Hot spot identification by means of classification methods employing wavelet transform-based features

Anna Tamulewicz (anna.tamulewicz@polsl.pl)
Politechnika Slaska
Ewaryst Tkacz
Politechnika Slaska
Ivo Provazník
Vysoke uceni technicke v Brne https://orcid.org/0000-0002-3422-7938

Research

Keywords: hot spot, Resonant Recognition Model, continuous wavelet transform, digital signal processing, protein interaction, classification

Posted Date: June 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34908/v1

RESEARCH

Hot spot identification by means of classification methods employing wavelet transform-based features

Anna Tamulewicz[1*], Ewaryst Tkacz[1] and Ivo Provazník[1,2]

[*]Correspondence:
anna.tamulewicz@polsl.pl
[1]Department of Biosensors and Biomedical Signals Processing, Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland
Full list of author information is available at the end of the article

Abstract

Background: Proteins can interact with one another by an interface composed of two proteins. Some of the interface residues – called hot spots – have the greatest impact on binding energy in a protein complex. This paper introduces the application of continuous wavelet transform based on the fast Fourier transform (CWTFT) to the analysis of hot spots in proteins.

Results: The algorithm was evaluated by using data sets containing 30 proteins. From the number of tested classifiers the best 10 models were preferred. The classifiers achieved sensitivity of 52%–71%, specificity of 70%–83% and accuracy of 70%–74%.

Conclusion: The analyses show that the method combining CWTFT and classification algorithms is able to identify hot spot residues with valuable results.

Methods: The basis of the algorithm is extraction of features from spectra obtained by using CWTFT with different wavelet functions including Morlet, m^{th} order derivative of Gaussian, Paul and Bump wavelets. Then, the classifiers that are able to separate hot spot from non-hot spot residues according to these features are applied.

Keywords: hot spot; Resonant Recognition Model; continuous wavelet transform; digital signal processing; protein interaction; classification

Background

Protein complexes control most of the biological processes in living cells. They can interact with one another at the interface, which is composed of two proteins linked by a noncovalent bond [1]. If the distance between any two atoms of two amino acid chains is less than the sum of their van der Waals radii plus 0.5 Å tolerance, then these amino acids can be deemed interface residues [2, 3].

One of the most important properties of an interface is that the energy is not uniformly distributed. Some of the interface residues have the greatest impact on binding energy in the protein complex. Such residues are called hot spots. Studies have shown that hot spots tend to cluster near the centre of the interface [2, 4, 5]. The hot spot residues are surrounded by energetically less important residues, which occlude the hot spots from the solvent. Such occlusion plays an important role in highly energetic interactions and determines the location of hot spots in the interface centre. Studies also show that hot spots usually comprise structurally conserved sequences because hot spot residues are found to mutate more slowly than the rest

of the protein surface. Furthermore, Bogan and Thorn [4] observed that tryptophan, arginine, and tyrosine are the preferred residues in hot spots.
Hot spot analysis is important because it can help to discover biological functions and structures of proteins and understand the interactions between them. Studies on energy distribution in proteins may also be useful in the detection of mutations and single-nucleotide polymorphism analysis. Another application of hot spot detection is the possibility of creating proteins with predefined functions, which can be useful in drug design. The basis of drug design is often mimmicking the regulation of interaction between proteins; hence, it is important to understand the rules governing these interactions [2, 4, 6].

Methods of hot spot identification

Experimentally, hot spot residues can be identified by measuring the change in binding free energy ($\Delta\Delta G$) by using a molecular biology technique called Alanine Scanning Mutagenesis (ASM) [7, 8]. In ASM experiments, each of the amino acids is mutated to alanine one by one, and the binding affinity of mutated chain is measured. The binding affinity can be determined by $\Delta\Delta G$ upon mutation of a given residue to alanine. In general, it is assumed that a residue can be deemed hot spot if it has a $\Delta\Delta G \geq 2\,kcal/mol$ when mutated to alanine. Mutations to alanine can cause changes in the biological properties of the proteins; thus, if the given residue is important for protein function, then replacing it with alanine should change the protein function. The $\Delta\Delta G$ from ASM experiments along with other data are deposited in the Alanine Scanning Energetic Database (ASEdb) [4, 9].
However, the ASM method is expensive and time-consuming, and thus a number of computational algorithms presenting different approaches were created to predict hot spot residues. Most of the developed models require information about the complex, such as the 3D structure of proteins (e.g. Robetta [6, 10], HotPoint [2, 3, 5], MAPPIS [1], KFC [11], SpotOn [12], PredHS [13], iPPHOT [14], the method using docking approach [15], HSPred [16], HEP [17]) or physico-chemical properties of their residues (e.g. method using random projection-based classifier [18], MSCA [19], method applying ensemble learning [20], DICFC [21], iFrag [22]). Most of the aforementioned algorithms require knowledge of the protein structure, which is a significant drawback of these methods because the protein structure has been determined only for a limited number of proteins. However, it is possible to analyse hot spots by using only the primary structure of proteins because the amino acid sequence contains all the information about the functions and structure of the protein. This assumption is used in algorithms applying digital signal processing methods to hot spot analysis [23].
Most such models try to find regions in the protein sequence where the characteristic frequency (described below) is dominant. The most direct algorithms apply Fourier transform in order to search for hot spots, either by changing the amplitude in the Fourier spectrum [24] or by simply subtracting the consensus spectrum from the amplitude spectrum of the given protein [25]. However, Fourier transform is insufficient for the analysis of protein sequences because such signals are non-stationary (their characteristics change in time). Thus, most researchers use signal processing methods by employing time-frequency tools, which are more suitable for

nonstationary signal analysis. Continuous wavelet transform (CWT) is one of the most popular methods of such analysis and was used in many research studies for the identification of hot spots [26, 27, 28, 29]. The wavelet-based approach helps to find areas of high energy that correspond to hot spots, but precise identification of hot spots is missing in most of these works. A more accurate method of hot spot detection was presented in a study by applying modified Gabor wavelet transform [30]. Other methods of hot spot analysis use short-time Fourier transform (STFT) [31] and S transform [32, 33], which allow for signal analysis in the time-frequency plane. An algorithm employing digital filtering was also presented [34]. Another technique uses statistically optimal null filters [35] with the smoothed pseudo Wigner–Ville distribution [36] incorporated. Another approach focuses on finding sequence-based frequency-derived descriptors capable of discriminating hot spot residues from others [37].

Resonant Recognition Model

Digital signal processing tools can be applied to the analysis of hot spots in proteins on the basis of the Resonant Recognition Model (RRM), which treats protein (and nucleic acid) sequences as discrete signals. The RRM assumes correlation between protein function and a numerical representation of its amino acids, which can be obtained by assigning each of the residues a numerical value relevant to the biological activity of the protein [23, 24]. A number of parameters were considered (including physicochemical, thermodynamic, structural and statistical parameters), and the electron-ion interaction pseudo-potential (EIIP, Table 1) was selected as the most suitable for the analysis of proteins by using digital signal processing methods [38]. EIIP is a physical quantity denoting average energy state of all delocalised electrons of a given amino acid.

The RRM shows that proteins with the same biological functions (i.e., the same target or receptor) have a common frequency component in the energy distribution of the delocalised electrons. The common frequency component of the biological process is called characteristic frequency. The cross-spectral function of the group of sequences with the same biological function is called a consensus spectrum. Peak frequencies in the consensus spectrum of the considered proteins correspond to frequencies common in all of the analysed signals (sequences), whereas frequency components that do not occur in all signals are removed from the consensus spectrum. Thus, the characteristic frequency can be obtained by finding the frequency of significant peak in the consensus spectrum of functionally related proteins. In a previous study [29], the authors showed that some problems with gathering a proper set of related proteins may occur; thus, the characteristic frequency could not always be properly determined [24, 23].

Results

The analysis of hot spots with the aid of continuous wavelet transform based on the fast Fourier transform was conducted according to the algorithm described in the Methods section. The model was validated by comparing the results with ASEdb, which contains hot spots from experimental techniques (ASM). In order to present the performance of the algorithm, the following evaluation metrics were calculated:

- Sensitivity:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

- Specificity:

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

- Accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP and TN are defined as the number of correctly identified hot spots and non-hot spots, respectively, whereas FP is defined as the number of non-hot spots incorrectly identified as hot spots, and FN denotes the number of hot spots incorrectly identified as non-hot spots.
In order to demonstrate the application of the algorithm to the identification of hot spots, 10 models that achieved the best results were chosen from all tested classifiers. Each classifier was validated with a five-fold cross-validation. To evaluate the models, 10 iterations were performed. The mean values for accuracy, sensitivity and specificity along with standard errors (se, eq. 4) acquired by using the presented algorithm are presented in Table 4.

$$se = \frac{s}{\sqrt{n}} \tag{4}$$

where s is a standard deviation and $n = 10$ is the number of iterations.
The classifiers were compared using one-way analysis of variance (ANOVA) for a significance level of $\alpha = 0.05$. P-values for the null hypothesis that the means of the groups (classifiers) are equal are presented in Tables 5, 6 and 7. Figure 4 presents 95% confidence interval for the comparison.

Discussion

The aim of this study was to create models that are able to predict the potential unknown location of hot spots, which could help researchers in performing ASM in the wet lab. ASM is expensive and time-consuming; hence, by identifying the possible locations of hot spots with the aid of computational methods, the researchers can perform alanine mutations only in amino acid locations identified by these methods. In order to correctly detect the greatest percentage of hot spots in the ASM method, the model should be described by as large a sensitivity (Sn) value as possible.
The results obtained by using the proposed algorithm and presented in Table 4 indicate that the Classifier 10 (RUSBoost algorithm) can be considered the best tool to identify hot spots. It returns the greatest Sn value ($\sim$ 71%) compared to other tested classifiers (Sn of other classifiers varies from $\sim$ 52% for Classifier 1 to $\sim$ 58% for Classifier 6; differences between them are not statistically significant,

Table 5), even though it shows a low Sp value. The accuracy (ACC) and specificity (Sp) values for Classifier 10 are both $\sim 70\%$; thus, it is the smallest Sp value of the tested classifiers, while the ACC is similar compared to the other classifiers (only Classifiers 2 and 3 exhibit significantly greater ACC values than Classifier 10, Table 7).

Comparing ACC values, it can be observed that Classifier 3 (decision tree) achieved the greatest $ACC \sim 74\%$; however, the statistical testing showed that the ACC value is significantly greater compared to only four classifiers (Table 7). Additionally, the Sp value is the greatest for the Classifier 3 ($\sim 83\%$, differences are statistically significant for three classifiers, Table 6). Finally, in addition to Classifier 10, Classifier 3 can also be considered a good tool for hot spot identification due to the greatest Sp and ACC values and $Sn \sim 58\%$.

It should be noted that non-hot spots were overrepresented in the analysed data set ($\sim$ 66% NH vs. $\sim$ 34% HS), therefore, some of classifiers could achieve the $ACC \sim 66\%$ by assigning all observations to one class (thus, $Sp = 100\%$ and $Sn = 0\%$). This observation resulted in focusing this study on finding the classifier achieving as great an Sn value as possible. The RUSBoost algorithm, used by Classifier 10, is effective for imbalanced data, which was shown by the performance of Classifier 10.

Classifier 10 is also described by a greater sensitivity value compared with other feature-based approaches introduced by Nguyen *et al.* ($\sim 59\% - 65\%$ [37], Table 8) and Cho *et al.* (58%) [56]. In addition, in [37], it was shown that changing the value of one sample does not significantly affect the spectrum of the analysed sequence; thus, a window of five residues centred at the position of a given residue was replaced by alanine. In this paper, only one residue (corresponding to a hot spot or non-hot spot) was substituted by alanine; thus, the effect of the change in spectra by a single residue was examined.

Even though the application of continuous wavelet transform to hot spot analysis was considered earlier [26, 27, 28], the exact identification of hot spots is missing in these works, and the results were presented only for a few example proteins, which did not allow for the evaluation of the performance of the proposed methods. A more accurate model employing modified Gabor wavelet transform (MGWT) was presented by Shakya *et al.* [30], which achieved accuracy of 67%, sensitivity of 70%, and specificity of 65% (Table 8). The presented algorithm using all of the classifiers showed greater Sp and ACC values than the MGWT-based method, whereas only Classifier 10 acquired slightly better results for the Sn metric. Although the algorithm does not improve the performance significantly compared with MGWT, it does not require knowledge of the characteristic frequency, which is a considerable advantage over the MGWT method. As mentioned before, gathering a proper set of related proteins can be difficult but is important for the determination of characteristic frequency, and thus for hot spot identification. Therefore, the superiority of the proposed approach over the other wavelet-based methods of hot spot identification was demonstrated.

Conclusion

In this paper, the analysis of hot spots by using classification algorithms and signal processing methods employing a time-frequency tool such as continuous wavelet transform based on the fast Fourier transform (CWTFT) was conducted. The protein data set consisted of 219 amino acid residues, and various wavelet functions were used for hot spot analysis. The numerical signals were acquired by converting the amino acid sequences into EIIP values. The obtained numerical signals were then processed by using CWTFT.

The employed approach focuses on finding classifiers that are able to separate hot spots from non-hot spots through features extracted from CWTFT spectra. Ten classifiers were considered. Classifier 10 (employing decision tree learners with random under-sampling boost ensemble method) was found to be the most appropriate tool for hot spot analysis because of the greatest sensitivity value ($\sim 71\%$) among all of the tested classifiers. Most of the classifiers achieved similar Sp values (from $\sim 77\%$ to $\sim 83\%$ for Classifier 3) except for Classifier 10 ($\sim 70\%$). The ACC values were also similar and ranged from $\sim 70\%$ for Classifier 10 to $\sim 74\%$ for Classifier 3. Thus, Classifier 3 (employing a decision tree) can also be used due to the greatest ACC and Sp values and sufficient Sn. The above leads to the conclusion that it is difficult to maintain a balance between all considered evaluation metrics.

This study shows that CWTFT is a proper digital signal processing tool for hot spot analysis and provides valuable results. Furthermore, the proposed algorithm does not require *a priori* knowledge of the characteristic frequency, which is a considerable advantage over other wavelet-based methods.

Methods

This study presents the analysis of hot spots by using one of the digital signal processing tools, that is continuous wavelet transform based on the fast Fourier transform and several classification methods.

Continuous wavelet transform based on the fast Fourier transform

The continuous wavelet transform (CWT) describes the correlation between the analysed signal x(t) and shifted and compressed or stretched versions of a function called wavelet ($\psi(t)$). The CWT can be defined as follows:

$$C(a,b) = \int_{-\infty}^{+\infty} x(t)\psi^*_{a,b}(t)dt = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) dt \tag{5}$$

where $\psi^*(t)$ denotes the complex conjugate of the analysing mother wavelet $\psi(t)$, and C(a,b) is the function of the parameters:

- a, which denotes the dilatation of the wavelet (scale),
- b, which defines the translation of the wavelet and determines the time localisation.

The CWT coefficients are affected by values of scale (a) and position (b) as well as the choice of wavelet. CWT transforms the signal in the time domain into the scale–time domain; however, CWT can be examined in the time-frequency plane by converting specific scale a into the pseudo-frequency value f_a by:

$$f_a = \frac{f_c}{a} \tag{6}$$

where f_c is the centre frequency of the mother wavelet at scale $a = 1$ and position $b = 0$.
The CWT is invertible, that is the original signal can be recovered from the wavelet transform coefficients by the inverse wavelet transform:

$$x(t) = \frac{1}{K_\psi} \int_{a=0}^{+\infty} \left(\int_{b=-\infty}^{+\infty} C(a,b)\psi_{a,b}(t) \frac{db}{a^2} \right) da \tag{7}$$

where K_ψ is a constant factor depending on the wavelet function.
The analysed signal $x(t)$ representing an amino acid sequence, is discrete; thus, the continuous wavelet transform was calculated on the discrete data [39, 40, 41]. The difference between CWT performed on discrete data and discrete wavelet transform is that CWT can be calculated for any value of scale. The CWT is also continuous to position, i.e. for the given scale, the analysing wavelet is shifted in the whole interval of the analysed function's domain [42].
CWT has one significant drawback – it is time-consuming – which induced the need to develop new more efficient algorithms. One of them uses fast Fourier transform to obtain CWT coefficients and is often referred to as continuous wavelet transform based on the fast Fourier transform (CWTFT). In the CWTFT algorithm, the CWT is expressed as an inverse Fourier transform:

$$C(a,b) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{X}(\omega)\hat{\Psi}^*_{a,b}(\omega)d\omega \tag{8}$$

where $\hat{X}(\omega)$ and $\hat{\Psi}^*_{a,b}(\omega)$ denote the Fourier transforms of the analysed signal *x(t)* and the wavelet at scale a and location b, respectively [43].
In the present study, six different wavelets were considered, which in Fourier domain can be defined as follows:

- the analytic Morlet wavelet:

$$\hat{\Psi}(a\omega) = \pi^{-1/4} e^{-(a\omega-\omega_0)^2/2} U(a\omega) \tag{9}$$

  where $U(\omega)$ is the Heaviside step function, and ω_0 is a frequency parameter.
- the nonanalytic Morlet wavelet:

$$\hat{\Psi}(a\omega) = \pi^{-1/4} e^{-(a\omega-\omega_0)^2/2} \tag{10}$$

- the m^{th} order derivative of Gaussian wavelets:

$$\hat{\Psi}(a\omega) = \frac{1}{\sqrt{\Gamma(m+1/2)}} (ja\omega)^m e^{-(a\omega)^2/2} \tag{11}$$

  where Γ is the gamma function. The second-order (Mexican hat wavelet) and the fourth-order derivatives of Gaussian wavelets were used.
- the m^{th} order Paul wavelet:

$$\hat{\Psi}(a\omega) = \frac{2^m}{\sqrt{m(2m-1)!}} (a\omega)^m e^{-a\omega} U(a\omega) \tag{12}$$

- Bump wavelet:

$$\hat{\Psi}(a\omega) = e^{(1-\frac{1}{1-(a\omega-\mu)^2/\sigma^2})} 1_{[(\mu-\sigma)/a,(\mu+\sigma)/a]} \tag{13}$$

where $1_{[(\mu-\sigma)/a,(\mu+\sigma)/a]}$ is the indicator function for the interval $(\mu-\sigma)/a \leq \omega \leq (\mu+\sigma)/a$ [41, 43].

Classification

Classification is a problem of machine learning that focuses on assigning observations to classes on the basis of features of these observations. The aim of classification is to construct a model that makes predictions based on the information obtained from learning carried out on observations. The training of the model is conducted by using the data set with known classes (the training set), and then the obtained model can be used for prediction of classes for the data set with unknown classes (test set) [44, 45, 46, 47].

In this work, the following classification methods were used:

- Support vector machine (SVM) – the data set is divided by finding the best hyperplane that separates data points belonging to one class from data points of other classes.
- k-nearest neighbours – assumes that observations in the data set are close to other observations with similar attributes. The algorithm searches for k nearest neighbours of the given observation and determines its label by finding the most frequent label among these neighbours.
- Decision trees – the method classifies observations by constructing a tree that assigns each observation to the class on the basis of its features. Each node of the tree corresponds to the feature of the given observation to be classified, and the branch represents the value that the node can receive [45].
- Ensemble classifiers – the basis of these algorithms is independent classification of observations by the group of classifiers. Ensemble classifiers need two kinds of information: ensemble aggregation method and a type of weak learner, such as a decision tree or k-nearest neighbours [47, 48]. Among the others, the following ensemble aggregation methods can be distinguished [49]:
  - Adaptive boosting (AdaBoost) – in each iteration the weighted classification error for all of the tested weak classifiers is measured. Then, in the next iteration, the weights of the misclassified observations are boosted [50].
  - Bootstrap aggregating (Bagging) – a decision tree is generated on the many bootstrap replicates of the analysed data set [51].
  - Random under-sampling boost (RUSBoost) – classes with a greater number of observations are under sampled; thus, for each of the weak learners, the subsets of observations of each class are analysed [52].

Data set

Primary sequences of the analysed proteins were obtained from the Protein Data Bank (PDB) [53] and UniProt database [54, 55]. The used data set (Table 2) consisted of 30 proteins [37]. The two proteins from the data set (PDB id = *1fc2* and

PDB id = *1jtg*) were excluded because of inconsistency between the acquired sequences and information given by the data set. For all residues in the data set, the change in the binding free energy ($\Delta\Delta G$) was characterised and is available in the ASEdb. Residues associated with a value $\Delta\Delta G \geq 2\,kcal/mol$ when mutated to alanine were deemed hot spots, whereas residues with $\Delta\Delta G < 0.4\,kcal/mol$ were considered as non-hot spots. Thus, the data set was found to be comprised of 219 residues of which 75 were hot spots (class HS) and 144 were non-hot spots (class NH). For each residue in the data set, a set of six features was determined according to the algorithm described below.

Algorithm used in hot spot identification

The algorithm demonstrating the application of CWTFT in the hot spot analysis is described below.

The method is based on an approach presented in [37] and [56]. It performs ASM, but computationally, by replacing the given residue (corresponding to the hot spot or non-hot spot) by alanine and looking for the differences between wild-type and mutated sequence spectra. It focuses on finding classifiers that are able to separate hot spots from non-hot spots through the given features. The features were selected from the spectra acquired by using CWTFT with the six different wavelets described above. The algorithm is presented as follows (Figure 1):

1. Mutate residue at a given position to alanine.
2. Convert wild-type and mutated amino acid sequences into numerical signals using EIIP values.
3. Compute the CWTFT coefficients (with different wavelets) of the wild-type and mutated sequences.
4. Compare the wild-type and mutated sequences by computing the difference of their CWTFT coefficients as follows:

$$D_{CWT} = |CWT_{wt} - CWT_{mut}| \tag{14}$$

   where CWT_{wt} and CWT_{mut} are the CWTFT coefficients of the wild-type and mutated sequences, respectively.
5. Extract a feature by finding the greatest difference in D_{CWT} (Figure 3).
6. Select the best classifiers that are able to discriminate hot spots from non-hot spots on the basis of the extracted features.

The performance of the algorithm is presented in Figures 2 and 3 and Table 3 using the example of Colicin E9 immunity protein ($1bxi$). Table 3 shows all considered positions of hot spots and non-hot spots of the $1bxi$ protein selected according to the value of the change in the binding free energy from the ASEdb database. Then, the sequence of the $1bxi$ protein (wild-type) and the signal obtained for this sequence are presented (Figure 2A) along with the sequence mutated for one of the hot spot positions (Glu 41) and its signal (Figure 2B). Finally, the amplitude spectra for the wild-type and mutated sequences are presented in the Figure 3A and Figure 3B, respectively. Then, the feature is extracted by computing the greatest value of the modulus of the difference between these spectra (Figure 3C). The procedure is repeated for all hot spots and non-hot spots for each of the proteins in the data set.

A number of classifiers were tested, and 10 that gave the best results were chosen [57]:

- Classifier 1 – a decision tree with many leaves that makes many fine distinctions between classes.
- Classifier 2 – a medium-complexity decision tree with fewer leaves.
- Classifier 3 – a simple decision tree with few leaves that makes coarse distinctions between classes.
- Classifier 4 – support vector machine with quadratic kernel function.
- Classifier 5 – support vector machine with Gaussian kernel function.
- Classifier 6 – k-nearest neighbours with city block distance metric.
- Classifier 7 – k-nearest neighbours with Chebyschev distance metric.
- Classifier 8 – ensemble classifier with decision tree learners and adaptive boosting ensemble method.
- Classifier 9 – ensemble classifier with decision tree learners and bagging ensemble method.
- Classifier 10 – ensemble classifier with decision tree learners and random under-sampling boost (RUSBoost) ensemble method.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Funding
Silesian University of Technology, Poland
Brno University of Technology, Czech Republic

Author's contributions
AT participated in design of study, investigation and the analysis of results. ET and IP helped to draft the manuscript and supervised the study. All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Author details
[1]Department of Biosensors and Biomedical Signals Processing, Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland. [2]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 12, 61600 Brno, Czech Republic.

References

1. Shulman-Peleg, A., Shatsky, M., Nussinov, R., Wolfson, H.J.: Spatial chemical conservation of hot spot interactions in protein-protein complexes. BMC Biol. **5**(1), 43 (2007)
2. Keskin, O., Ma, B., Nussinov, R.: Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J. Mol. Biol. **345**(5), 1281–1294 (2005)
3. Tuncbag, N., Keskin, O., Gursoy, A.: HotPoint: hot spot prediction server for protein interfaces. Nucleic Acids Res. **38**(Web Server issue), 402–406 (2010)
4. Bogan, A.A., Thorn, K.S.: Anatomy of hot spots in protein interfaces. J. Mol. Biol. **280**(1), 1–9 (1998)
5. Tuncbag, N., Gursoy, A., Keskin, O.: Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics **25**(12), 1513–1520 (2009)
6. Kortemme, T., Kim, D.E., Baker, D.: Computational alanine scanning of protein-protein interfaces. Sci. STKE **2004**(219), 2 (2004)
7. Clackson, T., Wells, J.A.: A hot spot of binding energy in a hormone-receptor interface. Science **267**(5196), 383–386 (1995)
8. Wells, J.A.: Systematic mutational analyses of protein-protein interfaces. Meth. Enzymol. **202**, 390–411 (1991)

9. Thorn, K.S., Bogan, A.A.: ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics **17**(3), 284–285 (2001)
10. Kortemme, T., Baker, D.: A simple physical model for binding energy hot spots in protein-protein complexes. Proc. Natl. Acad. Sci. U.S.A. **99**(22), 14116–14121 (2002)
11. Zhu, X., Mitchell, J.C.: KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins **79**(9), 2671–2683 (2011)
12. Moreira, I.S., Koukos, P.I., Melo, R., Almeida, J.G., Preto, A.J., Schaarschmidt, J., Trellet, M., Gümüş, Z.H., Costa, J., Bonvin, A.M.J.J.: SpotOn: High accuracy identification of Protein-Protein interface Hot-Spots. Scientific Reports **7**(1) (2017)
13. Deng, L., Guan, J., Wei, X., Yi, Y., Zhang, Q.C., Zhou, S.: Boosting prediction performance of protein-protein interaction hot spots by using structural neighborhood properties. Journal of Computational Biology **20**(11), 878–891 (2013)
14. Qiao, Y., Xiong, Y., Gao, H., Zhu, X., Chen, P.: Protein-protein interface hot spots prediction based on a hybrid feature selection strategy. BMC Bioinformatics **19**(1), 14 (2018)
15. Sable, R., Jois, S.D.: Surfing the protein-protein interaction surface using docking methods: Application to the design of ppi inhibitors. Molecules **20**(6), 11569–603 (2015)
16. Lise, S., Buchan, D., Pontil, M., Jones, D.T.: Predictions of hot spot residues at protein-protein interfaces using support vector machines. PLOS ONE **6**(2), 1–7 (2011)
17. Xia, J., Yue, Z., Di, Y., Zhu, X., Zheng, C.-H.: Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features. Oncotarget **7**(14), 18065–75 (2016)
18. Jiang, J., Wang, N., Chen, P., Zheng, C., Wang, B.: Prediction of protein hotspots from whole protein sequences by a random projection ensemble system. International Journal of Molecular Sciences **18**(7) (2017)
19. Arenas, A.F., Salcedo, G.E., Montoya, A.M., Gomez-Marin, J.E.: Msca: a spectral comparison algorithm between time series to identify protein-protein interactions. BMC Bioinformatics **16**(1), 152 (2015)
20. Liu, Q., Chen, P., Wang, B., Zhang, J., Li, J.: Hot spot prediction in protein-protein interactions by an ensemble system. BMC Systems Biology **12** (2018). doi:10.1186/s12918-018-0665-8
21. Hu, J., Zhang, X., Tang, J.: Prediction of hot regions in protein–protein interaction by combining density-based incremental clustering with feature-based classification. Computers in Biology and Medicine **61** (2015). doi:10.1016/j.compbiomed.2015.03.022
22. Garcia-Garcia, J., Valls-Comamala, V., Guney, E., Andreu, D., Muñoz, F.J., Fernandez-Fuentes, N., Oliva, B.: ifrag: A protein–protein interface prediction server based on sequence fragments. Journal of Molecular Biology **429**(3), 382–389 (2017). doi:10.1016/j.jmb.2016.11.034. Computation Resources for Molecular Biology
23. Veljkovic, V., Cosic, I., Dimitrijevic, B., Lalovic, D.: Is it possible to analyze DNA and protein sequences by the methods of digital signal processing? IEEE Trans Biomed Eng **32**(5), 337–341 (1985)
24. Cosic, I.: Macromolecular bioactivity: is it resonant interaction between macromolecules?–Theory and applications. IEEE Trans Biomed Eng **41**(12), 1101–1114 (1994)
25. Murakami, M.: Resonant recognition model of neuropeptide Y family: hot spot amino acid distribution in the sequences. J. Protein Chem. **19**(7), 609–613 (2000)
26. Cosic, I.: Analysis of HIV proteins using DSP techniques. In: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2886–2889. IEEE, Istanbul, Turkey (2001)
27. Pirogova, E., Fang, Q., Akay, M., Cosic, I.: Investigation of the structural and functional relationships of oncogene proteins. Proceedings of the IEEE **90**(12), 1859–1867 (2002)
28. Rao, K.D., Swamy, M.N.S.: Analysis of genomics and proteomics using DSP techniques. IEEE Trans Circuits Syst I Regul Pap **55-I**(1), 370–378 (2008)
29. Tamulewicz, A., Tkacz, E.: Human fibroblast growth factor 2 hot spot analysis by means of time-frequency transforms. In: Pietka, E., Badura, P., Kawa, J., Wieclawek, W. (eds.) Information Technologies in Medicine, pp. 147–159. Springer, Cham (2016)
30. Shakya, D.K., Saxena, R., Sharma, S.N.: Identification of hot spots in proteins using modified Gabor wavelet transform. Pertanika J Sci Technol **22**(2), 457–470 (2014)
31. Ramachandran, P., Antoniou, A., Vaidyanathan, P.P.: 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, pp. 1656–1660. IEEE, Pacific Grove, CA, USA, USA (2004)
32. Sahu, S.S., Panda, G.: Efficient localization of hot spots in proteins using a novel S-transform based filtering approach. IEEE/ACM Trans Comput Biol Bioinform **8**(5), 1235–1246 (2011)
33. Kasparek, J., Maderankova, D., Tkacz, E.: Protein hotspot prediction using s-transform. In: Pietka, E., Kawa, J., Wieclawek, W. (eds.) Information Technologies in Biomedicine, Volume 3, pp. 327–336. Springer, Cham (2014)
34. Ramachandran, P., Lu, W., Antoniou, A.: Filter-based methodology for the location of hot spots in proteins and exons in DNA. IEEE Trans Biomed Eng **59**(6), 1598–1609 (2012)
35. Kakumani, R., Ahmad, M.O., Devabhaktuni, V.: Prediction of hot-spots in protein sequences using statistically optimal null filters. In: Proceedings of the 10th IEEE International NEWCAS Conference, pp. 121–124. IEEE, Montreal, QC (2012)
36. Pirogova, E., Vojisavljevic, V., Caceres, J.L., Cosic, I.: Ataxin active site determination using spectral distribution of electron ion interaction potentials of amino acids. Med Biol Eng Comput **48**(4), 303–309 (2010)
37. Nguyen, Q.T., Fablet, R., Pastor, D.: Protein interaction hotspot identification using sequence-based frequency-derived features. IEEE Trans Biomed Eng **60**(11), 2993–3002 (2013)
38. Lazovic, J.: Selection of amino acid parameters for Fourier transform-based analysis of proteins. Comput. Appl. Biosci. **12**(6), 553–562 (1996)
39. The MathWorks, Inc.: Continuous Wavelet Transform and Scale-Based Analysis. https://www.mathworks.com/help/wavelet/gs/continuous-wave

let-transform-and-scale-based-analysis.html. (Online; Access date: April 14th, 2020)
40. Komorowski, D., Pietraszek, S.: The use of continuous wavelet transform based on the fast fourier transform in the analysis of multi-channel electrogastrography recordings. J Med Syst **40**(1), 10 (2015)
41. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc. **79**(1), 61–78 (1998)
42. Białasiewicz, J.: Falki i Aproksymacje. Wydawnictwa Naukowo-Techniczne, Warszawa (2000). (in Polish)
43. The MathWorks, Inc.: CWTFT. https://www.mathworks.com/help/wavelet/ref/cwtft.html. (Online; Access date: April 14th, 2020)
44. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, Berlin, Heidelberg (2006)
45. Kotsiantis, S.B.: Supervised machine learning: A review of classification techniques. In: Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies, pp. 3–24 (2007)
46. The MathWorks, Inc.: Supervised Learning Workflow and Algorithms. https://www.mathworks.com/help/stats/supervised-learning-machine-learning-workflow-and-algorithms.html. (Online; Access date: April 14th, 2020)
47. Woźniak, M.: Classifier ensemble – recent research directions. Mechanik (10), 833–4483348 (2015)
48. The MathWorks, Inc.: Framework for Ensemble Learning. https://www.mathworks.com/help/stats/framework-for-ensemble-learning.html. (Online; Access date: April 14th, 2020)
49. The MathWorks, Inc.: Ensemble Algorithms. https://www.mathworks.com/help/stats/ensemble-algorithms.html. (Online; Access date: April 14th, 2020)
50. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
51. Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996)
52. Seiffert, C., Khoshgoftaar, T., Van Hulse, J., Napolitano, A.: Rusboost: Improving classification performance when training data is skewed, pp. 1–4 (2008)
53. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. **28**(1), 235–242 (2000)
54. The UniProt Consortium: UniProt: a hub for protein information. Nucleic Acids Res. **43**(Database issue), 204–212 (2015)
55. European Bioinformatics Institute (EMBL-EBI), Swiss Institute of Bioinformatics (SIB), Protein Information Resource (PIR): The Universal Protein Resource (UniProt). http://www.uniprot.org. (Online; Access date: April 14th, 2020)
56. Cho, K.I., Kim, D., Lee, D.: A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res. **37**(8), 2672–2687 (2009)
57. The MathWorks, Inc.: Classification Learner App. https://www.mathworks.com/help/stats/choose-a-classifier.html. (Online; Access date: April 14th, 2020)

Figures

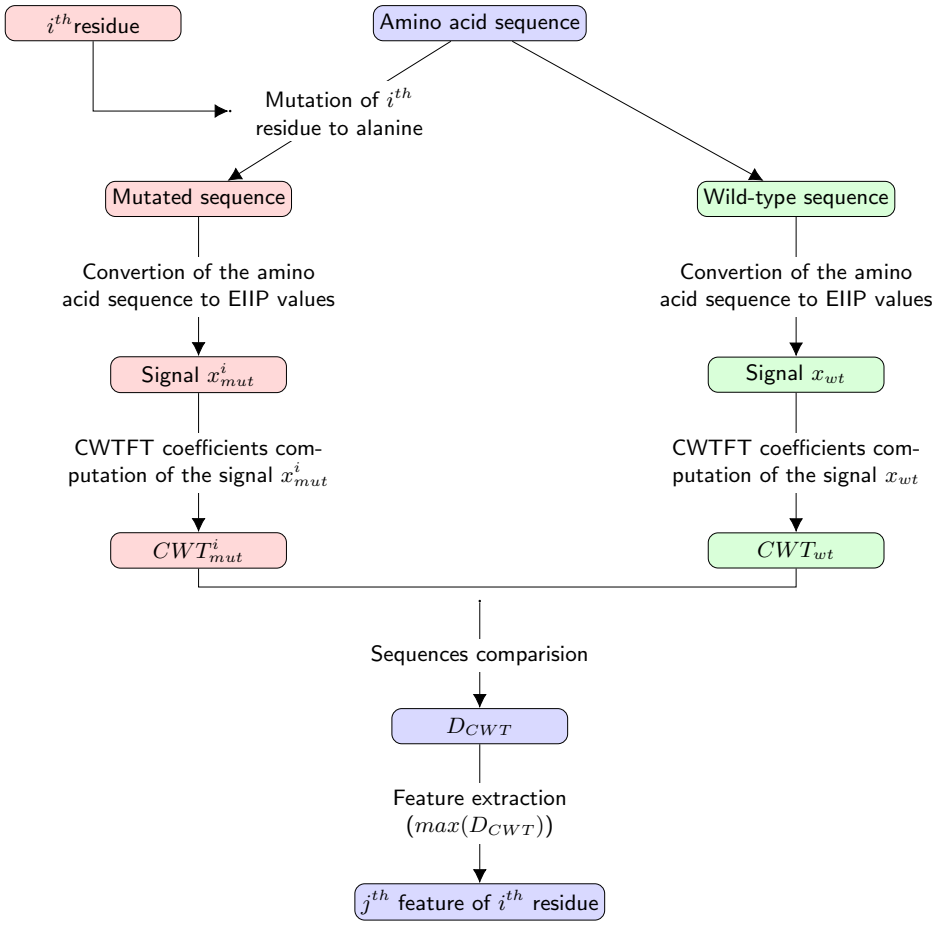

Figure 1 Algorithm used in hot spot identification. The extraction of j^{th} feature of the i^{th} amino acid residue.

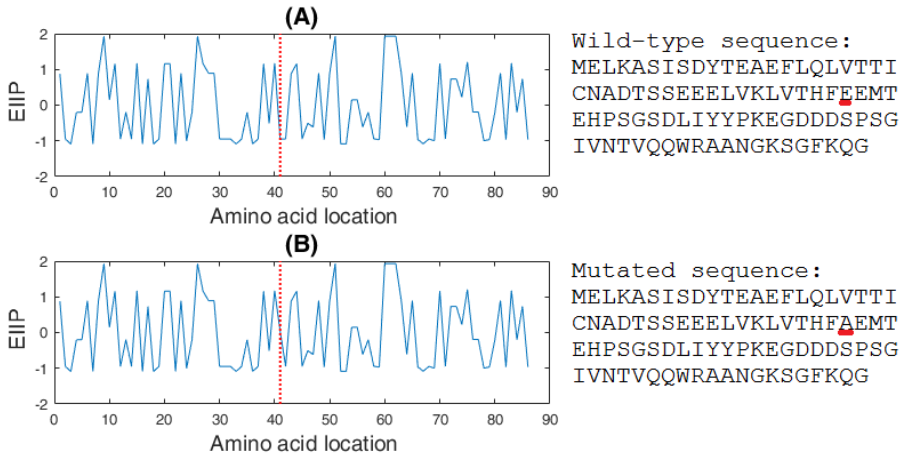

Figure 2 The signals obtained for the wild-type and mutated sequences of the example protein Colicin E9 immunity protein ($1bxi$). (A) The wild-type sequence and the signal obtained for this sequence. (B) The sequence mutated for one of the hot spot positions (Glu 41) and the signal obtained for this sequence. At the plot, the mutated position is marked by the red vertical line, while at the protein sequences, the hot spot position are underlined.

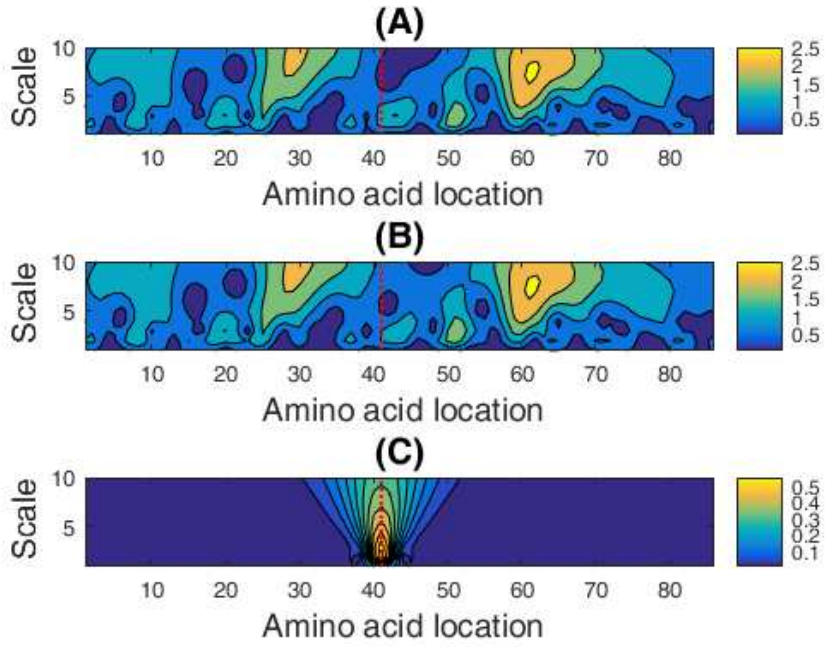

Figure 3 The feature extraction on the example of Colicin E9 immunity protein ($1bxi$). Modulus of CWTFT coefficients (amplitude spectra) obtained by using Paul wavelet for (A) the wild-type sequence and (B) the sequence mutated for one of the hot spot positions (Glu 41). (C) Modulus of the difference between CWTFT coefficients of the wild-type and mutated sequences. Mutated hot spot position is marked by the red vertical line.

Tables

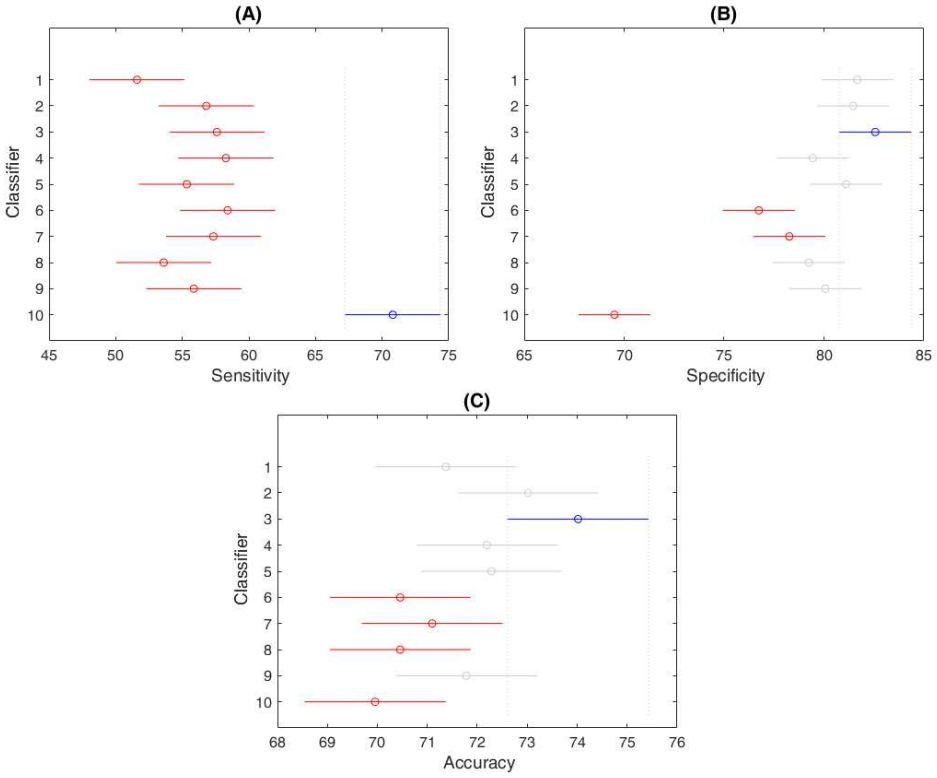

Figure 4 The confidence intervals for comparison of (A) sensitivities, (B) specificities and (C) accuracies of classifiers. The best classifier for each evaluation metrics is in blue. Red indicates the classifiers which means are significantly different than the mean of the classifier in blue. Each interval is a 95% confidence interval for the mean of a group.

Table 1 EIIP values of amino acids.

Amino acid	3-letter code	1-letter code	EIIP [Ry]
Alanine	Ala	A	0.0373
Cysteine	Cys	C	0.0829
Aspartic acid	Asp	D	0.1263
Glutamic acid	Glu	E	0.0058
Phenylalanine	Phe	F	0.0946
Glycine	Gly	G	0.0050
Histidine	His	H	0.0242
Isoleucine	Ile	I	0.0000
Lysine	Lys	K	0.0371
Leucine	Leu	L	0.0000
Methionine	Met	M	0.0823
Asparagine	Asn	N	0.0036
Proline	Pro	P	0.0198
Glutamine	Gln	Q	0.0761
Arginine	Arg	R	0.0959
Serine	Ser	S	0.0829
Threonine	Thr	T	0.0941
Valine	Val	V	0.0057
Tryptophan	Trp	W	0.0548
Tyrosine	Tyr	Y	0.0516

Table 2 Proteins in the dataset.

PDB id	Chain	Molecule	Number of hot spots	Number of non-hot spots
1a4y	A	Ribonuclease inhibitor	2	5
1a4y	B	Angiogenin	1	7
1a22	A	Growth hormone	3	14
1a22	B	Growth hormone receptor	4	15
1ahw	C	Tissue factor	1	3
1brs	A	Barnase	7	1
1brs	D	Barstar	3	0
1bxi	A	Colicin E9 immunity protein	7	3
1cbw	D	BPTI	1	2
1cbw	I	BPTI	1	4
1dan	L	Blood coagulation factor VIIA light chain	0	10
1dan	T	Soluble tissue factor	2	10
1dan	U	Soluble tissue factor	0	8
1dvf	A	FV D1.3	1	1
1dvf	B	FV D1.3	5	0
1f47	A	Cell division protein FtsZ	3	1
1fcc	C	Streptococcal protein G (C2 fragment)	4	2
1gc1	C	CD4	0	11
1jrh	I	Interferon-gamma receptor alpha chain	5	4
1jtg	B	Beta-lactamase inhibitory protein	2	1
1nmb	L	FAB NC10	0	1
1vfb	A	IGG1-KAPPA D1.3 Fv (light chain)	0	2
1vfb	B	IGG1-KAPPA D1.3 Fv (heavy chain)	2	2
1vfb	C	Hen egg white lysozyme	1	3
2ptc	I	Trypsin inhibitor	1	0
3hfm	H	HYHEL-10 IGG1 FAB (heavy chain)	4	1
3hfm	L	HYHEL-10 IGG1 FAB (light chain)	4	0
3hfm	Y	Hen egg white lysozyme	3	6
3hhr	A	Human growth hormone	3	17
3hhr	B	Human growth hormone receptor (hGHbp)	5	10

Table 3 The analysed residues of the $1bxi$ sequence. HS and NH denote hot spots and non-hot spots, respectively, whereas $\Delta\Delta G$ is the change in the binding free energy from the ASEdb database.

Amino acid type	Residue position	$\Delta\Delta G$	HS/NH
N	24	0,14	NH
S	28	0,17	NH
L	33	3,42	HS
V	34	2,58	HS
E	41	2,08	HS
S	48	0,01	NH
S	50	2,19	HS
D	51	5,92	HS
Y	54	4,83	HS
Y	55	4,63	HS

Table 4 The evaluation metrics for the identification of hot spots with the aid of the presented algorithm. Mean values with standard error (se) from 10 iterations are presented. The biggest values of sensitivity, specificity and accuracy are in bold.

	Sensitivity [%]	Specificity [%]	Accuracy [%]
Classifier 1	51.60 ($se = 1.45$)	81.67 ($se = 0.78$)	71.37 ($se = 0.77$)
Classifier 2	56.80 ($se = 2.04$)	81.46 ($se = 1.31$)	73.01 ($se = 0.64$)
Classifier 3	57.60 ($se = 2.18$)	**82.57** ($se = 1.29$)	**74.02** ($se = 0.63$)
Classifier 4	58.27 ($se = 0.82$)	79.44 ($se = 0.38$)	72.19 ($se = 0.33$)
Classifier 5	55.33 ($se = 1.41$)	81.11 ($se = 0.44$)	72.28 ($se = 0.60$)
Classifier 6	58.40 ($se = 2.35$)	76.74 ($se = 0.65$)	70.46 ($se = 0.73$)
Classifier 7	57.33 ($se = 1.26$)	78.26 ($se = 0.46$)	71.10 ($se = 0.62$)
Classifier 8	53.60 ($se = 0.91$)	79.24 ($se = 0.64$)	70.46 ($se = 0.51$)
Classifier 9	55.87 ($se = 1.25$)	80.07 ($se = 0.46$)	71.78 ($se = 0.61$)
Classifier 10	**70.80** ($se = 1.06$)	69.51 ($se = 0.78$)	69.95 ($se = 0.6$)

Table 5 P-values for the null hypothesis that the means of sensitivities of the classifiers are equal. The p-values below the significance level of $\alpha = 0.05$ are in bold.

Classifier	1	2	3	4	5	6	7	8	9
2	0.36								
3	0.18	1							
4	0.09	1	1						
5	0.79	1	0.99	0.94					
6	0.08	1	1	1	0.92				
7	0.23	1	1	1	0.99	1			
8	0.99	0.91	0.72	0.52	1	0.48	0.79		
9	0.64	1	1	0.98	1	0.98	1	0.99	
10	**1e-07**	**5e-07**	**2e-06**	**7e-06**	**1e-07**	**9e-06**	**1e-06**	**1e-07**	**2e-07**

Table 6 P-values for the null hypothesis that the means of specificities of the classifiers are equal. The p-values below the significance level of $\alpha = 0.05$ are in bold.

Classifier	1	2	3	4	5	6	7	8	9
2	1								
3	1	0.99							
4	0.60	0.73	0.15						
5	1	1	0.95	0.89					
6	**1e-03**	**2e-03**	**4e-05**	0.32	**0.01**				
7	0.08	0.13	**0.01**	0.99	0.25	0.93			
8	0.48	0.60	0.10	1	0.80	0.43	1		
9	0.91	0.96	0.43	1	0.99	0.10	0.83	1	
10	**1e-07**	**1e-07**	**1e-07**	**1e-07**	**1e-07**	**3e-07**	**1e-07**	**1e-07**	**1e-07**

Table 7 P-values for the null hypothesis that the means of accuracies of the classifiers are equal. The p-values below the significance level of $\alpha = 0.05$ are in bold.

Classifier	1	2	3	4	5	6	7	8	9
2	0.67								
3	0.08	0.98							
4	0.99	0.99	0.53						
5	0.99	1	0.60	1					
6	0.99	0.11	**3e-03**	0.60	0.53				
7	1	0.46	**0.04**	0.96	0.93	1			
8	0.99	0.11	**3e-03**	0.60	0.53	1	1		
9	1	0.92	0.24	1	1	0.88	1	0.88	
10	0.83	**0.02**	**4e-04**	0.24	0.20	1	0.95	1	0.53

Table 8 Comparison of results achieved by the best two classifiers (Classifier 3 and 10) selected from 10 described in this work and algorithms using similar methods.

	Sensitivity [%]	Specificity [%]	Accuracy [%]
Nguyen *et al.* [37]	58.9 - 64.9	89.6 - 91.5	79.0 - 82.4
Cho *et al.* [56]	58	89	*not reported*
Shakya *et al.* [30]	70	65	67
Classifier 3	57.60	**82.57**	**74.02**
Classifier 10	**70.80**	69.51	69.95

Figures

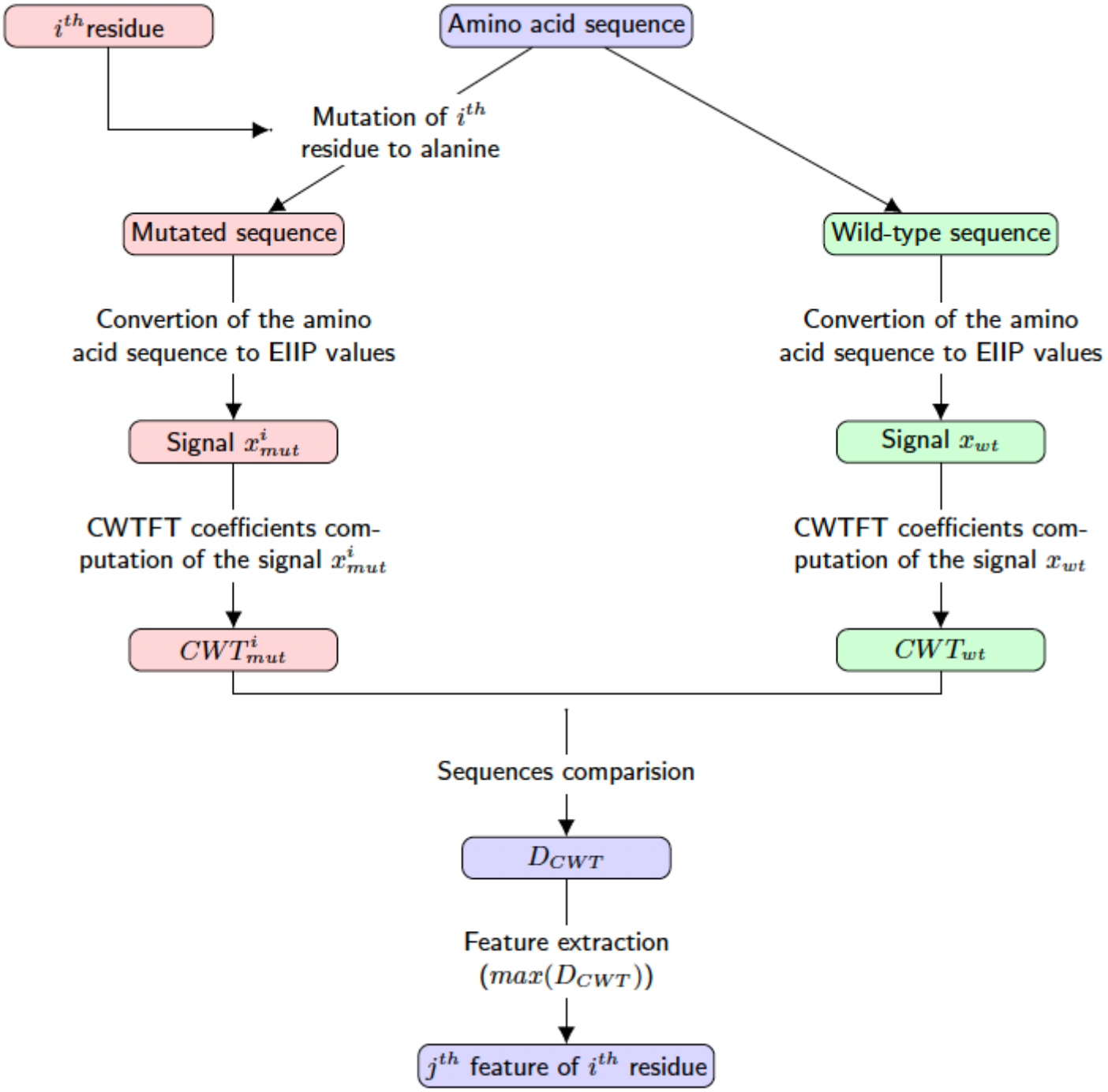

Figure 1

Algorithm used in hot spot identification. The extraction of jth feature of the ith amino acid residue.

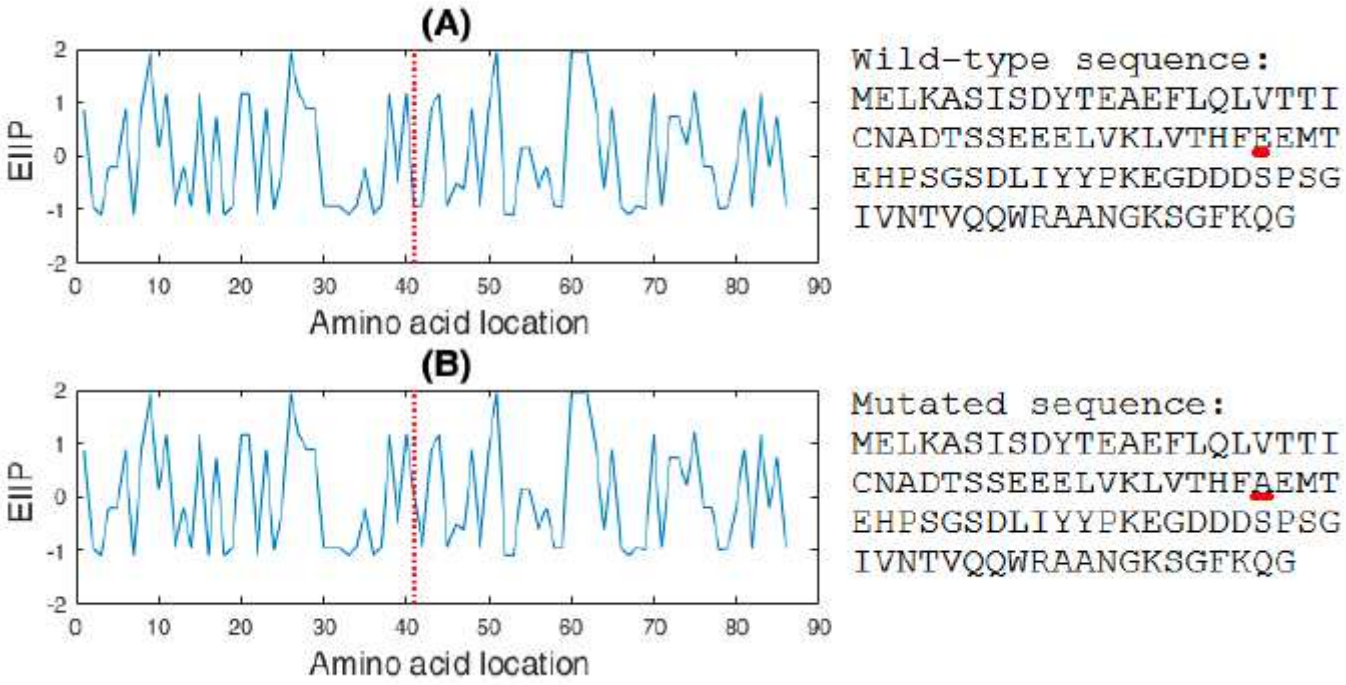

Figure 2

The signals obtained for the wild-type and mutated sequences of the example protein Colicin E9 immunity protein (1bxi). (A) The wild-type sequence and the signal obtained for this sequence. (B) The sequence mutated for one of the hot spot positions (Glu 41) and the signal obtained for this sequence. At the plot, the mutated position is marked by the red vertical line, while at the protein sequences, the hot spot position are underlined.

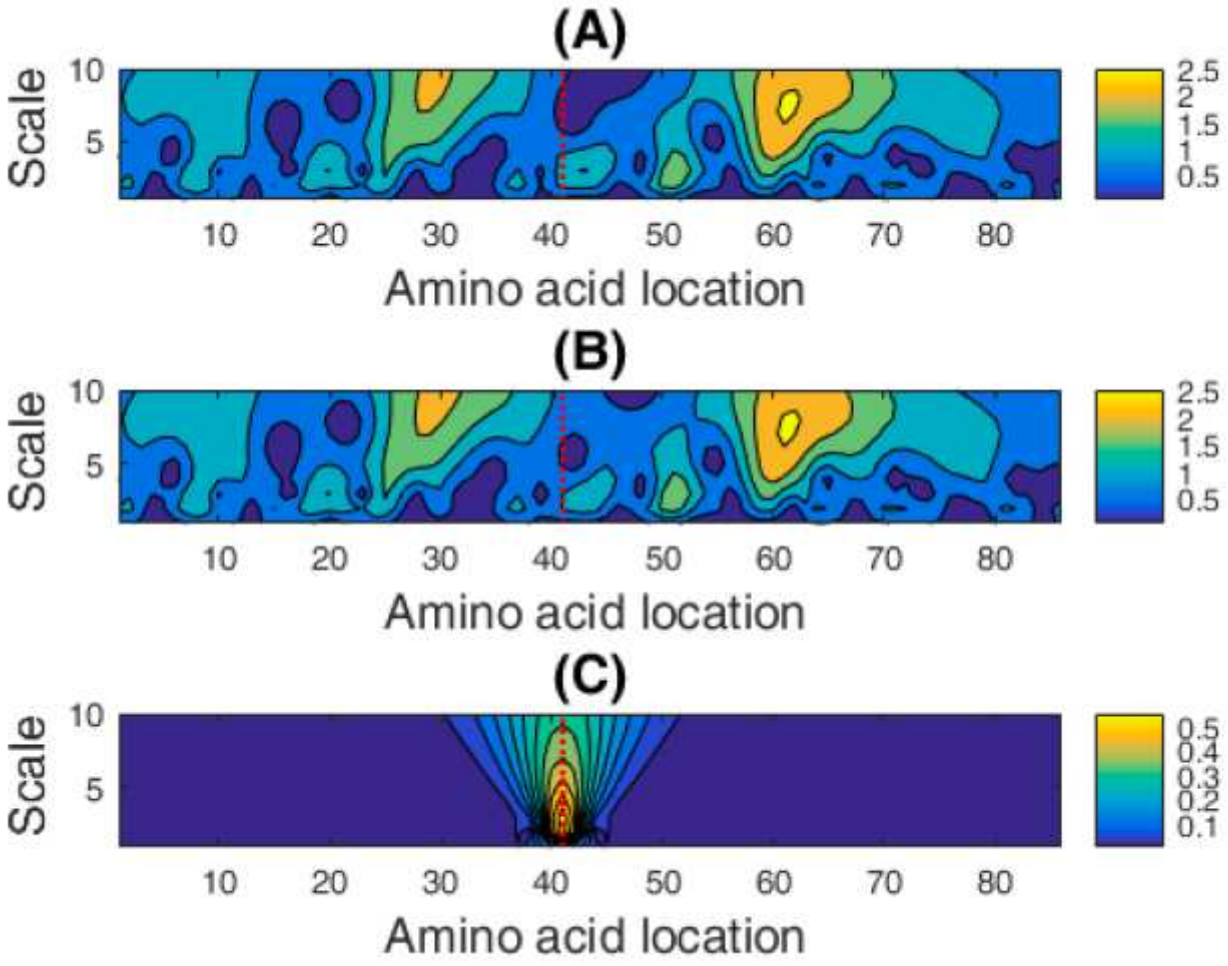

Figure 3

The feature extraction on the example of Colicin E9 immunity protein (1bxi). Modulus of CWTFT coefficients (amplitude spectra) obtained by using Paul wavelet for (A) the wild-type sequence and (B) the sequence mutated for one of the hot spot positions (Glu 41). (C) Modulus of the difference between CWTFT coefficients of the wild-type and mutated sequences. Mutated hot spot position is marked by the red vertical line.

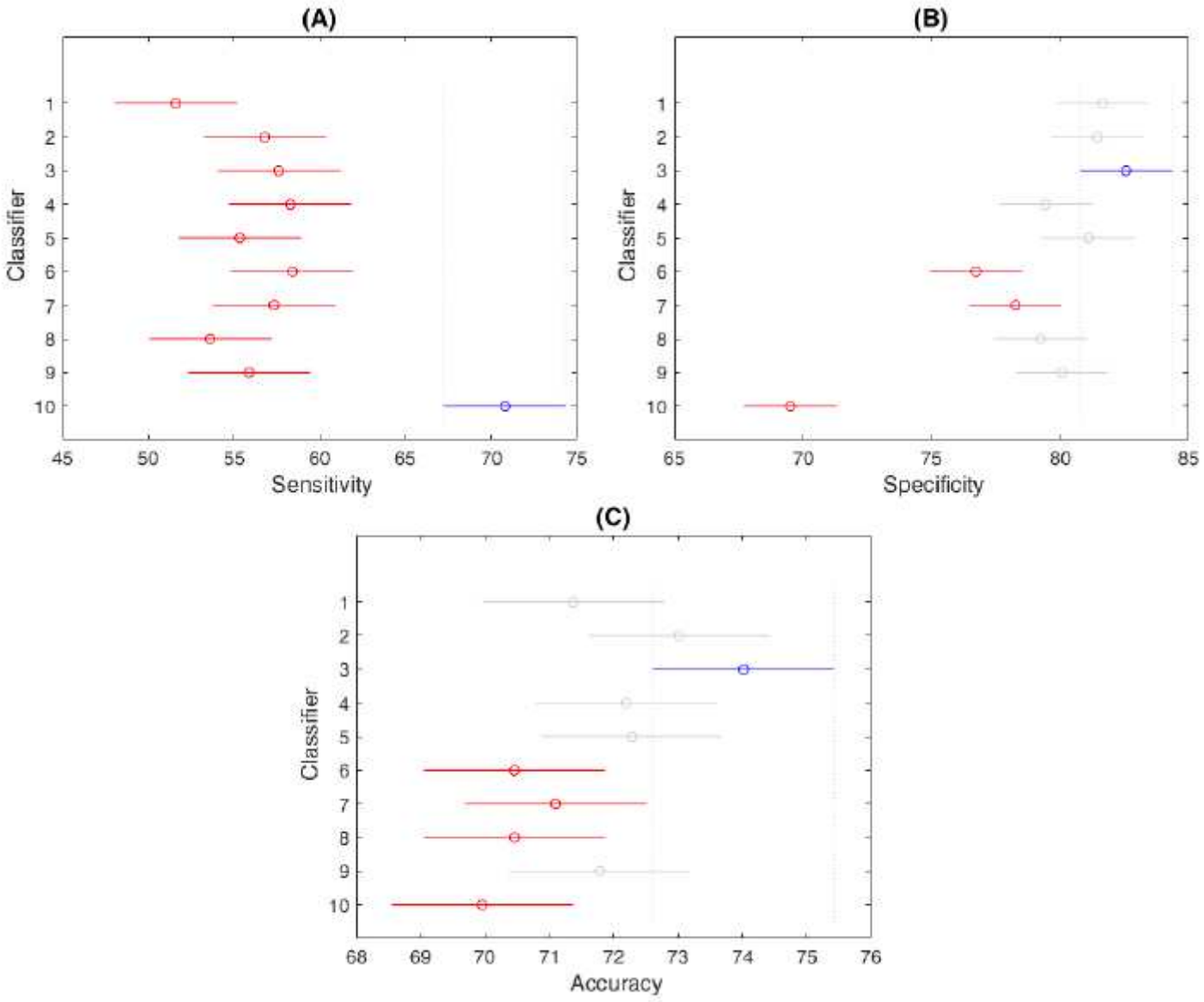

Figure 4

The confidence intervals for comparison of (A) sensitivities, (B) specificities and (C) accuracies of classifiers. The best classifier for each evaluation metrics is in blue. Red indicates the classifiers which means are significantly different than the mean of the classifier in blue. Each interval is a 95% confidence interval for the mean of a group.