

DIPN

A dictionary of the internal proteins nanoenvironments and their potential for transformation into agricultural assets

Ivan Mazoni | Goran Neshich

Introduction

Proteins play a vital role in supporting life. They are macromolecules resulting from the combination, through peptide bonds, of these 20 amino acids: alanine, arginine, aspartate, asparagine, cysteine, phenylalanine, glycine, glutamate, glutamine, histidine, isoleucine, leucine, lysine, methionine, proline, serine, tyrosine, threonine, tryptophan, and valine. Considering a linear combination between these 20 amino acids, the number of possible variations is 20^n , in which n is the amount of amino acid residues in the protein (as the amino acids lose some atoms when forming the peptide bond, it is common to call them amino acid residues, since they are part of a polypeptide chain). For example, for a protein with 100 amino acid residues, the number of possible combinations will equal $20^{100} = 1.27 \times 10^{130}$. In comparison, the estimated total number of atoms in the Universe is 9×10^{78} (Villanueva, 2009). Each organism, animal or vegetable, has thousands of different proteins. Among their various functions: structural, transport, protection, defense, control and regulation of expression, catalysis, movement, and storage stand out as some examples. For a better understanding of the relationship between the amino acid sequence in a protein, its three dimensional structure, and its function, came the proposition for analyses of the proteic nanoenvironment. It is also known as a proteic district or functional region, and it is where biologically functional elements are located.

The hypothesis that motivated the work of the Embrapa Digital Agriculture Computational Biology Research Group (CBRG) in Campinas (SP), during the 2010s, was an approach that assumed the existence of a “sign”, or that is, a variation in the values of the physico-chemical and structural descriptors that distinguish a specific site (or a protein substructure). This is where a certain element of secondary structure (or an active site, an interface, etc.) is inserted in the framework of a whole protein. Understanding how the subordinate structural elements to the biologically functional structure are formed and later maintained will open the way for us to understand how proteins assume their final structure and, consequently, their function. In our work we use STING_RDB, a unique database in the world, produced and maintained by the Embrapa CBRG, which gathers in a single repository more than 1300 physicochemical and structural descriptors of all amino acid residues for each chain of all protein structures deposited in the PDB (Protein Data Bank – a world public repository where all macromolecular structures deciphered so far were deposited).

Based on the obtained results, we conclude that a given nanoenvironment can be described not by a single descriptor, but by a set of descriptors, and that this set of descriptors varies according to the element of the protein structure selected from a hierarchically superior one. This differentiates a given nanoenvironment from the rest of the protein and even from other nanoenvironments in the same protein. The knowledge acquired from the study of different nanoenvironments allows specialists in different areas, such as experts in plant improvement, in search of new pesticides, or researchers in search of more sustainable fuels, to advance their work with greater molecular introspection and use of more precise and refined tools, working at the most fundamental level (molecular-atomic) of all biologically relevant processes for medicine, agriculture, livestock, etc.

Protein nanoenvironments and their characteristics

The local structural environment of proteins, here called the nanoenvironment (Neshich et al., 2015), characterizes the functional purpose of different protein districts, also known as “structural sites” in proteins. It is therefore suggested that the local environment at each protein point and/or region reflects not only its structural role, but also its contribution in providing the necessary characteristics for the functional purpose of each protein. For example, protein-protein communication is performed via protein interfaces: amino acid residues at the same site have some particular characteristics that not only differentiate them from other residues on the free surface of the protein, but also allow specific and selective binding between proteins and the realization of their biochemical function (Moraes et al., 2014). Similarly, the function of an enzyme is normally related to the activity of its catalytic amino acid residues (Catalytic Site Residues – CSR). These very peculiar residues are inserted in a very specific nanoenvironment, also defined by the contribution of the CSRs themselves. Consequently, the enzymatic function can be described by the characteristics of the CSRs and their surroundings (Salim, 2015). Based on these considerations, and assuming that the local nanoenvironment defines the protein function, this is a concept that can be used to obtain specific metrics to quantify and describe other nanoenvironments.

The exploration of nanoenvironments properties of can be done through a method that is both self-explanatory and intuitive. Suppose it is possible to insert an imaginary probe anywhere in a protein structure and obtain as a result, a diagnosis describing the characteristics of the environment in which the probe is inserted. This type of physical intervention cannot be carried-out, and therefore the probe needs to be replaced by calculating values, metrics, and forces that we want to quantify at each particular site/point. This approach resembles the GRID method for calculating molecular interaction

fields in drug development (Goodford, 1985; Von Itzstein et al., 1993), but with a different focus. Its advantage is that any amino acid residue, or any of its main or side chain atoms, can serve as the center for the probe. With this selected point, the interactions of all forces can be estimated, cataloged, and stored in an appropriate relational database – in our case, the STING_RDB (Oliveira, 2007). Once stored, the attributes and their respective values can be mapped back to the protein structure, the protein sequence, or even the nucleotide sequence of the gene that encodes that protein, and can be used for visual inspection or statistical and/or numerical analyses. Our hypothesis is that any specific environment (the nanoenvironment) has a precise tuning of the specific physicochemical and structural writers for the performance of its function and, thus, can be identified and classified accordingly. For example, interfaces for protein contacts, which are specific areas of the protein occupying part of their surface, can be expected to have characteristics sufficiently different from the amino acid residues found in free surface areas (Moraes et al., 2014). In fact, we consider such an assumption to be part of the biological requirements for performing a specific function: in this example, the function is actually a kind of “communication” between very specific protein partners. Therefore, a nanoenvironment is accurately characterized by its physicochemical and/or structural descriptors and their corresponding values, making it possible to distinguish it from the rest of the protein structure. It is also possible to predict the coordinates of these districts in other proteins (homologous or not) that have not yet been chemically and functionally characterized through computational techniques and machine learning statistics.

Among the most studied protein nanoenvironments, ten stand out, as follows:

- 1) Protein interfaces: These are intersections of protein surfaces, where the two proteins approach and touch, building a macromolecule homo or heterocomplex (Moraes et al., 2014).
- 2) Antibody and antigen interfaces: as in case 1, but the two proteins in question are an antibody and an antigen (Viart et al., 2016).
- 3) Protein surface hot spots: locations delimited from the surface area of the protein, obligatorily located at its interface, and with identified hydrophobic amino acids prone to interact with similar residues from the complementary interface of the other protein (Pereira, 2012).
- 4) Interfaces between proteins and DNA: as in case 1, but with the two molecules in question being a protein and a DNA molecule.
- 5) Interfaces between proteins and ligands: as in case 1, here the two molecules in question are a protein and a ligand (Borro et al., 2016).
- 6) Interfaces between proteins and membranes.
- 7) Amino acid residues from catalytic sites: identifying the amino acid residues that form the enzymes catalytic site, determining their function (Salim, 2015).
- 8) Allosteric sites: usually located on the protein surface. When occupied by a particular molecule, they control the speed of a chemical reaction that the protein performs, using, as a rule, its set of CSRs as part of its function.
- 9) Secondary structure elements: physicochemical and structural characterization of α -helices (Mazoni et al., 2018), β -sheets and turns.
- 10) The depth of range of local sensing between amino acids: a measure often used to delimit the distance over which atoms, with their charges (and other characteristics), still exert some influence in remote locations, but within the aforementioned limit (Silveira et al., 2009).

Items: 1 to 6 describe the interfaces in general; 7 and 8 describe chemical activity of proteins; and 9 and 10 describe structural characteristics of proteins in general.

List of physicochemical and structural descriptors that characterize specific nanoenvironments

Currently, the Blue Star STING (BSS) (Neshich et al., 2006) has 32 independent physicochemical and structural protein descriptor types or classes (Table 1) (Neshich et al., 2005), and a total of 1,307 variations of these descriptors are pre-calculated (using different parameterizations) and stored in the STING_RDB database (Oliveira, 2007). On May 18, 2020, the STING_RDB had 151,711 structures, with 467,038 chains and 95,148,233 amino acid residues. For each, 1,307 parameters were pre-calculated, totaling 12×10^9 records in the database. Among these, some were chosen to be used in the nanoenvironment characterization and in the composition of their dictionary, considering only those that are more likely to be associated with pattern recognition processes in the selected proteins. For an adequate definition of the catalytic residues nanoenvironment and which is generally valid also for the other mentioned nanoenvironments, based on the physicochemical and structural descriptors, the descriptors referring to the conservation of amino acids were initially discarded, since these parameters are a measure of a set of homologous proteins and do not reflect any feature present in the protein structure (Salim, 2015).

Table 1. List of the 32 physicochemical descriptors classes and Blue Star STING structures.

Blue Star STING Descriptor Classes	
1. ResBoxes	17. Hot spots
2. Intra-chain atomic contacts [ITC]	18. Sequence conservation [HSSP]
3. The inter-chain atomic contacts [IFC]	19. Sequence conservation [SH ₂ Q ⁺]
4. ITC contacts energy	20. Solvent accessibility
5. IFC contacts energy	21. Dihedral angles
6. Interface area [IF]	22. Pockets/cavities
7. Water contacting [WC]	23. Electrostatic potential
8. Ligand pocket forming [LP]	24. Hydrophobicity
9. Surface forming [SF] residues	25. Curvature
10. Prosite	26. Distance from the N-/C-terminal
11. ProTherm	27. Density
12. Secondary structure indicator [PDB]	28. Sponge
13. Secondary structure indicator [DSSP]	29. Order of cross presence
14. Secondary structure [STRIDE]	30. Order of cross link
15. Multiple occupancy	31. Rotamers
16. Temperature factor	32. Space clash

Contributions

What does knowledge about protein nanoenvironments entail?

The protein's structure defines its functionality. However, how this is performed and which structural features contribute to their function remains to be fully deciphered. To answer this question, it is necessary to consider the structural elements (also called protein districts or nanoenvironments) rather than considering the structure as a whole. These elements, on the other hand, must be understood based on the physicochemical and structural characteristics from the amino acid residue properties, which interact with each other and create a new hierarchical structural element. Only by considering these elements in the structural hierarchy can we understand that the functionality of proteins can be broken down into communication elements, such as interfaces, constructive elements, secondary structure, and elements of chemical activity. The latter normally give rise to the functionality and specificity of the protein as a whole. Following this reasoning, each element in the structural hierarchy has its distinctive local characteristic and, consequently, its local function. It is clear that a general and detailed knowledge about protein nanoenvironments is, basically, a dictionary with which we can construct complex expressions to describe the structural-functional protein relationships.

A dictionary of nanoenvironment descriptors will impact the variety of research aimed at innovation in areas such as agriculture, medicine, and biology in general

A compilation of the results of work done since 1998 – when the STING platform was launched in the US as an integral part of platforms offered for protein structural analysis at the Brookhaven National Laboratory, the headquarters of the Protein Structures Database (PDB) – it resulted in a website called: “Dictionary of Internal Protein Nanoenvironments” (DIPN)¹.

Figures 1 to 3 show the general interface of the new CBRG offered by Embrapa Digital Agriculture. It is an introductory page, with a general description of the purpose of this platform with detailed elements listed in a functional order.

Figure 1 shows the entry page of the Dictionary of Internal Protein Nanoenvironments (DIPN) platform, indicating the purpose of this product, the options for access, the site organization logistics, and the list of the ten most studied protein nanoenvironments.

In Figure 2 we have a Dictionary of Internal Protein Nanoenvironments (DIPN) platform page showing six of the ten available nanoenvironments, with a short description and access to details of the entry of each option: a) protein-DNA interfaces, b) protein interfaces-membrane, c) elements of secondary structure.

Figure 3 presents a Dictionary of Internal Protein Nanoenvironments (DIPN) platform page showing three more of the ten available nanoenvironments, with a short description and access to details of each entry option: a) residues from the catalytic site, b) allosteric sites, and c) depth of local sensing range between amino acids.

¹ Available at: <https://www.proteinnanoenvironments.cnptia.embrapa.br/index.html>



The concept of internal protein nanoenvironment

The lab's research is driven by a conviction that internal protein structural districts/neighbourhoods, or, as we named them, Internal Protein Nanoenvironments (IPN), contain a significant core of information about their ultimate function. Such information content, fully describing corresponding nanoenvironments, is selectable in form of an ensemble of specific descriptors and corresponding values. The ensemble of physical-chemical and structural parameters is peculiarly less sensitive to localized variation of sequence encoding for that structure, causing limited structural promiscuity regarding underlining protein sequences, explaining why sequences may vary to a limited extent while resulting function remains unchanged.

In conclusion: What we found is that for each nanoenvironment there is a specific ensemble of descriptors, making possible their cataloguing into a dictionary of IPNs.

Also, the lab is continually employing leading initiatives to encourage and facilitate the use of "big data" in large-scale research across the scientific and technological disciplines.

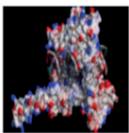
Ten most studied internal protein nanoenvironments

Share [+](#) [f](#) [t](#) [e](#) [in](#)

1. Protein-Protein Interfaces (PPI)
2. Hot spots (HS)
3. Antibody-antigen interfaces (AA)
4. Protein-Ligand interfaces (PL)
5. Protein-DNA interfaces (PD)
6. Protein-Lipid membrane interfaces (PLM)
7. Secondary structure elements (SSE)
8. Catalytic site residues (CSR)
9. Allosteric sites (AS)
10. Max distance reach for detection of AA Residue presence

Figure 1. Dictionary of Internal Protein Nanoenvironments (DIPN) platform entry page.

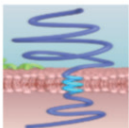
Source: Embrapa (2020).



Protein DNA interfaces

Protein-DNA interactions are key to control gene expression, transcription, DNA repair and DNA packing among all living beings. Due to its importance, several computational approaches that focus on the prediction of protein-DNA interacting protein residues are available in the scientific literature. Yet, only a fraction makes use of structural information and all of the available methods rely on amino acids conservation. The methods described here about this particular nanoenvironment are using statistical and machine learning approaches having at the input some physicochemical and structural descriptors from Blue Star Sting database that reflects the importance of each parameter in forming of the complex between proteins and DNA molecules. The developed approach is important in order to correctly assess protein function as well as identify important residues for composing protein-DNA interfaces. Likewise, DNA component is also included in terms of specific type of contacts established between two interacting macromolecules, an extremely valuable information for analysis of the protein-dna interface nanoenvironment.

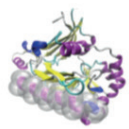
[Return later](#)



Lipid membrane -Protein Interfaces

Here, we analyze the nanoenvironment of interfaces formed among lipid membranes and proteins. Membrane proteins account for approximately one-third of the proteomes of all organisms and include receptors, structural proteins and channels. We focus on two types of membrane proteins: integral membrane proteins which are usually permanently anchored to the membrane, and so called peripheral membrane proteins, which are usually temporarily attached to a lipid bilayer. The nanoenvironments of such interfaces represent potential pharmacological targets of fundamental importance for a variety of diseases, with very important implications for the design and discovery of new drugs or peptides modulating or inhibiting relevant interactions.

[Return later](#)



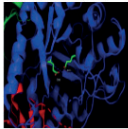
Secondary Structure Elements

Protein secondary structure elements (PSSEs) such as α -helices, β -strands, and turns are the basic building blocks of the tertiary protein structure. Our primary interest here is to reveal the characteristics of the nanoenvironment formed by both PSSEs and their surrounding amino acid residues (AARs), what might contribute to the general understanding of how proteins fold. The characteristics of such nanoenvironments must be specific to each secondary structure element, and we have set our goal here to gather the fullest possible description of the α -helical nanoenvironment first, and then for β -strands and turns. Here, published paper on this subject is presented as well as link to PhD theses with complete research data. Graphical and tabular information about nanoenvironments for α -helices, β -strands, and turns are offered for a user analysis.

[Learn More](#)

Figure 2. Dictionary of Internal Protein Nanoenvironments (DIPN) platform page.

Source: Embrapa (2020).

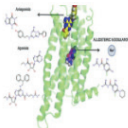


Catalytic Site Residues

The function of enzymes is determined by specific residues, called catalytic amino acid residues (CSR). The protein function is maintained for eons of selective pressure which preserves in its structure many physical-chemical and structural patterns. Frequently, enzymes from distinct organisms exert exactly the same biological function due to preservation of similar catalytic amino acid residues, even with evident low sequence similarity at the level of whole proteins. The majority of catalytic amino acid residues prediction methods use sequence conservation features to provide classification. Seeking to understand these conserved patterns in enzyme structures, that even after eons of evolution perform the same biological function, the present work searches to identify which protein structural descriptors (available in Blue Star STING platform) are capable of discriminating the amino acid catalytic residues from non-catalytic residues by means of their nanoenvironment properties.

Here too, we will offer to a user useful links such as the PhD theses elaborated at the University of Campinas, SP, Brazil. Viewing of the key biophysical, structural, biochemical and physical-chemical descriptors for CSR NANOENVIRONMENT is possible.

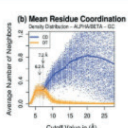
[Learn More](#)



Allosteric Sites

High-resolution structural data has been instrumental in characterizing and defining the stereochemical parameters that promote and define the binding of peptides, protein inhibitors and substrates at the active sites of the point that we now have a very comprehensive understanding of specific interactions and catalytic mechanisms which facilitates the design and development of highly selective synthetic inhibitors and drugs. On a more subtle level, protein surfaces often bristle with secondary binding sites (exosites) which serve key roles in regulating and modulating diverse activities ranging from gene expression, conformational stabilization, transport, inhibition and, by extension, the modulation and exhibition of distinct functions or moonlighting by the same enzyme in response to changes in physicochemical conditions is less well understood. Here we are compiling key descriptors for exosites so we may gain more insight on how those sites function and how are they regulating main protein function and / or are being regulated themselves.

[Return later](#)



Max Distance reach for detection of Amino Acid Residue presence

This particular entry in our dictionary of IPNs is not strictly speaking an environment. Rather, it is an feature of all nanoenvironments within a protein structure, which determines the extension or reach of presence of any amino acid residue "felt" across distance. For that, our work on atom coordination is presented as a useful tool for internal protein nanoenvironment understanding, in particular, establishment of contacts and "presence" among AAR - a very important descriptor in STING RDB.

Atom and residue contacts have been used in a wide range range of studies involving proteins and other biomolecules. Its correct and precise assignment comprise the touchstone of the most important structural analysis algorithms, which should be able to perform: packing calculations, functional similarities, evolutionary relationships, topological classifications, structural alignments, structural assessment, protein structure prediction, threading experiments, network contact analysis, empirical potentials, thermodynamic stability previews, folding inferences, protein-protein and protein-ligand interactions, and so forth.

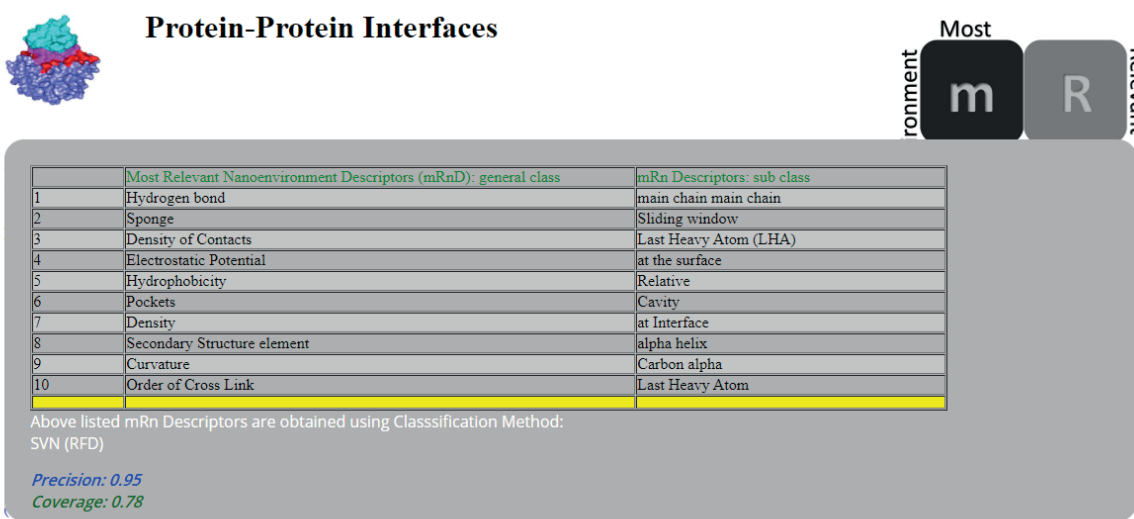
Here we will focus our attention on some methods that underlay contact characterizations in most of these applications.

[Learn More](#)

Figure 3. Dictionary of Internal Protein Nanoenvironments (DIPN) platform page.

Source: Embrapa (2020).

In Figure 4, the user can see the presentation details of one of the nanoenvironments: protein interfaces. The purpose of DIPN is to provide the user with information that indicates which are the most relevant descriptors that, with their specificity and broad coverage, describe the nanoenvironment selected for analysis. At the bottom of Figure 4 there is a table with the ten descriptors of the most relevant protein interfaces. These are: 1) main-chain main-chain hydrogen bonds; 2) spongicity (in a sliding window mode); 3) contact density between amino acids (centered on the last heavy atom of the amino acid side chain); 4) electrostatic potential on the protein surface; 5) hydrophobicity (on the relative scale); 6) structural pockets (cavity type); 7) atomic density on the surface; 8) element of the secondary structure present (α -helix); 9) curvature from α -carbon; and 10) the order of cross-linking (starting from the last heaviest atom in the side chain). These descriptors can be understood as main requirements that demand their inclusion so that a set of amino acids, not necessarily contiguous in the primary sequence, build a set that can be considered apt to form an interface with another protein. Then, the platform informs which statistical classification method was used to obtain this ranking of the importance of the descriptors (in this case: Support Vector Machine and Random Forest), and also informs with what precision and coverage the conclusions were reached. In this case, 0.95 and 0.78, respectively. On that



Protein-Protein Interfaces

Environment **m** **R** Relevant

	Most Relevant Nanoenvironment Descriptors (mRnD): general class	mRn Descriptors: sub class
1	Hydrogen bond	main chain main chain
2	Sponge	Sliding window
3	Density of Contacts	Last Heavy Atom (LHA)
4	Electrostatic Potential	at the surface
5	Hydrophobicity	Relative
6	Pockets	Cavity
7	Density	at Interface
8	Secondary Structure element	alpha helix
9	Curvature	Carbon alpha
10	Order of Cross Link	Last Heavy Atom

Above listed mRn Descriptors are obtained using Classification Method: SVN (RFD)

Precision: 0.95
Coverage: 0.78

Feedback

PhD Theses on Protein Protein Interfaces and corresponding nanoenvironment
Moraes, Fábio Rogério de, 2012

Characteristics of protein interface nano-environment revealed

[View PhD Theses](#)

Figure 4. Dictionary of Internal Protein Nanoenvironments (DIPN) platform page showing user choices.

Source: Embrapa (2020).

same page, you will find a variety of additional information, such as links to the doctoral thesis that generated the results and publications describing pertinent work to the subject (in Figure 5 we are illustrating the abstract of this publication). Lastly, there is a link so that the user can access the software if one wants to generate new data for a set of proteins for biological interest.

In Figure 4, we have the Dictionary of Internal Protein Nanoenvironments (DIPN) platform page showing user options once the protein interfaces item is selected. At the top of this figure, there is an indication of relevant publications to the subject and a list of software. Next, an abstract of the main publication can be seen describing our work with the nanoenvironment of protein interfaces, with a corresponding pointer to the original publication. On the upper right side, there is an icon with the title: MRND (Most Relevant Nanoenvironment Descriptors). By hovering over this icon, a window is opened with information indicated in the icon's title.

In Figure 6, we present the available items for accessing the software page which helps the user to prepare a list of descriptors for a set of proteins of interest. In Figure 7 we have the two main options for preparing protein interface data: LDA methodology (linear models for inferring the list of the most relevant descriptors of protein interfaces) and SHI option, an alternative methodology that determines the hydrophobicity index on the surface protein, an accurate interface indicator. The user can find a tutorial to find details about the software, datamart description for defining benchmarks and description of the complexes used in the training of the method using both homo and heteroprotein complexes. In Figures 1 to 7 we show only the most crucial entries of the DIPN platform.

The platform is complex and requires the knowledge of a trained computer biologist to process the data for a set of selected proteins. However, molecular biology specialists interested in knowing which descriptors are most relevant for each nanoenvironment listed in the DIPN platform can do so in a reasonable time, with minimal training, and know which characteristics of these nanoenvironments



Protein-Protein Interfaces

Publications and Software

Primary Publication and Software access

Fábio R. de Moraes, Izabella A. P. Neshich, Ivan Mazoni, Inácio H. Yano, José G. C. Pereira, José A. Salim, José G. Jardine, Goran Neshich

Improving Predictions of Protein-Protein Interfaces by Combining Amino Acid-Specific Classifiers Based on Structural and Physicochemical Descriptors with Their Weighted Neighbor Averages;

PLoS One. 2014 Jan 28;9(1):e87107.
doi: 10.1371/journal.pone.0087107.
eCollection 2014.

[View Abstract](#) [Access Software](#)

Abstract

Protein-protein interactions are involved in nearly all regulatory processes in the cell and are considered one of the most important issues in molecular biology and pharmaceutical sciences but are still not fully understood. Structural and computational biology contributed greatly to the elucidation of the mechanism of protein interactions. In this paper, we present a collection of the physicochemical and structural characteristics that distinguish interface-forming residues (IFR) from free surface residues (FSR). We formulated a linear discriminative analysis (LDA) classifier to assess whether chosen descriptors from the BlueStar STING database (<http://www.cbi.cnpqia.embrapa.br/SMS/>) are suitable for such a task. Receiver operating characteristic (ROC) analysis indicates that the particular physicochemical and structural descriptors used for building the linear classifier perform much better than a random classifier and in fact, successfully outperform some of the previously published procedures, whose performance indicators were recently compared by other research groups. The results presented here show that the selected set of descriptors can be utilized to predict IFRs, even when homologue proteins are missing (particularly important for orphan proteins where no homologue is available for comparative analysis/indication) or, when certain conformational changes accompany interface formation. The development of amino acid type specific classifiers is shown to increase IFR classification performance. Also, we found that the addition of an amino acid conservation attribute did not improve the classification prediction. This result indicates that the increase in predictive power associated with amino acid conservation is exhausted by adequate use of an extensive list of independent physicochemical and structural parameters that, by themselves, fully describe the nano-environment at protein-protein interfaces. The IFR classifier developed in this study is now integrated into the BlueStar STING suite of programs. Consequently, the prediction of protein-protein interfaces for all proteins available in the PDB is possible through STING_interfaces module, accessible at the following website: (<http://www.cbi.cnpqia.embrapa.br/SMS/predictions/index.html>).

See complete publication @:
10.1371/journal.pone.0087107

PhD Theses on Protein Protein Interfaces and corresponding nanoenvironment

Moraes, Fábio Rogério de, 2012

Characteristics of protein interface nano-environment revealed

[View PhD Thesis](#)

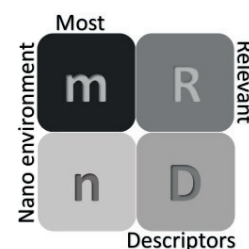


Figure 5. Page in the Dictionary of Internal Protein Nanoenvironments (DIPN) platform, with option for a quick view of the publication's abstract. In this case, an article published in a renowned journal in the field of computational biology about nanoenvironment.

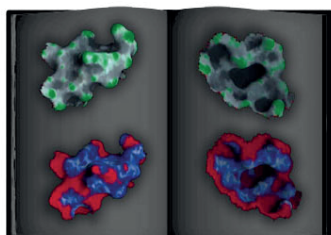
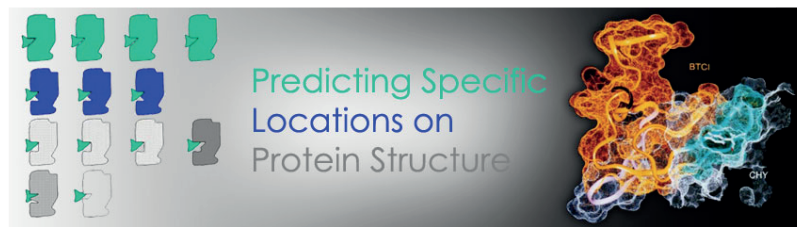
Source: Embrapa (2020).

are crucial. Thus, there are candidates that cannot perform modifications, for example in attempts that require site-directed mutations in the proteins of interest. The algorithm options for using or even accessing the source code are provided in order to offer a complete work environment, including for those computational biologists who wish to adapt the algorithms to their own requirements. This allows the sharing of work already carried out by Embrapa, and may be modified by colleagues in other laboratories for specific purposes.

Final considerations

With a dictionary of descriptors of the main protein nanoenvironments, a reality is built that guides researchers and enables advancement in areas aiming to intensify innovation in agriculture, medicine, and biology in general. It is understood that a compilation of the essential descriptors of the 10 most studied

Home	Interfaces	CSRs	Secondary Structure	BlueStar STING
------	------------	------	---------------------	----------------



Interfaces

Proteins and in particular enzymes, interact with their substrates and/or inhibitors through a specific area of their surfaces called **interface**. The **interfaces** are composed in a such way (from the 20 regular amino acids) so that there is a particular nano environment that they create and by doing so, they can be recognized by the substrate and/or inhibitor. In other words, **interface** acts as it is emitting a specific signal to the molecules in the solute indicating which one of those molecules can bind to a protein and on what location of its surface. The **interfaces** are defining specificity of enzymes. A catalytic site on the other hand defines the nature of chemical reaction that would be performed on substrate and consequently, the nano environment created by **Catalytic Site Residues (CSRs)** is responsible for identification of the protein function. The location of an **interface** on protein surface is a key factor which guides substrate docking to a protein and experiments designed to change the specificity of an enzyme need to start exactly with the detailed knowledge of where that **interface** is located. Specificity is therefore defined by composition and characteristics of the **interfaces** while function is generally defined by a fraction of the **interface** - the catalytic site.



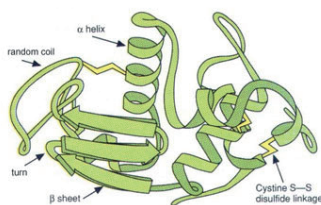
Catalytic Site Residues

In our [previous work](#) we were motivated to identify those amino acids with decreased accessibility to solvent after docking of different types of inhibitors to sub classes of serine proteases and then create a table (matrix) of all amino acid positions at the **interface** as well as their respective occupancies. Our goal was to establish a platform for analysis of the relationship between Interface Firming Residues (IFRs) characteristics and binding properties/specificity for bi-molecular complexes

In this work we expand the initial goal by first studying how often protein use the hydrophobic effect for oligomerization and then apply such basic knowledge to generate an algorithm for predicting the **interface** area on any protein structure, considering only the Surface Hydrophobicity Index - a new index we elaborated in order to measure how hydrophobic are protein surfaces and corresponding **interfaces**.

In addition, we studied the characteristics of the nano environment created by amino acids which constitute any given **interface** and by learning from those characteristics, we were able to create an algorithm which we can now use for predicting the location of an **interface**.

Finally, as the **catalytic site residue** occupy generally only a fraction of the positions among the **interface** residues, we focused our work to first catalogue and then understand the environment of **CSRs** and by doing so, elaborate the algorithm for identification of **CSRs** and creation of a sort of "Periodic Table of Protein Families", based exclusively on selection of few descriptors of sequence and structure (and their value ranges) which can then be used as a sole identifiers of **CSR** for each protein family.



Secondary Structure Elements

Figure 6. Dictionary of Internal Protein platform page of Nanoenvironments (DIPN), with access to software that ranks protein interface nanoenvironment descriptors, catalytic residues, and secondary structure elements of the most relevant proteins.

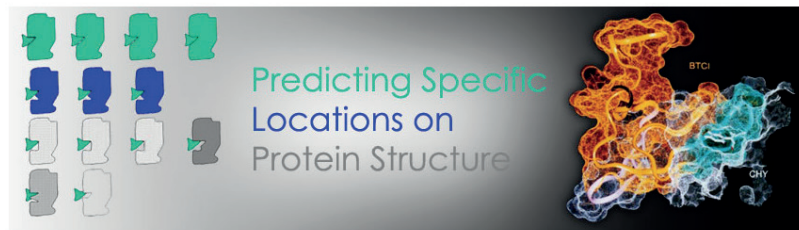
Source: Screen captured from the DIPN platform (available at: <https://www.proteinnanoenvironments.cnptia.embrapa.br/index.html>)

protein nanoenvironments may provide an optimized condition for the most accurate, effective, and effective design of new drugs, pesticides, vaccines, inhibitors, catalysts, and antibodies. We can use as an example, the applicability of the content presented in this chapter and mention some of the technologies which the CBRG of Embrapa Digital Agriculture managed to file. This resulted in the application for four patents over the years, focusing mainly on understanding, learning, and in the analysis of protein nanoenvironments which were crucial to the solution of some biologically relevant demands. The research group also focused on a path to the necessary impacts in the field for producers who needs to use the technology in order to avoid losses and improve its effectiveness. Some of them are listed below:

Fungicide: a method for designing a new fungicide by computationally designing new compounds with potential inhibitory function on the endopolygalacturonase enzyme, involved in invasion processes in plant cells. (Neshich et al., 2013a)

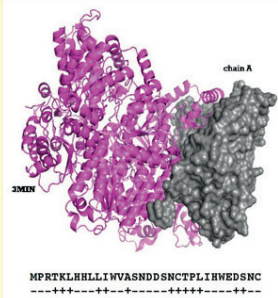
Biodiesel: method for predicting mutants that increase the surface hydrophobicity index of proteins. (Neshich et al., 2013b)

Home	Interfaces	CSRs	Secondary Structure	BlueStar STING
------	------------	------	---------------------	----------------



LDA STING Interfaces

[Linear Model for Protein - Protein Interface Prediction - Methodology Description](#)

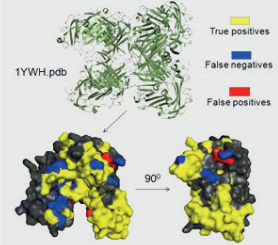


- Predict Protein Interface location by using specific protein structure and LDA_sting algorithm (PUBLIC PDB files)
- Predict Protein Interface location by using specific protein structure and LDA_sting algorithm (modelled and non-public PDB format files)

[Tutorial](#) Datasets: [DS30](#) [DS70](#) [DS95](#)
[Partial Source Code](#) [Hetero Complexes](#) [Homo Complexes](#)

USI-PEPI STING Interfaces

[USI-PePPI, a Systematic Neural Network-based Methodology for Predicting Protein-Protein Interfaces using STING Database descriptors](#)



- [Predict Interface location by using specific protein structure and USI_PePPI algorithm](#)
- [Supplementary material](#)

SHI STING Interfaces

[Surface Hydrophobicity Index \(SHI\): insights into the relationship between hydrophobic effect and oligomerization](#)



- [Predict Interface location by using specific protein structure and ΔSHI algorithm](#)
- [Supplementary material](#)

Figure 7. Page of the Dictionary of Internal Protein Nanoenvironments (DIPN) platform, with options for users who want to rank the most relevant protein interfaces descriptors by using a set of proteins of interest for a biological problem that requires their engagement.

Insecticide: computational design for new alpha-amylase inhibitors. (Neshich et al., 2013c)

Bactericide: identification of therapeutic targets for computational design of drugs against bacteria possessing the pilt protein. (Neshich et al., 2012)

These four technologies reflect the strong interdependence between the demands of modern agriculture and knowledge, calling for an innovative, interdisciplinary, and molecular approach, interconnected with mathematics, computation, and statistics for advances in the increasingly complex needs of the productive sector. The example of the CBRG at Embrapa Digital Agriculture is a manifestation of national possibilities for the potential of technological development at the highest and most competitive level. The research carried out by the CBRG at Embrapa Digital Agriculture drew the attention of international collaborators and colleagues from the most renowned universities, such as Oxford, Cambridge, MIT, followed by companies with great digital impact, such as Microsoft Research and companies in the field of agricultural pesticides, such as Bayer and BASF. Half a hundred publications in scientific journals with an average impact factor of 3, and several with impact factors above 11. There were hundreds of lectures and seminars, international courses, and workshops, as well as international meetings organized here in national territory with the participation of several Nobel Prize-winning scientists. Fifty software packages were published and made available for the scientific community, as well as dozens of databases in the field of computational structural biology, including the STING_RDB. Twenty-six projects were approved (90%) by external sources to Embrapa, with funding approaching 4 million dollars and total deliverables approaching 500 million. This entire library of results and professional awards was a stepping-stone for us to transform our acquired knowledge into something applicable to the production chain and developing these solutions into products for national and international markets. Therefore, the platform called Dictionary of Internal Protein Nanoenvironments was developed while considering the applications from our knowledge, but with patience and determination to stay on the path that requires time, learning, and basic science, since scientific applications do not exist without the former.

References

- BORRO, L.; YANO, I. H.; MAZONI, I.; NESHICH, G. Binding affinity prediction using a nonparametric regression model based on physicochemical and structural descriptors of the nano-environment for protein-ligand interactions. In: STRUCTURAL BIOINFORMATICS AND COMPUTATIONAL BIOPHYSICS, 2016, Orlando. **Proceedings...** Orlando: [s.n.], 2016. p. 116-117.
- EMBRAPA. Computational Biology Research Group. **Dictionary of Internal Protein NanoEnvironments**. Available at: <https://www.proteinnanoenvironments.cnptia.embrapa.br/index.html>. Accessed on: 18 May 2020.
- GOODFORD, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. **Journal of Medicinal Chemistry**, v. 28, n. 7, p. 849-857, July 1985. DOI: [10.1021/jm00145a002](https://doi.org/10.1021/jm00145a002).
- MAZONI, I.; BORRO, L. C.; JARDINE, J. G.; YANO, I. H.; SALIM, J. A.; NESHICH, G. Study of specific nanoenvironments containing α -helices in all- α and $(\alpha + \beta)$ proteins. **PLOS One**, v. 13, n. 7, p. 1-25, 2018. Artigo e0200018. DOI: [10.1371/journal.pone.0200018](https://doi.org/10.1371/journal.pone.0200018).
- MORAES, F. R. de; NESHICH, I. A. P.; MAZONI, I.; YANO, I. H.; PEREIRA, J. G. C.; SALIM, J. A.; JARDINE, J. G.; NESHICH, G. Improving predictions of protein-protein interfaces by combining amino acid-specific classifiers based on structural and physicochemical descriptors with their weighted neighbor averages. **PLOS ONE**, v. 9, n. 1, p. 1-15, 2014. DOI: [10.1371/journal.pone.0087107](https://doi.org/10.1371/journal.pone.0087107).
- NESHICH, G.; JARDINE, J. G.; NESHICH, I. A.; SALIM, J. A.; MAZONI, I. **EUA Patente Nº WO2013/110147A1**, 2013c.
- NESHICH, G. E. A.; BORRO, L. C.; HIGA, R. H.; KUSER, P. R.; YAMAGISHI, M. E.; FRANCO, E. H.; KRAUCHENCO, J. N.; FILETO, R.; RIBEIRO, A. A.; BEZERRA, G. B.; VELLUDO, T. M.; JIMENEZ, T. S.; FURUKAWA, N.; TESHIMA, H.; KITAJIMA, K.; BAVA, A.; SARAI, A. TOGAWA, R. C.; MANCINI, A. L. The diamond STING server. **Nucleic Acids Research**, v. 33, n. 2, p. W29-W35, July 2005. Supplement. DOI: [10.1093/nar/gki397](https://doi.org/10.1093/nar/gki397).
- NESHICH, G.; MAZONI, I.; OLIVEIRA, S. R. M.; YAMAGISHI, M. E. B.; KUSER-FALCÃO, P. R.; BORRO, L. C.; MORITA, D. U.; SOUZA, K. R. R.; ALMEIDA, G. V.; RODRIGUES, D. N.; JARDINE, J. G.; TOGAWA, R. C.; MANCINI, A. L.; HIGA, R. H.; CRUZ, S. A. B.; VIEIRA, F. D.; SANTOS, E. H.; MELO, R. C.; SANTORO, M. M. The Star STING server: a multiplatform environment for protein structure analysis. **Genetics and Molecular Research**, v. 5, n. 4, p. 717-722, 2006.
- NESHICH, G. E. A.; NESHICH, I. A. P.; MORAES, F.; SALIM, J. A.; BORRO, L.; YANO, I. H.; MAZONI, I.; JARDINE, J. G.; ROCCHIA, W. Using structural and physical-chemical parameters to identify, classify, and predict functional districts in proteins – the role

of electrostatic potential. In: ROCCHIA, W.; SPAGNUOLO, M. (ed.). **Computational electrostatics for biological applications: geometric and numerical approaches to the description of electrostatic interaction between macromolecules**. Cham: Springer, 2015. p. 227-254. DOI: [10.1007/978-3-319-12211-3_12](https://doi.org/10.1007/978-3-319-12211-3_12).

NESHICH, G.; JARDINE, J. G.; NESHICH, I. A.; SALIM, J. A.; MAZONI, I. **EUA Patente Nº WO2012/031343A2**, 2012.

NESHICH, G.; JARDINE, J. G.; NESHICH, I. A.; SALIM, J. A.; MAZONI, I. **EUA Patente Nº WO2013097012A1**, 2013a.

NESHICH, G.; JARDINE, J. G.; NESHICH, I. A.; SALIM, J. A.; MAZONI, I. **EUA Patente Nº WO2013/016794A1**, 2013b.

OLIVEIRA, S. R. de M.; ALMEIDA, G. V.; SOUZA, K. R. R.; RODRIGUES, D. N.; KUSER-FALCÃO, P. R.; YAMAGISHI, M. E. B.; SANTOS, E. H. dos; VIEIRA, F. D.; JARDINE, J. G.; NESHICH, G. Sting_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. **Genetics and Molecular Research**, v. 6, n. 4, p. 911-922, 2007.

PEREIRA, J. G. D. C. **Caracterização dos aminoácidos da interface proteína-proteína com maior contribuição na energia de ligação e sua predição a partir dos dados estruturais**. 2012. 106 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Genética e Biologia Molecular, Instituto de Biologia, Universidade Estadual de Campinas, Campinas.

SALIM, J. A. **Aplicação de técnicas de reconhecimento de padrões usando os descritores estruturais de proteínas da base de dados do software STING para discriminação do sítio catalítico de enzimas**. 2015. 214 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia Elétrica, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas.

SILVEIRA, C. H. da; PIRES, D. E. V.; MINARDI, R.; RIBEIRO, C.; VELOSO, C. J. M.; LOPES, J. C. D.; MEIRA JÚNIOR, W.; NESHICH, G.; RAMOS, C. H. I.; HABESCH, R.; SANTORO, M. M. Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. **Proteins: structure, function, and bioinformatics**, v. 74, n. 3, p. 727-743, Feb. 2009. DOI: [10.1002/prot.22187](https://doi.org/10.1002/prot.22187).

VIART, B.; DIAS-LOPES, C.; KOZLOVA, E.; OLIVEIRA, C. F. B.; NGUYEN, C.; NESHICH, G.; CHÁVEZ-OLÓRTEGUI, C.; MOLINA, F.; FELICORI, L. F. EPI-peptide designer: a tool for designing peptide ligand libraries based on epitope–paratope interactions. **Bioinformatics**, v. 32, n. 10, p. 1462-1470, May 2016. DOI: [10.1093/bioinformatics/btw014](https://doi.org/10.1093/bioinformatics/btw014).

VILLANUEVA, J. C. How many atoms are there in the Universe. **Universe Today**, July 30, 2009. Available at: <https://www.universetoday.com/36302/atoms-in-the-universe>. Accessed on: 18 May 2020.

VON ITZSTEIN, M.; WU, W.-Y.; KOK, G. B.; PEGG, M. S.; DYASON, J. C.; JIN, B.; VAN PHAN, T.; SMYTHE, M. L.; WHITE, H. F.; OLIVER, S. W.; COLMAN, P. M.; VARGHESE, J. N.; RYAN, D. M.; WOODS, J. M.; BETHELL, R. C.; HOTHAM, V. J.; CAMERON, J. M.; PENN, C. R. Rational design of potent sialidase-based inhibitors of influenza virus replication. **Nature**, v. 363, p. 418-423, 1993. DOI: [10.1038/363418a0](https://doi.org/10.1038/363418a0).