Bioinformatics, 31(15), 2015, 2571–2573
doi: 10.1093/bioinformatics/btv141
Advance Access Publication Date: 26 March 2015
Applications Note

Sequence analysis

MSA-PAD: DNA multiple sequence alignment framework based on PFAM accessed domain information

Bachir Balech[1], Saverio Vicario[2], Giacinto Donvito[3], Alfonso Monaco[3], Pasquale Notarangelo[3] and Graziano Pesole[1,4,*]

[1]Istituto di Biomembrane e Bioenergetica and [2]Istituto Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, 70126 Bari, Italy, [3]Istituto Nazionale di Fisica Nucleare, 70126 Bari, Italy and [4]Dipartimento Bioscienze, Biotecnologie e Biofarmaceutica, University of Bari, 70126 Bari, Italy

*To whom correspondence should be addressed.
Associate Editor: David Posada

Received on August 8, 2014; revised on March 6, 2015; accepted on March 9, 2015

Abstract

Summary: Here we present the MSA-PAD application, a DNA multiple sequence alignment framework that uses PFAM protein domain information to align DNA sequences encoding either single or multiple protein domains. MSA-PAD has two alignment options: gene and genome mode.
Availability and Implementation: MSA-PAD is available as a web application (https://recasgateway.ba.infn.it/) and as two Taverna workflows corresponding to two alignment modes (Gene mode: http://www.myexperiment.org/workflows/4549.html; Genome Mode: http://www.myexperiment.org/workflows/4551.html).
Contact: g.pesole@ibbe.cnr.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Multiple sequence alignment (MSA) is one of the oldest problems in bioinformatics and represents a key step in sequence analysis applications such as phylogenetic inference and comparative genomics. Algorithms such as *translatorX* (Abascal *et al.*, 2010) and *tranalign* (EMBOSS; Rice *et al.*, 2000), taking advantage of the higher conservation of proteins than their respective coding sequences, use the protein alignment to guide the reconstruction of accurate MSAs at nucleotide level. However, these methods do not make use of information embedded in protein domains. Moreover, when applied to genomic sequences, such approaches account neither for intron occurrence nor for gene order variations (e.g. mitochondrial genomes; D'Onorio de Meo *et al.*, 2012) resulting in apparently truncated protein domains or variations in their arrangement along the genome regions under investigation.

Here we present a DNA MSA framework (MSA-PAD), which conceptually translates DNA sequences into amino acids (based on user-defined genetic code and reading frame/s), uses information from conserved PFAM domains (Finn *et al.*, 2014) to assign the translated sequences to known protein domains, accounts for frameshifts when domain regions are split by introns, performs a domain-based protein alignment and then uses protein alignment information to generate the relevant nucleotide multiple alignment. The final MSA can be generated following two different strategies: (i) Gene and (ii) Genome mode. Gene mode alignment respects domain order organization from 5′ to 3′, and resolves the alignment of repetitive domains even when they are repeated in tandem. Genome mode alignment provides a super-gene-like alignment ignoring domain order constraints.

2 Methods

2.2 Implementation: a simple wrapper

MSA-PAD is available as a service on a RESTfull endpoint of Job Submission Tool (JST; Donvito *et al.*, 2012) Web service. To upload/download input and output files, the JST framework provides a

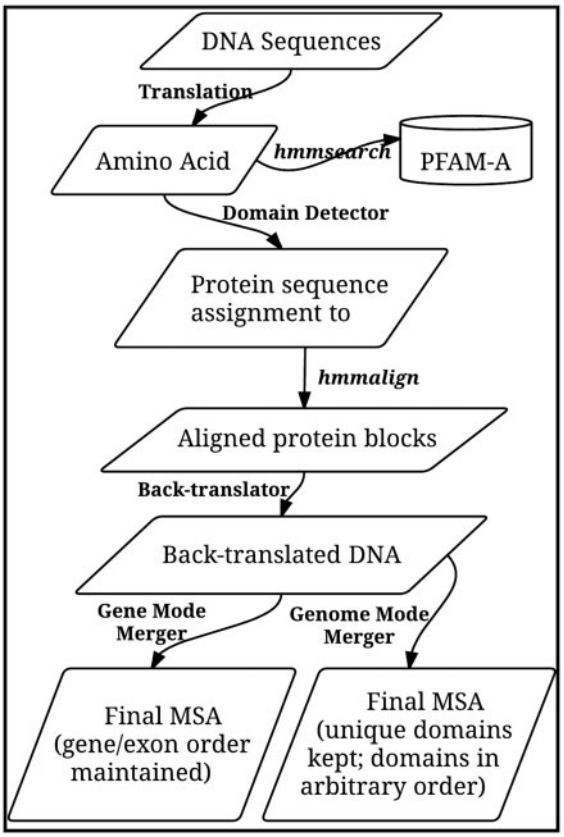

Fig. 1. MSA-PAD at work overview

WebDav endpoint. In this way, it is possible to use a Uniform Resource Identifier (URI) during the execution of the application. MSA-PAD takes as input a multiple DNA sequence data set (FASTA format) containing one (gene mode) or several genes (genome mode) coding for single or multiple protein domains and potentially including non-coding intervening regions. The Web service includes wrappers for HMMer3.0 (Eddy, 2011) and a series of python scripts to compute the analysis (Fig. 1). In detail, MSA-PAD first translates DNA sequences following a user-defined genetic code and one or more reading frame/s. All peptide fragments generated by the user-defined frame translation (i.e. all reading frames split by stop codons) are used as probes to search a local mirror of the PFAM-A conserved domain database using *hmmsearch* and then detected protein domain blocks are aligned using the *hmmalign* algorithms of HMMer3.0 (Eddy, 2011). Where frameshifts or a non-coding intervening sequences are present, the domain alignments are merged by a custom parser script that concatenates alignment chunks detected in multiple frames. In this way, multigenic nucleotide sequences (i.e. complete mitochondrial genomes) can be also submitted given that the correct reading frame for each coding region is automatically detected by domain assignment step. Later on, protein domain alignments are used to guide the relevant alignments at nucleotide level, a process we denote in the following as back-alignment. The back-alignments are finally merged according to the experimental mode specified. In Gene Mode, the most frequent domain organization pattern from 5′ to 3′ is detected and all sequences that are compatible with this domain order, are aligned, while others are eliminated. In this way, the back-aligned domains are merged according to the order of current domain organization. In Genome Mode, only unique protein domains are considered (avoiding domain paralogy conflict across repetitive domains) and the back-aligned domains are concatenated in arbitrary order. Together with the final MSA, the position of each domain in the alignment is given in a separate file. Additional outputs corresponding to the protein and the back-aligned blocks are also provided (complete information regarding the output data are provided in supplementary materials).

3 Results

We evaluated the ability of MSA-PAD to assign sequence fragments to the correct domain block. The alignment within blocks was not assessed given that it is entirely determined by the *hmmalign* algorithm (Eddy, 2011). The case studies consisted of mitochondrial genomes exhibiting either conserved or varied gene order, genes containing introns, sequences encoding repeated domains, genes demonstrating intron loss/gain and ORFs encoding single protein domains. We evaluated MSA-PAD for protein sequence identity compared with the reference, protein sequence assignment to the correct PFAM domain, the order of domain blocks merged—with particular reference to repeated domains (in Gene Mode) and the uniqueness of these domains (in Genome Mode). Our results show that MSA-PAD correctly translated DNA sequences, assigned them to the right PFAM domains (no false positive assignments) and assembled them as expected in both Gene and Genome modes (Supplementary Table S1). The unique feature of MSA-PAD compared to other methods (e.g. TranslatorX, tranalign) is that it is able to automatically generate an accurate MSA even where frameshifts or introns are present in the input sequences. Furthermore, it is able to generate a super-gene-like alignment when input sequences contain multiple genes (e.g. complete mitochondrial genome sequences).

However, MSA-PAD may have some limitations when no PFAM domain match is found for the input sequences or when a matched domain fragment is too short (e.g. in the case of a short exon) to produce a significant alignment score.

4 Conclusions

We have created a new DNA MSA framework, MSA-PAD, which uses PFAM-A profiles to generate DNA multiple alignments from sequences encoding either single or multiple (including repeated) protein domain/s. MSA-PAD will be useful for comparative genomics applications as it can account for genomic rearrangements and provide a ‘super-gene’ final alignment. In Gene Mode either complete homologous gene loci (including introns) and mRNAs (including UTRs) can be used as input as MSA-PAD automatically reconstructs the domain-based alignment. We are currently preparing new options allowing user defined HMM profile/s to replace or to be added to PFAM-A database. This will increase performance in terms of quality and/or speed.

Funding

This work was supported by Biodiversity Virtual eLaboratory (BioVeL) funded by the European Commission 7th Framework Programme (FP7), through the grant agreement number 283359, and Lifewatch-Italy. We thank David S. Horner for critical reading the manuscript and checking the English style.

Conflict of Interest: none declared.

References

Abascal,F.*et al.* (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7–W13.

D'Onorio de Meo,P. *et al*. (2012) MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. *Nucleic Acids Res.*, **40**, D1168–D1172.

Donvito,G. *et al.* (2012) The BioVeL Project: robust phylogenetic workflows running on the GRID. *Proceedings of the EGI Community Forum 2012/ EMI Second Technical Conference (EGICF12-EMITC2). 26-30 March, 2012*. Munich, Germany.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Finn,R.D. *et al*. (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.