Bioinformatika dan Pattern Recognition

(Laporan Kenkyukai Pattern Recognition & Media Understanding)

Anto Satriyo Nugroho
asnugroho@ieee.org
http://www.asnugroho.net

Banjir merupakan bencana alam yang seringkali menimbulkan kerugian harta benda, bahkan jiwa Masih teringat di benak kita bagaimana banjir besar yang melanda Jakarta menyebabkan kegiatan perekonomian di ibukota mati, sehingga menimbulkan kerugian bermilyar rupiah. Tulisan ini akan membahas juga mengenai banjir, tetapi bukan banjir karena lupaan air. Maksud banjir dalam tulisan ini adalah derasnya arus informasi dan data biologi, terutama pasca genome-project. Tulisan ini merupakan sebuah catatan diskusi dalam sebuah seminar yang bertemakan bioinformatika.

Salah satu keyword yang menjadi sangat populer pada era ini adalah bioinformatika. Sebagai suatu disiplin ilmu, bioinformatika melibatkan dua aspek, yaitu aspek teknologi informasi (TI) dan aspek biologi [1][2]. Aplikasi dari bioinformatika ini meliputi berbagai bidang, antara lain bidang farmasi, kedokteran dan pertanian. Bioinformatika merupakan suatu bidang yang melibatkan berbagai metode analisa, sehingga dalam melakukan penelitian di bidang ini, kuantitas dan kualitas data menjadi aspek penting.

Berkaitan dengan arus data yang deras mengalir ini, akan menarik untuk mencermati perbandingan perkembangan pesat semikonduktor dan genetika. Di bidang komputer, dikenal Moore's law yang memprediksi bahwa setiap 18 bulan, jumlah transistor per satuan area pada IC, selalu berlipat dua. Dengan kata lain, kemampuan komputer akan berlipat jadi dua kali setiap 18 bulan. Pengamatan Gordon Moore ini dikeluarkan tahun 1965, dan secara ajaib selalu terbukti berlaku, setidaknya selama dua dekade ini. Hal ini menjadi motivasi dan misi Intel Corporation untuk selalu memenuhi tuntutan dari ramalan Moore [3].

Analog dengan Moore's law, di dunia biologi, dikenal juga ramalan menarik dari Prof. R. Dawkins (Oxford University), yang mencoba menarik korelasi antara jumlah nucleotide-base yang bisa dibaca dengan dana £1000, terhadap waktu. Pada tahun 1965, diperlukan £1 untuk membaca 1 huruf pada RNA bakteri. Tahun 1975, £10 untuk satu huruf pada virus code. Pada 1985, membaca 1

huruf pada nematode memerlukan £1, dan pada tahun 2000 diperlukan 0.10 £ untuk membaca 1 huruf pada human genome project. Kalau ramalan ini benar, maka pada tahun 2012 diprediksi dengan £1000 dapat dianalisa E.coli yang terdiri dari 200 ribu bases. Sedangkan pada tahun 2050, diperlukan £1000 untuk membaca seluruh nucleotide base pairs manusia. Prediksi ini dikenal sebagai "Son of Moore's law for genetics", menggambarkan perkembangan yang pesat di dunia genetika [4].

Berangkat dari ketersediaan data genome dalam jumlah besar ini, terminologi *biological-datamining* menjadi sangat populer. Datamining didefinisikan sebagai proses otomatis mengekstrak suatu informasi dari sekumpulan data yang berjumlah besar. Salah satu aplikasi dari penerapan datamining di bioinformatika ini adalah pengembangan industri farmasi dan kedokteran. Informasi yang diekstrak ini dapat dimanfaatkan dalam industri medis, misalnya menekan resiko timbulnya efek samping dari terapi kanker.
Bagaimana cara ekstraksi informasi tersebut dilakukan ? Berbagai metode dikenal dalam Data Mining dan Knowledge Discovery. Diantaranya memakai metode softcomputing seperti artificial neural network, fuzzy, genetic algorithm. Melihat perhatian yang semakin meningkat di bidang ini, Pattern Recognition and Media Understanding Technical Group (PRMU), menyelenggarakan kenkyukai (seminar) dengan tema "Bioinformatics dan Pattern Recognition". PRMU yang merupakan satu bagian dari IEICE (The Institute of Electronics, Information, and Communication Engineers), Jepang. Organisasi ini menyelenggarakan kegiatan rutin bulanan yang disebut dalam bahasa Jepang "kenkyukai". Kenkyukai merupakan wahana peneliti di Jepang untuk saling bertukar ide dan berdiskusi mengenai perkembangan penelitian terakhir.

Pada bulan Juni 2003, kenkyukai diselenggarakan selama dua hari di Chiba University, yang berlokasi sekitar 1 jam dari Tokyo. Beberapa tema cukup menarik dibahas dalam event tersebut, diantaranya adalah pemakaian Hidden Markov Model untuk estimasi arah dan posisi sekuens, prediksi Operon memakai Bayesian Hierarchical model, dan feature selection diaplikasikan untuk analisa ekspresi gen. Peserta yang hadir sekitar 25 orang, sebagian besar berlatarbelakang computer science. Hanya sedikit yang berasal dari kalangan bio. Dalam diskusi sangat sedikit pertanyaan kritis diajukan dari aspek biologi. Terlepas dari kesibukan yang mungkin menjadi kendala utama bagi para peneliti, fenomena di atas mungkin juga terjadi karena tema yang dipresentasikan sangat spesifik ke arah TI, dan kurang menarik bagi peneliti yang berlatarbelakang biologi.

Secara umum, diskusi dalam kenkyukai ini membahas masalah-masalah dalam biologi dari sudut ilmu komputer, dan mencoba menawarkan solusi dari sudut pandang pattern recognition. Dalam seminar tersebut, seorang peneliti NTT yang juga dikenal sebagai salah seorang pakar di bidang Pattern Recognition, Dr. Eisaku Maeda, menyampaikan special lecture dengan judul "Gene expression analysis and Feature Selection". Menurutnya, setidaknya ada tiga hal dalam study microarray yang merupakan bahasan menarik ditinjau dari sudut pattern recognition.

1. Mungkinkah dengan data microarray, kita melakukan prediksi suatu kategori (misalnya klasifikasi jenis penyakit) ? Topik ini dapat dipecahkan dengan teknik supervised learning.

2. Kalau prediksi tersebut memungkinkan untuk dilakukan, berapakah kira-kira tingkat akurasi yang mungkin dicapai ? Pertanyaan ini merupakan masalah yang sangat penting, karena menjadi referensi bagi praktisi medis dalam clinical application. Dari sudut pattern recognition, hal ini merupakan masalah setting parameter yang tepat dalam mendesign suatu metode pattern recognition.

3. Penentuan kandidat gen yang memiliki potensi kedokteran/farmasi , bukan hanya penting dari sudut biologi, akan tetapi juga diperlukan untuk mengembangkan terapi medis yang murah dan efektif.

Selanjutnya Dr. Maeda menggarisbawahi tiga masalah yang wajib diperhatikan oleh seorang peneliti dalam melakukan analisa ekspresi gen:

1. Tingkat reliabilitas hasil pengukuran yang rendah

2. Jumlah sample yang sangat minim dibandingkan dimensi dari input vektor (jumlah gen)

3. Preprocessing data yang tepat diperlukan untuk mengantisipasi kelemahan data tergantung pada protokol pengukuran.

Pada seminar tersebut Dr. Maeda memperkenalkan metode *feature selection* yang dipakai dalam studinya. Feature Selection adalah salah satu metode pengolahan awal data (preprocessing), untuk menentukan subset feature yang merupakan kombinasi input informasi yang terbaik untuk diolah pada tahap berikutnya. Kalau input informasi tersebut berupa vector dengan dimensi N, maka banyaknya kombinasi yang memungkinkan dari input vector adalah 2^N. Vektor yang berdimensi 3, memiliki 8 kemungkinan kombinasi. Pencarian kombinasi yang optimal dengan melacak satu persatu disebut *exhaustive searching*. Akan tetapi pendekatan ini tidak realistis untuk input vector yang berdimensi besar, seperti sel manusia yang terdiri dari sekitar 30 ribu gen. Diskusi mengenai feature selection dapat dilihat antara lain pada referensi [5] dan [6].

Dalam studinya, Dr. Maeda memperkenalkan metode yang memilih secara acak sejumlah gen, untuk kemudian dievaluasi kualitasnya. Evaluasi ini diukur dengan classification rate Support Vector Machine dan k-Nearest Neighbour classifier. Proses ini dilakukan berulang-ulang dan hasil akhir merupakan rata-rata dari percobaan yang dilakukan. Terlepas dari minimnya jumlah sample yang dipakai, percobaan ini menunjukkan hasil yang cukup memuaskan dengan tingkat akurasi sekitar 90%.

Mengacu pada taxonomi metode feature selection yang dibuat oleh Prof. Anil K. Jain (Michigan State University) [6], metode Dr.Maeda bersifat stochastic single-solution. Disebut stochastic karena sifat seleksi yang random ini akan menghasilkan subset kombinasi gen yang berubah-ubah untuk tiap kali eksperimen. Disebut single-solution, karena proses tersebut menghasilkan satu kombinasi saja untuk tiap eksperimen. Pendekatan ini mirip dengan pemakaian genetic algorithm untuk feature selection. Hanya saja genetic algorithm menawarkan multi-solution yang dicerminkan oleh variasi kombinasi feature pada generasi terakhir algoritma tersebut.

Selain itu, dalam diskusi yang dilakukan di Chiba, terlihat bahwa banyak sekali masalah di bioinformatika yang sebenarnya bukan masalah baru bila ditinjau dari sudut Pattern Recognition. Selama ini metode tersebut dipakai juga di Pattern Recognition, hanya saja dengan objek lain yang memiliki karakteristik masalah yang serupa. Diperkirakan hal ini mungkin terjadi dikarenakan selama ini di antara para kalangan Bio dan kalangan TI kurang terdapat komunikasi. Mungkin sudah ditakdirkan, bahwa bioinformatika akan menjadi jembatan awal komunikasi dua dunia yang berbeda ini: biologi dan komputer, untuk saling mengisi dalam menguak tabir rahasia alam mikrokosmos yang berada dalam diri manusia.

Dari diskusi mengenai analisa data microarray, dapat ditarik beberapa kesimpulan. Pertama, penemuan microarray memberikan peluang untuk menganalisa ekspresi ribuan gen dalam satu waktu, dan membuka harapan besar aplikasinya di bio-industri. Kedua, analisa ekspresi gen dari data microarray saat ini menjadi salah satu tema menarik bagi para peneliti, terutama pembahasan dari sudut pattern recognition. Hal ini karena masalah analisa ekspresi gen memiliki peluang untuk ditinjau dari aspek yang sangat luas. Mulai dari metode statistical pattern recognition seperti Bayesian classifier, Hidden Markov Model, hingga metode-metode softcomputing seperti artificial neural network, fuzzy clustering, genetic algorithm. Kesimpulan ketiga menyoroti salah satu kelemahan utama dalam studi ini, yaitu minimnya jumlah sample dibandingkan dengan dimensi dari ruang vector yang dibentuk oleh ekspresi gen. Hal ini mengakibatkan kesimpulan yang diambil akan menjadi sangat sulit untuk dinilai validitasnya dari sudut pandang biologi.

Kesimpulan ketiga di atas akan menggiring kita pada suatu pertanyaan: berapakah jumlah sample yang diperlukan ? Sangat sulit untuk menentukan berapa minimal sample yang diperlukan untuk

memenuhi syarat dalam melakukan analisa ekspresi gen. Yang pasti, semakin banyak semakin baik. Akan tetapi hal ini berimplikasi pada diperlukannya dana yang sangat besar dalam penyediaan data eksperimen. Dalam beberapa paper, seperti eksperimen Golub [7], Khan [8] dsb. disebutkan bahwa jumlah sample yang dipakai berkisar 100 sample. Sedangkan jumlah sample yang dipakai dalam eksperimen Dr. Maeda berkisar antara 70-80 sample. Jumlah ini tentu saja jauh dari nilai minimal yang diperlukan untuk membuat suatu kesimpulan yang valid, tapi saat ini perkembangan riset pada tema ini memang baru sampai tahap tersebut.

Dalam satu studi teoritik yang berkaitan dengan jumlah sample dan dimensional data (d), setidaknya disyaratkan lebih dari 2(d+1). Kalau jumlah gen manusia sekitar 30 ribu, maka jumlah sample yang disyaratkan setidaknya menjadi sekitar 60 ribu [9]. Kalau biaya yang diperlukan untuk membaca ekspresi satu sample dari sel manusia, misalnya sekitar 800 USD, maka dapat dibayangkan betapa mahalnya eksperimen analisa data microarray [10]. Sebagai perbandingan, saat saya melakukan eksperimen dengan handwriting character recognition, untuk tiap character dipakai sekitar 1000 sampai 5000 sample. Padahal dimensional dari tiap data yang diolah hanya berkisar antara 100 sampai 800.

Untuk memecahkan masalah ini ada dua alternatif: (1) Menekan cost yang diperlukan untuk melakukan analisa microarray, dan; (2) mengembangkan perangkat lunak (metode pattern recognition) yang dapat mengolah data dengan jumlah sample relatif kecil dibandingkan dimensionalnya. Dalam konteks ini, saya bekerja sama dengan beberapa peneliti di Jepang dan di Indonesia, tengah melakukan penelitian mengenai potensi metode-metode softcomputing (a.l. artificial neural network, genetic algorithm, fuzzy) dan support vector machine untuk mencoba menjawab permasalahan di bidang bioinformatika.

Sebagai penutup bahasannya, Dr. Maeda memberikan analogi yang menarik. Setelah banjir yang biasanya menimbulkan dampak negatif bagi umat manusia, fenomena alam ini biasanya membawa juga dampak positif, misalnya bagi kesuburan tanah dsb. Harapan kita, arus deras data pada era pasca genome-project ini juga berakhir dengan happy ending, yaitu dengan ditemukannya berbagai macam metode dan aplikasi bagi kesejahteraan bagi umat manusia.

Referensi

1. Diskusi mengenai bioinformatika di milis biotek@yahoogroups.com antara lain pada bulan Mei 2003
2. A.B.Witarto, "Bioinformatika: Mengawinkan Teknologi Informasi dengan Bioteknologi", artikel di http://www.ilmukomputer.com/
3. Moore's Law http://www.intel.com/research/silicon/mooreslaw.htm
4. http://myweb.tiscali.co.uk/royalphil/rps/summaries/evolution.htm
5. A.K. Jain, D. Zongker,``Feature Selection: Evaluation, Application, and Small Sample Performance", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.19, No.2, pp.153-158, 1997
6. A. S. Nugroho, “Information Analysis using Softcomputing: The Applications to Character Recognition, Meteorological Prediction, and Bioinformatics Problems" , Doctoral Thesis, Grad.School of Engineering, Dept. of Electrical & Computer Engineering, Nagoya Institute of Technology, Japan, January 2003
7. T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science, Volume 286, 1999.
8. J. Khan et al., ‘‘Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine Vol.7, pp.673-679, 2001.
9. K. Ishii, N. Ueda, E.Maeda and H. Murase, "Introduction to Pattern Recognition", Ohmsha, 1998(Japanese ed.)
10. Contoh harga analisa satu sample RNA, dapat dilihat di http://tcag.bioinfo.sickkids.on.ca/microarray.html