PERBANDINGAN PROGRAM SEQUENCE ALIGNMENT

Asril Adi Sunarto

Program Studi Teknik Informatika STT Nusa Putra Sukabumi
Email : asril_adi83@yahoo.com

Abstrak

Multiple Sequence Alignment merupakan metode untuk menemukan kemiripan diantara banyak urutan biologis (*deoxyribonucleic acid* dan protein) yang salah satu algoritmanya menggunakan teknik *dynamic programming Needlemen–Wunsch* dengan kompleksitas $O(n^2)$. Penelitian ini bertujuan untuk mengembangkan program *Sequence Alignment* dengan menggunakan algoritma *Needlemen–Wunsch* yang menjajarkan banyak urutan. Data organisme berasal dari NCBI yaitu *Arabidopsis thaliana, Solanum lycopersicum 2, Lycopersicum, Solanum2 lycopersicum*. Program *beta* ini dibandingkan dengan program lainnya yang sudah ada. Hasilnya terdapat selisih hingga 10% perbedaan skor kemiripan antara program yang ada.

Keyword : Sequence Alignment, Needlemen-Wunsch, Dynamic Programming, Multiple Sequence Alignment.

Abstract

Multiple Sequence Alignment is a method to find similarities among many biological sequences (deoxyribonucleic acid and protein) which is one of the algorithm using dynamic programming techniques Needlemen-Wunsch with complexity O (n2). This research aims to develop a program Sequence Alignment using Needlemen-Wunsch algorithm that aligns many sequences. The data comes from NCBI organism like Arabidopsis thaliana, Solanum lycopersicum 2, Lycopersicum, Solanum2 lycopersicum. This beta program compared to other programs that already exist. The result is there is a difference of up to 10% difference between the *Similarity* scores of existing programs.

I. PENDAHULUAN

Sequence alignment (penjajaran barisan) adalah cara mengatur urutan *deoxyribonucleic acid* (DNA), *ribonucleic acid* (RNA), atau protein untuk mengidentifikasi daerah kesamaan yang mungkin menjadi konsekuensi dari hubungan fungsional, struktural, atau evolusi antara urutan (Sing *et al*, 2011).

Identifikasi DNA memerlukan *sequence alignment* tersendiri. Salah satu metode untuk *sequence alignment* adalah metode *Needlemen–Wunsch*. Metode ini mendefinisikan cara menemukan urutan global terbaik dari dua *sequence* / *Pair wise* (Needleman-Wunsch, 1970). *Pairwise alignment* (pasangan urutan) yang dijajarkan menggunakan *dynamic programming* dengan matrik dua dimensi sehingga kompleksitasnya $O(n^2)$.

Dynamic programming yang digunakan ketika *pairwise alignment* merupakan metode yang mempunyai strategi memecah masalah menjadi lebih kecil guna membangun solusi yang lebih besar (Jones dan Pevzner, 2004).

Susanti *et al* (2008) menganalisa hubungan filogenetik dari fragmen (HA) gen hemaglutinin dari virus flu burung (AIV) subtipe H5N1 yang diisolasi dari hewan dan isolat manusia di Indonesia. Abbas *et al* (2010) melanjutkan penelitian tentang analisis filogenetik yang mengungkapkan bahwa ada dua perkenalan H7 ke dalam wilayah Pakistan dan satu pengenalan N3.

Baik Susanti *et al* (2008) maupun Abbas *et al* (2010) keduanya menggunakan *multiple sequence alignment* untuk membuat dan menganalisa *phylogetic tree*. Do *et al* (2013) memperkenalkan *tool* untuk *multiple sequence alignment* yang dikenal dengan *probabilistic consistency*. *Tool* ini menggunakan *Pair Hidden Markov Model*. Hasilnya dalam 141 penjajaran secara rata-rata membutuhkan waktu 5:32 menit dengan *processor Pentium IV 3.3 GHz* dan *RAM 2 GB*.

Hadirnya Virus Avian Influenza (VAI) atau yang lebih dikenal dengan virus flu burung subtipe H5N1, H5N2, dan H5N9 telah menyebabkan ribuan unggas mati di Indonesia beberapa tahun lalu. Untuk mengetahui seberapa persen kemiripan atau membuat vaksin VAI antara subtype H5N1, H5N2, H5N9 memerlukan *sequence alignment*. Pentingnya *sequence alignment* dalam pembuatan vaksin, maka diperlukan suatu program *sequence alignment*. Pembuatan program ini menggunakan metode *Needlemen–Wunsch*. Sebagai uji program, hasil dari sequence alignment dibandingkan dengan program lainnya yang sudah ada seperti Clustal X version 2.0 (Larkin et al, 2007) dan EMBOSS (*http://www.ebi.ac.uk*).

II. METODE

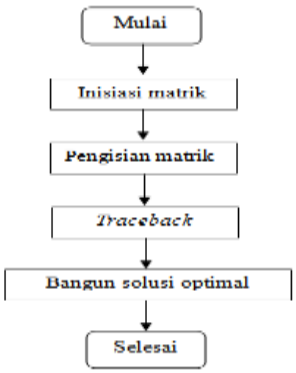
A. Data

Sebelum menganalisa data, hal yang harus dipersiapkan adalah data virus yang akan diteliti. Data penelitian ini diperoleh dari GeneBank NCBI (http://www.ncbi.nlm.nih.gov/genomes/FLU/Database). Virus yang akan dibandingkan berasal dari *NCBI* yaitu :

No	Nama Organisme
1	Arabidopsis thaliana
2	Solanum lycopersicum 2
3	Lycopersicum
4	Solanum2 lycopersicum

B. Metode

Sequence alignment DNA yang menggunakan Needleman-Wunsch diperkenalkan pada tahun 1970. Algoritma ini merupakan algoritma iteratif yang digambarkan pada sebuah matrik (*array*) dua dimensi untuk menemukan skor tertinggi dari sequence alignment. Kompleksitas yang dihasilkan mencapai $O(n^2)$. Metode dalam memproses *sequence alignment* dapat dilihat seperti berikut ini:



Gambar 1. Metode penelitian.

Formula untuk pengisian matrik didefnisikan sebagai berikut :

$$M[i,j] = max\begin{cases} M[i-1,j-1] + sub(S[i],B[j]); \\ M[i-1,j] + gap; \\ M[i,j-1] + gap; \end{cases}$$

Penelitian ini menggunakan parameter match = 7, mismatch = -3, dan gap = -2.

III. HASIL DAN PEMBAHASAN

Tahap pertama adalah proses inisiasi matrik. Matrik yang dipakai adalah matrik dua dimensi yang berisi dua sequence yang akan dibandingkan. Tahap ini memberikan nilai awal secara iteratif sesuai dengan nilai *gap* yang diberikan di tiap baris ke pertama dan kolom pertama. Proses inisiasi matrik ini dicontohkan terdapat *sequence* 1 = 'ATGCAG', *sequence* 2 = 'TAGCGA' dapat dilihat pada tabel 1 dibawah ini :

Tabel 1 : Inisiasi matrik

		A	T	G	C	A	G
	0	-2	-4	-6	-8	-10	-12
T	-2						
A	-4						
G	-6						
C	-8						
G	-10						
A	-12						

Implementasi dari formula diatas dapat dilihat pada tabel 2 dibawah ini :

Tabel 2 : Pengisian matrik

		A	T	G	C	A	G
	0	-2	-4	-6	-8	-10	-12
T	-2	-3	5	3	1	-1	-3
A	-4	5	3	2	0	8	6
G	-6	3	2	10	8	6	15
C	-8	1	0	8	17	15	13
G	-10	-1	-2	7	15	14	22
A	-12	-3	-4	5	13	12	20

Tahap ketiga adalah proses *traceback* (runut balik) yang dimulai dari M[i, j] menuju M[0,0]. Kompleksitas dari proses *traceback* ini adalah $O(i + j)$. Hasil dari traceback ini dapat dilihat pada Tabel 3 dibawah ini :

(1)

Tabel 3. Proses traceback

		A	T	G	C	A	G
	0	-2	-4	-6	-8	-10	-12
T	-2	-3	5	3	1	-1	-3
A	-4	5	3	2	0	8	6
G	-6	3	2	10	8	6	15
C	-8	1	0	8	17	15	13
G	-10	-1	-2	7	15	14	22
A	-12	-3	-4	5	13	12	20

Selanjutnya, proses terakhir adalah membangun solusi yang optimal dari hasil sebelumnya. Berdasarkan hasil sebelumnya proses akhir ini menjadi:

Tabel 4. Membangun Solusi Optimal

–	A	T	G	C	A	G	–
	I		I	I		I	
T	A	-	G	C	-	G	A

Pada Tabel 3 di atas, skor *sequence alignment* antara dua invidu yang dibandingkan mencapai 20. Skor ini tidak mempunyai arti apapun. Hasil *sequence alignment* yang menjadi informasi penting dapat dilihat pada Tabel 4 di atas.

Pada Tabel 4, dapat dihitung jumlah persentase kemiripan dengan cara menghitung jumlah karakter yang mirip dibandingkan dengan jumlah seluruh karakter yang ada yaitu $\frac{4}{8} x100\% = 50\%$.

Adapun hasil sequence alignment antara data-data yang telah disiapkan berturut-turut menggunakan program EMBOSS, ClustalX 2.1 dan program *beta* yang penulis sertakan diantaranya adalah :

a. 1-2

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: Arabidopsis
# 2: Solanum
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2456
# Identity:    1083/2456 (44.1%)
# Similarity:  1083/2456 (44.1%)
# Gaps:         772/2456 (31.4%)
# Score: 4419.0
#
#
#=======================================
```

ClustalX 2.1 : *Similarity* 40.6 %

Program *beta* : *Similarity* 48.47%

b. 1-3

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: Arabidopsis
# 2: lycopersicum
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2432
# Identity:    1025/2432 (42.1%)
# Similarity:  1025/2432 (42.1%)
# Gaps:         787/2432 (32.4%)
# Score: 4395.0
#
#
#=======================================
```

ClustalX 2.1 : *Similarity* 40.4%

Program *beta* : *Similarity* 51.16%

c. 1-4

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: Arabidopsis
# 2: Solanum2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2462
# Identity:    1129/2462 (45.9%)
# Similarity:  1129/2462 (45.9%)
# Gaps:         669/2462 (27.2%)
# Score: 4932.5
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 43 %

Program *beta Similarity* 53.05%

d. 2-3

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: Solanum
# 2: lycopersicum
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1871
# Identity:    1018/1871 (54.4%)
# Similarity:  1018/1871 (54.4%)
# Gaps:         245/1871 (13.1%)
# Score: 4684.5
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 52.32%

Program *beta Similarity* 59.84%

e. 2-4

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: Solanum
# 2: Solanum2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1927
# Identity:    1085/1927 (56.3%)
# Similarity:  1085/1927 (56.3%)
# Gaps:         179/1927 ( 9.3%)
# Score: 5191.5
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 54.08%
Program *beta Similarity* 59.78%

f. 3-4
EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: lycopersicum
# 2: Solanum2
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1956
# Identity:    1044/1956 (53.4%)
# Similarity:  1044/1956 (53.4%)
# Gaps:         300/1956 (15.3%)
# Score: 4784.5
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 50.73%
Program *beta Similarity* 58.38%

Hasil rekapitulasi *sequence alignment* diatas oleh beberapa program dapat dilihat pada Tabel 1 dibawah ini :

Tabel 1. Sequence Alignment I

Seq	1	2	3	4	
1		44.1	42.1	45.9	
2	44.1		54.4	56.3	
3	42.1	54.4		53.4	
4	45.9	56.3	53.4		

EMBOSS

Seq	1	2	3	4	
1		40.6	40.37	43	
2	40.6		52.32	54.08	
3	40.4	52.32		50.73	
4	43	54.08	50.73		

ClustalX 2.1

Seq	1	2	3	4	
1		48.47	51.16	53.05	
2	48.47		59.84	59.78	
3	51.16	59.84		58.38	
4	53.05	59.78	58.38		

Program *Beta*

Dari Tabel 1 diatas memperlihatkan kandidat *center star* adalah sequence *Solanum2 lycopersicum*. Namun urutan sequence *Solanum2 lycopersicum* yang mana yang dipilih menjadi center star mengingat sequence *Solanum2 lycopersicum* mempunyai hasil *sequence alignment* yang berbeda dengan semua sequence. Untuk itu diperlukan *sequence alignment* antara sequence *Solanum2 lycopersicum.* Dibawah ini merupakan hasil dari *sequence alignment* antara *sequence Solanum2 lycopersicum* hasil *sequence alignment* tahap pertama :

a. *Solanum2 lycopersicum 1-4 dengan Solanum2 lycopersicum 2-4*
EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1895
# Identity:    1895/1895 (100.0%)
# Similarity:  1895/1895 (100.0%)
# Gaps:           0/1895 ( 0.0%)
# Score: 11009.0
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 100 %
Program *beta Similarity* 97.34 %

b. *Solanum2 lycopersicum 1-4 dengan Solanum2 lycopersicum 3-4*
EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1895
# Identity:    1895/1895 (100.0%)
# Similarity:  1895/1895 (100.0%)
# Gaps:           0/1895 ( 0.0%)
# Score: 11009.0
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 99.55 %
Program *beta Similarity* 96.71%

c. *Solanum2 lycopersicum 2-4 dengan Solanum2 lycopersicum 3-4*

EMBOSS

```
#=======================================
#
# Aligned_sequences: 2
# 1: EMBOSS_001
# 2: EMBOSS_001
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1895
# Identity:    1895/1895 (100.0%)
# Similarity:  1895/1895 (100.0%)
# Gaps:           0/1895 ( 0.0%)
# Score: 11009.0
#
#
#=======================================
```

ClustalX 2.1 *Similarity* 100 %
Program *beta Similarity* 98.19 %

IV. KESIMPULAN

Dari hasil uji program diatas, skor tiap program mempunyai nilai yang berbeda meskipun objek penelitiannya sama. Hal ini bisa saja dipengaruhi teknik pemogramannnya. Perbedaan skor kemiripan antara program yang sudah ada dengan program yang dibuat hingga mencapai 10%. Selisih ini angka sangat besar sehingga patut kiranya program yang telah dibuat oleh penulis perlu perbaikan. Meskipun begitu, program ini layak untuk dikembangkan untuk proses *multiple sequece alignment*. Dalam *multiple sequence alignment* terdapat proses *sequence alignment* sebanyak $\frac{k(k-1)}{2}$ sehingga kompleksitas menjadi $O(k^2 n^2)$.

Perlu penelitian lanjutan agar komputasi waktu yang dibutuhkan menjadi lebih cepat. Salah satu caranya menggunakan teknik komputer paralel.

V. DAFTAR PUSTAKA

1. Do, Chuong B *et al*. 2013. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 2005 15: 330-340. doi:10.1101/gr.2821705.
2. Jones, Neil C and Pavel A Pevzner. 2004. An Introduction to Bioinformatics Algorithms. Massachusetts Institute of Technology, USA.
3. Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.
4. Sing, et al. 2011. *Role Of Bioinformatics In Agriculture And Sustainable Development*. Banaras Hindu University, India.
5. Susanti, et al. 2008. Filogenetik dan Struktur Antigenik Virus Avian Influenza Subtipe H5N1 Isolat Unggas Air. Veteriner Vol. 9 No. 3 : 99-106.