picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

PandoraGAN: Generating antiviral peptides using Generative Adversarial Network

Shraddha Surana [1], Pooja Arora [1,*], Divye Singh [1], Deepti Sahasrabuddhe [2] and Jayaraman Valadi [3]

[1] Engineering for Research, ThoughtWorks Technologies, Pune, 411006, India,
[2] Department of Chemical Engineering, Indian Institute of Technology, Bombay, 400076, India
[3] Department of Computer Science, Flame University, Pune, 411007, India.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The continuous increase in pathogenic viruses and the intensive laboratory research for development of novel antiviral therapies often poses challenge in terms of cost and time efficient drug design. This accelerates research for alternate drug candidates and contributes to recent rise in research of antiviral peptides against many of the viruses. With limited information regarding these peptides and their activity, modifying the existing peptide backbone or developing a novel peptide is very time consuming and a tedious process. Advanced deep learning approaches such as generative adversarial networks (GAN) can be helpful for wet lab scientist to screen potential antiviral candidates of interest and expedite the initial stage of peptide drug development. To our knowledge this is the first ever use of GAN models for antiviral peptides across the viral spectrum.

Results: In this study, we develop PandoraGAN that utilizes GAN to design bio active antiviral peptides. Available antiviral peptide data was manually curated for preparing highly active peptides data set to include peptides with lower IC50 values. We further validated the generated sequences comparing the physico-chemical properties of generated antiviral peptides with manually curated highly active training data.

Availability: Antiviral sequences generated by PandoraGAN are available on PandoraGAN server. https://pandora-gan.herokuapp.com/ . The code is available at https://gitlab.com/shraddha.surana/antiviral-peptide-predictions-using-gan

Contact: parora@thoughtworks.com

1 Introduction

Since few decades, antiviral peptides have emerged to be an attractive candidate against variety of viruses including RNA viruses like HIV, MERS etc (Vilas Boas *et al.*, 2019). Despite multiple reports on isolation and characterization of these compounds from various natural and synthetic sources, their utility in main stream therapeutics is highly limited due to their diverse structural and sequence cohort, lack of understanding of motifs and elusive mechanism of action. Use of successful peptide therapy is often obstructed by their lower stability and poor availability (Mahendran *et al.*, 2020) (Agarwal and Gabrani, 2020). With emergence of novel viral strain, discovering an effective peptide against specific virus requires longer time commitment compared to the other drug therapies. This is specially due to limited knowledge-base and lack of structure-activity or sequence-activity correlation. There are only couple of databases such as AVPDB (Qureshi *et al.*, 2014), APD3 (Wang *et al.*, 2016) listing the antiviral peptide information along with their efficacy from the literature. Many antimicrobial peptides, which are essential action force in innate immunity (Diamond *et al.*, 2009), may possess antiviral properties which are yet to be explored. Wide spectrum of viral hosts diversify the potential action of the peptide. This makes further rational alteration in the targets of such peptides even more challenging (Di Natale *et al.*, 2020). Without predefined structural patterns and defined physico-chemical properties preferences makes the designing and

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

prediction of the antiviral potential of the peptide very difficult. These pose serious bottlenecks in utilization of computational methods to predict peptide's properties. Different methods of calculating the potency of these compounds is another serious drawbacks in comparing the available information.

The recent viral pandemic has made the entire world fall on its knees. While the efforts for making vaccine are on the forefront, discovery of novel drug therapies is imperative to keep the disease at bay. However, the success rates for discovering a new chemical entity and efficacy of the repurposed drugs against viruses has been a major concern. This clearly portrays the dire need of finding alternate therapies for viral diseases. Advances in deep learning altered the realm of data driven biological problems. Computational advances could be further exploited to onset the research in antiviral peptide domain. During the COVID 19 times the Indian Government launched a drug discovery hackathon[1] in July 2020 to develop computational methods for COVID 19 drug discovery and extrapolate these algorithms to other generic drug discovery challenges. In this study we picked one of the problems statements which required to generate bioactive antiviral peptides using generative adversarial network.

2 Literature Survey

GANs have been previously used in biological domain to generate the novel protein sequences from a single family, biological staining images etc. Recently, PepGAN (Tucs *et al.*, 2020) has been used to generate antimicrobial peptides showing very high activity.

Killoran *et al.*, 2017 proposed a Wasserstein GAN (WGAN) based architecture for generating DNA and tune them to have desired properties. They presented three approaches: generating synthetic DNA sequences using a GAN; a DNA-based variant of the activation maximization design method; and a joint procedure which combines these two approaches together. Another report which focuses on generating synthetic DNA using GAN-based architecture is Feedback-GAN (Gupta and Zou, 2018) which implemented function analyzer to score the generated sequence and pass selected n highest scoring sequences to the discriminator's training dataset as "real" samples, replacing n oldest sequences in the "real" samples. GANs have also been used in the genomics space for example, Yelmen *et al.*, 2019 have used GANs and restricted Boltzmann machines to learn the high dimensional distributions of real genomics datasets and create high quality artificial genomes. Repecka *et al.*, 2019 introduced ProteinGAN for generating protein sequences. Their GAN architecture used ResNet blocks in both discriminator and generator model along a single self-attention layer in both. Anand and Huang, 2018 have used GAN to generate protein structures to expedite protein design.

3 Approach

In this study, we utilise the GAN's potential to generate novel antiviral peptides and compared them with the training dataset following certain physico-chemical properties such as hydrophobicity, hydrophilicity, net charge to name a few (Chang and Yang, 2013; Beltrán Lissabet *et al.*, 2019). The available non-redundant antiviral peptides databases were cumulatively studied to define certain acceptable range of selected physico-chemical properties based on sequences. These rules were later compared with the generated sequences to assess the performance of GAN. In order to generate bio active antiviral peptides, it is important to train it using bio active sequences. Our GAN was trained using highly active antiviral peptides based on their IC50/activity values. We utilised manually curated antiviral database including few peptides in clinical trials, as an input in GAN for higher confidence in the generated data. This study is pioneering in the field of antiviral peptides.

This study can broadly be summarised as:

- Collating anti-viral peptide data
- Training a suitable GAN
- Validating the generated peptides computationally

4 Materials & Methods

In this section, we elaborate on the data we have used, curation of the dataset, fitering process and details of PandoraGAN.

4.1 Collating anti-viral peptide data

We built our dataset utilizing existing experimentally validated databases like AVPDb (Qureshi *et al.*, 2014), AVPpred Benchmark data (Thakur *et al.*, 2012), CAMP (Thomas *et al.*, 2010), DRAMP (Kang *et al.*, 2019), APD3 (Wang *et al.*, 2016), DBAMP (Jhong *et al.*, 2018) as well as from the literature. From CAMP, DbAMP, DRAMP and APD3 we extracted the experimentally validated peptides with antiviral affinity quantified. We combined all the data and excluded the sequences which were more than 90 percent redundant to avoid sequence bias in the training data using CDHIT[2]. Sequences with unnatural and modified amino acids(B, J, O, U, X and Z) were removed as it is out of the designed study in this phase. After removing the redundant sequences we were left with AVPDb sequences. In order to train GAN efficiently we further selected sequences having lower IC50 values and created two datasets as active and highly active sequences based on their IC50 values.

4.1.1 Curation of active and highly active sequences

To generate efficient dataset as input for the GAN, antiviral potential was considered to be the key criteria. The sequences having an IC50 value less than 30 micro molar were considered as active sequences. The sequences having IC50 values below 15 micro molar were considered to be highly active sequences (Qureshi *et al.*, 2015).

Active sequences: We selected the peptide sequences having IC50 values less that 30 micro molar as active sequences creating the dataset of 423 active sequences.

Highly active sequences(HA-AVP): The same data is further curated manually based on the peptide's activity/IC50 values as well as information from the literature. Differential reporting of activity values limit the computational automated curation of the dataset based on the given range. The peptide concentrations showing viral inhibition were converted into uniform unit and the peptides in clinical trials as per current literature were included in the database. Peptides with the activity in the range of 0-15 μM were combined to form a subset of highly active peptides. Subsequently, the sequences with similar identity of more than 90% were removed to ensure equivalent representation of all the peptide classes. Such dataset of 130 peptides is further utilized while training GAN. The peptides taken from literature and clinical trials are listed in following references - HIV candidates (Liang *et al.*, 2020; Shi *et al.*, 2016; Chupradit *et al.*, 2017; Coffey *et al.*, 1997); SARS COV (Xia *et al.*, 2020); MERS (Zhao *et al.*, 2016);T20 (Ding *et al.*, 2017) ;Hepatitis B (Lempp *et al.*, 2016);

We train our GAN using 2 sets of data viz. highly active and combination of active and highly active antiviral peptides. We find that training the GAN on only highly active peptides gives better results.

[1] https://innovateindia.mygov.in/ddh2020/

[2] http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

4.2 Training a suitable GAN

In this section we discuss different types of GANs we used in our study and details about the final variant of GAN which gave the best results which is LeakGAN.

4.2.1 Generative Adversarial Models

Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) are deep learning based generative models which learn the underlying data distribution using gradient descent without a requirement of prior knowledge of the structure of the data. With generative and discriminator models working in parallel, GAN iterates between generating the data and testing its equivalence to the input data till the generated data does not appear to be unreal. The GAN model architecture involves two sub-models:

- **Generator model:** It produces samples that mimic the actual data distribution. It starts with generating random samples and improves over time by the feedback it receives from the discriminator module.
- **Discriminator model:** Discriminator distinguishes between samples generated by the generator module versus those belonging to the actual data. It takes an example (real or generated) as input and predicts a binary class label of real or fake (generated). The real example comes from the actual data on which the GAN is trained and the fake examples are output by the generator model.

The GAN is trained as a two player zero-sum (adversarial) game where the generator model and the discriminator model are considered as the two players from a game theory perspective. In this case, zero-sum means that each of the two players get rewarded when other player does not perform well. This leads to the formulation of minimax loss function on which the GAN is trained and this loss function is defined as

$$E_x[log(D(x))] + E_x[log(1 - D(G(z)))] \quad (1)$$

where $G(.)$ and $D(.)$ are generator and discriminator models respectively, x is an input example to the discriminator model and z is an input random vector for the generator model.

In order to generate peptides with variable lengths, we use a RNN or LSTM in the generator model - the same concept used for generating variable length textual sentences.

In case of generating text, GAN faces a limitation that the classical discriminative model for the discriminator can only assess a complete sequence. For a partially generated sequence, it is non-trivial to balance the current score and the future score once the entire sequence has been generated. Another problem that is common in text generation using GANs (adversarial modeling) is that the binary signal from the discriminator is not sufficiently informative. This may cause the generator to be inadequately trained and could result in mode collapse problem. Generally, a huge number of training and generated samples are required to improve the generator which becomes very challenging with limited number of experimentally validated AVPs available. To overcome these limitations there are some modified versions of GAN widely used in text generation such as - MaliGAN (Che *et al.*, 2017), SeqGAN (Yu *et al.*, 2017), LeakGAN (Guo *et al.*, 2017), RankGAN (Lin *et al.*, 2017), TextGAN (Zhang *et al.*, 2017) and GSGAN (Kusner and Hernández-Lobato, 2016).

We have modified the implementations made available by Texygen [3] which is a text generation model to generate antiviral peptides. In this work, we have implemented and validated the antiviral peptides generated with all the above mentioned variants of GAN. Of these, LeakGAN gives us promising results and hence we have used it in PandoraGAN. This is described in more detail in the next section.

[3] https://github.com/geek-ai/Texygen

4.3 LeakGAN

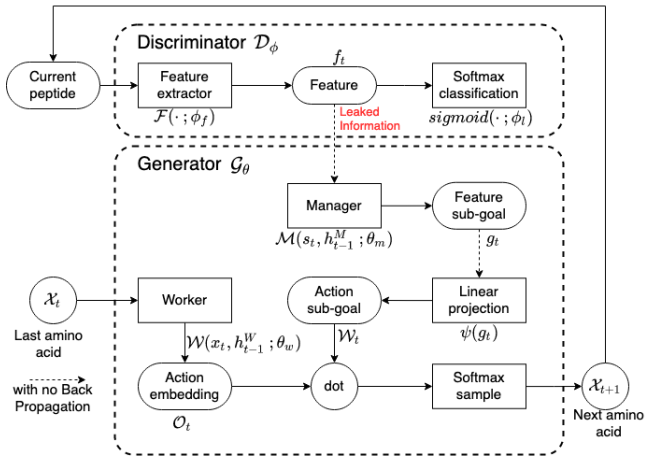

Fig. 1. An overview of the LeakGAN framework. The generator is responsible to generate the next amino acid, while the discriminator judges the generated sequence once it is complete. Unlike conventional adverserial training, the discriminator reveals it's internal state during the training process in order to guide the generator more informatively.

LeakGAN (Guo *et al.*, 2017) is a GAN architecture that takes recent advancements in hierarchical reinforcement learning to provide a richer information from the discriminator to the generator. The LeakGAN introduces a hierarchical generator G, which consists of a high-level `Manager` module and a low-level `Worker` module. This helps the architecture to overcome the problem of sparse binary guiding signal which the discriminator provides the generator to learn upon.

In every time step, the `Manager` module receives a high-level feature representation from the discriminator which is used to form the guiding goal for the `Worker` module in that time step. In an adversarial learning, the information within the discriminator and generator models are to be internally maintained. However, in the case of LeakGAN, high-level feature information is passed from discriminator model to generator. This information is said to be the leaked information.

The information flow in LeakGAN can be seen in Figure 1 where at each step, `Manager` receives the leaked information f_t i.e. the high level features of the input peptides, from the discriminator $\mathcal{D}_\phi$. This is used to form the goal vector g_t. To this goal vector a linear transformation ψ is performed to give a k-dimensional goal embedding vector $\mathcal{W}_t$. On the other hand, `Worker` takes the current amino acid residue, $\mathcal{X}_t$ as input and output a matrix $\mathcal{O}_t$ that represents the current vector for all amino acids. This matrix is combined with the goal embedding vector $\mathcal{W}_t$ to yield the final action space distribution under current state through a softmax.

Since the amount of antiviral data we have is very less, this action of leaking the learnt features from the discriminator to the generator is most likely the reason for LeakGAN generating the best set of possible antiviral peptides. PandoraGAN was trained on a manually curated highly active peptides as described earlier. With this training set, the underlying LeakGAN is able to learn a good representation of the implicit properties of anti viral peptides. It is able to generate sequences with similar properties as shown in the evaluation graphs.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

5 Results

Using the manually curated dataset of highly active peptides as training data, we implemented different GANs such as LeakGAN, MaliGAN, RankGAN, TextGAN, GSGAN and SeqGAN to generate novel antiviral peptides. The peptides generated by LeakGAN were most promising and closest to the highly active antiviral peptides. The top five peptides generated by LeakGAN are mentioned in Table 1 sorted by their molecular weight which was one of the evaluation criteria in the drug discovery hackathon.

Table 1. Antiviral peptides sequences generated by PandoraGAN

S.no	Sequence	Molecular weight	Net charge at pH7.4
1	ALLEKSKK	916.12	1.60
2	VSLVKAALKE	1057.28	0.52
3	LLDFKLSDAK	1149.33	-0.44
4	GAYTFKIRRK	1239.46	3.55
5	FKRAKKWVFR	1365.67	4.55

While training for these variety of GANs, we observed that training for a large number of iterations leads to overfitting and generation of peptides with many biased consecutive repetitions of a particular amino acid patterns. This was specifically the case for Text GAN and GSGAN. SeqGAN, RankGAN and MaliGAN intermittenly generated good sequences but the results were not consistent. Training these GANs further lead to overfitting. Although LeakGAN gave the best results, it took the longest time to train. The LeakGAN training is approximately 400 times longer compared to other GANs. All GANs were trained on a machine with 2.2GHz 6-Core Intel Core i7 processor and 2400MHz 32GB DDR4 memory.

We did not provide any rules based on specific physico-chemical or structural properties as features of the peptides to GAN, instead allowed GAN to learn from the sequences alone. The generated peptides were compared with the highly active peptides training dataset to validate the learnings of GAN. We utilised the data driven rules for physico-chemical properties and statistical inferences to assess similarity of the GAN's output to the training data.

5.1 Validation strategy

Although little elusive, characteristic physico-chemical properties of antiviral peptides have been reported in literature (Kang *et al.*, 2012). Various prediction programs use combination of properties like length, charge, hydrophilicity distribution to classify these peptides. Most of the antiviral peptides are of short length, high in cationic residues and have a spatial distribution of hydrophobic and hydrophillic residues etc. They also exhibit evolutionary conservation of short motifs which are mostly di, tri or tetra peptide long. These characteristics might be beneficial for exhibiting antiviral activity and reducing the viral load eventually. We utilized these physico-chemical properties and statistical methods to validate the performance of PandoraGAN by comparing it with the highly active peptides. The validation strategy is shown in Figure 2. The ranges of physico-chemical properties used in this study are mentioned in table 2. The parameters utilized in validation strategy were calculated using the python library Biopython (Cock *et al.*, 2009) and pfeature webserver (Pande *et al.*, 2019).

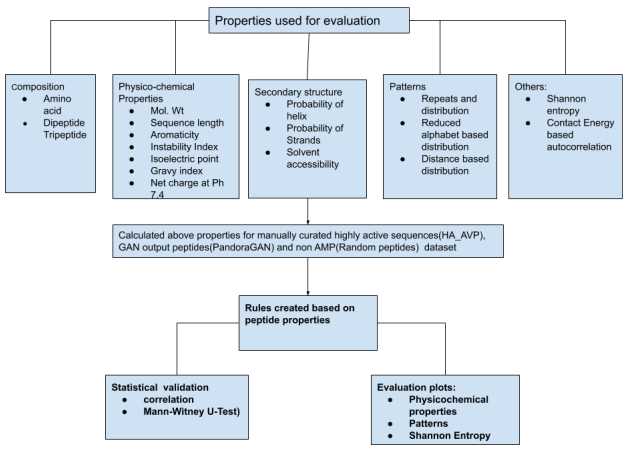

Fig. 2. Validation strategy for comparing the peptides using the various properties of highly active peptides.

5.1.1 Validation based on Physico-chemical properties, repeats and patterns

For evaluating efficiency of PandoraGAN's learning capabilities, we focus on few physico-chemical properties and compare the generated sequences with the highly active peptides as explained in the validation strategy in Figure 2.

Below we elaborate the various plots utilized to study the comparison of highly active antiviral peptides(HA-AVP) which were curated manually, PandoraGAN peptides and Random non-secretory peptides.

Table 2. Validation Strategy Rules

Properties	Rules	Values
Composition	Residues in more than 70% sequences	KR
	Residues in more than 80% sequences	I, L, V
	Residues in less than 20% sequences	M, N
	Residues in around 50% sequences	W, P
Physico-chemical	Sequence length	7 to 30
	Net charge at Ph7.4	-1 to 6
	GRAVY index	> -1
Secondary structure	Probability of Helix	> 0.35
	Probability of Strand	< 0.3
Repeats and patterns	Dipeptide repeats	FK, KK, LL, VV
	Tripeptide repeats	KKK, KKL, KKV, KKR
	Shannon Entropy of residue	Higher shannon entropy in case of AVP for M, N, P, Y as opposed to non-AVP
	Residue repeats	lower repeats for E, M, N, P as compared to non-AVP
Statistical Validation	Correlation of Shannon entropy	high correlation
	Mann-Witney U-Test on physio-chemical properties	p-value > 0.05

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

- PandoraGAN was able to identify the appropriate composition of amino acids for antiviral peptides. To exemplify our observation, AVP's are low on Methionine and comparatively high on Lysine and Arginine (Chang and Yang, 2013) which is also observed in PandoraGAN generated sequences as seen in Figure 3.
- We observed similar trend of net charge at pH 7.4, (-1 to +4) in the generated peptide as in the highly active training sequences.

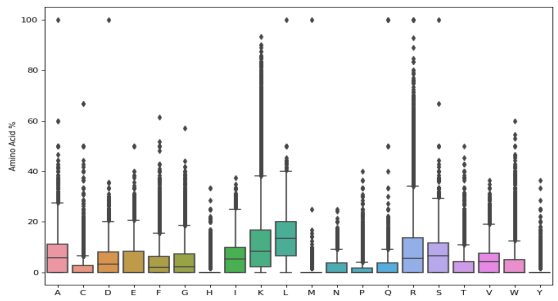

Fig. 3. Composition of amino acids int he peptides generated by PandoraGAN

- Similar to highly active AVPs, the peptides generated by PandoraGAN also have a higher propensity of alpha helices as compared to sheets as shown in Figure 4 and Figure 5 respectively.

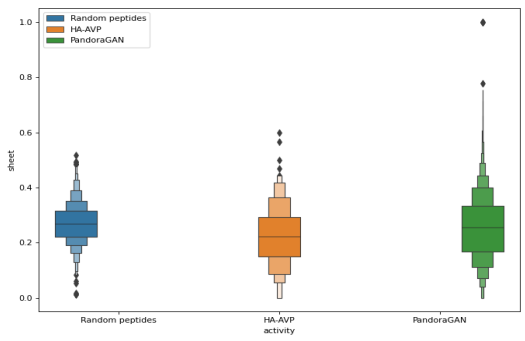

Fig. 4. Sheet forming propensity

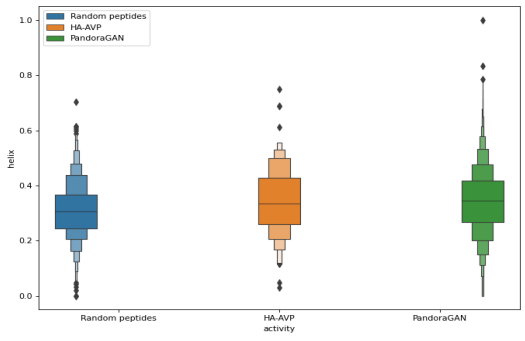

Fig. 5. Helix forming propensity

- Antimicrobial peptides are expected to have lower instability index which decreases their chances of in-vivo survival. PandoraGAN generated the peptides in similar range to highly active peptides as in Figure 6.

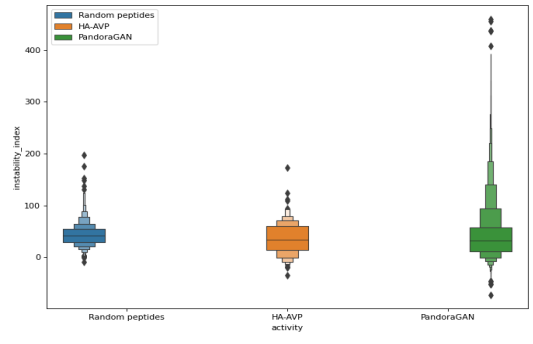

Fig. 6. instability Index

- Short amino acid motifs can be crucial for antiviral peptides' activity. We identified certain abundant di and tri peptide repeats from the input dataset which could be important criteria for assessment. We plotted Shannon entropy and residue repeats and it was observed that the entropy of conserved residues in highly active peptides and PandoraGAN is equivalent as shown in Figure 7 and Figure 8

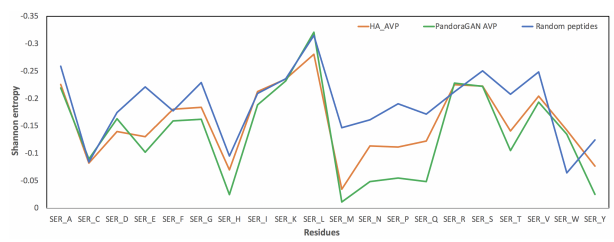

Fig. 7. Shannon entropy of all the residues

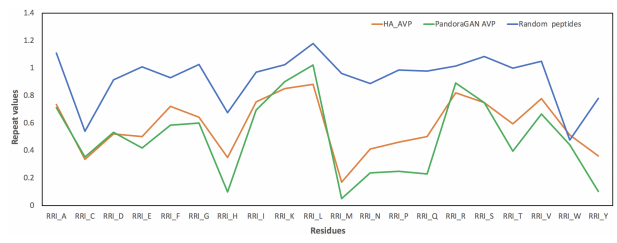

Fig. 8. Residue repeats patterns for generated sequences as compared to highly active input sequences.

These observations confirm the appropriate learning of GAN as the PandoraGAN generated sequences displaying various physico-chemical properties of antiviral peptides, conserved regions and patterns similar to HA-AVP data.

5.1.2 Statistical Inferences

The comparative analysis of PandoraGAN generated AVPs and the curated HA-AVPs, show that the peaks (high entropy) and valleys (low entropy)

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

for the shannon entropy of residues were highly correlated (Pearson's ρ = 0.96 with p-value < 1e-10). Further, Mann-Witney U-Test showed that aromaticity, instability index, isoelectric point, helix, turn, sheet, gravy and net charge at pH7.4 for PandoraGAN generated AVPs and HA-AVPs follow same distribution. These results show that the PandoraGAN has learned overall amino acid composition corresponding to physico-chemical properties of HA-AVPs. The properties are calculated using the biopython library.

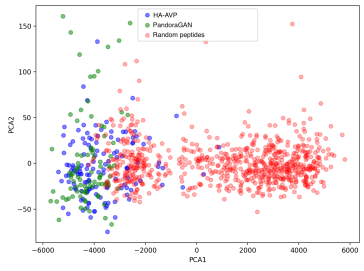

Fig. 9. Plot of PCA of the physio-chemical properties of highly active peptides, PandoraGAN generated sequences and random non-secretory peptides.

Further, the physico-chemical properties of highly active peptides, PandoraGAN generated sequences and random non-secretory peptides were projected onto first two principal components via principal component analysis. Figure 9 shows an overlap of HA-AVP peptides and PandoraGAN generated peptides while showing clear distinction between from random non-secretory peptides.

5.2 Cross-Validation of generated peptides using existing servers

We also validated the sequences generated by PandoraGAN by giving them as input to existing antiviral prediction servers like AVPpred (Thakur *et al.*, 2012) and MetaiAVP (Schaduangrat *et al.*, 2019). We observed that similar to HA-AVPs in the training data, 70% of the peptides generated by PandoraGAN were predicted as antiviral peptides by these servers with a probability above 0.9. This additionally confirms the pandoraGAN generated sequences to be AVPs based on the properties used by these external servers.

6 Discussion

The dearth in the knowledge of antiviral peptides limits potential application of these candidates as alternative therapeutics. Diversity of their host viral spectrum is another concern while generating a specific antiviral peptide. Studying an antiviral peptide in wet-lab setup starts with identification of viral family, determining the peptide backbone to work on, its source identification, bioinformatics analysis, multiple assays, mutational analysis etc. All these steps make the identification a novel peptide against specific virus family a laborious and time consuming process. Automation as well as computational advances were only partially successful in easing the load.

PandoraGAN gives a new outlook to the entire process by providing the first step of a peptide backbone to work on which can further be explored in the laboratory. From our observations, PandoraGAN produces sequences with similar physico-chemical properties to the HA-AVP training data and can be extremely agile depending on the dataset provided. One can also train PandoraGAN using their own training dataset. Computational generation of the peptides would also ensure exploring novel domains of peptides which were otherwise unexplored in the traditional approach.

In this study, PandoraGAN provides the generalised antiviral peptide candidates based on the training data active against the broader spectrum of viral hosts. The physico-chemical properties work as an initial validation for the first line of defense peptides. The amino acid composition, net charge, instability index, repeats and patterns were used to compare the PandoraGAN peptides to HA-AVP training data.

Here, we have explored the abilities of PandoraGAN to learn based on the given sequences and we have observed that PandoraGAN performs better when trained on highly active sequences vs active sequences or the sequences with unknown activity/IC50. Alongwith generating bioactive peptides another important aspect for the scientist is to know detailed properties of the generated sequences, understanding and providing that flexibility to filter and choose the required generated sequence which is another important aspect of PandoraGAN.

7 Conclusion and future work

We present PandoraGAN, a generative model for designing antiviral peptides and demonstrates it's ability to generate peptides which mimics the distribution of a very small dataset showing prominent characteristics. Physico-chemical properties of highly active peptides(training data) and generated peptides were analyzed and were found to be similar with high statistical significance. In our knowledge, this is the first time GANs have been used for generating antiviral peptides across the viral spectrum.

In future work, we plan to improve PandoraGAN to be able to use both highly active and also the active peptides to generate better distribution of residues in a peptide. For this, we will explore techniques that retains the quality of highly active peptides while exposing the GAN to higher number of peptides. We further plan to explore structural and evolutionary aspects of the peptide as reward functions for PandoraGAN to further tune its ability to generate bioactive peptides. This work can further be extrapolated towards identifying the activity of AVPs against specific viral families.

Conflict of interest statement

This study was conducted as part of the Drug discovery hackathon(DDH) 2020 organized by Govt. of India. The copyright and commercial aspects are to be followed as mentioned in DDH policies.

Acknowledgements

SS lead the development of the GAN architecture along with Divye S. PA conceptualised the problem and analysed the data. DS provided domain guidance as well as manual dataset curation and JV provided overall guidance.

References

Agarwal, G. and Gabrani, R. (2020). Antiviral Peptides: Identification and Validation. *Int J Pept Res Ther*, pages 1–20.

Anand, N. and Huang, P.-S. (2018). Generative modeling for protein structures. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7505–7516, Red Hook, NY, USA. Curran Associates Inc.

Beltrán Lissabet, J. F., Belén, L. H., and Farias, J. G. (2019). AntiVPP 1.0: A portable tool for prediction of antiviral peptides. *Comput Biol Med*, **107**, 127–130.

Chang, K. Y. and Yang, J.-R. (2013). Analysis and prediction of highly effective antiviral peptides based on random forests. *PloS one*, **8**(8), e70166.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Che, T., Li, Y., Zhang, R., Hjelm, R. D., Li, W., Song, Y., and Bengio, Y. (2017). Maximum-likelihood augmented discrete generative adversarial networks. *CoRR*, **abs/1702.07983**.

Chupradit, K., Moonmuang, S., Nangola, S., Kitidee, K., Yasamut, U., Mougel, M., and Tayapiwatana, C. (2017). Current peptide and protein candidates challenging hiv therapy beyond the vaccine era. *Viruses*, **9**.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Coffey, M. J., Woffendin, C., Phare, S. M., Strieter, R. M., and Markovitz, D. M. (1997). RANTES inhibits HIV-1 replication in human peripheral blood monocytes and alveolar macrophages. *Am J Physiol*, **272**(5 Pt 1), L1025–1029.

Di Natale, C., La Manna, S., De Benedictis, I., Brandi, P., and Marasco, D. (2020). Perspectives in peptide-based vaccination strategies for syndrome coronavirus 2 pandemic. *Frontiers in Pharmacology*, **11**, 1779.

Diamond, G., Beckloff, N., Weinberg, A., and Kisich, K. O. (2009). The roles of antimicrobial peptides in innate host defense. *Curr Pharm Des*, **15**(21), 2377–2392.

Ding, X., Zhang, X., Chong, H., Zhu, Y., Wei, H., Wu, X., He, J., Wang, X., and He, Y. (2017). Enfuvirtide (T20)-Based Lipopeptide Is a Potent HIV-1 Cell Fusion Inhibitor: Implications for Viral Entry and Inhibition. *J Virol*, **91**(18).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., and Wang, J. (2017). Long text generation via adversarial training with leaked information. *CoRR*, **abs/1709.08624**.

Gupta, A. and Zou, J. (2018). Feedback gan (fbgan) for dna: a novel feedback-loop architecture for optimizing protein functions. *ArXiv*, **abs/1804.01694**.

Jhong, J.-H., Chi, Y.-H., Li, W.-C., Lin, T.-H., Huang, K.-Y., and Lee, T.-Y. (2018). dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Research*, **47**(D1), D285–D297.

Kang, S. J., Kim, D. H., Mishig-Ochir, T., and Lee, B. J. (2012). Antimicrobial peptides: their physicochemical properties and therapeutic application. *Arch Pharm Res*, **35**(3), 409–413.

Kang, X., Dong, F., Shi, C., Liu, S., Sun, J., Chen, J., Li, H., Xu, H., Lao, X., and Zheng, H. (2019). Dramp 2.0, an updated data repository of antimicrobial peptides. *Scientific data*, **6**(1), 1–10.

Killoran, N., Lee, L. J., Delong, A., Duvenaud, D., and Frey, B. J. (2017). Generating and designing DNA with deep generative models. *CoRR*, **abs/1712.06148**.

Kusner, M. J. and Hernández-Lobato, J. M. (2016). Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*.

Lempp, F. A., Qu, B., Wang, Y.-X., and Urban, S. (2016). Hepatitis b virus infection of a mouse hepatic cell line reconstituted with human sodium taurocholate cotransporting polypeptide. *Journal of virology*, **90**(9), 4827–4831.

Liang, X., Zhang, X., Lian, K., Tian, X., Zhang, M., Wang, S., Chen, C., Nie, C., Pan, Y., Han, F., Wei, Z., and Zhang, W. (2020). Antiviral effects of Bovine antimicrobial peptide against TGEV in vivo and in vitro. *J Vet Sci*, **21**(5), e80.

Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-t. (2017). Adversarial ranking for language generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3155–3165. Curran Associates, Inc.

Mahendran, A. S. K., Lim, Y. S., Fang, C.-M., Loh, H.-S., and Le, C. F. (2020). The potential of antiviral peptides as covid-19 therapeutics. *Frontiers in Pharmacology*, **11**, 1475.

Pande, A., Patiyal, S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S., *et al.* (2019). Computing wide range of protein/peptide features from their sequence and structure. *bioRxiv*, page 599126.

Qureshi, A., Thakur, N., Tandon, H., and Kumar, M. (2014). AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res*, **42**(Database issue), D1147–1153.

Qureshi, A., Tandon, H., and Kumar, M. (2015). AVP-IC50 Pred: Multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC50). *Biopolymers*, **104**(6), 753–763.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Zrimec, J., Poviloniene, S., Rokaitis, I., Laurynenas, A., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M. K. M., and Zelezniak, A. (2019). Expanding functional protein sequence space using generative adversarial networks. *bioRxiv*.

Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., and Shoombuatong, W. (2019). Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. *Int J Mol Sci*, **20**(22).

Shi, S., Nguyen, P. K., Cabral, H. J., Diez-Barroso, R., Derry, P. J., Kanahara, S. M., and Kumar, V. A. (2016). Development of peptide inhibitors of hiv transmission. *Bioactive materials*, **1**(2), 109–121.

Thakur, N., Qureshi, A., and Kumar, M. (2012). AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res*, **40**(Web Server issue), 199–204.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., and Idicula-Thomas, S. (2010). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res*, **38**(Database issue), D774–780.

Tucs, A., Tran, D. P., Yumoto, A., Ito, Y., Uzawa, T., and Tsuda, K. (2020). Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks.

Vilas Boas, L. C. P., Campos, M. L., Berlanda, R. L. A., de Carvalho Neves, N., and Franco, O. L. (2019). Antiviral peptides as promising therapeutic drugs. *Cell Mol Life Sci*, **76**(18), 3525–3542.

Wang, G., Li, X., and Wang, Z. (2016). APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res*, **44**(D1), D1087–1093.

Xia, S., Liu, M., Wang, C., Xu, W., Lan, Q., Feng, S., Qi, F., Bao, L., Du, L., Liu, S., *et al.* (2020). Inhibition of sars-cov-2 (previously 2019-ncov) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell research*, **30**(4), 343–355.

Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., and Jay, F. (2019). Creating artificial human genomes using generative models. *bioRxiv*.

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient.

Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., and Carin, L. (2017). Adversarial feature matching for text generation. In *International Conference on Machine Learning*, pages 4006–4015. PMLR.

Zhao, H., Zhou, J., Zhang, K., Chu, H., Liu, D., Poon, V. K., Chan, C. C., Leung, H. C., Fai, N., Lin, Y. P., Zhang, A. J., Jin, D. Y., Yuen, K. Y., and Zheng, B. J. (2016). A novel peptide with potent and broad-spectrum antiviral activities against multiple respiratory viruses. *Sci Rep*, **6**, 22008.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture