Anotação, reanotação e consolidação de anotações de genomas com aplicação à classe Oomycetes

Aluno: Lucas Miguel de Carvalho
Orientador: Zanoni Dias
Coorientador: Guilherme Pimentel Telles

IC - UNICAMP

6 de Janeiro de 2014

Resumo

Os Oomycetes formam um grupo de centenas de organismos que incluem alguns agentes patogênicos de plantas. Algumas classes de Oomycetes, como, por exemplo, as Phytophthoras, começaram a ser sequenciadas e anotadas em 2006, mas suas anotações possuem grandes inconsistências. A reanotação de um genoma é uma tarefa necessária com o crescimento do número de bancos de dados de anotação. O objetivo deste trabalho é reanotar genomas da classe Oomycetes para obter maior precisão em suas anotações levando em conta que suas anotações atuais possuem muitos genes mal anotados ou sem anotação.

1 Introdução

Este texto tem como objetivo apresentar o plano de trabalho a ser desenvolvido ao longo do mestrado. Na Seção 2 serão introduzidos os conceitos básicos necessários para o seu entendimento. As seções 3 a 5 fornecerão detalhes sobre os casos de estudo. Por fim, as seções 6 e 7 tratarão a respeito da motivação e do cronograma a ser seguido durante o mestrado.

2 Conceitos básicos

Nesta seção faremos uma descrição sobre os conceitos básicos a serem utilizados durante o trabalho.

2.1 Genética

O nucleotídeo é formado por um açúcar (pentose), um grupo fosfato e uma base nitrogenada. O que diferencia um nucleotídeo de outro são suas bases nitrogenadas, sendo elas: adenina (A), citosina (C), uracila (U), timina (T) e guanina (G). A ligação de pentoses entre nucleotídeos forma uma cadeia, e a junção de várias cadeias forma um polinucleotídeo. Existem dois tipos de polinucleotídeos: DNA (ácido desoxirribonucléico) e RNA (ácido ribonucléico).

O DNA é formado por duas fitas, sendo elas ligadas pelas pontes de hidrogênio entre as bases nitrogênadas. A adenina forma ponte de hidrogênio com a timina, e a citosina com a guanina. As fitas do DNA são complementares, e é possível construir uma fita a partir da outra.

O RNA é formado por uma única fita. No RNA existe a base nitrogenada uracila ao invés da timina do DNA. Uma fita de RNA pode se dobrar de modo que suas bases se pareiem umas com as outras.

A genética é a ciência que estuda os genes. Os genes são unidades básicas que contêm informações físicas e comportamentais de um indivíduo, transmitidas por hereditariedade. Os genes são partes de moléculas de DNA que regulam a criação de proteínas e o funcionamento celular.

2.2 Genômica

A genômica é o estudo dos genomas de um organismo. Nela se aplicam métodos de sequenciamento de DNA, ferramentas de bioinformática para montagem e para análise funcional e estrutural do genoma, entre outros. O genoma é todo material genético que está localizado no núcleo de um organismo. Abaixo estão descritos métodos importantes para a melhor compreensão do estudo sobre genômica.

2.2.1 Métodos de sequenciamento

O sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõem. Atualmente existem várias tecnologias de sequenciamento, sendo elas diferenciadas pelo método utilizado e pelo tamanho das moléculas de DNA geradas (chamadas *reads*). As tecnologias mais conhecidas para sequenciamento de DNA geram sequências de vários tamanhos, de 25 a 1000 pares de base (bp). As tecnologias mais usadas são Sanger e pirosequenciamento (454, Solexa e Illumina).

2.2.2 Montagem

A montagem de um genoma é a combinação dos fragmentos de DNA (*reads*) para a obtenção da sequência original. Ela é usada para reconstruir a sequência do genoma dados vários fragmentos pequenos de sequência (*reads*). As tecnologias de sequenciamento geram uma grande quantidade de dados que posteriormente são processadas por programas montadores de genoma, como por exemplo SGA e Phrap, entre outros. Normalmente os fragmentos são gerados a partir de sequenciamento genômico *shotgun* ou genes transcritos (ESTs) [24].

Em sequenciamento de DNA, cobertura média representa o número médio de vezes que cada nucleotídeo foi lido durante o processo de sequenciamento, dado pela seguinte fórmula:

$$C = \frac{L \times N}{G}$$

onde C é a cobertura média, L é o tamanho médio dos *reads* (bp), N é o número de *reads* e G é o tamanho do genoma (bp).

Em genômica, isso significa que cada base foi sequenciada um certo número de vezes em média (2x, 4x, 20x, etc). Essa cobertura pode variar bastante, depedendo da região genômica.

2.3 Bancos de dados de anotação

A era da biologia em escala tem visto o acúmulo de dados biológicos, acompanhado pelo vasto crescimento de locais computacioanis para o armazenamento destes dados: os banco de dados biológicos. Para fazer o melhor uso dos bancos de dados biológicos e dos conhecimentos que eles contém, diferentes tipos de informação de diferentes fontes devem ser integradas de forma que façam sentido para os biólogos [2].

Esses bancos de dados podem fornecer várias informações, como, por exemplo, localização celular, função molecular, interação metabólica, interações proteína-proteína e componentes celulares. Alguns dos bancos de dados mais usados estão descritos abaixo.

2.3.1 GenBank [4]

O GenBank é um abrangente banco de dados público de sequências de nucleotídeos, suporte bibliográfico e anotação biológica. Ele é fornecido pelo

National Center for Biotechnology Information (NCBI), uma divisão da National Library of Medicine (NLM), localizado no campus do US National Institutes of Health (NIH), em Bethesda, nos Estados Unidos.

O NCBI constrói o GenBank principalmente a partir de dados de sequenciamento genômico e de Expressed Sequence Tags (ESTs), dados estes que são, geralmente, submetidos por centros de sequenciamento em larga escala.

Desde a sua criação, o GenBank tem crescido exponencialmente, e sua duplicação ocorre em um tempo de aproximadamente 18 meses. As divisões tradicionais do GenBank contêm mais de 156 bilhões de bases nitrogenadas de mais de 169 milhões de sequências individuais (dezembro/2013). Os genomas completos continuam a representar um segmento de rápido crescimento do banco de dados. O GenBank possui mais de 300 mil organismos nomeados no nível de gênero ou inferior.

Existem algumas maneiras de pesquisar e recuperar dados do GenBank, tais como: (i) pesquisar por identificadores de sequência e anotações com o *Entrez Nucleotide*, que é dividido em três divisões: CoreNucleotide, dbEST, e dbGSS, (ii) pesquisar e alinhar sequências do GenBank com uma sequência de consulta usando o BLAST (Basic Local Alignment Search Tool) [1].

2.3.2 Gene Ontology (GO) [21]

O projeto Gene Ontology (GO) é um esforço colaborativo para atender à necessidade de descrições consistentes de genes em diferentes bancos de dados. O projeto Gene Ontology fornece uma ontologia de termos definidos que representam as propriedades do gene. A ontologia abrange três domínios: (i) os componentes celulares (CC), as partes de uma célula ou do seu ambiente extracelular, (ii) as funções moleculares (MF), as atividades elementares de um produto do gene, a nível molecular e (iii) os processos biológicos (BP), que são operações ou conjuntos de eventos moleculares com um início e fim definidos, essenciais para o funcionamento de unidades integradas (células, tecidos, órgãos e organismos).

Anotação em termos de GO é a prática de prover as atividades e a localização de um gene, fornecendo referências e indicando qual o tipo de dado está disponível para apoiar as anotações. Ela indica se o gene tem uma função molecular em particular ou está envolvido em um processo biológico ou se tem um componente celular conhecido, além de determinar o método usado na anotação e fornecer a referência.

2.3.3 UniProt [29]

O UniProt (Universal Protein Resource) é um repositório abrangente para sequências de proteínas e dados de anotação. A missão do UniProt é proporcionar à comunidade científica um repositório de alta qualidade e de livre acesso formado por sequências de proteínas e informações funcionais. As bases de dados contidas no UniProt são UniProtKB, UniRef e UniParc. O UniProtKB é a base de dados de proteínas que consiste em duas seções: Swiss-Prot, que possui proteínas manualmente anotadas e revisadas em um total de 541.561 entradas, e TrEMBL, que possui proteínas automaticamente anotadas e não revisadas em um total de 44.746.523 de entradas (outubro/2013). O UniRef é um repositório de sequências usado para acelerar as pesquisas de similaridade de sequência. O UniParc é usado para manter o controle de sequências e os identificadores dessas sequências.

O UniProt fornece, além dos bancos de sequências, o uso da ferramenta BLAST, ferramenta de alinhamento de sequências, e uma ferramenta para converter dados de sequências de um banco de dados em outro (ID Mapping).

2.3.4 KEGG [15]

KEGG (Kyoto Encyclopedia of Genes and Genomes) é um repositório de banco de dados para a compreensão de funções de alto nível e do sistema biológico, como a célula, o organismo e o ecossistema, a partir de informações de nível molecular, especialmente conjuntos de dados moleculares em larga escala gerados pelo sequenciamento de genomas e outras tecnologias de alto rendimento.

O repositório KEGG é composto de informações moleculares de genes e proteínas (informação genômica) e substâncias químicas (informações químicas) que são integradas por ligações moleculares em diagramas de interação, de reação e de relações (sistemas de informação). Cada tipo de informação possui seus próprios bancos de dados. As informações genômicas são compostas pelos bancos de dados KEGG ORTHOLOGY, KEGG GENOME, KEGG GENES e KEGG SSDB. As informações químicas são compostas pelos bancos KEGG COMPOUND, KEGG GLYCAN, KEGG REACTION, KEGG RPAIR, KEGG RCLASS e KEGG ENZYME. Os sistemas de informação são compostos pelos bancos KEGG PATHWAY, KEGG BRITE, KEGG MODULE, KEGG DISEASE, KEGG DRUG e KEGG ENVIRON.

As informações mais usadas do KEGG são as redes moleculares (interações moleculares, reações e redes de relações que representam as funções

sistêmicas da célula e do organismo). O conhecimento experimental sobre tais funções sistêmicas é obtido a partir da literatura e organizado em três formas: (i) Pathway maps, em KEGG PATHWAY, (ii) Lista hierárquica (ontologia), em KEGG BRITE, (iii) Lista de membros, em KEGG MODULE e KEGG DISEASE.

Além de manter os aspectos de apoio à investigação básica, o KEGG está sendo expandido para aplicações que integram doenças humanas, drogas e outras substâncias relacionadas à saúde. O banco de dados KEGG possui 17.230 grupos de ontologia KEGG (KO), 2889 genomas de organismos (sendo 202 eucariotos, 2526 de bacterias e 161 archaea), sendo destes organismos um total de 11.534.269 genes e, além de 442.323 genes de genomas com montagem incompleta (outubro/2013).

3 A classe Oomycetes [10]

Os Oomycetes formam um grupo de centenas de organismos que incluem alguns agentes patogênicos de plantas. As doenças que eles causam incluem doenças de raízes, decaimento de folhas, podridão de raízes e caules, e aumento da suceptibilidade a pragas e mofos [8].

Os patógenos de plantas da classe Oomycetes incluem mais de 65 espécies de Phytophthora, uma centena de espécies de Pythium, e uma variedade de biótrofos obrigatórios (organismos que se alimentam de matéria viva), incluindo mofo felpudo e ferrugens brancas. Eles causam doenças devastadoras em numerosas culturas, plantas ornamentais e plantas nativas, e têm um enorme impacto sobre a agricultura. Algumas doenças notáveis são a requeima da batata, o míldio da videira e a morte repentina do carvalho.

Por causa de seu padrão de crescimento, nutrição por absorção e reprodução via esporos, os Oomycetes eram considerados pelos fitopatologistas como fungos inferiores [8]. No entanto, como a compreensão das relações evolucionárias cresceu, ficou claro que este grupo de organismos não está relacionado com os fungos verdadeiros. De fato, os fungos parecem mais próximos dos animais do que os Oomycetes, e os Oomycetes estão mais estritamente relacionadas com as algas e as plantas verdes [8]. Como ilustrado na Figura 1, os Oomycetes são classificados como Stramenopilhas ou Heterocontes que, juntamente com os Alveolates, formam os Chromalveolates, um dos cinco supergrupos na árvore de eucariotos.

Phytophthora literalmente significa destruidor de plantas, um nome cunhado por Anton de Bary em 1961, quando ele provou que um microorganismo, na época designado como um fungo, foi o agente causal da doença de

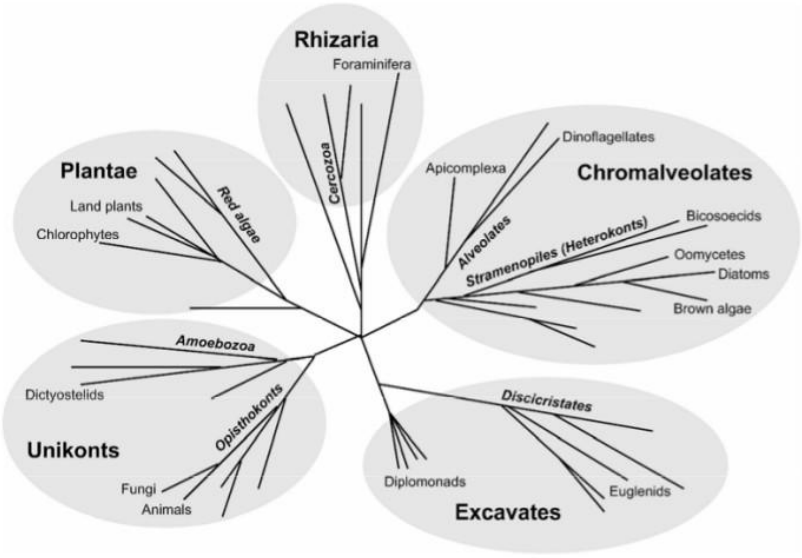

Figura 1: Um esquema da filogenia dos eucariotos, divididos em seus cinco superreinos [10].

planta conhecida como requeima da batata, que foi responsável pela grande fome da batata na Irlanda.

Devido às grandes perdas econômicas e à necessidade de solução para elas, os cientistas começaram projetos de sequenciamentos dos genomas de espécies de Phytophthora. As primeiras espécies a serem sequenciadas foram a *Phytophthora ramorum*, o agente causal da devastação de carvalhos na Califórnia em 2001, e a *Phytopthora sojae*, um agente causador da podridão da raiz na soja [28]. A *Phytophthora sojae*, a partir de então, foi usada como modelo para o sequenciamento e estudos de outras Phytophthoras ao longo dos anos, como por exemplo no sequenciamento, em 2009 [12], da *Pythophthora infestans*, o agente causador da grande fome da batata na Irlanda.

Mais recentemente, pelo fato de se iniciarem os estudos profundos envolvendo o gênero Phytophthora, se deu início ao sequenciamento de outras três espécies não menos importantes integrantes da classe Oomycetes: a *Hyaloperonospora arabidopsidis* (H. arabidopsidis) em 2010 [3] (um patógeno de *Arabidopsis thaliana*), a *Pythium ultimum*, sequenciada em 2010 [18] (um patógeno de várias doenças globais), e a *Phytophthora capsici*, sequenciada em 2012 [17] (um patógeno de hortaliças).

4 Genomas a serem estudados de Oomycetes

Nesta seção, discutiremos sobre cada espécie a ser estudada neste projeto de pesquisa, tratando de introduzir cada espécie e o *pipeline* usado na anotação inicial de seu genoma.

4.1 Phytophthora sojae e Phytophthora ramourum [28]

Phytophthora sojae (*P. sojae*) é um dos patógenos de soja predominantes nas regiões de produção com solos mal drenados. Ele pode infectar a soja em todas as fases de seu crescimento e provoca a podridão de suas sementes, raízes e caules. A gravidade da doença é proporcional ao tempo que os solos ficaram saturados de água e também está relacionada à susceptibilidade do cultivar [11].

A *P. sojae* sobrevive no solo como espóros chamados oósporos, que são produzidos em plantas infectadas. Os oósporos podem sobreviver por muitos anos no solo após os resíduos vegetais se decomporem [20], como mostra a Figura 2.

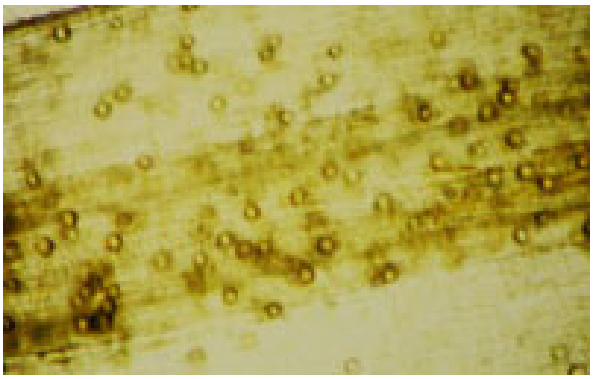

Figura 2: Oósporos na folha de soja [20].

O genoma da *P. sojae* foi descrito por Tyler et al. [28] em 2006. Para seu sequenciamento foi usado shotgun com cobertura de 9x, resultando em uma montagem de 95 Mbp.

Phytophthora ramorum (*P. ramorum*) é o patógeno conhecido por causar a morte repentina de carvalhos de diferentes espécies. Ela também provoca ferrugem e perecimento em muitas plantas ornamentais. A doença foi descoberta simultaneamente na Europa e na Califórnia na década de 1990, onde afetou culturas de carvalho e viveiros de mudas (Figura 3).

Figura 3: Carvalhos afetados pela *P. ramorum* [8].

O genoma da *P. ramorum* foi sequenciado em 2006, como descrito em Tyler et al. [28]. Para seu sequenciamento foi usado o sequenciamento shotgun com cobertura de 7-fold (7x), resultando em uma montagem de 65 Mbp.

Para a anotação da *P. sojae* e da *P. ramorum* foi utilizado o pipeline de anotação do JGI (Joint Genome Institute) que inclui vários métodos de anotação e predição. A maioria dos genes (75-80%) foram preditos *ab initio* usando os programas FGENESH [25] e GeneWise [5]. O último foi utilizado para corrigir modelos imperfeitos de genes ortólogos. Ao final da anotação foram preditos 19.027 genes de P. sojae e 15.743 genes de *P. ramorum* [28].

4.2 Phytophthora infestans [12]

Phytophthora infestans (*P. infestans*) é o agente patogênico mais destrutivo da batata e um organismo modelo para os Oomycetes. Como agente causador da grande fome da batata irlandesa em meados do século XIX, *P. infestans* teve um enorme efeito sobre a história humana, resultando em fome e dissipação da população.

Hoje em dia, ela afeta a agricultura mundial (Figura 4), já que a batata é a quarta maior cultura alimentar e uma alternativa às principais culturas de cereais para a alimentação da população mundial.

O genoma da *P. infestans* foi sequenciado em 2009, como descrito por Haas et al. [12]. Para o sequenciamento foi usado shotgun com cobertura de 9x, resultando em uma montagem de 240 Mbp.

Os genes preditos foram inicialmente anotados usando uma combinação de programas de localização gênica (Orthosearch [7] e GeneID [22]). Ao final

Figura 4: Plantação de batata afetada pela *P. infestans* [8].

da anotação foram preditos 17.797 genes.

4.3 Hyaloperonospora arabidopsidis [3]

O Oomycete *Hyaloperonospora arabidopsidis* (*H. arabidopsidis*) é um agente patogênico natural de Arabidopsis thaliana. A *H. arabidopsidis* pertence a um grupo de míldios patógenos, um grupo de Oomycetes patogênicos, compreendendo mais de 800 espécies que causam doenças em centenas de espécies de plantas. Míldios patógenos estão relacionados a outros Oomycetes patógenos de plantas (por exemplo, espécies de Phytophthora).

O genoma da *H. arabidopsidis* foi sequenciado usando os sequenciadores Sanger e Illumina. O sequenciamento feito no Sanger, com uma cobertura de 9.5x, gerou uma montagem de 77.8 Mbp. Uma análise adicional foi feita usando o Illumina, agora usando cobertura de 46x, que resultou em um acréscimo de 3.8 Mbp na montagem, gerando assim uma montagem final de 81.6 Mbp.

Os genes preditos em *H. arabidopsidis* foram identificados usando uma combinação de métodos *ab initio* e fazendo um BLASTx contra a base de dados nr do NCBI [26]. Um subconjunto de 458 genes nucleares de eucariotos (CEGs) foram adicionalmente anotados usando uma instalação local do pipeline CEGMA [23]. Estes genes são conservados em seis espécies de eucariotos (*Arabidopsis thaliana*, *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* e *Schizosaccharomyces pombe*). Ao final desta anotação foram preditos 14.543 genes.

4.4 Phytium ultimum [18]

Pythium ultimum (*P. ultimum*) pode causar uma grande variedade de doenças, como podridão das sementes, caimento de folhas e caules, e deixar seu hospedeiro mais sucetível a pragas. Espécies de Phytium são as principais causas da podridão da semente antes da germinação, por causa do solo mal drenado. Qualquer planta antes da germinação está exposta a ser infectada pela Phytium.

Espécies de Phytium são patógenos de plantas oportunistas que podem causar graves danos em plantas dependendo do seu estado de vulnerabilidade. Algumas espécies têm sido utilizadas como agentes de controle biológico, enquanto outros podem ser parasitas de animais, incluindo seres humanos.

O genoma da *Phytium ultimum* foi sequenciado usando o sequenciador Sanger, com bibliotecas de pirosequenciamento acopladas, com uma cobertura de 8x. O sequenciamento feito gerou uma montagem de 42.8 Mbp.

Para as anotações do genoma de *P. ultimum* foi utilizado o programa MAKER [6]. O MAKER foi ajustado para filtrar os modelos genéticos para predições de genes pequenos e parciais que produzem proteínas com menos de 28 aminoácidos. O *pipeline* MAKER foi criado para produzir previsões de genes *ab initio* tanto com repetições mascaradas quanto não mascaradas de sequências genômicas usando SNAP, FGENESH e GeneMark. Ao final da anotação foram preditos 15.290 genes.

4.5 Phytopthora capsici [17]

Phytophthora capsici (*P. capsici*) é um patógeno hemibiotrófico de hortaliças, causador de perdas significativas em todo o mundo. Seus principais hospedeiros são pimentas e cucurbitáceas (abóboras, melões, melâncias, pepinos, etc) (Figura 5). Durante os últimos 90 anos, a *P. capsici* tem se espalhado geograficamente ao mesmo tempo que o cultivo agrícola desses hospedeiros tem se intensificado, e ela tem se adaptado dinamicamente a fungicidas e a novos hospedeiros. Como outros membros deste gênero destrutivo (por exemplo, o patógeno causador da grande fome da batata irlandesa, a *P. infestans* e o patógeno causador da morte súbita de carvalhos, a *P. ramorum*), *P. capsici* tem uma epidemiologia explosiva, produzindo rapidamente um enorme número de esporos assexuais em hospedeiros infectados.

O sequenciamento do genoma foi processado combinando estratégias usando sequenciamento Sanger (cobertura de 5x) e 454 FLX Titanium (co-

Figura 5: Infecção de *P. capsici* em plantação de abóbora [14].

bertura de 30x) de sequências de DNA genômico, que resultou em uma montagem de 64 Mbp.

Para a anotação da *P. capsici* foi usado o *pipeline* de anotação do JGI [14], que combina vários preditores de genes: i) modelos genéticos baseados em cDNA; ii) modelos de genes baseados em proteína, preditos usando FGENESH+ [25] e GeneWise [5], juntamente com alinhamentos BLASTx contra o conjunto nr de sequências não redundantes do NCBI; e iii) os modelos de genes *ab initio* preditos usando FGENESH. Ao final dessa anotação foram preditos 19.805 genes.

5 Anotação de genomas

A anotação inclui a determinação dos genes, e a incorporação de informação proveniente de bases de dados biológicas já existentes a partir da identificação de similaridades entre sequências.

A anotação de um genoma pode ser manual ou automática. Por causa do grande número de genes e por tratar-se de uma atividade trabalhosa, a anotação automática é uma atividade importante nos projetos de sequenciamento de genomas.

Os programas de predição de genes são executados para localizar regiões que provavelmente são codificadoras de proteínas ou RNAs funcionais. Embora muito precisos para procariotos, esses programas ainda estão sujeitos a perder pequenos genes ou genes com uma composição atípica de nucleotídeos. Além disso, um aumento no número de genomas que estão sendo liberados antes da fase de finalização do projeto de montagem, com altas taxas de erros, está levando a erros nas predições de genes [19].

Depois da identificação dos genes e de seus produtos, espera-se transferir informação de bases de dados biológicas já existentes.

O processo de anotação é dependente da existência de bases de dados a partir das quais as informações sobre as seqüências possam ser transferidas.

Uma questão importante é a confiabilidade das informações. O problema é diferente em bases de dados curados e não curados. Um banco de dados é dito curado quando seus dados de entrada são cuidadosamente avaliados por um pesquisador, levando em conta os dados públicos ja existentes, com objetivo de validar a anotação. Já um banco de dados é dito não curado quando seus dados são gerados de algum modo por um método computacional, sem sofrer nenhum tipo de revisão posteriormente.

Apenas uma pequena parte dos genes de todos os genomas sequenciados ou presentes em bases de dados foram caracterizados em experimentos diretos, enquanto que a grande maioria são anotadas por transferência de informações das poucas sequências caracterizadas usando como base a similaridade de sequência.

Com bancos de dados curados, há boas razões para acreditar que, na maioria das vezes, esta transferência de informação é feita de forma responsável e conservadora. As bases de dados não-curadas, em sua maioria, são pesquisadas e utilizadas para a recuperação de sequências, que na maioria das vezes apresentam problemas causados por inconsistências em suas anotações. Essas anotações incorretas podem permanecer na base de dados por anos e portanto seria prudente exercer certa cautela antes de tirar conclusões de longo alcance exclusivamente a partir da sequência anotada [16].

A transferência de informação entre cadeias biológicas é feita com base em algum tipo de relacionamento entre elas. Esse relacionamento tipicamente é similaridade de seqüência primária ou similaridade de estrutura.

A reanotação é o processo no qual se anotam as proteínas de um organismo usando bases de dados mais recentes, com objetivo de consolidar as informações já existentes.

Com o aumento no número de banco de dados, a reanotação gênica é considerada uma tarefa necessária para melhor comparar os resultados obtidos em relação aos dados originais, buscando evoluções ou erros. O crescimento dinâmico dos banco de dados afeta diretamente a anotação gênica.

Ao longo do tempo, pode-se notar que a reanotação de um gênoma é frequente. Podemos citar, por exemplo, o genoma da *Arabidopsis thaliana* que foi sequenciado e anotado pela primeira vez no ano de 2000 [13]. Várias reanotações da *A. thaliana* foram feitas e, na maioria das vezes, resultou em números diferentes, como mostra a Tabela 1.

6 Motivação

A anotação de genomas é um processo sujeito a erros. Fontes de erros são montagem incorreta, determinação imprecisa da entrada mais similar em uma base de dados e transferência de erros da base de dados.

Por exemplo, para analisar a anotação feita para a *P. sojae*, buscamos todos os genes em que o produto (campo *Product Description*) contém as palavras *hypothetical* e *unknown*. Após essa análise, observamos que dos 26.585 genes de *P. Sojae*, 18.127 (68%) estão anotados com um desses termos hipotéticos. Desses, para 5.441 genes (30%) há alguma evidência de que algo melhor do que hipotético poderia ter sido predito (professor João Carlos Setubal, comunicação pessoal). Esse é um número muito grande, e portanto se utilizada alguma técnica para melhorar a anotação, pode haver um impacto significativo.

Tabela 1: Evolução na anotação gênica de *Arabidopsis thaliana* [27].

	Nature	TIGR1	TIGR2	TIGR3	TIGR4	TIGR5	TAIR6	TAIR7
Data de publicação	14/12/00	17/01/01	11/09/01	02/08/02	18/04/03	29/01/04	11/11/05	24/04/07
Tamanho do genoma (Mbp)	115.410	116.238	117.227	117.077	119.055	118.998	119.186	119.186
Proteínas codificadoras de genes	25.498	25.554	26.156	27.117	27.170	26.207	26.541	26.819
Transposons e pseudogenes	n/a	1274	1305	1967	2218	3786	3818	3889
Genes anotados com variante de splice alternativo	n/a	0	28	162	1267	2330	3159	3866
Densidade do gene (kb por gene)	4.50	4.55	4.48	4.32	4.38	4.54	4.48	4.44
Éxons/genes	5.20	5.23	5.25	5.24	5.31	5.42	5.64	5.79
Tamanho médio dos éxons (bp)	250	256	265	266	279	276	269	268
Tamanho médio dos íntrons (bp)	168	168	167	166	166	164	164	165

7 Objetivos

O objetivo deste projeto é investigar abordagens e avaliar múltiplas anotaçoes automáticas de um mesmo genoma ou de genomas próximos e tentar produzir anotações de consenso automaticamente. Adicionalmente, vamos avaliar a estratégia usando os genomas de Oomycetes citados.

Para o desenvolvimento do projeto primeiramente consideraremos as estratégias para anotação automática descritas na literatura.

Para a consolidação de anotações poderemos considerar estratégias baseadas nas cadeias, como clustering, ou estratégias baseadas nas próprias anotações, como a construção de um grafo de termos. A partir dos genomas disponíveis no FungiDB [9], e a aplicação de uma reanotação automática para sua consolidação, pode ser criado um grafo de termos para cada genoma, onde os vértices são compostos pelos seus genes e as arestas serão criadas caso haja algum vínculo entre eles.

A análise do grafo poderá mostrar, além da conectividade entre cada genoma e suas proximidades, se algum gene mal anotado tem alguma relação com algum outro gene de um outro organismo, mas pertencente à mesma classe, podendo haver transfêrencia de função biológica. As partes desconectadas podem imprimir um fator de não-proximidade com genomas da classe Oomycetes.

A reanotação pode evidenciar mudanças e melhorias nas predições das funções gênicas de um organismo. Nossa proposta é usar os genomas disponíveis de Oomycetes no FungiDB [9] e melhorar a sua anotação.

Primeiramente, usaremos *pipelines* automaticos de anotação de eucariotos para a atualização da anotação dos genomas de Oomycetes, usando bases de dados atualizadas. Estima-se que haverá diferença nos números obtidos pelas anotações, se comparados com os descritos na Seção 4.

Posteriormente, faremos um levantamento semelhante ao da *P. sojae*, como descrito na Seção 6, tentando evidenciar os genes que foram mal anotados. Para isso, usaremos ferramentas de comparação entre genomas.

A comparação das anotações entre os genomas da classe Oomycetes e cada nova anotação produzida pelos *pipelines* de anotação dessa classe pode fornecer novas informações sobre cada genoma e melhorar a anotação de cada gene.

8 Cronograma

As atividades que irão ser executadas ao longo deste trabalho estão descritas na Tabela 2.

- I. Disciplinas Obrigatórias
- II. Elaboração e defesa do Exame de Qualificação
- III. Programa de Estágio Docente (PED)
- IV. Pesquisa sobre os genomas de Oomycetes
- V. Verificação de possíveis genes mal anotados
- VI. Elaboração de um pipeline de anotação
- VII. Comparação de anotações entre os genomas de Oomycetes
- VIII. Finalização da dissertação
- IX. Defesa do mestrado

Tabela 2: Cronograma das atividades previstas.

Atividades	Meses																							
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
I	X	X	X	X		X	X	X	X															
II						X	X	X	X	X														
III													X	X	X	X								
IV									X	X	X													
V											X	X	X	X										
VI												X	X	X	X	X								
VII																X	X	X	X	X				
VIII																					X	X	X	
IX																								X

Referências

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] M. Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] L. Baxter et al. Signatures of Adaptation to Obligate Biotrophy in he Hyaloperonospora arabidopsidis Genome. *Science*, (330):1559–1561, 2010.

[4] D. A. Benson et al. GenBank. *Nucleic Acids Research*, 33:D34–D38, 2005.

[5] E. Birney, M. Clamp, and R. Durbin. GeneWise and Genomewise. *Genome Research*, 14:988–995, 2004.

[6] B. L. Cantarel et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18:188–196, 2008.

[7] S. M. S. Cruz et al. OrthoSearch: a scientific workflow approach to detect distant homologies on protozoans. *In: ACM Symposium on Applied Computing*, pages 1282–1286, 2008.

[8] W. E. Fry and N. J. Grunwald. Introduction to Oomycetes. *The Plant Health Instructor*, 2010.

[9] Fungal and Oomycetes Genomics Resources (FungiDB), 2013. http://fungidb.org/fungidb/.

[10] F. Govers and M. Gijzen. Phytophthora Genomics: The Plant Destroyers' Genome Decoded. *The American Pythopathological Society*, 19(12):1295–1301, 2006.

[11] C. R. Grau, A. E. Dorrance, J. Bond, and J. S. Russin. Fungal diseases. *Soybeans: Improvement, Production, and Uses*, pages 679–763, 2004.

[12] B. J. Hass et al. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature*, 461:393–398, 2009.

[13] Arabdopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408:796–815, 2000.

[14] Joint Genome Institute, 2005. http://genome.jgi-psf.org/PhycaF7/PhycaF7.home.html.

[15] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[16] E. V. Koonin and M. Y. Galperin. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. 2003.

[17] K. H. Lamour et al. Genome sequencing amd Mapping Reveal Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen Phytophthora capsici. *e-Xtra*, 25(10):1350–1360, 2012.

[18] C. A. Lévesque. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. *Genome Biology*, (11), 2010.

[19] C. Médigue and I. Moszer. Annotation, comparison and databases for hundreds of bacterial genomes. *Research in Microbiology*, 158:724–736, 2007.

[20] North Central Soybean Research Program. Phytophthora Root and Stem Rot, 2013. http://www.planthealth.info/prr_basics.htm.

[21] The Gene Ontology, 2013. http://www.geneontology.org/GO.doc.shtml.

[22] G. Parra, E. Blanco, and R. Guigo. GeneID in Drosophila. *Genome Research.*, 10:511–515, 2000.

[23] G. Parra, K. Bradnam, and I. Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, 2007.

[24] J. Pevsner. Bioinformatics and functional genomics, 2009.

[25] A. A. Salamov and V. V. Solovyev. Ab initio gene finding in Drosophila genomic DNA. *Genome Research*, 10:516–522, 2000.

[26] E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, pages D5–D15, 2009.

[27] D. Swarbreck et al. The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic Acids Research*, 36:D1009–D1014, 2008.

[28] B. M. Tyler et al. Phytophthora Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis. *Science*, (313):1261–1266, 2006.

[29] UniProt, 2013. http://www.uniprot.org/.