Mapping regulatory network from one organism to another

Rachita Sharma, Patricia Evans, Virendra Bhavsar

Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada E3B 5A3

Introduction

Determination of regulatory networks from available data is one of the major challenges in bioinformatics research. A regulatory network of an organism is represented by a set of genes and their regulatory relationships, which indicate how a gene or a group of genes affect (inhibit or activate) production of other gene products. Some organisms such as yeast, *A. thaliana* and fruit fly have been investigated very thoroughly by biologists as model organisms, being simpler and having shorter life cycles. We have developed a system to map the regulatory network from a model organism (source genome) to a non-model organism (target genome), about which less information is known.

Objectives

- Map regulatory elements and their relationships (links) from a model organism to a non-model organism
- Compare different methods used to map the regulatory links

Regulatory Elements Mapping

- Map transcription factors based on (Figure1)
  - sequence similarity - **TFbl**
  - protein family classification - **TFf**
  - protein sub-family classification - **TFsf**
- Map target genes based on (Figure 2)
  - sequence similarity - **TGbl**
  - using ranscription factor binding site (TFBS) motifs - **TGbs**
  - sequence similarity and TFBS motifs – **TGblbs**
  - using transcription factor binding site (TFBS) motifs on promoters- **TGpr**
  - finding transcription factor binding site (TFBS) motifs - **TGgalf**

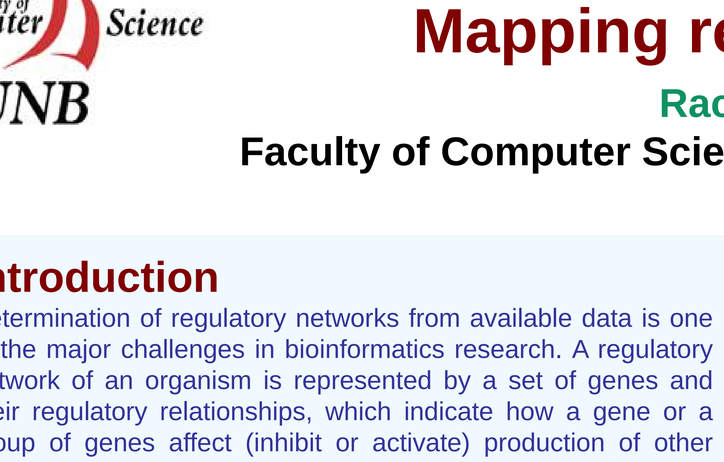


Figure 1: Method to map transcription factors from a source genome to a target genome

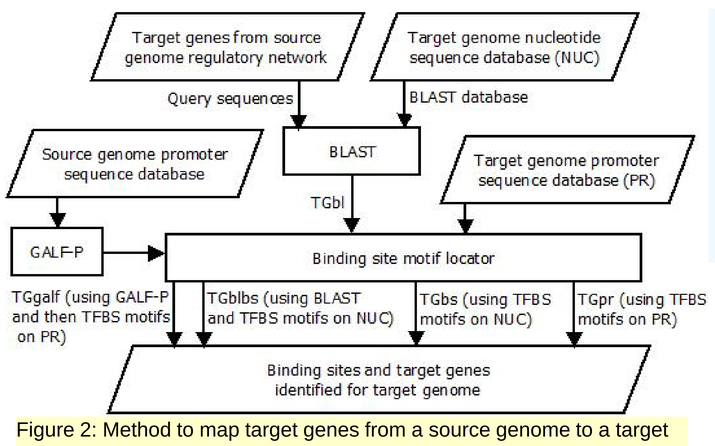
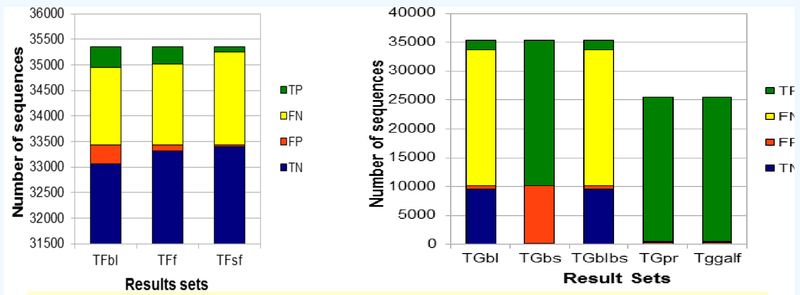


Figure 2: Method to map target genes from a source genome to a target genome

We have used *S. cerevisiae (*species of budding yeast*)* as the source genome and *A. thaliana* as the target genome for experimentation in this work. We evaluated the mapped transcription factors (TF) and target genes (TG) by comparing them to the available transcription factor data and binding site data of *A. thaliana*, respectively. The result sets are compared as shown in Figure 3(a) and 3(b) based on True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). We found that transcription factor mapping based on same protein family classification (TFf) has better performance than the other two result sets based on sequence similarity (TFbl) and both sequence similarity and same protein sub-family classification (TFsf). The improvement in accuracy over TGbs, TGbl and TGblbs makes TGpr (using TFBS on promoters) and TGgalf (finding TFBS) the best sets for mapping target genes.



(a) (b)

Figure 3: Comparing result sets of (a) transcription factor mapping methods and (b) target gene mapping methods

Regulatory Elements Integration

We integrate the mapped regulatory elements (TF and TG) to predict regulatory links for the target genome as shown in Figure 4.

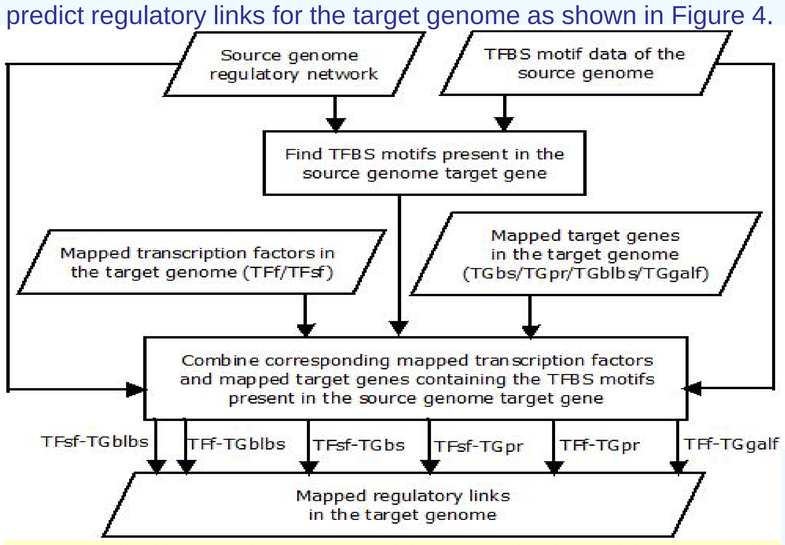


Figure 4: Method to integrate mapped regulatory elements into regulatory links for the target genome

Regulatory Links Verification

We specify rules to evaluate the predicted regulatory links using gene expression experiments. The expression values of both transcription factor and target gene from a regulatory link should be present in the experiment to evaluate that regulatory link. An experiment can either Confirm (*C*), Contradict ($\overline{C}$) or be Neutral (*N*) for any regulatory link as shown in Table 1.

Transcription Factor Expression	Gene Expression	Result
Yes	Yes	*C*
Yes	No	$\overline{C}$
No	Yes	*N*
No	No	--

(a)

Transcription Factor Expression	Gene Expression	Result
Yes	Yes	$\overline{C}$
Yes	No	*C*
No	Yes	*C*
No	No	$\overline{C}$

(b)

Table 1: Rules to verify predicted regulatory links with type (a) positive gene regulation and (b) negative gene regulation

The Confirmed value (*c*) for a regulatory link represents the number of experiments that support that link. The experiments contradicting the regulatory link are part of the Contradicted value ( ). The rest of the experiments that neither support nor contradict the regulatory link, but do provide additional information about the regulatory link, are represented by the Neutral value (*n*).

The results for the six predicted regulatory links sets are shown in Table 2. Rows 5 to 9 in the table show the different conditions used to evaluate the results based on Confirmed, Contradicted and Neutral values. The TF mapping result set TFf comprises the best results when we only consider how many TFs are mapped, but it does not produce the best regulatory links set when integrated with the mapped TG set TGblbs. The TGbs set, TGpr set and TGgalf set contain the best results of mapped TGs from the previous subsection, but they do not work well when used in the integration of regulatory elements to predict regulatory links. The additional predicted TGs in these sets lead to too many false regulatory links in the sets TFsf-TGbs, TFf-TGpr, TFsf-TGpr, and TFf-TGgalf. These false links indicate that many of the true TGs identified for the target genome in set TGbs, set TGpr and set TGgalf are, however, not the correctly mapped TGs linked to the right TF in their corresponding regulatory link result sets.

	TFsf-TGblbs	TFf-TGblbs	TFsf-TGbs	TFsf-TGpr	TFf-TGpr	TFf-TGgalf
Number of mapped TFs	118	434	118	118	434	434
Number of mapped TGs	2252	2252	35181	25259	25259	25388
Number of links mapped	43423	480524	536154	274400	3182551	242030
Number of links analyzed	3056	8621	6085	34529	236585	25227
$c \geq 2\overline{c}+n$	2090	2628	2995	2343	93287	3727
$c \geq 3\overline{c}$	2109	2375	2968	2323	101224	2012
$2\overline{c} \leq c < 3\overline{c}$	58	344	30	26	12737	1715
$\overline{c} \leq c < 2\overline{c}$	103	215	100	66	19037	503
$c < \overline{c}$	786	5687	2987	2050	87983	150

Table 2: Regulatory links confirmed for A. thaliana using 43 gene expression experiments

Table 2 shows that the set TFsf-TGblbs has better results than the other regulatory link sets, based on having a higher percentage of true regulatory links identified and a lower percentage of links contradicted more than they are confirmed. Therefore, integrating the mapped TFs based on protein sub-family classification along with the mapped TGs based on sequence similarity and TFBS motifs produces the best results for regulatory links. These results indicate that the regulatory relationships are conserved between genomes and can be mapped from one genome to another.

For future work, more information about gene regulation at different stages of gene expression can be incorporated, along with the new data that becomes available for the non-model organism, to map the regulatory network.

Publications:

Sharma, R.; Evans, P.A.; Bhavsar, V.C., "Transcription Factor mapping between Bacteria Genomes", *International Journal of Functional Informatics and Personalised Medicine*, **2009**, Vol. 2, 4, 424-441.

Sharma, R.; Evans, P.A.; Bhavsar, V. C., "Regulatory link mapping between organisms", Accepted by *BMC Bioinformatics*.

Canadian Potato Genome Project
BIO ATLANTECH
Atlantic Canada Opportunities Agency
New Nouveau Brunswick CANADA
Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada
Canada
GenomeCanada