Proposta de Dissertação de Mestrado

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

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

Instituto de Computação - Unicamp

18 de dezembro de 2013

Roteiro

Introdução

- Com o avanço da tecnologia, os vários dados biológicos gerados são armazenados em banco de dados biológicos.
- A anotação dos genes de um organismo requer bancos de dados com informações atualizadas.
- Motivação: Banco de dados normalmente apresentam erros ou inconsistências de anotação.
- Proposta: Reanotar genomas da classe Oomyecetes, visando melhorar suas anotações com pipelines automáticos e novas metodologias, já que, com os dados existentes sobre suas anotações, podem haver melhoras significativas nelas.

Introdução

- Com o avanço da tecnologia, os vários dados biológicos gerados são armazenados em banco de dados biológicos.
- A anotação dos genes de um organismo requer bancos de dados com informações atualizadas.
- Motivação: Banco de dados normalmente apresentam erros ou inconsistências de anotação.
- Proposta: Reanotar genomas da classe Oomyecetes, visando melhorar suas anotações com pipelines automáticos e novas metodologias, já que, com os dados existentes sobre suas anotações, podem haver melhoras significativas nelas.

Introdução

- Com o avanço da tecnologia, os vários dados biológicos gerados são armazenados em banco de dados biológicos.
- A anotação dos genes de um organismo requer bancos de dados com informações atualizadas.
- Motivação: Banco de dados normalmente apresentam erros ou inconsistências de anotação.
- Proposta: Reanotar genomas da classe Oomyecetes, visando melhorar suas anotações com pipelines automáticos e novas metodologias, já que, com os dados existentes sobre suas anotações, podem haver melhoras significativas nelas.

Introdução

- Com o avanço da tecnologia, os vários dados biológicos gerados são armazenados em banco de dados biológicos.
- A anotação dos genes de um organismo requer bancos de dados com informações atualizadas.
- Motivação: Banco de dados normalmente apresentam erros ou inconsistências de anotação.
- Proposta: Reanotar genomas da classe Oomyecetes, visando melhorar suas anotações com pipelines automáticos e novas metodologias, já que, com os dados existentes sobre suas anotações, podem haver melhoras significativas nelas.

Introdução

- Com o avanço da tecnologia, os vários dados biológicos gerados são armazenados em banco de dados biológicos.
- A anotação dos genes de um organismo requer bancos de dados com informações atualizadas.
- Motivação: Banco de dados normalmente apresentam erros ou inconsistências de anotação.
- Proposta: Reanotar genomas da classe Oomyecetes, visando melhorar suas anotações com pipelines automáticos e novas metodologias, já que, com os dados existentes sobre suas anotações, podem haver melhoras significativas nelas.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

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

- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

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

- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

    C é a cobertura média, L é o tamanho médio dos *reads*, N é o número de *reads* e G é o tamanho do genoma (bp).
- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

    C é a cobertura média, L é o tamanho médio dos *reads*, N é o número de *reads* e G é o tamanho do genoma (bp).
- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

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

- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Genômica

- Genômica: Estudo dos genes de um determinado organismo.
- Sequenciamento de um genoma fornece a disposição dos nucleotídeos que o compõe.
  - Cobertura média: Número médio de vezes que cada nucleotídeo foi lido no sequencimento.

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

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

- Montagem: Combinação de fragmentos de DNA (*reads*) para obtenção da sequência original.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

Base de dados

- Base de dados de anotação: Local de armazenamento de dados biológicos estruturado em diversas informações.
  - GenBank: 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).
  - Gene Ontology (GO): Fornece ontologias que representam as propriedades de um gene (componente celular (CC), função molecular (MF) e processo biológico (BP)).
  - UniProt: Repositório de sequências de proteínas e dados de anotação, constituído por UniProtKB, UniRef e UniParc.
  - KEGG: Repositório formado por informações genômicas, informações químicas e sistema de informação. Seu diferencial é a formação de redes moleculares.

A classe Oomycetes

- Oomycetes: Centenas de organismos que incluem patógenos de plantas que causam grande impacto na agricultura.

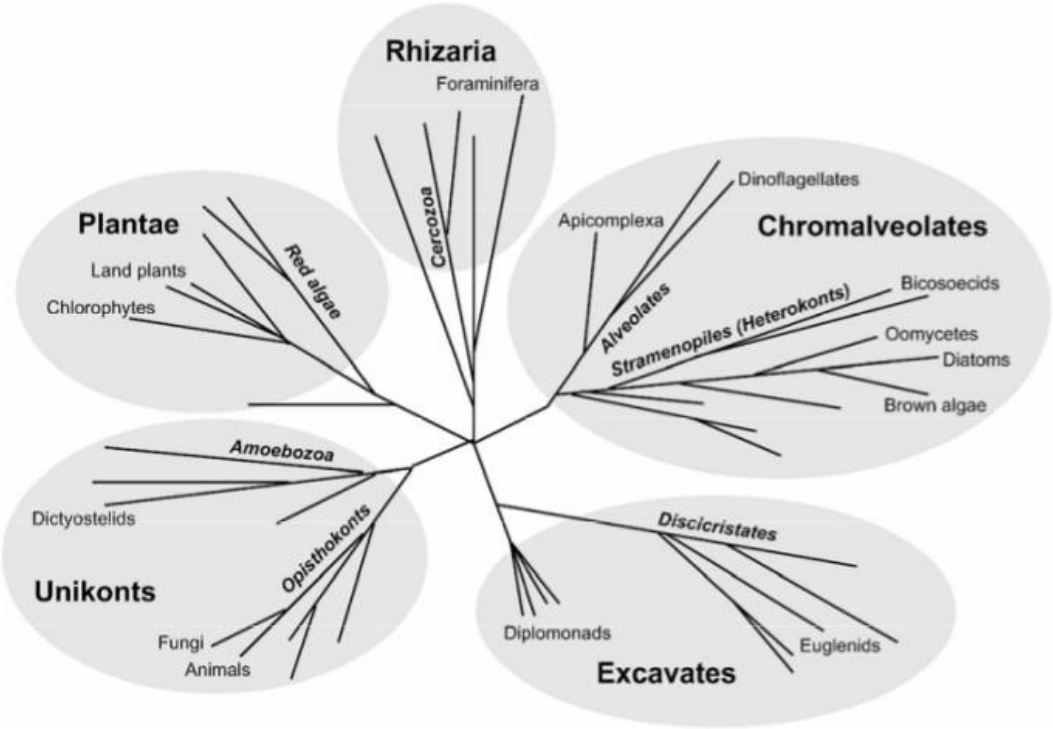

Figura: Um esquema da filogenia dos eucariotos, divididos em seus cinco superreinos (Tyler et al., 2006).

A classe Oomycetes

- Phytophthora: Literalmente significa destruidor de plantas, é um gênero da classe Oomycetes.
- Este estudo sobre Oomycetes será desenvolvido baseando em:
  - Quatro espécies de Phytophthora:
    - *Phytophthora sojae - P. sojae*
    - *Phytophthora ramorum - P. ramorum*
    - *Phytophthora capsici - P. capsici*
    - *Phytophthora infestans - P. infestans*
  - Uma espécie de Phytium: *Phytium ultimum - P. ultimum*
  - Uma espécie de Míldio: *Hyaloperonospora arabidopsidis - H. arabidopsidis*

A classe Oomycetes

- Phytophthora: Literalmente significa destruidor de plantas, é um gênero da classe Oomycetes.
- Este estudo sobre Oomycetes será desenvolvido baseando em:
  - Quatro espécies de Phytophthora:
    - *Phytophthora sojae - P. sojae*
    - *Phytophthora ramorum - P. ramorum*
    - *Phytophthora capsici - P. capsici*
    - *Phytophthora infestans - P. infestans*
  - Uma espécie de Phytium: *Phytium ultimum - P. ultimum*
  - Uma espécie de Míldio: *Hyaloperonospora arabidopsidis - H. arabidopsidis*

A classe Oomycetes

- Phytophthora: Literalmente significa destruidor de plantas, é um gênero da classe Oomycetes.
- Este estudo sobre Oomycetes será desenvolvido baseando em:
  - Quatro espécies de Phytophthora:
    - ★ *Phytophthora sojae - P. sojae*
    - ★ *Phytophthora ramorum - P. ramorum*
    - ★ *Phytophthora capsici - P. capsici*
    - ★ *Phytophthora infestans - P. infestans*
  - Uma espécie de Phytium: *Phytium ultimum - P. ultimum*
  - Uma espécie de Míldio: *Hyaloperonospora arabidopsidis - H. arabidopsidis*

A classe Oomycetes

- Phytophthora: Literalmente significa destruidor de plantas, é um gênero da classe Oomycetes.
- Este estudo sobre Oomycetes será desenvolvido baseando em:
  - Quatro espécies de Phytophthora:
    - ★ *Phytophthora sojae* - *P. sojae*
    - ★ *Phytophthora ramorum* - *P. ramorum*
    - ★ *Phytophthora capsici* - *P. capsici*
    - ★ *Phytophthora infestans* - *P. infestans*
  - Uma espécie de Phytium: *Phytium ultimum* - *P. ultimum*
  - Uma espécie de Míldio: *Hyaloperonospora arabidopsidis* - *H. arabidopsidis*

A classe Oomycetes

- Phytophthora: Literalmente significa destruidor de plantas, é um gênero da classe Oomycetes.
- Este estudo sobre Oomycetes será desenvolvido baseando em:
  - Quatro espécies de Phytophthora:
    - ★ *Phytophthora sojae* - *P. sojae*
    - ★ *Phytophthora ramorum* - *P. ramorum*
    - ★ *Phytophthora capsici* - *P. capsici*
    - ★ *Phytophthora infestans* - *P. infestans*
  - Uma espécie de Phytium: *Phytium ultimum* - *P. ultimum*
  - Uma espécie de Míldio: *Hyaloperonospora arabidopsidis* - *H. arabidopsidis*

Genomas de Oomycetes a serem estudados

- *Phytophthora sojae* e *Phytophthora ramorum*
  - *P. sojae* é um patógeno da soja e a *P. ramorum* é um patógeno que causa a morte repentina de carvalhos.
  - Ambos seus genomas foram descritos por Tyler et al. em 2006.
  - *P. sojae* foi sequenciada com uma cobertura de 9x gerando uma montagem de 95 Mbp.
  - *P. ramorum* foi sequenciada com uma cobertura de 7x gerando uma montagem de 65 Mbp.
  - Pipeline de anotação do JGI - Joint Genome Institute (19.027 genes preditos de *P. sojae* e 15.743 genes preditos de *P. ramorum*).

Genomas de Oomycetes a serem estudados

- *Phytophthora sojae* e *Phytophthora ramorum*
  - *P. sojae* é um patógeno da soja e a *P. ramorum* é um patógeno que causa a morte repentina de carvalhos.
  - Ambos seus genomas foram descritos por Tyler et al. em 2006.
  - *P. sojae* foi sequenciada com uma cobertura de 9x gerando uma montagem de 95 Mbp.
  - *P. ramorum* foi sequenciada com uma cobertura de 7x gerando uma montagem de 65 Mbp.
  - Pipeline de anotação do JGI - Joint Genome Institute (19.027 genes preditos de *P. sojae* e 15.743 genes preditos de *P. ramorum*).

Genomas de Oomycetes a serem estudados

- *Phytophthora infestans*
  - *P. infestans* é o agente patogênico mais destrutivo da batata.
  - Seu genoma foi descrito por Hass et al. em 2009.
  - *P. infestans* foi sequenciada com uma cobertura de 9x gerando uma montagem de 240 Mbp.
  - Foram usados os programas Orthosearch e GeneID para a predição dos genes (17.797 genes preditos).

Genomas de Oomycetes a serem estudados

- *Phytophthora infestans*
  - *P. infestans* é o agente patogênico mais destrutivo da batata.
  - Seu genoma foi descrito por Hass et al. em 2009.
  - *P. infestans* foi sequenciada com uma cobertura de 9x gerando uma montagem de 240 Mbp.
  - Foram usados os programas Orthosearch e GeneID para a predição dos genes (17.797 genes preditos).

Genomas de Oomycetes a serem estudados

- *Phytophthora capsici*
  - *P. capsici* é um agente patogênico de hortaliças. Seus principais hospedeiros são pimentas e curcubitáceas (abóbora, melão, melancia, abobrinha, pepino, etc).
  - Seu genoma foi descrito por Lamour et al. em 2012.
  - *P. capsici* foi sequenciada com uma cobertura de 5x (Sanger) e 30x (454 FLX Titanium) gerando uma montagem de 64 Mbp.
  - Para a anotação de seu genoma foi utilizado o pipeline de anotação do JGI (19.805 genes preditos).

Genomas de Oomycetes a serem estudados

- *Phytophthora capsici*
  - *P. capsici* é um agente patogênico de hortaliças. Seus principais hospedeiros são pimentas e curcubitáceas (abóbora, melão, melancia, abobrinha, pepino, etc).
  - Seu genoma foi descrito por Lamour et al. em 2012.
  - *P. capsici* foi sequenciada com uma cobertura de 5x (Sanger) e 30x (454 FLX Titanium) gerando uma montagem de 64 Mbp.
  - Para a anotação de seu genoma foi utilizado o pipeline de anotação do JGI (19.805 genes preditos).

Genomas de Oomycetes a serem estudados

- *Hyaloperonospora arabidopsidis*
  - *H. arabidopsidis* é o agente patogênico natural da *Arabidopsis thaliana*.
  - Seu genoma foi descrito por Baxter et al. em 2009.
  - *H. arabidopsidis* foi sequenciada com uma cobertura de 9.5x (Sanger) e 46x (Illumina) gerando uma montagem de 81.6 Mbp.
  - Para a anotação de seu genoma foram utilizados métodos *ab initio* (utilização do programa SNAP) e a utlização do BLASTx contra o banco de dados NR do NCBI (14.543 genes preditos).

Genomas de Oomycetes a serem estudados

- *Hyaloperonospora arabidopsidis*
  - *H. arabidopsidis* é o agente patogênico natural da *Arabidopsis thaliana*.
  - Seu genoma foi descrito por Baxter et al. em 2009.
  - *H. arabidopsidis* foi sequenciada com uma cobertura de 9.5x (Sanger) e 46x (Illumina) gerando uma montagem de 81.6 Mbp.
  - Para a anotação de seu genoma foram utilizados métodos *ab initio* (utilização do programa SNAP) e a utlização do BLASTx contra o banco de dados NR do NCBI (14.543 genes preditos).

Genomas de Oomycetes a serem estudados

- *Phytium ultimum*
  - *P. ultimum* pode causar uma grande variedade de doenças, como podridão das sementes e caimento de folhas e caules, além de deixar seu hospedeiro mais suscetível a pragas.
  - Seu genoma foi descrito por Lévesque et al. em 2010.
  - *P. infestans* foi sequenciada com uma cobertura de 8x gerando uma montagem de 42.8 Mbp.
  - Para a anotação de seu genoma foi usado o programa MAKER (15.290 genes preditos).

Genomas de Oomycetes a serem estudados

- *Phytium ultimum*
  - *P. ultimum* pode causar uma grande variedade de doenças, como podridão das sementes e caimento de folhas e caules, além de deixar seu hospedeiro mais suscetível a pragas.
  - Seu genoma foi descrito por Lévesque et al. em 2010.
  - *P. infestans* foi sequenciada com uma cobertura de 8x gerando uma montagem de 42.8 Mbp.
  - Para a anotação de seu genoma foi usado o programa MAKER (15.290 genes preditos).

Anotação de genomas

- Anotar é associar informações às cadeias obtidas no sequenciamento. Existem anotações manuais e automáticas, sendo as automáticas as mais importante em projetos de sequenciamento de genomas.
- Depende de um banco de dados. Existem dois tipos:
  - Curados: Se acredita que a transferência de informação é de forma responsável.
  - Não-curados: Apresentam problemas e inconsistências em suas anotações.
- Com o aumento do número de banco de dados, a reanotação - uso de banco de dados mais recentes para consolidar as anotações anteriores - é uma tarefa necessária para comparar evoluções ou erros em relação aos dados originais.

Anotação de genomas

- Anotar é associar informações às cadeias obtidas no sequenciamento. Existem anotações manuais e automáticas, sendo as automáticas as mais importante em projetos de sequenciamento de genomas.
- Depende de um banco de dados. Existem dois tipos:
  - Curados: Se acredita que a transferência de informação é de forma responsável.
  - Não-curados: Apresentam problemas e inconsistências em suas anotações.
- Com o aumento do número de banco de dados, a reanotação - uso de banco de dados mais recentes para consolidar as anotações anteriores - é uma tarefa necessária para comparar evoluções ou erros em relação aos dados originais.

Anotação de genomas

- Anotar é associar informações às cadeias obtidas no sequenciamento. Existem anotações manuais e automáticas, sendo as automáticas as mais importante em projetos de sequenciamento de genomas.
- Depende de um banco de dados. Existem dois tipos:
  - Curados: Se acredita que a transferência de informação é de forma responsável.
  - Não-curados: Apresentam problemas e inconsistências em suas anotações.
- Com o aumento do número de banco de dados, a reanotação - uso de banco de dados mais recentes para consolidar as anotações anteriores - é uma tarefa necessária para comparar evoluções ou erros em relação aos dados originais.

Anotação de genomas

- Anotar é associar informações às cadeias obtidas no sequenciamento. Existem anotações manuais e automáticas, sendo as automáticas as mais importante em projetos de sequenciamento de genomas.
- Depende de um banco de dados. Existem dois tipos:
  - Curados: Se acredita que a transferência de informação é de forma responsável.
  - Não-curados: Apresentam problemas e inconsistências em suas anotações.
- Com o aumento do número de banco de dados, a reanotação - uso de banco de dados mais recentes para consolidar as anotações anteriores - é uma tarefa necessária para comparar evoluções ou erros em relação aos dados originais.

Anotação de genomas

- Exemplo - Reanotação ao longo do tempo da *A. thaliana*

	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

Anotação de genomas

- Exemplo - Reanotação ao longo do tempo da *A. thaliana*

	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

Objetivos

- Avaliar múltiplas anotações automáticas de um mesmo genoma ou de genomas próximos e produzir anotações de consenso automaticamente utilizando novas metodologias de anotação.

Objetivos

- Avaliar múltiplas anotações automáticas de um mesmo genoma ou de genomas próximos e produzir anotações de consenso automaticamente utilizando novas metodologias de anotação.

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.
- Verificamos a anotação da *P. sojae*, buscando todos os genes em que o produto (campo *Product Description*) contém as palavras *hypothetical* e *unknown*.
  - Total 26.585 genes de *P. Sojae*.
  - 18.127 (68%) estão anotados com um desses termos hipotéticos.
  - Desses genes hipotéticos, 5.441 genes (30%) possuem alguma evidência de que algo melhor do que hipotético poderia ter sido predito.

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.
- Verificamos a anotação da *P. sojae*, buscando todos os genes em que o produto (campo *Product Description*) contém as palavras *hypothetical* e *unknown*.
  - Total 26.585 genes de *P. Sojae*.
  - 18.127 (68%) estão anotados com um desses termos hipotéticos.
  - Desses genes hipotéticos, 5.441 genes (30%) possuem alguma evidência de que algo melhor do que hipotético poderia ter sido predito.

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.
- Verificamos a anotação da *P. sojae*, buscando todos os genes em que o produto (campo *Product Description*) contém as palavras *hypothetical* e *unknown*.
  - Total 26.585 genes de *P. Sojae*.
  - 18.127 (68%) estão anotados com um desses termos hipotéticos.
  - Desses genes hipotéticos, 5.441 genes (30%) possuem alguma evidência de que algo melhor do que hipotético poderia ter sido predito.

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.
- Verificamos a anotação da *P. sojae*, buscando todos os genes em que o produto (campo *Product Description*) contém as palavras *hypothetical* e *unknown*.
  - Total 26.585 genes de *P. Sojae*.
  - 18.127 (68%) estão anotados com um desses termos hipotéticos.
  - Desses genes hipotéticos, 5.441 genes (30%) possuem alguma evidência de que algo melhor do que hipotético poderia ter sido predito.

Metodologia

- Levantar estratégias de anotação automática na literatura.
- Consolidar as anotações:
  - Estratégias baseadas em cadeia: Clustering.
  - Estratégias baseadas nas próprias anotações: Grafo de termos.
- Utilizar os genomas de Oomycetes disponíveis no FungiDB, e melhorar sua anotação.

Metodologia

- Levantar estratégias de anotação automática na literatura.
- Consolidar as anotações:
  - Estratégias baseadas em cadeia: Clustering.
  - Estratégias baseadas nas próprias anotações: Grafo de termos.
- Utilizar os genomas de Oomycetes disponíveis no FungiDB, e melhorar sua anotação.

Metodologia

- Levantar estratégias de anotação automática na literatura.
- Consolidar as anotações:
  - Estratégias baseadas em cadeia: Clustering.
  - Estratégias baseadas nas próprias anotações: Grafo de termos.
- Utilizar os genomas de Oomycetes disponíveis no FungiDB, e melhorar sua anotação.

Metodologia

- Levantar estratégias de anotação automática na literatura.
- Consolidar as anotações:
  - Estratégias baseadas em cadeia: Clustering.
  - Estratégias baseadas nas próprias anotações: Grafo de termos.
- Utilizar os genomas de Oomycetes disponíveis no FungiDB, e melhorar sua anotação.

Clustering

- Clustering: Agrupamento de um conjunto de cadeias em n-clusters, segundo uma característica em comum.

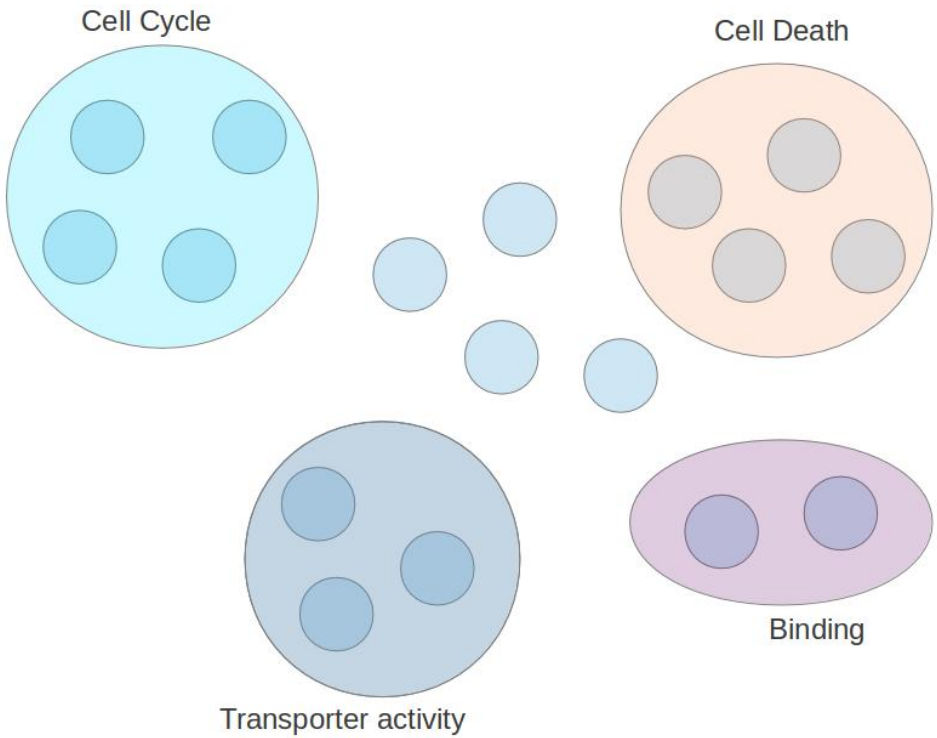

Figura: Exemplo de Clustering usando termos GO.

Grafo de termos

- Grafos de Termos: Exibe a conectividade entre cada genoma e suas proximidades. Verificação de transferência de função biológica por relações entre genes mal anotados e outros da mesma classe.

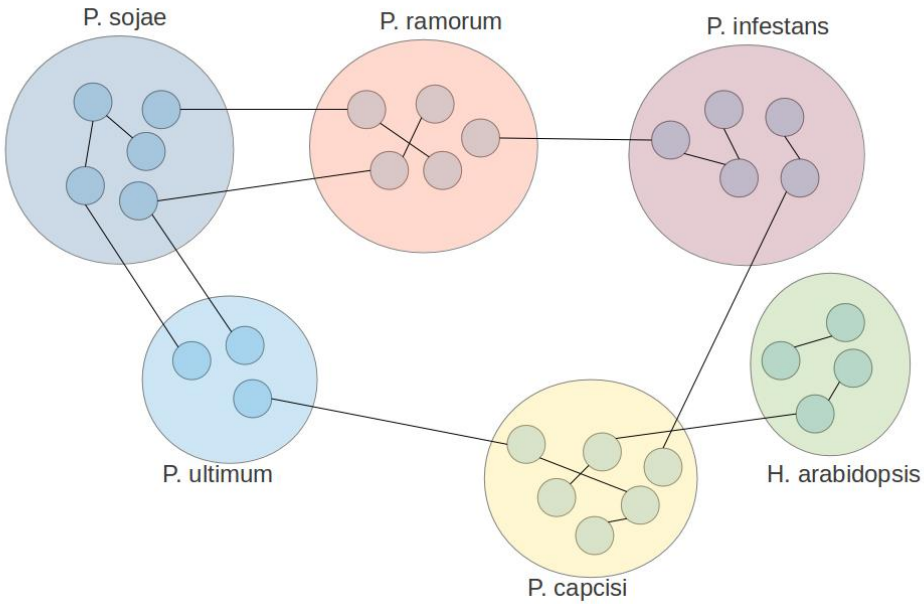

Figura: Exemplo de Grafo de Termos para os genomas de Oomycetes a serem estudados.

Metodologia

- Passos para a reanotação:
  - Uso de pipelines automáticos de anotação de eucariotos para atualizar as anotações dos Oomycetes.
  - Levantamento dos genes mal anotados das espécies citadas de Oomycetes (utilização de ferramentas de comparação de genomas).

Metodologia

- Passos para a reanotação:
  - Uso de pipelines automáticos de anotação de eucariotos para atualizar as anotações dos Oomycetes.
  - Levantamento dos genes mal anotados das espécies citadas de Oomycetes (utilização de ferramentas de comparação de genomas).

Metodologia

- Passos para a reanotação:
  - Uso de pipelines automáticos de anotação de eucariotos para atualizar as anotações dos Oomycetes.
  - Levantamento dos genes mal anotados das espécies citadas de Oomycetes (utilização de ferramentas de comparação de genomas).

Metodologia

- Passos para a reanotação:
  - Uso de pipelines automáticos de anotação de eucariotos para atualizar as anotações dos Oomycetes.
  - Levantamento dos genes mal anotados das espécies citadas de Oomycetes (utilização de ferramentas de comparação de genomas).

Resultados esperados

- Com o uso da nova metodologia sugerida, espera-se a obtenção de novas informações e melhoramento das anotações das espécies de interesse.

Resultados esperados

- Com o uso da nova metodologia sugerida, espera-se a obtenção de novas informações e melhoramento das anotações das espécies de interesse.

Cronograma

Tabela: 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

I. Disciplinas Obrigatórias.

Cronograma

Tabela: 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

II. Elaboração e defesa do Exame de Qualificação.

Cronograma

Tabela: 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

III. Programa de Estágio Docente (PED).

Cronograma

Tabela: 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

IV. Pesquisa sobre os genomas de Oomycetes.

Cronograma

Tabela: 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

V. Verificação de possíveis genes mal anotados.

Cronograma

Tabela: 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

VI. Elaboração de um pipeline de anotação.

Cronograma

Tabela: 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

VII. Comparação de anotações entre os genomas de Oomycetes.

Cronograma

Tabela: 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

VIII. Finalização da dissertação.

Cronograma

Tabela: 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

IX. Defesa do mestrado.

Agradecimentos

- Ao professor Dr. Zanoni Dias e ao professor Dr. Guilherme Telles pela oportunidade de pesquisa.
- Ao professor Dr. João Carlos Setubal pelos dados iniciais da pesquisa.
- A CNPq pela bolsa de mestrado.
- À Unicamp, mais especialmente ao Instituto de Computação, pela base e estrutura fornecida até agora no mestrado.