

**Computational Classification of Protein Subcellular Localization using Pattern Discovery
Methods**

Thesis proposal

for the degree of Master of Science in Computer Science

at Southern Connecticut State University

Nnamdi Ihuegbu

January 2005

Thesis Advisor: Dr.Taraneh Seyed

Major-Field Approval – The advisor and the department chairperson

Advisor

Date

Chairperson

Date

A. Title

Computational Classification of Protein Subcellular Localization using Pattern Discovery Methods.

B. Statement of Purpose

The purpose of this endeavor is two-fold: given the acquired dataset of proteins with known subcellular localization, to collect conserved patterns within the primary structure of proteins that may be responsible for localization (i.e. isolate localization ‘signals’); And to demonstrate how these repository of patterns within the amino acid sequence can be used to build a classifier that disaggregates arbitrary proteins into their correct subcellular localization.

C. Introduction

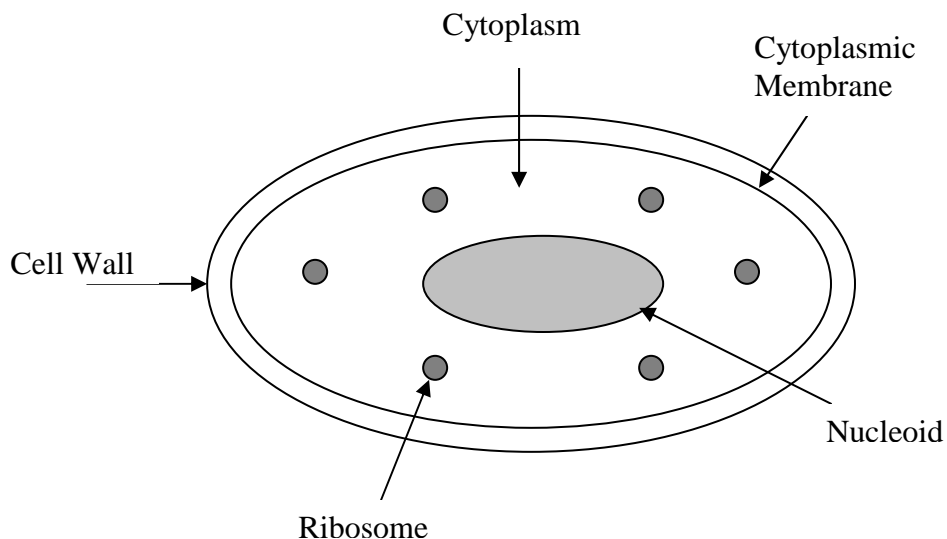
There is much effort to understand the parameters that influence the function of proteins. Fully understanding these functions have obvious ramifications in pursuits such as the mapping of disorders to responsible molecular and genomic defects, drug design and drug discovery.

To accomplish this, the structure and components of defective and responsive proteins are investigated. The primary structure of proteins are chains of amino acids sequences. There are twenty different amino acids that can be linked covalently by peptide bonds. The DNA in the cell is replicated and transcribed into RNA. Using three RNA bases at a time (i.e. codons), the RNA sequence is translated into each amino acid resulting in a chain of amino acids. Different combinations of linked amino acids confer characteristic properties and functions to the resulting protein.

Proteins can be attributed to every biological function of an organism. Some of these functions include:

- Motion and locomotion of cells and organisms. This function depends on contractile proteins such as muscles.
- The catalysis of all biochemical reactions. This function depends on enzymes which are made up by proteins.
- The encasing of cells and organelles in cell walls and membranes. These walls and membranes are largely made molecules such as Collagens that are made up by proteins.
- The transport of materials in body fluids depends of proteins that constitute blood.
- The receptors for hormones and other signaling molecules are made from proteins.
- The transcription factors that regulate the expression of genes are macromolecules that are constituted by proteins.

In addition to proteins, the cell structure is also investigated to study the effects of intervening therapies in the host system. Prokaryotes and Eukaryotes are amongst the major classifications of organisms that have significantly different cell structures. Prokaryotic organisms (such as bacteria), originated earlier, are usually smaller in size, and have the simpler cell structure. Like Eukaryotes, Prokaryotes have DNA as the basis of their genetic make-up, their cell structures are both enveloped by a membrane (a bi-lipid layer), and they both have ribosomes (the location for protein synthesis). Additionally, Eukaryotic cells contain a nucleus and membrane-bound organelles (small, dedicated structures within the cell that perform specific functions). Consequently, the genetic make-up (i.e. the DNA) in Eukaryotes tend to be more complex and extensive than in prokaryotes. This difference in complexity is also observed in the activity of the resulting proteins that are coded for in the DNA. The following diagram illustrates the common principal components in the structure of cells.



One of the major parameters responsible for protein function and overall cell stability is the localization of proteins. Protein Subcellular Localization (PSL) studies which compartments or conduits of the cell proteins are transported to in order to accomplish their function. For proteins whose function and activity pathways still remain intractable, knowing how they are transported in the cell and what local they eventually reside in is instrumental in predicting their function. Prediction of protein function is made possible since many of the proteins that localize in a specific location either have similar functions or contribute to a general function (Jenson et. al, 2002). Currently, manual determination of localization involves: cell fractionation, electron microscopy, and fluorescence microscopy, all of which constitute to a time consuming, subjective, and highly variable enterprise (Murphy, 2000). Automation and more accurate assessment of protein subcellular localization are needed if this branch of knowledge is to be feasibly incorporated into drug discovery and other like efforts in need.

Different proteins are localized in different compartments of the cell partly due to differences in their protein sequence. It has been shown that subsequences of amino acids in the protein sequence help direct and traffic proteins through the various cell conduits (Nakai, 2000). These subsequences,

known as sorting signals, participate in the localization of proteins at almost every subcellular location. The set of signals include: nuclear localization signals (i.e. signals that determine whether the protein would be localized in the nucleus), sorting signals associated with localization in the plasma membrane from the endoplasmic reticulum, and the endocytosis of proteins, N-Terminal signals—signals that determine locations in secretory pathway in eukaryotic cells (including Mitochondrial targeting peptides—mTP, chloroplast transit peptides—cTP, and signal peptides—SP). These signals are usually carried in the leader portion of the protein, in a transient region that is a precursor to the mature protein, called the preprotein.

Apart from subtle differences in polarity and hydrophobicity, existing studies show that there is very little difference in preproteins (Lewin, 1999). This apparent lack of significant difference in preproteins suggests that secondary and tertiary structural components of the protein are involved in the recognition of signals and their sorting. Yet, experiments show that the preprotein (signal-containing portion) encodes the necessary information to effect specific localization, and very little deviation—such as an extra dilucine motif—can alter this localization (Miranda et. al, 2001).

D. Present State of Knowledge & Existing Solutions

There are currently three main approaches to solve the Protein Subcellular Localization Problem. One approach, based on amino acid composition, involves utilizing biochemical properties (such as hydrophobicity) of each amino acid and recoding the sequence to reflect these properties (Bannai et al, 2002). Using the converted sequence of numbers, computational methods are designed to extract especial biochemical features in the sequences that are consistent with the properties at certain subcellular boundaries. Machine learning techniques such as Support Vector Machines and Neural Networks are then employed to classify the localization (Yu et al, 2004; Hua and Sun, 2001).

This approach has been successfully implemented in the prediction and classification of proteins bound to certain membranes. Yet, there is an obvious drop-off in performance when using this method to predict other types of localizations whose boundaries may not be well characterized. Furthermore, this method may not capture other vital information from global inter-relationships of amino acids (Lei and Dai, 2004).

A second approach is based on identifying protein subsequences approximately 3 to 70 amino acids (i.e. sorting signals) that direct proteins to specific locations (Yu, et. al, 2003). This approach entails detecting the exact portion of the protein sequence that may contain sorting signals, and classifying the localization of the protein based on localizations of proteins with this signal. Yet, the lack of homology in the primary sequence of signals and the incomplete known set of signals for specific localizations hamper general implementations of this method (Lewin, 1999).

Regardless, there have been many attempts to show homology and sequence conservation within protein sequences of known localizations. This effort represents the third approach, sequence homology. It generally entails looking for certain similarities between protein sequences and classifying other proteins based on these similarities. Studies have shown that sequences at a particular localization have more conserved regions than those at other localization and this conservation is generally more than expected (Nair and Rost, 2002).

Existing results suggest that no one method yields the highest accuracy for all localizations and that a combination or hybrid of these methods may be most beneficial. This observation is the major impetus for this project's adopted plan.

E. Research Plan

Using the training dataset, this project will generate and build a repository of the patterns contained in the proteins of a certain class. These conserved patterns would imply a base set which includes the specific signal. Instead of recoding the sequence to represent a biochemical scale and using a feature extraction method to remove seemingly redundant and insignificant patterns, in this method, *all* the peptide-patterns will be used to compare and classify the localization of proteins in the test dataset.

Despite this intention to *blindly* classify the proteins solely based on the contents of their sequence without reliance on any previous biochemical profile, there are obvious benefits to incorporating known biochemical attributes to the classification (as seen in the high accuracy when predicting membrane-bound proteins due to their well-established hydrophobic profile—Bannai et al, 2002; Lei and Dai, 2004). This glaring tension: to remain naïve so as not to miss any inconspicuous, but distinguishing pattern and at the same time improve accuracy and efficiency by using prior attributes, is a very important matter that has to be first resolved before progressing.

To this end, another layer of distinction would be added to the aforementioned naïve approach. For proteins whose final localization is in contention due to the presence of patterns from multiple classes of compartments, a pseudo-composition of the protein's polarity and hydrophobicity would be used as the deciding factor. The frequency pattern of these biochemical attributes with the most resemblance to the arbitrary protein's recoded sequence would designate the localization.

F. Methodology

The proposed solution can be summarized in four general steps: (1) Collect the amino acid sequences of proteins with known localizations into separate classes by localization. (2) Generate

patterns from the known sequences by class and store in a repository. Common patterns in each class may suggest markers in the protein code that lead to specific-site localizations. (3) Using the sequence-patterns as potential markers, other proteins of unknown localizations can be classified into specific subcellular compartments or conduits. (4) For protein's with a majority of patterns in multiple classes, recode for hydrophobicity, compute the Fourier Transform and compare results to the Fourier results from the classes.

Patterns in the training dataset would be extracted based on their frequency of occurrence and organized according to localization. The classification process would involve checking and counting the presence of certain signals against the established repository and determining which class has the most pattern-hits. If two or more classes have the same number of hits, then the hydrophobic Fourier analysis of the member sequences in the competing classes would be assessed to determine the most similar to the sequence in question.

It is envisaged that the entire system would be designed using a Windows-based graphical platform and, after initial successful implementations, made available as a web service.

The results of the project will be tested in a 5-fold cross-validation with PSORT-B (Gardy et al., 2003), the FFT-based approach by Lei and Dai (Lei and Dai, 2004), and CELLO by Yu et al, 2004. As of when this research began, Yu et al claim that CELLO's overall prediction accuracy of 89% is the highest ever reported. The 5-fold cross-validation will consist of predicting localization of proteins at five different subcellular compartments: cytoplasmic, inner membrane, periplasmic, outer membrane, and extra cellular. Predictive statistics such as the precision, recall, and Matthew's Correlation Coefficient will be used to compare the results of all four methods.

G. Dataset

The PSORT database holds 1302 proteins from Gram-negative bacteria that have experimentally-confirmed single localization sites: 248 cytoplasmic, 268 inner membrane, 244 periplasmic, 352 outer membrane, and 190 extra cellular. Two thirds of this dataset would be used to train the designed solution, and one third would be used for testing.

H. Contributions

There are multiple potential benefits from the successfully implementation of the aforementioned solution to the Protein Subcellular Localization problem. Accurate results from this endeavor will demonstrate the benefits of Pattern Recognition as an additional tool to solve the Protein Subcellular Localization problem. This additional perspective can be used to confirm results or reveal undiscovered signaling motifs that may have been missed by other methods.

Furthermore, since finding the localization of proteins is a common task in many biomedical and biopharmaceutical investigations, the results of this endeavor can be far-reaching. Classifying localization of a protein from its sequence is frequently conducted in studies that seek to discover molecular reasons for disorders and possible therapeutic solutions. For instance, in several forms of pancreatic cancer, researchers have found an excessive localization of Ras proteins in the membrane (Zhou et al, 2003). These proteins are member oncogenes of a cascade of enzymes that stimulate cell proliferation. The excessive localization stimulates abnormal cell proliferation in the organ. In these studies, the solution proposed here may lead to accurate classifications of proteins localized in the membrane from their sequence and enable the efficient development of genetic and molecular therapeutic options.

Finally, a successful implementation may support the adoption of more pattern-based recognition systems to classify the localization of proteins in much more complex eukaryotic organisms.

References

- Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., Miyano, S. (2002). Extensive Feature Detection of N-Terminal Protein Sorting Signals. *Bioinformatics* **18**: 298-305.
- Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., et al. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acid Res.* **31**: 3613-3617.
- GlaxoSmithKline (2004). *Genetics at GlaxoSmithKline: Genes and Drug Development*. Source retrieved August, 2004 from <http://genetics.gsk.com/role.htm>
- Hua, S, and Sun, Z. (2001). Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17**: 721-728.
- Jensen, L.J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H.H., Rapacki, K., Workman, C., et al. (2002). Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.* **319**: 1257-1265.
- Lewin, B. (1999). *Genes VII*. Oxford University Press.
- Miranda, K.C., Khromykh, T., Christy, P., Le, T.L., Gottardi, C.J., Yap, A.S., Stow, J.L., Teasdale, R.H. (2001). A Dileucine Motif Targets E-cadherin to the Basolateral Cell Surface in Madin-Darby Canine Kidney and LLC-PK1 Epithelial Cells. *J. Biol. Chem.* **276**: 22565-22572.
- Murphy, R.F., Boland, M.V. and Velliste, M. (2000). Towards A systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Bio.* **8**, 251-259.

Nair, R. and Rost, B. (2002). Sequence conserved for subcellular localization. *Protein Science*. **11**: 2836-2847.

Nakai, K. (2000). Protein sorting signals and prediction for subcellular localization. *Adv. Protein Chem.* **54**: 277-344.

Lei, Z. and Dai, Y. (2004). A Novel Approach for Prediction of Protein Subcellular Localization from Sequence Using Fourier Analysis and Support Vector Machines. *BIOKDD04: 4th Workshop on Data Mining in Bioinformatics (with SIGKDD Conference)*.

Yu, C.S., Lin, C.J., and Hwang, J.K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions. *Protein Science* **13**: 1402-1406.

Zhou, J., Zhu, X., Pan, Q., Liao, D., Li, Z., Liu, Z. (2003). Manumycin inhibits cell proliferation and the Ras signal transduction pathway in human hepatocellular carcinoma cells. *International Journal of Molecular Medicine* **11**: 767-771.