENIC_REGION_R.txt";
$file_30_F = "INTERGENIC_REGION_LESS_THAN_30_F.txt";
$file_30_R = "INTERGENIC_REGION_LESS_THAN_30_R.txt";
$file_50_F = "input_probabilities_50_F.txt";
$file_50_R = "input_probabilities_50_R.txt";
$file_100_F = "input_probabilities_100_F.txt";
$file_100_R = "input_probabilities_100_R.txt";
$file_200_F = "input_probabilities_200_F.txt";
$file_200_R = "input_probabilities_200_R.txt";
$file_300_F = "input_probabilities_300_F.txt";
$file_300_R = "input_probabilities_300_R.txt";

system ("/bin/mkdir input");


# OPEN INTERGENIC UPSTREAM SEQUENCES  FILE
open (IRFILE,"<$ARGV[0]") or die "Impossible open $ARGV[0] file: $! \n";
while (<IRFILE>)
{
   if ($_ =~ /^(\S+)\t(FORWARD|REVERSE)\t(\d+)\t(\d+)\t(\S+)\t([\-\d+]*)\t(\S+)/)
   {
      chomp;
      $gene = $1 ;
      $direction = $2 ;
      $size_group = $5 ;
      $size_intergenic_region = $6 ;
      $seq_intergenic = $7 ;
#     print "$size_group $size_intergenic_region \n";
      if (($size_group eq '<=0') && ($direction eq 'FORWARD'))
      {
         open (PROBAB,">>$file_NEG_F");
         print PROBAB $1."\t".$size_intergenic_region."\t".$seq_intergenic;
         print PROBAB "\n";
         close (PROBAB);
      }
```

```
if (($size_group eq '<=0') && ($direction eq 'REVERSE'))
{
   open (PROBAB,">>$file_NEG_R");
   print PROBAB $1."\t".$size_intergenic_region."\t".$seq_intergenic;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '1..30') && ($direction eq 'FORWARD'))
{
   open (PROBAB,">>$file_30_F");
   print PROBAB $1."\t".$size_intergenic_region."\t".$seq_intergenic;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '1..30') && ($direction eq 'REVERSE'))
{
   open (PROBAB,">>$file_30_R");
   print PROBAB $1."\t".$size_intergenic_region."\t".$seq_intergenic;
   print PROBAB "\n";
   close (PROBAB);
}


if (($size_group eq '~50') && ($direction eq 'FORWARD'))
{
   system ("/bin/cp $file_50_F input/$1.txt");
   open (PROBAB,">>input/$1.txt");
   print PROBAB $seq_intergenic ;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '~50') && ($direction eq 'REVERSE'))
{
   system ("/bin/cp $file_50_R input/$1.txt");
   open (PROBAB,">>input/$1.txt");
   print PROBAB $seq_intergenic ;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '~100') && ($direction eq 'FORWARD'))
{
   system ("/bin/cp $file_100_F input/$1.txt");
   open (PROBAB,">>input/$1.txt");
   print PROBAB $seq_intergenic ;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '~100') && ($direction eq 'REVERSE'))
{
   system ("/bin/cp $file_100_R input/$1.txt");
   open (PROBAB,">>input/$1.txt");
   print PROBAB $seq_intergenic ;
   print PROBAB "\n";
   close (PROBAB);
}
if (($size_group eq '~200') && ($direction eq 'FORWARD'))
{
   system ("/bin/cp $file_200_F input/$1.txt");
   open (PROBAB,">>input/$1.txt");
   print PROBAB $seq_intergenic ;
```

```
         print PROBAB "\n";
         close (PROBAB);
      }
      if (($size_group eq '~200') && ($direction eq 'REVERSE'))
      {
         system ("/bin/cp $file_200_R input/$1.txt");
         open (PROBAB,">>input/$1.txt");
         print PROBAB $seq_intergenic ;
         print PROBAB "\n";
         close (PROBAB);
      }
      if (($size_group eq '~300') && ($direction eq 'FORWARD'))
      {
         system ("/bin/cp $file_300_F input/$1.txt");
         open (PROBAB,">>input/$1.txt");
         if ($size_intergenic_region <= 350)
         {
            print PROBAB $seq_intergenic ;
            print PROBAB "\n";
         }
         else
         {
            $seq_observ = substr($seq_intergenic,-350,350);
            print PROBAB $seq_observ ;
            print PROBAB "\n";
         }
         close (PROBAB);
      }
      if (($size_group eq '~300') && ($direction eq 'REVERSE'))
      {
         system ("/bin/cp $file_300_R input/$1.txt");
         open (PROBAB,">>input/$1.txt");
         if ($size_intergenic_region <= 350)
         {
            print PROBAB $seq_intergenic ;
            print PROBAB "\n";
         }
         else
         {
            $seq_observ = substr($seq_intergenic,-350,350);
            print PROBAB $seq_observ ;
            print PROBAB "\n";
         }
         close (PROBAB);
      }
         }
}
close (IRFILE);

exit;
```

Apêndice D: Programa tata2HMM_for_all.pl

```
#!/usr/local/bin/perl -w

############################################################
# File: tata2HMM_for_all.pl
############################################################

# PARAMETER TESTING
if ($#ARGV != 1) {
  die "Usage: $0 <INPUT_FILES_DIR> <OUTPUT_FILES_DIR>\n";
  exit;
}

# VARIABLES INITIALIZATION
my ($dir_input_files,$dir_output_files) = "";


# OPEN DIRECTORY
$dir_input_files = $ARGV[0];
$dir_output_files = $ARGV[1];


open (VERIFY,'ls -1 '.$dir_input_files.'/ |') or die "Impossible open $ARGV[0]
directory: $! \n";

while(<VERIFY>)
{
    chomp();

    system ("/usr/local/bin/perl -w tata2.pl $dir_input_files/$_
>$dir_output_files/OUT_$_");

    print "$_ pronto!\n";
}


exit;
```

Apêndice E: Programa make_final_report.pl

```
#!/usr/local/bin/perl -w
########################################################################
# File: make_final_report.pl
########################################################################

# PARAMETER TESTING
if ($#ARGV != 2) {
  die "Usage: $0 <OUTPUT_FILES_DIR> <REPORT_OUTPUT_FILE_NAME>
<INTERGENIC_LIST_GENES>\n";
  exit;
}

# VARIABLES INITIALIZATION
my
($gene,$direction,$size_group,$observation_seq,$allleft,$tata_position_seq,$tata_seq) =
"";
my (@fileOUT,@tata_positions)=();
my
($size_seq_observ,$observation_length,$i,$j,$line,$numberOfHexamers,$Second_probability
,$tataLength,$start_TATA,$stop_TATA) = 0;
my ($Hex_10_start,$Hex_10_stop,$Hex_35_start,$Hex_35_stop,$diffHexamers,$h,$b) = 0;
my ($dir_output_files,$report_name) = "";

$dir_output_files = $ARGV[0];
$report_name = $ARGV[1];

open (VERIFY,'ls -1 '.$dir_output_files.'/ |') or die "Impossible open $ARGV[0]
directory: $! \n";
open (REPORT,">$report_name");
print REPORT
"GENE"."\t"."DIRECTION"."\t"."SIZE_GROUP"."\t"."OBSERVATION_LENGTH"."\t"."TATA_BOX_LENG
TH"."\t"."NUMBER_OF_HEXAMERS"."\t"."TATA_BOX_SEQUENCE"."\t"."HEX-
35_START_POS"."\t"."HEX-35_STOP_POS"."\t"."HEX-10_START_POS"."\t"."HEX-
10_STOP_POS"."\t"."SPACE-35-
10_HEXAMERS"."\t"."2nd_PROBABILITY"."\t"."OBSERVATION_SEQUENCE"."\n";

while(<VERIFY>)
{
    chomp();
    push (@fileOUT,$_) ;
    print $fileOUT[$#fileOUT]."\n";
}
close (VERIFY);

for ($j=0;$j<=$#fileOUT;$j++)
{
   open(OUTGENE,"<$dir_output_files/$fileOUT[$j]");
   $line=1;
   $numberOfHexamers = 0;
   while (<OUTGENE>)
   {
        if ($line==1)
           {
              $gene = "$fileOUT[$j]";
              $gene =~ s/OUT\_//g;
              $gene =~ s/\.txt//g;
```

```
        open (IRFILE,"<$ARGV[2]") or die "Impossible open $ARGV[2] file: $! \n";
        while (<IRFILE>)
        {
           if ($_ =~ /^($gene)\t(FORWARD|REVERSE)\t(\d+)\t(\d+)\t(\S+)\t([\-
\d+]*)\t(\S+)/)
           {
              $direction = $2;
              $sizeGroup = $5;
           }
        }
        close (IRFILE);
           }
        if ($line==5)
        {
           chomp;
           $observation_seq = $_ ;
           $observation_length = length($observation_seq);
        }
        if ($line==6)
        {
           chomp;

           $tataLength = 0;
           $Hex_10_stop = 0;
           $Hex_10_start =0 ;
           $Hex_35_stop = 0;
           $Hex_35_start = 0;
           $diffHexamers = 0;
           $h = 0;
          $b = 0;
           $i = 0;
           @tata_positions = ();
           $tata_seq = "";
           $tata_position_seq = $_ ;
           $tata_position_seq =~ s/\s/n/g;
           @tata_positions=split(//,$tata_position_seq);

           while (<@tata_positions>)
           {
              if (($tata_positions[$b] eq '#') && ($h==1.5))
           {
                 $Hex_10_start = (-$observation_length)+$b;
                 $h = 2;
           }
              if (($tata_positions[$b] eq 'n') && ($h == 2))
           {
                 $Hex_10_stop = (-($observation_length+1))+$b;
                 $stop_TATA = $b --;
                 $h = 2.5 ;
           }
           if (($tata_positions[$b] eq '#') && ($h == 2)) { }
              if (($tata_positions[$b] eq '#') && ($h == 1)) { }
              if (($tata_positions[$b] eq 'n' ) && ($h == 1))
              {
                 $Hex_35_stop = (-($observation_length+1))+$b;
                 $h =  1.5;
              }
              if (($tata_positions[$b] eq '#') && ($h == 0))
              {
                 $h = 1 ;
```

```
                $Hex_35_start = (-($observation_length))+$b;
                $start_TATA = $b ;
              }
              $b ++;
            }
            if (($tata_positions[$#tata_positions] eq '#' ) && ($h == 2))
            {
              $Hex_10_stop = (-($observation_length+1))+$b;
              $stop_TATA = $b --;
            }
            $numberOfHexamers = int ($h);
          if ($numberOfHexamers == 2)
          {
               $tataLength = $stop_TATA - $start_TATA;
              $tata_seq = substr($observation_seq,$start_TATA,$tataLength);
              $diffHexamers = ((-$Hex_35_stop)-(-$Hex_10_start))-1 ;
            }
          if ($numberOfHexamers == 0)
          {
            $tata_seq = "* * * * * no hexamer";
              $tataLength = 0;
              $Hex_10_stop = 0;
              $Hex_10_start = 0 ;
              $Hex_35_stop = 0;
              $Hex_35_start = 0;
              $diffHexamers = 0;

          }
          if ($numberOfHexamers == 1)
          {
            $tata_seq = "* * * * * unique hexamer";
              $tataLength = 0;
              $Hex_10_stop = 0;
              $Hex_10_start = 0 ;
              $Hex_35_stop = 0;
              $Hex_35_start = 0;
              $diffHexamers = 0;
          }
          }
          if ($line==9)
          {
            chomp;
            $Second_probability = $_ ;
          }
          $line++;
    }
    close (OUTGENE);
print REPORT
"$gene"."\t"."$direction"."\t"."$sizeGroup"."\t"."$observation_length"."\t"."$tataLengt
h"."\t"."$numberOfHexamers"."\t"."$tata_seq"."\t"."$Hex_35_start"."\t"."$Hex_35_stop"."
\t"."$Hex_10_start"."\t"."$Hex_10_stop"."\t"."$diffHexamers"."\t"."$Second_probability"
."\t"."$observation_seq"."\n";
}

close (REPORT);
exit;
```

Apêndice F: Programa make_mix_with_data_annotation.pl

```
#!/usr/local/bin/perl
# File: make_mix_with_data_annotation.pl
#

if ($#ARGV != 1) {
    print "Usage: $0 <TATA_FINAL_REPORT_FILE> <ORIGINAL_ANNOTATION_GENES_FILE> \>
         <OUTPUT_PATH>"."\n";
    exit(1);
}
print
         "GENE"."\t"."DIRECTION"."\t"."SIZE_GROUP"."\t"."OBSERVATION_LENGTH"."\t"."
         TATA_BOX_LENGTH"."\t"."NUMBER_OF_HEXAMERS"."\t"."TATA_BOX_SEQUENCE"."\t"."
         HEX-35_START_POS"."\t"."HEX-35_STOP_POS"."\t"."HEX-
         10_START_POS"."\t"."HEX-10_STOP_POS"."\t"."SPACE-35-
         10_HEXAMERS"."\t"."2nd_PROBABILITY"."\t"."OBSERVATION_SEQUENCE"."\t"."PROD
         UCT"."\t"."PRIMARY_CATEGORY"."\t"."GENE_START_POS"."\t"."GENE_STOP_POS"."\
         n";
#Open the TATA_FINAL_REPORT_FILE for search in EXPRESSION_GENES_FILE
open (TATA,"<$ARGV[0]") or die "Could not open input file $ARGV[0]: $!\n";

while (<TATA>)
{
         #### get Gene and the rest of line
         if ($_ =~ /^(\S+)\t([\w\W]*$)/)
         {
          $gene = $1;
          $restline = $2;
          chomp $restline ;
          # Open the EXPRESSION_GENES_FILE (SECOND file)
          open (ANNOTATION,"<$ARGV[1]") or die "Could not open input file $ARGV[1]:
         $!\n";

          $found = 0 ;

          while (<ANNOTATION>)
          {
             $line = $_ ;
             chomp $line;
             if ($line =~
         /^$gene\t\d*\t\d*\t(\d+)\t(\d+)\t[\w\s]*\t[\w\s]*\t[\w\s\W]*\t([\s\w\W]*)\
         t([\s\w\d\.\-]*)\t[\S\s]*$/ )
             {
                print
         "$gene"."\t"."$restline"."\t"."$3"."\t"."$4"."\t"."$1"."\t"."$2"."\n";
                $found = 1 ;
             }
          }
          close (ANNOTATION);

          if ($found == 0)
          {
             print "ERROR"."\n";
          }
         }
}
close (TATA);
exit 0;
```

Apêndice G: Programa cross_experimental_data.pl

```
#!/usr/local/bin/perl
#
# File: cross_experimental_data.pl
#


if ($#ARGV != 4) {
    print "Usage: $0 <MIX_TATA_FINAL_REPORT_FILE> <EXPRESSION_GENES_FILE_1>
<EXPRESSION_GENES_FILE_2> <EXPRESSION_GENES_FILE_3> <EXPRESSION_GENES_FILE_4>\>
<OUTPUT_PATH>"."\n";
    exit(1);
}

print
"N.A."."\t"."XAM1."."\t"."3_DPI"."\t"."5_DPI"."\t"."GENE"."\t"."DIRECTION"."\t"."SIZE_G
ROUP"."\t"."OBSERV_LENGTH"."\t"."TATABOX_LENGTH"."\t"."HEXAMERS_NR"."\t"."TATABOX_SEQ".
"\t"."-35_START_POS"."\t"."-35_STOP_POS"."\t"."-10_START_POS"."\t"."-
10_STOP_POS"."\t"."-35-
10_SPACE"."\t"."2nd_PROBABILITY"."\t"."OBSERVATION_SEQUENCE"."\t"."PRODUCT"."\t"."PRIMA
RY_CATEGORY"."\t"."GENE_START_POS"."\t"."GENE_STOP_POS"."\n";;

#Open the TATA_FINAL_REPORT_FILE for search in EXPRESSION_GENES_FILE
open (TATA,"<$ARGV[0]") or die "Could not open input file $ARGV[0]: $!\n";

while (<TATA>)
{
   if ($_ =~ /^(\S+)\t([\w\W]*$)/)
   {
      $gene = $1;
      $leftline = $2;
      chomp $leftline ;

      # Open the EXPRESSION_GENES_FILE_1
      open (EXP1,"<$ARGV[1]") or die "Could not open input file $ARGV[1]: $!\n";
      while (<EXP1>)
      {
         $line = $_ ;
         chomp $line;
         if (($line =~ /$gene/))
         {
            $col1 = $gene ;
            last;
         }
         else
         {
            $col1 = "no" ;
         }
      }
      close (EXP1);

      # Open the EXPRESSION_GENES_FILE_2
      open (EXP2,"<$ARGV[2]") or die "Could not open input file $ARGV[2]: $!\n";
      while (<EXP2>)
      {
         $line = $_ ;
         chomp $line;
```

```
         if (($line =~ /$gene/))
         {
            $col2 = $gene ;
            last;
         }
         else
         {
            $col2 = "no" ;
         }
      }
      close (EXP2);

      # Open the EXPRESSION_GENES_FILE_3
      open (EXP3,"<$ARGV[3]") or die "Could not open input file $ARGV[3]: $!\n";
      while (<EXP3>)
      {
         $line = $_ ;
         chomp $line;
         if (($line =~ /$gene/))
         {
            $col3 = $gene ;
            last;
         }
         else
         {
            $col3 = "no" ;
         }
      }
      close (EXP3);

      # Open the EXPRESSION_GENES_FILE_4
      open (EXP4,"<$ARGV[4]") or die "Could not open input file $ARGV[4]: $!\n";
      while (<EXP4>)
      {
         $line = $_ ;
         chomp $line;
         if (($line =~ /$gene/))
         {
            $col4 = $gene ;
            last;
         }
         else
         {
            $col4 = "no" ;
         }
      }
      close (EXP4);
      print
"$col1"."\t"."$col2"."\t"."$col3"."\t"."$col4"."\t"."$gene"."\t"."$leftline"."\n";
   }
}
close (TATA);
exit 0;
```