IDENTIFIKASI KLASTER TIPE KANKER BERDASARKAN ALGORITMA ITERASI KERNEL GAUSSIAN

Wasito
Jurusan Teknik Elektro, PS Teknik, Universitas Jenderal Soedirman, Purwokerto
E-mail: ito@gmx.co.uk

ABSTRAKSI

Diagnosa yang cepat dan tepat sangat memegang peranan penting dalam mengatasi masalah penyakit kanker. Melalui penggunaan teknologi microarray DNA, diharapkan diagnosa dan prediksi penyakit kanker akan dapat dilakukan secara lebih cepat dan akurat. Hanya saja, masalah utama dalam penggunaan database yang dihasilkan melalui microarray DNA ini adalah besarnya dimensi database yan dianalisis. Dalam laporan ini, akan diperkenalkan metode pengelompokan alternatif berbasis densitas fungsi kernel Gaussian dengan teknik K-nearest neighbour yang kami namakan sebagai algoritma Iterative Gaussian Local Klastering (ILGC).

Eksperimen untuk implementasi algoritma ILGC dilakukan menggunakan database yang diperoleh dari eksperimen studi tentang kanker ***Diffuse large B-cell lymphoma (DLBCL)****. Hasil eksperimen menunjukan bahwa banyaknya klaster database identik dengan hasil eksperimen masing-masing oleh Alizadeh et. al. (2000) dan Lossos te. al (2000). Secara umum hasil eksperimen menunjukan bahwa gene klaster yang ditemukan memiliki korelasi yang signifikan dengan parameter klinis dari penyakit kanker.*

Kata kunci: *penyakit kanker, teknologi microarray DNA, Algoritma Iterasi Kernel Gaussian*

1. PENDAHULUAN

Penyakit kanker sudah lama dikenal dimulai sejak ratusan tahun yang lalu. Pada umumnya tipe penyakit kanker ini sangat beragam dan sangat minim sifat-sifat biologinya yang sudah dapat diidentifikasi. Pada akhirnya hingga saat ini penyakit kanker ini sangat sulit dalam usaha penyembuhannya bahkan pencegahannya. Dalam rangka mempercepat proses identifikasi penyakit ini, saat ini para peneliti dalam bidang biologi molekuler melakukan terobosan dalam rangka mempelajari sifat-sifat biologi berbagai macam tipe kanker melalui penggunaan teknologi microarray DNA.

Pengembangan teknologi microrarray DNA ini sangat bermanfaat dalam menghasilkan sejumlah besar data ekspresi gen. Profiling ekspresi gen melalui microarray tersebut adalah teknik yang sangat efektif dalam usaha menghimpun ribuan tingkatan ekspresi gen secara simultan. Eksperimen yang di dasarkan ekspresi gen ini dapat di lakukan dengan dua cara: (1) Pengamatan setiap gen dari beberapa kondisi yang berbeda atau (2) Mengevaluasi setiap gen dari satu kondisi tetapi dalam tipe jaringan yang berbeda, khususnya jaringan sel yang mengandung kanker (Furey, et. al. 2000).

Selain untuk menginvestigasi sifat biologi dari sel yang telah di ketahui, profiling ekspresi gen ini juga bermanfaat untuk mengeksplorasi sifat-sifat biologi yang belum diketahui yang berkaitan dengan fungsi suatu gen (Brown dan Botsein, 1999). Sebagai contoh adalah sifat sifat yang belum banyak diketahui terkait dengan jaringan yang mengandung kanker *Colorectal carcinoma (*Notterman et. al, 2001*)* di Indonesia dikenal sebagai kanker usus yang merupakan salah satu dari tipe kanker pada manusia yang paling prevalen dan sudah sangat dikenal. Jenis tumor penyebab kanker ini belum dapat diindentifikasi sifat-sifat biologinya sehingga sampai kini tumor ini menjadi penyebab utama kematian pada manusia (Muro et. al. 2003).

Hasil dari penggunaan teknologi microarray DNA ini adalah tersedianya database hasil penelitian laboratorium biologi molekuler dalam jumlah yang sangat besar. Pertumbuhan volume database hasil penelitian bidang biologi molekuler yang cukup cepat ini mendorong perlu adanya metode ekstraksi informasi yang efektif dan efisien secara komputasi.

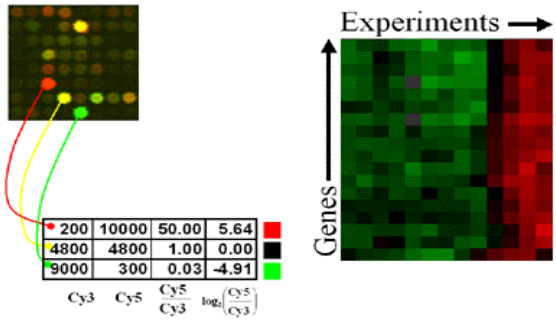
2. DATA MICROARRAY DNA

Suatu array DNA dapat dikategorikan sebagai microarray jika ukuran diameter dari *spot* DNA kurang dari 250 microns. Microarray DNA ini juga biasa disebut sebagai *chips-DNA.*

Chips-DNA terdiri dari ribuan deret DNA yang tercetak dalam array densitas tinggi pada sebuah gelas mikroskop menggunakan suatu tool yang disebut sebagai *robotic arrayer.* Chips-DNA ini selanjutnya di bagi menjadi dua kelompok samples DNA atau RNA yang akan dianalisa melalui monitoring perbedaan efek hibridasisasi (differential hybridization) dari dua kelompok sampel tersebut yang selanjutnya membentuk deret (sequences) pada array. Khusus untuk sampel dalam bentuk mRNA, maka kedua kelompok sampel tersebut perlu di konversi balik dalam bentuk cDNA. Setelah proses *hybridization*, kemudian dilakukan pembuatan image dari slides array DNA dengan menggunakan scanner untuk memperoleh pengukuran intensitas fluorescence. Ratio antara tingkat intensitas dari dua kelompok sampel tersebut di atas di sebut sebagai ekspresi gen yang dapat ditulis sebagai berikut:

 ISSN: 1907-5022

$$Ekspresi\ gen = \log_2 \frac{Int(Cy5)}{Int(Cy3)}$$



(a) (b)

Gambar 1. Proses pembentukan database ekspresi gen dari microarray DNA (chips-DNA).

3. APLIKASI PEMBELAJARAN MESIN UNTUK IDENTIFIKASI TIPE KANKER MENGGUNAKAN DATA MICROARRAY DNA

Seperti telah diuraikan di atas bahwa tantangan utama dalam database microarray DNA ini adalah volume database yang sangat besar, sehingga diperlukan teknik ekstraksi informasi yang efektif dan efisien dengan tingkat akurasi yang tinggi.

Pada makalah ini akan diperkenalkan penggunaan algoritma Iterative Gaussian Local Klastering (ILGC) (Wasito, 2005, Wasito dan Mirkin, 2006) pada proses identifikasi karakteristik tipe kanker melalui profiling ekpresi gen dari data microarray. Metode ini adalah bagian dari cabang disiplin ilmu Komputer yang di kenal sebagai Pembelajaran Mesin (Machine Learning) sebagai pendekatan alternatif dalam memecahkan masalah dalam bidang medis yaitu melalui pendekatan metodologi bidang informatika bukan dengan menggunakan pendekatan konvensional yang berbasis kegiatan laboratorium biokimia/biomolekuler yang tidak saja mahal dalam segi biaya tetapi juga membutuhkan waktu yang lama untuk memperoleh hasil yang memuaskan.

3.1 Algoritma Klaster Berbasis Pendugaan Densitas

3.1.1 Pendugaan Densitas dengan Teknik *K-Nearest Neighbour (K-NN)*

Pengelompokan objek berdasarkan pendugaan densitas berbasis K-*Nearest Neighbour* dimplementasikan oleh Tran et. al. (Tran et. al, 2003). Metode ini mudah implementasikan selain sangat jelas dalam interpretasi hasil komputasinya. Metode ini dilakukan melalui prosedur sebagai berikut (Theoridis dan Koutroumbas, 2003):

a. Tentukan K-NN dari profile gen X dan lambangkan dengan $K_{nn}(X)=\{X_j\}$, j=[1..k].
b. Jika volume data k-NN dari X dinyatakan sebagai $V_k(X)$, maka densitas dari klaster ke-j pada X dinyatakan sebagai:

$$f(X,C_i) = \frac{size(K_{nn}(X) \cap C_i)}{V_k(X)}$$

Sembarang nilai X dapat di masukkan ke dalam klaster ke-i, C_i, jika memenuhi kriteria (*K-NN rule*) berikut:

a. $size(K_{nn}(X) \cap C_c) = \underset{i=[1..c]}{Max}(size(K_{nn}(X) \cap C_i))$
b. Jika ada lebih dari satu *klaster* yang mengandung titik-titik maksimum (Max), maka X akan dimasukan pada klaster yang "*terdekat*" yang didefinisikan melalui jarak yang paling pendek.

Tran et. al. (Tran et. al, 2003) membuat modifikasi prosedur di atas melalui tahapan berikut:

a. Kelompokan X_i (i=1,..,N) menjadi N klaster dengan K-NN.
b. Implementasikan *K-NN rule*.
c. Jika tidak ada perubahan dari c klaster, maka iterasi selesai selainnya kembali ke langkah 2.

3.1.2 Fungsi Kernel Gaussian

Metode ini adalah pengembangan dari metode Pendugaan Densitas berbasis K-Nearest Neighbour yang diimplementasikan oleh Tranh (2003). Pada dasarnya metode ini menggunakan fungsi kernel Gaussian dengan rata-rata dan standard deviasi masing-masing bernilai 0 dan 1 sebagai basis fungsi densitas gen yang dinyatakan sebagai berikut:

$$f(x,c) = \frac{1}{n}\sum_{i=1}^{n} K_c(x,x_i)$$

dimana $K_c(x,x_i) = e^{1/2x_i^2}$, x_i adalah ekspresi gen ke *i* dan *n* adalah banyaknya gen yang terdapat pada klaster ke *c*. Suatu gen digolongkan termasuk ke klaster ke c, jika mayoritas k-neighbours dari gen tersebut membuat nilai *f*(*x*,*c*) menjadi maximum. Berbeda dengan *KNN-rule* di mana kriteria adalah semata-mata berdasarkan banyaknya *gen neighbours* yang termasuk di f(x,c).

4. ALGORITMA ILGC (ITERATIVE LOCAL GAUSSIAN KLASTERING)

Dengan alasan metode ini menggunakan Kernel Gaussian yang di bentuk secara local melalui teknik K-NN, maka metode ini kami namakan sebagai Iterative Local Gaussian Klastering (ILGC). Secara lengkap algoritma ILGC ini di lakukan melalui tahapan berikut:

Algoritma ILGC (Database, *k* neighbours)

1. *Inisialisasi jumlah klaster sesuai banyaknya informative gene yang digunakan.*
2. *Masukan gen x ke klaster c jika* $size(NN(x) \cap N(c)) = \underset{i=[1..n_c]}{Max}((f(x,c))$ *dimana NN(x) adalah nearest neighbour dari x*

berdasarkan jarak Euclidean, N(c) adalah Klaster ke c serta f(x,c) adalah fungsi Gaussian ke c.

3. *Bila tidak ada perubahan signifikan pada struktur klaster, iterasi telah konvergen, selainnya kembali ke langkah ke 2.*

Pada dasarnya algoritma ini merupakan kombinasi antara algoritma *Hierarchical Clustering* dan algoritma *Gaussian Mixture*. Berkaitan dengan algoritma Hierarchical Clustering pada proses penggabungan beberapa klaster menjadi klaster yang lebih besar berdasarkan nilai maximum dari fungsi kernel Gaussian seperti diperlihatkan pada langkah ke 2 algoritma ILGC. Berkaitan dengan algoritma Gausian-Mixture dalam kaitannya penggunaan kernel Gaussian untuk menduga densitas klaster suatu gen.

Kelebihan algoritma ini adalah dalam hal penggunaan fungsi kernel Gaussian sebagai basis penduga densitas struktur klaster. Bentuk kernel ini sangat sesuai dengan sebaran ekspresi gen dalam jumlah besar selain sangat umum dalam pendugaan densitas suatu sebaran data. Kelebihan lainnya dari algoritma ini adalah dalam penggunaan jumlah gen neighbours (*k*) sebagai satu-satunya parameter. Sehingga metode ini sangat mudah diimplementasikan secara komputasi.

Pemilihan Nilai *k* (Banyaknya *Neighbours*).

Masalah utama dalam implementasi algoritma ILGC adalah bagaimana menentukan nilai *k* untuk menentukan jumlah klaster yang optimum dalam database. Berdasarkan eksperimen dengan menggunakan database *Lymphoma*, nilai *k* ditentukan kurang lebih 5% dari banyaknya gen (baris) untuk banyaknya gen cukup besar (>200), gunakan nilai k=10 untuk banyaknya gen diantara 40-200, selainnya pilih *k=3*.

5. EXPERIMEN DENGAN DATABASE KANKER GETAH BENING (B-CELL DIFFUSE LARGE CELL LYMPHOMA)

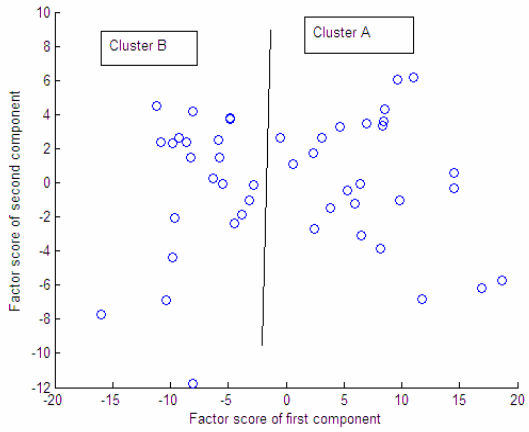
5.1 Database Kanker Getah Bening

Kanker lymphoma yang lebih populer dengan nama biologinya ***B-cell diffuse large cell lymphoma (B-DLCL)*** atau kanker limpoid atau getah bening mengandung beragam jenis tipe tumor yang secara signifikan telah dibuktikan dari hasil kajian secara morfologi, klinis dan respons terhadap treatment tertentu. Dari hasil penelitian Lossos *et. al* (2000) melalui penggunaan profiling ekspresi gen telah ditemukan bahwa B-DLCL ini memiliki dua klasters utama yaitu: klaster pertama adalah B-DLCL yang kurang aktif menyebar dan kurang berbahaya yang disebut sebagai ***germinal center B cell-like DLCL***. Klaster kedua lebih aktif dan lebih mematikan yang disebut sebagai activated B-cell like DLCL.

Eksperimen dengan database kanker lymphoma ini akan dilakukan pencarian banyaknya klaster dalam sampel (bagian kolom database) menggunakan suatu kelompok *informative genes* yang dipilih berdasarkan kriteria tertentu. *Informative genes* tersebut akan dianalisa lebih lanjut berdasarkan tingkat korelasi yang berkaitan dengan peluang hidup pasien yang mengidap salah satu jenis tipe kanker lymphoma ini. Dalam eksperimen ini, dipilih 226 *informative genes* yang dipilih berdasarkan beragamnya setiap gen dalam sampel serta memiliki dua tipe kanker lymphoma DLBCL (Lossos et. al, 2000). Database ini seluruhnya memiliki 4026 gen serta 47 sampel.

5.2 Hasil Eksperimen

Hasil proses penentuan klaster dengan menggunakan algoritma ILGC dengan banyaknya neighbours, *k=10*, serta taraf konvergensi=0.95 ditunjukan pada Gambar 1. Berdasarkan hasil analisa melalui Gambar 1 menunjukan bahwa terdapat dua klaster dalam sampel: klaster A (24 sampel) dan klaster B (23 sampel) . Hasil ini identik dengan hasil penelitian Alizadeh *et. al.* (2000) dan Lossos *et. al*. (2000). Menurut sifat biologinya, klaster A ini termasuk dalam tipe kanker *Germinal Center B-Like DLBCL* yang kurang berbahaya dibandingkan pada klaster B yang biasa disebut sebagai *Activated B-Like DLBCL* yang lebih aktif menyebarkan penyakit kanker. Pada Tabel 1 dapat dilihat keanggotaan sampel pada klaster A (*Germinal Center B-Like DLBCL*) dan B (*Activated B-Like DLBCL).*

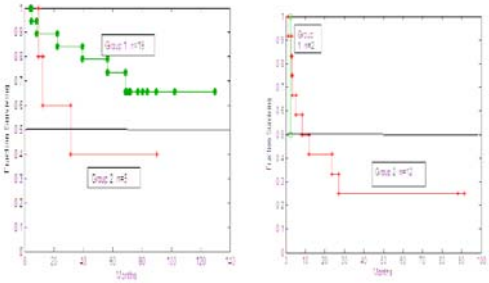


Gambar 1. Pembentukan Klaster sampel kanker Lymphoma (DLBCL) menggunakan Algoritma ILGC yang menghasilkan dua klaster, klaster A dan B, masing-masing memiliki 24 dan 23 sampel.

Untuk melihat korelasi antara 226 gen (*informative gene*) tersebut di atas dengan parameter klinis kanker Lymphoma, 226 gen tersebut di analisa pengaruhnya terhadap peluang hidup pasien menggunakan kurva Kaplan-Meier seperti yang terliha pada Gambar 9 berdasarkan nilai IPI (International Prognosis Indicator). Korelasi dilakukan berdasarkan 2 kategori nilai IPI, dimana jika nilai IPI diantara 0-2 maka kategori ini disebut sebagai berisiko rendah secara klinis (*low clinical riks patients*), sedangkan jika nilai IPI diantara 3-4 maka termasuk dalam beresiko tinggi secara klinis (*high clinical risks patients*).

Tabel 1. Dua type kanker DLBCL berhasil diidentifikasi dari implementasi algoritma ILGC

Type Kanker DLBCL	***Sampel***
GC B-Like DLBCL	DLCL-0012 DLCL-0024 DLCL-003 DLCL-0026 DLCL-0023 DLCL-0015 DLCL-0010 DLCL-0030 DLCL-0034 SUDHL-6 DLCL-0018 DLCL-0032 DLCL-0052 DLCL-0037 DLCl-0001 DLCL-0008 GC B GC Centroblast DLCL-004 DLCL-0029 DLCL-000 DLCL-0020 DLCL-0051 DLCL-0033
Activated B-Like DLBCL	DLCL-005 DLCL-0011 DLCL-0048 DLCL-0027 DLCL-0013 DLCL-0007 DLCL-0028 DLCL-0025 DLCL-0021 DLCL-0021 DLCL-0016 DLCL-0002 DLCL-0017 OCI-Ly3 DLCL-0040 DLCL-0014 DLCL-0031 DLCL-0036 DLCL-0039 DLCL-0042 OCI-Ly10 DLCL-0041 DLCL-0049 DLCL-0006



a. low clinical risks b. high clinical risks

Gambar 2. Peluang pasien bertahan hidup dengan indikator nilai IPI.

Gambar 2.a menunjukan peluang pasien bertahan hidup pada nilai IPI diantara 0-2 (*low IPI*), dan Gambar 2.b menunjukan peluang pasien bertahan hidup pada nilai IPI diantara 3-4 (*high IPI*).

Dari gambar 2a terlihat bahwa pada kategori *low clinical risks patients*, terdapat dua group yang berbeda nyata pada taraf 1%: pasien pada group 1 memiliki peluang hidup lebih tinggi (65.5%) dibandingkan dengan peluang untuk bertahan hidup pada group 2 (40%). Pada gambar 2b menunjukan bahwa peluang hidup dari dua group juga berbeda nyata pada taraf 1%, dimana pasien pada group 1 memiliki peluang hidup lebih tinggi (50%) dibandingkan pasien pada pada group 2 (25%).

Tabel 2. Perbandingan hasil identifikasi banyaknya sampel dalam setiap group berdasarkan nilai IPI. Supercript 1, 2 dan 3 menunjukkan hasil penelitian Wasito *et. al*. (2005), Alizadeh *et. al*. (2000) serta Hastie *et. al*. (2000).

	Nilai IPI[1]		Nilai IPI[2]		Nilai IPI[3]	
	0-2	3-4	0-2	3-4	0-2	3-4
Group1	19	2	13	5	7	7
Group2	5	12	6	12	11	7

Nilai korelasi antara group dari pasien untuk setiap kategori nilai indikator IPI dapat dilihat pada tabel 2. Sebagai bahan perbandingan diperlihatkan pula banyaknya sampel dari group 1 dan group 2 dari hasil penelitian Alizadeh et. al. (2000) dengan algoritma *Hierarchical Klastering* serta dari hasil penelitian Hastie et. al (2000) dengan metode *Gene Shaving.*

6. KESIMPULAN

Pada makalah ini telah diperkenalkan algoritma ILGC (Iterative Local Gaussian Klastering) yang pada dasarnya berbasis pada penggunaan nilai maximum densitas kernel Gaussian diantara *k nearest neighbours* sebuah ekspresi gen yang akan dimasukan (*assigned*) kepada sebuah klaster.

Berdasarkan hasil eksperimen dengan menggunakan database Lymphoma menunjukan bahwa terdapat dua klaster sampel dalam database. Hasil ini identik dengan banyaknya klaster yang ditemukan dalam penelitian Alizadeh *et. al.* (2000), Lossos *et. al* (2000) serta Hastie *et. al*. (2000). Dalam analisa peluang hidup seorang pasien yang memiliki parameter klinis Lymphoma berdasarkan nilai IPI (International Prognosis Indicator) menggunakan kurva Kaplan-Meier menunjukan bahwa hasil penelitian ini memiliki pola (pattern) yang serupa dengan hasil penelitian Alizadeh *et. al.* (2000).

DAFTAR PUSTAKA

Alizadeh, A.A. et. al. (2000). Distinct types of difuse large B-Cell lymphoma identified by gene expression profiling. *Nature*, 403, 3.

Brown, P.O. dan Botsein, D. (1999). Exploring a new world of the genome with DNA microarrays. *Nature Genetics Supplement*, Vol. 21.

Hastie, T. *et al.* (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expressions patterns. *Genome Biology*. vol 1, pp. 1-21.

Furey, T.S, et al. (2000). Support vector machine classification and validation cancer tissues samples using microarray expression data. *Technical report*, University of California, Santa Cruz.

Jain, A.K. dan R.C Dubes. (1988). *Algorithms for klastering data*, Prentice Hall.

Lossos, I.S, et al. (2000). Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. *PNAS*, vol. 97.

Mitchel, T. (1997). *Machine Learning*. McGraw-Hill.

Theoridis, S. dan Koutroumsbas. (2003). *Pattern Recognition*. Academic Press.

Tran, T. N, Wehrens, R. dan Buydens, M. C. (2003). KNN Density-based klastering for high dimensional multispectral images. *Proceeding 2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*, URBAN 2003, Berlin Germany.

Wasito et. al.. (2005). Identifikasi tipe kanker berdasarkan profiling ekspresi gen menggunakan piranti lunak berbasis pendugaan densitas dengan teknik K-Nearest Neighbour. *Laporan Penelitian Hibah Bersaing XIII*, Dikti, Depdiknas, Jakarta.

Wasito, I., Estri, M. N. dan Zabidi, S. A (2005). Selection of Number of Factors and Neighbours in Global-Local Iterative Least Squares Data Imputation (INI) Algorithm for Microarrays Gene Expression, *Proceeding of UKCI05*, London, UK.

Wasito, I. dan Mirkin, B. (2005). Nearest neighbour approach in the least-squares data imputation algorithms, *Information Sciences* vol.1, pp. 691-20.