http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 2 No. 1, June 2013 : 39 - 47
e-ISSN: 2289-2192

CARIAN PERSAMAAN PANGKALAN DATA DRUG : SATU ULASAN

SHEREENA M. ARIF
NURUL MALIM

ABSTRAK

Carian persamaan merupakan asas aplikasi dalam bidang informatik kimia yang mengguna kaedah informatik untuk menyelesai masalah kimia. Kertas ini membincang komponen prinsip dalam carian persamaan yang menumpu kepada penyaringan maya 2D. Perbincangan juga menekankan latar belakang sejarah tentang prinsip dan kaedah yang diguna dalam bidang ini.

Kata kunci: Carian Persamaan, Fingerprints, Prinsip Sifat Serupa, Perwakilan Struktur, Pekali Persamaan, Penilaian Persamaan.

ABSTRACT

Similarity searching served as the basis for chemoinformatics field where informatics methods are applied to solve chemical problem. This paper, discusses the principle components of similarity searching, which focus on 2D virtual screening. The discussion also highlights the historical background of the principles and methods used in the field.

Keywords: Similarity search, Fingerprints, Similar Properties Principle, Structure Representation, Similarity Coefficient, Similarity Evaluation

PENGENALAN

Carian persamaan bagi pangkalan data kimia melibatkan perbandingan antara keseluruhan struktur pertanyaan pengguna (disebut seterusnya sebagai sebatian pertanyaan) dengan struktur sebatian dalam pangkalan data kimia (disebut seterusnya sebagai sebatian pangkalan data) berdasarkan ukuran persamaan yang menilai darjah persamaan antara sebatian kimia. Willet et al. (1998) membahagi ukuran persamaan ini kepada tiga komponen, iaitu pemerihal struktur (*structural descriptor*), pekali persamaan (*similarity coefficient*) dan skema pemberat (*weighting scheme*). Bagaimanapun, dalam bidang kimia informatik, skema pemberat ini merupakan sebahagian daripada kaedah pempiawaian. Kertas ini tidak mengulas dengan lanjut tentang skema pemberat ini.

Carian persamaan berasaskan *fingerprint* diguna dalam konteks tertentu. Sebatian pertanyaan dan sebatian pangkalan data diwakili oleh set setara pemerihal molekul yang mewakili sifat sebatian, iaitu *fingerprints*. Persamaan antara set pemerihal ini dikira mengguna pekali persamaan, yang merupakan ukuran berangka berdasarkan kepada pengiraan darjah persamaan antara sebatian berdasarkan perwakilan bit dalam *fingerprints*.

PANGKALAN DATA DRUG

Pangkalan data drug dicirikan oleh dua set data yang berlainan, iaitu set aktif dan set tidak aktif. Kedua-dua set data ini disimpan dalam pangkalan data yang menyimpan maklumat

berkaitan dengan bioaktiviti iaitu sama ada telah disintesis dan terbukti aktif atau tidak aktif. Terdapat beberapa pangkalan data drug yang tersedia secara meluas seperti berikut:

1. *Comprehensive Medicinal Chemistry* (CMC) mengandungi lebih daripada 8400 sebatian farmaseutikal yang diperoleh daripada Kompendium Dadah dalam siri Comprehensive Medicinal Chemistry oleh Pergamon Press dan dikemas kini pada setiap tahun dengan sebatian yang dikenal pasti oleh *United States Approved Names* (USAN) (Jonsdottir et al., 2005).
2. *World Drug Index* (WDI) adalah koleksi sebanyak 80,*000* drug yang dipasar dan sedang dimaju di seluruh dunia yang diperoleh daripada 1200 jurnal saintifik dan persidangan prosiding (Lambert 2000).
3. *DrugBank* merupakan sebuah pangkalan data awam yang terdiri daripada sebatian yang diekstrak daripada pangkalan data rangkaian protein, buku teks perubatan kimia dan buku rujukan kimia (Wishart et al., 2006).
4. *MDL Drug Data Report* (MDDR) adalah sebuah pangkalan data komersial yang mengandungi 185,844 sebatian (penerbitan 2008.1) yang baharu dilancar atau sedang dimaju, diperoleh daripada penulisan paten, jurnal, mesyuarat dan kongres (Southan et al., 2009).
5. *National Cancer Institute* (NCI) adalah gabungan empat pangkalan data yang berjumlah sebanyak 213,000 sebatian (Jonsdottir et al., 2005).
6. *World of Molecular BioActivity* (WOMBAT) adalah sebuah pangkalan data komersial yang mengandungi 309,847 sebatian tulen yang diperoleh daripada 15,320 jurnal dan prosiding (Sunset-Molecular 2011).

KOMPONEN UTAMA UKURAN PERSAMAAN

PRINSIP SIFAT SERUPA DAN PERILAKU KETETANGGAAN

Carian persamaan adalah berdasarkan kepada konsep yang berasal daripada premis utama dalam kimia ubatan yang menyatakan bahawa sebatian yang berstruktur serupa mempunyai aktiviti biologi yang sama. Konsep ini dikenali sebagai Prinsip Sifat Serupa (Johnson & Maggiora, 1990). Prinsip ini bagi menentu tiga jenis persamaan dalam kimia: persamaan kimia, yang melibatkan perbandingan dan pengumpulan sistem kimia berkenaan dengan pelbagai sifat makroskopik; persamaan molekul, yang melibatkan perbandingan dan pengumpulan sesuatu sebatian berdasarkan struktur 2D atau 3D dan maklumat berkaitan dengan sifat sebatian tersebut; dan persamaan intra-molekul, yang melibatkan perbandingan dan pengumpulan sifat dalam sesuatu sebatian.

Kajian Patterson et al. (1996) mencadangkan satu konsep yang berkaitan dengan Prinsip Sifat Serupa, iaitu Perilaku Ketetanggaan. Kajian tersebut memperluas skop hipotesis sebelum ini, yang menyatakan bahawa sebatian dalam lingkungan kawasan yang sama cenderung untuk menunjukkan aktiviti yang serupa. Penemuan ini turut disokong oleh kajian yang menyimpul bahawa aktiviti biologi meningkat dengan persamaan struktur (Martin et al., 2002). Semakin tinggi potensi drug aktif yang diguna untuk carian tersebut, maka semakin meningkat pecahan aktif dalam sebatian yang serupa.

PERWAKILAN STRUKTUR

Perwakilan struktur yang juga dikenali sebagai pemerihal molekul, merupakan pencirian sesuatu sebatian kimia. Ia dibahagi kepada tiga kelas berdasarkan dimensi (D) dalam ruang

data kimia. Pemerihal 1D adalah vektor (nombor tunggal) yang mewakili sifat-sifat sebatian kimia, seperti berat molekul, bilangan atom, dan sebagainya. Pemerihal 2D mewakili sebatian kimia berdasarkan maklumat yang koordinat atom dan ditumpu kepada isu konformasi sebatian. Perbincangan tentang carian persamaan berasaskan *fingerprint* dalam artikel ini menumpu pada pemerihal 2D. Willet (2007) melakukan satu ulasan menyeluruh dan terkini mengenai pemerihal yang lain.

Fingerprints 2D adalah perwakilan perduaan (*binary*) satu sebatian dalam bentuk rentetan bit atau susunan Boolean. Rentetan bit adalah siri "0" dan "1". Setiap kedudukan bit dipadan dengan kehadiran serpihan kimia (*chemical fragment*) yang ditandai dengan 1, dan ketiadaan serpihan ditandai dengan 0. Ia dikira daripada graf 2D perwakilan suatu sebatian yang dikenali sebagai jadual sambungan.

***Kekunci Struktur* (*Structural Keys*)** adalah *fingerprint* berasaskan kamus dengan setiap kedudukan bit adalah sepadan dengan serpihan substruktur tertentu yang ditakrif dalam kamus serpihan (*fragment dictionary*). Kamus serpihan tersebut direka berdasarkan hasil analisis statistik sebatian yang dijangka disimpan dalam pangkalan data dan pertanyaan lazim yang mungkin dikemuka (Leach & Gillet, 2003). MACCS (Molecular Access System) dan ISIS (Integrated Scientific Information System) adalah antara kekunci struktur dan kamus serpihan pertama yang direka untuk carian substruktur oleh MDL Information Systems Inc. (Leach & Gillet, 2003).

Hashed Fingerprints merupakan alternatif kepada pendekatan berasaskan kamus. Ia dijana secara aljabbar dengan mengguna fungsi *hash*, yang merupakan satu kaedah berketentuan untuk mengubah sejenis data kepada nombor kecil yang bertindak sebagai "tandatangan" unik untuk data tersebut. Setiap bit dalam *hashed fingerprints* ditetapkan oleh kombinasi serpihan dan fungsi *hash* yang berbeza. Dalam kata lain, setiap serpihan boleh diproses mengguna beberapa fungsi *hash* yang berbeza. *Hashed fingerprints* yang pertama dilaksana oleh Daylight Chemical Information System Inc. (Leach & Gillet, 2003).

***Fingerprints Substruktur Membulat (Circular Substrcuture Fingerprints*)** dibangun untuk menciri serpihan yang tertumpu pada sesuatu atom dan ikatan yang melingkunginya (Hassan et al., 2006). Bremser (1978) mentakrifnya sebagai "suatu pencirian persekitaran sfera atom tunggal dan sistem gegelang yang lengkap". *Fingerprints* Kesalinghubungan Berlanjutan (*Extended Connectivity Fingerprints* - ECFP) dan *Fingerprints* Kesalinghubungan Fungsian (*Functional Connectivity Fingerprints* - FCFP) yang dibangun oleh SciTegic adalah contoh *fingerprints* dalam kelas ini. *Fingerprints* ini mewakili atom melalui rentetan nilai kesambungan berlanjutan yang dikira mengguna aljabbar Morgan (Hert et al., 2004). Antara kriteria yang membeza kedua-dua *fingerprints* ini ialah petua dalam penjenisan atom yang diguna untuk menghasilnya, sama ada mengguna petua varian atom (ECFP) ataupun mengguna petua kelas fungsian (FCFP) (Hassan et al., 2006).

Hologram adalah *fingerprint* bernilai integer. Setiap kedudukan bit dalam rentetan bit menyimpan kekerapan kewujudan serpihan berbanding nilai "1", yang hanya menandai kehadiran serpihan tanpa mengira berapa kali ia berlaku. Rasionalnya, jumlah kewujudan suatu struktur serpihan yang berkait rapat dengan aktiviti biologi boleh menjadi pengukur kepada sejauh mana dua struktur kimia mempunyai kualiti fungsian yang sama.

***Fingerprints Topologi Pharmacophore* (*Topological Pharmacophore Fingerprints*)** mengekod corak titik *pharmacophore* 3- atau 4- bersama dengan jarak antara titik yang sepadan (Hert et al., 2004). Corak ini adalah ciri-ciri *pharmacophore* (contohnya gegelang aromatik) yang diperlu untuk bioaktiviti. Ia boleh wujud dalam dua bentuk, iaitu apabila diguna dalam ruang 3D, jarak antara titik ditentu oleh jarak atom manakala apabila diguna

dalam ruang 2D, jarak antara titik ditentu oleh jarak melalui ikatan (Hert et al., 2004). Contoh *fingerprints* jenis ini ialah pemerihal Kekunci Similog dan *Chemically Advance Template Search* (CATS) *(*Hert et al., 2004).

PEKALI PERSAMAAN

Pekali persamaan mengukur tahap persamaan antara sebatian. Terdapat 4 kelas pekali seperti yang diterangkan oleh Sneath dan Sokal (1973) iaitu jarak, gabungan, korelasi, dan kebarangkalian.

Dalam perwakilan binari (dikotomi) sebatian, Willett et al. (1998) memerihal pekali gabungan (persamaan) dan jarak (perbezaan) berdasarkan empat terma. Katakan sebatian A dibanding dengan sebatian B dan kedua-duanya diwakili oleh bit binari. Pembolehubah *a,* adalah bilangan bit "yang wujud" dalam sebatian A. Pembolehubah *b,* adalah bilangan bit "yang wujud" dalam sebatian B. Pembolehubah *c,* adalah bilangan bit "yang wujud" dalam kedua-dua A dan B. Pembolehubah *d* pula merupakan bilangan bit "yang tidak wujud" dalam kedua-dua A dan B.

Oleh itu: $n = a + b - c + d$

dengan *n* adalah jumlah bilangan bit dalam satu rentetan bit. Persamaan antara sebatian A dan B adalah S(A, B) dan perbezaan antara keduanya adalah D(A, B).

Pekali Jarak (*Distance coefficients*) mengira perbezaan jarak antara dua sebatian yang sedang dipertimbang dalam suatu ruang pemerihal (Sneath & Sokal, 1973). Sebatian adalah sama jika kedudukan mereka selari, yang bermaksud bahawa jarak di antaranya adalah 0. Oleh itu, apabila nilai jarak antara dua sebatian meningkat, maka kebarangkalian sebatian untuk serupa menurun. Pekali jarak Euclidan dan Hamming adalah dua pekali jarak yang popular diguna dalam bidang informatik kimia.

Pekali gabungan (*Association coefficients*) mengukur keselarian antara dua sebatian (Sneath & Sokal, 1973). Pekali ini sesuai diguna ke atas data binari berbanding dengan data selanjar yang mana kehilangan maklumat mungkin berlaku. Bertentangan dengan pekali jarak, pekali gabungan mengukur persamaan antara dua sebatian dengan nilainya adalah antara 0 (menandai tidak ada ciri-ciri yang sama), hingga 1 (menandai pemerihal dengan kedudukan bit yang serupa) (Salim et al., 2003).

Pekali Padanan Mudah adalah antara pekali pertama yang diguna dalam informatik kimia. Ia adalah pekali tertua yang mula diperkenal dalam bidang taksonomi berangka (Sneath & Sokal, 1973), dengan mengambil kira aspek kehadiran dan ketiadaan ciri-ciri serpihan dalam pengiraannya. Ini kemudiannya dikritik oleh Willet et al. (1998) yang menyatakan ketiadaan ciri-ciri tidak menukar persamaan antara dua sebatian. Pekali yang popular adalah Pekali Tanimoto yang diguna secara meluas dengan *fingerprint* 2D. Pekali Tanimoto (T_c) diberi oleh formula berikut:

$$S_{A,B} = \frac{c}{a+b-c}$$

Nilai T_c adalah antara 0 hingga 1. Oleh itu, adalah mudah untuk menukarnya kepada ukuran ketidaksamaan, iaitu jarak Soergel, dan sebaliknya. Pekali gabungan lain yang popular adalah seperti pekali Dice, Kosinus, Fossum, Rusell-Rao dan Forbes.

Pekali korelasi (*correlation coefficients*) mengukur kekadaran dan kebebasan antara dua sebatian. Gasteiger dan Engel (2003) menyifatkan pekali korelasi sebagai satu ukuran komposit yang mengguna kedua-dua ukuran persamaan dan perbezaan terhadap perbandingan *fingerprints*, contohnya pekali Hamann, Yule dan Stiles.

Pekali kebarangkalian termasuklah statistik maklumat (iaitu taburan kekerapan pemerihal) yang mengukur kehomogenan sebatian melalui pemetaaan pangkalan data (Sneath & Sokal, 1973). Adamson dan Bush (1975) mendapati prestasi pekali ini adalah kurang memuaskan dalam penggunaan data struktur kimia. Pekali korelasi dan kebarangkalian ini jarang diguna dalam carian persamaan.

PENILAIAN PERSAMAAN

Ukuran prestasi sesuatu sistem carian persamaan kimia boleh ditentu daripada prestasi sistem dalam perolehan semula pangkalan data. Oleh itu, ia boleh dinilai berdasarkan dua asas, iaitu kecekapan berdasarkan penggunaan optimum sesuatu sumber dan keberkesanan yang menilai kebolehan sesuatu carian memenuhi keperluan pengguna.

Kecekapan seperti yang dinyatakan oleh (Gasteiger dan Engel, 2003) diukur dari jumlah masa yang diambil untuk suatu carian memperoleh penyelesaian dan jumlah memori komputer yang diperlu untuk melaksana carian tersebut. Namun, keberkesanan adalah prinsip yang kerap diguna dalam susastera untuk menilai prestasi sesuatu carian persamaan.
Willett (2005) mengkategori penilaian keberkesanan kepada tiga kelas, iaitu ukuran prestasi grafik, ukuran prestasi berangka dan data kuantitatif biologi. Dua kelas pertama hanya melibatkan aktiviti pembolehubah binari (iaitu aktif atau tidak aktif) dan kelas ketiga melibatkan pembolehubah yang bernilai sebenar (Willett, 2004). Artikel ini menghad skop kepada perbincangan yang melibatkan ukuran pembolehubah binari, sejajar dengan domain carian berasaskan *fingerprint* 2D.

Ukuran prestasi secara grafik termasuklah plot dapatan-kejituan dan plot dapatan kumulatif. Plot dapatan-kejituan mencarta ketepatan sesuatu carian terhadap dapatannya. Edgar et al. (2000) mendapati puncak yang ditanda pada plot menandakan kewujudan sebilangan besar sebatian aktif yang sama. Plot dapatan kumulatif sering diguna dalam carian pangkalan data yang biasa, yang mana dapatan bilangan sebatian yang diperoleh dalam carian yang dikenali sebagai pangkatan daftar kena (*hit-list*) diplot.

Terdapat beberapa ukuran prestasi berangka yang diuji oleh Edgar et al. (2000). Bagaimanapun, terdapat hanya dua ukuran yang sering diguna dalam susastera, iaitu faktor pengayaan dan skor G-H. Faktor pengayaan adalah nisbah bilangan aktif yang diperoleh relatif kepada bilangan yang diperoleh jika sebatian dipilih secara rawak dari pangkalan data. Ia diguna oleh Waszkowycz et al. (2001) untuk pemilihan ligan berskor tinggi untuk reseptor agonis estrogen. Dalam kajian ini, objektif utama kumpulan penyelidik tersebut adalah untuk mencari molekul yang sesuai dengan ikatan yang kukuh (i.e ligan) terhadap atom terpusat (i.e reseptor agonis estrogen) bertujuan membentuk satu kompleks koordinasi. Formula faktor pengayaan ini diberikan seperti berikut, dengan n adalah jiran terdekat dalam suatu daftar kena hasil carian, a merupakan sebatian aktif daripada n dan subset kepada A, manakala A adalah semua sebatian aktif dalam pangkalan data yang mempunyai N sebatian :

$$\frac{a/n}{A/N}$$

Skor G-H adalah purata pemberat bagi dapatan dan kejituan. Ia diwujud pada asalnya untuk menilai keberkesanan carian pangkalan data 3D. Bagaimanapun, Raymond dan Willett (2002) berjaya menggunanya untuk menilai keberkesanan graf 2D dan carian persamaan kimia berasaskan *fimgerprint*. Skor G-H ditakrif sebagai:

$$\frac{\alpha P + \beta R}{2}$$

dengan α and β adalah pemberat yang menerangkan kepentingan relatif dapatan (R) dan kejituan (P). Bagaimanapun, skor G-H menjadi kurang popular sejak kebelakangan ini.

Banyak metrik yang diwujud untuk menilai prestasi kaedah berpangkat dalam penyaringan maya (*virtual screening*). Ini termasuk metrik seperti *Receiver Operator Characteristics* (ROC) dan variannya. ROC adalah plot grafik kepada pecahan positif tulen (*true positives*) terhadap pecahan positif palsu pada keseluruhan kedudukan pangkatan. Setiap kaedah carian mempunyai plot lengkungan ROC yang diguna untuk membeza prestasi kaedah carian. Kaedah carian yang mempunyai lengkungan ROC menguasai lengkungan ROC kaedah lain dianggap lebih baik dalam perolehan sebatian aktif (Truchon dan Bayly, 2007). Bagaimanapun, kemampuan suatu kaedah carian untuk mengenal pasti sebatian aktif secepat mungkin (pengecaman awal) dengan menyenarai sebatian aktif pada kedudukan awal pangkatan adalah lebih penting berbanding dengan kemampuannya mendapat lebih banyak sebatian aktif tanpa mengira kedudukannya dalam pangkatan. Oleh itu, metrik yang diguna untuk menilai kaedah berpangkat ini juga perlu termasuk dalam kriteria ini. Atas alasan ini, Truchon dan Bayly (2007) mendakwa ROC tidak sensitif terhadap pengecaman awal. Justeru, metrik baharu yang merupakan generalisasi ROC, dikenali sebagai *Boltzmann-Enhanced Discrimination of ROC* (BEDROC) dicadang. Kajian Truchon dan Bayly kemudiannya diperlua bagi membanding tujuh metrik termasuk ROC, *Area Under Accumulation Curve* (AUAC), purata pangkat aktif, faktor pengayaan (EF), pemberat AUAC (wAUAC), *Robust Initial Enhancemet* (RIE), dan BEDROC. Butiran perbandingan ini boleh didapati daripada Truchon dan Bayly (2007).

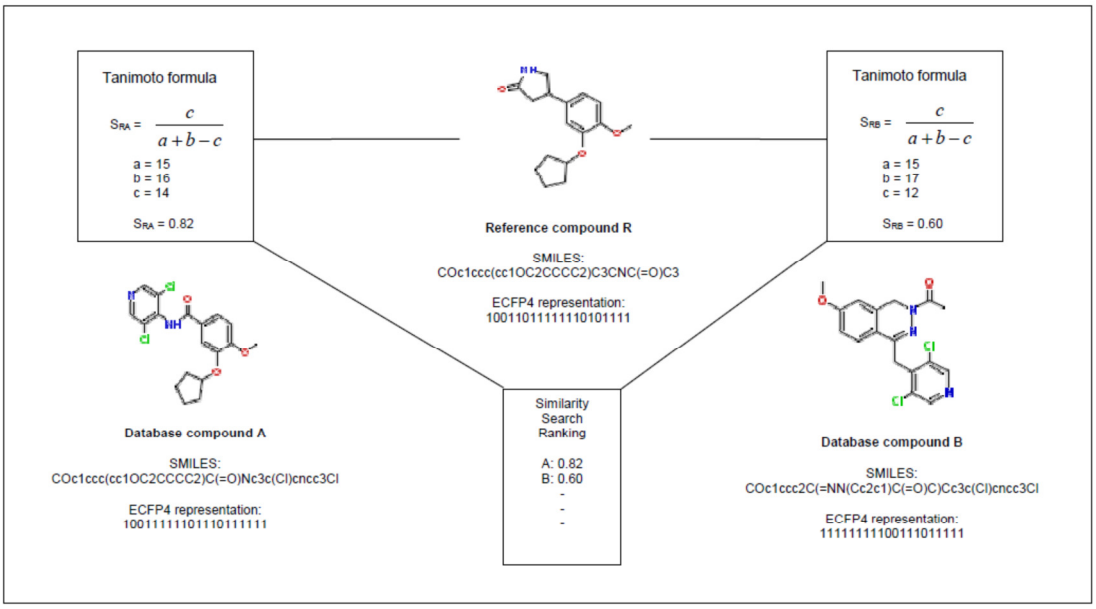
Clark dan Webster-Clark (2008) memperluas konsep wAUAC dengan memperkenal dua skema pemberat yang dipanggil aritmetik dan harmonik. Skema pemberat aritmetik menetapkan berat untuk setiap aktif berdasarkan saiz kelompok asalnya, manakala pemberat harmonik menetapkan berat untuk setiap aktif berdasarkan kedudukan sebatian aktif tersebut dalam kelasnya. Semua skema pemberat harmonik didapati berkeupayaan menangani isu pengecaman awal yang merupakan kepentingan utama dalam penyaringan maya. Skim ini memperoleh sebatian aktif daripada setiap kelas aktiviti pada awal lengkungan pemulihan (*recovery curve*).

Selain daripada ukuran yang disebut di atas, masih terdapat hujah mengenai pengukuran 'kebaikan' sesuatu kaedah persamaan. Satu pendekatan seperti yang dicadang oleh Sheridan (2007) ialah penggunaan ‘radius kejiranan’ (*neighbourhood radius*) yang diperkenal oleh Patterson et al. (1996) untuk mengukur tahap kemungkinan sebatian yang dikumpul mempunyai aktiviti yang sama. Selain daripada itu, konsep lompatan perancah (*scaffold hopping*) juga boleh diguna. Ia menunjukkan keupayaan suatu kaedah carian bagi mengenal pasti struktur sebatian aktif yang baharu yang merupakan pengubahsuaian struktur teras pusat sebatian aktif yang diketahui, dikenali sebagai sebatian pertanyaan (*query compound*). Pada masa yang sama, ia juga mengekal ciri-ciri penting yang diperlu untuk pengikatan (Bohm et al., 2004). Rasional di sebalik penilaian ini adalah perancah yang berbeza (iaitu struktur teras sebatian) mampu menawarkan pilihan dari segi ketercapaian kimia dan prospek untuk pengoptimuman *Lead* (Renner dan Schneider, 2006; Schneider et al., 2006). *Lead* merupakan sebatian kimia yang berpotensi tinggi menjadi drug yang boleh dipasar kepada pengguna. Bohm et al. (2004) juga menyokong konsep lompatan perancah dengan menyatakan bahawa analisis terhadap drug sedia ada menunjukkan terdapat kemungkinan untuk mencari satu set sebatian dengan pelbagai struktur yang terikat kepada reseptor yang sama.

Oleh kerana carian persamaan adalah berasaskan Prinsip Sifat Serupa, terdapat percanggahan antara konsep lompatan perancah dan carian persamaan. Lompatan perancah adalah satu konsep yang mencadang sebatian dengan struktur yang berbeza boleh memiliki sifat biologi yang serupa (iaitu kesamaan fungsi tidak semestinya memerlukan persamaan

struktur (Schneider et al, 2006)), manakala Prinsip Sifat Serupa menyatakan sebatian dengan sifat biologi yang serupa juga mempunyai struktur yang serupa. Brown dan Jacoby (2006) mengkaji semula lompatan perancah secara menyeluruh dan menghasil suatu matriks yang memisah carian persamaan dengan lompatan perancah. Beberapa ukuran objektif bagi lompatan perancah disenarai termasuk: kadar kena *chemotype (perancah)*, kiraan Bemis dan Perancah Murcko, kekunci MEQI, Min Persamaan Berpasangan (*Mean Pairwise Similarity - MPS*) dan julat *Tanimoto*. Walaupun carian persamaan dan lompatan perancah adalah berbeza, ukuran ini boleh diguna untuk menilai kemungkinan suatu aljabbar carian persamaan yang dapat melompat perancah. Sebagai contoh, jika MPS bagi satu set pangkatan algoritma carian persamaan adalah rendah (bermakna struktur sebatian dalam pangkatan secara puratanya adalah berbeza), tetapi terdapat tanda kewujudan lompatan perancah, ini juga menunjukkan aljabbar tersebut mempunyai merit tambahan iaitu keupayaan untuk melompat perancah.

Proses carian persamaan mengguna pemerihal struktur *fingerprint* ECFP4 dan metrik carian persamaan *Tanimoto* dalam Rajah 1 menunjukkan bagaimana komponen utama ini bekerja. Struktur yang diguna adalah sebatian sebenar yang diekstrak daripada keluarga *Phosphodiesterase* dalam pangkalan data '*World of Molecular Bioactivity*'. Kedua-dua perwakilan dijana mengguna perisian Pipeline Pilot yang dibinkan pada 20 bin (bagaimanapun secara realitinya kedua-duanya dibinkan sama ada kepada 1024 atau 2048 bin).



RAJAH 1. Illustrasi carian persamaan mengguna ECFP4 dan *Tanimoto* (Malim, 2011).

PENUTUP

Kertas ini membincang definisi setiap ukuran persamaan dan menerangkan bagaimana ukuran ini dilaksana bersama untuk memasti keberkesanan dan kemantapan carian persamaan berasaskan *fingerprint*. Komponen utama dalam ukuran carian persamaan berasaskan *fingerprint* adalah Prinsip Sifat Serupa yang mendokong konsep carian persamaan; *fingerprints* yang merupakan perwakilan struktur diguna untuk menyifat sebatian; pekali persamaan yang mengukur darjah perhubungan antara sebatian berdasarkan pemerihal molekul yang mewakili mereka (dalam kes ini, *fingerprints*); dan pengoptimuman ukuran

persamaan seperti penggunaan skema pemberat. Bagaimanapun, konsep pengoptimunan ukuran persamaan tidak dibincang.

Setiap ukuran persamaan memainkan peranan yang besar dalam prestasi carian persamaan berasaskan *fingerprint*. Sebagai contoh, *fingerprints* mengekod pelbagai jenis kandungan maklumat berdasarkan reka bentuknya. Oleh yang demikian, prestasi yang ditunjukkan oleh penggunaan pelbagai jenis *fingerprint* adalah berbeza, walaupun kadang kala hanya perbezaan yang kecil dapat diperhati.

RUJUKAN

Adamson, G. W. & J. A. Bush. 1975. A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures. *Journal of Chemical Information and Computer Sciences*, 15: 55 - 58.

Bohm, H.J., Flohr, A. & Stahl, M. 2004. Scaffold Hopping. *Drug Discovery Today: Technologies,* 1: 217 - 224.

Bremser, W. 1978. Hose - a Novel Substructure Code. *Analytica Chimica Acta*, 103: 355 - 365.

Brown, N. & Jacoby, E. 2006. On Scaffolds and Hopping in Medicinal Chemistry. *Medicinal Chemistry,* 6: 1217 - 1229.

Clark, R. D. & Webster-Clark, D. J. 2008. Managing Bias in ROC Curves. *Journal of Computer-Aided Molecular Design,* 22: 141 - 146.

Edgar, S. J., Holliday, J. D. & Willett, P. 2000. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *Journal of Molecular Graphics and Modelling,* 18: 343 - 357.

Gasteiger, J. & T. Engel. 2003. *Chemoinformatics a Textbook*, New York: Wiley-VCH.

Hassan, M., R. D. Brown, S. Varma-O'Brien & D. Rogers. 2006. Cheminformatics Analysis and Learning in a Data Pipelining Environment. *Molecular Diversity*, 10: 283 - 299.

Hert, J., P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby & A. Schuffenhauer 2004. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Journal of Organic & Biomolecular Chemistry*, 2: 3256 - 3266.

Johnson, M. A. & G. M. Maggiora. 1990. *Concepts and Applications of Molecular Similarity*, New York: John Wiley & Sons Inc.

Jonsdottir, S. O., Jorgensen, F. S. & Brunak, S. 2005. Prediction Methods and Databases within Chemoinformatics: Emphasis on Drugs and Drug Candidates. *Bioinformatics,* 21: 2145 - 2160.

Lambert, R. 2000. An Introduction to Derwent World Drug Index. *EuroMug 2000.* Daylight Chemical Information System.

Leach, A. R. & V. J. Gillet. 2003. *An Introduction to Chemoinformatics*, Germany: Kluwer Academic Publishers.

Martin, Y. C., J. L. Kofron & L. M. Traphagen. 2002. Do Structurally Similar Molecules Have Similar Biological Activity? *Journal of Medicinal Chemistry*, 45: 4350-4358.

Patterson, D. E., R. D. Cramer, A. M. Ferguson, R. D. Clark & L. E. Weinberger. 1996. Neighborhood Behavior: A Useful Concept for Validation Of "Molecular Diversity" Descriptor. *Journal of Medicinal Chemistry*, 39: 3049 - 3059.

Raymond, J. W. & Willett, P. 2002. Effectiveness of Graph-based and Fingerprint-based Similarity Measures for Virtual Screening of 2D Chemical Structure databases. *Journal of Computer-Aided Molecular Design*, 16: 59 - 71.

Renner, S. & Schneider, G. 2006. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem,* 1: 181 - 185.

Salim, N., J. Holliday & P. Willett. 2003. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *Journal of Chemical Information and Computer Sciences*, 43: 435 - 442.

Schneider, G., Schneider, P. & Renner, S. 2006. Scaffold-Hopping: How Far Can You Jump? *QSAR & Combinatorial Science,* 25: 1162 - 1171.

Sheridan, R. P. 2007. Chemical Similarity Searches: When Is Complexity Justified? *Expert Opinion in Drug Discovery,* 2: 423 - 430.

Sneath, P. H. A. & R. R. Sokal. 1973. *Numerical Taxonomy*, London: W. H. Freeman and Company.

Southan, C., Varkonyi, P. & Muresan, S. 2009. Quantitative Assessments of the Expanding Complementarity between Public and Commercial Databases of Bioactive Compounds. *Journal of Chemoinformatics,* 1*:* 10.

Sunset-Molecular. 2011. WOrld of Molecular BioAcTivity. Sunset Molecular Discovery LLC. http://www.sunsetmolecular.com/index.php [1 Septemeber 2011].
Truchon, J.-F. & Bayly, C. I. 2007. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling,* 47: 488 - 508.
Waszkowycz, B., Perkins, T. D. J., Sykes, R. A. & Li, J. 2001. Large-scale Virtual Screening for Discovering Leads in the Postgenomics Era. IBM System Journal, 40: 360-376.
Willett, P. 2005. The Evaluation of Retrieval Effectiveness in Chemical Database Searching. In Tait, J. I. (Ed.) *Charting a New Course: Natural Language Processing and Information Retrieval.* Germany: Springer-Verlag.
Willett, P. 2004. The Evaluation of Molecular Similarity and Molecular Diversity Methods using Biological Activity Data. *Methods in Molecular Biology,* 275: 51 - 63.
Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. 2006. DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucleic Acids Research,* 34: 668 - 672.
Willett, P. 2007. Similarity Methods in Chemoinformatics. *Annual Review of Information Science and Technology,* 42.
Willett, P., J. M. Barnard & G. M. Downs. 1998. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences*, 38: 983 - 996.

Shereena M. Arif
Pusat Pengajian Sains Maklumat,
Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia,
43600 Bangi, Selangor, Malaysia
shereen@ftsm.ukm.my

Nurul Malim
Pusat Pengajian Sains Komputer,
Universiti Sains Malaysia,
11800 Pulau Pinang, Malaysia
nurulhashimah@cs.usm.my