Towards exploring the activity landscape of peptide datasets using MAP4 fingerprint

Edgar López-López [1,2,*], Oscar Robles[3], Fabien Plisson[4] and José L. Medina-Franco[1*]

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, 04510 Mexico City, Mexico

[2] Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Section 14-740, 07000 Mexico City, Mexico

[3] Medicinal Chemistry and Chemogenomics Laboratory, Faculty of Bioanalysis-Veracruz, Universidad Veracruzana, 91700 Veracruz, Mexico

[4] Department of Biotechnology and Biochemistry, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Irapuato Unit, Irapuato 36824, Mexico.

*Corresponding authors: elopez.lopez@cinvestav.mx; medinajl@unam.mx

Abstract

Peptides are a re-emerged strategy to fight a plethora of diseases and their utility has been expanded to new areas. Now sequence-based peptide design opens up new possibilities to develop peptidic molecular entities. However, its methodological limitations (e.g., its inefficiency in designing large peptides and that do not allow the analysis of post-traductional modification) limit their applicability domain. In contrast, ligand-based molecular design approaches have demonstrated their extensive applicability domain, although the peptide design-based in this method continues been not exploited. The main limitation has been the complex molecular structure of peptides, which has not been studied using classical fingerprints tuned for small organic compounds. Towards this end, MAP4 is a recently developed universal fingerprint that allows quantifying the sequence/structure diversity of natural products or peptides. As part of the peptide design, there is a current trend to develop predictive models which are founded on the available structure-activity data available. Before developing such models, it is essential to characterize in detail the structure-activity relationship and identify if any activity cliffs: subtle structural modifications that have a large and unexpected effect on the biological activity. In this study, we map the structure-activity landscape of an exemplary dataset with 165 peptides (anti-methicillin-resistant Staphylococcus aureus peptides) using a similarity metric based on MAP4 fingerprint. Specifically, we characterized the activity landscape of this data set, and we identified key amino acids

(AAs) and structural motifs that play a key role in the activity of the anti-methicillin-resistant Staphylococcus aureus peptides. To the best of our knowledge, this is the first chemoinformatics approach to systematically explore the activity landscape of peptides emphasizing the quantification of the structural similarity. The approach is general and can be extended to analyze the presence of activity cliffs in any set of peptides. Identifying activity cliffs has practical implications during the development of predictive models.

Keywords: Activity landscape modeling, activity cliffs, chemoinformatics, chemical space, drug discovery, MAP4, peptide design, structure-property (activity) relationships, *Staphylococcus aureus.*

Introduction

Peptides play important roles in plant and animal physiology targeting various proteins including growth factors, ion channels, receptors, and enzymes. Many peptides exhibit a broad range of biological activities, all valuable starting points to treat human disorders [1–3]. The peptide space is vast; a peptide sequence of length N could lead to 20N possible mutations solely with canonical residues. Adding the post-translational modifications and the number of mutations becomes astronomical. It is not practical to synthesize all sequences or even to investigate all functionally interesting variants. A central goal for computational peptide design is to create novel sequences that carry the underlying properties of natural peptides with defined structural and functional properties. Multiple bioinformatic approaches have proven to be useful in accelerating peptide design learning either from their sequences or their tridimensional structures [4,5]. In addition, the automation of peptide synthesis on solid support or the heterologous expression of proteins across biological systems has reduced production costs, making peptide space exploration accessible. These in silico methods are predominantly learning from primary sequences from medium-large datasets, rather than their structures due to the high costs associated with solving structures experimentally [6]. However, current sequence-based approaches do not systematically study post-translational modifications (PTMs) that can significantly affect the physicochemical, chemical, or biological properties of peptides [7].

Historically, chemoinformatics and bioinformatics fields were considered independent between them. However, both fields use computational methods to analyze and manipulate large datasets to identify patterns and relationships in the data. The main historic difference is that chemoinformatics studies small chemical structures (e.g. drugs and natural products), in contrast to bioinformatics which allows the study of more complex molecules (e.g. peptides and proteins). But, in recent years new methodologies and technologies have been reducing the gap between chemoinformatics and bioinformatics fields. For example, new molecular representations based on atom-connectivity allow the systematic study of complex molecules, that could be applied to mapping the structural diversity of peptides and may help to understand the roles of PTMs in their physicochemical properties or biological activities [8].

Different computational strategies to develop peptides are based on the analysis of sequence alignments and physicochemical similarity metrics. [9]. However, the lack of data limits the use of alignment algorithms and the classification and prediction of secondary structures [10]. Fortunately, in the last ten years, new computational methods have contributed to decoding the structure-property relationships (SPR) of peptides (P-SPR) [11,12]. For example, new approaches that depend on the sequence and/or features of peptides (e.g., machine-learning methods, the *de novo* design, linguistic modeling, pattern insertion methods, and genetic algorithms) [13], are new research opportunities to explore P-SPR and guide a new era of peptide-based drug design.

Recently, computational drug design approaches have been used to decode the physicochemical and sequence-activity relationships on peptides [14,15]. However, has yet to be a reported study describing the relationship between small structural changes and their specific biological activity.

Additionally, physicochemical properties are used to compare, filter, and classify molecular structures of pharmaceutical interest. However, the main limitation of physicochemical properties is that these do not describe structural conformations, small chemical changes, or fold peptide differences. But, using consensus similarity metrics (that consider physicochemical and structural approaches) [13,16] could be implemented to complement the description and comparison between peptide structures using e.g., considering different peptide features: 3D structure, topology, backbone structure, drug-like properties, amino acid sequence, molecular fingerprints, etc.). From a conceptual point of view, a combination of descriptors and structural representations commonly used in chemoinformatics and bioinformatics would provide a comprehensive picture of the peptides [17]. For example, Plisson *et al.* recently demonstrated that it is possible to predict properties and design new peptide structures from the consensus description of known peptidic information [18]. In that work, the authors remark that “not all similar peptides conserve necessarily similar properties”. Such a highlight is related to the concept of “activity cliff” [19]: a pair of compounds (e.g. peptides) with high structural similarity but large and unexpected potency difference. Recently, analysis of activity cliffs has been used to decode structure-activity relationships of linear and circular peptides against different endpoints [20,21]. Also, the presence of activity cliffs in data sets reduces the performance of predictive models [22], including the recent machine and deep learning models [23].

This study discusses a new approach to aid in exploring and describing the activity landscape of peptides. As a case study, we use an exemplary data set of 165 peptides with reported activity against methicillin-resistant Staphylococcus aureus (MRSA) strains (one of the most important endpoints to developing new drug candidates according to the WHO) [24], including identifying anti-MRSA peptide activity cliffs. To this end, we employed an atom-connectivity fingerprint recently developed and well-suited to represent peptides. Finally, we discuss an interpretation of the peptide activity cliffs.

Methodology

Data set

To analyze the landscape of anti-MRSA peptides, we collected a total of 165 peptides sequences from the Antimicrobial Peptide Database [25], of which 59 examples (~35%) have a half minimum inhibitory concentration (MIC_{50}) value measured against clinical isolation of MRSA strains. In total, 106 peptides (~65%) have MIC_{50} value reported against at least one of the 34 characterized MRSA strains. We transformed all MIC_{50} values to a negative decimal listed logarithm scale as follows: ($-\mathit{log}\ MIC_{50}$). The values range from 3.89 to 6.69 and are listed in Supplementary material - Table S1 alongside the peptide sequences and SMILES representations. In some cases, the same peptide has been evaluated against different strains, we only kept the lower value of (p)MIC_{50}.

Activity landscape modeling

We studied the activity landscape of these peptides through two approaches that are frequently used with small organic compounds: Structure-Activity Similarity (SAS) map and the Structure-Activity Landscape Index (SALI). Both approaches are explained hereunder.

A (SAS) map is a low-dimension graph for analyzing the structure-activity relationships (SAR) of the compound dataset. SAS maps are one of the early approaches to studying activity landscapes and rapidly identifying activity cliffs. Activity cliffs are pairs of compounds with high structure similarity but significantly different biological activity [19].

SAS maps are based on systematic pairwise comparisons of the compounds in a data set. A general schematic representation of a SAS map is presented in Figure 1.

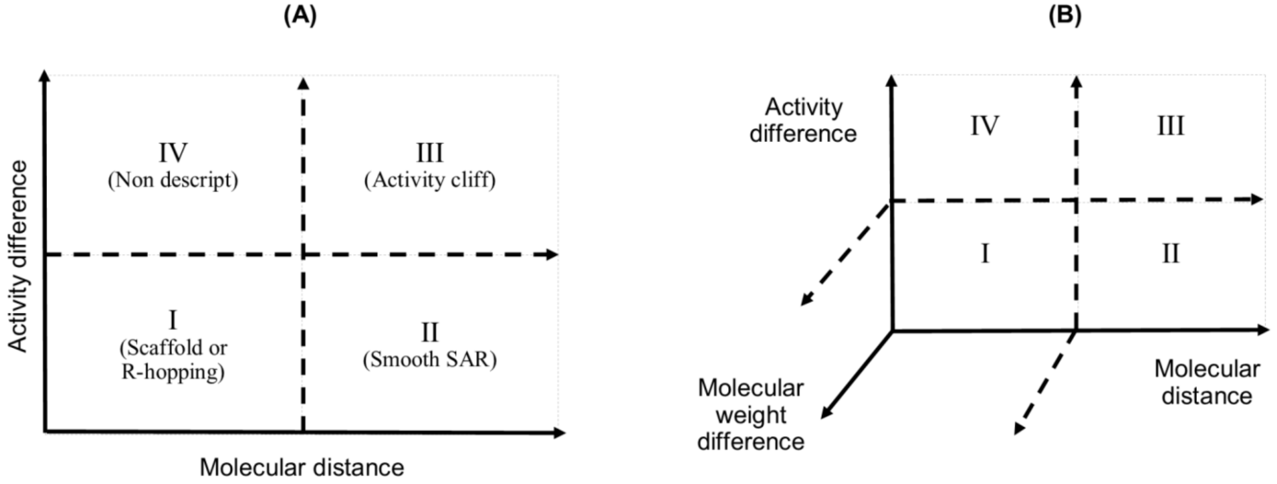


Figure 1. Graphical representation of a Structure-Activity Similarity (SAS) map (**A**) and an extension of a SAS map (**B**). A SAS map is based on a pairwise comparison of each compound on a data set. Each data point in the graph in the map represents a pair of compounds. SAS map is based on the activity differences of the pair of compounds against a specific biological endpoint, and their molecular distance. (**A**) Map with four regions: **I** identifies pair of compounds with low activity difference and low molecular distance (also called scaffold or R- hopping, o similarity cliffs); **II** represents pair of compounds with low activity difference and higher molecular distance (smooth SAR cases); **III** represents pair of compounds with higher activity differences and higher molecular distance (activity cliff); and **IV** represent pair of compounds with a discontinuous SAR [26] . (**B**) An extension of the conventional SAS map (extended SAS map) implemented in this study adds the molecular weight differences as a new axis.

SAS maps generated in this study represented all 13,533 pairwise comparisons between the 165 peptides of the dataset. The map calculated structure similarity with the MAP4 fingerprint [27] and the MinHashed distance function proposed by Capecchi *et al* [27], represented on the X-axis. The activity difference (i.e., the difference between them and the $pMIC_{50}$ values of the pair of peptides) was plotted on the Y-axis. The molecular weight differences between each pair of peptides are plotted on the Z-axis. The data points in the SAS maps were further colored by their SALI value. This index quantifies the activity landscape using the expression proposed by Guha and Van Drie [28,29] (Equation **(1)**):

$$\textbf{(1)}\quad \mathrm{SALI}\ i, j = |\mathrm{A}i - \mathrm{A}j| /\ 1 - \mathrm{sim}(i, j)$$

where A*i* and A*j* are the activities of the *i* and *j* molecules, and sim(*i*, *j*) is the similarity coefficient between *i* and *j*. Herein, sim(*i*,*j*) was computed with the MAP4 fingerprint and the MinHashed distance function. The SALI values were further mapped onto the SAS map using a continuous color scale from blue (low SALI values) to red (high SALI values that are associated with activity cliffs).

Chemical space of anti-MRSA peptides

A visual representation of the chemical space of anti-MRSA peptides was constructed using a Treemap (TMAP). TMAP allows the visual representation of many chemical compounds through the distance between the clusters and the cluster's detailed structure through Local Sensitive Hashing (LSH) forest data structure, enabling c-approximate k-nearest neighbors (k-NN) [30]. MAP4 fingerprints for peptides were encoded using the MinHash algorithm. The number of nearest neighbors, k = 50, and the factor used by the augmented query algorithm, kc = 10, were used to develop the TMAP graphs. The activity values were represented using a color scale from red (most active peptide; 6.69 $pMIC_{50}$) to blue (most inactive peptide; 3.89 $pMIC_{50}$).

Results

Data set description

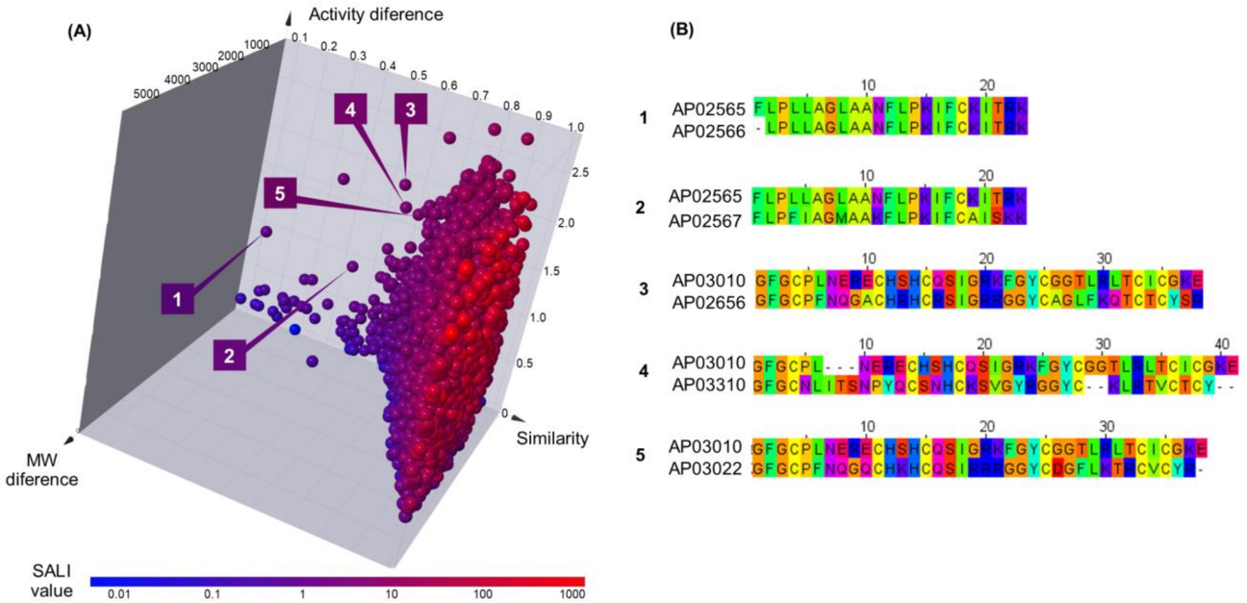
We collected 165 anti-MRSA peptides from APD03 alongside their $pMIC_{50}$ values. Most peptides were identified from amphibians (55 / 34%), followed by bacteria (32 / 20%), and arthropods (17 / 10%) - see Figure S1-A in the Supplementary material.

Regarding the types of structures reported in the set of peptides analyzed (Figure S1-B), the predominant structure (34%) in them is the alpha helix. The primary method used to elucidate the structure of anti-MRSA peptides with known structures was nuclear magnetic resonance (23 / 14% - Figure S1-C in the Supplementary material). Also remarkable is that a representative portion of the structures (41 / 25%) of the peptides studied herein are predicted computationally, and more than half of the peptidic structures of this data set have not been associated with any experimental or predicted structure (94 / 57%).

Activity landscape modeling

Figure 2-A shows an extended SAS map annotated with SALI values of 13,376 pairwise comparisons between the 165 peptides, which facilitates the identification of activity cliffs. Namely, the extended SAS map allows the identification of pairs of peptides with high structural similarity as determined by MAP4/MinHashed distance function (similarity > 0.40) but with a large anti-MRSA activity difference ($pMIC_{50}$ difference > 0.90), and with low molecular weight difference (MW difference < 150). With these criteria, we select, as representative examples, five peptide pairs (**1** - **5** in Figure 2-A).

Sequence alignment of the peptide pairs **1** - **5** (Figure 2-B) confirms that these pairs of peptides have comparable amino acid sequences (from 0.13 to 0.61). However, the distance calculated based on their peptide chemical structure does not necessarily have a linear relationship with the identity of peptide sequences (Figure 2-C). This observation suggests that the MAP4 fingerprints are helpful to compare the chemical structures of the peptides in addition to the AA sequence. Nevertheless, the 2D and 3D alignments of AA typically used in bioinformatics provides additional and intuitive information to decode the P-SPRs. For example, peptide pair **1** (Figure 2-C) shows a low fingerprint-based similarity (0.131) in contrast with their sequence-based identity (100%). This data suggests that the fingerprint-based similarity could be highly sensitive to small structural changes in short peptides (< 20 AA). In contrast, the peptide pairs **2** - **5** exhibit good relationships between their fingerprint-based similarity and their sequence-based identity (0.43 vs. 69%; 0.55 vs. 55%; 0.57 vs. 57%; 0.61 vs. 52%, respectively). This suggests that our fingerprint-based similarity metric could be useful to quantify the similarity between medium-large peptides (> 20 AA).



(C)

Pair	Peptide pair	Peptide activity 1	Peptide activity 2	Activity Difference ($pMIC_{50}$)	Fingerprint-based similarity	SALI value	MW difference	Percentage of identity
1	**AP02565 — AP02566**	318.4 µM	2.1 µM	0.96	0.131	1.11	147.2	100.00
2	**AP02565 — AP02567**	318.4 µM	2.4 µM	0.90	0.438	1.61	33.0	69.57
3	**AP03010 — AP02656**	32 µM	0.4 µM	1.90	0.546	4.19	38.0	55.26
4	**AP03010 — AP03022**	32 µM	0.63 µM	1.70	0.569	3.96	121.1	56.78
5	**AP03010 — AP03311**	32 µM	0.68 µM	1.66	0.612	4.31	90.2	52.78

Figure 2. Structural and sequence similarity of the dataset of 165 anti-MRSA peptides studied in this work. **A**) Modified (extended) structure-activity similarity map; each sphere represents a pairwise comparison of the chemical structure (quantified by means of MinHassed distance/MAP4 fingerprints), activity difference, and molecular weight difference. The spheres are colored according to the SALI values using a continuous scale from low values (blue) to high values (red). **B**) Sequence alignment of representative peptide activity cliffs; **C**) Summary characterization of selected peptide pairs. SALI: Structural-Activity Landscape Index.

Visualization of chemical space

In addition to the extended SAS map, we explore the anti-MRSA peptide landscape using a TMAP (Figure 3). A TMAP shows the □-nearest neighbors of each peptide (represented with a sphere) using as distance metric the MAP fingerprint and the MinHashed algorithm. Namely, the TMAP facilitates the identification and intuitive visualization of compounds (e.g., peptides) structurally related. For example, the pair of peptides **AP02565** and **AP02566** (pair **1** in Figure 2) have a 100% of AA sequence identity and are plotted close to each other. Note, however, that the pair of peptides do not have exactly the same coordinates since the distance that is measured in a TMAP does not depend on the % identity of AA but on an alternative representation that depends on structural fingerprints (*vide supra*). In contrast, the pair of peptides **AP02565** and **AP02567** (pair **2** in Figure 2) are structurally different e.g., 69% of AA sequence identity and 0.43 fingerprint-based similarity, and are plotted farthest apart as compared to the peptide pair **1**.

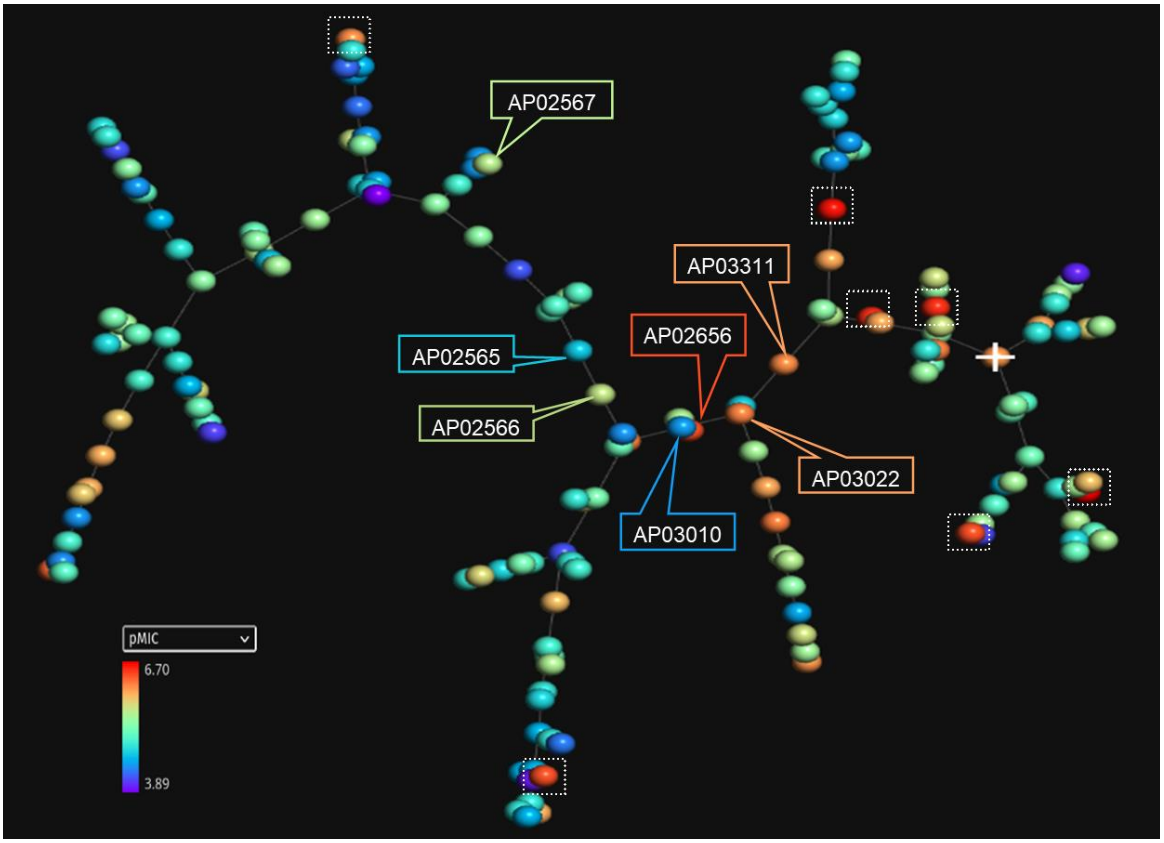


Figure 3. TMAP of 165 anti-MRSA peptides studied in this study. Each sphere represents a peptide, and the distance between each sphere represents the structural relationships between them. Each sphere was colored using a scale of red (higher pMIC$_{50}$ values; higher anti-MRSA activity) to blue (lower pMIC$_{50}$ values; lower anti-MRSA activity). Dotted squares indicate the most active anti-MRSA peptides.

Interestingly, the peptide pairs **1** - **5** have a medium-to-high structural similarity (AA sequence identity between 52% - 100%) but are associated with a large change in their pMIC$_{50}$ values. However, the fingerprint-based similarity values (as measured with MAP4) do have positive correlations with identity values (R^2 = 0.66, Figure S3 in the Supplementary material). This observation suggests that fingerprint-based similarity measures complement the insights derived from sequence alignments, but do not replace them. Namely, it is possible to use the small structural/sequence changes in peptides to help rationalize the P-SPR.

Figure 3 illustrates the most active peptides reported with anti-MRSA activity (e.g., **AP00505**, **AP02816**, **AP01320**) that could be used as starting points to design novel active peptide derivates. In addition, existing peptides that are constantly inactive could be used to decode the features related to the inactivity against MRSAs [12].

Overview of anti-MRSA peptides sequence alignments

Alignment analysis of the 165 anti-MRSA peptides resulted in a consensus sequence (“FLKKLLKLLGKLL” - Figure S2 in the Supplementary material) which is characterized by being rich in Leucines (Leu - L) and Lysines (Lys - K). The consensus sequence has a net charge of +4 and a 62% hydrophobic ratio [31]. The consensus sequence has been related to high anti-MRSA activity. However, it also has been associated with hemolytic properties due to the presence of the “LKL”, “KKLL” “LKKL” motifs [31–35].

We also study the sequence alignment of the twenty most active peptides within our dataset shows a consensus sequence ("GCRANKGACRAKKCKSHGGRGGKCF" - Figure S4 in the Supplementary material), having a high net charge of +8.25 [36–38] and 32% hydrophobic ratio. We remark on the presence of four cysteines (Cys - C) in the peptide sequence, which contributes to the stability of the tertiary structure [39]. The presence of four cysteines is also observed in anti-MRSA peptidic sequences, for example in peptides **AP03010**, **AP03022**, and **AP03311** in Figure 4 (*vide infra*).

In summary, these results suggest that the structural similarity calculations based on MAP4 fingerprint and MinHashed function help to explore the activity landscape of peptides. Methods such as extended SAS maps and SALI enable the landscape study of 165 anti-MRSA peptides rapidly uncovering small changes in structures associated with large modifications in the pMIC$_{50}$ values. Though this methodology is general and could be adapted to study any other properties of peptides, i.e., P-SPR. We remark that TMAPs are useful to visualize different features of the properties landscape of peptides (and other small molecules). However, we need to consider that TMAPs, as other visualization methods, highly depend on the structural representation (e.g., molecular fingerprint), and it is sensitive to the relative size of peptides that are compared. Thus, we recommend using only TMAP visualization when the peptide size is comparable.

Discussion

P-SPR is a re-emerged and important topic, now it is possible to synthesize them on a large scale and generate analog series from an interesting initial sequence. But, the synthetic methods have been updated more rapidly than the peptide design area. Although different efforts to reduce this methodological gap have been developed, it remains challenging to fully decode the P-SPR since it can be highly complex in nature. Interestingly, peptides could resolve a large list of clinical, biological, chemical, pharmacological, alimentary, and agrochemical current issues [3,40–44]. For this reason, it is important to use novel approaches to quantify and understand, as much as possible, their SPR.

Activity and property landscapes have been extensively studied, however, most of the studies are focused on small organic molecules. This fact could be explained by the lack of a proper fingerprint to represent peptides. However, bioinformatics approaches allow the identification of activity cliffs using uniquely the sequence alignment of peptides. Parallelly, in the last years, chemoinformatics approaches (e.g. molecular similarity metrics) based on topologies, connectivity, tridimensional features, and molecular properties offer a new alternative to study in depth more complex molecules [45], like peptides. Additionally, a previous study using different fingerprints (e.g. MACCs keys, ECFP4, ECFP6, and atom pairs) allowed the construction of peptide landscapes but using unique peptides with the same number of amino acids [46]. In contrast, this work shows an application of methods typically used in chemoinformatics to study small organic molecules to study the P-SPR using the concept of activity (property) landscapes.

The anti-MRSA peptide landscape explored in this work (Figure 2-A) indicates a total of 20 (~0.16%) peptide pairs in quadrant **I** (scaffold or R- hopping peptides); 9204 (~68.8%) in quadrant **II** (smooth SAR peptides); 4150 (~31.02%) in quadrant **III** (peptide activity cliffs, like pairs **2** - **5**); and 2 (~0.02%) in quadrant **IV** (peptides without a with a discontinuous SAR, like pair **1**). Namely, most of a third part of the peptide pairs have considered activity cliffs, which could limit the model ability of these data sets to be used to develop a predictive model of anti-MRSA activity.

We point out that it is possible to establish a direct SAR based on pairwise comparisons. For instance, in the peptide pair **1** (Figure 2), the terminal phenylalanine (Phe / F) could be associated with their biological activity difference. This is in agreement with X-ray diffraction studies that indicate that the terminal Phe on peptide structures improves the stability of helical conformation [47]. Furthermore, He et al.

confirm that the activity of antimicrobial peptides depends on the stability of their helical structure [48]. Therefore, the protocol presented here to describe the activity landscape of the 165 anti-MRSA peptides was capable of identifying small peptide differences involved in their activity.

Another key example that reflects the impact of one unique AA change on the peptide structure/sequence is the peptide activity cliff **2** (**AP02565** - **AP02567**, Figure 2), which suggests a key role of asparagine (Asn - N). Their tridimensional model (as generated with PEP-FOLD) (Figure 4) reflects the impact of this AA change on the stability of the helical peptide structures. Additionally, quantum methods confirm this observation and remark on the importance of Asn on peptide reactivity [49].

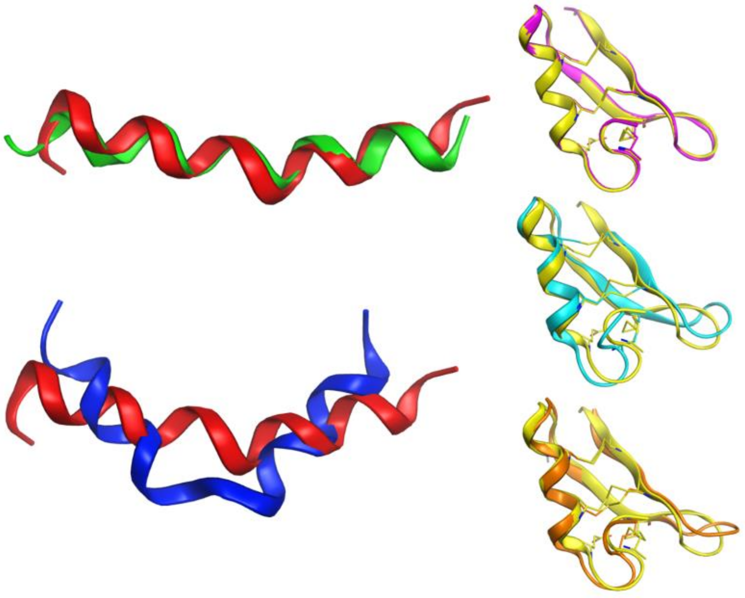


Figure 4. Conformational differences between selected peptides studied in this work. Each peptide is represented with a different color; red (**AP02565**), green (**AP02566**), blue (**AP02567**), yellow (**AP03010**), cyan (**AP03022**), and orange (**AP03311**). The tridimensional representation of each peptide was modeled by PEP-FOLD [50].

Although the predicted tridimensional structures of the peptides forming activity cliffs are similar (pairs **3** (**AP03010** - **AP02656**); **4** (**AP03010** - **AP03022**); and **5** (**AP03010** - **AP03311** in Figures 2 and 4), their values of volume of accessible surface area are different which suggest changes in their solubility (Table S2 in the Supplementary material). Such differences could be associated with changes in their biological activity [26]. Additionally, the differences between the cationic area [36,38,51,52] (involved in the membrane interaction on MRSA strains) of each peptide pair could be associated with their variations in biological activity (Table S2 in the Supplementary material).

These results indicate the dependency of the activity cliffs with the descriptors used to quantify the similarity between pairs of peptides [53]. For example, if one uses the volume of the accessible surface area instead of the MAP4 fingerprint as a descriptor, then the peptide pairs **2** - **5** would no longer be considered activity/property cliffs. Namely, these results indicate that the anti-MRSA activity does not depend uniquely on the peptide sequence and the features encoded on MAP4 fingerprints. The anti-MRSA activity also depends on other criteria, like the tridimensional similarity and the physicochemical properties. We remark that the selection of the molecular representations is a crucial step to decoding correctly the P-SPR.

During the past five years, the concept of SPR has been adapted to design and develop novel peptidic entities. The concept of P-SPR has been used to design and develop lipopeptides, and cyclic peptides [54,55], and to decode the membranolytic mechanism of different kinds of peptides [56]. However, there are complex challenges to resolve towards consolidating the *in silico* peptide design area [13,40–43,57]. Principally, the limited access to quality data and the balance of active and inactive reports are difficult the generation of new information and knowledge. However, methods that prioritize the selection of the most representative structure could resolve (almost in part) this issue. Additionally, the implementation of the "Structure-Sequence-Property relationships" concept on peptides is a key step forward to exploiting the potential of peptide data. Besides, the biological issues (i.e., immunogenicity, proteolytic degradation, permeability, and toxicity) have been shallowly explored.

Current methodologies that have been used to study P-SPR have limitations, and the activity landscape approximation presented in this work is not an exception. The fingerprint-based similarity (using MAP4 and the MinHassed distance) is a new method to explore and describe the landscape of any property in peptides. However, the results of this study suggest that this methodology could be high-sensitive to structural changes on peptides with less than 20 AA, which could limit their applicability, and remarks on the importance to develop new molecular representations focused on peptides. For this reason, we recommend using multiple criteria and methodologies to understand the P-SPR: activity land. complementing this method using classical alignment sequence analysis, and 3D approximation to help understand the P-SPR. In this regard, the present work contributes to establishing a helpful workflow based on structure similarity metrics to explore P-SPRs and quickly identify peptide activity cliffs.

Conclusions

This work presents a new method to explore and describe the landscape of any property of peptides based on the MAP4/MinHashed distance function. We construct and discuss the activity landscape of 165 anti-MRSA peptides. For the case study, it was concluded that: 1) the fingerprint-based similarity values (as measured with MAP4/MinHashed distance function) have a positive correlation with the sequence-based identity values (R^2 = 0.66) suggesting that fingerprint-based similarity measures complement the insights derived from sequence alignments, but do not replace them; 2) around 31% of paired anti-MRSA peptides were consider activity cliffs. These findings point out the challenge of this data set to develop predictive models.

As part of this work, we introduced the extended SAS map (using molecular weight differences values of each peptide pair) that facilitated the rapid identification of peptide activity cliffs. The fingerprint-based similarity using MAP4 is a good addition to starting a new peptide design/development campaign. However, as basically any in silico approach, each one has advantages and limitations. Therefore, activity landscape analysis should be in combination with classical sequence alignment, and physicochemical descriptors to explore in detail the SPR in peptides.

The main perspective of this work is to apply fingerprint-based similarity calculations to develop consensus virtual screening protocols (i.e., based on 2D and 3D structure similarity, chemical properties similarity, and sequence identity) to identify peptide structures with specific properties and apply the methodology present in this work to curate peptide datasets previously to use to develop artificial intelligent methods to predict peptide's properties.

Supplementary material

Figure S1. Descriptive analysis of the 165 anti-MRSA peptides studied in this work; **Figure S2**. Alignment analysis of the 165 anti-MRSA peptides studied in this work; **Figure S3**. Correlations of identity values and fingerprint-based similarity values; **Figure S4**. Alignment analysis of the 20 most potent anti-MRSA peptides; **Table S1**. Anti-MRSA peptides dataset; **Table S2**. Physicochemical properties of representative anti-MRSA peptides.

References

1 A. A. Vinogradov, Y. Yin and H. Suga, *J. Am. Chem. Soc.*, 2019, **141**, 4167–4181.

2 A. Isidro-Llobet, M. N. Kenworthy, S. Mukherjee, M. E. Kopach, K. Wegner, F. Gallou, A. G. Smith and F. Roschangar, *J. Org. Chem.*, 2019, **84**, 4615–4628.

3 V. Apostolopoulos, J. Bojarska, T.-T. Chai, S. Elnagdy, K. Kaczmarek, J. Matsoukas, R. New, K. Parang, O. P. Lopez, H. Parhiz, C. O. Perera, M. Pickholz, M. Remko, M. Saviano, M. Skwarczynski, Y. Tang, W. M. Wolf, T. Yoshiya, J. Zabrocki, P. Zielenkiewicz, M. AlKhazindar, V. Barriga, K. Kelaidonis, E. M. Sarasia and I. Toth, *Molecules*, 2021, **26**, 430.

4 P. Bhadra, J. Yan, J. Li, S. Fong and S. W. I. Siu, *Sci. Rep.*, 2018, **8**, 1697.

5 S. Decker, A. Taschauer, E. Geppl, V. Pirhofer, M. Schauer, S. Pöschl, F. Kopp, L. Richter, G. F. Ecker, H. Sami and M. Ogris, *Eur. J. Pharm. Biopharm.*, 2022, **176**, 211–221.

6 F. D. Prieto-Martínez, E. López-López, K. Eurídice Juárez-Mercado and J. L. Medina-Franco, in *In Silico Drug Design*, ed. K. Roy, Academic Press, 2019, pp. 19–44.

7 M. Leutert, S. W. Entwisle and J. Villén, *Mol. Cell. Proteomics*, , DOI:10.1016/j.mcpro.2021.100129.

8 F. I. Saldívar-González, V. D. Aldas-Bulos, J. L. Medina-Franco and F. Plisson, *Chem. Sci.*, 2022, **13**, 1526–1546.

9 P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen, H. Song, Y.-D. Cai and K.-C. Chou, *PLOS ONE*, 2011, **6**, e18476.

10 E. López-López, E. Fernández-de Gortari and J. L. Medina-Franco, *Drug Discov. Today*, 2022, **27**, 2353–2362.

11 B. I. Díaz-Eufracio, O. Palomino-Hernández, A. Arredondo-Sánchez and J. L. Medina-Franco, *Mol. Inform.*, 2020, **39**, 2000035.

12 C. D. Fjell, R. E. W. Hancock and A. Cherkasov, *Bioinformatics*, 2007, **23**, 1148–1155.

13 D. Kanduc, *J. Pept. Sci.*, 2012, **18**, 487–494.

14 B. Mishra and G. Wang, *J. Am. Chem. Soc.*, 2012, **134**, 12426–12429.

15 P. Charoenkwan, S. Kanthawong, N. Schaduangrat, P. Li', M. A. Moni and W. Shoombuatong, *ACS Omega*, 2022, **7**, 32653–32664.

16 R. P. Sheridan, S. B. Singh, E. M. Fluder and S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1395–1406.

17 E. López-López, J. Bajorath and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2021, **61**, 26–35.

18 F. Plisson, O. Ramírez-Sánchez and C. Martínez-Hernández, *Sci. Rep.*, 2020, **10**, 16581.

19 E. López-López, O. Rabal, J. Oyarzabal and J. L. Medina-Franco, *J. Comput. Aided Mol. Des.*, 2020, **34**, 659–669.

20 D. Erzina, A. Capecchi, S. Javor and J.-L. Reymond, *Angew. Chem. Int. Ed Engl.*, 2021, **60**, 26403–26408.

21 V. C. S. R. Chittepu, P. Kalhotra, T. Osorio-Gallardo, C. Jiménez-Martínez, R. R. Robles-de la

Torre, T. Gallardo-Velazquez and G. Osorio-Revilla, *Molecules*, 2019, **24**, 3887.

22 M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro and F. Borges, *Drug Discov. Today*, 2014, **19**, 1069–1080.

23 J. Bajorath, A. L. Chávez-Hernández, M. Duran-Frigola, E. Fernández-de Gortari, J. Gasteiger, E. López-López, G. M. Maggiora, J. L. Medina-Franco, O. Méndez-Lucio, J. Mestres, R. A. Miranda-Quintana, T. I. Oprea, F. Plisson, F. D. Prieto-Martínez, R. Rodríguez-Pérez, P. Rondón-Villarreal, F. I. Saldívar-Gonzalez, N. Sánchez-Cruz and M. Valli, *J. Cheminformatics*, 2022, **14**, 82.

24 F. Chaib, B. John and S. Hwang, Antimicrobial resistance crisis, https://www.who.int/news/item/29-04-2019-new-report-calls-for-urgent-action-to-avert-antimicrobial-resistance-crisis, (accessed May 23, 2022).

25 G. Wang, X. Li and Z. Wang, *Nucleic Acids Res.*, 2016, **44**, D1087–D1093.

26 J. L. Medina-Franco, J. J. Naveja and E. López-López, *Drug Discov. Today*, 2019, **24**, 2162–2169.

27 A. Capecchi, D. Probst and J.-L. Reymond, *J. Cheminformatics*, 2020, **12**, 43.

28 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

29 R. Guha and J. H. Van Drie, *J. Chem. Inf. Model.*, 2008, **48**, 646–658.

30 D. Probst and J.-L. Reymond, *J. Cheminformatics*, 2020, **12**, 12.

31 G. Wang, X. Li and Z. Wang, *Nucleic Acids Res.*, 2009, **37**, D933–D937.

32 K. Chaudhary, R. Kumar, S. Singh, A. Tuknait, A. Gautam, D. Mathur, P. Anand, G. C. Varshney and G. P. S. Raghava, *Sci. Rep.*, 2016, **6**, 22843.

33 S. L. Haynie, G. A. Crum and B. A. Doele, *Antimicrob. Agents Chemother.*, 1995, **39**, 301–307.

34 C. Wisdom, S. K. VanOosten, K. W. Boone, D. Khvostenko, P. M. Arnold, M. L. Snead and C. Tamerler, *J. Mol. Eng. Mater.*, 2016, **4**, 1640005.

35 S.-J. Kang, H.-S. Won, W.-S. Choi and B.-J. Lee, *J. Pept. Sci. Off. Publ. Eur. Pept. Soc.*, 2009, **15**, 583–588.

36 B. Mishra, J. Lakshmaiah Narayana, T. Lushnikova, X. Wang and G. Wang, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 13517–13522.

37 Z. Jiang, A. I. Vasil, J. D. Hale, R. E. W. Hancock, M. L. Vasil and R. S. Hodges, *Pept. Sci.*, 2008, **90**, 369–383.

38 J. J. López Cascales, S. Zenak, J. García de la Torre, O. G. Lezama, A. Garro and R. D. Enriz, *ACS Omega*, 2018, **3**, 5390–5398.

39 A. Rodriguez, M. Ø. Pedersen, E. Villegas, B. Rivas-Santiago, J. Villegas-Moreno, C. Amero, R. S. Norton and G. Corzo, *Proteins Struct. Funct. Bioinforma.*, 2020, **88**, 175–186.

40 M. Muttenthaler, G. F. King, D. J. Adams and P. F. Alewood, *Nat. Rev. Drug Discov.*, 2021, **20**, 309–325.

41 A. Levin, T. A. Hakala, L. Schnaider, G. J. L. Bernardes, E. Gazit and T. P. J. Knowles, *Nat. Rev. Chem.*, 2020, **4**, 615–634.

42 R. J. Malonis, J. R. Lai and O. Vergnolle, *Chem. Rev.*, 2020, **120**, 3210–3229.

43 S. Mondal, S. Das and A. K. Nandi, *Soft Matter*, 2020, **16**, 1404–1454.

44 D. Bhandari, S. Rafiq, Y. Gat, P. Gat, R. Waghmare and V. Kumar, *Int. J. Pept. Res. Ther.*, 2020, **26**, 139–150.

45 D. Stumpfe, D. Dimova and J. Bajorath, *J. Chem. Inf. Model.*, 2014, **54**, 451–461.

46 B. I. Díaz-Eufracio, O. Palomino-Hernández, R. A. Houghten and J. L. Medina-Franco, *Mol. Divers.*, 2018, **22**, 259–267.

47 S. Aravinda, N. Shamala, C. Das, A. Sriranjini, I. L. Karle and P. Balaram, *J. Am. Chem. Soc.*, 2003, **125**, 5308–5315.

48 Y. He and T. Lazaridis, *PLOS ONE*, 2013, **8**, e66440.

49 C. Soriano-Correa, C. Barrientos-Salcedo, L. Campos-Fernández, A. Alvarado-Salazar and R. O. Esquivel, *Chem. Phys.*, 2015, **457**, 180–187.

50 A. Lamiable, P. Thévenet, J. Rey, M. Vavrusa, P. Derreumaux and P. Tufféry, *Nucleic Acids Res.*, 2016, **44**, W449–W454.

51 M. Mihajlovic and T. Lazaridis, *Biochim. Biophys. Acta*, 2012, **1818**, 1274–1283.

52 M. Bacalum and M. Radu, *Int. J. Pept. Res. Ther.*, 2015, **1**, 47–55.

53 J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2013, **81**, 553–556.

54 S. E. Jujjavarapu and S. Dhagat, *Probiotics Antimicrob. Proteins*, 2018, **10**, 129–141.

55 S. Wang, K. Krummenacher, G. A. Landrum, B. D. Sellers, P. Di Lello, S. J. Robinson, B. Martin, J. K. Holden, J. Y. K. Tom, A. C. Murthy, N. Popovych and S. Riniker, *J. Chem. Inf. Model.*, 2022, **62**, 472–485.

56 T. Rončević, D. Vukičević, L. Krce, M. Benincasa, I. Aviani, A. Maravić and A. Tossi, *Biochim. Biophys. Acta BBA - Biomembr.*, 2019, **1861**, 827–834.

57 S. R. Eddy, *Nat. Biotechnol.*, 2004, **22**, 1035–1036.