Square: Um novo software para anotação automática de genomas microbianos

FREDERICO SCHMITT KREMER[1]; MARCUS REDÜ ESLABÃO[2]; LUCIANO DA[3] SILVA PINTO; ODIR ANTÔNIO DELLAGOSTIN[4]

[1]Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Núcleo de Biotecnologia, Laboratório de Bioinformática – fred.s.kremer@gmail.com
[2]Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Núcleo de Biotecnologia, Laboratório de Bioinformática – marcus.eslabao@yahoo.com.br
[3]Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Núcleo de Biotecnologia, Laboratório de Biotecnologia Vegetal – ls_pinto@hotmail.com
[4]Universidade Federal de Pelotas, Centro de Desenvolvimento Tecnológico, Núcleo de Biotecnologia, Vacinologia Molecular – odirad@terra.com.br

1. INTRODUÇÃO

O sequenciamento genômico é um procedimento de grande relevância para a pesquisa biológica, sendo a sua popularização um dos marcos iniciais da chamada *era pós-genômica* (HOMER *et al*, 2010). O crescimento na demanda por dados genômicos, sobretudo a partir do começo dos anos 2000, trouxe a necessidade de novas abordagens, além de softwares capazes de analisar os dados gerados. Desta forma surgiram os sequenciadores de nova geração, como o SOLiD, Illumina e Roche 454 (ZHANG, 2011), capazes de gerar um grande volume de dados em um espaço de tempo curto e por um custo muito inferior a abordagem tradicional. Estas novas técnicas vêm sendo utilizadas para o sequenciamento de genomas de diversos organismos, o que se reflete no número crescente de entradas em bancos de dados públicos (HOU *et al*, 2013).

Com um número cada vez maior de dados genômicos é necessário fazer a identificação das suas regiões funcionais (Ex: regiões codificantes de proteínas, tRNAs, rRNAs, regiões repetitivas, promotores) em um processo chamado anotação. Diferentes combinações de ferramentas podem ser utilizadas para esta finalidade, o que inclui preditores de genes como GLIMMER (DELCHER *et al*, 1999) e Prodigal (HYATT *et al*, 2010), ferramentas para busca por similaridade como BLAST (ALTSCHUL *et al*, 1990) e HMMER (EDDY, 2011), preditores de RNAs não codificantes como RNAmmer (LAGESEN *et al*, 2007) e trnascan-SE (LOWE *et al*, 1997), dentre outros. Além disso, alguns servidores online também possibilitam a anotação automatizada, como o servidor RAST (AZIZ *et al*, 2008) e o software BLAST2GO (CONESA *et al,* 2005). O uso de ferramentas de uso local apresenta como vantagem um maior controle do processo pelo pesquisador, tirando a necessidade de filas de espera que existe ao se usar ferramentas web, mas por outro lado exige conhecimento de programação para a construção da pipeline.

Desta forma, o presente trabalho teve por objetivo desenvolver um *software* de fácil utilização e de uso local, capaz de realizar a anotação de genomas microbianos através de ferramentas de predição de genes e identificação de proteínas. Esta ferramenta, denominada Square, será distribuída de forma livre para plataformas Windows e, futuramente, para sistemas Linux/Unix.

2. METODOLOGIA

O Square foi desenvolvido em linguagem Python versão 3.2 (www.python.org) e utiliza os programas Prodigal e BLASTx como base para

realizar a anotação dos genomas. Em sua *pipeline*, um arquivo em formato FASTA contendo a sequência a ser anotada é analisada pelo Prodigal que gera um arquivo temporário em formato SCO, onde estão identificadas as *Open Read Frames* (ORFs) e suas respectivas posições. Com base no arquivo SCO gerado o Square utiliza o programa BLASTx para comparar os dados das ORFs com um banco de dados de proteínas para a inferência das suas funções. O resultado do BLASTx é um arquivo XML, que é lido e tem seus dados cruzados com os resultados do Prodigal para a geração de um arquivo em formato GenBank com todos os genes e suas respectivas funções e posições no genoma.

Os bancos de dados de proteínas utilizados durante o processo de anotação são derivados do Uniprot. Em sua primeira versão são disponibilizados os bancos Swissprot (curado e não-redundante) e trEMBL (curado e redundante), que possuem respectivamente 540.958 e 42.821.879 sequências.

O código-fonte do Square foi escrito com auxílio da ferramenta Sublime Text 2, compilado com a biblioteca cx_freeze e compactado pelo software UPX. Um instalador para ser utilizado em computadores que possuem sistema operacional Windows foi desenvolvido com a ferramenta Install Creator.

Para avaliar o funcionamento do Square foi realizada a anotação do cromossomo I da *Leptospira interrogans* L1-130 (NCBI, GI: 45655914) (que possui 3.394 regiões codificadoras de proteínas), com o banco de dados Swissprot a partir do arquivo FASTA derivado do GenBank sendo os dados comparados com a anotação original presente no arquivo GBK. O tempo necessário para o processo de anotação e o número de genes encontrados também foram avaliados. A análise dos resultados foi feita através de scripts escritos em linguagem Python. O processo foi executado em um Notebook Dell Inspiron 14R com processador i5 de quatro núcleos com 2.7 Ghz, 6 Gb de memória RAM e 1 Tb de HD rodando Windows 7.

3. RESULTADOS E DISCUSSÃO

O Square consiste em uma ferramenta executada através de linha de comando (Figura 1), onde os comandos para a execução são definidos através de argumentos. Os argumentos disponibilizados em sua primeira versão e suas respectivas funções estão representados na tabela 1.

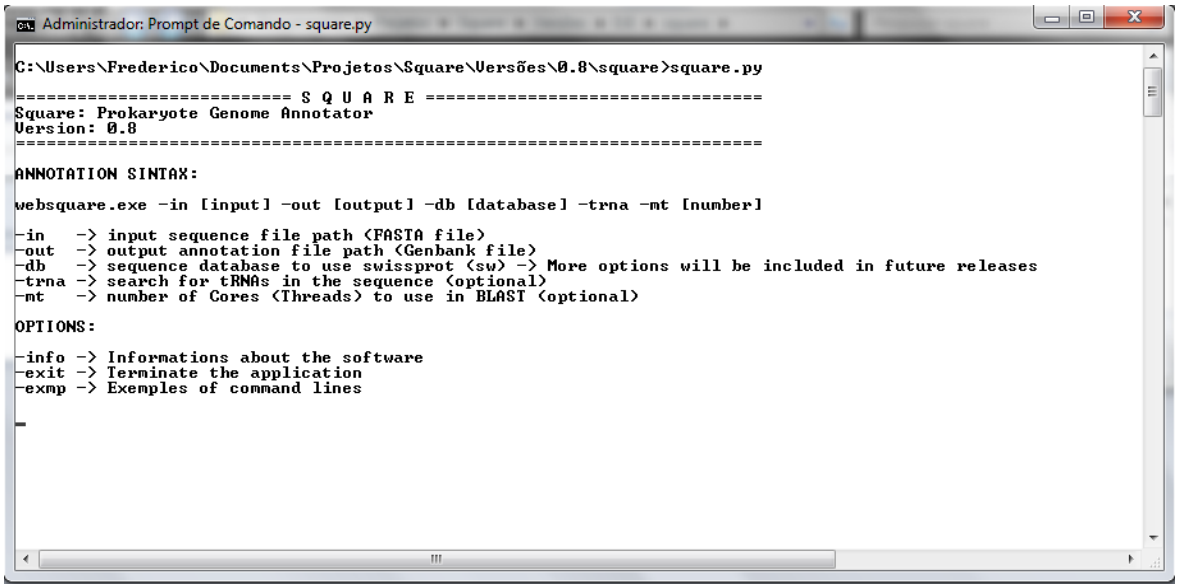

Figura 1. Square sendo executado via *prompt de comandos* em um computador rodando Windows 7.

Tabela 1. Indicação dos argumento permitidos na utilização do Square. *= A função de busca por tRNAs relacionada ao argumento "-trna" ainda está em fase de implementação.

Argumento	Função
-in	Indica o caminho para o arquivo FASTA a ser anotado
-out	Indica o caminho para o arquivo Genbank a ser gerado
-db	Indica o banco de dados a ser utilizado. ('sw' para Swissprot e 'tr' para trEMBL).
-trna*	Argumento opcional para indicar a busca por tRNAs.
-mt	Argumento opcional para indicar o número de núcleos que serão utilizados pelo programa BLAST.

No teste de funcionamento realizado com a sequência do cromossomo I da *L. interrogans* L1-130, o Square, utilizando o banco de dados Uniprot-Swissprot, identificou 1.848 CDSs, o que representa 54,4% dos dados presentes no anotação original. O tempo necessário para a anotação foi de 12 horas, 6 minutos e 56 segundos. O número de CDSs identificadas inferior ao previamente conhecido pode ser explicado pela natureza do banco de dados utilizado, sendo novas analises necessárias para a inferência da capacidade de anotação do Banco de Dados Uniprot-trEMBL.

4. CONCLUSÕES

No presente trabalho foi apresentado o *Square*, um software de bioinformática desenvolvido para realizar a anotação de genomas procariotos de forma automatizada. Apesar de simples, mostrou-se funcional nos testes realizados, sendo capaz de identificar genes codificadores de proteínas em genomas bacterianos. Futuramente, funções adicionais, como predição de tRNAs e rRNAs por HMM, poderão ser adicionadas de forma a aumentar o volume de dados gerados durante a análise. Versões binárias, instaladores e documentações para Windows e Debian/Linux estarão disponíveis para download através do endereço http://200.132.101.131/square/ após a finalização dos testes.

REFERÊNCIAS BIBLIOGRÁFICAS

ALTSHUL, S. GHISH, W. MILLER, W. MYERS, E.W. LIPMAN, D.J. Basic Local alignment search tool**. Journal of Molecular Biology**, V. 215, n. 3, p. 403-10, 1990.

AZIZ, R. K. BARTELS, D. BEST, A. A. DEJONGH, M. TERRENCE, D. 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. PUSCH, G. D. REICH, C. STEVENS, R. VASSEVA, O. VONSTEIN, V. WILKE, A. ZAGNITKO, O. The server: rapid annotation using subsystem technology. **BMC Genomics**, V. 9, n. 75, 2008.

DELCHER, A. L. HARMON, D. KASIF, S. WHITE, O. SALZBERG, S. L. Improved microbial gene identification whith Glimmer. **Nucleic Acids Research**. v. 27, n. 23, p. 4636 – 4641, 1999.

HOMER, D. S. PAVESI, G. CASTRIGNANO, T. MEO, P. D. O. LIUNI, S. SAMMETH, M. PICARDI, E. PESOLE, G. Bioinformatics approaches for Genomics and post Genomics applications of next-generation sequencing. **Briefings in bioinformatics**. v. IL, n. 2, p. 181 – 197, 2009.

HOU, r. YANG, Z. LI, M. XIAO, H. Impact of the next-generation sequencing data depth on various biological result inferences. **Science China**, v. 56, n. 2, p. 104-109, 2013.

HYATT, D. CHEN, G. LOCASDO, P. F. LAND, M. L. LARIMER, F. W. HAUSER, L. J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**, v. 11. 2011.

LAGESEN, K. HALLIN, P. RODLAND, E. A. STAERFELDT, H.H. ROGNES, T. USSERY, D. W. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9. 2007.

LOWE, T. M. EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, V. 25, N. 5, p. 955-964, 1997.

Python Programming Language. Python Software Foundation. Acessado em 10 de Setembro de 2013. Online. Disponível em: http://www.python.org/

ZHANG, J. CHIDINI, R. BADR, A. ZHANG, G. The impact of next-generation sequencing on genomics. **Journal of Genetics and Genomics**, V. 38, p. 95-104, 2011.