A MULTIPLE SITE PREDICTOR FOR SUBCELLULAR LOCALIZATION OF FUNGAL PROTEINS

MICHEL NATHAN

A THESIS
IN
THE DEPARTMENT
OF
COMPUTER SCIENCE
AND
SOFTWARE ENGINEERING

PRESENTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE
CONCORDIA UNIVERSITY
MONTREAL, QUEBEC, CANADA

SEPTEMBER 2006
© MICHEL NATHAN, 2006

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-20780-2
Our file Notre référence
ISBN: 978-0-494-20780-2

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

A Multiple Site Predictor for Sub-cellular Localization of Fungal Proteins

Michel Nathan

In this work, we build a system that uses a decision tree to predict fungal protein localization based on physiochemical properties of proteins calculable from their primary sequences. The training examples that serve as basis for learning are obtained from experimentally validated localizations. Although there is clear evidence of presence of the same protein in more than one sub-cellular compartment, almost all existing automated systems restrict their predictions to single-site localization. Here, we attempt to address this issue and for proteins that are reported to target more than one sub-cellular location, our system predicts as many localization sites as possible.

When localizing among 17 sub-cellular compartments, in 64% of the cases our system successfully predicts at least one of the experimentally reported localizations. In addition, our results indicate that all the reported localizations are correctly predicted in 49% of the cases. We also report 76 fungal protein features implicated in localization and indicate those with the highest relative discriminatory power. Finally, we report on necessary conditions for localization to specific sub-cellular sites.

Acknowledgement

I would like to thank my supervisor, Professor Gregory Butler, for his valuable guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF MODULES

CHAPTER 1. INTRODUCTION

1.1. Problem Statement

Sub-cellular localization allows biologists to make inference on the functions of a protein and its annotation. It also provides important hints on the pathway the protein is involved in and the class of proteins it may interact with. Biochemical, cytological and genetic methods are used for functional characterization of known proteins. Cell Fractionation, Electron Microscopy and Fluorescence Microscopy have been extensively used for experimental sub-cellular localization. The results of such experiments provide accurate and reliable information on the sub-cellular compartments to which proteins are targeted. These experiments are, however, labour intensive and manual annotation cannot keep up with the increasing number of gene products that become available.

In an effort to aid in accelerating the localization process, numerous automated predictors have been built in recent years. These systems use classification schemes that are based on one or more of the following approaches: (i) Ab-initio method that use compositional, bio-chemical or structural features of proteins to predict their localization; (ii) Methods based on sorting signals that determine a protein's target location; (iii) Signature-based methods that use motifs and profiles to characterize protein families and family domains. This information then will serve in localization, and (iv) Homology-based methods. Representative predictors for these categories are (i) PLOC [Park and Kanehisa, 2003] that uses compositions of amino acids to predict localization by support vector machines, (ii) TargetP [Emanuelsson et al., 2000] that uses N-terminal targeting sequences to predict localization, (iii) PSLT [Scott et al., 2004] that uses InterPro motifs with Bayesian networks, and (iv) Proteome Analyst [Lu et al., 2004] that uses proteins of known localization in Swiss-Prot together with annotation of homologous proteins in

the same database in order to build Bayes classifiers to predict localization for animal, plant, fungal, and bacterial proteins. There are also predictors that are based on a combination of the 4 categories mentioned above. The best known of such hybrid systems is PSORT [Nakai and Kanehisa, 1992] and its extensions (PSORT II, PSORTb, iPSORT and WoLF PSORT) that use a knowledge-based system to predict localization based on a protein's overall amino acid composition, targeting signals and motifs.

It is not known a priori which specific features or combination of features of a protein play determining roles in its localization. Therefore, ideally all biochemical and environmental factors should be considered as potentially representative and should undergo scrutiny. There is also clear evidence of the presence of the same protein in more than one sub-cellular compartment. Although this fact has been widely recognized by biologists, none of the developed automated protein localizers except the one recently developed by Chou and Cai [Chou and Cai, 2005] attempts to predict more than one possible destination for a given protein. This shortcoming needs to be addressed and a predictor should ideally be able to handle multi-site localization.

We propose to build a classifier that learns, from experimentally derived localization information, which bio-chemically meaningful features of proteins are relevant to their localization. The system then uses this knowledge to determine, given an amino acid sequence of a protein, the sub-cellular locations to which the protein is destined. Two important aspects of our system need to be emphasized: (i) the use of a hybrid classification approach based on bio-chemical properties of proteins captured through their amino acid compositions, targeting signals as well as signature-based motifs, and (ii) multiple-site localization, i.e., the capacity to determine all the destinations within the cell to which a given protein may be targeted.

Fungi were chosen as organism group for our investigation because they are endowed with numerous organelles and sub-cellular locations. Moreover, extensive experimental data related to fungi may be found in the literature.

1.2. Overview of the Work

We obtain a significant number of experimental examples localizing various fungal proteins in numerous compartments of the cell. For each element of the set thus obtained, i.e. for each protein of known sub-cellular localization, we compute a set of pre-determined characteristic features purely based on the information contained in its amino acid sequence. Using the reported localizations and the features values calculated from the collected examples, we build a features-location matrix that associates each set of feature values with its corresponding localization. This matrix is input into a learning tool that generates a decision tree that classifies the proteins according to their targeted locations. We evaluate the performance of the system and compare the results to the State of the Art. Furthermore, by analyzing the decision tree obtained from our training set, we shall attempt to determine: (i) which features have no or little impact on the targeting of proteins to sub-cellular locations, (ii) what is the ordering in which various feature values are to be tested in order to obtain the shortest path to a localization, (iii) which sub-cellular locations cannot be localized given our set of examples, and (iv) necessary conditions, in terms of feature values, that a protein should satisfy in order to localize to some specific sub-cellular sites.

1.3. Summary of the State of the Art

State of the Art localization predictor, Proteome Analyst [Lu et al., 2004], achieves a precision of 87% when set to predict localization of fungal proteins among 9 sub-cellular sites. We note,

however, that Proteome Analyst's classifier selects the most probable localization and does not report multiple-site localizations.

Similarly, many other existing predictors do not consider this multiplicity and predict the best possible localization. Among the systems that do output more than one prediction for a given protein is the PSORT [Nakai and Kanehisa, 1992] family of localizers. These localizers provide as output, the probabilities of targeting to individual sub-cellular compartments.

WoLF PSORT, the most recent member of the PSORT family, specializes in animal, plant and fungal proteins. An overall prediction accuracy of 83% has been reported for a 14 compartment localization of yeast by WoLF PSORT [Horton et al., 2006]. These 14 compartments consist of 10 single-site locations and 4 dual localization sites (Cytoplasm and Nucleus, Cytoplasm and Mitochondrion, Cytoplasm and Peroxisome and Mitochondrion and Nucleus). WoLF PSORT analyzes the sequence using pre-determined rules (mostly from PSORT) having as pre-conditions information on the type of targeting signal motifs that direct proteins to various sub-cellular locations. Thus, WoLF PSORT has numerous modules each of which can predict one or more localization sites. Predictions from these modules are weighted and integrated to generate a final prediction. WoLF PSORT also uses some correlative sequence features such as amino acid content from iPSORT [Bannai et al., 2002] and sequence length. Although the latter features contain non-causal ab-initio information about the sorting signals, causal components of localization reflected through rules continue to drive the decision process.

There is a risk involved in attempting to guide the localization process through rules and pre-conditions as many of such rules are incomplete. In fact, as it will be shown later in this work, in many cases, the presence of a targeting motif is not a sufficient cause for localization to a particular site. Moreover, there are ambiguities involved in determining the exact pre-condition to

which some of these targeting motifs should correspond. For example, a trans-membrane helix is indicative of the presence of a targeting motif that favours localization to the plasma membrane or to the membrane of an organelle. On the other hand, the presence of a secretory signal peptide implies direction of the protein to the extra-cellular region. Now, it is hard to distinguish between these two signals. In fact, it has been shown that the presence of one signal in a protein consistently interferes with the detection of the other [Lao et al., 2002].

The only existing single-specie predictor that does consider multi-site localization in a rigorous manner is the one devised by Chou and Cai [Chou and Cai, 2005]. This system uses a data set of 3875 proteins that are experimentally localized into one or more of 22 sub-cellular locations in budding yeast [Huh et al., 2003]. This amounts to a total of 5132 reported localizations or <p, s> pairs where p stands for a protein and s for a single sub-cellular site. To define a protein's feature space, this system uses a dimensional vector of 9772 entries consisting of 3 categories of information: a) "InterPro-mapped-to-GO" dimensional space containing 1930 serialized GO numbers, one for each entry of InterPro functional domain database, b) "functional domain composition" dimensional space containing 7785 entries, and c) "pseudo amino acid composition" dimensional space containing 57 entries, 20 of which are ordinary amino acids and the other 37 are sequence correlation factors, representing the effect of sequence order. The feature vector corresponding to a given data set localized protein is determined using the following procedure:

If InterProScan of protein contains a GO term then the feature set is represented by:
$(a_1, a_2, \ldots, a_{1930}, 0, \ldots, 0)$ where $a_i = 1$ if i^{th} GO term is among the result
Otherwise, if InterProScan of protein result contains a functional domain entry then the feature set is represented by:
$(0, \ldots, 0, b_1, b_2, \ldots, b_{7785}, 0, \ldots, 0)$ where $b_i = 1$ if i^{th} functional domain is in the result
Otherwise, the feature set is represented by:
$(0, \ldots, 0, c_1, c_2, \ldots, c_{57})$ where $c_i = 1$ if the protein contains i^{th} pseudo amino acid.

For a given query protein, p_q, its feature vector, v_q, is determined as per the above procedure (subscript q stands for query). v_q is then compared to the set of feature vectors, V, of 5132 training proteins. The reported localization(s) of those elements of V that show the highest similarity with v_q are designated as the system prediction(s) for the query protein. For example, if v_1=<p, s_1>, v_2=<p, s_2> and v_3=<p, s_3> (v_1, v_2 and v_3 being in the training set vector space V) depict the highest similarity with v_q then p_q is predicted to localize to the set $\{s_1 \cup s_2 \cup s_3\}$ (i.e. the union of s_1, s_2 and s_3).

The performance measure reported by Chou and Cai [Chou and Cai, 2005] consists of the standard per site *sensitivity* measure TP / (TP + FN) where TP stands for true positive and FN stands for false negative. The reported values are 70%, 84% and 90% (respectively) when the highest-ranking, the two highest-ranking, and the three highest-ranking predictions (respectively) in terms of similarity with experimental results are taken into consideration. No mention of false positives (FP) and the impact of considering more than one highest-ranking prediction on the number of FP has been made in their work.

1.4. Summary of Results

When localizing among 17 sub-cellular compartments, for 64% of the proteins some of their reported localizations and for 49% of the proteins all of their reported localizations were successfully predicted. The level of these "partial" and "total" predictions varies depending on the number of sites to which a given protein is reported to localize. As such, for single-site, double-site and triple-site proteins our measured values for "partial" and "total" predictions are (58%, 58%), (82%, 20%) and (84%, 5%) respectively.

Standard per site sensitivity measures obtained by our system are 55% and 60% for localization to 17 and 9 sites respectively. As for standard per site specificity measures TN / (TN + FP) where TN stands for true negative, we obtained a performance of 91% and 92% respectively.

CHAPTER 2. BIOLOGICAL BACKGROUND

2.1. Proteins

In this chapter, we provide some background information on the subject of interest, proteins. A brief description of general function and structure of proteins is followed by a more detailed introduction to the eukaryotic cell. Cell membrane and major intracellular compartments are summarily described. The main part of this chapter (section 2.3) deals with localization of the proteins within the cell. Different pathways the proteins participate in, molecular requirements for their recognition, as well as various interactions they have with other sub-cellular elements are described in the context of protein sub-cellular localization.

2.1.1 Protein Function

Proteins are long un-branched polymers of amino acids (generally between 50 and 2000) joined head to tail by covalent peptide bonds. Condensation reaction (formation and expulsion of a water molecule) allows polymerization of amino acids. Proteins form most of a cell's dry mass and execute nearly all cell functions. Each protein molecule performs a specific function according to its own genetically specified sequence of amino acids. Following are some examples of these functions:

- Catalyze reactions (ex: the enzyme glucoamylase completely hydrolyzes starch to glucose)
- Maintain structure (ex: membrane proteins form channels in plasma membrane that permit passage of molecules in and out of cells)
- Control of cellular activity (ex: regulatory proteins control DNA transcription)
- Generate movement (ex: kinesin propels organelles through cytoplasm)
- Sense signals (ex: receptors detect signal molecules and initiate an appropriate response)

Many protein structures contain inorganic ions, water and small inorganic or organic molecules. These ligands participate directly in the function of the protein. There are typical binding sites on proteins for inorganic ions (ex: copper), small organic molecules (ex: substrates, inhibitors, effectors), other proteins and nucleic acids. Most of a protein's functions are performed in various sub-cellular locations and are determined by the protein's sequence and structure [Lesk, 2004].

2.1.2 Protein Structure

A protein has a polypeptide backbone consisting of a repeating sequence of atoms along its core. Each amino acid in this backbone contains one of 20 different side chains and it is the number and kind of these side chains that give a protein its specific properties. Each protein has a 3D structure that depends on its amino acid sequence. The process of folding into this higher order structure depends on a large set of weak interactions produced by non-covalent forces between atoms. These forces are of 4 types: ionic bonds, hydrogen bonds, van der Waals attractions and an interaction between non-polar groups caused by their hydrophobic expulsion from water [Lesk, 2001].

Protein structures form a hierarchy: (i) ***Primary structure***: amino acid sequence of the main chain, (ii) ***Secondary structure***: helices and sheets formed by the H-bonding pattern of the main chain, (iii) ***Super-secondary structure***: recurrent patterns of interactions between helices and sheets close together in the sequence. Examples are α-helix hairpin, β-hairpin, and β-α-β, (iv) ***Domain***: compact units within the folding pattern of a single chain that seem to have independent stability, (v) ***Tertiary structure***: assembly of helices and sheets, (vi) ***Quaternary structure***: assembly of monomers (for proteins composed of more than one sub-unit)

Under physiological conditions of solvent and temperature, most proteins fold spontaneously to an active native state (3D form) dictated by their amino acid sequence. This fact is concluded from experiments whereby a protein is denatured (unfolded) through treatment with a solvent that disrupts the non-covalent interactions that hold the folded chain together. When the denaturing solvent is removed, the protein often refolds spontaneously (re-natures) into its original conformation indicating that all the information needed for specifying the 3D shape of a protein is contained in its amino acid sequence [Alberts et al., 2002, Chapter 3]. Mutations / additions / deletions in the amino acid sequences perturb the protein conformation but selection can impose constraints on the structure to preserve function. Therefore, the core of the structure is well conserved. Any possible conformation of the polypeptide chain of a protein places different sets of residues in proximity. The interaction of the side chains and main chain, with one another and with solvents and ligands, determine the energy of the conformation. Proteins have evolved so that one folding pattern of the chain produces a set of interactions that is significantly more favourable than all others. Formation of the native state requires all the protein because many of the stabilizing interactions involve parts that are distant in the polypeptide chain but are brought into spatial proximity by the folding [Lesk, 2001].

Two regular folding patterns often recur in proteins: ***α-helix*** and ***β-sheet***. An α-helix is formed when a rigid cylinder is generated by twisting a single polypeptide chain around itself. Each residue at position i of the main chain is H-bonded to the residue at position (i+4). There is also a rotation of 100 degrees around the helix axis from one main chain residue to the next. Cell membrane proteins such as transport proteins and receptors particularly abound in short regions of α-helix. A β-sheet is formed when separate strands that may arise from regions distant in the sequence are H-bonded together. A pair of adjacent strands in a β-sheet may interact in parallel or anti-parallel so β-sheets may be in parallel, anti-parallel or mixed forms. The core of many proteins contains extensive regions of β-sheet. Proteins adopt α-helix or β-sheet conformation

because individual residues are in a low energy form when in α or β conformation. To form compact structures, globular proteins have regions of helix and strand linked by turns or loops. The former transverse the structure whereas the latter are on the surface.

In addition to simple α-helix and β-sheet patterns, two adjacent β-sheets may be connected by a bridge segment (often α-helix) leading to a ***β-α-β conformation***, and two adjacent anti-parallel β-sheets may be connected by a turn (this is called a ***β–hairpin***). Hydrogen bonding is different between parallel sheets and anti-parallel sheets in that in the former bonding alternates with right and left neighbours but in the latter bonding is with neighbours on both sides. In some other proteins, two or three α-helices wrap around each other forming what has become known as ***coiled-coil*** structure that is particularly stable. Table-1 depicts possible conformations of proteins according to their secondary and tertiary structures [Lesk, 2001]:

Class	Characteristics	Examples
α–helical	almost exclusively α–helical	Citrate synthase
β–sheet	almost exclusively β–sheet	Chymotrypsin
α+β	α-helical and β-sheets separated in different parts of molecule	Papain
α/β	helices and sheets assembled from β-α-β units	flavodoxin
coiled- coil	2 or 3 α–helices wrapped around each other	α -keratin

Table-1: Protein conformations according to secondary and tertiary structures

Another important structural unit is the ***protein domain***, which is a sub-structure produced by any part of a polypeptide chain that can fold independently into a compact, stable structure. A domain usually contains between 40 and 350 amino acids and is often associated with a distinct function. A protein may contain one or more (up to several dozens) domains all connected to each other. Gene duplication has occurred often in the course of evolution, allowing each gene

copy to evolve independently in order to perform new functions. This has led to formation of ***protein families*** where within each family the members have amino acid sequences and 3D conformations that resemble each other. Larger protein molecules may contain more than one polypeptide. This occurs when a binding site of a protein molecule recognizes the binding site of another protein and binds tightly to it. Several protein sub-units may bond to each others in this way forming a ***protein complex***.

2.2. Cell

2.2.1. Cell Theory

Cells are morphological and physiological units of all living organisms and the properties of a given organism depend on those of its individual cells. Moreover, cells originate only from other cells and continuity is maintained through the genetic material that is precisely partitioned during the process of cell division. Each cell is bounded by an amphipathic, bilayer plasma membrane that acts as a selective barrier to enable the cell to retain the products it synthesizes for its own use and excrete waste products. The transport proteins in the membrane determine which molecules enter the cell and catalytic proteins inside the cell determine the reactions that these molecules undergo. Therefore, by specifying the set of proteins that a cell is to manufacture, the genetic information recorded in its DNA sequence dictates the entire chemistry of the cell [Alberts et al., 2002, Chapter 1].

2.2.2. Cell Compartments

A eukaryotic cell is divided into membrane-bound compartments called organelles, each of which has its specialized enzymes and is hence functionally distinct. Enzymes catalyze the reactions in these organelles and also serve markers and transporters that allow exchange of molecules

between the compartments. The main intracellular organelles in eukaryotes are depicted in Figure-1.

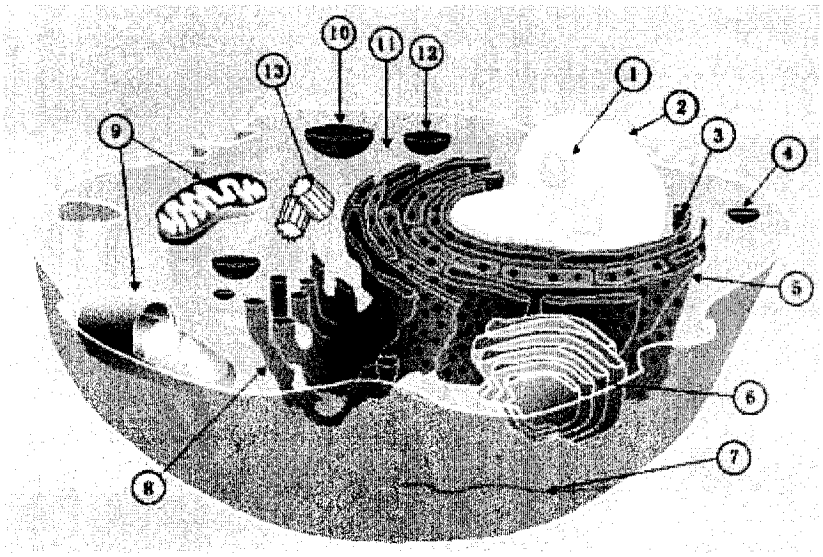

Figure-1: Schematic of a eukaryotic cell
Source: http://www.netlexikon.akademie.de/images/en/thumb/a/a7/400px-Biological_cell.png

The nucleus (2) contains the main genome and is the site of synthesis of DNA and RNA. Assembly and synthesis of RNA & proteins that will make up ribosomes are accomplished in the nucleolus (1) that is contained within the nucleus. The nucleus is surrounded by cytoplasm (11). The latter accommodates the cytosol and various cytoplasmic organelles. The cytosol is the site of most protein synthesis. The endoplasmic reticulum (ER) is a large membrane-bound organelle consisting of rough ER (5) with many ribosomes (3) attached to its cytosolic side and smooth ER (8) with no attached ribosomes. The ER is involved in the synthesis of soluble and membrane proteins. Many of these proteins are sent to the Golgi apparatus (6). The Golgi consists of disk-like cisternae that receive proteins and lipids from the ER, modifies them and sends them in transport vesicles (4) to various part of the cell. Mitochondria (9) are membrane-bound organelles that generate most of the energy required to run many of the chemical reactions in the cell. Lysosomes (12) are another set of organelles that degrade dead cell parts and macromolecules ingested from outside of the cell using digestive enzymes contained within their

lumen. These macromolecules pass through endosomes on their way to lysosomes. Peroxisomes contain oxidative enzymes. Vacuoles (10) perform a variety of secretory, excretory, and storage functions. Most of the intracellular organelles are created by pinching off membrane from the plasma membrane or the ER. The interior of these organelles is thus topologically equivalent to the exterior of the cell. Thanks to microtubules and other cytoskeleton elements (7), membrane-bound organelles have characteristic positions in the cytosol. For example, Golgi is near the center of the cell, close to centrioles (13) whereas ER is spread almost everywhere in the cytosol.

2.2.3. Cell Membrane

The plasma membrane encloses the cell and serves as the boundary between the cytosol (within the cell) and the cell's extra-cellular environment. Similar segregation is achieved within the cell by the membranes of the ER, Golgi, mitochondria and other organelles that allow the contents of these organelles to remain separated from the cytosol. The plasma membrane is a thin film consisting of a lipid bilayer that allows it to act as a relatively impermeable barrier to the passage of most water-soluble molecules. Proteins embedded in this lipid bilayer are responsible for most of the functions of the plasma membrane (such as transport of specific molecules across the membrane itself). There are often some oligosaccharides that are attached to proteins in the membrane facing toward the cell exterior. The exterior surface of the cell therefore is covered with a cell coat made of carbohydrates that protects it from mechanical and chemical damage. The cell coat also helps mediate cell-cell adhesion [Alberts et al., 2002, Chapter 10].

Approximately 20% of proteins in yeast are identified as trans-membrane proteins [Alberts et al., 2002, Chapter 10]. Trans-membrane proteins are amphipathic structures, having hydrophobic regions that interact with the hydrophobic tails of lipid bilayer, away from water, and hydrophilic regions that remain exposed to water on either side of the membrane.

There are various ways in which membrane proteins are associated with the membrane lipid bilayer. They may pass through the lipid bilayer (i) as a single α-helix, (ii) as multiple α-helices, (iii) as a rolled-up β sheet (also called a β-barrel), (iv) anchored to the cytosolic face of the membrane by an amphipathic part of their structure, or be (v) attached covalently to the cytosolic monolayer by a lipid chain, (vi) attached covalently to the outer monolayer of the membrane via a ***glycosylphosphatidylinositol*** (GPI) anchor, or (vii) simply bound non-covalently on the cytosolic or outer face of the membrane to another membrane protein. The proteins that adopt the latter mechanism (bind to other existing membrane proteins) are called ***peripheral membrane proteins***. Whereas, the proteins that traverse the membrane or are tightly attached or anchored to it are called ***integral membrane proteins*** [Alberts et al., 2002, Chapter 10].

Within the trans-membrane proteins, the hydrophobic segments that are in contact with the membrane lipid bilayer have non-polar side chains. Since peptide bonds themselves are polar and water is absent within the membrane, all peptide bonds form H-bonds with one another. As H-bonding is maximized in a regular α–helix, most membrane spanning segments of membrane proteins are in α–helix form. Segments of 20 to 30 hydrophobic amino acids are long enough to span the membrane lipid bilayer as a α–helix. As for β-barrels, 10 amino acids or fewer are enough to traverse the membrane lipid bilayer as an extended β strand [Alberts et al., 2002, Chapter 10].

2.2.4. Membrane Transport Proteins

Small non-polar molecules can cross the cell membrane by simple diffusion. Lipid bilayers are, however, highly impermeable to most polar molecules (ions, sugars, amino acids, etc.) and ***membrane transport proteins*** are needed to let these molecules cross the membrane. These proteins are also needed for the transport of polar molecules into and out of intra-cellular membrane-enclosed organelles. Each membrane transport protein usually transports a certain

type of a particular class of molecules. These proteins form continuous protein pathways that enable the passage of hydrophilic solutes without any contact with membrane lipid [Alberts et al., 2002, Chapter 11].

There are two main types of membrane transport proteins. ***Channel proteins*** form aqueous pores through the membrane lipid bilayer allowing passive diffusion of solutes down their concentration gradient. If these solutes are also charged then it is the sum of the electrical and concentration gradient (i.e. electro-chemical gradient) that determines the direction of the diffusion. ***Carrier proteins*** bind a specific solute (to be transported) on one side of the membrane. This binding makes the carrier proteins undergo a conformational transformation that enables them to transfer the solute across the membrane and make it accessible to the other side of the membrane. Some carrier proteins can only transport passively the solutes down their electro-chemical gradient, whereas, some others can perform active transport of solutes (against their gradient) using ATP hydrolysis or other sources of energy [Alberts et al., 2002, Chapter 11].

2.3. Transport Pathways and Protein Localization

2.3.1. Inter-Compartmental Movement of Proteins

Apart from the few proteins that are synthesized on ribosomes of mitochondria, the great majority of proteins are synthesized on cytosolic ribosomes. Once generated, the proteins that do not carry a ***sorting signal*** remain in the cytosol while others are directed to the nucleus, ER, mitochondria, or other locations within the cell or are excreted outside the cell. There are 3 main transport methods that may be deployed by proteins to move between various compartments of the cell: (i) ***Gated transport***, the protein moves between two topologically equivalent organelles. For example, the nuclear pore complexes actively transport specifically selected proteins between the nucleus and the cytosol. (ii) ***Transmembrane transport***, specific proteins are directly

translocated across an organelle membrane between the cytosol and the organelle lumen. The transported proteins should be unfolded by transmembrane proteins to cross the membrane. Transport into and out of the ER, mitochondria and peroxisomes is achieved by this mode. (iii) ***Vesicular transport*** is carried out by membrane-enclosed transport vesicles that pinch off part of the membrane of some compartment with some materials from the lumen entrapped inside and carry these to some other compartment. The vesicle then fuses with the new compartment and thus passes its cargo to it. This mode of transport takes place between the ER and the Golgi as well as between the Golgi, lysosomes and the cell surface [Alberts et al., 2002, Chapter 12].

There are two types of sorting signals that are found on the transported proteins that guide protein transport. First, there are ***signal sequences*** that consist of a continuous stretch of 15-60 amino acids. These sequences are usually cleaved once the sorting is complete. The second type of sorting signal is a ***signal patch*** that consists of a specific 3D structure on the surface of the protein. Although the amino acids that lead to this specific conformation may be distinct from one protein to another, the 3D specificity persists. Some of the signals that direct proteins from cytosol to nucleus and direct degradative enzymes into lysosomes are signal patches. Sorting signals are recognized by complementary sorting receptors that guide them to appropriate destinations [Alberts et al., 2002, Chapter 12].

In general, during cell division, a complete set of specialized cell membranes is passed from mother cell to daughter cell. This allows regeneration of most of the organelles in the daughter cell. Some organelles (like ER and mitochondria) cannot be constructed from scratch. For example, many proteins that function in the ER are the product of the ER itself. Such organelles are therefore passed intact to the daughter cells during cell division [Alberts et al., 2002, Chapter 12].

2.3.2. Molecular Mechanisms of Protein Transport

The ***Endocytic pathway*** is a pathway by which macro-molecules are ingested from the cell environment and delivered to the lysosome to be metabolized using digestive enzymes. This pathway runs from the plasma membrane toward the early endosome toward the late endosome and lysosome. There are also some endocytosed molecules that are retrieved from early endosomes and return to the cell surface. The ***Bio-synthetic-secretory pathway***, on the other hand, runs from the ER to the Golgi and from the Golgi to the cell surface and the lysosome. This pathway is thus responsible for delivery of newly synthesized proteins, carbohydrates and lipids to the extra-cellular space. This process is usually performed by modification and storage of the molecules produced until they are needed, followed by exocytosis to exit the cell. In this pathway some molecules are retrieved from late endosomes and returned to the Golgi and some others are retrieved from the Golgi and returned to the ER. Figure-2 schematically summarizes these transport pathways for proteins [Alberts et al., 2002, Chapter 13].

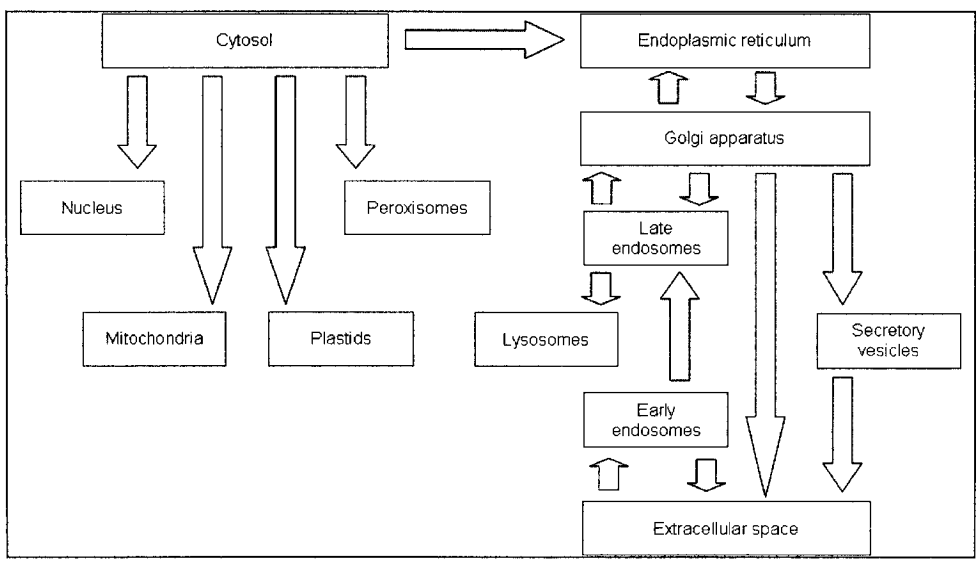

Figure-2: Intra-cellular protein transport pathways
Source: [Alberts et al., 2002, Chapter 12]

The Lumens of the membrane-enclosed compartments along the mentioned pathways are topologically equivalent and communicate through ***transport vesicles*** that continually bud off from one membrane and selectively fuse with another membrane that displays the appropriate receptor (that can recognize the vesicle) on the cytosolic side of its membrane. In this manner these vesicles transport membrane and soluble molecules between different organelles. When a transport vesicle buds off, it is covered on its cytosolic side with a coat. An appropriate ***chaperone*** removes this coat once the vesicle is released from the originating membrane. There are 3 types of coated vesicles: (i) Clathrin-coated vesicles that mediate transport from the Golgi to the late endosome and from the plasma membrane to early endosomes and from the plasma membrane to the Golgi, (ii) COPII-coated vesicles that transport from the ER to the Golgi, and (iii) COPI-coated vesicles that transport from the Golgi to the ER, from the Golgi to the plasma membrane, between different cisternae of the Golgi, and from early endosomes to late endosomes [Alberts et al., 2002, Chapter 13] (Figure-3).

With all of these transport vesicles moving between various organelles, the specificity of transport becomes crucial. Each transport vesicle contains a surface marker (vesicle SNARE or v-SNARE) that indicates the type and origin of the cargo it carries. Another protein of the same family (target-SNARE or t-SNARE) is found on the surface of a target membrane. There are many different SNAREs associated with each membrane-enclosed organelle. When a v-SNARE is recognized by a complementing t-SNARE, the helical domains of one wrap around those of the other, forming a stable trans-SNARE complex that allows the fusion of the two membranes. SNARE proteins are therefore responsible for specificity in recognition and catalytic fusion in vesicular transport [Alberts et al., 2002, Chapter 13].

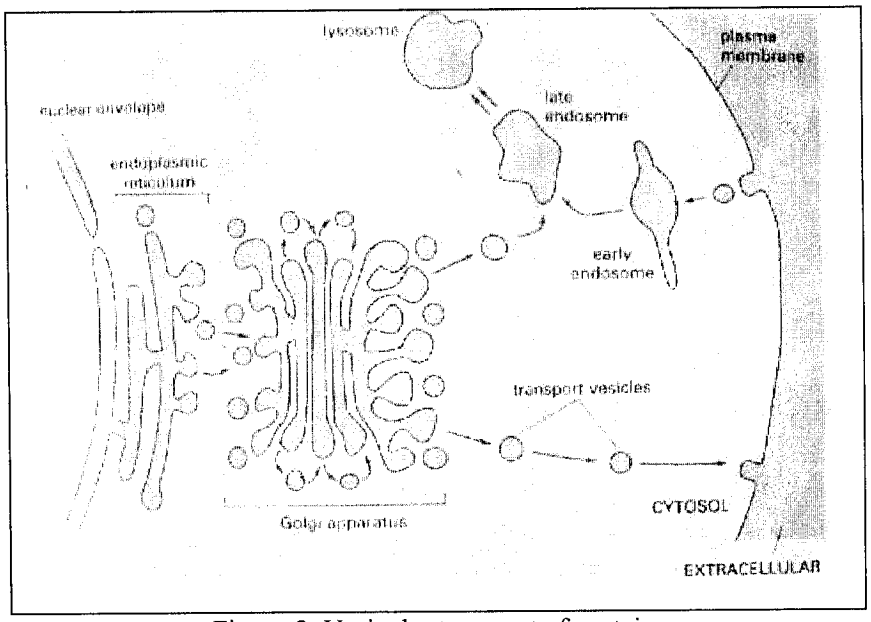

Figure-3: Vesicular transport of proteins
Source: http://137.222.110.150/calnet/cellbio/image/vesicular%20traffic-export%20&%20inport%20pathways.jpg

2.3.3. Nucleus – Cytosol Transport

The nucleus is separated from the cytoplasm by a double membrane nuclear envelope. The outer nuclear membrane is studded with ribosomes engaged in protein synthesis. The newly synthesized protein is transported to the ***perinuclear space*** (between inner and outer nuclear membrane) that is continuous with the ER lumen (Figure-4).

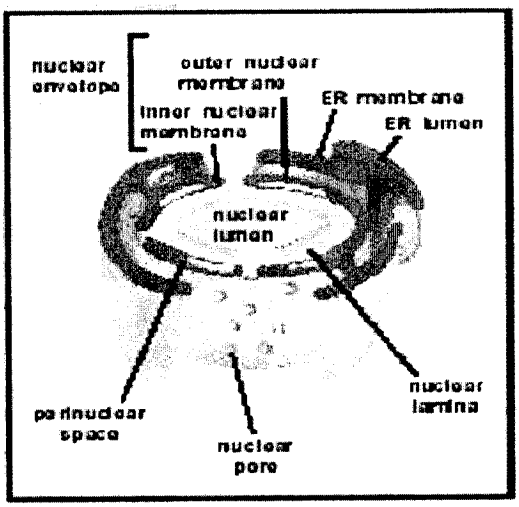

Figure-4: Nuclear compartment and its sub-structures
Source: http://www.biologyreference.com/images/biol_03_img0319.jpg

Selective import into the nucleus (of gene regulatory proteins for example) and selective export out of the nucleus (of ribosomal subunits for example) take place through ***nuclear pore complexes***. These are large structures made up of more than 50 different proteins (called ***nucleoporins***). Unlike translocation through membranes of other organelles, the traffic through the nuclear envelope occurs while proteins are in their fully folded form. This is possible because of the presence of the large aqueous pore in the nuclear pore complex. Selectivity of the proteins that are imported into the nucleus is assured by ***nuclear localization signals*** (NLS). These signals can be in the form of signal sequences or signal patches and their precise location within the chain of amino acids is not important. Cytosolic soluble import receptors bind to an NLS and to nucleoporins (latter binding is made with or without the aid of adapter proteins) and allow import of protein into the nucleus. Once in the nucleus, the import receptor is dissociated from the cargo and is returned to the cytosol to take part in a new round of import. In a similar process, recognition of ***nuclear export signals*** (NES) by nuclear export receptors initiates the export from the nucleus. Some proteins shuttle between the cytosol and the nucleus. An example of such proteins is the protein receptor of newly synthesized mRNA. It has both an NLS and an NES and its steady-state equilibrium can only be achieved through balancing the relative rate of import and export. This is regulated by turning on and off the mentioned signals, i.e. denying access to receptors or through the usage of some other regulatory proteins [Alberts et al., 2002, Chapter 12].

At the beginning of cell division, nuclear pore complexes are disassembled and dispersed through the cytosol. Later on during this process, the nuclear envelope reassembles and wraps around chromosomes until a new sealed envelope is reformed. At this stage, nuclear proteins need to be re-imported from the cytosol. For this reason, unlike in many other organelles, when a protein

has been imported into the nucleus, its NLS is not cleaved, as it will be needed during each subsequent cell division [Alberts et al., 2002, Chapter 12].

2.3.4. Cytosol – Mitochondria Transport

Mitochondria are double-membrane-bounded organelles that are the major site of energy synthesis (in form of ATP) in eukaryotic cells. Although they are endowed with their own DNA and ribosomes, they only synthesize some of the proteins they use and import most of the proteins they need from the cytosol. The intermembrane space and the matrix (Figure-5) constitute two distinct sub-compartments in mitochondria each with their own characteristic proteins.

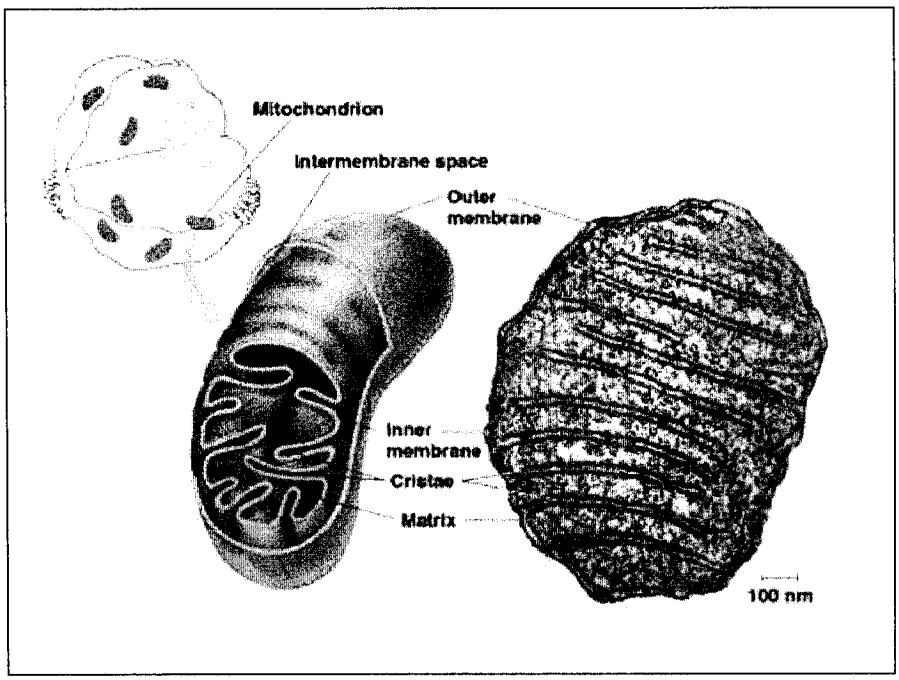

Figure-5: Mitochondrion and its sub-compartments
Source: http://www.uni-kl.de/FB-Biologie/AG-Neuhaus/Forschung/mitochondria1.jpg

Mitochondria are produced by growth (through uptake of proteins from the cytosol) followed by fission to generate daughter organelles. Proteins are imported into mitochondria post-translationally. This means that the protein precursors are fully synthesized in the cytosol prior to their import into mitochondria. Sorting signals for this import are generally N-terminal on the protein to be transported. This sequence folds into an amphipathic α–helix that has its positively

charged residues clustered on one side of the helix and uncharged hydrophobic residues on the other side. It is this particular configuration and not the precise location of the amino acids in the sequence that allows recognition of the marker by the signal receptor that initiates the import process [Alberts et al., 2002, Chapter 12].

Multi-protein complexes are responsible for translocation of nucleus-encoded proteins into mitochondria. These proteins that are kept in unfolded conformation by cytosolic chaperones are first translocated into the mitochondrial matrix across the mitochondrial inner and outer membrane contact sites. This transport is achieved by the translocation machinery consisting of translocase of the outer and the inner mitochondrial membranes (TOM and TIM complexes respectively). Once in the mitochondrial matrix, further transport across or into the inner membrane becomes possible if the protein has a second hydrophobic internal signal sequence that is unmasked once the (matrix targeting) N-terminal signal is cleaved. Therefore, transport into the mitochondrial inner membrane or the mitochondrial inter-membrane space requires two signal sequences. Once in the mitochondrial inner membrane, the protein may be further pulled by TOM to release its second hydrophobic sequence in the inner membrane and end up in the inter-membrane space [Alberts et al., 2002, Chapter 12].

2.3.5. Transport Into and Out of Peroxisomes

The peroxisome is a single-membrane-enclosed organelle that specializes in oxidative reactions. It usually contains enzymes that oxidize organic substances and produce H_2O_2. Catalase (an enzyme also found in peroxisome) then uses H_2O_2 to oxidize various undesired substances (ex: ethanol, formic acid, etc.). A specific C-terminal tri-peptide signal or some other N-terminal signal sequence on a protein is recognized by both a soluble receptor protein in the cytosol and by docking proteins on the cytosolic surface of the peroxisomes. A complex of 23 proteins, called ***peroxins***, is involved in the import process whereby proteins are translocated across the

membrane into the peroxisome without being unfolded. New peroxisomes are generated from growth and fission of pre-existing organelles [Alberts et al., 2002, Chapter 12].

2.3.6. Endoplasmic Reticulum

The ER consists of a large set of branching tubules and flattened sacs spread across the cytosol all inter-connected to form a continuous organelle. Trans-membrane lipids and proteins for most of the cell's organelles are produced in ER membrane. Protein complexes composed of more than one subunit are usually assembled in the ER. This assembly is a pre-requisite for transport out of the ER [Rose and Doms, 1988]. Most of the secretable proteins as well as those destined to the lumen of the ER and other organelles are initially sent to the ER lumen. Transmembrane proteins are partially translocated across the ER membrane and become embedded in it (either permanently as ER membrane protein residents or transitionally before being sent to other organelles). Soluble proteins, on the other hand, are fully translocated across the membrane and become released into the ER lumen.

Import into the ER may be co-translational or post-translational. In the co-translational mode of import, the process starts before the importable protein is completely synthesized. Membrane-bound ribosomes that are attached to the cytosolic side of (rough) ER are responsible for this type of translation. As the protein is never released before the translation is complete, no chaperone is required to keep the protein in un-folded form prior to import. ***Signal recognition particle*** (SRP) in cytosol and ***SRP receptor*** found in the ER membrane are both necessary for recognition of the ER signal sequence. As protein is being synthesized in a ribosome, SRP binds both to the ER signal sequence (as soon as it emerges from the ribosome) and to the ribosome. The latter causes a pause in the translation process that, in turn, allows the ribosome to bind to the ER membrane before the synthesis is completed. SRP receptor then binds to the SRP-ribosome complex and directs the translocation. SRP and SRP receptor are released only after the ribosome is fully

engaged with the translocator in the ER membrane and thus the nascent protein is never exposed to the cytosol and traverses the membrane by passing through the aqueous pore of the translocator (Figure-6). An N-terminal ER signal sequence serves two purposes. First, it directs the protein to the ER membrane where it is recognized by SRP. Secondly, it serves as a start transfer signal that triggers the opening of the pore. It is also possible for proteins that are already fully synthesized in the cytosol to be imported into the ER post-translationally. In post-translational import, uni-directional feeding of protein across the pore is achieved with the help of chaperones (called binding proteins) which are cyclically generated and released onto the protein chain as it appears across the pore thus pulling the protein into the ER lumen [Alberts et al., 2002, Chapter 12].

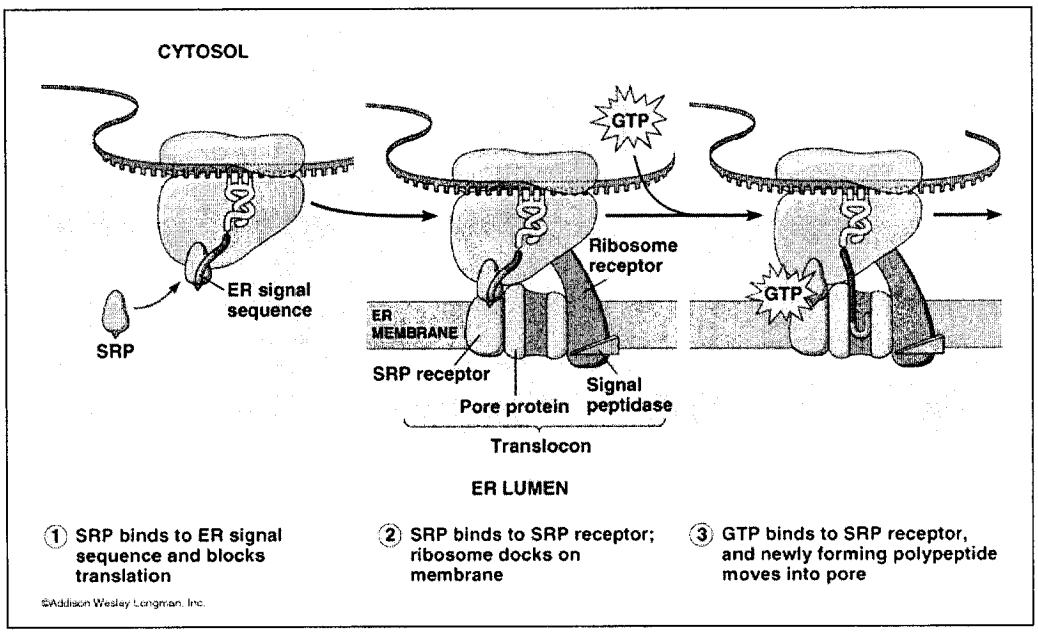

Figure-6: Signal mechanism of cotranslational import
Source: http://www.mun.ca/biology/desmid/brian/BIOL2060/CellBiol20/CB20_16.html

The above description applies to import of soluble proteins across the ER membrane. As for trans-membrane proteins that are imported to the ER membrane, they all contain one or more spanning hydrophobic α–helical regions. ***Single-pass transmembrane proteins*** are inserted into the ER membrane in one of three ways depending on the location of the signal sequence. An N-terminal signal sequence initiates the translocation (as in the case of soluble protein import) but

an additional hydrophobic segment in the polypeptide chain stops the transfer process before the entire protein is translocated. When the N-terminal signal sequence is released from the translocator and is cleaved off, the stop-transfer signal anchors the trans-membrane protein in the ER membrane. The translocator then opens laterally and the stop-transfer signal is released laterally into the ER membrane where it stands as a single-pass α-helical membrane protein in the ER membrane with its N-terminus on the lumenal side and its C-terminus on the cytosolic side. This is the first type of single-pass membrane protein insertion into ER membrane. There are also cases where an internal signal sequence is recognized by SRP and acts as a start-transfer signal that initiates the translocation process. Once released from the translocator, this internal start-transfer signal remains within the lipid bilayer as a single-pass α-helical protein membrane in ER membrane. Now, depending on the orientation of the inserted signal two cases are possible. When N-terminus is on the lumenal side, the protein segment preceding the start-transfer signal is moved into the lumen. When the C-terminus is on the lumenal side, the protein segment following the start-transfer signal is moved into the lumen. These two constitute the second and the third type of single-pass membrane proteins insertion into ER membrane. ***Multi-pass transmembrane proteins*** have many hydrophobic α–helices that span the lipid bilayer many times back and forth. This is achieved by numerous hydrophobic α-helical regions located along their polypeptide chains alternately serving as start-transfer and stop-transfer signals [Alberts et al., 2002, Chapter 12].

Smooth ER is that region of the ER that lacks attached ribosomes and is usually the ER exit site where the transport vesicles bud off towards the Golgi. In the ER lumen, proteins fold and oligomerize, disulfide bonds are formed and N-linked oligosaccharides are added to the proteins. N-linked glycosylation is used to indicate the extent of protein folding. In fact, many proteins that are translocated into the lumen of ER fail to fold properly. These are recognized through their N-linked glycosylation and are sent back to cytosol to be degraded. Certain ER lumen

enzymes covalently add a GPI anchor to the C-terminus of selected membrane proteins. These proteins will be destined for the plasma membrane where they will be attached to the exterior of the plasma membrane by their GPI anchors [Alberts et al., 2002, Chapter 12].

2.3.7. ER - Golgi Transport

Various membrane and soluble lumenal proteins are packaged into COPII-coated vesicles and leave the ER at special ER exit sites (in the Smooth ER region) for the Golgi. Many of these exiting proteins display transport signals on their surface and are actively recognized by complementary receptor proteins that become trapped in the budding vesicles. Only proteins that are properly folded may leave the ER in this manner. Those that are not properly folded will be transported to the cytosol for degradation. Many of the transport vesicles that are released from the ER fuse together to form ***vesicular tubular clusters***. As these short-lived structures move along microtubules from the ER to the Golgi, COPI-coated vesicles bud from them and take the escaped proteins back to the ER in a ***retrograde transport***. The contents of these clusters therefore matures continually. Once the clusters reach the Golgi, they fuse and release their contents into the Golgi [Alberts et al., 2002, Chapter 13].

Resident ER proteins that are retrieved back to the ER display characteristic sorting signal sequences at their C-termini. These signals consist of KKXX (i.e. double lysine followed by any two other amino acids) for membrane proteins [Schutze et al., 1994], and KDEL (or similar sequences) for soluble proteins [Nilsson et al., 1989]. KKXX-carrying proteins are directly recognized by COPI-coat but KDEL-carrying proteins require additionally a multi-pass trans-membrane protein (KDEL receptor) to help them package into COPI-coated retrograde transport vesicles. KDEL receptors cycle continually between the ER and the Golgi and have a higher affinity for KDEL sequences near the Golgi than they do near the ER (due to the slightly acidic

pH of the Golgi as opposed to the neutral pH of the ER) thus ensuring that most of the proteins displaying such sorting signals are retrieved back to the ER [Alberts et al., 2002, Chapter 13].

In addition to retrieval signal sequences, there is another mechanism that ensures the retention of ER resident proteins in the ER. ER resident proteins bind to one another (in a process called ***kin-recognition***) forming a complex that is too large to enter transport vesicles. This mechanism also applies to many Golgi enzymes that need to function together and are therefore retained in the Golgi [Alberts et al., 2002, Chapter 13].

The Golgi apparatus is a set of disc-shaped (flattened), membrane-enclosed cisternae near the center of the cell that is held in position by microtubules. Each Golgi stack has three functionally distinct compartments (named cis, medial and trans cisternae). Cis-Golgi-network (CGN) is a set of inter-connected tubular and cisternal structure associated with the cis-cisterane. Similarly, trans-Golgi-network (TGN) is associated with the trans-cisternae (Figure-7). The Golgi is the major site of carbohydrate synthesis in the cell. Most of the glycosylation of proteins (that produce proteoglycans) occurs in the Golgi. The Golgi also modifies, sorts and distributes ER products to various destinations. In the Golgi, soluble and membrane vacuolar proteins are segregated from proteins destined for the plasma membrane.

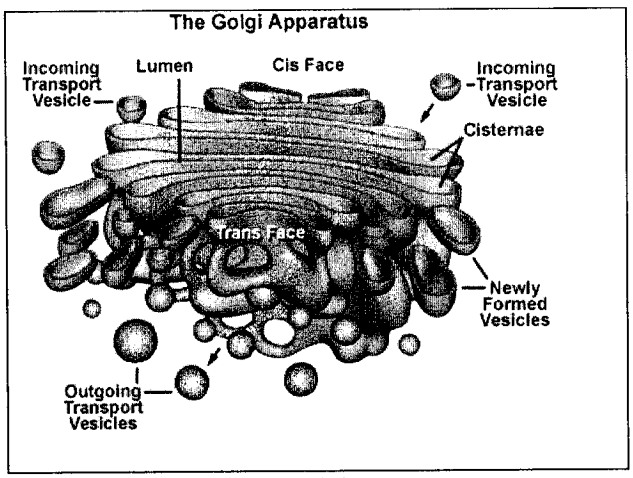

Figure-7: Golgi apparatus
Source: http://micro.magnet.fsu.edu/cells/golgi/images/golgifigure1.jpg

In forward transport, proteins enter the Golgi at the cis side and exit from it at the trans side. These proteins are modified as they pass from one cisterna to the next across a stack. Forward transport across the Golgi is accomplished by one or both of the following mechanisms: (i) ***vesicular transport method***, whereby forward-moving transport vesicles bud from one cisterna and fuse to the next, thus passing the molecules across the CGN, cis cistera, medial cisterna, trans cisterna and TGN that are all considered to be static, and (ii) ***cisternal maturation method***, whereby each Golgi cisterna matures as it moves outwards through the stack. Golgi proteins that are carried forward in a cisterna are moved backward in COPI-coated vesicles to earlier compartments thus causing the cisterna to mature. Finally, when a cisterna becomes TGN, COPI-coated vesicles bud off from it until TGN disappears and becomes replaced by a maturing cisterna behind it [Alberts et al., 2002, Chapter 13].

2.3.8. Golgi – Lysosome Transport

Lysosomes are single-membrane-enclosed organelles containing hydrolytic enzyme and responsible for digestion of intra-cellular debris, phagocytosed micro-organisms, etc. All enzymes in lysosomes are acid-hydrolases, requiring an acidic environment to function. This is provided in the lysosomal lumen that is kept at a pH of 5.0 using a H+ pump in the lysosomal membrane. Products of lysosome degradation are transported to the cytosol where they may be degraded or re-used. Fungi and plants have one or more large, fluid-filled vesicles called ***vacuoles*** that are functionally related to lysosomes in animal cells. A vacuole may occupy some 25% of cellular volume in yeast [Wiemken and Durr, 1974] and is the default compartment for membrane proteins [Roberts et al., 1992]. Membrane proteins are delivered to the vacuole through a) the normal secretory pathway via Golgi, b) the extra-secretory pathway (cytoplasm to vacuole pathway, c) through endocytosis via the plasma membrane, and d) autophagy (non-selective macrophagy). In the late Golgi compartment, soluble vacuolar proteins are actively sorted away from secretory proteins [Stevens et al., 1982].

Lsosomal hydrolases are synthesized in the ER and are transported to the TGN. A ***mannose 6-phosphate*** (M6P) group is attached to the N-linked oligosaccharides of the precursors of these enzymes when they pass through the CGN lumen. This addition segregates these enzymes from all other proteins that reach the TGN since M6P receptors that are trans-membrane proteins in TGN recognize the M6P group and bind to the enzymes on the lumenal side of the membrane and assemble a clathrin coat on the cytosolic side, thus packaging the enzymes into vesicles that bud from the TGN. These vesicles transport the enzymes to the late endosome where at a low pH (6.0 compared with 6.5 at TGN) the enzymes disassociate from M6P receptors. As the pH drops further by maturation of the late endosome, hydrolases become active and start digesting endocytosed materials already delivered to the late endosome. M6P receptors are returned to TGN for re-use [Alberts et al., 2002, Chapter 13] (Figure-8).

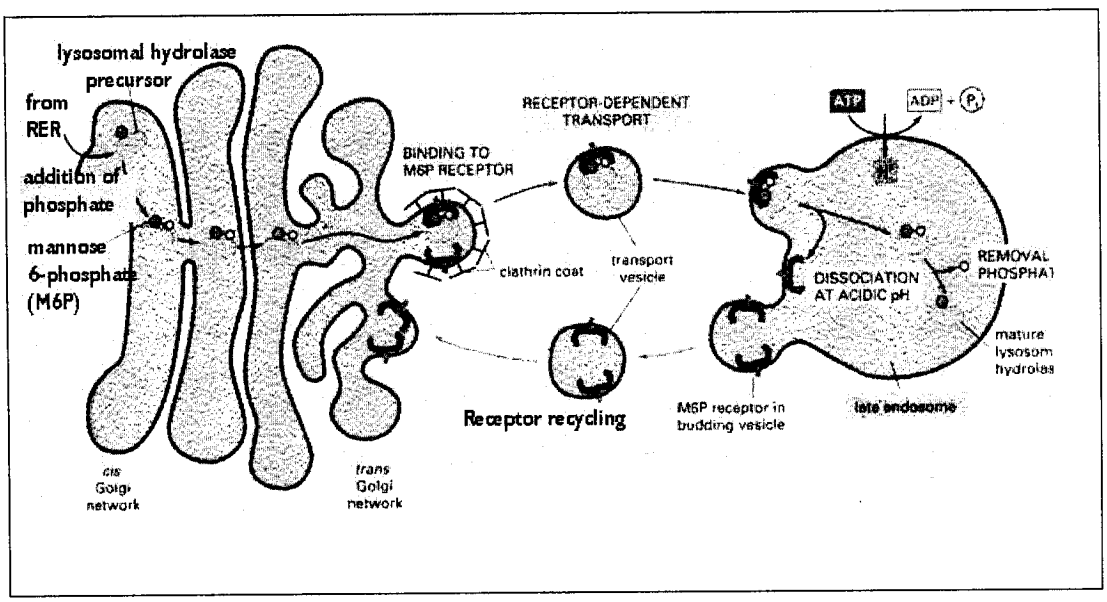

Figure-8: Golgi sorting of lysosomal enzymes
Source: http://cellbio.utmb.edu/cellbio/lys4.jpg

Materials to be digested are delivered to lysosomes following three pathways [Alberts et al., 2002, Chapter 13]:

- Endocytosis: Endocytosed macromolecules from the extra-cellular region are delivered to early endosomes. They are then transported to late endosome. The lumen of late endosomes has a pH of 6 and contains hydrolases so the process of digestion starts. As late endosomes mature to become lysosomes with a further fall in pH, the endocytosed materials become fully digested.
- Autophagy: Organelles or parts of organelles that are no longer needed are enclosed by some membrane forming what is called an ***autophagosome***. The autophagosome fuses with the late endosome or lysosome for digestion of its contents.
- Phagocytosis: Some cells are specialized to phagocytose large microorganisms (ex: bacteria). These cells are called phagocytes. They engulf the object forming a phagosome. Phagosome is then converted to a lysosome.

2.3.9. Endocytosis

A small portion of the plasma membrane invaginates and engulfs the material to be ingested. It then pinches off forming an endocytic vesicle. Some integral membrane proteins that have reached the plasma membrane can also be taken up by endocytosis. There are two main types of endocytosis [Alberts et al., 2002, Chapter 13]:

- ***Phagocytosis***: ingestion of large particles (ex: dead cells) by forming a large vesicle (phagosome) that will eventually fuse with lysosome for degradation. This is an active process requiring transmission of signals by activated receptors (ex: antibodies) to initiate the mechanism. Phagocytosis is mostly performed by specialized cells.
- ***Pinocytosis***: ingestion of fluid and solutes by forming a small pinocytic clathrin-coated vesicle. As soon as the vesicle is released, its coat is shed and it fuses with early endosomes. ***Multivesicular bodies*** are then formed from the endosomes as they move inward along the microtubules. Multivesicular bodies fuse with each other or with pre-existing late endosome and mature into late endosome. The TGN sends vesicles to

deliver appropriate proteins to convert late endosome to lysosome. Pinocytosis is a constitutive process that occurs often in all eukaryotic cells.

2.3.10. Exocytosis

Some transport vesicles leave the TGN and fuse with the plasma membrane in a process called exocytosis. There are two main types of secretory pathways:

- ***Constitutive secretory pathway*** applies to all cells by default. Any soluble protein in the lumen of the Golgi is automatically carried by this default pathway to the cell surface unless retrieved to the ER, retained in the Golgi, or selected for lysosome or regulated secretion.
- **Regulated secretory pathway** occurs mostly in cells specialized for rapid secretion on demand (ex: digestive enzymes). These cells concentrate and store their products in ***secretory vesicles***. These vesicles are released from the TGN and are positioned near the plasma membrane. In response to extra-cellular signals (ex: hormones) secretory vesicles fuse with the plasma membrane and release their contents to the extra-cellular region by exocytosis.

Membrane components of secretory vesicles are removed from the plasma membrane by endocytosis almost as fast as they are added to it by exocytosis to maintain the composition of the cell membranes [Alberts et al., 2002, Chapter 13].

CHAPTER 3. METHODS AND MATERIALS

This chapter describes the methodology we propose for developing an automated localization predictor. Section 3.1 explains how we select a representative set of sub-cellular components for our study. Section 3.2 details our approach to identify factors potentially implicated in protein localization. ID3 decision tree is briefly explained and justification for its selection as the learning method of choice is described in section 3.3. Section 3.4 details the data set and the criteria applied for its obtention. Section 3.5 describes the method used to compute the mentioned features for the example proteins in our data set. Building of the decision tree and its utilization for prediction are further detailed in sections 3.6 and 3.7.

3.1. Localization Sites Selection

We obtained a directed graph, having Cell as the root and depicting the hierarchy of sub-cellular structures as they appear in the Cellular Component portion of Gene Ontology (GO) from the QuickGO Browser site at http://www.ebi.ac.uk/ego/. This graph was subdivided into disjoint subgroups each containing a distinct major sub-cellular component and all sub-components thereof. For the purpose of the current work, we selected one localization site from within each sub-graph. The criteria used to guide the selection were: (i) Biological importance of the compartment - for example nucleolus is considered as a distinct site of localization, (ii) Availability of evidences of experimentally discovered localizations from existing literature, and (iii) An effort to keep the total number of predictable sites within a manageable limit. The result of this exercise (Table-2) is the set of fungal sub-cellular components for which localization may be attempted.

Sub-cellular Site	GO Id Number
Extracellular region	GO:0005576
Cell wall	GO:0009277
Plasma membrane	GO:0005886
Nucleus	GO:0005634
Nuclear envelope	GO:0005635
Nucleolus	GO:0005730
Cytoplasm	GO:0005737
Cell cortex	GO:0005938
Cytoplasmic membrane-bound vesicle	GO:0016023
Endosome	GO:0005768
Golgi apparatus	GO:0005794
Peroxisome	GO:0005777
Ribosome	GO:0005840
Endoplasmic reticulum	GO:0005783
Endoplasmic reticulum lumen	GO:0005788
Endoplasmic reticulum membrane	GO:0005789
Vacuole	GO:0005773
Vacuolar lumen	GO:0016023
Vacuolar membrane	GO:0005768
Mitochondrion	GO:0005794
Mitochondrial inner membrane	GO:0005777
Mitochondrial intermembrane space	GO:0005840
Mitochondrial outer membrane	GO:0005783
Mitochondrial matrix	GO:0005788

Table-2: Representative cellular components selected from Gene Ontology

In the above table, the hierarchical containment of sub-cellular components (i.e. "is part of" relation in the ontology) is depicted by nesting of the sub-cellular sites. Additional notes on grouping of localization sites are found in Appendix–1.

3.2. Feature Selection

Eukaryotic cells are endowed with many sub-cellular organelles and non-organelle structures. Each of these structures exists within a unique biochemical environment and performs one or more specific functions. Abstracting from the environmental parameters that are external to the

proteins and are hard to itemize, we propose to consider three main categories of characteristic features as factors important in determining protein locations.

3.2.1. Physiochemical Features

Length, molecular weight, average hydrophobicity and isoelectric point are four important measurable physical features that characterize proteins. As for chemical features, it has been shown that there is a correlation between amino acid composition and cellular location of proteins [Cedano et al., 1997]. We propose to consider the following distinctive chemical features: a) amino acids percentage composition, b) distribution of residue size (tiny, small, medium-sized and bulky), c) basic, acidic, uncharged-polar, non-polar, aromatic and aliphatic composition, d) hydrophobic, weakly-hydrophobic and hydrophilic content, and e) dipeptide composition.

Studies of dipeptide frequencies in the literature reveal that GF, GY, NG, NT are among the dipeptides that occur most frequently in proteins [Campion et al., 2001] however, to our knowledge, no results have been reported for this occurrence specifically in fungi. We use the amino acid sequences of a set of representative fungi to find the dipeptides that predominantly occur in fungal proteins. Calculations performed on a large set (120734) of fungal proteins from 20 different fungal species from RefSeq database of NCBI (http://www.ncbi.nlm.nih.gov/RefSeq/ Release 7) revealed that 5 specific dipeptides (SS, LL, LS, AA, SL) occur more often than others in fungal protein sequences. The occurrence of these dipeptides is some 20% more frequent than that of other 395 dipeptides (please see Table-3 for a list of 10 dipeptides occurring most often in fungi). Based on these findings, we select only the 5 most frequently occurring dipeptides as features to study for sub-localization.

Dipeptide	Frequency
SS	587762
LL	515712
LS	459168
AA	453526
SL	452435
LA	414964
AL	402812
AS	395638
SA	367342
ST	345803

Table-3: 20 most predominantly occurring dipeptides in fungal proteins

In total 42 features were selected for this category (Table-A2 in Appendix–2).

3.2.2. Protein Signature

Most of the proteins can be grouped, on the basis of similarities in their structure and function, into a limited number of families and family domains. A ***domain*** is a conserved protein region that can independently fold as a structural unit and often contains a hydrophobic core [Ponting, 2001]. A ***motif*** is a region of a domain containing conserved active or binding site residues, or else a conserved sequence outside a domain that may adopt folded conformation only in association with its binding ligands [Ponting, 2001]. A motif is usually located in a short well-conserved region - typically an enzyme catalytic site, prosthetic group and an attachment site [Sigrist et al., 2002]. These considerations indicate that *motifs* have a higher frequency of occurrence than *domains*. Moreover, a *motif* is a pattern or regular expression that is a quantitative descriptor: it either matches or it does not. We therefore choose *motifs* (as opposed to *domains*) to be used as protein signature for determination of localization. PROSITE database (http://ca.expasy.org/prosite/) is a compilation of protein families. It uses one or more patterns (functional motifs) to represent each protein family. Through interaction with environmental factors within the cell, these functional motifs confer to the proteins the specific functionality that

characterizes the family. We propose to extract these functional motifs from PROSITE and use them as characteristic features for fungal protein localization.

On release 19 of PROSITE, we found 124 patterns associated with fungal proteins. We scanned the totality of fungal protein sequences in RefSeq database from NCBI (mentioned above) to ensure that all these 124 selected protein motifs do occur in some proteins. The result of this scanning revealed 5 motifs (PS00203, PS00269, PS00574, PS0812 and PS60011) that are not contained in any sequence and could therefore be done without. The total number of PROSITE motifs that are considered for training is therefore 119 (Table-A3 in Appendix–3).

3.2.3. Targeting Signals

Proteins that take part in a given biological process are often from diverse families. Therefore, family-specific functional motifs alone are not sufficient for targeting proteins in the pathways in which they are involved. The final address of proteins that enter the secretory pathway is largely specified by short signals (motifs) within the protein that determine interaction with particular elements of the pathway [Pringle et al., 1997]. Indeed, numerous experimental studies provide well-documented information on how the sorting receptors within the cell recognize various motifs in proteins' amino acids sequences and selectively target them to appropriate destinations. Through review of findings in the literature, we selected a set of 17 motifs that have been recurrently associated with targeting of specific cellular compartments. Table-4 provides a list of selected targeting motifs and the corresponding source of the information.

Feature name	Feature description	Hypothesized Targeting Site
Cleavable signal Peptide (CLSP)	Signal peptide that is cleaved prior to being exported out of the cell	ExtraCellular Region [von Heijne, 1986], [von Heijne, 1987]
Transmembrane segment (TMS)	Transmembrane segment	Plasma membrane [Nakai and Horton, 1999]
Mature protein Transmembrane Segment (MTMS)	Mature protein (protein with signal peptide cleaved) containing a trans-membrane segment	Plasma membrane [Nakai and Horton, 1999]

Feature name	Feature description	Hypothesized Targeting Site
Length of MTMS (MTMSLEN)	Length of mature protein transmembrane segment Values: (>20, 18-20, <=17)	
ER retention / retrieval signal (ERRS)	xAIAKE or KDEL at C-terminus or HDEL at C-terminus	ER [Pelham et al., 1988], [Wrzeszczynski and Rost, 2004]
Endosomal signal (ES)	DAKSS	Endosome [Rohrer et al., 1993]
Vacuolar targeting signal (VTS)	QRPL near N-terminus or [TIK]LP[LKI]	Vacuole [Valls et al., 1987], [Nakai and Horton, 1999]
GPI attaching signal (GAS)	[AGST]x[AGST][DEHKRNQST](5-10)[ACGPYILMVFW](15-20)	Cell Wall [Gerber et al., 1992]
ER transmembrane segment (ERTMS)	MTMS length <= 17 and at least one of the following motifs: xKKxx or x(1,3)RR or x(1,2)RxR	ER membrane [Nilsson et al., 1989], [Schutze et al., 1994], [Townsley et al., 1993]
Vacuolar transmembrane segment (VTMS)	MxHCxMxM	Vacuole [Bellemare et al., 2002]
Nuclear membrane localization signal (NMLS)	GLFG{KRHDE}(1-20)GLFG or FGFG	Nuclear Envelope [Pante et al., 1994], [Davis and Blobel, 1986]
Nuclear localization signal (NLS)	[KR](4-6)[PG] or [KR](3)[HP][PG] or Px(1-3)[KR](3) or Px(1-3)[KR](2)x[KR] or Px(1-3)[KR]x[KR](2) or [KR](2)x(9-12)[KR](3) or [GR](2)	Nucleus [Christophe et al., 2000] [Dono et al., 1998]
Peroxisomal targeting signal (PTS)	[SAC][KRH]L> or [RK][LVI]x(5)[HQ][LA]	Peroxisome [Gould et al., 1990], [Swinkels et al., 1991]
Peroxisomal membrane signal (PMS)	[DEHKNQR](20)	Peroxisomal membrane [Dyer et al., 1996]
Mitochondrial transfer peptide (MTP)	xRx(30) or x(2)Rx(29)	Mitochondrial inner or outer membrane [Emanuelsson et al., 2001]
Mitochondrial matrix transport signal (MMTP)	MLSLRQSIRFFKPATRTLCSSRYLL	Mitochondrial Matrix [Drin et al., 2003]
Vesicular signal (VS)	YQRL	Cytoplasmic membrane-bound Vesicles [Bos et al., 1993]

Table-4: Protein targeting motif features and their corresponding hypothesized target sites

Considering the three mentioned categories, a total number of 178 (42+119+17) features were thus selected for localization study.

3.3. Selection of Learning Tool

In analyzing various characterizing features of proteins, we observed the frequent presence of nominal values (for example: hydrophobicity, length, etc.). Moreover, there may be cases where data are missing (for example the length of transmembrane segment). These types of data lend themselves better to a decision tree type of algorithm than to support vector machines.

ID3 is a machine learning method that performs a Hill-Climbing strategy in the possible search space of a decision tree, examining, at each stage of the search, all tests that could be made to extend the tree and selecting greedily the branch that moves the longest distance towards a goal state. Its heuristic is based on the assumption that the smallest tree that correctly classifies all the training examples is the one most likely to make correct future classification [Luger, 2002].

Our choice of ID3 as the learning tool for sub-cellular localization was motivated by the following considerations: a) Our training examples are known at the time of building the classifier so there is no need for incremental learning, b) ID3 has not been directly explored to this date as the learning method of choice in sub-cellular localization of protein and our effort may lead to some insight into its applicability in this domain, c) ID3 uses all training examples at each step in the search to make statistically based decisions regarding how to refine its current hypothesis. One advantage of using statistical features of all the examples is that the resulting search is much less sensitive to errors in individual training examples [Mitchell, 1997], d) Our training set is quite large and is assumed to be representative of the fungal protein population. This favours the ID3's heuristic as for shorter (less deep) trees there will necessarily be less generalization made that will not be supported by the data and e) ID3's inductive bias has been proven to provide a high efficiency for very large sizes of training sets [Quinlan, 1983].

3.4. Data Set

3.4.1. Choice of Species

We have selected three well-studied fungal species to serve as representative organisms of fungi: (i) *Saccharomyces cerevisiae* (budding yeast) is used in baking and brewing. It has been extensively studied as a eukaryotic model organism in molecular biology, (ii) *Candida albicans* is a fungus from the Saccharomycetaceae family. It is an important cause of mortality in immunocompromised humans, (iii) *Schizosaccharomyces pombe* (fission yeast) is almost as easily cultured and manipulated as yeast. It is well characterized as to classical and molecular genetics, its nuclear genome has been sequenced, and it is an alternative fungal model system, comparable to that of the budding yeast. There are three publicly available databases that provide detailed information on proteins belonging to these species. They are accessible at www.yeastgenome.org/ , www.candidagenome.org/ , and www.sanger.ac.uk/Projects/S_pombe/ for budding yeast, Candida albicans, and fission yeast, respectively.

3.4.2. Source of Data

We downloaded localization data for Saccharomyces cerevisiae, Candida albicans and Schizosaccharomyces pombe from the annotation section of Gene Ontology website (Table-5).

http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.sgd.gz http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.cgd.gz http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.GeneDB_Spombe.gz Release validated by GO on 13 Feb. 2006

Table-5: Source of localization information for fungal proteins

The current experimental finding of fungal protein localizations is far from complete and most of the localizations found in the selected sources (as well as elsewhere in the literature) are inferred from electronic annotation (IEA) or from sequence or structural similarity (ISS). In order to choose more biologically significant localizations for our training set we restricted our examples to those obtained from experimental sources. Therefore, from the downloaded data we solely

selected the localizations evidenced as IDA, IGI, IMP and IPI (inferred from direct assay, genetic Interaction, mutant phenotype and physical interaction respectively) and mapped these entries to the designated sub-cellular sites listed in Table-2. In the process of mapping, redundant duplicate protein IDs could be generated. For example, YGL047W is reported to localize to "UDP-N acetylglucosamine transferase complex" as well as to "extrinsic to ER membrane" sites. Since both these sites are part of the ER membrane, two entries are generated both reporting the mentioned protein to localize to the ER membrane. We removed all such duplicate protein IDs entries. Moreover, as the number of available localized examples having the mentioned evidence codes for some sites was considered insufficient (less than 15) for training purposes, we combined ER lumen with ER membrane (to constitute a single site, ER) and combined vacuolar lumen with vacuolar membrane (to form a single site, vacuole). We also included ribosome and cytoplasmic membrane-bound vesicle proteins with those of cytoplasm. We therefore compiled, as our data set, 5659 reported localizations for the 17 sites listed in Table-6.

Localization Site	Count of reported localizations per site
Cell Cortex	72
Cell Wall	140
Cytoplasm	2821
Endosome	39
ER	74
Extracellular Region	35
Golgi	50
Mitochondrial inner membrane	73
Mitochondrial Matrix	42
Mitochondrial outer membrane	23
Mitochondrial Intermembrane Space	26
Nuclear Envelope	81
Nucleus	1755
Nucleolus	147
Peroxisome	24
Plasma membrane	211
Vacuole	46
Total	5659

Table-6: Data set of experimentally reported fungal protein localizations sites

In this set there are proteins that are reported to localize to two or more distinct locations. For example, YAL016W is reported with IDA evidence to localize to nucleus and cytoplasm

(Saccharomyces Genome Database, SGD [Cherry et al., 1997]). Table-7 depicts the frequency of single-site and multi-site proteins in our data set for localization to 17 sites.

Number of sites of Localization	Total number of proteins per category	Total number of localizations reported
4	2	8
3	55 (~1%)	165
2	1014 (~22%)	2028
1	3458 (~76%)	3458
ALL	4529	5659

Table-7: Frequency of single-site and multi-site localized examples in the data set

3.5. Feature Calculation

Appropriate modules (Appendix-4) were developed to compute the values of the selected features for each amino acid sequence in our data set.

3.5.1. Calculation of Physiochemical Features

For the discretization of nominal values of compositional features, we first determined, for each compositional feature, the range of values it takes within the given set of training examples. We then divided this range into three equal segments representing regions of low, medium and high abundance. For example, the range of calculated values for isoelectric point of our training set proteins is [5.16 – 7.18]. Any feature value falling into the region [5.16 – 5.83], [5.84 – 6.50], and [6.51 – 7.18] (respectively) is considered to have a low, medium and high (respectively) isoelectric point. Each compositional feature of each of our data set example proteins was thus assigned a representative discrete value based on the region on which its nominal value fell.

For each compositional features used, Appendix–2 displays the range of its nominal and discretized values.

3.5.2. Calculation of Protein Signature

Each example protein, in our data set was scanned for each of the functional motifs and was thereby assigned a corresponding value (yes/no) for the presence or absence (respectively) of the given motif.

3.5.3. Calculation of Targeting Signals

Table-4 lists the hypothesized localization site(s) that correspond to each proposed targeting motif. We scanned our example proteins sequences for these motifs and counted and/or determined the presence or absence of each motif in each of the example proteins. Each feature of each example protein was thus assigned a discrete value based on this evaluation.

The results of the 178 feature values extracted for all proteins in our data set thus constitute a feature matrix. Each tuple of this matrix consists of a protein name, a reported localization and values of the 178 selected features.

3.6. Building of Decision Tree

A decision tree is built from the set of training examples in a top-down recursive manner. At each node of the tree (starting at the root) one of the protein features is selected following ID3's heuristic [Quinlan, 1983], in such a way as to minimize a certain measure of disorder. This disorder represents the expected cost of completing the tree if the split is based on the selected feature [Luger, 2002]. Training examples are then divided into disjoint subsets such that in each subset all the examples have the same value for the selected feature. We generate as many children at each node as there are such subsets and assign each subset to the corresponding child node. When all the example proteins associated with a given node have the same localization, we have reached a ***terminal node*** for which no further splitting is required and the recursive process

may terminate. It may so happen that at a certain node no decision could be made as to which feature to split upon next. This could occur, for example, when the set of examples associated with that node consists of 2 or more proteins with identical feature values but different localization sites. We consider such a node as a ***multi-site terminal node*** to which we associate the mentioned localization sites. It is important to note that not all potentially possible values of the splitting feature are always represented in examples associated with a given node. No branch corresponding to the missing feature values therefore emanates from such nodes. This may lead to the inability of the system to predict localization of proteins presenting certain combination of features values, as we will see in the next section.

3.7. Localization Based on the Decision Tree

Classification of a query protein is accomplished by starting at the root of the decision tree and traversing the tree until a terminal node is reached. Localization site associated with that terminal node represents the prediction for the query protein. During the traversal, the branch taken at each node is determined by the value of the query protein for the feature associated with that node. If a node is reached at which no emanating branch corresponds to the feature value of the query protein, then no prediction could be made for this query protein. If the tree traversal leads to a multi-site terminal node (as defined in section 3.6) then the query protein is predicted to target all the localizations associated with that node.

CHAPTER 4. RESULTS

In this chapter, we report the results obtained from using a decision tree to localize fungal proteins as well as our findings from the analysis of the said tree. Section 4.1 describes the per protein performance measures used to evaluate our system. It also lists the proteins for which our system could not provide a prediction. Analysis of the built decision tree leads to various findings of biological interest: Sections 4.2 and 4.3 detail the actual features that are found to play a role in protein localization and enumerate the ones with the highest discriminatory power. Section 4.4 proceeds with the validation of the proposed targeting motifs. Finally, section 4.5 identifies those protein features among the proposed ones that are always present in localization to certain sites. These features are, therefore, expected to constitute necessary conditions for such localizations.

4.1. Performance

We built a decision tree that classifies the proteins based on their extracted feature values and then used this tree to predict the localization(s) of query proteins using the method described in sections 3.6 and 3.7. To evaluate the performance of the system, we followed a 5-fold cross-validation. We divided our data set into five subsets (S_1, S_2, S_3, S_4 and S_5) of approximately equal size ensuring that: a) single-site, double-site and triple-site proteins are each equally distributed among the five subsets, b) quadruple-site proteins, whose number is less than 5, are distributed into distinct subsets, and c) each subset contains approximately the same proportion of examples located to each sub-cellular site. In each of the 5 cross validations, a distinct set among these five subsets (for example set S_2) was used as the query set and the totality of the remaining 4 subsets (for example $\{S_1, S_3, S_4, S_5\}$) was used as training set. The set of experimentally reported

localizations was then compared to the corresponding prediction set using the following performance criteria:

- *Partial prediction*: system can predict SOME of the reported localizations. We determine the percentage of proteins for which at least one true positive (TP) prediction was made.
- *Total prediction*: system can predict ALL of the reported localizations but may also predict some that are not reported. We determine the percentage of proteins for which no false negative (FN) prediction was made.
- *Identical prediction*: system can predict ALL of the reported localizations and none other. We determine the percentage of proteins for which neither FN nor false positive (FP) prediction was made.

Table-8 and Table-10 depict the results of these calculations for localizations to 17 and 9 sites respectively. Table-9 and Table-11 list the proteins for which no localization could be predicted.

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
CC	49	4	8%	4	8%	1	2%
CW	101	46	46%	46	46%	30	30%
C	1878	1290	69%	1290	69%	700	37%
E	20	1	5%	1	5%	0	0%
ER	38	4	11%	4	11%	1	3%
XR	23	9	39%	9	39%	7	30%
GA	20	5	25%	5	25%	3	15%
MI	48	8	17%	8	17%	2	4%
MS	11	2	18%	2	18%	0	0%
MM	25	0	0%	0	0%	0	0%
MO	15	0	0%	0	0%	0	0%
N	981	551	56%	551	56%	275	28%
NE	33	4	12%	4	12%	1	3%
NL	51	12	24%	12	24%	2	4%
P	7	0	0%	0	0%	0	0%
PM	133	59	44%	59	44%	22	17%
V	25	4	16%	4	16%	0	0%
All Single-Site Proteins	**3458**	**1999**	**58%**	**1999**	**58%**	**1044**	**30%**
CC – CW	1	0	0%	0	0%	0	0%
CC – C	14	11	79%	0	0%	0	0%
CC – N	4	4	100%	0	0%	0	0%
CC – PM	1	0	0%	0	0%	0	0%

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
CW – C	20	18	90%	3	15%	2	10%
CW – ER	1	0	0%	0	0%	0	0%
CW – XR	7	5	71%	0	0%	0	0%
CW – PM	2	1	50%	0	0%	0	0%
C – E	9	6	67%	0	0%	0	0%
C – ER	26	17	65%	1	4%	1	4%
C – XR	2	1	50%	0	0%	0	0%
C – GA	22	18	82%	4	18%	1	5%
C – MI	20	13	65%	0	0%	0	0%
C – MS	13	9	69%	0	0%	0	0%
C – MM	14	13	93%	0	0%	0	0%
C – MO	6	2	33%	1	17%	0	0%
C – N	646	577	89%	157	24%	90	14%
C – NE	19	12	63%	4	21%	0	0%
C – NL	6	4	67%	1	17%	0	0%
C – P	14	12	86%	0	0%	0	0%
C – PM	49	37	76%	7	14%	2	4%
C – V	9	9	100%	1	11%	0	0%
E – GA	6	1	17%	1	17%	0	0%
E – V	1	0	0%	0	0%	0	0%
ER - NE	2	0	0%	0	0%	0	0%
ER - PM	2	1	50%	0	0%	0	0%
MI - MM	1	0	0%	0	0%	0	0%
MI - MO	1	0	0%	0	0%	0	0%
MI – N	1	0	0%	0	0%	0	0%
MS - MO	1	0	0%	0	0%	0	0%
N – NE	14	9	64%	1	7%	0	0%
N – NL	72	50	69%	16	22%	8	11%
N – PM	2	1	50%	0	0%	0	0%
NE - NL	1	0	0%	0	0%	0	0%
NE - PM	1	0	0%	0	0%	0	0%
PM – V	4	3	75%	2	50%	2	50%
All Double-Site Proteins	**1014**	**834**	**82%**	**199**	**20%**	**106**	**10%**
CC – C - N	2	2	100%	0	0%	0	0%
CC – C - PM	1	1	100%	0	0%	0	0%
CW – C - MS	1	0	0%	0	0%	0	0%
CW – C - N	3	3	100%	0	0%	0	0%
CW – C - PM	3	3	100%	0	0%	0	0%
C - E - PM	1	1	100%	0	0%	0	0%
C - E - V	1	1	100%	0	0%	0	0%
C – ER - N	1	0	0%	0	0%	0	0%
C – ER - NE	2	1	50%	0	0%	0	0%
C – XR - NE	2	0	0%	0	0%	0	0%
C – GA - PM	1	1	100%	0	0%	0	0%
C – MI - MM	2	1	50%	0	0%	0	0%
C - N - NE	5	4	80%	0	0%	0	0%
C - N - NL	15	14	93%	1	7%	0	0%
C - N - P	1	1	100%	0	0%	0	0%
C - N - PM	4	4	100%	1	25%	0	0%

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
C - N - V	2	2	100%	1	50%	0	0%
C - P - PM	2	2	100%	0	0%	0	0%
C – PM - V	3	3	100%	0	0%	0	0%
ER - PM - V	1	1	100%	0	0%	0	0%
N – NE - NL	2	1	50%	0	0%	0	0%
All Triple-Site Proteins	**55**	**46**	**84%**	**3**	**5%**	**0**	**0%**
CW-C-XR-PM	1	1	100%	0	0%	0	0%
C-E-ER-GA	1	1	100%	0	0%	0	0%
All Quadruple-Site Proteins	**2**	**2**	**100%**	**0**	**0%**	**0**	**0%**
All Proteins	**4529**	**2881**	**64%**	**2201**	**49%**	**1150**	**25%**

Legend: CC: cell cortex, CW: cell wall, C: cytoplasm, E: endosome, ER: endoplasmic reticulum, XR: extracellular region, GA: Golgi apparatus, MI: mitochondrial inner membrane, MS: mitochondrial inter-membrane space, MM: mitochondrial matrix, MO: mitochondrial outer membrane, N: nucleus, NE: nuclear envelope, NL: nucleolus, P: peroxisome, PM: plasma membrane, V: vacuole.

Table-8: Per protein performance evaluation measures for localizations to 17 sites

YOR383C,YOR009W,YER087C-B,YDR379C-A,YPR063C,YGL103W,YKR092C,YPL096C-A, YAL015C,SPCC1840.02C,YDR071C,YDL167C,YBR264C,YDR231C,SPCP31B10.03C,YPL163C, YFL017W-A,YBR258C,YBR193C,SPAC6G9.11,YDL085C-A,YGR204W,YJR002W,YML129C, YER074W-A,ORF19.5636,YOR142W,YLR066W,YGL221C,YGL126W,YGL011C,SPAC20G4.04C, YOR159C,YJL056C,YGR074W,YKL046C,YOL154W,YDR363W-A,SPAC16.01,ORF19.3014, YJL078C,YER177W,YPR133W-A,YOR181W,YOR089C,YNR017W,YJL151C,YER043C, YEL048C,YDR493W,YDR086C,SPBC1677.02,YIL136W,YKL054C,YJR052W,YLR332W,YLR347C, YLR150W,YKL122C, YDR439W,YIL008W,YDR378C,SPBC13E7.09,SPAC3A12.14,ORF19.6975, YER053C-A,YPR020W,ORF19.3138,YML028W,YJL143W,YHR051W,YGR076C,YGL191W, YDR032C,SPCC1840.01C,YLR074C

Table-9: 76 query proteins with no prediction in localization to 17 sites

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
CW	101	47	47%	47	47%	36	36%
C	1991	1361	68%	1361	68%	903	45%
ER	38	2	5%	2	5%	0	0%
XR	23	10	43%	10	43%	9	39%
GA	20	2	10%	2	10%	2	10%
M	102	9	9%	9	9%	3	3%
N	1154	624	54%	624	54%	404	35%
PM	133	48	36%	48	36%	24	18%
V	25	4	16%	4	16%	0	0%
All Single-Site Proteins	**3587**	**2107**	**59%**	**2107**	**59%**	**1381**	**39%**
CW - C	21	17	81%	2	10%	1	5%
CW - ER	1	0	0%	0	0%	0	0%

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
CW - XR	7	3	43%	0	0%	0	0%
CW - PM	2	1	50%	0	0%	0	0%
C – ER	26	18	69%	0	0%	0	0%
C – XR	2	2	100%	0	0%	0	0%
C – GA	28	19	68%	4	14%	2	7%
C – M	55	35	64%	5	9%	0	0%
C – N	698	643	92%	146	21%	128	18%
C – PM	54	38	70%	5	9%	3	6%
C – V	11	9	82%	1	9%	1	9%
ER – N	2	0	0%	0	0%	0	0%
ER – PM	2	1	50%	0	0%	0	0%
M – N	1	0	0%	0	0%	0	0%
N - PM	3	2	67%	0	0%	0	0%
PM – V	4	3	75%	2	50%	2	50%
All Double-Site Proteins	**917**	**791**	**86%**	**165**	**18%**	**137**	**15%**
CW-C-M	1	1	100%	0	0%	0	0%
CW-C-N	3	3	100%	0	0%	0	0%
CW-C-PM	3	2	67%	0	0%	0	0%
C-ER-GA	1	1	100%	0	0%	0	0%
C-ER-N	3	3	100%	0	0%	0	0%
C-XR-N	2	1	50%	0	0%	0	0%
C-GA–PM	1	1	100%	0	0%	0	0%
C-N– PM	4	3	75%	0	0%	0	0%
C-N–V	2	2	100%	0	0%	0	0%
C-PM–V	3	3	100%	1	33%	1	33%
ER-PM–V	1	1	100%	0	0%	0	0%
All Triple-Site Proteins	**24**	**21**	**88%**	**1**	**4%**	**1**	**4%**
CW-C-XR–PM	1	1	100%	0	0%	0	0%
All Quadruple-Site Proteins	**1**	**1**	**100%**	**0**	**0%**	**0**	**0%**
All Proteins	**4529**	**2920**	**64%**	**2273**	**50%**	**1519**	**34%**
Legend: CW: cell wall, C: cytoplasm, ER: endoplasmic reticulum, XR: extracellular region, GA: Golgi apparatus, M: mitochondrion, N: nucleus, PM: plasma membrane, V: vacuole.							

Table-10: Per protein performance evaluation measures for localizations to 9 sites

YPR024W,YNL146W,YMR132C,YLR395C,YLR064W,YIL124W,YGL256W,YDR106W,YAL020C, SPCC962.06C,YIL008W,YEL007W,YFL014W,SPCP1E11.04C, SPAC1805.08,YOL052C-A, YGL226C-A,YNL015W,YMR252C,YLR066W,YKL187C,YKL152C,YJL054W,YGR120C, YPR086W,YOR062C,YML022W,YHR040W,YBR252W,YMR251W-A,ORF19.220,ORF19.3548.1, YER053C-A,YPL231W,YOR045W,YNL024C,YML012W,YER043C,SPBC1718.03,YOR159C, YHR089C,YGR074W,YDR378C,YCL054W,YNL098C,YKR020W,ORF19.5558,YER087C-B, YNR002C,YNL211C,YHR051W,YGR285C,YBR096W,YFL017W-A,YJL056C,SPAC23H4.18C, YCR090C,YKR092C,YDR123C,YPL163C,YGR192C,YGL058W,YDL125C

Table-11: 63 query proteins with no prediction in localization to 9 sites

4.2. Characteristic Features Determination

In order to determine the features that are effective in localization, a full decision tree was built for 5659 examples using 178 features (Appendix-2, Appendix-3 and Table-4) and 17 localization sites (Table-6). We found that 76 feature values were actually used in the decision tree in order to determine the localizations. These features are therefore expected to play a role in sorting the target locations. Table-12 provides a breakdown of these features in different category.

Feature Type	Feature Name
18 Physio-chemical	ACID, ALIP, AROM, BASIC, BULKY, HYDPHB, HYDPHBC, HYDPHL, IEP, LEN MED, MW, NONPOL, POLUNCH, SMALL, TINY, TMS, WKHYDPHB
3 Dipeptide	LL, LS, SL
23 Functional Motifs	PS00068, PS00079, PS00331, PS00406, PS00414, PS00442, PS00455 PS00585, PS00591, PS00614, PS00719, PS00776, PS00777, PS00789 PS00885, PS00931, PS00934, PS00955, PS01028, PS01070, PS01095 PS01101, PS01224
20 Residue Composition	A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
12 Targeting Motifs	RLCLSP, RLERRS, RLERTMS, RLES, RLGAS, RLMTP, RLNLS, RLNMLS, RLPMS, RLPTS, RLVS, RLVTS

Table-12: 76 Characteristic fungal protein features implicated in sub-cellular localization

4.3. Decision Tree Discriminatory Power

The decision tree obtained using the evidence-specific examples also indicates the relative order in which various physical and biochemical features may be searched when attempting to localize a protein. The higher up a property lies in the tree, the more informative is its testing in narrowing down the possibilities towards the final localization(s), hence the higher is its discriminatory power. Table-13 depicts the features found in the top 4 levels of the decision tree. The order in which various features appear in different levels of the tree is indicative of their relative discriminatory power in localization.

Level 1	Level 2	Level 3	Level 4
Hydrophilic content	Leucine content	Polar uncharged content	Alanine content
			Threonine content
			Aliphatic content
			Small residues content
		Tryptophan content	Hydrophobicity
			Isoleucine content
	Tiny residues content	Transmembrane Segment	Tyrosine content
			Polar uncharged content
		Polar uncharged content	Transmembrane Segment
			Tyrosine content
		Arginine content	Cysteine content
			Alanine content
	Lysine content	Asparagine content	Glutamine content
			PMS targeting singal
			Tyrosine content
		Polar uncharged content	Alanine content
			Ribosomal protein S5 motif
		Alanine content	Isoleucine content
			Histidine content

Table-13: Top 4 levels of decision tree

4.4. Targeting Motifs Validation

The constructed decision tree is used to validate the hypothesized targeting motifs (Table-4). Targeting motif feature values attached to each node of the decision tree built from our data set are extracted and compared to values proposed to favour specific targets as per the literature. Table-14 shows the frequency of occurrence of 15 targeting motifs in proteins that localize to 12 corresponding sub-cellular sites. Only 12 out of the 17 designated sub-cellular sites (Table – 6) are presented in this table because we did not consider any targeting motif for the remaining 5 sites (cell cortex, cytoplasm, Golgi, mitochondrial inter-membrane space and nucleolus).

Localization Site	Count of protein per site	Targeting Motif(s)	Count of proteins containing motif(s)	Percentage of proteins with targeting motif(s) localized to hypothesized site
Mitochondrial inner membrane	73	MTP	73	100
Mitochondrial outer membrane	23	MTP	23	100
Nucleus	1755	NMLS / NLS	1663	95
ER	74	ERRS / ERTMS	62	84
Plasma membrane	211	TMS / MTMS	170	81
Extracellular Region	35	CLSP	25	71
Peroxisome	24	PTS / PMS	9	38
Vacuole	46	VTS / VTMS	9	20
Nuclear Envelope	81	NMLS	5	6
Mitochondrial Matrix	42	MMTP	0	0
Endosome	39	ES	0	0
Cell Wall	140	GAS	0	0

Table-14: Occurrence of targeting motifs in proteins of the targeting set

From the result in the last six rows of Table-14 (indicating that only less than or equal to 38% of the analyzed proteins contain the proposed targeting motifs), even after considering the margin of error, we can claim that the stated targeting motifs are not necessary for localization of fungal proteins to peroxisome, vacuole, nuclear envelope, mitochondrial matrix, endosome and cell wall. But, are there some features that are necessary conditions for localization to these sites?

4.5. Necessary Conditions for Localization to Specific Sites

We scan for feature values that uniformly take the same value across ALL the examples of a given localization site. Our data set of available examples (5659 proteins from Table-6) was used to search such occurrences among values of all 178 studied features, regardless of any hypothesized association between a given feature value and a sub-cellular location site. The results are compiled in Table-15.

Location Site	Set of Protein Features
Cell Cortex	Low in Cysteine, Glycine, Threonine and dipeptides SS, LL and AA Mitochondrial Transfer Peptide
Cell Wall	Low in Methionine and dipeptide LL
Endosome	Low in Alanine, Cysteine, Glutamine, Serine, Threonine and dipeptides SS and AA Medium content of Aliphatic residues ER Trans-membrane segment Mitochondrial Transfer Peptide
ER	Low in Proline, Glutamine and dipeptides SS and AA
Extra-cellular Region	Low in Cysteine, Glutamic Acid, Methionine, Proline, Glutamine, Arginine and dipeptides LL and AA Short Sequence Length
Golgi Apparatus	Low in Alanine, Cysteine, Methionine, Proline, Glutamine, Serine, Threonine and dipeptides SS and AA Short Sequence Length Mitochondrial Transfer Peptide
Mitochondrial inner membrane	Low in Cysteine, Serine and dipeptides SS and AA Short Sequence Length Mitochondrial Transfer Peptide
Mitochondrial Inter-membrane Space	Low in Asparagine, Serine, Threonine and dipeptides SS, AA, LS and SL Medium content of Non-polar residues Short Sequence Length Mitochondrial Transfer Peptide
Mitochondrial Matrix	Low in Cysteine, Proline, Serine and dipeptides SS, LL and AA Medium content of Non-polar residues Short Sequence Length and Medium Hydrophobicity Mitochondrial Transfer Peptide
Mitochondrial outer membrane	Low in Cysteine, Glycine, Methionine, Proline, Glutamine, Serine, Threonine and dipeptides SS, LL and AA Medium content of Non-polar residues Short Sequence Length; Medium Isoelectric point Nuclear localization signal Mitochondrial Transfer Peptide
Nuclear membrane	Low in Cysteine and dipeptides SS and AA Mitochondrial Transfer Peptide
Nucleus	Low in dipeptides SS and AA
Nucleolus	Low in Cysteine, Proline, Glutamine, Threonine and dipeptide AA
Peroxisome	Low in Cysteine, Glycine, Methionine, Proline, Glutamine, Serine, Threonine and dipeptides SS, LL and AA Medium content of Aliphatic, Non-polar, small, hydrophilic and hydrophobic residues Short Sequence Length; Medium Hydrophobicity and Isoelectric point Mitochondrial Transfer Peptide
Plasma membrane	Low in Cysteine
Vacuole	Low in Cysteine, Proline, Serine, Threonine and dipeptides SS and AA Medium content of Non-polar residues

Table-15: Necessary protein features for localization to specific sites

The above table reveals some interesting information about composition of proteins targeted to specific sites. For example Mitochondrial intermembrane space is the only sub-cellular

compartment, among those considered, that accommodates proteins with a low composition of SL dipeptide. The results also indicate certain chemical and physical properties that characterize proteins found in specific locations. For example Mitochondrial outer membrane and peroxisome proteins have medium isoelectric points. As for the hypothesized targeting motifs, this exercise shows that ERTMS occurs in all proteins localized to Endosome. The hypothesis that the features enumerated in Table-15 do constitute necessary conditions for localization to specific sites is consistent with the known localized examples as the decision tree is built based on the very examples. As we have used a large number of examples in our data set, we can expect them to be significant.

CHAPTER 5. DISCUSSION

In this chapter we discuss the significance of the performance evaluation reported and comment on the type and extent of potential errors our system is exposed to. Section 5.1 interprets and explains the results. It also evaluates the performance and coverage of the system by comparing its predicting power to that of similar existing systems. Section 5.2 explains how the hierarchical structure of the sub-cellular compartment is an impediment to successful localization of proteins to distinct sites. Section 5.3 enumerates the type of errors that may impact the performance of the system.

5.1. System Evaluation

5.1.1. Overall System Performance

In Table-8 (section 4.1), we observed a large dispersion between the performance measures corresponding to different localization sites. For example, "partial prediction" ranges from 0 to 100% and "total prediction" and "identical prediction" ranges from 0 to 50%. This result can be attributed to the variability of the number of examples to train the system. In fact, by sorting the results presented in the first section of Table-8 in descending order of the percentage of correct localization, we obtain Table-16 that clearly shows a correlation between the number of training examples and the proportion of correct localization. In the lower extreme, we see that no correct prediction could be made for mitochondrial matrix and outer membrane as well as peroxisome. As a general finding, all the proteins (single-site or multi-site localized) for which the system could not correctly predict any of the reported localizations had 25 or less examples in the training set. There were 3 single-site, 11 double-site and 3 triple-site localized proteins with such conditions.

Reported Localization Site	# Examples	# Correct Prediction	% Correct Prediction
Cytoplsm	1878	1290	69%
Nucleus	981	551	56%
Cell Wall	101	46	46%
Plasma Membrane	133	59	44%
Extracellular Region	23	9	39%
Golgi Apparatus	20	5	25%
Nucleolus	51	12	24%
Mitochondrial Inter-membrane Space	11	2	18%
Mitochondrial Inner Membrane	48	8	17%
Vacuole	25	4	16%
Nuclear Envelope	33	4	12%
Endoplasmic Reticulum	38	4	11%
Cell Cortex	49	4	8%
Endosome	20	1	5%
Mitochondrial Matrix	25	0	0%
Mitochondrial Outer Membrane	15	0	0%
Peroxisome	7	0	0%

Table-16: Percentage of correctly localized single-site proteins and its correlation with the number of example proteins in the training set for 17 site localization

Table-10 shows that combining the cellular compartments in order to predict among 9, instead of 17, localization sites results mainly in an improvement of "identical prediction" performance measure that passes from 25% to 34%. The capacity of the system to predict all the experimentally reported localization sites nonetheless does not exceed 50%. This result may be in part attributed to the use of "part-of" relationship when we compiled our data based on GO classification. For example a protein reported to localize to the nucleolus but predicted to go to the nucleus or one reported to target the cell cortex but predicted to go to the cytoplasm are each counted as a FP and a FN. On the other hand, the fact that the system predicts, quite correctly, that localization occurs in the nucleus and cytoplasm (respectively) is not considered as a TP result. This containment factor is a result of the hierarchical nature of the cellular compartments as depicted in Table-2. We propose to measure the extent of impact of such containment by considering a given predicted localization site as TP when it is identical to or when it contains the experimentally reported localization site (as per the hierarchy shown in Table-2). Table-17 depicts the performance measures obtained when such containments are considered in 9-site localization.

Localizations	# of Proteins	# of Partial Prediction	% of Partial Prediction	# of Total Prediction	% of Total Prediction	# of Identical Prediction	% of Identical Prediction
CW	101	47	47%	47	47%	36	36%
C	1991	1361	68%	1361	68%	903	45%
ER	38	26	68%	26	68%	0	0%
XR	23	10	43%	10	43%	9	39%
GA	20	17	85%	17	85%	2	10%
M	102	81	79%	81	79%	3	3%
N	1154	624	54%	624	54%	404	35%
PM	133	48	36%	48	36%	26	20%
V	25	18	72%	18	72%	0	0%
All Single-Site Proteins	**3587**	**2232**	**62%**	**2232**	**62%**	**1383**	**39%**
CW – C	21	17	81%	2	10%	1	5%
CW – ER	1	1	100%	0	0%	0	0%
CW – XR	7	3	43%	0	0%	0	0%
CW – PM	2	1	50%	0	0%	0	0%
C – ER	26	18	69%	18	69%	11	42%
C – XR	Deux	2	100%	0	0%	0	0%
C – GA	28	19	68%	18	64%	11	39%
C – M	55	35	64%	32	58%	20	36%
C – N	698	643	92%	146	21%	128	18%
C – PM	54	38	70%	5	9%	3	6%
C – V	11	9	82%	8	73%	4	36%
ER – N	2	2	100%	0	0%	0	0%
ER – PM	2	1	50%	0	0%	0	0%
M – N	1	1	100%	0	0%	0	0%
N – PM	3	2	67%	0	0%	0	0%
PM – V	4	4	100%	2	50%	2	50%
All Double-Site Proteins	**917**	**796**	**87%**	**231**	**25%**	**180**	**20%**
CW – C - M	1	1	100%	0	0%	0	0%
CW – C - N	3	3	100%	0	0%	0	0%
CW – C - PM	Trois	2	67%	0	0%	0	0%
C - ER - GA	1	1	100%	1	100%	0	0%
C - ER - N	Trois	3	100%	2	67%	2	67%
C - XR - N	2	1	50%	0	0%	0	0%
C – GA - PM	1	1	100%	0	0%	0	0%
C - N - PM	4	3	75%	0	0%	0	0%
C - N – V	2	2	100%	0	0%	0	0%
C – PM - V	3	3	100%	1	33%	1	33%
ER - PM - V	1	1	100%	0	0%	0	0%
All Triple-Site Proteins	**24**	**21**	**88%**	**4**	**17%**	**3**	**13%**
CW-C-XR-PM	1	1	100%	0	0%	0	0%
All Quadruple-Site Proteins	**1**	**1**	**100%**	**0**	**0%**	**0**	**0%**
All Proteins	**4529**	**3050**	**67%**	**2467**	**54%**	**1566**	**35%**
Legend: CW: cell wall, C: cytoplasm, ER: endoplasmic reticulum, XR: extracellular region, GA: Golgi apparatus, M: mitochondrion, N: nucleus, PM: plasma membrane, V: vacuole.							

Table-17: Performance evaluation measures for localizations to 9 sites with cellular containment taken into account

The result indicates that containment factor is only partially responsible for inability of the system to predict some of the localizations ("total prediction" percentage passing from 50% to 54%).

5.1.2. Performance Comparison with Other Existing Multi-site Localizers

The performance measure reported by Chou and Cai [Chou and Cai, 2005] consists of the standard per site sensitivity measure which obtains 70%, 84% and 90% (respectively) when the highest, the two highest, and the three highest ranking predictions in terms of similarity with experimental results are taken into consideration. For the purpose of comparing our system's performance with that of Chou and Cai, we also proceeded with per-site evaluation of sensitivity of our system. For each protein, we counted the number of sites that were correctly predicted, summed this number over all the proteins and divided the result by the total number of sites reported. We thus obtained 55% and 56% for sensitivity in localization to 17 and 9 sites respectively. Chou and Cai did not, however, report the number of FP and how this number increases as they consider a larger number of top ranking candidates for the purpose of localization based on similarity. As a performance measure that reflects FP we also calculated the per-site specificity of our system and obtained the result of 91% and 92% for localization to 17 and 9 sites specificity.

It is important to note that Chou and Cai's system deals solely with budding yeast and does not perform multi-species prediction as is attempted in this work.

5.1.3. System Coverage

Three types of coverage need to be considered in evaluating a sub-cellular localization problem:

(i) Location coverage: sub-regions that are supported by a predictor. In this work, particular attention was made to select sub-cellular compartments that were representative of all cellular

components (please see section 3.1). Gene ontology cellular component database, that is used as the source, is currently the best annotated structured controlled vocabulary for genes and gene products with respect to their cellular location; (ii) Sequence coverage: ratio of sequences for which a prediction is made to the total number of sequences of interest. As reported in Table-9, the total number of query proteins for which the system could not predict any localization was 76. Out of a total of 4529 query proteins, this amounts to (76/4529 =) 1.7% of the total set considered, hence a sequence coverage of 98.3%; (iii) Taxonomic coverage: measures the range of organisms that the predictor covers. In our work, we have considered 3 main species in fungi kingdom for which we have found extensive experimental localization information.

5.2. Hierarchy and Multiple Classification Issues

One major difficulty we encountered in subdividing the network of cellular components into disjoint segments was the "many-to-many" relationships that exist between some sub-cellular compartments. This is due to nature of this network that is a directed graph and not a tree. For example, based on the sub-cellular hierarchy graph of polarisome (GO quick browser at EBI website: http://www.ebi.ac.uk/ego/) a protein reported to localize to polarisome may be classified as belonging to either cell cortex or "site of polarized growth". Both of these major locations have polarisome as direct descendant. Another difficulty is that some prominent sub-cellular compartments are fully contained within other compartments. Nucleolus found within nucleus and cell cortex contained within cytoplasm are two such examples. In many cases there is no apparent solution to this hierarchy issue. For example, nuclear proteins that are not localized to the nucleolus or nuclear envelope are not reported as such. Instead, they are rather simply identified as belonging to nucleus. All these reported localizations are nonetheless significant and need to be equally considered as representative examples. But then, how can a predictor distinguish between such classes? Not only a predictor that is seeking differences in protein

features that would explain distinct targeting could not utilize such examples, the information contained in such examples hampers classification of proteins into a single category. These considerations necessitate multiple localizations of the same protein to different locations, the problem that we address in this work.

5.3. Error Analysis

There are three main categories of errors that impact the accuracy of the system: (i) General Computing Errors, (ii) Errors caused by our choice of data source, and (iii) Error due to our method.

5.3.1. General Computing Errors

There are general types of error that apply to any computing system including ours. Examples of such errors are the ones introduced during data entry, in manipulation of files and databases as well as calculation and programming errors. Extensive usage of scripting language, in particular, exposes our system to a higher risk of data and file handling error.

5.3.2. Errors due to Data Source

Lack of sufficient protein examples localized to one or more sub-cellular sites is the main source of error that we may identify. As evidenced by the results in Tables 8 and 10, localization to some sites may not be predicted due to insufficient examples.

Reported localizations downloaded from the GO site may contain errors. Even though we selectively chose the ones with evidence codes IDA, IGI, IMP, and IPI, the data may still contain a small margin of error due to imprecision of experimental measurements as well as errors in

reporting and compilation of annotations. Similar remarks apply to 69 megabytes of fungal protein sequence data downloaded from RefSeq Site.

We used all the localizations presenting the specified experimental evidence codes that we obtained from the GO site for the 3 selected species assuming that such a set of data is representative of fungal proteins for these 3 species and across the fungal kingdom in general.

Errors in the data source have a major impact on accuracy of the training examples and hence on the predictions made based on the learning.

5.3.3. Errors Specific to our Method

The basic premise in ID3 is that the simplest decision tree that covers all the training examples is the one that is most likely to correctly classify the unknown objects. In order to construct the simplest tree, at each internal node, ID3 selects one feature, among all existing features, such that if the tree is split on that feature, the expected cost of completion of the tree is minimized. The calculation of this expected cost is entirely dependent on the classifications and feature values of examples found in the training set. Even if the data set that we have used is fully representative of fungal proteins, there is no guarantee that the most efficiently obtainable solution (one on the shortest path) is the correct solution.

CHAPTER 6. CONCLUSION

This chapter summarizes the work and its contributions. It also provides some indication of potential future enhancement to the system.

6.1. Contribution

Proteins target those sub-cellular locations where they can perform the functions they are intended for. They are guided to their respective destinations through interaction with other proteins (Protein-Protein interaction, PPI). Specificity of such interactions is presumed to be due to physiochemical and biological characteristics inherent to the very protein considered. Molecular functions of a protein as well as the biological processes in which it takes part are also expected to influence the choice of its target destination. Specificity of such aspects need to be characterized by appropriate protein family signatures as well as targeting sequence signals that allow recognition of the proteins by specific receptors of a given process.

In this work, we have started with a set of features that can potentially impact PPI, protein's functions and biological processes. A uniform, non-causal method has then been deployed to determine the features that truly correlate with any of the sub-cellular localizations. Our developed algorithm is non-causal in the sense that no pre-established rule has been introduced that would specifically favour the localization of proteins with certain motifs to particular compartments. We have contributed towards elucidation of protein characteristics that correlate with localization to sub-cellular sites by identifying and reporting 76 features that are used in our decision tree.

In spite of our simple approach, using a classical machine learning method, and our restriction of the feature space to utilize solely the knowledge elicitable from the protein's primary sequence,

the developed system succeeded in correctly predicting 64% of the multiply-localized proteins. This result is indicative of our distinctive strategy that uses the decision tree and an ab-initio approach to handle the problem of sub-cellular localization.

Lastly, in contrast to other multi-site localizers built and evaluated based on a single organism (yeast), in this work we have attempted to take advantage of the variability of the sorting mechanisms that may be found in different species. This work is a first attempt in multi-site, multi-species localization of fungal proteins.

6.2. Possible Future Enhancements

The system developed in the present work could be enhanced in different ways:

- Enhance the degree to which the data set represents fungal protein localizations by incorporating the information on localized fungal proteins from species other than the three used here
- Augment the set of targeting motifs by reviewing a wider range of works related to fungal protein localization in the biological literature
- Incorporate secondary, super-secondary and tertiary structures among the features evaluated for their implication in localization. Knowledge obtained from the evolving field of *protein structure determination* can prove useful in identification of signal patches (Please see section 2.3.1)
- Search for additional characteristic features to be correlated to localization by generating multiple sequence alignments (MSA) of proteins homologous to sequences of known localization. This is especially useful in the case of localization sites for which no targeting motifs have been discovered.

BIBLIOGRAPHY

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (2002) Molecular Biology of The Cell, 4th edition, Garland Science Publishing, USA.

[Bellemare et al., 2002] Bellemare, D. P., Shaner, L., Morano, K. A., Beaudoin, J., Langlois, R., Labbé, S. (2002) Ctr6, a Vacuolar Membrane Copper Transporter in Schizosaccharomyces pombe. J. Biol. Chem., Vol. 277, Issue 48, 46676-46686

[Bonnai et al., 2002] Bonnai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S. (2002) Extensive feature detection of N-terminal protein sorting signals, Bioinformatics, 18:298-305.

[Bos et al., 1993] Bos, K., Wraight, C., Stanley, K. K. (1993) TGN38 is maintained in the trans-Golgi network by a tyrosine-containing motif in the cytoplasmic domain. EMBOJ. 12: 2219–2228

[Campion et al., 2001] Campion, S. R., Ameen, A. S., Lai, L., King, J. M., Munzenmaier, T. N. (2001) Dipeptide frequency/bias analysis identifies conserved sites of nonrandomness shared by cysteine-rich motifs. Proteins 2001;44:321-328. 2001 Wiley-Liss, Inc.

[Cedano et al., 1997] Cedano, J., Aloy P., Perez-Pons, J. A ., Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. JMB, 266(3): 594-600.

[Cherry et al., 1997] Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature* 387(6632 Suppl):67-73

[Chou and Cai, 2005] Chou, K. C., Cai, Y. D. (2005) Predicting protein localization in budding Yeast. Bioinformatics. Vol. 21, No. 7, 944 – 950

[Christophe et al., 2000] Christophe, D., Christophe-Hobertus, C., Pichon, B. (2000) Nuclear targeting of proteins: how many different signals? Cellular Signaling 12:337-341

[Davis and Blobel, 1986] Davis, L. I., Blobel, G. (1986) Identificaion and characterization of a nuclear pore complex protein. Cell 45:699

[Dono et al., 1998] Dono, R., James, D., Zeller, R. (1998) A GR-motif functions in nuclear accumulation of the large FGF-2 isoforms and interferes with mitogenic signaling, Oncogene, Vol. 16, 2151-2158

[Drin et al., 2003] Drin, G., Cottin, S., Blanc, E., Rees, A. R., Temsamani, J. (2003) Studies on the Internalization Mechanism of Cationic Cell-penetrating Peptides. (2003) J. Biol. Chem., Vol. 278, Issue 33, 31192-31201

[Dyer et al., 1996] Dyer, J. M., McNew, J. A., Goodman, J. M., (1996) The sorting sequence of the peroxisomal integral membrane protein PMP47 is contained within a short hydrophilic loop. J. Cell Biol. 133: 269

[Emanuelsson et al., 2000] Emanuelsson, O., Henrik Nielsen, H., Brunak, S., von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300: 1005-1016.

[Emanuelsson et al., 2001] Emanuelsson, O., von Heijne, G., Schneider, G. (2001) Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biology, Vol. 65, 175-187

[EMBOSS, 2000] EMBOSS: The European Molecular Biology Open Software Suite (2000). Rice,P. Longden,I. and Bleasby,A. (2000) Trends in Genetics 16, (6) pp276--277

[Gerber et al., 1992] Gerber, L. D. Kodukula, K., Udenfriend, S. (1992) Phosphatidylinositol glycan (PI-G) anchored membrane proteins. Amino acid requirements adjacent to the site of cleavage and PI-G attachment in the COOH-terminal signal peptide. *J. Biol. Chem.* 267: 12168–12173

[Gould et al., 1990] Gould, S. J., Keller, G. A., Schneider, M., Howell, S. H., Garrard, L. J., Goodman, J. M., Distel, D., Tabak, H., Subramani, S. (1990) Peroximal protein import is conserved between yeast, plants, insects and mammals. EMBO J. 9: 85

[Horton et al., 2006] Horton, P., Park, K. J., Obayashi, T., Nakai, K. (2006) Protein Subcellular Localization Prediction with WoLF PSORT, Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06, Taipei, Taiwan. pp. 39-48.

[Huh et al., 2003] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast, Nature, 425, 686-691.

[Hulo et al., 2004] Hulo, N., Sigrist, C. J. A., Le Saux, V., Langendijk, P., S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., Bairoch, A. (2004) Recent improvements to the PROSITE database. Nucleic Acids Res., 32, D134 – D137

[Konopka et al., 1988] Konopka, J. B., Jenness, D. D., Hartwell, L. H. (1988) The C-terminus of the S. Cerevisiae alpha-pheromone receptor mediates and adaptive response to pheromone. Cell, 54: 609

[Krijnse-Locker et al., 1994] Krijnse-Locker, J., Ericsson, M., Rottier, PJ., Griffiths, G. (1994) Characterization of the budding compartment of mouse hepatitis virus: evidence that transport from RER to the Golgi complex requires only one vesicular transport step. J. Cell Biol. 124:55-70

[Lao et al., 2002] Lao, D. M., Okuno, T., Shimizu, T. (2002) Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction. In Silico Biology 2, 485-494.

[Lesk, 2001] Lesk, A. M. (2001) Introduction to Protein Architecture. The structural biology of proteins, Oxford University Press

[Lesk, 2004] Lesk, A. M. (2004) Introduction to Protein Science. Architecture, Function, and Genomics, Oxford University Press

[Lu et al., 2004] Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D. S., Poulin, B., Anvik, J., Macdonell, C., Eisner, R (2004) Predicting Sub-cellular localization of proteins using machine-learned classifers. Bioinformatics, Vol. 20, no. 4, pages 547-556

[Luger, 2002] Luger, G. F. (2002) Artificial Intelligence: Structures and Strategies, 5th edition, Pearson Education, Essex, England

[Mitchell, 1997] Mitchell, T. M., (1997) Machine Learning, McGraw-Hill Higher Education, USA

[Nakai and Horton, 1999] Nakai K, Horton P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci. 1999 Jan;24(1):34-6.

[Nakai and Kanehisa, 1992] Nakai, K., Kanehisa, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14: 897-911.

[Nilsson et al., 1989] Nilsson, T., Jackson, M., Peterson, P. A., (1989) Short cytoplasmic sequences serve as retention signals for tranmembrane proteins in the ER. Cell 58: 707-718

[Pante et al., 1994] Pante N., Bastos R., McMorrow I., Burke B., Aebi U. (1994) Interactions and three-dimensional localization of a group of nuclear pore complex proteins. J Cell Biol. 1994 Aug;126(3):603-17.

[Park and Kanehisa, 2003] Park, K. J., Kanehisa, M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19: 1656-1663.

[Park et al., 1987] Park, M. K., D'Onofrio, M., Willingham, M. C., Hanover, J. A. (1987) A monoclonal antibody against a family of nuclear pore proteins : O-kubjed B-acettkgkycisanube us oart if the immunodeterminant. Proc. Natl. Acad. Sci. 84: 6462

[Pelham et al., 1988] Pelham, H. R. B., Hardwick, K. G., Lewis, M. J. (1988) Sorting of soluble ER proteins in yeast. EMBOJ. 7: 1757

[Persson and Argos, 1994] Persson, B., Argos, P. (1994) Prediction of transmembrane segments in proteins utilizing multiple sequence alignments J. Mol. Biol. 237, 182-192

[Ponting, 2001] Ponting C. P. Issues in Predicting Protein Functino from Sequence. Briefings in Bioinformatics. Vol 2, No. 1.

[Pringle et al., 1997] Pringle, J. R., Broach, J. R., Jones, E. W. (1997) The molecular and cellular biology of the yeast saccharomyces. Cold Spring Harbor Laboratory Press, USA

[Quinlan, 1983] Quinlan, J. R. (1983) Learning efficient classification procedures and their application to chess end games. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, Machine Learning - An Artificial Intelligence Approach, pages 463--482. Tioga, Palo Alto, CA, USA

[Roberts et al., 1992] Roberts, C. J., Nothwehr, S. F., Stevens, T. H., (1992) Membrane protein sorting in the yeast secretory pathway: evidence that the vacuole may be the default compartment. J. Cell Biol. 119:69

[Rohrer et al., 1993] Rohrer, J., Benedetti, H., Zanolari, B., Riezman, H. (1993) Identification of a novel sequence mediating regulated endocytosis of the G protein-coupled alpha-pheromone receptor in yeast. Mol. Biol. Cell 4: 511

[Rose and Doms, 1988] Rose, J., Doms, R. W. (1988) Regulation of protein export from the ER. Annu. Rev. Cell Biol. 4: 257

[Schutze et al., 1994] Schutze, M. P., Peterson, P. A., Jackson, M. R., (1994) An N-terminal double-arginine motif maintains type II membrane proteins in ER. EMBO J. 13: 1696-1705

[Scott et al., 2004] Scott, M. S., Thomas, D. Y., Hallett, M. T. (2004) Predicting Subcellular Localization via Protein Motif Co-Occurrence, Genome Research 14:1957-1966

[Sigrist et al., 2002] Sigrist C.J.A., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, 3, 265–274

[Stevens et al., 1982] Stevens, T., Esmon, B., Schekman, R. (1982) Early stages in yeast secretory pathway are required for transport of carboxypeptidase Y to the vacuole. Cell, 30: 439

[Swinkels et al., 1991] Swinkels B.W., Gould S.J., Bodnar A.G., Rachubinski R.A., Subramani S. (1991) A novel, cleavable peroxisomal targeting signal at the amino-terminus of the rat 3-ketoacyl-CoA thiolase. EMBO J. 1991 Nov;10(11):3255-62.

[Townsley et al., 1993] Townsley, F. M., Wilson, D. W., Pelham, H. R. B. (1993) Mutational analysis of the human KDEL receptor: distinct structural requirements for Golgi retention, ligand binding and retrograde transport. EMBOJ. 12: 2821

[Valls et al., 1987] Valls, L. A., Hunter, C. P., Rothman, J. H., Stevens, T. H., (1987) Protein sorting in yeast: The localization determinant of yeast vacuolar carboxypeptidase Y resides in the propeptide. Cell 48: 887

[von Heijne, 1986] von Heijne, G. (1986) "A new method for predicting signal sequence cleavage sites" Nucleic Acids Res.: 14:4683

[von Heijne, 1987] von Heijne, G. (1987) "Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit" (Acad. Press, (1987), 113-117)

[Wiemken and Durr, 1974] Wiemken, A., Durr, M. (1974) Characterization of amino acid pools in the vacuolar component of Saccharomyces Cerevisiae. Arch. Microbiol. 101: 45

[Wilcox et al., 1992] Wilcox, C. A., Redding, K., Wright, R., Fuller, R. S., (1992) Mutation of a tyrosine localization signal in the cytosolic tail of Yeast Kex2 protease disrupts Golgi retntion and results in default transpor to the vacuole. Mol. Biol. Cell 3: 1353

[Wrzeszczynski and Rost, 2004] Wrzeszczynski, K. O., Rost, B., (2004) Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes. Cell Mol Life Sci., 61(11):1341-53

APPENDICES

APPENDIX–1: LOCALIZATION SITES SELECTION

The following are additional notes on grouping of sub-cellular localization sites:

- One of the localization sites is cytoplasmic membrane-bound vesicle that includes endocytic and other protein carrying vesicles (ER to Golgi, Golgi to ER, Golgi to vacuole, Golgi to plasma membrane). There is no distinct set of vesicular transport that is deployed to reach the ER-Golgi intermediate compartment (ERGIC) [Krijnse-Locker et al., 1994], we do not consider ER to Golgi as a separate compartment and group it under cytoplasmic membrane-bound vesicle. As for Golgi to ER, Golgi to vacuole or Golgi to plasma membrane, lack of sufficient experimentally discovered factors that distinguish them from other transport vesicles is an incentive to also group them under cytoplasmic membrane-bound vesicle.
- Given the relatively small number of experimentally localized proteins available in various components of the Golgi apparatus, we designate a single localization site, Golgi apparatus, to include all Golgi cisterna (cis, medial and trans), all Golgi faces (cis and trans) as well as Golgi membrane and Golgi transport complex.
- We do not distinguish between endosomal lumen and endosomal membrane. Both are grouped under endosome. The same applies to peroxisome that includes peroxisomal membrane and peroxisomal matrix.
- Nucleolus has been designated as a distinct localization site due to its particular biological importance. Obviously, proteins targeted to this site are also localized at the nucleus.

APPENDIX–2: PYSIO-CHEMICAL FEATURES

Table-A2 depicts the compositional features that are selected to represent protein's physiochemical properties. Amino acids specific to each category of property (ex: basic) are listed using one-letter-abbreviations [Alberts et al., 2002]. The range of numerical values that a feature may take (ex: 0 to 28.7% for Alanine composition) is calculated from the reported values of the corresponding feature in the full set of examples sequences. Furthermore, the average molecular weight of a protein, its average scaled hydrophobicity, and its average isoelectiric point are calculated from data at http://www.ionsource.com/virtit/VirtualIT/aainfo.htm.

Feature name	Feature description	Discretization Range
A-Content	Composition Percent of Alanine	Low: [0.0-9.5], medium: [9.6-19.1], high: [19.2-28.7]
C-Content	Composition Percent of Cysteine	Low: [0.0-6.5], medium: [6.6-13.1], high: [13.2-19.7]
D-Content	Composition Percent of Aspartic Acid	Low: [0.0-6.5], medium: [6.6-13.1], high: [13.2-19.7]
E-Content	Composition Percent of Glutamic Acid	Low: [0.0-8.1], medium: [[8.2-16.3], high: [[16.4-24.5]
F-Content	Composition Percent of Phenylalanine	Low: [0.0-5.5], medium: [5.6-11.1], high: [11.2-16.7]
G-Content	Composition Percent of Glycine	Low: [0.0-9.3], medium: [9.4-18.8], high: [18.9-28.1]
H-Content	Composition Percent of Histidine	Low: [0.0-3.4], medium: [3.5-6.9], high: [7.0-10.2]
I-Content	Composition Percent of Isoleucine	Low: [0.0-5.8], medium: [5.9-11.6], high: [11.7-17.5]
K-Content	Composition Percent of Lysine	Low: [0.0-9.0], medium: [9.1-18.1], high: [18.2-27.1]
L-Content	Composition Percent of Leucine	Low: [0.0-7.4], medium: [7.5-14.8], high: [14.9-22.2]
M-Content	Composition Percent of Methionine	Low: [0.0-4.6], medium: [4.7-9.2], high: [9.3-13.9]
N-Content	Composition Percent of Asparagine	Low: [0.0-7.8], medium: [7.9-15.6], high: [15.7-23.5]
P-Content	Composition Percent of Proline	Low: [0.0-8.5], medium: [8.6-17.0], high: [17.1-25.6]
Q-Content	Composition Percent of Glutamine	Low: [0.0-8.9], medium: [9.0-17.9], high: [18.0-26.9]
R-Content	Composition Percent of Arginine	Low: [0.0-5.7], medium: [5.8-11.5], high: [11.6-17.3]
S-Content	Composition Percent of Serine	Low: [0.0-16.1], medium: [16.2-32.2], high: [32.3-48.3]
T-Content	Composition Percent of Threonine	Low: [0.0-9.7], medium: [9.8-19.5], high: [19.6-29.3]

Feature name	Feature description	Discretization Range
V-Content	Composition Percent of Valine	Low: [0.0-5.2], medium: [5.3-10.4], high: [10.5-15.6]
W-Content	Composition Percent of Tryptophan	Low: [0.0-2.4], medium: [2.5-4.9], high: [5.0-7.4]
Y-Content	Composition Percent of Tyrosine	Low: [0.0-37.3], medium: [37.4-74.9], high: [75.0-112.1]
PosChrg-Content	Composition Percent of Basic residues (K, R, H)	Low: [1.1-13.7], medium: [13.8-26.4], high: [26.5-39.0]
NegChrg-Content	Composition Percent of Acidic residues (D, E)	Low: [0.0-13.1], medium: [13.2-26.3], high [26.4-39.5]
PolUnchrg-Content	Composition Percent of Polar and uncharged residues [NQSTY]	Low: [11.1-27.8], medium: [27.9-44.7], high: [44.8-61.6]
NonPol-Content	Composition Percent of Non-polar residues [ALPMGVIFWC]	Low: [15.3-35.2], medium: [35.3-55.2], high: [55.3-75.4]
Alip-Content	Composition Percent of Aliphatic residues [ILV]	Low: [1.9-15.3], medium: [15.4-28.8], high: [28.9-42.4]
Arom-Content	Composition Percent of Aromatic residues [FWYH]	Low: [0.0-8.3], medium: [8.4-16.7], high: [16.8-25.0]
TinyResidue-Content	Composition Percent of Tiny residues (MW < 90 D) [AG]	Low: [0.0-12.3], medium: [12.4-24.6], high: [24.7-37.0]
SmlResidue-Content	Composition Percent of Small residues (MW <115) [AGST]	Low: [11.4-30.2], medium: [30.3-49.1], high: [49.2-68.1]
MdmResidue-Content	Composition Percent of Med-sized residues [LVIPMCNQDEK]	Low: [26.0-42.9], medium: [43.0-59.9], high: [60.0-77.1]
BklResidue-Content	Composition Percent of Bulky residues (150 <= MW) [FHRWY]	Low: [1.3-12.0], medium: [12.1-22.8], high: [22.9-33.7]
HdrphbRes-Content	Composition Percent of Hydrophobic residues (1<hydrophobicity) [ILMVFW]	Low: [3.7-21.8], medium: [21.9-40.0], high: [40.1-58.3]
WkHdrphbRes-Content	Composition Percent of Weakly hydrophobic residues (-1.5 < hydrophobicity <= 1) [ACGPYST]	Low: [16.5-35.4], medium: [35.5-54.4], high: [54.5-73.4]
HdrphlRes-Content	Composition Percent of Hydrophilic residues (hydrophobicity <= -1.5) [DEHKRNQ]	Low: [10.6-28.0], medium: [28.1-45.5], high: [45.6-63.2]
SS-Content	Composition Percent of SS dipeptide	Low: [0.0-10.6], medium: [10.7-21.3], high: [21.4-32.0]
LL-Content	Composition Percent of LL dipeptide	Low: [0.0-2.7], medium: [2.8-5.5], high: [5.6-8.3]
LS-Content	Composition Percent of LS dipeptide	Low: [0.0-1.3], medium: [1.4-2.6], high: [2.7-3.9]
AA-Content	Composition Percent of AA dipeptide	Low: [0.0-4.0], medium: [4.1-8.0], high: [8.1-12.0]
SL-Content	Composition Percent of SL dipeptides	Low: [0.0-1.3], medium: [1.4-2.6], high: [2.7-3.9]
LEN	residue sequence length	Low: [36-1660], medium: [1661-3284], high: [3285-4910]
MW	Computed average molecular weight	Low: [94.2-103.9], medium: [104.0-113.7], high: [113.8-123.6]
Hdrphb	Computed average hydrophobicity	Low: [-2.4—1.4], medium: [-1.3-0—0.3], high: [-0.2-0.8]
Iep	Computed average isoelectiric point	Low: [5.16 – 5.83], medium: [5.84 – 6.50], and high: [6.51 – 7.18]

Table-A2: Proteins Physiochemical Features

APPENDIX–3: FUNCTIONAL MOTIFS

Table-A3 depicts the fungal proteins functional motif features obtained from PROSITE database:

Feature name	Feature description
14-trois-trois proteins	[RA]-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA]
14-trois-trois proteins	Y-K-[DE]-[SG]-T-L-I-[IML]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-[TANS]-[SAD]
ATP phosphoribosyltransferase	E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]
Actin and Actin-like	[FY]-[LIV]-[GV]-[DE]-E-[ARV]-[QLAH]-x(1,2)-[RKQ](2)-[GD]
Actin and Actin-like	W-[IVC]-[STAK]-[RK]-x-[DE]-Y-[DNE]-[DE]
Actin and Actin-like	[LM]-[LIVMA]-T-E-[GAPQ]-x-[LIVMFYWHQPK]-[NS]-[PSTAQ]-x(2)-N-[KR]
Adenylate kinase	[LIVMFYWCA]-[LIVMFYW](2)-D-G-[FYI]-P-R-x(3)-[NQ]
Amino acid permeases	[STAGC]-G-[PAG]-x(2,3)-[LIVMFYWA](2)-x-[LIVMFYW]-x-[LIVMFWSTAGC](2)-[STAGC]-x(3)-[LIVMFYWT]-x-[LIVMST]-x(3)-[LIVMCTA]-[GA]-E-x(5)-[PSAL]
Aminotransferases class-V pyridoxal-phosphate attachment site	[LIVFYCHT]-[DGH]-[LIVMFYAC]-[LIVMFYA]-x(2)-[GSTAC]-[GSTA]-[HQR]-K-x(4,6)-G-x-[GSAT]-x-[LIVMFYSAC]
ArgE / dapE / ACY1 / CPG2 / yscS family	[LIV]-[GALMY]-[LIVMF]-{Q}-[GSA]-H-x-D-[TV]-[STAV]
ArgE / dapE / ACY1 / CPG2 / yscS family	[GSTAI]-[SANQCVIT]-D-x-K-[GSACN]-x(1,2)-[LIVMA]-x(2)-[LIVMFY]-x(12,17)-[LIVM]-x-[LIVMF]-[LIVMSTAGC]-[LIVMFA]-x(2)-[DNGM]-E-E-x(0,1)-[GSTNE]
Aspartate-semialdehyde dehydrogenase	[LIVM]-[SADN]-x(2)-C-x-R-[LIVM]-x(4)-[GSC]-H-[STA]
Casein kinase II regulatory subunit	C-P-x-[LIVMYAT]-x-C-x(5)-[LI]-P-[LIVMCA]-G-x(9)-V-[KRM]-x(2)-C-[PA]-x-C
Carboxylesterases type-B	F-[GR]-G-x(4)-[LIVM]-x-[LIV]-x-G-x-S-[STAG]-G
Carboxylesterases type-B	[ED]-D-C-L-[YT]-[LIV]-[DNS]-[LIV]-[LIVFYW]-x-[PQR]
Chitinases family 18 active site	[LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E
Chorismate synthase	G-[DES]-S-H-[GC]-x(2)-[LIVM]-[GTIV]-x-[LIVT]-[LIV]-[DEST]-[GH]-x-[PV]
Chorismate synthase	[GE]-x(2)-S-[AG]-R-x-[ST]-x(3)-[VT]-x(2)-[GA]-[STAVY]-[LIVMF]
Chorismate synthase	R-[SHF]-D-[PSV]-[CSAVT]-x(4)-[SGAIVM]-x-[IVGSTAPM]-[LIVM]-x-E-[STAHNCG]-[LIVMA]
Copper amine oxidase	[LIVM]-[LIVMA]-[LIVMF]-x(4)-[ST]-x(2)-N-Y-[DE]-[YN]
Copper amine oxidase	T-x-[GS]-x(2)-H-[LIVMF]-x(3)-E-[DE]-xP
Cutinase active sites	P-x-[STA]-x-[LIV]-[IVT]-x-[GS]-G-Y-S-[QL]-G
Cutinase active sites	C-x(3)-D-x-[IV]-C-x-G-[GST]-x(2)-[LIVM]-x(2,3)-H
Cytochrome b5 family, heme-binding domain	[FY]-[LIVMK]-{I}-{Q}-H-P-[GA]-G
Cytochrome c oxidase subunitVIa	[LIV]-R-x-K-x-[FYW]-x-W-[GS]-D-G-x-[KH]-[ST]-x-F-xN
D-amino acid oxidases	[LIVM](2)-H-[NHA]-Y-G-x-[GSA](2)-x-G-x(5)-G-x-A
D-isomer specific 2-hydroxyacid dehydrogenases	[LIVMA]-[AG]-[IVT]-[LIVMFY]-[AG]-x-G-[NHKRQGSAC]-[LIV]-G-x(13,14)-[LIVMFT]-{A}-x-[FYWCTH]-[DNSTK]
D-isomer specific 2-hydroxyacid dehydrogenases	[LIVMFYWA]-[LIVFYWC]-x(2)-[SAC]-[DNQHR]-[IVFA]-[LIVF]-x-[LIVF]-[HNI]-x-P-x(4)-[STN]-x(2)-[LIVMF]-x-[GSDN]
D-isomer specific 2-hydroxyacid dehydrogenases	[LMFATCYV]-[KPQNHAR]-x-[GSTDNK]-x-[LIVMFYWR]-[LIVMFYW](2)-N-x-[STAGC]-R-[GP]-x-[LIVH]-[LIVMCT]-[DNVE]
DNA photolyases class 1	T-G-x-P-[LIVM](2)-D-A-x-M-[RA]-x-[LIVM]

Feature name	Feature description
DNA photolyases class 1	[DN]-R-x-R-[LIVM]-[LIVMN]-x-[STA]-[STAQ]-F-[LIVMFA]-x-K-x-L-x(2,3)-W-[KRQ]
DNA polymerase family B	[YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-{LA}-[LIVMSTAC]
DNA/RNA non-specific endonucleases active site	D-R-G-H-[QLIM]-x(3)-[AG]
Dehydroquinase class I active site	D-[LIVM]-[DE]-[LIVMN]-x(18,20)-[LIVM](2)-x-[SC]-[NHY]-H-[DN]
Dehydroquinase class II	[LIVM]-[NQHS]-G-P-N-[LVI]-x(2)-[LT]-G-x-R-[QED]-x(3)-[FY]G
EPSP synthase	[LIVF]-{LV}-x-[GANQK]-[NLG]-[SA]-[GA]-[TAI]-[STAGV]-x-R-x-[LIVMFYAT]-x-[GSTAP]
EPSP synthase	[KR]-x-[KH]-E-[CSTVI]-[DNE]-R-[LIVMY]-x-[GSTAVLD]-[LIVMCTF]-x(3)-[LIVMFA]-x(2)-[LIVMFCGANY]G
ER lumen protein retaining receptor	G-[LIV]-S-x-[KR]-x-[QH]-x-L-[FY]-x-[LIV](2)-[FYW]-x(2)-RY
ER lumen protein retaining receptor	L-E-[SA]-V-A-I-[LM]-P-Q-[LI]
Endoplasmic reticulum targeting sequence	[KRHQSA]-[DENQ]-E-L
Eukaryotic and viral aspartyl proteases	[LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-{GQ}-[LIVMFSTNC]-{EGK}-[LIVMFGTA]
Fatty acid desaturases	G-E-x-[FYN]-H-N-[FY]-H-H-x-F-P-x-DY
Fructose-bisphosphate aldolase class-II	[FYVMT]-x(1,3)-[LIVMH]-[APNT]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-[GACH]
Fructose-bisphosphate aldolase class-II	[LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-xE
GMC oxidoreductases	[GA]-[RKNC]-x-[LIVW]-G(2)-[GST](2)-x-[LIVM]-N-x(3)-[FYWA]-x(2)-[PAG]-x(5)-[DNESHQ]
GMC oxidoreductases	[GS]-[PSTA]-x(2)-[ST]-[PS]-x-[LIVM](2)-x(2)-S-G-[LIVM]G
Glucosamine/galactosamine-6-phosphate isomerases	[LIVM]-x(3)-[GNH]-x(0,1)-[LITCRV]-x-[LIVWF]-x-[LIVMF]-x-[GS]-[LIVM]-G-x-[DENV]-G-[HN]
Glutamine amidotransferases class-I active site	[PAS]-[LIVMFYT]-[LIVMFY]-G-[LIVMFY]-C-[LIVMFYN]-G-x-[QEH]-x-[LIVMFA]
Glycosyl hydrolases family 10 active site	[GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-[LIVMF]
Glycosyl hydrolases family 11 active sites	[PSA]-[LQ]-x-E-Y-Y-[LIVM](2)-[DE]-x-[FYWHN]
Glycosyl hydrolases family 11 active sites	[LIVMF]-x(2)-E-[AG]-[YWG]-[QRFGS]-[SG]-[STAN]-G-x-[SAF]
Glycosyl hydrolases family 2	N-x-[LIVMFYWD]-R-[STACN](2)-H-Y-P-x(4)-[LIVMFYWS](2)-x(3)-[DN]-x(2)-G-[LIVMFYW](4)
Glycosyl hydrolases family 2	[DENQLF]-[KRVW]-N-[HRY]-[STAPV]-[SAC]-[LIVMFS](3)-W-[GS]-x(2,3)-NE
Glycosyl hydrolases family 3 active site	[LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI]
Glycosyl hydrolases family 32 active site	H-x(2)-P-x(4)-[LIVM]-N-D-P-NG
Glycosyl hydrolases family 35 putative active site	G-G-P-[LIVM](2)-x(2)-Q-x-E-N-E-[FY]
Glycosyl hydrolases family 45 active site	[STA]-T-R-Y-[FYW]-D-x(5)-[CA]
Glycosyl hydrolases family 5	[LIV]-[LIVMFYWGA](2)-[DNEQG]-[LIVMGST]-{SENR}-N-E-[PV]-[RHDNSTLIVFY]

Feature name	Feature description
Glycosyl hydrolases family 6	V-x-Y-x(2)-P-x-R-D-C-[GSAF]-x(2)-[GSA](2)-xG
Glycosyl hydrolases family 6	[LIVMYA]-[LIVA]-[LIVT]-[LIV]-E-P-D-[SAL]-[LI]-[PSAG]
Glycosyl hydrolases family 9 active sites	[STV]-x-[LIVMFY]-[STV]-x(2)-G-x-[NKR]-x(4)-[PLIVM]-H-xR
Glycosyl hydrolases family 9 active sites	[FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA]
Glyoxalase I	[HQ]-[IVT]-x-[LIVFY]-x-[IV]-x(4)-{E}-[STA]-x(2)-F-[YM]-x(2,3)-[LMF]-G-[LMF]
Glyoxalase I	G-[NTKQ]-x(0,5)-[GA]-[LVFY]-[GH]-H-[IVF]-[CGA]-x-[STAGLE]-x(2)-[DNC]
HIT domain	[NQAR]-x(4)-[GSAV]-x-[QFLPA]-x-[LIVM]-x-[HWYRQ]-[LIVMFYST]-H-[LIVMFT]-H-[LIVMF]-[LIVMFP]-[PSGAW]
Heat shock hsp90 proteins family	Y-x-[NQH]-K-[DE]-[IVA]-F-[LM]-R-[ED]
Heme peroxidase	[DET]-{LIVMTA]-x(2)-[LIVM]-[LIVMSTAG]-[SAG]-[LIVMSTAG]-H-[STA]-[LIVMFY]
Heme peroxidase	[SGATV]-{D}-x(2)-[LIVMA]-R-[LIVMA]-x-[FW]-H-{V}-[SAC]
Histidinol dehydrogenase	[IVT]-[DE]-x(2)-[AYEP]-G-[PT]-[ST]-E-[LIVST]-[LIVMAECGF]-[LIVMA]-[LIVMAYF]-[ACNDSTI]-x(3)-[ACNGVST]-x(4)-[LIVMA]-[AVLKI]-[SACLYWNRMT]-[DE]-[LIVMFC]-[LIVMKF]-[SAG]-x(2)-EH
Hydroxymethylglutaryl-coenzyme A reductase	[RKH]-x(2)-{I}-x-{I}-x-D-x-M-G-x-N-x-[LIVMA]
Hydroxymethylglutaryl-coenzyme A reductase	[LIVM]-G-x-[LIVM]-G-G-[AG]T
Hydroxymethylglutaryl-coenzyme A reductase	A-[LIVM]-x-[STAN]-x(2)-[LI]-x-[KRNQ]-[GSA]-H-[LM]-x-[FYLH]
Hydroxymethylglutaryl-coenzyme A synthase active site	N-x-[DN]-[IV]-E-G-[IV]-D-x(2)-N-A-C-[FY]-xG
Hypothetical hesB/yadR/yfhF family	F-x-[LIVMFY]-x-[NH]-[PGT]-[NSKQR]-x(4)-C-x-C-[GS]-x-SF
Imidazoleglycerol-phosphate dehydratase	[LIVMY]-[DE]-x-H-H-x(2)-E-x(2)-[GCA]-[LIVM]-[STAVCL]-[LIVMF]
Imidazoleglycerol-phosphate dehydratase	[GW]-x-[DNIE]-x-H-H-x(2)-E-[STAGC]-x-[VMFYH]K
Indole-3-glycerol phosphate synthase	[LIVMFY]-[LIVMC]-x-E-[LIVMFYC]-K-[KRSPQV]-[STAHKRYC]-S-P-[STRK]-x(3,4)-[LIVMFYST]
Inorganic pyrophosphatase	D-[SGDN]-D-[PE]-[LIVMF]-D-[LIVMGAC]
Isocitrate and isopropylmalate dehydrogenases	[NSK]-[LIMYTV]-[FYDNH]-[GEA]-[DNGSTY]-[IMVYL]-x-[STGDN]-[DN]-x(1,2)-[SGAP]-x(3,4)-[GE]-[STG]-[LIVMPA]-[GA]-[LIVMF]
Isocitrate lyase	K-[KR]-C-G-H-[LMQR]
Lipoate-protein ligase B	R-G-G-x(2)-T-[FYWCAH]-H-x(2)-[GH]-Q-x-[LIVMT]-xY
Malate dehydrogenase active site	[LIVM]-T-[TRKMN]-L-D-x(2)-R-[STA]-x(3)-[LIVMFY]
Malate synthase	[KR]-[DENQ]-[HN]-x(2)-G-L-N-x-G-x-W-D-Y-[LIVM]F
Malic enzymes	[FM]-x-[DV]-D-x(2)-[GS]-T-[GSA]-x-[IV]-x-[LIVMAT]-[GAST](2)-[LIVMFA]-[LIVMFY]
Multicopper oxidases	G-x-[FYW]-x-[LIVMFYW]-x-[CST]-x-{PR}-x(5)-{LFH}-G-[LM]-x(3)-[LIVMFYW]
Multicopper oxidases	H-C-H-x(3)-H-x(3)-[AG]-[LM]
N-acetyl-gamma-glutamyl-phosphate reductase active site	[LIVMA]-[GSA]-x-[PA]-G-C-[FYN]-[AVP]-T-[GSAC]-x(3)-[GTAC]-[LIVMCA]-xP
Nitrilases / cyanide hydratase	G-x(2)-[LIVMFY](2)-x-[IF]-x-E-x(2)-[LIVM]-x-G-YP
Nitrilases / cyanide hydratase	G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS]-x-[KR]

Feature name	Feature description
PdxS/SNZ family	[LV]-P-[VI]-[VTPI]-[NQLHT]-[FL]-[ATVS]-[AS]-G-G-[LIV]-[AT]-T-P-[AQS]-D-[AGVS]-[AS]-[LM]
PdxT/SNO family	[GA]-L-I-[LIV]-P-G-G-E-S-T-[STA]
Phenylalanine and histidine ammonia-lyases active site	[GS]-[STG]-[LIVM]-[STG]-[SAC]-S-G-[DH]-L-x-P-L-[SA]-x(2,3)-[SAGVT]
Phosphopantetheine attachment site	[DEQGSTALMKRH]-[LIVMFYSTAC]-[GNQ]-[LIVMFYAG]-[DNEKHS]-S-[LIVMST]-{PCFY}-[STAGCPQLIVMF]-[LIVMATN]-[DENQGTAKRHLM]-[LIVMWSTA]-[LIVGSTACR]-{LPIY}-{VY}-[LIVMFA]
Polyprenyl synthetases	[LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH]
Polyprenyl synthetases	[LIVMFY]-G-x(2)-[FYL]-Q-[LIVM]-x-D-D-[LIVMFY]-x-[DNG]
Profilin	<x(0,1)-[STA]-x(0,1)-W-[DENQH]-x-[YI]-x-[DEQ]
Putative AMP-binding domain	[LIVMFY]-{E}-{VES}-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-x-[PASLIVM]-[KR]
RNA 3'-terminal phosphate cyclase	[RH]-G-x(2)-P-x-G(3)-x-[LIV]
Ribosomal protein L15	K-[LIVM](2)-[GASL]-x-[GT]-x-[LIVMA]-x(2,5)-[LIVM]-x-[LIVMF]-x(3,4)-[LIVMFCA]-[ST]-x(2)-A-x(3)-[LIVM]-x(3)G
Ribosomal protein L23	[RK](2)-[AM]-[IVFYT]-[IV]-[RKT]-L-[STANEQK]-x(7)-[LIVMFT]
Ribosomal protein L27e	G-K-[NS]-x-W-F-F-x(2)-L-[RH]-F>
Ribosomal protein L30	[IVTAS]-[LIVM]-x(2)-[LF]-x-[LI]-x-[KRHQEG]-x(2)-[STNQH]-x-[IVTR]-x(10)-[LMSN]-[LIV]-x(2)-[LIVA]-x(2)-[LMFY]-[IVT]
Ribosomal protein S12e	[AS]-[LI]-[KREQP]-x(2)-[LIVM]-x(2)-[SA]-x(3)-[DNG]-G-[LIV]-x(2)-G-[LIV]
Ribosomal protein S26e	[YH]-C-[VI]-[SA]-C-A-IH
Ribosomal protein S28e	E-[ST]-[EA]-R-E-A-[RK]-x-[LI]
Ribosomal protein S5	G-[KRQE]-x(3)-[FYVI]-x-[ACVTI]-x(2)-[LIVMA]-[LIVM]-[AG]-[DN]-x(2,3)-G-x-[LIVMA]-[GS]-x-[SAG]-x(5,6)-[DEQG]-[LIVMARF]-x(2)-A-[LIVMFR]
Ribulose-phosphate 3-epimerase family	[LIVMF]-H-[LIVMFY]-D-[LIVM]-x-D-x(1,2)-[FY]-[LIVM]-x-N-x-[STAV]
Ribulose-phosphate 3-epimerase family	[LIVMA]-x-[LIVM]-M-[ST]-[VS]-x-P-x(3)-[GN]-Q-x(0,1)-[FMK]-x(6)-[NKR]-[LIVMC]
SAICAR synthetase	[LIVMRPA]-[LIVFY]-[PLNRKG]-[LIVMF]-E-x-[IV]-[LVCATI]-R-x(3)-[TAEYSI]-G-[ST]
SAICAR synthetase	[LI]-[IVCAP]-D-x-K-[LIFY]-E-[FI]G
Shikimate kinase	[KR]-x(2)-E-x(3)-[LIVMF]-x(8,12)-[LIVMF](2)-[SA]-x-G(3)-x-[LIVMF]
Short-chain dehydrogenases/reductases family	[LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-{PC}-[SAGFYR]-[LIVMSTAGD]-x-{K}-[LIVMFYW]-x(2)-{YR}-[LIVMFYWGAPTHQ]-[GSACQRHM]
Squalene and phytoene synthases	Y-[CSAM]-x(2)-[VSG]-A-[GSA]-[LIVAT]-[IV]-G-x(2)-[LMSC]-x(2)-[LIV]
Squalene and phytoene synthases	[LIVM]-G-x(3)-Q-x(2,3)-[ND]-[IFL]-x-[RE]-D-[LIVMFY]-x(2)-[DE]-x(4,7)-R-x-[FY]-xP
Terpene synthases	[DE]-G-S-W-x-[GE]-x-W-[GA]-[LIVM]-x-[FY]-x-Y-[GA]
Tryptophan synthase alpha chain	[LIVM]-E-[LIVM]-G-x(2)-[FYCHT]-[STP]-[DEK]-[PA]-[LIVMYG]-[SGALIMY]-[DE]-[GN]
Tryptophan synthase beta chain pyridoxal-phosphate attachment site	[LIVMYAHQ]-x-[HPYNVF]-x-G-[STA]-H-K-x-N-x(2)-[LIVM]-x-[QEH]
Zinc-containing alcohol dehydrogenases	G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVSAC]

Feature name	Feature description
Zinc-containing alcohol dehydrogenases	[GSD]-[DEQHKM]-x(2)-L-x(3)-[SAG](2)-G-G-x-G-x(4)-Q-x(2)-[KRS]

Table-A3: Proteins Functional Motifs

APPENDIX–4: MODULES

A4.1. Feature Selection Modules

The following modules are developed in order to select protein features used for training and classification:

Fungal Protein Sequence files Download
13 fasta files of all existing fungal proteins from RefSeq database (Release 7) and downloaded from http://www.ncbi.nlm.nih.gov/RefSeq/ and merged into a single file (refseqfungiprot.faa).

Module : SelectDipeptides.cpp
Description: Reads a fasta file of all existing fungal proteins and calculates and output the 20 most predominantly occurring dipeptides.
Input: refseqfungiprot.faa
Output: Dipep.dat

Module : SelectMotifs
Description: Reads fasta files of all existing fungal proteins and determines which of the existing known motifs do actually occur in fungi
Input: refseqfungiprot.faa
Output: PSxxx. (119 files)

Module : SelectSignals
Description: Reads fasta files of all existing fungal proteins and determines which of the proposed signals (found in pre-conditions of rules) actually occur in fungi
Input: refseqfungiprot.faa
Output: EXISTxxx. (17 files)

A4.2. Training Examples Selection Modules

The following modules are developed in order to prepare training examples:

Training Fungal Protein Sequence files Download
Candida albicans protein sequences were downloaded from:
http://www.candidagenome.org/download/sequence/genomic_sequence/orf_protein/
(filename: orf_trans_all.fasta.gz)
Saccharomyces cerevisiae protein sequences were downloaded from: ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/
(filename: orf_trans_all.fasta.gz)
Schizosaccharomyces pombe protein sequences were downloaded from:
http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/gen_list.cgi?genome=po
The 3 files were then uniformized and merged into a single file: ProteinSequences.faa

Training Fungal Protein Localization files Download
We obtained localizations of fungal proteins for 3 different species (SC, CA and SP) from GO site: http://www.geneontology.org/GO.current.annotations.shtml from the files: gene_association.sgd.gz, gene_association.cgd.gz, and gene_association.GeneDB_Spombe.gz

These files were filtered to get only IDA, IGI, IPI, IMP evidenced localizations, further uniformized to retain only the required fields, and were finally merged into a single file (LocalizedExamples.dat) with 8604 entries.

Module : ExtractExamples
Description: Extracts non-repeating name of proteins from list of localized proteins
Input: LocalizedExamples.dat
Output: Examples.dat

Module : ExtractSequences.cpp
Description: Reads the names of example proteins (and extract their fasta sequences from the file containing protein sequences of all proteins for considered fungal species.
Input: Examples.dat, ProteinSequences.faa
Output: Examples.faa

A4.3. Feature Calculation Modules

The following three programs were downloaded from [EMBOSS, 2000] and used for protein feature extraction:

Module : tmap (Trans-membrane segments)
Function: This program predicts trans-membrane segments in proteins.
References: EMBOSS, [Persson and Argos, 1994]
Call format: tmap –auto –sequence "input-file-name" –graph none –outfile "output-file-name"
Description: It reads in one or more aligned protein sequences. Two sets of propensity values are then used for the calculations: one for the middle, hydrophobic portion and one for the terminal regions of the trans-membrane sequence spans. Average propensity values are calculated for each position along the alignment, with the contribution from each sequence weighted according to its dissimilarity relative to the other aligned sequences. Eight-residue segments are considered as potential cores of trans-membrane segments and elongated if their middle propensity values are above a threshold. End propensity values are also considered as stop signals. Only helices with a length of 15 to 29 residues are allowed and corrections for strictly conserved charged residues are made.

Module : sigcleave (Signal Peptide Cleavage)
Function: Predict protein signal cleavage sites
Reference: EMBOSS, [von Heijne, 1986], [von Heijne, 1987]
Call format: sigcleave –auto –sequence "input-file-name" –minwight 3.5 –outfile "output-file-name"
Description: sigcleave predicts the site of cleavage between a signal sequence and the mature exported protein. It uses the method of von Heijne as modified by von Heijne in his later book where treatment of positions -1 and -3 in the matrix is slightly altered. We can also use -send 50 on command line to make program check only the first 50 residues. We use the value of minweight of 3.5, at which the method correctly identifies 95% of signal peptides, and rejects 95% of non-signal peptides.

Module : fuzzpro (Pattern Search)
Function: Detects a specific protein residues pattern
Reference: [EMBOSS, 2000]
Call format: fuzzpro –auto –sequence "input-file-name" –pattern "searchable-subsequence" –mismatch 0 –outfile "output-file-name"

Description: It uses PROSITE style patterns to search protein sequences for a pattern.
Patterns are specifications of a (typically short) length of sequence to be found. They can specify a search for an exact sequence or they can allow various ambiguities, matches to variable lengths of sequence and repeated subsections of the sequence.

The following modules were developed in order to calculate feature values for proteins:

Module: SelectCompositionRange.cpp
Description: Reads fasta files of all fungal protein examples and determines the range of values for each of 42 composition features (i.e. min and max % contents). This will be used to calculate the thresholds allowing discretization of feature values
Input: Examples.faa
Output: CompositionRange.dat

Module: ExtractComposition.cpp
Description: Reads fasta file of example fungal proteins and assigns a discretized range value, based on pre-determined thresholds, to each of the example proteins for each of the 42 physiochemical features.
Input: Examples.faa, CompositionRange.dat
Output: CompositionRangeValues.dat

Module : ScanMotifs
Description: Reads fasta files of our example proteins and determines which of the functional motifs occurs in which proteins among our set of examples.
Input: Examples.faa
Output: PSxxx (119 files)

Module : ExtractMotifs
Description: Reads all PSxxx files generated by ScanMotifs script and creates a file listing proteins in which one or more of these functional motifs occur.
Input: PSxxx (119 files)
Output: Motifs.dat

Module : ScanSignals
Description: Reads fasta files of our example proteins and determines which of the targeting motifs occurs in which proteins among our set of examples.
Input: Examples.faa
Output: RLxxx (26 files)

Module : ExtractSignals
Description: Reads all RLxxx files generated by ScanSignals script and creates files listing proteins in which each targeting motif occurs.
Input: Examples.faa
Output: CLSPSignal.dat, ERRSSignal.dat, ESSignal.dat, VTSSignal.dat, GASSignal.dat, ERTMSSignal.dat, VTMSSignal.dat, NMLSSignal.dat, NLSSignal.dat, PTSSignal.dat, PMSSignal.dat, MTPSignal.dat, MMTPSignal.dat, VSSignal.dat.

Module : ScanTransmembraneSignals
Description: Splits the sequence file of example fungal proteins into single sequence files that are subsequently scanned for trans-membrane signals using tmap module from [EMBOSS, 2000].
Input: Examples.faa

Output: xxx.tms (4912 files)

Module : ExtractTransmembraneSignals
Description: Reads all xxx.tms files generated by ScanTransmembraneSignals script and creates a file listing proteins that have a transmembrane segment.
Input: xxx.tms (4912 files)
Output: TMSSignal.dat

Module : ScanMatureTransmembraneSignals
Description: Reads the file listing the example proteins with cleavable signal peptides, obtains a file of names of these proteins, and splits the latter into single sequence files that are subsequently scanned for existence of trans-membrane signals using tmap from [EMBOSS, 2000].
Input: RLCLSPOUT
Output: xxx.mtms (6648 files)

Module : ExtractMatureTransmembraneSignals
Description: Reads all xxx.mtms files generated by ScanMatureTransmembraneSignals script and creates a file listing mature proteins (with their signal peptides cleaved) that have a transmembrane segment.
Input: xxx.mtms (6648 files)
Output: MTMSSignal.dat

Generation of Mapping File
go_hierarchy.xrt from fungal genome project was accessed and all GO mappings to each of the 17 designated sites were extracted into GOMapping.dat. Each entry of the extracted file contains two pieces of information: first detailed reported localization site and second the ancestor parent of this detailed site. In cases where no mapping to any of the 17 designated sites is possible, the localization is considered as localized to GO term "Cellular Component Unknown".

Module : ExtractFeatures.cpp
Description: Reads the files containing information about the example proteins (their localization, their compositional range, the presence or absence of each of the functional motifs and targeting motifs) and produces a matrix file containing association between localizations and feature values for each of the example proteins. Each tuple of this generated matrix includes a protein name, a reported localization and values for each of the 178 selected features.
Input: CLSPSignal.dat, ERRSSignal.dat, ESSignal.dat, VTSSignal.dat, GASSignal.dat, ERTMSSignal.dat, VTMSSignal.dat, NMLSSignal.dat, NLSSignal.dat, PTSSignal.dat, PMSSignal.dat, MTPSignal.dat, MMTPSignal.dat, VSSignal.dat, TMSSingal.dat, MTMSSignal.dat, Motifs.dat, SearchableMotifs.dat, GOMapping.dat, LocalizedExamples.dat, and CompositionRangeValues.dat.
Output: Matrix.dat.

A4.4. Classification Module

The following module first builds a decision tree using training examples and then predicts localization(s) for query proteins based on their feature values using the trained tree.

Module : Localizer.cpp
Description: This module, first, builds the decision tree. It reads all the example proteins in the training file. Each protein example consists of a name, a localization, and a set of 178 feature values.

A linked list of these examples is then created and is associated with the root of the tree.
Then, sub-trees are built, recursively and starting at the root, in the following manner:

- If all examples associated with a given node have the same localization, then we have reached a terminal node (all examples and hence their localizations are already associated with this node, and no more processing need to be performed on this node).
- Otherwise, this node is a branching point and we must select a feature to split upon. The choice of a different splitting feature leads to a different degree of disorder in the forthcoming sub-trees produced by the split and the selection is made, following ID3's heuristic, in such a way as to minimize the disorder. The disorder is numerically measured by evaluating the expected cost of completing the tree if the split is based on the selected feature. This splitting generates new children nodes (as many nodes as there are possible values for this feature) and each example is then allocated to the child node that has the feature value that corresponds to the value of the feature for that example. The same exercise is repeated on all the children in a recursive fashion until the complete tree is built.

This module also predicts the localizations for query proteins. For each query, it reads the protein's feature values and traverses the already built decision tree, based on these feature values, to determine the localization(s) for the query. Thus, at each decision node, the branch that corresponds to the value of the feature that corresponds to that node is taken and the tree traversal continues until a terminal node (one with no child) is reached. The localization(s) corresponding to this node are selected as the predicted localization(s). If a terminal node is reached for which there is no associated example, hence no localization, the system cannot predict the localization for such a query and a message is printed in the output.
Input: Train.dat, Query.dat
Output: Prediction.dat, Tree.dat

A4.5. PerformanceEvaluationModules

The following modules and scripts are used to prepare the files for performing a 5-fold cross-validation and proceeding with evaluation of the classifier. With minor modifications, the following descriptions apply equally well to both locatlizations (17 sites and 9 sites).

Module: PrepareCrossValidationFiles
Description: This script performs the following tasks:

- Invokes **LumpTo17Sits** script to compress features-localization matrix file (Matrix.dat) into a file (MatrixUnique) containing only 17 localization sites
- Runs **DividePerNumSites.cpp** module to read the file of training data (MatrixUnique2, first 2-columns version of MatrixUnique) and to distribute it into 4 files (S1.dat, S2.dat, S3.dat, and S4.dat) containing protein Id and reported localization sites for training proteins reported to localize to 1, 2, 3 and 4 sites respectively. Each of the output files thus obtained contains lines with format: *Protein_name Site_1_flag Site_2_flag . . . Site_17_flag* where, each *flag* indicates if, yes/no, the protein is localized to the corresponding site
- Invoke the shell command: **sort** -r --key=2,18 that sorts the files S1.dat to S4.dat in descending order of fields 2 to 18 to obtain the files: S1Sorted.dat, . . . , S4Sorted.dat
- Extracts the first field (protein name) of the latter 4 files and creates a shell script that allows distribution of corresponding protein examples, into the files CV1, CV2, CV3, CV4 and CV5 in a work directory in such a way that each consecutive example goes to the next CVx File
- Runs **BuildTrainQuery** script to gererate 5 pairs of Train.dat and Query.dat.

Input: Matrix.dat
Output: Train.dat.i, Query.dat.i for i=1,..,5

Module: CrossValidate

Description: Performs the following tasks for each of the 5 training-query pairs of files:

- Invokes **Localizer.cpp** that builds a decision tree and classifies the proteins in a query file
- Invokes **SummarizePrediction** that removes duplicate occurrences in prediction file, conserves first 2 columns of the query file and prepares actual and predicted localization files
- Renames the outputs to get Actuals.dat.i and Predictions.dat.i for i = 1 to 5
- Merges these files into Predictions.dat and Actuals.dat

Input: Train.dat.i, Query.dat.i for i=1,..,5
Output: Predictions.dat, Actuals.dat

Module: EvaluatePerformance.cpp

Description: Compares the reported localizations with the predicted ones. It generates 2 binary matrices representing the actual sites and predicted sites for each of the 17 sub-cellular sites. It compares these matrices and calculates the number of the cases where a) at least one of the reported sites is predicted, b) when all the reported sites are predicted, and c) when all the reported sites are predicted and there is no false positive.
Input: Predictions.dat, Actuals.dat
Output: Performance.dat

Module: EvaluateSensitivity.cpp

Description: Determines the per-site sensitivity of the system when localization is made to 17 sites. It calculates and outputs TP/(TP + FN) that shows the percentage of the total number of reported localizations that system can successfully predict.
It starts by reading the files of actual and predicted localizations. It generates 2 binary matrices representing the actual sites and predicted sites for each of the 17 sub-cellular sites.
By comparing the corresponding elements of these two matrices, it calculates the number of sites truly predicted (TP) for each protein. It sums all the numbers thus calculated over all the proteins to obtain the total number of sites truly calculated. It finally calculates the measure of Sensitivity as a percentage of this number to the total number of all sites reported for all the proteins.
Input: Predictions.dat, Actuals.dat
Output:Sensitivity.dat

Module: EvaluateSpecificity.cpp

Description: Determines the per-site specificity of the system when localization is made to 17 sites. It calculates and outputs TN/(TN + FP) that shows the percentage of the total number of sites that system can successfully predict as not being the localization site.
It starts by reading the files of actual and predicted localizations. It generates 2 binary matrices representing the actual sites and predicted sites for each of the 17 sub-cellular sites.
By comparing the corresponding elements of these two matrices, it calculates the number of sites truly predicted as negative (TN) for each protein. It sums all the numbers thus calculated over all the proteins to obtain the total number of sites truly predicted as negative. It finally calculates the measure of Specificity as a percentage of this number to the total number of all negative sites reported for all the proteins.
Input: Predictions.dat, Actuals.dat
Output:Specificity.dat

Module: EvaluatePerformanceWithContainment

Description: Compares the reported localizations with the predicted ones. It generates 2 binary matrices representing the actual sites and predicted sites for each of the 17 sub-cellular sites. It compares these matrices and calculates the number of the cases where a) at least one of

the reported sites is predicted, b) when all the reported sites are predicted, and c) when all the reported sites are predicted and there is no false positive. Here the sub-cellular containment is also taken into consideration. As such if a query protein is predicted to localize to a particular site and this site is an ancestor of the actual reported site (as per hierarchy of cellular components) then this prediction is considered as a true positive as well.
Input: Predictions.dat, Actuals.dat
Output: PerformanceWithContainment.dat

A4.6. Tree Analysis: Characteristic Features Determination Modules

The following modules determine the features of fungal proteins implicated in localization as per information contained in the decision tree:

Module: DeriveFullTree
Description: This script performs the following:

- Merges a pair of training and query files (ex: Train.dat.1 and Query.dat.1) to obtain a full set of available examples (Train.dat.full)
- Runs **localizer.cpp** on Train.dat.full to produce a full tree of localizations (excludes distribution information). This result (Tree.dat.full) represents the full decision tree for 17-site localization of 4529 proteins into 5659 sub-cellular sites

Input: Train.dat.1, Query.dat.1
Output: Tree.dat.full
Module: FindSplittingFeatures
Description: Scans the full decision tree and determines all the features associated with internal nodes (all nodes including root on which branching takes place).
Input: Tree.dat.full
Output: SplittingFeatures.dat

A4.7. Tree Analysis: Targeting Motifs Validation Modules

The following modules validate the proposed targeting motifs:
Module: FindRuleFeatureValues
Description: Scans the path from the root of the full decision tree to each of the terminal nodes. It determines the values of all the features based on which the splitting along the path took place. It then outputs the localization associated with each terminal node together with the corresponding feature values.
Input: Tree.dat.full
Output: ExRuleFeatureValues.dat

Module: ValidateRules.cpp
Description: HypRuleFeatureValues.dat is a features-localization correspondance matrix that has one row per localization site. The ith position in each row of this matrix indicates whether, as per the targeting motifs hypotheses in Table-4, the ith targeting motif needs to be present for localization to the localization site corresponding to the row. **ValidateRules** module compares the list of hypothesized targeting motifs for localization to each sub-cellular site (HypRuleFeatureValues.dat) with the actual feature values of the proteins localized to the corresponding site (obtained from the terminal nodes of the decision tree) in order to determine the proportion of the cases that confirm the targeting motif localization rules.
Input: ExRuleFeatureValues.dat, HypRuleFeatureValues.dat
Output: ValidRulesResults.dat

A4.8. Tree Analysis: Discriminatory Power Determination Module

The following module determines the features with highest discriminatory power for localization:

Module: FindTop4Levels
Description: Scans the full decision tree and determines the features associated with each of the nodes that are less than five levels deep with respect to the root. It then outputs these features in form of a hierarchical tree 4 levels deep.
Input: Tree.dat.full
Output: TreeTop4Levels.dat

A4.9. Tree Analysis: Localization Necessary Conditions Determination Modules

The following modules determine the necessary conditions for localization to specific sites:

Module: CategorizeExmplesPerSite
Description: This script reads the training set of protein examples for localization to 17 sites and generates 17 individual files each containing all localization examples for a specific site.
Input: MatrixUnique
Output: CellC, CellW, Cytop, Endos, EndRe, ExtRg, Golgi, MitIM, MitSp, MitMt, MitOM, Nucls, NucEn, Nucol, Perox, Plasm, Vacuo

Module: FindLocalizationNecessaryFeatures.cpp
Description: By scanning all the proteins localized to each sub-cellular site, this module determines the feature values that uniformly occur in all proteins localized to each specific site. These constitute the necessary feature values for localization to corresponding sites.
Input: CellC, CellW, Cytop, Endos, EndRe, ExtRg, Golgi, MitIM, MitSp, MitMt, MitOM, Nucls, NucEn, Nucol, Perox, Plasm, Vacuo
Output: NecessaryFeatureValues.dat

APPENDIX–5: DATA SET CONTENT

Table-A5 enumerates the 4529 protein IDs that were used as the data set to evaluate the system:

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
1	6	ORF19.1059	YBL046W	YDR397C	YHR030C	YLR022C	YNL234W
7	12	ORF19.1065	YBL051C	YDR398W	YHR031C	YLR023C	YNL236W
13	18	ORF19.1232	YBL052C	YDR399W	YHR033W	YLR024C	YNL238W
19	24	ORF19.125	YBL054W	YDR400W	YHR034C	YLR025W	YNL239W
25	30	ORF19.1263	YBL055C	YDR407C	YHR036W	YLR026C	YNL240C
31	36	ORF19.1321	YBL056W	YDR408C	YHR037W	YLR028C	YNL241C
37	42	ORF19.1378	YBL057C	YDR409W	YHR038W	YLR032W	YNL245C
43	48	ORF19.1597	YBL058W	YDR410C	YHR039C	YLR033W	YNL246W
49	54	ORF19.1614	YBL059C-A	YDR411C	YHR040W	YLR034C	YNL247W
55	60	ORF19.1623	YBL059W	YDR412W	YHR041C	YLR035C	YNL249C
61	66	ORF19.1738	YBL060W	YDR416W	YHR042W	YLR038C	YNL250W
67	72	ORF19.1770	YBL063W	YDR420W	YHR043C	YLR039C	YNL251C
73	78	ORF19.1779	YBL064C	YDR422C	YHR045W	YLR040C	YNL252C
79	84	ORF19.1816	YBL068W	YDR423C	YHR046C	YLR042C	YNL253W
85	90	ORF19.1837	YBL071W-A	YDR425W	YHR047C	YLR043C	YNL254C
91	96	ORF19.1896	YBL074C	YDR427W	YHR049W	YLR044C	YNL255C
97	102	ORF19.1944	YBL076C	YDR429C	YHR050W	YLR045C	YNL256W
103	108	ORF19.1957	YBL078C	YDR430C	YHR051W	YLR050C	YNL257C
109	114	ORF19.1973	YBL079W	YDR432W	YHR052W	YLR051C	YNL258C
115	120	ORF19.2014	YBL080C	YDR434W	YHR053C	YLR052W	YNL259C
121	126	ORF19.2060	YBL082C	YDR437W	YHR055C	YLR054C	YNL260C
127	132	ORF19.2062	YBL088C	YDR439W	YHR056C	YLR055C	YNL262W
133	138	ORF19.2075	YBL090W	YDR440W	YHR058C	YLR059C	YNL263C
139	144	ORF19.2179	YBL093C	YDR441C	YHR059W	YLR063W	YNL264C
145	150	ORF19.220	YBL095W	YDR444W	YHR063C	YLR064W	YNL265C
151	156	ORF19.2248	YBL097W	YDR446W	YHR064C	YLR066W	YNL267W
157	162	ORF19.242	YBL098W	YDR448W	YHR066W	YLR067C	YNL268W
163	168	ORF19.2531	YBL099W	YDR449C	YHR067W	YLR068W	YNL272C
169	174	ORF19.2557	YBL101C	YDR450W	YHR068W	YLR069C	YNL273W
175	180	ORF19.2559	YBL102W	YDR451C	YHR069C	YLR071C	YNL274C
181	186	ORF19.2560	YBL103C	YDR452W	YHR070W	YLR072W	YNL275W
187	192	ORF19.2614	YBL104C	YDR453C	YHR072W	YLR073C	YNL277W
193	198	ORF19.2706	YBL105C	YDR454C	YHR072W-A	YLR074C	YNL278W
199	204	ORF19.2770.1	YBL106C	YDR456W	YHR073W	YLR077W	YNL281W
205	210	ORF19.2803	YBL107C	YDR457W	YHR074W	YLR078C	YNL282W
211	216	ORF19.2877	YBR001C	YDR460W	YHR075C	YLR079W	YNL283C
217	222	ORF19.2884	YBR002C	YDR461W	YHR076W	YLR080W	YNL286W
223	228	ORF19.2990	YBR003W	YDR463W	YHR079C-A	YLR082C	YNL287W
229	234	ORF19.2992	YBR005W	YDR464W	YHR080C	YLR084C	YNL288W
235	240	ORF19.3010.1	YBR006W	YDR465C	YHR081W	YLR085C	YNL290W
241	246	ORF19.3013	YBR008C	YDR469W	YHR082C	YLR086W	YNL291C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
247	252	ORF19.3014	YBR010W	YDR470C	YHR083W	YLR087C	YNL292W
253	258	ORF19.3111	YBR014C	YDR472W	YHR085W	YLR088W	YNL297C
259	264	ORF19.3138	YBR015C	YDR473C	YHR086W	YLR089C	YNL299W
265	270	ORF19.3188	YBR016W	YDR476C	YHR087W	YLR090W	YNL300W
271	276	ORF19.3340	YBR017C	YDR477W	YHR088W	YLR091W	YNL305C
277	282	ORF19.3347	YBR018C	YDR479C	YHR089C	YLR092W	YNL306W
283	288	ORF19.3366	YBR021W	YDR481C	YHR090C	YLR094C	YNL308C
289	294	ORF19.3370	YBR023C	YDR482C	YHR091C	YLR095C	YNL309W
295	300	ORF19.350	YBR024W	YDR484W	YHR094C	YLR096W	YNL310C
301	306	ORF19.3548.1	YBR025C	YDR485C	YHR097C	YLR098C	YNL313C
307	312	ORF19.3575	YBR026C	YDR486C	YHR098C	YLR100W	YNL314W
313	318	ORF19.3618	YBR028C	YDR487C	YHR099W	YLR103C	YNL316C
319	324	ORF19.3651	YBR030W	YDR489W	YHR101C	YLR105C	YNL317W
325	330	ORF19.367	YBR034C	YDR490C	YHR102W	YLR106C	YNL318C
331	336	ORF19.3708	YBR036C	YDR492W	YHR103W	YLR107W	YNL322C
337	342	ORF19.3829	YBR037C	YDR493W	YHR104W	YLR108C	YNL323W
343	348	ORF19.3838	YBR039W	YDR494W	YHR105W	YLR109W	YNL325C
349	354	ORF19.3895	YBR040W	YDR496C	YHR106W	YLR110C	YNL326C
355	360	ORF19.395	YBR041W	YDR497C	YHR107C	YLR113W	YNL327W
361	366	ORF19.3967	YBR042C	YDR498C	YHR108W	YLR114C	YNL328C
367	372	ORF19.3997	YBR043C	YDR499W	YHR109W	YLR115W	YNL329C
373	378	ORF19.4015	YBR044C	YDR505C	YHR111W	YLR116W	YNL330C
379	384	ORF19.4035	YBR046C	YDR508C	YHR112C	YLR118C	YNL336W
385	390	ORF19.4045	YBR047W	YDR510W	YHR113W	YLR119W	YNR001C
391	396	ORF19.4152	YBR052C	YDR511W	YHR114W	YLR120C	YNR002C
397	402	ORF19.4192	YBR054W	YDR513W	YHR115C	YLR121C	YNR003C
403	408	ORF19.4308	YBR055C	YDR514C	YHR116W	YLR126C	YNR004W
409	414	ORF19.4328	YBR056W	YDR515W	YHR117W	YLR129W	YNR006W
415	420	ORF19.4424	YBR057C	YDR516C	YHR119W	YLR130C	YNR007C
421	426	ORF19.4477	YBR059C	YDR517W	YHR120W	YLR131C	YNR009W
427	432	ORF19.4555	YBR061C	YDR518W	YHR121W	YLR132C	YNR010W
433	438	ORF19.4660	YBR065C	YDR519W	YHR122W	YLR134W	YNR012W
439	444	ORF19.4773	YBR067C	YDR520C	YHR124W	YLR135W	YNR014W
445	450	ORF19.4774	YBR070C	YDR522C	YHR127W	YLR136C	YNR015W
451	456	ORF19.4784	YBR071W	YDR523C	YHR131C	YLR138W	YNR016C
457	462	ORF19.4821	YBR072W	YDR524C	YHR132C	YLR139C	YNR017W
463	468	ORF19.4892	YBR077C	YDR525W-A	YHR132W-A	YLR142W	YNR018W
469	474	ORF19.4930	YBR078W	YDR527W	YHR135C	YLR143W	YNR019W
475	480	ORF19.4980	YBR079C	YDR528W	YHR136C	YLR145W	YNR021W
481	486	ORF19.4987	YBR080C	YDR529C	YHR137W	YLR146C	YNR022C
487	492	ORF19.5076	YBR081C	YDR530C	YHR142W	YLR147C	YNR024W
493	498	ORF19.5089	YBR082C	YDR531W	YHR143W	YLR148W	YNR026C
499	504	ORF19.5100	YBR084W	YDR532C	YHR144C	YLR150W	YNR027W
505	510	ORF19.5107	YBR085C-A	YDR534C	YHR147C	YLR151C	YNR028W
511	516	ORF19.5112	YBR086C	YDR538W	YHR148W	YLR153C	YNR029C
517	522	ORF19.5117	YBR088C	YDR539W	YHR150W	YLR154C	YNR030W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
523	528	ORF19.5119	YBR089C-A	YDR540C	YHR152W	YLR154W-C	YNR031C
529	534	ORF19.5305	YBR090C	YEL001C	YHR154W	YLR155C	YNR032W
535	540	ORF19.5343	YBR093C	YEL002C	YHR155W	YLR163C	YNR033W
541	546	ORF19.5383	YBR094W	YEL003W	YHR156C	YLR165C	YNR034W
547	552	ORF19.5389	YBR095C	YEL004W	YHR158C	YLR166C	YNR034W-A
553	558	ORF19.5423	YBR096W	YEL005C	YHR159W	YLR167W	YNR035C
559	564	ORF19.548	YBR097W	YEL006W	YHR160C	YLR168C	YNR037C
565	570	ORF19.5526	YBR101C	YEL007W	YHR162W	YLR170C	YNR038W
571	576	ORF19.5542	YBR102C	YEL009C	YHR163W	YLR172C	YNR039C
577	582	ORF19.5558	YBR103W	YEL011W	YHR164C	YLR174W	YNR040W
583	588	ORF19.5585	YBR105C	YEL012W	YHR165C	YLR175W	YNR044W
589	594	ORF19.5636	YBR106W	YEL013W	YHR167W	YLR176C	YNR045W
595	600	ORF19.5672	YBR107C	YEL015W	YHR168W	YLR177W	YNR046W
601	606	ORF19.5714	YBR108W	YEL018W	YHR169W	YLR179C	YNR047W
607	612	ORF19.5716	YBR109C	YEL019C	YHR170W	YLR180W	YNR050C
613	618	ORF19.5741	YBR110W	YEL020C	YHR171W	YLR181C	YNR051C
619	624	ORF19.5788	YBR111C	YEL020W-A	YHR172W	YLR182W	YNR052C
625	630	ORF19.5928	YBR111W-A	YEL022W	YHR174W	YLR183C	YNR053C
631	636	ORF19.6000	YBR112C	YEL024W	YHR175W	YLR186W	YNR054C
637	642	ORF19.6001	YBR114W	YEL025C	YHR176W	YLR187W	YNR055C
643	648	ORF19.6010	YBR117C	YEL026W	YHR178W	YLR188W	YNR060W
649	654	ORF19.6034	YBR119W	YEL029C	YHR179W	YLR189C	YNR061C
655	660	ORF19.6124	YBR120C	YEL030W	YHR181W	YLR190W	YNR067C
661	666	ORF19.6176	YBR121C	YEL031W	YHR182W	YLR192C	YNR070W
667	672	ORF19.6190	YBR122C	YEL032W	YHR183W	YLR193C	YNR074C
673	678	ORF19.6214	YBR125C	YEL034W	YHR186C	YLR194C	YNR075W
679	684	ORF19.6367	YBR126C	YEL036C	YHR187W	YLR195C	YOL004W
685	690	ORF19.637	YBR127C	YEL037C	YHR188C	YLR196W	YOL006C
691	696	ORF19.6403.1	YBR129C	YEL038W	YHR189W	YLR197W	YOL007C
697	702	ORF19.6515	YBR130C	YEL039C	YHR192W	YLR199C	YOL008W
703	708	ORF19.6540	YBR131W	YEL040W	YHR193C	YLR200W	YOL010W
709	714	ORF19.657	YBR132C	YEL043W	YHR194W	YLR201C	YOL011W
715	720	ORF19.6582	YBR135W	YEL044W	YHR195W	YLR203C	YOL012C
721	726	ORF19.6673	YBR136W	YEL046C	YHR196W	YLR204W	YOL013C
727	732	ORF19.6763	YBR137W	YEL047C	YHR197W	YLR205C	YOL016C
733	738	ORF19.6814	YBR138C	YEL048C	YHR198C	YLR208W	YOL017W
739	744	ORF19.6854	YBR139W	YEL052W	YHR199C	YLR210W	YOL019W
745	750	ORF19.689	YBR140C	YEL053C	YHR200W	YLR211C	YOL020W
751	756	ORF19.6975	YBR141C	YEL055C	YHR201C	YLR212C	YOL021C
757	762	ORF19.7021	YBR142W	YEL056W	YHR202W	YLR213C	YOL022C
763	768	ORF19.7030	YBR143C	YEL058W	YHR204W	YLR214W	YOL023W
769	774	ORF19.709	YBR145W	YEL059C-A	YHR205W	YLR215C	YOL025W
775	780	ORF19.7111.1	YBR146W	YEL061C	YHR206W	YLR216C	YOL026C
781	786	ORF19.7114	YBR149W	YEL063C	YHR207C	YLR218C	YOL027C
787	792	ORF19.7178	YBR150C	YEL065W	YHR208W	YLR220W	YOL030W
793	798	ORF19.7188	YBR151W	YEL066W	YHR211W	YLR221C	YOL031C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
799	804	ORF19.7394	YBR152W	YEL071W	YHR215W	YLR222C	YOL032W
805	810	ORF19.7398.1	YBR155W	YER001W	YHR216W	YLR223C	YOL033W
811	816	ORF19.7417	YBR158W	YER002W	YIL001W	YLR225C	YOL034W
817	822	ORF19.7436	YBR159W	YER003C	YIL003W	YLR226W	YOL038W
823	828	ORF19.7447	YBR160W	YER004W	YIL004C	YLR227C	YOL041C
829	834	ORF19.7551	YBR161W	YER005W	YIL005W	YLR228C	YOL042W
835	840	ORF19.7586	YBR162C	YER006W	YIL006W	YLR229C	YOL043C
841	846	ORF19.7622	YBR162W-A	YER007C-A	YIL007C	YLR231C	YOL044W
847	852	ORF19.7668	YBR163W	YER008C	YIL008W	YLR233C	YOL045W
853	858	ORF19.895	YBR164C	YER009W	YIL009C-A	YLR237W	YOL047C
859	864	ORF19.922	YBR165W	YER011W	YIL010W	YLR238W	YOL048C
865	870	ORF19.940	YBR166C	YER012W	YIL011W	YLR239C	YOL051W
871	876	ORF19.97	YBR167C	YER014W	YIL016W	YLR242C	YOL052C
877	882	SPAC1002.15C	YBR168W	YER015W	YIL017C	YLR244C	YOL052C-A
883	888	SPAC1006.08	YBR169C	YER016W	YIL019W	YLR245C	YOL054W
889	894	SPAC1071.01C	YBR170C	YER017C	YIL022W	YLR246W	YOL057W
895	900	SPAC1093.06C	YBR171W	YER018C	YIL026C	YLR247C	YOL058W
901	906	SPAC1142.05	YBR172C	YER019W	YIL027C	YLR248W	YOL059W
907	912	SPAC11G7.02	YBR173C	YER020W	YIL030C	YLR249W	YOL060C
913	918	SPAC12G12.03	YBR175W	YER021W	YIL031W	YLR250W	YOL061W
919	924	SPAC12G12.04	YBR176W	YER022W	YIL033C	YLR251W	YOL064C
925	930	SPAC12G12.14C	YBR177C	YER023W	YIL036W	YLR253W	YOL065C
931	936	SPAC13A11.01C	YBR179C	YER024W	YIL038C	YLR254C	YOL066C
937	942	SPAC13A11.03	YBR181C	YER025W	YIL039W	YLR256W	YOL067C
943	948	SPAC13C5.03	YBR182C	YER027C	YIL040W	YLR257W	YOL068C
949	954	SPAC13D6.05	YBR183W	YER028C	YIL041W	YLR258W	YOL069W
955	960	SPAC13F5.02C	YBR185C	YER029C	YIL043C	YLR259C	YOL070C
961	966	SPAC13F5.06C	YBR187W	YER030W	YIL044C	YLR262C-A	YOL071W
967	972	SPAC140.02	YBR188C	YER031C	YIL045W	YLR263W	YOL072W
973	978	SPAC1420.03	YBR189W	YER033C	YIL046W	YLR265C	YOL077C
979	984	SPAC1486.04C	YBR192W	YER034W	YIL047C	YLR266C	YOL077W-A
985	990	SPAC1486.05	YBR193C	YER035W	YIL048W	YLR268W	YOL078W
991	996	SPAC14C4.02C	YBR194W	YER036C	YIL051C	YLR270W	YOL080C
997	1002	SPAC14C4.03	YBR195C	YER037W	YIL053W	YLR271W	YOL081W
1003	1008	SPAC14C4.09	YBR197C	YER038C	YIL056W	YLR272C	YOL082W
1009	1014	SPAC14C4.14	YBR198C	YER040W	YIL057C	YLR274W	YOL084W
1015	1020	SPAC1565.06C	YBR201W	YER042W	YIL061C	YLR275W	YOL087C
1021	1026	SPAC1565.08	YBR202W	YER043C	YIL063C	YLR277C	YOL089C
1027	1032	SPAC15A10.15	YBR204C	YER044C	YIL064W	YLR278C	YOL090W
1033	1038	SPAC15E1.09	YBR207W	YER045C	YIL065C	YLR281C	YOL091W
1039	1044	SPAC15F9.02	YBR208C	YER046W	YIL066C	YLR283W	YOL092W
1045	1050	SPAC16.01	YBR212W	YER047C	YIL067C	YLR284C	YOL093W
1051	1056	SPAC16.02C	YBR214W	YER048W-A	YIL068C	YLR285W	YOL095C
1057	1062	SPAC1610.03C	YBR216C	YER049W	YIL070C	YLR286C	YOL096C
1063	1068	SPAC167.03C	YBR218C	YER050C	YIL071C	YLR287C	YOL098C
1069	1074	SPAC1687.01	YBR221C	YER052C	YIL072W	YLR289W	YOL100W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
1075	1080	SPAC1687.13C	YBR222C	YER053C	YIL074C	YLR290C	YOL103W
1081	1086	SPAC16A10.05C	YBR223C	YER053C-A	YIL075C	YLR291C	YOL104C
1087	1092	SPAC16A10.06C	YBR227C	YER056C	YIL076W	YLR292C	YOL107W
1093	1098	SPAC16A10.07C	YBR229C	YER057C	YIL077C	YLR293C	YOL109W
1099	1104	SPAC16E8.09	YBR230C	YER062C	YIL078W	YLR295C	YOL110W
1105	1110	SPAC1751.03	YBR231C	YER063W	YIL079C	YLR297W	YOL111C
1111	1116	SPAC1782.09C	YBR233W	YER064C	YIL083C	YLR298C	YOL115W
1117	1122	SPAC1782.10C	YBR236C	YER067W	YIL084C	YLR300W	YOL116W
1123	1128	SPAC1783.04C	YBR238C	YER068W	YIL087C	YLR301W	YOL117W
1129	1134	SPAC1783.05	YBR239C	YER069W	YIL090W	YLR303W	YOL119C
1135	1140	SPAC1783.07C	YBR241C	YER070W	YIL091C	YLR304C	YOL122C
1141	1146	SPAC1786.03	YBR242W	YER071C	YIL092W	YLR305C	YOL123W
1147	1152	SPAC17A5.04C	YBR244W	YER072W	YIL093C	YLR309C	YOL124C
1153	1158	SPAC17C9.01C	YBR245C	YER073W	YIL094C	YLR310C	YOL125W
1159	1164	SPAC17C9.08	YBR247C	YER074W	YIL096C	YLR312W-A	YOL126C
1165	1170	SPAC17C9.13C	YBR249C	YER074W-A	YIL097W	YLR315W	YOL129W
1171	1176	SPAC17D4.02	YBR251W	YER075C	YIL098C	YLR316C	YOL130W
1177	1182	SPAC17G6.16C	YBR252W	YER076C	YIL099W	YLR318W	YOL132W
1183	1188	SPAC17G8.03C	YBR253W	YER077C	YIL101C	YLR319C	YOL133W
1189	1194	SPAC17G8.09	YBR254C	YER078C	YIL103W	YLR320W	YOL135C
1195	1200	SPAC17G8.10C	YBR255W	YER079W	YIL104C	YLR321C	YOL137W
1201	1206	SPAC17G8.13C	YBR257W	YER080W	YIL105C	YLR323C	YOL138C
1207	1212	SPAC17H9.05	YBR258C	YER081W	YIL106W	YLR324W	YOL139C
1213	1218	SPAC17H9.06C	YBR261C	YER082C	YIL108W	YLR327C	YOL141W
1219	1224	SPAC17H9.10C	YBR262C	YER083C	YIL110W	YLR328W	YOL142W
1225	1230	SPAC17H9.19C	YBR263W	YER086W	YIL111W	YLR329W	YOL143C
1231	1236	SPAC1805.04	YBR264C	YER087C-B	YIL112W	YLR330W	YOL145C
1237	1242	SPAC1805.07C	YBR265W	YER088C	YIL113W	YLR332W	YOL146W
1243	1248	SPAC1805.08	YBR267W	YER089C	YIL114C	YLR335W	YOL148C
1249	1254	SPAC1805.15C	YBR269C	YER090W	YIL119C	YLR336C	YOL149W
1255	1260	SPAC1805.17	YBR271W	YER091C	YIL120W	YLR343W	YOL151W
1261	1266	SPAC18B11.02C	YBR272C	YER092W	YIL121W	YLR344W	YOL154W
1267	1272	SPAC18B11.04	YBR273C	YER094C	YIL122W	YLR345W	YOL155C
1273	1278	SPAC18G6.02C	YBR274W	YER095W	YIL123W	YLR346C	YOL159C-A
1279	1284	SPAC18G6.03	YBR275C	YER099C	YIL124W	YLR347C	YOR001W
1285	1290	SPAC18G6.15	YBR278W	YER100W	YIL125W	YLR348C	YOR002W
1291	1296	SPAC1952.07	YBR279W	YER101C	YIL126W	YLR350W	YOR004W
1297	1302	SPAC19A8.12	YBR281C	YER103W	YIL127C	YLR351C	YOR006C
1303	1308	SPAC19D5.01	YBR282W	YER104W	YIL128W	YLR355C	YOR007C
1309	1314	SPAC19D5.03	YBR283C	YER105C	YIL129C	YLR356W	YOR008C
1315	1320	SPAC19E9.01C	YBR286W	YER106W	YIL130W	YLR357W	YOR009W
1321	1326	SPAC19E9.02	YBR287W	YER107C	YIL131C	YLR362W	YOR011W
1327	1332	SPAC19G12.01C	YBR290W	YER109C	YIL132C	YLR363C	YOR014W
1333	1338	SPAC19G12.14	YBR293W	YER110C	YIL134W	YLR363W-A	YOR017W
1339	1344	SPAC1A6.07	YBR294W	YER111C	YIL135C	YLR364W	YOR018W
1345	1350	SPAC1B1.03C	YBR296C	YER112W	YIL136W	YLR368W	YOR020C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
1351	1356	SPAC1B2.05	YBR302C	YER113C	YIL137C	YLR369W	YOR022C
1357	1362	SPAC1D4.01	YCL001W	YER119C	YIL139C	YLR372W	YOR023C
1363	1368	SPAC1D4.12	YCL004W	YER120W	YIL142W	YLR373C	YOR026W
1369	1374	SPAC1F3.06C	YCL005W	YER122C	YIL144W	YLR375W	YOR027W
1375	1380	SPAC1F7.03	YCL005W-A	YER123W	YIL145C	YLR376C	YOR028C
1381	1386	SPAC1F7.05	YCL008C	YER125W	YIL149C	YLR377C	YOR033C
1387	1392	SPAC20G4.02C	YCL009C	YER126C	YIL150C	YLR378C	YOR035C
1393	1398	SPAC20G4.04C	YCL010C	YER129W	YIL153W	YLR380W	YOR036W
1399	1404	SPAC20G8.01	YCL011C	YER132C	YIL154C	YLR381W	YOR037W
1405	1410	SPAC20G8.05C	YCL017C	YER133W	YIL155C	YLR382C	YOR038C
1411	1416	SPAC20H4.10	YCL018W	YER134C	YIL157C	YLR383W	YOR040W
1417	1422	SPAC222.10C	YCL026C-A	YER139C	YIL158W	YLR384C	YOR042W
1423	1428	SPAC222.12C	YCL026C-B	YER140W	YIL161W	YLR385C	YOR044W
1429	1434	SPAC227.08C	YCL027W	YER141W	YIL162W	YLR386W	YOR045W
1435	1440	SPAC22A12.07C	YCL028W	YER143W	YIL172C	YLR387C	YOR046C
1441	1446	SPAC22A12.11	YCL029C	YER145C	YIL173W	YLR390W-A	YOR047C
1447	1452	SPAC22A12.15C	YCL031C	YER146W	YIR001C	YLR392C	YOR048C
1453	1458	SPAC22E12.07	YCL032W	YER147C	YIR002C	YLR393W	YOR049C
1459	1464	SPAC22E12.08	YCL034W	YER148W	YIR004W	YLR394W	YOR051C
1465	1470	SPAC22E12.11C	YCL035C	YER150W	YIR005W	YLR395C	YOR052C
1471	1476	SPAC22F3.02	YCL038C	YER152C	YIR006C	YLR396C	YOR056C
1477	1482	SPAC22F3.12C	YCL039W	YER153C	YIR007W	YLR398C	YOR059C
1483	1488	SPAC22F8.10C	YCL042W	YER154W	YIR009W	YLR399C	YOR060C
1489	1494	SPAC22G7.02	YCL044C	YER155C	YIR011C	YLR401C	YOR062C
1495	1500	SPAC22G7.09C	YCL045C	YER156C	YIR012W	YLR403W	YOR064C
1501	1506	SPAC22G7.10	YCL050C	YER157W	YIR014W	YLR408C	YOR065W
1507	1512	SPAC22H10.03C	YCL052C	YER159C	YIR015W	YLR409C	YOR067C
1513	1518	SPAC22H12.02	YCL054W	YER161C	YIR017C	YLR410W	YOR069W
1519	1524	SPAC23A1.06C	YCL056C	YER162C	YIR019C	YLR411W	YOR070C
1525	1530	SPAC23C11.04C	YCL057W	YER163C	YIR021W	YLR412W	YOR073W
1531	1536	SPAC23C11.05	YCL059C	YER164W	YIR022W	YLR414C	YOR074C
1537	1542	SPAC23C11.14	YCL061C	YER165W	YIR024C	YLR417W	YOR076C
1543	1548	SPAC23C11.15	YCL063W	YER166W	YIR025W	YLR418C	YOR077W
1549	1554	SPAC23C11.16	YCL064C	YER167W	YIR026C	YLR419W	YOR078W
1555	1560	SPAC23C4.15	YCL069W	YER168C	YIR031C	YLR420W	YOR079C
1561	1566	SPAC23D3.06C	YCR004C	YER170W	YIR033W	YLR421C	YOR081C
1567	1572	SPAC23E2.01	YCR005C	YER171W	YIR034C	YLR422W	YOR083W
1573	1578	SPAC23E2.02	YCR009C	YER172C	YIR035C	YLR423C	YOR084W
1579	1584	SPAC23G3.01	YCR010C	YER173W	YIR036C	YLR424W	YOR085W
1585	1590	SPAC23H3.02C	YCR011C	YER175C	YIR038C	YLR426W	YOR086C
1591	1596	SPAC23H3.05C	YCR012W	YER177W	YIR039C	YLR427W	YOR087W
1597	1602	SPAC23H3.08C	YCR016W	YER178W	YJL001W	YLR430W	YOR089C
1603	1608	SPAC23H4.12	YCR017C	YER179W	YJL002C	YLR432W	YOR090C
1609	1614	SPAC23H4.18C	YCR018C	YER182W	YJL003W	YLR435W	YOR091W
1615	1620	SPAC24B11.06C	YCR020W-B	YER183C	YJL004C	YLR436C	YOR092W
1621	1626	SPAC24B11.11C	YCR021C	YFL001W	YJL005W	YLR437C	YOR095C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
1627	1632	SPAC24C9.05C	YCR024C	YFL003C	YJL006C	YLR438C-A	YOR096W
1633	1638	SPAC24H6.05	YCR024C-A	YFL004W	YJL008C	YLR438W	YOR098C
1639	1644	SPAC24H6.06	YCR027C	YFL005W	YJL010C	YLR440C	YOR099W
1645	1650	SPAC24H6.09	YCR028C	YFL007W	YJL011C	YLR442C	YOR100C
1651	1656	SPAC25B8.16	YCR028C-A	YFL008W	YJL012C	YLR447C	YOR101W
1657	1662	SPAC25G10.02	YCR031C	YFL009W	YJL013C	YLR449W	YOR103C
1663	1668	SPAC25G10.04C	YCR032W	YFL010C	YJL014W	YLR451W	YOR106W
1669	1674	SPAC25G10.07C	YCR033W	YFL011W	YJL016W	YLR452C	YOR107W
1675	1680	SPAC25G10.08	YCR034W	YFL013C	YJL019W	YLR453C	YOR108W
1681	1686	SPAC25G10.09C	YCR035C	YFL014W	YJL023C	YLR454W	YOR112W
1687	1692	SPAC26A3.01	YCR036W	YFL016C	YJL026W	YLR455W	YOR113W
1693	1698	SPAC26A3.12C	YCR042C	YFL017C	YJL029C	YLR457C	YOR115C
1699	1704	SPAC26A3.15C	YCR043C	YFL017W-A	YJL030W	YLR459W	YOR117W
1705	1710	SPAC27D7.03C	YCR044C	YFL018C	YJL034W	YML001W	YOR118W
1711	1716	SPAC27D7.04	YCR046C	YFL021W	YJL035C	YML004C	YOR119C
1717	1722	SPAC27D7.05C	YCR047C	YFL022C	YJL041W	YML005W	YOR120W
1723	1728	SPAC27D7.12C	YCR048W	YFL023W	YJL043W	YML006C	YOR123C
1729	1734	SPAC27D7.13C	YCR051W	YFL024C	YJL044C	YML007C-A	YOR128C
1735	1740	SPAC27E2.09	YCR052W	YFL025C	YJL045W	YML007W	YOR129C
1741	1746	SPAC27F1.04C	YCR053W	YFL026W	YJL047C	YML008C	YOR130C
1747	1752	SPAC27F1.09C	YCR054C	YFL028C	YJL048C	YML010W	YOR131C
1753	1758	SPAC29A4.07	YCR057C	YFL029C	YJL050W	YML011C	YOR132W
1759	1764	SPAC29A4.10	YCR059C	YFL030W	YJL052W	YML012W	YOR136W
1765	1770	SPAC29A4.11	YCR060W	YFL033C	YJL053W	YML013W	YOR138C
1771	1776	SPAC29A4.18	YCR061W	YFL034C-B	YJL054W	YML014W	YOR140W
1777	1782	SPAC29B12.02C	YCR063W	YFL036W	YJL055W	YML015C	YOR141C
1783	1788	SPAC29B12.03	YCR065W	YFL037W	YJL056C	YML016C	YOR142W
1789	1794	SPAC2C4.07C	YCR066W	YFL038C	YJL058C	YML017W	YOR143C
1795	1800	SPAC2C4.11C	YCR067C	YFL039C	YJL059W	YML018C	YOR144C
1801	1806	SPAC2C4.15C	YCR068W	YFL041W	YJL060W	YML019W	YOR145C
1807	1812	SPAC2F3.06C	YCR071C	YFL044C	YJL061W	YML021C	YOR147W
1813	1818	SPAC2F3.15	YCR072C	YFL045C	YJL062W	YML022W	YOR148C
1819	1824	SPAC2F7.03C	YCR073W-A	YFL046W	YJL062W-A	YML023C	YOR150W
1825	1830	SPAC2F7.04	YCR075C	YFL047W	YJL063C	YML027W	YOR153W
1831	1836	SPAC2F7.11	YCR077C	YFL048C	YJL065C	YML028W	YOR156C
1837	1842	SPAC2G11.08C	YCR079W	YFL049W	YJL066C	YML030W	YOR157C
1843	1848	SPAC30C2.02	YCR082W	YFL050C	YJL068C	YML031W	YOR158W
1849	1854	SPAC30D11.07	YCR083W	YFL062W	YJL069C	YML032C	YOR159C
1855	1860	SPAC30D11.09	YCR084C	YFR001W	YJL070C	YML035C	YOR160W
1861	1866	SPAC31A2.11C	YCR086W	YFR002W	YJL071W	YML037C	YOR162C
1867	1872	SPAC31A2.14	YCR087C-A	YFR004W	YJL072C	YML038C	YOR163W
1873	1878	SPAC31G5.09C	YCR088W	YFR005C	YJL073W	YML041C	YOR164C
1879	1884	SPAC31G5.13	YCR089W	YFR006W	YJL074C	YML042W	YOR165W
1885	1890	SPAC343.03	YCR090C	YFR009W	YJL076W	YML046W	YOR171C
1891	1896	SPAC343.11C	YCR092C	YFR010W	YJL078C	YML048W	YOR172W
1897	1902	SPAC3A11.05C	YCR093W	YFR011C	YJL079C	YML049C	YOR173W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
1903	1908	SPAC3A11.08	YCR094W	YFR013W	YJL080C	YML051W	YOR174W
1909	1914	SPAC3A11.14C	YCR095C	YFR014C	YJL081C	YML052W	YOR175C
1915	1920	SPAC3A12.07	YCR097W	YFR015C	YJL082W	YML053C	YOR177C
1921	1926	SPAC3A12.11C	YCR098C	YFR016C	YJL084C	YML054C	YOR179C
1927	1932	SPAC3A12.14	YDL001W	YFR017C	YJL085W	YML055W	YOR180C
1933	1938	SPAC3C7.08C	YDL002C	YFR019W	YJL087C	YML056C	YOR181W
1939	1944	SPAC3C7.12	YDL003W	YFR021W	YJL088W	YML058W	YOR184W
1945	1950	SPAC3G6.02	YDL004W	YFR024C-A	YJL089W	YML058W-A	YOR185C
1951	1956	SPAC3G9.04	YDL005C	YFR027W	YJL093C	YML059C	YOR188W
1957	1962	SPAC3G9.07C	YDL006W	YFR028C	YJL094C	YML060W	YOR189W
1963	1968	SPAC3H1.01C	YDL007W	YFR029W	YJL096W	YML061C	YOR190W
1969	1974	SPAC3H1.12C	YDL010W	YFR031C	YJL097W	YML062C	YOR191W
1975	1980	SPAC3H8.10	YDL012C	YFR032C-A	YJL098W	YML067C	YOR193W
1981	1986	SPAC458.07	YDL014W	YFR033C	YJL099W	YML069W	YOR194C
1987	1992	SPAC4A8.03C	YDL015C	YFR034C	YJL100W	YML070W	YOR195W
1993	1998	SPAC4A8.09C	YDL019C	YFR037C	YJL102W	YML071C	YOR196C
1999	2004	SPAC4A8.11C	YDL020C	YFR040W	YJL104W	YML072C	YOR197W
2005	2010	SPAC4D7.03	YDL021W	YFR041C	YJL106W	YML074C	YOR201C
2011	2016	SPAC4F10.11	YDL022W	YFR044C	YJL109C	YML075C	YOR205C
2017	2022	SPAC4F10.20	YDL024C	YFR046C	YJL111W	YML076C	YOR206W
2023	2028	SPAC4F8.13C	YDL027C	YFR047C	YJL112W	YML077W	YOR208W
2029	2034	SPAC4F8.14C	YDL028C	YFR048W	YJL113W	YML078W	YOR209C
2035	2040	SPAC4G8.13C	YDL030W	YFR050C	YJL115W	YML079W	YOR211C
2041	2046	SPAC4G9.09C	YDL031W	YFR051C	YJL117W	YML080W	YOR212W
2047	2052	SPAC4G9.10	YDL033C	YFR052W	YJL122W	YML081C-A	YOR213C
2053	2058	SPAC56F8.04C	YDL035C	YGL001C	YJL123C	YML081W	YOR214C
2059	2064	SPAC57A10.02	YDL036C	YGL002W	YJL124C	YML082W	YOR215C
2065	2070	SPAC57A10.10C	YDL040C	YGL003C	YJL125C	YML085C	YOR219C
2071	2076	SPAC57A10.12C	YDL042C	YGL004C	YJL128C	YML086C	YOR220W
2077	2082	SPAC57A7.04C	YDL043C	YGL005C	YJL130C	YML088W	YOR221C
2083	2088	SPAC57A7.08	YDL044C	YGL006W	YJL131C	YML091C	YOR222W
2089	2094	SPAC57A7.11	YDL045C	YGL008C	YJL133W	YML092C	YOR226C
2095	2100	SPAC589.08C	YDL045W-A	YGL010W	YJL134W	YML093W	YOR227W
2101	2106	SPAC5D6.02C	YDL046W	YGL011C	YJL137C	YML094W	YOR228C
2107	2112	SPAC5D6.05	YDL047W	YGL012W	YJL138C	YML096W	YOR229W
2113	2118	SPAC637.12C	YDL048C	YGL013C	YJL139C	YML097C	YOR230W
2119	2124	SPAC644.06C	YDL049C	YGL014W	YJL140W	YML098W	YOR232W
2125	2130	SPAC644.12	YDL051W	YGL016W	YJL141C	YML099C	YOR233W
2131	2136	SPAC664.01C	YDL052C	YGL018C	YJL143W	YML101C	YOR238W
2137	2142	SPAC664.02C	YDL053C	YGL020C	YJL144W	YML102W	YOR241W
2143	2148	SPAC664.07C	YDL055C	YGL022W	YJL145W	YML103C	YOR242C
2149	2154	SPAC664.10	YDL056W	YGL023C	YJL146W	YML105C	YOR243C
2155	2160	SPAC664.11	YDL059C	YGL025C	YJL147C	YML106W	YOR244W
2161	2166	SPAC688.04C	YDL060W	YGL026C	YJL148W	YML107C	YOR245C
2167	2172	SPAC688.06C	YDL063C	YGL027C	YJL151C	YML108W	YOR246C
2173	2178	SPAC688.11	YDL064W	YGL028C	YJL153C	YML109W	YOR247W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
2179	2184	SPAC6B12.11	YDL065C	YGL029W	YJL154C	YML110C	YOR250C
2185	2190	SPAC6B12.16	YDL066W	YGL030W	YJL156C	YML111W	YOR251C
2191	2196	SPAC6F12.02	YDL067C	YGL033W	YJL157C	YML112W	YOR252W
2197	2202	SPAC6F12.09	YDL069C	YGL035C	YJL158C	YML113W	YOR253W
2203	2208	SPAC6F12.10C	YDL072C	YGL036W	YJL159W	YML114C	YOR254C
2209	2214	SPAC6F12.15C	YDL074C	YGL037C	YJL161W	YML115C	YOR258W
2215	2220	SPAC6F6.08C	YDL076C	YGL038C	YJL162C	YML116W	YOR259C
2221	2226	SPAC6F6.17	YDL077C	YGL039W	YJL166W	YML117W	YOR260W
2227	2232	SPAC6G10.02C	YDL078C	YGL040C	YJL168C	YML120C	YOR261C
2233	2238	SPAC6G9.06C	YDL080C	YGL043W	YJL170C	YML121W	YOR262W
2239	2244	SPAC6G9.11	YDL084W	YGL044C	YJL171C	YML123C	YOR265W
2245	2250	SPAC806.07	YDL085C-A	YGL047W	YJL172W	YML124C	YOR266W
2251	2256	SPAC821.04C	YDL085W	YGL048C	YJL174W	YML125C	YOR267C
2257	2262	SPAC821.06	YDL086W	YGL049C	YJL179W	YML127W	YOR269W
2263	2268	SPAC821.07C	YDL087C	YGL051W	YJL183W	YML128C	YOR270C
2269	2274	SPAC824.04	YDL088C	YGL053W	YJL186W	YML129C	YOR271C
2275	2280	SPAC824.08	YDL091C	YGL054C	YJL187C	YML130C	YOR272W
2281	2286	SPAC890.02C	YDL092W	YGL056C	YJL191W	YML131W	YOR273C
2287	2292	SPAC890.07C	YDL097C	YGL057C	YJL192C	YML132W	YOR274W
2293	2298	SPAC8C9.17C	YDL098C	YGL058W	YJL197W	YML133C	YOR275C
2299	2304	SPAC8E11.02C	YDL099W	YGL059W	YJL200C	YMR001C	YOR279C
2305	2310	SPAC8F11.10C	YDL100C	YGL060W	YJL201W	YMR002W	YOR280C
2311	2316	SPAC9.03C	YDL101C	YGL061C	YJL203W	YMR003W	YOR281C
2317	2322	SPAC926.09C	YDL103C	YGL062W	YJL204C	YMR004W	YOR283W
2323	2328	SPAC959.09C	YDL104C	YGL064C	YJL207C	YMR005W	YOR284W
2329	2334	SPAC977.10	YDL105W	YGL066W	YJL208C	YMR006C	YOR285W
2335	2340	SPAC9E9.08	YDL106C	YGL067W	YJL209W	YMR008C	YOR286W
2341	2346	SPAC9E9.10C	YDL107W	YGL068W	YJL212C	YMR009W	YOR288C
2347	2352	SPAC9G1.04	YDL108W	YGL071W	YJL217W	YMR010W	YOR289W
2353	2358	SPAC9G1.09	YDL110C	YGL073W	YJR001W	YMR011W	YOR292C
2359	2364	SPAC9G1.11C	YDL111C	YGL075C	YJR002W	YMR012W	YOR294W
2365	2370	SPACUNK4.07C	YDL112W	YGL077C	YJR003C	YMR013C	YOR295W
2371	2376	SPAP27G11.15	YDL115C	YGL078C	YJR004C	YMR014W	YOR296W
2377	2382	SPAP8A3.06	YDL116W	YGL079W	YJR007W	YMR016C	YOR297C
2383	2388	SPAPB17E12.02	YDL117W	YGL080W	YJR008W	YMR020W	YOR298C-A
2389	2394	SPAPB1A10.09	YDL120W	YGL082W	YJR009C	YMR021C	YOR299W
2395	2400	SPAPB1E7.09	YDL121C	YGL083W	YJR010C-A	YMR023C	YOR304C-A
2401	2406	SPAPB24D3.10C	YDL122W	YGL084C	YJR010W	YMR024W	YOR304W
2407	2412	SPAPB2B4.02	YDL123W	YGL085W	YJR014W	YMR025W	YOR305W
2413	2418	SPAPB8E5.02C	YDL124W	YGL086W	YJR015W	YMR027W	YOR307C
2419	2424	SPAPJ696.01C	YDL125C	YGL087C	YJR016C	YMR030W	YOR308C
2425	2430	SPAPYUG7.03C	YDL126C	YGL091C	YJR017C	YMR031C	YOR310C
2431	2436	SPBC106.09	YDL128W	YGL092W	YJR019C	YMR033W	YOR311C
2437	2442	SPBC106.20	YDL129W	YGL093W	YJR022W	YMR035W	YOR315W
2443	2448	SPBC1105.04C	YDL131W	YGL094C	YJR024C	YMR036C	YOR316C
2449	2454	SPBC1105.06	YDL135C	YGL095C	YJR025C	YMR037C	YOR317W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
2455	2460	SPBC11B10.05C	YDL138W	YGL098W	YJR031C	YMR038C	YOR319W
2461	2466	SPBC11C11.03	YDL139C	YGL099W	YJR032W	YMR042W	YOR320C
2467	2472	SPBC11C11.08	YDL140C	YGL100W	YJR033C	YMR043W	YOR322C
2473	2478	SPBC1271.02	YDL141W	YGL101W	YJR039W	YMR044W	YOR323C
2479	2484	SPBC1289.02C	YDL142C	YGL103W	YJR040W	YMR047C	YOR324C
2485	2490	SPBC1289.04C	YDL143W	YGL104C	YJR041C	YMR048W	YOR330C
2491	2496	SPBC1289.07C	YDL144C	YGL105W	YJR042W	YMR049C	YOR334W
2497	2502	SPBC1289.11	YDL145C	YGL107C	YJR043C	YMR052W	YOR335C
2503	2508	SPBC12C2.10C	YDL146W	YGL110C	YJR044C	YMR053C	YOR337W
2509	2514	SPBC12D12.01	YDL147W	YGL111W	YJR045C	YMR054W	YOR342C
2515	2520	SPBC13E7.02	YDL148C	YGL112C	YJR046W	YMR055C	YOR350C
2521	2526	SPBC13E7.09	YDL149W	YGL113W	YJR048W	YMR058W	YOR351C
2527	2532	SPBC13E7.10C	YDL153C	YGL115W	YJR049C	YMR059W	YOR352W
2533	2538	SPBC13G1.08C	YDL154W	YGL119W	YJR050W	YMR060C	YOR353C
2539	2544	SPBC146.07	YDL155W	YGL120C	YJR051W	YMR061W	YOR354C
2545	2550	SPBC146.13C	YDL156W	YGL122C	YJR052W	YMR064W	YOR355W
2551	2556	SPBC14C8.01C	YDL157C	YGL123W	YJR053W	YMR065W	YOR356W
2557	2562	SPBC14C8.12	YDL159W	YGL124C	YJR054W	YMR066W	YOR357C
2563	2568	SPBC14F5.03C	YDL160C	YGL125W	YJR055W	YMR067C	YOR359W
2569	2574	SPBC14F5.08	YDL164C	YGL126W	YJR056C	YMR068W	YOR360C
2575	2580	SPBC15D4.03	YDL165W	YGL127C	YJR057W	YMR069W	YOR361C
2581	2586	SPBC15D4.04	YDL166C	YGL128C	YJR059W	YMR070W	YOR362C
2587	2592	SPBC15D4.10C	YDL167C	YGL129C	YJR060W	YMR071C	YOR363C
2593	2598	SPBC1604.08C	YDL168W	YGL130W	YJR062C	YMR072W	YOR368W
2599	2604	SPBC1604.10	YDL171C	YGL131C	YJR064W	YMR073C	YOR370C
2605	2610	SPBC1604.20C	YDL173W	YGL133W	YJR066W	YMR074C	YOR371C
2611	2616	SPBC1604.21C	YDL174C	YGL136C	YJR067C	YMR075W	YOR372C
2617	2622	SPBC1677.02	YDL175C	YGL137W	YJR069C	YMR076C	YOR373W
2623	2628	SPBC1685.08	YDL178W	YGL139W	YJR070C	YMR077C	YOR374W
2629	2634	SPBC1685.15C	YDL180W	YGL141W	YJR072C	YMR078C	YOR375C
2635	2640	SPBC16A3.01	YDL182W	YGL143C	YJR073C	YMR079W	YOR377W
2641	2646	SPBC16A3.19	YDL185W	YGL145W	YJR074W	YMR080C	YOR381W
2647	2652	SPBC16C6.09	YDL189W	YGL148W	YJR075W	YMR083W	YOR382W
2653	2658	SPBC16H5.05C	YDL190C	YGL150C	YJR077C	YMR086W	YOR383C
2659	2664	SPBC16H5.06	YDL193W	YGL151W	YJR078W	YMR088C	YOR384W
2665	2670	SPBC16H5.07C	YDL194W	YGL153W	YJR080C	YMR089C	YOR385W
2671	2676	SPBC16H5.11C	YDL197C	YGL156W	YJR082C	YMR090W	YOR386W
2677	2682	SPBC1703.06	YDL201W	YGL157W	YJR084W	YMR091C	YOR388C
2683	2688	SPBC1706.01	YDL203C	YGL162W	YJR085C	YMR092C	YPL001W
2689	2694	SPBC1709.04C	YDL204W	YGL163C	YJR086W	YMR093W	YPL002C
2695	2700	SPBC1709.08	YDL205C	YGL164C	YJR088C	YMR095C	YPL004C
2701	2706	SPBC1709.15C	YDL207W	YGL166W	YJR089W	YMR097C	YPL005W
2707	2712	SPBC1709.17	YDL208W	YGL169W	YJR090C	YMR098C	YPL006W
2713	2718	SPBC1718.02	YDL209C	YGL170C	YJR093C	YMR099C	YPL008W
2719	2724	SPBC1718.03	YDL210W	YGL172W	YJR094C	YMR101C	YPL009C
2725	2730	SPBC1718.06	YDL211C	YGL173C	YJR096W	YMR104C	YPL010W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
2731	2736	SPBC1718.07C	YDL212W	YGL174W	YJR097W	YMR108W	YPL011C
2737	2742	SPBC1778.02	YDL213C	YGL175C	YJR098C	YMR110C	YPL012W
2743	2748	SPBC17D11.01	YDL215C	YGL178W	YJR101W	YMR111C	YPL013C
2749	2754	SPBC17F3.01C	YDL216C	YGL179C	YJR102C	YMR112C	YPL014W
2755	2760	SPBC17F3.02	YDL217C	YGL180W	YJR104C	YMR113W	YPL015C
2761	2766	SPBC17G9.04C	YDL218W	YGL181W	YJR105W	YMR114C	YPL017C
2767	2772	SPBC1861.01C	YDL219W	YGL183C	YJR110W	YMR115W	YPL018W
2773	2778	SPBC18H10.06C	YDL220C	YGL184C	YJR111C	YMR116C	YPL019C
2779	2784	SPBC1921.02	YDL222C	YGL185C	YJR113C	YMR117C	YPL020C
2785	2790	SPBC1921.03C	YDL224C	YGL186C	YJR117W	YMR119W	YPL023C
2791	2796	SPBC1921.06C	YDL225W	YGL187C	YJR118C	YMR120C	YPL024W
2797	2802	SPBC19C7.09C	YDL226C	YGL190C	YJR119C	YMR122W-A	YPL026C
2803	2808	SPBC19C7.10	YDL230W	YGL191W	YJR121W	YMR123W	YPL029W
2809	2814	SPBC19G7.05C	YDL232W	YGL194C	YJR122W	YMR124W	YPL030W
2815	2820	SPBC19G7.09	YDL233W	YGL195W	YJR125C	YMR125W	YPL031C
2821	2826	SPBC19G7.15	YDL234C	YGL197W	YJR127C	YMR127C	YPL032C
2827	2832	SPBC1D7.04	YDL236W	YGL200C	YJR129C	YMR128W	YPL036W
2833	2838	SPBC1D7.05	YDL238C	YGL201C	YJR130C	YMR129W	YPL037C
2839	2844	SPBC20F10.01	YDL240W	YGL202W	YJR131W	YMR131C	YPL038W
2845	2850	SPBC21.04	YDL248W	YGL203C	YJR132W	YMR132C	YPL040C
2851	2856	SPBC21.06C	YDR001C	YGL207W	YJR133W	YMR134W	YPL045W
2857	2862	SPBC211.02C	YDR002W	YGL208W	YJR134C	YMR135C	YPL046C
2863	2868	SPBC211.04C	YDR003W	YGL210W	YJR135C	YMR138W	YPL047W
2869	2874	SPBC211.06	YDR004W	YGL211W	YJR135W-A	YMR139W	YPL048W
2875	2880	SPBC216.05	YDR005C	YGL212W	YJR136C	YMR140W	YPL050C
2881	2886	SPBC21B10.04C	YDR006C	YGL213C	YJR137C	YMR144W	YPL051W
2887	2892	SPBC21C3.11	YDR009W	YGL219C	YJR138W	YMR145C	YPL055C
2893	2898	SPBC23G7.16	YDR011W	YGL220W	YJR139C	YMR146C	YPL058C
2899	2904	SPBC244.01C	YDR012W	YGL221C	YJR140C	YMR149W	YPL059W
2905	2910	SPBC24C6.07	YDR013W	YGL222C	YJR144W	YMR150C	YPL060W
2911	2916	SPBC24C6.11	YDR016C	YGL223C	YJR145C	YMR152W	YPL061W
2917	2922	SPBC25D12.03C	YDR017C	YGL225W	YJR147W	YMR153W	YPL063W
2923	2928	SPBC25D12.04	YDR019C	YGL226C-A	YJR148W	YMR157C	YPL064C
2929	2934	SPBC26H8.06	YDR020C	YGL226W	YJR149W	YMR158W	YPL065W
2935	2940	SPBC26H8.10	YDR022C	YGL227W	YJR150C	YMR160W	YPL066W
2941	2946	SPBC27.02C	YDR023W	YGL229C	YJR151C	YMR161W	YPL067C
2947	2952	SPBC28E12.01C	YDR026C	YGL231C	YJR152W	YMR162C	YPL068C
2953	2958	SPBC28F2.02	YDR027C	YGL232W	YJR153W	YMR163C	YPL069C
2959	2964	SPBC28F2.04C	YDR028C	YGL233W	YJR154W	YMR165C	YPL070W
2965	2970	SPBC28F2.07	YDR031W	YGL234W	YJR161C	YMR167W	YPL071C
2971	2976	SPBC29A10.15	YDR032C	YGL236C	YKL002W	YMR168C	YPL074W
2977	2982	SPBC2A9.12	YDR033W	YGL237C	YKL003C	YMR170C	YPL076W
2983	2988	SPBC2D10.12	YDR034C	YGL238W	YKL004W	YMR171C	YPL081W
2989	2994	SPBC2F12.11C	YDR035W	YGL240W	YKL009W	YMR172W	YPL082C
2995	3000	SPBC2F12.13	YDR036C	YGL241W	YKL010C	YMR173W	YPL083C
3001	3006	SPBC2G2.04C	YDR039C	YGL243W	YKL011C	YMR176W	YPL084W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
3007	3012	SPBC2G5.06C	YDR040C	YGL244W	YKL012W	YMR177W	YPL085W
3013	3018	SPBC2G5.07C	YDR041W	YGL245W	YKL014C	YMR178W	YPL086C
3019	3024	SPBC30B4.05	YDR043C	YGL246C	YKL016C	YMR179W	YPL087W
3025	3030	SPBC31E1.02C	YDR045C	YGL247W	YKL018C-A	YMR180C	YPL090C
3031	3036	SPBC31E1.03	YDR046C	YGL249W	YKL018W	YMR183C	YPL091W
3037	3042	SPBC31F10.09C	YDR047W	YGL250W	YKL020C	YMR184W	YPL092W
3043	3048	SPBC31F10.11C	YDR049W	YGL251C	YKL023W	YMR186W	YPL093W
3049	3054	SPBC31F10.13C	YDR050C	YGL252C	YKL024C	YMR188C	YPL094C
3055	3060	SPBC32F12.04	YDR051C	YGL253W	YKL025C	YMR189W	YPL096C-A
3061	3066	SPBC32H8.02C	YDR054C	YGL255W	YKL027W	YMR190C	YPL096W
3067	3072	SPBC32H8.10	YDR055W	YGL256W	YKL029C	YMR192W	YPL097W
3073	3078	SPBC336.08	YDR056C	YGR001C	YKL032C	YMR195W	YPL099C
3079	3084	SPBC336.09C	YDR059C	YGR002C	YKL033W	YMR196W	YPL100W
3085	3090	SPBC354.02C	YDR060W	YGR003W	YKL034W	YMR197C	YPL101W
3091	3096	SPBC354.03	YDR061W	YGR004W	YKL035W	YMR198W	YPL103C
3097	3102	SPBC36.05C	YDR062W	YGR006W	YKL040C	YMR199W	YPL104W
3103	3108	SPBC36.09	YDR063W	YGR007W	YKL041W	YMR200W	YPL105C
3109	3114	SPBC365.06	YDR065W	YGR010W	YKL042W	YMR203W	YPL107W
3115	3120	SPBC365.13C	YDR066C	YGR012W	YKL043W	YMR204C	YPL108W
3121	3126	SPBC365.15	YDR067C	YGR013W	YKL045W	YMR205C	YPL109C
3127	3132	SPBC3B8.09	YDR068W	YGR014W	YKL046C	YMR207C	YPL110C
3133	3138	SPBC3B9.11C	YDR069C	YGR015C	YKL047W	YMR211W	YPL112C
3139	3144	SPBC3B9.16C	YDR070C	YGR017W	YKL049C	YMR212C	YPL116W
3145	3150	SPBC3B9.21	YDR071C	YGR021W	YKL051W	YMR213W	YPL118W
3151	3156	SPBC3D6.04C	YDR074W	YGR024C	YKL052C	YMR214W	YPL119C
3157	3162	SPBC3D6.09	YDR075W	YGR028W	YKL053C-A	YMR215W	YPL121C
3163	3168	SPBC3E7.02C	YDR076W	YGR029W	YKL054C	YMR216C	YPL122C
3169	3174	SPBC4.04C	YDR077W	YGR030C	YKL055C	YMR218C	YPL123C
3175	3180	SPBC409.20C	YDR079C-A	YGR031W	YKL056C	YMR219W	YPL124W
3181	3186	SPBC428.01C	YDR079W	YGR033C	YKL057C	YMR221C	YPL125W
3187	3192	SPBC428.02C	YDR080W	YGR036C	YKL058W	YMR222C	YPL126W
3193	3198	SPBC428.13C	YDR083W	YGR037C	YKL059C	YMR223W	YPL127C
3199	3204	SPBC428.17C	YDR084C	YGR038W	YKL060C	YMR224C	YPL128C
3205	3210	SPBC428.20C	YDR086C	YGR040W	YKL061W	YMR226C	YPL129W
3211	3216	SPBC4C3.07	YDR087C	YGR042W	YKL062W	YMR227C	YPL130W
3217	3222	SPBC4C3.12	YDR088C	YGR043C	YKL063C	YMR229C	YPL132W
3223	3228	SPBC4F6.06	YDR091C	YGR044C	YKL065C	YMR231W	YPL133C
3229	3234	SPBC4F6.16C	YDR092W	YGR046W	YKL067W	YMR232W	YPL134C
3235	3240	SPBC530.04	YDR093W	YGR048W	YKL068W	YMR233W	YPL135W
3241	3246	SPBC530.10C	YDR096W	YGR049W	YKL069W	YMR235C	YPL137C
3247	3252	SPBC530.14C	YDR097C	YGR052W	YKL070W	YMR236W	YPL138C
3253	3258	SPBC582.03	YDR098C	YGR054W	YKL071W	YMR237W	YPL139C
3259	3264	SPBC582.06C	YDR099W	YGR056W	YKL073W	YMR238W	YPL141C
3265	3270	SPBC646.04	YDR100W	YGR057C	YKL074C	YMR239C	YPL144W
3271	3276	SPBC646.05C	YDR101C	YGR058W	YKL075C	YMR240C	YPL145C
3277	3282	SPBC646.06C	YDR103W	YGR060W	YKL077W	YMR241W	YPL146C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
3283	3288	SPBC646.09C	YDR104C	YGR061C	YKL078W	YMR243C	YPL147W
3289	3294	SPBC646.14C	YDR105C	YGR062C	YKL081W	YMR244C-A	YPL148C
3295	3300	SPBC649.05	YDR106W	YGR063C	YKL082C	YMR246W	YPL149W
3301	3306	SPBC651.05C	YDR108W	YGR065C	YKL084W	YMR250W	YPL151C
3307	3312	SPBC685.09	YDR110W	YGR071C	YKL086W	YMR251W-A	YPL153C
3313	3318	SPBC6B1.09C	YDR111C	YGR072W	YKL087C	YMR252C	YPL154C
3319	3324	SPBC713.02C	YDR113C	YGR074W	YKL088W	YMR253C	YPL155C
3325	3330	SPBC725.17C	YDR116C	YGR075C	YKL091C	YMR255W	YPL157W
3331	3336	SPBC776.02C	YDR119W	YGR076C	YKL094W	YMR256C	YPL160W
3337	3342	SPBC776.13	YDR120C	YGR077C	YKL095W	YMR257C	YPL161C
3343	3348	SPBC800.03	YDR121W	YGR078C	YKL096W	YMR258C	YPL162C
3349	3354	SPBC800.07C	YDR122W	YGR081C	YKL096W-A	YMR259C	YPL163C
3355	3360	SPBC800.13	YDR123C	YGR083C	YKL099C	YMR263W	YPL164C
3361	3366	SPBC83.03C	YDR125C	YGR084C	YKL101W	YMR264W	YPL166W
3367	3372	SPBC83.04	YDR127W	YGR086C	YKL103C	YMR267W	YPL167C
3373	3378	SPBC83.07	YDR128W	YGR087C	YKL106W	YMR268C	YPL168W
3379	3384	SPBC887.14C	YDR130C	YGR090W	YKL112W	YMR269W	YPL169C
3385	3390	SPBC8D2.05C	YDR132C	YGR091W	YKL113C	YMR271C	YPL173W
3391	3396	SPBC8D2.10C	YDR134C	YGR092W	YKL114C	YMR272C	YPL174C
3397	3402	SPBC902.04	YDR135C	YGR093W	YKL116C	YMR273C	YPL175W
3403	3408	SPBC902.06	YDR138W	YGR094W	YKL117W	YMR274C	YPL176C
3409	3414	SPBC947.12	YDR139C	YGR095C	YKL119C	YMR275C	YPL178W
3415	3420	SPBC9B6.10	YDR140W	YGR096W	YKL120W	YMR277W	YPL179W
3421	3426	SPBP19A11.06	YDR141C	YGR097W	YKL122C	YMR278W	YPL180W
3427	3432	SPBP22H7.07	YDR142C	YGR098C	YKL126W	YMR282C	YPL183C
3433	3438	SPBP23A10.04	YDR143C	YGR099W	YKL128C	YMR283C	YPL184C
3439	3444	SPBP23A10.08	YDR144C	YGR100W	YKL130C	YMR284W	YPL186C
3445	3450	SPBP23A10.13	YDR145W	YGR102C	YKL132C	YMR287C	YPL188W
3451	3456	SPBP26C9.02C	YDR146C	YGR103W	YKL134C	YMR288W	YPL190C
3457	3462	SPBP35G2.05C	YDR147W	YGR104C	YKL135C	YMR289W	YPL191C
3463	3468	SPBP35G2.06C	YDR148C	YGR105W	YKL138C	YMR290C	YPL192C
3469	3474	SPBP4H10.03	YDR150W	YGR106C	YKL139W	YMR291W	YPL193W
3475	3480	SPBP4H10.06C	YDR152W	YGR108W	YKL140W	YMR292W	YPL194W
3481	3486	SPBP4H10.20	YDR153C	YGR111W	YKL143W	YMR293C	YPL196W
3487	3492	SPBP8B7.23	YDR155C	YGR112W	YKL145W	YMR296C	YPL199C
3493	3498	SPCC1020.02	YDR156W	YGR113W	YKL146W	YMR297W	YPL200W
3499	3504	SPCC1020.04C	YDR158W	YGR116W	YKL150W	YMR298W	YPL202C
3505	3510	SPCC1183.06	YDR159W	YGR117C	YKL151C	YMR300C	YPL204W
3511	3516	SPCC11E10.02C	YDR160W	YGR119C	YKL152C	YMR301C	YPL206C
3517	3522	SPCC11E10.05C	YDR161W	YGR120C	YKL154W	YMR303C	YPL207W
3523	3528	SPCC1223.06	YDR162C	YGR121C	YKL155C	YMR305C	YPL208W
3529	3534	SPCC1235.06	YDR163W	YGR122W	YKL156W	YMR306W	YPL209C
3535	3540	SPCC1235.10C	YDR164C	YGR123C	YKL157W	YMR307W	YPL210C
3541	3546	SPCC126.03	YDR165W	YGR124W	YKL160W	YMR308C	YPL212C
3547	3552	SPCC1322.06	YDR166C	YGR125W	YKL162C	YMR309C	YPL213W
3553	3558	SPCC1322.08	YDR167W	YGR126W	YKL163W	YMR310C	YPL214C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
3559	3564	SPCC1322.12C	YDR168W	YGR128C	YKL164C	YMR311C	YPL217C
3565	3570	SPCC1442.10C	YDR169C	YGR129W	YKL165C	YMR312W	YPL221W
3571	3576	SPCC1450.05C	YDR170C	YGR130C	YKL168C	YMR313C	YPL222W
3577	3582	SPCC1450.06C	YDR171W	YGR132C	YKL170W	YMR314W	YPL223C
3583	3588	SPCC1450.14C	YDR172W	YGR133W	YKL171W	YMR315W	YPL224C
3589	3594	SPCC1494.01	YDR173C	YGR134W	YKL172W	YMR316W	YPL225W
3595	3600	SPCC162.08C	YDR174W	YGR135W	YKL173W	YMR319C	YPL226W
3601	3606	SPCC1672.08C	YDR175C	YGR136W	YKL174C	YMR323W	YPL227C
3607	3612	SPCC1672.10	YDR176W	YGR138C	YKL175W	YNL001W	YPL228W
3613	3618	SPCC1682.02C	YDR177W	YGR140W	YKL176C	YNL002C	YPL229W
3619	3624	SPCC1682.04	YDR179C	YGR142W	YKL178C	YNL003C	YPL231W
3625	3630	SPCC16A11.17	YDR181C	YGR144W	YKL179C	YNL004W	YPL235W
3631	3636	SPCC16C4.07	YDR183W	YGR145W	YKL180W	YNL005C	YPL236C
3637	3642	SPCC16C4.08C	YDR184C	YGR147C	YKL181W	YNL006W	YPL237W
3643	3648	SPCC16C4.09	YDR185C	YGR150C	YKL182W	YNL007C	YPL239W
3649	3654	SPCC16C4.18C	YDR186C	YGR152C	YKL183W	YNL008C	YPL240C
3655	3660	SPCC1739.03	YDR188W	YGR154C	YKL184W	YNL009W	YPL243W
3661	3666	SPCC1739.11C	YDR189W	YGR155W	YKL185W	YNL010W	YPL245W
3667	3672	SPCC1739.12	YDR190C	YGR156W	YKL186C	YNL012W	YPL246C
3673	3678	SPCC1739.14	YDR191W	YGR158C	YKL187C	YNL014W	YPL247C
3679	3684	SPCC1753.03C	YDR192C	YGR159C	YKL188C	YNL015W	YPL249C
3685	3690	SPCC1795.01C	YDR194C	YGR161C	YKL191W	YNL016W	YPL252C
3691	3696	SPCC1795.03	YDR195W	YGR162W	YKL192C	YNL021W	YPL253C
3697	3702	SPCC1795.10C	YDR196C	YGR163W	YKL193C	YNL022C	YPL254W
3703	3708	SPCC1795.11	YDR200C	YGR165W	YKL194C	YNL024C	YPL255W
3709	3714	SPCC18.04	YDR201W	YGR166W	YKL195W	YNL026W	YPL256C
3715	3720	SPCC18.11C	YDR202C	YGR169C	YKL196C	YNL027W	YPL259C
3721	3726	SPCC1840.01C	YDR204W	YGR171C	YKL197C	YNL032W	YPL260W
3727	3732	SPCC1840.02C	YDR205W	YGR172C	YKL203C	YNL035C	YPL262W
3733	3738	SPCC1840.03	YDR207C	YGR173W	YKL205W	YNL036W	YPL263C
3739	3744	SPCC188.03	YDR208W	YGR175C	YKL206C	YNL037C	YPL268W
3745	3750	SPCC188.04C	YDR211W	YGR177C	YKL208W	YNL040W	YPL269W
3751	3756	SPCC188.07	YDR212W	YGR178C	YKL210W	YNL041C	YPL270W
3757	3762	SPCC18B5.03	YDR213W	YGR179C	YKL211C	YNL042W	YPL271W
3763	3768	SPCC18B5.07C	YDR214W	YGR180C	YKL212W	YNL044W	YPL273W
3769	3774	SPCC1906.01	YDR216W	YGR181W	YKL213C	YNL045W	YPL274W
3775	3780	SPCC191.09C	YDR221W	YGR183C	YKL214C	YNL046W	YPL283C
3781	3786	SPCC191.11	YDR222W	YGR184C	YKL215C	YNL047C	YPR002W
3787	3792	SPCC1919.03C	YDR223W	YGR185C	YKL216W	YNL048W	YPR003C
3793	3798	SPCC1919.10C	YDR224C	YGR187C	YKL217W	YNL049C	YPR004C
3799	3804	SPCC290.03C	YDR227W	YGR188C	YKL220C	YNL051W	YPR005C
3805	3810	SPCC297.03	YDR228C	YGR189C	YKR001C	YNL052W	YPR006C
3811	3816	SPCC306.03C	YDR229W	YGR191W	YKR002W	YNL053W	YPR007C
3817	3822	SPCC306.04C	YDR231C	YGR192C	YKR007W	YNL054W	YPR008W
3823	3828	SPCC330.08	YDR232W	YGR193C	YKR008W	YNL055C	YPR009W
3829	3834	SPCC330.10	YDR233C	YGR194C	YKR010C	YNL056W	YPR016C

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
3835	3840	SPCC417.07C	YDR234W	YGR195W	YKR011C	YNL058C	YPR018W
3841	3846	SPCC4B3.07	YDR235W	YGR196C	YKR013W	YNL059C	YPR019W
3847	3852	SPCC4B3.15	YDR236C	YGR198W	YKR014C	YNL062C	YPR020W
3853	3858	SPCC4G3.19	YDR238C	YGR200C	YKR016W	YNL063W	YPR022C
3859	3864	SPCC550.02C	YDR239C	YGR202C	YKR018C	YNL064C	YPR023C
3865	3870	SPCC550.12	YDR240C	YGR203W	YKR019C	YNL065W	YPR024W
3871	3876	SPCC594.05C	YDR243C	YGR204W	YKR020W	YNL066W	YPR025C
3877	3882	SPCC5E4.03C	YDR244W	YGR205W	YKR021W	YNL068C	YPR026W
3883	3888	SPCC613.12C	YDR245W	YGR206W	YKR022C	YNL070W	YPR029C
3889	3894	SPCC622.16C	YDR246W	YGR207C	YKR023W	YNL071W	YPR030W
3895	3900	SPCC663.01C	YDR247W	YGR208W	YKR024C	YNL072W	YPR031W
3901	3906	SPCC663.12	YDR248C	YGR209C	YKR026C	YNL073W	YPR032W
3907	3912	SPCC736.04C	YDR252W	YGR210C	YKR027W	YNL074C	YPR033C
3913	3918	SPCC736.11	YDR253C	YGR211W	YKR028W	YNL075W	YPR034W
3919	3924	SPCC736.14	YDR254W	YGR212W	YKR029C	YNL077W	YPR035W
3925	3930	SPCC737.09C	YDR255C	YGR215W	YKR030W	YNL078W	YPR040W
3931	3936	SPCC74.02C	YDR256C	YGR217W	YKR031C	YNL081C	YPR041W
3937	3942	SPCC74.06	YDR257C	YGR218W	YKR034W	YNL085W	YPR042C
3943	3948	SPCC825.03C	YDR258C	YGR222W	YKR035W-A	YNL086W	YPR045C
3949	3954	SPCC830.03	YDR259C	YGR223C	YKR036C	YNL088W	YPR046W
3955	3960	SPCC895.05	YDR260C	YGR224W	YKR037C	YNL091W	YPR047W
3961	3966	SPCC895.07	YDR261C	YGR225W	YKR038C	YNL093W	YPR052C
3967	3972	SPCC962.02C	YDR262W	YGR227W	YKR039W	YNL096C	YPR054W
3973	3978	SPCC962.03C	YDR263C	YGR229C	YKR041W	YNL097C	YPR055W
3979	3984	SPCC962.06C	YDR266C	YGR231C	YKR042W	YNL098C	YPR056W
3985	3990	SPCC965.07C	YDR267C	YGR232W	YKR043C	YNL099C	YPR057W
3991	3996	SPCC970.04C	YDR268W	YGR233C	YKR044W	YNL100W	YPR060C
3997	4002	SPCC970.07C	YDR270W	YGR234W	YKR045C	YNL101W	YPR061C
4003	4008	SPCC970.09	YDR272W	YGR235C	YKR046C	YNL103W	YPR062W
4009	4014	SPCP1E11.04C	YDR275W	YGR236C	YKR048C	YNL104C	YPR063C
4015	4020	SPCP1E11.07C	YDR276C	YGR237C	YKR049C	YNL107W	YPR065W
4021	4026	SPCP31B10.03C	YDR279W	YGR239C	YKR052C	YNL108C	YPR066W
4027	4032	SPCPJ732.01	YDR280W	YGR240C	YKR053C	YNL110C	YPR067W
4033	4038	YAL001C	YDR281C	YGR243W	YKR059W	YNL111C	YPR069C
4039	4044	YAL003W	YDR283C	YGR244C	YKR060W	YNL112W	YPR070W
4045	4050	YAL005C	YDR284C	YGR245C	YKR061W	YNL115C	YPR072W
4051	4056	YAL007C	YDR285W	YGR247W	YKR063C	YNL116W	YPR073C
4057	4062	YAL008W	YDR288W	YGR248W	YKR064W	YNL117W	YPR075C
4063	4068	YAL010C	YDR291W	YGR250C	YKR065C	YNL118C	YPR079W
4069	4074	YAL011W	YDR292C	YGR251W	YKR066C	YNL119W	YPR081C
4075	4080	YAL012W	YDR293C	YGR252W	YKR067W	YNL122C	YPR082C
4081	4086	YAL015C	YDR294C	YGR253C	YKR068C	YNL123W	YPR086W
4087	4092	YAL016W	YDR295C	YGR254W	YKR070W	YNL125C	YPR088C
4093	4098	YAL017W	YDR296W	YGR255C	YKR071C	YNL126W	YPR091C
4099	4104	YAL019W	YDR298C	YGR257C	YKR072C	YNL129W	YPR093C
4105	4110	YAL020C	YDR299W	YGR260W	YKR074W	YNL132W	YPR094W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
4111	4116	YAL021C	YDR300C	YGR262C	YKR075C	YNL133C	YPR095C
4117	4122	YAL025C	YDR301W	YGR263C	YKR076W	YNL134C	YPR097W
4123	4128	YAL026C	YDR303C	YGR264C	YKR077W	YNL135C	YPR098C
4129	4134	YAL027W	YDR304C	YGR266W	YKR078W	YNL136W	YPR100W
4135	4140	YAL028W	YDR305C	YGR267C	YKR079C	YNL137C	YPR101W
4141	4146	YAL029C	YDR307W	YGR268C	YKR080W	YNL138W-A	YPR103W
4147	4152	YAL030W	YDR308C	YGR270W	YKR081C	YNL139C	YPR104C
4153	4158	YAL031C	YDR310C	YGR271C-A	YKR082W	YNL141W	YPR105C
4159	4164	YAL032C	YDR311W	YGR271W	YKR083C	YNL142W	YPR107C
4165	4170	YAL033W	YDR312W	YGR272C	YKR084C	YNL144C	YPR108W
4171	4176	YAL035W	YDR313C	YGR274C	YKR085C	YNL145W	YPR111W
4177	4182	YAL036C	YDR314C	YGR275W	YKR087C	YNL146W	YPR112C
4183	4188	YAL039C	YDR315C	YGR276C	YKR088C	YNL147W	YPR113W
4189	4194	YAL040C	YDR316W	YGR277C	YKR089C	YNL148C	YPR114W
4195	4200	YAL041W	YDR318W	YGR278W	YKR090W	YNL149C	YPR115W
4201	4206	YAL042W	YDR320C	YGR279C	YKR091W	YNL152W	YPR116W
4207	4212	YAL043C	YDR322C-A	YGR281W	YKR092C	YNL153C	YPR118W
4213	4218	YAL044C	YDR322W	YGR283C	YKR093W	YNL154C	YPR119W
4219	4224	YAL047C	YDR323C	YGR284C	YKR094C	YNL155W	YPR120C
4225	4230	YAL048C	YDR324C	YGR285C	YKR095W	YNL156C	YPR124W
4231	4236	YAL049C	YDR325W	YGR286C	YKR096W	YNL157W	YPR125W
4237	4242	YAL051W	YDR328C	YGR287C	YKR097W	YNL158W	YPR127W
4243	4248	YAL053W	YDR329C	YGR295C	YKR099W	YNL161W	YPR128C
4249	4254	YAL054C	YDR330W	YHL002W	YKR101W	YNL162W-A	YPR129W
4255	4260	YAL056W	YDR331W	YHL003C	YLL001W	YNL163C	YPR131C
4261	4266	YAL058W	YDR332W	YHL004W	YLL002W	YNL167C	YPR133C
4267	4272	YAL059W	YDR333C	YHL006C	YLL003W	YNL168C	YPR133W-A
4273	4278	YAL060W	YDR334W	YHL008C	YLL006W	YNL169C	YPR134W
4279	4284	YAL061W	YDR335W	YHL009C	YLL007C	YNL173C	YPR135W
4285	4290	YAL062W	YDR337W	YHL011C	YLL009C	YNL175C	YPR137W
4291	4296	YAR002C-A	YDR341C	YHL013C	YLL010C	YNL176C	YPR138C
4297	4302	YAR003W	YDR342C	YHL014C	YLL011W	YNL177C	YPR139C
4303	4308	YAR007C	YDR343C	YHL016C	YLL012W	YNL180C	YPR140W
4309	4314	YAR008W	YDR346C	YHL017W	YLL013C	YNL181W	YPR141C
4315	4320	YAR015W	YDR347W	YHL018W	YLL014W	YNL182C	YPR143W
4321	4326	YAR019C	YDR349C	YHL019C	YLL015W	YNL186W	YPR144C
4327	4332	YAR027W	YDR350C	YHL021C	YLL018C	YNL187W	YPR145W
4333	4338	YAR029W	YDR351W	YHL022C	YLL018C-A	YNL188W	YPR147C
4339	4344	YAR031W	YDR352W	YHL024W	YLL022C	YNL189W	YPR148C
4345	4350	YAR033W	YDR354W	YHL028W	YLL023C	YNL190W	YPR149W
4351	4356	YAR035W	YDR356W	YHL029C	YLL024C	YNL191W	YPR151C
4357	4362	YAR042W	YDR357C	YHL030W	YLL026W	YNL192W	YPR152C
4363	4368	YAR050W	YDR358W	YHL032C	YLL027W	YNL194C	YPR154W
4369	4374	YAR071W	YDR359C	YHL034C	YLL028W	YNL195C	YPR155C
4375	4380	YBL001C	YDR361C	YHL035C	YLL029W	YNL197C	YPR156C
4381	4386	YBL004W	YDR363W	YHL038C	YLL032C	YNL200C	YPR159W

Count (From)	Count (To)	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID	Protein ID
4387	4392	YBL005W	YDR363W-A	YHL039W	YLL034C	YNL201C	YPR161C
4393	4398	YBL008W	YDR364C	YHL040C	YLL036C	YNL202W	YPR165W
4399	4404	YBL010C	YDR365C	YHL044W	YLL040C	YNL206C	YPR166C
4405	4410	YBL011W	YDR367W	YHL048W	YLL043W	YNL207W	YPR168W
4411	4416	YBL013W	YDR368W	YHR001W-A	YLL046C	YNL208W	YPR169W
4417	4422	YBL015W	YDR369C	YHR002W	YLL048C	YNL209W	YPR172W
4423	4428	YBL016W	YDR370C	YHR003C	YLL051C	YNL210W	YPR173C
4429	4434	YBL018C	YDR371W	YHR004C	YLL052C	YNL211C	YPR174C
4435	4440	YBL020W	YDR372C	YHR005C	YLL055W	YNL212W	YPR175W
4441	4446	YBL022C	YDR375C	YHR005C-A	YLL060C	YNL214W	YPR178W
4447	4452	YBL023C	YDR376W	YHR006W	YLL061W	YNL215W	YPR179C
4453	4458	YBL024W	YDR377W	YHR007C	YLL062C	YNL216W	YPR180W
4459	4464	YBL026W	YDR378C	YHR008C	YLR001C	YNL217W	YPR181C
4465	4470	YBL028C	YDR379C-A	YHR009C	YLR002C	YNL219C	YPR182W
4471	4476	YBL029W	YDR380W	YHR011W	YLR003C	YNL220W	YPR183W
4477	4482	YBL030C	YDR381W	YHR012W	YLR005W	YNL221C	YPR184W
4483	4488	YBL032W	YDR384C	YHR013C	YLR007W	YNL222W	YPR186C
4489	4494	YBL033C	YDR386W	YHR014W	YLR008C	YNL224C	YPR189W
4495	4500	YBL034C	YDR390C	YHR015W	YLR011W	YNL225C	YPR191W
4501	4506	YBL035C	YDR391C	YHR017W	YLR015W	YNL227C	YPR192W
4507	4512	YBL040C	YDR392W	YHR018C	YLR016C	YNL229C	YPR193C
4513	4518	YBL041W	YDR393W	YHR024C	YLR017W	YNL231C	YPR198W
4519	4524	YBL042C	YDR394W	YHR027C	YLR018C	YNL232W	YPR199C
4525	4529	YBL045C	YDR395W	YHR028C	YLR020C	YNL233W	

Table-A5: Content of the data set used for system evaluation