Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 498031, 9 pages
doi:10.1155/2012/498031

Research Article

Artificial Neural Network for the Prediction of Tyrosine-Based Sorting Signal Recognition by Adaptor Complexes

Debarati Mukherjee, Claudia B. Hanna, and R. Claudio Aguilar

Department of Biological Sciences, Purdue Center for Cancer Research and Center for Science of Information, 201 S University Street, Hansen Life Sciences Building, West Lafayette, IN 47907, USA

Correspondence should be addressed to R. Claudio Aguilar, claudio@purdue.edu

Received 12 July 2011; Accepted 3 November 2011

Academic Editor: Alejandro Giorgetti

Sorting of transmembrane proteins to various intracellular compartments depends on specific signals present within their cytosolic domains. Among these sorting signals, the tyrosine-based motif (YXXØ) is one of the best characterized and is recognized by μ-subunits of the four clathrin-associated adaptor complexes (AP-1 to AP-4). Despite their overlap in specificity, each μ-subunit has a distinct sequence preference dependent on the nature of the X-residues. Moreover, combinations of these residues exert cooperative or inhibitory effects towards interaction with the various APs. This complexity makes it impossible to predict *a priori*, the specificity of a given tyrosine-signal for a particular μ-subunit. Here, we describe the results obtained with a computational approach based on the Artificial Neural Network (ANN) paradigm that addresses the issue of tyrosine-signal specificity, enabling the prediction of YXXØ-μ interactions with accuracies over 90%. Therefore, this approach constitutes a powerful tool to help predict mechanisms of intracellular protein sorting.

1. Introduction

A defining characteristic of eukaryotic cells is the presence of membrane-bound intracellular compartments. These membranous structures host specific biochemical processes by virtue of their distinctive lipid and protein composition [1]. Nevertheless, in order to be able to contribute to the physiology of the cell, this array of processing stations needs to be linked and coordinated by a robust trafficking system of membranous carriers [1, 2]. Indeed, the transport of cargo by this system plays a crucial role in the establishment/maintenance of each compartment's identity and in the delivery of substrates [1, 2].

Given the outstanding relevance of protein trafficking for the onset of diseases, as well as the significance of trafficking in pathogenic infection [3, 4], understanding the mechanisms by which the cell targets its proteins to the appropriate compartment has been the focus of multiple labs [5–9]. A landmark achievement resulting from these efforts was the realization that some transmembrane proteins contain sorting signals embedded in the aminoacid sequence of their cytoplasmic segments [9]. These signals are recognized by intracellular receptors that mediate the protein inclusion in, or exclusion from, trafficking carriers [9]. Among this signal-recognition machinery, the tetrameric clathrin-associated Adaptor Proteins (APs) emerge as major players in the protein trafficking system [9, 10]. Four different AP complexes (AP-1 through AP-4) with distinctive intracellular localizations have been identified and they are believed to mediate different protein sorting events from and/or to several compartments [11, 12]. Whereas other subunits are engaged in interactions with various molecules, the medium AP μ subunit is in charge of recognizing tyrosine-based sorting signals fitting a XXXYXXØ consensus (where X = any amino acid; Y = tyrosine and Ø = residues with a bulky hydrophobic side chain such as phenylalanine, leucine, isoleucine, methionine, and valine) [6, 9, 13, 14].

Although the Y and Ø residues within these signals are critical for μ subunit binding, it is known that the less conserved X-positions play an important role in defining the specificity of different Y-signals for different AP complexes [14, 15]. In fact, the differential interaction of signals with

APs is responsible for the ultimate intracellular localization of the corresponding cargo.

The two-hybrid technology was used by the Bonifacino lab at NIH to conduct the most comprehensive study of μ subunit specificity for Y-signals available to date [14, 15]. Specifically, this group used the different μ subunits (μ1–μ4 from AP-1 through AP-4, resp.) as "baits" to screen a two-hybrid XXXYXXØ signal-library. The sequences of the signals selected by each μ subunit were established and the data was statistically analyzed. Further, each set of signals selected by a particular μ subunit was tested against the other μ chains generating a vast amount of data about the signal binding preferences of APs. These investigations provided unique and extremely valuable information about the signal specificity of μ subunits [14, 15]. However, they also highlighted the complexity of μ/Y-signal recognition process; particularly by indicating that combinations of residues at certain X-positions display (positive or negative) cooperative effects, thereby affecting the overall ability of signals to interact with μ subunits [14, 15]. Unfortunately, these interdependence effects made it impossible to extract explicit rules for predicting recognition of Y-signals by AP μ subunits. A classical alternative to rule-based analytical models is the Artificial Neural Network (ANN) paradigm [16–18]. ANNs analyze existing examples of the phenomena under study and, through an iterative process ("training" or "learning"), mathematically encode their behavior for predictive purposes [19–21]. A critical requirement for the success of ANN approaches is that a critical mass of information be available for training [22]. Since this precondition is satisfied in the case of Y-signal recognition by μ subunits [14], we designed, trained, and validated ANNs for the prediction of μ/Y-signal interactions.

Our results indicate that trained ANNs were capable of predicting the experimental outcomes of previously published two-hybrid experiments with over 90% accuracy. Further, ANNs also successfully forecasted the results from novel two-hybrid experiments involving Lamp2 and CD63 mutant signals with μ subunits. Importantly, ANNs were proficient for correctly predicting two-hybrid results even in the presence of positive or negative cooperativity effects among residues within a Y-signal. Indeed, the ANNs' predictions were correlated with the intracellular localization of transmembrane proteins bearing analyzed signals.

In summary, our results demonstrate that application of the ANN paradigm is suitable for the prediction of μ/Y-signal interactions and providing a solution to this important problem in cell biology. To further improve the system performance, we encourage our colleagues to submit their own experimental results to be used in future rounds of training and validation.

2. Materials and Methods

2.1. Plasmids and Strains

2.1.1. DNA Constructs. Plasmids used in this study were prepared using standard techniques and following the general design described in [14]. Thus, XXXYXXØ signals were cloned in-frame with the TGN38 cytoplasmic tail in the multiple-cloning site of the two-hybrid vector pGBT9 (Clontech).

Site directed mutagenesis was done using the QuikChange kit (Stratagene, La Jolla, CA).

2.1.2. Yeast Culture Conditions and Transformation Procedures. Yeast two-hybrid strain AH109 (Clontech) was grown in standard yeast extract-peptone-dextrose (YPD) or synthetic medium with dextrose lacking appropriate aminoacids for plasmid maintenance at 30°C for 3-4 days unless indicated otherwise. Transformations were performed by standard Li-Acetate transformation procedures (Clontech yeast handbook).

2.2. HeLa Cell Culture and Transfection. HeLa cells (American Type Culture Collection, Manassas, VA) were cultured in DMEM supplemented with 10% (vol/vol) FBS/100 units/mL penicillin/100 mg/mL streptomycin (Biofluids, Rockville, MD). The night before transfection, cells were seeded onto six-well plates (Costar) in 2 mL of medium. The following day, the cells were transfected with the TAC constructs in pXS using Fugene-6 reagent (Roche Molecular Biochemicals). Twenty-four hours after transfection, cells were fixed and analyzed for expression of the TAC constructs by immunofluorescence microscopy with the 7G7 anti-TAC monoclonal antibody.

2.3. Immunofluorescence Microscopy. HeLa cells transiently transfected with TAC constructs were grown on coverslips, fixed with 4% formaldehyde and incubated with the 7G7 mouse monoclonal anti-TAC antibody diluted 1 : 500 in DMEM, 10% FCS, 0.1% saponin for 1 h at room temperature. After washing with PBS, coverslips were incubated with a goat anti-mouse IgG antibody conjugated to Alexa488 for 1 h. Coverslips were washed with PBS and mounted on slides using Aqua-PolyMount (Polysciences) and imaged in a Zeiss Axiovert 200 M microscope.

2.4. Two-Hybrid Experiments and Result Coding. Potential interactions between XXXYXXØ signals and a given AP μ subunit was tested using the two-hybrid technology as previously described [14]. Briefly, plasmid DNA encoding for GAL4 DNA Binding Domain (G4BD)-XXXYXXØ and Gal4 Activation Domain (G4AD)-μ fusion proteins were transformed into AH109 yeast cells bearing GAL4-based reporter genes. If the μ moiety is capable of binding the Y-signal of the DNA-bound G4BD-XXXYXXØ fusion, then the G4AD-μ will be recruited to the reporter gene leading to gene activation (Figure 1). The presence of the reporter gene product, for example His3 (an enzyme involved in the biosynthesis of the aminoacid histidine), will allow the cells to grow in selective media, that is, plates lacking histidine (−His, see Figure 1). Therefore, cell growth in −His media, visualized as yeast colony formation, constitutes the experimental readout that corresponds to μ/Y-signal

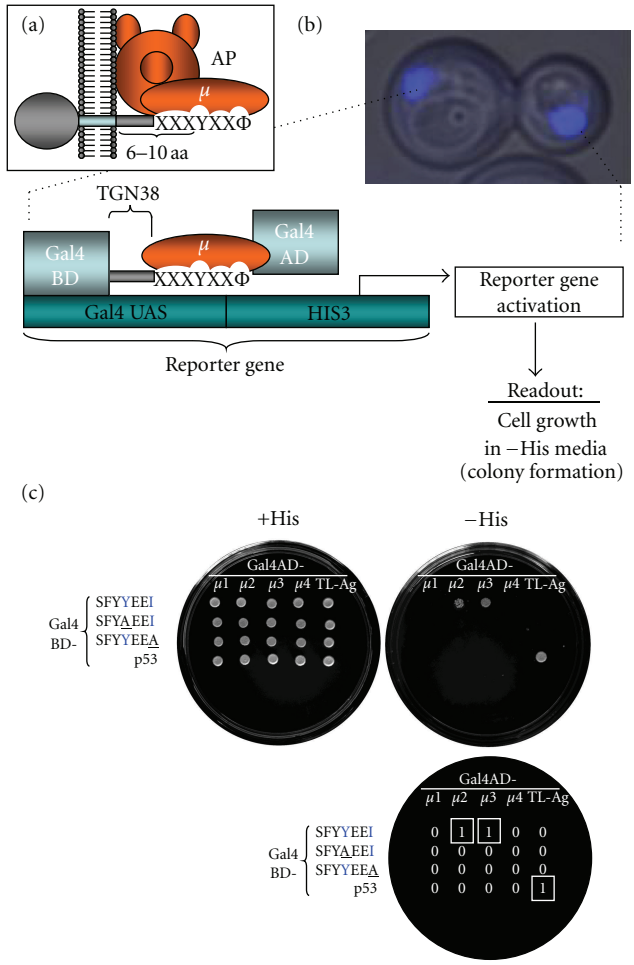

Figure 1: Two-hybrid approach and result coding. (a) Clathrin-associated adaptor complexes bind Y-signals. Scheme depicts a Y-signal (fitting into a XXXYXXØ consensus) within the cytoplasmic tail of a transmembrane protein bound by an adaptor complex (AP in orange). X represents any aminoacid and Ø a residue with a bulky hydrophobic side chain (F, M, I, L and V). The AP's μ-subunits bind signals located at about 6–10 aminoacids from the transmembrane domain. (b) Two-hybrid strategy used in this study. Yeast two-hybrid strain (AH109) bearing-integrated reporter genes were transformed with plasmids expressing the Gal4 binding domain (BD) fused to a XXXYXXØ-signal (via a TGN38-derived spacer) and the Gal4 activation domain (AD) fused to the C-terminus of an AP μ subunit. The *GAL4* upstream activating sequences (UAS) within the reporter gene are bound by the Gal4BD-Y signal fusion. If the expressed μ subunit binds the featured signal, then the Gal4AD activates the *HIS3* open reading frame. His3 production allows the cells to grow in absence of the aminoacid histidine (−His), leading to the formation of colonies. (c) Result coding: The colony formation two-hybrid readout was coded as follows: growth in −His (μ/Y-signal interaction) = 1, whereas absence of growth in −His (lack of interaction) = 0. The SFYYEEI signal used as example was isolated in a combinatorial two-hybrid screen. Signal's critical Y and Ø (I in this signal) are indicated in blue and were alternatively mutated to A. The interacting pair mouse p53 and SV40 T-large antigen (TL-Ag) was used as a positive control and as negative control when cotransformed with any other construct.

interaction. The two-hybrid results were coded as follows: when visible colonies were formed an *Interaction Value*, $V = 1$ was assigned; if no colonies were observed the *Interaction Value* was 0 (Figure 1).

2.5. Data Sets. In this work, we used AP μ-subunit/ Y-signal interaction data coming from two-hybrid library screens, most of which have been previously published [14, 15].

(a) Training Set: We used extensive collections of about 200 μ/Y-signal interaction data per μ subunit [14, 15] to train neural networks for the prediction of the interaction of XXXYXXØ sorting motifs with different adaptor μ subunits. Since it has been recently demonstrated that μ4 is capable of binding two types of sorting signals via two different binding sites [23], we did not train an ANN for prediction of Y signal interactions with this medium subunit. However, we used data corresponding to the analysis of cross-reactivity of other μ-subunits with Y-signals isolated in a μ4 screen.

(b) Validation Set: In order to test the generalization capabilities of our neural network, we used a second set of μ-sorting signal interaction data including a reserved group (not used for training) from the published screens [14] and also naturally occurring Y-based targeting motifs previously tested by using the two-hybrid technology [15, 24–26].

3. Results and Discussion

Here we describe a novel approach to the analysis of protein trafficking mediated by sorting signals. Specifically, we describe the design and application of an artificial intelligence approach based on the neural network paradigm.

We trained three different ANNs, which predict whether a given Y-based sorting signal will be recognized or not by three adaptor medium subunits (μ1, μ2, and μ3). Although it is clear that μ4 binds to Y-signals in a Y- and Ø-dependent manner, it recognizes at least two kinds of sorting signals [23]. Therefore, since μ4 two-hybrid screens for Y-signals may have produced mixed results corresponding to more than one type of signal selected, we excluded this medium subunit from the current development. Following training, ANNs (one per adaptor medium subunit) were assembled in a single system. Algorithm and current weight sets are freely available upon request.

3.1. Design of ANN for the Prediction of μ/Y-Signal Two-Hybrid Interaction. ANNs are algorithms capable of predicting the outcome of complex processes not viable for deconvolution into simple sets of rules [19, 20]. Therefore, we reasoned that these approaches would be suitable for the analysis and prediction of μ/Y-signal two-hybrid *Interaction Values* (see Figure 1 and Section 2.4).

A typical artificial neural network (see Figure 2 for an example) is made up of independent computing units ("neurons") organized in "layer" groups. Following adjustment by the corresponding "connection weights", the computing

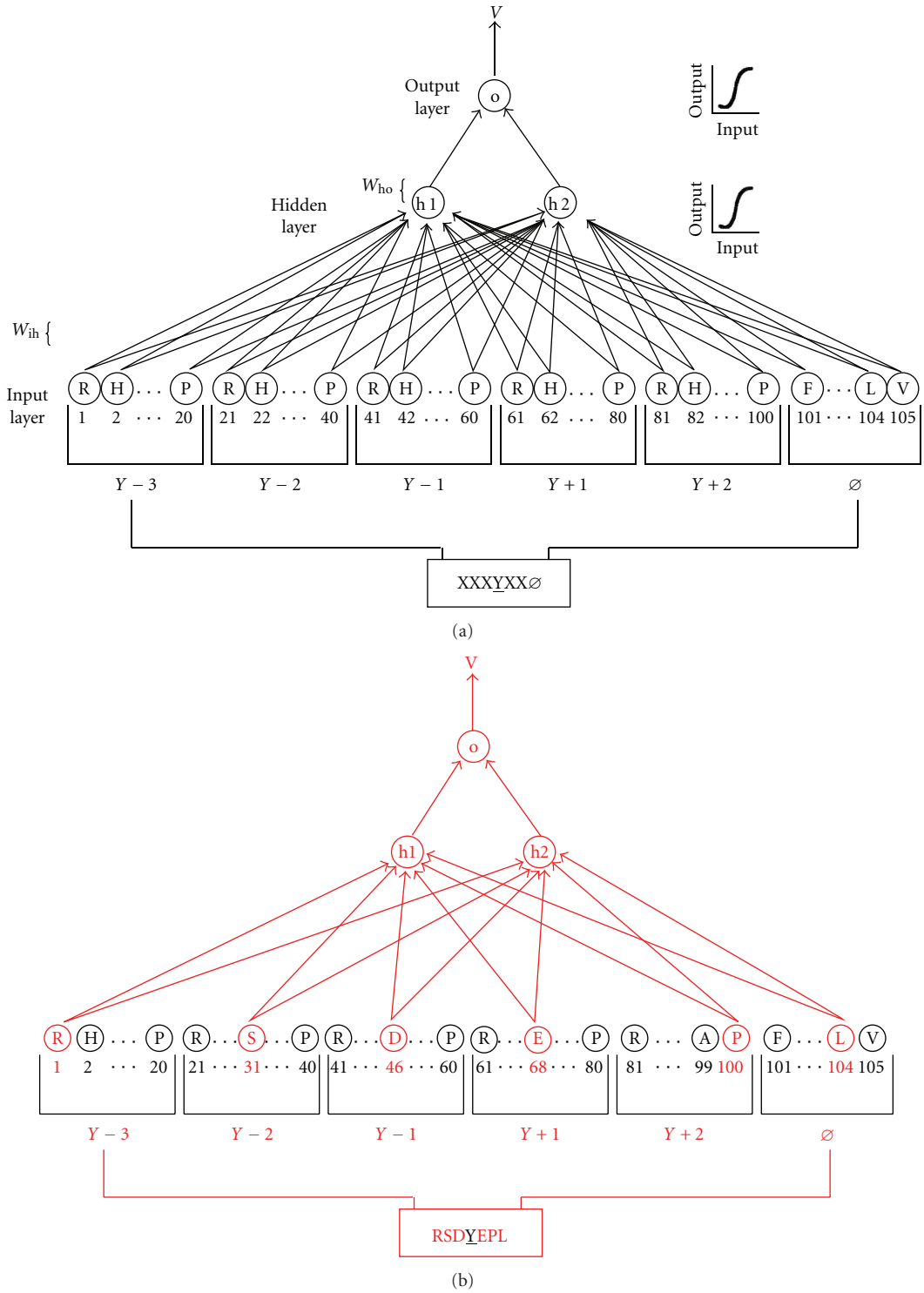

Figure 2: Neural networks for the analysis of Y-signals. (a) ANN architecture: the neurons in the network are represented by circles and the connections between units by arrows. The input layer is made up of 5 clusters (one for each X position within the XXXYXXØ motif) containing 20 nodes each (representing the 20 possible residues—only 3 per cluster is shown) plus the Ø cluster with only 5 nodes (for F, M, I, L, and V), yielding 105 neurons in total. Neurons from the hidden layer are labeled h1 and h2, whereas the output neuron is marked o, both types of units rely on a logistic activation function, depicted as a sigmoidal output-input response. Final network output is denoted as V (*Interaction Value*). The weights associated to the input hidden layer and hidden-output layer connections are indicated as W_{ih} and W_{ho}, respectively. The Bias neurons are not showed. (b) ANN Signal analysis. Sequences (a hypothetical RSDYEPL signal is shown in red) are analyzed at every position. Within each of the 6 input clusters, only the neuron representing the aminoacid present at that position is activated (represented in red). This group of input neurons "fire" to the each hidden neuron according to the corresponding connection weight. Each hidden neuron compiles a total input and elaborates a sigmoidal output that is sent to the o-neuron, which in turn produces the network output V.

results from lower layers are used by the upper layer neurons as inputs for their own calculations.

During "training", a neural network uses iterative processes to adjust its internal parameters (connection weights) so that its output function can produce the expected response (e.g., *interaction value*) for each element of a large set of known experimental data. If properly trained and validated, the network will predict unknown experimental outcomes.

The tyrosine-signal neural network (TySNN) is a feed-forward ANN designed to address the question: *"Does this AP μ subunit bind this Y-signal?"* by predicting an *interaction value, V*.

After trying several network architectures (not shown), we concluded the most robust system consisted of one hidden layer containing 2 neurons fully connected with the input layer as well as with the unique node within the *output layer* (Figure 2(a)). Therefore, TySNN is made up of three neuron layers: an input layer (106 neurons), a hidden layer (2 neurons, h1, and h2), and one output (o) neuron (Figure 2(a)). The input layer is comprised of 5 clusters that represent each X-position in a XXXYXXØ signal. Each cluster contains 20 neurons representing the 20 possible aminoacids that can be found at that specific X-position. A sixth cluster of 5 neurons represents the 5 possible aminoacids (F, M, I, L, and V) to be found at the Ø-position (Figure 2(a)). An extra, constitutively activated, "bias"-neuron [20] was added yielding a total amount of 106 input neurons.

The network reads each position of the XXXYXXØ signal and sends inputs to every neuron in the corresponding position-cluster. Within a cluster, an input = 0 is sent to all neurons except to the one representing the aminoacid found at the position and that receives an input = 1 (Figure 2(b)). All neurons from the input layer send an output value to both hidden neurons equal to their input multiplied by the corresponding connection weights (W_{ih}, Figure 2(a)). The resulting values constitute the input to the hidden layer. Each hidden neuron compiles a total input and elaborates an output following a sigmoidal activation function (see Appendix and [19, 20] that is transmitted to the output neuron according to their corresponding W_{ho} weights (Figure 2(a)). In turn, the output neuron sums the inputs coming from both h1 and h2 and elaborates the network output (predicted *Interaction Value, V*) through its own sigmoidal activation function. The network predicted *V* values are translated from a real number in the range (0.0; 1.0) into an appropriate binary output. Thus, an arbitrary output value >0.5 is considered a "yes" result while any value ≤0.5 means "no" (i.e., *There is* or *there is not* an interaction between the sorting signal and the μ subunit, resp.).

3.2. Evaluation of the Artificial Neural Network Performance. The networks were initialized using small weight values randomly generated and following a normal distribution with mean = 0.00 and standard deviation = 1/[number of neurons]1/2 (i.e., ≈0.10) ([20] and Figure 3(a)). During training, the predicted binary *V* values (see above) were compared to the known experimental results (training set, [14]) and the weights were modified to minimize the differences (see appendix for details). More specifically, training was performed following a "batch" scheme; that is, the weight changes were accumulated and only applied after one run of the whole set of training examples or "epoch" (see appendix for further details on the algorithm and network architecture). The process was repeated until convergence was attained (Figure 3(b)).

Two parameters were used to measure the performance of the neural networks.

(1) *Accuracy* (*A*). Represents the ratio between the number of correctly predicted outcomes (*C*) and the total number of examples (*N*).

$$A = \frac{C}{N}. \tag{1}$$

(2) *Mathews' correlation coefficient* (MCC) [27].

$$\text{MCC} = \frac{pn - uo}{(n+u)(n+o)(p+u)(p+o)} \times \frac{1}{2}, \tag{2}$$

where *p* is the number of true positives predictions, *n* the number of true negatives predictions, *u* the number of false positives, and *o* the number of false negatives. MCC is used as a reliable performance indicator that is independent of the proportion of positive and negative results in the training set [28].

Accuracy and the total error E (see appendix) were also used to monitor the evolution of network learning during training (see Figure 3(c) for an example).

In general, the shape of the curves obtained indicated the presence of local minima (Figure 3). In fact, some of our networks' current weight sets may correspond to low local, rather than global, minima.

Table 1 summarizes the performance of the networks following training. In all cases we observed above 90% accuracy in predicting the result of a potential μ/Y-signal interaction. These values support the suitability of the ANN paradigm for predicting Y-signal specificity for clathrin-associated adaptor complexes.

We believe the accuracy of the networks can be further improved with subsequent training, aiming to reach the global minima. However, in order to avoid overtraining with a single data set, new results should be used. Therefore, we encourage our colleagues to participate in this effort by submitting their own μ/Y-signal binding results. In addition, the spreadsheet macro that runs the ANN algorithm is freely available upon request.

Table 1: TySNN performance.

	μ1[a]	μ2[a]	μ3[a]
A[b]	0.95	0.96	0.93
MCC[c]	0.98	0.99	0.96

a. ANN trained for the Y-signal preference of the indicated μ subunit.
b. Accuracy: ratio between correct and total number of examples. Value range: [0; 1].
c. Mathews' Correlation Coefficient: see text for details. Value range: [−1; 1].

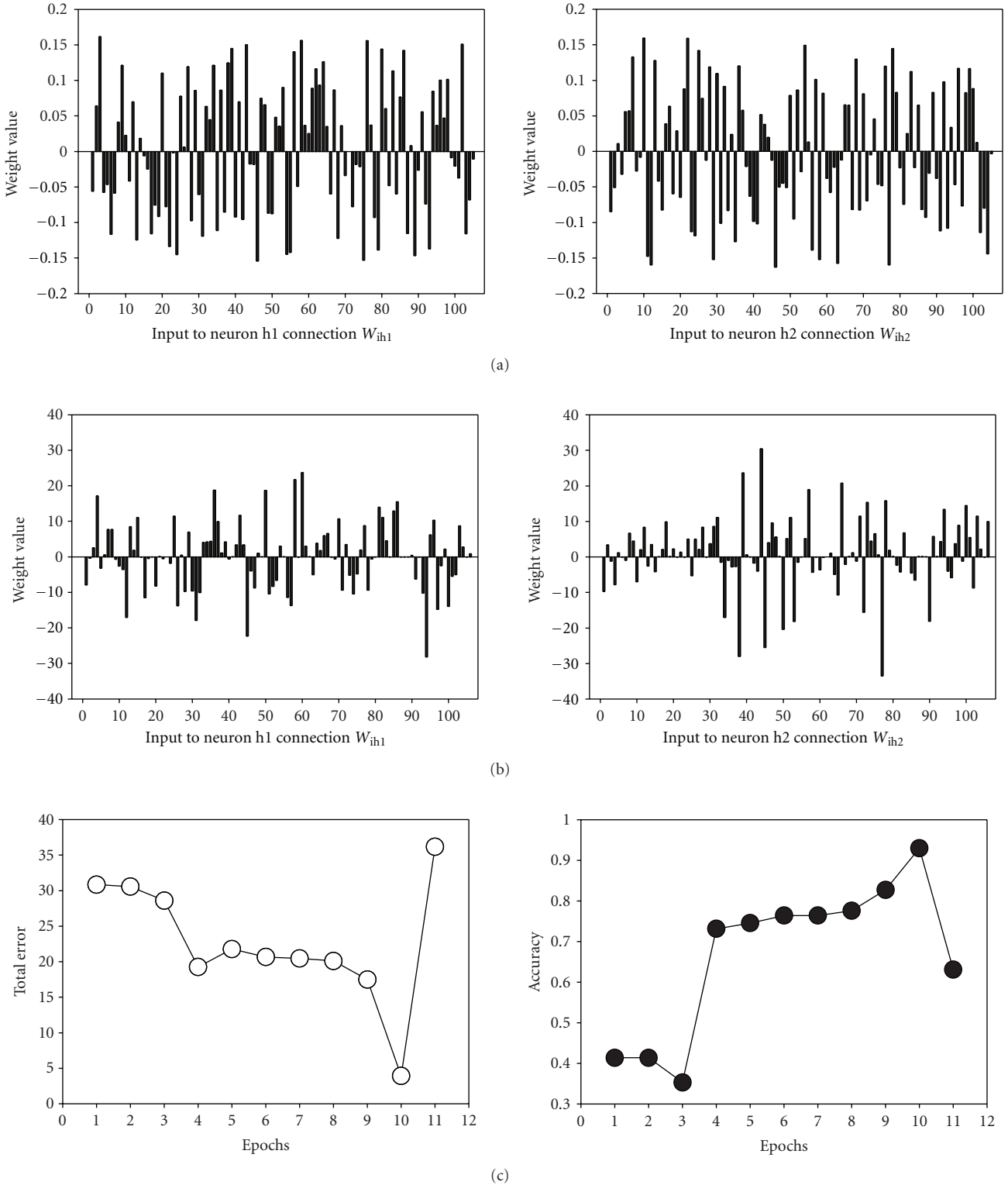

Figure 3: ANN training. Connection weights between the Input and hidden layers before (a) and after training (b): values shown were taken from the training of $\mu 1$ ANN. weights were initialized with small random quantities (0.00 ± 0.10) as shown in (a) and converged into a broader value range (b). (c) ANN training. The evolution of the total error and accuracy (see main text and Appendix for details) was monitored as a function of the number of Epochs (i.e., number of iterations of a complete set of sequences).

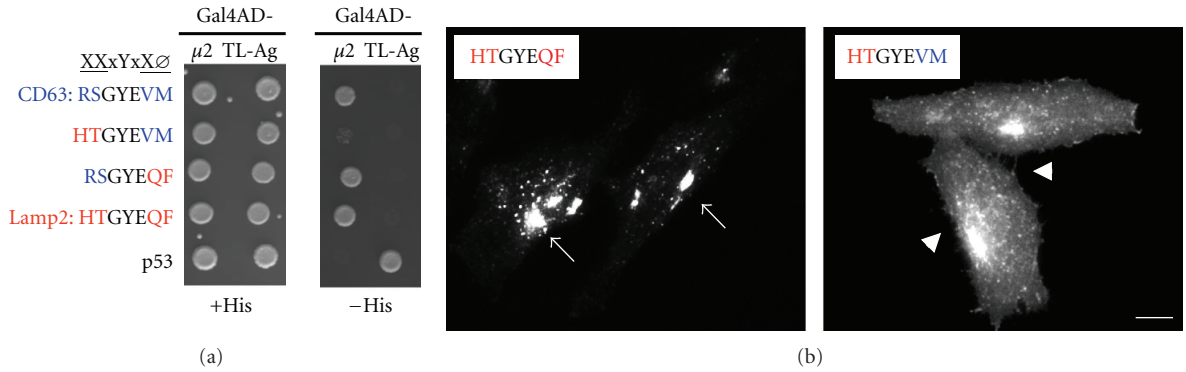

FIGURE 4: Binding of Cd63-Lamp2 chimeric Y-signals to $\mu 2$. (a) Two-hybrid experiments between $\mu 2$ subunit and WT or chimeric Y-signals were performed as described in Figure 1 and Section 2.4. Cd63-specific and Lamp2-specific residues are denoted in blue and red, respectively. (b) HeLa cells transiently transfected with TAC-fusion proteins were fixed, permeabilized, and incubated with mouse anti-TAC antibody followed by an Alexa 448-conjugated secondary antibody. Representative cells showing TAC fusion protein localization are shown. Arrows (left panel) point to intracellular structures observed in TAC-HTGYEQF (Lamp2 signal) fusion; Arrowheads (right panel) highlight plasma membrane enrichment of TAC-HTGYEMV chimeric signal fusion protein. Scale bar: 1 μm.

3.3. Biologically Relevant Predictions and Detection of Cooperative Effects among Residues within a Signal. ANNs described in this work were trained using two-hybrid interaction data. Therefore, ANNs predict two-hybrid interaction values from experiments performed under similar conditions (see Section 2.4). It should be noted that two-hybrid results can significantly correlate with the targeting behavior of proteins expressed in cells [15].

Analysis of the relative relevance of residues within the signal suggests that positions Y − 3, Y − 2, Y + 2, and Ø usually have major effects on the overall ability of the Y-signal to interact with μ subunit.

Importantly, TySNN was able to correctly predict the specificity of a subset of naturally occurring signals, including the sorting signals for lamp2 (HTGYEQF) and CD63 (RSGYEVM). Interestingly, these signals display a similar interaction pattern against the different μ subunits: both could bind $\mu 2$ and $\mu 3$ but showed negligible interaction with $\mu 1$ [15]. Although the residues immediately flanking the critical Y within these signals are identical (Y − 1 and Y + 1), the ones occupying the positions Y − 3, Y − 2, Y + 2, and Ø are different (Figure 4(a)).

In order to test the relevance of these residues for the interaction of these naturally-occurring and highly similar Y-signals with μ subunits, we asked TySNN to predict the specificity of chimeric signals as indicated in Figure 4. Surprisingly, TySNN predicted negligible reactivity of the chimeric signal HTGYEVM with $\mu 2$. This prediction was surprising as $\mu 2$ has been described as the medium subunit with the most relaxed specificity [14]. Also, through this result, TySNN indicated the existence of negative cooperative effects among residues at different positions within a signal. Importantly, we tested this prediction experimentally and observed a complete correspondence with actual two-hybrid results (Figure 4(a)).

Further, we introduced both lamp2 (HTGYEQF) and the chimeric (HTGYEMV) signal into the cytoplasmic tail of interleukin-2 receptor α-subunit (also known as TAC) and expressed them in heLa cells. Intracellular localization of TAC-fusion proteins can be easily detected by immunofluorescence with an anti-TAC antibody (7G7). In fact, the TAC-Lamp2 fusion protein showed a largely intracellular, perinuclear immunofluorescence staining, compatible with a late endosomal-lysosomal localization (Figure 4(b)). In contrast, the TAC-chimeric signal fusion protein showed a strong plasma membrane staining compatible with deficient internalization due to impaired recognition by $\mu 2$ (Figure 4(b)). These results support the applicability of the predictions of the ANN system to *in vivo* intracellular˜trafficking problems.

4. Conclusions

Our results indicate that ANNs can handle the complexity of the μ/Y-signal interaction process. Therefore, candidate protein cargo with a suitable Y-signal within their cytoplasmic tail can be identified based on their predicted ability to interact or not with the various μ subunits. However, the investigator should be aware that for a YXXØ motif to be recognized by APs *in vivo,* it must also satisfy other requirements, for example, proper spacing from the corresponding transmembrane domain [9]. As mentioned in previous sections, further training with additional naturally occurring Y-sorting signals should enhance the predictive power of this approach towards cytoplasmic domains of transmembrane proteins.

Importantly, trained ANNs have been successfully used to extract information about the principles ruling the phenomenon under study [29]. Therefore, we anticipate that upon further developments, results obtained with TySNN will contribute to the establishment of explicit rules for the analysis of Y-based sorting signals. In fact, this work already reports the conclusions concerning the relative importance of certain X-positions for the recognition of the Y-signal by the different AP medium subunits. Moreover, improvements to the algorithm reported here will be directed to provide for the capability to analyze quantitative data rather than

binary "Yes/No" results. Specifically, ANNs can be trained to predict the strength of μ/Y-signal interaction based on β-galactosidase activity or cell growth in the presence of different concentrations of the competitive inhibitor 3AT in two-hybrid experiments [24].

Finally, we envision that this approach may be used in the analysis of results from future screens. For example, there is almost no information regarding the specificity of APs for signals in plants and *Saccharomyces cerevisiae*. Therefore, we believe a systematic study of μ/Y-signal interactions, like the ones conducted by the Bonifacino lab [13–15], should be pursued in yeast and plants.

Along the same lines, a screen to define the specificity of APs for dileucine signals is also lacking. The Bonifacino lab also developed a successful three-hybrid approach [30] that should be adapted for the screening of putative combinatorial dileucine signal libraries. Further, a similar ANN-based approach can be adopted for screens involving other signal/motif receptors than APs. We anticipate that use of the ANN paradigm would be of great benefit for rapidly utilizing the information generated by all these efforts and for the analysis of data from other challenging endeavors in the area of vesicle trafficking.

Appendix

Neural Network Architecture and Data Flow

As described in Section 3.1, the identity of the residues within the XXXYXXØ signal determines which neuron within the X- and Ø-position clusters (at the input layer) is turned on (i.e., "on" output value = 1; "off" output value = 0). Then, each input neuron i sends a "message" to each hidden neuron j, equal to its off/on output value (O_i) times the connection weight (W_{ij}). Thus, the total net input (I) received by each hidden neuron is

$$I_j = \sum O_i W_{ij}. \tag{A.1}$$

In turn, neurons from the hidden layer as well as the unique node in the output layer (Figure 2) produce a response according to a sigmoidal activation function

$$O_j = \frac{1}{\left(1 + e^{-\alpha I_j}\right)}, \tag{A.2}$$

where O_j represents the output response from a given hidden or output neuron j receiving a net input I_j (modulated by an α factor) [21]. The final output (O) is then compared with the expected interaction value (V) (from the training data set) by using an error function (E) (Figure 2(a))

$$E = \sum k\,(V - O)^2, \tag{A.3}$$

where k is the number of examples in the training data set.

A weight correction to minimize the error function is estimated according to

$$\Delta W_{ij\,(n)} = -\eta\left(\frac{dE}{dW_{ij}}\right) + m\Delta W_{ij\,(n-1)}, \tag{A.4}$$

where $\Delta W_{ij\,(n)}$ and $\Delta W_{ij\,(n-1)}$ represent the change of the weights calculated at iterations n and $n-1$, respectively. η is the learning rate parameter and m is the momentum constant [31]. The weight corrections are implemented, on the initially random W_{ij}, in the opposite direction to data flow (back-propagation), and then another feed-forward run is started by using the newly updated W_{ij} values.

The learning rate η is continuously optimized according to a "line search" algorithm [32] for maximal convergence efficiency to an E minima (Figure 3). The iterations will continue until convergence is reached, leading the network to learn by backpropagation [33–35].

Acknowledgments

The authors are indebted to Dr. Juan S. Bonifacino (NIH) for sharing his lab data and for useful discussions. We also thank Dr. Darwin Reyes (National Institute of Standard and Technologies), Dr. Lymarie Maldonado-Baez (NIH), Dr. Henry Chang (Purdue University), and members of the Aguilar and Chang labs for stimulating discussions and critical reading of the paper. Special thanks to Saikat Banerjee (Indian Institute of Management, Bangalore, India) for writing the spreadsheet macro that runs the ANN algorithm (freely available upon request). This work was supported by start-up funds from the department of Biological Sciences, Purdue University and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under GRANT agreement CCF-0939370.

References

[1] C. Enns, "Overview of protein trafficking in the secretory and endocytic pathways," *Current Protocols in Cell Biology*, vol. 15, p. 15.1, 2001.

[2] J. B. Dacks, A. A. Peden, and M. C. Field, "Evolution of specificity in the eukaryotic endomembrane system," *International Journal of Biochemistry and Cell Biology*, vol. 41, no. 2, pp. 330–340, 2009.

[3] M. L. Wei, "Hermansky-Pudlak syndrome: a disease of protein trafficking and organelle function," *Pigment Cell Research*, vol. 19, no. 1, pp. 19–42, 2006.

[4] J. Gruenberg and F. G. van der Goot, "Mechanisms of pathogen entry through the endosomal compartments," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 7, pp. 495–504, 2006.

[5] P. Ghosh and S. Kornfeld, "The GGA proteins: key players in protein sorting at the trans-Golgi network," *European Journal of Cell Biology*, vol. 83, no. 6, pp. 257–262, 2004.

[6] F. Nakatsu and H. Ohno, "Adaptor protein complexes as the key regulators of protein sorting in the post-Golgi network," *Cell Structure and Function*, vol. 28, no. 5, pp. 419–429, 2003.

[7] L. M. Traub, "Common principles in clathrin-mediated sorting at the Golgi and the plasma membrane," *Biochimica et Biophysica Acta*, vol. 1744, no. 3, pp. 415–437, 2005.

[8] J. L. Urbanowski and R. C. Piper, "Ubiquitin sorts proteins into the intralumenal degradative compartment of the late-endosome/vacuole," *Traffic*, vol. 2, no. 9, pp. 622–630, 2001.

[9] J. S. Bonifacino and L. M. Traub, "Signals for sorting of transmembrane proteins to endosomes and lysosomes," *Annual Review of Biochemistry*, vol. 72, pp. 395–447, 2003.

[10] L. M. Traub, "Tickets to ride: selecting cargo for clathrin-regulated internalization," *Nature Reviews Molecular Cell Biology*, vol. 10, no. 9, pp. 583–596, 2009.

[11] M. Boehm and J. S. Bonifacino, "Genetic analyses of adaptin function from yeast to mammals," *Gene*, vol. 286, no. 2, pp. 175–186, 2002.

[12] H. Ohno, "Physiological roles of clathrin adaptor AP complexes: lessons from mutant animals," *Journal of Biochemistry*, vol. 139, no. 6, pp. 943–948, 2006.

[13] J. S. Bonifacino and E. C. Dell'Angelica, "Molecular bases for the recognition of tyrosine-based sorting signals," *Journal of Cell Biology*, vol. 145, no. 5, pp. 923–926, 1999.

[14] H. Ohno, R. C. Aguilar, D. Yeh, D. Taura, T. Saito, and J. S. Bonifacino, "The medium subunits of adaptor complexes recognize distinct but overlapping sets of tyrosine-based sorting signals," *Journal of Biological Chemistry*, vol. 273, no. 40, pp. 25915–25921, 1998.

[15] R. C. Aguilar, M. Boehm, I. Gorshkova et al., "Signal-binding specificity of the μ4 subunit of the adaptor protein complex AP-4," *Journal of Biological Chemistry*, vol. 276, no. 16, pp. 13145–13152, 2001.

[16] R. Nair and B. Rost, "Protein subcellular localization prediction using artificial intelligence technology," in *Functional Proteomics: Methods and Protocols*, pp. 435–463, 2008.

[17] G. Schneider and U. Fechner, "Advances in the prediction of protein targeting signals," *Proteomics*, vol. 4, no. 6, pp. 1571–1580, 2004.

[18] J. Hawkins and M. Bodén, "Detecting and sorting targeting peptides with neural networks and support vector machines," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 1, pp. 1–18, 2006.

[19] A. Krogh, "What are artificial neural networks?" *Nature Biotechnology*, vol. 26, no. 2, pp. 195–197, 2008.

[20] R. Rojas, *Neural Networks—A Systematic Introduction*, Springer, New York, NY, USA, 1996.

[21] J. Zou, Y. Han, and S.-S. So, "Overview of artificial neural networks," in *Methods in Molecular Biology*, D. J. Livingstone, Ed., pp. 15–23, Humana Press, Totowa, NJ, USA, 2008.

[22] M. M. Poulton, "Neural networks as an intelligence amplification tool: a review of applications," *Geophysics*, vol. 67, no. 3, pp. 979–993, 2002.

[23] P. V. Burgos, G. A. Mardones, A. L. Rojas et al., "Sorting of the Alzheimer's disease amyloid precursor protein mediated by the AP-4 complex," *Developmental Cell*, vol. 18, no. 3, pp. 425–436, 2010.

[24] R. C. Aguilar, H. Ohno, K. W. Roche, and J. S. Bonifacino, "Functional domain mapping of the clathrin-associated adaptor medium chains μ1 and μ2," *Journal of Biological Chemistry*, vol. 272, no. 43, pp. 27160–27166, 1997.

[25] E. C. Dell'Angelica, V. Shotelersuk, R. C. Aguilar, W. A. Gahl, and J. S. Bonifacino, "Altered trafficking of lysosomal proteins in Hermansky-Pudlak syndrome due to mutations in the β3A subunit of the AP-3 adaptor," *Molecular Cell*, vol. 3, no. 1, pp. 11–21, 1999.

[26] N. R. Gough, M. E. Zweifel, O. Martinez-Augustin, R. C. Aguilar, J. S. Bonifacino, and D. M. Fambrough, "Utilization of the indirect lysosome targeting pathway by lysosome-associated membrane proteins (LAMPs) is influenced largely by the C-terminal residue of their GYXXΦ targeting signals," *Journal of Cell Science*, vol. 112, no. 23, pp. 4257–4269, 1999.

[27] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[28] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[29] I. A. Taha and J. Ghosh, "Symbolic interpretation of artificial neural networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 3, pp. 448–463, 1999.

[30] K. Janvier, Y. Kato, M. Boehm et al., "Recognition of dileucine-based sorting signals from HIV-1 Nef and LIMP-II by the AP-1 γ-σ1 and AP-3 δ-σ3 hemicomplexes," *Journal of Cell Biology*, vol. 163, no. 6, pp. 1281–1290, 2003.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[32] S. L. Hu, Z. H. Huang, and N. Lu, "A non-monotone line search algorithm for unconstrained optimization," *Journal of Scientific Computing*, vol. 42, no. 1, pp. 38–53, 2010.

[33] M. Tusar, J. Zupan, and J. Gasteiger, "Neural networks and modeling in chemistry," *Journal de Chimie Physique et de Physico-Chimie Biologique*, vol. 89, no. 7-8, pp. 1517–1529, 1992.

[34] J. Gasteiger and J. Zupan, "Neural networks in chemistry," *Angewandte Chemie*, vol. 32, no. 4, pp. 503–527, 1993.

[35] V. Simon, J. Gasteiger, and J. Zupan, "A combined application of two different neural network types for the prediction of chemical reactivity," *Journal of the American Chemical Society*, vol. 115, no. 20, pp. 9148–9159, 1993.