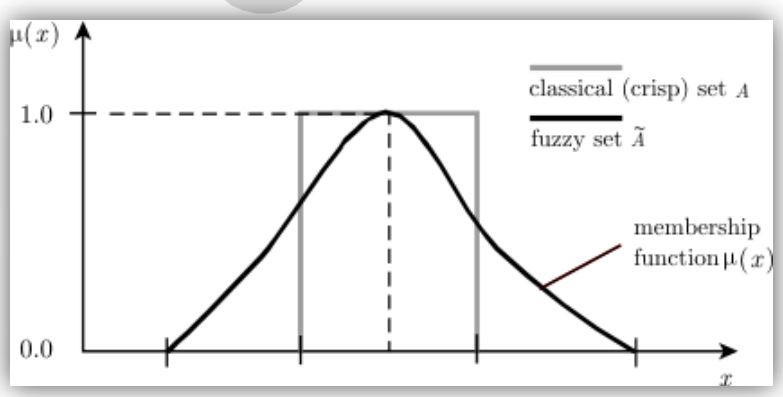
BAB II LANDASAN TEORI

Bab ini menjelaskan tentang berbagai teori yang digunakan untuk melakukan penelitian ini. Penjelasan meliputi teori himpunan *fuzzy*, pengukuran *fuzzy*, DNA sebagai himpunan *fuzzy* serta perhitungan ruang vektor polinukleotida. Selain itu, bab ini menjelaskan pengenalan mengenai DNA.

II.1. Himpunan Fuzzy

Himpunan *fuzzy* adalah konsep yang mendasari lahirnya logika *fuzzy*. Himpunan *fuzzy* adalah sebuah himpunan yang anggotanya memiliki derajat keanggotaan tertentu [ZAD65]. Setiap anggota memiliki derajat keanggotaan tertentu yang ditentukan oleh fungsi keanggotaan (*membership function*) tertentu atau disebut juga fungsi karakteristik (*characteristik function*).

Himpunan *crisp* adalah himpunan klasik yang telah dikenal secara umum. Himpunan *crisp* membedakan anggotanya dengan nilai nol atau satu, anggota himpunan atau bukan. Sebagai contoh himpunan yaitu, pada himpunan manusia. Himpunan wanita atau himpunan laki-laki dapat direpresentasikan dengan mudah dengan cara himpunan klasik. Akan tetapi, bagaimana merepresentasikan himpunan pada manusia muda atau tua. Muda atau tua itu cukup relatif tidak langsung terpisah hanya karena berbeda satu hari. Dalam hal ini himpunan *fuzzy* dapat memberikan mengelompokkan dengan memberi nilai derajat tertentu. Berbeda dengan himpunan klasik, keanggotaan himpunan *fuzzy* dapat bernilai parsial.



Gambar 1. Perbandingan Fungsi Keanggotaan Himpunan *Fuzzy* Terhadap Himpunan *Crips*.

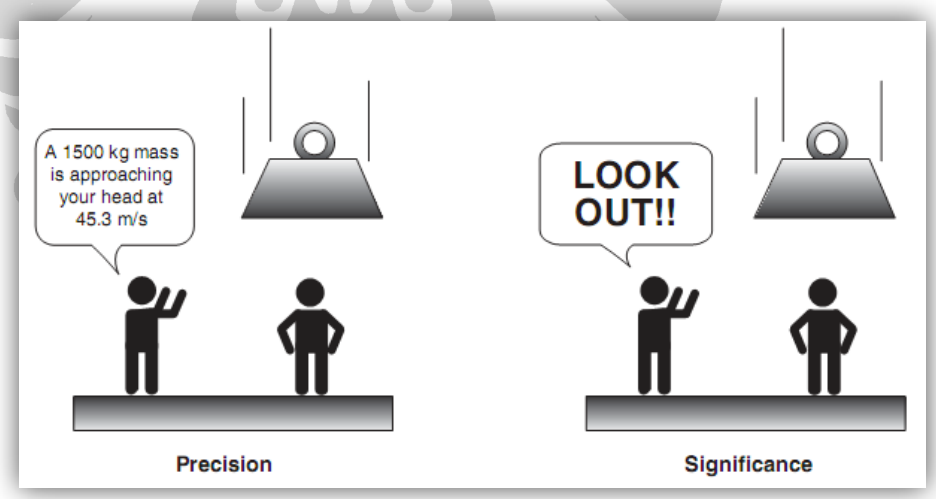
Fungsi keanggotaan didefinisikan sebagai berikut: Jika X adalah himpunan semesta, maka fungsi keanggotaan μ_A (fungsi keanggotaan /fungsi karakteristik A pada X) yang didefinisikan oleh himpunan *fuzzy* A memiliki ketentuan berikut:

$$\mu_A : X \rightarrow [0, 1] \qquad (2.1)$$

dimana [0,1] adalah interval bilangan real dari nol sampai dengan satu. Dua himpunan A dan B dinyatakan sama jika dan hanya jika $\mu_A(x) = \mu_B(x)$. Jika $\mu_A(x)$ bernilai nol, berarti x bukan anggota dari himpunan *fuzzy* A. Jika $\mu_A(x)$ bernilai satu, menunjukkan x adalah anggota penuh dari himpunan *fuzzy* A. Sementara nilai antara nol hingga satu menunjukkan bahwa x merupakan anggota dari himpunan *fuzzy* A secara parsial.

II.2. Logika Fuzzy

Logika *Fuzzy* adalah logika yang berbasiskan pada teori himpunan *fuzzy* dan diperkenalkan oleh Lotfi Zadeh [ZAD65]. Pada logika *fuzzy*, terdapat proses pemetaan dari suatu ruang input ke dalam suatu ruang output. Logika *fuzzy* terdiri dari tiga operator, yaitu *fuzzy negation*, t-norm, dan s-norm.



Gambar 2. Ilustrasi ketelitian (*precision*) dibandingkan dengan kepentingan (*significance*) [MAT08].

Logika *fuzzy* lebih mendekati masalah kepentingan (*significance*) dibandingkan masalah ketelitian (*precision*). Meskipun logika *fuzzy* memiliki ketelitian yang kurang teliti, tetapi lebih dekat dengan intuisi manusia.

II.2.1. Fuzzy Negation

Fuzzy negation adalah operasi negasi yang digunakan di logika *fuzzy* dan dituliskan dengan notasi $^{(n)}$. Berdasarkan definisi, *fuzzy negation* adalah sebuah fungsi $^{(n)}$: [0,1] → [1,0] yang memenuhi sifat – sifat berikut:

1. $0^{(n)} = 1$
2. $x_1^{(n)} > x_2^{(n)}$ jika $x_1 < x_2$
3. $(x^{(n)})^{(n)} = x$ (2. 2)

II.2.2. T-norm

T-norm adalah operasi konjungsi yang digunakan di logika *fuzzy*. Pada laporan ini, t-norm dituliskan dengan simbol T. Selain dapat melakukan operasi konjungsi di logika *fuzzy*, t-norm juga dapat digunakan sebagai basis untuk operator agregasi pada operasi himpunan *fuzzy*. Berdasarkan definisi, t-norm adalah sebuah fungsi T: [0,1] x [0,1] → [1,0] yang memenuhi sifat – sifat berikut:

1. $T(x,0) = 0$ dan $T(x,1) = x$
2. $T(x_1, x_2) = T(x_2, x_1)$
3. $T(x_1, T(x_2, x_3)) = T(T(x_1, x_2), x_3)$
4. $T(x_1, x_3) \leq T(x_2, x_3)$ jika $x_1 \leq x_2$ (2. 3)

II.2.3. S-norm

S-norm (juga dikenal sebagai T-conorm) adalah operasi disjungsi yang digunakan di logika *fuzzy*. Pada laporan ini, s-norm dituliskan dengan simbol ⊥. Berdasarkan t-norm, S-norm dapat didefinisikan sebagai ⊥(a,b) = 1 – T(1-a , 1-b)). Berdasarkan definisi, s-norm adalah sebuah fungsi ⊥: [0,1] x [0,1] → [1,0] yang memenuhi sifat – sifat berikut:

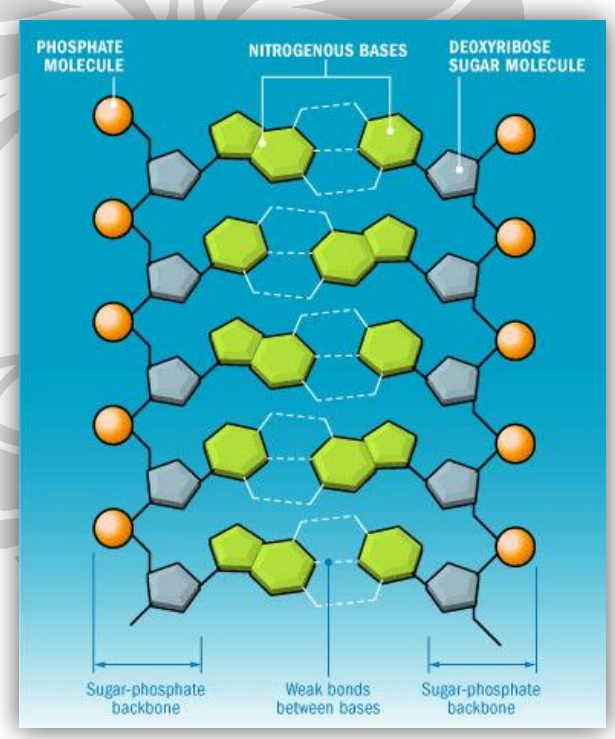
1. $\perp (x, 0) = x$, $\perp(x, 1) = 1$
2. $\perp(x_1, x_2) = \perp (x_2, x_1)$
3. $\perp(x_1, \perp (x_2, x_3)) = \perp (\perp (x_1, x_2), x_3)$
4. $\perp(x_1, x_3) \leq \perp (x_2, x_3)$ jika $x_1 \leq x_2$ (2. 4)

II.3. Materi Genetis

Materi genetis makhluk hidup dikenal dengan asam nukleat (*nucleic acid*). Ada dua tipe asam nukleat yaitu asam deoksiribonukleat (DNA) dan asam ribonukleat (RNA) [SAD00]. Setiap materi genetis tersebut tersusun dalam struktur tertentu dan memiliki fungsi tertentu.

II.3.1. DNA dan RNA

DNA adalah material genetis yang dimiliki oleh semua organisme ber-sel tunggal, banyak sel dan beberapa tipe virus yang diturunkan oleh induknya. Akan tetapi, beberapa virus lainnya memiliki RNA sebagai material genetisnya. DNA dan RNA keduanya merupakan materi genetis yang menyimpan informasi tentang sifat-sifat makhluk hidup.

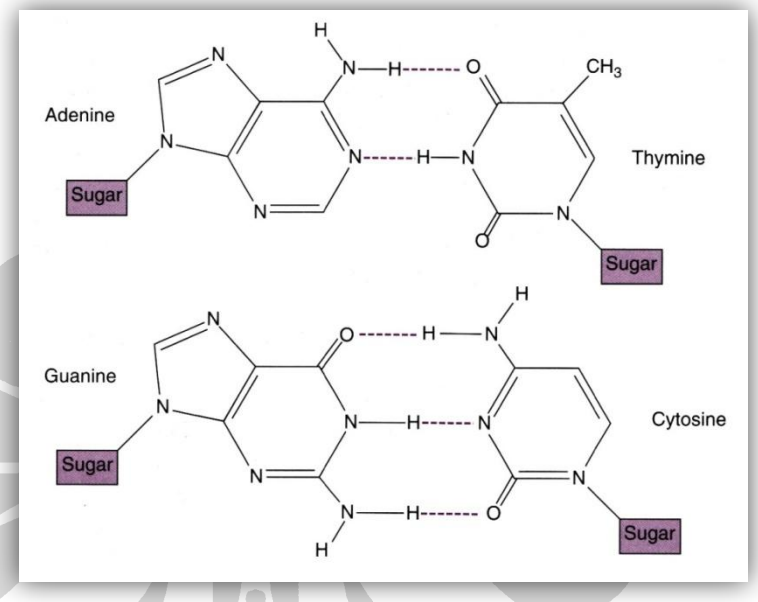


Gambar 3. Ilustrasi Pasangan DNA.

(sumber: howstuffworks.com**)**

DNA ada di setiap sel organisme. Di dalam struktur kimia yang dikandungnya terkandung informasi mengenai organisme tersebut dan juga semua aktivitas organisme. Walaupun demikian, DNA tidak secara langsung terlibat dalam

aktifitas dalam sel dan organisme. Sintesis DNA secara langsung menjadi RNA disebut messenger RNA (mRNA). mRNA kemudian berinteraksi dengan pembuat protein. Proses produksi ini mengatur hidup dan matinya sel dan organisme karena enzym *intraseluler* – sebagai spesifik protein – bertanggung jawab untuk sintesis semua zat kimia dalam sel.



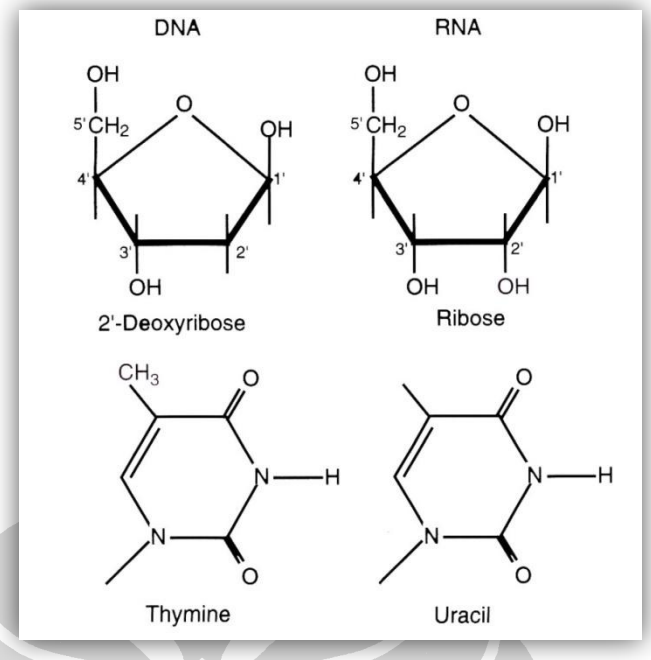
Gambar 4. Struktur Pasangan DNA [DAL03].

Nukletioda sendiri tersusun atas tiga zat penyusun yang lebih kecil, yaitu: gula, asam fosfat dan sebuah basa mengandung nitrogen. (Lihat gambar). Di dalam rantai polinukleotida DNA dan RNA, sebuah monomer nukleotida memiliki kelompok fosfat yang dibatasi oleh gula pada sambungan nukleotida berikutnya. Jadi rantai tersebut memiliki tulang punggung gula dan fosfat dengan tambahan komponen di dalamnya. Komponen tambahan tersebut terdiri dari empat basa yang mengandung nitrogen yang mungkin yang disebut:

- *Thymine* = T
- *Adenine* = A
- *Cytosine* = C
- *Guanine* = G

Di dalam DNA, sedangkan di dalam RNA thymine digantikan oleh:

- *Uracil* = U



Gambar 5. Perbedaan DNA dan RNA [DAL03].

II.3.2. Protein dan Kodon

Protein adalah sebuah *polimer*. *Polimer* adalah sebuah *makromolekul* yang besar terdiri dari banyak blok yang identik atau mirip. Blok-blok tersebut dinamakan *monomer*. *Monomer-monomer* tersebut berhubungan membentuk sebuah rantai. Beberapa *monomer* protein adalah asam amino diantaranya *alanine*, *glycine*, *serine*, dll.

Gambar 6. Kode Genetik Berdasarkan Kodon mRNA dan Beberapa Berdasarkan Kodon DNA.

(sumber: howstuffworks.com)

Asam amino terbentuk kodon yang tersusun oleh tiga nukleotida atau sering disebut *triplet* kodon. Beragam kemungkinan protein dapat terjadi dari kombinasi kodon. Urutan asam amino dalam rantai protein menentukan fungsi protein dalam metabolisme. Sebagai contoh dua protein seperti *serine-alanine-glycine* dan *serine-glycine-alanine* memiliki fungsi yang berbeda. Struktur kimia mRNA yang memproduksi protein menjadi tugas urutan asam amino tertentu.

II.4. DNA Dan RNA Sebagai Himpunan Fuzzy

Subbab ini membahas teori yang dikemukan pada *fuzzy* genomes [SAD00]. Pembahasan meliputi representasi data asam nukleat DNA dan RNA sebagai himpunan *fuzzy* yang terurut. Pembuatan representasi data ini dibuat untuk dapat

menggali informasi lebih jauh di dalam rantai DNA dan RNA. Selain itu, representasi data ini supaya memungkinkan menerapkan teori logika *fuzzy* untuk menganalisis rantai DNA dan RNA. Teori dalam bab ini merupakan sebuah cara yang ditawarkan oleh Sadegh-Zadeh [SAD00].

II.4.1. Rumusan Barisan Polinukleotida

DNA terdiri dari banyak nukleotida oleh karena itu disebut dengan polinukleotida. Sadegh-Zadeh [SAD00] menawarkan pembuatan representasai data dari polinukleotida dengan menggunakan deret himpunan *fuzzy*. Representasi data ini dibuat supaya dapat mengimplementasikan teori *fuzzy* untuk melakukan analisis dan perbandingan terhadap rantai DNA [SAD07].

Sebuah himpunan *fuzzy* terurut (*ordered fuzzy set*) adalah himpunan *fuzzy* yang disusun pada himpunan dasar (*ground set*) $\Omega = <x_1, x_2, x_3, \ldots>$. Sebagai contoh, misal sebuah himpunan <0, 1, 2, 3, 4> adalah himpunan deret lima bilangan cacah pertama. Didalamnya terdapat pula himpunan bilangan prima, tingkat keprimaan dimasukkan sebagai tambahan untuk mengukur keprimaan suatu bilangan. Himpunan deret tersebut menjadi: <(0, 0), (1, 0), (2, 1), (3,1), (4,0)>. Tingkat keprimaan bernilai 0 mengindikasikan bahwa bilangan tersebut bukan bilangan prima, sedangkan 1 mengindikasikan bahwa bilangan terebut adalah bilangan prima.

Definisi 1.

1. Apabila $<S_1, \ldots, S_n>$ adalah alfabet sebuah bahasa (*language*) dengan $n \geq 1$ tanda *(signs)* $S_1, \ldots, S_n$, adalah sebuah anggota dari tanda Sj $\in <S_1, \ldots, S_n>$ disebut juga sebagai string atau barisan pada $<S_1, \ldots, S_n>$ dengan panjang satu.
2. Apabila $\mathfrak{s}_1$ dan $\mathfrak{s}_2$ adalah barisan pada $<S_1, \ldots, S_n>$ dengan panjang p dan q, maka penggabungan rangkaian (*concatination*) keduanya $\mathfrak{s}_1\mathfrak{s}_2$ di dalam $<S_1, \ldots, S_n>$ dengan panjang $p + q$.

Sebagai contoh, pada frasa “GENE” adalah sebuah barisan dengan panjang 4 berada di dalam alfabet huruf latin <A, B, C, …, Z>, kemudian frasa

TACTGT

adalah sebuah sequence di dalam DNA alphabet dengan panjang enam di dalam alfabet DNA. TACTGT terdiri dari dua kodon, TAC untuk asam amino *tyrosine* dan TGT untuk asam amino *cysteine*.

Alfabet DNA = <T, C, A, G> dan

Alfabet RNA = <U,C,A,G>

Selanjutnya pembahasan mencakup RNA dan DNA, namun simbol yang digunakan selanjutnya adalah simbol untuk DNA saja. Perubahan terjadi pada basa *uracil* (U) dan *thymine* (T).

II.4.2. Ilustrasi Kode Fuzzy

Beberapa gambaran penggunaan *fuzzy* dalam pengkodean DNA akan dijelaskan pada bagian ini. Barisan DNA yang diambil sebagai contoh adalah barisan:

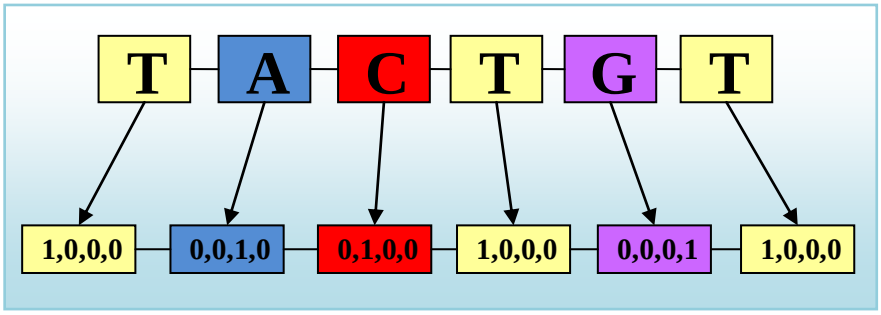
TACTGT

Barisan tersebut ditransformasi menjadi informasi dalam bentuk barisan bit yang setara (*equivalent*), sebagai contoh barisan yang hanya memiliki nilai binary 0 dan 1. Representasi yang digunakan adalah sebagai berikut.

T ∈ <T, C, A, G >	direpresentasikan sebagai	<1, 0, 0, 0>	atau	1000
C ∈ <T, C, A, G >	direpresentasikan sebagai	<0, 1, 0, 0>	atau	0100
A ∈ <T, C, A, G >	direpresentasikan sebagai	<0, 0, 1, 0>	atau	0010
G ∈ <T, C, A, G >	direpresentasikan sebagai	<0, 0, 0, 1>	atau	0001

Dengan menggunakan representasi ini maka hasil transformasi yang diperoleh dari TACTGT menjadi barisan dengan panjang 24 bit:

100000100100100000011000



Gambar 7. Ilustrasi Pengkodean Polinukleotida.

Jika ditulis dalam notasi vektor adalah:

(1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)

Vektor berdimensi 24 bit vektor ini merepresentasikan kode *fuzzy* pada TACTGT. Vektor dengan dimensi 24 bit ini adalah sebuah titik pada *hypercube* dengan dimensi 24 bit (lihat subbab II.5). Ilustrasi yang dijelaskan pada bab ini adalah sebuah pengantar untuk memahami lebih dalam dari ide pengkodean *fuzzy* untuk nukleotida.

II.4.3. Pengkodean Fuzzy Informasi Genetik

Vektor dimensi 24 dalam contoh sebelumnya hanya memiliki dua nilai himpunan yaitu {0, 1}. Contoh tersebut memiliki keterbatasan dan tidak alami. Pada kenyataannya barisan polinukleotida dapaat memiliki nilai *fuzzy*, seperti:

(0.2, 0.5, 0,3, …, 0, 0.1, 1)

Dalam bentuk ini setiap nilai vektor bernilai bilangan riil dengan interval [0, 1]. Informasi penting dan menarik dapat digali dari fakta yang ada ini. Untuk menjelaskan lebih dalam, terlebih dahulu kita kenalkan dengan notasi dari alfabet *fuzzy*.

Notasi “alfabet” di sini diartikan dalam arti kata yang seluas-luasnya. Sebuah alfabet $<S_1, ..., S_n>$ adalah segala bentuk dari tanda prototipe $S_1, ...,S_n$ yang memiliki anggota dari $s_1, ..., s_m$ dapat dioperasikan dengan operasi konkatinasi. Sebagai contoh yaitu kode morse, alfabet latin, struktur kimia, dan penggabungan suku kata pada bahasa natural.

Dalam transformasi barisan dasar $\mathfrak{s} = s_1 \ldots s_m$ menjadi himpunan *fuzzy*, sebuah himpunan dasar Ω harus diisi dengan $\mathfrak{s}$ dimana $\mathfrak{s}$ adalah sebuah himpunan *fuzzy*. Sebagai contoh, misal diberikan sebuah alfabet <S_1, ..., S_n> pada barisan $\mathfrak{s}$, himpunan dasar Ω dapat dibentuk sebagai himpunan terurut dari semua kemungkinan kombinasi dalam barisan alfabet dengan tanda $S_1, \ldots, S_n$, yaitu $< 1, \ldots, m > \times < S_1, \ldots, S_n >$ dimana m-*tuple* $< 1, \ldots, m >$ merepresentasikan nomor posisi pada barisan. Himpuan dasar ini diproduksi oleh fungsi *gmatr* dengan cara sebagai berikut:

Definisi 2. Jika $\mathfrak{s} = s_1 \ldots s_m$ adalah barisan pada $< S_1, \ldots, S_n >$, maka $gmatr(\mathfrak{s}) = ground\_matrix$ seperti:

$$ground\_matrix = \begin{Bmatrix} 1 \\ \cdot \\ \cdot \\ m \end{Bmatrix} * (S_1, \ldots, S_n) = \begin{Bmatrix} S_1 1, \ldots, S_n 1 \\ \ldots \\ \ldots \\ S_1 m, \ldots, S_n m \end{Bmatrix} \qquad (2.5)$$

Ground_matrix ini adalah sebuah *outer product* dari vektor kolom (1, …,*m*) dari barisan yang menunjukkan nomor posisi dengan vektor baris $< S_1, \ldots, S_n >$ dari alfabet. Sebuah nilai "$S_i j$" dalam sebuah matriks dibaca "Tanda S_i dari sebuah alfabet pada posisi j pada barisan". Di dalam matriks, isi setiap barisnya adalah himpunan dasar terurut $\Omega = < S_1 1, \ldots, S_i j, \ldots, S_n m >$, matriks inilah yang kita dicari. Sebagai contoh, diberikan sebuah triplet kodon:

TAC

anggota alfabet DNA <T, C, A, G>, maka barisan tersebut dimasukkan ke dalam himpunan dasar Ω pada matriks 3 x 4:

$$ground\_matrix(\text{TAC}) = \begin{Bmatrix} 1 \\ 2 \\ 3 \end{Bmatrix} * (\text{T},\ \text{C},\ \text{A}, \text{G}) = \begin{Bmatrix} \text{T di 1},\ \text{C di 1},\ \text{A di 1}, \text{G di 1} \\ \text{T di 2},\ \text{C di 2},\ \text{A di 2}, \text{G di 2} \\ \text{T di 3},\ \text{C di 3},\ \text{A di 3}, \text{G di 3} \end{Bmatrix}$$

Ground_matrix diatas berisikan himpunan dasar terurut $\Omega = < S_1 1,\ \ldots, S_i j,\ \ldots,\ S_n m >$ memungkinkan untuk membuat barisan $\mathfrak{s} = s_1 \ldots s_m$ himpunan *fuzzy* terurut dengan cara yang sederhana. Hal yang dibutuhkan adalah sebuah fungsi keanggotaan untuk barisan μ_s.

Sebuah fungsi keanggotaan μ_s memetakan ground_matrix pada nilai dengan interval [0, 1] dengan setiap $\mu_s(S_i j)$ bernilai untuk alfabet S_i yang ada pada posisi j di dalam barisan. Semua pasangan $(S_i j,\ \mu_{\mathfrak{s}}(S_i j))$ kemudian dikumpulkan dalam sebuah himpunan *fuzzy* terurut $< (S_1 1,\ \mu_{\mathfrak{s}}(S_1 1), \ldots, (S_n m,\ \mu_{\mathfrak{s}}(S_n m))$. Himpunan *fuzzy* terurut ini adalah hasil pembuatan *fuzzy* barisan $\mathfrak{s}$. Hal ini dapat dicapai dengan dua langkah dengan Definisi 3-4.

Definisi 3. $\mu_{\mathfrak{s}}$: ground_matrix→[0, 1] maka $\mu_{\mathfrak{s}}(S_i j) = \mu_{\mathfrak{s}_i}(s_j)$ untuk semua $S_i j \in$ ground_matrix.

Sebuah fungsi keanggotaan dibutuhkan untuk membentuk sebuah himpunan *fuzzy*. Sebuah fungsi keanggotaan μ_s dipetakan dari matriks dasar pada nilai dengan interval [0, 1] dengan setiap $\mu_s(S_i j)$. Berdasarkan contoh yang diberikan TAC pada alfabet DNA <T, C, A, G>, hasil yang berdasarkan definisi XX adalah sebagai berikut.

$\mu_s(S_1 1) = \mu_{S1}(S_1) = 1$ yaitu $\mu_{\text{TAC}}(\text{T di } 1) = 1$

$\mu_s(S_2 1) = \mu_{S2}(S_1) = 0$ $\mu_{\text{TAC}}(\text{C di } 1) = 0$

$\mu_s(S_3 1) = \mu_{S3}(S_1) = 0$ $\mu_{\text{TAC}}(\text{A di } 1) = 0$

$\mu_s(S_4 1) = \mu_{S4}(S_1) = 0$ $\mu_{\text{TAC}}(\text{G di } 1) = 0$

$\mu_s(S_1 2) = \mu_{S1}(S_2) = 0$ $\mu_{\text{TAC}}(\text{T di } 2) = 0$

$\mu_s(S_2 2) = \mu_{S2}(S_2) = 0$ $\mu_{\text{TAC}}(\text{C di } 2) = 0$

$\mu_s(S_3 2) = \mu_{S3}(S_2) = 1$ $\mu_{\text{TAC}}(\text{A di } 2) = 1$

$\mu_s(S_4 2) = \mu_{S4}(S_2) = 0$ $\mu_{\text{TAC}}(\text{G di } 2) = 0$

$\mu_s(S_1 3) = \mu_{S1}(S_3) = 0$ $\mu_{\text{TAC}}(\text{T di } 3) = 0$

$\mu_s(S_2 3) = \mu_{S2}(S_3) = 1$ $\mu_{\text{TAC}}(\text{C di } 3) = 1$

$\mu_s(S_3 3) = \mu_{S3}(S_3) = 0$ $\mu_{\text{TAC}}(\text{A di } 3) = 0$

$\mu_s(S_4 3) = \mu_{S4}(S_3) = 0$ yaitu $\mu_{\text{TAC}}(\text{G di } 3) = 0$

Fungsi keanggotaan global ditentukan oleh nilai keanggotaan lokal $\mu_{s_1}(s_1), \dots, \mu_{s_n}(s_m)$ dengan tanda $s_1, \dots, s_m$. Hasil perhitungan dari cara yang digunakan dari ground_matrix adalah sebuah matriks dengan nilai:

$$\begin{Bmatrix} \mu_s(S_1 1), \dots, \mu_s(S_1 1) \\ \dots \\ \dots \\ \mu_s(S_1 m), \dots, \mu_s(S_1 m) \end{Bmatrix}$$

Setiap nilai $\mu_s(S_i j)$ di dalam matriks ditentukan oleh tingkat kemiripan berdasarkan Definisi 3. Berikutnya fungsi baru, *fset*, mengkombinasikan semua komponen $S_i j$ pada matriks dasar dengan tingkat keanggotaan $\mu_s(S_i j)$ dan mengembalikan pasangan $(S_i j, \mu_s(S_i j))$. matriks *fuzzy* yang dibentuk yaitu:

Definisi 4. $fset(ground\_matrix) = fuzzy\_matrix$ seperti

$$fuzzy\_matrix = \begin{Bmatrix} (S_1 1, \mu_s(S_1 1)), \dots, (S_n 1, \mu_s(S_n 1)) \\ \dots \\ \dots \\ (S_1 m, \mu_s(S_1 m)), \dots, (S_n m, \mu_s(S_n m)) \end{Bmatrix}$$

Di dalam *fuzzy*_matrix terdapat $m \times n$ elemen himpunan *fuzzy* terurut di dalam seluruh barisnya:

$$< (S_1 1, \mu_s(S_1 1)), \dots, (S_i j, \mu_s(S_i j)), \dots, (S_n m, \mu_s(S_n m)) >$$

Himpunan *fuzzy* terurut ini merepresentasikan asal barisan s pada alfabet $< S_1, \dots, S_n >$. Berikutnya transformasi barisan menjadi sebuah himpunan *fuzzy* terurut dalam dua langkah:

$$fset(gmatr(s)) = < (S_1 1, \mu_s(S_1 1)), \dots, (S_i j, \mu_s(S_i j)), \dots, (S_n m, \mu_s(S_n m)) > \quad \textbf{(2. 6)}$$

Himpunan *fuzzy* ini menggambarkan bagaimana urutan sebuah tanda alfabet muncul dalam barisan dasar. Sebagai contoh triplet UAC pada alfabet DNA <U, C, A, G>.

$$fset(gmatr(\text{TAC})) = < (\text{T di } 1,\text{T}),\ (\text{C di } 1,0),\ (\text{A di } 1,0),\ (\text{G di } 1,0)$$
$$(\text{T di } 2,0),\ (\text{C di } 2,0),\ (\text{A di } 2,1),\ (\text{G di } 2,0)$$
$$(\text{T di } 3,0),\ (\text{C di } 3,1),\ (\text{A di } 3,0),\ (\text{G di } 3,0) >$$

Di dalam sebuah himpunan *fuzzy* terurut $< (x_1, a_1), \dots, (x_m, a_m) >$, sebuah m-*tuple* $(a_1, \dots, a_m)$ dari tingkat keanggotaannya dapat dijadikan sebagai vektor *fuzzy*. Vektor *fuzzy* dari *fuzzy*_matrix didapat dengan menggunakan fungsi *fvector* dengan jalan sebagai berikut:

Definisi 5. $fvector(fuzzy\_matrix) = < \mu_{\mathfrak{s}}(s_1 1), \dots, \mu_{\mathfrak{s}}(s_n m) >$ **(2. 7)**

Jumlah elemen pada *fuzzy*_matrix adalah $m \times n$ elemen, maka vektor yang dimiliki adalah vektor berdimensi $m \times n$, dalam bentuk:

$$(r_1, \dots, r_{m \times n})$$

Dengan $r_1 \epsilon [0,1]$dimana m adalah panjang dari barisan dasar $\mathfrak{s} = s_1 \dots s_m$ dan *n* adalah panjang dari alfabet $< S_1, \dots, S_n >$. Hasil yang di dapat adalah vektor *fuzzy* dari himpunan *fuzzy* terurut (2. 7) diatas merepresentasikan sumber dari barisan. Dengan kata lain, kode *fuzzy* untuk barisan polinukleotida $\mathfrak{s}$:

Definisi 6. Jika $\mathfrak{s}$ adalah sebuah barisan, maka $fcode(\mathfrak{s}) = fvector(fset(gmatr(\mathfrak{s})))$. **(2. 8)**

Sebuah *single-strand* polinukleotida dapat memiliki sebuah kode *fuzzy* dengan cara ini. Berikut contoh untuk TACTGT adalah:

$$fcode(\text{TACTGT}) = fvector\left(fset(gmatr(\text{TACTGT}))\right)$$

= (1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)

Seperti yang telah disebutkan sebelumnya, kode *fuzzy* pada rantai basa polinukleotida dapat berisi nilai bilangan riil antara 0 dan 1. Pada kenyataannya ketika terdapat keruksakan pada rantai, maka nilai mungkin dapat memiliki beberapa kjelas *fuzzy*, atau ketika basa tidak dapat diidentifikasi dengan pasti atau belum teridentifikasi. Dalam kasus seperti disebutkan, kemungkinan-kemungkinan dapat menentukan tingkat keanggotaan untuk diisi pada kode *fuzzy*.

II.5. Geometri Polinukleotida

Polinukleotida dapat direpresentasikan sebagai sebuah titik pada sebuah unit *hypercube* melalui kode *fuzzy* yang dimilikinya. Unit *hypercube* dapat memiliki sebuah titik polinukleotida dengan menggunakan ruang vektor *fuzzy* polinukleotida. Melalui geometri polinukleotida ini memungkinkan untuk melakukan pendekatan pada taksonomi dan diagnosis. Pada bagian ini akan dijelaskan mengenai geometri yang akan digunakan. Selain itu juga, penggunaan representasi data dengan menggunakan *hypercube*.

II.6.1. Ruang Vektor Fuzzy Polinukleotida

Diberikan sebuah himpunan terhingga (*finite*) $\Omega = \{x_1, \ldots, x_n\}$ dengan $n \geq 1$ anggota, *power set*-nya adalah $F(2^{\Omega})$ membentuk sebuah unit *hypercube* berdimensi *n*. Oleh karena itu, setiap anggota $F(2^{\Omega})$, sebuah himpunan *fuzzy*, adalah sebuah titik dalam kubus. Beberapa penjelasan dasar akan dijelaskan pada subbab ini.

Sebuah unit dengan interval [0, 1] adalah sebuah garis dengan panjang 1. Sebuah sistem koordinat terdiri dari dua *axis x* dan *y*. Kedua axis tersebut berada dalam interval [0, 1] sebuah unit segi empat. Penulisan keduanya dengan cara [0, 1] × [0, 1] atau lebih pendeknya $[0, 1]^2$. Sebuah sistem koordinat dengan tiga axis x, y dan z dan semuanya pada interval [0, 1] maka akan ada di dalam sebuah kubus. Penulisan ketiganya dengan cara [0, 1] × [0, 1] × [0, 1] atau lebih singkatnya $[0, 1]^3$. Secara umum, sistem koordinat yang terdiri dari n koordinat axis $x_1, \ldots, x_n$ dengan semua dengan nilai dalam interval [0, 1] adalah sebuah kubus berdimensi *n*. Koordinat ini disebut juga dengan *hypercube* dan ditulis [0, 1] × … × [0, 1] atau singkatnya $[0, 1]^n$. Sebuah kubus berdimensi n adalah:

- Sebuah unit garis antara 0 dan 1 untuk n = 1
- Sebuah unit segi empat untuk n = 2
- Sebuah kubus biasa untuk n = 3
- Sebuah *hypercube* $[0, 1]^n$ untuk $n \geq 1$

Setiap *hypercube* $[0, 1]^n$ memiliki sebanyak 2^n sudut. Sebuah himpunan *fuzzy* adalah sebuah titik pada sebuah unit *hypercube* berdimensi n. Beberapa gambaran akan diberikan pada bahasan berikut.

Jika $\{(x_1, a_1), \dots, (x_n, a_n)\}$ adalah himpunan *fuzzy* dengan $n \geq 1$ anggota dan n-tuple berurut $(a_1, \dots, a_n)$ berisi tingkat keanggotaan dari vektor *fuzzy* dimana a_i adalah derajat keanggotaan objek x_i pada himpunan tersebut, untuk $i \geq 1$.

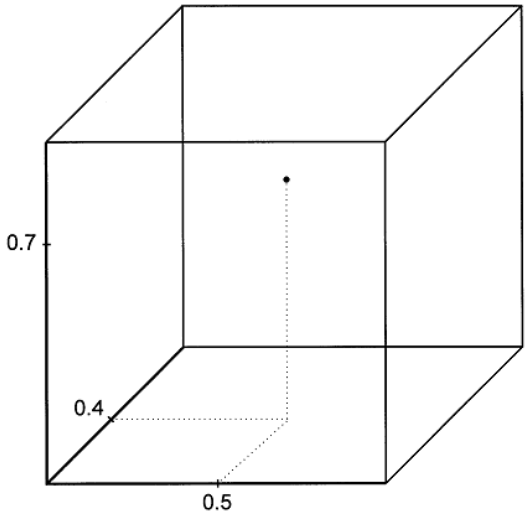
Misalkan $\Omega = \{x_1, \dots, x_n\}$ adalah sebuah himpunan dasar. Penulisan berikutnya untuk Ω dalam urutan yang sama sebanyak n kolom $x_1, x_2, \dots, x_n$.

Apabila himpunan dasar $\Omega = \{x_1, x_2, x_3\}$, maka ditulis:

(a_1, a_2, a_3) untuk himpunan $fuzzy$ $\{(x_1, a_1), (x_2, a_2), (x_3, a_3)\}$

Seperti:

- $(1, 1, 1)$ untuk himpunan $fuzzy$ $\{(x_1, 1), (x_2, 1), (x_3, 1)\}$
- $(0.2, 0.8, 0.6)$ untuk himpunan $fuzzy$ $\{(x_1, 0.2), (x_2, 0.8), (x_3, 0.6)\}$
- $(1, 0, 1)$ untuk himpunan $fuzzy$ $\{(x_1, 1), (x_2, 0), (x_3, 1)\}$



Gambar 8. Interpretasi Titik Himpunan Fuzzy Berdimensi Tiga
$\boldsymbol{A = \{(x_1, 0.5), (x_2, 0.7), (x_3, 0.4)\}}$.

Komponen ke-i pada a_i pada kolom $i \geq 1$ seperti pada himpunan vektor *fuzzy* $(a_1, \dots, a_n)$ merepresentasikan derajat keanggotaan $\mu_A(S_i) = a_i$ berkorenpondensi

dengan objek x_i. Tiga contoh himpunan diatas adalah vektor berdimensi tiga. Sebuah fungsi μ_A mendefinisikan sebuah himpunan *fuzzy* A sebagai vektor berdimensi n A = $(\mu_A(x_i), \dots, \mu_A(x_n)) = (a_1, \dots, a_n)$ dengan $a_i \epsilon [0,1]$. Pada penerapannya geometri vektor berdimensi n $(a_1, \dots, a_n)$dengan nilai riil [0, 1] mendefinisikan:

- Sebuah titik pada sebuah garis jika n = 1
- Sebuah titik pada sebuah persegi empat jika n = 2
- Sebuah titik pada sebuah kubus jika n = 3
- Sebuah titik pada sebuah *hypercube* jika n ≥ 1

II.6.2. Kode Genetik Dimensi 12

Representasi data secara nyata terbatas hanya hingga pada dimensi tiga $[0, \infty]^3$. Variable waktu menjadi dimensi keempat yang bisa diajukan oleh Einstein. Kode genetik sendiri dapat dilihat sebagai vektor berdimensi 12 karena *triplet* kodon XYZ memiliki $3 \times 4 = 12$ dimensi kode *fuzzy* $(a_1, \dots, a_{12})$.

Semua *triplet* kodon dapat direpresentasikan sebagai vektor berdimensi 12. Dalam vektor berdimensi 12 ini informasi genetik dapat disimpan karena setiap triplet kodon menyimpan informasi tersendiri.

II.6. Pengukuran Kemiripan dan Perbedaan Antar Ruang Vektor Fuzzy Polinukleotida

Terdapat dua pendekatan dalam pengukuran kemiripan berdasarkan ruang vektor polinukleotida [LIM07]. Pendekatan tersebut yaitu pendekatan yang ditawarkan oleh Sadegh-Zadeh [SAD00] dan Tores et al [NIE06]. Masing-masing memiliki rumus tersendiri dalam menghitung kemiripannya. Cara yang ditawarkan Tores et al merupakan pengembangan dari teori *fuzzy* genom yang ditwarkan Sadegh-Zadeh.

II.6.1. Pengukuran Sadegh-Zadeh

Perbedaan antara dua rantai DNA diusulkan oleh Sadegh-Zadeh [SAD00] dibangun dari bagian jarak geometris antara dua titik dalam *fuzzy* polinukleotida *hypercube* $[0, 1]^n$. Sebagai contoh diberikan dua rantai seperti berikut TACTGT

dan CACTGT, setiap dari huruf tersebut menjadi elemen dalam 24 dimensi. Sebagai contoh lain:

ς_1 = TACTGT	kode untuk	*Tyrosine/cystein*
ς_2 = CACTGT		*Histidine/cystein*
ς_3 = CTCTGT		*Leucine/cystein*

Rantai ς_2 di dalam kubus berlokasi lebih dekat dengan lokasi rantai ς_1 dibandingkan dengan rantai ς_3 karena antara ς_1 dan ς_2 hanya memiliki satu perbedaan dalam sedangkan ς_1 dan ς_3 memiliki dua perbedaan huruf. Pada akhirnya ruang rantai DNA ini dijadikan ruang pengukuran (*metric space*).

Sebuah *metric space* didefinisikan sebagai pasangan <*X*, d>, X adalah himpunan tidak kosong dan d adalah fungsi biner dari $X \times X$ pada bilangan riil seperti untuk semua $x, y, z \in X$ pada:

$d(x, y) \geq 0$	non-negatif	**(2. 9)**
d(x, y) = 0, jika dan hanya jika x = y	Aturan identitas	**(2. 10)**
d(x, y) = d(y, x)	Simetris	**(2. 11)**
$d(x, y) + d(y, z) \geq d(x, z)$	Aturan segitiga.	**(2. 12)**

Fungsi *d* disebut sebagai fungsi jarak atau pengukuran dalam *X*. Sebagai contoh, misal *X* adalah himpunan semua kode *fuzzy* n-dimensi rantai DNA dengan panjang $n \geq 1$. Diberikan dua kode $(a_1, \ldots, a_n) = x$ dan $(b_1, \ldots, b_n) = y$ sebagai sebuah titik pada ruang polynucletiode $[0, 1]^n$, setiap element ℓ^p pada perhitungan Minkowski:

$$\ell^p(x, y) = \left(\sum_i |a_i - b_i|^p\right)^{1/p} \text{ untuk } 1 \leq i \leq n \text{ dan } p \geq 1 \qquad \textbf{(2. 13)}$$

Ketika p =1, perhitungan ini disebut juga dengan Jarak Hamming, rumusnya menjadi:

$$\ell^1(x, y) = \sum_i |a_i - b_i|^p \text{ untuk } 1 \leq i \leq n, \qquad \textbf{(2. 14)}$$

dan jarak euclidian ketika p = 2:

$$\ell^2(x, y) = (\sum_i |a_i - b_i|^2)^{1/2} \text{ untuk } 1 \leq i \leq n \quad \textbf{(2. 15)}$$

Sebuah contoh sederhana, jarak hamming diantara 2 vektor berdimensi tiga x = (0.9, 0.2, 0.4) dan y = (0.3, 0.5, 1) adalah

$$\ell^1(x, y) = |0.9 - 0.3| + |0.2 - 0.5| + |0.4 + 1| = 1.5$$

sedangkan jarak euclidian-nya adalah

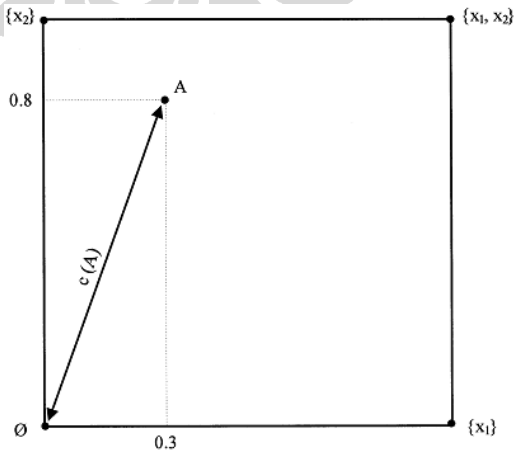
$$\ell^2(x, y) = (|0.9 - 0.3|^2 + |0.2 - 0.5|^2 + |0.4 - 1|^2)^{1/2} = (0.81)^{1/2} = 0.9$$

Ruang vektor (metrics) n-dimensional yang menggunakan ℓ^1, dinotasikan dengan $<[0, 1]^n, \ell^1>$. Berikutnya notasi menghitung nilai *fuzzy* pada himpunan A, dinotasikan denga c(A). Jika A = $(\mu_A(x_1), \ldots, \mu_A(x_n))$ adalah himpunan *fuzzy* yang direpresentasikan dalam notasi vektor, maka jumlah atau nilai, c(A), adalah jumlah semua nilai anggota vektor tersebut.

Definisi 7. $c(\text{A}) = \sum_i \mu_A(x_i)$ untuk $1 \leq i \leq n$. **(2. 16)**

Dari definisi yang telah dibuat, dalam hipercube unit, jumlah himpunan A adalah jarak hamming dengan jarak himpunan kosong $\emptyset$ pada titik origin kubus, yaitu $\ell^1(\text{A}, \emptyset)$.

Teorema 1. $c(\text{A}) = \ell^1(\text{A}, \emptyset)$.



Gambar 9. Ilustrasi Pengukuran Jarak Dalam Ruang Dua Dimensi.

Simbol yang ditawarkan Zadeh untuk menghitung perbedaan yaitu *differ*(*A*, *B*) = *r* dan dibaca "tingkat perbedaan antara himpunan A dan himpunan B adalah r". Definisi yang digunakan adalah sebagai berikut:

Definisi 7. $differ(A, B) = \frac{\sum_i \max\left(0,\mu_A(x_i)-\mu_B(x_i)\right)+\sum_i \max\left(0,\mu_B(x_i)-\mu_B(x_i)\right)}{c(A\cup B)}$ **(2. 17)**

Secara sederhana, menghitung selisih dari dua nilai vektor:

$$\max\big(0, \mu_A(x_i) - \mu_B(x_i)\big) + \max\big(0, \mu_B(x_i) - \mu_B(x_i)\big)$$

Dilakukan untuk semua anggota himpunan:

$$\sum_i \max\big(0, \mu_A(x_i) - \mu_B(x_i)\big) + \sum_i \max\big(0, \mu_B(x_i) - \mu_B(x_i)\big)$$

Kedua himpunan A dan B di-normalisasi dengan jumlah mereka $c(A \cup B)$:

$$\frac{\sum_i \max\big(0,\ \mu_A(x_i) - \mu_B(x_i)\big) + \sum_i \max\big(0,\ \mu_B(x_i) - \mu_B(x_i)\big)}{c(A \cup B)}$$

Rumus inilah yang dijadikan dasar pengukuran yang ditawarkan Sadegh-Zadeh. Berikutnya didefinisikan

Definisi 8. $similar(A, B) = 1 - differ(A, B)$ **(2. 18)**

Dengan beberapa perhitungan rumus similar sama saja dengan:

$$similar(A, B) = \frac{c(A \cap B)}{c(A \cup B)} \quad \textbf{(2. 19)}$$

Rumus (2. 18) diatas yang berikutnya digunakan untuk melakukan penelitian ini.

II.6.1. Pengukuran Tores

Nieto Tores et al [TOR02, NIE06] mengembangkan pengukuran yang ditawarkan oleh Sadegh-Zadeh [SAD00]. Pengembangan tersebut yaitu dengan menerapkan metode statistik pada rantai yang panjang. Rantai yang panjang dikonversi menjadi hanya vektor berdimensi 12. Konversi dilakukan dengan cara mengambil rata-rata semua kodon.

Pengembangan lainnya juga yaitu pembarian beberapa alternatif rumus pengukuran antara dua vektor. Berdasarkan unit *hypercube* berdimensi 12, I^{12}. Apabila $p = (p_1, \dots, p_{12})$ dan $q = (q_1, \dots, q_{12}) \in I^{12}$ adalah dua buah *fuzzy* polinukleotida yang berbeda, maka jarak element antara *p* dan *q* adalah

$$d(p,q) = \frac{\sum_{i=1}^{12} |p_i - q_i|}{\sum_{i=1}^{12} \max(p_i, q_i)} \quad (2.20)$$

Apabila $p = q = \emptyset = (0, \dots, 0)$, maka $d(\emptyset, \emptyset) = 0$

$$d_1(p,q) = \frac{d(p,q)}{1+d(p,q)} \quad (2.21)$$

Untuk setiap ruang vektor (X, ρ) formula

$$\rho_1(x,y) = \frac{\rho(x,y)}{1+\rho(x,y)} \quad (2.22)$$

Didefinisikan pengukuran terhadap himpunan *X*

$$d_2(p,q) = \frac{\sqrt{\sum_{i=1}^{12}(p_i - q_i)}}{\sqrt{12}} \quad (2.23)$$

$$d_3(p,q) = \frac{d(p,q)}{1+d(p,q)} \quad (2.24)$$

$$d_4(p,q) = \frac{\sum_{i=1}^{12} |p_i - q_i|}{12} \quad (2.25)$$

$$d_5(p,q) = \frac{\sum_{i=1}^{12} |p_i - q_i|}{1+\sum_{i=1}^{12} |p_i - q_i|} \quad (2.26)$$

Dengan menggunakan definisi yang telah dibuat diatas, beberapa perhitungan dilakukan pada beberapa nukleotida untuk melihat sifat dari pengukuran jarak tersebut.

Tabel 1. Hasil Perhitungan Beberapa Metode Distance yang Diusulkan Tores et al [NIE06].

X	Y	d(X, Y)	d1(X, Y)	d2(X, Y)	d3(X, Y)	d4(X, Y)	d5(X, Y)
CAT	CCG	0.8	0.444444	0.57735	0.366025	0.333333	0.8
CAT	TCG	1	0.5	0.707107	0.414214	0.5	0.857142
CGT	CAG	0.8	0.444444	0.57735	0.366025	0.333333	0.8

CAT	CGT	0.5	0.333333	0.408248	0.289898	0.166667	0.66666
AAA	GGG	1	0.5	0.707107	0.414214	0.5	0.857142

Berbeda metode perhitungan ternyata menghasilkan nilai yang berbeda pula. Metode yang dianalisis dalam penelitian ini adalah metode dengan rumus (**2.20**).

II.7. Representasi data DNA

DNA ketika di simpan dalam komputer sebagai data memiliki format tertentu yang telah terstandarisasi. Format tersebut digunakan untuk mempermudah pembuatan representasi struktur data di dalam komputer.

II.7.1. Kode Asam Nukleat Berdasarkan IUPAC

Berdasarkan aturan dari *International Union of Pure and Applied Chemistry* (IUPAC) terdapat beberapa konvensi dalam penggunaan huruf yang digunakan untuk merepresentasikan data DNA.

Tabel 2. Huruf Untuk Merepresentasikan DNA berdasarkan IUPAC.

A = adenine	*M = A* atau *C (amino)*
C = cytosine	*S = G* atau *C*
G = guanine	*W = A* atau *T*
T = thymine	*B = G, T,* atau *C*
U = uracil	*D = G, A,* atau *T*
R = G atau *A (purine)*	*H = A, C,* atau *T*
Y = T atau *C (pyrimidine)*	*V = G, C* atau *A*
K = G atau *T (keto)*	*N = A, G, C* atau *T*

II.7.2. Format Rantai DNA

Beberapa standarisasi format representasi DNA yang ada diantaranya (Genomatix 2009):

- *Plain Sequence Format*. Format ini hanya terdiri dari huruf IUPAC dan spasi, tanpa angka. Dalam satu *file* hanya terdapat satu rantai DNA saja. Contoh dari *plain sequence format* yaitu:

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
```

```
TTTAATTACAGACCTGAA
```

- Format EMBL. Dalam sebuah *file* dengan menggunakan format EMBL dapat menyimpan beberapa rantai berbeda. Sebuah rantai dimulai dengan baris yang berisi tulisan "ID", diikuti dengan baris penjelasannya. Awal dari rantai DNA-nya sendiri diawali dengan sebuah baris berisi "SQ" dan diakhiri dengan dua buah garis miring "//"

Contoh data DNA dalam Format EMBL:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
240
     gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
360
     gacctgaa
368
//
```

- Format FASTA. Format ini dikenal juga sebagai format Pearson. Dalam satu *file* dapat memuat banyak rantai DNA. Setiap rantai DNA diawali dengan satu baris deskripsi yang harus diawali dengan tanda lebih besar ">".

Contoh data DNA dalam format FASTA:

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin
like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

- Format GCG. Setiap *file* dalam format GCG memiliki tepat satu rantai DNA. *File* diawali dengan barisan penjelasan dan awal dari rantai ditandai dengan barisan penjelasan yang diakhiri dengan karakter dua titik "..". Pada barisan ini juga terdapat informasi panjang dan *checksum*.

Contoh dari format GCG:

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
AB000263  Length: 368  Check: 4514  ..
       1  acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
      61  ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
     121  caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
     181  aggccagtgc cgggccccтc ataggagagg aagctcggga ggtggccagg cggcaggaag
     241  gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
     301  agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
     361  gacctgaa
```

- Format Genbank. Format genbank dapat menyimpan beberapa rantai dalam satu *file*. Sebuah rantai dalam format GenBank diawali dengan baris mengandung kata LOCUS dan jumlah baris penjelasan. Awal dari rantai ditandai dengan baris yang dimulai dengan kata "ORIGIN" dan diakhiri dengan dua garis miring "//".

Contoh dari format GenBank:

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggccccтc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcaccccc ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

- Format IG. Sebuah *file* berformat IG dapat menyimpan beberapa rantai DNA. Setiap rantai terdiri beberapa baris komentar yang harus diawali dengan titik koma ";", sebuah baris menunjukkan nama dari rantai (tidak

boleh mengandung spasi) dan rantainya itu sendiri dengan tanda berhenti "1" untuk rantai yang linier atau "2" untuk rantai yang melingkar.

Berikut contoh data DNA dengan format IG:

```
; comment
; comment
AB000263
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA1
```

Dalam penelitian tugas akhir ini format data DNA yang digunakan adalah format FASTA. Hal ini disebabkan format inilah yang digunakan oleh NCBI sebagai sumber data pengambilan virus.