The PicTar Sequence Scoring Algorithm

Translational repression by microRNAs involves synergistic and combinatorial effects: several different microRNAs can act together to repress the same gene, while one microRNA often has multiple target sites within one 3'UTR and targets several different genes. This obviously resembles the known mechanism of transcriptional regulation, which was successfully modeled by a hidden Markov model (HMM)[1], and it inspired us to design the algorithm that scores the statistical likelihood for each 3'UTR sequence to be targeted for repression by a given set of microRNAs $\mu_i, i = 1$ to M (see Fig.1b in the main text).

Hidden Markov models can be thought of as sequence-generating models. A random sequence of states is generated by a set of independent "drawings", governed by the transition probabilities between the states. At each state, a certain nucleotide is chosen ("emitted") according to emission probabilities of this particular state. This way, any sequence can be generated with a certain probability. Given the particular sequence S, one can calculate the probability $P(S)$ that S was generated by the specific model (defined by the transition and emission probabilities). Another important task can be to determine the parameters of the model that will maximize the probability that S was generated (this procedure is usually referred to as Expectation Maximization).

Our goal was to predict the efficacy of translational repression of a gene by a given set of microRNAs. Our algorithm follows the general logic of Ahab [1], an experimentaly validated method for detection of transcriptional regulatory modules. We modeled the 3'UTR of a gene as generated by an HMM whose states are binding sites of each of the microRNAs from the set and the background. The microRNA binding sites are represented by the 7-8 nucleotides long "nuclei" - mRNA stretches basepaired with the 5' end of the microRNA. Different microRNAs, or, more precisely, their nuclei compete with each other and background for binding (it was experimentally shown [2] that two overlapping nuclei can not act cooperatively). The final score of a sequence is the log ratio of the probability of the sequence being generated by this model, versus the probability that it was generated by background process alone, after the optimal parameters were found using the Expectation Maximization procedure. The sequence is generated in the following way: at each step one of the states of the system is chosen with prior probability (the equivalent of transition probabilities) ρ_i ($i = 0$ to M, where ρ_0 is the prior probability for the background, and M is the total number of different microRNAs whose combinatorial effect we are assessing). Depending on the nature of the state, a certain sequence will be emitted: in the background state, one nucleotide will be emitted, and in microRNA target site state the 7-mer or 8-mer will be emitted. The emitted sequence is appended to the sequence previously generated, and the process proceeds to the next step until the required 3'UTR length is reached. The sequence of states chosen from positions 1 to L is called the "parse". Obviously, the same sequence of nucleotides can be created by different parses. Background is modeled with the Markov model of order 0 (Methods).

Given the set θ of microRNAs μ_i and their prior probabilities of binding ρ_i, $i = 1$ to M, the probability for the particular sequence to be generated is the sum of probabilities of all possible "parses" of this sequence:

$$P(S \mid \theta) = \sum_{T} P(S \mid \theta, T)$$

Here the probability of a particular parse T, $P(S \mid \theta, T)$, is simply a product of transition probabilities for each of the states in this particular parse and the emission probabilities of subsequences present in those states (N = total number of states in the parse):

$$P(S \mid \theta, T) = \prod_{i=1}^{N(T)} \rho_i \cdot m_i(s)$$

The emission probabilities $m(s)$ for the microRNA binding sites (i.e. the probabilities that a certain mRNA subsequence will be the microRNA binding site) were modeled to best fit the experimental data, as described in Methods. Only a small set of possible sites has a non-zero probability: among them the perfectly base-paired nuclei, and the "imperfect" nuclei with at most one insertion or mutation (compared to the perfect ones) that does not lead to the wobble base-pair, and whose free energy of binding is not higher than that of a perfect nucleus (it was shown that the free energy of binding of the nucleus correlates with the efficacy of the translational repression [2]). Among these allowed binding sites, the perfect nuclei are assigned the probability p (we used $p = 0.8$), while the imperfect nuclei get the probability $(1-p)$ divided by the total number of allowed imperfect nuclei. Another condition the binding site needs to satisfy in order to be assigned a non-zero probability is that the free energy of the entire microRNA:mRNA duplex is below a certain cutoff value (see Methods). This choice of emission probabilities reflects the correlation between the binding free energy of the nucleus and the contribution to translational repression of a particular site [2], and it realistically models the fact that "imperfect" binding sites can contribute to translational repression if there is at least one strong (perfect) binding site present in the sequence. Our current ways of assigning probabilities to binding sites are almost certainly incomplete. However, as more target sites are experimentally validated, we will be able to improve the criteria that we currently use in this scheme.

The set of optimal prior probabilities is determined by maximizing the probability $P(S)$ to observe the sequence we are scoring. Since the prior probabilities in principle represent the probability that a given microRNA is present in the cell (and sequence emission probabilities in the target site state represent the probability that microRNA will actually bind to this particular subsequence), we can say that by finding the priors that maximize the likelihood to observe our sequence, we are actually finding the relative concentrations of different microRNAs in the cell that will maximize the probability that this gene is translationally repressed. The optimization is performed using the Baum-Welch algorithm for Expectation Maximization, as described in [3] and [4]. This procedure is performed because we in principle do not have sufficient information about concentrations of all messenger RNA and microRNA molecules present in the cell. One can imagine, however, that this will change in the future (e.g. availability of more microarray data), and it can be in a natural way incorporated into the model.

Since the sequence is modeled by the HMM, consecutive steps while generating it are independent, which means that one can use the forward equation to calculate $P(S \mid \theta)$ in time proportional

to the length of the sequence. Posterior decoding is used to calculate the probability for each of the positions in the sequence to be bound by a microRNA.

The advantage of the HMM treatment of the microRNA target prediction problem is that it naturally captures many characteristics of microRNA:mRNA interaction that experiments seem to show: coordinate action of several microRNAs, amplification of repression efficacy in case of multiple binding sites, competition of microRNAs in case of overlapping binding sites etc. In case of multiple binding sites (either from one or from several microRNAs) the PicTar score will reflect synergistic action: it will be higher than a simple sum of scores for single sites. The length of the sequence influences the score as well: the longer UTR containing n binding sites is statistically less significant than the shorter one with the same number of sites. Beside the features of translational repression that we described in connection to our model, different factors in the structure of microRNA binding sites have begun to emerge that can affect the repression efficacy [5, 6]. In addition, there is a possibility that the process can be affected by the correlations in the positions of the target sites relative to each other [2, 5]. It was already shown in case of the transcriptional regulation model described in [3] that these correlations can be incorporated in the HMM.

Our knowledge of details of the mechanism of microRNA mediated translational repression is still quite limited, but this is expected to begin changing in the near future. The PicTar method for evaluating the efficacy of translational repression of a gene is based on a clean and intuitive mathematical model that is flexible enough to incorporate future insights into this topic.

References

[1] Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E.D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* **3**, 30 (2002).

[2] Doench, J.G. & Sharp, P.A. Specificity of microRNA target selection in translational repression. *Genes Dev.* **18**, 504–511 (2004).

[3] Sinha, S., van Nimwegen, E. & Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics* **19**, i292–301 (2003).

[4] Durbin, R., Eddy S., Krogh A. and Mitchinson, G., Markov chains and hidden Markov models. in *Biological Sequence Analysis.* 46 - 79 (Cambridge University Press, Cambridge, 1998).

[5] Vella, M.C., Reinert, K. & Slack, F.J. Architecture of a validated microRNA::target interaction *Chemistry & Biology* **11**, 1619–1623 (2004).

[6] Kiriakidou, M. et al. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev* **18** 1165–1178 (2004).