APPLICATIONS NOTE

Manuscript Title

Co-responding Author[,*], Co-author[2] and Co-Author[2]
[1]Department of XXXXXXX, Address XXXX etc.

ABSTRACT

Summary: Here we present a set of Web services, compliant with MOBY specifications, that can be connected together to allow the implementation of a computational protocol to analyse the promoter regions of a given set of co-expressed genes. The main goal of this study is to show that in-silico analysis of genomic data can be done through Web resources in an automatized manner.

Availability: Web interface submission page, http://genome.imim.es/webservices/workflows/gene_clustering.html

Contact: arnaud.kerhornou@crg.es

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

1 INTRODUCTION

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

In-silico experiments can be seen as ...

In this regard, Web services architecture (Ref: 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 (Stevens R. et al, 2003).

We have applied such technology to the problem of gene promoter regulation. Promoter analysis is essential to understand the regulation of the expression of genes. It has already been shown that co-expressed genes present similar regulatory elements. These elements, known as Transcription Factor Binding Sites (TFBSs), are located in the intergenic sequence, upstream of the transcription start site (TSS) of a gene (Wray GA et al. 2003).

Here we present the implementation of a set of Web services, compliant with MOBY specifications (Wilkinson M. et al., 2002), that can be applied to implement a computational pipeline for analysing the promoter region of co-regulated genes and execute it in an automatic manner.

2 PROTOCOL

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 WW et al, 2004). These methods take into account background information to report statistically overrepresented motifs (Clover, Frith M. et al., 2004) or cross-species conservation data (PhyloGibbs, Siddharthan R. et al., 2005). Here we present a protocol based on the processing of the pairwise alignments of TFBSs, using the implementation of an algorithm that has been described by Blanco E. et al, 2005.

The final purpose of this protocol is to define transcriptional modules, corresponding to clusters of genes regulated by the same set of transcription factors (Leyfer D et al., 2005).

Given a set of gene identifiers, the following steps are performed (Figure 1.):

1/ Upstream sequence extraction. The upstream sequence of the input gene identifiers is automatically extracted from the Ensembl database (Birney et al., NAR, 2006).

2/ Search for putative TFBSs in public databases. You can either search the position weight matrices (PWMs) collection from Jaspar (Matys V. et al., 2003) or the one from Transfac (v6.4) (Sandelin A. et al., 2004). This can be done using MatScan software (E. Blanco, unpublished).

3/ Perform the pairwise alignments using meta-alignment software (E. Blanco et al.)

4/ Parsing of the pairwise alignment scores from meta-alignment outputs.

5/ Using this score matrix as input, perform a K-means clustering to partition the gene space into clusters according to the score of the TFBSs maps pairwise alignments. We are using Cluster v3.0 software (de Hoon M.J.L., 2004), originally written by Eisen M. (PNAS, 1998).

6/ For each cluster, run multiple meta-alignment software (E. Blanco et al., submitted) to define the consensus “transcriptional” regulatory pattern, common to a set of a co-regulated genes.

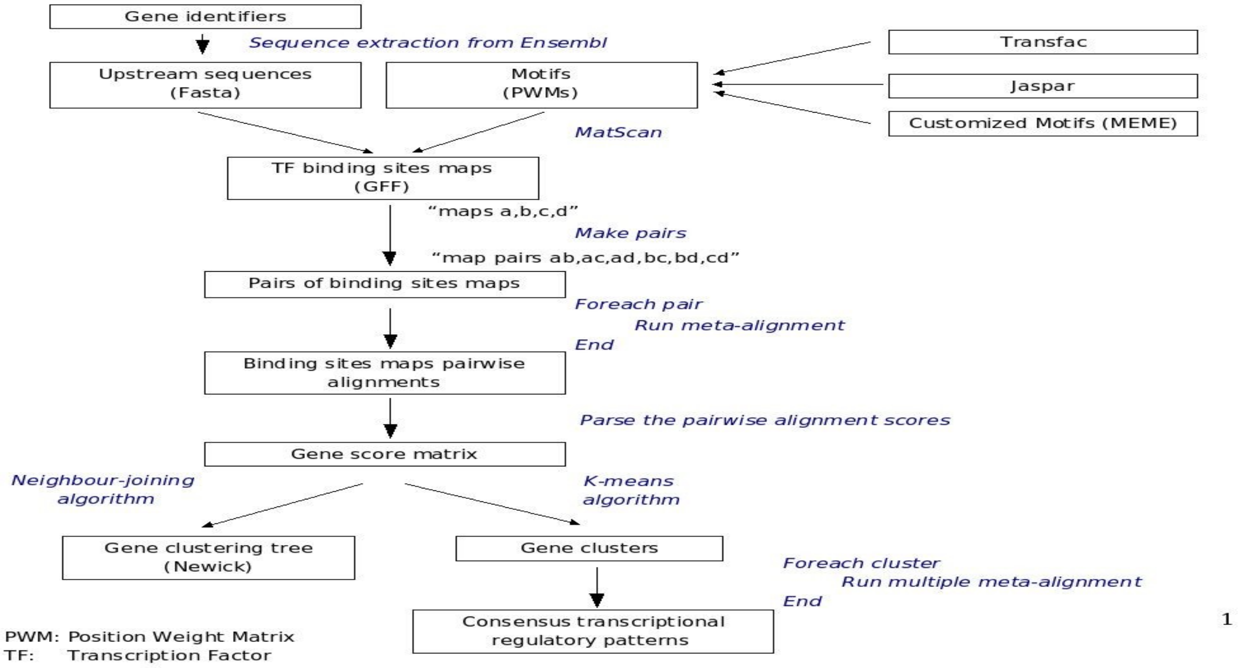
Gene identifiers
Sequence extraction from Ensembl
Upstream sequences (Fasta)
Motifs (PWMs)
Transfac
Jaspar
Customized Motifs (MEME)
MatScan
TF binding sites maps (GFF)
“maps a,b,c,d”
Make pairs
“map pairs ab,ac,ad,bc,bd,cd”
Pairs of binding sites maps
Foreach pair
Run meta-alignment
End
Binding sites maps pairwise alignments
Parse the pairwise alignment scores
Gene score matrix
Neighbour-joining algorithm
K-means algorithm
Gene clustering tree (Newick)
Gene clusters
Foreach cluster
Run multiple meta-alignment
End
Consensus transcriptional regulatory patterns
PWM: Position Weight Matrix
TF: Transcription Factor
1

Fig. 1. Protocol Description

3 IMPLEMENTATION AND RESULTS

1/ Web services

Each step of the protocol described below can be perceived as a service that can be formally described. A service can be defined as a resource that takes one or more inputs, process them and returns one or more outputs. For example, let us take the step two of our protocol. The input is a DNA sequence and the output is a map of putative TFBSs. The service defines two optional parameters, the first one is the database of TFBSs represented as PWMs (An enumeration of two values, “Jaspar” and “Transfac”). The second parameter is a threshold.

The implementation of such a service was done in Perl, following MOBY specifications.

2/ Workflows

There are various ways to enact the workflow. To facilitate the execution of the protocol, a Web submission page has been setup. One could also use the Standalone application, Taverna or Web-based, Remora. The advantage of these tools is that you can make your own workflows, by combining Web resources.

Fig. 2 => Workflow implementation using taverna 1.4 (In supplementary materials).

We have applied such pipeline to the genes involved in the X pathway in *Drosophila melanogaster*.

Fig 3 or table 1 (in supplementary materials) => Results on our dataset

4 DISCUSSION

Web services technology look promising to federate computational Bioinformatics resources through the Web. Here we have applied the MOBY framework to develop a set of Web services that can be combined together to allow the execution of a pipeline of analysis of the promoter regions of co-expressed genes. Such developments would strongly contribute to the reproducibility of in-silico experiments. Furthermore, because Web services are formally described, this would also contribute in facilitating their discovery. Finally, unified framework would facilitate programmatic access to remote resources.

ACKNOWLEDGEMENTS

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

REFERENCES

Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design Patterens Applied. Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog.