J. Math. and Its Appl.
E-ISSN: 2579-8936
P-ISSN: 1829-605X
Vol. 14, No. 2, Desember 2017, 89-101

Pencarian Rongga Berpotensi Binding Site pada Protein dengan Menggunakan Support Vector Machine (SVM)

Umi Mahdiyah[1]

Universitas Nusantara PGRI Kediri, Jl Ahmad Dahlan, Mojoroto Gg I, Kota Kediri, umimahdiyah@gmail.co m

Abstrak

Bioinformatika merupakan ilmu multidisipliner yang melibatkan berbagai bidang ilmu. Salah satu aplikasi dari bioinformatika adalah dalam proses desain obat berbantuan komputer. Dalam desain obat berbantuan komputer salah satu langkah awal yang dibutuhkan adalah mencari suatu rongga pada protein, rongga tersebut nantinya untuk melekat suatu ligan (partikel kecil) maupun protein yang merupakan partikel dari calon obat. Dalam penelitian ini untuk pencarian rongga delakukan dengan menggunakan metode klasifikasi dengan *Support Vector Machine*. Hasil dari pencarian rongga dengan metode ini menunjukkan akurasi *G-Mean* yang cukup tinggi yaitu 0,903 atau 90,3
Katakunci: *Bioinformatika, binding site, protein, SVM*

1 Pendahuluan

Bioinformatika merupakan ilmu multidisipliner yang melibatkan berbagai bidang ilmu, yang meliputi biologi molekuler, matematika, ilmu komputasi, kimia molekul, fisika, dan beberapa disiplin ilmu lainnya [8]. Selama ini, bioinformatika banyak diaplikasikan dalam berbagai masalah, salah satunya adalah masalah desain obat. Dalam desain obat berbantuan komputer salah satu langkah awal yang dibutuhkan adalah mencari suatu rongga pada protein, rongga tersebut nantinya untuk melekat suatu ligan (partikel kecil) maupun protein yang merupakan partikel atau protein dari calon obat.

Machine Learning sudah banyak digunakan pada data bioinformatika [6] serta menunjukkan hasil yang baik untuk prediksi *binding site* [7]. Penelitian bioinformatika dengan *machine learning* yang telah digunakan, diantaranya adalah "*Fast prediction of protein–protein interaction sites based on Extreme Learning Machines*" adalah penelitian yang dilakukan Debby D. Wang dkk[9]. Dalam paper tersebut dilakukan pencarian rongga pada permukaan protein sebagai tempat melekat protein yang lain. Selanjutnya, "*Protein Sequence Classification Using Extreme Learning Machine*" merupakan penelitian yang dilakukan oleh Dianhui Wang dan Guang-Bin Huang [10], penelitian tersebut mengklasifikasikan *sequence* protein dengan *Extreme Learning Machine*.

Support Vector Machine merupakan salah satu *machine learning* yang bagus. SVM adalah metode learning machine yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada *input space*, sehingga SVM ini dapat melihat atau mentransformasi suatu objek ke dimensi yang lebih tinggi.

Prediksi dari *binding site* dapat dirumuskan sebagai masalah klasifikasi biner, yaitu untuk membedakan lokasi *binding site* dan bukan *binding site*. Sehingga dalam penelitian ini digunakan SVM sebagai algoritma klasifikasi untuk pencarian binding site pada suatu protein.

2 TINJAUAN PUSTAKA

2.1 Protein

Protein adalah salah satu bio-makromolekul yang penting peranannya dalam makhluk hidup. Protein tersusun dari 20 macam asam amino alami. Asam amino penyusun protein mengandung beberapa atom kimia seperti carbon(C), nitrogen (N), dan hidrogen (H), kecuali *cyteine* dan *methionine* juga mengandung sulfur (S).

Konsep desain obat pada bioinformatika didasarkan pada fungsionalitas dari protein, yaitu pencarian sebuah senyawa untuk mengaktifkan atau

menghambat fungsi protein target, sebab protein dapat mengekspresikan suatu informasi genetik. Sebagaimana terdapat ribuan gen di dalam inti sel, masing-masing mencirikan satu sifat nyata dari organisme, di dalam sel terdapat ribuan jenis protein yang berbeda, masing-masing membawa fungsi spesifik yang ditentukan oleh gen yang sesuai [13].

2.2 Rongga pada Permukaan Protein (Binding Site)

Binding site adalah bagian permukaan *reseptor* (protein) yang berfungsi untuk melekatnya suatu obat. *Binding site* protein-ligan umumnya berada di *pocket* (celah, galur) pada permukaan protein[4]. Penentuan *pocket* merupakan langkah penting menuju desain obat dalam penemuan senyawa baru.

2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang *pattern recognition*. Sebagai salah satu metode *pattern recognition*, usia SVM terbilang masih relatif muda. Walaupun demikian, evaluasi kemampuannya dalam berbagai aplikasinya menempatkannya sebagai *state of the art* dalam *pattern recognition*, dan dewasa ini merupakan salah satu tema yang berkembang dengan pesat. SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada *input space*, usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM. [5]

Data yang tersedia dinotasikan sebagai $\vec{x} \in \Re^d$ sedangkan label masing-masing dinotasikan $y_i \in \{-1,+1\}$untuk $i = 1,2,\dots,l$, yang mana l adalah banyaknya data. Diasumsikan kedua *class* -1 dan +1dapat terpisah sempurna oleh *hyperplane* berdimensi d, yang didefinisikan

$$\bar{w}\bar{x} + b = 0 \qquad (1)$$

Pattern $\bar{x}_i$ yang termasuk class -1 dapat dirumuskan sebagai pattern yang memenuhi pertidaksamaan

$$\bar{w}\bar{x} + b \leq -1 \quad (2)$$

Sedangkan pattern $\bar{x}_i$ yang termasuk class +1

$$\bar{w}\bar{x} + b \geq -1 \quad (3)$$

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara *hyperplane* dan titik terdekatnya, yaitu $1/\|\bar{w}\|$. Hal ini dapat dirumuskan sebagai *Quadratic Programming* (QP) problem, yaitu mencari titik minimal, dengan memperhatikan .

$$\min_{\bar{w}} \tau(w) = \frac{1}{2}\|\bar{w}\|^2 \quad (4)$$

$$y_i(\bar{x}_i.w + b) - 1 \geq 0, \forall i$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, di antaranya *Lagrange Multiplier*.

$$L(\bar{w}, b, \alpha) = \frac{1}{2}\|\bar{w}\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i\left((\bar{x}_i.\bar{w} + b) - 1\right)\right) \quad (5)$$

$$(i = 1,2, \dots, l)$$

α_i adalah *Lagrange multipliers*, yang bernilai nol atau positif $(a_i \geq 0)$. Nilai optimal dari persamaan di atas dapat dihitung dengan meminimalkan L terhadap $\bar{w}$ dan b ,dan memaksimalkan L terhadap α_i . Dengan memperhatikan sifat bahwa pada titik optimal *gradient* $L = 0$, persamaan di atas dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung α saja, sebagaimana persamaan (6).

$$\sum_{I=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \bar{x}_i \bar{x}_j \quad (6)$$

Dengan

$$\alpha_i \geq 0 (i = 1,2..)$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0$$

Dari hasil dari perhitungan ini diperoleh a_i yang kebanyakan bernilai positif. Data yang berkorelasi dengan a_i yang positif inilah yang disebut sebagai *support vector*.[1]

3 Metode Penelitian

3.1 Pengambilan Data

Data yang digunakan dalam penelitian ini merupakan data eksperimental, yaitu data protein yang didapat dari webserver RCSB Protein Data Bank. Selanjutnya dari data PDB dicari nilai setiap atributnya dengan menggunakan program online LISE[11]. Dalam program LISE tersebut dapat diperoleh 2 atribut atau variabel, yaitu nilai *score conservation* dan nilai *potential energy*.

3.2 Data dan Pembagian Data

Data yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

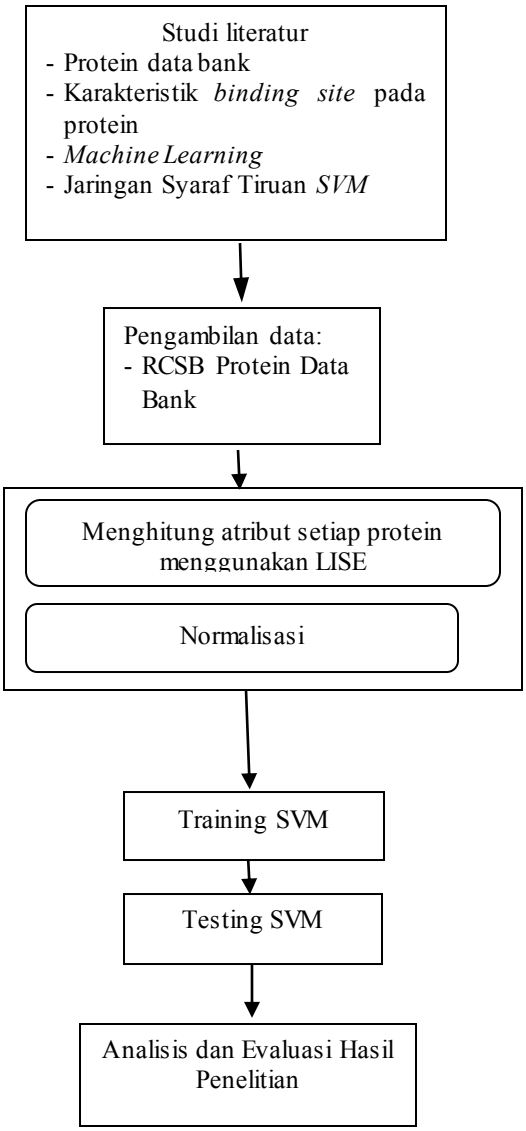
Tabel 1 Data Protein berdasarkan jenisnya

No	Protein ID	Banyak Data +	Banyak Data -
1	4TPI	776	2266
2	2ZAL	688	4598
3	2V8L	1030	1030
4	1WYW	827	8789
5	1RN8	918	1315
6	1C1P	704	4093
7	3D4P	988	5216
8	1A4U	737	3926
9	2WLA	846	1598
10	2GGA	316	3830
11	1SQF	934	3431
12	1O26	1332	7940
13	1G6C	724	3780
14	1BJ4	477	3728
15	1U7Z	731	5413
16	1ADE	805	8267

Untuk membagi data *training* dan data *testing* dalam penelitian ini digunakan 5-*fold cross validation* artinya 80% data untuk *training* dan 20% *testing*.

3.3 Alur Penelitian

Alur penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Alur Penelitian

Studi literatur tentang Protein data bank, karakteristik *binding site* pada protein, *Machine Learning,* Jaringan Syaraf Tiruan *SVM*, dilakukan untuk memahami tenang protein, *binding site*, dan SVM. Setelah itu, dilakukan pengambilan data protein pada web RCSB Protein Data Bank. Selanjutnya atribut yang dibutuhkan untuk klasifikasi didapat pada web LISE dengan menginputkan protein ID atau mengupload file .pdb yang diperoleh sebelumnya.

Data dengan atributnya yang diperoleh dari LISE dinormalisasi dengan normalisasi minmax. [12]

$$x_{normalized} = \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \times \left(\max_{new} - \min_{new} \right) \right) + \min_{new} \qquad (7)$$

Selanjutnya dari data yang ada dilakukan *5-fold cross validation* untuk dilakukan training dan testing, kemudian dianalisis. Untuk mengukur performa dari klasifikasi data pengujian *imbalance* dalam hal ini digunakan *confusion matrix*, *precision* dan *recall*, *specificity*, dan *G-mean*[3]. Tabel *confusion matrix* Dapat dilihat pada tabel 2, sedangkan untuk pengukuran performa seperti pada Persamaan 8 sampai persamaan 11.

Tabel 2. *Confusion Matrix*

		Nilai Sebenarnya	
		True	*False*
Prediksi	*True*	TP *(True Positive)*	FP *(False Positive)*
	False	FN *(False Negative)*	TN *(True Negative)*

Precision adalah presentase dari data yang diprediksi benar oleh *classifier* yang bernilai benar.

$$precision = \frac{TP}{TP + FP}. \qquad (8)$$

Recall adalah porsi dari data sampel yang diprediksi benar oleh *classifier*.

$$recall = \frac{TP}{TP + FN}. \qquad (9)$$

$$sensitifity = recall$$

Geometric mean telah digunakan beberapa peneliti untuk mengevaluasi *classifier* pada dataset yang *imbalanced* . *G-mean* mengindikasikan keseimbangan antara kinerja klasifikasi pada kelas mayoritas dan minoritas. Ukuran *G-mean* diambil berdasarkan s*ensitifity* (akurasi dari data positif) dan s*pecificity* (akurasi data negatif). [3]

$$specificity = 1 - \frac{TP}{FP + TN} \tag{10}$$

$$G - mean = \sqrt{sensitivity \times specificity} \tag{11}$$

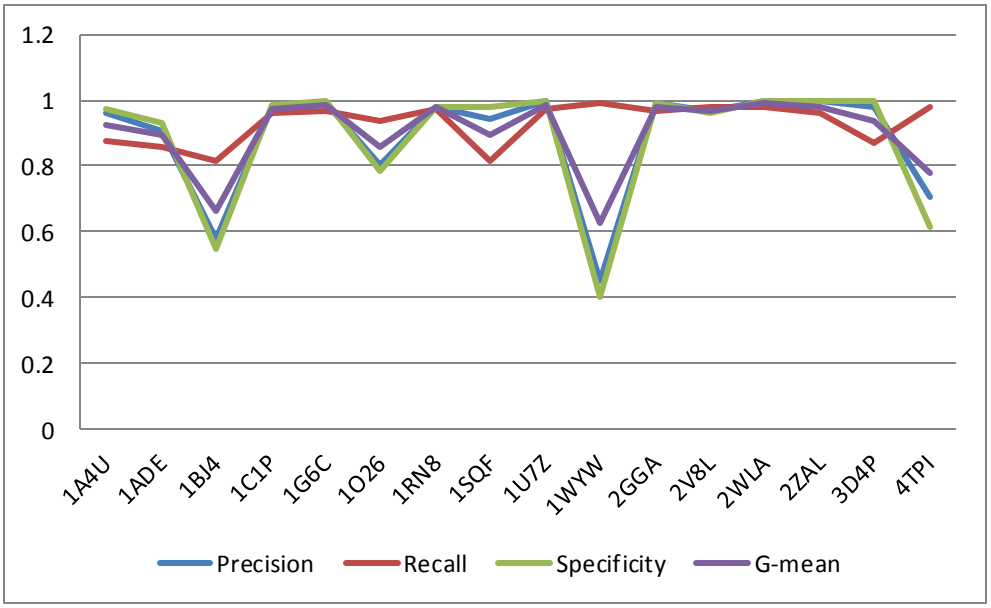
4 Hasil dan Pembahasan

Dalam penelitian ini, data yang digunakan merupakan data data protein hasil eksperimen yang dipublish dalam web RCSB Protein data bank, yang juga merupakan data open source. Data protein dengan *binding site* nya adalah data yang mempunyai karakter *imbalance*, sehingga untuk mengkur performa algoritma yang digunakan untuk klasifikasi tidak dapat menggunakan perhitungan akurasi biasa. Dalam penelitian ini ukuran performa dilihat dari *precision*, *recall*, *specificity*, dan *G-mean*. Hasil dari penelitian ini dapat dilihat pada Tabel 3, sedangkan grafik dari hasil penelitian dapat dilihat pada Gambar 2.

Tabel 3. Hasil Penelitian

Protein ID	Nilai					
	Precision	Recall	Specificity	G-mean	CPU Time (Training)	CPU Time (Testing)
1A4U	0.9586	0.87585	0.9741	0.92369	12.090	0.0676
1ADE	0.9063	0.85934	0.9312	0.89453	146.014	0.0546
1BJ4	0.5756	0.81289	0.5452	0.66572	19.032	0.0312
1C1P	0.9809	0.96076	0.9844	0.97249	10.816	0.0546
1G6C	0.9930	0.96616	0.9981	0.98202	18.876	0.0585
1O26	0.8036	0.93525	0.7815	0.85493	18.447	0.0520
1RN8	0.9777	0.97222	0.9815	0.9765	11.050	0.0338
1SQF	0.9420	0.81263	0.9772	0.89111	13.416	0.0546
1U7Z	0.9971	0.97538	10.000	0.98761	17.940	0.0832
1WYW	0.4497	0.98864	0.3991	0.62814	12.610	0.0754
2GGA	0.9890	0.96756	0.9907	0.97908	19.071	0.1755
2V8L	0.9690	0.98026	0.9578	0.96895	11.128	0.1170
2WLA	0.9970	0.97674	10.000	0.9883	12.948	0.0364
2ZAL	0.9949	0.96076	10.000	0.98018	12.766	0.0442
3D4P	0.9794	0.87247	0.9996	0.93388	58.656	0.0936

Protein ID	Nilai					
	Precision	Recall	Specificity	G-mean	CPU Time (Training)	CPU Time (Testing)
4TPI	0.7073	0.98174	0.6133	0.77593	10.764	0.1326
Rata-rata	**0.891959**	**0.93311**	**0.8866**	**0.9032**	**2.4443**	**0.0784**



Gambar 2. Grafik hasil penelitian

Dari tabel dan grafik di atas dapat dilihat bahwa pencarian rongga pada protein dengan metode klasifikasi menggunakan *Support Vector Machine* mempunyai nilai akurasi yang cukup tinggi. Hal tersebut dapat dilihat bahwa rata-rata dari hasil pengukuran akurasi, hampir di semua data protein mempunyai akurasi baik *precision, recall, specificity,* maupun *G-Mean* lebih dari 60% atau 0,6.

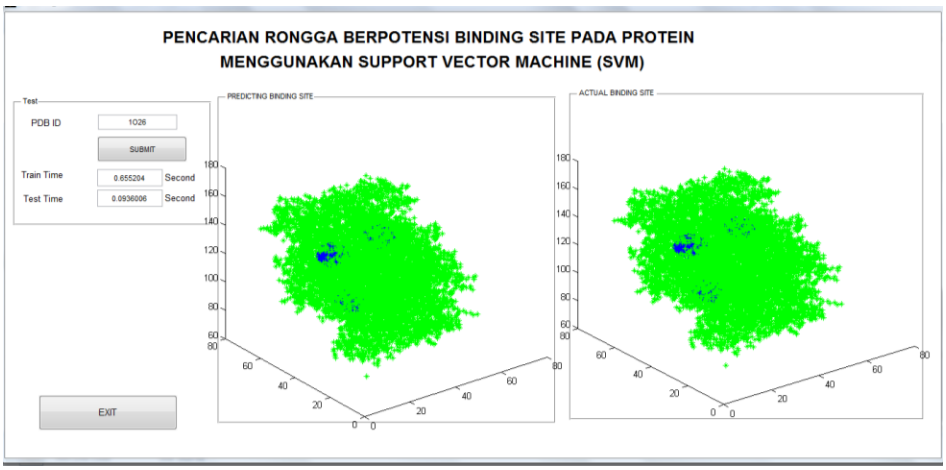
Precision dari pencarian rongga pada protein dengan metode klasifikasi menggunakan *Support Vector Machine* memiliki rata-rata 0.891959. *Recall, specificity,* dan *G-Mean* memiliki rata-rata berturut turut 0.93311, 0.8866, 0.9032.

Pada Gambar 2 dapat dilihat bahwa pada beberapa protein memiliki perbedaan akurasi yang sangat signifikan. Berdasarkan pengamatan terhadap

jenis nya hal tersebut terjadi karena data yang diambil dan pembagian data *training* serta *testing* dalam penelitian ini tidak memperhatikan jenis proteinnya. Sehingga untuk data training dan data testing di beberapa kasus dilakukan training dan testing untuk jenis protein yang berbeda.

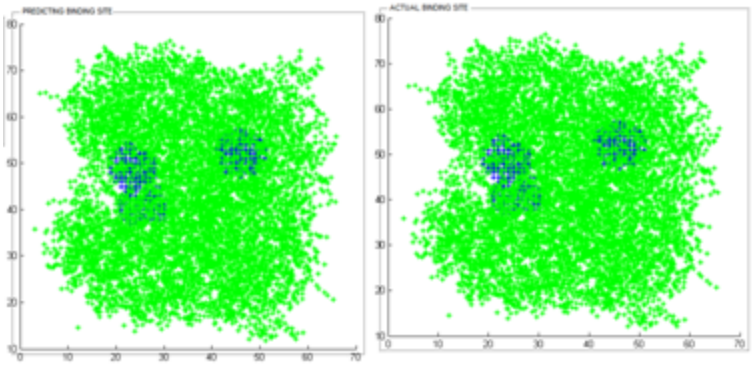
4.1 Tampilan Program

Gambar 2 merupakan tampilan dari program untuk pencarian rongga pada permukaan protein secara 3 dimensi atau dilihat dari sumbu x, y, z. Warna hijau menunjukkan permukaa protein, sedangkan warna biru menunjukkan lokasi rongga tersebut. Sebelah kanan (*actual binding site*) merupakan gambar asli dari protein dan rongganya, sedangkan sebelah kiri (*predicting binding site*) merupakan hasil pencarian dengan klasifikasi SVM.

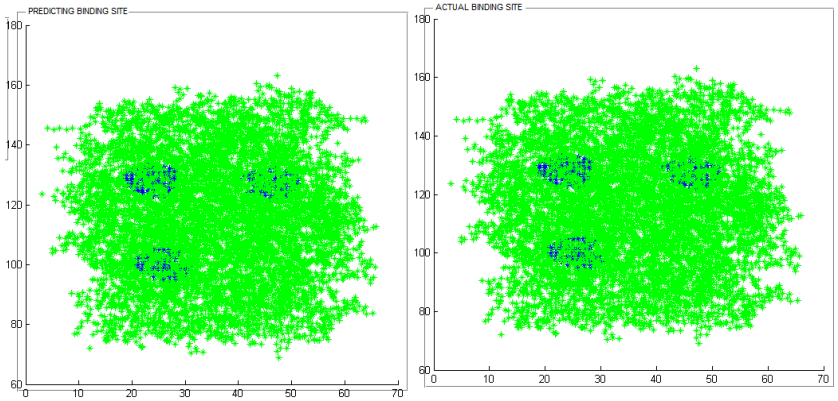


Gambar 3. Tampilan Program dengan contoh running protein ID 1O26

Gambar 4 merupakan tampilan protein dan rongganya dilihat secara dua dimensi dari sumbu x dan y. Dari gambar tersebut dapat dilihat bahwa hanya terdapat 2 rongga pada permukaan protein 1O26. Sedangkan pada Gambar 5 jika dilihar dari sumbu x-z terlihar terdapat 3 rongga protein.



Gambar 4. Tampilan protein dilihat secara 2 dimensi dari sumbu x-y



Gambar 5. Tampilan protein dilihat secara 2 dimensi dari sumbu x-z

5 Kesimpulan

Dari paparan hasil penelitian dan analisis data dapat dilihat bahwa SVM memiliki akurasi yang tinggi jika diaplikasikan untuk pencarian rongga pada permukaan protein. Rata-rata dari *Precision, Recall, specificity,* dan *G-Mean* memiliki rata-rata berturut turut 0.891959, 0.93311, 0.8866, 0.9032.

Dalam penelitian ini juga dapat dilihat bahwa untuk klasifikasi data protein harus diperhatikan berdasarkan jenis-jenisnya. Tidak semua data dapat digunakan data latih/*training* untuk keseluruhan protein

6 Pustaka

[1] Batuwitage, Manohara Rukshan Kannangara, "*Enhanced Class Imbalance Learning Methods for Support Vector Machines", Thesis of Doctor of Philosophy Hilary Term* 2010, St. Cross College, 2010.

[2] Bekkar, Mohamed, dan Taklit Akrouf Alitouche, "*Imbalanced Data Learning Approaches Review", International Journal of Data Mining & Knowledge Management Process* (IJDKP), Vol. 4, No. 4, hal. 15-33, 2013.

[3] H. He, E.A. Garcia, "*Learning from imbalanced data*". *IEEE Trans. Knowl. Data Eng*, Vol. 21, no.9, hal. 1263-1284,2009.

[4] Hendlich, Manfred, Rippmann, Friedrich dan Gerhard Barnickel, "LIGSITE: Automatic and efficient detection of potential small molecule-*binding site*s in proteins", *Journal of Molecular Graphics and Modelling*, Vol.15, hal. 359 –363, , (1997),

[5] Nugroho, SA, Witarto, AB, Handoko, D, *"Application of Support Vector Machine in Bioinformatics", Proceeding of Indonesian Scientific Meeting in Central Japan,* December 20, 2003.

[6] Mahdiyah, Umi, Irawan, Isa, dan Imah,EM, "*Study Comparison Backpropogation, Support Vector Machine, and Extreme Learning Machine for Bioinformatics Data", Journal of Computer Science and Information*, Vol 8: No 1, 53-59, 2015a.

[7] Mahdiyah, Umi, Imah,EM, dan Irawan, "*Integrating Data Selection and Extreme Learning Machine to Predict Protein-Ligand Binding Site", Contemporary Engineering Sciences,* Vol. 9, no. 16, 791 – 797,2016.

[8] Shen, Shiyi dan Jack A. Tuszynski, *Theory and Mathematical Methods for Bioformatics*, Springer, Verlag Berlin Heidelberg, 2008.

[9] Wang, Debby D., Wang, Ran dan Hong Yan, "*Fast prediction of protein–protein interaction sites based on Extreme Learning Machines"*, *Neurocomputing*,Vol. 128, hal. 258–266, 2014.

[10] Wang, Dianhui dan Guang-Bin Huang, "*Protein Sequence Classification Using Extreme Learning Machine", Proceedings of International Joint Conference on Neural Networks,* Montreal, Canada,hal. 1406-1411, 2005.

[11] Xie, Z.R. and Hwang, M.J. (2012) Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics, 28, 1579-1585.

[12] Zhu, Chengzhang, Yin, Jianping dan Qian Li, *"A Stock Decision Support System Based on ELM", Proceedings of the International Conference on Extreme Learning Machines (ELM2013)*, (eds) Sun, F., Toh, K.-A., Romay, M.G., Mao, K., Beijing, hal.67-79, 2013.

[13] Zvelebil, Marketa dan Jeremy O. Baum, "*Understanding Bioinformatics"*, Garland Science, New York, 2008.