MATHunesa
Jurnal Ilmiah Matematika *Volume 8 No. 2 Tahun 2020*
ISSN 2716-506X

APLIKASI *SUPPORT VECTOR MACHINE* (SVM) UNTUK PENCARIAN *BINDING SITE* PROTEIN-LIGAN

Antri Wulandari
Jurusan Matematika, FMIPA, Universitas Negeri Surabaya
e-mail : antriwulandari3@gmail.com

Abstrak

Drug design (desain obat) telah banyak dikembangkan dengan berbantuan komputer. Langkah awal dalam desain obat berbantuan komputer yaitu dengan mencari daerah *binding site* suatu protein. *Binding site* adalah suatu rongga pada permukaan protein yang berperan sebagai tempat melekatnya suatu ligan. Dalam penelitian ini, prediksi *binding site* protein-ligan dirumuskan sebagai klasifikasi biner, yaitu sebagai pembeda daerah berpotensi mengikat ligan dan daerah yang tidak berpotensi mengikat ligan. Dataset yang akan digunakan dalam penelitian ini yaitu diambil dari *webserver* RCSB *Protein Data Bank*) sebanyak 14 data protein. Untuk menyelesaikan masalah klasifikasi tersebut, dipilihlah metode *Support Vector Machine* (SVM). Hasil dari penelitian ini diperoleh rata-rata *training* untuk akurasi 99,02%, *precision* 99,04%, *recall* 99,02%, *f-measure* 99,02%, *MCC* 96,84%, *ROC* 92,82%, *PRC* 93,71%, dan rata-rata waktu *training* sebesar 18,92 detik. Serta didapat rata-rata akurasi *testing* sebesar 95,98 % dan rata-rata waktu *test* sebesar 0,03642 detik.
Kata kunci: *drug design, Support Vector Machine (SVM)*

Abstract

Drug design (drug design) has been developed by using computer aids. The initial step in the design of computer-aided drugs is to look for a protein binding site. The binding site is a cavity on the surface of the protein where the ligand attaches. In this study, the prediction of protein-ligand binding sites is formulated as a binary classification, namely as a differentiator of regions that are bound to ligand binding and areas that are not binding to ligand binding. The data to be used in this study was taken from the RCSB Protein Data Bank web server as many as 14 protein data. To solve this classification problem, the Support Vector Machine (SVM) method was chosen. The results of this study obtained an average training for achieving 99.02%, precision 99.04%, remember 99.02%, f-size 99.02%, VFD 96.84%, ROC 92.82%, RRC 93 , 71%, and an average training time of 18.92 seconds. Thus the average test obtained was 95.98% and the average test time of 0.03642 seconds.
Keywords: *drug design, Support Vector Machine (SVM)*

1. PENDAHULUAN

Bioinformatika adalah ilmu multi disiplin yang melibatkan berbagai bidang ilmu seperti komputasi, matematika, biologi molekuler modern, dan penelitian medis (Ng & Wong, 2004). Salah satu aplikasi dari bioinformatika adalah *drug design* (desain obat), yaitu pencarian suatu daerah *binding site* (rongga) pada protein yang berperan sebagai tempat melekatnya suatu ligan (partikel kecil) atau calon obat. Setiap molekul protein dihubungkan dengan jalur biokimia seluler spesifik yang hanya akan mengikat struktur ligan tertentu. Sinyal kimia suatu ligan yang berikatan dengan molekul protein akan menyebabkan suatu respon jaringan, yaitu mengaktifkan atau menghambat jalur biokimia yang terkait protein tersebut (Mahdiyah, 2017).

Protein adalah rantai asam amino yang bergabung dengan ikatan peptida yang berperan penting dalam mengatasi berbagai masalah dalam tubuh manusia dan merupakan penyusun utama seluruh sel tubuh. Fungsi protein antara lain adalah membentuk enzim dan hormon, membentuk sel darah, dan membuat antibodi untuk melindungi tubuh dari penyakit dan infeksi (Horton, dkk., 2011). *Binding site* protein-ligan adalah kantung pada protein yang mengikat atau membentuk ikatan kimia dengan molekul dan ion lain (ligan) (Mahdiyah, 2015). Pengikatan protein oleh *binding site* sering *reversible* dan dapat stabil atau tidak stabil bergantung pada struktur dan aktivitasnya. Banyak ilmuwan mencoba melakukan eksperimen dan penelitian mengenai *binding site* untuk menemukan ligan atau obat yang cocok agar bisa mengobati suatu penyakit tertentu. Beberapa penelitian tersebut diantaranya adalah yang dilakukan oleh Gu dkk. (2014), White dkk. (2008), dan Carrasco dkk. (2016).

Drug Design (desain obat) dikategorikan menjadi dua jenis, yaitu desain obat berbasis struktur dan desain obat berbasis ligan. Desain obat berbasis struktur merupakan sebuah pendekatan yang didasarkan pada informasi struktur geometri dan kimia dari protein. Desain obat berbasis ligan merupakan sebuah pendekatan berbantuan komputer yang didasarkan pada informasi dari ligan dan digunakan ketika informasi 3D reseptor tidak ada (Aparoy, dkk., 2012). Pada dasarnya desain obat dilakukan atas informasi dari struktur protein untuk mencari ligan yang cocok (Wong, dkk., 2013). Informasi dari struktur protein merupakan hasil dari analisis geometri, *sequence*, dan energi dari protein yang didapat dari struktur tiga dimensi dari target, serta *binding site*

protein-ligan yang ditemukan merupakan dasar untuk pencarian rongga (*binding site*) (Mahdiyah, 2015).

Penelitian menggunakan pendekatan komputasi berbasis struktur dan sequence untuk memprediksi *binding site* telah banyak dilakukan (Oh, dkk., 2009), antara lain: *Predicting Functionally Important Residues From Sequence Conservation* (Capra & Singh, 2007), *Prediction of Protein Binding site in Protein Structures Using Hidden Markov Support Vector Machine* (Liu, dkk., 2009), *Predicting Protein-Ligand Binding site Using Support Vector Machine with Protein Properties* (Wong, dkk., 2013), dan *Integrating Data Selection and Extreme Learning Machine to Predict Protein-Ligand Binding site* (Mahdiyah, dkk., 2016).

Machine Learning (ML) merupakan sebuah bidang riset yang menggabungkan matematika, statistika, logika inferensi, analisa, dan visualisai data (Mahdiyah, 2015). *Machine Learning* (ML) telah banyak digunakan untuk memprediksi *binding site* dan telah memberikan hasil yang baik. Dapat dilihat pada beberapa penelitian yang ditulis oleh Capra dan Singh (2007), Liu dkk. (2009), Wong dkk. (2013), dan Mahdiyah dkk. (2016).

Support Vector Learning (SVM) merupakan sebuah metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) yaitu bertujuan menemukan *hyperplane* terbaik yang memisahkan dua buah class pada *input space*. Pada penelitian ini akan dilakukan pencarian suatu daerah *binding site* protein-ligan untuk *drug design* menggunakan *Support Vector Learning* (SVM) (Mahdiyah, 2017).

2. KAJIAN TEORI

Drug Design

Desain obat dilakukan untuk menemukan obat baru berdasarkan informasi dari target biologis. Desain obat sangatlah kompleks, sehingga perlu analisis secara mendalam dari segi kimia dan biologi. Keberhasilan suatu obat menyembuhkan penyakit tertentu sangat ditentukan oleh proses desain obat. Aplikasi desain obat semakin berkembang, salah satu diantaranya ialah desain obat berbantuan komputer yaitu dengan memanfaatkan informasi tiga dimensi struktur biomolekul target (protein) (Zhou & Zhong, 2017).

Protein adalah rantai asam amino yang bergabung dengan ikatan peptida yang berperan penting dalam mengatasi berbagai masalah dalam tubuh manusia. Aktivitas protein atau fungsi biokimia ditentukan oleh struktur tiga dimensinya. Konformasi tiga dimensi protein adalah hasil dari *X-ray crystallography* atau *Nuclear Magnetic Resonance* (NMR) yang berupa titik koordinat (x, y, z) (Horton, dkk., 2011).

Pada dasarnya desain obat dilakukan atas informasi dari struktur protein untuk mencari ligan yang cocok (Wong, dkk., 2013). Informasi dari struktur protein merupakan hasil dari analisis geometri, *sequence*, dan energi dari protein yang didapat dari struktur tiga dimensi dari target, serta *binding site* protein-ligan yang ditemukan merupakan dasar untuk pencarian rongga (*binding site*) (Mahdiyah, 2015).

Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi bertipe *supervised learning* yang bekerja dengan mencari *hyperplane* terbaik. *Hyperplane* (batas keputusan) bertujuan untuk memisahkan jarak antar kelas. Usaha untuk menemukan lokasi *hyperplane* terbaik merupakan inti dari proses pelatihan dalam SVM.

Langkah-langkah algoritma SVM (Mahdiyah, 2017):

1. Hitung nilai hyperplane berdimensi d dengan persamaan berikut:
   $\vec{w} \cdot \vec{x}_i + b = 0$ (1)
   $\vec{x}_i$ adalah data, dengan $i = 1,2,\cdots,l \cdot$
   l adalah banyaknya data.
2. Pola $\vec{x}_i$ yang memenuhi persamaan (2) termasuk kelas -1,
   $\vec{w} \cdot \vec{x}_i + b \leq -1$ (2)
   Sedangkan pola $\vec{x}_i$ yang memenuhi persamaan (3) termasuk kelas +1,
   $\vec{w} \cdot \vec{x}_i + b \geq -1$ (3)
3. Maksimalkan nilai jarak antar *hyperplane* dan titik terdekat,
   $\frac{1}{\|\vec{w}\|}$ (4)
4. Hitung titik minimal (*Quadratic Programming*),
   $\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\|^2$ (5)
   $y_i(\vec{x}_i \cdot w + b) - 1 \geq 0, \forall_i$ (6)
   y_i adalah label data, dengan $y_i \epsilon \{-1, +1\}$ untuk $i = 1,2,\cdots,l \cdot$
5. Hitung nilai *Lagrange Multiplier*,
   $L(\vec{w}, b, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{1} a_i \left(y_i((\vec{x}_i \cdot \vec{w} + b) - 1)\right)$(7)
   $(a_i \geq 0, i = 1,2,\cdots,l)$
6. Minimalkan L terdahap $\vec{w}$ dan b, serta maksimalkan L terhadap a_i,
   $\sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \vec{x}_i \cdot \vec{x}_j$ (8)
   $\sum_{i=1}^{l} a_i y_i = 0$
7. Data yang berkorelasi dengan a_i yang positif disebut *support vector*.

3. METODE

Penelitian ini menggunakan dataset dari *webserver* RCSB Protein Data Bank dalam bentuk pdb yang diakses secara online melalui *http://www.rcsb.org.html*. Analisis atribut pada protein menggunakan *webserver* LISE, yaitu dengan mengupload file pdb ke *webserver* LISE.

Dataset ini terdiri dari 14 data dari lima jenis protein berbeda yang digunakan, seperti yang disajikan dalam Tabel 1. Dalam prediksi *binding site* protein ligan perlu mempertimbangkan segi kimia dan biologi dari protein, baik sebelum ataupun sesudah proses klasifikasi. Oleh sebab itu dilakukan pemisahan data berdasarkan jenis protein untuk mengatasi masalah perbedaan karakter masing-masing jenis protein, energi interaksi, dan *sequence* pada protein.

Tiap jenis protein dibagi kembali atas data *training* dan *testing*. Untuk pembagian kelompok data dari 14 data yang ada, dilakukan dengan rasio data untuk *oxidoreductase* 3:1, *ligase* 1:1, *transferase* 4:1, dan *hydrolase* 5:1. Setiap data protein pernah menjadi data *training* maupun data *testing*. Dengan kata lain data yang

telah dikelompokkan berdasarkan jenisnya, pada satu jenis protein diambil 1 protein untuk *testing* dan sebanyak sisa data protein dalam satu jenis tersebut digunakan untuk proses *training*.

Proses *testing* dimaksudkan untuk memprediksi *binding site* protein-ligan pada suatu protein. Selain itu, data *training* diambil dalam jumlah besar bertujuan agar *classifier* yang dilatih dapat menemukan pola pemetaan yang cukup akurat. Jika data *training* terlalu sedikit, dikhawatirkan *classifier* tersebut kurang mampu melakukan generalisasi, sehingga performa yang diberikan akan kurang baik saat dipakai untuk mengenali data pada *testing set.*

Tabel 1. Jenis dan Ukuran Data Protein

No	Jenis Protein	PDB ID	Ukuran Data	Banyak Data Positif	Banyak Data Negatif
1	Oxidoredu-ctase	3D4P	6.204	988	5.216
2		2WLA	2.444	844	1.600
3		1A4U	4.828	737	4.091
4	Ligase	1U7Z	6.144	731	5.413
5		1ADE	10.793	805	9.988
6	Transferase	2GGA	4.146	316	3.830
7		1SQF	4.365	934	3.431
8		1G6C	4.504	724	3.780
9		1BJ4	4.205	477	3.728
10	Hydrolase	4TPI	3.042	776	2.266
11		2V8L	2.060	1.030	1.030
12		1WYW	3.398	860	2.538
13		1RN8	2.235	918	1.317
14		1C1P	4.797	705	4.092

Alur Penelitian

Rancangan penelitian disusun dan disajikan dengan diagram alur pada Gambar 1. Studi literatur meliputi Protein dan karakteristik *binding site*, PDB (*Protein Data Bank*), *Machine Learning*, dan SVM (*Support Vector Machine*). Selanjutnya pengambilan data, data diambil dari *webserver* RCSB *Protein Data Bank*. Data tersebut selanjutnya dianalisis dengan LISE dan akan diprediksi bagian *binding site* proteinnya dengan menggunakan SVM. Hasil perhitungan LISE kolom jarak titik grid ke atom ligan terdekat dan *grid score* selanjutnya dinormalisasi.

Metode normalisasi yang digunakan dalam penelitian ini adalah normalisasi *mapstd*.

$$y = (x - x_{mean}).\frac{y_{std}}{x_{std}} + y_{mean} \quad (3.1)$$

Selanjutnya dengan rasio data yang telah diatur sebelumnya, data dilakukan proses *training* dan *testing*. Data protein akan diklasifikasikan dalam kelas 1 untuk daerah *binding site* dan kelas 0 untuk daerah bukan *binding site*. Tabel *confusion matrix* disajikan pada Tabel 2.

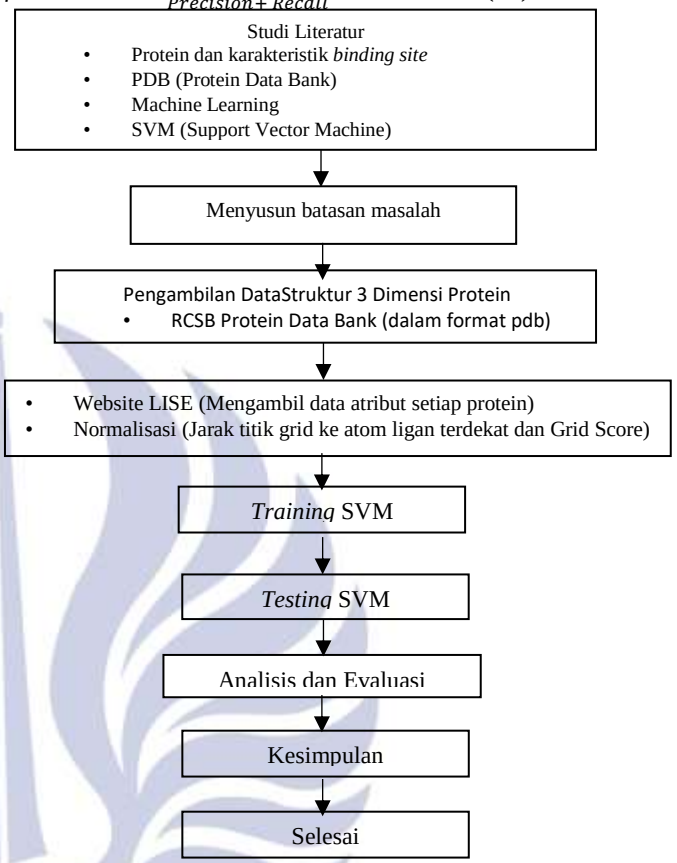
Sedangkan untuk mengukur performa SVM, dihitung akurasi, *precision, recall, f-measure, MCC, ROC area, PRC area,* dan waktu. Berikut adalah persamaan-persamaan yang digunakan:

$$\text{Akurasi} = \frac{BB + SS}{BB+BS+SB+SS} \quad (9)$$

$$Precision = \frac{BB}{BB+BS} \quad (10)$$

$$Recall = \frac{BB}{BB+SB} \quad (11)$$

$$f\text{-}measure = 2.\frac{Precision \times Recall}{Precision + Recall} \quad (12)$$



Gambar 1. Diagram Alur Penelitian

Tabel 2. *Confusion Matrix*

		Target	
		Benar	Salah
Prediksi	Benar	BB	BS
	Salah	SB	SS

4. PEMBAHASAN

Hasil *training* dan *testing* pencarian *binding site* protein ligan pada tiap data protein disajikan dalam Tabel 3 dan Tabel 4. Dari Tabel 3 dan Tabel 4 diperoleh rata-rata *training* untuk akurasi 99,02%, *precision* 99,04%, *recall* 99,02%, *f-measure* 99,02%, *MCC* 96,84%, *ROC* 92,82%, *PRC* 93,71%, dan rata-rata waktu *training* 18,92 detik. Serta didapat rata-rata akurasi *testing* sebesar 95,98 %.

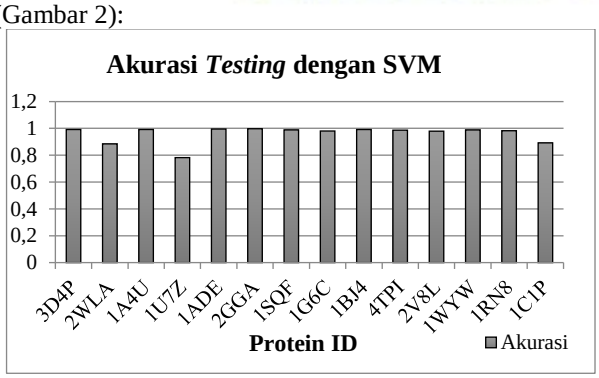
Tabel 3. Hasil *training* dengan SVM

No	Jenis Protein	PDB ID	Nilai							
			Akurasi	***Precision***	***Recall***	***F-Measure***	***MCC***	***ROC Area***	***PRC Area***	**Waktu**
1	Oxidore-ductase	1A4U 2WLA	0,9900	0,9900	0,9900	0,9900	0,9710	0,9770	0,9820	25
2		1A4U 3D4P	0,9920	0,9920	0,9920	0,9920	0,9700	0,9740	0,9850	13
3		2WLA 3D4P	0,9899	0,9900	0,9900	0,9900	0,9700	0,9650	0,8860	7
4	Ligase	1ADE	0,9956	0,9960	0,9960	0,9960	0,9680	0,8200	0,9540	11
5		1U7Z	0,9926	0,9930	0,9930	0,9930	0,9650	0,6080	0,5650	13
6	Transfe-rase	1SQF, 1G6C 1BJ4	0,9872	0,9870	0,9870	0,9870	0,9690	0,9770	0,9780	23
7		2GGA, 1G6C 1BJ4	0,9927	0,9930	0,9930	0,9930	0,9650	0,9690	0,9860	16
8		2GGA,1SQF 1BJ4	0,9921	0,9920	0,9920	0,9920	0,9660	0,9710	0,9850	21
9		2GGA, 1SQF 1G6C	0,9938	0,9940	0,9940	0,9940	0,9700	0,8570	0,9460	10
10	Hydrola-se	2V8L, 1WYW 1RN8, 1C1P	0,9876	0,9880	0,9880	0,9880	0,9700	0,9780	0,9790	24
11		4TPI, 1WYW 1RN8, 1C1P	0,9887	0,9890	0,9890	0,9890	0,9690	0,9770	0,9800	26
12		4TPI, 2V8L 1RN8, 1C1P	0,9872	0,9870	0,9870	0,9870	0,9690	0,9770	0,9780	27
13		4TPI, 2V8L 1WYW, 1C1P	0,9883	0,9890	0,9880	0,9880	0,9690	0,9770	0,9800	24
14		4TPI, 2V8L 1WYW, 1RN8	0,9853	0,9860	0,9850	0,9850	0,9670	0,9690	0,9360	25
Rata-rata			**0,9902**	**0,9904**	**0,9902**	**0,9902**	**0,9684**	**0,9282**	**0,9371**	**18,92**

Tabel 4. Hasil *testing* dengan SVM

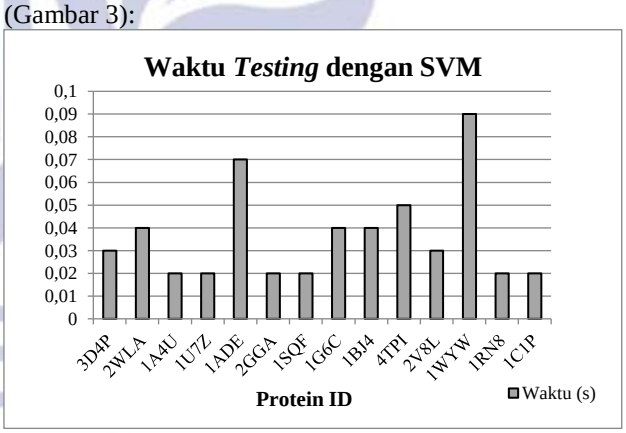
Type of protein	Test	Akurasi	Waktu (s)
Oxidoreductase	3D4P	0,9919	0,03
	2WLA	0,8848	0,04
	1A4U	0,9920	0,02
Ligase	1U7Z	0,7826	0,02
	1ADE	0,9956	0,07
Transferase	2GGA	0,9978	0,02
	1SQF	0,9890	0,02
	1G6C	0,9808	0,04
	1BJ4	0,9919	0,04
Hydrolase	4TPI	0,9871	0,05
	2V8L	0,9796	0,03
	1WYW	0,9888	0,09
	1RN8	0,9829	0,02
	1C1P	0,8924	0,02
Rata-rata		**0,9598**	**0,03642**

Berikut adalah grafik akurasi *testing binding site* protein ligan dari tiap data protein menggunakan SVM (Gambar 2):



Gambar 2. Grafik akurasi *testing binding site* protein ligan menggunakan SVM

Berikut adalah grafik waktu *testing binding site* protein ligan dari tiap data protein menggunakan SVM (Gambar 3):



Gambar 3. Grafik waktu *testing binding site* protein ligan menggunakan SVM

5. PENUTUP

Simpulan

Berdasarkan hasil penelitian, dapat dilihat bahwa SVM memiliki akurasi yang tinggi saat diaplikasikan dalam prediksi *binding site* protein ligan, yaitu dalam 14 data protein yang diuji menghasilkan rata-rata akurasi *test* sebesar 95,98 % dan rata-rata waktu *test* sebesar 0,03642 detik. Sehingga, metode SVM dapat dipertimbangkan sebagai langkah dalam menyelesaikan masalah prediksi *binding* site protein ligan untuk *drug design*.

DAFTAR PUSTAKA

Aparoy, P., Reddy, K. K., Reddanna, P. (2012). Structure and Ligand Based Drug Design Strategies in the Development of Novel 5- LOX Inhibitors. *Journal of Current Medicinal Chemistry*, 19(22), 3763–3778.

Capra, J. A., Singh, M. (2007). Predicting Functionally Important Residues From Sequence Conservation. *Journal of Bioinformatics*, 23 (15),1875–1882.

Carrasco, J P. C., Parra, T. C., Tudela, B. I., Luna, A. J. B., Ghasemi, F., Meseguer, J. M. V., Luque, I., Azam, S. S., Henden, S. T., Sanchez, H. P. (2016). Application of Computational Drug Discovery Techniques for Designing New Drugs against Zika Virus. *Journal of Drug Des*, 5 (2), 1-2.

Gu, W. G., Zhang, X., Yuan, J. F. (2014). Anti HIV Drug Development Through Computational Methods. *Journal of The AAPS*. 1-8.

Horton, H., Moran, L., Scrimgeour, K., Perry, M. (2011). *Study Guide for Principles of Biochemistry*. Fifth Edition. North Carolina: Pearson Education.

Liu, B., Wang, X., Lin , L. Tang, B., Dong, Q., Wang, X. (2009). Prediction of Protein *Binding site* in Protein Structures Using Hidden Markov Support Vectore Machine. *Journal of BMC Bioinformatics*, 10 (381), 1-14.

Mahdiyah, U. (2015). *Integrasi Seleksi Data dan Extreme Learning Machine (ELM) untuk Prediksi Binding site Protein-Ligan.* Tesis tidak diterbitkan. Surabaya: PPs Institut Teknologi Sepuluh Nopember.

Mahdiyah, U. (2017). Pencarian Rongga Berpotensi *Binding site* pada Protein dengan Menggunakan Support Vector Machine (SVM). *Jurnal Matematika, 1*4 (2), 1–13.

Mahdiyah, U., Irawan, M. I., Imah, E. M., S. (2016). Integrating Data Selection and Extreme Learning Machine to Predict Protein-Ligand *Binding site*. *Journal of Contemporary Engineering Science*, 9 (16), 791–797.

Ng, S. K., Wong, L. (2004). Accomplishments and Challenges in Bioinformatics. Dalam IT Pro, Januari. Singapura.

Oh, M., Joo, K., Lee, J. (2009). Protein-*Binding site* Prediction Based on Three-Dimensional Protein Modeling. *Journal of Proteins: Structure, Function and Bioinformatics*, 77 (9), 152–156.

White, A. W., Westwell, A. D., Brahemi, G. 2008. Protein-protein Interactions as Targets for Small Molecule Therapeutics. *Journal of Expert Reviews in Molecular Medicine*, 10 (8), 1-14.

Wong, G. Y., Leung, F. H. F., Ling, S. H. (2013). Predicting Protein-Ligand *Binding site* Using Support Vector Machine with Protein Properties. *Journal of Transaction on Computational Biology and Bioinformatic*, 10 (6), 1517–1529.

Zhou, S., Zhong, W. (2017) . Drug Design and Discovery: Principles and Applications. *Journal of Molecules,*22 (279), 1-6.