Sequence Alignment Menggunakan Algoritma Smith Waterman[1]

[1]Inte Christinawati Bu'ulölö, [2]Nopelina Simamora, [3]Sabar Tampubolon, [4]Allan Pinem
Politeknik Informatika Del
Jl. Sisingamangaraja, Sitoluama
Kabupaten Tobasa, Sumater Utara
Email: [1]inte@del.ac.id, { [2]if07002, [3]if07072, [4]if07089}@students.del.ac.id

Abstrak

Analisis terhadap data genetik dilakukan untuk mengetahui struktur dan fungsi data genetik tersebut. Analisis pada data genetik protein dapat dilakukan dengan melakukan penyejajaran sekuen (*sequence alignment)* yaitu proses menyejajarkan suatu sekuen dengan satu atau beberapa sekuen lain sehingga diperoleh tingkat kesamaan di antaranya (*sequence similarity*). Satu sekuen terdiri dari sejumlah simbol (residu) yang mewakili data genetik protein. Dalam proses *alignment* sejumlah sekuen-sekuen, dibutuhkan nilai konstanta kesamaan dan ketidaksamaan antara dua residu, sehingga dapat dihitung nilai kesamaan antar dua atau lebih sekuen tersebut. Pada makalah ini dibahas bagaimana melakukan *sequence alignment* pada sekuen protein menggunakan algoritma *Smith-Waterman* secara terkomputerisari, serta menentukan nilai konstanta yang digunakan untuk kesamaan dan ketidaksamaan residu.

Kata kunci: ***sequence alignment,* sekuen protein, *Smith-Waterman, matrix BLOSUM*.**

1. Pendahuluan

Latar Belakang

Bioinformatika merupakan ilmu perpaduan antara ilmu biologi dan ilmu informatika untuk penyimpanan data (*storage*), pencarian informasi (*retrieval*), manipulasi data, dan distribusi informasi yang direlasikan dengan makromolekul biologi, seperti DNA, RNA, dan protein [1]. Salah satu tulisan mengenai *sequence alignment* pada data genetik DNA [2] menarik perhatian penulis untuk mengkaji lebih lanjut mengenai *sequence alignment* pada data genetik protein menggunakan algoritma *Smith-Waterman*.

Tujuan

Oleh karena itu tujuan dari TA ini adalah membangun sebuah prototipe aplikasi berbasis *desktop* yang dapat *melakukan sequence alignment pada* sekuen protein dengan algoritma *Smith-Waterman*.

Lingkup

Lingkup kajian yang dibahas dalam karya tulis ini adalah melakukan *sequence alignment* pada protein untuk mengetahui tingkat kesamaan dari satu sekuen protein terhadap satu sekuen protein lainnya dengan menggunakan algoritma *Smith-Waterman*.

2. Landasan Teori

Data Genetik Protein

Protein adalah senyawa organik yang terbuat dari *asam amino* yang disusun dalam sebuah rangkaian *linear* dan dibungkus ke dalam sebuah bentuk melingkar. Susunan dari *asam amino* dalam sebuah protein didefinisikan sebagai sekuen dari sebuah gen, yang dituliskan ke dalam kode genetik [3].

Protein merupakan rangkaian *polymer* yang terdiri dari 20 jenis asam amino yang berbeda. Terdapat sekitar 20 hingga lebih dari 5000 deretan dalam asam amino, namun rata – rata panjang protein sekitar 350 asam amino [4]. Sebuah individual asam amino disebut residu [5].

Sequence Alignment

Sequence Alignment adalah proses penyusunan/ pengaturan dua atau lebih sekuen sehingga

[1] Makalah ini disarikan dari Tugas Akhir (TA) Diploma Tiga Politeknik Informatika Del [6].

persamaan sekuen-sekuen tersebut tampak nyata [1]. *Sequence alignment* pada sekuen protein dilakukan untuk mencari tingkat kesamaan di antara sekuen-sekuen yang disejajarkan. Dengan mengetahui *similarity* sekuen protein, maka akan lebih mudah untuk mencari struktur dan fungsi dari protein tersebut. Struktur protein terkait dengan struktur tiga dimensi (3D) yang dimiliki protein. Fungsi dari suatu protein yang baru dapat diprediksi dari fungsi protein lain yang memiliki kesamaan dengan protein tersebut. Namun pada makalah ini tidak dibahas bagaimana menentukan struktur dan fungsi suatu sekuen protein dari tingkat kesamaan yang diperoleh.

Berikut adalah contoh *alignment* protein dari dua sekuen pendek protein yang berbeda, "**MNENLFAS**" dan "**MMENGGGLFAS**" (tanda "**:**" menunjukkan kecocokan atau *match* di antara kedua sekuen).

```
M N E  N - - - L F A S
:   :  :       : : : :
M M E  N G G G L F A S
```

Sequence alignment digunakan untuk mempelajari evolusi sekuen-sekuen dari leluhur yang sama (*common ancestor*). Ketidakcocokan (*mismatch*) dalam *alignment* diasosiasikan dengan proses mutasi, sedangkan kesenjangan (*gap*, tanda "–") diasosiasikan dengan proses insersi (bertambahnya sekumpulan asam amino yang baru dalam sekuen.) atau delesi (sejumlah asam amino yang hilang dalam suatu sekuen.).

Algoritma Smith-Waterman

Sequence alignment dilakukan dengan menggunakan algoritma, salah satunya adalah algoritma *Smith-Waterman* yang khusus meneliti *local alignment* dengan cara membandingkan *segment – segment* dari dua buah sekuen. *Segment* merupakan panjang deretan sekuen dari panjang keseluruhan sekuen protein yang memiliki nilai kesamaan tertinggi. Contoh algoritma lain adalah *Needleman-Wunsch* yang khusus meneliti *global alignment*.

Algoritma *Smith-Waterman* adalah salah satu algoritma yang termasuk *pairwise alignment* dan dalam prosesnya melakukan *local alignment.* Dalam penyejajaran protein, algoritma *Smith-Waterman* membandingkan setiap *segment* dari *sequence* protein. Algoritma *Smith-Waterman* memiliki beberapa kelebihan dibandingkan dengan algoritma lainnya, yaitu:

1. Algoritma ini dapat digunakan untuk *sequence alignment* yang sangat panjang, tergantung berapa lama proses perbandingan yang diinginkan dan tingkat generalisasi *sequence* yang ingin dibandingkan. Semakin panjang *sequence* berarti bahan yang dibandingkan semakin spesifik jenisnya, sebaliknya semakin pendek *sequence* berarti bahan yang dibandingkan semakin umum jenisnya.
2. Algoritma ini dapat digunakan untuk model *sequence alignment* pada beberapa data genetik, seperti *sequence alignment* DNA dan protein.
3. Algoritma *Smith-Waterman* menghasilkan hasil perbandingan yang lebih optimal karena algoritma ini membandingkan setiap *segment* secara *sensitive* dan *recursive*, sehingga proses perbandingan relatif lebih lama dibandingkan dengan algoritma lain. Algoritma ini dikatakan *sensitive* karena membandingkan setiap karakter dalam sekuen protein, sedangkan *recursive* karena menggunakan fungsi yang sama untuk membandingkan satu karakter dengan karakter lain.

Secara umum tahap-tahap dalam melakukan *sequence alignment* dengan algoritma *Smith – Waterman* adalah sebagai berikut:

1. Buat sebuah *matrix* dengan ukuran (x + 1) * (y + 1), di mana x menyatakan panjang sekuen pertama (S) dan y menyatakan panjang sekuen kedua (T).
2. *Set* semua elemen pada baris pertama dan kolom pertama pada *matrix* dengan nilai nol (0). Isi setiap kolom dengan menggunakan Persamaan(1) dan (2) [4].

   Basis : (1)

   $V(i,0) = 0,$
   $V(0,j) = 0,$
   $E(i,0) = -\infty$, untuk $i > 0,$
   $F(0,j) = -\infty$, untuk $j > 0$

   Rekurens: Untuk $i>0$ dan $j>0$ (2)

   $V(i,j) = \max(0, G(i,j), F(i,j), E(i,j)),$
   $G(i,j) = V(i-1,j-1) + \sigma(S[i], T[j]),$
   $F(i,j) = \max(F(i-1,j) – Ws, V(i-1,j) – Wg - Ws),$
   $E(i,j) = \max(E(i-1,j) – Ws, V(i-1,j) – Wg - Ws).$

3. Cari nilai tertinggi dari semua sel yang telah terisi pada *matrix.*

4. Lakukan *traceback* dimulai dari sel dengan nilai tertinggi hingga menemukan sel dengan nilai 0. Lakukan penghitungan nilai optimal dengan Persamaan (3) [4] sehingga dapat ditentukan tingkat kesamaan sekuen.

$$\sigma(S',T') = \sum_{i=1}^{l} \sigma(S[i],T[j]) \qquad (3)$$

3. Metodologi Penelitian

Berikut pendekatan yang dilakukan dalam menyelesaikan topik ini adalah:

1. Mempelajari literatur mengenai protein, *sequence alignment* dan algoritma *Smith-Waterman*.
2. Analisis mengenai permasalahan terkait penggunaan algoritma Smith-Waterman dalam melakukan *sequence alignment*.
3. Analisis dan perancangan kebutuhan data dan fungsi yang perlu dalam membangun prototipe aplikasi *sequence alignment* (SeqAP) pada protein.
4. Implementasi dari rancangan yang telah dibuat sebelumnya.
5. Pengujian dan Evaluasi terhadap hasil implementasi.

4. Analisis dan Desain

File FASTA

File FASTA merupakan salah satu tipe *file* sekuen dari beberapa tipe *file* format sekuen protein yang tersedia di GenBank (*database* penyedia sekuen protein). Tipe *file* FASTA lebih popular dibandingkan format sekuen protein lainnya karena cukup sederhana dan formatnya lebih mudah dibaca oleh banyak program komputer yang digunakan untuk analisis bioinformatika.

Meskipun demikian, *file* FASTA memiliki bermacam macam ekstensi format *file* seperti *.fa* dan *.fsa*. Namun NCBI[2] (National Center of Biotechnology Information) yang merupakan salah satu *database* penyedian data sekuen protein, mendistribusikan sekuen ke dalam empat ekstensi *file* FASTA yang berbeda, *.fna* digunakan untuk sekeun DNA, .faa untuk protein *coding sequence* (CDS), *.ffn* untuk sekuen yang tidak ditranslasikan untuk setiap CDS, dan *.fra* untuk sekuen fitur RNA yang direlasikan.

File Blossum

File BLOSUM merupakan *file* yang berisi *matrix* penyimpanan nilai pasangan residu asam amino. BLOSUM sangat cocok digunakan dalam melakukan *sequence alignment* pada protein agar dapat mengetahui struktur dan evolusi suatu protein tersebut [1]. *Matrix* BLOSUM berbeda dengan *unitary matrix* karena pada matrix ini telah ditentukan nilai dari setiap pasangan asam amino. Pada *Unitary Matrix* hanya ada beberapa nilai antara lain nilai *match*, *mismatch*, dan *gap*.

Ketika melakukan *traceback* pada *matrix* BLOSUM diperhatikan apakah kedua sekuen bersimbol sama, jadi nilai tertinggi tidak selalu menjadi patokan untuk tahapan selanjutnya. Misalnya nilai pada sel dengan *i* = 9 dan *j* = 11 nilainya 39, meskipun angka tersebut adalah angka yang tertinggi (*traceback* seharusnya *horizontal*), namun karena kedua simbol pada sel ke *i* dan ke *j* adalah sama maka *traceback* langsung ke arah *diagonal*.

Gambar 1 adalah contoh penghitungan *optimal alignment* pada algoritma *Smith – Waterman* dengan menggunakan matriks BLOSUM.

	j	0	1	2	3	4	5	6	7	8	9	10	11	12
i			M	N	N	E	N	E	N	M	L	E	E	N
0		0	0	0	0	0	0	0	0	0	0	0	0	0
1	M	0	9	2	0	0	0	0	0	7	3	0	0	0
2	N	0	2	14	9	4	7	2	7	2	3	3	0	7
3	N	0	0	9	21	16	11	7	9	5	0	3	3	7
4	E	0	0	4	16	27	22	17	12	7	2	6	9	4
5	N	0	0	7	11	22	34	29	24	19	14	9	6	16
6	M	0	7	2	6	17	29	32	27	31	26	21	16	11
7	L	0	3	3	1	12	24	27	28	30	36	31	26	21
8	M	0	7	2	1	7	19	22	25	35	33	34	29	24
9	E	0	2	7	2	7	14	25	22	30	32	39	40	35
10	N	0	0	9	14	9	14	20	32	27	27	34	39	47
11	M	0	7	4	9	12	9	15	27	39	34	29	34	42

Gambar 1 Contoh penghitungan *optimal alignment*

Nilai *optimal alignment*-nya adalah sebagai berikut:

S' =	M	-	N	-	N	E	N	M	L	M	E	N
T' =	M	N	N	E	N	E	N	M	L	E	E	N

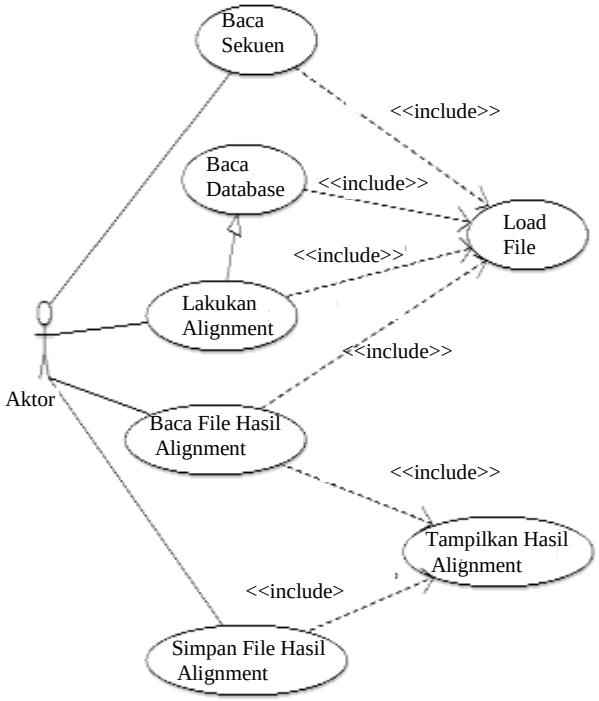
σ (S',T') = 2MM + 4NN + 2EE + (-E) + (-N) + LL + (ME)

σ (S',T') = 2(7) + 4(7) + 2(6) + (-5) + (-5) + 5 + (-2)

σ (S',T') = 47

[2] www.ncbi.nlm.nih.gov

Use Case



Gambar 2 Use Case SeqAp

5. Hasil

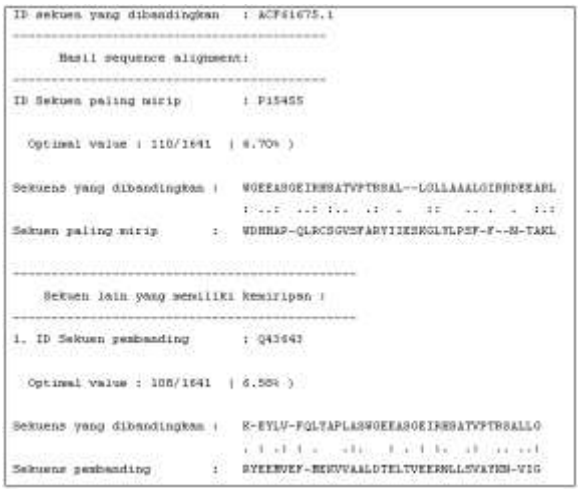
Pada prototipe aplikasi SeqAP, sekuen protein dibaca dari *file* dengan format Fasta. Nilai konstanta pada matriks BLOSUM50 digunakan dalam menentukan nilai sepasang residu yang disejajarkan.

Hasil *sequence alignment* dapat ditampilkan pada sebuah *file* buatan berekstension *.psw*.

Gambar 3 adalah contoh hasil *sequence alignment* menggunakan SeqAP.

Evaluasi pada aplikasi ini dilakukan dalam tiga pengujian, yaitu:

1. Pengujian aplikasi terhadap fungsi-fungsi yang telah dicakup dalam *use case* yang ditampilkan pada Gambar 2. Untuk membuktikan bahwa fungsi-fungsi pada aplikasi berjalan dengan baik dan mencakup fungsi-fungsi pada *use case*.
2. Pengujian membandingkan hasil penyejajaran sekuen pada aplikasi dengan hasil penyejajaran sekuen melalui penghitungan manual.
3. Pengujian membandingkan hasil penyejajaran sekuen pada aplikasi dengan hasil penyejajaran sekuen dari modul **BioPerl** (website PIR[3]).



Gambar 3 Hasil *Sequence Alignment* dengan SeqAP

Setelah pengujian diperoleh bahwa prototipe SeqAp telah memenuhi semua fungsi yang telah dirancang sebelumnya. Sedangkan hasil pengujian lainnya ditampilkan pada Tabel 1 dan Tabel 2.

Tabel 1 Perbandingan hasil *sequence alignment* antara SeqAP dan penghitungan manual.

No	Id Sekuen yang Dibandingkan	Id sekuen Pembanding	Optimal Value		Kesimpulan
			Penghitungan Manual	Aplikasi	
1	P21215	P13813	24	24	Output sama
2	P21215	Q197F8	24	24	Output sama
3	P21215	P13744	16	16	Output sama
4	P21215	P19084	15	15	Output sama
5	P21215	Q8GBW6	30	30	Output sama

Tabel 2 Perbandingan hasil *sequence alignment* antara SeqAP dan modul BioPerl (Web PIR).

Id Sekuen yang Dibandingkan	Nomor Urut	Tingkat Kemiripan		Kesimpulan
		Modul BioPerl	Aplikasi	
P21215	1	Q84J55	Q84J55	Sama
	2	P93207	P13813	Berbeda
	3	Q8GBW6	Q8GBW6	Sama
	4	Q43643	Q9U408	Berbeda
	5	P85938	Q43643	Berbeda
	6	Q197F8	Q197F8	Sama
	7	P13813	P85938	Berbeda
	8	Q9U408	P93207	Berbeda
	9	P13744	P13744	Sama
	10	P19084	P19084	Sama

[3] http://pir.georgetown.edu/

6. Pembahasan Hasil

Pada Gambar 3 ditampilkan hasil keluran SeqAp yaitu nilai *similarity* yang diurutkan secara *descending*.

Pada tahap ini ada isu yang muncul yaitu panjang sekuen yang dibandingkan berbeda-beda. Untuk itu penghitungan *optimal value* dimodifikasi sebagai berikut: pada hasil penyejajaran sekuen ditampilkan ID sekuen yang memiliki nilai optimal paling tinggi yang terdapat dalam *database*. Angka 110 dari hasil 110/1641 merupakan nilai optimal dari penyejajaran sekuen dengan Id ACF61675.1 dan sekuen dengan Id P15455. Sementara angka 1641 merupakan nilai optimal dari sekuen dengan Id ACF61675.1 dibandingkan dengan sekuen itu sendiri. Angka 6.70 % merupakan presentasi nilai optimal. Pada urutan selanjutnya ditampilkan sekuen – sekuen lain yang juga memiliki tingkat kemiripan dengan sekuen yang dibandingkan.

Data pada Tabel 1 menampilkan hasil yang sama antara *sequence alignment* dari SeqAP dan dari penghitungan manual. Sementara pada Tabel 2 ditampilkan hasil yang bervariasi. Hal ini terjadi karena setelah dipelajari lebih dalam, modul BioPerl pada web PIR secara *default* menggunakan *global alignment* (mengimplementasikan algoritma *Needleman-Wunsch*) namun sama-sama menggunakan *matrix* BLOSUM50. Penyejajaran sekuen menampilkan hasil yang bervariasi disebabkan karena pada beberapa kasus, hasil dari *traceback* algoritma *Smith-Waterman* berada pada jalur *traceback* algoritma *Needleman-Wunsch.*

7. Kesimpulan dan Saran

Dari TA ini diperoleh bahwa Algoritma *Smith – Waterman* dapat digunakan untuk *sequence alignment* pada sekuen protein. Dalam melakukan *sequence alignment* digunakan *matrix* untuk menghitung nilai dari pasangan residu.

Matrix yang digunakan pada *sequence alignment* dapat menggunakan *unitary matrix* dan *matrix* BLOSUM. Namun, *unitary matrix* lebih cocok digunakan pada *sequence alignment* DNA karena suatu karakter pada DNA sama sekali tidak memiliki keterikatan dengan karakter DNA yang lain. Sedangkan *matrix* BLOSUM lebih cocok digunakan pada *sequence alignment* protein karena suatu karakter pada protein kemungkinan memiliki keterikatan dengan karakter protein yang lain.

Pada *local sequence alignment*, penghitungan nilai optimal tidak tergantung pada panjang pendeknya karakter sekuen karena *local alignment* hanya menghitung *optimal value* dari setiap *seqment* yang memiliki kesamaan.

Untuk proses pengurutan hasil *sequence alignment* perlu dipertimbangkan panjang dari sekuen-sekuen yang dibandingkan.

Hasil *sequence alignment* antara algoritma *Smith-Waterman* dan *Needleman-Wunsch* bervariasi, dalam kondisi tertentu dapat memperlihatkan hasil yang sama. Hal ini terjadi apabila hasil *traceback* pada algoritma *Smith-Waterman* berada pada jalur *traceback* algoritma *Needleman-Wunsch*.

Pada makalah ini dikaji mengenai penggunaan algoritma *Smith-Waterman* pada sekuen protein. Pada prosesnya hanya dilakukan analisis pada format *file* FASTA dari GenBank (NCBI) sebagai penyimpanan sekuen protein. Karena itu sebaiknya dilakukan juga analisis untuk *file* FASTA yang berasal dari *database* penyedia sekuen protein lainnya seperti DDBJ, EMBL, PIR, karena *file* FASTA pada setiap *database* memiliki struktur informasi sekuen yang berbeda. Aplikasi ini juga dapat dikembangkan menjadi aplikasi yang berbasis *web* agar dapat diakses dan digunakan dari tempat yang berbeda.

8. Penutup

Dari Tugas Akhir (TA) ada beberapa hal yang diidentifikasi oleh pembimbing dari bidang, tipe, orisinalitas, dan tingkat kesulitan.

1. Bidang kajian dari TA ini termasuk Bioinformatika, dengan tipe TA yaitu eksplorasi dan implementasi.
2. Topik ini tentunya sudah pernah dilakukan oleh orang lain, sisi orisinalitasnya adalah mahasiswa mengerjakan sendiri topik ini dengan menggunakan pengetahuan yang sudah diperolehnya selama masa perkuliahan.
3. Tingkat kesulitannya dinilai cukup tinggi karena mahasiswa mengimplementasikan ilmu yang dimiliki pada bidang baru yang belum pernah dipelajari sebelumnya yaitu bioinformatika. Terkait dengan bidang baru ini, mahasiswa menghabiskan cukup banyak waktu hanya untuk memahami.

Meskipun lingkup dari TA ini adalah implementasi suatu algoritma, namun topik ini dipandang layak diberikan kepada mahasiswa karena di sini mahasiswa dituntut melakukan analisis serta tahu aspek apa, bagaimana, dan mengapa, dari topik yang mereka kerjakan. Selain itu mahasiswa juga perlu memahami sedikit tentang bioinformatika khususnya

sequence alignment. Memang disadari bahwa pada kurikulum PI Del tidak dicakup matakuliah tentang bioinformatika, oleh karena itu pada masa awal TA pembimbing memberikan tutorial singkat tentang bioinformatika khususnya *sequence alignment*.

9. Daftar Pustaka

[1] Xiong, Jin; 2006; *Essential Bioinformatics; Cambridge*.

[2] Bu'ulölö, Inte. *Studi dan Implementasi Algoritma SST dan Smith-Waterman menggunakan Perl, Proceeding of 1st Conference on Applied Information Technology, POLBAN, Bandung* 2007.

.

[3] Ridley, M. *Genome*. New York, NY: *Harper Perennial*: 2006.

[4] Tompa, Martin; *Biological Sequence Analysis; Winter:* 2000.

[5] Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipurksy SL, Darnell J. *Molecular Cell Biology* 5th ed. WH Freeman and Company: New York, NY: 2004.

[6] Nopelina Simamora, Sabar Tampubolon, Allan Pinem, *Sequence Alignment* Menggunakan Algoritma *Smith-Waterman*, Tugas Akhir Diploma Tiga Politeknik Informatika Del Program Studi Teknik Informatika, 2010.