Bioinformatics and Biosystems 2006, Vol. 1, No. 2, pp. 99-102

생물정보시스템을 이용한 Local Animal BLAST Search System 구축

(Development of Local Animal BLAST Search System Using Bioinformatics Tools)

김 병 우[1], 이 근 우[2, 3], 김 효 선[1], 노 승 희[1], 이 윤 호[2], 김 시 동[4], 전 진 태[1], 이 지 웅[4], 조 용 민[4], 정 일 정[4], 이 정 규[1]

[1]경상대학교 응용생명과학부 · 농업생명과학연구원, [2]경상대학교 생명과학부 생화학전공, [3]경상대학교 응용생명과학부 · 국가핵심연구센터, [4]농촌진흥청 축산연구소

초 록

BLAST(Basic Local Alignment Search Tool)는 서열 데이터베이스 탐색을 위하여 가장 많이 사용되는 프로그램이다. 전체 서열간의 최적 글로벌 정렬을 수행하는 대신에 지역적 유사성이 있는 부분을 찾아 서열 짝짓기를 수행하는 특징을 갖는다. 일반적인 연구자들은 서열 상동성 검색을 위해 NCBI에 접속하여 웹 브라우저를 통해 온라인으로 BLAST를 수행하게 되는데, 이 경우 사용자 각각의 네트워크 환경이나 입력할 데이터양에 따른 검색속도의 지연 및 제한 등과 같은 여러 문제에 부딪히게 되고, 또한 보안유지가 필요한 서열 데이터의 유출 가능성이 존재한다. 그러므로 대량의 서열 데이터에 대하여 빠르고 안전하게 BLAST 상동성 검색이 가능한 Local BLAST 검색 시스템의 필요성이 증대되고 있다. 본 연구에서는 NCBI의 Genbank에서 공개된 동물의 발현 유전자 단편들(ESTs)에 대한 데이터를 이용하여 소, 돼지, 닭, 등의 경제형질과 연관된 유용 유전자만을 추출하여 이들만으로 구성된 새로운 데이터베이스를 구축하였고, 또한 이들을 사용할 수 있는 새로운 검색시스템을 개발하였다. 자체 제작한 Perl script를 사용하여 필요한 데이터를 축종별로 추출 하여 새로운 DB를 구축하였으며 이 속에는 소의 경우 650,046개, 돼지의 경우 368,120개, 닭의 경우 693,005개의 발현 유전자 단편들(ESTs)이 포함된다. 또한 이들 DB 분석이 가능한 Local Animal BLAST Web 검색시스템(http://bioinfo.kohost.net)을 고성능 병렬 PC Cluster 시스템과 연동하도록 자체 구축함으로써 본 시스템이 보다 효율적인 생물정보학 연구수행이 기여할 것으로 기대된다.

키워드: 생물정보학, 발현 유전자 단편 DB, Local Animal BLAST, PC 클러스터

Abstract

The Basic Local Alignment Search Tool (BLAST) is one of the most established software in bioinformatics research and it compares a query sequence against the libraries of known sequences in order to investigate sequence similarity. Expressed Sequence Tags (ESTs) are single-pass sequence reads from mRNA (or cDNA) and represent the expression for a given cDNA library and the snapshot of genes expressed in a given tissue and/or at a given developmental stage. Therefore, ESTs can be very valuable information for functional genomics and bioinformatics researches. Although major bio database (DB) websites including NCBI are providing BLAST services and EST data, local DB and search system is demanding for better performance and security issue. Here we present animal EST DBs and local BLAST search system. The animal ESTs DB in NCBI Genbank were divided by animal species using the Perl script we developed. and we also built the new extended DB search systems for the new data (Local Animal BLAST Search System: http://bioinfo.kohost.net), which was constructed on the high-capacity PC Cluster system for the best performance. The new local DB contains 650,046 sequences for Bos taurus(cattle), 368,120 sequences for Sus scrofa (pig), 693,005 sequences for Gallus gallus (fowl), respectively.

Keywords: Bioinformatics, ESTs DB, Local Animal BLAST, PC Cluster

서 론

생물정보학(Bioinformatics)의 가장 기본적인 작용 중 하나는 새롭게 서열을 알게 된 DNA와 이미 기록되어 있는 DNA 서열 간의 유사성(Similarity)이나 상응관계(Homology)를 검색하는 것이다. 이를 통해서 연구자들은 새롭게 얻어

진 유전자 서열에 의해 부호화된 단백질의 종류를 예측할 수 있다. 이러한 서열 유사성 검색 도구들 중 가장 대표적인 것으로 1988년 개발된 FASTA(Pearson 등, 1988; Wilbur 등, 1983, 1984)와 이와 유사한 목적으로 1990년 개발된 BLAST(Basic Local Alignment Search Tool, Altschul 등, 1990, 1997)가 있다. 웹 상에서 가장 대중화되고 이용자에게 친숙한 서열 유사성 검색 도구인 BLAST는 SWISS-PORT나 PDB 등과 같은 모든 주요 서열 DB를 검색할 수 있는 도구로 Karlin과 Altschul이 개발한 통계이론을 이용하고 있다 (Karlin 등, 1990, 1993). BLAST 의 서열정렬 알고리즘은 Smith and Waterman방법(1981)을 기반으로 하며 전체 서열 간의 최적 글로벌 정렬을 수행하는 대신에 지역적 유사성 (local similarity)이 있는 부분을 찾아 서열 짝짓기를 수행하는 특징을 갖는다.

현재 국내의 동물분야 생물정보학 환경은 아주 열악한 상태라고 할 수 있다. 일례로 생물정보학에 대한 수요시장이 형성되어 있지 않아 개발에 대한 동기부여가 미약할 뿐만 아니라, 대부분의 연구가 단순한 연구수준에 머물러 있고, 기술 간의 질적인 수준 차이가 커 이 또한 장애 요인 중 하나이다. 그리고 대량의 원시데이터를 생산, 저장, 그리고 관리하는 곳이 적고, 국외의 대학이나 비영리기관에서 개발하여 제공하는 무료서비스가 많다는 것도 개발의 제한요소이다.

동물분야 생물정보학 관련 연구에 대한 필요성은 증대되고 있으나, 아직까지 연구는 초기단계에 머물고 있는 실정이다. 동물분야에서도 유용유전자 탐색을 위하여 일반적으로 NCBI나 EBI 같은 주요서버에 웹 브라우저를 통해 온라인으로 접속하여 BLAST를 수행하게 되는데, 이 경우 사용자 각각의 네트워크 환경이나 입력할 데이터양에 따른 검색속도의 지연 및 제한 등과 같은 여러 문제에 부딪히게 되고, 또한 보안유지가 필요한 서열 데이터의 유출 가능성이 존재한다. 그리고 NCBI 등 국외 전문 database는 동물분야 연구에 적합하도록 축종별로 검색 시스템을 제공하고 있지 않아 동물분야 연구에 있어 효율성이 다소 떨어진다고 판단된다. 그러므로 대량의 서열 데이터에 대하여 빠르고 안전하게 BLAST 상동성 검색이 가능한 Local Animal BLAST 검색 시스템의 필요성이 증대되고 있다.

본 연구에서는 동물분야의 생물정보학 연구의 기초연구 단계로 동물관련 유전자를 한곳에 모아 유용유전자 탐색연구의 효율성을 증대시키기 위하여 자체적인 local 검색 시스템을 구축하였다. NCBI의 Genbank에서 공개된 동물의 발현 유전자 단편(ESTs)들에 대한 데이터를 이용하여 소, 돼지, 닭의 경제형질과 연관된 유전자만을 모아 자체 제작한 Perl script(Tisdall, 2001)를 사용하여 동물분야 연구에 적합하도록 축종별로 새롭게 database를 재구성 하였으며, 대용량 자료 분석이 가능한 Local Animal BLAST Web 검색시스템을 구축하였다. 동물분야에서 유용유전자의 발굴 및 탐색을 위해서는 수많은 반복 작업, 많은 시간과 인력이 필요한 방대한 연구이므로 본 시스템이 효율적인 동물분야 생물정보학 연구의 기반을 조성하는데 도움이 되었으면 한다.

교신저자 : 이정규 (Email: jglee@gnu.ac.kr)
본 연구는 농촌진흥청 "바이오그린21사업(#. 20050303103441)"과 과학기술부/한국과학재단의 "환경생명과학 국가핵심연구센터(EB-NCRC)"사업의 지원(#: R15-2003-012-02001-0)에 의해 이루어진 연구결과의 일부임.

방 법

NCBI에서는 동물분야 생물정보학 연구에 적합하도록 축종별로 세분화 하여 서비스 하지 않는 점을 본 연구의 목적으로 하였고, Local Animal BLAST Search System 구축을 위하여 경상대학교 동물유전육종학 연구실 리눅스(Linux) 서버에 Red Hat Linux 8.0과 MySQL을 설치하고, NCBI 공개 FTP 서버(ftp.ncbi.nih.gov/blast/db/FASTA/)에서 2004년 12월 31일자 est_others.Z(2.14 GB)와 nt.Z(3.05 GB)파일을 다운로드 받아 본 연구진이 제작한 Perl script를 이용하여 소, 돼지, 닭으로 추출 및 분류하여 자체 DB화 하였다.

Perl script를 이용하여 각각 축종별로 Bos taurus (cattle), Sus scrofa (pig), Gallus gallus (chicken, fowl)이 하나라도 포함된 자료만을 분류 DB화하여 Web 검색 시스템을 자체구축 하였다. 또한, 대용량 자료분석이 가능하도록 메인서버와 PC Cluster(Red Hat Linux 8.0)시스템을 구축하여 자료의 전산처리속도를 향상시켰다.

그림 1은 대용량 자료처리가 가능하도록 PC를 병렬로 연결(29대)한 PC Cluster 연결 모식도와 제어시스템의 화면모습이다. 유전공학의 발달과 고성능 분석기계의 도입으로 대용량의 유전자 관련 자료가 양산되고 있다. 대용량의 유전자의 구조 및 기능을 분석하기 위해서는 고성능 대형 서버가 필요한 실정이다. 그러나 PC Cluster 시스템은 대용량 자료를 분산처리한 후 메인컴퓨터가 계산값을 취합하는 장점과 전산실 등 자주 사용하지 않은 PC를 효율적으로 활용할 수 있어 자료분석 시간 단축과 비용 절감 효과 등 동물분야 생물정보학(Bioinformatics) 연구에 도움이 될 것으로 기대된다.

결과 및 고찰

표 1에서부터 표 3까지 축종별로 확대 구축한 DB의 간략한 내용을 정리 하였다. 표 1에서는 소와 관련된 DB 자료에 대한 요약이 정리되어 있다. Bos taurus (cattle)이 하나라도 포함된 자료만을 분류 cattle_est DB를 만든 후 formatdb 명령을 사용하여 blast search에 적합하도록 Database파일을 컴파일 하였다. NCBI의 nt DB에도 소와

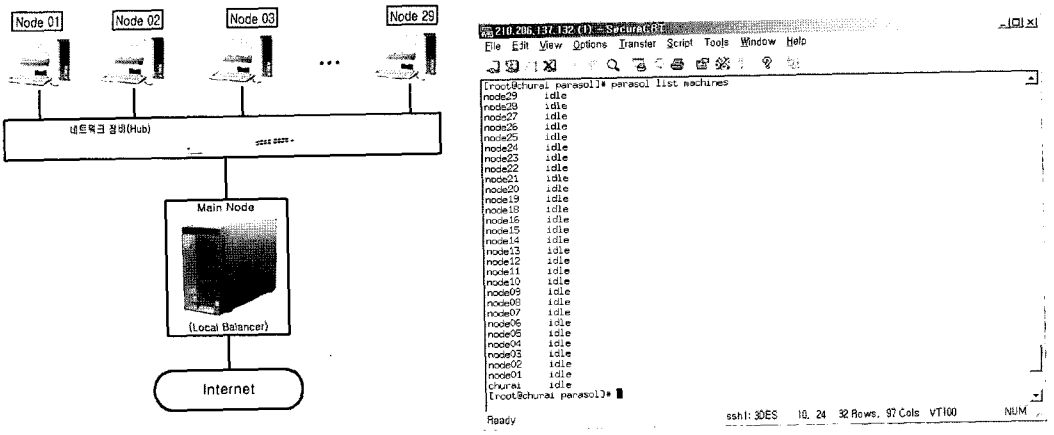

그림 1. PC Cluster System 모식도 와 제어시스템 화면모습

연관된 ESTs가 있으므로 위와 같은 방법으로 소와 관련 있는 자료를 추출하여 cattle_nt를 만들었으며, 소와 연관된 모든 자료를 한곳에 모으기 위하여 cattle_est와 cattle_nt DB를 합하여 cattle_all DB를 만들었다. 소와 관련된 자료는 모두 951,007개(cattle_all)였다.

표 2와 3에는 돼지와 닭에 대한 DB 자료에 대하여 간략히 요약 정리하였다. 소의 경우와 유사하게 NCBI의 est_others DB를 이용하여 Sus scrofa (pig) 이나 Gallus gallus (chicken, fowl)이 하나라도 포함된 자료만을 뽑아 같은 방식으로 분류 정리하였다. 돼지와 관련된 자료는 모두 381,969개(pig_all)였으며 닭과 관련된 자료는 모두 746,096개(fowl_all)였다. 이들 자료는 앞으로 계속하여 Update 할 예정이다. 자료의 분석 효율성을 높이기 위하여 다양한 DB를 확대구축 하였으며 또한, 대용량 자료 분석이 가능하도록 메인서버와 PC Cluster 시스템을 구축하여 자

표 1. Construction of Local Cattle ESTs Ddatabase Overview

DB name	Sequence	Total letters	Size	Source DB name
cattle_est	650,046	321,794,657	404 MB	est_others
cattle_nt	300,961	292,239,278	336 MB	nt
cattle_all	951,007	614,033,935	740 MB	cattle_est, cattle_nt
cattle_estmRNA	664,946	339,327,882	423 MB	cattle_mRNA + cattle_est
cattle_mRNA	14,900	17,533,225	19 MB	cattle_nt (mRNA | cDNA)

표 2. Construction of Local Porcine ESTs Database Overview

DB name	Sequence	Total letters	Size	Source DB name
pig_est	368,120	187,601,524	229 MB	est_others
pig_nt	13,849	30,106,728	32 MB	nt
pig_all	381,969	217,708,252	261 MB	pig_est, pig_nt
pig_estmRNA	373,166	194,002,264	236 MB	pig_mRNA + pig_est
pig_mRNA	5,046	6,400,740	7 MB	pig_nt (mRNA | cDNA)

표 3. Construction of Local Fowl ESTs Database Overview

DB name	Sequence	Total letters	Size	Source DB name
fowl_est	693,005	429,836,228	523 MB	est_others
fowl_nt	53,091	98,575,801	105 MB	nt
fowl_all	746,096	528,412,029	628 MB	fowl_est, fowl_nt
fowl_estmRNA	737,903	498,660,126	597 MB	fowl_mRNA + fowl_est
fowl_mRNA	44,898	68,823,898	74 MB	fowl_nt (mRNA | cDNA)

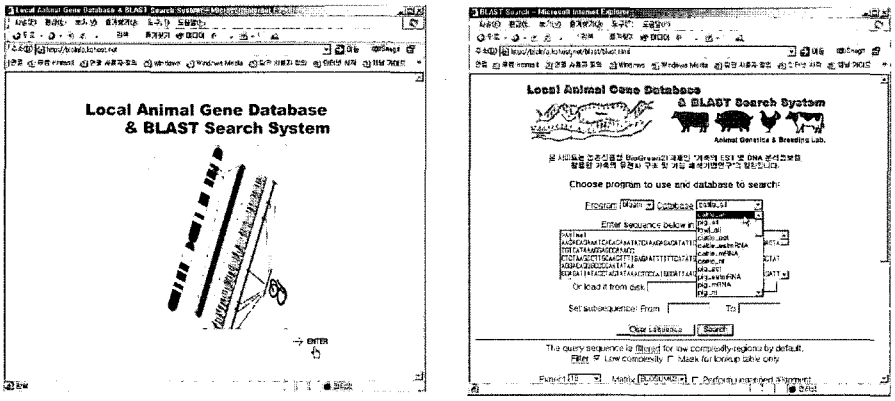

그림 2. Local Animal ESTs Database and Web-Based Data Search System

료의 전산처리속도를 향상시켰다.

대용량 자료 분석을 위한 PC Cluster 시스템과 연동 가능한 자체 구축한 동물유전체 DB 및 검색 시스템(http://bioinfo.kohost.net)은 그림 2와 같다.

참고문헌

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403-10.

[2] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-402.

[3] Tisdall JD. (2001) *Beginning Perl for Bioinformatics*, O'Reilly

[4] Karlin S, Altschul SF. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA,* 87:2264-8.

[5] Karlin S, Altschul SF. (1993) Applications and statistics for Multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA,* 90:5873-77.

[6] Pearson WR, Lipman DJ. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* 85:2444-8.

[7] Smith TF, Waterman MS. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-7.

[8] Wilbur WJ, Lipman DJ. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA,* 80:726-30.

[9] Wilbur WJ, Lipman DJ. (1984) The context dependent comparison of biological sequences. *Siam. J. Appl. Math,* 44:557-67.