BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 22 2014, pages 3266–3267
doi:10.1093/bioinformatics/btu544

Genome analysis

Advance Access publication August 12, 2014

Figmop: a profile HMM to identify genes and bypass troublesome gene models in draft genomes

David M. Curran[1,2,*], John S. Gilleard[2] and James D. Wasmuth[1]

[1]Department of Ecosystem and Public Health and [2]Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, T2N 4Z6, Canada

Associate Editor: John Hancock

ABSTRACT

Motivation: Gene models from draft genome assemblies of metazoan species are often incorrect, missing exons or entire genes, particularly for large gene families. Consequently, labour-intensive manual curation is often necessary. We present Figmop (Finding Genes using Motif Patterns) to help with the manual curation of gene families in draft genome assemblies. The program uses a pattern of short sequence motifs to identify putative genes directly from the genome sequence. Using a large gene family as a test case, Figmop was found to be more sensitive and specific than a BLAST-based approach. The visualization used allows the validation of potential genes to be carried out quickly and easily, saving hours if not days from an analysis.

Availability and implementation: Source code of Figmop is freely available for download at https://github.com/dave-the-scientist, implemented in C and Python and is supported on Linux, Unix and MacOSX.

Contact: curran.dave.m@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2014; revised on July 25, 2014; accepted on August 7, 2014

1 INTRODUCTION

The majority of published metazoan genome assemblies are in draft form, representing an *in silico* prediction of the *in vivo* genome. Errors include sequence reads that are not clustered into contigs, and contigs that are misplaced on scaffolds, either in the wrong location or orientation. This can have a significant impact on finding the genes in the genome, an already complex process in eukaryotes that may have tiny exons and introns of variable length (Picardi and Pesole, 2010). The best algorithms get many of the gene models mostly correct, but this may not be enough to make and test evolutionary and functional hypotheses in many cases, particularly for large and complex gene families.

Here, we present Figmop (Finding Genes using Motif Patterns), which will guide the user to identify the correct gene models, or provide a measure of the accuracy of the current gene models of their gene family. The software extends the use of MEME and MAST to identify regions of the genome containing genes of interest (Bailey *et al.*, 2006). It uses amino acid motifs to capture conservation within short stretches of sequence in the user's gene family, giving an architecture for that family as a specific pattern of those motifs. For highly variable regions, alternative or optional motifs can be included. Figmop implements a profile hidden Markov model (pHMM) that conducts a fuzzy match of this motif architecture against the given genome sequence, accounting for variation and introns as random or unmatched motifs.

Figmop has proved invaluable in our efforts to manually curate cytochrome P450 (CYP) gene family members, which have high sequence variability and differing intron/exon structure, within a draft genome assembly of the parasitic nematode *Haemonchus contortus* (Laing *et al.*, 2013). Figmop is a general tool, and besides CYPs, it has also proven useful in identifying other diverse gene families, including glutathione-S-transferase, UDP-glucuronosyl transferase and ABC-transporters.

2 IMPLEMENTATION

2.1 Motif generation and detection

A set of protein sequences of the user's gene family is the starting point. These sequences should be from related species and/or confirmed full-length proteins from the test species. The user runs the MEME software to generate a set of motifs and specifies a pattern from these. Figmop is then run, where it first uses the associated MAST software to detect these motifs across the test genome sequence (Bailey *et al.*, 2006). This is the most computationally intensive step of Figmop, taking ~8 min to process a 370 Mb genome on a 2.6 GHz computer, but must be run only once per genome per full set of motifs.

2.2 Figmop pHMM

The Figmop configuration file is automatically generated and contains the motif pattern and a set of configurable probabilities that constitute a pHMM (Supplementary Fig. S1). The software scans the motif complement of the test genome from the previous step and uses the Viterbi algorithm to detect significant architectures. This step is coded in C, and takes ~9 s to run on a 370 Mb genome. Different architectures or probabilities using the same set of motifs can be searched in the same amount of time without the need to run MAST on the genome again.

2.3 Output

The genomic regions containing the detected architectures are collected, and the sequences are extracted and saved in FASTA

*To whom correspondence should be addressed.

format. MAST is then run on these sequences to produce an HTML output with which the significant architectures can be visualized.

3 RESULTS

3.1 Validation

Our original motivation for developing Figmop was to find the CYP genes in nematode genome assemblies. A total of 530 nematode CYP genes from *Caenorhabditis* species and *Pristionchus pacificus* were collected and an architecture of 13 MEME-derived motifs was generated (Fig. 1A). This pattern of motifs was then used by Figmop to search the genome assembly of *Drosophila melanogaster*, where it returned all of the 85 defined CYP genes, as well as two probable pseudogenes (results not shown).

We then used Figmop to search the current genome assembly of the parasitic nematode *H.contortus* (project ID PRJEB506, version CAVP000000000.1; Laing *et al*., 2013) and found 28 regions (e.g. Fig. 1B). However, all of the published gene model predictions for these likely CYP genes were truncated or fragmented (e.g. Fig. 1C). Further inspection revealed that all of the expected sequence motifs for this CYP were present in the genome (Fig. 1B), but were not identified as coding sequence by the gene prediction software. This was also found to be the case for the rest of the 28 putative CYP genes.

We do not discount the possibility that the disparity between Figmop and the *H.contortus* gene models may be because of Figmop detecting pseudogenes, but this is unlikely to be the case for all 28 regions. We have also observed a similar disparity in the genomes of the parasitic nematodes *Ascaris suum*, *Brugia malayi* and *Strongyloides ratti* (data not shown).

3.2 Comparison with BLAST

A common approach to find a gene of interest in a genome in the absence of gene models is to compile a set of known homologous proteins, and run TBLASTN (version 2.2.27) against the genomic sequence, accepting those high-scoring segment pairs (HSPs) that satisfy an *E*-value threshold (Camacho *et al*., 2009). When searching a whole genome a small *E*-value should be used to avoid false positives; 1E-40 might be suitable in this case. At this threshold, only 4 of the 28 regions found by Figmop were matched by one or more HSPs, but there were an additional seven HSPs that, following manual inspection, were probably false positives, as the genomic regions contained no evidence of the CYP pattern described previously. The *E*-value threshold had to be relaxed to 1E-10 before all 28 regions were found by at least one HSP, at which point, there were an additional 432 HSPs that were likely false positives. A reciprocal BLAST approach was used to improve the specificity, as described in the Supplementary Information. Using this approach, TBLASTN still required an *E*-value threshold of 1E-10 before it could detect all 28 putative CYP regions, though it still returned eight false positives. Even at this very relaxed threshold, many of the CYP regions were matched only by a single HSP (Supplementary Fig. S2). A threshold of 1E-25 was required to exclude all false positives, though at this point only 11 of the 28 putative CYP regions were returned. One *H.contortus* scaffold was shown to contain five putative CYPs by Figmop. A detailed investigation of the TBLASTN results on this scaffold is described in the Supplementary Information.

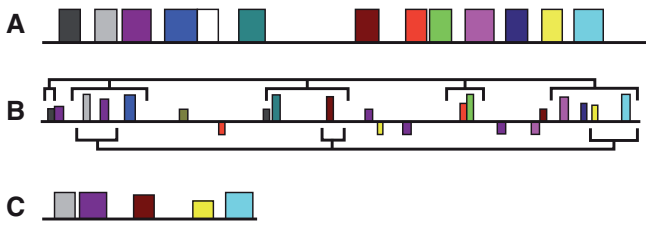

Fig. 1. Figmop in action: improving a CYP gene model. The coloured bars represent sequence motifs identified by MAST. (**a**) shows the pattern of motifs common to all CYP genes, while (**c**) shows one of the fragmented gene models identified by AUGUSTUS. (**b**) shows the corresponding genomic region from (c), where the bottom indicators map the previously predicted coding regions and the top indicators map the CYP pattern from (a) as found by Figmop

4 CONCLUSION

We have created Figmop, a software that uses a pHMM to compare the motif patterns of a gene family against a test genome. We have used it for the manual curation of several large divergent gene families and for these cases, found it to be preferable to BLAST in sensitivity and specificity, but mostly in terms of ease of use and time.

ACKNOWLEDGEMENTS

The authors thank Aude Gilabert for breaking earlier versions of the program.

Funding: This work is supported by NSERC CREATE (Natural Sciences and Engineering Research Council of Canada Collaborative Research and Training Experience) programme in Host–Parasite Interactions (#413888-2012).

Conflict of interest: none declared.

REFERENCES

Bailey,T.L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Laing,R. *et al.* (2013) The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. *Genome Biol.*, **14**, R88.

Picardi,E. and Pesole,G. (2010) Computational methods for *ab initio* and comparative gene finding. *Methods Mol. Biol.*, **609**, 269–284.