

Prediction of microRNA targets in *Caenorhabditis elegans* using a self-organizing map

Liisa Heikkinen^{1,2}, Mikko Kolehmainen³ and Garry Wong^{1,2,*}¹Department of Biosciences, ²Department of Neurobiology, A.I.Virtanen Institute for Molecular Sciences, Biocenter Finland and ³Department of Environmental Science, University of Eastern Finland, Kuopio, Finland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs that regulate transcriptional processes via binding to the target gene mRNA. In animals, this binding is imperfect, which makes the computational prediction of animal miRNA targets a challenging task. The accuracy of miRNA target prediction can be improved with the use of machine learning methods. Previous work has described methods using supervised learning, but they suffer from the lack of adequate training examples, a common problem in miRNA target identification, which often leads to deficient generalization ability.

Results: In this work, we introduce mirSOM, a miRNA target prediction tool based on clustering of short 3'-untranslated region (3'-UTR) substrings with self-organizing map (SOM). As our method uses unsupervised learning and a large set of verified *Caenorhabditis elegans* 3'-UTRs, we did not need to resort to training using a known set of targets. Our method outperforms seven other methods in predicting the experimentally verified *C.elegans* true and false miRNA targets.

Availability: mirSOM miRNA target predictions are available at <http://kokki.uku.fi/bioinformatics/mirsom>.

Contact: liisa.heikkinen@uef.fi

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 25, 2010; revised on February 20, 2011; accepted on March 12, 2011

1 INTRODUCTION

MicroRNAs (miRNAs) are an abundant class of small, non-coding RNAs found in diverse organisms (Bartel, 2009; Lee *et al.*, 1993). miRNAs direct the post-transcriptional repression and degradation of mRNAs by binding to their sequence with partially complementary sites and thus impact a variety of biological functions, including cell differentiation, organism growth, development and tumor suppression (Kloosterman and Plasterk, 2006). While the amount of published miRNAs is increasing and a large number of genes is predicted to be regulated by miRNAs, only a small fraction of miRNA–target genes are experimentally verified. For example, there are now 175 *Caenorhabditis elegans* miRNAs annotated in miRBase database (Griffiths-Jones *et al.*, 2008), but only 15 entries for six of these miRNAs in the latest release of TarBase (Papadopoulos *et al.*, 2009). As experimental identification

of miRNA targets is difficult, the computational prediction of the target genes is a valuable tool to investigate miRNA functions and in guiding related wetlab experiments.

Occurrence of perfect seed matches (seed = the nucleotides 1–8 of the miRNA 5'-end) from the 3'-untranslated region (3'-UTR) of the mRNA, evolutionary conservation of the miRNA–target relationship and the free energy ΔG of the miRNA:mRNA duplex are the most often used determinants in miRNA target recognition (Brennecke *et al.*, 2005; Enright *et al.*, 2003; Krek *et al.*, 2005; Lewis *et al.*, 2003). However, in animals, miRNA binding to the mRNA is imperfect, and a perfect seed matching site is neither necessary nor sufficient for downregulation (Didiano and Hobert, 2006; Vella *et al.*, 2004). Thus, the importance of miRNA 3' base pairing to compensate for weak seed pairing has become another factor in identifying potential miRNA targets (Brennecke *et al.*, 2005). Also the sequence context of the miRNA target site and the amount of sites in a 3'-UTR are suggested to have effect on the miRNA binding efficacy (Didiano and Hobert, 2008; Grimson *et al.*, 2007; Sætrom *et al.*, 2007; Vella *et al.*, 2004). In all, the predictions given by different tools are diverse, and the amount of overlapping miRNA:target gene predictions is small (Bartel, 2009).

The use of machine learning methods can improve the accuracy of miRNA target prediction (Kim *et al.*, 2006; Wang and El Naqa, 2008; Yousef *et al.*, 2007). However, the performance of these methods is dependent on the quantity and quality of the dataset used in the training. The number of experimentally validated miRNA true and false targets is small, and while we have some knowledge about miRNA binding to the target site, our knowledge regarding false target sites is limited. This problem can be avoided by using unsupervised learning. In this work, we show how the self-organizing map (SOM) (Kohonen, 1995) can be used to identify potential miRNA target sites from *C.elegans* 3'-UTR sequences. The SOM is a neural network algorithm widely used to categorize large, high-dimensional datasets by mapping the data into a smaller dimension, typically into a 2D lattice of interconnected neurons. Each neuron of the SOM contains a reference model, which represents a local domain in the input space. In bioinformatics, SOM is applied to problems like gene expression data analysis (Törönen *et al.*, 1999), study of codon usage (Kanaya *et al.*, 2001), clustering of protein sequences (Kohonen and Somervuo, 2002), gene finding (Mahony *et al.*, 2004) and identification of transcription factor binding sites (Mahony *et al.*, 2005). We trained the SOM using short substrings of *C.elegans* 3'-UTR sequences so that the putative target sites for each miRNA were clustered in one or in two adjacent neurons in the lattice. As the clustering is based on the

*To whom correspondence should be addressed.

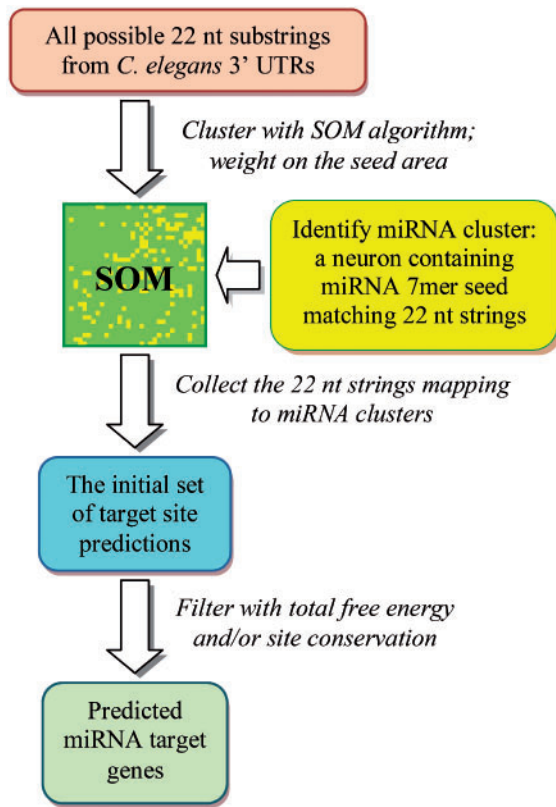


Fig. 1. Flow chart for the miRNA target prediction using mirSOM.

overall similarity of the training strings, the SOM captures not only those verified miRNA target sites with perfect 7mer and 8mer seed matches, but also those sites with more imperfect matches. The SOM contains the whole repertoire of putative miRNA target sites ordered according to similar seed complementarity, so it likely includes also the target predictions for currently unknown miRNAs.

2 METHODS

Our goal was to find putative miRNA target sites by clustering all the substrings extracted from the regions of experimentally verified *C.elegans* 3'-UTRs. To do this, we constrained the target sites of all miRNAs to have the same length. As the length of miRNA target site is close to the length of the binding miRNA and 95% of *C.elegans* miRNAs are 20–24 nt long, with an average length of 22 nt, we chose 22 nt as the length of training sequences. Although the length of the most miRNAs differ from 22 nt within a few nucleotides, this has a minor effect on the result since the last nucleotides in the 3'-end are not as important as the seed area in the miRNA–target site interaction. A process flow chart describing the mirSOM method is given in Figure 1.

Caenorhabditis elegans 3'-UTR sequence data used for SOM training: we downloaded all the experimentally verified *C.elegans* 3'-UTR sequences contained in WormBase release WS195 (Harris et al., 2010). If there were several transcripts available for a gene, the transcript with longest 3'-UTR was selected, and those sequences that were <50 nt long were discarded. The final sequence set used for training the SPM contained 8980 verified 3'-UTR sequences, 50–1912 nt in length. These sequences were then segmented into every overlapping 22 nt substrings, resulting a training set with 1 813 599 22 nt 3'-UTR substrings.

Numerical coding of the input sequences: in order to use SOM algorithm to cluster short sequences, the nucleotides were coded as follows: A=[1 0 0 0]^T, C=[0 1 0 0]^T, G=[0 0 1 0]^T and T=[0 0 0 1]^T, so each 22 nt sequence was represented by a 4×22 matrix, for example substring S=AGTCAATTTTTTTTAATTTTCT was represented by matrix:

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

miRNA sequences: the sequences of *C.elegans* and *C.briggsae* mature miRNAs and miRNA*s were downloaded from miRBase release 15 (Griffiths-Jones et al., 2008). This dataset includes 176 miRNA and 57 miRNA* sequences for *C.elegans*, and 121 miRNA and 10 miRNA* sequences for *C.briggsae*. This set contains 92 orthologous miRNAs with a conserved hairpin precursor.

SOM architecture: the structure of the SOM is a 32×32 lattice of interconnected neurons. Each neuron contains a 4×22 model matrix, which is initialized with a random sequence, but is through an unsupervised learning process modified to represent the position-specific characteristics of a larger set of repeated features in the training data, for example target sites of a miRNA. As Watson–Crick base-pairing of a target site to the miRNA seed is crucial for targeting, the last 8 nucleotides (nucleotides 15–22) in the neuron model matrix representing the seed were weighted so that the input sequence distance from the model matrix in nucleotides 15 and 22 was multiplied by a factor of two and in nucleotides 16–21 with a factor of five, while the distances in the rest of the nucleotides were not weighted. Training of the SOM was carried out sequentially, so that after finding the most similar neuron weight matrix for a randomly picked training sequence, this 'winner' weight matrix and the weight matrices of its neighbor neurons were immediately updated towards the training sequence. Thus, during the training, the weight matrix in each neuron was adjusted to represent a position-specific frequency matrix of 3'-UTR substrings clustered to it. The SOM was trained with 20 epochs with the whole training dataset. In the ordering phase (epoch 1), the neighborhood of the best matching unit contained the whole net, but was then decreased linearly so that during the last fine-tuning phase (epoch 20), it contained only the closest neurons around the winner node. The learning rate parameter was held constant in the first (lr=0.9) and last epoch (lr=0.0001), but was decreased linearly in between. The size of the SOM and the seed weighting schema used in the final network were found to work best when all networks containing 16×16, 24×24 and 32×32 neurons with different emphasis in the seed area were trained and their capability to separate the miRNA clusters was analyzed. The SOM algorithm was implemented with Java 2 SE, the source code is available as Supplementary Material.

miRNA clusters: for each miRNA, we searched all the 22nt sequences in our training set that had perfect Watson–Crick base-pairing in their section of the last 8 nts with a miRNA 7mer seed. Neurons that contained these seed matching sequences for a miRNA were attached to the cluster of that miRNA, and so each *C.elegans* miRNA cluster came to include 1–3 neurons. For *C.briggsae*, the miRNA clusters were defined similarly.

Initial set of miRNA target sites: a larger sequence set containing all the verified 3'-UTRs for *C.elegans* coding transcripts in WS213 was extracted from WormBase. Again, only the transcript with the longest verified 3'-UTR for each gene was kept for analysis, leading to the set of 12 866 verified 3'-UTRs. All of the 22 nt substrings of these 3'-UTRs were then introduced to the trained SOM and the substrings hitting the miRNA cluster neurons were collected, yielding the sets of initial target site predictions for each miRNA.

Orthologous 3'-UTR sequences and the initial set of miRNA target predictions for *C.briggsae*: the predicted *C.briggsae* WormBase orthologs for all *C.elegans* protein coding genes were downloaded from WS215. This set contained at least one predicted *C.briggsae* ortholog for 11 851 *C.elegans* genes included in our miRNA target prediction gene set. The predicted 3'-UTR sequences for these orthologs were constructed by extracting 1000 nt

8296	11406		3169	1610	2520	3298	1525	4997	2127	3852	1633	3322	3164	2164	2752	2450	3258	3540	2139	4152	2106	3224	2454	2715	621	3765	1645	3033	1915	5531	
3912			1900	995	1769	2057	1626	346	1072	1068	2631	1	3017	1	1057	532	7	1530	1117	1107	381	1937		1580	2036		1456	1789	319	2214	
2094	1062	2093	6531		5550	2096	4	2937	857	1612	2439	2036	2237	1476	3830		3669	2	2487	1948	1170	950	3187	1393	2106	1846	2666	1163	1955	3710	
3804	2354	1377				2018	3982		1941	1801	437	1901	644	2357		2559	1007	2111	2839	789	3589	994	298	2138	597	1363	1206	835	1134	1821	
5195		1761	1414	6761	1505	4922		2557	1499	2570	1510	1787	2112	22	4631	991		2477	1234	6	3780		4610	630	2006	1941	2957	1339	1373	1430	3775
	5855	1417				3819	1453	887	2505	1196	3604	3071	3470		2990	4899		3019	3612	663	4272	305	2294	2471		856	562	2203	717	1349	
3			5944	2414	2374	1924			2002		2544		4112		2914	1127	17	2359	326	1804	2090	2560	680	741	2463	1646	1889	2144	2410	1497	3463
10425	9887		1823			9414	1896		2759	1786	2487		2102	3212	707	2733	1023	2172	680	1514	631	1171	1955	1143	1548	862	2038	3071	1616	4183	
			3179	6597	1165	2549			6684		3580	1744	2559	527	1336	1831	631	1948	561	2050	2323	2856	1327	2150	1387	2362	767	1508	2506	699	4080
5594	5860	1929	1344			4137	3165	735	17	913	556	2594	78	3296	1595	1835	869	3995	1382	1265	955	689	1376	1868	592	2992	1120	1649	1776	1364	
			4816		6458	1284	2560	1369	568	1494	1468	2477	1669	1311		2158	834	3102	520	1067	1597	1996	2609	1599	2551	1588	1794	1441	2283	2082	2115
7808			3295		1670	2302	2085	2839	2527	801	1065	1843	999	1927	2508	1364	1892	2131	1620	1160	2141	597	3018		833	1222	1339	1444	3409	1094	4714
3425		5567		2145	2905	489	2972		1017	1766	1044	1876	1882	2529	50	1749	55	2673	753	2230	854	1995	1470	2370	3151	4173	2368		1774	1021	1826
3065	433		2633		2564	1227	2603	51	2915	2157	1929	22	2674	966	2636	663	3792	390	2672	127	2567	1807	2757	255	3365	1	1581	2339	2752	1239	2472
2915	2498	2446	392	3282	1105	1113	2173	816	858		1792	1900	760	1718	845	2	1560	1082	753	2704	1254	183	2702		2323	1111	2319	944	777	1717	2786
3632		1489	1829	1298	1052		2133	938	1312	2391	2384	1245	2043	2015	1268	1094	3220	1378	1261	756	752	1937	3054	2400		1635		2836	1524		1093
1593	2038	2772	1483	2925	2636	3795	1302	1307	2674		1748	325	2489		4788		1906		3625	1174	3476	576	1134		2091	1661	1886	1262	1334	1699	3444
2390		2309	674	1097		1178	86	2411		4621	1828	1425	1823	1105		4675		2117		2183	388	2700	1469	1532	1392	458	1846	1115	2087	554	2689
4883		4654	1732	1090	2982	1760	1790	1431	824	1037		2135	1954	2779	2396	2055		1472	3217		3201	615	1170	2174	774	1994	556	2469	1682	818	2482
804		1477	840	1523	1698		2434	846	2589	1720	2427	1080	678		274	3282	3704	983	2005	1533	2050	1413	3772	59	2298	2184	439	2758	1478	1356	2393
4452	1146	951	2206	1449	1260	2028	1348	479	2493		3362		2296	3897	3524		479	803	1290		1814	669	1124	2588	634	2148	1906	1352	2770		3624
2100	1828	2097	1197	1376	1221	2302	120	2904	1618	2277	1348	4000	711	540		2605	4472	472	3843	1072	3360		2521	1164	2564	1797	389	1973		2400	3640
3063	303	613	1453	3415	846	1726	2212	1918	351	2835	1442		2602	1941	4096		1228	929	1201	1060	1805		1197	2036	1031	848		3806	1363	1492	
1429	1610	3023		2258		1754		1890	2160		2359	1791	2414	1499		6546	831	4935	8	1298	2030	770	2601	1892		2613	1659	2581		3194	3128
3926		3170	1864		446	2585	2809	2066	87	2620	1951	1243	795	3067			1019	1147	2656	718	2226	1905	1073	2798		737	1380	1658		1201	
	5317		961	3826	3092	807		2340	1749	1385	848	1842	1988		4462	2198	2886	993	1158		2389	663	2581	2167	475	2930		3195	1401	1351	4028
4477		4282		1237		864	3845	1665	2317	2135	1281	3314	1088	2846					1928	2927	2166	1003	1361		2581		4294		2494	3512	
4687	54		4168	3102	2461	2139		1132	1604		1154	983	628	2901	3521	3746	570	3031		2534	480	1679	4022		3253		3330	3833		3247	
	1635	4165	2224	566	1605		859	1185	1613	2360	2310	2762	1954	3279					2255	1909	2084	763	2153	1917	2292	1552	1455	422	979	679	1716
7269		2004	858	3025	1738	2019	4645	1198	1592	1716		1	1085		5498	2915	4926	2654	2095	2870	778	2697	1386	3105	2775	747	3026	2513	3280	1727	2976
	3282		442	3337	298	1258	519		992	2988	946	2574	1835				54			1030	1971		2150	321		750	489	356		1687	
7800		5266	2724	3233	4996	1240	3941	2400	2674	1514	3654	2574	2211	2708	2184	4382	2412	4123	1717	2990	2038	2170	3015	1519	4981	4166	2612	3307	6521	805	6075

Fig. 2. The SOM trained with *C.elegans* 3'-UTR substrings. The yellow colored cells indicate the locations of *C.elegans* miRNA cluster neurons in the lattice. The numbers display the amount of training sequences mapped to each neuron of the trained map, empty cells are neurons with no hits.

downstream from the end of the last exon of each *C.briggsae* gene included. The initial *C.briggsae* miRNA target sites were obtained by collecting the 22 nt substrings whose most similar neuron in the SOM was a *C.briggsae* miRNA cluster neuron.

Total free energy for miRNA binding: the total free energy was calculated for each miRNA–target site pair in the initial prediction set of the two worms. First, a larger sequence containing the 22 nt target site and 70 nt upstream and downstream from it was extracted from the 3'-UTR. This length of 70 nt was chosen based on the fact that secondary structure base-pairing interaction between nucleotides that are separated by >70 nt is unlikely (Kertesz *et al.*, 2007). If the first nucleotide of the site was located closer than 70 nt from the 3'-UTR start, a part of the last exon was attached to get the 70 nt upstream. If the last nucleotide of the site was closer than 70 nt from the 3'-UTR end, the downstream part was truncated to this smaller size. To get the energy needed to open the site area for miRNA binding, ΔG_{open} , we calculated the free energy of the thermodynamic ensemble using RNAfold (Hofacker, 2003), first for this 162 nt sequence as such, and then when the target site in the middle of it was forced to be open. ΔG_{open} is the difference of these energies. Next, we calculated the free energy of the thermodynamic ensemble for the miRNA–target site duplex with RNAfold (Hofacker, 2003) and obtained the total energy freed in miRNA binding to the site with equation $\Delta \Delta G = \Delta G_{\text{duplex}} + \Delta G_{\text{open}}$.

3 RESULTS

3.1 SOM trained with *C.elegans* 3'-UTR substrings

The SOM trained with 22 nt substrings of *C.elegans* experimentally verified 3'-UTRs is illustrated in Figure 2. Each short sequence in the training data is assigned to one, most similar neuron in the net. Sequences mapped to the same neuron are similar with each other, and especially similar they are in the region of the last 8 nt

complementary to the miRNA 8mer seed area. In Figure 2, the number of training sequences mapped to each neuron is shown, and the yellow color indicates the locations of miRNA cluster neurons. For *C.elegans*, there are 160 cluster neurons in total, containing an average of 2148 training data strings each.

In Figure 3, the miRNA names are attached to their clusters. miRNAs that share the same 8mer seed sequence are always clustered to the same neuron and their names are coupled together in a single miRNA name ending with an 'F' representing 'family'. To save space, we used this abbreviation also in some cases where two miRNAs with the same cluster share a 7mer seed. For example *let-7F* stands for *let-7*, *miR-241*, *miR-48*, *miR-795* and *miR-84* sharing the 8mer seed UGAGGUAG and also for *miR-793* (UGAGGUAG) and *miR-794* (UGAGGUAA) which share a 7mer seed with the first five miRNAs. The abbreviations used in the Figure 3 are clarified in Supplementary Table S1. The seven *let-7F* miRNAs and two other miRNAs, *miR-124* (UAAGGCAC) and *miR-2211* (UCAGGUAG) share the same, one and only cluster neuron of the SOM lattice. Altogether, 211 of 233 miRNAs contain only one neuron in their cluster and 117 neurons of these are clusters for only one miRNA. Twenty-one miRNAs have two neurons in their clusters, while the cluster for one miRNA star sequence, *cel-lin-4** contains three neurons. When there is more than one neuron in a miRNA cluster, these neurons are adjacent in the map, except for *mir-240* and *mir-63** for which these neurons are more distant but still quite near to each other. Note that only the upper right quarter of the SOM is shown in Figure 3. For the whole network, see Supplementary Figure S1.

To visualize how the different miRNA 8mer seed complementary sequences are distributed in the SOM lattice, we presented the

	miR-238F		miR-355		miR-61*	miR-261									
				miR-76	miR-76										
		miR-220b*						miR-243*	miR-71*	miR-2F miR-250				miR-72F miR-73F miR-64*	
					miR-253	miR-247F miR-2218b							miR-49 miR-83	miR-2207*	miR-784
miR-1832b		miR-267			miR-253		miR-65*		miR-1822 miR-255	miR-1822 miR-2213					miR-248
	miR-1829a*		miR-2221 miR-785F				miR-65*	lin-4F miR-789	miR-359				lin-4*		
		miR-265								miR-51F miR-52F	miR-1829a	lin-4*	miR-35F miR-271 miR-360 lin-4*		miR-249
	let-7F miR-124 miR-2211				miR-1832		miR-56* miR-2213*	miR-2214*			miR-57	miR-1817	miR-2211* miR-2218a*	miR-273	miR-2210 miR-2212 miR-34F
	miR-1821	miR-1821								miR-1829b		miR-58*			miR-2208a* miR-2208b-3p
	miR-1830 miR-1833	miR-1818								miR-242					
miR-240*			miR-245	miR-42*		miR-360*								miR-46F miR-392	
		miR-264 miR-37*												miR-2216	
					miR-2208F miR-2209*F		miR-124*			miR-786				miR-2216	miR-246
										miR-786					
					miR-244*	miR-240 miR-358*					miR-70	miR-236			miR-50F
	miR-1829c miR-1832b* miR-247*			miR-1823 miR-60*		miR-230*	miR-2212* miR-240								

Fig. 3. *Caenorhabditis elegans* miRNA cluster locations. The cluster of each miRNA is shown by the miRNA name. A name ending with ‘F’ indicates a miRNA family, a couple of miRNAs with the same 8mer or 7mer seed with the same cluster neuron. Only the upper right quarter of the SOM is shown. The entire network can be seen in Supplementary Figure S1.

consensus sequences of the substrings mapped to each neuron, including only the last 8 nt (see Fig. 4 for the upper-right quarter of the SOM and Supplementary Figure S2 for the entire SOM). Consensus sequences that contain an abundance of nucleotide T are located in the upper left corner of the net while sequences in the opposite, right lower corner of the net are rich of A:s instead. The clusters of about a third of all known *C.elegans* miRNAs are located in the upper right quarter of the net. The consensus sequences often contain an N in the first and in the last place, for example the consensus sequence of the *let-7F* cluster is NTNCCTNN. When there are fewer miRNAs attached to a neuron, the consensus sequence often is more precise. For example, NTACAAAN for the *lsy-6* sole cluster neuron.

3.2 miRNA target prediction using mirSOM

All the 22 nt 3'-UTR substrings mapped to a miRNA cluster neuron in the SOM were considered as the initial set of target sites for that miRNA. The total free energy showing the accessibility of the site for the miRNA binding was calculated for each site in the initial prediction set. In order to optimize the predictions made by our tool, its capability to correctly pick out the 20 experimentally verified

miRNA–target genes and discard the 13 well-characterized *cel-lsy-6* false targets (for a list of genes see Supplementary Table S2) was measured as the function of the total free energy cut-off used (Fig. 5). The optimal threshold was -7.2 kcal/mol, when 17 of 20 verified true targets were found and 12 of 13 false targets were rejected. Then, for each miRNA in both nematode species, we left out those sites in the initial prediction set whose total free energy value exceeded this threshold. For the orthologous miRNAs, the conservation of each miRNA–target gene relationship in *C.elegans* and *C.briggsae* was studied: if also the ortholog gene 3'-UTR included a predicted site for the miRNA located closer than 20nt from the site in the *C.elegans* 3'-UTR, the miRNA–target gene relationship was classified as conserved. As phylogenetic conservation is a very strong indicator for a miRNA–target relationship, we restored to the set of final predictions target genes which included a perfect 7mer or 8mer seed match with too large total energy, but instead a conserved predicted site in *C.briggsae*, thus resulting in the final set of mirSOM target predictions. The sensitivity and specificity of mirSOM are 0.90 and 0.92, when calculated using the sets of experimentally verified miRNA true and false target genes. It is noteworthy that while the initial prediction set for all miRNAs with the same cluster neuron were the same, this is not the case in the

NATACATN	NNWACANN	NNAACANN	NAAACANN	NAAACNNN	NNAACNNN	NNAASCNN	NNAAGANN	NNAAGAAN	NNNAGAAN	NTKWGAAN	TTTTGAAN	WNTTGAAN	WATTGANN	NATTGNNN	NATTGTNN
NANACANN	NNNACANN	NNAACANN	NWAAACANN	NNAACNNN	NNAACNNN	NNAASNNN	NNAAGNNN	NNNAGAAN	NNGAGAAN	NTGWGAAN	TTKTGAAN	TTTTGAWN	WNTTGATN	WATTGNNN	WATTGCNN
NNCACANN	NTCACANN	NTNACANN	NTWACNNN	NTAACTNN	NNAACTNN	NNAAGTNN	NNNAGTNN	NNGAGNNN	NNGAGANN	NNGWGAAN	NTNTGATN	TTTTGATN	TTTTGNTN	TNTTGCTN	WNTTGCNN
NNWCANN	NTCACANN	NTNACNNN	NTNACNNN	NTTACTNN	NNTACTNN	NNTAGTNN	NTNAGNNN	NNNAGNNN	NNGAGTNN	NNCAGTNN	NTCTGATN	NTNTGNTN	NTNTGNTN	NTGTGCTN	NTGTGCNN
NTCCANN	NTCMCMNN	NTCACNNN	NTNACTNN	NWYACTNN	NNTACTNN	NNTASNNN	NTNAGNNN	NTYAGNNN	NNCAGTNN	NNCAGTNN	NNCTGTTN	NTCTGTTN	NTGTGTTN	NNGTGTTN	NNGTGTNN
NTCCANN	TTCCNNNN	TTCCMNNN	NTCACTNN	NNCACTNN	NNYACNNN	NNTACNNN	NTTASNNN	NNCAGNNN	NNCAGTNN	NNCAGTNN	NNCNGTTN	NNCNGTTT	NNNGGTTN	NTNGGTTN	NNKGTTNN
NTNCCNNN	NTCCNNNN	NTCCCTNN	NNCACTNN	NCCACTNN	NNCACNNN	NNCACNNN	NNCASCNN	NNCMGNNN	NNCCGNNN	NNCCGTTN	NNCGGNTT	NNTGNTNN	NNTGNTNN	NNCGGTTN	NNCTGTSN
NTNCCNTN	NTNCCNTN	NYCCCTNN	NCCCCTNN	NCCMCCNN	NCCMCCNN	NCCMCCNN	NNCCGCMN	NNCCGANN	NNCCGATN	NNCCGNTN	NNNSGNTN	NNTGNNNN	NNTGNNNN	NNCGNNNN	NNCTGNNN
NTTCTNN	NNNCCCTN	NCNCCCTN	NCNCCNNN	NCCCCTNN	NCCCCCCTN	NCCCCCCTN	NCYCMGNN	NNTCGANN	NNTCGATN	NNNCGNTN	NNNCGNTN	NNNGGNNN	NNNGGAAN	NNCKGAAN	NNCTGAAN
NNTCTNN	NAWCCTNN	NMACCANN	NCNCCANN	NCCCCTNN	NCCCCCCTN	NCYCMGNN	NNTCRCAN	WTTTCGANN	NTTCGANN	NNWCGNKN	NNACGNNN	NNANGAAN	NNAGGAAN	NNATGAAN	NNATGAAN
NNTCTNN	NAWCCTNN	NANCCANN	NMACCANN	NCNCCANN	NCNCCANN	NNTCACAN	NNTCANAN	WTTTCGANN	NTTCGANN	NNWCGTNN	NNACGTTN	NAACGNNN	NAATGANN	NNATGNNN	NNATGNNN
NNTSCWNN	NATCCNNN	NANCCNNN	NAAACNNN	NNACMCAN	NNACACAN	NNWCACAN	NTTCATAN	NTTCATNN	NNTCRTNN	NNTCRTTN	NANCRTTT	NAANGNTN	NAATGNNN	NNATGTTN	NNATGTNN
NTTSCNEN	NTTCCCN	NAYCCCN	NANCMCN	NNACACNN	NNACACNN	NNACATAN	NTTCATAN	NTTCATNN	NNTCATYNN	NNTCATT	NNMCATT	NAAYATT	NAATRTTN	NAATGTNN	NNATGTAN
TTTSCNN	NTTCCCN	NWTCCCN	NANACYN	NNNCACNN	NNACAGNN	NNACATNN	NNNCATNN	NTYCATNN	NTTCATT	NNWCATT	NNACATT	NNATATT	NAATATT	NAATATNN	NNATATAN
TTTCCCTN	TTTCCCYN	TTTCNYYN	NWTCCACYN	NNNCAGNN	NNNCAGNN	NNCCANNN	NNCCATNN	NNCCATT	NTNCATT	NTNCATT	NTATATT	NNATATT	NAATATT	NAATATNN	NNATATNN
TTTSTCTN	TTTCTCTN	TTTCACTN	TTTCANTN	NTNCARTN	NTCCARNN	NNCCANNN	NCCCATTN	NNCCATT	NNCCATT	NTCTATT	NTNTATT	NTGTATTN	NWGTATTN	NNATAKNN	NNATAGNN

Fig. 4. The distribution of 8mer seed complementary sequences in the SOM. Yellow color indicates the clusters of *C.elegans* miRNAs. Only the upper right quarter of the SOM is shown. The entire network can be seen in Supplementary Figure S2.

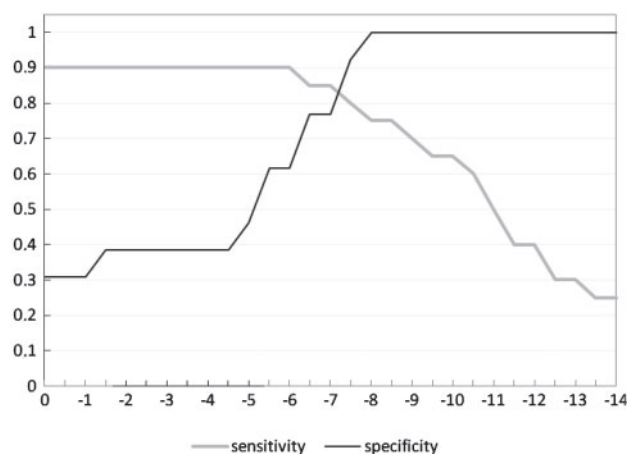


Fig. 5. The sensitivity and specificity of the model as the function of the total energy threshold used. The verified true (20 genes) and false (13 genes) miRNA–target relationships were used to optimize the output of the SOM model. The optimal energy threshold is -7.2 kcal/mol, when 17 of 20 true targets were accepted and 12 of 13 false targets rejected.

final predictions. For example, if we consider miRNA *let-7*, 56% of its final predicted targets are shared with the miRNAs with the same seed and the same cluster neuron, and only 17% of its targets are shared with all eight miRNAs with the same cluster neuron.

3.3 Comparison of mirSOM with other methods

We compared the mirSOM performance with seven other miRNA target prediction tools available: NBmiRTar (Yousef *et al.* 2007);

PITA (Kertesz *et al.*, 2007); rna22 (Miranda *et al.*, 2006); MicroCosm (Griffiths-Jones *et al.*, 2008); TargetScanWorm (Ruby *et al.*, 2006); PicTar (Lall *et al.*, 2006); and mirWIP (Hammell *et al.*, 2008). Of these, NBmiRTar uses supervised machine learning by a naïve Bayes classifier applied to the output of miRanda program (John *et al.* 2004). NBmiRTar is a model generated from sequence and miRNA:mRNA duplex information from validated targets and artificially generated negative examples. Rna22 uses short patterns extracted from mature miRNA sequences to find eligible places for miRNA targeting, and then pairs up these islands with a miRNA. PITA first identifies the initial seeds for each miRNA in the 3'-UTR, then calculates the site accessibility and combines the sites for the same miRNA to get the total interaction score. NBmiRTar, rna22 and PITA (version PITA All) do not apply any cross-species sequence conservation filter to the target predictions. MicroCosm (earlier known as miRBase Targets), TargetScanWorm and PicTar are the three tools most commonly used for miRNA target prediction in *C.elegans*. While MicroCosm searches for maximal local complementarity alignments between a miRNA and a 3'-UTR and uses the conservation filter afterwards, TargetScanWorm and PicTar predict conserved miRNA targets using genome-wide alignments of the 3'/UTRs of *C.elegans*, *C.briggsae* and *C.remanei*. mirWIP is a recent tool based on contextual features of miRNA binding sites enriched in a set of genes predicted by immunoprecipitation to be targets of a set of conserved *C.elegans* miRNAs (Zhang *et al.*, 2007). mirWIP differs from the other six computational tools compared in that it is based on unique biological experiments.

First, we compared the overlap of the miRNA target sites predicted with these tools. NBmiRTar was excluded from this part of the

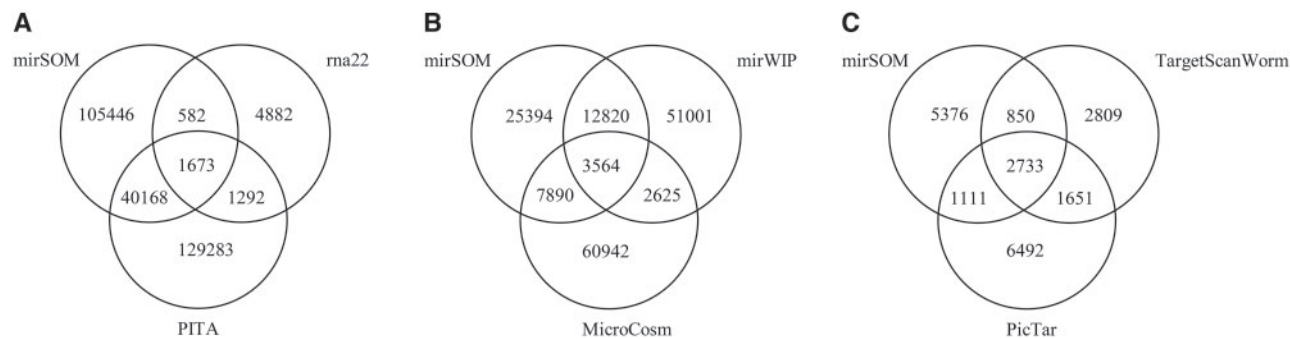


Fig. 6. The degree of overlap between the predictions of mirSOM compared with six other miRNA target prediction tools. The number of mirSOM predictions included in each comparison depends on the scope. **(A)** All mirSOM predictions compared with rna22 and PITA, **(B)** mirSOM predictions for orthologous miRNAs compared with mirWIP and MicroCosm and **(C)** mirSOM conserved predictions compared with TargetScanWorm and PicTar.

comparison, because there is no bulk download available for its prediction data. As mirSOM has also the conserved site prediction aspect, we made these comparisons separately for tools that do not require conservation and for tools with a conservation filter. The Venn diagram in Figure 6a shows the number of predicted sites common to mirSOM, PITA and rna22. In comparison to the number of predictions made in all, the size of target gene set predicted with all the three tools is small. Approximately 28% of mirSOM predictions are common with PITA, but PITA also contains the largest number of predictions in all. The low number of common predictions with rna22 may be due to the small number of predictions downloadable for rna22 which contains sites from <3000 genes.

When we compared mirSOM with tools utilizing a conservation filter, only the mirSOM predictions for the orthologous miRNAs were taken into calculations. The number and intersections of miRNA–target gene interactions predicted by mirSOM, mirWIP and MicroCosm are shown in the Venn diagram in Figure 6b. About 30% of mirSOM predictions are common with mirWIP and ~50% of the mirSOM predictions are found with either of the tools. The results of a similar comparison made between mirSOM, TargetScanWorm and PicTar are shown in Figure 6c. As TargetScanWorm and PicTar search for conserved targets, only the 10 070 conserved miRNA–target gene pairs contained in mirSOM predictions were taken along. Forty percent of these mirSOM predictions are shared with either PicTar or TargetScanWorm, and about one-third of the mirSOM predictions are common to all of the three tools, thus underlining the ability of them all to find the perfect seed matching, conserved miRNA target sites. Altogether, ~35% of the miRNA–target gene interactions predicted by mirSOM are included in the prediction set of at least one of the six other tools. Among the predictions made only with mirSOM are the predicted targets for 57 star miRNAs and 23 miRNAs that are not included in any of the other tools. Also the gene sets and 3'-UTR sequences incorporated in these tools are unequal, which has an impact to the comparison results.

Secondly, we tested the ability of these tools to find the experimentally verified *C.elegans* miRNA–target genes and to reject the *lsy-6* false target genes (Supplementary Table S2). For seven of the 33 genes studied, there was no verified 3'-UTR available; so for them we extracted the 1000nt sequence downstream from the last exon to act as the predicted 3'-UTR. mirSOM found 18 of 20 true target genes in the list (Table 1). The two verified targets not

Table 1. The experimentally verified *C.elegans* miRNA–target genes found by mirSOM, NBmiRTar, PITA, rna22, MicroCosm, TargetScan, PicTar and mirWIP

miRNA	Gene name	mir Som	NBmiR Tar	PITA	rna22	Micro Cosm	Target Scan	PicTar	mirWIP
<i>let-7</i>	C35E7.4	X	X	X	–	–	–	–	–
<i>let-7</i>	<i>ceh-16*</i>	X	X	–	–	–	–	–	X
<i>let-7</i>	<i>daf-12</i>	X	X	X	X	X	X	X	X
<i>let-7</i>	<i>die-1</i>	X	X	X	–	X	–	–	–
<i>let-7</i>	<i>hbl-1</i>	X	X	X	X	X	X	X	X
<i>let-7</i>	<i>let-60</i>	–	X	–	–	–	–	–	X
<i>let-7</i>	<i>lin-41</i>	–	–	–	X	–	–	X	X
<i>let-7</i>	<i>let-526</i>	X	X	X	–	–	–	–	–
<i>let-7</i>	<i>nhr-23</i>	X	–	–	–	–	–	–	X
<i>let-7</i>	<i>nhr-25</i>	X	X	X	–	–	–	–	X
<i>let-7</i>	<i>nhr-4</i>	X	X	X	–	–	X	X	–
<i>let-7</i>	<i>pha-4</i>	X	–	–	–	–	–	–	X
<i>let-7</i>	<i>T14B1.1</i>	X	X	X	X	X	X	X	X
<i>let-7</i>	<i>uba-1</i>	X	X	X	–	–	–	–	X
<i>let-7</i>	<i>unc-129</i>	X	X	X	–	X	X	X	X
<i>lin-4</i>	<i>lin-14*</i>	X	X	–	–	–	X	X	X
<i>lin-4</i>	<i>lin-28</i>	X	X	X	–	X	X	X	X
<i>lsy-6</i>	<i>cog-1</i>	X	–	X	–	X	X	X	X
<i>mir-273</i>	<i>die-1</i>	X	X	X	–	–	–	–	–
<i>mir-61</i>	<i>vav-1</i>	X	X	X	–	–	X	X	X
		18	16	14	4	7	9	10	15

X' = the gene is included; '–' = the gene is not included in the predictions. Asterisks () next to gene name refers to targets which did not have a verified 3'-UTR in WS213, and the 1000 nt downstream sequence was used instead to test the mirSOM method.

found are *let-7* target genes *let-60* and *lin-41*. Interestingly, *let-60* and *lin-41* have a couple of sites in neurons adjacent to *let-7* cluster. Of these, the *let-60* sites in positions 479–500 and 128–149 in its 3'-UTR have also acceptable total free energy values, –13.9 kcal/mol and –9.1 kcal/mol, respectively. Thus, if we would include also the adjacent neurons to *let-7* cluster, mirSOM would correctly predict all but one of the true targets. However, even with the 18 true targets found, mirSOM outperforms all the other methods. Of the 13 well-characterized *lsy-6* false targets mirSOM rejects 12, together with rna22. NBmiRTar is the only tool that rejects all the false targets, when used with the default values of its parameters. The number of

Table 2. Verified *C.elegans* *lxy-6* non-target genes found by mirSOM, NBmiRTar, PITA, rna22, MicroCosm, TargetScan, PicTar and mirWIP

miRNA	Gene name	mir Som	NBmiR Tar	PITA	rna22	Micro Cosm	Target Scan	PicTar	mirWIP
<i>lxy-6</i>	C27H6.3	–	–	–	–	X	X	X	–
<i>lxy-6</i>	<i>ptp-1</i>	–	–	X	X	–	–	–	–
<i>lxy-6</i>	<i>fkh-8</i>	–	–	X	–	X	X	X	X
<i>lxy-6</i>	<i>nsy-1</i>	–	–	X	–	–	–	–	–
<i>lxy-6</i>	<i>acl-5</i>	–	–	X	–	–	X	X	–
<i>lxy-6</i>	<i>aex-4</i>	–	–	X	–	–	X	X	–
<i>lxy-6</i>	T20G5.9	–	–	–	–	–	X	X	–
<i>lxy-6</i>	<i>glb-1</i>	–	–	–	–	X	X	X	–
<i>lxy-6</i>	<i>hlh-8*</i>	–	–	–	–	–	X	X	–
<i>lxy-6</i>	F55G1.12*	–	–	–	–	–	X	X	–
<i>lxy-6</i>	T04C9.2*	X	–	–	–	–	X	X	X
<i>lxy-6</i>	T05C12.8*	–	–	–	–	–	X	X	–
<i>lxy-6</i>	T23E1.1*	–	–	–	–	–	–	X	–
		1	0	5	1	3	10	11	2

'X' = the gene is included; '–' = the gene is not included in the predictions. Asterisks (*) next to gene name refers to targets which did not have a verified 3'-UTR in WS213, and the 1000 nt downstream sequence was used instead to test the mirSOM method.

rejections made by the other five tools is between 3 and 11 (Table 2). For mirWIP and PITA, the largest set of predictions available is used in the comparison, which has an effect to the number of false target predictions.

4 DISCUSSION

We have here introduced a new machine learning-based method, mirSOM, for miRNA target finding. Our method is founded on the basic knowledge about miRNA target sites: these sites have imperfect complementarity with the miRNA sequence, where the most crucial part is the miRNA seed, these sites are preferentially located in the 3'-UTR of the transcript and they are accessible to miRNA binding. This means that a chunk of short substrings of the 3'-UTRs are miRNA target sites. We sought these sites by clustering short 3'-UTR substrings with the SOM. The training data consisted of all possible 22-nt long substrings extracted from verified *C.elegans* 3'-UTRs, and the trained SOM contains, besides the information about the target sites of miRNAs known today, also a view of the mutual similarity of miRNA sequences. The fact that 90% of the miRNA clusters reside in a single neuron of the lattice implies that the SOM is well organized, and it is quite probable that also the clusters for miRNAs yet to be found from *C.elegans* are located in one or two nearby neurons of the map. Once trained, the SOM can be applied, not only to find putative targets for new *C.elegans* miRNAs, but also to find the targets for the orthologous miRNAs in other worms thus enabling the study of target conservation. As mirSOM uses only the filtering by site accessibility to post-process the initial target sites obtained from the SOM, it opens up a simplification to the miRNA target prediction, but still works well in finding the verified miRNA target sites, while keeping the number of predictions moderate.

A common problem in miRNA target finding is the small number of experimentally verified miRNA–target genes, which has hindered efforts to develop machine learning methods for target prediction tasks. In particular, this makes it hard to build a supervised model

whose performance is highly dependent on the quantity and quality of the positive and negative training datasets used. The unsupervised learning algorithm used by mirSOM clusters the putative target sites for each miRNA objectively, using no knowledge about the verified true and false targets until the optimal total free energy threshold for the site accessibility is defined. When compared with NBmiRTar, a method using supervised machine learning, mirSOM finds 18 of the 20 experimentally verified targets while NBmiRTar predicts 16 of 20. These are the two best results in the comparison (Tables 1 and 2). Since both of these tools are also very specific in their predictions, it suggests that miRNA target prediction truly can be improved by the use of machine learning methods.

The standard SOM algorithm has some limitations, like the non-adaptable fixed architecture, which is relative to the expected number and structure of the clusters. Instead of using an algorithm that would adaptively determine the number of clusters (Mavroudi *et al.*, 2002; Hsu *et al.*, 2003), we searched experimentally for such size and weighting schema for the standard SOM that would work best in this particular problem. As the solution does not necessarily need a more sophisticated architecture, we decided to keep the original SOM algorithm which is well known and easy to understand.

mirSOM finds well those verified miRNA targets which include a perfect 7mer or 8mer seed match, and also most of those target genes with a shorter or otherwise imperfect miRNA seed match. It also successfully rejects the sites from known false target genes with a perfect seed match by using the total energy filter to the initial prediction set. The energy threshold used, which relies on just a few verified miRNA–target gene interactions presently available, should be adjusted as more miRNA–target genes in *C.elegans* are verified. Also, the SOM used in the initial target site finding step is trained with verified 3'-UTR data, so it may be not suitable for finding miRNA target sites from other sequence data, for example from 5'-UTRs and the coding region. This is because the sequence characteristics, like nucleotide composition and functional elements, of these regions of mRNA differ from the 3'-UTRs, and the clustering result using these sequences would likely be different. We demonstrate here that mirSOM can identify targets in nematodes and it should be possible to construct SOMs for other species. Future studies will be aimed at verifying targets in more complex organisms such as human and mice.

ACKNOWLEDGEMENTS

Jussi Paananen, Mitja Kurki and Petri Törönen are acknowledged for technical assistance, comments and discussion.

Funding: Saastamoinen Foundation; Finnish Cultural Foundation, North Savo Regional fund; Finnish Cultural Foundation; Doctoral Program in Molecular Medicine at the University of Eastern Finland.

Conflict of Interest: none declared.

REFERENCES

- Bartel, D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
- Brennecke, J. *et al.* (2005) Principles of microRNA–target recognition. *PLoS Biol.*, **3**, e85.
- Didiano, D. and Hobert, O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA–target interactions. *Nat. Struct. Mol. Biol.*, **13**, 849–851.

- Didiano,D. and Hobert,O. (2008) Molecular architecture of a miRNA-regulated 3'UTR. *RNA*, **14**, 1297–1317.
- Enright,A.J. et al. (2003) microRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
- Griffiths-Jones,S. et al. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
- Grimson,A. et al. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
- Hammell,M. et al. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods*, **5**, 813–819.
- Harris,T. et al. (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
- Hofacker,I. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
- Hsu,A. et al. (2003) An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data. *Bioinformatics*, **19**, 2131–2140.
- John,B. et al. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
- Kanaya,S. et al. (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene*, **276**, 89–99.
- Kertesz,M. et al. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.*, **39**, 1278–1284.
- Kim,S.K. et al. (2006) miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, **7**, 411.
- Kloosterman,W.P. and Plasterk,R.H. (2006) The diverse functions of microRNAs in animal development and disease. *Dev. Cell*, **11**, 441–450.
- Kohonen,T. (1995) *Self-Organizing Maps*. Springer, Berlin.
- Kohonen,T. and Somervuo,P. (2002) How to make large self-organizing maps for nonvectorial data. *Neural Netw.*, **15**, 945–952.
- Krek,A. et al. (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
- Lall,S. et al. (2006) A genome-wide map of conserved microRNA targets in *C.elegans*. *Curr. Biol.*, **16**, 460–471.
- Lee,R.C. et al. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, **75**, 843–854.
- Lewis,B.P. et al. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
- Mahony,S. et al. (2004) Gene prediction using the self-organizing map: automatic generation of multiple gene models. *BMC Bioinformatics*, **5**, 23.
- Mahony,S. et al. (2005) Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, **21**, 1807–1814.
- Mavroudi,S. et al. (2002) Gene expression analysis with a dynamically extended self-organized map that exploits class information. *Bioinformatics*, **18**, 1446–1453.
- Miranda,K.C. et al. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
- Papadopoulos,G. et al. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
- Ruby,J.G. et al. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *Caenorhabditis elegans*. *Cell*, **127**, 1193–1207.
- Sætrom,P. et al. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.*, **35**, 2333–2342.
- Törönen,P. et al. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**, 142–146.
- Vella,M. et al. (2004) The *C. elegans* microRNA *let-7* binds to imperfect *let-7* complementary sites from the *lin-41* 3'UTR. *Genes Dev.*, **18**, 132–137.
- Wang,X. and El Naqa,I. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, **24**, 325–332.
- Yousef,M. et al. (2007) Naive Bayes for microRNA target predictions—machine learning for microRNA targets. *Bioinformatics*, **23**, 2987–2992.
- Zhang,L. et al. (2007) Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell.*, **28**, 598–613.