



Available online freely at [www.isisn.org](http://www.isisn.org)

# Bioscience Research

Print ISSN: 1811-9506 Online ISSN: 2218-3973

Journal by Innovative Scientific Information & Services Network



RESEARCH ARTICLE

BIOSCIENCE RESEARCH, 2019 16(3): 3210-3216.

OPEN ACCESS

## Detecting noncrystallographic symmetry in Icosahedral Viruses using deep learning approach

Nora Abd El-Hameed Mohamed<sup>1</sup>, Mohammad Nassef<sup>1</sup>, Ahmed Farouk Al-Sadek<sup>2</sup> and Amr A. Badr<sup>1</sup>

<sup>1</sup>Department of Computer Science, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt.

<sup>2</sup>Agricultural Research Center, Giza, Egypt.

\*Correspondence: [n.abdelhameed@fci-cu.edu.eg](mailto:n.abdelhameed@fci-cu.edu.eg) Accepted: 06 Sep. 2019 Published online: 30 Sep 2019

Assembly of capsid virus is a crucial step in virus life cycle. Without this step, virus will not replicate itself to hijack other cells and its life cycle would end. Many researchers studied virus structural shape and its dynamics to understand the behavior of the virus. So, this paper focuses on the structural shape of Icosahedral viruses and prediction of symmetries in their capsids. A small virus capsid contains identical asymmetric units that are packed in regular manner. Every icosahedral virus has two types of symmetry, regular symmetry and non-crystallographic symmetry. So, one asymmetric unit and some rotation matrices are needed to form the whole capsid. These rotation matrices define the location of adjacent asymmetric unit. In this paper, deep learning approach is followed to create a layered model that predicts non-crystallographic symmetry in virus capsid. Through visualization technique, the results were promising; the accuracy was 89% for assembling the capsid in icosahedral viruses using dataset taken from the Protein Data Bank (PDB).

**Keywords:** Viral Capsids; Asymmetric Unit; Non-crystallographic Symmetry; Icosahedral Virus; Deep Learning

### INTRODUCTION

The world is full of dangerous viruses. In order to understand how they work. We need to study their structural shape and life cycle (Twarock, 2006). The virus consists of genetic material (RNA/DNA) that is enclosed within a protein shell (Capsid) (Lodish, et al., 1999).

The virus hijacks the cells and disassembles its capsid to let the genetic materials be translated by the host cell's machinery (Cann, 2005). Then, the genetic materials is replicated and the capsid is assembled around the new replicated genetic material (Pelczar, 1977). Therefore, it is very important to study the assembly and disassembly of the capsid structure.

Viruses capsids consist of multiple copies of repeating subunits (CRICK & WATSON, 1956). These subunits follow symmetries to form the

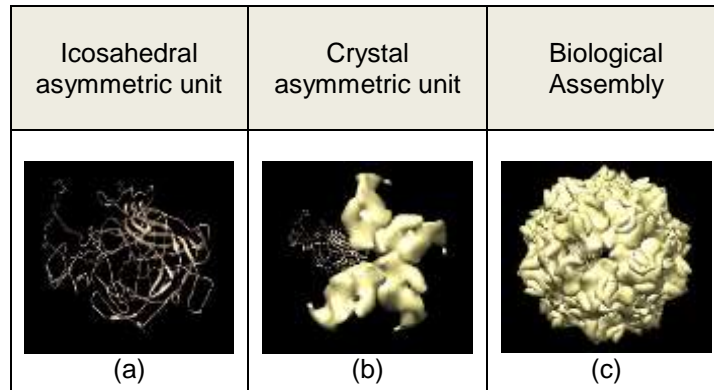
capsid virus.

There are two types of symmetry (Lawson, et al., 2008). The first type is a local symmetry known as Non-crystallographic symmetry (NCS). NCS is a transformation to construct crystal asymmetric unit (Fig. 1b) from these subunits which are asymmetric units (ASU) (Fig. 1a).

The second type is a global symmetry which is the transformation required to build a full capsid. These global symmetries construct the biological assembly (Fig. 1c) from the crystal asymmetric units (Fig. 1b).

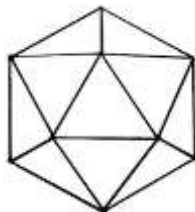
These symmetries can be understood by checking the periodicity pattern. If there is a periodicity, then these are global symmetries known as regular symmetries. Such symmetries are 2, 3, 4, 6 folds ...etc. If the periodicity found in higher dimensions, then these are local

symmetries, such as 5,7 folds ...etc.



**Figure 1 ; The biological assembly of a virus (1fpv) (Agbandje & Rossmann, 1994) is composed of crystal asymmetric units, which in turn is composed of icosahedral asymmetric (Berman, et al., 2003).**

For example, Icosahedral viruses consist of 20 triangular faces with twofold, threefold and fivefold axes of symmetry (Fig. 2). The twofold and threefold are regular symmetry while the fivefold is NCS symmetry. In this study, we will focus on Icosahedral viruses.



**Figure 2; 20-sided icosahedron face.**

Find NCS is a program was created to find non-crystallographic symmetry from heavy atoms in proteins in general. This program adopts a method for systematic search to find the best NCS. It takes into consideration the space group and ranks its results accordingly (Lu, 1999).

A model presented by (Lawson, et al., 2008) to remediate PDB archive studied the regular symmetry of the capsid. While NCS, that defines the crystal asymmetric unit, was considered as an input to the model.

Deep learning approach reached state-of-the-art in many fields. However, it was not adopted before in this problem. So, a model is built with deep neural network. The Model input is all atoms in ASU provided by PDB bank. The input focuses on the spatial geometries and biological characteristics of the virus. The Model output is the rotation NCS matrix of the crystal asymmetric

unit.

Our deep learning based model achieved a promising results. The model and a full explanation of the evaluation process will be explained in this paper.

## BACKGROUND

The capsid is constructed of ASUs which are similar to each other with the difference in their spatial geometries. Group of icosahedral asymmetric units following the Group Theory constructs crystal asymmetric unit (Fig. 1) (Senechal, 2009). Crystal asymmetric unit is considered in the proposed model as the first cap of the virus. A cap of an Icosahedral virus consists of five ASUs with one rotation matrix that defines the adjacent ASU.

The relationship between any ASU in cap and its adjacent ASU is represented with the following equation:

$$B = R * A \text{ (eq.1)}$$

$$C = R * B \text{ (eq.2)}$$

Where A is an ASU and B is the adjacent ASU of A. C is the adjacent ASU of B. R is the rotation matrix that defines this Crystal Asymmetric Unit.

Our deep learning model tries to predict the three angles in radius. Then, it constructs the rotation matrix while ignoring the translation vector. This model is trained on the first cap of the capsid where a translation vector does not affect the cap.

## MATERIALS AND METHODS

### Dataset preparation:

The dataset is extracted from RCSB PDB which is updated to the PDB archive each week. It can be accessed via ftp at ftp://ftp.wwpdb.org (Berman, et al., 2003). The size of the taken dataset is around 200 Icosahedral viruses. This dataset is divided into 70% training and 10% validation while testing dataset is 20%.

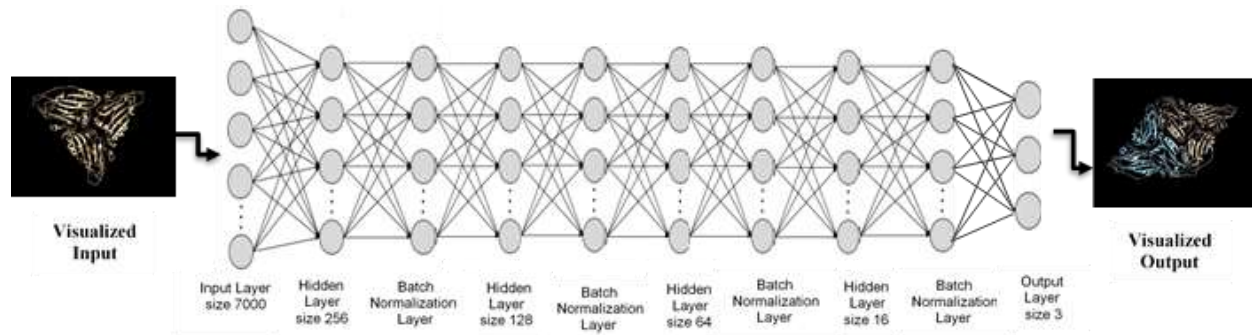
The training dataset is divided into input and output. The model input is fed by an ASU That consists of list of atoms. Spatial geometry and biological characteristics are features of an atom in ASU. To illustrate, an atom can have float points of (x, y, z, e); where x, y and z are the three value axes and e represents the biological characteristics. To handle input size inconsistency, we limited our atoms to 7000. This number represents the average number of atoms

in the dataset between viruses. Viruses with less than 7000 are padded with zero values in spatial geometries and biological characteristics.

The output found in PDB bank is a matrix of twelve float numbers. The last column in this matrix represents the translation vector and the rest of the matrix (3x3 matrix) represents a rotation matrix around the axes. The rotation matrix was converted to three angles around the axes. Such angles represent the output of our model. So every ASU input has three angles output. These angles represent the adjacent ASU in crystal asymmetric unit. Noting that viruses that have non-invertible matrix are excluded from the dataset.

### Model Architecture:

The proposed model is a fully connected neural network with five dense layers and four batch normalization layers (Fig. 3).



**Figure 2 ;The proposed model: a layered deep neural network.**

The activation function used is a rectified linear unit (ReLU), while the used optimizer is Stochastic Gradient Descent (SGD) with learning rate 0.01. The mean absolute error is used as the loss function.

The Input layer (L) is the ASU atoms. Its size is calculated as follows:

$$L = N * A \text{ (eq. 3)}$$

Where N represents number of atoms, and A represents the atom features. Atom features include spatial geometry and the biological characteristics.




The output layer is the predicted three-angle rotation around the X, Y and Z axes. To visualize the predicted output. A rotation matrix is constructed from these three angles with

translation vector (0,0,0,1). Such rotation matrix is applied on the input ASU to get the adjacent ASU.

### Evaluation:

This model has been tested by two criteria. The first criterion is subjective as it visualizes the results and compares them with the expected adjacent ASU position.

The calculation of the score for each test case is based on multiple error categories. Error categories consider the following errors: translation error (Fig. 4.a), rotation and translation error (Fig. 4.b), and minor error (Fig. 4.c). In order to calculate these errors we used two programs: Visual Molecule Dynamics (VMD) (Humphrey, et al., 1996) and UCSF Chimera (Pettersen, et al., 2004).

Translation error	Translation and rotation error	Minor error
		
(a)	(b)	(c)

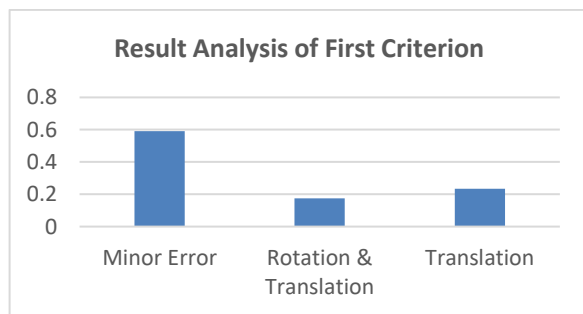
**Figure 3; Error Categories examples.** (a) Virus 1u1y where the predicted ASU (in blue) needs minor translation to completely overlap with the expected ASU (in brown) (Horn, et al., 2004). (B) Virus 2zi8 where the predicted ASU (in blue) has translation and rotation errors to overlap with the expected ASU (in brown) (Sabini, et al., 2008). (c) Virus 2cas where both the predicted and expected overlap, hence represents the perfect result (Wu & Rossmann, 1994).

The second criterion of evaluating the result is by visualizing how the model forms a crystal asymmetric unit. If the predicted crystal asymmetric unit has no gaps, and its ASUs do not overlap on each other as the predicted crystal asymmetric unit, then these are accepted outputs

**RESULTS AND DISCUSSION**

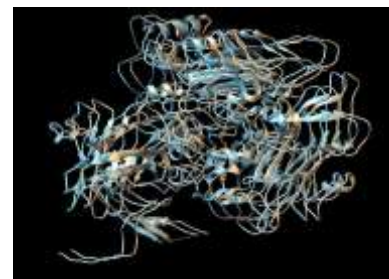
The results of the proposed model were tested against the two criteria explained earlier in the evaluation section. In the first criterion, the proposed model achieved 83% accuracy on the test data.

ASUs have the same rotation yet require translation to identically overlap the expected ASUs. Check the viruses in (Fig. 7) and (Fig. 8) as examples. Such translation error is evaluated based on Euclidian distance between center of the predicted ASUs and the center of expected ASUs. While 18% of the predicted ASUs failed to match both translation and rotation of the expected ASUs.



**Figure 5; The percentage of the viruses according to the error category in the first criterion.**

As per above (Fig. 5), we found that around 60% of the predicted ASUs have the same position and rotation as the expected adjacent ASUs which we consider a minor error. For example, virus 1ncq in PDB (Fig. 6) presents ideal results in this model. Also, 24% of the predicted



**Figure 6; An accepted case in virus with code 1ncq where the predicted ASU (in blue) is almost exactly as the expected ASU (in brown) (Chakravarty, et al., 2000).**

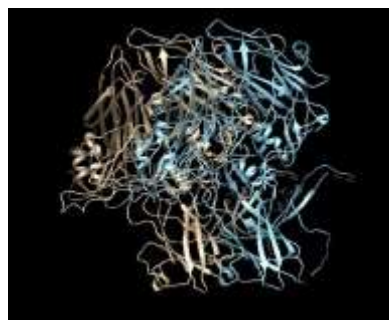


Figure 7; An accepted case in virus with code 1r09 where the predicted ASU (in blue) intersects with the expected ASU (in brown) by more than half of the ASU (Chapman, et al., 1991).



Figure 8; Virus with code 1c8m where the predicted ASU (Blue) is far from the expected ASU (Brown) (Chakravarty, et al., 2000).

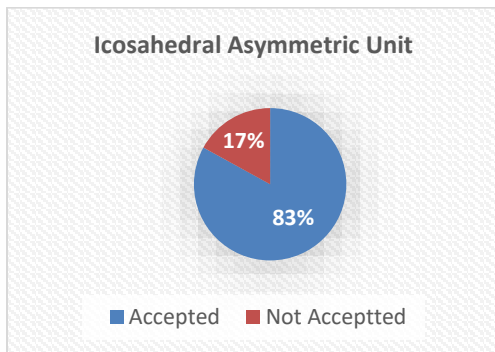


Figure 9; The percentage of the accepted ASUs in the testing dataset.

Since the translation vector is not counted as an error in this model, then such translation errors are accepted. Based on this fact, the accepted results according to Icosahedral Asymmetric units are 83% as shown in (Fig. 9).

In the second criterion, we visualized the results and found that some crystal asymmetric units position are the same as the original units (Fig. 10), and some crystal asymmetric units are constructed far from the original position due to ignoring the translation vector in this model (Fig. 11), however, it is accepted output due to forming no gaps. Accordingly, the accepted results are 89% as in (Fig. 12).

For future work, we aim to build a model that predicts the translation vector and the regular symmetry in order to construct the whole capsid using deep learning approach.

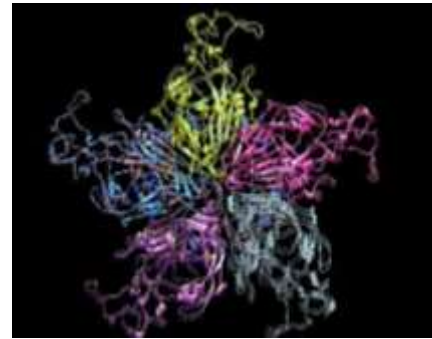


Figure 10; Virus with code 2rr1 where the expected crystal asymmetric unit position and the predicted crystal asymmetric unit position lays on top of the other (Badger, et al., 1990).

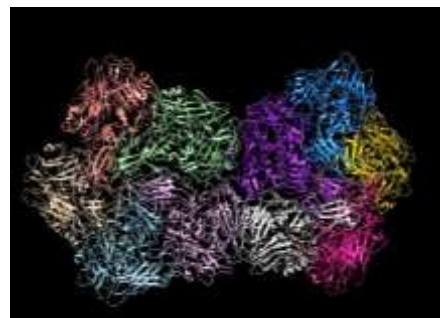


Figure 11; Virus with code 1mec where on the right is the expected crystal asymmetric unit position while the predicted crystal asymmetric unit position is on the left. and this error is accepted in the model (Rossmann, 1994).

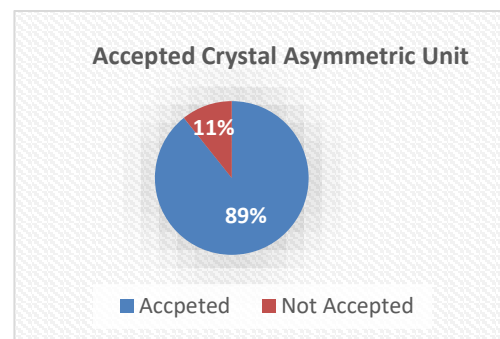


Figure 12; The percentage of the accepted crystal asymmetric unit in the proposed model.

**CONCLUSION**

Virus capsid protects its genatic material.

Capsid assembles and disassembles to coat/protect or release the genetic material for replication process in the host cell. Understanding capsid assembly and disassembly is a vital for breaking the virus replication cycle. Various models are built for studying the assembly of the virus capsid such as mathematical or simulation models. Yet few models found in computer science. This paper proposes a deep neural network model for virus capsid. This model predicts noncrystallographic symmetry found in capsid virus. Results were promising as it reached 89%. Also, results have been verified visually using VMD and UCSF Chimera. This model maybe a step in developing a neural network that predicts the full capsid with its two types of symmetry (noncrystallographic and regular symmetries).

### CONFLICT OF INTEREST

The authors declared that present study was performed in absence of any conflict of interest.

### AUTHOR CONTRIBUTIONS

NAM and AFA designed this model. NAM tested the model and also wrote the manuscript. MN and AAB contributed to the writing, amending and approving of manuscript. All authors read and approved the final version.

---

### Copyrights: © 2019 @ author (s).

This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and source are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

---

### REFERENCES

- Agbandje, m. & rossmann, m. G., 1994. Structure determination of feline panleukopenia virus empty particles. S.I.:protein data bank, rutgers university.
- Badger, j., smith, t. J. & rossmann, m. G., 1990. Structural analysis of antiviral agents that interact with the capsid of human rhinoviruses. s.I.:Protein Data Bank, Rutgers University.
- Berk, A. et al., 1999. Molecular Cell Biology. s.I.:W H Freeman & Co (Sd).
- Berman, H., Henrick, K. & Nakamura, H., 2003. Announcing the worldwide Protein Data Bank. Nature Structural & Molecular Biology, 12, Volume 10, pp. 980-980.
- Berman, H., Henrick, K., Nakamura, H. & Markley, J. L., 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Research, 1, Volume 35, pp. D301--D303.
- Bourne, C. R. & Zlotnick, A., 2006. Human hepatitis B virus T=4 capsid strain adyw complexed with assembly effector HAP1. s.I.:Protein Data Bank, Rutgers University.
- Cann, A. J., 2005. Principles of Molecular Virology (Standard Edition) (Cann, Principles of Molecular Virology). s.I.:Academic Press.
- Caspar, D. L. D. & Klug, A., 1962. Physical Principles in the Construction of Regular Viruses. Cold Spring Harbor Symposia on Quantitative Biology, 1, Volume 27, pp. 1-24.
- Chakravarty, S. et al., 2000. Refined crystal structure of human rhinovirus 16 complexed with vp63843 (pleconaril), an anti-picornaviral drug currently in clinical trials. s.I.:Protein Data Bank, Rutgers University.
- Chapman, M. S. et al., 1991. Human rhinovirus 14 complexed with antiviral compound r 61837. s.I.:Protein Data Bank, Rutgers University.
- CRICK, F. H. C. & WATSON, J. D., 1956. Structure of Small Viruses. Nature, 3, Volume 177, pp. 473-475.
- Fermi, G. & Perutz, M. F., 1984. The crystal structure of human deoxyhaemoglobin at 1.74 angstroms resolution. S.I.:protein data bank, rutgers university.
- Fry, e. E. Et al., 1999. Foot-and-mouth disease virus/ oligosaccharide receptor complex.. S.I.:protein data bank, rutgers university.
- Hagan, m. F., 2014. Modeling viral capsid assembly. In: advances in chemical physics. S.I.:john wiley & sons, inc., pp. 1-68.
- Horn, w. T. Et al., 2004. Crystal structure of a complex between wt bacteriophage ms2 coat protein and an f5 aptamer rna stemloop with 2aminopurine substituted at the-10 position. S.I.:protein data bank, rutgers university.
- Humphrey, w., dalke, a. & schulten, k., 1996. Vmd: visual molecular dynamics. Journal of molecular graphics, 2, volume 14, pp. 33-38.
- Lawson, c. L. Et al., 2008. Representation of viruses in the remediated pdb archive. Acta crystallographica section d biological crystallography, 7, volume 64, pp. 874-882.
- Lodish, h. Et al., 1999. Molecular cell biology. S.I.:w. H. Freeman.
- Lu, g., 1999. : a program to detect non-

- crystallographic symmetries in protein crystals from heavy-atom sites. *Journal of applied crystallography*, 4, volume 32, pp. 365-368.
- Nespolo, m., souvignier, b. & litvin, d. B., 2008. About the concept and definition of noncrystallographic symmetry. *Zeitschrift für kristallographie*, 1.volume 223.
- Pelczar, m. J., 1977. *Microbiology*. S.I.:mcgraw-hill.
- Pettersen, e. F. Et al., 2004. Ucsf chimera? a visualization system for exploratory research and analysis. *Journal of computational chemistry*, volume 25, pp. 1605-1612.
- Rossmann, m. G., 1994. Conformational variability of a picornavirus capsid: ph-dependent structural changes of mengo virus related to its host receptor attachment site and disassembly. S.I.:protein data bank, rutgers university.
- Sabini, e., hazra, s., konrad, m. & lavie, a., 2008. C4s-e247a dck variant of dck in complex with cladribineadp. S.I.:protein data bank, rutgers university.
- Senechal, m., 2009. *Quasicrystals and geometry*. S.I.:cambridge university press.
- Twarock, r., 2006. *Mathematical virology: a novel approach to the structure and assembly of viruses*. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, 10, volume 364, pp. 3357-3373.
- Wu, h. & rossmann, m., 1994. The canine parvovirus empty capsid structure. S.I.:protein data bank, rutgers university.