LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN TUM

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Bachelor Thesis
in Bioinformatik

Species-Specific microRNA Regulatory Networks

Andre Aberer

Aufgabensteller:	Prof. Hans-Werner Mewes
Betreuer:	Dominik Lutter, Carsten Marr, Fabian J. Theis
Abgabedatum:	15.10.2008

Ich versichere, dass ich diese Bachelor Thesis selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15.10.2008 _____

Andre Aberer

Abstract

This thesis deals with the generation and analysis of species-specific microRNA regulatory networks based upon target prediction algorithms that do not use evolutionary information. For one prediction tool a notable bias in its predictions is identified. The networks of nine organisms are examined using various topological measures and are compared with focus on how network topology and attributes changed through evolution. It is revealed, that in such basic properties as the degree distributions the nets resemble each other very well and that higher developed organisms show more regulatory activity. It turns out that the amount of microRNAs that regulate a transcript mainly depends on the length of the 3'UTR of the transcript. Beyond comparing whole-network properties the networks are rendered comparable through homology information. This way, it can be shown that the networks among mammals are closely related, while comparisons of evolutionary distant organisms only show few similarity. Functional annotation of transcripts is used in order to create a functional categorisation of microRNAs. The categorisation is utilised to show that homolog microRNA tend to keep their functions, while acquiring new target sites or losing older ones throughout evolution.

Diese Arbeit behandelt die Generierung und Analyse von speziesspezifischen microRNA-Regulationsnetzwerken, die auf Wirkort-Vorhersagealgorithmen beruhen, welche sich keine evolutionäre Information zunutze machen. Für ein Vorhersageprogramm wird eine sichtliche Ausrichtung in der Vorhersage identifiziert. Die Netzwerke von neun Organismen werden mit Hilfe verschiedener topologischer Maße untersucht. Dabei wird insbesondere beachtet, wie sich Netzwerktopologie und -eigenschaften im Laufe der Evolution verändert haben. Es wird gezeigt, dass sich die Netze in solch einfachen Eigenschaften wie den Gradverteilungen stark ähneln und dass in höher entwickelten Organismen microRNAs regulatorisch aktiver sind. Es stellt sich heraus, dass die Anzahl von microRNAs die ein Transkript regulieren von der Länge der 3'UTR des Transkripts abhängt. Neben der Untersuchung von Eigenschaften, die ganze Netzwerke auf einmal vergleichen, wird Homologieinformation angewandt, um die Netzwerke vergleichbar zu machen. Auf diese Weise kann gezeigt werden, dass die Netzwerke auf dem taxonomischen Niveau der Mammalia eng verwandt sind, wohingegen bei Vergleichen zwischen evolutionär weit entfernten Organismen nur wenig Ähnlichkeit zum Vorschein kommt. Funktionelle Annotation von Transkripten wird für eine funktionelle Einordnung der microRNAs verwendet. Diese wird benützt, um zu zeigen, dass homologe microRNAs ihre Funktionen beibehalten, während sie durch Evolution neue Wirkorte erhalten und alte verlieren.

Contents

1 Introduction

One of the major disillusions for the public that came after the releases of the human genome was the fact that plain sequence and locations of genes alone would not suffice to get a deeper insight into the secrets of life. Instead it crystalized that these DVD-filling sequences are controlled by immensely advanced, fragile regulation machineries. For achieving the higher goal of all biosciences a shift of interest towards them would be inevitable. RNA interference (RNAi) based upon microRNAs (miRNAs) is a comparatively lately discovered regulatory mechanism. Given the small size of miRNAs ($\sim$22 nucleotides), they surely dominate the ranking *amount of attention per size* closely following transcription factor binding sites (TFBS). Different than transcription factors their main function is a dampening of the expression level of a transcript (mRNA) that bears parts of the reverse complementary sequence of the miRNA in its $3'$ untranslated region ($3'$UTR) [1]. The regulation itself is performed by the RNA induced silencing complex (RISC) which includes proteins of the argonaute family and affect about 30% of all transcripts. It is assumed that miRNA act as switches or fine tuning instances for developmental programs respectively enhance the robustness of these programs. Furthermore, they influence the threshold of expression levels of regulatory units that trigger proliferation or apoptosis. Relations of miRNA misexpression to various diseases (particularly cancer) were observed [2].

In animals miRNAs occur in peculiar genes or introns being removed from pre-mature mRNA during the process of splicing. In the first case they are transcribed from the genome by RNA-Polymerase II just like a normal protein-coding gene. This transcript called pri-miRNA already shows the typical hairpin form. The non-hybridised 5' and 3' ends are subsequently cleaved by the microprocessor (containing the RNase III enzyme Drosha) yielding the pre-miRNA. This hairpin is then transfered into the cytoplasm by Exportin-5 where it is processed into the final miRNA and its complement (the miRNA*) by the Dicer complex. The miRNA (sometimes also its complement from the opposing side) is now ready for insertion into the RISC and regulatory activity.

When it comes to analyse the relationship between a particular miRNA and a transcript, *in vivo* experiments can provide phenotypes for miRNA mutants or general expression fluctuations. However, determining whether a single transcript is target of a certain miRNA has not been automated, yet. As the amount of interactions to be tested grows quadratic with the number of genes and miRNAs, the manual approach therefore is applied more exceptionally than generally. This is where prediction tools come into play. Since the observation of the first miRNA-based interaction in *D. melanogaster* between *lin-4* and *lin-14* [3], a set of rules was found out, that allows for automated *in silico* prediction tools. For most algorithms it is essential that the 5' end of the miRNA shows exact sequence complementarity. These six to eight nucleotides are termed the miRNA's seed region. Moreover, the mRNA-miRNA free hybridisation energy can be determined by RNA folding algorithms and the amount of the miRNA's target sites on the 3'UTR is taken into account [4]. Another important fact harnessed by most tools is the conservation of specific target sites among closely related species. Consideration of conservation at first helps to restrict false positive rates (e.g. down to about 30% for

PicTar [4]). Given the amount and size of a species' 3'UTRs it seems probable that a 7-mer (as putative seed) shows sufficient complementarity and good energy values but nevertheless has no biological function. But it is unlikely that the same effect is observed coincidently in related species. This meets the bioinformatic paradigm that conservation implies function. The downside is, that there is no chance to predict non-conserved and therefore species-specific interactions, which according to TarBase can be estimated to at least a quarter of all interactions [4]. Although ignored by the established tools these interactions are not neglectable: a study examining the consequences of miRNA regulation on protein expression levels [5] showed that among the interaction sites with the highest impact non-conserved sites are well represented.

In their approach to create preferably complete models of regulatory relationships of all influences that alter protein expression levels, the system biologists' results summon another fraction of scientists, namely graph analysts. Graph theory does not focus on specific interactions but instead is interested in the general picture and derived properties and measures which can unveil facts about genesis or behaviour of a network. For instance there was an approach to characterise the local and global properties of an as complete as possible regulatory network (miRNA and TFBS information was combined into a single graph) [6]. Another study analysed the over-representation of patterns in comparison to randomised graphs and came to the conclusion that gene regulatory networks (GRN) might be part of a entirely different class of networks than for instance food chain networks [7]. The restrictions shaping a network in its genesis are in a GRN similar as in neuronal networks (information processing), but different compared to food chain networks (energy flow). The over-representation of a pattern compared to randomised networks was coined a *motif* in [7] although in graph theory this is no finally defined term and thus could be deployed for nearly any network property.

In contrast to the graphs from the analyses above, a miRNA regulatory network is represented as a bipartite graph. This means that there are links between members from the partition *miRNAs* to members of the partition *transcripts*, but no intra-partition links are allowed. In conclusion the discussion of miRNA regulatory graphs works different than with fully connected graphs.

The approach of this work shall be the analysis of motifs and properties of miRNA regulatory networks among different species. Therefore it is necessary to neglect the conservation filter, which novel algorithms might allow with reasonable false positive rates. However, it is inevitable to keep an eye on the tools' behaviour as the analysis depends entirely on the quality of the predicted nets. This species-specific networks would allow to analyse the intrinsic network attributes of the whole miRNA regulatory machinery – not just the conserved part of it (which cannot a priori assumed to be representative). Though these properties can also can be compared with each other for every network, this enables an inter-species comparison of the whole networks, too. That means, that both partitions of the graph can be mapped to their orthologs of other species. The resulting networks then denotes, if the very interaction is conserved in both species and gives an impression how the whole network changed throughout evolution. Thus, it is the goal of this work to compare miRNA regulatory networks at various analytic levels.

The analysis part of this thesis is divided into two parts, chapter 2 and 3. In the first part, it is discussed, how the species-specific miRNA regulatory networks were created. Afterwards graph properties of the whole species graphs are discussed and compared among the species. The second part continues the comparison after rendering the graphs comparable down to their nodes via homolog information.

2 MiRNA regulatory networks of two non-conservative tools

The first part of this chapter covers all considerations relevant for the generation of the species-specific networks. Straight from the beginning it aims to show how two different approaches of miRNA prediction result in different graphs. While one of the prediction tools was applied, some interesting biases were observed. This is the content of chapter 2.2. After that in 2.3 the results of a short validation are discussed. The actual analysis of graph properties starts in 2.4 with a rather simple measure called bipartite density. In chapters 2.5 and 2.6 two topological properties of the graphs are compared. The last part covers the analysis of an advanced graph property, the bipartite clustering coefficient, for the miRNA regulatory networks.

2.1 Generation

As it is to compare species-specific miRNA regulatory networks the first step is to build these graphs. Unfortunately it is not possible to compare a really huge amount of organisms covering all corners of the putative *tree of life*. The possibilities here are limited by the underlying data that can later be interpreted as nodes in the networks. One partition of nodes is the miRNAs that are verified to be active in the organism, the other consists of the particular 3'UTRs of the transcriptome (as this is where the RISC targets and acts repressively). Table 1 shows organisms ranked by the amount of miRNAs they have. In the ranking only organisms occur, that have at least 100 miRNAs. The used data resource is the miRBase version 11.0 [8, 9, 10] that was released in April 2008. The so-called *star sequences* were not taken into account, as even miRBase itself lists them in a separate file. When the hairpin structure of a miRNA precursor is cleaved by the dicer, two sequences are freed. The star sequence usually is fast degenerated such that *in vivo* the expression levels of miRNA:miRNA* are 1:100 and it is assumed that in general their impact on translational repression is irrelevant.

As mentioned in chapter 1 for the generation of these networks, tools working without evolutionary information are needed. Otherwise only conserved interactions would be found and a comparison of the nets would be futile. The first choice for this job was a new in-house algorithm [11] called TargetSpy. It currently is in mature developmental stage and utilises machine learning techniques on a broad spectrum of biological features composed from knowledge about mRNA-miRNA binding. The prediction was provided as is for human, mouse, rat, fish, worm, fly and chicken based upon miRNA data from miRBase version 8 and 3'UTR data of the RefSeq gene set [12].

name	# miRNA	# in RefSeq	CR	# in Ensembl	CR
H. sapiens	677	24355	0.97	39673	0.48
M. musculus	491	19279	0.91	31929	0.57
M. mulatta	483	234	0.47	30223	0.58
R. norvegicus	293	11792	0.82	25321	0.51
*O. sativa**	*275*				
*P. patens**	*263*				
*P. trichocarpa**	*234*				
*A. thaliana**	*199*				
D. rerio	193	11586	0.92	25187	0.54
C. familiaris	176	627	0.69	19362	0.61
X. tropicalis	154	6680	0.97	24005	0.29
C. elegans	154	10928	0.47	14939	0.27
D. melanogaster	147	16653	0.80	18536	0.74
*V. vinifera**	*140*				
G. gallus	124	3636	0.87	18635	0.68
B. taurus	117	9195	0.97	21831	0.64
M. domestica	109	31	0.19	23141	0.5

Table 1: MiRNA and 3'UTR statistics for organisms with more than 100 miRNAs. Star marks plants and hash sign stands for the amount of 3'UTRs. Completeness ratio (CR) columns list the ratio of mRNAs in a database that have annotated 3'UTRs. Table is ranked by amount of miRNAs.

As the TargetSpy tool itself was not available and as tools with different paradigms can differ widely in their predictions, a second set of graphs was predicted. For this task the Probability of Interaction by Target Accessibility (PITA) [13] algorithm was applied. Its developers gained attention for proving that the mRNA secondary structure of the target 3'UTR has to be taken into account. The developers of PITA come to the conclusion that for efficient repression target accessibility is as important as base pair matching between miRNA and mRNA. In consequence they even provide on their website a prediction abstaining from evolutionary filters for four organisms.

Using the RefSeq genes, this amount can reasonably be extended to nine organisms: human, mouse, rat, fish, worm, fly, chicken and cow. The evolutionary relationship between them is shown as dendrogram in figure 1. This information was extracted from the NCBI taxonomy database [14]. With mouse and rat being close relatives to the best annotated organism human (concerning 3'UTRs and amount of miRNAs) and organisms from distant clades as nematodes and insects, this selection allows for analysis of differences that were created very recently or many speciations ago by evolution. Furthermore this set represents the most common animal model organisms.

Prediction of targets in the plant organisms shown in table 1 was not under consideration as the plants' RNAi machinery works in a different manner. Although the Ensembl [15] gene set would allow for the Rhesus Macaque and dog to be examined, RefSeq was

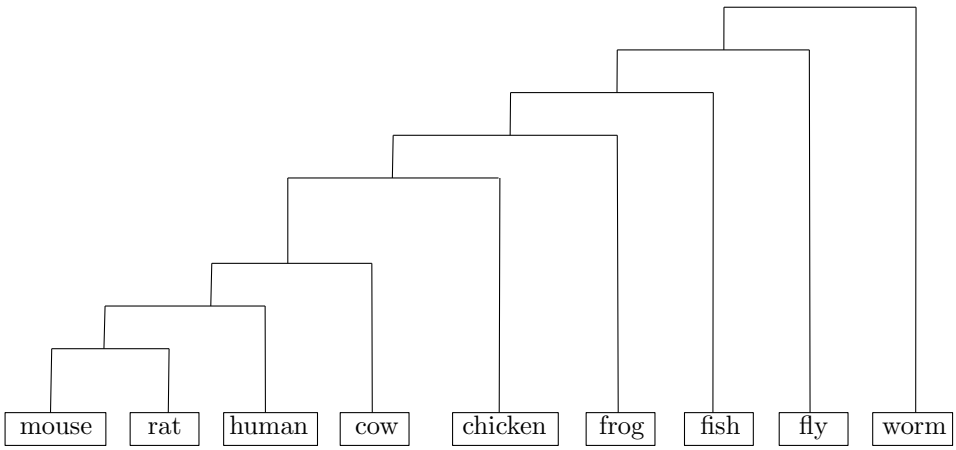

Figure 1: Evolutionay relationship between examined organisms

chosen as standard data set as it is better annotated (in sense of completeness as defined in table 1) and is the common set to be used among miRNA prediction tools.

The RefSeq data were retrieved via the UCSC table browser [16] and for comparison Ensembl data were downloaded via BioMart [17]. PITA was applied on both gene sets with its recommended parameters and functions as follows: it first searches for a seed of length six to eight. The case of an almost perfect 8mer seed match, just containing mismatch between a Guanine and Uracyl – a so-called GU-wobble – is also accepted. Then it calculates the energy difference ($\Delta\Delta G$) between the amount of energy needed to unfold the mRNA and the amount of energy that is freed when miRNA and mRNA hybridise. For calculating the secondary structure of the mRNA a 15 bases long upstream and a 3 bases long downstream flank beside the target site are considered. For the returned list of scored target sites, all those with a $\Delta\Delta G$ score below -10 were chosen to be considered valid target sites. If no cutoff were chosen the PITA method would be reduced to a simple seed search. The threshold is a problem insofar setting it too high would allow for a big false negative rate. On the other side even target prediction with positive scores could be valid interaction for instance at high expression levels of miRNA and mRNA. Furthermore it should be accentuated that a high $\Delta\Delta G$ only denotes the tendency of less repression. Yet, binary relationships not quantified ones are needed here.

2.2 Relationship between energy, cutoff, degree and GC-content

That the -10 cutoff is merely vaguely suggested [18] by the publishers of PITA gave the impulse for some experimentation with it. As TargetSpy has been provided as an already balanced prediction, it seems reasonable to measure the distance between both predictions as a function of the cutoff. As a second similarity measure the ratio of the average density of both scores is taken (with the natural optimum 1). The average density [19] is defined as the amount of links in the graph divided by the amount of all possible links. The distance d between the graphs A and B with the $n \times m$ adjacency

matrices $(a)_{ij}$ and $(b)_{ij}$ is defined as

$$d(A,B) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}|(a)_{ij} - (b)_{ij}|}{n \cdot m} \tag{1}$$

and measures the percent of matrix fields in the adjacency matrix that differ.

Figure 2 shows the results for the chicken graphs and surprisingly enough both optima appear near the suggested energy cutoff of -10.

Having two predictions based on entirely different methods brings one major advantage: one might expect that one method should identify targets that the other one cannot find and vice versa. But as the truth should lie somewhere between both predictions, their predicted set of targets for each miRNA should not be entirely different. So it comes a bit astonishing that in fact the predictions for each organisms have only few targets in common.

Table 2 shows an overlap score for each organism that consists of the amount of targets predicted by both tools divided by the amount of all target genes predicted by either PITA or TargetSpy. This score favours differing tendencies, as for the amount in the divisor the predictions can stem from both predictions. But nevertheless in any case an overlap of 10% is not achieved.

Looking for an explanation for this rather wide disparity, an interesting fact was observed: as PITA's main ingredients are the miRNA-mRNA hybridisation energy and

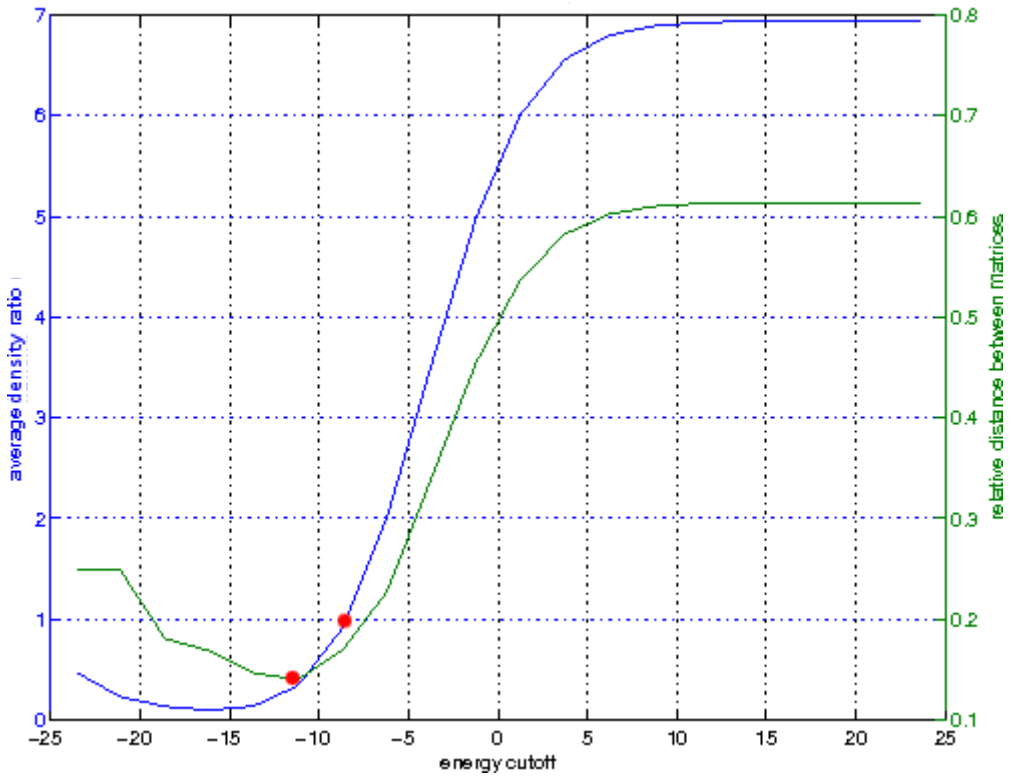

Figure 2: Distance (green) and average connectivity ratio (blue) of TargetSpy and PITA prediction under varying PITA cutoff with optima marked (red).

organism	worm	fly	fish	chicken	human	mouse	rat
overlap[%]	8.7	6.3	7.2	9.2	8.0	9.2	8.8

Table 2: Amount of common targets divided by all targets predicted by either PITA or TargetSpy

the energy necessary to break up the secondary structure of the mRNA, both these values strongly depend on the guanine-cytosine content (GC-content) of those sequences. The GC-content is simply defined as the percentage of guanine and cytosine per sequence. As these two bases form stronger Watson-Crick bindings with each other than the two remaining bases, the amount of free energy is higher when a GC-rich miRNA hybridises and it costs more energy to unwind the secondary structure of a GC-rich mRNA.

The effect for the PITA method can clearly be seen in figure 3. Obviously, there is a strong inverse correlation between the miRNA's GC-content and its $\Delta\Delta G$ score with a Pearson-Correlation-Coefficient (PCC) of -0.947 and a P-value near zero. Figures for other organisms were nearly identical. With such a strong inverse correlation the GC-content of the miRNAs is presumably the main necessary condition for a miRNA for being predicted with a high score. Of course, AT-rich miRNAs might be predicted in the same amount, but in average they achieve a much lower score. Therefore they fall below the suggested -10 cutoff. Thus, PITA prefers GC-rich miRNAs in its prediction, but if this correctly reflects the biological circumstances remains questionable.

To illustrate the full effect of this tendency on the prediction, the target site predictions were ranked by their energy score. Then this ranking was divided into non-overlapping sets of equal size (438 as shown here in fish). For each of this set the average GC-content of the predicted miRNAs and the amount of different miRNAs in the set was calculated. Figure 3 shows the result of the computation: Sets with predominantly GC- or AT-rich miRNAs contain less different miRNAs than with a mixed combination or miRNAs with an average GC-content. Combined with the fact that GC-rich sets should have a higher $\Delta\Delta G$ score and vice versa for AT-rich ones, it becomes clear, how a PITA prediction is composed: the energy ranking is dominated by target sites that are easy accessible and are targeted by GC-rich miRNAs. The AT-richer the miRNA of a target site becomes, the better accessible the target site must be in order to yield a good energy score. Yet, both miRNAs (the AT-rich and the GC-rich) might match the sequence in the mRNA equally well. With such strong influence of GC-content on the prediction, it is likely that a valid target site with an AT-rich miRNA gets the same $\Delta\Delta G$ score as a falsely predicted GC-rich miRNA, that for instance does not match the sequence in its target so well, but anyway gets a better hybridisation energy score. Recently, it was shown that another tool seems to have a similar bias [20]. Here also AT-rich miRNAs are predicted less often, which they state, is especially a problem for scientists working with organisms that have generally low GC-contents in their sequences.

TargetSpy uses a much wider range of features and does not concentrate to that extent on the GC-content as a feature. Figure 4 is a direct comparison between TargetSpy and PITA. Each scatter represents a human miRNA. It becomes once again clear, how the

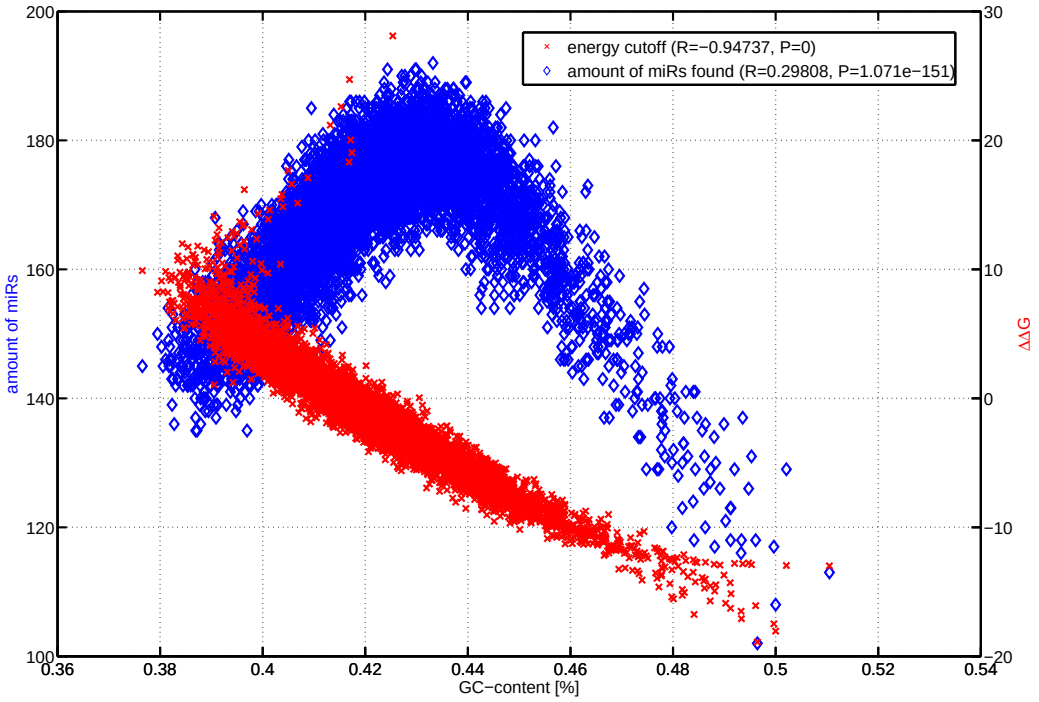

Figure 3: Effect of GC-content of miRNAs on the $\Delta\Delta G$ score (red) and the amount of different miRNAs found in part of a target list (detailed explanation in text) ranked by energy score (blue) in fish. Correlation coefficients and P-values in brackets.

amount of targets predicted depends on the GC-content. This amount is represented in figure 4 as the ratio of transcripts regulated (relative degree) in order to render both predictions comparable. While as already shown the PITA prediction favours GC-rich miRNAs and predicts them to have generally more targets than AT-rich miRNAs, this is completely different for the TargetSpy prediction where the data show moderate inverse correlation. Especially for TargetSpy this observation does not hold for every other organism (in mouse for instance there is a slight correlation). This surely is also a effect of the learning techniques applied in TargetSpy. However, the GC preference of PITA holds in any other organism. Thus, miRNAs with utterly different attributes are predicted to have many targets among the two tools.

In any case this reveals a facette that makes it more imaginable how the considerable lack of overlap as observed in table 2 can occur. These biases should be kept in mind during any further analysis. Yet, having two such different prediction methods also offers two very different perspectives on the general target prediction problem.

It must not be forgotten, that miRNAs are just half of the cake. The GC-content of the 3'UTRs is also of importance for hybridisation. The histograms in figure 5 reveal two interesting tendencies. At first with human and cow, two of the four mammalian

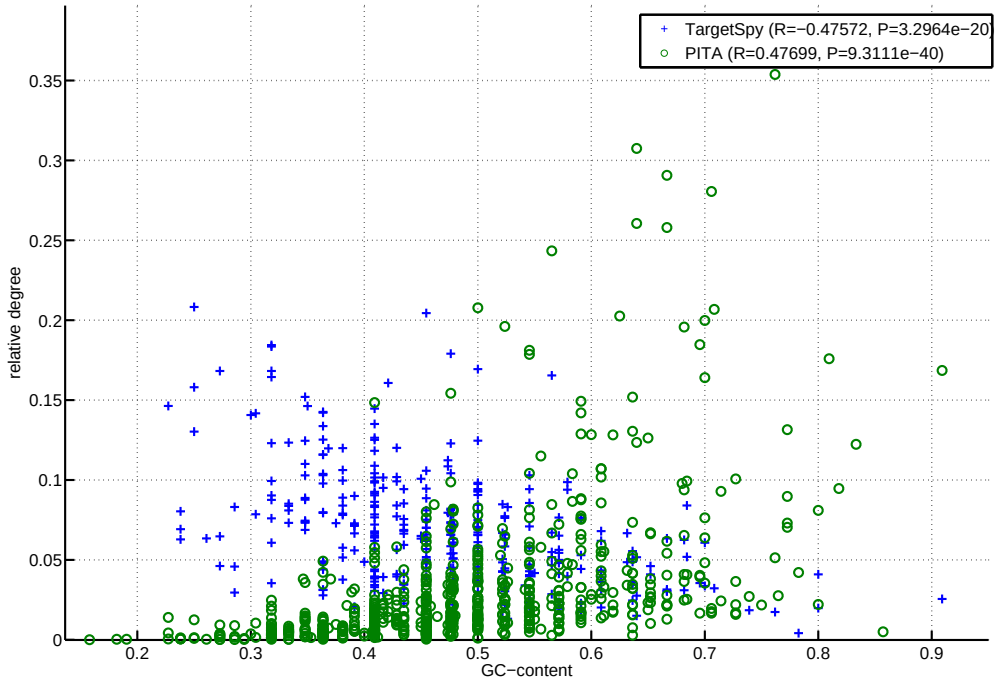

Figure 4: Connection between relative degree and GC-content of miRNAs in PITA (blue) and TargetSpy (green) prediction on human sequences

organisms have two peaks of very frequent GC-contents. For human there seems to be an interesting explanation, that was found in [21]. Human genes can be separated into two partitions: a set with high GC-content and a set with AT-rich 3'UTRs. When in the study functional annotation was added to the analysis, it revealed that genes with GC-rich 3'UTRs have predominantly functions like signal transduction and posttranslational protein modification, while genes with AT-rich 3'UTRs are often involved in transcription and translational processes.

The second thing to mention about the histograms is that the GC-content decays steadily with the assumed evolutionary age of the organism. Respectively, the GC-rich peak vanishes in older organisms. In the study mentioned here, it was also stated that the functions of the genes with the AT-rich 3'UTRs are evolutionary older. This seems to be clear as translational and transcriptional processes are essential for almost any form of life. Extrapolated from the point of view of the human genes one might now hypothesise that genes with AT-rich 3'UTRs are present among all species and cover the most basic activities, while in the higher organisms (or at least in the mammals) there emerged an additional set of genes for higher (or maybe mammal-specific) functions. In any case this set for mammal-specific functions would also draw different regulatory activity of miRNAs upon themselves.

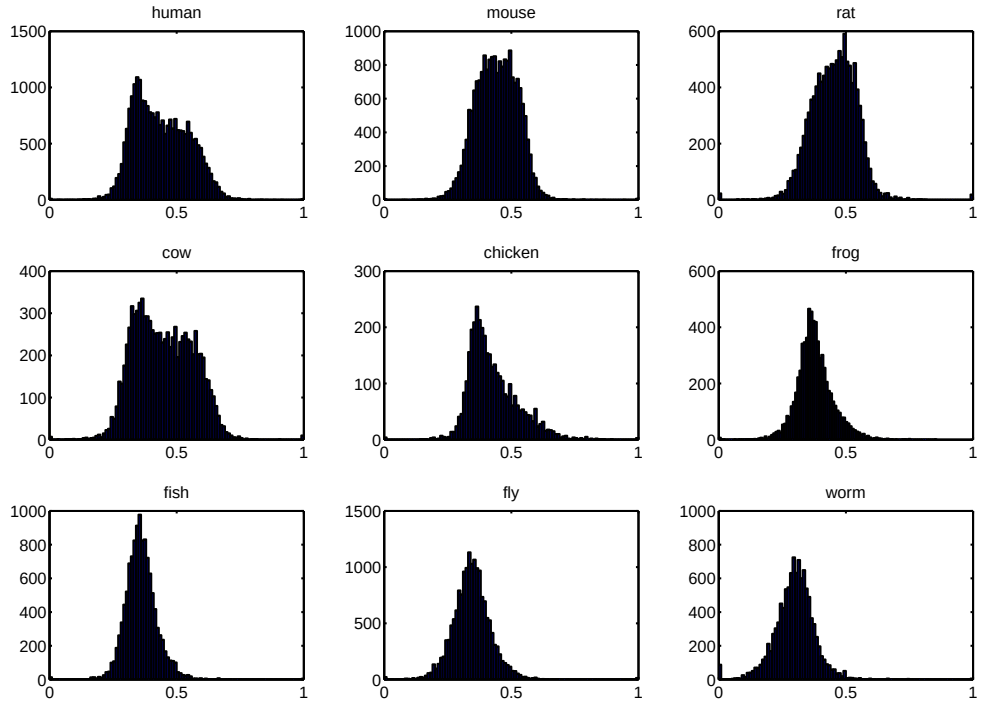

Figure 5: Histogram of GC-contents of 3'UTRs among species

2.3 Verification

Verification of miRNA regulatory networks is still very hard as the amount of experimentally verified miRNAs is very low and badly organised as most people working with miRNAs use various prediction tools instead of experimentation. Thus, MirBase for instance, often used for queries about miRNA targets, offers various predictions from various tools (including miRanda [22]). In order to fill this gap, TarBase [23] was created. This is a database containing only experimentally verified interactions, but as it is composed from numerous sources, the quality is varying. After filtering (only human interactions with positive or negative support and no data from micro-arrays were taken into account) and mapping of the gene symbols to RefSeq identifiers, a positive set of 142 interaction and negative set of 12 interactions remained.

	TargetSpy	PITA
positive set (total: 142)	34 (24.1%)	13 (16.6%)
negative set (total: 12)	2 (16.7%)	1 (8.3%)

Table 3: Verification with TarBase interactions

The size of the sets does not suffice for serious false positive/negative statistics as this is a just a evanescent detail of the whole picture (with for comparison the human

PITA prediction having $5.4 \cdot 10^5$ links). Although the results on the positive set are not overwhelming, the amount of false positives is calmingly low. This is particularly valuable as false positives are the main problem when doing without a conservation filter.

2.4 Network size and bipartite density

The most basic property of miRNA regulatory networks is, that they are bipartite networks with the two partitions *transcripts* and *miRNAs*. This means, that in the theoretical model there never might be a link between two transcripts or two miRNAs, although in reality surely a gene product can act as activator or repressor for further gene products. With two separated partitions it is not meaningful to search for regulatory patterns that are overrepresented in comparison to randomised models as it was done with TFBS networks. For a distinct amount of nodes we have in a bipartite graph fewer possibilities how these nodes could be connected and thus it would be necessary to examine more nodes at once that could constitute a valid motif as defined in the study already mentioned [7]. On the other side, if more than three or four nodes were taken (as this was the case in the TFBS-study), a combinatoric explosion would be the consequence. For instance there are 343 possibilities how two partitions with each three nodes can be connected when every node shall be connected by at least one link. Searching for these motifs would be utterly time consuming and even more important, their biological meaning would be questionable. Instead the theory of affiliation networks and hypergraphs offers orientation.

As proposed in the referenced hypergraph paper [19], the discussion of bipartite graphs could begin with the amount of nodes in the graphs. These amount have already been presented implicitly in table 1. Yet, one could rightly state, that there is not much mystery in the reason, why for instance human has on the one hand more miRNAs and on the other hand more transcript nodes in the predicted graphs: the human genome simply is stronger annotated, as human is maybe the most interesting model organism. As the identification of miRNAs is still in process, it is very likely that many miRNAs still remain to be discovered and thus the current miRNA regulatory graphs are just more or less big subgraphs of the true constitution.

The number of links in the graphs strongly depends on the amount of nodes in both partitions and in conclusion, there is few sense in comparing these measures. However this amount can be normalised by the product of the amount of nodes in both partitions. This is what could be called the *bipartite density* and in effect it tells how many of all possible links are in fact predicted. Figure 6 shows the bipartite densities for the organisms in the PITA prediction. Values range from circa 0.5% starting with worm up to 3% in human. Interestingly, it is human that achieves the highest bipartite density with some distance followed by the other mammals (cow is to a certain extent an outlier here). In contrary the densities of the evolutionary most distant organisms fish, fly and worm are very low. The tendency seems to be clear: the more complex an organism is, the denser is its regulatory network.

It should be mentioned that the observation described here, does not really hold for the TargetSpy prediction. In TargetSpy the amount of links (and thus also the bipartite

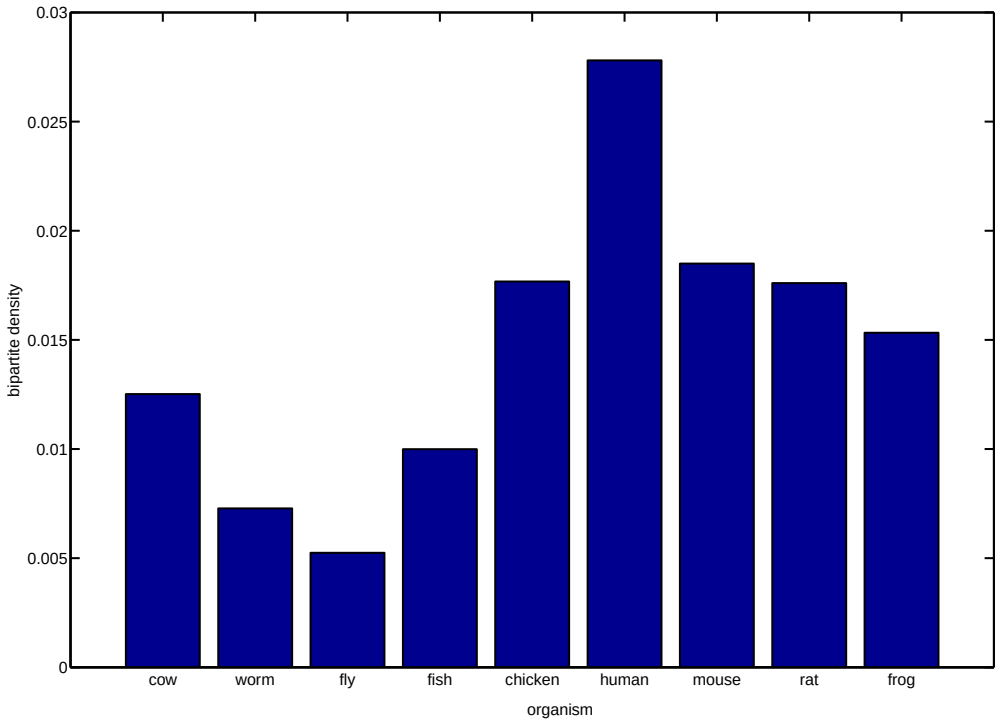

Figure 6: Bipartite density among organisms in PITA

density) is in average around twofold higher than in PITA. This is mainly the effect of a decision that has not yet been discussed here. PITA allows for some parameters like size of the seed to be searched for, amount of GU-wobbles, mismatches and loops in the seed. In their FAQ [18] they state that if called with default parameters no mismatch is allowed in the seed. But if the program is called without any arguments, it will also search for 8-mer seeds that include a mismatch. This might be in fact a correct assumption as currently it is believed that if the rest of a miRNA hybridises well, seed complementarity might not be that important. But if this option had been chosen, the amount of links predicted would have been even higher than TargetSpy's prediction by factor at least two. This would have yielded unhandily dense networks and would have enforced the problem of false positives. Beside that, the bipartite densities in TargetSpy did not show a structure that can be explained so well with evolutionary distances. Instead in TargetSpy the bipartite density in chicken is about twice as high as in other organisms. A probable error source of errors in this case might be the fact, that it is not entirely known what set of transcripts and miRNAs was used for the predictions, as it was provided as is. Because of that, for most graph properties the examination will concentrate on PITA.

2.5 Degree distributions

Transcript degree distributions

The strong implications that come with the statement that higher level organisms also have denser networks deserves closer examination. After all, the bipartite density is a rather coarse measure. Much more details can be seen in the *degree distributions*. The degree of a node in the graph is its amount of edges that connects it with other nodes (of the other partition). In the biological context here, the degree of a miRNA is the amount of targets it regulates and vice versa the degree of a transcript is the amount of miRNAs that regulate it. Respectively, the relative degree is defined as the amount of nodes of the other partition that are linked to the node divided by the amount of all nodes of the other partition. When comparisons among various nets are performed, it makes generally more sense to work with relative degrees. Furthermore, it is common in graph theory to term the nodes with the highest degree as *hubs*. Biologically, a miRNA hub would be a miRNA that regulates a large percentage of genes products. Conclusively, it is the transcript hubs that can be seen in the rightmost part of the transcript degree distributions, shown in figure 7.

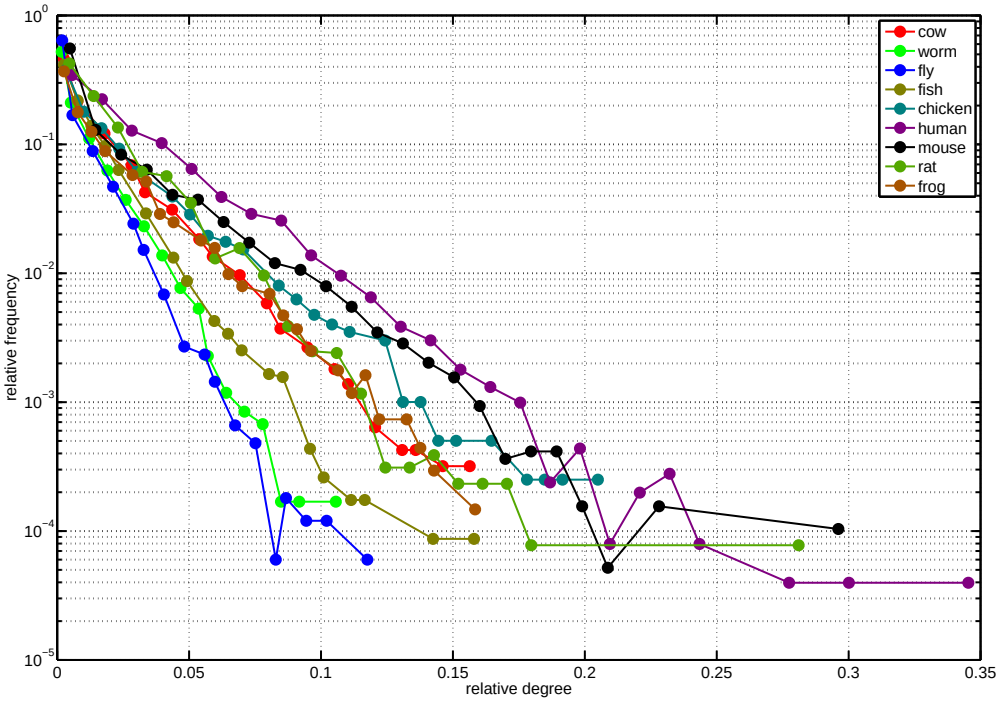

Figure 7: Relative degree distributions of transcript nodes among species

As already discussed, the nets vary in their link amounts and therefore the normalisation to relative frequency (thus showing the percentage of transcript with a certain

relative degree) comes handy. With logarithmic y-axis it becomes easy visible that the transcript degree distributions of all nets are exponential distributions. Thus, the frequency of a link with degree x is proportional to $e^{-\lambda x}$, where λ is constant. As this applies to all nets under examination, this topological attribute seems to be well conserved, respectively it is an implicit network property of miRNA regulatory networks. In biological terms it appears logical that many transcripts are only regulated by a few miRNAs, after all it is estimated that just 30% is regulated at all. There are genes whose products are needed at any time and thus regulation of these genes would not make sense. On the other side, some proteins are only of necessity in certain tissues, certain developmental stages or special circumstances (initiation of apoptosis for instance). These could be the transcript hubs, which are repressed by many miRNAs in order to avoid a waste of translationary resources and material. They would only become active, when there are no or very few of its regulatory miRNAs present. Yet, it is hard to imagine, for example more than 60 miRNAs (which according to PITA is more rule than exception) can be concerted into a useful signal that leads to a gene product when needed and supplies enough repression when it is not needed. Of course, here is also redundancy involved as will be discussed later.

In fact, the transcripts with the highest degree in human (more than 230 regulating miRNAs according to PITA) indeed are themselves involved in regulatory activity. One of them is a HOX-gene (gene symbol: ONECUT2), that is itself a transcription factor. AAK1 and HIPK2 are kinases, the latter interacting with homeodomains and the fourth notable highly regulated transcript is a kinase suppressor (KSR2). As kinases often appear in multiple signaling pathways, it is important that they only are active when their pathway is meant to conduct information.

As the relative degree gets higher, a clear tendency develops: organisms like human and mouse have in general more high relative degree transcripts than other organisms. The ranking is reminiscent of figure 6, which makes instantly sense. Organisms with higher relative degrees of course also need more links in their nets in average. Just as in the bipartite densities evolutionary distant organisms like fly and worm generally have the least regulated transcripts. Furthermore these organisms with more high relative degree transcripts also reach higher relative degrees. This is not an artifact of the binning applied here as it was created with bins of equal size over the whole range of values. As known from table 1 the species are differently well annotated. Thus the multiple testing problem might occur when comparing for instance human with cow (around 25.000 and 9.000 transcripts). As human has more transcripts the probability is higher that some of these transcripts are by chance the target of more miRNAs than when there are only few transcripts as in cow. The multiple testing problem might indeed be responsible for organisms like human and mouse achieving the highest relative degrees in figure 6. However it still would be possible that for instance worm has higher relative degree transcripts in the middle field (5 to 10%), which is definitely not the case. Despite the multiple testing problem, chicken achieves the fourth highest relative degrees although chicken is one of the poorest organisms concerning annotated transcripts. Fly on the other side has the overall lowest degree hubs and its amount of annotated transcripts is comparable to mouse. This strongly indicates that the mammals (dominated by

human) have the highest relative degree hubs and having generally more high relative degree transcripts is not simply an artifact or the result of multiple testing. Instead this really might have a biological meaning. The deeper sense of this tendency could be, that these species in fact have a more elaborate posttranscriptional regulatory system.

As a last notice it should be mentioned, that there are also transcripts that are not regulated at all by miRNAs (with good reasons as stated above). Table 4 shows exactly that ratio for each organism. With human having almost all of its transcripts targeted by at least one miRNA, this once more shows how incomplete data for the other organisms still might be. Surprisingly enough, even in mouse many transcripts are not regulated according to PITA in spite of the high amount of miRNAs that has already been identified there. The very low values of human and rat are consistent with recent presumptions that more than 30% of all gene products are posttranscriptionally regulated. However, in this case the additional false positve rate of a conservation free prediction might have contributed to that.

organism	worm	fly	fish	chicken	cow	human	mouse	rat	frog
not regulated [%]	52	64	39	38	46	6	40	17	37

Table 4: Ratio of transcripts not regulated at all

Target site density on 3'UTRs remains equal

However, there must be a reasonable explanation, why on the average 3'UTR of a human there are a lot more target sites than on that of worm. If all species' 3'UTRs were equally long, then the human target sites would be located very densely on the 3'UTRs and this might raise the question, if there is actually enough space between them, such that two neighbours could be biological active at the same time. Yet, this problem does not show up, as in reality the 3'UTRs are not equally long. The ranking of average 3'UTR lengths pretty much reflects the tendency shown in figure 7. Organisms, whose transcripts have higher relative degrees, also have longer 3'UTRs. When figure 7 is normalised by the average 3'UTR length this yields a measure, signifying how dense the various target sites are distributed on the 3'UTRs. And in fact as can be seen in figure 8 these densities do not vary greatly. Instead the 3'UTR length explains the differences between the species observed in figure 7 very well. Or at least pretty well with the exception of worm, that seems to have many more target sites on its short 3'UTRs than expected from comparison to the other organisms. This is really astonishing, because of the 3'UTRs also having a much lower GC-content as we know from figure 5. In effect PITA actually should rather predict fewer targets and thus the density should actually be lower than in worm. However when talking about GC-contents of 3'UTRs the GC-content of the miRNAs must not be forgotten. The mean GC-content in human (star-sequences neglected) is around 49%, while in worm the mean is around 4% lower, which would allow for more target sites with good hybridisation energy. It is questionable, if this is enough, to create such a higher density of target sites on the 3'UTRs of worm but it surely enforces the tendency.

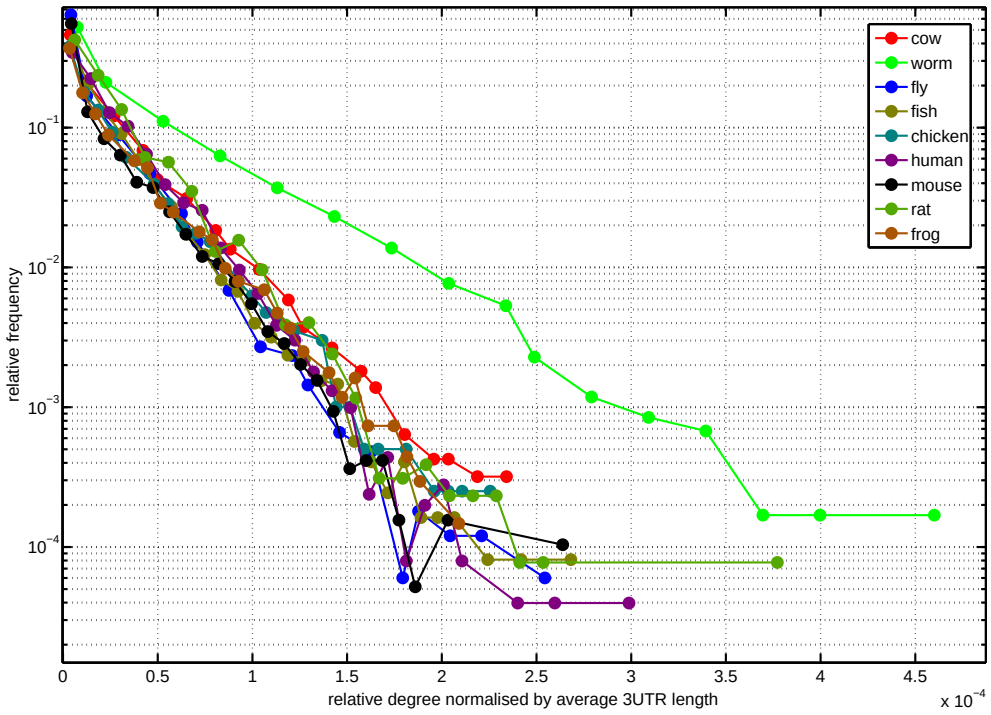

Figure 8: Relative degree distributions of transcript nodes normalised by average 3'UTR length

Degree distributions of miRNAs

The degree distributions of the other partition, the miRNAs, shows to a certain extent a similar picture as in figure 7. The relative degree distributions of the miRNA nodes of the networks specify how many miRNAs regulate which fraction of all of the organism's transcripts. The main problem with these distributions is, that the number of nodes this time is about two orders of magnitude lower, which make it necessary to apply a very coarse binning in order to even the stronger fluctuations in the distributions. With around 100 data points as we have it here for organisms with fewer annotated miRNAs, it is harder to get a straight distribution. However the well annotated mammal organisms provide a picture that can be explained coherently. Of course, there are no miRNAs without any targets as this mostly would be the product of too restrictive parameters in the prediction. Depending on the species there is mostly a low amount of miRNAs having a very low relative degree, which cannot be seen well in figure 9 due to the coarse binning. Instead the peak can be found at a relative degree of 0.1 to 1% depending on the species. After this maximum the distribution strongly resembles a power law relationship. In mathematical terms this means that the amount of miRNAs regulating k transcripts is proportional to $k^{-\alpha}$, a property that is often observed in bipartite and fully connected real world networks [19]. The special attribute of power law degree distributions is scale

invariance and thus networks, where they occur are also termed *scale-free* networks. Scale-free miRNA degree distributions seem to be a well conserved topological property, at least among the mammal organisms. It seems natural that evolutionary more distant organisms share this network motif, too (in fact fly even does so), but there are currently not enough miRNA to tell that for sure. Furthermore the whole distributions in human, mouse and rat resemble each other very well, beside the fact that they are shifted. This once more is an effect of differently high bipartite densities in the species-specific graphs.

In a scale-free network, the factor α denotes the networks heterogenity. Different than in figure 7 there is not much difference in the slope of the appropriate organisms data points which means that heterogenity from the miRNAs' point of view is also well conserved among the three well annotated mammals.

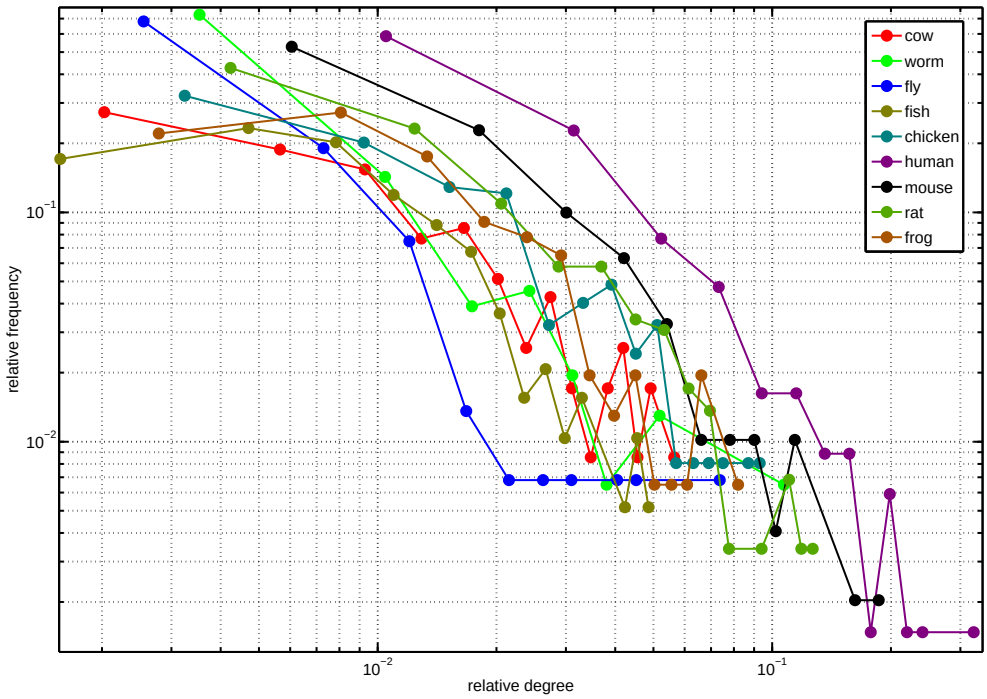

Figure 9: Relative degree distributions of miRNA nodes among species

2.6 Neighborhood correlations

So far the amount of the nodes with a certain relative degree were examined among the organisms and for both partitions. From this point of view it would be interesting, if there is any preference of a node with a certain degree to regulate nodes that have a particular high or low degree. In biological terms this equals the question, whether for instance miRNA hubs tend to regulate transcripts that are regulated by many or few

miRNAs. In graph theory all nodes that are connected with a node are called the nodes *neighbourhood* [19]. Thus, all transcripts that are regulated by miRNA α are called first neighbours of α, while miRNAs co-regulating a transcript with α are second neighbours of α, as the shortest path between the two nodes is of length two.

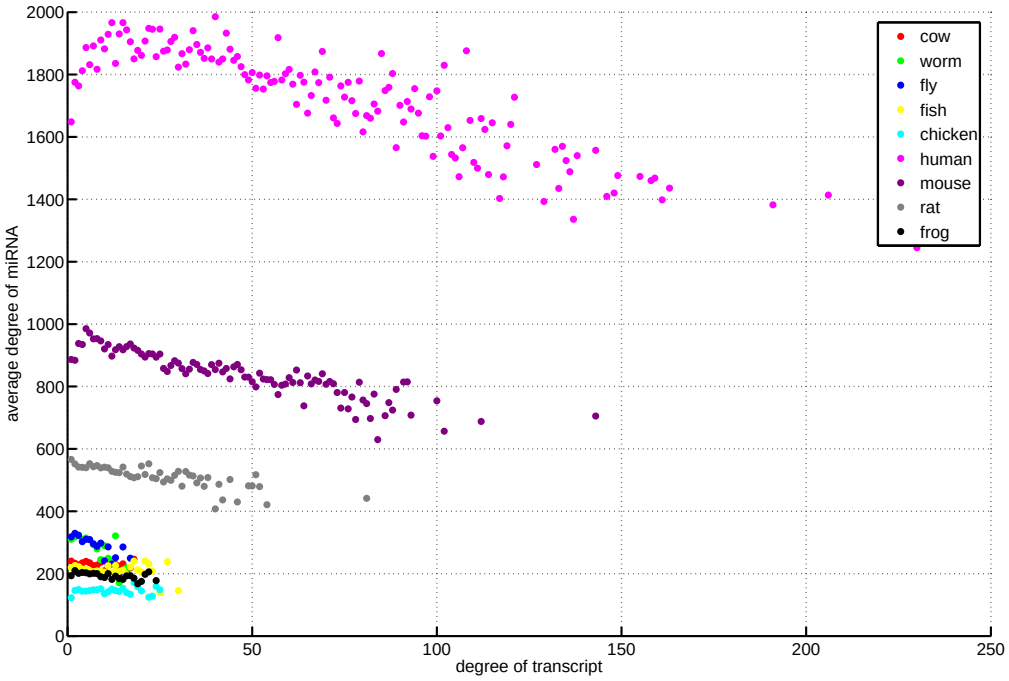

Figure 10: MiRNA degrees compared to average degree of regulated transcripts

First neighbour degree correlation

Figure 10 shows what could be called a first neighbour degree correlation. For all miRNAs of a certain degree, the average is taken of all first neighbour degrees. The plot shows a moderate inverse correlation of these data points for most organisms. For very low degree miRNAs the transcript's average degree does not start at its maximum (especially in human and mouse). Then the rule seems to hold that the higher the miRNA degree, the lower is the average degree of the transcript regulated by these high degree miRNAs (hubs).

The same plot also exists for the transcripts, where the tendency could be seen, too, though a bit more slightly. In essence, it appears logical that both plots resemble each other as it is only two different point of views from which the same phenomenon is observed. This is, because of the symmetry of the problem. Each link in the network connects two nodes with two specific degree values. When the transcripts neighbourhood degree correlation is examined, we always take one side of each link and the average of all other sides of the link and vice versa for miRNAs.

The biological implication of figure 10 is that when a transcript is regulated by many miRNAs, then these miRNAs tend to have lower degrees. A possible interpretation could be, that transcript hubs are in average fewer regulated by miRNA hubs. It can be supposed that miRNA hubs do not regulate very specific. As already stated above transcript hubs are likely to need very specific regulation. This specific regulation is provided by for instance tissue specific miRNAs that as consequence of their tissue specifity do not target as many miRNAs. A very simplified example would be a transcript that is only to be translated in kidney. Then in any other tissue it should be repressed, which is performed by tissue specific miRNAs for all of the other tissues. As this effect appeared more or less strong in any organism, so far, it seems as this once more is an intrinsic network property of miRNA regulatory networks.

It is common in graph theory to compare a result as that of figure 10 to randomised graphs. When randomised graphs with the same degree distributions show the same effect as observed in the natural one, then it is likely that the phenomenon under examination is not an intrinsic property of the network, but instead merely a product of the present degree distribution.

One randomisation step of the algorithm applied [24] here basically consists of drawing two nodes of each partition, swapping their link and assuring, that the degrees of all nodes involved did not change. With growing bipartite density it becomes more difficult to find four nodes where swapping the link is possible. Because of this time bottle neck, it merely was possible to create one randomised graph for every species. One question to clarify concerning the algorithm still remained open: how many successful randomisation steps are necessary in order to provide sufficient randomisation? The bipartite density was observed to be the major determinant. In figure 11 the distance (defined as in equation 1) between the original and the product after an amount of randomisation steps is shown. For the test 10 small graphs with different bipartite densities were created at random. The test shows that when circa three randomisation steps are performed for every link in the graph, the distance of the product graph from the original stops to grow and thus best possible randomisation is achieved.

When the first neighbour degree correlation test was performed on the randomised networks, a very similar effect was observed as in the natural graphs. After all, the inverse correlation in the randomised nets even turned out to be greater. The inverse correlation is possibly just the product of the application of the mean as measure in the plots. The higher the transcript degree becomes, the more miRNAs are its neighbours. But as known from figure 7 the amount of miRNA hubs decays with increasing degree. Thus, transcripts with increasing degree are not merely regulated by high degree miRNAs (which regulate tautologically most of the transcripts) but also by the lower degree miRNAs. While in the mean this appears as inverse correlation, there is only very mild inverse correlation when the PCC is calculated for both degree values on every links. The PCCs calculated in table 5 shows at first, that there is almost no correlation and secondly that the PCCs in the randomised nets are generally even higher than in reality. Although in these PCCs hubs are overrepresented (as they simply have more links), this result renders it hard to infer biological meaningful conclusion from it.

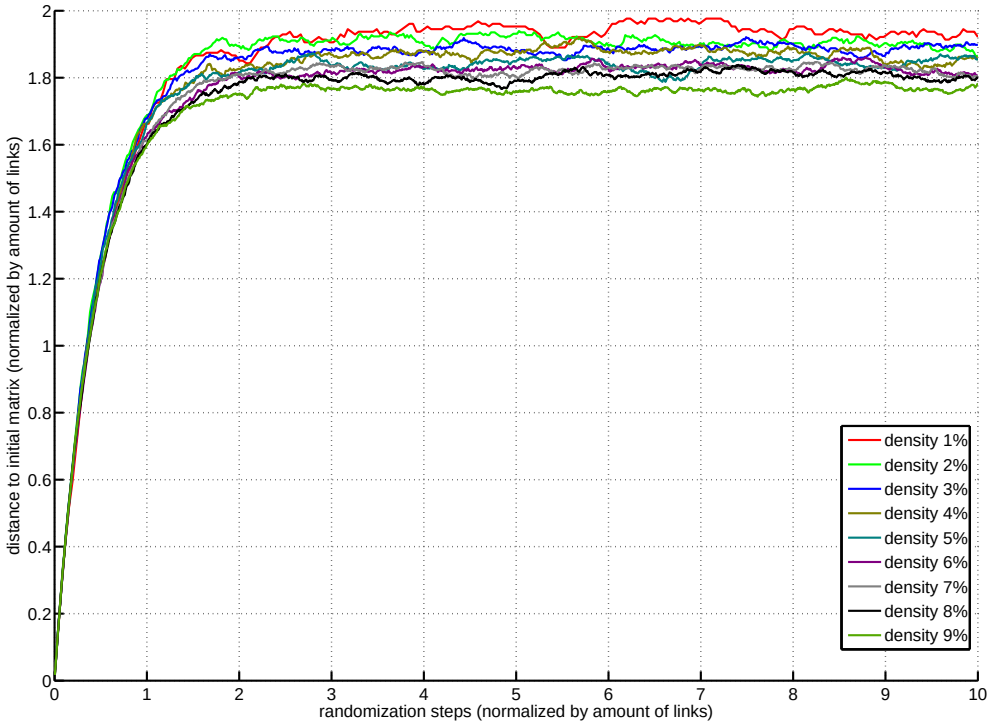

Figure 11: Growing distance of randomized graph from original graph. Both axis normalised by the amount of links in the network.

Second neighbour degree correlation

The particularly interesting point about the second neighbour degree correlation was the miRNA partition. As already mentioned above, the question here is, whether miRNAs that co-regulate transcripts (the second neighbours of the miRNA α) have a degree preference depending on the degree of α. However, in this case there are two possibilities for scoring the second neighbourhoods' degrees: firstly, simply all second neighbours can be determined for the degree correlation and secondly this degree of the second neighbours can be weighted with the amount of times that the miRNAs are second neighbours. In this scenario the weight would consist of the amount of transcripts that are co-regulated by two miRNAs. From a purely graph theoretical point of view

	worm	fly	fish	chicken	cow	human	mouse	rat	frog
real-world [%]	−4	−3	−3	2	−2	−5	−6	−4	−4
randomised [%]	−6	−1	−5	−5	−9	−8	−5	−5	−4

Table 5: First neighbour degree correlation coefficients for every link's two degree values in the real-world and randomised networks

the unweighted correlation is more interesting, because when the described weight is considered, the analysis becomes distorted by hubs that have more links than low degree miRNA who are respectively fewer times second neighbours and thus count less.

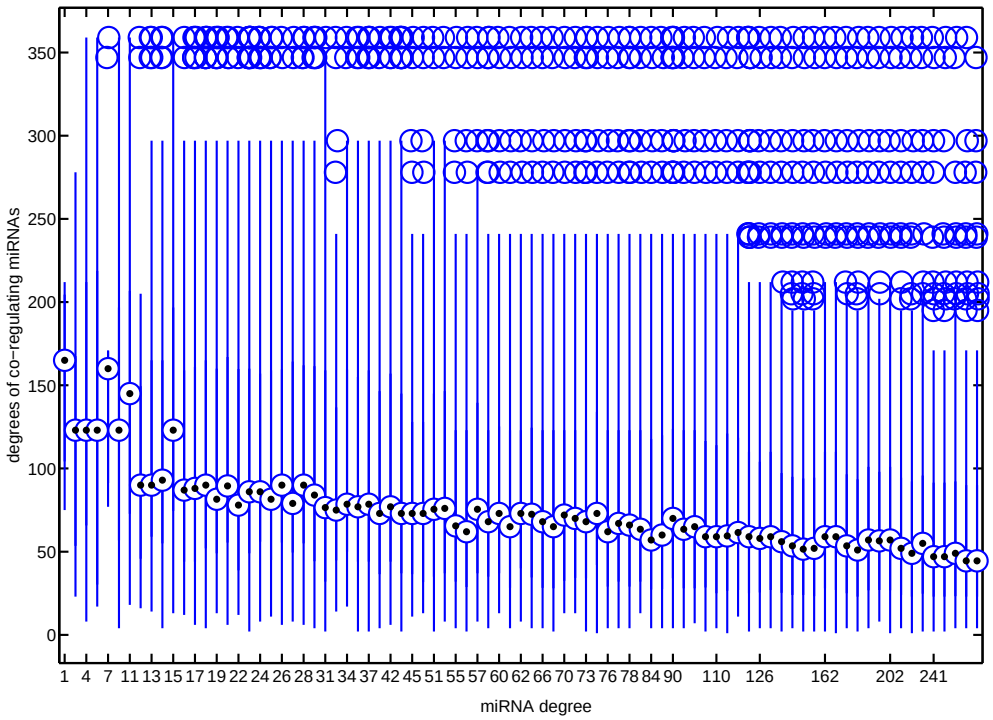

Figure 12: Box-and-whisker statistics for the second neighbour degree correlation in chicken

In order to avoid the problem that was observed with the first neighbour degree correlations, this time the full statistics of a box-and-whisker plot is shown for all degrees of all second neighbours of nodes with a certain degree. For all species the plots look similar as in figure 12, though in this organism (chicken) the tendency can be seen best. The medians of the co-regulating miRNA decay as the degree of miRNA under examination rises. Once again for every second neighbour relationship between two miRNAs the PCCs were calculated. However the correlation between the miRNAs degree and the degrees of its second neighbours can according to table 6 only be named a very slight inverse correlation.

	worm	fly	fish	chicken	cow	human
unweighted	−0.13	−0.13	−0.12	−0.16	−0.15	−0.03
weighted	0.01	−0.01	0.10	0.12	0.09	n.a.

Table 6: Second neighbour PCCs

Furthermore in bigger networks as mouse or human the effect appears in a different manner: the initial medians are high as in figure 12, but quickly the medians reach a plateau and do not decay any further. This means that at a certain degree a miRNA has almost all other miRNAs as second neighbours and therefore the medians stop to decay. This together with the not really high PCCs of table 6 render it questionable, if the phenomenon is not once more just a mathematical artifact. Nevertheless from a purely graph theoretic point of view this topological attribute is well conserved among the species.

In any case the second score where the correlation is weighted according to how often a miRNA is second neighbour of a miRNA would be the biological more meaningful one. However the PCCs for the weighted correlation as shown in 6 are either entirely non-existent or instead of a slight inverse correlation, there is a mild positive correlation. The box-and-whisker statistics did not show a coherent picture and thus it is reasonable to come to the conclusion that there is no biological meaningful preference for miRNAs co-regulating a transcript at all. In order to validate this result by a second measure the assortativities [25] of the nets were calculated (implementation provided by adviser). The assortativity measures, if nodes with a certain degree tend to be connected to nodes with either the same degree (which leads to positive assortativity with the maximum 1) or different degrees. In the last case the graph would be called *dissortative* and the assortativity score would be negative with minimum -1. For the species examined here, the assortativity scores varied between -0.05 and 0.05 and thus this score also assures, that there is no specific degree preference of nodes with a certain degree for having second neighbours with a particular high or low degree.

2.7 Bipartite clustering coefficient

Beyond degree statistics there are advanced network motifs like the so-called *bipartite cluster coefficient* (BCC). In fully connected networks the cluster coefficient of a node is a measure for how close a node and its neighbours are to forming a clique, where all nodes are connected to each other. As in a bipartite graph this can never be the case, the BCC of two nodes can be defined as the overlap of their neighbours [19]. Applied to miRNA regulatory networks for the miRNA partition this would mean, how many of their transcript targets two miRNAs have in common. For the sake of comparability it is suitable to normalise the neighbour overlap of two nodes by either the total amount of neighbours of both miRNAs, the minimum or the maximum of neighbours that one of the nodes has got. Without normalisation there would be no difference between two miRNA hubs sharing a low amount of their neighbours and two low degree miRNAs having a huge percentage of their targets in common.

In figure 13 the BCCs of each possible pair of miRNAs was computed for every organism in the PITA prediction. As normalisation the overlap here was divided by the amount of all targets of either the one or the other miRNA. Thus, the BCCs are generally lower than in the other two normalisations. The distribution of this property score once again is exponential between the very beginning and the really high BCCs. The amount of clustering coefficients grows quadratically with the size of the nets and so

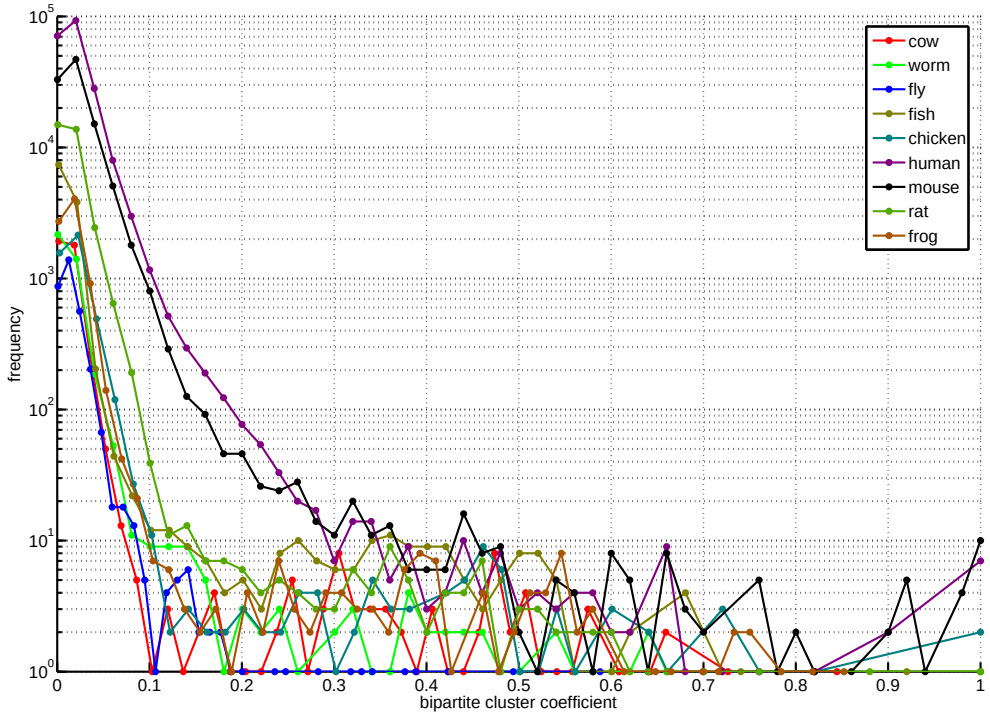

Figure 13: BCCs of almost all species follow an exponential distribution until in the rightmost part, where higher overlaps occur more frequently as product of miRNAs stemming from the same precursor

human and mouse have far more BCCs than the organisms with smaller nets. It does not come surprisingly, that the most frequent BCC is not zero but instead some small value near zero. Instead it is expectable, as most miRNAs have a relative degree between 10 and 20%, that miRNAs have some targets in common just by chance. In order to clarify, what would be expected by chance, the BCCs were calculated for the randomised networks, too. Figure 14 shows, what distribution of percental overlap would have been expected in randomised networks with the same degree distribution, although it is necessary to keep in mind that these randomised networks are but one sample. There is a much clearer exponential trend in the distributions. In general the BCCs are much lower than in the real world nets. A look on the x-scales of both plots reveals, that the BCCs in the real nets are in general far higher than those in the randomised networks. Of course it is expectable that some very high BCCs are achieved due to the fact that especially for organisms with many already explored miRNAs there are more or less redundant miRNAs. For example hsa-miR-let7a and hsa-miR-let7c are almost identical, just differing in one nucleotide far away from the seed region. In consequence their BCC according to PITA is comparably high with 0.66. These related miRNA explain the fluctuations in the right hand end of figure 13 well.

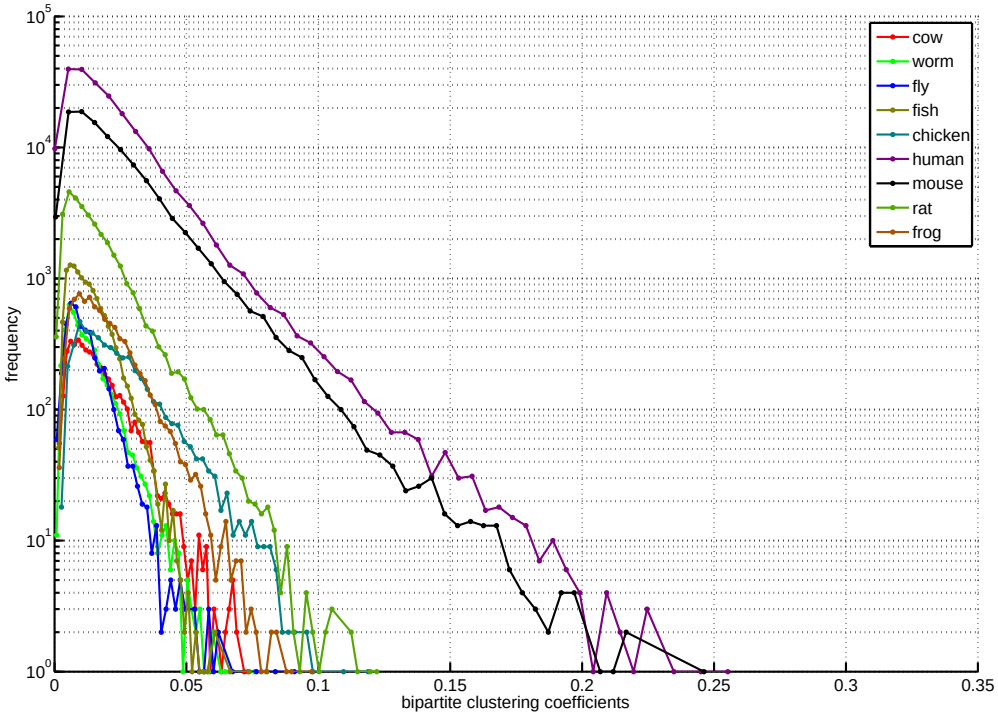

Figure 14: BCCs in randomised nets show a more well formed exponential distribution and generally lower BCCs

However it is still astonishing that even in the exponential part of the distributions in figure 13, which now can be considered the random part of the distributions, reaches further than in the random nets. Especially in mouse and rat the BCCs being higher than those that are achieved in randomised nets, seem to keep their exponential property but with a changed slope. This together with the fact of generally higher BCCs in the real world nets provides indices that there might be a modularity beneath the trivial modularity caused by simple sequence similarity of related miRNAs. It would be interesting to see, whether these modules could be revealed, if the analysis would be extended into the dimensions developmental time and tissue location.

3 Inter-species network analysis

The following chapter aims to compare the networks on a deeper level. When homology information is added to the networks every link can be compared among the species as far as it is conserved. Thus, conserved interactions can be compared to the non-conserved. Therefore chapter 3.1 at first explains the thoughts necessary in order to perform a homology mapping of both partitions. The following chapter is dedicated to the question, how well the relative degrees are conserved among homolog miRNAs.

Chapter 3.3 covers the comparison between the amount of conserved and non-conserved target sites among the species. At last in chapter 3.4 the graphs are enriched with functional annotation in a short attempt to examine, how the functions of the miRNAs changed throughout evolution.

3.1 Mapping the partitions

Concerning the relative degree distributions it has been shown so far, that the shape of the distributions is well conserved. The general higher relative degrees in the younger organisms could be well explained by 3'UTR lengths. Nevertheless the distributions only yield information about the entity of all miRNAs and not about how the amount of targets of a single miRNA changed through various speciation events. Currently, it is assumed that miRNA regulatory networks have undergone *extensive rewiring* during evolution [26] and mainly miRNA target relationships with vital functions have been deeply conserved. It would be thinkable, that many target sites on 3'UTRs of a miRNA hub in worm mutated and left a more specific single-purpose miRNA or vice versa. For this analysis the area of classical network analysis with its whole graph properties needs to be left behind and the networks must be rendered comparable link-by-link.

For enabling this kind of comparability there is the need for a homology mapping of both partitions (although for the first test actually it is not necessary to map the transcript partition) and for all organisms. Homology mappings for the transcripts are not difficult to acquire as many groups already have treated this problem with different approaches. However not all of the homology mappings available cover all of the species under examination here and most of them use for their prediction the Ensembl gene identifiers. Thus, it is particular easy to find a homology mapping with good benchmark results [27]. Although currently InParanoid receives best critics, it is originally intended for two species comparison. The approach of creating a clustering with MultiParanoid [28] for all of the nine organisms was neglected, as it turned out to be too time consuming and presumably it would be rather error prone, too. Instead, for the Ensembl gene set OrthoMCL [29], which uses a Markov cluster algorithm, was applied. A mapping between the Ensembl and RefSeq gene set would not be possible without a huge loss of information, as both sets are far too different. Thus, a RefSeq tailored homology mapping needed to be found and after some recherche only HomoloGene [30], provided by the NCBI that also takes care of the RefSeq gene set, could be determined as possible candidate.

For the mapping actually necessary for the first analysis – a comparison of the relative degrees for each miRNA homolog cluster – homology information about miRNAs was needed. No web resource or group could be found that already have solved the problem of clustering mature miRNAs into families. Clustering the miRNAs by their similarity among the orthologs in the organisms would have been the intuitive bioinformatic approach. But the obvious problem hereby is, that miRNA with analog regulatory activity need not to be ortholog. First of all it is not really simple to define what a family of miRNAs should be. For instance the well performing algorithm TargetScanS [31] considers all miRNAs sharing the same seed as stemming from the same family. A look in

the matrix of clustering coefficients for PITA and TargetSpy reveals, that this mapping would be way to coarse for the two algorithms. As mentioned in chapter 2.7 the clustering coefficients of two miRNAs only differing in one nucleotide can be as low as 0.66. When the union of all targets of all miRNAs with the same seed are combined into the collective of targets regulated by this family, this would mean: as the number of family members grows, so does the amount of targets until with enough family members almost all transcripts are targeted by the family. Best comparability among species is achieved, when only miRNAs are considered to be homolog that share the exact sequence. Then the problem of having too many targets for a family vanishes, too. The downside is an immense loss of sensibility that would only allow for comparison between human, mouse and rat.

However, there is a second possibility that is commonly utilised [32]: a family mapping based on homology knowledge about the miRNA's precursor. In fact the naming of the miRNAs is based upon this precursor homology. The name mapping itself cannot be applied without significant errors, as for instance bta-miR-98 and dre-let-7h are identical except for one nucleotide at the less important 3' end. Fortunately, this clear sequence homology is covered by the miRBase precursor homology information. This mapping was taken as initial point. As information about genomic origin is not available for all miRNA and the precursor mapping does not resolve whether the 5p- or 3p-arm of the miRNA is in a homology cluster, manual post-annotation was necessary. This less restrictive mapping was created with steadily having an eye on sequence conservation. For all miRNAs in one cluster it was required that the seed is exactly conserved with the sole exception that a shift by one nucleotide is accepted (this means for example an extra base at the beginning of the sequence). This clause was introduced, as it is not unalterable dogma that the seed ranges from bases two to eight. Instead the seed might be much more multi-facetted and it is not fallacious to consider for instance bases one to seven the seed region [33]. However if more than three bases differed, the differing sequence was considered a separate cluster.

3.2 Conservation of miRNA-specific relative degrees

As mentioned before, the first thing to be examined in the inter-species networks was the conservation of the relative degrees of each of the miRNAs. Figure 15 shows a pairwise comparison between rat and mouse for the TargetSpy prediction. Each symbol in the plot represents a miRNA. MiRNAs that are located at the x- or y-axis (thus having degree zero in the respective other organism) are species-specific miRNAs for which according two the mapping no homolog miRNA exists. However the assumption that hubs are more likely to be conserved among species - which could be derived from all plots mapped with the less restrictive mapping - is misleading. This is because of the less restrictive homology mapping that leads to higher relative degrees for bigger families as explained in chapter 3.1. Thus, the immense relative degree (~ 0.53 in mouse) of the let-7 family mainly has two reasons: first, this is an utterly big family with nine members and second for an unknown reason, the clustering coefficients of the let-7 family members in TargetSpy are near noise level, though normally the homology

information of the family mapping brings high clustering coefficients with it. Thus, the proposition that miRNA hub sequences are better or well conserved cannot hold.

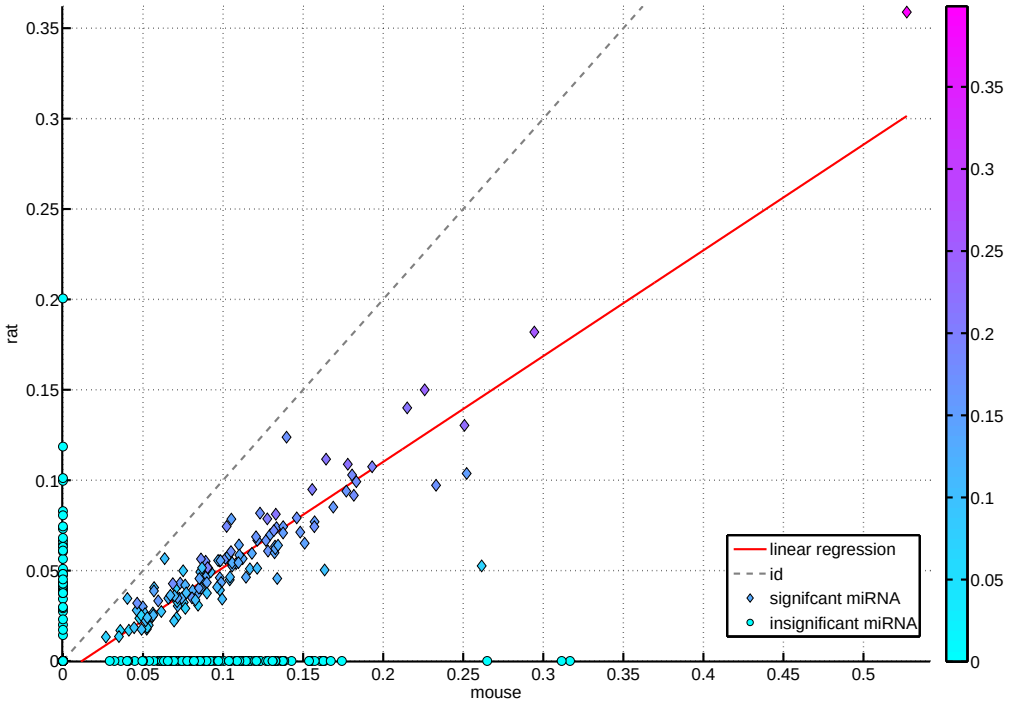

Figure 15: Comparison of relative degrees among rat and mouse for TargetSpy shows good conservation of relative degrees

The actual hypothesis, that the relative degrees of the particular miRNAs are conserved, can be approved. It seems, as the relative degrees are not just well conserved for either hubs or low degree miRNAs but instead it generally holds that the relative degree in one organism resembles the relative degree in other organisms. With a PCC of 0.93 the data in figure 15 are very strong correlated. The overall trend of the relative degrees (see the linear regression) differs from the bisectrix varyingly for each organism. This represents the general higher relative degrees in evolutionary younger organisms caused by longer 3'UTRs as discussed in chapter 2.5.

As expected the relative degrees are even stronger correlated, when the more restrictive mapping is applied. When miRNA homology clusters are of different sizes among the species, this strongly influences the relative degrees and thus the PCC is reduced. The relative degree conservation also holding for the restrictive mapping assures that the high PCCs are not just mere artifacts of some big equal-sized families. Interestingly, in comparison to the TargetSpy prediction the relative degrees for PITA are not just very strongly correlated, but they almost form a stringent line. This once more seems to come from the two different approaches of the algorithms. PITA is more seed-depending and

for a target site high seed identity is a necessary condition. When there is high 3'UTR sequence similarity between for instance mouse and rat, then the PITA prediction will yield very similar results, while in TargetSpy the determination of a target site still depends on some other features. However, as both predictions yield these high PCCs, it can be stated that relative degrees are well conserved among species. If TargetSpy produced a correct and PITA was entirely noisy, then the results of PITA would at least allow to say, that the amount of seeds is very well conserved on the 3'UTRs, but through evolution some of them became inactivated.

Another question that comes up quite naturally: are the relative degrees less conserved for species that are evolutionary more distant? To clarify this issue, a pairwise distance measure is needed. The PCC with its big influence of values that are distant from the main field is perhaps not the best option for this purpose. Instead the root mean spare deviation (RMSD) was applied. The RMSD measures the distance between two vectors, thus it can be used for pairwise species comparison. In this case the vector values are the respective relative degrees of the homolog miRNAs. However, if the RMSD was applied without modification the problem would arise, that human and worm would be most distant per default, as their tendency (linear regression) dissents most from the bisectrix. As it is not this, that should be measured, but instead something like the deviation from the linear regression, the relative degrees of a species at first must be normalised by their mean. For the relative degree vectors $\mathbf{x}$ and $\mathbf{y}$ of length n and with means $\mathbb{E}[\mathbf{x}]$ and $\mathbb{E}[\mathbf{y}]$ this yields the formula:

$$\mathrm{RMSD}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{\sum_{i=1}^{n} ((x_i - \mathbb{E}[\mathbf{x}]) - (y_i - \mathbb{E}[\mathbf{y}]))^2}{n}} \cdot 100 \tag{2}$$

Here only species from the less restrictive mapping were compared, as otherwise many species are not comparable at all as they share too few fully conserved miRNAs.

Table 7 shows the normalised RMSD values for all organisms except frog, where no appropriate transcript ortholog mapping was available. Though there are only few miRNAs comparable for worm and fly, they are most distant from other organisms according to this score, while comparably close to each other (and to fish, too). The four mammals are rather close to each other and surprisingly, chicken and fish are not that distant - the gap to worm and fly is much higher. Suitably, mouse and rat are among the less distant organisms. So far, according to this overview of RMSD values it is reasonable to state, that the conservation of relative degrees is stronger among evolutionary close organisms and vice versa.

The hypothesis about the PITA prediction being unfeasible mentioned above does not comes without reason. In order to validate the results from this and the succeeding examination, a PITA prediction with random miRNAs was created for human, rat, mouse and worm. A hundred miRNA were randomly created from the four letter alphabet not regarding the GC distributions of the miRNAs. Target sites for these fake miRNA then were searched with the PITA tool using the suggested parameters. The results come astonishingly: the relative degree plots as seen in figure 15 look almost identical for the random miRNAs. The relative degrees are at least as correlated as for the verified miRNAs, yielding PCCs around 0.99. Even the height of the relative degrees remains

	fly	fish	chicken	cow	human	mouse	rat	
worm	4.2	13	12	n.a.	11	14	9.7	TargetSpy
	0.7	1.6	4.0	3.6	6.2	4.1	4.3	PITA
fly		8.8	9.0	n.a.	8.9	9.5	6.5	TargetSpy
		1.1	2.6	2.4	3.8	2.6	2.7	PITA
fish			3.6	n.a.	3.4	3.8	3.5	TargetSpy
			1.3	1.2	1.9	1.2	1.3	PITA
chicken				n.a.	3.7	2.7	2.7	TargetSpy
				1.3	1.9	1.3	1.3	PITA
cow					n.a.	n.a.	n.a.	TargetSpy
					1.3	0.5	0.7	PITA
human						4.5	3.2	TargetSpy
						1.4	1.2	PITA
mouse							3.0	TargetSpy
							0.48	PITA

Table 7: Normalised RMSDs among species for PITA and TargetSpy (less restrictive mapping applied)

in the same level although the hubs generally have a bit lower relative degrees. This can be considered okay, as in the real world case there are more than 100 miRNAs for the comparison. So the chance is higher, that some hubs achieve higher relative degrees. However, as these fake miRNA are sequences without biological function, it could have been expected that their relative degrees would be lower.

There are two possible explanations for the fact that a prediction of random miRNAs resembles the real world case that much. Either in fact PITA is not potent of finding biological valid target sites and only represents the conservation of the 3'UTR. For evolutionary distant species this would mean that their less conserved 3'UTRs would result in greater variance of their relative degrees. The results of figure 16 fit well into that picture as the PCC is only 0.66 for these random miRNA comparison among the distant organisms human and worm. As the tool itself was not available, no random TargetSpy prediction could be created. If it is assumed, that a PITA prediction without conservation filter has modest false positive rates, then a second thought would explain the phenomenon: that even random 22mers could theoretically act as miRNA, if they only were expressed or induced into the organism or a RISC favouring the 22mer's energetic and sequence-specific conditions would be available. Anyway this does not falsify the statement that the relative degrees of each miRNA is a conserved network property. At least it is astonishing, why there seems to be no need that possible target sites for random sequences should be removed by mutation. Maybe the answer is really as simple as: "because the random sequences do not occur in reality" respectively that RNA sequences have to offer a lot more features in order to produce a valid miRNA [34].

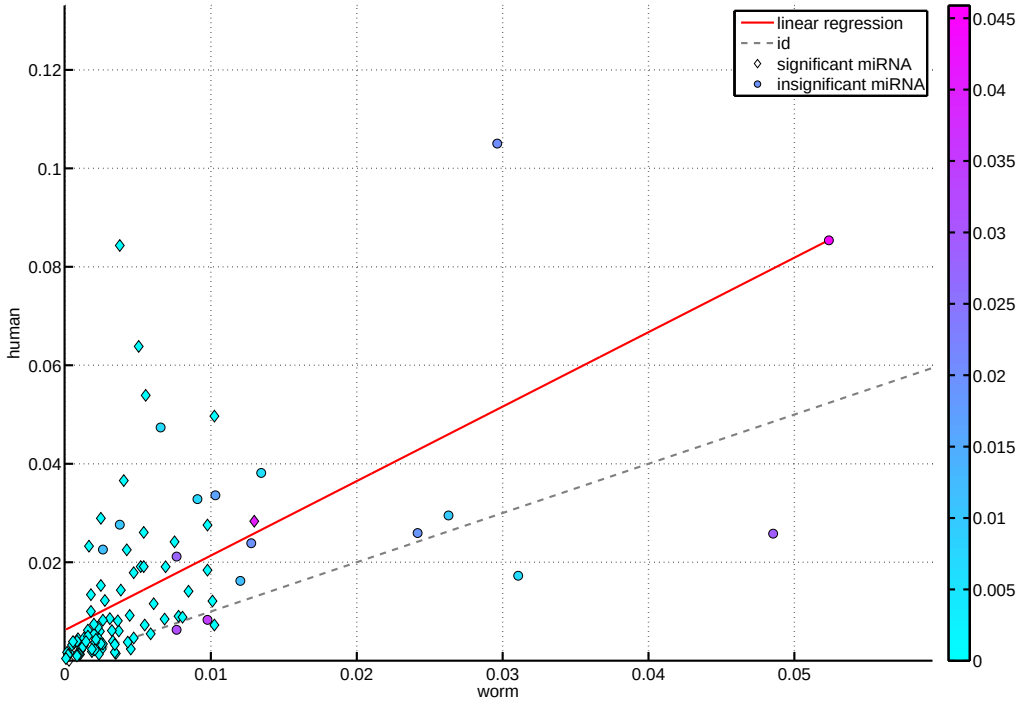

Figure 16: Comparison of relative degrees for 100 fake miRNAs among worm and human. As in the natural nets there is fewer conservation.

3.3 Amount of conserved target sites

The figures 15 and 16 already contain information that has not been brought up yet. As so far all analysis aimed to abstract the very miRNA target site regulatory relationship into graph properties. But conservation-free miRNA regulatory networks – as far as they are correct – also provide the chance of a comparison between conserved and non-conserved target sites. In detail, the next examination of the homolog miRNAs is to compare the amount of conserved target sites against the amount of species-specific target sites. Thus, in figures 15 and 16 the colour coding of the miRNAs shows the amount of target sites conserved divided by the amount of target sites for that miRNA occurring either in the one or the other species (ergo this measure resembles the BCC) among the transcripts for which an homology mapping was available. A glance in figure 15 makes clear that in general these conservation ratios are rather low. The general tendency among the species shows that low degree miRNAs have utterly low conservation ratios (with many of them having no conserved target sites at all) and as the degree rises, hubs tend to show rather decent conservation ratios.

However the last applications of this BCC resembling scores have shown that a normalisation by the amount of target sites occurring in either the one or the other vector, can be misleading. Two vectors with constant amount of non-zero entries, can be compared

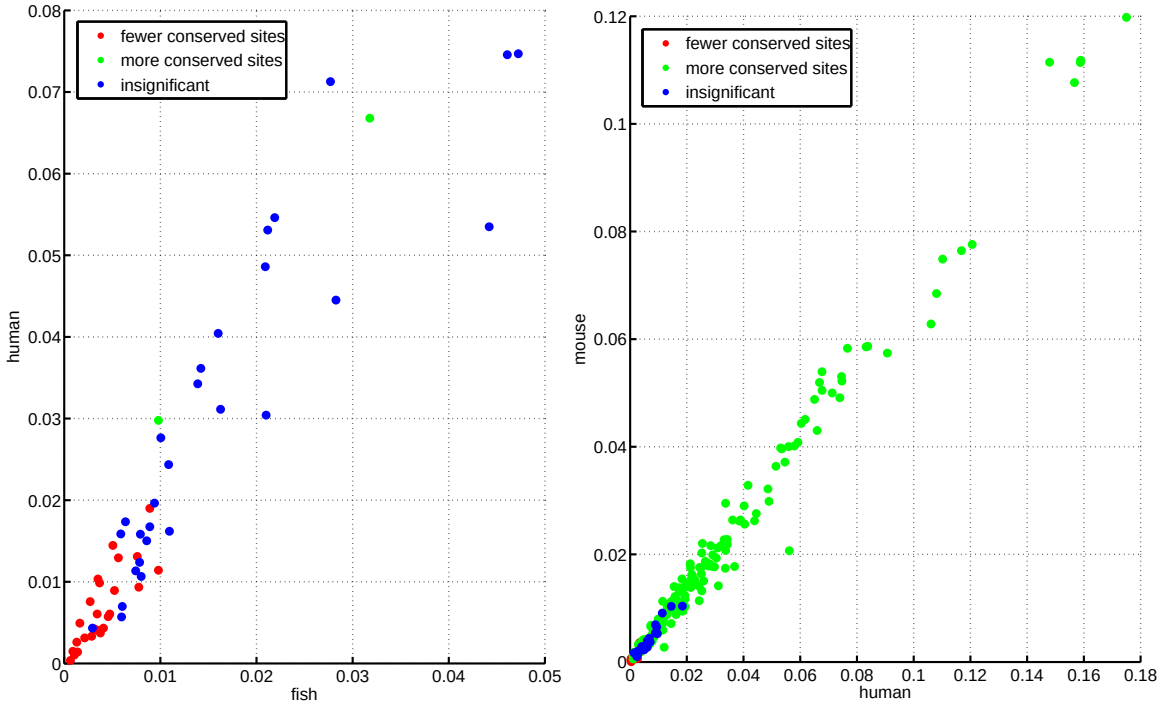

Figure 17: Significance of conservation ratios for each miRNA in human, mouse and fish. While in the human-mouse comparison more target sites are conserved than expected, in human-fish there are fewer conserved target sites.

this way. Then with rising divergence every time the denominator is decremented by one, the divisor becomes incremented by two. Thus, in order to avoid premature conclusions about the amount of orthologs conserved being low, it was about to test whether the amount of conserved target sites was higher or lower than would be expected by chance. For a significance test with the alternative hypotheses of the amount of conserved target sites being higher or lower than expected by chance, an appropriate null model was needed. The first null model to be applied for the tests was the assumption that the amount of ortholog transcripts being targeted by a miRNA follows a hypergeometric distribution. Abstractly, the hypergeometric distribution yields the probability that with k draws without replacement from a basic population of n objects containing m objects with a special property, l objects with the special property are drawn. Assumed, that miRNA target transcripts randomly, in the present scenario the hypergeometric distribution determines the probability for targeting l transcripts from the set of ortholog transcripts with size m, when all in all the miRNA targets k transcripts and could target any of the total of n transcripts (which are either orthologs or transcripts only available in one of the species). The result of this significance test was, that the null hypothesis could be rejected in favour of the alternative hypothesis ("conservation

ratio is lower than by chance") almost for any miRNA in any pairwise comparison with extremely low P-values. These better-than-expected results rendered it questionable, if this null model was the adequate one for the test. With conserved relative degrees the miRNAs do not really have the possibility to target just transcripts of the organisms and none of the other one. Thus, some experimentation was performed with applying the hypergeometric distribution on two basic populations (each the amount of transcripts of only one organisms) with amount of draws equal to the number of targets of each single organism. However the immensely significant P-values were not about to change. Obviously, the problem with the null model is, that it is not capable of representing, how probable it is to draw the same objects two times when the amount of draws and the basic population changes.

When a problem is too complex for simply applying a distribution the right way, it is often appropriate to use the Monte Carlo approach. For the present case, this means that for each miRNA 1000 random samples are created. A sample consists of simulating for each of the two species that the homolog miRNA would target the same amount of transcripts at random. From such a sample then can be extracted, how many of the set of orthologs are regulated by the miRNA in both species. The 1000 samples yield a distribution that allows to assign a P-value for each miRNA in a pairwise comparison. An implicit assumption that could corrupt this null model is the ignorance of the degree distribution. If miRNA are more likely to target ortholog transcripts in general (which does not mean that the target sites also need to be conserved) this should be included in the null model. To clarify, if this is the case, 3'UTR lengths where compared between transcripts occurring in the homolog mapping and those who do not. As shown in chapter 2.5, this is a necessary condition for higher degrees. However the distributions did not show a difference worth mentioning, neither for the HomoloGene nor the OrthoMCL mapping. Solely in rat the 3'UTR lengths were notably higher for both mappings, which however was not regarded in the null model.

MiRNAs for which the null hypothesis ("among orthologs as many targets are conserved as could be expected by chance") could be rejected on a confidence level of 99% are already marked in figures 16 and 15. However the picture gets clearer after a look on the plots of figure 17, which represent the collective of all plots created this way well. For evolutionary close organisms such as human and mouse there are more target sites conserved than expected by pure chance. This holds mostly for hubs, the lowest degree nodes tend to become insignificant or if there is no target site conserved at all, they have often fewer conserved target sites, as this case might not have been achieved by one of the 1000 samples. The more distant two organisms are, the higher the degree of the miRNAs that are insignificant or have fewer conserved target sites than expected by chance. The conclusion that can be drawn from this examination contains interesting implications about the possible evolution of miRNA regulatory networks. The mammals among each other have significant more conserved target sites than expected. This implies that their nets are closely related. All other nets except the closely related mammals have among each other as much in common as expected by chance. This indicates that the miRNA regulatory nets function very similar at a taxonomic level of the mammals but have undergone massive reorganisation since the speciation of nematodes, insects and birds.

For completeness it needs to be mentioned that the 100 random miRNAs also follow the same trend as described in the section above. Thus, the significance plots for figure 16 shows most miRNAs having fewer conserved target sites than expected by chance. Obviously this tendency is anchored in the 3'UTRs. This appears particularly reasonable, when it is assumed that mainly the 3'UTRs evolve and the miRNA underly higher evolutionary pressure to keep their sequences. This is because of changes in miRNA sequences would have a large impact on the miRNA regulatory net. However, if the PITA prediction is right, there is no need for putative sites that could be targeted by random sequences to be removed through mutation.

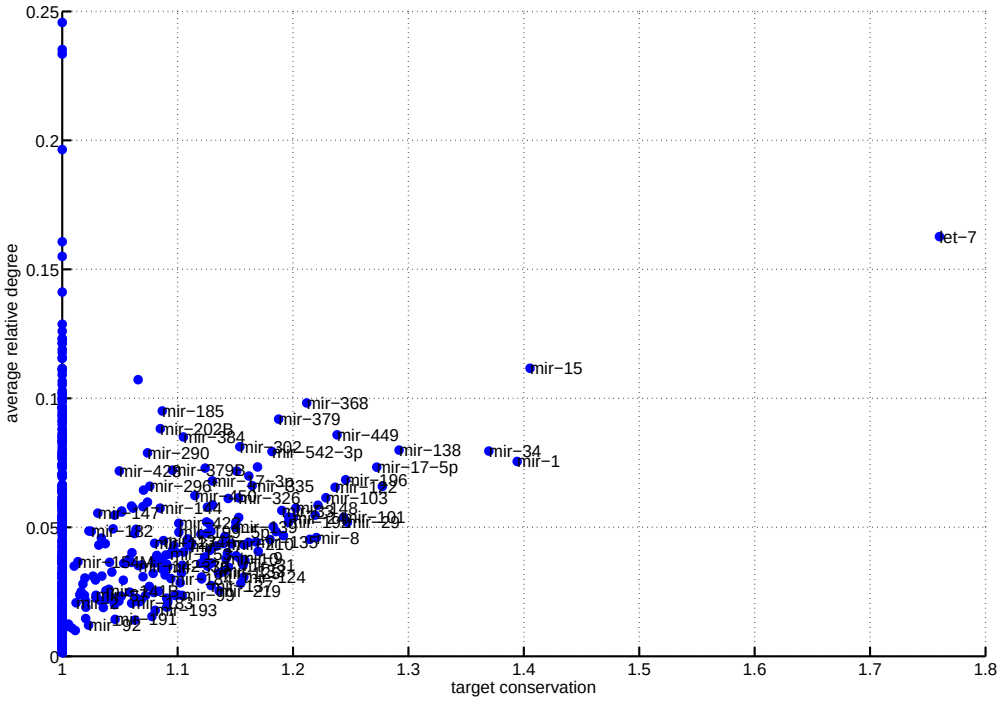

Figure 18: Correlation between conservation of target sites and relative degree in TargetSpy

From the significance and relative degree plots the hypothesis was deduced, that miRNA hubs tend to have more conserved target sites. However, this was derived from comparing many pairwise plots. In order to represent this tendency among all species at once and thus providing further validation, figure 18 was created. It shows the correlation between the relative degree of a miRNA (averaged among the species in which homologs of the family occur) and a conservation score. For each miRNA this score measures in how many species the regulatory relationship between homolog miRNA and transcript ortholog are in average conserved (see equation 3 for the average

conservation of miRNA α with the amount of the union of all orthologs z).

$$\text{average\_conservation } (\alpha) \ = \frac{\sum_{i=1}^{z} \#\text{species interaction } \alpha\text{-}i \text{ is conserved}}{\#\text{interactions between } \alpha \text{ and any } z} \tag{3}$$

Thus, a score of 2 would signify that the mean target sites is conserved among two species. With a PCC of 0.66 the data in figure fully approve the above stated thesis. Interestingly, there are the let-7 and mir-1 family among those with the highest conservation score. For these two miRNA families the evolutionary most distant conserved targets were found [35]. It is not surprising that the best conserved target sites in both predictions are regulated by these two and the mir-34 family. The transcripts regulated by these deeply conserved relationships included the growth factors CTGF and TGBR1, the two oncogenes RRAS2 and RAB10, a membrane protein (TRAM1) and some genes of which the products are involved in RNA processing (TPP, PTBP1 and SERBP1). These results come quite natural as the repression of oncogenes is essential in any case, also as the regulation of proteins as important as those involved in the mRNA processing. That growth factors appear among these most conserved target sites coherently fits to statements according to which miRNAs play a vital role in development.

3.4 Retention of the functional profile of miRNAs

The most important aspect of miRNAs still is their function. As shown in chapter 3.3 many target sites are conserved among closely related species and few among distant species. But what does this mean for the functionality of miRNAs?

In order to gain insight into this issue, the change in functionality of a miRNA was compared to how distant the homolog miRNAs are in their regulatory profile. In pairwise species comparisons for each miRNA family the RMSDs of their regulatory profiles were computed. In contrast to the RMSD definition from equation 2 this time there was no normalisation by any average $\mathbb{E}$. Here, the regulatory profile of a miRNA is defined as a binary vector $\mathbf{x}$, with $x_i = 1$, if the i-th ortholog transcript is regulated by the miRNA family and a null entry elsewise. This way the RMSD of the regulation profiles in two species can range from 0 to 1.

Change of function of a miRNA means that the homolog miRNA represses in one organisms transcripts having different functions than those repressed in the other organism. In order to compare this, functional annotation for the transcripts was obtained from the GOA project's website [36]. For species without GOA annotation this data was completed by annotations from Gene Ontology [37]. The Gene Ontology (which is also used by GOA) provides a fix vocabulary called GO-terms for describing (amongst others) the biological processes a gene product is involved in. The GO-terms are organised in a directed acyclic graph (DAG) and walking along the edges each node becomes more specific. There is a huge amount of GO-terms that might only vary slightly in their meaning. Therefore it is suitable to generalise all the GO-terms into few categories. The Gene Ontology provides a dedicated subset for this task: 53 so-called generic GO-slim-terms (including terms like "response to stress", "cell death" or "transcription"). All of the specific GO-terms were mapped to the more general GO-slim-terms. The mapping

yields a binary bipartite matrix specifying the relationship between GO-slim-terms and transcripts. In order to provide functional annotation for miRNAs for each species a $n \times m$ matrix $(a)_{ij}$ created, where n is the amount of miRNA and m the amount of GO-slim-terms. With x, y and z being transcripts, equation 4 defines the values of $(a)_{ij}$:

$$(a)_{ij} = \frac{|\{x|\ x \text{ is target of } i \text{ and has annotation } j\}|}{|\{y|\ y \text{ is target of } i\}| \cdot |\{z|\ z \text{ is annotated with } j\}|} \tag{4}$$

The amount of targets of a miRNA with a certain annotation must be normalised by two factors. The first is the amount of targets of the miRNA in order to avoid that miRNA hubs are functionally more active by default. The other factor is the amount of times that the GO-slim-term occurs in the mapping. This removes the perturbation that terms which are often annotated would bring with them. The matrix normalised this way specifies for each miRNA how specifically it influences biological processes.

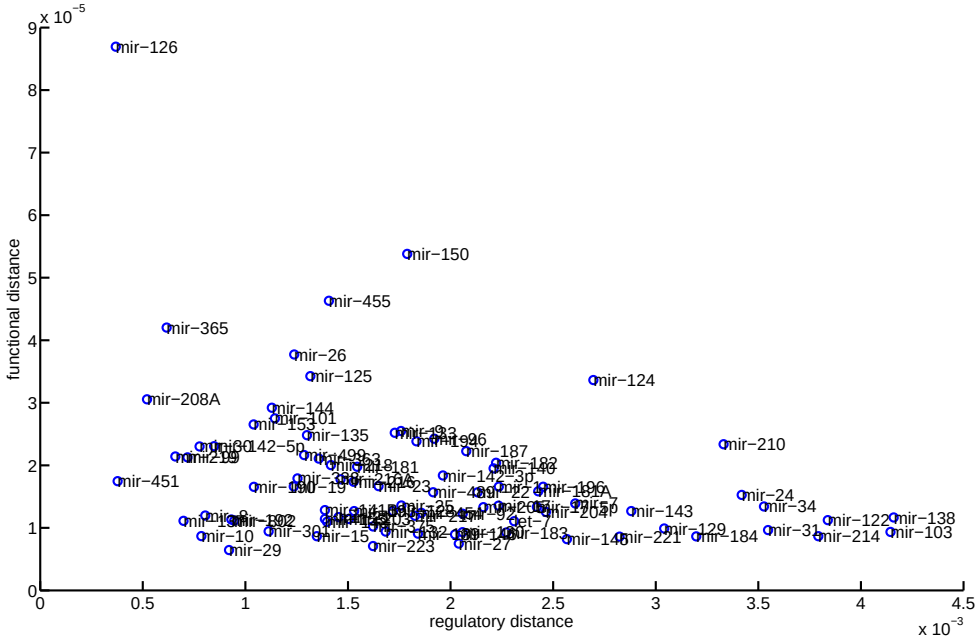

Figure 19: Homolog miRNAs having few target sites in common (regulatory distance) mostly show same functional distance as miRNAs with many conserved target sites

The functional profiles of homolog miRNAs from two species are compared using the RMSD which results in a functional distance. For each homolog miRNA among two species the functional and regulatory distance were calculated. Figure 19 shows the comparison for miRNAs of human and fish in the PITA prediction. In all plots created this way, as in the example shown here, most of the miRNAs remain functionally similar. Their RMSDs are in the same order of magnitude as mean and median of all entries of a matrix. This would imply that miRNAs tend to retain their functionality through

evolution. Figure 19 furthermore shows that the miRNAs even keep their functions, when they loose target sites and acquire new ones through evolution.

This scenario with the 53 GO-slim-terms might be over-simplified. Thus, instead of mapping all GO-terms to the GO-slim subset, two additional mappings were tried. Once the terms were mapped to the subset that is achieved after two edges from the root ("biological process") term, the second subset results from three steps starting in the root. The subsets achieved this way are more coarse (containing 221 and 1793 terms), however the plots showed comparable results.

In fact, the relationship between functional and regulatory distance in most plots is a inverse correlation with PCCs ranging from -0.1 to -0.6. The conclusion of these PCCs would be, that miRNAs have very similar functions, if they and their target sites mutated heavily and the other way round. This implausible tendency mainly is the product of an artifact, that once again involves the degree of the miRNAs. In figure 19 the outlier most distant from the main field is mir-126. In human this miRNA family only has 10 targets, ergo mir-126 is a low degree miRNA. A degree as low as this one naturally yields a low regulatory RMSD. However the few transcripts regulated in either fish or human seem to have very different annotations. This way low degree miRNA distort PCCs such that they maybe should not be regarded really biological meaningful. More important for the hypothesis (retention of function with rising regulatory distance) is, that the main field with functional distances between the extremes is poorly correlated. However, it could be that through the manifold mappings and by reducing the functional annotation (intended for many transcript) to few miRNAs a huge noise level was produced. Then naturally the functional distances between homolog miRNAs would also be noise-distorted and would be most likely rather low.

4 Summary and outlook

This thesis began with clarifying how many organisms' miRNA regulatory networks can be compared in a useful manner. As the miRNA research field evolves rapidly, it is probable that soon widest parts of the miRNA partition are identified for more organisms. However, the sequencing of higher organisms still takes time and therefore it is hard to say, when the analyses performed here could be repeated in larger dimensions.

While predicting networks with the PITA tool, a GC-bias was noticed. MiRNA having a more GC-rich sequence achieve higher energy scores and are more likely to be predicted as hubs. As many prediction tools concentrate much on the miRNA-transcript hybridisation energy, this bias exists in many target predictions. In any case for PITA it leads to more false positive predictions for GC-rich miRNAs and therefore it could be, that the degree of the biggest hubs is over-estimated. Machine learning prediction tools like TargetSpy have fewer problems with this bias. For the more sequence- and energy-based prediction tools like PITA a general approach, reducing this influence would be desirable.

The discussion of the bipartite densities came to the conclusion that higher developed organisms like the mammals have denser networks than the others. This could mean that their miRNA regulatory networks might be higher developed as well. Surely the

developmental programs of mammals can be expected to be more complicated than those of nematodes for example. Thus, it sounds reasonable that much more miRNA activity is needed in higher organisms.

In spite of the strong variations in the bipartite densities, the comparison of miRNA and transcript degree distributions showed that the networks are from that point of view very similar. An exponential transcript degree distribution and the scale invariance on the miRNA side were found out to be intrinsic properties of these networks. They strongly differ from the binomial distributions that are observed in random nets. Furthermore it turned out that the amount of target sites a 3'UTR bears mostly depends on its length. However, it could not be solved, why worm had a target site density on its 3'UTRs that deviates such strongly from the rest. As the organism is smaller, it possibly had pressure to fit to different natural circumstances or there are economical reasons.

The examination of the first neighbour degree correlations revealed that the higher the degree of a transcript, the lower is the average degree of the miRNAs regulating it. As the same result was also observed in randomised nets, the biological significance of this finding remains doubtable. From a graph theoretic point of view, this is a topological attribute, that occurs in all nets, but the comparison to the random nets showed that it also occurs in any net with the same degree distribution. The analysis of the second neighbour degree correlations came to the conclusion, that there is no notable tendency of miRNAs with certain degrees co-regulating transcripts. However, the fact that PCCs in randomised nets were higher, allows for thoughts about whether there is a biological meaning behind this fact. Possibly it could be important for the regulation mechanisms that miRNAs of very different degrees co-regulate transcripts.

When the BCCs of the networks were analysed it was a priori clear that there might occur high BCCs for miRNAs stemming from the same precursor. However, it was surprising that the randomised networks yielded generally lower BCCs. There seems to be more modularity in the natural networks than can be expected by chance. This could be a good starting point for further analyses that try to identify such modules. Maybe if information about tissue and developmental stages was added, interesting relationships could be revealed. For these tasks binary bipartite networks which were used in this thesis barely seem to be the appropriate data structure.

For the mapping two well-defined homology mappings for mature miRNAs were created. However, literature still lacks here a coherent definition of a miRNA family and a common standard would help to avoid mistakes. For the analysis done in the second part of the thesis the mapping performed well except for the let-7 family in TargetSpy. With its help it was shown that the relative degrees are strongly conserved among closely related species. The more distant two species are in the tree of life, the more different are the relative degrees of its miRNAs. This means, that the rewiring of the miRNA regulatory networks through speciation events is not such strong that for instance a miRNA hub has only few target sites left in the new organism. Instead the nature of a miRNA remains mostly the same.

A similar picture was observed when examining the amount of conserved target sites. It was proven in a statistically significant manner, that in mammals more target sites are

conserved than expected by chance. The more distant organisms are, the fewer homolog miRNAs pass this test. It was concluded that the miRNA regulatory nets are rather similar on the taxonomic level of mammals, but for organisms of different clades there seems to be only few similarity. In the same context it was shown, that the target sites of miRNA hubs are in average conserved in more organisms than this is the case for low degree miRNAs. This indicates, that not the target sites of specific miRNAs need to be conserved but instead the general influence of miRNA hubs is important enough for justifying conservation.

At last it was determined which kind of functions the miRNAs influence with their repressive activity. It could be shown that even if homolog miRNAs have only few conserved target sites and acquired many new ones, the miRNAs still take care of the same function. This could imply that the new target sites are acquired in 3'UTRs of transcripts with similar functions as the 3'UTRs where target sites were lost. It motivates the hypothesis that miRNAs during evolution can change their targets but with the focus on performing a similar function. This would strengthen the hypothesis that mainly the 3'UTRs evolve, while miRNAs only rarely mutate.

References

[1] N. Bushati and S. Cohen. miRNA functions. *Annu Rev Cell Dev Biol*, 23:175–205, 2007.

[2] Y. Sun Lee and A. Dutta. microRNAs in cancer. *Annu Rev Pathol*, Sep 2008.

[3] R. C. Lee, R. L. Feinbaum and V. Ambros. The C. elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, Dec 1993.

[4] G. Martin, K. Schouest, P. Kovvuru and C. Spillane. Prediction and validation of microRNA targets in animal genomes. *J Biosci*, 32(6):1049–1052, Sep 2007.

[5] D. Baek, J. Villén, C. Shin, F.D. Camargo, S.P. Gygi and David P Bartel. The impact of microRNAs on protein output. *Nature*, Jul 2008.

[6] R. Shalgi, D. Lieber, M. Oren and Y. Pilpel. Global and local architecture of the mammalian microRNA-transcription factor regulatory network. *PLoS Comput Biol*, 3(7):e131, Jul 2007.

[7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002.

[8] S. Griffiths-Jones. The microRNA registry. *Nucleic Acids Res*, 32(Database issue):D109–D111, Jan 2004.

[9] S. Griffiths-Jones, H. Kaur Saini, S. van Dongen and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36(Database issue):D154–D158, Jan 2008.

[10] S. Griffiths-Jones, R. J Grocock, S. van Dongen, A. Bateman and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, Jan 2006.

[11] M. Sturm. TargetSpy. unpub.

[12] K. D. Pruitt, T. Tatusova and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–D65, Jan 2007.

[13] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul and E. Segal. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284, Oct 2007.

[14] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova and B. A. Rapp. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 28(1):10–14, Jan 2000.

[15] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Gräf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kähäri, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J P Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal and S. Searle. Ensembl 2008. *Nucleic Acids Res*, 36(Database issue):D707–D714, Jan 2008.

[16] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler and W. J. Kent. The UCSC table browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493–D496, Jan 2004.

[17] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox and E. Birney. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, 14(1):160–169, Jan 2004.

[18] http://genie.weizmann.ac.il/pubs/mir07/mir07_notes.html.

[19] M. Latapy, C. Magnien and N. Del Vecchio. Basic notions for the analysis of large affiliation networks / bipartite graphs, 2006.

[20] N. Davis, N. Biddlecom, D. Hecht and G. B. Fogel. On the relationship between GC content and the number of predicted microRNA binding sites by microInspector. *Comput Biol Chem*, 32(3):222–226, Jun 2008.

[21] H. Robins and W. H. Press. Human microRNAs target a functionally distinct population of genes with at-rich 3' UTRs. *Proc Natl Acad Sci U S A*, 102(43):15557–15562, Oct 2005.

[22] B. John, A. J Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks. Human microRNA targets. *PLoS Biol*, 2(11):e363, Nov 2004.

[23] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2):192–197, Feb 2006.

[24] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, May 2002.

[25] A. Trusina, S. Maslov, P. Minnhagen, and K. Sneppen. Hierarchy measures in complex networks. *Phys Rev Lett*, 92(17):178702, Apr 2004.

[26] C. S. Chan, O. Elemento and S. Tavazoie. Revealing posttranscriptional regulatory elements through network-level conservation. *PLoS Comput Biol*, 1(7):e69, Dec 2005.

[27] F. Chen, A. J. Mackey, J. K. Vermunt and D. S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383, 2007.

[28] A. Alexeyenko, I. Tamas, G. Liu and E. L. L. Sonnhammer. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22(14):e9–15, Jul 2006.

[29] L. Li, C. J Stoeckert and D. S. Roos. OrthoMcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003.

[30] http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene.

[31] B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, Dec 2003.

[32] S. Roush and F. J. Slack. The let-7 family of microRNAs. *Trends Cell Biol*, Sep 2008.

[33] D. Gaidatzis, E. van Nimwegen, J. Hausser and M. Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007.

[34] C. Yu Chan and Y. Ding. Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. *J Math Biol*, 56(1-2):93–105, Jan 2008.

[35] K. Chen and N. Rajewsky. Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harb Symp Quant Biol*, 71:149–156, 2006.

[36] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler. The gene ontology annotation (GOA) database: sharing knowledge in uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–D266, Jan 2004.

[37] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock. Gene Ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, May 2000.