Korean Journal of Malacology, Vol. 20(2): 165-169, December, 2004

연체동물 전용 서열 블라스트 서버구축

이용석, 조용훈[1], 김대수, 김대원, 김민영, 최상행, 연제오, 변인선, 강보라, 정계헌[1], 박홍석

한국생명공학연구원 유전체연구센터, [1]순천향대학교 생명과학부

Construction of BLAST Server for Mollusks

Yong-Seok Lee, Yong-Hun Jo[1], Dae-Soo Kim, Dae-Won Kim, Min-Young Kim, Sang-Haeng Choi, Jei-Oh Yon, In-Sun Byun, Bo-Ra Kang, Kye-Heon Jeong[1], and Hong-Seog Park

Genome Research Center, Korea Research Institute of Bioscience and Biotechnology, 52, Eoeun-dong, Yuseong-gu, Daejeon, 305-333, Korea

[1]Department of Biology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam, 336-745 Korea

ABSTRACT

The BLAST server for the mollusk was constructed on the basis of the Intel Server Platform SC-5250 dual Xeon 2.8 GHz cpu and Linux operating system. After establishing the operating system, we installed NCBI (National Center for Biotechnology Information) WebBLAST package after web server configuration for cgi (common gate interface) (http://chimp.kribb.re.kr/mollusks). To build up the stand alone blast, we conducted as follows: First, we downloaded the genome information (mitochondria genome information), DNA sequences, amino acid sequences related with mollusk available at NCBI. Second, it was translated into the multifasta format that was stored as database by using the formatdb program provided by NCBI. Finally, the cgi was used for the Stand Alone Blast server. In addition, we have added the vector, *Escherichia coli,* and repeat sequences into the server to confirm a potential contamination. Finally, primer3 program is also installed for the users to design the primer. The stand alone BLAST gave us several advantages: (1) we can get only the data that agree with the nucleotide sequence directly related with the mollusks when we are searching BLAST; (2) it will be very convenient to confirm contamination when we made the cDNA or genomic library from mollusks; (3) Compared to the current NCBI, we can quickly get the BLAST results on the mollusks sequence information.

Keywords: BLAST, Mollusk, Sequence, Information.

Received November 2, 2004; Accepted December 8, 2004
Corresponding author: Park, Hong-Seog
Tel: (82) 42-879-8132 e-mail: hspark@kribb.re.kr
1225-3480/20210

서 론

바이오테크놀로지와 관련된 기계, 전자 산업의 급속한 발전에 따라 생물학데이터는 대량화 되고 있으며, 이 대량화된 정보를 처리하기 위해 컴퓨터의 이용은 필수적인 요소가 되어 버렸다. 이러한 데이터는 유전체 서열정보 (genome), 유전자 서열정보 (gene), 아미노산 서열정보를 비롯하여, 단백질의 3차 구조, 또 이러한 정보들과 관련된 문헌정보 까지도 연결되어 종합되고 있으며 이러한 정보들은 NCBI (National Center for Biotechnology Information), EMBL (The European Molecular Biology Laboratory) 등의 공공 웹싸이트 들을 통해 무상으로 공개되어 제공되고 있다. 이렇게 제공되어지는 데이터베이스에서 우리가 원하는 데이터를 찾는 방법 중 가장 많이 사용되는 방법은 상동성 검사 (homology test) 이며, 이를 수행하는데 가장 잘 알려진 프로그램은 BLAST (basic local alignment search tool) 이다 (Altschul *et al.*, 1990; McGinnis and Madden, 2004). 하지만 초기의 BLAST 프로그램은 Unix type의 운영체제에서 명령어 방식 (command line interface) 으로만 운영이 가능하여 일반 연구자들이 사용하기엔 조금 어려움이 있었으며 비교할 서열 중간에 들어있

는 gap 을 계산하지 못하여 상동성이 적은 서열은 비교하지 못하는 단점이 있었다. 1900년대 후반에 들어서면서 이러한 단점을 보완하기 위해 gapped and PSI (position-specific iterated) - BLAST가 발표되었고 (Altschul *et al.*, 1997; Muller *et al.*, 1999), 그래픽 유저 인터페이스 (GUI; graphic user interface) 가 일반화 되고 인터넷 사용이 범용화 되어 많은 사람들이 컴퓨터를 사용하게 되면서 WebBLAST가 등장하였다 (Ferlanti *et al.*, 1999). WebBLAST의 출현으로 많은 사람들이 웹브라우저를 통해 서열의 상동성 검사를 매우 편리하게 사용할 수 있게 되었다.

사람, 마우스, rat, 복어 등 게놈프로젝트가 끝난 동물들의 경우에는 게놈정보를 바탕으로 한 많은 유전자 및 아미노산 정보가 밝혀져 있어, 연구자들이 유전자에 대한 연구를 할 때 많은 도움을 받을 수 있다. 하지만 무척추동물 특히 연체동물을 경우엔 분류에 필요한 서열들은 조금 연구가 되어 있으나 게놈 연구가 되어 있는 종은 없고, 기능성 유전자에 대한 연구도 거의 되어 있질 않아 연구에 많은 애로사항이 있다. 이러한 유전자원을 대량으로 확보하는 방법에는 expressed sequence tag (EST) 방법이 있으나 비용이 많이 들어 일반 연구자들의 경우 이미 많은 연구가 되어진 척추동물의 데이터를 이용하여 관련 연체동물의 데이터를 찾기 위해 BLAST를 통한 상동성을 조사하는 방법을 사용하는 사례가 많다. 하지만, 척추동물 유전자를 대상으로 하여 NCBI BLAST를 이용하면 유사도가 높은 다른 척추동물의 데이터가 먼저 나오므로 정작 찾고자 하는 연체동물의 데이터는 아래 부분에 나오거나, HSP (high-scoring segment pair) 숫자에 들어가지 못해 데이터를 찾을 수 없는 경우가 많다. 그러므로 AnoXcel (*Anopheles gambiae* 단백질 관련 데이터베이스; Ribeiro *et al.*, 2004) 등과 같이 특정한 생물의 서열정보를 다루거나, PathoGene database (질병 유발유전자 데이터 베이스; Ng *et al.*, 2004), Apoptosis database (세포자살관련 유전자 데이터베이스; Doctor *et al.*, 2003) 등과 같이 특정한 연구목적에 관련된 정보 데이터베이스 등이 독립적으로 구성되고 있으며, NCBI 에서도 이와 비슷하게 적용하여 많은 연구가 되어진 사람, 닭, 돼지, 개, 양, 마우스, rat, 고양이를 비롯하여 곤충, 선형동물, 식물, 곰팡이, 말라리아 등에 관해 독립적으로 BLAST를 실행 할 수 있도록 제공하고 있다. 하지만 연체동물만을 대상으로 한 데이터베이스는 아직 없어 연체동물을 대상으로 한 유전자 연구를 수행 할 때 많은 어려움이 따르고 있다.

본 연구에서는 연체동물을 단독으로 한 BLAST 서버를 구축하여 앞으로 많은 연구가 되어질 연체동물을 대상으로 한 유전자 연구에 도움이 되고자 하였다.

재료 및 방법

1. 서버 구축

사용된 서버는 Intel Server Platform SC-5250에 dual Xeon 2.8 GHz cpu 시스템이며, 운영체제 (operating system) 는 Linux Enterprise AS-3를 사용하였다. 운영체제 설치 후 Apache 웹서버의 설정에서 일반 사용자가 cgi (common gate interface) 를 사용할 수 있도록 환경설정을 한 후 WebBLAST 패키지를 설치하였다.

2. 데이터베이스 구축

NCBI 에 등록되어 있는 연체동물과 관련된 genome정보 (미토콘드리아 게놈정보), 유전자서열 정보, 아미노산 서열정보를 taxonomy browser와 연계하여 모두 다운 받은 후, multifasta 형태의 정보로 만든 후 NCBI에서 제공하는 formatdb 프로그램을 사용하여 BLAST용 데이터베이스로 만들었으며 부가적으로 실험 후 데이터 확인시 필요한 벡터서열, *Escherichia coli* 서열, 반복서열 등을 모두 데이터베이스에 포함하여, 실험 데이터를 검증할 때 용이하도록 하였다. 그리고 primer3 등 실험시 부가적으로 필요한 웹용 프로그램

Table 1. Ongoing genome projects related with mollusks. (http://www.genomesonline.org)

Species	Type	Main Institution
Argopecten irradians (bay scallop)	EST	Marine Biological Laboratory
Biomphalaria glabrata	Genome	NHGRI
Biomphalaria glabrata (bloodfluke planorbis)	Genome	International Consortium (Univ. of New Mexico)
Crassostrea virginica	EST	Auburn Univ.
Lottia scutum	Genome	JGI / Univ. of California, Berkeley / Univ. of Washington Univ. of Arizona / Univ of Queensland, Australia
Spisula solidissima (clam)	Genome	Marine Biological Laboratory / Hebrew Univ Technion-Israel Institute of Technology

을 설치하여 연구자들이 편리하게 이용하도록 하였다.

3. 웹 인터페이스 구축

Nucleotide 서열정보, 아미노산 서열정보, mitochondrial genome서열 데이터베이스를 독립적으로 검색이 가능하도록 구성하였으며 query 및 데이터베이스가 허용하는 한 blastp, blastn, blastx, tblastn, tblastx 모두 수행이 가능하도록 하였다. Vector, *E-coli,* repeat 서열을 따로 검색 할 수 있도록 하였으며, multi DB 메뉴를 만들어 라이브러리 확인 (insert size 측정) 등을 할 때 용이하도록 하였다. 또한 Perl script를 기반으로 한 검색엔진을 설치하여 연체동물관련 서열정보를 종 이름, 유전자 이름 및 NCBI accession number 등을 query 로 하여 찾을 수 있도록 하였다.

결과 및 고찰

세계적으로 총 1238 개의 genome project가 진행중이거나 완성되었다 (2004년 12월 7일 기준). 그중 236 개 생물종에 대한 genome project는 완성되어 논문으로 발간되었으며 537개 원핵생물 게놈 및 436 개 진핵생물의 genome project 가 현재 진행중이다 (2종은 정보가 공개되지 않았음). 이렇듯 많은 생물에 대한 게놈 및 EST 연구가 진행되고 있으나 연체동물의 경우 이러한 연구가 매우 미진하여 4종에 대한 게놈 프로젝트 및 2 종에 대한 EST 프로젝트 전부였다 (Table 1). 이는 전 세계의 게놈 연구의 0.5% 이하에 지나지 않는다. 이러한 결과는 NCBI에 등록되어진 염기서열의 숫자에도 그대로 반영된다. NCBI에 등록되어진 연체동물 관련 염기서열 등록 숫자는 2004년 12월 13일 기준으로 총 45,379 개로 NCBI 전체 염기서열 38,000,000 개의 약 0.1% 에 지나지 않아 연체동물의 서열정보는 다른 생물군에 비해 매우 빈약함을 알 수 있었다. 그러므로 BLAST를 이용한 상동성 검색을 하는 경우 연체동물의 서열을 찾기 힘든 것은 너무나 당연하다는 사실을 알 수 있었다.

이번 연구에 의해 구축된 연체동물 전용 BLAST 서버를 이용하여 BLAST 검색시 연체동물관련 nucleotide 정보와 일치된 결과만을 따로 얻을 수 있었고, 아직 데이터베이스가 그리 크지 않아 매우 빠른 속도로 검색을 할 수 있었다.

이러한 사실을 입증하기 위하여 NCBI에 등록되어진 사람의 몇 가지 유전자 염기서열을 대상으로 NCBI BLAST 및 연체동물 전용 데이터베이스에 blastx 및 tblastx 방법을 이용하여 시험하였다. (NCBI의 경우 데이터 포맷시간은 제외하고 시험하였다). 테스트에 사용된 데이터는 Table 2 와 같다.

Cathepsin CDs 서열을 가지고 시험한 경우 NCBI BLAST 에서 결과가 나온 후 연체동물의 sequence를 찾을 수 없었지만, 연체동물 전용 데이터베이스를 이용할 경우 blastx 결과에서 *Mytilus galloprovincialis* cathepsin을 검색 할 수 있었으며, tblastx 결과에서 *Crassostrea virginica, Argopecten irradians, Aplysia californica* 등의 생물의 EST 데이터와 일치함을 매우 빠른 시간에 검색 할 수 있었다. Ferritin mRNA 서열을 가지고 시험을 한 경우엔 NCBI 에서 blastx를 사용한 경우 100개의 HSP 중 85번째로 검색되어 결과는 나왔으나 연체동물의 데이터를 식별하는데 약간의 어려움이 있었으며, tblastx 결과는 매우 오랜 시간이 걸렸으나 연체동물 관련 데이터를 찾을 수 없었다. 하지만 연체동물 전용데이터베이스에서 blastx를 사용한 경우 *Lymnaea stagnalis*의 soma ferritin, *Crassostrea gigas*의 ferritin GF1, ferritin GF2, *Pinctada fucata*의 ferritin-like protein, *Octopus dofleini*의 ferritin 등의 결과를 찾을 수

Table 2. Query sequences used for the following test.

species	description	type	accession No	length
Human	cathepsin	complete CDs	U20280.1	1597 bp
Human	ferritin	mRNA	NM_177478.1	894 bp
Human	cytochrome P450	mRNA	NM_000782.3	3295 bp

Table 3. Comparison of response time between NCBI and mollusks database.

Method / Query	NCBI			Mollusks-DB		
	blastx	tblastx	result	blastx	tblastx	result
U20280.1	1 min 30 sec	16 min	×	2 sec	5 sec	○
NM_177478.1	1 min	12 min	○	1 sec	4 sec	○
NM_000782.3	1 min 15 sec	Error	×	1 sec	8 sec	○

있었으며, tblastx 결과에선 EST연구가 많이 되어진 *Crassostrea gigas, Crassostrea virginica, Lymnaea stagnalis, Pinctada fucata, Biomphalaria glabrata*의 EST 결과와 상동성이 있음을 찾을 수 있었다. 마지막으로 cytochrome P450 mRNA 서열을 가지고 시험을 한 경우엔 NCBI 에서 blastx를 사용한 경우 역시 연체동물 관련 서열 정보를 찾을 수 없었으며 tblastx 는 에러로 인해 테스트를 할 수 없었다. 하지만 연체동물 전용데이터베이스에서 blastx 결과에선 *Lymnaea stagnalis, Haliotis rufescens, Mercenaria mercenaria, Mytilus galloprovincialis* 등의 cytochrome P450을 tblastx 결과에선 *Crassostrea virginica*의 아가미 (gill) 및 간췌장(hepatopancreas) 의 EST 서열들을 검색 할 수 있었다.

이러한 검색결과의 차이뿐만 아니라 검색시간의 경우도 NCBI에서 blastx 로 검색한 경우 보통 1분 이상 tblastx 로 검색 한 경우 10분 이상이거나 데이터가 큰 경우 에러로 인해 테스트 할 수 없는 경우가 많았으나 연체동물 전용 데이터베이스에선 10초 이내에 결과를 얻을 수 있었다 (Table 3). 또한 vector 및 *Escherichia coli* 시퀀스와 동시에 검색이 가능하므로 library 데이터 검증에 매우 용이하였으며, 특히 여러 개의 sequence를 동시에 검색할 수 있다는 장점도 있었다.

그리고 연체동물의 mitochondrial genome 정보는 굴족강 2 종, 다판강 1 종, 두족강 4 종, 복족강 6 종, 부족강 4 종 총 17 종 (Table 4.) 에 대한 정보가 NCBI 에 공개되어 있었는데, 이를 따로 모아 데이터베이스를 만든 결과 mitochondrial genome 내의 coding region을 query로 하여 검색하는 경우 매우 빠른 시간에 17 종의 mitochondrial genome 과 비교할 수 있었다. 이는 유전자를 이용한 베타분류 연구자의 경우 매우 유용한 정보가 될 수 있을 것이라 생각된다.

요 약

본 연구를 통해서 http://chimp.kribb.re.kr/mollusks 에 연체동물 전용 서열 BLAST 데이터베이스가 구축되었다. 예비 실험을 통해 본 결과와 마찬가지로 연체동물을 대상으로 한 유전자 정보만을 매우 빠른 속도로 얻을 수 있었다. 본 시스템을 사용하여 앞으로 많은 연구가 진행되어질 연체동물 유전자 연구 및 EST 연구에 많은 도움이 되리라고 사료된다.

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, **25**: 3389-3402.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**: 403-410.

Table 4. Mitochondrial genome sequences used for mollusk BLAST server.

species	Acc. No.	length	class
Biomphalaria glabrata	NC_005439.1	13670 bp	Gastropoda
Tbdarodes pacificus	NC_006354.1	20254 bp	Cephalopoda
Aplysia californica	NC_005827.1	14117 bp	Gastropoda
Octopus vulgaris	NC_006353.1	15744 bp	Cephalopoda
Graptacme eborea	NC_006162.1	14492 bp	Scaphopoda
Mytilus edulis	NC_006161.1	16740 bp	Pelecypoda
Haliotis rubra	NC_005940.1	16907 bp	Gastropoda
Siphonodentalium lobatum	NC_005840.1	13932 bp	Scaphopoda
Crassostrea gigas	NC_001276.1	18224 bp	Pelecypoda
Cepaea nemoralis	NC_001816.1	14100 bp	Gastropoda
Lampsilis ornata	NC_005335.1	16060 bp	Pelecypoda
Roboastra europaea	NC_004321.1	14472 bp	Gastropoda
Venerupis (Ruditapes) philippinarum	NC_003354.1	22676 bp	Pelecypoda
Loligo bleekeri	NC_002507.1	17211 bp	Cephalopoda
Pupa strigosa	NC_002176.1	14189 bp	Cephalopoda
Katharina tunicata	NC_001636.1	15532 bp	Polyplacophora
Albinaria coerulea	NC_001761.1	14130 bp	Gastropoda

Doctor, K.S., Reed, J.C., Godzik, A. and Bourne, P.E. (2003) The apoptosis database. *Cell Death and Differentiation*, 10: 621-633.

Ferlanti, E., Ryan, J., Makalowska, I. and Baxevanis, A. (1999) WebBLAST 2.0: An integrated solution for organizing and analyzing sequence data. *Bioinformatics,* 15: 422-423.

McGinnis, S. and Madden, T.L. (2004) BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research,* 32: W20-W25.

Muller A, MacCallum RM, Sternberg MJE (1999) Benchmarking PSI-BLAST in genome annotation. *Journal of Molecular Biology,* 293: 1257-1271.

Ng, K.W., Lawson, J. and Garner, H.R. (2004) PathoGene: A pathogen coding sequence discovery and analysis resource. *Biotechniques,* 37(2): 218, 220-212.

Ribeiro, J.M., Topalis, P. and Louis, C. (2004) AnoXcel: An *Anopheles gambiae* protein database. *Insect Molecular Biology,* 13: 449-457.