RECIIS
Revista Eletrônica de Comunicação Informação & Inovação em Saúde
[www.reciis.cict.fiocruz.br]
ISSN 1981-6278

SUPLEMENTO – BIOINFORMÁTICA E SAÚDE

Artigos de revisão

Comparando genomas: bancos de dados e ferramentas computacionais para a análise comparativa de genomas procarióticos

DOI: 10.3395/reciis.v1i2.Sup.105pt

Marcos Catanho
Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz da Fundação Oswaldo Cruz, Rio de Janeiro, Brasil
mcatanho@fiocruz.br

Antonio Basílio de Miranda
Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz da Fundação Oswaldo Cruz, Rio de Janeiro, Brasil
antonio@fiocruz.br

Wim Degrave
Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz da Fundação Oswaldo Cruz, Rio de Janeiro, Brasil
wdegrave@fiocruz.br

Resumo
Desde a década de 1990, os esforços internacionais no sentido de obter seqüências genômicas completas levaram à determinação de todo o código genético de mais de 600 organismos, entre estes, procariotos, leveduras, protozoários, plantas, invertebrados e vertebrados, incluindo o próprio *Homo sapiens*. Atualmente, mais de 2.000 outros *projetos genoma* estão em andamento, representando interesses médicos, comerciais, ambientais e industriais, ou contemplando organismos-modelos importantes para o desenvolvimento de pesquisas científicas. Aliada ao vertiginoso avanço da computação nas últimas décadas, a obtenção de seqüências genômicas completas de inúmeros organismos têm permitido o uso de abordagens holísticas e ao mesmo tempo inovadoras no estudo da estrutura, organização e evolução dos genomas e na predição e classificação funcional de genes, entre outros. Inúmeros bancos de dados e ferramentas computacionais de acesso público ou privado têm sido criados na tentativa de organizar e permitir acesso eficiente e rápido a estas informações através da *internet*. Nesta revisão apresentamos os principais recursos disponíveis publicamente na *internet* para a análise comparativa de genomas procarióticos, especialmente de genomas micobacterianos, grupo que contém importantes patógenos humanos e de animais. A Bioinformática e a Biologia Computacional, áreas do conhecimento responsáveis pelo desenvolvimento e aplicação de tais instrumentos computacionais, são também abordadas, enfatizando-se suas origens e contribuições para o desenvolvimento da ciência.

Palavras-chave
Bioinformática, biologia computacional, banco de dados, genoma, procariotos

Princípio de uma nova era: o surgimento da Bioinformática e da Biologia Computacional

A **Bioinformática** e a **Biologia Computacional** têm suas origens na década de 1960, quando os computadores emergiram como ferramentas importantes na Biologia Molecular. Este surgimento, segundo Hagen (2000), teria sido motivado por três fatores principais: (i) pelo crescente número de seqüências protéicas disponíveis, que representavam, ao mesmo tempo, uma fonte de dados e um conjunto de problemáticas importantes, porém intratáveis sem o auxílio de um computador; (ii) pela idéia de que as macromoléculas carregam informação ter se tornado parte fundamental do modelo conceitual da Biologia Molecular; (iii) pela disponibilidade de computadores mais velozes nas universidades e centros de pesquisa.

Até o final dos anos 1960, diversas técnicas computacionais (**algoritmos** e programas de computador) para análise da estrutura, função e evolução moleculares, bem como bancos de dados rudimentares de seqüências protéicas, já haviam sido desenvolvidos (HAGEN, 2000; revisto por OUZOUNIS e VALENCIA, 2003). Novas técnicas e abordagens foram desenvolvidas nas décadas seguintes, destacando-se os algoritmos para **alinhamento de seqüências**, a criação de bancos de dados de acesso público, a implementação de sistemas rápidos de busca em bancos de dados, o desenvolvimento de sistemas mais sofisticados para a predição de estrutura de proteínas, de ferramentas para anotação e comparação de **genomas** e de sistemas para análise funcional de genomas (OUZOUNIS, 2002).

Foi somente na década de 1980, no entanto, que a Bioinformática e a Biologia Computacional tomaram forma de disciplinas independentes, com seus próprios problemas e conquistas, sendo a primeira vez em que algoritmos eficientes foram desenvolvidos para lidar com o volume crescente de informação e que implementações destes algoritmos (programas) foram disponibilizadas para toda a comunidade científica (OUZOUNIS e VALENCIA, 2003). A afirmação definitiva destas novas disciplinas aconteceu na década de noventa, com o surgimento dos *projetos genoma*, ***transcriptoma*** e ***proteoma*** (sustentados por avanços importantes nos métodos de seqüenciamento de ADN, no desenvolvimento de ***microarrays*** e ***biochips*** e na espectrometria de massa), das redes de computadores em escala mundial (*internet*), de bancos de dados biológicos imensos, de supercomputadores e de computadores pessoais bastante robustos.

De fato, a obtenção de seqüências genômicas completas de inúmeros organismos, de dados de expressão gênica e protéica de células, tecidos e órgãos inteiros aliada ao desenvolvimento de tecnologias de computação de alto desempenho e de algoritmos mais eficientes, permitiu o uso de abordagens holísticas (que consideram integralmente todo o corpo de informações disponíveis, como por exemplo, todos os genes codificados por um grupo de genomas analisados) no estudo da estrutura, organização e evolução de genomas, no estudo da expressão diferencial de genes e proteínas, na análise da estrutura tridimensional de proteínas, no processo de reconstrução metabólica e na predição funcional de genes. Como resultado, a Bioinformática e a Biologia Computacional produziram ao longo destes anos pelo menos duas possíveis constatações gerais (que sintetizam diversas observações experimentais) aplicáveis aos sistemas biológicos, considerando-se a existência de várias deduções decorrentes destas com aplicação direta no campo da pesquisa biológica: (i) as estruturas tridimensionais de moléculas protéicas são muito mais conservadas do que suas funções bioquímicas; (ii) a comparação do número total de genes codificados em um dado genoma com o número total de genes codificados em outros genomas não reflete a filogenia das espécies, mas a comparação entre suas seqüências genômicas sim (OUZOUNIS, 2002).

Novos desafios, novas abordagens: a análise comparativa de genomas procarióticos

A iniciativa pioneira do Departamento de Energia Norte-Americano (DOE) de obter uma seqüência genômica humana de referência que pudesse atender melhor os seus propósitos de compreender os riscos potenciais para a saúde e para o meio ambiente da produção e do uso de novas fontes de energia e novas tecnologias, culminou no lançamento do Projeto Genoma Humano, em 1990. Mais tarde, os recursos tecnológicos gerados por este projeto estimularam o desenvolvimento de muitos outros *projetos genoma*, tanto por setores públicos quanto por setores privados (HGP 2001).

Atualmente, além do mapa completo do genoma humano (VENTER et al., 2001; LANDER et al., 2001) e de alguns outros vertebrados e plantas, totalizando 70 genomas, 47 arqueobactérias e 543 eubactérias já tiveram seus genomas inteiramente seqüenciados e outros 2.258 projetos estão em andamento (GOLD, 2007). Entre as **micobactérias**, 16 representantes já tiveram seus genomas inteiramente seqüenciados e outros 23 estão em curso (Tabela 1).

Tabela 1 – *Projetos Genoma* de Micobactérias

Espécie ou cepa	Importância	Centro de Pesquisa	URL	Status
M. tuberculosis H37Ra	Médica; patógeno de animais e humanos; causadora de tuberculose.	Beijing Genomics Institute	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=21081	Completo
M. tuberculosis F11 (ExPEC)	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Completo

cont.

Tabela 1 – *Projetos Genoma* de Micobactérias (cont.)

M. bovis BCG Pasteur 1173P2	Médica; patógeno de animais, gado e humanos; causadora de tuberculose.	Institut Pasteur	http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html	Completo
M. ulcerans Agy99	Médica; patógeno humano; causadora de úlcera de Buruli.	Institut Pasteur	http://www.pasteur.fr/recherche/unites/Lgmb/mycogenomics.html	Completo
M. flavenscens PYR-GCK	Biotecnológica; isolada de solo.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/mycfl/mycfl.home.html	Completo
M. vanbaalenii PYR-1	Biotecnológica; isolada de solo.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/mycva/mycva.home.html	Completo
Mycobacterium sp JLS	Biotecnológica; isolada de solo contaminado por creosoto.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_j/myc_j.home.html	Completo
Mycobacterium sp KMS	Biotecnológica; isolada de solo contaminado por creosoto.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_k/myc_k.home.html	Completo
Mycobacterium sp MCS	Biotecnológica; isolada de solo contaminado por creosoto.	Joint Genome Institute	http://genome.jgi-psf.org/finished_microbes/myc_k/myc_k.home.html	Completo
M. tuberculosis H37Rv	Médica; patógeno de animais e humanos; causadora de tuberculose.	Sanger Institute	http://www.sanger.ac.uk/Projects/M_tuberculosis/	Completo
M. bovis AF2122/97	Médica; patógeno de animais, gado e humanos; causadora de tuberculose.	Sanger Institute/ Institut Pasteur	http://www.sanger.ac.uk/Projects/M_bovis/	Completo
M. leprae TN	Médica; patógeno humano; causadora da hanseníase.	Sanger Institute/ Institut Pasteur	http://www.sanger.ac.uk/Projects/M_leprae/	Completo
M. avium 104	Médica; patógeno de animais; causadora de infecções respiratórias.	The Institute for Genomic Research	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=20086	Completo
M. smegmatis MC2 155	Médica; patógeno humano; oportunista.	The Institute for Genomic Research	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gms	Completo
M. tuberculosis CDC1551	Médica; patógeno de animais e humanos; causadora de tuberculose.	The Institute for Genomic Research	http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=gmt	Completo
M. avium paratuberculosis k10	Médica; patógeno de animais e gado; causadora da doença de Johne, paratuberculose e enterite.	University of Minnesota	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Overview&list_uids=380	Completo
M. tuberculosis A1	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	-	Incompleto
M. tuberculosis C	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incompleto
M. tuberculosis Ekat-4	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	-	Incompleto
M. tuberculosis Haarlem	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incompleto
M. tuberculosis KZN 1435 (MDR)	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incompleto
M. tuberculosis KZN 4207 (DS)	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incompleto
M. tuberculosis KZN 605 (XDR)	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	http://www.broad.mit.edu/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html	Incompleto

cont.

Tabela 1 – *Projetos Genoma* de Micobactérias (cont.)

M. tuberculosis Peruvian1	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	-	Incompleto
M. tuberculosis Peruvian2	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	-	Incompleto
M. tuberculosis W-148	Médica; patógeno humano; causadora de tuberculose.	Broad Institute	-	Incompleto
M. ulcerans	Médica; patógeno humano; causadora de úlcera de Buruli.	Clamson University	http://www.genome.clemson.edu/projects/stc/m.ulcerans/MU__Ba/index.html	Incompleto
M. bovis BCG Moreau[a]	Médica; patógeno de animais, gado e humanos; causadora de tuberculose.	Fundação Oswaldo Cruz / Fundação Ataulpho de Paiva	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=18279	Incompleto
M. abscessus CIP 104536	Médica; patógeno humano; causadora de infecções bronco-pulmonares e respiratórias.	Genoscope	http://www.genoscope.cns.fr/externe/English/Projets/Projet_LU/organisme_LU.html	Incompleto
M. chelonae CIP 104535	Médica; patógeno humano; causadora de infecções bronco-pulmonares e respiratórias.	Genoscope	http://www.genoscope.cns.fr/externe/English/Projets/Projet_LU/organisme_LU.html	Incompleto
Mycobacterium sp. Spyr1	Biotecnológica; isolada de solo contaminado por creosoto.	Joint Genome Institute / University of Ioannina	-	Incompleto
M. liflandii 128FXT	Médica; patógeno de sapo e outros animais; causadora de infecção sistêmica.	Monash University	-	Incompleto
M. marinum DL240490	Médica; patógeno de peixe e humanos; causadora de infecção semelhante à tuberculose em peixes e infecção de pele.	Monash University	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=20229	Incompleto
M. ulcerans 1615	Médica; patógeno humano; causadora de úlcera de Buruli.	Monash University	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=20231	Incompleto
M. africanum GM041182	Médica; patógeno humano, de gado e de animais; causadora de tuberculose.	Sanger Institute	http://www.sanger.ac.uk/sequencing/Mycobacterium/africanum/	Incompleto
M. canetti CIPT140010059	Médica; patógeno de humanos, gado e animais; causadora de tuberculose.	Sanger Institute	http://www.sanger.ac.uk/sequencing/Mycobacterium/canetti/	Incompleto
M. microti OV254	Médica; patógeno de animais, gado e humanos; causadora de tuberculose.	Sanger Institute / Institut Pasteur	http://www.sanger.ac.uk/Projects/M_microti/	Incompleto
M. marinum M	Médica; patógeno de animais e humanos; causadora de tuberculose.	Sanger Institute / University of Washington / Institut Pasteur / Monash University / University of Tennessee	http://www.sanger.ac.uk/Projects/M_marinum/	Incompleto
M. tuberculosis 210	Médica; patógeno humano; causadora de tuberculose.	The Institute for Genomic Research	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genomeprj&cmd=ShowDetailView&TermToSearch=273	Incompleto

Fontes: Genomes Online Database (GOLD 2007), NCBI Entrez Genome Project Database (Genome Project 2007) e Comprehensive Microbial Resource (CMR 2007).

Seqüências genômicas completas constituem uma fonte de dados única pois, em princípio, elas representam tudo o que é necessário para criar um organismo, juntamente com **fatores epigenéticos** e através de sua interação com estes fatores. No entanto, não é imediatamente óbvio o que se pode fazer com toda esta informação. Acredita-se, por exemplo, que a análise sistemática de todo o conteúdo gênico de um organismo tem o potencial de levar à compreensão integral da genética, da bioquímica, da fisiologia e da patogênese dos microrganismos (BROSCH et al., 2001). Entretanto, argumenta-se que este potencial só é capaz de concretizar-se através do estudo comparativo dos genomas ou de **regiões sintênicas** de duas ou mais espécies, subespécies ou cepas, porque a visão isolada do ADN de um único organismo, fora do contexto filogenético do processo evolutivo, nos permite uma compreensão apenas parcial destas questões (WEI et al., 2002).

Neste sentido, FRASER et al. (2000) deram exemplos claros de como a perspectiva evolutiva pode beneficiar estas análises genômicas, tais como auxiliar na identificação da função biológica de novos genes, na inferência de padrões de recombinação nas espécies, na ocorrência de transferência lateral de genes entre diferentes espécies e na perda de material genético, além de contribuir para a distinção entre similaridades devidas a **homologia** e similaridades originadas por **convergência**. Por outro lado, KONDRASHOV (1999) e KOONIN et al. (2000) destacaram a importância dos resultados obtidos com a análise comparativa de genomas para a Biologia Evolutiva. Segundo KONDRASHOV (1999), os produtos destas análises têm fornecido as melhores evidências disponíveis para alguns fenômenos evolutivos e, em alguns casos, levado ao refinamento de antigos conceitos. Mais recentemente, novas abordagens de análise filogenética que tentam explorar todo o conteúdo gênico de genomas inteiramente seqüenciados têm sido desenvolvidas, e diferentes métodos de calcular a distância entre os genomas de distintas espécies têm sido propostos (OTU e SAYOOD 2003; HENZ et al. 2005; KUNIN et al. 2005a e referências contidas neste trabalho; KUNIN et al. 2005b; TEKAIA et al. 2005), superando problemas antigos e comuns aos métodos tradicionais de análise filogenética, como por exemplo, a saturação de determinadas posições nos códons, a escolha de marcadores evolutivos apropriados e desvios nas análises provocados por estes fatores. Há, portanto, uma *alça de retro-alimentação* entre as análises evolutivas e genômicas, como afirmaram FRASER et al. (2000).

É importante ressaltar que nos últimos anos, desde o seqüenciamento dos primeiros genomas bacterianos em 1995, análises comparativas de genomas procarióticos têm nos revelado cada vez mais a natureza complexa da estrutura e organização destes genomas e a enorme diversidade genética entre estes organismos (muito acima daquela esperada, mesmo entre isolados de uma mesma espécie), levando a questionamentos importantes sobre os mecanismos pelos quais estes microrganismos evoluem e como devem ser classificados taxonomicamente (COENYE et al. 2005; BINNEWIES et al. 2006).

No que se refere aos microrganismos patogênicos e às micobactérias em especial, várias aplicações potenciais da análise comparativa de genomas têm sido reportadas, visando sobretudo à prevenção (através do desenvolvimento de vacinas mais eficazes), o tratamento (pelo desenvolvimento de novas drogas) e o diagnóstico (através da criação de métodos mais rápidos, sensíveis e específicos) da tuberculose e outras doenças causadas por micobactérias. Algumas dessas aplicações incluem: a identificação de genes únicos de uma espécie em particular, a identificação de fatores de virulência e a reconstrução metabólica (GORDON et al. 2002); a caracterização de patógenos, a identificação de novos alvos para diagnóstico e para procedimentos terapêuticos (FITZGERALD e MUSSER, 2001); a investigação sobre a origem molecular da patogênese, do espectro de hospedeiros e das diferenças fenotípicas entre isolados clínicos e populações naturais de patógenos (BEHR et al. 1999; BROSCH et al. 2001; COLE 2002; KATO-MAEDA et al. 2001) e a investigação dos fundamentos genéticos da virulência e da resistência a drogas de micobactérias causadoras de tuberculose (RANDHAWA e BISHAI, 2002).

A análise comparativa de genomas é uma abordagem relativamente recente, tendo início com o seqüenciamento dos primeiros genomas na década de 1990. No entanto, suas ferramentas mais importantes têm origem nas técnicas clássicas de análise de seqüências: algoritmos de **alinhamento global** e **local** de pares de seqüências ou de múltiplas seqüências, métodos de análise filogenética e as implementações destes métodos e algoritmos (NEEDLEMAN e WUNSCH, 1970; SMITH e WATERMAN, 1981; LIPMAN e PEARSON, 1985; PEARSON e LIPMAN, 1988; FENG e DOOLITTLE, 1987; ALTSCHUL et al. 1990; 1997; THOMPSON et al., 1994; FELSENSTEIN, 1981; 1989). De fato, ela se beneficia não somente de ferramentas desenvolvidas no passado, mas também da criação de novas ferramentas e do aperfeiçoamento das ferramentas já existentes, estimulados pela imensa, diversificada e complexa quantidade de dados produzida com os projetos de seqüenciamento em larga escala.

Análises comparativas de genomas podem ser feitas em diferentes níveis de abordagem, oferecendo múltiplas perspectivas acerca dos organismos estudados (revisto por WEI et al., 2002): (i) comparação da estrutura genômica, incluindo a descrição de parâmetros estruturais do ADN, a análise de repetições e de regiões de baixa complexidade em geral, a identificação de rearranjos tanto ao nível do ADN quanto ao nível dos genes, a identificação de sintenia e a análise de regiões limítrofes entre regiões sintênicas vizinhas (*breakpoints*); (ii) comparação das regiões codificantes, abrangendo a identificação destas regiões, a comparação dos conteúdos gênico e protéico, a identificação de regiões conservadas entre os genomas comparados, a análise da conservação de grupos de seqüências e de genes **ortólogos**, da conservação de famílias de genes **parálogos** e da conservação da localização dos genes entre as diferentes espécies estudadas e a análise da ocorrência de eventos de **fusão** e/ou ligação funcional entre genes; (iii) comparação de regiões não codificantes, envolvendo a identificação de elementos regulatórios.

Sendo os genomas basicamente longas seqüências, poder-se-ia analisá-los alinhando-os como se fossem seqü-

ências comuns, utilizando um dos algoritmos de análise de seqüências citados anteriormente. No entanto, isto só pode ser feito com genomas de espécies muito próximas, uma vez que mudanças na estrutura do ADN (inserções, deleções, inversões, rearranjos, trocas e duplicações) ocorrem com uma taxa muito elevada. Além disto, por tratar-se de seqüências de tamanho extremo, torna-se computacionalmente inviável a análise de mais de um par de genomas de uma só vez, mesmo com o uso de algoritmos e programas eficazes, especialmente desenvolvidos para esta finalidade (MORGENSTERN et al., 1998; 1999; 2002; JAREBORG et al., 1999; DELCHER et al., 1999; 2002; KENT e ZAHLER, 2000; BATZOGLOU et al., 2000; MA et al., 2002; BRAY et al., 2003; 2004; SCHWARTZ et al., 2003b; BRUDNO et al., 2003a; b; KURTZ et al., 2004). Portanto, na maioria das vezes as análises comparativas entre genomas são feitas em um nível de abordagem mais modular, tomando-se as partes que compõem tais seqüências, como por exemplo, o conjunto completo de genes codificados pelas espécies em estudo.

A etapa crucial deste tipo de análise é determinar se as seqüências comparadas são ou não homólogas, ou seja, se descendem ou não de uma seqüência ancestral comum, estabelecendo-se equivalência entre as partes comparadas. O resultado obtido permite, entre outras coisas, a predição de função, já que é presumido que seqüências homólogas tendem a ter funções similares (BORK e KOONIN, 1998) e também determinar quais os genes correspondentes entre os pares ou grupos de genomas analisados. Esta tarefa nada trivial é feita comparando-se uma ou mais seqüências de entrada (*query sequences*), com outras inúmeras seqüências depositadas em um banco de dados (*subject sequences*), através do alinhamento consecutivo de cada seqüência de entrada com cada seqüência depositada no banco, com a utilização de um algoritmo de alinhamento local (SMITH e WATERMAN, 1981; PEARSON e LIPMAN, 1988; ALTSCHUL et al., 1997). Para cada alinhamento, calcula-se o número de pontos obtidos (*score*), com base em uma **matriz de substituição** (PAM ou BLOSUM normalmente) e em valores arbitrados de penalidade para a abertura e extensão de espaços nas seqüências alinhadas (*gap opening/extension penalties*), e o número de alinhamentos esperados ao acaso com pontuação igual ou superior ao do alinhamento em questão (*E-value*), a partir da pontuação normalizada (*bitscore*) e do tamanho e composição do banco de dados. A homologia é inferida com base nos valores calculados dos diferentes parâmetros do alinhamento, alguns deles já mencionados: pontuação, pontuação normalizada, número de alinhamentos esperados ao acaso com pontuação igual ou superior ao do alinhamento em questão, percentual de identidade, percentual da extensão de cada seqüência no par alinhado que contribui para o alinhamento, diferença de tamanho entre as seqüências alinhadas etc. A existência de domínios (módulos que constituem unidades distintas do ponto de vista evolutivo, funcional e estrutural) em proteínas é um fator complicador nestas análises, que deve ser tratado com atenção.

Comparando genomas: os recursos computacionais disponíveis para a análise comparativa de genomas procarióticos

Inúmeros bancos de dados e ferramentas computacionais de acesso público (na grande maioria) ou privado têm sido criados na tentativa de organizar e permitir acesso eficiente e rápido às informações geradas pelos projetos de larga escala mencionados anteriormente (revisto de forma exaustiva em HIGGINS e TAYLOR, 2000), bem como permitir a análise comparativa dessa quantidade maciça de dados (Tabela 2). A criação e manutenção de bancos de dados biológicos são por si só um desafio, devido não só à imensa quantidade de dados, mas sobretudo devido à dificuldade de desenvolver esquemas e estruturas que representem de forma exata ou bastante aproximada a complexa relação existente entre os diversos componentes dos sistemas biológicos (MACÊDO et al., 2003). Outra dificuldade é a criação de mecanismos eficientes de busca e obtenção de dados nestes bancos, que permitam a elaboração e execução de consultas complexas e maciças, através de uma interface amigável para o usuário. É importante ressaltar que, em muitos casos, os criadores e curadores destes bancos recebem pouca ou nenhuma remuneração pelos seus esforços e conseguir financiamento para a criação e manutenção de bancos de dados biológicos ainda é uma tarefa difícil nos dias atuais (GALPERIN, 2005).

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos

Nome	Descrição	Referência(s)	URL
BANCOS DE DADOS			
Genéricos e multifuncionais			
BacMap	Atlas (coleção de mapas genômicos de alta resolução) interativo para a exploração de genomas bacterianos. Contém extensa anotação de genes e oferece, para cada genoma, gráficos representando estatísticas globais, como composição de bases e de aminoácidos, distribuição do tamanho das seqüências protéicas, preferência por fita de ADN, entre outras.	STOTHARD et al., 2005	http://wishart.biology.ualberta.ca/BacMap/

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

CMR	*Comprehensive Microbial Resource*. Oferece acesso a ampla gama de informações e análises sobre todos os genomas procarióticos já seqüenciados. Buscas podem ser feitas por genes, genomas, regiões genômicas e propriedades dos genes. Comparações entre múltiplos genomas podem ser executadas com base em diferentes critérios, tais como similaridade de seqüência e atributos dos genes.	PETERSON et al., 2001	http://cmr.tigr.org/
Genome Atlas Database	Desenvolvido para a visualização e comparação de características estruturais do ADN de genomas microbianos seqüenciados (composição de bases, energia de empilhamento, posição preferencial, sensibilidade a DNase I, curvatura intrínseca, entre outras).	HALLIN e USSERY, 2004	http://www.cbs.dtu.dk/services/GenomeAtlas/
IMG	*Integrated Microbial Genomes*. Plataforma para análise comparativa de genomas seqüenciados pelo grupo *Joint Genome Institute* pertencente ao DOE. Foi desenvolvido para facilitar a visualização e exploração de genomas a partir de uma perspectiva funcional e evolutiva.	MARKOWITZ et al., 2006	http://img.jgi.doe.gov
MBGD	*Microbial Genome Database*. Permite a criação de tabelas de classificação de genes ortólogos usando algoritmo próprio, dados pré-computados de similaridade e grupos de organismos e parâmetros selecionados pelo usuário. Oferece análise de perfis filogenéticos, comparação da ordem e estrutura dos genes e classificação funcional.	UCHIYAMA, 2003; 2006	http://mbgd.genome.ad.jp/
MicrobesOnline	Banco de dados para análise comparativa de genomas procarióticos. Integra várias ferramentas disponíveis para análise genômica e de seqüências, oferecendo dados pré-computados de predição de óperons e seqüências regulatórias, e de grupos de ortólogos, para centenas de genomas procarióticos.	ALM et al., 2005	http://www.microbesonline.org/
PLATCOM	Plataforma para genômica comparativa computacional. Ambiente onde os usuários podem escolher livremente qualquer combinação entre centenas de genomas e compará-los através de um conjunto de ferramentas computacionais para a análise de seqüências, interconectadas entre si e com bancos de dados internos, estabelecendo seu próprio protocolo experimental para investigar similaridades de seqüência e sintenia, vias metabólicas conservadas e potenciais eventos de fusão gênica.	CHOI et al., 2005	http://platcom.informatics.indiana.edu/platcom/
PUMA2	Sistema interativo e integrado de bioinformática para análises maciças de seqüências e reconstrução metabólica. Oferece estrutura para análises comparativas e evolutivas de genomas e redes metabólicas, em um contexto taxonômico e fenotípico. Contém mais de 1.000 genomas procarióticos e eucarióticos, além de genomas virais e mitocondriais.	MALTSEV et al., 2006	http://compbio.mcs.anl.gov/puma2/
	Organismos ou grupos – específicos		
GenoList	Coleção de bancos de dados dedicados à análise de genomas microbianos, individualmente ou em conjunto. Oferecem um conjunto completo de dados de seqüências protéicas e nucleotídicas destas espécies, relacionados às respectivas anotações e classificações funcionais, permitindo ao usuário navegar através destes dados e obter informações usando diferentes critérios de busca e análise de seqüência: nome do gene, localização, palavra-chave, categoria funcional etc. e busca por padrões ou por similaridade de seqüência.	FANG et al., 2005	http://genolist.pasteur.fr/

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

GenoMycDB	*Banco de dados relacional* para análise comparativa de genes micobacterianos. O banco armazena parâmetros e valores computados de similaridade entre todas as seqüências protéicas preditas codificadas pelos genomas de seis diferentes micobactérias. Oferece para cada uma destas proteínas a sua localização subcelular predita, sua classificação em COG(s), descrição dos genes correspondentes e ligações com diversos outros bancos de dados. Através de uma interface amigável, tabelas de pares ou grupos de proteínas homólogas potenciais, entre as espécies selecionadas, podem ser geradas dinamicamente com critérios definidos pelo próprio usuário.	CATANHO et al., 2006	http://www.dbbm.fiocruz.br/GenoMycDB
LEGER	Banco de dados para análise comparativa de genomas do gênero *Listeria*. Reúne dados pré-computados de comparações genômicas e listas de genes ortólogos potenciais obtidas com parâmetros pré-definidos. Permite análises funcionais e de vias metabólicas, busca e mineração de dados através de sistemas próprios de integração e obtenção de dados, entre outros. Disponibiliza, de forma integrada, dados experimentais resultantes de análises proteômicas.	DIETERICH et al., 2006	http://leger2.gbf.de/cgi-bin/expLeger.pl
MolliGen	Banco de dados para análise comparativa de genomas de Mollicutes. Reúne dados pré-computados de comparações genômicas e listas de genes ortólogos potenciais obtidas com parâmetros pré-definidos. Permite análises funcionais e de vias metabólicas, busca e mineração de dados através de sistemas próprios de integração e obtenção de dados, entre outros.	BARRÉ et al., 2004	http://cbi.labri.fr/outils/molligen/
ShiBASE	Banco de dados para análise comparativa de genomas do gênero *Shigella*. Reúne dados pré-computados de comparações genômicas e listas de genes ortólogos potenciais obtidas com parâmetros pré-definidos. Permite análises funcionais e de vias metabólicas, busca e mineração de dados através de sistemas próprios de integração e obtenção de dados, entre outros. Disponibiliza, de forma integrada, dados experimentais resultantes de análises comparativas de hibridação em escala genômica.	YANG et al., 2006	http://www.mgc.ac.cn/ShiBASE/
xBASE	Coleção de bancos de dados dedicados à análise comparativa de genomas bacterianos. Reúnem dados pré-computados de comparações entre genomas de gêneros específicos e relacionados, listas de genes ortólogos potenciais, anotações funcionais, referências e resultados de análises de utilização de códons, composição de bases, CAI - *codon adaptation index*, hidrofobicidade e aromaticidade de proteínas. As buscas são orientadas por genoma e podem ser feitas através de diferentes critérios: anotação, nome do gene, localização etc.	CHAUDHURI e PALLEN, 2006	http://xbase.bham.ac.uk/
	Especializados		
COG	*Clusters of Orthologous Groups*. Representa uma tentativa de classificação filogenética de grupos de proteínas preditas codificados por genomas procarióticos (e também eucarióticos), integralmente seqüenciados. Através de inúmeras páginas navegáveis o usuário tem acesso a diversos dados pré-computados, como por exemplo, padrões filogenéticos, classificações funcionais, listas de grupos de genes ortólogos (COG) por categoria funcional ou por via metabólica, entre outros.	TATUSOV et al., 1997; 2003	http://www.ncbi.nlm.nih.gov/COG

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

FusionDB	Banco de dados que oferece uma análise bastante densa sobre eventos de fusão gênica em procariotos, proporcionando uma base para a busca de potenciais interações entre proteínas e redes de regulação metabólica.	SUHRE e CLAVERIE, 2004	http://igs-server.cnrs-mrs.fr/FusionDB/
HAMAP	*High-Quality Automated and Manual Annotation of Microbial Proteomes*. Coleção de famílias de proteínas ortólogas microbianas, geradas manualmente por especialistas (curadores). Oferece para cada família, extensa anotação, alinhamentos, **perfis** e atributos computados (regiões transmembranares, sinais para exportação, entre outros).	GATTIKER et al., 2003	http://www.expasy.org/sprot/hamap/
Hogenom	Banco de dados de seqüências homólogas de genomas completamente seqüenciados. Permite a seleção de genes homólogos entre espécies e a visualização de alinhamentos múltiplos e árvores filogenéticas.	DUFAYARD et al., 2005	http://pbil.univ-lyon1.fr/databases/hogenom.html
IslandPath	Sistema que incorpora características comumente associadas à presença de ilhas genômicas (grupos de genes que foram potencialmente transferidos horizontalmente, incluindo ilhas de patogenicidade) - tais como anomalias no conteúdo GC, desvios na composição dinucleotídica, entre outros -, em uma representação gráfica do genoma de procariotos, auxiliando a detecção de tais estruturas.	HSIAO et al., 2003	http://www.pathogenomics.sfu.ca/islandpath/
KEGG	*Kyoto Encyclopedia of Genes and Genomes*. Grande plataforma que integra vários bancos de dados diferentes reunidos em três categorias principais: redes de interação molecular (vias bioquímicas) em processos biológicos, informação sobre o universo de genes e proteínas e informação sobre a vasta gama de componentes químicos e reações. A primeira contém uma coleção de mapas manualmente elaborados, representando o conhecimento atual sobre interação molecular e redes de interação. A seção dedicada às informações genômicas baseia-se em resultados pré-computados de comparação de seqüências, busca de **motivos** e padrões e agrupamento de genes ortólogos.	KANEHISA, 1997; KANEHISA e GOTO, 2000; KANEHISA et al., 2006	http://www.genome.jp/kegg
MetaCyc	Banco de dados não redundante de vias metabólicas experimentalmente elucidadas, abrangendo 700 vias de mais de 600 organismos diferentes. Contém vias metabólicas, reações enzimáticas, enzimas, compostos químicos, genes e revisões. Informações sobre enzimas incluem especificidade de substrato, propriedades cinéticas, ativadores, inibidores e outros. Oferece uma variedade de aplicações, tais como predição computacional de vias metabólicas, análise comparativa de redes bioquímicas, entre outras.	CASPI et al., 2006	http://metacyc.org/
OMA Browser	Interface *web* que permite explorar pares ou grupos de ortólogos em um banco de dados resultante do projeto OMA de identificação de ortólogos em genomas completamente seqüenciados.	SCHNEIDER et al., 2007	http://omabrowser.org/
ORFanage	Banco de dados desenvolvido para investigar e classificar genes órfãos (genes exclusivos de uma espécie, família ou linhagem). Consiste em ORF (fases abertas de leitura) preditas computacionalmente em genomas totalmente seqüenciados, permitindo buscas orientadas por classes de genes órfãos (únicos, parálogos e ortólogos).	SIEW et al., 2004	http://www.cs.bgu.ac.il/~nomsiew/ORFans/

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

OrphanMine	Banco de dados desenvolvido para a análise de genes órfãos (taxonomicamente restritos) de forma comparativa. Construído a partir da comparação par a par entre todas as proteínas preditas codificadas nos genomas de mais de 300 espécies bacterianas. Permite a detecção de genes órfãos com base em diferentes critérios (similaridade de seqüência, tamanho, conteúdo GC, entre outros).	WILSON et al., 2005	http://www.genomics.ceh.ac.uk/orphan_mine/faq.php
OrthoMCL-DB	Banco de dados de grupos de ortólogos preditos para 55 espécies diferentes, incluindo procariotos e eucariotos. Os grupos são formados com base em similaridade de seqüência, através de algoritmo próprio (OrthoMCL). A busca e a obtenção de dados podem ser executadas através de palavras-chaves e similaridade de seqüência, entre outros. Oferece visualização e análise de perfis filogenéticos, arquitetura de domínios, similaridade de seqüência e outros, através de representações gráficas.	CHEN et al., 2006	http://orthomcl.cbil.upenn.edu
ProtRepeatsDB	Banco de dados de diferentes tipos de repetições de aminoácidos presentes em seqüências protéicas de centenas de genomas completamente seqüenciados. Oferece um conjunto de ferramentas para identificação rápida e em larga escala de repetições aminoacídicas, facilitando a análise comparativa e evolutiva destas repetições.	KALITA et al., 2006	http://bioinfo.icgeb.res.in/repeats/
RoundUp	Repositório de grupos de genes ortólogos entre centenas de espécies e suas respectivas distâncias evolutivas, computados com algoritmo próprio (*Reciprocal Smallest Distance*). Oferece busca e obtenção de dados por genes ou genomas, apresentando os resultados na forma de perfis filogenéticos, acompanhados de anotação dos genes e funções moleculares.	DELUCA et al., 2006	https://rodeo.med.harvard.edu/tools/roundup/
SEED	Banco de dados extensamente curado e não redundante desenvolvido pela organização chamada *Fellowship for Interpretation of Genomes* (FIG), através da compilação de dados obtidos de diversas fontes (GenBank, RefSeq, UniProt, KEGG e de centros seqüenciadores de genomas). Oferece uma plataforma de apoio a análise comparativa de genomas, aberta à contribuição de toda a comunidade científica, na qual a anotação dos genomas é orientada por subsistemas (vias bioquímicas inteiras ou parciais, grupos de genes relacionados funcionalmente entre si).	OVERBEEK et al., 2005	http://theseed.uchicago.edu/FIG/index.cgi
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*. Banco de dados de interações preditas ou já conhecidas entre proteínas. As interações incluem associações diretas (físicas) e indiretas (funcionais), derivadas de quatro fontes diferentes: contexto genômico, experimentos de alto desempenho, co-expressão e conhecimento experimental prévio. O banco integra quantitativamente os dados de interações obtidos destas fontes para centenas de organismos e transfere informação entre eles, quando possível.	VON MERING et al., 2005; 2007	http://string.embl.de/
TransportDB	Banco de dados que descreve proteínas transportadoras de membrana celular preditas em organismos cujo genoma já foi inteiramente seqüenciado. As proteínas identificadas são classificadas em diferentes tipos e famílias, de acordo com a topologia predita, família protéica, bioenergética e especificidade de substrato. Oferece busca por similaridade de seqüência, comparação entre sistemas de transporte em diferentes organismos, árvores filogenéticas de famílias de transportadores em particular, entre outros.	REN et al., 2004,2007	http://www.membranetransport.org/

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

Filogenômicos			
BPhyOG	*Bacterial Phylogenies Based on Overlapping Genes*. Servidor *web* interativo destinado à reconstrução de filogenias de genomas bacterianos completamente seqüenciados, com base no conteúdo de genes com sobreposição compartilhados entre as espécies analisadas.	LUO et al., 2007	http://cmb.bnu.edu.cn/BPhyOG/
PHOG	*Phylogenetic Orthologous Groups*. Banco de dados de genes homólogos entre dezenas de espécies diferentes, incluindo procariotos e eucariotos, construído de forma automática a partir do conteúdo protéico predito nestes genomas e de forma orientada por cada nó da árvore taxonômica que representa este grupo de espécies, ou seja, através de uma abordagem evolutiva criteriosa.	MERKEEV et al., 2006	http://bioinf.fbb.msu.ru/phogs/index.html
Phydbac	*Phylogenomic Display of Bacterial Genes*. Oferece visualização e comparação interativas de perfis filogenéticos de seqüências protéicas de centenas de bactérias, permitindo a detecção de proteínas funcionalmente relacionadas entre si e padrões de conservação entre diversos organismos.	ENAULT et al., 2004	http://igs-server.cnrs-mrs.fr/phydbac/
SHOT	Sistema desenvolvido para a reconstrução de filogenias genômicas. Oferece construção de árvores filogenéticas para centenas de organismos cujos genomas foram completamente seqüenciados, com base no conteúdo de genes compartilhados ou na conservação da ordem dos genes entre os genomas dos organismos selecionados.	KORBEL et al., 2003	http://www.Bork.EMBL-Heidelberg.de/SHOT
Metadados genômicos			
Genome Properties	Sistema desenvolvido para pesquisa do conteúdo genético de procariotos, com aplicação em microbiologia, anotação de genomas e genômica comparativa. Buscas e comparações podem ser executadas com base em numerosos atributos de genomas procarióticos cujos estados podem ser descritos por valores numéricos ou por termos pertencentes a um vocabulário controlado.	HAFT et al., 2005; SELENGUT et al., 2007	http://www.tigr.org/Genome_Properties/
GenomeMine	Banco de dados que integra informações gerais sobre todos os genomas completamente seqüenciados. As informações são obtidas de diversas fontes, incluindo os bancos de dados *Genome* (NCBI) e GOLD (*Genomes Online Database*), ou computadas a partir das seqüências genômicas. Comparações podem ser executadas com base em numerosos atributos dos genomas.	-	http://www.genomics.ceh.ac.uk/GMINE/
SACSO	*Systematic Analysis of Completely Sequenced Organisms*. Banco de dados que consiste na análise comparativa entre organismos cujos genomas foram completamente seqüenciados. Inclui composição de bases e de aminoácidos, duplicação e conservação ancestrais e classificação dos organismos, obtidas a partir da comparação do proteoma predito destes organismos, com uso de análise de correspondência para sintetizar estas informações.	TEKAIA et al., 2002	http://www.pasteur.fr/~tekaia/sacso.html
FERRAMENTAS COMPUTACIONAIS			
Navegação interativa de genomas			
ABC	*Application for Browsing Constraints*. Programa para exploração interativa de dados de alinhamentos múltiplos de seqüências genômicas. Permite a visualização simultânea de diversos dados quantitativos (por exemplo, similaridade de seqüência e taxas evolutivas) e de anotação (localização dos genes, repetições, entre outros).	COOPER et al., 2004	http://mendel.stanford.edu/sidowlab/downloads.html

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

ACT	*Artemis Comparison Tool*. Permite a visualização interativa de comparações entre seqüências genômicas e suas anotações. As comparações podem ser geradas com diferentes programas de alinhamento, possibilitando a identificação de regiões sintênicas, inversões e rearranjos.	CARVER et al., 2005	http://www.sanger.ac.uk/Software/ACT/
AutoGRAPH	Servidor *web* interativo para análises comparativas entre genomas de múltiplas espécies, a partir de dados fornecidos pelo próprio usuário ou a partir de dados públicos pré-computados. O programa destina-se à construção e visualização de mapas de sintenia entre duas ou três espécies, à determinação e representação de relações de macro e micro sintenia entre as mesmas e à evidência de regiões de ruptura (*breakpoints*), facilitando a identificação de rearranjos cromossômicos.	DERRIEN et al., 2007	http://genoweb.univ-rennes1.fr/tom_dog/AutoGRAPH/
CGAT	*Comparative Genome Analysis Tool*. Programa para visualização interativa e comparação de pares de genomas alinhados, juntamente com suas anotações. O programa oferece uma estrutura genérica para processar alinhamentos em escala genômica com uso de vários programas de alinhamento já existentes e a visualização perpendicular (*dot plot*) ou horizontal (linhas paralelas) dos dados.	UCHIYAMA et al., 2006	http://mbgd.genome.ad.jp/CGAT/
Cinteny	Servidor para a identificação de sintenia e análise de rearranjos genômicos em dados pré-computados ou fornecidos pelo próprio usuário. O programa permite a comparação automática de pares de genomas e executa análises para detecção de blocos de sintenia e para o subseqüente cálculo de distâncias reversas.	SINHA e MELLER, 2007	http://cinteny.cchmc.org/
ComBo	*Comparative Genome Browser*. Programa para visualização interativa e comparação de pares de genomas alinhados, juntamente com suas anotações. O programa aceita alinhamentos e anotações em diferentes formatos e oferece visualização perpendicular (*dot plot*) ou horizontal (linhas paralelas) dos dados.	ENGELS et al., 2006	http://www.broad.mit.edu/annotation/argo/
DNAVis	Pacote de programas que oferece visualização interativa de anotações genômicas de forma comparativa.	FIERS et al., 2006	http://www.win.tue.nl/dnavis/
GECO	Programa desenvolvido para visualização linear de múltiplos genomas procarióticos, que permite a detecção de transferência horizontal de genes, pseudogenes e eventos de inserção/deleção em espécies relacionadas. É capaz de evidenciar relações de ortologia, estabelecidas com o algoritmo implementado no programa BLASTCLUST que faz parte do pacote de programas NCBI BLAST, e identificar irregularidades ao nível genômico através de anomalias no conteúdo GC.	KUENNE et al., 2007	http://bioinfo.mikrobio.med.uni-giessen.de/geco2/GecoMainServlet
GenColors	Programa desenvolvido para melhorar e acelerar a anotação de genomas procarióticos, através do uso de informações disponíveis sobre genomas relacionados que já foram totalmente seqüenciados e do uso extensivo de comparação genômica. As ferramentas de comparação incluem detecção de melhores *hits* bidirecionais, análise de conservação gênica e sintenia, entre outros.	ROMUALDI et al., 2005	http://gencolors.imb-jena.de
GeneOrder3.0	Programa para comparação da ordem dos genes e sintenia em pares de genomas bacterianos pequenos.	CELAMKOTI et al., 2004	http://binf.gmu.edu/genometools.html

cont.

Tabela 2 – Principais bancos de dados e ferramentas computacionais disponíveis para a análise comparativa de genomas procarióticos (cont.)

GenomeViz	Ferramenta para a visualização interativa e comparação de múltiplos genomas ou seqüências genômicas a partir de diversas fontes de informação qualitativa e quantitativa derivadas de estudos de anotação/classificação de genes, conteúdo GC, ilhas genômicas, *microarrays*, entre outros.	GHAI et al., 2004	http://www.uniklinikum-giessen.de/genome/genomeviz/intro.html
G-InforBIO	Sistema integrado para genômica microbiana. Permite a importação de dados genômicos (anotações e seqüências) de diferentes fontes e formatos, criando um banco de dados local com estas informações. Oferece diversas opções de busca e obtenção de dados, exportação de dados, e ferramentas para visualização e análises comparativas, através de uma interface gráfica amigável.	TANAKA et al., 2006	http://rhodem17.ddbj.nig.ac.jp/inforbio/
inGeno	Sistema integrado para visualização de ortólogos e comparação de pares de genomas. Permite a visualização interativa de comparações entre seqüências genômicas e suas anotações. As comparações podem ser geradas com diferentes programas de alinhamento, possibilitando a identificação de regiões sintênicas, inversões e rearranjos.	LIANG e DANDEKAR, 2006	http://ingeno.bioapps.biozentrum.uni-wuerzburg.de/
MuGeN	Programa para a exploração visual interativa de múltiplos segmentos genômicos anotados. Aceita diversos tipos de formatos de anotação, além de informações personalizadas, fornecidas pelo usuário.	HOEBEKE et al., 2003	http://genome.jouy.inra.fr/MuGeN/
SynBrowse	*Synteny Browser for comparative sequence analysis*. Programa para visualização e análise comparativa de genomas alinhados. Possibilita a identificação de seqüências conservadas, regiões sintênicas, inversões e rearranjos.	PANE et al., 2005	http://www.synbrowse.org/
SynView	Programa interativo e personalizável para visualização e análise comparativa de múltiplos genomas. Possibilita a identificação de seqüências conservadas, regiões sintênicas, inversões e rearranjos.	WANG et al., 2006	http://www.ApiDB.org/apps/SynView/
	Comparação de seqüências genômicas em larga escala		
BioParser	Programa que oferece um conjunto de interfaces gráficas amigáveis para manipulação e análise de dados obtidos com alinhamentos locais entre seqüências em larga escala. As comparações podem ser obtidas com diversos programas de alinhamento local. Permite que pares ou grupos de seqüências sejam selecionados dinamicamente, com base em múltiplos critérios estabelecidos pelo usuário (parâmetros calculados de similaridade, anotação, tamanho da seqüência, entre outros).	CATANHO et al., 2006	http://www.dbbm.fiocruz.br/BioParser.html
BSR	*The BLAST Score Ratio Analysis Tool*. Permite a visualização do grau de similaridade entre o proteoma predito em 3 genomas diferentes (incluindo sintenia), com base em uma classificação obtida através de algoritmo próprio (*BLAST Score Ratio*).	RASKO et al., 2005	http://www.microbialgenomics.org/BSR/
COMPAM	Programa para visualização e comparação de múltiplos genomas, baseado na combinação de todos os alinhamentos par a par dos genomas estudados.	LEE et al., 2006	http://bio.informatics.indiana.edu/projects/compam/
GenomeBlast	Programa disponível via *web* para a análise comparativa de múltiplos genomas de tamanho pequeno, a partir de dados fornecidos pelo próprio usuário. A ferramenta permite a identificação de genes únicos e genes homólogos, visualização da distribuição dos mesmos entre os genomas comparados e reconstrução filogenética em nível genômico.	LU et al., 2006	http://bioinfo-srv1.awh.unomaha.edu/genomeblast/

cont.

GenomeComp	Ferramenta para manipulação e comparação visual de dados obtidos com alinhamentos locais (BLAST somente) entre seqüências genômicas de múltiplos organismos em larga escala. Permite a detecção de repetições, inversões, deleções e rearranjos de segmentos genômicos.	YANG et al., 2003	http://www.mgc.ac.cn/GenomeComp/
GenomePixelizer	Ferramenta de visualização genômica que gera imagens personalizadas a partir de coordenadas físicas ou genéticas de grupos de genes específicos em segmentos genômicos ou genomas inteiros e das matrizes de similaridade destas seqüências, permitindo a detecção de ortólogos e parálogos.	KOZIK et al., 2002	http://www.atgc.org/GenomePixelizer/
M-GCAT	*Multiple Genome Comparison and Alignment Tool*. Programa para alinhamento múltiplo e visualização de genomas inteiros, ou grandes segmentos de ADN, de forma interativa e computacionalmente rápida e eficiente, através de algoritmo próprio.	TREANGEN e MESSEGUER, 2006	http://alggen.lsi.upc.es/recerca/align/mgcat/intro-mgcat.html
MUMmer	Sistema para alinhamento múltiplo e visualização de genomas inteiros, ou grandes segmentos de ADN, de forma computacionalmente rápida e eficiente, através de algoritmo próprio (*Space efficent suffix trees*).	KURTZ et al., 2004	http://www.tigr.org/software/mummer/
PipMaker, PipTools, MultiPipMaker, zPicture	Conjunto de ferramentas para alinhamento e visualização, em diversos formatos, de segmentos genômicos ou genomas inteiros. Permite a geração de perfis de conservação e identificação de regiões evolutivamente conservadas de forma dinâmica.	SCHWARTZ et al., 2000 2003a; ELNITSKI et al., 2002; OVCHARENKO et al., 2004	http://bio.cse.psu.edu/
PyPhy	Conjunto de ferramentas para a reconstrução automática e em larga escala de relações filogenéticas entre genomas microbianos completamente seqüenciados.	Sicheritz-Ponten & Andersson 2001	http://www.cbs.dtu.dk/staff/thomas/pyphy/
VISTA	Conjunto de ferramentas computacionais para genômica comparativa. Oferece algoritmos para alinhamento de grandes segmentos genômicos e visualização destes alinhamentos, com suas respectivas anotações funcionais.	Frazer et al. 2004; Brudno et al. 2007	http://www-gsd.lbl.gov/vista/

De uma forma geral, os bancos de dados que permitem análises comparativas de genomas procarióticos podem ser divididos em cinco categorias principais, segundo seus propósitos e funcionalidades: (i) genéricos e multifuncionais; (ii) organismos ou grupos – específicos; (iii) especializados; (iv) filogenômicos; e de (v) **metadados genômicos** (Tabela 2). As ferramentas computacionais, por sua vez, podem ser agrupadas em (i) programas para navegação interativa de genomas e (ii) programas que utilizam comparação de seqüências genômicas em larga escala (Tabela 2). Entretanto, é importante lembrar que esta classificação não é, sob nenhuma circunstância, definitiva ou quiçá a mais adequada, devido ao grande número de sobreposições existente entre os propósitos e funcionalidades destes bancos e ferramentas. Portanto, outras formas de classificação são possíveis e igualmente válidas (FIELD et al., 2005; GALPERIN, 2005).

Os bancos de dados genéricos e multifuncionais, em sua grande maioria, se propõem a abranger o universo de espécies procarióticas (e em alguns casos eucarióticas também) cujos genomas foram completamente seqüenciados e a oferecer os meios necessários para a busca e obtenção de dados pré-computados (na maior parte das vezes) e/ou obtidos experimentalmente (pelos próprios desenvolvedores ou compilados de outras fontes) para cada espécie (BacMap, CMR, Genome Atlas, IMG, MBGD, Microbes Online, PLATCOM, PUMA2). Os dados disponíveis variam bastante de um banco para outro, podendo compreender propriedades/atributos físico-químicos, estruturais, estatísticos, funcionais, evolutivos, taxonômicos, fenotípicos, entre outros, associados aos genomas inteiros ou às regiões codificantes e/ou não codificantes nestes genomas (Figura 1). As ferramentas de análise e de consulta oferecidas por estes e pelos demais bancos de dados também variam consideravelmente, podendo incluir busca por palavras-chaves, por nome/identificador do gene/região codificante e/ou espécie, comparação entre genomas inteiros, seqüências genômicas ou regiões codificantes através de algoritmos de alinhamento local ou global, entre outros. Igualmente, tudo isto se aplica àqueles bancos classificados como organismos ou grupos – específicos (GenoList, GenoMycDB, LEGER, MolliGen, ShiBASE, xBASE), com a diferença de os mesmos dedicarem-se à análise de genomas microbianos particulares, individualmente ou em conjunto.

Por outro lado, há um número crescente de bancos de dados dedicados à análise comparativa de característi-

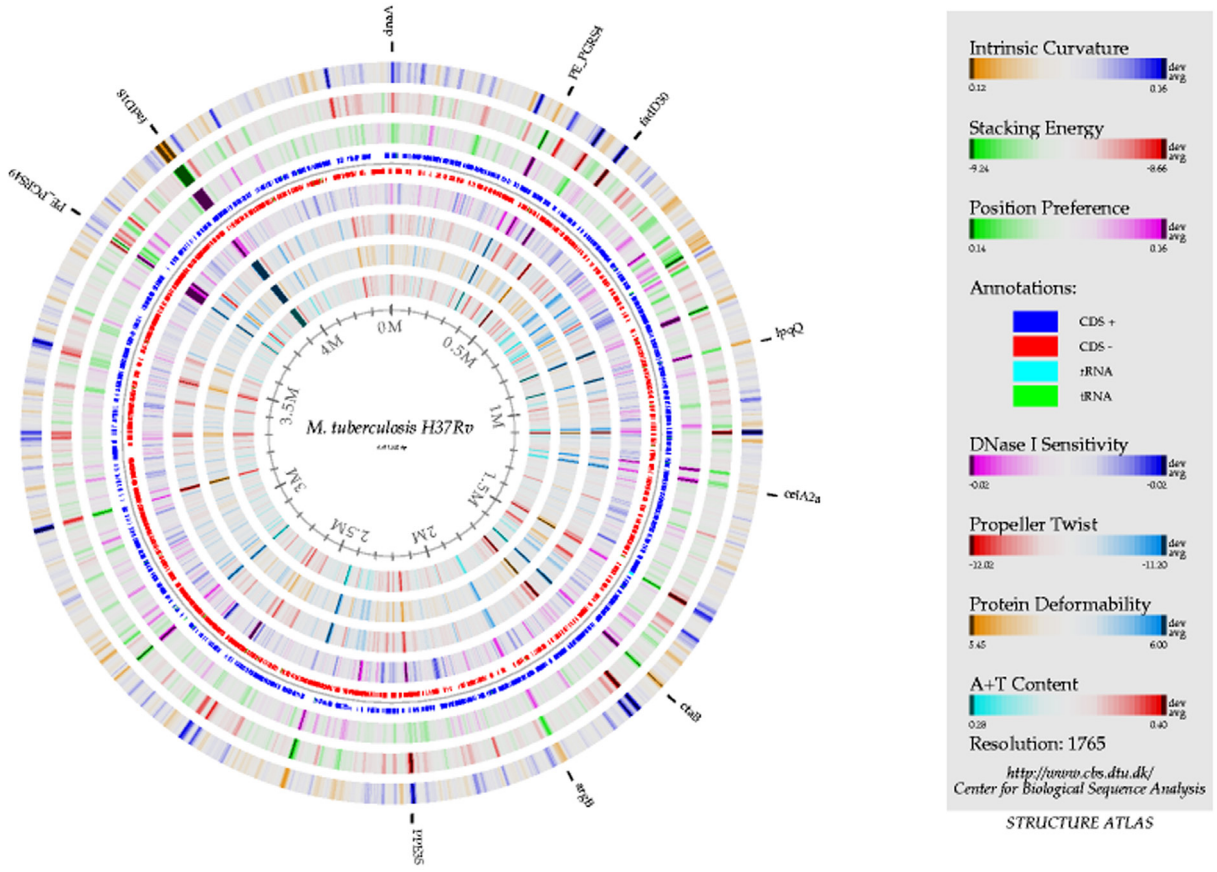


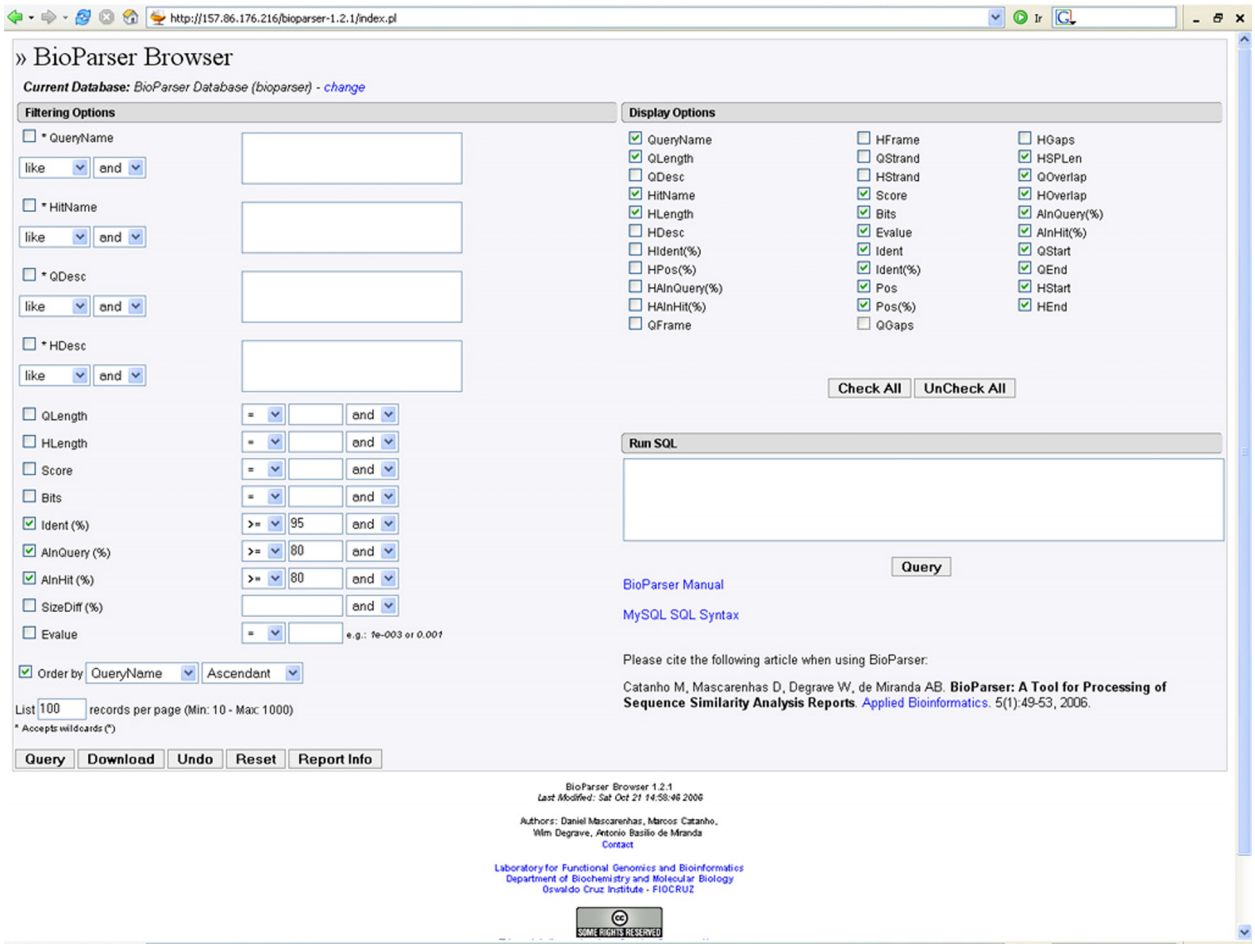
Figura 1 – Atlas estrutural do genoma de *Mycobacterium tuberculosis* H37Rv. Os círculos concêntricos representam sete diferentes características estruturais das moléculas de ADN (ver legenda na própria figura). O quarto e quinto círculos, do círculo mais externo em direção ao centro, representam a distribuição das regiões codificantes anotadas na fita de ADN (fita positiva, em azul, e fita negativa, em vermelho, respectivamente) e a distribuição de regiões que codificam ARN ribossômico (azul-claro) e transportador (verde) no genoma. Os valores de cada parâmetro estrutural medido são representados por escalas de cores, permitindo a visualização de sua variação ao longo do genoma. Mapas semelhantes a este, representando estas e outras características biológicas, podem ser facilmente obtidos (ou gerados a partir de dados fornecidos pelo próprio usuário) no banco de dados *Genome Atlas Database* (GenomeAtlas 2007) e, posteriormente, comparados de forma visual. Explicações detalhadas sobre os parâmetros estruturais calculados e a importância de cada um deles podem ser encontradas no próprio site do banco *Genome Atlas*.

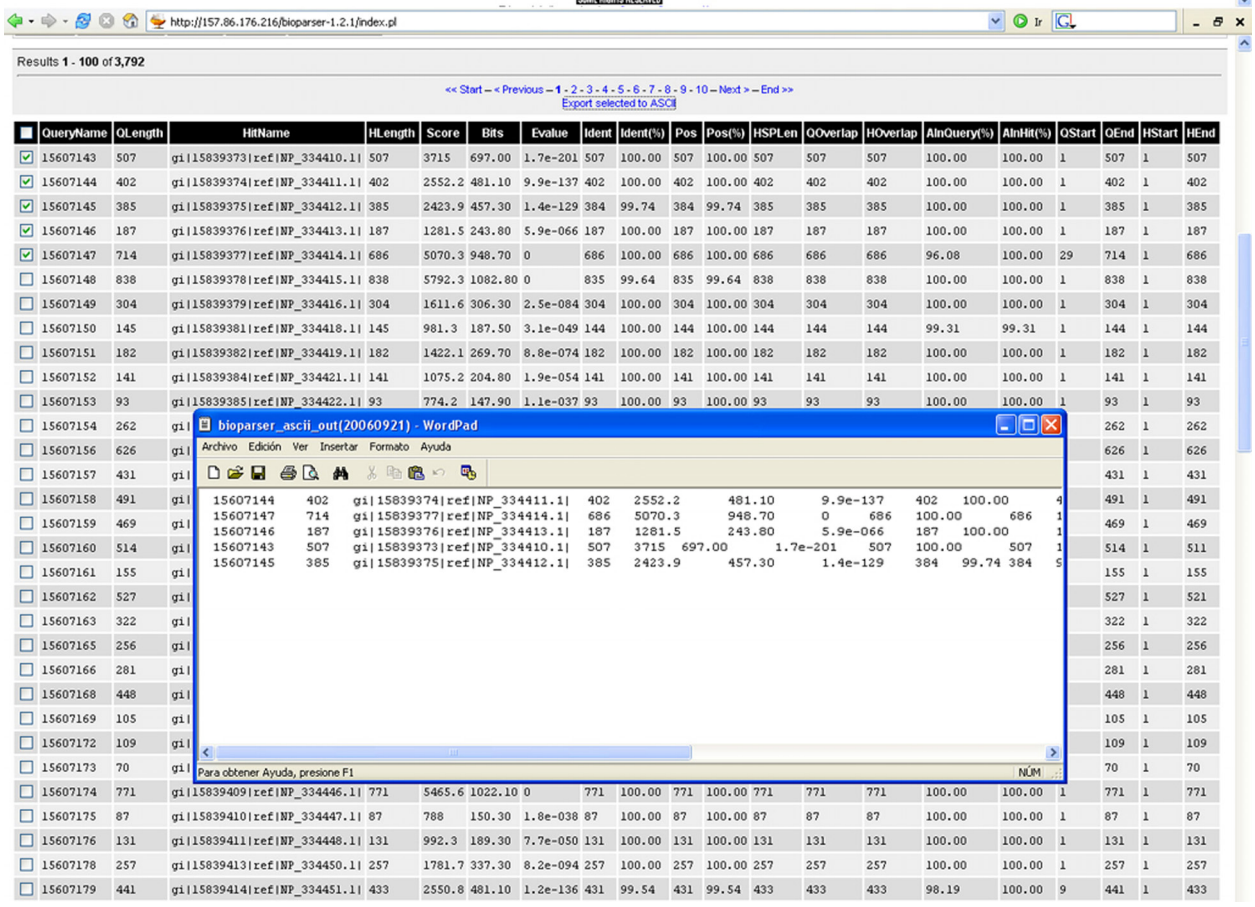
cas particulares associadas aos genomas e seus componentes. Entre as características exploradas por estes bancos especializados, destacam-se a conservação de genes (ou proteínas) ortólogos (COG, HAMAP, Hogenom, OMA Browser, OrthoMCL-DB, RoundUp); eventos de fusão gênica (FusionDB); ocorrência de ilhas genômicas (IslandPath); presença de repetições de aminoácidos em proteínas (ProtRepeatsDB); ocorrência e classificação de genes órfãos (ORFanage, OrphanMine) ou de grupos funcionais, como genes pertencentes a subsistemas celulares (SEED) ou ainda proteínas transportadoras de membrana (TransportDB); formação de redes de interação entre proteínas (STRING); ocorrência e conservação de vias bioquímicas (KEGG, MetaCyc).

Nos últimos anos, com o desenvolvimento de métodos filogenéticos que empregam não apenas genes marcadores, mas sim todo o conteúdo gênico de genomas completamente seqüenciados, surgiram bancos de dados denominados filogenômicos, os quais permitem a visualização e comparação de perfis filogenéticos (Phydbac), a reconstrução de filogenias com base no conteúdo gênico compartilhado (BPhyOG, SHOT) ou na conservação da ordem dos genes nos genomas (SHOT), ou ainda a análise de grupos de proteínas ortólogas entre inúmeras espécies, construídos de forma orientada pela classificação taxonômica destes organismos (PHOG).

Também recentemente, bancos de dados dedicados à comparação de metadados genômicos têm sido desenvolvidos através da análise de informações associadas aos genomas e grupos particulares de genes de centenas de espécies microbianas e também, em parte, através de informações compiladas de trabalhos científicos já publicados, permitindo que relações entre o estilo de vida, a história evolutiva e características genômicas possam ser exploradas (Genome Properties, GenomeMine, SACSO).

No que se refere às ferramentas computacionais desenvolvidas para a análise comparativa de genomas a maior





	QueryName	QLength	HitName	HLength	Score	Bits	Evalue	Ident	Ident(%)	Pos	Pos(%)	HSPLen	QOverlap	HOverlap	AlnQuery(%)	AlnHit(%)	QStart	QEnd	HStart	HEnd
☑	15607143	507	gi|15839373|ref|NP_334410.1|	507	3715	697.00	1.7e-201	507	100.00	507	100.00	507	507	507	100.00	100.00	1	507	1	507
☑	15607144	402	gi|15839374|ref|NP_334411.1|	402	2552.2	481.10	9.9e-137	402	100.00	402	100.00	402	402	402	100.00	100.00	1	402	1	402
☑	15607145	385	gi|15839375|ref|NP_334412.1|	385	2423.9	457.30	1.4e-129	384	99.74	384	99.74	385	385	385	100.00	100.00	1	385	1	385
☑	15607146	187	gi|15839376|ref|NP_334413.1|	187	1281.5	243.80	5.9e-066	187	100.00	187	100.00	187	187	187	100.00	100.00	1	187	1	187
☑	15607147	714	gi|15839377|ref|NP_334414.1|	686	5070.3	948.70	0	686	100.00	686	100.00	686	686	686	96.08	100.00	29	714	1	686
☐	15607148	838	gi|15839378|ref|NP_334415.1|	838	5792.3	1082.80	0	835	99.64	835	99.64	838	838	838	100.00	100.00	1	838	1	838
☐	15607149	304	gi|15839379|ref|NP_334416.1|	304	1611.6	306.30	2.5e-084	304	100.00	304	100.00	304	304	304	100.00	100.00	1	304	1	304
☐	15607150	145	gi|15839381|ref|NP_334418.1|	145	981.3	187.50	3.1e-049	144	100.00	144	100.00	144	144	144	99.31	99.31	1	144	1	144
☐	15607151	182	gi|15839382|ref|NP_334419.1|	182	1422.1	269.70	8.8e-074	182	100.00	182	100.00	182	182	182	100.00	100.00	1	182	1	182
☐	15607152	141	gi|15839384|ref|NP_334421.1|	141	1075.2	204.80	1.9e-054	141	100.00	141	100.00	141	141	141	100.00	100.00	1	141	1	141
☐	15607153	93	gi|15839385|ref|NP_334422.1|	93	774.2	147.90	1.1e-037	93	100.00	93	100.00	93	93	93	100.00	100.00	1	93	1	93
☐	15607154	262	gi|																262	1	262
☐	15607156	626	gi|																626	1	626
☐	15607157	431	gi|																431	1	431
☐	15607158	491	gi|																491	1	491
☐	15607159	469	gi|																469	1	469
☐	15607160	514	gi|																514	1	511
☐	15607161	155	gi|																155	1	155
☐	15607162	527	gi|																527	1	521
☐	15607163	322	gi|																322	1	322
☐	15607165	256	gi|																256	1	256
☐	15607166	281	gi|																281	1	281
☐	15607168	448	gi|																448	1	448
☐	15607169	105	gi|																105	1	105
☐	15607172	109	gi|																109	1	109
☐	15607173	70	gi|																70	1	70
☐	15607174	771	gi|15839409|ref|NP_334446.1|	771	5465.6	1022.10	0	771	100.00	771	100.00	771	771	771	100.00	100.00	1	771	1	771
☐	15607175	87	gi|15839410|ref|NP_334447.1|	87	788	150.30	1.8e-038	87	100.00	87	100.00	87	87	87	100.00	100.00	1	87	1	87
☐	15607176	131	gi|15839411|ref|NP_334448.1|	131	992.3	189.30	7.7e-050	131	100.00	131	100.00	131	131	131	100.00	100.00	1	131	1	131
☐	15607178	257	gi|15839413|ref|NP_334450.1|	257	1781.7	337.30	8.2e-094	257	100.00	257	100.00	257	257	257	100.00	100.00	1	257	1	257
☐	15607179	441	gi|15839414|ref|NP_334451.1|	433	2550.8	481.10	1.2e-136	431	99.54	431	99.54	433	433	433	98.19	100.00	9	441	1	433

Figura 2: Comparação entre o conteúdo protéico total codificado nos genomas de duas cepas de *M. tuberculosis*, H37Rv e CDC1551, através de uma versão para uso local da ferramenta BioParser. As proteínas preditas nos genomas destas micobactérias foram obtidas no banco de dados Reference Sequence (REFSEQ, 2007) (número de acesso NC_000962 e NC_002755, respectivamente), e foram comparadas localmente, todas contra todas, usando o programa FASTA de busca por similaridade (UVA FASTA SERVER, 2007). O arquivo resultante desta comparação foi processado com o BioParser e as informações obtidas foram inseridas automaticamente em um banco de dados local, criado e configurado de acordo com as instruções fornecidas no manual do programa. Em seguida, foi elaborada uma consulta neste banco através da interface gráfica de acesso oferecida, o BioParser Browser, que consistiu em retornar somente os pares alinhados cujo percentual de posições idênticas no alinhamento é maior ou igual a 95% e cuja fração percentual do tamanho de ambas as seqüências no par alinhado é maior ou igual a 80% (parte superior da figura). Apenas algumas das opções de formatação do resultado oferecidas foram selecionadas (*Display Options*) e somente parte dos 100 primeiros resultados obtidos de um total de 3.792 pares alinhados que satisfizeram às condições impostas na consulta são mostrados, ordenados pelos nomes das seqüências de entrada usadas na busca por similaridade (parte inferior da figura). Os cinco primeiros pares alinhados foram selecionados e exportados para um arquivo texto, através da ferramenta *Export selected to ASCII*. Para arquivos resultantes de buscas por similaridade de seqüência com até 5 megabytes, o processamento e análise podem ser feitos remotamente através de um servidor *web* (BIOPARSERWEB, 2007). Detalhes sobre a construção, aplicações, uso e instalação local da ferramenta podem ser encontradas na página do programa (BIOPARSER, 2007) e no artigo no qual ela é descrita (CATANHO et al., 2006).

parte dedica-se a visualização/navegação interativa e comparativa de pares (ATC, Cinteny, DNAVis, GeneOrder3.0, G-InforBIO, inGeno, SynBrowse) ou grupos (AutoGRAPH, GECO, GenColors, GenomeViz, MuGeN, SynView) de genomas ou seqüências genômicas em diferentes ambientes gráficos, ou ainda a exploração interativa de dados de alinhamentos múltiplos de seqüências genômicas (ABC, CGAT, ComBo). Outro grupo de ferramentas baseia-se em comparações de seqüências em larga escala entre múltiplos genomas, através do uso de algoritmos de alinhamento local (BioParser, BSR, COMPAM, GenomeBlast, GenomeComp) (Figura 2) ou global (M-GCAT, MUMmer, PipMaker/PipTools/MultiPipMaker/zPicture, VISTA, PyPhy), ou ainda a partir de coordenadas físicas ou genéticas de grupos de genes específicos em segmentos genômicos ou genomas inteiros e das matrizes de similaridade destas seqüências (GenomePixelizer). Similarmente aos bancos de dados, as opções de busca, obtenção de dados e análise oferecidas por estas ferramentas são extremamente variáveis, havendo sobreposições em muitos casos. Entre elas, destacam-se: busca por palavra-chave, nome/identificador do gene/região codificante e/ou espécie; obtenção das anotações funcionais dos genes descritos; reconstrução filogenética; detecção de colinearidade, sintenia, duplicação gênica, grupos de genes ortólogos e parálogos, rearranjos, repetições, inversões, inserções, deleções, sítios de restrição, motivos e perfis, entre outros. Estas ferramentas encontram-se disponíveis como serviços *on-line* e/ou programas independentes para uso local (*stand-alone applications*).

Pensando o amanhã: conclusões e perspectivas para o futuro

Como foi visto, a análise comparativa de genomas possui variadas aplicações em diferentes campos do conhecimento, desde a análise da estrutura, organização e evolução dos genomas até o desenvolvimento de métodos mais eficientes de prevenção, tratamento e diagnóstico de doenças parasitárias, por exemplo. Vimos também que esta abordagem holística se serve de dados obtidos com o desenvolvimento e aplicação de tecnologias de alto desempenho como a genômica, a proteômica e a transcriptômica, e que os métodos, algoritmos e ferramentas empregados neste tipo de abordagem têm suas raízes no surgimento e consolidação de ciências como a Computação, a Bioinformática e a Biologia Computacional. Entretanto, apesar de toda a sua relevância científica, a comparação maciça de dados genômicos traz consigo uma gama de desafios técnicos e científicos importantes, tais como capacidade de armazenamento de dados, estrutura e representação adequada dos mesmos, facilidade de acesso e manipulação destes dados pelo usuário, velocidade de processamento, diferentes formatos de arquivos e integração de múltiplas ferramentas.

Numerosos bancos de dados e ferramentas computacionais têm sido desenvolvidos para permitir o acesso de toda a comunidade científica aos diferentes dados genômicos disponíveis, bem como a análise comparativa dos mesmos. Variadas opções de visualização, busca, obtenção e análise destes dados são oferecidas, permitindo a aquisição de conhecimento cada vez mais detalhado sobre os genomas e seus respectivos organismos. No entanto, todo esse conhecimento encontra-se fragmentado, disperso através de todos estes recursos computacionais, muitas vezes de forma redundante, necessitando ser unificado, de tal forma que nós possamos ter uma visão integrada e global da biologia de todos estes genomas e espécies estudados. Idealmente, as bases de dados e as ferramentas computacionais futuras deveriam oferecer informações integradas, permitindo a análise de genomas sob múltiplas perspectivas; combinar dados obtidos *in silico* com dados curados, ampliando a qualidade de nossos estudos; ter estrutura, armazenamento e processamento de dados eficiente, possibilitando visualização, busca, obtenção e análise de dados de maneira dinâmica, flexível e rápida, através de uma interface gráfica amigável para o usuário; descrever os dados através de um vocabulário controlado e disponibilizá-los em arquivos com formatos padronizados, proporcionando intercâmbio e integração plena da informação entre si e com outras fontes de dados. Dessa forma, abrir-se-ia um campo fértil para interações e colaborações amplas entre pesquisadores de diferentes áreas, necessárias à interpretação e análise dessa imensa e variada quantidade de dados, através de uma abordagem verdadeiramente multidisciplinar.

Glossário

Algoritmo. Procedimento organizado (passos e instruções) para executar um determinado tipo de cálculo ou solucionar um determinado tipo de problema.

Alinhamento de seqüências. Processo de alinhar (colocar lado a lado) duas ou mais seqüências do mesmo tipo (nucleotídicas ou protéicas) de forma a obter o máximo de identidade entre elas com o propósito de determinar o grau de similaridade.

Alinhamento global. Alinhamento de pares de seqüências nucleotídicas ou protéicas ao longo de toda a extensão das mesmas.

Alinhamento local. Alinhamento de uma ou mais partes de duas seqüências nucleotídicas ou protéicas.

Banco de dados relacional. Sistema de banco de dados no qual a base de dados é organizada e acessada de acordo com o relacionamento existente entre os itens que compõem a base. O relacionamento entre estes itens é expresso através de tabelas.

Biochip. *Microarrays* de proteínas. Quantidades maciças de diferentes agentes de captura, freqüentemente anticorpos monoclonais, depositados sobre a superfície de uma matriz sólida de vidro ou silício em miniatura (*e.g.* lâmina de microscópio), usados para determinar a presença e/ou quantidade de proteínas em amostras biológicas.

Bioinformática e **Biologia Computacional**. Em 17 de julho de 2000, o *National Institutes of Health* (NIH), uma das agências do departamento de saúde norte-americano com reconhecimento internacional na área de pesquisa médica, divulgou sua definição de trabalho para Bioinformática e para Biologia Computacional, elaborada pelo *Biomedical Information Science and Technology Initiative Consortium* (BISTIC) *Definition Committee*. De acordo com este documento

> "A bioinformática e a biologia computacional têm suas raízes nas ciências da vida bem como nas ciências da computação e informação e na tecnologia. Ambas estas abordagens interdisciplinares se beneficiam de disciplinas específicas, tais como a matemática, a física, as ciências da computação e a engenharia, a biologia e as ciências do comportamento. Cada uma delas mantém interações muito estreitas com as ciências da vida para concretizar todo o seu potencial. A bioinformática aplica princípios das ciências da informação e da tecnologia para tornar os vastos, diversificados e complexos dados produzidos pelas ciências da vida mais compreensíveis e úteis. A biologia computacional usa abordagens matemáticas e computacionais para resolver questões teóricas e experimentais na biologia. Embora a bioinformática e a biologia computacional sejam distintas, há significativa sobreposição e atividade em suas interfaces. (...) Bioinformática: pesquisa, desenvolvimento ou aplicação de ferramentas e abordagens computacionais para ampliar o uso de dados de origem biológica, médica, comportamental ou de saúde, incluindo adquirir, armazenar, organizar, arquivar, analisar ou visualizar tais dados. Biologia Computacional: desenvolvimento e aplicação de métodos analíticos e teóricos de dados e técnicas de modelagem matemática e simulação computacional para o estudo de sistemas biológicos, comportamentais e sociais." (BISTIC Definition Committee, 2000). [Tradução livre do autor].

Convergência. Processo que dá origem à analogia, ou seja, relação entre dois caracteres quaisquer que descendem (por convergência) de caracteres ancestrais não relacionados entre si (FITCH, 1970; 2000).

Fatores epigenéticos. Fatores responsáveis pelo controle temporal e espacial da atividade de todos os genes necessários para o desenvolvimento de um organismo complexo desde o zigoto até a fase adulta (citado por STROHMAN, 1997).

Fusão gênica. Foi observado que determinados pares de proteínas funcionalmente relacionadas entre si, presentes em certos organismos, têm homólogos em outros organismos fundidos em uma única cadeia protéica (MARCOTTE et al., 1999; ENRIGHT et al., 1999). O processo de formação destas proteínas é chamado de fusão gênica.

Genoma. Termo criado, em 1920, por Hans Winkler, professor de Botânica na Universidade de Hamburgo. Designa toda a informação hereditária de um organismo que está codificada no seu ADN (ou, em alguns vírus, no ARN). Isto inclui tanto os genes como as sequências não codificadoras (conhecidas como ADN-lixo).

Homologia. Relação entre dois caracteres (traços genéticos, estruturais ou funcionais de um organismo) quaisquer que descendem de um caractere ancestral comum, normalmente com divergência (Fitch 1970, 2000).

Matrizes de substituição. Matrizes que representam todas as possíveis trocas entre aminoácidos, nas quais um valor é atribuído a cada uma destas trocas. Estes valores são proporcionais à probabilidade de ocorrência de cada troca, tomando-se como base um determinado modelo evolutivo. PAM – **P**ercent **A**ccepted **M**utation (DAYHOFF et al., 1978). BLOSUM - **BLO**cks **SU**bstitution **M**atrix (HENIKOFF e HENIKOFF, 1992).

Metadados genômicos. Dados que descrevem ou resumem outros dados genômicos, ou seja, todas as informações que podem ser usadas para descrever seqüências genômicas, como por exemplo, conteúdo GC, número de regiões codificantes e tamanho do genoma, ou para descrever a espécie da qual elas se originam, como por exemplo, taxonomia, habitat e nível trófico (FIELD et al., 2005).

Micobactérias. O gênero *Mycobacterium* (familia *Mycobacteriaceae*, ordem *Actinomycetales*), um dos mais antigos e bem conhecidos gêneros de bactéria, foi introduzido por Lehmann e Neumann em 1896, para incluir os agentes causadores da hanseníase e da tuberculose, bactérias que haviam sido anteriormente classificadas como *Bacterium leprae* e *Bacterium tuberculosis*, respectivamente (Goodfellow & Minnikin, 1984). Os organismos pertencentes a este gênero são aeróbios, imóveis e não formam endósporos ou esporos; têm forma de bastonetes delgados, retos ou ligeiramente encurvados, com raras formas ramificadas. Seu ADN é rico em guanina (G) e citosina (C) (de 62 a 70% G+C, com exceção

de *Mycobacterium leprae* que tem 57.8% de GC). As micobactérias possuem ainda características peculiares como álcool-ácido resistência (uma vez coradas por corantes básicos, resistem à descoloração por soluções álcool-ácidas sendo, portanto, denominadas *bacilos álcool-ácido resistentes*) e resistência incomum à dessecação e a agentes químicos.

Microarrays. Também conhecidos como *DNA chips*. Quantidades maciças de moléculas de ADN clonadas depositados sobre a superfície de uma matriz sólida de vidro ou silício em miniatura (*e.g.* lâmina de microscópio), usadas em experimentos de hibridação molecular, com a finalidade de determinar padrões de expressão gênica ou a seqüência nucleotídica de moléculas de ADN ou ARN.

Motivos. Elemento (porção) conservado de um alinhamento de seqüências protéicas, normalmente correlacionado com uma função em particular.

Ortólogos. Genes homólogos em espécies diferentes originados de um gene ancestral comum, durante a especiação (FITCH, 1970; 2000).

Parálogos. Genes homólogos em uma espécie em particular originados por duplicação (FITCH 1970; 2000).

Perfis. Perfis de seqüências são tabelas que contêm as freqüências de cada aminoácido em cada posição de uma proteína. As freqüências são calculadas a partir de alinhamentos múltiplos de seqüências que contêm um domínio de interesse (GRIBSKOV et al., 1987).

Proteoma. Conjunto completo de proteínas expressas por uma célula, tecido ou organismo, em um dado momento e sob certas circunstâncias ambientais.

Regiões sintênicas. Sintenia foi um termo originalmente cunhado para designar a presença de dois ou mais *loci* gênicos (próximos ou não) no mesmo cromossomo. Atualmente, refere-se também a duas regiões de genomas distintos que mostram considerável grau de similaridade de seqüência entre si e algum grau de conservação da ordem dos genes nestas regiões e que, portanto, têm probabilidade de descender de um ancestral comum.

Transcriptoma. Conjunto de todos os ARN mensageiros (transcriptos) de uma célula, tecido ou organismo, em um dado momento e sob certas circunstâncias ambientais.

Referências bibliográficas

ALM, E.J. et al. The MicrobesOnline Web site for comparative genomics. **Genome Research**, v.15, n.7, p.1015-22, 2005.

ALTSCHUL S.F.et al.. Basic local alignment search tool. **Journal of Molecular Biology**, v.215, n.3, p.403-10, 1990.

ALTSCHUL S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, v.25, n.17, p.3389-402, 1997.

BARRE A., de DA; BLANCHARD, A. MolliGen, a database dedicated to the comparative genomics of Mollicutes. **Nucleic Acids Research**, v.32(Database issue), p.D307-D310, 2004.

BATZOGLOU, S. Human and mouse gene structure: comparative analysis and application to exon prediction. **Genome Research**, v.10, n.7, p.950-8, 2000.

BEHR, M.A. et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. **Science**, v.284, n.5419, p.1520-3, 1999.

BENSON, D.A. GenBank. **Nucleic Acids Research**, v.33(Database issue), p.D34-D38, 2005.

BINNEWIES, T.T. et al. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. **Functional & Integrative Genomics**, v.6, n.3, p.165-85, 2006.

BIOPARSER. Disponível em: <http://www.dbbm.fiocruz.br/BioParser> Acesso em: 8 out. 2007.

BIOPARSERWEB. Disponível em: <http://www.dbbm.fiocruz.br/BioParserWeb> Acesso em: 8 out. 2007.

BISTIC Definition Committee. NIH working definition of bioinformatics and computational biology. 2000. Disponível em: <http://www.bisti.nih.gov/CompuBioDef.pdf> Acesso em: 8 out. 2007.

BOECKMANN, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Research**, v.31, n.1, p.365-70, 2003.

BORK, P.; KOONIN, E.V. Predicting functions from protein sequences--where are the bottlenecks? **Nature Genetics**, v.18, n.4, p.313-8, 1998.

BRAY, N.; DUBCHAK, I.; PACHTER, L. AVID: A global alignment program. **Genome Research**, v.13, n.1, p.97-102, 2003.

BRAY, N.; PACHTER, L. MAVID: constrained ancestral alignment of multiple sequences. **Genome Research**, v.14, n.4, p.693-9, 2004.

BROSCH, R. et al. The evolution of mycobacterial pathogenicity: clues from comparative genomics. **Trends Microbiol**, v.9, n.9, p.452-8, 2001.

BRUDNO, M. et al.. Fast and sensitive multiple alignment of large genomic sequences. **BMC Bioinformatics**, v.4, n.1, p.66, 2003a.

BRUDNO, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. **Genome Research**, v.13, n.4, p.721-31, 2003b.

BRUDNO, M. et al. Multiple whole genome alignments and novel biomedical applications at the VISTA portal. **Nucleic Acids Research**, v.35, p.W669-W674, 2007.

CARVER, T.J. et al. ACT: the Artemis Comparison Tool. **Bioinformatics**, v.21, n.16, p.3422-3, 2005.

CASPI, R. et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. **Nucleic Acids Research**, v.34, p.D511-D516, 2006.

CATANHO, M. et al. GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. **Genetic Molecular Research**, v.5, n.1, p.115-26, 2006.

CATANHO, M. et al. AB. BioParser: a tool for processing of sequence similarity analysis reports. **Applied Bioinformatics**, v.5, n.1, p.49-53, 2006.

CELAMKOTI S. et al. GeneOrder3.0: software for comparing the order of genes in pairs of small bacterial genomes. **BMC Bioinformatics**, v.5, p.1, p.52, 2004.

CHAUDHURI, R.R.; PALLEN, M.J. xBASE, a collection of online databases for bacterial comparative genomics. **Nucleic Acids Research**, v.34, p.D335-D337, 2006.

CHEN, F. et al. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. **Nucleic Acids Research**, v.34, p.D363-D368, 2006.

CHOI, K. et al. PLATCOM: a Platform for Computational Comparative Genomics. **Bioinformatics**, Mar 15, 2005.

CMR. Comprehensive Microbial Resource. Disponível em: <http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl> Acesso em: 8 out. 2007.

COENYE, T. et al. Towards a prokaryotic genomic taxonomy. **FEMS Microbiology Reviews**, v.29, n.2, p.147-67, 2005.

COLE, S.T. Comparative mycobacterial genomics as a tool for drug target and antigen discovery. **European Respiratory Journal**, v.36, Suppl., p.78s-86s, 2002.

COOPER, G.M.; SINGARAVELU, S.A.; SIDOW, A. ABC: software for interactive browsing of genomic multiple sequence alignment data. **BMC Bioinformatics**, v.5, n.1, p.192, 2004.

DAYHOFF, M.O.; SCHWARTZ, R.M.; ORCUTT, B.C. A model of evolutionary change in proteins. In: DAYHOFF, M.O. (ed.) **Atlas of Protein Sequence and Structure**. Washington DC: National Biomedical Research Foundation, 1978. v.5. Suppl.3. p.345-352.

DELCHER, A.L. et al. Alignment of whole genomes. **Nucleic Acids Research**, v.27, n.11, p.2369-76, 1999.

DELCHER, A.L. et al. Fast algorithms for large-scale genome alignment and comparison. **Nucleic Acids Research**, v.30, n.11, p.2478-83, 2002.

DELUCA, T.F. et al. Roundup: a multi-genome repository of orthologs and evolutionary distances. **Bioinformatics**, v.22, n.16, p.2044-6, 2006.

DERRIEN, T. et al. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. **Bioinformatics**, v.23, n.4, p.498-499, 2007.

DIETERICH G, et al.. LEGER: knowledge database and visualization tool for comparative genomics of pathogenic and non-pathogenic Listeria species. **Nucleic Acids Research**, v.34, p.D402-D406, 2006.

DUFAYARD, J.F. et al. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. **Bioinformatics**, v.21, n.11, p.2596-603, 2005.

ELNITSKI, L. et al. PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. **Genomics**, v.80, n.6, p.681-90, 2002.

ENAULT, F. et al. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. **Nucleic Acids Research**, v.32, p.W336-W339, 2004.

ENGELS, R. et al. Combo: a whole genome comparative browser. **Bioinformatics**, v.22, n.14, p.1782-3, 2006.

ENRIGHT, A.J. et al. Protein interaction maps for complete genomes based on gene fusion events. **Nature**, v.402, n.6757, p.86-90, 1999.

FANG, G, et al. Specialized microbial databases for inductive exploration of microbial genome sequences. **BMC Genomics**, v.6, n.1, p.14, 2005.

FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. **Journal of Molecular Evolution**, v.17, n.6, p.368-76, 1981.

FELSENSTEIN, J. PHYLIP -- Phylogeny Inference Package (Version 3.2). **Cladistics**, v.5, p.164-6, 1989.

FENG; D.F.; DOOLITTLE, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **Journal Molecular Evolution**, v.25, n.4, p.351-60, 1987.

FIELD, D.; FEIL, E.J.; WILSON, G.A. Databases and software for the comparison of prokaryotic genomes. **Microbiology**, v.151, n.Pt 7, p.2125-32, 2005.

FIERS, M.W. et al. DNAVis: interactive visualization of comparative genome annotations. **Bioinformatics**, v.22, n.3, p.354-5, 2006.

FITCH, W.M. Distinguishing homologous from analogous proteins. **Systematic Zoology**, v.19, n.2, p.99-113, 1970.

FITCH, W.M. Homology a personal view on some of the problems. **Trends in Genetics**, v.16, n.5, p.227-31, 2000.

FITZGERALD, J.R.; MUSSER, J.M. Evolutionary genomics of pathogenic bacteria. **Trends Microbiol**, v.9, n.11, p.547-53, 2001.

FRASER, C.M. et al. Comparative genomics and understanding of microbial biology. **Emerging Infectious Diseases**, v.6, n.5, p.505-12, 2000.

FRAZER, K.A. et al. VISTA: computational tools for comparative genomics. **Nucleic Acids Research**, v.32, p.W273-W279, 2004.

GALPERIN, M.Y. The Molecular Biology Database Collection: 2005 update. **Nucleic Acids Research**, v.33, p.D5-24, 2005.

GATTIKER, A. et al. Automated annotation of microbial proteomes in SWISS-PROT. **Comput Biol Chem**, v.27, n.1, p.49-58, 2003.

GENOMEATLAS. CBS Genome Atlas Database. Disponível em: <http://www.cbs.dtu.dk/services/GenomeAtlas/>. Acesso em: 8 out. 2007.

GENOME PROJECT. NCBI Entrez Genome Project Database. Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj>. Acesso em: 8 out. 2007.

GHAI, R.; HAIN, T.; CHAKRABORTY, T. GenomeViz: visualizing microbial genomes. BMC Bioinformatics, v.5, n.1, p.198, 2004.

GOLD. Genomes Online Database. Disponível em: <http://www.genomesonline.org/>. Acesso em: 8 out. 2007.

GOODFELLOW, M.; MINNIKIN, D.E. Circunscription of the genus. In: KUBICA, G.P.; WAYNE, L.G. (eds.) **The Mycobacteria: A Source Book**. New York: Marcel Dekker; 1984. p.1-24.

GORDON, S.V. et al. Royal Society of Tropical Medicine and Hygiene Meeting at Manson House, London, 18th January 2001. Pathogen genomes and human health. Mycobacterial genomics. **Transactions of the Royal Society of Tropical Medicine and Hygiene**, v.96, n.1, p.1-6, 2002.

GRIBSKOV, M.; MCLACHLAN, A.D.; EISENBERG, D. Profile analysis: detection of distantly related proteins. **Proceedings of National Academy of Science**, v.84, n.13, p.4355-8, 1987.

HAFT, D.H. et al. Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. **Bioinformatics**, v.21, n.3, p.293-306, 2005.

HAGEN, J.B. The origins of bioinformatics. **Nature Reviews Genetics**, v.1, n.3, p.231-6, 2000.

HALLIN, P.F.; USSERY, D.W. CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. **Bioinformatics**, v.20, n.18, p.3682-6, 2004.

HENIKOFF, S.; HENIKOFF, J.G. Amino acid substitution matrices from protein blocks. **Proceedings of National Academy of Science**, v.89, n.22, p.10915-9, 1992.

HENZ, S.R. et al. Whole-genome prokaryotic phylogeny. **Bioinformatics**, v.21, n.10, p.2329-35, 2005.

HGP. HUMAN GENOME PROGRAM (USA). U.S. Department of Energy. Genomics and Its Impact on Medicine and Society: A 2001 Primer; 2001.

HIGGINS, D.; TAYLOR, W.R. Bioinformatics sequence, structure, and databanks: a practical approach. Oxford: Oxford University Press, 2000.

HSIAO, W. et al.. IslandPath: aiding detection of genomic islands in prokaryotes. **Bioinformatics**, v.19, n.3, p.418-20, 2003.

HOEBEKE, M.; NICOLAS, P.; BESSIERES, P. MuGeN: simultaneous exploration of multiple genomes and computer analysis results. **Bioinformatics**, v.19, n.7, p.859-64, 2003.

JAREBORG, N.; BIRNEY, E.; DURBIN, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res, v.9, n.9, p.815-24, 1999.

KALITA, M.K. et al. ProtRepeatsDB: a database of amino acid repeats in genomes. BMC **Bioinformatics**, v.7, p.336, 2006.

KANEHISA, M. A database for post-genome analysis. **Trends Genet**, v.13, n.9, p.375-6, 1997.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, v.28, n.1, p.27-30, 2000.

KANEHISA, M. et al. From genomics to chemical genomics: new developments in KEGG. **Nucleic Acids Research**, v.34, p.D354-7, 2006.

KATO-MAEDA, M. et al. Comparing genomes within the species Mycobacterium tuberculosis. Genome Res, v.11, n.4, p.547-54, 2001.

KENT, W.J.; ZAHLER, A.M. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res, v.10, n.8, p.1115-25, 2000.

KONDRASHOV, A.S. Comparative genomics and evolutionary biology. **Current Opinion in Genetics and Development**, v.9, n.6, p.624-9, 1999.

KOONIN, E.V.; ARAVIND, L.; KONDRASHOV, A.S. The impact of comparative genomics on our understanding of evolution. Cell, v.101, n.6, p.573-6, 2000.

KORBEL, J.O. et al. SHOT: a web server for the construction of genome phylogenies. **Trends in Genetics**, v.18, n.3, p.158-62, 2002.

KOZIK, A.; KOCHETKOVA, E.; MICHELMORE, R. GenomePixelizer--a visualization program for comparative genomics within and between species. **Bioinformatics**, v.18, n.2, p.335-6, 2002.

KUENNE, C.T. et al. GECO--linear visualization for comparative genomics. Bioinformatics, v.23, n.1, p.125-126, 2007.

KUNIN, V. et al. Measuring genome conservation across taxa: divided strains and united kingdoms. **Nucleic Acids Research**, v.33, n.2, p.616-21, 2005a.

KUNIN, V. et al. The net of life: reconstructing the microbial phylogenetic network. **Genome Research**, v.15, n.7, p.954-9, 2005b.

KURTZ, S. et al. Versatile and open software for comparing large genomes. **Genome Biology**, v.5, n.2, R12, 2004.

LANDER, E.S. et al. Initial sequencing and analysis of the human genome. **Nature**, v.409, n.6822, p.860-921, 2001.

LEE D. et al. COMPAM: visualization of combining pairwise alignments for multiple genomes. **Bioinformatics**, v.22, n.2, p.242-4, 2006.

LIANG, C.; DANDEKAR, T.; inGeno--an integrated genome and ortholog viewer for improved genome to genome comparisons. **BMC Bioinformatics**, v.7, p.461, 2006.

LIPMAN, D.J.; PEARSON, W.R. Rapid and sensitive protein similarity searches. **Science**, v.227, p.4693, p.1435-41, 1985.

LU, G. et al. GenomeBlast: a web tool for small genome comparison. **BMC Bioinformatics**, v.7, Suppl 4, p.S18, 2006.

LUO, Y. et al. BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. **BMC Bioinformatics**, v.8, p.266, 2007.

MA, B.; TROMP, J.; LI, M. PatternHunter: faster and more sensitive homology search. **Bioinformatics**, v.18, n.3, p.440-5, 2002.

MACÊDO, J.A. et al. A Molecular Biology Conceptual Model for Information Integration. **Revista Tecnologia da Informação**, v.3, n.2, p.41-8, 2003.

MALTSEV, N. et al. PUMA2--grid-based high-throughput analysis of genomes and metabolic pathways. **Nucleic Acids Research**, v.34, p.D369-D372, 2006.

MARCOTTE, E.M. et al. Detecting protein function and protein-protein interactions from genome sequences. **Science**, v.285, n.5428, p.751-3, 1999.

MARKOWITZ, V.M. et al. The integrated microbial genomes (IMG) system. **Nucleic Acids Research**, v.34, p.D344-D348, 2006.

MERKEEV, I.V.; NOVICHKOV, P.S.; MIRONOV, A.A. PHOG: a database of supergenomes built from proteome complements. **BMC Evoltionary Biology**, v.6, p.52, 2006.

MORGENSTERN, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. BIOINFORMATICS, v.15, n.3, p.211-8, 1999.

MORGENSTERN, B. DIALIGN: finding local similarities by multiple sequence alignment. **Bioinformatics**, v.14, n.3, p.290-4, 1998.

MORGENSTERN, B. et al. Exon discovery by genomic sequence alignment. **Bioinformatics**, v.18, n.6, p.777-87, 2002.

NEEDLEMAN, S.B.; WUNSCH, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v.48, n.3, p.443-53, 1970.

OTU, H.H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. **Bioinformatics**, v.19, n.16, p.2122-30, 2003.

OUZOUNIS, C. Bioinformatics and the theoretical foundations of molecular biology. **Bioinformatics**, v.18, n.3, p.377-8, 2002.

OUZOUNIS, C.A.; VALENCIA, A. Early bioinformatics: the birth of a discipline--a personal view. **Bioinformatics**, v.19, n.17, p.2176-90, 2003.

OVCHARENKO, I. et al. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. Genome Res 2004 Mar;14(3):472-7.

OVERBEEK, R. et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. **Nucleic Acids Research**, v.33, n.17, p.5691-702, 2005.

PAN, X.; STEIN, L.; BRENDEL, V. SynBrowse: a synteny browser for comparative sequence analysis. **Bioinformatics**, v.21, n.17, p.3461-8, 2005.

PEARSON, W.R.; LIPMAN, D.J. Improved tools for biological sequence comparison. **Proceedings of National Academy of Science**, v.85, n.8, p.2444-8, 1988.

PETERSON, J.D. et al. The Comprehensive Microbial Resource. **Nucleic Acids Research**, v.29, n.1, p.123-5, 2001.

RANDHAWA, G.S.; BISHAI, W.R. Beneficial impact of genome projects on tuberculosis control. **Infectious Disease Clinics of North America**, v.16, n.1, p.145-61, 2002.

RASKO, D.A.; MYERS, G.S.; RAVEL, J. Visualization of comparative genomic analyses by BLAST score ratio. **BMC Bioinformatics**, v.6, n.1, p.2, 2005.

REFSEQ. NCBI Reference Sequence. Disponível em: <http://www.ncbi.nlm.nih.gov/RefSeq/> Acesso em: 8 out. 2007.

REN, Q.; KANG, K.H.; PAULSEN, I.T. TransportDB: a relational database of cellular membrane transport systems. **Nucleic Acids Research**, v.32, p.D284-D288, 2004.

Ren Q, Chen K, Paulsen IT. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 2007 Jan;35(Database issue):D274-279.

ROMUALDI, A. et al. GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. **Bioinformatics**, v.21, n.18, p.3669-71, 2005.

SCHWARTZ, S. et al. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. **Nucleic Acids Research**, v.31, n.13, p.3518-24, 2003a.

SCHWARTZ, S. et al. Human-mouse alignments with BLASTZ. **Genome Research**, v.13, n.1, p.103-7, 2003b.

SCHWARTZ, S. et al. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res, v.10, n.4, p.577-86, 2000.

SCHNEIDER A.; DESSIMOZ, C.; GONNET, GH. OMA Browser--exploring orthologous relations across 352 complete genomes. **Bioinformatics**, v.23, n.16, p.2180-2182, 2007.

SELENGUT, J.D. et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* 2007 January;35(Database issue): D260-D264.

SICHERITZ-PONTEN, T.; ANDERSSON, S.G. A phylogenomic approach to microbial evolution. **Nucleic Acids Research**, v.29, n.2, p.545-52, 2001.

SIEW, N.; AZARIA, Y.; FISCHER, D. The ORFanage: an ORFan database. **Nucleic Acids Research**, v.32, p.D281-D283, 2004.

SINHA, A.U.; MELLER, J. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. **BMC Bioinformatics**, v.8, p.82, 2007.

SMITH, T.F.; WATERMAN, M.S. Comparison of Biosequences. **Advances in Applied Mathematics**, v.2, p.482-9, 1981.

STOTHARD, P.; et al. BacMap: an interactive picture atlas of annotated bacterial genomes. **Nucleic Acids Research**, v.33, p.D317-D320, 2005.

STROHMAN, R.C. The coming Kuhnian revolution in biology. **Nature Biotechnology**, v.15, n.3, p.194-200, 1997.

SUHRE, K.; CLAVERIE, J.M. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. **Nucleic Acids Research**, v.32, p.D273-D276, 2004.

TANAKA, N. et al. G-InforBIO: integrated system for microbial genomics. BMC Bioinformatics 2006;7:368.

TATUSOV, R.L. et al. The COG database: an updated version includes eukaryotes. **BMC Bioinformatics**, v.4, n.1, p.41, 2003.

TATUSOV, R.L.; KOONIN, E.V.; LIPMAN, D.J. A genomic perspective on protein families. **Science**, v.278, n.5338, p.631-7, 1997.

TEKAIA, F.; YERAMIAN, E.; DUJON, B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. **Gene**, v.297, n.1-2, p.51-60, 2002.

TEKAIA, F.; YERAMIAN, E. Genome trees from conservation profiles. **PLoS Computational Biology**, v.1, n.7, p.e75, 2005.

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

TREANGEN, T.J.; MESSEGUER, X. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. **BMC Bioinformatics**, v.7, p.433, 2006.

UCHIYAMA, I. MBGD: microbial genome database for comparative analysis. **Nucleic Acids Research**, v.31, n.1, p.58-62, 2003.

UCHIYAMA, I.; HIGUCHI. T.; KOBAYASHI, I. CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. BMC Bioinformatics, v.7, p.472, 2006.

UVA FASTA SERVER. Disponível em: <http://fasta.bioch.virginia.edu/> Acesso em: 08 out. 2007.

VENTER, J.C. et al. The sequence of the human genome. **Science**, v.291, n.5507, p.1304-51, 2001.

VON MERING, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. **Nucleic Acids Research**, v.33, p.D433-D437, 2005.

VON MERING, C. et al. STRING 7: recent developments in the integration and prediction of protein interactions. **Nucleic Acids Research**, v.35, p.D358-D362, 2007.

WANG, H. et al. SynView: a GBrowse-compatible approach to visualizing comparative genome data. **Bioinformatics**, v.22, n.18, p.2308-9, 2006.

WEI, L. et al. Comparative genomics approaches to study organism similarities and differences. **Journal of Biomedical Informormatics,** v.35, n.2, p.142-50, 2002.

WILSON, G.A. et al. Orphans as taxonomically restricted and ecologically important genes. **Microbiology**, v.151, n.Pt 8, p.2499-501, 2005.

YANG, J. et al. ShiBASE: an integrated database for comparative genomics of Shigella. **Nucleic Acids Research**, v.34, p.D398-D401, 2006.

YANG, J. et al. GenomeComp: a visualization tool for microbial genome comparison. **Journal of Microbiological Methods**, v.54, n.3, p.423-6, 2003.

Sobre os autores

Marcos Catanho

Mestre em Biologia Celular e Molecular pelo Instituto Oswaldo Cruz (IOC/FIOCRUZ) e Bacharel em Farmácia pela Universidade Federal do Rio de Janeiro (UFRJ). Atualmente é doutorando em Biologia Celular e Molecular pelo Instituto Oswaldo Cruz (IOC/FIOCRUZ), onde desenvolve tese na área de Bioinformática. Tem experiência nas áreas de Biologia Molecular e Bioinformática, atuando principalmente nos seguintes temas: análise comparativa de genomas e evolução e desenvolvimento de algoritmos e aplicativos para genômica comparativa e funcional de procariotos.

Antonio Basílio de Miranda

Farmacêutico pela Universidade Federal do Rio de Janeiro (UFRJ), Mestre e Doutor em Ciências (Departamento de Genética, UFRJ). Pós-Doutorado no Sanger Institute (UK). Possui experiência em Biologia Molecular e Bioinformática, atuando principalmente nas áreas de Genômica Comparativa e Evolução Molecular.