PENJAJARAN SEKUEN JAMAK PROTEIN DENGAN ALGORITMA GENETIKA

Ceerinda Macro Jingga[1], Lailil Muflikhah[2], Marji[3]

[1] Mahasiswa Program Studi Teknik Informatika Universitas Brawijaya
[2] Dosen Program Studi Teknik Informatika Universitas Brawijaya
[3] Dosen Program Studi Teknik Informatika Universitas Brawijaya
Email: [1]ceerinda@gmail.com, [2]lailil@ub.ac.id, [3]marji@ub.ac.id

Abstrak

Penjajaran sekuen adalah proses dimana sekuen dibandingkan dengan mencari pola karakter yang paling umum dan berhubungan antar sekuen. Penjajaran sekuen jamak adalah proses penjajaran lebih dari dua sekuen yang merupakan dasar dari pencarian kemiripan pada database dan penjajaran banyak sekuen (*multiple sequence alignment*). Penjajaran sekuen memiliki masalah dalam hal waktu karena semakin panjang sebuah sekuen maka dibutuhkan waktu yang semakin banyak, selain itu sulit mengoptimasi atau sering disebut dengan masalah kombinatorial. Salah satu metode yang dapat menyelesaikan masalah ini adalah algoritma genetika. Algoritma genetika membentuk populasi dari solusi acak lalu menggunakan konsep seleksi alami, *crossover* dan mutasi yang digunakan untuk mengembangkan solusi tersebut. Algoritma genetika berhasil menghasilkan suatu solusi untuk masalah optimasi yang sulit. Keuntungan yang ditawarkan oleh algoritma ini adalah mengoperasikan beberapa solusi dengan mengkombinasikan solusi terakhir. Pada penelitan sebelumnya jika menggunakan algoritma genetika, menghasilkan hasil yang optimal dan lebih cepat dibandingkan dengan metode pe*mr*ogaman dinamis serta algoritma genetika ini inheren parallel sehingga dapat diimplementasikan sangat efisien pada komputer secara parallel. Pada kasus penjajaran sekuen tanpa *gap*s dan adanya *gap*s algortima genetika ini menghasilkan solusi yang baik. Bioinformatika adalah bidang ilmu biologi dan ilmu komputer. Data molekuler di dalam ilmu biologi sangat besar dan terus berkembang termasuk database pada protein. Protein adalah salah satu biomolekul yang sangat penting pada manusia yang berperan dalam proses seluler, fungsi enzim, antibody, *hormone* dan transport molekul. Pada penelitian ini, kami menampilkan bagaimana algoritma genetika dapat membantu dalam menyelesaikan penjajaran sekuen jamak protein. Data sekuen yang digunakan pada pengujian sebanyak 4 dan panjang sekuen maksimal 36. Penghitungan nilai *fitness* yang digunakan adalah *sum of pairs* pada *blosum62* serta skema affine yang terdiri dari *gap* open bernilai -11 dan *gap* extension bernilai -1. Pada penelitian ini algoritma genetika mampu menghasilkan hasil optimasi yang baik dari permasalahan yang komplek dari penjajaran sekuen jamak protein.

Kata kunci: affine, algoritma genetika, bioinformatika, *blosum62*, penjajaran sekuen jamak, protein

Abstract

Sequence Alignment is the process whereby a sequence compared by looking for patterns of the most common character and inter-related sequences. Multiple sequence Alignment is the process of more than two sequences alignment that the basis of similarity search on the database. Multiple Sequence Alignment have a problem in terms of time because the longer a sequence so it takes more and more, but it is difficult to optimize or often called combinatorial problems. One method that can solve this problem is a genetic algorithm. Genetic algorithms form the population of random solutions and using the concept of natural selection, crossover and mutation are used to develop the solution. Genetic algorithms managed to produce a solution to a difficult optimization problems. The benefits offered by this algorithm is operating several solutions by combining the final solution. In the previous research has been generated if using a genetic algorithm, produce optimal results and faster than the dynamic programming method and the genetic algorithm is inherently parallel so that it can be implemented very efficiently on parallel computers. In the case of sequence alignment without gaps and the gaps genetic algorithms produce good solutions. Bioinformatics is a field the sciences and computer science. Molecular data in the science of biology very large and continued to grow including a database in proteins. Protein is one of biomolekul very important for human being who participate in the process cellular, function an enzyme, antibody, hormone and transport molecules. In this study, we show how a genetic algorithm can help in solving the multiple sequence alignment of protein. Data sequence used for testing is 4 and long sequence maximum 36 .The calculation of value fitness used is sum of pairs with blosum62 and the scheme of affine transformations consisting of gap open -11 and gap extension -1 .This research prove that algorithm genetics capable of produce results good optimize from trouble a complex of the multiple sequence alignment of protein

Keywords*: sequence alignment, genetic algorithms, bioinformatics, protein*

1. PENDAHULUAN

Bioinformatika merupakan perpaduan antara ilmu biologi dan ilmu komputer [1]. Bioinformatika membahas tentang pengolahan, analisis, prediksi, penyimpanan dan pencarian data biologi molekuler seperti DNA, RNA dan protein menggunakan komputer Salah satu cakupan bidang ilmu bioinformatika adalah tentang analisis protein. Protein merupakan senyawa yang ada di dalam suatu organisme yang terdiri dari asam amino. Protein merupakan senyawa yang ada di dalam suatu organisme yang terdiri dari 20 macam asam amino. Protein menjalankan sebagian besar proses yang ada di dalam sel [2]. Protein memiliki sekuen yakni karakter-karakter yang melambangkan asam amino pembentuknya. Sekuen ini bisa digunakan untuk mengetahui fungsi dari suatu protein. Salah satu teknik yang digunakan untuk memprediksi suatu sekuen protein adalah algoritma genetika.

Masalah yang didapat adalah ketika melakukan analisis terhadap sekuen protein tersebut. Analisis sekuen yang dimaksud meliputi pencarian sekuen yang mirip terhadap sekuen yang ada di database yang telah diketahui fungsinya *(sequence alignment)* dan pencarian fungsi pada sekuen protein serta pencarian motif dari beberapa sekuen protein (*multiple sequence alignment*) dan diasumsikan memiliki hubungan evolusi dimana diturunkan dari sekuen terdahulu sehingga bisa dsimpulkan memiliki fungsi yang sama. Pada protein dengan tingkat kemiripan yang tinggi maka dapat diasumsikan bahwa kedua protein tersebut memiiki fungsi yang sama. Pada 8 Oktober 2015 kumpulan data yang dimiliki oleh National Center for Biotechnology Information (NCBI) ada 51.933.925 [3] dan bertambah setiap harinya.

Pada penelitian yang dilakukan oleh Pankaj yang berjudul "Alignment of Multiple Sequences using GA method". Dalam penelitian ini peneliti sudah menggunakan berbagai metode *crossover*, mutasi dan seleksi skema untuk *multiple alignment*. Hasil setiap *alignment* cenderung membaik yang ditunjukkan adanya peningkatan pada nilai fitness dan peningkatan jumlah iterasi [4].

Berdasarkan paparan informasi sebelumnya, maka akan dibuat perancangan sistem "Penjajaran Sekuen Jamak Protein Dengan Algoritma Genetika" yang dapat melakukan proses algoritma genetika dan mencari nilai fitness terbaik dari kumpulan beberapa sekuen.

Tujuan dari penelitian ini adalah membuat rancangan aplikasi penjajaran sekuen jamak protein dengan algoritma genetika dan mencari nilai fitness terbaik dari kumpulan beberapa sekuen dan hasil penjajarannya.

Manfaat dari penelitian ini adalah dapat memahami bagaimana merancang aplikasi penjajaran sekuen jamak protein dengan algoritma genetika dan dapat memahami bagaimana menghitung nilai fitness dari penjajaran sekuen dengan menggunakan algoritma genetika.

2.PENJAJARAN SEKUEN

Penjajaran sekuen adalah proses penyusunan dua atau lebih sekuen. Penjajaran sekuen jamak adalah penjajaran lebih dari dua sekuen. Penjajaran ini bertujuan agar sekuen-sekuen tersebut dapat dikenali kemiripannya dengan protein yang ada di database. Ketidakcocokan dalam penjajaran diasosiasikan dengan proses mutasi, sedangkan kesenjangan (*gap*, tanda "–") diasosiasikan dengan proses insersi atau delesi[5]. Penjajaran sekuen memberikan hipotesis atas proses evolusi yang terjadi dalam sekuen-sekuen tersebut. Penjajaran juga menunjukkan posisi-posisi yang dipertahankan selama evolusi dalam sekuen-sekuen protein. Ini menunjukkan posisi-posisi tersebut penting dan bisa jadi fungsi dari protein tersebut.

2.2 Algoritma Genetika

Algoritma genetika adalah teknik optimasi komputasi yang diadaptasi dari fenomena evolusi alam spesies untuk menyelesaikan masalah optimasi. Individu secara terus – menerus mengalami perubahan agar dapat bertahan pada lingkungannya dalam proses evolusi. Pendekatan algoritma genetika didasarkan pada simulasi dalam proses evolusi pada populasi adalah solusi potensial yang berkembang menjadi solusi terbaik[6]. Kromosom akan berevolusi menghasilkan anak terbaik yang menjadi solusi terbaik dari sebuah permasalahan.

Semua individu pada algoritma genetika memiliki nilai *fitness*nya. Nilai *fitness* digunakan untuk mengukur seberapa baik individu (Mahmudy, 2013). Nilai *fitness* terbaik merupakan representasi dari individu terbaik yang akan menjadi solusi terbaik dari sebuah permasalahan.

Dalam *Genetic Algorithms and the Multiple Sequence Alignment Problem in Biology* menjelaskan metodologi algoritma genetika dapat diterapkan untuk menghasilkan solusi yang optimal atau mendekati optimal untuk masalah *Multiple Sequence Alignment*. Algoritma genetika menghasilkan solusi yang cukup baik [4].

2.2.1 Struktur Algoritma Genetika

I. Inisialisasi

Proses inisialisasi menciptakan populasi awal secara acak. Populasi tersebut berisi kumpulan individu (kromosom). Dalam kasus penjajaran sekuen ini, inisialisasi untuk pembangkitan populasi awal dilakukan dengan pengkodean permutasi berdasarkan pada sekuen protein. berikut contoh inisialisasi :

Preprocessing

Terdapat dua sekuen dengan panjang berbeda

>MTVK (panjang sekuen 4)
>MTVKARCILS (panjang sekuen 10)
>MVKRCI (panjang sekuen 6)

Pada sekuen dua terdapat sepuluh gen. jumlah gen tersebut dikalikan dengan 1,2[3] untuk menentukan maxLength

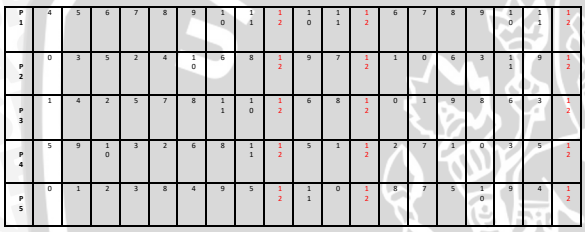
10 x 1.2 = 12

(populasi atau Popsize 5)

Setelah mendapatkan nilai maxlength, sekuen membentuk kromosom dengan panjang maxlength dengan mengisikan gen dengan *gap*

Sekuen 1	M	T	V	K	-	-	-	-	-	-	-	-
Sekuen 2	M	T	V	K	A	R	C	I	L	S	-	-
Sekuen 3	M	V	K	R	C	I	-	-	-	-	-	-
Index array	0	1	2	3	4	5	6	7	8	9	10	11

Langkah selanjutnya sekuen di*encode* dari index array yang berisi *gap* dan dibatasi dengan maxLength untuk inisialisai populasi.

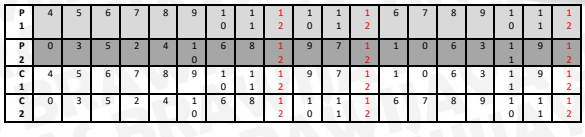


II. Reproduksi

Reproduksi merupakan proses yang dilakukan untuk menghasilkan keturunan (*offspring*) yang berasal dari induk kromosom yang dipilih secara random. Terdapat dua operator genetika dalam reproduksi yaitu *crossover* dan mutasi.

A. Tukar Silang (*Crossover*) dengan Metode *One-Cut Point*

Persilangan pada algoritma genetika melibatkan dua buah individu yang dipilih secara acak menjadi parent yang akan membentuk kromosom baru. Proses ini diawali dengan penentuan tingkat *crossover* (cr). Nilai ini menentukan rasio dari keturunan (*offspring*) yang dihasilkan dari proses *crossover*. Hasil *offspring* dari proses *crossover* diperoleh dengan mengalikan cr dengan ukuran populasi (*popsize*). Berikut contoh *crossover* :

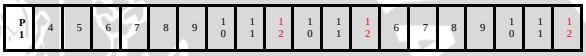


B. Mutasi dengan Metode Reciprocal Exchange Mutation

Mutasi merupakan operator algoritma genetika yang mengubah salah satu atau lebih gen pada individu. Proses mutasi diawali dengan penentuan nilai *mutation rate*(mr) yang akan menentukan rasio *offspring* yang dihasilkan dari proses mutasi. Anak (*offspring*) yang dihasilkan yaitu dengan mengalikan nilai mr dengan ukuran populasi (*popsize*). Contoh mutasi sebagai berikut :

III. Evaluasi

Pada proses ini dilakukan proses evaluasi terhadap nilai fitness dari setiap kromosom. Fungsi fitness adalah fungsi yang dimiliki oleh masing-masing individu yang menentukan tingkat kesesuaian individu tersebut dengan kriteria yang ingin dicapai. Pada tahap evaluasi dilakukan perhitung nilai fitness sum of pairs masing-masing sekuen dari table matriks Blosum62 dengan menggunakan gap penalty skema affine dengan nilai gap open -11 dan gap extension -1. Contoh perhitungan fitness pada C3 :



Angka dikembalikan dalam bentuk semula (decode) dan diisi dengan tanda *gap*

Sek 1	M	T	V	K	-	-	-	-	-	-	-	-
Sek2	M	T	V	K	A	R	C	I	L	S	-	-
Sek3	M	V	K	R	C	I	-	-	-	-	-	-
Index array	0	1	2	3	4	5	6	7	8	9	10	11
Bobot fitness	5	5	0	7	-11	-4	-6	-6	-6	-6	-13	-3	-38

IV Seleksi

Tahap seleksi bertujuan untuk mendapatkan calon induk (*parent*) yang baik. Induk yang baik akan menghasilkan anak yang baik. Untuk proses seleksi dilakukan dengan menghitung nilai fitness. Semakin tinggi nilai fitness-nya maka semakin tinggi pula kesempatan individu untuk terpilih. Terdapat beberapa jenis seleksi dalam algoritma genetika, yaitu roulette wheel, binary tournament dan elitis. Pada penelitian ini, seleksi dilakukan dengan elitis. Seleksi elitis (elitism) berarti usaha untuk mempertahankan individu-individu terbaik yang telah diperoleh di suatu generasi ke generasi selanjutnya. Sehingga dalam generasi berikutnya, individu-individu terbaik tersebut akan tetap ada dan dipertahankan.

Induk	Nilai Fitness
P1	-38
P2	-142
P3	-112
C1	-82
C2	-89
C3	-110

Maka fitness terbaik adalah P1 dengan nilai fitness -38. Hasil penjajaran sekuennya adalah :

> MTVK−−−−−−−−−
> MTVKARCILS−−

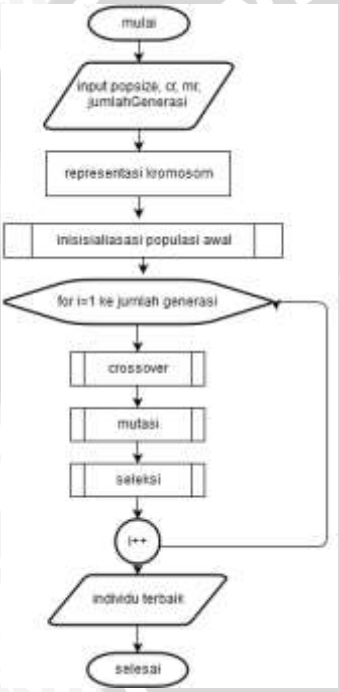
> MTVKRCI−−−−−
Karena menggunakan seleksi elitism maka yang diambil adalah individu dengan nilai terbaik sebanyak popsize yang akan diproses ke generasi berikutnya.

Induk Lama	Induk Baru	Nilai Fitness
P1	>P1	-38
P2	>C1	-82
P3	>C2	-89
C1	>C3	-110
C2	>P3	-112
C3	>P2	-142

P1, C1, C2, C3, dan P3 yang menjadi P1,P2,P3,P4,P5 yang baru.

3. PERANCANGAN

Proses penjajaran sekuen protein menggunakan algoritma genetika ditunjukkan pada Gambar 1.



Gambar 1. Proses Algoritma Genetika

4. PENGUJIAN DAN ANALISIS

4.1 Hasil Pengujian Kombinasi *Crossover Rate* dan *Mutation Rate*

Penentuan kombinasi nilai crossover rate dan mutation rate sangat penting dilakukan untuk mendapatkan solusi yang mendekati optimum. Nilai dari kombinasi cr dan mr yang digunakan pada pengujian ini yaitu 0.1:0.9, 0.2:0.8, 0.3:0.7, 0.4:0.6, 0,5:0.5, 0.6:0.4, 0.7:0.3, 0.8:0.2 dan 0.9:0.1 dengan masing-masing 10 kali pengujian. Perbandingan nilai cr dan mr tersebut akan menghasilkan jumlah anak yang sama pada masing-masing parameter, sehingga proses pembandingan masing masing parameter kombinasi cr dan mr akan seimbang.

Untuk ukuran populasi yang digunakan yaitu 100 dan untuk jumlah generasinya yaitu 100. Gambar merupakan grafik hasil pengujian kombinasi cr dan mr.

.

Gambar 2. Grafik Hasil Pengujian Kombinasi Cr Mr.

Pada Gambar dapat dilihat pada ukuran crossover rate dan mutation rate 0,3 dan 0,7 menghasilkan nilai fitness tertinggi yaitu -685,2. Setelah itu mengalami penurunan.

4.2 Hasil Pengujian Jumlah Populasi

Pada pengujian ini, ukuran Populasi yang digunakan dalam pengujian ini dimulai dari 100 – 4000 dengan interval 100 serta masing-masing 10 kali pengujian. Pada pengujian ini nilai parameter crossover rate dan mutation rate 0.3 dan 0,7 sesuai ukuran populasi optimal pada pengujian pertama. Untuk ukuran generasi yaitu 100. Gambar merupakan grafik hasil pengujian terhadap jumlah generasi.

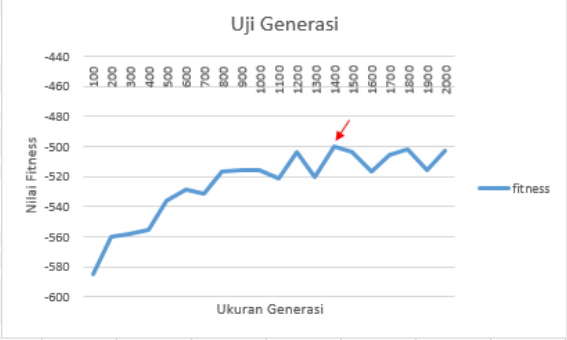
Gambar 3. Grafik Hasil Pengujian Jumlah Populasi.

Pada gambar diatas dapat disimpulkan bahwa semakin tinggi ukuran populasi maka berpengaruh terhadap rata-rata nilai fitness yang didapatkan, namun semakin tinggi ukuran populasi juga berpengaruh pada waktu pemrosesan algoritma genetika yang semakin lama (Suprayogi, dkk, 2015). Namun, dengan melihat grafik pada Gambar 3, nilai fitness mengalami kenaikan, dan pada ukuran populasi 3800 menghasilkan nilai fitness terbaik yaitu -584,6.

4.3 Hasil Pengujian Ukuran Generasi

Pengujian ukuran generasi ini untuk mengetahui bahwa pada generasi keberapa-kah sistem penjajaran sekuen jamak protein dengan algoritma genetika ini akan menghasilkan fitness terbaik. Dalam pengujian ukuran generasi ini menggunakan parameter jumlah populasi 300, karena telah kita hitung pada pengujian sebelumnya bahwa pada jumlah populasi tersebut menghasilkan nilai fitness terbaik.

Parameter crossover rate (cr) dan mutation rate (mr) yang digunakan adalah 0,3 dan 0,7, karena telah kita ketahui bahwa pada pengujian sebelumnya kombinasi crossover rate (cr) dan mutation rate (mr) pada jumlah tersebut menghasilkan nilai fitness terbaik. Ukuran generasi yang diuji yaitu dari 100 sampai 1000 dengan interval 100. Pengujian dilakukan 10 kali lalu dihitung nilai rata-rata fitness di setiap jumlah generasi. Gambar 4 merupakan grafik hasil pengujian kombinasi cr dan mr.



Gambar 4. Grafik Hasil Pengujian Ukuran Generasi.

4.3 Hasil Pengujian Faktor Pengali

Pengujian faktor pengali ini digunakan untuk mengetahui apakah pada faktor pengali 1,2 sudah menghasilkan nilai fitness tertinggi pada sistem penjajaran sekuen jamak protein dengan algoritma genetika ini. Pada pengujian factor pengali ini menggunakan parameter generasi dan populasi atau popsize 100 serta nilai crossover rate 0,3 dan mutation rate 0,7. Pengujian ini dilakukan 10 kali dan diambil rata-rata. Gambar 5 merupakan grafik hasil pengujian kombinasi cr dan mr.

Gambar 5. Grafik Hasil Pengujian Faktor Pengali.

Pada grafik diatas dapat kita lihat bahwa nilai faktor pengali 1 menghasilkan nilai fitness terbaik dan setelahnya mengalami penurunan. Hal ini terjadi karena semakin besar factor pengali pada penghitungan maxlength maka semakin banyak gap yang ada dan semakin sedikit kemungkinan penjajaran sekuen jamak sejajar dengan sekuen yang sama.

5. KESIMPULAN

1. Algoritma genetika dapat menghasilkan nilai *fitness* yang optimal dari langkah yang diawali dengan mencari *maxlength* dari sekuen yang disejajarkan untuk insialisasi kromosom, kemudian sekuen dilakukan *encode* untuk inisialisasi populasi dari index array yang berisi *gap* dengan dibatasi oleh *maxlength* dan dilakukan pengacakan index array yang berisi *gap* sebanyak populasi. Setelah itu dilakukan reproduksi yaitu *crossover* metode one-cut point serta mutasi metode reciprocal exchange dengan jumlah *offspring* atau anak dari *crossover rate* dan *mutation rate* yang sudah ditentukan. Tahap evaluasi atau disini peneliti melakukan *decode* yaitu mengembalikan index array berisi *gap* ke dalam bentuk semula dengan diisi *gap* untuk dilakukan penghitungan nilai *fitness* menggunakan *sum of pairs* menggunakan *blosum62* serta skema affine yang berisi *gap* open dan *gap* extension. Tahap terakhir adalah seleksi *elitism* yaitu individu atau kromosom dengan nilai *fitness* terbaik sejumlah populasi atau *popsize* akan dipertahankan untuk dilakukan proses ke generasi selanjutnya.

2. Telah dilakukan beberapa pengujian untuk mengetahui seberapa optimalkah sistem ini berjalan. Pengujian yang dilakukan adalah pengujian pada parameter algoritma genetika yaitu *crossover rate* dan *mutation rate*, populasi atau *popsize* dan generasi. Hasil Pengujian ukuran populasi didapatkan parameter ukuran populasi (*popsize*) yang optimal pada populasi ke-3800 dengan nilai *fitness* -584,6. Pada pengujian jumlah generasi, optimal pada pengujian generasi ke-1300 yaitu sebesar -505,4. Untuk pengujian kombinasi *crossover rate* (*cr*) dan *mutation rate* (*mr*) optimal pada pengujian *crossover rate* 0,3 dan *mutation rate* 0,7 dengan nilai *fitness* -685,2. Pengujian factor pengali 1 menghasilkan nilai *fitness* yang terbaik dengan nilai -491,5.

6. DAFTAR PUSTAKA

[1] Dr. Pankaj Agarwal. Alignment of Multiple Sequences using GA method. International Journal of Emerging Technologies in Computational and Applied Science (IJETCAS), 2008.

[2] Jin Xiong, Essential Bioinformatics. New York, United States of America: Cambridge University Press, 2006.

[3] Lei Xie and Philip E. Bourne. San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California,2007.

[4] Mahmudy, W. F. Modul Algoritma Evolusi. Malang: s.n, 2013.

[5] Suprayogi, D. A. & Mahmudy, W. F. Penerapan Algoritma Genetika Traveling Salesman Problem with TimePenerapan Algoritma Genetika Traveling Salesman Problem with Time. 6(2), pp. 121-130, 2014.

[6] Wargasetia, T. L. Peran Bioinformatika dalam Bidang Kedokteran. Jurnal Kesehatan Masyarakat, 5(2), pp. 59-72,2006.