Application of the Smith Waterman and Jukes Cantor Algorithm

by Aang Darmawan

Submission date: 23-Jul-2023 09:20PM (UTC+0700)
Submission ID: 2135351428
File name: 10887-37080-1-PB.pdf (426.19K)
Word count: 3616
Character count: 18425

JOURNAL OF
NATURAL SCIENCES AND MATHEMATICS RESEARCH

Available online at http://journal.walisongo.ac.id/index.php/jnsmr

Application of the Smith Waterman and Jukes Cantor Algorithm in the Arrangement of the SARS CoV-2 Virus

Tony Yulianto[1], Mohamad Tafrikan[2], Rica Amalia[1], Emi Yunita[3], Moch. Haikal[4], Fathorrozi Ariyanto[5], Zuhrotul Hasanah[1]

[1] Department of Mathematics, Universitas Islam Madura, Indonesia
[2] Department of Mathematics, Universitas Islam Negeri Walisongo Semarang, Indonesia
[3] Department of Midwifery, Universitas Islam Madura, Indonesia
[4] Department of Biology Education, Universitas Islam Madura, Indonesia
[5] Department of Informatics Engineering, Universitas Islam Madura, Indonesia

Corresponding author:
toniyulianto65@gmail.com
Received : 10 Jan 2022
Revised : 10 May 2022
Accepted : 25 June 2022

Abstracts

In early 2020, the world was shocked by an outbreak of a new pneumonia that started in Wuhan, Hubei Province, which then spread rapidly to more than 190 countries and territories. This outbreak was named coronavirus disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). The spread of this disease has had a wide social and economic impact. There is still a lot of controversy surrounding this disease, including in the aspects of diagnosis, treatment, and prevention. Therefore, a study was carried out on studies related to COVID-19 that have been widely published since the beginning of 2020 until the end of March 2020.So to overcome this problem, the Smith Waterman Jukes Cantor Algorithm was made to align Covid19 by taking the a pair of DNA and RNA sequences to align protein sequences. From this alignment, the percentage of identical and mutations will be known. The identical percentage in the genetic code will prove that although the symptoms caused by the disease are almost the same, the protein sequences are not necessarily the same. Based on the simulation results of the distance between sequences that produce a phylogenetic tree using the jukes cantor method, it was obtained that 4 groups of 26 sequences were divided into groups, namely, group 1 consists of 16 sequences, group 2 consists of 6 sequences, group 3 consists of 2 sequences, group 4 consists of 2 sequences. Based on these groups, it turns out that the China Wuhan sequence (sequence MT291826) is located in group 1 and other countries that are almost similar to the sequence in China Wuhan, namely the country of Timoe Leste with the sequence MT641766 also located in group 1.

Keywords: Covid19; DNA; RNA; Jukes Cantor; Smith Waterman Algorithm

1. Introduction

Although genes are getting more attention for research and discussion, it is actually proteins that have a major role in carrying out life functions and make up the majority of cellular structures. If a gene is disturbed, causing the protein it encodes to be unable to carry out its normal function, it will result in a genetic defect [10, 11, 18, 20, 27, 32].

To find some of the causes of a disorder in the body's organs that are difficult to study what the cause is, DNA is used as one of the main tools in various research in the field of biology. Genetic diversity based on mitochondrial DNA is currently highly developed because mitochondrial DNA has a high number of descendants. In addition, RNA is also needed for a ribonucleic acid contained in the genetic information flow of organisms in the form of the central dogma of DNA->RNA->protein, namely DNA is transcribed into RNA, and then RNA is translated into protein. [1, 2, 8, 9, 17].

As for in this study focused on the application of alignment with the algorithm smith-Waterman. Smith Waterman's algorithm is a type of local alignment algorithm. From this alignment, the percentage of identical and mutations will be known. The identical percentage in the genetic code will prove that although the symptoms caused by the disease are almost the same, the protein sequences are not necessarily the same. Although it looks simple to develop algorithms based on dynamic programming with appropriate local alignment, this algorithm is very important in bioinformatics [12, 13, 14, 19, 23].

Along with the development of bioinformatics as a science that applies computational techniques to manage and analyze biological information, bioinformatics also includes the application of mathematical, statistical and informatics methods to solve biological problems, especially by using information contained in a DNA or protein sequence. By aligning and analyzing the protein sequence and the level of the virus will be known. The fundamental contribution to a field of science given in this research is the contribution to the field of Bioinformatics. As it is known that bioinformatics is a combination of mathematical, statistical, and informatics methods to process biological data [21, 30, 24, 25].

This algorithm tries to find as many similarities as possible from a pair of DNA and RNA sequences, by assigning a negative value to the base pair that is not the same (mismatch), and a positive value to the same base pair (match). So it will get the maximum positive value as the end of the alignment, and the minimum value as the beginning of the alignment. From a previous study, alignment results show similarity of 84%, with a gap of 3% [32, 34].

2. Experiments Procedure

This research will be divided into several stages as follows:

1. Literature Study
   At this stage, supporting theories will be studied such as: Bioinformatics [24, 26], DNA and protein [8, 18, 30, 34], protein sequence alignment, the Smith Waterman algorithm, the application of the Smith Waterman algorithm for sequence alignment [26, 35].
2. Genbank data retrieval
   The data used were taken from the National Center for Biotechnology Information [27, 36].
3. At this stage, the Smith Waterman method will be applied to the Covid data by aligning the Covid data [3, 4, 5, 6]. The Smith Waterman Algorithm is as follows [15, 16]:
   a. Alignment of protein sequences using the Smith Waterman algorithm using formula (1) [29, 30]:

$$s(i,j) = max \begin{cases} 0 \\ s(i-1,j-1) + s(x,y_i) \\ s(i-1,j) - d \\ s(i,j-1) - d \end{cases} \quad (1)$$

   with a series of processes and trace back algorithms.

b. Calculates the total time used to perform the alignment.
c. Calculates the identical percentage of the two sequences using Jukes Cantor's model [20, 24, 31] in equation (2).

$$d = -\frac{3}{4}\ln\left(1 - \frac{3}{4}\,p\right) \quad (2)$$

Where:

p: different nucleotide proportions in two sequences

d : score penalty from virtual symbol

d. Indicates the mutation that occurs in the virus [7, 9, 27].
e. Forming a phylogenetic tree [34, 35, 37]

4. Program Implementation
After the Smith Waterman method was carried out, it was then applied to the Matlab software. The steps are as follows [29, 30, 31]:
a. Designing a menu interface to facilitate communication between the user and the computer (GUI)
b. Entering the FASTA code of all data sequences in a txt file.
5. Analysis and Discussion
The results of the alignment will later be tabled and analyzed, what is the identical percentage of each alignment sequence, and the spread of the virus to Indonesia from any country [28].
6. Conclusion Drawing
This stage is the last stage in completing the research. After the research got the results from the application of the Smith Waterman method. Then the conclusions and suggestions of this research are drawn [32, 33].

3. Result and Discussion

Data Identification

During the SARS epidemic, many groups of scientists isolated and published SARS sequences. Virus that was originally recognized as the coronavirus that causes SARS, it turns out to have a different sequence with other coronaviruses that attack humans. Hence, the origin of coronavirus is thought to have come from animals [36].

In this study, we will compare some of the suspected DNA and RNA as a coronavirus host by using protein sequences that can downloaded from GenBank database [10]. GenBank is a storage site largest genetic sequence database today. GenBank is managed by the NIH (National Institute of Health) America, which is a composite of the database sequence database international nucleotides consisting of DNA DataBank of Japan (DDJB), European Molecular Biology Laboratory (EMBL) and GenBank itself in National Center of Biotechnology Information (NCBI) [26].

Mutation Result From Alignment Process

The mutation results obtained from the alignment between sequence 1 and sequence 5, and obtained 7 different sequences with the following results:

1. At the 684th position mutation from C to T
2. At position 1858 mutation from T to G
3. At the 11046th position mutation from T to G
4. At position 21674 mutation from T to C
5. At position 24288 mutation from A to G
6. At position 26107 mutation from T to G
7. At position 29658 mutation from G to A

As for the others, the mutation results obtained from the alignment between sequence 1 and sequence 10, and obtained 10 different sequences with the following results:

1. At position 701 mutation from C to T
2. At position 1875 mutation from T to G
3. At position 4382 mutation from T to C
4. At position 5042 mutation from G to T
5. At position 8762 mutation from C to T
6. At position 11063 mutation from T to G
7. At position 21691 mutation from T to C
8. At position 26124 mutation from T to G
9. At position 28124 mutation from T to C
10. At position 29675 the mutation from G to A.

Results of Phylogenetic Tree Formation

Cycle 1

a. Input: Cycle 1 evolutionary distance matrix [32, 34, 38, 40].

Step 1: ,

Calculate following the formula: S_i

$$S_i = \frac{1}{N-2}\sum_{k=1}^{N} D_{ik}$$

Where N is the number of OTUs, D_{ik} is the distance from i to k on its evolutionary matrix. While N in this Cycle is 26.

Step 2:

Find the minimum value for each pair of sequences:

$$M_{ij} = D_{ij} - S_i - S_j$$

When written in full, the matrix M_{ij} it is as follows: The smallest M pair is obtained = -0.0139M_{RX}

Step 3:

Define a new OTU i.e. U_1 which replaces the smallest pair (R and X). Furthermore, these taxa are combined as U_1 follow the formula:

$$S_{RU} = 0.5\,(D_{RX} + S_R - S_X) = -0.03421$$

$$S_{XU_1} = 0.5\,(D_{RX} + S_X - S_R) = 0.03421$$

Step 4:

Connect taxa U_1 with R and U_1 with X, respectively by following the edge length or distance as the result calculation in step 3.

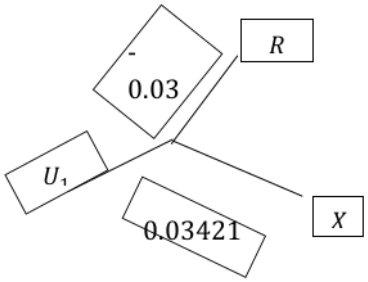

Figure 1. Tree in Cycle1

In the tree Figure 1 the length of the branch describes the evolutionary distance.

Step 5:

Merge new distances from all taxa to U_1

$$D_{AU_1} = 0.5\,(D_{RA} + D_{XA} - D_{RX}) = 0.00675$$

$$D_{BU_1} = 0.5\,(D_{RB} + D_{XB} - D_{RX}) = 0.0018$$

Result of new distance U_1 to all taxa for further inclusion in the new evolutionary distance matrix. The next calculation step is the same as in cycle 1, the value is different because N is different, the smallest M_{RX} is different, which finally the new distance to each taxa is also different.

After all the branches are combined, the following phylogenetic tree is obtained [22, 27, 28, 38, 39, 40]:

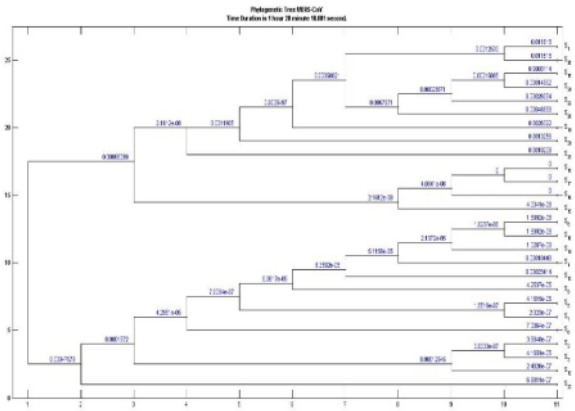

Figure 2. Group and Distance

4. Conclusion

From the discussion that has been carried out, the conclusions obtained are as follows: the results of mutations between sequences produce differences, sometimes mutations occur and there are no mutations at all. As for the mutation, the susceptibility is very small because only the location is different and the virus is on average the same, namely the type of SAR-COV-2 virus.

Based on the simulation results of the distance between sequences that produce a phylogenetic tree using the Jukes Cantor method, 26 sequences were obtained consisting of 4 groups of them, group 1 consisted of 16 sequences, group 2 consisted of 6 sequences, group 3 consisted of 2 sequences, group 4

consisted of 2 sequences. Based on these groups, it turns out that the China Wuhan sequence (sequence MT291826) is located in group 1 and other countries that are almost similar to the sequence in China Wuhan, namely the country of Timoe Leste with the sequence MT641766 also located in group 1. The accuracy obtained is 1 hour 28 minutes 10,080 seconds

Acknowledgement

The research is supported by mathematics department, mathematics and science faculty and LPPM Universitas Islam Madura who has provided support in the form of assistance to conduct a research.

References

[1] Y. S. Ismail, Febrian, C. Yulvizar and R. Ramadhani, "Identification Of The Bacterium Isolate From Mackerel Fish (Rastrelliger sp.) Using 16S rRNA Gene," *IOP Conference Series: Earth and Environmental Science,* 2019.

[2] . Sundari and . Khadijah, "The Application Of Barcode DNA RbcL Gene For Identification Of Medicinal," *IOP Conf. Series: Journal of Physics: Conf. Series,* 2019.

[3] C. Kirana and Samsu, "The Effect of Climate On The Outbreak Of Covid-19: A Review," *IOP Conf. Series: Earth and Environmental Science,* 2021.

[4] V. Gallego, H. Nishiura, R. Sah and A. J. R. Morales, "The COVID-19 outbreak and implications for the Tokyo 2020 Summer Olympic Games," *Travel Med Infect Dis,* vol. 34, no. 101604, 2020.

[5] M. Gupta, A. Abdelmaksoud, M. Jafferany, T. Lotti, R. Sadoughifar and M. Goldust, "COVID-19 and economy," *Dermatologic Therapy,* p. 1, 2020.

[6] W. C. W. Chan, "Nano Research for COVID-19," *ACS Nano,* vol. 14, no. 4, pp. 3719-3720, 2020.

[7] . World Health Organization, COVID-19 Weekly Epidemiological Update, All The World: National Authorities, 2020.

[8] A. R. Poetsch, "The genomics of oxidative DNA damage, repair, and resulting mutagenesis," *Computational and Structural Biotechnology Journal,* vol. 18, pp. 207-219, 2020.

[9] T. R. F. Smith, A. Patel, S. Ramos, D. Elwood, X. Zhu, J. Yan, E. N. Gary, S. N. Walker, K. Schultheis, M. Purwar, . Z. Xu, J. Walters, P. Bhojnagarwala, M. Yang, . N. Chokkalingam, P. Pezzoli, E. Parzych, E. L. Reuschel, A. Doan, N. Tursi, M. Vasquez, . J. Choi, E. T. Ruiz, I. Maricic, . M. A. Bah, Y. Wu, D. Amante, D. H. Park, Y. Dia, A. R. Ali, F. I. Zaidi, A. Generotti, K. Y. Kim, T. A. Herring, S. Reeder, V. M. Andrade, K. Buttigieg, G. Zhao, . J.-M. Wu, D. Li, L. Bao, J. Liu, W. Deng, C. Qin, A. S. Brown, M. Khoshnejad, N. Wang, J. Chu, D. Wrapp, J. S. McLellan, K. Muthumani, B. Wang, M. W. Carroll, J. J. Kim, J. Bover, D. W. Kulp, L. M. P. F. Humeau, D. B. Weiner and K. E. Broderick, "Immunogenicity of a DNA vaccine candidate for COVID-19," *Nature Communications,* vol. 11, no. 2601, pp. 1-13, 2020.

[10] Y. Kwon, J. M. Daley and P. Sung, "Reconstituted System for the Examination of Repair DNA Synthesis in Homologous Recombination," *Methods in Enzymology,* vol. 591, pp. 307-325, 2017.

[11] A. M. Fleming, Y. Ding and C. J. Burrows, "Sequencing DNA for the Oxidatively Modified Base 8-Oxo-7,8-Dihydroguanine," *Methods in Enzymology,* vol. 591, pp. 187-210, 2017.

[12] Y. Zhang, J. Wu, M. Li, J. Lin and Z. Wang, "A Three-Level Scoring System for Fast Similarity Evaluation Based on Smith-Waterman Algorithm," *2020 IEEE*

International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2020.

[13] . Alhadi, G. Ardaneswari, H. Tasman and D. Lestari, "Performance evaluation of fast smith-waterman algorithm for sequence database searches using CUDA GPU-based parallel computing," *Journal of Next Generation Information Technology,* vol. 5, no. 2, pp. 38-46, 2014.

[14] Z. Xia, Y. Cui, A. Zhang, T. Tang, L. Peng, C. Huang, C. Yang and X. Liao, "A Review of Parallel Implementations for the Smith–Waterman Algorithm," *Interdisciplinary Sciences: Computational Life Sciences,* pp. 1-14, 2021.

[15] R. Barnes, A Review of the Smith-Waterman GPU Landscape, Berkeley: Electrical Engineering and Computer Sciences University of California, 2020, pp. 1-23.

[16] L. Li, J. Lin and Z. Wang, "PipeBSW: A Two-Stage Pipeline Structure for Banded Smith-Waterman Algorithm on FPGA," *2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI),* 2021.

[17] K. Hammad, Z. Wu, E. G. Zadeh and S. Magierowski, "A Scalable Hardware Accelerator for Mobile DNA Sequencing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 29, no. 2, pp. 273 - 286, 2021.

[18] M. G. Awan, J. Deslippe, A. Buluc, O. Selvitopi, S. Hofmeyr, L. Oliker and K. Yelick, "ADEPT: a domain independent sequence alignment strategy for gpu architectures," *BMC Bioinformatics,* vol. 21, no. 406, pp. 1-29, 2020.

[19] M. J. Pallen, "Microbial Bioinformatics 2020," *Microbial Biotechnology,* vol. 9, no. 5, pp. 681-686, 2016.

[20] M. I. Irawan, I. Mukhlash, A. Rizky and A. R. Dewi, "Application of Needleman-Wunch Algorithm To," *IOP Conf. Series: Journal of Physics: Conf. Series,* 2019.

[21] R. A. Purba, S. Suparno and M. Giatman, "The Optimalization of Cosine Similarity Method in Detecting Similarity," *IOP Conf. Series Materials Science and Engineering,* 2020.

[22] D. Rahmalia, T. Herlambang, A. M. Rohmah and A. Muhith, "Weights Optimization Using Firefly Algorithm On," *Journal of Physics Conference Series,* 2020.

[23] K. N. Goswami and K. A. Srivastav, "Mathematical Modeling of Zika Virus Diasease With Non Linear Incidence and Optimal Control," *IOP Conf. Series: Journal of Physics: Conf. Series,* 2018.

[24] Q. Zou, G. Lin, X. Jiang, X. Liu and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics,* vol. 21, no. 1, pp. 1-10, 2018.

[25] J. J. Davis, . A. . R. Wattam, . R. K. Aziz, T. Brettin, R. Butler, R. M. Butler, . P. Chlenski, N. Conrad, A. Dickerman, E. M. Dietrich, J. L. Gabbard, S. Gerdes, A. Guard, R. W. Kenyon, D. Machi, C. Mao, . D. M. Olson, M. Nguyen, E. K. Nordberg, G. J. Olsen, R. D. Olson, . J. C. Overbeek, . R. Overbeek, B. Parrelloh, G. D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A. S. Warren, F. Xia, D. Xie, H. Yoo and R. Stevens, "The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities," *Nucleic Acids Research,* vol. 48, no. 1, p. 606–612, 2020.

[26] F. Gabler, S. Z. Nam, S. Till, M. Mirdita, M. Steinegger, J. Söding, A. N, L. and V. Alva, "Protein Sequence Analysis Using the MPI Bioinformatics Toolkit," *Current Protocols*

in Bioinformatics, vol. 72, no. 108, pp. 1-30, 2020.

[27] A. Poran, D. Harjanto, M. Malloy, C. M. Arieta, D. A. Rothenberg, D. Lenkala, M. M. v. Buuren, T. A. Addona, M. S. Rooney, L. Srinivasan and R. B. Gaynor, "Sequence-based prediction of SARS-CoV-2 vaccine targets using a mass spectrometry-based bioinformatics predictor identifies immunogenic T cell epitopes," *Genome Medicine,* vol. 12, no. 70, pp. 1-15, 2020.

[28] B. Robson, "Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus," *Computers in Biology and Medicine,* vol. 119, no. 103670, pp. 1-19, 2020.

[29] B. Xu, C. Li, H. Zhuang, J. Wang, Q. Wang and X. Zhou, "Efficient Distributed Smith-Waterman Algorithm Based on Apache Spark," *2017 IEEE 10th International Conference on Cloud Computing (CLOUD),* pp. 608-615, 2017.

[30] Y. Liu, T.-T. Tran, F. Lauenroth and B. Schmidt, "SWAPHI-LS: Smith-Waterman Algorithm on Xeon Phi coprocessors for Long DNA Sequences," *2014 IEEE International Conference on Cluster Computing (CLUSTER),* pp. 257-265, 2014.

[31] S. K. Zahid, L. Hasan, A. A. Khan and S. Ullah, "A novel structure of the Smith-Waterman Algorithm for efficient sequence alignment," *2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC),* pp. 6-9, 2015.

[32] F. Muhamad, R. Ahmad, S. Asi and M. Murad, "Performance Analysis Of Needleman-Wunsch Algorithm (Global) And Smith-Waterman Algorithm (Local) In Reducing Search Space And Time For Dna Sequence Alignment," *Journal of Physics: Conference Series,* vol. 1019, no. 012085, pp. 1-8, 2018.

[33] S. A. M. A. Junid, M. F. M. Idros, A. H. A. Razak, F. N. Osman and N. M. Tahir, "Parallel processing cell score design of linear gap penalty smith-waterman algorithm," *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA),* pp. 299-302, 2017.

[34] S. Röhling, A. Linne, J. Schellhorn, M. Hosseini, T. Dencker and B. Morgenstern, "The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances," *PLOS ONE,* pp. 1-18, 2020.

[35] P. K. Pandey, Y. S. Singh, P. S. Tripathy, R. Kumar, S. K. Abujam and J. Parhi, "DNA barcoding and phylogenetics of freshwater fish fauna of Ranganadi River, Arunachal Pradesh," *Gene,* vol. 754, no. 144860, pp. 1-28, 2020.

[36] J. Yavarian, N. Z. S. Jandaghi, K. Sadeghi, S. S. Malekshahi, V. Salimi, A. Nejati, F. A. Minejad, N. Ghavvami, F. Saadatmand, S. Mahfouzi, G. Fateminasab, N. Parhizgari, A. Ahmadi, K. Razavi, S. Ghabeshi, M. Saberian, E. Zanjani, F. Namazi, T. Shahbazi, F. Rezaie, H. Erfani, M. M. Gouya, M. N. Dadras and T. M. Azad, "First Cases of SARS-CoV-2 in Iran, 2020: Case Series Report," *Iran Journal Public Health,* vol. 49, no. 8, pp. 1564-1568, 2020.

[37] S. Awasthi, A. K. Mahadani, G. Sanyal and P. Bhattacharjee, "Modified indel treatment for accurate Phylogenetic Tree construction," *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM),* 2020.

[38] J. Rusinko and M. McPartlon, "Species tree estimation using Neighbor Joining," *Journal of Theoretical Biology,* vol. 414, no. 7, pp. 5-7, 2017.

[39] T. Le, A. Sy, E. K. Molloy, Q. Zhang, S. Rao and T. Warnow, "Using Constrained-INC for Large-Scale Gene Tree and Species Tree Estimation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 18, no. 1, pp. 2-15, 2020.

[40] H. Prasetya, Performance Comparison Between Kimura 2-Parameters and Jukes-Cantor Model in Constructing Phylogenetic Tree of Neighbour Joining (NJ), Bogor: IPB (Bogor Agricultural University), 2011.

Application of the Smith Waterman and Jukes Cantor Algorithm

ORIGINALITY REPORT

19%	9%	15%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ www.coursehero.com

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		