IMPLEMENTASI ALGORITMA *SMITH–WATERMAN* P ADA *LOCAL ALIGNMENT* DALAM PENCARIAN KESAMAAN PENSEJAJARAN BARISAN DNA (STUDI KASUS : DNA TUMOR WILMS)

[1]Ernawati, [2]Diyah Puspitaningrum, [3]Ambar Pravitasari

[1]Program Studi Teknik Infomatika, Fakultas Teknik, Universitas Bengkulu.
Jl. WR. Supratman Kandang Limun Bengkulu 38371A INDONESIA
(telp: 0736-341022; fax: 0736-341022)

[1]w_ier_na@yahoo.com
[2]diyahpuspitaningrum@gmail.com
[3]ambarpravitasari@yahoo.com

Abstrak : Pencarian kesamaan pensejajaran barisan DNA dengan menggunakan algoritma *Smith – Waterman* pada *local alignment* ini digunakan untuk mencari persentase kesamaan yang dihasilkan untuk mengetahui tingkat kesamaan pensejajaran 2 buah sekuen. *DNA* merupakan suatu makro molekul yang tersusun oleh nukleotida sebagai molekul dasar yang membawa sifat gen. Aplikasi pencarian kesamaan pensejajaran barisan DNA ini diharapkan dapat memberikan kemudahan dalam melakukan pensejajaran 2 buah sekuen sehingga menampilkan tingkat kemiripan dari kedua sekuen DNA tersebut. Pada penelitian ini, sebagai masukannya menggunakan string barisan DNA yang di peroleh dari *National Center for Biotechnology Information* (**www.ncbi.nlm.nih.gov**) yang memiliki keluaran berupa persentase kesamaan serta lamanya waktu proses dijalankan.
Kata kunci : *local alignment, Smith – Waterman, DNA*

Abstract: ***Smith-waterman algorithm implementation in local alignment are used to find similarity percentage which also then be used to find the accuracy level of the two sequences. DNA itself is a macro molecul from nucleotide. This application of DNA parallel line similarity finder makes the alignment of 2 sequences of DNA easier so this application can show similarity level from the two sequences. In this research, DNA string line from National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov) used as input of application and the outputs are similarity level percentage and elapsed time of similarity process.***
Keywords: local alignment, Smith – Waterman, DNA

I. PENDAHULUAN

Bioinformatika merupakan gabungan antara ilmu Biologi dan ilmu Informatika yang akan menghasilkan suatu sistem komputasi dan analisa untuk menangkap dan menginterpretasikan data – data biologi. DNA atau *Deoxyribo Nucleic Acid* terdiri atas dua untai yang berpilin membentuk struktur heliks ganda. Pada struktur heliks ganda, orientasi rantai nukleotida untai lainnya. Hal ini disebut sebagai antipararel. Masing-masing untai terdiri dari rangka utama, sebagai struktur utama, dan basa nitrogen, yang berinteraksi dengan untai DNA satunya pada heliks. Kedua untai pada heliks ganda DNA disatukan ikatan hidrogen antara basa-basa yang terdapat pada kedua untai tersebut. Empat basa yang ditemukan pada DNA adalah *adenine* (A), *cytosine* (C), *guanine*(G), *thymine*(T). Adenin berikatan higrogen dengan timin, sedangkan guanine berikatan dengan sitosin.

Barisan DNA atau yang lebih dikenal dengan sekuen DNA merupakan suatu untaian nukleotida. Barisan DNA inilah yang selanjutnya akan dilakukan pensejajaran. Pensejajaran sekuen dapat digunakan untuk mempelajari serta

mengidentifikasi kemiripan (*similarity*) dari barisan DNA sehingga dapat dilakukan klasifikasi pada DNA yang sejenis. Klasifikasi ini penting dilakukan agar tidak terjadi kesalahan fungsi dalam penanganan suatu kasus.

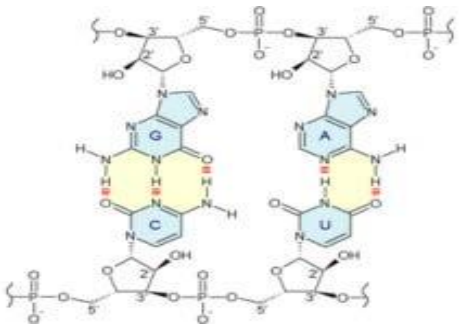
Untuk melakukan penyejajaran barisan DNA, pada artikel ini penulis menggunakan Smith - Waterman. Algoritma Smith - Waterman dirasa cukup baik untuk digunakan didalam pensejajaran 2 sekuen DNA ini. Selain memiliki akurasi waktu yang cukup baik, Algoritma Smith - Waterman merupakan algoritma yang paling baik apabila digunakan dalam melakukan penyejajaran secara *Local Alignment*. Menurut Referensi [2] *Local alignment* merupakan penyejajaran yang dilakukan dengan membagi sekuen menjadi beberapa sub bagian. Sub bagian itulah yang akan disejajarkan untuk mendapatkan persamaannya.

II. LANDASAN TEORI

A. *Bioinformatika*

Bioinformatika merupakan kajian yang memadukan disiplin biologi molekul, matematika dan teknik informasi (TI). Ilmu ini didefinisikan sebagai aplikasi dari alat komputasi dan analisa untuk menangkap dan menginterpretasikan data-data biologi molekul.

Genbank merupakan salah satu wadah bagi para peneliti untuk mempublikasikan sekuen basa nukleotida dan protein sebagai hasil proses translasi basa nukleotida yang ditemukan melalui kerja di laboratorium. Nukleotida menurut Referensi [1] merupakan struktur pembentuk inti sel – DNA dan RNA yang penting untuk perkembangan sel, fungsi-fungsi tubuh dan penggantian jaringan yang rusak, dapat dilihat pada Gambar 1. Nukleotida tersebut terdapat di semua sel tubuh. Nukleotida juga berperan dalam metabolisme sel.



Gambar 1. Nukleotida [1]

B. *DNA (Deoxyribo Nucleic Acid)*

DNA atau *Deoxyribo Nucleic Acid* merupakan suatu makro molekul yang tersusun oleh nukleotida sebagai molekul dasar yang membawa sifat gen. DNA terbentuk dari empat tipe nukleotida, yang berikatan secara kovalen, yang direpresetasikan oleh sejumlah huruf – huruf alphabet yaitu A, C, G dan T. Setiap huruf adalah inisial dari asam nukleat atau nukleotida (*nucleotides)* penyusun , yaitu *adenine* (A), *cytosine* (C), *guanine*(G), *thymine*(T). Molekul DNA memiliki struktur dua pita yang terjalin (*double helix*) yang bersatu dan berfungsi sebagai penyusun materi genetic [4].

C. *Wilms Tumor*

Tumor adalah pertumbuhan jaringan tubuh dimana terjadi proliferasi yang abnormal dari sel–sel. Tumor berupa masa padat atau berisi cairan yang ukurannya membesar. Tumor terjadi ketika sel membelah dan tumbuh berlebihan dalam tubuh. Sel–sel baru diciptakan untuk menggantikan sel yang lebih tua agar dapat melakukan fungsi–fungsi baru. Sel yang rusak atau tidak diperlukan lagi akan mati dan membuat ruang. Jika keseimbangan pertumbuhan sel dan kematian sel terganggu, maka akan terbentuklah apa yang dinamakan tumor. Wilm Tumor merupakan jenis tumor ginjal yang sering terjadi pada anak–anak. Tumor ini lebih banyak muncul pada anak usia 3–8 tahun. Wilms tumor yang dikenal dengan nephroblastoma,

diambil dari nama seorang ahli bedah Jerman yaitu Max Wilms, yang pertama kali mendeskripsikan tumor ini pada abad ke 19 [3]. Adanya massa besar di abdomen, terutama pada anak–anak usia 1–5 tahun harus menimbul-kan kecurigaan adanya Wilms Tumor. Terdapat 500 kasus baru tiap tahun di Amerika Serikat dan sebanyak 6 % Wilm Tumor menjangkit kedua buah ginjal [3].

D. Algoritma Smith – Waterman

1. Pengertian dan Fungsi

Algoritma Smith – Waterman muncul pada tahun 1981 dan sangat mirip dengan Algoritma Needleman – Wunsch. Namun algoritma Smith – Waterman digunakan pada local sequence alignment. Algoritma Smith – Waterman merupakan perluasan algoritma pencocokan *string* pada teks atau barisan sebagai salah satu implementasi program dinamis. Algoritma ini akan membandingkan keseluruhan panjang 2 sekuen yang terbagi menjadi sub bagian untuk mendapatkan kesamaan tertinggi antara kedua sekuen [2].

2. Penerapan Algoritma Smith–Waterman dalam Pensejajaran DNA

Prosedur algoritma Smith–Waterman adalah sebagai berikut, Pertama-tama, tetapkan nilai untuk setiap kecocokan karakter, ketidakcocokan karakter, serta nilai penalti apabila salah satu karakter dari kedua teks yang dibandingkan digeser sehingga diganti dengan karakter celah kosong. Nilai kecocokan (apabila dua karakter sama) haruslah ditetapkan sebagai suatu nilai positif. Hal ini disebabkan dua teks dikatakan semakin mirip jika nilai kecocokannya tinggi, sementara nilai kecocokan kedua teks tinggi apabila nilai kecocokan tiap karakter tinggi.

Sebaliknya, nilai ketidakcocokan ditetapkan sebagai nilai negatif atau nol. Nilai ketidakcocokan karakter juga harus berlaku simetris, artinya jika ketidakcocokan karakter a dan b bernilai -10, maka ketidakcocokan karakter b dan a juga harus bernilai -10. Serupa dengan nilai kecocokan, nilai ketidakcocokan harus bernilai negatif atau nol karena ketidakcocokan karakter mengurangi kemiripan kedua teks.

Adapun nilai penalti apabila salah satu karakter dari kedua teks yang dibandingkan digeser sehingga diganti dengan karakter celah kosong juga harus bernilai negatif. Hal ini disebabkan kita memerlukan upaya tambahan untuk menggeser karakter-karakter setelah celah yang disisipkan. Pada penelitian ini, nilai *match* 2, *mismatch* 0, dan *penalty* (*gap*) adalah -1.

Setelah ketiga nilai tadi ditetapkan, langkah selanjutnya adalah membentuk matriks berukuran jumlah baris = panjang teks pertama + 1, dan jumlah kolom = panjang teks kedua + 1. Ketika mengisi matriks, salah satu dari nilai-nilai matriks tidak boleh menjadi negatif, dengan demikian kita menganggap 0 berpotensi menjadi nilai maksimum. Dengan tidak membiarkan salah satu dari nilai-nilai berada di bawah nol. Hal ini memungkinkan algoritma Smith-Waterman untuk fokus hanya pada daerah-daerah dari sekuen yang serupa. Dan pada nilai positif diberikan nilai 1 sehingga nilai yang ada pada tabel hanya merupakan bilangan biner, yaitu 0 dan 1.

Pensejajaran Local dengan sekuen ACGT terhadap AGT dengan nilai match 2, mismatch 0 dan gap -1. Pada algoritma ini, perhitungan menggunakan bilangan binary. Sehingga setiap yang bernilai positif memiliki nilai 1 dan yang bernilai *negative* memiliki nilai 0.

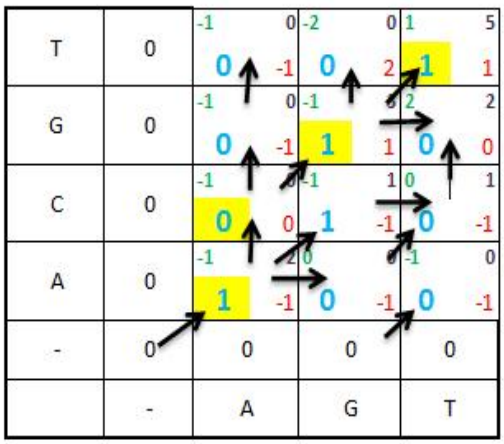
T	0			
G	0			
C	0			
A	0			
-	0	0	0	0
	-	A	G	T

Untuk mengisi nilai disetiap kolom dan baris, maka ingatlah ketetapan berikut:

a. Beside box (inputkan nilai gap) = hijau
b. Bottom box (inputkan nilai gap) = merah
c. Diagonal box (inputkan nilai match/mismatch + nilai diagonal) = hitam

T	0	-1	0 -1	-2	0 2	1	5 1
G	0	-1	0 -1	-1	5 1	2	2 0
C	0	-1	0 0	-1	1 -1	1	1 -1
A	0	-1	2 -1	0	0 -1	-1	0 -1
-	0	0		0		0	
	-	A		G		T	

Langkah selanjutnya, tentukan nilai paling besar pada setiap baris dan kolom. Maka diperolehlah hasil berikut,ingat nilai positif bernilai 1 dan nilai negative bernilai 0.



Langkah terakhir adalah mencocokkan karakter–karakter kedua teks yang bersesuaian.

Sehingga diperoleh hasil pensejajaran A C G T

| | |

A - G T

Hasil persentase :

$$\frac{jumlah\ huruf\ sejajar}{jumlah\ total\ huruf}\ x\ 100\ \% = \frac{3}{3} = 100\%$$

3. Pseudo – Code

```
function GetSWAlign(sf1,sf2 : string): string;
setlength(H,length(f1)+1,length(f2)+1);
//tetapkan panjang matrik H berdasarkan panjang sekuen input f1 dan f2
Algoritma
  for j := 0 to length(f2) do
  begin
      H[0,j] := 0;
  end;
  for i := 0 to length(f1) do
  begin
      H[i,0] := 0;
  end;
  for i := 1 to length(f1) do
  begin
      for j := 1 to length(f2) do
      begin
         match := H[i - 1,j - 1] + w;
         mismatch := H[i - 1,j - 1] + gap;
         delete := H[i - 1,j] + gap;
         insert := H[i,j - 1] + gap;
         if f1[i - 1] = f2[j - 1] then begin
              m := match;
          end else
          begin
              m := mismatch;
          end;
          if ((m > delete) and (m > insert) and (m > 0)) then
          begin
              H[i,j] := m;
          end else
          if ((delete > insert) and (delete > 0)) then
          begin
              H[i,j] := delete;
          end else if (insert > 0) then
          begin
              H[i,j] := insert;
          end else
          begin
              H[i,j] := 0;
          end;
      end;
  end;
  for i := 0 to length(f1) do
  begin
      for j := 0 to length(f2) do
      begin
      end;
  end;

  Alignmentf1 := '';
  Alignmentf2 := '';
  Alignmentf3 := '';
  for i := 0 to length(f1) do
```

```
begin
    for j := 0 to length(f2) do
    begin
        if (H[i,j] > max) then
        begin
            max := H[i,j];
            line := i;
            colum := j;
        end;
    end;
end;

while ((i > 0) and (j > 0)) do
begin
    Score := H[i,j];
    ScoreDiag := H[i - 1,j - 1];
    ScoreLeft := H[i - 1,j];
    jtot := jtot + 1;
    if (f1[i - 1] = f2[j - 1]) then
    begin
        w := 2;
    end else
    begin
        w := -1;
    end;
    if (Score = (ScoreDiag + w)) then
    begin
        Alignmentf1    :=    f1[i    -
1]+Alignmentf1;
        Alignmentf2    :=    f2[j    -
1]+Alignmentf2;
        if f1[i - 1] = f2[j - 1] then
        begin
          Alignmentf3    :=    '|'    +
Alignmentf3;
          jsama := jsama + 1;
        end else
          Alignmentf3    :=    '.'    +
Alignmentf3;
        begin

        end;
```

III. METODOLOGI

Pengumpulan data memiliki beberapa metode. Adapun metode yang digunakan dalam penelitian ini adalah metode studi pustaka dengan jenis data sekunder.

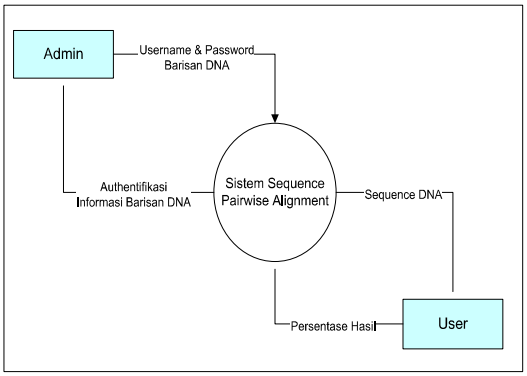
Studi kepustakaan dilakukan dengan mengumpulkan data dan informasi yang digunakan sebagai acuan dalam pembuatan aplikasi yang dapat menunjang dalam melakukan implementasi algoritma Smith – Waterman dalam pensejajaran sekuen DNA. Data dan informasi dapat berupa buku-buku ilmiah, laporan penelitian, skripsi, jurnal dan sumber-sumber tertulis lainnya yang berhubungan dengan pemahaman metode yang digunakan.

Jenis data yang akan digunakan adalah data sekunder. Dengan memperoleh data dari *National Center for Biotechnology Information* (NCBI) sebagai penyedia sumber informasi terkait perkembangan biologi molekuler. *National Center for Biotechnology Information* (NCBI) merupakan suatu institusi yang menyediakan sumber informasi terkait perkembangan biologi molekuler. NCBI membuat database yang dapat diakses oleh publik dan mengembangkan software penganalisis data genom. Situs NCBI dapat diakses pada website ***www.ncbi.nlm.nih.gov***. Sedangkan Database nukleotida merupakan suatu koleksi sekuen dari beberapa sumber, termasuk diantaranya GenBank. GenBank merupakan database sekuen genetik dari NIH (*National Institutes of Health*), berupa koleksi sekuen yang dapat diketahui oleh publik.

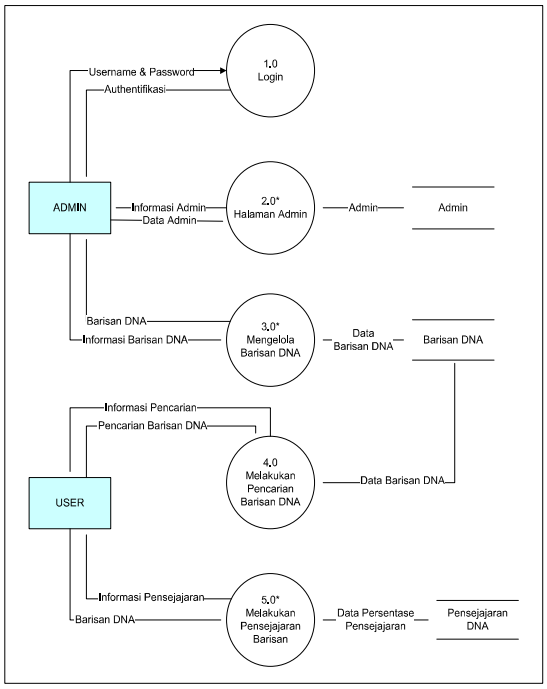
IV. ANALISIS DAN PERANCANGAN SISTEM

Analisis sistem merupakan bagian penelitian yang menganalisis sistem yang ada untuk merancang sistem baru atau memperbaharui sistem yang ada. Bagian ini merupakan bagian yang penting dikarenakan hasil dari sistem yang akan dibuat tergantung dari analisis yang dilakukan

A. *Diagram Alir Sistem*



Gambar 2. Diagram Konteks atau Diagram Level 0 Sistem Sequence Pairwise Alignment



Gambar 3. Diagram Level 1 Sistem pensejajaran Barisan DNA

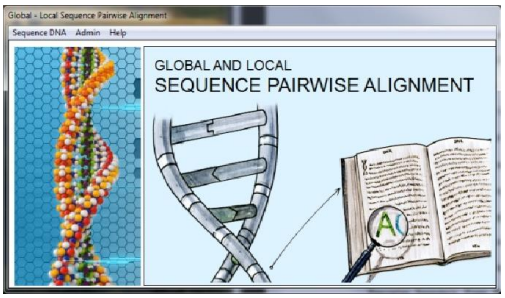
Pada Gambar 2 diagram konteks diatas, terdapat 2 *user* yang dapat menggunakan sistem, yaitu *admin* dan *user* biasa. Admin harus melakukan login terlebih dahulu untuk mengakses halaman admin dan menginputkan data sekuen barisan DNA. Sedangkan *user* biasa tidak perlu melakukan login terlebih dahulu untuk mengakses system, sehingga *user* biasa dapat melakukan pensejajaran barisan dan melihat data barisan secara langsung.

Pada Gambar 3 diatas, terdapat 5 proses yang dimiliki oleh sistem pensejajaran barisan DNA Tumor Wilms dan 2 entitas sistem. Dimana login, halaman *admin* dan mengelola barisan DNA dilakukan oleh Admin dan melakukan pencarian barisan dan melakukan pensejajaran dilakukan oleh *user* biasa.

Tanda bintang (*) pada proses 2 dan proses 5 menunjukkan bahwa pada proses tersebut ada proses lain yang lebih rinci.

V. HASIL DAN PEMBAHASAN

A. Implementasi Antar Muka Sistem



Gambar 4. Halaman Utama

Gambar 4 merupakan tampilan beranda sistem. User yang akan mengakses sistem, akan menemui halaman ini. Halaman yang ditunjukkan oleh Gambar 4 memiliki beberapa menu yang dapat digunakan oleh admin dan *user*.

Gambar 5. Halaman Pensejajaran *Sequence* DNA

Gambar 5 merupakan contoh dari pensejajaran barisan DNA Tumor Wilms pada locus AF484671 dan Locus AK290854 dengan menghasilkan output berupa persentase kesamaan dari hasil pensejajaran dan lamanya waktu proses yang dibutuhkan oleh sistem.

B. Hasil Pengujian Penerapan Algoritma SW dengan persentase jumlah karakter yang sama

Tabel 1 merupakan hasil rata – rata yang dihasilkan dari sejumlah *sequence* yang telah di sejajarkan dengan persentase jumlah karakter yang

sama dengan menggunakan 30 pasang sampel locus yang berbeda.

Tabel 1. Pengujian dengan Persentase Jumlah Karakter Sama

Persentase Barisan	Persentase Local	Runtime Local (ms)
100	58.61	178
90	58.62	123
80	58.78	98
70	58.93	73
60	58.99	61
50	59.23	55
40	59.15	39
30	59.34	24
20	58.39	13
10	59.45	7

Pada pensejajaran lokal, perhitungan dilakukan berdasarkan bagian sekuen yang memiliki tingkat kemiripan yang cukup tinggi. Semakin sedikit karakter yang disejajarkan, maka akan semakin besar persentasi kesamaan yang dihasilkan.

Setiap DNA yang akan disejajarkan, akan memiliki kombinasi sekuen DNA dan panjang DNA yang berbeda pula. Hal ini juga yang menjadi faktor utama mengapa persentase pensejajaran dan waktu proses yang dihasilkan memiliki tingkatan yang berbeda – beda. Semakin panjang sekuen DNA yang akan disejajarkan, maka akan semakin lama pula waktu proses yang dibutuhkan dalam melakukan pensejajaran.

C. Hasil Pengujian Penerapan Algoritma SW Dengan Persentase Jumlah Karakter yang Berbeda

Pada Tabel 2 merupakan hasil rata – rata yang dihasilkan dari sejumlah sequence yang telah di sejajarkan dengan persentase jumlah karakter berbeda dengan menggunakan 30 pasang sampel locus yang berbeda.

Tabel 2. Pengujian dengan Persentase Sequence 2, 100%

% Barisan 1	% Barisan 2	% Local	Runtime Local (ms)
100	100	50.14	203
90	100	50.18	99
80	100	50.61	89
70	100	50.65	82
60	100	51.30	69
50	100	52.80	61
40	100	53.06	48
30	100	51.81	35
20	100	51.57	23
10	100	57.38	18

Pada pengujian ini, peneliti ingin mengetahui selisih persentase yang dihasilkan pada pensejajaran local dengan cara memotong salah satu barisan DNA saja yang akan dibandingkan dengan barisan DNA utuh. Mengapa hanya salah satu? Karena jika dibandingan antara sekuen 1 terhadap sekuen 2 dan sekuen 2 terhadap sekuen 1, maka akan menghasilkan persentase kesamaan dan waktu proses yang sama.

Jika salah satu barisan dipotong secara berkala, maka hal tersebut akan berpengaruh terhadap hasil persentase yang dihasilkan, baik dari total persentase kesamaan maupun waktu proses yang dibutuhkan. Semakin panjang sekuen DNA yang akan disejajarkan, maka akan semakin lama pula waktu proses yang dibutuhkan dalam melakukan pensejajaran.

VI. KESIMPULAN

Algoritma Smith – Waterman merupakan algoritma yang digunakan untuk melakukan pensejajaran barisan secara local ini memiliki pensejajaran yang tinggi. Dari 30 pasang sampel yang diujikan, algoritma ini memiliki kestabilan pada hasil persentase pensejajarannya, mencakup karakter yang diujikan hanya 10% dari total jumlah keseluruhan karakter.

VII. SARAN

1. Pada aplikasi pensejajaran ini, d*atabase* yang digunakan merupakan database yang hanya di inputkan secara manual oleh admin. Diharapkan untuk pengembangannya, aplikasi ini dapat langsung terhubung ke database DNA (GenBank), sehingga pengguna akan lebih leluasa untuk mencari DNA yang ingin di sejajarkan.
2. Aplikasi ini hanya dapat melakukan pensejajaran dua buah sekuen. Diharapkan pada penelitian selanjutnya dapat melakukan pensejajaran lebih dari dua sekuen secara langsung.

REFERENSI

[1] Khulaipi, A. (2011). *Nukleutida.* [Online] Available at : http://viechumchumz.wordpress.com/2011/01/07/nukleotida/. [Accessed Februari 2014]

[2] Chan, A. (2004). *An Analysis of Pairwise Sequence Alignment Algorithm Complexities* . [Online] Available at : *http://biochem218.stanford.edu/ Projects%202004/Chan.pdf* [Accessed Januari 2014]

[3] Chrestella, J. (2009). Wilms Tumor. *Wilms Tumor* . [Online] Available at : *http://repository.usu.c.id/bitstream/123456789/ 2045/1/10E00542.pdf.* [Accessed Februari 2014]

[4] Sumardjo, Darmin. (2009). Pengantar Kimia : Buku Panduan Kuliah Mahasiswa Kedokteran dan Program Strata 1 Fakultas Biosekta. Jakarta. Penerbit Buku Kedokteran EGC.