Genix: Uma nova pipeline automática para anotação de genomas bacterianos

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

[1] Laboratório de Bioinformática, Universidade Federal de Pelotas – fred.s.kremer@gmail.com
[2] Laboratório de Vacinologia Molecular, Universidade Federal de Pelotas
[4] BioPro Lab. Universidade Federal de Pelotas – ls_pinto@hotmail.com

1. INTRODUÇÃO

O advento do sequenciamento de DNA de nova geração (NGS - *Next Generation Sequencing*), representado pelas plataformas como o Illumina Solexa, ABI SOLiD, Iontorrent PGM e o Roche 454, resultou em uma drástica redução no custo necessário para a obtenção de sequências genômicas, o que possibilitou que diversos centros de pesquisas iniciassem projetos envolvendo a análise genômica de diversos organismos (LIU et al., 2012). Desta forma, o volume de dados de sequências biológicas em bancos públicos, como o Genbank, cresceu exponencialmente (MARDIS, 2008). O processo de anotação consiste na identificação das regiões funcionais (*features*) presentes em um determinado conjunto de sequências (KISAND; LETTIERI, 2013). No caso de genomas bacterianos, este tipo de análise é normalmente realizado sistemas automatizados, denominados *pipelines*, que integram softwares para a identificação de regiões de DNA codificantes (CDS), genes para RNAs não codificantes, regiões regulatórias, dentre outras (BECKLOFF et al., 2012).

As pipelines para anotação podem ser de uso local ou *web*. No caso das *pipelines* de uso local, os softwares são executados no computador do usuário e é possível um maior controle dos dados gerados, apesar de ser necessário um maior conhecimento técnico, como implementado nas ferramentas Prokka (SEEMANN, 2014) e Eugene-PP (SALLET; GOUZY; SCHIEX, 2014). Já no caso das pipelines web, como o RAST (AZIZ et al., 2008), BASys (VAN DOMSELAAR et al., 2005) e xBASE (CHAUDHURI; PALLEN, 2006), os *softwares* tendem a ser de utilização mais facilitada, mas oferecem um menor número de ferramentas de análise e filtragem o que pode acarretar uma perda de dados relevantes e um maior número de dados falso positivos.

A eficiência de uma *pipeline* de anotação pode variar de acordo com as ferramentas que a integram, assim como com os bancos de dados que são usados como base. No presente trabalho é apresentado o *software* Genix, uma nova plataforma *web* para anotação automática de genomas bacterianos.

2. METODOLOGIA

O Genix recebe como entrada um arquivo em formato FASTA, contendo as sequências do genoma, e um tax_id, um código de identificação taxonômica padronizado e utilizado por diferentes bancos de dados biológicos, como Genbank e o Uniprot. A partir do tax_id, uma lista de sequências de proteínas é obtida a partir do Uniprot e usada para criar um banco de dados bruto (“raw database”), que é processado pelo *software* CD-HIT (LI; GODZIK, 2006) para a remoção das redundâncias, gerando-se um banco de dados final, que por sua vez

é formato pelo software makeblastdb para a criação de um banco de dados para os softwares do pacote BLAST (ALTSCHUL et al., 1990).

Para a anotação de regiões, o *software* Prodigal é utilizado como *gene finder*, sendo seus resultados comparados pelo BLASTp do pacote NCBI-BLAST+ (CAMACHO et al., 2009) utilizando-se o banco de dados gerado no passo anterior. No caso de proteínas que apresentam correspondentes, uma análise pelo HMMER (EDDY, 2011) utilizando o banco de dados do AntiFam (EBERHARDT et al., 2012) é realizada, de forma a se remover possíveis proteínas falso positivas (*spurious ORFs*) utilizando-se um *e-value* mínimo de 0.0005. Para os RNAs não codificantes (ncRNAs), as ferramentas tRNAscan-SE (LOWE; EDDY, 1997), RNAmmer (LAGESEN et al., 2007) e Aragorn (LASLETT, 2004) são usadas para a predição de genes de tRNAs, rRNAs e tmRNAs, respectivamente. Além destas, uma busca por outros grupos de ncRNAs é realizada pela ferramenta BLASTn utilizando-se o banco de dados de famílias de RNAs Rfam (GRIFFITHS-JONES et al., 2003), sendo as família encontradas posteriormente realinhadas contra o genoma com o *software* INFERNAL (NAWROCKI; KOLBE; EDDY, 2009). No final da anotação, o Genix integra os resultados dos diferentes programas e corrige regiões conflitantes, como genes sobrepostos e quebrados nas extremidades 5’ e 3’. A pipeline foi implementada em linguagem Python e utilizada as bibliotecas BioPython e MySQLdb, assim como alguns scripts escritos em Perl e Bash. Para o gerenciamento dos dados, MySQL é usado como sistema de banco de dados principal, armazenando informações das requisições e usuários, e um pequeno banco de dados SQLite é usado para os dados intermediários da anotação. O sistema roda em um servidor Dell PowerEdge de 8 núcleos com 24 Gb de RAM com sistema operacional Ubuntu e usa o Apache HTTP como servidor *web*.

Para a avaliação da pipeline, foi realizada a re-anotação do cromossomo I da *Leptospira interrogans* sorovar Copenhageni cepa L1-130 utilizando nossa pipeline e os softwares RAST (AZIZ et al., 2008), xBASE (CHAUDHURI; PALLEN, 2006) e BASys (VAN DOMSELAAR et al., 2005). As anotações foram comparadas entre si e com o genoma de referência para a identificação de genes faltantes (ausentes em relação à referência), novos (não presentes na referência) e exclusivos (presentes em apenas uma das anotações). Com base nos genes que apresentaram diferenças, anotações funcionais com a ferramenta BLAST2GO (CONESA et al., 2005) foram realizadas para aferir a o impacto da ausência de cada gene na anotação do genoma. Para fins de comparação, cada anotação também foi compara com o banco de dados do AntiFam (EBERHARDT et al., 2012) com a ferramenta HMMER (EDDY, 2011) para a identificação de *spurious ORFs*. Por fim, um cálculo de discrepância entre as re-anotações e o genoma de referência foi realizada se dividindo a soma de genes faltantes e novos pela soma de genes no genoma de referência e na re-anotação.

3. RESULTADOS E DISCUSSÃO

A ferramenta Genix está disponível na forma se um *webserver*, sendo constituído por um *front-end* desenvolvido em HTML/Javascript/CSS e um *back-end* constituído por scripts em Python, Perl e Bash que se comunicam com bancos de dados MySQL e SQLite para gerenciamento e anotação dos genomas submetidos. A utilização da ferramenta é livre, mas um cadastro é solicitado de

forma a prover um maior controle no acesso aos resultados de cada submissão. Além das sequências a serem anotadas e do *tax_id*, o usuário também pode controlar parâmetros como o grau de identidade usado pelo CD-HIT, o tipo de informação usado para a ligação das *scaffolds*, dados do organismo (Ex: cepa, sorovar, sorotipo), assim como enviar arquivos de *template* de submissão para o Genbank. As submissões são alocadas em uma fila de espera e processadas uma de cada vez. No final, arquivos *Genbank* (.gb) e *Feature Table* (.tbl) são gerados. Caso um template de submissão seja provido durante a submissão, um arquivo *Sequin* (.sqn) é gerado e disponibilizado em uma pasta compactado junto aos relatórios do software tbl2asn.

Na comparação das pipelines, BASys apresentou 3222 CDS (436 faltantes, 946 novos, 258 exclusivos e 151 *spurious ORFs*), RAST 4352 CDS (349 faltantes, 1044 novos, 470 exclusivos e 292 *spurious ORFs*), xBASE 4078 CDS (449 faltantes, 869 novos, 161 exclusivos e 131 *spurious ORFs*) e o Genix 3183 (475 faltantes, 155 novos, 20 exclusivos, e nenhuma *spurious ORFs*). Na análise funcional, BASys apresentou 123 faltantes, 34 novos e 5 exclusivos, RAST apresentou 121 faltantes, 31 novos e nenhum exclusivo, xBASE apresentou 126 faltantes, 33 novos e 2 exclusivos, e Genix apresentou 115 faltantes, 29 novos e 0 exclusivos. Na análise de discrepância, as ferramentas BASys, RAST, xBASE e Genix apresentaram 20.09%, 17.39%, 17.04% e 9.21%, respectivamente.

4. CONCLUSÕES

A pipeline Genix para anotação automática de genomas bacterianos se mostrou uma ferramenta capaz de gerar anotações com grande acurácia, mais próxima em relação ao genoma de referência usado para testes (menor discrepância), com menor taxo de spurious ORFs e menor número de genes funcionais faltantes. No presente trabalho também foi apresentada uma abordagem para comparação de anotações, denominada *discrepância de anotação*. O Genix está disponível para uso através do endereço http://labbioinfo.ufpel.edu.br/genix.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 5 out. 1990.

AZIZ, R. K. et al. The RAST Server: rapid annotations using subsystems technology. **BMC genomics**, v. 9, p. 75, jan. 2008.

BECKLOFF, N. et al. Bacterial genome annotation. **Methods in molecular biology (Clifton, N.J.)**, v. 881, p. 471–503, jan. 2012.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC bioinformatics**, v. 10, n. 1, p. 421, jan. 2009.

CHAUDHURI, R. R.; PALLEN, M. J. xBASE, a collection of online databases for bacterial comparative genomics. **Nucleic acids research**, v. 34, n. Database issue, p. D335–7, 1 jan. 2006.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics (Oxford, England)**, v. 21, n. 18, p. 3674–6, 15 set. 2005.

EBERHARDT, R. Y. et al. AntiFam: a tool to help identify spurious ORFs in protein annotation. **Database : the journal of biological databases and curation**, v. 2012, p. bas003, jan. 2012.

EDDY, S. R. Accelerated Profile HMM Searches. **PLoS computational biology**, v. 7, n. 10, p. e1002195, out. 2011.

GRIFFITHS-JONES, S. et al. Rfam: an RNA family database. **Nucleic acids research**, v. 31, n. 1, p. 439–41, 1 jan. 2003.

KISAND, V.; LETTIERI, T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. **BMC genomics**, v. 14, p. 211, jan. 2013.

LAGESEN, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic acids research**, v. 35, n. 9, p. 3100–8, jan. 2007.

LASLETT, D. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. **Nucleic Acids Research**, v. 32, n. 1, p. 11–16, 2 jan. 2004.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics (Oxford, England)**, v. 22, n. 13, p. 1658–9, 1 jul. 2006.

LIU, L. et al. Comparison of next-generation sequencing systems. **Journal of biomedicine & biotechnology**, v. 2012, p. 251364, jan. 2012.

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. 955–64, 1 mar. 1997.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual review of genomics and human genetics**, v. 9, p. 387–402, jan. 2008.

NAWROCKI, E. P.; KOLBE, D. L.; EDDY, S. R. Infernal 1.0: inference of RNA alignments. **Bioinformatics (Oxford, England)**, v. 25, n. 10, p. 1335–7, 15 maio 2009.

SALLET, E.; GOUZY, J.; SCHIEX, T. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. **Bioinformatics (Oxford, England)**, v. 30, n. 18, p. 2659–61, 15 set. 2014.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics (Oxford, England)**, v. 30, n. 14, p. 2068–9, 15 jul. 2014.

VAN DOMSELAAR, G. H. et al. BASys: a web server for automated bacterial genome annotation. **Nucleic acids research**, v. 33, n. Web Server issue, p. W455–9, 1 jul. 2005.