Supplementary Material (ESI) for Molecular BioSystems

Genome-Wide Survey of MicroRNA - Transcription Factor Feed-Forward Regulatory Circuits in Human

Angela Re #, Davide Corá #, Daniela Taverna and Michele Caselle *

# equal contribution

* corresponding author, email: caselle@to.infn.it, phone: +39-011-6707205, fax: +39-011-6707214

Supporting Information

Oligos analysis

We searched for statistically significant motifs in promoters and 3'-UTRs using the algorithms developed in our previous works [1, 2], with minor modifications. Here we report the main features of these algorithms for completeness.

Motif search in the promoter regions

We preprocessed the promoter regions by merging overlapping fragments, in order to build a non-redundant set of sequences for the evaluation of the background probabilities.

Both the original dataset of promoter regions and the set of non overlapping blocks were then separated according to their CG content. Following what we did in ref [2] we associated to each gene a label (CG-rich or CG-poor) based on the CG content of the corresponding promoter region using the median of the CG content of the whole set of promoter regions as threshold.

Motif search in the 3'-UTR regions

We applied the same pipeline to the set of 3'-UTRs. We first built a set of non-redundant sequences by merging the overlapping ones and then divided both datasets (the original and the non-redundant one) into CG-rich and CG-poor subsets. The only difference with respect to the promoter case was that we merged the sequences only when they were located on the same strand.

Once equipped with those dataset, we constructed, separately for human and mouse, the sets $S(w)$ of genes such that the oligo w is overrepresented in the corresponding promoter or 3'-UTR. This was done for all 6 to 9 nt oligos for promoter regions and for 7 nt oligos for 3'-UTRs through the following steps:

- We computed the overall frequency $f(w)$ as the ratio

$$f(w) = \frac{N(w)}{N} \tag{1}$$

  where $N(w)$ is the number of times w occurs in the collection of all the non-redundant sequences, and $N = \sum_w N(w)$.

- For each gene g let $n_g(w)$ be the number of occurrences of w in the original (i.e. before merging overlapping sequences) promoter or 3'-UTR of g. We computed the overrepresentation Pvalue as

$$P_g(w) = \sum_{k=n_g(w)}^{n_g} \binom{n_g}{k} f(w)^k (1 - f(w))^{n_g - k} \tag{2}$$

  where

$$n_g = \sum_w n_g(w) \tag{3}$$

  is the total number of oligos of the same length as w that can be read in the promoter or the 3'-UTR of g. Self-overlapping matches of the same oligo were discarded [3]. Motifs were counted on both strands in the promoter case and only on the transcribed strand in the 3'-UTR case.

- The genes for which $P_g(w) < 0.01$ were included in the set $S(w)$. Notice that no biological significance was ascribed to these sets before they are selected for evidence of evolutionary conservation as explained later: therefore the choice of the cutoff on P can be arbitrarily lenient

and in particular no correction for multiple testing was applied in this step.

The procedure described above was performed separately for CG-rich and CG-poor genes, so as to identify overrepresented words with respect to the appropriate background frequencies. For each word w the sets $S(w)$ computed for CG-rich and CG-poor genes were then joined to obtain a single set $S(w)$. This procedure was performed separately for human and mouse.

Conservation of overrepresentation

We define an oligo w as "conserved overrepresented" if the sets of genes $S_{human}(w)$ and $S_{mouse}(w)$ contain a significantly larger number of orthologous genes than expected by chance. Pairs of human-mouse orthologous genes were obtained from Ensembl, selecting only orthologous defined as Unique Blast Reciprocal Hit to obtain one-to-one orthology relationships. For miRNA genes, orthology relatioships were downloaded from [4].

Let M be the total number of human genes represented in our sequences which have a mouse ortholog. Given an oligo w and the set $S_{human}(w)$, let m be the number of human genes in $S_{human}(w)$ which have a mouse ortholog, N the number of genes in $S_{mouse}(w)$ with a human ortholog, and n the number of genes in $S_{human}(w)$ with a mouse ortholog in $S_{mouse}(w)$. We then compute the Pvalue

$$P = \sum_{k=n}^{m} F(M, m, N, k) \tag{4}$$

where

$$F(M, m, N, k) = \frac{\binom{m}{k}\binom{M-m}{N-k}}{\binom{M}{N}} \tag{5}$$

Multiple testing was taken into account with a Benjamini-Yekutiely FDR approach [5], and conserved overrepresentation was defined to be significant when the Benjamini-Yekutiely corrected Pvalues were less than 0.1. Note that a similar approach was used e.g. in [6], and termed "Network conservation"; basically the main difference with respect to our approach is that in [6] the simple presence of a k-mer is used instead of statistical overrepresentation.

Search of putative TF binding sequences in the promoter motifs database

We searched for putative TF binding sequences in the set of conserved overrepresented promoter motifs with two different and complementary approaches.

- **TRANSFAC analysis**

  We performed a weight matrix search of TF binding sites among our motifs using the Match web-based tool that is integrated in TRANSFAC Professional (release 11.2). In a few cases the motif turned out to be compatible with more than one weight matrix. In these cases we associated all the potential binding factors to the motif. Weight matrix profiles adopted in our search were limited to the vertebrate-specific subset of high quality matrices with predefined cutoffs for core and matrix similarity optimized in order to minimize the sum of both false positive and false negative rates.

  To restrict the search, we specified the predefined human site set and allowed no mismatching nucleotide in a match between search pattern and motif.

- **Comparison with the Xie et al.[7] list of binding sequences**

  A second approach was the direct comparison with the consensus sequences for vertebrate TFs published in [7] (see Supplementary table S3 of their work). The association between motif and TF was accepted only if our motif exactly overlapped (according to the IUPAC alphabet) the consensus taken from the [7] list.

In this way we could obtain a list of putative target genes for the TF's contained in TRANSFAC and in the [7] paper: for each TF we simply merged together the genes contained in the sets $S(w)$ associated to all the words w compatible with the TRANSFAC weight matrix or the [7] consensus.

Search of putative miRNA binding sequences in the 3'-UTR motifs database

We obtained human mature miRNA sequences from the miRBase Sequence Database (release 9.2) in flat file form from `ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/`. We followed standard seed parameter settings and considered 7 nts long

seeds, beginning within the third position from the miRNA 5' end (nt 1 or 2 or 3). We identified a motif (a miRNA binding site) if it had a perfect Watson-Crick match with at least one of the possible miRNA seeds [2].

In this way we could obtain a list of putative target genes for each one of the 193 mature miRNAs contained in our database: for each miRNA we merged together the genes contained in the sets $S(w)$ associated to all the words w identified as miRNA binding sites.

Randomizations for network motifs analysis

We adopted three different randomization protocols to study the statistical properties of our mixed FFLs.

Random reshuffling of miRNA promoters and seeds We constructed a randomized versions of miRNA promoters and seeds and then constructed the putative FFLs following the same pipeline discussed in the text. All the other parameters of the pipeline were left unchanged and we performed a total of 50 randomizations and evaluated the mean value and standard deviation of the number of FFLs. For each randomization, the catalogue of random miRNA promoters was build performing $NlogN$ swaps between two randomly selected nucleotides in the original promoter sequence, preserving the repeat-masking (being N the effective length of the miRNA promoter sequence). In this way, the CG content of each miRNA promoter was kept unchanged. To build the set of random 7-mers used as miRNA seeds, we started with the catalogue of real seed regions used in this work and performed, as above, $NlogN$ swaps between two randomly selected nucleotides, and retaining only oligos not already identified with any known miRNA seed.

Edge switching, (see [8]). In this type of randomization, we started from the transcriptional and post-transcriptional regulatory networks obtained with our sequence analysis pipeline (and reported in Supplementary files S3 and S5), and applied the following randomization strategy: we randomly selected two TFs (or miRNAs) and two of their target genes, one for each of them. We then swapped the two target genes, only if the selected genes were not already present in the list of regulated genes of the two TFs (or miRNAs). In that way, the out-degree of a certain TF and the in-degree of a certain target gene were preserved, whilst the connections were randomized; the same for a miRNA and its targets. For each complete

randomization cycle, a total of 10^6 exchanges was executed. The randomization protocol was performed 1000 times and the mean value and standard deviation of the number of FFLs was evaluated.

Complete node replacement, (see again [8]). In this type of randomization, we started again from our transcriptional and post-transcriptional regulatory networks, and applied a less constrained randomization strategy: we randomly selected two TFs (or miRNAs) and two of their target genes, one for each of them. We then swapped the two target genes, without any filter. In this way, the global network topology of the original regulatory network was completely modified. For each complete randomization cycle, a total of 10^6 exchanges was executed. Also this randomization protocol was performed 1000 times and the mean value and standard deviation of the number of FFLs was then evaluated.

List of Supplementary files

Supplementary file 1: **List of human and mouse pre-miRNAs included as transcriptional units (TU) representatives.**

Supplementary file 2: **List of human and mouse mature miRNAs included in our study.**

Supplementary file 3: **Transcriptional regulatory network: conserved-overrepresented oligos in promoter regions of protein-coding and microRNA genes and their association to known transcription factor binding sites.**

Supplementary file 4: **Transcriptional regulatory network: conserved-overrepresented oligos in promoter regions of microRNA genes and their association to known transcription factor binding sites.**

Supplementary file 5: **Post-transcriptional regulatory network: conserved-overrepresented oligos in 3'-UTRs of protein-coding genes and their association to known microRNA seeds.**

Supplementary file 6: **Raw data of circuits.**

Supplementary file 7: **Raw data of circuits related to unknown Transcription Factors.**

Supplementary file 8: **Circuits final datasets with complete annotations.**

Supplementary file 9: **Circuits relevant to cancer, related to known Transcription Factors.**

Supplementary file 10: **Circuits relevant to cancer, but related to unknown Transcription Factors.**

Supplementary file 11: **Raw data of circuits for the alternative choice of the promoter region: (-500/+100) around the TSS.**

Supplementary Figure S1: **Randomization results for the network motifs analysis of mixed feed-forward loops.** We plotted the number of single-target mixed Feed-Forward Loops (FFLs) obtained in the real network and associated to known Transcription Factors (blue line) alongside the distributions (normalized histograms) of the number of single-target mixed FFLs detected for the three randomization strategies adopted. The red data refer to the results obtained with the *Random reshuffling of microRNA promoters and seeds*, whilst the light green refers to the *Edge switching* randomization strategy and the dark green to the *Complete node replacement* one. The figure is divided into two separated panels: panel A) contains results relative to the (-900/+100) nts window for the definition of promoter of miRNAs and protein-coding genes, whereas panel B) contains results relative to the (-500/+100) nts case. See Supporting Text for details.

Supporting website.

All the Supplementary files and raw data are available at:
`http://personalpages.to.infn.it/~cora/circuits/index.html`.

References

[1] Corà D, Herrmann C, Dieterich C, Di Cunto F, Provero P, et al. (2005) Ab initio identification of putative human transcription factor binding sites by comparative genomics. BMC Bioinformatics 6:110.

[2] Corà D, Di Cunto F, Caselle M, Provero P (2007) Identification of candidate regulatory sequences in mammalian 3' utrs by statistical analysis of oligonucleotide distributions. BMC Bioinformatics 8:174.

[3] van Helden J, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827–842.

[4] Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, et al. (2007) A mammalian microrna expression atlas based on small rna library sequencing. Cell 129:1401–14.

[5] Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. ANN STAT 29:1165–1188.

[6] Chan C, Elemento O, Tavazoie S (2005) Revealing posttranscriptional regulatory elements through network-level conservation. PLoS Comput Biol 1:e69.

[7] Xie X, Lu J, Kulbokas E, Golub T, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. Nature 434:338–345.

[8] Martinez N, Ow M, Barrasa M, Hammell M, Sequerra R, et al. (2008) A c. elegans genome-scale microrna network contains composite feedback motifs with high flux capacity. Genes Dev 22:2535–49.