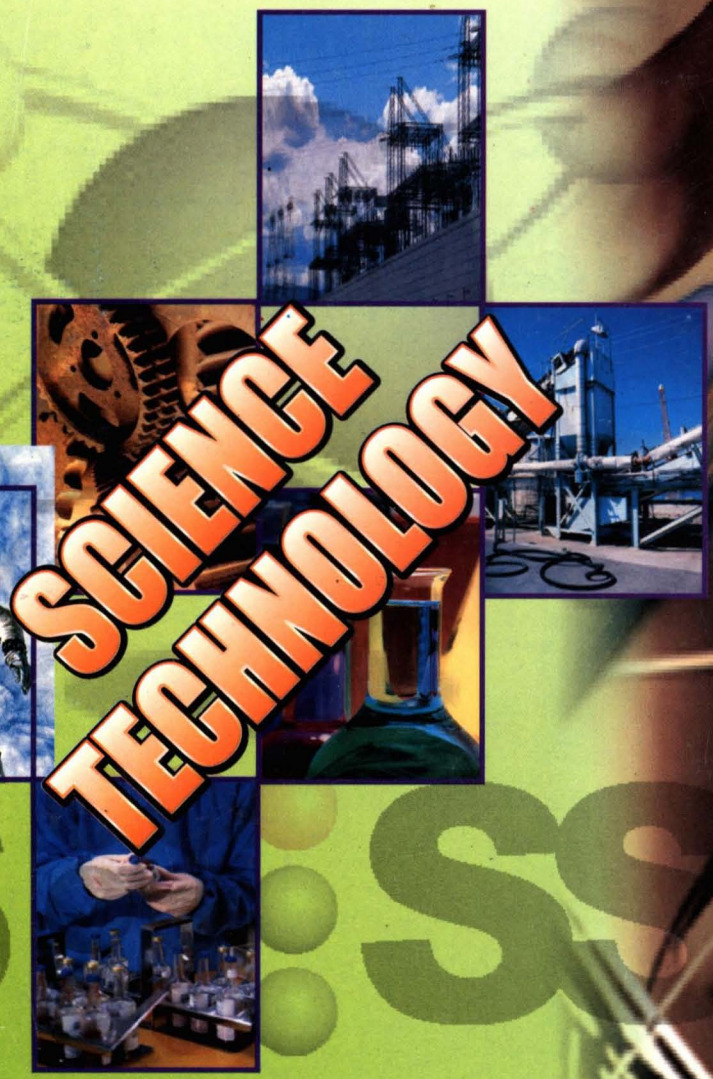
UNIVERSITI TEKNOLOGI MARA

ISO 9001:2000 No Sijil :0500132

Globalising Knowledge and Information



NATIONAL SEMINAR ON SCIENCE TECHNOLOGY & SOCIAL SCIENCES 2006

30-31 May 2006

Swiss Garden Resort & Spa

Kuantan, Pahang

Pengoptimuman Kelompok Sebatian Kimia Menggunakan Teknik Algoritma Genetik (GA)

Rosmayati Mohamed
Zuriana Abu Bakar
Naomie Salim

ABSTRAK

Proses mengenalpasti molekul-molekul yang berpotensi untuk dijadikan ubat merupakan satu proses yang rumit dan lama. Kewujudan berjuta-juta molekul dalam pangkalan data kimia menambahkan lagi kesukaran kepada ahli kimia untuk memprosesnya. Disebabkan oleh faktor kos dan masa, algoritma pemilihan sebatian telah digunakan di mana hanya satu set perwakilan kecil sebatian dipilih daripada sebuah pangkalan data kimia yang besar. Salah satu teknik dalam pemilihan sebatian adalah dengan menggunakan pengelompokan. Algoritma pengelompokan Ward untuk mengelompokkan molekul-molekul yang diwakilkan dalam bentuk perwakilan molekul 2D sering digunakan dalam penyelidikan kimiainformatik. Dalam kajian ini, teknik algoritma genetik (GA) digunakan untuk mengoptimumkan kelompok-kelompok yang diperolehi melalui pengelompokan Ward dan hasil ujikaji dibandingkan dengan pengelompokan Ward. Perbandingan di antara kelompok-kelompok yang dihasilkan menggunakan Ward dan kelompok-kelompok yang dioptimumkan menggunakan GA mendapati kelompok yang dioptimumkan menggunakan GA menunjukkan peningkatan dari segi jarak ketidakserupaan antara kelompok. Ini menunjukkan penggunaan GA dalam pengelompokan Ward mempunyai kecenderungan dalam menghasilkan pemilihan sebatian yang lebih pelbagai.

Kata kunci: *Pengelompokan, algoritma genetik, dan kimia informatik*

Pengenalan

Faktor kos dan masa memainkan peranan penting dalam proses mencari molekul-molekul yang berpotensi untuk dijadikan ubat. Beribu-ribu juta molekul dengan setiap satunya mempunyai ciri-ciri yang berbeza dalam pangkalan data kimia perlu dianalisis untuk melihat struktur aktiviti biologinya. Oleh itu, teknologi maklumat digunakan untuk mencari sebatian aktif yang baru bagi menjimatkan kos dan masa dalam proses penemuan ubat (Salim 2002). Salah satu pendekatan yang digunakan ialah teknik pengelompokan dalam pemilihan subset perwakilan molekul kimia (Brown & Martin 1996). Pengelompokan membenarkan hanya sebahagian kecil subset perwakilan molekul yang mewakili keseluruhan aktiviti biologi dipilih untuk diuji.

Analisis kelompok merupakan proses meletakkan sekumpulan objek (molekul atau sebatian) ke dalam kelas-kelas tertentu, atau kelompok di mana darjah keserupaan yang tinggi dalam kelompok dan darjah ketidakserupaan yang tinggi di antara kelompok. Pengelompokan perwakilan molekul kimia didasarkan pada prinsip keserupaan yang menyatakan molekul yang serupa dari segi strukturnya akan mempunyai ciri-ciri biologi yang sama (Brown & Martin 1996; Gillet et al. 1998). Maka, hanya sebahagian perwakilan molekul yang perlu dipilih untuk diuji bagi setiap kelompok.

Kajian-kajian yang telah dilakukan oleh Brown & Martin (1996) dan Borosy et al. (2001), mendapati pengelompokan Ward merupakan teknik yang terbaik dalam pengelompokan sebatian kimia. Mereka telah membuktikan kaedah berhierarki timbunan dapat mengasingkan sebatian aktif dan tidak aktif secara konsisten. Dalam kertas kerja ini, kami mengkaji penggunaan algoritma genetik (GA) untuk meningkatkan keberkesanan (optimum) setiap kelompok sebatian kimia yang dihasilkan daripada teknik Ward. GA telah digunakan untuk pengelompokan dalam pelbagai bidang seperti rangkaian, dokumen, geografi, biologi, pemprosesan imej, dan sebagainya. Kajian dalam pengelompokan menggunakan GA menghasilkan keputusan yang menggalakkan apabila ianya digunakan untuk mengoptimumkan fungsi objektif dalam algoritma K-Means (Painho & Bacao 2000). Terdapat kajian yang menggabungkan pengelompokan berhierarki dan GA menunjukkan hasil yang baik dari segi mengasingkan set data ke dalam kumpulan yang betul (Green 2003). Setakat itu, menurut Salim (2002) kaedah pengelompokan terbaik didapati masih tidak sempurna dan perlu dikaji sama ada ianya boleh diperbaiki dengan menggunakan teknik-teknik lain seperti GA, fuzzy, rangkaian neural dan sebagainya.

Tujuan utama kertas kerja ini adalah untuk melihat keberkesanan teknik GA dalam mengoptimumkan jarak ketidakserupaan antara kelompok yang dihasilkan daripada teknik pengelompokan Ward. Untuk mencapai tujuan ini, teknik ini perlu mampu memisahkan dan meletakkan struktur molekul yang berbeza ke dalam kelompok yang berasingan untuk meningkatkan kepelbagaian set sebatian seperti yang telah dilakukan oleh Willet (1997).

Algoritma

Pengelompokan Ward

Konsep utama pengelompokan Ward adalah untuk mengurangkan bilangan kelompok daripada n kepada n-1 dan pada masa yang sama dapat meminimakan kehilangan maklumat (Ward 1963). Algoritma pengelompokan Ward dimulakan dengan n kelompok di mana setiap kelompok mengandungi satu sebatian. Bilangan kelompok dikurangkan kepada n-1 dengan meletakkan dua sebatian yang mempunyai struktur molekul yang hampir serupa ke dalam kelompok yang sama. Keserupaan di antara kelompok diukur dengan menggunakan pekali Euclidean $D_{A,B}$.

$$D_{A,B} = \frac{\sqrt{a+b-2c}}{n}$$

dengan a ialah bilangan posisi bit yang disetkan bernilai "1" dalam molekul A, b ialah bilangan posisi bit yang disetkan bernilai "1" di dalam molekul B, c ialah bilangan posisi bit yang disetkan bernilai "1" hasil kesatuan di antara molekul A dan B sementara n ialah jumlah posisi bit dalam rentetan bit yang mewakili dua molekul yang dibandingkan.

Algoritma Genetik (GA)

Dalam kertas kerja ini, kami menggunakan algoritma genetik yang ringkas. Satu populasi awalan yang mengandungi c kromosom dijana secara rawak, $P(c)$. Setiap kromosom mewakili n-ahli subset atau molekul. Setiap n-ahli subset dikumpukkan dengan i kelompok. Untuk mengukur kesesuaian setiap kromosom adalah dengan menggunakan purata jarak ketidakserupaan antara kelompok. Berikut merupakan fungsi kesesuaian yang digunakan:

$$PD = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} T_{ij}}{n^2}$$

dengan T_{ij} mewakili pekali Tanimoto yang mengukur jarak keserupaan di antara pusat kelompok i dan j manakala n merujuk kepada bilangan pusat kelompok (Turner et al. 1997).

Proses *crossover* dilaksanakan untuk menghasilkan dua individu baru yang unik. Dua kromosom (dipanggil ibu dan bapa) dipilih berdasarkan kaedah pemilihan berasaskan penyusunan (*rank-based selection)* untuk melakukan proses *crossover*. *Crossover* satu titik dilakukan pada kromosom ibu $\{k_1, ..., k_C\}$ dan kromosom bapa $\{k'_1, ..., k'_C\}$ dengan kebarangkalian p_c. Titik bagi melakukan proses *crossover* pada setiap kromosom dipilih secara rawak dengan memilih integer i di antara 1 dan n. Kedua-dua kromosom akan dibahagikan kepada dua bahagian pada kedudukan i. Bahagian $\{k_{i+1}, ..., k_C\}$ dan $\{k'_{i+1}, ..., k'_C\}$ saling ditukar di antara satu sama lain untuk menghasilkan dua individu atau kromosom yang baru (dipanggil anak).

Proses mutasi dilaksanakan pada gen dalam kromosom dengan kebarangkalian kecil p_m untuk menghasilkan satu individu baru. Satu nombor integer rawak j di antara 1 dan n dijana untuk menentukan *locus* gen di dalam kromosom yang perlu dimutasikan. Kaedah elitism digunakan bagi memastikan b ahli kromosom terbaik dalam populasi semasa akan kekal untuk populasi seterusnya. Di sini, kromosom yang mempunyai fungsi kesesuaian yang tertinggi akan disalin ke dalam generasi seterusnya. Proses ini akan berterusan sehingga bilangan generasi yang ditetapkan dicapai.

Rekabentuk Eksperimen

Set Data

Set data yang digunakan dalam kajian ini ialah set data AIDS yang diperolehi daripada pangkalan data Institut Kanser Kebangsaan (NCI). Set data ini mengandungi sebanyak 1000 sebatian yang terdiri daripada 247 sebatian aktif (CA), 253 sebatian sederhana aktif (CM) dan 500 sebatian tidak aktif (CI). Setiap struktur di dalam set data ini diberikan dengan perwakilan molekul dalam bentuk rentetan bit BCI dari *Barnard Chemical Information* menggunakan yang mempunyai panjang rentetan 1052 bit.

Parameter

Proses dimulakan dengan populasi yang mengandungi kromosom yang mewakili molekul dan kelompok yang diperolehi daripada teknik Ward. Bilangan kromosom dalam populasi awalan adalah 50 kromosom yang mewakili 50 subset. Kromosom pertama diperolehi daripada teknik Ward dan selebihnya dipilih secara rawak. Setiap kromosom terdiri daripada molekul-molekul yang diwakilkan dalam setiap gen. Setiap *allele* mewakili nombor kelompok. Gen yang mempunyai nilai *allele* yang sama berada di dalam kelompok yang sama.

Kadar mutasi (p_m) dan *crossover* (p_c) ialah masing-masing 0.09 and 0.6. Mutasi melibatkan proses menukar satu elemen di dalam kromosom kepada elemen baru yang mewakili molekul untuk membentuk kromosom yang berlainan di dalam subset. *Crossover* satu titik diubahsuai supaya setiap kromosom baru yang terbentuk tidak sama dengan kromosom yang sudah sedia ada.

Kelompok yang diperolehi daripada teknik Ward dan kelompok-kelompok yang dioptimumkan dengan menggunakan GA dianalisis keberkesanannya dengan menggunakan ukuran purata ketidakserupaan antara kelompok (persamaan 2).

Hasil Eksperimen

Jadual 1 menunjukkan perbandingan berdasarkan ukuran purata ketidakserupaan antara kelompok yang diperolehi daripada kelompok yang dijana melalui teknik Ward dan kelompok yang dioptimumkan dengan menggunakan GA. Pengujian dilakukan daripada 10 kelompok sehingga 100 kelompok. Hasil eksperimen berdasarkan pada ukuran ini menunjukkan kelompok-kelopok sebatian kimia yang dioptimumkan menggunakan teknik GA mempunyai potensi untuk memperolehi keputusan yang lebih baik jika dibandingkan dengan teknik Ward.

Peningkatan hasil ujikaji dari segi jarak ketidakserupaan antara kelompok menunjukkan keupayaan teknik GA untuk memisahkan dan meletakkan sebatian yang mempunyai struktur yang hampir sama ke dalam kelompok yang sama. Di samping itu, teknik GA juga memastikan kelompok-kelompok yang berbeza mempunyai jarak ketidakserupaan yang optimum di antara satu sama lain.

Jadual 1: Perbandingan berdasarkan purata ketidakserupaan antara kelompok di antara Ward dan GA

Bilangan Kelompok	Purata Ketidakserupaan antara Kelompok	
	Ward	**GA**
10	0.73665	0.75034
20	0.74219	0.74896
30	0.74723	0.75260
40	0.75186	0.75774
50	0.74919	0.75406
60	0.74788	0.75279
70	0.74637	0.75034
80	0.75062	0.75364
90	0.75238	0.75524
100	0.75297	0.75549

Kesimpulan

Perbandingan di antara kelompok-kelompok yang dihasilkan melalui teknik Ward dan kelompok-kelompok yang dioptimumkan menggunakan GA menunjukkan sedikit peningkatan dari segi ketidakserupaan antara kelompok. Keputusan ini membuktikan kemampuan dan potensi teknik GA dalam menghasilkan set pilihan sebatian yang lebih pelbagai.

Banyak kajian tambahan yang boleh dilakukan iaitu dengan menggunakan set data yang lebih besar dan mengandungi pelbagai aktiviti. Selain itu, pelbagai kaedah pemilihan dalam algoritma genetik boleh diaplikasikan seperti *roulette wheel*, *tournament* dan *spatial*.

Penghargaan

Penghargaan kepada University of Sheffield kerana sumbangan set data NCI yang digunakan dalam kajian ini.

Rujukan

Borosy, A., Csizmadia, F. and Volford, A. (2001). Structure Based Clustering of NCI's Anti-HIV Library. *First Symposium of the European Society of Combinatorial Science, Budapest.*

Brown, R. D. and Martin, Y. C. (1996). Use of Structure-Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *Journal of Chemical Information and Computer Sciences*. 36: 572-584.

Gillet, V. J., Wild, D. J., Willett, P. and Bradshaw, J. (1998). Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *The Computer Journal*. 41: 547-558.

Green, A. W. (2003). Unsupervised Hierarchical Clustering Via a Genetic Algorithm. *Proceedings of the 2003 Congress on Evolutionary Computation, December 9-12, 2003, Canberra, Australia. IEEE Press*. 998-1005.

Painho, M. and Bacao, F. (2000). Using Genetic Algorithm in Clustering Problems. (online). http://www.geocomputation.org/2000/GC015/Gc015.htm (19 November 2003).

Salim, N. (2002). *Analysis and Comparison of Molecular Similarity Measures*. Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom.

Turner, D. B., Tyrell, S. M., Willet, P. (1997). Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *Journal of Chemical Information and Computer Sciences*. 37: 18-22.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. 58: 236-244.

Willet, P (1997). Computational Tools for the Analysis of Molecular Diversity. Perspective in Drug Discovery and Design. 7/8:1-11

ROSMAYATI MOHAMED & ZURIANA ABU BAKARFakulti Sains & Teknologi, Kolej Universiti Sains & Teknologi Malaysia (KUSTEM), Mengabang Telipot, 21030, Kuala Terengganu, Terengganu.

, Fakulti Sains & Teknologi, Kolej Universiti Sains & Teknologi Malaysia (KUSTEM), Mengabang Telipot, 21030, Kuala Terengganu, Terengganu.

NAOMIE SALIM, Fakulti Sains Komputer & Sistem Maklumat,Universiti Teknologi Malaysia, 81310, Skudai, Johor Bahru, Johor.