


Time-Frequency Approach Applied to Finding Interaction Regions in Pathogenic Proteins

Ailan F Arenas^{1,2} , Nicolás Arango-Plaza²,
Juan Camilo Arenas^{1,2} and Gladys E Salcedo²

¹Grupo de Estudio en Parasitología Molecular (Gepamol), Universidad del Quindío, Armenia, Colombia. ²Grupo de Investigación y Asesoría en Estadística, Universidad del Quindío, Armenia, Colombia.

Bioinformatics and Biology Insights
Volume 13: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1177932219850172



ABSTRACT: Protein-protein interactions govern all molecular processes for living organisms, even those involved in pathogen infection. Pathogens such as virus, bacteria, and parasites contain proteins that help the pathogen to attach, penetrate, and settle inside the target cell. Thus, it is necessary to know the regions in pathogenic proteins that interact with host cell receptors. Currently, powerful pathogen databases are available and many pathogenic proteins have been recognized, but many pathogenic proteins have not been characterized. This work developed a program in MATLAB environment based on the time-frequency analysis to recognize important sites in proteins. Our program highlights the highest energy patches in proteins from their time-frequency distribution and matches the corresponding frequency. We sought to know if this approach is able to recognize stretches residues related to interaction. Our approach was applied to five study cases from pathogenic co-crystallized structures that have been well characterized. We searched the frequencies that characterize interaction regions in pathogenic proteins and with this information tried to identify new interaction patches in either paralogs or orthologs. We found that our program generates a well-interpretable graphic under several descriptors that can show important regions in proteins even those related to interaction. We propose that this MATLAB program could be used as a tool to explore outstanding regions in uncharacterized proteins.

KEYWORDS: non-stationary series, interaction motifs, protein-protein interaction, pathogenic proteins, time-frequency analysis

RECEIVED: April 16, 2019. **ACCEPTED:** April 18, 2019.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Awarded by COLCIENCIAS Colombia through grant 1113-744-55483.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ailan F Arenas, Grupo de Estudio en Parasitología Molecular (Gepamol), Universidad del Quindío, Carrera 15 Calle 12N, Armenia, Quindío, 630001, Colombia. Email: aylanfarid@yahoo.com

Introduction

Protein-protein interactions (PPIs) govern all molecular processes of living organisms, including infections. When pathogens such as virus, bacteria, and parasites invade a host cell, they use membrane proteins to attach to some receptors in the host and these contacts allow the pathogens to penetrate the target cell.^{1,2} Once inside, the pathogen releases protein factors that interact with intra-cell proteins and kidnap the host mechanisms for their own benefits.^{3–5} The driving forces that rule the interaction between two proteins lie in certain regions hidden in the primary structure of the proteins; therefore, the discovery of these regions in pathogenic proteins is essential to implement future therapies that could block pathogen infections.⁶ Most computational tools to infer interacting regions in proteins require three-dimensional (3D) structure information from protein complexes; unfavorably, the majority of PPI complexes have no crystallography information, which is why inferences for interaction regions should be predicted from the primary structure of proteins stored in the pathogen databases.⁷ Many of the interaction regions in proteins rely on their short linear motifs (SLiMs). They are short stretches of amino acids normally located in the intrinsic disorder region.^{8–10} Most are conserved in eukaryotes even in pathogens, but most of the pathogenic proteins that participate in host cell invasion are often highly divergent from eukaryote homologs.¹¹

The informational spectrum method (ISM) is based on the primary structure of a protein, where each amino acid of the primary chain is translated into a numerical index to obtain a numerical sequence and each numerical index represents a particular physical or biochemical property for the 20 amino acids. Afterward, a Fourier transform (FT) is applied to the numerical sequence obtained; therefore, the information defined by the amino acid sequence itself can be observed in the form of informational spectrum (IS). In the IS of a protein, the maximum amplitude correlates with the highest repetition pattern in the sequence and the frequency for this amplitude carries relevant information that can represent either a functional or an interaction relation. Thus, when comparing the IS from two proteins and both have at least one common frequency with a higher amplitude, it means that both proteins share some information that could be either functional or structural.^{2,12–14} Most of the literature where IS was applied merely compare the frequencies obtained in the IS, where proteins belonging to the same family or performing the same function share at least one frequency peak with the highest amplitude. This approach has been used to classify and predict the function/structure of unknown proteins or peptides.^{2,15–17} The next step maps the region in a protein that is responsible for that particular frequency. Hence, we evaluated if the time-frequency analysis (TFA) approach is capable of recognizing interaction regions in a protein with a particular frequency/amplitude.



This work expands the ISM approach by including the TFA. The TFA was applied in Hassani Saadi et al¹⁸ to find local structure periodicities in DNA, but we expanded the search of this approach to look for interacting regions in pathogenic proteins from intraspecific recognized pathogen PPI. Finally, we developed a program in MATLAB that generates a well-interpretable graphic that could show interaction regions either in paralog or ortholog pathogenic proteins. Our program was assessed in five case studies.

Materials and Methods

Dataset

We performed an exploratory analysis from six Protein Data Bank (PDB) structures extracted from the National Center for Biotechnology Information (NCBI) structure summary (www.ncbi.nlm.nih.gov/Structure) to locate the exposed regions from each protein. The PDB used are the following: 3ZLD, 5NQG, 5NQF, 3ZWZ, 4Z80, 4LV5. Although the PDB structures show all the interaction regions exposed to their substrate, we also supported all the information about the PPI with the literature reported for each PDB.

Time-frequency analysis

Since the distance between amino acid residues in a protein sequence is about 3.8 Å, most of corresponding numeric representations can be analyzed as an equidistance realization (or *time series*) from some stochastic process that can be stationary or not. Stationary can be strong or weak; strong stationarity establishes the same probability distribution of $\{x_{t_1}, x_{t_2}, \dots, x_{t_n}\}$ and $\{x_{t_1+k}, x_{t_2+k}, \dots, x_{t_n+k}\}$, $\forall n, k \in \mathbb{N}$, ie, the n -dimensional distributions are time invariants. Process is weakly stationary when the unconditional expectations and variances are also time invariants and the correlation structures between x_t and x_s depend solely on the delay $k = |s - t|$.

The second-order structure most used in practice to analyze a stationary time series $\{x_t, t \in \mathbb{Z}\}$ is the autocorrelation function:

$$\rho_x(k) = \frac{\gamma_x(k)}{\sigma_x^2} = \frac{E[(x_{t+k} - \mu_x)(x_t - \mu_x)]}{\sigma_x^2}, \quad k \in \mathbb{Z}$$

where $\gamma_x(k)$ is the autocovariance function of $\{x_t, t \in \mathbb{Z}\}$; μ_x is the unconditional expectation, and σ_x^2 is the unconditional variance of the process. In the *frequency-domain*, if $\sum_{k=-\infty}^{\infty} |\gamma_x(k)| < \infty$, the covariance structure is represented by the spectral density function defined by

$$f_x(\omega) = \sum_{k=-\infty}^{\infty} \gamma_x(k) e^{-i2\pi\omega k}, \quad \omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

The spectrum $f_x(\omega)$ of a stationary process describes the power-frequency distribution for the whole process. Analogously, for two time series $\{x_t, t \in \mathbb{Z}\}$ and $\{y_t, t \in \mathbb{Z}\}$,

the correlation structure in the frequency-domain can be analyzed from the coherence function. The expectation

$$\gamma_{xy}(k) = E[(x_{t+k} - \mu_x)(y_t - \mu_y)], \quad k \in \mathbb{Z}$$

represents the cross-covariance between $\{x_t\}$ and $\{y_t\}$. In the frequency-domain, if $\sum_{k=-\infty}^{\infty} |\gamma_{xy}(k)| < \infty$, this cross-covariance is represented by the cross-spectrum:

$$f_{xy}(\omega) = \sum_{k=-\infty}^{\infty} \gamma_{xy}(k) e^{-i2\pi\omega k}, \quad \omega \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

Because of the relationship $\gamma_{yx}(k) = \gamma_{xy}(-k)$, it follows that $f_{yx}(\omega) = \overline{f_{xy}(\omega)}$, and the squared coherence function is defined by

$$\rho_{xy}^2(\omega) = \frac{|f_{xy}(\omega)|^2}{f_x(\omega)f_y(\omega)}$$

and satisfies $0 \leq \rho_{xy}^2(\omega) \leq 1$.

On the other hand, since several studies have reported the non-stationarity feature of genomic and biomolecular sequences,¹⁸ time-dependent spectra are an useful tool to identify localized characteristics of a protein. For instance, “hot spot” aminoacids or motifs that most contribute to a specific frequency that describes either a biological function or an interaction.¹

For a real-valued signal $x(t)$, $t \in \mathbb{R}$, Ville¹⁹ in 1948 introduced the analytical signal concept and a quadratic transform previously studied by Wigner²⁰ on quantum thermodynamic and rediscovered by Cohen in 1966²¹ for applications in statistical mechanics and signal processing of light waves. The Wigner-Ville (WV) transform of $x(t)$, $t \in \mathbb{R}$ is given by

$$\begin{aligned} W_x(t, \omega) &= \frac{1}{2\pi} \int x^* \left(u - \frac{\tau}{2}\right) x \left(u + \frac{\tau}{2}\right) e^{-i\tau\omega} d\tau \\ &= \frac{1}{2\pi} \int R_x(u, \tau) e^{-i\tau\omega} d\tau \end{aligned} \quad (1)$$

where $x^*(t)$ is the analytic signal associated with $x(t)$. Function $R_x(u, \tau) = x^*(u - \tau/2)x(u + \tau/2)$ represents a form of local autocovariance and measures the covariance between values at time points separated by an interval τ and symmetrically placed about the time $t = u$. In this transform, time and frequency have a symmetric role, so by applying the Parseval formula,²² this time-frequency distribution (TFD) can also be rewritten as a frequency integration, ie,

$$W_x(t, \omega) = \frac{1}{2\pi} \int f_x^* \left(\omega - \frac{\lambda}{2}\right) f_x \left(\omega + \frac{\lambda}{2}\right) e^{i\lambda t} d\lambda$$

Even though WV transform looks like a powerful tool to analyze the time-frequency features of a signal, this is not the case due to the interferences created by the cross terms

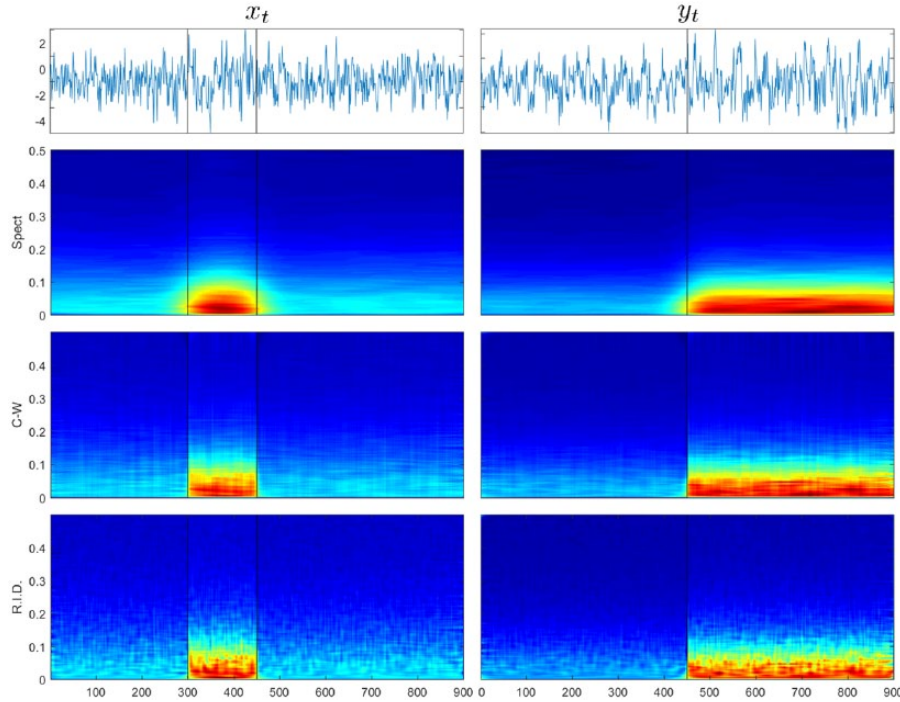


Figure 1. One simulated signal x_t (on top) and average of the estimated TFDs for the 1000 replications (left column) and one simulated signal y_t (on top) and average of the estimated TFDs for the 1000 replications (right column). C-W indicates Choi-Williams; R.I.D., reduced interference distribution; Spect, Spectrogram.

in equation (1). These interferences can be attenuated by smoothing the WV transform as proposed by Cohen²¹; however, the consequence of this is a decrease of the time and frequency resolutions, and more generally a loss of theoretical properties. The general family of Cohen's quadratic TFDs is

$$C_x(t, \omega) = \frac{1}{4\pi^2} \iiint \phi(\theta, \tau) R_x(u, \tau) e^{-i\theta t - i\tau\omega + i\theta u} du d\tau d\theta$$

where $\phi(\theta, \tau)$ is a function independent of time and frequency that acts as a smoothing kernel. By choosing different kernels, we obtain different distributions as well, and the mathematical properties of $C_x(t, \omega)$ depend on kernel chosen. If $\phi(\theta, \tau) = 1$, we obtain the WV distribution which satisfies many desirable properties as energy conservation, time and frequency marginals, convolution, real-valued, time and frequency shifts, group delay, among others. In order to balance both properties and resolutions, in this work, we used the *Spectrogram*, the *Choi-Williams* distribution with kernel $\phi(\theta, \tau) = e^{-\theta^2 \tau^2 / \sigma}$, and the *reduced interference* distribution with a kernel based on the Hanning window.^{23,24}

A simulation example. To illustrate how local covariations can be detected from TFA, we generate 1000 bivariate time series $(x_t, y_t)', t = 1, 2, \dots, 900$, with different structures of local dependence, from bivariate vector autoregression (VAR)¹ process of order 1:

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = A \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}_{I_{\{1 \leq t \leq 300\}}} + B \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}_{I_{\{301 \leq t \leq 450\}}} + C \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}_{I_{\{451 \leq t \leq 700\}}} + \begin{pmatrix} a_{1,t} \\ a_{2,t} \end{pmatrix} \quad (2)$$

$$A = \begin{pmatrix} 0.4 & 0 \\ 0 & 0.5 \end{pmatrix}, B = \begin{pmatrix} 0.4 & 0.6 \\ 0 & 0.5 \end{pmatrix}, C = \begin{pmatrix} 0.4 & 0 \\ 0.8 & 0.5 \end{pmatrix}$$

and errors $(a_{1,t}, a_{2,t})' : N(\mathbf{0}, I_2)$, where I_2 is the identity matrix of order 2. Note that for $1 \leq t < 300$, two time series are non-correlated; x_t depends on y_{t-1} from $t = 301$ to 450 and y_t depends on x_{t-1} for $t \geq 451$. For each simulated time series, we estimate the above-mentioned distributions. On top of Figure 1, we plot two random signals $x_t, y_t, t = 1, \dots, 900$, and the average of the 1000 estimations of each TFD. Note that from time series plots, it is not possible to identify the local covariations; however, the TFDs exhibit a strong energy at low frequencies, $\omega < 0.05$, approximately. Both situations are expected the characteristic of the VAR generator process²³ and the local variations induced by the type of linear dependence between x_t and y_t . Finally, Figure 2A shows the squared coherence function between the two random signals x_t and y_t from Figure 1. In this case, we may reject the hypothesis of no coherence for values of $C_{0.001} > 0.27$.²⁵ Figure 2B exhibits the average of the 1000

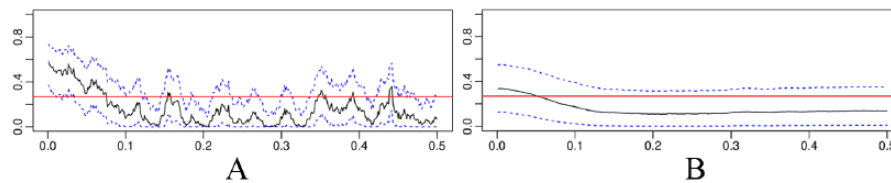


Figure 2. Coherence between the two random signals x_t and y_t (A) and the average of the 1000 estimated coherence functions (B).

estimated coherence functions with confidence bands²⁵ where it is clear that there exists some form of interaction between the simulated time series at frequencies lower than 0.05.

ISM procedure

The MATLAB program was loaded with 631 molecular descriptors (Supplementary Material 1). The proteins were translated for each molecular descriptor obtaining 631 numerical series for each protein. Thereafter, we applied a discrete Fourier transform (DFT) for each numerical series and then TFA was applied for each FT. Finally, we obtained 631 matrices for each protein evaluated. Here, we introduced a threshold value to highlight the highest energy density sites for each matrix and search if the densities obtained for each matrix lie on a position that matches the interaction region and a particular frequency. The frequency value matching the highest density energy in an interaction region is then used to search interaction regions in either ortholog or paralog proteins. Our program was assessed in five case studies. Supplementary information 2 includes the user manual for the MATLAB program. The MATLAB program is stored in Additional file 1.

Results

Case studies

Application of the MATLAB TFA program to the TgRON2 protein looking for an interaction region. First, we downloaded the complete protein sequence for *Toxoplasma* RON2 TgRON2 in FASTA format, then we searched for the interaction regions in the sequence from the PDB 3ZLD, which describes the interaction between the TgAMA1 protein and a peptide derived from TgRON2.²⁶ The authors considered that a peptide in TgRON2 that covers the amino acids from 1003 to 1028 1003-FLTDSGMKAIEDCSWNPIMQQMACVV-1028 interacts with TgAMA1.²⁶ Therefore, we applied the TFA program to the TgRON2 sequence to observe if the energy density lies in the location 1003-1028. We found that the MATLAB TFA program highlighted the TgRON2 1003-1028 region in eight descriptors with 0.80 threshold. The energy patch obtained matches with a particular interval frequency (0.346 ± 0.001), where the interaction peptide in TgRON2 is located (Table 1 and Figure 3).

The frequency in the PVRON2 interaction peptide matches the ortholog PFRON2 interaction peptide. Similar to the case above in TgRON2, the interaction region in PVRON2 was obtained

from PDB 5NQG. This is a peptide 31 residues long from the PVRON2 C-terminus region 2039-HATDIGMGPATSCYT-STIPPPKQVCIQQAVK-2069.²⁷ We took the complete protein sequence of PVRON2 and applied the MATLAB TFA program to extract the frequency that highlights the interaction peptide 2039-2069. We found two descriptors with 0.6 threshold that highlighted the 2039-2069 region with a frequency around 0.33 ± 0.005 in PVRON2 (Figure 4, Table 1, and Additional file 2). Then, we explored if the frequency (0.33) also matches the ortholog PFRON2 interaction peptide 2028-DIGAGPVAS-CFTTRMSPPQQICLNSVNN-2055 (see PDB 3ZWZ).²⁸ Applying our bio-informatics approach to the complete PFRON2 sequence, we found that effectively the frequency (0.33) also highlights the interaction peptide 2028-2025 with the same threshold (Figure 4). PVRON2 and the ortholog PFRON2 only shared two descriptors that showed the frequency (0.33) in their respective interaction peptides (Table 1, and Additional file 2). However, the frequency (0.33) in PFRON2 was also found in five other descriptors (Additional file 2).

The frequency in the PVAMA1 interaction region matches the ortholog PFAMA1 interaction region. By applying the same protocol above to find interaction regions, we obtained the interaction region in *Plasmodium vivax* AMA1 PVAMA1 from PDB 5NQG; the authors who described this structure recognized an interaction region in PVAMA1 that covers the 168-SFVMA-172 amino acids.²⁷ Then, we applied our MATLAB program to PVAMA1 complete sequence to find a frequency that highlights the 168-172 region. We found seven descriptors that indicated an interval frequency (0.39 ± 0.01) that highlighted the 168-172 region with 0.6 threshold (Table 1 and additional file 2). Thereafter, we used this information to find interaction regions in the ortholog *Plasmodium falciparum* AMA1 PFAMA1. We found a relevant interaction region recognized in PFAMA1 from the amino acid 222 to 227 in nine descriptors (Additional file 2). This 222-GNMNPD-227 patch in PFAMA1 was previously recognized as an interface interaction in the PDB 5NQF.²⁷ PVAMA1 and the ortholog PFAMA1 only shared one descriptor that showed the frequency (0.39) in their respective interaction patches (Table 1) (see all the graphics for PFAMA1 and PVAMA1 in Additional file 2).

The frequency TgAMA1 interaction region matches the paralog TgAMA4 interaction region. We recognized the interaction region in *Toxoplasma gondii* AMA1 TgAMA1 from PDB 3ZLD that covered two patches, 183-QVYTS-187 and 222-TIAV-225,²⁶ and our MATLAB program showed a frequency around

Table 1. Description of all the results obtained in the five case studies.

PROTEIN	PDB	TRSHL	DESCPTS	S.F.	RFC	ORGANISM	INTERACTION DOMAIN/MOTIF
TgRON2	3ZLD	0.85	2, 29, 30, 32, 153, 156, 380, 450	0.346 ± 0.001	Poukchanski et al ²⁶	<i>Toxoplasma gondii</i>	FLTDSGMKAIEDCSWNPIMQMACVV 1003-1028
PVAMA1	5NQG	0.6	19, 68, 370, 394, 443, 576	0.39 ± 0.01	Vuiliez-Le Normand et al ²⁷	<i>Plasmodium vivax</i> (orthologs)	168-SFVMA-172
PFAMA1	5NQ		250, 356, 368, 369, 370, 381, 383, 445, 460			<i>Plasmodium falciparum</i>	225-GNMNPD-227
PVRON2	5NQG	0.6	68, 672	0.33 ± 0.005	Vuiliez-Le Normand et al ²⁷	<i>P vivax</i> (orthologs)	HATDIGMGFATSCYTSTIPPPKQVCIGQAVK 2039-2069
PFRON2	3ZWZ		68, 70, 96, 672, 617, 626, 629		Vuiliez-Le Normand et al ²⁸	<i>P falciparum</i>	DIGAGPVASCFTTRMSPPQICLNSVVN 2028-2055
TgAMA1	3ZLD	0.6	1, 3, 46, 68, 83, 92, 152, 672, 618	0.217 ± 0.001	Poukchanski et al ²⁶	<i>T gondii</i> (paralogs)	183-QVYTS-187 222-TIAV-225
TgAMA4	4Z80		250, 373, 374, 672		Parker et al ¹¹		209-YTLHCPYVNVYRQD-223
IRGa6 cim-IRGb2/b1 lab-IRGb2/b1 (<i>Helix</i>) 4	4LV5	0.7	319, 365, 368, 374, 380, 388, 391, 438, 391, 391	0.454 ± 0.001	Reese et al ⁵	<i>Mus musculus</i> (orthologs)	208-DIRLNCVNTFREN-220 194-NRENILKSLFRNCISSNLKEC-213 194-NRENILKSIRICLSSNLKER-213

Abbreviations: Descpts, descriptors; PDB, Protein Data Bank; Rfc, references; S.F., shared frequencies; Trshl, threshold. Includes all the regions chosen for the studies, descriptors that showed similar frequencies for each study case, and related information. Descriptors shared in the paralog/ortholog proteins appear in italics.

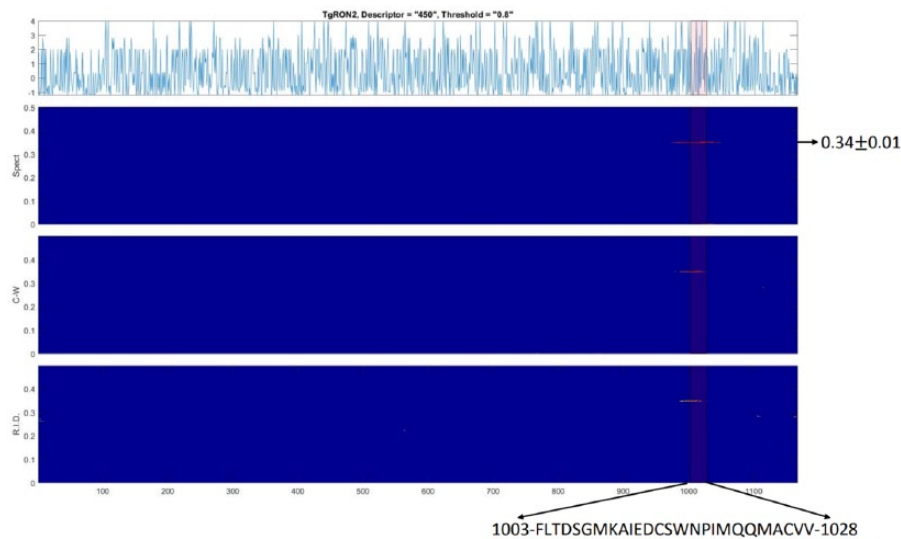


Figure 3. MATLAB TFA application in TgRON2. The graphic shows the higher energy patch for the three distributions in an interval frequency (0.346 ± 0.001) that covers the 1003-FLTDSGMKAIEDCSWNPIMQQMACVV-1028 region where the TgRON2 interaction peptide is located. This finding was obtained under the descriptor (450) (Supplementary Material 1). The graphics obtained for the eight descriptors in TgRON2 are in Additional file 2. C-W indicates Choi-Williams; R.I.D., reduced interference distribution; Spect, Spectrogram.

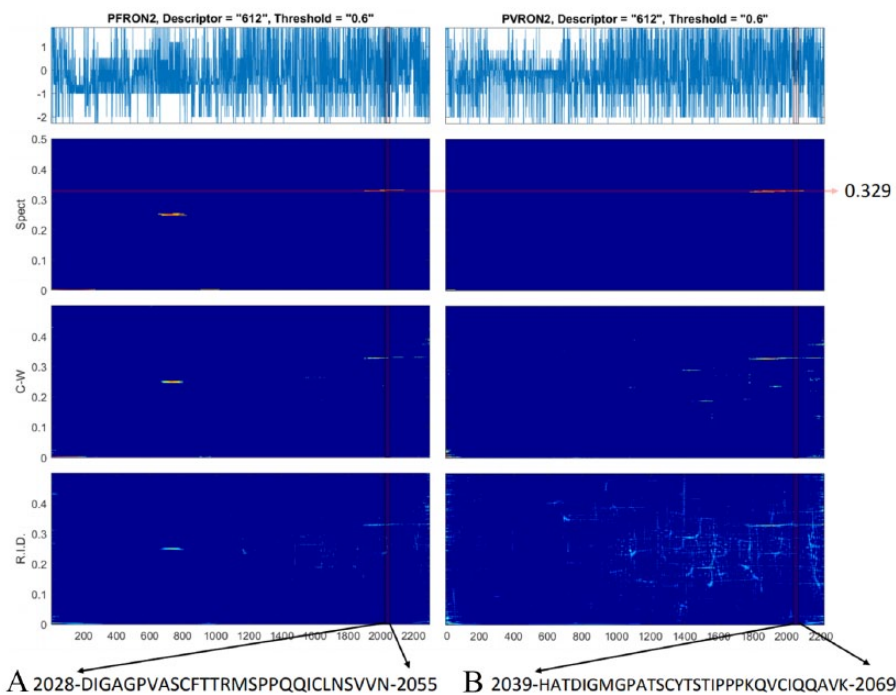


Figure 4. MATLAB TFA application in PVRON2 and the ortholog PFRON2. The graphic shows both energy patches in an interval frequency (0.33 ± 0.005) that covers the (B) 2039-HATDIGMGPATSCYTSTIPPPQVCIQQAVK-2069 and (A) 2028-DIGAGPVASCFTTRMSPPQQICLNSVVN-2055 regions for PVRON2 and PFRON2, respectively. Both peptides were considered interaction interfaces in Vulliez-Le Normand et al.^{27,28} These findings were obtained under descriptors (68 and 612) (Supplementary Material 1). The seven graphics for PFRON2 and the two graphics for PVRON2 are in Additional file 2. C-W indicates Choi-Williams; R.I.D., reduced interference distribution; Spect, Spectrogram.

(0.217 ± 0.001) that highlights both 183–187 and 222–225 patches in nine descriptors with 0.6 threshold (Figure 5A and Additional file 2). We used this information to look for the interaction region in TgAMA4. We found the 209-YTLHCPYVNVYRQD-223 interaction patch in four descriptors in TgAMA4 (Figure 5B and additional file 2). The authors who described TgAMA4 published

that amino acids 209, 211, 215, and 223 are part of the interaction interface of TgAMA4, and these amino acids interact by hydrogen bonds with their respective protein substrate.¹¹ In this case, TgAMA1 and the paralog TgAMA4 shared one descriptor that showed the frequency (0.217) in their respective interaction regions (Table 1) (additional file 2).

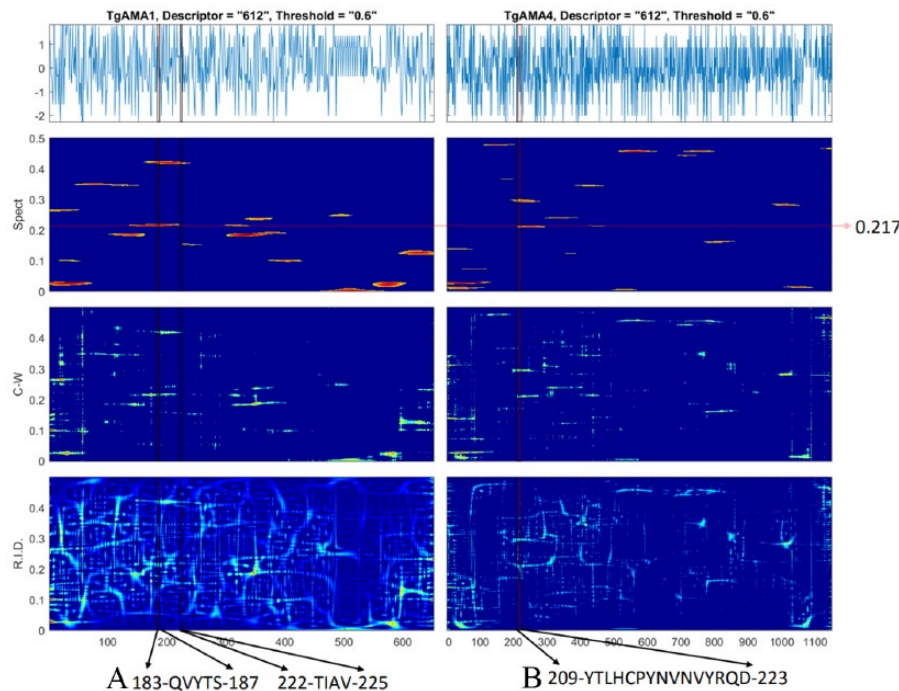


Figure 5. MATLAB TFA application in TgAMA1 and the paralog TgAMA4. The graphic shows an energy density in a frequency (0.217 ± 0.001) that covers patches 183-QVYTS-187 and 222-TIAV-225 in TgAMA1 (A), and patch 209-YTLHCPYNNVYRQD-223 in TgAMA4 (B). The three patches were considered interaction regions in the previous works.^{11,26} These findings were obtained under descriptor (612) (Supplementary Material 1). The nine graphics for TgAMA1 and the four graphics for TgAMA4 are in Additional file 2. C-W indicates Choi-Williams; R.I.D., reduced interference distribution; Spect, Spectrogram.

The frequency in IRGa6 helix 4 matches the paralog *cimIRGb2-b1* helix 4 interaction region, but not in the paralog *labIRGb2-b1* helix 4. Immunity-related guanosine triphosphatases (IRGs) are interferon-inducible proteins that mediate cell autonomous resistance against intracellular pathogens.^{3-5,29-32} These IRGs are well characterized in mice and can accumulate onto vacuolar membrane-coated parasites. *Toxoplasma gondii* is a well-adapted parasite able to avoid immune responses in susceptible mice; the process is driven by toxoplasma kinase and pseudokinase proteins called ROP and now it is known that the interaction mechanisms between ROPs and IRGs produce the balance between virulence and resistance in mice.^{3-5,29-32} For instance, in susceptible mice (*lab* mice), the parasite secretes the pseudokinase ROP5 and this protein directly interacts with IRGa6 blocking its clearing activity onto the parasite vacuole allowing the parasite replication.^{3-5,29-32} In the case of the CIM strain, which is a toxoplasma natural resistance strain of mice, an allele of the IRGb2-b1 protein, which also interacts with ROP5, allows the IRGa6 to break the parasite vacuole, preventing parasite replication.^{4,29-32} For susceptible mice, they have also an allele copy for IRGb2-b1, but mutations in this copy cause no interaction with ROP5 allowing mice death.^{4,29-32} So then, we wanted to analyze this resistance/virulence mechanism with our MATLAB program, so we extracted the IRGa6 interfaces from the PDB 4LV5, which describes the interaction between IRGa6 and ROP5B³⁻⁵; the authors who analyzed this interaction suggest that helix 3 and helix 4 in IRGa6 mediate the interaction with ROP5B. According to our bio-informatics approach, we

found that 11 descriptors showed a frequency (0.454 ± 0.001) that highlights the amino acids that lie in helix 4 in IRGa6 208-DIRLNCVNTFREN-220 with 0.7 threshold (Table 1 and Figure 6A). In Lilue et al,⁴ the authors suggest that helix 4 in IRGb2-b1 CIM mouse is the interface region that also interacts with ROP5; then, we searched if the frequency (0.454) obtained in the IRGa6 helix 4 also highlights helix 4 in either *cimIRGb2-b1* (resistance mice) or *labIRGb2-b1* (susceptible mice). Interestingly, similar to what other authors reported, a high-energy patch was found that lies in helix 4 in *cimIRGb2-b1* mouse sequence, but it was not found in helix 4 in *labIRGb2-b1* mouse sequence (Figure 6B and C). IRGa6 and the paralog *cimIRGb2-b1* shared the frequency (0.454) in their respective helix 4 interaction regions in one descriptor (Table 1 and additional file 2).

Discussion

Looking for interaction regions in proteins is a hard task even by means of bio-informatics approaches; this is because interaction information lies in small regions in the protein even on a few single amino acids. For that reason, it is necessary to design computational approaches that help researchers to get hints about the sites in the proteins that promote interactions. Our MATLAB program was designed seeking to obtain important regions in proteins and projecting a clearer visualization of the information. We suggest that our MATLAB program could be a complementary approach to include in protein analysis.

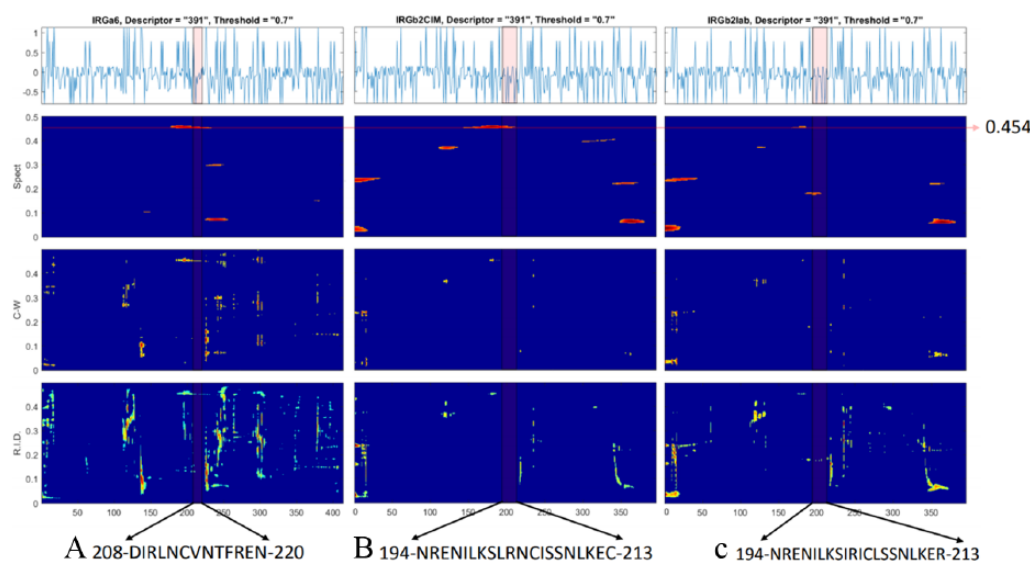


Figure 6. MATLAB TFA application in IRGa6 and their paralogs cimIRGb2-b1 and labIRGb2-b1. The graphic shows two energy patches in a frequency (0.454 ± 0.001) that covers both helix 4 208-DIRLNCVNTFREN-220 in IRGa6 (A) and helix 4 194-NRENILKSLRNCISSNLKEC-213 in cimIRGb2-b1 (B), but there was no energy patch found in helix 4 194-NRENILKSIRICLSSNLKER-213 in labIRGb2-b1 (C). These findings were obtained under descriptor (391) (Supplementary Material 1). The 11 graphics for IRGa6 and the graphics for cimIRGb2-b1 and labIRGb2-b1 are in additional file 2. C-W indicates Choi-Williams; R.I.D., reduced interference distribution; Spect, Spectrogram.

We observed that the three TFA distributions showed energy in the interaction region suggested in all the examples, but the spectral distribution was clearer distinguishing the energy patches in proteins. Although only a few descriptors had been able to locate energy regions related to interactions, we suggest that if a number of descriptors highlights the same region in a protein in the same range of frequency, this region must be relevant in that protein. We also realized that as we increase the threshold, the more prominent energy patches rise up in the figure. For instance, we observed in the first case study that the TFA approach was capable of highlighting the specific interaction peptide in TgRON2 with a very specific frequency (0.346 ± 0.001) in eight descriptors with 0.80 threshold (Figure 3 and additional file 2). Thus, we could locate important regions in proteins where we would have no knowledge otherwise. Similarly, in our second study, the PFRON2 protein showed a specific frequency (0.33), also shared with the ortholog PVRON2 (Figure 4).

We also observed in the figures 5 and 6 more prominent energy patches in protein regions probably not related to interaction (Figures 5 and 6). That means large regions that conserve more local periodicities, like secondary structures, may stand out more than interaction patches because interaction regions lie in only a few amino acids with not enough periodic information.

We suggest that the frequencies and the energy patches shared either in the ortholog or in the paralog interaction regions are not because of sequence similarities given that no sequence conservation exists between regions analyzed (Table 1). Even in case study 5, where we compared helix 4 from cimIRGb2-b1 and labIRGb2-b1, both helices are highly conserved at sequence level (Table 1); however, the MATLAB program showed that

the energy density in both helices is quite different (Figure 6B and C and additional file 2). It may mean that the TFA program is highly sensitive even to smaller differences in the primary sequence of the proteins and it can reflect the importance of these sites in the overall energy density in a particular region in a protein.

We consider that our program works best when the region to find is larger than 20 amino acids and it is inconclusive if the region we are looking for is a smaller stretch of amino acids. Most computational programs to identify interaction regions in proteins come from either profile-based methods or conservation of clue amino acids in the intrinsic disorder regions.⁸⁻¹⁰ Because most pathogenic proteins follow a specialized co-evolutionary process regarding their host, we would not expect profile-based methods to find conserved interaction motifs for these kinds of proteins. For that reason, we used co-crystallized PDB structures to locate experimentally recognized interaction interfaces in pathogenic proteins to evaluate our program. We consider that these few examples analyzed were promising and suggest that the MATLAB program works suitably, given that it was able to find large interaction patches proposed in the PDB analyzed.

Conclusions

The MATLAB TFA program generates a well-interpretable graphic that can show important regions in proteins, even those related to protein interactions. We propose that this MATLAB program can be a starting point analysis tool to locate important regions in proteins, especially those lacking 3D structure information or without characterization. Our program can also be applied to a different context not only to pathogenic paralog/ortholog proteins.

Author Contributions

NAP, GES, JCA, and AFA designed the approach and implemented the algorithm; all authors contributed to the discussion and revision of the manuscript, and all authors read and approved of the final manuscript.

Availability of Data and Material

Additional file 1: contains the MATLAB program.

Additional file 2: contains all the graphics obtained with the MATLAB program in the 5 case studies.

Supplementary material 1: contains the numerical descriptors used in this work.

Supplementary material 2: contains the user manual for the MATLAB program.

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Ailan F Arenas  <https://orcid.org/0000-0002-0331-5556>

REFERENCES

- Arenas AF, Salcedo GE, Montoya AM, Gomez-Marín JE. MSCA: a spectral comparison algorithm between time series to identify protein-protein interactions. *BMC Bioinformatics*. 2015;16:152.
- Nourani E, Khunjush F, Durmus S. Computational approaches for prediction of pathogen-host protein-protein interactions. *Front Microbiol*. 2015;6:94–10.
- Fleckenstein MC, Reese ML, Konen-Waisman S, Boothroyd JC, Howard JC, Steinfeldt T. A toxoplasma gondii pseudokinase inhibits host IRG resistance proteins. *PLoS Biol*. 2012;10:e1001358.
- Lilue J, Muller UB, Steinfeldt T, Howard JC. Reciprocal virulence and resistance polymorphism in the relationship between toxoplasma gondii and the house mouse. *Elife*. 2013;2:e01298.
- Reese ML, Shah N, Boothroyd JC. The toxoplasma pseudokinase ROP5 is an allosteric inhibitor of the immunity-related gtpases. *J Biol Chem*. 2014;289:27849–27858.
- Escotte-Binet S, Huguénin A, Aubert D, et al. Metallopeptidases of toxoplasma gondii: in silico identification and gene expression. *Parasite*. 2018;25:26.
- Aurrecochea C, Barreto A, Basenko EY, et al. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Research*. 2016;45:D581–D591.
- Ren S, Uversky VN, Chen Z, Dunker AK, Obradovic Z. Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics*. 2008;9:S26.
- Gould CM, Diella F, Via A, et al. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res*. 2009;38:D167–180.
- Davey NE, Haslam NJ, Shields DC, Edwards RJ. SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res*. 2011;39:W56–W60.
- Parker ML, Penarete-Vargas DM, Hamilton PT, et al. Dissecting the interface between apicomplexan parasite and host cell: insights from a divergent AMA-RON2 pair. *Proc Natl Acad Sci U S A*. 2016;113:398–403.
- Chrysostomou C, Seker H, Aydin N. CISAPS: complex informational spectrum for the analysis of protein sequences. *Adv Bioinformatics*. 2015;2015:909765.
- Koma T, Veljkovic V, Anderson DE, et al. Zika virus infection elicits auto-antibodies to cIq. *Sci Rep*. 2018;8:1882.
- Schmier S, Mostafa A, Haarmann T, et al. In silico prediction and experimental confirmation of HA residues conferring enhanced human receptor specificity of H5N1 influenza A viruses. *Sci Rep*. 2015;5:11434.
- Glisic S, Cavanaugh DP, Chittur KK, Sencanski M, Perovic V, Bojic T. Common molecular mechanism of the hepatic lesion and the cardiac parasympathetic regulation in chronic hepatitis c infection: a critical role for the muscarinic receptor type 3. *BMC Bioinformatics*. 2016;17:139.
- McCullough C, Wang M, Rong L, Caffrey M. Characterization of influenza hemagglutinin interactions with receptor by NMR. *PLoS ONE*. 2012;7:e33958.
- Veljkovic V, Veljkovic N, Muller CP, et al. Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. *BMC Struct Biol*. 2009;9:21.
- Hassani Saadi H, Sameni R, Zollanvari A. Interpretive time-frequency analysis of genomic sequences. *BMC Bioinformatics*. 2017;18:154.
- Ville J. Theorie et application de la notion de signal analytique. *Cables et Trans*. 1948;2:61–74.
- Wigner EP. On the quantum correction for thermodynamic equilibrium. In: Wightman, AS, ed. *Part I: Physical Chemistry. Part II: Solid State Physics*. Berlin, Germany: Springer; 1932:110–120.
- Cohen L. Generalized phase-space distribution functions. *J Math Phys*. 1966;7:781–786.
- Mallat S. *A Wavelet Tour of Signal Processing: The Sparse Way*. San Diego, CA: Academic Press; 2008.
- Auger F, Flandrin P, Gonçalvès P, Lemoine O. *Time-Frequency Toolbox*. Vol. 46. Paris; Houston, TX: CNRS; Rice University; 1996.
- Flandrin P. *Time-Frequency/Time-Scale Analysis*. Vol. 10. San Diego, CA: Academic Press; 1998.
- Shumway RH, Stoffer DS. *Time Series Analysis and Its Applications: With R Examples*. Berlin, Germany: Springer; 2017.
- Poukchanski A, Fritz HM, Tonkin ML, Trecek M, Boulanger MJ, Boothroyd JC. Toxoplasma gondii sporozoites invade host cells using two novel paralogues of RON2 and AMA1. *PLoS ONE*. 2013;8:e70637.
- Vulliez-Le Normand B, Saul FA, Hoos S, Faber BW, Bentley GA. Cross-reactivity between apical membrane antigen 1 and rhoptry neck protein 2 in P. vivax and P. falciparum: a structural and binding study. *PLoS ONE*. 2017;12:e0183198.
- Vulliez-Le Normand B, Tonkin ML, Lamarque MH, et al. Structural and functional insights into the malaria parasite moving junction complex. *PLoS Pathog*. 2012;8:e1002755.
- Bekpen C, Hunn JP, Rohde C, et al. The interferon-inducible p47 (IRG) gtpases in vertebrates: loss of the cell autonomous resistance mechanism in the human lineage. *Genome Biol*. 2005;6:R92.
- Hermanns T, Muller UB, Konen-Waisman S, Howard JC, Steinfeldt T. The toxoplasma gondii rhoptry protein ROP18 is an IRGA6-specific kinase and regulated by the dense granule protein GRA7. *Cell Microbiol*. 2016;18:244–259.
- Howard JC, Hunn JP, Steinfeldt T. The irg protein-based resistance mechanism in mice and its relation to virulence in toxoplasma gondii. *Curr Opin Microbiol*. 2011;14:414–421.
- Muller UB, Howard JC. The impact of toxoplasma gondii on the mammalian genome. *Curr Opin Microbiol*. 2016;32:19–25.