On Identifying Metabolic Functions of Noncoding RNAs in *S. cerevisiae*

Eric Struminger  
*The University of Illinois, Urbana*  
strum1@illinois.edu

Max H. Garzon  
*The University of Memphis*  
mgarzon@memphis.edu

Sungchul Ji  
*Rutgers University*  
sji@rci.rutgers.edu

**Abstract**

We produce putative biological functions of over 150 non-coding RNAs in *S. cerevisiae*, out of over 2,800 unknowns RNAs, together with an analysis that provides confidence levels, obtained using two major computational intelligence techniques, multilayer perceptrons (MLPs) and Self-Organizing Maps (SOMs). The identifications fall in two groups, depending on the level of confidence with which the function is being assessed (high and low). In the remaining group of RNAs (over 2,700), “hard core” RNAs remain elusive to identification. The first two groups of putative categories may be representing a new ontology worthy of further research and validation by the biological community, given other successes in the application of MLPs and SOMs as less researcher-biased classification tools. Although analyses of microarray data are plentiful by other techniques, a novel contribution of this paper is that the analyses has been carried out in the researcher-independent ontology implicit in the inherent properties of neural network, which are based solely on the given data.

*Key Words:* *S. cerevisiae*, biological function, noncoding RNA, microarray data analysis, neural networks.

**1. Introduction**

The human genome project of the 1990s marked a critical transition in the study of biological organisms and has transformed theory and practice of experimental biology. While sequencing has been in itself a challenge, enormous progress has brought genome sequencing to the verge of a commodity that can be had for well under $1,000 in the near future [1].

As already anticipated by many, this progress has brought to the front the second and more important phase of the post-genome project era, i.e. the elucidation of the molecular mechanisms underlying the genotype-phenotype coupling. Despite the enormous amount of data generated by genome sequencing, they pale in comparison by the extraordinary amount of analytic and computational resources required to do the bioinformatics of assembling an accurate picture of the complex molecular interactions among genes, RNAs and proteins in living cells that sustain life. A primary problem in this program is the identification of the metabolic functions of long non-coding RNAs (ncRNA) usually defined as non-polyadenylated RNAs with greater than 200 nucleotides [8,9,10]. Intense research over the past decade or so has demonstrated that many noncoding RNAs participate in regulating cell functions including RNA splicing, RNA editing, transcription factor transport, translation, and transcript degradation [9].

In this paper, we focus on identifying the metabolic functions of non-coding RNAs (ncRNAs) in one of the biologist’s favorite organisms, *Saccharomyces cerevisiae* (*S. c. hereafter*), or baker’s yeast [11]. This organism was sequenced in 1996 (the first eukaryotic genome that was fully sequenced, annotated, and made publicly available) and shown to consist of over 6,000 genes [6][11]. Of these, about 3,000 genes code for proteins with known metabolic functions, but the remaining genes code for RNAs that do not encode proteins and hence their metabolic functions are unknown. The RNAs encoded by such genes are referred to as non-coding RNAs [9]. Using two independent techniques described in Section 3, we have identified the possible metabolic functions of over 170 of these ncRNAs with greater than 90% confidence and propose possible molecular mechanisms underlying their suggested functions in Section 4. Finally, some discussion of the credibility of this assessment, its interpretation and general biological significance is presented in Section 5.

**2. The measurement of the transcriptome**

In this Section we describe the tools used and the data
utilized to train and enable neural network to make the predictions subject of this paper.

DNA microarrays have been used as a means to ascertain the possible functions of non-protein coding RNAs through analyses of their transcript levels (TL). In the study reported in [3], the S.c. strain BQS252 was grown overnight at 28 °C in YPD medium (2% glucose, 2% peptone, 1% yeast extract) to exponential growth phase (OD$_{600}$ = 0.5) [3]. Cells were recovered by centrifugation, resuspended in YPGal medium (2% galactose, 2% peptone, 1% yeast extract), and allowed to grow in YPGal medium for 14-15 hours. Cell samples were taken at 0, 5, 120, 360, 450 and 850 minutes after the glucose-galactose shift. The 850 minute sampling time corresponded to the exponential growth phase in the YPGal medium. The TLs were measured with DNA arrays as described in [3]. The total amount of poly(A) mRNA per cell was measured and used to normalize the microarray signals. The original data contained readings of 5,914 mRNAs (ORFs) which were assigned to 531 metabolic pathways (henceforth referred to as "categories") as illustrated in Fig. 1. Therefore, there were 2,817 remaining RNAs (and their transcripts designated as ncRNAs, or noncoding RNAs) whose metabolic functions remain unknown. Each RNA in the data is identified by a name, a category (perhaps "Unknown"), and mRNA expression level readings taken at the six different time points.

![Figure 1. Histogram of the genes assigned to the 531 biological functions (categories) in S.c](image)

3. Genomics with Multi-layer Perceptrons

To determine the biological function of the ncRNA (i.e., RNAs with unknown functions), several methods exist in the literature that would allow an "educated extrapolation" based on a computational analysis of their microarray expression profiles (also called RNA trajectories, or RNA traces). Techniques vary from statistical approaches, to neural networks, to evolutionary algorithms, to ad-hoc approaches such as chaos theory. Neural networks appear most appropriate for this task because of their proven generalization ability, based solely on a large data corpus, as is the case here. Fig. 1 shows the distribution of the frequency of these genes across the 531 categories. Most common among these are protein synthesis, transcription, transport, cytoskeleton, DNA replication, mRNA splicing, cell wall genesis, protein degradation, glycosylation, and signaling.

Neural networks can be obtained through the use of learning algorithms that “discover” patterns in the known data and enable them to extrapolate answers to unknown data. We used two types of neural networks, multi-layer perceptrons (MLPs, for which supervised learning algorithms such as the well-known backpropagation are available [7, Chap 4]). Another approach with self-organizing maps (SOMs, for which unsupervised learning algorithm are available) is described in Section 4. Both were trained on the data consisting of the six-feature vectors describing the expression profile of a given RNA molecule. The reader is referred to any textbook on neural nets ([7], for example) for further background details about these types of neural networks.

Input RNAs must be encoded as so-called features (i.e., numerical vectors) to train a neural network. Given the mRNA expression profile data available, the easiest way was to use the 6-feature mRNA expression level for each RNA as a 6D input, and the categories (biological functions) an integer 1-531 as the output.

To fit the model of a multi-layer perceptron, the known data is usually partitioned into a training set and a testing set. Through trial and error, it was determined that selecting 33% of the mRNAs with known categories for training and the remaining 67% for testing gave the best results. In order to preserve the proportion of RNAs in the various categories, a random selection was made from data in each category in these proportions. To avoid unintended patterns in the data (due to alphabetic presentation by names, for example) the exemplars were presented at random in the learning phase. The RNAs with unknown categories were stripped down to only their 6D-feature vector so they could be used as an input after testing to determine results. In order to improve training, categories with less than 30 mRNAs in them were excluded, as they would probably lead to memorization of the inputs by the MLP and thus poor generalization (more below.) Although that reduced the number of RNAs to attempt identification to 1,696 in 25 categories, the levels of confidence for the predictions increases substantially by avoiding poor generalization.

3.1 Training Phase

There were several possible approaches to training a network to predict ncRNA function (as a category). The ideal result is a single network able to correctly
classify all RNAs in S.c. This approach failed for such a large number of RNAs and functions despite many attempts in strategy for coding, for architecture selection and for training. However, the alternate approach of training an individual network for each category separately, was very successful for all categories. That means that 531 networks were trained on the same data (creating separate exemplar sets by changing the desired answer to “1” for those RNAs in the most frequent category “376” and “0” for all other RNAs, for example). Therefore a neural network was created by backpropagation for the purpose of determining whether an RNA (with known or unknown function) is a member of a given category or not. Since all the networks received the same input, these 531 networks can in fact assembled to produce a single MLP with 6 inputs and 531 outputs, each feature value in the output coarsely coding for the categories as a 531D-Boolean vector) whose 1 values may produce appropriate categories for every mRNA.

To create a specific neural network for a given category, the standard training procedure for MLPs was followed: a neural network was first created with randomly assigned weights with mean 0 and small std (standard deviation.) In addition, multiple architectures were used in which only the number of nodes in the hidden layer(s) were changed. Neurons in these layers were assigned sigmoid functions as transfer functions (such as the inverse trigonometric tangent, arctan) for the input and all hidden layers and a pure linear function for the output neuron/node, producing a continuous value in the interval [0,1] for an output to other neurons, or as output of the neural network.

In backpropagation training, there is usually an optimal number of epochs (i.e., repeated presentations of the training data) that gives high values for both. Many choices for the number of epochs to train with were used until each architecture’s optimal number of epochs was determined. This was achieved by choosing the number of epochs where the training and testing percentages were the highest but there was no large drop off from training percentage to testing percentage, i.e., there was little evidence of “memorizing” answers. If the testing percentage is high (> 90%), it was considered a success because the probability is greater than 90% that the predicted category for each RNA is the correct category.

**3.2 Testing Phase**

Once optimal training and generalization rates were obtained, the networks were then put to use in the testing phase. The ncRNA in the testing set put aside were, naturally, coded in the same form (as a 6D vector array of mRNA expression levels) and given to the network for a specific mRNA as an input in order to produce a putative category in which the network would put each. The output determines an answer to membership in a given category by assigning a certain threshold (0.5, consistent with the data) in order to predict whether the RNA belongs in a given category (>=0.5) or not (< 0.5). The generalization performance (i.e., testing accuracy) is based on the percentage of RNAs the neural network predicts correctly for a given category.

Once the training was complete and a satisfactory network was obtained for each mRNA from various architectures using a MLP, we selected the top three architectures that worked effectively on the training and testing data for most mRNAs, as can be seen in Table 1 and Fig. 2. These rates are considered very good to excellent for the typical performance of MLPs in this type of problem. The mean square error (MSE) over all exemplars in Fig. 2 was used along with the training and testing accuracy percentages to determine how well the neural network is predicting the categories among the known data, and so build some confidence interval for the prediction phases, as shown in Fig. 3. The Mean Squared Error (MSE) is the usual average error between predicted values (0 or 1) and actual network outputs for a given RNA. Therefore, the MSE is a measure of the overall quality of the predictions by the neural network.

**Table 1. Average Training and Testing Accuracies for three best MLP classification of ncRNA in S.c.**

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Avg. Train %</th>
<th>Avg. Test %</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6 4 1]</td>
<td>99.84</td>
<td>99.82</td>
</tr>
<tr>
<td>[6 6 3 1]</td>
<td>99.83</td>
<td>99.83</td>
</tr>
<tr>
<td>[6 12 9 1]</td>
<td>99.84</td>
<td>99.80</td>
</tr>
</tbody>
</table>

**Figure 2. Overall performance of the MLPs on the 3,082 known mRNAs, given by the accuracy and MSE in the training and testing phases.** Consistently with the training data, a prediction is considered accurate if the MLP produces a response at or over 0.5 for a target 1, whereas it is in error if the value is under 0.5 for a target 0. Low MSEs indicate confident predictions overall.
3.3 Prediction Phase

Once optimal architectures with high training and generalization rates were obtained on the known data, a composite neural network was used to predict categories based on the 6D-vector of the ncRNA with unknown categories as opposed to mRNA with known categories as in testing. The results are displayed in Fig. 3. The results from the three architectures are not equal, as expected from the results in the testing phase, or even within the same architecture for a particular mRNA, so some more detail is required.

The predictions can be classified into three basic groups. The first group consists of 153 ncRNAs that are being claimed to be in a unique category (i.e., biological process.) Fig. 4 (top) shows them, as sorted by the level of confidence (as defined above) with which the network makes the prediction; Fig. 4 (bottom) shows the corresponding categories.

The second group consists of 391 mRNAs that are being claimed by more than one category. This may appear contrary to the data, in which every mRNA gets assigned a unique category. Upon reflection, however, it makes sense biologically because an mRNA may be involved in several biological processes. We interpret RNAs in this second group as being so, on the evidence presented by the corpus of data.

The third group consists of the remaining 1,596 mRNAs, so-called “orphans” because they were not assigned to any given category by the MLP. This conclusion can be interpreted in two different ways.

One possibility is that the MLPs are not “smart” enough to tell in which category they are. The alternative possibility is that there are yet unknown biological processes that are not present in the original data, so that the MLP is actually discovering hitherto unknown processes at play in S.c., or that these mRNAs do not share many features with the known ones to allow MLPs to ascertain one category.

Table 2. Classification of biological function in S.c. by three best performing MLP architectures.

<table>
<thead>
<tr>
<th>mRNA Group</th>
<th>Known</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Unique Cat)</td>
<td>171</td>
<td>74</td>
</tr>
<tr>
<td>2 (Various Cats)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3 (Orphans)</td>
<td>2,908</td>
<td>2,740</td>
</tr>
<tr>
<td>Totals (5,899)</td>
<td>3,082</td>
<td>2,817</td>
</tr>
</tbody>
</table>

4. Genomics with Self-Organizing Maps (SOMs)

A second set of predictions was produced by categorizing the ncRNAs (i.e., RNAs with unknown functions) using another neural network technique, the so-called Self Organizing Map (SOM) [7, Chap 9]. SOMs afford an unsupervised learning algorithm. While supervised learning imposes a specific assignment of biological function by requiring a label (“teacher”) on each input for training, SOMs takes the
inputs with no labeling and thus produces a classification with no \textit{a priori} assumptions about how the inputs should be clustered together, based solely on any patterns of similarity that might be identified in the course of training. Therefore they have the potential to identify objective classifying criteria that may be complementary (or even conflicting) with MLPs or even individual researchers’ ontologies, but which, on the other hand, might suggest some more objective criteria on the transcriptomics of \textit{S. c.} for further analyses.

SOM training uses the same input as the multi-layer perceptron (a 6D-feature vector) but produces a so-called \textit{topological map} of “locations” or “nodes” on a metric space (such as a common geometric plane or 3D space) whose relative proximity (or distance) is arranged to capture the relative similarities (or differences, respectively) among the clusters represented by the various locations/nodes. Therefore, the features that are mapped in the prediction phase to the same node are close enough to be considered to belong to the same category.

\subsection*{4.1 Training Phase}

Various topologies were tried to develop a SOM of the given data, including 1D and 2D architectures of various sizes (18x18, 24x24, 30x30, and 40x40). Eventually, as before, we selected the top performing couple of architectures, namely 40x40 locations/nodes. The 6D-feature vectors of the entire set of 91 RNAs representing approximately the top 50%+ of the RNAs with known categories with 30+ mRNAs were taken from the original data and used as the input to the SOM for training. The training consisted in presenting an input vector to the SOMs, identifying the location with maximum output on that input (called “the winner”), strengthening the connections to it so that next time the same will happen again, and weakening the connections to other nodes proportionally to their topological distance (here in the plane) from the winner (hence the name “winner-takes-all” used to describe this type of network.) The SOM maps were trained for 15,000 epochs (3,000 in the ordering phase and the rest in the converging phase [7, Chap. 7]).

Once trained, the SOM will do a “forward” pass and produce a classification into a \textit{unique} location (the winner for that input) for any given node. RNAs mapped to the same node can be regarded as belonging to the same category, which then has to be identified using prior knowledge about the data.

\subsection*{4.2 Labeling Phase}

Once the SOM is trained, it is necessary to figure out the meaning of the classification being made by it, before proceeding to the prediction phase. That requires inspection of the results in light of preliminary prior experience with the data in order to figure out what input patterns each node may be capturing.

Ideally, every location/node in the SOM should be regarded as defining a single category (here, a biological process), although RNAs in a category could be mapped to several locations, which together would represent that category. In particular, if a location gets only RNAs from a single category, it is clear it should represent that category. Locations capturing an overwhelming number of mRNAs (over 80%) from a single category were also considered to be “uniquely” labeled by that category. The full category itself is thus represented by all such locations. There were 932 such locations and they turn out to represent about 244 unique categories shown in Fig. 5, a hit rate of over 94%. The top categories labeling locations with high confidence are shown by the distinct labels in Table 3.

Figure 5. Results of classification by a 40x40 SOM for mRNAs in 91 biological functions capturing all unknown mRNAs, trained over 15,000 epochs. Shown are the 932 uniquely labeled locations (as described in the text) representing 244 (out of 259, over 94%) categories (biological functions) over the 1,600 locations. The radii of the circles are proportional to the number of mRNAs in a category being mapped to each location.

The remaining locations get RNAs from more than one category, and therefore the label (the category they represent) is not obvious. They might represent mRNAs involved in several higher-level biological functions. More sophisticated analyses or additional data may be required to produce putative single categories for these mRNAs.
4.3 Prediction Phase

Once the SOM has been labeled, we proceeded to obtain its classification for the unknown ncRNAs by presenting them to the network. This time, the unknown ncRNAs can be placed in two groups, those for which a highly confident prediction is made, and those for which the prediction is unclear. Some of the results are shown in Table 3. Once again, further analyses or additional data may be required to produce putative categories for the latter mRNAs.

Table 3. Summary of predicted unique identification with high confidence of most important biological function for 74 mRNAs in S.c. by MLPs and/or SOMS architectures described above. The prediction for boldfaced mRNAs are matched by both types of architectures and therefore can be assigned a very high level of confidence. Other putative predictions are made by the MLP (second column) or SOM (third column) architecture only and so are made with less confidence (not all of them are shown for lack of space.)

<table>
<thead>
<tr>
<th>Gene</th>
<th>MLP</th>
<th>SOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycolysis</td>
<td>YOR324C</td>
<td>YOR324C</td>
</tr>
<tr>
<td>Protein Synthesis</td>
<td>YAR116C</td>
<td>YOR116C</td>
</tr>
<tr>
<td>Transport</td>
<td>YNL056C</td>
<td>YNL056C</td>
</tr>
<tr>
<td>Protein Folding</td>
<td>YDL120C</td>
<td>YDL120C</td>
</tr>
<tr>
<td>Transcription</td>
<td>YLL110C</td>
<td>YLL110C</td>
</tr>
</tbody>
</table>

5. Discussion and Conclusions

We have used two major computational intelligence techniques, MultiLayer perceptrons (MLPs) and Self-Organizing Maps (SOMs), to identify putative biological functions of over 170 non-coding RNAs in S.c. together with an analysis that provides confidence levels about the identifications being made, in three groups. In the first, 174 of the unknown RNAs are given putative biological functions by one of the two networks. In the second group, 26 of the remaining RNAs are given a putative biological function that is less certain but still worthy of consideration given that none has been hitherto been suspected. In the remaining group of RNAs, “hard core” RNAs remain elusive to classification. The first two groups of putative categories may now be subjected to further validation by the biological community, and may be representing a new ontology worthy of further research, given other successes in the application of MLPs and SOMs [7].

Some of the results presented here can be ported to other cell systems. The theory of grand unification [4] holds that information about a shared gene and associated proteins contributes to our understanding of all the diverse organisms that share it, so that knowledge of such roles illuminate and provide strong inference of its role in other organisms. For example, about 12% of the worm genes (~18,000 genes) encode proteins whose biological roles could be inferred from their similarity to their putative orthologues in S.c. (or about 27% of the S.c. genes) [6]. Further, most of these proteins have been found to have a role in the ‘core biological processes’ common to all eukaryotic cells, such as DNA replication, transcription and metabolism [4]. It would not be surprising if the same is true of genes with noncoding RNAs.

Finally, a word of caution is in order in assessing the results presented in this paper in the proper context. It is important to keep in mind (i) that a given category refers to either a gene (which is a DNA molecule) or its transcript (which is an RNA molecule), depending on the context, and (ii) that the intracellular level of an RNA molecule at any given time (referred to as the transcript level, TL) is determined by the balance of two factors – the transcription rate (TR) and transcript degradation rate (TD), which can be algebraically represented by the following equation:

\[ \frac{dTTL}{dt} = TR - TD \]  

where \( \frac{dTTL}{dt} \) indicates the rate of change in TL with time [2]. Garcia-Martinez et al. [3] measured both genome-wide TL and TR simultaneously in S. cerevisiae following glucose-galactose shift. Some examples of their data are plotted in Figure 6. As evident in Fig. 6 (top), the average behaviors of the glycolytic and oxidative phosphorylation, RNA trajectories (also called TL kinetics, TL traces, or gene expression profiles) reflect the metabolic functions of the genes coding for the RNA molecules in a goal-directed manner. The glycolytic RNAs decrease, since they are no longer needed due to the removal of glucose, while the oxphos RNAs are required to metabolize ethanol left over from the previous glucose metabolism and the new nutrient galactose [2, 5].
One of the most surprising findings of Garcia-Martinez et al. [3] is that the rate of change in TL, i.e., dTL/dt, can vary independently of TR. This observation is supported by the fact that, although TR can change in the same direction for the glycolytic and oxphos pathways, their TL kinetics can be opposite (see the TL and TR traces between 5 and 360 minutes in Fig. 6, top and bottom.) This finding cannot be explained unless we take into account the transcript degradation rate (TD), in agreement with Eq. (1). Thus, an important conclusion one can draw from Eq. (1) is that “It is impossible to infer the genes responsible for metabolic functions solely based on analyzing gene expression profiles.” [9]. However, theoretical considerations strongly indicate that it should be possible to infer the metabolic functions of unknown RNAs based on the similarity of their TL traces with those of known RNAs [5]. Although large-scale prediction [8] have been used before in analyses of microarray data, a novel contribution of this paper is that the analyses has been carried out in a researcher-independent ontology implicit in the inherent properties of neural networks, which are based solely on the given data.

6. References


