미생물 다양성 분석을 위한 웹 기반의 생물정보도구 개발

Web-based Research Assistant Tools for Analysis of Microbial Diversity

강병철[1], 김현진[1,2], 박준형[1,2], 박희경[3], 김철민[1,2,3]
[1] 부산대학교 대학원 생물정보협동과정
[2] 부산지놈센터
[3] 부산대학교 의과대학 생화학교실

Byeong-Chul Kang[1], Hyun-Jin Kim[1,2], Jun-Hyung Park[1,2],
Hee-Kyung Park[3], and Cheol-Min Kim[1,2,3]
[1] Interdisciplinary Program of Bioinformatics,
Graduate School, Pusan National University.
[2] Busan Genome Center.
[3] Dept. of Biochemstry, College of Med., Pusan National University.
E-mail : bckang@pusan.ac.kr

요 약

생태학, 환경공학, 임상진단 등 여러 생물학 분야에서 미생물의 다양성 연구의 중요성이 대두되고 그 연구가 점증하고 있다. 특히 16S rRNA를 분자지표로한 DNA 염기서열 분석방법이 널리 사용되고 있다. 본 논문에서는 16S rRNA의 염기서열 분석과정을 각 단계별로 자동화하고, 생물학자들의 결과 판단이나 사용상의 편의를 도모하기 위하여 웹기반의 미생물 다양성 분석 어플리케이션을 개발하였다. 개발을 위하여 단계별 자동화 및 인터페이스 개발에 적합한 폴더-프로세스-필터 모델을 고안하고 적용하였다. 제공되는 생물정보분석도구는 서열입력, 서열방향교정, 다중서열정렬 및 가시화, 서열동정 등의 분석등이 있으며, 각 결과는 계통분류도구와 호환가능하도록 하였다. 또한 신생아의 장내 세균총에 대한 분석을 수행하여 개발된 도구의 유용성을 확인하였다.

개발된 웹 에플리케이션은 리눅스 시스템 상에서 Perl 과 CGI를 이용하였으며, http://home.pusan.ac.kr/~genome/tools/rat.htm 으로 접속하여 사용할 수 있다.

1. 서론

미생물 군집에서의 유전자형에 대한 연구는 생태학, 환경공학, 임상진단 등의 몇몇 분야에서 그 중요성이 대두되고 있다[1]. 종래의 방법으로는 미생물의 군집을 직접배양하여 형태학적, 생화학적 분류법으로 그 다양성을 분석하였다. 그러나 순수배양의 어려움, 세균의 작은 크기, 형태학적 구분의 제한성, 자동화의 어려움으로 대규모의 연구에는 많은 한계가 있다. 근래에는 이러한 전통적인 방법의 단점을 극복하기 위해서 미생물을 직접 배양하지 않고 분석하는 분자생물학적 방법들이 많이 제안되고 있다. 대표적인 분석방법들로는 전체 염색체 DNA에 대한 PFGE(Pulsed Field Gel Electrophoresis), Southern blotting 과 RFLP(Restriction Fragment Length Polymorphism), PCR-based locus specific RFLP, REP(Repetitive Extragenic Palindromic) PCR, CFLP(Cleavase Fragment Length Polymorphism), AFLP(Amplified Fragment

Length Polymorphism) 시험 그리고 염기서열분석(DNA Sequencing) 등이 있다[1,2].

대표적인 분자지표인 16S rRNA는 단백질 합성에 핵심적인 역할을 하는 리보좀 유전자 중의 하나이다. 16S rRNA는 종과 속간의 분화에 따른 다양성이 큰 부분을 가지고 있어서 특정 분류군에만 존재하는 염기서열을 가지고 있다. 동시에 이 유전자는 생명현상 유지에 필수적인 이차구조를 가지는 서열 부위가 있기 때문에 대부분의 생명체에 공통적으로 보존되어 있다. 이러한 특성으로 인하여 16S rRNA를 분자지표로하여 염기서열을 결정하고 계통분류학적으로 분석하면 다양한 분류군의 상호비교가 가능하다[3].

16S rRNA 유전자 중심의 다양성 분석에는 염기서열의 결정, 프라이머(primer) 부위 제거, DNA 서열의 방향 교정, 대표 OTU(operation taxonomic units) 선정, 키메라(chimera) 서열 검사, 서열동정, 다중서열정렬(multiple alignment), 진화거리 계산, 계통수 작성 및 통계적 검증 등 다양한 단계의 생물정보학적 분석이 필요하다. 각각의 과정에 대한 공개 소프트웨어들이 여러 연구자들에 의해서 제공되고 있지만, 사용자 인터페이스가 일관되지 않으며, 일반 생물학자에게는 익숙하지 않은 유닉스 환경으로 제공되는 경우가 많다. 윈도 환경의 상용 소프트웨어가 있지만, 고가이거나 일괄처리를 하지 못하는 경우가 많다. 온라인 상에서 이를 지원하는 대표적인 데이터베이스는 RDP-II(Ribosomal Database Project - II)[4]로서 95,000건 이상의 미생물의 16S rRNA 서열과 인간을 포함한 다수의 진핵생물의 리보좀 서열을 제공하고 있다. 동시에 웹기반의 정렬 프로그램과 Phylip[5] 인터페이스를 제공하여 본격적인 계통분류를 수행할 수 있도록 하였다. 그러나, 자동화가 절실한 염기서열의 전처리(다중서열정렬 이전까지의 처리과정)에 대한 개발이 미비한 실정이다.

본 논문에서는 서열 처리의 자동화와 사용자 요구에 의한 단계별(stepwise) 인터페이스 개발을 위해서 폴더-프로세스-필터(FPF) 모델을 고안하고 적용하였다. 개발된 웹기반의 생물정보분석도구는 서열입력, 서열방향교정, 다중서열정렬 및 가시화, 서열동정 등의 분석기능을 내장하고 있으며, 그 결과는 손쉽게 RDP-II에 연계되도록 하였다. 또한 신생아의 장내 세균총에 대한 분석을 수행하여 개발된 도구의 유용성을 확인하였다.

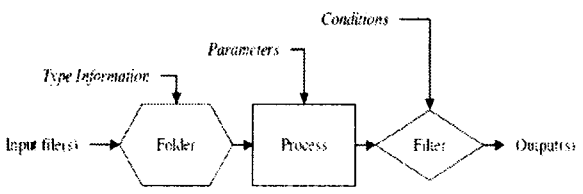

그림 1 최소FPF 객체의 구성

2. 시스템 구성 및 개발방법

2.1 Modeling

유전체 분석을 위한 파이프라인 구축과 유전자 기능 탐색 분야에 많은 선행연구가 있다. PEDANT는 유전체 규모의 서열을 입력하면, 유전자 부위를 찾고 그 기능을 분석하는 전과정을 자동화하고 웹 인터페이스를 통해서 결과를 확인할 수 있도록하였다[6]. 이는 대용량의 데이터 처리에 적합하지만, 분석과정의 단계별 확인이 필요한 분석업무에는 부적절하며 복잡한 시스템 구성으로 유지보수에 어려움이 있다. Bioperl[7]의 세부 프로젝트의 하나인 Biopipe(Bioperl pipeline framework)는 매우 유연성이 높은 방법으로 알려져 있다[8]. 그러나 Biopipe도 PEDANT와 같이 많은 분석 절차의 자동화를 위한 구조를 채택하여 각 단계에 대한 인터페이스 구현에 어려움이 있다.

본 연구에서 제안하는 FPF(*Folder-Process-Filter*) 모델은 각 단계별 인터페이스를 웹상에서 쉽게 개발할 수 있도록 고안되었다. 일반적인 하나의 생물정보분석 과정은 그림 1과 같은 최소FPF(minimal FPF)에 매핑 시킬수 있고, 일련의 생물정보분석과정은 FPF망(FPF network)으로 대치시킬 수 있다. FPF망의 구성은 XML로 된 기술규약에 의해서 정의되도록 하였다. 즉, 단계별 생물정보분석 파이프라인은 최소FPF의 연결에 의해서 구현된다. 최소FPF는 하나의 폴더, 프로세스, 필터 객체를 가지고 있다. 폴더 객체는 프로세스에 입력될 데이터의 저장과 형식 인증을 수행한다. 프로세스 객체는 BLAST나 ClustalW 등 외부 프로그램이

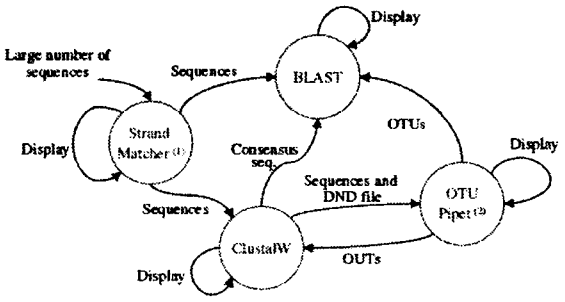

그림 2 개발된 웹기반의 생물정보분석도구의 FPF망. (1) 자체 개발된 도구로 *E. coli* 16S rRNA을 기준으로 입력서열의 DNA방향을 교정한다; (2) 기준 역치값보다 유사성을 보이는 서열은 하나의 서열로 간주하여 대표 OTU를 선정한다

나 각 개인이 개발한 분석 프로그램과의 연결자 역할을 하거나 개발자 고유의 알고리즘을 내장한다. 필터 객체는 프로세스의 결과값을 조건에 따라 변환 및 저장하며 다음 FPF와 연결정보를 가진다.

2.2 Framework for Stepwise Web Application

웹 어플리케이션의 기본 코드는 XML에 의해서 정의된 FPF망 정보로부터 자동으로 생성된다. 즉, 사용자 요구에 따른 생물분석도구를 선정하고 기본적인 파이프라인의 구성을 FPF망으로 표현한다. 고안된 프레임워크(framework)는 FPF망 정보에 따라 CGI::Application[9]로부터 상속된 코드를 생성한다. 생성된 코드에는 CGI의 기본제어 구조, 웹 입출력을 위한 템플레이트 객체, 데이터 저장을 위한 객체 정보를 포함하고 있다. 따라서, 템플레이트를 이용한 웹 프로그래밍과 펄(perl)이나 외부 프로그램을 이용한 생물정보처리 프로그래밍이 완전히 구분되어진다. 전통적인 MVC 모델링에 대비하여 보면, 제어부(Controller)는 CGI::Application의 상속에 의해서 구현되고, 모델(Model)은 펄을 이용한 자체 구현 또는 외부 프로그램을 호출하여 구현된다. 마지막으로 뷰(View)는 Template Toolkit 2[10]와 HTML에 이용한 웹 프로그래밍에 의해서 구현된다.

3. 결과 및 고찰

3.1 Implementation

미생물의 다양성 분석시에 필요한 전처리 과정을 그림 2은 같이 FPF망으로 설계하고 프레임워크를 이용하여 기본 코드를 생성하였다. 서열의 동정을 위해서 BLAST[11]를 이용하였고 기초적인 계통분류를 위해서 ClustalW[12]를 활용하였다. 각 서열의 방향 조정과 OTU선정을 위해서는 자체 개발한 알고리즘을 구현하였다.

다중서열정렬 결과는 MVIEW[13]에 의해 보여지고 이것의 계층구조는 SVG[14] 형식의 그래프파일로 제공되도록 구현하였다. 개발된 각 웹 어플리케이션의 실제 화면을 그림 3에 보이며, http://home.pusan.ac.kr/~genome/tools/rat.htm으로 접속하여 사용할 수 있다.

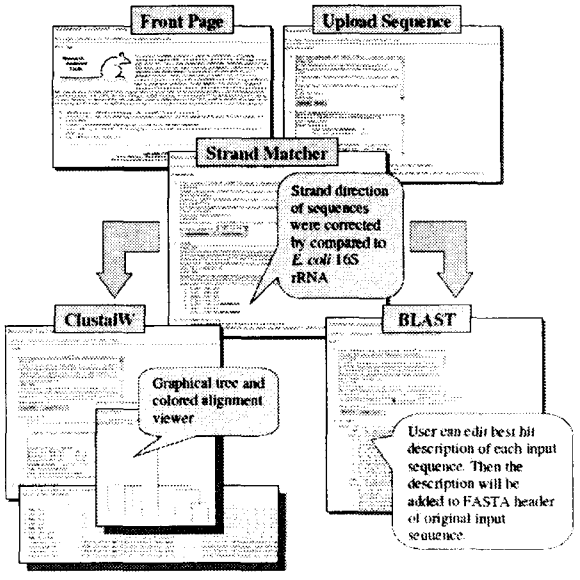

그림 3 개발된 웹 도구의 화면 구성과 단계별 사용방법

3.2 Materials and Methods

개발된 웹기반의 생물정보도구의 유용성을 확인하기 위하여 시험 데이터를 분석하였다. 시험 목적은 신생아의 장내 세균 분포를 확인하는 것이면 이를 위해서 생후 1일째 신생아의 태변을 수집하였다. 분변에서 DNA를 분리한 뒤 중합효소연쇄반응에 의해 16S rDNA 유전자를 선택적으로 증폭하여 클론화 라이브러리 구축한 뒤 각 클론의 염기서열분석을 수행하였다. 이를 통해 얻은 전체 염기서열 수는 총 124개며, 이후 개발된 웹 도구와 계통분류 프로그램을 활용하여 분석하였다.

3.2 Analysis of Sample Data

전체 124개의 16S rRNA 서열을 분석한 결과 20개의 OzU를 얻었으며, RDP 8.1 데이터베이스와 검색 결과, 그림 4와 같이 7종의 분류군으로 나뉨을 확인하였다. *Enterobacter sp.*, *Lactococcus latis*, *Leuconostoc citrem*, *Streptococcus mitis*를 동정 하였으며 가장 큰 분류군인 *Lactococcus latis* 는 66.9%에 달했고, 미동정그룹이 약 18% 였다.

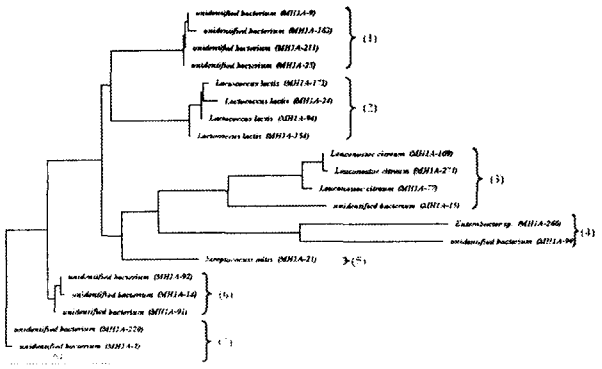

그림 4 시험데이터의 계통수 작성 결과

4. 결론 및 향후 연구

지금까지 우리는 16S rRNA 서열을 바탕으로 미생물의 다양성 분석을 지원하기 위한 웹 에플리케이션과 단계별 웹 에플리케이션 개발을 위한 모델을 제시하였다.

그러나 제안된 FPF 모델은 단계별 웹 인터페이스의 개발과 유지보수를 용이하게 할 것으로 기대한다. 또한 미생물 다양성 분석용 웹 도구는 실제 실험데이터를 분석하여 그 유용성을 확인하였다.

제안된 본격적인 계통분류 패키지를 포함하고 있지 않으므로 Phylip, ARB[15], RDP(Ribosomal Database Project)와 연계하여 활용하면 좋을 것으로 기대한다.

향후 개발은 프로세스 관리모듈(process manager)을 추가하여 시간이 많이 소요되는 분석과정의 자동화를 가능하게 하고 Phylip 인터페이스를 포함하여 엄밀한 계통분류가 가능하도록 하겠다.

5. 참고문헌

[1] Dahllof, I., "Molecular community analysis of microbial diversit", *Curr. Opin. Biotechnol.*, Vol. 13, pp.213-217, 2002.

[2] Olive, D.M and Bean, P., "Principles and Applications of Methods for DNA-Based Typing of Microbial Organisms", *J. of Clin. Microbiol.*, Vol. 37, pp.1661-1669. 1999.

[3] Pace, B., and Campbell, L.L. "Homology of ribosomal ribonucleic acid of diverse bacterial species with Escherichia coli and Bacillus stearothermophilus", *J. Bacteriol.*, Vol. 107, pp.543-547. 1971.

[4] Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarrell, D.M., Schmidt, T.M., Garrity, G.M., and Tiedje, J.M., "The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy.", *Nucleic Acids Res.*, Vol. 31, pp.442-443, 2003.

[5] Felsenstein, J., PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle, 1993.

[6] Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., Mewes, H.W., "Functional and structural genomics using PEDANT", *Bioinformatics*, Vol. 17, pp.44-57, 2001.

[7] Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E.D., Wilkinson, M., and Birney, E., "The Bioperl Toolkit: Perl modules for the life sciences", *Genome Research*, Vol. 12, pp.1161-1168, 2002.

[8] Letondal, C.A., "Web interface generator for molecular biology programs in Unix", *Bioinformatics.*, Vol. 17, pp.73-82, 2001.

[9] Erlbaum, J., http://search.cpan.org/~markstos/CGI-Application-3.22/lib/CGI/Application.pm

[10] Wardley, A., http://search.cpan.org/~abw/Template-Toolkit-2.13/lib/Template.pm

[11] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., "A basic local alignment search tool", *Journal of Molecular Biology*, Vol. 215, pp.403-410, 1990.

[12] Thompson, J.D., Higgins, D.G., and Gibson, T.J., "CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.", *Nucleic Acids Res., Vol.* 22, pp.4673, 1994.

[13] Brown, N.P., Leroy, C., and Sander, C., "MView: A Web compatible database search or multiple alignment viewer", *Bioinformatics*, Vol. 14, pp.380-381, 1998.

[14] SVG, http://www.adobe.com/svg/main.html

[15] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., and Schleifer, K-H., "ARB: a software environment for sequence data.", *Nucleic Acids Res.*, Vol. 32(4), pp.1363-1371, 2004.