Laboratório **N**acional de **C**omputação **C**ientífica – **LNCC**
Programa de Pós-Graduação em Modelagem Computacional
Curso de Mestrado em Modelagem Computacional com Ênfase em Bioinformática
Biologia Computacional

IMPLEMENTAÇÃO DE UM BANCO DE DADOS DE PROTEOMAS DE BACTÉRIAS ASSOCIADAS A PLANTAS: PROBACTER

Por

Fernanda Nascimento Almeida

sob orientação da

Profa. Dra. Claudia Barros Monteiro-Vitorello

e sob co-orientação da

Profa. Dra. Ana Tereza Ribeiro de Vasconcelos

Março de 2007

Petrópolis, RJ - Brasil

IMPLEMENTAÇÃO DE UM BANCO DE DADOS DE PROTEOMAS DE BACTÉRIAS ASSOCIADAS A PLANTAS: PROBACTER

Fernanda Nascimento Almeida

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MODELAGEM COMPUTACIONAL COM ÊNFASE EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL.

Aprovada por:

_____

Profa. Claudia Barros Monteiro-Vitorello, D.Sc.

(Presidente)

_____

Profa. Ana Tereza Ribeiro de Vasconcelos, D.Sc.

_____

Prof. Fabiano Lopes Thompson, Ph.D.

_____

Profa. Marie-Anne Van Sluys, Ph.D.

PETRÓPOLIS, RJ - BRASIL

MARÇO DE 2007

ALMEIDA, FERNANDA N.

Implementação de um Banco de Dados de Proteomas de Bactérias Associadas a Plantas: ProBacter

[Petrópolis] 2006

XIV, 88, 29,7cm (MCT/LNCC, M.Sc., Modelagem Computacional com Ênfase em Bioinformática e Biologia Computacional, LNCC.

Tese – Laboratório Nacional de Computação Científica, LNCC.

1. Genômica Comparativa de Genomas Bacterianos
2. Banco de Dados Genômico de Bactérias

I. MCT/LNCC II. Título (Série)

Às minhas saudosas e amadas avós
Nellyr e Yolanda,
DEDICO.

AGRADECIMENTOS

Gostaria de deixar meus agradecimentos aos amigos queridos e acima de tudo companheiros, em especial a Saul, Mônika, Chandra, Reinaldo e Luiz Gonzaga pela contribuição que me deram, fornecendo apoio durante os momentos alegres e mais difíceis no período desta dissertação.

Agradeço a minha querida orientadora que se tornou uma grande amiga, a Profa. Dra. Claudia Barros Monteiro-Vitorello, por sempre ter acreditado em mim e no meu trabalho, e por ter compartilhando comigo uma "casquinha" do seu conhecimento.

Agradeço a Profa. Dra. Ana Tereza Ribeiro de Vasconcelos por ter me incorporado ao Laboratório de Bioinformática e me dado esta oportunidade.

Aos técnicos do Laboratório de Bioinformática e ao suporte do Laboratório Nacional de Computação Científica que forneceram assistência e dividiram comigo algumas de suas experiências.

As secretárias da pós-graduação do Laboratório Nacional de Computação Científica, Ana Paula e Ana Néri, pela presteza, profissionalismo, dedicação e carinho sempre oferecidos.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de mestrado e ao Laboratório Nacional de Computação Científica (LNCC) pela grande oportunidade.

Especialmente aos meus pais, Marcio e Tereza, pelo amor incondicional e por estarem ao meu lado em todos os momentos da minha vida. Não poderia esquecer da minha querida irmã, Caroline, por ser minha melhor e eterna amiga e também por ter me presenteado com o meu sobrinho e afilhado, Victor Hugo, uma "figurinha" que hoje tem apenas dois aninhos e me enche de felicidade e orgulho.

E a Deus pela vida, fé e paz.

Resumo da Dissertação apresentada ao MCT/LNCC como parte dos requisitos necessários para a obtenção de grau de Mestre em Ciências (M.Sc.)

IMPLEMENTAÇÃO DE UM BANCO DE DADOS DE PROTEOMAS DE BACTÉRIAS ASSOCIADAS A PLANTAS: PROBACTER

Fernanda Nascimento Almeida

26 de Março de 2007.

Orientadora: Profa. Dra. Claudia Barros Monteiro-Vitorello.
Co-orientadora: Profa. Dra. Ana Tereza Ribeiro de Vasconcelos.

Programa: Modelagem Computacional

Esta dissertação resultou na implementação de uma abordagem computacional para a análise comparativa entre informações de genomas completamente seqüenciados de bactérias associadas à planta. O sistema desenvolvido foi denominado de Probacter e é composto de um banco de dados relacional e de ferramentas computacionais para a análise de seqüências, teve por finalidade agrupar as informações disponíveis em vários bancos de dados em um único ambiente, oferecer uma padronização às informações disponibilizadas e fornecer ferramentas para análises comparativas e de seqüências. O banco de dados contém informações provenientes de diversas fontes, incluindo as bases GenBank, Swiss-Prot, TrEMBL, Interpro, COG e GO. As proteínas foram organizadas dentro de grupos, utilizando a metodologia de BBH (*Bidirectional Best Hit*) e a anotação padronizada de acordo com a classificação funcional anteriormente descrita para o Projeto Genoma de bactérias do gênero *Xanthomonas*. Cada entrada disponibilizada pelo sistema numa interface amigável corresponde a uma ficha contendo informações sobre o gene e a proteína por ele codificada, incluindo a categorização funcional, a predição de domínios, a seqüência de aminoácidos da proteína, a ligação com os grupos gerados pelo BBH, referências direta a outros bancos

de dados, e as publicações científicas. O sistema oferece uma interface de busca comum a bancos de dados, utilizando consultas pré-definidas. Para consultas mais elaboradas, foi desenvolvida uma interface para ser utilizada sem que o usuário tenha conhecimento prévio de linguagens como SQL e/ou da arquitetura desta base. Ferramentas de alinhamento múltiplo ClustalW e T-Coffee e o programa BLASTP também foram integradas a este sistema, permitindo que sejam feitas comparações entre seqüências internas e externas ao banco. O ProBacter integra ferramentas de visualização gráfica, que permite disponibilizar o posicionamento dos genes pertencentes a grupos no genoma de cada organismo e que permite visualizar as ligações durante a formação dos grupos formados pelo BBH. Por fim, um campo aberto é disponibilizado para que seja possível a intervenção de usuários na anotação de novas informações em determinada entrada, sendo as informações novas oferecidas gravadas diretamente no banco de dados.

Dissertation Abstract Presented to MCT/LNCC as a Partial Fulfillment of the Requirements for the degree of Master of Science (M.Sc.)

IMPLEMENTATION OF A PLANT-ASSOCIATED BACTERIA PROTEOME DATABASE: PROBACTER

Fernanda Nascimento Almeida

March 26st, 2007.

Advisor: Claudia Barros Monteiro-Vitorello, D.Sc.

Co-advisor: Ana Tereza Ribeiro de Vasconcelos, D.Sc.

Department: Computational Modeling

This dissertation offers a computation approach to comparative analysis between completely sequenced genomes of plant-associated bacteria. The created system was denominated ProBacter and it is composed of a relational database and computational tools for sequence analysis. The database was created from a diverse data source, including information from GenBank, TrEMBL, Interpro, COG and GO. The proteins were organized into clusters through the BBH (Bidirectional Best Hits) methodology and categorized according to the functional classification of the *Xanthomonas* Genome Project. Each entry displayed by the system in a friendly user interface corresponds to an information sheet with the gene and protein sequence, functional category, domain prediction, and related scientific publications, in addition to the group that it belongs, and external links. The system offers a search interface similar to other database systems with pre-formatted queries. For advanced queries, the user has access to an interface that can be used without previous knowledge of the SQL language or ProBacter's database arquiteture. The BLASTP program and two multiple sequence alignment tools, namely ClustalW and T-Coffee, were integrated

into the system as well, allowing internal and external sequence comparison. In addition, the system makes available visualization tools capable of displaying the gene position inside a genome and BHH links of clusters. Also, the user is capable of adding new information for each gene in the system. ProBacter's goal is to collect information available from a large source of databases into one computational environment, organize this information and offer comparative tools for sequence analysis.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O grande volume de dados gerados da análise de seqüência de genomas levou a novas aplicações para a informática no contexto da biologia. Hoje, de acordo com o banco de dados GOLD[1] (*Genomes Online Database*) existem 506 projetos com genomas totalmente seqüenciados, e um total de 2.412 projetos de seqüenciamento estão em andamento (estatística de fevereiro de 2007). Entre os organismos com genomas completamente seqüenciados estão as bactérias associadas às plantas[2]**.** Estas bactérias vivem durante parte ou todo seu ciclo de vida, em contato com diferentes partes das plantas, podendo ser restrita aos vasos xilemáticos, viver entre e/ou dentro das células de diversos tecidos ou ainda na parte aérea ou na raiz. Algumas destas bactérias são patogênicas, sendo responsáveis por causar doenças que ocasionam perdas economicamente relevantes para a agricultura, tais como: o cancro que manifesta-se através de lesões em folhas, frutos e ramos; a murcha, principal doença vascular de plantas, fazendo com que a planta murche e seque completamente, em estágios avançados da doença; ou o raquitismo, que na cana-de-açúcar causa atrofia dos colmos e internódios curtos , reduzindo a produtividade agrícola; outras estabelecem uma relação de simbiose, como é o caso das bactérias fixadoras de nitrogênio.

Organismos de espécies relacionadas são muitas vezes escolhidos para projetos de seqüenciamento, porque a comparação entre as seqüências tem o potencial de revelar o conteúdo diferencial de cada genoma. É neste conteúdo específico que podem estar presentes os genes associados à patogenicidade da bactéria, aqueles que estão potencialmente envolvidos na interação com os seus hospedeiros e também, aqueles

[1] Disponível em http://www.genomesonline.org (Último acesso em 11/02/2007).

[2] Bactérias associadas a plantas, são definidas neste trabalho, como àquelas que vivem em contato com plantas.

necessários à sobrevivência no ambiente em que vivem. Nas análises comparativas, ferramentas computacionais são fundamentais para promover o processamento e o cruzamento das informações biológicas geradas em projetos independentes. Por exemplo, a criação de bases de dados permite o armazenamento, a administração, a extração e a difusão generalizada e sistemática da informação biológica, disponibilizando, ferramentas computacionais desenvolvidas para atualizar, pesquisar e recolher dados armazenados no sistema[3]. Assim, maximizando a quantidade de informações biológicas extraídas a partir das seqüências de DNA facilitando a comparação dos dados. Durante o processo de anotação (etapa onde as informações biológicas são associadas às seqüências de aminoácidos), cada projeto define a melhor maneira de classificar os genes preditos em categorias funcionais. De modo geral, os campos de anotação não são padronizados, dificultando a comparação direta entre os dados gerados de forma independente. A análise comparativa pode ser facilitada quando as informações estão organizadas em um único ambiente, possibilitando o acesso livre e fácil à informação (através da Internet) e disponibilizando um método para extrair a informação necessária.

O objetivo deste trabalho foi implementar uma base de dados para armazenar as seqüências e as informações biológicas (vias metabólicas, dados sobre a anotação dos genes, dentre outros), extraíndo-as a partir de projetos genoma e bancos de dados públicos. A intenção é auxiliar na administração e organização destes dados, permitindo o estabelecimento de relações entre parâmetros que foram definidos independentemente para a classificação funcional dos genes, a entrada de algoritmos de análise e que evitasse a redundância dos dados armazenados. A base de dados criada como resultado

[3] Definimos aqui como sistema os bancos de dados que integraram ferramentas externas, como as de alinhamento de seqüências e de visualização, e as disponibilizam publicamente.

deste trabalho denominada de ProBacter (Proteomas[4] de Bactérias Associadas à Plantas) foi especificamente desenvolvida para conter os dados da análise de genomas completamente seqüenciados de bactérias associadas à plantas. As proteínas preditas foram organizadas dentro de grupos de genes, utilizando a metodologia de BBH (*Bidirectional Best Hit*) (OVERBEEK *et al.*, 1999) e a anotação padronizada de acordo com a classificação funcional descrita para o Projeto Genoma de bactérias do gênero *Xanthomonas* (DA SILVA *et al.*, 2002). O sistema ProBacter foi criado usando MySQL, um SGBD (Sistema Gerenciador de Banco de Dados) que utiliza a linguagem SQL (*Structured Query Language*), como interface. O banco de dados foi criado para auxiliar nas análises comparativas de genomas de bactérias associadas às plantas, na tentativa de organizar e permitir o acesso eficiente e rápido a estas informações via *Web* para auxiliar pesquisadores interessados nas relações planta-bactéria no desenvolvimento de hipóteses para o estudo sobre a biologia desses organismos.

[4] O termo *Proteoma* foi usado para indicar que o genoma teve sua seqüência completamente seqüenciada e seus genes anotados.

2. REVISÃO BIBLIOGRÁFICA

Esta revisão contém dois focos principais, sendo o primeiro o de apresentar as bactérias que vivem associadas à plantas e as informações que foram geradas nos projetos de seqüenciamento de genomas completo; e o segundo referente aos bancos de dados biológicos disponíveis até o momento, considerando o tipo de informação armazenada, a abrangência e os objetivos de cada um.

2.1. Genômica de Bactérias Associadas à Plantas

As bactérias que vivem em contato com as plantas podem ser encontradas (*i*) no exterior (PRUST *et al*., 2005), na parte aérea ou na raiz (**epifíticas**), (*ii*) infectando os tecidos internos em interações que provocam doenças nos hospedeiros (**patógenos**) (GOODNER *et al*., 2001; WOOD *et al*., 2001; SALANOUBAT *et al*., 2002; BELL *et al*., 2004; JOARDAR *et al.*, 2005; FEIL *et al*., 2005; BUELL *et al*., 2003; DA SILVA *et al*., 2002; QIAN *et al.*, 2005; THIEME *et al.*, 2005; LEE *et al.*, 2005; OCHIAI *et al.*, 2005; SIMPSON *et al.*, 2000; VAN SLUYS *et al.*, 2003; MONTEIRO-VITORELLO *et al*., 2004), (*iii*) habitando o interior de tecidos vegetais, sem causar danos aparentes, sendo que em muitos casos desempenham funções importantes no processo de adaptação das plantas. Tais organismos podem promover o crescimento vegetal, conferir ao seu hospedeiro resistência a estresse e as alterações fisiológicas, ou proteger contra herbívora e organismos patogênicos (NETO *et al.*, 2004). Estão presentes em todas as espécies vegetais, permanecendo em estado de latência ou colonizando ativamente os tecidos de forma local ou sistêmica (**endofíticas**) (NIERMAN *et al*., 2001; BLATTNER *et al*., 1997; PAULSEN *et al*., 2005; NELSON *et al*., 2002; IKEDA

et al., 2003; BENTLEY *et al*., 2002), e (*iv*) ainda interagindo com o hospedeiro numa relação em que bactéria e planta podem se beneficiar mutuamente (**simbiontes**) (KANEKO *et al*., 2000; GALIBERT *et al*., 2001). Toda essa diversidade está presente também na composição de genes do genoma dessas bactérias (BINNEWIES *et al.*, 2006). Por exemplo, bactérias de vida livre apresentam muitas vezes um genoma maior com um maior número de genes (VAN SLUYS *et al.*, 2002). De maneira geral, dispõem uma quantidade maior de genes que codificam proteínas transportadoras e reguladores da transcrição (SETUBAL, MOREIRA & DA SILVA, 2005). Uma exceção interessante é a bactéria heterotrófica marinha, *Pelagibacter ubique*, também de vida-livre e que possui um genoma bastante reduzindo (1,308,759 pares de bases) (GIOVANNONI *et al.*, 2005). Genomas reduzidos estão presentes na maioria dos casos em bactérias que são parasitas intracelulares obrigatórios como descrito para o cromossomo da espécie *Mycoplasma genitalium* (FRASER *et al*., 1995), o menor genoma de bactéria seqüenciado até o momento.

Nos últimos anos, o seqüenciamento do genoma de bactérias que ocupam os mais diversos nichos teve como um dos objetivos comparar o conteúdo de genes entre eles e definir aqueles que são específicos a um determinado grupo de organismos. Assim, possibilitando a criação de hipóteses sobre os genes essenciais à sobrevivência em determinada condição, que pode ser, por exemplo, ambiental, de patogenicidade, simbiose, ou outra condição qualquer de interesse.

Fleischmann e colaboradores (1995) foram responsáveis pelo primeiro seqüenciamento completo do genoma de um organismo procarioto, a bactéria *Haemophilus influenzae,* causadora de doenças em humanos. Desde então, o número total de genomas completamente seqüenciados cresceu rapidamente (Figura 1). Esse avanço foi possível em grande parte devido à automação dos métodos de

seqüenciamento e a estratégia conhecida como *Shotgun* que se baseia em fragmentar o genoma em pedaços pequenos possibilitando o seqüenciamento (FLEISCHMANN *et al.*, 1995), que permitiu uma redução no tempo e nos custos do seqüenciamento de genomas.

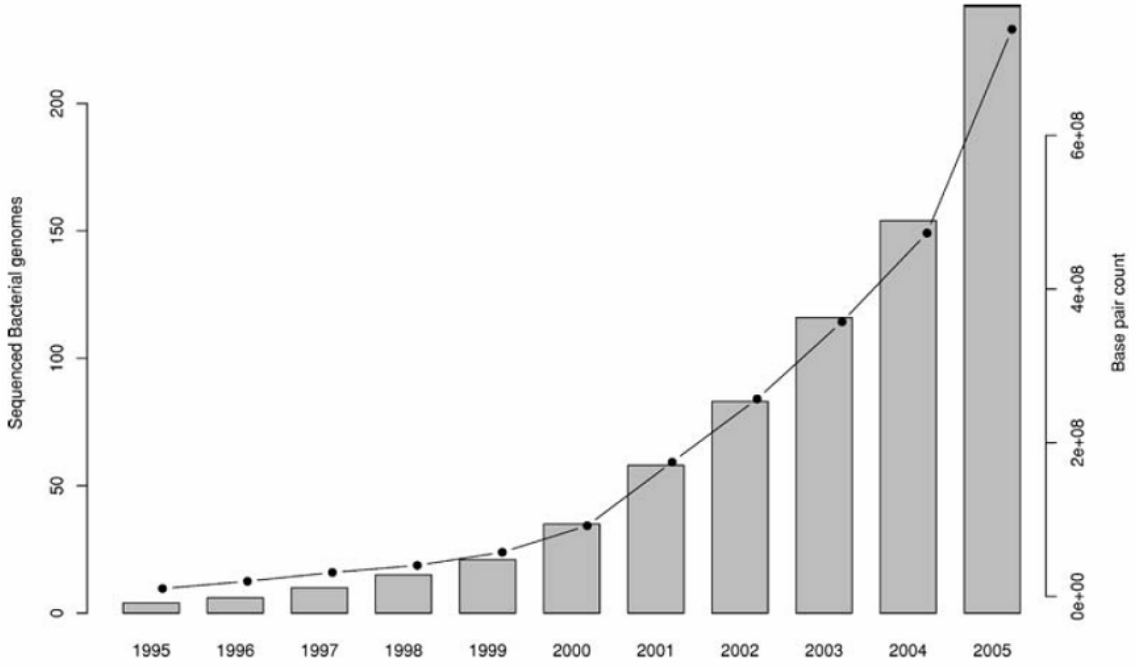

Figura 1. Número de genomas publicados na década passada (1995 – 2005) e número total de pares de bases (BINNEWIES *et al*., 2006).

Entre os organismos que tiveram seu genoma completamente sequenciado, a primeira bactéria com importância econômica para a agricultura a ter a sua seqüência de DNA determinada foi o agente causador da clorose variegada dos citrus (CVC) vulgarmente conhecida como praga do amarelinho, *Xylella fastidiosa* (SIMPSON *et al.*, 2000). Hoje, existem 18 bactérias associadas às plantas com genomas completamente seqüenciados (Tabela 1) e a Tabela 2 resume as principais características estruturais desses genomas.

Tabela 1. Listas dos organismos utilizados como grupo de estudo deste trabalho.

Organismo	Grupo taxonômico	Coloração Gram	Doença ou Aplicações	Hospedeiro Principal	Referência
Agrobacterium tumefaciens C58 (Cereon e Wash)	Alpha-proteobacteria	-	Galha-da-coroa	Plantas dicotiledôneas em geral.	GOODNER *et al.*, 2001 e WOOD *et al.*, 2001
Mesorhizobium loti MAFF303099	Alpha-proteobacteria	-	Fixadora de nitrogênio	Plantas leguminosas	KANEKO *et al.*, 2000
Sinorhizobium meliloti 1021	Alpha-proteobacteria	-	Fixadora de nitrogênio	Plantas leguminosas	GALIBERT *et al.*, 2001
Ralstonia solanacearum GMI1000	Beta-proteobacteria	-	Murcha Bacteriana	Batateiros	SALANOUBAT *et al.*, 2002
*Erwinia carotovora s*ubsp. *atroseptica* SCRI1043	Gamma-proteobacteria	-	Talo oco dos batateiros	Batateiros	BELL *et al.*, 2004
Pseudomonas syringae phaseolicola 1448[a]	Gamma-proteobacteria	-	Podridão dos feijoeiros	Feijoeiro	JOARDAR *et al.*, 2005
Pseudomonas syringae pv. B728a	Gamma-proteobacteria	-	Queima bacteriana do feijão	Feijoeiro	FEIL *et al.*, 2005
Pseudomonas syringae pv. *tomato* DC3000	Gamma-proteobacteria	-	Mancha bacteriana pequena	Tomateiros/Arabidopsis	BUELL *et al.*, 2003
Xanthomonas axonopodis pv. *citri* str 306	Gamma-proteobacteria	-	Cancrose A	Plantas de citros	DA SILVA *et al.*, 2002
Xanthomonas campestris pv. *campestris* str. 8004	Gamma-proteobacteria	-	Podridão negra das crucíferas	Arabidopsis e Brassica	QIAN *et al.*., 2005
Xanthomonas campestris pv. *campestris* str. ATCC33913	Gamma-proteobacteria	-	Podridão negra das crucíferas	Arabidopsis e Brassica	DA SILVA *et al.*, 2002
Xanthomonas campestris pv. *vesicatoria* str. 85-10	Gamma-proteobacteria	-	Mancha bacteriana	Pimenteiro e tomateiro	THIEME *et al*., 2005
Xanthomonas oryzae pv. *oryzae* KACC10331	Gamma-proteobacteria	-	Estria bacteriana do arroz	Plantas de arroz	LEE *et al*., 2005
Xanthomonas oryzae pv. *oryzae* MAFF 311018	Gamma-proteobacteria	-	Estria bacteriana do arroz 1	Plantas de arroz	OCHIAI *et al*., 2005
Xylella fastidiosa 9a5C	Gamma-proteobacteria	-	Clorose variegada dos citros	Plantas de citros	SIMPSON *et al*., 2000
Xylella fastidiosa Temecula1	Gamma-proteobacteria	-	Doença de Pierce	Videiras	VAN SLUYS *et al*., 2003
Leifsonia xyli subsp. *xyli* str. CTCB07	Actinobacteria	+	Raquitismo da soqueira	Cana-de-açúcar	MONTEIRO-VITORELLO *et al*., 2004

Tabela 2. Características gerais dos genomas listados na Tabela 1.

	Composição do Genoma							
Organismos	**Tamanho (Mpb)**	**Conteúdo de CG (%)**	**Seqüências codificadoras**	**com função putativa**	**hipotéticos**	**operon de rRNA**	**tRNA**	**Plasmídeos**
A. tumefaciens C58 (Cereon e Wash)	5,7	58	5.419	1944	708	4	53	2
M. loti	2,7	52	6.752	3.799	1.759	2	50	2
S. meliloti	6,7	62	5.748	4.155	548	3	54	2
R. solanacearum	5,8	66	5.129	2.261	848	4	58	1
*E. carotovora s*ubsp. *atroseptica*	5,0	50	4.491	3.217	1.255	7	76	-
P. syringae phaseolicola 1448A	5,9	57	5.170	3.948	1.192	5	64	2
P. syringae pv. B728a	6,1	59	5.136	3.840	1.297	5	64	-
P. syringae pv. *tomato* DC3000	6,5	58	5.615	3.402	610	5	63	2
X. axonopodis pv. *citri* str 306	5,3	65	4.428	2.770	1.658	2	54	2
X. campestris 8004	5,2	65	4.273	2.671	1.602	2	53	-
X. campestris ATCC33913	5,1	65	4.182	2.708	1.474	2	53	-
X. campestris pv. *vesicatoria* str. 85-10	5,0	64	4.637	3.340	1.297	2	56	4
X. oryzae KACC10331	5,0	64	4.637	3.340	1.297	2	54	-
X. oryzae MAFF 311018	4,8	63	4.372	2.776	1.596	2	53	-
X. fastidiosa 9a5C	2,8	53	2.848	1314	1.534	2	49	2
X. fastidiosa Temecula1	2,5	52	2.066	1.362	704	2	49	1
L. xyli pv. *Xyli*	2,5	68	2.044	307	321	1	45	-

A árvore filogenética (figura 2) foi gerada com base na região 16S rRNA do gene ribossomal dos organismos utilizados no banco de dados, juntamente com o genoma completo de espécies consideradas modelo em estudos biológicos e que representam cada um dos grandes grupos taxonômicos foram adicionados ao trabalho (Tabela 3). Entre elas estão *Escherichia coli* (BLATTNER *et al.*, 1997) para bactérias Gram-negativas a *Mycobacterium tuberculosis* (COLE *et al.*, 1998) para Gram-positivas.

A maioria das bactérias pertence ao grupo taxonômico Proteobacteria, sendo a classe Gamma a mais representada (12 dessas bactérias). Proteobacteria compreende um grupo de bactérias Gram-negativas que possuem grande diversidade de estilo de vida, ecologia e metabolismo (STUDHOLMES *et al.*, 2004). O grupo taxonômico Actinobactéria que engloba bactérias Gram-positivas é o menos representativo, entretanto, compreende a bactéria fastidiosa, *Leifsonia xyli* (MONTEIRO-VITORELLO *et al*., 2004). Este organismo é o único que esta no banco de dados, até o momento, que representa as Gram-positivas causadoras de doenças em plantas e que teve seu genoma completamente seqüenciado.

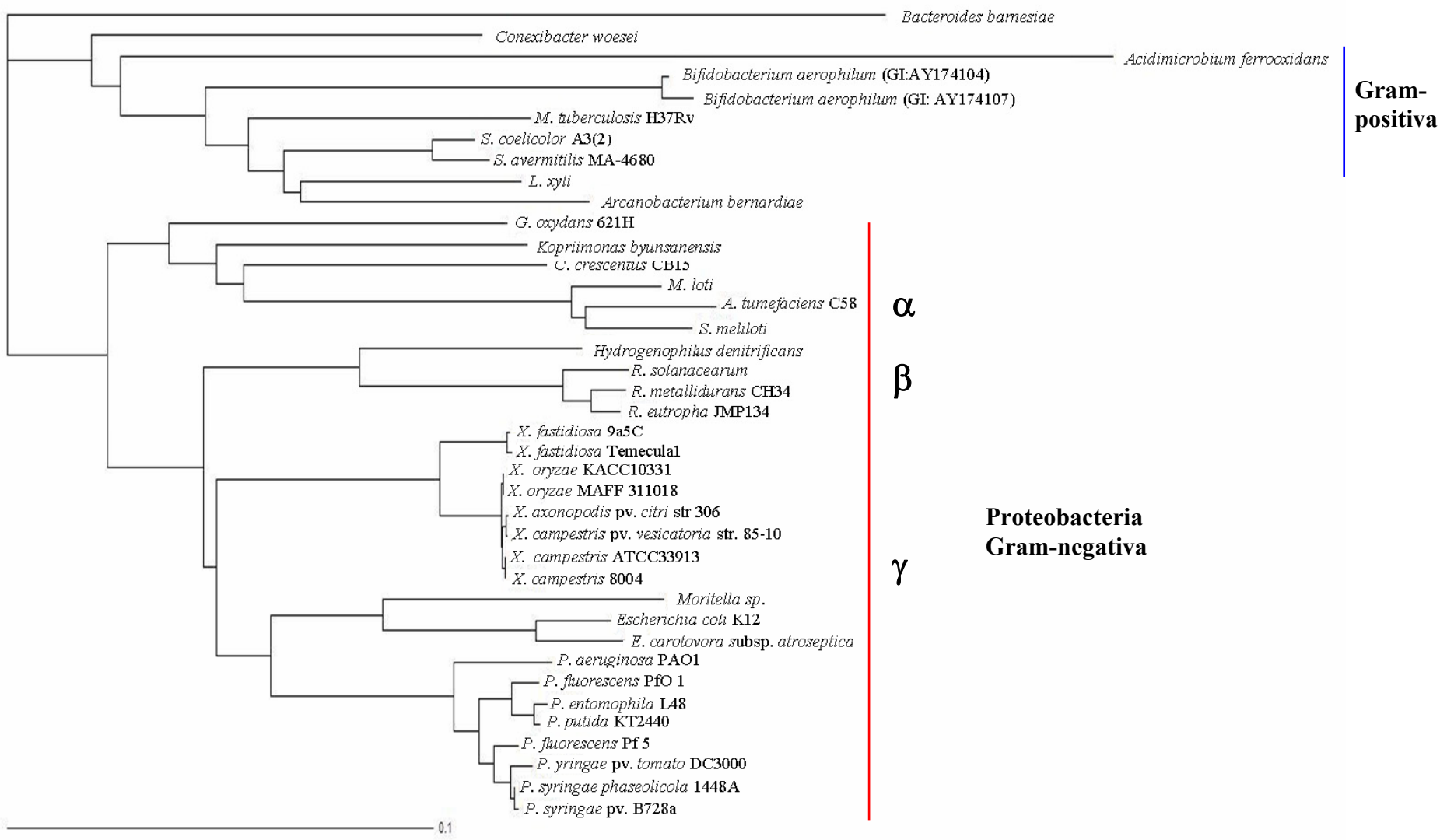

Figura 2. Árvore filogenética das regiões SSU rRNA (16S) do gene ribossomal dos organismos que estão hoje incluídos no sistema ProBacter. A mesma foi construída usando o programa PAUP e consenso neighbor-joining, as bactérias: *Bacteroides barnesiae, Conexibacter woesei, Acidimicrobium ferrooxidans, Bifidobacterium aerophilum, Bifidobacterium aerophilum, Arcanobacterium bernardiae, Kopriimonas byunsanensis, Kopriimonas byunsanensis, Moritella sp*. foram utilizadas para calibrar a árvore filogenética.

Tabela 3. Listas dos organismos utilizados como referência em análises comparativas.

Organismo	Grupo taxonômico	Linhagem Gram	Doença ou Aplicações	Hospedeiro Principal	Referências
Caulobacter crescentus CB15	Alpha-proteobacteria	-	Vida-livre	-	NIERMAN et al., 2001
Escherichia coli K12	Gamma-proteobacteria	-	Vida-livre	-	BLATTNER et al., 1997
Gluconobacter oxydans 621H	Alpha-proteobacteria	-	Epifítica	-	PRUST et al., 2005
Mycobacterium tuberculosis H37Rv	Actinobacteria	+	Tuberculose	Humanos	COLE et al., 1998
Pseudomonas aeruginosa PAO1	Gamma-proteobacteria	-	Infecção oportunista	Humanos	STOVER et al., 2000
Pseudomonas entomophila L48	Gamma-proteobacteria	-	Morte de Insetos	Insetos	VODOVAR et al., 2006
Pseudomonas fluorescens Pf 5	Gamma-proteobacteria	-	Vida-livre	-	PAULSEN et al., 2005
Pseudomonas fluorescens PfO 1	Gamma-proteobacteria	-	Vida-livre	-	Não publicada
Pseudomonas putida KT2440	Gamma-proteobacteria	-	Vida-livre	-	NELSON et al., 2002
Ralstonia eutropha JMP134	Beta-proteobacteria	-	Vida-livre	-	Não publicada
Ralstonia metallidurans CH34	Beta-proteobacteria	-	Vida-livre	-	Não publicada
Streptomyces avermitilis MA-4680	Actinobacteria	-	Vida-livre	-	IKEDA et al., 2003
Streptomyces coelicolor A3(2)	Actinobacteria	-	Vida-livre	-	BENTLEY et al., 2002

2.2. Bancos de dados

Como discutido no item anterior, as informações obtidas a partir das seqüências completas dos genomas têm ajudado a entender e desenvolver hipóteses a respeito dos mecanismos de patogenicidade e sobre a biologia das bactérias que vivem associadas às plantas. O volume crescente de informações disponíveis extraídas de genomas completamente seqüenciados levou ao desenvolvimento de novos algoritmos, bancos de dados e softwares sofisticados que auxiliam nas análises e comparações entre as seqüências disponíveis e o cruzamento entre as informações biológicas inferidas a partir dessas seqüências. Entretanto, vale ressaltar que não é uma tarefa fácil criar e manter um banco de dados, pois existem dificuldades para o desenvolvimento de plataformas que consigam representar fielmente ou aproximadamente as relações que podem ser feitas entre os componentes de um ou de vários sistemas biológicos. Outra dificuldade encontrada pelos desenvolvedores de bancos de dados são os tipos de consultas que podem ser feitas, pois estas devem ter como requisito básico proporcionar ao usuário, uma interface de fácil entendimento, a execução de consultas complexas e mais elaboradas, assim como apresentar os resultados de uma maneira organizada.

Atualmente, existem 968 bancos de dados na área de biologia molecular, segundo informações publicadas na revista *Nucleics Acids Research* (NAR)[5] (GALPERIN, 2007), alguns deles podem ser vistos na Tabela 4. Esta grande diversidade de *sistemas de informação* (*ou sistemas*)[6] que integram tanto dados de seqüências genômicas como dados de seqüências de proteínas, receberam uma classificação didática pela NAR sendo que estes são subdivididos em categorias de

[5] A revista NAR (disponível em http://nar.oxfordjournals.org/) faz um apanhado dos bancos de dados de biologia molecular existentes, e publica anualmente uma coleção deles.
[6] Sistemas de informação ou sistemas mencionados neste trabalho são os bancos de dados que integram as seqüências dos genomas que desejam estudar e ferramentas computacionais para auxiliar nas análises.

acordo com a finalidade a que pertencem, como: (*i*) aqueles que armazenam todas as seqüências; (*ii*) seqüências genômicas; (*iii*) estrutura e seqüência de proteínas; (*iv*) aqueles dedicados a vias metabólicas; e (*v*) aqueles com finalidade específica.

Entre os bancos de dados mais importantes que se encaixam na primeira classificação citada e dos primeiros a serem estabelecidos estão os que armazenam as informações de seqüência de DNA e proteínas, independentes do organismo, tamanho da seqüência ou função. São as bases de dados: GenBank[7] (BILOFSKY *et al.*, 1986; BENSON *et al.*, 2007); EMBL (*European Molecular Biology Laboratory)*, também conhecido como *EMBL-Bank* (STOESSER *et al.*, 1998; KULIKOVA *et al.*, 2007); e DDBJ (*DNA Data Bank of Japan*) (SUGAWARA *et al.*, 1999, 2007).

Pertencentes a classe de seqüências genômicas, estão também os sistemas de informação que procuram integrar as informações entre os genomas de bactérias completamente seqüenciados, como o sistema CMR (*Comprehensive Microbial Resource*) (PETERSON *et al.*, 2001) do consórcio TIGR[8] (*The Institute for Genomic Research*). O Ominione é banco de dados associado ao sistema CMR que divide as informações em níveis (nível genômico, nível do gene) contendo dados sobre conteúdo de CG do DNA até propriedades químicas das proteínas, tais como, peso molecular ou hidrofobicidades, categoria funcional (retirada do sistema COG), função exercida pelo gene, grupo taxonômico ao qual o organismo pertence e ligações com outros bancos de dados públicos (PETERSON *et al.*, 2001). A comparação dos genomas no CMR pode ser feita através de buscas pré-definidas ao banco de dados. Objetivos semelhantes têm o banco de dados IMG (*The Integrated Microbial Genomes System*) (MARKOWITZ *et al.*, 2006). Este banco de dados integra e administra informações de genomas microbianos e alguns genomas selecionados de eucariotos provenientes de diversas

[7] Banco de dados de seqüência vinculado ao NCBI (*National Center for Biotechnology Information*) e este se encontra em www.ncbi.nlm.nih.gov/.
[8] TIGR disponível em www.tigr.org/ é um centro dedicado a armazenar genomas.

fontes. Desta forma, o mesmo incorpora informações de seqüências, modelos de genes preditos computacionalmente e manualmente curados, relações de similaridade pré-computadas entre as seqüências, anotação funcional e dados de vias metabólicas. Neste caso a busca aos dados pode ser realizada em três níveis: genoma (organismo), funções (termos e vias) e gene. Uma análise comparativa dos genomas é fornecida pelo sistema IMG através de um conjunto de ferramentas que permitem que os genomas sejam comparados, disponibilizando os dados estatísticos de genes organismo-específico e seqüências conservadas (MARKOWITZ *et al.*, 2006). Ainda dentro deste contexto, o banco de dados MBGD (*Microbial Genome Database for Comparative Analysis*) (UCHIYAMA, 2003) tem como objetivo a análise comparativa de genomas completos de microrganismos baseados na classificação de genes ortólogos geradas por um algoritmo de classificação hierárquico desenvolvido especificamente para este sistema (UCHIYAMA, 2007). O sistema permite que o usuário gere sua própria tabela de classificação usando o algoritmo hierárquico, bastando especificar um grupo de organismos e alguns parâmetros para que a tabela, que é armazenada temporariamente no banco de dados, possa ser explorada em detalhes. Diferente do IMG e outros bancos de dados de grupos de ortólogos construídos por processos automatizados, o MBGD permite que o usuário classifique os genes dinamicamente. (UCHIYAMA, 2007).

O banco de dados COG (*Cluster of Orthologous Groups*) (TATUSOV *et al.*, 1997, 2003) compreende grupo de proteínas preditas codificadas por genomas procarióticos e atualmente eucarióticos, cujos genomas foram completamente seqüenciados. Este sistema fornece uma classificação filogenética destas proteínas, caracterizando uma importante fonte de informação em genômica funcional e evolutiva. Permite que o usuário tenha acesso aos dados pré-computados por meio de várias páginas que são navegáveis, como por exemplo, padrões filogenéticos, classificações

funcionais, e uma lista contendo os grupos de genes ortólogos por categoria funcional ou por via metabólica e co-ocorrência em COGs. Os genes no banco de dados COG são agrupados de acordo com o critério de similaridade definido pelo algoritmo desenvolvido para o sistema, o COGNITOR (TATUSOV *et al.*, 1997). Na mesma linha de estudo, a base de dados HOBACGEN (*Homologous Bacterial Genes*), foi desenvolvida para comparar os genes em genomas de microrganismos. Este sistema contém todas as seqüências de proteínas preditas disponíveis de bactérias, provenientes de outros bancos de dados, e estas são classificadas dentro de famílias de genes homólogos determinadas pela sua similaridade (PERRIÈRE *et al.*, 2006). Entretanto, os dois sistemas de informação citados possuem limitações quanto aos tipos de busca disponíveis, fazendo com que o usuário tenha que percorrer por grande parte do banco de dados para obter sua informação, além de não ser possível fazer consultas mais complexas a base de dados.

Outros sistemas de informação que integram diferentes bancos de dados e que também reunem informações a respeito das interações moleculares em processos biológicos, dados sobre as características do gene e proteínas e dados sobre a vasta gama de componentes químicos e reações. Exemplos desse tipo de banco de dados são o MetaCyc: *A multiorganism database of metabolic pathways and enzymes* (CASPI *et al.*, 2006) e o KEGG: *Kyoto Encyclopedia of Genes and Genomes* (KANEHISA & GOTO 2000); os quais reúnem informações de genes, genomas e vias metabólicas de diversos organismos. Outros bancos de dados com informações vindas de diferentes genomas são: MicrobesOnline (ALM *et al.*, 2005), Entrez Genome (WHEELER *et al.*, 2006), PUMA2 (MALTSEV *et al.*, 2006), BacMap (STOTHARD *et al.*, 2005), dentre outros. O que pode ser visto nestes bancos de dados, assim como descrito para o COG, existem limitações quanto a complexidade de consultas.

Existem bancos dedicados a armazenar, administrar e disponibilizar informações de seqüência e função das proteínas. Com o número crescente de genomas completamente seqüenciados e de projetos que analisam o transcriptoma de uma grande variedade de organismos, a atenção está hoje voltada para a identificação e função das proteínas codificadas por esses genomas. Recentemente, o UniProt (*Universal Protein Resource*) que engloba um catálogo contendo informações de proteínas (APWEILER *et al*., 2004 e *THE UNIPROT CONSORTIUM*, 2007) foi criado e é hoje um repositório central de informações de proteínas, unificando os dados contidos no SWISS-PROT (*Swiss-Prot Knowledege Database*) (BOECKMANN *et al.*, 2003), TrEMBL (*Translated EMBL*) (BOECKMANN *et al.*, 2003), e PIR (*Protein Information Reseource*) (WU *et al.*, 2003).

O SWISS-PROT é um banco de dados que concentra as informações biológicas disponíveis a seqüência de proteína através de programas de anotação automática e predição de estrutura e domínios, e também através de uma detalhada anotação manual, dirigida por especialistas em cada um dos projetos: HPI (*Human Proteomics Initiative*), IPI (*The Internation Protein Index*), HAMAP (*High-quality Automated and Manual Annotation of Microbial Proteomes*), PPAP (*Plant Proteome Annotation Project*) e NEWT (*The New Taxonomy Database*).

O projeto HAMAP foi criado há cinco anos devido ao grande número de genomas de bactérias completamente seqüenciados que estão disponíveis em bancos de dados públicos (GATTIKER *et al.*, 2003). A cada novo genoma completo, todas as proteínas preditas traduzidas a partir da seqüência de nucleotídeos são incorporadas ao banco TrEMBL. O banco de dados TrEMBL concentra mais do que 500.000 proteínas oriundas desses genomas, que precisam ser anotadas manualmente. O HAMAP é um sistema que anota uma porcentagem de cada proteoma de bactéria automaticamente; o

sistema está direcionado para a anotação de proteínas que pertencem a famílias ou subfamílias bem definidas. O principal objetivo é produzir entradas de alta qualidade de acordo com os padrões da anotação manual UniProt/SWISS-PROT. Para alcançar esse objetivo, são criadas manualmente regras de famílias, que permitem definir o nível e a extensão da anotação que é propagada para cada membro da família por similaridade. Toda a informação é manualmente adicionada depois de uma revisão detalhada da literatura disponível (GATTIKER *et al*., 2003).

Considerando ainda a anotação de proteínas preditas, o InterPro (MULDER *et al.*, 2007) é um banco de dados que acumula informações sobre domínios, motivos e regiões conservadas nas proteínas e famílias de proteínas disponibilizadas por outros bancos de dados que incluem: PROSITE (HULO *et al.*, 2004), PRINTS (ATTWOOD *et al.*, 1999), ProDom (CORPET *et al.*, 2000), Pfam (BATEMAN *et al.*, 2004), SMART (SCHULTZ *et al*., 2000; LETUNIC *et al.*, 2006), TIGRFAMs (SELENGUT *et al.*, 2007), PIRSF (WU *et al.*, 2004), UPERFAMILY (WILSON *et al.*, 2007) e PANTHER (MI *et al.*, 2007). Todas as bases de dados citadas organizam seus dados por meio de algoritmos próprios.

Outro banco de dados que integra informações sobre proteínas, especificamente informações de estruturas tridimensionais é o PDB (*Protein Data Bank*), que contém uma coleção de estruturas de proteínas comprovadas experimentalmente de diferentes organismos, bem como as informações gerais e específicas e os métodos usados para a determinação das mesmas (BERMAN *et al*., 2000).

Existem também bancos de dados que integram informações provenientes de outros sistemas de informação, proporcionando uma visão mais analítica a respeito do tema a que se referem, como por exemplo, o banco PseudoCAP (*Pseudomonas aeruginosa Genome Database* [WINSOR *et al.*, 2005]) usado pelo Projeto Genoma da

Pseudomonas aeruginosa, para submissão de dados de anotação, acesso e consulta por informações referentes a esta bactéria, possuindo ligações com bancos de dados correlacionados. Semelhante a esse, é o banco de dados LEGER desenvolvido para análises funcionais dos genomas provenientes do projeto de seqüenciamento de diferentes linhagens de *Listeria*, com o objetivo de melhorar as anotações originais (DIETERICH *et al.*, 2006).

Com o objetivo de analisar comparativamente seqüências de genômas de bactérias associadas a plantas está o sistema PABdb (VAN SLUYS *et al*., 2002; DIAGIAMPETRI, MEDEIROS & SETUBAL, 2003). Desenvolver bancos de dados com um propósito específico é uma pratica cada vez mais comum hoje em dia, já que várias espécies de um determinado gênero ou o genoma de diversas linhagens da mesma espécie tem sido seqüenciado. Neste sentido, o PABdb (*Plant Associated Bacteria database*) foi construído para promover a análise comparativa de oito genomas de bactérias associadas às plantas. Este banco de dados usou como base para a comparação entre os genomas a formação de agrupamentos de genes em famílias utilizando o programa BLASTP (*protein – protein Basic Local Alignment Search Tool* [ALTSCHUL *et al.*, 1990]). As análises comparativas obtidas com a utilização deste banco de dados foram descrita por Van Sluys e colaboradores (2002). No entanto, não é um banco de dados disponível para o acesso público.

Foram descritos aqui sistemas de informação que possuem relevância para o presente estudo e dados adicionais dos mesmos podem ser obtidos na Tabela 3. O banco de dados criado como resultado deste projeto se encaixa na categoria por último descrita, banco de dados com objetivos específicos a determinado grupo de organismos, como os sistemas PseudoCAP, o LEGER e o PABdb. No entanto, o sistema desenvolvido além de auxiliar nas análises comparativas, integrar ferramentas

computacionais e informações de outros bancos de dados, tenta disponibilizar ao usuário métodos de consulta mais robustos e que permitam buscas mais complexas e elaborada obtendo os resultados de forma rápida e organizada.

Tabela 4. Lista de bancos de dados citados nesta revisão e disponíveis para análise comparativa de genomas microbianos.

Nome do banco de dados	Nome completo ou descrição	Referências	URL
Bancos de dados que armazenam todas as seqüências			
GenBank	Banco de dados de seqüências genômicas	BENSON *et al.*, 2007	www.ncbi.nlm.nih.gov/Genbank/
EMBL	European Molecular Biology Laboratory	KULIKOVA *et al.*, 2007	www.ebi.ac.uk/embl/
DDBJ	DNA Data Bank of Japan	SUGAWARA *et al.*, 2007	www.ddbj.nig.ac.jp/
Bancos de dados dedicados a seqüências genômicas			
CMR	Comprehensive Microbial Resource	PETERSON *et al.*, 2001	http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi
IMG	The Integrated Microbial Genomes System	MARKOWITZ *et al.*, 2006	http://img.jgi.doe.gov/cgi-bin/pub/main.cgi
MBGD	Microbial Genome Database for Comparative Analysis)	UCHIYAMA, 2003, 2007	http://mbgd.genome.ad.jp/
COG	Cluster of Orthologous Groups	TATUSOV *et al.*, 1997, 2003	www.ncbi.nlm.nih.gov/COG
Bancos de dados dedicados a vias metabólicas			
HOBACGEN	Homologous Bacterial Genes	PERRIÈRE *et al.*, 2006	http://pbil.univ-lyon1.fr/databases/hobacgen.html
KEGG	Kyoto Encyclopedia of Genes and Genomes	KANEHISA & GOTO 2000	www.genome.jp/kegg/
MetaCyc	Banco de dados de vias metabólicas e enzimas	CASPI *et al.*, 2006	http://metacyc.org/
MicrobesOnline	Ferramenta para comparar genomas procariotos	ALM *et al.*, 2005	http://www.microbesonline.org/
Entrez Genome	Banco de dados que fornece informações archeas, bactérias, vírus, eucariotos, viroids e plasmídeos	WHEELER *et al.*, 2006	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
PUMA2	Análise evolucionária do Metabolismo	MALTSEV *et al.*, 2006	http://compbio.mcs.anl.gov/puma2/
BacMap	Atlas interativo para explorar genomas microbianos	STOTHARD *et al.*, 2005	http://wishart.biology.ualberta.ca/BacMap/
Bancos de dados dedicados à estrutura e seqüência de proteínas			
UniProt	Universal Protein Resource	THE UNIPROT CONSORTIUM, 2007	www.uniprot.org/
InterPro	Dedicado a familia de proteínas, domínios e regiões funcionais	MULDER *et al.*, 2007	http://www.ebi.ac.uk/interpro/

Continua.

SWISS-PROT	Swiss-Prot Knowledege Database	BOECKMANN *et al.*, 2003	http://ca.expasy.org/sprot/
TrEMBL	Translated EMBL	BOECKMANN *et al.*, 2003	www.ebi.ac.uk/trembl/
PIR	Protein Information Resource	WU et al., 2003	http://pir.georgetown.edu/
HAMAP	High-quality Automated and Manual Annotation of Microbial Proteomes	GATTIKER *et al.*, 2003	http://ca.expasy.org/sprot/hamap/
NEWT	The New Taxonomy Database	PHAN *et al.*, 2003	http://www.ebi.ac.uk/newt/display
PROSITE	Bando de dados de domínio de proteínas, famílias e sítios ativos	HULO *et al.*, 2004	http://ca.expasy.org/prosite/
PRINTS	Banco de dados de domínio de proteínas	ATTWOOD *et al.*, 1999	http://www.bioinf.manchester.ac.uk /dbbrowser/PRINTS/
ProDom	Ferramenta para análise de dmínio de proteína e comparação de genomas completos	CORPET *et al.*, 2000	http://prodom.prabi.fr/prodom/current/html/home.php
Pfam	Protein families	BATEMAN *et al.*, 2004	www.sanger.ac.uk/Software/Pfam/
SMART	Simple Molecular Architeture Research Tool	SCHULTZ *et al.*, 2000; LETUNIC *et al.*, 2006	http://smart.embl-heidelberg.de/
TIGRFAMs	Banco de dados dedica a família de proteínas	SELENGUT *et al.*., 2007	www.tigr.org/TIGRFAMs/
PIRSF	Sistema de classificação de famílias de proteínas	WU *et al*., 2004	http://pir.georgetown.edu/pirsf/
SUPERFAMILY	Designação de homologia a seqüência genômicas usando uma biblioteca de cadeia de Markov oculta para estrutura de proteínas	WILSON *et al.*, 2007	http://supfam.org/SUPERFAMILY/
PANTHER	Protein ANalysis THrough Evolutionary Relationship	MI *et al.*., 2007	http://www.pantherdb.org/
PDB	Protein Data Bank	BERMAN *et al.*., 2000	www.rcsb.org/pdb/
Bancos de dados com finalidade específica			
PABdb	Banco de dados construído para promover analise comparativa de genomas de bactérias associadas às plantas	Não foi publicado.	Não esta disponível publicamente.
PseudoCAP	Pseudomonas aeruginosa Genome Database	WINSOR *et al.*., 2005	http://www.pseudomonas.com/
LEGER	Banco de dados para análises funcionais dos genomas do projeto de seqüenciamento de Listeria	DIETERICH *et al.*, 2006	http://leger2.gbf.de/cgi-bin/expLeger.pl

3. OBJETIVOS

O objetivo geral desse trabalho foi desenvolver um sistema computacional para auxiliar a análise comparativa dos genomas de bactérias associadas às plantas armazenadas no banco de dados. Os objetivos específicos foram:

1. Criar um banco de dados, o ProBacter, que contenha as informações das proteínas preditas nos genomas completamente seqüenciados;
2. Agrupar as proteínas em grupos funcionais;
3. Padronizar a categorização funcional das proteínas com base na classificação descrita para o Projeto Genoma da *Xanthomonas*;
4. Disponibilizar as informações geradas publicamente através de uma interface amigável.

4. METODOLOGIA

Neste capítulo será apresentado o banco de dados ProBacter, que é uma base de dados sobre a qual o estudo desta dissertação foi aplicado. Um modelo de dados deve conter todas as entidades e atributos necessários assim como os relacionamentos existentes entre estas entidades para que os genomas possam ser armazenados e posteriormente comparados. Neste banco de dados foram definidas as seguintes entidades: organismos e genes (Figura 3). Estas entidades formam a base utilizada para desenhar o modelo do banco de dados, que atendeu as necessidades deste trabalho.

Figura 3. Diagrama Entidade e Relacionamento (ER) do primeiro modelo de dados do ProBacter. ***Organisms*** refere-se ao organismo que teve o genoma seqüenciado e ***Gene_card*** refere-se às informações sobre o gene, obtidas dos projetos de seqüenciamento.

4.1. O Modelo de Dados

Um banco de dados foi desenvolvido com base no modelo demonstrado anteriormente que expressa um relacionamento entre duas entidades, onde cada uma destas entidades dá origem a uma tabela ou relação. Assim obteve-se: ***Organisms*** (contendo informações referentes à bactéria) e ***Gene_card*** (constituído de dados do conjunto de genes pertencentes ao genoma) (Figura 4).

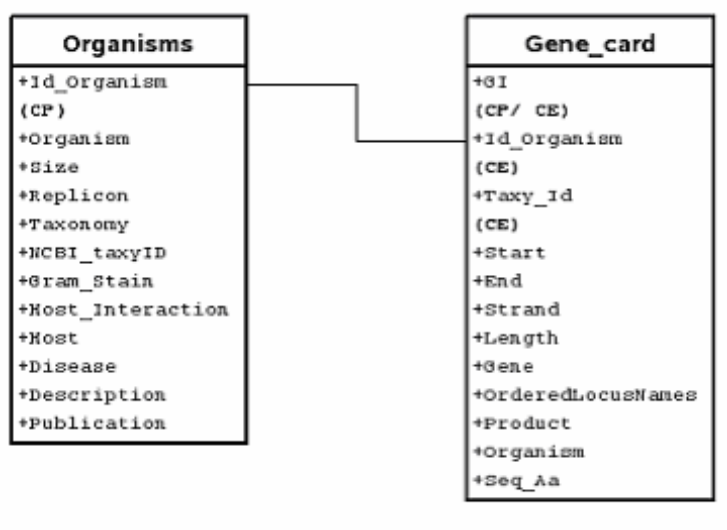

Figura 4. Modelo de dados usado como base para o sistema ProBacter. CP representa as *chaves primárias* e CE representam as *chaves estrangeiras*. O campo GI esta indicado também como chave estrangeira, pois estabelecerá ligações entre as outras tabelas que serão descritas a seguir.

A definição dos atributos de cada entidade é incluída em uma tabela que armazena informações relevantes sobre estes genomas. A tabela ***Organisms*** inclui informações sobre as características individuais do genoma das bactérias. A seleção de seus atributos foi feita considerando as características biológicas de cada organismo como nos trabalhos publicados que apresentam os dados de seqüenciamento. Assim, foram obtidos os seguintes campos: nome do organismo (*Organism*), replicons (*Replicon*), grupo taxonômico a que pertence (*Taxonomy*), bem como seu número de identificação (*NCBI_Taxy_Id*), tipo de coloração Gram (*Gram_Stain*), tamanho do genoma (*Size*), tipo de interação com o hospedeiro (*Host_Interaction*), doença causada (quando patogênico) (*Disease*), tipo de hospedeiro (*Host*), breve descrição da bactéria (*Description*), principal publicação (*Publication*) e número de identificação do organismo (*Id_Organism*).

A tabela ***Gene_card*** é composta pelas informações do conjunto de genes que compõem um dado genoma. O procedimento para definir os atributos foi baseado no banco de dados do NCBI, que possui uma coleção de genomas completamente seqüenciados, incluindo dados de anotação como: posição do gene dentro do genoma (*Start* e *End*), orientação do gene (*Strand*), tamanho do gene (*Length*), nome do gene (*Gene*), sinônimo (*OrderedLocusNames*), produto (*Product*), nome do organismo (*Organism*), seqüência de aminoácidos da proteína (*Seq_Aa*), número de identificação do gene (*GI*), e o número taxonômico (*Taxy_Id*) e do organismo (*Id_Organism*).

Para compreender um banco de dados baseado no modelo relacional, é de fundamental importância conhecer e entender o funcionamento dos diferentes tipos de "chaves" que podem ser usadas. Quando se faz menção ao termo chave primária (CP) está se referindo ao menor conjunto de campos/atributos de uma tabela que a representam univocamente; ou seja, cada um dos registros de uma relação possui somente um significado ou interpretação. Um atributo que se repete em diferentes tabelas pode ser considerado redundante, porém esta redundância existe com um propósito: ela auxilia em buscas onde se faz necessário consultar diferentes tabelas. Estes atributos podem ser definidos como chave estrangeira (CE), que é formada através de um relacionamento com a chave primária de outra tabela, definindo assim, um relacionamento entre tabelas, podendo esta se repetir muitas vezes (Figura 4).

4.2. O Sistema ProBacter

O banco de dados ProBacter foi implementado usando o SGBD MySQL versão 3.23.46. As interfaces gráficas do sistema foram hospedadas em um servidor SUN-Fire e com servidor *Web* Apache. A conexão com o banco de dados foi feita através de

scripts programados em PERL versão 5.6.1. Este sistema tem como objetivo comparar genomas de bactérias associadas às plantas. Hoje são administrados 31 genomas completos, incluindo bactérias de diversos nichos ecológicos como: patogênicas e não patogênicas, simbiontes e de vida-livre, grupos taxonômicos distintos (ver Tabela 1 e 2). O propósito do sistema ProBacter é capturar as informações importantes sob o ponto de vista biológico, fornecer uma descrição de todas as proteínas como categoria funcional, referências cruzadas e predição de domínios que foram resultantes da análise do genoma dessas bactérias, e também de análises realizadas por outros grupos de pesquisa e disponibilizadas publicamente (Figura 5).

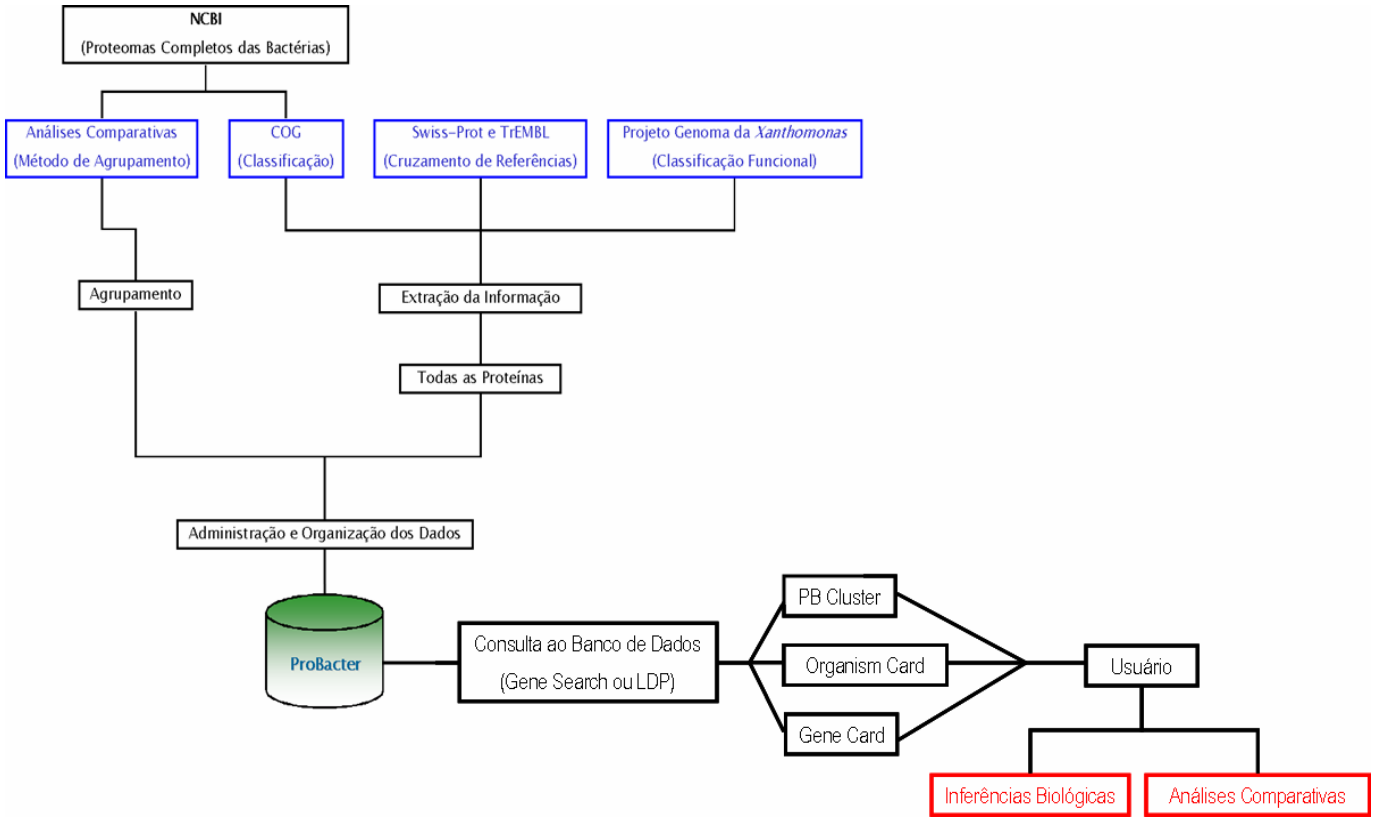

Figura 5. Representação do modo como as informação do modelo de dados ProBacter foram organizadas até chegarem ao usuário final.

4.2.1. Cruzamento de Referências

Como mencionado anteriormente, o banco de dados ProBacter tem por característica integrar informações provenientes de outras bases. Isto é necessário para que diferentes tipos de dados sejam "cruzados", aumentando a abrangência das referências de um dado gene. Optou-se por extrair informações dos seguintes bancos de dados: ***NCBI***, ***EMBL***, ***HAMAP***, ***InterPro***, ***PDB***, ***Pfam***, ***Swiss-Prot***, ***KEGG*** e ***PubMed***, e a partir de cada um desses uma tabela é criada contendo seus devidos atributos, a saber: números de acesso ao banco de dados selecionado e categorias funcionais (quando ocorrer). Estas informações provêm das bases de dados Swiss-Prot, TrEMBL, GO e COG, e foram agregadas ao ProBacter automaticamente. Dessa forma, foi possível gerar um segundo modelo parcial de dados (Figura 6).

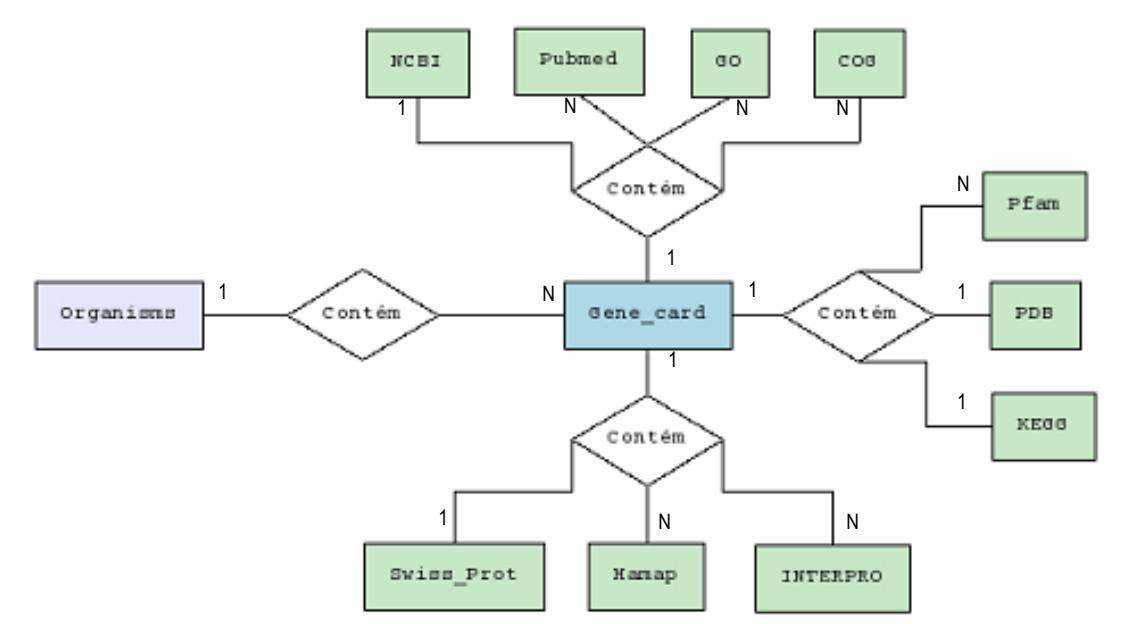

Figura 6. Diagrama ER para o segundo modelo parcial dos dados do ProBacter, incluindo os relacionamentos do cruzamento das referências (em verde).

4.2.2. Agrupamento dos Genes

O agrupamento dos genes foi feito usando a metodologia BBH (*Bidirectional Best Hit*) para organizar as proteínas dentro de grupos de genes. Este método é uma estratégia para detecção de grupos de genes conservados, que tem como base a seguinte definição: dado dois genes Xa e Xb de dois genomas Ga e Gb, Xa e Xb são chamados de melhor *hit* bi-direcional (BBH), se e somente se, houver similaridade reconhecida entre eles (OVERBEEK *et al.*, 1999); ou seja, se Xa possuir melhor *hit* com Xb e este possui melhor *hit* com Xa (Xa ↔ Xb). Entende-se aqui por similaridade reconhecida, o parâmetro "valor esperado" (*e-value*[9]) do programa de alinhamento de seqüências BLASTP (*protein – protein Basic Local Alignment Search Tool*) (ALTSCHUL *et al.*, 1990), onde esse valor deve ser menor que 1.0×10^{-5} e o alinhamento entre as seqüências possuir pelo menos 60% de cobertura. O agrupamento é construído por meio do alinhamento de seqüências de cada genoma contra todos os outros genomas através do programa BLASTP. Todos os *hits* obtidos são então armazenados em uma tabela do MySQL e, em seguida, é utilizado um algoritmo para a formação dos grupos de genes cuja ligações são definidas pelo BBH.

Como a estratégia acima havia sido implementada anteriormente para outros trabalhos desenvolvidos no Laboratório de Bioinformática do Laboratório Nacional de Computação Científica (ALMEIDA *et al*., 2004), esta implementação foi utilizada no sistema ProBacter. Optou-se por formalizar esta estratégia de formação de agrupamentos nesta dissertação para a discussão de sua complexidade e funcionamento, descrevendo-a no algoritmo a seguir (Algoritmo 1).

---

[9] O *e-value* é um parâmetro que descreve o número de *hits* que são "esperados" ao acaso quando se faz consulta em um banco de dados de tamanho particular.

A análise de complexidade deste algoritmo depende fortemente da escolha de como as estruturas de dados L, L_2 e D do algoritmo foram implementadas. Caso esta escolha seja ignorada temporariamente, este algoritmo possui complexidade O(n.m), onde "n" representa o número total de genes e "m" o número máximo de ligações que um gene pode fazer. Os dois primeiros laços deste algoritmo contêm os mesmos genes, ou seja, tudo que for inserido em L_2 será retirado de L antes de sair do laço interior, sendo esta a razão de sua complexidade. Sobre as escolhas das estruturas L, L_2 e D, as duas primeiras podem ser implementadas por meio de uma árvore binária *AVL*[10] (ou árvore balanceada pela altura, ou seja, uma árvore de busca binária autobalanceada), com deleção preguiçosa, já que serão feitos inúmeras inserções e deleções. Para a estrutura D, que possui relativamente poucos elementos, um exemplo de implementação poderia ser uma tabela "*hash*[11]". Se o tamanho e a função da tabela forem escolhidos de forma adequada para evitar colisões, obtêm-se inserções e deleções de ordem O(1). Portanto, se estas escolhas forem feitas, a complexidade final do algoritmo seria de O(n.m.*log*(n)), onde não foi levado em consideração a complexidade da busca dos *hits* para cada gene no MySQL. Entretanto, isso não quer dizer que a complexidade do algoritmo independe da busca pelo SQL.

Os resultados dos grupos formados provenientes da aplicação do algoritmo de agrupamento foram armazenados em duas tabelas: ***Clusters_Elements,*** que contém as informações de um elementos do grupo, cujos atributos são: nome do organismo, valor esperado (*e-value*), produto gênico e identificação do grupo; e ***Clusters_Links***, que contém as informações presentes em todas as ligações de Xa com Xb e Xb com Xa, o produto gênico e o valor esperado (*e-value*). Com isso, foi possível gerar um novo diagrama ER (Entidade-Relacionamento) parcial do ProBacter (Figura 7).

---

10 O nome AVL vem de seus criadores Adelson Velsky e Landis.

11 A tabela "*hash"* também é conhecida por tabelas de espalhamento. Seu objetivo é, a partir de uma chave simples, fazer uma busca rápida e obter o valor desejado.

Os agrupamentos construídos foram utilizados também para a identificação de parálogos dentro de um genoma, ou seja, se dois genes do mesmo organismo estiverem presentes em um mesmo agrupamento, estes genes têm o potencial de serem parálogos. Desta forma, para facilitar a identificação dos possíveis parálogos foi feita após a formação dos grupos, uma busca por genes que se repetem múltiplas vezes dentro do mesmo genoma e esta informação foi armazenada no banco de dados.

Algoritmo 1: Agrupamento de Proteínas.

```
L:= lista completa de proteínas de todos os genomas;
i:= 0;
Enquanto (tamanho(L) > 0)
{
        x:= primeiro elemento da lista L;
        L2:= lista vazia;
        D := lista vazia;
        d := falso;
        colocar x em L2;
        Enquanto (tamanho(L2) > 0)
        {
                y := primeiro elemento da lista L2;
                remover y de L e L2;
                colocar y em D;
                Para cada h que é hit de y e tem similaridade reconhecida
                {
                        Se y é hit de h e tem similaridade é reconhecida
                        {
                                guardar ligação h ↔ y com identificador de grupo = i;
                                coloque h em L2 se não estiver em D;
                                d ← verdadeiro;
                        }
                }
        }
        Se d é verdadeiro incremente i;
}
```

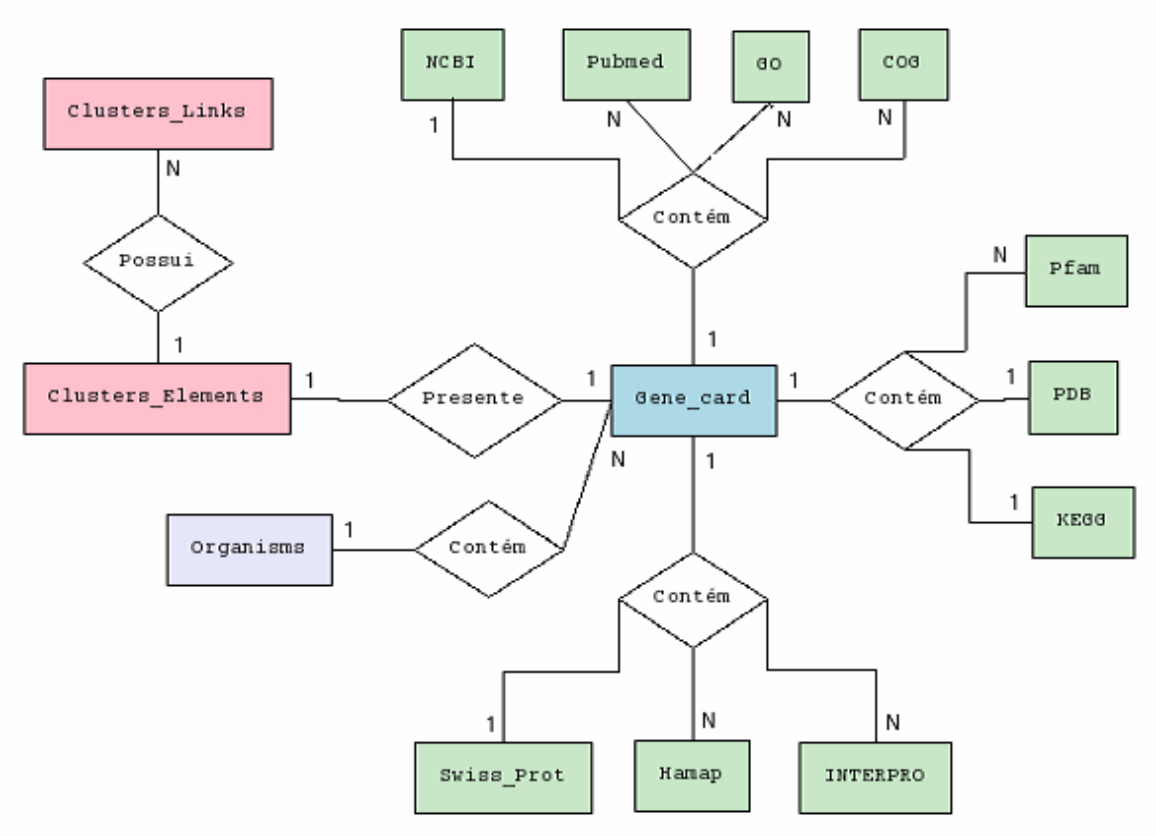

Figura 7. Diagrama ER para o terceiro modelo conceitual parcial do ProBacter, incluindo as relações dos agrupamentos formados (em rosa).

4.2.3. Categorização Funcional dos Genes

Uma categoria (ou classe) funcional se refere ao processo biológico, função molecular ou componente celular que o produto de um dado gene pode desempenhar. Para este banco de dados foi utilizado o sistema de categorias funcionais definido para o projeto genoma da *Xanthomonas axonopodis* pv. c*itri e Xanhomonas campestris pv.campestris* (DA SILVA *et al.*, 2002) (Tabela 4).

A propagação da categoria funcional para cada produto dos genes do ProBacter foi feita automaticamente, onde as categorias foram transferidas para os membros dos grupos de proteínas que continham pelo menos uma proteína dos genomas das *Xanthomonas* acima mencionadas. As categorias do projeto *Xanthomonas* foram utilizadas para padronizar as categorias dos produtos dos genes do ProBacter devido a;

(*i*) sua abrangência na classificação, (*ii*) ter sido manualmente curada, (*iii*) por ser uma bactéria associada à planta, com um genoma de aproximadamente cinco mil genes, onde existe a categorização das proteínas associadas à patogenicidade, virulência e adaptação ao hospedeiro.

Vale ressaltar que existem grupos que não contém proteínas para as *Xanthomonas* citadas, e assim não recebendo uma categorização funcional. Para solucionar este problema deverão ser propagadas categorias funcionais de genomas de bactérias associadas às plantas que foram anotadas e manualmente curadas.

Tabela 5. Tabela com as instâncias e as subdivisões correspondentes das classes.

Descrição das Categorias Funcionais
I. Intermediary metabolism
A. Degradation
1. Degradation of polysaccharides and oligosaccharides
2. Degradation of small molecules
3. Degradation of lipids
B. Central intermediary metabolism
1. Amino sugars
2. Entner-Douderoff
3. Gluconeogenesis
4. Glyoxylate bypass
5. Miscellaneous glucose metabolism
6. Non-oxidative branch, pentose pathway
7. Nucleotide hydrolysis
8. Nucleotide interconversions
9. Phosphorus compounds
10. Pool, multipurpose conversions
11. Sugar-nucleotide biosynthesis, conversions
12. Sulfur metabolism
C. Energy metabolism, carbon
1. Aerobic respiration
2. Anaerobic respiration and fermentation
3. Electron transport
4. Glycolysis
5. Oxidative branch, pentose pathway
6. Pyruvate dehydrogenase
7. TCA cycle
8. ATP-proton motive force interconversion
D. Regulatory functions
1. Two component systems
2. Activators-Repressors
3. Kinases-Phosphatases
4. Sigma factors and other regulatory components

Continua.

5. Not used
II. Biosynthesis of small molecules
A. Amino acids biosynthesis
1. Glutamate family|nitrogen assimilation
2. Aspartate family, pyruvate family
3. Glycine-serine family|sulfur metabolism
4. Aromatic amino acid family
5. Histidine
B. Nucleotides biosynthesis
1. Purine ribonucleotides
2. Pyrimidine ribonucleotides
3. 2'-Deoxyribonucleotides
4. Salvage of nucleosides and nucleotides
C. Sugars and sugar nucleotides biosynthesis
D. Cofactors, prosthetic groups, carriers biosynthesis
1. Biotin
2. Folic acid
3. Lipoate
4. Molybdopterin
5. Pantothenate
6. Pyridoxine
7. Pyridine nucleotides
8. Thiamin
9. Riboflavin
10. Thioredoxin, glutaredoxin, glutathione
11. Menaquinone, ubiquinone
12. Heme, porphyrin
13. Biotin carboxyl carrier protein (BCCP)
14. Cobalamin
15. Enterochelin
16. Biopterin
17. Others
E. Fatty acid and phosphatidic acid biosynthesis
F. Polyamines biosynthesis
III. Macromolecule metabolism
A. DNA metabolism
1. Replication
2. Structural DNA binding proteins
3. Recombination
4. Repair
5. Restriction, modification
B. RNA metabolism
1. Ribosomal and stable RNAs
2. Ribosomal proteins
3. Ribosomes - maturation and modification
4. Aminoacyl tRNA synthetases, tRNA modification
5. RNA synthesis, modification, DNA transcription
6. RNA degradation
C. Protein metabolism
1. Translation and modification
2. Chaperones
3. Protein degradation
D. Other macromolecules metabolism
1. Polysaccharides
2. Phospholipids

Continua.

3. Lipoprotein
IV. Cell structure
A. Membrane components
1. Inner membrane
2. Outer membrane constituents
B. Murein sacculus, peptidoglycan
C. Surface polysaccharides, lipopolysaccharides, and antigens
D. Surface structures
V. Cellular processes
A. Transport
1. Amino acids, amines
2. Anions
3. Carbohydrates, organic acids, alcohols
4. Cations
5. Nucleosides, purines, pyrimidines
6. Protein, peptide secretion
7. Other
B. Cell division
C. Chemotaxis and mobility
D. Osmotic adaptation
E. Cell killing
VI. Mobile genetic elements
A. Phage-related functions and prophages
B. Plasmid-related functions
C. Transposon- and intron-related functions
VII. Pathogenicity, virulence, and adaptation
A. Avirulence
B. Hypersensitive response and pathogenicity
C. Toxin production and detoxification
D. Host cell wall degradation
E. Exopolysaccharides
F. Surface proteins
G. Adaptation, atypical conditions
H. Other
VIII. Hypothetical
A. Conserved hypothetical proteins
B. Hypothetical proteins (includes no hits or only low score hits)
C. Xanthomonas conserved hypothetical
IX. ORFs with undefined category

Após terem sido transferidas, as classes foram armazenadas em uma entidade denominada ***Category***, contendo os seguintes atributos: nome da categoria funcional, identificador dos grupos de proteínas e identificador do gene. Assim, esta última entidade torna completo o modelo de dados do ProBacter (Figura 8).

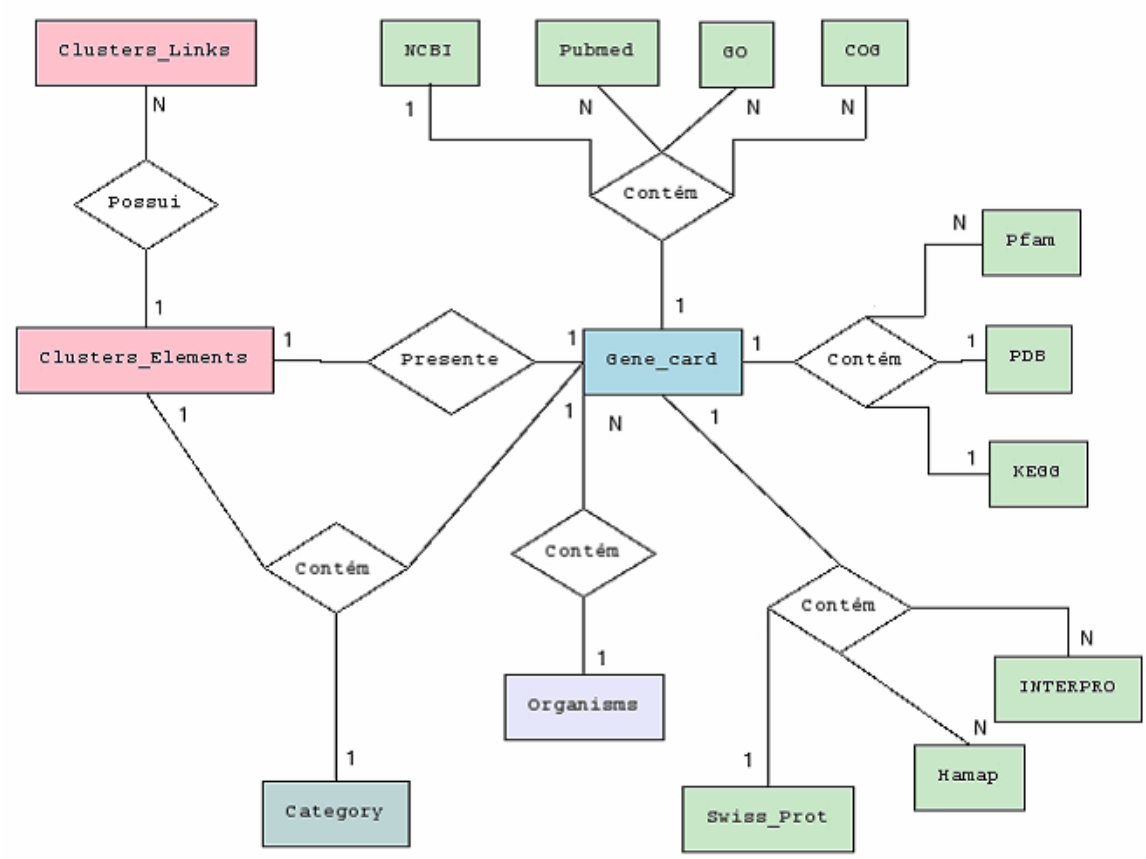

Figura 8. Diagrama ER para o modelo conceitual final do ProBacter, incluindo os relacionamentos da categoria funcional (em verde escuro).

4.2.4. Módulos de Visualização *Web*

Nas seções anteriores discutiu-se a construção do sistema ProBacter. No entanto, para que as informações do banco de dados possam ser visualizadas e utilizadas, foi preciso desenvolver interfaces amigáveis *Web*. As informações foram organizadas em módulos de visualização distintos, tomando como base às tabelas do banco de dados (Figura 9). Assim, os seguintes tipos de visualização dos dados foram criados: *Organism Card*, *Gene Card* e *PB Cluster*. Vale ressaltar que a maioria dos campos desses módulos possui informações adicionais, possibilitando que o usuário obtenha dados adicionais de diferentes bases de dados. Os módulos de visualização citados serão discutidos a seguir.

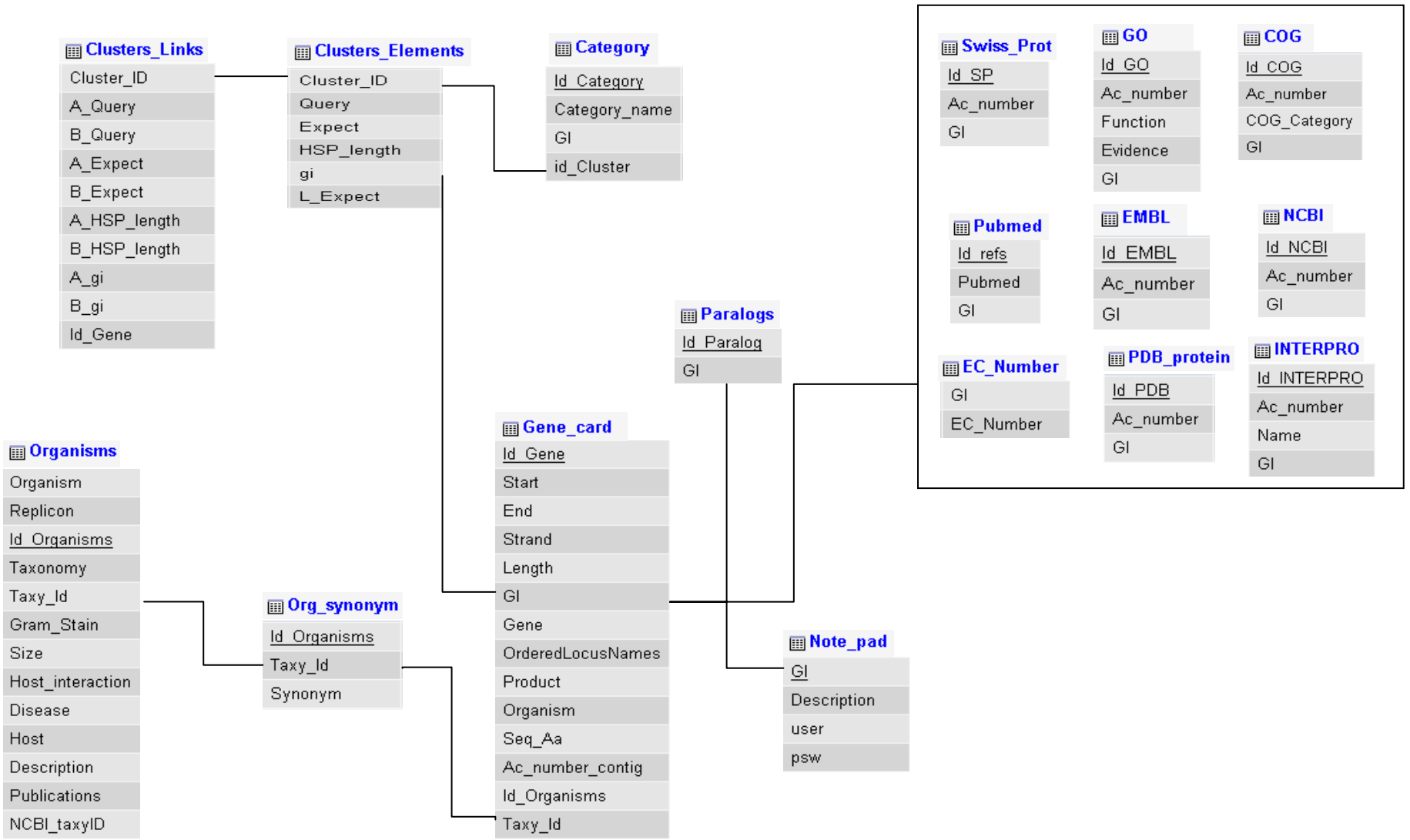

Figura 9. Descrição completa do banco de dados ProBacter, demonstrando que todas as tabelas (com seus respectivos campos) estão interligadas.

4.2.4.1. *Organism Card*

O programa de visualização *Organism Card* (Figura 10) contém informações gerais sobre o genoma dos organismos, tais como: taxonomia, tipo de coloração Gram (maneira pela qual a forma da bactéria pode ser observada, dividindo-as em dois grupos: Gram-positivas e Gram-negativas, aproximadamente iguais em número e importância), tamanho do genoma, tipo de interação com o hospedeiro, tipo do hospedeiro, doença associada (se for o caso) e publicação relevante. Estes dados foram extraídos dos campos da tabela ***Organisms*** descritas na Figura 9.

Organism Card

Organism Name: ***Xanthomonas axonopodis pv. citri str 306***			
Gram Stain -	Host Interaction Pathogenic	Host Plant	Disease Cintrus canker
Size 5.27	NCBI Taxy ID 92829	Taxonomy Gammaproteobacteria	
Replicon	Consist one circular chromosome and two plasmids (Xanthomonas citri plasmid pXAC33 and Xanthomonas citri plasmid pXAC64).		
Description	This strain contains genes that are similar to genes from another citrus pathogen, Xylella fastidiosa. It has 2 different type II secretion system for the export of extracellular enzymes that degrade the plant cell wall including cellulases. The genome encodes a single type III secretion system (Hrp) that is important for pathogenicity as well as two type IV secretion system, one on the chromosome and one found on one of the two plasmids this organism contains, pXAC64.		
Publication	Nature. 2002 May 23;417(6887):459-63.		

Figura 10. Módulo de visualização *Organism Card* contendo as características gerais de um determinado organismo. Neste caso, foi usado como exemplo a *Xanthomonas axonopodis* pv. *citri* str. 306.

4.2.4.2. *Gene Card*

As informações de cada gene podem ser visualizadas no *Gene Card* que inclui: (*i*) categoria funcional que foi transferida do Projeto Genoma do gênero *Xanthomonas* e que neste sistema recebeu o nome de *PB Functional Category*; (*ii*) cruzamento das referências e predição de domínios que foram nomeados seguindo a nomenclatura dos bancos de dados a que pertencem; (*iii*) mapa indicando a posição do gene (destacado por chaves vermelhas) dentro do genoma e sua vizinhança. A coloração dos genes corresponde a categorias funcionais do COG (Figura 11). Neste mapa o usuário pode percorrer o genoma, bastando somente selecionar o gene da figura. Neste caso as seguintes informações adicionais sobre o gene selecionado ficam disponíveis ao usuário; (*i*) seqüência de aminoácidos da proteína, onde o usuário pode compará-la com outras proteínas presentes no banco através do programa BLAST; e (*ii*) publicações correspondentes obtidas a partir da página do Entrez[12] (PubMed).

[12] Disponível em www.ncbi.nlm.nih.gov/**entrez**/

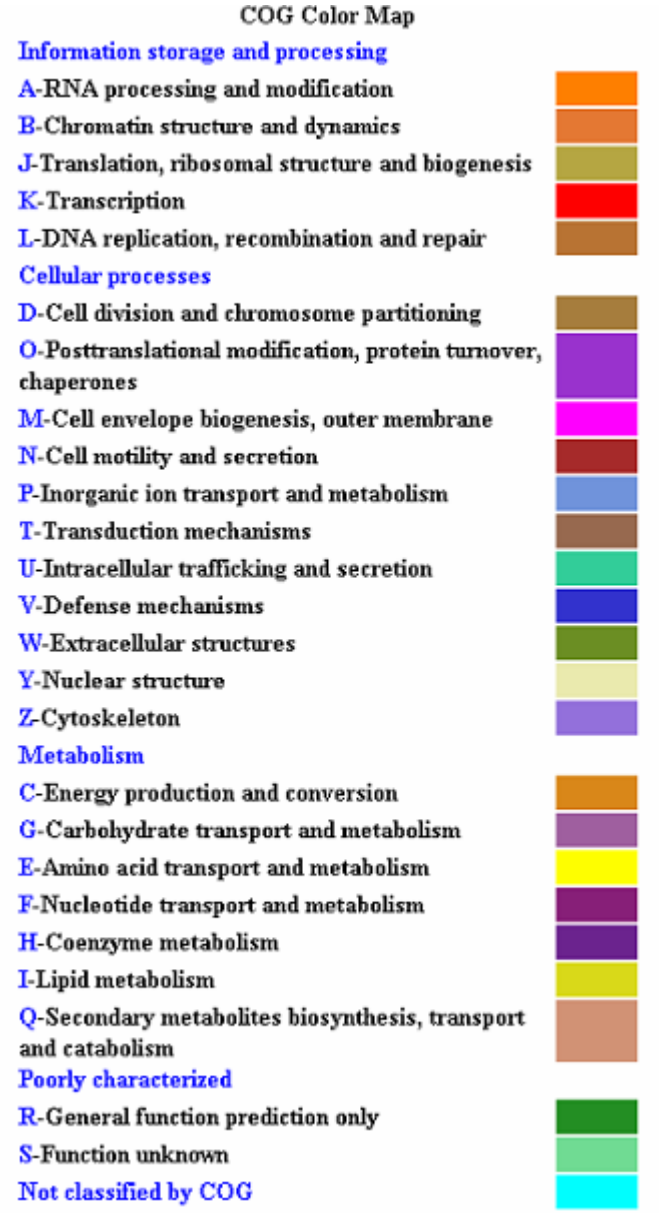

Figura 11. Esta figura esta disponível no sistema ProBacter e informa quais as categorias presentes no banco de dados COG, bem como as cores que as correspondem. Estas cores, como já foram mencionadas, são utilizadas nas figuras que indicam a posição dos genes.

4.2.4.3. *PB Cluster*

O módulo de visualização *PB Cluster* (Figura 12), contém informações sobre as proteínas que foram agrupadas pelo método BBH, já descrito. Neste módulo estão disponíveis também, o organismo ao qual o gene pertence, o produto codificado pelo gene, a classificação de acordo com o COG, e a ligação com menor valor esperado (*e-value*) ligando o gene ao grupo. Além de mostrar as proteínas que participam dos agrupamentos e o número total que o compõe, foram agregadas outras que permitem explorar as interações entre estes genes, tais como: ferramentas de visualização de alinhamento múltiplo de seqüências e um modo de re-agrupamento dos grupos formados. Estas ferramentas serão discutidas mais detalhadamente a seguir.

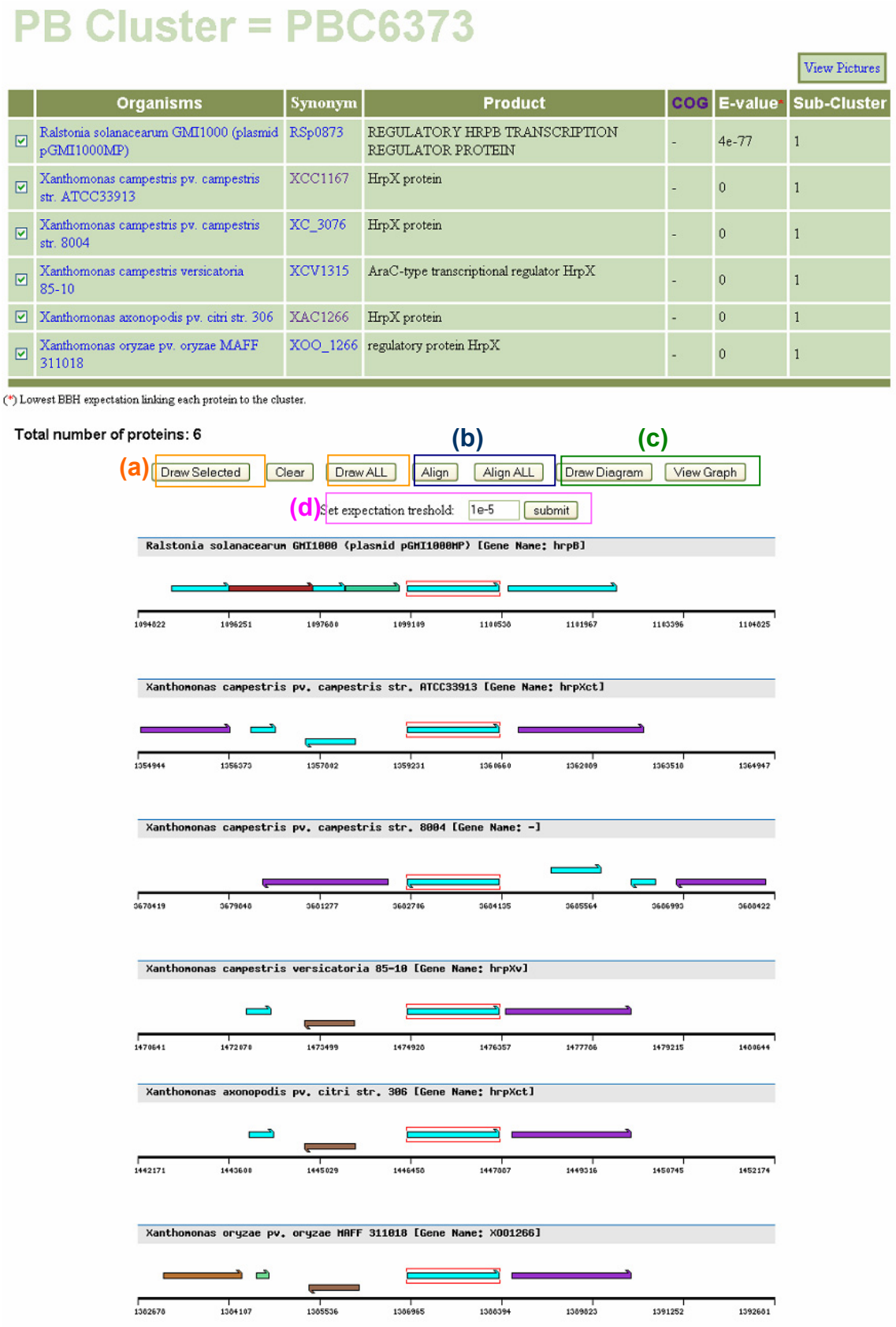

Figura 12. Módulo de visualização da ferramenta *PBCluster*. a) Os retângulos em laranja indicam a localização dos genes nos grupos em um mapa. Caso opte por visualizar todos os genes apresentados basta selecionar *Draw ALL*, ou se preferir por selecionar apenas alguns dos

genes é necessário indicar nas caixas citadas na frente do organismo e em seguida selecionar *Draw Selected*; (b) o retângulo em azul indica que a ferramenta *Align* pode ser selecionada, escolhendo alguns genes ou alinhando todos os que foram apresentados; (c) o retângulo em verde mostra as duas ferramentas usadas para visualizar a conformação estrutural de um cluster; e (d) o retângulo em rosa indica a ferramenta de re-agrupamento dos genes, onde o usuário pode indicar o valor desejado, o *e-value* 1e-5 é a opção padrão do programa. A figura que indicada em (a) é uma modificação da utilizada pelo programa SABIA (ALMEIDA *et al.*, 2004) para atender aos propósitos do sistema ProBacter.

4.2.4.3.1. *Draw Selected* e *Draw ALL*

As ferramentas *Draw Selected* e *Draw ALL* foram utilizadas para demonstrar os mapas de localização dos genes presentes no agrupamento dentro de seus respectivos genomas (Figura 12a).

4.2.4.3.2. *Align* e *Align ALL*

Foram integradas ao sistema ProBacter os programas de alinhamento múltiplo de seqüências de aminoácidos ClustalW versão 1.81 (THOMPSON *et al.*, 1994) e T-Coffee versão 4.45 (NOTREDAME *et al.*, 2000), sendo estas opções disponíveis nas ferramentas em questão (Figura 12b). Os parâmetros padrões de cada programa mencionado podem ser alterados quando necessário (Figura 13). É possível também visualizar a árvore filogenética e o alinhamento gerado mostrando as bases conservadas, a qualidade e consenso, por meio do programa *Jalview*[13] versão 2.1.1 (CLAMP *et al.*, 2004).

[13] Disponível em www.**jalview**.org/

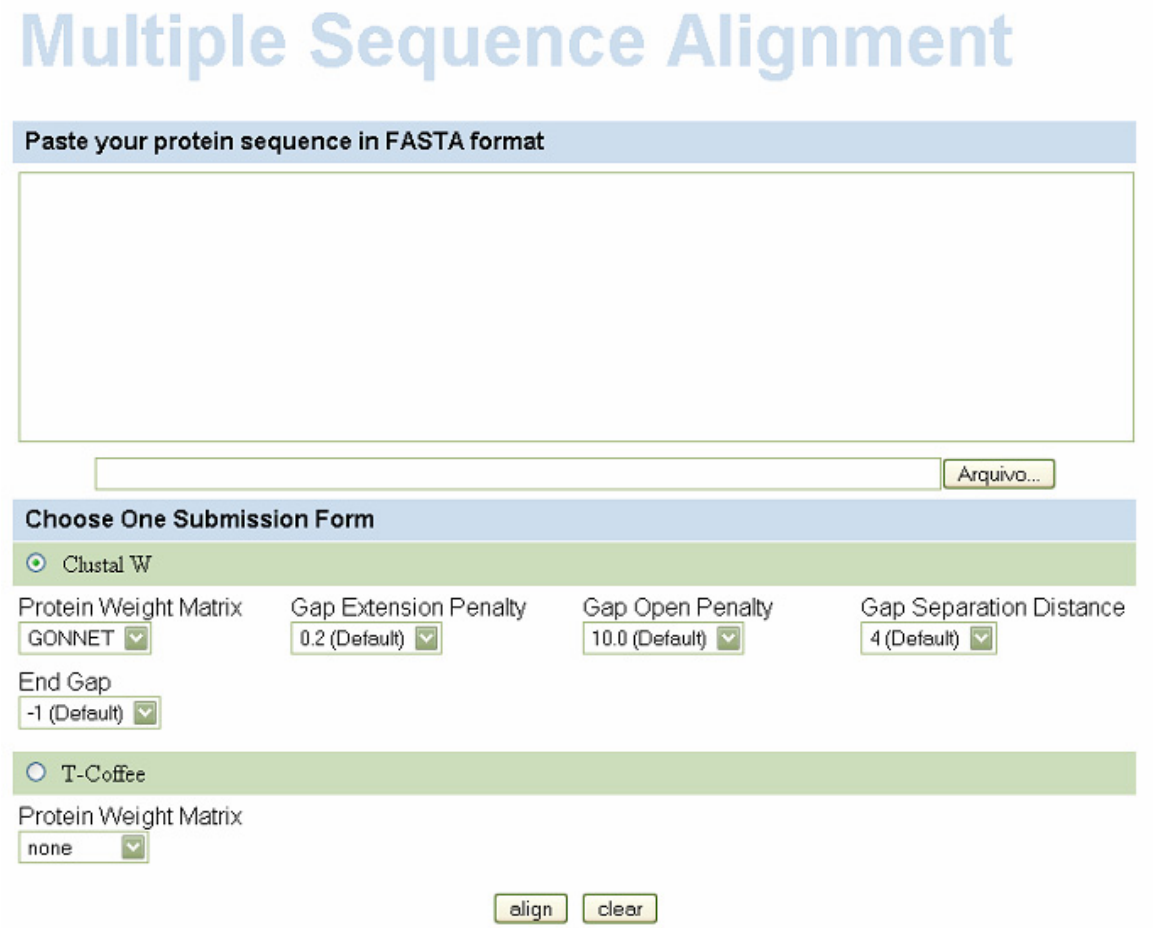

Figura 13. Módulo de visualização das ferramentas integradas ClustalW (opção padrão) e T-Coffee. Estas ferramentas não somente estão atreladas ao módulo *PBCluster*, possibilitando alinhar uma ou mais seqüências externas com as que podem ser selecionadas no referido módulo, como também permite que o usuário alinhe a seqüência que desejar, bastando somente selecionar a opção *Multiple Alignment*.

4.2.4.3.3. *Draw Diagram* e *View Graph*

As ferramentas *Draw Diagram* e *View Graph* são módulos gráficos de visualização que apresentam a composição estrutural formada pelos agrupamentos das proteínas. A opção *Draw Diagram* (Figura 12c) mostra, em formato de diagrama, a estrutura do agrupamento gerado pelo programa *Graphviz*[14] (*Graph Visualization Software*). Já a ferramenta *View Graph* (Figura 12c) permite a visualização da disposição das proteínas no agrupamento de uma forma interativa, utilizando o

[14] O programa *Graphviz* está disponível em www.graphviz.org/

programa *LinkBrownser*[15] versão 1.20. Todos os genes, se selecionados, direcionam para o módulo de visualização *Gene Card* que contém informações particulares a respeito dos mesmos.

4.2.5.2.4. Re-Agrupamento

A ferramenta de re-agrupamento (Figura 12d) foi desenvolvida com o intuito de possibilitar que os agrupamentos já formados pudessem ser re-agrupados. Este módulo tem por objetivo aumentar o ponto de corte sobre o valor esperado (*e-value*) das ligações que cada proteína faz ao grupo. Desta forma, genes com ligações cujo valor esperado é superior ao ponto de corte estipulado serão temporariamente removidos do grupo. Assim, um novo agrupamento é criado, onde as ligações entre proteínas presentes possuem valores iguais ou inferiores àquele que foi estipulado. Além de descartar proteínas que estão fora do escopo, novos subgrupos poderão ser gerados. Estes novos subgrupos recebem uma numeração que é indicada na última coluna da tabela de resultados.

4.3. Métodos de Consulta e Anotação Manual

Para que os dados contidos no banco de dados fossem acessados e visualizados foram implementadas consultas de fácil entendimento. Também, para que o usuário pudesse acrescentar novas informações ao banco uma ferramenta de anotação manual foi criada. Essas ferramentas serão detalhadas nas seções que se seguem e podem ser acessadas em ***http://www.probacter.lncc.br*** selecionando a opção *Database Search*.

[15] O programa *LinkBrownser* está disponível em http://www.touchgraph.com/

4.3.1 Gene Search

A ferramenta de busca *Gene Search* foi desenvolvida para permitir consultas ao banco de dados por meio de palavras-chave, inserindo-as em caixas pré-determinadas conforme indicado na Figura 14. Estas caixas pré-determinadas (quando selecionadas) são convertidas em uma consulta SQL (*Structured Query Language*) que é em seguida interpretada pelo MySQL, retornando o resultado da busca. Os resultados da consulta são mostrados em formato tabular ordenados pelo nome do organismo, com cada entrada representada pelo nome do gene, sinônimo, produto do gene, organismo, número identificador do agrupamento e tamanho do grupo (estes dois últimos somente são vistos quando a informação do agrupamento for consultada). Estes resultados, quando selecionados, direcionam para os módulos de visualização já descritos.

Figura 14. Módulo de visualização da ferramenta *Gene Search*. Neste método de consulta o usuário pode estipular quantos resultados deseja visualizar, indicando o valor na opção *Show results*.

4.3.2. LDP - Linguagem Declarativa do ProBacter

Na seção anterior foi descrito um tipo de busca para obtenção e visualização dos dados armazenados no sistema ProBacter. Contudo, nem sempre é possível responder a todas as perguntas desejadas somente através de buscas pré-formatadas como é feito no *Gene Search*. Buscas mais elaboradas a base de dados necessitam de relações mais complexas entre as tabelas, e seria necessária a implementação destas consultas usando diretamente a linguagem SQL (*Structured Query Language*). No entanto, não se pode esperar que todos os usuários do sistema estejam familiarizados com este tipo de linguagem de busca, como também com as tabelas existentes no banco de dados. Além disso, não é seguro disponibilizar acesso à linguagem SQL na rede já que pode causar danos ao banco de dados se usado de forma maliciosa, este é um caso comum de falha de segurança chamado de *SQL injection*. Assim, foi desenvolvida uma nova linguagem de busca denominada LDP (Linguagem Declarativa do ProBacter), apelidada de *Probactish* (Figura 15). Esta linguagem tem o propósito de estabelecer uma "ponte" entre os usuários que acessam o banco de dados e a SQL.

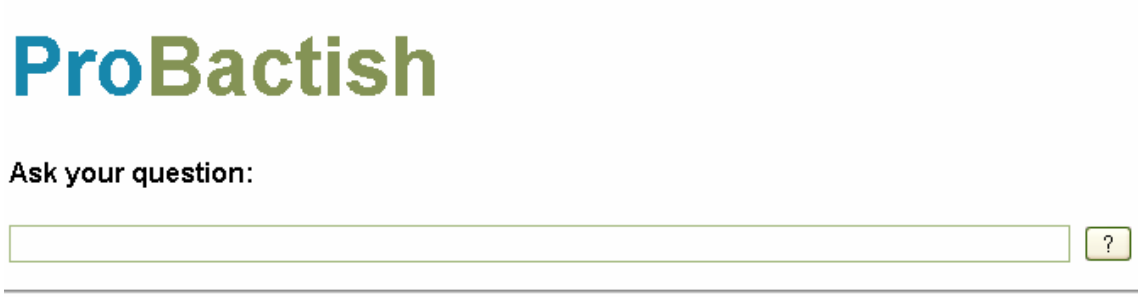

Figura 15. Módulo de visualização da Linguagem Declarativa do ProBacter (LDP) ou *ProBactish*. Os resultados provenientes deste módulo de consulta são fornecidos em formato tabular e também possuem ligações com os outros módulos de visualização já descritos.

A LDP é uma linguagem de consulta construída sobre a SQL para facilitar o acesso dos usuários às informações armazenadas no banco de dados. A própria linguagem SQL foi originalmente concebida para facilitar a consulta. Através dela, o usuário concentra-se somente na busca de interesse e não nos detalhes de como ela está sendo executada. Este tipo de linguagem de programação de alto-nível[16] é algumas vezes denomina de busca declarativa. Dentro desse paradigma, a LDP foi construída em um mais alto-nível, de maneira que a arquitetura do banco de dados ficasse totalmente escondida. Além disso, a linguagem foi definida a partir de palavras comumente utilizadas pela comunidade científica, tornando-a ainda mais acessível.

A implementação da LDP consistiu em definir uma gramática livre de contexto (Tabela 5) e, em seguida, desenvolver um sistema tradutor que recebe as sentenças em LDP e posteriormente as traduz para SQL. O procedimento de definir uma gramática e utiliza-la para criar um tradutor (ou compilador) é uma prática considerada comum no campo da Ciência da Computação para o desenvolvimento de novas linguagens de programação (HOPCROFT, ULLMAN, 1979). Linguagens de buscas delineadas especificamente para determinado tipo de aplicação não são temas inovadores em bioinformática e já foram utilizados em outros trabalhos (TATA *et al*., 2006; PATEL, HUDDLER, HAMMEL, 2005).

[16] Linguagem de programação de alto nível é uma nomenclatura usada para linguagens com um nível de abstração elevado, longe do código de máquina e mais próximo à linguagem humana.

Tabela 6. Definição da gramática para a Linguagem Declarativa do ProBacter.

query	→	QUERYTYPE agents action
agents	→	AGENT | agents , AGENT
action	→	VERB relations | action , BOOLOP VERB relations
relations	→	relation | relations BOOLOP relation
relation	→	relation_agent BIOP information
Relation_agent	→	AGENT | REL_AGENT
information	→	NUMBER | STRING
QUERYTYPE	→	*Which* | *how many*
AGENT	→	*gene(s)* | *protein(s)* | *organism(s)* | *product(s)* | *synonym(s)* | *gene length* | *gene start* | *gene end* | *cluster(s)* | *cluster description* | *cluster size* | *interpro* | *pubmed* | *pfam* | *replicon(s)* | *taxonomy(s)* | *gram stain* | *host interaction* | *organism size* | *disease(s)* | *host(s)* | *organism description*| *publication reference(s)* | *cateory(ies)*
VERB	→	*Are associated with*
BOOLOP	→	*and* | *or*
REL_AGENT	→	*gi* | *cog* | *taxy_id*
BIOP	→	*like* | *not like*| *=* | *!=* | *>* | *<* | *<=* | *>=*
NUMBER	→	*Um número qualquer.*
STRING	→	*Uma seqüência de letras entre aspas.*

Observando a definição da gramática dada acima, a linguagem é constituída de dois tipos de buscas (identificado na gramática acima por QUERYTYPE): a primeira irá listar os resultados e é iniciada pela palavra "*which"*, a segunda informa quantidades, esta é iniciada com a palavra "*how many"*. Denomina-se verbo a palavra que descreve o

tipo de ação que se está interessado na busca. Atualmente, foi somente implementado um tipo de verbo (identificado por VERB), e este é indicado pelas palavras "*are associated with"*. Este verbo está relacionado com a consulta no banco de dados. A LDP foi construída de forma que outros verbos poderão ser posteriormente inseridos tornando-a extensível, podendo, desta forma, ir além de buscas a base de dados. Os agentes (AGENT) são palavras que descrevem campos das tabelas contidas na base de dados, os quais o usuário possa ter interesse de visualizar; já os agentes de relação (REL_AGENT) são campos que contém identificadores geralmente usados como chaves de referência. Dessa forma, permite que o usuário tenha acesso a todas as tabelas presentes na base de dados.

Para implementar o tradutor foram usadas ferramentas clássicas na construção de compiladores como, o *LEX*[17] (*Lexical Analyser Generator*), um programa que extrai padrões de textos através de expressões regulares definidas pelo usuário, e o *YACC*[17] (*Yet Another Compiler-Compiler*), um programa que gera um *parser* a partir da definição de uma gramática livre de contexto. Abaixo seguem alguns exemplos da utilização da LDP com suas respectivas traduções em SQL. Estes exemplos foram usados para gerar alguns dos resultados obtidos e que foram discutidos na seção seguinte.

Exemplo 1. *Quais organismos estão associados à doença podridão negra?*

LDP: **which organisms are associated with disease like "black rot"**

SQL: SELECT DISTINCT Org_synonym.Synonym FROM Gene_card, Org_synonym, Organisms WHERE Org_synonym.Taxy_Id = Gene_card.Taxy_Id AND Organisms.Taxy_Id = Org_synonym.Taxy_Id AND (Organisms.Disease like "%black rot%")

[17] Estes programas estão disponíveis em http://dinosaur.compilertools.net/.

Exemplo 2. *Quantos agrupamentos possuem 10 ou mais proteínas?*

LDP: **how many cluster are associated with cluster size >= 10**

SQL: SELECT count(DISTINCT Clusters_Elements.Cluster_ID) FROM Gene_card, Clusters_Elements, Clusters_Info WHERE Gene_card.GI = Clusters_Elements.GI AND Clusters_Elements.Cluster_ID = Clusters_Info.Cluster_ID AND (Clusters_Info.size >= 10)

Exemplo 3. *Quantos agrupamentos são da categoria funcional "intermediary metabolism" em Ralstonia solanacearum?*

LDP: **how many cluster are associated with organism like "ralstonia solanacearum" and category like "intermediary metabolism"**

SQL: SELECT count(DISTINCT Clusters_Elements.Cluster_ID) FROM Gene_card, Org_synonym, Organisms, Clusters_Elements, Category WHERE Org_synonym.Taxy_Id = Gene_card.Taxy_Id AND Organisms.Taxy_Id = Org_synonym.Taxy_Id AND Gene_card.GI = Clusters_Elements.GI AND Clusters_Elements.Cluster_ID = Category.id_Cluster AND (Org_synonym.Synonym like "%ralstonia solanacearum%" and Category.Category_name like "%intermediary metabolism%")

Para permitir acesso externo a esta linguagem foi criada uma interface *Web* que permite desenvolver perguntas na linguagem LDP (Figura 15), e os resultados obtidos são dispostos em formato tabular, possuindo ligações para os diferentes módulos de visualização.

4.3.3. Anotação Manual

A anotação manual pode ser feita no sistema ProBacter através da inserção de informações em um campo, o ***NotePad***, permite que o usuário insira no banco de dados informações de anotação para uma determinada proteína e que estas fiquem disponíveis para o acesso público. Entretanto, para que sua anotação seja feita é necessário o

cadastramento deste usuário ou grupo de pesquisa, assim ele receberá um nome de usuário e senha para acessar a ferramenta *NotePad* e inserir seus dados (Figura 16).

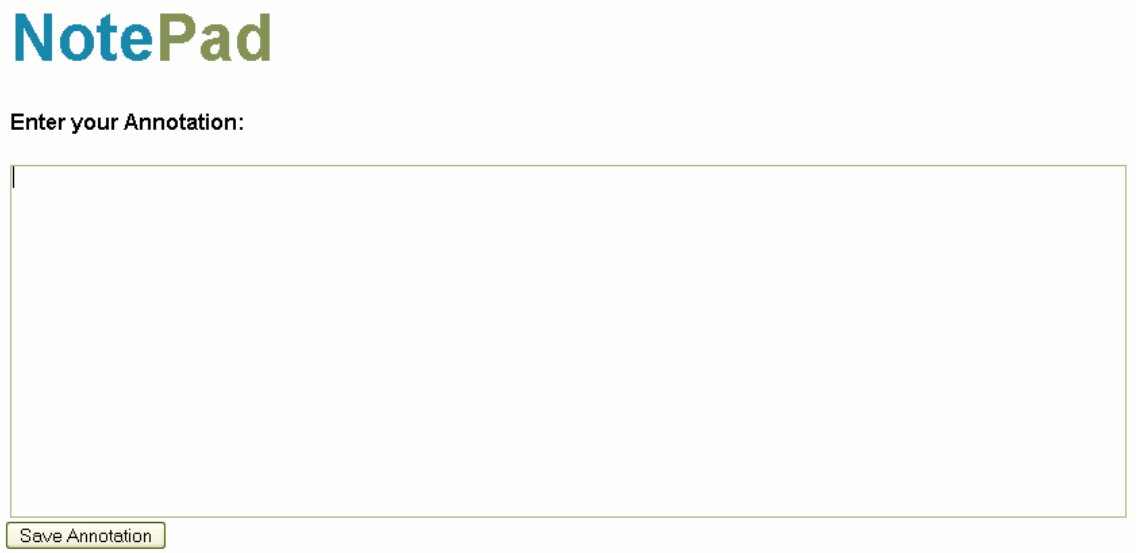

Figura 16. Visualização da ferramenta *NotePad*, onde é permitido que o usuário insira seus próprios dados de anotação manual.

5. RESULTADOS E DISCUSSÃO

No capítulo anterior foi apresentado como o sistema ProBacter foi construído. Nesta seção serão apresentados os módulos de consulta, as aplicações desse sistema e os diferentes métodos de análise, discutindo os resultados obtidos e fornecendo uma visão comparativa com outros trabalhos. Esse banco de dados é uma re-edição do banco de dados PABdb usado para capturar informações de genomas procariotos, tendo como principais entidades o organismo e a família do gene. O sistema PABdb foi desenvolvido para auxiliar nas análises comparativas de oitos genomas de bactérias que vivem em associação com plantas e os resultados desse trabalho foram publicados no ano de 2002 por Van Sluys e colaboradores.

5.1. Análise do Sistema ProBacter

Hoje, o sistema ProBacter contém informações sobre o genoma de 31 organismos, compreendendo um total de 153.712 entradas que representam os genes e as proteínas descritas para cada um desses genomas. A metodologia de agrupamento gerou um total de 13.506 grupos que abrangem 126.577 (82,3%) de todas as proteínas do banco (Figura 17).

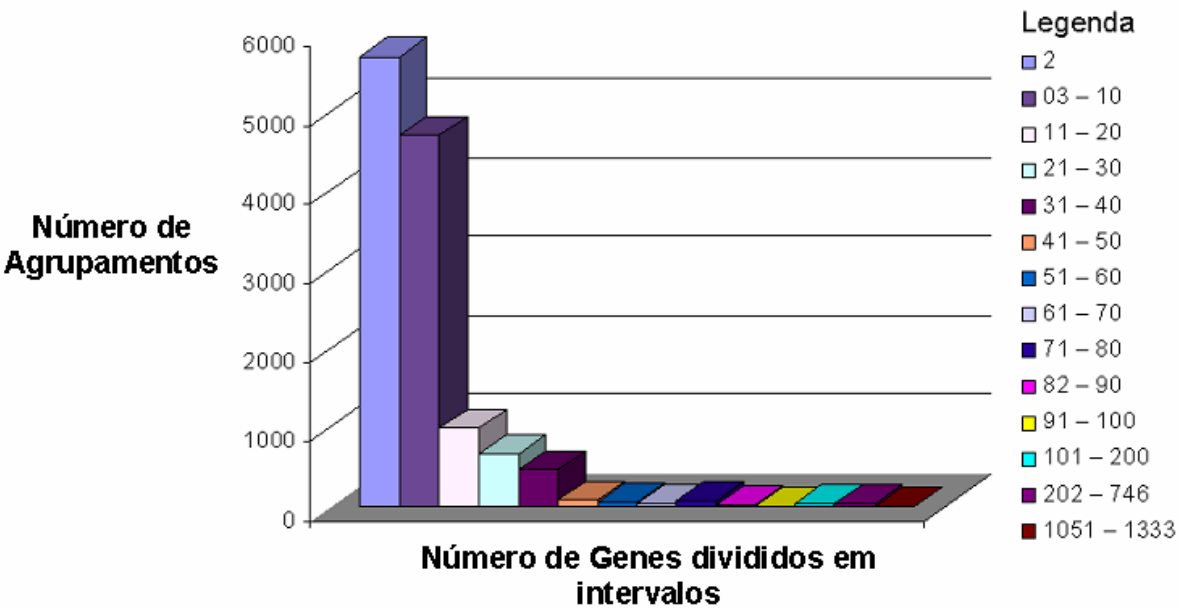

Figura 17. Gráfico representa o número de agrupamentos pelo número de genes que estão divididos em intervalos de acordo com a legenda.

Uma grande parcela dos agrupamentos formados (10.926) possuem até 10 proteínas, sendo que na maioria dos casos os agrupamentos incluíram proteínas do mesmo grupo taxonômico. Uma observação interessante é que um único agrupamento (PBC6539) apresenta somente proteínas dos gêneros *Xylella* e *Xanthomonas*, esse agrupamento é composto por 12 proteínas que foram anotados como regulador de virulência, estando presente em duas cópias em cada um dos genomas das bactérias.

Um total de 2.581 proteínas estão presentes em agrupamentos com mais de 10 proteínas, essas proteínas puderam ser dividas nas seguintes categorias funcionais: 375 agrupamentos apresentam proteínas categorizadas como I. Metabolismo intermediário; 284 agrupamentos, II. Biossíntese de Pequenas Moléculas; 365 agrupamentos, III. Metabolismo de Macromoléculas; 134 agrupamentos, IV. Estrutura Celular; 211 agrupamentos, V. Processo Celular; 20 agrupamentos, VI. Elementos Genéticos Móveis; 159 agrupamentos, VII. Patogenicidade, Virulência e Adaptação; 525 agrupamentos, VIII. Hipotéticos; e 62 agrupamentos, IX. ORFs com Categoria

Indefinida (Figura 18). Avaliando agrupamentos com mais de 100 proteínas (Figura 19) verifica-se que estes contêm proteínas principalmente da categoria I (42 genes) e V (31 genes).

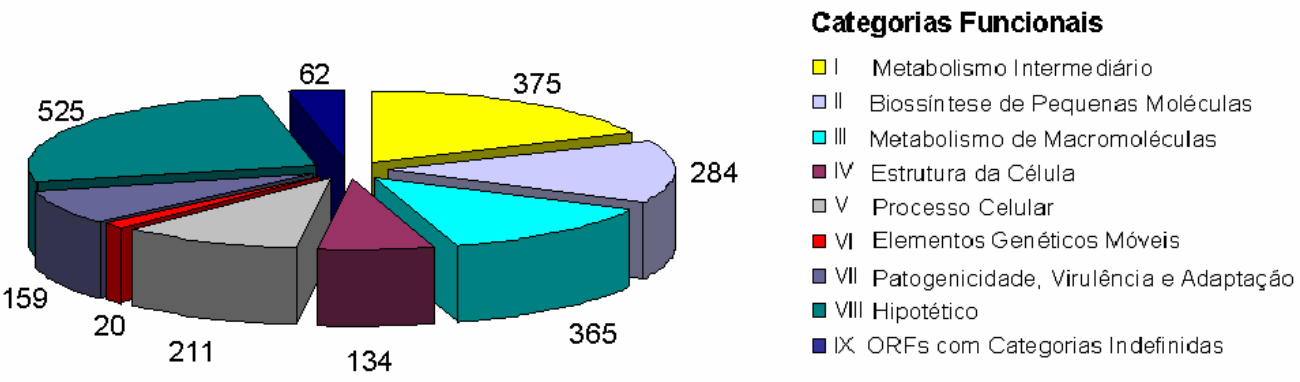

Figura 18. Categorias funcionais presentes em agrupamentos com mais de 10 proteínas.

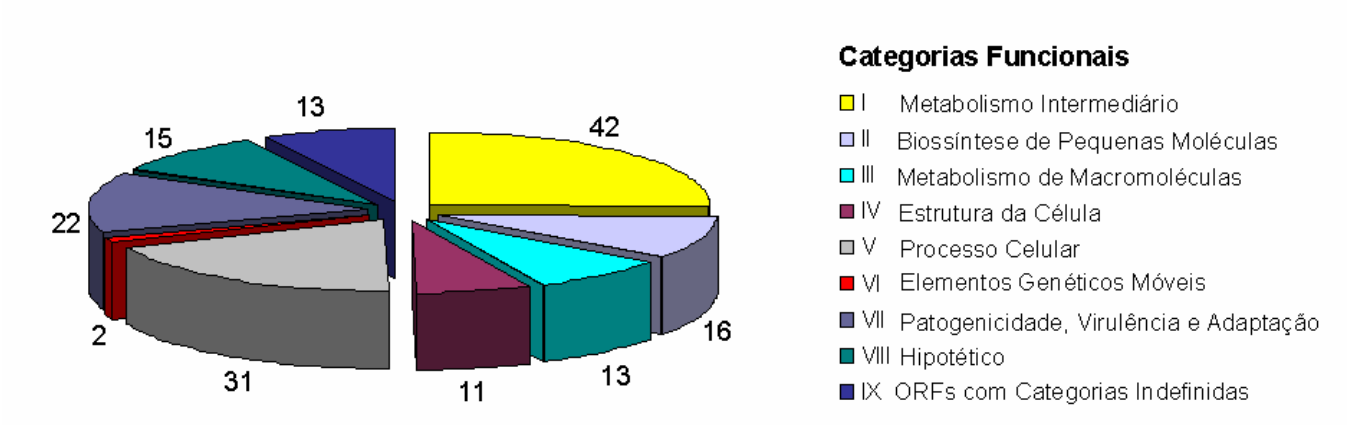

Figura 19. Categorias funcionais presentes em agrupamentos com mais de 100 proteínas.

Os dois agrupamentos que contém mais de 1000 proteínas são PBC278 e PBC206. O agrupamento PBC278 contém reguladores da transcrição de diversas famílias, e uma análise detalhada é necessária para re-agrupar essas proteínas. Já o agrupamento PBC206 contém proteínas que apresentam diferentes funções, no entanto

com base na descrição do produto, a maioria das proteínas está associada à membrana, como por exemplo, transportadores (*MFS permease, transportador ABC, transporter, membrane spanning protein, membrane protein, major facilitator family transporter*).

Existem 10.248 agrupamentos que contém pelo menos uma proteína de uma bactéria associada à planta e desses, 9.640 pelo menos uma proteína de bactérias patogênicas de plantas. Dentre esses agrupamentos, 4.194 agrupamentos continham pelo menos uma proteína do gênero *Xanthomonas*, sendo que essas proteínas puderam ser categorizadas. O sistema conseguiu então categorizar aproximadamente 23,8% dos clusters formados. Como o banco é constituído por bactérias dos mais diversos grupos taxonômicos esse resultado não é surpreendente. Em média cada organismo presente no banco de dados possui 66,9% de suas proteínas em agrupamentos classificados, dentro das associadas às plantas esta média é de 72%, e considerando apenas as bactérias fitopatogênicas, 73,5%. Quando as categorias VIII. Hipotéticas e IX. ORFs com Categoria Indefinida não são consideradas, estas médias são 63%, 67,6%, e 68,8%, respectivamente. Existem agrupamentos com mais de uma classificação funcional, sendo que 37% das proteínas com categorias VIII ou IX também pertencem a uma outra categoria. Além disso, 70,2% dos grupos com 10 ou mais proteínas foram categorizados, enquanto que, 11,7% dos grupos com menos de 10 proteínas foram categorizados. Uma maneira de ser mais eficiente na categorização automática seria combinar a categorização manual de vários genomas de diferentes bactérias e ter um conjunto de genes referência que pudessem ser utilizados durante o processo de agrupamento.

Quando os agrupamentos exclusivos foram analisados, foi verificado que 3.858 proteínas estão exclusivamente presentes em bactérias que estão associadas às plantas e 2.936 em bactérias patogênicas de plantas. Vale ressaltar que estes valores foram

retirados considerando apenas as bactérias presentes no banco, incluindo as de referência. Considerando somente as bactérias associadas às plantas, pode-se observar que 426 proteínas estão presentes na categoria de patogenicidade e adaptação. Fazendo o mesmo para as proteínas dentre esses agrupamentos pertencentes na categoria funcional de patogenicidade e adaptação exclusivas em bactérias patogênicas de plantas (categoria VII), foi obtido um total de 355 proteínas distribuídas em 56 grupos (Tabela 7). Para filtrar ainda mais esses resultados, uma análise poderia ser feita através da comparação dessas seqüências contra o banco de dados NR do GenBank[18].

Tabela 7. Agrupamentos e o produto gênico pertencentes a categoria funcional de patogenicidade, virulência e adaptação em bactérias patogênicas de plantas. Foram retirados os genes hipotéticos e genes com mesma nomenclatura.

Identificação do Agrupamento	Produto
PBC4812	putative signal peptide protein, VirK protein, virA/G regulated gene, PD0855 VirK protein, VirK
PBC5336	general stress protein, truncated general stress protein
PBC5574	acetyltransferase, AttT protein
PBC6304	AMP-ligase, peptide synthase, PD1311 peptide synthase
PBC6339	general secretion pathway protein I, XpsI, general secretion pathway protein GspI, type II secretory pathway pseudopilin
PBC6341	general secretion pathway protein L, general secretory pathway protein L, PD0739 general secretory pathway protein L,PSPTO3310 general secretion pathway protein L, putative, Fimbrial assembly, general secretion pathway protein Gs pL, putative, general secretion pathway protein
PBC6342	general secretion pathway protein M, general secretion pathway protein GspM, putative
PBC6372	pathogenicity-related protein, PD0310 pathogenicity-related protein
PBC6373	HrpX protein, regulatory hrpb transcription regulator protein, AraC-type transcriptional regulator HrpX, regulatory protein HrpX
PBC6386	HpaB protein
PBC6387	HrpD5 protein, hrpW transmembrane protein, HrcD protein
PBC6388	HpaA protein, HRPV PROTEIN, HpaA
PBC6389	HpaP protein, hrpC3, HpaC protein
PBC6390	HrpB1 protein, hrpk protein
PBC6391	HrpB2 protein, hrpj protein
PBC6392	HrpB4 protein, hrph protein
PBC6393	HrpB7 protein, hrpd protein

[18] Informações mais detalhadas a respeito desse banco de dados podem ser encontradas em http://www.ncbi.nlm.nih.gov/Genbank/index.html.

Continua.

PBC6472	polygalacturonase precursor (pectinase) signal peptide protein
PBC6481	GumK protein, glucuronosyltransferase GumK, xanthan biosynthesis glucuronosyltransferase GumK
PBC6482	GumE protein, xanthan biosynthesis exopolysaccharide polymerase GumE
PBC6539	virulence regulator, histone-like nucleoid-structuring protein, putative histone-like nucleoid-tructuring protein
PBC6554	VirB6 protein, component of type IV secretion system
PBC6574	cellulase, endo-1,4-beta-glucanase, extracellular endoglucanase precursor
PBC6576	cellulase, extracellular endoglucanase precursor
PBC6621	TonB-like protein
PBC6640	OmpA-related protein, TonB-dependent outer membrane receptor (C-terminal fragment)
PBC6656	cellulase precursor
PBC6663	methicillin resistance protein, putative penicillinase repressor family protein
PBC6666	avirulence protein AvrBs2
PBC6699	avirulence protein, type III effector HopX1
PBC6709	HrpE, HrpE1 protein
PBC6710	HrpD6 protein
PBC6711	Hpa1 protein, Xanthomonas outer protein A, Hpa1 homolog
PBC6717	nuclease, putative secreted protein
PBC6854	outer membrane component of multidrug efflux pump
PBC6855	RND efflux membrane fusion protein
PBC6856	Toxin secretion ABC transporter ATP-binding protein
PBC6868	TonB-like protein
PBC6884	general stress protein
PBC6995	VirB3 protein
PBC6996	VirB2 protein
PBC7165	mitomycin resistance protein
PBC7209	OmpA-related protein, TonB-dependent outer membrane receptor
PBC10069	avirulence protein
PBC10746	Avirulence protein, type III effector AvrB4
PBC11014	pectate lyase E
PBC11069	Nisin-resistance protein
PBC11098	antifreeze glycopeptide AFGP related protein
PBC11101	virulence protein, Xanthomonas outer protein D
PBC11115	Beta-lactamase
PBC11136	avirulence protein
PBC11151	avirulence protein, avirulence protein AvrBs1
PBC11152	avirulence protein
PBC11182	RpfH protein, putative membrane protein RpfH
PBC11234	RNA polymerase sigma factor, RNA polymerase ECF-type sigma factor
PBC11277	avirulence protein

Caso sejam levadas em conta apenas as bactérias que vivem em associação com plantas, a distribuição dos genes em cada uma das categorias funcionais adotadas para este trabalho é demonstrada na Tabela 8. De maneira geral, a porcentagem das proteínas classificadas dentro das categorias funcionais nos diferentes genomas foi coincidente. Não havendo grande discrepância entre as bactérias listadas.

Tabela 8. Análise comparativa da distribuição dos genes (normalizada pelo tamanho do genoma) por categoria dos genomas de bactérias associadas às plantas.

Organismos	Categorias Funcionais (%) *								
	I. Metabolismo intermediário	**II. Biossíntese de Pequenas Moléculas**	**III. Metabolismo de Macromoléculas**	**IV. Estrutura Celular**	**V. Processo Celular**	**VI. Elementos Genéticos Móveis**	**VII. Patogenicidade, Virulência e Adaptação**	**VIII. Hipotéticos**	**IX. ORFs com Categoria Indefinida**
A. tumefaciens C58 (Cereon)	34,7	17,1	17,5	8,9	19,7	0,5	11	20,2	9,9
A. tumefaciens C58 (Wash)	34,7	17,2	17,9	8,9	19,6	0,6	11,1	19,9	9,9
M. loti	35,9	18,7	18,6	8,0	16,8	1,0	12,0	21,7	10,2
S. meliloti	36,5	17,6	17,2	8,3	18,7	0,6	11,0	20,2	9,6
R. solanacearum	30,9	16,2	17,3	9,1	16,8	2,1	14,1	21,4	7,9
*E. carotovora s*ubsp. *atroseptica*	27,6	16,7	18,6	8,7	20,3	1,8	12,2	18,6	7,4
P. syringae phaseolicola 1448A	29,7	16,3	18,2	9,9	18,5	1,1	12,9	21,2	9,0
P. syringae pv. B728a	30,6	16,3	18,1	9,6	18,	0,7	12,7	21,3	9,2
P. yringae pv. *tomato* DC3000	30,3	16	17,9	9,4	18,6	1,4	12,2	21,6	8,8
X. axonopodis pv. *citri* str 306	22,6	11,9	15,5	8	12,9	1,2	10,9	35,3	5,7
X. campestris 8004	20,6	10,7	14,6	7,1	11,4	2	10	38,8	5,8
X. campestris ATCC33913	20,5	10,6	14,4	7,1	11,3	2,5	9,8	38,9	5,7
X. campestris pv. *vesicatoria* str. 85-10	22,8	12	15,7	7,8	12,7	1,4	11	34,5	5,9
X. oryzae KACC10331	20,7	12,6	16,8	7,7	12,7	2,2	10,6	33	5,2
X. oryzae MAFF 311018	20,4	12,5	16,9	7,7	12,3	2,6	10,3	33,6	5,1
X. fastidiosa 9a5C	18,1	16	23,7	9,7	10,8	1,6	10,6	25,8	4,6
X. fastidiosa Temecula1	17,2	16,4	23,8	9,8	10,4	1,7	9,9	26,2	4,1
L. xyli	25,8	19,2	28,1	7,6	15,9	0,8	8	16,3	6,1

*** Existem proteínas que são classificadas em mais de uma categoria funcional.**

As bactérias *A. tumefaciens*, *M. loti*, *S. meliloti*, *R. solanacearum* e as cepas do gênero *Pseudomonas* apresentam valores acima de 30% das proteínas dentro da categoria de metabolismo intermediário. Quando estes valores são comparados aos descritos por Van Sluys e colaboradores (2002), embora os valores percentuais dos trabalhos sejam diferentes, existe uma correlação nos dados informados, já que os organismos *A. tumefaciens* e *R. solanacearum* são descritos como aqueles de maiores valores (23,2% e 17,4%, respectivamente). Este fato provavelmente está associado à presença de um conjunto de vias metabólicas necessárias à sua adaptação a diferentes condições ambientais. Interessante notar que, embora existam pequenas diferenças no número de genes da categoria de elementos genéticos móveis entre os genomas há uma distinção no tipo de elemento presente em cada organismo. É conhecida a presença de elementos de transposição no gênero *Xanthomonas* e quando comparada a outro organismo como, por exemplo, as do gênero *Xylella*, fica claro a presença de genes que codificam proteínas de fagos (VAN SLUYS *et al.*, 2002).

A lista de genes apresentada (Tabela 7) e os dados citados nessa seção demonstram o potencial desse banco de dados em apresentar proteínas que possam ser mais bem estudadas em análises funcionais. Alguns dos resultados citados acima podem ser obtidos utilizando a Linguagem Declarativa do ProBacter, como por exemplo, a quantidade de agrupamentos com tamanhos maiores que um determinado número de proteínas (Exemplo 2, Seção 5.1.2), informações a respeito de quantas proteínas de um certo organismo estão presentes em uma dada categoria (Exemplo 3, Seção 5.1.2), quantidade de múltiplas cópias presentes no banco de dados (Exemplo 4, Seção 5.1.2) ou de um determinado organismo, dentre outros.

5.2. Aplicação

Para exemplificar a utilização do sistema ProBacter optou-se por apresentar alguns casos que possam demonstrar o funcionamento do sistema desenvolvido.

5.2.1. Análise da Proteína *hrpX*

A proteína codificada pelo gene *hrpX*, está envolvida na regulação da expressão do grupo de genes responsáveis por codificar proteínas associadas a formação do sistema de secreção do tipo III (TTSS). O TTSS é essencial para a virulência em muitas bactérias Gram-negativas que infectam plantas, animais e humanos. Esses patógenos utilizam o TTSS para injetar proteínas conhecidas como efetoras dentro da célula hospedeira para sobrepor as defesas do hospedeiro (HUECK, 1998; GALLAN & COLMER, 1999). Os patógenos de plantas são capazes de translocar essas moléculas efetoras mesmo através da parede celular, ausente em células de animais. Os componentes desse sistema em plantas são codificados por genes *hrp* (*hypersentive response and pathogenicity*). As proteínas efetoras secretadas têm a habilidade de elicitar à reposta de hipersensibilidade (HR) em plantas resistentes e de patogenicidade em plantas hospedeiras (ALFANO & COLLMER, 1997). Fora do agrupamento de genes *hrp* estão genes com função regulatórias, como por exemplo, *hrpX* e *hrpG*. Essas proteínas já foram descritas em espécies de diferentes dos gêneros *Xanthomonas* (WENGELNIK & BONAS, 1996; NOËL *et al*., 2002) e *Ralstonia* (CUNNAC *et al*., 2004) como reguladores da expressão de vários dos operons encontrados dentro do agrupamento de *hrp*.

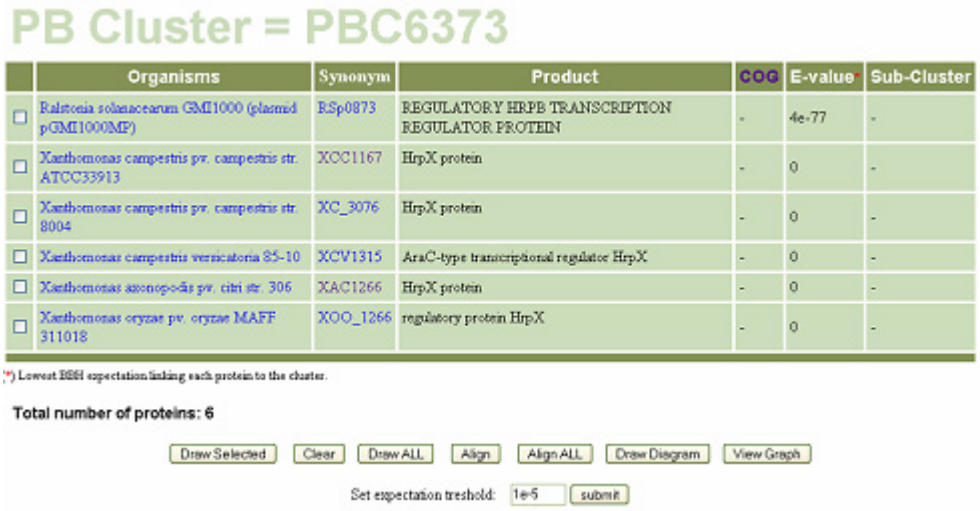

Figura 20. Demonstração do agrupamento PBC6373 que contém o gene *hrpX* descrito.

Uma consulta ao banco de dados utilizando o *Gene Search* procurando por *hrpX* em C*lusters* resulta de dois agrupamento PBC6373 e PBC1483. O agrupamento PBC6373 contém o gene *hrpX* descrito, e uma variação com relação anotação nos diferentes organismos pode ser verificada (Figura 20). O agrupamento PBC1483 contém 15 entradas, e potencialmente contém genes que codificam um regulador da transcrição que poderia ser estudado mais detalhadamente quanto a sua função em relação ao cluster de genes *hrp* (Figura 21). Como pode ser observado, o resultado da anotação dessas proteínas nos diversos projetos é bastante diferente, incluindo uma anotação como proteína hipotética. Neste caso, o sistema permite que o usuário avalie a melhor anotação para o gene comparando-o com outros organismos. Com relação as proteínas encontradas nas diferentes espécies de *Xanthomonas*, foram selecionadas as seqüências para fazer um alinhamento múltiplo diretamente na página. O resultado do alinhamento utilizando o programa CLUSTALW (Figura 22) indica que pode existir um problema na predição do códon de iniciação do gene *hprX* da *Xanthomonas oryzae* quando comparada com ortólogos de outras espécies de *Xanthomonas*. Analisando

comparativamente os genes vizinhos nos genomas dos diferentes organismos pode-se verificar que este gene está localizado em uma região altamente conservada, no entanto, esta região está inserida em um local de inversão no genoma de *Xanthomonas campestris* pv. *campestris* quando comparada a *Xanthomonas axonopodis* pv. *citri* e a *Xanthomonas campestris* pv. *vesicatoria* (Figura 22).

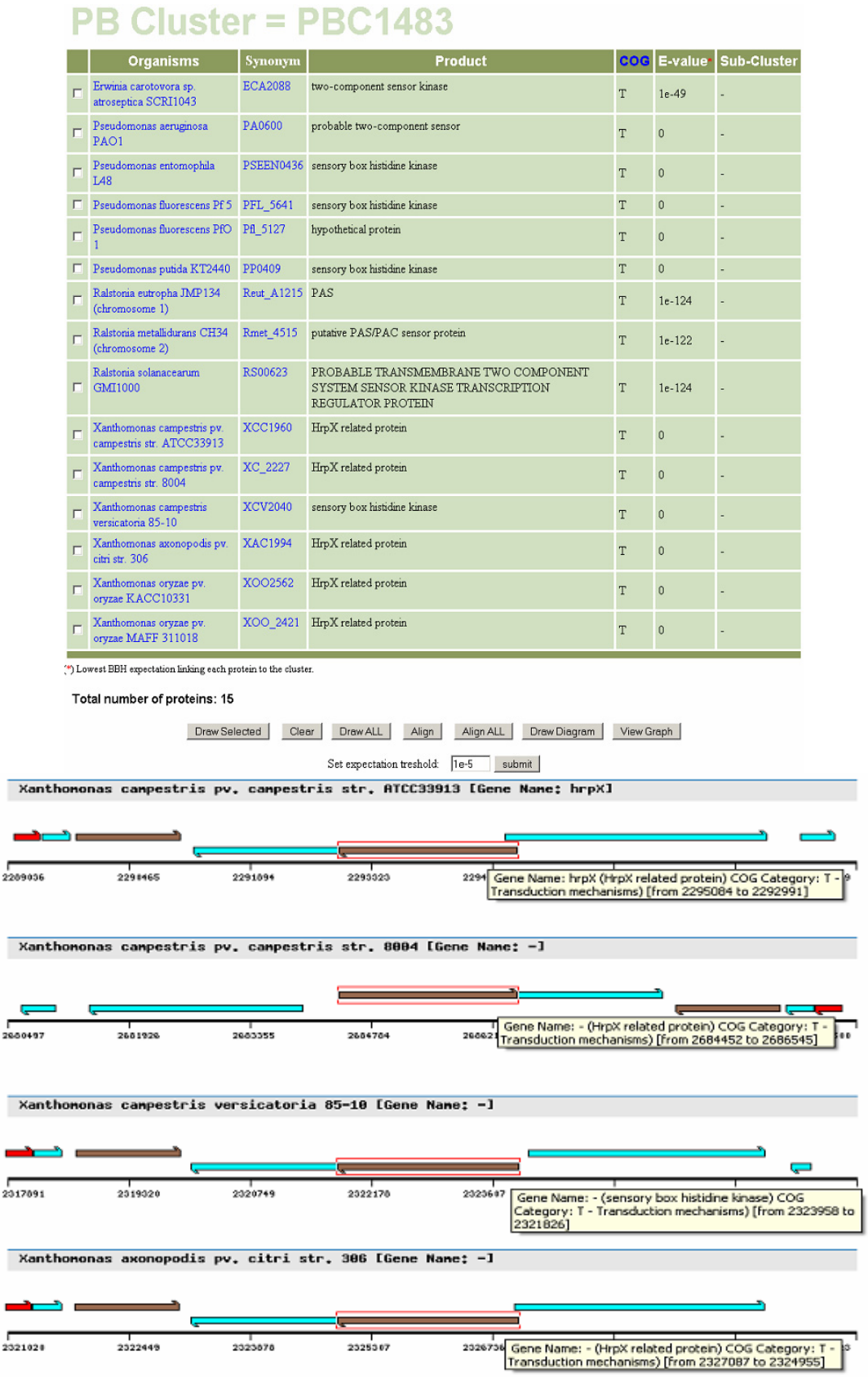

Figura 21. Demonstração do agrupamento PBC1483 contendo 15 entradas e que potencialmente contém genes que codificam um regulador da transcrição.

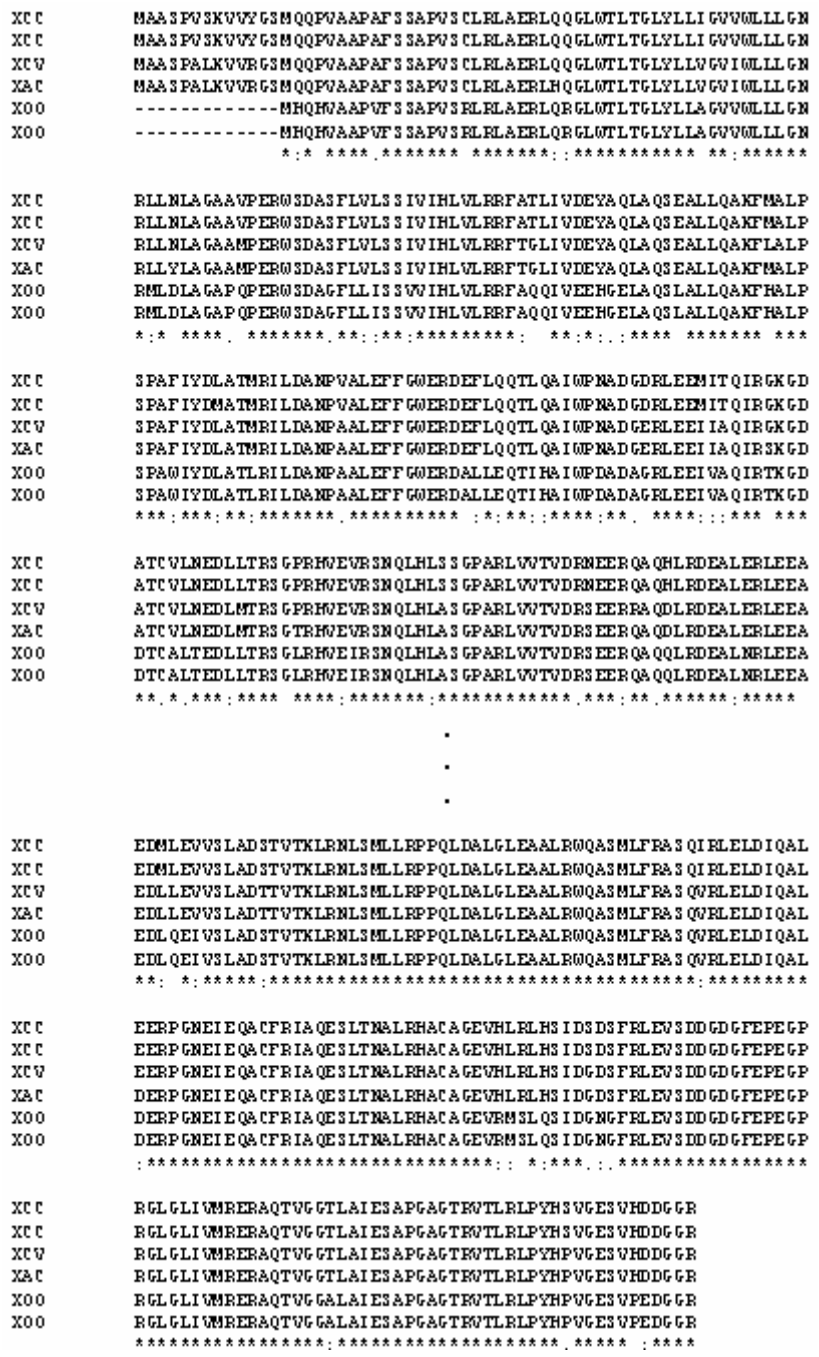

Figura 22. Resultado do alinhamento múltiplo utilizando o programa ClustalW, integrado ao sistema ProBacter, de cepas do gênero *Xanthomonas*.

5.2.2. Análise da proteína *hrcU*

Os genes *hrp* que codificam componentes conservados do TTSS foram renomeados para *hrc* (HR *e conserved*) (BOGDANOVE *et al*., 1996). Entre esses genes está o *hrcU*, que codifica uma das principais unidades estruturais do sistema. Um dos mais importantes métodos de análise do ProBacter é a ferramenta de agrupamento. Utilizando o gene *hrcU* nas buscas por agrupamentos, resultou no grupo PBC509 contendo 39 entradas (Figura 23). O resultado do agrupamento (Figura 24 e 25a) pode ser submetido a um novo agrupamento, gerando grupos menores com um maior grau de similaridade. O re-agrupamento dividiu o agrupamento PBC509 em 4 grupos menores: (*i*) contendo proteínas somente de espécies do gênero *Pseudomonas*, (*ii*) proteínas somente de cepas de *Xanthomonas* e dois grupos menores contendo proteínas de (*iii*) *Agrobacterium* e *Sinorhizobium* e (*iv*) contendo as proteínas de *E. coli*, *Erwinia* e *Ralstonia*. Esses grupos menores podem então ser analisados separadamente (Figura 25b).

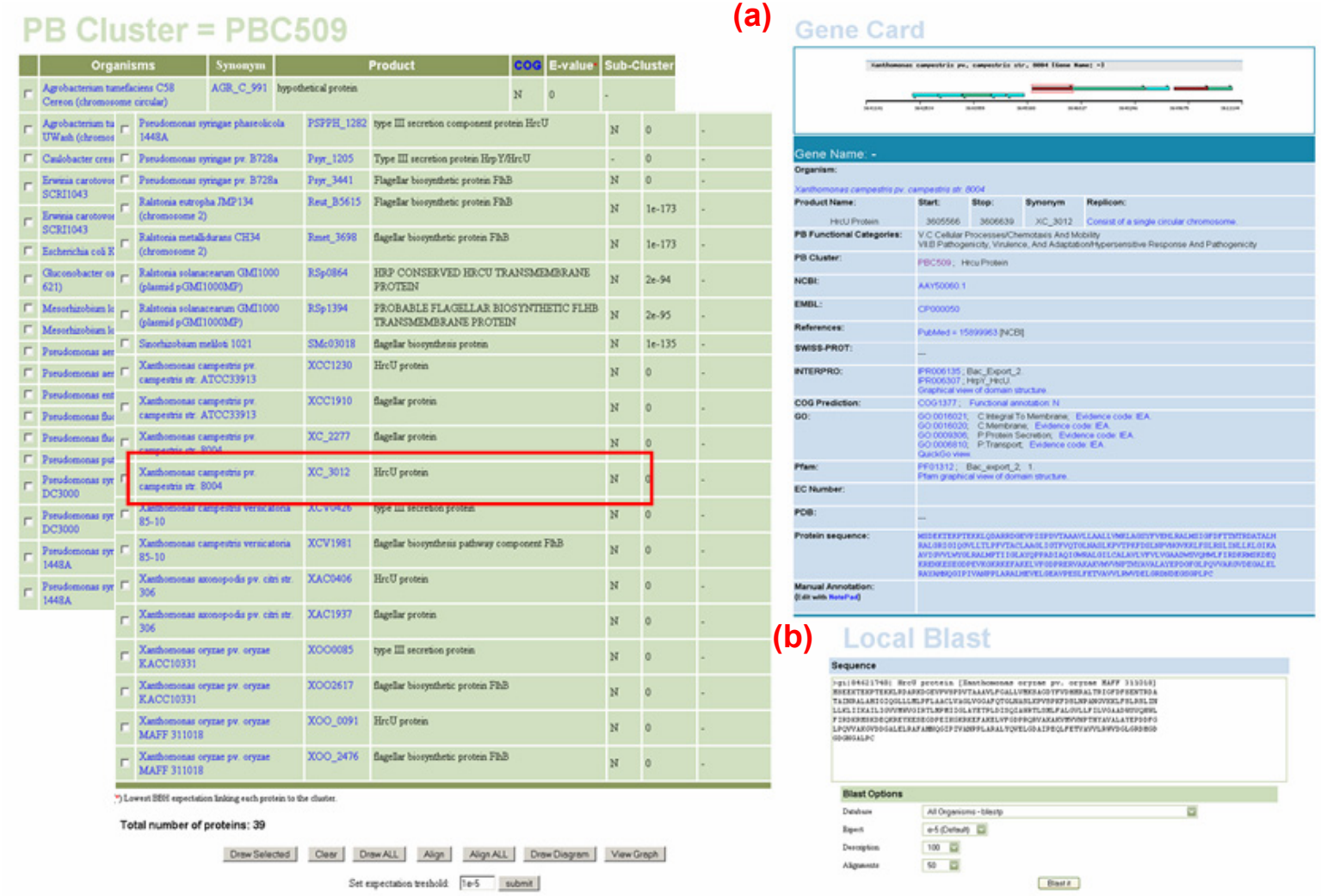

Figura 23. Visão geral dos resultados obtidos a partir da consulta por agrupamentos com *HrcU*, indicando os seguintes valores: o agrupamento presente no banco de dados, o grupo de organismos, o produto, classificação de acordo com o COG, valor esperado e o sub-cluster (neste caso foi indicado o valor 1e-100 no campo *Set expectation treshold*). As figuras (a) e (b) são vistas quando o usuário pede pela informação específica de um gene, sendo então direcionado para o módulo de visualização *Gene Card* (a), onde pode fazer BLAST local (b) com a seqüência de proteína contra parte ou todo o banco de dados. A interface do *Local Blast* foi modificada daquela utilizada pelo programa SABIA (ALMEIDA et al., 2004) para atender aos propósitos do sistema ProBacter.

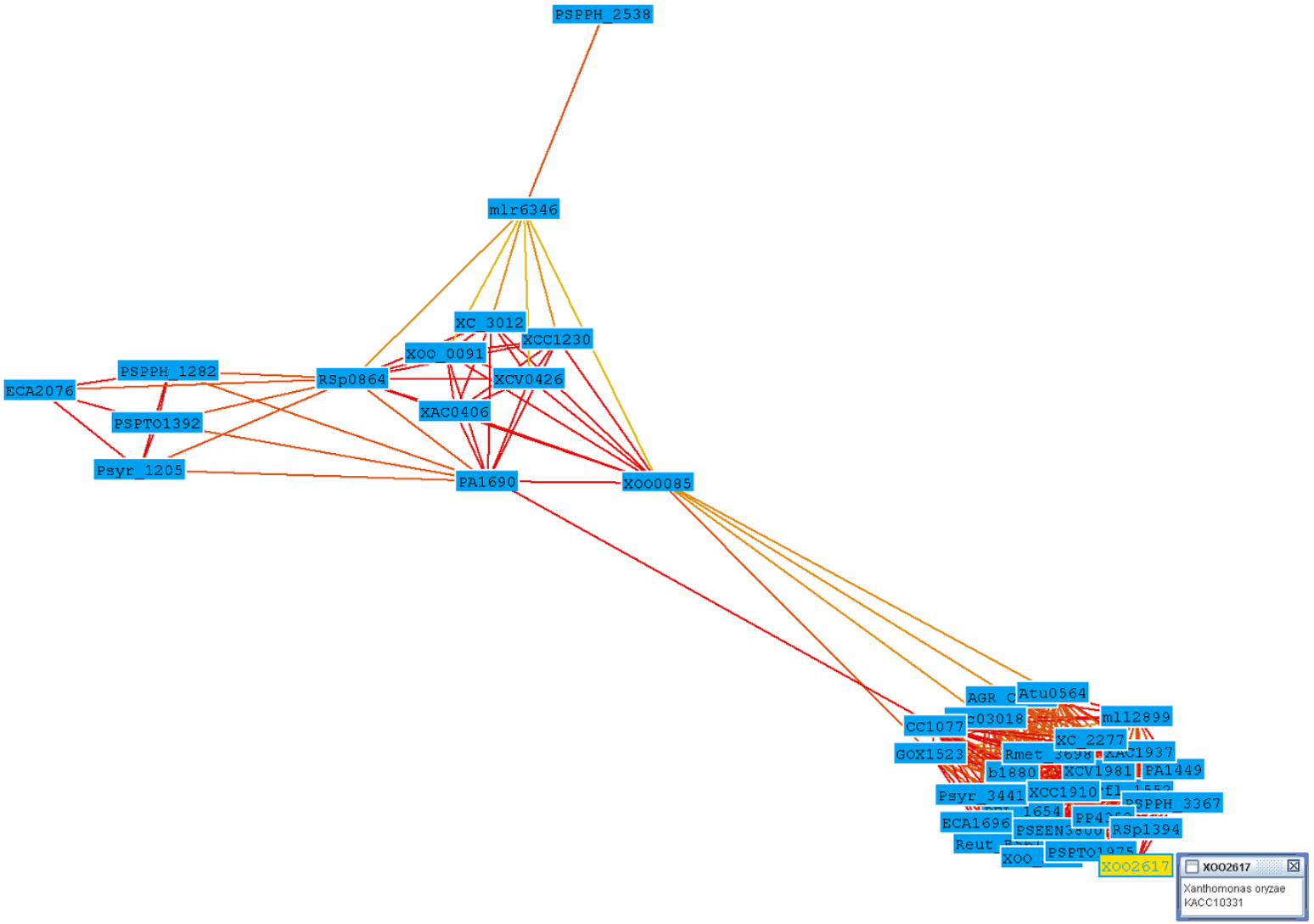

Figura 24. Interface interativa de visualização dos agrupamentos. Cada ligação entre dois nós corresponde a um BBH. A qualidade de cada ligação é representada por uma escala de cores que varia entre a cor laranja (alto e-value) à cor vermelha (baixo e-value). Quando algum gene é selecionado um boxe aparece (como o indicado) informando o organismo a que pertence, e também é direcionado ao *Gene Card*.

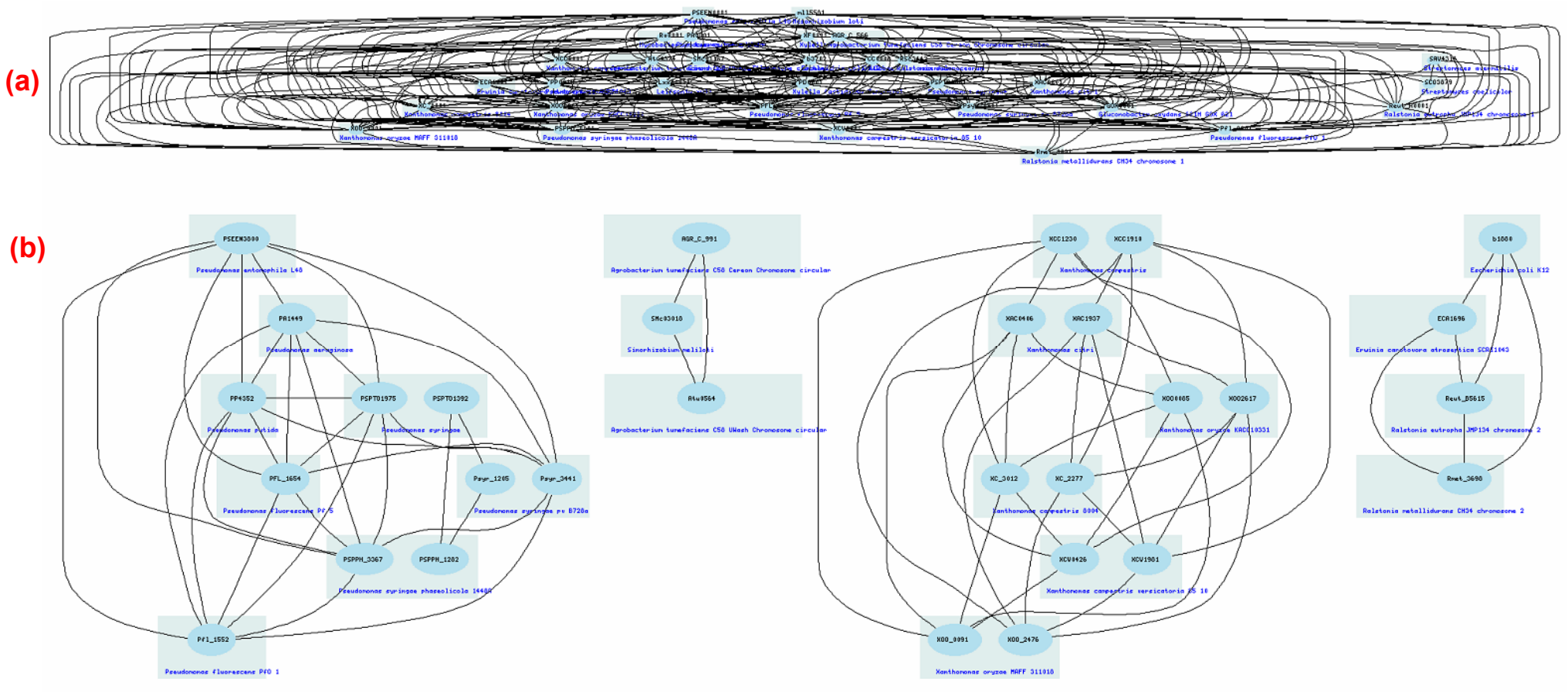

Figura 25. Diagrama que mostra a relação entre os genes dado pelo agrupamento.Cada ligação entre os nós acima corresponde ao um BBH entre os dois genes. (a) Esta figura, assim como a apresentada na figura 23, demonstra a disposição do agrupamento formado; (b) Demonstra o resultado do re-agrupamento, considerando um e-value de e-100.

6. CONCLUSÃO

Sistemas de informação evoluem com as mudanças de contexto e a necessidade dos usuários. Isso é ainda mais crítico em bioinformática, devido ao avanço rápido nessa área. Para atender essa demanda, estes sistemas precisam ser extensíveis e flexíveis. O modelo de dados ProBacter, um sistema de informação, contém as quatro características básicas que compõem este tipo de sistema: armazena as informações utilizando um sistema gerenciador de banco de dados, utiliza ferramentas para administração dos dados, fornece módulos de busca, como por exemplo, *Gene Search* e LDP, e o disponibiliza em um ambiente gráfico amigável. Esta base de dados atendeu também aos requisitos de extensibilidade, permitindo que outros sistemas sejam futuramente integrados, e os de flexibilidade, ou seja, foi capaz de acomodar a adição de novos campos e ainda promover a indexação de tabelas já estruturadas. Estes são aspectos essenciais para a construção de um modelo de dados.

Este banco de dados oferece um sistema para análise comparativa de genomas microbianos, em particular de bactérias associadas às plantas. Os métodos de consulta implementados possibilitaram a obtenção de informações tanto por consultas pré-definidas como por perguntas "livres" (LDP) ao banco de dados, permitindo que o usuário desenvolva sua consulta sem que tivesse conhecimento de como esta base de dados foi implementada. O sistema ProBacter possibilitou que fossem feitas buscas por informações de similaridade e alinhamento de seqüências, devido a sua integração com ferramentas como BLASTP, ClustalW e T-Coffee. Ele forneceu também um método de análise dos agrupamentos gerados, permitindo o acesso à lista de pares ou grupos de seqüências homólogas, identificando possíveis relações evolutivas, estruturais e funcionais existentes entre as seqüências. Ele permitiu a anotação manual de um dado

gene, possibilitando que este sistema torne-se uma ferramenta para auxiliar na identificação de erros propagados por anotação automática.

Através do sistema ProBacter podemos chegar as seguintes conclusões:

- O banco de dados foi criado para conter as informações de maneira uniforme de genomas completos de bactérias associadas a plantas e permitir a análise comparativa entre as informações armazenadas;
- Para cada uma das proteínas depositadas no banco de dados, foram incluídas informações de referência cruzada com os bancos de dados: PDB, Swiss-Prot, GenBank, PubMed, COG, GO, EMBL, InterPro, Ec_Number. Com um total de 9 bancos de dados e 1.242.781 referências cruzadas;
- Os métodos de agrupamento e de transferência automática da categorização funcional permitiram a categorização de 64,6% das proteínas no banco de dados. Sendo que a maioria dos agrupamentos formados apresentou menos de 10 proteínas;
- Os métodos de buscas implementados permitiram a extração de informações de análise comparativa, tanto de maneira quantitativa quanto qualitativa. Entre as informações obtidas, destacamos os 56 agrupamentos compostos por 355 proteínas pertencentes exclusivamente a bactérias fitopatogênicas e classificadas como associadas à patogenicidade e adaptação ao hospedeiro;
- Foram integradas ao banco de dados, ferramentas que auxiliam na análise comparativa de seqüências como BLAST, ClustalW, T-Coffee; ferramentas que permitem a visualização gráfica e comparativa do posicionamento do gene no genoma, assim como a formação de agrupamentos pelo método utilizado.

Assim, o presente trabalho visou responder aos principais questionamentos já mencionados de forma clara e fácil, tendo o potencial de abordar temas recorrentes como a interação hospedeiro-patógeno, patogenicidade, a determinação das funções de genes hipotéticos por meio dos agrupamentos, dentre outros, através de uma avaliação qualitativa e quantitativa estando baseada na similaridade de seqüências.

6.1. Perspectivas Futuras

- Aplicar outras técnicas de agrupamento dos genes mais elaboradas, para gerar grupos de proteínas mais específicos.
- Implementar algoritmos de mineração de dados e reconhecimento de padrões, bem como o desenvolvimento de novos algoritmos de acordo com a necessidade para a extração de novas informações do banco de dados.
- Estender a Linguagem Declarativa do ProBacter (LDP) com o intuito de ir além da linguagem SQL, integrando os algoritmos de mineração de dados.
- Desenvolver scripts de atualização automática para que novas informações sejam integradas ao sistema.
- Fornecer, dentro do contexto da análise comparativa, detecção e correção de erros da anotação.
- Expandir os métodos de anotação manual, como por exemplo, a possibilidade de edição das categorias funcionais e o armazenamento do histórico da alteração das informações provenientes das anotações.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALFANO, J. R., COLLMER, A. (1997). **Free in PMC The type III (Hrp) secretion pathway of plant pathogenic bacteria: trafficking harpins, Avr proteins, and death.** *J Bacteriol.* 179(18), 5655-62.

ALM, E. J., HUANG, K. H., PRICE, M. N., KOCHE, R. P., KELLER, K., DUBCHAK, I. L. & ARKIN, A. P. (2005). **The MicrobesOnline Web site for comparative genomics.** *Genome Res.* 15(7), 1015–1022.

ALMEIDA, L.G., PAIXAO, R., SOUZA, R.C., COSTA, G.C., BARRIENTOS, F.J., SANTOS, M.T., ALMEIDA, D.F., VASCONCELOS, A.T. (2004) **A System for Automated Bacterial (genome) Integrated Annotation--SABIA**. *Bioinformatics*. 20(16):2832-3.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990). **Basic local alignment search tool.** *J Mol Biol* 215, 403-10.

ATTWOOD. T. K., FLOWER, Dr., LEWIS, A. P., MABEY, J. E., MORGAN, S. R., SCORDIS, P., SELLEY, J. N. & WRIGHT, W. (1999). **PRINTS prepares for the new millennium.** *Nucleic Acids Res*. 27, 220-225.

BAIROCH C. H., WU W. C., BARKER B., BOECKMANN S. A., FERRO E., GASTEIGER H., HUANG R., LOPEZ M., MAGRANE M. J., MARTIN D. A., NATALE C. O. & RSQUO; DONOVAN, NICOLE REDASCHI, AND LAI-SU L. Yeh. (2004). **UniProt: the Universal Protein knowledgebase**. Nucleic Acids Res. 32: D115 - D119.

BATEMAN, A., COIN, L., DURBIN, R., FINN, R. D., HOLLICH, V., GRIFFITHS-JONES, S., KHANNA, A., MARSHALL, M., MOXON, S., SONNHAMMER, E. L. L., STUDHOLME, D. J., YEATS, C. & EDDY, S. R. (2004). **The Pfam protein families database.** *Nucleic Acids Res*. 32, 138-141.

BELL, K.S., SEBAIHIA, M., PRITCHARD, L., HOLDEN, M.T., HYMAN, L.J., HOLEVA, M.C., THOMSON, N.R., BENTLEY, S.D., CHURCHER, L.J., MUNGALL, K., ATKIN, R., BASON, N., BROOKS, K., CHILLINGWORTH, T., CLARK, K., DOGGETT, J., FRASER, A., HANCE, Z., HAUSER, H., JAGELS, K., MOULE, S., NORBERTCZAK, H., ORMOND, D., PRICE, C., QUAIL, M.A., SANDERS, M., WALKER, D., WHITEHEAD, S., SALMOND, G.P., BIRCH, P.R., PARKHILL, J., TOTH, I.K. (2004**). Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors.** Proc. Natl. Acad. Sci. 101(30), 11105-10.

BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & WHEELER, D. L. (2007). **GenBank** *Nucl. Acids Res* 35, 21-25

BENTLEY, S. D., CHATER, K. F., CERDENO-TARRAGA, A. M., CHALLIS, G. L., THOMSON, N. R., JAMES, K. D., HARRIS, D. E., QUAIL, M. A., KIESER, H., HARPER, D., BATEMAN, A., BROWN, S., CHANDRA, G., CHEN, C. W., COLLINS, M., CRONIN, A., FRASER, A., GOBLE, A., HIDALGO, J., HORNSBY, T., HOWARTH, S., HUANG, C. H., KIESER, T., LARKE, L., MURPHY, L., OLIVER, K., O'NEIL, S., RABBINOWITSCH, E., RAJANDREAM, M. A., RUTHERFORD, K., RUTTER, S., SEEGER, K., SAUNDERS, D., SHARP, S., SQUARES, R., SQUARES, S., TAYLOR, K., WARREN, T., WIETZORREK, A., WOODWARD, J., BARRELL, B. G., PARKHILL, J., HOPWOOD, D. A. (2002). **Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2).** *Nature*. 417(6885), 141-7.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. (2000). **The Protein Data Bank**. *Nucleic Acids Res*. 28, 235-242.

BINNEWIES, T.T., MOTRO, Y., HALLIN, P.F., LUND, O., DUNN, D., LA, T., HAMPSON, D.J., BELLGARD, M., WASSENAAR, T.M., USSERY, D.W. (2006). **Ten years of bacterial genome sequencing: comparative-genomics-based discoveries**. Funct. Integr. Genomics 6(3), 165-85.

BLATTNER, F. R., PLUNKETT, G. 3RD., BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B., SHAO, Y. (1997). **The complete genome sequence of *Escherichia coli* K-12.** *Science.* 277(5331), 1453-74

BOECKMANN B., BAIROCH A., APWEILER R., BLATTER M.-C., ESTREICHER A., GASTEIGER E., MARTIN M. J., MICHOUD K., O'DONOVAN C., PHAN I., PILBOUT S. & SCHNEIDER M. (2003). **The SWISS-PROT protein knowledgebase and its supplement TrEMBL.** *Nucleic Acids Res*. 31, 365-370.

BOGDANOVE, A.J., BEER, S.V., BONAS, U., BOUCHER, C.A., COLLMER, A., COPLIN, D.L., CORNELIS, G.R., HUANG, H.C., HUTCHESON, S.W., PANOPOULOS, N.J. & VAN GIJSEGEM, F. (1996). **Unified nomenclature for broadly conserved hrp genes of phytopathogenic bacteria**. Mol. Microbiol. 20(3), 681-3.

BUELL, C.R., JOARDAR, V., LINDEBERG, M., SELENGUT, J., PAULSEN, I.T., GWINN, M.L., DODSON, R.J., DEBOY, R.T., DURKIN, A.S., KOLONAY, J.F., MADUPU, R., DAUGHERTY, S., BRINKAC, L., BEANAN, M.J., HAFT, D.H., NELSON, W.C., DAVIDSEN, T., ZAFAR, N., ZHOU, L., LIU, J., YUAN, Q., KHOURI, H., FEDOROVA, N., TRAN, B., RUSSELL, D., BERRY, K., UTTERBACK, T., VAN AKEN, S.E., FELDBLYUM, T.V., D'ASCENZO, M., DENG, W.L., RAMOS, A.R., ALFANO, J.R., CARTINHOUR, S., CHATTERJEE, A.K., DELANEY, T.P., LAZAROWITZ, S.G., MARTIN, G.B., SCHNEIDER, D.J., TANG, X., BENDER, C.L.,

WHITE, O., FRASER, C.M., COLLMER, A. (2003). **The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000.** Proc. Natl. Acad. Sci. 100(18), 10181-6.

BURKS, C, FICKETT, J. W., GOAD, W. B, LEWITTER F. I, RINDONE W. P., SWINDELL C. D, AND TUNG C. S. (1986). **The GenBank genetic sequence Databank**. *Nucleic Acids Res*. 14: 1 - 4.

CASPI, R., FOERSTER, H., FULCHER, C. A., HOPKINSON, R., INGRAHAM, J., KAIPA, P., KRUMMENACKER, M., PALEY, S., PICK, J., RHEE, S. Y., TISSIER, C., ZHANG, P. & KARP, P. D. (2006). **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res*. 34, 511-516.

CLAMP, M., CUFF, J., SEARLE, S. M. & BARTON, G. J. (2004) **The Jalview Java Alignment Editor.** *Bioinformatics*. 20, 426-7**.**

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S. V., EIGLMEIER, K., GAS, S., BARRY, C. E. 3RD., TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES, S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, A., MCLEAN, J., MOULE, S., MURPHY, L,, OLIVER, K., OSBORNE, J., QUAIL, M. A., RAJANDREAM, M. A., ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S., SULSTON, J. E., TAYLOR, K., WHITEHEAD, S., BARRELL, B. G. (1998). **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature*. 393(6685), 515-6.

CORPET, F., SERVANT, F., GOUZY, J., KAHN, D. (2000). **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res.* 28:267-269

CUNNAC, S., BOUCHER, C., GENIN, S. (2004). **Characterization of the cis-acting regulatory element controlling HrpB-mediated activation of the type III secretion system and effector genes in *Ralstonia solanacearum*.** *J Bacteriol.* 186(8), 2309-18.

DA SILVA, A. C., FERRO, J. A., REINACH, F. C., FARAH, C. S., FURLAN, L. R., QUAGIO, R. B., MONTEIRO-VITORELLO, C. B., VAN SLUYS, M. A., ALMEIDA, N. F., ALVES, L. M., DO AMARAL, A. M., BERTOLINI, M. C., CAMARGO, L. E., CAMAROTTE, G., CANNAVAN, F., CARDOZO, J., CHAMBERGO, F., CIAPINA, L. P., CICARELLI, R. M., COUTINHO, L. L., CURSINHO-SANTOS, J. R., EL-DORRY, H., FARIA, J. B., FERREIRA, A. J., FERREIRA, R. C., FERRO, M. I., FORMIGHIERI, E. F., FRANCO, M. C., GREGGIO, C. C., GRUBER, A., KATSUYAMA, A. M., KISHI, L. T., LEITE, R. P., LEMOS, E. G., LEMOS, M. V., LOCALI, E. C., MACHADO, M. A., MADEIRA, A.M., MARTINEZ-ROSSI, N.M., MARTINS, E. C., MEIDANIS, J., MENCK, C. F., MIYAKI, C. Y., MOON, D. H., MOREIRA, L. M., NOVO, M. T., OKURA, V. K., OLIVEIRA, M. C., OLIVEIRA, V. R.,

PEREIRA, H. A., ROSSI, A., SENA, J. A., SILVA, C., DE SOUZA, R. F., SPINOLA, L. A., TAKITA, M. A., TAMURA, R. E., TEIXEIRA, E. C., TEZZA, R. I., TRINDADE DOS SANTOS, M., TRUFFI, D., TSAI, S. M., WHITE, F. F., SETUBAL, J. C. & KITAJIMA, J. P. (2002). **Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities.** *Nature* 417, 459-63

DIETERICH, G., KÄRST, U., FISCHER, E., WEHLAND, J. & JÄNSCH, L. (2006). **LEGER: knowledge database and visualization tool for comparative genomics of pathogenic and non-pathogenic Listeria species.** *Nucleic Acids Res*. 34, 402-406.

DIGIAMPIETRI, L. A., MEDEIROS, C. B. & SETUBAL, J. C. (2003). **A data model for comparative genomics.** *Ver. Tecnol. Informat.* 3, 35-40.

FEIL, H., FEIL, W.S., CHAIN, P., LARIMER, F., DIBARTOLO, G., COPELAND, A., LYKIDIS, A., TRONG S., NOLAN, M., GOLTSMAN, E., THIEL, J., MALFATTI, S., LOPER, J.E., LAPIDUS, A., DETTER, J.C., LAND, M., RICHARDSON, P.M., KYRPIDES, N.C., IVANOVA, N., LINDOW, S.E. (2005). **Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. tomato DC3000**. Proc. Natl. Acad. Sci. 102(31), 11064-9.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A., MERRICK, J. M., *et al.* (1995). **Whole-genome random sequencing and assembly of Haemophilus influenzae Rd**. *Science.* 269(5223):496-512.

FRASER, C. M., GOCAYNE, J. D., WHITE, O., ADAMS, M. D., CLAYTON, R. A., FLEISCHMANN, R. D., BULT, C. J., KERLAVAGE, A. R., SUTTON, G., KELLEY, J. M., FRITCHMAN, R. D., WEIDMAN, J. F., SMALL, K. V., SANDUSKY, M., FUHRMANN, J., NGUYEN, D., UTTERBACK, T. R., SAUDEK, D. M., PHILLIPS, C. A., MERRICK, J. M., TOMB, J. F., DOUGHERTY, B. A., BOTT, K. F., HU, P. C., LUCIER, T. S., PETERSON, S. N., SMITH, H. O., HUTCHISON, C. A. 3RD, VENTER, J. C. (1995). **The minimal gene complement of Mycoplasma genitalium.** *Science***.** 270(5235):397-403.

GÁLAN, J. E. & COLMER, A. (1999). **Type III secretion machines: bacterial devices for protein delivery into host cells.** *Science.* 284, 1322-1328.

GALIBERT, F., FINAN, T.M., LONG, S.R., PUHLER, A., ABOLA, P., AMPE, F., BARLOY-HUBLER, F., BARNETT, M.J., BECKER, A., BOISTARD, P., BOTHE, G., BOUTRY, M., BOWSER, L., BUHRMESTER, J., CADIEU, E., CAPELA, D., CHAIN, P., COWIE, A., DAVIS, R.W., DREANO, S., FEDERSPIEL, N.A., FISHER, R.F., GLOUX, S., GODRIE, T., GOFFEAU, A., GOLDING, B., GOUZY, J., GURJAL, M., HERNANDEZ-LUCAS, I., HONG, A., HUIZAR, L., HYMAN, R.W., JONES, T., KAHN, D., KAHN, M.L., KALMAN, S., KEATING, D.H., KISS, E., KOMP, C., LELAURE, V.,

MASUY, D., PALM, C., PECK, M.C., POHL, T.M., PORTETELLE, D., PURNELLE, B., RAMSPERGER, U., SURZYCKI, R., THEBAULT, P., VANDENBOL, M., VORHOLTER, F.J., WEIDNER, S., WELLS, D.H., WONG, K., YEH, K.C., BATUT, J. (2001). **The composite genome of the legume symbiont *Sinorhizobium meliloti***. Science 293(5530), 668-72.

GALPERIN, M. Y. (2007). **The Molecular Biology Database Collection: 2007 update.** *Nucl. Acids Res.* 35, 3-4.

GATTIKER, A., MICHOUD, K., RIVOIRE, C., AUCHINCLOSS, A. H., COUDERT, E., LIMA, T., KERSEY, P., PAGNI, M., SIGRIST, C. J. A., LACHAIZE, C., VEUTHEY, A-L., GASTEIGER, E. & BAIROCH, A. (2003). **Automated annotation of microbial proteomes in SWISS-PROT.** *Computational Biology and Chemistry*. 27, 49-58.

GIOVANNONI S.J., TRIPP H.J., GIVAN S., PODAR M., VERGIN K.L., BAPTISTA D., BIBBS L., EADS J., RICHARDSON T.H., NOORDEWIER M., RAPPE M.S., SHORT J.M., CARRINGTON J.C. & MATHUR E.J. (2005). **Genome streamlining in a cosmopolitan oceanic bacterium**. *Science*. 309(5738):1242-5.

GOODNER, B., HINKLE, G., GATTUNG, S., MILLER, N., BLANCHARD, M., QUROLLO, B., GOLDMAN, B.S., CAO, Y., ASKENAZI, M., HALLING, C., MULLIN, L., HOUMIEL, K., GORDON, J., VAUDIN, M., IARTCHOUK, O., EPP, A., LIU, F., WOLLAM, C., ALLINGER, M., DOUGHTY, D., SCOTT, C., LAPPAS, C., MARKELZ, B., FLANAGAN, C., CROWELL, C., GURSON, J., LOMO, C., SEAR, C., STRUB, G., CIELO, C., SLATER, S. (2001). **Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58**. Science 294(5550), 2323-8.

HOPCOFT, J. E, ULLMAN, J. D., 1979, Introduction to Automata Theory, Languages, and Computation. 1st. London, Addison-Wesley.

HUECK, C. J. (1998). **Type III Protein Secretion Systems in Bacterial Pathogens of Animals and Plants**. *Microbiol Mol Biol Rev*. 62, 379-433.

HULO, N., SIGRIST, C. J. A., LE SAUX, V., LANGENDIJK-GENEVAUX, P. S., BORDOLI, L., GATTIKER, A., DE CASTRO, E., BUCHER, P. & BAIROCH, A. (2004). **Recent improvements to the PROSITE database.** *Nucleic Acids Res*. 32, 134-137.

IKEDA, H., ISHIKAWA, J., HANAMOTO, A., SHINOSE, M., KIKUCHI, H., SHIBA, T., SAKAKI, Y., HATTORI, M., OMURA, S. (2003). **Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis.*** *Nat Biotechnol.* 21(5), 505-6.

JOARDAR, V., LINDEBERG, M., JACKSON, R.W., SELENGUT, J., DODSON, R., BRINKAC, L.M., DAUGHERTY, S.C., DEBOY, R., DURKIN, A.S., GIGLIO, M.G., MADUPU, R., NELSON, W.C., ROSOVITZ, M.J.,

SULLIVAN, S., CRABTREE, J., CREASY, T., DAVIDSEN, T., HAFT, D.H., ZAFAR, N., ZHOU, L., HALPIN, R., HOLLEY, T., KHOURI, H., FELDBLYUM, T., WHITE, O., FRASER, C.M., CHATTERJEE, A.K., CARTINHOUR, S., SCHNEIDER, D.J., MANSFIELD, J., COLLMER, A., BUELL, C.R. (2005). **Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition.** J Bacteriol. 187(18), 6488-98.

KANEHISA, M. & GOTO, S. (2000). **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 28, 27-30

KONDRASOV, A. S. (1999). **Comparative genomics and evolutionary biology.** *Curr Opin Genet*. 9(6), 624-9.

KULIKOVA, T., AKHTAR, R., ALDEBERT, P., ALTHORPE, N., ANDERSSON, M., BALDWIN, A., BATES, K., BHATTACHARYYA, S., BOWER, L., BROWNE, P., CASTRO, M., COCHRANE, G., DUGGAN, K., EBERHARDT, R., FARUQUE, N., HOAD, G., KANZ, C., LEE, C., LEINONEN, R., LIN, Q., LOMBARD, V., LOPEZ, R., LORENC, D., MCWILLIAM, H., MUKHERJEE, G., NARDONE, F., PASTOR, M. P. C., PLAISTER, S., SOBHANY, S., STOEHR, P., VAUGHAN, R., WU, D., ZHU, W. & APWEILER R. (2007). **EMBL Nucleotide Sequence Database in 2006.** *Nucl. Acids Res*. 35, 16-20.

LEE, B.M., PARK,Y.J., PARK, D.S., KANG, H.W., KIM, J.G., SONG, E.S., PARK, I.C., YOON, U.H., HAHN, J.H., KOO, B.S., LEE, G.B., KIM, H., PARK, H.S., YOON, K.O., KIM, J.H., JUNG, C.H., KOH, N.H., SEO, J.S., GO, S.J. (2005). **The genome sequence of *Xanthomonas oryzae* pathovar *oryzae* KACC10331, the bacterial blight pathogen of rice**. Nucleic Acids Res. 33(2), 577-86.

LETUNIC, I., COPLEY, R. R., PILS, B., PINKERT, S., SCHULTZ, J. & BORK, P. (2006**). SMART 5: domains in the context of genomes and networks**. *Nucleic Acids Research*. 34, 257-260.

MAGLOTT, D., OSTELL, J., PRUITT, K. D. & TATUSOVA, T. (2007). **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Research* 35, 26-D31.

MALTSEV, M., GLASS, E., SULAKHE, D., RODRIGUEZ, A., SYED, M. H., BOMPADA, T., ZHANG, Y. & D'SOUZA, M. (2006**). PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways.** *Nucleic Acids Res*. 34, 369-372.

MARKOWITZ, V. M., KORZENIEWSKI, F., PALANIAPPAN, K., SZETO, E., WERNER, G., PADKI, A., ZHAO, X., DUBCHAK, I., HUGENHOLTZ, P., ANDERSON, I., LYKIDIS, A., MAVROMATIS, K., IVANOVA, N. &

KYRPIDES, N. C. (2006). **The integrated microbial genomes (IMG) system.** *Nucleic Acids Res.* 34, 344-348.

MI, H., GUO, N., KEJARIWAL, A. & THOMAS , P. D. (2007). **PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways.** *Nucl. Acids Res.* 35, 247-252.

MOREIRA, L. M., DE SOUZA, R. F., ALMEIDA, N. F., JR., SETUBAL, J. C., OLIVEIRA, J. C., FURLAN, L. R., FERRO, J. A. & DA SILVA, A. C. (2004). **Comparative genomics analyses of citrus-associated bactéria.** *Annu Ver Phytapathol* 42, 163-84.

MOREIRA L. M., DE SOUZA, R. F., DIGIAMPIETRI, L. A., DA SILVA, A. C. & SETUBAL, J. C., (2005). **Comparative analyses of *Xanthomonas* and *Xylella* complete genomes.** *Omics* 9, 43-76.

MONTEIRO-VITORELLO, C.B., CAMARGO, L.E., VAN SLUYS, M.A., KITAJIMA, J.P., TRUFFI, D., DO AMARAL, A.M., HARAKAVA, R., DE OLIVEIRA, J.C., WOOD, D., DE OLIVEIRA, M.C., MIYAKI, C., TAKITA, M.A., DA SILVA, A.C., FURLAN, L.R., CARRARO, D.M., CAMAROTTE, G., ALMEIDA, N.F. JR, CARRER, H., COUTINHO, L.L., EL-DORRY, H.A., FERRO, M.I., GAGLIARDI, P.R., GIGLIOTI, E., GOLDMAN, M.H., GOLDMAN, G.H., KIMURA, E.T., FERRO, E.S., KURAMAE, E.E., LEMOS, E.G., LEMOS, M.V., MAURO, S.M., MACHADO, M.A., MARINO, C.L., MENCK, C.F., NUNES, L.R., OLIVEIRA, R.C., PEREIRA, G.G., SIQUEIRA, W., DE SOUZA, A.A., TSAI, S.M., ZANCA, A.S., SIMPSON, A.J., BRUMBLEY, S.M., SETUBAL, J.C. (2004). **The genome sequence of the gram-positive sugarcane pathogen *Leifsonia xyli* subsp. *xyli*.** Mol. Plant Microbe Interact. 17(8), 827-36.

KANEKO, T., NAKAMURA, Y., SATO, S., ASAMIZU, E., KATO, T., SASAMOTO, S., WATANABE, A., IDESAWA, K., ISHIKAWA, A., KAWASHIMA, K., KIMURA, T., KISHIDA, Y., KIYOKAWA, C., KOHARA, M., MATSUMOTO, M., MATSUNO, A., MOCHIZUKI, Y., NAKAYAMA, S., NAKAZAKI, N., SHIMPO, S., SUGIMOTO, M., TAKEUCHI, C., YAMADA, M., TABATA, S. (2000). **Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*.** DNA Res. 7(6), 331-8.

MULDER, N. J., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BORK, P., BUILLARD, V., CERUTTI, L., COPLEY, R., COURCELLE, E., DAS, U., DAUGHERTY L., DIBLEY, M., FINN, R., FLEISCHMANN W., GOUGH, J., HAFT, D., HULO, N., HUNTER, S., KAHN, D., KANAPIN, A., KEJARIWAL, A., LABARGA, A., LANGENDIJK-GENEVAUX, P. S., LONSDALE, D., LOPEZ, R., LETUNIC, I., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., NIKOLSKAYA, A. N., ORCHARD, S., ORENGO, C., PETRYSZAK, R., SELENGUT, J. D., SIGRIST, C. J. A., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C, H. & YEATS, C.

(2007). **New developments in the InterPro database.** *Nucleics Acids Research* 35, 224-228.

NELSON, K. E., WEINEL, C., PAULSEN, I. T., DODSON, R. J., HILBERT, H., MARTINS DOS SANTOS, V. A., FOUTS, D. E., GILL, S. R., POP, M., HOLMES, M., BRINKAC, L., BEANAN, M., DEBOY, R. T., DAUGHERTY, S., KOLONAY, J., MADUPU, R., NELSON, W., WHITE, O., PETERSON, J., KHOURI, H., HANCE, I., CHRIS, L. P., HOLTZAPPLE, E., SCANLAN, D., TRAN, K., MOAZZEZ, A., UTTERBACK, T., RIZZO, M., LEE, K., KOSACK, D., MOESTL, D., WEDLER, H., LAUBER, J., STJEPANDIC, D., HOHEISEL, J., STRAETZ, M., HEIM, S., KIEWITZ, C., EISEN, J. A., TIMMIS, K. N., DUSTERHOFT, A., TUMMLER, B., FRASER, C. M. (2002). **Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440.** *Environ Microbiol.* 4(12), 799-808.

NETO, P. A. S. P., AZEVEDO, J. L., CAETANO, L. C. (2004). **Microrganismos endofíticos em plantas: status atual e perspectivas**. *Boletin Latinoamericano y del Caribe de Plantas Medicinales y Aromatica*, v. 3, pag. 69-72.

NIERMAN, W. C., FELDBLYUM, T. V., LAUB, M. T., PAULSEN, I. T., NELSON, K. E., EISEN, J. A., HEIDELBERG, J. F., ALLEY, M. R., OHTA, N., MADDOCK, J. R., POTOCKA, I., NELSON, W. C., NEWTON, A., STEPHENS, C., PHADKE, N. D., ELY, B., DEBOY, R. T., DODSON, R. J., DURKIN, A. S., GWINN, M. L., HAFT, D. H., KOLONAY, J. F., SMIT, J., CRAVEN, M. B., KHOURI, H., SHETTY, J., BERRY, K., UTTERBACK, T., TRAN, K., WOLF, A., VAMATHEVAN, J., ERMOLAEVA, M., WHITE, O., SALZBERG, S. L., VENTER, J. C., SHAPIRO, L., FRASER, C. M. (2001**). Complete genome sequence of *Caulobacter crescentus*.** *Proc Natl Acad Sci U S A.* 98(7), 4136-41.

NOËL, L; THIEME, F; NENNSTIEL, D. & BONAS, U. (2002). **Two novel type III system-secreted proteins of *Xanthomonas campestris* pv. *versicatoria* are encoded within the hrp pathogenicity island.** *Journal of Bacteriology.* 184, 1340-1348.

NOTREDAME, C., HIGGINS, D., HERINGA, J. (2000). **T-Coffee: A novel method for multiple sequence alignments**. *Journal of Molecular Biology.* 302, 205-217.

OVERBEEK, R., FONSTEIN, M., D'SOUZA, M., PUSCH, G. D., MALTSEV, N. (1999). **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci U S A.* 96(6):2896-901

OCHIAI,H., INOUE,Y., TAKEYA,M., SASAKI,A. AND KAKU,H. (2005). **Genome sequence of *Xanthomonas oryzae* pv. *oryzae* suggests contribution of large numbers of efector genes and insertion sequences to its race diversity**. JARQ 39, 275-287.

PATEL, J.M., HUDDLER, D.P., HAMMEL, L.. Declarative and Efficient Querying on Protein Secondary Structures. In: Data Mining in Bioinformatics, Advanced Information and Knowledge Processing, pp. 243-273.

PAULSEN, I. T., PRESS, C. M., RAVEL, J., KOBAYASHI, D. Y., MYERS, G. S., MAVRODI, D. V., DEBOY, R. T., SESHADRI, R., REN, Q., MADUPU, R., DODSON, R. J., DURKIN, A. S., BRINKAC, L. M., DAUGHERTY, S. C., SULLIVAN, S. A., ROSOVITZ, M. J., GWINN, M. L., ZHOU, L., SCHNEIDER, D. J., CARTINHOUR, S. W., NELSON, W. C., WEIDMAN, J., WATKINS, K., TRAN, K., KHOURI, H., PIERSON, E. A., PIERSON, L. S. 3RD., THOMASHOW, L. S., LOPER, J. E. (2005). **Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5.** *Nat Biotechnol.* 23(7), 823-4.

PERRIÈRE, G., DURET, L. & GOUY, M. (2006). **HOBACGEN: Database System for Comparative Genomics in Bacteria.** *Genome Res* 10, 379-385.

PETERSON, J. D., UMAYAM, L. A., DICKINSON, T., HICKEY, E. K. & WHITE. O. (2001). **The Comprehensive Microbial Resource.** *Nucleic Acids Res.* 29, 123-125.

PHAN, I.Q., PILBOUT, S.F., FLEISCHMANN, W., BAIROCH, A. (2003). **NEWT, a new taxonomy portal**. Nucleic Acids Res. 31(13), 3822-3.

PRUST, C., HOFFMEISTER, M., LIESEGANG, H., WIEZER, A., FRICKE, W. F., EHRENREICH, A., GOTTSCHALK, G., DEPPENMEIER, U. (2005). **Complete genome sequence of the acetic acid bacterium *Gluconobacter oxydans*.** *Nat Biotechnol.* 23(2), 186-7.

QIAN, W., JIA, Y., REN, S. X., HE, Y. Q., FENG, J. X., LU, L. F., SUN, Q., YING, G., TANG, D. J., TANG, H., WU, W., HAO, P., WANG, L., JIANG, B. L., ZENG, S., GU, W. Y., LU, G., RONG, L., TIAN, Y., YAO, Z., FU, G., CHEN, B., FANG, R., QIANG, B., CHEN, Z., ZHAO, G. P., TANG, J. L. & HE, C. (2005). **Comparative and funciotional genomic analyses of the pathogenicity of phytopathogen *Xanthomonas campestris* pv. *campestris*.** *Genome Rev* 15, 757-67

SALANOUBAT, M., GENIN, S., ARTIGUENAVE, F., GOUZY, J., MANGENOT, S., ARLAT, M., BILLAULT, A., BROTTIER, P., CAMUS, J.C., CATTOLICO, L., CHANDLER, M., CHOISNE, N., CLAUDEL-RENARD, C., CUNNAC, S., DEMANGE, N., GASPIN, C., LAVIE, M., MOISAN, A., ROBERT, C., SAURIN, W., SCHIEX, T., SIGUIER, P., THEBAULT, P., WHALEN, M., WINCKER, P., LEVY, M., WEISSENBACH, J., BOUCHER, C.A. (2002). **Genome sequence of the plant pathogen *Ralstonia solanacearum***. Nature 415(6871), 497-502.

SCHULTZ, J., MILPETZ, F., BORK, P. & PONTING, C. P. (2000). **SMART, a simple modular architecture research tool: Identification of signaling domains.** *PNAS,* 95, 5857-5864.

SELENGUT, J. D., HAFT, D. H., DAVIDSEN, T., GANAPATHY, A., GWINN-GIGLIO, M., NELSON, W. C., RICHTER, A. R. & WHITE, O. (2007). **TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes.** *Nucl. Acids Res.* 35, 260-264.

SETUBAL, J. C., MOREIRA, L. M. & DA SILVA, A. C. (2005). **Bacterial phytopathogens and genome science.** *Curr Opin Microbiol* 8, 595-600.

SIMPSON, A. J., REINACH, F. C., ARRUDA, P., ABREU, F. A., ACENCIO, M., ALVARENGA, R., ALVES, L. M., ARAYA, J. E., BAIA, G. S., BAPTISTA, C. S., BARROS, M. H., BONACCORSI, E. D., BORDIN, S., BOVE, J. M., BRIONES, M. R., BUENO, M. R., CAMARGO, A. A., CAMARGO, L. E., CARRARO, M. D., CARRER, H., COLAUTO, N. B., COLOMBO, C., COSTA, M. C., COSTA-NETO, C. M., COUTINHO, L. L., CRISTOFANI, M., DIAS-NETO, E., DOCENA, C., EL-DORRY, H., FACINCANI, A. P., FERREIRA, A. J., FERREIRA, V. C., FERRO, J. A., FRAGA, J. S., FRANCA, J. S., FRANCA, S. C., FRANCO, M. C., FROHME, M., FURKAN, L. R., GARNIER, M., GOLDMAN, G. H., GOLDMAN, M. H., GOMES, S. L., GRUBER, A., HO, P. L., HOHEISEL, J. D., JUNQUEIRA, M. L., KEMPER, E. L., KITAJINA, J. P., KRIGER, J. E., KURAMAE, E. E., LAIGRET, F., LAMBAIS, M. R., LEITE, L. C., LEMOS, E. G., LEMOS, M. V., LOPES, S. A., LOPES, C. R., MACHADO, J. A., MACHADO, M. A., MADEIRA, A. M., MADEIRA, H. M., MARINO, C. L., MARQUES, M. V., MARTINS, E. A., MARTINS, E. M., MATSUKUMA, A. Y., MENCK, C. F., MIRACCA, E. C., MIYACA, C. Y., MONTERIRO-VITORELLO, C. B., MOON, D. H., NAGAI, M. A., NASCIMENTO, A. L., NETTO, L. E., NHANI, A., JR., NOBREGA, F. G., NUNES, L. R., OLIVEIRA, M. A., DE OLIVEIRA, M. C., DE OLIVEIRA, R. C., PALMIERI, D. A., PARIS, A., PEIXOTO, B. R., PEREIRA, G. A., PEREIRA, H. A., JR., PESQUERO, J. B., QUAGGIO, R. B., ROBERTO, P. G., RODRIGUES, V., DE, M. R. A. J., DE ROSA, V. E., JR., DE AS, R. G., SANTELLI, R. V., SAWASAKI, H. E., AS SILVA, A. C., DA SILVA, A. M., DA SILVA, F. R., DA SILVA, W. A., JR., *et al* (2000). **The genome sequence of the plant pathogen *Xylella fatidiosa*. The *Xylella fatidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis.** *Nature* 406, 151-7.

STOTHARD, P., VAN DOMSELAAR, G., SHRIVASTAVA, S., GUO, A., O'NEILL, B., CRUZ, J., ELLISON, M. & WISHART, D. S. (2005). **BacMap: an interactive picture atlas of annotated bacterial genomes.** *Nucleic Acids Res.* 33, 317-320.

STOVER, C. K., PHAM, X. Q., ERWIN, A. L., MIZOGUCHI, S. D., WARRENER, P., HICKEY, M. J., BRINKMAN, F. S., HUFNAGLE, W. O., KOWALIK, D. J., LAGROU, M., GARBER, R. L., GOLTRY, L., TOLENTINO, E., WESTBROCK-WADMAN, S., YUAN, Y., BRODY, L. L., COULTER, S. N., FOLGER, K. R., KAS, A., LARBIG, K., LIM, R., SMITH, K., SPENCER, D., WONG, G. K., WU, Z., PAULSEN, I. T., REIZER, J., SAIER, M. H., HANCOCK, R. E., LORY, S., OLSON, M. V. (2000). **Complete genome**

sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature.* 406(6799), 959-64.

STOESSER G., MOSELEY M.A., SLEEP J., MCGOWRAN M., GARCIA-PASTOR M., AND STERK P. (1998). **The EMBL nucleotide sequence database**. Nucleic Acids Res. 26: 8 - 15.

STUDHOLME, D. J., DOWNIE, J. A., PRESTON, G. M. (2004). **Protein domains and architectural innovation in plant-associated Proteobacteria.** *BMC Genomics.* 6(1):17.

SUGAWARA H., MIYAZAKI S., GOJOBORI T., AND TATENO Y. (1999). **DNA Data Bank of Japan dealing with large-scale data submission.** Nucleic Acids Res. 27: 25- 28.

SUGAWARA, H., ABE, T., GOJOBORI, T. & TATENO Y. (2007). **DDBJ working on evaluation and classification of bacterial genes in INSDC.** *Nucl. Acids Res*. 35, 13-15.

TATA, S., PATEL, J.M., FRIEDMAN, J.S, SWAROOP, A. (2006). Declarative Querying for Biological Sequence Databases, Proceeding of the International Conference on Data Engineering, Georgia, USA, 3-8 April**.**

TATUSOV, R. L., FEDOROVA, N. D., JACKSON, J. D., JACOBS, A. R., KIRYUTIN, B., KOONIN, E. V., KRYLOV, D. M., MAZUMDER, R., MEKHEDOV, S. L., NIKOLSKAYA, A. N., RAO, B. S., SMIRNOV, S., SVERDLOV, A. V., VASUDEVAN, S., WOLF, Y. I., YIN, J. J., NATALE, D. A. (2003). **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics.* 11, 4:41.

TATUSOV, R.L., GALPERIN, M.Y., NATALE, D.A., KOONIN, E.V. (2000). **The** COG **database: a tool for genome-scale analysis of protein functions and evolution**. Nucleic Acids Research 28(1), 33-36.

THE UNIPROT CONSORTIUM. (2006). **The Universal Protein Resource (UniProt)**. Nucleic Acids Research 35, D193-D197.

THIEME, F., KOEBNIK, R., BEKEL, T., BERGER, C., BOCH, J., BUTTNER, D., CALDANA, C., GAIGALAT, L., GOESMANN, A., KAY, S., KIRCHNER, O., LANZ, C., LINKE, B., MCHARDY, A. C., MEYER, F., MITTENHUBER, G., NIES, D. H., NIESBACH-KLOSGEN, U., PATSCHKOWSKI, T., RUCKERT, C., RUPP, O., SCHNEIKER, S., SCHUSTER, S. C., VORHOLTER, F. J., WEBER, E., PUHLER, A., BONAS, U., BARTELS, D. & KAISER, O. (2005). **Insights into genome plasticity and the plant pathogenic bacterium *Xanthomonas campestris* pv. *vesicatoria* revealed by the complete genome sequence.** *J Bacteriol* 187, 7254-66.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994). **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 22, 4673-80.

UCHIYAMA, I. (2003). **MBGD: microbial genome database for comparative analysis.** *Nucleic Acids Research* 31, 58-62.

UCHIYAMA, I. (2007). **MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups.** *Nucl. Acids Res*. 35, 343-346.

VAN SLUYS, M. A., DE OLIVEIRA, M. C., MONTEIRO-VITORELLO, C. B., MIYAKI, C. Y., FURLAN, L. R., CAMARGO, L. E., DA SILVA, A. C., MOON, D. H., TAKITA, M. A., LEMOS, E. G., MACHADO, M. A., FERRO, M. I., DA SILVA, F. R., GOLDMAN, M. H., GOLDMAN, G. H., LEMOS, M. V., EL-DORRY, H., TSAI, S. M., CARRER, H., CARRARO, D. M., DE OLIVEIRA, R. C., NUNES, L. R., SIQUEIRA, W. J., COUTINHO, L. L., KIMURA, E. T., FERRO, E. S., HARAKAVA, R., KURAMAE, E. E., MARINO, C. L., GIGLIOTI, E., ABREU, I. L., ALVES, L. M., DO AMARAL, A. M., BAIA, G. S., BLANCO, S. R., BRITO, M. S., CANNAVAN, F. S., CELESTINO, A. V., DA CUNHA, A. F., FENILLE, R. C., FERRO, J. A., FORMIGHIERI, E. F., KISHI, L. T., LEONI, S. G., OLIVEIRA, A. R., ROSAV. E., JR., SASSAKI, F. T., SENA, J. A., DE SOUZA, A. A., TRUFFI, D., TSUKUMO, F., YANAI, G. M., ZAROS, L. G., CIVEROLO, E. L., SIMPSON, A. J., ALMEIDA, N. F., JR., SETUBAL, J. C. & KITAJIMA, J. P. (2003). **Comparative analyses of the complete genome sequence of Pierce's disease and itrus variegated chlorosis strains of *Xylella fatidiosa*.** *J Bacteriol* 185, 1018-26.

VAN SLUYS, M. A., MONTEIRO-VITORELLO, C. B., CAMARGO, L. E., MENCK, C. F., DA SILVA, A. C., FERRO, J. A., OLIVEIRA, M. C., SETUBAL, J. C., KITAJIMA, J. P. & SIMPSON, A. J. (2002). **Comparative genomic analysis of plant-assosciated bactéria.** *Annu Ver Phytopathol* 40, 169-89.

VODOVAR, N., VALLENET, D., CRUVEILLER, S., ROUY, Z., BARBE, V., ACOSTA, C., CATTOLICO, L., JUBIN, C., LAJUS, A., SEGURENS, B., VACHERIE, B., WINCKER, P., WEISSENBACH, J., LEMAITRE, B., MEDIGUE, C., BOCCARD, F. (2006). **Complete genome sequence of the entomopathogenic and metabolically versatile soil bacterium *Pseudomonas entomophila*.** *Nat Biotechnol.* 24(6), 660-1.

WENGELNIK, K. & BONAS, U. (1996). **HrpXv, na AraC-type regulator expression of fiver the six loci in *hrp* cluster of *Xanthomonas campestris* pv. *vesicatoria*.** *Journal of Bacteriology*. 178, 3462-3469.

WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., GEER, L. Y., KAPUSTIN, Y., KHOVAYKO, O., LANDSMAN, D., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R.,

OSTELL, J., MILLER, V., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., TATUSOV, R. L., TATUSOVA, T. A., WAGNER, L. & YASCHENKO, E. (2006). **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research*, 00, 1,8.

WILSON, D., MADERA, M., VOGEL, C., CHOTHIA, C. & GOUGH, J. (2007). **The SUPERFAMILY database in 2007: families and functions.** *Nucl. Acids Res*. 35, 308-313.

WINSOR, G. L., LO, R., SUI, S. J. H., UNG, K. S. E., HUANG, S., CHENG, D., CHING, W. H., HANCOCK, R. E. W. & BRINKMAN, F. S. L. (2005). ***Pseudomonas aeruginosa*** **Genome Database and PseudoCAP: facilitating community-based, continually updated, genome annotation.** *Nucleic Acids Res.* 33, 338-343.

WOOD, D.W., SETUBAL, J.C., KAUL, R., MONKS, D.E., KITAJIMA, J.P., OKURA, V.K., ZHOU, Y., CHEN, L., WOOD, G.E., ALMEIDA, N.F. JR, WOO, L., CHEN, Y., PAULSEN, I.T., EISEN, J.A., KARP, P.D., BOVEE, D. SR., CHAPMAN, P., CLENDENNING, J., DEATHERAGE, G., GILLET, W., GRANT C., KUTYAVIN, T., LEVY, R., LI, M.J., MCCLELLAND E., PALMIERI, A., RAYMOND, C., ROUSE, G., SAENPHIMMACHAK, C., WU, Z., ROMERO, P., GORDON, D., ZHANG, S., YOO, H., TAO, Y., BIDDLE, P., JUNG, M., KRESPAN, W., PERRY, M., GORDON-KAMM, B., LIAO, L., KIM, S., HENDRICK, C., ZHAO, Z.Y., DOLAN, M., CHUMLEY, F., TINGEY, S.V., TOMB. J.F., GORDON, M.P., OLSON, M.V., NESTER, E.W. (2001). **The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58**. Science 294(5550), 2317-23.

WU, C.H., HUANG, H., YEH, L.S., BARKER, W.C. (2003). **Protein family classification and functional annotation**. Computational Biology and Chemistry 27, 37-47.

WU, C. H., NIKOLSKAYA, A., HUANG, H., YEH, L. L., NATALE, D. A., VINAYAKA, C. R., HU, Z., MAZUMDER, R., KUMAR, S., KOURTESIS, P., LEDLEY, R. S., SUZEK, B. E., ARMINSKI, L., CHEN, Y., ZHANG, J., CARDENAS, J. L., CHUNG, S., CASTRO-ALVEAR, J., DINKOV, G. & BARKER, W. C. (2004). **PIRSF: family classification system at the Protein Information Resource.** *Nucleic Acids Res*. 32, 112-114.

ANEXO 1

DISSERTAÇÃO DE MESTRADO

FERNANDA NASCIMENTO ALMEIDA

SÚMULA CURRICULAR

SÚMULA CURRICULAR

Fernanda Nascimento Almeida

Data de nascimento: 18/06/1978, Rio de Janeiro/RJ, Brasil.

FORMAÇÃO ACADÊMICA/TITULAÇÃO

1998 – 2002

Graduação em Ciências Biológicas com Ênfase em Biomedicina.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil
Título: Exemplo de Junção Experimental e Identificação de Transcritos (ORESTES) Genes Humanos.
Orientadoras: Dra. Mônica Rosa Bertão e Dra. Marina Passeto Nóbrega.

2005 – 26/03/2007.

Mestrado em Modelagem Computacional.com ênfase em Bioinformática e Biologia Computacional.
Laboratório Nacional de Computação Científica, LNCC, Petrópolis, Brasil
Título: Implementação de um Banco de Dados de Proteomas de Bactérias Associadas a Plantas: ProBacter.
Orientadoras: Dra. Claudia Barros Monteiro-Vitorello e Dra. Ana Tereza R. Vasconcelos.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

2007 – em andamento

Doutorado em Bioinformatica.
Universidade de São Paulo, USP, São Paulo, Brasil
Título: (Em discussão)
Orientador: Dra. Aline Maria da Silva

FORMAÇÃO COMPLEMENTAR

1999 - 1999

Curso de curta duração em Técnicas Em Genética Molecular e Suas Aplicações.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil

2001 - 2001

Curso de curta duração em Fundamentos e Procedimentos da Pesquisa Científica.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil

2002 - 2002

Curso de curta duração em A Bioinformática na Análise da Seqüência de DNA.
Sociedade Brasileira de Genética, SBG, Brasil

2003 - 2003

Extensão universitária em Biotecnologia.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil

Extensão universitária em Profissional LINUX Programando.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil

Extensão universitária em Genética da conservação de espécies abóreias.
Universidade Estadual de Santa Cruz, UESC, Ilhéus, Brasil

2004 - 2004

Curso de curta duração em Bioinformática Aplicada a Proteômica.
Laboratório Nacional de Computação Científica, LNCC, Petrópolis, Brasil
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

Curso de curta duração em Bioinformática Aplicada a Genômica.
Laboratório Nacional de Computação Científica, LNCC, Petrópolis, Brasil
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

2006 - 2006

Curso de curta duração em Introdução ao Linux.
Laboratório Nacional de Computação Científica, LNCC, Petrópolis, Brasil

PRODUÇÃO BIBLIOGRÁFICA

ALMEIDA, F. N., SOUZA, R. C., PASCHOAL, A. R., VASCONCELOS, A. T. R., MONTEIRO-VITORELLO, C. B. **Database PROBACTER: proteomes of plant-associated bacteria**. In: ISMB 2006 and 2and Annual Meeting of the International Society for Computacional Biology, 2006, Fortaleza. ISMB 2006 and 2and Annual Meeting of the International Society for Computacional Biology. ISCB International Society for Computacional Biology, 2006.

ALMEIDA, F. N., SOUZA, R. C., VASCONCELOS, A. T. R., MONTEIRO-VITORELLO, C. B. **Building PPROBACTER: a database of plant-associated bacteria complete proteomes**. In: X-Meeting 1st Internationl Conference of the AB³C, 2005, Caxambu. X-Meeting 1st Internationl Conference of the AB³C. AB³C, 2005.

VIDAL, Ramon Oliveira, ALMEIDA, F. N., SUARÉZ, Diego Gervásio Frías, CARELS, Nicolas, CASCARDO, Júlio Cézar de Mattos. **Promoter Finding Produce Applied to the *Crinipellis perniciosa* Fungus Incomplete**

Genome. In: International Conference on Bioinformatics and Computacional Bioilogy, 2003, Ribeirão Preto. International Conference on Bioinformaticsand Computational Biology. , 2003.

MOREIRA, J. C., FERREIRA, L. E., BOGOSSIAN, A. P., ALMEIDA, F. N., NÓBREGA, M. P. **Exemplo de Junção Experimental e Identificação de Transcritos (ORESTES) Genes Humanos**. In: VI Encontro de Iniciação Científica e II Encontro de Pós Graduação, UNIVAP, 2002, São José dos Campos. VI Encontro de Iniciação Científica e II Encontro de Pós Graduação. São José dos Campos: UNIVAP, 2002. p.58 – 62

MOREIRA, J. C., FERREIRA, L. E., BOGOSSIAN, A. P., ALMEIDA, F. N., NÓBREGA, M. P. **Validação de Novos Transcritos Humanos do Grupo VPO**. In: 48° Congresso Nacional de genética, 2002, Águas de Lindóia. 48° Congresso Nacional de genética. , 2002. p.93

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo