APPLICATIONS NOTE

BioMoby Web services to support promoter analysis protocols

Arnaud Kerhornou[*,1,2] and Roderic Guigó[1,2]

[1]Centre de Regulacio Genomica, Pg. Maritim de la Barceloneta, 08003 Barcelona, Catalonia, Spain.

[2]Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Pg. Maritim de la Barceloneta, 08003 Barcelona, Catalonia, Spain.

ABSTRACT

Summary: Here we present a computational protocol to analyse the promoter regions of a given set of co-expressed genes, and its implementation through the use of Web services technologies. This protocol aims to cluster a set of co-regulated genes in subsets of genes showing similar configurations of Transcription Factor Binding Sites. All the steps of this protocol have been developed as Web services that are compliant with BioMoby specifications.

Availability: {{http://genome.imim.es/cgi-bin/moby/GeneClustering_DataSubmission.cgi}}

Contact: arnaud.kerhornou@crg.es, roderic.guigo@crg.es

Supplementary information: {{http://genome.imim.es/webservices/index.html}}

1 INTRODUCTION

With the completion of many sequencing projects, there are tremendous amount of data and, coming along, of analysis methods that are being made available through the Web to the scientific community. While these resources are of great help for retrieval of data, and quick hypothesis verification, they are mostly used through manual execution and can not be applied for automated tasks. This implies some drawbacks, such as slowness and being error-prone when executed repeatedly.

In silico experiments, on the other hands, are described in protocols that can be seen as an orchestrated execution of atomic steps. Such computational protocols are commonly implemented using a script language such as Perl. The various steps may be executed on local resources, but increasingly often using remote resources.

In this regard, Web services architecture (Web Services Architecture specifications document, {{http://www.w3.org/TR/ws-arch/}}) have emerged to provide programmatic access to remote resources, thus allowing users to perform *in silico* experiments through the Web in an automatic manner (Stevens *et al.*, 2004).

We have applied such technology to develop a pipeline for the characterization of the promoter regions of co-regulated genes. It is generally assumed that genes with similar transcriptional regulatory programs also exhibit similar configurations of Transcription Factor (TF) Binding Sites (TFBS) in their promoter regions upstream of the Transcription Start Site (TSS) (Wray *et al.*, 2003). Because TFBSs are short DNA motifs (8 to 15bp in range), they can occur by chance very often in DNA sequences, thus producing a high level of false positives. To differentiate false positive predictions from truly functional elements, new methods have been proposed (for a review, see Wasserman *et al.*, 2004). In addition, promoter elements bound by the same TF may not show sequence similarity and, therefore, sequence comparisons between promoter elements of co-expressed genes often fail to reveal the underlying common regulatory domains. To address this limitation, Blanco *et al.* (2005) introduced TF-map alignments. In these, TFBS on promoter sequences are labeled according to the corresponding TF, and the comparison is performed between the sequence of labels. TF-map alignments have been shown to uncover common regulatory domains, which can not be detected by typical sequence comparisons. Here we have developed and automated a protocol which clusters a given set of co-regulated genes in subsets of genes showing similar configurations of regulatory elements as revealed by TF-map alignments.

2 PROTOCOL

The protocol is schematized in Figure 1. Given a set of gene identifiers, in the first step, the upstream sequences of the genes are automatically extracted from the Ensembl database (Birney *et al.*, 2006). It is also possible to directly provide the upstream sequences in FASTA format. The second step is the search for putative TFBSs in the sequences. Two public position weight matrices (PWMs) libraries are available, Jaspar (Vlieghe *et al.*, 2006) and Transfac (v6.4) (Matys *et al.*, 2006). This step is performed using MatScan software (E. Blanco, unpublished). The third step performs the pairwise alignments of the TFBSs maps using the TF-alignment software (Blanco *et al.*, 2005). In the fourth step, the pairwise alignment scores are parsed to generate a score matrix. In the fifth step, the SOTA clustering algorithm (Herrero *et al.*) is applied to partition the gene space into clusters according to the score of the alignments of the TFBSs maps. Finally, for each gene cluster, the sixth step consists in running the multiple TF-map alignment software (E. Blanco, unpublished) to define a consensus "transcriptional regulatory pattern". To facilitate the analysis of the results, a graphical representation of the multiple TF-map alignment is produced using gff2ps tool (Abril *et al.*, 2003).

3 IMPLEMENTATION

At each step of this procedure corresponds a Web service that has been implemented following the BioMoby specifications (Wilkinson *et al.*, 2002). To facilitate the execution of the procedure, a data submission page has been setup at the following URL, {{http://genome.imim.es/cgi-bin/moby/GeneClustering_DataSubmission.cgi}}.

All the Web services that compose our procedure have been registered in the official BioMoby registry, as well as the one maintained by the INB ({{http://www.inab.org}}), under the following authority, *genome.imim.es*. We have also prepared a

workflow implementation to allow users to execute it using the stand-alone application called Taverna (Hull *et al.*, 2006).
Furthermore, the various Web services can also be executed separately or integrated into other pipelines of analysis. To this effect, one can use Taverna or Web-based tools such as Remora (Carrere *et al.*, 2006) or MOWServ (Navas-Delgado *et al.*, 2006).

4 DISCUSSION

The search for modules of *cis*-regulatory elements associated with co-expressed genes is still a challenging task. E. Blanco *et al.* have shown that comparisons of annotations of higher order domains can be more meaningful to characterize the underlying functionality of sequences than direct comparisons at the sequence level. Based on this method, here we have presented a fully automatized implementation of a protocol of analysis of co-expressed genes. The different steps of the pipeline may be executed in different distant computational resources, but this is totally transparent to the user. The pipeline encompasses many steps that the users would otherwise need to perform individually, and ensure, therefore the repeatability of this *in silico* experiment. Furthermore, because BioMoby Web services are formally described and their description is published in a central registry, this would also contribute in facilitating their discovery and their integration in other pipelines of analysis.

ACKNOWLEDGEMENTS

We thank Miguel Pignatelli for useful discussions during this work, Enrique Blanco for feedback and helpful comments on the manuscript, and Òscar Gonzalez for the technical support.
The work described here has been developed under grants from the Spanish Instituto Nacional de Bioinformática and the Spanish Ministerio de Educación y Ciencia.

REFERENCES

Abril, J.F. *et al.* (2003) gff2aplot: Plotting sequence comparisons. *Bioinformatics*, **19**, 2477-79.
Birney, E. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, Database issue, D556-61.
Blanco, E. *et al.* (2006) Transcription factor map alignment of promoter regions. *PLoS Comput. Biol.*, **2**, e49.
Carrere, S. *et al.* (2006) REMORA: a pilot in the ocean of BioMoby web-services. *Bioinformatics*, **22**, 900-1.
Herrero, J. *et al.* (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126-36.
Hull, D. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, Web Server issue, W729-32.
Matys, V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, Database issue, D108-10.
Navas-Delgado, I. *et al.* (2006) Intelligent client for integrating bioinformatics services., Bioinformatics, **22**, 106-11.
Stevens, R.D. *et al.* (2004) Exploring Williams-Beuren syndrome using myGrid. *Bioinformatics*, **20** Suppl 1, I303-I310.
Vlieghe, D. *et al.* (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, Database issue, D95-7.
Wasserman, W.W. *et al.* (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276-87.
Wilkinson, M.D. *et al.* (2004) BioMOBY: an open source biological web services proposal. *Brief. Bioinform.*, **3**, 331-41.
Wray, G.A. *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377-419.

Fig. 1. Co-expressed genes promoter analysis protocol description