PENELUSURAN PANGKALAN DATA TEKS BIDANG BIOLOGI

Siti Elly Faisholyah

Sub Bidang Dokumentasi dan Informasi
Bidang Jasa dan Informasi
Puslitbang Bioteknologi LIPI

ABSTRAK

Pangkalan data teks bidang biologi molekuler dapat ditelusuri melalui sarana Entrez, Sequence Retrieval System (SRS) dan DBGET. Entrez merupakan pangkalan data terpadu sebagai titik entri penjelajah data secara nyata. Sementara SRS adalah *interface* yang homogen pada lebih dari 80 pangkalan data biologi sehingga user dapat mneghemat waktu dalam penelusurannya. Sedangkan DBGET menyediakan akses sekitar 20 pangkalan data yang dapat langsung menghubungkan informasi yang dibutuhkan oleh user sebagai tambahan pada daftar hasil penelusuran yang dilakukannya.

PENDAHULUAN

Jumlah informasi biologi yang dapat diperoleh melalui *World Wide Web* (WWW) sangat banyak, karena volume datanya bertambah dengan sangat cepat. Para ilmuwan selalu mengharapkan dapat memperoleh informasi yang dibutuhkannya dengan mudah dan cepat sebagai pembanding terhadap penelitian yan pernah dilakukannya. Penelusuran data biologi pada *WWW* di atas dapat dilaksanakan dengan menggunakan kata-kata deskriptif untuk menelusur pangkalan data teks atau dengan menggunakan "sequens protein" untuk menelusur pangkalan data khusus.

Ada 3 sarana yang penting dalam proses penelusuran data yang dimaksud yaitu *Entrez, Sequence Retrieval System* (SRS) dan *DBGET.* Ketiga sarana tersebut dapat dipakai dalam kegiatan penelusuran pangkalan data teks biologi molekuler. Selain itu sarana-sarana tersebut dapat juga dipakai untuk melakukan hubungan informasi yang relevan dalam proses entri data yang cocok dalam proses penelusuran. Meskipun banyak pangkalan

data yang dapat diakses melalui proses penelusuran teks lengkap (*full-text*), strategi penelusuran melalui sarana di atas diharapkan dapat diterapkan secara lebih luas dan dapat digunakan untuk menelusur sumber-sumber organisma penting, seperti Saccharomyces Genome Database (SGD) dan Mouse Genome Database (MGD).

Sistem temu kembali tersebut sangat penting bagi para ilmuwan dalam menelusur informasi. Pemakaian sistem yaitu dengan cara memasukkan bilangan tertentu dari suatu sequens yang terbit atau cakupan pangkalan data penelusuran melalui istilah-istilah khusus. Keuntungan sarana Entrez, SRS dan DBGET adalah bahwa sistem temu kembali tersebut tidak hanya mencocokkan kembali sebuah permasalahan, tetapi juga menyediakan petunjuk yang mudah digunakan sebagai informasi tambahan penting yang berhubungan dengan pangkalan data. Kinerja ketiga sistem pangkalan data yang dimaksud berbeda antara yang satu dengan yang lainnya, namun seluruhnya dapat dipakai untuk menelusur berbagai informasi yang berhubungan dengan sumber-sumber informasi lainnya.

Konsep Penelusuran

Penelusuran dengan Boolean logic - Teknik ini dipakai untuk menelusur informasi yang menggunakan dua istilah atau lebih pada saat penelusuran. Istilah-istilah tersebut dikombinasikan dengan mempergunakan operator DAN (operator default), ATAU dan TIDAK.

Penelusuran secara luas – Teknik ini dipakai untuk menelusur informasi sesuai dengan keinginan pemakai (*user*). Penelusuran dilakukan dengan melakukan perulangan kata, apabila istilah yang dicari tidak dikenali oleh pangkalan data yang digunakan.

Penelusuran secara terbatas - Apabila hasil penelusuran menghasilkan terlalu banyak temuan, maka user harus merubah atau menambah istilah-istilah lain yang lebih relevan.

Penelusuran proximity – Cara ini dipakai untuk menelusur berbagai kata atau istilah, dengan menempatkan kata secara berulang di seputar istilah yang dicari.

Wird card - karakter * dapat ditambahkan di akhir kata untuk menelusur suatu istilah dalam proses penelusuran, agar penelusuran tersebut menjadi lebih spesifik. Sebagai contoh, untuk mencari seluruh nama pengarang yang memiliki nama akhir Zau, penelusuran dilakukan dengan menggunakan Zau*.

Entrez

Entrez, merupakan pangkalan data dan sistem temu kembali dalam bidang biologi molekuler, telah dikembangkan oleh National Center for Biotechnology Information (NCBI). Pangkalan data tersebut merupakan titik entri untuk menjelajah secara nyata, juga merupakan pangkalan data terpadu. Sistem Entrez menyediakan akses pangkalan data pada nukleotid dan pangkalan data sequens protein, sebuah pangkalan data contoh terstruktur 3-D (MMDB), suatu pangkalan data genom, peta dan literatur. Pangkalan data literatur (PubMed), menyediakan akses yang sangat baik dan mudah pada artikel-artikel di MEDLINE dan pra-MEDLINE (tidak diindeks secara penuh). Pangkalan data taksonomi tersebut berisi lebih dari 23.000 spesies yang berbeda dan memungkinkan temu kembali sequens protein dan DNA untuk setiap kelompok taksonomi.

Dari ketiga sistem pangkalan data teks di atas, Entrez merupakan yang termudah untuk digunakan, tetapi pangkalan data tersebut menyajikan informasi penelusuran secara lebih terbatas.

Contoh penelusurannya berdasarkan teks

Tujuan

Untuk menemukan homolog manusia dari gen *Drosophila per.*

Prosedur

Di Entrez, mulailah dengan menelusur "All Fields" dalam pangkalan data nukleotid. Masukkan istilah *human per* dalam boks teks. Penelusuran ternyata menghasilkan tiga temuan, dimana tidak seorangpun dapat melakukan sesuatu dengan homolog manusia dari gen *Drosophila per*. Hal ini ditemukan karena istilah tersebut lebih membahas topisomerase manusia, dimana *Per* merupakan nama akhir salah satu pengarangnya. Kemudian, penelusuran harus 'meluas' dengan menggunakan istilah period daripada per. Period yang dimaksud lebih menggambarkan istilah yang lebih luas tentang fungsi gen dimana hal ini berlawanan dengan per yang merupakan nama gen yang aktual dalam *Drosophila*. Hasilnya ternyata ditemukan sebanyak 24 temuan, namun sejumlah 13 temuan diperoleh apabila penelusuran tersebut dibuat dengan istilah period dan drosophila. Hasil tersebut kemudian dapat dilihat atau penelusuran kemudian dapat dibatasi dengan mencantumkan ketiga istilah : drosophila, human dan period, sebagaimana homolog manusia dari gen periode *Drosophila* yang ditelusur. Proses ini menghasilkan 3 temuan, dimana seluruh temuan menunjukkan bahwa permintaan mulai ditemukan. Gene dalam humans hasil temuan tersebut kemudian disebut sebagai RIGUI.

Penelusuran dapat diulang dengan menggunakan SRS dan DBGET melalui *search engines* yang berbeda. Berdasarkan sarana SRS ternyata ditemukan dua cantuman. Sementara dengan menggunakan DBGET, penelusuran akhir tidak menghasilkan temuan dalam pangkalan data nukleotid, dan penelusuran lain tidak menemukan homolog manusia, meskipun pada kenyataannya entri RIGUI manusia dapat ditemukan secara langsung dengan menggunakan perintah BGET. Masalah yang timbul adalah bahwa tidak semua cakupan subjek pangkalan data tersedia untuk penelusuran.

Penelusuran dapat dilakukan dalam setiap pangkalan data dengan memasukkan cantuman yang sesuai dengan kehendak *user*. Berbagai cantuman (yang berdekatan) yang ditelusur dalam pangkalan data tersebut dapat disusun (berhubungan) dalam pangkalan data Entrez lain, sehingga informasi yang diinginkan ditemukan kembali dengan mudah. Hubungan tersebut kemudian disambungkan dengan pangkalan data eksternal seperti Online Inheritance in Man (OMIM) dan MGD. Istilah-istilah yang berdekatan dan berhubungan lalu didaftar dalam susunan yang sama. Kesamaan ini didasarkan pada analisis pra-perhitungan sequens, struktur dan literatur. Sebagai contoh, dalam hal sequens, analisis pra-penghitungan merupakan hasil penelusuran BLAST.

Salah satu bagian penting dalam Entrez adalah kemampuannya untuk menemukan kembali serangkaian data yang didasarkan pada beberapa kriteria dan untuk proses simpan datanya *(download)* pada komputer lokal, memungkinkan sequens tersebut untuk dipekerjakan pada alat analitis yang tersedia di komputer tersebut. Hal ini dikenal dengan sejumlah Entrez yang memungkinkan proses temu kembali DNA atau sequens protein khusus yang sejajar. Selain itu, seluruh entri untuk organisma khusus berdasarkan pangkalan data taksonomi dapat ditemukan kembali, atau suatu penelusuran Boolean dapat dicantumkan untuk mendeskripsi proses penelusuran guna menemukan kembali sequens tersebut sesuai dengan kehendak *user*.

SRS

SRS adalah *interface* yang homogen pada lebih dari 80 pangkalan data biologi yang telah dikembangkan di European Bioinformatics Institute (EBI) di Hinxton, UK. Jenis-jenis pangkalan data yang tercakup adalah sequens dan sequens yang berhubungan, jalan kecil metabolik, faktor-faktor transkripsi, hasil-hasil

genom, pemetaan, mutasi umum dan mutasi khusus. Isi pangkalan data ini dapat diakses secara luas oleh user. Halaman Web SRS mendaftar seluruh pangkalan data yang berhubungan untuk sebuah halaman deskripsi tentang pangkalan data dan mencakup tanggal terakhir kali diperbaharui. Lebih dari 30 versi SRS yang telah diluncurkan dalam *World Wide Web (WWW)*. Setiap SRS meliputi *subset* pangkalan data yang berbeda dan perangkat analitis yang terhubung.

Meskipun terdapat beberapa pangkalan data potensial untuk ditelusur, pangkalan data SRS telah diindeks dengan baik, dengan demikian user dapat mneghemat waktu penelusuran. Isi cakupan data dalam setiap pangkalan data dibagi dalam komponen dan kata-kata yang terpilih yang disarikan dan disisipkan ke dalam indeks. Setiap cakupan pada umumnya memiliki indeks tersendiri. User dapat menelusur seluruh cakupan subjek dengan menggunakan pilihan "all text," SRS juga menyajikan bentuk permintaan alternatif yang diinginkan oleh user dengan menggunakan fasilitas Boolean yang lebih kompleks.

DBGET

DBGET/LinkDb adalah suatu sistem temu kembali pangkalan data terpadu yang dikembangkan oleh Institute for Chemical Research, Kyoto University dan Human Genome Center of the University of Tokyo dan tersedia melalui GenomeNet. DBGET menyediakan akses pada sekitar 20 pangkalan data. DBGET langsung dapat menghubungkan informasi yang diinginkan user, sebagai tambahan pada daftar hasil penelusuran. Pangkalan data LinkDB juga dapat ditelusur secara langsung dengan entri khusus dan menyediakan daftar hubungan pada seluruh entri pangkalan data dengan seluruh informasi yang dimilikinya. Hal yang unik dari DBGET yaitu memiliki hubungan dengan pangkalan data Kyoto Encyclopedia of Genes and Genomes (KEGG), dimana pangkalan data ini merupakan suatu pangkalan data metabolik yang diatur dan dikembangkan oleh kelompok yang sama.

DBGET lebih sederhana, namun metode penelusurannya lebih terbatas dibandingkan SRS atau Entrez. Untuk DBGET, user dapat menelusur pangkalan data pilihan dengan menggunakan salah satu dari dua perintah. Perintah *bfind* memungkinkan penelusuran berdasarkan istilah teks. Sebagai jawaban atas hal ini, istilah yang cocok dengan permintaan user akan disajikan secara bersama-sama. Sedangkan perintah *bget* ditelusur melalui nama entri atau nomor tambahan.

Contoh

Setiap sistem dalam penelusuran dengan DBGET dapat digunakan untuk menemukan P04391 SWISS-PROT, suatu protein ornithine carbamoyltranferase dalam *Escherichia coli*. Sedangkan dengan Entrez, user dapat memasukkan nama P04391 dalam bentuk permintaan pangkalan data protein dan melihat entri serta menyusun informasi lain yang berhubungan. Sementara pada SRS, pertama kali user dapat memilih pangkalan data SWISS-PROT, kemudian memasukkan P04391 dalam bentuk permintaan dan sekali waktu entri/istilah dicantumkan lagi selalu dipamerkan, menelusur berbagai hubungan pada pangkalan data lain yang terkait. Meskipun demikian, cara tercepat untuk mengumpulkan informasi yang berhubungan dengan entri ini adalah dengan menelusur melalui Link DB. Dengan memasukkan Swiss-Prot:P04391 secara sederhana, daftar seluruh hubungan pada seluruh pangkalan data yang terkait dapat dipamerkan.

Penelusuran berdasarkan teks tergantung pada kualitas data, anotasi dan indeks yang ditelusur. Jika entri tidak dianotasi secara penuh atau secara konsisten, akan sulit menemukan seluruh entri yang relevan untuk suatu topik. Teks dapat berbentuk kosa kata bebas atau kosa kata terkendali, masing-masing dapat mengakibatkan masalah yang berbeda untuk suatu

akibatkan masalah yang berbeda untuk suatu penelusuran. Sebagai contoh, jika menelusur teks bentuk-bebas, kesalahan ejaan dalam teks dapat menghasilkan entri hasil penelusuran yang relevan. Ketidak konsistenan juga terdapat dalam tanda penghubung, misalnya. Pada suatu tempat, frase mungkin mengandung tanda penghubung, tetapi di lain tempat tidak; dengan demikian entri akan hilang pada saat penelusuran. Masalah potensial lainnya yang perlu disadari oleh user adalah keterbatasan penelusuran melalui istilah yang terdapat dalam "Kata kunci". Oleh sebab itu user perlu mengetahui apakah kata kunci yang dicantumkan sesuai dengan topik permasalahan yang sedang dicari. User perlu berhati-hati pada fasilitas indeks kata kunci yang memiliki bilangan penting, yang menunjukkan hanya satu entri, atau kata kunci yang menunjukkan tambahan pengarang, yang dapat menghasilkan informasi yang berbeda dan bahkan tidak bermakna sama sekali. Jika hal ini terjadi, maka sebaiknya user melakukan penelusuran 'teks bebas'. Apabila user ingin menelusur subjek dengan kosa kata terkendali (contoh MeSH), maka user harus mengerti dengan benar organisasi dan hirarkinya.

Disadur dari :

Lewitter, Frank. 1998. "Text-based database searching". *Trends Guide to Bioinformatics.* (Supplement): 3-5.

--ooo0ooo--