Universidad Nacional de Colombia
Sede Manizales

Master's Thesis

Methodology for predicting semantic annotations of protein sequences by feature extraction derived of statistical contact potentials and continuous wavelet transform

Author:
Gustavo Alonso Arango Argoty

Supervisor:
Dr. Cesar German Castellanos Dominguez

A thesis submitted in fulfillment of the requirements for the degree of Master's on Engineering - Industrial Automation

in the

Department of Electronic, Electric Engineering and Computation
Signal Processing and Recognition Group

June 2014

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MANIZALES

TESIS DE MAESTRÍA

Metodología para predecir la anotación semántica de proteínas por medio de extracción de características derivadas de potenciales de contacto y transformada wavelet continua

Autor:
GUSTAVO ALONSO ARANGO ARGOTY

Tutor:
DR. CESAR GERMAN CASTELLANOS DOMINGUEZ

Tesis presentada en cumplimiento a los requerimientos necesarios para obtener el grado de Maestría en Ingeniería en Automatización Industrial

en el

Departamento de Ingeniería Eléctrica, Electrónica y Computación
Grupo de Procesamiento Digital de Senales

Enero 2014

UNIVERSIDAD NACIONAL DE COLOMBIA

Abstract

Faculty of Engineering and Architecture
Department of Electronic, Electric Engineering and Computation

Master's on Engineering - Industrial Automation

Methodology for predicting semantic annotations of protein sequences by feature extraction derived of statistical contact potentials and continuous wavelet transform

by Gustavo Alonso Arango Argoty

In this thesis, a method to predict semantic annotations of the proteins from its primary structure is proposed. The main contribution of this thesis lies in the implementation of a novel protein feature representation, which makes use of the pairwise statistical contact potentials describing the protein interactions and geometry at the atomic level. Initially, a protein sequence is decomposed into a numerical series by a contact potential. From the interactions between adjacent amino acids, the wavelet transform can easily detect and characterize subsequences at specific position along the protein sequence. Then, all subsequences are grouped into clusters and a Hidden Markov Model (HMM) profile is built for each one of the groups. Finally, the modeled profiles HMM are used as features in order to build a feature space with the aim to train and evaluate a support vector machine classifier. Evaluations of the proposed methodology are driven against three different views 1) known protein features 2) motif-domain based features (PFam terms) and 3) performance evaluation over several methods for protein annotation prediction. As result, The method have acquired the highest performance prediction in most of the study cases. Thus, this efficiency suggest our approach as an alternative method for the characterization of protein sequences. Although, the research in this thesis focuses on the classification problem, the scientific community can make use of the methodology in two different ways: 1) as a protein predictor and 2) as a motif finding tool. Finally, the source code of the method is free available for download at SourceForge http://sourceforge.net/projects/wamofi/?source=navbar.

UNIVERSIDAD NACIONAL DE COLOMBIA

Resumen

Facultad de Ingeniería y Arquitectura
Departamento de Ingeniería Eléctrica, Electrónica y Computación

Maestría en Ingenieria Automatización Industrial

Metodología para predecir la anotación semántica de proteínas por medio de extracción de características derivadas de potenciales de contacto y transformada wavelet continua

by Gustavo Alonso Arango Argoty

En esta tesis se propone un método para la predicción de anotaciones de proteínas a partir de la estimación de características en secuencias biológicas. Dicha estimación emplea información sobre la estructura de las proteínas a partir de las estadisticas de contactos potenciales entre pares de amino ácidos. Inicialmente, una proteína es transformada a una serie numerica por medio de estos contactos potenciales. Debido a las interacciones entre amino ácidos cercanos, la transformada wavelet puede fácilmente detectar las subsecuencias pertenecientes a posiciones específicas a lo largo de la proteína. Así, todas las subsecuencias son agrupadas de acuerdo a su distribución y éstos grupos son modelados empleando perfiles de Modelos Ocultos de Markov. Finalmente, los perfiles son usados como características donde proteínas de análsis son mapeadas generando así un espacio de representación que es usado para entrenar un clasificador basado en vectores de soporte. La metodolgía ha sido rigurosamente evaluada y comparada con tres diferentes criterios de caracterización: 1) características globales comunmente usadas para representar proteínas, 2) características específicas como motivos y dominios, y por último 3) evaluación de el renimiento de varios programas construidos para la predicción de anotación de proteínas. Como resultado el método propuesto ha logrado los mas altos puntajes de predicción en la mayoría de los casos de estudio. De manera que éstas predicciones sugieren a nuestro método como una alternativa a los comunmente usados algoritmos de caracterización. Por otra parte, a pesar de que el enfoque de la metodología esta diseñada para resolver problemas de clasificación, la comunidad científica puede hacer uso de ella en dos diferentes enfoques: 1) como un predictor de anotaciones en proteínas y 2) como una herramienta para encontrar motivos. Por último, el código fuente del método se encuentra para libre descarga en: http://sourceforge.net/projects/wamofi/?source=navbar.

keywords continuous wavelet transform, statistical contact potentials, protein prediction, support vector machine, sequence alignment.

palabras-clave transformada wavelet continua, potenciales de contacto estadisticos, prediccion de proteins, maquinas de vectores de soporte, alineamiento de secuencias.

Acknowledgements

To my siblings Manuel, Miguel and Martha Arango who emotionaly helped me in the achievement of this goal. To my advisers German Castellanos and Jorge Jaramillo who guide me on the process of the development of this thesis and I specially want to dedicate this work to my parents (*passed away*) Gladys Argoty and Miguel Arango who are my main inspiration for the development of my academic and personal life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The framework of this thesis is located in the field of bioinformatics. Particularly in the area associated to machine learning and pattern recognition, where a methodology for predict semantic annotations of proteins based on a robust feature extraction is proposed. The first part of this thesis covers a review of important concepts on molecular biology and machine learning algorithms applied to protein sequence data sets. The second part comprises all the methods used in the proposed methodology for extracting and modeling of features derived of the protein sequences and the classification of proteins using those features. The third part of the thesis lies on the evaluation of the method using a reported data set on Gram negative bacterial proteins. Finally, we perform a robust testing over several methods employed to characterize and annotate protein sequences on Gram negative bacterial organisms.

Most of the popular methods used to characterize protein sequences reduce the sequence to a feature vector which lose the spatial information of the amino acid arrangement, for instance a simple count of the total number amino acids on the protein sequence. These type of protein representations are commonly known as *global features* and have been widely used for predict semantic annotations of the proteins [1, 2, 3, 4, 5, 6, 7, 8]. On the other hand, several methods use information of the specific arrangement of the nucleic acids along the protein. These attributes are known as *protein motifs or local features* which depicts specific portions of the proteins that generally play an important roll [6, 9, 10, 11, 12]. For instance, PsortB ([6]) a method developed to predict subcellular localizations on Gram negative bacterial proteins uses a set of distinctive motifs as features on each cellular compartment [6]. PFam [11] and PROSITE [10] are data bases of protein domains, motifs and sites. Interestingly, those models have been used in machine learning application as features in for protein semantic annotation [6, 13, 14]. The principal limitations of the global features and the motif-based descriptors lies in

A) the high number of features. For instance, the number of terms for the amino acid composition descriptor growth exponentially depending on the length of the term by 20^n. B) the low sensitivity of the motif profiles as features. This fact is proved in this thesis when the Pfam terms are used as features. Finally, C) those features (local and global) are based on the primary sequence information and in some predictors secondary structure information, however, they don't take into account hints given by structural information. In this research we have proposed the inclusion of structural information of the proteins driven by the statistical contact potentials [15, 16].

Chapter 2

Objectives

2.1 Main objective

Develop a methodology oriented to predict protein sequence annotations based on the extraction of relevant features captured from the primary structure of the proteins.

2.2 Specific objectives

- Characterize and localize relevant features or $motifs$ distributed along the protein sequences by using pairwise statistical contact potentials and continuous wavelet transform.
- Collect and model protein *motifs* by sequence clustering and profile Hidden Markov Models.
- Predict protein sequence annotations on Gram Negative bacterial proteins employing profiles HMMs as features.

Chapter 3

Background

In this background chapter we provide a general overview about molecular biology and machine learning algorithms which have been applied to characterize and predict protein annotations.

3.1 Proteins

Proteins perform most of the work of living cells and they are enrolled in virtually every cellular process as: DNA replication and transcription, production, processing and secretion of other proteins, controlling cell division, metabolism, flow of materials and information into and out of the cell [17].

3.1.1 Protein synthesis

Proteins are synthesized through the information stored in genes by three processes **1) Replication**: One double stranded *deoxyribonucleic acid* (DNA) molecule produces identical copies of itself. **2) Transcription**: Part of the DNA is copied and encoded into a molecule called messenger *ribonucleic acid* (mRNA) and **3) translation** in which the mRNA is transported out of the nuclear membrane acting as template and converted into a chain of amino acids by the ribosomes in the cytoplasm (**Figure 3.1**).

During the translation process, mRNA is decoded into a specific amino acid chain in which a three-RNA codon specifies a single amino acid. There are 20 different amino acids that make up essentially all proteins of the living organisms. Each amino acid has a fundamental design composed of a central carbon (α-carbon) bonded to a) a hydrogen, b) A carboxyl group, c) An amino group or d) A unique side chain or R-group (**Figure**

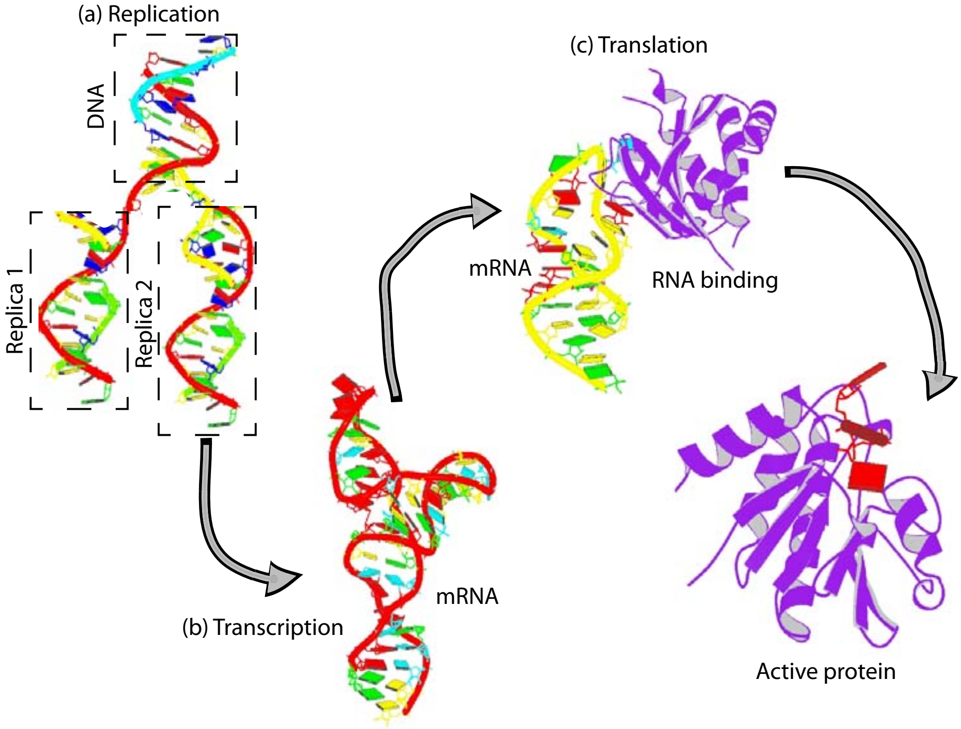


Figure 3.1: The basis of cell and molecular biology: **a) DNA replication** one double-stranded DNA molecule produces two identical copies of the DNA, **b) RNA transcription** a segment of DNA is copied into RNA by the enzyme RNA polymerase, if the gene transcribed encodes a protein, the result of transcription is the messenger RNA (*mRNA*). **c) Protein translation** mRNA is decoded by the ribosome to produce a specific amino acid chain, which, later will fold into an active protein.

3.2). Then, the only characteristic that distinguishes the amino acids is its unique side chain which at the same time dictates the chemical properties of the amino acids [17, 18, 19].

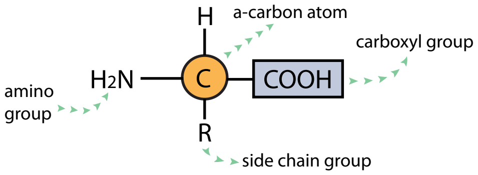


Figure 3.2: General formula of the amino acid showing the positions of amino ($-NH_2$) group, carboxyl group ($-COOH$), α-carbon atom and the side chain R that can be any of the 20 different chains

3.1.2 Protein structure

The properties of a protein are defined by its atomic configuration. Thus, the structure of proteins can be discussed at four levels of organizations:

- **Primary structure.** Describes the arrangement of the polypeptide chain that is made from amino acids in the ribosomes. The chain starts with an amino acid with a free amino group (the N terminus) and ends with one with a free carboxyl group (the C terminus) **Figure** 3.3-A.

- **Secondary structure**. A polypeptide chain always has a spatial organization. The distribution in the chain of amino acids with charged side-groups causes it to be folded into areas of helices (alpha helix) which can be left handed or right handed, beta strands or beta sheets (two or more beta strands are arranged in rows) and turns connecting the helices and strands. Such helices and sheets form the secondary structure which can also combine with one another to form *motifs* or super-secondary structures **Figure** 3.3-B.

- **Tertiary structure**. This is the full description of a folded polypeptide chain. The tertiary structure is referred to three dimensional structure of the proteins and it is stabilized by side chains of amino acids. However, proteins are often not completely stable, so this three dimensional structure often corresponds to the most observed state or the crystallized state of the protein **Figure** 3.3-C.

- **Quaternary structure**. The final functional protein consists of more than one polypeptide chain, in which, all chains have their own primary, secondary and tertiary structures. This association of polypeptides is called the quaternary structure **Figure** 3.3-D. In the example the NIMA-family protein kinases a complex of 4 chains: Nek9/Nercc1, Nek6 and Nek7 that constitute a signaling module activated in early mitosis involved in the control of spindle organization [20].

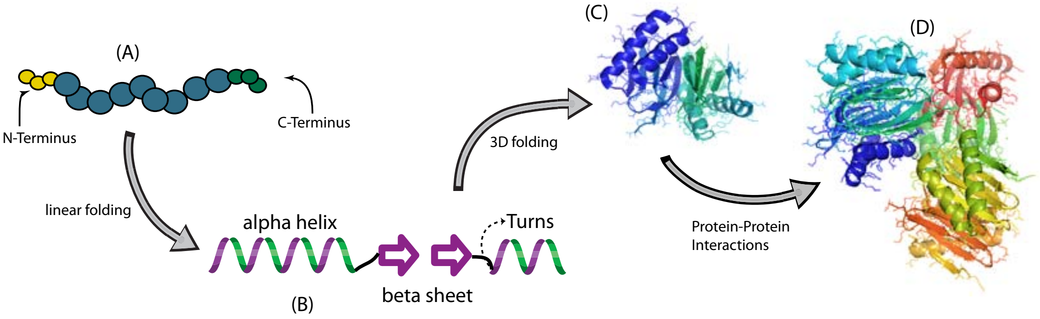


Figure 3.3: Levels of protein structure: A) Primary structure is an arrangement of amino acids. B) Secondary structure is a linear folding of the polypeptides into alpha helices, beta sheets, turns and combinations of those. C) Tertiary structure is the full description of the folding of amino acid chain in a 3D-space. D) Quaternary structure is a complex of different polypeptide chains to accomplish a specific function.

3.1.3 Protein folding

Anfinsen and co-workers studied the effect of the unfolding (denature) and folding (renature) procedures in the ribonuclease protein. They observed that different conformations of the unfolded polypeptide chain always fold into the same native state, thus, they postulated the *thermodynamical hypothesis*. The hypothesis establishes that the native state of a protein in its normal environment is the structure with lowest Gibbs free energy. This property is fundamental for the understanding of protein folding and is why it is believed that the native state of a protein can be predicted just from the knowledge of its amino acid sequence. However it is observed that some proteins receive help to fold from specialized proteins called chaperones [17, 21]

3.1.4 Protein domains and motifs

A **domain** is a compact region of protein structure that is often made up for a continuous segment of the amino-acid sequence, which, in most of the cases retains part of the biochemical function of the larger protein from which they are derived. However, Not all domains consist of continuous stretches of polypeptide. In some proteins, a domain is interrupted by a block of sequence that folds into a separate domain, after which the original domain continues [22]. Domains vary in size but are usually no larger than the largest single-domain protein, about 250 amino acids, and most are around 200 amino acids or less. Forty-nine per cent of all domains are in the range 51 to 150 residues. The largest single-chain domain so far has 907 residues, and the largest number of domains found in a protein to date is 13 [17, 23].

The term *motifs* is referred in two different ways in biology: A) a particular amino acid sequence that is characteristic of a specific biochemical function. For instance, the zinc finger motif, which is found in a widely varying family of DNA-binding proteins [24]. These **sequence motifs** can often be recognized by inspection of the amino acid arrangement of a protein providing strong evidence of the protein biochemical function [25]. B) A **functional motif** that refers to a set of contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain [26]. An example is the helix-turn-helix motif found in many DNA-binding proteins. For instance, In **Figure** 3.4 an example is shown: A) **Functional domain:** *Lac* repressor tetramer binding to each DNA monomer of the *Lac* repressor is made up of a tetramerization domain and a DNA binding domain. B) Schematic diagram of the **domain arrangement** of a number of signal transduction proteins. The different modules have different functions; Pro = proline-rich regions that bind SH3 domains; P = phosphotyrosine-containing regions that bind SH2 domains;

PH = pleckstrin homology domains that bind to membranes; PTPase = phosphatase domain; kinase = protein kinase domain; G-kinase = guanylate kinase domain; GAP = G-protein activation domain; PLC = phospholipase C catalytic domain. The function of the individual modules is sometimes, but not always, independent of the order in which they appear in the protein. Different proteins can contain the same domains/motifs. C) **Motif:** A zinc finger is a small protein structural motif that is characterized by the coordination of one or more zinc ions in order to stabilize the fold. D) **Structural domain:** Helix-turn-helix The DNA-binding domain of the bacterial gene regulatory protein lambda repressor, with the two helix-turn-helix motifs. The two helices closest to the DNA are the reading or recognition helices, which bind in the major groove and recognize specific gene regulatory sequences in the DNA.

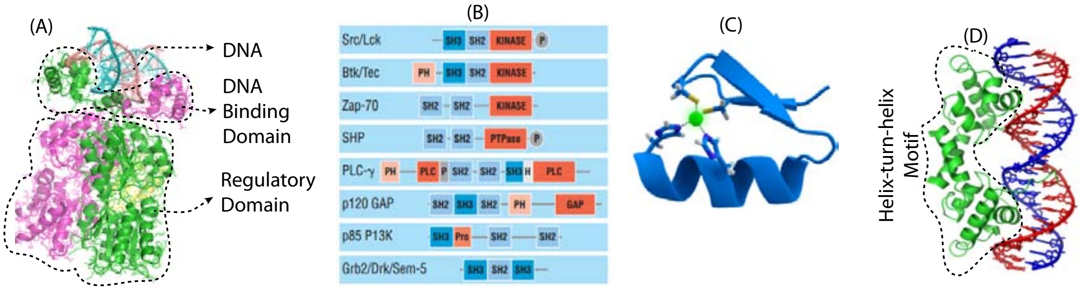


FIGURE 3.4: Domains and motifs.

3.2 Protein annotation prediction

Molecular biology approaches often result in the accumulation of abundant biological sequence data. Ideally, the function of individual proteins predicted using such data would be determined experimentally. However, if a gene of interest has no predictable function or if the amount of data is too large to experimentally assess individual genes, bioinformatics techniques may provide additional information to allow the inference of function [27, 28].

3.2.1 Protein Similarity

Similarity is a concept commonly used for inferring the relationship between diverse things. For example, classification of living organisms can be depicted in terms of hierarchy: Kingdom, Phylum, Class, Order, Family, Genus, Species. This classification is based on the observed similarity which widely reflects a biological ancestry [28]. The characteristics derived from a common ancestor are called *homologous.* For example a bat's wing, a seal's flipper, a cat's limb and a human arm have a common basic anatomy which was present in their last common ancestor and so therefore are homologous. Proteins that are derived from a common ancestor are called homologous [29].

Sequence similarity analysis is the measurement procedure used to infer homology. Essentially, if a new protein sequence is found, it's function and structure can be deduced indirectly by finding similar sequences whose features are known [28]. Methods used to infer homology commonly employ a scoring scale in which higher values depict *similarity*, and a lower values represent *divergence*. Usually sequence similarity is carried out by sequence alignment, which group the amino acid sequences in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships [30, 31, 32].

The methods employed to estimate sequence similarity can be grouped in two major types. **Global sequence alignment** consisting of compare two sequences along their entire length and try to find the the best alignment of the two sequences across their whole length. In general, global sequence alignment methods are most applicable to highly similar sequences of approximately the same size. However, if the degree of the sequence similarity decrease, they tend to miss important biological relationships between sequences [30, 33].

A global alignment may be viewed as a path through a directed *path graph* where each alignment corresponds to a unique path [34]. The similarity between the positions X_i and Y_i of the sequences X and Y is depicted by:

$$SIM(i,j) = max \begin{cases} SIM(i-1,j-1) + s(x_i,y_j) & x_i \; and \; y_j \; aligned \\ SIM(i-i,j) + g & x_i \; aligned \; with \; a \; null \\ SIM(i,j-1) + g & y_j \; aligned \; with \; a \; null \end{cases} \tag{3.1}$$

where: g the gap score for aligning any letter to a null.
X_i is the partial sequence consisting of the first i letters of X.
Y_i is the partial sequence consisting of the first i letters of Y.
$s(a,b)$ is the substitution score for aligning letters a and b, which is usually given by a matrix of size $20X20$. In bioinformatics the most popular substitution matrices are BLOSUM and PAM [35].

Finally, the score of an alignment is the sum of the scores of the edges it traverses (path), but which is the better alignment?. The answer depends on how the matches, substitutions and indels (insertion or deletion mutations) are measured. The cost of an alignment can be calculated by the following general formula:

$$score(path) = \sum_{\forall i,j \in path} SIM(i,j) \tag{3.2}$$

If two sequences in an alignment are derived from a common ancestor, mismatches can be interpreted as point mutations and gaps as indels. The degree of similarity between amino acids occupying a certain position in the sequences can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions in a particular region of the sequence, suggest that this region has structural or functional importance [28, 36].

The second type of sequence similarity scoring is the **Local sequence alignment** in which the sequence comparison is intended to find the most similar regions in the two sequences being aligned rather than finding the best way to align the entire length of the two sequences [37]. The local sequence alignment is capable of finding subsequences within the compared sequences that may have a biological relationship. This type of alignment is the best for sequences of different lengths [30, 33, 37].

The local alignment between two sequences X and Y of length m and n respectively can be computed using dynamic programing as the Smith-Waterman algorithm [30, 38]. Thus, given an alignment matrix SIM the score for each position is computed as:

$$SIM(i,j) = max \begin{cases} max_{1 \leq k \leq i}(SIM(i-k,j) - g(k)) \\ max_{1 \leq k \leq i}(SIM(i,j-k) - g(k)) \\ \quad SIM(i-1,j-1) + s(x_i, y_i) \end{cases} \tag{3.3}$$

Here g stands for the gap penalty function and s represents the cost function. The score for the best alignment is the maximum value in the matrix SIM and the corresponding alignments are paths to this maximal value in SIM from a cell with zero value.

3.2.2 Machine learning and feature-based classification

The major focus of machine learning research is to extract information from data automatically employing computational and statistical methods [39, 40, 41]. Machine learning covers many applications including natural language processing, syntactic pattern recognition, search engines, medical diagnosis, speech and handwriting recognition, object recognition in computer vision among others [42, 43]. Depending of the type of algorithms, the machine learning techniques can be divided into two major classes: **a) the supervised** learning, which attempts to make a generalization of the input data by the training a function using a set of features [42, 44]. Thus, the output of the function can be a continuous value (regression) or can predict a class label of the input object (classification). And **b) the unsupervised** learning in which the labels of the inputs are not used (clustering) [44, 45]. For instance in **Figure** 3.5 there is a description of the two types of learning. **A) Supervised learning** with the goal to construct a function (or model) to accurately predict the target output. In **B) Unsupervised learning or clustering**, the goal is to partition the training samples into subsets (clusters) so the data in each cluster has a high level of proximity. Opposed to supervised learning, the labels of the data are not used or are not available in clustering.

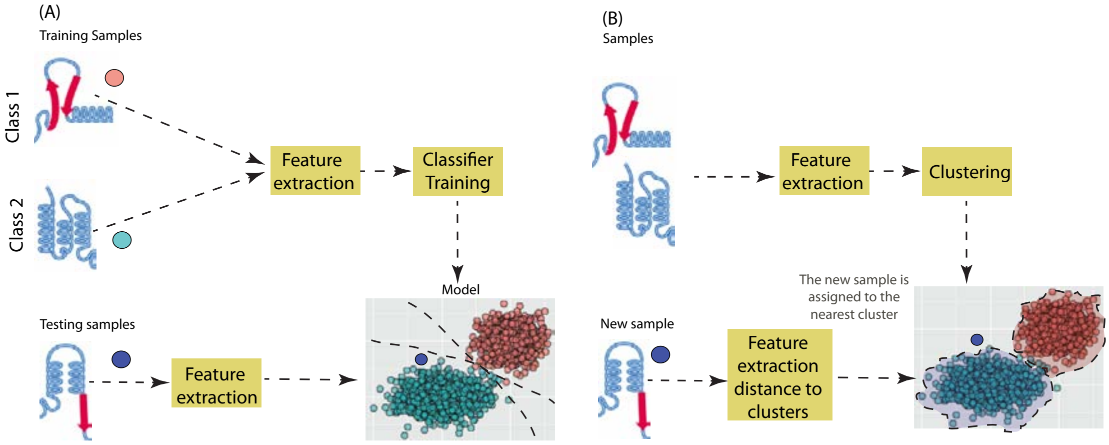


FIGURE 3.5: Supervised and unsupervised learning.

Pattern recognition is a sub-topic of machine learning and aims to classify data (patterns) based on the statistical information extracted from the patterns, which are usually groups of measurements or observations (*feature extraction*) defining points in a multidimensional space [46]. The performance of machine learning algorithms comprises two stages, training and testing. Usually, the labeled samples are split into two parts: one is used to adjust the parameters of the learner algorithm, and the other is used to estimate the generalization error (*evaluation*).

3.2.2.1 Protein classification

In general, the classification of a new protein is performed finding similar sequences whose cellular functions are experimentally determined. Local alignments methods such as BLAST and PSI-BLAST are commonly used to find homologous to the unknown protein from public databases [47]. In many cases these methods perform accurate inferences. However, is well known that a high similarity not necessarily imply the same function [48]. Also, proteins that lack of sequence similarity can achieve the same role. Thus, homology based annotations has important drawbacks such as propagation of errors, threshold relativity, and low sensitivity/specificity. On the other hand, feature-based approaches model the differences between positive and negative samples by the extraction of protein properties arranged into a feature space (see section:3.2.2.2) which is used as input to a classifier. Finally, there are hybrid approaches, which consist on the use of homology methods collectively with feature and local-feature based characterizations [49, 50, 51].

The prediction of protein annotations is usually considered as a classification problem, where a set of features are collected for each protein, and the learning algorithm is used to infer the association rule between the features and the annotations. Among the supervised learning algorithms, support vector machines have been particularly popular due to their good performance and strong statistical background [52]. For instance, signal peptide prediction represents one of the big successes in the protein classification field. Algorithms are approaching a performance level comparable to the quality of the underlying experimental data, even in some cases better [53]. Several applications of machine learning applied to protein annotation prediction have been developed. Among the most popular methods for protein prediction we have: SignalP [54], LOCTree[55], CellPLoc [13] among others. Details about protein annotation prediction can be consulted in the follow reviews: [49, 52, 53, 56].

3.2.2.2 Feature extraction

Feature extraction is the process used for capturing information from the protein sequences. Depending of the type of characterization, the features can be categorized as global feature extraction and local feature extraction. **Global feature extraction** extract information from the protein sequences regardless the order in which the amino acids occur along the protein. Some of the most popular global features are:

- The **k-peptide composition (KCC)**[57] computes the frequency of the k-peptide along the protein sequence and it is expressed as:

  $$f(r_1, .., r_k) = \frac{N_{r_1,...,r_k}}{N-1}, \tag{3.4}$$

  where, $r_1, .., r_k = 1, 2, ..., 20$, $N_{r_1,...,r_k}$ is the number of the $k-$peptide composed of the amino acids $r_1, ..., r_k$.

- The **global descriptor (GD)**method was proposed to predict protein folding classes and human Pol II promoter sequences, also this has been used to distinguish coding from non-coding sequences in a prokaryote complete genome [1]. The global descriptor contains three parts: composition (Comp), transition (Tran) and distribution (Dist). *Comp* describes the overall composition of a given symbol in the new symbol sequence. *Tran* characterizes the percentage frequency that amino acids of a particular symbol are followed by a different one. *Dist* measures the chain length within which the first, 25, 50, 75 and 100% of the amino acids of a particular symbol are located.

- The **Lempel-Ziv (LZ)** complexity is one of the conditional complexity measures of symbol sequences. The LZ complexity has been successfully employed to construct phylogenetic tree and predict protein structural classes [58].

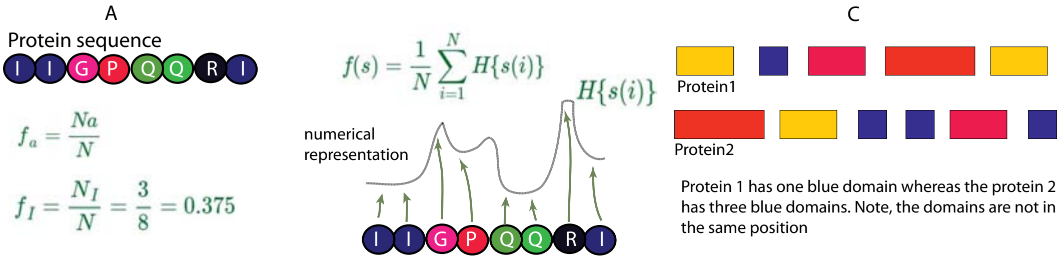
- **Autocorrelation descriptors (AD)** are all defined based on the value distributions of 30 physicochemical properties of amino acids along a protein sequence [59]. The three widely-used autocorrelation descriptors are: normalized Moreau-Broto autocorrelation descriptors, Moran autocorrelation descriptors and Geary autocorrelation descriptors [59].

- Features based on amino acid composition such as: Amino acid composition, pseudo amino acid composition and quasi-sequence order descriptors [1, 2, 3].

- **Physiochemical properties of the amino acids** have been used as features in basically two different approaches: 1) *Global features* proteins are represented as numerical series using the physiochemical properties of the amino acids. Then, some statistics as the mean are extracted from this numerical arrangement [4]. 2) *Localized features* proteins that accomplish the same function can have difference in length given by some specific deletions at genomic level [60], or by the structure of the proteins, in which one protein is multi-domain whereas a related protein is composed by one domain [61]. In [5] a method to avoid this problem is proposed, in which, the protein is compressed into a fixed length using a slide window taking the average of the amino acids value among this interval. The amino acid index *aaindex* is a comprehensive data base of 544 amino acid properties clustered into different areas: alpha and turn propensities, beta propensity, composition, hydrophobicity, physiochemical properties, and others [62].

- Proteins can be represented in terms of families, domains, motifs and functional sites ([9], [10], [11]). Some methods use annotated protein domains from different databases to build feature vectors [6, 9, 12].

An example of protein characterization and feature extraction is given in **Figure** 3.6: A) From the amino acid sequence a set of features are extracted by the amino acid composition, in the example, the Isoleucine composition $I = 0.375$. **B)** Numerical representation of the amino acids, from the series a set of features can be extracted, for example if the index is the hydrophobicity, the mean of the series represents the average hydrophobicity of the protein. Other features can be extracted from this representation as is shown in [5]. **C)** Domain arrangement of the proteins, some related proteins may have the same domain in different positions along its sequence. Methods based on this approach use the domains as features, setting a score if the domain is or not present along the protein.



Protein 1 has one blue domain whereas the protein 2 has three blue domains. Note, the domains are not in the same position

FIGURE 3.6: Protein feature representation

Chapter 4

Methodology

In this thesis, a method to predict semantic annotations of protein sequences making use of the features distributed along the protein sequence represented by motifs is proposed. These features are detected by a continuous wavelet transform that has been reported as a powerful tool for the characterization of protein motifs [63, 64, 65, 66, 67, 68]. However, the most important aspect in the wavelet analysis is the protein's numerical representation; here, the use the pairwise protein contact potentials from the aaindex database is proposed [62] with the aim of extracting protein structural information for representing protein primary structures. In the proposed method, sequences are described as numerical representations from residue-residue interactions by the contact potentials, then the wavelet transform decodes these interactions allowing the identification of groups of amino acids with similar contact (free energy) distribution, thus, the protein is splitted into those groups and so it can be represented as a set of subsequences. If a group of proteins has related distributions, they will produce similar subsequences, so, by using a clustering procedure those subsequences can be grouped into a set of homology/related ones, which in turn can be modeled by a Hidden Markov Model (HMM). Biologically, if a set of proteins has a motif/domain in common and it is well characterized by a contact potential, a continuous wavelet transform can easily identify and localize this motif regardless of whether it is distributed in different positions among the set of proteins or even if the motif has mutations [63, 68]. Thus, this outcome to identify conserved motifs in a specific protein categorization is used in this thesis. Two datasets are defined in order to avoid bias and overtraining on the classification: A protein modeling dataset, used to detect the expressed motifs (profiles HMMs) and a control data set used to evaluate the performance of the method. Then, a set of single-class-SVM classifiers is used as predictor. The method is divided into two main stages as follows:

4.1 Profile descriptor

The profile descriptor models a set of proteins from a *modeling dataset* $C = \{C_1, .., Cm\}$, where, m is the number of classes) as a group of profiles based on the usage of HMMs. The name modeling dataset is given because these sequences are only and exclusively used to build the profile HMMs. The profile descriptor assumes that recognizable amino-acid sub-sequences are found in different proteins that usually indicate biochemical function, for instance, the so-called zinc finger motif which is determined by this three-dimensional structure, however, it can also be recognized based on the primary structure of the protein [69]. These sub-sequences are known as motifs. Furthermore, motifs are assumed to be frequently located among proteins with the same function (see section 3.1.4). The scheme of the profile descriptor is shown in **Figure** 4.1. Let's set $C_c = \{S_1, ..., S_r\}$ as the protein dataset defined for the class c. Then, the numerical representation of a protein S_i is obtained by the decomposition of the protein sequences by the statistical contact potentials (see 4.1.1) as is explained in section 4.1.3.1. Next, subsequences X_s^k are extracted from the protein sequences by using the continuous wavelet transform (see 4.1.2) where k is the total number of sections where the protein S is broken. Sections *motif detection* (4.1.3.2) and *projection* (4.1.3.3) describe the methods to split the protein into several variable-length sequences. Note, the subsequence detection is conducted for each one of the proteins in the class C_c. Thus, all the sequences for this class are splitted and represented by $X_c = \{X_{s_1}^{k_1}, .., X_{s_i}^{k_i}\}$, where s_i is the ith protein on the class c and k_i is the number of subsequences for this protein. Then, all the subsequences in X_c are grouped using the progressive alignment software *CLUSTAL*Ω [70] into a set of clusters $\zeta_c = \{\zeta_1, ..., \zeta p\}$, where p is the total number of clusters on the class c (see section 4.1.4 for details on progressive alignment). Clusters are defined by cutting branches, where higher cutoff levels relate groups with a large number of elements being mostly non-correlated, whereas lower levels reveal correlated groups. In order to identify the optimal branch break, the *DynamicCutTree*, a fast and accurate method for cutting dendrograms, is used [71]. Finally, each one of the clusters in ζ are modeled by a profile HMM (see section 4.1.5 for details on Hidden Markov Models) using the software package *HMMER3.0* [72], thus, the full sequences on the class C_c are reduced as a set of profile HMMs $H_c = \{h_1, .., h_p\}$.

Finally, if the profile descriptor process is carried out over the whole set of classes, a set of profile HMMs $\Theta = \{H_1, ..H_m\}$ depicts the entire database where m is the total number of classes. Thus, the transformation from the whole set sequences C is represented as a set of profiles HMMs Θ.

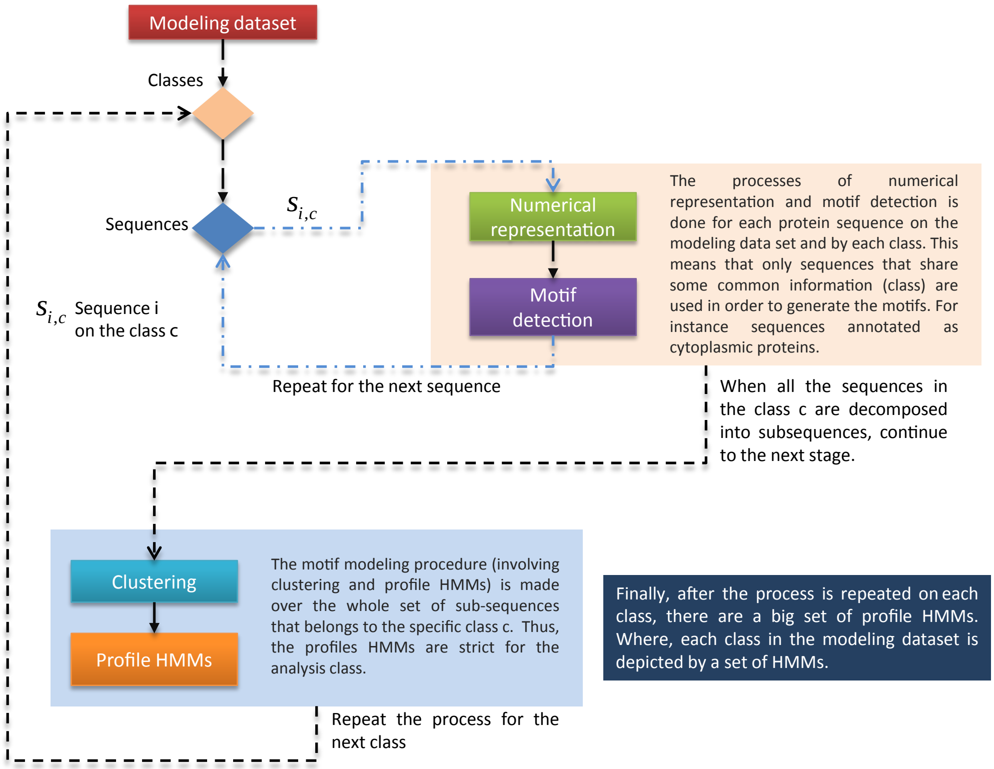

Modeling dataset
Classes
Sequences
$S_{i,c}$
$S_{i,c}$ Sequence i on the class c
Numerical representation
Motif detection
The processes of numerical representation and motif detection is done for each protein sequence on the modeling data set and by each class. This means that only sequences that share some common information (class) are used in order to generate the motifs. For instance sequences annotated as cytoplasmic proteins.
Repeat for the next sequence
When all the sequences in the class c are decomposed into subsequences, continue to the next stage.
Clustering
Profile HMMs
The motif modeling procedure (involving clustering and profile HMMs) is made over the whole set of sub-sequences that belongs to the specific class c. Thus, the profiles HMMs are strict for the analysis class.
Finally, after the process is repeated on each class, there are a big set of profile HMMs. Where, each class in the modeling dataset is depicted by a set of HMMs.
Repeat the process for the next class

FIGURE 4.1: Decomposition of a set of sequences into a ensemble of clusters depicted as profile HMMs. Subsequences are extracted by the use of the continuous wavelet transform and the decomposition from the pairwise statistical contact potentials.

4.1.1 Statistical contact potentials

Statistical contact potentials are energy functions derived from analysis of proteins over their tertiary structures [15]. These energies are widely used in computer applications such as: folding, docking, or protein identification. They are derived from: (a) observed pairing frequencies of the 20 amino acids in databases of known protein structures, and (b) approximations and assumptions about the physical process that these quantities measure [15, 16].

Possible features to which an energy can be assigned including: torsion angles (such as the ϕ, ψ angles of the Ramachandran plot), solvent exposure or hydrogen bond geometry and pairwise amino acid contacts or distances [16]. For pairwise amino acid contacts, a statistical potential Y is formulated as an interaction matrix that assigns a weight or energy value $Y[i, j]$ to each possible pair of standard amino acids i, j. In **Figure** 4.2 a typical interaction between two amino acids is shown. Commonly, the energy of a particular structural model is then the combined energy of all pairwise contacts (defined as two amino acids within a certain distance of each other) in the structure. The energies are determined using statistics on amino acid contacts in a database of known protein structures.

The *AAindex* is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids [62]. AAindex consists of three sections: AAindex1 for the amino acid index of 20 numerical values, AAindex2 for the amino acid mutation matrix and AAindex3 for the statistical protein contact potentials (see Chapter 9 table 9). Those statistical contact potentials comprise: Energy of interactions on buried environment, energy transfered of amino acids from water to the protein, statistical contact potentials and contacts derived from x-ray crystal structures, interactions energies derived from side chain contacts, distant dependent potentials, potentials derived from perceptron criterion, energy derived from protein-protein complexes, optimization-derived potential obtained from a set of decoys, quasichemical potential derived from parallel, antiparallel and intermediate orientations and environment-dependent residue contact energies [15, 16, 48].

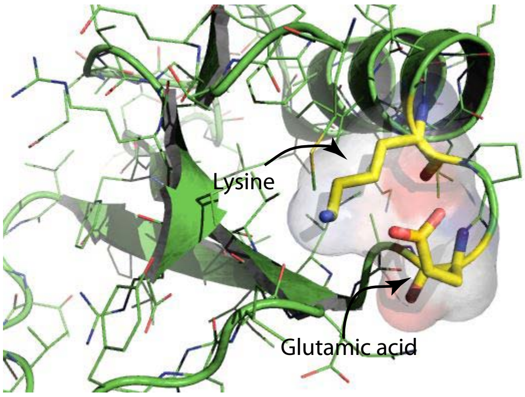


FIGURE 4.2: Interaction between Lysine and Glutamic acid in a protein complex. Contact potentials are derived from the interactions between a couple of amino acids

4.1.2 Continuous wavelet transform

The wavelet transform is a mathematical function that converts a one dimensional signal into a two dimensional representation. This conversion reveals hidden features within the original signal and represent the original signal more succinctly [63, 73]. A mother wavelet is needed in order to realize the wavelet transform which is a small oscillating wave with its energy concentrated in time [64, 74].

The difference between a wave and a wavelet is that a wave is usually smooth and regular in shape, whereas, a wavelet may be irregular in shape, and normally lasts only for a limited period of time [64]. A wave (e.g., sine and cosine) is typically used as a deterministic template in the Fourier transform for representing a signal that is time-invariant or stationary [75]. In comparison, a wavelet can serve as both a deterministic and nondeterministic template for analyzing time-varying or nonstationary signals by decomposing the signal into a 2D, time-frequency domain [64, 65].

Mathematically, a wavelet is a square integrable function ψ that satisfies the admissibility condition [65].

$$\int_{-\infty}^{\infty} \frac{|\Psi(f)|^2}{(f)} df < \infty \tag{4.1}$$

Where, $\Psi(f)$ is the Fourier transform (i.e. frequency domain expression) of the wavelet function ψ (in time domain). The admissibility condition implies that the Fourier transform of the function ψ wear off at zero frequency; in other words,

$$|\Psi(f)|^2 \,|_{f=0} = 0 \tag{4.2}$$

This means that the wavelet transform must have a band-pass like spectrum. Also the average value of the wavelet ψ in the time domain is zero:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{4.3}$$

This means that the wavelet transform must be of oscillatory nature [65].

Using a dilation (s) and translation (τ) parameters. A family of translated and scaled wavelets can be defined as follows:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi \frac{(t-\tau)}{s}, s > 0, t \in R \tag{4.4}$$

The purpose of the $\frac{1}{\sqrt{s}}$ factor implies that the energy of the wavelet family will remain constant under different scales [65]. If the energy of the wavelet function $\psi(t)$ is assumed as:

$$\epsilon = \int_{-\infty}^{\infty} |\psi(t)|^2 dt \tag{4.5}$$

The energy of the scaled and translated wavelets can be calculated dividing by s as follows:

$$\epsilon = \frac{1}{s}\int_{-\infty}^{\infty} |\psi(\frac{t-\tau}{s})|^2 dt \tag{4.6}$$

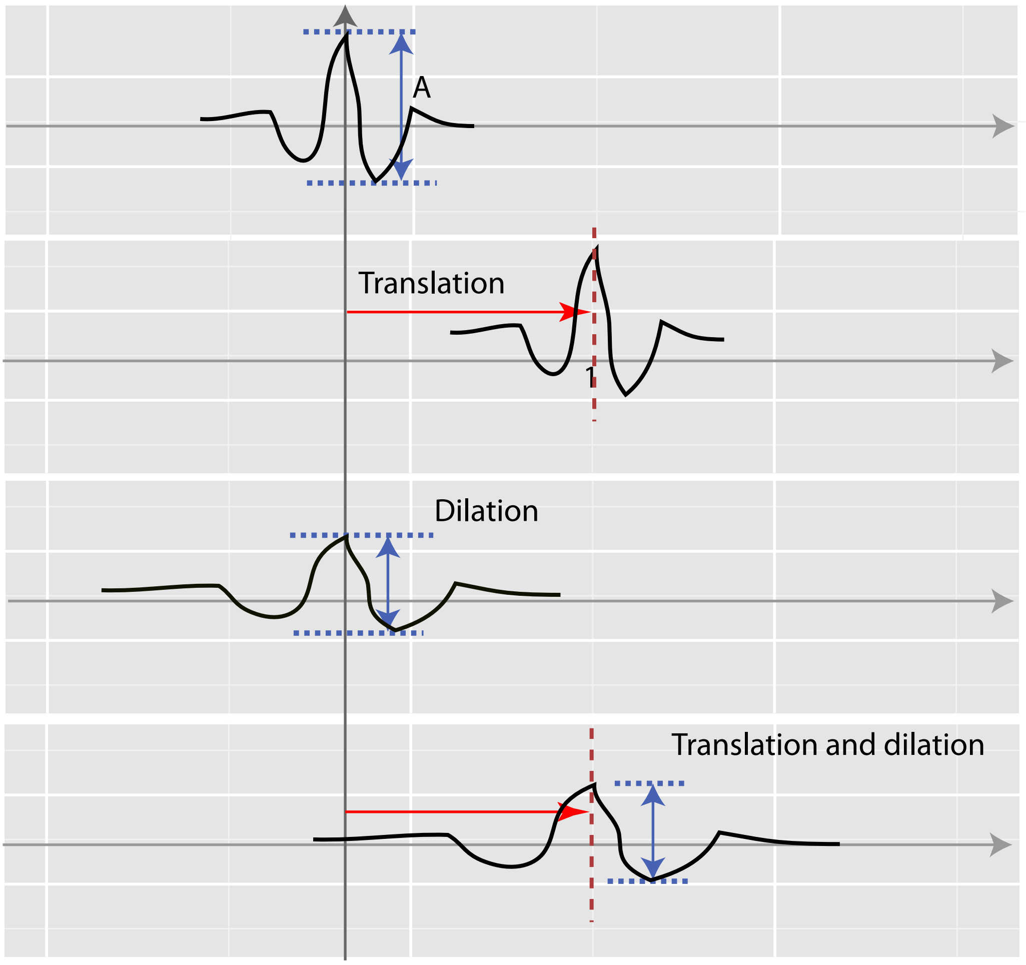


FIGURE 4.3: Illustration of the translation and dilation of the wavelet

As a result, the energy of the original base wavelet ϵ and the scaled and translated wavelets remains the same. This relationship is illustrated in **Figure** 4.3. The process through which a signal is decomposed by analyzing it with a family of scaled and translated wavelets is called the wavelet transform [63, 64, 65, 66, 67, 68]. The continuous wavelet transform (CWT) of a signal $x(t)$ is defined as:

$$wt(s,t) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-\tau}{s})dt \tag{4.7}$$

where $\psi^*(.)$ is the complex conjugate of the scaled and shifted wavelet function $\psi(.)$. Thus, the CWT is an integral transformation as the Fourier transform, due to the integration operation will be performed in both transforms [64, 65]. Also, as the wavelet contains two parameters (scale parameter s and translation parameter t), transforming a signal with the wavelet basis means that such a signal will be projected into a 2D, time-scale plane, instead of the 1D frequency domain in the Fourier transform [75]. Furthermore, because of the localization nature of the wavelet, the transformation will extract features from the signal in the time-scale plane that are not revealed in its original form, for example, what specific bearing defect-related spectral components existed at what time [64, 65].

4.1.2.1 Resolution of the continuous wavelet transform

The continuous wavelet transform enables variable window sizes in analyzing different frequency components within a signal [66]. This is carried out by comparing the signal with a set of template functions obtained from the *scaling* (dilation and contraction) and shift (translation along the time axis) of a base wavelet $\psi(t)$ and looking for their similarities, as shown in **Figure** 4.4. The resolution on the CWT is good at high frequencies, however, the bandwidth of mother wavelet is wide for these frequencies and as consequence the scale resolution is not good. On the other hand, for low frequencies, the mother wavelet is wide on time and has a concentration on high frequency allowing characterize low frequency components but with a low resolution in time [64, 65, 66]. The time-frequency space division is not uniform. Moreover, the wavelet components represented by the rectangles in the **Figure** 4.4 keep a constant area. The uncertainty principle is applied to the wavelet transform [67].

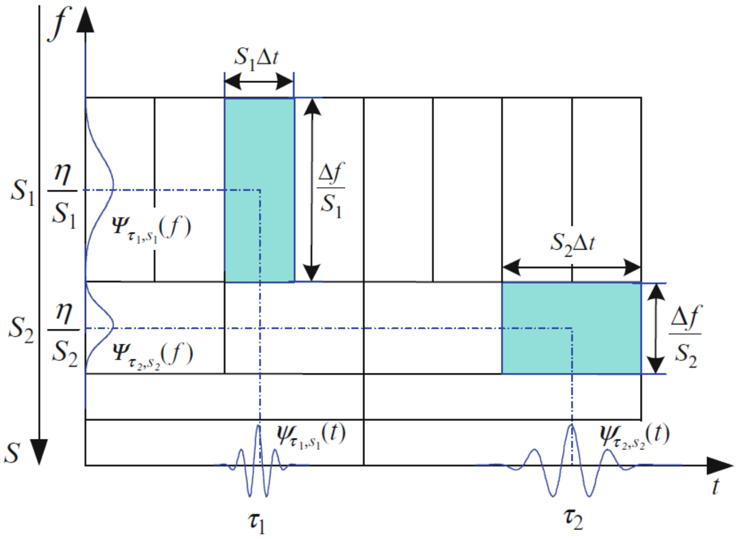


FIGURE 4.4: Illustration of the translation and dilation of the wavelet

4.1.3 Feature detection by wavelet transform

The continuous wavelet transform have been used to characterize protein sequences specially for detecting conserved regions inside the protein [63, 68, 73]. In [63] the continuous wavelet transform have been used to identify and characterize repeating motifs. They have proved the ability of the wavelet transform to detect concrete motifs on a set of specific proteins. For their numerical representation they have used relative accessible surface area (rASA) [76] and Kyte-Doolittle hydrophaty scale [62]. Both, indicators of the protein's overall geometry. In this thesis we have extended the work developed by [63] in several aspects: 1) we have used pairwise statistical contact potentials instead of rASA or Hydrophaty, the reason is because the potentials reveals preferences of contacts between pairwise amino acids in a set of related proteins at atomic level. Then, those potentials increase the specificity of the protein's geometry. 2) Based on the claim that protein domains/motifs can be placed in a diverse set of sequences at different locations (see **Section** 3.1.4). We have included a cluster and HMM steps with the purpose to merge domains/motifs which are not highly conserved. 3) we have employed those domains/motifs in order to predict annotations of the proteins. Thus, our method can be applied to any set of proteins. The mathematical formulation of the motif detection is described as follows:

4.1.3.1 Numerical representation

Let's the protein sequence $S(t) = \{s_1, ...s_i, ...s_t\}$ of length t, s_i express the ith residue of the protein S belonging to one of the 20 native amino acids, represented by the numerical signal $F(t) = \{f_1, ..., f_i, ..., f_{t-1}, ..., f_t\}$, where, $F(t)$ is defined as the distribution of the amino acids along the protein given the pairwise score from the contact potential matrix Y.
Each element f_i in the series $F(t)$ is defined as follows:

- $f_i = Y[s_i, s_{i+1}]$. The energy/score between the i and $i + 1$ amino acids, where $i < t$.

In the special case of t (last amino acid in the sequence) f_t is defined as $f_t = Y[s_t, s_{t-1}]$.

In the **Figure** 4.5 the conversion from the sequence to numerical series is shown. The sequence S is decomposed into the distribution series F, where each couple of adjacent amino acids $S[i, j]$ is decomposed by the contact potential $Y[i, j]$ (blue arrow) and the score is represented as $F[i]$.

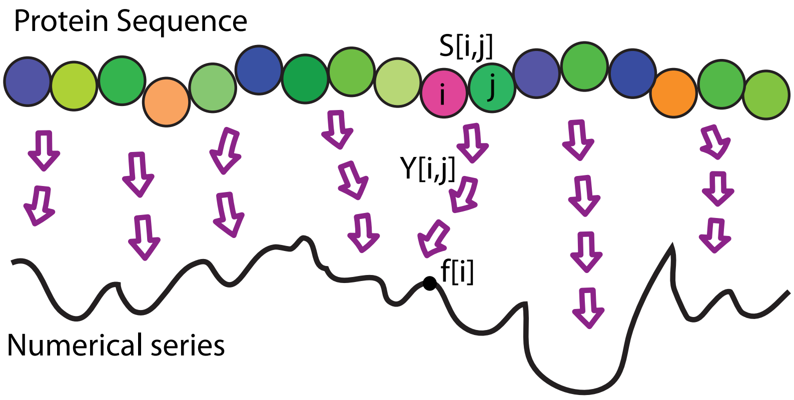


FIGURE 4.5: Numerical representation of a protein sequence

4.1.3.2 Motif detection

Given the numerical sequence $F(t)$ the CWT $WT_F(s, \tau)$ (Equation:4.7) provides a representation of the interactions between adjacent amino acids. These patterns or subsequences are distributed along the protein sequence at different scales in the CWT domain as shown in **Figure** 4.6.

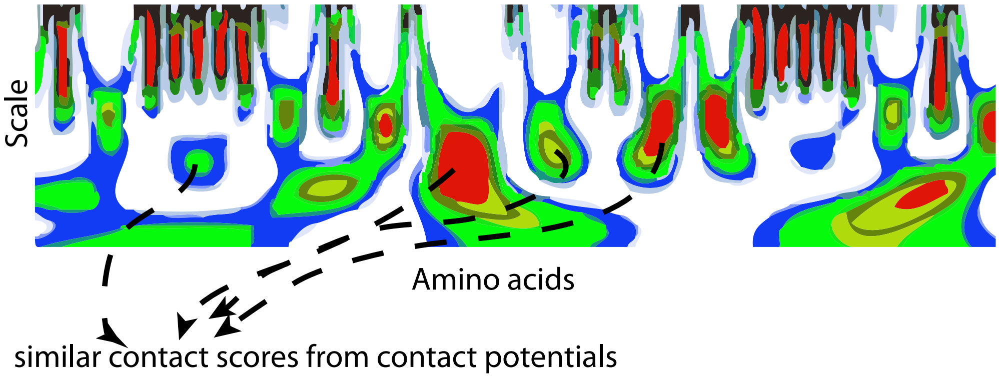


FIGURE 4.6: Numerical representation of the protein is decomposed into the sequence-scale domain by the CWT, then adjacent amino acids with similar level of energy from the contact potentials are grouped

Patterns localized along the protein sequence are computed separating the wavelet space matrix $WT_F(s, \tau)$ into two binary matrices WT_F^+ and WT_F^- which are determined by the score given by the contact potentials as shown in **Figure** 4.7. For instance, if the score from the potential is an energy approximation, the negative regions indicate that the amino acid contacts have high scores/energy (e.g., binding affinities) and thus they could represent binding sites or common contacts on the structures of the proteins. Positive scores are associated to non specific contacts on the statistical contact potential.

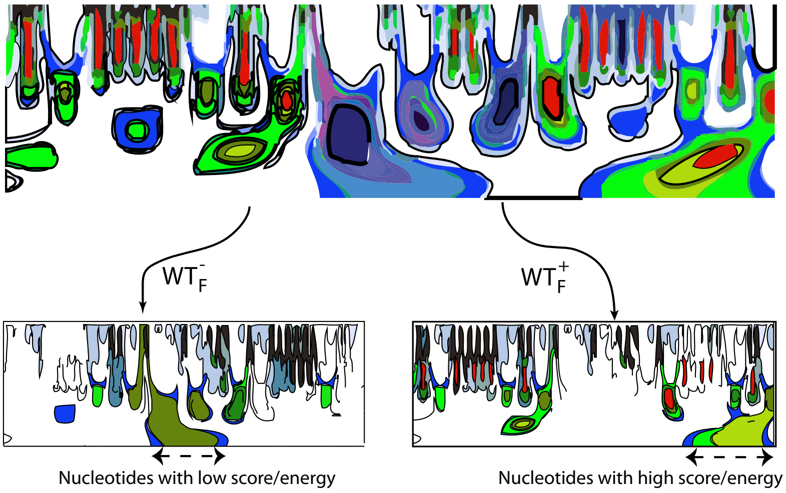


FIGURE 4.7: Identification of conserved regions in the wavelet space

WT_F^+ and WT_F^- matrices are expressed as follows:

$$WT_F^+ = \begin{cases} 1 & W_f \geq thr \\ 0 & other \end{cases} \tag{4.8}$$

$$WT_F^- = \begin{cases} 1 & W_f \leq thr \\ 0 & other \end{cases} \tag{4.9}$$

Where thr is the threshold for which the wavelet space WT_F is splitted. This threshold is defined as the mean value of the all amino acid pairwise interactions over the sequence S, expressed as:

$$thr = \frac{1}{N_a t} \sum_{j=1}^{t} \sum_{i=1}^{N_a} WT_F(i,j) \tag{4.10}$$

N_a is the total number of scales and t is the length of the protein.

4.1.3.3 Projection

This step consists on the extraction of the amino acid sequence that corresponds to each regions. Thus, in order to model the possible motifs in the sequences it is necessary to recover the sequence of the subsequences. Then, each region $r[k]$ on either WT_F^+

and WT_F^- is projected into the sequence S. Where, the position of each region $r[k]$ corresponds to the sequence position as shown in the **Figure** 4.8. Then, each region is expressed as the subsequence X_s^k shown in **Figure** 4.8.

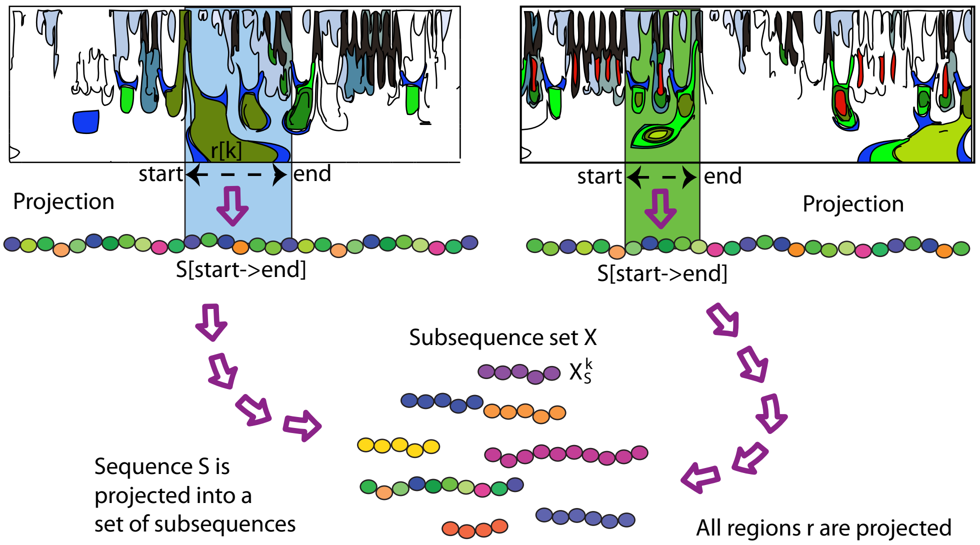


Figure 4.8: Protein feature detection by the wavelet transform

4.1.4 Multiple sequence alignment

In bioinformatics analysis, the multiple sequence alignments have been widely used specially to compare homologous sequences. The exact way to compute an optimal alignment among N sequences has a computational complexity of $O(L^N)$ for sequences of length L making it improper even for a small set of sequences. However, *Hogeweg P and coworkers* [77] proposed the progressive alignment, which aligns sequences in larger and larger sub-alignments, following the branching order in a guided tree. This method usually involves two sets of parameters: a gap penalty and a substitution matrix assigning scores to the alignment of each possible pair of amino acids. This scores are based on the similarity of the chemical properties of the amino acids and the evolutionary probability of the mutation [30, 31, 32]. First the most closely related sequences are aligned, then, the more distant sequences are added gradually. This approach has a computational complexity of $O(N^2)$ and works well when the protein set consists of sequences of different degrees of divergence. Pairwise alignment of very closely related sequences can be carried out very accurately. However, if the identity among the sequences is less than $\tilde{2}5$ - 30% this progressive approach becomes much less reliable [30].

The two major problems of the progressive alignment are:

1. The local minimum problem: following the guided tree the algorithm adds sequences together. However, there is no guarantee that the global optimal solution (multiple alignment quality) will be found. In other words, any misaligned regions made early in the alignments process cannot be corrected later as new information or sequences are added.

2. Alignment parameters: Commonly two parameters are chosen, a weight matrix and two gap penalties (one for opening a new gap and one for extending an existing gap). When the sequences are all closely related, this works well over all parts of all the sequences in the data set. The first reason is because all residue weight matrices give most weight to identities. Then, if identities dominate an alignment, almost any weight matrix will find approximately the correct solution. In the case of divergent sequences the scores given to non-identical residues will become critically important. Because, there will be more mismatches than identities. Different weight matrices will be optimal at different evolutionary distances or for different classes of proteins. The second reason lies in the range of the gap penalty values that will find the best possible solution. This value can be very broad for highly similar sequences [30].

The multiple alignment algorithm proposed on the CLUSTAL family [30, 31, 32] consists of three main stages: (1) all pairs of sequences are aligned separately in order to calculate a distance matrix based on the divergence of each pair of sequences; (2) a guide tree is calculated from the distance matrix; (3) the sequences are progressively aligned according to the branching order in the guide tree as illustrated in **Figure** 4.9.

4.1.4.1 The distance matrix

In the CLUSTAL programs, the pairwise distances were computed using a fast approximate method [78] allowing alignments for a large number of sequences. However, CLUSTAL offers the option to use another method calculating scores from full dynamic programming alignments using two gap penalties (for opening or extending gaps) and a full amino acid weight matrix. These scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded). Both of these scores are initially calculated as per cent identity scores and are converted to distances by dividing by 100 and subtracting from 1.0 to give number of differences per site [30].

4.1.4.2 The guide tree

The trees used to guide the final multiple alignment process are calculated from the distance matrix (previous subsection) using the Neighbor Joining method [79]. This produces unrooted trees with branch lengths proportional to estimated divergence along each branch. The root is placed by a mid-point at a position where the means of the branch lengths on either side of the root are equal [30]. These trees are also used to derive a weight for each sequence [79]. The weights are dependent upon the distance from the root of the tree but sequences which have a common branch with other sequences share the weight derived from the shared branch.

4.1.4.3 Progressive alignment

The basic idea of this stage is to align larger and larger groups of sequences by a series of pairwise alignments, following the branching order in the guide tree. The alignment proceeds from the root of the rooted tree. At each stage a full dynamic programming [38] algorithm is used with a residue weight matrix and penalties for opening and extending gaps. Each step consists of aligning two existing alignments or sequences. In the basic algorithm, new gaps that are introduced at each stage. In order to calculate the score between a position from one sequence or alignment from another, the average of all the pairwise weight matrix scores from the amino acids in the two sets of sequences is used. Sequence weights are calculated directly from the guide tree. The weights are normalized such that the biggest one is set to 1.0 and the rest are all less than 1.0. Groups of closely related sequences receive lowered weights because they contain much duplicated information. Highly divergent sequences without any close relatives receive high weights. These weights are used as simple multiplication factors for scoring positions from different sequences or prealigned groups of sequences [30].

4.1.5 Hidden Markov Models

Hidden Markov Models (HMMs) are statistical models which are generally applicable to time series or linear sequences. They have been used in speech recognition applications [80]. HMMs have been introduced to computational biology analysis in the late 80's [81]. A HMM can be viewed as a finite state machine. Where a finite state machine can move through a series of states and produce an output, either when the machine has reached a particular state or when it is moving from state to state [82, 83]. The HMM generates a protein sequence by emitting amino acids as it progresses through a series of states. Each state has a table of amino acid emission probabilities, and transition probabilities

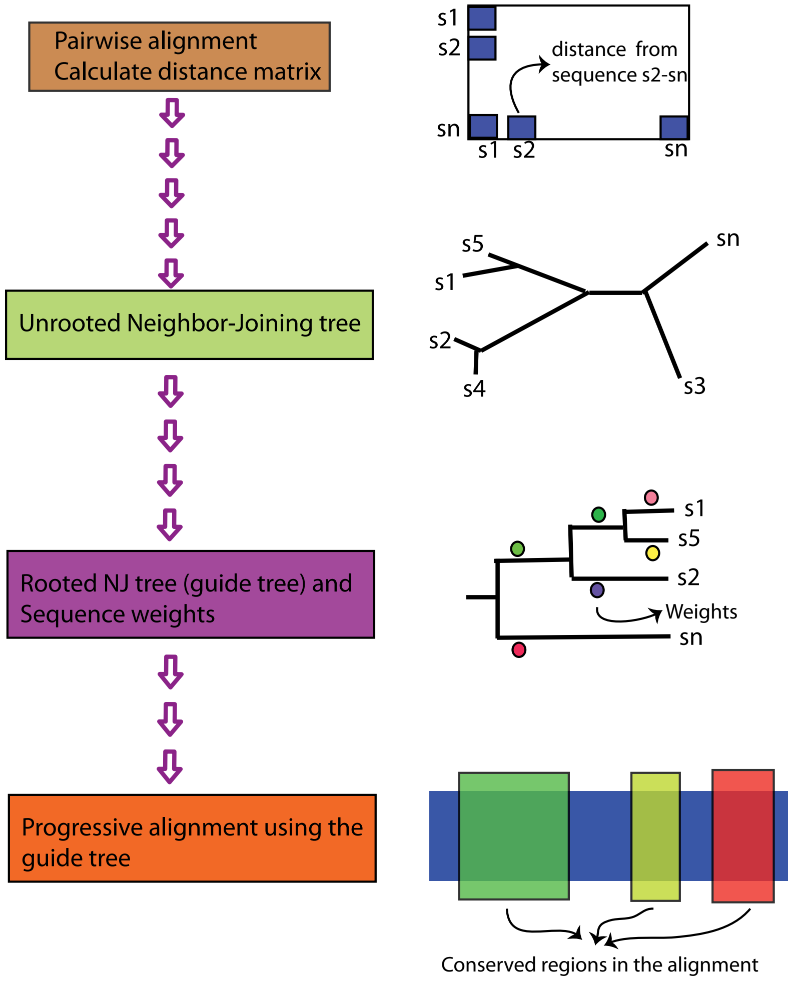

Pairwise alignment
Calculate distance matrix
s1
s2
sn
s1 s2
sn
distance from
sequence s2-sn
Unrooted Neighbor-Joining tree
s5
s1
s2
s4
sn
s3
Rooted NJ tree (guide tree) and
Sequence weights
s1
s5
s2
Weights
sn
Progressive alignment using the
guide tree
Conserved regions in the alignment

FIGURE 4.9: Progressive alignment algorithm implemented in the CLUSTAL programs

for moving from state to state. Transition probabilities define a distribution over the possible next states [82, 83, 84].

An example of a simple HMM that models sequences composed of two letters (a, b) is shown in **Figure** 4.10. This model is appropriate for a problem in which the sequences started with one residue composition a then switched once to a different residue composition b. The HMM consists of two states connected by state transitions. Each state has a symbol emission probability distribution for generating (matching) a symbol in the alphabet. An HMM can be viewed as a generator of sequences. Starting in an initial state, a new state with some transition probability is chosen by staying in state 1 with transition probability $t_{1,1}$, or moving to state 2 with a transition probability $t_{1,2}$. Then, a residue with an emission probability specific to that state is generated. The transition/emission process is repeated until an state is reached. Finally, there is a hidden state sequence that is not observed and a symbol sequence which is observed [72, 82, 83, 84, 85]. The states of the HMM are often associated with a meaningful biological labels [82]. Any sequence can be represented by a path through the model. This path follows the Markov assumption, that is, the choice of the next state is only dependent on the choice of the current state. Inferring the alignment of the observed sequence to the hidden sequence is like labeling the sequence with relevant biological information.

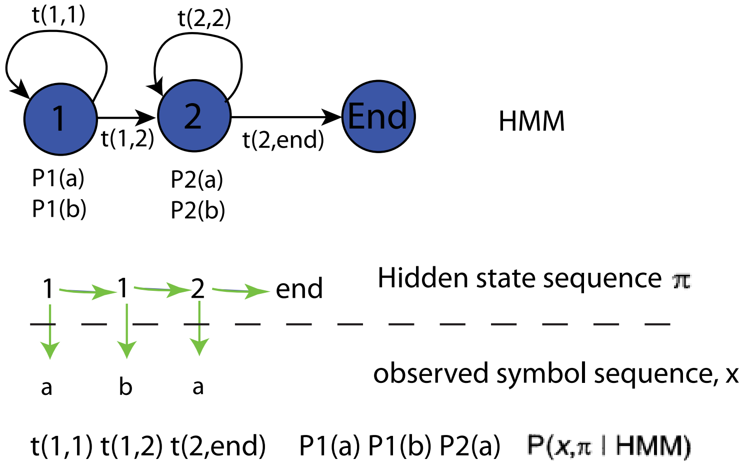


FIGURE 4.10: A HMM modeling sequences of a and b as two regions of potentially different residue composition. The model, blue circles as states and narrows as transition states. A state sequence is generated from the model and a possible symbol sequence. The joint probability $P(x, \pi|HMM)$ of the sequence and the state sequence is a product of all the transition and emission probabilities.

Parameters on the HMM can be set by (i) training of the HMM from an initially unaligned set of sequences or (ii) from a set of pre-aligned sequences. In the latter case, the parameter estimation consists of converting observed counts of symbol emissions and state transitions into probabilities. Thus, for the building of a profile HMM an existing multiple alignment is given. Training a profile HMM is similar to make a multiple sequence alignment [82].

The commonly used HMM training algorithms are: Baum-Welch expectation maximization, gradient descent algorithms, Gibbs sampling, simulated annealing, genetic algorithm [82, 86, 87, 88, 89]. These training algorithms are local optimizers, thus, it is recommended to use a set of pre-aligned sequences [82]. Specially for complicated HMMs, the parameter space may be complex, with many local optima that trap a training algorithm.

Another important aspect of the HMMs is its architecture depicted by the number of states and how they are connected by state transitions. Profile HMMs and HMM-based gene finders have been the most successful applications of HMM's in computational biology [82, 87, 90, 91, 92, 93, 94]

4.1.5.1 Profile HMMs

Krogh and collaborators [95] proposed a HMM architecture to represent profiles of multiple sequence alignments. For each consensus column of the multiple alignment, a *match* state models the distribution of residues allowed in the column. An *insert* state and *delete* state at each column allow for insertion of one of more residues between that column and the next, of for deleting the consensus residue. Profile HMMs are strongly linear, left right models, unlike the general HMM case. **Figure** 4.11 shows a profile HMM corresponding to a short multiple sequence alignment.

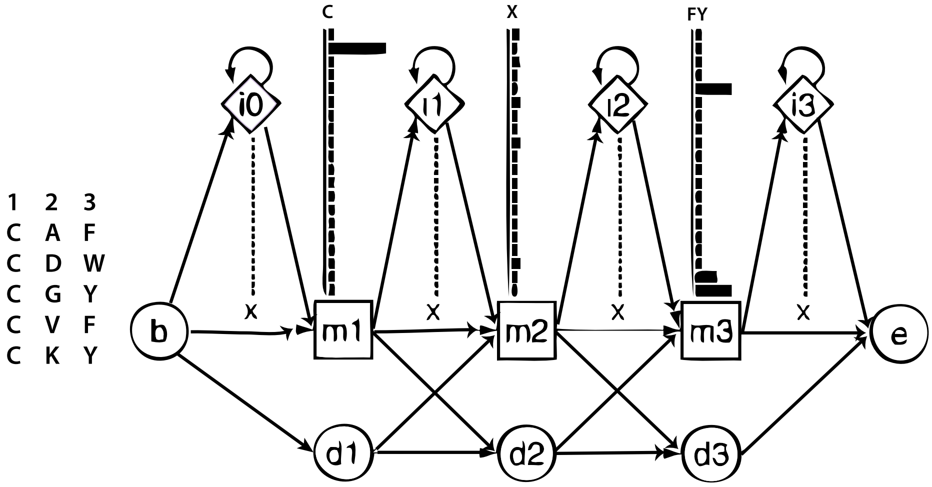


Figure 4.11: A small profile HMM (right) representing a short multiple alignment of five sequences (left) with three consensus columns. The three columns are modeled by three match states (squares labeled m1, m2 and m3), each of which has 20 residue emission probabilities, shown with black bars. Insert states (diamonds) also have 20 emission probabilities each. Delete states (circles) are null states that have no emission probabilities. A begin and end state are included (b,e). State transitions are shown as arrows.

The probability parameters in a profile HMM are usually converted to additive log-odds scores before the alignment and the score of a query sequence [96]. The scores for aligning a residue to a profile match state are comparable to the derivation of BLAST score. Suppose p_x as the probability of the match state emitting the residue x and the expected background frequency of residue x in the sequence data base is f_x, the score for the residue x at this match state is $log(\frac{p_x}{f_x})$

For other scores, profile HMM treatment diverges from the standard sequence alignment scoring. In the common gapped alignment, an insert of x residues is typically scored with a affine gap penalty, $a + b(x - 1)$, where, a is the score of the first residue and b is the score of each subsequent residue in the insertion. In a profile HMM, for an insertion of length x there is a state transition into an insert state which costs $log(t_{MI})$, where MI is the transition probability for moving from the match state to the insert state, $(x-1)$ state transitions for each subsequent insert state that costs $log(T_{IM})$. This is close to the traditional affine gap penalty, with the gap open cost $a = log(t_{MI} + log(t_{IM})$ and the gap extend cost as $b = log(t_{II})$.

Several HMM software packages have been implemented. The main difference between these packages is the model architecture that they adopt (**Figure** 4.12). There is a difference between a profile and motif models. The first class is related to insert and delete states associated to each match state, allowing insertion and deletion anywhere in a target sequence. By motif models are denominated by strings of match states (modeling ungapped blocks of sequence consensus) separated by a small number of insert states modeling the spaces between ungaped blocks. In the **Figure** 4.12 several architectures are shown, state transitions are shown as arrows and emission distributions are not represented.

SAM [97], HMMER [72], PFTOOLS [98] and HMMPro [99] implement models based on the original profile HMMs of Krogh [95]. These packages have augmented that simple model to deal with multiple domains, sequence fragments and local alignments (**Figure** 4.12). As an example the model architecture adopted by HMMER containing local versus global alignment which is not intrinsic on the algorithm. But, can be found as part of the model architecture [82]. Local alignments with respect to the model are allowed by non-zero state transition probabilities from a being state or internal match states, and from internal match states to an end state (dotted lines **Figure** 4.12). Local alignments with respect to the sequence are allowed by non-zero state transitions on the flanking insert states (shaded in the HMMER architecture in **Figure** 4.12).

These profile HMMs allow insertions and deletions anywhere in a sequence relative to the consensus model. They should be more sensitive than ungapped models. However, in practice the complexity of a model can overfit the training data and falling to generalize to other sequences. SAM and HMMER use mixture Dirichlet priors on most distributions to help to avoid overfitting and to limit the effective number of free parameters [100]. HMMER and PFTOOLS are used to build database search models from pre-existing alignments, such as in the Pfam and PROSITE profiles database. PROBE, META-MEME and BLOCKS assume motif models in which alignments consist of one or more ungapped blocks, separated by intervening sequences that are assumed to be random (**Figure** 4.12). The motif models can be viewed as special cases of profile HMMs; indeed, HMMER, SAM and PFTOOLS have various options for creating motif-like models [82].

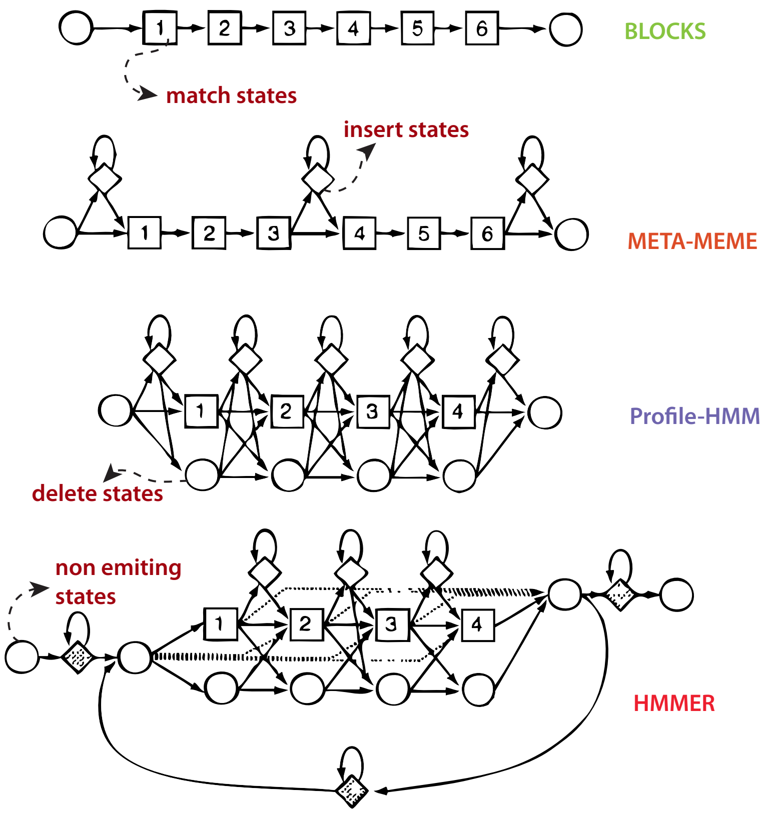


Figure 4.12: Different model architectures used in current methods.

4.2 Classification framework

This framework makes use of the profile HMMs Θ (described in Section 4.1) as features in order to train a classifier. Then, if there are similarities between profiles and query sequences, they should produce similar distributions. Thus, classification algorithms such as SVMs may discriminate those distributions and hence make an appropriate prediction.

A feature space can be viewed as the distribution of the set of profiles Θ along the query sequence S_p, in which, the profile-sequence relationship $P(S_p|h_j^c)$ is the probability that the profile j from the class c is part of the protein at a level of identity. Some proteins may enclose repeating motifs [63]. For instance, the ANK1 (Ankyrin 1, erythrocitic) protein involved in binding, consists of two alpha helices, repeated frequently from four to six times [101]. Then, if the protein S_p has a repeating motif, the score of its repeated profile h_j^c is set to the maximum likelihood among all repeats $\aleph(i,j,c) = max_{1..n}\{P(S_p|h_j^c\}$. In other words, when a profile is matched several times along the sequence, the most probable match is selected (see **Figure** 4.13). The matching profile is carried out by the software *HMMER* using the tool *hmmscan/hmmsearch*. Thus, the probability $P(S_i|h_j^c)$ is the mean posterior probability of aligned residues in the maximum expected accuracy alignment; a measure of how reliable the overall alignment is (from 0 to 1, with 1.00 indicating a completely reliable alignment according to the model) [72].

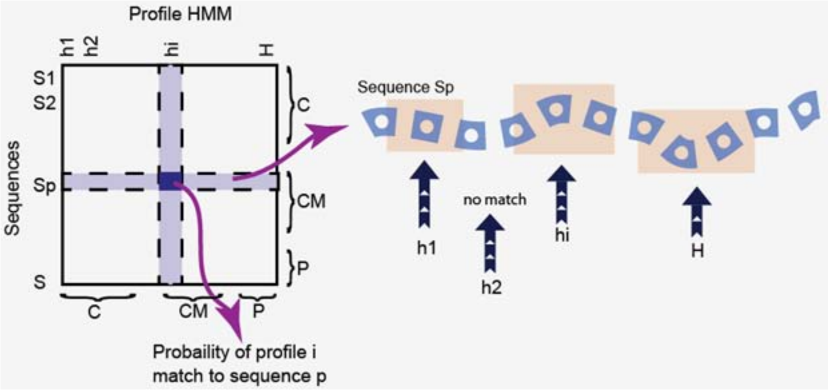
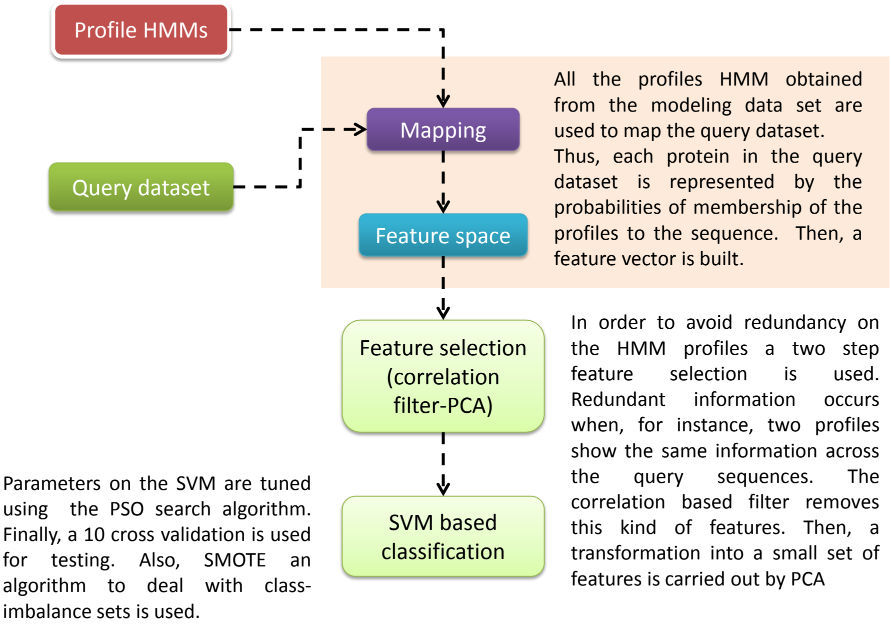


FIGURE 4.13: Representation space used to train and test the SVM. A query protein is depicted by a set of features (motifs). If the protein belongs to a specific compartment is expected that the associated profiles to the compartment will show a higher value than the other profiles



All the profiles HMM obtained from the modeling data set are used to map the query dataset. Thus, each protein in the query dataset is represented by the probabilities of membership of the profiles to the sequence. Then, a feature vector is built.

In order to avoid redundancy on the HMM profiles a two step feature selection is used. Redundant information occurs when, for instance, two profiles show the same information across the query sequences. The correlation based filter removes this kind of features. Then, a transformation into a small set of features is carried out by PCA

Parameters on the SVM are tuned using the PSO search algorithm. Finally, a 10 cross validation is used for testing. Also, SMOTE an algorithm to deal with class-imbalance sets is used.

FIGURE 4.14: Classification schema. A control dataset (query sequences) is used to test the method. Profile HMM are used as features for the SVMs.

Once the protein sequences from a control/validation data set are mapped into the profiles HMMs, a SVM is further used as a predictor with a 10 fold cross validation. Redundant information on the profile space is removed using a fast correlation-based filter algorithm [102]. In addition, Principal Component Analysis (PCA) is applied to this reduced space and the first 5 principal components are selected as meaningful (holding 80% of the explained variance). SVMs are designed following the one-against-all strategy which produces a strong class imbalance. Therefore, to avoid this class imbalance issue, the Synthetic Minority Over-sampling Technique SMOTE is employed [103]. Parameters of the SVM are tuned using the Particle Swarm Optimization algorithm [104, 105].

4.2.1 Support Vector Machines (SVMs)

Support vector machines are supervised learning models, introduced by Vapnik and coworkers in 1992 [106]. The basic idea of a SVM lies in the classification problem, a given input is predicted into several classes depending on the training of the SVMs. The learning problem setting for SVMs consists of an unknown and nonlinear dependency (mapping, function) $y = f(x)$ between some high-dimensional input vector x and the scalar output y. Where, there is not information about the underlying joint probability functions. The only information available is the training data set $D = \{(x_i, y_i) \in X * Y\}, i = 1, l$, where l is the number of the training data pairs and is equal to the size of the training data set D. Parameters for learning of the SVMs (selection, identification, estimation, training or tuning) are not predefined and their value depends on the training data used [107]. This is a basic paradigm of the structural risk minimization introduced by Vapnik and Chervonenkis and their coworkers leading to a new learning algorithm [107, 108, 109].

In the design of the SVM model is common to keep the value of training error fixed (approximation error, empirical error) and minimize the confidence interval. The resulting model should resolve the trade-off between under-fitting and over-fitting the training data. The final structure should ideally *match the learning machines capacity with training data complexity.* The risk functional R applied in the developing of support vector machines is described as:

$$R = \sum_{l}^{i=1} L_{\epsilon} + \Omega(l, h) \tag{4.11}$$

Where, For classification problems L_{ϵ} is a 0-1 loss function, Ω is a function bounding the capacity of the learning machine and h is the VapnikChervonenkis (VC) dimmension. In the simplest pattern recognition tasks, support vector machines use a linear separating hyperplane to create a classifier with a maximal margin. Then, the learning problem for the support vector machine will be cast as a constrained nonlinear optimization problem [107]. Thus, the cost function will be quadratic and the constraints linear. In cases when the classes cannot be linearly separated in the original input space, the SVM first transforms the original input space into a higher dimensional feature space. This transformation can be acquired by several nonlinear mappings as: polynomial, sigmoidal (multilayer perceptrons), radial basis functions (Gaussians, multiquadrics or different spline functions). After this nonlinear transformation step, the task of a support vector machine in finding the linear optimal separating hyperplane in this feature space is relatively trivial. In other words, the optimization problem to solve in a feature space will be of the same kind as the calculation of a maximal margin separating hyperplane in the original input space for linearly separable classes. Thus, after the specific nonlinear transformation, nonlinearly separable problems in input space can become linearly separable problems in a feature space [107].

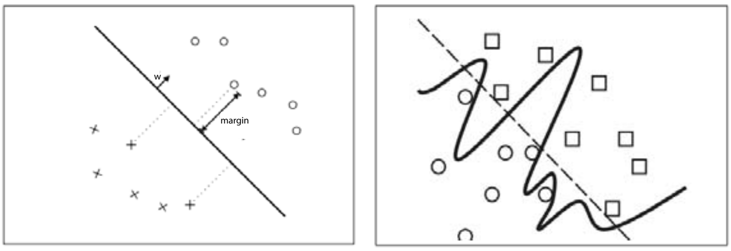


FIGURE 4.15: A simple linear support vector machine (Left). Overfitting in the case of linearly separable classification problem (empty circles and squares) (Right).

From the training dataset D, the input vector $x = \{x_1, ..., x_l\}$ in the space $X \subseteq \Re^d$ and their labels $y = \{y_1, ..., y_l\}$ are defined. Where, $y_i \in \{-1, 1\}$. In the simplest form SVMs where hyperplanes try to separate the training data by a maximal margin 4.15, all vectors lying on one side of the hyperplane are labeled as -1, and all vectors lying on the other side are labeled as 1 [110]. The training instances that lie closest to the hyperplane are called *support vectors.* In a more general definition, SVMs project the original training space X to a higher dimensional feature space Γ via the kernel operator K. Thus, the set of classifiers are depicted by:

$$f(x) = \sum_{i=1}^{n} \alpha K(x_i, x) \tag{4.12}$$

When K satisfies the Mercer's condition [111], it can be written as: $K(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u}).\phi(\mathbf{v})$, where, $\phi : X \longrightarrow \Gamma$ and "." denotes the inner product. Thus, f (4.12) can be expressed as:

$$f(x) = \mathbf{w}.\phi(x) \tag{4.13}$$

$$w = \sum_{i=1}^{n} \alpha_i \phi(x_i) \tag{4.14}$$

Thus, by K the training dataset is projected into a different (often higher dimension) feature space Γ. The SVM then computes the $\alpha_i s$ that correspond to the maximal margin hyperplane in γ. By choosing different kernel functions the training data can be projected into different spaces, for which their hyperplanes correspond to more complex boundaries in the original space X [107, 110, 112].

4.2.2 Selecting the best contact potential

Proteins from different classes have different properties as result of the environment in which they interact, making it difficult to characterize and discriminate. Therefore, with the aim to achieve the best representation per class, all 47 statistical contact potentials from AAindex are used independently. This means, each class is described by 47 representation/classifiers with their respective performances. Then, the ”best” predictor given by the highest performance is selected. This process does not make any inference about the training-testing or parameters on the SVM. It is just designed with the purpose of finding the best representation (statistical contact potential) for each class.

Chapter 5

Experimental framework

In this chapter the methodology described in the Chapter 4 have been evaluated for the prediction of subcellular localizations on Gram-Negative Bacterial proteins. Organisms such as *Proteobacteria*, Escherichia coli *(E. coli)*, *Salmonella, Shigella, Pseudomonas, Moraxella, Stenotrophomonas* and numerous others have been widely studied given their central role in several infections including pneumonia, bloodstream infections, wound or surgical site infections, and meningitis [113]. Gram-negative bacteria are resistant to multiple drugs and are increasingly resistant to most available antibiotics [114]. Thus, these bacteria have built-in abilities to find new ways to be resistant. Therefore, computational prediction of the subcellular localization of proteins is a valuable tool for genome analysis and annotation [6].

5.1 Protein subcellular localization on Gram-Negative bacterial proteins

Protein subcellular localizations can indicate how and what kind of cellular environments the proteins interact, helping to elucidate their function and role in biological processes [13]. Knowledge of the cellular compartment, where a protein probably resides, can help in the design of protein isolation experiments. For example, the identification of cell-surface-exposed proteins in a bacterial genome can help the discovery of therapeutic intervention points [56]. Experimental techniques such as immunolocalization, fluorescent and tagged isotopes are accurate, but they are slow and labor-intensive [115]. Besides, the high number of protein sequences with missing annotations makes it impossible to carry out experimental validations over all of them. To cope with this drawback, several computational approaches have been developed as an alternative to predict subcellular localizations on Gram negative bacterial proteins. In general, the classification of a new bacterial protein is performed by finding similar sequences, whose cellular compartments are experimentally determined. Usually, local alignments, such as BLAST or PSI-BLAST, are used to find a protein homologous to the unknown protein from public databases. In many cases, these methods perform accurate inferences [50, 116, 117, 118, 119]. However, it is well known that some proteins with a high sequence similarity do not achieve the same molecular function. As a result, homology based annotations have significant drawbacks, namely, propagation of errors, threshold relativity, and low sensitivity/specificity [120].

On the other hand, feature-based approaches model the differences between positive and negative samples by the extraction of protein properties arranged in a feature space, which is used as input to a classifier [1, 4, 6, 9, 12, 59]. In this regard, the feature extraction can be based on local information on the proteins. The purpose of the local featuring is the extraction of conserved protein subsequences known as motifs. However, this characterization has the following implications: not all proteins are related by the same motif, not all motifs are highly conserved, finding motifs is a very difficult task, and the presence of insignificant motifs may reduce the classifier performance [120]. Lastly, there are hybrid approaches consisting of the use of homology, theatre and local-feature based characterizations. Among these approaches to classify bacterial proteins, the following are widely used: PSORTb v.3 [6], CELLO [121], PSLpred [116], LOCtree [55], P-CLASSIFIER [122], and GNeg-mPLoc [3], which cover different types of algorithms such as support vector machines (SVM), amino-acid composition, Bayesian networks, signal peptides, motif matching, homology based prediction, hidden Markov models (HMM), and text labeling. In general terms, they all report adequate performance, but, in spite of the low false positive rate in most of them, a high false negative rate remains. The main limitation of the listed methods is the protein representations that they adopt, for instance: Psortb, CELLO, LOCtree, P-classifier and GNEg-mPloc use feature vectors based on different classes of amino acid compositions such as n-grams, physio-chemical properties of the amino acids and Gene Ontology annotation. This kind of characterization misplaces the distribution of the polypeptide chain, and therefore loses the information contained by the continuous segments of the amino acid sequence known as protein domains.

5.2 Protein control-evaluation data sets for Gram negative bacteria

Five subcellular localizations have been selected from Gram negative bacterial proteins. The modeling data set comprises 500 cytoplasmic proteins (**C**), 500 inner membrane proteins (**CM**), 359 periplasmic proteins (**P**), 349 outer membrane proteins (**OM**), and 288 extracellular proteins (**E**) selected from ePSORTdb [6], omitting sequences with an identity superior to 60%. As a control dataset, the reported in [56] is used for testing purposes. This dataset holds 299 protein sequences distributed as follows: 145 cytoplasmic proteins, 69 cytoplasmic membrane proteins, 29 periplasmic proteins, 38 outer membrane proteins, and 18 extracellular proteins. In addition, any proteins sharing <60% identity of modeling data set with respect to the control dataset were removed. All identity filters were carried out using the cdHit software [123].

Chapter 6

Results and discussion

In order to test the performance of the wavelet approach, three different views have been analyzed: 1) Performance evaluation of three currently active services for subcellular localization prediction in Gram-negative bacteria: Psortb, CELLO, and SOSUIGramN; 2) Assessment of classical protein representations; 3) A subsequence/profile based approach that is the closest approach to the one proposed here (the difference lies on how the motifs-features are obtained). In the 2-3) views, all evaluations are carried out following the same classification strategy proposed in this paper.

- The methods CELLO version 2.5 and SOSUIGramN (web versions) are used to predict the subcellular localizations in addition to the standalone version of Psortb V3.0.2 (stand alone version). In order to ensure that the PsortB performance is not biased, the modeling data set is used in the blast module and the predictions are accomplished with the control data set. The presence of a sequence in the training data set of any of the listed servers would cause biasing on their classifications, so, it is necessary to clarify that it is not possible to verify whether test sequences are in the training set of CELLO as well as in SOSUIGramN servers.

- Many protein feature descriptors have been developed for protein prediction, such as amino acid composition, pseudo-amino acid composition, amphiphilic pseudo amino acid composition, autocorrelation, Composition/Transition/Distribution (CTD) descriptors, quasi-sequence order descriptor, among others (see Table 6). As a result, three tests based on global feature descriptors have been formulated for evaluating the hypothesis whether proteins can be depicted more accurately using local features (motifs, domains or sites) than by the use of global features (molecular weight, amino acid composition). This formulation proves the statement that global descriptors such as amino acid composition dispel the protein structure. In order to evaluate the quality of the characterization and make an accurate comparison, the same classification strategy is adopted for wavelet and global featuring approaches

- The Pfam profile test is constructed with the purpose of demonstrating the discriminant ability of the wavelet HMM profiles over a set of annotated HMM terms. Thus, the testing data set is submitted to Pfam sequence batch search (Parameters: E-value=20, PfamA, PfamB) allowing a wide range of matches. A total of 646 Pfam profiles (domains, motifs and families) were found. All profiles are used as features to evaluate the pfam-profile discriminance. Performance predictions are carried out in the same way as in the wavelet method. Pfam HMMs were treated by the same process as it is described in section:4.2.

Featuring type	**Shortcut**	**No Features**	**Description**
Amino acid composition	ACC,DC,TC	8420	frequency of the 1,2 and 3 n-grams
Correlation descriptor	Geary, Moran, Moreau	720	Autocorrelation
Pseudo amino acid composition	PAAC,APAAC	80	Pseudo Amino Acid Composition
Sequence-Order descriptor	QSO,SOCN	160	Quasi-Sequence-Order descriptors
Triad descriptor	Triad	343	Triad method abstracts

Table 6.1: Different sets of global features extracted to the proteins

In order to estimate the performance prediction of the different methods, three classifier performance parameters are used, sensitivity (S_n), specificity (S_p), and Matthews Correlation Coefficient (MCC), respectively, given as follows:

$$S_n = \frac{T_P}{T_P + F_N} \tag{6.1}$$

$$S_p = \frac{T_N}{T_N + F_P} \tag{6.2}$$

$$MCC = \frac{T_P * T_N - F_P * F_N}{\sqrt{(T_P + F_N)(T_P + F_P)(T_N + F_P)(T_N + F_N)}} \tag{6.3}$$

where notations T_P, F_P, T_N and F_N stand for the true positive, false positive, true negative, and false negative values, respectively. Performance prediction of the wavelet-based method across all contact potentials is shown in Figure:6. As seen from the obtained results, the best contact potentials per class are shown in Table 6.1 and the corresponding comparisons are presented in Table 6.3

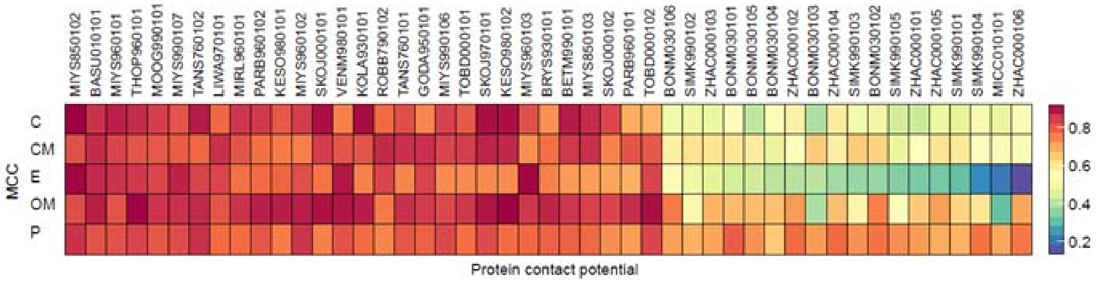


Figure 6.1: Performance predictions over all subcellular localizations using all 47 contact potentials. Each box depicts the MCC for the specific SCL and contact potential. Accurate predictions are highlighted in dark red scales whereas wrong predictions are marked with blue scales. For each SCL the best descriptor (statistical contact potential) is selected.

6.1 Comparison with cutting edge predictors

Performance of the individual methods reveals that the wavelet approach achieves the highest overall sensitivity. This fact is highlighted in cytoplasmic, periplasmic and extracellular localizations, in which our method significantly outperforms, by more than 10% the psort, SOSUIGramN, and CELLO methods. Also, the specificities for these classes are basically the same in all methods showing that the wavelet approach can improve the true positive rate while holding a low false positive rate (see Table 6.3). For cytoplasmic proteins, our approach shows the highest sensitivity (0.94) followed by CELLO (0.93), SOSUIGramN (0.9) and psort (0.82). However, both CELLO and SOSUIGramN have a low specificity (0.83 and 0.88, respectively), that is, this performed value can be interpreted as a high false positive rate. For this reason, we infer that the proposed method achieves the best MCC performance of 0.85 in cytoplasmic proteins followed by psort 0.79, SOSUIGramN 0.78 and CELLO 0.76. A protein can remain in the cytoplasm or be targeted into different sites by a transport system, thus, proteins associated to the cytoplasm localization are highly diverse and comprise a big variety of domains. This diversity is also the case of transmembrane proteins, which are simultaneously located on both sides of the membrane and transport molecules from one side to the other, making it difficult to characterize these kinds of proteins through local features or motifs. Accordingly, both cytoplasmic membrane and outer membrane are the classes with the lowest performances of sensitivity in comparison to psortb and SOSUIGramN, respectively. For cytoplasmic membrane PsortB shows an upper sensitivity of 5% better compared to the wavelet-based method while the specificity remains nearly the same and the MCC is close to ours (2%) making the prediction comparable. For outer membrane proteins, SOSUIGramN achieves the best sensitivity of 92% followed by our method and psortb with a sensitivity of 0.81 and 0.87, respectively. SOSUIGramN achieves a MCC of 0.92, a 5% upper than wavelet method with a 0.87.

SCL	Location	aaindex ID	Description
C	Cytoplasm	TANS760102	Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins
CM	Cytoplasmic Membrane	THOP960101	Mixed quasichemical and optimization-based protein contact potential
P	Periplasm	MIYS960101	Quasichemical energy of transfer of amino acids from water to the protein environment
OM	Outer Membrane	SKOJ970101	Statistical potential derived by the quasichemical approximation. Derivation and testing of pair potentials for protein folding
E	Extracell	MIYS850102	Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures Quasi-Chemical Approximation

Table 6.2: Best statistical contact potential per subcellular localization

Statistical contact potentials turn to be out a useful representation of the proteins and are used to subtract local features by the continuous wavelet transform. Thus, if a protein set has similar interactions among adjacent amino acids at any position in the proteins, the wavelet transform can efficiently detect those interactions. Unlike SOSUIGramN, CELLO, and Psortb in which several types of protein representations had been proposed (amino acid composition, partitioned amino acid composition, local amino acid composition, SCL-blast, signal peptides, N and C-terminal composition, profile motifs among others), the wavelet-based method involves just the local feature representation. Psortb uses a set of known profile motifs per subcellular localization in contrast to the proposed method which generates its own set of profiles. SOSUIGramN consists of a set of filters, in which proteins are divided into ten segments, and the average values of physiochemical properties over those segments are computed. CELLO divides the sequence into subsequences of equal length and each partition is encoded by a particular amino acid composition. On the other hand, the proposed method uses the protein contact potential plus the wavelet transform to detect core local features encoded by the amino acid sequence, thus making use of the main protein information contained in the amino acid distribution. The results obtained are in accordance with previous works suggesting that the wavelet transform is a powerful tool for motif detection and characterization [63, 68].

6.2 Comparison with global featuring methods

Global featuring has been widely used to describe proteins from their primary arrangement. For example, amino acid composition has shown to characterize well intra and extra cellular proteins, where aliphatic and charged residues occur more frequently in intracellular proteins than extracellular proteins [124]. In our study as table 3 shows, amino acid composition and pseudo amino acid composition have shown good performance (MCC) compared to correlation, sequence order and triad descriptors for cytoplasmic, cytoplasmic membrane, and outer membrane compartments. Cytoplasmic was the cellular compartment with the highest MCC score (0.69) compared against all the global featuring methods. However, the wavelet method has shown a performance (MCC) of 0.85. In general, none of the global featuring results are comparable with our proposed method and nor with the cutting edge predictors.

6.3 Comparison with a profile based method

Pfam is a large collection of protein families represented by multiple sequence alignments and Hidden Markov Models. From these families 14831 have been manually curated (Pfam A) and 80% of all proteins in Uniprot contain a match , at least, to the Pfam domain [11]. Based on the prediction performance, the wavelet method achieves the highest performance over all subcellular compartments compared to the profile method (Table 6.3), suggesting the wavelet-profiles as potential features to describe Gram negative proteins better than the annotated pfam terms.

	Wavelet			**PsortB**			**CELLO**			**SOSUIGRAMN**		
SCL	Sens	Spec	MCC	Sens	Spec	MCC	Sens	Spec	MCC	Sens	Spec	MCC
C	0.94	0.91	0.85	0.82	0.97	0.79	0.93	0.83	0.76	0.9	0.88	0.78
CM	0.77	0.99	0.83	0.82	0.98	0.85	0.62	0.99	0.72	0.72	0.99	0.8
P	0.93	0.99	0.89	0.79	0.99	0.84	0.5	0.97	0.57	0.59	0.98	0.64
OM	0.87	0.98	0.85	0.81	1	0.89	0.55	0.96	0.56	0.92	0.99	0.91
E	0.83	0.99	0.88	0.77	0.99	0.81	0.44	0.96	0.38	0.5	0.99	0.66
	AAC,DC,TC			**Triad**			**Geary,Moran,Moreau**			**PAAC,APAAC**		
SCL	Sens	Spec	MCC	Sens	Spec	MCC	Sens	Spec	MCC	Sens	Spec	MCC
C	0.83	0.7	0.53	0.76	0.73	0.49	0.83	0.73	0.56	0.86	0.82	0.69
CM	0.58	0.95	0.59	0.59	0.87	0.46	0.68	0.82	0.46	0.57	0.83	0.38
P	0.68	0.7	0.25	0.72	0.67	0.24	0.34	0.79	0.09	0.72	0.64	0.22
OM	0.74	0.92	0.58	0.74	0.8	0.41	0.84	0.85	0.55	0.79	0.87	0.54
E	0.89	0.81	0.39	0.72	0.8	0.29	0.83	0.83	0.4	0.72	0.81	0.29
	QSO,SOCN			**Pfam**								
SCL	Sens	Spec	MCC	Sens	Spec	MCC						
C	0.73	0.58	0.31	0.8	0.76	0.56						
CM	0.52	0.77	0.27	0.92	0.62	0.46						
P	0.68	0.58	0.16	0.55	0.74	0.19						
OM	0.58	0.79	0.28	0.68	0.8	0.37						
E	0.27	0.61	0.05	0.5	0.71	0.1						

Table 6.3: Performance prediction of the wavelet method and the different tests.

Our results suggest this novel characterization as a powerful tool for representing protein sequences from their primary structures (Table 6.3). The contact potential MIYS850102 has shown a good efficiency over all subcellular compartments (Figure:6). This potential depicts the contact energies between residues in globular proteins estimated from the amount of residue-residue contacts observed in protein crystal structures by regarding them as statistical averages in the quasi-chemical approximation [62]. The contact potential ZHAC000106 has shown a regular performance prediction in all subcellular localizations as shown in Figure:6 except for extracellular medium in which the performance (MCC) rounds the 0.2. This potential characterizes the pairwise amino acid interactions limited to the context of secondary structural environments (helix, strand and coil) making this potentially less sensitive to the pairwise amino acid interactions than the previous one. Our results based on contact potentials from sequence-dependent, sequence-independent features and structural classes show average performances (BONM,ZHAC,SIMK,MICC) [62] for all classes, except for extracellular medium. On the other hand, contact potentials based on quasi-chemical energy approximations (MIYS,BASU,THOP,TANS,LIWA, MOOG, SKOJ, KOLA) have shown good performance (MCC ¿ 0.7). In this analysis, a set of contact potentials that described each subcellular localization with high levels of accuracy are found (Figure:6) of those, only the best contact potential per class is selected (Table:6.1). Yet, it is worth noting that this selection process is out of the SVM classification strategy and does not interfere in the tuning and evaluation process.

Chapter 7

Conclusions

In this thesis a methodology for protein annotation prediction have been implemented employing diverse tools such as statistical contact potentials, continuous wavelet transform, hidden Markov models, dynamic clustering and support vector machine classifiers. The pairwise statistical contact potentials are representations of the interactions of the amino acids in a set of related proteins by its structural configurations. This information is useful for inferring the function of a protein from its tertiary structure because they model the geometry of the proteins. On the other hand, there are a lot of proteins that contain only the primary structure information. In this context, the proposed method uses the structural information comprised on the contact potentials and then describe the interactions of the arrangement of amino acid sequences. With the aim to decode the data given by this representation, the wavelet transform allows the identification of patterns localized at specific positions. Thus, motifs that are relevant in the conformation of the protein structures can be easily identified. The wavelet method leads a powerful tool for protein motif identification. In this thesis the focus of the problem is the characterization of proteins and the prediction of them in different terms using the motifs (profiles) as features. However, this is an starting point, applications of this method can be extended to many problems including: motif-domain detection, structural prediction, even hints of folding. All of them are a conception of future work.

The evaluation of the method is carried out on a set of subcellular localizations on Gram-Negative bacterial proteins. Results suggest the method as a powerful tool for the characterization of proteins. Also, each subcellular localization has an specific contact potential which can be inferred as a correlation between the structural information and the localization by the motifs on the sequences. Specifically, for the five major subcellular localizations in Gram-negative bacterial proteins, the wavelet method shows the best performance prediction by decreasing the false negative rate whereas the false positive rate is maintained.

One of the main advantages of the method is its capability to find correlated and variable length motifs which usually give clues about the protein function. The variability on length of the subsequences and the posterior alignment by CLUSTAL allows the possibility to merge similar motifs even if the motif-sequences have gaps or mutations. For instance, a sequence of length 20 is found in some protein and a couple of motifs of length 5 are found in another protein, if the small sequences have a correspondence on the big sequence, they are merged, then the profile HMM model the motif as a big motif with an enrichment in the positions in which the small sequences match with the big one. Also, profiles HMM are a powerful tool to represent motifs in divergent sequences.

In conclusion, The proposed contact-potential characterization is an alternative to the classic models based on the amino acid composition and physiochemical properties due to the use of structural information from the potentials over primary protein structures. The wavelet method, unlike Psortb, CELLO and SOSUIGramN, uses only one protein characterization. Thus, in terms of prediction, an implementation of other representations such as physiochemical properties, amino acid composition, or homology modules as blast can be used in order to improve the performance of the predictions.

Chapter 8

Appendix A: Wavelet Transform

8.1 Properties of the Continuous Wavelet Transform

From the chapter 4 equation 4.7 can be inferred that the CWT is a linear transformation, characterized by the next properties.

Superposition Property Suppose $x(t)$, $y(t)$ in the euclidean space of $L^2(R)$ and k_1 and k_2 are constants. If the CWT of $x(t)$ is $wt_x(s,\tau)$ and the CWT of $y(t)$ is $wt_y(s,\tau)$, then the CWT of $z(t) = k_1 x(t) + k_2 y(t)$ can be defined by:

First $wt_x(s,\tau)$, $wt_y(s,\tau)$ and $wt_z(s,\tau)$ are defined as:

$$wt_x(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t)\psi^*(\frac{t-\tau}{s})dt \tag{8.1}$$

$$wt_y(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} y(t)\psi^*(\frac{t-\tau}{s})dt \tag{8.2}$$

$$wt_z(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} z(t)\psi^*(\frac{t-\tau}{s})dt \tag{8.3}$$

$$\tag{8.4}$$

Then, replacing $z(t)$ in wt_z, we have:

$wt_z(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} [k_1 x(t) + k_2 y(t)]\psi^*(\frac{t-\tau}{s})dt$
$wt_z(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} k_1 x(t)\psi^*(\frac{t-\tau}{s})dt + \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} k_2 y(t)\psi^*(\frac{t-\tau}{s})dt$

Thus the expression is reduced to:

$$wt_z(t) = k_1 wt_x(t) + k_2 wt_y(t) \tag{8.5}$$

Which corresponds to the superposition property of the CWT.

Covariant under translation Suppose the CWT of $x(t)$ is $wt_x(s,\tau)$; then the CWT of $x(t-t_0)$ is $wt_x(s,\tau-t_0)$ In order to demonstrate this property let's:

$x'(t) = x(t-t_0)$, then: $wt_{x'}(s,\tau) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t-t_0)\psi^*(\frac{t-\tau}{s})dt$

and $t' = t - t_0$
$wt_{x'}(s,\tau) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t')\psi^*(\frac{t'+t_0-\tau}{s})dt$
Thus, the CWT of $x(t-t_0)$ is $wt_x(s,\tau-t_0)$. This means that the wavelet coefficients of $x(t-t_0)$ can be obtained by translating the wavelet coefficients of $x(t)$ along the time axis with t_0

Covariant under dilation Suppose the CWT of $x(t)$ is $wt_x(s,\tau)$, let $x'(t) = x(\frac{t}{a})$, then:

$$wt_{x'}(s,\tau) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(\frac{t}{a})\psi^*(\frac{t-\tau}{s})dt \tag{8.6}$$

let $t' = \frac{t}{a}$; then (8.6) can be expressed as:

$$wt_{x'}(s,\tau) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t')\psi^*(\frac{at'-\tau}{s})d(at') \tag{8.7}$$

$$wt_{x'}(s,\tau) = \frac{\sqrt{a}}{\sqrt{s}}\int_{-\infty}^{\infty} x(t')\psi^*(\frac{at'-\frac{\tau}{a}}{\frac{s}{a}})d(at') = \sqrt{a}wt_x(\frac{s}{a},\frac{\tau}{a}) \tag{8.8}$$

Equation 8.8 indicates that, when a signal is dilated by a, its corresponding wavelet coefficients are also dilated by a along both the scale and time axes.

Moyal Principle Suppose $x(t)$, $y(t)$. If the CWT of $x(t)$ is $wt_x(s,\tau)$ and the CWT of $y(t)$ is $wt_y(s,\tau)$; that is,

$$wt_x(s,\tau) = \langle x(t), \psi_{s,\tau}(t)\rangle \tag{8.9}$$

$$wt_y(s,\tau) = \langle y(t), \psi_{s,\tau}(t)\rangle \tag{8.10}$$

Then,

$$\langle wt_x(s,\tau), wt_y(s,\tau)\rangle = C_\psi \langle x(t), y(t)\rangle \tag{8.11}$$

where $C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df$ is the admissibility condition of the wavelet.

8.2 Mother Wavelets Commonly Used

This section introduces several commonly used mother wavelets for performing the continuous wavelet transform analysis. **Mexican Hat Wavelets** The mexican hat wavelet is a normalized, second derivative of a Gaussian function, defined as [63, 66, 68]:

$$\psi(t) = \frac{1}{\sqrt{2\pi}\sigma^3}\left(1 - \frac{\sigma^2}{t^2}\right) e^{\frac{-t^2}{2\sigma^2}} \tag{8.12}$$

σ is a width constant.

The Mexican hat wavelet is frequently called the Ricker wavelet in geophysics, where frequently is used to model seismic data [65, 67].

Morlet Wavelet The Morlet wavelet is defined as [65, 67]

$$\psi_M(t) = \frac{1}{\sqrt{\pi f_b}} e^{j2\pi f_c t} e^{-\frac{t^2}{f_b}} \tag{8.13}$$

where f_b is the bandwidth parameter and f_c denotes the wavelet center frequency.

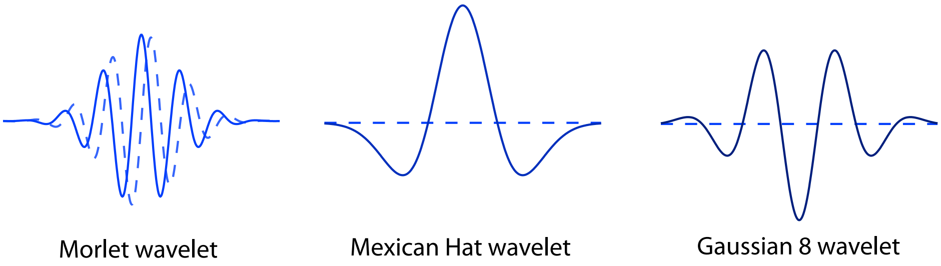


FIGURE 8.1: Different wavelets used on the continuous wavelet decomposition, Morlet, Mexican Hat and Gaussian wavelets

The Morlet wavelet has been widely used for identifying transient components embedded in a signal, for example, bearing defect-induced vibration [65, 67].

Gaussian wavelet

The Gaussian function is expressed by (Teolis 1998):

$$f(t) = e^{-jt}e^{-t^2} \tag{8.14}$$

If the Nth derivative of this function is taken the wavelet can be expressed as:

$$\psi_G = C_N \frac{d^{(N)} f(t)}{dt^N} \tag{8.15}$$

Where N is an integer parameter (≥ 1) and denotes the order of the wavelet, and C_N is a constant introduced to ensure that $||f^{(N)}(t)||^2 = 1$. The Gaussian wavelet is often used for characterizing singularity that exists in a signal [63, 65, 67]). In the **Figure** 8.1 are illustrated the wavelet forms of some of the common used mother wavelets.

Chapter 9

Appendix B: List of Pairwise Contact Potentials

Potential	Description
TANS760101	Statistical contact potential derived from 25 x-ray protein structures
TANS760102	Number of contacts between side chains derived from 25 x-ray protein structures
ROBB790102	Interaction energies derived from side chain contacts in the interiors of known protein
BRYS930101	Distance-dependent statistical potential (only energies of contacts within 0-5 Angstro
THOP960101	Mixed quasichemical and optimization-based protein contact potential
MIRL960101	Statistical potential derived by the maximization of the harmonic mean of Z scores
VENM980101	Statistical potential derived by the maximization of the perceptron criterion
BASU010101	Optimization-based potential derived by the modified perceptron criterion
MIYS850102	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS850103	Quasichemical energy of interactions in an average buried environment
MIYS960101	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS960102	Quasichemical energy of interactions in an average buried environment
MIYS960103	Number of contacts between side chains derived from 1168 x-ray protein structures
MIYS990106	Quasichemical energy of transfer of amino acids from water to the protein environment
MIYS990107	Quasichemical energy of interactions in an average buried environment
LIWA970101	Modified version of the Miyazawa-Jernigan transfer energy
KESO980101	Quasichemical transfer energy derived from interfacial regions of protein-protein compl
KESO980102	Quasichemical energy in an average protein environment derived from interfacial regions
MOOG990101	Quasichemical potential derived from interfacial regions of protein-protein complexes
BETM990101	Modified version of the Miyazawa-Jernigan transfer energy
TOBD000101	Optimization-derived potential obtained for small set of decoys
TOBD000102	Optimization-derived potential obtained for large set of decoys
PARB960101	Statistical contact potential derived by the quasichemical approximation
PARB960102	Modified version of the Miyazawa-Jernigan transfer energy
KOLA930101	Statistical potential derived by the quasichemical approximation
GODA950101	Quasichemical statistical potential derived from buried contacts
SKOJ970101	Statistical potential derived by the quasichemical approximation
SKOJ000101	Statistical quasichemical potential with the partially composition-corrected pair scale
SKOJ000102	Statistical quasichemical potential with the composition-corrected pair scale
BONM030101	Quasichemical statistical potential for the antiparallel orientation of interacting sid
BONM030102	Quasichemical statistical potential for the intermediate orientation of interacting sid
BONM030103	Quasichemical statistical potential for the parallel orientation of interacting side gr
BONM030104	Distances between centers of interacting side chains in the antiparallel orientation
BONM030105	Distances between centers of interacting side chains in the intermediate orientation
BONM030106	Distances between centers of interacting side chains in the parallel orientation
MICC010101	Optimization-derived potential
SIMK990101	Distance-dependent statistical potential (contacts within 0-5 Angstrooms)
SIMK990102	Distance-dependent statistical potential (contacts within 5-7.5 Angstrooms)
SIMK990103	Distance-dependent statistical potential (contacts within 7.5-10 Angstrooms)
SIMK990104	Distance-dependent statistical potential (contacts within 10-12 Angstrooms)
SIMK990105	Distance-dependent statistical potential (contacts longer than 12 Angstrooms)
ZHAC000101	Environment-dependent residue contact energies (rows = helix, cols = helix)
ZHAC000102	Environment-dependent residue contact energies (rows = helix, cols = strand)
ZHAC000103	Environment-dependent residue contact energies (rows = helix, cols = coil)
ZHAC000104	Environment-dependent residue contact energies (rows = strand, cols = strand)
ZHAC000105	Environment-dependent residue contact energies (rows = strand, cols = coil)
ZHAC000106	Environment-dependent residue contact energies (rows = coil, cols = coil)

Table 9.1: List of the 47 pairwise contact potentials from AAindex database

Bibliography

[1] Inna Dubchak, Ilya Muchnik, Stephen R Holbrook, and Sung-Hou Kim. Prediction of protein folding class using global description of amino acid sequence. *Proceedings of the National Academy of Sciences*, 92(19):8700–8704, 1995.

[2] Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

[3] Kuo-Chen Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and Biophysical Research Communications*, 278(2):477–483, 2000.

[4] Manoj Bhasin and GPS Raghava. Eslpred: Svm-based method for subcellular localization of eukaryotic proteins using dipeptide composition and psi-blast. *Nucleic acids research*, 32(suppl 2):W414–W419, 2004.

[5] Deepak Sarda, Gek H Chua, Kuo-Bin Li, and Arun Krishnan. pslip: Svm based protein subcellular localization prediction using multiple physicochemical properties. *Bmc Bioinformatics*, 6(1):152, 2005.

[6] Y Yu Nancy, Matthew R Laird, Cory Spencer, and Fiona SL Brinkman. Psortdban expanded, auto-updated, user-friendly protein subcellular localization database for bacteria and archaea. *Nucleic acids research*, 39(suppl 1):D241–D244, 2011.

[7] Hiroshi Nakashima and Ken Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of Molecular Biology*, 238(1):54–61, 1994.

[8] Xiao Nan, Dongsheng Cao, Qingsong Xu, and Yizeng Liang. protr: Protein sequence feature extraction with r.

[9] Rabie Saidi, Mondher Maddouri, and Engelbert Mephu Nguifo. Protein sequences classification by means of feature extraction with substitution matrices. *BMC bioinformatics*, 11(1):175, 2010.

[10] Kay Hofmann, Philipp Bucher, Laurent Falquet, and Amos Bairoch. The prosite database, its status in 1999. *Nucleic Acids Research*, 27(1):215–219, 1999.

[11] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[12] GA Arango-Argoty, JF Ruiz-Munoz, JA Jaramillo-Garzon, and CG Castellanos-Dominguez. An adaptation of pfam profiles to predict protein sub-cellular localization in gram positive bacteria. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5554–5557. IEEE, 2012.

[13] Kuo-Chen Chou and Hong-Bin Shen. Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols*, 3(2):153–162, 2008.

[14] Kuo-Chen Chou and Hong-Bin Shen. Plant-mploc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PloS one*, 5 (6):e11335, 2010.

[15] Paul D Thomas and Ken A Dill. Statistical potentials extracted from protein structures: how accurate are they? *Journal of molecular biology*, 257(2):457–469, 1996.

[16] Piotr Pokarowski, Andrzej Kloczkowski, Robert L Jernigan, Neha S Kothari, Maria Pokarowska, and Andrzej Kolinski. Inferring ideal amino acid interaction forms from statistical protein contact potentials. *PROTEINS: Structure, Function, and Bioinformatics*, 59(1):49–57, 2005.

[17] Harvey Lodish. *Molecular cell biology*. Macmillan, 2008.

[18] Avinash Upadhyay and Kakoli Upadhyay. *MOLBIO Fundamentals of Molecular Biology*. 2010.

[19] Hasker P Davis and Larry R Squire. Protein synthesis and memory: a review. *Psychological bulletin*, 96(3):518, 1984.

[20] Steven K Hanks and Tony Hunter. Protein kinases 6. the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *The FASEB Journal*, 9(8):576–596, 1995.

[21] R John Ellis and Saskia M Van der Vies. Molecular chaperones. *Annual review of biochemistry*, 60(1):321–347, 1991.

[22] Scott A Burchett. Regulators of g protein signaling. *Journal of neurochemistry*, 75(4):1335–1351, 2000.

[23] Iain D Campbell and A Kristina Downing. Building protein structure and function from modular units. *Trends in biotechnology*, 12(5):168–172, 1994.

[24] A Klug and JW Schwabe. Protein motifs 5. zinc fingers. *The FASEB journal*, 9 (8):597–604, 1995.

[25] Paul S Freemont, Isabel M Hanson, and John Trowsdale. A novel gysteine-rich sequence motif. *Cell*, 64(3):483–484, 1991.

[26] Bostjan Kobe and Andrey V Kajava. The leucine-rich repeat as a protein recognition motif. *Current opinion in structural biology*, 11(6):725–732, 2001.

[27] Alan Wee-Chung Liew, Hong Yan, and Mengsu Yang. Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recognition*, 38 (11):2055–2073, 2005.

[28] Susana Vinga and Jonas Almeida. Alignment-free sequence comparisona review. *Bioinformatics*, 19(4):513–523, 2003.

[29] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.

[30] Desmond G Higgins and Paul M Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, 1988.

[31] Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. Multiple sequence alignment with the clustal series of programs. *Nucleic acids research*, 31(13):3497–3500, 2003.

[32] David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3):275–282, 1992.

[33] Rohit Singh, Jinbo Xu, and Bonnie Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. In *Research in computational molecular biology*, pages 16–31. Springer, 2007.

[34] Xiaoqui Huang. On global sequence alignment. *Computer applications in the biosciences: CABIOS*, 10(3):227–235, 1994.

[35] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[36] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217, 2000.

[37] Ralf Bundschuh. Rapid significance estimation in local sequence alignment with gaps. *Journal of Computational Biology*, 9(2):243–260, 2002.

[38] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3): 403–410, 1990.

[39] Henrik Nielsen, Søren Brunak, and Gunnar von Heijne. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 12(1):3–9, 1999.

[40] Thomas G Dietterich. Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pages 15–30. Springer, 2002.

[41] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[42] Sotiris B Kotsiantis, ID Zaharakis, and PE Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[43] Pat Langley and Herbert A Simon. Applications of machine learning and rule induction. *Communications of the ACM*, 38(11):54–64, 1995.

[44] Krzysztof J Cios, Witold Pedrycz, and RM Swiniarsk. Data mining methods for knowledge discovery. *Neural Networks, IEEE Transactions on*, 9(6):1533–1534, 1998.

[45] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[46] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

[47] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25 (17):3389–3402, 1997.

[48] Deborah Goldman, Sorin Istrail, and Christos H Papadimitriou. Algorithmic aspects of protein structure similarity. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 512–521. IEEE, 1999.

[49] Gaurav Pandey, Vipin Kumar, and Michael Steinbach. Computational approaches for protein function prediction: A survey. *Twin Cities: Department of Computer Science and Engineering, University of Minnesota*, 2006.

[50] Y Yu Nancy, James R Wagner, Matthew R Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S Cenk Sahinalp, Martin Ester, Leonard J Foster, et al. Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615, 2010.

[51] Y Yu Nancy, Matthew R Laird, Cory Spencer, and Fiona SL Brinkman. Psortdban expanded, auto-updated, user-friendly protein subcellular localization database for bacteria and archaea. *Nucleic acids research*, 39(suppl 1):D241–D244, 2011.

[52] Xing-Ming Zhao, Luonan Chen, and Kazuyuki Aihara. Protein function prediction with high-throughput data. *Amino Acids*, 35(3):517–530, 2008.

[53] Agnieszka S Juncker, Lars J Jensen, Andrea Pierleoni, Andreas Bernsel, Michael L Tress, Peer Bork, Gunnar Von Heijne, Alfonso Valencia, Christos A Ouzounis, Rita Casadio, et al. Sequence-based feature prediction and annotation of proteins. *Genome Biol*, 10(2):206, 2009.

[54] Jannick Dyrløv Bendtsen, Henrik Nielsen, Gunnar von Heijne, and Søren Brunak. Improved prediction of signal peptides: Signalp 3.0. *Journal of molecular biology*, 340(4):783–795, 2004.

[55] Rajesh Nair and Burkhard Rost. Mimicking cellular sorting improves prediction of subcellular localization. *Journal of molecular biology*, 348(1):85–100, 2005.

[56] Jennifer L Gardy and Fiona SL Brinkman. Methods for predicting bacterial protein subcellular localization. *Nature Reviews Microbiology*, 4(1):741–751, 2006.

[57] Inna Dubchak, Ilya Muchnik, Christopher Mayor, Igor Dralyuk, and Sung-Hou Kim. Recognition of a protein fold in the context of the scop classification. *Proteins: Structure, Function, and Bioinformatics*, 35(4):401–407, 1999.

[58] Y Diao, D Ma, Z Wen, J Yin, J Xiang, and M Li. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and lempel-ziv complexity. *Amino Acids*, 34(1):111–117, 2008.

[59] Ze-Rong Li, Hong Huang Lin, LY Han, L Jiang, X Chen, and Yu Zong Chen. Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, 34(suppl 2):W32–W37, 2006.

[60] E Barbarese, PEf Braun, and JH Carson. Identification of prelarge and presmall basic proteins in mouse myelin and their structural relationship to large and small basic proteins. *Proceedings of the National Academy of Sciences*, 74(8):3360–3364, 1977.

[61] Heidi E Hamm. The many faces of g protein signaling. *Journal of Biological Chemistry*, 273(2):669–672, 1998.

[62] Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000.

[63] Kevin B Murray, Denise Gorse, and Janet M Thornton. Wavelet transforms for the characterization and detection of repeating motifs. *Journal of molecular biology*, 316(2):341–363, 2002.

[64] C Sidney Burrus, Ramesh A Gopinath, Haitao Guo, Jan E Odegard, and Ivan W Selesnick. *Introduction to wavelets and wavelet transforms: a primer*, volume 23. Prentice hall Upper Saddle River, 1998.

[65] Alfred Karl Louis, Peter Maaß, and Andreas Rieder. *Wavelets: theory and applications*. Wiley New York, 1997.

[66] Stephane Mallat and Sifen Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on pattern analysis and machine intelligence*, 14(7): 710–732, 1992.

[67] Albert Cohen and Jelena Kovacevic. Wavelets: The mathematical background. *Proceedings of the IEEE*, 84(4):514–522, 1996.

[68] Pietro Lio. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.

[69] Aaron Klug. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual review of biochemistry*, 79:213–231, 2010.

[70] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1), 2011.

[71] Peter Langfeldera, Bin Zhangb, and Steve Horvatha. Dynamic tree cut: in-depth description, tests and applications. 2007.

[72] Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl 2):W29–W37, 2011.

[73] Arun Krishnan, Kuo-Bin Li, and Praveen Issac. Rapid detection of conserved regions in protein sequences using wavelets. *In Silico Biology*, 4(2):133–148, 2004.

[74] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *Information Theory, IEEE Transactions on*, 36(5):961–1005, 1990.

[75] Ronald Newbold Bracewell and RN Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[76] Byungkook Lee and Frederic M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379–IN4, 1971.

[77] P Hogeweg and B Hesper. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, 20(2):175–186, 1984.

[78] Mansoor AS Saqi and Michael JE Sternberg. A simple method to generate non-trivial alternate alignments of protein sequences. *Journal of molecular biology*, 219 (4):727–732, 1991.

[79] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[80] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[81] Gary A Churchill. Stochastic models for heterogeneous dna sequences. *Bulletin of mathematical biology*, 51(1):79–94, 1989.

[82] Sean R. Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[83] Naila Mimouni, Gerton Lunter, and Charlotte Deane. Hidden markov models for protein sequence alignment.

[84] Kevin Karplus, Christian Barrett, and Richard Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[85] Johannes Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

[86] Andrew F Neuwald, Jun S Liu, David J Lipman, and Charles E Lawrence. Extracting protein alignment models from the sequence database. *Nucleic Acids Research*, 25(9):1665–1677, 1997.

[87] Richard Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[88] William J Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13(10):1368–1374, 1996.

[89] Rachel Karchin and Richard Hughey. Weighting hidden markov models for maximum discrimination. *Bioinformatics*, 14(9):772–782, 1998.

[90] Anders Krogh. Two methods for improving performance of an hmm and their application for gene finding. *Center for Biological Sequence Analysis. Phone*, 45: 4525, 1997.

[91] David Kulp David Haussler and Martin G Reese Frank H Eeckman. A generalized hidden markov model for the recognition of human genes in dna. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology, St. Louis*, pages 134–142, 1996.

[92] Chris Burge and Samuel Karlin. Prediction of complete gene structures in human genomic dna. *Journal of molecular biology*, 268(1):78–94, 1997.

[93] John Henderson, Steven Salzberg, and Kenneth H Fasman. Finding genes in dna with a hidden markov model. *Journal of Computational Biology*, 4(2):127–141, 1997.

[94] Alexander V Lukashin and Mark Borodovsky. Genemark. hmm: new solutions for gene finding. *Nucleic acids research*, 26(4):1107–1115, 1998.

[95] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sjölander, and David Haussler. Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531, 1994.

[96] Christian Barrett, Richard Hughey, and Kevin Karplus. Scoring hidden markov models. *Computer applications in the biosciences: CABIOS*, 13(2):191–199, 1997.

[97] Richard Hughey and Anders Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *Computer applications in the biosciences: CABIOS*, 12(2):95–107, 1996.

[98] Laurent Falquet, Marco Pagni, Philipp Bucher, Nicolas Hulo, Christian JA Sigrist, Kay Hofmann, and Amos Bairoch. The prosite database, its status in 2002. *Nucleic acids research*, 30(1):235–238, 2002.

[99] Pierre Baldi, Yves Chauvin, Tim Hunkapiller, and Marcella A McClure. Hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences*, 91(3):1059–1063, 1994.

[100] Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I Saira Mian, and David Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer applications in the biosciences: CABIOS*, 12(4):327–345, 1996.

[101] Peter Michaely, Diana R Tomchick, Mischa Machius, and Richard GW Anderson. Crystal structure of a 12 ank repeat stack from human ankyrinr. *The EMBO journal*, 21(23):6387–6396, 2002.

[102] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, 2003.

[103] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 2011.

[104] James Kennedy. Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer, 2010.

[105] C Bendtsen. pso: Particle swarm optimization, 2012, r package version 1.0. 3.

[106] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[107] Lipo Wang. *Support Vector Machines: theory and applications*, volume 177. Springer, 2005.

[108] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 2000.

[109] Vladimir N Vapnik. Statistical learning theory. 1998.

[110] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2: 45–66, 2002.

[111] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[112] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels.* The MIT Press, 2002.

[113] Stephen A Kruth. Gram-negative bacterial infections. *Infections diseases of dog and cat*, 2:217–222, 2006.

[114] David L Paterson. Resistance in gram-negative bacteria: Enterobacteriaceae. *The American journal of medicine*, 119(6):S20–S28, 2006.

[115] Tom PJ Dunkley, Rod Watson, Julian L Griffin, Paul Dupree, and Kathryn S Lilley. Localization of organelle proteins by isotope tagging (lopit). *Molecular & Cellular Proteomics*, 3(11):1128–1134, 2004.

[116] Manoj Bhasin, Aarti Garg, and GPS Raghava. Pslpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, 21(10):2522–2524, 2005.

[117] Manoj Bhasin and GPS Raghava. Eslpred: Svm-based method for subcellular localization of eukaryotic proteins using dipeptide composition and psi-blast. *Nucleic acids research*, 32(suppl 2):W414–W419, 2004.

[118] Alejandro A Schäffer, L Aravind, Thomas L Madden, Sergei Shavirin, John L Spouge, Yuri I Wolf, Eugene V Koonin, and Stephen F Altschul. Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, 29(14):2994–3005, 2001.

[119] Dan Xie, Ao Li, Minghui Wang, Zhewen Fan, and Huanqing Feng. Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using svm and profile of psi-blast. *Nucleic Acids Research*, 33(suppl 2):W105–W110, 2005.

[120] Omer Sinan Sarac, Özge Gürsoy-Yüzügüllü, Rengul Cetin-Atalay, and Volkan Atalay. Subsequence-based feature map for protein function classification. *Computational Biology and Chemistry*, 32(2):122–130, 2008.

[121] Chin-Sheng Yu, Yu-Ching Chen, Chih-Hao Lu, and Jenn-Kang Hwang. Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, 64(3):643–651, 2006.

[122] Jiren Wang, Wing-Kin Sung, Arun Krishnan, and Kuo-Bin Li. Protein subcellular localization prediction for gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC bioinformatics*, 6 (1):174, 2005.

[123] Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[124] Hiroshi Nakashima and Ken Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of Molecular Biology*, 238(1):54–61, 1994.