Aplikasi Algoritma Pencocokan String Knuth-Morris-Pratt dalam Pendeteksian Kemiripan Sifat

Irene Wiliudarsan 13513002[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]*irene@s.itb.ac.id*

***Abstrak*—Kemiripan sifat dari dua individu yang berbeda dapat dideteksi dari DNA kedua individu tersebut. DNA adalah sejenis biomolekul yang menyimpan instruksi-instruksi genetika setiap organisme dan berbagai jenis virus. Pencocokan string adalah salah satu metode untuk mendeteksi kemiripan dari dua buah DNA tersebut. Pencocokan string dapat dilakukan secara komputasi dengan menggunakan program. Salah satu algoritma terkenal yang dapat digunakan adalah algoritma Knuth-Morris-Pratt.**

Kata Kunci—Kemiripan Sifat, Genetika, DNA, Pencocokkan String, Algoritma Knuth-Morris-Pratt.

I. PENDAHULUAN

Berdasarkan pengamatan secara sederhana, dapat disimpulkan bahwa terdapat kesamaan dalam sifat tertentu yang dimiliki tetua dengan sifat pada anaknya atau sesama saudara kandung dan atau sesama saudara tiri. Demikian pula ternak yang berkerabat dekat terdapat beberapa kesamaan sifat-sifat di antara mereka. Disimpulkan pula bahwa terdapat sifat-sifat yang pewarisannya lebih mantap dari sifat yang lain. Sejak itu, manusia diduga telah mencoba mencari hukum-hukum yang mengatur pewarisan sifat.

Hal inilah yang kemudian membuahkan ilmu genetika modern. Manusia diperkirakan telah berusaha mengembangkan cara-cara pemuliaan sederhana yang efektif. Namun baru setelah hukum-hukum Mendel ditemukan kembali pada awal abad ini, Ilmu Pemuliaan Ternak Moderen mulai berkembang dengan pesat.

Ilmu Genetika merupakan salah satu cabang ilmu yang mempelajari seluk beluk gen sebagai unit dasar biologis yang mengontrol pewarisan sifat. Karena gen memegang peran utama dalam kehidupan, menyebabkan ilmu genetika memiliki banyak kaitan dengan cabang ilmu lain dalam bidang biologi. Pada dasarnya genetika mempelajari dua aspek yang saling kontradiksi, yaitu kemiripan anak dengan tetuanya dan perbedaan antara anak dengan tetuanya serta perbedaan sesama anak. Jadi genetika mempelajari tentang pewarisan dari kesamaan dan variasi sifat antar individu.

Dengan berdasarkan ilmu genetika tersebut, dapat dilakukan pendeteksian terhadap kemiripan sifat dua individu. Pendeteksian tersebut dapat dilakukan dengan menggunakan ilmu komputasi dengan menggunakan program dimana kemiripan sifat dideteksi dengan menggunakan program yang dapat melakukan pencocokan string dengan salah satu algoritma yang cukup terkenal, yaitu Knuth-Morris-Pratt.

III. DASAR TEORI

A. Pencocokan String

Algoritma pencarian string (atau terkadang disebut algoritma pencocokan string) adalah algoritma untuk melakukan pencarian semua kemunculan string yang berukuran pendek dalam sebuah string yang berukuran jauh lebih panjang. String yang berukuran pendek tersebut biasa disebut dengan *pattern* dan string yang berukuran jauh lebih panjang disebut dengan teks. Algoritma ini adalah salah satu algoritma yang sangat penting dalam algoritma string.

Algoritma pencarian string dapat dibagi menjadi beberapa jenis berdasarkan jumlah pattern yang digunakan, yaitu algoritma dengan *pattern* tunggal, algoritma dengan beberapa *pattern* yang berhingga jumlahnya, dan algoritma dengan *pattern* yang tidak berhingga jumlahnya. Beberapa contoh algoritma yang dapat digunakan untuk melakukan pencarian string dengan *pattern* tunggal adalah algoritma pencarian string Naïve, Rabin-Karp, *Finite-state automaton*, Knuth-Morris-Pratt (KMP), Boyer-Moore, dan Bitap. Sedangkan contoh algoritma pencarian string yang dapat digunakan untuk beberapa *pattern* adalah algoritma pencocokan string Aho-Corasick, Commentz-Walter, dan Rabin-Karp. Pada algoritma pencocokkan string dengan *pattern* yang tak berhingga, *pattern* pada umunya direpresentasikan sebagai *regular grammar* atau *regular expression.*

B. Algoritma Knuth-Morris-Pratt

Algoritma KMP dikembangkan oleh D. E. Knuth,

bersama-sama dengan J. H. Morris dan V. R. Pratt. Algoritma Knuth-Morris-Pratt melakukan perbandingan dari bagian kiri ke kanan. Algoritma ini merupakan hasil modifikaasi dari algoritma *Brute Force*. Pada algoritma *Brute Force*, setiap kali ditemukan ketidakcocokan *pattern* dengan teks, maka *pattern* digeser satu karakter ke kanan. Sedangkan pada algoritma Knuth-Morris-Pratt, kita memanfaatkan informasi yang digunakan untuk melakukan sejumlah pergeseran. Algoritma menggunakan informasi tersebut untuk membuat pergeseran yang lebih jauh, tidak hanya satu karakter per karakter seperti pada algoritma *Brute Force*. Dengan memanfaatkan algoritma Knuth-Morris-Pratt ini, waktu pencarian menjadi dapat lebih singkat. Perbedaan yang terjadi cukup signifikan.

Berikut adalah definisi dari algoritma Knuth-Morris-Pratt.

Misalkan A adalah alfabet dan x = x1x2…xk , k ∈N, adalah string yang panjangnya k yang dibentuk dari karakter-karakter di dalam alfabet A. Awalan (prefix) dari x adalah upa-string
(substring) u dengan

$u = x1x2…xk – 1$, $k \in \{1, 2, …, k – 1\}$

dengan kata lain, x diawali dengan u.

Akhiran (suffix) dari x adalah upa-string (substring) u dengan

$u = xk – b\ xk – b + 1\ …xk$, $k \in \{1, 2, …, k – 1\}$

dengan kata lain, x diakhiri dengan v.

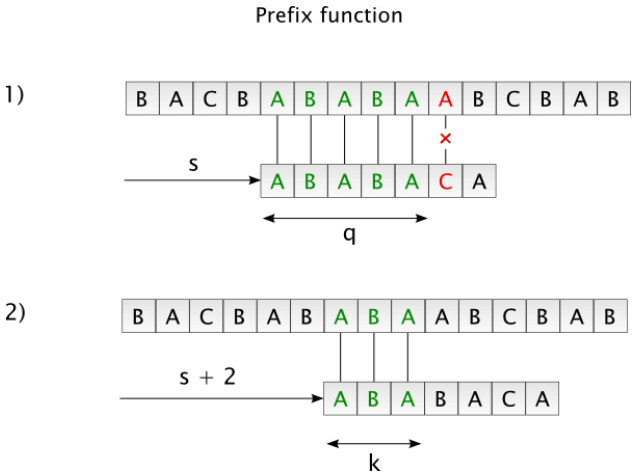
Pinggiran (border) dari x adalah upa-string r sedemikian sehingga

$r = x1x2…xk – 1$ dan $u = xk – b\ xk – b + 1\ …xk$, $k \in \{1, 2, …, k – 1\}$

dengan kata lain, pinggiran dari x adalah upastring yang keduanya awalan dan juga akhiran
sebenarnya dari x.

Fungsi pinggiran *(Border Function*) b(j) didefinisikan sebagai ukuran awalan terpanjang dari P yang merupakan akhiran dari P[1..j]. Sebagai contoh, tinjau pattern P = ababaa. Nilai F untuk setiap karakter di dalam P adalah sebagai berikut:

J	1	2	3	4	5	6
P[j]	a	b	a	b	a	a
b[j]	0	0	1	2	3	1



Gambar 1: Ilustrasi algoritma Knuth-Morris-Pratt
Sumber:
http://www.codeproject.com/KB/recipes/boolean-text-search/image2.png diakses tanggal 4 Mei 2015

Secara sistematis, langkah-langkah yang dilakukan algoritma Knuth-Morris-Pratt pada saat mencocokkan string:

1. Algoritma Knuth-Morris-Pratt mulai mencocokkan pattern pada awal teks.
2. Dari kiri ke kanan, algoritma ini akan mencocokkan karakter per karakter pattern dengan karakter di teks yang bersesuaian, sampai salah satu kondisi berikut dipenuhi:
    1. Karakter di pattern dan di teks yang dibandingkan tidak cocok (mismatch).
    2. Semua karakter di pattern cocok. Kemudian algoritma akan memberitahukan penemuan di posisi ini.
3. Algoritma kemudian menggeser pattern berdasarkan tabel next, lalu mengulangi langkah 2 sampai pattern berada di ujung teks.

Berikut adalah algoritma untuk menghitung fungsi pinggiran

```
procedure HitungPinggiran(input m :
integer, P : array[1..m] of
char, output b : array[1..m] of
integer)
{ Menghitung nilai b[1..m] untuk
pattern P[1..m] }
Deklarasi
      k,q : integer
Algoritma
      b[1]←0
      q←2
      k←0
      for q←2 to m do
            while ((k > 0) and (P[q] ≠
            P[k+1])) do
                  k←b[k]
            endwhile
            if P[q]=P[k+1] then
                  k←k+1
            endif
```

```
            b[q]=k
      endfor
```

Berikut adalah *pseudocode* algoritma KMP.

```
  procedure KMPsearch(input m, n : integer, input P : array[1..m] of char,
  input T : array[1..n] of char, output idx : integer)
  { Mencari kecocokan pattern P di dalam teks T dengan algoritma
  Knuth-Morris-Pratt. Jika ditemukan P di dalam T, lokasi awal
  kecocokan disimpan di dalam peubah idx.
  Masukan: pattern P yang panjangnya m dan teks T yang panjangnya n.
  Teks T direpresentasika sebagai string (array of character)
  Keluaran: posisi awal kecocokan (idx). Jika P tidak ditemukan, idx = -1. }
  Deklarasi
    i, j : integer
    ketemu : boolean
    b : array[1..m] of integer
    procedure HitungPinggiran(input m : integer, P : array[1..m] of char, output b : array[1..m] of integer)
    { Menghitung nilai b[1..m] untuk pattern P[1..m] }

  Algoritma
    HitungPinggiran(m, P, b)
    j←0
    i←1
    ketemu←false
    while (i ≤ n and not ketemu) do
      while((j > 0) and (P[j+1]≠T[i])) do
        j←b[j]
      endwhile
      if P[j+1]=T[i] then
        j←j+1
      endif
      if j = m then
        ketemu←true
      else
       i←i+1
      endif
    endwhile
    if ketemu then
      idx←i-m+1 { catatan: jika indeks array dimulai dari 0, maka idx←i-m }
    else
      idx←-1
  endif
```

Kompleksitas algoritma KMP adalah *O(m+n)*, dimana m adalah panjang *pattern* dan n adalah panjang teks. Kompleksitas berikut berasal dari *O(m)* adalah kompleksitas yang dibutuhkan untuk menghitung fungsi pinggiran dan *O(n)* yang dibutuhkan untuk melakukan pencarian string.

C. Genetika

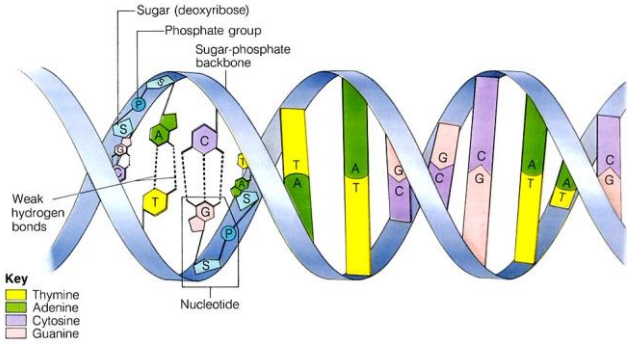
Genetika adalah cabang ilmu biologi yang mempelajari pewarisan sifat pada organismemaupun suborganisme (seperti virus dan prion). Secara singkat dapat juga dikatakan bahwa genetika adalah ilmu tentang gendan segala aspeknya. Istilah "genetika" diperkenalkan oleh William Bateson pada suatu surat pribadi kepada Adam Chadwick dan ia menggunakannya pada Konferensi Internasional tentang Genetika ke-3 pada tahun 1906.

Gambar 2: Genetika
Sumber: http://www.encyclopedia.com/doc/1G2-3045302452.html diakses tanggal 4 Mei 2015

D. DNA

DNA (deoxyribonucleic acid atau asam deoksiribonukleat) adalah sejenis biomolekul yang menyimpan dan menyandi instruksi-instruksi genetikasetiap organisme dan banyak jenis virus. Instruksi-instruksi genetika ini berperan penting dalam pertumbuhan, perkembangan, dan fungsi organisme dan virus. DNA merupakan asam nukleat; bersamaan dengan protein dankarbohidrat, asam nukleat adalahmakromolekul esensial bagi seluruhmakhluk hidup yang diketahui. Kebanyakan molekul DNA terdiri dari dua unting biopolimer yang berpilin satu sama lainnya membentuk heliks ganda. Dua unting DNA ini dikenal sebagai polinukleotida karena keduanya terdiri dari satuan-satuan molekul yang disebut nukleotida. Tiap-tiap nukleotida terdiri atas salah satu jenis basa nitrogen (guanina (G), adenina (A), timina (T), atau sitosina (C)), gula monosakarida yang disebutdeoksiribosa, dan gugus fosfat. Nukleotida-nukelotida ini kemudian tersambung dalam satu rantai ikatan kovalen antara gula satu nukleotida dengan fosfat nukelotida lainnya. Hasilnya adalah rantai punggung gula-fosfat yang berselang-seling. Menurut kaidah pasangan basa (A dengan T dan C dengan G), ikatan hidrogen mengikat basa-basa dari kedua unting polinukleotida membentuk DNA unting ganda.

Dua unting DNA bersifat anti-paralel, yang berarti bahwa keduanya berpasangan secara berlawanan. Pada setiap gugus gula, terikat salah satu dari empat jenis nukleobasa. Urutan-urutan empat nukleobasa di sepanjang rantai punggung DNA inilah yang menyimpan kode informasi biologis. Melalui proses biokimia yang disebut transkripsi, unting DNA digunakan sebagai templat untuk membuat unting RNA. Unting RNA ini kemudian ditranslasikan untuk menentukan urutan asam amino protein yang dibangun.



Gambar 3: Struktur DNA
Sumber:
https://karimedalla.files.wordpress.com/2012/11/dna-structure.jpg pada tanggal 4 Mei 2015

Struktur kimia DNA yang ada membuatnya sangat cocok untuk menyimpan informasi biologis setiap makhluk hidup. Rantai punggung DNA resisten terhadap pembelahan kimia, dan kedua-dua unting dalam struktur unting ganda DNA menyimpan informasi biologis yang sama. Karenanya, informasi biologis ini akan direplikasi ketika dua unting DNA dipisahkan. Sebagian besar DNA (lebih dari 98% pada manusia) bersifat non-kode, yang berarti bagian ini tidak berfungsi menyandikan protein.

Dalam sel, DNA tersusun dalam kromosom. Semasa pembelahan sel, kromosom-kromosom ini diduplikasi dalam proses yang disebut replikasi DNA. Organisme eukariotik (hewan, tumbuhan, fungi, dan protista) menyimpan kebanyakan DNA-nya dalam inti sel dan sebagian kecil sisanya dalam organel seperti mitokondria ataupun kloroplas. Sebaliknya organisme prokariotik (bakteri dan arkaea) menyimpan DNA-nya hanya dalam sitoplasma. Dalam kromosom, protein kromatin seperti histon berperan dalam penyusunan DNA menjadi struktur kompak. Struktur kompak inilah yang kemudian berinteraksi antara DNA dengan protein lainnya, sehingga membantu kontrol bagian-bagian DNA mana sajakah yang dapat ditranskripsikan.

DNA merupakan sebuah polimer yang terdiri dari satuan-satuan berulang yang disebut nukleotida. Tiap-tiap nukleotida terdiri dari tiga komponen utama, yakni gugusfosfat, gula deoksiribosa, dan basa nitrogen (nukleobasa). Pada DNA, nukleobasa yang ditemukan adalah Adenina (A), Guanina (G), Sitosina (C) dan Timina (T). Nukleobasa yang terhubung dengan sebuah gugus gula disebut sebagai nukleosida, dan nukleosida yang terhubung dengan satu atau lebih gugus fosfat disebut sebagainukleotida. Polimer yang terdiri dari nukleotida yang saling terhubung menjadi satu rantai disebut sebagai polinukleotida. Sehingga DNA termasuk pula ke dalam polinukleotida.

Rantai punggung unting DNA terdiri dari gugus fosfat dan gula yang berselang-seling. Gula pada DNA adalah gula pentosa (berkarbon lima), yaitu 2-deoksiribosa. Dua gugus gula terhubung dengan fosfat melalui ikatan fosfodiester antara atom karbon ketiga pada cincin satu gula dan atom karbon kelima pada gula lainnya. Ikatan yang tidak simetris ini membuat DNA memiliki arah atau orientasi tertentu. Pada struktur heliks ganda, orientasi rantai nukleotida pada satu unting berlawanan dengan orientasi nukleotida unting lainnya. Hal ini disebut sebagai*antiparalel*. Kedua ujung asimetris DNA disebut sebagai 5' (lima prima) dan 3' (tiga prima). Ujung 5' memiliki gugus fosfat terminus, sedangkan ujung 3' memiliki gugus hidroksi terminus. Salah satu perbedaan utama DNA dan RNA adalah gula penyusunnya, yakni gula 2-deoksiribosa pada DNA digantikan gula ribosa pada RNA.

Dalam organisme hidup, DNA biasanya ditemukan dalam bentuk berpasangan dan terikat kuat. Dua unting DNA saling berpilin membentuk heliks ganda. Heliks ganda ini distabilisasi oleh dua gaya utama: ikatan hidrogen antar nukleotida dan interaksi tumpukan antar nukleobasa aromatik. Dalam lingkungan sel yang berair, ikatan πkonjugasi antar basa nukleotida tersusun tegak lurus terhadap sumbu pilinan DNA. Hal ini meminimalisasi interaksi dengan cangkang solvasi, dan sehingganya menurunkan energi bebas Gibbs.

Struktur DNA semua jenis spesies terdiri dari dua rantai heliks yang berpilin dengan jarak antar putaran heliks 34 Å (3,4 nanometer) dan jari-jari 10 Å (1.0 nanometer). Menurut kajian lainnya, ketika diukur menggunakan larutan tertentu, rantai DNA memiliki lebar 22-26 Å (2,2-2,6 nanometer) sedangkan satu satuan nukleotida memiliki panjang 33 Å (0,33 nm). Walaupun satuan nukleotida ini sangatlah kecil, polimer DNA dapat memiliki jutaan nukleotida yang terangkai seperti rantai. Misalnya, kromosom 1 yang merupakan kromosom terbesar pada manusia mengandung sekitar 220 juta pasangan basa.

IV. ANALISIS

Seperti yang telah kita ketahui, struktur DNA yang merupakan kumpulan basa nitrogen dapat dimodelkan dengan rantai berurutan. Rantai dari suatu organisme dapat dimodelkan dengan teks atau string yang dalam kasus ini akan menjadi tempat pencarian DNA.

Dengan menjalankan algoritma Knuth-Morris-Pratt, kita dapat mengetahui letak kesamaan dari dua buah DNA. Untuk melakukan hal ini, pertama-tama perlu dicari DNA kedua individu yang akan dijadikan sampel. Berikut adalah sampel DNA yang akan digunakan.

Sampel 1:

```
cctattagaa cgcgaatcgc gaacgcgaat
atctgtaaaa agcggaatct
```

Sampel 2:

```
gaacgcgaat gcctctctct
```

Dengan menganggap sampel 1 adalah sampel utama dan sampel 2 adalah sampel yang berisi *pattern* yang ingin dicocokkan. Dilakukan pencarian kemungkinan seluruh substring pada sampel 2. Berikut adalah hasilnya.

```
gaacgcgaat
aacgcgaatg
acgcgaatgc
cgcgaatgcc
…
gcctctctct
```

Untuk setiap kemungkinan *pattern* yang ada pada sampel 2, dilakukan pencarian pada sampel 1 seluruh kemungkinan *pattern* yang ada. Apabila ternyata *pattern* ditemukan pada sampel 1, maka simpan lokasi hasil temuan *pattern* pada string tersebut. Apabila ternyata tidak ditemukan, lakukan pencarian untuk kemungkinan *pattern* berikutnya.

Untuk setiap kali pencocokan *pattern* di dalam string, jalankan algoritma Knuth-Morris-Prath dengan menghitung fungsi pembatas dari *pattern*, dan kemudian melakukan pencarian string. Apabila ternyata karakter pertama tidak sama dengan karakter pertama pada *pattern*, lakukan pergeseran ke karakter berikutnya sesuai dengan nilai dari fungsi pembatar pasa karakter pertama. Apabila ternyata karakter tersebut sama, lakukan pencocokan untuk karakter kedua pada sampel dengan karakter kedua pada *pattern.*

Apabila ternyata *pattern* ditemukan pada sampel 1, maka simpan lokasi hasil temuan *pattern* pada string tersebut. Apabila ternyata tidak ditemukan, lakukan pencarian untuk kemungkinan *pattern* berikutnya.

Sebagai contoh, untuk kemungkinan pertama lakukan pencarian *pattern* `gaacgcgaat` pada string `cctattagaa cgcgaatcgc gaacgcgaat atctgtaaaa agcggaatct`.

Berikut adalah fungsi pembatas dari *pattern*.

j	1	2	3	4	5	6	7	8	9
P[i]	g	a	a	c	g	c	a	a	t
b[j]	0	0	1	1	2	2	2	2	2

Berikut adalah hasi dari algoritma Knuth-Morris-Pratt.

```
Ditemukan kesamaan pada:
gaacgcgaat
gaacgcgaat
```

Dari hasil tersebut ditemukan kesamaan DNA dari kedua individu. Setelah didapatkan kesamaan tersebut, dapat dilakukan analisis lebih lanjut terhadap hasil algoritma dengan menggunakan metode DNA *testing*, sesuai dengan kebutuhan yang diinginkan.

V. KESIMPULAN

Kemiripan sifat antara dua individu dapat dideteksi dengan menggunakan algoritma pencocokkan string Knuth-Morris-Prath dengan mencari seluruh *pattern* yang sama pada DNA kedua individu.

VI. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Tuhan Yang Maha Esa yang telah menyertai dalam pembuatan makalah ini. Penulis juga mengucapkan terima kasih kepada Dr. Ir. Rinaldi Munir, M.T. dan Dr. Nur Ulfa Maulidevi, S.T, M.Sc. selaku dosen pembimbing mata kuliah IF2211 Strategi Algoritma, Program Studi Teknik Informatika, Institut Teknologi Bandung yang telah memberi berbagai pegetahuan, terutama dalam bidang algoritma pencocokan string dan kepada seluruh pihak yang turut membantu dalam pembuatan makalah ini.

REFERENSI

[1] Munir, Rinaldi, *Strategi Algoritma*. Bandung: Percetakkan ITB, 2004.

[2] http://www-igm.univ-mlv.fr/~lecroq/string/index.html, diakses tanggal 4 Mei 2015.

[3] http://www.personal.kent.edu/~rmuhamma/Algorithms/MyAlgorithms/StringMatch/kuthMP.htm, diakses tanggal 4 Mei 2015.

[4] http://bunghatta.ac.id/artikel-137-dasar-fisologis-pewarisan-sifat.html, diakses tanggal 4 Mei 2015.

[5] http://www.encyclopedia.com/doc/1G2-3045302452.html, diakses tanggal 4 Mei 2015.

[6] http://www.sersc.org/journals/IJAST/vol47/2.pdf, diakses tanggal 4 Mei 2015.

PENYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 11 Desember 2014

Irene Wiliudarsan
13513002