UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Biologia Animal

IN SILICO ANALYSIS OF MIRNA PROMOTERS

Fernando Manuel Magalhães Martins

Mestrado em Bioinformática e Biologia Computacional

2011

UNIVERSIDADE DE LISBOA
Faculdade de Ciências
Departamento de Biologia Animal

IN SILICO ANALYSIS OF MIRNA PROMOTERS

Fernando Manuel Magalhães Martins

Trabalho orientado pelo Prof. Doutor Francisco J. Enguita e pela Doutora Ângela Inácio

Mestrado em Bioinformática e Biologia Computacional

2011

Resumo

Os microRNAs (miRNAs) contribuem de uma forma abundante para a fracção de RNAs não-codificantes eucariotas. Estes estão envolvidos na regulação negativa pós-transcricional da expressão genética através da ligação com a região 3'-UTR dos transcritos de mRNA nascente, conjuntamente com várias outras proteínas ajudantes. Em mamíferos, manifesta-se principalmente através da inibição da síntese proteica. Actualmente, sabe-se que estas moléculas de RNA são reguladores moleculares mestre envolvidos em processos celulares que englobam a diferenciação, transdução de sinal, divisão celular e cancro.

A expressão dos microRNAs parece ter uma assinatura específica para cada um dos tecidos. Ainda não está claro quais são os principais factores que controlam esta especificidade, porém vários autores têm postulado a existência de circuitos de regulação entre os factores de transcrição que controlam a expressão de miRNA e a regulação exercida pelo miRNA sobre a expressão do factor de transcrição.

Recentemente, as sequências de DNA de todos os promotores de miRNA humanos foram caracterizados por imunoprecipitação da cromatina por Marson *et al* **[1]**. Começamos com estes dados e a primeira coisa que se fez foi recolher todas estas sequências, usando a versão do UCSC Genome Browser indicada no estudo anterior e tendo em conta as posições nele indicadas para cada um dos 550 promotores. Para este efeito, foi necessário escrever um pequeno programa.

O presente trabalho tem como objectivo principal realizar uma caracterização *in silico* de todos estes promotores, estudando os factores de transcrição que possivelmente controlam a expressão de miRNAs. Procurou-se factores de transcrição que regulassem a expressão de cada um destes miRNAs e que, simultaneamente, fossem proteínas codificadoras alvo desses mesmos miRNAs.

O primeiro passo na análise dos circuitos de regulação entre os microRNA e os factores de transcrição foi a predição dos locais de ligação (TFBS) destes últimos para todas as sequências de promotores de miRNA obtidas. Ou seja, dadas as sequências de promotores de cada um dos miRNAs, era necessário saber quais os factores de transcrição que a elas se poderiam ligar e regular sua transcrição dos respectivos miRNAs.

Actualmente, existem vários programas disponíveis. No entanto, apesar de todos os esforços, esses algoritmos às vezes produzem muitos falsos positivos ou falsos negativos. Assim, um dos maiores problemas ainda existentes é como encontrar o software apropriado. Consequentemente, os investigadores costumam usar vários dos programas existentes. Nós

usamos o TFSEARCH 1.3, MAPPER 2, Match 1,0, Patch 1.0, P-Match 1.0, PROMO 3.0.2 e o TFBind.

A primeira diferença entre todos estes programas é a maneira como as sequências dos promotores lhes podem ser enviadas. O MAPPER 2, foi o único que foi capaz de processar um arquivo FASTA contendo todas as sequências de promotores. Para o TFSEARCH 1.3 foi possível descarregar o EZRetrieve. Esta é uma ferramenta gratuita que se baseia no TFSEARCH e também processou o arquivo FASTA completo. Para o TFBind concebemos uma ferramenta similar ao EZRetrieve. Este programa lê um arquivo FASTA e envia cada sequência à ferramenta TFBind que está disponível *online*. Em seguida, guarda os ficheiros HTML que podem ser obtidos quando se realizam as pesquisas *online*.

Para todas as outras ferramentas, era necessário um registo prévio nos locais onde elas se encontram disponíveis e, como tal, é necessário fazer o *login* antes de começar a usar essas ferramentas. Por este motivo, não foi possível conceber qualquer ferramenta para realizar esta pesquisa automaticamente. A única solução foi dividir nosso arquivo FASTA em vários arquivos pequenos e submeter cada um deles a cada uma dessas ferramentas.

Tendo esta quantidade enorme de dados proveniente dos sete programas, foi necessário, então, uniformizá-los e prepará-los para serem analisados, tendo sido necessário desenvolver diversos programas para o efeito. As principais questões surgidas durante este processo foram o facto de algumas das aplicações usadas não permitirem restringir os resultados a genes de Homo Sapiens e, para além disso, a identificação dos genes não ser feita de forma uniforma, em virtude de os mesmos terem diversas designações.

Para o efeito, descarregamos todos os genes de Homo Sapiens existentes na base de dados GenBank do NCBI. Além dos símbolos oficiais de cada gene, esta base de dados também contém os seus sinónimos. Depois de comparar os nomes dos genes, foi possível identificar a maioria dos genes obtidos nas aplicações de TFBS. No entanto, muitos deles permanecem por classificar ou não são genes de Homo Sapiens.

Hoje em dia, é evidente que os processos pós-transcricionais desempenham um papel muito mais importante na regulação da expressão génica do que o anteriormente esperado. Assim, um passo crucial para a análise de funções reguladoras dos miRNAs é a previsão de seus alvos. Actualmente, existem diversos programas e bases de dados disponíveis. Nós usamos o Diana micro-T, Miranda, miRWalk, miRTarBase e uma base de dados publicada em 2010 por Saito T e P Sætrom **[44]**.

Por comparação com o processo de análise das bases de dados de TFBS, estas revelaram uma melhoria considerável na forma de identificação dos genes, pois algumas delas usam identificadores únicos, quer sejam do GenBank ou do sistema Ensemble. Dado que os dados dos genes extraídos do GenBank também incluem os identificadores Ensemble, esta questão da identificação dos genes nas bases de dados de *targets* não obrigou a tanto esforço e permitiu certamente resultados mais fiáveis.

A principal questão surgida com a análise das bases de dados de *targets* foi o volume de dados das mesmas. Estas bases de dados contêm geralmente milhões de registos e, apesar de os formatos das mesmas serem de muito mais fácil tratamento, obrigam a que se desenvolvam ferramentas para a extracção dos dados pretendidos. Refira-se que a maior destas bases de dados por nós usadas contém cerca de 20 milhões de registos.

Depois de analisar todos os dados seleccionados, encontramos 38.773 *loops*, cobrindo 285 diferentes factores de transcrição e 417 miRNAs distintos. Estes *loops* envolvem factores de transcrição que regulam a expressão de um miRNA e que, simultaneamente, são proteínas codificadoras alvo desse mesmo miRNA. No entanto, cada *loop* é composto por um único factor de transcrição e um único miRNA.

Uma vez que um único miRNA pode regular múltiplos genes e um único gene pode ser regulado por múltiplos miRNAs, é bastante natural pensar que miRNAs e factores de transcrição possam cooperar na regulação dos genes-alvo tanto a nível transcricional como a nível pós-transcricional. Na verdade, factores de transcrição e miRNAs funcionam juntos em redes reguladoras de genes que ainda não estão completamente identificadas nem compreendidas. Consequentemente, todos os *loops* identificados por este estudo devem ser vistos como componentes de módulos reguladores, em vez de *loops* isolados. Embora isto seja verdade, também podemos analisar individualmente cada um destes *loops*.

Tendo em mente o facto de que esta é uma análise *in silico*, devemos estar cientes que a grande maioria de todos os *loops* detectados têm uma probabilidade muito baixa de serem *loops* reais. Portanto, futuras investigações devem começar pela definição de critérios de fiabilidade de todos os dados obtidos. Na verdade, todos estes dados exigem futuras investigações e necessitam de validações experimentais.

Assim, este trabalho permitiu reunir e catalogar *loops* de regulação mistos entre factores de transcrição e miRNAs, em humanos, tendo sido todos os dados processados e armazenados numa base de dados relacional. Além disso, foi desenvolvida uma plataforma *web* de modo a permitir futuras investigações, pois apesar de ainda não compreendermos

perfeitamente o significado biológico destes circuitos, eles são provavelmente um importante mecanismo de regulação da expressão génica. Esta base de dados tem 36 tabelas e armazena mais de 2,5 milhões de registos. A interface *web* permite a procura de *loops* usando vários critérios de pesquisa e permite a análise de todos os detalhes de cada um dos *loops*, tais como os TFBS previstos, os *targets*, as pontuações associadas a cada previsão, etc.

Palavras-Chave: microRNAs, Factores de Transcrição, Circuitos de Regulação, Bases de Dados de TFBS, Bases de Dados de *Targets*

Abstract

MicroRNAs (miRNAs) are an abundant class of eukaryotic non-coding RNAs. They are involved in the negative post-transcriptional regulation of gene expression. Their inhibitory action is exerted by binding to the 3'-UTR region of nascent mRNA transcripts together with several other helper proteins, and in mammals it is observed mainly as an inhibition of protein synthesis. These non-protein coding RNA molecules are master molecular regulators that have been found to be involved in cellular processes ranging from differentiation, cell division, signal transduction and cancer.

MicroRNAs expression appears to have a tissue specific signature in which specific miRNAs are expressed preferentially in some tissues or organs. It remains unclear which are the main factors that control this tissue-specificity, however several authors have postulated the existence of a regulatory feedback loop between transcription factors controlling miRNA expression and the regulatory control exerted by miRNA over the transcription factor expression.

Recently, the DNA sequences of all the human miRNA promoters have been characterized by chromatin-immunoprecipitation **[1]**. The present work has the main objective of performing an in silico characterization of all these promoters, studying the possible transcription factors controlling miRNA expression. We were looking for transcription factors regulating miRNA expression and being simultaneously the target protein-coding gene of that same miRNA. Despite the fact that we cannot yet understand the biological significance of these regulation loops, this must be an important mechanism of genes regulation.

The purpose of this work was to assemble and characterize a catalogue of such mixed transcription factor/miRNA regulation loops in humans. All data was processed and stored in a relational database. Furthermore, a web platform was developed in order to enable further investigations.

Keywords: microRNAs, Transcription Factors, Regulation Loops, TFBS Databases, Targets Databases

Contents

List of Tables

List of Figures

Acronyms

DNA – Deoxyribonucleic Acid

FASTA – FAST-All File

GIGO – Garbage In, Garbage Out

GO – Gene Ontology Database

GTP – Guanosine-5'-Triphosphate

HS – Homo Sapiens

NCBI – National Center for Biotechnology Information

PCA – Principal Component Analysis

RNA – Ribonucleic Acid

RISC – RNA-Induced Silencing Complex

SVM – Support Vector Machines

TF – Transcription Factor

TFBS – Transcription Factor Binding Sites

TSS – Transcription Start Site

UCSC – University of California Santa Cruz Genome Browser

Background

MicroRNAs (miRNAs) are small (≈22 nucleotides), non-protein coding RNA molecules known to regulate the expression of genes by binding to the 3'-untranslated regions (3'-UTR) of mRNAs. The first microRNA molecules, lin-4 and let-7, were identified in 1993 **[2]** and, since then, there has been a rapid progress in identifying more miRNAs and understanding their biogenesis, functionality and their target gene regulation.

The majority of the miRNAs identified in the first 10 years were located in the noncoding regions between genes and transcribed by unidentified promoters. These miRNAs that are produced from their own genes are also known as intergenic miRNAs. In 2003, Ambros *et al* **[3]** also discovered some tiny noncoding RNAs derived from the intron regions of gene transcripts; these are intronic miRNAs, i.e., miRNAs produced from introns. A schematic description of miRNAs biogenesis is illustrated in **Figure 1**.

Transcription factors (TFs) are proteins that either activate or repress genes transcription by binding to short cis-regulatory elements called transcription-factor binding sites. These binding sites are located in the upstream region of genes – the promoter region, which is located around the transcription start site (TSS). Post-transcriptionally, microRNAs repress mRNA translation by binding to partially complementary sites, called miRNA binding sites, in their target mRNAs. In animals, miRNA-mediated repression is often relatively weak, whereas transcription-factor-mediated repression can be much stronger **[4]**.

Similarly to TFs, a single miRNA can regulate multiple genes, and a single gene can be regulated by multiple miRNAs. Thus, it seems quite natural to think that both miRNAs and TFs may cooperate in regulating the same target genes at the transcriptional and post-transcriptional levels. However, the molecular mechanism and nature of this interaction has not yet been understood.

TFs are essential for transcription by binding to transcription-factor binding sites. The resulting transcript is capped with a specially-modified nucleotide at the 5' end, and polyadenylated with multiple adenosines - a poly(A) tail, at the 3' end **[5]**. In the case of the miRNAs, this initial transcript, also known as primary miRNA (pri-miRNA), can be hundreds to thousands of nucleotides long and may contain several miRNA precursors. Each one is a hairpin loop structure composed by 60 to 80 nucleotides.

The double-stranded hairpin loop RNA structure is then recognized by a nuclear protein known as DGCR8 or “Pasha”. Pasha associates with the enzyme Drosha and orients this last one to excise the hairpin structure. The resulting hairpin, known as pre-miRNA, is exported from the nucleus to the cytoplasm in a process mediated by Exportin-5 protein. This transportation is energy-dependent, using GTP bound to the Ran protein **[6]**.

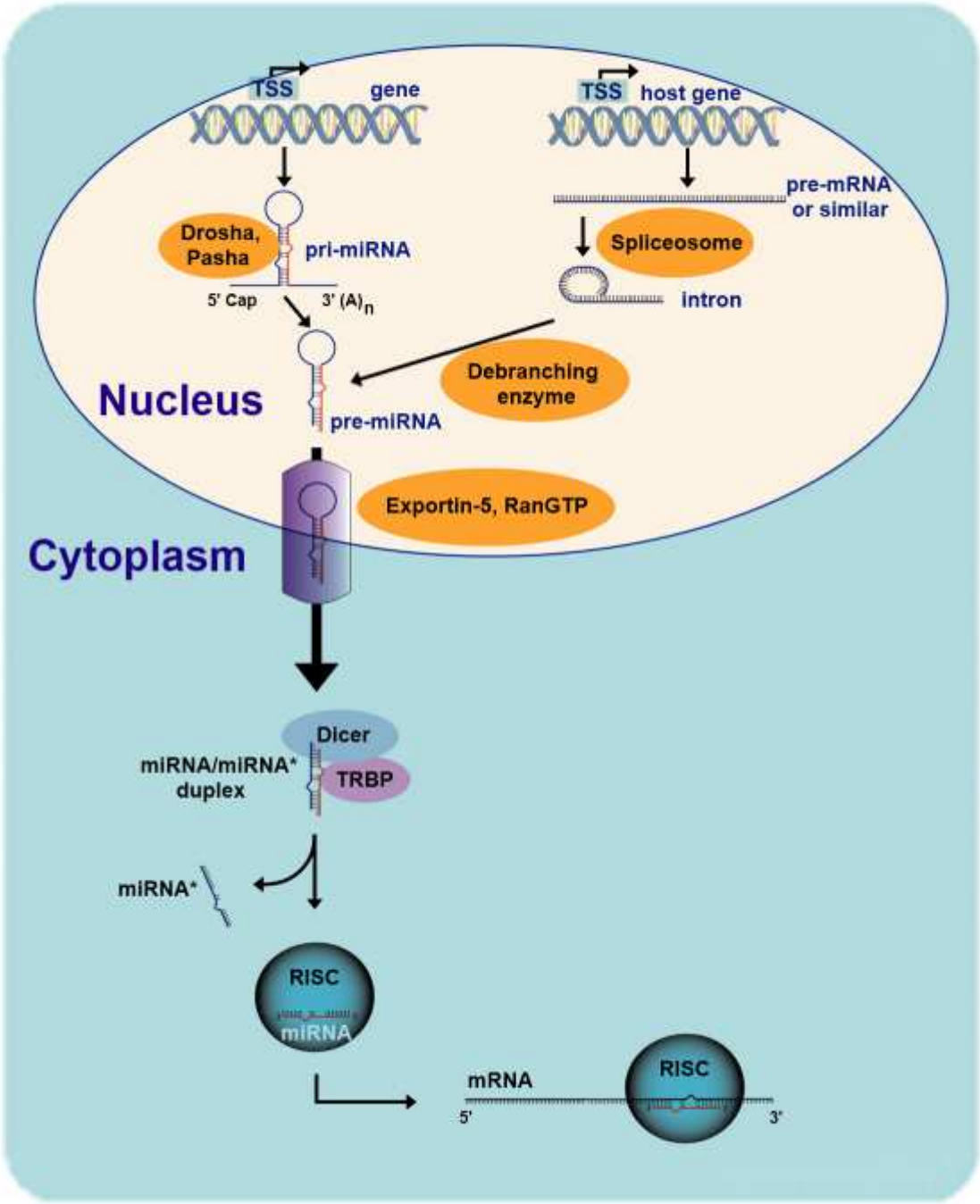


Figure 1. MicroRNAs biogenesis: MicroRNAs are produced from either their own genes or from introns

In the cytoplasm, the pre-miRNA hairpin is recognized and cleaved by the Dicer enzyme, and its binding partners, TRBP protein included. This complex removes the loop

region of the hairpin structure, releasing the miRNA:miRNA* duplex which is approximately 22 nucleotides long. The strand of the miRNA duplex that is less thermodynamically stable is preferentially loaded into the RNA-induced silencing complex (RISC) **[7]**, which includes Dicer, TRBP and Argonaute proteins. The strand loaded into the RISC complex is called the guide strand and directs the RISC complex to its mRNA target. The other strand, the passenger strand or miRNA*, is subsequently degraded by an unknown mechanism **[8]**.

The mature miRNA loaded into to the RISC complex guides both to their mRNA target and usually binds to the 3'-UTR of the mRNA. This association may result in either cleavage or translational inhibition of the target mRNA, depending on the base pair complementarity between the miRNA and the mRNA target region. Perfect complementarity usually results in mRNA cleavage by the RISC complex, whereas imperfect base pairing leads to translation repression **[8]**.

Predicting transcription factor binding sites

The first step in the analysis of the transcription factor/microRNA regulation loops was to predict the transcription factor binding sites (TFBS) for all sequences of miRNA promoters published by Marson *et al* **[1]**. Given the miRNAs promoters sequences, it was necessary to know which TFs could bind to those promoters and regulate their transcription.

Currently, there are several programs available, e.g. AliBaba 2.1 **[9]**, TFSEARCH 1.3 **[10]**, Genomatix MatInspector **[11]**, MAPPER 2 **[12]**, Match 1.0 **[13]**, P-Match 1.0 **[14]**, PROMO 3.0.2 **[15]** and TFBind **[16]**. Predicting TFBS using position weight matrices (PWM) is widely used and theoretically supported by Berg and von Hippel **[17]**. Each matrix relates a consensus sequence to the four bases and each score is proportional to the binding energy for the protein–DNA interaction **[18]**. **Figure 2** illustrates this.

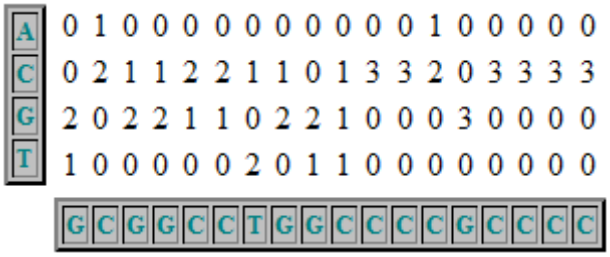


Figure 2. Sp1 [T00757] Matrix on TRANSFAC 8.3

Matrices and TFBS have been collected into databases such as TRANSFAC **[19]** and JASPAR **[20]**. However not only all matrices have their own specificity, as prediction also requires the quantification of the similarity between the each weight matrix and a potential TFBS detected in the sequence.

In order to achieve a greater degree of accuracy, when comparing to the existing ones, several algorithms have been proposed in the last years. However, despite all efforts, these algorithms sometimes produce many false positives or false negatives. Thus, one of the major remaining problems is how to find the appropriate software. Consequently, investigators often use several of the existing programs.

Predicting microRNA targets

Nowadays it is evident that post-transcriptional processes play a much more important role in the regulation of gene expression than previously expected. So, a crucial step for the analysis of regulatory roles of miRNAs is the prediction of their targets. Although we do not know exactly the precise way how miRNAs play their role, it is known that, in animals, miRNAs are able to repress the translation of target genes by binding to a small region of nucleotides that are present at the 3'-UTR region of the regulated gene **[21]**. This region, called "seed", is located at positions 2-8 of the 5' end of miRNAs and is known to contribute significantly to target recognition **[22]**. That is why most existing algorithms start by trying to find regions of 3' UTR target gene that have strong Watson-Crick base pairing complementary to the miRNA seed sites.

Since this initial step usually results in thousands of potential target sites and many false positives, most algorithms also use other prediction criteria such as conservation of the miRNA target sites in homologous genes and local miRNA-mRNA interaction with a positive balance of minimum free energy **[23]**. However, several other features have been experimentally and computationally identified, considering an individual target site level as well a global mRNA level **[24]**.

Currently, there are several programs available, e.g. Diana micro-T **[25]**, miRanda **[26]**, PicTar **[27]**, PITA **[28]**, RNA22 **[29]** and TargetScan **[30].** The several algorithms provide different predictions, and the degree of overlap between them is often poor or null **[31]**. Using GO (The Gene Ontology Consortium, 2000) has become a standard way to validate the functional coherence of genes in a target list. Nevertheless, this type of validation usually requires a statistical analysis to confirm statistical significance.

Additionally, databases such as miRWalk **[32]** and miRTarBase **[33]** have been published. These databases aggregate target predictions from several programs and/or also store experimentally validated targets.

Results and Discussion

It is known that the cell's machinery is designed in order to minimize energy consuming, so why should a gene regulate the expression of a miRNA and being simultaneously his target, usually resulting in its own translational repression?

The existence of such regulatory loops seems to reveal a complex mechanism of genes regulation. Therefore, we were looking for transcription factors regulating the expression of a miRNA and being simultaneously the target protein-coding gene of that same miRNA. **Figure 3** illustrates this.

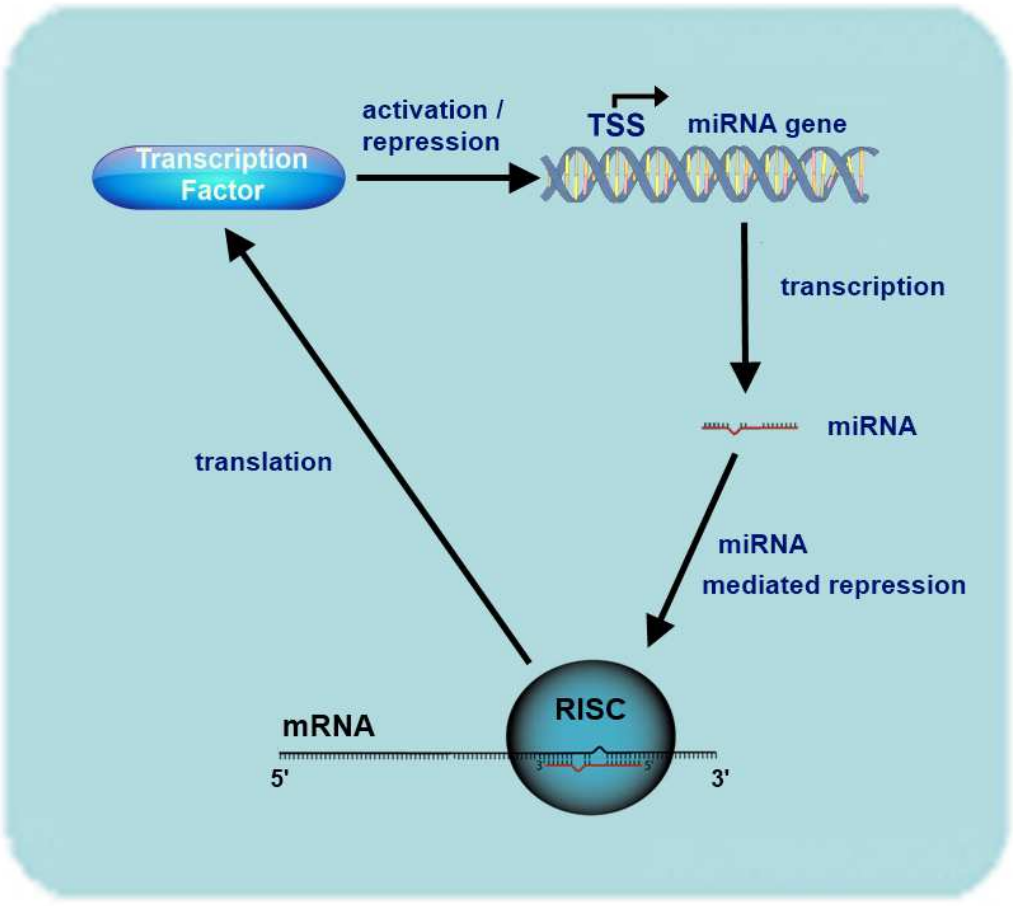


Figure 3. Mixed transcription factor/miRNA regulation loops

After analyzing all selected data (see **Materials and Methods**), we found 38773 of such loops, covering 285 distinct transcription factors and 417 distinct miRNAs. Despite the fact that we cannot yet understand the biological significance of these regulatory loops, their existence seems to be evident and should be experimentally validated.

Since a single miRNA can regulate multiple genes and a single gene can be

regulated by multiple miRNAs, it is quite natural to think that both miRNAs and TFs may cooperate in regulating the same target genes at the transcriptional and post-transcriptional levels. In fact, the co-regulation of transcription factors and microRNAs in transcriptional regulatory networks is a subject that has been investigated by several authors **[34] [35] [36] [37] [38] [39] [40] [41]**.

Clearly, miRNAs cannot independently perform a single task in cells. Instead, miRNAs regulate cellular networks as network components in many cellular functions **[42]**. In fact, TFs and miRNAs function together in gene regulatory networks that are not yet completely identified and understood. Consequently, all loops identified by this investigation should be seen as components of regulatory modules, instead of isolated loops. Although this is true, we can also analyze each one of these individual loops.

A similar loop was found in the developing of Drosophila melanogaster eye **[43]**. Author's investigation revealed that, in nonstimulated cells, Yan represses miR-7 transcription, whereas miR-7 RNA represses Yan protein expression in photoreceptors, by binding to sequences within its mRNA 3' UTR. This mutually inhibitory relationship helps to partition the expression of Yan into eye progenitor cells and that of miR-7 into differentiating photoreceptors, contributing to these two alternative fates. According to the authors' conclusion, this mechanism can explain how signal transduction activity can robustly generate a stable change in gene-expression patterns.

Keeping in mind the fact that this is an *in silico* analysis, we should be aware that the vast majority of all detected loops have a very low probability of being real loops. Therefore, further investigations should start by defining reliability criteria. As demonstrated in the **Materials and Methods** section, prediction of both TFBS and targets varies widely among all tools. To reduce the number of predictions and to try to raise the reliability of predicted results, the usual procedure is to consider only those results that are predicted by several algorithms. From now on, we will briefly analyze some of the results.

Using databases concordance as reliability criteria, the pair hsa-mir-9/NFKB1 is the only loop predicted by all seven TFBS tools and five of the six miRNA targets databases used. However, because the average number of both predicted binding and target sites is very low, this result is not as good as it appears to be. An average of 10.86 binding sites per application and only 3.7 target sites were predicted. Nevertheless, there are several investigations relating NFKB1 with hsa-mir-9.

If NFKB1 is involved in the loops with highest databases concordance, MYB transcription factor is involved in the most loops with the highest target sites average (see

Table 1 for details), considering TFBS and targets predicted by at least three databases. In fact, MYB is involved in twenty one of the first twenty five loops in these conditions.

miRNA	TF	#TFBS Apps	Avg. TFBS	# Targets Apps	Avg. Targets
mir-150	MYB	4	9.25	5	32.33
mir-7	SP1	5	97.00	3	24.00
mir-182	MYB	5	21.00	3	21.00
mir-124	SP1	5	766.80	3	21.00
mir-519a	MYB	4	34.00	3	19.00
mir-7	NF1	4	18.75	3	18.00
mir-497	MYB	4	1.50	3	17.00
mir-424	MYB	5	15.40	3	17.00

Table 1. Regulation loops that have the highest target sites average

At this point, it is important to say that average target sites were calculated using only four databases, because the other two used databases do not indicate the number of target sites. MirTarBase contains experimentally validated targets and mirWalk contains published targets only.

Considering mirTarBase as a reliable source of miRNA targets and selecting only loops with targets predicted by mirTarBase and whose TFBS were predicted by at least 4 tools, we have the 26 regulation loops listed below in **Table 2**.

mirNA	TF	Avg. #TFBS	#TFBS Apps	#Targets Apps
mir-9	NFKB1	10.86	7	5
mir-15a	NFKB1	3.00	7	4
let-7a	NFKB1	8.57	7	3
mir-106a	RUNX1	76.00	6	4
mir-23b	PLAU	2.33	6	4
mir-146a	NFKB1	2.50	6	3
mir-101	FOS	6.80	5	5
mir-429	ZEB1	8.40	5	5
mir-16	MYB	2.40	5	4
mir-122	SRF	16.60	5	4
mir-200b	ZEB1	8.40	5	4
mir-200c	ZEB1	3.20	5	4
mir-218	SP1	32.80	5	3
mir-124	SP1	766.80	5	3
mir-141	ZEB1	3.20	5	3
mir-124	AHR	30.80	5	3
mir-200a	ZEB1	8.40	5	3
mir-27a	SP1	170.80	5	3
mir-612	TP53	16.50	4	5
mir-124	NR3C1	48.00	4	5
mir-150	MYB	9.25	4	5
mir-221	FOS	11.75	4	4

mir-222	FOS	11.75	4	4
mir-101	MYCN	5.00	4	4
mir-103	CREB1	6.25	4	3
mir-29b	SP1	41.25	4	3

Table 2. Regulation loops predicted by mirTarBase and at least four TFBS applications

As we can see in **Table 2**, the average number of TFBS for the hsa-mir-124/SP1 loop is much higher than all other loops. This is because both Patch 1.0 and TFBind predicted hundreds of TFBS in this case. Nevertheless, this is not a unique case. Considering five TFBS tools and at least two miRNA targets databases, SP1 is also involved in regulations loops with several other miRNAs, besides hsa-mir-124, namely hsa-mir-425, hsa-mir-92b, hsa-mir-607, hsa-mir-505, hsa-mir-148a, hsa-mir-345 and hsa-mir-24. All these interactions have in common an average number of predicted TFBs much higher than usual (in this case, greater than 200 binding sites).

In fact, as illustrated in **Table 3**, the total number of TFBS predicted in all loops involving SP1 is incomparably higher than any other transcription factor. The second TF in this list is RUNX1 and has less than half of predicted TFBS when compared with SP1. However the number of predicted loops is almost the same, considering both SP1 and RUNX1.

TF	Total BS	Total loops	Avg. BS
SP1	98363	407	241.68
RUNX1	47194	404	116.82
REL	28645	206	139.05
POU2F1	25739	294	87.55
REPIN1	23023	299	77.00
CREB1	19556	364	53.73
FOS	18966	262	72.39
PAX5	18285	344	53.15
ELK1	15647	287	54.52
TP53	14452	338	42.76

Table 3. Top 10 of TFs by sum of predicted TFBS

Since the number of predicted binding sites is a good indicator for the probability of a TF to regulate the transcription of a miRNA promoter sequence, further investigations should take into consideration the predicted TFBS average. Remarkably, the fifth place of this ranking is occupied by GATA1 that only has 67 predicted loops, each one with an average of 87.04 predicted TFBS. Listing all loops predicted for GATA1, we can observe that there are six miRNAs for which all TFBS tools have predicted exactly the same number of binding sites. These miRNAs are hsa-mir-498, hsa-mir-518c, hsa-mir-520a, hsa-mir-520d, hsa-mir-524 and hsa-mir-525 and they are all mapped to chromosome 19. This is interesting, since it

seems to reveal that GATA1 similarly regulates the transcription of these six miRNAs as if they were members of a regulatory module.

Analyzing now the sum of predicted TFBS per miRNA, we detected a miRNA that is in a similar situation as GATA1. As we can see in **Table 4**, hsa-mir-450a isn't listed at top 10 of miRNAs sorted by sum of predicted TFBS. In fact, hsa-mir-450a is the 47th of this rating. However, when sorted by the average of predicted TFBS, hsa-mir-450a is in the fourth position. This miRNA has only 81 predicted loops, each one with an average of 65.01 predicted TFBS.

miRNA	Total BS	Total loops	Avg. BS
mir-124	24782	215	115.27
mir-365	10587	163	64.95
mir-194	10145	182	55.74
mir-425	9818	147	66.79
mir-182	9703	160	60.64
mir-191	9603	139	69.09
mir-92b	9592	177	54.19
mir-148a	9225	187	49.33
mir-183	8858	151	58.66
mir-96	8806	158	55.73

Table 4. Top 10 of miRNAs by sum of predicted TFBS

A closer look to the hsa-mir-450a predicted loops reveals the reason of this situation: Patch 1.0 predicted 802 TFBS for SP1 transcription factor. On the other end, only one application predicted SP1 as a target of hsa-mir-450a and this prediction only has four target sites. Once all miRNAs with loops sorted by the average of TFBS, we can see that hsa-mir-124 has an average of 115.27 binding sites per loop, which is significantly higher than all the others. Second place is occupied by hsa-mir-191 but only has 69.09 predicted binding sites per loop.

All these predictions rely on several other tools and, as postulated by GIGO (garbage in, garbage out) axiom, if invalid data is entered into a system, the resulting output will also be invalid. Therefore, it is important to start by defining validation criteria for all these predictions. Best validation would be to compare all predictions with experimentally validated targets. However, such datasets are too small to be used as benchmarks.

Nevertheless, we can compare predictions of all databases in order to find differences and similarities. One possible way to do this is by using principal component analysis (PCA). We can also use clustering techniques and compare all resulting clusters. We started by selecting 55 of the most probable miRNAs with loops. For that, we started by computing an overall score for each loop (see **Supplementary Material** for details). This score uses all

scores calculated by each database (when available), the total number of databases with that prediction and also the average number of binding sites and targets.

Using this score we selected the top 55 miRNAs and TFs. In order to be able to compare our results with some of the already known clusters, we added eight additional miRNAs, namely, hsa-miR-17, hsa-miR-18a, hsa-miR-19a, hsa-miR-20a, hsa-miR-15a, hsa-miR-16, hsa-miR-34b and hsa-miR-34c. Subsequently, for each database, we collected all predictions for the 63 miRNAs and TFs with loops. For TFBS predictions we used the number of bind sites and for target predictions we used the number of predicted targets in every loop.

After applying PCA, we can visually analyze how miRNAs are related to each other concerning the TFs that control their transcription (**Figure 4**; see also **Supplementary Material** for details), as predicted by each one of the databases. We can cluster these results, measuring the Euclidean distance of all miRNAs (for example). However, we can also cluster all data used to perform PCA analysis and get a cluster dendrogram as illustrated in **Figure 5** (see also **Supplementary Material** for details).

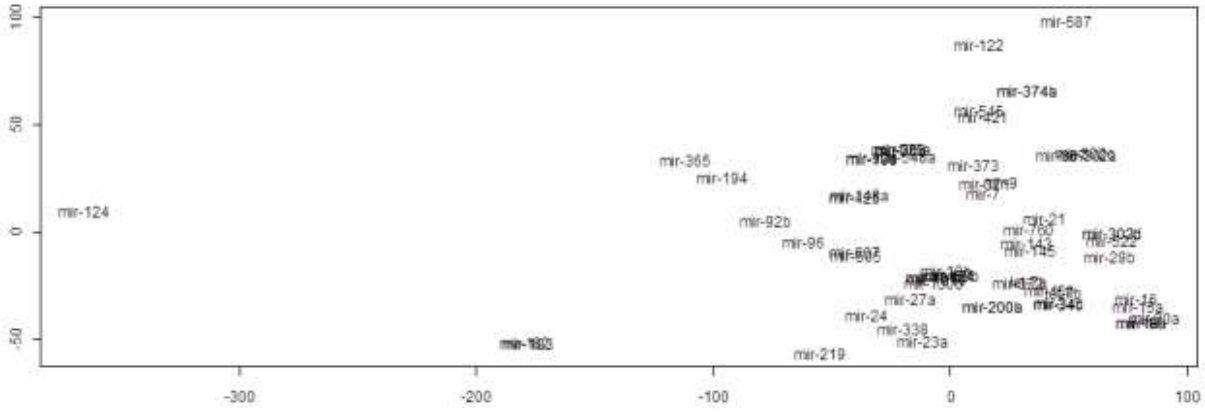


Figure 4. PCA analyses using PROMO 3.0.2 TFBS predictions

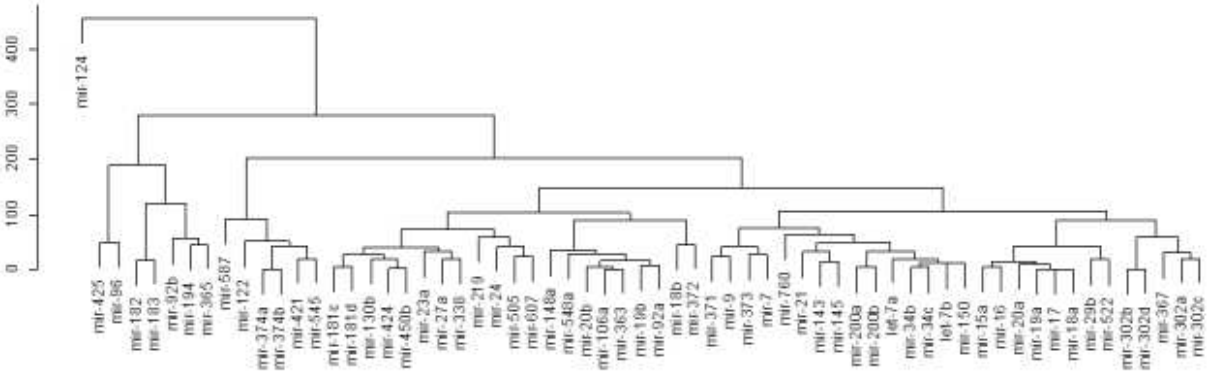


Figure 5. Cluster dendrogram using PROMO 3.0.2 TFBS predictions

Having this additional information, it is possible, for example, to detect unrealistic predictions. After comparing all these graphical information we we're able to identify several differences and similarities among all TFBS predictions. First similarity is that all predictions separate hsa-mir-124 from all other miRNAs. However, TF Search and Patch 1.0 have grouped hsa-mir-124 with hsa-mir-96, hsa-mir-182 and hsa-mir-183. Match 1.0 predictions are slightly different from all others because this is the only cluster dendrogram with two distinct major groups of miRNAs, having hsa-mir-124 grouped in one of them, however separated from all other miRNAs of that group. We must remember that these clusters were obtained using a subset of TFs (top 55).

We used three known clusters in order to validate these predictions:

- mir-15a/mir-16
- mir-34b/mir-34c
- mir-17-92 cluster, which includes mir-17, mir-18a, mir-19a, mir-20a, mir-19b and mir-92a

None of these databases, using top 55 TFs, completely predicted the mir-17-92 cluster. All of them grouped just four miRNAs, namely, mir-17, mir-18a, mir-19a, mir-20a. PROMO 3.0.2 and Patch 1.0 were able to group mir-19b and mir-92a in other cluster. TFBind has clustered these two miRNAs with mir-106a. The mir-15a/mir-16 cluster was predicted by all TFBS databases except Match 1.0 and P-Match 1.0. The mir-34b/mir-34c cluster was predicted by all TFBS databases except P-Match 1.0. We were also able to detect three other groups of miRNAs that were clustered by all TFBS databases. These clusters are:

- mir-302a, mir-302b, mir-302c and mir-302d
- mir-181c and mir-181d
- mir-374a and mir-374b

All, except Match 1.0 have also grouped:

- mir-200a and mir-200b
- mir-23a and mir-27a

Performing this very same analysis for targets databases reveals much more differences than similarities. For example, when comparing Diana micro-T **[25]** with SVM **[44]**, there are five exactly equal clusters:

- mir-181c and mir-181d
- mir-374a and mir-374b
- mir-17 and mir-20a

- mir-130b and mir-148a
- mir-106a and mir-20b

After comparing several clusters with the already known ones, Diana micro-T predictions reveal more accuracy then SVM predictions because it was able to predict the mir-15a/mir-16 cluster. MirTarBase and mirWalk are not comparable with the previous two databases because the first one contains experimentally validated targets and the other contains published targets only. MiRanda predictions are not comparable with any of the previous ones either, because each one of these databases contains a subset of the 55 selected miRNAs. One of them contains conserved miRNAs and the other non-conserved miRNAs. Because of this fact, these databases are not even comparable with each other.

Since TFs and miRNAs function together in gene regulatory networks and some of these networks are partially identified, we can use this knowledge to validate these predictions as well to analyze new findings. Besides the number of target and binding sites and clustering information, a deeper analysis should also include databases scores, when available. For example, miRanda uses mirSVR scores **[46]**. However, different score calculation methods are used among both targets and transcription factors binding sites databases. Since scores are not comparable among different databases, this analysis requires previous normalization or should be done separately for each prediction in every database.

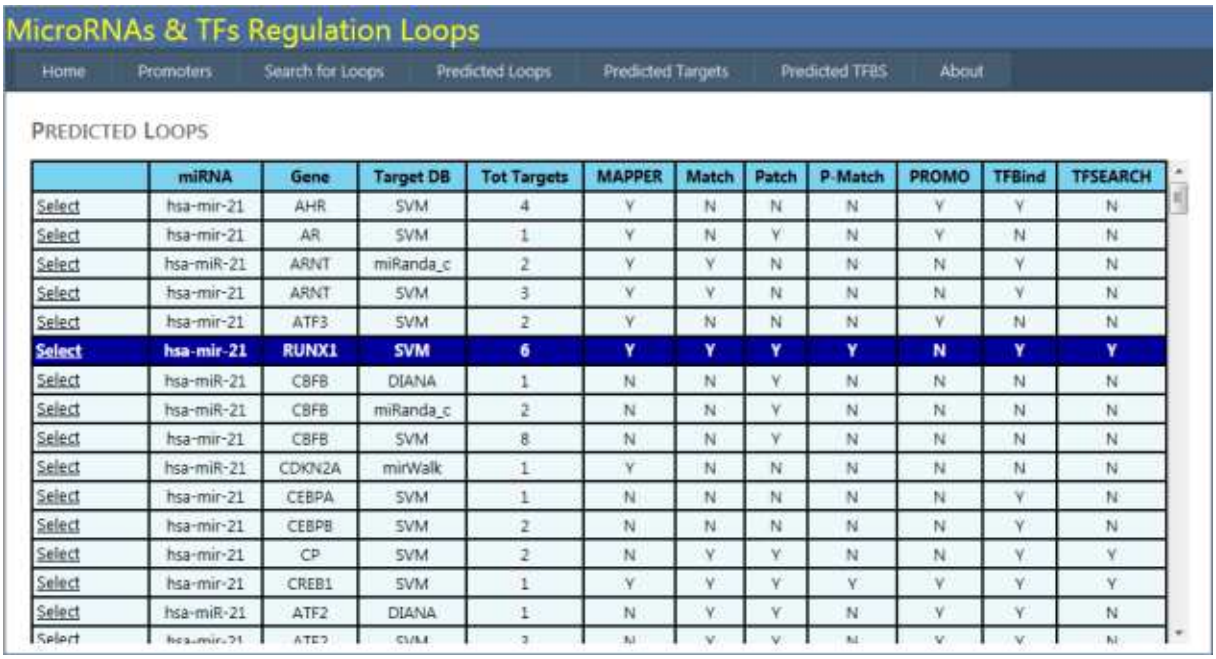
Nevertheless this is a very useful type of analysis because it can reveal loops that are hidden when we just look for the number of target sites. For example, after analyzing all loops of miRanda conserved miRNAs with highest scores we were able to find a loop with a single target site than can be very promising due to its high score. Six of the seven TFBS applications used predicted that FOS regulates hsa-mir-148a transcription and the average number of binding sites is 46.5. However, only SVM and miRanda conserved databases predicted that FOS is a target of hsa-mir-148a. Worse than that, both databases predicted that a single target site and SVM score is very far from being one of highest of that database. However, the mirSVR score is one of the highest for all targets predicted by miRanda and involved in the potential regulation loops detected.

We found 38773 potential regulation loops, most of them predicted by a single database and some others predicted by several databases. To reduce the number of predictions, investigators often consider only those predictions that are common to several databases and assume this overlap as a higher-quality subset of predictions. However, this is not necessarily true. In fact, as indicated by Ritchie W et al. **[45]**, this can be a trap. They

suggest that searching for overlaps between miRNA target prediction algorithms should be discouraged owing to a lack of utility and rationale. For this reason and because we did not want to restrict future investigations, we decided to publish results from all used databases, despite the certainty that the vast majority of these predictions are not real loops.

All these data demand for further investigations and experimental validations. However, the ultimate goal of this investigation was to identify transcription factors regulating miRNA expression and being simultaneously the target protein-coding gene of that same miRNA. As result of this work, we assembled and characterized a catalogue of such mixed transcription factor/miRNA regulation loops in human. All data was stored in a relational database and a web platform was developed in order to enable further investigations (**Figure 6** is a screen shot of this platform).

This database has 36 tables and stores over than 2.5 million records. The web interface allows a search for loops using several search criteria and analyzes all details of every loop such as predicted TFBS and targets, scores of each prediction, etc.



	miRNA	Gene	Target DB	Tot Targets	MAPPER	Match	Patch	P-Match	PROMO	TFBind	TFSEARCH
Select	hsa-mir-21	AHR	SVM	4	Y	N	N	N	Y	Y	N
Select	hsa-mir-21	AR	SVM	1	Y	N	Y	N	Y	N	N
Select	hsa-miR-21	ARNT	miRanda_c	2	Y	Y	N	N	N	Y	N
Select	hsa-mir-21	ARNT	SVM	3	Y	Y	N	N	N	Y	N
Select	hsa-mir-21	ATF3	SVM	2	Y	N	N	N	Y	N	N
Select	**hsa-mir-21**	**RUNX1**	**SVM**	**6**	**Y**	**Y**	**Y**	**Y**	**N**	**Y**	**Y**
Select	hsa-miR-21	CBFB	DIANA	1	N	N	Y	N	N	N	N
Select	hsa-miR-21	CBFB	miRanda_c	2	N	N	Y	N	N	N	N
Select	hsa-mir-21	CBFB	SVM	8	N	N	Y	N	N	N	N
Select	hsa-miR-21	CDKN2A	mirWalk	1	Y	N	N	N	N	N	N
Select	hsa-mir-21	CEBPA	SVM	1	N	N	N	N	N	Y	N
Select	hsa-mir-21	CEBPB	SVM	2	N	N	N	N	N	Y	N
Select	hsa-mir-21	CP	SVM	2	N	Y	Y	N	N	Y	Y
Select	hsa-mir-21	CREB1	SVM	1	Y	Y	Y	Y	Y	Y	Y
Select	hsa-miR-21	ATF2	DIANA	1	N	Y	Y	N	Y	Y	N

Figure 6. Screen shot of regulation loops web platform

Conclusions

Since cell's machinery is designed in order to minimize energy consuming, it would be unlikely for a gene to regulate the expression of a miRNA and being simultaneously his target, usually resulting in its own translational repression at a post-transcriptional level. However, this *in silico* analysis has found 38773 potential loops, covering 285 distinct transcription factors and 417 distinct miRNAs. Some of these loops have a great probability of being experimentally confirmed. Although not being the ultimate goal of this investigation, we also computed a score for each predicted loop. With this or any other scoring system it is possible to guide experimental validations of predicted loops.

Despite the fact that we cannot yet understand the biological significance of these regulatory loops, their existence seems to be evident and this must be an important mechanism of genes regulation. In order to enable further investigations, we developed a web platform through which all data can be analyzed.

Materials and Methods

In 2008, Marson *et al* characterized the DNA sequences of all human miRNA promoters by chromatin-immunoprecipitation **[1]**. Their work provided, among other data and information, a table with human miRNA promoters and associated proteins and genomic features (Table S7). All human coordinate information upon which this investigation was based it was downloaded in January 2005 from the UCSC Genome Browser (hg17, NCBI build 35).

We started from these data and the first thing done was to collect all sequences from the indicated version of UCSC Genome Browser, according to the TSS positions of all 550 promoters. For that purpose it was necessary to write a small program. One of the sequences (hsa-mir-142) was later discarded due to its huge length (406435 nucleotides).

Having all these promoters' sequences, it was then necessary to predict TFBS for all of them. For that, we initially used nine programs, namely AliBaba 2.1, Genomatix MatInspector, Mapper 2, Match 1.0, Patch 1.0, P-Match 1.0, PROMO 3.0.2, TFBind and TFSEARCH 1.3. Each program has its own specificities and it was necessary to deal with that in order to harmonize both inputs and outputs.

Their first difference is the way how promoter sequences can be send to them. MAPPER 2 it was the only one that was able to process a FASTA file containing all promoter sequences. For TFSEARCH 1.3 we were able to download EZRetrieve. This is a free tool that relies on TFSearch and has also processed the complete FASTA file. For TFBind we conceived a tool similar to EZRetrieve. This program reads a FASTA file and sends each sequence to the TFBind tool that is available online. Then saves the HTML outputs that can be seen when we perform the online search.

For all the others, a previous register on the sites where these tools are available was necessary. Therefore, it is necessary to login before starting to use these tools. Because of that, it was not possible to conceive any tool to perform this search automatically. The only solution it was to split our FASTA file into several small files and submit each one of them to each one of these tools.

Having all these huge amount of data, it was then necessary to prepare it to be analyzed. AliBaba 2.1 results were discarded because of output complexity and outdated version of TRANSFAC. Genomatix MatInspector results were also discarded because they use matrices of their own and it is not a free software tool.

Consequently, it was necessary to analyze outputs from seven programs. EZRetrieve produced a table indicating the number of binding sites for each pair of predicted transcription factor and miRNA promoter sequence given to it as input. Since the number of binding sites is a good indicator for the probability of a TF to regulate the transcription of a miRNA promoter sequence, we decided to write a tool to parse all output files of each prediction program in order to count all binding sites for each pair transcription factor/miRNA promoter. **Figure 7** is a screen shot of this tool.

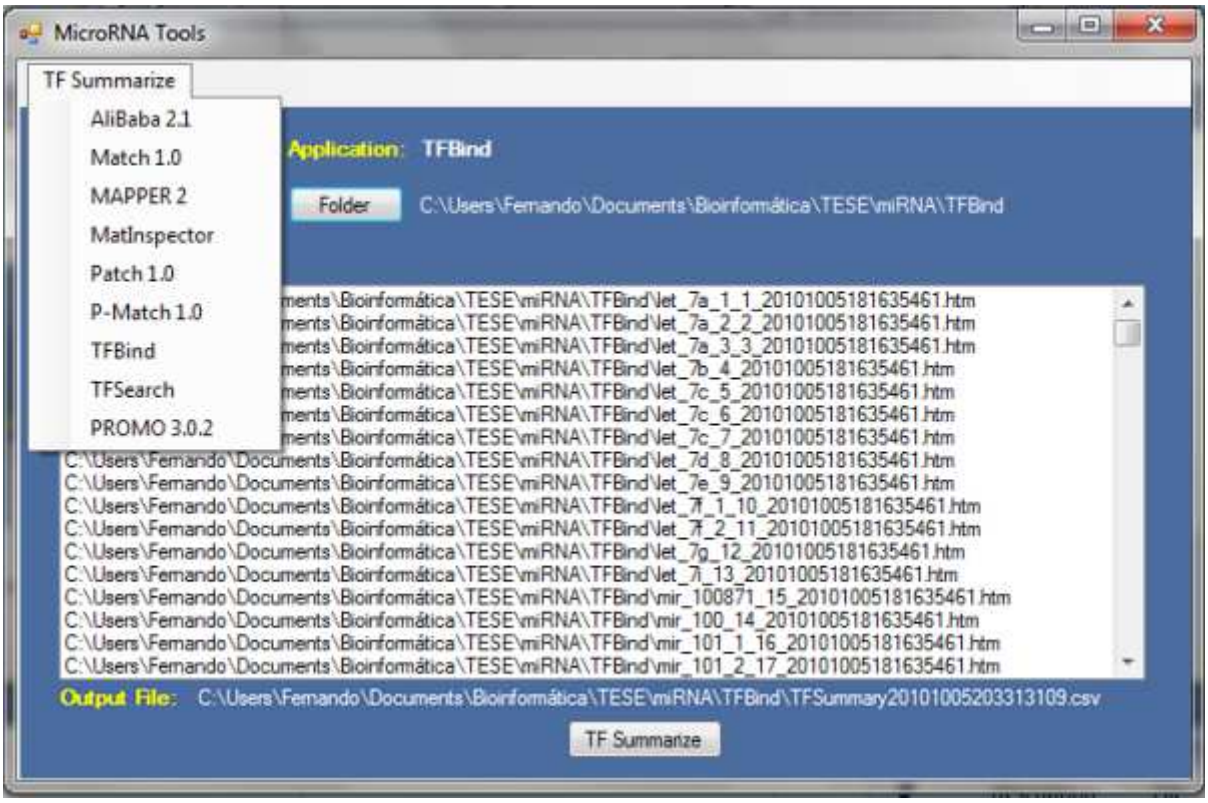


Figure 7. Screen shot of TF summarize tool

Besides the specificities of each output, this tool had to deal with the fact the we were only interested in results from Homo Sapiens (HS) and some programs gave us more than that. Thus, when not indicated in the output result, it was necessary to test each matrix against matrices databases in order to verify if we were in the presence of a human matrix or not. Same verification was performed with gene symbols, when necessary. Applied these filters and totalized all binding sites, it became obvious that there are significant differences among all prediction programs (**Table 5**; see also **Supplementary Material** for details).

Application	TFBS HS Genes	TFBS Other Genes	Total TFBS	Pct.
TFSEARCH 1.3	22064	19084	41148	2.40
MAPPER 2	80318	96734	177052	10.33
P-Match 1.0	103935	6638	110573	6.45
Match 1.0	138336	54381	192717	11.25
PROMO 3.0.2	111377	108704	220081	12.85
Patch 1.0	185326	106627	291953	17.04
TFBind	502931	176858	679789	39.68

Table 5. Total of binding sites per application

After joining data from all seven TFBS prediction tools we found 64701 distinct pairs of human TF/miRNA promoters. However the vast majority (75.47%) of all TFBS human predictions is predicted by only 1 or 2 applications. The way these pairs are distributed by the number of applications that have simultaneously predicted them is illustrated in **Table 6** (see also **Supplementary Material** for details).

# applications	TF/miRNA pairs	Pct.
1	36461	56.35
2	12370	19.12
3	8012	12.38
4	3830	5.92
5	2842	4.39
6	902	1.39
7	284	0.44

Table 6. Pairs of TF/miRNA promoters simultaneously predicted by TFBS prediction applications

Another issue related with these predictions is genes names. These outputs usually indicate a gene name and the identification of which matrix was used to get each prediction. However, genes names are not always compatible among the several databases, because most genes have more than one name. This is illustrated in the next table with some examples obtained in NCBI online database.

Gene	Also known as
FOS	AP-1; C-FOS
HOXD10	HOX4; HOX4D; HOX4E; Hox-4.4
MYB	fg; Cmyb; c-myb; c-myb_CDS
MYC	MRTL; c-Myc; bHLHe39
NFKB1	p50; KBF1; p105; EBP-1; MGC54151; NFKB-p50; NF-kappaB; NFKB-p105; NF-kappa-B; DKFZp686C01211
TP53	P53; LFS1; TRP53; FLJ92943
LHX1	LIM1; LIM-1; MGC126723; MGC138141
MYCN	NMYC; ODED; MODED; N-myc; bHLHe37
RELA	p65; NFKB3; MGC131774
RUNX1	AML1; CBFA2; EVI-1; AMLCR1; PEBP2aB; AML1-EVI-1

Table 7. Different gene names

We downloaded all Homo Sapiens genes registered in the NIH genetic sequence database GenBank from the NCBI site. Besides the official gene symbols, this file also contains their synonyms or alias. After comparing gene names, we were able to identify most of the genes listed in the outputs of the TFBS applications. However, many of them remain unclassified or are not Homo Sapiens genes. This is illustrated in **Table 8** (see also **Supplementary Material** for details).

Application	HS Genes	Other Genes
TFSEARCH 1.3	61	49
MAPPER 2	232	357
P-Match 1.0	33	16
Match 1.0	83	46
PROMO 3.0.2	58	32
Patch 1.0	139	164
TFBind	89	47

Table 8. Total number of different genes predicted by TFBS prediction applications

The next step was the prediction of miRNA targets. For this purpose, we started by using miRWalk target published predictions. A file with all miRNAs names was sent to mirWalk and this application returned a total of 7307 targets representing 2654 different genes. All these gene names were compared with Homo Sapiens genes predicted by all TFBS applications and with their synonyms as well.

After comparing mirWalk gene names, we found 163 genes with the same name as the names of transcription factors predicted by TFBS tools. These 163 distinct genes, according to mirWalk predictions, are targets of 102 distinct miRNAs. Loops were found for 82 distinct transcription factors, covering 85 distinct miRNAs (**Table 9**; see also **Supplementary Material** for details). It is important to remember that our miRNAs list is a subset of all miRNAs, since we are analyzing the sequences from Marson *et al* **[1]**. In fact, all targets databases also predicted targets related to other miRNAs.

Application	TFs	miRNAs	loops
TFSEARCH 1.3	20	24	43
MAPPER 2	53	49	140
P-Match 1.0	9	31	39
Match 1.0	27	45	92
PROMO 3.0.2	29	50	122
Patch 1.0	42	50	124
TFBind	45	73	186

Table 9. Distinct TFs/miRNAs with loops, using mirWalk

The next database used was Diana micro-T v3.0. This database has about 2.5 million records and targets are identified by Ensemble ID. Since GenBank also contains Ensemble

IDs, we wrote a program in order to extract from Diana database all records in which the target gene is one of the genes predicted by the TFBS applications. After comparing gene ID's, we found 279 genes with the same Ensemble ID as the ones of transcription factors predicted by TFBS tools. According to Diana micro-T predictions, these 279 distinct genes are targets of 417 distinct miRNAs. Loops were found for 259 distinct transcription factors, covering 346 distinct miRNAs (**Table 10** ; see also **Supplementary Material** for details).

Application	TFs	miRNAs	loops
TFSEARCH 1.3	51	233	682
MAPPER 2	201	324	3571
P-Match 1.0	31	306	1133
Match 1.0	74	327	2278
PROMO 3.0.2	57	319	1765
Patch 1.0	116	338	3513
TFBind	84	344	4313

Table 10. Distinct TFs/miRNAs with loops, using Diana

We also analyzed miRanda databases. There are four of them, combining good and non-good mirSVR scores with conserved and non-conserved miRNAs. However, we only analyzed good mirSVR scores databases. In these databases genes are identified by GeneBank ID (NCBI Entrez ID) and we started by writing a program in order to extract from these databases all records in which the target gene is one of the genes predicted by the TFBS applications.

The one with non-conserved miRNAs has about 3.3 million targets and, after comparing gene ID's, we found 288 genes with the same gene ID as the ones of transcription factors predicted by TFBS tools. According to this database, these 288 distinct genes are targets of 336 distinct miRNAs. Loops were found for 252 distinct transcription factors, covering 144 distinct miRNAs (**Table 11** ; see also **Supplementary Material** for details).

Application	TFs	miRNAs	loops
TFSEARCH 1.3	38	83	241
MAPPER 2	184	132	1063
P-Match 1.0	28	116	367
Match 1.0	70	123	737
PROMO 3.0.2	56	136	714
Patch 1.0	112	137	1159
TFBind	86	143	1587

Table 11. Distinct TFs/miRNAs with loops, using miRanda non-conserved miRNAs

The miRanda database with conserved miRNAs has about one million targets and, after comparing gene ID's, we found 284 genes with the same gene ID as the ones of

transcription factors predicted by TFBS tools. According to this database, these 284 distinct genes are targets of 228 distinct miRNAs. Loops were found for 259 distinct transcription factors, covering 193 distinct miRNAs (**Table 12**; see also **Supplementary Material** for details).

Application	TFs	miRNAs	loops
TFSEARCH 1.3	47	150	542
MAPPER 2	193	187	2424
P-Match 1.0	29	175	675
Match 1.0	74	186	1460
PROMO 3.0.2	57	190	1360
Patch 1.0	121	189	2269
TFBind	86	192	2870

Table 12. Distinct TFs/miRNAs with loops, using miRanda conserved miRNAs

This analysis was also performed using mirTarBase, a database with experimentally validated targets. As expected, numbers are much lower. After comparing gene names, we found 90 genes with the same name as the names of transcription factors predicted by TFBS tools. According to mirTarBase, these 90 distinct genes are targets of 93 distinct miRNAs. Loops were found for 58 distinct transcription factors, covering 70 distinct miRNAs (**Table 13** ; see also **Supplementary Material** for details).

Application	TFs	miRNAs	loops
TFSEARCH 1.3	11	19	25
MAPPER 2	27	36	49
P-Match 1.0	6	15	15
Match 1.0	17	31	37
PROMO 3.0.2	21	35	50
Patch 1.0	27	39	62
TFBind	32	51	73

Table 13. Distinct TFs/miRNAs with loops, using mirTarBase

Very recently, Saito T and Sætrom P have also published a database with miRNAs targets **[44]**. This database has almost 20 million target sites and was created using a two-step Support Vector Machines (SVM). We decided to incorporate this database in our analysis and, as expected, numbers are much higher. Similarly to the previous database we started by writing a program in order to extract from these databases all records in which the target gene is one of the genes predicted by the TFBS applications.

After comparing gene names, we found 283 genes with the same name as the names of transcription factors predicted by TFBS tools. These 283 distinct genes, according to these predictions, are targets of 412 distinct miRNAs. Loops were found for 278 transcription factors, covering all 412 miRNAs (**Table 14**; see also **Supplementary Material** for details).

Application	TFs	miRNAs	loops
TFSEARCH 1.3	59	386	2684
MAPPER 2	218	412	15461
P-Match 1.0	32	411	4513
Match 1.0	78	411	9082
PROMO 3.0.2	57	411	8137
Patch 1.0	131	411	14305
TFBind	84	411	18191

Table 14. Distinct TFs/miRNAs with loops, using SVM

As demonstrated by previous tables, prediction of both TFBS and targets varies widely among all tools.

Supporting Information

Supplementary file 1: Marson_Cell08_S7.xlsx
Human miRNA promoters and associated TSS positions.

Supplementary file 2: mirna_prom.fasta
Sequences of miRNA promoters that have been collected from the UCSC Genome Browser (hg17, NCBI build 35), taking into consideration chromosomes and TSS positions indicated in supplementary file 1.

Supplementary file 3: HSGenes.xlsx
All identified genes, aliases and matrices.

Supplementary file 4: LOOPS01_stats.xlsx
All predictions about TFs regulating miRNA expression and being simultaneously the target protein-coding gene of that same miRNA. This file also contains scores and some statistics for each loop.

Supplementary file 5: LOOPS02_by_targetsDB.xlsx
All predicted loops organized by targets database.

Supplementary file 6: TFBS01_TFSEARCH.xlsx
TFBS predicted by TFSEARCH.

Supplementary file 7: TFBS02_MAPPER.xlsx
TFBS predicted by MAPPER 2.

Supplementary file 8: TFBS03_TFBIND.xlsx
TFBS predicted by TFBind.

Supplementary file 9: TFBS04_MATCH.xlsx
TFBS predicted by Match 1.0.

Supplementary file 10: TFBS05_PMATCH.xlsx
TFBS predicted by P-Match 1.0.

Supplementary file 11: TFBS06_PATCH.xlsx
TFBS predicted by Patch 1.0.

Supplementary file 12: TFBS07_PROMO.xlsx
TFBS predicted by PROMO 3.0.2.

Supplementary file 13: TARGETS01_mirWalk.xlsx
Targets predicted by mirWalk for all identified genes.

Supplementary file 14: TARGETS02_mirTarBase.xlsx
Targets validated by mirTarBase for all identified genes.

Supplementary file 15: TARGETS03_svm.xlsx
Targets predicted by SVM for all identified genes.

Supplementary file 16: TARGETS04_DianaMicroT.xlsx
Targets predicted by Diana micro-T v3.0 for all identified genes.

Supplementary file 17: TARGETS05_miRanda_cons.xlsx
Targets predicted by miRanda conserved miRNAs for all identified genes.

Supplementary file 18: TARGETS06_miRanda_nonc.xlsx
Targets predicted by miRanda non-conserved miRNAs for all identified genes.

Supplementary file 19: PCA_clusters.xlsx
Cluster dendrograms and graphics from PCA analysis.

References

[1] Marson et al., 2008, Cell, 134, 521-533 [PMID: 18692474]

[2] Lee RC, Feinbaum RL, Ambros V, 1993, Cell, 75, 843-854 [PMID: 8252621]

[3] Ambros et al., 2004, Methods Mol Biol, 265, 131–158 [PMID: 15103073]

[4] K Chen and N Rajewsky, 2007, Nature Reviews, 8:93 [PMID: 17230196]

[5] Cai X, Hagedorn CH, Cullen BR, 2004, RNA 10 (12): 1957–66 [PMID: 15525708]

[6] Murchison E, Hannon G, 2004, Current opinion in cell biology 16 (3): 223–229 [PMID: 15145345]

[7] Schwarz DS et al., 2003, Cell. Oct 17;115(2):199-208 [PMID: 14567917]

[8] Okamura K et al, 2008, Cell Cycle. Sep 15;7(18):2840-5 [PMID: 18769156]

[9] Grabe N, 2002, In Silico Biology, 2(1):S1-15 [PMID: 11808873]

[10] Yutaka Akiyama, 1995, Kyoto University, http://www.cbrc.jp/research/db/TFSEARCH.html

[11] Cartharius K et al., 2005, Bioinformatics, Jul 1;21(13):2933-42 [PMID: 15860560]

[12] Marinescu VD, Kohane IS, Riva A, 2005, BMC Bioinformatics, 6:79 [PMID: 15799782]

[13] Kel AE et al., 2003, Nucleic Acids Res., 31(13):3576-9 [PMID: 12824369]

[14] Chekmenev DS, Haid C, Kel AE, 2005, Nucleic Acids Res., 33(Web Server issue):W432-7 [PMID: 15980505]

[15] Farré D et al., 2003, Nucleic Acids Res., 31(13):3651-3 [PMID: 12824386]

[16] Tsunoda T, Takagi T, 1999, Bioinformatics, 15(7-8):622-30 [PMID: 10487870]

[17] Berg OG, von Hippel PH, 1987, J Mol Biol, 193(4):723-50 [PMID: 3612791]

[18] Fields DS et al, 1997, J Mol Biol., 271(2):178-94 [PMID: 9268651]

[19] Heinemeyer T et al., 1998, Nucleic Acids Res., 26(1):362-7 [PMID: 9399875]

[20] Sandelin A et al, 2004, Nucleic Acids Res., 32(Database issue):D91-4 [PMID: 14681366]

[21] Lai EC, 2002, Nature Genetics 30, 363 - 364 [PMID: 11896390]

[22] Rajewsky N, 2006, Nature Genetics 38, Suppl:S8-13 [PMID: 16736023]

[23] Barbato C at al., 2009, J Biomed Biotechnol., 2009:803069 [PMID: 19551154]

[24] Saito T, Sætrom P, 2010, BMC Bioinformatics, 11:612 [PMID: 21194446]

[25] Maragkakis M et al., 2009, Nucleic Acids Res., 37(Web Server issue):W273-6 [PMID: 19406924]

[26] John B e tal., 2004, PLoS Biol., 2(11):e363 [PMID: 15502875]

[27] Krek A et al., 2005, Nature Genetics, 37(5):495-500 [PMID: 15806104]

[28] Kertesz M et al., 2007, Nature Genetics, 39(10):1278-84 [PMID: 17893677]

[29] Miranda KC et al., 2006, Cell, 22, 126(6):1203-17 [PMID: 16990141]

[30] Lewis BP, Burge CB, Bartel DP, 2005, Cell, 120(1):15-20 [PMID: 15652477]

[31] Sethupathy P, 2006, Nature Methods, 3(11):881-6. [PMID: 17060911]

[32] Dweep H et al., 2011, Journal of Biomedical Informatics, [PMID: 21605702]

[33] Hsu SD et al., 2011, Nucleic Acids Res., 39(Database issue):D163-9 [PMID: 21071411]

[34] Chen CY et al., 2011, BMC Bioinformatics, Feb 15;12 Suppl 1:S41 [PMID: 21342573]

[35] Flynt AS, Lai EC, 2008, Nat Rev Genet. Nov;9(11):831-42 [PMID: 18852696]

[36] El Baroudi M et al., 2011, PLoS One 3;6(3):e14742 [PMID: 21390222]

[37] Shalgi R et al., 2007, PLoS Comput Biol. Jul;3(7):e131 [PMID: 17630826]

[38] Su N et al., 2010, BMC Syst Biol. Nov 8;4:150 [PMID: 21059252]

[39] Tsang J et al., 2007, Mol Cell. Jun 8;26(5):753-67 [PMID: 17560377]

[40] Qiu C et al., 2010, BMC Syst Biol. Jun 29;4:90 [PMID: 20584335]

[41] Tran DH et al., 2010, Bioinformation. Feb 28;4(8):371-7 [PMID: 20975901]

[42] Yichen Li et al., 2011, Plos One, 6(10): e26302 [doi:10.1371/journal.pone.0026302]

[43] Li X, Carthew RW, 2005, Cell Dec 29;123(7):1267-77 [PMID: 16377567]

[44] Saito T, Sætrom P, 2010, BMC Bioinformatics, 11:612 [PMID: 21194446]

[45] Ritchie W et al., 2009, Nature Methods, 6(6):397-8 [PMID: 19478799]

[46] Betel D et al., 2010, Genome Biol., 11(8):R90 [PMID: 20799968]