https://doi.org/10.17576/apjitm-2014-0302-05

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 3 No. 2, December 2014: 63 - 70
e-ISSN: 2289-2192

PRESTASI JEJAK JARI 2D DALAM PANGKALAN DATA KIMIA PUBCHEM

SHEREENA M. ARIF
FATIMAH ZAWANI ABDULLAH
NURUL HASHIMAH AHAMED HASSAIN MALIM

ABSTRAK

Kajian carian persamaan molekul dalam pangkalan data kimia semakin berkembang terutama dalam bidang penemuan kimia. Carian persamaan merupakan satu aplikasi dalam bidang kimia informatik untuk mengukur persamaan antara struktur molekul yang dikenali sebagai kueri dan struktur sebatian kimia dalam pangkalan data. Carian persamaan juga merupakan salah satu pendekatan dalam penyaringan maya yang melibatkan teknik komputeran dan penskoran kebarangkalian aktiviti. Objektif utama kajian adalah untuk menentu jejak jari (fingerprint) terbaik dalam kalangan enam jejak jari yang dipilih dalam kajian ini mengguna pangkalan data kimia PubChem. Kertas ini membincang proses carian persamaan yang dijalankan mengguna 6 jenis pemerihal iaitu ECFP4, ECFC4, FCFP4, FCFC4, SRECFC4 dan SRFCFC4 ke atas 15 kelas aktiviti dalam set data PubChem mengguna pekali persamaan Tanimoto untuk mengira persamaan antara struktur kueri dengan setiap struktur pangkalan data. Hasil kajian mendapati ECFP4 menunjukkan prestasi yang terbaik untuk diguna dengan pekali persamaan Tanimoto dalam set data PubChem.
Kata kunci: Jejak jari 2D, Tanimoto, PubChem, Carian Persamaan dan Kimia Informatik.

ABSTRACT

The study of molecular similarity search in chemical database is increasingly widespread, especially in the area of drug discovery. Similarity search is an application in the field of Chemoinformatics to measure the similarity between the molecular structure which is known as the query and the structure of chemical compounds in the database. Similarity search is also one of the approaches in virtual screening which involves computational techniques and scoring the probabilities of activity. The main objective of this work is to determine the best fingerprint among six fingerprints selected in this study using PubChem chemical dataset. This paper will discuss the similarity searching process conducted using 6 types of descriptors, which are ECFP4, ECFC4, FCFP4, FCFC4, SRECFC4 and SRFCFC4 on 15 activity classes of PubChem dataset using Tanimoto coefficient to calculate the similarity between the query structures and each of the database structure. The results suggest that ECFP4 performs the best to be used with Tanimoto coefficient in the PubChem dataset.
Keywords: 2D Fingerprints, Tanimoto, PubChem, Similarity Searching and Chemoinformatics.

PENGENALAN

Kimia informatik merujuk kepada aplikasi kaedah komputeran untuk membincang permasalahan kimia dengan memberi penekanan terhadap manipulasi maklumat struktur kimia (Leach & Gillet, 2007). Malah, kimia informatik adalah satu bidang baharu yang muncul daripada beberapa bidang sebelum ini seperti kimia komputeran, *Chemometrics* dan maklumat kimia (Xu & Hagler, 2002). Perbezaan antara pemprosesan data kimia dengan pemprosesan data yang lain adalah keperluan kepada struktur kimia dalam menghasilkan hasil akhir. Keperluan ini membawa kepada pengenalan pendekatan khas untuk mewakil, menyimpan dan mencapai semula makumat dalam sistem komputer. Terdapat banyak perkakasan dan teknik telah diguna dalam bidang ini yang mempunyai pendekatan atau kaedah, kekuatan dan batasan tersendiri.

Penyaringan maya melibatkan penggunaan teknik komputeran dalam pemilihan molekul untuk proses-proses yang terlibat dalam pembangunan dadah seperti pemilihan lead, penyaringan hits dan ujian farmakokinetik bagi menentukan tindak balas terapeutik (Willett, 2011). Terdapat banyak kaedah penyaringan maya yang berbeza, yang menyusun aktiviti kebarangkalian struktur pangkalan data dalam urutan menurun seterusnya menganggap struktur yang berada di atas mempunyai aktiviti kebarangkalian yang tinggi. Carian persamaan merupakan salah satu pendekatan penyaringan maya yang diguna secara meluas dalam membanding molekul aktif (juga dikenali sebagai struktur rujukan) dengan setiap molekul dalam pangkalan data (Hert et al., 2004). Pengukuran persamaan struktur dikira, seterusnya persamaan skor molekul pangkalan data disusun secara menurun. Molekul yang tersusun dalam senarai pangkatan teratas, yang dirujuk sebagai jiran terhampir kepada struktur rujukan kemudiannya dipapar kepada pengguna sebagai hasil daripada carian persamaan.

Pengukuran persamaan kimia terdiri daripada tiga komponen iaitu pemerihal untuk mewakilkan struktur; skema pemberat untuk menetapkan berat terhadap bahagian perwakilan yang berbeza yang mempengaruhi darjah kepentingan; dan pekali persamaan untuk mengukur tahap persamaan antara kedua-dua perwakilan. Terdapat beberapa teknik untuk mewakil dan mengekod struktur molekul kimia (Willett, 2011). Tiga kelas umum pemerihal adalah pemerihal keseluruhan molekul (dikenali sebagai 1D); pemerihal yang diukur daripada molekul perwakilan 2D dan pemerihal yang diukur daripada perwakilan 3D. Kertas ini hanya melibatkan molekul perwakilan 2D.

Perbandingan ciri molekul melibatkan pelbagai teknik seperti binari atau bukan binari (Khalifa et al., 2009). Kebanyakan jejak jari adalah binari iaitu setiap bit diwakili dengan nilai 1 atau 0 berdasarkan kewujudannya. Bagaimanapun, terdapat beberapa jejak jari yang diguna secara meluas yang mengambil kira kekerapan dalam molekul yang dikenali sebagai hologram seperti ECFC dan FCFC (Arif et al., 2009). Bilangan dan variasinya dikenali sebagai skema pemberat. Jejak jari ini juga boleh dilanjutkan sama ada kepada logaritme asli atau punca kuasa dua kepada nilai bilangan untuk setiap bit dalam molekul. Jejak jari ini dicadang untuk memaksimum keberkesanan penyaringan maya dan boleh memberi prestasi yang baik dalam penyaringan.

Aspek lain dalam pengukuran persamaan adalah pekali persamaan yang boleh dikelas kepada tiga kumpulan umum. Tiga kumpulan tersebut ialah pekali gabungan, pekali korelasi dan pekali jarak (Salim et al., 2002). Kajian ini mengguna pekali Tanimoto iaitu salah satu pekali gabungan yang diguna secara meluas dalam carian persamaan.

Pangkalan data kimia terdiri daripada pangkalan data komersial dan pangkalan data akses terbuka. Pangkalan data komersial mengenakan bayaran dan langganan daripada pengguna manakala pangkalan data akses terbuka menyedia akses dan muat turun data secara percuma kepada pengguna. Kajian pengukuran persamaan kurang dijalankan terhadap pangkalan data akses terbuka. Justeru, kajian ini dijalankan untuk menguji prestasi jejak jari dan pekali persamaan terpilih mengguna pangkalan data akses terbuka iaitu pangkalan data PubChem.

KAJIAN LEPAS

Satu kajian dijalankan oleh Shuib et al. (2013) untuk menentu pekali terbaik yang diguna dalam carian persamaan bagi memperoleh keputusan yang optimum. Kajian tersebut membanding tiga pekali persamaan iaitu Tanimoto, Russell/Rao dan Euclidean yang memfokus kepada jejak jari ECFP4 dan UNITY dengan mengguna 5 kelas aktiviti dalam set data MDDR. Hasil kajian mendapati pekali Tanimoto harus diguna dalam penyaringan maya untuk mendapat keputusan yang optimum. Selain daripada itu, pekali Tanimoto juga menunjukkan prestasi yang baik

dalam pelbagai bidang termasuk bidang carian persamaan. Justeru, kajian ini memilih untuk mengguna pekali Tanimoto untuk penilaian persamaan.

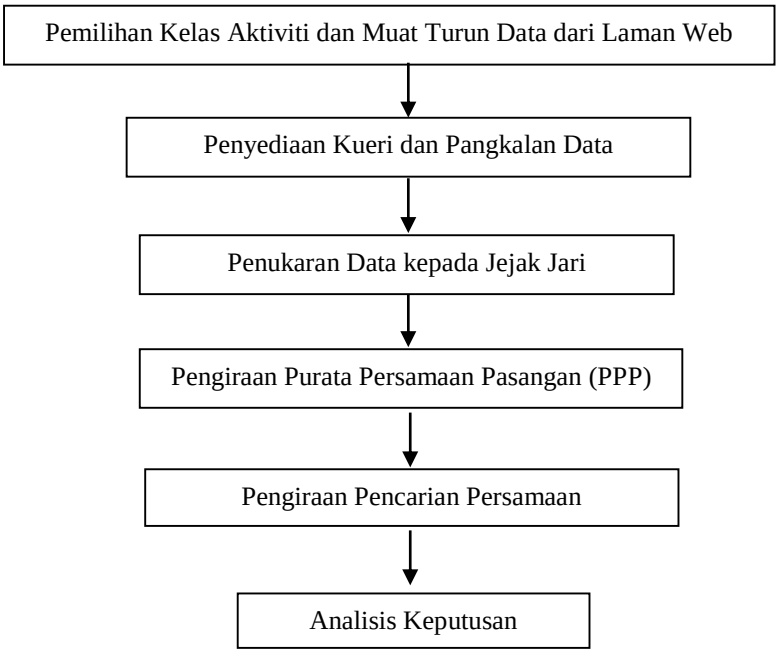
Banyak perbandingan dijalankan terhadap prestasi jejak jari dalam bidang carian persamaan. Kajian Abdo et al. (2010) melibatkan jejak jari yang ditetapkan pemberat mengguna set data MDDR dan WOMBAT serta mencadang penggunaan punca kuasa dua kepada fragmen kekerapan kasar akan memberi prestasi yang baik dalam penyaringan maya.

Kajian Todeschini et al. (2012) pula mengguna set data MUV. Set data tersebut mengandungi 17 kelas aktiviti yang diambil daripada pangkalan data PubChem. Bagaimanapun, setiap kelas hanya mengandungi 30 struktur aktif. Kajian ini hanya melibatkan jejak jari ECFP4. Sama seperti kajian Abdo et al. (2010), Todeschini et al. mengguna set data MUV tetapi terhad kepada dua jejak jari sahaja iaitu ECFC4 dan SRECFC4.

Beberapa kajian mengguna punca kuasa dua kepada bilangan kasar tetapi tidak melibatkan pangkalan data PubChem dan ada juga yang mengguna set data MUV dengan anggapan set data MUV adalah berkaitan dengan pangkalan data PubChem tetapi tidak merangkumi pelbagai jenis pemerihal. Kajian ini mengguna 6 jenis pemerihal untuk menguji prestasi masing-masing mengguna set data PubChem.

METOD

Kajian ini mengguna pangkalan data kimia terbuka iaitu PubChem (diperoleh daripada http://pubchem.ncbi.nlm.nih.gov). Pangkalan data kimia PubChem merupakan pangkalan data eksperimen yang menyedia platform akses secara terbuka untuk kajian mengenai aktiviti biologi molekul kecil (Bolton et al., 2008). Rajah 1 menunjukkan aliran dan langkah yang dijalankan dalam kajian ini.



RAJAH 1. Carta aliran kajian

Kajian ini melibatkan 15 kelas aktiviti daripada pangkalan data kimia PubChem yang dipilih daripada kajian lepas. Han et al. (2008) mengguna AID612, AID567, AID372, AID565 untuk membeza sebatian bioaktiviti manakala Weis et al. (2008) mengguna AID798, AID846 untuk meramal aktiviti sebatian mengguna pengkelas Mesin Vektor Sokongan (MVS). Kelas aktiviti yang lain diambil daripada kajian untuk menentu kualiti ramalan (Chen et al., 2010).

Jadual 1 menunjukkan maklumat terperinci mengenai kelas aktiviti yang diguna dalam kajian ini.

JADUAL 1. Senarai kelas aktiviti yang diguna dalam kajian

Kelas Aktiviti	Bilangan Molekul Aktif	Purata Persamaan Pasangan	Maklumat Kelas Aktiviti
AID1	2104	0.1103	Anti-Kanser
AID53	1737	0.1133	Anti-Kanser
AID129	1659	0.1119	Anti-Kanser
AID248	1559	0.1031	Anti-Tumor
AID573	1420	0.1336	Antibiotik Baharu
AID565	1250	0.1270	Anti-HIV
AID256	983	0.0957	Anti-Tumor
AID372	769	0.1432	Perencatan Ribonuclease H
AID612	416	0.1326	Anti-Murung
AID567	366	0.1337	Anti-Murung
AID798	302	0.1381	Terapi Antikoagulan Baharu
AID376	250	0.1056	Pengutuban Semula Sel Jantung
AID823	137	0.1336	Anti-tumor
AID846	91	0.1766	Terapi Antikoagulan Baharu
AID607	34	0.1351	Pengawalan Kepekatan Sel Nukleotida

Enam jejak jari dipilih untuk diguna dalam kajian ini iaitu ECFP4, FCFP4, FCFC4, ECFC4, SRECFC4 dan SRFCFC4. Semua jejak jari ini mempunyai perwakilan yang berbeza iaitu dalam bentuk binari atau bukan binari. Jenis perwakilan struktur yang berbeza memberi kesan terhadap keseluruhan tahap persamaan. Kajian Arif et al. (2009) menunjukkan kelebihan pemberat berdasar kekerapan (dikenali sebagai hologram seperti ECFC4 dan FCFC4). Kajian tersebut mencadang pengekodan kekerapan kasar dengan mengguna punca kuasa dua terhadap kedua-dua fragmen struktur rujukan dan pangkalan data, dikenali sebagai SRECFC4. Justeru, kajian ini melibatkan jejak jari SRECFC4 dan SRFCFC4 untuk diuji.

Terdapat dua tugasan utama yang perlu dijalankan dalam uji kaji ini. Tugasan pertama ialah mengira Purata Persamaan Pasangan (PPP). PPP diguna untuk menentu persamaan antara setiap molekul dalam kelas aktiviti menggunakan pekali Tanimoto. Willett (2011) menyatakan pekali Tanimoto daripada kumpulan pekali gabungan adalah pekali terbaik dan menunjukkan prestasi yang baik dalam pelbagai bidang termasuk carian persamaan. Lantaran itu, pekali Tanimoto diguna dalam uji kaji ini untuk mengira PPP. Nilai PPP yang tinggi menunjukkan persamaan yang tinggi antara setiap pasangan molekul dalam kelas aktiviti, yang dikenali sebagai kelas homogen. Nilai PPP yang rendah dikenali sebagai heterogen.

Pekali Tanimoto untuk data binari dan bukan binari masing-masing ditakrif mengguna rumus PAB iaitu persamaan antara molekul A (struktur rujukan) dengan molekul B (struktur pangkalan data) seperti (1) dan (2) berikut:

$$\mathrm{PAB} = \frac{c}{a+b-c} \qquad (1)$$

$$\mathrm{PAB} = \frac{\sum_{i=1}^{N} x_i A x_i B}{\sum_{i=1}^{N} (x_i A)^2 + \sum_{i=1}^{N} (x_i B)^2 - \sum_{i=1}^{N} x_i A x_i B} \qquad (2)$$

iaitu:

$a = \sum_{i=1}^{N} (x_i A)$: Jumlah bit yang "on" dalam struktur rujukan

$b = \sum_{i=1}^{N} (x_i B)$: Jumlah bit yang "*on*" dalam struktur pangkalan data

$c = \sum_{i=1}^{N} x_i A x_i B$: Jumlah bit yang "*on*" dalam struktur rujukan dan pangkalan data

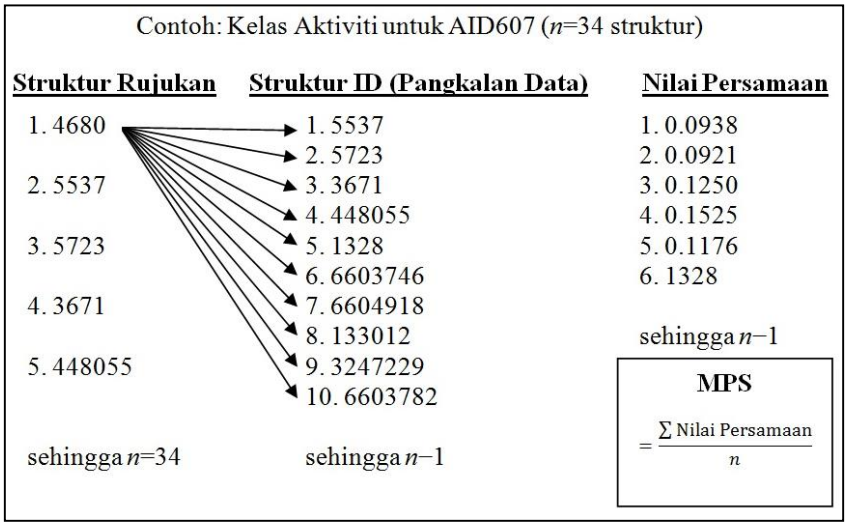
Mengguna data binari dengan bilangan bit sebanyak 10 sebagai contoh, Molekul A dan Molekul B ditakrifkan seperti berikut:

Molekul A: 1 1 1 0 1 1 0 1 1 0

Molekul B: 1 0 0 1 0 1 1 0 1 0

Jadi, nilai *a* adalah 7, *b* adalah 5 dan *c* adalah 3.

Rajah 1 menunjukkan aliran pengiraan PPP. Algoritme membanding persamaan bit binari daripada setiap struktur dengan struktur yang lain menggunakan pekali persamaan Tanimoto. Contoh ini menunjukkan bahawa hanya jejak jari ECFP4 akan diguna dalam uji kaji ini.



RAJAH 2. Aliran pengiraan purata persamaan pasangan (PPP)

Tugasan kedua adalah mengira purata dapatan semula dan Ukuran F. Dalam tugasan ini, langkah mengira nilai PPP diulang tetapi kelas aktiviti diganti dengan semua struktur dalam pangkalan data (semua struktur duplikasi disingkir). Langkah uji kaji diringkaskan seperti berikut:

PENYEDIAAN PEMERIHAL

Kajian ini melibatkan 6 pemerihal (ECFP4, ECFC4, FCFP4, FCFC4, SRECFC4, dan SRFCFC4) untuk mewakil struktur kimia dan pekali Tanimoto untuk mengira persamaan antara dua struktur kimia. Pemerihal melibatkan dua jenis jejak jari (ECFP4, FCFP4) dan dua jenis skema pemberat, masing-masing adalah bilangan kasar (ECFC4, FCFC4) dan punca kuasa dua kepada bilangan kasar (SRECFC4, SRFCFC4). Kajian ini mengguna 6 jenis pemerihal iaitu ECFP4, ECFC4, FCFP4, FCFC4, SRECFC4 dan SRFCFC4.

Terdapat 9312 struktur sebatian dalam set data ini, gabungan daripada 15 kelas aktiviti yang dipilih dan semua struktur duplikasi disingkir. Untuk struktur rujukan, 10 struktur dipilih secara rawak daripada setiap kelas aktiviti. Struktur rujukan dan struktur pangkalan data ini ditukar kepada 6 pemerihal mengguna perisian Pipeline Pilot (diperoleh daripada Accelrys Software Inc. di http://www.accelrys.com). Pada akhir langkah ini, carian persamaan mengguna pekali Tanimoto untuk 6 pemerihal, setiap pemerihal mewakili 10 struktur kueri dan 9312 struktur pangkalan data daripada 15 kelas aktiviti. Justeru, kajian carian persamaan ini melibatkan pengiraan sebanyak 8,380,800 (iaitu 6 jejak jari × 10 kueri × 9312 struktur × 15 kelas aktiviti).

Ukuran Ukuran

$$\text{Purata Dapatan Semula} = \frac{\text{Bilangan Positif Benar}}{\text{Bilangan Molekul Aktif dalam Kelas Aktiviti}}$$

$$\text{Purata Kejituan} = \frac{\text{Bilangan Positif Benar}}{\text{Bilangan Struktur yang Dicapai}}$$

$$\text{Ukuran F} = \frac{2\ (\text{Purata Dapatan Semula} * \text{Purata Kejituan})}{(\text{Purata Dapatan Semula} + \text{Purata Kejituan})}$$

ANALISIS

PPP, purata dapatan semula dan Ukuran F dikira untuk dianalisis. Berdasarkan Jadual 1, AID846 mencatat nilai PPP tertinggi dan menunjukkan kebanyakan struktur dalam kelas aktiviti adalah hampir sama antara satu dengan yang lain (homogen) berbanding 14 kelas aktiviti yang lain. Nilai PPP terendah adalah daripada kelas aktiviti AID256. Keputusan ini juga menunjukkan nilai PPP untuk set data PubChem adalah rendah iaitu kurang daripada 0.18 berbanding set data lain seperti MDDR dan WOMBAT (Abdo et al., 2010; Todeschini et al., 2012). Ini menunjukkan PubChem merupakan set data yang mencabar untuk pengujian carian persamaan dan bidang berkaitan yang lain.

Jadual 2 menunjukkan keputusan pengiraan purata dapatan semula dan Ukuran F. Purata dapatan semula tertinggi dalam kalangan kelas aktiviti ditanda dengan huruf tebal dan diwarna gelap manakala purata dapatan semula terendah diwarnakan kurang gelap. AID846 kerap mempunyai nilai tertinggi kecuali untuk jejak jari SRECFC4 yang mana nilai tertinggi terbanyak adalah AID607. Untuk nilai terendah, AID1 kerap menunjukkan nilai terendah dalam semua kelas aktiviti kecuali jejak jari FCFC4 dan SRFCFC4. Analisis ini boleh dikaitkan dengan saiz kelas aktivti. AID846 mempunyai saiz kelas aktiviti yang kecil dan menunjukkan nilai purata dapatan semula yang tinggi. Manakala, AID1 mempunyai saiz kelas aktiviti yang besar dan menunjukkan nilai purata dapatan semula yang rendah. Justeru, saiz kelas aktiviti menyumbang kepada keupayaan mencari struktur yang lebih relevan dan sama.

Pengiraan Ukuran F pula melibatkan pengiraan kekerapan mempunyai nilai tertinggi dalam kalangan jejak jari. Nilai tertinggi Ukuran F ditanda dengan huruf tebal dan diwarnakan gelap. Berdasarkan Jadual 2, jejak jari ECFP4 menunjukkan kekerapan tertinggi dalam Ukuran F berbanding jejak jari lain iaitu sebanyak 6 kali manakala jejak jari FCFC4 pula tidak sesuai digunakan dalam kajian ini. Keputusan ini boleh disusun berdasarkan kekerapan mempunyai nilai tertinggi dalam urutan menurun seperti berikut:

ECFP4 (6) > SRECFC4 (3) = SRFCFC4 (3) > ECFC4 (2) > FCFP4 (1) > FCFC4 (tiada)

JADUAL 2. Nilai purata dapatan semula dan Ukuran F untuk pekali Tanimoto

Kelas Aktiviti	Purata Dapatan Semula						Ukuran F					
	ECFP4	ECFC4	FCFP4	FCFC4	SR ECFC4	SR FCFC4	ECFP4	ECFC4	FCFP4	FCFC4	SR ECFC4	SR FCFC4
AID1	0.0209	0.0205	0.0197	0.0159	0.0206	0.0192	**0.0400**	0.0393	0.0378	0.0305	0.0394	0.0367
AID53	0.0258	0.0250	0.0240	0.0219	0.0270	0.0260	0.0490	0.0475	0.0456	0.0416	**0.0513**	0.0493
AID129	0.0239	0.0236	0.0225	0.0213	0.0240	0.0241	0.0453	0.0447	0.0426	0.0404	0.0454	**0.0457**
AID248	0.0248	0.0226	0.0257	0.0238	0.0239	0.0266	0.0469	0.0426	0.0484	0.0449	0.0452	**0.0502**
AID256	0.0347	0.0281	0.0316	0.0281	0.0320	0.0307	**0.0634**	0.0513	0.0578	0.0513	0.0586	0.0561
AID372	0.0277	0.0350	0.0311	0.0320	0.0312	0.0333	0.0494	**0.0624**	0.0555	0.0571	0.0557	0.0594
AID376	0.0236	0.0208	0.0132	0.0200	0.0276	0.0180	0.0344	0.0303	0.0192	0.0292	0.0402	0.0262
AID565	0.0302	0.0234	0.0266	0.0254	0.0300	0.0255	**0.0561**	0.0436	0.0494	0.0472	0.0558	0.0475
AID567	0.0309	0.0276	0.0287	0.0292	0.0287	0.0290	**0.0492**	0.0440	0.0458	0.0466	0.0458	0.0462
AID573	0.0221	0.0242	0.0200	0.0241	0.0235	0.0232	0.0415	**0.0453**	0.0375	0.0452	0.0442	0.0435
AID607	0.1147	0.0853	0.1029	0.0765	**0.1176**	0.0971	0.0614	0.0457	0.0551	0.0409	**0.0630**	0.0520
AID612	0.0416	0.0361	0.0416	0.0317	0.0469	0.0481	0.0680	0.0589	0.0680	0.0519	0.0766	**0.0786**
AID798	0.0470	0.0387	0.0430	0.0288	0.0437	0.0414	**0.0719**	0.0592	0.0658	0.0441	0.0668	0.0633
AID823	0.0358	0.0255	0.0343	0.0307	0.0307	0.0321	**0.0426**	0.0304	0.0409	0.0365	0.0365	0.0383
AID846	**0.1209**	**0.0989**	**0.1308**	**0.0868**	0.1121	**0.1154**	0.1196	0.0978	**0.1293**	0.0858	0.1109	0.1141

Berdasarkan perbandingan antara nilai PPP dengan nilai purata dapatan semula, AID846 yang mempunyai saiz kelas yang kecil cenderung untuk mempunyai nilai PPP yang tinggi dan sebatian dalam kelas aktiviti tersebut adalah hampir sama antara satu dengan yang lain. Justeru, dapat disimpul bahawa kelas aktiviti yang homogen berkemungkinan mempunyai nilai purata dapatan semula lebih tinggi seperti yang ditunjukkan dalam Jadual 2.

Berdasarkan nilai PPP yang tinggi, AID846 sering menunjukkan nilai tertinggi untuk pengiraan purata dapatan semula. Mengambil kira saiz kelas aktiviti dalam formula dapatan semula, AID846 menunjukkan nilai terbaik dalam kalangan kelas aktiviti yang lain. Keputusan ini mencadang jejak jari ECFP4 menunjukkan prestasi yang baik berbanding 5 jejak jari yang lain berdasarkan kekerapan menjadi tertinggi dalam setiap jejak jari. Berdasarkan analisis, didapati jenis jejak jari molekul dan saiz kelas aktiviti memberi kesan kepada pengiraan persamaan molekul dalam set data PubChem mengguna pekali Tanimoto.

PENUTUP

Pangkalan data kimia PubChem merupakan set data yang mencabar untuk diuji berdasarkan nilai MPS yang rendah. Hasil ini disokong oleh pengenalan set data MUV yang merupakan gabungan kelas aktiviti daripada set data PuChem. Set data MUV dibina untuk menguji kedah kimia informatik dalam menangani cabaran set data berdasarkan kajian lepas (Abdo et al., 2010; Todeschini et al., 2012).

Keputusan daripada uji kaji ini menunjukkan nilai PPP dan saiz kelas aktiviti memainkan peranan dalam pengiraan purata dapatan semula. Saiz kelas aktiviti yang kecil menyumbang kepada nilai PPP yang tinggi seterusnya memberi kesan kepada pengiraan purata dapatan semula. Selain daripada itu, uji kaji ini juga mendapati jejak jari ECFP4 adalah terbaik berbanding dengan 5 jejak jari yang lain berdasarkan kekerapan tertinggi mengguna pekali Tanimoto.

Kajian ini dilanjut mengguna pekali persamaan yang lain untuk melihat kesan pekali yang berlainan terhadap capaian struktur kimia PubChem. Pekali persamaan terdiri daripada jenis pekali persamaan yang lain seperti pekali jarak.

RUJUKAN

Abdo, A., Chen, B., Mueller, C., Salim, N. & Willett, P. 2010. Ligand-Based Virtual Screening Using Bayesian Networks. *Journal of Chemical Information and Modelling,* 50(6): 1012-1020.

Arif, S.M., Holliday, J.D.& Willett, P. 2009. Analysis and use of fragment occurrence data in similarity based virtual screening. *Journal of Computer Aided Molecular Design,* 23(9): 655-668.

Bolton, E. E.,. Wang, Y., Thiessen, P. A. & Bryant S. H. 2008. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry,* 4(1): 217-241.

Chen, B & Wild, D.J. 2010.PubChem BioAssays as a data source for predictive models. *Journal of Molecular Graphics and Modelling,* 28 (5):420-426.

Han, L., Wang, Y. & Bryant, S. H. 2008. Developing and Validating Predictive Decision Tree Models from Mining Chemical Structural Fingerprints and High Throughput Screening Data in PubChem. *BMC Bioinformatics 2008,* 9(401):401 408.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E. & Schuffenhauer, A.2004. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic & Biomolecular Chemistry,* 22(2): 3256-3266.

Khalifa, A. A., Haranczyk, M. & Holliday, J. 2009. Comparison on nonbinary similarity coefficients for similarity searching, clustering and compound selection. *Journal of Chemical Information and Modelling,* 49(5): 1193-1201.

Leach, A. R. & Gillet, V. L. 2007. *An Introduction to Chemoinformatics*. Dordrecht: Springer.

Salim, N., J. Holliday& P. Willett. 2003. Combination of Fingerprint-based Similarity Coefficients Using Data Fusion. *Journal of Chemical Information and Computer Sciences,* 43(2): 435 - 442.

Shuib, M., Arif, S. & Malim, N. 2013.Comparison of Similarity Coefficients for Chemical Database Retrieval. *Proceeding of the First International Conference on Artificial Intelligence, Modelling & Simulation*.

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M. & Willett, P. 2012. Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated real datasets. *Journal of Chemical Information and Modelling,* 52(11): 2884-2901.

Weis, D. C., Visco, D. P. J. & Faulon, J. L. 2008. Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors. *Journal of Molecular Graphics and Modelling,* 27(4):466-475.

Willet, P. 2011. Similarity searching using 2D structural fingerprints. *Chemoinformatics and Computational Chemical Biology.* Springer.

Xu, J. & Hagler, A. 2001.Chemoinformatics and drug discovery. *Molecules,* 7(8): 566-600.

Shereena M. Arif
Fatimah Zawani Abdullah
Pusat Penyelidikan Teknologi Kecerdasan Buatan,
Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia,
shereena.arif@ukm.edu.my, fatimahzawani@gmail.com.

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

Received: 17 October 2014
Accepted: 13 November 2014