Prediction of Protein-Protein Interactions from Sequences using a Correlation Matrix of the Physicochemical Properties of Amino Acids

Charlemagne N'Diffon **Kopoin**[1] , Armand Kodjo **Atiampo**[2], Behou Gerard **N'Guessan**[2] **and** Michel **Babri**[1]

[1]Institut National Polytechnique Felix Houphouet Boigny, Cote d'Ivoire; [2]Université Virtuelle de Cote d'Ivoire, Cote d'Ivoire charlemagnekopoin@gmail.com

Summary

Detection of protein-protein interactions (PPIs) remains essential for the development of therapies against diseases. Experimental studies to detect PPI are longer and more expensive. Today, with the availability of PPI data, several computer models for predicting PPIs have been proposed. One of the big challenges in this task is feature extraction. The relevance of the information extracted by some extraction techniques remains limited. In this work, we first propose an extraction method based on correlation relationships between the physicochemical properties of amino acids. The proposed method uses a correlation matrix obtained from the hydrophobicity and hydrophilicity properties that it then integrates in the calculation of the bigram. Then, we use the SVM algorithm to detect the presence of an interaction between 2 given proteins. Experimental results show that the proposed method obtains better performances compared to the approaches in the literature. It obtains performances of 94.75% in accuracy, 95.12% in precision and 96% in sensitivity on human HPRD protein data.

Key words:

Feature extraction, bigram, NLP, protein-protein interaction

1. Introduction

Protein is the essential building block of the living organism and is involved in various processes of life activities such as metabolism, signal transduction, hormonal regulation, transcription, and DNA replication. In general, proteins perform their complex functions by interacting with other proteins. The study of protein-protein interactions (PPIs) not only helps to understand the process of life, but also to explore the parthenogenesis of diseases and helps target drugs for new diseases such as covid-19. Certain high-throughput proteomic techniques such as proteomic chips [1], immunoprecipitation, the two yeasts hybrid technique [2] were invented to detect PPIs. All these limitations are at the root of the motivation for developing computer models to predict PPIs on a large scale and efficiently.

To date, many computational approaches have been proposed to predict PPIs from different types of data, including genetic ontology and gene annotation [3], 3D structural information, and so on. However, these approaches are not universal, and their accuracy and reliability depend heavily on prior information collected on proteins. In practice, the 3D structure of many proteins is unknown, the ontology and annotation of genes are incomplete, and PPIs for many species are rarely available. With the rapid development of sequencing techniques, protein sequence information is collected and stored in large quantities in databases such as the Human Protein Reference Database (HPRD) [4], the Protein Interaction Database (PID) [5], the Molecular Interaction Database (IntAct) [6] and the Biomolecular Interaction Network Database (BIND) [7]. However, one of the major difficulties in setting up a computer model for PPI prediction that uses protein sequence information is feature extraction. This essential step consists in transforming the information of the protein sequence generally coded with letters of the alphabet into useful numerical data. Chou [8] proposed the pseudo amino acid composition method (PAAC) to improve the quality of prediction of subcellular localization of proteins and membrane protein types. Its goal is to continue to use a discrete model to represent a protein without losing completely information about its sequence order. The PAAC method generates 20 + *m* components, of which the first 20 are the 20 components of the amino acid composition and the last *m* components are the sequential order components. Guo et al. [9] applied the autocovariance (AC) method to discover information in discontinuous amino acid sequence segments. As a result of classification, they obtained 86.55% in accuracy on the PPIs of the S. Cerevisiae dataset. You et al. [10] used an amino acid substitution matrix to extract characteristics and then applied rotation forest set classifiers [11] to predict PPIs. This method achieved an accuracy of 90.06%, sensitivity of 85.74% and specificity of 94.37% in the yeast protein dataset. Pan et al. [12] proposed a new hierarchical model (LDA-RF) to directly predict protein-protein interactions in primary protein sequences. Their approach allows the extraction of internal structures hidden in amino acid sequences. Experimental results show that this model can efficiently predict potential protein interactions.

The Bigram method is a simple NLP (Natural Language Processing) [13] method used in the extraction of features from sequences. It allows to have two-by-two combinations of amino acid residues along a sequence. For

Manuscript received March 5, 2021
Manuscript revised March 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.3.6

example, the bigram of amino acid *i* and amino acid *j* will be represented by the frequency of occurrence of the transition from the *i*-th amino acid to the *j*-th amino acid. However, when applied directly to the primary sequence of the protein, it can produce a characteristic vector with many zeros, which can cause numerical instability during the training phase of machine learning algorithms. To deal with it, some authors [14]–[16] use PSSM values to replace residues. PSSM is a method that was first introduced by Gribskov et al. in 1987 [17]. Its particularity is that it provides information on the probability of substitution of a given amino acid according to its specific position along the sequence with the 20 amino acids of the genetic code. PSSM values can be obtained from the online PSI-BLAST tool. Göktepe and Kodaz [14] used the triad and bigram methods combined with PSSM values to extract the features and used an SVM to predict human PPIs from the HPRD database. Their model obtained 93.45% in accuracy, 89.84% in precision, 89.29% in sensitivity and 85.71% in Mcc. In Kopoin et al. [18], a distance function from the values of hydrophobicity [19] and hydrophilicity [20] of amino acids was proposed to replace amino acid residues in the calculation of bigram. However, this method takes a long time to perform.

In this work, we propose a method that quickly calculates the bigram with correlation values from the physicochemical properties of amino acids instead of amino acid residues. We have developed a correlation matrix obtained using the values of the hydrophobicity and hydrophilicity properties. This matrix is then used for the calculation of the bigram to have a much more informative and interesting characteristic vector.

The rest of the document is organized as follows. We present in section 2 the data sets used. Section 3 is devoted to the detailed study of the proposed solution. Section 4 deals with the analysis of the results obtained and section 5 is devoted to the conclusion and future work.

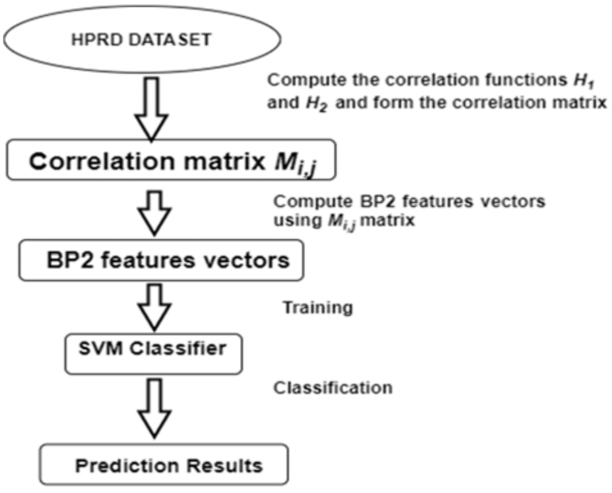
HPRD DATASET
Compute the correlation functions H_1 and H_2 and form the correlation matrix
Correlation matrix $M_{i,j}$
Compute BP2 features vectors using $M_{i,j}$ matrix
BP2 features vectors
Training
SVM Classifier
Classification
Prediction Results

Fig. 1 Flowchart of our study

2. Materials

2.1 Benchmark dataset

In this work, we propose the implementation of a PPI prediction model using only protein sequences. To do this, we need datasets that contain positive PPIs (interacting proteins) and negative PPIs (non-interacting proteins). The HPRD dataset is a reference dataset of human PPIs built from the work of Pan et al. [12], which can be viewed at csbio server[1]. The HPRD database is one of the most comprehensive databases for human PPI data. To ensure accurate representativeness of the data, double pairs of PPIs and pairs with sequences consisting of less than 50 amino acids or multiple locations have been excluded. The positive data for this dataset (a total of 36630 positive interactions from 9630 different human proteins) were collected from the Human Protein Reference Database (HPRD, 2007 version). To form negative pairs, proteins in distinct subcellular locations were paired. This localization information was obtained by selecting only human proteins from version 57.3 of the Swiss-Prot database[2]. Indeed, one of the most common methods for obtaining negative PPI pairs is based on cell localization annotations [21]. Information on the cellular localization of proteins tells us that a protein can be divided into several types of proteins: nucleus, cytoplasm, endoplasmic reticulum, mitochondrion, Golgi apparatus, vacuole, and peroxisome. This information can be obtained from Swiss-Prot. PPI negative pairs are obtained by pairing proteins of one localization with proteins of another localization. Thus, 36480 negative interactions were obtained from 1773 proteins from 6 subcellular locations.

2.2 Other datasets

To show the robustness of our method, we used four other different PPI datasets that are commonly used in the prediction of PPIs. The first set of IPP data is Homo Sapien dataset (H. Sapien), also collected from the HPRD data, described by Huang et al. [22], which consists of 8161 pairs of human proteins (3899 interacting pairs and 4262 non-interacting pairs). The second dataset is the IPP dataset described by You et al., [23]. This dataset is collected from the basic subset of S. Cerevisiae in the Protein Interaction Database (PID). This dataset consists of 5594 positive pairs and 5594 negative pairs, for a total of 11188 protein pairs. The third set of data is H. Pylori dataset that proposed by Martin et al. [24], consisting of 2916 pairs of proteins including 1458 pairs which interact

[1] http://www.csbio.sjtu.edu.cn/bioinf/LR_PP/Data.htm
[2] https://www.expasy.org/sprot/

and 1458 pairs which do not interact. The last datasets namely E. Coli [25] consist of 6954 positive pairs.

The various IPP datasets are in FASTA [26] format. The FASTA format facilitates manipulation and analysis of sequences using word processing tools and scripting languages such as the R programming language, Python, Ruby, and Perl.

3. Methods

3.1 Proposed extraction method

The extraction method that we propose in this study will allow us to transform a sequence of characters into a sequence of numerical values reflecting the characteristic binding properties of any pair of amino acids. It combines the bigram technique and a correlation matrix obtained from the values of two physicochemical properties of amino acids which are hydrophobicity and hydrophilicity. This method, called BP2, is a variant of the BP method [18] which uses the values of a distance function obtained from the hydrophobicity and hydrophilicity values of the amino acids to calculate the bigram. Know that the hydrophobicity and hydrophilicity of the amino acids of a protein play a very important role in its folding, its interaction with other molecules, its structure as well as its catalytic function [27].

To calculate the bigram values, we do as follow:

Consider a protein *P* composed of *L* amino acid residues:

$$R_1 R_2 R_3 \ldots R_{L-1} R_L \tag{1}$$

with *R*₁, the residue at position 1 of the chain, *R*₂, the residue at position 2 of the chain, *R*₃, the residue at position 3 of the chain and so on. First, we use the derived values of hydrophobicity and hydrophilicity [20] to calculate their correlated functions (the original values can be found in [28]). Suppose that H_1^0 and H_2^0, are the original hydrophobicity value and the original hydrophilicity value of amino acid *R*i (i = 1, 2 ..., 20), respectively, the derived values are calculated as follows:

$$\begin{cases} H_1^*(R_i) = \dfrac{H_1^0(R_i)-\mu_1}{\sqrt{\sum_{i=1}^{20}\left[H_1^0(R_i)-\mu_1\right]^2 / 20}} \\ H_2^*(R_i) = \dfrac{H_2^0(R_i)-\mu_2}{\sqrt{\sum_{i=1}^{20}\left[H_2^0(R_i)-\mu_2\right]^2 / 20}} \end{cases} \tag{2}$$

where μ_1 and μ_2 are the mean of the hydrophobicity and hydrophilicity values of the 20 amino acids, respectively, H_1^* and H_2^* are the correlated hydrophobicity and hydrophilicity functions defined as follows:

$$H_{i,j}^1 = H_1^*(R_i) \times H_1^*(R_j) \quad ; \quad H_{i,j}^2 = H_2^*(R_i) \times H_2^*(R_j) \tag{3}$$

Next, we will add the correlated hydrophobicity and hydrophilicity functions, denoted *CF*, according to the equation below:

$$CF_{i,j} = H_{i,j}^1 + H_{i,j}^2 \tag{4}$$

Now, to represent the different correlations along the protein of length L, we define a matrix M calculated as follows:

$$M_{i,j} = \frac{1}{i} CF_{i,j}, \qquad 1 \le i \le L; \; 1 \le j \le 20 \tag{5}$$

where $\frac{1}{i}$ represents is a rank weighting function.

Finally, the frequency of occurrence BP2 of the transition from the *i-th* amino acid to the *j-th* amino acid is calculated as follows:

$$BP_{2(ij)} = \sum_{t=1}^{L-1} M_{t,i} \times M_{t+1,j}\,, \qquad 1 \le i,j \le 20 \tag{6}$$

with L the length of the sequence, $M_{t,i}$, the value of the correlation matrix in the *t-th* row and *i-th* column and $M_{t+1,j}$, the value of the correlation matrix in the (*t*+1)-*th* row and *j-th* column.

This method applied to a protein sequence generates a 400-D vector. To represent the pair of proteins, we concatenate the vector of each protein, resulting in a final 800-D vector.

3.2 Classification with SVM

In this study, we used the SVM algorithm [9], [28]–[30], widely used in the literature for PPIs prediction. A SVM can be a problem of finding a hyperplane of equation $f(x) = \omega x + b$, (with ω, a weight function) allowing to separate positive and negative observations. To do this, the notion of geometric margin is introduced and represents the distance between the separating band and the nearest points, called support vectors. The hyperplane is chosen to maximize the margin to better generalize to new observations. The SVM algorithm belongs to the class of supervised learning and kernel methods [31]. There are four main types of kernels including linear kernel, Polynomial kernel, Laplacian kernel, and Gaussian kernel.

The classification performance of an SVM model strongly depends on three main parameters, including the capacitance parameter C, the gamma parameter (γ) and the kernel type *k*. The parameter C controls the trade-off between a fluid decision limit and error minimization. The gamma parameter (γ) determines the extent of influence of a single training example. The kernel *k* determines the learning ability of the SVM. For our case study, we used the Gaussian kernel because it has been shown to be an optimized option in most cases by previous studies, especially when the number of cases far exceeds the number of characteristics [12]. The parameters *C* and (γ) were optimized via a grid search. Finally, we have best parameter *C* equal to 2 and (γ) equal to 1.

4. Results and analysis

The codes were made with the python 3.7 language. The experiments were carried out on a machine with an i7 processor with 8 GB of RAM.

To evaluate our model, we used the following measures: Accuracy (Acc), Precision (Pre), Sensitivity (Sen), Matthews Correlation Coefficient (Mcc), Area under the ROC curve (AUC). Some of these measures are defined as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$Pre = \frac{TP}{TP+FP} \tag{8}$$

$$Sen = \frac{TN}{TN+FN} \tag{9}$$

$$Mcc = \frac{TP \times N - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \tag{10}$$

TP (true positive) is the number of predicted positive PPIs, i.e., interact really, FP (false positive) is the number of predicted positive PPIs, but are negative really, TN (true negative) is the number of PPI predicted negative, and which are negative really, and FN (false negative) is the number of PPI predicted negative, but are positive really. MCC is a measure of the quality of the binary classification, which is a correlation from the coefficient between observed and predicted results. It returns a value between -1 (is considered a false prediction) and +1 (is considered an interesting prediction). The ROC curve and AUC value graphically illustrate the performance of a binary classification system.

First, we used 5 cross-validations on human PPI data to evaluate the performance of our model and to avoid overlearning.

In Table 1, we can see that the best performance is obtained in Step 4 with 95.68% in accuracy, 95.35% in precision, 96.03% in sensitivity, 96.03% in Mcc and 95.58% in Auc. For all 5 cross-validations, we obtain respectively 94.75%, 95.12%, 96%, 95.01% and 95.55% in accuracy, precision, sensitivity, Mcc and AuC, respectively.

Table 1: Results of cross-validation on the HPRD dataset

Fold	*Acc*	*Pre*	*Sen*	*Mcc*	*Auc*
1	94.48%	94.98%	96%	94.77%	95.48%
2	95.51%	95.01%	95.58%	94.13%	95.51%
3	93.67%	95.15%	95.67%	94.77%	95.57%
4	95.68%	95.35%	95.98%	95.02%	95.58%
5	94.42%	95.05%	96.03 %	95.69%	95.52%
mean	**94.75%**	**95.12%**	**96%**	**95.01%**	**95.55%**
std	**± 0.17%**	**± 0.23%**	**± 0.14%**	**± 0.07%**	**±0.12%**

Std mean standard deviation

In the following, we compare the performances of our method with those of certain methods of the literature that we have implemented: PAAC [8], APAAC [27], AC [9], CTD [32], BP [18] and Res2Vec [33]. The source codes for the PAAC and APAAC methods are available through csbio server (see section2). These two methods take a parameter λ, which indicates the order-sequence level. Both methods are computationally intensive as the parameter λ increases. λ is equal to 20 in the experiment as in [14]. The AC method uses the values of the physicochemical properties as well. In our case we have chosen six physicochemical properties as in [28], include hydrophobicity, hydrophilicity, polarity, polarizability, solvent-accessible surface area and net charge index of side chains. For the Res2vec method, we acquired the source code through GitHub[3]. This method uses two parameters including the size of the residue and the size of the window. For our case, we used residual dimension equal to 20 and a window size equal to 4 as in [33].

In the table below, we compare the performance results of our method with those of the other methods cited above on the HPRD data. We can see that our model performs 96.96% in accuracy, 95.97% in precision, 96.09% in sensitivity and 94.76% in Mcc. Overall, our method shows superior performance with 0.36% more in accuracy, 0.5% more in precision, 0.07% more in sensitivity and 0.03% more in Mcc.

Table 2: Performance results comparison on HPRD dataset

Method	*Acc*	*Pre*	*Sen*	*Mcc*
AC	94.48%	95.08%	95,10%	94.73%
CTD	80.23%	81.44%	86.21%	80.12%
APAAC	95.75%	95.71%	95.78%	94.73%
BP	94.67%	95.05%	94.67%	94.37%
PAAC	83.64%	82.35%	84.03%	83.02%
Res2Vec	94.44%	95.15%	93.77%	94.09%
Our method	**96.16%**	**95.97%**	**96.09%**	**94.76%**

We present in the figure below the performance obtained with the ROC curve on HPRD dataset. The ROC area of our method is approximately 1% higher than the other approaches mentioned. All results show the reliability of the features extracted by our method that can improve the accuracy of predictions.

[3] https://github.com/xal2019/DeepFE-PPI

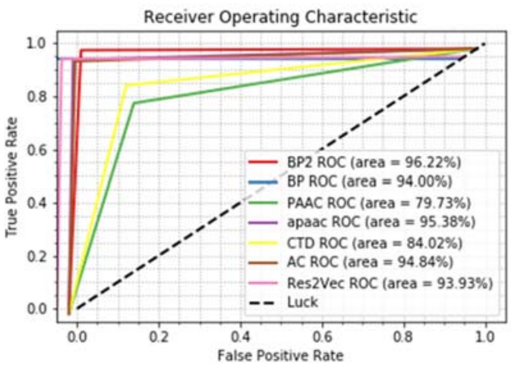

Fig. 2 The ROC curve of different methods

We also compared the performance results obtained by our method with those of the other methods cited above on the H. Sapiens data. For our method, we obtained 92.5% in accuracy, 91.70% in precision, 92% in sensitivity and 90.89% in Mcc. The best precision is obtained by the APAAC method with 93.10%.

We also compared the performance obtained on the S.Cerevisiae data set. The performances obtained by our method are 93.75%, 92.12%, 94% and 92.12% in accuracy, precision, sensitivity and Mcc, respectively. Our method performs well in almost all metrics with +0.3% more than the others, +0.4% more in sensitivity. On the other hand, the best accuracy is obtained by the Res2Vec method with 93.05% against 92.12% for our method.

The figure below gives us the performance results obtained on the H. Pylori data. Most of the methods show performances around 74% to 93% on about all metrics. These poor performances can be explained by the fact that the H. Pylori data are not large. The performances obtained by our method are 89.12%, 85.45%, 92.06% and 79.93% respectively in accuracy, precision, sensitivity and Mcc, which surpasses those obtained by some methods in the literature.

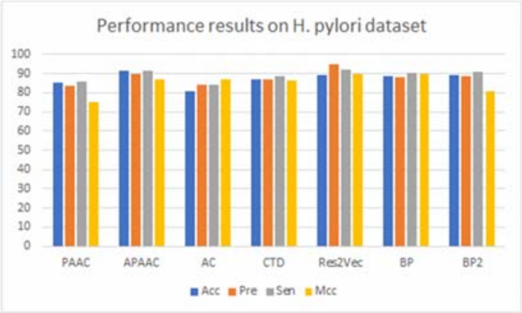

Fig. 3 The performance results comparison of different methods on S.Cerevisiae data

For E. Coli data, (figure 4) , our method is outperformed by the APAAC method with 93.79% compared to 92.98% for our method. The APAAC method extracts order-sequence information, which is a reliable characteristic for the prediction of PPIs. However, compared to the AC, CTD, BP, Res2Vec and PAAC methods, our method has a better performance.

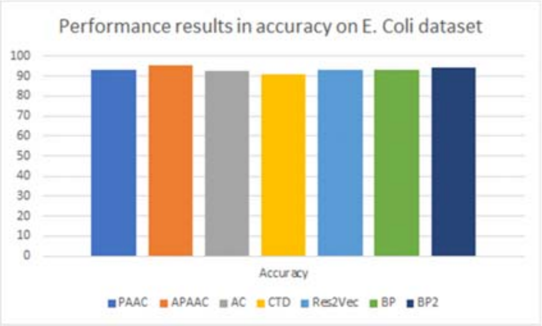

Fig. 4 The performance results comparison of different methods on E. Coli data

Here, we give the performance results obtained by our model and the model of other authors on HPRD data. We compared our results with those of Göktepe and Kodaz [14], You et al. [34], Xu et al. [35], Pan et al. [12] and You et al. [23]. The performances obtained by our model are 96.16%, 95.97%, 96.29% and 94.09% in accuracy, precision, sensitivity and Mcc, respectively.

Table 1: Results of cross-validation on the HPRD dataset

Method	***Acc***	***Pre***	***Sen***	***Mcc***
Kodaz [14]	93.45%	89.84%	89.29%	85.71%
You et al. [34]	84.8%	85.47%	84.08%	74.22%
Xu et al. [35]	90.67%	89.15%	91.69%	91.77%
Pan et al. [12]	97.95%	N/A	96.26%	95.76
ELM [23]	91.68%	91.35%	93%	89.02%
Our method	**96.16%**	**95.97%**	**96.29%**	**94.76%**

N/A mean Not Available.

We can see in table 3 that overall, our model presents good performances which are better than the performances displayed by other authors in the literature. However, the LDA-RF model has a slightly better performance than ours in accuracy and Mcc.

5. Conclusion

Protein-protein interactions play an important role in therapeutic targeting. With new diseases such as covid19 , research to identify protein-protein interactions is of great help in the search for drug solutions. As part of this work, we have proposed a feature extraction method for a better prediction of protein-protein interactions based on protein sequences. We tested our method on reference data sets including the HPRD, Homo Sapien and S. cerevisiae datasets which are widely used for the prediction of PPIs. The results obtained allow us to say that our method is an interesting tool for feature extraction in large data.

Acknowledgments

References

[1] H. Zhu *et al.*, ‘Global Analysis of Protein Activities Using Proteome Chips’, *Science*, vol. 293, no. 5537, pp. 2101–2105, Sep. 2001, doi: 10.1126/science.1062191.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, ‘A comprehensive two-hybrid analysis to explore the yeast protein interactome’, *PNAS*, vol. 98, no. 8, pp. 4569–4574, Apr. 2001, doi: 10.1073/pnas.061034498.

[3] C. D. Nguyen, K. J. Gardiner, and K. J. Cios, ‘Protein annotation from protein interaction networks and Gene Ontology’, *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 824–829, Oct. 2011, doi: 10.1016/j.jbi.2011.04.010.

[4] T. S. Keshava Prasad *et al.*, ‘Human Protein Reference Database--2009 update’, *Nucleic Acids Research*, vol. 37, no. Database, pp. D767–D772, Jan. 2009, doi: 10.1093/nar/gkn892.

[5] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, ‘DIP: the Database of Interacting Proteins’, *Nucleic Acids Res*, vol. 28, no. 1, pp. 289–291, Jan. 2000.

[6] B. Aranda *et al.*, ‘The IntAct molecular interaction database in 2010’, *Nucleic Acids Res.*, vol. 38, no. Database issue, pp. D525-531, Jan. 2010, doi: 10.1093/nar/gkp878.

[7] G. D. Bader, D. Betel, and C. W. Hogue, ‘BIND: the biomolecular interaction network database’, *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[8] K.-C. Chou, ‘Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology’, *Current Proteomics*, vol. 6, no. 4, pp. 262–274, Dec. 2009, doi: 10.2174/157016409789973707.

[9] Y. Guo, L. Yu, Z. Wen, and M. Li, ‘Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences’, *Nucleic Acids Res*, vol. 36, no. 9, pp. 3025–3030, May 2008, doi: 10.1093/nar/gkn159.

[10] Z.-H. You, X. Li, and K. C. Chan, ‘An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers’, *Neurocomputing*, vol. 228, pp. 277–282, Mar. 2017, doi: 10.1016/j.neucom.2016.10.042.

[11] L. Wong, Z.-H. You, S. Li, Y.-A. Huang, and G. Liu, ‘Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor’, in *International Conference on Intelligent Computing*, 2015, pp. 713–720.

[12] X.-Y. Pan, Y.-N. Zhang, and H.-B. Shen, ‘Large-Scale Prediction of Human Protein−Protein Interactions from Amino Acid Sequence Based on Latent Topic Features’, *J. Proteome Res.*, vol. 9, no. 10, pp. 4992–5001, Oct. 2010, doi: 10.1021/pr100618t.

[13] T. Beysolow II, *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing*. Berkeley, CA: Apress, 2018.

[14] Y. E. Göktepe and H. Kodaz, ‘Prediction of Protein-Protein Interactions Using An Effective Sequence Based Combined Method’, *Neurocomputing*, vol. 303, pp. 68–74, Aug. 2018, doi: 10.1016/j.neucom.2018.03.062.

[15] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, ‘A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition’, *Journal of Theoretical Biology*, vol. 320, pp. 41–46, Mar. 2013, doi: 10.1016/j.jtbi.2012.12.008.

[16] A. Dehzangi *et al.*, ‘PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction’, *Journal of Theoretical Biology*, vol. 425, pp. 97–102, Jul. 2017, doi: 10.1016/j.jtbi.2017.05.005.

[17] M. Gribskov, A. D. McLachlan, and D. Eisenberg, ‘Profile analysis: detection of distantly related proteins’, *PNAS*, vol. 84, no. 13, pp. 4355–4358, Jul. 1987, doi: 10.1073/pnas.84.13.4355.

[18] C. N. Kopoin, Nt. Tchimou, B. K. Saha, and M. Babri, ‘A Feature Extraction Method in Large Scale Prediction of Human Protein-Protein Interactions using Physicochemical Properties into Bi-gram’, in *2020 IEEE International Conf on Natural and Engineering Sciences for Sahel’s Sustainable Development - Impact of Big Data Application on Society and Environment (IBASE-BF)*, Feb. 2020, pp. 1–7, doi: 10.1109/IBASE-BF48578.2020.9069594.

[19] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus, ‘Hydrophobicity of amino acid residues in globular proteins’, *Science*, vol. 229, no. 4716, pp. 834–838, Aug. 1985, doi: genetic.

[20] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, ‘Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition’, *Journal of Biomolecular Structure and Dynamics*, vol. 34, no. 9, pp. 1946–1961, Sep. 2016, doi: 10.1080/07391102.2015.1095116.

[21] C. J. Shin, S. Wong, M. J. Davis, and M. A. Ragan, ‘Protein-protein interaction as a predictor of subcellular location’, *BMC Syst Biol*, vol. 3, no. 1, p. 28, Feb. 2009, doi: 10.1186/1752-0509-3-28.

[22] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, ‘Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding’, *BMC Bioinformatics*, vol. 17, no. 1, p. 184, Dec. 2016, doi: 10.1186/s12859-016-1035-4.

[23] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, ‘Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis’, *BMC Bioinformatics*, vol. 14, no. S8, p. S10, May 2013, doi: 10.1186/1471-2105-14-S8-S10.

[24] S. Martin, D. Roe, and J.-L. Faulon, ‘Predicting protein–protein interactions using signature products’, *Bioinformatics*, vol. 21, no. 2, pp. 218–226, Jan. 2005, doi: 10.1093/bioinformatics/bth483.

[25] 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res*, vol. 44, no. Database issue, pp. D7–D19, Jan. 2016, doi: 10.1093/nar/gkv1290.

[26] P.-A. Binz *et al.*, 'Proteomics standards initiative extended FASTA format', *Journal of proteome research*, vol. 18, no. 6, pp. 2686–2692, 2019.

[27] K.-C. Chou, 'Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes', *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005, doi: 10.1093/bioinformatics/bth466.

[28] Z.-H. You, J.-Z. Yu, L. Zhu, S. Li, and Z.-K. Wen, 'A MapReduce based parallel SVM for large-scale predicting protein–protein interactions', *Neurocomputing*, vol. 145, pp. 37–43, Dec. 2014.

[29] S. B. Rakhmetulayeva, K. S. Duisebekova, A. M. Mamyrbekov, D. K. Kozhamzharova, G. N. Astaubayeva, and K. Stamkulova, 'Application of Classification Algorithm Based on SVM for Determining the Effectiveness of Treatment of Tuberculosis', *Procedia Computer Science*, vol. 130, pp. 231–238, Jan. 2018, doi: 10.1016/j.procs.2018.04.034.

[30] A. J. González and L. Liao, 'Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines', *BMC Bioinformatics*, vol. 11, no. 1, p. 537, Oct. 2010, doi: 10.1186/1471-2105-11-537.

[31] A. Ben-Hur and W. S. Noble, 'Kernel methods for predicting protein–protein interactions', *Bioinformatics*, vol. 21, no. suppl_1, pp. i38–i46, Jun. 2005, doi: 10.1093/bioinformatics/bti1016.

[32] Z.-H. You, L. Zhu, C.-H. Zheng, H.-J. Yu, S.-P. Deng, and Z. Ji, 'Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set', *BMC Bioinformatics*, vol. 15, no. 15, p. S9, Dec. 2014, doi: 10.1186/1471-2105-15-S15-S9.

[33] Y. Yao, X. Du, Y. Diao, and H. Zhu, 'An integration of deep learning with feature embedding for protein–protein interaction prediction', *PeerJ*, vol. 7, p. e7126, Jun. 2019, doi: 10.7717/peerj.7126.

N'Diffon Charlemagne Kopoin is a 3rd year PhD student at the Ecole Doctorale Polytechnique of the Institut National Polytechnique Felix Houphouet Boigny in Yamoussoukro, Côte d'Ivoire. He holds a master's degree in computer science with a major in computer methodology applied to management from Nangui Abrogoua University. He is a member of the MIABD (Modeling Artificial Intelligence and Database) team of the Computer Science and Telecommunications Research Laboratory of the Institut National Polytechnique Houphouët Boigny (INP-HB), Abidjan, Côte d'Ivoire. His research interests include mathematical modeling, machine learning, and bioinformatics. His work focuses on the modeling of biological processes.

Armand Kodjo ATIAMPO holds a PhD in Computer Science, specialization in Image Processing obtained at the Institut national Polytechnique Félix Houphouet-Boigny of Cote d'Ivoire. He also holds a Master of Science in Computer Science obtained at the University Nangui Abrogoua in 2014 and a diploma in Network and Telecommunication Systems Engineering obtained in 2005. He is currently an associate researcher at the Laboratory of Research in Computer Science and Telecommunications (LARIT) of the INPHB. His interests are statistical image processing, Deep Learning applied to computer vision, detection of changes in satellite images and blockchain applications to the development of intelligent sites. He is currently working at the Virtual University of Cote d'Ivoire where he is the deputy head of the Signal, Image Processing and Multimedia research team.

Behou Gerard N'Guessan has a doctorate in computer engineering. He holds a master's degree in media engineering from the University of May 08, 1945 in Guelma, Algeria. He obtained his PhD at the University Nangui Abrogoua, Abidjan-Côte d'Ivoire in the Faculty of Applied Basic Sciences. He is a member of the Research Laboratory in Computer Science and Telecommunications of the Institut National Polytechnique Houphouët Boigny (INP-HB), Abidjan, Cote d'Ivoire, member of the Laboratory of Data Engeering and Artificial Intelligence and associate member of the Research Unit and Digital Expertise of the Virtual University of Côte d'Ivoire. His research interests include mathematical modeling, media engineering, traditional medicine, and application inventor. His work focuses on their method of research and training in traditional medicine. He is currently working as a Master Assistant at the Virtual University of Côte d'Ivoire in Abidjan (Côte d'Ivoire).

Michel Babri obtained his PhD from the Blaise Pascal University of Clermont-Ferrand, in1995. He is a full professor at the Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), Yamoussoukro Cote d'Ivoire. His research interests focus on parallel and distributed systems, cloud computing and massive data. In this capacity, he supervises several doctoral research projects. He is also the current director of the Computer Science and Telecommunications Research Laboratory (LARIT) of the INP-HB.