Using random forest to find Nuclear Export Signals (NESs) in proteins of Arabidopsis thaliana

Claudia Rubiano¹, Thomas Merkle² and Tim W. Nattkemper³

¹Chemistry department, National University of Colombia, Bogotá, Colombia. Former member of the Graduate School of Bioinformatics and Genome Research, Bielefeld University, Germany.

²Institute for Genome Research & Systems Biology, Faculty of Biology, Bielefeld University, Germany.

³Biodata Mining group, Faculty of Technology, Bielefeld University, Germany.

ccrubianoc@unal.edu.co, crubiano@cebitec.uni-bielefeld.de, tmerkle@cebitec.uni-bielefeld.de, tim.nattkemper@uni-bielefeld.de

Keywords: Nuclear export signals - Arabidopsis thaliana - random forest.

Abstract: This paper presents a new computational strategy for predicting Nuclear Export Signals (NESs) in proteins of the model plant Arabidopsis thaliana based on a random forest classifier. NESs are amino acid sequences that enable a protein to interact with a nuclear receptor and in this way to be exported from the nucleus to the cytoplasm.

The proposed classifier uses two kinds of features, the sequence of the NESs expressed as the score obtained from a HMM profile and physicochemical properties of the amino acid residues expressed as amino acid index values. Around 5000 proteins from the total of proteins sequences from Arabidopsis were predicted as containing NESs. A small group of these proteins was experimentally tested for the actual presence of an NES. 11 out of 13 tested proteins showed positive interaction with the receptor Exportin 1 (XPO1a) from Arabidopsis in yeast two-hybrid assays, which indicates they contain NESs.

The experimental validation of the nuclear export activity in a selected group of proteins is an indicative of the potential usefulness of the tool. From the biological perspective, the nuclear export activity observed in those proteins strongly suggest that nucleo-cytoplasmic partitioning could be involved in regulation of their functions.

1 INTRODUCTION

Nucleo-cytoplasmic shuttling refers to the transport of proteins and other molecules into and out of the cell nucleus. It plays an important regulatory role in key cellular processes like transcription, RNA processing and cell cycle. The process is usually mediated by a family of transport receptors known as karyopherins (Ström and Weis, 2001) that directly or indirectly bind to their cargoes via signals like the nuclear localization signal (NLS) for nuclear import or the nuclear export signal (NES) for export to form a transport complex (Pemberton and Paschal, 2005). In the case of nuclear export, several pathways have been identified (Ossareh-Nazari et al., 2001). To date, the best studied pathway depends on the presence of a leucine-rich NES in the cargo.

Leucine-rich NESs have been experimentally verified in proteins from diverse organisms (mainly S. cerevisiae, H. sapiens and virus). Most of them were compiled in the database NESbase 1.0 (la Cour et al., 2003) and used to build a predictor for leucine-rich NES by a combination of Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) (la Cour et al., 2004).

The existence of a nuclear export pathway for proteins carrying a leucine-rich NES in plants has been demonstrated (Haasen et al., 1999; Merkle, 2001). This process plays an important role in the nucleo-cytoplasmic partitioning of proteins and hence in the regulation of many signalling processes in plants (Merkle, 2004; Merkle, 2011). However, since the available tool to predict NES (la Cour et al., 2004) very often fails to recognize them in plant proteins, it is very likely that there are additional features specific for plants that are not included in that tool. Since finding NES sequences in proteins by experimental methods is expensive and time consuming, an efficient computational prediction is

¹http://www.cbs.dtu.dk/services/NetNES/
of great interest.

It has been shown that identifying leucine-rich NESs is not a trivial task. This is because consensus patterns alone are not sufficient for prediction and, additionally, Leucine is one of the most abundant amino acids in proteins. Furthermore, NESs share sequence similarity to regions that form the hydrophobic core of many proteins (Cook et al., 2007). Since a machine learning method has the potential to detect very divergent signals that a consensus pattern is unable to detect, this approach is used here. Supervised machine learning methods have been widely used in bioinformatics prediction applications like: subcellular location of proteins (Hua and Sun, 2001; Schneider and Fechner, 2004; Bendtsen et al., 2004; Garg et al., 2005; Lei and Dai, 2005; Pazos and Jung Wook Bang, 2006; Brameier et al., 2007; Verma et al., 2008; Habib et al., 2008; Gromiha and Yabuki, 2008; Kumar and Raghava, 2009), protein function (Lee et al., 2009), protein secondary structure (Riis and Krogh, 1996), protein binding sites (Liu et al., 2009), protein-protein interaction (Bock and Gough, 2001), and special features in proteins like ubiquitylation (Tung and Ho, 2008) and glycosylation (Caragea et al., 2007).

Random forest is a classifier consisting of a collection of many decision trees where each tree is grown using a (bootstrap) subset of the training dataset. Bootstrapping is a resampling technique where a number of bootstrap training sets are drawn randomly from the original training set with replacement. Each tree induced from bootstrap samples grows to full length and the number of trees in the forest is adjustable. To classify an instance of unknown class label, each tree casts a unit classification vote. The forest selects the classification having the most votes over all the trees in the forest. Compared with the decision tree classifier, random forests have better classification accuracy, are more tolerant to noise and are less dependent on the training datasets (Breiman, 2001).

2 METHODS

2.1 Data Sets

A positive data set was conformed with 107 experimentally confirmed NES sequences, including those contained in the NES database already available (la Cour et al., 2003) together with sequences from Arabidopsis, which have been experimentally confirmed (T. Merkle, unpublished). The length of the sequences used as positive NESs was defined by taking as a reference the last hydrophobic amino acid within the NES relative to the C-terminal of each protein sequence, and counting 10 amino acids towards the N-terminal and 4 towards the C-terminal, which makes a total length of 15 amino acids. The amino acid taken as reference has been shown to be necessary and critical for the interaction of the NES with the Exportin receptor (Görlich and Kutay, 1999; Kaffman and O’Shea, 1999; Haasen et al., 1999; Ossareh-Nazari et al., 2001).

On the other hand, a negative data set was conformed with protein regions without nuclear export activity. It was done by excluding from the proteins used in the positive data set, those amino acid regions for which some evidence for nuclear export activity was available. Around 10000 sequences of 15 amino acids length were considered from which some subsets were randomly selected.

2.2 Feature Calculation

The elements of the data sets are amino acid sequences, therefore they could not be used directly in a machine learning task. It was necessary to find some properties that could be expressed numerically to generate feature vectors as a representation of each sequence. This study assessed two kinds of properties: amino acid sequence and physicochemical properties.

2.2.1 Amino Acid Sequence

The possibility of using amino acid residue order as one of the elements of the feature vector was explored by constructing a distance matrix to reveal the similarity among all the sequences. The pairwise alignment score obtained by comparing each sequence to each other with the program ALIGN was used as similarity measure. ALIGN computes the global alignment of two sequences using a modification of the algorithm described by (Myers and Miller, 1988).

The distance matrix obtained showed that the order of the amino acid residues could be used to distinguish positive (NES) from negative (nonNES) sequences. To express that in a numerical way, a profile HMM was built with HMMER ver 2.3.2\(^2\) using the NESs sequences from Arabidopsis intending to capture specific features from plant sequences.

The multiple sequence alignment used to construct the profile HMM was obtained with the NES sequences of Arabidopsis using CLUSTALW (Chenna et al., 2003) and QALIGN (Sammeth et al., 2003).

\(^2\)http://hmmer.janelia.org/
2.2.2 Physicochemical Properties

Since the physicochemical properties of the amino acid residues are the most important feature for biochemical reactions, the amino acid index values were used to extract additional features that are independent of the amino acid order in the sequence. An amino acid index (aaindex) is a set of 20 numerical values representing any of the different physicochemical and biochemical properties of each amino acid residue.

Many of the published index values are collected in the AAindex database (Kawashima and Kanehisa, 2000; Kawashima et al., 2008). There are 544 attributes in the AAindex1 database Version 9.1, therefore one can calculate such a number of features. The aaindex values for each sequence were calculated by the sum of the respective index values of the amino acid residues present in the sequence, as follows:

Each aaindex is represented as: \[ AA_j = (AA_{j1}, \ldots, AA_{j20}) \] where \( j \) corresponds to each aaindex value and varies from 1 to 544.

For each sequence \( s \) of length \( l \) amino acid residues \( a \) represented as: \( s = a_1, \ldots, a_l \), the value of the corresponding aaindex value \( x_{s,j} \) is obtained by adding the individual aaindex value of each amino acid: \( x_{s,j} = \sum_{k=1}^{l} AA_j(a_k) \).

Aaindex features, like described above, have been used in other proteomics contexts as well to encode molecular features for instance to predict mass spectrometry signals (Timm et al., 2008).

Finally, the profile HMM score \( (h_s) \) of a sequence \( s \) (see above) is appended to the aaindex values to conform the final feature vector for each sequence: \( x_{s,545} = h_{mm} \).

2.3 Training

Data pre-processing included a step of unsupervised filtering to remove high correlated features as well as centering and scaling. In addition to random forest, two other learning algorithms were trained, namely k-Nearest Neighbour (k-NN) and Support Vector Machine (SMV). The caret (classification and regression training) package (Kuhn, 2008b; Kuhn, 2008a) under the statistical platform R (Ihaka and Gentleman, 1996; R Development Core Team, 2005) was used for that purpose.

The training process was carried out using a combination of repetitive hold-out or splitting method (with \( p = 0.25 \)) in the complete data set and classical resampling methods (10-fold CV, LOOCV and .632+ bootstrap) applied only to the training set. After the training, the respective test set was used to evaluate the performance of the classifiers. The complete process (hold-out plus resampling) was performed multiple times along the complete data set.

2.4 Performance Evaluation Criteria

2.4.1 Classification Measures

Based on the class predicted by the trained classifiers for every element of the test set and its actual class, there are four possible outcomes. If the sample is positive (is an NES) and it is classified as NES, it is counted as true positive (TP); if it is classified as nonNES, it is counted as a false negative (FN). If the sample is negative and it is classified as nonNES, it is counted as a true negative (TN); if it is classified as NES, it is counted as a false positive (FP). Based on these possibilities, a classical two-by-two confusion matrix or contingency table was used as reference to calculate some performance metrics (Baldi et al., 2000) (Accuracy (ACC), True positive rate (TPR), False positive rate (FPR), Specificity and Precision), as well as two correlation measures (Matthews correlation coefficient (MCC) and F-score).

2.4.2 Receiver Operating Characteristic Curves (ROCs)

With the trained classifiers, it is possible to produce a continuous output (directly or by transformation of a discrete output). It means that the outcome of the classifier is an estimated confidence value. Thus, depending on the confidence threshold value applied, the results of the confusion matrix can change, which implies that some of the performance measurements described before are valid only at a particular probability threshold value.

To assess the performance of the trained classifiers in a broad range of probability threshold values, receiver operating characteristic (ROC) curves were used. A ROC is a two-dimensional graph where the proportion of correctly classified positive samples i.e., true positive rate (TPR) is plotted as a function of the proportion of incorrectly classified negative instances i.e., false positive rate (FPR). Each point on the ROC curve represents a classification threshold \( \theta \in [0,1] \) and corresponds to particular values of TPR and FPR. Varying the threshold gives a tradeoff between TPR and FPR. The construction of ROCs allows to calculate an additional measure called area under the ROC curve (AUC). This value has an important statistical property: the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly
chosen positive sample higher than a randomly chosen negative sample (Fawcett, 2004). The range of AUC values is $[0, 1]$: 1 represents the perfect classification and 0.5 a quite random one.

In this study the ROCs were constructed in R using the package rocr (Sing et al., 2005) and the AUC value was calculated using the function aucRoc from the caret package.

2.5 Pipeline Construction

To deploy the finished classifier for prediction of NES in new protein sequences, it was necessary to process the new sequences in the same way as the training and test sequences. It is convenient to have a mechanism that uses a standard format (for instance amino acid sequences in fasta format) as input. For this, the classifier was integrated into a pipeline, which was implemented using PERL and R.

For the prediction of NESs, each protein is initially split into overlapping fragments of 15 amino acid residues length. Then the full set of features is calculated (profile HMM scores and aaindex values) for each fragment. Next, the resulting feature matrix is passed to the actual classifier and after the classification process, the original sequence is reassembled with probability values for the two classes (NES and nonNES) assigned to each amino acid residue. The output of the pipeline is a list of the proteins containing NES(s) with the position where the possible signal is located in the sequence. This output can be modulated by changing the probability value used as threshold for the class assignment.

2.6 Prediction of NESs in New Protein Sequences and Experimental Verification

A data set containing 33410 protein sequences, obtained from the Arabidopsis Information Resource website (TAIR)\(^4\) was used as target for the prediction. Since one requirement for a protein to be exported from the nucleus is its interaction with the nuclear export receptor Exportin 1, a group of 24 proteins was selected out of the total predicted to be experimentally tested for the presence of an actual NES. Selection of proteins to be assessed was carried out using Gene Ontologies (GO) (The Gene Ontology Consortium, 2000) and some experimental constraints of the Yeast-two-Hybrid (YTH) plasmid vectors used (Clontech Matchmaker LexA system). The GO terms used were taken from the categories Biological Process and Molecular Function, focusing on those related with transcription and/or nucleic acid metabolism.

For YTH, the respective cDNA from the protein to be tested was amplified by PCR using specifically designed oligonucleotides. The amplified fragments were cloned in the vector pB42AD and confirmed by sequencing. The pB42AD plasmids containing the cDNAs investigated, together with pGilda plasmid containing the cDNA of Arabidopsis XPO1a (Haasen et al., 1999) were used in the final interaction assays.

3 RESULTS AND DISCUSSION

3.1 Preliminary Analysis

One of the most important points when developing a classification tool is to look for properties that allow the separation between the classes. Intuitively, the first property in this case could be the order and identity of the amino acid residues in each class (NES and nonNES sequences). Fig. 1 shows a comparison all against all of NES and nonNES sequences, the presence of a darker zone in comparison to the rest of the matrix is clearly visible. This area corresponds to the region where NES sequences are compared to other NES sequences. It means that an NES sequence is more similar to another than is also NES than to another that is nonNES. Therefore, the identity and order of the amino acid residues in the sequences could be used as one of the features to separate the two classes.

To extract a measurable feature from that property, a profile HMM was used. Since profile HMMs use position specific scores for the amino acid residues and position specific penalties for opening and extending an insertion or deletion, they capture important information about the degree of conservation and the varying degree to which gaps and insertions are permitted at various positions in a multiple alignment (Eddy, 1998).

3.2 Assessment and Selection of the Optimal Classifier

Performance Assessment:

Fig. 2 presents the results of the performance metrics evaluated for the trained classifiers. Regarding the sensitivity value, the $k$-NN classifier had a small advantage over the other two. Nevertheless, this classifier was the least specific and least precise, and

\(^4\)http://www.arabidopsis.org - release TAIR9
showed also in correspondence the highest values for false positive rate and classification error. RF was comparable to SVM regarding sensitivity, however it showed slightly higher values in accuracy, specificity and precision, as well as lower false positive rate and error than SVM. It is also noticeable that $k$-NN exhibits a higher degree of dispersion in specificity and false positive rate, compared to RF and SVM.

In addition to the performance measures shown in Fig. 2, two correlation measures, Matthews Correlation Coefficient (MCC) and F-score were also evaluated. Concerning MCC, values of 0.55, 0.77 and 0.66 were obtained for $k$-NN, RF and SVM, respectively. Since for MCC a value of “1” is regarded as perfect prediction and “0” indicates a completely random prediction, in this case the three classifiers predict much better than random being RF and SVM slightly superior to $k$-NN. Similar results were obtained regarding the F-score, values of 0.75, 0.86 and 0.80 were obtained for $k$-NN, RF and SVM, respectively.

Receiver Operating Characteristics (ROC) Curves:

The outcome of the classification process can be seen as class probability values for every classified sample. Therefore, the performance metrics can change depending on the cutoff value used. In order to assess the relation between sensitivity (expressed as true positive rate (TPR)) and true negative rate (TNR) across different cutoff values of class probabilities, receiver operating characteristics (ROC) curves were constructed.

The ROCs for the trained classifiers are shown in Fig. 3, where the indicator area under the curve (AUC) is also included. According to the ROCs the three classifiers can predict much better than random, which can be seen in the localization of the curve in the ROC space, in the shape of the curves and also in the AUC value which is $>0.5$ in all the cases. According to this parameter it seems that RF performs better than the other two classifiers. However, this conclusion can not be drawn using only the ROCs since the class distribution of the samples (proportion of positive (NES) compared to negative (nonNES) sequences) is not considered. Hence, for a direct comparison of the three classifiers in the ROC space, the ROC convex hull (ROCCH) method, described by Provost and Fawcett (Provost and Fawcett, 2001) was used. Two scenarios were considered: first, when the sample contains same proportion of positives and negative samples and second,
when the sample has 20% of positive examples and 80% of negative ones. According to this approach, RF would be the best classifier under the two circumstances considered (results not shown due to space constrains).

From the set of 33410 protein sequences used as target for the prediction, 5156 sequences corresponding to individual loci were predicted as NES-containing proteins. From this set of predictions, 24 proteins were selected as described before and finally 13 of them were cloned and experimentally tested for interaction with the receptor XPO1a of Arabidopsis.

The outcome from the YTH assays for these 13 proteins is shown in Fig 4. A positive result in this assay can be taken as an indicative that the tested protein has a functional NES since such a protein interacts with the nuclear export receptor XPO1a. That was the case for 11 out of the 13 tested proteins. The identity of those proteins is not given here since further functional analysis are been carried out with them and the results are not published yet.

### 3.3 Classification of New Samples and Experimental Verification

From the set of 33410 protein sequences used as target for the prediction, 5156 sequences corresponding to individual loci were predicted as NES-containing proteins. From this set of predictions, 24 proteins were selected as described before and finally 13 of them were cloned and experimentally tested for interaction with the receptor XPO1a of Arabidopsis.

The outcome from the YTH assays for these 13 proteins is shown in Fig 4. A positive result in this assay can be taken as an indicative that the tested protein has a functional NES since such a protein interacts with the nuclear export receptor XPO1a. That was the case for 11 out of the 13 tested proteins. The identity of those proteins is not given here since further functional analysis are been carried out with them and the results are not published yet.

---

**Figure 3:** Receiver operating characteristics (ROC) curves

Considering the results of the performance measurements and ROC curves, RF was selected as the best method to classify NESs. Therefore, it was used to predict NESs in new protein sequences. It is worth to mention that, as part of the algorithm, FR returns measures of variable importance. It was noticeable that although the highest relevance values corresponded to the score from the profile HMM as well as to `aaindex` values related with hydrophobicity measures, any of those features alone was enough to make an accurate classification of the NES and nonNES sequences.

One of the intended uses of this classifier was to predict NES-containing proteins in the whole available sequences of Arabidopsis. For an application like this, it is very important to have a high specificity even if the sensitivity decreases i.e., it is desirable to minimize the number of false positives even if that means that some true positives are missed. One way to achieve that is by adjusting the probability `cutoff` value that the classifier uses to assign the class label to new samples.

It was seen that probability `cutoff` values higher than 0.5 can give a better specificity at the cost of some decrease in accuracy and sensitivity. Consequently, for the screening of the whole available protein sequences of Arabidopsis using the RF classifier, a `cutoff` value of 0.7 was selected as a trade-off between gaining in specificity without loosing too much in sensitivity.

**Figure 4:** Receptor binding activity for selected proteins. Yeast two-hybrid assays for 13 proteins selected from the total of predicted as containing NESs. The respective cDNA fragments amplified by PCR were ligated into the vector pB42AD (Vector 1) and tested in the yeast strain EGY48[p8op-LacZ] for interaction with Arabidopsis nuclear export receptor XPO1a. Blue/green color indicates a positive interaction between the protein tested and the receptor.
4 CONCLUSIONS

The foremost contribution of this work was the development of an accurate tool for predicting NESs in proteins of Arabidopsis based on a Random Forest classifier. This conclusion is based on two facts. First, the high values obtained in the performance metrics, correlation measures and ROCs used as evaluation criteria. Second, the experimental verification of the nuclear export activity in a selected group from the total of predicted proteins that confirmed that the developed tool is accurate for the intended use: the detection of NESs in proteins of Arabidopsis.

From the computational point of view, two major challenges were addressed: finding the appropriate features to represent the NESs and dealing with a low number of available samples. The first problem was managed with the combination of a profile HMM and physicochemical properties expressed as amino acid index values. On the other hand, to deal with the limited availability of samples, a mixed resampling approach was used for the training and testing. This approach has turned out to be effective.

An important characteristic of the developed tool is that the random forest classifier was integrated into a pipeline where it is possible to adapt the probability threshold value according to the application which allows to modify the trade-off between specificity and sensitivity. In other words: for an application like the screening of a big set of protein sequences, could be advisable to use an astringent threshold value i.e., the specificity is more important. However, if the aim is to look for the possible position of an NES in a protein with known or suspected nuclear export activity, it would be better to low the threshold value to gain more sensitivity.

From a biological perspective, the prediction of around 5000 proteins that possibly contain NESs implies that approximately 18% of the total of proteins of Arabidopsis could have an NES, which is an indicative of high potential of the nucleo-cytoplasmic partitioning as a regulation mechanism in Arabidopsis.

The results of this work raise new challenges for further investigation. The nuclear export activity detected in the proteins tested calls to be determined and characterized in planta. Additionally, the experimental localization of the NESs is necessary to determine if they are in accordance with the predicted positions. On the other hand, in the total set of proteins predicted as NES-containing there are still many waiting to be tested. As soon as more proteins are experimentally verified, the classifier could be re-trained using the new data to improve the performance even more.

The developed prediction tool was directed to Arabidopsis proteins, however the extension to other plants or related organisms is thinkable. To facilitate that, it would be desirable to extend the usability of the tool. Since currently the prediction tool is available for individual use only, one of the perspectives for the near future is to make it available as a web application with both, a graphical interface and an application server interface.

ACKNOWLEDGEMENTS

Thanks to the Bioinformatics Resource Facility (BRF) of the Bielefeld University for the technical cooperation.

REFERENCES


Fawcett, T. (2004). ROC graphs : Notes and practical considerations for researchers. Technical re-


