INSTITUTO OSWALDO CRUZ

Mestrado em Biologia Computacional e Sistemas

Área de Concentração: Genômica Funcional, Evolução e Filogenômica

GENOMYCDB 2.0 – ATUALIZAÇÃO DE TECNOLOGIA E AMPLIAÇÃO DE ESCOPO

Leandro Cesar Monteiro

Rio de Janeiro

2015

Ficha catalográfica elaborada pela
Biblioteca de Ciências Biomédicas/ ICICT / FIOCRUZ - RJ

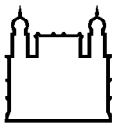
M775 Monteiro, Leandro Cesar

GENOMYCDB 2.0: atualização de tecnologia e ampliação de escopo / Leandro Cesar Monteiro. – Rio de Janeiro, 2015.
xi,84 f. : il. ; 30 cm.

Dissertação (Mestrado) – Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2015.
Bibliografia: f. 59-62

1. Micobactérias. 2. Genoma. 3. Banco de dados biológico. 4. Genômica comparativa. I. Título.

CDD 572.8628



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Computacional e Sistemas

Autor: Leandro Cesar Monteiro

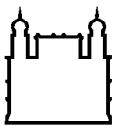
Título da Dissertação: GenoMycDB 2.0 – Atualização de Tecnologia e Ampliação de Escopo

Dissertação apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Mestre em Biologia Computacional e Sistemas

Orientadores: Dr. Antônio Basílio de Miranda e Dr. Marcos Catanho

RIO DE JANEIRO

2015



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: Leandro Cesar Monteiro

Título da Dissertação

GenoMycDB 2.0 – Atualização de Tecnologia e Ampliação de Escopo

Orientadores: Dr. Antônio Basílio De Miranda

Dr. Marcos Catanho

Programa: Biologia Computacional e Sistemas

Examinadores:

Dr. Nicolas Carels (Presidente)

Dra. Ana Carolina Ramos Guimarães (Membro)

Dr. Diogo Tschoeke (Membro)

Dr. Luis Fernando Encinas Ponce (Suplente)

Dr. Floriano Paes Silva Jr. (Suplente)

Rio de Janeiro, 15 de Maio de 2015

Dedicatória

A minha família, namorada e amigos,

minha sincera gratidão.

“Curai os enfermos, limpai os leprosos, ressuscitai os mortos, expulsai os demônios; de graça recebestes, de graça dai.”

Mateus 10:8

Agradecimentos

Aos meus pais, a quem devo tudo.

Aos meus orientadores, que acreditaram em mim.

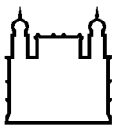
Aos colegas Elaine Carvalho e Pedro Erthal pelo apoio.

Aos meus colegas de curso, que trilharam o mesmo caminho.

A coordenação da Pós-Graduação e Corpo Docente pela organização.

Aos amigos Marcio Albuquerque, Altair Mentzingen e Antonio Antilon, pela contribuição.

E ao meu amor, a quem dedico a plenitude deste trabalho.



Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

GenoMycDB 2.0 – Atualização de Tecnologia e Ampliação de Escopo

RESUMO

DISSERTAÇÃO DE MESTRADO

Leandro Cesar Monteiro

Esta dissertação trata do desenvolvimento de uma nova versão do sistema GenoMycDB, desde a aquisição dos dados brutos até sua construção e análise dos resultados. Esta versão contempla a análise comparativa dos genomas de 63 linhagens de micobactérias através do processamento em diferentes softwares de análises biológicas e a organização dos dados gerados em um banco de dados relacional, acessível através de uma interface desenvolvida em tecnologia WEB atual e restrito a filtros de natureza biológica.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

GenoMycDB 2.0 –Tecnology Updrage and Scope Expansion

ABSTRACT

DISSERTAÇÃO DE MESTRADO

Leandro Cesar Monteiro

This dissertation deals with the development of a new version of GenoMycDB system, from acquisition of raw data to its construction and analysis of results. This version includes a comparative analysis of the genomes of 63 mycobacteria lineages by processing software's in different biological tests and the organization of the data generated in a relational database, accessible through an interface developed in current web technology and restricted by filters of a biological nature.

Sumário

Lista de Figuras

1. Introdução

Passado todo o fascínio com o qual o projeto Genoma Humano e seus desdobramentos foram apresentados ao mundo em meados dos anos 1990, a comunidade científica ultrapassou a barreira do sequenciamento do código genético dispendioso e demorado, aprendendo a lidar com os resultados deste esforço na década imediatamente posterior a este avanço (Pevsner 2009).

Em conformidade com o desenvolvimento exponencial de métodos e recursos de computação, diferentes iniciativas de instituições de pesquisa e grupos de estudo começaram a convergir ideias e amadurecer maneiras de lidar com esta enormidade de dados biológicos gerados a partir do processo de sequenciamento.

Os produtos resultantes deste esforço em conjunto foram aprimorados sob a forma de softwares cada vez mais eficientes para análises de dados brutos e repositórios de informações disponibilizados para consulta pública (Falda et al. 2012).

Desde então, diversos organismos objetos de pesquisa tiveram seus genomas completamente sequenciados, gerando uma grande quantidade de dados que foi depositada em diferentes bancos de dados biológicos cujos objetivos são organizar estas informações a partir da sua natureza e aplicação (Fraser et al. 2000).

Com a disponibilização eficiente, rápida e cada vez mais confiável destes dados, a comunidade científica pôde abordar sob variados aspectos, o estudo da estrutura, organização e evolução de genomas e a predição e classificação funcional de genes, alcançando altos níveis de complexidade e eficiência.

As aplicações finais destes estudos permitiram um salto qualitativo no tratamento e prevenção de doenças, um melhor entendimento de relações filogenéticas e uma melhor compreensão da evolução das espécies em nível molecular (Souza et al. 2014).

No ano de 2006 foi desenvolvido no Instituto Oswaldo Cruz, uma iniciativa onde os resultados referentes à comparação do genoma de seis micobactérias foram disponibilizados para consulta em um bem sucedido projeto chamado GenoMycDB, conduzido pelos Drs. Antônio Basílio, Marcos Catanho, Wim

Degrave e Rafael Mascarenhas (Catanho et al. 2006).

Os capítulos a seguir apresentam um breve histórico sobre o desenvolvimento da tecnologia de processamento de dados biológicos e sua aplicação no estudo da genômica comparativa.

No capítulo 2 são apresentados os fundamentos teóricos que norteiam o desenvolvimento deste trabalho, bem como uma contextualização do cenário atual, exemplificando por fim o conceito de bancos de dados biológicos, suas características e aplicações.

Na parte deste estudo relacionada à biologia computacional, são referenciadas a importância da tecnologia no estudo de doenças causadas por micobactérias, bem como um breve referencial teórico sobre a natureza destes organismos.

No capítulo 3 são apresentados os objetivos do projeto. No capítulo 4 demonstramos a metodologia e os meios pelos quais se tornou possível atingir o objetivo de construir o software e alcançar os resultados, expressos ao fim da dissertação, no capítulo 5.

No capítulo 6 é apresentada uma proposta de discussão e comentários com possíveis desdobramentos do projeto e eventual ampliação do escopo de pesquisa, onde também apresentamos um estudo de caso.

2. Fundamentos teóricos

Com o surgimento da Bioinformática e da Biologia Computacional por volta da década de 1970, diversas iniciativas de bancos de dados e ferramentas computacionais foram desenvolvidas a fim de disponibilizar para a comunidade científica os meios necessários para acessar e interpretar uma série de dados biológicos.

A contribuição destas áreas tornou-se bastante evidente nas décadas de 1990 e 2000 quando o desenvolvimento de computadores com grande poder de processamento para uso pessoal se tornou viável e o desenvolvimento de redes de computadores em escala global revolucionou o compartilhamento da informação no mundo.

Podemos assumir que o estudo da bioinformática permite principalmente aos pesquisadores coletar, processar e interpretar dados originados do sequenciamento de genomas, com o objetivo de estudar proteínas, enzimas, vias metabólicas e outras estruturas e funções biológicas (Enright et al. 2002).

Entre os produtos gerados pelo desenvolvimento da bioinformática podemos destacar a dedução da forma e a função de proteínas a partir de sequências de aminoácidos, a busca pelos genes e proteínas em um determinado genoma e a determinação de áreas na estrutura da proteína onde as moléculas de novas drogas podem aderir (Cohen 2004).

Mais especificamente, no campo da Biologia, houve particular avanço com a organização de tecnologias de alta capacidade, nomeadas coletivamente como ômicas - por exemplo, genômica, transcriptômica e proteômicas (Souza et al. 2014).

Atualmente existe uma série de recursos disponíveis na Internet ao público em geral, com o objetivo de organizar, integrar e fornecer acesso eficiente à quantidade crescente de informações biológicas produzidas ao longo de décadas de pesquisa.

Nos últimos anos, inúmeros projetos científicos, dentre os quais podemos destacar o Projeto ENCODE (Feingold et al. 2004), utilizaram estas tecnologias de alta capacidade de processamento com o objetivo de visualizar, pesquisar, recuperar e analisar uma enorme quantidade de dados biológicos.

Estas novas abordagens permitiram à comunidade científica adquirir um significativo conhecimento sobre os genomas dos organismos pesquisados.

No entanto, a criação e manutenção de tais ferramentas e sua disponibilização na Internet revelaram-se um desafio, não só porque é preciso lidar com grandes quantidades de dados, mas principalmente porque estes dados exigem o estudo e modelagem de esquemas e estruturas que representam com precisão a complexidade dos sistemas biológicos, o que, por si só, torna- se uma tarefa muito difícil de ser realizada.

Ao contrário de modelos que tratam de resultados exatos e precisos, a natureza dos sistemas biológicos pressupõe uma enormidade de variáveis que aumentam exponencialmente a complexidade dos estudos e análise de dados.

Além disso, é preciso garantir uma qualidade considerável para que sejam construídos sistemas robustos e eficientes de recuperação e manipulação de dados eficiente, utilizando interfaces inteligentes e amigáveis.

2.1. Histórico e contexto atual

Desde que a humanidade começou a escrever a sua própria história, doenças causadas por micobactérias são representadas das mais variadas formas sob diferentes aspectos histórico-religiosos ao redor do mundo.

Podemos encontrar registros da doença de Hansen, causada pelo bacilo *Mycobacterium leprae* e descoberto por Gerhard Armauer Hansen em 1874, em textos religiosos como a bíblia cristã, onde a palavra hebraica “tzaraat” foi traduzida para o grego “lepra”, que remete à ideia de “escamoso” ou “portador de escamas”, uma referência ao aspecto degradado e muitas vezes repulsivo das lesões na pele.

Na idade média em especial, a Hanseníase se instalou como uma epidemia na Europa (Pinto 1995) (Figura 2.1-1).

Figura 2.1-1: Representação do tratamento da hanseníase na idade média.
Fonte: Escola Politécnica Federal de Lausanne, França.

Embora tenha sido atribuída a diferentes doenças que causavam algum tipo de deformidade aparente da pele, a Hanseníase foi relacionada à ideia do pecado na cultura hebraica, o que pode ser interpretado como uma alusão ao castigo divino sobre indivíduos portadores da doença, seja pelo contágio e convivência com outros doentes ou pelas más condições de higiene em discordância com os rituais de purificação e limpeza religiosa daquela cultura.

No contexto atual países como Índia, Brimânia, Nepal, Brasil, Madagascar, Moçambique e Tanzânia representam 90% dos casos de Hanseníase no mundo (Rodrigues e Lockwood 2011); segundo dados da OMS (Nota descritiva Nº 101 de Janeiro de 2014), o Brasil é o único país das Américas que não conseguiu reduzir o número de novos casos para um em cada 10.000 habitantes.

A Hanseníase figura na lista da OMS (Alberts et al. 2011) como uma das doenças consideradas negligenciadas, ou seja, doenças radicais endêmicas cujas medidas preventivas e de tratamento não estão disponíveis ao livre acesso nas áreas mais pobres do mundo (Figura 2.1-2).

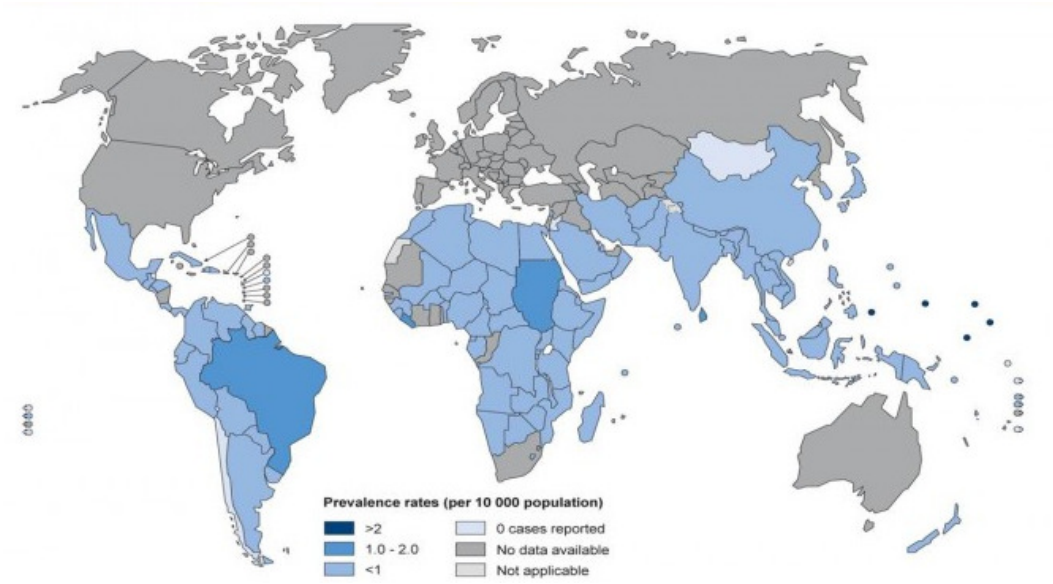


Figura 2.1-2: Taxas de prevalência da hanseníase no Brasil e no mundo em 2011. Fonte: Organização Mundial da Saúde.

A tuberculose é outra doença que foi muito bem documentada durante a história. Embora tenha vitimado milhões de seres humanos nas mais diferentes localidades do mundo, teve somente na metade do século XIX o seu agente causador, o bacilo *Mycobacterium tuberculosis*, isolado pelo pesquisador alemão *Robert Koch* em 1882.

Figura 2.1-3: Múmia húngara do século 18 onde foram encontradas cepas do bacilo M. tuberculosis*. Fonte: Museu Húngaro de História Natural.*

Esta doença foi relatada por médicos clássicos Greco-Romanos como Hipócrates, Plínio, o velho, ou Areteu da Capadócia (Oster 1928) e mais recentemente, a tuberculose foi diagnosticada em ossos humanos datados de 8.000 a.C., em restos mortais mumificados do antigo Egito e em múmias húngaras do século 18 (Kay et al. 2015) (Figura 2.1-3).

Assim como a hanseníase, a tuberculose é um grave problema de saúde pública no brasil. A cada ano, segundo dados do Ministério da Saúde[1], são notificados aproximadamente 70 mil novos casos, levando 4,6 mil pessoas a óbito em decorrência desta doença, fazendo do Brasil o 17º colocado entre os 22 países responsáveis por 80% do total de casos de Tuberculose no mundo.

Ao longo dos anos estas doenças tiveram seus métodos de diagnósticos e tratamentos aperfeiçoados e foram praticamente erradicadas em países desenvolvidos. Entretanto, o surgimento de linhagens multirresistentes provocou o ressurgimento destas mesmas doenças em vários países (Silva e Boéchat 2004).

Podemos citar ainda como potencial agente patogênico, o complexo *Mycobacterium avium* - associação dos bacilos *Mycobacterium avium* e *Mycobacterium intracellulare* - causador de doenças respiratórias e sistêmicas como a Síndrome de Lady Windermere em pessoas imunodeprimidas[2].

Apesar do esforço milenar que a humanidade emprega no tratamento e prevenção de doenças causadas por micobactérias, ainda hoje novas formas e manifestações infecciosas são objetos de crescente interesse e pesquisa, como por exemplo a doença pulmonar provocada por micobactérias não tuberculosas (MNT), possivelmente por se constatar no presente um aumento expressivo neste tipo de incidência (Griffith et al. 2007).

[1] http://portalsaude.saude.gov.br/index.php?option=com_content&view=article&id=11045&Itemid=674 – Acesso em 13/04/2015

[2] http://www.cdc.gov/ncidod/dbmd/diseaseinfo/mycobacteriumavium_t.htm - Acesso em 13/04/2015

2.2. 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* (Figura 2.2-1), respectivamente (Goodfellow e Minnikin 1984).

Pertencentes ao gênero de actinobactérias bacilares, são organismos aeróbios, sem mecanismos de locomoção próprios e que não formam endosporos ou esporos; têm forma de bastonetes delgados, retos ou ligeiramente encurvados, com raras formas ramificadas, além de elevado índice de patogenia (Madigan et al. 2010).

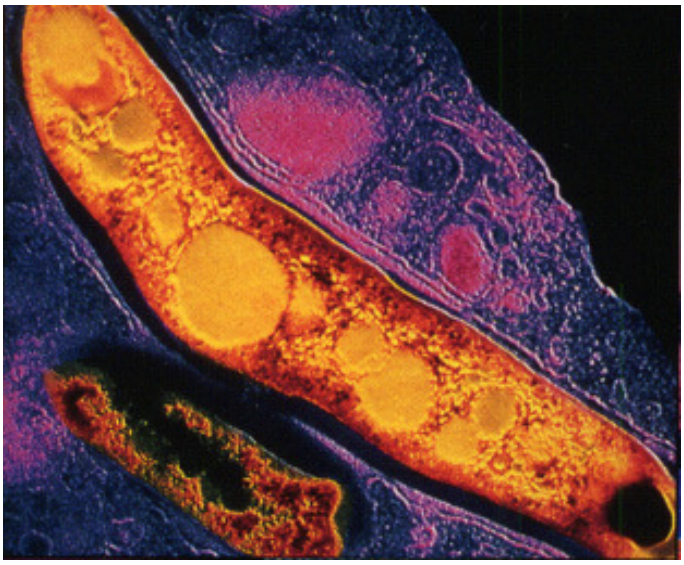


Figura 2.2-1: Mycobacterium tuberculosis. *Fonte:* University of Wisconsin-Madison.

As micobactérias possuem ainda características peculiares como álcool-ácido resistência, explicado pela elevada taxa de concentração de ácido micólico na sua parede celular (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 (Kenneth e Ray 2004).

A maior parte destas propriedades pode ser explicada pelo elevado teor de lipídios destas bactérias, sobretudo na parede celular, que afeta drasticamente a permeabilidade destes microrganismos à água, a soluções corantes, a agentes químicos e a nutrientes (Goodfellow e Minnikin 1984).

Estes organismos vivem basicamente na água e em fontes de alimento vegetal/animal, podendo em alguns casos, incluindo os bacilos da hanseníase e da tuberculose, serem encontrados na forma de parasitas obrigatórios.

Algumas espécies de micobactérias adaptaram-se para um crescimento em substratos simples, utilizando amônia ou amino ácidos como fontes de nitrogênio e de glicerol como fonte de carbono na presença de sais minerais (Madigan et al. 2010). As diferentes espécies possuem temperaturas de crescimento ideal variáveis entre 25 ° C a mais de 50 ° C .

Através de testes filogenéticos (Figura 2.2-2) foi possível distinguir diferentes espécies e linhagens, classificadas entre grupos de crescimento lento e rápido, onde o crescimento de colônias pode ser observado em meio de subcultura em poucos dias.

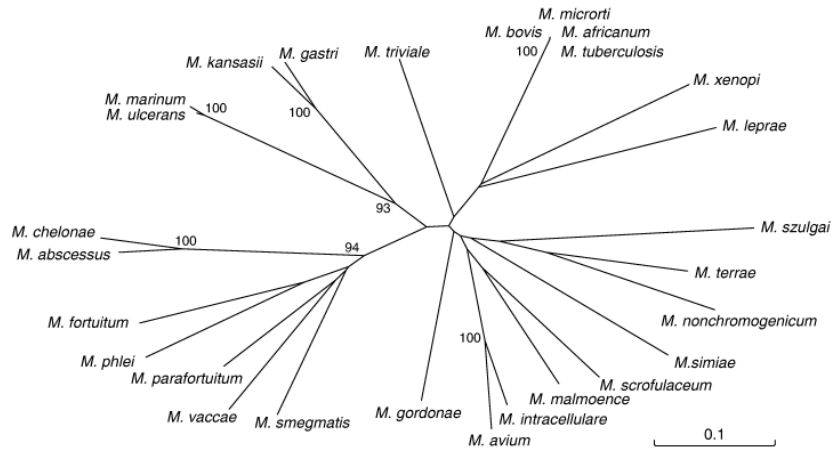


Figura 2.2-2: Árvore filogenética do gênero Mycobacterium. *Fonte:* Identification of Mycobacterium species by comparative analysis of the dnaA gene *(Mukai Tetsu et al. 2006).*

Existe ainda uma classificação médica internacional que divide as espécies em grupos (ou complexos) a partir de sua patogenicidade com o propósito de facilitar o diagnóstico e tratamento.

Entre estes complexos, podemos destacar:

- Complexo *Mycobacterium tuberculosis* (MTBC), formado pelas espécies *M. tuberculosis*, *M. bovis*, *M. africanum*, e *M. microti*, que são as causadoras da tuberculose humana e animal. Complexo *Mycobacterium avium* (MAC)**,** grupo que costuma figurar entre causadores de óbito em pacientes com AIDS, formado pelas espécies *M. avium*, *M. avium paratuberculosis*, *M. avium silvaticum*, *M. avium "hominissuis"*, *M. colombiense* e *M. indicus pranii*.
- *M. leprae* e *M. lepromatosis*, causadores da Hanseníase.
- Micobactérias não tuberculosas (MNT) que são todas as outras micobactérias que podem causar doenças similares à Tuberculose ou doenças de pele similares à hanseníase, além de sintomas não completamente exemplificados (Griffith et al. 2007).

2.3. Bioinformática

O termo bioinformática foi originalmente usado por Paulien Hogeweg e Ben Hesper no começo dos anos 1970 para representar o estudo dos processos biológicos utilizando recursos computacionais (Hesper e Hogeweg 1970).

O estudo da bioinformática contempla conhecimentos de biologia, ciência da computação, física, química, e matemática/estatística e suas ramificações para processar os dados biológicos ou biomédicos (Figura 2.3-1).

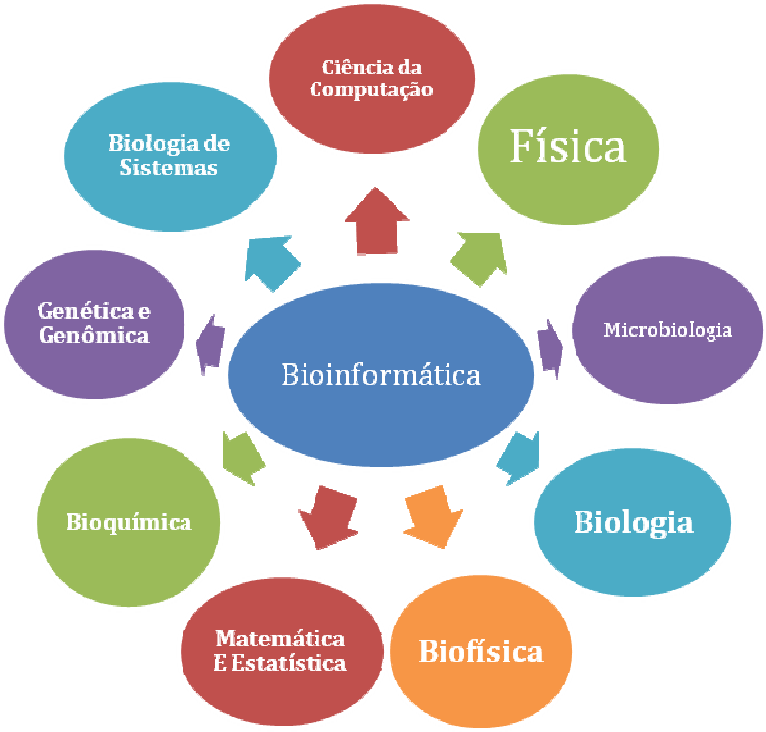


Figura 2.3-1: Multidisciplinaridade da bioinformática.

Ao processar estes dados utilizando softwares específicos é possível alcançar resultados como identificação de genes, previsão de estruturas tridimensionais de proteínas, identificação de inibidores de enzimas, organização e relacionamento de informação biológica, simulação de processos celulares, construção de árvores filogenéticas, comparação de comunidades microbianas para construção de bibliotecas genômicas, analise de experimentos de expressão gênica entre outras inúmeras aplicações (Attwood et al. 2011).

2.4. Genômica comparativa

Um genoma é uma sequência de DNA que contém toda a informação hereditária de um organismo. Cada sequência é constituída de genes, que na classificação da genética clássica é a unidade fundamental da hereditariedade do ser vivo.

O gene pode ser descrito como uma sequência de nucleotídeos distintos cuja informação pode ser transcrita na forma de RNA, que por sua vez é a molécula responsável por sintetizar uma determinada proteína na célula.

Para ser decodificado um genoma precisa ser sequenciado, ou seja, ser submetido a uma série de métodos bioquímicos para relacionar a composição do DNA e

representá-la através dados digitais, resultantes do processamento computacional da informação molecular. Na maioria das vezes, depois de sequenciado, o genoma de uma espécie é gravado em arquivos digitais e disponibilizado para a consulta pública.

A partir de uma iniciativa pioneira do Departamento de Energia dos EUA (DOE) para obter uma sequência de referência do genoma humano, foi lançado em 1990 o Projeto Genoma Humano [3].

O plano inicial era obter uma melhor compreensão dos potenciais riscos para a saúde humana e para o ambiente, causados pela produção e utilização de novos recursos e tecnologias energéticas. Mais tarde, os recursos tecnológicos gerados por este projeto estimularam o desenvolvimento de outras iniciativas públicas e privadas com objetivos parecidos.

Desde os anos 1990 (Figuras 2.4-1 e 2.4-2), os códigos genéticos completos de quase 7.500 organismos vivos, e em alguns casos já extintos[4] foram decifrados representando uma enorme variedade de dados de interesse médico, comercial, ambiental, acadêmico e industrial.

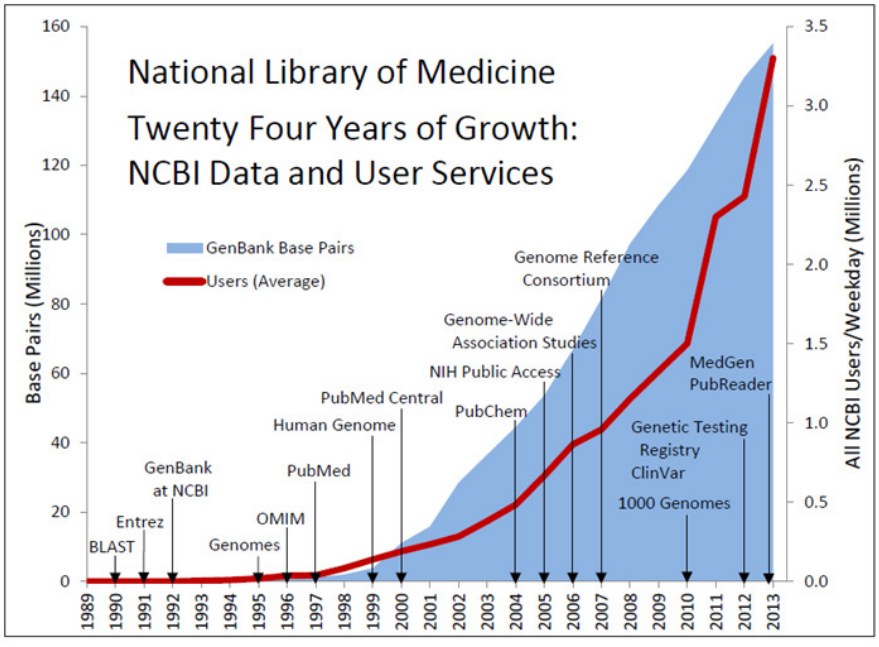


Figura 2.4-1: Crescimento da disponibilização de pares de base sequenciados nos últimos anos. Fonte: http://www.nlm.nih.gov

[3] What was the Human Genome Project? Genome.gov. NHGRI. NIH. http://www.genome.gov/12011238
[4] Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth http://www.cell.com/current-biology/abstract/S0960-9822%2815%2900420-0

Assim, com o tempo de execução de cada um destes projetos se tornando cada vez mais curto devido aos novos avanços nas técnicas de sequenciamento e processamento de dados, a viabilidade em obter e analisar sequências genômicas (totais ou parciais) de comunidades microbianas inteiras gerou novas e importantes descobertas científicas que puderam ser antecipadas do distante futuro previsto.

De forma conjunta, o sequenciamento e análise de inúmeros genomas completos, os dados de expressão gênica e proteica de células, tecidos e órgãos (apoiados por outras tecnologias de alta capacidade, como transcriptômica e proteômica, respectivamente), aliados ao desenvolvimento da computação em paralelo, tecnologias e algoritmos mais eficientes permitiram novas abordagens que puderam ser utilizadas no estudo da estrutura do genoma, organização, evolução e na análise de expressão diferenciais de genes e proteínas, em previsões de estrutura de proteína tridimensional no processo de reconstrução metabólica e na predição funcional de genes.

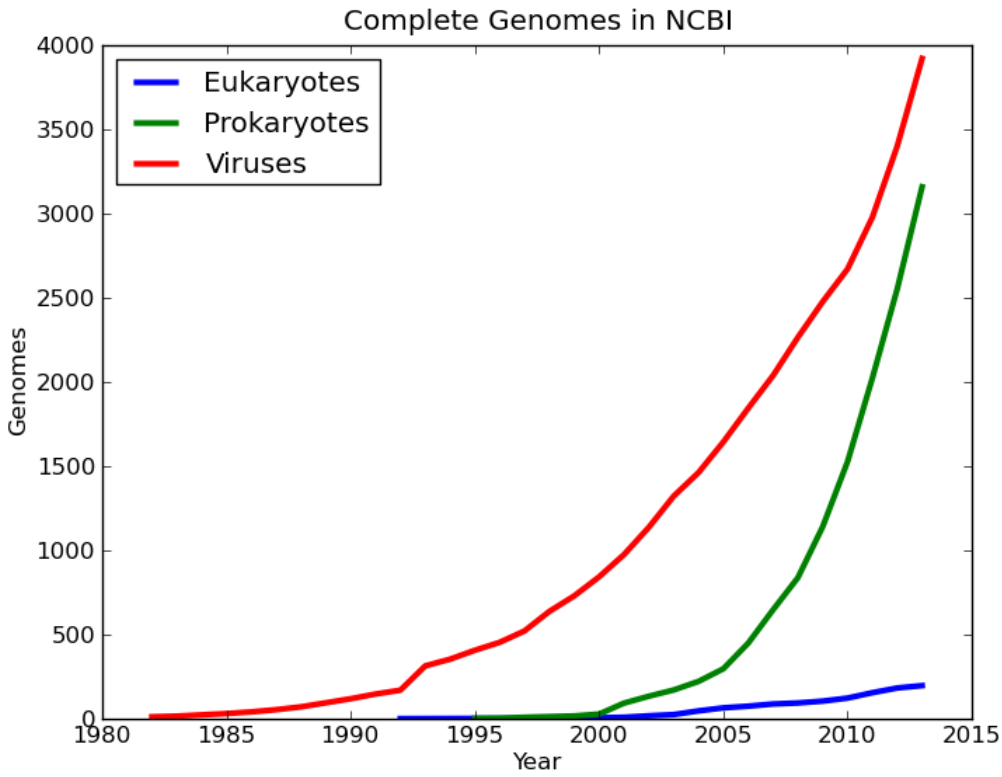


Figura 2.4-2: Quantidade de genomas completamente sequenciados até o momento.

Fonte: http://gregoryzynda.com/ncbi/genome/python/2014/03/31/ncbi-genome.html

A partir da disponibilidade destes dados, novos estudos (Touchman 2010) demonstraram a importância de analisar os genomas de forma comparativa, identificando as funções do DNA a partir de critérios como a anotação e expressão gênica além da identificação de homologias entre genes de diferentes organismos.

Assim podemos entender a genômica comparativa como uma maneira de analisar e comparar materiais genéticos de diversas espécies ou variedades, com o objetivo de investigar a estrutura, organização e evolução destes organismos, bem como revelar a função dos genes e regiões não codificadoras contidas nestes genomas.

A importância deste tipo de estudo se dá pelo fato de que um genoma é uma fonte única de informação, uma vez que representa, em princípio, todas as informações necessárias para a natureza construir um organismo (Figura 2.4-3).

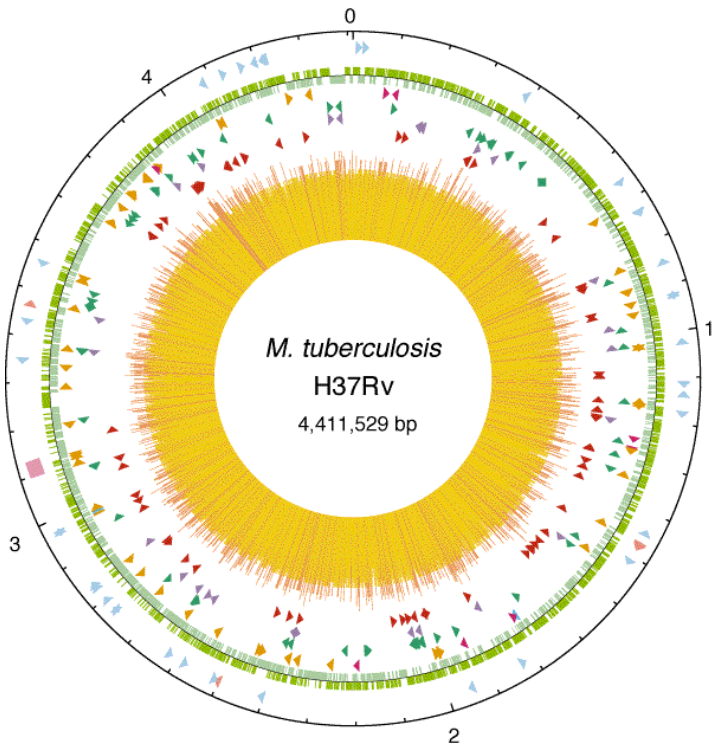


Figura 2.4-3: Mapa circular do cromossomo H37Rv - M. tuberculosis.

No entanto, não é claramente óbvio o que podemos fazer com toda esta grande quantidade de informação. Acredita-se, por exemplo, que a análise global dos genomas tem o potencial de fornecer um entendimento completo da genética, bioquímica, fisiologia e patogênese de microrganismos (Brosch et al. 2002).

Do ponto de vista prático, tecnologias de alta capacidade de processamento computacional permitem a comunidade científica a oportunidade de não só ampliar o nosso conhecimento sobre a biologia, mas também desenvolver novos sistemas de diagnóstico, medicamentos, vacinas mais eficientes, marcadores prognósticos e uma variedade de aplicações biotecnológicas.

Na atualidade existe o argumento de que este potencial só podem ser alcançado através do estudo comparativo dos genomas, regiões compartilhadas ou genes de duas ou mais espécies, subespécies ou variedades, porque um genoma sozinho, sem a estrutura filogenética do processo evolutivo, proporciona apenas uma compreensão parcial e incompleta destas questões (Ounit et al. 2015).

Neste sentido, surgiram exemplos claros (Fraser et al. 2000) de que a observação das informações evolutivas pode beneficiar análises genômicas, auxiliando a identificação de funções biológica de novos genes e ajudando a inferir padrões de recombinação nas espécies.

É possível também identificar a ocorrência de transferência lateral de genes entre diferentes espécies e a perda de material genético, além de contribuir para a distinção entre similaridades devido a homologia ou convergência evolutiva.

As análises comparativas dos genomas podem ser interpretadas a partir de variados tipos de abordagem, oferecendo uma visão diversificada acerca dos organismos estudados (Wei et al. 2002).

Entre as metodologias de comparação genômica que podem ser aplicadas no estudo e interesse biológico, podemos citar comparações:

- Envolvendo a estrutura genômica global de dois ou mais organismos;
- Entre regiões codificantes identificadas em diversos genomas;
- Envolvendo regiões não codificantes de diferentes genomas;

As análises genômicas comparativas microbianas são uma contribuição importante para a elucidação de aspectos fundamentais da genética, da bioquímica e da evolução de numerosas espécies.

Quanto aos microrganismos patogênicos em geral e micobactérias em específico, têm sido relatadas uma série de potenciais aplicações da análise comparativa

do genoma, visando especialmente à prevenção (desenvolvimento de vacinas mais eficientes), diagnóstico (desenvolvimento de métodos mais rápidos e mais precisos) e tratamento (desenvolvimento de novas drogas) da tuberculose e outras doenças causadas por micobactérias.

Algumas dessas aplicações incluem: identificação de genes únicos e fatores de virulência, reconstrução do metabolismo; caracterização de agentes patogénicos e identificação de novos alvos terapêuticos de diagnóstico; investigação das bases moleculares da patogênese e gama de hospedeiros, diferenciação de fenótipo entre isolados clínicos e populações naturais de patógenos; investigação da base genética de virulência e resistência aos medicamentos em bactérias causadoras de tuberculose.

Analisando comparativamente os genomas das micobactérias, é possível identificar regiões codificadoras e assinaturas proteicas conservadas unicamente no gênero *Mycobacterium* (Gao e Gupta 2012) o que demonstra a ancestralidade deste tipo de ramificação na evolução das espécies bacterianas.

2.5. FASTA

O formato FASTA é um tipo de arquivo baseado em texto simples utilizado para representar sequências de nucleotídeos e sequências proteicas, no qual estes são escritos utilizando letras como representação.

Este formato de arquivo permite que informações como nomes, comentários e identificadores, sejam inseridas precedendo a sequência, facilitando a compreensão e categorização destes dados.

A simplicidade desta organização (Figura 2.5-1*)* permite que dados complexos sejam processados por softwares específicos e scripts de tratamento de texto.

O formato FASTA foi muito bem aceito pela comunidade científica e atualmente é um padrão da área de bioinformática.

Embora não exista um padrão para a extensão de um arquivo de texto contendo sequências formatadas em FASTA, os arquivos referentes aos sessenta e três organismos estudados foram organizados com a extensão ".faa", que representa sequências codificadoras traduzidas em amino ácidos.

```
>gi|15607143|ref|NP_214515.1| chromosomal replication initiator protein DnaA
[Mycobacterium tuberculosis H37Rv]
MTDDPGSGFTTVWNAVVSELNGDPKVDDGPSSDANLSAPLTPQQRAWLNLVQPLTIVEGFALLSVPSSFV
QNEIERHLRAPITDALSRRLGHQIQLGVRIAPPATDEADDTTVPPSENPATTSPDTTTDNDEIDDSAAAR
GDNQHSWPSYFTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRFAHAAALAIAEAPARAYNPLFIWGES
GLGKTHLLHAAGNYAQRLFPGMRVKYVSTEEFTNDFINSLRDDRKVAFKRSYRDVDVLLVDDIQFIEGKE
GIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKKAQMER
LAVPDDVLELIASSIERNIRELEGALIRVTAFASLNKTPIDKALAEIVLRDLIADANTMQISAATIMAAT
AEYFDTTVEELRGPGKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAQRKILSEMAERREV
```

Figura 2.5-1: Primeira sequência do arquivo FASTA referente ao organismo M. tuberculosis.

Os bancos de dados primários utilizam em sua maioria o formato FASTA para disponibilização de conteúdo.

2.6. BLAST

O BLAST+ (acrônimo de *Basic Local Aligment Search Tool*) é um software utilizado para comparar sequências de dados biológicos como nucleotídeos, proteínas e aminoácidos (Altschul et al. 1997).

Existem várias versões de BLAST para comparação de sequências proteicas e de nucleotídeos, tais como BLASTp, BLASTn, BLASTx, tBLASTn, tBLASTx que podemos representar genericamente como BLAST+.

Uma pesquisa utilizando o BLAST+ permite a comparação de uma ou mais sequências contra uma biblioteca ou base de dados de sequências, identificando nestas as similaridades entre a sequência pesquisada e as outras contidas.

O grau de similaridade pode variar de acordo com os parâmetros informados pelo usuário, onde este pode atribuir maior ou menor grau de semelhança.

Numa situação hipotética, após descobrir um gene anteriormente desconhecido em um organismo, um usuário poderia elaborar uma pesquisa utilizando o BLAST+ no genoma de outro organismo para verificar se estes dois organismos compartilham algum gene semelhante.

Este software foi desenvolvido pelo NCBI – URL: http://www.ncbi.nlm.nih.gov na década de 1990 e possui uma ampla comunidade de usuários e colaboradores ao redor do mundo.

Após a obtenção, organização e tratamento dos arquivos FASTA, estes foram

importados para um banco de dados utilizando uma ferramenta da suíte BLAST+, visando preparar a base de consulta.

Finalmente, utilizando o BLAST+, para cada uma das sequências de cada um dos organismos alvo, foi realizada uma comparação par a par entre todas as sequências proteicas obtidas a partir das regiões anotadas como codificadoras nos organismos contemplados neste estudo.

2.7. Inferência de homologia

Na biologia, a homologia é o estudo das semelhanças entre estruturas evolutivas de diferentes organismos que possuem a mesma origem ontogenética e filogenética. Estas estruturas podem ou não ter a mesma função[5].

O estudo da homologia entre espécies corrobora sob diferentes aspectos a teoria da evolução, pois sugere que existe uma ancestralidade comum entre organismos diferentes a partir da identificação de estruturas semelhantes com a mesma origem.

Existem três tipos de homologia: (i) ortologia, quando as sequências têm um único e mesmo ancestral comum; (ii) paralogia, quando se originam de uma duplicação gênica; e (iii) xenologia, quando se originam por transferência lateral (ou horizontal) (Matioli e Fernandes 2012).

Em escala genética seria mais preciso definir que duas estruturas biológicas que possuem relação evolutiva podem ser consideradas homogenéticas (Amorim 2002).

Ainda segundo (Matioli e Fernandes 2012), para fornecer informações filogenéticas e reconstruir o histórico evolutivo de determinados organismos, apenas sequências ortólogas devem ser levadas em consideração.

Existe ainda uma subcategorização entre genes ortólogos, que são os chamados co-ortólogos, ou seja, genes que duplicaram em uma espécie e são simultaneamente ortólogos de genes em outra espécie (Mudado 2007).

[5] "Homology". *Encyclopædia Britannica Online* . Visitado em 01/04/2015

No estudo genético de organismos patogênicos, é particularmente importante a possibilidade de inferir a homologia entre as sequências e a relação filogenética entre espécies para que sejam determinadas estratégias de controle de doenças, diagnósticos e novas drogas.

A partir deste embasamento teórico, para interpretar os dados resultantes da comparação par a par realizada pelo software BLAST+, utilizamos o OrthoMCL – URL: http://www.orthomcl.org (Li et al. 2003).

O OrthoMCL é um software desenvolvido para realizar o agrupamento de sequências de proteínas ortólogas e parálogas em escala genômica e determinar um perfil filogenético dos organismos pesquisados, sendo uma ferramenta bastante experimentada para anotação de genomas eucarióticos cujo algoritmo de análise é utilizado para o agrupamento de sequências de proteínas previamente processadas.

Este algoritmo identifica os melhores hits recíprocos entre pares, dentro do mesmo genoma, como potenciais parálogos e melhores hits recíprocos em dois genomas como potenciais pares ortólogos, de forma que as proteínas relacionadas estão interligadas por um gráfico de similaridade (Figura 2.7-1).

Figura 2.7-1: Método para agrupar proteínas ortólogas.

Em seguida, o MCL (Markov Clustering Algortim, Van Dongen 2000) URL: http://www.micans.org/mcl, é utilizado para dividir os grandes agrupamentos.

Este passo é baseado nos pesos entre cada par de proteínas e por isso são normalizados antes da execução do MCL, de modo a corrigir as diferenças na distância evolutiva (Figura 2.7-2).

Figura 2.7-2 Método para agrupar proteínas ortólogas.

Este processo é análogo à avaliação manual dos agrupamentos ortólogos de grupos de proteínas (COGs) (Tatusov et al. 2008).

O OrthoMCL é semelhante ao algoritmo INPARANOID (Remm et al. 2001), mas vai além no que diz respeito a agrupamentos de genes ortólogos de múltiplas espécies.

Os agrupamentos são coerentes com grupos identificados por EGO (Lee et al. 2002), e, segundo análises utilizando o esquema de classificação numérica para as enzimas EC number, baseado em reações químicas que catalisam, é atribuído um alto grau de confiabilidade a estes (Li et al. 2003).

2.8. Localização subcelular

A localização subcelular de uma proteína pode fornecer pistas quanto à sua função em um organismo. Por isso, a previsão de sua localização subcelular através de análises computacionais, é uma valiosa contribuição para a análise do genoma e sua anotação (Hartwell et al. 2010).

Mais especificamente, para agentes patogênicos bacterianos, a previsão de proteínas na superfície da célula é de particular interesse, devido ao potencial de tais proteínas serem alvos de drogas ou vacinas.

A localização subcelular de uma proteína é influenciada por vários elementos presentes dentro da estrutura primária da proteína, tais como a presença de um peptídeo sinal ou alfa-hélices que atravessam a membrana.

Vários algoritmos foram desenvolvidos para analisar estas características individuais. Para este estudo utilizamos a suíte de sistemas PSORTb - URL: http://www.psort.org/psortb (Gardy et al. 2005) que é capaz de analisar de uma só vez várias características da sequência, utilizando as informações obtidas a partir de cada uma dessas análises para gerar uma previsão geral do sítio de localização.

Por exemplo, as sequências nos formato FASTA, podem ser analisadas com o PSORT empregando o modelo de construção para bactérias GRAM- positivas.

Consequentemente a classificação subcelular de cada uma das proteínas de cada um dos organismos estudados podem ser relacionadas com as propriedades de homologia.

2.9. Gene ontology

O *gene ontology*, ou GO, é parte de um esforço internacional de classificação de ontologias chamado *Open Biomedical Ontologies* ou OBO – URL: http://obofoundry.org, cujo objetivo é controlar e regulamentar as definições e utilização de vocabulário das informações biomédicas.

Trata-se de uma importante iniciativa para unificar a representação de genes e suas expressões, tais como RNA e proteínas, para todas as espécies.

Este projeto tem como objetivo principal manter e desenvolver o vocabulário controlado de atributos de genes e produtos de genes, anotar genes e produtos de genes, assimilar e divulgar dados de anotação e fornecer ferramentas para facilitar o acesso a todos os aspectos dos dados fornecidos pelo projeto e permitir a interpretação funcional de dados experimentais.

Uma ontologia consiste em uma representação ou classificação utilizada como um meio para categorizar ou agrupar as informações em classes (Smith et al. 2010) (Figura 2.9-1).

```
id:         GO:0000016
name:       lactase activity
namespace:  molecular_function
def:        "Catalysis of the reaction: lactose + H2O = D-
glucose + D-galactose." [EC:3.2.1.108]
synonym:    "lactase-phlorizin hydrolase activity" BROAD
[EC:3.2.1.108]
synonym:    "lactose galactohydrolase activity" EXACT
[EC:3.2.1.108]
xref:       EC:3.2.1.108
xref:       MetaCyc:LACTASE-RXN
xref:       Reactome:20536
is_a:       GO:0004553 ! hydrolase activity, hydrolyzing O-
glycosyl compounds
```

Figura 2.9-1: Exemplo de termo do gene ontology

O projeto *gene ontology* fornece ontologias de termos definidos que representam três domínios. :

- Componente celular: Partes de uma célula ou do seu ambiente extracelular.
- Função molecular: Atividades elementares de um produto do gene no nível molecular, tais como a ligação ou a catálise.
- Processo biológico: Operações ou conjuntos de eventos moleculares com um começo e fim definidos, pertinentes para o funcionamento de unidades vivas integradas, tais como células, tecidos, órgãos e organismos.

Para atribuir a ontologia aos resultados esta análise, e categorizá-los em algum destes domínios, se pode utilizar o sistema A.R.G.O.T2 - URL: http://www.medcomp.medicina.unipd.it/Argot2.

Este sistema é disponibilizado para funcionar on-line e tem o objetivo de atribuir uma função potencial baseada na devida classificação e anotação do GO através da similaridade da sequência pesquisada.

Assim, é possível utilizar o A.R.G.O.T2 para atribuir as propriedades acima descritas em proteínas codificadas nos genomas estudados como um dos filtros de pesquisa.

Para isso, basta processar os arquivos FASTA utilizando o BLAST+, porém desta vez confrontando os resultados com a base de dados do UniProt - URL:

http://www.uniprot.org.

O UniProt é um dos maiores e mais completos bancos de dados de proteínas disponíveis para a consulta pública, e por isso pôde servir como fonte para fornecer a correta anotação das sequências pesquisadas.

2.10. Bancos de dados biológicos

Bancos de dados podem ser classificados como primários, cujo conteúdo é resultado de pesquisa básica, ou secundários, onde as informações contidas são resultado de análises a partir das fontes de dados contidos nos primários.

Não tardou para que a disseminação deste conteúdo por meio da Internet se tornasse uma popular maneira de suportar a troca de informação e a cooperação científica a partir de softwares específicos e interação com estes repositórios desenvolvidos exclusivamente para este fim.

Podemos citar entre os diferentes tipos de dados: sequências de nucleotídeos, sequências de proteínas, estruturas 3D de macromoléculas, expressão de genes, vias metabólicas, entre outros.

Hoje, os principais centros de pesquisa do mundo possuem grandes bancos de dados biológicos, dentre os quais podemos destacar:

- **GeneBank** – URL: http://www.ncbi.nlm.nih.gov/genbank – Um dos principais e mais populares bancos de dados de sequências anotadas de DNA e proteínas.
- **PDB** – URL: http://www.rcsb.org/pdb/home/home.do - Repositório de dados que armazena informações sobre estruturas em 3D de grandes moléculas biológicas.
- **Kyoto Encyclopedia of Genes and Genomes – KEGG** – URL - http://www.genome.jp/kegg - outro banco de dados gerado a partir de sequenciamento e tecnologias experimentais.
- **UniProt consortium - UniPro**t - URL: http://www.uniprot.org – Banco de dados de sequências de proteínas e informação funcional.

- **Reference Sequence Database - RefSeq -** URL: http://www.ncbi.nlm.nih.gov/refseq, banco de dados de referência de sequências e fonte de dados utilizada para este trabalho.

Estes repositórios de dados estão disponibilizados na Internet para acesso público e são mantidos e tutorados por grandes centros de pesquisa como o *National Center for Biotechnology Information* – NCBI - URL: www.ncbi.nlm.nih.gov, o *Research Collaboratory for Structural Bioinformatics* – RCSB - URL: http://home.rcsb.org, o *European Molecular Biology Laboratory* – EMBL – URL: http://www.embl.org/, e os centros *Bioinformatics Center do Institute for Chemical Research and Human Genome Center do Institute of Medical Science* da Universidade de Kyoto no Japão.

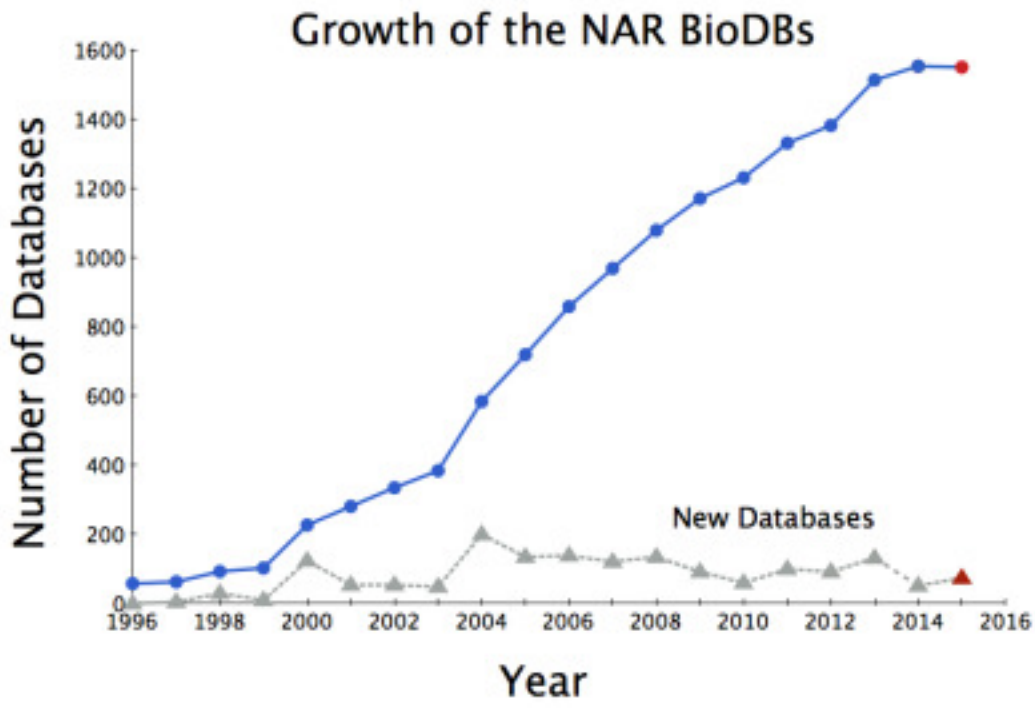


Figura 2.10-1 Crescimento de bancos de dados biológicos.
Fonte: http://scienceblogs.com/digitalbio/2015/01/30/bio-databases-2015/

A disponibilização pública de bancos de dados biológicos apresentou um enorme crescimento nos últimos 10 anos (Figura 2.10-1).

Logo, a importância e o sucesso de um estudo estão diretamente relacionados com a confiabilidade da fonte dos dados utilizados na pesquisa e por isso estas organizações tornaram-se responsáveis pelo controle de qualidade do conteúdo,

sendo estes administrados por especialistas que checam constantemente a consistência, a redundância, os conflitos e atualizações decorrentes da colaboração científica entre os demais centros de pesquisa.

Outra grande dificuldade apresentada é que em muitas ocasiões os autores e curadores de tais repositórios recebem pouca ou nenhuma remuneração pelos seus esforços produtivos, de forma que a quase inexistência de apoio financeiro para a criação e manutenção de bancos de dados biológicos acaba por inviabilizar grande parte destas iniciativas (Tatusov et al. 2000).

Para o estudo que se segue, apresentamos os principais recursos disponíveis na Internet totalmente ou parcialmente dedicados a espécies de micobactérias.

Cada sistema é classificado de acordo com a sua finalidade e funcionalidade, e os recursos computacionais apresentados estão todos disponíveis publicamente como serviços on-line, podendo ser utilizados para a pesquisa e produção de novos medicamentos, antígenos vacinais ou diagnósticos para a tuberculose, entre muitas outras aplicações.

2.11. RefSeq

As sequências de proteínas preditas codificadas pelos genomas de micobactérias completamente sequenciadas e as características correspondentes dos seus genes foram obtidas a partir do banco de dados de referência de genomas completos RefSeq (URL: http://www.ncbi.nlm.nih.gov/RefSeq) (Pruitt et al. 2007) do Centro Nacional de Informações sobre Biotecnologia (URL: http://www.ncbi.nlm.nih.gov).

Este repositório fornece um conjunto abrangente, integrado, não-redundante e bem anotado de dados de sequências, incluindo DNA genômico, transcrições, e proteínas (Pruitt e Maglott 2001).

As sequências do RefSeq formam uma fonte para os estudos médicos, funcionais e de diversidade biológica, fornecendo uma referência estável para anotação do genoma, identificação e caracterização de genes, mutação e análise de polimorfismo, estudos de expressão e análises comparativas.

Até o momento o RefSeq (Release 47 – 06/05/2015) disponibiliza mais de 12 mil

genomas distintos abrangendo vírus, procariotos e organismos eucariontes.

Estas informações estão disponíveis à comunidade científica sem qualquer tipo de restrição e podem ser pesquisadas através de buscas no próprio endereço eletrônico.

Os arquivos referentes aos organismos objeto deste estudo, foram gravados à partir de acesso FTP aos servidores do NCBI RefSeq e armazenados em local específico para processamento.

Os dados dos arquivos foram então tratados utilizando scripts de ajuste e filtros para servirem de entrada para o software BLAST+.

2.12. Sistemas e bancos de dados para estudo de micobactérias

Vários bancos de dados e ferramentas computacionais foram desenvolvidos com o objetivo de organizar, integrar e analisar a riqueza de informações geradas pelos projetos de sequenciamento em larga escala de genomas de micobactérias e os de outros organismos.

Podemos destacar entre estes bancos de dados:

MyBase - URL: http://mybase.psych.ac.cn/ - é uma plataforma integrada para o estudo funcional e genômico evolutivo do gênero *Mycobacterium*, compreendendo extensa revisão da literatura e anotação de dados sobre polimorfismos nos genomas de micobactérias, fatores de virulência e genes essenciais.

TBDB - URL: http://www.tbdb.org/ - é um repositório de dados sobre os genomas de *M. tuberculosis* e outras bactérias relacionadas. O TBDB combina sequências do genoma e dados de anotação e de expressão genética e disponibiliza uma plataforma de análise com ferramentas computacionais desenvolvidas para auxiliar os estudos genômicos comparativos e de expressão genética destes microrganismos. São disponibilizadas anotações características de genes e genomas, previsão de grupos ortólogos e blocos de sintenia, bem como epítopos imunológicos preditos e revisão de padrões de expressão genética.

MycoBrowser - URL: http://mycobrowser.epfl.ch/ - O portal *The Mycobacterial Browser* é um extenso repositório de dados de genômica e proteômica para quatro micobactérias relacionadas:

- *M. tuberculosis H37Rv*
- *M. leprae TN*
- *M. marinum M*
- *M. smegmatis MC2*

O sistema fornece informações sobre a sequência completa do genoma destes organismos revisados manualmente. Como parte deste portal, o banco de dados TubercuList integra um conjunto de informações sobre o genoma de *M. tuberculosis*, tais como anotações genômicas e proteínas e características, dados de medicamentos e transcriptoma, anotação mutante e óperon, e genômica comparativa.

MycoDB - URL: http://www.xbase.ac.uk/ - O xBASE é uma outra coleção de bancos de dados dedicado à análise comparativa do genoma bacteriano. Este sistema fornece dados pré- computados de análises de genoma comparados entre gêneros de bactérias, bem como grupos de ortólogos inferidos e anotações funcionais. Também proporciona análises pré-computadas de utilização de códons, composição de bases, o índice de adaptação de códons (CAI), de hidropatia e aromaticidade da proteínas codificadas nestas bactérias.

O MycoDB compreende atualmente dados comparativos de 61 genomas de micobactérias completamente sequenciados.

Mycobacterium tuberculosis Comparative Database – Este Banco de dados compreende análises comparativas de dados pré-computados do genoma de oito linhagens de *M. tuberculosis* isolados com fenótipos clínicos relevantes e epidemiologia da doença (grau variado de propagação, resistência aos medicamentos e gravidade clínica):

- *M. tuberculosis* F11
- *M. tuberculosis* Haarlem
- *M. tuberculosis* KZN 4207 (DS)
- *M. tuberculosis* KZN 1435 (MDR)
- *M. tuberculosis* KZN 605 (XDR)
- *M. tuberculosis* C
- *M. tuberculosis* 98-R604 INH- RMP-EM

- *M. tuberculosis* W-148

Entre os dados comparativos fornecidos por este repositório podemos citar: determinação de famílias de genes ortólogos, matrizes gráficas genômicas bidimensionais, mapeamento genômico comparativo e várias funcionalidades de anotações e comparação genômica.

2.13. GenoMycDb 1.0

Com o objetivo de proporcionar um recurso on-line para a classificação funcional de proteínas de micobactérias, bem como para a análise da estrutura do genoma, organização e evolução em tais espécies, em 2006 a equipe de bioinformática do Laboratório de Genômica Funcional e Bioinformática (Departamento de Bioquímica e Biologia Molecular do Instituto Oswaldo Cruz) desenvolveu um sistema chamado GenoMycDB – URL: http://www.dbbm.fiocruz.br/GenoMycDB - , um banco de dados relacional para análises comparativas em grande escala de genomas sequenciados de micobactérias com base nas suas proteínas preditas.

Apesar de, à época do desenvolvimento, existirem diversas iniciativas com objetivo similar, estes softwares, com raras exceções, não permitiam comparações dinâmicas e em larga escala entre estes dados.

O GenoMycDB 1.0 foi desenvolvido com o objetivo de suprir esta demanda em relação aos genomas micobacterianos, permitindo a realização de estudos de classificação funcional de proteínas em micobactérias, análises da estrutura, organização e evolução dos genomas destes organismos e a anotação de dados genômicos de outras espécies relacionadas.

Nesta versão, o banco de dados foi composto pelos genomas até então sequenciados de seis micobactérias:

- *M. tuberculosis* (cepas H37Rv e CDC1551)
- *M. bovis* AF2122 / 97
- *M. avium* subsp. paratuberculosis K10
- *M. leprae* TN
- *M. smegmatis* MC2 155

Para cada uma destas sequências de proteínas codificadas era fornecida a localização subcelular predita, o *cluster* atribuído de grupos ortólogos (CPV), características do gene correspondente e links para vários bancos de dados importantes.

Além disso, pares ou grupos de homólogos entre espécies selecionadas/cepas podem ser dinamicamente inferidos com base em critérios definidos pelo usuário.

Neste software, a seleção dos pares alinhados com atributos específicos pode ser realizada com base em um ou vários parâmetros de alinhamento, de forma que o usuário pode escolher, de acordo com seu objeto de pesquisa, um campo ou uma combinação de campos para interrogar o sistema (Figura 2.13-1).

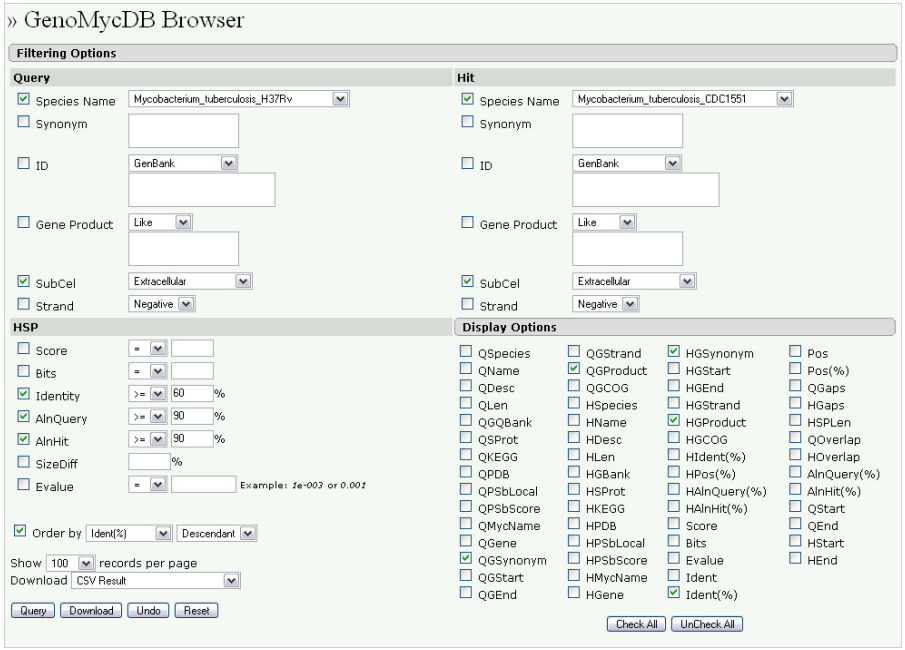


Figura 2.13-1: Visão geral da interface de pesquisa do GenoMycDB 1.0.
Fonte: http://www.dbbm.fiocruz.br/GenoMycDB

Desenvolvido utilizando um banco de dados MySQL e a linguagem de programação Perl, o GenoMycDB foi construído como parte de um estudo que citava o desenvolvimento de abordagens computacionais e ferramentas para a análise comparativa de genomas microbianos (Catanho et al. 2006).

Como resultado foi concebido um bem sucedido projeto de construção de software que permitiu a diversos estudos subsequentes, a identificação de relações evolutivas estruturais e funcionais entre os genomas dos organismos sequenciados naquele momento e compreendidos pelo estudo.

Esta versão apresentou avanços importantes em relação a outras iniciativas similares, tais como:

- Listas de pares ou grupos de potenciais sequências parálogos e ortólogos gerados dinamicamente com base nos diferentes filtros e parâmetros de similaridade entre as sequências alinhadas.
- Para cada par de proteínas alinhadas no resultado, ligações com os principais bancos de dados biológicos de sequencia (GenBank e SwissProt/TrEMBL), de estrutura tridimensional (PDB), de vias metrabólicas (KEGG) e de análise dos dados.
- Alinhamentos globais dos pares selecionados gerados instantaneamente em nível proteico ou nucleotídico, permitindo a inspeção mais detalhada das sequencias comparadas.
- Diferentes opções de exportação dos resultados

3. Objetivo

3.1. Objetivo geral

O objetivo principal deste estudo é ampliar o escopo de pesquisa e atualizar a tecnologia empregada para a nova versão do projeto.

3.2. Objetivos específicos

- Exibição de relatórios específicos contendo grupos de famílias de genes ortólogos entre diferentes espécies.

- Ampliação do escopo de pesquisa de seis (6) para sessenta e três (63) de micobactérias cujos genomas se encontram completamente sequenciados no banco de dados biológico RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) do Centro Nacional de Informações sobre Biotecnologia (http://www.ncbi.nlm.nih.gov/).

- Atualização da tecnologia empregada, visando acompanhar o desenvolvimento tecnológico global e atendimento os padrões atuais de sistemas, contemplando facilidades como escalabilidade, reuso e portabilidade de código.

- Possibilidade de comparar sequências proteicas particulares contra a base de dados reunida, através de uma ferramenta de alinhamento local (BLAST), inserida como uma funcionalidade padrão do novo sistema.

4. Material e métodos

Neste capítulo são descritos os recursos necessários para alcançar os objetivos do estudo. A aquisição e processamento dos dados e o desenvolvimento do sistema.

Para o processamento destes dados foi utilizado um servidor Silicon Graphics UV2000 de alto desempenho, cuja configuração disponibiliza 320 processadores Intel® Xeon® E5-4650 v2 2.4GHz-3.3GHz com 2 TB de memória RAM, que utiliza por sua vez o sistema operacional Linux, distribuição CentOS, versão 6.6.

O conjunto de dados de sequências de proteínas referentes aos genomas dos organismos pesquisados foi gravado, processado e submetido a diferentes tipos de análises utilizando os softwares:

- NCBI BLAST+ versão 2.2.30 - URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi
- OrthoMCL versão 5 - URL: http://www.orthomcl.org/orthomcl/
- PSORTb versão 3.0.2 - URL: http://www.psort.org/psortb
- A.R.G.O.T2 - URL: http://www.medcomp.medicina.unipd.it/Argot2/

4.1. Desenvolvimento

Podemos definir um software ou sistema de informação como um conjunto de componentes inter-relacionados, trabalhando juntos para coletar, recuperar, processar, armazenar e distribuir informações, com a finalidade de facilitar o planejamento, o controle, a coordenação, a análise e o processo decisório em organizações estudadas (Laudon e Laudon 1999).

Para atingirmos o objetivo deste estudo, utilizamos como referência a primeira versão do GenoMycDB que, conforme explicado no capítulo 2.8, se trata de uma interface WEB integrada à um banco de dados que, através de consultas utilizando filtros pré-definidos pelo usuário, retorna os resultados desejados.

Desde o desenvolvimento da primeira versão em 2006, a tecnologia de construção de softwares evoluiu bastante, principalmente na questão da mobilidade de equipamentos e escalabilidade de versões.

De lá para cá, a Internet se transformou numa gigantesca plataforma de desenvolvimento suportando verdadeiras comunidades de desenvolvedores e

permitindo o compartilhamento de recursos e plataformas tecnológicas gratuitamente, ajudando a formar o conceito de WEB 2.0.

Segundo Tim O`Reilly (2005)

> *Web 2.0 é a mudança para uma internet como plataforma, e um entendimento das regras para obter sucesso nesta nova plataforma. Entre outras, a regra mais importante é desenvolver aplicativos que aproveitem os efeitos de rede para se tornarem melhores quanto mais são usados pelas pessoas, aproveitando a inteligência coletiva.*

Assim, as etapas do desenvolvimento do sistema foram compostas pelo levantamento dos requisitos, definição da tecnologia, modelagem dos dados, construção do banco de dados, programação, implementação e testes.

Neste capítulo pretendemos demonstrar como foi possível construir o software, nomeado como GenoMycDB 2.0, com o objetivo de integrar os dados biológicos gerados, processá-los e exibir, através de relatórios dinâmicos, as análises comparativas entre os genomas das espécies estudadas.

4.2. Levantamento de requisitos

As atividades de levantamento e análise de requisitos estão presentes na etapa de definição do software, independentemente do modelo de ciclo de vida adotado. Como essa é uma fase que apresenta uma série de dificuldades, pois há o reconhecimento de que é na atividade de descoberta que surgem os problemas mais dispendiosos e de maior impacto negativo (Cruz 2004).

A Engenharia de Requisitos visa aplicar técnicas de engenharia em métodos de definição e análise de requisitos para garantir o atendimento das necessidades de informatização de processos através do software projetado (Azevedo Junior e Campos 2008).

Assim, o levantamento de Requisitos é normalmente atingível por meio da interação com um cliente. Neste caso, assumimos que o Laboratório de Genômica Funcional do IOC é o patrocinador do projeto e, portanto a quem deve ser entregue o produto final, resultante deste estudo.

Para isso é peremptório compreender as necessidades do cliente e assim tornar possível o desenvolvimento da solução atendendo satisfatoriamente às suas expectativas.

Esta atividade objetiva definir um nome para o sistema, descrever a finalidade do projeto, resumir o processo padrão adotado, descrever as expectativas, quais as funcionalidades que o projeto do sistema deverá contemplar e ainda identificar se o sistema possuirá interface com algum já existente (Lima 2007).

Com a meta de identificar quais serão as funções do sistema é primeiro necessário definir quais são os seus requisitos. Entende-se como requisito de sistema, uma descrição das necessidades ou desejos para a versão final de um produto.

Iniciamos então este processo com o objetivo de avaliar as inconsistências, ambiguidades, riscos e prioridades dos requisitos classificando-os em dois grupos distintos:

4.2.1. Requisitos funcionais

Sommerville (2003) afirma que os requisitos funcionais são declarações de funções fornecidas pelo sistema. Essas declarações dizem como deve ser a reação do sistema mediante entradas específicas e como deve ser seu comportamento em certas situações. Em algumas destas, requisitos desse tipo também podem expressar restrições explícitas daquilo que o sistema não deve fazer. Neste estudo, identificamos como requisitos funcionais:

- Consultar Resultados: Deve ser permitido através de um formulário realizar a consulta principal do sistema. Este formulário deve conter os campos de seleção da espécie de origem (uma única opção), espécies de destino (uma ou mais opções) e os filtros de Localização Subcelular, Processo Molecular e Tipo de Homologia (todos contendo as opções carregadas pelo banco de dados). Não deve ser permitida a consulta com qualquer um dos campos disponíveis sem preenchimento.

- Exibir Resultados: Os resultados devem ser exibidos em um formato de tabela que permita a paginação dos mesmos e a pesquisa de um tempo em específico dentre eles. Cada coluna da tabela precisa poder ser ordenada e a quantidade de linhas de resultados por página, variando de 10 a 100.

- Exibir detalhes: Ao escolher uma linha de resultado, o sistema deve apresentar todos detalhes disponíveis sobre a comparação e o gene bem como a própria sequência proteica.

- Efetuar Download: O sistema deve permitir descarregar o resultado da consulta em arquivo eletrônico.

- Exibir Estatísticas do Sistema: O sistema deve apresentar as estatísticas dos dados carregados totais, apresentando os quantitativos segregados por espécies e seus respectivos filtros.

- Executar Consulta BLAST+: O sistema deve apresentar a função de incluir uma consulta BLAST+ contra os organismos que fazem parte do estudo e retornar os resultados comparados.

4.2.2. Requisitos não funcionais

Os requisitos não funcionais constituem restrições sobre serviços ou funções fornecidos pelo sistema (Sommerville 2007). Esses tipos de requisitos aparecem de acordo com a necessidade dos usuários seja devido a restrições de orçamento, de políticas organizacionais ou à necessidade de interoperabilidade com outros sistemas, bem como por necessidades do negócio. Requisitos que não incluem interfaces externas, restrições de desempenho, banco de dados, plataforma de desenvolvimento, documentação para o usuário final. Para este estudo, utilizamos:

- Usabilidade: Como o sistema tem como ambiente de uso a Internet, precisa ser desenvolvido com tecnologias de interface que promovam uma experiência agradável com o usuário através de um Navegador multiplataforma.

- Segurança: O sistema deve garantir a segurança dos dados, bem como as permissões de acesso às suas funcionalidades.

- Confiabilidade: O sistema precisa estar disponível 24 horas por dia, 7 dias por semana e estar resguardado com rotina de backup e sob uma política de segurança da informação.

- Desempenho: O sistema precisa ser rápido no processamento das consultas e na exibição da resposta.

- Hardware e Software: Uma vez disponibilizado na Internet, o sistema precisa poder ser acessado dos principais navegadores atuais.

4.3. Modelagem dos dados

A modelagem de dados é uma técnica usada para especificar as regras de negócio e estruturas de dados de um banco de dados. Faz parte do ciclo de desenvolvimento do sistema e é de extrema importância para alcançar o objetivo do projeto. Modelar dados consiste em criar um desenho do sistema, identificando entidades lógicas e suas dependências.

Este processo, também chamado de modelagem de banco de dados, envolve uma série de aplicações teóricas e práticas, visando construir um modelo de dados consistente, não redundante e perfeitamente aplicável em qualquer SGBD moderno.

A metodologia utilizada nesse trabalho seguiu as seguintes etapas:

- **Coleta de dados:** Envolveu obtenção, armazenamento e disponibilização dos dados, conforme descrito no capítulo 3.1. Após a obtenção dos arquivos FASTA, realizamos a ordenação dos dados, que não indicou nenhuma inconsistência e não apresentou valores nulos ou muito discrepantes. Depois, realizamos uma avaliação descritiva, gerando os seguintes totalizadores:

    1. Quantidade de Genomas: 63
    2. Quantidade de Sequências: 279.403
    3. Tamanho total dos arquivos: 115 Mb

- **Tratamento de dados:** Este procedimento foi realizado para gerar uma base de dados consistente em relação ao sistema. Essa etapa permitiu identificar pontuais erros de coleta de dados, além de aumentar o conhecimento do sistema em estudo. Identificamos as informações relevantes e suas posições nos arquivos FASTA e nos outputs dos sistemas que utilizamos para o processamento. Assim, pudemos filtrar o que era necessário e descartar as informações irrelevantes para este estudo.

- **Inferência:** É a identificação da distribuição dos dados processados. Essa distribuição pode ser teórica ou empírica, conforme a aderência de modelos teóricos aos dados. No caso dos dados processados, optamos por uma análise empírica ao comparar os resultados.

Por fim, depois de tratados, iniciamos a identificação e rotulação dos dados conforme as figuras 4.3-1, 4.3-2, 4.3-3 e 4.3-4:

Arquivos Fasta

```
>gi|15607143|ref|NP_214515.1| chromosomal replication initiator protein
DnaA [Mycobacterium tuberculosis H37Rv]
MTDDPGSGFTTVWNAVVSELNGDPKVDDGPSSDANLSAPLTPQQRAWLNLVQPLTIVEGFALLSVPSSFV
QNEIERHLRAPITDALSRRLGHQIQLGVRIAPPATDEADDTTVPPSENPATTSPDTTTDNDEIDDSAAAR
GDNQHSWPSYFTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRFAHAAALAIAEAPARAYNPLFIWGES
GLGKTHLLHAAGNYAQRLFPGMRVKYVSTEEFTNDFINSLRDDRKVAFKRSYRDVDVLLVDDIQFIEGKE
GIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKKAQMER
LAVPDDVLELIASSIERNIRELEGALIRVTAFASLNKTPIDKALAEIVLRDLIADANTMQISAATIMAAT
AEYFDTTVEELRGPGKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAQRKILSEMAERREV
```

Espécie: Mycobacterium tuberculosis H37Rv

Id da Proteína: 15607143

Tipo da Proteína: NP_214515.1

Descrição da Proteína: chromosomal replication initiator protein DnaA

Sequencia:

```
MTDDPGSGFTTVWNAVVSELNGDPKVDDGPSSDANLSAPLTPQQRAWLNLVQPLTIVEGFALLSVPSSFV
QNEIERHLRAPITDALSRRLGHQIQLGVRIAPPATDEADDTTVPPSENPATTSPDTTTDNDEIDDSAAAR
GDNQHSWPSYFTERPHNTDSATAGVTSLNRRYTFDTFVIGASNRFAHAAALAIAEAPARAYNPLFIWGES
GLGKTHLLHAAGNYAQRLFPGMRVKYVSTEEFTNDFINSLRDDRKVAFKRSYRDVDVLLVDDIQFIEGKE
GIQEEFFHTFNTLHNANKQIVISSDRPPKQLATLEDRLRTRFEWGLITDVQPPELETRIAILRKKAQMER
LAVPDDVLELIASSIERNIRELEGALIRVTAFASLNKTPIDKALAEIVLRDLIADANTMQISAATIMAAT
AEYFDTTVEELRGPGKTRALAQSRQIAMYLCRELTDLSLPKIGQAFGRDHTTVMYAQRKILSEMAERREV
```

ID da Espécie: AAA (Atribuímos arbitrariamente uma sigla para identificar internamente a espécie).

Figura 4.3-1: Exemplo de extração de dados de um arquivo Fasta.

Resultado OrthoMCL

Linha: AAA 507419080 AAB 169627617 1.155 c

Espécie 'Query': AAA

Gene 'Query': 507419080

Espécie 'Target': AAB

Gene 'Target': 169627617

Score: 1.155

Relação de Homologia: c

Figura 4.3-2: Exemplo de extração de dados do resultado do processamento utilizando o OrthoMCL.

Resultado PSORT

Linha: AAA 507418943 Cytoplasmic 9.95

Espécie: AAA

Gene: 507418943

Localização: Cytoplasmic

Figura 4.3-3: Exemplo de extração de dados do resultado do processamento utilizando o PSORT.

Resultado Argot2

Linha: 41408515 C GO:0005886 plasma membrane

Espécie: AAA (Atribuímos arbitrariamente uma sigla para identificar internamente a espécie).

Id da Proteína: 41408515

Processo Molecular: C

GO: GO:0005886

Descrição GO: plasma membrane

Figura 4.3-4: Exemplo de extração de dados do resultado do processamento utilizando o Argot2.

Assim, a partir da identificação dos dados, o próximo passo foi relacioná-los através de um modelo lógico visando a construção do banco de dados.

O resultado desta modelagem pôde ser organizado segundo o esquema da figura 4.3-5.

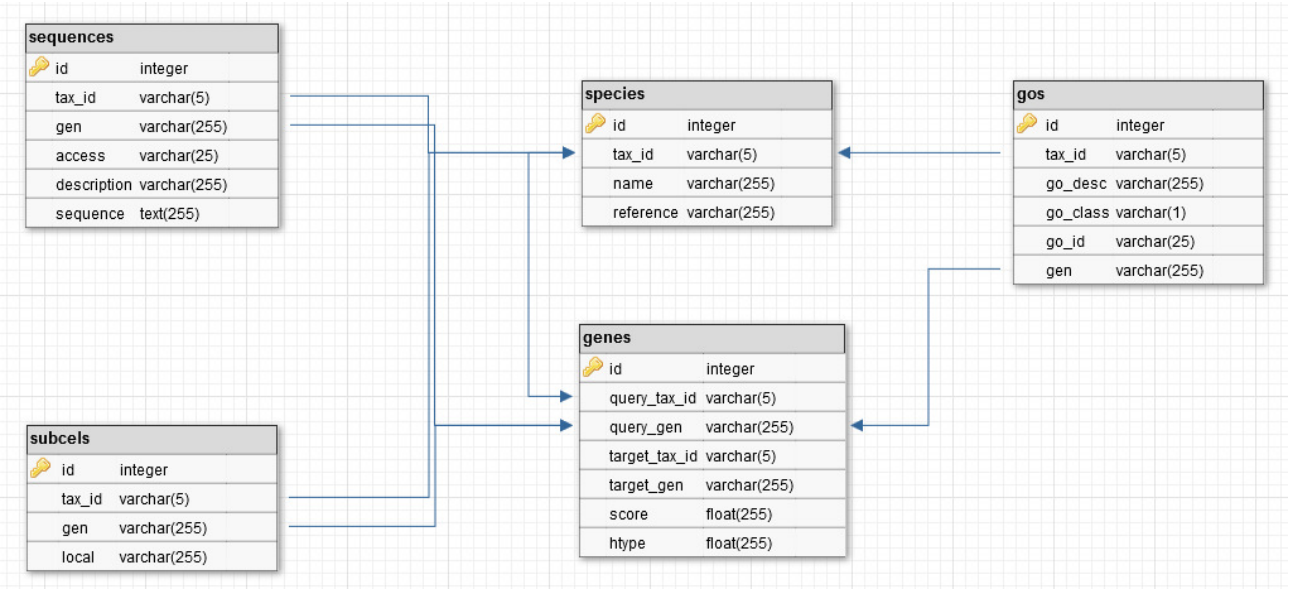


Figura 4.3-5: Modelagem lógica dos dados.

4.4. Definição da tecnologia

Para definir a tecnologia empregada na nova versão do sistema, levamos em consideração a predileção por softwares e ambiente formatados em código aberto e distribuição gratuita, bem como um alinhamento tecnológico com o laboratório onde foi desenvolvido o projeto. Neste sentido, para o desenvolvimento do sistema utilizamos:

4.4.1. Linguagem de programação - Ruby on Rails:

Por ser um framework livre e desenvolvido com o intuito de aumentar velocidade de desenvolvimento e facilidade na programação orientada a banco de dados, optamos por utilizar esta tecnologia, uma vez que é possível criar aplicações com base em estruturas pré-definidas e suporte de uma grande comunidade de desenvolvimento internacional. O Ruby on Rails é um projeto de código aberto escrito na linguagem de programação Ruby, cujos sistemas são desenvolvidos com base no padrão de arquitetura MVC (Model-View-Controller).

4.4.2. Bibliotecas para construção de interface:

- **JQuery** é uma biblioteca JavaScript desenvolvida para simplificar os scripts client side que interagem com o HTML. Usada por cerca de 80% dos 10 mil sites de internet mais visitados do mundo, jQuery é a mais popular das bibliotecas JavaScript de código aberto e possui licença dual, fazendo uso da Licença MIT ou da General Public License versão 2.4.

- **Twitter Bootstrap** é uma coleção de ferramentas para criação de websites e aplicações web utilizando as tecnologias HTML e CSS, que tem como objetivo facilitar o desenvolvimento/manutenção de um projeto utilizando reaproveitamento de código (bibliotecas).

- ***SGDB*** - MySQL versão 5.1.73. É um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (Linguagem de Consulta Estruturada, do inglês *Structured Query Language*) como interface. É atualmente um dos bancos de dados mais populares do mundo por ser reconhecidos a partir de características positivas como a portabilidade, compatibilidade, desempenho, facilidade e bom desempenho, além de ser considerado um software livre levando em consideração que o software que acessa o MySQL seja livre também).

5. Resultado

Os dados primários foram obtidos através de repositório público e as análises realizadas utilizando os softwares resultantes de pesquisas científicas e disponibilizados na internet.

Para isso foi desenvolvida uma nova versão do software agregando e trabalhando os dados referentes aos genomas das 63 micobactérias disponibilizados, através de outros softwares de análise à seguir especificados.

Esta nova versão utiliza um banco de dados relacional, construído a partir dos resultados obtidos com alinhamentos par a par de todas as sequências proteicas preditas codificadas pelos genomas utilizados.

O sistema armazena os parâmetros e valores computados de similaridade entre cada par de sequências alinhadas e também oferece, para cada uma destas proteínas, a sua localização sub-celular predita, sua classificação em COG(s) e a descrição dos genes correspondentes nos genomas.

Através de uma interface construída para ser minimalista e intuitiva, tabelas de pares ou grupos de proteínas homólogas potenciais podem ser geradas dinamicamente com critérios definidos pelo usuário, baseados em diferentes parâmetros de similaridade entre as sequências alinhadas.

As buscas podem ainda ser filtradas de acordo com a localização sub-celular predita das proteínas, com a localização do gene correspondente na fita de DNA e/ou de acordo com a descrição da proteína.

As consultas são feitas de forma maciça e os resultados obtidos podem ser exportados facilitando o processamento e a análise das informações.

Os principais itens a serem abordados foram:

- Os dados de 63 genomas de micobactérias foram obtidos e processados de forma a identificar, entre as espécies que fazem parte deste estudo, genes compartilhados por ancestralidade (ortólogos), genes originados por eventos de duplicação (parálogos), genes co-ortólogos e genes encontrados exclusivamente em espécies ou taxa particulares (genes taxonomicamente

restritos).

- Foram realizadas análises para identificar as relações entre os genes atribuindo filtros de pesquisa à partir da confrontação destes resultados com o processamento de softwares específicos.

- O sistema foi inteiramente reescrito utilizando tecnologia atual, o que permitiu reutilização de código no desenvolvimento, garantindo maior agilidade nas atualizações e eventuais manutenções.

- Foi desenvolvida uma nova interface para a consulta aos dados utilizando uma versão do software BLAST+, que por sua vez tem como base de consulta os genomas disponibilizados no projeto.

Como requisito para acessar o sistema, basta o usuário acessar o endereço - http://157.86.120.137/genomycdb - a partir de qualquer dispositivo conectado à internet.

O sistema reúne os dados processados e relacionados no banco de dados e dispõe de uma interface de consulta desenvolvida para ser simples e funcional.

5.1. Busca de resultados

Ao acessar o endereço eletrônico, o usuário é remetido à tela inicial do sistema (Figura 5.1-1), que apresenta o formulário principal de consulta com os campos de seleção:

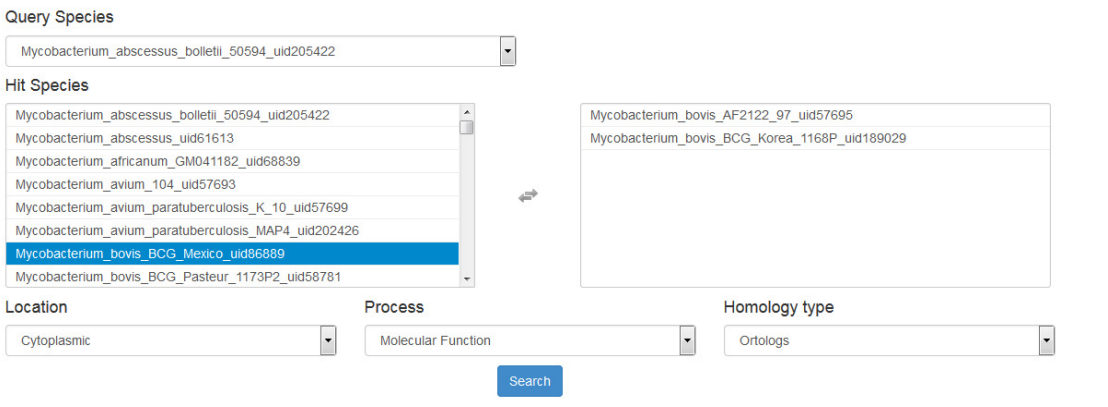


Figura 5.1-1: Tela inicial do sistema.

- **Seleção de espécie *Query*:**

A espécie *query* é a espécie que utilizaremos para comparar a relação de homologia com as espécies *targets*. Este campo (Figura 5.1-2) em formato lista disponibiliza cada uma das espécies contempladas no estudo.

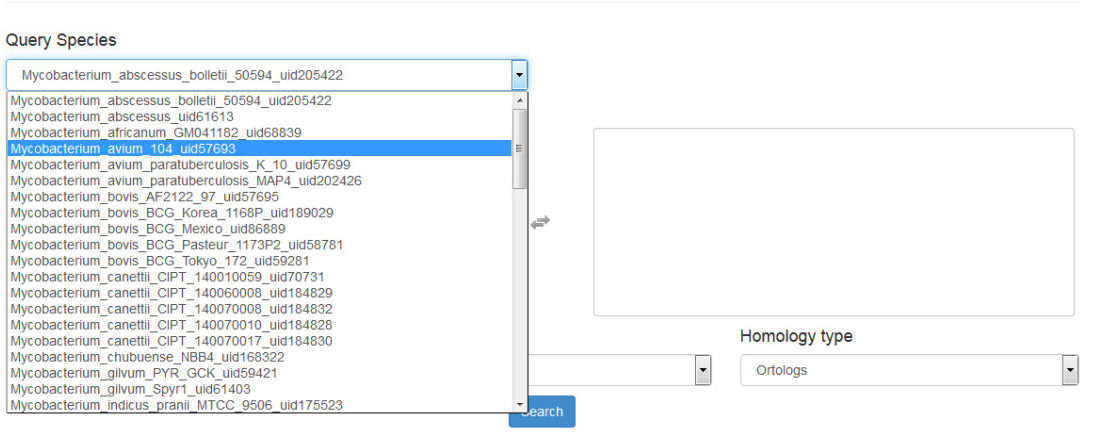


Figura 5.1-2: Seleção da espécie query.

- **Seleção de espécies *Target*:**

As espécies *target* são aquelas que apresentarão a relação de homologia com a espécie *query*. São representadas em uma lista (Figura 5.1-3) contendo todas as espécies disponíveis no sistema e podem ser selecionadas em quantidade desejada pelo usuário.

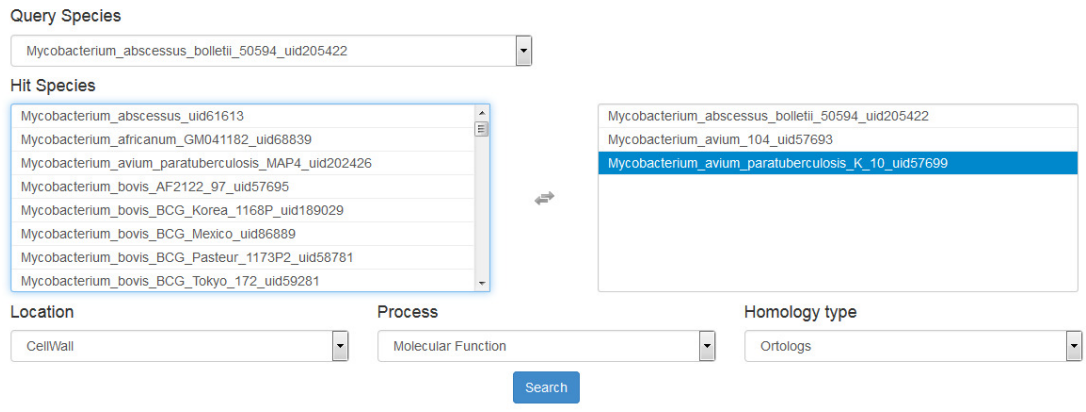


Figura 5.1-3: Seleção da espécie target.

- **Filtro de localização subceluar:**

O filtro de seleção “Location” (Figura 5.1-4) permite ao usuário selecionar entre as opções disponíveis, qual a localização subcelular onde determinado gene da espécie *query* se encontra. Como opções disponibiliza a parede celular, citoplasma, membrana citoplasmática, localização extra celular e localização desconhecida.

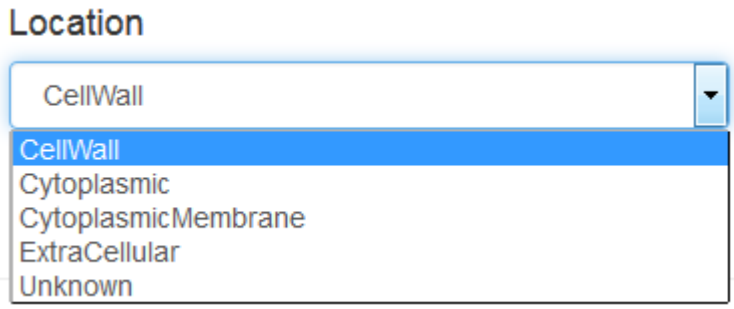


Figura 5.1-4: Seleção do filtro de localização subcelular.

- **Filtro de processo biológico:**

O filtro de seleção *Process* (Figura 5.1-5) permite ao usuário refinar os resultados em relação à função biológica de determinado gene. Como opções é disponibilizada a função molecular, o processo biológico e o componente celular.

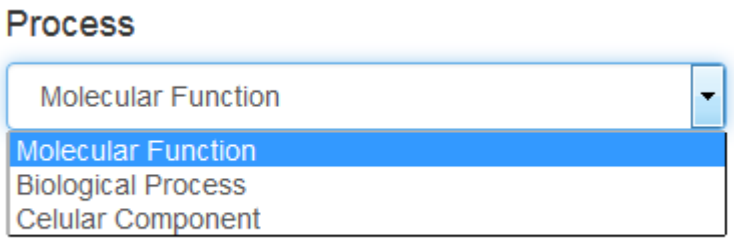


Figura 5.1-5: Seleção do filtro de processo biológico.

- **Relação de homologia:**

O filtro de relação de homologia (Figura 5.1-6) permite ao usuário selecionar o tipo de relação entre os genes comparados. Entre as opções podemos selecionar relações de coortologia, paralogia, ortologia nenhuma homologia, ou seja, genes únicos.

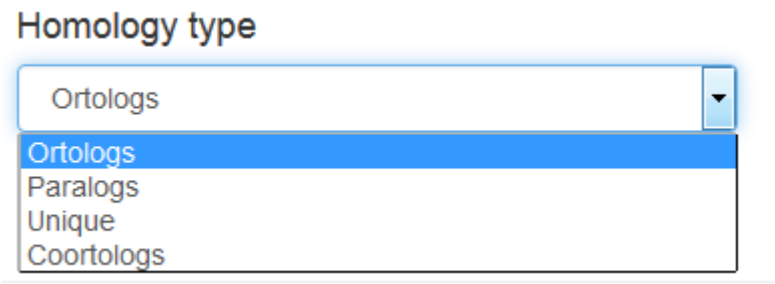


Figura 5.1-6: Seleção do filtro de relação de homologia.

Ao submeter o formulário, o sistema remete o usuário para a seguinte tela de resultados:

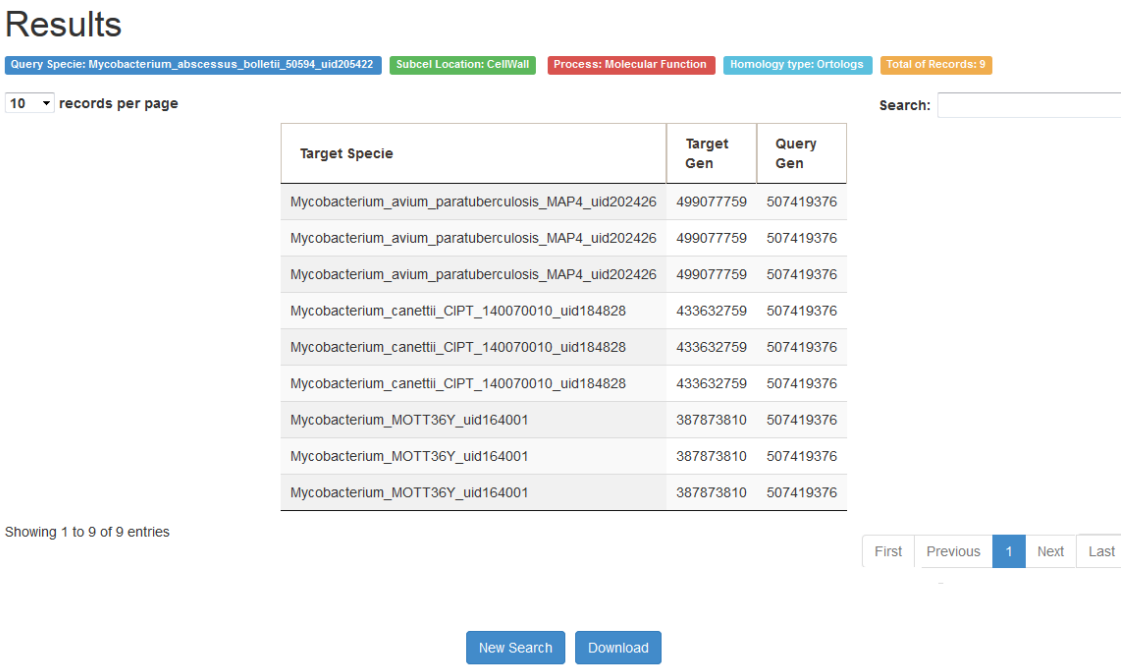


Figura 5.1-7: Tela de resultado do sistema.

Os resultados apresentam os filtros selecionados (Figura 5.1-7), as espécies *query* e *target*, os genes de ambas as espécies e o link para detalhamento dos genes.

Os resultados podem ser ordenados por cada uma das colunas apresentadas e filtrados novamente por qualquer texto inserido no campo *search*, onde o sistema realizará uma sub-busca pelo termo pesquisado, somente entre os resultados retornados.

Para um detalhamento do gene retornado na consulta, o sistema permite a visualização de todas as informações disponibilizadas no banco de dados, relacionados àquele gene.

Para isso, basta clicar na identificação do gene na linha de resultados e o sistema retorna ao usuário os detalhes do gene pesquisado (Figura 5.1-8).

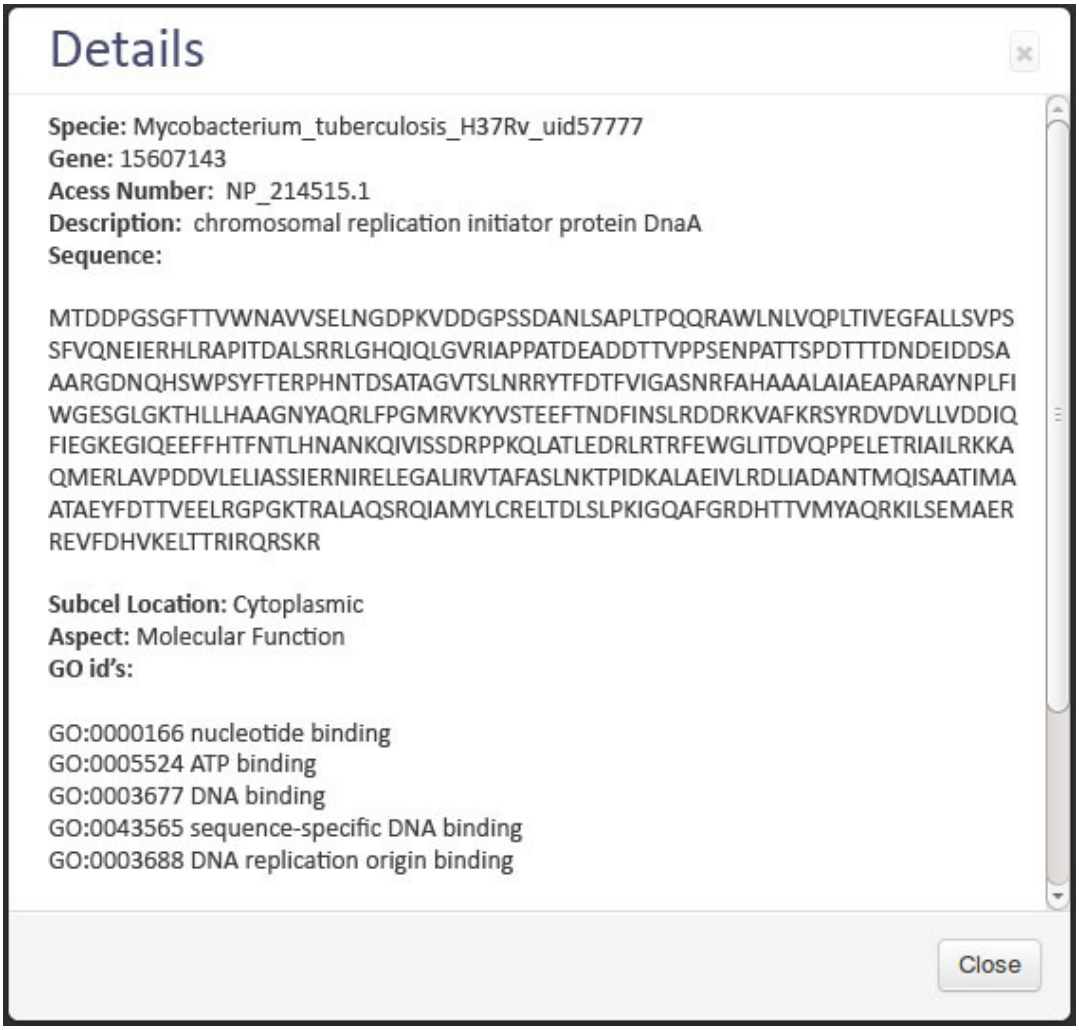


Figura 5.1-8: Seleção de detalhamento do resultado.

- *Download* – O sistema permite ainda a gravação dos resultados em um arquivo com extensão .CSV de forma que o mesmo possa ser trabalhado em planilhas eletrônicas ou mesmo importado para novas análises em outros sistemas.

Para isso, basta o usuário clicar no botão *Download* na tela de resultados e o arquivo será descarregado em seguida (Figura 5.1-9).

```
Mycobacterium_abscessus_uid61613;507418943;169627109;1.178
Mycobacterium_abscessus_uid61613;507418944;169627110;1.178
Mycobacterium_abscessus_uid61613;507418945;169627111;1.178
Mycobacterium_abscessus_uid61613;507418946;169627112;1.178
Mycobacterium_abscessus_uid61613;507418947;169627113;0.823
Mycobacterium_abscessus_uid61613;507418948;169627114;1.178
Mycobacterium_abscessus_uid61613;507418949;169627115;1.178
Mycobacterium_abscessus_uid61613;507418950;169627116;0.148
Mycobacterium_abscessus_uid61613;507418951;169627117;0.517
Mycobacterium_abscessus_uid61613;507418952;169627118;1.178
Mycobacterium_abscessus_uid61613;507418953;169627119;0.471
Mycobacterium_abscessus_uid61613;507418954;169627120;0.519
Mycobacterium_abscessus_uid61613;507418955;169627121;1.178
Mycobacterium_abscessus_uid61613;507418956;169627122;0.293
Mycobacterium_abscessus_uid61613;507418957;169627123;0.407
Mycobacterium_abscessus_uid61613;507418960;169627125;0.889
Mycobacterium_abscessus_uid61613;507418961;169627126;1.178
Mycobacterium_abscessus_uid61613;507418962;169627127;1.178
Mycobacterium_abscessus_uid61613;507418963;169627128;1.178
Mycobacterium_abscessus_uid61613;507418964;169627129;1.178
```

Figura 5.1-9: Exemplo de download *dos resultados.*

5.2. Quantitativos

O processamento dos dados brutos e seus relacionamentos geraram os seguintes quantitativos significativos e uma enorme massa de dados que pode ser interpretada de variadas formas em conjunto.

É possível inclusive detectar a grande quantidade de genes ainda não anotados e relações interessantes de genes com mais de uma anotação ou mesmo relações de genes homólogos com anotações diferentes entre si.

Entre os quantitativos gerados podemos destacar o total de sequências anotadas (Tabela 5.2-1), a localização subcelular dos genes (Tabela 5.2-2), as relações de homologia (Tabela 5.2-3), a descrição dos genes (Tabela 5.2-4) e a classificação no gene ontology (Tabela 5.2-5).

Tabela 5.2-1: Quantitativos de anotação.

Total de Espécies: 63	
Total de Sequencias	279.403
Anotados	219.558
Não anotados	59.945

Tabela 5.2-2: Quantitativos de localização subcelular.

Total de Genes: 279.403	
Parede Celular	802
Citoplasma	141.483
Membrana do Citoplasma	74.416
Extracelular	3.964
Desconhecida	58.738

Tabela 5.2-3: Quantitativos de comparação.

Total de Comparações: 4.169.278	
Ortólogos	4.133.568 comparações de 248.019 genes
Parálogos	14.770 comparações de 8.643 genes
Genes Únicos	11.111 genes
Coortólogos	9.829 comparações de 3.587 genes

Tabela 5.2-4: Quantitativos de genes únicos.

Total de Genes Únicos: 11.111	
Hipotéticos	5.931
Possíveis	16
Putativos	395
Outros	4769

Tabela 5.2-5: Classificação gene ontology.

Total de Ontologias: 219.558	
Função Molecular	190.143
Processo Biológico	196.292
Componente Celular	85.226

5.3. Consultas BLAST

Utilizando o software sequenceserver[6] disponibilizamos uma interface (Figura 5.3-1) para consultas BLAST, onde a base de dados para comparações é formada pelos genomas das espécies contempladas neste estudo.

Para realizar uma consulta, basta inserir no campo de pesquisa a sequencia a ser comparada.

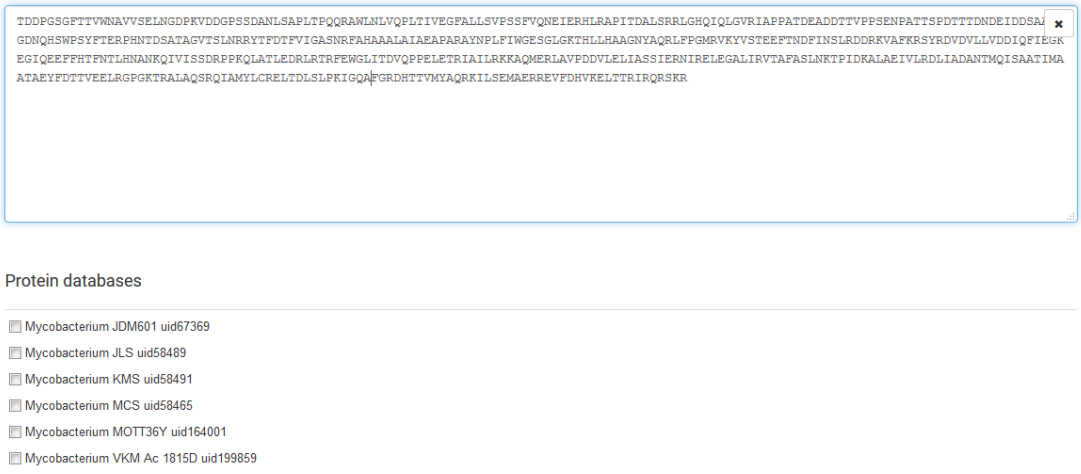
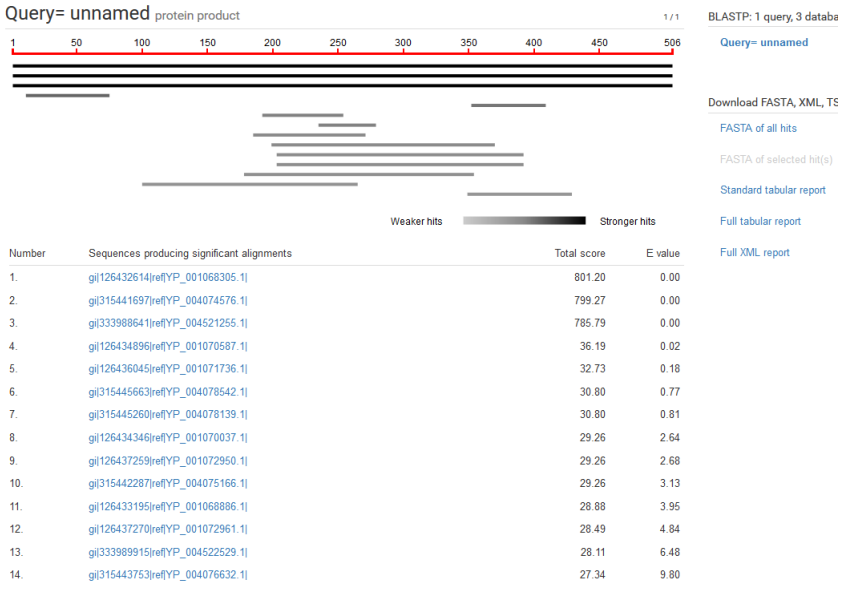


Figura 5.3-1 : Tela de pesquisa BLAST.

[6] Disponível em http://www.sequenceserver.com/ - Acesso em 01/05/2015

O sistema retorna como resultado as sequências com alinhamento significativo, bem como o score e o evalue alcançados (Figuras 5.3-2 e 5.3-3).



Number	Sequences producing significant alignments	Total score	E value
1.	gi|126432614|ref|YP_001068305.1|	801.20	0.00
2.	gi|315441697|ref|YP_004074576.1|	799.27	0.00
3.	gi|333988641|ref|YP_004521255.1|	785.79	0.00
4.	gi|126434896|ref|YP_001070587.1|	36.19	0.02
5.	gi|126436045|ref|YP_001071736.1|	32.73	0.18
6.	gi|315445663|ref|YP_004078542.1|	30.80	0.77
7.	gi|315445260|ref|YP_004078139.1|	30.80	0.81
8.	gi|126434346|ref|YP_001070037.1|	29.26	2.64
9.	gi|126437259|ref|YP_001072950.1|	29.26	2.68
10.	gi|315442287|ref|YP_004075166.1|	29.26	3.13
11.	gi|126433195|ref|YP_001068886.1|	28.88	3.95
12.	gi|126437270|ref|YP_001072961.1|	28.49	4.84
13.	gi|333989915|ref|YP_004522529.1|	28.11	6.48
14.	gi|315443753|ref|YP_004076632.1|	27.34	9.80

Figura 5.3-2: *Resultado da consulta* BLAST.

Figura 5.3-3: Resultado da consulta BLAST – *Continuação.*

O presente estudo analisou os genomas de 63 espécies e entre outras informações, identificou a quantidade de genes destas de forma comparativa.

Na figura 5-4 podemos identificar a quantidade total de genes por espécie.

Na figura 5-5 podemos identificar a quantidade total de genes únicos por espécie.

A figura 5-6 apresenta uma sobreposição das duas informações por espécie.

A figura 5-7 apresenta o número total de genes x número total de genes únicos para o complexo tuberculosis.

O estudo permitiu também, a partir das informações extraídas do banco de dados construído, extrair a mesma informação para apenas um conjunto de espécies, neste caso, o complexo *tuberculosis* (Figura 5-4).

Para exemplificar melhor a capacidade de extração de informação do sistema, realizamos um estudo do core genoma, ou seja, de todos os genes compartilhados por todas as espécies compreendidas no estudo.

O resultado retornou os 31 genes da lista:

```
50S ribosomal protein L4
invasion protein
hypothetical protein MAF_15660
hypothetical protein MAF_16220
hypothetical protein MAF_16480
hypothetical protein MAF_17110
hypothetical protein MAF_17280
isocitrate lyase
hypothetical protein MAF_21900
hypothetical protein MAF_21960
gamma-glutamyl phosphate reductase
signal peptidase I
50S ribosomal protein L19
hypothetical protein MAF_30120
hydrolase
DNA polymerase III subunit gamma/tau
thioredoxin ThiX
succinyl-CoA synthetase subunit alpha
hypothetical protein MASS_1341
HNH endonuclease
nucleoside diphosphate kinase
translation initiation factor IF-3
50S ribosomal protein L20
putative FeS assembly protein SufB
crossover junction endodeoxyribonuclease RuvC
fumarylacetoacetate hydrolase
hypothetical protein MASS_3553
hypothetical protein MASS_3555
putative RNA methyltransferase
```

```
tRNA (guanine-N(7)-)-methyltransferase
chromosome partitioning protein ParB
```

Este resultado apresenta proteínas envolvidas em processos básicos da célula, no entanto, 11 das 31 proteínas apresentam-se como hipotéticas. Esta quantidade pode sugerir uma necessidade destes genes serem melhor anotados.

Esta mesma análise pode ser feita restringindo a quantidade e selecionando as espécies a fim de determinar o conjunto de genes compartilhados. Por isso, realizamos a mesma análise para as espécies contidas no complexo tuberculosis.

O resultado gerado retornou 1056 (Ver anexo 1) genes compartilhados entre as 22 espécies selecionadas.

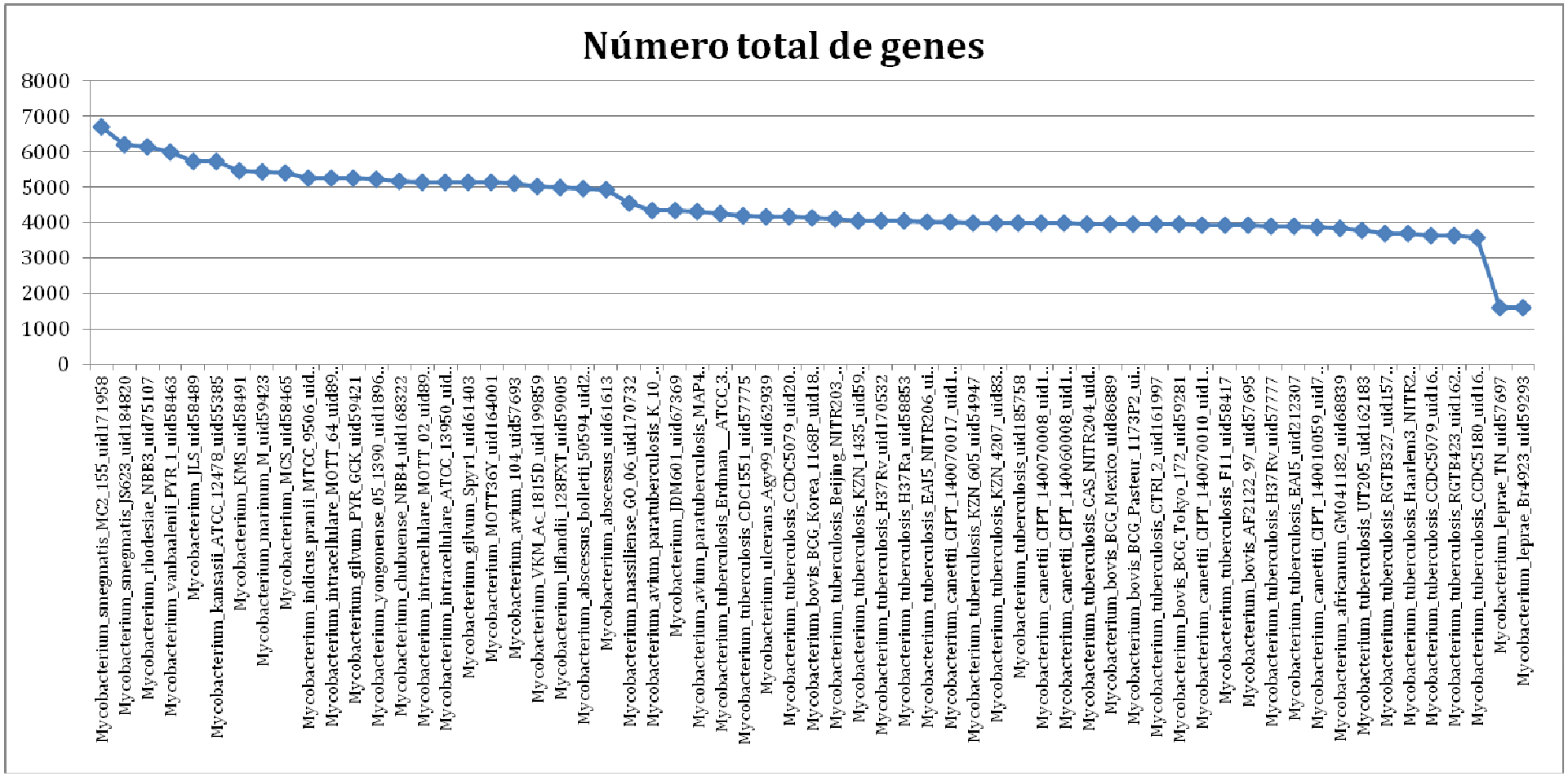


Figura 5-4: Número total de genes por espécie.

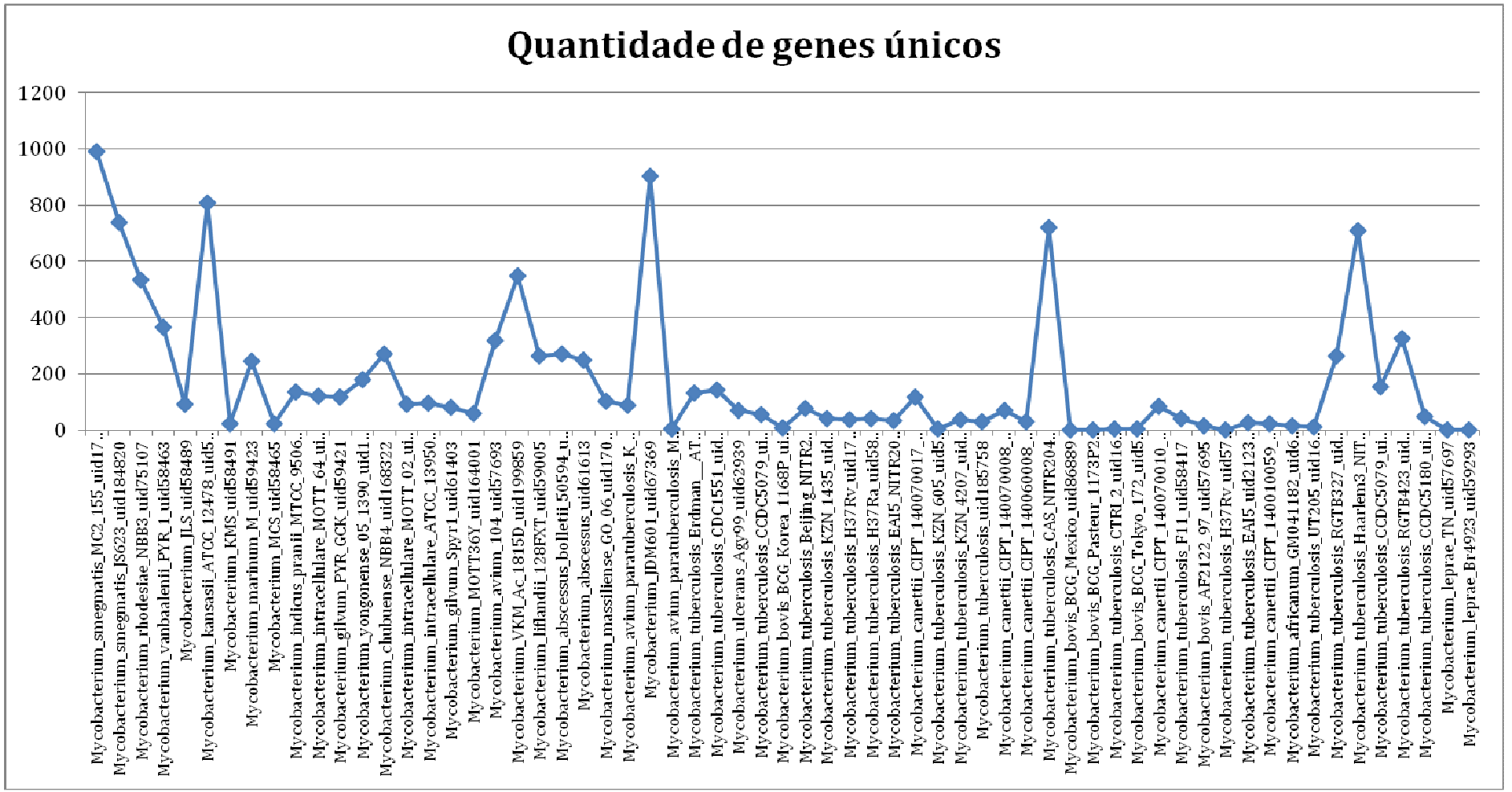


Figura 5-5: Número total de genes únicos por espécie.

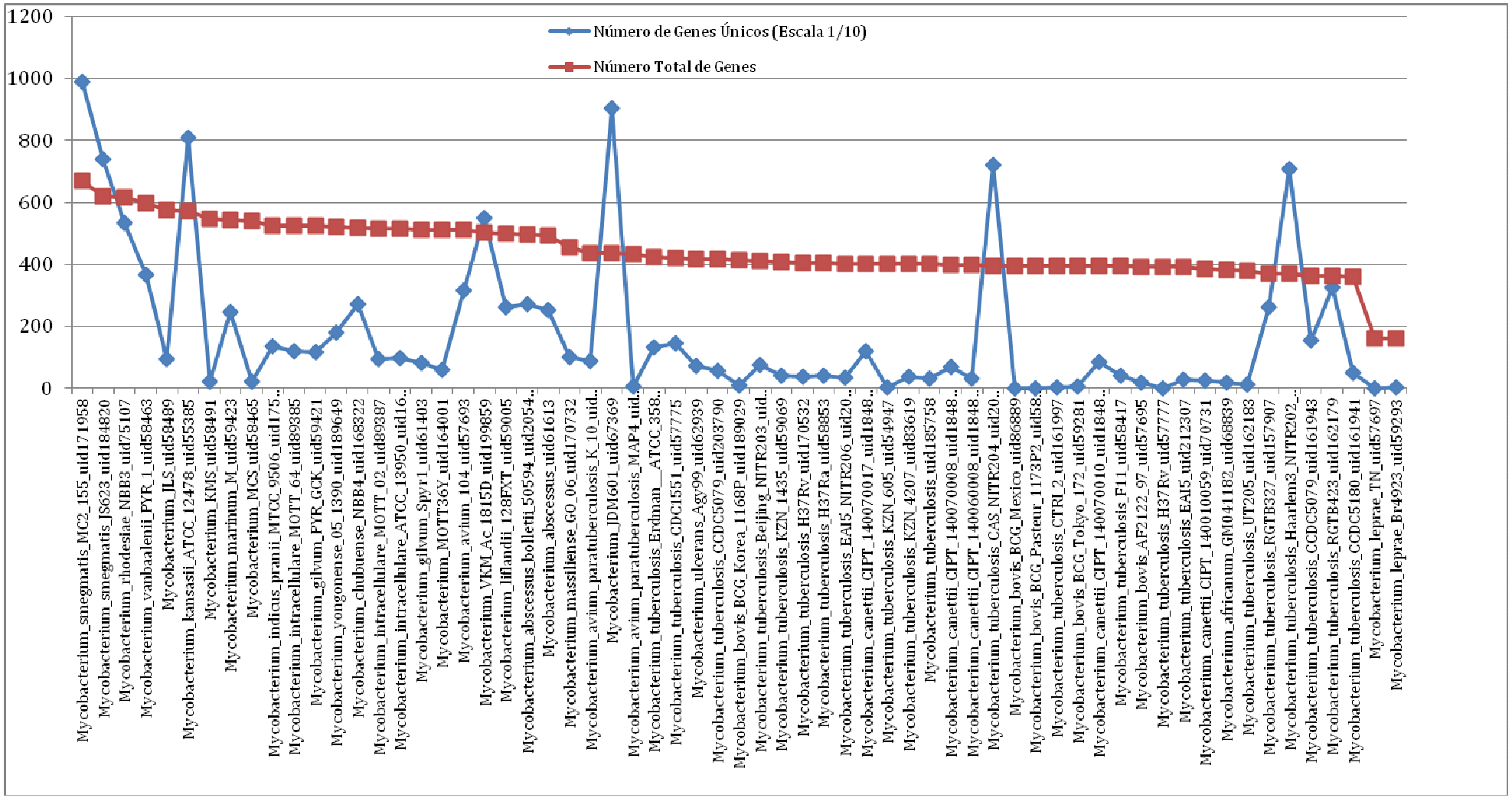


Figura 5-6: Sobreposição de gráficos.

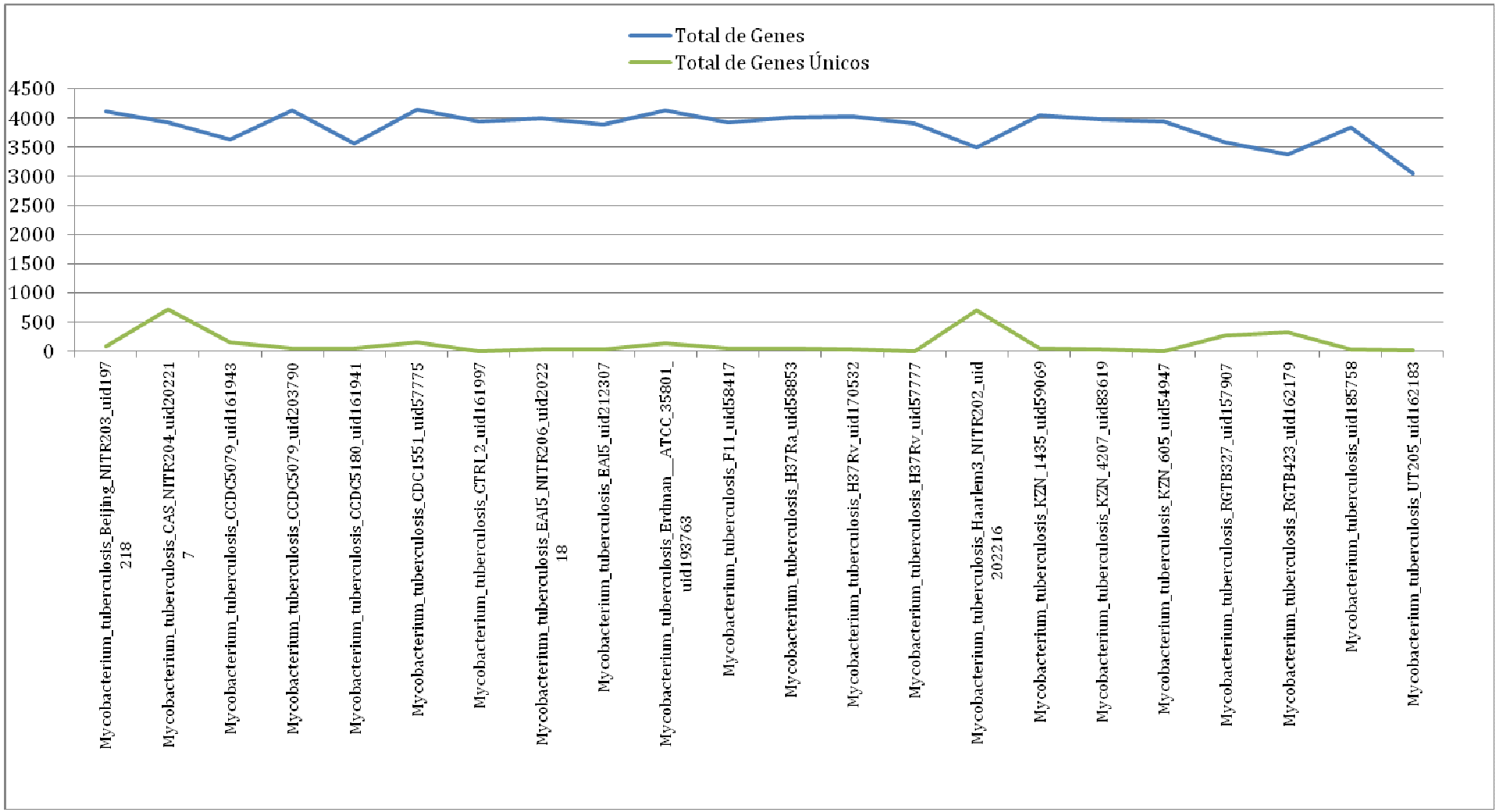


Figura 5-7: Número total de genes x número total de genes únicos para o complexo tuberculosis.

6. Discussão

Dado que este é um sistema funcional e que precisará ser mantido, foi desenvolvido visando escalabilidade e ampliação das suas funções.

Entre as novas oportunidades podemos destacar a inclusão de novas espécies de Micobactérias e a extrapolação para outros gêneros de bactérias, como *Escherichia* ou até mesmo a inclusão de vias metabólicas.

A inclusão destes novos organismos pode seguir os processos descritos neste estudo somando-os aos resultados gerados pelo sistema sem necessidade de novo desenvolvimento específico para uma nova espécie.

Porém, novos desenvolvimentos também estão previstos, principalmente na forma de relatórios amigáveis e novas formas de apresentação dos resultados como, por exemplo, a inclusão de um *heatmap* dinâmico que apresente cores em vez de números em uma matriz de relações de homologia entre espécies.

Esta nova função (Figura 6-1) permitiria uma melhor visualização dos resultados e identificação de áreas de convergência ou distanciamento filogenético.

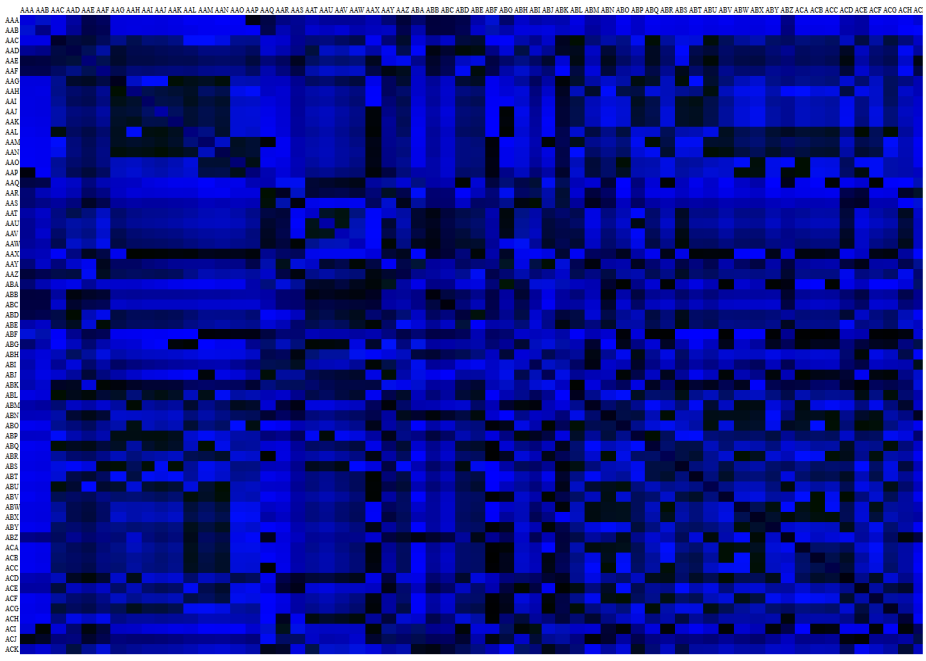


Figura 6-1: Heat map *gerado com os dados do estudo.*

Além do *heatmap*, é prevista também a construção dinâmica de árvores filogenéticas dinâmicas (Figura 6-2) baseadas no relacionamento de homologias identificado nas comparações.

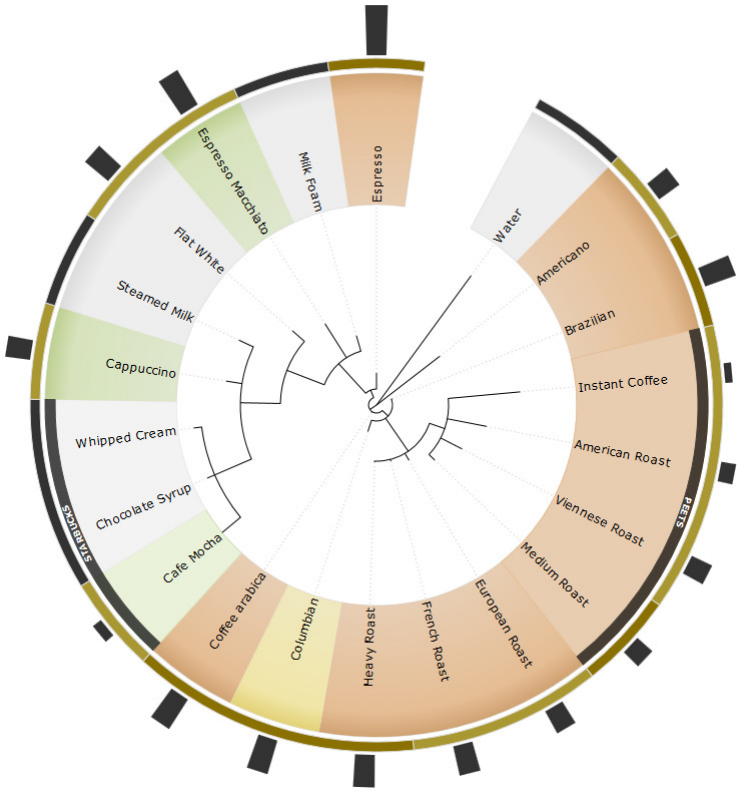


Figura 6-2: Exemplo de árvore filogenética que pode ser implementada pela mesma tecnologia.

Em um comparativo com os outros sistemas descritos neste estudo, podemos destacar como avanços significativos deste projeto a utilização de uma interface minimalista que retorna resultados integrados de diversos softwares de processamento de dados biológicos. Entre estes resultados, a comparação de genes em formato todos contra todos e a determinação e correlação de filtros de natureza biológica.

Outro ponto bastante interessante é a presença da interface BLAST+ que permite a comparação de qualquer sequência contra para os genomas compreendidos neste estudo.

O versão 2.0 do GenoMycDb, apresentada neste trabalho, oferece uma excelente abordagem para a análise comparativa de genomas microbianos.

Referências

Alberts CJ, Smith WCS, Meima A, Wang L, Richardus JH. Potential effect of the World Health Organization's 2011-2015 global leprosy strategy on the prevalence of grade 2 disability: a trend analysis. Bull. World Health Organ. 2011;89(March 2011):487–95.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

Amorim D de S. Fundamentos de Sistemática Filogenética. Holos; 2002.

Attwood TKG a EN-E, E B-R, T.K. Attwood a. G, Eriksson N-E, Bongcam-Rudloff E. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. Bioinforma. - Trends Methodol. [Internet]. 2011;1–36. Recuperado de: http://teacher.bmc.uu.se/SLUBIOINFO2011/SLUBIOINFO2011/Lectures_files/InTech-Concepts_historical_milestones_and_the_central_place_of_bioinformatics_in_modern_biology_a_european_perspective.pdf

Azevedo Junior DP de, Campos R de. Definição de requisitos de software baseada numa arquitetura de modelagem de negócios. Produção. 2008;18(1):26–46.

Brosch R, Gordon S V, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, et al. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc. Natl. Acad. Sci. U. S. A. 2002;99(6):3684–9.

Catanho M, Mascarenhas D, Degrave W, de Miranda AB. BioParser: a tool for processing of sequence similarity analysis reports. [Internet]. Appl. Bioinformatics. 2006. p. 49–53. Recuperado de: http://www.ingentaconnect.com/content/adis/abi/2006/00000005/00000001/art00007\nhttp://www.ncbi.nlm.nih.gov/pubmed/16539538\nhttp://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00822942-200605010-00007

Cohen J. Bioinformatics---an introduction for computer scientists. ACM Comput. Surv. 2004;36(2):122–58.

Cruz POS. Heurísticas para Identificação de Requisitos de Sistemas de Informações a partir de Modelos de Processos. UFRJ; 2004.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. [Internet]. 2002;30(7):1575–84. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101833&tool=pmcentrez&rendertype=abstract

Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, et al. Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene

Ontology terms. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2012;13(Suppl 4):S14. Recuperado de: http://www.biomedcentral.com/1471-2105/13/S4/S14

Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science [Internet]. 2004;306(October):636–40. Recuperado de: http://discovery.ucl.ac.uk/167456/

Fraser CM, Eisen J, Fleischmann RD, Ketchum K a, Peterson S. Comparative genomics and understanding of microbial biology. Emerg. Infect. Dis. [Internet]. 2000;6(5):505–12. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2627966&tool=pmcentrez&rendertype=abstract

Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. Microbiol. Mol. Biol. Rev. 2012;76(1):66–112.

Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, et al. PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics [Internet]. 2005;21(5):617–23. Recuperado de: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti057

Goodfellow M, Minnikin DE. Circumscription of the genus. The Mycobacteria: A Source Book. Nova Iorque: Dekker; 1984.

Griffith DE, Aksamit T, Brown-Elliott BA, Catanzaro A, Daley C, Gordin F, et al. An Official ATS/IDSA Statement: Diagnosis, Treatment, and Prevention of Nontuberculous Mycobacterial Diseases. Am. J. Respir. Crit. Care Med. [Internet]. 2007;175(4):367–416. Recuperado de: http://www.atsjournals.org/doi/abs/10.1164/rccm.200604-571ST

Hartwell L, Hood L, Goldberg M, Reynolds AE, Silver L. Genetics: From Genes to Genomes (Hartwell, Genetics). 4º ed. McGraw-Hill Education; 2010.

Hesper B, Hogeweg P. Bioinformatica: een werkconcept. 1º ed. Kameleon; 1970.

Kay GL, Sergeant MJ, Zhou Z, Chan JZ-M, Millard A, Quick J, et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat. Commun. [Internet]. 2015;6:6717. Recuperado de: http://www.nature.com/doifinder/10.1038/ncomms7717

Kenneth JR, Ray CG. Sherris Medical Microbiology [Internet]. Vasa. 2004. Recuperado de: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf

Laudon K, Laudon JP. Sistema da Informação com Internet. 4º ed. Rio de Janeiro: LTC; 1999.

Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Res [Internet]. 2002;12(3):493–502. Recuperado de: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Cita

tion&list_uids=11875039

Li L, Stoeckert CJJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research. Genome Res. [Internet]. 2003;13(9):2178–89. Recuperado de: http://genome.cshlp.org/cgi/content/full/13/9/2178

Lima RR De. Metodologia de Desenvolvimento de Sistemas de Informação baseados em OO. 2007;1–30.

Madigan MT, Martinko JM, Dunlap P V., Clark DP. Microbiologia de Brock. 12º ed. Artmed; 2010.

Matioli S, Fernandes F. Biologia Molecular e Evolução. 2º ed. Ribeirão Preto: Holos; 2012.

Mudado MDA. Caracterização de expressão gênica e mineração de dados em projetos transcriptoma. Bibliotecadigital.Ufmg.Br [Internet]. 2007; Recuperado de: http://scholar.google.com/scholar?q=intitle:Uso+da+Base+de+Dados+Secund?ria+KOG+como+Ferramenta+para+Caracteriza??o+de+Express?o+G?nica+e+Minera??o+de+Dados+em+Projetos+Transcriptoma#0

Oster W. The Principles and Practice of Medicine: Designed for the Use of Practitioners and Students of Medicine. 1928.

Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics [Internet]. 2015;16(1):236. Recuperado de: http://www.biomedcentral.com/1471-2164/16/236

Pevsner J. Bioinformatics and Functional Genomics. 2º ed. Wiley, Blackwell, organizadores. 2009.

Pinto PGH da R. O estigma do pecado a lepra durante a Idade Média. Physis - Rev. Saúde Coletiva. 1995;133–4.

Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. [Internet]. 2001;29(1):137–40. Recuperado de: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29787&tool=pmcentrez&rendertype=abstract

Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. [Internet]. 2007;35(Database):D61–5. Recuperado de: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl842

Remm M, Storm C, Sonnhammer E. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 2001;314:1041–52.

Rodrigues LC, Lockwood DNJ. Leprosy now: Epidemiology, progress, challenges, and research gaps. Lancet Infect. Dis. [Internet]. Elsevier Ltd; 2011;11(6):464–70.

Recuperado de: http://dx.doi.org/10.1016/S1473-3099(11)70006-8

Silva JRL, Boéchat N. O ressurgimento da tuberculose e o impacto do estudo da imunopatogenia pulmonar. J. Bras. Pneumol. [Internet]. 2004;30(4):388–94. Recuperado de: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1806-37132004000400014&lng=pt&nrm=iso&tlng=pt

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. NIH Public Access. 2010;25(11).

Sommerville I. Engenharia de Software. 8º ed. Addison, Wesley, organizadores. 2007.

Souza LDELIZ, Rhoden SA, Pamphile JA, Biológicas C, Estadual U. A importância das ômicas como ferramentas Para o estudo da prospecção de Microrganismos: perspectivas e desafios. Rev. UNINGÁ. 2014;18(2):16–21.

Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28(1):33–6.

Tatusov RL., Koonin EV., Lipman DJ. A Genomic Perspective on Protein Families. Science (80-.). 2008;278(5338):631–7.

Touchman J. Comparative Genomics [Internet]. Nat. Educ. Knowl. 2010 [citado 12 de maio de 2015]. p. 13. Recuperado de: http://www.nature.com/scitable/knowledge/library/comparative-genomics-13239404

Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. J. Biomed. Inform. 2002;35:142–50.

Anexo I

Lista de genes compartilhados entres as 22 espécies do complexo *tuberculosis*.

339632358	[NAD]-dependent malate oxidoreductase
339632211	1-acylglycerol-3-phosphate O-acyltransferase
339633867	10 KDA culture filtrate antigen EsxB
339632913	16S rRNA processing protein RimM
339633757	19 KDa lipoprotein antigen precursor Lpq
339633605	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
339632040	20-beta-hydroxysteroid dehydrogenase
339632565	3-dehydroquinate synthase
339632993	3-isopropylmalate dehydratase small subunit
507420159	3-oxoacyl-[acyl-carrier-protein] reductase
339630769	30S ribosomal protein S10
339632915	30S ribosomal protein S16
339630779	30S ribosomal protein S17
339633463	30S ribosomal protein S4
339630789	30S ribosomal protein S5
339633447	30S ribosomal protein S9
507420161	4-carboxymuconolactone decarboxylase
339632125	5'-3' exonuclease
339630777	50S ribosomal protein L16
339632910	50S ribosomal protein L19
507421191	50S ribosomal protein L20
507422760	50S ribosomal protein L24
339631076	50S ribosomal protein L25
339630771	50S ribosomal protein L4
339630784	50S ribosomal protein L5
339630787	50S ribosomal protein L6
339631188	6-phosphogluconate dehydrogenase,decarboxylating
339631512	6-phosphogluconolactonase
339631740	ABC transporter
339631050	ABC transporter
339631741	ABC transporter ATP-binding protein
339630215	acetyltransferase
339633317	acid phosphatase
339630890	acyl-ACP desaturase
339631160	acyl-ACP desaturase
339631987	acyl-CoA dehydrogenase
339631733	acyl-CoA dehydrogenase
339630821	acyl-CoA dehydrogenase
339631674	acyl-CoA thioesterase
339633383	acylamidase
339631621	acyltransferase
339633809	acyltransferase
339633810	acyltransferase
339631321	acyltransferase
339633588	adenine glycosylase
339630802	adenylate kinase

507423384 alkyl hydroperoxide reductase AhpD
339632455 alkyl hydroperoxide reductase C
339630407 alpha-D-glucose-1-phosphate thymidylyltransferase
339631288 alternative RNA polymerase sigma factor SIGE
339633232 alternative RNA polymerase sigma-E factor
339633313 amidohydrolase
507419171 amino acid ABC transporter ATP-binding protein
118463223 aminodeoxychorismate synthase component I
339630158 aminotransferase
339632315 aminotransferase
339631665 anthranilate synthase sununit I
339631434 anti-anti-sigma factor RsfA
339633679 anti-anti-sigma factor RsfB
339633090 AraC family transcriptional regulator
339632696 arsenic-transport integral membrane protein ARSA
339630398 ArsR family transcriptional regulator
339633735 ArsR family transcriptional regulator
339632074 ArsR family transcriptional regulator
339633564 arylamine N-acetyltransferase
118465079 arylsulfatase
339630733 arylsulfatase
507421980 AsnC family transcriptional regulator
339630291 AsnC family transcriptional regulator
339632349 AsnC family transcriptional regulator
339633701 aspartokinase
339631372 ATP synthase subunit A
339631374 ATP synthase subunit B
339631379 ATP synthase subunit epsilon
339632489 ATP-dependent CLP protease proteolytic subunit 1
339633370 ATP/GTP-binding protein
339631864 ATPase P
507419069 bacterioferritin BfrB
339632547 bacterioferritin comigratory protein BCP
339633289 bifunctional protein BirA: biotin operon repressor + biotin--[acetyl-CoA-carboxylase] synthetase
339631867 C-5 sterol desaturase
339632668 cadmium inducible protein CadI
339630445 carbon monoxyde dehydrogenase medium subunit
339630444 carbon monoxyde dehydrogenase small subunit
339631777 carboxylase
339631875 CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
339632758 CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
339630505 CDP-diacylglycerol--serine O-phosphatidyltransferase
339633633 cell filamentation protein FIC
339631171 cholesterol dehydrogenase
339633221 chromosome partitioning protein ParA
507423892 chromosome partitioning protein ParB
339632524 citrate lyase subunit beta
339630961 citrate synthase
507421758 class A beta-lactamase
339633550 CoA-transferase subunit beta
339632236 cobalamin 5'-phosphate synthase
507419272 cobalamin synthesis protein

169628921	Conserved hypothetical protein
169631930	Conserved hypothetical protein
169629165	Conserved hypothetical protein (transglutaminase?)
507421767	crossover junction endodeoxyribonuclease RuvC
339632022	cutinase precursor
339633456	cutinase precursor CUT3
339631661	cyclase
339633385	cyclase
339632883	cytochrome C biogenesis protein DipZ
339632221	cytochrome C oxidase subunit III
339630401	cytochrome P450
339633067	cytochrome P450
339631839	cytochrome P450
339631933	cytochrome p450 140 CYP140
339632987	D-alanine--D-alanine ligase
339632917	D-alanyl-D-alanine carboxypeptidase
507420772	D-alanyl-D-alanine dipeptidase
339631958	D-amino acid oxidase
507423116	D-tyrosyl-tRNA(Tyr) deacylase
339632944	daunorubicin-DIM-transport integral membrane protein ABC transporter DRRB
339632945	daunorubicin-DIM-transport integral membrane protein ABC transporter DRRC
339633718	dehydrogenase
339630908	dehydrogenase
339630767	dehydrogenase
339630508	dehydrogenase/reductase
339633074	DeoR family transcriptional regulator
339633606	dihydroneopterin aldolase
339633607	dihydropteroate synthase
507423346	dimethyladenosine transferase
339631072	dimethyladenosine transferase
339630088	DNA gyrase subunit B
339633273	DNA methylase
507419248	DNA polymerase III subunit gamma/tau
339631457	DNA-directed RNA polymerase subunit omega
339631690	drug efflux membrane protein
339633469	dTDP-glucose 4,6-dehydratase RMLB
339633305	endonuclease VIII NEI
339630700	enoyl-CoA hydratase
339633024	ESAT-6 like protein ESXQ
339633450	ESAT-6 like protein ESXU
339632306	esterase
339630296	esterase
339630299	esterase
31794920	excisionase
339632335	excisionase
339632680	exported hypothetical protein
339631252	fatty-acid--CoA ligase
339632948	fatty-acid-CoA ligase
339632955	fatty-acid-CoA ligase
339630917	fatty-acid-CoA ligase
339630473	fatty-acid-CoA ligase
339631593	fatty-acid-CoA ligase
339632618	fatty-acid-CoA ligase

339631244	ferredoxin
339632044	ferredoxin FDXA
339630758	ferredoxin reductase
507421334	ferric uptake regulation protein FurA
339632385	ferric uptake regulation protein FURB
507420788	fluoroquinolones export permease protein
507420789	fluoroquinolones export permease protein
339630168	formate hydrogenlyase
507420247	formyltetrahydrofolate deformylase
507423233	fructose-bisphosphate aldolase
339631624	fumarate reductase
339631625	fumarate reductase
339631623	fumarate reductase iron-sulfur subunit
507422186	fumarylacetoacetate hydrolase
507423852	GABA permease GabP
339632454	gamma-glutamyl phosphate reductase
507419860	gap like protein
339632498	globin GlbO
339631396	glucanase
339630156	glutamine ABC transporter ATP-binding protein
339632508	glycerol-3-phosphate acyltransferase
339630391	glycerophosphoryl diester phosphodiesterase
339631879	glycine cleavage system protein H
339632238	glycine cleavage system protein T
339631440	glycolipid sulfotransferase
118463402	glycosyl hydrolase 3
339633623	glycosyltransferase
339631597	glycosyltransferase
339630126	GntR family transcriptional regulator
339630654	GntR family transcriptional regulator
339633608	GTP cyclohydrolase
339631456	guanylate kinase
339632607	haloalkane dehalogenase DHAA (1-chlorohexane halidohydrolase)
339632071	heat shock protein HspX
339631368	HemK protein
339631656	histidinol-phosphate aminotransferase
507420605	HNH endonuclease
339632549	holo-ACP synthase
339631365	homoserine kinase
339630169	hydrogenase
339633668	hydrolase
339632777	hydrolase
339632609	hydroxyacylglutathione hydrolase
169627876	Hypothetical protein MAB_0775
339630093	hypothetical protein MAF_00100
339630107	hypothetical protein MAF_00240
339630110	hypothetical protein MAF_00270
339630111	hypothetical protein MAF_00280
339630112	hypothetical protein MAF_00290
339630113	hypothetical protein MAF_00300
339630120	hypothetical protein MAF_00370
339630123	hypothetical protein MAF_00400
339630142	hypothetical protein MAF_00590

339630148 hypothetical protein MAF_00650
339630162 hypothetical protein MAF_00790
339630204 hypothetical protein MAF_01220
339630206 hypothetical protein MAF_01240
339630210 hypothetical protein MAF_01280
339630211 hypothetical protein MAF_01290
339630212 hypothetical protein MAF_01300
339630222 hypothetical protein MAF_01400
339630223 hypothetical protein MAF_01410
339630245 hypothetical protein MAF_01640
339630246 hypothetical protein MAF_01650
339630248 hypothetical protein MAF_01680
339630256 hypothetical protein MAF_01760
339630258 hypothetical protein MAF_01780
339630266 hypothetical protein MAF_01860
339630268 hypothetical protein MAF_01890
339630273 hypothetical protein MAF_01940
339630278 hypothetical protein MAF_02000
339630279 hypothetical protein MAF_02010
339630280 hypothetical protein MAF_02020
339630288 hypothetical protein MAF_02100
339630298 hypothetical protein MAF_02200
339630317 hypothetical protein MAF_02410
339630323 hypothetical protein MAF_02470, partial
339630334 hypothetical protein MAF_02590
339630340 hypothetical protein MAF_02650
339630351 hypothetical protein MAF_02770
339630357 hypothetical protein MAF_02840
339630364 hypothetical protein MAF_02910
339630367 hypothetical protein MAF_02940
339630374 hypothetical protein MAF_03010
339630383 hypothetical protein MAF_03110
339630384 hypothetical protein MAF_03120
339630388 hypothetical protein MAF_03160
339630394 hypothetical protein MAF_03220
339630397 hypothetical protein MAF_03250
339630403 hypothetical protein MAF_03320
339630412 hypothetical protein MAF_03420
339630418 hypothetical protein MAF_03490
339630420 hypothetical protein MAF_03510
339630436 hypothetical protein MAF_03680
339630438 hypothetical protein MAF_03700
339630441 hypothetical protein MAF_03730
339630459 hypothetical protein MAF_03920
339630463 hypothetical protein MAF_03960
339630464 hypothetical protein MAF_03970
339630465 hypothetical protein MAF_03980
339630467 hypothetical protein MAF_04000
339630488 hypothetical protein MAF_04210
339630495 hypothetical protein MAF_04280
339630500 hypothetical protein MAF_04330
339630503 hypothetical protein MAF_04360
339630510 hypothetical protein MAF_04430

339630516 hypothetical protein MAF_04500
339630523 hypothetical protein MAF_04570
339630528 hypothetical protein MAF_04620
339630530 hypothetical protein MAF_04640
339630533 hypothetical protein MAF_04670
339630535 hypothetical protein MAF_04690
339630543 hypothetical protein MAF_04770
339630557 hypothetical protein MAF_04910
339630575 hypothetical protein MAF_05090
339630579 hypothetical protein MAF_05130
339630586 hypothetical protein MAF_05200
339630603 hypothetical protein MAF_05380
339630609 hypothetical protein MAF_05450
339630626 hypothetical protein MAF_05630
339630649 hypothetical protein MAF_05870
339630652 hypothetical protein MAF_05900
339630653 hypothetical protein MAF_05910
339630663 hypothetical protein MAF_06020
339630666 hypothetical protein MAF_06050
339630677 hypothetical protein MAF_06160
339630681 hypothetical protein MAF_06200
339630686 hypothetical protein MAF_06250
339630692 hypothetical protein MAF_06310
339630693 hypothetical protein MAF_06320
339630705 hypothetical protein MAF_06440
339630728 hypothetical protein MAF_06670
339630731 hypothetical protein MAF_06700
339630735 hypothetical protein MAF_06740
339630749 hypothetical protein MAF_06880
339630750 hypothetical protein MAF_06890
339630762 hypothetical protein MAF_07010
339630811 hypothetical protein MAF_07530
339630830 hypothetical protein MAF_07720
339630845 hypothetical protein MAF_07870
339630846 hypothetical protein MAF_07880
339630865 hypothetical protein MAF_08090
339630868 hypothetical protein MAF_08120
339630870 hypothetical protein MAF_08140
339630879 hypothetical protein MAF_08230
339630883 hypothetical protein MAF_08270
339630888 hypothetical protein MAF_08320
339630904 hypothetical protein MAF_08480
339630912 hypothetical protein MAF_08560
339630921 hypothetical protein MAF_08650
339630922 hypothetical protein MAF_08660
339630940 hypothetical protein MAF_08840
339630942 hypothetical protein MAF_08860
339630944 hypothetical protein MAF_08880
339630952 hypothetical protein MAF_08960
339630958 hypothetical protein MAF_09020
339630975 hypothetical protein MAF_09210
339630982 hypothetical protein MAF_09280
339631003 hypothetical protein MAF_09500

339631016 hypothetical protein MAF_09630
339631022 hypothetical protein MAF_09690
339631025 hypothetical protein MAF_09720
339631027 hypothetical protein MAF_09740
339631029 hypothetical protein MAF_09760
339631053 hypothetical protein MAF_10000
339631054 hypothetical protein MAF_10010
339631055 hypothetical protein MAF_10020
339631062 hypothetical protein MAF_10090
339631068 hypothetical protein MAF_10160
339631077 hypothetical protein MAF_10260
339631082 hypothetical protein MAF_10310
339631111 hypothetical protein MAF_10600
339631121 hypothetical protein MAF_10700
339631127 hypothetical protein MAF_10760
339631128 hypothetical protein MAF_10770
339631136 hypothetical protein MAF_10850
339631142 hypothetical protein MAF_10910
339631145 hypothetical protein MAF_10940
339631163 hypothetical protein MAF_11120
339631166 hypothetical protein MAF_11150
339631220 hypothetical protein MAF_11720
339631223 hypothetical protein MAF_11750
339631238 hypothetical protein MAF_11900
339631257 hypothetical protein MAF_12090
339631259 hypothetical protein MAF_12110
339631271 hypothetical protein MAF_12230
339631276 hypothetical protein MAF_12280
339631282 hypothetical protein MAF_12340
339631295 hypothetical protein MAF_12470
339631301 hypothetical protein MAF_12530
339631309 hypothetical protein MAF_12610
339631331 hypothetical protein MAF_12840
339631336 hypothetical protein MAF_12890
339631338 hypothetical protein MAF_12910
339631339 hypothetical protein MAF_12920
339631343 hypothetical protein MAF_12960
339631352 hypothetical protein MAF_13050
339631357 hypothetical protein MAF_13100
339631360 hypothetical protein MAF_13130
339631371 hypothetical protein MAF_13250
339631380 hypothetical protein MAF_13340
339631389 hypothetical protein MAF_13440
339631406 hypothetical protein MAF_13610
339631408 hypothetical protein MAF_13630
339631410 hypothetical protein MAF_13650
339631411 hypothetical protein MAF_13660
339631416 hypothetical protein MAF_13710
339631421 hypothetical protein MAF_13760
339631426 hypothetical protein MAF_13810
339631431 hypothetical protein MAF_13860
339631432 hypothetical protein MAF_13870
339631435 hypothetical protein MAF_13900

339631436 hypothetical protein MAF_13910
339631437 hypothetical protein MAF_13920
339631439 hypothetical protein MAF_13940
339631464 hypothetical protein MAF_14190
339631468 hypothetical protein MAF_14230
339631484 hypothetical protein MAF_14390
339631485 hypothetical protein MAF_14400
339631486 hypothetical protein MAF_14410
339631488 hypothetical protein MAF_14430
339631491 hypothetical protein MAF_14460
339631495 hypothetical protein MAF_14500
339631498 hypothetical protein MAF_14530
339631500 hypothetical protein MAF_14550
339631506 hypothetical protein MAF_14610
339631511 hypothetical protein MAF_14660
339631526 hypothetical protein MAF_14810
339631533 hypothetical protein MAF_14880
339631544 hypothetical protein MAF_14990
339631557 hypothetical protein MAF_15120
339631560 hypothetical protein MAF_15150
339631564 hypothetical protein MAF_15190
339631571 hypothetical protein MAF_15260
339631572 hypothetical protein MAF_15270
339631586 hypothetical protein MAF_15410
339631589 hypothetical protein MAF_15440
339631590 hypothetical protein MAF_15450
339631602 hypothetical protein MAF_15580
339631610 hypothetical protein MAF_15660
339631612 hypothetical protein MAF_15680
339631613 hypothetical protein MAF_15690
339631617 hypothetical protein MAF_15730
339631629 hypothetical protein MAF_15850
339631632 hypothetical protein MAF_15880
339631653 hypothetical protein MAF_16090
339631654 hypothetical protein MAF_16100
339631666 hypothetical protein MAF_16220
339631671 hypothetical protein MAF_16270
339631680 hypothetical protein MAF_16360
339631691 hypothetical protein MAF_16470
339631692 hypothetical protein MAF_16480
339631704 hypothetical protein MAF_16610
339631724 hypothetical protein MAF_16870
339631727 hypothetical protein MAF_16900
339631730 hypothetical protein MAF_16930
339631739 hypothetical protein MAF_17020
339631744 hypothetical protein MAF_17070
339631748 hypothetical protein MAF_17110
339631751 hypothetical protein MAF_17150
339631762 hypothetical protein MAF_17260
339631764 hypothetical protein MAF_17280
339631775 hypothetical protein MAF_17390
339631788 hypothetical protein MAF_17520
339631792 hypothetical protein MAF_17570

339631801 hypothetical protein MAF_17670
339631802 hypothetical protein MAF_17680
339631832 hypothetical protein MAF_18000
339631834 hypothetical protein MAF_18020
339631836 hypothetical protein MAF_18040
339631851 hypothetical protein MAF_18190
339631863 hypothetical protein MAF_18320
339631866 hypothetical protein MAF_18350
339631868 hypothetical protein MAF_18370
339631878 hypothetical protein MAF_18470
339631891 hypothetical protein MAF_18600
339631898 hypothetical protein MAF_18670
339631900 hypothetical protein MAF_18690
339631913 hypothetical protein MAF_18820
339631914 hypothetical protein MAF_18830
339631916 hypothetical protein MAF_18850
339631923 hypothetical protein MAF_18920, partial
339631924 hypothetical protein MAF_18930
339631926 hypothetical protein MAF_18950
339631938 hypothetical protein MAF_19070
339631940 hypothetical protein MAF_19090
339631951 hypothetical protein MAF_19210
339631952 hypothetical protein MAF_19220
339631956 hypothetical protein MAF_19260
339631957 hypothetical protein MAF_19270
339631959 hypothetical protein MAF_19290
339631960 hypothetical protein MAF_19300
339631966 hypothetical protein MAF_19360
339631973 hypothetical protein MAF_19430
339631975 hypothetical protein MAF_19450
339631977 hypothetical protein MAF_19470
339632000 hypothetical protein MAF_19710
339632005 hypothetical protein MAF_19760
339632007 hypothetical protein MAF_19780
339632009 hypothetical protein MAF_19800
339632014 hypothetical protein MAF_19850
339632020 hypothetical protein MAF_19930
339632027 hypothetical protein MAF_20000
339632039 hypothetical protein MAF_20130
339632041 hypothetical protein MAF_20150
339632043 hypothetical protein MAF_20170
339632049 hypothetical protein MAF_20240
339632052 hypothetical protein MAF_20270
339632053 hypothetical protein MAF_20280
339632055 hypothetical protein MAF_20300
339632056 hypothetical protein MAF_20310
339632059 hypothetical protein MAF_20340
339632060 hypothetical protein MAF_20350
339632066 hypothetical protein MAF_20410
339632072 hypothetical protein MAF_20470
339632076 hypothetical protein MAF_20510
339632077 hypothetical protein MAF_20520
339632082 hypothetical protein MAF_20570

339632086 hypothetical protein MAF_20610
339632094 hypothetical protein MAF_20690
339632116 hypothetical protein MAF_20940
339632119 hypothetical protein MAF_20970
339632146 hypothetical protein MAF_21250
339632166 hypothetical protein MAF_21460
339632169 hypothetical protein MAF_21490
339632174 hypothetical protein MAF_21540
339632176 hypothetical protein MAF_21560
339632200 hypothetical protein MAF_21820
339632203 hypothetical protein MAF_21850
339632204 hypothetical protein MAF_21860
339632208 hypothetical protein MAF_21900
339632209 hypothetical protein MAF_21910
339632210 hypothetical protein MAF_21920
339632214 hypothetical protein MAF_21960
339632215 hypothetical protein MAF_21970
339632217 hypothetical protein MAF_22000
339632227 hypothetical protein MAF_22100
339632231 hypothetical protein MAF_22140
339632247 hypothetical protein MAF_22310
339632265 hypothetical protein MAF_22500
339632266 hypothetical protein MAF_22510
339632268 hypothetical protein MAF_22530
339632270 hypothetical protein MAF_22550
339632279 hypothetical protein MAF_22640
339632283 hypothetical protein MAF_22680
339632295 hypothetical protein MAF_22800
339632308 hypothetical protein MAF_22940
339632314 hypothetical protein MAF_23000
339632316 hypothetical protein MAF_23020
339632318 hypothetical protein MAF_23040
339632325 hypothetical protein MAF_23120
339632332 hypothetical protein MAF_23190
339632339 hypothetical protein MAF_23260
339632348 hypothetical protein MAF_23350
339632352 hypothetical protein MAF_23390
339632355 hypothetical protein MAF_23420
339632356 hypothetical protein MAF_23430
339632361 hypothetical protein MAF_23490
339632362 hypothetical protein MAF_23500
339632366 hypothetical protein MAF_23540
339632373 hypothetical protein MAF_23610
339632386 hypothetical protein MAF_23740
339632421 hypothetical protein MAF_24090
339632430 hypothetical protein MAF_24180
339632433 hypothetical protein MAF_24210
339632434 hypothetical protein MAF_24220
339632442 hypothetical protein MAF_24300
339632447 hypothetical protein MAF_24350
339632450 hypothetical protein MAF_24380
339632452 hypothetical protein MAF_24400
339632501 hypothetical protein MAF_24900

339632506 hypothetical protein MAF_24950
339632520 hypothetical protein MAF_25090
339632533 hypothetical protein MAF_25220
339632540 hypothetical protein MAF_25290
339632553 hypothetical protein MAF_25420
339632556 hypothetical protein MAF_25450
339632569 hypothetical protein MAF_25580
339632574 hypothetical protein MAF_25630
339632576 hypothetical protein MAF_25650
339632598 hypothetical protein MAF_25870
339632599 hypothetical protein MAF_25880
339632602 hypothetical protein MAF_25910
339632603 hypothetical protein MAF_25920
339632616 hypothetical protein MAF_26050
339632624 hypothetical protein MAF_26130
339632625 hypothetical protein MAF_26140
339632627 hypothetical protein MAF_26160
339632649 hypothetical protein MAF_26390
339632653 hypothetical protein MAF_26450
339632654 hypothetical protein MAF_26460
339632657 hypothetical protein MAF_26490
339632663 hypothetical protein MAF_26550
339632695 hypothetical protein MAF_26870
339632705 hypothetical protein MAF_26970
339632706 hypothetical protein MAF_26980
339632716 hypothetical protein MAF_27080
339632721 hypothetical protein MAF_27130
339632726 hypothetical protein MAF_27180
339632729 hypothetical protein MAF_27210
339632741 hypothetical protein MAF_27330
339632747 hypothetical protein MAF_27390
339632763 hypothetical protein MAF_27560
339632764 hypothetical protein MAF_27570
339632769 hypothetical protein MAF_27620
339632771 hypothetical protein MAF_27640
339632774 hypothetical protein MAF_27670
339632783 hypothetical protein MAF_27760
339632805 hypothetical protein MAF_27990
339632806 hypothetical protein MAF_28000
339632809 hypothetical protein MAF_28030
339632812 hypothetical protein MAF_28060
339632814 hypothetical protein MAF_28080
339632822 hypothetical protein MAF_28160
339632824 hypothetical protein MAF_28180
339632831 hypothetical protein MAF_28270
339632833 hypothetical protein MAF_28290
339632872 hypothetical protein MAF_28680
339632878 hypothetical protein MAF_28740
339632881 hypothetical protein MAF_28770
339632884 hypothetical protein MAF_28800
339632907 hypothetical protein MAF_29050
339632933 hypothetical protein MAF_29310
339632953 hypothetical protein MAF_29510

339632960 hypothetical protein MAF_29580
339632961 hypothetical protein MAF_29590
339632975 hypothetical protein MAF_29740
339632979 hypothetical protein MAF_29780
339632989 hypothetical protein MAF_29880
339632996 hypothetical protein MAF_29950
339633006 hypothetical protein MAF_30050
339633012 hypothetical protein MAF_30110
339633013 hypothetical protein MAF_30120
339633015 hypothetical protein MAF_30140
339633041 hypothetical protein MAF_30400
339633044 hypothetical protein MAF_30430
339633072 hypothetical protein MAF_30710
339633075 hypothetical protein MAF_30740
339633077 hypothetical protein MAF_30760
339633078 hypothetical protein MAF_30770
339633083 hypothetical protein MAF_30820
339633089 hypothetical protein MAF_30880
339633111 hypothetical protein MAF_31100
339633128 hypothetical protein MAF_31270
339633130 hypothetical protein MAF_31300
339633141 hypothetical protein MAF_31420
339633173 hypothetical protein MAF_31740
339633186 hypothetical protein MAF_31870
339633188 hypothetical protein MAF_31890
339633190 hypothetical protein MAF_31910
339633193 hypothetical protein MAF_31940
339633202 hypothetical protein MAF_32030
339633218 hypothetical protein MAF_32190
339633220 hypothetical protein MAF_32210
339633231 hypothetical protein MAF_32320
339633242 hypothetical protein MAF_32430
339633251 hypothetical protein MAF_32520
339633253 hypothetical protein MAF_32540
339633264 hypothetical protein MAF_32650
339633268 hypothetical protein MAF_32690
339633282 hypothetical protein MAF_32830
339633287 hypothetical protein MAF_32880
339633291 hypothetical protein MAF_32930
339633298 hypothetical protein MAF_33000
339633320 hypothetical protein MAF_33240
339633328 hypothetical protein MAF_33320
339633345 hypothetical protein MAF_33490
339633372 hypothetical protein MAF_33790
339633377 hypothetical protein MAF_33840
339633389 hypothetical protein MAF_33970
339633395 hypothetical protein MAF_34030
339633405 hypothetical protein MAF_34130
339633408 hypothetical protein MAF_34170
339633409 hypothetical protein MAF_34180
339633413 hypothetical protein MAF_34220
339633427 hypothetical protein MAF_34360
339633439 hypothetical protein MAF_34490

339633468 hypothetical protein MAF_34780
339633477 hypothetical protein MAF_34870
339633483 hypothetical protein MAF_34930
339633485 hypothetical protein MAF_34950
339633489 hypothetical protein MAF_34990
339633495 hypothetical protein MAF_35050
339633496 hypothetical protein MAF_35060
339633512 hypothetical protein MAF_35230
339633523 hypothetical protein MAF_35360
339633526 hypothetical protein MAF_35390
339633539 hypothetical protein MAF_35530
339633571 hypothetical protein MAF_35850
339633575 hypothetical protein MAF_35890
339633585 hypothetical protein MAF_35990
339633586 hypothetical protein MAF_36000
339633592 hypothetical protein MAF_36060
339633602 hypothetical protein MAF_36160
339633604 hypothetical protein MAF_36180
339633613 hypothetical protein MAF_36270
339633614 hypothetical protein MAF_36280
339633615 hypothetical protein MAF_36290
339633621 hypothetical protein MAF_36360
339633643 hypothetical protein MAF_36580
339633666 hypothetical protein MAF_36820
339633678 hypothetical protein MAF_36940
339633680 hypothetical protein MAF_36960
339633688 hypothetical protein MAF_37050
339633690 hypothetical protein MAF_37070
339633693 hypothetical protein MAF_37100
339633696 hypothetical protein MAF_37130
339633706 hypothetical protein MAF_37230
339633708 hypothetical protein MAF_37250
339633710 hypothetical protein MAF_37270
339633714 hypothetical protein MAF_37310
339633726 hypothetical protein MAF_37440
339633740 hypothetical protein MAF_37580
339633760 hypothetical protein MAF_37790
339633761 hypothetical protein MAF_37800
339633762 hypothetical protein MAF_37810
339633764 hypothetical protein MAF_37830
339633776 hypothetical protein MAF_37950
339633783 hypothetical protein MAF_38020
339633785 hypothetical protein MAF_38040
339633792 hypothetical protein MAF_38110
339633814 hypothetical protein MAF_38340
339633817 hypothetical protein MAF_38370
339633826 hypothetical protein MAF_38460
339633833 hypothetical protein MAF_38540
339633837 hypothetical protein MAF_38580
339633842 hypothetical protein MAF_38630
339633857 hypothetical protein MAF_38790
339633858 hypothetical protein MAF_38800
339633863 hypothetical protein MAF_38850

339633869 hypothetical protein MAF_38910
339633872 hypothetical protein MAF_38950
339633886 hypothetical protein MAF_39110
339633895 hypothetical protein MAF_39200
507419168 hypothetical protein MASS_0226
507419224 hypothetical protein MASS_0282
507419312 hypothetical protein MASS_0370
507419388 hypothetical protein MASS_0446
507419395 hypothetical protein MASS_0453
507419573 hypothetical protein MASS_0631
507419607 hypothetical protein MASS_0665
507419688 hypothetical protein MASS_0746
507419689 hypothetical protein MASS_0747
507419783 hypothetical protein MASS_0841
507420152 hypothetical protein MASS_1210
507420243 hypothetical protein MASS_1301
507420283 hypothetical protein MASS_1341
507420458 hypothetical protein MASS_1516
507420568 hypothetical protein MASS_1626
507420599 hypothetical protein MASS_1657
507420675 hypothetical protein MASS_1733
507420706 hypothetical protein MASS_1764
507420831 hypothetical protein MASS_1889
507421022 hypothetical protein MASS_2080
507421062 hypothetical protein MASS_2120
507421317 hypothetical protein MASS_2375
507421351 hypothetical protein MASS_2409
507421383 hypothetical protein MASS_2441
507421499 hypothetical protein MASS_2557
507421737 hypothetical protein MASS_2795
507422009 hypothetical protein MASS_3067
507422116 hypothetical protein MASS_3174
507422133 hypothetical protein MASS_3191
507422367 hypothetical protein MASS_3425
507422495 hypothetical protein MASS_3553
507422497 hypothetical protein MASS_3555
507422741 hypothetical protein MASS_3799
507422867 hypothetical protein MASS_3925
507422881 hypothetical protein MASS_3939
507422928 hypothetical protein MASS_3986
507423019 hypothetical protein MASS_4077
507423332 hypothetical protein MASS_4390
507423649 hypothetical protein MASS_4707
507423661 hypothetical protein MASS_4719
507423845 hypothetical protein MASS_4903
118462871 hypothetical protein MAV_1869
118467240 hypothetical protein MAV_2538
118464542 hypothetical protein MAV_2768
118465695 hypothetical protein MAV_3499
31791507 hypothetical protein Mb0336c
31791731 hypothetical protein Mb0564c
31792930 hypothetical protein Mb1771
31793257 hypothetical protein Mb2100c

31793809	hypothetical protein Mb2656
31794583	hypothetical protein Mb3436c
31794704	hypothetical protein Mb3558c
507422182	IclR family transcriptional regulator
507420761	immunogenic protein MPT64
339631660	inositol-monophosphatase
339632855	integral membrane efflux protein EFPA
31793517	integral membrane transport protein
118464288	integrase
339631637	inv protein
339631545	invasion protein
339633596	iron-regulated LSR2 protein precursor
339630410	iron-sulfur-binding reductase
339633223	isochorismate synthase
339630149	isocitrate dehydrogenase [NADP]
339631969	isocitrate lyase
507423265	Isoniazid inducible protein iniA
507419869	isonitrile hydratase, ThiJ/PfpI family protein
339631798	isopentenyl-diphosphate delta-isomerase
339630796	L-fuculose phosphate aldolase
339630152	L-serine dehydratase
339633574	LacI family transcriptional regulator
339631048	large-conductance ion mechanosensitive channel MSCL
339630124	leucyl-tRNA synthetase
507419612	linocin-M18
339631976	lipase
339631466	lipase
339632801	lipid-transfer protein LTP1
507421989	lipoprotein LppU
339632882	lipoprotein P23
507423642	lipoprotein-releasing system ATP-binding protein LolD
118467102	LpqG protein
339630959	LuxR family transcriptional regulator
339630275	LuxR family transcriptional regulator
339630561	LuxR family transcriptional regulator
339633387	LYTB-related protein LYTB1
339631306	magnesium and cobalt transporter
339631890	malate synthase
507422045	malate:quinone oxidoreductase
339630719	malonyl CoA-ACP transacylase
339630125	MarR family transcriptional regulator
339630661	MCE-family lipoprotein LPRL
339630250	MCE-family protein MCE1A
339630251	MCE-family protein MCE1B
339630252	MCE-family protein MCE1C
339630253	MCE-family protein MCE1D
339630255	MCE-family protein MCE1F
339630657	MCE-family protein MCE2A
339630659	MCE-family protein MCE2C
339630660	MCE-family protein MCE2D
339633500	MCE-family protein MCE4C
339630424	MerR family transcriptional regulator
339633346	MerR family transcriptional regulator

507421856 methionine-R-sulfoxide reductase
339630712 methoxy mycolic acid synthase
339630713 methoxy mycolic acid synthase
339631383 methylated-DNA--protein-cysteine methyltransferase
507422489 methylated-DNA--protein-cysteine methyltransferase
339632959 methyltransferase
339631567 methyltransferase
339631594 methyltransferase
507423673 methyltransferase type 11
339630637 methyltransferase/methylase
339630432 Mg2+ transporter
507423058 MmpS family protein
339631910 molybdate-binding lipoprotein MODA
339633127 molybdenum cofactor biosynthesis protein E
339632363 molybdopterin biosynthesis protein MoeW
339631735 molybdopterin biosynthesis protein MoeX
339630623 muconate cycloisomerase
339633404 multifunctional geranylgeranyl pyrophosphate synthetase: dimethylallyltransferase + geranyltranstransferase + farnesyltranstransferase
339630539 mycolic acid synthase
339633799 mycolyl transferase
339631146 mycothiol conjugate amidase
339630461 NADH dehydrogenase
339633152 NADH dehydrogenase I subunit A
339633154 NADH dehydrogenase I subunit C
339633165 NADH dehydrogenase I subunit N
339632448 nicotinate-nucleotide adenylyltransferase
339631229 nitrate reductase subunit delta
339631230 nitrate reductase subunit gamma
339630337 nitrite extrusion protein
339632354 nitrite extrusion protein 1
507422124 nitrogen regulatory protein P-II
339633060 NRDI protein
507420644 nucleoside diphosphate kinase
339631584 nucleotide-sugar epimerase
339632537 oligoribonuclease
339631452 orotidine 5'-phosphate decarboxylase
507419170 osmoprotectant ABC transporter ProW
339630964 outer membrane protein OmpA
339630200 oxidative stress response regulatory protein OxyS
339632899 oxidoreductase
339631908 oxidoreductase
339632958 oxidoreductase
339632978 oxidoreductase
339630166 oxidoreductase
339630439 oxidoreductase
339631991 oxidoreductase
339631781 oxidoreductase
339631828 oxidoreductase
118463264 P450 heme-thiolate protein
339630408 PE family protein
339633737 PE family protein
339631702 PE family protein

339631235	PE family protein
339632781	PE family protein
339630233	PE family protein
339632545	PE family protein
339630242	PE family protein
507419862	PE protein
339631856	PE-PGRS family protein
339632873	penicillin-binding lipoprotein
507421832	peptidase, M48 family protein
339631350	peptide ABC transporter
339631351	peptide ABC transporter
507419378	peptide ABC transporter DppC
339630092	peptidyl-prolyl cis-trans isomerase
339630624	peroxidase
339633478	peroxidase
339631664	peroxidoxin
339632394	PhoH protein
339630995	phosphate ABC transporter substrate-binding protein
339633309	phosphate-transport system transcriptional regulator PHOU homolog 1 PHOY1
339632188	phospho-N-acetylmuramoyl-pentappeptidetransferase
507423465	phosphoenolpyruvate carboxykinase
339633222	phosphoglycerate mutase
339632374	phospholipase C
339632971	phosphopantetheine adenylyltransferase
339630850	phosphoribosylaminoimidazole-succinocarboxamide synthase
507419103	phosphotransferase
339630887	PhoY family transcriptional regulator
339633403	phytoene synthase PHYA
339631599	polyketide synthase
507423075	polyphosphate kinase 2
339633388	polyprenyl synthetase
339633143	PPE family protein
339632636	PPE family protein
339631101	PPE family protein
339632141	PPE family protein
31794068	PPE family protein
339631430	PPE family protein
339632457	PPE family protein
339633537	PPE family protein
339632782	PPE family protein
339631759	PPE family protein
339631760	PPE family protein
339631806	PPE family protein
339633866	PPE family protein
339631843	PPE family protein
339632110	precorrin-4 C11-methyltransferase
339633832	prephenate dehydratase PHEA
339630708	preprotein translocase subunit SecE
339632614	preprotein translocase subunit SecF
169629383	Probable methyltransferase
339630905	proline iminopeptidase
339633659	protease
339632252	protease

339632143 proteasome subunit beta
339631291 protein TatB
339633297 protein USFY
507419336 putative acyltransferase
507420767 putative beta-1,3-glucanase
433632219 Putative cytochrome P450 141 cyp141
507420286 putative dTDP-glucose-4,6-dehydratase-related protein
507421638 putative FeS assembly protein SufB
507419975 putative formamidopyrimidine-DNA glycosylase
507420293 putative glycosyltransferase
507422129 putative integral membrane transporter
507421707 putative integration host factor
169631459 Putative membrane protein, MmpS
507421370 putative oxidoreductase
507420160 putative oxidoreductase
169628009 Putative phenylacetic acid degradation protein PaaE/phenylacetate-CoA oxygenase/reductase, PaaK subunit
507420881 putative phosphoesterase
507422368 putative PLP-dependent transferase
507421410 putative proline-specific permease
507423237 putative RNA methyltransferase
507419029 putative transcriptional regulator
507422951 putative transcriptional regulatory protein
507419483 putative tRNA/rRNA methyltransferase
507419835 putative tRNA/rRNA methyltransferase
339632522 pyruvate dehydrogenase subunit E1
339633145 pyruvate formate lyase activating protein
507423572 quinone reductase
339631521 quinone reductase
507422961 R2-like ligand binding oxidase
507421941 RecA protein
339633707 recombination protein RecR
339632732 REPRESSOR LEXA
339632893 resolvase
339630673 resolvase
339633150 response regulator
507420635 resuscitation-promoting factor
507423032 resuscitation-promoting factor RpfA
339631937 resuscitation-promoting factor RPFC
339632683 riboflavin biosynthesis protein RibD
339631479 riboflavin synthase subunit alpha
339631483 riboflavin synthase subunit beta
339632463 ribokinase
339632908 ribonuclease HII
339631409 ribonuclease PH
339632019 ribonucleoside-diphosphate reductase subunit beta
339631078 ribose-phosphate pyrophosphokinase
339633425 ribosomal-protein-alanine acetyltransferase
339632890 ribosome recycling factor
339630804 RNA polymerase sigma factor
339632129 SEC-independent protein translocase membrane-bound protein TATA
339632128 Sec-independent protein translocase transmembrane protein TatC
118464269 secretion protein Snm4

339632205 Ser/Thr protein kinase L
339632360 serine acetyltransferase
339630366 serine protease
339633875 serine protease
339631935 short-chain dehydrogenase/reductase
339630916 short-chain dehydrogenase/reductase
339631993 short-chain dehydrogenase/reductase
339633787 short-chain dehydrogenase/reductase
339632778 short-chain dehydrogenase/reductase
339630230 short-chain dehydrogenase/reductase
339632535 short-chain dehydrogenase/reductase
339632866 short-chain dehydrogenase/reductase
339631955 sialic acid-transport integral membrane protein NANT
339633524 siderophore-binding protein
339632909 signal peptidase I
339632844 Sn-glycerol-3-phosphate transport integral membrane protein ABC transporter UGPA
339632843 Sn-glycerol-3-phosphate transport integral membrane protein ABC transporter UGPE
507422128 soluble secreted antigen MPT53
507423895 spoIIIJ-associated protein Jag
339633108 SSRA-binding protein SMPB
339633325 succinate dehydrogenase
339630324 succinate dehydrogenase
507419991 succinyl-CoA synthetase subunit alpha
507419990 succinyl-CoA synthetase subunit beta
339632529 succinyl-CoA:3-ketoacid-coenzyme A transferase subunit beta
339632080 sugar ABC transporter
339632342 sugar ABC transporter
339631305 sugar ABC transporter ATP-binding protein
339632078 sugar ABC transporter ATP-binding protein
339630720 sugar kinase
339630766 sugar transferase
339631588 sugar transferase
339632424 sulfate ABC transporter
339631353 sulfate adenylyltransferase
339630501 superoxide dismutase
507419061 superoxide dismutase (Mn)
339633181 TetR family transcriptional regulator
339630377 TetR family transcriptional regulator
339630402 TetR family transcriptional regulator
339633259 TetR family transcriptional regulator
339633303 TetR family transcriptional regulator
339630240 TetR family transcriptional regulator
339633573 TetR family transcriptional regulator
339633063 TetR family transcriptional regulator
339633849 TetR family transcriptional regulator
339631080 TetR family transcriptional regulator
339630315 TetR family transcriptional regulator
339631285 tetronasin ABC transporter ATP-binding protein
339630486 thiamin biosynthesis protein ThiG
339630483 thiamine-phosphate pyrophosphorylase
339632935 thioesterase TESA

507420231	thiol peroxidase
339630598	thioredoxin
507421627	thioredoxin
339633903	thioredoxin
507419653	thioredoxin ThiX
339633322	thymidine phosphorylase
339633582	transcription factor
507420372	transcription termination factor Rho
339631095	transcriptional regulator
339632892	transcriptional regulator
339632894	transcriptional regulator
339630348	transcriptional regulator
339631401	transcriptional regulator
339630893	transcriptional regulator
339631428	transcriptional regulator
339630419	transcriptional regulator
339630956	transcriptional regulator
507419071	transcriptional regulator
339631471	transcriptional regulator
339631728	transcriptional regulator
339631729	transcriptional regulator
339632757	transcriptional regulator
339631233	transcriptional regulator
339632008	transcriptional regulator
339630217	transcriptional regulator
339632028	transcriptional regulator
339630761	transcriptional regulator
339631542	transcriptional regulator
339632054	transcriptional regulator
339631044	transcriptional regulator
339631827	transcriptional regulator
339631830	transcriptional regulator
339631322	transcriptional regulator
339630812	transcriptional regulator
339633131	transcriptional regulator
339631088	transcriptional regulator
339631605	transcriptional regulator
507421431	transcriptional regulatory protein ArsR
339633042	transferase
507421189	translation initiation factor IF-3
479314470	transmembrane transport protein
31792708	transmembrane transport protein MmpL12
339632484	transporter
339632487	transporter
339633818	transporter MMPL8
339633632	transposase
339633380	trehalose 6-phosphate phosphatase
339632704	TRK system potassium uptake protein CEOC
339633032	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase TRMU
507421851	tRNA (Guanine-N(1)-)-methyltransferase
507423470	tRNA (guanine-N(7)-)-methyltransferase
339633228	two component sensor kinase
339633256	two component sensory transduction transcriptional regulator MTRA

339631370	UDP-phosphate alpha-N-acetylglucosaminyltransferase
507421443	universal stress protein
507421286	urease subunit beta UreB
118464539	virulence factor Mce
118463194	virulence factor Mce
507422680	WhiB family transcriptional regulator
339632793	zinc protease
339630244	zinc-type alcohol dehydrogenase subunit