게놈 프로젝트를 위한 생물정보학

생물학연구정보센터 김양석

최근 많은 분자 생물학 기술의 발달과 급속히 진행되고 있는 Genome Project를 통해 대량의 유전자 서열 정보 및 새로운 형태의 생물학 정보들이 산출되고 있다. 이러한 생물학 정보의 양과 질의 변화를 통해 생물학은 'gene era'에서 'whole genome era'로의 새로운 paradigm shift가 일어나고 있다. 이러한 변화를 유도하고 새로운 많은 개념의 많은 지식들을 창출하는 과정에서 가장 급부상하고 있는 생물학 분야는 bioinformatics이다.

Bioinformatics는 통계적 이론, 전산 기술들을 이용하여 생물학 정보들을 저장, 분석, 및 해석을 수행하는 학문이다. 게놈 프로젝트를 통해 산출되는 대량의 유전자 서열 정보 및 새로운 형태의 생물학 자료(서열, image 등)들의 산출은 기존의 저장 및 분석 방식으로는 처리가 불가능하기 때문에 초기의 bioinformatics 연구는 대량의 정보 저장을 위한 데이터베이스의 개발이 중요한 문제로 대두되었다. 또한 실험실에서 새로이 밝힌 유전자의 기능을 예측하기 위해 데이터베이스에 저장된 서열들과의 유사성을 분석하는 도구들이 많이 개발되었고, 그 대표적인 예로 FASTA와 BLAST를 들 수 있다[1,2]. 또한 서열의 특성 및 진화적 관계를 파악(gene prediction, multiple alignment, secondary structure prediction, phylogenetic analysis 등)하기 위한 많은 알고리듬 및 프로그램들이 개발되었다[3](그림 1).

하지만 human 및 model organism들의 full-genome sequencing을 목표로 하는 1단계 human genome project가 성공리에 마치게 되고 게놈 프로젝트의 전체적 방향이 'Full-genome sequencing'에서 'Functional Genomics'로 바뀌게 됨에 따라 bioinformatics의 연구 방향도 바뀌고 있다. 즉, 특정 개체의 full genome sequence와 관련된 정보들이 밝혀짐에 따라, 생물학, 의학, 약학 연구자들은 이미 유전 정보의 홍수에 직면하여 몇 개의 큰 문제들-새로이 밝혀진 유전자들의 기능은 무엇인가, 새로운 유전자들은 어떻게 발현되어지고 통제되어지는가?, genome 구조의 비밀은 무엇인가?-을 풀기 위해 노력하고 있다. 생물정보학자들 또한 이러한 문제를 풀기 위한 핵심적인 기술들-연구 정보들의 효율적 통합과 검색, 새로운 분석 알고리듬의개발, genetic network의 구성 및 이해, data mining을 통한 새로운 생물학 지식의 창출-을 개발하기 위해 많은 노력들을 기울이고 있다. 또한 이러한 연구 결과들은 급속히 성장하고 있는

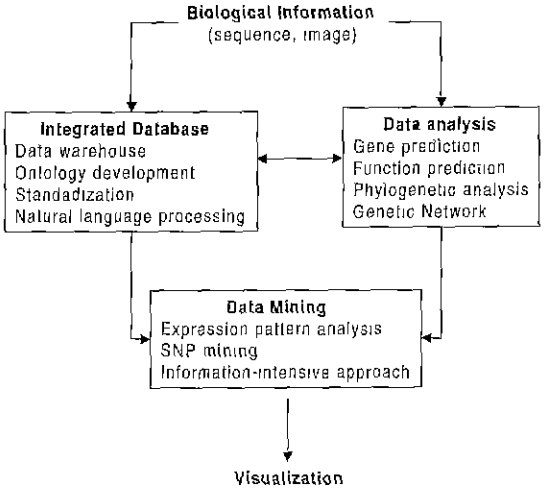

그림 1 Genome 분석을 위한 Bioinformatics 기술들

생물 산업에 직접적으로 응용되어 신약 개발, 유전병의 진단 등 많은 관련 산업들의 발전에 기여하고 있다. 본 논문에서는 post genome 시대를 맞이하여 발전하고 있는 bioinformatics의 동향과 그 발전 가능성에 대해 기술하고자 한다.

1. 연구 정보들의 효율적인 통합과 검색

현재 GenBank, EMBL 등 1차 유전자 정보를 가진 데이터베이스를 포함하여 인터넷을 통해 이용할 수 있는 생물학 데이터베이스의 수는 500여 개 정도로 알려져 있다. 하지만 생물학자가 특정 주제에 관련된 모든 정보를 얻기 위해 500여 개의 데이터베이스를 모두 검색한다는 것은 불가능하다. 특히 functional genomics의 연구를 위해서는 많은 유전자에 대한 관련 정보를 자동으로 추출하는 기술들이 필수적인데 각각의 포맷이 다른 데이터베이스로부터 이러한 작업을 수행한다는 것은 불가능하다.

현재 여러 전산 기술을 이용하여 데이터베이스의 통합 검색을 위한 시도가 이루어 지고 있다. 그 대표적인 예는 EBI(European Bioinformatics Institute)에서 구성한 SRS(Sequence Retrieval System)를 들 수 있다[4]. SRS는 전 세계에서 구축한 50여개의 데이터베이스로부터 데이터들의 Link 링크 정보들을 추출한 후 인덱싱 과정을 거쳐 다시 저장함으로써 빠른 속도로 특정 keyword키워드에 대해 검색을 수행 할 수 있다. 이러한 방식의 또 다른 데이터베이스로는 일본 Kyoto 대학의 Kanehisa 박사 team팀이 구축한 DBGet 시스템을 들 수 있다[5]. 하지만 이러한 데이터베이스의 연결들도 결국은 데이터베이스 상호 link링크를 이용한 것이므로 총체적인 정보를 얻기 위해서는 많은 한계를 가진다. 현재 이러한 한계를 극복하기 위해 데이터베이스 및 검색 프로그램들간의 연결을 매개해 주는 미들웨어 시스템을 구축하여 전체 데이터베이스를 좀 더 효율적으로 통합하고자 하는 노력들이 시도되고 있으며, 대표적인 방법은 CORBA(Common Object Request Broker Architecture)를 이용한 데이터베이스의 통합이다[6]. 다른 접근 방법으로는 data warehousing과 data muning 기법을 도입하여, 특정 주제에 관련된 총체적인 정보들을 여러 데이터베이스로부터 수집한 후 새로운 통합 데이터베이스를 구축하는 방법도 사용되고 있으며, 인간의 질병에 관련된 총체적인 정보를 수집하여 만든 Weizmann 연구소의 GeneCard를 예로 들 수 있다[7]. 또한 데이터베이스의 통합보다는 특수한 연구 목적을 위해 1차 정보의 재가공을 통한 '특수 목적 데이터베이스의 구축' 들도 새로이 구축되고 있다[8].

2. 분석 알고리듬의 개발

생물 정보의 분석에 활용되는 알고리듬들도 대량의 genome 데이터들이 산출됨에 따라 새로운 방향으로 발전 해 나가고 있다. 초기의 분석 알고리듬은 일정한 법칙(rule-based)을 기반으로 서열을 분석하는 방법들을 많이 이용하였으나, 생물학의 개체 특이적이고, 유전자 특이적인 성질을 모두 만족시킬 수 있는 법칙을 찾는다는 것은 거의 불가능하기 때문에 많은 알고리듬들이 한계를 가지고 있었다. 하지만 대량의 서열 정보들과 관련 정보들이 밝혀짐에 따라, 컴퓨터 학습 방법을 통해 새로운 서열의 특징을 컴퓨터가 보다 정확하게 분석 및 예측할 수 있는 machine-learning 기법들이 많이 도입되고 있다. 대표적인 예가 전산학의 음성 인식에 많이 활용되고 있는 HMM(hidden markov rule)을 들 수 있으며, HMM은 이미 gene prediction, multiple alignment 등 Bioinformatics의 많은 분야에 적용되고 있다[10, 11, 12]. 현재 생물학 지식의 축적 속도를 고려하면 학습을 시킬 수 있는 training data set은 빠른 속도로 증가할 것으로 예상된다. 따라서 이러한 machine-learning 기법에 의한 예측 방법은 점점 더 그 정확성이 높아지며 더 많은 범위에 활용될 것으로 기대된다.

또한 model organism들의 full genome sequence가 공개됨에 따라 bioinformatics를 이용한 genome의 구조, 특성의 비교 분석이 가능하게 되었다. 한 개체의 전체 genome의 특징-oligonucleotide bias, repeat sequence의 distribution, coding region과 non-coding region의 비교 등-을 computer를 이용하여 파악할 수 있게 되었다. 즉, 각 개체에 대한 정보들을 다른 개체들과 genome level에서 비교할 수 있고, 각 유전자들의 개체간의 변이와 진화에 관련된 많은

정보들을 비교 분석할 수 있게 되었다. 이러한 연구들은 각 개체들의 진화에 대한 많은 정보들을 제공하고 있으며, comparative genomics라는 새로운 genome 연구의 영역으로 발전하고 있다[13, 14].

통계 열역학 등 여러 방법을 이용한 단백질의 3 차 구조 예측은 bioinformatics가 해결해야 할 큰 난제 중의 하나이다. 초기의 구조 예측을 위한 시도에서는 단백질의 folding이 Anfinsen's thermodynamic principle에 따른다면 해결 할 수 있을 것이라 생각했으나 현재는 단백질의 folding은 단순한 열역학 문제로만 풀 수 없고 단백질 주위의 복잡한 molecular interaction을 모두 고려해야 풀 수 있을 것으로 고려되어지고 있다. 따라서 계산에 의한 구조 예측은 현재에도 어려운 난제로 남아있다. 하지만 최근의 genome project로부터 산출된 많은 유전자 데이터들은 새로운 방식의 3차 구조 예측 방법을 제시하고 있다. 즉 서열의 유사성은 구조의 유사성을 반영한다는 가정으로부터, 주어진 서열과 같은 family에 속한 서열의 구조가 밝혀진 경우 그 서열의 구조를 참고로 주어진 서열의 구조를 예측하는 방법이다. 이러한 방법을 comparative protein modeling이라 부르고 SWISS-PROT에서는 이 방법을 이용하여 SWISS-PROT에 등록된 서열들에 대한 구조 예측 작업을 통해 전체 entry의 15 %에 해당하는 단백질들의 구조를 예측하여 SWISS-MODEL 데이터베이스를 구축하였다[15].

대량으로 산출되는 유전자 서열 정보들의 functional annotation을 생물학 전문가가 일일이 분석하는 것은 불가능하므로, functional annotation에 관련된 bioinformatics 연구들은 post-genome 시대에 특히 그 중요성이 부각된다고 할 수 있다. 이미 genome sequencing project를 수행하고 있는 외국 기관들의 경우 산출된 서열들을 automatic annotation 할 수 있는 정보 인프라의 구축은 필수적인 작업으로 인식되고 있다.

3. Genetic network의 이해

생명체 내에 복잡하게 존재하는 genetic network을 이해한다는 것은 생물학자들의 큰 소망중의 하나이다. 비교적 연구가 많이 진행된 metabolic pathway를 중심으로 이러한 연구가 진행되어 왔으나, 생물학 지식의 한계로 인해 genetic network의 이해는 요원한 문제로 인식되어 왔다. 하지만 개체에 포함된 모든 유전자들의 발현 양상을 파악할 수 있는 DNA chip 기술의 발전은 genetic network의 이해를 실현 가능한 문제로 만들었다. 즉 일정 시간 간격으로 실험이 진행된 DNA chip의 결과는 한 개체에서 진행된 genetic network의 결과를 직접적으로 반영하므로, DNA chip 결과 분석을 통해 개체의 genetic network을 예측 해 낼 수 있다. 이미 Baysian network이나 Boolean Network 과 같은 통계적 방법들을 이용한 DNA chip 분석 결과를 토대로 genetic network을 이해하려는 시도가 이루어지고 있으며, 이러한 연구는 21세기 Bioinformatics 연구의 핵심 과제중의 하나로 발전할 것이다[16, 17](그림 2).

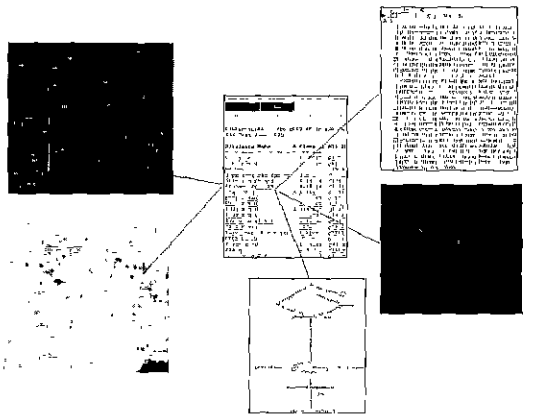

그림 2 Molecular Network의 분석을 위한 데이터들의 통합

4. 데이터마이닝

데이터마이닝이란 경영학에서 나온 이론으로 대량의 데이터로부터 여러가지 분석 도구들을 이용하여 기존에 알지 못했던 새로운 지식을 창출하는 과정(Knowledge-discovery in databases)을 의미한다. 따라서 대량으로 산출되는 생물학 정보의 분석을 위해 데이터마이닝이 도입되는 것은 당연하다고 볼 수 있다. 현재 데이터마이닝의

생물학에의 적용은 초기 단계라 볼 수 있지만 이미 DNA chip data의 분석, EST *mining*을 통한 Single Nucleotide Polymorphism(SNP)의 발굴 등 몇 몇 분야에 기술들이 적용되고 있다.

DNA chip의 경우 데이터마이닝의 적용은 image data의 clustering 부분과 clustering data의 재분석 부분의 두 가지로 나눌 수 있다. Chip image data의 clustering은 한 개의 DNA chip data 내에서 발현 양상이 비슷한 유전자들을 여러 가지 통계적 기법을 이용하여 분류하는 clustering 기법과 여러 번의 DNA chip image data의 분석을 통해 expression profile을 작성하고, 비슷한 expression profile을 가진 유전자들을 grouping 하는 기술로 나눌 수 있다. 이렇게 clustering 된 데이터들은 다시 생물학적 의미를 파악하기 위해 기존에 알려진 다른 데이터(functional catalog, structure information 등)들과의 비교 분석을 위해 data mining 기법의 도입이 필요하다. Decision tree 기법을 이용하여 효모 DNA chip의 expression profile과 promoter region의 상관 관계를 분석한 것을 대표적인 예로 들 수 있다[19].

SNP는 같은 종의 genome에 존재하는 한 염기쌍의 차이(single base-pair variation)로 DNA sequence polymorphism 중 가장 많이 존재하는 형태이다. 인간의 경우 1000 bp당 1 개 정도의 SNP가 존재할 것으로 예측되고 있으며, 질병 연구 및 신약 개발에 중요한 정보들을 제공할 것으로 예측되고 있다. 현재 유전자 데이터베이스에 등록된 서열들의 대부분을 차지하는 Expressed Sequence Tags(ESTs)들은 SNP 발견의 중요한 source들로 활용될 수 있으므로, 많은 bioinformatics tool들과 통계적 기법들이 SNP mining을 위해 적용되고 있다. SNP 발굴을 위한 EST mining은 크게 EST data *cleaning*, *paralog sequence*의 제거, putative SNP의 검출의 세 단계로 나눌 수 있다. EST data cleaning은 다시 sequence histogram을 통해 각 서열의 정확성에 대한 신뢰도를 계산하는 과정, *non-complex region*의 제거, EST의 clustering 등의 기술을 통해 수행 할 수 있다. Paralog의 제거는 SNP 발생 확률과 paralog와 ortholog의 유사성 정도의 비교를 통해 mining 기법을 통해 수행할 수 있고, 마지막으로 putative SNP는 전체 서열에서 SNP 발생 확률을 예측함으로써 얻을 수 있다[20].

5. 결 론

이와 같이 bioinformatics는 1차적인 생물학 실험 결과 혹은 생물학 데이터베이스로부터 많은 지식 기반 방법들을 이용하여 새로운 생물학적 지식을 창출해내기 위해 끊임없이 발전하고 있다. 더욱이 실험 방법들이 점점 더 자동화되고, 결과 산출 속도가 가속화됨에 따라 그 중요성은 점점 더 부각될 것으로 예측된다. 특히 post genome 시대의 주역으로 떠오르고 있는 functional genomics 연구는 신약 개발, 유전병 진단 및 치료 등 기술 집약적인 미래 산업과 밀접한 관계를 가지고 있으므로, *bioinformatics*는 미래 생물학을 선도할 가장 핵심 분야로 발전할 것으로 기대된다.

참고문헌

[1] http://www.ncbi.nlm.nih.gov/BLAST/

[2] http://www.ebi.ac.uk/fasta3/

[3] Benton D. Bioinformatics-principles and potential of a new multidisciplinary tool. *Trends Biotechnol* 1996 14261-72.

[4] http://srs.ebi ac.uk/

[5] http://www.genome.ad. jp/dbget/dbget.html

[6] Barıllot E, Leser U, Lijnzaad P, Cussat-Blanc C, Jungfer K, Guyon F, Vaysseix G, Helgesen C, and Rodriguez-Tome P. A proposal for a standard CORBA interface for genome maps. *Bioinformatics* 1999 15: 157-169.

[7] http://bioinformatics.weizmann.ac.il/cards/

[8] Andreas DB. The Molecular Biology Database Collection: an online compilation of relevant database resources, *Nucleic Acids Res.* 2000 28: 1-7.

[9] http://www.infobiogen.fr/services/dbcat

[10] Eddy SR, Profile hidden Markov models, *Bioinformatics* 1998 14: 755-763.

[11] Hughey R and Krogh A, Hidden

Markov models for sequence analysis: extension and analysis of the basic method, *Comput. Appl. Biosci.* 1996 12: 95-107.

[12] Grundy WN. Bailey TL, Elkan CP, and Baker ME, Meta-MEME: motif-based hidden Markov models of protein families, *Comput. Appl. Biosci.* 1997 13: 397-406.

[13] Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S. Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P. Celniker SE, Chervitz SA, Coates D, Cravchik A. Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Zhong F, Zhong W, Gibbs R, Venter JC, Adams MD, Lewis S,Comparative genomics of the eukaryo tes. *Science* 2000 24;287(5461): 2204-15.

[14] Perriere G. Duret L, and Hobacgen GM. database system for comparative genomics in bacteria. *Genome Res* 2000 Mar;10(3):379-85.

[15] Guex N, and_Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis.* 1997 ;18(15): 2714-23.

[16] Butte AJ, and Kohane IS. Mutual Information relevance network: Functional genomic clustering using pairwise entropy measurement. *PSB 2000.* 418-429.

[17] Akutsu T, Miyano S, and Kuhara S. Algorithms for identifying Boolean Networks and related biological networks based on matrix multiplication and fingerprint function. *RECOMB 2000*, 8-14.

[18] Berry MJA, and Linoff G. Data Mining Techniques. 1997. Wiley Computer Publishing.

[19] http://industry.ebi.ac.uk/~brazma/dm.html

[20] Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, and Gish WR. A general approach to single-nucleotide polymorphism discovery. *Nat Genet.* 1999 23(4):452-6.

김 양 석

1996 생물학연구정보센터 시스템 총괄 담당
1998 생물학연구정보센터 Bioinformatics 팀장
1999 한국 Bioinformatics 학회 기획 간사

E-mail:yskim@bric.postech.ac.kr