Sequence analysis

**ACTG: novel peptide mapping onto gene models**

Seunghyuk Choi¹, Hyunwoo Kim² and Eunok Paek¹,*

¹Department of Computer Science, Hanyang University, Seongdong-gu, Seoul, Korea and ²Scientific Data Technology Lab, Korea Institute of Science and Technology Information, Yuseong-gu, Daejeon, 34141, Korea

*To whom correspondence should be addressed.

Received on September 4, 2016; revised on December 4, 2016; editorial decision on December 5, 2016; accepted on December 7, 2016

Abstract

Summary: In many proteogenomic applications, mapping peptide sequences onto genome sequences can be very useful, because it allows us to understand origins of the gene products. Existing software tools either take the genomic position of a peptide start site as an input or assume that the peptide sequence exactly matches the coding sequence of a given gene model. In case of novel peptides resulting from genomic variations, especially structural variations such as alternative splicing, these existing tools cannot be directly applied unless users supply information about the variant, either its genomic position or its transcription model. Mapping potentially novel peptides to genome sequences, while allowing certain genomic variations, requires introducing novel gene models when aligning peptide sequences to gene structures. We have developed a new tool called ACTG (Amino aCids To Genome), which maps peptides to genome, assuming all possible single exon skipping, junction variation allowing three edit distances from the original splice sites, exon extension and frame shift. In addition, it can also consider SNVs (single nucleotide variations) during mapping phase if a user provides the VCF (variant call format) file as an input.

Availability and Implementation: Available at http://prix.hanyang.ac.kr/ACTG/search.jsp.

Contact: eunokpaek@hanyang.ac.kr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Proteogenomics is a newly developing field of research, situated at the interface between genomics and proteomics. In many proteogenomic applications, peptide sequences identified from proteomics are used to propose a novel annotation of a gene. This requires mapping peptides onto reference genome sequences so that the genomic loci of peptide sequences can be determined.

There are many software tools such as TransVar (Zhou et al., 2015) and PGx (Askenazi et al., 2016) that can be used to map peptides onto genome. However, these tools cannot map ‘novel’ peptides that result from structural variations such as exon skipping, splice junction variation, frame shift and intron retention, unless such novel transcript models are given as a user input. We present a new software tool called ACTG (Amino aCids To Genome, http://prix.hanyang.ac.kr/ACTG/search.jsp), that maps novel peptide sequences to a reference genome/transcriptome model, with emphasis on peptides resulting from structural variations. Unlike other existing approaches, our method does not take genome coordinates for peptide mapping as a user input. Potential splice variants predicted from matching RNA-seq results (Gascoigne et al., 2012) need not be provided, either.

ACTG provides a user friendly web interface. Users simply upload a peptide list (up to 100 KB), and an optional VCF file (up to 20 MB) if one wants to apply sample-specific point mutations to the splice graph before mapping peptides. One can select a mapping method, a reference database to be used for filtering out known sequences, and what variations they consider (Supplementary Note 1). The output of the mapping can be downloaded from the web site.
ACTG consists of two phases (Fig. 1): a variant splice graph construction phase and a peptide mapping phase. The variant splice graph is built from reference genome/transcriptome databases, thus can be pre-compiled.

During the construction phase (Supplementary Note 1), ACTG takes reference genome sequences in FASTA file format and reference transcriptome models in GTF. Each transcript model is represented as a splice graph (Woo et al., 2014). It is a direct acyclic graph where nodes represent nucleotide sequences for exons and edges connect neighboring exons. They are merged into a single graph if they belong to the same gene. After merging, the splice graph is modified to incorporate four novel genomic events that we postulate: single exon skipping (any single exon may be spliced out while forming mature mRNA); splice junction variation (three edit distance shifts from original splice sites if there is a consensus sequence of ‘GT’ at a splice donor site or ‘AG’ at a splice acceptor site); frame shift; and exon-extension (translation is extended into intron regions at exon-intron junction).

During the peptide mapping phase, ACTG takes two inputs: the serialized file of the variant splice graph and peptide sequences. An optional VCF file can be given as an input if a user wants to consider SNVs (Supplementary Note 2). For sequence matching, Aho-Corasick algorithm is applied for fast multiple pattern matching. In case of multiple occurrences of a single peptide, we report a list of genomic coordinates together with annotations (Supplementary Note 3).

We took 1,577 novel peptides that had been previously reported (Kim et al., 2014; Park et al., 2014; Sheynkman et al., 2013) as resulting from novel coding regions (106), pseudogenes (183), novel n-termini (198), exon-extension (60), exon skipping (12), splice junction variation (22) and SNV (996). A variant splice graph built from GRCh37 human reference genome and Ensembl v75 transcriptome models was used. ACTG automatically mapped 1564 (99.2%) peptides successfully and missed only 13 peptides (Supplementary Note 4).

We compared ACTG with PGx, the most recent and powerful tool. Only three types of aforementioned novel peptides were used because PGx cannot handle the rest. ACTG used GRCh37 human reference genome and RefSeq transcriptome model from UCSC Table Browser (Karolchik et al., 2004) because PGx uses RefSeq.

PGx web application (pgx.fenyolab.org) was applied. Table 1 shows the number of peptides with the same annotations as the previous report (Supplementary Note 5).

PGx was designed to map peptides against potential novel transcript model inferred from RNA-seq results. Thus, when applied against RefSeq database, it showed poor performance. In multiple mapping cases, PGx returned different genomic loci from the previous annotation (Supplementary Note 5).

Proteogenomic approaches have been applied to annotate genome information of MS/MS-certified peptides. In such applications, ACTG can be a very useful tool for mapping novel peptides to genome, which will give an insight of a potential novel gene model.

This work was supported by the National Research Foundation of Korea [NRF-2012M3A9B9036676, NRF-2014R1A2A1A11054147, NRF-2012M3A9D1054452] and the Brain Korea 21 Plus Project.

Conflict of Interest: none declared.

PGx was designed to map peptides against potential novel transcript model inferred from RNA-seq results. Thus, when applied against RefSeq database, it showed poor performance. In multiple mapping cases, PGx returned different genomic loci from the previous annotation (Supplementary Note 5).

References


