Ferramentas para comparação genômica

Nalvo F. Almeida Jr.[1*], **João Carlos Setubal**[2]

[1]Dep. de Computação e Estatística – Universidade Federal de Mato Grosso do Sul
Caixa Postal 549 – 79070-900 - Campo Grande, MS

[2]Instituto de Computação – Universidade Estadual de Campinas
Caixa Postal 6176 – 13083-970 - Campinas, SP

nalvo@dct.ufms.br, setubal@ic.unicamp.br

Abstract. *With increasing availability of published genome sequences, we need to analyse them in order to understand functional and evolutionary issues of the organisms. A genome project, in particular for prokaryotes, consists of three main phases: sequencing, annotation and analysis. The last phase consists of getting an overview of the genome from the annotation and other analysis, like comparison to other genomes, for example. This work is about genome comparison. We propose methodologies and implementations for detailed comparison of two genomes, at the DNA and their genes levels. The main goal is to provide a set of tools for functional characterization of organisms, serving also as an auxiliar tool for annotation. The methodologies and implementations have been used successfully in several genome projects.*

Resumo. *Com o crescente número de genomas seqüenciados, precisamos analisar as seqüências geradas, com o objetivo de entender melhor caracterizações funcionais e aspectos dos organismos estudados. Um projeto genoma, em especial de um procarioto, consiste de três grandes fases: o seqüenciamento, anotação e análise. A última etapa consiste na tentativa de se obter uma visão global do genoma a partir da anotação e a partir de outras análises, como por exemplo a comparação com outros genomas. E é nesse contexto, comparação de genomas, que este trabalho se insere. Propomos metodologias para comparação detalhada de dois genomas, tanto no nível de DNA, quanto no de seus genes, assim como a implementação dessas metodologias. O objetivo é fornecer um conjunto de ferramentas para caracterização funcional do organismo estudado, servindo também como ferramental auxiliar na anotação. As metodologias propostas e respectivas ferramentas já foram utilizadas com sucesso em diversos projetos genoma.*

1. Introdução

Uma forma de auxiliar na descoberta de informações biológicas relevantes, a partir dos dados gerados nos projetos genoma, passa pela determinação dos papéis que os mais diversos objetos envolvidos num genoma desempenham. Esses papéis estão muitas vezes

*Este trabalho recebeu apoio de: CAPES (programa PICDT), CNPq/Pronex 664107/1997-4, Fulbright Comission e Fundect-MS.

relacionados às características estruturais de cada objeto. As funções de uma proteína, por exemplo, são determinadas diretamente pela sua forma e estruturação. Assim, é de se esperar que a comparação entre objetos, nas suas formas mais primárias, nos tragam pistas de relacionamentos entre eles e, por conseqüência, entre suas funcionalidades. No caso de genomas é de se esperar que a comparação entre seqüências de DNA ou de genes seja útil na determinação de funcionalidades comuns. A comparação de genomas tem como principais objetivos: detecção de similaridades e diferenças entre genomas completos, no nível de DNA; identificação de genes ou grupos de genes envolvidos em diversas funções; identificação de genes ou grupos de genes responsáveis por características fenotípicas peculiares a um genoma particular; identificação de genes homólogos (genes descendentes de um mesmo gene anscestral); anotação de genes de genomas não completos; e inferência de relações filogenéticas entre os organismos.

O resultado principal do nosso trabalho é um conjunto de metodologias para a comparação de genomas, tanto no nível de DNA, quanto no nível de seus genes. Como resultado das metodologias, implementamos as seguintes ferramentas:

- BACON - Bacterial Comparator – compara dois genomas no nível de DNA;
- SBACON - Self BACON – encontra repetições aproximadas num genoma; e
- EGG - Extended Genome-Genome comparison – compara genes de dois genomas.

As metodologias e os programas propostos na tese foram utilizados em alguns projetos genoma, como os das bactérias *Xylella fastidiosa* (Pierce's disease) [11], *Xanthomonas axonopodis* pv. *citri* e *Xanthomonas campestris* pv. *campestris* [5], *Agrobacterium tumefaciens* [13] e *Paracoccidioides brasiliensis* [7]. Nossas participações nesses projetos aconteceram paralelamente ao desenvolvimento das metodologias, o que nos possibilitou aprimorá-las de acordo com as demandas exigidas nos projetos.

O trabalho resultou na publicação [9], além da participação em outras seis [4, 5, 7, 11, 12, 13]. Propusemos ainda um algoritmo baseado em programação dinâmica para comparação de proteínas, resultando na publicação [2]. O texto original da tese encontra-se disponível em `http://www.dct.ufms.br/~nalvo/publications/`. A utilização das ferramentas desenvolvidas nos projetos, bem como as publicações com alto índice de impacto obtidas, evidenciam a multidisciplinaridade e a contribuição do trabalho.

Este texto está organizado como segue. Na seção 2., propomos uma metodologia baseada em repetições aproximadas e árvores de sufixos para comparar dois genomas no nível de DNA, além dos programas BACON e SBACON. Na seção 3., descrevemos nossa metodologia para a comparação de proteomas, assim como o programa resultante, EGG. Na seção seguinte, propomos uma forma alternativa de uso da programação dinâmica na comparação de proteínas. Finalmente, na seção 5., fazemos alguns comentários finais.

2. Comparação de DNA

Um problema na comparação de dois genomas no nível de DNA é que as seqüências são longas, com ocorrências de subsqüências com similaridades aproximadas. A presença de tais similaridades aproximadas sugere o uso da programação dinâmica na comparação. No entanto, seus altos custos de espaço e tempo inviabilizam sua utilização para tal fim.

Uma alternativa para o tratamento de seqüências longas é o uso de *Árvore de Sufixos*, que nos permite determinar muito rapidamente as repetições exatas de uma seqüência [8]. A idéia é então, a partir das repetições encontradas na seqüência formada pela concatenação dos dois genomas, obtermos as regiões similares entre eles. Nossa contribuição nesse caso é uma metodologia que permite uma comparação aproximada dos genomas, a partir das repetições exatas fornecidas pela árvore de sufixos. Em particular, estamos interessados em obter as repetições maximais aproximadas, descritas em [8], que ocorrem em dois genomas. Repetições maximais nos interessam porque são uma forma de não gerarmos dados redundantes, já que elas podem envolver repetições menores.

Para obtermos as repetições maximais aproximadas, fazemos uso de estruturas intermediárias como pares maximais exatos, repetições maximais exatas, e pares maximais aproximados, também descritas em [8]. A metodologia aqui consiste basicamente num algoritmo eficiente para percorrer a árvore de sufixos de dois genomas, tal que os pares maximais exatos são facilmente identificados. A partir dos pares maximais encontrados, as repetições maximais exatas e os pares maximais aproximados são determinados. Pelo fato de podermos ver um par aproximado como sendo um par exato ou aproximado com a inclusão de algumas diferenças, o algoritmo determina as repetições aproximadas pela junção de pares e repetições exatas próximos uns dos outros, com regras bem definidas de proximidade, que levam em conta o tamanho das ocorrências da repetição, e o tamanho dos intervalos que separam as respectivas ocorrências.

O programa resultante dessa metodologia se chama BACON, de *Bacterial Comparator*. A entrada para o BACON são as seqüências de DNA dos dois genomas, no formato FASTA (formato padrão para seqüências biológicas). Como saída, BACON reporta: repetições maximais exatas e aproximadas que ocorrem nos dois genomas; repetições *tandem* exatas; diferenças singulares; e blocos de inserção e remoção. SBACON é uma outra ferramenta desenvolvida, que segue exatamente a mesma linha de BACON, só que dessa vez o objetivo é comparar um genoma contra ele mesmo.

No trabalho desenvolvido por Delcher e outros [6], o programa MUMMER compara dois genomas. MUMMER, assim como BACON, é baseado em árvore de sufixos; e usa o conceito de *maximal unique match* (MUM), uma seqüência que ocorre exatamente uma vez em cada genoma, e que pode ser encontrado olhando num nó interno da árvore de sufixos, tal que esse nó interno tem apenas dois fihos que são folhas e que representam posições em genomas diferentes. Além das diferenças metodológicas entre BACON e MUMMER, BACON tem a vantagem de reportar repetições múltiplas (que ocorrem mais que duas vezes), ao invés de reportar apenas os MUMs. Isso é possível porque BACON faz uso de vértices internos mais gerais na árvore. Além disso, BACON tem alguns subprodutos que MUMMER não tem, como por exemplo blocos únicos, repetições tandem e diferenças singulares.

3. Comparação de proteomas

Comparações de genomas no nível de DNA, tais como feitas por BACON, apenas mostram regiões que se repetem nos genomas, sem levarem em conta aspectos funcionais das regiões similares detectadas. Além disso, tais comparações são muito úteis apenas quando os genomas são evolutivamente próximos. Nossa contribuição aqui consiste numa meto-

dologia para a comparação de dois proteomas e a respectiva implementação (o programa EGG). Um **proteoma** é o conjunto de genes de um genoma.

Um **gene**, para efeitos deste texto, é uma seqüência traduzida para aminoácidos de uma ORF (*Open Read Frame*) predita como pertencendo a um gene. A **ordem dos genes** de um proteoma é dada pela ordem não-decrescente das coordenadas de início dos genes. Dois genes g_i e g_j de um mesmo genoma G são **parálogos** se são descendentes de um mesmo gene ancestral. Uma **região de genes consecutivos (RGC)** é um conjunto de genes consecutivos num proteoma, de acordo com suas coordenadas de início, independente da fita. Note que o próprio proteoma é uma RGC.

Considere a comparação entre os proteomas dos genomas G e H, e os genes g, g' de G e h, h' de H. Dois genes g e h são **ortólogos** se são descendentes de um mesmo gene ancestral. Um gene g é **específico** em relação a H se não existir h em H tal que g e h são ortólogos. Uma **região específica (RE)** de G em relação a H é uma região de G predominantemente formada por genes específicos. Uma **região ortóloga (RO)** é um par (α, β) tal que: α é uma RGC em G; β é uma RGC em H; α e β são descendentes de uma mesma região ancestral; e α e β contêm aproximadamente o mesmo número de genes. Dois pares de ortólogos (g, h) e (g', h') formam um **cruzamento** quando a ordem de g e g' em G e a ordem de h e h' em H são invertidas. A **espinha dorsal de duas RGCs** α de G e β de H é uma seqüência de pares de ortólogos do tipo (g, h), tal que: cada gene de α tem no máximo um gene ortólogo a ele em β, e vice-versa; e não existe cruzamentos entre os pares da seqüência.

Nossa metodologia consiste de passos e/ou critérios que devem ser seguidos para cada um dos objetivos específicos citados abaixo.

1. encontrar genes específicos de um proteoma em relação ao outro;
2. encontrar regiões específicas de um proteoma em relação ao outro;
3. encontrar pares de genes ortólogos;
4. encontrar regiões ortólogas;
5. determinar a espinha dorsal dos proteomas; e
6. determinar as famílias de genes parálogos de um proteoma.

Para encontrar genes específicos, nossa metodologia sugere o seguinte critério: o gene g de G é **específico** com relação ao genoma H se, e somente se, a medida de significância $s(g, h)$ da similaridade entre eles é maior que um limite fixo S', para todo gene h de H, tal que $S' > S$.

A metodologia que propomos para encontrar uma RE é baseada no problema computacional conhecido como *subcadeia de máxima soma*. Esse problema tem como entrada uma seqüência L de inteiros e como saída a subcadeia de L cuja soma dos elementos é máxima, entre todas as subcadeias de L. Para isso atribuímos um valor a cada um dos genes do proteoma, digamos ε para um genes não específicos e ε' para genes específicos, tais que $\varepsilon' > \varepsilon$. A seqüência de entrada para o problema da subcadeia de máxima soma é a seqüência dos valores atribuídos aos genes. Dessa forma, o algoritmo deve encontrar todas as subcadeias cuja soma dos valores é maior ou igual a um valor limite.

Nossa metodologia para a obtenção dos pares de genes ortólogos é tal que: (g, h) é um par de genes ortólogos se, e somente se: a medida de significância estatística da similaridade entre g e h seja menor ou igual a um limite fixo; e o alinhamento de g e h cubra pelo menos um certo percentual mínimo de cada uma das duas seqüências.

Um algoritmo que compara dois genes e que fornece a similaridade e a significância estatística entre eles é suficiente para a implementação dessa parte da metodologia. Nossa implementação fez uso do programa BLAST [3], usando o *e-value* (fornecido por BLAST) como medida de significância estatística. Assim, faz sentido usarmos um limite superior de significância estatística para similaridade (quanto menor o *e-value*, maior a significância).

Para as ROs, precisamos do conceito de *run*. Seja α uma RGC de G formada pelos genes $g_i, \ldots g_k$ e β uma RGC de H formada pelos genes $h_j, \ldots h_l$, tais que $k - i + 1 = l - j + 1$, $k > i$, e $l > j$. Dizemos que α e β formam um **run** se uma das seguintes seqüências de pares de genes ortólogos acontece: ou $(g_i, h_j), (g_{i+1}, h_{j+1}), \ldots, (g_k, h_l)$; ou $(g_i, h_l), (g_{i+1}, h_{l-1}), \ldots, (g_k, h_j)$.

A estratégia para a determinação dos *runs* é baseada na varredura dos pares de ortólogos contíguos. Isso pode ser feito com o uso de uma matriz binária, que indica a ortologia para cada par de genes dos proteomas. Nosso critério para a determinação de uma RO consiste basicamente na junção de *runs* próximos, com critérios bem definidos para proximidade. Para tanto, basta percorrermos todos os *runs*, da esquerda para a direita (segundo a ordem dadas pelos genes de um dos proteomas), e juntarmos cada vez que são próximos, de acordo com o segundo critério da definição de RO acima.

Para a determinação da espinha dorsal de dois proteomas, obtemos um alinhamento global dos proteomas. Para tanto, vamos exigir que um par de genes (g, h) seja candidato a se alinhar somente se g e h forem ortólogos. Estamos tentando minimizar a interferência de parálogos, que obviamente pode embaralhar o alinhamento. O objetivo mais específico é o de obter o maior alinhamento possível (com maior número de pares de genes ortólogos) sem que haja cruzamentos. É natural imaginarmos que, quanto mais próximos filogeneticamente forem os genomas, maior a espinha dorsal. Note que a espinha dorsal de duas RGCs não é necessariamente única. Assim, nossa tarefa é a de determinar a espinha dorsal que mais consiga evidenciar o quanto os proteomas são próximos, considerando um alinhamento. A estratégia que propomos para a construção da espinha dorsal é baseada no conhecido problema computacional chamado de *subseqüência comum mais longa*. Neste caso, cada símbolo corresponde ao número seqüencial do gene no proteoma e dois símbolos das cadeias são iguais se, e somente se, os respectivos genes são ortólogos.

Para encontrarmos famílias de genes parálogos, que são famílias de genes que supostamente evoluíram a partir de um mesmo ancestral, por duplicação, usamos critérios semelhantes aos utilizados na definição de ortologia, quais sejam, significância estatística da similaridade e cobertura de alinhamento.

A estratégia para a determinação de uma família de parálogos tem basicamente duas grandes fases. Na primeira fase encontramos todas as cliques maximais do grafo onde os vértices são os genes e dois vértices são vizinhos se, e somente se, os dois genes correspondentes são ortólogos. Em seguida, na segunda fase, genes não pertencentes a uma família podem ser acrescentados a ela desde que seja ortólogo a algum gene da família, agora seguindo um critério menos exigente.

Vale lembrar que os critérios propostos são "aproximações operacionais" de conceitos biológicos. São aproximações porque estamos usando similaridade para inferir

ancestralidade comum. São operacionais porque estamos propondo critérios que nos permitem escrever programas que automatizam a detecção dessa ortologia aproximada.

A implementação das metodologias propostas resultaram num programa chamado EGG, de Extended Genome-Genome comparison. EGG foi inicialmente proposto por nós em [1], e depois reformulado em [9]. A descrição feita na tese é uma segunda reformulação. O objetivo geral de EGG, portanto, é o de comparar dois proteomas. Em linhas gerais, EGG constrói um grafo bipartido onde os vértices são os genes de cada proteoma e as arestas representam ortologia entre os genes. Após a construção desse grafo bipartido, EGG encontra estruturas organizacionais envolvendo os genes.

EGG ainda permite a visualização gráfica dos *runs* e regiões ortólogas encontradas. Isto foi muito importante durante a execução dos projetos genoma, para que os participantes pudessem entender melhor quais e como os genes se mantêm agrupados de um genoma para o outro. A figura 1 mostra um exemplo de RO.

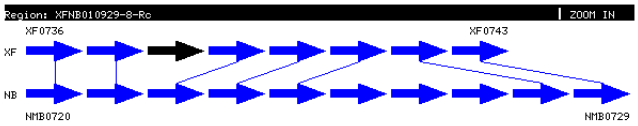

Figura 1: Exemplo de uma RO encontrada na comparação de Xylella fastidiosa e Neisseria meningitidis MC58, resultante da junção de 3 runs. Genes para os quais não se sabe a função são marcados com a cor mais escura.

4. Comparação de proteínas

Muita similaridade entre proteínas é um forte indício de que são homólogas. No entanto, quando trata-se de descobrir-se homologia entre proteínas com seqüências com pouca similaridade, o problema torna-se mais complicado. Essas homologias menos aparentes, que formam o que é chamada de *twilight zone*, costumam ser diluídas em ruídos capazes de confundir determinadas técnicas de comparação.

A contribuição desta seção, publicada em [2], consiste num algoritmo de programação dinâmica, adaptado dos já conhecidos, que usa o conceito de *regiões curingas*. Regiões curingas (formalmente definidas na tese) são regiões nas quais nós não estamos preocupados com a qualidade do alinhamento, por serem regiões supostamente menos conservadas. A idéia então é pontuarmos tais regiões de maneira mais relaxada, em detrimento de pontuarmos mais rigidamente regiões mais conservadas, que tendem a abrigar estruturas secundárias e onde acontecem menos substituições, inserções e remoções.

Os alinhamentos baseados em informações estruturais mostram que regiões conservadas contêm relativamente poucas diferenças, enquanto que as não-conservadas apresentam muitas diferenças. Um bom algoritmo deveria então usar este conhecimento e aplicar o conjunto de penalidades usual somente àquelas regiões supostamente conservadas; regiões não-conservadas, uma vez detectadas, deveriam ser excluídas do alinhamento. É isto que nosso algoritmo faz, atribuindo penalidades diferenciadas a regiões menos conservadas.

O algoritmo alinha duas seqüências usando função afim para penalizar buracos [10], acrescentando uma segunda função de pontuação para as colunas das regiões

curingas. Ele opera em dois modos. Num modo aplica as penalizações usuais para iniciar e estender um buraco. No outro usa: um parâmetro de inicialização de região curinga; um parâmetro de extensão de região curinga com casamentos e substituições; e um parâmetro de extensão de região curinga com buraco. A troca entre os modos é feita automaticamente pelas recorrências da programação dinâmica. A computação do valor do alinhamento, bem como sua construção, é feita da mesma forma que é feita pela programação dinâmica padrão. Nós usamos seis matrizes, ao invés de três, usualmente usadas. A complexidade do algoritmo é $O(mn)$, onde m, n são os tamanhos das seqüências.

5. Conclusão

Este trabalho tem como principal resultado um conjunto de metodologias e respectivas implementações para a comparação de genomas, tanto no nível de DNA, quanto no nível de seus genes. A metodologia proposta para a comparação de genomas no nível de DNA utiliza, como principal ferramenta, uma árvore de sufixos para as seqüências genômicas. A principal contribuição está no uso desse tipo de árvore, essencialmente preparada para lidar com repetições exatas, na busca por repetições aproximadas envolvendo os genomas.

No nível de genes, propomos uma metodologia para a comparação dos proteomas de dois genomas. A principal contribuição neste caso é a determinação de estruturas organizacionais envolvendo os genes que apresentam conservação de ordem e funcionalidade. Além disso, propomos uma metodologia para determinar a espinha dorsal de dois proteomas, na forma de um alinhamento, assim como para encontrar as famílias de genes parálogos de um genoma. Propomos também um algoritmo adaptado da programação dinâmica usual para comparação de proteínas, que faz uso das chamadas regiões curingas.

As metodologias propostas, assim como os programas, foram utilizados com sucesso em vários projetos genoma, como já dissemos. Além disso, nosso trabalho resultou nas publicações [2, 9] e co-autoria em [4, 5, 7, 11, 12, 13]. Vale ressaltar que o aspecto comparativo foi de extrema relevância em todos os projetos citados, principalmente nos casos do *Agrobacterium tumefaciens* (publicado na revista Science [13]) e no projeto das duas *Xanthomonas* (publicado na revista Nature [5]). No caso das *Xanthomonas*, por exemplo, a figura principal do artigo, que é a espinha dorsal dos dois genomas, e que de certa forma resume o trabalho de comparação, foi feita a partir dos dados gerados pelo programa EGG. A figura 2 mostra uma pequena parte dela.

Uma versão atualizada do pacote BACON/EGG estarão disponíveis brevemente no endereço `http://www.dct.ufms.br/~nalvo/baconegg/`, incluindo manuais e códigos-fonte, enquanto que a versão final da tese pode ser obtida no endereço `http://www.dct.ufms.br/~nalvo/publications/`, em vários formatos.

Referências

[1] N.F. Almeida and J.C. Setubal. A set of tools for detailed syntatic pairwise comparison of whole bacterial genomes. manuscript, 2000.

[2] N.F. Almeida, J.C. Setubal, and M. Tompa. On the use of don't care regions for protein sequence alignment. Technical Report 99-07, Institute of Computing, University of Campinas, Brazil, 1999.

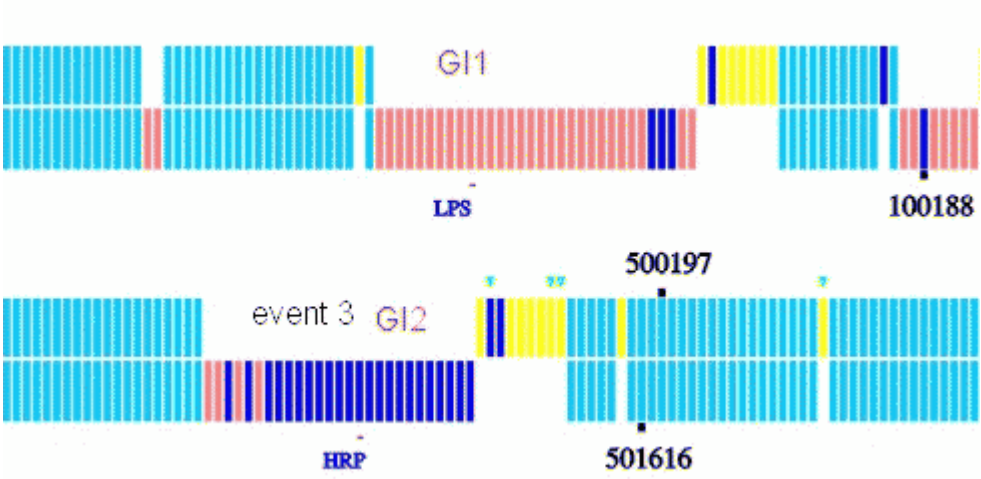

Figura 2: Pequena parte da principal figura do artigo das Xanthomonas, publicado na revista Nature [5].

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and psi-BLAST: a new generation of protein database search programs. *Nucleic Acid Research*, 25:3389–3402, 1997.

[4] G.S. Araújo and N.F. Almeida. Phylogeny from whole genome comparison. In *Proc. of the 1st Brazilian Workshop on Bioinformatics*, pages 9–15, October 2002.

[5] A.C. Rasera da Silva, J.C. Setubal, and N.F. Almeida et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*, 417(6887):459–463, 2002.

[6] A.L. Delcher, S. Kasif, R.D. Fleischmann, O. White J. Peterson, and S.L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.

[7] M.S.S. Felipe and N.F. Almeida et al. Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis. *Yeast*, 20:263–271, 2003.

[8] D. Gusfield. *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*. Cambridge University Press, 1997.

[9] J.C. Setubal and N.F. Almeida. Detection of related genes in procaryotes using syntenic regions. In *DIMACS Workshop on Whole Genome Comparison*. DIMACS Center, Rutgers University, February 2001.

[10] J.C. Setubal and J. Meidanis. *Introduction to computational molecular biology*. PWS Publishing Co., 1997.

[11] M.A. Van Sluys, J.C. Setubal, and N.F. Almeida et al. Comparative analyses of the complete genome sequences of Pierce's Disease and Citrus Variegated Chlorosis strains of *Xylella fastidiosa*. *J. Bacteriology,*, 185(3):1018–1026, 2003.

[12] D.W. Wood, J.C. Setubal, N.F. Almeida, and et al. Sequencing and analysis of the *Agrobacterium tumefaciens* genome. In *10th Int'l congress on Molecular plant-microbe interactions*, 2001. Madison, WI (poster).

[13] D.W. Wood, J.C. Setubal, and N.F. Almeida et al. The genome of *Agrobacterium tumefaciens*: insights into the evolution and evolution of a natural genetic enginner. *Science*, 294:2317–2323, December 2001.