Algoritma Brute Force, Knuth Morris Pratt, dan Boyer Moore untuk deteksi sel DNA kanker lebih dini

Enjella Melissa Nababan, NIM 13510109
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
13510109@itb.ac.id

***Abstract*— Bioinformatika adalah aplikasi teknologi komputer untuk manajemen dan analisis data biologis. Hasilnya adalah bahwa komputer yang digunakan untuk mengumpulkan, menyimpan, menganalisis dan menggabungkan data biologis. Tujuan dari bio-informatika adalah untuk mengungkap kekayaan informasi biologis tersembunyi dalam massa data dan memperoleh suatu jelas wawasan biologi dasar organisme. yang paling aplikasi terkenal dari bioinformatika adalah analisis urutan. Pada analisis urutan, DNA urutan berbagai penyakit yang disimpan didatabase untuk memudahkan pengambilan dan perbandingan. Ketika kita tahu urutan tertentu merupakan penyebab penyakit,melacak urutan dalam DNA dan jumlah kejadianurutan mendefinisikan intensitas penyakit. Karena DNA adalah database yang besar, diusulkan algoritma String dan Pola pencocokanuntuk mengetahui urutan tertentu dalam DNA yang diberikan. Makalah ini berfokus pada pendekatan baru untuk pendeteksian dalam database gen. Makalah ini juga menekankan tentang bagaimana penyakit ini dapat berubah dari orang tua untuk anak-anak mereka dan efisien metode untuk mengidentifikasi adanya penyakit pada herediter dasar dan dampaknya.**

***Index Terms*— Bioinformatika, pencocokan pola, analisis urutan, algoritma KMP, DNA**

I. PENDAHULUAN

Bioinformatika adalah penelitian para ilmuwan yang merupakan antarmuka ilmu biologi dan komputasi. Tujuan utama dari bioinformatika adalah untuk mengungkap kekayaan informasi biologis tersembunyi dalam massa data dan memperoleh suatu kejelasan wawasan biologi dasar organisme. Imu pengetahuan ini bisa memiliki dampak besar pada bidang beragam seperti kesehatan manusia, pertanian, lingkungan, energi dan bioteknologi. ada banyak lainnya aplikasi bioinformatika, termasuk memprediksi untai protein keseluruhan, belajar bagaimana gen mengekspresikan diri mereka ke berbagai spesies, dan membangun model yang kompleks dari sel keseluruhan.

Untuk meningkatkan daya komputasi informasi genetik dan molekuler, bioinformatika membawa perubahan drastis dan memungkinkana untuk membuat sesuatu model yang kompleks dan luar biasa. Deoxyribonucleic acid atau yang sering disingkat dengan DNA adalah untaian DNA mengandung tiga komponen, yaitu :

1. deoksiribosa,
2. serangkaian fosfat
3. empat basa nitrogen

Empat basa yang ada dalam DNA adalah :

1. adenine (A),
2. thymine (T),
3. guanine (D),
4. cytosine (C).

Gabungan deoksiribosa dengan fosfat akan membentuk “tulang punggung” DNA. Thymine dan Adenine selalu datang dalam pasangan. Demikian juga, guanine dan cytosine datang bersama-sama.

Setiap manusia memiliki gen yang unik. Gen tersebut dibuat dari DNA. Karena itu, urutan DNA dari setiap manusia sangatlah unik. Namun yang mengherankan, sekuen DNA dari semua manusia tingkat kemiripannya adalah 99,9 % identik, yang berarti hanya ada 0,1 % perbedaan. DNA terkandung dalam setiap sel yang hidup dari suatu organisme dan hal ini merupakan masalah kode genetic organisme.Kode genetik adalah seperangkat urutan yang mendefinisikan apa protein untuk membangun dalam seorang organisme.

Organisme harus mereplikasi dan / atau mereproduksi jaringan untuk melanjutkan kehidupan, yang pastinya harus ada beberapa cara dari pengkodean kode genetik yang unik untuk protein yang digunakan dalam pembuatan jaringan tersebut. Kode genetik adalah informasi yang akan dibutuhkan untuk pertumbuhan biologis organisme dan reproduksi gen warisan.

Urutan DNA adalah representasi dari kode genetic yang terkandung dalam suatu organisme. Para peneliti molekuler biologi memerlukan usaha yang besar untuk membandingkan urutan DNA. Urutan DNA adalah representasi dari kesamaan antara dua atau lebih bagian dari kode genetik. Hal ini digunakan untuk dibandingkan dalam sebuah kuantitatif. Hal inilah yang digunakan untuk perbandingan dalam menemukan divergensi evolusi, asal-usul penyakit, dan cara untuk menerapkan kode genetik dari satu organisme ke yang lain.

Urutan DNA memberikan representasi dari string nukleotida yang terkandung dalam untai DNA

Contohnya: A T G C G AT A C A A G T T G T A
Nukleotida yang dapat diperoleh adalah A, G, C, dan T.

Istilah urutan DNA meliputi metode biochemical untuk menentukan urutan basa nukleotida, adenine, guanine, cytosine, dan thymine dalam sebuah DNA oligonucleotide. Urutan DNA diwariskan inti informasi genetik dalam plasmids, mitokondria, dan kloroplas yang membentuk dasar untuk para program perkembangan dari semua organisme hidup. Menentukan urutan DNA berguna dalam penelitian dasar mempelajari proses biologis mendasar, serta diterapkan bidang forensik diagnostik penelitian.

Karena DNA merupakan kunci untuk semua hidup organisme, pengetahuan dari urutan DNA dapat berguna dalam hampir setiap subjek biologis.

Kebanyakan database biologis terdiri dari panjang senar nukleotida (guanina, adenina, timina, uralic sitosina dan) dan / atau asam amino(threonine, serine, glisina, dll.). Secara berurutan nukleotida atau asam amino mewakili gen tertentu atau protein.

II. PENYAKIT

Sebuah gejala yang tidak sehat atau sebuah penyakit spesifik dalam tubuh adalah disebut sebagai penyakit. Istilah penyakit merujuk untuk kondisi abnormal sebuah organisme yang rusak fungsinya. Pada manusia, kata “penyakit” ini sering digunakan lebih luas lagi untuk merujuk untuk salah satu kondisi yang menyebabkan ketidaknyamanan, disfungsi, tertekan, masalah sosial, atau kematian untuk orang lain.Secara luas, penyakit dapat mencakup hal-hal seperti luka, cacat, gangguan , sindrom infeksi, gejala, perilaku, menyimpang dan rariasi khas lain dari struktur dan fungsi tubuh organisme

III. PENYAKIT PADA DNA

Nukleotida dikelompokkan untuk membentuk kata-kata dan kata-kata ini akan dibuat untuk membuat kalimat. Kalimat yang dimaksud dalam hal ini adalah gen. Gen dapat membuat sebuah sel melakukan sebuah fungsi yang spesifik. Ketika sel tidak menurut pada perintah gen maka telah terjadi sesuatu hal yang tidak diinginkan atau sel bermutasi yang dapat menyebabkan penyakit.

Setiap penyakit akan mempunyai familinya sendiri dengan urutan DNA yang dia suka sehingga intensitas penyakit yang terjadi bermutasi di urutan DNA yang ada. Contohnya penyakit kanker. Kanker adalah penyakit yang disebabkan karena terjadinya pertumbuhan sel yang tidak terkendali pada sel tersebut yang merupakan hasil dari perubahan atau mutasi dalam bentuk material secara genetik. Lebih tepatnya, munculnya kanker mungkin memerlukan akumulasi multiple mutasi dimana sel dimungkinkan untuk diambil alih dan keluar dari jaringan regulasi yang memastikan kerjasama antara sel dan gen terputus.

Konsep ini adalah dimaksud sebagai multi-stage Carcinogenesis. Sekali sebuah sel kanker telah dibuat itu dapat menjalani proses yang dikenal sebagai clonal ekspansi. Hal ini memberikan kewenangan untuk keturunan sel dalam hal pembelahan sel dan populasi sel secara mandiri.

III. INSTABILITAS GENERIK

Banyak sel-sel kanker menunjukkan sejumlah besar perubahan genetik yang berkisar dari mutasi skala kecil hingga besar aberrations kromosom. Sementara ini sebuah observasi yang menarik tidak membuktikan bahwa secara genetic sel-sel yang tidak stabil. Perubahan tersebut bisa datang melalui berbagai faktor, seperti paparan luas kerusakan di beberapa titik dalam waktu, atau spesifik kondisi selektif. Ketidakstabilan genetik didefinisikan sebagai peningkatan tingkat di sel yang mana dapat mengakuisisi kelainan sel yang memiliki cacat.

Dalam perbaikan spesifik , memang telah membuktikan bahwa sel-sel kanker banyak ditandai dengan peningkatan perubahan genetik.

IV. MENDETEKSI PENYAKIT

Ketika kita tahu sebuah urutan tertentu adalah penyebab untuk sebuah penyakit, para peneliti melacak dari urutan di DNA dan jumlah kejadian. Dari urutan dapat didefinisikan intensitas penyakit. Sebuah database besar tentang DNA memerlukan algoritma yang efisien untuk mencari tahu sebuah urutan tertentu dalam bentuk yang diberikan DNA. Kita harus menemukan jumlah pengulangan yang dari dan awal dan akhir indeks index dari urutan sehingga dapat digunakan untuk diagnosis penyakit dan juga intensitas penyakit dari menghitung jumlah dari pola pencocokan string yang terjadi di sebuah database gen DNA.

V. TRANSFORMASI PENURUNAN SIFAT GENETIK

Sejak anak-anak mewarisi gen mereka dari orang tua mereka, mereka dapat mewarisi setiap cacat genetik. Anak-anak dan saudara kandung dari seorang pasien umumnya memiliki 50 % kesempatan untuk juga menjadi pengaruh terhadap sebuah penyakit.

Pengujian genetik dapat mengidentifikasi orang-orang anggota keluarga yang membawa familial mutasi yang tidak biasa dan harus menjalani tahunan skrining tumor dari usia dini.

Sebaliknya, pengujian genetik dapat juga mengidentifikasi anggota keluarga yang tidak membawa familial tidak membawa sifat mutasi dan tidak butuh untuk menjalani meningkatnya tumor

Surveillance recom-mended diberikan untuk pasien dengan mutasi. yang tidak biasa Pola pemandangan yang

tidak biasa dalam bentuk untaian DNA yang mencerminkan peningkatan mutasi yang tidak biasa dalam sel. Semua familial sindrom kanker yang disebabkan oleh cacat dalam gen penting untuk dicegah perkembangan tumornya.

Semua orang membawa dua salinan dari gen kanker ini dalam setiap sel, dan pembangunan tumor hanya terjadi jika kedua gen salinan menjadi cacat di rentan tertentu. Penurunan dari kedua gen salinan dalam rentan sel membutuhkan dua peristiwa independen utnuk mempengaruhi sesama gen. Tidak mungkin ada peristiwa pasien dengan mutasi yang tidak biasa, namun, sudah membawa cacat pada salah satu dari salinan gen dalam tiap sel di tubuh mereka. Hanya satu acara tambahan mempengaruhi se pasien tersebut. Gen yang utuh menyalin di sel-sel rentan tetrtentu yang diperlukan untuk memungkinkan pembangunan. Karena itu, pasien dengan mutasi yang tidak biasa dapat mengembangkan risiko dengan benjolan-benjolan yang tidak biasa yang terkait dengan mutasi dalam tubuhnya.

Pengujian genetik dapat membantu untuk mendiagnosis dan mendeteksi cacat pada para pasien yang selnya bermutasi.

VI. MASALAH POLA PADAN

Masalah klasik dalam masalah pola padan kita lambangkan dengan pemisalah sebagai berikut. T merupakan text dengan panjang n yang merepresentasikan urutan DNA yang akan diperiksa. Sedangkan P merupakan pola dengan panjang m yang merepresentasikan potongan DNA untuk mencari substring dari text yang memilikii pola yang sama.

Misalkan kita diberikan text sebagai berikut :

T = abacaabaccabacabaabb

dan pola P sebagai berikut :

P = abacab

Dengan kasat mata kita dapat menghitung dan mengetahui bahwa P merupakan substring dari T dengan P = T [10…15].

Ada beberapa algoritma pola padan yang dapat digunakan dalam hal ini. Algoritma ini dapat digunakan dalam mencocokkan P terhadap T yang ada di database DNA yang dimiliki. Algoritma tersebut adalah :

1. Algoritma Brute- Force
2. Algoritma Boyer-Moore
3. Algoritma Knuth Morris Pratt

VII. POLA PADAN ALGORITMA BRUTE FORCE

Brute force merupakan pola desain algoritmik dengan sebuah teknik yang kuat. Ketika kita punya sesuatu dan berharap untuk dapatr melalkuakn pencarian tanpa teknik khusus dan pasti menghasilkan jawaban yang diharapkan terlepas dari kapasitas waktu yang diperlukan untuk mendapatkan jawaban merupakan ciri khas gaya algoritma Brute force.

Algoritma *brute force* umumnya tidak “cerdas” dan tidak mangkus, karena ia membutuhkan jumlah komputasi yang besar dalam penyelesaiannya. Kata “force” mengindikasikan “tenaga” ketimbang “otak”. Kadang-kadang algoritma *brute force* disebut juga algoritma naif (*naïve algorithm*).

Algoritma *brute force* lebih cocok untuk masalah yang berukuran kecil. Pertimbangannya: sederhana dan implementasinya mudah . Algoritma *brute force* sering digunakan sebagai basis pembanding dengan algoritma yang lebih mangkus.

Contoh implementasi algoritma brute force.

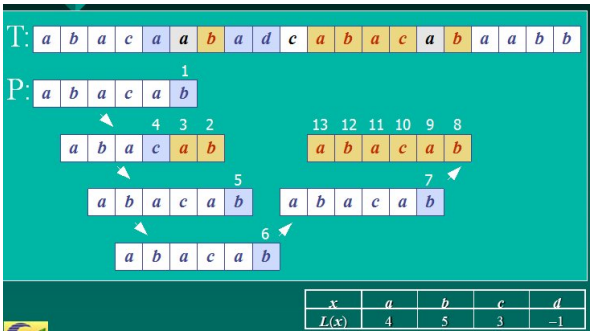
Pattern: NOT
Teks: NOBODY NOTICED HIM

NOBODY **NOT**ICED HIM
1 NOT
2 NOT
3 NOT
4 NOT
5 NOT
6 NOT
7 NOT
8 **NOT**

VIII. POLA PADAN ALGORITMA BOYER MOORE

Algoritma *Boyer-Moore* adalah algoritma pencarian *string* yang mencari dengan cara membandingkan sebuah huruf dengan huruf yang ada di *pattern* yang dicari, dan menggeser *pattern* tersebut hingga posisinya sama dengan teks yang dicari dan membandingkan kata tersebut. Cara ini disebut *character jump.*

Gambaran kinerja algoritma KMP adalah sebagai berikut :



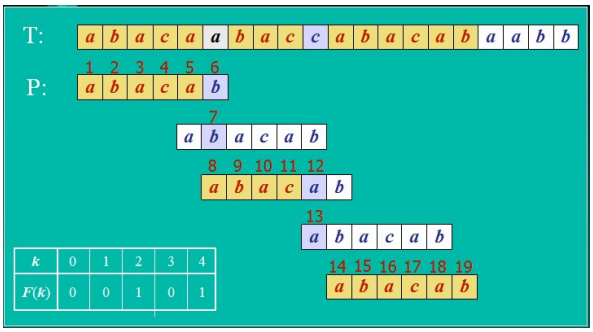
Contoh Algoritma *Boyer-Moore* dalam *Java:*

```
public static int bmMatch(String text,
String pattern)
{int last[] = buildLast(pattern);
int n = text.length();
int m = pattern.length();
int i = m-1;
if (i > n-1)
return -1;
int j = m-1;
do {
if (pattern.charAt(j) ==
text.charAt(i))
if (j == 0)
return i; // match
else { i--;
j--;
}
else { int lo = last[text.charAt(i)];
i = i + m - Math.min(j, 1+lo);
j = m - 1;
}
} while (i <= n-1);
return -1;
}
public static int[] buildLast(String
pattern)
{
int last[] = new int[128]; // ASCII
char set
for(int i=0; i < 128; i++)
last[i] = -1; // initialize array
for (int i = 0; i < pattern.length();
i++)
last[pattern.charAt(i)] = i;
return last;
}
```

IX. POLA PADAN ALGORITMA KNUTH MORRIS PRATT

Algoritma *Knuth-Morris-Pratt* adalah algoritma pencarian *string* yang mencari dengan cara menghitung dari dimulai dari ketidakcocokan ditemukan, dari ketidakcoockan tersebut akan dhitung dari mana pencarian selanjutnya sebaiknya dimulai. Algoritma *Knuth-Morris-Pratt* menggunakan fungsi pembatas (*border function)* yang digunakan untuk menghitung urutan keberapa perbandingan harus dilakukan. *Border function* dihitung dengan menghitung *panjang prefix* yang ada disebuah *pattern* yang sama dengan *suffix*-nya.

Gambaran kinerja algoritma KMP adalah sebagai berikut :



Contoh Algoritma KMP dalam Java:

```
public static int kmpMatch(String
text, String pattern)
{
int n = text.length();
int m = pattern.length();
int fail[] = computeFail(pattern);
int i=0;
int j=0;
while (i < n) {
if (pattern.charAt(j) ==
text.charAt(i)) {
if (j == m - 1)
return i - m + 1; // match
i++;
j++;
}
else if (j > 0)
j = fail[j-1];
else
i++;
}
return -1; // no match
}
public static int[] computeFail(String
pattern)
{
int fail[] = new
int[pattern.length()];
Makalah IF3051 Strategi Algoritma –
Sem. I Tahun 2011/2012
fail[0] = 0;
int m = pattern.length();
int j = 0;
int i = 1;
while (i < m) {
if (pattern.charAt(j) ==
pattern.charAt(i)) { //j+1 chars match
fail[i] = j + 1;
i++;
j++;
}
else if (j > 0) // j follows matching
prefix
j = fail[j-1];
else { // no match
fail[i] = 0;
i++;
}
}
```

```
return fail;
```

X. KESIMPULAN

Bioinformatika dapat digunakan dalam mendiagnosis penyakit para pasien dengan cara dengan mengidentifikasi pola yang tidak biasa dalam DNA mereka.Database DNA tersebut dapat memberikan gambaran yang mewakili intensitas penyakit yang ada dalam tubuh orang tersebut.

Biouinformatika juga dapat membantu dalam deteksi dari kehadiran pola yang tidak biasa untuk para anak-anak dari terinfeksi orang, dimana ada hereditariliti yang ditransfer.

Dengan ini, aplikasi ini dapat diharapkan untuk membantu pendeteksian penyakit, khususnya kanker pada pusat penelitian dan di rumah sakit

REFERENCES

[1] Fast Pattern matching in strings, SIAM Journal of computer science, pp323 – 350, 1977, Knuth D., Morris J. and Pratt V.

[2] A Minimum Cost Process in Searching for a Set of similar DNA Sequence, International conference on Telecommunications and Informatics, pp348 – 353, Saman, Rahman, Ahmad, Osman.

[3] Fast practical Exact and Approximate Pattern Matching in Protein Sequences, C. S. Iliopoulos, Inuka Jayasekera1, and L. Mouchard.

[4] Whole – Genome DNA Sequencing, IEEE Computer society,2012, pp33 – 43, Gene Myers.

[5] A fast string-searching algorithm, Comm. Assoc. Comput. Mach., pp762 – 772, 1977, R S Boyer & J S Moore.

[6] Occurrences Algorithm for string searching based on Brute-force Algorithm, Journal of Computer Science, 82 – 86, 2012.

[7] Brute Force Algorithm, Christian Charras.

PERNYATAAN

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, 21 Desember 2012

Enjella Melissa Nababan