Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 1, Januari 2019, hlm. 702-707 http://j-ptiik.ub.ac.id

Pengelompokan Fungsi Aktif Senyawa Data SMILES (*Simplified Molecular Input Line Entry System*) Menggunakan Metode K-Means Dengan Inisialisasi Pusat Klaster Menggunakan Metode *Heuristic O(N LogN)*

Sherly Witanto[1], Dian Eka Ratnawati[2] , Syaiful Anam[3]

[1,2]Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
[3]Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Brawijaya
Email: [1]sherlywitanto@gmail.com, [2]dian_ilkom@ub.ac.id, [3]syaiful@ub.ac.id

Abstrak

Senyawa aktif mempunyai salah satu kegunaan sebagai bahan obat-obatan yang mampu mencegah maupun menyembuhkan penyakit. Sebagian senyawa aktif sudah ditemukan fungsinya dan sebagian lagi masih dalam tahap penelitian. Saat ini di Indonesia masih belum ada program yang mampu mengklasifikasi senyawa kimia sebagai obat untuk penyakit tertentu. Notasi SMILES merupakan konversi senyawa kimia dalam bentuk notasi baris. Notasi SMILES mampu memberikan kemudahan pada proses komputerisasi pada klasifikasi senyawa kimia. Klasifikasi atau pengelompokan notasi SMILES dilkukan dengan mengambil nilai 11 fitur atom B,S,N,O,I,F,C,P,Cl,Br dan OH yang ada pada senyawa tersebut. Sebelum diproses, untuk mendapatkan nilai fitur dilakukan proses dengan membagi masing-masing jumlah atom dengan panjang senyawanya. Algoritme K-Means merupakan metode klastering yang paling banyak digunakan karena bersifat mudah dan sederhana. Pengelompokan fungsi aktif menggunakan metode K-Means mempunyai kelemahan pada proses inisialisasi klaster yang bersifat *random*, sehingga digunakan metode *heuristic o(n logn)* untuk mendapatkan inisial klaster dengan nilai yang lebih baik. Berdasarkan perangkat lunak yang telah dibuat, pengujian dilakukan dengan menggunakan data latih sebanyak 512 dan data uji sebanyak 128. Akurasi yang diperoleh dari pengujian yaitu sebesar 63% dan pengujian menggunakan *K-Fold Cross Validation* dengan 10 kali pengujian menghasilkan akurasi rata-rata sebesar 52,58%. Pengujian menggunakan K-Means dengan *heuristic o(n logn)* menghasilkan akurasi yang lebih baik dibandingkan dengan K-Means konvensional.
Kata kunci: SMILES, K-Means, *Heuristic O(N LogN)*

Abstract

Active compounds have function as a medicine that can prevent or cure diseases. Some of the active compounds have been known the function and some are still in the research stage. Currently in Indonesia there is still no program that capable to classifying chemical compounds as drugs for certain diseases. SMILES notation is the conversion of chemical compounds in the form of line notation. Notation SMILES able to provide convenience to the process of computerization on the classification of chemical compounds. The classification of the SMILES notation is carried out by taking the values of the B, S, N, O, I, F, C, P, Cl, Br and OH atoms present in the compound. Before being processed, to get the value of the feature is done by dividing the process of each atom with the length of the compound. K-Means algorithm is the most widely used clustering method because it is easy and simple. The grouping of active function using K-Means method has weakness in random cluster initialization process, so that heuristic method o (n logn) is used to get the cluster initials with better value. Based on the software that has been made, the test is done using 512 of training data and test data as much as 128. Accuracy obtained from the test that is equal to 63% and testing using ¬K-Fold Cross Validation with 10 times the test produces an average accuracy of 52,58 %. Testing using K-Means with heuristic o (n logn) yielded better accuracy compared to conventional K-Means.

Keywords: *SMILES, K-Means, Heuristic O(N LogN)*

1. PENDAHULUAN

Senyawa adalah zat tunggal yang terdiri dari dua atau lebih unsur yang berbeda dan membentuk ikatan, sehingga terbentuklah senyawa sebagai zat baru yang mempunyai fungsi tertentu. Senyawa dibagi menjadi dua, yaitu senyawa aktif dan tidak aktif. Senyawa aktif mempunyai farmakologis yang berfungsi sebagai obat tertentu. Sedangkan senyawa tidak aktif tidak mempunyai peran signifikan dan hanya berfungsi sebagai zat tambahan/pengikat. Sebagian senyawa aktif sudah ditemukan fungsinya dan sebagian lagi belum ditemukan dan masih dalam tahap penelitian. (Rizki *et al*, 2015).

Bagi orang awam terutama bidang IT (*Information and Technology*) sangat sulit memahami senyawa kimia untuk mengetahui kegunaan senyawa tersebut. Sehingga diperlukan senyawa dengan bentuk yang mudah untuk dipahami yaitu dengan menggunakan kode SMILES (Simplified Molecular Input Line Entry System). Kode SMILES mengkonversi senyawa dalam bentuk notasi baris untuk menggambarkan senyawa kimia (Weininger, 1987). Saat ini di Indonesia masih belum ada program yang dapat mengklasifikasi senyawa kimia sebagai obat untuk penyakit tertentu. Masih sedikit orang yang mengetahui notasi SMILES dan bagaimana melakukan pengelompokan senyawa tersebut. Notasi SMILES sendiri dapat memberikan kemudahan pada proses komputerisasi dan penyimpanan data di komputer supaya mempermudah pengelompokan senyawa kimia.

K-Means merupakan metode klastering paling umum, sederhana dan mudah. Hal ini dikarenakan K-Means dapat mengelompokkan data dalam jumlah banyak dan dalam waktu yang relatif cepat. Algoritme K-Means konvensional memiliki kelemahan dikarenakan pembentukan awal pusat klaster yang bersifat *random*. Hasil dan kecepatan proses pengelompokan data bergantung pada inisialisasi tersebut, sehingga untuk menghindari hasil yang kurang maksimal diperlukan metode yang membantu dalam penentuan inisial awal pusat klaster (Tahta, 2012). Berbeda dengan K-Means konvensional yang inisialisasi pusat klasternya dilakukan secara random, K-Means dengan inisialisasi klaster menggunakan metode *heuristic o (n logn)* atau disebut juga *improved* K-Means melakukan inisialisasi pusat klaster dengan distribusi data (Nazeer, 2011).

2. TINJAUAN PUSTAKA

2.1. SMILES (*Simplified Molecular Input Line Entry System*)

SMILES merupakan notasi kimia yang dirancang khusus untuk ahli kimia dalam penggunaan komputer. Hal ini cukup fleksibel untuk menginterpretasi notasi kimia independen dan spesifik. Sistem SMILES dirancang agar interaktif dari segi pengguna komputer, ahli kimia maupun sistem itu sendiri. Penggunaan kode SMILES yang sederhana memungkinkan pengguna mengkodekan struktur kimia yang mudah digunakan (Weininger, 1987).

Tata cara penulisan notasi kode SMILES yaitu (Junaedi, 2011) :

1. Penulisan Atom

   Penulisan atom disesuaikan dengan simbol atomic senyawa. Penulisan atom dilakukan dengan cara menuliskan huruf besar. Apabila memiliki simbol lebih dari satu huruf, maka huruf pertama huruf besar dan diikuti dengan huruf kecil.
2. Penulisan Ikatan

   Ikatan antar atom terbagi menjadi tiga macam, yang pertama yaitu ikatan tunggal dilambangkan dengan notasi "-", yang kedua yaitu ikatan rangkap yang dilambangkan dengan notasi "=" dan yang ketiga yaitu ikatan rangkap tiga yang dilambangkan dengan notasi "#".
3. Penulisan Percabangan

   Penulisan notasi pada percabangan ditandai dengan kurung buka dan kurung tutup "()".

2.2 K-Means

Analisa klaster merupakan kegiatan yang menganalisa kumpulan obyek untuk menemukan kesamaan dan perbedaan sehingga membentuk suatu klaster yang sama maupun berbeda dengan obyek tersebut (Hermawati, 2013). Pengklasteran bertujuan untuk mengelompokkan dan memahami struktur data. Klasterisasi hanya tahap awal untuk kemudian dilanjutkan dengan pengolahan inti dan pelabelan kelas pada tiap kelompok. Hal ini nantinya dapat digunakan sebagai data latih.

Algoritme klastering K-Means dapat membagi data berdasarkan jarak antar data pada kelompok yang telah ditetanpkan. Algoritme ini bergantung pada fungsi untuk mengukur data yang mempunyai ciri khas sama. Jarak itu sendiri dihitung menggunakan fungsi *euclidean*. Kemudian data dimasukkan dalam kelompok yang mempunyai jarak terdekat (Santosa, 2007). Langkah-langkah pengelompokan data adalah (Santosa, 2007) :

1. Pilih jumlah klaster.
2. Inisialisasi awal dan pusat klaster dilakukan secara random.
3. Setiap data ditempatkan ke pusat klaster terdekat berdasarkan jarak antar obyek. Pada tahap ini jarak dihitung dengan menentukan kemiripan atau ketidakmiripan data dengan metode jarak *Euclidean* (*Euclidean Distance*) dengan rumus seperti pada persamaan 1 :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

   dimana :
   $d(x,y)$ = ukuran ketidakmiripan
   $x_i = (x_1, x_2, \dots, x_i)$ yaitu variabel data
   $y_i = (y_1, y_2, \dots, y_j)$ yaitu variabel pada titik pusat.
4. Hitung pusat klaster yang baru dengan keanggotaan yang baru dengan cara menghitung rata-rata obyek pada klaster. Penghitungan bisa juga dengan menggunakan median.
5. Hitung kembali jarak tiap objek dengan pusat klaster yang baru, hingga klaster tidak berubah, maka proses pengklasteran selesai.

2.3. *Heuristic O(n logn)*

Algoritme *heuristic n(o logn)* digunakan untuk inisialisasi pusat klaster. Berbeda dengan algoritme K-Means Konvensional yang inisialisasi pusat klasternya ditentukan secara random, K-Means dengan *heuristic n (o logn)* pusat klasternya sudah ditentukan di awal dan tidak berubah-ubah. Cara kerja metode ini adalah dengan mempartisi inputan ke sejumlah klaster, kemudian dirata-rata untuk digunakan sebagai nilai awal pusat klaster (Nazeer, 2011). Langkah-langkah metode ini adalah :

1. Pada setiap kolom fitur dataset ditentukan nilai terbesar dan terkecil elemen.
2. Menentukan jarak setiap kolom fitur dengan mencari selisih antara nilai terbesar dan terkecil pada poin 1.
3. Data diurutkan dari nilai terkecil ke terbesar berdasarkan nilai jarak terbesar yang sudah dicari pada poin 2.
4. Mempartisi data sejumlah klaster menjadi bagian sama banyak.
5. Menghitung rata-rata tiap fitur untuk masing-masing klaster yang sudah dipartisi pada poin 4.
6. Lakukan perhitungan jarak minimum menggunakan *euclidean distance* berdasarkan pusat klaster awal yang sudah didapat dari poin 5 dengan masing-masing data.

3. METODOLOGI

3.1. Deskripsi Umum Sistem

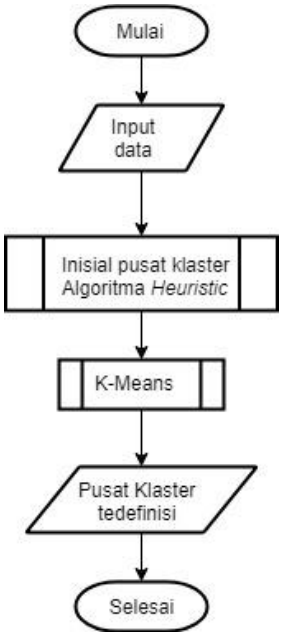
Sistem yang akan dibuat betujuan untuk mengetahui penerapan metode K-Means dengan inisialisasi pusat klasternya menggunakan metode heuristic o (n logn). Fitur yang digunakan sebagai masukan ada 11, yaitu jumlah masing-masing elemen atom yang kemudian akan dibagi dengan panjang kode SMILES. Sistem akan mengolah masukan dengan cara pembelajaran. SIstem ini mempunyai 2 proses, yaitu proses pelatihan dan proses pengujian. Pada proses pelatihan dan pengujian masing-masing memerlukan masukan berupa data latih dan data uji. Proses yang akan dijalankan oleh sistem adalah :

1. Proses Pelatihan
   Proses ini bertujuan untuk mendapatkan nilai pusat klaster dari metode K-Means konvensional dan *improved* K-Means. Nilai pusat klaster didapat nantinya akan digunakan pada proses pengujian dengan data uji. Data latih digunakan untuk mendapatkan nilai pusat klaster yang berjumlah 2 klaster.
2. Proses Pengujian
   Proses pengujian menggunakan nilai pusat klaster yang sudah didapatkan dari proses pelatihan yang akan dilakukan proses pengelompokan dengan data uji. Keluaran yang dihasilkan berupa data yang sudah dikelompokkan berdasarkan 2 klaster penyakit.

3.2. Perancangan Sistem

Perancangan dibuat untuk mengetahui bagaimana sistem dengan metode K-Means dengan inisialisasi pusat klaster *heuristic o(n*

logn) bekerja pada data senyawa aktif SMILES. Metode *heuristic o(n logn)* digunakan untuk menginisialisasi pusat klaster awal, selanjutnya algoritme K-Means akan memproses pengelompokan data. Alur perancangan sistem ditunjukkan pada Gambar 1.



Gambar 1. Alur Perancangan Sistem

Sistem mempunyai 2 proses yaitu *improved* K-Means dan K-Means konvensional. Gambar 1 menunjukkan alur kerja dari *improved* K-Means. Perbedaannya terletak pada inisialisasi pusat klaster awal. Pada K-Means konvensional, inisialisasi pusat klaster diambil secara random, sedangkan pada *improved* K-Means, inisialisasi dilakukan menggunakan algoritme *heuristic o(n logn).*

3.3. Basis Pengetahuan

Basis pengetahuan berisi pengetahuan yang digunakan untuk memahami, merumuskan dan memecahkan masalah. Basis pengetahuan merupakan representasi pengetahuan dari hasil analisis data SMILES. Terdapat 11 fitur masukan yang digunakan sebagai perhitungan pengelompokan senyawa kode SMILES meliputi jumlah masing-masing elemen B, C, N, O, P, S, F, Cl, Br, I dan OH. Masing-masing fitur akan dibagi dengan panjang senyawa kode SMILES sebelum diproses.

3. HASIL

Pengujian menggunakan seluruh data dengan komposisi 2 kelas yaitu metabolisme dan kanker yang berjumlah total 640 data. Pengujian yang dilakukan meliputi pengujian validitas program, pengujian data latih dan data uji serta pengujian *K-Fold Cross Validation.*

4.1. Pengujian Validitas Program

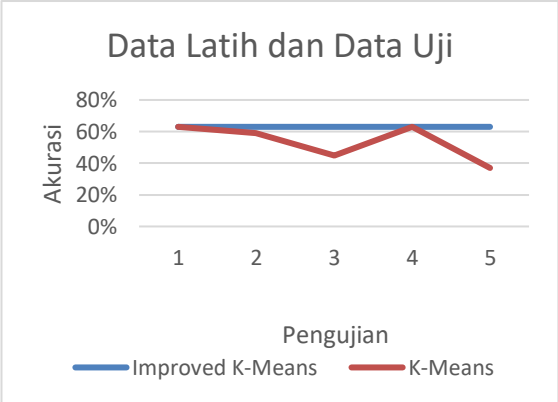
Pengujian ini dilakukan dengan tujuan mengetahui kesesuaian program dengan perhitungan manual. Hasil yang didapatkan mengunakan 20 data latih dan 10 data uji sebesar 40% sudah sesuai dengan perhitungan manual.

4.2. Pengujian Data Latih dan Data Uji

Pengujian data latih dan data uji dilakukan dengan menggunakan seluruh data yang ada yaitu sebanyak 512 data latih dan 128 data uji. Data latih terdiri dari 350 data metabolisme dan 162 data kanker, sedangkan data uji terdiri dari 87 data metabolisme dan 41 data kanker. Hasil pengujian yang dilakukan sebanyak 5 kali mendapatkan nilai akurasi rata-rata improved K-Means sebesar 63%, sedangkan rata-rata K-Means sebesar 53.4%. Hasil pengujian dapat dilihat pada Tabel 1 dan Gambar 2.

Tabel 1. Pengujian Data Latih dan Data Uji

Pengujian	*Improved* K-Means	K-Means
1	63%	63%
2	63%	59%
3	63%	45%
4	63%	63%
5	63%	37%
Rata-rata	63%	53.4%



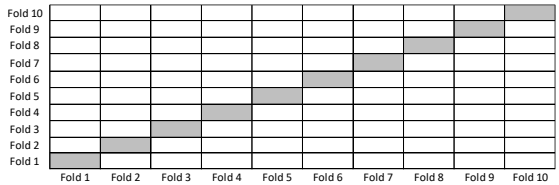
Gambar 2. Grafik Pengujian Data Latih dan Data Uji

Berdasarkan grafik pengujian pada Gambar 2 yang didapat dari 5 kali pengujian menggunakan data latih dan data uji yang sama, didapat kesimpulan bahwa metode improved K-

Means mempunyai hasil yang konstan, sedangkan K-Means mempunyai hasil yang berbeda. Hal ini dikarenakan inisialisasi yang dilakukan pada K-Means bersifat random, sehingga dapat bergantung pada inisial yang dilakukan. Berbeda dengan improved K-Means yang inisialisasi pusat klaster awalnya dilakukan dengan menggunakan metode heuristic o(n logn), hasil yang didapat tidak berubah, sehingga akurasi maksimum bisa langsung diketahui.

4.2. Pengujian Data Latih dan Data Uji

Pengujian ini menggunakan data latih dan data uji dengan jumlah data latih sebanyak 512 dan data uji sebanyak 66. Dataset dibagi menjadi 10 bagian, kemudian masing-masing bagian akan dijadikan data uji, sedangkan bagian yang lain akan menjadi data latih. Gambar 3 merupakan skenario pembagian dataset setiap Fold.



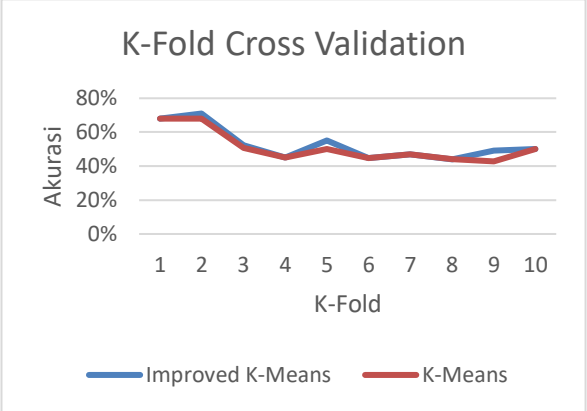
Gambar 3. Pembagian Dataset

Pengujian *K-Fold Cross Validation* mempunyai hasil akurasi rata-rata *improved* K-Means sebesar 52.58%, lebih baik dibandingkan K-Means konvensional dengan rata-rata akurasi sebesar 51%. Tabel 2 menunjukkan hasil pengujian *K-Fold Cross Validation.*

Tabel 2. Pengujian *K-Fold Cross Validation*

K-Fold	*Improved* K-Means	K-Means
1	68%	68%
2	71%	68%
3	52.3%	50.7%
4	45%	45%
5	55%	50%
6	44.6%	44.6%
7	46.9%	46.9%
8	44%	44%
9	49%	42.8%
10	50%	50%
Rata-rata	52.5%	51%

Dengan pengujian sebanyak 10 *Fold*, masing-masing hasil akurasi menunjukkan bahwa *Improved K-Means* mempunyai akurasi sama atau lebih baik dari pada K-Means Konvensional. Akurasi terbesar *Improved K-Means* yaitu pada Fold ke 2 sebesar 71% dan akurasi terkecil pada Fold ke 8 sebesar 44%. Sedangkan pada K-Means akurasi terbesar pada Fold ke 1 dan 2 yaitu 68% dan akurasi terkecil pada Fold ke 9 sebesar 42.8%. Grafik hasil pengujian dapat dilihat pada Gambar 4.



Gambar 4. Grafik *K-Fold Cross Validation*

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan kesimpulan yang didapat adalah :

1. Penerapan metode *heuristic o(n logn)* untuk inisialisasi pusat klaster K-Means didapatkan dengan mencari nilai maksimum dan minimum setiap atribut dataset. Data diurutkan berdasarkan kolom yang mempunyai nili rentang terbesar. Data dibagi menjadi ‘k’ bagian sama banyak, kemudian dihitung nilai rata-rata tiap bagian untuk ditetapkan sebagai nilai pusat klaster awal.
2. Penelitian dengan menerapkan inisial pusat klaster K-Means menggunakan metode *heuristic o(n logn)* pada senyawa aktif kode SMILES mampu meningkatkan akurasi lebih baik dibandingkan K-Means konvensional. Pengujian dilakukan menggunakan 2 kelas dataset yaitu kelas kanker dan metabolisme. Hasil yang didapatkan meliputi :
   a. Pengujian validasi program yang bertujuan untuk mengetahui kesesuaian program dengan perhitungan manual sudah sama. Hasil akurasi yang diperoleh dari pengujian validasi program yaitu sebesar 40%.
   b. Pengujian data latih dan data uji dengan menggunakan seluruh data latih dan

data uji yaitu sebanyak 512 data dan 128 data memperoleh hasil akurasi sebesar 63%.

c. Pengujian *K-Fold Cross Validation* pada *improved K-Means* memperoleh rata-rata akurasi sebesar 52.58% dari 10 kali percobaan. Sedangkan rata-rata akurasi untuk K-Means sebesar 51%.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah :

Data pada notasi SMILES bersifat acak dan kurang menunjukkan karakteristik pada tiap kelasnya. Hal ini ditunjukkan dengan adanya notasi yang tidak terdapat dalam pemilihan fitur. Untuk meningkatkan akurasi, dapat dilakukan pemilihan data dan kelas yang berpengaruh pada fitur yang digunakan.

DAFTAR PUSTAKA

Christopher D. Manning, P. R. H. S., 2009. Introduction to Information Retrieval. England: Cambridge University Press.

Hanifa Maria Salni, R. W. M., 2011. Isolasi Senyawa Antibakteri Dari Daun Jengkol (Pithecolobium Iobatum Benth) dan Penentuan Nilai KHM-nya. Jurnal Penelitian Sains, Volume 14, pp. 38-41.

Hermawati, F. A., 2013. Data Mining. Yogyakarta: Penerbit Andi.

Junaedi, H., 2011. Penggambaran Rantai Karbon Dengan Menggunakan Simplified Molecular Input Line System (SMILES). Sekolah Tinggi Teknik Surabaya.

K A Abdul Nazeer, S. M. K. M. P. S., 2011. Enhancing the K-means Clustering Algorithm by Using a O(n logn) Heuristic Method for Finding Better Initial Centroids. International Conference on Emerging Applications of Information Technology- EAIT 2011.

Muhammad Ikhwan Rizki, E. M. H., 2015. Aktivitas Farmakologis, Senyawa Aktif, dan Mekanisme Kerja Daun Salam (Syzygium Polyanthum). Prosiding Seminar Nasional & Workshop "Perkembangan Terkini Sains Farmasi & Klinik", Volume 5, pp. 239-244.

Prasetyo, E., 2012. Klasifikasi Metode-Metode Pilihan. Yogyakarta: Penerbit Andi.

Rinadewi Astuti, D. E. R. B. D. S., 2015. Implementasi Algoritme K-Means Clustering dengan Inisialisasi Centroid Menggunakan Metode Heuristic O(N LogN). Repositori Jurnal Mahasiswa PTIIK UB, 6(17).

Santosa, B., 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi. Yogyakarta: Graha Ilmu.

Tahta Alfina, B. S. A. R. B., 2012. Analisa Perbandingan Metode Hierarchical Clustering,K-Means dan Gabungan Keduanya Dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). Jurnal Teknik POMITS, Volume 1, pp. 1-5.

Weininger, D., 1988. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. J. Chem. Inf. Comput. Sci., pp. 31-36.