User Guide

Cupid v1.0

Baylor College of Medicine

Texas Children's Hospital

Last update: 8/21/2014 4:36:41 PM

Table of Contents

Chapter I. Outline

Cupid is an integrative framework for the context-specific inference of miRNA targets. It integrates sequence-based evidence and functional clues derived from RNA and miRNA expression analysis, predicting candidate miRNA binding sites and associated target genes using ensemble machine learning classifiers that are trained on validated interactions. Candidate interactions emerging from this step are then refined based on independent, context specific clues, including their predicted ability to mediate competitive endogenous RNA (ceRNA) interactions, where mRNA compete for shared miRNA regulators. Thus, Cupid simultaneously infers both interaction types (ceRNA and miRNA-target interactions). In addition, Cupid considers evidence for combinatorial regulation by multiple miRNA species and for indirect miRNA regulation through effector proteins. Specifically, Cupid is implemented in three sequential steps:

First, Cupid re-evaluates candidate miRNA binding sites in 3' UTRs, as inferred by TargetScan [1], miRanda [2] and PITA [3]. This is accomplished by integrating features including their algorithm-specific scores, their location in the 3' UTR, and their cross-species conservation. The result is a context-independent binding-site prediction; interactions are then integrated using a support vector machine (SVM) algorithm [4] implemented as LIBSVM; you will need to download LIBSVM in order to implement Cupid Step I.

Then, in Cupid Step II, miRNA-target interactions are predicted by further integrating information about selected sites, their multiplicity, and the statistical dependency between the expression profiles of miRNA and putative targets. Likelihoods for each predictive feature are computed based on a positive gold standard set of experimentally confirmed miRNA-target interactions, representing binding sites in databases including TarBase [5], TRANSFAC [6] and miRecords [7]; curated confirmed miRNA-target interactions described in Chiu et. al. (Genome Res., submitted 2014) are provided in this packages. Interactions are then integrated using a support vector machine (SVM) algorithm [4] implemented as LIBSVM; you will need to download LIBSVM in order to implement Cupid Step II. The only context-specific input at this step is statistical dependency between the miRNA and target expression profiles, which is computed as

normalized mutual information (NMI) and Spearman's correlation coefficient (for sign); the script to compute normalized mutual information is provided here.

Finally, Cupid assesses whether inferred targets compete for their predicted miRNA regulators by predicting miRNA mediated mRNA-mRNA interactions.

In addition to these three steps, we provide functions to evaluate other evidence for context-specific regulation, including evidence for combinatorial regulation by multiple miRNA species, and evidence for indirect regulation through effectors. In the following sections, we describe each of Cupid's execution steps, including input files and predictions based on gene expression profiles taken from TCGA breast cancer samples [8], as described in Chiu et. al., Genome Res., (submitted 2014). First, we describe data files and MATLAB programs included in the package. Architecture-specific decisions and scripts will need to be designed to take advantage of the provided files.

Chapter II. Provided Data Files

Below we list all files provided with the package.

Data directory

- **3PrimeUTR_20491transcripts_18093genes.txt**: 3' UTRs of RefSeq transcripts; these were used to predict miRNA sites and interactions. Each 3' UTR description includes an official gene symbol, RefSeq ID, 3' UTR length, and 3' UTR sequence (5'->3').
- **tablePos_1481sites.txt**: sites associated with curated gold standard interactions. This flat file contains a scored table, of predicted sites associated with gold standard interactions, after integrating scores and data used as input for Cupid Step I site prediction. Table columns include
    - AvgProb[0,1] – probability of binding according to SVM testing, averaged across 1000 training/testing runs.
    - AvgBin[0,1] – frequency of inclusion in the predicted sites set in 1000 training-testing runs.
    - GeneSymbol – official name of the candidate miRNA target
    - RefSeqID – RefSeq ID of the candidate miRNA target
    - miRBaseID – miRBase ID of the candidate regulator miRNA
    - SitePos(7mer) – site position in the target 3' UTR
    - 3'UTRLength – length of the target 3' UTR
    - SiteSeq(5'->3') – 7-base sequence of the target site
    - F1:RelDist(from 5') – relative site distance from the 3' UTR start site (position divided by 3' UTR length)
    - F2:RelDist(from 3') – relative site distance from the 3' UTR end site (position divided by 3' UTR length)
    - F3:miRanda[0,1] – normalized site score according to miRanda
    - F4:PITA[0,1] – normalized site score according to PITA
    - F5:TargetScan[0,1] – normalized site score according to TargetScan
    - F6:Conservation[0,1] – normalized cross-species score according to PhastCons [9]

- **tableNeg_36986648sites.txt**: predicted sites, not associated with gold standard interactions. Format follows above description.
- **tablePos_588pairs.txt**: curated gold-standard interactions. This flat file contains a scored table of curated gold-standard interactions, after integrating scores and data used as input for Cupid Step II interaction prediction. Table columns include
    - AvgProb[0,1] – probability of an interaction according to SVM testing, averaged across 1000 training/testing runs.
    - AvgBin[0,1] – frequency of inclusion in the predicted interactions set in 1000 training-testing runs.
    - GeneSymbol – official name of the candidate miRNA target
    - RefSeqID – RefSeq ID of the candidate miRNA target
    - miRBaseID – miRBase ID of the candidate regulator miRNA
    - NumSite – number of tested binding sites
    - SiteScore – a ';'-separated list of site scores given by Cupid Step I probabilities
    - F01:NormMI – signed normalized mutual information between miRNA and target. The sign is from Spearman's correlation coefficient.
    - F02:Max – maximum site score
    - F03:Med – median site score
    - F04:MidRge – medium range site score (max+min)/2
    - F05:Sum – sum of site scores
    - F06:Prod – product of site scores
    - F07:Avg – average of site scores
    - F08:GeoMean – geometric mean of site scores
    - F09:HarMean – harmonic mean of site scores
    - F10:RMS – root mean of site scores
    - F11:WtdMean – weighted mean of site scores, where weights are proportional to the minimum distance from start and end of the 3' UTR
    - F12:SumSq – sum of squares of site scores
    - F13:SumLog – sum of natural logs of site scores
    - F14:SumExp – sum of natural exponents of site scores

  - F15:AvgSq – average of site-score squares
  - F16:AvgLog – average of the natural logs of site scores
  - F17:AvgExp – average of the natural exponents of site scores
  - F18:NumSite – number of tested binding sites
  - F19:SiteWdt – the genomic distance from the most upstream to the most downstream site
  - F20:MinDist – the genomic distance between the closest sites
  - F21:MaxDist – the genomic distance between the furthest adjacent sites
  - F22:AvgDist – the average distance between adjacent sites
- **tableNeg_11542856pairs.txt**: predicted interactions, not associated with gold standard interactions. Format follows above description.

Example directory

The example directory contains example input and output for main functions described in the following section. The directory includes the following files.

For the function **cupidcerna.m**: expr1, score, and output.cerna

For the function **cupidcombinat.m**: expr2, and output.combinat

For the function **cupidindirect.m**: expr3, regulon, and output.indirect,

Chapter III. Provided MATLAB Programs

All MATLAB programs are documented and example inputs and outputs are provided. We are not providing scripts to for executing Cupid Step I & II, as these are highly architecture specific; instead, pipeline and example files accompany this user guide. Example files include context-independent tables that can be reused for predicting context-specific interactions using Cupid Step III and for evaluating other functional evidence for regulation in given contexts. Below we outline program provided as MATLAB functions.

Main functions

- **cupidcerna.m**: given miRNA-target interaction probabilities from Cupid Step II, and expression vectors for 2 ceRNA candidates followed by miRNA regulators predicted by Cupid Step II, cupidcerna evaluates the two potential directed ceRNA interactions and identifies miRNA mediators; optional variables include output file name and mediator cutoff; see description in later section. Output consists of ceRNA evaluation and miRNA mediator prediction. Input file examples, in the example directory, include “expr1” and “score”, where expr1 has expression profiles for candidate ceRNAs ESR1 and HIF1A, in addition to miRNA expression profiles. The first two rows in the output file example “output.cerna” describe the two potential directed interactions between the ceRNA candidates. Each of the two interaction rows includes a p value for the significance of the shared miRNA program, and an integrated p value for the modulation of this shared miRNA program targeting one ceRNA candidate by the other; integration is computed using Brown’s method. The interaction rows also include a list of selected miRNA mediators. The evaluation of candidate miRNA mediators is provided in following rows, and includes CMI value, corresponding p value, miRNA-target probability scores from Cupid Step II, and mediator score and status; selected mediators for the interaction have status “Yes”. Inclusion decisions about the interactions should rely on the significance of the common miRNA program and integrated expression-based evidence for modulation, and each direction should be evaluated independently.

- **cupidcombinat.m**: evidence for combinatorial regulation by miRNA species was obtained from a process using ARESLab's multivariate adaptive regression with splines (MARS) [10, 11]. We used MARS to predict target expression from the expression profiles of its miRNA regulators predicted by Cupid Step II. Non-linear interaction predictions between significantly predictive expression profiles of miRNA species was taken as evidence for combinatorial regulation [12]. Output from cupidcombinat.m can be parsed to evaluate predictive significance and identify multiplicative relationships between miRNA expression profiles. Only miRNA modules that are composed of multiplicative interactions are said to have evidence for combinatorial (or synergistic) regulation. Example input file "expr2" includes target expression profile (ESR1), followed by expression profiles of its predicted miRNA regulators from Cupid Step II; output is given in output.combinat and it includes a target identifier, count of predicted miRNA regulators from Cupid Step II, the number of samples used, mean squared error of the predictive function, associated generalized cross validation (GCV) and its p value, and number of basis functions used to predict target expression. Following rows describe the predictive function and its components, its expansion reveals combinatorial regulation; note that for in this example, permutation testing with only 10 randomized instances was ran (min p is 0.1), while Chiu et. el. (Genome Res., submitted 2014) ran MARS on 1000 permuted instances per target.
- **cupidindirect.m**: the process for obtaining evidence for indirect regulation through effectors requires regulatory networks, curated or predicted by a reverse engineering program, as input. Required input includes (1) the expression profile of the regulating miRNA followed by the expression profile of the effector, and expression profiles of all other profiled genes, and (2) the identities of the targets of the effector. We provide example input files "expr3" and "regulon" in the example directory, focusing on predicted ESR1 regulation by hsa-miR-17-5p, and using ARCNe target prediction for ESR1. The cupidindirect function calculates normalized mutual information (NMI) between the expression profile of the miRNA and the expression profiles of all other genes, and then searches for evidence of enrichment for high NMI between miRNA expression and the expression of effector targets. The function

outputs miRNA and effector name, regulon size (as provided), NMI cutoff selected, Fisher's exact test p value at the cutoff, and the adjusted p value after multiple test correction; see "output.indirect" in "example" directory for example output.

Subroutines

- **cupiddownload.m**: a script to download the 3rd party programs "fexact.m" and "ARESLab", which were not included in the package; fexact.m implements Fisher's exact test, which is needed for Cupid Step III (ceRNA prediction) and for predicting indirect interactions; ARESLab includes an implementation for MARS, which is used to predict combinatorial interactions between miRNA species [10, 11].
- **cupidkdb.m**: returns the bandwidth for a kernel density estimator, given and expression vector. The included function is used as a subroutine for estimating mutual information.
- **cupidmi.m**: calculates mutual information (MI), using kernel density estimators [13], for two expression vectors. The included function, takes as input 2 column vectors x and y of length m and two floating point variables sx and sy, which are corresponding variances of the underlying normal distributions used to calculate mutual information between x and y. The function return a floating point number, the mutual information between x and y.
- **cupidnmi.m**: calculates normalized mutual information (NMI), using cupidmi.m, between an expression vector of length m and each of *n* expression vectors. The input consists of and expression vector *x*, and expression matrix, where each column represents an expression vector of the same size as *x*, and a variable v, which indicates verbose output. The output is a vector of length *n*, providing NMI values in [0,1] between *x* and each of the corresponding expression vectors in the matrix.
- **cupidcmi.m**: given an m-by-3 matrix (m>1), containing the expression profiles of 3 genes (x,y,z) in order, this function will return an estimate of the conditional mutual information I(x;y|z) using adaptive partitioning [14]. No missing information is allowed.

- **cupidbrown.m**: this subroutine is used by cupidcerna.m to integrate p values associated with CMIs using Brown's method. Function input include p-values, weights associated with the p-values, and a correlation matrix associated with p-value computations; see Chiu et. al. (Genome Res., submitted 2014) for a detailed description.

Chapter IV. Site Prediction

Preface

The files tablePos_1481sites.txt and tableNeg_36986648sites.txt (data directory) include sites predicted and scored by TargetScan, miRanda, and PITA using default parameters in RefSeq-defined 3' UTRs on December 3rd, 2010, which include 20,491 transcripts for 18,093 genes (3PrimeUTR_20491transcripts_18093genes.txt in data directory). We predicted binding sites for 1,218 miRNAs in miRBase Release 16 in 20,491 3' UTRs of transcripts for 18,093 genes. Tables table tablePos_1481sites.txt and tableNeg_36986648sites.txt can be implicitly used to predict miRNA-target interactions in any context. In total the prediction table include, 36,986,648 sites, corresponding to 11,542,856 interactions, with no evidence from curated literature. Replacing these tables or adding to them requires rerunning the machine learning process, trained on this or another gold standard interaction table.

Prediction scores were normalized to produce scores in [0, 1]. Each site was associated with multiple predictive features. Features include:

- Quantile-normalized site scores, as given by TargetScan, miRanda and PITA.
- [0,1]-normalized distance from the start and end of the 3' UTR.
- PhastCons cross-species conservation score based on the binding-site seed (i.e., the 3'UTR regions aligned to position 2 – 8 of the cognate miRNA) using alignment of 46 vertebrate genomes [9].

Candidate site features were used for site scoring with LIBSVM, trained on equivalent features for previously validated sites and sites in 3' UTRs of previously validated miRNA target genes (table given in tablePos_1481sites.txt). For efficiency, sites were first clustered using K-means into 1481 clusters, matching the number of sites representing validated interactions. Euclidean-distance clustering was performed on feature vectors associated with sites. For classification with support vector machines, each cluster was represented by at least one randomly selected site, and large clusters were proportionally represented: *x* representatives were selected for a cluster that is *x* times the size of the smallest cluster.

Rebuilding the site prediction table (Cupid Step I)

The site prediction tables can be rebuilt and rescored using results from any number of site prediction programs, namely, given a set of target regions, as in the case of RefSeq-defined 3' UTRs, follow the described Cupid Step I protocol to identify putative sites. We, however, did not include scripts to rebuild the table in Cupid v1.0. Instead, we describe the pipeline used and note that its implementation is domain specific.

Using our approach, each identified site must be scored using all programs. A site that is identified by one program and is partially overlapping a site predicted by a 2nd program can be considered to be identified and scored by the other program. In order to use LIBSVM following our approach, site scores should be normalized to produce scores in [0, 1], and [0,1]-normalized distances from the start and end of the 3' UTR should be included in the table. Finally, site conservation, also in [0,1], should be produced. We used PhastCons to score all 7-base windows, obtaining PhastCons (46 vertebrate genomes) probability scores Θ and a geometric average score μ over Θ; the final score for $\theta \in \Theta$ was taken as $-\log_{10}\frac{\mu}{\theta}$, and these scores were normalized to [0,1].

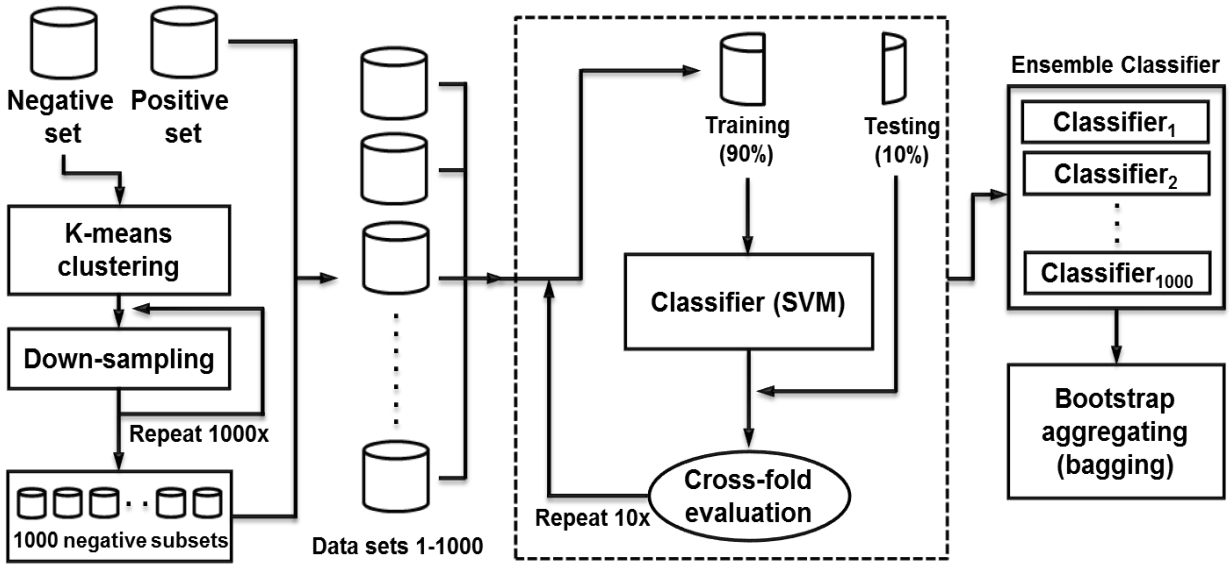

Figure 1: Cupid's learning site and interaction features. Cupid selects sites and produces probabilistic site scores for each candidate site and interaction after comparing predictive features of candidates to those of verified interactions. The process begins with sampling 1% of candidates and clustering them according to the number of verified interactions. An SVM is then trained on cluster representatives together with validated interactions within a 10-fold cross validation framework to produce probabilistic scores for each candidate interaction. The process is repeated 1000 times and candidates are scored through consensus decisions across bootstrap runs.

Compiling these data produces a table with no missing information, with a (miRNA, target) pair associated with a set of values in [0,1]. Sites associated with validated interactions and any additional curated interactions should be processed in the same way as predicted interactions and included in a site table. When training and testing with LIBSVM, the first column in the LIBSVM input table should be set to 0 or 1, denoting predicted (0) and gold standard (1) sites.

To score the resulting table, given that the number of candidates dwarfs the number of gold standard sites, down sampling may be required to effectively distinguish between candidates with similar properties to those previously identified. When predicting sites (and interactions as described later), we randomly sampled 1% of candidates (370K sites and 115K interactions) and proceeded to cluster them according to their predictive properties. First cluster predicted (0 sites) sites into as many clusters as there are gold standard sites (1 sites) using K-means. Then repeat the following process 1000 times:

(1) Randomly select a representative from each cluster

(2) Train the SVM within a 10-fold cross validation framework to produce a test probability and an exclusion/inclusion decision for each binding-site candidate.

Note that, here, 10-fold cross validation is used to score each site during the testing phase, so each site receives a single probability score in one of the cross validation rounds. When predicting both sites and interactions, we used LIBSVM [4] to score candidates. When building SVM classifiers with ten-fold cross validation, select a (cost, γ) combination for a final classifier that will be used to score all candidates (including sites excluded by the sampling process). To fine tune parameter selection, use accuracy maximization, evaluated using a Radial Basis Function kernel and a grid search process. Probability estimates are a confidence measure for the classification using the final classifier [15], trained on all cluster representatives and using the optimal (cost, γ) combination.

Repeating this process 1000 times will produce 1,000 inclusion decisions and probability scores per (miRNA, target) pair. Binding site selection follows a majority vote amongst the 1000 inclusion decisions (bagging), and binding-site scores are set to be the average probability across runs. Figure 1 depicts the learning process. Chiu et. al. used consensus inclusion (>0.5 inclusion frequency) as a determinant of site prediction.

Chapter V. Interaction Prediction

All candidate binding sites can be used to determine the probability of interactions, independently of selection in Step I. For each candidate interaction, where multiple candidate binding sites for the same miRNA were identified on a specific target, additional predictive features, including the number of binding sites, their density, their location, and their scores — computed as site probabilities by Cupid step I — can be integrated using summary functions, including trivial integration when only one binding-site candidate was identified for a specific interaction. In addition to sequence-based features, candidate interactions can be evaluated for context-specific statistical dependency and inverse correlation between the miRNA expression and the expression of the candidate-target gene using NMI. Interactions are then predicted by a support vector machine trained on previously validated interactions, using the same features as those of candidate interactions. Featured we included when scoring interactions include:

- Signed normalized mutual information between miRNA and target. The sign is from Spearman's correlation coefficient
- Maximum site score
- Median site score
- Medium range site score (max+min)/2
- Sum of site scores
- Product of sites scores, taken as [1-(1-S1)*(1-S2)*...*(1-Sn)]
- Average of sites scores
- Geometric mean of site scores
- Harmonic mean of site scores
- Root mean square of site scores
- Weighted mean of site scores, where weights are proportional to the minimum distance from start and end of the 3' UTR
- Sum of site-score squares
- Sum of natural logs of site scores
- Sum of natural exponents of site scores
- Average of site-score squares

- Average of the natural logs of site scores
- Average of the natural exponents of site scores
- The number of sites
- The genomic distance from the most upstream to the most downstream site
- The genomic distance between the closest sites
- The genomic distance between the furthest adjacent sites
- The average distance between adjacent sites

Rebuilding the interaction prediction table (Cupid Step II)

As was the case for Cupid Step I scores, miRNA-target interactions can be re-scored by rebuilding and rescoring tables tablePos_588pairs.txt and tableNeg_11542856pairs.txt using a machine learning process. The process is computationally expensive and, for this reason, may be of little value. The only context specific component of Cupid Step II is inverse correlation between miRNA expression profiles and the expression of its candidate-target genes using NMI; this feature adds relatively little to the total score. For labs with limited person and CPU time, we recommend using the provided tables and rerunning only the highly context specific Cupid Step III. However, we describe the Cupid II process to help rebuild the interaction prediction table, which may be necessary if the site table was rebuilt, or if you'd like to reevaluate interactions using context-specific correlation between miRNA expression and expression of its candidate-target genes. Note that another alternative is to rerun interaction prediction without using miRNA-target anti-correlation.

The input table to Cupid Step II can be rebuilt with custom features, in addition or instead of the ones we provided. As in the case for site predictions, each row describes features for a miRNA-target pair and the first column indicates whether the interaction is in the gold standard, with '1' or '0' for present or absent from the gold standard table, respectively. The learning process follow the protocol described for rebuilding the site prediction and outlined in Figure 1. Namely, we suggest to randomly sample 1% of candidates (115K interactions in our tables) and proceed to cluster them according to their predictive properties. First cluster predicted (0) sites into as many clusters as there

are gold standard (1) sites using K-means. Then repeat the following process 1000 times:

(3) Randomly select a representative from each cluster

(4) Train the SVM within a 10-fold cross validation framework to produce a test probability and an exclusion/inclusion decision for each binding-site candidate.

As described for site prediction, 10-fold cross validation can be used to score each site during the testing phase, so each site receives a single probability score in one of the cross validation rounds. When predicting interactions, we used LIBSVM [4] to score candidates. When building SVM classifiers with ten-fold cross validation, select a (cost, γ) combination for a final classifier that will be used to score all candidates (including interactions excluded by the sampling process). To fine tune parameter selection, use accuracy maximization, evaluated using a Radial Basis Function kernel and a grid search process. Probability estimates are a confidence measures for the classification using the final classifier [15], trained on all cluster representatives and using the optimal (cost, γ) combination. Repeating this process 1000 times will produce 1,000 inclusion decisions and probability scores per (miRNA, target) pair. Interaction selection follows a majority vote amongst the 1000 inclusion decisions (bagging), and interaction scores are set to be the average probability across runs. Chiu et. al. used consensus inclusion (>0.5 inclusion frequency) as a determinant of interaction prediction.

Chapter VI. Predicting Evidence for Competition for MiRNA Regulation

Cupid Step III is context specific and results will vary dramatically across contexts, but less so for replicate datasets from the same context. Evaluation requires genome-wide ceRNA interaction prediction, which is time and CPU intensive. We provide MATLAB code to predict ceRNA interactions between ceRNA candidate pairs, given their ceRNA candidate expression and the expression of their candidate miRNA regulators and their interaction probabilities from Cupid Step II. Each ceRNA interaction has the potential to predict new miRNA-target interactions using the supplied MATLAB script cupidcerna.m.

ceRNA prediction

We provide MATLAB code (cupidcerna.m) to independently evaluate each potential ceRNA interaction. Given Cupid Step II scores, which can be extracted from tablePos_588pairs.txt and tableNeg_11542856pairs.txt or analogous tables, and a file containing mRNA expression profiles for two target genes and miRNA expression profiles, provided code will evaluate the ceRNA interaction and identify predicted miRNA mediators. Note that expression profiles should have no missing information; the first two lines should give target expression and the following line give miRNA expression profiles, as shown in the example file “expr1”. All miRNA expression profiles provided should have targeting probabilities for both ceRNA candidates as in example file “score”; note that because miRNA targeting of one candidate ceRNA and not the other have influence on the evaluation of the candidate ceRNA interaction we suggest to include all scores interactions as input. ceRNA-pair evaluation includes miRNA mediator selection and integrated (across miRNA mediators) p-value assignments for each pair. Cupidcerna associates each candidate ceRNA pair with a list of miRNA mediators. Mediator selection is subject to a mediator cutoff, which is specified as input to cupidcerna (variable: cutoff), with a default value of 0.05. Mediator candidate miRNA M_m for the interaction $T_i \rightarrow T_j$, where T_i is a candidate ceRNA regulator of T_j is scored as $P_{i\rightarrow j}^{m}{}^{\sqrt{S_i^m S_j^m}}$, where $S_i^m \in [0,1]$ is the Cupid Step II interaction score for M_m and target T_i, derived from Cupid Step II interaction SVM inclusion decisions, and $p_{i\rightarrow j}^{m}$ is the *p*-

value of the test $I[M_m; T_j|T_i] > I[M_m; T_j]$; variables indicate the expression of the corresponding RNA species.

Each output file produced by cupidcerna produces both ceRNA and miRNA-target predictions. P values produced can be used to evaluate the two directed ceRNA interactions, and identified interactions indicate ceRNA mediators and act as evidence for competition for miRNA regulation. The miRNA-target interactomes is constructed by collecting ceRNA mediators from significant ceRNA interactions, where significant regulation of RNA1 by RNA2, mediated by miRNAs as evidence for regulation of RNA1 and RNA2 by miRNAs.

Chapter VII. Other Evidence for Functional Regulation by MiRNAs

Chiu et. al. tested candidate miRNA-target interactions for evidence for combinatorial regulation by multiple miRNA species, and evidence for indirect regulation through effectors (Figure 2). Evidence for combinatorial regulation is complementary to evidence derived from expression correlation between a miRNA and its target. Similarly, evidence for indirect regulation by miRNAs examines the correlation between the expression of the miRNA and a set of predicted indirect targets; these were not used to predict direct miRNA-target interactions and are considered complementary evidence. These lines of evidence can be used to supplement predictions from Cupid Step III, and consequently, can be used to support functional regulation by Cupid Step II predicted interactions.

Evidence for combinatorial regulation by multiple miRNA species

Evidence for combinatorial regulation by multiple miRNA species is collected through prediction of target mRNA expression profiles using ARESLab's multivariate adaptive regression with splines (MARS) [10, 11], when trained on the expression profiles of its candidate miRNA regulators, as predicted by Cupid Step II. MARS is used to identify non-linear dependence between expression profiles of miRNA species and the expression of their common target; non-linear dependencies point to miRNA modules that act combinatorially to regulate target RNAs, and evidence for the predictive power of these modules is independent of other analysis and lines of evidence used in Cupid.

The wrapper cupidcombinat.m is provided to run MARS, and its output mimics that of MARS, while adding information about the predictive function and its variables. The complexity of predictor functions was set during backward passes that minimized GCV [16], following piecewise-linear forward construction of up to 21 basis functions with a maximum degree of 3. Here, predictor functions are linear combinations of basis functions, and basis functions model multiplicative or combinatorial

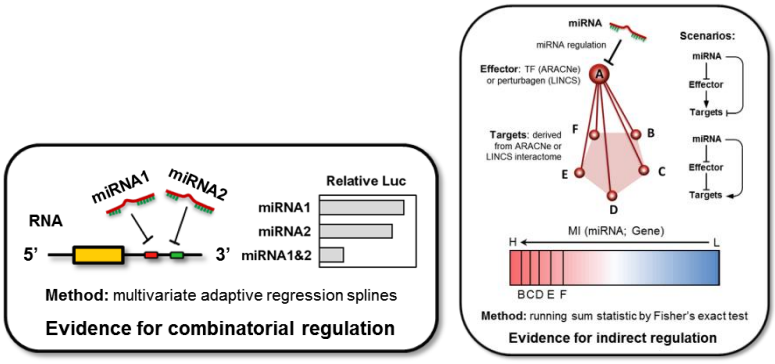

Figure 2: Evidence for combinatorial regulation (left) and indirect regulation through effectors (right) can support context-specific miRNA-target prediction.

relationships between miRNA species. We term sets of miRNAs that form non-linear basis functions miRNA *modules*. MARS was used to construct classifiers of up to 21 basis functions of the form $\{1, \max(M_m - \varepsilon_i, 0), \max(\varepsilon_i - M_m, 0)\}$, where M_m is the expression profile of a predicted miRNA regulator and ε_i is a constant termed knot. Classifiers had a maximum degree of 3, and self-interactions were excluded. Backwards construction was used to reduce the classifier to λ basis functions by minimizing the generalized cross validation (GCV) error, $GCV(\lambda)$, which penalizes for model complexity [16]. Namely we minimize

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}\left[y_i - \hat{f}_\lambda(M_i)\right]^2}{[1 - M(\lambda)/N]^2}$$

where N is the number of samples in the dataset, y_i is the expression estimate of the target gene in tumor sample i, calculated as transcripts per million (TPM) [17], and $\hat{f}_\lambda(M_i)$ is its miRNA-expression based prediction (miRNA expression is in reads per million); $M(\lambda)$ is the effective number of parameters as estimated by randomized trace method. Example input (expr2 in example) and output (output.combinat in example) are provided. In order to identify miRNA module, output needs to be parsed and mined for predictive multiplicative relationships between miRNA expression profiles.

Evidence for indirect regulation through effectors

Evidence for indirect regulation can help identify miRNA targets whose regulation is harder to detect using RNA expression profiles alone. Considering each miRNA m and a predicted direct target T_i^m from Cupid Step II, cupidindirect can help evaluate correlation between the expression profile of miRNA m and the expression profiles of predicted (direct or indirect) targets of T_i^m; we term T_i^m *effector*, and its predicted downstream targets are its *regulon*. The total RNA abundance of the regulon may be affected following miRNA-mediated inhibition of the effector, even if the effector's RNA expression is only weakly altered. Execution of cupidindirect requires both expression profiles of the miRNA, effector, and all other profiles genes (expr3 in example), and the identity if the effector's regulon (regulon in example), i.e. effector targets. Output (see output.indirect in example) includes an evaluation of the interaction between miRNA and effector, based on gene set enrichment of the regulon as a function of NMI between

miRNA expression and the expression profiles of all profiled genes. The comparison uses a running sum statistic based on Fisher's exact test, where we compare, for decreasing NMI cutoffs within the regulon, the number of included and excluded regulon genes and non-target genes. To correct for multiple testing, we use Bonferroni correction for the p-value obtained from the *n*th iteration of the test, considering this p-value as a selection from *n* trials.

In Chiu et. el. (Genome Res., submitted 2014), we used regulons of transcription factors as predicted by ARACNe [18] and genes perturbed by shRNA in Library of Integrated Network-based Cellular Signatures (LINCS) [19]. ARACNe [18] was used to measure mutual information using adaptive partitioning, with interaction p-value cutoff 1E-07, DPI coefficient 0, and using consensus predictions from 100 bootstraps. Regulons for genes perturbed by three or more targeting shRNAs in LINCS were collected by identifying genes with high and low fold change in response to shRNA transfection relative to both (1) other profiled genes in response to the same perturbation and (2) the gene's responses to other perturbations. See Chiu et. al. for a detailed description.

References

1. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
2. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.
3. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.
4. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**(3): p. 27:1--27:27.
5. Papadopoulos, G.L., et al., *The database of experimentally supported targets: a functional update of TarBase.* Nucleic Acids Res, 2009. **37**(Database issue): p. D155-8.
6. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
7. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.
8. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
9. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
10. Smith, A.D., et al., *DNA motifs in human and mouse proximal promoters predict tissue-specific expression.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6275-80.
11. Friedman, J.H. and C.B. Roosen, *An introduction to multivariate adaptive regression splines.* Stat Methods Med Res, 1995. **4**(3): p. 197-217.
12. Smith, A.D., P. Sumazin, and M.Q. Zhang, *Tissue-specific regulatory elements in mammalian promoters.* Mol Syst Biol, 2007. **3**: p. 73.
13. Moon, Y.I., B. Rajagopalan, and U. Lall, *Estimation of mutual information using kernel density estimators.* Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 1995. **52**(3): p. 2318-2321.
14. Darbellay , G. and I. Vajda, *Estimation of the Information by an Adaptive Partitioning of the Observation Space.* IEEE Trans. on Information Theory, 1999. **45**: p. 1315--1321.
15. Wu, T.-f., C.-J. Lin, and R.C. Weng, *Probability Estimates for Multi-class Classification by Pairwise Coupling.* Journal of Machine Learning Research, 2003. **5**: p. 975--1005.
16. Craven, P. and G. Wahba, *Smoothing noisy data with spline functions.* Numer. Math, 1979. **31**: p. 377-403.
17. Li, B., et al., *RNA-Seq gene expression estimation with read mapping uncertainty.* Bioinformatics, 2010. **26**(4): p. 493-500.
18. Margolin, A., et al., *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.* BMC Bioinformatics, 2006. **7**(Suppl 1): p. S7.
19. Peck, D., et al., *A method for high-throughput gene expression signature analysis.* Genome Biol, 2006. **7**(7): p. R61.