Study of Contribution of Chemical and Physical Properties of Molecules towards Their Activities against Ebola Virus Using Data Mining Techniques

Rungsang Nakrumpai

Abstract—**Ebola virus disease is a highly fatal hemorrhagic disease. To date, there is no approved drug to cure Ebola virus disease. Thus there is still the need to search for an effective drug. This study used various data mining techniques to investigate the relationship of chemical and physical properties of chemical molecules towards their activities against Ebola virus. The results suggest that these properties contribute to their activities against Ebola virus. Thus they could be used together with bioinformatics and cheminformatics techniques to help speed up the drug discovery process of Ebola virus disease.**

Index Terms—**Biological activity, chemical and physical properties, data mining, Ebola virus.**

I. Introduction

Ebola virus disease is a fatal hemorrhagic fever disease with high mortality rate of 70% to 90% [1], [2]. It is caused by Ebola virus belonging to Filoviridae family. The family consists of three genera: Ebola virus, Marburgvirus, and Cuevavirus [2], [3]. There are the total of five species of Ebola viruses: Zaire Ebola virus, Bundibugyo Ebola virus, Sudan Ebola virus, Tai forest Ebola virus (formerly referred to as Côte d'Ivoire Ebola virus) and Reston Ebola virus [1], [2]. The latest Ebola virus disease outbreak in Africa was caused by Zaire Ebola virus, the most virulent Ebola virus [2], [3].

Ebola virus has a filamentous single-stranded RNA genome. The genome is non-segmented and of negative sense. The size of the genome is 19 kilobases [2]. It encodes the following proteins: nucleoprotein (NP), polymerase cofactor VP35 and VP40, glycoprotein (GP), transcription activator VP30 and VP24, and RNA-dependent RNA polymerase (L). These proteins are vital to Ebola virus life cycle [2].

Ebola virus membrane is spiked with glycoproteins responsible for the attachment and entry of the virus to its target host cell [4]. The glycoprotein activation that enables cellular entry of the virus is done by its host cell cysteine proteases cathepsin B and L [5]. A secreted form of Ebola virus glycoprotein is also found though its role in Ebola virus pathogenesis is still unknown [4].

Ebola virus can infect many cell types. This broad cell tropism may due to the fact that Ebola virus cell entry involves several cell surface molecules including the receptor-type tyrosine kinases (RTKs), calcium-dependent lectins, and β1 integrin [6].

VP30 acts as the transcription initiator [7]. Polymerase cofactor VP35 and the viral RNA polymerase L are involved in the synthesis of Ebola virus new genome [7]. Ebola viral protein VP24 acts as a type I interferon (IFN) antagonists and thus it is also an important virulence factor [7]. VP40 is a major matrix protein of Ebola virus. It plays a vital role in the budding of Ebola virus from plasma membrane of the host cell [8], [9].

Healthy people can be infected with Ebola virus if come into contact with body fluids from infected patients such as blood, saliva, sweat, urine, semen, vomit, mucus, vaginal fluids, including feces [1], [3], [10]. Incubation time of the disease ranges from 2 days to 3 weeks [10]. Symptoms include sudden onset of fever, chills, myalgia, and malaise. Then nasal discharge, cough, breath shortness, nausea, vomiting, diarrhea, and abdominal pain would follow [11], [12]. In severe cases, hemorrhagic symptoms would then occur.

Although the latest Ebola virus disease epidemic largely occurred in West-Africa, there were many cases of the disease reported in other continents including Europe and North America [13]. These patients contracted the disease in Africa before they went to other continents.

Reverse transcriptase polymerase chain reaction (RT-PCR) technique can be used to confirm the infection of Ebola virus by detecting the presence of Ebola virus RNA [14], [15]. Serologic tests for detecting host antibodies against the virus can also be used [15]. Certain tests to detect Ebola virus antigen proteins can also be used to ascertain the virus infection [15].

It was reported that Ebola virus RNA still persisted in many patients long after their recovery [14]. It has also been suggested that Ebola virus RNA could persist for a period of time even after the dead of its host organism [16].

According to World Health Organization (WHO), currently there is still no approved vaccine or drug to cure Ebola virus disease. Thus the search for a drug that can effectively cure Ebola virus disease is still needed. Without any effective cure or prevention, future outbreak of Ebola virus would still incur severe problems to the infected countries as already encountered in the latest outbreak [3], [11]. The resulted unavoidable fatalities, weaken healthcare systems, and economic disruptions would inevitably interrupt the development of the infected countries.

Manuscript received March 3, 2017; revised May 7, 2017. This work was funded by Naresuan University.

R. Nakrumpai is with the Department of Biochemistry, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand (e-mail: rungsangn@nu.ac.th).

R. Nakrumpai is also with the Centre of Excellent in Medical Biotechnology, Faculty of Medical Science, Naresuan University, Phitsanulok, Thailand (e-mail: rungsangn@nu.ac.th).

doi: 10.18178/ijmlc.2017.7.3.617

PubChem is a public chemical database provided by the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH) [17]. It contains chemical structures of small molecules and their biological activities in three of its interconnected sub-databases [18]. Related web services are also provided to facilitate PubChem data access [19].

In PubChem, chemical molecules can be searched using related keywords. Comprehensive information related to the searched molecules would subsequently be provided. These information include PubChemCID, molecular formula, molecular weight, 2D structure, chemical vendors, related literatures, as well as chemical and physical properties of the molecules [17], [18], [20]. Three-dimensional conformers of most chemical molecules are also available [21]. PubChem data have been used in various researches including Quantitative Structure–Activity relationship (QSAR) studies [21], [22].

PubChem BioAssay is one of PubChem sub-databases [23], [24]. It contains information regarding biological activities of small molecules. Currently, more than a million records are available [25]. Entries are cross-linked to their chemical structures as well as related literatures. A simple keyword search as well as an advanced search are available. Data can also be downloaded in various formats including XML, ASN, JSON, and CSV. The database also contains data from high-throughput screening (HTS) useful for drug discovery process [26], [27].

II. Materials and Methods

A. Retrieving the Chemical Molecules

The list of chemical molecules that have been reported about their activities against Ebola virus was obtained from PubChem BioAssay. Some of these chemical molecules were FDA-approved drugs used for other medical purposes. Their designated activities could be active or inactive against Ebola virus. For example, some active molecules have been shown to block the entry of Ebola virus while inactive molecules have been reported to have no effect towards Ebola virus. As each molecule could come from different vendors and tests, thus only unique molecules with reported activities towards Ebola virus (active or inactive) were used. This resulted in 2,541 molecules being used in this study. Conversions of chemical structural formats were done using Open Babel [28]. Molecular visualization was done using Pc3D Viewer from PubChem [29].

B. Extracting the Physical and Chemical Properties

Physical and chemical properties of a chemical molecule influence its function and activity [24], [25]. Thus chemical and physical properties of the studied chemical molecules were obtained and used as data mining attributes in this study. These properties were: compound complexity, number of heavy atoms, molecular weight, charge, number of chiral atom, number of chiral bond, number of isotope atom, number of covalent unit, number of hydrogen bond acceptor, number of hydrogen bond donor, number of rotatable bond, polar surface area, exact mass, monoisotopic mass, and number of tautomer.

C. Data Mining

The effect of these properties towards the activities of molecules against Ebola virus was studied using three data mining techniques: Logistic Regression, Random Forest, and Na ïve Bayes. Three cross validation methods were used for each data mining technique: *k*-fold, leave one out, and repeated random subsampling. For *k*-fold cross validation, the number of *k* was set to 10 (this cross validation will subsequently be referred to as 10-fold cross validation). For repeated random subsampling cross validation, also called Monte Carlo cross validation, the size of training set was 70% of the total samples and the resampling was repeated ten times.

Exact mass and monoisotopic mass of a chemical molecule could be considered as similar to molecular weight of the molecule. These two properties were thus removed and data mining analyses were then repeated using the three data mining techniques and the three cross validation methods previously described.

D. Software and Tools

The software used for data mining was Orange3 [30]. Data preparation and analysis were done using Microsoft Excel. Notepad++ was used for inspecting and viewing data. Computer programming was done using Python 3.4 and Microsoft Visual Basic for Applications (VBA).

III. Results and Discussion

Table I to Table III show the results of employing the three data mining techniques using all the chemical and physical properties of the molecules. But Table IV to Table VI show the results of employing the three techniques with two properties removed, exact mass and monoisotopic mass. The results obtained with the removed properties were slightly worse than those with all properties used. However, the differences were seemingly quite small. Thus when necessary, these two properties of chemical molecules could be omitted out of the analysis.

TABLE I: Results Obtained from Employing Repeated Random Subsampling with Various Data Mining Techniques. The Training Size Was 70% of the Total Samples. The Resampling Was Repeated ten Times. All the Chemical and Physical Properties of the Molecules Were Used.

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.788	0.795	0.773
Recall	0.812	0.813	0.716
AUC	0.621	0.663	0.677
CA	0.812	0.813	0.716
F1	0.889	0.886	0.805

Among the three techniques used, Random Forest gave the best precision and recall values for all cases, as shown in Table I to Table VI. Different cross validation methods seemed to have minimal effects on the performances of the three techniques in all analyses. Although Na ïve Bayes method seemed to give the worst results among all techniques employed, the actual differences in the results were quite small. In overall, the precision values for all techniques were good and almost reached 80%. The recall percentages were slightly varied. Na ïve Bayes also gave the worst recall values among all the techniques used in all cases. Classification accuracy (CA) scores and F1 scores were quite high for all methods. Receiver Operating Characteristic (ROC) analysis with the target class as active against Ebola virus is shown in Fig. 1. ROC analysis with the target class as inactive against Ebola virus is displayed in Fig. 2.

TABLE II: RESULTS OBTAINED FROM EMPLOYING LEAVE ONE OUT WITH VARIOUS DATA MINING TECHNIQUES. ALL THE CHEMICAL AND PHYSICAL PROPERTIES OF THE MOLECULES WERE USED

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.783	0.788	0.758
Recall	0.808	0.807	0.692
AUC	0.614	0.655	0.655
CA	0.808	0.807	0.692
F1	0.886	0.882	0.786

TABLE III: RESULTS OBTAINED FROM EMPLOYING 10-FOLD CROSS VALIDATION WITH VARIOUS DATA MINING TECHNIQUES. ALL THE CHEMICAL AND PHYSICAL PROPERTIES OF THE MOLECULES WERE USED

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.789	0.796	0.758
Recall	0.812	0.814	0.689
AUC	0.619	0.662	0.654
CA	0.812	0.814	0.689
F1	0.889	0.887	0.784

TABLE IV: RESULTS OBTAINED FROM EMPLOYING REPEATED RANDOM SUBSAMPLING WITH VARIOUS DATA MINING TECHNIQUES. THE TRAINING SIZE WAS 70% OF THE TOTAL SAMPLES. THE RESAMPLING WAS REPEATED TEN TIMES. EXACT MASS AND ISOTOPIC WEIGHT WERE NOT USED

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.785	0.796	0.774
Recall	0.810	0.815	0.787
AUC	0.613	0.656	0.651
CA	0.810	0.815	0.787
F1	0.888	0.888	0.868

TABLE V: RESULTS OBTAINED FROM EMPLOYING LEAVE ONE OUT WITH VARIOUS DATA MINING TECHNIQUES. EXACT MASS AND ISOTOPIC WEIGHT WERE NOT USED

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.779	0.801	0.769
Recall	0.806	0.818	0.781
AUC	0.608	0.666	0.646
CA	0.806	0.818	0.781
F1	0.885	0.89	0.864

TABLE VI: RESULTS OBTAINED FROM EMPLOYING 10-FOLD CROSS VALIDATION WITH VARIOUS DATA MINING TECHNIQUES. EXACT MASS AND ISOTOPIC WEIGHT WERE NOT USED

Values of	Data Mining Techniques		
	Logistic Regression	Random Forest	Na ïve Bayes
Precision	0.781	0.797	0.766
Recall	0.807	0.815	0.778
AUC	0.609	0.665	0.641
CA	0.807	0.815	0.778
F1	0.886	0.887	0.862

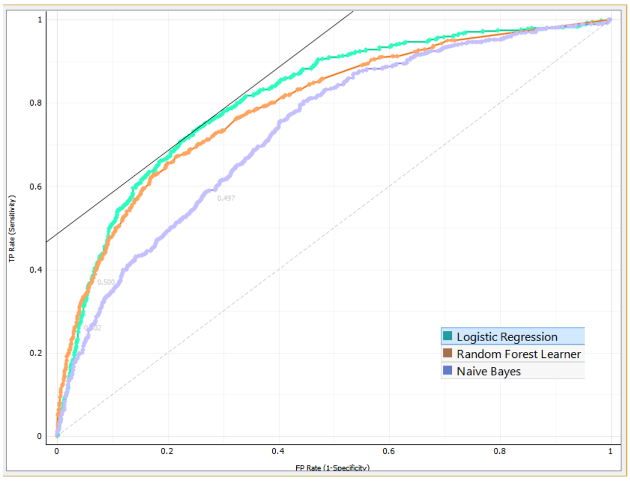

Fig. 1. ROC analysis with the target class as active against Ebola virus.

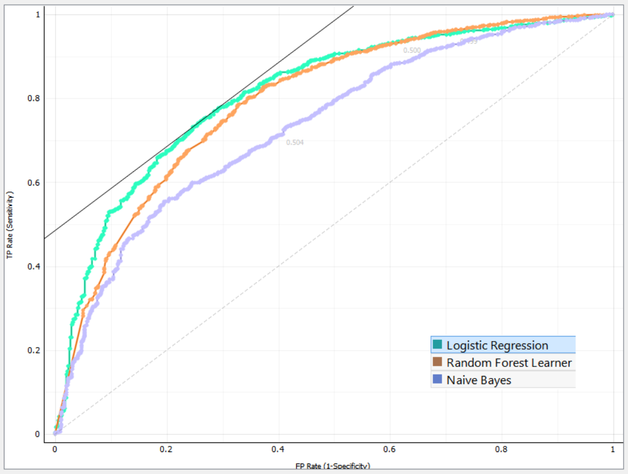

Fig. 2. ROC analysis with the target class as inactive against Ebola virus.

IV. CONCLUSION

The results in this study suggest that the chemical and physical properties of chemical molecules contribute towards their activities against Ebola virus. This knowledge could thus be used to aid in searching for potentially effective chemical molecules against Ebola virus. For example, it could be used to help speeding up the virtual screening process employed in related bioinformatics and cheminformatics methods.

REFERENCES

[1] B. Singh, A. Ganguly, and H. Sunwoo, “Current and future diagnostic tests for ebola virus disease,” *J. Pharm. Pharm. Sci.*, vol. 19, no. 4, pp. 530-551, December 2016.

[2] S. Omilabu, O. Salu, B. Oke, and A. James, “The West African Ebola virus disease epidemic 2014-2015: A Commissioned review,” *Niger Postgrad Med J.*, vol. 23, no. 2, pp. 49-56, April-June 2016.

[3] H. Shiwani, R. Pharithi, B. Khan, C. Egom, P. Kruzliak, V. Maher, and E. Egom, “An update on 2014 Ebola outbreak in Western Africa,” *Asian Pac. J. Trop. Med.*, vol. 10, no. 1, pp. 6-10, January 2017.

[4] J. Lee and E. Sapphire, “Ebolavirus glycoprotein structure and mechanism of entry,” *Future Virol.*, vol. 4, no. 6, pp. 621–635, March 2009.

[5] K. Gnir, A. K ühl, C. Karsten, I. Glowacka, S. Bertram, F. Kaup, H. Hofmann, and S. P öhlmann, “Cathepsins B and L activate Ebola but not Marburg virus glycoproteins for efficient entry into cell lines and macrophages independent of TMPRSS2 expression,” *Virol.*, vol. 424, pp. 3–10, January 2012.

[6] M. Shimojima, Y. Ikeda, and Y. Kawaoka, “The Mechanism of Axl-Mediated Ebola Virus Infection,” *J. Infect. Dis.*, vol. 196, no. Suppl 2, pp. S259-S263, November 2007.

[7] B. Zawilińska and M. Kosz-Vnenchak, “General introduction into the Ebola virus biology and disease,” *Folia. Med. Cracov*., vol. 54, no. 3, pp. 57-65, 2014.

[8] V. Karthick, N. Nagasundaram, C. Doss, C. Chakraborty, R. Siva, A. Lu, G. Zhang, and H. Zhu, “Virtual screening of the inhibitors targeting at the viral protein 40 of Ebola virus,” *Infect. Dis. Poverty*., vol. 5, no. 12, pp. 1-10, February 2016.

[9] E. Adu-Gyamfi, K. Johnson, M. Fraser, J. Scott, S. Soni, K. Jones, M. Digman, E. Gratton, C. Tessier, and R. Stahelin, “Host cell plasma membrane phosphatidylserine regulates the assembly and budding of Ebola virus,” *J. Virol.*, vol. 89, no. 18, pp. 9440–9453, September 2015.

[10] G. Matua, D. Van der Wal, and R. Locsin, “Ebola hemorrhagic fever outbreaks: strategies for effective epidemic management, containment and control,” *Braz. J. Infect. Dis*., vol. 19, no. 3, pp. 308-313, May-June 2015.

[11] C. Rio, A. Mehta, G. Lyon, and J. Guarner, “Ebola Hemorrhagic Fever in 2014: The Tale of an Evolving Epidemic,” *Ann. Intern. Med*., vol. 161, no. 10, pp. 746-748, November 2014.

[12] B. Billioux, B. Smith, and A. Nath, “Neurological Complications of Ebola Virus Infection,” *Neurotherapeutics*, vol. 13, pp. 461–470, July 2016.

[13] T. Uyeki, A. Mehta *et al*., “Clinical management of Ebola Virus disease in the United States and Europe,” *N. Engl. J. Med.*, vol. 374, no. 7, pp. 636–646, February 2016.

[14] P. Cherpillod, M. Schibler, G. Vieille, S. Cordey, A. Mamin, P. Vetter, and L. Kaiser, “Ebola virus disease diagnosis by real-time RT-PCR: A comparative study of 11 different procedures,” *J. Clin. Virol.*, vol. 77, pp. 9-14, April 2016.

[15] M. Broadhurst, T. Brooks, and N. Pollock, “Diagnosis of Ebola Virus Disease: Past, present, and future,” *Clin. Microbiol. Rev*., vol. 29, no. 4, pp. 773-793, October 2016.

[16] J. Prescott, T. Bushmaker, R. Fischer, K. Miazgowicz, S. Judson, and V. Munster, “Postmortem Stability of Ebola Virus,” *Emerg. Infect. Dis.*, vol. 21, no. 5, pp. 856–859, May 2015.

[17] T. Cheng, Y. Pan, M. Hao, Y. Wang, and S. Bryant, “PubChem applications in drug discovery: a bibliometric analysis,” *Drug Discov. Today*, vol. 19, no. 11, pp. 1751-1756, November 2014.

[18] S. Kim, P. Thiessen, E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. Bryant, “PubChem Substance and Compound databases,” *Nucl. Acids Res.*, vol. 44, no. D1, pp. D1202-D1213, January 2016.

[19] S. Kim, P. Thiessen, E. Bolton, and S. Bryant, “PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem,” *Nucl. Acids Res.*, vol. 43, no. W1, pp. W605-W611, April 2015.

[20] S. Kim, P. Thiessen, T. Cheng, B. Yu, B. Shoemaker, J. Wang, E. Bolton, Y. Wang, and S. Bryant. (June 2016). Literature information in PubChem: associations between PubChem records and scientific articles. J. Cheminform. [Online]. 8(32). pp. 1-15. Available: https://jcheminf.springeropen.com/articles/10.1186/s13321-016-0142-6

[21] S. Kim, L. Han, B. Yu, V. Hähnke, E. Bolton, and S. Bryant, “PubChem structure–activity relationship (SAR) clusters,” *J. Cheminform*., vol. 7, no. 33, pp. 1-22, July 2015.

[22] A. Zakharov, M. Peach, M. Sitzmann, and M. Nicklaus, “QSAR modeling of imbalanced high-throughput screening data in pubchem,” *J. Chem. Inf. Model*., vol. 54, no. 3, pp. 705–712, March 2014.

[23] Y. Wang, J. Xiao, T. Suzek, J. Zhang, J. Wang, Z. Zhou, L. Han, K. Karapetyan, S. Dracheva, B. Shoemaker, E. Bolton, A. Gindulyte, and S. Bryant, “PubChem's Bioassay database,” *Nucl. Acids Res*., vol. 40 (Database issue), pp. D400-D412, January 2012.

[24] Y. Wang, S. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. Shoemaker, P. Thiessen, S. He, and J. Zhang, “PubChem BioAssay: 2017 Updates,” *Nucl. Acids. Res.*, vol. 45, no. D1, pp. D955-D963, November 2016.

[25] Y. Wang, T. Cheng, and S. Bryant. (January 2017). PubChem BioAssay. SLAS Discov. [Online]. pp. 1-12. Available: http://journals.sagepub.com/doi/full/10.1177/2472555216685069

[26] L. Balderud, D. Murray, N. Larsson, U. Vempati, S. Schürer, M. Bjäreland, and O. Engkvist, “Using the bioassay ontology for analyzing high-throughput screening data,” *SLAS Discov*., vol. 20, no. 3, pp. 402–415, December 2014.

[27] K. Helal, M. Maciejewski, E. Gregori-Puigjané, M. Glick, and A. Wassermann, “Public domain HTS fingerprints: Design and evaluation of compound bioactivity profiles from PubChem's bioassay repository,” *J. Chem. Inf. Model*., vol. 56, no. 2, pp. 390-398, February 2016.

[28] P. Mazzatorta, L. Tran, B. Schilter, and M. Grigorov, “Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity,” *J. Chem. Inf. Model.*, vol. 47, no. 1, pp. 34-38, January-February 2007.

[29] E. Bolton, J. Chen, S. Kim, L. Han, S. He, W. Shi, V. Simonyan, Y. Sun, P. Thiessen, J. Wang, B. Yu, J. Zhang, and S. Bryant, “PubChem3D: a new resource for scientists,” *J. Cheminform*., vol. 3, no. 32, pp. 1-15, September 2011.

[30] J. Demšar, T. Curk *et al*., “Orange: Data mining toolbox in Python,” *J. Mach. Learn. Res.*, vol. 14, pp. 2349-2353, August 2013.

R. Nakrumpai is an assistant professor at the Department of Biochemistry, Faculty of Medical Science, Naresuan University, Thailand and the Centre of Excellent in Medical Biotechnology, Faculty of Medical Science, Naresuan University, Thailand. After receiving a Ph.D. degree from the Department of Molecular Biology and Biotechnology, University of Sheffield, UK (under joint supervision with the Chemoinformatics Research Group, Information School, University of Sheffield, UK) from a thesis about using bioinformatics to study proteins, current research interests are of bioinformatics, cheminformatics, and systems biology.