PENGGALIAN INFORMASI MENGGUNAKAN MODEL TERDEKOMPOSISI: APLIKASI PADA RISET PENEMUAN ANTIBIOTIK

Nyoman Juniarta
Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember
Jalan Teknik Kimia, Gedung Teknik Informatika, Kampus ITS Sukolilo, Surabaya, 60111
e-mail: nyoman14@mhs.if.its.ac.id

ABSTRAK

Penemuan obat-obatan antibiotik adalah salah satu tantangan pada bidang kemo-informatika. Dibutuhkan antibiotik baru secara cepat dan efektif karena banyak bakteri menjadi kebal terhadap antibiotik lama. Molekul-molekul kimia yang tersimpan di beberapa perusahaan dan laboratorium menyediakan kandidat yang berpotensi sebagai antibiotik baru. Tetapi, terlalu banyak kandidat yang harus diteliti. Untuk mengatasinya, dibutuhkan pencarian informasi yang dapat mendeteksi kandidat-kandidat penting melalui atribut mereka. Jumlah atribut tersebut sangatlah besar. Tujuan penelitian ini adalah mempelajari atribut-atribut tersebut dan menentukan atribut yang penting, dengan kata lain, untuk mereduksi dimensi data molekul. Fokus penelitian ini ditujukan pada molekul-molekul antibiotik yang sudah ada di pasaran, dengan sekitar 500 atribut yang diperoleh dari penelitian sebelumnya. Sebagai prosedur seleksi fitur, penelitian ini menggunakan analisis log-linear untuk menemukan asosiasi di antara atribut. Karena jumlah atribut mencapai ratusan, maka digunakan Chordalysis yang bekerja pada model log-linear yang bisa didekomposisi. Penelitian ini menemukan bahwa atribut-atribut dari penelitian sebelumnya memiliki beberapa asosiasi. Dengan demikian, beberapa atribut yang redundan dapat dieliminasi.

***Kata Kunci**: antibiotik, model probabilistik grafis, penggalian informasi, seleksi fitur.*

ABSTRACT

Antibacterial drug discovery is one of the emerging challenges in chemo-informatics. There is an urgent need for finding new effective drugs faster because many bacteria become resistant to the old drugs. The chemical molecules stored in companies and laboratories' databases provide potential candidates for developing new drugs. However, there are far too many candidates to investigate. This is where knowledge discovery could be of help, by sifting through the known properties of the molecule to select the most promising candidates for further experiments. The number of properties, or descriptors, characterizing the molecules is rather large. The aim of the present work is to study these descriptors and find out which ones really matter, that is, to reduce the dimension of the description space. We focus on a subset of antibacterial molecules already on the market, with around 500 descriptors obtained from the selection process in the previous work. As our feature selection procedure, we use log-linear analysis (LLA) to discover associations among descriptors. Given that the number of descriptors is high, we study Chordalysis that focuses on a specific subset of log-linear models: decomposable models. We find that the selected descriptors from the previous work still have many associations among them. Therefore, a number of redundant descriptors can still be left out.

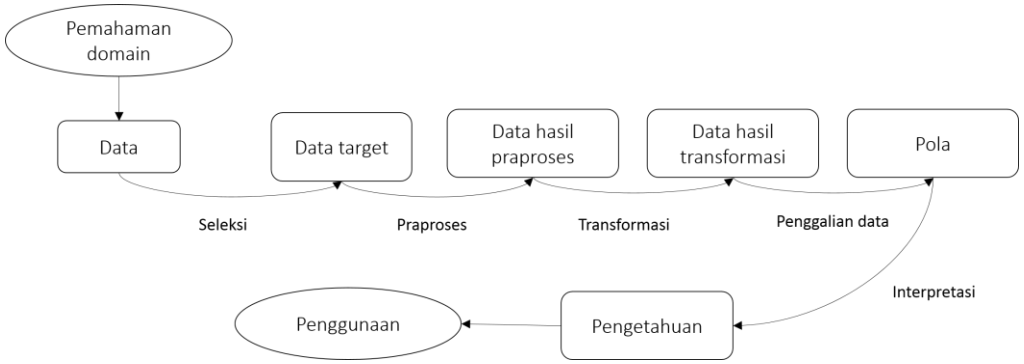
***Keywords**: antibacterial drug, feature selection, knowledge discovery, probabilistic graphical model.*

I. PENDAHULUAN

PENYAKIT yang berasal dari bakteri dan parasit merupakan salah satu penyebab utama kematian di seluruh dunia. Karena aman dan efektif dalam menangani penyakit-penyakit menular tersebut, antibiotik banyak digunakan tanpa proses diagnosis dan uji kerentanan. Dengan semakin banyaknya penggunaan antibiotik, beberapa bakteria mengalami evolusi, mutasi, dan seleksi, sehingga mereka menjadi kebal terhadap antibiotik tersebut [1]. Kemudian, seiring dengan munculnya "*superbug*" yang kebal antibiotik, seperti *Staphylococcus aureus* (MRSA), antibiotik-antibiotik tradisional menjadi tidak efektif.

Munculnya bakteri-bakteri yang kebal antibiotik menyebabkan banyak perusahaan obat-obatan mengurangi penelitian di bidang tersebut. Riset untuk menemukan antibiotik baru kini dijalankan oleh komunitas akademik, termasuk para peneliti di ilmu komputer. Kerja sama antara ilmu komputer dan kimia komputasional melahirkan bidang ilmu baru, kemoinformatika. Kemoinformatika menggabungkan bermacam-macam area penelitian dan teknologi, termasuk pembelajaran mesin, pengenalan pola, dinamika molekul, mekanika kuantum, dan statistika. Bidang ilmu ini banyak dieksplorasi seiring dengan bertumbuhnya kemampuan kalkulasi komputer dan volume data. Salah satu tujuannya adalah menghasilkan metode yang lebih efisien sebagai metode penemuan obat antibiotik baru.

Dengan berlimpahnya data tentang molekul yang tersimpan di perusahaan-perusahaan dan di laboratorium, dimensi yang sangat besar dari tiap molekul merupakan suatu keuntungan sekaligus kerugian. Volume yang besar



Gambar 1. Diagram proses KDD [2].

dari basis data tersebut menjamin adanya molekul yang bisa menjadi antibiotik, tetapi juga memunculkan kesulitan dalam menemukan satu dari ribuan molekul. Untuk mengatasinya, peneliti dapat menggunakan subset dari seluruh molekul, dengan ribuan atribut untuk tiap molekul.

Struktur dan properti molekul merupakan sumber informasi yang sangat berharga untuk mendefinisikan suatu antibiotik. Selain informasi fisikokimia dan properti dari graf kimia, informasi lain bisa diperoleh dari perubahan struktur dinamis dan dari interaksi antara senyawa dan target. Agar informasi yang berlimpah tersebut dapat digunakan secara efektif, informasi kimia perlu dikonversi menjadi data yang penuh arti dan berguna. Tiap molekul dapat direpresentasikan oleh suatu himpunan nilai numerik atau kategorikal, yang disebut atribut. Atribut ini menunjukkan properti fisikokimia dan topologi. Data tersebut dapat dihitung melalui berbagai macam metode, kompleksitas informasi, dan waktu eksekusi. Beberapa atribut berkaitan dengan grafik molekul, sedangkan yang lainnya berkaitan dengan representasi tiga dimensi. Hasilnya, dari suatu himpunan molekul, didapatkan matriks dengan baris sebagai ID molekul, dan kolom sebagai atribut.

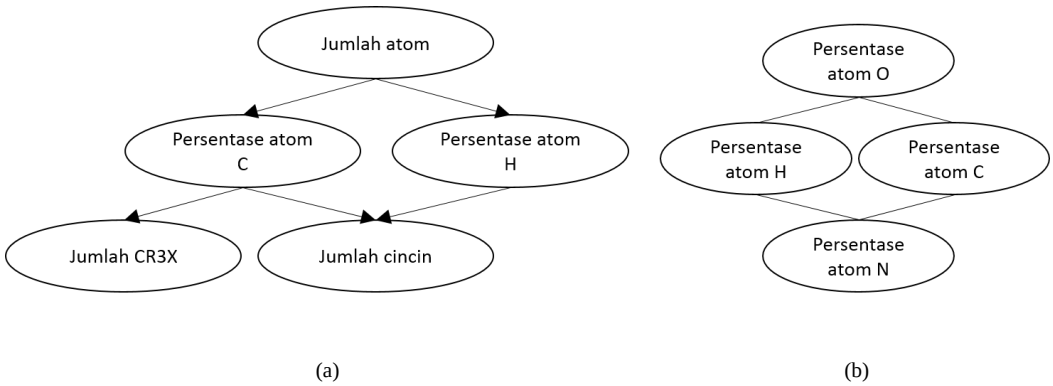
Dengan adanya variasi atribut kimia yang dapat digunakan (terdapat ribuan atribut), maka muncul permasalahan mengenai pemilihan atribut untuk analisis kemiripan kimia secara akurat. Atribut-atribut yang terpilih harus dapat digunakan untuk mendeskripsikan molekul-molekul dengan tingkah laku yang sama. Ekstraksi informasi tersebut dapat membantu menentukan, di antara himpunan data molekul yang sangat besar, molekul-molekul yang membutuhkan penanganan yang sama.

Dengan demikian, kesulitan yang ditemui adalah bagaimana mereduksi dimensi dari matriks atribut (dari ribuan menjadi ratusan) tanpa mengurangi informasi penting. Cara yang umum adalah dengan mendeteksi korelasi antar atribut untuk menghindari redundansi. Tetapi, diketahui bahwa analisa statistik dasar tidak cukup efisien sehingga dibutuhkan metode baru.

II. KAJIAN PUSTAKA

Penelitian ini berfokus pada permasalahan reduksi dimensi, lebih tepatnya pada seleksi fitur. Tujuan dari penelitian ini adalah mengurangi dimensi dari data kimia dengan memilih sebagian atribut untuk menghindari redundansi. Reduksi dimensi ini merupakan salah satu subproses dari *knowledge discovery in database* (KDD) [2]. Seperti ditunjukkan pada Gambar 1, KDD terdiri dari beberapa langkah. Langkah pertama dari KDD adalah proses memahami domain permasalahan dan keinginan klien. Kemudian, pada proses seleksi, peneliti memilih satu atau beberapa data set yang akan digunakan, disebut data target. Pada langkah ketiga, dilakukan praproses pada data target. Praproses tersebut dapat berupa penghilangan derau atau penanganan *missing values*. Keempat, peneliti mengurangi kompleksitas data dengan transformasi, yaitu membentuk representasi dari data. Proses transformasi ini dapat dilakukan dengan mengeliminasi atau mentransformasi beberapa atribut.

Tiga langkah berikutnya merupakan proses penggalian data dari data yang telah ditransformasi. Langkah kelima adalah mencocokkan keinginan klien terhadap metode penggalian data (klasifikasi, pengelompokan, regresi, dan sebagainya). Setelah menentukan metode, pada langkah keenam, peneliti kemudian menentukan algoritma yang akan digunakan. Pada langkah ketujuh, peneliti menjalankan proses penggalian data. Dari langkah tersebut, peneliti dapat menemukan pola-pola yang cukup berarti dari data. Langkah berikutnya adalah interpretasi. Setelah menemukan pola, peneliti harus menginterpretasikan pola tersebut. Salah satu cara adalah dengan visualisasi. Langkah terakhir adalah mengintegrasikan pengetahuan (*knowledge*) ke proses KDD lain, melaporkannya ke



Gambar 2. Contoh representasi grafis dari jaringan Bayesian (a) dan jaringan Markov (b).

pihak-pihak yang berkepentingan, atau langsung menggunakannya. Peneliti juga dapat menjalankan kembali beberapa langkah sebelumnya untuk meningkatkan kualitas pengetahuan.

KDD dapat dibedakan menjadi dua, yaitu numerik dan simbolis. Metode yang digunakan pada penelitian ini merupakan model probabilistik grafis, salah satu bagian dari metode numerik.

A. Seleksi Fitur

Untuk menemukan antibiotik baru, peneliti memeriksa sejumlah molekul kimia, untuk dapat mengetahui molekul mana yang dapat berlaku sebagai antibiotik. Untuk menjalankannya, peneliti dapat melihat properti dari tiap molekul seperti jumlah atom, adanya atom-atom tertentu, adanya cincin, polaritas, eksentrisitas, dan lain-lain. Properti-properti tersebut berjumlah ribuan, yang berarti bahwa peneliti membutuhkan waktu dan memori yang sangat besar.

Untuk meminimalkan kebutuhan ruang dan waktu, sebelum penggalian data, peneliti sebaiknya mereduksi dimensi data, sebagai proses transformasi dari KDD. Proses ini bisa dilakukan dengan menggabungkan beberapa atribut menjadi satu, atau bisa juga dengan mengeliminasi beberapa atribut.

Dari ribuan atribut untuk tiap molekul, dapat ditemukan beberapa redundansi. Persentase atom H berkorelasi dengan banyak atom H, banyak atom C berkorelasi dengan banyak struktur CR3X, dan beberapa korelasi tersembunyi lainnya. Dengan demikian, suatu atribut dapat dieliminasi bila atribut tersebut bisa diwakilkan oleh atribut lain. Atribut-atribut yang terpilih harus dapat mendefiniskan molekul tanpa kehilangan informasi secara signifikan.

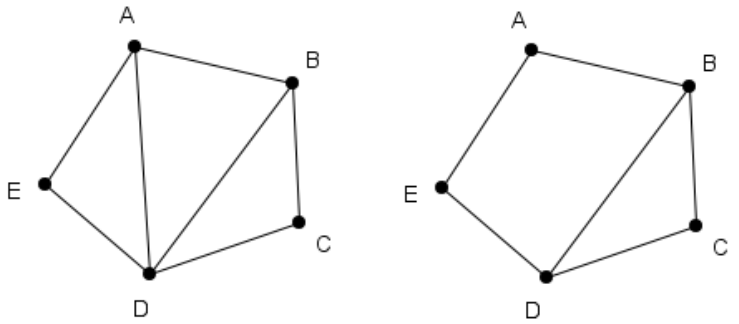
Selain redundansi, peneliti juga dapat mengeliminasi atribut yang tidak deskriptif. Sebagai contoh, jika terdapat atribut yang menyatakan banyak phospane dalam suatu molekul, tetapi semua molekul tidak memiliki phospane, maka atribut tersebut tidak deskriptif, yang berarti tidak dapat membedakan antara molekul antibiotik dengan molekul non-antibiotik.

B. Model Probabilistik Grafis

Salah satu aspek dari KDD numerik adalah model probabilistik. Sebagai contoh, untuk menemukan antibiotik baru, penting untuk mengetahui probabilitas bahwa suatu molekul dapat berperan sebagai antibiotik. Probabilitas ini bergantung pada atribut tiap molekul. Tiap molekul memiliki kira-kira 4000 atribut, mengenai struktur, elemen, massa, dan sebagainya. Tiap atribut dapat memiliki lebih dari dua nilai, sehingga dimensi probabilitas memiliki minimal 2^{4000} nilai. Peneliti kemudian harus mengetahui jumlah kemunculan tiap nilai tersebut untuk mendapatkan probabilitas gabungan.

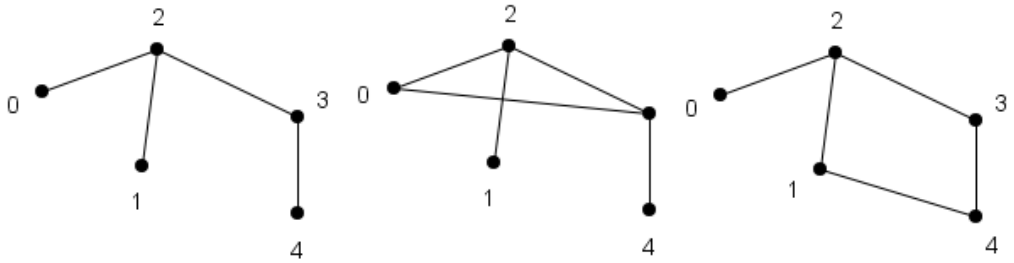
Dengan banyaknya kemungkinan nilai pada data kimia, penyusunan distribusi gabungan akan menjadi sangat kompleks. Untuk mendeskripsikan distibusi yang kompleks tersebut, digunakan model grafis probabilistik, yang menggambarkan suatu distribusi sebagai graf.

Terdapat dua tipe model probabilistik grafis: jaringan Bayesian dan jaringan Markov [3]. Graf dari kedua tipe tersebut memiliki titik sebagai atribut, dan garis sebagai interaksi probabilistik. Garis pada jaringan Bayesian merupakan garis berarah, sedangkan garis pada jaringan Markov merupakan garis tak berarah. Contoh graf dari dua tipe ini digambarkan pada Gambar 2. Pada graf-graf tersebut, terdapat beberapa independensi antar atribut, disimbolkan dengan $\perp$. Pada Gambar 2a, banyak atom pada suatu molekul tidak berinteraksi langsung dengan banyak CR3X. Hal ini menunjukkan, jika diketahui persentase atom C dan H, maka distribusi banyak cincin dapat diketahui, tanpa harus mengetahui banyak atom. Independensi ini ditulis sebagai (cincin $\perp$ atom|persentase C,



(a) (b)

Gambar 3. Contoh graf *chordal* (a) dan non-*chordal* (b).



(a) (b) (c)

Gambar. 4. Contoh *M** pada suatu iterasi (a), kandidat yang memenuhi syarat (b), dan kandidat yang tidak memenuhi syarat (c).

persentase H). Selain itu, jika diketahui banyak atom, maka persentase C independen terhadap persentase H.

Sementara itu, pada Gambar 2b, terdapat dua independensi, (C ⊥ H|N, O) dan (N ⊥ O|C, H).

C. Analisis Log-linear

Analisis log-linear (ALL) adalah suatu ekstensi dari analisis frekuensi multiarah yang mencoba menemukan relasi statistik antara tiga atau lebih variabel diskrit [4]. Untuk melakukan analisis frekuensi multiarah dengan ALL, dibangun sebuah model linear terhadap logaritma frekuensi harapan untuk tiap kombinasi nilai variabel. Seiring dengan bertambahnya jumlah variabel, jumlah asosiasi juga bertambah. Jika terdapat tiga variabel, maka terdapat tujuh asosiasi: satu asosiasi tiga arah, tiga asosiasi dua arah, dan tiga asosiasi satu arah. Untuk meminimalkan kompleksitas sebuah model, ALL mencoba menentukan asosiasi mana yang akan dieliminasi. Asosiasi yang akan dieliminasi adalah asosiasi yang tidak signifikan. Dengan ribuan variabel, banyak asosiasi akan menjadi sangat besar sedemikian hingga tidak praktis untuk menguji semua asosiasi. Batasan ini dapat diselesaikan menggunakan Chordalysis, suatu model probabilistic grafis yang akan dijelaskan kemudian.

Suatu model log-linear bersifat grafis jika, kapanpun terdapat seluruh asosiasi dua arah yang berasal dari orde yang lebih tinggi, model tersebut juga mengandung orde tinggi tersebut. Karena adanya asosiasi dua arah, maka model grafis bisa digambarkan sebagai graf, dengan titik sebagai variabel dan garis sebagai asosiasi dua arah antar variabel.

Model log-linear grafis disebut *decomposable* bila representasi grafnya bisa didekomposisi. Dekomposisi graf *G* adalah partisi dari titik-titik *V* menjadi tiga subset independen (*A*, *B*, *S*), dengan:

- $A \neq \emptyset$ dan $B \neq \emptyset$,
- *S* membentuk subgraf komplet,
- *A* dan *B* tidak terhubung dalam *G* – *S*.

Setiap graf dapat didekomposisi secara rekursif sampai pada subgrafnya yang bersifat prima maksimal, yaitu subgraf yang tidak bisa didekomposisi lebih jauh. Suatu graf disebut *decomposable* jika graf tersebut komplet atau jika graf tersebut bisa didekomposisi menjadi graf-graf lain yang *decomposable*. Dengan definisi tersebut, maka seluruh graf prima maksimal dari graf *decomposable* adalah *clique*.

Selain itu, suatu graf bersifat *decomposable* jika dan hanya jika graf tersebut *chordal*. Pada graf *chordal*, setiap siklus dengan panjang lebih dari tiga memiliki *chord*, yaitu suatu garis yang bukan bagian dari siklus tetapi menghubungkan dua titik pada siklus. Graf pada Gambar 3a termasuk graf *chordal*, tetapi graf pada Gambar 3b

bukan merupakan graf *chordal* karena siklus (A,B,D,E) tidak memiliki *chord*.

Model *decomposable* adalah satu-satunya model log-linear yang memiliki *maximum likelihood estimates* yang tertutup [5]. Selain itu, model *decomposable* juga memiliki keuntungan mengenai *separator* minimal dan *clique* maksimal. Properti-properti tersebut dapat diperoleh dalam waktu linear, menggunakan *Lexicographic Breadth First Search* atau menggunakan *Maximum Cardinality Search* [6]. Keuntungan lain adalah mengenai statistik G^2, yang dapat diperoleh melalui observasi struktur grafnya.

III. METODOLOGI PENELITIAN

A. *Chordalysis*

Terdapat dua cara untuk menentukan asosiasi-asosiasi yang akan dipilih dalam model log-linear, eliminasi mundur dan seleksi maju. Eliminasi mundur dimulai dari model jenuh dan satu persatu mengeliminasi asosiasi-asosiasi yang tidak signifikan. Seleksi maju dimulai dari model kosong dan menambah asosiasi secara iteratif sampai penambahan tidak lagi signifikan. Metode ALL yang telah ada saat ini mempertimbangkan semua kemungkinan asosiasi untuk menentukan yang mana yang akan ditambahkan atau dieliminasi. Metode tersebut menjadi tidak praktis jika jumlah variabel bertambah besar, karena jumlah asosiasi bertambah secara eksponensial terhadap jumlah variabel.

Sebagai salah satu metode numerik grafis, Chordalysis kemudian membantu ALL dalam menentukan asosiasi mana yang akan dimasukkan ke dalam model log-linear [7]. Chordalysis berfokus pada model log-linear yang *decomposable*, karena model tersebut memiliki beberapa kelebihan, seperti telah dijelaskan pada bab sebelumnya. Metode ini merupakan pendekatan seleksi maju, karena bermula dari graf kosong, lalu secara iteratif menambahkan satu garis yang membuat graf tetap *chordal*.

1) *Pembangkitan Model Kandidat*

Langkah pertama pada tiap iterasi adalah pembangkitan model-model kandidat. Salah satu dari mereka akan menggantikan model terbaik dari iterasi sebelumnya (M^*). Suatu kandidat M^c merupakan M^* yang ditambahi satu garis. Penambahan satu garis tersebut harus membuat graf tetap *chordal*. Untuk menemukan garis tersebut, dilihat konektivitas antara dua titik di ujungnya. Suatu garis (a, b) bisa ditambahkan jika:

- a dan b tidak terhubung (mereka ada dalam komponen yang berbeda), atau
- a dan b terhubung, dengan semua jalur tanpa *chord* memiliki panjang dua.

Karena suatu garis menggambarkan asosiasi antara dua variabel, maka penambahan satu garis relevan dengan penambahan satu asosiasi dua arah (atau orde yang lebih tinggi) pada model log-linear. Salah satu contoh

TABEL I
FILTER ATRIBUT

Filter	Standar deviasi lebih besar dari	*Pair correlation* lebih kecil dari	Jumlah atribut terpilih
1	0,010	0,4	87
2	0,010	0,8	576
3	0,001	0,8	588
4	0,100	0,8	524

TABEL II
REKAPITULASI HASIL CHORDALYSIS PADA TIGA DATA UJI

Data Uji	Jumlah Atribut	Representasi Grafis dari Hasil		
		Jumlah Garis	Jumlah Komponen Terkoneksi	Jumlah Atribut yang Tidak Ditampilkan
D2	576	88	21	467
D3	588	89	21	478
D4	524	85	20	419

TABEL III
HASIL SELEKSI FITUR PADA TIGA DATA UJI

Data Uji	Jumlah Atribut	Jumlah Atribut Terpilih		
		Independen	Komponen Terkoneksi	Total
D2	576	467	43	510
D3	588	478	44	522
D4	524	419	42	461

ditunjukkan pada Gambar 4. Anggap bahwa pada suatu iterasi, diperoleh model [2][12] (Gambar 4a). Suatu kandidat haruslah bersifat *decomposable*. Dengan demikian, Chordalysis tidak membangkitkan seluruh model-model dengan menambahkan satu garis, tetapi hanya beberapa model yang *chordal*, contohnya [12] pada Gambar 4b. Model-model yang tidak hierarkis tidak terpilih sebagai kandidat, karena grafnya tidak *chordal* (Gambar 4c). Proses seleksi ini memperkecil ruang pencarian secara signifikan.

2) Skor Kandidat

Seperti ALL, Chordalysis juga menggunakan statistik G^2 untuk menghitung skor suatu model. Skor G^2 dari suatu model M adalah:

$$G^2(M) = 2 \cdot \sum_{x \in Dom(V)} O_x \ln\left(\frac{O_x}{E_x}\right) \quad (1)$$

Untuk model *decomposable*, (1) dapat disederhanakan menjadi:

$$G^2(M) = 2 \cdot N\left(\sum_{C \in C'} H(C) - \sum_{S \in S'} H(S) - H(V)\right) \quad (2)$$

dengan C' adalah himpunan *clique* maksimal, dan S' adalah himpunan *separator* minimal [8].

Pada Chordalysis, M^c dan M^* bersifat hierarkis dan hanya berbeda pada satu garis. Menurut ALL, perbedaan G^2 mereka dapat dianggap sebagai G^2 tersendiri. Karena itu, untuk mendapatkan signifikansi dari penggantian M^* dengan M^c (M^* vs. M^c), perbedaan G^2 mereka didapatkan dari pengurangan:

$$G^2(M^* \text{vs.} M^c) = G^2(M^*) - G^2(M^c)$$

$$G^2(M^* \text{vs.} M^c) = 2 \cdot N\left(\sum_{C \in C^*} H(C) - \sum_{S \in S^*} H(S) - \sum_{C \in C^c} H(C) + \sum_{S \in S^c} H(S)\right) \quad (3)$$

Jumlah komponen pada (3) akan banyak tereduksi, karena banyaknya pembatalan [7]. Jika terdapat dua model *decomposable* M^* dan M^c yang berbeda hanya pada satu garis (a, b), maka:

$$G^2(M^* \text{vs.} M^c) = 2 \cdot N\left(H\left(S_{ab} \cup \{a\}\right) + H\left(S_{ab} \cup \{b\}\right) - H\left(S_{ab} \cup \{a,b\}\right) - H\left(S_{ab}\right)\right) \quad (4)$$

3) Pemilihan Kandidat

Setelah perhitungan skor, Chordalysis memilih M^c terbaik. Kandidat terbaik tersebut adalah kandidat dengan *p-value* terkecil, yang dapat diketahui dari urutan garis di q. Skor tersebut kemudian dibandingkan dengan suatu nilai ambang. Jika skor terkecil lebih rendah dari nilai ambang, maka M^* diganti dengan M^c, lalu dijalankan iterasi berikutnya. Jika tidak, maka M^* adalah model akhir, yang tidak lagi membutuhkan penambahan asosiasi.

B. Prioritized Chordalysis

Pada Chordalysis, di setiap iterasi, dibangkitkan model-model *decomposable* yang berbeda pada satu garis saja dengan M^*. Dengan kata lain, satu garis ditambahkan pada M^*, yang membuat graf tetap *chordal*. Dalam mencari garis-garis tersebut, dihitung skor tiap garis. Banyaknya perhitungan skor tersebut dapat direduksi dengan mengetahui bahwa skor dari beberapa garis tidak berubah antar iterasi.

Prioritized Chordalysis [9] mengetahui bahwa untuk garis (a, b) yang sama, skornya juga sama, kecuali *separator* minimal S_{ab} berbeda. Dengan demikian, pada suatu iterasi, suatu garis butuh dihitung kembali skornya apabila *separator* minimalnya berubah. Kemudian, untuk menentukan skor yang harus dihitung ulang, seluruh garis ditinjau, kemudian dilihat apakah *separator* minimalnya berubah. Tetapi, untuk ribuan variabel, cara tersebut tidaklah efisien. Untuk mengatasinya, *Prioritized* Chordalysis menggunakan *clique-graph* [10] yang berkaitan dengan graf *chordal*.

C. Data Uji

Data mengenai molekul-molekul antibiotik dan non-antibiotik diperoleh dari antibiotik yang ada di pasaran, dan juga dari MDDR dan Life Chemical Inc. Dari pasar, didapatkan 150 antibiotik. Dari MDDR, didapatkan 2854 antibiotik dan 57179 non-antibiotik. Dari Life Chemical Inc, diperoleh 38907 antibiotik dan 52604 non-antibiotik. Perangkat lunak Dragon [11] digunakan untuk mendefinisikan 4885 atribut tiap molekul. Nilai tiap atribut dihitung menggunakan Corina [12].

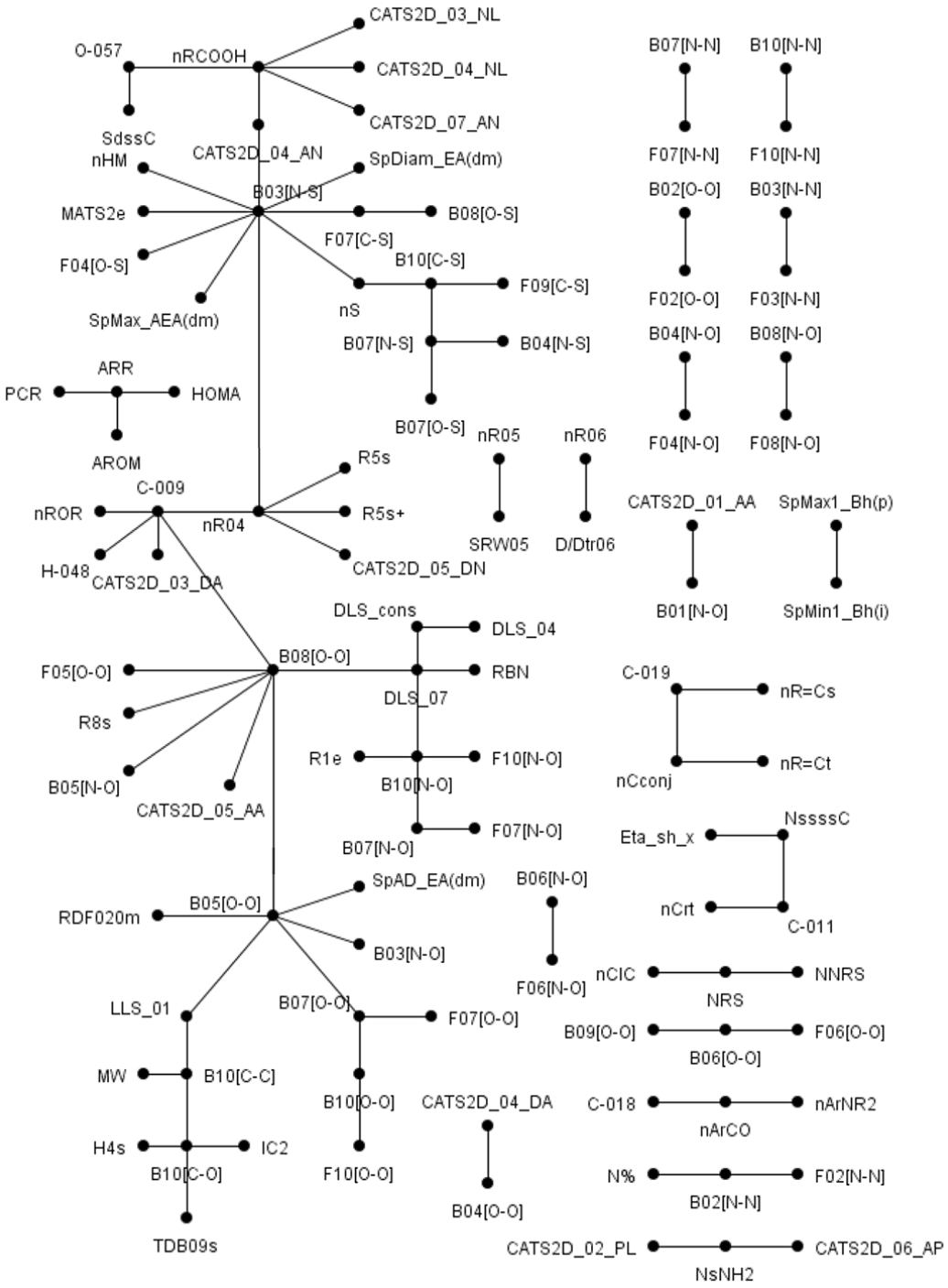
Pertama-tama, jumlah atribut direduksi melalui beberapa prosedur. Atribut yang memiliki *missing value* dihapus. Selain itu, jika terdapat grup atribut yang berkorelasi sempurna, hanya satu dari mereka yang diambil. Kedua prosedur penghapusan atribut ini menghasilkan 4532 atribut.

Dari data yang sudah tereduksi ini, empat filter digunakan secara independen untuk mendapatkan jumlah atribut yang lebih sedikit, dengan memperhatikan standar deviasi dan *pair correlation*. Parameter dan hasil dari filter-filter

tersebut (disebut Filter 1, Filter 2, Filter 3, dan Filter 4) ditunjukkan pada Tabel 1. Untuk penelitian ini, digunakan hasil dari Filter 2, 3, dan 4 dari penelitian sebelumnya, karena hasil filter tersebut memiliki jumlah atribut yang besar, lebih dari 500.

D. Diskritisasi

Beberapa atribut bersifat numerik. Karena ALL dan Chordalysis bekerja pada variabel diskrit, maka atribut-atribut numerik tersebut harus melalui praproses setelah proses filter sehingga mereka menjadi diskrit. Praproses ini dijalankan pada atribut-atribut yang memiliki lebih dari 10 macam nilai. Kemudian, digunakan *equal-width discretization* dari Weka, dengan 10 *bin* sebagai hasil diskritisasi.



Gambar 5. Graf hasil Chordalysis pada D2.

IV. HASIL PENGUJIAN DAN ANALISIS

A. *Penelitian Sebelumnya*

Tujuan dari penelitian sebelumnya [13] adalah menguji hasil dari enam algoritma dalam mengklasifikasikan molekul antibiotik dan non-antibiotik. Keenam algoritma tersebut adalah: Support Vector Machine (SVM) dengan kernel linear, random forest, regresi logistik, gradient boosted trees, naïve Bayes, dan pohon keputusan.

B. *Hasil Chordalysis*

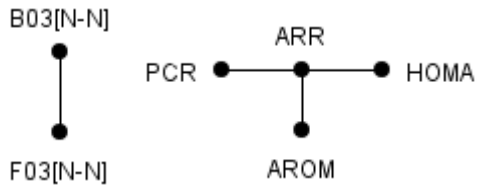
Chordalysis diaplikasikan pada ketiga data uji. Hasil grafis dari data-data tersebut ditunjukkan pada Gambar 5 (untuk D2) dan Tabel 2 (untuk D2, D3, dan D4). Jumlah garis menunjukkan jumlah asosiasi yang ditemukan. Terdapat atribut-atribut yang tidak muncul dalam graf. Sebagai contoh, untuk D2, dari 576 atribut, graf yang dihasilkan memiliki 109 atribut. Hal ini berarti bahwa 467 atribut lainnya tidak memiliki asosiasi. Komponen-komponen terkoneksi memiliki jumlah titik yang berbeda-beda. Komponen terkoneksi terbesar dari D2, D3, dan D4 secara berturut-turut memiliki 58, 59, dan 58 titik, sementara komponen-komponen terkoneksi lainnya memiliki paling banyak 4 titik.

C. *Seleksi Fitur*

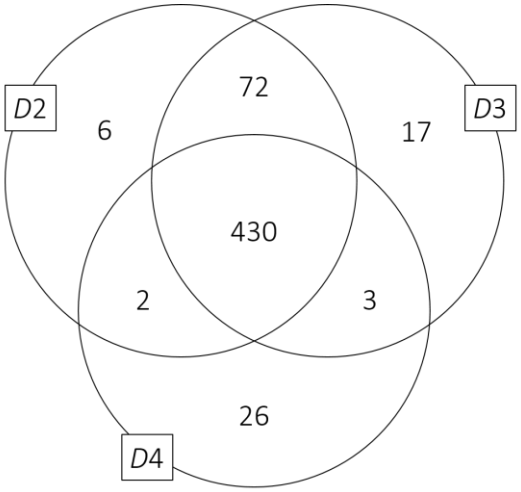
Setelah memperoleh tiga graf yang menggambarkan asosiasi antar atribut, seleksi fitur dijalankan dengan melihat graf-graf tersebut. Hasil dari seleksi ini ditunjukkan pada Tabel 3. Atribut-atribut yang independen (tidak memiliki asosiasi) diambil, karena mereka tidak bisa diwakilkan oleh atribut lain. Jumlah atribut-atribut ini ditunjukkan pada kolom "independen".

Seleksi fitur dijalankan pada tiap komponen terkoneksi. Atribut-atribut yang dihapus adalah atribut yang memiliki hanya satu asosiasi. Dengan demikian, jika ditemui komponen terkoneksi seperti Gambar 6b, atribut yang terpilih adalah ARR. Hal ini berarti bahwa ARR dapat merepresentasikan atribut PCR, HOMA, dan AROM. Di sisi lain, jika suatu komponen terkoneksi hanya memiliki 2 titik, seperti Gambar 6a, satu atribut dipilih dan lainnya dihapus. Jumlah atribut yang terpilih dari proses ini ditunjukkan pada kolom "komponen terkoneksi" pada Tabel 3. Di antara hasil-hasil seleksi fitur dari D2, D3, dan D4 (Gambar 7), terdapat 430 atribut yang terpilih dari ketiga data uji.

Selain itu, ditunjukkan bahwa proses seleksi fitur pada penelitian sebelumnya dapat ditingkatkan. Pada tahap awal dari penelitian tersebut, atribut-atribut yang berkorelasi sempurna dihapus. Setelah mengaplikasikan



Gambar 6. Contoh komponen terkoneksi dengan 2 titik (a) dan lebih dari 2 titik (b).



Gambar 7. Diagram venn atribut-atribut yang terpilih dari D2, D3, dan D4.

Chordalysis, diketahui bahwa masih ada beberapa asosiasi (lebih dari 80). Oleh sebab itu, dari D2, D3, dan D4, secara berturut-turut dihapus 66, 66, dan 63 atribut.

V. KESIMPULAN

Pada penelitian ini, ditunjukkan bahwa hasil dari Chordalysis dapat digunakan untuk seleksi fitur. Gagasan awalnya adalah bahwa seleksi dapat dijalankan dengan menghilangkan atribut yang bisa diwakilkan oleh atribut lain. Dengan demikian, dibutuhkan suatu metode yang dapat menemukan asosiasi antar atribut. Chordalysis dapat melakukan hal tersebut dalam waktu yang layak, dan dari hasil Chordalysis, dapat ditentukan atribut-atribut yang bisa dieliminasi.

Penelitian ini berfokus pada molekul-molekul antibiotik yang terdapat di pasaran. Untuk penelitian selanjutnya, dapat dilakukan seleksi fitur berbasis Chordalysis untuk data uji yang lebih besar, yaitu molekul-molekul antibiotik dan non-antibiotik, termasuk dari MDDR dan Life Chemical Inc.

Chordalysis bekerja pada variabel diskrit dan mengeliminasi variabel-variabel yang memiliki *missing values*. Untuk itu, diperlukan suatu metode yang tepat untuk proses diskritisasi. Selain itu, dibutuhkan juga pengolahan *missing values* agar variabel yang akan dipertimbangkan menjadi lebih beragam.

UCAPAN TERIMA KASIH

Penelitian ini didanai oleh Kementerian Pendidikan dan Kebudayaan Republik Indonesia dan *Centre National de la Recherche Scientifique* (CNRS) Prancis.

DAFTAR PUSTAKA

[1] P. Fernandes, "The global challenge of new classes of antibacterial agents: an industry perspective," *Current Opinion in Pharmacology*, vol. 24, hal. 7–11, 2015.
[2] U. Fayyad, G. Piatetsky-Shapiro, dan P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, hal. 37–54, 1996.
[3] *Probabilistic Graphical Models*, The MIT Press, Cambridge, 2009.
[4] *Using Multivariate Statistics*, edisi kelima, Pearson Education Inc., Upper Saddle River, 2007, hal. 858–912.
[5] *The Analysis of Frequency Data*, University of Chicago Press, Chicago, 1977.
[6] A. Berry dan R. Pogorelnick, "A simple algorithm to generate the minimal separator and the maximal cliques of a chordal graph," *Information Processing Letters*, vol. 111, no. 11, hal. 508–511, 2011.
[7] F. Petitjean, G.I. Webb, dan A.E. Nicholson, "Scaling log-linear analysis to high-dimensional data," *IEEE 13th International Conference on Data Mining*, hal. 597–606, 2013.
[8] F. Malvestuto, "Approximating discrete probability distributions with decomposable models," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 21, no. 5, hal. 1287–1294, 1991.
[9] F. Petitjean dan G.I. Webb, "Scaling log-linear analysis to datasets with thousands of variables," dalam *Proc. SDM15*, Vancouver, Kanada, 2015, hal. 469–477.
[10] P. Galinier, M. Habib, dan C. Paul, "Chordal graphs and their clique graphs," *Graph-Theoretic Concepts in Computer Science, Lecture Notes in Computer Science*, hal. 358–371, 1995.
[11] *Molecular Description for Chemoinformatics*, edisi kedua, Wiley-VCH, Weinheim, 2009.
[12] J. Sadowski, J. Gasteiger, dan G. Klebe, "Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures," *J. Chem. Inf. Comput. Sci.*, vol. 34, hal. 1000–1008, 1994.
[13] J. Hung, "An experiment about the classification of antibacterial molecules", Internal Technical Report, University of Lorraine, Nancy, 2015.