Pendekatan Teori Graf untuk Analisis Jaringan Interaksi Protein-Protein

(*Graph Theory Approach to Network Analysis of Protein-Protein Interactions*)

Yustina Sri Suharini*, Muhamad Ramli, Sulistyowati, Endang R.D.

Program Studi Teknik Informatika, Institut Teknologi Indonesia
Jl. Raya Puspiptek, Serpong, Kota Tangerang Selatan, Provinsi Banten 15320

Abstrak

Jaringan interaksi protein-protein merupakan hal penting pada setiap proses yang terjadi dalam sel biologis karena dapat digunakan untuk mempelajari kondisi fisiologis sel ketika berada dalam keadaan normal atau tidak normal. Di sisi lain, infrastruktur komputasi telah berada di era yang cukup memadai untuk menyimpan data hasil eksperimen dari berbagai tempat dan waktu. Namun data yang terkumpul perlu diolah dan dianalisis dengan cara yang tepat agar menghasilkan pengetahuan atau wawasan baru yang bermanfaat. Penelitian ini bertujuan melakukan pendekatan agar data jaringan interaksi protein-protein yang terkumpul di database menjadi informasi yang bermakna. Pendekatan dilakukan menggunakan teori graf dengan studi kasus data protein virus SARS-Cov-2. Metode yang digunakan adalah metode in-silico dengan data sekunder berasal dari database bereputasi yang dapat diakses publik. Hasil penelitian berupa daftar protein-protein paling berpengaruh pada virus SARS-Cov-2 berdasarkan parameter-parameter umum yang digunakan dalam ilmu jaringan.

Kata Kunci : Sentralitas, Cytoscape, Ilmu Jaringan, String

Abstract

The protein-protein interaction network is important in every process that occurs in biological cells because it can be used to study the physiological conditions of cells when they are in normal or abnormal conditions. On the other hand, the infrastructure is in an adequate era to store experimental data from various places and times. However, the collected data needs to be processed and analyzed in an appropriate way to produce useful new knowledge or insights. This study aims to take an approach so that the protein-protein interaction network data collected in the database becomes meaningful information. The approach is carried out using graph theory with case studies of the SARS-Cov-2 virus protein data. The method used is an in-silico method with secondary data coming from a reputable database that can be accessed by the public. The results of the research are in the form of a list of the most influential proteins on the SARS-Cov-2 virus based on general parameters used in network science.

Keyword : Centrality, Cytoscape, Network Science, String

* Telp: +62 21 7561092; fax: +62 21 7560542
Alamat E-mail : yustina.ss@iti.ac.id, ramli@iti.ac.id, sulistyowati.if@iti.ac.id, endangrd@iti.ac.id

1. Pendahuluan

Perkembangan teknologi informasi dan telekomunikasi berpengaruh ke bidang-bidang ilmu yang lain, termasuk bidang biologi molekuler. Apalagi dengan ditemukannya untaian genom lengkap atau utuh manusia, serta infrastruktur komputasi yang memungkinkan data biologi disimpan dan diolah untuk keperluan penelitian di kemudian hari, membuka peluang akan pemikiran dan pemahaman baru akan penyakit beserta respon sel tubuh atas penyakit serta pengobatan untuk terapi penyembuhannya. Salah satu perkembangan penting yang terjadi dalam belasan tahun terakhir adalah penemuan jaringan interaksi protein-protein yang dapat digunakan untuk mempelajari kondisi fisiologis sel hidup, misalnya untuk mengidentifikasi ada atau tidak ada gangguan yang mengarah kepada penyakit tertentu.

Persoalan utamanya adalah, jaringan interaksi protein-protein sangat kompleks [1]–[3]. Para peneliti menggunakan berbagai metode untuk melakukan pendekatan terhadap sistem kompleks itu. Ada yang mencocokkan satu per satu untaian protein, seperti yang dilakukan oleh Zaki N dan teman-teman di publikasi mereka pada tahun 2009 ketika untaian *the human whole genome* belum lama ditemukan [4]. Namun setelah ada usulan *deep learning* di tahun berikutnya, dan beberapa tahun kemudian para investor khususnya Google memberi fasilitas berupa infrastruktur komputasi yang lengkap, termasuk *framework* dan bahasa pemrograman yang mendukung paradigma tersebut, maka banyak peneliti yang mencoba menggunakan *deep learning* untuk melakukan analisis jaringan interaksi protein-protein [5]–[13]. Di samping pendekatan-pendekatan tersebut, ada juga pendekatan dari sudut pandang lain, yaitu yang berdasarkan teori graf. Peneliti pendahulu menggunakan graf antara lain untuk mencari hubungan obat dan target penyakit hepatitis [14]. Sedangkan teori graf yang lebih lanjut digunakan antara lain untuk mengklasifikasikan fungsi protein [15]. Penelitian ini bertujuan menggunakan teori graf untuk mengungkap jaringan interaksi protein-protein dengan studi kasus protein virus SARS-CoV-2 atau yang dikenal dengan Covid-19.

2. Teori Dasar

Protein berinteraksi dengan protein-protein lain dalam sel membentuk jaringan yang sangat besar dan kompleks. Apabila protein dinyatakan sebagai simpul (*vertex*) dan interaksi dinyatakan dalam garis (*edge*) maka jaringan interaksi protein-protein dapat dinyatakan sebagai sebuah graf G = (V, E) dengan V merupakan himpunan tidak kosong dari simpul-simpul pada graf, dinyatakan dengan V = {v_1, v_2, v_3, .. v_n}. sedangkan E merupakan himpunan garis-garis yang menghubungkan simpul satu dengan simpul lainnya, dituliskan sebagai E = {e_1, e_2, e_3, .. e_m}.

Pada jaringan interaksi protein-protein, garis tidak berarah, sehingga berlaku jumlah derajat semua simpul adalah genap, yaitu dua kali jumlah garis pada graf, sesuai dengan teori jabat tangan. Jika derajat (*degree*) dinyatakan dengan d maka

$$d\ (v) = \sum_{w \neq v} e(v, w) \quad \text{....(1)}$$

dan

$$\sum_{v \in V} d(V) = 2\ |E| \quad \text{....(2)}$$

Derajat (*degree*) pada simpul merupakan besaran penting dalam jaringan interaksi protein-protein. Besaran-besaran lain yang juga penting, sehingga sering digunakan adalah *betweenness centrality*, *closeness centrality*, *average shortest path length*, dan *neighborhood connectivity* [16]–[18].

Betweeness centrality (BC) adalah besaran yang digunakan untuk menunjukkan seberapa kuat sebuah protein berperan sebagai jembatan bagi protein-protein lain.

$$BC(v) = \sum_{u \neq v \neq w} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \quad \text{....(3)}$$

Closeness centrality (CC), merupakan besaran yang menggambarkan kedekatan sebuah protein ke protein lain, dinyatakan dalam

$$CC(v) = \frac{n-1}{\sum_{v \neq w} d(v,w)} \quad \text{....(4)}$$

Panjang rata-rata jalur terpendek (*average shortest path length*), sesuai dengan namanya, dihitung berdasarkan jumlah semua jarak terpendek simpul ke simpul lainnya dibagi jumlah jarak.

$$L(v) = \frac{\sum_{v \neq w} d(v,w)}{n-1} \quad \text{....(5)}$$

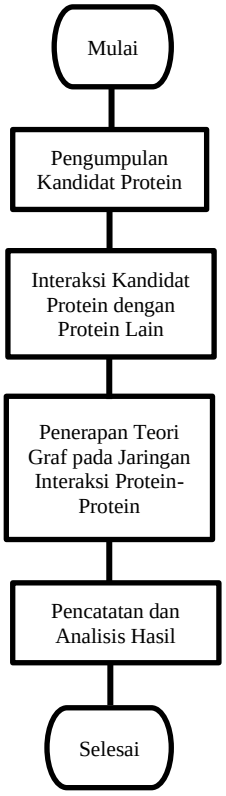
Neighborhood connectivity adalah jumlah tetangga pada sebuah protein, semakin tinggi nilainya maka semakin banyak protein lain yang bertetangga dengan protein yang sedang diselidiki ini.

$$NC(v) = \frac{\sum_{k \in N(v)} |N(v)|}{|N(v)|} \quad \text{....(6)}$$

Masing-masing besaran dapat digunakan dengan kriteria yang berbeda-beda sesuai dengan keperluan penelitian. Sebagai contoh, besaran *betweenness centrality* (BC) yang mencerminkan posisi sebuah protein menjadi jembatan bagi protein-protein lain, dapat diartikan bahwa protein tersebut berpengaruh kuat terhadap protein-protein lain, sehingga dalam jaringan interaksi yang sedang diselidiki akan dicari nilai BC yang tertinggi. Sebaliknya besaran rata-rata jalur terpendek (*average shortest path length*), akan dipilih dari yang nilainya terendah atau minimum.

3. Metodologi

Metode yang digunakan pada penelitian secara garis besar terdiri atas empat bagian, seperti ditunjukkan pada Gambar 1.



Gambar 1. Tahapan penelitian

Adapun keempat bagian metode penelitian secara garis besar dapat dijelaskan sebagai berikut.

1. Langkah pertama adalah pengumpulan kandidat protein. Pengumpulan data dilakukan pada artikel-artikel terkini yang membahas tentang SARS-Cov-2 atau Covid-19, yaitu artikel yang dipublikasikan pada tahun 2022 [19]–[24].
2. Daftar kandidat protein yang didapat dari langkah 1 diolah menggunakan aplikasi String [25], [26] untuk melihat interaksinya dengan protein-protein lain. Interaksi-interaksi yang disimpan pada String database merupakan kumpulan hasil eksperimen para peneliti sejak tahun 2003 dan hingga saat ini masih secara terus menerus diperbaharui isi dan fitur-fiturnya. Hasil langkah kedua berupa daftar baru (superset) dari daftar kandidat protein yang diperoleh melalui langkah pertama.
3. Langkah ketiga adalah penerapan teori graf pada data baru yang dihasilkan dari langkah kedua. Pada langkah ini digunakan aplikasi Cytoscape [27], yaitu sebuah perangkat lunak yang banyak digunakan untuk membuat model jaringan.
4. Langkah keempat adalah pencatatan hasil, analisis hasil, serta pelaporan.

4. Hasil dan Pembahasan<10 pt bold>

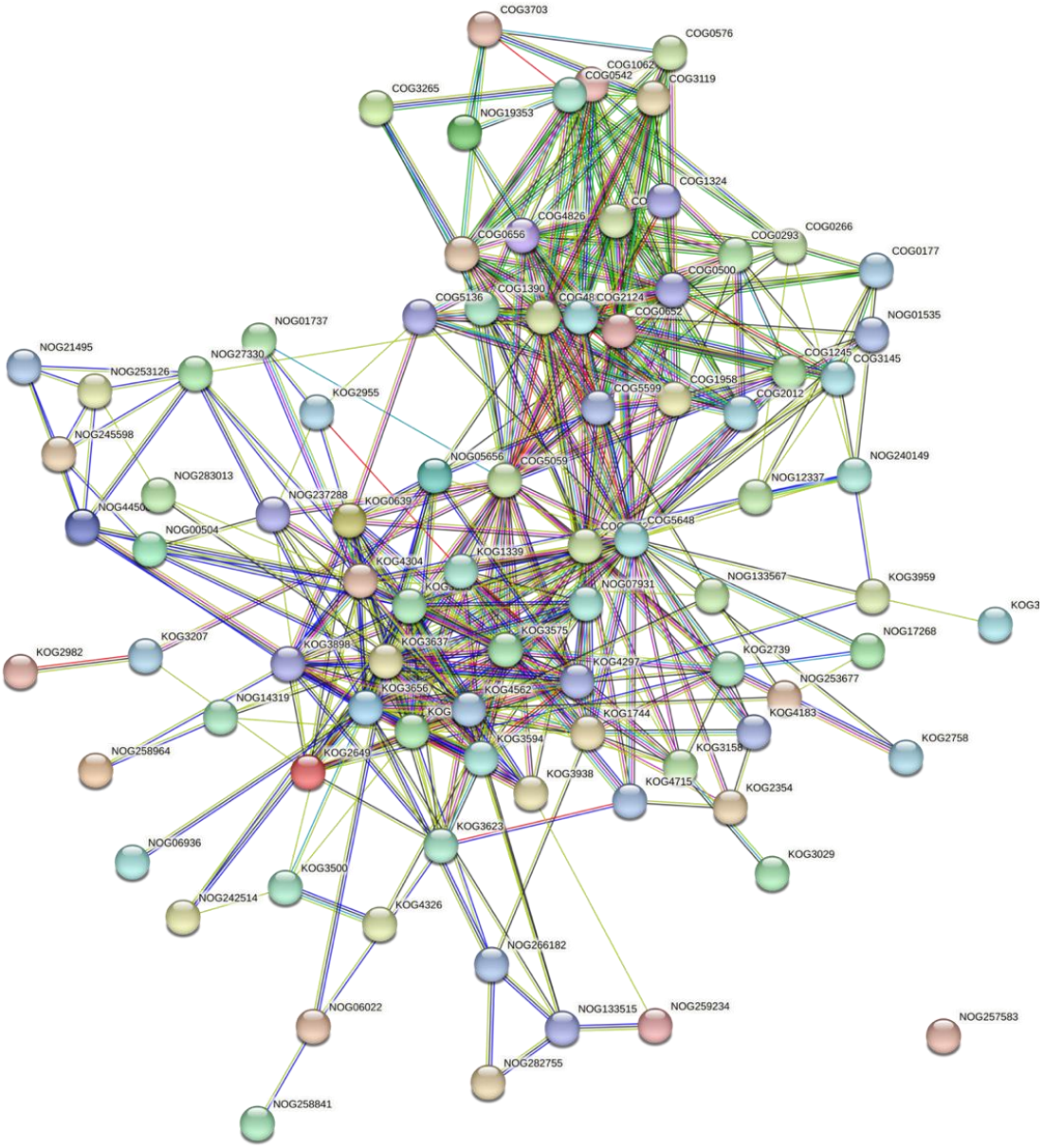
Daftar kandidat protein potensial yang dikumpulkan dari artikel-artikel tentang SARS-Cov-2 atau Covid-19 berdasarkan publikasi tahun 2022 ditunjukkan pada Tabel 1.

Tabel 1. Daftar Kandidat

Kandidat ke	Nama Protein atau Kompleks
1	cap(0)-RTC
2	E
3	N (cryo-STA)
4	N CTD
5	N NTD
6	Nsp1
7	Nsp10/Nsp16
8	Nsp13-RTC
9	Nsp15
10	Nsp21-276
11	Nsp3
12	Nsp3 Macro Domain
13	Nsp5
14	Orf3a
15	Orf7a
16	Orf8
17	Orf9b
18	S (Postfusion)
19	S Prefusion (RBD Down)
20	S Prefusion-ACE2
21	S Prefusion(one RBD up)
22	SN501Y Prefusion-Fab ab1
23	SN501Y Prefusion-VH ab8

Interaksi antara 23 kandidat protein dengan protein-protein lain disajikan pada Gambar 2. Terdapat 86 protein atau kompleks, serta 843 jalur interaksi yang dinyatakan sebagai garis (*edge*). Kemudian jaringan interaksi yang terbentuk di Gambar 2 diolah menggunakan teori graf sehingga menghasilkan urutan protein paling penting menurut besaran-besaran dalam ilmu jaringan. Karena jumlahnya cukup banyak untuk dimuat semuanya pada halaman artikel, maka disajikan 10 protein teratas sesuai besaran masing-masing. Tabel 2, Tabel 3, Tabel 4, dan Tabel 5 merupakan representasi dari 10 protein paling penting menurut masing-masing perhitungan besaran tersebut.

Meskipun terdapat sebuah simpul yang tidak terhubung pada interaksi yang dihasilkan oleh aplikasi String, namun simpul itu tidak dihilangkan, karena secara komputasi masih mempunyai nilai tertentu, yang bisa saja bermanfaat bagi penelitian berikutnya.



Gambar 2. Jaringan interaksi protein-protein yang terbentuk.

Tabel 2. Daftar 10 Protein Teratas Berdasarkan *Betweeness Centrality* (BC)

Nilai BC	Nama Protein atau Kompleks
0.895053544	KOG1744
0.874770354	COG0500
0.824307405	COG5599
0.780651784	KOG3898
0.767592905	KOG2649
0.763296635	KOG3959
0.759394091	NOG133515
0.671245802	NOG266182
0.668863112	KOG3938
0.659225207	NOG44508

Tabel 3. Daftar 10 Protein Teratas Berdasarkan *Closeness Centrality* (CC)

Nilai CC	Nama Protein atau Kompleks
0.636363636	COG5648
0.571428571	COG5059
0.559999999	COG5262
0.545454545	KOG3656
0.538461538	KOG3900
0.531645569	KOG1215
0.528301886	KOG3637
0.518518518	KOG3575
0.515337423	COG4886
0.512195121	KOG4562

Sedangkan berdasarkan nilai *closeness centrality* pada Tabel 3, justru protein KOG1744 tidak muncul di 10 teratas. Menggunakan parameter ini, nilai tertinggi 0.6 berada pada protein COG5648 yang berarti bahwa protein ini adalah yang mempunyai kedekatan paling dekat dengan protein-protein lain dalam jaringan interaksi protein-protein.

Tabel 4. Daftar 10 Protein Teratas Berdasarkan *Average Shortest Path Length*

Nilai *Average Shortest Path Length*	Nama Protein atau Kompleks
1.5714285714	COG5648
1.75	COG5059
1.7857142857	COG5262
1.8333333333	KOG3656
1.8571428571	KOG3900
1.8809523809	KOG1215
1.8928571428	KOG3637
1.9285714285	KOG3575
1.9404761904	COG4886
1.9523809523	KOG4562
1.9523809523	KOG4304

Nilai *average shortest path length* pada Tabel 4, menunjukkan bahwa protein COG5648 mempunyai nilai terendah yaitu 1.57 atau mempunyai kedekatan paling dekat dengan protein-protein lain dalam jaringan interaksi protein-protein.

Tabel 5. Daftar 10 Protein Teratas Berdasarkan *Neighborhood Connectivity* (NC)

Nilai *Neighborhood Connectivity*	Nama Protein atau Kompleks
27.0	KOG2758
25.0	NOG06936
22.75	NOG14319
21.5714285	COG1245
21.3636363	NOG07931
20.8888888	KOG3938
20.8571428	KOG0639
20.8	NOG05656
20.8	KOG4715
20.0	NOG01535

Pada besaran neighborhood connectivity yang ditunjukkan oleh Tabel 5, protein KOG2758 merupakan protein yang mempunyai konektivitas paling tinggi terhadap protein-protein lain dalam jaringan interaksi protein-protein. Nilai konektivitasnya adalah 27.

5. Kesimpulan

Penelitian menghasilkan daftar urutan protein-protein atau kompleks teratas SARS-Cov-2 atau dikenal dengan Covid-19, yang dihitung melalui penerapan teori graf dalam ilmu jaringan. Empat besaran terkait yang digunakan adalah *betweenness centrality*, *closeness centrality*, *average shortest path length*, dan *neighborhood connectivity*. Hasil penelitian yang berupa urutan protein ini dapat digunakan sebagai masukan bagi penelitian lain di bidang biologi atau kedokteran, misalnya untuk keperluan pencegahan atau pengobatan pasien terkait virus SARS-Cov-2 atau Covid-19.

Ucapan Terima Kasih

Terima kasih kepada panitia Technopex Institut Teknologi Indonesia 2022 yang memberi kesempatan kepada kami untuk memaparkan hasil penelitian pada Seminar Nasional Technopex ITI tanggal 21 Oktober 2022.

Daftar Pustaka

[1] I. Lee dan H. Nam, "Identification of drug-target interaction by a random walk with restart method on an interactome network," *BMC Bioinformatics*, vol. 19, 2018, doi: 10.1186/s12859-018-2199-x.

[2] H. Feng *dkk.*, "Machine Learning Analysis of Cocaine Addiction Informed by DAT, SERT, and NET-Based Interactome Networks.," *J Chem Theory Comput*, vol. 18, no. 4, hlm. 2703–2719, Apr 2022, doi: 10.1021/acs.jctc.2c00002.

[3] M. F. M. Bjorbækmo, A. Evenstad, L. L. Røsæg, A. K. Krabberød, dan R. Logares, "The planktonic protist interactome: where do we stand after a century of research?," *ISME Journal*, vol. 14, no. 2, hlm. 544–559, 2020, doi: 10.1038/s41396-019-0542-5.

[4] N. Zaki, S. Lazarova-Molnar, W. El-Hajj, dan P. Campbell, "Protein-protein interaction based on pairwise similarity.," *BMC Bioinformatics*, vol. 10, no. 1, hlm. 150, Mei 2009, doi: 10.1186/1471-2105-10-150.

[5] W. Wardah *dkk.*, "Predicting protein-peptide binding sites with a deep convolutional neural network," *J Theor Biol*, vol. 496, 2020, doi: 10.1016/j.jtbi.2020.110278.

[6] M. A. Rezaei, Y. Li, D. Wu, X. Li, dan C. Li, "Deep Learning in Drug Design: Protein-Ligand Binding Affinity Prediction," *IEEE/ACM Trans Comput Biol Bioinform*, hlm. 1–1, 2021, doi: 10.1109/TCBB.2020.3046945.

[7] F. Wahab Khattak, Y. Salamah Alhwaiti, A. Ali, M. Faisal, dan M. H. Siddiqi, "Protein-Protein Interaction Analysis through Network Topology (Oral Cancer)," *J Healthc Eng*, vol. 2021, hlm. 1–9, Jan 2021, doi: 10.1155/2021/6623904.

[8] H. Y. Yuen dan J. Jansson, "Better Link Prediction for Protein-Protein Interaction Networks," dalam *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, Okt 2020, hlm. 53–60. doi: 10.1109/BIBE50027.2020.00017.

[9] Y. Wang, S. Wu, Y. Duan, dan Y. Huang, "A Point Cloud-Based Deep Learning Strategy for Protein-Ligand Binding Affinity Prediction," *Brief Bioinform*, vol. 23, no. 1, Jul 2021, doi: 10.1093/bib/bbab474.

[10] G. Diez, D. Nagel, dan G. Stock, "Correlation-Based Feature Selection to Identify Functional Dynamics in Proteins.," *J Chem Theory Comput*, vol. 18, no. 8, hlm. 5079–5088, Agu 2022, doi: 10.1021/acs.jctc.2c00337.

[11] S. Wang *dkk.*, "SE-OnionNet: A Convolution Neural Network for Protein–Ligand Binding Affinity Prediction," *Front Genet*, vol. 11, hlm. 607824, Feb 2021, doi: 10.3389/fgene.2020.607824.

[12] A. Dhakal, C. McKay, J. J. Tanner, dan J. Cheng, "Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions," *Brief Bioinform*, vol. 23, no. 1, Jan 2022, doi: 10.1093/bib/bbab476.

[13] R. Nandakumar dan V. Dinu, "Developing a machine learning model to identify protein–protein interaction hotspots to facilitate drug discovery," *PeerJ*, vol. 8, 2020, doi: 10.7717/peerj.10381.

[14] F. Prado-Prado *dkk.*, "Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica.," *Eur J Med Chem*, vol. 46, no. 4, hlm. 1074–94, Apr 2011, doi: 10.1016/j.ejmech.2011.01.023.

[15] A. Ben Rejab dan I. Boukhris, "FAST Community Detection for Proteins Graph-Based Functional Classification," dalam *Advances in Intelligent Systems and Computing*, 2020, hlm. 587–596. doi: 10.1007/978-3-030-16660-1_57.

[16] A. Manuscript, I. Networks, dan H. Disease, "WoodenBoat-logo.ai," vol. 144, no. 6, hlm. 986–998, 2012, doi: 10.1016/j.cell.2011.02.016.Interactome.

[17] J. Menche *dkk.*, "Uncovering disease-disease relationships through the incomplete interactome," *Science (1979)*, vol. 347, no. 6224, hlm. 841, 2015, doi: 10.1126/science.1257601.

[18] H. C. Rustamaji *dkk.*, "A network analysis to identify lung cancer comorbid diseases," *Appl Netw Sci*, vol. 7, no. 1, 2022, doi: 10.1007/s41109-022-00466-y.

[19] S. Abubaker Bagabir, N. K. Ibrahim, H. Abubaker Bagabir, dan R. Hashem Ateeq, "Covid-19 and Artificial Intelligence: Genome sequencing, drug development and vaccine discovery.," *J Infect Public Health*, vol. 15, no. 2, hlm. 289–296, Feb 2022, doi: 10.1016/j.jiph.2022.01.011.

[20] T. Singh, R. Khan, dan S. Srivastava, "A Review on AI-Based Techniques for Tackling COVID-19," dalam *Studies in Computational Intelligence*, 2022, hlm. 325–336. doi: 10.1007/978-981-16-8012-0_25.

[21] A. Khamis *dkk.*, "AI and Robotics in the Fight Against COVID-19 Pandemic," dalam *Studies in Systems, Decision and Control*, 2022, hlm. 57–85. doi: 10.1007/978-3-030-72834-2_3.

[22] G. Floresta, C. Zagni, D. Gentile, V. Patamia, dan A. Rescifina, "Artificial Intelligence Technologies for COVID-19 De Novo Drug Design.," *Int J Mol Sci*, vol. 23, no. 6, hlm. 3261, Mar 2022, doi: 10.3390/ijms23063261.

[23] R. Gupta, A. Gupta, M. Bedi, dan S. K. Pal, "Application of Deep Learning Techniques for COVID-19 Management," dalam *Studies in Computational Intelligence*, 2022, hlm. 165–197. doi: 10.1007/978-3-030-74761-9_8.

[24] S. Monteleone, T. F. Kellici, M. Southey, M. J. Bodkin, dan A. Heifetz, "Fighting COVID-19 with Artificial Intelligence.," *Methods Mol Biol*, vol. 2390, hlm. 103–112, 2022, doi: 10.1007/978-1-0716-1787-8_3.

[25] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, dan B. Snel, "STRING: a database of predicted functional associations between proteins.," *Nucleic Acids Res*, vol. 31, no. 1, hlm. 258–61, Jan 2003, doi: 10.1093/nar/gkg034.

[26] D. Szklarczyk *dkk.*, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res*, vol. 39, no. Database, hlm. D561–D568, Jan 2011, doi: 10.1093/nar/gkq973.

[27] 1 Paul Shannon *dkk.*, "Cytoscape: A Software Environment for Integrated Models," *Genome Res*, vol. 13, no. 22, hlm. 426, 1971, doi: 10.1101/gr.1239303.metabolite.