Protein Interaction Hotspot Identification Using Sequence-based Frequency-derived Features

Quang-Thang Nguyen, Ronan Fablet, and Dominique Pastor

***Abstract*—Finding good descriptors, capable of discriminating hotspot residues from others, is still a challenge in many attempts to understand protein interaction. In this paper, descriptors issued from the analysis of amino acid sequences using digital signal processing (DSP) techniques are shown to be as good as those derived from protein tertiary structure and/or information on the complex. The simulation results show that our descriptors can be used separately to predict hotspots, via a random forest classifier, with an accuracy of 79% and a precision of 75%. They can also be used jointly with features derived from tertiary structures to boost the performance up to an accuracy of 82% and a precision of 80%.**

***Index Terms*—Hotspots, protein interaction, sequence-based features, DSP-based features, electron-ion interaction pseudo-potential (EIIP), ionization constant (IC), resonant recognition model (RRM)**

I. Introduction

UNDERSTANDING the structure and the biological function of proteins, the elementary building blocks of all living organisms, is among the most important topics in biology [1]. Scientists are working together to answer the question on how the primary amino acid sequence of the protein defines its conformation and function [1]–[3]. Solving this issue could open a new era in biology where most bioactivities can be controlled, including curing diseases by newly designed proteins with pre-defined functions (see [2] amongst others).

Studies in biology have shown that proteins form certain active three-dimensional structures to interact with other molecules through their interfaces [1]. “Most interfaces are composed of two relatively large protein surfaces with good shape and electrostatic complementarity for one another” [4]. It has also been shown that the distribution of binding energies on these interfaces is not uniform [4]. Some residues are more important than others as they comprise only a small fraction of the interface but contribute most of the necessary energies to the interaction [3]. If they are mutated, the interaction may be affected and, as a result, the protein function may be altered. These critical residues are commonly referred to as *hotspots* [4], [5]. Fig.1 shows an example of such protein hotspots. The characterization, detection and identification of *hotspots* are then keys to the understanding of protein interactions and functions. Much research, both experimental and computational, has been conducted to shed light on these critical residues of the interfaces [3], [4], [6]–[17].

Experimentally, the energy contribution of a given residue to the interaction of a protein with its target can be determined by measuring the change in binding free energy when this residue is *in vitro* mutated to alanine. When the measured change in binding free energy is large enough, this residue is deemed as a hotspot [5]. This method, also known as alanine scanning mutagenesis (ASM), was used by Thorn and Borgan to analyze hotspots and the database that they established is referred to as the Alanine Scanning Energetics database (ASEdb) [18]. Unfortunately, such a widely accepted experimental method requires significant effort and hence induces low throughput [3] [6].

In the search for lower-cost methods applicable to high-throughput analysis, computational approaches have been proposed to identify hotspot residues in protein interfaces. In this respect, Kortemme and Baker [19] introduced a simple physical model for binding free energy. This model takes into account packing interaction, polar interaction involving ion pairs and hydrogen bonds, and solvation. Hotspots are then identified by computational alanine scanning (Robetta) [6], which involves the numerical evaluation of the change in this binding free energy of protein-protein complexes due to computational alanine mutations. These computationally identified hotspots are shown to be in agreement with those identified by *in vitro* experiments and reported in the ASEdb database. Motivated by these works, other energy-based methods have been proposed in [7] [8]. Other computational approaches also investigated molecular dynamics (MD) simulations [9], graph analysis [10] and machine learning [3], [11], [12]. Among all the aforementioned methods, the most successful ones require the structure of the complex — or, at least, the three-dimensional structure of the protein — to be known. The docking approach in [13], which requires simulating thousands of possible docking poses for the protein complex, is among the most popular in this respect.

Although the biological functions of proteins relate to certain active tertiary structures, it is assumed that all information about their structures and, thus their functions, is primarily embedded in amino acid sequences [1]. In other words, knowledge of the three-dimensional structure of the protein or of the complex is expected to be more than sufficient to identify hotspots of the interfaces. In [3], Ofran and Rost showed that hotspots can probably be predicted using only amino acid sequence information. Albeit less accurate than methods based on available three-dimensional (3D) structure information, their sequence-based hotspot identification method yielded

The three authors are with the Department of Signal and Communications, Institut Télécom; Télécom Bretagne, 29238 BREST, FRANCE e-mails: {quang.nguyen, dominique.pastor, ronan.fablet}@telecom-bretagne.eu

This work was funded by Institut Télécom; Télécom Bretagne.

Manuscript received [...]; revised [...].

relevant results. On the other hand, the introduction of the Resonant Recognition Model (RRM) by I. Cosic in [20] pointed out the existence of a *characteristic frequency*, which represents a certain periodicity within the energy distribution of valence electrons along the protein molecule. This finding has inspired many attempts to detect hotspots by using digital signal processing (DSP) methods, such as those based on Short-time Fourier transform (STFT) [14], digital filters [15], wavelet transform [16] and S-Transform filtering [17]. Though tested on only a few individual sequences, these approaches suggest time-series analysis as a relevant framework for hotspot identification.

In this paper, we propose a new family of frequency-based descriptors derived solely from the protein primary amino acid sequence. These descriptors are extracted using a simple *in silico* alanine scanning and DSP techniques based on the discrete Fourier transform. To assess the relevance of the proposed descriptors, a machine-learning-based classification is carried out. The underlying idea is that once a classifier successfully separates hotspots from non-hotspots via certain given features, these features are then considered to be capable of discriminating hotspots from non-hotspots. In other words, these features are actually relevant to the hotspot identification problem. In this study, Random Forests is used since it has been shown to be one of the most powerful machine learning methods [21]. The results on the dataset show that these descriptors can be used to achieve an accuracy of 79% and a precision of 75%. Without information on the protein three-dimensional structure and/or the complex, our descriptors can achieve performance comparable to that reported in [6] [22] where such information is required. This is a key feature since knowing the protein 3D structure, either computationally or experimentally, is not straightforward, and actually, most protein sequences are available without 3D structure information. The experimental results also show that our sequence-based frequency-derived descriptors can boost the prediction up to an accuracy of 82% and a precision of 80% when combined with the 3D structure-based features proposed in [22]. Moreover, using DSP techniques, our method requires very little computational load and thus can be applied to large-scale analysis.

This paper is organized as follows. Section II will introduce our sequence-based frequency-derived features. The learning-based hotspot identification, the selected descriptors and the ground-truth dataset will be presented in Section III with results reported in Section IV. Finally, Section V and Section VI will bring the overall discussion, conclusion and perspectives.

II. Sequence-based frequency-derived features

A. Conversion to numerical sequence

The primary structure of a protein is given by the associated sequence of amino acids. This sequence is often represented by a string of characters sampled from an alphabet of 20 single characters representing the 20 different amino acids. By properly mapping these character strings into numerical sequences, time series analysis can be applied to design very high throughput methods. This conversion from symbolic to numerical sequences may rely on assigning to each amino acid numerical values that represent its physico-chemical and biochemical properties. A number of such indices have been introduced in the literature (more than 500 indices can be found in the AAIndex database [23]). Among them, the *electron-ion interaction pseudo-potential (EIIP)* values [20] and the *ionization constant (IC)* parameters [24] are shown to be very relevant to the protein bioactivity. For each amino acid, the EIIP value describes the average energy states of all valence electrons of its atoms, while the IC value measures its acid dissociation constant from the corresponding ionization reaction. The EIIP and IC values for the 20 amino acids occurring in nature are listed in TABLE I. These two indices have been shown to be very successful in the so-called Resonant Recognition Model [20] [2] [24] (cf. Section V-B) to get an insight into the physical characterization of protein interactions as well as protein hotspots. In our work, these indices will be used to obtain numerical sequences for further DSP analysis.

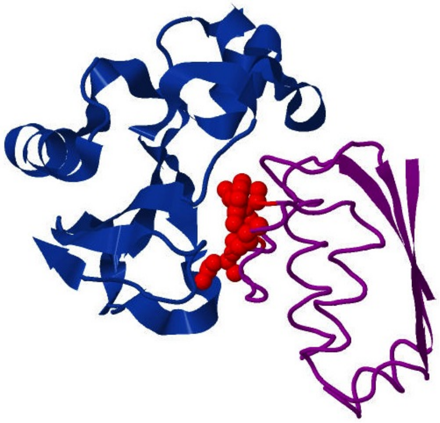

Fig. 1. An example of a protein with hotspots. In this figure, the barstar molecule (right/violet) with hotspots (red) is shown to be in interaction with barnase (left/blue), forming the complex barstar-barnase. The three-dimensional structures of barstar and its target, barnase, are represented in terms of basic secondary structure motifs (α-helices, β-sheets, turns) while red balls indicate atoms of hotspot residues. The structure of the complex was retrieved from the Protein Data Bank (PDB) using its identity 1brs. On the other hand, information on the hotspot residues involved in this interaction was provided by ASEdb.

B. In-sillico alanine scanning and frequency-based features

Experimental alanine scanning mutagenesis has been shown to be an extremely useful tool for analyzing interactions in protein interfaces (see [5], [6] amongst others). This technique involves mutating an amino acid residue to alanine (i.e. deleting the sidechain beyond C_β carbon atom) and then evaluating the effects of this mutation on the affinity of the protein interaction. These effects can be measured by the change in binding free energy ($\Delta\Delta G$) of the protein-target complex. Although experimental ASM is very powerful in identifying hotspot residue, it is still too expensive and laborious to be easily applied to large-scale analysis, despite many advances in molecular biology.

Here we investigate an alternative based on a purely computational approach. More specifically, we propose an *in silico*

TABLE I
EIIP AND IC NUMERICAL VALUES

Amino acid	3-Letter Code	1-Letter Code	EIIP	IC
Leucine	LEU	L	0.0000	2.4000
Isoleucine	ILE	I	0.0000	2.4000
Asparagine	ASN	N	0.0036	2.2000
Glycine	GLY	G	0.0050	2.4600
Valin	VAL	V	0.0057	2.3500
Glutamic Acid	GLU	E	0.0058	2.3000
Proline	PRO	P	0.0198	2.0000
Histidine	HIS	H	0.0242	2.3000
Lysine	LYS	K	0.0371	2.2000
Alanine	ALA	A	0.0373	2.3000
Tyrosine	TYR	Y	0.0516	2.2000
Tryptophan	TRP	W	0.0548	2.3700
Glutamine	GLN	Q	0.0761	2.0600
Methionine	MET	M	0.0823	2.1700
Serine	SER	S	0.0829	2.1000
Cysteine	CYS	C	0.0829	1.9600
Threonine	THR	T	0.0941	2.0900
Phenylalanine	PHE	F	0.0946	1.9800
Arginine	ARG	R	0.0959	1.8200
Aspartic Acid	ASP	D	0.1263	1.8800

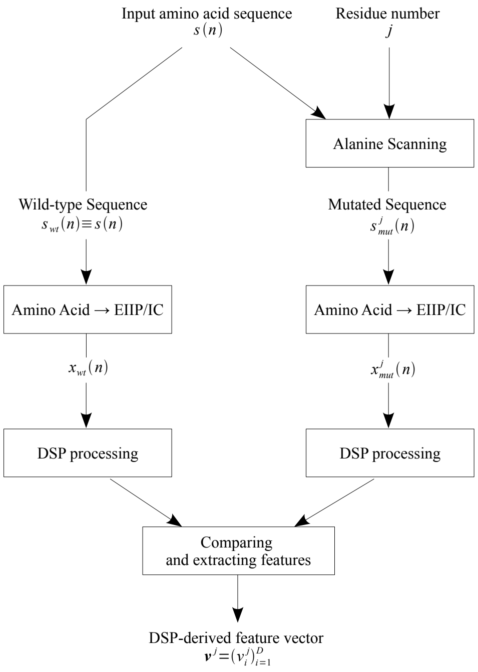

Fig. 2. Computational alanine scanning and DSP-based features deriving

alanine scanning approach inspired from the experimental ASM. We proceed as in ASM, but computationally, by replacing subsequences of residues by alanines and looking for frequency-related changes in the overall sequence. The approach is very similar to the computational alanine scanning method described in [6]. However, instead of investigating a physical model or a single measure that relates to binding free energy as in [6], we analyze changes in the frequency spectrum caused by computational mutagenesis.

The proposed framework is sketched in Fig.2. Our alanine scanning module computationally mutates residues around a given position j of the input amino acid sequence $s(n)$ to alanines. Instead of replacing residue $s(j)$ only, a window of residues centered at position j is processed. All the residues of the window are thus computationally mutated to alanines since changing the value of one single sample will not significantly affect the spectrum of the sequence. On the other hand, the O-ring theory also claims that hotspots are surrounded by other residues, less important in binding energy, but whose role is likely to occlude bulk solvent from central residues to form high affinity interactions [4]. To take these surrounding residues into account, a window of length $L = 5$ — the tested residue $s(j)$ itself and two residues on each side — has been empirically chosen. Furthermore, this choice is reasonable with respect to cases where hotspots are very close to each other.

After computational mutation, both the wild-type sequence $s_{wt}(n)$ and the mutated one $s^j_{mut}(n)$ are converted into numerical sequences ($x_{wt}(n)$ and $x^j_{mut}(n)$, respectively) using either EIIP or IC values. These two numerical sequences will then be analyzed by the same DSP scheme and their associated frequency-based characteristics will be further compared to derive the proposed descriptor vector $\mathbf{v}^j$. Various DSP techniques, both traditional and modern, are thinkable, including Fast Fourier Transform (FFT), Short-time Fourier transform (STFT) or wavelet transform. Similarly, many characteristics could be considered, including peak frequencies, sub-bands energies, and so on. Within our framework, as comparison criteria, we consider spectrum peak changes, sub-band energy changes and global energy changes. These features can be regarded as the analysis at different levels of resolution, from local to global, of the frequency spectrum.

a) Spectrum peak changes: Both the wild-type and the computationally mutated numerical sequences (i.e. $x_{wt}(n)$ and $x^j_{mut}(n)$, respectively) are transformed into the frequency domain by FFT. Peak frequencies are defined as the local maximum points of the wild-type sequence frequency amplitude spectrum. For discrete sequences, we define the set I of these peak frequencies as:

$$I = \{0 < k < N : |X_{wt}(k)| \geq \max(|X_{wt}(k-1)|, |X_{wt}(k+1)|)\}$$

where $X_{wt} = FFT(x_{wt})$ is the FFT of x_{wt} and the FFT size N is chosen to be equal to the sequence length. The DC component is removed from the input sequence before FFT to avoid any spurious peak at the null frequency. Since the amplitude spectrum is symmetric, only one half of it is considered. In terms of the RRM [2], these peak frequencies are regarded as potential characteristic frequencies of the protein functions (cf. our discussion in Section V-B). Changes in the amplitude spectrum at peak frequencies caused by computational mutation are regarded as potential signatures of hotspots. More precisely, we compute the following features:

$$PeakChange^j_k = \frac{|X_{wt}(k)|}{|X^j_{mut}(k)|}$$

where $X^j_{mut} = FFT(x^j_{mut})$ and k is among the considered peak frequencies. In this study, only the set of the three highest peak changes will be retained and will be taken as descriptors.

b) Sub-band energy changes: In addition to amplitude changes at peak frequencies, local energy-changes in frequency subbands are also considered. Specifically, sequences are transformed into time-frequency representations using STFT with a sliding window of length $(\frac{N}{4}+1)$, where the number N of FFT points is now chosen to be the smallest power of two greater than or equal to the sequence length. This value is the default configuration of the Time-Frequency Toolbox (http://tftb.nongnu.org/) that we use to perform time-frequency analysis. To achieve a relevant time-frequency analysis, an analyzing window with small side-lobes is required. According to [25], the 4-term Blackman-Harris window is adopted here for its trade-off between the main-lobe width and the side-lobe levels. Other windows with low side-lobe levels such as the Blackman and the Gaussian windows were also tested and provided similar results. Moderate windows, such as Hamming and Hanning, were shown to be less efficient. After the STFT, since the frequency spectra $S^j_{mut}(j,.)$ and $S_{wt}(j,.)$ at mutated position j are also symmetric, the higher halves can be discarded. The retained lower halves are then evenly divided into 8 equal sub-bands. The change in energy due to computational mutation will be considered in these 8 sub-bands by computing

$$SBEnergyChange^j_m = \frac{\sum_{\nu \in SB_m} |S_{wt}(j,\nu)|^2}{\sum_{\nu \in SB_m} |S^j_{mut}(j,\nu)|^2}, m = 1..8$$

where

$$S_{wt} = STFT(x_{wt})$$
$$S^j_{mut} = STFT(x^j_{mut})$$

and SB_m is the m-th sub-band

$$SB_m = \{k : (m-1)\frac{N}{16} \leq k < m\frac{N}{16}\}.$$

c) Global energy changes: Global energy change is defined as the ratio of the mutated sequence energy to that of the wild-type one:

$$EnergyChange^j = \frac{\sum_{n=1}^{L} |x^j_{mut}(n)|^2}{\sum_{n=1}^{L} |x_{wt}(n)|^2}$$

where L is the sequence length. Of course, this energy ratio can be equivalently computed in the frequency domain.

III. Learning-based hotspot identification

In order to evaluate the relevance of the proposed descriptors for hotspot identification, a learning-based recognition scheme is developed. In this study, we exploit Random Forest (RF) [21] as the learning-based classifier since it is among the most powerful techniques for supervised classification issues. This section first highlights the key features of the RF classifier. We then present the evaluated features and the considered hotspot dataset.

A. Learning-based recognition setting

Before presenting the RF classifier and discussing its advantages for hotspot identification, we begin with a brief introduction to classification trees, the elementary components of any RF.

1) Classification tree: A classification or decision tree [26] is a tree-structured predictive model in which each internal node is associated with a decision rule based on object features $\mathbf{v} = (v_i)_{i=1}^{i=D} \in V$ (V is the so-called feature space) and each terminal leaf is assigned to a class y ($y \in \{0,1\}$ for a binary classification problem such as hotspot identification). Given a decision tree, the class of an object is predicted by filtering its features through the successive decision rules of the internal nodes until a terminal leaf is reached. The class of the terminal leaf is then assigned to the object. In the considered random forest setting, decision trees are binary and the decision rule at each internal node of the tree is a test on only one of the object features, say v_i. In this test, v_i is compared to its associated threshold λ_i. Objects with feature v_i less (resp. greater) than the threshold λ_i will be filtered to the left (resp. right) child node. Fig.3 shows an example of a decision tree.

The construction of a binary decision tree is generally performed on the basis of training samples. Starting from the root node with all the training samples $\{(\mathbf{v}^j, y^j), j = 1..L\}$, the decision tree is grown by recursively splitting nodes in such a way that at each node t_p, the training samples are divided into two subsets (corresponding to two children nodes, t_L and t_R) with maximum class homogeneity according to a decision rule. The determination of the decision rule associated with each split amounts to seeking the best feature v_i and its best threshold λ_i that maximize the information gain G defined by:

$$G = I(t_p) - p_L I(t_L) - p_R I(t_R)$$

where p_L (resp. p_R) is the fraction of samples in t_p that will be sent to the child left node t_L (resp. the right node t_R) and $I(t)$ is the *impurity* of node t [26]. For binary classification problems, node impurity can be interpreted as the proportion of the less frequent class in the sample subset associated with that node. In practice, Shannon's entropy and Gini's diversity index are usually used [27] [26]. Using the aforementioned splitting rules, the decision tree is recursively grown until maximum homogeneity, i.e. minimum impurity, is obtained in the terminal leaf nodes. The construction of a decision tree can be regarded as an adapted quantization of the feature space V into homogeneous regions, in which most training samples are of the same class — and this class will be assigned to any new sample observed in that region (cf. Fig.3).

2) Random forests: Random Forest [21] is an ensemble classifier that combines N decision trees. These trees are constructed using subsets of individuals that are independently and randomly sampled from the original training set. The search for the optimal splitting rule of each node is optimized with respect to a randomly selected subset of features. The classification of an input is obtained by aggregating the votes of the individual trees in the forest. By combining two sources of randomness, i.e. the random selection of training samples and the random selection of features for the determination of each splitting criterion, classification performance of RF greatly improves compared to a single decision tree [21]. RF has been shown to be among the most efficient machine learning schemes for a variety of issues, including mass spectrometry data analysis [28], microarray data analysis [29]

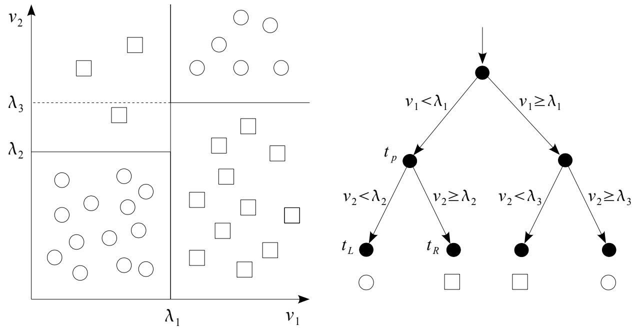

Fig. 3. Example of a classification tree. This example involves a two-class dataset of objects in a 2-dimensional feature space (left). From training samples of each class, represented as squares and circles in the left figure, the classification (decision) tree in the right figure is built. The solid lines in the left figure show the division of the feature space into homogeneous regions in which only samples of the same class are present.

and protein interaction prediction [30]. The construction of an RF involves only two parameters: the number *nbTrees* of trees and the number *mTry* of randomly selected features for the determination of each optimal splitting criterion. These key characteristics make RF a good choice for hotspot data and, particularly, for our purpose of assessing and comparing the relevance of the descriptors.

B. *Evaluated features*

Our sequence-based frequency-derived features will be used in place of or together with 3D structure-based descriptors as the input of the RF classifier. In this way, we aim at showing their capability of discriminating hotspot residues from non-hotspot ones.

1) Frequency-derived features of amino acid sequences: The frequency-based features presented in Section II-B, that is, the 3 highest spectrum peak changes, the 8 sub-band energy changes and the global energy changes, are considered. Using these measures with both EIIP and IC values, a set of 24 different features is computed. The descriptors that best discriminate hotspots from other residues will be selected. This can help reduce the dimensionality of the feature space, without affecting the original semantics of the descriptors, thus providing the ability to interpret the result by domain experts [31]. In this study, such a selection is performed by using a decision tree-based feature ranking technique [32]. The technique involves growing a decision tree based on a sample set (cf. section III-A for more details) then pruning it at a certain level. During the growing process, a decision tree, by its nature, selects the best feature (in the sense of maximizing the information gain) each time a node is split. In the pruning phase, nodes that provide less entropy gain are eliminated. Therefore, the features associated with internal nodes after pruning are considered as the most relevant features. Using the MATLAB *treefit* routine, the decision tree based on samples extracted from [22] showed that the 3 highest spectrum peak changes using EIIP, the energy change in the 7-th sub-band using EIIP and the global energy band using IC are the most appropriate candidates. These selected descriptors form a 5-dimensional vector called the sequence-based frequency-derived features in the sequel.

2) Structure-based features: For comparison purposes, we consider the 3D-structure-based features proposed in [22], namely, the solvent accessibility (accessible surface area (ASA)), the pair potentials and the computational binding free energy change in Robetta [6]. The conservation score is not considered since it is sequence-based and was not included in the best decision rule reported in [22]. It should be noted that the conservation score is seemingly not discriminating enough between hotspot and non-hotspot residues [12].

a) Solvent accessibility: The relative ASA in the complex state and the relative difference ASA between the complex and the monomer states of residue j are defined as in [22]:

$$relCompASA^j = \frac{ASA^j_{comp}}{ASA^j_{max}} \times 100$$

$$relDiffASA^j = \frac{ASA^j_{mono} - ASA^j_{comp}}{ASA^j_{max}} \times 100$$

where ASA^j_{mono} (resp. ASA^j_{comp}) is the ASA of the j-th residue in monomer (resp. complex) state and ASA^j_{max} is its maximum ASA in a tri-peptide state.

b) Pair potentials: The contact potential of residue j is defined as:

$$Potential^j = abs(\sum_{k=1}^{L} Pair(j,k))$$

where L is the number of residues and $Pair(j,k)$ is the contact potential of residues j and k. Two residues are considered to be in contact if they are closer than 7.0Å to each other in space and are separated by at least 3 residues in sequence [22]. We thus have

$$Pair(j,k) = \begin{cases} p(j,k) & \text{if } d(j,k) \leq 7.0 \text{ and } |k-j| \geq 4 \\ 0 & \text{otherwise} \end{cases}$$

in which $p(j,k)$ is the knowledge-based solvent-mediated potential [33] between two residues at positions j and k while $d(j,k)$ is the distance between their centers.

c) Computational binding free energy change (Robetta): These values, given by the Robetta server [6], are changes in computational binding free energy. The calculation is based on the energy function, proposed in [19], which takes into account Lennard-Jones potential, hydrogen bonding and solvation interaction.

The first three structure-based features can be retrieved through the HOTPOINT server [34] and the fourth one from the Robetta server [6].

C. *Dataset*

The evaluation is performed on the union of ground-truth datasets considered in recent works [12], [22] dedicated to hotspot detection. In this union, we consider only the experimental alanine scanning data with available measured values

TABLE II
AMINO ACID CHAINS PRESENT IN DATASET

PDB id	Chain id	Molecule
1a4y	A	RNase inhibitor
	B	Angiogenin
1ahw	C	Tissue factor
1brs	A	Barnase
	D	Barstar
1bxi	A	Colicin E9 immunity protein
1cbw	D	BPTI Trypsin inhibitor
1dan	L	Blood coagulation factor VIIA
	T, U	Soluble tissue factor
1dvf	A, B	FV D1.3
1f47	A	Cell division protein FTSZ
1fc2	C	Fragment B of protein A complex
1fcc	C	Streptococcal protein G (C2 fragment)
1gc1	C	CD4
1jrh	I	Interferon-gamma receptor alpha chain
1jtg	A	Beta-lactamase tem
	B	Beta-lactamase inhibitory protein
1nmb	L	FAB NC10
1vfb	A	IGG1-KAPPA D1.3 FV (light chain)
	B	IGG1-KAPPA D1.3 FV (heavy chain)
	C	Hen egg white lysozyme
2ptc	I	Trypsin inhibitor
3hfm	H	HYHEL-10 IGG1 FAB (heavy chain)
	L	HYHEL-10 IGG1 FAB (light chain)
	Y	Hen egg white lysozyme
3hhr	A	Human growth hormone
	B	Human growth hormone receptor (hGHbp)

of $\Delta\Delta G$. These data were extracted by Tuncbag [22] and Cho [12] from the ASEdb [18] and the published dataset of [19], after removing redundancy that could bias the training and/or the classification performance measurements. More specifically, they excluded homologous proteins with more than 35% sequence identity. Furthermore, in [12], proteins with high structural similarity (structure alignment score is higher than 80) were also discarded. Data from BID (Binding Interface Database) [35] are not included because they do not provide the measured values of the change in binding free energy ($\Delta\Delta G$).

To label the residues of the dataset, we proceed as in [22]. Specifically, residues associated with a value of $\Delta\Delta G$ greater than or equal to 2.0 kcal/mol when mutated to alanines are deemed as hotspots and those with $\Delta\Delta G$ less than 0.4 kcal/mol are regarded as non-hotspots. The other residues are not included in the dataset in order to better discriminate the two classes. The final two-class dataset[1] contains 221 residues in which 76 are hotspots and 145 are non-hotspots. This dataset is somewhat unbalanced with the hotspot class representing only 34% of the samples. The amino acid chains considered in the dataset are listed in TABLE II. The detailed information on these sequences can be obtained from the Protein Data Bank (PDB) [36] via their entry identities (PDB ids) and chain identities (Chain ids).

IV. RESULTS

A. Hotspot identification performance assessment

To assess the identification performance, we consider six evaluation measures: Accuracy (A), Precision (P), Recall (R), Specificity (Sp), F-measure ($F1$) and Matthews correlation coefficient (MCC). These measures are defined as follows:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$F1 = 2 \times \frac{P \times R}{P + R}$$

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

where: TP (resp. TN) is the number of true positives (resp. true negatives), defined as the number of samples that are correctly predicted as hotspots (resp. non-hotspots); FP (false positive) is the number of non-hotspots that are falsely predicted as hotspots, and FN (false negative) is the number of hotspots that are not detected.

Because of the unavoidable trade-off between precision and recall on the one hand, and between recall and specificity on the other hand, both $F1$ and MCC are very usual in machine learning as quality measures of binary classification. The F-measure ($F1$) balances precision P and recall R only, whereas the Matthews correlation coefficient (MCC) takes into account the four terms TP, TN, FP, FN of the confusion matrix. Let us note that a predictor should not perform worse than the 'random guess', 'all-are-positives' and 'all-are-negatives' ones. Therefore, it should satisfy the following conditions:

$$MCC > MCC_{rand}$$

$$F1 > \max(F1_{rand}, F1_{pos}, F1_{neg})$$

where MCC_{rand} and $F1_{rand}$ are expected values of MCC and $F1$ scores for the 'random guess' predictor; $F1_{pos}$ and $F1_{neg}$ are F-measure values for the 'all-are-positives' and the 'all-are-negatives' predictors, respectively. In case of a dataset with p positives and n negatives, these conditions can easily be proved to become $MCC > 0$ and $F1 > 2p/(2p+n)$. With a simple calculation, the significant thresholds of MCC and $F1$ can be found to be $MCC_{thres} = 0$ and $F1_{thres} = 0.51$ for our evaluation dataset of 76 hotspots and 145 non-hotspots.

B. Results

To demonstrate the relevance of our sequence-based frequency-derived features (1DFreq), we compare their predictive performance with that of 3D structure-based ones (3DStruct), i.e. relCompASA, relDiffASA, Potential and Robetta, in terms of the six aforementioned measures, especially $F1$ and MCC. The results are also compared with those obtained using the empirical rule introduced by Tuncbag in HOTPOINT [34], which is shown in [22] to provide similar results to Robetta [6] and outperform other state-of-the-art methods including KFC (Knowledge-based FADE and

[1] The dataset will be available at http://perso.telecom-bretagne.eu/quangnguyen/ upon the acceptance of the paper for publication.

Contacts) [11]. This empirical model only requires two out of the four 3D structure-based features to achieve hotspot recognition:

$$isHotspot = (relCompASA \leq 20\%)\text{AND}(Potential \geq 18.0)$$

In the sequel, for the sake of convenience, the group of these two features, i.e. relCompASA and Potential, will be referred to as 3DHotpoint. The recognition results obtained by combining structure-based features (3DStruct or 3DHotpoint) with our sequence-based 1DFreq are also presented.

Quantitative evaluation is obtained through repeated 10-fold cross-validations. In a 10-fold cross-validation, the dataset is first randomly partitioned into 10 mutually exclusive subsets (or folds) of nearly equal size. This partition is processed in such a way that all folds contain approximately the same proportion of hotspots and non-hotspots as the original dataset. By such stratified sampling, each fold is a good representative of the whole dataset. Given a partition, 10 training-testing iterations are subsequently performed so that within each iteration a different fold is taken as the test-set and the remaining 9 folds serve as the training-set. The results from the 10 iterations are then combined to produce a single estimation of the classification performance. To obtain a better estimation, the 10-fold cross-validation can be repeated multiple times with different stratified partitions. In this study, the 100×10-fold cross-validation is used. The results[2] for the considered dataset are reported in TABLE III. The prediction performance of HOTPOINT for the same dataset is also presented for reference. In Fig.4, the boxplots of $F1$ and MCC scores yielded by different groups of features are included for better comparison. The statistical significance of the results is further assessed by examining the p-values obtained using Student's t-tests. The statistical significance level is set to $\alpha = 0.01$. TABLE IV provides the results of different t-tests obtained using the MATLAB Statistics Toolbox.

V. Discussion

A. Relevance of sequence-based frequency-derived features with respect to previous work

1) Sequence-based descriptors can predict hotspots: The reported quantitative evaluation demonstrates the relevance of the proposed frequency-based protein sequence features for hotspot recognition compared to previous work. Their recognition performance is actually better than that of 3DStruct with respect to all six performance measures. More hotspots are detected (59% compared to 54%) and they are detected more precisely (75% compared to 67%). These features then yield higher $F1$ and MCC scores than 3DStruct (0.66 and 0.52 compared to 0.60 and 0.43 respectively). The t-tests on $F1$ and MCC stress the statistical significance of this improvement with p-values ≈ 0 ($\ll 0.01$). With an $F1$ score of 0.66 ($>$0.51) and an MCC score of 0.52 ($>$0), the hotspot recognition based on the proposed protein sequence features is meaningful.

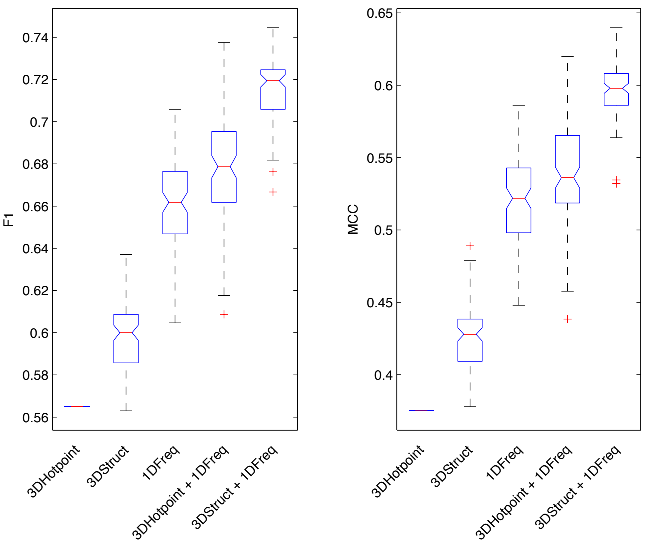

Fig. 4. Boxplots of $F1$ (left) and MCC (right) score values yielded by different sets of features. These boxplots were obtained using the MATLAB *boxplot* routine with the default parameters. For a given boxplot, the extremes of the triangular notch represent the endpoints of the so-called comparison interval of the median at the 5% significance level. Two medians are considered to be significantly different if their comparison intervals do not overlap.

Besides the improvement of hotspot recognition performance with respect to previous works, especially those relying on features extracted from the 3D protein structure, an additional advantage of the proposed approach is its low complexity. It only relies on the analysis of the numerical representations of the 1D sequence of amino acid using frequency analysis. By contrast, the reconstruction of the 3D structure of a protein is a complex task requiring complex experimental expertise, especially regarding protein crystallization to achieve a 3D imaging of the protein structure. Such crystallization issues are particularly complex for large compounds [1]. Relying only the 1D sequence, we enlarge the potential application field of hotspot recognition techniques, especially for newly-sequenced proteins presenting weak homologies to proteins with known 3D structures [37], [38].

2) The combination of 3D structure characteristics and 1D frequency-based features improves the recognition of hotspots: We also evaluated the combination of the proposed 1D sequence features and descriptors of the 3D structure. As reported in TABLE III, the combination [3DStruct+1DFreq] leads to significant recognition statistics (p-values$<$0.01) with an accuracy of 82% and a precision of 80%. It is proved to reach better recognition performance than the 1D sequence features (i.e. 1DFreq) alone or the combination [3DHotpoint+1DFreq] (respectively, 82% vs. 79% and 80% for recognition accuracy and 80% vs. 75% and 75% for recognition precision). It is also worth noticing that [3DStruct+1DFreq] returns a significant gain for all six assessment indices.

These results show that the proposed frequency-based 1D sequence features provide discriminative information complementary to the descriptors issued from the classical local characteristics of the 3D structure of the protein. It then provides the means to improve recognition performance for a subset

[2] The Matlab code yielding these results will be available at http://perso.telecom-bretagne.eu/quangnguyen/ upon the acceptance of the paper for publication.

TABLE III
CLASSIFICATION PERFORMANCE RESULTS (MEAN(±STANDARD DEVIATION))

Features	Accuracy (A)	Precision (P)	Recall (R)	Specificity (Sp)	$F1$	MCC
3DHotpoint [a, d]	0.729	0.629	0.513	0.841	0.565	0.375
3DStruct [b]	0.751(±0.010)	0.672(±0.021)	0.541(±0.018)	0.861(±0.012)	0.599(±0.016)	0.427(±0.024)
1DFreq [c]	0.790(±0.013)	0.748(±0.025)	0.589(±0.029)	0.896(±0.014)	0.659(±0.023)	0.518(±0.031)
3DHotpoint+1DFreq	0.798(±0.014)	0.751(±0.025)	0.616(±0.031)	0.893(±0.013)	0.676(±0.025)	0.537(±0.033)
3DStruct+1DFreq	0.824(±0.009)	0.801(±0.017)	0.649(±0.017)	0.915(±0.009)	0.716(±0.015)	0.597(±0.020)

[a] 3DHotpoint: relCompASA and Potential.
[b] 3DStruct: relCompASA, relDiffASA, Potential and Robetta.
[c] 1DFreq: our proposed sequence-based frequency-derived features.
[d] The results presented in this row are obtained by HOTPOINT while others are yielded by using RF with *nbTrees*= 1000 classification trees. For RF, all possible values of *mTry* are tested and the best results are provided.

TABLE IV
RESULTS GIVEN BY DIFFERENT t-TESTS [a]

Null hypothesis (H_0)	Alternative hypothesis (H_1)	$F1$		MCC	
		Accept (h)	p-value	Accept (h)	p-value
1DFreq $\leq$ 3DHotpoint [b]	1DFreq $>$ 3DHotpoint	H_1	2.62×10^{-63}	H_1	2.08×10^{-69}
1DFreq $\leq$ 3DStruct	1DFreq $>$ 3DStruct	H_1	5.04×10^{-50}	H_1	7.70×10^{-58}
[3DHotpoint+1DFreq] $\leq$ 1DFreq	[3DHotpoint+1DFreq] $>$ 1DFreq	H_1	2.89×10^{-07}	H_1	2.60×10^{-05}
[3DStruct+1DFreq] $\leq$ 1DFreq	[3DStruct+1DFreq] $>$ 1DFreq	H_1	1.21×10^{-48}	H_1	1.37×10^{-50}

[a] Right-side t-tests were performed. In this table, the notation FeasA $>$ FeasB (resp. FeasA $\leq$ FeasB) implies that the mean performance score provided by FeasA is greater than (resp. less than or equal to) that yielded by FeasB.
[b] The results reported in this row are obtained using one-sample t-test while others are provided by two-sample ones.

of protein sequences whose 3D structures are known. It may also provide the basis for similar improvements for protein sequences having high homology (typically, greater than 35% of residue identity) with a protein whose 3D structure is known. For such a homology level, it is indeed generally assumed that the 3D structure of the analyzed protein can be inferred from its homologue [37]. One may expect that the combination of the proposed 1D sequence features and of 3D features extracted from the inferred structure could also lead to substantial improvement of hotspot recognition compared to 1D sequence features alone.

B. Physico-chemical interpretation of the proposed features

The analysis of frequency-based features of 1D numerical representations of the protein amino acid sequence was initially motivated by the RRM [20], a physico-mathematical model which was originally introduced as an attempt to get an insight into the selectivity of protein interactions. By assigning to each amino acid a physical parameter value relevant to the protein bioactivity and analyzing the resulting numerical sequence, the RRM has successfully revealed the existence of frequency characteristics that characterize how a protein can recognize its target in an interaction. From the RRM perspective, proteins of the same family, sharing the same biological function, also share some frequency-based features. In particular, their frequency spectra exhibit a common *characteristic frequency* [2]. This characteristic frequency was identified from the consensus spectrum, which is defined as the multiple cross-spectrum function of the Fourier transforms of all the sequences of the protein family as in [2]:

$$M(n) = |X_1(n)|.|X_2(n)|...|X_K(n)|, n = 0, 1, ..., N-1$$

where $X_i(n), i = 1, 2, ..., K$ are the discrete Fourier transform coefficients of the numerical representation of the i-th protein sequence of the family, K is the number of family sequences and N is the length of the longest sequence. Shorter sequences are filled up with their mean value to have the same length N. Fig.5 reports the consensus spectrum of the fibroblast growth factor (FGF) family. This consensus spectrum clearly exhibits a characteristic frequency at $f_c = 0.4567$, which is significantly present in all the sequences of the FGF family.

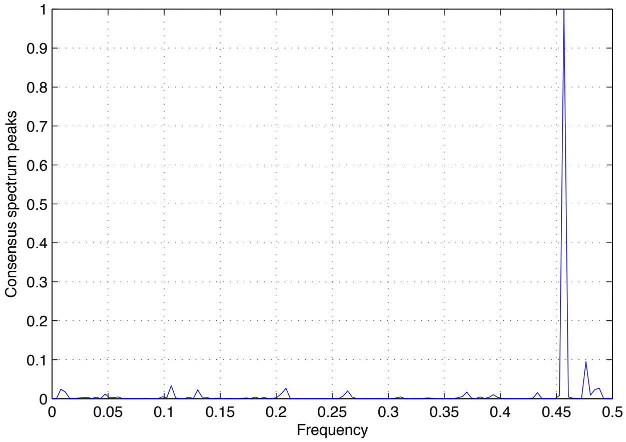

Fig. 5. Characteristic frequency of the fibroblast growth factor (FGF) protein family: the consensus Fourier spectrum shows that the FGF protein family members share a common characteristic frequency at $f_c = 0.4567$

It was conjectured in [2] that these characteristic frequencies are associated with the common function of the proteins of a given family. Since hotspots are referred to as the key positions that determine the protein function, they were defined by Cosic et al. [2] as the residues that are most affected by any change made to the amplitude spectrum at the characteristic frequency corresponding to the protein biological function. Although some evidence of the correlation between the hotspots defined by RRM and those detected by ASM were reported [14], [15], [17], the recognition performance was limited to very few examples. Besides, earlier applications of the RRM required the functional family of the protein to be known to compute the corresponding characteristic frequency.

Compared to this previous work, our contribution is twofold. First, whereas the determination of RRM-based hotspots initially requires the computation of the characteristic frequency of a family of proteins, we do not impose such a constraint. Second, rather than a purely DSP-based approach as in [2], [14]–[17] aimed at detecting local residues associated with the characteristic frequency, we combine DSP tools and mutagenesis principles. We locally determine frequency-related energy changes resulting from the computational mutation of residue subsets to alanines. Considering the alanine mutations as a reference model, our procedure can be applied to newly sequenced or unclassified proteins, which might enlarge its potential application domain. Moreover, we have reported an actual evaluation of hotspot recognition performance with respect to a reference database of experimental ASM hotspots.

Our results bring new evidence to support the conjecture of Cosic et al. [2] that protein hotspots are associated with frequency features of physico-chemical characteristics of the amino acid sequence. Whereas this statement was analyzed in [2] for the RRM model associated with electron-ion interaction potentials, we have shown here that protein hotspots may also involve specific frequency-related features for other physico-chemical characteristics such as ionization constants. Future work should further investigate, from both the computational and the biophysical point of view, the characterization and the interpretation of such frequency-related properties of protein and associated hotspots.

VI. Conclusion

In this paper, an *in-silico* alanine scanning framework with frequency-derived features of numerical representations of the amino acid sequences has been introduced for protein hotspot recognition. It outperforms previous work on a ground-truth database of protein hotspots [12], [22]. We have also shown that improved recognition performance can be achieved when the 3D structure of the protein is available, i.e. from the combination of the proposed 1D frequency-related features and local descriptors of the 3D structure.

The reported experiments support the assumption that all functionalities of a protein are basically encoded into its primary amino acid sequence. But how this encoding is performed is still an open question. In this respect, it could be profitable to get a better insight into the physico-chemical meaning of the frequency-related descriptors introduced in this paper.

From an engineering point of view, the analysis of one-dimensional (1D) sequences requires very little computational load, making our approach much less complex than those based on docking, MD simulations, graph analysis or 3D structure information derived descriptors. As a result, our method should be capable of dealing with large-scale datasets, which become a crucial problem as more and more proteomic data are available in the public domain [39] [16].

As mentioned in Section II-B, other time series analyses can be involved in the proposed framework to provide new hotspot descriptors. The use of DSP techniques such as those in [14]–[17] might be investigated to derive descriptors that could further be compared to and/or combined with ours for proteins belonging to the same functional family.

The main focus of this paper was not the classifier itself, but rather the relevance — assessed by classification performance measurements — of the proposed descriptors as hotspots signatures. Therefore, it can be expected that the classification performance could perhaps be even further improved by combining RF with other classifiers such as SVM, neural networks and so forth. An exhaustive study of this type could be addressed in future work.

Acknowledgments

The authors are very grateful to the editor and the reviewers for their insightful comments and suggestions that helped to significantly improve the quality of this paper. Many thanks to Patrick McLaughlin for his help in improving the English writing of the paper.

References

[1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential Cell Biology*, 3rd ed. Galand Science, 2010.

[2] I. Cosic, *The resonant recognition model of macromolecular bioactivity: theory and applications*. Birkhauser Verlag, 1997.

[3] Y. Ofran and B. Rost, “Protein-protein interaction hotspots carved into sequences,” *PLoS Comput Biol*, vol. 3, no. 7, p. e119, Jul. 2007.

[4] A. Bogan and K. Thorn, “Anatomy of hot spots in protein interfaces,” *Journal of molecular biology*, vol. 280, pp. 1–9, 1998.

[5] J. Wells, “Systematic mutational analyses of protein-protein interfaces,” *Methods in enzymology*, vol. 202, pp. 390–411, 1991.

[6] T. Kortemme, D. E. Kim, and D. Baker, “Computational Alanine Scanning of Protein-Protein Interfaces,” *Sci. STKE*, vol. 2004, no. 219, pp. pl2–, 2004.

[7] R. Guerois, J. Nielsen, and L. Serrano, “Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations,” *Journal of molecular biology*, vol. 320, no. 2, pp. 369–387, Jul. 2002.

[8] Y. Gao, R. Wang, and L. Lai, “Structure-based method for analyzing protein-protein interfaces,” *Journal of molecular modeling*, vol. 10, no. 1, pp. 44–54, Feb. 2004.

[9] D. Rajamani, S. Thiel, S. Vajda, and C. J. Camacho, “Anchor residues in protein-protein interactions,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 31, pp. 11 287–11 292, 2004.

[10] K. V. Brinda, N. Kannan, and S. Vishveshwara, “Analysis of homodimeric protein interfaces by graph-spectral methods,” *Protein engineering*, vol. 15, no. 4, pp. 265–77, Apr. 2002.

[11] S. Darnell, D. Page, and J. Mitchell, “Automated Decision-Tree Approach to Predicting Protein-Protein Interaction Hot Spots,” *Proteins*, vol. 68, no. 4, pp. 813–823, 2007.

[12] K.-i. Cho, D. Kim, and D. Lee, “A feature-based approach to modeling protein-protein interaction hot spots,” *Nucleic Acids Research*, vol. 37, no. 8, pp. 2672–87, May 2009.

[13] J. Fernández-Recio, M. Totrov, and R. Abagyan, “Identification of protein-protein interaction sites from docking energy landscapes,” *Journal of molecular biology*, vol. 335, no. 3, pp. 843–865, Jan. 2004.

[14] P. Ramachandran, A. Antoniou, and P. Vaidyanathan, “Identification and location of hot spots in proteins using the short-time discrete fourier transform,” in *Conference Record of the Thirty-Eighth Asilomar, Conference on Signals, Systems and Computers, 2004*, vol. 2, nov. 2004, pp. 1656–1660.

[15] P. Ramachandran and A. Antoniou, “Identification of hot-spot locations in proteins using digital filters,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 378–389, Jun. 2008.

[16] K. Deergha Rao and M. Swamy, “Analysis of genomics and proteomics using DSP techniques,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 55, no. 1, pp. 370 –378, feb. 2008.

[17] S. Sahu and G. Panda, “A new approach for identification of hot spots in proteins using s-transform filtering,” in *IEEE International Workshop on Genomic Signal Processing and Statistics, 2009 (GENSIPS 2009)*, may. 2009, pp. 1–4.

[18] K. S. Thorn and A. A. Bogan, “ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions,” *Bioinformatics*, vol. 17, no. 3, pp. 284–285, 2001.

[19] T. Kortemme and D. Baker, “A simple physical model for binding energy hot spots in protein-protein complexes,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14 116–14 121, 2002.

[20] I. Cosic, “Macromolecular bioactivity: is it resonant interaction between macromolecules?-theory and applications,” *Biomedical Engineering, IEEE Transactions on*, vol. 41, no. 12, pp. 1101–1114, dec. 1994.

[21] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[22] N. Tuncbag, A. Gursoy, and O. Keskin, “Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy,” *Bioinformatics*, vol. 25, no. 12, pp. 1513–1520, 2009.

[23] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, “AAindex: amino acid index database, progress report 2008.” *Nucleic acids research*, vol. 36, no. Database issue, pp. D202–5, Jan. 2008.

[24] I. Cosic and E. Pirogova, “Application of ionisation constant of amino acids for protein signal analysis within the resonant recognition model,” in *Proceedings of the 20th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, 1998*, vol. 2, oct. 1998, pp. 1072–1075 vol.2.

[25] F. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51 – 83, Jan. 1978.

[26] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*, 1st ed. Chapman and Hall/CRC, January 1984.

[27] R. J. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[28] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, “Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data,” *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[29] R. Diaz-Uriarte and S. Alvarez de Andres, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[30] X.-W. Chen and M. Liu, “Prediction of protein-protein interactions using random decision forest framework,” *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[31] Y. Saeys, I. Inza, and P. Larraaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[32] C. Cardie, “Using decision trees to improve case-based learning,” in *In Proceedings of the Tenth International Conference on Machine Learning*. Morgan Kaufmann, 1993, pp. 25–32.

[33] O. Keskin, I. Bahar, a. Y. Badretdinov, O. B. Ptitsyn, and R. L. Jernigan, “Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions,” *Protein Science*, vol. 7, no. 12, pp. 2578–2586, Dec. 1998.

[34] N. Tuncbag, O. Keskin, and A. Gursoy, “HotPoint: hot spot prediction server for protein interfaces,” *Nucleic Acids Research*, vol. 38, no. suppl 2, pp. W402–W406, 2010.

[35] T. B. Fischer, K. V. Arunachalam, D. Bailey, V. Mangual, S. Bakhru, R. Russo, D. Huang, M. Paczkowski, V. Lalchandani, C. Ramachandra, B. Ellison, S. Galer, J. Shapley, E. Fuentes, and J. Tsai, “The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces,” *Bioinformatics*, vol. 19, no. 11, pp. 1453–1454, 2003.

[36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[37] B. Rost, “Twilight zone of protein sequence alignments,” *Protein Engineering*, vol. 12, no. 2, pp. 85–94, 1999.

[38] A. Pandini, G. Mauri, A. Bordogna, and L. Bonati, “Detecting similarities among distant homologous proteins by comparison of domain flexibilities,” *Protein Engineering Design and Selection*, vol. 20, no. 6, pp. 285–299, 2007. [Online]. Available: http://peds.oxfordjournals.org/content/20/6/285.abstract

[39] P. P. Vaidyanathan and B.-J. Yoon, “The role of signal-processing concepts in genomics and proteomics,” *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 111–135, 2004.