UNIVERSIDADE FEDERAL DE PELOTAS

Programa de Pós-Graduação em Biotecnologia

Tese

Square: uma plataforma gráfica e intuitiva para anotação de genomas bacterianos

Marcus Redü Eslabão

Pelotas, 2016

Marcus Redü Eslabão

Square: uma plataforma gráfica e intuitiva para anotação de genomas bacterianos

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciências (área do Conhecimento: Bioinformática).

Orientador: Odir Antonio Dellagostin

Pelotas, 2016

Dados de catalogação na fonte:
Maria Beatriz Vaghetti Vieira – CRB-10/1032
Biblioteca de Ciência & Tecnologia - UFPel

E76s Eslabão, Marcus Redü

Square: uma plataforma gráfica e intuitiva para anotação de genomas bacterianos / Marcus Redü Eslabão. – 48f. – Tese (Doutorado). Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas. Centro de Desenvolvimento Tecnológico. Pelotas, 2016. – Orientador Odir Antonio Dellagostin.

1.Biotecnologia. 2. Anotação de genomas. 3. Genômica. 4. Sequenciamento d DNA. I.Dellagostin, Odir Antonio. II. Título.

CDD: 574.88

BANCA EXAMINADORA

Prof. Dr. Alan John Alexander McBride (UFPel, CDTec)

Prof. Dr. Luciano da Silva Pinto (UFPel, CDtec)

Prof. Dr. Marilton Sanchotene de Aguiar (UFPel, CDTec)

Prof. Dr. Odir Antonio Dellagostin (Orientador, UFPel, CDTec)

Agradecimentos

Gostaria de agradecer aos membros da equipe de Bioinformática, Frederico Kremer, Fernanda Valiati, Jessica Plaça, Mariana Brutschin, Monize Provisor, Amanda Guimarães, Julia Labonde, Rafael Woloski, Paulo Porto e ao Prof. Luciano Pinto, por proporcionar todo o necessário para realização do meu trabalho, desde a parte material, até a parte de convivência e emocional, gerando um ambiente positivamente propicio para o desenvolvimento intelectual e pessoal. Um agradecimento mais que especial ao meu orientador Prof. Odir Antonio Dellagostin, que desde a graduação serve como guia e exemplo, acreditando na minha capacidade e fornecendo de forma direta e indireta o conhecimento e recursos necessários para minha formação. Aos colaboradores Timóteo Rico, Augusto Schmidt, Prof. Marilton Aguiar e Prof. Alan McBride, que interagem com o nosso laboratório, proporcionando um ganho de conhecimento, novas ideias e soluções.

A minha mãe Francisca, ao meu pai Felizardo, ao meu irmão Rafael e aos demais membros da minha família, pela atenção e apoio necessários para minha evolução e realização dos meus sonhos. Em especial a minha esposa Mirian, pessoa com a qual decidi compartilhar a minha vida, que tem me apoiado de todas formas possíveis e imagináveis, e ao meu futuro filho Vitor, que mesmo estando a caminho, já serve de inspiração para o meu futuro.

Ao Núcleo de Biotecnologia, por me acolher e dar a infraestrutura necessária para o desenvolvimento do meu trabalho, e aos órgãos de fomento CAPES e FAPERGS, que apoiaram minhas atividades.

Obrigado a todos.

"Nada na vida deve ser temido, somente compreendido.”

Marie Curie

Resumo

ESLABÃO, Marcus. **Square: uma plataforma gráfica e intuitiva para anotação de genomas bacterianos.** 2016. 46f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

O sequenciamento de DNA é uma técnica que fornece uma fonte vasta de informações sobre diversos organismos. Atualmente, novas metodologias de sequenciamento conhecidas como *Next-Generation Sequencing*, estão fazendo com que esta técnica fique inúmeras vezes mais rápida, precisa e economicamente acessível, tornando-se popular e disseminada no meio científico. Com a popularização do sequenciamento de genomas, laboratórios que não possuem ênfase em sequenciamento de DNA, utilizam desta abordagem para complementar seus estudos. Porém, a facilidade em obter a sequência do DNA contrasta com a dificuldade em processar, analisar e anotar o genoma, para que então seja possível obter informações biológicas relevantes sobre aquele organismo. Para auxiliar os pesquisadores que se utilizam desta técnica, alguns softwares estão disponíveis, porém, geralmente são pagos, não realizam toda a tarefa ou são de difícil utilização, neste último caso, por serem em sua grande maioria executados através de terminais de comando, que não contam com um ambiente gráfico para guiar os usuários. Com base nesta problemática, o presente trabalho teve por objetivo criar um software de anotação de genomas de fácil utilização e com interface gráfica amigável, gratuito e que anote com as informações necessárias para submissão ao GenBank. Para implementação do software, denominado Square, as linguagens de programação Python e Object Pascal foram utilizadas. Os algoritmos Prodigal, NCBI BLAST e tRNAscan-SE também foram integrados no software. Ao final da etapa de desenvolvimento, o Square foi testado com três genomas e comparado com dois anotadores populares: o RAST e o BASys. O resultado mostrou que o Square possui maior precisão que os dois outros anotadores, por se aproximar mais do resultado depositado no NCBI, e mais rápido, por ser executado localmente com rapidez. O Square demonstrou-se uma boa alternativa para usuários que não estão acostumados com o terminal de comando Linux e está disponível no endereço http://sourceforge.net/projects/sqgenome/.

Palavras-chave: Anotação de Genomas, Genômica, Sequenciamento de DNA.

Abstract

ESLABÃO, Marcus. **Square: a graphical and intuitive platform for annotation of bacterial genomes.** 2016. 46f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

DNA sequencing is a technique that provides a vast source of information on various organisms. Currently, new sequencing methods known as Next-Generation Sequencing, are making this technique many times more rapid, accurate and affordable, making it popular and widespread in the scientific community. With the popularization of genome sequencing, laboratories that do not have an emphasis on DNA sequencing, are using this approach to complement their studies. However, the ease in obtaining a DNA sequence contrasts with the difficulty to process, analyze and annotate the genome, in order to obtain relevant biological information. To assist researchers who use this technique, several programs are available, however, they are generally not free, do not perform all the necessary analysis or are difficult to use, mainly because a considerable number of them make use of command line to be executed, which is not intuitive. The objective of this study was to create a genome annotation software easy to use, with a user friendly interface, free and able to provide all the necessary information for the annotated genome to be submitted to GenBank. For software implementation named Square, Python and Object Pascal programming languages were used. The Prodigal algorithms, NCBI BLAST and tRNAscan-SE were also integrated in the software. At the end of the development stage, Square was tested with three genomes and compared to two popular annotators: RAST and BASYS. The result showed that the Square has higher accuracy than the other two annotator programs, as the results are similar to what is deposited in NCBI, and produce the result in a shorter time, as it runs locally. The Square proved to be a good alternative for users not familiar with the Linux command terminal and is available in http://sourceforge.net/projects/sqgenome/ address.

Keywords: Genome annotation, Genomics, DNA sequencing.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

CDS – Coding DNA Sequence (Região codificante do DNA)

DNA – Deoxyribonucleic Acid (Ácido desoxirribonucleico)

NGS - Next-Generation Sequencing (Sequenciamento de nova geração)

ORF - Open Reading Frame (Fase de leitura aberta)

RAM - Random Access Memory (Memória de acesso randômico)

RNA – Ribonucleic Acid (Ácido ribonucleico)

tRNA – Transfer Ribonucleic Acid (Ácido ribonucleico de transferência)

Sumário

1 INTRODUÇÃO GERAL

O sequenciamento de DNA é um método chave na obtenção de informações em diversas áreas. Em 1977 um método diferenciado de sequenciamento foi criado por Frederick Sanger (Sanger et al., 1977). Esta técnica manual permaneceu inalterada até meados dos anos 80, quando surgiram os primeiros sequenciadores automáticos (Smith et al., 1986). Nove anos mais tarde, uma nova atualização foi realizada nesta técnica, com o desenvolvimento de sequenciadores com capilares, e sua utilização e comercialização permanecem até os dias de hoje. O sequenciamento com eletroforese capilar permitiu um novo método de sequenciamento denominado *Whole genome shotgun sequencing*, que permite que um genoma seja fragmentado, sequenciado e sua sequência de DNA montada novamente, proporcionando um paralelismo da técnica, onde diversos fragmentos são sequenciados ao mesmo tempo, agilizando o processo de obtenção da sequência genômica. Seu primeiro uso ocorreu com o organismo *Haemophilus influenzae* (Fleischmann et al., 1995), sendo posteriormente empregado no projeto genoma humano. Na mesma proporção que a capacidade de gerar sequências foi sendo aprimorada, uma problemática ficou mais evidente, ter o código genético sequenciado em mãos não estava sendo a solução. Estudos eram necessários para desvendar a função de cada parte de uma sequência e um lugar público para depositar estas informações se fazia necessário. Em 1992 um centro para desenvolvimento de ferramentas em bioinformática, o NCBI, passou a gerenciar o banco de dados biológicos GenBank (Benson et al., 2012), tornando-se o primeiro banco de dados de DNA público, onde, deste então, é possível depositar sequências e através de um algoritmo denominado BLAST (Altschul et al., 1997) buscar sequências depositadas neste banco de dados. Com um banco de dados de DNA e um mecanismo de busca eficiente, tornava-se possível e fácil a inferência de dados sobre uma sequência, levando em consideração a homologia, porém, genomas possuem centenas de regiões codificadoras (ORFs) que precisavam ser anotadas de forma individual, demandando uma equipe especializada e um grande tempo de trabalho.

Em 2005, uma nova geração de sequenciadores foi desenvolvida, conhecida como *next-generation sequencing* (NGS), e pela primeira vez, em 28 anos, um

sequenciador comercial tinha uma metodologia diferente do método de Sanger, sendo a nova técnica conhecida como pirosequenciamento (Ronaghi et al., 1996). O custo por par de base sequenciada com método de Sanger nos anos 90, que era de U$10, caiu para U$0,03 no ano de 2005 com o pirosequenciamento (Service, 2006). A capacidade de sequenciamento subiu de 0,85MB em 3 horas no método de Sanger (Illumina, 2012) para cerca de 180MB em 3 horas com esta técnica (Metzker, 2010). Logo após o surgimento do pirosequenciamento, outras técnicas de NGS foram apresentadas a comunidade científica, tais como *bridge amplification, ligation reaction for sequencing, true single molecule sequencing* e *four-colour real-time sequencing* (Hui, 2012; Metzker, 2010). Com o NGS, obter as sequências de DNA tornou-se rápido e barato, proporcionando um aumento no número de projetos e consórcios para sequenciamento de genomas em massa. Com isso, houve a necessidade da criação de bancos de dados para informações de metagenomas (Barrett et al., 2012). Juntamente com o aumento da capacidade de gerar sequências e o aumento do volume de dados a problemática inicial da inferência manual de informações sobre estes genomas aumentou. Para contornar essa problemática diversos programas de predições e bancos de dados com localização celular da proteína, família e domínio proteico, rotas metabólicas, dentre outros, foram criados. Porém, somente com a criação de anotadores automatizados ou *pipelines*, que visam unir softwares e bancos de dados já existentes, o objetivo de agilizar a anotação de genomas começou a ser alcançado.

Basicamente, o processo de anotação de um genoma consiste em atribuir o máximo de informações a cada uma das possíveis regiões codificadoras (ORF). Inicialmente as ORFs são preditas, podendo levar em conta similaridade com outro organismo já anotado, cálculos *ab initio*, ou simplesmente procurar por um códon de iniciação e outro de terminação, com um número de bases entre eles suficientes para codificar uma proteína. Usualmente, esses métodos são fundidos, permitindo o aumento da confiabilidade da predição de ORFs. Ao término do processo de localização das ORFs, diversas predições podem ser feitas de forma individual a cada uma das regiões codificadoras, como por exemplo, predições de RNA transportadores e ribossomais. Da mesma forma individualizada, é possível atribuir funções com base em bancos de dados específicos, levando em consideração a similaridade entre a sequência da ORF do novo organismo e do organismo já sequenciado (Aziz et al., 2008).

Alguns softwares automatizados para anotação de genomas bacterianos estão disponíveis, dentre eles o RAST (Aziz et al., 2008), AGeS (Kumar et al., 2011), Blast2GO (Conesa et al., 2005), KAAS (Moriya et al., 2007), Diya (Stewart et al., 2009), xBASE2 (Chaudhuri et al., 2008), entre outros. Uma análise geral dos anotadores disponíveis revela alguns problemas pontuais, tais como, pouca informação na anotação, demora para anotação *online* devido a filas de espera, presença de dados com pouca relevância e principalmente interface designada a usuários experientes. Estes problemas forçam grupos de pesquisa a desenvolver *pipelines* próprios e a usuários leigos a desperdiçar tempo aderindo a um dos softwares de anotação disponíveis ou a se privar do sequenciamento como uma ferramenta de apoio.

O objetivo deste trabalho foi desenvolver uma *pipeline* local e gratuita para anotação de genomas bacterianos, com uma interface gráfica e amigável, um modo passo a passo para usuários leigos e um modo avançado para usuários experientes. Os bancos de dados biológicos poderão ser transferidos para o computador pessoal ou ser utilizado via internet. Um módulo do *pipeline* será exclusivo para o *download* e atualização automatizados dos softwares presentes no anotador, facilitando a instalação e manutenção. Com estas medidas, é esperado acompanhar a crescente facilidade de sequenciar genomas bacterianos e atrair um maior número de pesquisadores que utilizam o sequenciamento apenas como uma ferramenta esporádica.

2 REVISÃO BIBLIOGRÁFICA

Com o intuito de reduzir o tempo de anotação, disponibilizando ao pesquisador mais tempo para análise do genoma, alguns softwares anotadores de genomas procariotos gratuitos estão disponíveis. De acordo com determinadas características, pode-se classificá-los em alguns grupos, tais como linguagem de programação utilizada, sistema operacional, funções, custo e interações com outros programas. Para facilitar a descrição, os softwares observados foram classificados em três grandes grupos, de acordo com a plataforma onde este software é executado, sendo dividido em localmente, em servidores na *internet* ou híbridos que fazem parte do trabalho localmente e parte *online*.

2.1 Softwares de execução local

Os softwares de execução local possuem a característica de serem executados em um *hardware* fisicamente próximo ao usuário, não necessitando estar conectado a Internet para utilizá-lo, podendo este ser executado em um ou mais sistemas operacionais.

2.1.1 BG7

O BG7 (Pareja-Tobes et al., 2012) é uma *pipeline* de anotação voltado para dados de sequenciadores de nova geração. Seu código é escrito em Java, executando somente por meio de linha de comando. Seus pré-requisitos são o NCBI BLAST+ e o Java JDK instalados, bem como um computador com arquitetura 64 bits e com mais de 7 GB de memória RAM disponível. Além do processamento local, é possível contratar o *Amazon Web Services* para melhorar o desempenho do programa. Além de um site próprio contendo a documentação e explicando o funcionamento, a sua distribuição é feita através do *site* GitHub. Seus arquivos de entrada podem ser em formato FASTA ou em XML, no caso do último deverá ser seguido um modelo disponibilizado no site. Já para saída, os formatos são GFF3, GenBank, XML e tabular. Como principais diferenciais, ele não se utiliza da forma clássica de anotação, onde ocorre primeiramente uma busca de genes e posteriormente a anotação. Para que sejam encontrados os genes, ele utiliza-se de um conjunto de dados de referência retirado do Uniprot (Magrane & Consortium,

2011), ou seja ele conhece todos os genes que são esperados em um determinado genoma, o que descarta a necessidade de predição.

2.1.2 Diya

O *Do-It-Yourself Annotator ou Diya* (Stewart et al., 2009) é uma *pipeline* escrita em Perl para anotação de genomas procariotos. Ele é indicado para os sistemas operacionais Linux e OS X, executando em terminal e possuindo como dependência o Perl e a biblioteca BioPerl. Tanto a entrada quanto a saída de dados pode ser feita através dos formatos FASTA, GFF3 e GenBank. Sua distribuição é feita através do GitHub, não possuindo diferenciais quanto a sua utilização e realizando uma anotação básica e podendo utilizar os bancos de dados padrão NCBI BLAST

Esta *pipeline,* denominada EuGene-PP (Sallet et al., 2014), é específica para genomas procariotos e voltado para o RNAseq, possuindo também outras versões para eucariotos. Escrita em Perl, ela executa em terminal Linux e é basicamente um facilitador para chamar o anotador EuGene-P (Sallet et al., 2013) escrito em C++. A *pipeline* utiliza cadeias de Markov para achar similaridade com um conjunto de proteínas de referência, o Prodigal (Hyatt et al., 2010) para predizer as possíveis regiões codificantes e o tRNAscan-SE (Lowe & Eddy, 1997) e RNAmmer (K Lagesen et al., 2007) para predição dos *RNAs.* Seu arquivo padrão de entrada é o FASTA e o de saída é o GFF3.

2.1.4 Maker

O Maker (Cantarel et al., 2008) é uma *pipeline* de anotação com foco em pequenos genomas eucariotos, mas funciona para anotação de genomas procariotos. Ele possui também uma versão denominada Maker-P específico para genomas de plantas (Campbell et al., 2014). Escrito em Perl, roda em terminais Linux e possui também uma versão *online.* Sua instalação possui diversas dependências e bibliotecas, descritas em seu artigo, que devem ser instaladas manualmente, além de um arquivo de configuração onde devem ser indicados os diretórios de trabalho e instalação do mesmo, porém disponibiliza um tutorial bem

detalhado explicando sua instalação, configuração e uso. Seu arquivo de entrada padrão é o formato FASTA e o de saída é o GFF3.

2.1.5 Prokka

A ferramenta Prokka (Seemann, 2014) é uma *pipeline* de anotação de genomas procariotos criada com o intuito de ser extremamente rápida. Segundo seus autores, um genoma médio de quatro milhões de pares de base pode ser anotado em dez minutos com computador quad-core. Escrito em linguagem Perl, roda em sistemas Unix e utiliza a biblioteca BioPerl como pré-requisito. Para predição de regiões codificantes utiliza-se do Prodigal (Hyatt et al., 2010), para predições de *RNAs* os softwares RNAmmer (K Lagesen et al., 2007), Aragorn (Laslett & Canback, 2004) e Infernal (Nawrocki & Eddy, 2013), e a predição dos peptídeos sinais fica por conta do SignalP (Petersen et al., 2011). Seu arquivo de entrada é o formato FASTA, e apresenta mais de dez formatos para arquivos de saída, dentre eles o FASTA, GFF3 e GenBank.

2.1.6 Características compartilhadas

Dentre os anotadores gratuitos disponíveis para execução local, pode-se observar como característica dominante a tendência de execução voltada para o sistema operacional Linux, especialmente o Mint, o Debian e o Ubuntu, que foram os mais acessados no mundo entre os anos de 2015 e 2016 (Distrowatch, 2016). Esta inclinação para utilização do sistema operacional Linux na área da bioinformática também pode ser observada em alguns fóruns de discussão (Researchgate, 2016; Biostars, 2016), mostrando-se ser uma linha a ser seguida.

Outra característica comum a todos os anotadores é a falta de uma *interface* gráfica. Isso acarreta em vantagens e desvantagens. Dentre as principais vantagens pode-se observar a facilidade de utilizar o programa remotamente em servidores e em scripts, e a utilização de uma pequena fração de memória RAM e do processamento, pela ausência da parte gráfica. Por outro lado, isso dificulta a utilização por usuários não familiarizados com este sistema operacional. Além disso,

a quantidade de memória e processamento liberado é praticamente irrelevante, dado a potência dos computadores atuais.

2.2 Softwares de execução em servidores na internet

Dentre os quinze anotadores gratuitos encontrados na internet, dez deles possui um acesso via internet, ou seja, o serviço é disponibilizado *online,* sem a necessidade de *download* e instalação de aplicativos. A diferença real entre eles é pouca, pesando mais ao usuário a questão da agilidade do processo de anotação e as filas de espera para iniciar o processo. Abaixo observa-se uma descrição sobre os anotadores gratuitos *online* mais utilizados em artigos de anotação de novos genomas e reanotações.

2.2.1 RAST

O *Rapid Annotations using Subsystems Technology* ou RAST (Aziz et al., 2008) é um dos anotadores mais utilizados no mundo. Sua *interface* limpa e o passo a passo guiam o usuário de forma fácil e rápida por todas as etapas de submissão dos arquivos a serem anotados. Mesmo possuindo essa interface sucinta, permite a escolha entre dois diferentes algoritmos de anotação: o padrão RAST (Overbeek et al., 2014) e o novo RASTtk (Brettin et al., 2015). Além disso, ele também possibilita ajustar todos os parâmetros que esses modelos apresentam, o que permite aos usuários mais experientes o ajuste fino do processo de anotação. Seu arquivo de entrada é o FASTA e para saída apresenta diversos, dentre eles EMBL, GenBank e GFF3. Possui também um visualizador integrado permitindo visualizar o genoma e também gera um gráfico estilo pizza representando a quantidade de genes em cada função biológica.

2.2.2 BASys

Bacterial Annotation System ou BASys (Van Domselaar et al., 2005) é um serviço *online* de anotação de genomas bacterianos. Sua submissão é feita através de um único formulário e não necessita que o usuário esteja registrado para efetuar a tarefa, requerendo somente um endereço de *e-mail* para alertar sobre o andamento do processo de anotação. Seu arquivo padrão, tanto de entrada quanto de saída, são no formato FASTA. O BASys também gera uma representação circular do genoma. Além da predição padrão pelo Glimmer (Delcher et al., 2007), é possível submeter um arquivo tabular com a predição ou do GeneMark HMM (Lukashin & Borodovsky, 1998).

2.2.3 KAAS

KAAS ou KEGG *Automatic Annotation Server* (Moriya et al., 2007) diferencia-se dos outros anotadores *online* por possuir um reconstrutor de rotas metabólicas incluso. Utiliza o banco de dados do KEEG e o algoritmo do KEEG Orthology para identificação das similaridades. Possui um formulário simples para submissão dos genes a serem anotados e não necessita de cadastro prévio para utilização dos serviços.

3 HIPÓTESE E OBJETIVOS

3.1 Hipótese

É possível facilitar a interação de novos usuários com o processo de anotação de genomas bacterianos, criando um anotador com *interface* gráfica e amigável, de fácil instalação, sem comprometer o desempenho do processo e a quantidade da anotação.

3.2 Objetivo Geral

Desenvolver uma *pipeline* de anotação de genomas bacterianos gratuito, de fácil instalação e com *interface* gráfica, que conte com as informações necessárias para submissão do genoma ao GenBank.

3.3 Objetivos Específicos

- Analisar as metodologias empregadas para anotação de genomas;
- Buscar por softwares para cada uma das fases de anotação;
- Avaliar os programas que irão compor a *pipeline*
- Desenvolver a *pipeline*
- Criar uma *interface* gráfica amigável
- Encontrar uma maneira fácil para instalar e executar o software em ambiente Linux
- Testar o novo software em genomas já anotados, para verificar sua eficiência
- Comparar os resultados de anotação do Square com outros anotadores populares.

4 MATERIAL E MÉTODOS

Para o desenvolvimento do Square, o modelo clássico de anotação foi adotado, contando com um preditor de genes, um preditor de ácido ribonucleico, um algoritmo para busca de similaridade e um banco de dados. Para melhor compreensão, cada uma destas etapas foi dividida em um item abaixo, e o conjunto dos processos é apresentado na Figura 1.

- Predição das possíveis regiões codificantes: Em um genoma não anotado, algumas predições são necessárias para inferência de dados, tais como localização das possíveis regiões codificantes ou CDS dentro de um genoma, localização celular da proteína, família e domínio proteico, RNAs, dentre outros. Alguns algoritmos de predições estão disponíveis, cada um com sua metodologia distinta. Dentre os populares estão o GenemarkHMM (Lukashin & Borodovsky, 1998), Genemark (Besemer & Borodovsky, 2005), EasyGene (Larsen & Krogh, 2003), MED (Zhu et al., 2007), Glimmer (Delcher et al., 2007) e o Prodigal (HYATT et al., 2010). O preditor Glimmer é o mais recorrente nas *pipelines* de anotação descritas neste trabalho, porém outro software chamou a atenção, o Prodigal (Hyatt et al., 2010), que de acordo com os testes apresentados em seu artigo, demonstrou-se mais rápido e com alta sensibilidade para encontrar regiões codificantes. Além destas características funcionais, o Prodigal possui uma fácil instalação, sendo uma ótima escolha para *pipelines* instaláveis. Por esses motivos, o Prodigal passou a ser o preditor padrão do Square.

- Predição de *RNAs:* Para predição dos *tRNAs*, o tRNAscan-SE (Lowe & Eddy, 1997) foi escolhido, possuindo grande sensibilidade em encontrar possíveis sequências de RNA em sequências de DNA, atingindo uma precisão entre 99% e 100%, sendo um dos mais utilizados no mundo.

- Inferência de dados: Uma forma eficiente de inferir dados sobre uma região de leitura de um genoma é compará-lo, com uma sequência de um organismo já sequenciado e anotado ou fragmentos anotados de um genoma. Para este fim, o software NCBI BLAST+ (Camacho et al., 2009) foi utilizado, percorrendo o banco de dados apontado pelo usuário, encontrando informações de cada um das CDS encontradas pelo Prodigal. Os resultados saem em formato xml, com tags padrões do NCBI BLAST+ contendo as sequências mais similares para cada uma das CDS. O Square captura a sequência mais similar e adiciona em um banco de dados SQLite3 (SQLite, 2015).

- Bancos de dados: Diversos bancos de dados biológicos estão disponíveis, com distintos tipos de informações. Nesta fase da pesquisa, um genoma já anotado teve suas informações removidas e uma nova anotação foi feita utilizando diferentes combinações de bancos de dados *online* e gratuitos, dentre eles o GenBank (BENSON et al., 2012), Uniprot (MAGRANE & CONSORTIUM, 2011) e Gene Ontology (ASHBURNER et al., 2000). Os melhores resultados foram obtidos com os bancos de dados do Uniprot, por possuir um cabeçalho padronizado, e no caso do Swiss-Prot, ainda conta com a curadoria manual das informações. No Square há uma opção para *download* automático dos bancos de dados Swiss-Prot e TrEMBL, e a criação automatizada de bancos de dados próprios a partir de um arquivo FASTA.

- Núcleo de processamento: Para o processamento dos dados foi utilizada a linguagem de programação Python (G. van Rossum and F.L. Drake, 2001) e o banco de dados SQLite (SQLite, 2015).

- Interface Gráfica: Para implementação do ambiente gráfico de trabalho, a linguagem Pascal foi utilizada através da IDE Lazarus 1.3 (Lazarus, 2016).

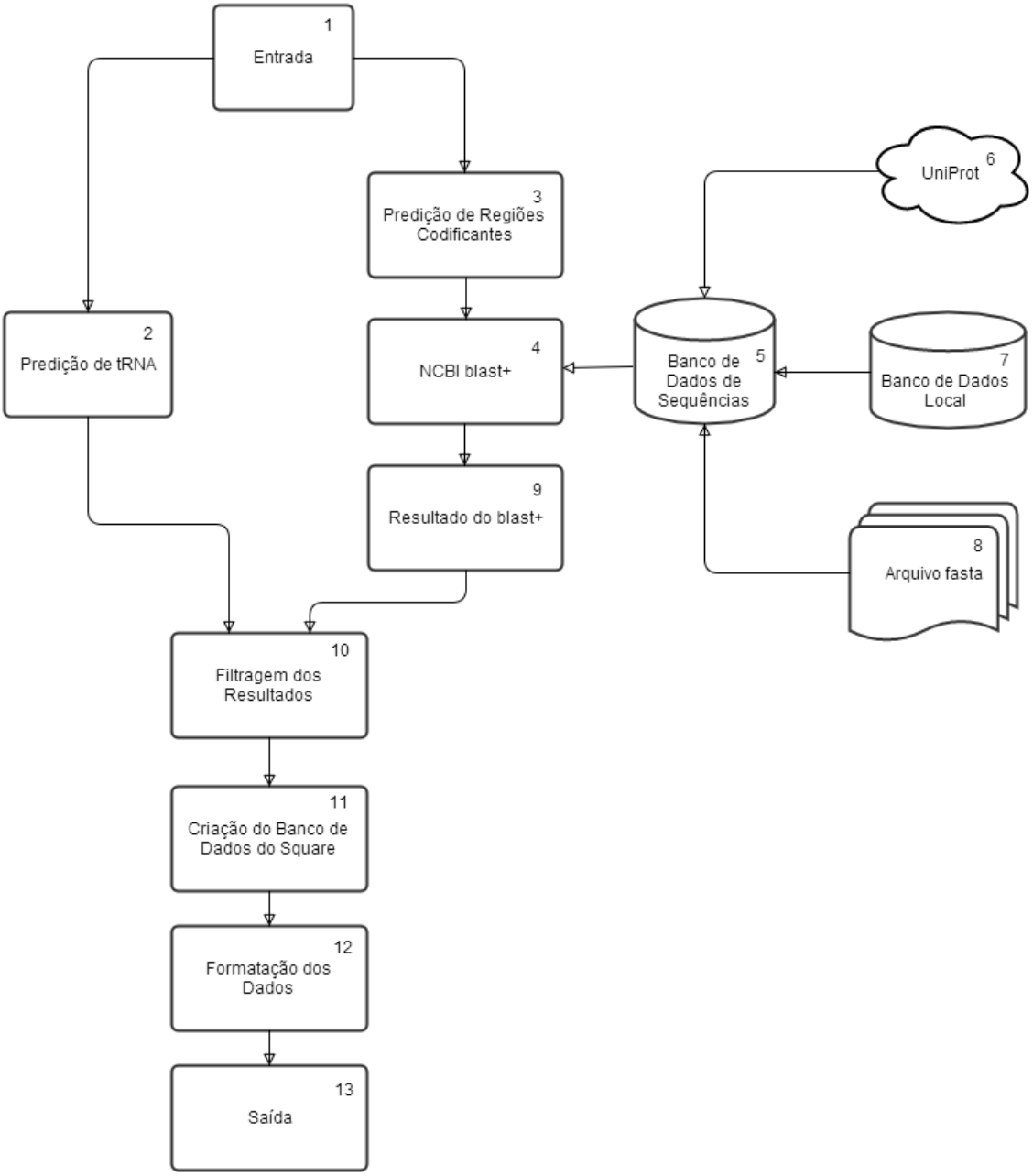

Figura 1. Diagrama de fluxo de dados dentro da pipeline Square.

- A Figura 1 demonstra o fluxo de dados da pipeline Square, onde um arquivo FASTA é apresentado na entrada (1), este arquivo é enviado ao preditor de tRNA (2) e ao preditor de CDs (3), para cada gene predito o NCBI BLAST+ (4) é executado utilizando-se de banco de dados (5) que pode ser de origem *online* através de dados do UniProt (6), onde os usuários podem filtrar pelo nome de um gênero e/ou organismo, e também pelo código taxonômico do organismo, outra fonte de dados é um banco de dados local já formatado (7) e também através de um arquivo FASTA

(8). Ao finalizar a comparação com o NCBI BLAST+, os dados resultantes deste processo(9) são filtrados, obtendo-se o melhor resultado para cada gene(10), estes resultados são encaminhados para um banco de dados SQLite(11) para posterior formatação(12) em formato Genbank, resultando em um arquivo com extensão .gbk na saída do Square(13)

5 RESULTADOS

Ao final do desenvolvimento foi obtida uma *pipeline* em duas versões, uma em modo texto (Figura 2), que pode ser rodado facilmente em terminais Linux, de forma local ou remota, e uma versão gráfica (Figura 3), que pode ser executada com um clique em ambiente gráfico Linux sem nenhuma necessidade de utilizar o terminal para executá-lo como super usuário. Para garantir as permissões necessárias, uma janela é aberta através do pacote Xdialog (Godefroy, 2016), que pergunta ao usuário a senha root, fornecendo ao Square a permissão total para ler, criar e modificar os arquivos de anotação durante todo o processo. Sua distribuição é feita através do endereço http://sourceforge.net/projects/sqgenome/, contando, além das versões gráficas e modo texto, com um tutorial explicando o passo a passo do *download*, instalação e utilização. O registro deste software está feito no INPI através do código BR 51 2013 001055 1.

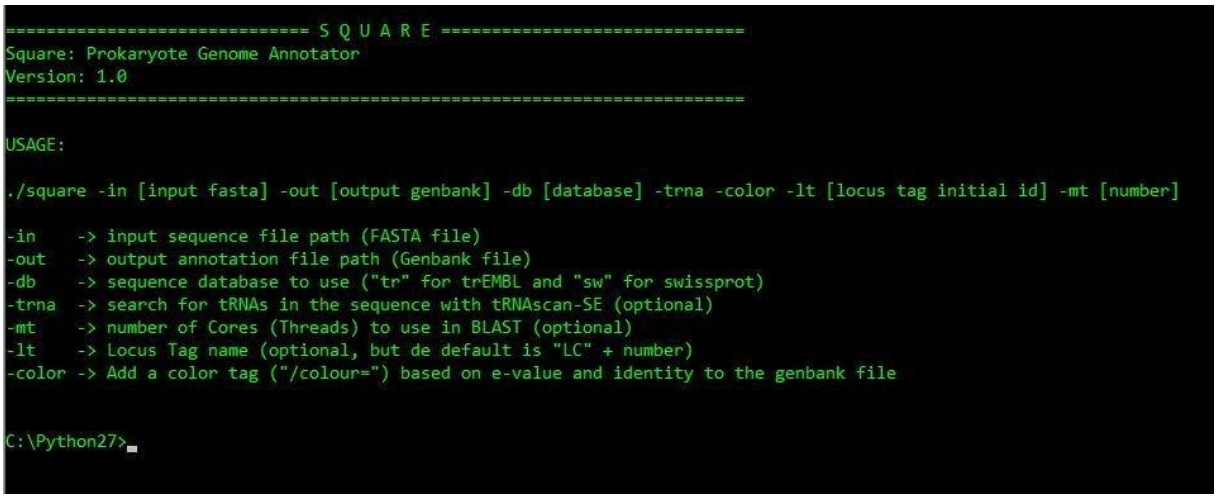

Figura 2. Square em modo texto.

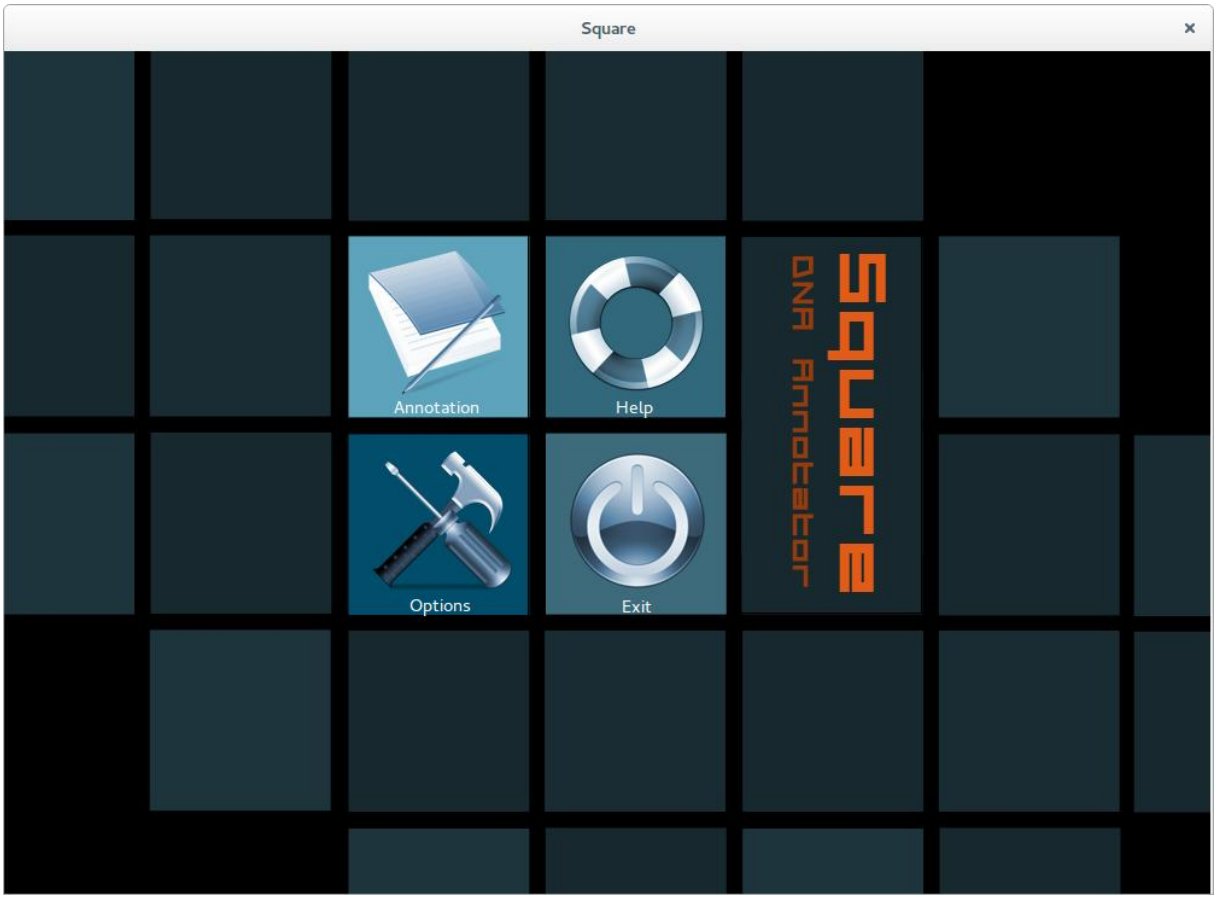

Figura 3. Interface gráfica do programa Square.

Além disso, funcionalidades como *download*, instalação e formatação dos bancos de dados com somente um clique, permitem ao usuário baixar os bancos de dados Swiss-Prot e TrEMBL, bem como criar os próprios bancos de dados a partir de um grupo de sequências de proteínas em formato FASTA. Para o processo de anotação, o software possui um passo a passo que guia o usuário, impedindo que informações erradas sejam inseridas e que o usuário fique perdido, e uma checagem no formato do arquivo inserido, impede que erros ocorram por incompatibilidade no formato ou informações erradas. Ao final da anotação, um arquivo no formato GenBank é criado com as informações necessárias para submissão ao NCBI GenBank, e um sistema de cores para cada um dos genes encontrados indica a porcentagem de identidade e informações encontradas. A cor verde indica a identidade maior ou igual a 95% e *e-value* 0.0 e todas as informações foram encontradas, a cor amarela indica a identidade entre 85% e 95% e *e-value* 0.0, ou anotações com alguma informação faltante, e por último, a cor vermelha indica uma identidade menor que 85% e *e-value* maior que 0.0. A visualização dos

dados pode ser feito por qualquer software visualizador que aceite o formato GenBank, como por exemplo o Artemis (Rutherford et al., 2000).

Para testar o Square, foram escolhidos dois anotadores populares utilizados por muitos dos usuários iniciantes: o RAST e o BASys. Como genomas a serem anotados, foram escolhidos o maior genoma anotado, o menor genoma anotado e um com o tamanho médio, todos encontrados no GenBank, sendo eles respectivamente Sorangium_cellulosum_So0157_2 , Candidatus_Hodgkinia_cicadicola_strain_TETUND1 e Burkholderia_pseudomallei_Pasteur_52237. O computador utilizado para rodar o Square possui um processador Intel i3-2100, 4gb de memória RAM e sistema operacional Ubuntu 14.03, onde cada genoma foi anotado 20 vezes, sem apresentar divergências em nenhum dos genomas. O resultado da comparação pode ser observado na Tabela 1.

Alguns pontos relevantes na Tabela 1 podem ser observados, o primeiro é o número total de genes anotados, referindo-se ao total de genes preditos, onde o Square apresentou um número bem próximo ao depositado no NCBI. Cada uma destas predições ao ser confrontado com o genoma base gerou uma porcentagem de identidade, ou seja, quão idêntico a sequência predita é parecida com o gene depositado, gerando assim um índice de qualidade que foi dividido em dois grupos, um com identidades superiores a 70% e outro com identidades inferiores a 70%, onde as predições que apresentaram menos de 70% da sequência idêntica são considerados de baixa qualidade, podendo revelar alguns problemas como, por exemplo, predições equivocadas. A coluna Total NCBI – Identidade (=>70%) pode dar uma ideia de quantos genes corretos devem estar entre as predições de baixa qualidade, enquanto que, a coluna Total NCBI X Total Software, mostra quantos genes a mais, que os informados no genoma apresentado no NCBI, foram preditos. Incluso na tabela também pode ser observado o tempo total de anotação de cada genoma em cada um dos anotadores, contando do final da submissão nos serviços *online* e do clique no botão anotar do Square até a obtenção do resultado.

Tabela 1. Avaliação comparativa do Square, RAST e BASys contra os dados depositados no GenBank

Software / Organismo	Genes anotados	Identidade >70%	Identidade =<70%	Genes não encontrados	Genes exclusivos	Tempo
Candidatus_Hodgkinia_cicadicola_strain						
GENBANK	121	121	0	0	0	0
Square	266	48	218	0	1	0h02m21s
RAST	365	51	314	1	4	03h44m
BASys	462	49	413	1	5	56h22m
Burkholderia_pseudomallei_Pasteur_522						
GENBANK	3647	3647	0	0	0	0
Square	3572	3430	142	1	0	02h19m
RAST	4284	3475	809	1	1	04h16m
BASys	5793	2004	3789	16	45	32h03m
Burkholderia_pseudomallei_Pasteur_522						
GENBANK	2520	2520	0	0	0	0
Square	2439	2303	136	3	1	01h45m
RAST	3294	2379	915	1	5	03h58m
BASys	4300	1876	2424	5	25	54h59m
Sorangium_cellulosum_So0157_2						
GENBANK	10400	10400	0	0	0	0
Square	10954	9772	1182	11	5	02h16m
RAST	13571	9777	3794	7	11	07h07m
BASys	16317	6238	10079	47	49	176h11m

6 DISCUSSÃO E PERSPECTIVAS FUTURAS

Tendo em vista a facilidade do sequenciamento de DNA, vários pesquisadores têm se aventurado nesta área, usando-a como complemento de suas pesquisas. Porém, a ausência de softwares amigáveis torna-se um empecilho, forçando muitas vezes o pesquisador a investir seu tempo em estudos na área de tecnologia da informação, o que acaba reduzindo o tempo para sua pesquisa central. O Square apresenta a possibilidade de realização da anotação de genomas bacterianos, sem a necessidade de um conhecimento prévio específico e com uma interface amigável.

Atualmente tem se observado uma expansão, no número de projetos com genomas procariotos. Segundo Richardson & Watson (Richardson & Watson, 2013), isso se deve à popularização dos sequenciadores de nova geração, que reduziram custos e agilizaram o processo para gerar dados de sequências biológicas. Porém, após a aquisição das sequências, elas passam por um processo de anotação, que consiste basicamente em rodar alguns programas de predição e de busca de similaridade, para posteriormente realizar uma curadoria manual (Stothard & Wishart, 2006). Essa metodologia tem se tornado padrão em várias ferramentas de anotação, onde *pipelines* tem sido utilizadas para unir programas e criar um fluxo de dados automatizado.

Dependendo do interesse do pesquisador, a anotação pode ser, por exemplo, para deposição em um banco de dados público ou para elucidar algumas questões. Reed et al. (2006) classifica a anotação em unidimensional, que seria uma anotação identificativa para o depósito em bancos de dados e posterior estudo, e multidimensional, que analisaria cada uma das interações e participação celular, de cada proteína contida em um determinado genoma. Ao observar os programas para anotação disponíveis, em sua grande maioria o foco é a anotação unidimensional. Isso se deve pela grande variedade de análises que um pesquisador pode querer realizar, necessitando de um grande número de ferramentas e um conhecimento razoável na área de tecnologia da informação. O Square, assim como a maioria dos anotadores, toma como base a anotação unidimensional, e o foco no depósito do genoma no GenBank (Benson et al., 2012). Porém, para o depósito neste banco de dados, certas regras devem ser seguidas, como descritas no *The NCBI Handbook*

(Tatusova, 2013). Isso auxilia na padronização do banco de dados, agilizando as buscas e futuras anotações.

O desenvolvimento do Square, tem o foco em seu usuário, buscando facilitar e melhorar a experiência do mesmo, sem esquecer de sua função primordial, a anotação. Para demonstrar seu desempenho quanto a precisão e velocidade, os anotadores RAST (Aziz et al., 2008) e BASys (Van Domselaar et al., 2005) foram escolhidos para uma comparação. Mesmo sendo serviços *online,* estes são os anotadores mais citados, e possuem uma interface amigável, assim como o programa deste trabalho. Para execução do Square durante os testes, foi utilizado um *desktop* com uma configuração popular comumente encontrado em vários desktops e notebooks.

Ao observar a Tabela 1 pode-se identificar alguns pontos pertinentes. O Square foi o que mais se aproximou do número de genes depositados no NCBI. Nos gêneros *Candidatus* e *Sorangium,* o Square encontrou um número superior de genes, podendo ser eles falso positivos ou novos genes. O mesmo pode ser observado pelos outros anotadores, com uma quantidade ainda maior ao predito pelo Square. Já no gênero *Burkholderia,* os anotadores RAST e BASys encontraram uma quantidade maior de genes do que depositado no NCBI, enquanto que o Square encontrou um número levemente inferior de genes.

A característica do número de genes encontrados é diretamente ligada ao preditor. No caso do anotador BASys, o preditor padrão é o Glimmer, no RAST o padrão é um algoritmo próprio, e no Square o Prodigal é utilizado como padrão, o que explica a diferença entre o número de genes preditos. Uma indicação que este gene pode ter sido predito erroneamente é a identidade inferior a 70%, em relação a sequência depositada no NCBI, o que indica que o códon de iniciação pode ter sido predito de forma errada ou que este não é um gene transcrito. A quantidade de genes interfere na demanda de tempo pela equipe de curadoria. Uma quantidade muito diferente de genes demanda muito tempo para verificar se esses genes são verdadeiros ou falsos, e até mesmo faltante. Atualmente a curadoria, mesmo possuindo um alto custo em recursos humanos, continua sendo fundamental para aumentar a credibilidade dos dados depositados (Ten Hoopen et al., 2016).

No caso do gênero *Burkholderia,* a quantidade inferior de genes preditos pelo Square pode ser causada por três motivos diferentes. A primeira delas é o preditor Prodigal que pode não ter sido o mais adequado para este genoma; a segunda é que os dados depositados no NCBI estão equivocados e genes a mais estão presentes; e por último, características típicas deste genoma, como regiões repetitivas ou quebradas, genes curtos ou transposases, podem ter interferido na predição. Para responder essa questão é necessário estudar melhor esta característica, observando quais os genes faltantes na predição e os que estão a mais depositados no NCBI. Além disso é necessário testar o Square, o RAST e o BASys em outros genomas para verificar se essas características se repetem.

O tempo de anotação é maior nos servidores *online* por possuir filas, sendo altamente dependente da quantidade de solicitações no servidor e tamanho dos genomas submetidos para anotação. Estes anotadores *online* apresentam uma grande variação no tempo total entre a submissão e a obtenção do resultado. Já no Square, por rodar localmente, os fatores determinantes do tempo de anotação são o número de genes de um genoma, o tamanho do banco de dados utilizado na anotação e a capacidade de processamento da máquina utilizada. Essas diferenças explicam o tempo reduzido do Square perante os outros anotadores testados.

Como perspectiva futura, outras ferramentas podem ser adicionadas, como por exemplo, o Glimmer (Delcher et al., 2007), dando ao usuário a escolha de outro preditor, além do Prodigal, podendo assim obter resultados ligeiramente diferentes para um mesmo genoma. Ao final do processo haveria a possibilidade do usuário escolher a melhor solução para a espécie que se está trabalhando. Outra ferramenta interessante é o banco de dados AntiFam (Eberhardt et al., 2012), capaz de aumentar a precisão da anotação, já que esta ferramenta apresenta um conjunto de informações sobre genes falso positivo do UniProt, sendo possível identificar falsas ORFs através do algoritmo BLAST, reduzindo assim o tempo de anotação automática e da curadoria manual, por excluir os genes falsos da fila de anotação.

Outro ponto que pode ser enriquecido é a predição de RNAs utilizando as ferramentas Infernal (Nawrocki & Eddy, 2013), RNAmmer (Karin Lagesen et al., 2007) e Aragorn (Laslett & Canback, 2004). Da mesma forma que os preditores de ORFs, os preditores de RNAs possuem metodologias diferentes, podendo gerar

resultados ligeiramente diferentes, que podem ser confrontados gerando um consenso do resultado final.

Por fim, a possibilidade de tornar o Square um software multi plataforma é real. Sua base escrita em Python, por ser interpretada, já possui suporte em vários sistemas operacionais. Já a interface gráfica, criada na IDE Lazarus, permite que o Square seja desenvolvido em Linux, Mac OS e Windows. O único empecilho é que os aplicativos que compõe a *pipeline* tem que ser compatíveis com a plataforma que será executado o Square.

7 CONCLUSÃO

O Square é uma nova alternativa para anotação de genomas bacterianos, sendo atualmente a única gratuita com *interface* gráfica, especialmente atrativo a pesquisadores que estão iniciando na área de sequenciamento de genomas. O Square é mais preciso que os demais softwares testados, pois, foi o que chegou mais perto do número total de genes anotados depositados no NCBI, bem como o que proporcionalmente encontrou um número maior de genes, com identidade superior a 70%, e também um tempo menor de anotação.

Sendo assim presente trabalho atingiu o seu objetivo com o desenvolvimento do Square. Com o foco em novos usuários, sua *interface* gráfica foi simplificada ao máximo, permitindo uma anotação em poucos cliques. Seu passo a passo impede que o usuário tenha dúvidas da ação a ser tomada, e verificações impedem que dados no formato errado e/ou incompleto sejam inseridos. Sua precisão demonstrou-se melhor em alguns genomas, como nos gêneros *Candidatus* e *Sorangium, e levemente inferior no Burkholderia,* precisando de mais testes com diversos genomas para melhorar o desempenho nos genomas que obtiverem resultados insatisfatórios. Alguns incrementos podem ser feitos para melhorar a sensibilidade de predição e a facilidade do usuário, podendo tornar-se um software de referência para novos usuários.

8 REFERÊNCIAS

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*. 25(17). p. 3389–3402.

AZIZ, R.K., BARTELS, D., BEST, A.A., DEJONGH, M., DISZ, T., EDWARDS, R.A., FORMSMA, K., GERDES, S., GLASS, E.M., KUBAL, M., MEYER, F., OLSEN, G.J., OLSON, R., OSTERMAN, A.L., OVERBEEK, R.A., MCNEIL, L.K., PAARMANN, D., PACZIAN, T., PARRELLO, B., ET AL. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC.Genomics*. 9. p. 75–.

BARRETT, T., CLARK, K., GEVORGYAN, R., GORELENKOV, V., GRIBOV, E., KARSCH-MIZRACHI, I., KIMELMAN, M., PRUITT, K.D., RESENCHUK, S., TATUSOVA, T., YASCHENKO, E. & OSTELL, J. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*. 40(Database issue). p. D57–D63.

BENSON, D.A., KARSCH-MIZRACHI, I., CLARK, K., LIPMAN, D.J., OSTELL, J. & SAYERS, E.W. (2012). GenBank. *Nucleic Acids Res.*. 40(1362-4962 (Electronic)). p. D48–D53.

BESEMER, J. & BORODOVSKY, M. (2005). GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research*. 33(SUPPL. 2). p. 451–454.

BIOSTARS (2016). Bioinformatics Explained. Available at: www.biostars.org.

BRETTIN, T., DAVIS, J.J., DISZ, T., EDWARDS, R.A., GERDES, S., OLSEN, G.J., OLSON, R., OVERBEEK, R., PARRELLO, B., PUSCH, G.D., SHUKLA, M., THOMASON, J.A., STEVENS, R., VONSTEIN, V., WATTAM, A.R. & XIA, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*. 5. p. 8365. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25666585\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4322359.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T.L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*. 10. p. 421.

CAMPBELL, M.S., LAW, M., HOLT, C., STEIN, J.C., MOGHE, G.D., HUFNAGEL, D.E., LEI, J., ACHAWANANTAKUN, R., JIAO, D., LAWRENCE, C.J., WARE, D., SHIU, S.-H., CHILDS, K.L., SUN, Y., JIANG, N. & YANDELL, M. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant physiology*. 164(2). p. 513–24. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84893466287&partnerID=tZOtx3y1.

CANTAREL, B.L., KORF, I., ROBB, S.M.C., PARRA, G., ROSS, E., MOORE, B., HOLT, C., ALVARADO, A.S. & YANDELL, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*. 18(1). p. 188–196.

CHAUDHURI, R.R., LOMAN, N.J., SNYDER, L.A., BAILEY, C.M., STEKEL, D.J. & PALLEN, M.J. (2008). xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.*. 36(Database issue). p. D543–D546.

CONESA, A., GOTZ, S., GARCIA-GOMEZ, J.M., TEROL, J., TALON, M. & ROBLES, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.*. 21(18). p. 3674–3676.

DELCHER, A.L., BRATKE, K.A., POWERS, E.C. & SALZBERG, S.L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*. 23(6). p. 673–679.

DISTROWATCH (2016). Page Hit Ranking. Available at: http://distrowatch.com/.

VAN DOMSELAAR, G.H., STOTHARD, P., SHRIVASTAVA, S., CRUZ, J.A., GUO, A., DONG, X., LU, P., SZAFRON, D., GREINER, R. & WISHART, D.S. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*. 33(Web Server issue). p. W455–W459.

EBERHARDT, R.Y., HAFT, D.H., PUNTA, M., MARTIN, M., O'DONOVAN, C. & BATEMAN, A. (2012). AntiFam: A tool to help identify spurious ORFs in protein annotation. *Database*. 2012.

FLEISCHMANN, R.D., ADAMS, M.D., WHITE, O., CLAYTON, R.A., KIRKNESS, E.F., KERLAVAGE, A.R., BULT, C.J., TOMB, J.F., DOUGHERTY, B.A., MERRICK, J.M. & . (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*. 269(5223). p. 496–512.

G. VAN ROSSUM AND F.L. DRAKE (2001). Python Reference Manual.

GODEFROY, T. (2016). Xdialog. Available at: http://xdialog.free.fr/.

TEN HOOPEN, P., AMID, C., LUIGI BUTTIGIEG, P., PAFILIS, E., BRAVAKOS, P., CERDEÑO-TÁRRAGA, A.M., GIBSON, R., KAHLKE, T., LEGAKI, A., NARAYANA MURTHY, K., PAPASTEFANOU, G., PEREIRA, E., ROSSELLO, M., LUISA TORIBIO, A. & COCHRANE, G. (2016). Value, but high costs in post-deposition data curation. *Database*. 2016. p. BAV126. Available at: http://database.oxfordjournals.org/lookup/doi/10.1093/database/bav126.

HUI, P. (2012). Next Generation Sequencing: Chemistry, Technology and Applications. *Top.Curr.Chem.*. p. -.

HYATT, D., CHEN, G.L., LOCASCIO, P.F., LAND, M.L., LARIMER, F.W. & HAUSER, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC.Bioinformatics.*. 11. p. 119–.

ILLUMINA (2012). An Introduction to Next-Generation Sequencing Technology. . p. -. Available at: http://www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf.

KUMAR, K., DESAI, V., CHENG, L., KHITROV, M., GROVER, D., SATYA, R. V, YU, C., ZAVALJEVSKI, N. & REIFMAN, J. (2011). AGeS: a software system for microbial genome sequence annotation. *PLoS.One.*. 6(3). p. E17469–.

LAGESEN, K., HALLIN, P., RODLAND, E.A., STAERFELDT, H.H., ROGNES, T. & USSERY, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Researsh*. 35(9). p. 3100–3108.

LAGESEN, K., HALLIN, P., RØDLAND, E.A., STAERFELDT, H.-H., ROGNES, T. & USSERY, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic acids research*. 35(9). p. 3100–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1888812&tool=pmcentrez&rendertype=abstract.

LARSEN, T.S. & KROGH, A. (2003). EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC bioinformatics*. 4(1). p. 21. Available at: http://www.biomedcentral.com/1471-2105/4/21.

LASLETT, D. & CANBACK, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*. 32(1). p. 11–16.

LAZARUS (2016). www.lazarus-ide.org. Available at: www.lazarus-ide.org.

LOWE, T.M. & EDDY, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*. 25(5). p. 955–64. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146525&tool=pmcentrez&rendertype=abstract.

LUKASHIN, A. V & BORODOVSKY, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic acids research*. 26(4). p. 1107–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147337&tool=pmcentrez&rendertype=abstract.

MAGRANE, M. & CONSORTIUM, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database.(Oxford)*. 2011. p. BAR009–.

METZKER, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Microbiology*. 11(1). p. 31–46.

MORIYA, Y., ITOH, M., OKUDA, S., YOSHIZAWA, A.C. & KANEHISA, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*. 35(Web Server issue). p. W182–W185.

NAWROCKI, E.P. & EDDY, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.*. 29(22). p. 2933–2935.

OVERBEEK, R., OLSON, R., PUSCH, G.D., OLSEN, G.J., DAVIS, J.J., DISZ, T., EDWARDS, R.A., GERDES, S., PARRELLO, B., SHUKLA, M., VONSTEIN, V., WATTAM, A.R., XIA, F. & STEVENS, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*. 42(Database issue). p. D206–14. Available at: http://nar.oxfordjournals.org/content/42/D1/D206.short.

PAREJA-TOBES, P., MANRIQUE, M., PAREJA-TOBES, E., PAREJA, E. & TOBES, R. (2012). BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PloS one*. 7(11). p. E49239. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3504008&tool=pmcentrez&rendertype=abstract.

PETERSEN, T.N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*. 8(10). p. 785–786. Available at: http://dx.doi.org/10.1038/nmeth.1701.

REED, J.L., FAMILI, I., THIELE, I. & PALSSON, B.O. (2006). Towards multidimensional

genome annotation. *Nature reviews Genetics*. 7(2). p. 130–141.

RESEARCHGATE (2016). Advance your research. Available at: www.researchgate.net.

RICHARDSON, E.J. & WATSON, M. (2013). The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*. 14(1). p. 1–12. Available at: http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbs007.

RONAGHI, M., KARAMOHAMED, S., PETTERSSON, B., UHLEN, M. & NYREN, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Anal.Biochem.*. 242(1). p. 84–89.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.A. & BARRELL, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics.*. 16(10). p. 944–945.

SALLET, E., GOUZY, J. & SCHIEX, T. (2014). EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics (Oxford, England)*. 30(18). p. 2659–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24880686.

SALLET, E., ROUX, B., SAUVIAC, L., JARDINAUD, M.-F., CARRÈRE, S., FARAUT, T., DE CARVALHO-NIEBEL, F., GOUZY, J., GAMAS, P., CAPELA, D., BRUAND, C. & SCHIEX, T. (2013). Next-generation annotation of prokaryotic genomes with EuGene-P: application to Sinorhizobium meliloti 2011. *DNA research : an international journal for rapid publication of reports on genes and genomes*. 20(4). p. 339–54. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3738161&tool=pmcentrez&rendertype=abstract.

SANGER, F., NICKLEN, S. & COULSON, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.U.S.A*. 74(12). p. 5463–5467.

SEEMANN, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*. 30(14). p. 2068–2069.

SERVICE, R.F. (2006). Gene sequencing. The race for the $1000 genome. *Science*. 311(5767). p. 1544–1546.

SMITH, L.M., SANDERS, J.Z., KAISER, R.J., HUGHES, P., DODD, C., CONNELL, C.R., HEINER, C., KENT, S.B. & HOOD, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*. 321(6071). p. 674–679.

SQLITE (2015). www.sqlite.org.

STEWART, A.C., OSBORNE, B. & READ, T.D. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. *Bioinformatics*. 25(7). p. 962–963. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp097.

STOTHARD, P. & WISHART, D.S. (2006). Automated bacterial genome analysis and annotation. *Current Opinion in Microbiology*. 9(5). p. 505–510.

TATUSOVA, T. (2013). *Prokaryotic Genome Annotation Pipeline*,

ZHU, H., HU, G.-Q., YANG, Y.-F., WANG, J. & SHE, Z.-S. (2007). MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC bioinformatics*. 8. p. 97.

9 ANEXOS

Anexo A – Registro de Software

INPI

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

protocolo
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
PROTOCOLO GERAL
03/10/2013 12:37 DERS
016130003604
BR 51 2013 001055 1

IDENTIFICAÇÃO DO PEDIDO (Para uso do INPI)

Número do Pedido	Protocolo, Data e Hora

DADOS DO AUTOR DO PROGRAMA

Nº de Autores: 7 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF*: 015.121.840-47
Nome: MARCUS REDÜ ESLABÃO
Nome Abreviado, pseudônimo ou sinal convencional (se houver):
Data de Nascimento: 05/06/1986 — Nacionalidade: BRASILEIRO
Endereço: AV. PINHEIRO MACHADO 901 CASA 16
Cidade: PELOTAS — UF: RS — País: BRASIL
CEP: 96.040-500 — Telefone: 5332211624 — FAX:
E-mail: marcus.eslabao@yahoo.com.br

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Nº de Titulares: 1 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF/CNPJ*: 92242080000100
Nome/Razão Social: UNIVERSIDADE FEDERAL DE PELOTAS
Nome abreviado, pseudônimo ou sinal convencional (se houver):
Data de Nascimento: — Nacionalidade/Origem: BRASIL
Endereço: RUA GOMES CARNEIRO, Nº 01
Cidade: PELOTAS — UF: RS — País: BRASIL
CEP: 96.010-610 — Telefone: -5332293090 — FAX:
E-mail: agtufpel@gmail.com

☒ SIM, este Titular é Pessoa Jurídica. Caso afirmativo, assinale a melhor classificação:

☐ Órgão Públi o ☐ Sociedade com Intuito não Econômico ☐ Microempresa ☐ Software House
☒ Instituição Pública de Ensino ou Pesquisa ☐ Instituição Privada de Ensino ou Pesquisa ☐ Outras

ENDEREÇO PARA CORRESPONDÊNCIA E CONTATO (Preencha apenas o necessário)

Toda correspondência será enviada para: ☐ O Procurador ou ☒ O Titular acima ou
☐ Escaninho nº ☐ Representação INPI em: ☐ O Endereço abaixo:

Nome:
Endereço:
Cidade: — UF: — País:
CEP: — Telefone: — FAX:
E-mail:

Modelo I (folha 1/2) E

DADOS DO PROGRAMA

Título	Square				
Data de Criação do Programa	05/02/2013	Regime de Guarda	☐ COM SIGILO	☒ SEM SIGILO	
Linguagens	Python	Pascal			
Classificação do Campo de Aplicação	SD - 09	BL - 01	BL - 02	BL - 04	-
Classificação do Tipo de Programa	IT - 03	TC - 01	-	-	-

☐ **SIM**, este Programa é Modificação Tecnológica ou Derivação. Caso afirmativo, informe Título do Programa Original e (se houver) Número de Registro:

Título do Programa Original

☐ **SIM**, este Registro é composto por obra(s) de outra(s) natureza(s) de ordem intelectual. Caso afirmativo assinale-a(s) abaixo:

☐ Literária ☐ Musical ☐ Artes Plásticas ☐ Áudio-Visual ☐ Arquitetura ☐ Engenharia

DOCUMENTOS ANEXADOS (Informe as quantidades de documentos, não o número de páginas)

Quant	Nome	Quant	Nome
	Guia de Recolhimento		Contrato de Trabalho/Prestação de Serviço
	Procuração		Invólucros/mídia eletrônica Utilizados
	Termo de Cessão		Contrato/Estatuto Social e Alterações (ou equivalente)
	Termo de Autorização para Modificações Tecnológicas ou Derivações		Autorização para Cópia do CD
			Outros(especificar)

DECLARAÇÕES

DECLARO, PARA TODOS OS FINS DE DIREITO:

A) que estou ciente de **TODAS AS RECOMENDAÇÕES** constantes do "Manual do Usuário de Registro de Programas de Computador", **ESPECIALMENTE NO QUE TANGE AO TÍTULO E AOS DOCUMENTOS DO PROGRAMA**, bem como da legislação pertinente ao assunto, constante dos anexos "A" ; "B" ; "C" ; "E" e "F", do referido Manual;

B) que se deixar de solicitar a prorrogação do sigilo, nos casos necessários, estarei desistindo desse caráter de guarda dos documentos de programa do presente depósito, na forma do art. 3º, § 2º, da Lei 9.609, de 12 de fevereiro de 1998;

C) que, se devido à qualidade do papel ou à qualidade gráfica dos documentos sigilosos anexos ao presente, houver deterioração ou perda de seu conteúdo, nenhuma responsabilidade caberá ao INPI, desde que mantida a inviolabilidade dos invólucros (ressalvadas as hipóteses de serem abertos por ordem judicial ou motivo de força maior);

D) que em caso de perda do SIGILO ou dos documentos, por culpa exclusiva do INPI, a indenização por perdas e danos, porventura cabível, estará limitada a 20 (vinte) salários mínimos;

E) que devo manter guardado, em segurança e inviolado, o COMPARTIMENTO "3" do invólucro especial para depósito, que é restituído pelo INPI, para fins de recomposição do arquivo do Instituto, no caso de sua destruição total ou parcial por algum tipo de sinistro;

F) que deverei manter endereço atualizado junto à Divisão de Registro de Programa de Computador, a fim de garantir o recebimento das comunicações relativas ao andamento do meu pedido/registro, ressalvando o INPI de qualquer responsabilidade decorrente da não observação deste preceito.

DADOS DO PROCURADOR

CPF/CNPJ*	45048584015	Código do Procurador (se houver)	
Nome	GLENIO DO COUTO PINTO JUNIOR		
Endereço	RUA ANDRADE NEVES, Nº 1529		
Cidade	PELOTAS	UF: RS	País: BRASIL
CEP	96.020-080	Telefone: -5332293090	FAX:
E-mail	gleniocoutopinto@gmail.com		

DECLARO, SOB AS PENAS DA LEI, SEREM VERDADEIRAS AS INFORMAÇÕES PRESTADAS

Porto Alegre, 03/12/2013
Local/Data

Assinatura/Carimbo

SIAPE 2351077
ADMINISTRADOR

Modelo I (folha 2/2) E

Anexo B – Formulário de requisição de registro de software

UNIVERSIDADE FEDERAL DE PELOTAS
GABINETE DO REITOR
AGÊNCIA DE GESTÃO TECNOLÓGICA

Pelotas, 18 de fevereiro de 2016

De: Marcus Redü Eslabão, Doutorando PPGB-UFPel
Para: Diretor da Agência de Gestão Tecnológica e Propriedade Intelectual
Pró-Reitoria de Pesquisa e Pós-Graduação
Universidade Federal de Pelotas

Assunto: Pedido de Registro de Software

Prezado Diretor:

Venho pelo presente encaminhar a esta Agência o formulário em anexo para análise do pedido de registro de Software denominado "Título identificador do Software"

Atenciosamente,

Marcus Redü Eslabão
Doutorando/PPGB-UFPel

FORMULÁRIO PARA ANÁLISE DE REGISTRO DE SOFTWARE DADOS DOS AUTORES					
Nome civil completo: Marcus Redü Eslabão					
Unidade: CDTec		Departamento: Biotecnologia			
Fone comercial: (53) 3275-7350	E-mail: marcus.eslabao@yahoo.com.br		Fax:		
Identidade: 1083052504	Órgão expedidor: SJS		Data de emissão:		
CPF: 01512184047	Data nascimento: 05/06/1986		Estado Civil: Solteiro		
Nacionalidade: Brasileiro		Naturalidade: Brasileiro			
Endereço Residencial Completo: Av. Pinheiro Machado 901 CASA 16					
Bairro: Fragata		CEP: 96040-500			
Telefone Residencial: (53) 32211624		Celular: (53) 81013060			
Vínculo com a UFPEL:		Professor	Aluno:	X	Doutorado
		Técnico-administrativo			Mestrado
					Especialização
					Graduação
Participante Externo:	Profissão:		Instituição:		
% Contribuição no presente invento: 20%					

FORMULÁRIO PARA ANÁLISE DE REGISTRO DE SOFTWARE DADOS DO INVENTO
Título do Programa de Computador: Square
Data de Criação do Programa de Computador: 12/01/13
Linguagem(s) de Programação na(s) qual (is) foi desenvolvido e está disponibilizado o programa: Perl, Pascal
O presente programa de computador é uma modificação tecnológica ou derivação (nova versão) de outro já existente? **Caso afirmativo, informe o título do programa original** **Não**
Descrição funcional do programa de computador: O presente programa visa automatizar o processo de anotação de genomas de procariotos inferindo informações necessárias para depósito destes em bancos de dados públicos como NCBI. Possui um interface avançada em modo linha de comando e uma interface gráfica simples e fácil de usar, auxiliando pesquisadores não familiarizados com este tipo de processo.
Informe trechos do programa ou outros elementos essenciais do programa que sejam capazes de caracterizar a criação independente e identificar o programa: Localização de regiões codificadoras, busca de similaridade com outros organismos utilizando banco de dados de terceiros, inferência de informações como nome de gene e produto gênico em cada região codificadora e predição de RNAs.
Informe o campo de aplicação (tabela em anexo) do presente programa de computador, definindo áreas de aplicação: BL01,SD09,BL02,BL04
Informe a classificação do tipo de programa (tabela em anexo): IT03, TC01
Quais os problemas que o programa de computador resolve? Anotação manual de genomas.
Quais as vantagens que o programa de computador apresenta? Agilidade na anotação de genomas, redução do número pessoas envolvidas na anotação de genomas, redução de erro humano, maior precisão e quantidade de dados inferidos.
Qual o uso presente e futuro do programa de computador? Atualmente o presente software é utilizado na anotação de genomas de organismos procariotos, no futuro o programa deverá ser capaz de anotar genomas de organismos eucariotos.

Conhece outro software com característica similar? Caso afirmativo, informe o nome Blast2go, MAKER, VESPA, AGeS, RAST
Há pesquisa bibliográfica relacionada com o software? Sim
O presente software já foi revelado fora da Universidade? **Se afirmativo, informe detalhadamente as circunstâncias e anexe cópia do trabalho.** Não
Esteve pessoalmente envolvido em outro processo de registro de software? **Se afirmativo, informe quando, onde, e que tipo de software:** Não
O presente software já foi revelado à indústria? Não
Foi demonstrado interesse comercial? **Se afirmativo, informe nome, contato e telefone da empresa:** Não

FORMULÁRIO PARA ANÁLISE DE REGISTRO DE SOFTWARE DADOS DA PESQUISA
Órgãos de Fomento Envolvidos (Apoio CNPq, CAPES, FAPERGS, etc): CAPES
Foi feito contrato com órgão financiador ou gerido de acordo com um Termo de Confidencialidade? Não
O órgão financiador foi informado do invento? Não
Suporte Interno (Fundos de Pesquisa da UFPEL, Unidade ou do Departamento): Não

TRANSFERÊNCIA DE TECNOLOGIA
Comente a potencialidade de comercialização do presente software, incluindo sugestões a longo prazo e especificando as áreas de aplicação que possam utilizar o programa: Em curto prazo não há pretensão da comercialização do presente software, caso fosse vendido em larga escala, necessitaria de mais pessoas, possibilitando o oferecimento de suporte, atualizações e adequações do software aos clientes.
Cite mercados ou empresas que poderiam ter interesse em conhecer esta nova tecnologia: Universidades, empresas de sequenciamento de DNA, fabricantes de sequenciadores de DNA.

Anexo C – Esboço do Artigo de Anúncio do Square

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome analysis

Square: a graphical interface genome annotator

Marcus Redü Eslabão[1], Frederico Schmitt Kremer[1] and Odir Antonio Dellagostin[1,*]

[1]Laboratório de Proteômica e Bioinformática, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, Brazil

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The Square is designed to help students and junior researchers in the DNA sequencing to assembly and annotate bacterial genomes. Presenting a graphical and intuitive interface, integrating a database creator and output annotation with color, where indicates the accuracy and relevance of the data recorded.

Results: This software was more accurate in find and annotate CDs when compared to RAST and BASys, both services used by users who **don't** have Linux terminal knowledge or seeking practicality. The accuracy reduces the number of CDs found, reducing the annotation time and post curated annotation. Others relevant factors are a friendly graphical interface and local machine running, this last allows to escape from online queues. Square can be installed on basic computer with little processing power as an Intel Pentium 4 with 1GB of RAM.

Contact: odir@ufpel.edu.br

Supplementary information: Supplementary data are available at Bioinformatics online.

3.1 Introduction

1 The new generation of DNA sequencing has brought advances and new possibilities in the genomics area, increasing the sequencing speed and reducing costs (Demming 2015) currently in the third generation is possible to sequence a bacterial genome in a few hours (Land *et al.* 2015), Observing this, many researchers turned their attention to this tool, both to complement their areas of expertise or to expand their fields of research. Currently companies provide sequencing service and genome assembly (Nucleics 2015), but the final annotation and curation need to be carried out by the researchers, which preferably have a good know the sequenced organism. Some free software perform annotation of genomes with good accuracy and great deal of information (Kisand and Lettieri 2013), but the most of them needs to run on Linux terminal, another part offers online annotation service with the convenience of a graphical interface, but depending on the demand of the servers may result in queues, making it a setback in the life of a researcher.

2 Observing these necessities the Square was created with a intention to providing a user-friendly alternative that run instantly on a personal computer, without requiring a prior knowledge of command line and database download and formatting, also providing the improvement of annotation quality compared with popular annotation software among researchers, even facilitating the visualization of data recorded with a color scheme indicating the accuracy of annotation.

3.2 Methods

3 To develop the Square were used two languages: Python (G. van Rossum and F.L. Drake 2001) and the Lazarus IDE (Lazarus 2015). Python makes its possible, with an ease of development and speed in biological data processing, in addition, enabling the Square run in a command line on remote servers, while Lazarus IDE provides a light and friendly graphical environment, calling the processing core written in Python and adding some functions such as automated download and database creation.

The gene prediction was performed using the Prodigal predictor (Hyatt *et al.* 2010), compared to EasyGene (Larsen and Krogh 2003), Glimmer (Delcher *et al.* 2007) and GenemarkHMM (Besemer and Borodovsky 2005), Prodial proved to be more sensitive and quick to find true CDS. The tRNA prediction was made from tRNAscan (Lowe and Eddy). Soon after the prediction one SQLite3 database (SQLite 2015) is created and the possible CDS are separated, for each one CDs the NCBI BLAST + (Camacho *et al.* 2009) is performed, using the database selected by the user. At the end of the comparison with NCBI BLAST the best results are filtered and selected and a file in GenBank format is created containing the annotation information and a color indicating the accuracy and quality of information of annotation.

3.3 Results

The software of this article has been successfully tested in genomes of Leptospira genus, Corynebacterium and Escherichia, with good speed and accuracy for the final test was selected three complete genomes and annotated NCBI genome database, this selection was used criteria of size, where the smallest genome, the largest genome and a medium size between the two were selected for annotation, being respectively Candidatus Hodgkinia cicadicola TETUND1 strain, Sorangium cellulosum So0157 2 and Burkholderia pseudomallei Pasteur 52237. The computer used to run Square features an Intel i3-2100 processor, 4GB of ram and operation system Ubuntu 14. For comparison we selected two popular software alternatives between those who seek applications with graphical interface, namely, RAST and BASYS. The table containing the comparison can be observed in the Supplementary Data 1.

Watching the genomes tested the Square had the lowest number of false genes and getting closer to the data deposited in NCBI and why run locally was the fastest.

3.4 Conclusions

The Square presents a new alternative for genome annotation among the command line and webservers, with easy double click installation via Debian package. Their graphical interface is user friendly and allows user the option of creating their own databases either from own fasta or download automatically from UniProt Database (EMBL, SIB Swiss Institute of Bioinformatics and Protein Information Resource (PIR) 2013), providing yet refined search for specific genus and species or taxonomy ID, the automated creation of databases, facilitate for the user, speeding up the process of annotation by reduce of database size compared to a more extensive database as TrEMBL. Your light execution can be performed from a notebook or basic desktop with an Intel Core2Duo or higher and 1GB ram processor, and higher the speed of the processor, faster will be the genome annotation process, also has the possibility to be running by command line on servers that have no graphical interface. The Square accuracy is consistent with the others annotators tested yet presenting a lower number of false positives, reducing the work of the curation team.

Funding

This work has been supported by the CAPES and FAPERGS

Conflict of Interest: none declared.

References

Besemer J, Borodovsky M. GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 2005;**33**:451–4.

Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.

Delcher AL, Bratke KA, Powers EC *et al.* Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007;**23**:673–9.

Demming A. DNA sequencing: nanotechnology unravels the code for life. *Nanotechnology* 2015;**26**:310201.

EMBL, SIB Swiss Institute of Bioinformatics, Protein Information Resource (PIR). UniProt. *Nucleic Acids Research*. 2013, 41: D43–7.

G. van Rossum and F.L. Drake. Python Reference Manual. 2001.

Hyatt D, Chen GL, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMCBioinformatics* 2010;**11**:119 – .

Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC Genomics* 2013;**14**:211.

Land M, Hauser L, Jun S-R *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* 2015;**15**:141–61.

Larsen TS, Krogh A. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 2003;**4**:21.

Lazarus. www.lazarus-ide.org. 2015.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Researsh* **25**:955–64.

Nucleics. DNA Sequencing Service Reviews. *www.nucleics.com/DNA_sequencing_support/sequencing-service-reviews.html* 2015.

SQLite. www.sqlite.org. 2015.