Nuclear Export Signals (NESs) in *Arabidopsis thaliana*

—

Development and experimental validation of a prediction tool

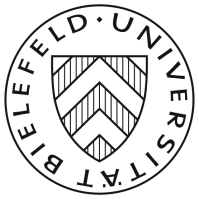
Dissertation

to obtain the academic title
Doctor of Natural Sciences (Dr. rer. nat.)
at the Faculty of Biology
in the Bielefeld University

presented by

Claudia Consuelo Rubiano Castellanos
from Moniquirá (Colombia)

in April 2010



Claudia Consuelo Rubiano Castellanos
Graf-von-Stauffenberg-Straße 10c
33615 Bielefeld
crubiano@cebitec.uni-bielefeld.de

Supervisors: Prof. Dr. Thomas Merkle
Prof. Dr. Tim Nattkemper

Printed on acid-free paper.
printed on ageing resistant paper - ISO 9706.

Summary

It is well-established that nucleo-cytoplasmic shuttling regulates not only the localization but also the activity of many proteins like transcription factors, cell cycle regulators and tumor suppressor proteins just to mention some. Also in plants the nucleo cytoplasmic partitioning of proteins emerges as an important regulation mechanism for many plant-specific processes. One requirement for a protein to shuttle between nucleus and cytoplasm lies in its nuclear export activity. The widely used mechanism for export of proteins from the nucleus involves the receptor Exportin 1 and the presence of a nuclear export signal (NES) in the cargo protein. Given the big amount of sequence data available nowadays the possibility to use a computational tool to predict the proteins potentially containing an NES would help to facilitate the screening and experimental characterization of NES-containing proteins. However, the computational prediction of NESs is a challenging task. Currently there is only one NES prediction tool and that is unfortunately not accurate for predicting these signals in proteins of plants. In that direction, this study aimed mainly at developing a prediction method for identifying NESs in proteins from Arabidopsis and to validate its usefulness experimentally. It included also the definition of the influence of the NES protein context in the nuclear export activity of specific proteins of Arabidopsis.

Three machine-learning algorithms (i.e. k-NN, SVM and Random Forests) were trained with experimentally validated NES sequences from proteins of Arabidopsis and other organisms. Two kinds of features were included, the sequence of the NESs expressed as the score obtained from an HMM profile constructed with the NES sequences of proteins from Arabidopsis, and physicochemical properties of the amino acid residues expressed as amino acid index values. The Random Forest classifier was selected among the three classifiers after evaluation of the performance by different methods. It showed to be highly accurate (accuracy values over 85%, classification error around 10%, MCC around 0.7 and area under the ROC curve around 0.90) and performed better than the other two trained classifiers.

Using the Random Forest classifier around 5000 proteins from the total of protein sequences from Arabidopsis were predicted as containing NESs. A group of these proteins was selected by using Gene Ontologies (GO) and from this last group, 13 proteins were experimentally tested for nuclear export activity. 11 out of those 13 proteins showed positive interaction with the receptor Exportin 1 (XPO1a) from Arabidopsis in yeast two-hybrid assays. The proteins showing nuclear export activity include 9 transcription factors and 2 DNA metabolism-related proteins. Furthermore, it was established that the amino acid residues located between the hydrophobic residues in the NES as well as the protein structure of the regions around the NES can modify the nuclear export activity of some proteins.

In conclusion, this work presents a new prediction tool for NESs in proteins of Arabidopsis based on a Random Forest classifier. The experimental validation of the nuclear export activity in a selected group of proteins is an indicative of the usefulness of the tool. From the biological point of view, the nuclear export activity observed in those proteins strongly suggest that nucleo-cytoplasmic partitioning could be involved in the regulation of their functions. For the follow up research the further characterization of the proteins showing positive nuclear export activity as well as the validation of additional predicted NES-containing proteins is envisioned. In the near future, the developed tool is going to be available as a web application to facilitate and promote its further usage.

Keywords: Nuclear export signals (NESs),
Machine learning applications,
Random Forest,
Arabidopsis thaliana.

Contents

List of Figures

List of Tables

CHAPTER 1

Motivation and overview

Transport between the cell nucleus and cytoplasm has captured the attention of researchers ever since the discovery of the cell nucleus by Robert Brown in the early 1830s. In eukaryotes, transcription and translation are spatially separated by the double membranes of the nuclear envelope. This fact alone necessitates a large amount of nuclear transport: the export of protein-encoding RNAs and the import of transcription factors, to name a few. Transport into and out of the nucleus generates a differential distribution of macromolecules that contributes to the regulation of numerous cellular functions, including gene expression, protein translation, cell division and nuclear dynamics.

The general mechanisms of nuclear transport are quite conserved among eukaryotes. Nevertheless, it looks like many of the cargo proteins differ and are more organism-specific: the transport system is the same but the passengers are different. But, how does the system recognize the passengers? Most proteins travelling into or out of the nucleus carry special signals that are recognized by receptor proteins. These are generically called nuclear localization signals (NLSs) and nuclear export signals (NESs).

An increasing number of studies point out that regulation of nuclear transport can be exerted at the level of the cargo. Hence, the identification of the proteins containing nuclear import and export signals is an important step toward the understanding of the interactions between all the components of the nucleocytoplasmatic transport process.

The need to analyse the massive accumulation of biological data generated by high-throughput genome projects has stimulated the development of new and rapid computational methods. Computational approaches for predicting and classifying protein functions are essential in determining the functions of unknown proteins in a faster and more cost-effective manner, because experimentally determining protein function is both costly and time-consuming. In some cases, it is essential that the prediction methods are appropriately calibrated for their respective target species, as the signals could differ between the organisms. In the case of nuclear import and export signals, NLSs are more recognizable than NESs and maybe because of that, nowadays there are some computational approaches to identify NLSs in protein sequences but only one for NESs.

This study was motivated by the need to identify proteins carrying NESs in *Arabidopsis thaliana* and the fact that the only computational tool available at present does not recognize them, reflecting the requirement of a species-specific prediction tool. Thus, the main objective was to develop a computational prediction method for NESs in proteins from Arabidopsis. The project scope also included the experimental verification of the nuclear export activity for some of the predicted proteins as well as the experimental assessment of the nuclear export activity of some NES-containing proteins. This last point was aimed at revealing special features of true positives and false positives, inside and outside the NES that might influence the nuclear export activity of a protein.

The following chapters describe the methodology used to reach the proposed goals, the results obtained as well as the conclusions extracted from them. **Chapter 2** gives an introduction into the concepts and terminology related to the biological and computational aspects used in this dissertation. The process of nucleocytoplasmic transport is explained in detail as well as the NESs. From the computational perspective, a brief introduction into the machine learning classification approach is given together with the foundations of the supervised classification algorithms used in this work.
Chapter 3 is dedicated to the explanation of the methodology used. It starts with the general description of the computational approach, the data sets used for the development of the predictor and the vector representation of the amino acid sequences used. Next, the training, evaluation and selection process of the classifier are presented, followed by the the prediction of NES containing proteins in the whole *A. thaliana* protein sequences. The second part of **Chapter 3** describes the most relevant experimental procedures used.

Chapter 4 presents the body of results obtained in this work. It is divided in two parts following the same order used in **Chapter 3**. The outcome from the computational part concerning the development of the LR-NES predictor are exposed in the first part and the experimental result in the second one.
Chapter 5 analyses and discusses the results as a whole and **Chapter 6** outlines the main contributions of this dissertation and presents and discusses possible future directions of new aspects to be explored in follow up research.

CHAPTER 2

Background

Contents

Overview

The objective of this chapter is to introduce the topic to the reader presenting the main theoretical background and fundamental concepts that will be used throughout this dissertation. It is divided into two main parts. Section 2.1 introduces the major biological principles of nuclear transport of macromolecules across the nuclear envelope and the proteins involved, it focuses then on the nuclear export process and associated signals (nuclear export signals, NESs). It includes aspects

of the regulation of the mechanisms of nuclear transport involving NESs as well as the importance of nucleo-cytoplasmic partitioning as a regulatory tool for signaling. Next, the facts already known about the nuclear export process in plants are presented, followed by the description of the currently available approaches to identify nuclear import and export signals in protein sequences. From this, Section 2.2 leads the topic to the main concepts of machine learning and focuses on the supervised classification algorithms that are used through this work. It also includes the basic concepts of Hidden Markov Models (HMM) and profiles. Details of the specific methods will be seen in Chapter 3.

2.1 Nucleo-cytoplasmic transport

In eukaryotic cells there is a physical separation of the nuclear genomic material from the other intracellular compartments, which implies also that fundamental processes like DNA replication and RNA biogenesis occurring in the nucleus are separate from protein synthesis taking place in the cytoplasm. To accomplish this, molecules such as RNAs, ribosomal subunits, transcription factors and many different proteins need to travel continuously between these two compartments. The current list of proteins shuttling between the cytoplasm and the nucleus includes transport receptors and adaptors, receptors of steroid hormones, transcription factors, cell cycle regulators, and a large number of RNA-binding proteins (Görlich and Kutay, 1999; Haché *et al.*, 1999; Pines, 1999; Yang and Kornbluth, 1999; Nakielny and Dreyfuss, 1999; Shyu and Wilkinson, 2000). The nucleo-cytoplasmic distribution of such proteins calls for continuous regulation and coordination that it needed for normal cellular functions. Often, it is found that the reasons of cancer transformation of cells include distortions in the distribution of proteins between the nucleus and cytoplasm (Poon and Jans, 2005). Also, the aberrant cytoplasmic localization of transcription factors and other proteins due to malfunction in the regulation of nucleo-cytoplasmic transport, has been implicated in Alzheimer, Parkinson, and Lewy body diseases as well as amyotrophic lateral sclerosis, and human immunodeficiency virus encephalitis (Chu *et al.*, 2007). This explains the attention paid lately to the studies of the nucleo-cytoplasmic transport, its mechanisms and analyses of the regulation of protein partitioning between the nucleus and the cytoplasm (Sorokin *et al.*, 2007).

The nucleo-cytoplasmic traffic has become functionally and mechanistically diversified, serving not only to permit operation of the basal replication, transcription,

and processing machinery but also to regulate the cell cycle, transcriptional activation and repression, circadian rhythms, and many other processes (Macara, 2001). Additionally, since the nuclear transport system is involved in various stages of cell differentiation, defects in nuclear transport may cause severe developmental disorders (Yasuhara *et al.*, 2009).

The transport between the nucleus and the cytoplasm occurs only through the nuclear pore complexes (NPCs) (Feldherr, 1962; Feldherr *et al.*, 1984; Richardson *et al.*, 1988; Dworetzky and Feldherr, 1988; Corbett and Silver, 1997; Görlich and Kutay, 1999; Ryan and Wente, 2000), which are embedded in the nuclear envelope (NE) that separates the nucleus from the cytoplasm. The NPC is a large proteinaceous structure, which contains approximately 30 structural proteins called nucleoporins (Nups) in yeast (Rout *et al.*, 2000) and in vertebrates (Cronshaw *et al.*, 2002), and has a molecular mass ranging from 4466 MDa in yeast (Rout and Blobel, 1993; Yang *et al.*, 1998) to 60125 MDa in vertebrates (Reichelt *et al.*, 1990; Cronshaw *et al.*, 2002). Studies of the NPCs in various organisms suggest that these structures are well conserved among eukaryotes (Vasu and Forbes, 2001; Cronshaw *et al.*, 2002; Meier, 2005).

Nucleo-cytoplasmic transport comprises a multitude of substrates. Not only must all nuclear proteins, such as histones and transcription factors, be imported from the cytoplasm, but also transfer RNA (tRNA), ribosomal RNA (rRNA), and messenger RNA (mRNA) that are synthesized by transcription in the nucleus need to be exported to the cytoplasm where they function in translation. The biogenesis of, for instance, ribosomes even involves multiple crossings of the NE: ribosomal proteins are first imported into the nucleus, assembled in the nucleolus with rRNAs, and finally are exported as ribosomal subunits to the cytoplasm (Görlich and Kutay, 1999). In addition, some molecules accomplish a cyclic movement between the nucleus and the cytoplasm, a process known as nucleo-cytoplasmic shuttling (Izaurralde and Adam, 1998). The transport used for these substrates to cross the NPC could be grouped into two general mechanisms, *passive diffusion* and *receptor-mediated transport* also called *facilitated translocation*.

Passive diffusion: This kind of transport can be used by molecules having a size inferior to the effective diameter of the NPC (9 nm) (Moore and Horowitz, 1975; Görlich and Kutay, 1999), travelling in favor of the concentration gradient. On a physiological scale, passive diffusion is exceedingly slow in the case of bovine serum albumin (∼7 nm diameter, 68 kDa), negligible still for ovalbumin (∼6 nm diameter, 46 kDa), and reasonably fast only for small proteins (size less than 20-

30 kDa) (Görlich and Kutay, 1999). Yet even proteins or RNAs that are smaller than 20-30 kDa, such as histones (Breeuwer and Goldfarb, 1990; Jäkel *et al.*, 1999) and tRNAs (Zasloff, 1983; Arts *et al.*, 1998a,b; Kutay *et al.*, 1998) normally cross the NPC in an active and carrier-mediated fashion.

Receptor-Mediated Transport: This kind of active transport is used for the majority of the molecules travelling through the NPC in both directions, to the nucleus or to the cytoplasm. Active transport is a selective process triggered by specific transport receptors and signals, it can proceed against concentration gradients using energy. Active nucleo-cytoplasmic transport is mostly, but not in all the cases, mediated by a family of homologous transport receptors belonging to the importin β-like family, also called karyopherin β receptors or importins/exportins if they participate in **nuclear import** (transport from cytoplasm to nucleus) or **nuclear export** (transport from nucleus to cytoplasm), respectively (Görlich, 1997; Görlich and Kutay, 1999; Ström and Weis, 2001; Pemberton and Paschal, 2005).

2.1.1 Nuclear protein transport mediated by importin β-like receptors

Members of the importin β-like family are diverse, and more than 22 members have been identified in humans until now, including importins and exportins for proteins and RNAs (Harel and Forbes, 2004). 14 importin β-like proteins are encoded in the yeast genome (Ström and Weis, 2001; Weis, 2003) and *Arabidopsis thaliana* contains 17 genes for proteins of this family, although this last group has not been yet fully characterized (Bollman *et al.*, 2003; Merkle, 2003). The function of these proteins is restricted to either import or export, with the exception of two family members (one in yeast and one in human) that function in both import and export (Fried and Kutay, 2003; Weis, 2003; Mosapparast and Pemberton, 2004). There are many more cargoes than receptors, suggesting that each receptor has multiple cargoes.

Inherent to the function of importin β-like nuclear transport receptors in nuclear import and export is the ability of these proteins to bind their cargoes directly or indirectly, interact with NPC proteins (Nups) and interact with the the small GTPase Ras-related nuclear protein (Ran) in its GTP-bound form (RanGTP)

(Fried and Kutay, 2003; Weis, 2003; Mosapparast and Pemberton, 2004).
The importin β-like receptors recognize signals in the cargo proteins that are called nuclear localization signals (NLSs) in the case of nuclear import and nuclear export signals (NESs) for nuclear export.
The energy for nucleocytoplasmic transport is largely provided by a steep concentration gradient of the GTPase Ran, which also ensures the directionality of nuclear transport (Izaurralde *et al.*, 1997; Ossareh-Nazari *et al.*, 2001). Ran is highly enriched in the nucleus in its GTP-bound form, and GTP hydrolysis by Ran is directly coupled to the import/export cycle (Weis, 2003). Nuclear import complexes, upon arriving on the nucleoplasmic side of the NPC, are induced to disassemble when RanGTP binds to the receptor. In contrast, RanGTP is used to assemble export complexes which are in turn destabilized by hydrolysis of GTP on Ran in the cytoplasm. Figure 2.1 illustrates the principles of nuclear import and export mediated by importins/exportins, which will be explained below with special emphasis in the nuclear export process.

Nuclear Import

Nuclear protein import mediated by the importin α/β heterodimer (Figure 2.1A) was the first pathway to be understood in detail (Imamoto *et al.*, 1995; Görlich *et al.*, 1995a,b; Yasuhara *et al.*, 2009). In human, the importin α family has at least 6 family members encoded by distinct genes, all of which interact with importin β. On the other hand, importin β can act as a transport factor without importin α due to the direct recognition of specific cargo molecules (Lam *et al.*, 1999; Nagoshi *et al.*, 1999).
In addition to their function in nuclear import, both importins are also involved in other cellular processes. Importin α plays a role in functions such as mitotic spindle assembly (Gruss *et al.*, 2001; Schatz *et al.*, 2003; Ems-McClung *et al.*, 2004) and nuclear membrane formation (Askjaer *et al.*, 2002; Geles *et al.*, 2002). Importin β can also function as a global regulator of cellular functions distinct from nuclear transport, including mitotic spindle assembly, centrosome dynamics, nuclear membrane formation, and NPC assembly (Harel and Forbes, 2004).

Importins bind to cargo proteins carrying NLSs. The first kind of characterized NLS was a single cluster of basic amino acid residues (PKKKRKV) from the simian virus SV40 large tumor (T-)antigen (Kalderon *et al.*, 1984). The second type was the bipartite NLS of nucleoplasmin, which has two clusters of basic

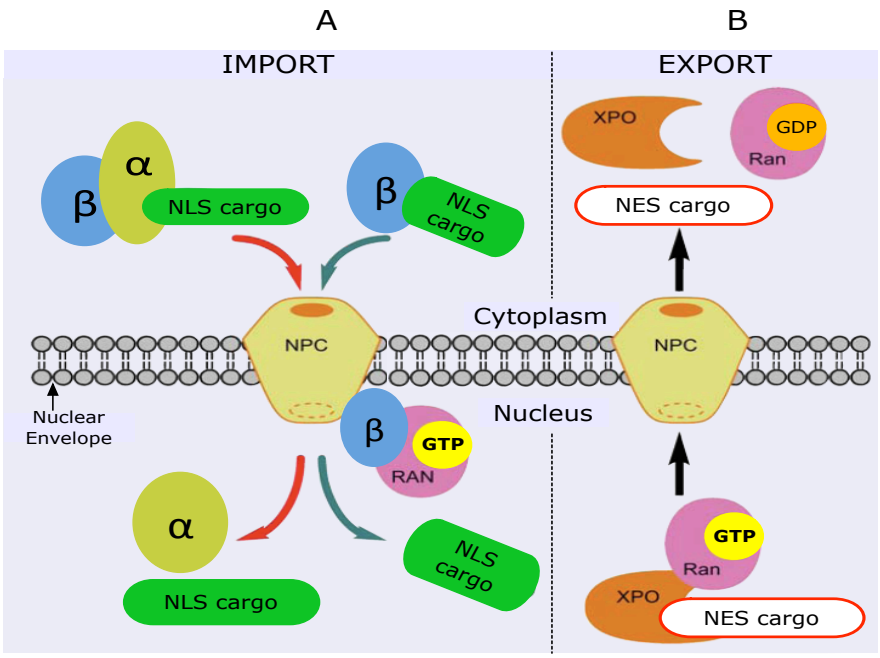


Figure 2.1: The processes of nuclear import and export

Basic model of importin/Exportin-mediated nucleocytoplasmic transport of proteins. **A**: Nuclear Import: Importin α binds to the NLS within a protein cargo in the cytoplasm and forms a ternary complex with importin β to enter into the nucleus. Some cargo molecules carrying NLS can also bind directly to importin β. In the nucleus, binding of RanGTP (the GTP-bound form of Ran) to importin β triggers the dissociation of the complex. **B**: Nuclear export: RanGTP stimulates binding of Exportin (XPO) to an NES-containing protein in the nucleus and the complex is exported to the cytoplasm, where hydrolysis of GTP on Ran results in complex disassembly. *NPC: nuclear pore complex, NLS: nuclear localization signal, NES: nuclear export signal*

amino acid residues with a spacer of conventionally 10-12 residues between them (KRPAATKKAGQAKKKK) (Robbins *et al.*, 1991). These two NLSs are considered as "classical" NLSs not only because they were described first but also because they are the two ones that are known best (Lange *et al.*, 2007). However, more types of NLSs do exist and some of them are even recognized by a different receptor (Fries *et al.*, 2007; Wagstaff and Jans, 2009).

Basic NLSs are generally recognized by importin α, which can enter the nucleus with or without cargo (Miyamoto *et al.*, 2002; Kotera *et al.*, 2005) but usually requires importin β, to deliver the cargo into the nucleus (Goldfarb *et al.*, 2004). The trimeric complex comprised of the cargo, importin α, and importin β, translocates through the NPC docked to the cytoplasmic face of the NPC and targeted to its core through the affinity of importin β for the NPC proteins (Nups) (Chi

et al., 1995; Enenkel *et al.*, 1995; Görlich *et al.*, 1995a; Radu *et al.*, 1995; Kraemer *et al.*, 1995; Nigg, 1997; Görlich, 1997). Once in the nucleus, dissociation of the complex containing both importins and the cargo protein is triggered by binding of Ran to importin β (Rexach and Blobel, 1995; Gilchrist *et al.*, 2002) and the importin α is released from the cargo with the help of the Nu Npap60 (Matsuura and Stewart, 2005).

Nuclear Export

Similar to nuclear import, substrates exported from the nucleus use targeting sequences called nuclear export signal (NESs), and specific receptors (Exportins) that recognize them. Of all characterized exportins, the export receptor Exportin 1 (CRM1 (for chromosome region maintenance) in vertebrates; Crm1p/Xpo1p in yeast) has the broadest known substrate range. Crm1p was originally identified in the fission yeast *Schizosaccharomyces pombe* in a genetic screen unrelated to nuclear transport (Adachi and Yanagida, 1989). The protein is encoded by an essential gene originally identified as a protein required for chromosome region maintenance (Adachi and Yanagida, 1989; Nishi *et al.*, 1994) and it was shown to be the target of the cytotoxic drug leptomycin B (LMB) (Nishi *et al.*, 1994; Wolff *et al.*, 1997) which makes a covalent modification at a cysteine residue in the central region of the protein (Kudo *et al.*, 1998, 1999a). Human CRM1 was identified based on its ability to bind nucleoporins and in particular CAN/Nup214 (Fornerod *et al.*, 1997b). The low but significant sequence relatedness between the N-terminal domains of CRM1 and the the emerging members of the karyopherin/importin β family of import receptors suggested that CRM1 was likely to be a transport receptor. This property together with the LMB sensitivity of CRM1 in *S. pombe*, were consistent with the possibility that CRM1 could be the NES receptor. Ultimately, studies from different laboratories demonstrated that CRM1 directly interacts with NES-containing substrates in a RanGTP dependent and LMB-sensitive manner (Fornerod *et al.*, 1997a; Fukuda *et al.*, 1997; Stade *et al.*, 1997).

Whereas Exportin 1 has a wide range of export substrates (proteins containing a hydrophobic NES, explained below), some other receptors from the importin β-like family are implicated in the nuclear export of specific cargoes. For example, CAS (cellular apoptosis susceptibility, Cse1p in yeast) exports the receptor importin α once it has carried in its protein cargo to the nucleus (Kutay *et al.*, 1997; Herold *et al.*, 1998; Hood and Silver, 1998; Solsbacher *et al.*, 1998). Exportin 4 functions

as an export receptor for eukaryotic translation initiation factor eIF5A (Lipowsky *et al.*, 2000). Exportin 5 transports microRNA precursors to the cytoplasm by recognizing the RNA hairpin structure with a 3' overhang as the NES (Yi *et al.*, 2003; Kim, 2004; Zeng and Cullen, 2004) whereas Exportin 6/RanBP20 exports profiling-actin complexes (Stüven *et al.*, 2003). RanBP16, designated Exportin 7, confines p50RhoGAP and 14-3-3 to the cytoplasm, in addition to export other substrates from the nucleus by using a NES different to the one present in the Exportin 1 cargoes (Mingot *et al.*, 2004). Exportin-t exports tRNA by recognizing a part of its structure as its NES (Lei and Silver, 2002; Rodriguez *et al.*, 2004). Concerning RNAs export, the case of mRNA is atypical in that it occurs by a mechanism that is distinct from that of proteins, tRNA or microRNA. Bulk mRNA is not exported by a member of the importin β-like family and does not rely on the RanGTP gradient. Instead, the transport involves a heterodimer formed of Nxf1 (metazoan; also known as TAP; Mex67 in yeast) and Nxt1 (metazoan; also known as p15; Mtr2 in yeast) (Rodriguez *et al.*, 2004; Cole and Scarcelli, 2006; Carmody and Wente, 2009).

Exportin 1 is the major receptor for the export of proteins out of the nucleus including cell cycle regulators, transcription factors, RNA binding proteins and many others. This Exportin also contributes to nuclear export of different classes of cellular RNAs or ribonucleoproteins (RNPs), for instance U snRNAs, ribosomal RNAs, signal recognition particle (SRP) and certain mRNP complexes (Fridell *et al.*, 1996; Ciufo and Brown, 2000; Popa *et al.*, 2002). More recently, Exportin 1 has also been associated with the export of mature microRNAs (Castanotto *et al.*, 2009). For the export of some cargoes, like the pre-60S ribosomal subunit, an NES-containing adapter protein that bridges its interaction with the export receptor is needed. In the case of the pre-60S ribosomal subunit this adapter protein is designated NMD3 (for non-sense-mediated mRNA decay), which contains an NES that recruits Exportin 1 (Zemp and Kutay, 2007). Another example of adapter protein is the NES-containing protein PHAX (phosphorylated adaptor for RNA export) (Ohno *et al.*, 2000), which serves as a bridge between the cap binding complex (CBC)-bound U snRNA and Exportin 1/RanGTP.

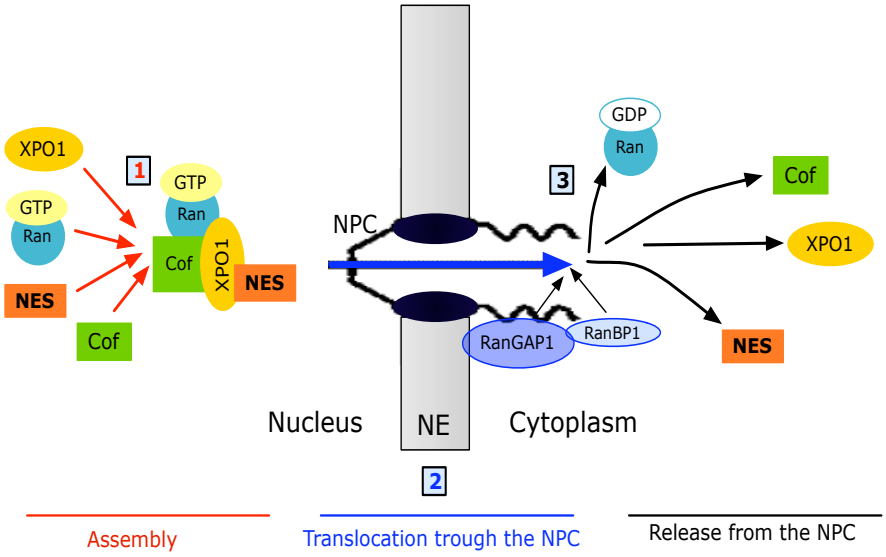


Figure 2.2: Nuclear export of proteins mediated by Exportin 1 receptor

The receptor Exportin 1 binds directly to a cargo protein containing a nuclear export signal (NES), co-operatively with RanGTP that is abundant in the nucleus (step 1). For some cargoes, an additional cofactor (Cof) is necessary to stabilize the complex. The export complex docks to the nuclear pore complex (NPC), and is translocated to the cytoplasmic side via direct interactions of Exportin 1 with FG-nucleoporins (step 2). On the cytoplasmic side, two regulators of the Ran GTPase cycle, Ran-binding protein 1 (RanBP1) and Ran GTPase-activating protein (RanGAP) that act together to hydrolyze GTP on Ran. As a result, the export complex disassembles and the NES-containing cargo is released into the cytoplasm (step 3). NE, nuclear envelope.

The nuclear export process mediated by the Exportin 1 (Figures 2.1B and 2.2), involves three main steps:

- **Assembly of nuclear export complexes**

  Assembly of cargo-Exportin 1 complexes requires the GTPase Ran. Exportins are characterized by a high affinity to their cargo substrates in the presence of RanGTP, which becomes a part of the export complex that forms in the nucleus. The formation of the complex between Exportin 1 and RanGTP occurs only in the presence of the cargo, and interaction between Exportin 1 and cargo depends on RanGTP. Thus, Exportin 1, NES-containing cargo, and RanGTP undergo cooperative binding, resulting in the formation of a trimeric export complex.
  Since the affinity of Exportin 1 for cargoes is generally low (explained below) (Askjaer *et al.*, 1999; Paraskeva *et al.*, 1999; Maurer *et al.*, 2001), additional cofactors or accessory proteins, appears to be required to stabilize the in-

teraction between Exportin 1 and some NES-containing substrates (Lindsay *et al.*, 2001; Englmeier *et al.*, 2001; Dasso, 2002). One example of this kind of proteins is the RanGTP-binding protein RanBP3 (Yrb2p in Yeast). The binding of RanBP3 results in formation of a quaternary complex consisting of RanBP3, RanGTP, Exportin 1 and the export NES-containing substrate (Figure 2.2). While RanBP3 does not bind to export substrates like an adaptor protein, it binds to Exportin 1 and increases its affinity for both RanGTP and NES-containing cargo (Englmeier *et al.*, 2001; Lindsay *et al.*, 2001). It has been suggested that RanBP3 has a function beyond the export complex formation in the nucleus, possibly accompanying Exportin 1 to the cytoplasm (Fried and Kutay, 2003).

- **Translocation of the export complex through the NPC**

  This step involves interactions between Exportin 1 and a subset of NPC proteins (FG nucleoporins) characterized by stretches of degenerative FG repeats enriched in phenylalanine and glycine residues and separated by hydrophilic linkers. These FG repeats are likely to be critical for the process by which transport receptor-cargo complexes gain selective access to the NPC channel. All known nuclear transport receptors can bind to FG-containing Nups, and interactions between transport receptors and FG repeats are essential for translocation through the NPC (Weis, 2003). However, the biophysical details of how these FG filaments contribute to the selective permeability of the NPC have been a matter of debate. Recently, Frey and Görlich (2007) proposed the formation of a "saturated hydrogel" within the NPC, in which all the FG domains engage in a maximum number of interactions to form a highly ordered mesh with very even pore size. In this model, the transport receptors are thought to dissolve the FG mesh and thus catalyze the entry and translocation of cargo through the NPC channel. Although this model answers some open questions concerning the NPC passage step of the nuclear transport, the issue remains whether such a hydrogel exists within the NPC and whether the model reflects the *in vivo* physiology of nuclear transport (Meier, 2007).

- **Release of the export complex from the NPC**

  After translocation, the export complex reaches the cytoplasmic side of the NPC and encounters two cytoplasmic regulators of the Ran GTPase cycle, Ran-binding protein 1 (RanBP1) and Ran GTPase-activating protein

(RanGAP) that act together to hydrolyze GTP on Ran (Figure 2.2). In humans, the large nucleoporin RanBP2 (NUP358) performs a similar function to RanBP1 and is also localized at the cytoplasmic face of the NPC. As a consequence of these topological arrangements, GTP hydrolysis on Ran occurs when an export complex reaches the cytoplasmic side of the nuclear envelope after its passage through the NPC. Since the affinity of exportins for their cargo substrates is very low in the absence of RanGTP, GTP hydrolysis on Ran results in the dissociation of export complexes and hence leads to the release of export cargo, RanGDP, and the Exportin 1. Exportin 1 recycles back to the nucleus on its own due to its ability to interact with FG-nucleoporins Merkle (2008).

2.1.2 Nuclear export signals (NESs)

Description

The best understood and widely distributed NES is the so-called leucine-rich NES (LR-NES), which consist of a short leucine-rich stretch of amino acid residues in which the leucine residues are critical for function (Mattaj and Englmeier, 1998; Görlich and Kutay, 1999; Kaffman and O'Shea, 1999; Ossareh-Nazari *et al.*, 2001).

The LR-NES was discovered originally in two proteins, the cellular protein kinase inhibitor (PKI) that terminates PKA-dependent signaling to nuclear targets by exporting PKA out of the nucleus (Wen *et al.*, 1995), and the HIV-1 Rev protein (Fischer *et al.*, 1995). Rev binds to a *cis*-acting RNA sequence, the Rev responsive element (RRE), which is present in unspliced and partially spliced viral mRNAs. Rev facilitates nuclear export of these RNAs to the cytoplasm of infected cells, a function that is necessary for viral protein expression and packaging of unspliced RNA into virions. In the absence of Rev, these viral transcripts are retained in the nucleus where they are either fully spliced or degraded. Two domains of Rev are necessary for its export activity. The first domain directly contacts the Rev-response element in the RNA, and the second domain contains an LR-NES sequence similar to that found in PKI (Nigg, 1997). Deletion and mutational analysis of PKI and Rev revealed that these proteins contain a 10-amino acid sequence essential for export function, which is enriched in hydrophobic residues, specially leucines. These studies provided evidence that the LR-NES is a necessary, suffi-

cient, and transferable signal capable of directing highly efficient nuclear export (Fischer *et al.*, 1995). Nuclear export mediated by the LR-NES was shown to be saturable, suggesting it is a receptor-mediated process. This view was validated with the discovery of Exportin 1 as the NES receptor (Fornerod *et al.*, 1997a; Fukuda *et al.*, 1997; Stade *et al.*, 1997).

By comparing several functional LR-NESs and by applying a randomization-selection approach, a consensus for the LR-NES was initially proposed as:

[L]-$x_{2\text{-}3}$-[F,I,L,V,M]-$x_{2\text{-}3}$-L-x-[L,I] (Bogerd *et al.*, 1996)

where x corresponds to any amino acid residue. With the description of new proteins containing the LR-NES, it was evident that hydrophobic amino acid residues other than leucines were also present in the first and/or third hydrophobic position and the initial consensus was lightly modified (Fornerod and Ohno, 2002; Engelsma *et al.*, 2004; Kutay and Güttinger, 2005). In a more recent publication a yeast selection method for the identification of proteins containing the LR-NES *in vivo* was used (Kosugi *et al.*, 2008). Based on the results, the authors proposed three different consensus sequences which differ basically in the number of spacing amino acid residues present between the hydrophobic ones and in the exclusion of proline. The proposed consensus sequences for the LR-NES are summarized in Figure 2.3 together with LR-NESs of some example proteins. It is worth to mention that besides the canonical LR-NES, less well-defined sequences exist that mediate interaction with Exportin 1 in a RanGTP-dependent manner (Klemm *et al.*, 1997; Paraskeva *et al.*, 1999; Macara, 2001). This dissertation is focused on the LR-NES that is exported by Exportin 1.

Remarkable features of the LR-NES

It has been shown that the consensus sequences alone do not provide a safe way of predicting Exportin 1 substrates from protein sequences. One reason is that leucine is statistically the most abundant amino acid residue in proteins, and the probability of finding such leucine-rich motifs by chance in a given protein is high. In addition, not all sequence motifs that look like an LR-NES also function as such in their original protein context. For example, LR-NES-like sequences have been suggested in yeast Importin β (Iovine and Wente, 1997) and in human Importin α (Boche and Fanning, 1997), but they turned out to be not functional.

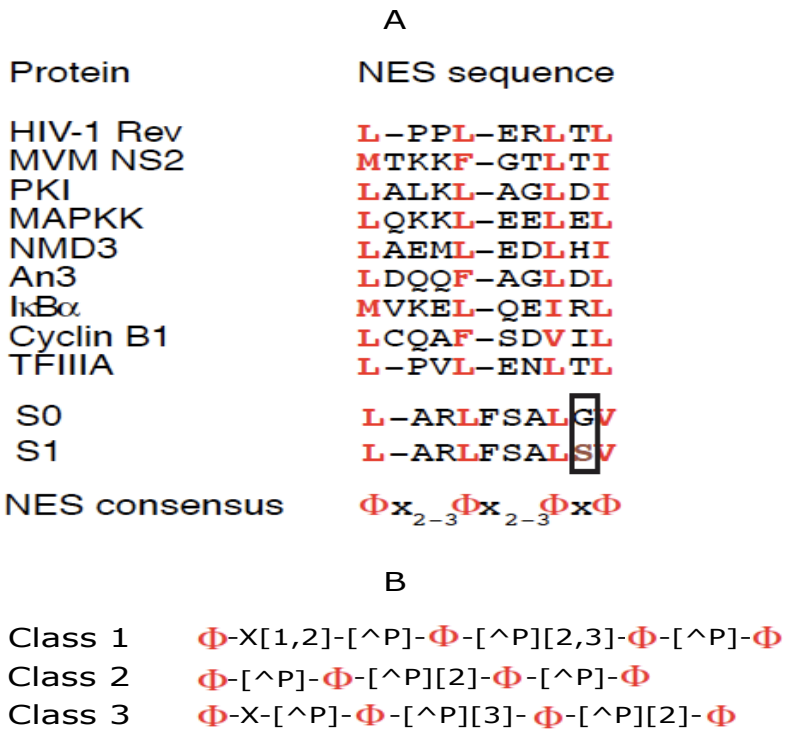


Figure 2.3: Leucine-rich nuclear export signal (LR-NES)

The NES recognized by CRM1 is defined as a short amino acid sequence of regularly spaced hydrophobic residues, of which leucine is statistically the most abundant. **A**: Sequence alignment of identified natural LR-NESs (HIV-1 Rev (Fischer *et al.*, 1995), MVM NS2, PKI (Wen *et al.*, 1995), MAPKK (Fukuda *et al.*, 1996), Nmd3 (Thomas and Kutay, 2003), *Xenopus* An3 (Askjaer *et al.*, 1999), I_k Bα, Cyclin B1 and TFIIIA),together with the artificial LR-NES S1, which has a high affinity for CRM1, binds independently of RanGTP and cannot be released in the cytoplasm (Engelsma *et al.*, 2004). The LR-NES consensus is included below the sequences. Hydrophobic residues are shown in red in the sequences and as ϕ in the consensus, x denotes any amino acid residue. **B**: Three classes of LR-NESs proposed by Kosugi *et al.* (2008), based on a yeast selection method. [$\wedge$P][2,3]: any two or three amino acid residues except proline; ϕ: L, I, V, M, F, C, W, A or T; X3: any three amino acid residues, where C, T, A and W are allowable only at one of the four positions.

Another point to considerer is that not all LR-NESs are exactly alike and individual, isolated NES segments are exported by Exportin 1 with different efficiencies (Henderson and Eleftheriou, 2000). In addition, different affinities of natural LR-NESs for Exportin 1 have been reported together with qualitative differences in the speed of the nuclear export (Askjaer *et al.*, 1999; Heger *et al.*, 2001). Thereby, an excess of proteins containing a fast NES inhibited the export and the biological activity of proteins containing a slower NES, but not vice-versa (Heger *et al.*, 2001), indicating a possible competition of different NES-containing substrates for export factors (Fried and Kutay, 2003).

It has been shown that natural LR-NESs bind to Exportin 1 with relatively low affinity. In fact, high-affinity NESs binding to Exportin 1 impairs the efficient release of Exportin-NES cargo export complexes from the NPC (Fornerod and Ohno, 2002; Kutay and Güttinger, 2005). In this sense, it was shown that although some artificial LR-NES sequences exhibit Exportin 1 binding affinities in the low nM range, 100 to 500-fold higher than the usual NES-Exportin affinity (Askjaer *et al.*, 1999; Paraskeva *et al.*, 1999), they are too strong to be optimal *in vivo* (Engelsma *et al.*, 2004; Kutay and Güttinger, 2005). These high affinity NES-Exportin complexes were trapped at the NPC when over-expressed in living cells (Engelsma *et al.*, 2004). The mentioned high-affinity NES (termed S1 and shown in Figure 2.3) differs from a low affinity LR-NES (termed S0 in the same study), only in a serine instead of a glycine occupying the penultimate amino acid position. It has been proposed that the low affinity between LR-NESs and Exportin 1 possibly enables clearance of the export receptor from the NPC (Kutay and Güttinger, 2005). Nevertheless, in a recent publication, a high Exportin 1 affinity NES (called "supraphysiological" in that report) in a viral protein (MVM NS2) that is required for viral nuclear export was identified in natural host cells (Engelsma *et al.*, 2008). In addition, NMD3, the nuclear export adaptor for 60S pre-ribosomal proteins, has also been reported to behave in a manner reminiscent of a unusual high Exportin 1 affinity NES (West *et al.*, 2007). Hence, the authors speculate that large cargoes (like a viral particle or a ribosomal subunit) require an adaptor with high-affinity interaction for Exportin 1 *in vivo* in order to be exported (Engelsma *et al.*, 2008).

Proteins containing the LR-NES

As already mentioned, the LR-NES confer binding to Exportin 1 and export from the nucleus. This signal functions in a great number of proteins executing quite heterogeneous biological functions (Heger *et al.*, 2001). Those include RNA transport (Pollard and Malim, 1998; Sandri-Goldin, 1998; Krätzer *et al.*, 2000), cell cycle and transcriptional control (Toyoshima *et al.*, 1998; Roth *et al.*, 1998; Stommel *et al.*, 1999; Begitt *et al.*, 2000), regulation of kinase activity (Wen *et al.*, 1995; Fukuda *et al.*, 1996; Engel *et al.*, 1998) or even the controlled localization of cytoskeletal proteins (Wada *et al.*, 1998). Until 2003, at least 75 LR-NES-containing proteins from viruses, yeast and vertebrates had been reported and experimentally validated and they were collected in the database NESbase (la Cour *et al.*, 2003).

LR-NESs are also present in a wide variety of cancer related proteins including p53, c-Abl, FOXO-3A and survivin (Nishi *et al.*, 1994; Altieri, 2006; Knauer *et al.*, 2007a), which use Exportin 1 as receptor. Many important tumor suppressors and transcription factors protect cells by regulating cell growth and apoptosis, and their cytoplasmic localization can serve as an inactivation mechanism resulting in uncontrolled growth and the onset of disease (Vousden and Woude, 2000). One strategy to prevent cytoplasmic localization of those factors is to inhibit the proteins responsible for their nuclear export. Therefore, it has been proposed that the prevention or inhibition of nuclear export of tumour supressors as drug targets could be useful in the treatment of cancer (Turner and Sullivan, 2008). In this sense, export inhibitors, e.g., LMB, have been proposed for anticancer therapy (Vigneri and Wang, 2001), additionally, nuclear export inhibitors (NEI) with the potency of LMB but with better tolerance than LMB *in vivo* have shown efficacy in some models (Mutka *et al.*, 2009). Nevertheless, Exportin 1 directed inhibitors can not be used in therapeutic applications due to their toxic side effects by blocking all Exportin 1 mediated transport pathways (Knauer *et al.*, 2007b). Hence, the interference of the LR-NES binding has been proposed as another strategy to inhibit the nuclear export. Since the exposition of the LR-NES could be modulated by phosphorylation (explained below), blocking of protein modifications (epecially phosphorylation) could prove useful. Furthermore, since NESs can be grouped into specific categories according to their activity *in vivo* (Heger *et al.*, 2001), these differences may represent an attractive opportunity to selectively block export and the biological functions of proteins by the generation of NES specific inhibitors (Knauer *et al.*, 2007a).

Summarizing, Exportin 1 is the nuclear export receptor for most of the known protein cargoes travelling from nucleus to cytoplasm, from which, the majority poses the LR-NES. A summary of the receptors and cargoes so far described is presented in Table 2.1.

2.1.3 NESs in the regulation of nuclear transport

The multi-stage nature of the nucleo-cytoplasmic transport process gives many possibilities for regulation in response to, e.g., environmental, cell cycle, apoptotic and developmental signals (Hood and Silver, 2000; Fried and Kutay, 2003). Many cellular processes from apoptosis to circadian rhythms and from signal transduction to the cell cycle are regulated, at least in part, by modulating NLSs and

Yeast		**Humans**		**Plants***
Export Receptor	*Cargo(es)*	*Export Receptor*	*Cargo(es)*	*Export Receptor*
Crm1p	**LR-NES**	**CRM1**	**LR-NES**	**XPO1a** **XPO1b**
Kap109(Cse1p)	Kap60(Sr1p)	CAS	Importin α	CAS(Exp 2)
Kap127(Los1p)	tRNAs	Exportin-t	tRNAs	Paused(PSD)
-	-	Exportin 4	eIF-5A	-
Kap142(Msn5) *(bidirectional)*	Cdh1*(export)*, Pho4, Crz1, Prot A*(import)*	Exportin 5	pre-miRNA	Hasty(HST)
-	-	Exportin 6	Profilin, Actin	-
-	-	RanBP16 (Exportin 7)	p50-RhoGAP	-

Table 2.1: Exportin receptors and their cargoes

The main receptor for nuclear export is Exportin 1 (shown in blue letters), it is conserved in yeast, vertebrates and plants (**Arabidopsis thaliana*). It is called Crm1p/Xpo1p in yeast, CRM1 in humans and XPO1 in Arabidopsis, which has two of these receptors XPO1a and XPO1b (explained in Section 2.1.4). The most prevalent NES is the LR-NES (colored red), which is present in the proteins being exported by Exportin 1. There are also other Exportin receptors for specific cargoes and some cases of bidirectional receptors that carry out nuclear export as well as nuclear import.

NESs (Macara, 2001). In general, the nucleo-cytoplasmic distribution of proteins can be regulated at the level of the NPC, transport receptors, and signals (NLSs and/or NESs) in individual cargoes. The examples described until now can be grouped in a number of similar common mechanisms (Jans *et al.*, 2000; Hogarth *et al.*, 2005; Poon and Jans, 2005; Sorokin *et al.*, 2007; Terry *et al.*, 2007), some of them directly related with the cargoes signals (NLS and/or NES).

The basic factor determining the nucleo-cytoplasmic distribution of proteins may be the regulation of transport complexes upon their formation due to modulation of receptor-signal (Exportin-NES) interactions, which are very sensitive to conformational changes in the NES regions and in the substrate-binding sites of Exportin. For that reason, the major part of known examples of nucleo-cytoplasmic distribution regulation involves changes in the substrate that affect the exposition of the signal (Poon and Jans, 2005) and in this way, the contact with the Exportin. These mechanisms include NES masking, enhancement of the Exportin-NES interaction and retention of the NES substrate in nucleus. In addition to these,

there are other regulation processes involving cotransport, changes in the cargo-binding properties of karyopherin, in the variety of importins and exportins and in the variety of nucleoporines (Sorokin *et al.*, 2007; Terry *et al.*, 2007). Here, the attention will be focused on the regulation mechanisms involving the NES directly, which are described thereafter.

Masking of the NES from recognition by Exportin

The masking of NESs is the most widespread mechanism of regulation of the nuclear export. Two general classes of masking can be recognized: Intramolecular and intermolecular.

In **intramolecular masking**, upon introduction of charge or conformational changes to the NES-containing region of the protein the access of Exportin to the NES vanishes. An example of this mechanism is the integrase interactor 1 (INI1) from the human SNF5 chromatin-remodeling complex. The C-terminal of this protein masks the NES making it inaccessible for Exportin 1 and thus prevents its nuclear export (Craig *et al.*, 2002).
Intramolecular masking can be also caused by phosphorylation near NES(s) or within them. One example is provided by the telomerase reverse transcriptase, to which 14-3-3 protein can bind in a non-phosphoserine dependent manner and block an NES (Seimiya *et al.*, 2000). Another example, under osmotic stress protein Hog1p (high osmolarity glycerol pathway-signaling protein) is phosphorylated by Pbs2p kinase at Thr174 and Tyr176, which renders the NES inaccessible for binding to Exportin 1 and leads to inhibiting the export of Hog1p from the nucleus (Figure 2.4A) (Ferrigno *et al.*, 1998).
The masking of NESs can be as well the result of conformational changes due to the formation of disulfide bonds between cysteine residues. Thus, under oxidative stress a disulfide bond between Cys598 and Cys620 is formed in the transcription factor Yap1p, which makes its NES inaccessible for interaction with Exportin 1 (Figure 2.4A) (Kuge *et al.*, 2001). A similar mechanism of export regulation was also shown for the transcription factor Pap1 (Kudo *et al.*, 1999b).
Some proteins have more that one NES, which can be differentially regulated. One example of this situation occurs in the protein p53. The tumor suppressor p53 shuttles between the nucleus and the cytoplasm in a cell cycle dependent manner (Stommel *et al.*, 1999) and its nuclear localization is regulated by different mechanisms. One of them is associated with phosphorylation of Ser15/20 in p53 in response to DNA damage that leads to masking of NES1. Another

mechanism consists in tetramerization of protein p53 within the nucleus in response to the DNA damage which causes NES2 masking. Dissociation of this tetramer is required for the export of the protein from the nucleus (Stommel *et al.*, 1999). Additionally, the nucleocytoplasmic shuttling of p53 is further abrogated by stress induced poly ADP-ribosylation that prevents p53 interaction with Exportin 1 (Kanai *et al.*, 2007). Defects in p53 nuclear retention are associated with a number of neoplasms (Jimenez *et al.*, 1999), illustrating the importance of proper cellular localization for single cargoes (Terry *et al.*, 2007).

The **intermolecular masking** consists in the distortion of Exportin-NES interactions caused by the binding of the NES-containing protein to another protein or nucleic acid. As an example, the protein calcineurin can be cited. At high $Ca2^{+}$ concentration, calcineurin ($Ca2^{+}$-responsive phosphatase) binds to the transcription factor NF-AT4 and masks its NES from interaction with Exportin 1, which suppresses nuclear export of the factor (Figure 2.4B). At low $Ca2^{+}$ concentration, calcineurin dissociates from NF-AT4 and unmasks its NES (Zhu and McKeon, 1999).
The binding of the ligand may also cause masking of the NES as was demonstrated for the androgen receptor whose NES is in the ligand-binding domain (Saporita *et al.*, 2003). In the presence of the ligand (androgen), the NES is masked and Exportin 1 cannot recognize it. The receptor is translocated only after the dissociation of androgen (Figure 2.4C).

Enhancement of the Exportin/NES binding affinity

In contrast to the NES masking, as a result of which interactions of Exportin and an NES are distorted, another regulation mechanism consists in the enhancement of the Exportin binding to the NES.
An example of such regulation occurs in the transcription factor Pho4, phosphorylation at Ser114 and Ser128 raises its affinity to Exportin 4 and stimulates its nuclear export (Figure 2.4D) (Komeili and O'Shea, 1999). Some other proteins, like cyclin D1 need to be phosphorylated for binding to Exportin 1 (Benzeno and Diehl, 2004). Although cyclin D1 accumulates in the nucleus during the G1 interval, it relocalizes to the cytoplasm during S phase. The essential functions of cyclin D1 require its nuclear localization, and thus the redistribution of cyclin D1 complexes to the cytoplasm following G1 implies that regulation of cyclin D1 nucleocytoplasmic distribution is necessary for maintaining cellular homeostasis.

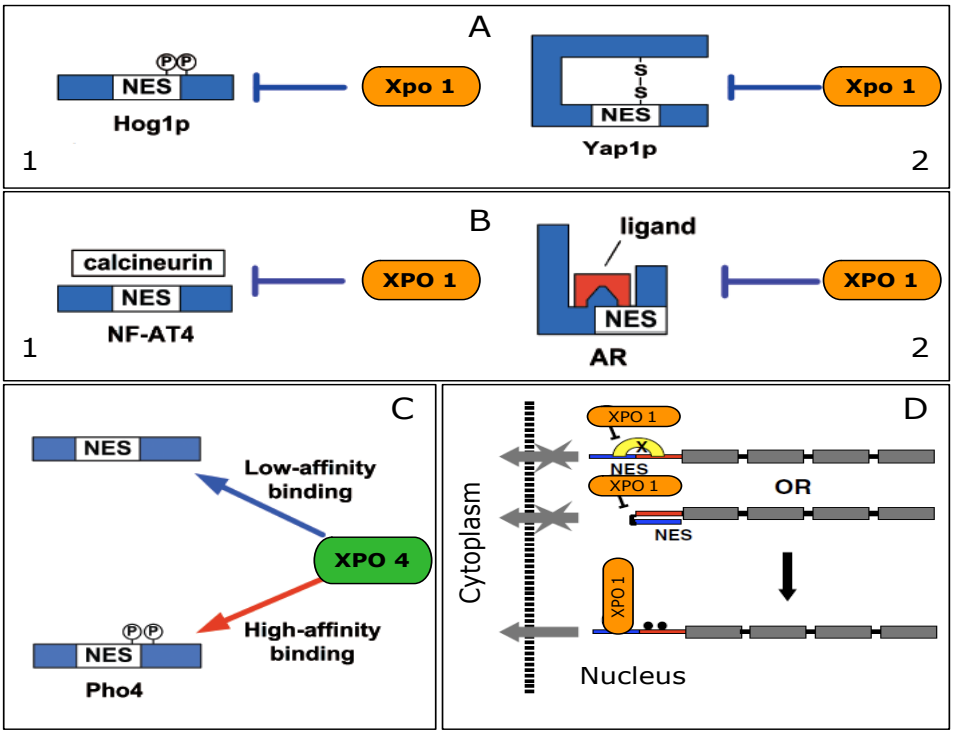


Figure 2.4: Nuclear export regulation involving NESs

Examples of some mechanisms of nuclear export regulation directly related with the NES. **A**: Intramolecular masking. **1**: Phosphorylation of Hog1p at Thr174/Tyr176 masks the NES from recognition by Exportin 1 (Xpo 1). Cytoplasmic translocation is found after dephosphorylation of the residues. **2**: Oxidative stress leads to disulfide linkage in protein Yap1p. The formation of this bond inhibits the binding of Xpo 1 to the NES and leads to accumulation of the protein in the nucleus. **B**: Intermolecular masking. **1**: At high Ca^{2+} concentrations, calcineurin binds to NF-AT4 and mask its NES from interaction with Xpo 1. Nuclear export of NF-AT4 can take place only after dissociation of calcineurin in response to decreasing concentration of Ca^{2+}. **2**: The NES of the Androgen Receptor (AR) is localized to the ligand-binding domain. In presence of ligand (androgen), the NES is masked and Xpo 1 cannot recognize it. The export of AR to the cytoplasm can take place only after androgen dissociation. **C**: Enhancement of NES binding to Exportin. Phosphorylation of Pho at Ser114/128 enhances the RanGTP-dependent interactions of the NES with Exportin 4 (Xpo 4) and leads to its nuclear export. **D**: Nuclear retention. Hypothetical model of how annexin A2 is sequestered in the nucleus, when annexin A2 is in the nucleus, the NES is masked by a protein interacting (X: unknown nuclear factor) with the region containing the NES, or it is masked by that region per se. When interaction domain is modified, the NES is exposed and exported. Figures A to C were modified from (Poon and Jans, 2005; Sorokin *et al.*, 2007), and D from (Liu and Vishwanatha, 2007).

Phosphorylation of cyclin D1 at Thr-286 by GSK-3 promotes Exportin 1 binding, which then shuttles cyclin D1 to the cytoplasm (Alt *et al.*, 2000) for subsequent degradation via the 26S proteasome (Diehl *et al.*, 1998).

Retention in the nucleus or cytoplasm

Another mechanism of nuclear export regulation is realized through the binding of the NES-containing protein to specific nuclear factors that retain proteins in the nucleus or in the cytoplasm in the case of nuclear import. This mechanism plays an important role in the regulation of nucleocytoplasmic distribution of proteins in response to stimulation of different signal pathways in the cell. (Sorokin *et al.*, 2007). There are many examples of this mechanism involving NLSs but few cases have been reported involving NESs directly. For example, the glucocorticoid receptor is retained in the cytoplasm through complexation with Hsp90 in the absence of the ligand. When binding to the hormone (ligand), the glucocorticoid receptor is dissociated from Hsp90 and imported into the nucleus by an NLS dependent mechanism (Tago *et al.*, 2004). Similarly, the tumor repressor p53 is retained in the cytoplasm by protein Parc (Parkin-like ubiquitin ligase). A nuclear retention example is Human annexin A2, it is proposed that the NES of this protein is masked by an interacting protein in the region tethering the NES, or is masked by that region per-se and in this way is sequestered in nucleus. When the interaction domain is modified, the NES is exposed and the export occurs subsequently (Liu and Vishwanatha, 2007). Another example involving retention in nucleus occurs in the tumor suppressor Rb protein. It was shown that Rb shuttles between nucleus and cytoplasm with similar rates, and is partially immobile in both cellular compartments, in part due to Rb binding to the microtubule network (Roth *et al.*, 2009).
Nuclear or cytoplasmic retention can additionally be regulated by phosphorylation. For example, the nuclear retention of IFI16 is enhanced after phosphorylation of NLS by kinase CK II (Briggs *et al.*, 2001). In cyclin B1, phosphorylation of its NES at the onset of mitosis leads to nuclear retention by blocking the interaction with Exportin 1 (Yang *et al.*, 2001).

2.1.4 Nuclear export in *Arabidopsis thaliana*

Analysis of plant proteins involved in nuclear export has shown that although this process is highly conserved between organisms from yeast to vertebrates (Ward

and Lazarowitz, 1999; Merkle, 2003; Meier, 2005), there are some plant-specific features.

The existence of a nuclear export pathway for proteins carrying the LR-NES in plants has been demonstrated (Haasen *et al.*, 1999a). In this way, the orthologue of Exportin 1, XPO1, has been caracterized in Arabidopsis (Haasen *et al.*, 1999a,b). The plant XPO1 interacts with RanGTP and with the LR-NES of Arabidopsis proteins like RanBP1a and proteins of other organisms like HIV Rev suggesting a high degree of conservation of this nuclear export pathway between organisms. Point mutation of specific hydrophobic residues within the LR-NES of AtRanBP1a abolished the interaction with XPO1 (Merkle, 2003). The interaction with proteins containing the LR-NES is also blocked by Leptomycin B (LMB) (Haasen *et al.*, 1999a), by the same mechanism as in Exportin 1 from vertebrates (Kudo *et al.*, 1999a). In addition, nuclear export activity in plants was demonstrated *in vivo* by using green fluorescent protein (GFP) fusion proteins. In tobacco BY-2 protoplasts, the NES of RanBP1a or the NES of HIV Rev was sufficient to confer nuclear export to GFP fusion proteins and the nuclear export activity was sensitive to LMB (Haasen *et al.*, 1999a). As a special feature and in contrast to the situation in yeast and vertebrates, Arabidopsis contains two genes encoding highly similar XPO1 proteins. These two genes were shown to be essential for development and function of the gametophytes (Blanvillain *et al.*, 2008). The proteins coded by the two XPO1 genes are called XPO1a and XPO1b (see Table 2.1).

In addition to XPO1, three more exportins have been characterised in Arabidopsis. One of them is Hasty (HST), which encodes the putative orthologue of human Exportin 5/yeast Msn5p. HST interacts with the GTPase Ran in the yeast two-hybrid system, and a fusion protein of HST with β-glucuronidase (GUS) was shown to be localised to the periphery of nuclei in transgenic plants (Bollman *et al.*, 2003). The loss of HST provokes a variety of developmental phenotypes (Bollman *et al.*, 2003). The second Exportin is Paused (PSD), the Arabidopsis orthologue of Exportin-t/Los1p. Evidence of PSD function as a nuclear export receptor for tRNA was provided by complementation of the los1-1 mutant of *S. cerevisiae* (Hunter *et al.*, 2003). Like HST, mutants of PSD have pleiotropic effects in plan development (Hunter *et al.*, 2003; Li and Chen, 2003). The third of the exportins characterized so far in Arabidopsis is CAS, the nuclear export receptor for importin α (Haasen and Merkle, 2002).

As explained before, the action of exportins alone is not sufficient to drive the transport across the NE, in this sense, some components of the Ran cycle have been identified in plants. Ran has been found in a variety of plant species (Ach and Gruissem, 1994; Merkle *et al.*, 1994; Saalbach and Christov, 1994). Genes encoding the GTPase Ran have been isolated from several plant species (Merkle and Nagy, 1997). Four Ran GTPases are present in Arabidopsis (Vernoud *et al.*, 2003). AtRAN1, AtRAN2, and AtRAN3 has been identified by sequence similarity and isolated (Xia *et al.*, 1996; Haizel *et al.*, 1997), whereas the fourth gene, AtRAN4, is annotated as "salt stress-inducible small GTP-binding protein Ran1-like protein", but so far no information has been published (Vernoud *et al.*, 2003). At the protein level, AtRAN1, AtRAN2, and AtRAN3 are nearly identical (95%-96% of identity) differing only in their C-terminal regions, whereas AtRAN4 is more divergent with only 65% identity to the other AtRAN sequences. AtRAN1, AtRAN2, and AtRAN3 are able to interact with Arabidopsis RanBP1 in the yeast two-hybrid system, using Ran mutants that are permanently blocked in the GTPbound form. All AtRan GTPases contain sequence motifs involved in GTP binding/hydrolysis and an effector-binding domain for interaction with RanGAPs. This effector-binding motif is 100% identical in AtRAN1 to AtRAN3 and in tomato and tobacco Ran GTPases but diverges strikingly in AtRAN4, which does not have any typical conserved C-terminal acidic domain at all. Since in animals this acidic domain is necessary for interaction with Ran-binding proteins (Haizel *et al.*, 1997), it is likely that AtRAN1 to AtRAN3 are involved in nucleocytoplasmic transport whereas AtRAN4 may have distinct functions in Arabidopsis (Vernoud *et al.*, 2003).

Regarding Ran binding proteins (RanBP), three genes encoding RanBP1 proteins were isolated from Arabidopsis (Xia *et al.*, 1996; Haizel *et al.*, 1997). AtRanBP1 proteins contain a Rev-type LR-NES in the C-terminal that confers interaction with Arabidopsis XPO1. A functional NES was necessary for the nuclear exclusion of a GFP-AtRanBP1a fusion protein in protoplasts (Haasen *et al.*, 1999a). The presence of an NES in AtRanBP1c has also been demonstrated, additionally, it was shown that this protein is functional as a co-activator of RanGAP *in vitro* (Kim and Roux, 2003). It has been established that AtRAN1 interacts with AtXPO1 and AtRanBP1a (Haizel *et al.*, 1997; Haasen *et al.*, 1999a), suggesting that the nuclear export machinery may be functionally conserved in plants (Haasen *et al.*, 1999a).

Two RanGAP sequences have been identified in Arabidopsis: AtRanGAP1 and AtRanGAP2. Both of them complemented yeast RanGAP mutants (Pay *et al.*, 2002), suggesting that these proteins are functional orthologs of the yeast RanGAP. AtRanGAP-GFP fusions associate with the nuclear envelope, and this localization is dependent upon a unique N-terminal domain (Rose and Meier, 2001). AtRanGAP1 localization is consistent with a role for AtRAN GTPases in nucleocytoplasmic transport. These facts together suggest that plant Ran, RanBP1 and RanGAP proteins have functions that are similar to those of the vertebrate and yeast proteins (Vernoud *et al.*, 2003).

Concerning the signals, animal and yeast NLSs and NESs have been found to be functional in plants (Smith *et al.*, 1997; Ward and Lazarowitz, 1999; Merkle, 2001), and endogenous NLSs and NESs have been identified in some plant proteins. For example, the movement protein BR1 of the squash leaf curl virus has a functional LR-NES and two basic NLSs (Ward and Lazarowitz, 1999). The NES of Arabidopsis RanBP1a is functionally indistinguishable from the NES on the HIV-1 Rev Protein (Haasen *et al.*, 1999a).

2.2 Computational prediction of NESs

The general components of the nuclear export process are quite conserved among eukaryotes, however, many of the NES-containing proteins are species-specific. This fact together with the functional diversity of the NES-containing proteins, make their identification of great interest. The approaches used to identify proteins containing the LR-NES are traditionally experimental and include techniques like yeast two-hybrid screenings using the receptor Exportin as a bait. Given the growing availability of sequence data for many organisms, many studies have been directed to develop computational tools that can predict specific signals from the amino acid sequences and in this way support and speed up the experimental approaches.

The task has some inherent difficulties in the case of predicting LR-NESs. As was already exposed, leucine is the most abundant amino acid residue in proteins from eukaryotes, the signal is relatively short (around 10 amino acid residues) and is has been already shown that the consensus alone does not provide an efficient method for prediction. Nevertheless, one NES prediction tool is currently available (la Cour *et al.*, 2004). It was developed using a combination of Hidden

Markov Models (HMM) and Neural Networks (NN). This tool has been proved to be accurate in the prediction of LR-NESs in proteins of organisms like viruses and yeast, but it is not useful to identify the signals in proteins from plants. This fact was evidenced since some proteins from Arabidopsis with an experimentally validated LR-NES (previous work in laboratory of Dr. Thomas Merkle) were not predicted as containing any LR-NES when screened with that tool.

In contrast to NESs, considerable work has been done in developing computational prediction methods for NLSs. At present there are several tools available for finding NLSs in protein sequences (Hawkins *et al.*, 2007). Some of them are:

PredictNLS	`http://www.rostlab.org/services/predictNLS/` (Cokol *et al.*, 2000)
NLSdb	`http://cubic.bioc.columbia.edu/db/NLSdb/` (Nair *et al.*, 2003)
NucPred	`http://www.sbc.su.se/maccallr/nucpred` (Brameier *et al.*, 2007)
NLStradamus	`http://www.moseslab.csb.utoronto.ca/NLStradamus/` (Ba *et al.*, 2009)

However, NESs are more difficult to identify compared with classical NLSs because, in addition to the reasons exposed before, they share sequence similarity to regions that form the hydrophobic core of many proteins (Cook *et al.*, 2007).

Hidden Markov Models (HMM) and supervised machine learning classification approaches have been widely employed for devising rules that can be used for targeting signal prediction and they will be used in this work as well. As has been already mentioned, consensus sequences are not robust enough to be used directly to find LR-NESs in amino acid sequences, therefore profile HMMs can be an alternative.
On the other hand, the goal of identifying an LR-NES in a protein sequence can be regarded as a two-class classification problem: given an amino acid sequence, the goal is to decide if it could be an LR-NES or not, that means, "classify" the sequences as "positive" or "negative" wrt. LR-NES. This task can be addressed using statistical classification techniques from the field of machine learning.
The general principles of these techniques are presented below together with a general introduction to Hidden Markov Models (HMM) and profiles since they are used as well in this work.

2.2.1 Supervised machine learning classification

The term machine learning refers to a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data (Tarca *et al.*, 2007). Two main branches exist in the field of machine learning: **supervised** and **unsupervised** learning (Duda *et al.*, 2001; Tarca *et al.*, 2007; Hastie *et al.*, 2009).
Supervised learning is a technique for deducing a function from training data. The training data consist of pairs of input objects (typically vectors), and desired outputs. The output of the function can be a continuous value (called regression), or can predict a class label of the input object (called classification). The task of the supervised learner is to predict the value of the function for any valid input object after having seen a number of training examples (i.e. pairs of input and target output).
In contrast to the supervised framework, in **unsupervised learning**, no predefined outputs (values or labels) are available for the objects under study. In this case, the goal is to explore the data and discover similarities between objects, referred to as *clusters*.
The goal of identifying LR-NESs in protein sequences is achieved in the frame of this work by using supervised classification learning, therefore, it will be described in more detail.

Supervised machine learning methods have been widely used in bioinformatics prediction applications. The following are just some relevant examples: subcellular location of proteins (Reinhardt and Hubbard, 1998; 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.*, 2009a), 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, 2008a) and glycosylation (Caragea *et al.*, 2007).

For the formal description, vector notation ($\mathbf{x}$ denotes an ordered p-tuple of numbers for some integer p) and matrix notation (X denotes a rectangular array of numbers, where x_{ij} denotes the number in the ith row and the jth column of X) will be used here.

In supervised classification, a classification function is learned from a so called *training set* of items with known class labels. The general task is to classify a collection of objects $i = 1, ...n$ into K predefined classes. In the case of the intended classification of amino acid sequences as NES or nonNES, $K = 2$. The class labels can be for instance $+1$ and -1.
Data and features can be organized in a matrix $X = (x_{ij})$, where x_{ij} represents the measured value of the feature j in the sample i. Every row of the matrix X is therefore a vector $\mathbf{x}_i$ with n features to which a class label y_i is associated, being $y = 1, ..., K$. In a classification problem, a classifier $C(\mathbf{x})$ may be viewed as a collection of K discriminant functions $g_c(\mathbf{x})$ such that the object with feature vector $\mathbf{x}$ will be assigned to the class c for which $g_c(\mathbf{x})$ is maximized over the class labels $c \in \{-1, +1\}$. The feature space X is thus partitioned by the classifier $C(\mathbf{x})$ into K disjoint subsets.
There are two main approaches to the identification of the discriminant functions $g_c(\mathbf{x})$ (Webb, 2005; Hastie *et al.*, 2009). The first assumes knowledge of the underlying class-conditional class probability density functions (the probability density function of $\mathbf{x}$ for a given class) and assigns $g_c(\mathbf{x}) = f(p(\mathbf{x} \mid y = c))$, where f is a monotonic increasing function, for example the logarithmic function. Intuitively, the resulting classifier will classify an object $\mathbf{x}$ in the class in which it has the greatest membership probability. In practice $p(\mathbf{x} \mid y = c)$ is unknown and therefore needs to be estimated from a set of correct classified samples, the training set. Parametric (for example, linear and quadratic discriminants) and non parametric methods (for example, the k-Nearest Neighbor (k-NN) decision rule) can be used for that end.
The second approach is to use the data to estimate the class boundaries directly, without explicit calculation of the probability density functions. Examples of algorithms in this category include decision trees, neural networks and support vector machines (SVM).

The process described above is called training of the classifier. Subsequently, the trained classifier, i.e., the learned classification function is applied to classify items with unknown class affiliation. In this process, a class label $y \in Y$ is assigned to new items $\mathbf{x} \in X$, called test items or *test set*.
If a learned classification function is able to nicely reproduce the class labels of the training set it is called well fit to the training data. On the other hand, the ability of a classifier to correctly predict the class labels of so far unseen items, which were not contained in the training set is called generalization. A non trivial

task in machine learning is to find a good balance between a classifier that is well fit to the training data and at the same time has a good generalization ability. For example, if a training set contains outliers (e.g. items with wrong class labels) a complex, perfectly fitted classifier might achieve only a poor generalization ability. Such a classifier is then called overfitted or overtrained.

A main goal of supervised classification is to learn a classification function $g_c(\mathbf{x})$ based on the training set that minimizes the *error rate (Err)*, which can be defined as the average number of misclassified samples among the total number of objects. Conversely the *accuracy (Acc)* of the classifier can be defined as $Acc = 1 - Err$ and represents the fraction of samples successfully classified and can be taken as an indicative of the classifier performance if the class distribution is balanced (the proportion of samples pertaining to each class is similar in the complete population).

Resampling methods

The goal behind developing classification models is to use them to predict the class membership of new samples. If the data used to build the classifier is also used to compute the error rate, then the resulting error estimate, called the resubstitution or training error, will be optimistically biased (Efron, 1983). Thus, an essential issue in machine learning relates to judge generalization capability or its ability of correctly predicting unseen examples of the learning method. The degree of generalization capability is evaluated by the closeness between the learned function and the true function, measured by the prediction or generalization error. In machine learning problems, a good classifier is one that minimizes the prediction error (produces good predictions) and not the training error on a particular data set.
In the absence of a large, independent test set, there are some techniques for assessing prediction error by implementing some form of partitioning or resampling of the original observed data. Each of these techniques involves dividing the data into a *training set* and a *test set*. For purposes of model fitting, the training set can be further divided into a training set and a validation set (Hastie *et al.*, 2009). The methods most commonly used are described below.

- **Split sample**: This method, also known as the training-test split or hold-out (McLachlan, 1992), entails a single partition of the data into a training an a test set based on a predetermined p. For example $p = \frac{1}{3}$ allots two-thirds of the data to the learning set and one-third to the test set.

- ***v*-fold cross validation (CV)**: This method randomly assigns the n observations to one of v partitions such that the partitions are near-equal size. Subsequently, the training set contains all but one of the partitions which is labeled as the test set. The error is assessed for each of the v test set and averaged over v.

- **Leave-one-out-cross-validation (LOOCV)** This is the most extreme case of v-CV. In this method each observation is individually assigned to the test set, i.e. $v = n$ and $p = \frac{1}{n}$ (Stone, 1974, 1978). That means, a single object is removed from the training set and the classifier is trained with the remaining data. Subsequently, in each step a single item is classified and the generalization error measured.

- **Bootstrap**: This technique uses sampling *with replacement* to form the training set. Several variations of the bootstrap method have been introduced to estimate the generalization error. The leave-one-out bootstrap ($\hat{\theta}_n^{BS}$) is based on a random sample drawn with replacement from the total of observations (Efron, 1983). For each draw, the observations left out ($\approx .368n$) serve as the test set. The training set has $\approx .632n$ unique observations which leads to an overestimation of the prediction error. To correct for this, two estimators have been suggested: the .632 bootstrap and the .632+ estimator. Both rectify the prediction error by adding the underestimated resubstitution error $\hat{\theta}_n^{RS}, \omega\hat{\theta}_n^{BS} + (1-\omega)\hat{\theta}_n^{RS}$. For the .632 bootstrap the weight is constant ($\omega = 0.632$), whereas for the .632+ bootstrap ω is determined based on the "no-information error rate" (Efron and Tibshirani, 1997). The .632+ estimator is the most used in literature and the most robust across different classification algorithms (Efron and Tibshirani, 1997).

2.2.2 Machine learning algorithms used in this work

In this work, three algorithms for supervised classification are used: k-Nearest Neighbor (k-NN), Random Forest (RF) and support vector machine (SVM).

k-Nearest Neighbor (*k*-NN)

The k-NN classifier can be seen as a nonparametric method of density estimation i.e. it does not assume an underlying distribution of the data (Duda *et al.*, 2001;

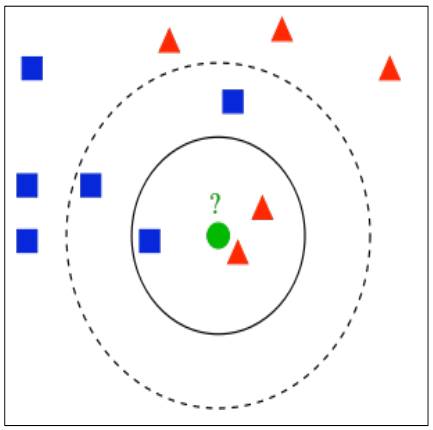


Figure 2.5: Principle of the *k*-Nearest Neighbor (*k*-NN) algorithm
The k-Nearest Neighbor (k-NN) algorithm is illustrated in this 'toy problem'. The data corresponds to two-dimensional input vectors $\mathbf{x}$, and associated y class values which can be either 'square' or 'triangle'. Given, a a new two-dimensional point $\mathbf{z}$, depicted below as a green circle, the goal is to decide whether it corresponds to 'square' or 'triangle' class. k-NN simply looks at the new point, and finds the closest points in the training set (the nearest neighbors) in order to decide how to classify the new point. The inner circle demonstrates that, when the closest 3 points (k = 3) are used, the algorithm will predict 'triangle'. However, the outer circle shows that, when k = 5 is used, it will predict 'square'. Since the distance function used by k-NN (i.e. Euclidean distance) can work with vectors of any dimension, this principle can be applied to input vectors of any size.

Hastie *et al.*, 2009), except for the continuity of the feature variables. It is one of the oldest and simplest methods for statistical classification, developed more that forty years ago (Cover and Hart, 1967).

The k-NN classifier does not require model fitting but simply stores the training dataset with all available vector prototypes of each class. A new object is classified by a majority vote of its neighbors, with the item being assigned to the most common class among its k nearest neighbors. So, when a new object $\mathbf{z}$ needs to be classified, the first step in the algorithm is to compute the distance between $\mathbf{z}$ and all the available objects in the training set, $\mathbf{x}_i, i = 1, ..., n$. A widely used distance metric is the Euclidean distance: $d_{euc}(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{j=1}^{p}(x_j - z_j)}$.

The distances are ordered and the top k training samples (closest to the new object to be predicted) are retained. Let n_c denote the number of objects in the training dataset among the k ones which belong to the class c. The k-NN classification rule classifies the new object $\mathbf{z}$ in the class that maximizes n_c, i.e., the class that is most common among the closest k neighbors (Figure 2.5). The k-NN discriminant functions can be written as $g_c(\mathbf{x}) = n_c$. When the two classes of the example (NES or nonNES) are equally represented in the vicinity of the point $\mathbf{z}$,

the class whose prototypes have the smallest average distance to $\mathbf{z}$ may be chosen. In the case of the k-NN classifier, the number of k nearest neighbors to be considered in the classification step becomes relevant. The choice of the parameter k depends upon the classification problem. In general, larger values of k will increase the bias and reduce the variance of the classifier and vice versa. Small values of k result in decision boundaries with higher variance that fit well the training set, while large values achieve smooth and stable decision boundaries that avoid overfitting and are more robust (Hastie *et al.*, 2009).

Random Forests (RF)

A Random Forest (RF) is a combination (in machine learning referred to as *ensemble*) of many decision trees, where each tree is grown using a (bootstrap) subset of the training dataset (Breiman, 2001). A decision tree is a special type of classifier (Breiman, 1984), which is trained by an iterative selection of individual features that are the most salient at each node in the tree.
In a classical decision tree, the input space X is repeatedly split into descendant subsets, starting with X itself. There are several heuristic methods for constructing decision-tree classifiers. They are usually constructed top-down, beginning at the root node and successively partitioning the feature space. The construction involves three main steps. (i) Selecting a splitting rule or decision criteria for each internal node, i.e., determining the feature together with a threshold that will be used to partition the dataset at each node. (ii) Determining which nodes are terminal nodes. This means that for each node a decision must be achieved whether to continue splitting or to make the node terminal and assign to it a class label. (iii) Assigning class labels to terminal nodes by minimizing the estimated error rate. The most commonly used decision tree classifiers are binary. They use a single feature at each node, resulting in decision boundaries that are parallel to the feature axes (Figure 2.6A). The Random Forest method improves on this procedure by building a large number decision trees (a forest of them) using random subsets of the training data for each one (Figure 2.6B). The resulting forest of trees can then be used to "vote" the most likely class of new data points. This use of random subsets of the training data avoids the problem of overfitting in which the classifier follows too closely the peculiarities of the training data to accurately classify new datasets.
The algorithm for inducing a Random Forest was developed by Leo Breiman and Adele Cutler (Breiman, 2001), and *Random Forests* is their trademark. The term

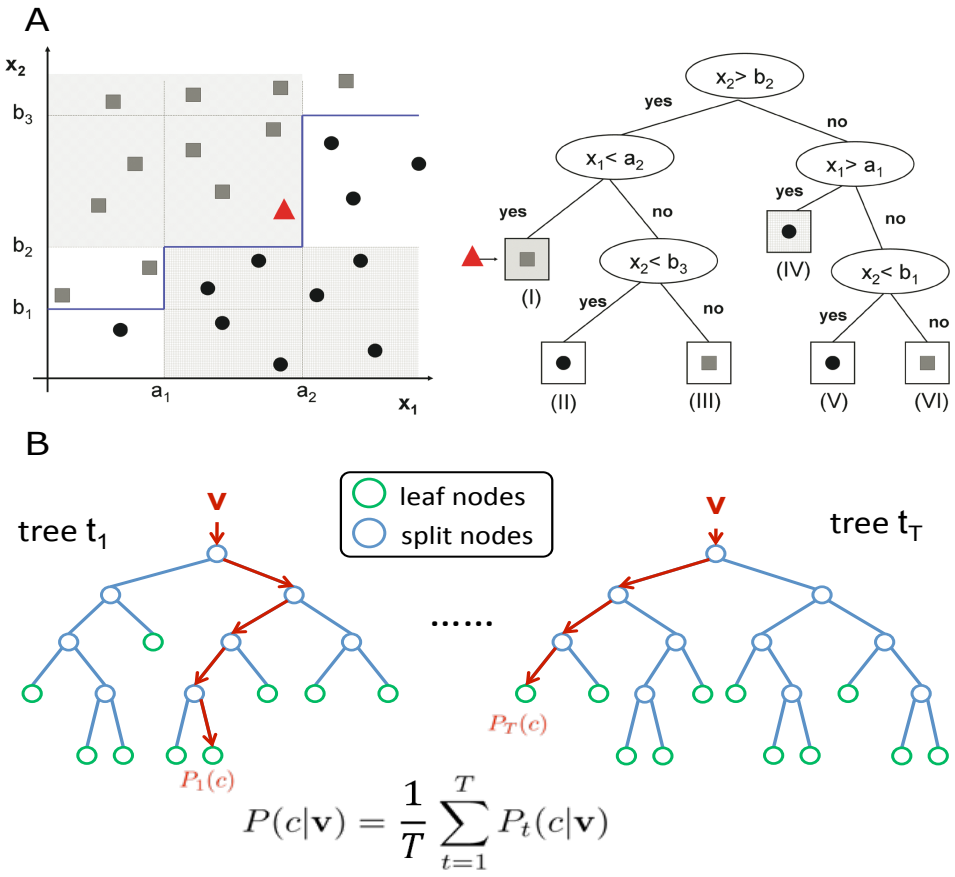


Figure 2.6: Principle of decision trees and Random Forest

A: Foundation of a binary decision tree, the left panel shows the data for a two-class decision problem, with dimensionality $p = 2$. The points known to belong to classes 1 and 2 are displayed with filled circles and squares, respectively. The decision boundary is shown as the blue thick line. The triangle designates a new point, $\mathbf{v}$ to be classified. The right panel shows the decision tree derived from this dataset whereas the new point $\mathbf{v}$ is classified in class 2 (squares). The regions in the input space covered by nodes I and IV in the tree are represented by the dashed areas at the top and bottom of the left panel respectively. This part of the figure was modified from Tarca *et al.* (2007).

B: In supervised machine learning, a Random Forest is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. During the construction of a single classification tree, logical if-then conditions are determined for predicting or classifying cases. At the end, the class label for a new sample $\mathbf{v}$ is assigned by averaging the class votes among all the trees in the forest.

came from the random decision forests that were first proposed by Tin Kam Ho (Ho, 1995). Random Forest combines a special model averaging approach called bagging (acronym for **b**ootstrap **agg**regat**ing**) (Breiman, 1996) with random selection of features (Ho, 1998) to construct a collection of decision trees with controlled variation. Given a standard training set D of size n, bagging generates m

new training sets D_i of size $n' \leq n$, by sampling examples from D uniformly and with replacement. By sampling with replacement it is likely that some examples will be repeated in each D_i. The m models are fitted using the above m bootstrap samples and combined by averaging the voting (Figure 2.6B).

Support Vector Machine (SVM)

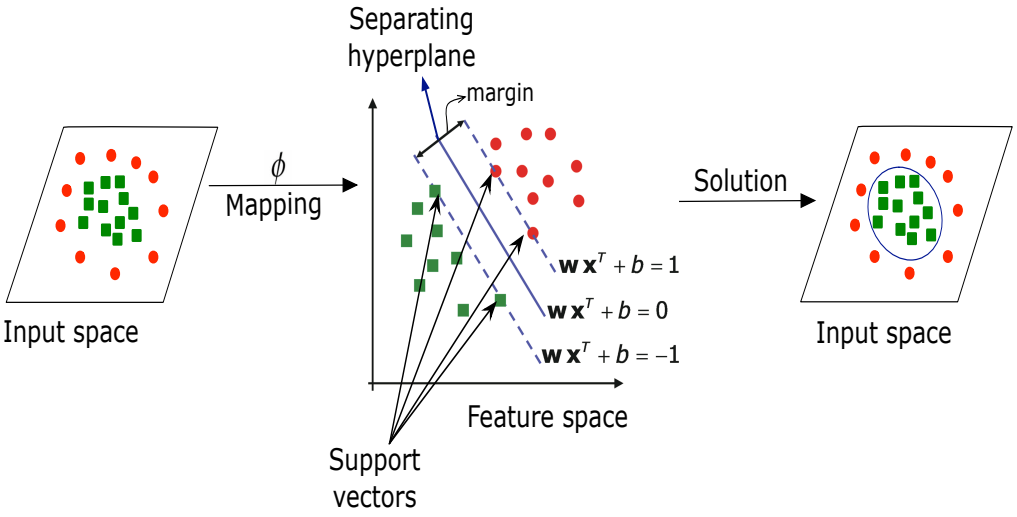


Figure 2.7: Principle of the Support Vector Machine (SVM) algorithm
The classical SVM is a data-driven method for binary classification. SVMs are generated by a two step procedure: first, the sample data vectors $\mathbf{x}$ are mapped or projected to a high dimensional space by using the function ϕ. Then, the algorithm finds a hyperplane in this high-dimensional space with the largest margin separating classes of data. The points classified by SVM can be divided into two groups, support vectors and nonsupport vectors. Nonsupport vectors are classified correctly by the hyperplane and are localized outside the separating margin. The Parameters of the hyperplane do not depend on them, and even if their position is changed, the separating hyperplane and margin will remain unchanged. By the contrast, the support vectors determine the exact position of the hyperplane. After that, the solution can be applied on the original input space.

Compared to k-NN and RF classifiers, the Support Vector Machine (SVM) algorithm was proposed (Boser *et al.*, 1992) and developed (Vapnik, 1995, 1998) more recently. The SVM was developed as a binary (two class) classifier, implementing the following idea: it maps the input vectors $\mathbf{x}$ into a high-dimensional feature space Z through some non-linear mapping, chosen *a priori*. In this space an optimal separating hyperplane is constructed (Vapnik, 1998) and in such a way, a non-linear classifier can be accomplished in the original input space (Yang, 2004; Noble, 2006; Tarca *et al.*, 2007; Ivanciuc, 2007) (See Figure 2.7).

When considering a binary classification problem, SVMs find the decision boundary that achieves maximum margin between the two classes. From statistical learning theory, the decision functions derived by maximizing the margin minimize the theoretical upper bound on the expected risk and are thus expected to generalize well (Vapnik, 1998). The margin is defined as the distance between a planar decision surface that separates two classes and the closest training samples to the decision surface (See Figure 2.7). Let denote with $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{N_T}, y_{N_T})$ the labelled training data set where $\mathbf{x}_i \in \Re^p$, $y_i \in \{-1, +1\}$, being -1 and $+1$ the class labels. SVMs find an optimal hyperplane $\mathbf{w}\mathbf{x}^T + b = 0$, where $\mathbf{w}$ is the p-dimensional vector perpendicular to the hyperplane and b is the bias. The objective of training SVMs is to find $\mathbf{w}$ and b such that the hyperplane separates the data and maximizes the margin $\frac{1}{\|\mathbf{w}\|^2}$ (Figure 2.7, central panel). By introducing non-negative slack variables ξ_i and a penalty function measuring classification errors, the linear SVM problem is formulated as follows:

$$\min_{\mathbf{w}}(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N_T}\xi_i) \tag{2.1}$$

$$f(\mathbf{x}) = sign(\mathbf{w}\ \mathbf{x}^T + b) = sign(\sum_i \alpha_i y_i(\mathbf{x}_i\ \mathbf{x}^T) + b) \tag{2.2}$$

where C is a parameter (usually referred to as *cost parameter*) to be set by the user, which controls the penalty to errors. The optimization problem can be reduced to a dual problem with solutions given by solving a quadratic programming problem (Vapnik, 1998) and the decision function is given by:

$$f(\mathbf{x}) = sign(\mathbf{w}\ \mathbf{x}^T + b) = sign(\sum_i \alpha_i y_i(\mathbf{x}_i\ \mathbf{x}^T) + b) \tag{2.3}$$

where α_i are coefficients that can be solved through the dual problem. Data points with non-zero α_i are called support vectors (SVs) (see Figure 2.7, central panel). In SVMs, only SVs contribute to the construction of the decision boundaries (Chen *et al.*, 2005; Ben-Hur *et al.*, 2008). The linear SVMs can be extended to nonlinear SVMs where more sophisticated decision boundaries are needed. This is done by applying a kernel transformation, i.e., replacing every matrix product $(\mathbf{x}_i\ \mathbf{x}^T)$ in linear SVMs with a non-linear kernel function evaluation $K(\mathbf{x}_i\ \mathbf{x})$. This is equivalent to transforming the original input space X non-linearly into a high-dimensional feature space Z, as denoted above. The training data that are not linearly separable in the original feature space can be linearly separated in the

transformed feature space. Consequently, the decision boundaries are linear in the projected high-dimensional feature space and non-linear in the original input space. Two commonly used kernels include polynomial

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}\mathbf{z}^T + 1)^d \tag{2.4}$$

and radial basis function (RBF)

$$K(\mathbf{x}, \mathbf{z}) = \exp(-\sigma\|\mathbf{x} - \mathbf{z}\|^2) \tag{2.5}$$

The kernel functions return larger values for arguments that are closer together in feature space. In constructing linear SVMs for classification, the only parameter to be selected is the cost parameter C. C controls the tradeoff between errors of SVMs on training data and the margin. For nonlinear SVMs, the learning parameters include C and parameters associated with the kernels used, e.g., σ, in radial basis function (RBF) kernels. In practice, learning parameters are selected through resampling methods like cross-validation.

2.2.3 Hidden Markov Models and profiles

Hidden Markov Models

Hidden Markov Models (HMMs) are a general statistic modelling for "lineal" problems like sequences or time series (Eddy, 1998). They provide a formal foundation for making probabilistic models of linear sequence "labelling" problems (Baldi and Brunak, 2001; Eddy, 2004; Durbin, 2006). The key idea is that an HMM is a finite model that describes a probability distribution over an infinite number of possible sequences (Eddy, 1998). The HMM is composed of a number of states, which might correspond to for example, columns of a multiple sequence alignment. Each state "emits" symbols (i.e amino acid residues) according to symbol-emission probabilities, and the states are interconnected by state-transition probabilities that describe the linear order in which the state is expected to occur.
The sequence of states is a Markov chain, because the choice of the next state to occupy depends on the identity of the current state. However, this state sequence is not observed: it is hidden. Only the symbol sequence that these hidden states generate is observed, hence the models are called 'hidden' (Eddy, 1996, 1998, 2004). Once a HMM is drawn, regardless of its complexity, the same standard

dynamic programming algorithms can be used for aligning and scoring sequences with the model. These algorithms are called Forward (for scoring) and Viterbi (for alignment) (Durbin, 2006). Parameters can be set for an HMM in two ways. An HMM can be trained from initially unaligned (unlabelled) sequences. Alternatively, an HMM can be built from pre-aligned (pre-labelled) sequences (i.e. where the state paths are assumed to be known). In the latter case, the parameter estimation problem is a matter of converting observed count of symbols emissions and state transitions into probabilities.

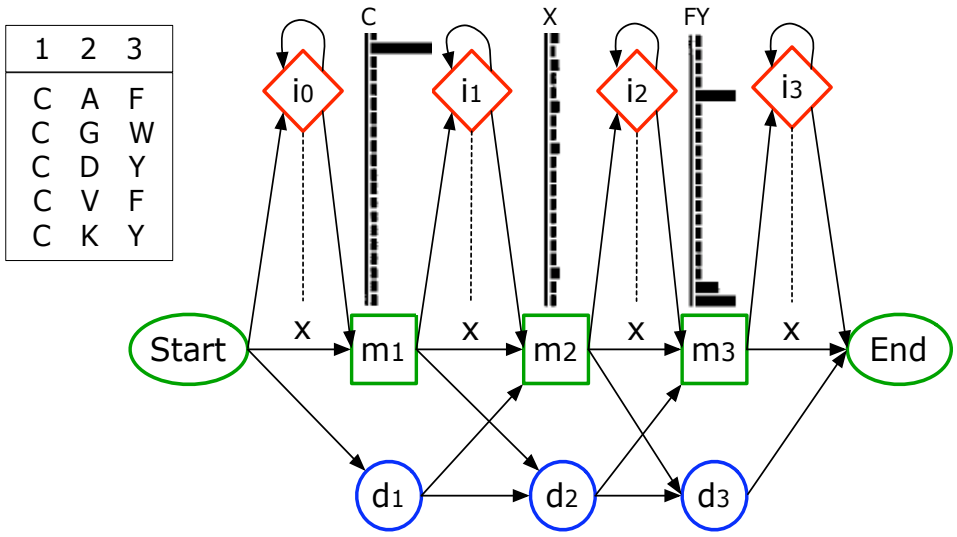


Figure 2.8: Profile HMM

A small profile HMM (right) representing a short multiple alignment of five amino acid sequences (left) with three consensus columns. The three columns are modeled by three match states (squares labeled m_1, m_2 and m_3), each of which has 20 residue emission probabilities, shown wit black bars. Insert states (diamonds labeled i_0 to i_3) also have 20 emission probabilities each. Delete states (circles labeled d_1 to d_3) are 'mute' states that have no emission probabilities. A *start* and *end* state are included. State transmission probabilities are shown as arrows.
This figure was lightly modified from (Eddy, 1998).

Profile HMMs

Profile HMMs are statistical tools that can model the commonalities of the amino acid sequences for a family of proteins. Considered to be more expressive than a standard consensus sequence or a regular expression, profile HMMs allow position dependent insertion and deletion penalties, as well as the option to use a separate distribution for inserted portions of the amino acid sequence. The architecture was introduced by Krogh *et al.* (1994) and is considered to be well suited

for representing profiles of multiple sequence alignments (Eddy, 1998). For each consensus column of the multiple alignment, a 'match' state models the distribution of residues allowed in the column. States 'insert' and 'delete' at each column allow for insertion of one o more residues between that column and the next, or for deleting the consensus residue (see example in Figure 2.8). Profile HMMs are strongly linear, left-right models, unlike the general HMM case.

The probability parameters in a profile HMM are usually converted to additive log-odds scores before aligning and scoring a query sequence (Barrett *et al.*, 1997). The scores for aligning a residue to a profile match state are therefore comparable to the derivation of BLAST or FASTA score: if the probability of the state emitting residue x is p_x, and the expected background frequency of residue x in the sequence database is f_x, the score for residue x at this match state is $log\frac{p_x}{f_x}$ (Eddy, 1998).

CHAPTER 3

Methods

Contents

Overview

This chapter is divided into two main sections: predictor development and laboratory testing. Section 3.1 describes the sequential methods that were used to construct the proposed predictor for NES sequences, it includes a general description of the process as well as a small introduction concerning the terms used.

This part of the chapter includes the links to the respective outcomes of each step, which are presented in Chapter 4.
Section 3.2 forms the second part of the chapter, it describes the experimental procedures that were used to test the predictions obtained, as well as for testing the nuclear export activity of some already known NES-containing proteins from *Arabidopsis thaliana*.

3.1 Development of a prediction tool for NESs

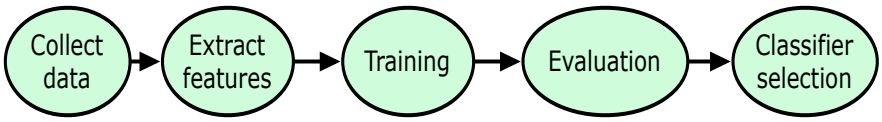


Figure 3.1: Simplified view of the pattern classification process
This scheme shows essential steps and main designations that are used in this section and further chapters.

Figure 3.1 is a simplified representation of the procedures that are described along this section. The raw data was composed of a set of *samples* (protein sequences in this case), which are also called *instances*, *elements* or *examples*. Given this set of samples belonging to different classes, the main goal was to construct a classifier that assigns a class label to new unclassified samples. In this case, each sample belonged to one of two classes: NES or nonNES. The samples labelled as NES were called the *positive set* and the samples belonging to the nonNES class were the *negative set*.

For constructing a classifier, the first task is to identify certain characteristics or properties, called *features* that could be measured to represent each sample. All the properties measured for every sample conform the *feature vector*. Formally, *features* are the individual measurable properties of the samples and a *feature vector* is the m-dimensional vector of numerical features that represent those properties. In a supervised classification task, a sample is composed of a feature vector and a class label. After the *feature* calculation, each sample is represented as a point (*feature vector*) in an m-dimensional feature space, where m is the total number of measured *features*. The samples are then evaluated by different classification algorithms in the training or learning phase, the samples used in this phase constituted the *training set*. After the training, the next phase consists of

evaluating the performance of the trained classifiers using samples not included in the training phase, called *test set*. Finally, one of the classifiers is selected taking into account the results of the evaluation process and the intended function of the classifier.

These concepts can be summarized as follows:

Let

$$S = (s_1, ..., s_n) \tag{3.1}$$

be the data set of n labelled samples. Each sample consists of two parts:

$$s_i = (\mathbf{x}_i, y_i) \tag{3.2}$$

where $\mathbf{x}_i$ is a vector (feature vector) of m dimensions (number of features measured) and y_i is the class label (for example $+1$ and -1, for NES and nonNES respectively).
The goal is to predict a future unlabelled sample $\mathbf{x}_0$ by the prediction rule:

$$P(\mathbf{x}_0|S) = y_o \begin{cases} +1 & \text{NES,} \\ -1 & \text{nonNES} \end{cases} \tag{3.3}$$

built on the observed data set S.

An overview of the classifier construction work-flow followed in this study is presented in Figure 3.2. The consecutive steps are explained in detail in the next subsections. This part of the work was done using the programming language PERL (version 5.8.6) and the statistical computing environment R (version 2.7.2) (R Development Core Team, 2005). The procedures described in the pre-processing, training and evaluation phases, were carried out with the packages *caret* (**c**lassification **a**nd **re**gression **t**raining) (Kuhn, 2008a,b), *randomForest* (Breiman, 2001), *kernlab*, *ipred* (Peters *et al.*, 2002) and *rocr* (Sing *et al.*, 2005), available from the BioConductor repository (Gentleman *et al.*, 2004).

3.1.1 Establishment of data sets

The positive data set $S_{pos}(pos = 1, 2, ..., p)$, contained $p = 107$ experimentally confirmed NES sequences. It included those contained in the NES database already available (75 NESs) (la Cour *et al.*, 2003) together with sequences from

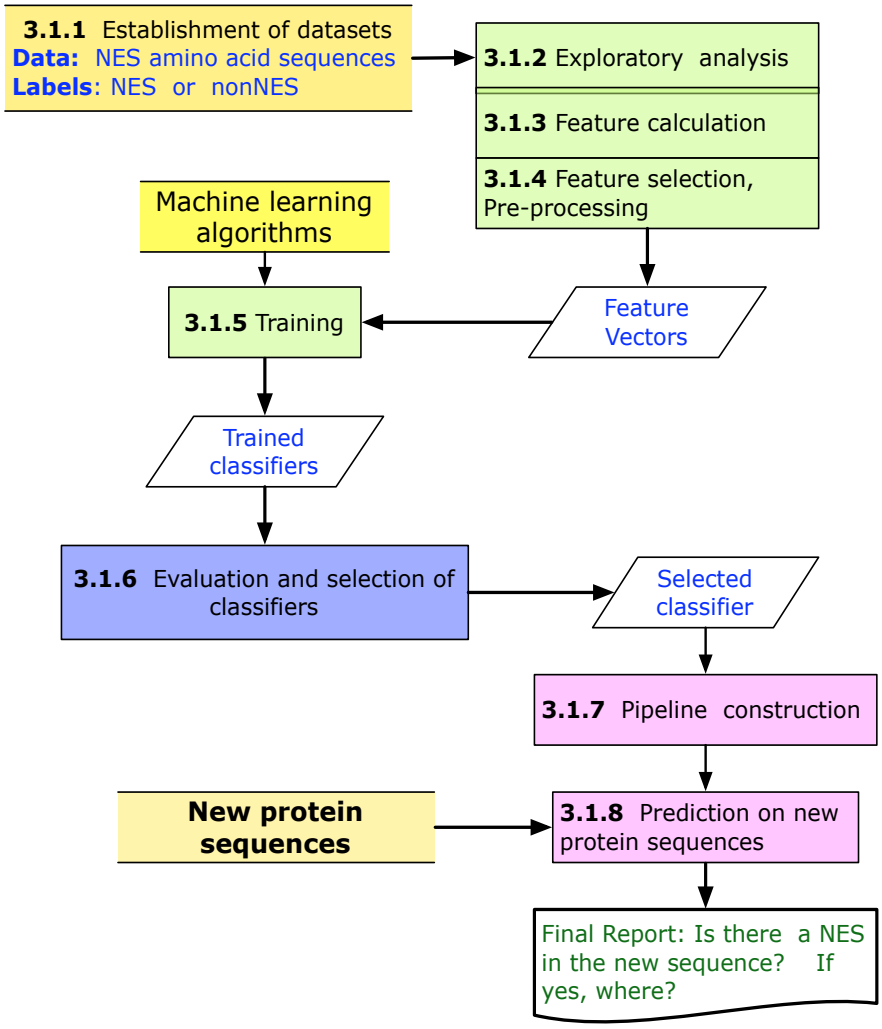


Figure 3.2: Development of a prediction tool for NESs: General flow chart

The first stage corresponded to the selection and analysis of the data as well as the calculation of the elements of the *feature vector*: the amino acid residues from the sequences were translated into numerical values. After that, a preparation step involved a selection or filtering of the features to be used and the splitting of the data. Next came the training of the models in supervised mode where the tuning parameter for each model was done. Next, evaluation using data not included in the training phase was carried out and one of the classifiers was selected. Finally, the selected classifier was integrated in a pipeline and used to classify new protein sequences. The numbers inside the boxes refer to the section in the text where the item is further explained.

Arabidopsis thaliana (32 NESs), which have been experimentally confirmed in the group of Dr. Thomas Merkle but not published yet. The length of the sequences to be 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, which has been shown to be necessary and critical for the interaction of the NES with the Exportin receptor (Mattaj and Englmeier, 1998; Görlich and Kutay, 1999; Kaffman and O'Shea, 1999; Ossareh-Nazari *et al.*, 2001).

The negative data set $S_{neg}(neg = 1, 2, ..., q)$ was conformed with the same proteins included in the positive data set, excluding the region(s) associated with nuclear export activity. To do this, all the regions of the proteins that have some experimental evidence associated with nuclear export activity were excluded and a sliding window was used along the rest of the sequence. In this way, around 10000 sequences were generated from which various subsets of size q were randomly selected. The training of the classifiers was performed using $q = 150$.

3.1.2 Exploratory analysis

The elements of the data sets are amino acid sequences (s), 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, x_i, as a representation of each sequence s_i. This study assessed two kinds of properties: amino acid sequence and physicochemical properties.

The first approach was to examine the amino acid sequence of the NESs contained in the positive data set by constructing sequence logos. Sequence logos are a graphical representation of an amino acid or nucleic acid multiple sequence alignment developed by Schneider and Stephens (1990). Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position (measured in bits of information), while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. The letters of each stack are ordered from most to least frequent, so that one may read the consensus sequence from the tops of the stacks. In this work, the sequence logos were generated using the application WebLogo at `http://weblogo.berkeley.edu/` (Crooks *et al.*, 2004) with multiple sequence alignments of the NES obtained with CLUSTALW (Thompson *et al.*, 1994; Chenna *et al.*, 2003) and QALIGN (Sammeth *et al.*, 2003). The results obtained are presented in Section 4.1.1

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). For this application the default scoring matrix (PAM250) was used. The reported score corresponds to a number between 0 and 1, 1 being the value for identical sequences (self alignment for each sequence) and 0, the value for totally different sequences. Using this approach, a matrix of dimensions $(p+q) \times (p+q)$ was generated, where $(p+q) = n$, n being the total number of sequences (positive and negative data sets). The outcome can be seen in Section 4.1.1.

3.1.3 Feature calculation

Profile Hidden Markov Model (Profile HMM)

The distance matrix obtained with the above approach 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 using the program HMMER ver 2.3.2 from `http://hmmer.janelia.org/`, which is an implementation of profile HMMs for biological sequence analysis. Profile HMMs use *position specific* scores for the amino acid residues and position specific penalties for opening and extending an insertion or deletion. In contrast, traditional pairwise alignment like BLAST (Altschul *et al.*, 1990), FASTA (Lipman *et al.*, 1989), or the Smith/Waterman algorithm (Smith and Waterman, 1981), use *position independent* scoring parameters. Because of this property, a profile HMM captures 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).
The multiple sequence alignment used to construct the profile HMM was obtained with the NES sequences of Arabidopsis using CLUSTALW (Thompson *et al.*, 1994; Chenna *et al.*, 2003) and QALIGN (Sammeth *et al.*, 2003). The `hmmsearch` function from the program HMMER was used to generate a *score* value for every sequence from the positive and negative data set, which was used as one of the elements to include in the feature vector. The distribution of the *score* values obtained is shown in Section 4.1.1.

Amino acid index values

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 not dependable on the order of the amino acid residues 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 at `http://www.genome.ad.jp/dbget/aaindex.html` (Kawashima *et al.*, 1999; Kawashima and Kanehisa, 2000; Kawashima *et al.*, 2008). This database is a flat file that consists of three sections: AAindex1 for the amino acid index values, AAindex2 for the amino acid mutation matrices and AAindex3 for the statistical protein contact potentials.

In this study, the section AAindex1 from the AAindex database was used. The idea was to calculate every *aaindex* value for all the sequences (NES and nonNES) and to use these values as additional features for the classification. There are 544 attributes in the AAindex1 database Version 9.1, therefore one can calculate 544 such features. The aaindex values for each sequence were calculated as the sum of the respective index values of the amino acid residues present in the sequence, the calculation was made as follows:

Each *aaindex* was represented as:

$$AA_j = (AA_{j1}, ..., AA_{j20}) \tag{3.4}$$

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:

$$s = a_1, ...a_l \tag{3.5}$$

the value of the corresponding *aaindex* value $x_{s,j}$ was obtained by adding the individual *aaindex* value of each amino acid:

$$x_{s,j} = \sum_{k=1}^{l} AA_j(a_k) \tag{3.6}$$

Finally, the HMM score (*hmm*) described before was appended to the *aaindex* values to conform the final feature vector for each sequence:

$$\mathbf{x}_{s,545} = hmm_s \tag{3.7}$$

3.1.4 Feature selection and data pre-processing

When considering a large number of features for classification, it is possible that some of these are noisy in nature and irrelevant for prediction. Thus, the use of all predictors to build the classifier some times can suppress or reduce the performance. Selection of a subset of the most informative predictors, called *feature selection*, is commonly addressed in classification problems (Blum and Langley, 1997). One approach to feature selection is removing irrelevant predictors according to some pre-determined criteria. As the sample labels are not taken into account for deciding which predictors are eliminated, this is a kind of *unsupervised* filtering.
In this study, predictors with the highest inter-correlations were eliminated. For that, a correlation matrix from all the predictors was calculated and the function `findCorrelation` from the package *caret* was used to flag predictors for removal. This function uses the following algorithm:

repeat
| Find the pair of predictors with the largest absolute correlation;
| For both predictors, compute the average correlation between each predictor and all the others;
| Flag the predictor with the largest mean absolute correlation for removal;
| Remove this column from the correlation matrix;
until *no correlations are above of a pre-defined threshold*;

The idea of this procedure was to exclude predictors with redundant characteristics so that the pairwise correlation was below a specified threshold. In this case, three correlation threshold values were tested: 0.9, 0.8 and 0.7. After removal of predictors, the three classifiers (k-NN, RF and SVM) were trained several times using the complete data set and the complete and reduced sets of predictors. The classification error was estimated through resampling (10 fold cross-validation and .632+ bootstrap) using the function `errorest` from the package *ipred* on the R platform (Peters *et al.*, 2002). The obtained classification error values were used as criterion to select the number of predictors to use. These results are presented in Section 4.1.2.

Once the final set of predictors was determined, the values were transformed before being used to train the models. The predictor values were `centered` and `scaled` using the predictor means and standard deviations from the training set, this two options provide location and scale transformations of each predictor. For that purpose, the function `preProcess` from *caret* was used.

3.1.5 Training of classifiers

In this study, the training process was carried out using a combination of repetitive hold-out or splitting method 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.

In the hold-out approach, the complete data set was divided into a *training* set and a *test* set using $p = 0.25$, which means that 75% of the data was used for training and the remaining 25% for testing. The test set was used only to evaluate performance and was not included in the classifier training. The partition was carried out with the function `createDataPartition` from the *caret* package. This function creates random splits within each class so that the overall class distribution is preserved.

Three supervised classification algorithms were used: k-Nearest Neighbors, Random Forest and Support Vector Machine. The classifiers were fitted to each of the the training sets by resampling using 10-fold CV, LOOCV and bootstrapping. The optimal classifier in each case was selected based on the highest accuracy value. The tuned parameters for each model are described below and summarized in Table 3.1.

- ***k*-Nearest Neighbors (*k*-NN)**

  In the case of k-NN the fitted parameter was k, which corresponds to the number of neighbors to be considered. In k-NN an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. Then a new vector is classified by assigning the label

which is most frequent among the k training samples nearest to that query point.

- **Random Forest (RF)**

  The fitted parameter for RF was the value *mtry*, which is the number of variables randomly sampled at each node to be considered for splitting. One noticeable gain during the training of this algorithm was the estimation of importance of variables in determining classification. This result is presented in Section 4.1.3.
  The importance of a variable for classification may be due to its (possibly complex) interaction with other variables. Hence, the definition of variable importance is generally problematic. Nonetheless, the importance of a variable is estimated in the RF algorithm by computing the increase in prediction error when the data for that variable is permuted while all others are left unchanged. The percent increase in misclassification rate for a permuted variable does reflect the importance of that variable for the overall classification. To obtain importance estimates for the complete data set, all input variables are consecutively and randomly permuted in the test set.

- **Support Vector Machine (SVM)**

  The SVM algorithm uses a kernel function to map the original feature space to some higher-dimensional feature space where the training set is separable. In this study, the SVM algorithm was trained with the radial kernel function (RBF for Radial Basis Function).
  There are two tuning parameters for SVM: *sigma* and *cost* (C). *Sigma* is a parameter for the kernel function that can be used to expand/contract the distance function and C is the cost parameter. C controls the trade-off between model complexity and proportion of non-separable samples, allowing training errors and forcing rigid margins. It creates a soft margin that permits some misclassifications. Increasing the value of C increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well. For the training of SVM in R, the function `sigest` in the *kernlab* package is used internally during the training to provide a good estimate of the *sigma* parameter, so here only the C parameter needed to be tuned.

Model	Fitted parameter	R Package	R Method
k-NN	k	*caret*	`knn`
RF	*mtry*	*randomForest*	`rf`
SVM	C	*kernlab*	`svmRadial`

Table 3.1: Fitted parameters for the trained classifiers

3.1.6 Performance evaluation criteria

The *test set*, previously described, was used to assess the performance of the classifiers. Based on the class predicted by the trained classifiers for every element of the *test set* and its actual class (the true class is previously known, every element is either NES or nonNES), there are four possible outcomes. If the sample is positive (is a NES) and it is classified as NES, it is counted as *true positive*; if it is classified as nonNES, it is counted as a *false negative*. If the sample is negative and it is classified as nonNES, it is counted as a *true negative*; if it is classified as NES, it is counted as a *false positive*. Based on these possibilities, a two-by-two confusion matrix or contingency table (Table 3.2) was used as reference to calculate the performance metrics (Baldi *et al.*, 2000).

		True Class: NES	True Class: nonNES
Predicted Class	NES	**TP** True Positive	**FP** False Positive
	nonNES	**FN** False Negative	**TN** True Negative

Table 3.2: Confusion Matrix

Based on the notation used in the confusion matrix (Table 3.2), the performance measurements described below were obtained to assess the performance of the classifiers used in this study. The area under the receiver operating characteristic curve (AUC) was also used, it is described in the next subsection. The results concerning the performance metrics and correlation measurements used are presented in Section 4.1.4.

- **Accuracy (ACC)**: Fraction of correctly predicted NES and nonNES sequences in whole data set.

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3.8}$$

- **True positive rate (TPR)**: Fraction of correctly classified NES sequences. It is also referred to as *recall*, *sensitivity* or *hit rate*.

$$TPR = Sensitivity = \frac{TP}{(TP + FN)} \tag{3.9}$$

- **False positive rate (FPR)**: Fraction of misclassified NES sequences. It is also called *false alarm rate*.

$$FPR = \frac{FP}{(FP + TN)} \tag{3.10}$$

- **Specificity**: Fraction of correctly detected nonNES sequences. It is equivalent to the True Negative Rate (TNR).

$$Specificity = TNR = \frac{TN}{(TN + FP)} = 1 - FPR \tag{3.11}$$

- **Precision**: Fraction of correctly predicted NES among the total of samples predicted as NES. It is also called *positive predictive value*.

$$Precision = \frac{TP}{TP + FP} \tag{3.12}$$

- **Matthews correlation coefficient (MCC)**: This parameter corresponds to the statistical Pearson correlation, called MCC since it was first used in Matthews (1975). It assesses the quality of prediction and takes care of unbalanced data (Baldi *et al.*, 2000). The MCC value is always between -1 and 1, a value of -1 indicates the worst possible prediction, 1 is regarded as perfect prediction and 0 indicates a completely random prediction.

$$MCC = \frac{((TP \times TN) - (FN \times FP))}{\sqrt{((TP + FN)(TN + FP)(TP + FP)(TN + FN))}} \tag{3.13}$$

- **F-score**: This value, also called F1-score or F-measure, combines recall (TPR or sensitivity) and precision. It corresponds to the harmonic mean of these two measures and can be interpreted as a weighted average of them, 1 being its best value and 0 the worst.

$$F\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.14}$$

Receiver operating characteristic (ROC) curves

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 *probability* value. Thus, depending on the probability *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 asses 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. To compare the classifiers using the ROC approach, the ROC *convex hull* (ROCCH) method, described by Provost and Fawcett (2001), was used. The ROCCH graphs were constructed with the tool ROCOn version 2.0 at `http://www.cs.bris.ac.uk/Research/MachineLearning/rocon/`. The results for ROC curves and ROCCH graphs are presented in Section 4.1.4.

3.1.7 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 work-flow, in bioinformatics commonly referred to as *pipeline*, which was implemented using PERL and R. For the prediction of NESs in a new proteins, each protein is initially split into overlapping fragments of 10 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 treshold for the class assignation.

3.1.8 Prediction on new protein sequences

Prediction set

A data set containing 33410 protein sequences, obtained from the Arabidopsis Information Resource website (TAIR, `http://wwww.arabidopsis.org`) was used as target (TAIR9 Genome Release, June 2009).

Selection of proteins to be experimentally tested

A group of the proteins predicted as containing NES were selected to be tested in the laboratory to find out if they contain or lack the predicted NES. The proteins were selected on the basis of their Gene Ontologies (GO).
The GO annotations include three main categories: cellular localization, molecular function and cellular process `http://www.geneontology.org/`, (The Gene Ontology Consortium, 2000). Since these three categories are represented as directed acyclic graphs (DAGs) or networks, a child term may have more than one parent term. That means that one protein could be in more than one

sub-category. Therefore, the common sequences among sub-categories were inspected. For that, the GO terms for every predicted protein were extracted from the Arabidopsis GO annotations (Berardini *et al.*, 2004) available at TAIR: http://www.arabidopsis.org/tools/bulk/go. The proteins sharing at least two sub-categories were selected and some of them were tested in the laboratory, according to procedures described in the next section. The steps described until now are summarized in Figure 3.3.

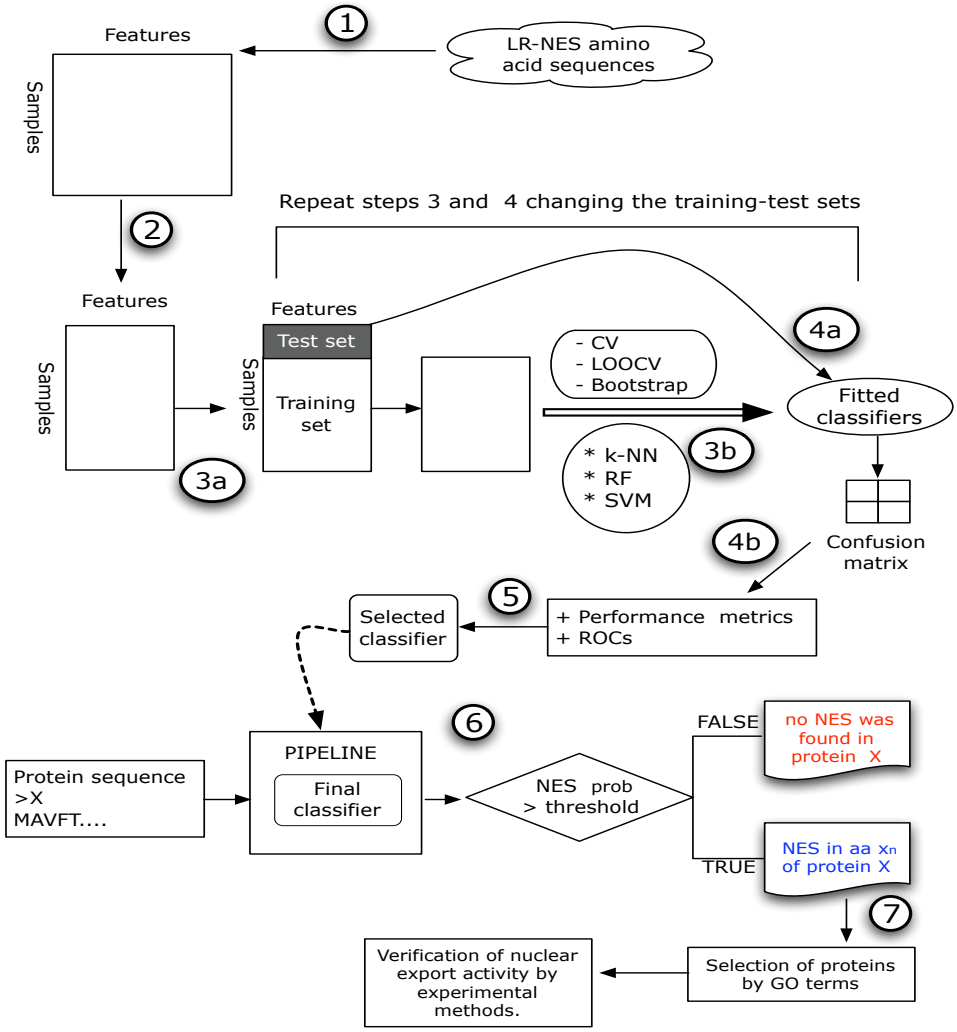


Figure 3.3: Summary of the development of the NESs prediction tool

This figure is intended to summarize and integrate the procedures described until now. The numbers in the figure represent each one of the principal steps of the process: **1**: The amino acid sequences were transformed into numerical features. **2**: The number of features was reduced by eliminating high correlated features. **3a and 3b**: Three algorithms for supervised classification were trained: k-Nearest Neighbor(k-NN), Random Forest (RF) ans Support Vector Machine (SMV). The training was carried out by combining repeated hold-out in the whole data set and resampling (10-fold Cross Validation (CV), Leave One Out Cross Validation (LOOCV) and .632+ bootstrap) only in the respective training set for each hold-out round. **4a and 4b**: The test sets of every hold-out round were used to evaluate the classifiers through performance metrics and receiver operating characteristics (ROC). **5**: A classifier was selected based on the results of the evaluation process. **6**: The selected classifier was integrated in a complete workflow or pipeline, which was used to predict possible NESs in new protein sequences. **7**: A subset of the predicted NES-containing proteins was selected by using Gene Ontology (GO) terms and from that selection, some proteins were experimentally tested by procedures described in Section 3.2.

3.2 Experimental assessment of the nuclear export activity

This section forms the second major part of the chapter. The first two subsections contain the general laboratory methods and the third one the composition of media and reagents.

3.2.1 General methods for DNA manipulation

Routine techniques, such as DNA agarose gel electrophoresis, DNA precipitation, DNA ligation, DNA cleavage with restriction endonucleases and DNA concentration measurement were done according to Sambrook and Russel (Sambrook and Russell, 2001).

Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) was employed to amplify DNA fragments for cloning, for screening of transformed bacterial colonies (colony PCR) and for performing site-directed mutagenesis by overlap-extension PCR, which is explained below.
Standard reaction conditions and an amplification profile are given in Table 3.3. Deployed enzymes were Taq DNA polymerase for colony PCR and a proof reading DNA polymerase (PWO DNA polymerase (Roche) or Phusion High-Fidelity DNA polymerase (Finnzymes, Finland)) for amplifying DNA fragments to be cloned. The amount of DNA used as template varied according to the application, for PCRs from cDNA library, 1 μg was used.

Purification of PCR products

The GFX^{TM} PCR DNA and Gel Band Purification Kit (GE Biosciences) was used for removal of undesired dNTPs, primers and the polymerase from PCR reactions as well as for purification of DNA fragments from agarose gels.

(a) General conditions

Enzyme buffer	1x
dNTPs	0.2 mM
Oligonucleotides	10 pmol of each one
DNA Polymerase	1-2 U
DNA template	50 ng to 1μg

(b) Temperature profile

Initial denaturation	94 °C	2 minutes
Amplification	94 °C	30 seconds
30 cycles	55-65 °C	30 seconds
	72 °C	1 minute/kb
Final extension	72 °C	1 minute

Table 3.3: General conditions for PCR

Site directed mutagenesis by overlap-extension PCR

To test the effect of changing specific amino acid residues on the nuclear export activity of some proteins, a PCR-based site directed mutagenesis was used (Figure 3.4). The process involved two steps: first, two separate PCRs were performed with primers that overlap at the position of the desired mutation. After that, a new PCR was performed using the amplicons from the previous reactions as templates and the external primers to amplify the whole fragment, now with the mutation included. The integrity of the DNA fragment and the presence of the desired mutation were confirmed by sequencing.

3.2.2 Detection of protein-protein interactions by yeast two-hybrid assays

Method overview

The yeast two-hybrid assay (Y2H) is a molecular biology technique used to reveal protein-protein interactions *in vivo* by testing for physical interactions (binding) between two proteins.
The assay is based on the fact that many eukaryotic transcriptional activators consist of two physically separable functional modules: one acts as the DNA-binding

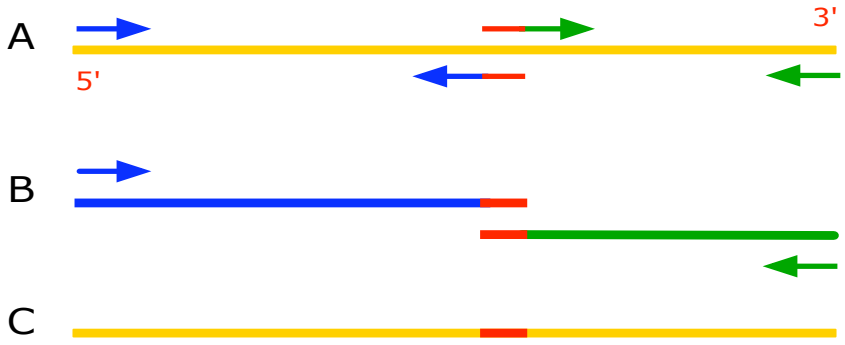


Figure 3.4: Site-directed mutagenesis by overlap-extension PCR
Strategy used to change specific amino acid residues in the NES of selected proteins. **A**: Two separate PCRs with primers that overlap at the position of the desired mutations (red coloured). **B**: New PCR using external primers and the amplicons from the last two reactions as template. **C**: Amplified DNA fragment with the mutations included.

domain (DNA-BD) and the other one functions as the transcriptional activation domain (AD). The DNA-BD localizes the transcription factor to specific sequences present in the upstream regions of genes that are regulated by this factor, on the other side the AD contacts components of the transcription machinery required to initiate transcription. Both domains are required for the activation of gene function, and normally the two domains are part of the same protein. However, it has been shown that a functional activator can be assembled *in vivo* from separate domains of the same or unrelated transcription factors residing in two proteins. The first protein of interest, also called bait protein, is fused to the DNA-BD, while the possible partner, usually called prey protein, is fused to the AD. If the two proteins interact, then both modules (DNA-BD and AD) are in close proximity and will activate the transcription of a reporter gene, as is schematically shown in Figure 3.5.

For the detection of interactions between the receptor Exportin 1 (XPO1a) from *Arabidopsis thaliana* and other proteins, the LexA-based Matchmaker system was used. In this system, the DNA-BD is provided by the prokaryotic protein LexA, and the AD is an 88-residue peptide (B42) from *E. coli* that can activate transcription in yeast. A overview of the procedure is shown in Figure 3.6.

The receptor XPO1a from *Arabidopsis thaliana* (TAIR:AT5G17020), was used as bait protein. Its was fused to the DNA-BD from LexA protein in the vector pGilda, which carries the HIS3 gene for selection in His^- auxotrophic yeast strains. The cDNA of the prey proteins was cloned in the pB42AD vector which contains the

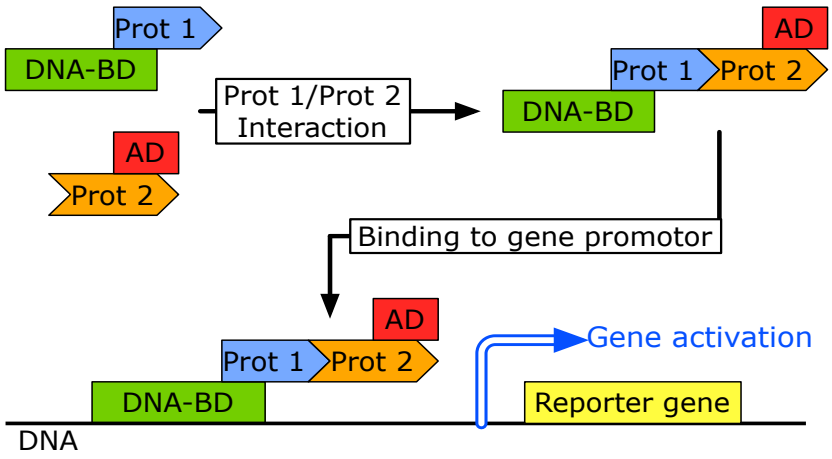


Figure 3.5: principle of the yeast two-hybrid assay

The potential interaction between two proteins (Prot1 and Prot2) can be assayed using this test. In the scheme Prot1 is fused to and DNA-Binding Domain (DNA-BD) whereas Prot2 contains an transcription Activation Domain (AD), if Prot1 and Prot2 interact then, the two domains (DNA-BD and AD) will be in the same unit and could activate the transcription of a reporter gene.

B42-AD and one gene for Tryptophan biosynthesis.

The yeast strain EGY48[p8op-lacZ] was co-transformed with the vectors pGilda and pB42AD, according to the protocol from the provider (Clontech). The vector p8op-lacZ contained in strain EGY48 includes eight LexA-operators, to which the LexA protein can bind and one gene for biosynthesis of uracil that allows the selection in an U^- medium.

Yeast cells that contain the three vectors (p8op-LacZ, pGilda and pB42-AD) are selected in a medium lacking histidine, tryptophan and uracil (SD Gluc HWU^- medium). In addition, since the expression of prey protein-AD in pB42AD and bait protein-LexA in pGilda is under the control of *GAL1* promotors, it is necessary to induce the expression with galactose in a glucose free medium. The interaction between XPO1a and prey proteins activates the transcription of the reporter gene *lacZ* and the yeast cells produce β-galactosidase which can be detected directly on the plate by including the substrate bromo-chloro-indolyl-galactopyranoside (X-gal or BCIG, qualitative assay) or in liquid medium by using the substrate ortho-nitrophenyl-β-D-galactopyranoside (ONPG, quantitative assay).

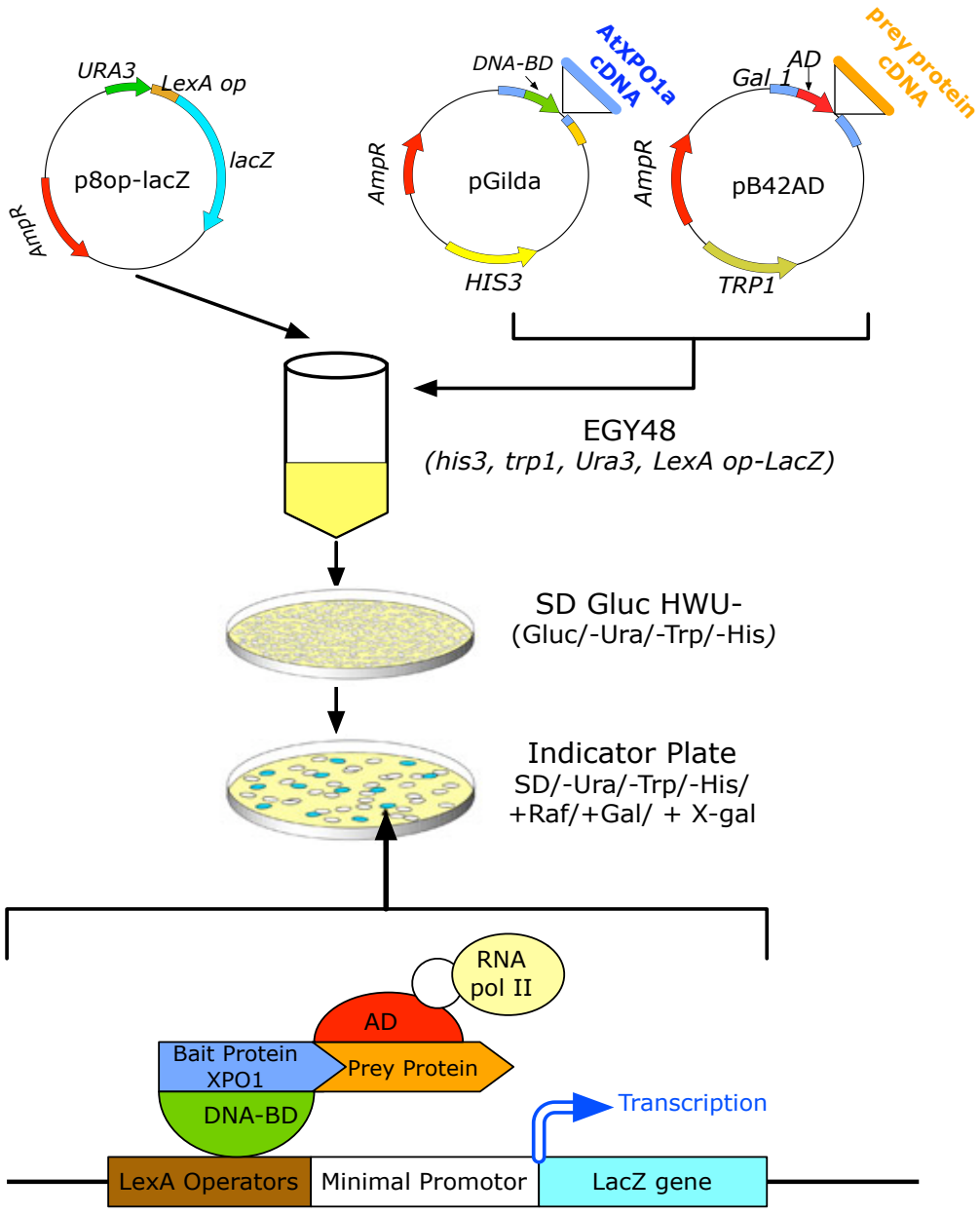


Figure 3.6: Identification of XPO1a-interacting proteins with the Matchmaker LexA Two-Hybrid System

Plasmids containing the *LacZ* reporter gene (p8op-lacZ), the bait XPO1a cDNA (in pGilda), and the cDNA of a putative prey interacting partner (in pB42AD) are propagated in yeast cells (EGY48) on selective media (SD Gluc HWU$^-$). Interaction of the tested proteins activates the reporter gene, *LacZ* whose expression can be detected in indicator plates that include galactose and X-gal.

Qualitative Y2H assay

In a qualitative version of the Y2H assay, the development of blue color indicates the positive interaction between the tested proteins. The intensity of the color is an indicative of the interaction strength. However, the results can only be evaluated in a qualitative way (for example: −, +, +++, to indicate: negative, positive weak and positive strong interaction, respectively).

The procedure was as follows:

- **Growth of yeast cells**
  Yeast cells from strain EGY48[p8op-lacZ] were inoculated in 50-100 mL of SD Gluc Ura$^-$ medium and incubated (overnight, 30 °C, 200 rpm). On the following day, enough yeast cells from the overnight culture were inoculated into 300 mL of YPD medium to produce an OD600 of 0.2 to 0.3 followed by further incubation (3 h, 30 °C, 200 rpm). After that, the cells were harvested (5 min, RT, 1000 *g*), washed once with sterile water and re-suspended in 1.5 mL of 1X TE/LiAc.

- **Transformation**
  For each transformation reaction, 100 μg of salmon sperm DNA, around 1 μg of each plasmid (pGilda and pB42AD), 100 μL of the yeast cells and 600 μL of PEG/TE/LiAc solution were mixed and incubated (30 min, 30 °C, 200 rpm). After that, 70 μL of DMSO were added, mixed and yeast cells were heat shocked (15 min, 42 °C). The cells were then cooled, harvested and re-suspended in 300 μL of 1X TE.

- **Plating**
  100 μL of the above cells were plated on SD HWU$^-$ medium and incubated (3 days, 30 °). On the third day, the yeast colonies were removed from the plates with 0.5-1 mL of sterile water, washed and dotted onto indicator plates (HWU$^-$ plates containing galactose, BU salts and X-gal). Then, the plates were incubated (1-2 days, RT or 30 °) and photographed afterwards.

- **Controls setting**
  Since in Y2H assays false positive results may arise if the test proteins have intrinsic DNA-BD or AD activities, empty vector controls were always included. To test if the prey protein has a DNA-BD, transformations with

empty pGilda vectors were carried out. In the same way, to find out if the bait protein has an AD, the empty pB42AD vector was used. The empty version of both plasmids was tested as well.

Quantitative Y2H assay

A quantitative version of the Y2H assay was used to compare the magnitude of the XPO1a interaction among different proteins or among the wild type and NES-mutated versions of the same protein.
The activity of β-galactosidase can be assayed by measuring hydrolysis of the chromogenic substrate, *o*-nitrophenyl-β-D-galactopyranoside (ONPG) as shown in Figure 3.7 (Miller, 1972). ONPG is colorless, while the product, ortho-nitrophenol (ONP) is yellow (λmax = 420 nm). Therefore, enzyme activity can be measured by the rate of appearance of yellow color using a spectrophotometer.

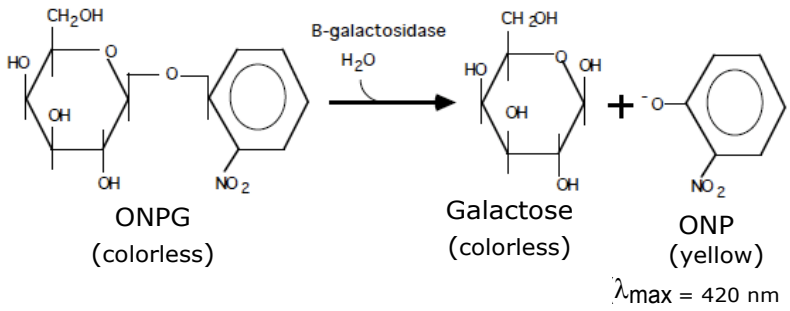


Figure 3.7: Transformation of ONPG into ONP by β-galactosidase

The procedure was as follows:

- **Galactose induction**
  For every interaction to test, three independent cultures were inoculated in 2 mL of SD Gluc HWU$^-$ (2 days, 30 °C, 200 rpm). Afterwards the galactose induction took place. For that, 0.5 mL of every culture was transferred to 5 mL of SD Gluc HWU$^-$ gal raf medium and incubated (4 h, 30 °C, 200 rpm). The growth was stopped by placing the cultures on ice, the cells were collected by centrifugation (10 min, RT, 1000 *g*), washed and re-suspended in 0.5 mL of Z buffer. At this point, the OD_{600} was measured and 3 aliquots of 100 μL for every initial culture were frozen in liquid nitrogen.

- **ONPG assay**
  The frozen cells were thawed (2 min, 37 °C) and frozen again, this procedure was repeated three times. Then, 700 μL of Z buffer + βME were added, followed by 160 μL of ONPG solution, the same was done for the blank tubes (3 tubes with 100 μL of Z buffer). Next, the samples were incubated for exactly 1h at 30 °C and the reactions were stopped by adding 400 μL of 1M Na_2CO_3. After centrifugation (10 min, 14000 rpm), 200 μL of the supernatant were transferred to a 98 well plate and the OD_{420} value was measured in an ELISA reader. The results were expressed as *Miller units*, calculated as:

$$Miller\ units = 1000 \times \frac{Abs_{420}}{OD_{600} \times mL \times min}$$

3.2.3 Reagents and media composition

To support traceability of the experimental assessments, Table 3.4 lists the composition of reagents and media used for the procedures during the laboratory experiments.

Reagent/Medium	**Composition**	
YPD medium	10 g Bacto pepton	
	5 g Yeast extract	
	10 g Glucose	
	ddH_2O to 500 mL	
Dropout HWU^-	L-Isoleucine	300 mg/L
	L-Valine	1500 mg/L
	L-Alanine	200 mg/L
	L-Arginine HCl	200 mg/L
	L-Leucine	1000 mg/L
	L-Lysine HCl	300 mg/L
	L-Methionine	200 mg/L
	L-Phenylalanine	500 mg/L
	L-Threonine	2000 mg/L
	L-Tyrosine	300 mg/L

continued on next page...

... continued from previous page

Reagent/Medium	**Composition**
SD Gluc HWU$^-$ medium	3.35 g YNB 10 g Glucose 325 mg *dropout* HWU$^-$ ddH_2O to 500 mL
SD Gluc HWU$^-$ plates	2.68 g YNB 8 g Glucose 7.2 g Bacto agar 260 mg *dropout* HWU$^-$
SD Gluc U$^-$ medium	3.35 Yeast nitrogen base (YNB) 10 g Glucose 325 mg *dropout* HWU$^-$ ddH_2O to 495 mL After autoclaving, add: 10 mg Tryptophan 10 mg Hystidine For plates add 1.8% Bacto agar
SD Gal HWU$^-$ plates	2.80 g YNB 7.2 g Bacto agar 260 mg *dropout* HWU$^-$ 325 mL H_2O After autoclaving, add: 40 mL 20% Galactose 20 mL 20x BU salts 0.65 mL 50 mg/mL X-Gal
20x BU salts	30 g NaH_2PO_4 x $2H_2O$ 70 g Na_2HPO_4 pH 7.0 ddH_2O to 1000 mL

continued on next page...

... continued from previous page

Reagent/Medium	**Composition**
Gal Raf HWU^- medium	3.35 g YNB 275 mg *dropout* HWU^- 5 g Raffinose 445 mL H_2O After autoclaving, add: 50 mL 20% Galactose
10x TE buffer	0.1 M Tris HCl (pH 7.5) 10 mM EDTA
10x Li-Acetate	1 M Li-Acetate pH 7.5 adjusted with acetic acid
PEG/Li-Acetate	40 % (w/v) PEG 1x TE buffer 1x Li-Acetate
Z buffer	40 mM NaH_2PO_4 x H_2O 60 mM Na_2HPO_4 x $2H_2O$ 10 mM KCl 1 mM $MgSO_4$ x $7H_2O$ pH 7.0
Z buffer + BME	0.27 mL β-Mercaptoetanol Z Buffer to 100 mL 10 mM EDTA
ONPG	4 mg/mL ONPG in Z buffer
Na_2CO_3	1 mM Na_2CO_3 in H_2O

Table 3.4: Reagents and media composition

CHAPTER 4

Results

Contents

Overview

In this chapter the outcomes of the steps described previously in Chapter 3 are presented in detail. It is divided in two main parts, the first one covers the development of the prediction tool for NESs and the second is dedicated to the results obtained in the experimental part of this work.

Section 4.1 starts with the preliminary analysis of the amino acid sequences that were used and continues with the calculation of features and the tuning and training of the three tested classifiers (k-NN, RF and SVM). After that, one of the

pivotal parts of the first section, the evaluation of the classifiers and posterior selection of one of these is presented. The section ends presenting the outcomes of using the selected classifier to predict which proteins encoded in the complete genome of Arabidopsis could contain NESs. The second part of the chapter, i.e. Section 4.2, highlights the experimental verification of the nuclear export activity in a group of proteins selected from the set of predicted as NES-containing proteins in Arabidopsis. It also includes the assessment of the nuclear export activity in some proteins already known to contain NESs, to explore the influence of amino acid residues inside and outside the NES on the Exportin 1 receptor binding activity.
In this chapter, the use of "nuclear export signal" and its acronym "NES" refers always to the leucine-rich NES.

4.1 Development of a prediction tool for NESs

4.1.1 Exploratory analysis

Comparison of NESs from Arabidopsis and from other organisms

As introduced in Chapter 1, one of the main motivations of this work was to develop a prediction tool to identify NESs in proteins of Arabidopsis. The first step followed in that direction was to explore if there are differences in the amino acid sequences of NESs in proteins of Arabidopsis and those that were used to construct the prediction tool already available (la Cour *et al.*, 2004), which are from proteins mainly from virus, yeast and human. One way to compare these two groups of NESs was to construct the sequence logos shown in Figure 4.1.
Figure 4.1 indicates that there are some differences between NES sequences from Arabidopsis proteins (Figure 4.1A) and NESs from virus, yeast and humans (Figure 4.1B). These dissimilarities are mainly in the identity and degree of conservation of the hydrophobic residues as well as in the number of amino acid residues between them.

Comparison of NES sequences to nonNES sequences

In this work, the goal was to separate NES (*positive*) sequences from nonNES (*negative*) sequences. One of the most important points when developing a classi-

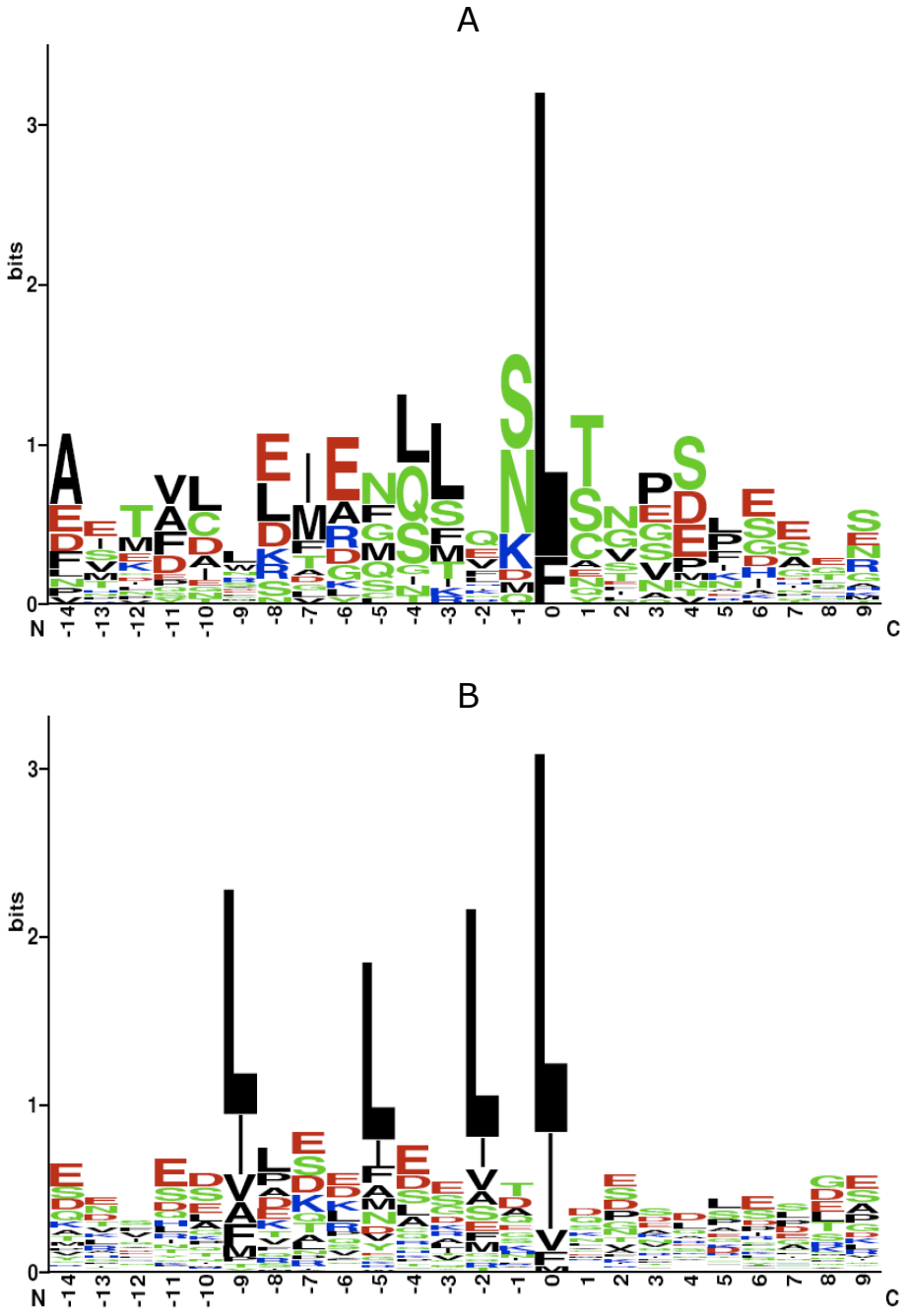


Figure 4.1: Sequence logos for nuclear export signals (NESs)

These logos are a graphical representation of the sequence conservation of amino acid residues in NESs, they were created from multiple sequence alignments of NESs in proteins of, **A**: Arabidopsis and, **B**: other organisms, mainly virus, yeast and human (la Cour *et al.*, 2003). This kind of representation shows how well the residues are conserved at each position: the fewer the number of residues, the higher the letters are, because the conservation is better at that position. Different residues at the same position are scaled according to their frequency. The amino acid residues are coloured according to their polarity: *black*: non polar (hydrophobic), *green*: polar without charge, *red*: acid and *blue*: basic.

fication 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). To see if there were differences between the classes NES and nonNES regarding this parameter, the distance matrix shown in Figure 4.2 was constructed by comparing all sequences to each other. In this matrix the intensity of the color is an indicator of the degree of similarity of each pair of sequences.

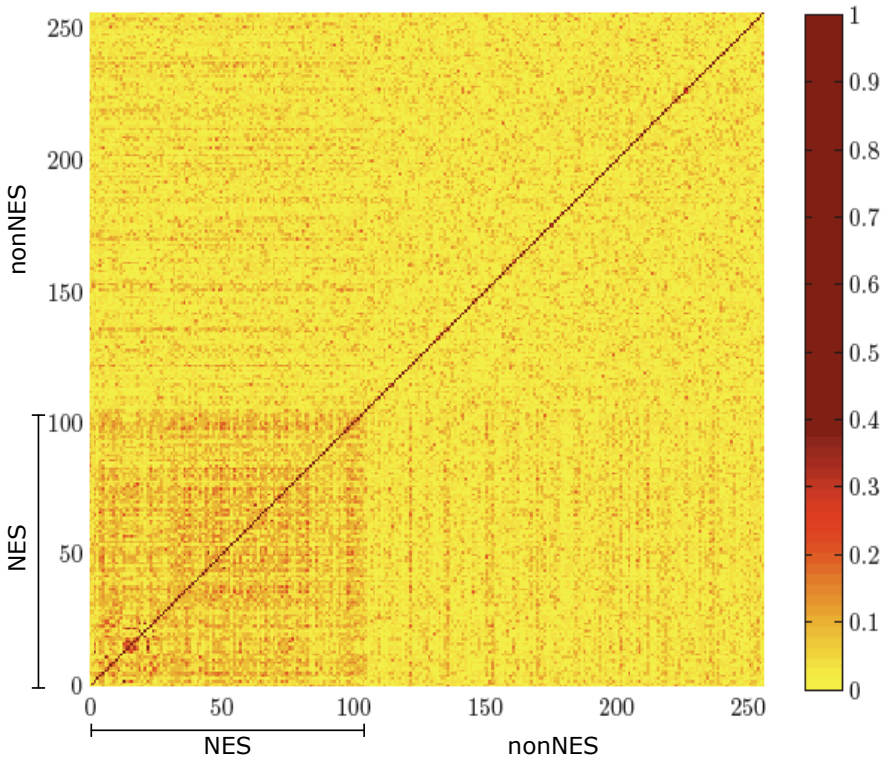


Figure 4.2: Similarity matrix from positive(NES) and negative(nonNES) sequences
This distance matrix shows the similarity between the amino acid sequences labelled as NES or nonNES. Every point in the matrix corresponds to a similarity score whose value varies between $[0, 1]$. This value is represented with the color bar code shown in the right side of the figure. The similarity score used to construct this matrix, corresponds to the identity value obtained from aligning each pair of amino acid sequences with the program ALIGN. According to that, the intensity of the color is directly related with the strength of the similarity and hence the central diagonal line represents the self-alignment of each sequence. The region 0...107 contains the NES or positive sequences, and 108...257 the nonNES or negative sequences.

In Figure 4.2, 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 that 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.

Extraction of features

According to the results shown in Figure 4.2, the NES sequences are different from the nonNES sequences concerning the amino acid identity and order. This property had to be expressed in a numerical value. For that purpose, a profile HMM for NESs was constructed as described in Chapter 3. Every sequence was then compared to this profile HMM and the score produced was used as the mentioned numerical value.

The next question to be answered was if the constructed profile HMM alone could separate the NES class from the nonNES class. This point can be addressed by looking at Figure 4.3, which shows the distribution of the *score* values obtained for all the sequences in both classes. Although there are differences between the HMM score values obtained for NES and those for nonNES sequences, no single HMM score value could be used to unambiguously discriminate between the two classes. Because of that, the profile HMM score was used as one of the elements of the feature vector but in addition, some extra properties namely the amino acid index values (*aaindex*) were included.

4.1.2 Tuning and training of classifiers

For every amino acid sequence labelled as NES or nonNES, 545 properties or features were calculated. One of these features corresponds to the profile HMM score and the other 544 are *aaindex* values. Given this high number of features, it was necessary to eliminate some of them. The criteria used for that was the correlation among them, since features that are highly correlated are sometimes non-informative and instead of improving the classification could increase the noise and also the time necessary for the training process.

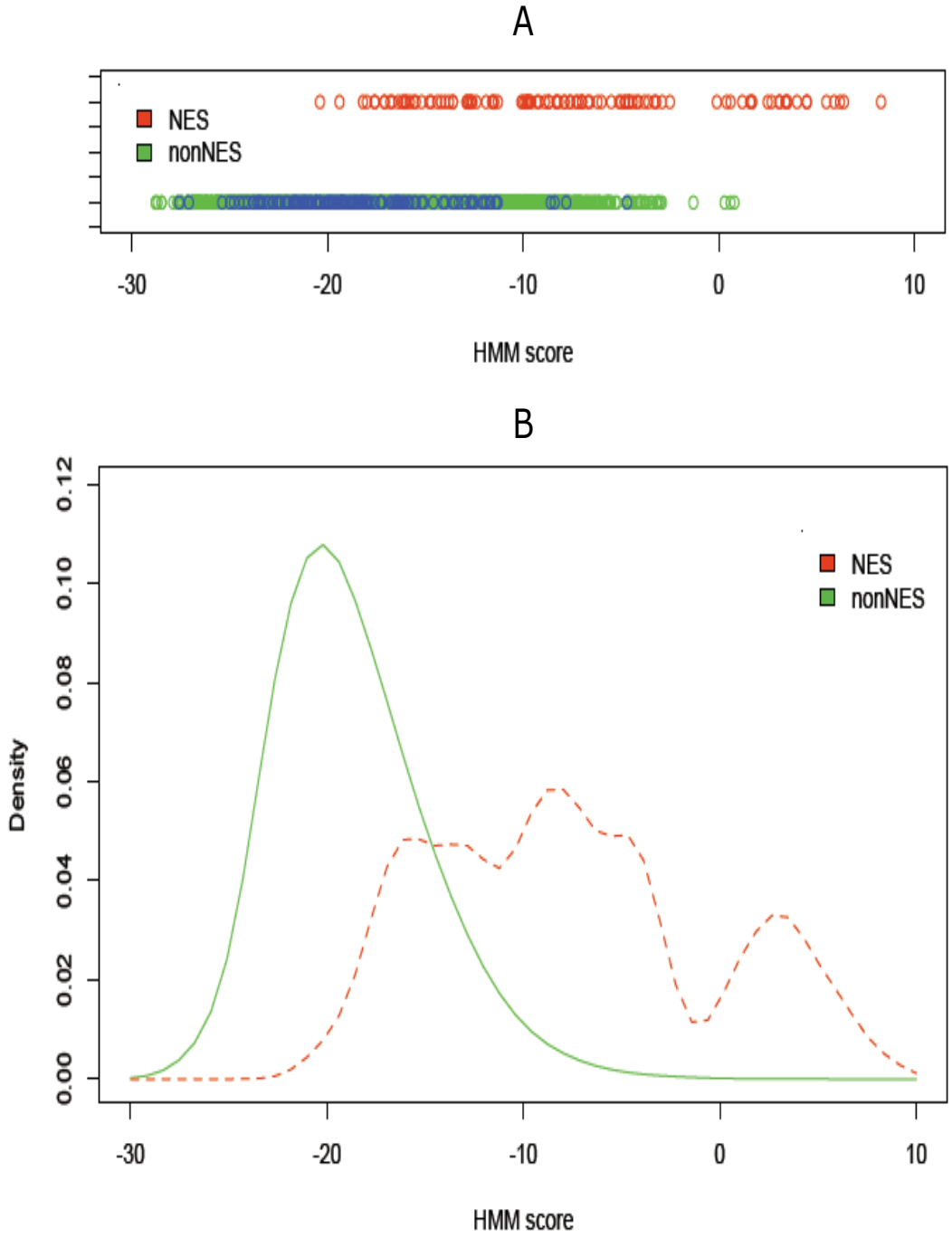


Figure 4.3: Distribution of the HMM score values for NES and nonNES sequences
Two ways of presenting the values of HMM score associated to each sequence. **A**: Each circle represents one sequence, belonging to the NES (red color) or nonNES (green and blue color) class. In the group of nonNES, the green color indicates **all** the negative sequences obtained (around 10000), whereas blue, represents the sequences that were randomly selected to conform the negative set (150). **B**: Each curve corresponds to the density distribution for each class (NES in red and **all** nonNES in green) estimated with the package *sm* in R.

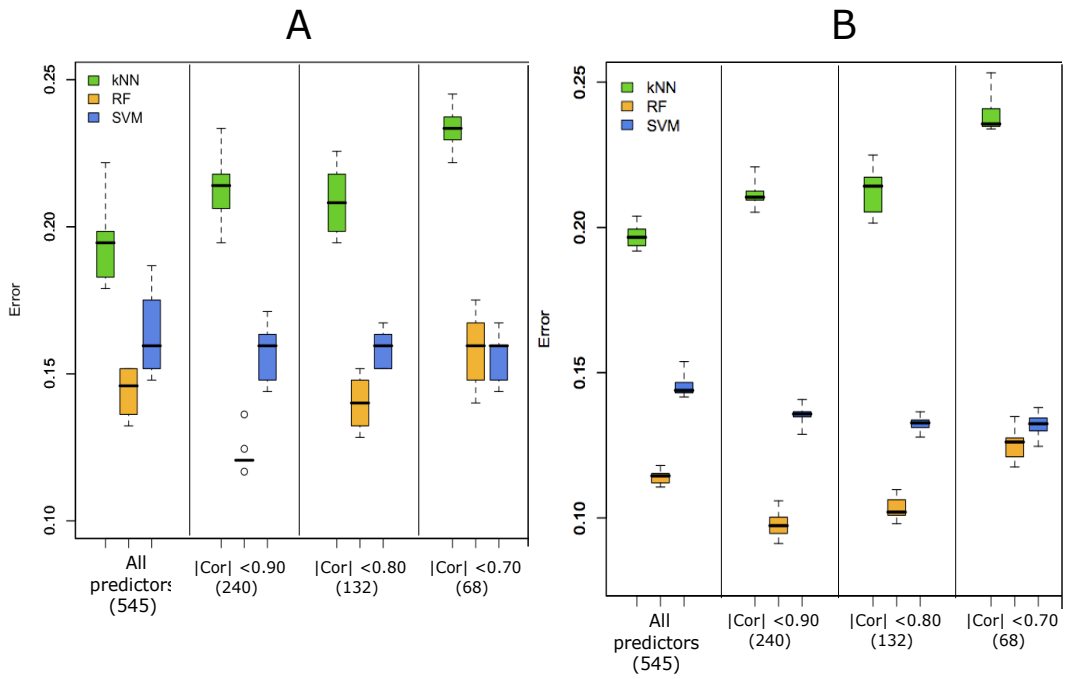


Figure 4.4: Influence of the number of features on the classification error

Boxplots showing the distribution of classification error values obtained by 10-fold cross-validation (**A**) and .632+ boostrap (**B**), for the three classifiers: k-Nearest Neighbor (k-NN), Random Forest (RF) and Support Vector Machine (SVM) with different number of predictors. The number of predictors used in each situation is shown in parenthesis. In the first scenario, *All features* were used and in the other three, a group of predictors with correlations above a certain threshold were eliminated. The features used in each situation are indicated as $|Cor| < 0.9$, $|Cor| < 0.8$ *and* $|Cor| < 0.7$, where the correlation threshold was 0.9, 0.8 and 0.7 respectively. In this graph, the whiskers extend from -1.5 to +1.5 of interquartile range (IQR), the dark horizontal line inside each box indicates the median of the sample (50^{th} percentile) and the limits of the box represent the lower and upper quartiles (25^{th} and 75^{th} percentiles) respectively. The outliers, if any, are represented as individual circles outside the whiskers. The values were calculated using the function `errorest` from the package *ipred* (Peters *et al.*, 2002) under the R platform .

For deciding which features would be used in the final classifier, the three classifiers were trained with all of them (545) and with features with correlation less than 0.90 (240), 0.80 (132) and 0.70 (68) and the classification error was evaluated for each feature set. The training was accomplished by resampling using 10-fold cross-validation and .632+ bootstrap. Figure 4.4 shows the classification error for the three classifiers (k-NN, RF and SVM) with the four features sets described before and the two resampling methods used. Each classifier reacted in a different way to the elimination of highly correlated features. The k-NN classifier produced the highest values for classification error and the error became even bigger for this

classifier when more features were excluded. In the case of SVM, the classification error values were much smaller than using k-NN and remained in the same range with all features sets. In the case of RF, there was a diminution of the error value after removal of the first group of features (correlation threshold 0.90) but further eliminations increased the error value again. Beyond that, RF produces the lowest classification error values, excepting the case of the smallest feature set where the error values for RF and SVM are about equal. The observed behaviour did not depend on the resampling procedure used since the tendency was similar when using 10-fold CV or .632+ bootstrap. As a whole, the values for classification error under the tested conditions (different number of predictors) were lightly lower and showed less variance, by using .632+ bootstrap compared to 10-fold CV. Based on these results, all the features were used in the case of k-NN and 240 (with correlation < 0.90) in the case of RF and SVM.

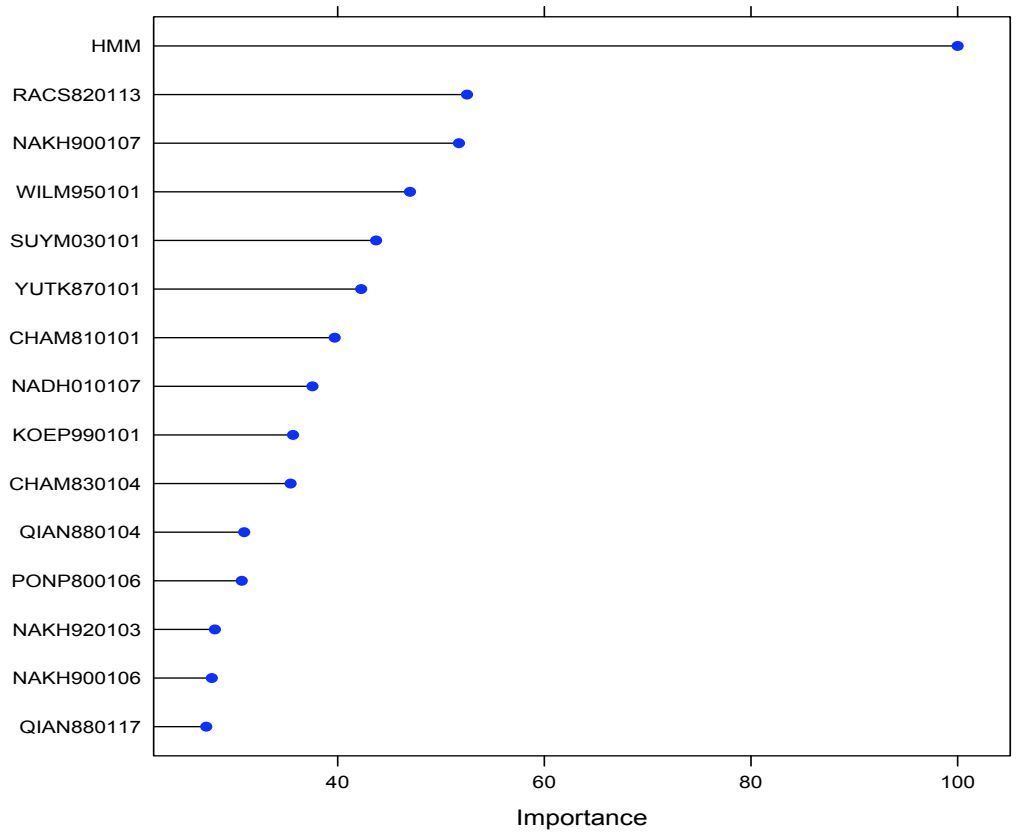


Figure 4.5: Variable importance estimated with the Random Forest(RF) algorithm
The 15 most important variables or features estimated by Random Forest. The most important variable is set to value 100 and the others are scaled accordingly. The name of the respective variable is given in the left side, *HMM* corresponds to the profile-HMM score and the other are designations for different *aaindex* values.

4.1.3 Variable importance

As part of the algorithm, RF returns measures of variable importance. This is done by computing the increase in prediction error when the data for that variable is permuted while all others are left unchanged. The percentage increase in misclassification rate for a permuted variable reflects the importance of that variable for the overall classification.

The variable importance measures for the complete data set (without partition) is shown in Figure 4.5. It is noticeable that the score from the profile HMM (HMM in Figure 4.5) had the highest relevance and also, *aaindex* values related with hydrophobicity measures (for example NAKH900107, WILM950101, YUTK870101, PONP800106 and NAKH900106) were included in the selection.

4.1.4 Classifier assessment and selection

Performance assessment

To obtain a preliminary idea of the classifiers performance, the accuracy was measured in the training phase (using resampling across the training set) and in the test phase (using the test set). The values obtained for the three classifiers are shown in Figure 4.6. It can be seen that the accuracy values for the three classifiers differ between training phase and test phase. In all cases, the values obtained were higher when the training set and resampling was used than when the test set was used. If only the resampling scheme were used, the RF classifier would have an accuracy value of 100% in all the cases, which is an extremely optimistic scenario. In the same way, for the other classifiers the accuracy values would be over-estimated if the values reported were only produced by resampling. Taking in mind this consideration, the evaluation of the classifiers was carried out only with the test set. The partition of the whole data set into training and test sets (hold-out method) was carried out more than once and the complete procedure of training and testing was repeated independently as described in Subsection 3.1.5.

The performance metrics accuracy, sensitivity, specificity, precision, false positive rate (FPR) and classification error (with respect to the test-set) obtained for the three classifiers are presented in Figure 4.7. Since three resampling schemes were used for the training with four different partitions (training-set test-set), twelve values for every performance metric were obtained for each classifier.

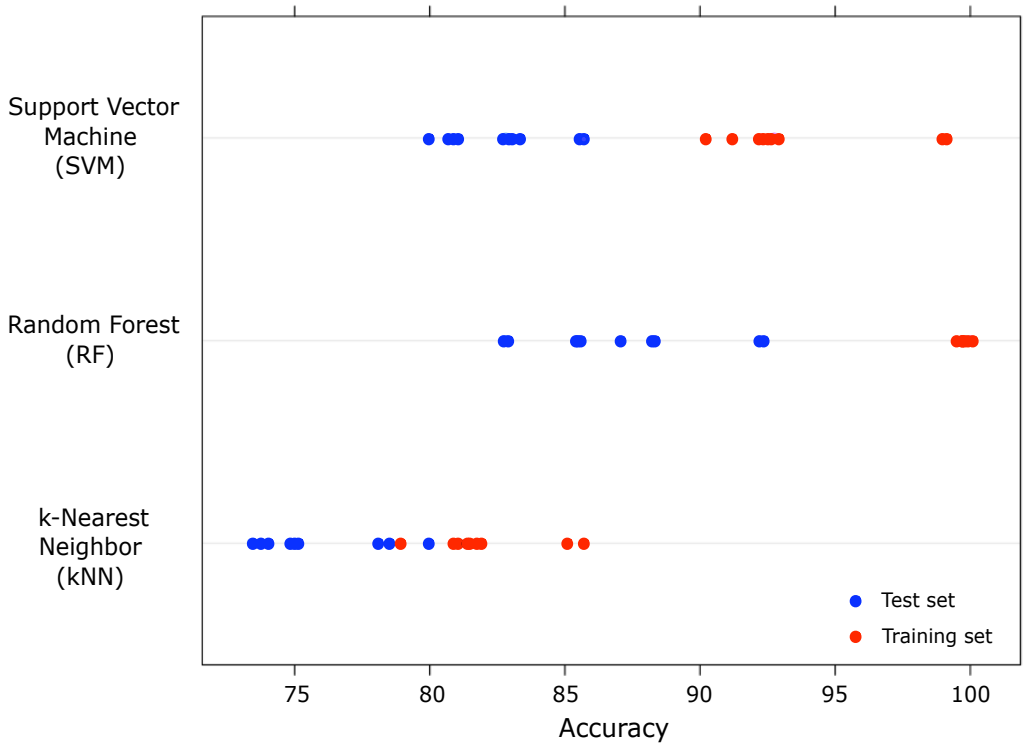


Figure 4.6: Accuracy compared: training and test sets

The accuracy values (in percentage) for the three classifiers were estimated using resampling within the training set (red circles) or using only the predictions obtained for the test set (blue circles). For each classifier twelve measures were obtained, corresponding to three resampling schemes and four different partitions of the whole data into training and test sets.

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 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.

The boxplots presented in Figure 4.7, allow not only comparing the values obtained for the three classifiers, but also heeding the degree of dispersion (spread) and skewness in the data, and identifying outliers. It is noticeable, for example that k-NN exhibits a higher degree of dispersion of the data in specificity and false positive rate, compared to RF and SVM.

In addition to the performance measures shown in Figure 4.7, two correlation measures were also evaluated: Matthews correlation coefficient (MCC) and F-score.

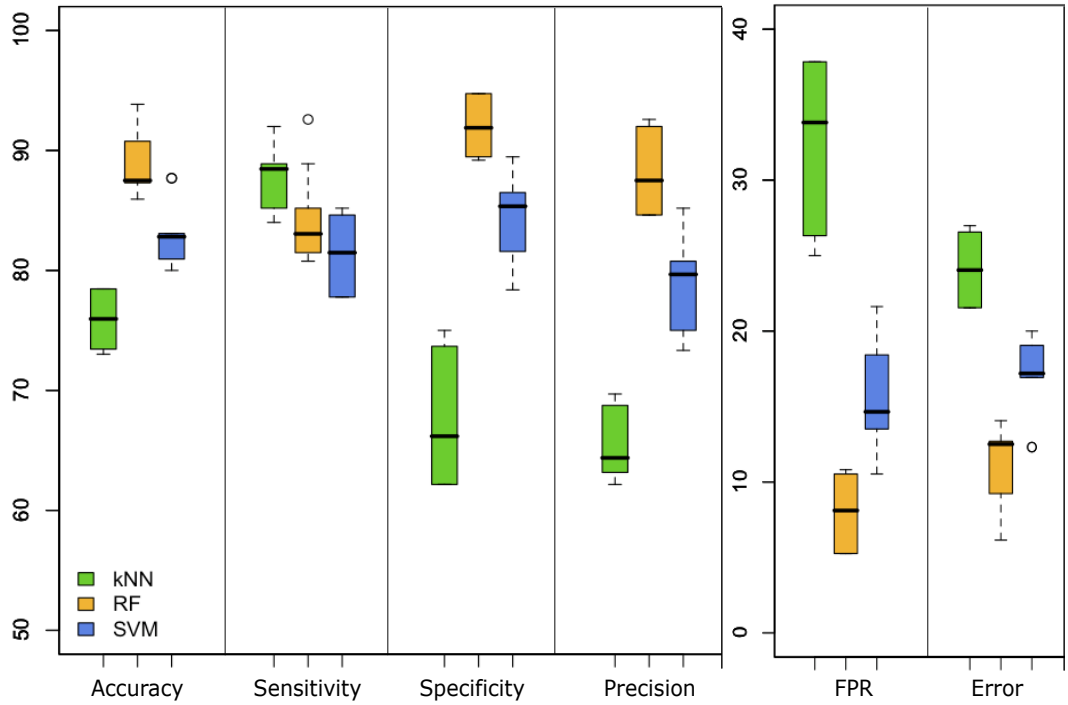


Figure 4.7: Comparative performance of the trained classifiers (I)

Boxplot graph showing the results for some of the performance metrics (in percentage) used to evaluate the classifiers k-Nearest Neighbor(k-NN), Random Forest (RF) and Support Vector Machine (SVM). Every measure corresponds to twelve independent experiments using different combinations of training and test sets. For the values of *accuracy*, *sensitibity*, *specificity* and *precision*, the higher the value, the better the classifier. By the contrast, for *false positive rate (FPR)* and *error* lower values indicate better performance.
The whiskers extend from -1.5 to +1.5 of interquartile range (IQR), the dark horizontal line inside each box indicates the median of the sample (50^{th} percentile) and the limits of the box represent the lower and upper quartiles (25^{th} and 75^{th} percentiles) respectively. The outliers, if any, are represented as individual circles outside the whiskers.

Classifier	MCC	F-score
k-NN	0.55 (0.03)	0.75 (0.01)
RF	0.77 (0.06)	0.86 (0.03)
SVM	0.66 (0.05)	0.80 (0.03)

Table 4.1: Comparative performance (II): correlation measures

Average values for two correlations measures: Matthews correlation coefficient (MCC) and F-score, obtained for the classifiers k-Nearest Neighbor (k-NN), Random Forest (RF) and Support Vector Machine (SVM). The standard deviation in each case is indicated in parenthesis.

The respective values obtained are presented in Table 4.1. For MCC a value of "1" is regarded as perfect prediction and "0" indicates a completely random prediction. From this point of view, the three classifiers predict better than random being RF and SVM slightly superior to k-NN. Similar observations were obtained regarding the F-score values.

Receiver operating characteristics (ROC) curves

The outcome of the classification process with the three trained classifiers 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 three trained classifiers are shown in Figure 4.8, where the indicator *area under the curve (AUC)* is also included.
According to the ROCs shown in Figure 4.8 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. The result is shown in Figure 4.9 where each point corresponds to one classifier. In this approach, the points that are closer to the *convex hull* represent the optimal classifiers under different scenarios. For example, in Figure 4.9 A and B, the blue lines indicate two possible class distributions situations: **A**, when the data set contains the same proportion of positives and negative samples and **B**, when the data set contains 20% of positive samples and 80% of negative samples. According to this approach, RF would be the best classifier under the two circumstances considered. An interesting observation is that k-NN would be a good classifier if the data set would contain a much higher proportion of positive samples in comparison to the negatives.

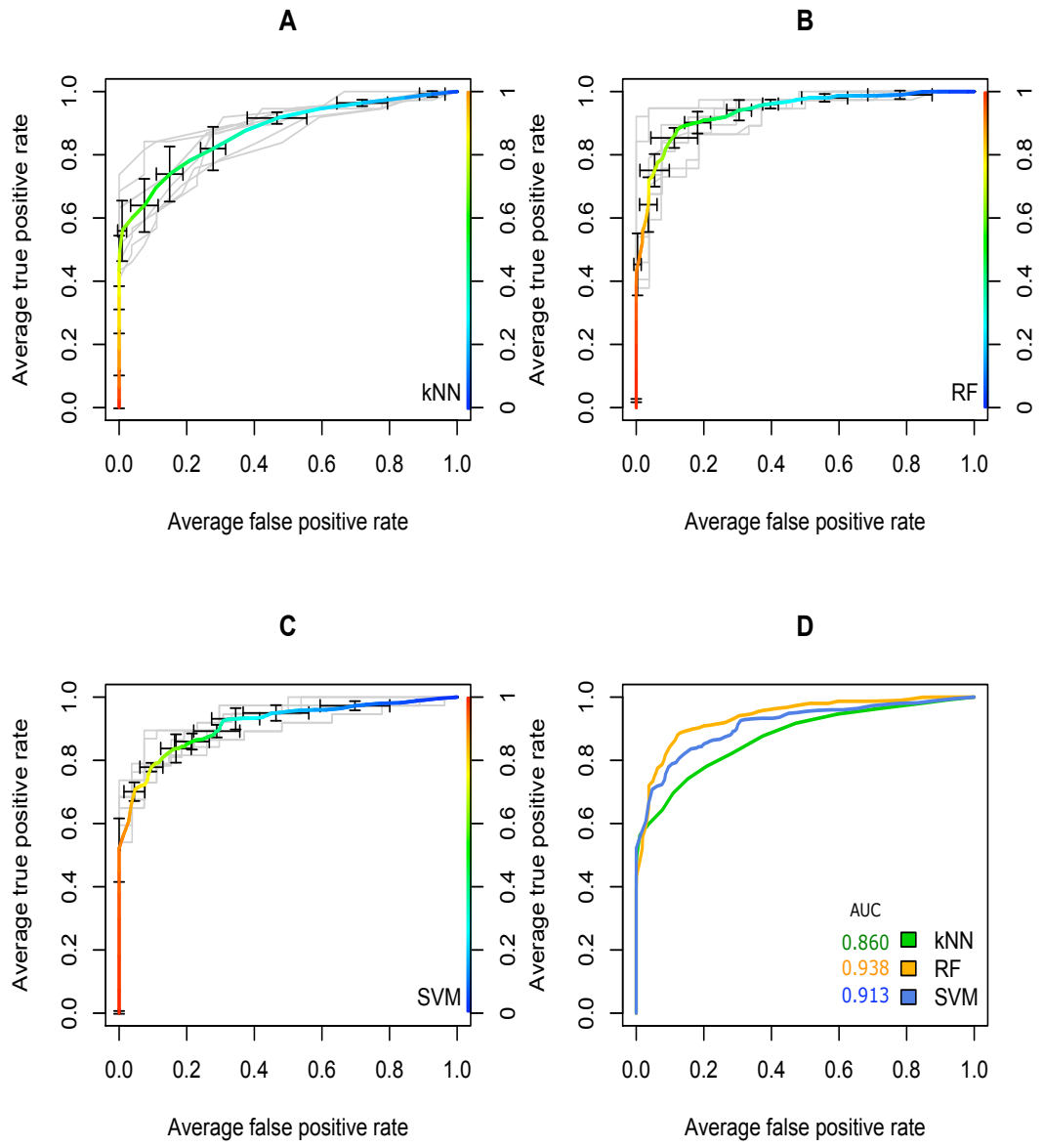


Figure 4.8: Comparative performance (III): receiver operating characteristics (ROC) curves

Receiver operating characteristics (ROC) curves of the three trained classifiers: **A**: k-Nearest Neighbor(*k-NN*), **B**: Random Forest (*RF*), **C**: Support Vector Machine (*SVM*). In **A-C**, the average curves are shown in color whereas the individual curves that were used to calculate the average are shown in grey, the horizontal and vertical lines along the curve correspond to the respective standard deviation. The differential color used for the average curves corresponds to a probability *cutoff* value, the scale is on the right side of each curve.
D shows the average curves for the three classifiers in the same graph and the average area under the curve (AUC). Each color corresponds to a classifier according to the scheme shown in the right down corner.

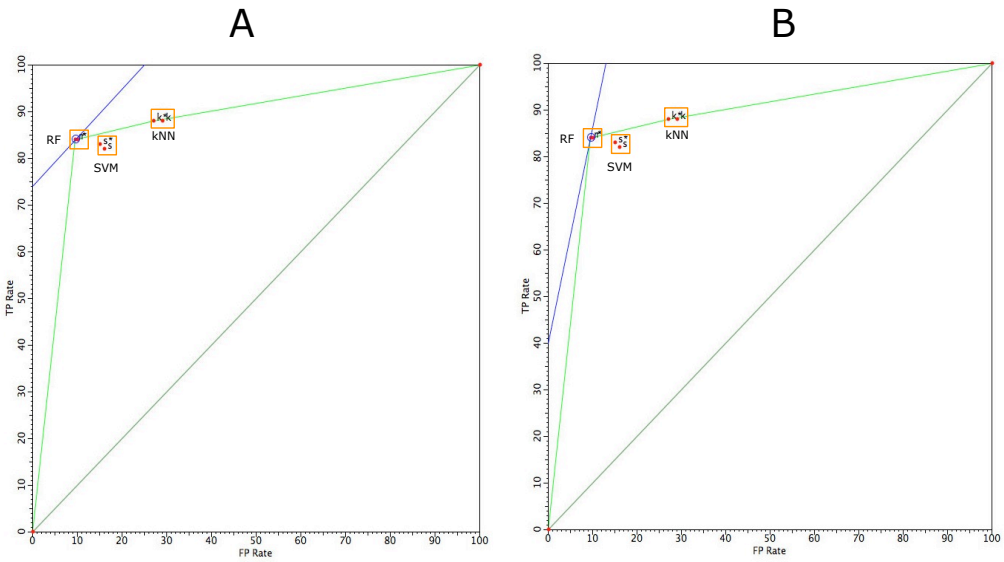


Figure 4.9: Comparative performance (IV): receiver operating characteristics convex hull (ROCCH)

The Random Forest (RF) classifier performed better than Support Vector Machine (SVM) and k-Nearest Neighbor(k-NN) and was the optimal classifier according to the receiver operating characteristics convex hull (ROCCH) approach. In this plot each red point represents one classifier whose values of true positive rate (TPR) and true negative rate (TNR) are taken on a specific probability *cutoff* value (0.5 in this case). The convex hull curve is shown in green, the potential optimal classifiers are those that lie on to that line. The blue lines show the potential optimal classifier under two different conditions of class distributions (proportion of positive and negative samples): **A**: Equal class distribution and **B**: Unbalanced class distribution, 20% of positive samples and 80% of negatives samples.

According to the results of the performance measurements and ROC curves, RF was selected as the best method to classify NESs. Therefore, the next step was to use it to predict NESs in new protein sequences. One of the intended uses of this classifier was to predict possible NES-containing proteins in the whole available sequences of Arabidopsis. For an application like this, it is more 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 positives are missed. One way to achieve that is to adjust the probability *cutoff* value that the classifier uses to assign the class label to new samples.

Fig 4.10 shows the variation of the RF classifier performance across different probability *cutoff* values. It can be 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 to much in sensitivity.

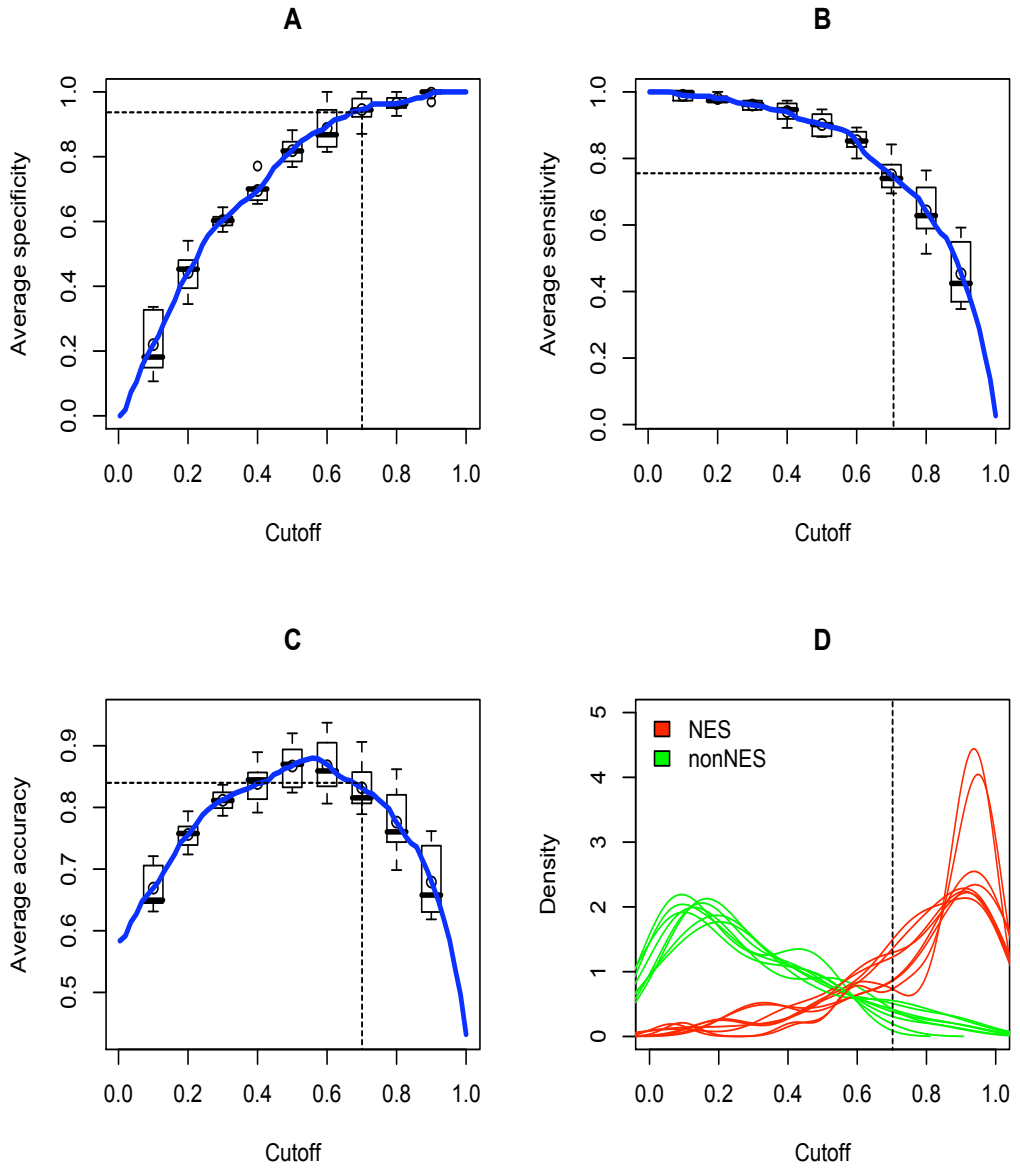


Figure 4.10: Probability *cutoff* value selection for the Random Forest classifier
Plots showing the relationship between **A**: Specificity, **B**: Sensitivity, **C**: Accuracy and **D**: class separation, and the probability *cutoff* value for the Random Forest(RF) classifier. The value indicated with a dashed line (0.7) was used as the probability *cutoff* value for screening the whole data set of protein sequences of Arabidopsis with the RF classifier.

4.1.5 Classification of new samples

The pipeline constructed was used to classify 33410 protein sequences, obtained from the Arabidopsis information resource website TAIR, using TAIR9 Genome Release, June 2009 (http://www.arabidopsis.org). From this data set 5156 sequences corresponding to individual loci were predicted as NES-containing proteins. The length of the sequences classified as NES was in a range from 6 to 20 amino acid residues being most of them between 6 and 11 (Figure 4.11).

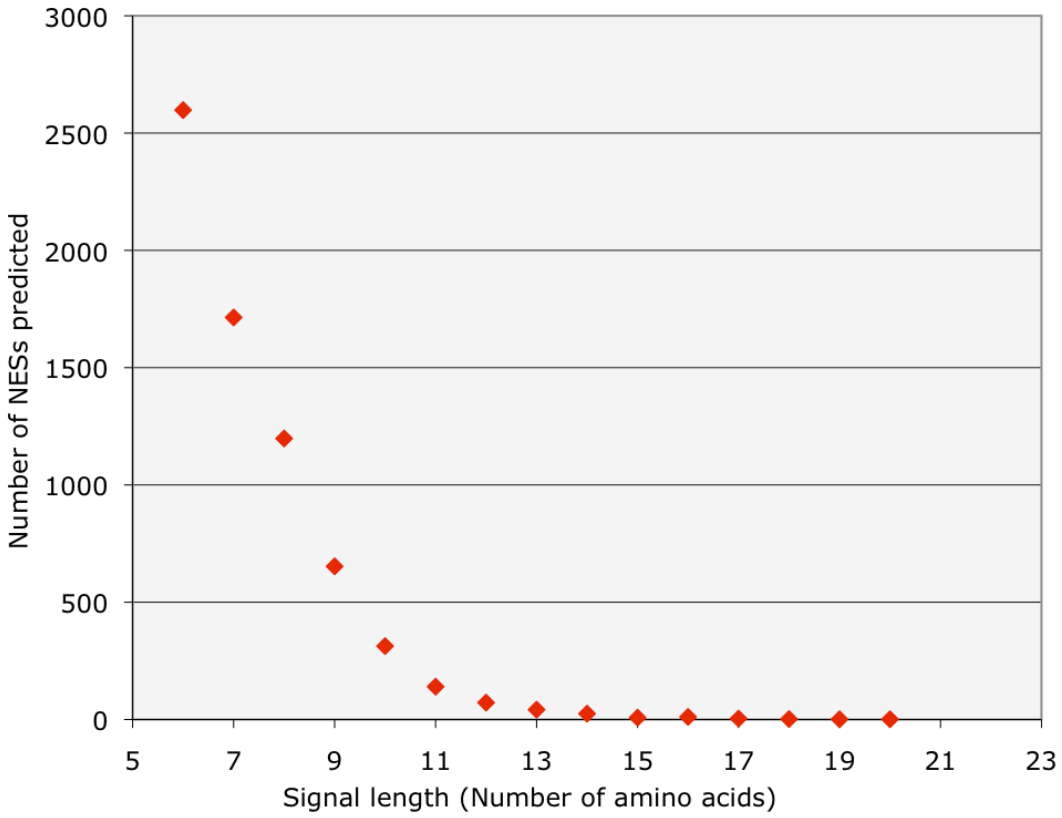


Figure 4.11: Length distribution of the predicted NESs in Arabidopsis
This graph shows the number of NESs predicted in relation to their length i.e., number of amino acid residues. Most of the signals predicted are between 6 and 11 amino acid residues long. Since more than one NES could be predicted in the same protein, the number of predicted signals is not necessarily equivalent to the number of predicted NES-containing proteins.

Selection of proteins to be experimentally tested

The next step after the prediction of the proteins possibly containing NESs in Arabidopsis was to test a group of these proteins to investigate if they had nuclear export activity. The first way to select the proteins to be tested in the laboratory was to use Gene Ontologies (GO) (The Gene Ontology Consortium, 2000). The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products: cellular component, biological process and molecular function. A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions.

Figure 4.12 shows the distribution of the predicted NES-containing proteins across the GOs molecular function and cellular process. Since many of the predicted NES-containing proteins do not have an associated GO in the categories molecular function or cellular process, it is highly probable that they are not characterized yet.

The targeted group of proteins pre-selected for experimental validation included all the sequences associated with the GOs "transcription factor activity" and "DNA or RNA binding" from the molecular function category and, "DNA or RNA metabolism" and "transcription" from the biological process category (Figure 4.12).

Since the three major GO categories (function, process and component) are represented as directed acyclic graphs (DAGs) or networks, a child node may have more than one parent node. That means that one protein could be in more than one sub-category. Therefore, the common sequences between sub-categories were examined. The distribution of the protein sequences across the selected sub-categories is shown as Venn diagrams in Figure 4.13. The protein sequences sharing at least 2 sub-categories (264 in total) were pre-selected for experimental validation.

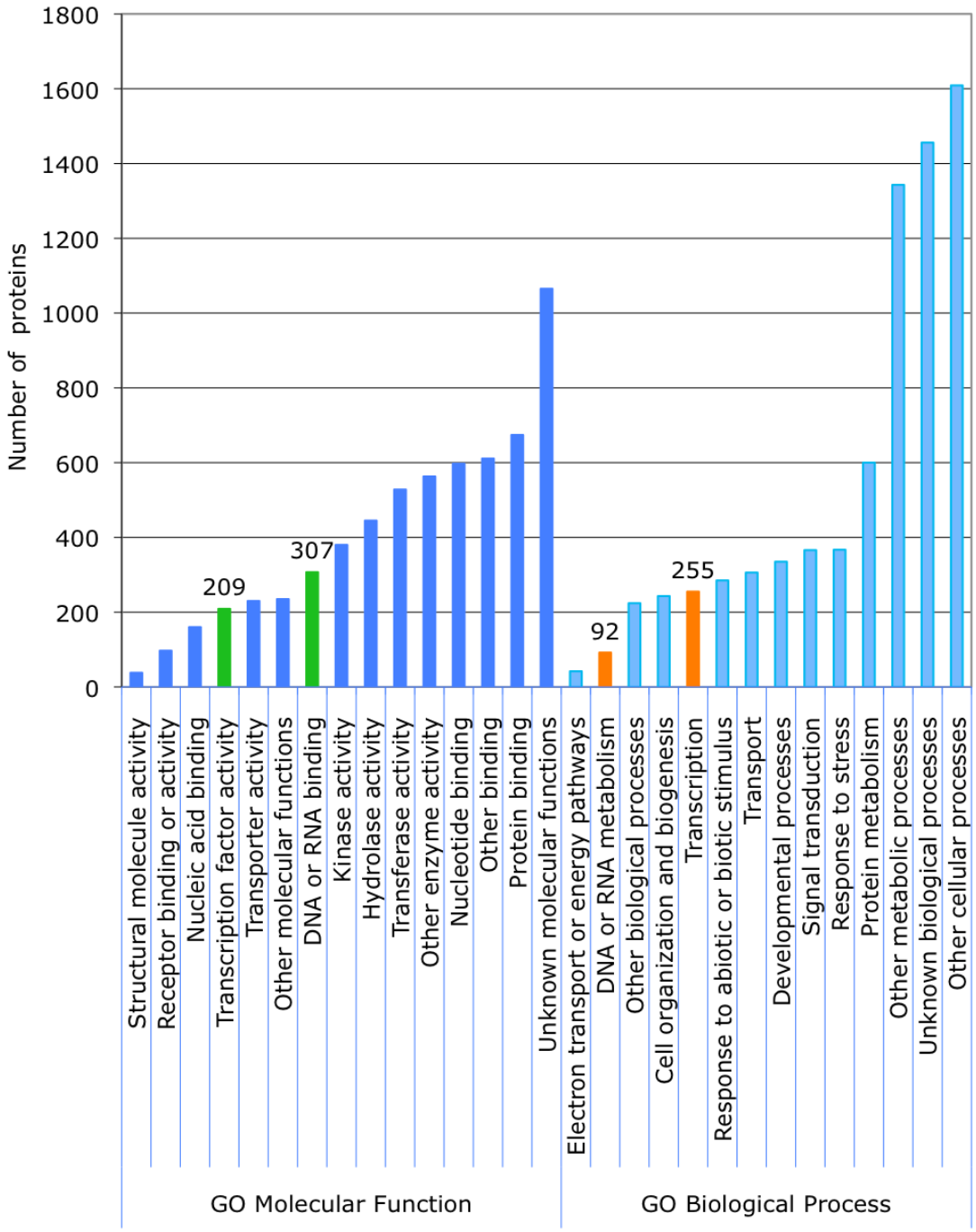


Figure 4.12: Distribution of the predicted NES-containing proteins according to gene ontologies (GO)

Distribution of the Arabidopsis proteins predicted as NES-containing among the gene ontologies (GO) molecular function and biological process. The bars colored in green correspond to the groups of proteins pre-selected for experimental testing of the nuclear export activity, the number of proteins in each of these groups are indicated above the respective bar.

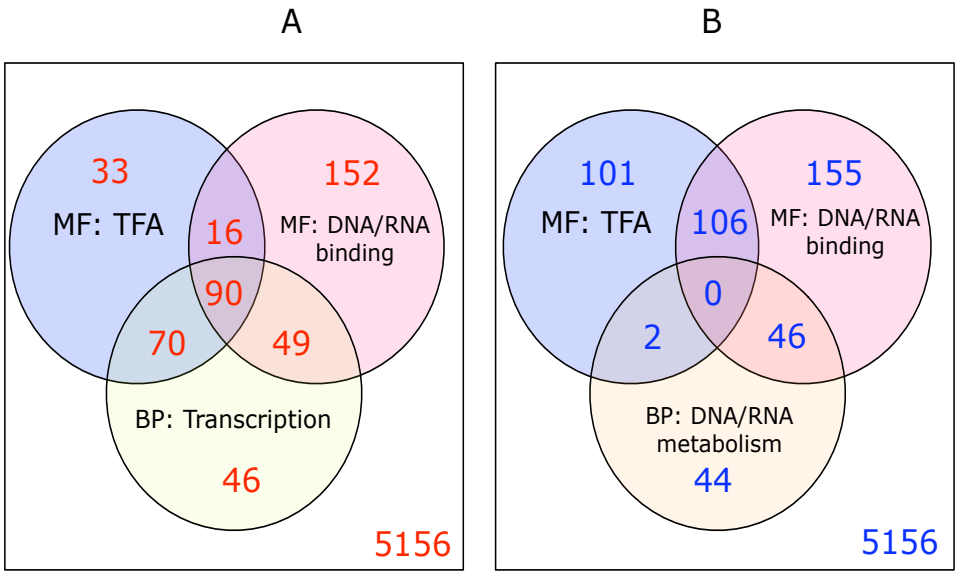


Figure 4.13: Distribution of the predicted NES-containing proteins among selected gene ontologies

These Venn diagrams show the number of proteins sequences belonging to specific gene ontologies (GOs). **A**: Distribution of the proteins for the GOs, molecular function:transcription factor activity (*MF:TFA*) (GO:0003700), molecular function: DNA/RNA binding (GO:0003677 or GO:0003723) and biological process: transcription (GO:0006350). **B**: Distribution of the proteins for the GOs, molecular function:transcription factor activity, molecular function: DNA/RNA binding and biological process: DNA/RNA metabolism (GO:0006259 or GO:0006403). A group of 264 protein sequences shared at least two of these GOs, they were pre-selected for experimental testing.

4.2 Experimental assessment of the nuclear export activity

4.2.1 NES verification in predicted proteins

The presence of an NES in a protein can be assessed by testing if the proteins interacts with the export receptor. In this case the possible interaction with the Arabidopsis Exportin 1a (XPO1a) receptor was tested in a group of proteins selected from the total predicted. The selection of the proteins to be experimentally tested was made in principle by using GOs, as described above. Ideally the experimental verification should be done in a random sample of the predicted proteins.

However, to facilitate the cloning process and obtain results for a larger number of proteins, an additional selection was made according to the experimental criteria described below.

- **Absence of specific restriction sites in the cDNA sequence**

  Since the cDNAa of the proteins that would be tested should be cloned in the plasmid pB42, they should not contain internal restriction sites for both of the enzymes used for the ligation in that vector: *EcoRI* and *XhoI*. If the cDNA contained one or more sites for only one of those enzymes, the correct orientation of the cDNA had to be verified additionally.

- **Size of cDNA**

  To facilitate the PCR amplification and posterior cloning, cDNAs with sizes greater than 2.5 kb were excluded.

Using these criteria 24 proteins from the group of 264 pre-selected by GOs were selected to be experimentally verified. The respective cDNA 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 were used in yeast two-hybrid (Y2H) assays.

A positive result in this test indicates that the tested protein interacts with XPO1a, which is a confirmation of the NES presence in that protein. The results for the examined proteins are summarized in Table 4.2 and Figure 4.14. From the 24 proteins selected for testing, 13 were evaluated by Y2H and from them 11 showed positive interaction with XPO1a.

Protein	AGI code	cDNA amplification	pB42 cloning	Y2H test	Encoded protein
1	AT5G41200	NO			
2	AT3G61150	NO			
3	AT1G09530	YES	YES	+	PIF3 (PHYTOCHROME INTERACTING FACTOR 3)
4	AT5G54260	YES	YES	-	MRE11 (MEIOTIC RECOMBINATION 11)
5	AT5G02820	YES	YES	+++	RHL2 (ROOT HAIRLESS 2)
6	AT2G43010	YES	YES	++	PIF4 (PHYTOCHROME INTERACTING FACTOR 4)
7	AT1G26260	NO			
8	AT4G35550	YES	YES	+++	WOX13 (WUSCHEL-RELATED HOMEOBOX 13)
9	AT3G62420	YES	YES	-	ATBZIP53 (BASIC REGION/LEUCINE ZIPPER MOTIF 53)
10	AT4G21750	NO			
11	AT5G46280	NO			
12	AT5G53210	YES	YES	++	SPCH (SPEECHLESS)
13	AT1G80190	YES	YES	++	PSF1 (PARTNER OF SLD FIVE 1)
14	AT5G60120	YES	YES	+++	TOE2 (TRANSCRIPTION FACTOR)
15	AT2G36010	YES	YES	++++	E2F3 (E2F TRANSCRIPTION FACTOR 3)
16	AT4G09820	NO			
17	AT1G05230	YES	NO		
18	AT1G15570	NO			
19	AT1G31360	NO			
20	AT5G67110	YES	YES	+	ALC (ALCATRAZ)
21	AT1G49720	YES	YES	+++	ABF1 (ABSCISIC ACID RESPONSIVE ELEMENT-BINDING FACTOR 1)
22	AT4G29940	NO			
23	AT5G45400	YES	NO		
24	AT1G75080	YES	YES	++	BZR1 (BRASSINAZOLE-RESISTANT 1)

Table 4.2: Summary of the results for the tested proteins.

Oligonucleotides were designed for 24 of the predicted NES-containing proteins. cDNA fragments generated from positive PCR experiments were ligated into the vector pB42AD and tested in the yeast strain EGY48[p8op-LacZ] for interaction with Arabidopsis XPO1a. This table presents the summary of the results for all the 24 proteins selected. 13 of them were evaluated using the yeast two-hybrid assay (highlighted rows). The remaining 11 failed either in the cDNA amplification or the cloning process. The sequence of the oligonucleotides are included in the appendix.

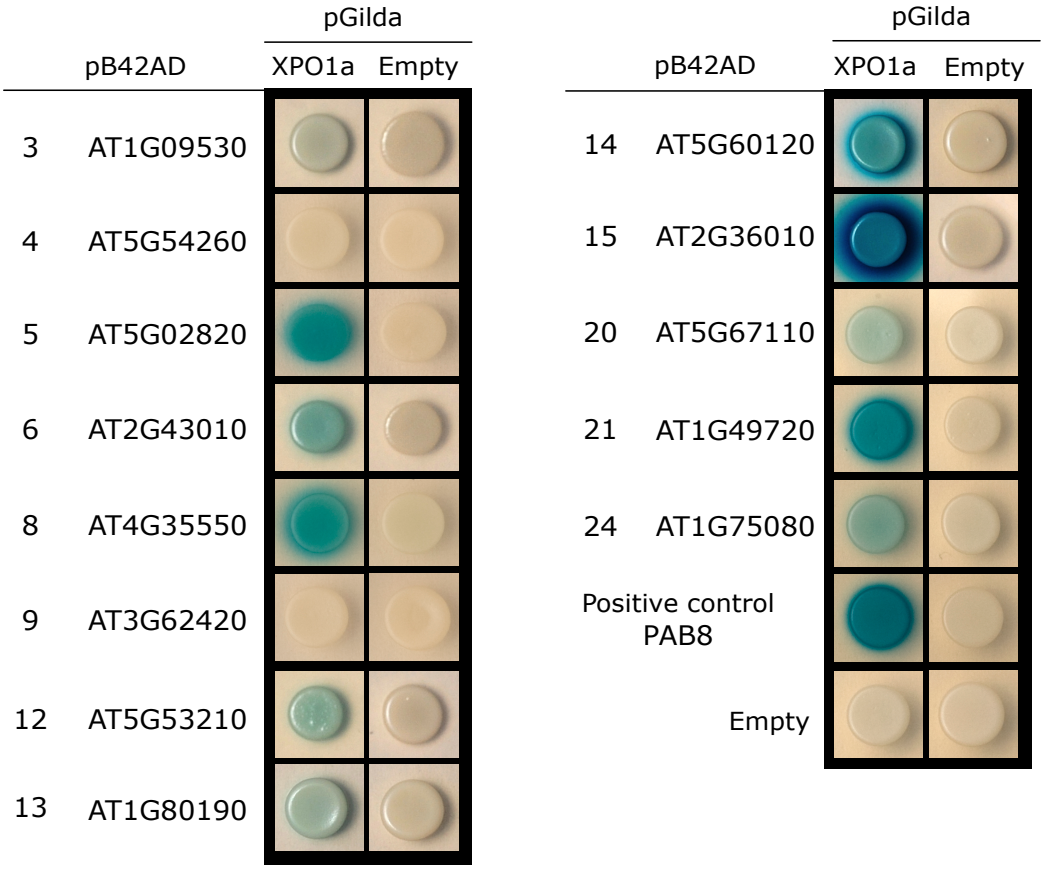


Figure 4.14: XPO1a binding activity for selected proteins out of the total predicted
Yeast two-hybrid assays for 13 out of the 24 proteins selected from the protein predicted as containing NES. The respective cDNA fragments amplified by PCR were ligated into the vector pB42AD and tested in the yeast strain EGY48[p8op-LacZ] for interaction with *A. thaliana* XPO1a. This group of proteins correspond to the highlighted cells in Table 4.2, the number on the left side next to the accession number corresponds to the same used in that table. The protein whose cDNA was cloned in plasmid pB42AD is labelled at the left of each picture panel and the content of plasmid pGilda (XPO1a cDNA or empty vector) is indicated on top. Blue color indicates positive interaction, 11 proteins out of the 13 assayed showed a positive interaction with XPO1a. For only two proteins, number 4 (AT5G54260) and number 9 (AT3G62420), no positive interaction was detected.

4.2.2 Further analysis of some NES-containing proteins

The influence of regions inside and outside of the NES on the XPO1a binding activity was assessed using proteins from Arabidopsis that were already known to contain a functional NES. Some of them were included in the group of NESs used for classifier development.

Effect of changes inside the NES

The effect of changing amino acid residues others than the hydrophobic ones inside the NES region was evaluated by using the NES of the PAB7 protein. Poly(A)-binding proteins (PABPs) are a family of proteins characterised by their ability to bind to poly(A) RNA. They bind the poly(A) tails of newly synthesized or mature mRNAs and appear to act as *cis*-acting effectors of specific steps in the polyadenylation, export, translation, and turnover of the transcripts to which they are bound. Lacking any evident catalytic activity, PABPs provide a scaffold for the binding of factors that mediate these steps and also apparently act as antagonists to the binding of factors that enable the terminal steps of mRNA degradation (Mangus *et al.*, 2003; Gorgoni and Gray, 2004; Kühn and Wahle, 2004). Based on intracellular location and phylogeny, PABPs have been divided into two broad categories, nuclear (PABPNs) and cytoplasmic (PABPCs). Both classes are ubiquitous in eukaryotic organisms, but PABPNs bear little resemblance to their cytoplasmic counterparts also in their functions (Wahle, 1991; Nemeth *et al.*, 1995; Kühn and Wahle, 2004).

PABPCs contain four consecutive RNA recognition motifs (RRMs), located in the N-terminal region, connected to a conserved C-terminal domain referred to as the PABC or CTC (Mangus *et al.*, 2003). Arabidopsis contains eight genes for PABPCs, all of which are expressed (Belostotsky, 2003). This is an unexpectedly large number in comparison to other eukaryotes whose genomes have been sequenced. For example, PABPCs are encoded by single genes in *Saccharomyces cerevisiae* (Sachs *et al.*, 1987), *Schizosaccharomyces pombe* (Thakurta *et al.*, 2002) and *Drosophila melanogaster* (Sigrist *et al.*, 2000); two genes are present in *Caenorhabditis elegans* and four in humans (Mangus *et al.*, 2003). Using phylogenetic comparisons coupled with expression analyses, the eight PABPCs of Arabidopsis were grouped into four classes (Belostotsky, 2003). The expression of class I (PAB3 and PAB5) is limited to reproductive tissue; class II members (PAB2, PAB4 and PAB8) are highly and broadly expressed; class III PABPs

(PAB6 and PAB7) have a restricted, weak expression pattern; and the sole member of class IV (PAB1) has low, tissue-specific expression.

In a previous study in the laboratory of Dr. Thomas Merkle, the XPO1a binding activity of the Arabidopsis PABPCs was assessed (Roessiger, 2008). Some of the PABPCs (PAB2, PAB8, PAB4) exhibited strong interaction whereas others (like PAB7) showed a weaker interaction and some other no interaction at all (PAB3). Comparing the NES from PAB7 (weak XPO1a interaction) with PAB8, PAB4 and PAB2 (strong XPO1a interaction), some differences can be noted in the spacer amino acid residues of the signal, i.e. the residues located between the hydrophobic ones. To test if these amino acid residues could influence the interaction with XPO1a, the NES of PAB7 was mutated in order to convert its weak NES into a strong one. The description of these mutations as well as a comparison of the NES sequence and the XPO1a interaction activity between PAB7 and PAB8 are shown in Figure 4.15.

The XPO1a interaction activity was assessed in Y2H assays for the wild type PAB7 and the PAB7 NES mutants. These results are presented in Figure 4.16. Modifications of some of the amino acid residues others than the hydrophobic ones in the NES sequence of PAB7 produced changes in the extent of the XPO1a binding activity compared to the PAB7 wild type. All the changes of individual amino acid residues yielded an increase of the XPO1a binding activity that is more remarkable when the Serine in position 601 is exchanged for Glutamate. When the four amino acid residues were simultaneously changed so that the NES of PAB7 was almost identical to the one of PAB2 and PAB4 concerning the spacing residues, the increase of the activity was even higher, in this case twice as high as compared to the PAB7 wild type NES.

Effect of changes in regions outside of the NES

To asses if the XPO1a binding activity could also be influenced by regions outside of the NES, the proteins CID 11 (AT1G32790) and CID 12 (AT4G10610) from Arabidopsis were used. CID (for **C**TC **i**nteracting **d**omain) proteins are potential interaction partners of the PABPCs (Bravo *et al.*, 2005). Two of them (CID 1 and CID 7) were initially isolated in a yeast two-hybrid screening using the PABC region of PAB2 and eleven members more were identified in a database search using the domain PAM2 (PABP interaction motif 2) as bait (Bravo *et al.*, 2005). The motif PAM2 is present in the PABP interacting proteins Paip1 and Paip2

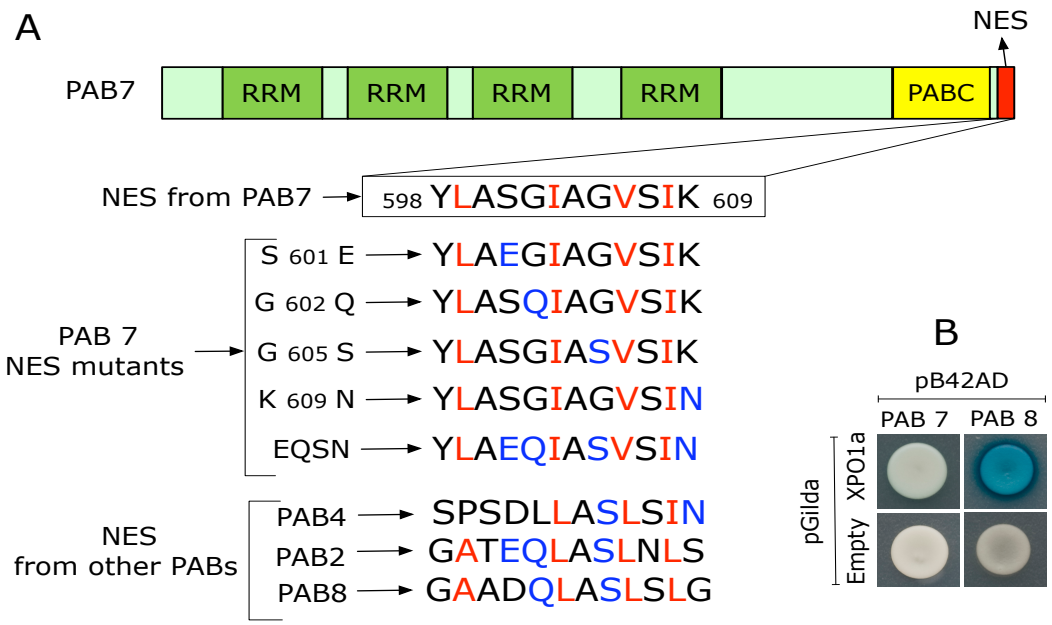


Figure 4.15: Comparison of the NES of some Arabidopsis PABPC proteins

The NES of PAB7 differs from that of PAB2, PAB4 and PAB8. Since PAB7 interacted weakly with XPO1a and PAB2, PAB4 and PAB8 interacted strongly, some amino acid residues inside the NES from PAB7 where changed trying to emulate the NES from the proteins that had a strong XPO1a interaction activity. **A**: A scheme of PAB7, the main sequence motives are shown and the NES region is enlarged. The hydrophobic amino acid residues in the NESs are shown in red, the amino acid residues changed in the respective mutant are shown in blue, the NES of PAB2, PAB4 and PAB8 are also presented. **B**: Interaction of PAB7 and PAB8 with XPO1a assayed by yeast two-hybrid assays. The assays were carried out using plasmids pB42AD containing the cDNA of proteins PAB7 or PAB8 and pGilda containing XPO1a cDNA or the empty plasmid as a control. Blue colour indicates positive interaction between the two proteins.

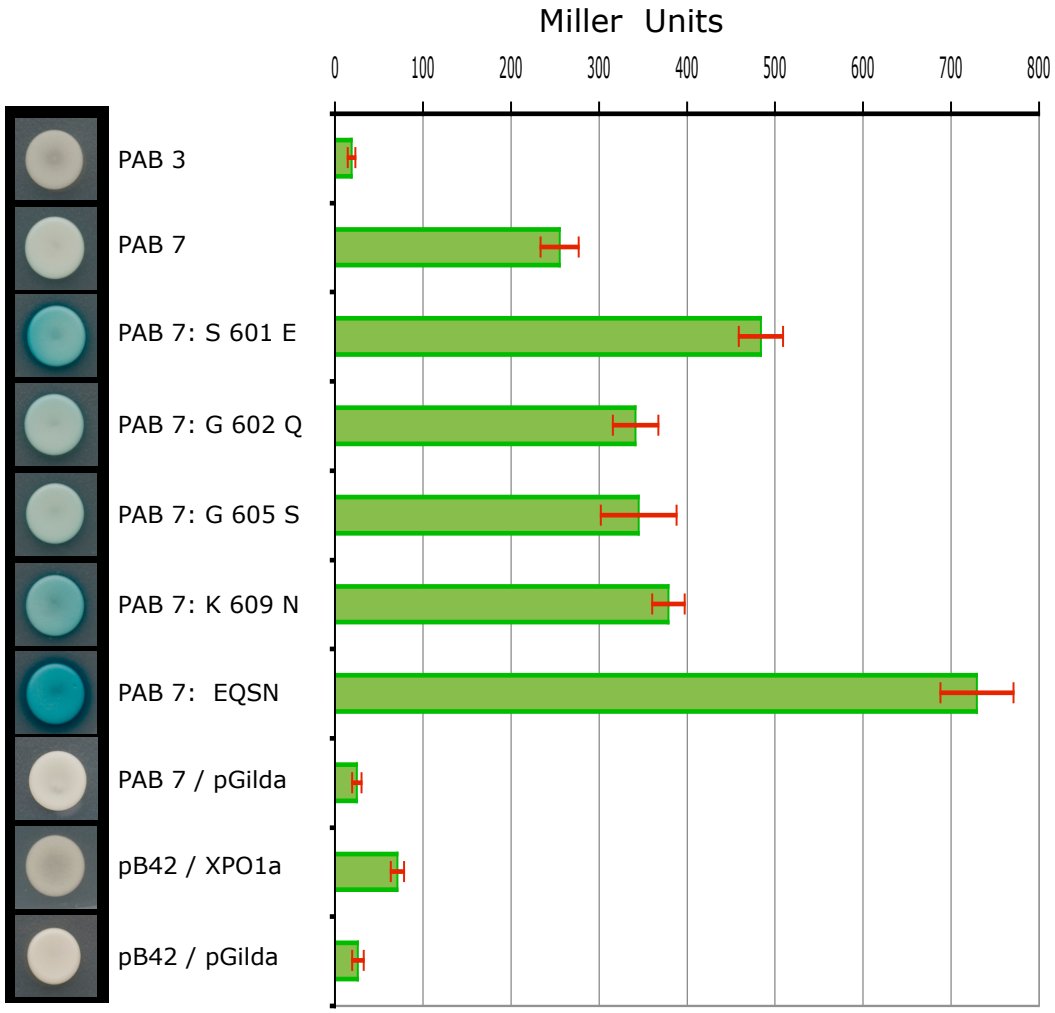


Figure 4.16: Yeast-two hybrid assays for PAB3 and proteins derived from PAB7

Changes in some amino acid residues inside the NES of PAB7 increased its XPO1a binding. The XPO1a binding activity of the proteins described in Figure 4.15 was assessed by yeast two-hybrid assays. The cDNAs of the proteins indicated in the central panel of the figure were cloned in plasmid pB42AD and tested in the yeast strain EGY48[p8op-LacZ] for interaction with XPO1a from *A. thaliana*, whose cDNA was cloned in plasmid pGilda. The interaction was evaluated by production of β-galactosidase in plates with the X-gal substrate (qualitative assay, left side of the figure) or in solution by using the ONPG-assay (quantitative assay, right side of the figure). In the qualitative assay, the development of a blue color indicates a positive interaction between the two tested proteins. PAB7/pGilda and pB42/pGilda correspond to controls using cDNA of the protein PAB7 or the plasmid pB42AD empty respectively, together with the plasmid pGilda empty. pB42/XPO1a corresponds to a control using the plasmid pB42AD empty and the cDNA from XPO1a in plasmid pGilda.

from humans and mediates their interaction with PABPs (Khaleghpour *et al.*, 2001; Roy *et al.*, 2002).
The thirteen CID genes from Arabidopsis were grouped in four classes: A, B, C and D (Bravo *et al.*, 2005). CID 11 and CID 12 are both in the class D, these encode highly related RBPs (RNA binding proteins) containing two RRMs and a basic region that resembles a bipartite NLS. CID 11 and CID 12 are closely related in sequence (Figure 4.17A), their NES differs just in two amino acid residues that do not belong to the group of hydrophobic ones. Nevertheless, these two proteins displayed a very different XPO1a interaction activity. In yeast two-hybrid assays, CID 11 interacted strongly with XPO1a whereas CID 12 interacted weakly (Figure 4.17B).

To elucidate why CID 11 and CID 12 interacted so differently with XPO1a, the mutant proteins whose sketch is shown in Table 4.3, were constructed and tested in Y2H assays. The proteins assayed were (the names used in Table 4.3 appear in *italics*):

- *CID 11* and *CID 12* wild type proteins.
- *CID 11 NES mut*: CID 11 with a mutated version of the NES in which two of the leucine residues were changed to alanine, obtained by overlap-extension PCR as described in Section 3.2.1.
- *CID 11 NES CID 12*: Version of CID 11 containing the NES sequence of CID 12, it was obtained by overlap-extension PCR.
- *CID 12-11 NES CID 12* and *CID 11-12 NES CID 11*: Two chimeric proteins obtained by restriction of CID 11 and CID 12 with the enzyme *Hind III* followed by ligation. This enzyme cuts at the end of the NES at identical positions in both proteins (shown in Figure 4.17). The first chimeric protein contains the N-terminal end of CID 12 together with its NES and the rest of CID 11 (after the NES). In the same way, the second protein contains the N-terminal end from CID 11 with its NES and the rest of CID 12.
- *CID 11 minus C-end* and *CID 12 minus C-end*: Shorter versions of CID 11 and CID 12 obtained by PCR, both proteins contain around 100 amino acid residues less than the wild type.
- *CID 11-12 (II)* and *CID 11-12 (II)*: Again two chimeric proteins, this time the C-terminal end of CID 11 and CID 12 was interchanged by using overlap-extension PCR.

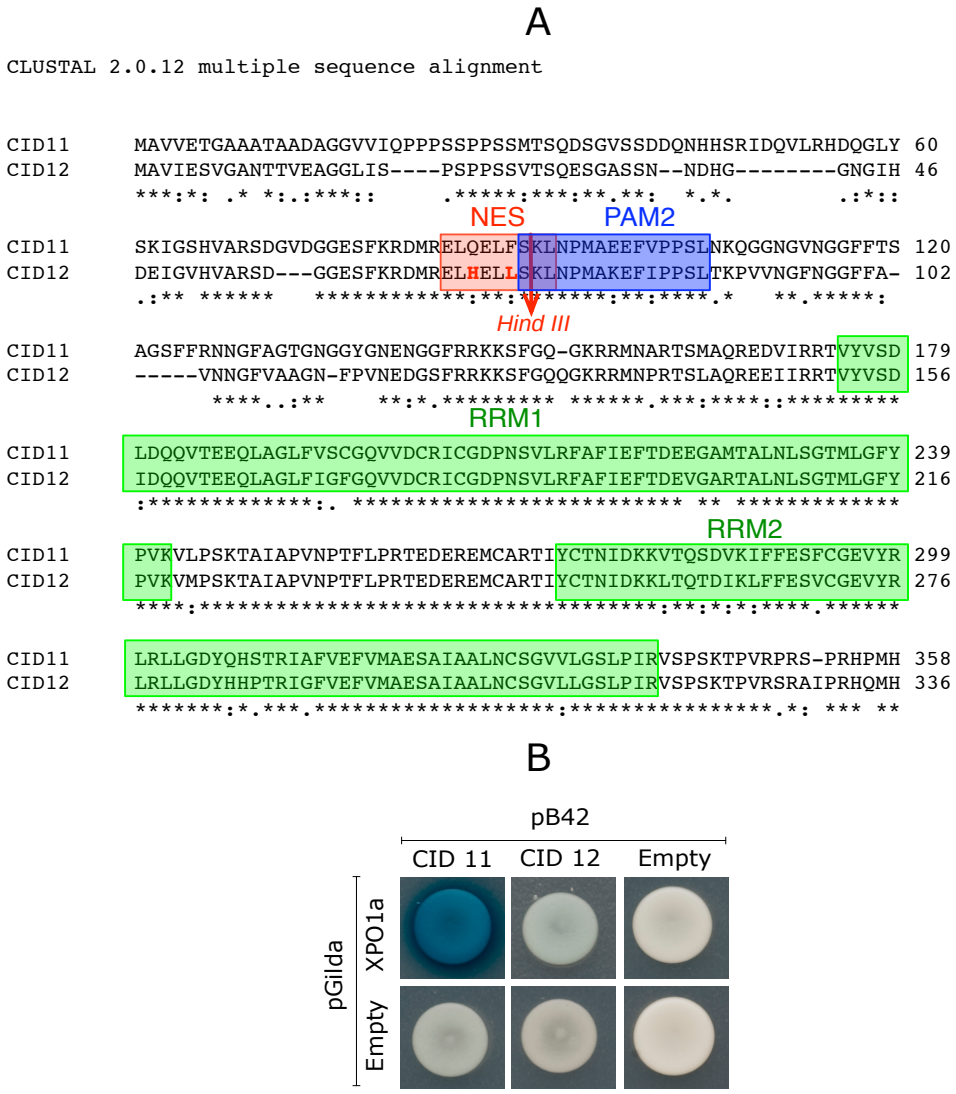


Figure 4.17: Comparison of XPO1a interaction of CID 11 and CID 12

Proteins CID 11 and CID 12 from *A. thaliana* are highly similar, however they showed different interaction with XPO1a in the yeast two-hybrid assay. **A**: Sequence alignment of proteins CID 11 and CID 12. Nuclear export signal (NES), RNA-recognition motives (RRM) and PAB interacting motif 2 (PAM2), are highlighted in red, green and blue, respectively. The restriction site for the enzyme *Hind III* is marked in red. This enzyme was used to construct chimeric proteins by combination of different regions of each protein. **B**: Yeast two-hybrids assay with proteins CID 11 and CID 12. The assays were carried out using plasmids pB42AD containing CID 11 cDNA or CID 12 cDNA and pGilda containing XPO1a cDNA or the empty plasmid as control. Blue colour indicates interaction between the two tested proteins.

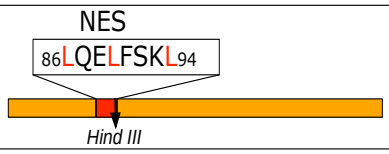
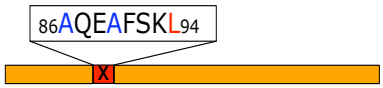
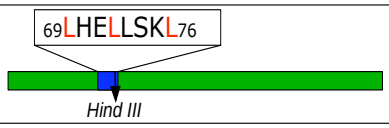
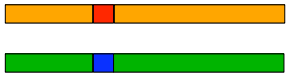
Name	Scheme	Protein description
CID 11		CID 11 wild type
CID 11 NES mut		CID 11 with mutated NES
CID 12		CID 12 wild type
CID 11 NES CID 12		CID 11 with NES of CID 12
CID 12-11 NES CID 12 CID 11-12 NES CID 11		Fusion proteins where the region after the NES was interchanged
CID 11 minus C end CID 12 minus C end		Short versions of CID11 and CID12 containing around 100 residues less in the C-terminal end
CID 11-12 (II) CID 12-11 (II)		Fusion proteins where the C-terminal end was interchanged

Table 4.3: Chimeric proteins obtained from CID 11 and CID 12

Description and schematic presentation of proteins CID 11 and CID 12 tested in Y2H assays. The left column contains the name of each protein, as shown in the text and in Figure 4.17, the central column presents the architecture of each protein and the right column explains briefly the characteristics of each protein. Protein CID 11 is represented in orange color, CID 12 in green, the NES from CID 11 in red and the NES from CID 12 in blue.

The results from the Y2H assays with the described proteins are shown in Fig 4.18. As was already known, the XPO1a binding activity of CID 11 is stronger than the one of CID 12. Furthermore, it is evident that the mutation of the hydrophobic residues in the NES of CID 11 eliminated such activity.
The XPO1a binding activity of CID 11 was diminished when its NES was replaced by the one of CID 12, a reduction approximately of the same extent was also observed when the C-terminal end of CID 11 was removed but the high activity was recovered if its C-terminal end was replaced by the one from CID 12.
The interchange of the regions after the NES between CID 11 and CID 12 decreased the original high XPO1a binding activity of CID 11 but did not eliminate it completely. Both of these proteins (*CID 12-11 NES CID 12* and *CID 11-12 NES CID 11*) showed less activity than CID 11 but more than CID 12, being the activity of the one that contains the NES from CID 11 more than twice higher than the one with the NES from CID 12.

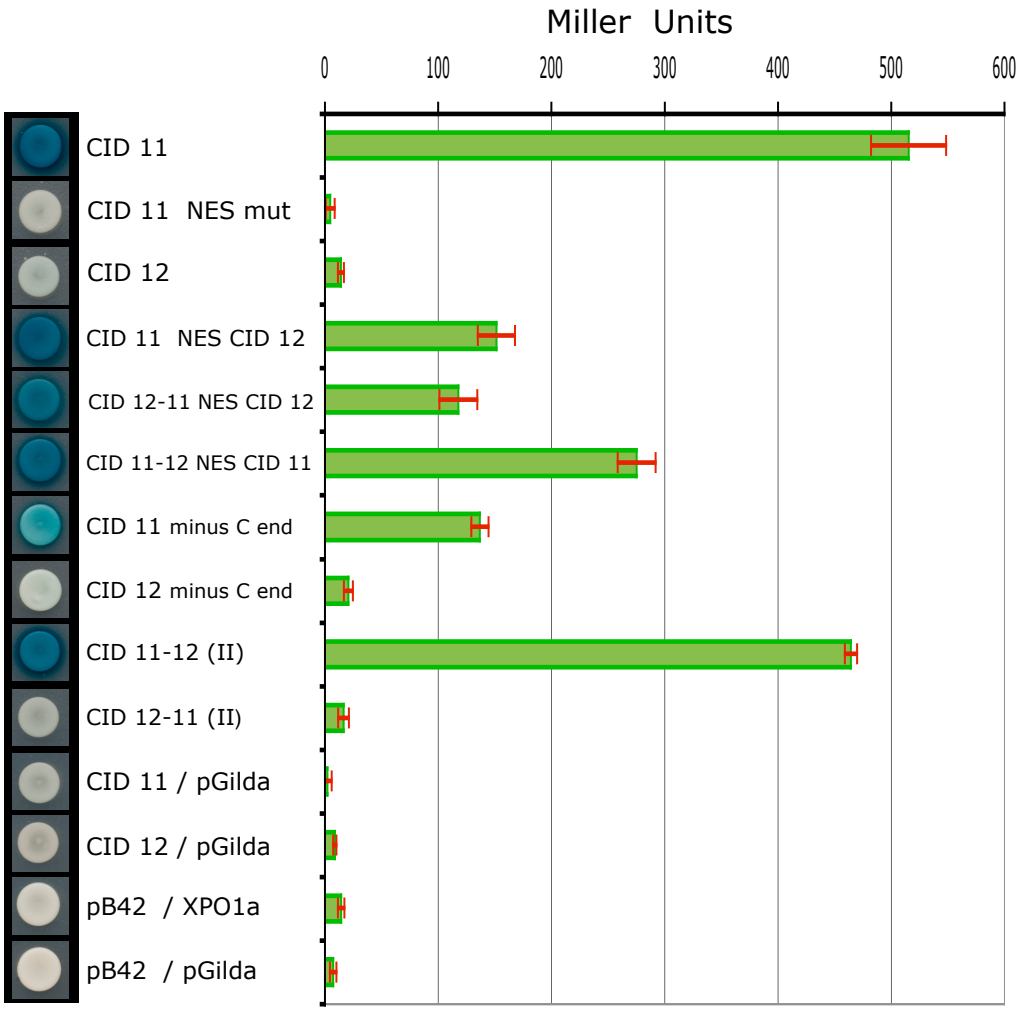


Figure 4.18: Y2H assays for proteins derived from CID 11 and CID 12

Domain swapping experiments between the proteins CID 11 and CID 12 produced variations in the original XPO1a binding activity of these two proteins. The XPO1a binding activity of the proteins described in Table 4.3 was assessed in yeast two-hybrid assays. The cDNAs of the proteins indicated in the central panel of the figure were cloned in plasmid pB42AD and tested in the yeast strain EGY48[p8op-LacZ] for interaction with XPO1a from *A. thaliana* whose cDNA was cloned in plasmid pGilda. The interaction was evaluated by production of β-galactosidase in plates with the X-gal substrate (qualitative assay, left side of the figure) or in solution by using the ONPG-assay (quantitative assay, right side of the figure). In the qualitative assay, the development of a blue color indicates a positive interaction between the two tested proteins. CID 11/pGilda, CID 12/pGilda and pB42/pGilda correspond to controls using cDNA of the proteins CID 11 or CID 12 or the plasmid pB42AD empty respectively, together with the plasmid pGilda empty. pB42/XPO1a corresponds to a control using the plasmid pB42AD empty and the cDNA from XPO1a in plasmid pGilda. pB42/pGilda corresponds to a control with both empty plasmids.

CHAPTER 5

Discussion

Contents

5.1 Analysis of LR-NESs of proteins from Arabidopsis

The sequences of LR-NESs of proteins from Arabidopsis were compared to LR-NES of proteins from viruses, yeast and humans that were already published (la Cour *et al.*, 2003) by using sequence logos. This kind of representation provides a visual yet precise description of sequence similarity that is superior to consensus sequences and can rapidly reveal significant features of the alignment that are otherwise difficult to perceive (Crooks *et al.*, 2004).

There are visible differences between the LR-NES of proteins from Arabidopsis and those from other organisms (Figure 4.1). The conservation of the four hydrophobic residues, mainly leucines, of a typical LR-NES is more evident in the sequences of organisms like yeast and humans than in the sequences from Arabidopsis. It

is also noteworthy that the spacing and the amino acid residues that are found between the hydrophobic residues differ between these two groups of sequences. For example, the pattern:

$$\boldsymbol{\phi}\text{-}\boldsymbol{x}_{\mathbf{2\text{-}3}}\text{-}\boldsymbol{\phi}\text{-}\boldsymbol{x}_{\mathbf{2}}\text{-}\boldsymbol{\phi}\text{-}\boldsymbol{x}\text{-}\boldsymbol{\phi}$$

with (2-3), 2 and 1 amino acid residues between the hydrophobic residues ($\boldsymbol{\phi}$), can be noted in Figure 4.1 B, but not in the sequences from Arabidopsis (Figure 4.1A). On the other hand, a definite preference for glutamate, aspartate and serine residues is observed in LR-NES positions not occupied by hydrophobic residues in both group of sequences, which is in agreement with previous studies (la Cour *et al.*, 2004; Kosugi *et al.*, 2008). Additionally, asparagine and glutamine are also present in the Arabidopsis proteins at the spacing positions. Two recent publications describe the crystal structure of a the complex between Exportin 1 and an NES cargo (the human protein Snurportin, SNUPN1 or SNP1) (Dong *et al.*, 2009a,b; Monecke *et al.*, 2009). SPN1 is a nuclear import adapter for cytoplasmically assembled spliceosomal uridine-rich small nuclear ribonucleoprotein particles (UsnRNPs). The protein is recognized by CRM1 via two distinct NES epitopes. Epitope I corresponds to the consensus LR-NES motif whereas the second epitope comprises a long patch of basic residues unrelated to the LR-NESs (Paraskeva *et al.*, 1999; Dong *et al.*, 2009a).
In the determined structures, the solvent accessible face of the LR-NES helix is composed of polar residues (Glu 2, Glu 3, Ser 5, Gln 6 and Ser 10 in the case of SPN1). Besides, it was determined that the acidic NESs side chains Glu 2 and Glu 3 make electrostatic contacts with the basic Exportin-1 side chains Lys 560 and Lys 522 that flank the hydrophobic groove (Dong *et al.*, 2009a). Hence, one side of the LR-NES is exposed to solvent, which explains the presence of polar residues between the hydrophobic positions, whereas the basic surface that flanks the N-terminal half of the exportin-1 groove explains the preference for acidic and electronegative residues.
The differences observed between the LR-NESs from Arabidopsis and from other eukaryotes could explain why the current tool (la Cour *et al.*, 2004) is not useful to identify LR-NESs in proteins from plants (i.e. Arabidopsis). It highlights also the importance of developing species-specific or kingdom-specific prediction systems.

Additional features of LR-NESs of proteins from Arabidopsis were analyzed by using qualitative and quantitative yeast two-hybrid assays. First, it is noticeable that the results from the qualitative and quantitative assays are in accordance

since the intensity in the blue color in the qualitative assays is correspondent with the value of the activity of β-galactosidase expressed in Miller units (Figures 4.16 and 4.18). That means that the darker and lighter blue yeast spots in the qualitative assay correspond to the higher and lower values in activity in the quantitative assay, respectively.

The experiments carried out with the protein PAB7 of Arabidopsis confirmed the importance of polar and acidic amino acid residues located between the hydrophobic residues of the LR-NES. The modifications in the LR-NES of PAB7 (Figure 4.15) were directed to transform a weak NES into a strong one (like that found in PAB4, PAB2 and PAB8 proteins). In fact, the results shown in Figure 4.16 indicate that the interaction between the LR-NES-containing protein and the receptor XPO1a was favoured when polar residues were placed instead of non-polar (for instance: G to Q in position 602 or G to S in 605), charged acidic residues instead of polar (S to E in 601) or a basic charged residue was changed (K to N in 609). The increase in XPO1a interaction activity was even higher when all the residues were changed simultaneously in comparison to individual modifications: in Figure 4.16, the XPO1a binding activity of PAB7:EQSN was around three times higher than the activity of the PAB7 wild type protein. Similar effects have been also reported by Dong *et al.* (2009a), where mutation of electronegative residues Glu 2, Glu 3 and Ser 11 in the LR-NES helix of the protein SPN1 decreased the interaction with Exportin 1. Furthermore, the quality of exportin-binding and the kinetics of intracellular transport have been also shown to be affected by the hydrophilic spacing amino acid residues in some LR-NES (Heger *et al.*, 2001; Engelsma *et al.*, 2004; Geisberger *et al.*, 2009). Taken together, these results suggest a role for polar contacts at the interface of the LR-NES and exportin receptor interaction and explain why the protein PAB7 of Arabidopsis exhibited a weak interaction with the receptor XPO1a.

The influence of the spacing residues was also observed in the experiments carried out with the proteins CID 11 and CID 12. The LR-NES from CID 12 contains a histidine instead of the glutamine after the first leucine (see alignment in Figure 4.17). This could explain partially that CID 12 showed a weaker XPO1a interaction activity as compared to CID 11. This explanation is only partially true because if that were the only reason, the XPO1a interaction activity of CID 11 containing the LR-NES of CID 12 would be as low as the one of CID 12. This was not the case (Figure 4.18, “CID 11 NES CID 12” compared to “CID 11” and “CID 12”). These findings can be interpreted as follows: the LR-NES of CID 12

is weaker than the one form CID 11 possibly due to the presence of histidine in the LR-NES of CID 12, but there is an additional characteristic of CID 12 that makes the XPO1a interaction activity even weaker. The chimeric protein "CID 12-11 NES CID 12" showed a similar XPO1a interaction activity as compared to "CID 11 NES CID 12", which is in agreement with the previous statement and eliminates the possibility that the N-terminal region of CID 12 is an "LR-NES activity inhibitory region". The region containing the two RRMs as well as the PAM2 motif are practically identical in the two proteins, hence it could not be a plausible reason for the different XPO1a interaction activities of CID 11 and CID 12. It is noticeable, however, that the C-terminal region is needed for an optimal activity, since the XPO1a interaction activity of CID 11 without the C-terminal is lower as compared to the full-length protein (Figure 4.18, "CID 11 minus C end" compared to "CID 11"). Also, the high interaction activity of CID 11 is maintained also with the C-terminal region of CID 12 (Figure 4.18, "CID 11-12 (II)" compared to "CID 11").

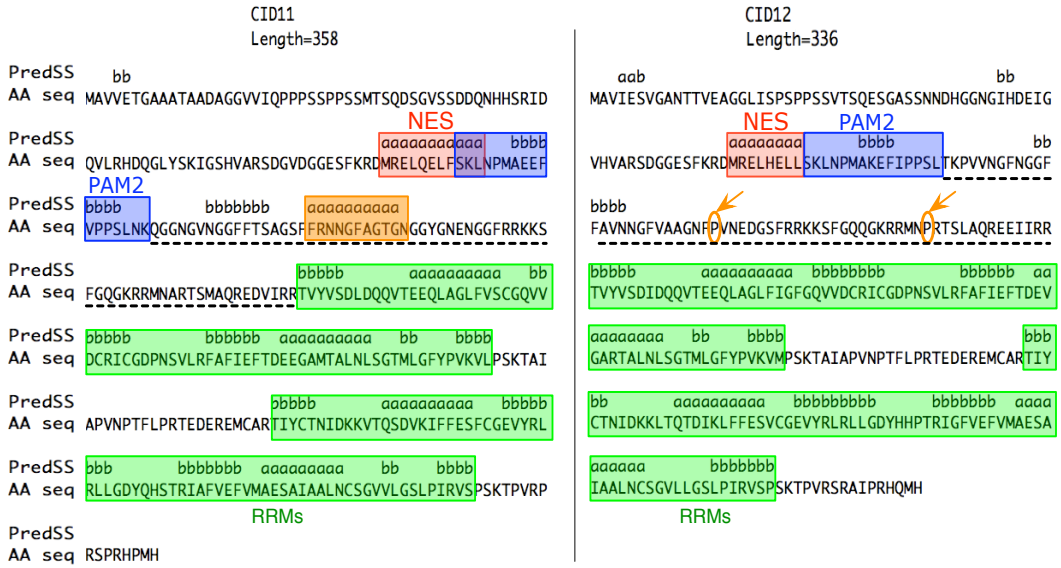


Figure 5.1: Secondary structure prediction for the proteins CID 11 and CID 12
The secondary structure of proteins CID 11 and CID 12 was predicted with SSPRED at http://linux1.softberry.com/. The predicted secondary structure elements are shown above the amino acid sequences as *a* for α-helix, *b* for β-sheet or *any letter* for coiled regions. The regions in the amino acid sequence corresponding to the RNA recognition motifs (RRMs), nuclear export signal (NES) and PABP interaction motif 2 (PAM2) are highlighted in green, red and blue color, respectively. The region of the protein CID 11 highlighted in orange corresponds to a predicted α-helix that is not predicted for CID 12. Similarly, the proline residues postulated to be the reason for the absence of the α-helix in CID 12 are shown with orange arrows.

This analysis restricts the possible "inhibitory characteristic" of CID 12 to a region comprised between the end of the PAM2 motif and the beginning of the first RRM. This statement is consistent with the decrease in activity observed in the chimeric protein "CID 11-12 NES CID 11" compared to "CID 11". This chimeric protein contains the N-terminal region of CID 11, including its LR-NES (a strong NES) fused to the region of CID 12 after the LR-NES. Sequence and structural analyses of LR-NES-containing proteins suggested that the signal needs to be accessible and flexible (la Cour *et al.*, 2004; Dong *et al.*, 2009a). So, one possibility is that in the context of the protein, the LR-NES is not exposed and because of that is not available to interact with XPO1a. To explore this aspect further, the amino acid sequences of CID 11 and CID 12 were analysed with the program SSPRED for secondary structure prediction from `http://linux1.softberry.com/` (Figure 5.1). The helical structure predicted for both LR-NES regions is consistent with previous reports (la Cour *et al.*, 2004; Dong *et al.*, 2009a). Specially the structural analysis in Dong *et al.* (2009a) shows that in the case of SPN1, the LR-NES region adopts a helix conformation. Additionally, the RRMs exhibit a pattern of helices and sheets connected by short coiled regions also in agreement with other RNA binding proteins containing that motif. It is interesting that the amino acid residues comprised between the end of the PAM2 motif and the beginning of the first RRM (dashed region in Figure 5.1) presents some differences in the predicted secondary structures for both proteins. On one side, CID 11 presents an additional predicted helix that is not predicted in CID 12. On the other side, CID 12 has two prolines in that region that could be the reason for the adoption of a coiled structure in this area instead of a helix. This observation could have consequences in the tertiary structure of CID 12 that limit the exposure of the LR-NES. Thus, the CID 12 LR-NES could be additionally buried in the three-dimensional structure of the protein.

These results indicate that, in addition to the identity of the spacing amino acid residues in the LR-NES region, residues flanking the NES apparently contribute to the interaction, as has been indicated by some authors (Paraskeva *et al.*, 1999; Petosa *et al.*, 2004; Dong *et al.*, 2009a). The NES needs an appropriate protein context to adopt a conformation required for high-affinity exportin receptor binding .

5.2 Development of the prediction tool for LR-NESs

The first step of the development process was the extraction or calculation of the features from the initial amino acid sequences data. Although the most intuitive characteristic to consider could be the amino acid order i.e. primary sequence, in the case of LR-NESs there were some additional points to be taken into account. First, LR-NESs are short sequences (7-12 amino acid residues, depending of the number of spacing residues) where the most outstanding residue uses to be leucine. Leucine is the most abundant amino acid in proteins (9.97 % in *H. sapiens*, 9.52 % in *A. thaliana*, 9.58 % in *S. cerevisiae*, (Pruess *et al.*, 2003; Schneider and Fechner, 2004) and additionally it is present in many hydrophobic signals in proteins that are not related to nuclear export. Thus an indication was needed that positive signals "NES" could be distinguished from negative "nonNES". The pair-wise alignment of all the amino acid sequences (Figure 4.2) showed that there was more similarity among the NES sequences than among non related sequences (nonNES). Hence, the sequence of the NESs could be used as one parameter to distinguish between "NES" and "nonNES" samples.
One possible way to express the amino acid order in a numerical feature could be the comparison of every sequence with some of the LR-NES consensus sequences available (Bogerd *et al.*, 1996; Kosugi *et al.*, 2008), which can generate some similarity measure. However, it has been well documented that the consensus sequences alone are not an optimal approach to detect NESs. Furthermore, as was discussed before, the NESs from Arabidopsis showed differences compared to the NESs from other organisms and the available consensus sequences did not include sequences of plant NESs. Instead of that, a profile HMM was used. Profile HMMs turn a multiple sequence alignment into a position-specific scoring system for searching for remotely homologous sequences (Eddy, 1998). Furthermore, they have a solid theoretical background and have been widely used alone or in combination with other techniques in many bioinformatics applications.

On the other hand, the use of amino acid index values (*aaindex*) as a source of features is a simple way to reflect the possible influence of some physical and chemical properties of the amino acid residues on the characteristics of an NES. In addition, the use of such values is well documented in the literature in a broad range of bioinformatics applications, which include sequence comparisons (Tomii and Kanehisa, 1996), prediction of structure and function of proteins (Bu *et al.*, 1999; Lee *et al.*, 2009b), prediction of specific binding sites (Tung and Ho, 2007),

identification of peptides in proteomic studies (Sanders *et al.*, 2007) and prediction of ubiquitylation sites (Tung and Ho, 2008b).

After the feature calculation, the NES classification became a high-dimensional problem where the number of features (that is the dimension of the feature vector $\mathbf{x}$) was larger than the number of samples N. In that situation, high variance and overfitting are usually the major concern (Hastie *et al.*, 2009). To reduce the dimensionality of the problem, an unsupervised filtering step was introduced before the training phase. Since the behaviour of a classification algorithm under variable dimensions for a specific problem is not predictable and depends on the characteristics of the data, it was necessary to evaluate the effect of reducing the number of predictors on the performance of each algorithm before the training phase (Figure 4.4).

The use of a combined scheme for the training-testing phase (repeated hold-out and resampling by 10-fold CV, LOOCV and bootstrap, only in the training set) yielded more realistic measures than those obtained using only resampling (Figure 4.6). The high accuracy values obtained when using only resampling, especially in the case of SVM and RF, might be an indicative of overfitting. This behaviour does not depend on the resampling method used since similar results were obtained with the three resampling schemes applied. 10-fold CV, LOOCV and .632+ bootstrap have been suggested as resampling methods when dealing with small datasets in classifiers like kNN or classification trees (Molinaro *et al.*, 2005; Kim, 2009). Besides that, the .632+ bootstrap estimator is in general known to have better performance for small samples because of its small variance (Efron, 1983; Efron and Tibshirani, 1997). That was observed also here in the evaluation of the effect of reducing the number of features on the classification error (Figure 4.4).
In the case of RF, there is no explicit need for resampling or a separate test set to get an unbiased estimate of the test set error. It is estimated internally with the out of bag (oob) error estimate (Breiman, 2001). However, when estimating the test error using both methods (the combined approach and the oob estimate), only slightly differences in the results were obtained. Because of that, the combined scheme was used with all the classifiers to have the same number of results in the three cases.

In this study, three assessment parameters were evaluated to select the "best" classifier: performance metrics based on the confusion matrix obtained for every test set, correlation measures, and receiver operating characteristics (ROC) curves.

A classifier with high specificity was preferred over one with high sensitivity. In other words, the focus was to obtain few false positives (in Equations 3.11 and 3.12, specificity and precision are higher if the false positives (FP) decrease) even at the cost of an increased number of false negatives (in Equation 3.9, sensitivity is lower if the false negatives (FN) increase). For the intended application of the classifier to predict potential NES-containing proteins in the complete proteome of Arabidopsis, it was preferable to have fewer predictions but with a low number of false positives, which would bind effort within the subsequent experimental analysis.

Although accuracy is not an appropriate measure in the event of imbalanced data, in this case it was taken as a preliminary indicative of performance since the data set was only slightly imbalanced. The accuracy values obtained for the three classifiers (ranging from around 75% in the case of kNN to around 89% in the case of RF, Figure 4.7), are comparable to the accuracy values obtained for some classifiers designed to predict protein targeting signals (Schneider and Fechner, 2004). Since Matthews correlation coefficient (MCC) and F-score are much better indicators of the performance of a classifier than sole accuracy (specially under imbalanced scenarios), these two measures were also included. In general, a correlation measure reflects the degree of linear relationship between two variables. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications whereas the F-measure correlates precision and recall. While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the MCC is generally regarded as being one of the best such measures. Although there is not a *threshold* value to be considered as "good" or "bad" when developing a classifier, it is accepted that above 0, which means a random prediction, the higher the value the more accurate the classifier. Some typical MCC values reported for classification problems are around 0.3-0.4 (Duan *et al.*, 2008; Liu *et al.*, 2009), 0.5-0.6 (Caragea *et al.*, 2007; Hawkins *et al.*, 2007; Ba *et al.*, 2009), 0.7-0.8 (Garg *et al.*, 2005; Kumar and Raghava, 2009) and for F-score 0.66 (Caragea *et al.*, 2007). The values obtained in this work for MCC were between 0.55 for k-NN and 0.77 for RF, similarly for the F-score values higher than 0.7 were obtained for the three classifiers (Table 4.1). These values might be considered as "high" when compared to other studies and can be taken as an additional indicative of the classifiers performance.

With any classifier, it is possible to make a trade-off between sensitivity and specificity. Hence, it is much more informative to compare the receiver operating

characteristic (ROC) curves, which show the trade-off between true positive and false positive predictions over the entire range of possible values, than to compare the performance of the classifiers for a particular choice of the trade-off, which corresponds to a specific point on the ROC curve (Fawcett, 2004). Furthermore, the construction of ROC curves offers an extra comparison parameter: the area under the ROC curve (AUC). The AUC is an effective means of comparing the overall prediction performance of different methods because it provides a single measure of overall threshold-independent accuracy. For the AUC, a value higher than 0.5 indicates better performance than random. In this case the three classifiers were quite superior with values between 0.8 and 0.9 (Figure 4.8). The differences between the three classifiers observed in the ROCs (Figure 4.8) were less obvious than when comparing the values of the performance metrics (Figure 4.7). Nevertheless, the superiority of the RF classifier was also evident when comparing the ROCs and was confirmed in the comparative performance using the ROC convex hull (ROOCH) method (Figure 4.9).

In the ROOCH method, the classifiers positioned in the left side of the ROC space represent "conservative" classifiers i.e., it is more important to preserve a low false positive rate (FPR) although the true positive rate (TPR) could be not so high (Provost and Fawcett, 2001). That was the case with the RF classifier. On the other hand, classifiers positioned in the right part of the ROC space are considered more "liberal" approaches i.e., it is more important not to leave out any potentially positive sample although the FPR could be also high (Provost and Fawcett, 2001). That was the case with the k-NN classifier, which is consistent with the high sensitivity results obtained for this classifier but it also showed the highest FPR values. The SVM classifier showed a performance close to RF as a whole, which makes this classifier also a good alternative for the problem of NES classification.

The purpose of the evaluation and comparison of the three trained classifiers was to choose one of them for predictions of NES on new protein sequences i.e., not included in the training. Thus, as a result of the evaluation phase the RF classifier was selected as the one with the best performance and hence, it was used for the actual prediction.

Random Forest is an ensemble learning algorithm and is known to be more robust against noise than many no-nensemble learning models. This algorithm has been also successfully used in other recent classification applications (Han *et al.*, 2009; Lee *et al.*, 2009a; Sikić *et al.*, 2009; Liu *et al.*, 2009).

5.3 Analysis of the predicted Arabidopsis LR-NES-containing proteins

In 2000, the entire genome sequence of Arabidopsis was determined and based on sequence conservation with known DNA binding domains Riechmann *et al.* (2000) reported that around 1.500 genes encode transcription factors (TFs). More recent analyses have recognized approximately 2.000 TF genes in the Arabidopsis genome (Davuluri *et al.*, 2003; Iida *et al.*, 2005; Guo *et al.*, 2005; Riaño-Pachón *et al.*, 2007). Based on these and other comparative studies, Mitsuda and Ohme-Takagi (2009) suggested that transcriptional regulation could play more important roles in plants than in animals. Since transcriptional regulation is the first step of gene expression and could affect various "-omes", namely the proteome, metabolome and phenome, the functional analysis of TFs is important and necessary for "omics" studies and for the elucidation of whole functional networks in plants. Taking these considerations in mind, the selection of the NES-containing proteins that were predicted in this work to be tested in the laboratory was targeted towards TFs. Protein sequences associated with the GOs transcription, transcriptional activity, DNA or RNA binding and DNA or RNA metabolism were pre-selected from the total group of the predicted protein sequences (Figures 4.12 and 4.13).
Since TFs play their role in a nuclear process (transcription), it was thought that they were just imported into the nucleus after translation in cytoplasm. However, an important concept that has emerged is the dynamic nature of transport between the nucleus and the cytoplasm, also for TFs. This concept is particularly relevant for understanding how transcription is coordinated with other cellular processes. Cells process a great deal of information from both intracellular and extracellular sensors. This information is conveyed to the nucleus and used, ultimately, to determine the rate of transcription of specific genes. Shuttling of TFs between the nucleus and the cytoplasm and the regulation of nuclear import and nuclear export provides multiple mechanisms to control the actual nuclear abundance of TFs (Figure 5.2).

The prediction of NESs in many TFs and proteins associated with DNA/RNA metabolism and the experimental verification of the nuclear export activity in a group of these proteins gives a clear insight that nucleo-cytoplasmic partitioning is involved in the regulation of TFs in Arabidopsis. The shuttling of TFs and other proteins whose main activity is nuclear has been widely reported, being a notable

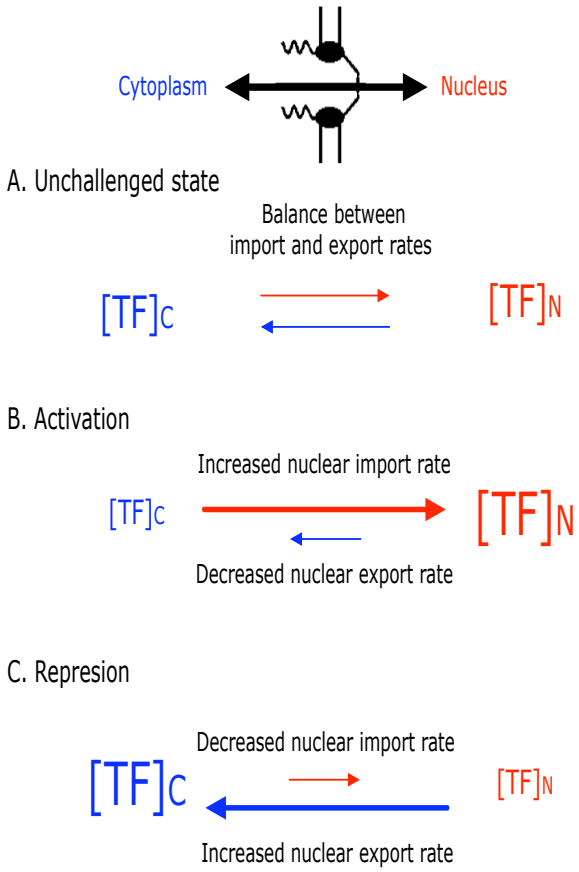


Figure 5.2: Nucleo-cytoplasmic shuttling of transcription factors

Schematic illustration of the net nuclear abundance of transcription factors. In all the cells **A**, the net nuclear abundance of transcription factors is determined by a balance between nuclear import and nuclear export processes. This steady state distribution can be altered by either increasing the rate of nuclear import and decreasing the rate of nuclear export, as shown in **B** or by decreasing the rate of nuclear import and increasing the rate of nuclear export, as shown in **C**. [TF]C, transcription factor concentration in the cytoplasm; [TF]N, transcription factor concentration in the nucleus.

example the case of many tumour suppressor proteins and cell-cycle regulators. Also in plants, some proteins have been reported to be nucleo-cytoplasmic shuttling proteins. For instance, the tomato heat stress TF HsfA2, a shuttling protein with dominant cytoplasmic localization as a result of a nuclear import combined with an efficient export (Heerklotz *et al.*, 2001).

In the present study, from eleven proteins tested and showing a positive interaction with XPO1a in yeast two-hybrid assays, nine correspond to TFs and the other two are proteins related to DNA metabolism. Four of those TFs, AT1G09530, PIF3 (phytocrome interacting factor 3, also called bHLH008); AT2G43010, PIF4 (phy-

tocrome interacting factor 4, also called bHLH009) ; AT5G67110, ALC (alcatraz) and AT5G53210, SPCH (speechless) belong to the family of basic-helix-loop-helix (bHLH) TFs in Arabidopsis. The first three are grouped in the subfamily 15 of the hHLH family, whereas SPCH is located in the subfamily 3 (Toledo-Ortiz *et al.*, 2003).

The Arabidopsis genome codes for more than 150 putative bHLH class TFs (Toledo-Ortiz *et al.*, 2003; Bailey *et al.*, 2003). Interestingly all the bHLH proteins involved in light signalling belong to a single evolutionarily related subclass: the subfamily 15 (Toledo-Ortiz *et al.*, 2003; Bailey *et al.*, 2003; Heim *et al.*, 2003). These bHLH proteins are known as PIF (phytochrome interacting factor) or PIL (phytochrome interacting factor-like) (Yamashino *et al.*, 2003; Khanna *et al.*, 2004) (Figure 5.3). These bHLH class proteins have closely related bHLH domains (Castillon *et al.*, 2007), and most of them carry a small conserved N-terminal domain called active phytochrome binding (APB), which is necessary and sufficient for mediating the interaction with phyB Pfr (maximally absorbing far red form, considered to be the active form for most phytochrome responses) (Khanna *et al.*, 2004; Duek and Fankhauser, 2005) (Figure 5.3).

PIF3 was the first bHLH protein to be identified as a phytochrome-interacting protein (Ni *et al.*, 1998), it binds to phytochromes A and B (phyA and phyB) in a light-dependent manner (Zhu *et al.*, 2000; Khanna *et al.*, 2004). Activation of phytochrome results in PIF3 phosphorylation (Al-Sady *et al.*, 2006) and subsequent degradation (Bauer *et al.*, 2004; Park *et al.*, 2004) in a mechanism that appears to be common to this class of signaling protein (Lorrain *et al.*, 2008; Shen *et al.*, 2008). Although there seems to be broad consensus on what is known about the molecular events after phytochrome interaction with PIF3, there is less certainty about how PIF3 is functioning in photomorphogenesis. This has led to the hypothesis that PIF3 has a dual function, acting early and positively as a transcription factor, but acting later to regulate phyB abundance and repress light induced inhibition of hypocotyl elongation (Monte *et al.*, 2007; Al-Sady *et al.*, 2008). More recently, PIF3, together with PIF1 were proposed to be negative regulators that function to integrate light and circadian control in the regulation of chloroplast development (Stephenson *et al.*, 2009). PIF4 interacts preferentially with phyB and functions negatively in the phyB-mediated inhibition of hypocotyl elongation (Huq, 2006) (Figure 5.3). PhyB and PIF4 both display positive roles in regulating stomatal development in response to light quantity (Casson *et al.*, 2009) and and it has been suggested that PIF4 also plays a role in the circadian clock (Castillon *et al.*, 2007). .

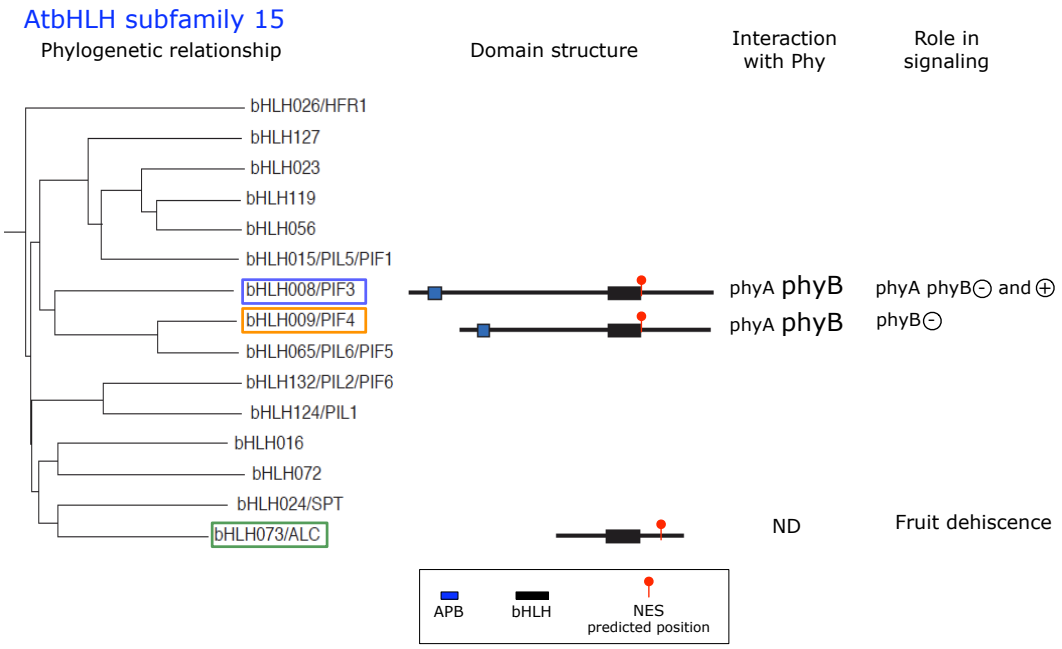


Figure 5.3: The subfamily 15 of bHLH transcription factors in Arabidopsis
A nuclear export signal (NES) was predicted and the nuclear export activity experimentally validated for three transcription factors (TFs) belonging to the subfamily 15 of the bHLH TFs family in Arabidopsis. The complete subfamily 15 is presented as a phylogenetic tree (Duek and Fankhauser, 2005), and the three TFs predicted as containing NES and tested in this work are highlighted. The central panel shows the domain distribution for those proteins including the predicted position for the NESs, the interaction with the phytochromes A (phyA) and B (phyB) is also indicated, as well as the role of the proteins in signalization. The size of the symbols phyA and phyB indicates a weak or strong interaction. The domain APB is a motif necessary and sufficient for binding to the biologically active Pfr form of phyB (Khanna *et al.*, 2004). The figure was adapted from Figure 2 in Duek and Fankhauser (2005).

The third member of the bHLH TF family whose NES was predicted and validated is alcatraz (ALC). This TF also belongs to the subfamily 15 of the bHLH TFs in Arabidopsis but in contrast to the PIFs, no interaction with the phytochomes has been determined and it does not contain an ABP domain (Figure 5.3). The function of ALC has been associated to cell separation during fruit dehiscence (Rajani and Sundaresan, 2001). Similar to PIF3 and PIF4, ALC showed a weak XPO1a interaction activity (Figure 4.14). In ALC the predicted location for the NES is not at the end of the bHLH domain, as is the case for PIF3 and PIF4, but in the C-terminal region of the protein (Figure 5.3).
The presence of a functional NES in these three bHLH TFs potentially enable them to be exported from the nucleus which could have implications for the regulation of their nuclear functions. Nucleo-cytoplasmic shuttling has been also reported

for TFs of the bHLH family in other organisms, for example, the aryl hydrocarbon receptor (AhR) (Kanno *et al.*, 2007) and the protein BMAL1 (Kwon *et al.*, 2006).

The fourth member of the bHLH TF family with nuclear export activity according to this study is SPCH (AT5G53210, speechless), which belongs specifically to the subfamily 3 of TFs in Arabidopsis (Toledo-Ortiz *et al.*, 2003). SPCH is a stomatal regulator that contains a unique MAPK phosphorylation target domain not present even in other TFs closely related that are also involved in stomatal development control (Lampard *et al.*, 2008). The MAPK phosphorylation domain mediates repression of SPCH, therefore it has been proposed as the effector to explain the use of a general mechanism, the MAPKs phosphorylation, in a specific biological event, the stomatal development in general (Lampard *et al.*, 2008) and also in response to fluctuating environmental conditions (Lampard, 2009). In the case of SPCH, the predicted NES is located in the C-terminal end of the protein, downstream from the MAPK phosphorylation target domain. In this putative position it is highly probable that the NES is freely exposed for the interaction with the receptor XPO1.

Another TF for which an NES was predicted and verified in this work is BRZ1 (AT1G75080, brassinazole-resistant 1 (Figure 4.14). This protein has been recently reported to shuttle between nucleus and cytoplasm in Arabidopsis (Gampala *et al.*, 2007; Ryu *et al.*, 2007) and also in rice (Bai *et al.*, 2007), being an example to demonstrate the potential of the nucleo-cytoplasmic partitioning as a regulatory mechanism in plants (Merkle, 2003, 2008). It has been shown that the nucleo-cytoplasmic localization and phosphorylation status of BRZ1 are regulated in a brassinosteroid(BR)-dependent manner (Ryu *et al.*, 2007). The nuclear export of BRZ1 has been associated with phosphorylation and interaction of the phosphorylated forms with 14-3-3 proteins which promote the nuclear export and/or cytosolic retention of BZR1 (Gampala *et al.*, 2007; Ryu *et al.*, 2007). However, the presence of a functional NES (which was predicted in this study in the C-terminal region of the protein) offers the alternative that BRZ1 can be exported from the nucleus by its own, which deserves further analysis.

The homeodomain (HD) TF WOX13 (AT4G35550, **w**uschel-related homeob**ox** 13) showed a strong interaction with the receptor XPO1a in this study (Figure 4.14). The members of the HD family of TF are key regulators implicated in the determination of cell fate and cell differentiation in both plants and animals, however WOX13 belongs to a special class present only in plants. In Angiosperms, a gene called wuschel (WUS) was isolated from many different species. WUS was the

first identified member of the wuschel-related homeobox (WOX) subfamily that is only found in plants (Haecker *et al.*, 2004; Nardmann and Werr, 2006). A recent study (Deveaux *et al.*, 2008) supports the existence of a WOX13 orthologous group (WOX13 OG), containing genes from many different members of the plant kingdom. The function of the WOX13OG Arabidopsis members was linked to organ initiation and development, most likely by preventing premature differentiation as shown for other WOX proteins (Deveaux *et al.*, 2008). Figure 5.4 shows the results presented in Deveaux *et al.* (2008) concerning the domain organization of WOX13, together with the position of the predicted NES of the present study, located at the end of the protein. Also in the case of this TF, the nuclear export activity could implicate that nucleo-cytoplasmic partitioning plays a role in its regulatory functions.

Figure 5.4: The WOX13 protein from Arabidopsis

Motif composition of the WOX13 protein from Arabidopsis as presented in Deveaux *et al.* (2008), including the predicted position of the NES (this study). The motifs are denoted by colored boxes on a bar representing the length of the protein (268 amino acid residues), the position of the predicted NES is indicated in red (position 232). MOG: Main ortologue group.

From the proteins experimentally tested for nuclear export activity in this study, the TF E2F3 (AT2G36010, E2F transcription factor 3) showed the strongest activity (Figure 4.14). It has been shown that the basic regulatory circuits governing cell cycle progression in animal cells are remarkably conserved in higher plants. In particular, plant cells possess all the key components of the cyclin D/retinoblastoma(RBR)/E2F pathway, which in animal cells is a major regulator of cell proliferation and is part of a critical checkpoint controlling the progression from G1 to S phase of the cell cycle (Stals and Inzé, 2001). The Arabidopsis E2F TF family is composed of eight members, six E2Fs (E2FA to E2FF) and two DPs (DPA and DPB) Mariconti *et al.* (2002). This family can be divided into two groups that differ both structurally and functionally. E2FA to E2FC possess all the features of typical E2Fs, including a DNA-binding domain, a marked box, a DP heterodimerization domain, a transactivation domain and an RBR-binding region (Figure 5.5). In the case of AtE2F3/AtE2FA, Kosugi and Ohashi (2002)

reported a non exclusive nuclear localization for this protein, therefore, the nuclear export through an NES in AtE2F3/AtE2FA was proposed as one of the mechanisms causing retention of the protein in the cytoplasm.
In the present work, two NESs were predicted for AtE2F3/AtE2FA, one of them would be located inside the DNA binding motif, which functions also as dimerization motif, and the other one after that motif presenting a partial overlap with the DB binding domain (Figure 5.5). The sequence position of these two NESs would be consistent with the fact that when AtE2F3/AtE2FA binds to DB, it localizes mainly in nucleus Kosugi and Ohashi (2002) due to the blockade of one of the NESs. The presence of two NESs in AtE2F3/AtE2FA as well as the sequence, although not the localization, are similar to the NESs in the TF E2F4 from human (Gaubatz *et al.*, 2001). The human E2F4 shuttles between the nucleus and the cytoplasm and has two NESs, one located in the N-terminal region, in front of the DNA binding domain and the other inside that domain (Gaubatz *et al.*, 2001).
AtE2F3/AtE2FA has been implicated in cell division and plant development by assuming a bimodal function in balancing the expression of both positive and negative regulators involved in cell division and growth (He *et al.*, 2004; de Jager *et al.*, 2009; Sozzani *et al.*, 2010). Consistent with these functions, it is expected that AtE2F3/AtE2FA shuttles between the nucleus and the cytoplasm. In that direction, the nuclear export activity of AtE2F3/AtE2FA was probed in this work. It remains to be established if the position of the NES(s) in AtE2F3/AtE2FA corresponds to the predicted ones.

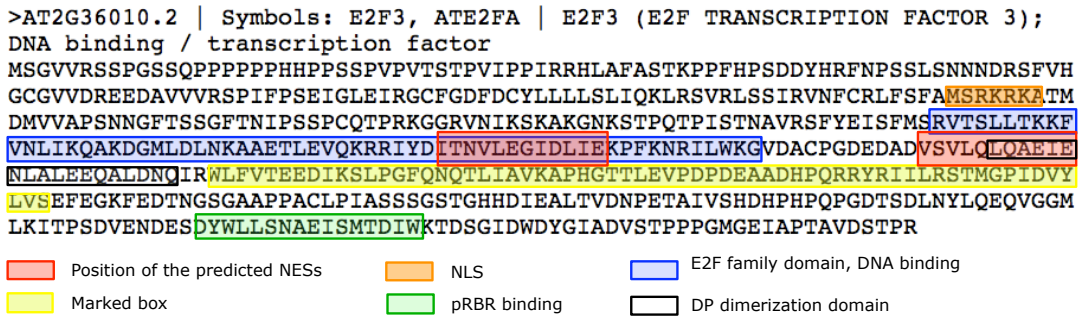


Figure 5.5: Predicted localization of two NESs in AtE2F3
Domain distribution of the protein E2F3 from Arabidopsis (AtE2F3). Each domain is displayed in a color box, which is explained below the sequence. The positions of the two NESs predicted in this work are also shown.

The TF ABF1 (AT1G49720, absidic acid responsive element binding factor1), showed a strong interaction with XPO1a (Figure 4.14). This protein, together

with other ABFs are catalogued as absidic acid (ABA)/stress-inducible TFs belonging to a distinct subfamily of basic leucine zipper (bZIP) class TFs (Choi *et al.*, 2000; Finkelstein *et al.*, 2002). ABFs bind to a *cis*-acting promotor element designated as the ABA-responsive element (ABRE), that is a subset of the G-box sequence. ABF1 has been proposed as a possible redundant mediator of seed or seedling ABA, together with ABF3 (Finkelstein *et al.*, 2005).
Two NESs were predicted in ABF1 (positions 135-142 and 340-347, being the length of the protein 392 amino acid residues), the NES close to the C-terminal overlaps with the position of the four heptad repeats of leucine, a motif that identifies ABF1 as a bZIP protein (Landschulz *et al.*, 1988). Therefore, the experimental definition of the position of the NES in ABF1 could decide whether or not the second predicted NES corresponds to a "false positive" masked with another motif rich in leucine residues but not related with nuclear export. On the other hand, if the two predicted NESs are functional, this could be an interesting example of regulation and/or promotion of the nuclear export activity by cooperative binding of multiple NESs, as has been suggested by some authors (Dong *et al.*, 2009a).

The group of TFs showing nuclear export activity in this study includes also the TF TOE2 (AT5G60120, target of early activation tagged (EAT) 2). This TF belongs to the Apetala 2-like (AP2)-ethylene responsive element binding protein (EREBP) family of TFs, which is present only in plants (Mitsuda and Ohme-Takagi, 2009). The TOE2 gene is one of the targets of the microRNA miR172 (Aukerman and Sakai, 2003), and together with other factors (like TOE1) is involved in the timing of flowering in Arabidopsis. To date, there is no evidence of nuclear export for this TF, but the verification of the NES predicted could be an indicative of nucleo-cytoplasmic shuttling which should be investigated further.

In addition to TFs, two proteins related with DNA metabolism that were predicted as NES-containing showed also nuclear export activity. The protein RHL2 (AT5G02820, root hairless 2) interacted strongly with XPO1a in the yeast two-hybrid assay (Figure 4.14). Depending on the genetic screening, this gene has lead to the identification of the proteins root hairless 2 (RHL2) (Schneider *et al.*, 1997), AtSPO11-3 (Hartung and Puchta, 2001) and brassinosteroid insensitive 5 (BIN5) (Yin *et al.*, 2002). The Spo11 protein is an eukaryotic homologue of the archaeal DNA topoisomerase VIA subunit (topo VIA). In archaea it is involved, together with its B subunit (topo VIB), in DNA replication. However, most eukaryotes, including yeasts, insects and vertebrates have a single gene encoding Spo11/topo

VIA and contain no homologues for topo VIB. In these organisms, Spo11 mediates DNA double-strand breaks that initiate meiotic recombination. Plants are the exception to this rule, as many plant genomes, including Arabidopsis and rice, code for three Spo11 homologues (Grelon *et al.*, 2001; Hartung and Puchta, 2001, 2000). The three Spo11 homologues in Arabidopsis have two very discrete functions. Both AtSPO11-1 and AtSPO11-2 play key roles in meiotic recombination (Grelon *et al.*, 2001; Stacey *et al.*, 2006), while AtSPO11-3/RHL2/BIN5, is involved in DNA endoreduplication (Hartung *et al.*, 2002; Sugimoto-Shirasu *et al.*, 2002; Corbett and Berger, 2003), a common process in eukaryotes that involves DNA amplification without corresponding cell divisions (Edgar and Orr-Weaver, 2001). A classical NLS was functionally identified at the N-terminal of AtSPO11-3/RHL2/BIN5, which was in agreement with a diffuse, but not exclusive, nuclear localization of the protein (Sugimoto-Shirasu *et al.*, 2002). In that sense, a non-exclusive nuclear localization of this protein could be explained by the nuclear export activity due to the predicted NES. As in the other proteins, the localization of the NES in the protein remains to be confirmed.

The protein PSF (AT1G80190, partner of SLD five 1) is to date, an uncharacterized protein identified in a genome-wide analysis of the core DNA replication machinery in plants as one of the components of the GINS complex (Shultz *et al.*, 2007). In eukaryotes, the GINS complex (the name is an acronym for go-ichi-ni-san, the Japanese for 5-1-2-3, after the four subunits of the complex Sld5, Psf1, Psf2 and Psf3) was recently identified as a novel factor essential for both the initiation and elongation stages of the replication process (MacNeill, 2010). The predicted NES is located in the middle of the protein (amino acid 103-112 from 201 residues) in a region predicted to adopt a helix conformation. Since the characterization of this protein is still at the beginning, its potential nucleo-cytoplasmic partitioning merits to be investigated further.

CHAPTER 6

Conclusions and outlook

The foremost contribution of this work is the development of an accurate tool for predicting NESs in proteins of *Arabidopsis thaliana* based on a Random Forest classifier. This conclusion is based on two facts. First, the results of the performance assessment during the classifier selection procedure were very promising (Section 4.1.4). Second, the experimental verification of the nuclear export activity in a selected group from the total of predicted proteins 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 has important implications because it 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 *A. thaliana* 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 the potential of the nucleo-cytoplasmic partitioning as a regulation mechanism in Arabidopsis. Moreover, the experimental validation of the nuclear export activity in a group of selected proteins, mainly transcription factors, corroborate such a potential.
Furthermore, this work addressed additional points concerning the nuclear export activity of some proteins of Arabidopsis. It was shown that although the hydrophobic amino acid residues present in the NES are indispensable for the nuclear export activity, the identity of the spacing residues is also important. Polar and negatively charged residues produced a higher receptor binding activity compared to neutral or positively charged. This observation has been reported in other organisms but not in proteins of Arabidopsis. Similarly, the nuclear export activity of some proteins (like CID 12 in *A. thaliana*) could possibly be modified due to structural constrains of the regions close to the NES.

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 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 performace 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.

APPENDIX A

Oligonucleotide sequences

1	F	GATGAGCGTGAGATGTGTGCAAGAACTATCTACTG	AT1G32790 AT4G10610
2	R	TCTTGCACACATCTCACGCTCATCTTCAGTCCT	AT1G32790 AT4G10610
3	F	GCTTCACGAGCTTCTCTCTAAGCTTAATCCTATGGCT	AT1G32790
4	R	CTTAGAGAGAAGCTCGTGAAGCTCTCTCATATCACGC	AT1G32790
PAB7	F	CCGTGTGGAAGCTAGAAAAAGCAGC	AT2G36660
PAB7	R	GCTTTTTCTAGCTTCCACACGGTTC	AT2G36660
K 609 N	R	GAGCTCGAGTCAGTTAATCGAAACGCCAGCAATGCCAGA	AT2G36660
G 605 S	R	GAGCTCGAGTCACTTAATCGAAACGGAAGCAATGCCAGA	AT2G36660
G 602 Q	F	GTCTCAGATTGCTGGCGTTTCGA	AT2G36660
G 602 Q	R	ACGCCAGCAATCTGAGACGCCAAGTAATCAG	AT2G36660
S 601 E	F	TTGGCGGAGGGCATTGCTGGCGTT	AT2G36660
S 601 E	R	AGCAATGCCCTCCGCCAAGTAATCAGAACG	AT2G36660
EQSN	F	GAGCAGATTGCTTCCGTTTCGATTAACTGACTCGAG	AT2G36660
EQSN	R	GTTAATCGAAACGGAAGCAATCTGCTCCGCCAAGTAATCAGAACG	AT2G36660

Table A.1: Oligonucleotides (I)

Oligonucletides used to amplify the cDNA of the modified versions of the proteins CID 11 (At1g32790), CID 12(At4g10610) and PAB7(At2g36660) (Subsection 4.2.2, Table 4.3 and Figures 4.18, 4.15 and 4.16).
1 and *2* and amplify the cDNA of the C-terminal end of CID 11 and CID 12. *7* and 8 were used in the overlap-extension PCR experiments to obtain the cDNA of the protein referred to in Table 4.3 as *CID 11 NES mut*.
In the case of PAB7 wild type and mutated, the names of the oilgonucleotides correspond to the same used in Figures 4.15 and 4.16.
The second column from the left gives the orientation of each oligonucleotide, F: forward and R: reverse and the last column to the right gives the AGI code for the respective encoded protein.

1	F	ATCGAATTCATGACTATGCGATCATCTTCACCT	AT5G41200
1	R	GAGCTCGAGTTAACAAAAGCCTGCATAACCGTAACC	AT5G41200
2	F	ATCGAATTCATGAATTTCAACGGTTTTCTCGACGACGGT	AT3G61150
2	R	GAGCTCGAGTCAGGTGCTATCACAATGAAGAGCAGC	AT3G61150
3	F	ATCGAATTCATGCCTCTGTTTGAGCTTTTCAGGCTC	AT1G09530
3	R	GAGCTCGAGTCACGACGATCCACAAAACTGATCAG	AT1G09530
4	F	ATCGAATTCATGTCTAGGGAGGATTTTAGTGATACACTTCGAG	AT5G54260
4	R	GAGCTCGAGTTATCTTCTTAGAGCTCCATAGTTCCTTG	AT5G54260
5	F	ATCGAATTCATGGCGGATAAGAAGAAGCGAAAGCGATCA	AT5G02820
5	R	GAGCTCGAGTCAGAGCCAATCCTGCTGCTGCAGTTTCAG	AT5G02820
6	F	ATCGAATTCATGGAACACCAAGGTTGGAGTTTTGAGGAG	AT2G43010
6	R	GAGCTCGAGCTAGTGGTCCAAACGAGAACCGTCGGTG	AT2G43010
7	F	ATCGAATTCATGAGTGACAAAGACGAGTTTGCCGCAAAG	AT1G26260
7	R	GAGCTCGAGCTACGGCTCCACCTTCATGTCAGCTGT	AT1G26260
8	F	ATCGAATTCATGATGGAATGGGATAATCAGCTACAACCCA	AT4G35550
8	R	GAGCTCGAGTCAGCCTGACATGCCATAATCTTCAACATG	AT4G35550
9	F	ATCGAATTCATGGGGTCGTTGCAAATGCAAACAAGTC	AT3G62420
9	R	GAGCTCGAGTCAGCAATCAAACATATCAGCAGAAGCTCTG	AT3G62420
10	F	ATCGAATTCATGTTCGATATGACGCCGAAAAACTCCGA	AT4G21750
10	R	GAGCTCGAGTTAGGCTCCGTCGCAGGCCAGAGCG	AT4G21750
11	F	ATCGAATTCATGGATGTGCCAGAGGAGACGAGGCTTC	AT5G46280
11	R	GAGCTCGAGTCAAATGATATGAACTTTGCCATCGCTG	AT5G46280
12	F	ATCGAATTCATGCAGGAGATAATACCGGATTTTCTTG	AT5G53210
12	R	GAGCTCGAGCTAGCAGAATGTTTGCTGAATTTGTTGAGCC	AT5G53210
13	F	ATCGAATTCATGTACGGGAGAAAAGGGTATCAGCT	AT1G80190
13	R	GAGCTCGAGTCATCCTGTCAGCTCCTCCATTTGG	AT1G80190
14	F	ATCGAATTCATGCTGGATCTCAATCTCGACGTCGACTC	AT5G60120
14	R	GAGCTCGAGCTATGGTGGTGGTTGTGGGCGGTTCATG	AT5G60120
15	F	ATCGAATTCATGTCCGGTGTCGTACGATCTTCTC	AT2G36010
15	R	GAGCTCGAGTCATCTCGGGGTTGAGTCAACAGCTGTTG	AT2G36010
16	F	ATCGAATTCATGGATGAATCAAGTATTATTCCGGCAGAG	AT4G09820
16	R	GAGCTCGAGCTATAGATTAGTATCATGTATTATGACTTGGTGG	AT4G09820
17	F	GAGCTCGAGATGTTCGAGCCAAATATGCTGCTTGCG	AT1G05230
17	R	GAGCTCGAGTCAAGCAGTCTCACAAGACATTGAAGC	AT1G05230
18	F	GAGCTCGAGATGGGGAAGGAAAATGCTGTGTCTCGG	AT1G15570
18	R	GAGCTCGAGTCAGAATAGCGTGTCAAGTAGCTTTGG	AT1G15570
19	F	GAGCTCGAGATGCTTCGTGGTGGGACAACTCTTGTTG	AT1G31360
19	R	GAGCTCGAGTCAAGCTTCTTCTTCTCCGCTGCTCTC	AT1G31360
20	F	ATCGAATTCATGGGTGATTCTGACGTCGGTGATCGT	AT5G67110
20	R	GATGAATTCTCAAAGCAGAGTGGCTGTGGAAAAGCA	AT5G67110
21	F	GAGCTCGAGATGGGTACTCACATTGATATCAACAACTTAGGC	AT1G49720
21	R	GAGCTCGAGTTACCACGGACCGGTAAGGGTTCTTCTC	AT1G49720
22	F	ATCGAATTCATGGAGGAGAGTGAAACAAAGGGGAGAATC	AT4G29940
22	R	GATGAATTCTTACCTTTTCTCCTTGATCTCTGCTATTGGA	AT4G29940
23	F	GAGCTCGAGATGGCGGTGAGTTTGACGGAGGGAGTG	AT5G45400
23	R	GAGCTCGAGTCAGTAGCTGCCTACATGTTGCCTTG	AT5G45400
24	F	ATCGAATTCATGACTTCGGATGGAGCTACGTCGACA	AT1G75080
24	R	GATGAATTCTCAACCACGAGCCTTCCCATTTCCAAG	AT1G75080

Table A.2: Oligonucleotides (II)

Oligonucletides used to amplify the cDNA of the predicted NES-containing proteins that were tested in the laboratory for nuclear export activity. The numbers in the left column correspond to the same used in Table 4.2 and Figure 4.14 (Section 4.2.1). The second column from the left gives the orientation of each oligonucleotide, F: forward and R: reverse. The last column to the right contains the AGI code for the respective encoded protein. Sequences are given in direction 5' to 3'.

List of Abbreviations

ABA	absidic acid
ABF	absidic acid responsive element binding factor
ABRE	ABA-responsive element
ACC	accuracy
AD	activation domain
APB	active phytochrome binding
AUC	area under the curve
bHLH	basic-helixloophelix
BME	β-Mercaptoetanol
CAS	cellular apoptosis susceptibility
CBC	cap binding complex
cDNA	coding DNA
CID	CTC interacting domain
CRM	chromosome region maintenance
CTC	C-terminal domain
CV	cross validation
DAG	directed acyclic graph
DP	dimerization parter
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNA-BD	DNA-binding domain
dNTP	deoxynucleic triphosphate
EAT	early activation tagged
EDTA	ethylene diamine tetraacetic acid
ELISA	enzyme-linked immunosorbent assay
EREBP	ethylene responsive element binding protein
FG	phenylalanine glycine repeats
FPR	false positive rate

GINS	go-ichi-ni-san
GDP	guanosine-5'-diphosphate
GO	Gene Ontology
GTP	guanosine-5'-triphosphate
HD	homeodomain
HIV	human immunodeficiency virus
HMM	Hidden Markov Model
HWU^-	w/o histidine, tryptophan and uracil
k-NN	k-nearest neighbors
LMB	leptomycin B
LOOCV	leave-one-out CV
LR-NES	leucine-rich nuclear export signal
MAPK	mitogen-activated protein kinase
MCC	Matthews correlation coefficient
MOG	main ortologue group
mRNA	messenger ribonucleic acid
NE	nuclear envelope
NES	nuclear export signal
NLS	nuclear localization signal
NPC	nuclear pore complex
Nup	nucleoporin
ONP	ortho-nitrophenol
ONPG	ortho-nitrophenyl-β-D-galactopyranoside
PABP	poly(A)-binding protein
PABPC	cytoplasmic PABP
PABPN	nuclear PABP
PCR	polymerase chain reaction
PEG	polyethylene glycol
PHAX	phosphorylated adaptor for RNA export
PIF	phytochrome interacting factor
PIL	phytochrome interacting factor-like
PKA	phosphokinase A
PKI	protein kinase inhibitor
Ran	Ras-related nuclear protein
RanBP	Ran-binding protein
RBF	radial basis function
Rev	regulator of virion
RF	Random Forest
RNA	ribonucleic acid
ROC	receiver operating characteristic

ROCCH	receiver operating characteristic convex hull
RRE	rev responsive element
RRM	RNA recognition motif
rRNA	ribosomal RNA
snRNA	small nuclear RNA
SVM	Support Vector Machine
TAIR	The Arabidopsis Information Resource
TF	transcription factor
TOE	target of EAT
TPR	true positive rate; sensitivity
tRNA	transfer ribonucleic acid
UsnRNP	uridine-rich small nuclear ribonucleoprotein particle
WOX	wuschel-related homeobox
XPO	exportin
Y2H	yeast-2-hybrid
YNB	yeast nitrogen base

Bibliography

Ach R. A., Gruissem W.: A small nuclear GTP-binding protein from tomato suppresses a schizosaccharomyces pombe cell-cycle mutant. *Proc Natl Acad Sci USA*, 91(13):5863–7, (1994). 26

Adachi Y., Yanagida M.: Higher order chromosome structure is affected by cold-sensitive mutations in a Schizosaccharomyces pombe gene crm1+ which encodes a 115-kD protein preferentially localized in the nucleus and its periphery. *J Cell Biol*, 108(4):1195–207, (1989). 11

Al-Sady B., Kikis E. A., Monte E., Quail P. H.: Mechanistic duality of transcription factor function in phytochrome signaling. *Proc Natl Acad Sci USA*, 105(6):2232–7, (2008). 110

Al-Sady B., Ni W., Kircher S., Schäfer E., Quail P. H.: Photoactivated phytochrome induces rapid PIF3 phosphorylation prior to proteasome-mediated degradation. *Mol Cell*, 23(3):439–46, (2006). 110

Alt J. R., Cleveland J. L., Hannink M., Diehl J. A.: Phosphorylation-dependent regulation of cyclin D1 nuclear export and cyclin D1-dependent cellular transformation. *Genes Dev*, 14(24):3102–14, (2000). 24

Altieri D. C.: Targeted therapy by disabling crossroad signaling networks: the survivin paradigm. *Mol Cancer Ther*, 5(3):478–82, (2006). 19

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.: Basic local alignment search tool. *J. Mol. Biol*, 215:403–410, (1990). 46

Arts G. J., Fornerod M., Mattaj I. W.: Identification of a nuclear export receptor for tRNA. *Curr Biol*, 8(6):305–14, (1998a). 8

Arts G.-J., Kuersten S., Romby P., Ehresmann B., Mattaj I.-W.: The role of exportin-t in selective nuclear export of mature tRNAs. *EMBO J*, 17(24):7430–41, (1998b). 8

Askjaer P., Bachi A., Wilm M., Bischoff F. R., Weeks D. L., Ogniewski V., Ohno M., Niehrs C., Kjems J., Mattaj I. W., Fornerod M.: RanGTP-regulated interactions of CRM1 with nucleoporins and a shuttling DEAD-box helicase. *Mol Cell Biol*, 19(9):6276–85, (1999). 13, 17, 18

Askjaer P., Galy V., Hannak E., Mattaj I. W.: Ran GTPase cycle and importins alpha and beta are essential for spindle formation and nuclear envelope assembly in living caenorhabditis elegans embryos. *Mol Biol Cell*, 13(12):4355–70, (2002). 9

Aukerman M. J., Sakai H.: Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, 15(11):2730–41, (2003). 115

Ba A. N. N., Pogoutse A., Provart N., Moses A. M.: NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics*, 10:202, (2009). 28, 106

Bai M.-Y., Zhang L.-Y., Gampala S. S., Zhu S.-W., Song W.-Y., Chong K., Wang Z.-Y.: Functions of OsBZR1 and 14-3-3 proteins in brassinosteroid signaling in rice. *Proc Natl Acad Sci USA*, 104(34):13839–44, (2007). 112

Bailey P. C., Martin C., Toledo-Ortiz G., Quail P. H., Huq E., Heim M. A., Jakoby M., Werber M., Weisshaar B.: Update on the basic helix-loop-helix transcription factor gene family in Arabidopsis thaliana. *Plant Cell*, 15(11):2497–502, (2003). 110

Baldi P., Brunak S.: *Bioinformatics, the machine learning approach*. MIT Pr., Cambridge, Mass. [u.a.] (2001). 38

Baldi P., Brunak S., Chauvin Y., Andersen C. A., Nielsen H.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–24, (2000). 51, 52

Barrett C., Hughey R., Karplus K.: Scoring Hidden Markov Models. *Comput Appl Biosci*, 13(2):191–9, (1997). 40

Bauer D., Viczián A., Kircher S., Nobis T., Nitschke R., Kunkel T., Panigrahi K. C. S., Adám E., Fejes E., Schäfer E., Nagy F.: Constitutive photomorphogenesis 1 and multiple photoreceptors control degradation of phytochrome interacting factor 3, a transcription factor required for light signaling in Arabidopsis. *Plant Cell*, 16(6):1433–45, (2004). 110

Begitt A., Meyer T., van Rossum M., Vinkemeier U.: Nucleocytoplasmic translocation of Stat1 is regulated by a leucine-rich export signal in the coiled-coil domain. *Proc Natl Acad Sci USA*, 97(19):10418–23, (2000). 18

Belostotsky D. A.: Unexpected complexity of poly(A)-binding protein gene families in flowering plants: three conserved lineages that are at least 200 million years old and possible auto- and cross-regulation. *Genetics*, 163(1):311–9, (2003). 89

Ben-Hur A., Ong C. S., Sonnenburg S., Schölkopf B., Rätsch G.: Support Vector Machines and kernels for computational biology. *PLoS Comput Biol*, 4(10):e1000173, (2008). 37

Bendtsen J. D., Nielsen H., von Heijne G., Brunak S.: Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–95, (2004). 29

Benzeno S., Diehl J.-A.: C-terminal sequences direct cyclin D1-CRM1 binding. *J Biol Chem*, 279(53):56061–6, (2004). 22

Berardini T., Mundodi S., Reiser R., Huala E., Garcia-Hernandez M., Zhang P., Mueller L., Yoon J., Doyle A., Lander G., Moseyko N., Yoo D., Xu I., Zoeckler B., Montoya M., Miller N., Weems D., Rhee S.: Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):1–11, (2004). 55

Blanvillain R., Boavida L. C., McCormick S., Ow D. W.: Exportin1 genes are essential for development and function of the gametophytes in Arabidopsis thaliana. *Genetics*, 180(3):1493–500, (2008). 25

Blum A.-L., Langley P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, (1997). 48

Boche I., Fanning E.: Nucleocytoplasmic recycling of the nuclear localization signal receptor alpha subunit in vivo is dependent on a nuclear export signal, energy, and RCC1. *J Cell Biol*, 139(2):313–25, (1997). 16

Bock J. R., Gough D. A.: Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–60, (2001). 29

Bogerd H. P., Fridell R. A., Benson R. E., Hua J., Cullen B. R.: Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel in vivo randomization-selection assay. *Mol Cell Biol*, 16(8):4207–14, (1996). 16, 104

Bollman K. M., Aukerman M. J., Park M.-Y., Hunter C., Berardini T. Z., Poethig R. S.: HASTY, the arabidopsis ortholog of exportin 5/MSN5, regulates phase change and morphogenesis. *Development*, 130(8):1493–504, (2003). 8, 25

Boser B. E., Guyon I. M., Vapnik V. N.: A training algorithm for optimal margin classifiers. In: *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, ACM, New York, NY, USA (1992). 36

Brameier M., Krings A., MacCallum R. M.: NucPred–predicting nuclear localization of proteins. *Bioinformatics*, 23(9):1159–60, (2007). 28, 29

Bravo J., Aguilar-Henonin L., Olmedo G., Guzmán P.: Four distinct classes of proteins as interaction partners of the PABC domain of Arabidopsis thaliana Poly(A)-binding proteins. *Mol Genet Genomics*, 272(6):651–65, (2005). 90, 93

Breeuwer M., Goldfarb D.: Facilitated nuclear transport of histone H1 and other small nucleophilic proteins. *Cell*, 60(6):999–1008, (1990). 8

Breiman L.: *Classification and regression trees*. Wadsworth, Belmont, Calif. (1984). 34

Breiman L.: Bagging predictors. *Machine Learning*, 24:123–140, (1996). 35

Breiman L.: Random Forests. *Machine Learning*, 45(5-32):1–28, (2001). 34, 43, 105

Briggs L. J., Johnstone R. W., Elliot R. M., Xiao C. Y., Dawson M., Trapani J. A., Jans D. A.: Novel properties of the protein kinase CK2-site-regulated nuclear- localization sequence of the interferon-induced nuclear factor IFI 16. *Biochem J*, 353(Pt 1):69–77, (2001). 24

Bu W., Feng Z., Zhang Z., Zhang C.: Prediction of protein (domain) structural classes based on amino-acid index. *Eur J Biochem*, 266(3):1043–9, (1999). 104

Caragea C., Sinapov J., Silvescu A., Dobbs D., Honavar V.: Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, 8:438, (2007). 29, 106

Carmody S. R., Wente S. R.: mRNA nuclear export at a glance. *Journal of Cell Science*, 122(12):1933–1937, (2009). 12

Casson S. A., Franklin K. A., Gray J. E., Grierson C. S., Whitelam G. C., Hetherington A. M.: phytochrome B and PIF4 regulate stomatal development in response to light quantity. *Curr Biol*, 19(3):229–34, (2009). 110

Castanotto D., Lingeman R., Riggs A. D., Rossi J. J.: CRM1 mediates nuclear-cytoplasmic shuttling of mature microRNAs. *Proc Natl Acad Sci USA*, 106(51):21655–9, (2009). 12

Castillon A., Shen H., Huq E.: Phytochrome interacting factors: central players in phytochrome-mediated light signaling networks. *Trends Plant Sci*, 12(11):514–21, (2007). 110

Chen P.-H., Lin C.-J., Schölkopf B.: A tutorial on Support Vector Machines: Research articles. *Appl. Stoch. Model. Bus. Ind.*, 21(2):111–136, (2005). 37

Chenna R., Sugawara H., Koike T., Lopez R., Gibson T., Higgins D., Thompson J.: Multiple sequence alignment with the CLUSTAL series of programs. *Nucleic Acids Res*, 31:3497–3500, (2003). 45, 46

Chi N. C., Adam E. J., Adam S. A.: Sequence and characterization of cytoplasmic nuclear protein import factor p97. *J Cell Biol*, 130(2):265–74, (1995). 10

Choi H., Hong J., Ha J., Kang J., Kim S. Y.: ABFs, a family of ABA-responsive element binding factors. *J Biol Chem*, 275(3):1723–30, (2000). 115

Chu C. T., Plowey E. D., Wang Y., Patel V., Jordan-Sciutto K. L.: Location, location, location: altered transcription factor trafficking in neurodegeneration. *J Neuropathol Exp Neurol*, 66(10):873–83, (2007). 6

Ciufo L. F., Brown J. D.: Nuclear export of yeast signal recognition particle lacking Srp54p by the Xpo1p/Crm1p NES-dependent pathway. *Curr Biol*, 10(20):1256–64, (2000). 12

Cokol M., Nair R., Rost B.: Finding nuclear localization signals. *EMBO Rep*, 1(5):411–5, (2000). 28

Cole C. N., Scarcelli J. J.: Unravelling mRNA export. *Nat Cell Biol*, 8(7):645–7, (2006). 12

Cook A., Bono F., Jinek M., Conti E.: Structural biology of nucleocytoplasmic transport. *Annu Rev Biochem*, 76:647–71, (2007). 28

Corbett A. H., Silver P. A.: Nucleocytoplasmic transport of macromolecules. *Microbiol Mol Biol Rev*, 61(2):193–211, (1997). 7

Corbett K. D., Berger J. M.: Emerging roles for plant topoisomerase VI. *Chem Biol*, 10(2):107–11, (2003). 116

Cover T., Hart P.: Nearest neighbor patter classification. *IEEE Transactions*, 13:21–27, (1967). 33

Craig E., Zhang Z.-K., Davies K. P., Kalpana G. V.: A masked NES in INI1/hSNF5 mediates hCRM1-dependent nuclear export: implications for tumorigenesis. *EMBO J*, 21(1-2):31–42, (2002). 21

Cronshaw J. M., Krutchinsky A. N., Zhang W., Chait B. T., Matunis M. J.: Proteomic analysis of the mammalian nuclear pore complex. *J Cell Biol*, 158(5):915–27, (2002). 7

Crooks G. E., Hon G., Chandonia J.-M., Brenner S. E.: WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–90, (2004). 45, 99

Dasso M.: The Ran GTPase: theme and variations. *Curr Biol*, 12(14):R502–8, (2002). 14

Davuluri R. V., Sun H., Palaniswamy S. K., Matthews N., Molina C., Kurtz M., Grotewold E.: AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, 4:25, (2003). 108

de Jager S. M., Scofield S., Huntley R. P., Robinson A. S., den Boer B. G. W., Murray J. A. H.: Dissecting regulatory pathways of G1/S control in Arabidopsis: common and distinct targets of CYCD3;1, E2Fa and E2Fc. *Plant Mol Biol*, 71(4-5):345–65, (2009). 114

Deveaux Y., Toffano-Nioche C., Claisse G., Thareau V., Morin H., Laufs P., Moreau H., Kreis M., Lecharny A.: Genes of the most conserved WOX clade in plants affect root and flower development in Arabidopsis. *BMC Evol Biol*, 8:291, (2008). 113

Diehl J. A., Cheng M., Roussel M. F., Sherr C. J.: Glycogen synthase kinase-3beta regulates cyclin D1 proteolysis and subcellular localization. *Genes Dev*, 12(22):3499–511, (1998). 24

Dong X., Biswas A., Chook Y.-M.: Structural basis for assembly and disassembly of the CRM1 nuclear export complex. *Nat Struct Mol Biol*, 16(5):558–60, (2009a). 100, 101, 103, 115

Dong X., Biswas A., Süel K. E., Jackson L. K., Martinez R., Gu H., Chook Y. M.: Structural basis for leucine-rich nuclear export signal recognition by CRM1. *Nature*, 458(7242):1136–1141, (2009b). 100

Duan M., Huang M., Ma C., Li L., Zhou Y.: Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures. *Protein Sci*, 17(9):1505–12, (2008). 106

Duda R. O., Hart P. E., Stork D. G.: *Pattern classification*. Wiley, New York [u.a.] (2001). 29, 32

Duek P. D., Fankhauser C.: bHLH class transcription factors take centre stage in phytochrome signalling. *Trends Plant Sci*, 10(2):51–4, (2005). 110, 111

Durbin R.: *Biological sequence analysis, probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press, Cambridge [u.a.] (2006). 38, 39

Dworetzky S. I., Feldherr C. M.: Translocation of RNA-coated gold particles through the nuclear pores of oocytes. *J Cell Biol*, 106(3):575–84, (1988). 7

Eddy S. R.: Hidden Markov Models. *Curr Opin Struct Biol*, 6(3):361–5, (1996). 38

Eddy S. R.: Profile Hidden Markov Models. *Bioinformatics*, 14(9):755–63, (1998). 38, 39, 40, 46, 104

Eddy S. R.: What is a Hidden Markov Model? *Nat Biotechnol*, 22(10):1315–6, (2004). 38

Edgar B. A., Orr-Weaver T. L.: Endoreplication cell cycles: more for less. *Cell*, 105(3):297–306, (2001). 116

Efron B.: Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, (1983). 31, 32, 105

Efron B., Tibshirani R.: Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, (1997). 32, 105

Ems-McClung S. C., Zheng Y., Walczak C. E.: Importin alpha/beta and RanGTP regulate XCTK2 microtubule binding through a bipartite nuclear localization signal. *Mol Biol Cell*, 15(1):46–57, (2004). 9

Enenkel C., Blobel G., Rexach M.: Identification of a yeast karyopherin heterodimer that targets import substrate to mammalian nuclear pore complexes. *J Biol Chem*, 270(28):16499–502, (1995). 11

Engel K., Kotlyarov A., Gaestel M.: Leptomycin B-sensitive nuclear export of MAPKAP kinase 2 is regulated by phosphorylation. *EMBO J*, 17(12):3363–71, (1998). 18

Engelsma D., Bernad R., Calafat J., Fornerod M.: Supraphysiological nuclear export signals bind CRM1 independently of RanGTP and arrest at Nup358. *EMBO J*, 23(18):3643–52, (2004). 16, 17, 18, 101

Engelsma D., Valle N., Fish A., Salomé N., Almendral J. M., Fornerod M.: A supraphysiological nuclear export signal is required for parvovirus nuclear export. *Mol Biol Cell*, 19(6):2544–52, (2008). 18

Englmeier L., Fornerod M., Bischoff F. R., Petosa C., Mattaj I. W., Kutay U.: RanBP3 influences interactions between CRM1 and its nuclear protein export substrates. *EMBO Rep*, 2(10):926–32, (2001). 14

Fawcett T.: ROC graphs : Notes and practical considerations for researchers. Tech. rep., HP Laboratories, MS 1143, 1501 Page Mill Road, Palo Alto CA 94304 (2004). 53, 107

Feldherr C. M.: The nuclear annuli as pathways for nucleocytoplasmic exchanges. *J Cell Biol*, 14:65–72, (1962). 7

Feldherr C. M., Kallenbach E., Schultz N.: Movement of a karyophilic protein through the nuclear pores of oocytes. *J Cell Biol*, 99(6):2216–22, (1984). 7

Ferrigno P., Posas F., Koepp D., Saito H., Silver P. A.: Regulated nucleo/cytoplasmic exchange of HOG1 MAPK requires the importin beta homologs NMD5 and XPO1. *EMBO J*, 17(19):5606–14, (1998). 21

Finkelstein R., Gampala S. S. L., Lynch T. J., Thomas T. L., Rock C. D.: Redundant and distinct functions of the ABA response loci ABA-insensitive(ABI)5 and ABRE-binding factor (ABF)3. *Plant Mol Biol*, 59(2):253–67, (2005). 115

Finkelstein R. R., Gampala S. S. L., Rock C. D.: Abscisic acid signaling in seeds and seedlings. *Plant Cell*, 14 Suppl:S15–45, (2002). 115

Fischer U., Huber J., Boelens W., Mattaj I., RLührmann: The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell*, 82(3):475–83, (1995). 15, 16, 17

Fornerod M., Ohno M.: Exportin-mediated nuclear export of proteins and ribonucleoproteins. *Results Probl Cell Differ*, 35:67–91, (2002). 16, 18

Fornerod M., Ohno M., Yoshida M., Mattaj I. W.: CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell*, 90(6):1051–60, (1997a). 11, 16

Fornerod M., van Deursen J., van Baal S., Reynolds A., Davis D., Murti K. G., Fransen J., Grosveld G.: The human homologue of yeast CRM1 is in a dynamic subcomplex with CAN/Nup214 and a novel nuclear pore component Nup88. *EMBO J*, 16(4):807–16, (1997b). 11

Frey S., Görlich D.: A saturated FG-repeat hydrogel can reproduce the permeability properties of nuclear pore complexes. *Cell*, 130(3):512–23, (2007). 14

Fridell R. A., Fischer U., Lührmann R., Meyer B. E., Meinkoth J. L., Malim M. H., Cullen B. R.: Amphibian transcription factor IIIA proteins contain a sequence element functionally equivalent to the nuclear export signal of human immunodeficiency virus type 1 Rev. *Proc Natl Acad Sci USA*, 93(7):2936–40, (1996). 12

Fried H., Kutay U.: Nucleocytoplasmic transport: taking an inventory. *Cell Mol Life Sci*, 60(8):1659–88, (2003). 8, 9, 14, 17, 19

Fries T., Betz C., Sohn K., Caesar S., Schlenstedt G., Bailer S. M.: A novel conserved nuclear localization signal is recognized by a group of yeast importins. *J Biol Chem*, 282(27):19292–301, (2007). 10

Fukuda M., Asano S., Nakamura T., Adachi M., Yoshida M., Yanagida M., Nishida E.: CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature*, 390:308–311, (1997). 11, 16

Fukuda M., Gotoh I., Gotoh Y., Nishida E.: Cytoplasmic localization of mitogen-activated protein kinase kinase directed by its NH2-terminal, leucine-rich short amino acid sequence, which acts as a nuclear export signal. *J Biol Chem*, 271(33):20024–8, (1996). 17, 18

Gampala S. S., Kim T.-W., He J.-X., Tang W., Deng Z., Bai M.-Y., Guan S., Lalonde S., Sun Y., Gendron J. M., Chen H., Shibagaki N., Ferl R. J., Ehrhardt D., Chong K., Burlingame A. L., Wang Z.-Y.: An essential role for 14-3-3 proteins in brassinosteroid signal transduction in Arabidopsis. *Dev Cell*, 13(2):177–89, (2007). 112

Garg A., Bhasin M., Raghava G. P. S.: Support Vector Machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem*, 280(15):14427–32, (2005). 29, 106

Gaubatz S., Lees J. A., Lindeman G. J., Livingston D. M.: E2F4 is exported from the nucleus in a CRM1-dependent manner. *Mol Cell Biol*, 21(4):1384–92, (2001). 114

Geisberger R., Rada C., Neuberger M. S.: The stability of AID and its function in class-switching are critically sensitive to the identity of its nuclear-export sequence. *Proc Natl Acad Sci USA*, 106(16):6736–41, (2009). 101

Geles K. G., Johnson J. J., Jong S., Adam S. A.: A role for Caenorhabditis elegans importin IMA-2 in germ line and embryonic mitosis. *Mol Biol Cell*, 13(9):3138–47, (2002). 9

Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y. H., Zhang J.: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, (2004). 43

Gilchrist D., Mykytka B., Rexach M.: Accelerating the rate of disassembly of karyopherin cargo complexes. *J Biol Chem*, 277(20):18161–72, (2002). 11

Goldfarb D.-S., Corbett A.-H., Mason D.-A., Harreman M.-T., Adam S.-A.: Importin alpha: a multipurpose nuclear-transport receptor. *Trends Cell Biol*, 14(9):505–14, (2004). 10

Gorgoni B., Gray N. K.: The roles of cytoplasmic poly(A)-binding proteins in regulating gene expression: a developmental perspective. *Brief Funct Genomic Proteomic*, 3(2):125–41, (2004). 89

Görlich D.: Nuclear protein import. *Curr Opin Cell Biol*, 9(3):412–9, (1997). 8, 11

Görlich D., Kostka S., Kraft R., Dingwall C., Laskey R. A., Hartmann E., Prehn S.: Two different subunits of importin cooperate to recognize nuclear localization signals and bind them to the nuclear envelope. *Curr Biol*, 5(4):383–92, (1995a). 9, 11

Görlich D., Kutay U.: Transport between the cell nucleus and the cytoplasm. *Annu Rev Cell Dev Biol*, 15:607–60, (1999). 6, 7, 8, 15, 45

Görlich D., Vogel F., Mils A., Hartmann E., Laskey R.: Distinct functions for the two importin subunits in nuclear protein import. *Nature*, 377(6546):246–8, (1995b). 9

Grelon M., Vezon D., Gendrot G., Pelletier G.: AtSPO11-1 is necessary for efficient meiotic recombination in plants. *EMBO J*, 20(3):589–600, (2001). 116

Gromiha M. M., Yabuki Y.: Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics*, 9:135, (2008). 29

Gruss O. J., Carazo-Salas R. E., Schatz C. A., Guarguaglini G., Kast J., Wilm M., Bot N. L., Vernos I., Karsenti E., Mattaj I. W.: Ran induces spindle assembly by reversing the inhibitory effect of importin alpha on TPX2 activity. *Cell*, 104(1):83–93, (2001). 9

Guo A., He K., Liu D., Bai S., Gu X., Wei L., Luo J.: DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, 21(10):2568–9, (2005). 108

Haasen D., Köhler C., Neuhaus G., Merkle T.: Nuclear export of proteins in plants: AtXPO1 is the export receptor for leucine-rich nuclear export signals in Arabidopsis thaliana. *Plant J*, 20(6):695–705, (1999a). 25, 26, 27

Haasen D., Merkle T.: Characterization of an Arabidopsis thaliana homologue of the nuclear export receptor CAS by its interaction with importin alpha. *Plant Biol*, 4:432–439, (2002). 25

Haasen D., Neuhaus G., Merkle T.: Isolation and sequence analysis of a genomic clone (accesion no. Y18470) of the nuclear export receptor AtXPO1 (AtCRM1) from Arabidopsis. *Plant Physiol*, 121(1):311, (1999b). 25

Habib T., Zhang C., Yang J. Y., Yang M. Q., Deng Y.: Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. *BMC Genomics*, 9 Suppl 1:S16, (2008). 29

Haché R. J., Tse R., Reich T., Savory J. G., Lefebvre Y. A.: Nucleocytoplasmic trafficking of steroid-free glucocorticoid receptor. *J Biol Chem*, 274(3):1432–9, (1999). 6

Haecker A., Gross-Hardt R., Geiges B., Sarkar A., Breuninger H., Herrmann M., Laux T.: Expression dynamics of WOX genes mark cell fate decisions during early embryonic patterning in Arabidopsis thaliana. *Development*, 131(3):657–68, (2004). 113

Haizel T., Merkle T., Pay A., Fejes E., Nagy F.: Characterization of proteins that interact with the GTP-bound form of the regulatory GTPase Ran in Arabidopsis. *Plant J*, 11(1):93–103, (1997). 26

Han P., Zhang X., Norton R. S., Feng Z.-P.: Large-scale prediction of long disordered regions in proteins using random forests. *BMC Bioinformatics*, 10:8, (2009). 107

Harel A., Forbes D. J.: Importin beta: conducting a much larger cellular symphony. *Mol Cell*, 16(3):319–30, (2004). 8, 9

Hartung F., Blattner F. R., Puchta H.: Intron gain and loss in the evolution of the conserved eukaryotic recombination machinery. *Nucleic Acids Res*, 30(23):5175–81, (2002). 116

Hartung F., Puchta H.: Molecular characterisation of two paralogous SPO11 homologues in Arabidopsis thaliana. *Nucleic Acids Res*, 28(7):1548–54, (2000). 116

Hartung F., Puchta H.: Molecular characterization of homologues of both subunits A (SPO11) and B of the archaebacterial topoisomerase 6 in plants. *Gene*, 271(1):81–6, (2001). 115, 116

Hastie T., Tibshirani R., Friedman J.: *The elements of statistical learning, data mining, inference, and prediction*. Springer, New York, NY (2009). 29, 30, 31, 33, 34, 105

Hawkins J., Davis L., Bodén M.: Predicting nuclear localization. *J Proteome Res*, 6(4):1402–9, (2007). 28, 106

He S. S., Liu J., Xie Z., O'Neill D., Dotson S.: Arabidopsis E2Fa plays a bimodal role in regulating cell division and cell growth. *Plant Mol Biol*, 56(2):171–84, (2004). 114

Heerklotz D., Döring P., Bonzelius F., Winkelhaus S., Nover L.: The balance of nuclear import and export determines the intracellular distribution and function of tomato heat stress transcription factor HsfA2. *Mol Cell Biol*, 21(5):1759–68, (2001). 109

Heger P., Lohmaier J., Schneider G., Schweimer K., Stauber R. H.: Qualitative highly divergent nuclear export signals can regulate export by the competition for transport cofactors in vivo. *Traffic*, 2(8):544–55, (2001). 17, 18, 19, 101

Heim M. A., Jakoby M., Werber M., Martin C., Weisshaar B., Bailey P. C.: The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol Biol Evol*, 20(5):735–47, (2003). 110

Henderson B. R., Eleftheriou A.: A comparison of the activity, sequence specificity, and CRM1-dependence of different nuclear export signals. *Exp Cell Res*, 256(1):213–24, (2000). 17

Herold A., Truant R., Wiegand H., Cullen B. R.: Determination of the functional domain organization of the importin alpha nuclear import factor. *J Cell Biol*, 143(2):309–18, (1998). 11

Ho T. K.: Random decision forest. *Proccedings of the 3rth international conference on document analysis and recognition*, pages 278–282. 35

Ho T. K.: Thwe random subspace method for constructing decision forest. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, (1998). 35

Hogarth C., Itman C., Jans D. A., Loveland K. L.: Regulated nucleocytoplasmic transport in spermatogenesis: a driver of cellular differentiation? *Bioessays*, 27(10):1011–25, (2005). 20

Hood J. K., Silver P. A.: Cse1p is required for export of Srp1p/importin-alpha from the nucleus in Saccharomyces cerevisiae. *J Biol Chem*, 273(52):35142–6, (1998). 11

Hood J. K., Silver P. A.: Diverse nuclear transport pathways regulate cell proliferation and oncogenesis. *Biochim Biophys Acta*, 1471(1):M31–41, (2000). 19

Hua S., Sun Z.: Support Vector Machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721–728, (2001). 29

Hunter C. A., Aukerman M. J., Sun H., Fokina M., Poethig R. S.: PAUSED encodes the Arabidopsis exportin-t ortholog. *Plant Physiol*, 132(4):2135–43, (2003). 25

Huq E.: Degradation of negative regulators: a common theme in hormone and light signaling networks? *Trends Plant Sci*, 11(1):4–7, (2006). 110

Iida K., Seki M., Sakurai T., Satou M., Akiyama K., Toyoda T., Konagaya A., Shinozaki K.: RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res*, 12(4):247–56, (2005). 108

Imamoto N., Shimamoto T., Kose S., Takao T., Tachibana T., Matsubae M., Sekimoto T., Shimonishi Y., Yoneda Y.: The nuclear pore-targeting complex binds to nuclear pores after association with a karyophile. *FEBS Lett*, 368(3):415–9, (1995). 9

Iovine M. K., Wente S. R.: A nuclear export signal in Kap95p is required for both recycling the import factor and interaction with the nucleoporin GLFG repeat regions of Nup116p and Nup100p. *J Cell Biol*, 137(4):797–811, (1997). 16

Ivanciuc O.: Applications of Support Vector Machines in chemistry. *Reviews in computational chemistry*, 23:291–400, (2007). 36

Izaurralde E., Adam S.: Transport of macromolecules between the nucleus and the cytoplasm. *RNA*, 4(4):351–64, (1998). 7

Izaurralde E., Kutay U., von Kobbe C., Mattaj I. W., Görlich D.: The asymmetric distribution of the constituents of the Ran system is essential for transport into and out of the nucleus. *EMBO J*, 16(21):6535–47, (1997). 9

Jäkel S., Albig W., Kutay U., Bischoff F. R., Schwamborn K., Doenecke D., Görlich D.: The importin beta/importin 7 heterodimer is a functional nuclear import receptor for histone H1. *EMBO J*, 18(9):2411–23, (1999). 8

Jans D. A., Xiao C. Y., Lam M. H.: Nuclear targeting signal recognition: a key control point in nuclear transport? *Bioessays*, 22(6):532–44, (2000). 20

Jimenez G. S., Khan S. H., Stommel J. M., Wahl G. M.: p53 regulation by post-translational modification and nuclear retention in response to diverse stresses. *Oncogene*, 18(53):7656–65, (1999). 22

Kaffman A., O'Shea E.-K.: Regulation of nuclear localization: a key to a door. *Annu Rev Cell Dev Biol*, 15:291–339, (1999). 15, 45

Kalderon D., Roberts B. L., Richardson W. D., Smith A. E.: A short amino acid sequence able to specify nuclear location. *Cell*, 39(3 Pt 2):499–509, (1984). 9

Kanai M., Hanashiro K., Kim S.-H., Hanai S., Boulares A. H., Miwa M., Fukasawa K.: Inhibition of Crm1-p53 interaction and nuclear export of p53 by poly(ADP-ribosyl)ation. *Nat Cell Biol*, 9(10):1175–83, (2007). 22

Kanno Y., Miyama Y., Takane Y., Nakahama T., Inouye Y.: Identification of intracellular localization signals and of mechanisms underlining the nucleocytoplasmic shuttling of human aryl hydrocarbon receptor repressor. *Biochem Biophys Res Commun*, 364(4):1026–31, (2007). 112

Kawashima S., Kanehisa M.: AAindex: amino acid index database. *Nucleic Acids Res*, 28374. 47

Kawashima S., Ogata H., Kanehisa M.: AAindex: amino acid index database. *Nucleic Acids Res*, 27:368–369, (1999). 47

Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.: AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*, 36(Database issue):D202, (2008). 47

Khaleghpour K., Kahvejian A., Crescenzo G. D., Roy G., Svitkin Y. V., Imataka H., O'Connor-McCourt M., Sonenberg N.: Dual interactions of the translational repressor Paip2 with poly(A) binding protein. *Mol Cell Biol*, 21(15):5200–13, (2001). 93

Khanna R., Huq E., Kikis E. A., Al-Sady B., Lanzatella C., Quail P. H.: A novel molecular recognition motif necessary for targeting photoactivated phytochrome signaling to specific basic helix-loop-helix transcription factors. *Plant Cell*, 16(11):3033–44, (2004). 110, 111

Kim J.-H.: Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics and Data Analysis*, 53(11):3735–3745, (2009). 105

Kim S.-H., Roux S. J.: An Arabidopsis Ran-binding protein, AtRanBP1c, is a co-activator of Ran GTPase-activating protein and requires the C-terminus for its cytoplasmic localization. *Planta*, 216(6):1047–52, (2003). 26

Kim V.-N.: MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol*, 14(4):156–9, (2004). 12

Klemm J. D., Beals C. R., Crabtree G. R.: Rapid targeting of nuclear proteins to the cytoplasm. *Curr Biol*, 7(9):638–44, (1997). 16

Knauer S. K., Krämer O. H., Knösel T., Engels K., Rödel F., Kovács A. F., Dietmaier W., Klein-Hitpass L., Habtemichael N., Schweitzer A., Brieger J., Rödel C., Mann W., Petersen I., Heinzel T., Stauber R. H.: Nuclear export is essential for the tumor-promoting activity of survivin. *FASEB J*, 21(1):207–16, (2007a). 19

Knauer S. K., Mann W., Stauber R. H.: Survivin's dual role: an export's view. *Cell Cycle*, 6(5):518–21, (2007b). 19

Komeili A., O'Shea E. K.: Roles of phosphorylation sites in regulating activity of the transcription factor Pho4. *Science*, 284(5416):977–80, (1999). 22

Kosugi S., Hasebe M., Tomita M., Yanagawa H.: Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic*, 9(12):2053–62, (2008). 16, 17, 100, 104

Kosugi S., Ohashi Y.: Interaction of the Arabidopsis E2F and DP proteins confers their concomitant nuclear translocation and transactivation. *Plant Physiol*, 128(3):833–43, (2002). 113, 114

Kotera I., Sekimoto T., Miyamoto Y., Saiwaki T., Nagoshi E., Sakagami H., Kondo H., Yoneda Y.: Importin alpha transports CaMKIV to the nucleus without utilizing importin beta. *EMBO J*, 24(5):942–51, (2005). 10

Kraemer D. M., de Castillia C. S., Blobel G., Rout M. P.: The essential yeast nucleoporin NUP159 is located on the cytoplasmic side of the nuclear pore complex and serves in karyopherin-mediated binding of transport substrate. *J Biol Chem*, 270(32):19017–21, (1995). 11

Krätzer F., Rosorius O., Heger P., Hirschmann N., Dobner T., Hauber J., Stauber R. H.: The adenovirus type 5 E1B-55K oncoprotein is a highly active shuttle protein and shuttling is independent of E4orf6, p53 and Mdm2. *Oncogene*, 19(7):850–7, (2000). 18

Krogh A., Brown M., Mian I. S., Sjölander K., Haussler D.: Hidden Markov Models in computational biology. applications to protein modeling. *J Mol Biol*, 235(5):1501–31, (1994). 39

Kudo N., Matsumori N., Taoka H., Fujiwara D., Schreiner E. P., Wolff B., Yoshida M., Horinouchi S.: Leptomycin B inactivates CRM1/exportin 1 by covalent modification at a cysteine residue in the central conserved region. *Proc Natl Acad Sci USA*, 96(16):9112–7, (1999a). 11, 25

Kudo N., Taoka H., Toda T., Yoshida M., Horinouchi S.: A novel nuclear export signal sensitive to oxidative stress in the fission yeast transcription factor Pap1. *J Biol Chem*, 274(21):15151–8, (1999b). 21

Kudo N., Wolff B., Sekimoto T., Schreiner E., Yoneda Y., Yanagida M., Horinouchi S., Yoshida M.: Leptomycin B inhibition of signal-mediated nuclear export by direct binding to CRM1. *Exp Cell Res*, 242(2):540–7, (1998). 11

Kuge S., Arita M., Murayama A., Maeta K., Izawa S., Inoue Y., Nomoto A.: Regulation of the yeast Yap1p nuclear export signal is mediated by redox signal-induced reversible disulfide bond formation. *Mol Cell Biol*, 21(18):6139–50, (2001). 21

Kuhn M.: Building predictive models in R using the caret package. *JSS Journal of Statistical Software*, 28(5):1–26, (2008a). 43

Kuhn M.: *Documentation for package caret version 3.45. [http://caret.r-forge.r-project.org/]* (2008b). 43

Kühn U., Wahle E.: Structure and function of Poly(A) binding proteins. *Biochim Biophys Acta*, 1678(2-3):67–84, (2004). 89

Kumar M., Raghava G. P. S.: Prediction of nuclear proteins using SVM and HMM models. *BMC Bioinformatics*, 10:22, (2009). 29, 106

Kutay U., Bischoff F.-R., Kostka S., Kraft R., Görlich D.: Export of importin alpha from the nucleus is mediated by a specific nuclear transport factor. *Cell*, 90(6):1061–71, (1997). 11

Kutay U., Güttinger S.: Leucine-rich nuclear-export signals: born to be weak. *Trends Cell Biol*, 15(3):121–4, (2005). 16, 18

Kutay U., Lipowsky G., Izaurralde E., Bischoff F., Schwarzmaier P., Hartmann E., Görlich D.: Identification of a tRNA-specific nuclear export receptor. *Mol Cell*, 1(3):359–69, (1998). 8

Kwon I., Lee J., Chang S. H., Jung N. C., Lee B. J., Son G. H., Kim K., Lee K. H.: BMAL1 shuttling controls transactivation and degradation of the CLOCK/BMAL1 heterodimer. *Mol Cell Biol*, 26(19):7318–30, (2006). 112

la Cour T., Gupta R., Rapacki K., Skriver K., Poulsen F.-M., Brunak S.: NESbase version 1.0: a database of nuclear export signals. *Nucleic Acids Res*, 31(1):393–6, (2003). 18, 43, 69, 99

la Cour T., Kiemer L., Mølgaard A., Gupta R., Skriver K., Brunak S.: Analysis and prediction of leucine-rich nuclear export signals. *Protein Eng Des Sel*, 17(6):527–36, (2004). 27, 68, 100, 103

Lam M. H., Briggs L. J., Hu W., Martin T. J., Gillespie M. T., Jans D. A.: Importin beta recognizes parathyroid hormone-related protein with high affinity and mediates its nuclear import in the absence of importin alpha. *J Biol Chem*, 274(11):7391–8, (1999). 9

Lampard G. R.: The missing link?: Arabidopsis SPCH is a MAPK specificity factor that controls entry into the stomatal lineage. *Plant Signal Behav*, 4(5):425–7, (2009). 112

Lampard G. R., Macalister C. A., Bergmann D. C.: Arabidopsis stomatal initiation is controlled by MAPK-mediated regulation of the bHLH speechless. *Science*, 322(5904):1113–6, (2008). 112

Landschulz W. H., Johnson P. F., McKnight S. L.: The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, 240(4860):1759–64, (1988). 115

Lange A., Mills R. E., Lange C. J., Stewart M., Devine S. E., Corbett A. H.: Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J Biol Chem*, 282(8):5101–5, (2007). 10

Lee B. J., Shin M. S., Oh Y. J., Oh H. S., Ryu K. H.: Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome science*, 7:27, (2009a). 29, 107

Lee B.-J., Shin M.-S., Oh Y.-J., Oh H.-S., Ryu K.-H.: Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome science*, 7:27, (2009b). 104

Lei E.-P., Silver P.-A.: Protein and RNA export from the nucleus. *Dev Cell*, 2(3):261–72, (2002). 12

Lei Z., Dai Y.: An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, 6:291, (2005). 29

Li J., Chen X.: PAUSED, a putative exportin-t, acts pleiotropically in Arabidopsis development but is dispensable for viability. *Plant Physiol*, 132(4):1913–24, (2003). 25

Lindsay M., Holaska J., Welch K., Paschal B., Macara I.: Ran-binding protein 3 is a cofactor for Crm1-mediated nuclear protein export. *J Cell Biol*, 153(7):1391–402, (2001). 14

Lipman D. J., Altschul S. F., Kececioglu J. D.: A tool for multiple sequence alignment. *Proc Natl Acad Sci USA*, 86(12):4412–5, (1989). 46

Lipowsky G., Bischoff F. R., Schwarzmaier P., Kraft R., Kostka S., Hartmann E., Kutay U., Görlich D.: Exportin 4: a mediator of a novel nuclear export pathway in higher eukaryotes. *EMBO J*, 19(16):4362–71, (2000). 12

Liu B., Wang X., Lin L., Tang B., Dong Q., Wang X.: Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC Bioinformatics*, 10:381, (2009). 29, 106, 107

Liu J., Vishwanatha J. K.: Regulation of nucleo-cytoplasmic shuttling of human annexin A2: a proposed mechanism. *Mol Cell Biochem*, 303(1-2):211–20, (2007). 23, 24

Lorrain S., Allen T., Duek P. D., Whitelam G. C., Fankhauser C.: Phytochrome-mediated inhibition of shade avoidance involves degradation of growth-promoting bHLH transcription factors. *Plant J*, 53(2):312–23, (2008). 110

Macara I.: Transport into and out of the nucleus. *Microbiol Mol Biol Rev*, 65(4):570–94, table of contents, (2001). 7, 16, 20

MacNeill S. A.: Structure and function of the GINS complex, a key component of the eukaryotic replisome. *Biochem J*, 425(3):489–500, (2010). 116

Mangus D. A., Evans M. C., Jacobson A.: Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol*, 4(7):223, (2003). 89

Mariconti L., Pellegrini B., Cantoni R., Stevens R., Bergounioux C., Cella R., Albani D.: The E2F family of transcription factors from Arabidopsis thaliana. novel and conserved components of the retinoblastoma/E2F pathway in plants. *J Biol Chem*, 277(12):9911–9, (2002). 113

Matsuura Y., Stewart M.: Nup50/Npap60 function in nuclear protein import complex disassembly and importin recycling. *EMBO J*, 24(21):3681–9, (2005). 11

Mattaj I. W., Englmeier L.: Nucleocytoplasmic transport: the soluble phase. *Annu Rev Biochem*, 67:265–306, (1998). 15, 45

Matthews B.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405(2):442–451, (1975). 52

Maurer P., Redd M., Solsbacher J., Bischoff F. R., Greiner M., Podtelejnikov A. V., Mann M., Stade K., Weis K., Schlenstedt G.: The nuclear export receptor Xpo1p forms distinct complexes with NES transport substrates and the yeast Ran binding protein 1 (Yrb1p). *Mol Biol Cell*, 12(3):539–49, (2001). 13

McLachlan G.: *Discriminant analysis and statistical pattern recognition*. Wiley, New York [u.a.] (1992). 31

Meier I.: Nucleocytoplasmic trafficking in plant cells. *International review of cytology: a survey of cell biology*, 244:95–135, (2005). 7, 25

Meier I.: Composition of the plant nuclear envelope: theme and variations. *J Exp Bot*, 58(1):27–34, (2007). 14

Merkle T.: Nuclear import and export of proteins in plants: a tool for the regulation of signalling. *Planta*, 213:499–517, (2001). 27

Merkle T.: Nucleo-cytoplasmic partitioning of proteins in plants: implications for the regulation of environmental and developmental signalling. *Curr Genet*, 44(5):231–60, (2003). 8, 25, 112

Merkle T.: *Plant Cell Monographs: Functional organization of the plant nucleus*, chap. Nuclear export of proteins and RNA. Springer-Verlag Berlin HD NY (2008). 15, 112

Merkle T., Haizel T., Matsumoto T., Harter K., Dallmann G., Nagy F.: Phenotype of the fission yeast cell cycle regulatory mutant pim1-46 is suppressed by a tobacco cDNA encoding a small, Ran-like GTP-binding protein. *Plant J*, 6(4):555–65, (1994). 26

Merkle T., Nagy F.: Nuclear import of proteins: putative import factors and development of in vitro import systems in higher plants. *Trends Plant Sci*, 2(12):458–464, (1997). 26

Miller J.: *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratory, NY (1972). 63

Mingot J.-M., Bohnsack M. T., Jäkle U., Görlich D.: Exportin 7 defines a novel general nuclear export pathway. *EMBO J*, 23(16):3227–36, (2004). 12

Mitsuda N., Ohme-Takagi M.: Functional analysis of transcription factors in Arabidopsis. *Plant and Cell Physiology*, 50(7):1232–48, (2009). 108, 115

Miyamoto Y., Hieda M., Harreman M. T., Fukumoto M., Saiwaki T., Hodel A. E., Corbett A. H., Yoneda Y.: Importin alpha can migrate into the nucleus in an importin beta- and Ran-independent manner. *EMBO J*, 21(21):5833–42, (2002). 10

Molinaro A. M., Simon R., Pfeiffer R. M.: Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–7, (2005). 105

Monecke T., Guttler T., Neumann P., Dickmanns A., Gorlich D., Ficner R.: Crystal structure of the nuclear export receptor CRM1 in complex with Snurportin1 and RanGTP. *Science*, 324(5930):1087–1091, (2009). 100

Monte E., Al-Sady B., Leivar P., Quail P. H.: Out of the dark: how the PIFs are unmasking a dual temporal mechanism of phytochrome signalling. *J Exp Bot*, 58(12):3125–33, (2007). 110

Moore C., Horowitz S.: Nuclear envelope permeability. *Nature*, 13(254):109–14, (1975). 7

Mosapparast N., Pemberton L.: Karyopherins: from nuclear-transport mediators to nuclear-function regulators. *Trends Cell Biol*, 14(10):547–56, (2004). 8, 9

Mutka S. C., Yang W. Q., Dong S. D., Ward S. L., Craig D. A., Timmermans P. B. M. W. M., Murli S.: Identification of nuclear export inhibitors with potent anticancer activity in vivo. *Cancer Research*, 69(2):510–7, (2009). 19

Myers E. W., Miller W.: Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–17, (1988). 46

Nagoshi E., Imamoto N., Sato R., Yoneda Y.: Nuclear import of sterol regulatory element-binding protein-2, a basic helix-loop-helix-leucine zipper (bHLH-Zip)-containing transcription factor, occurs through the direct interaction of importin beta with HLH-Zip. *Mol Biol Cell*, 10(7):2221–33, (1999). 9

Nair R., Carter P., Rost B.: NLSdb: database of nuclear localization signals. *Nucleic Acids Res*, 31(1):397–9, (2003). 28

Nakielny S., Dreyfuss G.: Transport of proteins and RNAs in and out of the nucleus. *Cell*, 99(7):677–90, (1999). 6

Nardmann J., Werr W.: The shoot stem cell niche in angiosperms: expression patterns of WUS orthologues in rice and maize imply major modifications in the course of mono- and dicot evolution. *Mol Biol Evol*, 23(12):2492–504, (2006). 113

Nemeth A., Krause S., Blank D., Jenny A., Jenö P., Lustig A., Wahle E.: Isolation of genomic and cDNA clones encoding bovine poly(A) binding protein II. *Nucleic Acids Res*, 23(20):4034–41, (1995). 89

Ni M., Tepperman J. M., Quail P. H.: PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. *Cell*, 95(5):657–67, (1998). 110

Nigg E. A.: Nucleocytoplasmic transport: signals, mechanisms and regulation. *Nature*, 386(6627):779–87, (1997). 11, 15

Nishi K., Yoshida M., Fujiwara D., Nishikawa M., Horinouchi S., Beppu T.: Leptomycin B targets a regulatory cascade of crm1, a fission yeast nuclear protein, involved in control of higher order chromosome structure and gene expression. *J Biol Chem*, 269(9):6320–4, (1994). 11, 19

Noble W. S.: What is a support vector machine? *Nat Biotechnol*, 24(12):1565–7, (2006). 36

Ohno M., Segref A., Bachi A., Wilm M., Mattaj I.-W.: PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell*, 101(2):187–98, (2000). 12

Ossareh-Nazari B., Gwizdek C., Dargemont C.: Protein export from the nucleus. *Traffic*, 2(10):684–9, (2001). 9, 15, 45

Paraskeva E., Izaurralde E., Bischoff F.-R., Huber J., Kutay U., Hartmann E., Lührmann R., Görlich D.: CRM1-mediated recycling of snurportin 1 to the cytoplasm. *J Cell Biol*, 145(2):255–64, (1999). 13, 16, 18, 100, 103

Park E., Kim J., Lee Y., Shin J., Oh E., Chung W.-I., Liu J. R., Choi G.: Degradation of phytochrome interacting factor 3 in phytochrome-mediated light signaling. *Plant Cell Physiol*, 45(8):968–75, (2004). 110

Pay A., Resch K., Frohnmeyer H., Fejes E., Nagy F., Nick P.: Plant RanGAPs are localized at the nuclear envelope in interphase and associated with microtubules in mitotic cells. *Plant J*, 30(6):699–709, (2002). 27

Pazos F., jung Wook Bang: Computational prediction of functionally important regions in proteins. *Current Bioinformatics*, 1(1):15–23, (2006). 29

Pemberton L.-F., Paschal B.-M.: Mechanisms of receptor-mediated nuclear import and nuclear export. *Traffic*, 6(3):187–198, (2005). 8

Peters A., Hothorn T., Lausen B.: ipred: Improved predictors. *R News* - ***http://CRAN.R-project.org/doc/Rnews/***, 2(2):33–36, (2002). 43, 48, 73

Petosa C., Schoehn G., Askjaer P., Bauer U., Moulin M., Steuerwald U., Soler-López M., Baudin F., Mattaj I. W., Müller C. W.: Architecture of CRM1/Exportin1 suggests how cooperativity is achieved during formation of a nuclear export complex. *Mol Cell*, 16(5):761–75, (2004). 103

Pines J.: Four-dimensional control of the cell cycle. *Nat Cell Biol*, 1(3):E73–9, (1999). 6

Pollard V. W., Malim M. H.: The HIV-1 Rev protein. *Annu Rev Microbiol*, 52:491–532, (1998). 18

Poon I.-K., Jans D.-A.: Regulation of nuclear transport: Central role in development and transformation? *Traffic*, 6(3):173–186, (2005). 6, 20, 23

Popa I., Harris M. E., Donello J. E., Hope T. J.: CRM1-dependent function of a cis-acting RNA export element. *Mol Cell Biol*, 22(7):2057–67, (2002). 12

Provost F., Fawcett T.: Robust classification for imprecise environments. *Machine Learning*, 42:203–231, (2001). 53, 78, 107

Pruess M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E., Mittard V., Mulder N., Phan I., Servant F., Apweiler R.: The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. *Nucleic Acids Res*, 31(1):414–7, (2003). 104

R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2005), ISBN 3-900051-07-0. 43

Radu A., Moore M. S., Blobel G.: The peptide repeat domain of nucleoporin Nup98 functions as a docking site in transport across the nuclear pore complex. *Cell*, 81(2):215–22, (1995). 11

Rajani S., Sundaresan V.: The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. *Curr Biol*, 11(24):1914–22, (2001). 111

Reichelt R., Holzenburg A., Buhle E. L., Jarnik M., Engel A., Aebi U.: Correlation between structure and mass distribution of the nuclear pore complex and of distinct pore complex components. *J Cell Biol*, 110(4):883–94, (1990). 7

Reinhardt A., Hubbard T.: Using neural networks for prediction of the sub-cellular location of proteins. *Nucleic Acid Research*, 26:2230–2236, (1998). 29

Rexach M., Blobel G.: Protein import into nuclei: association and dissociation reactions involving transport substrate, transport factors, and nucleoporins. *Cell*, 83(5):683–92, (1995). 11

Riaño-Pachón D. M., Ruzicic S., Dreyer I., Mueller-Roeber B.: PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, 8:42, (2007). 108

Richardson W. D., Mills A. D., Dilworth S. M., Laskey R. A., Dingwall C.: Nuclear protein migration involves two steps: rapid binding at the nuclear envelope followed by slower translocation through nuclear pores. *Cell*, 52(5):655–64, (1988). 7

Riechmann J. L., Heard J., Martin G., Reuber L., Jiang C., Keddie J., Adam L., Pineda O., Ratcliffe O. J., Samaha R. R., Creelman R., Pilgrim M., Broun P., Zhang J. Z., Ghandehari D., Sherman B. K., Yu G.: Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, 290(5499):2105–10, (2000). 108

Riis S., Krogh A.: Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J Comput Biol*, 3:163–183, (1996). 29

Robbins J., Dilworth S. M., Laskey R. A., Dingwall C.: Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. *Cell*, 64(3):615–23, (1991). 10

Rodriguez M. S., Dargemont C., Stutz F.: Nuclear export of RNA. *Biol Cell*, 96(8):639–55, (2004). 12

Roessiger T.: *Funktionelle Analyse von Poly(A)-bindenden Proteinen in Arabidopsis thaliana*. Master's thesis, Universität Bielefeld- Fakultät für Biologie (2008). 90

Rose A., Meier I.: A domain unique to plant RanGAP is responsible for its targeting to the plant nuclear rim. *Proc Natl Acad Sci USA*, 98(26):15377–82, (2001). 27

Roth D. M., Harper I., Pouton C. W., Jans D. A.: Modulation of nucleocytoplasmic trafficking by retention in cytoplasm or nucleus. *J Cell Biochem*, 107(6):1160–7, (2009). 24

Roth J., Dobbelstein M., Freedman D. A., Shenk T., Levine A. J.: Nucleocytoplasmic shuttling of the hdm2 oncoprotein regulates the levels of the p53 protein via a pathway used by the human immunodeficiency virus rev protein. *EMBO J*, 17(2):554–64, (1998). 18

Rout M. P., Aitchison J. D., Suprapto A., Hjertaas K., Zhao Y., Chait B. T.: The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol*, 148(4):635–51, (2000). 7

Rout M. P., Blobel G.: Isolation of the yeast nuclear pore complex. *J Cell Biol*, 123(4):771–83, (1993). 7

Roy G., Crescenzo G. D., Khaleghpour K., Kahvejian A., O'Connor-McCourt M., Sonenberg N.: Paip1 interacts with poly(A) binding protein through two independent binding motifs. *Mol Cell Biol*, 22(11):3769–82, (2002). 93

Ryan K., Wente S.: The nuclear pore complex: a protein machine bridging the nucleus and cytoplasm. *Curr Opin Cell Biol*, 12:361–371, (2000). 7

Ryu H., Kim K., Cho H., Park J., Choe S., Hwang I.: Nucleocytoplasmic shuttling of BZR1 mediated by phosphorylation is essential in Arabidopsis brassinosteroid signaling. *Plant Cell*, 19(9):2749–62, (2007). 112

Saalbach G., Christov V.: Sequence of a plant cDNA from Vicia faba encoding a novel Ran-related GTP-binding protein. *Plant Mol Biol*, 24(6):969–72, (1994). 26

Sachs A. B., Davis R. W., Kornberg R. D.: A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Mol Cell Biol*, 7(9):3268–76, (1987). 89

Sambrook J., Russell D.: *Molecular Cloning: A Laboratory Mannual*. CSHL press, 3th Edn. (2001). 57

Sammeth M., Rothgänger J., Esser W., Albert J., Stoye J., Harmsen D.: QAlign: quality-based multiple alignments with dynamic phylogenetic analysis. *Bioinformatics*, 19(12):1592–1593, (2003). 45, 46

Sanders W.-S., Bridges S.-M., McCarthy F.-M., Nanduri B., Burgess S.-C.: Prediction of peptides observable by mass spectrometry applied at the experimental set level. *BMC Bioinformatics*, 8 Suppl 7:S23, (2007). 105

Sandri-Goldin R. M.: ICP27 mediates HSV RNA export by shuttling through a leucine-rich nuclear export signal and binding viral intronless RNAs through an RGG motif. *Genes Dev*, 12(6):868–79, (1998). 18

Saporita A. J., Zhang Q., Navai N., Dincer Z., Hahn J., Cai X., Wang Z.: Identification and characterization of a ligand-regulated nuclear export signal in androgen receptor. *J Biol Chem*, 278(43):41998–2005, (2003). 22

Schatz C. A., Santarella R., Hoenger A., Karsenti E., Mattaj I. W., Gruss O. J., Carazo-Salas R. E.: Importin alpha-regulated nucleation of microtubules by TPX2. *EMBO J*, 22(9):2060–70, (2003). 9

Schneider G., Fechner U.: Advances in the prediction of protein targeting signals. *Proteomics*, 4(6):1571–80, (2004). 29, 104, 106

Schneider K., Wells B., Dolan L., Roberts K.: Structural and genetic analysis of epidermal cell differentiation in Arabidopsis primary roots. *Development*, 124(9):1789–98, (1997). 115

Schneider T. D., Stephens R. M.: Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, (1990). 45

Seimiya H., Sawada H., Muramatsu Y., Shimizu M., Ohko K., Yamane K., Tsuruo T.: Involvement of 14-3-3 proteins in nuclear localization of telomerase. *EMBO J*, 19(11):2652–61, (2000). 21

Shen H., Zhu L., Castillon A., Majee M., Downie B., Huq E.: Light-induced phosphorylation and degradation of the negative regulator phytochrome-interacting factor1 from Arabidopsis depend upon its direct physical interactions with photoactivated phytochromes. *Plant Cell*, 20(6):1586–602, (2008). 110

Shultz R. W., Tatineni V. M., Hanley-Bowdoin L., Thompson W. F.: Genome-wide analysis of the core DNA replication machinery in the higher plants Arabidopsis and rice. *Plant Physiol*, 144(4):1697–714, (2007). 116

Shyu A. B., Wilkinson M. F.: The double lives of shuttling mRNA binding proteins. *Cell*, 102(2):135–8, (2000). 6

Sigrist S. J., Thiel P. R., Reiff D. F., Lachance P. E., Lasko P., Schuster C. M.: Postsynaptic translation affects the efficacy and morphology of neuromuscular junctions. *Nature*, 405(6790):1062–5, (2000). 89

Sikić M., Tomić S., Vlahovicek K.: Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput Biol*, 5(1):e1000278, (2009). 107

Sing T., Sander O., Beerenwinkel N., Lengauer T.: ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940, (2005). 43, 53

Smith H. M., Hicks G. R., Raikhel N. V.: Importin alpha from Arabidopsis thaliana is a nuclear import receptor that recognizes three classes of import signals. *Plant Physiol*, 114(2):411–7, (1997). 27

Smith T. F., Waterman M. S.: Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, (1981). 46

Solsbacher J., Maurer P., Bischoff F. R., Schlenstedt G.: Cse1p is involved in export of yeast importin alpha from the nucleus. *Mol Cell Biol*, 18(11):6805–15, (1998). 11

Sorokin A. V., Kim E. R., Ovchinnikov L. P.: Nucleocytoplasmic transport of proteins. *Biochemistry Mosc*, 72(13):1439–57, (2007). 6, 20, 21, 23, 24

Sozzani R., Maggio C., Giordo R., Umana E., Ascencio-Ibañez J. T., Hanley-Bowdoin L., Bergounioux C., Cella R., Albani D.: The E2FD/DEL2 factor is a component of a regulatory network controlling cell proliferation and development in Arabidopsis. *Plant Mol Biol*, 72(4-5):381–95, (2010). 114

Stacey N. J., Kuromori T., Azumi Y., Roberts G., Breuer C., Wada T., Maxwell A., Roberts K., Sugimoto-Shirasu K.: Arabidopsis SPO11-2 functions with SPO11-1 in meiotic recombination. *Plant J*, 48(2):206–16, (2006). 116

Stade K., Ford C., Guthrie C., Weis K.: Exportin 1 (Crm1p) is an essential nuclear export factor. *Cell*, 90(6):1041–50, (1997). 11, 16

Stals H., Inzé D.: When plant cells decide to divide. *Trends Plant Sci*, 6(8):359–64, (2001). 113

Stephenson P. G., Fankhauser C., Terry M. J.: PIF3 is a repressor of chloroplast development. *Proc Natl Acad Sci USA*, 106(18):7654–9, (2009). 110

Stommel J. M., Marchenko N. D., Jimenez G. S., Moll U. M., Hope T. J., Wahl G. M.: A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *EMBO J*, 18(6):1660–72, (1999). 18, 21, 22

Stone M.: Cross-validation and multinomial prediction. *Biometrika*, 61(3):509, (1974). 32

Stone M.: Cross-validation:a review. *Statistics*, 9(1):127–139, (1978). 32

Ström A. C., Weis K.: Importin-beta-like nuclear transport receptors. *Genome Biol*, 2(6):REVIEWS3008, (2001). 8

Stüven T., Hartmann E., Görlich D.: Exportin 6: a novel nuclear export receptor that is specific for profilin.actin complexes. *EMBO J*, 22(21):5928–40, (2003). 12

Sugimoto-Shirasu K., Stacey N. J., Corsar J., Roberts K., McCann M. C.: DNA topoisomerase VI is essential for endoreduplication in Arabidopsis. *Curr Biol*, 12(20):1782–6, (2002). 116

Tago K., Tsukahara F., Naruse M., Yoshioka T., Takano K.: Regulation of nuclear retention of glucocorticoid receptor by nuclear Hsp90. *Mol Cell Endocrinol*, 213(2):131–8, (2004). 24

Tarca A. L., Carey V. J., wen Chen X., Romero R., Drăghici S.: Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, (2007). 29, 35, 36

Terry L. J., Shows E. B., Wente S. R.: Crossing the nuclear envelope: hierarchical regulation of nucleocytoplasmic transport. *Science*, 318(5855):1412–6, (2007). 20, 21, 22

Thakurta A. G., Yoon J. H., Dhar R.: Schizosaccharomyces pombe spPABP, a homologue of Saccharomyces cerevisiae Pab1p, is a non-essential, shuttling protein that facilitates mRNA export. *Yeast*, 19(9):803–10, (2002). 89

The Gene Ontology Consortium: Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, (2000). 54, 83

Thomas F., Kutay U.: Biogenesis and nuclear export of ribosomal subunits in higher eukaryotes depend on the CRM1 export pathway. *Journal of Cell Science*, 116(Pt 12):2409–19, (2003). 17

Thompson J. D., Higgins D. G., Gibson T. J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, (1994). 45, 46

Toledo-Ortiz G., Huq E., Quail P. H.: The Arabidopsis basic/helix-loop-helix transcription factor family. *Plant Cell*, 15(8):1749–70, (2003). 110, 112

Tomii K., Kanehisa M.: Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng*, 9(1):27–36, (1996). 104

Toyoshima F., Moriguchi T., Wada A., Fukuda M., Nishida E.: Nuclear export of cyclin B1 and its possible role in the DNA damage-induced G2 checkpoint. *EMBO J*, 17(10):2728–35, (1998). 18

Tung C.-W., Ho S.-Y.: POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, 23(8):942–9, (2007). 104

Tung C.-W., Ho S.-Y.: Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics*, 9:310, (2008a). 29

Tung C.-W., Ho S.-Y.: Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinformatics*, 9:310–325, (2008b). 105

Turner J., Sullivan D.: CRM1-mediated nuclear export of proteins and drug resistance in cancer. *Current Medicinal Chemistry*, 15:2648–2655, (2008). 19

Vapnik V.: *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA (1995). 36

Vapnik V.: *Statistical Learning Theory*. Wiley-Interscience (September 1998). 36, 37

Vasu S. K., Forbes D. J.: Nuclear pores and nuclear assembly. *Curr Opin Cell Biol*, 13(3):363–75, (2001). 7

Verma R., Tiwari A., Kaur S., Varshney G. C., Raghava G. P.: Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics*, 9:201, (2008). 29

Vernoud V., Horton A. C., Yang Z., Nielsen E.: Analysis of the small GTPase gene superfamily of Arabidopsis. *Plant Physiol*, 131(3):1191–208, (2003). 26, 27

Vigneri P., Wang J. Y.: Induction of apoptosis in chronic myelogenous leukemia cells through nuclear entrapment of BCR-ABL tyrosine kinase. *Nat Med*, 7(2):228–34, (2001). 19

Vousden K. H., Woude G. F.: The ins and outs of p53. *Nat Cell Biol*, 2(10):E178–80, (2000). 19

Wada A., Fukuda M., Mishima M., Nishida E.: Nuclear export of actin: a novel mechanism regulating the subcellular localization of a major cytoskeletal protein. *EMBO J*, 17(6):1635–41, (1998). 18

Wagstaff K. M., Jans D. A.: Importins and beyond: Non-conventional nuclear transport mechanisms. *Traffic*, 10(9):1188–1198, (2009). 10

Wahle E.: A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. *Cell*, 66(4):759–68, (1991). 89

Ward B. M., Lazarowitz S. G.: Nuclear export in plants. use of geminivirus movement proteins for a cell-based export assay. *Plant Cell*, 11(7):1267–76, (1999). 24, 27

Webb A.: *Statistical pattern recognition*. John Wiley & Sons, Ltd (2005). 30

Weis K.: Regulating access to the genome: nucleocytoplasmic transport throughout the cell cycle. *Cell*, 112(4):441–51, (2003). 8, 9, 14

Wen W., Meinkoth J.-L., Tsien R.-Y., Taylor S.-S.: Identification of a signal for rapid export of proteins from the nucleus. *Cell*, 82(3):463–73, (1995). 15, 17, 18

West M., Hedges J. B., Lo K.-Y., Johnson A. W.: Novel interaction of the 60S ribosomal subunit export adapter Nmd3 at the nuclear pore complex. *J Biol Chem*, 282(19):14028–37, (2007). 18

Wolff B., Sanglier J. J., Wang Y.: Leptomycin B is an inhibitor of nuclear export: inhibition of nucleo-cytoplasmic translocation of the human immunodeficiency virus type 1 (HIV-1) Rev protein and Rev-dependent mRNA. *Chem Biol*, 4(2):139–47, (1997). 11

Xia G., Ramachandran S., Hong Y., Chan Y. S., Simanis V., Chua N. H.: Identification of plant cytoskeletal, cell cycle-related and polarity-related proteins using Schizosaccharomyces pombe. *Plant J*, 10(4):761–9, (1996). 26

Yamashino T., Matsushika A., Fujimori T., Sato S., Kato T., Tabata S., Mizuno T.: A link between circadian-controlled bHLH factors and the APRR1/TOC1 quintet in Arabidopsis thaliana. *Plant Cell Physiol*, 44(6):619–29, (2003). 110

Yang J., Kornbluth S.: All aboard the cyclin train: subcellular trafficking of cyclins and their CDK partners. *Trends Cell Biol*, 9(6):207–10, (1999). 6

Yang J., Song H., Walsh S., Bardes E. S., Kornbluth S.: Combinatorial control of cyclin B1 nuclear trafficking through phosphorylation at multiple sites. *J Biol Chem*, 276(5):3604–9, (2001). 24

Yang Q., Rout M. P., Akey C. W.: Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. *Mol Cell*, 1(2):223–34, (1998). 7

Yang Z. R.: Biological applications of Support Vector Machines. *Briefings in Bioinformatics*, 5(4):328–338, (2004). 36

Yasuhara N., Oka M., Yoneda Y.: The role of the nuclear transport system in cell differentiation. *Semin Cell Dev Biol*, 20(5):590–9, (2009). 7, 9

Yi R., Qin Y., Macara I.-G., Cullen B.-R.: Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17(24):3011–6, (2003). 12

Yin Y., Cheong H., Friedrichsen D., Zhao Y., Hu J., Mora-Garcia S., Chory J.: A crucial role for the putative Arabidopsis topoisomerase VI in plant growth and development. *Proc Natl Acad Sci USA*, 99(15):10191–6, (2002). 115

Zasloff M.: tRNA transport from the nucleus in a eukaryotic cell: carrier-mediated translocation process. *Proc Natl Acad Sci USA*, 80(21):6436–40, (1983). 8

Zemp I., Kutay U.: Nuclear export and cytoplasmic maturation of ribosomal subunits. *FEBS Lett*, 581(15):2783–93, (2007). 12

Zeng Y., Cullen B.-R.: Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res*, 32(16):4776–85, (2004). 12

Zhu J., McKeon F.: NF-AT activation requires suppression of Crm1-dependent export by calcineurin. *Nature*, 398(6724):256–60, (1999). 22

Zhu Y., Tepperman J. M., Fairchild C. D., Quail P. H.: Phytochrome B binds with greater apparent affinity than phytochrome A to the basic helix-loop-helix factor PIF3 in a reaction requiring the PAS domain of PIF3. *Proc Natl Acad Sci USA*, 97(24):13419–24, (2000). 110

Acknowledgments

First I wish to thank Dr. Thomas Merkle and Dr. Tim Nattkemper for the opportunity to do my PhD thesis under their supervision. I am very grateful for their academic and personal support.

I would like to acknowledge the International Graduate School in Bioinformatics and Genome Research for funding the PhD project.

I also thank all the people of both groups where I worked during this time, the genome research group of the Faculty of Biology and the data mining and neuroinformatics group of the faculty of Technology. Thanks specially to Sandra Niemeier, Ute Bürstenbinder, Rima Bachmann and Julia Herold.

I wish to thank also the people of the Bioinformatics Resource Facility (BRF), specially to Dr. Heiko Neuweger for his constant help and support.

Furthermore, I would like to express my gratitude to all my friends who have accompanied and helped me in this project academic and personal. Thanks Andrea Rorher, Narda Forero, Oliver General and Narittza Diaz.

Finally I wish to thank Björn for his company, help, support and all the things that the words can not say.

Claudia Rubiano

Bielefeld, April 2010

ERKLÄRUNG

Ich, Claudia Consuelo Rubiano Castellanos, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe.

Bielefeld, den 20. April 2010

Claudia Rubiano