말 데이터베이스 구축

김대수[1] · 조운종[1] · 허재원[2] · 최은상[2,4] · 조병욱[3,4] · 김희수[1,2,4]*

[1]부산대학교 자연과학대학 생물정보학협동과정, [2]부산대학교 자연과학대학 생명과학부, [3]부산대학교 생명자원과학대학 동물자원학과, [4]부산대학교 말과학연구소

Received February 22, 2006 / Accepted May 4, 2006

HorseDB; an Integrated Horse Resource and Web Service. Dae-Soo Kim[1], Un-Jong Jo[1], Jae-Won Huh[2], Eun-Sang Choe[2,4], Byung-Wook Cho[3,4] and Heui-Soo Kim[1,2,4]*. [1]*PBBRC, Interdisciplinary Research Program of Bioinformatics, College of Natural Sciences, Pusan National University, Busan 609-735, Korea,* [2]*Division of Biological Sciences, College of Natural Sciences, Pusan National University, Busan 609-735, Korea,* [3]*Department of Animal Science, College of Life Sciences, Pusan National University, Miryang 627-706, Korea,* [4]*Horse Science Research Center, Pusan National University, Miryang 627-706, Korea* – We have built a database server called HorseDB which contains the genome annotation information and biological information for horse from public database entries. The aims of HorseDB are the integration of biological information and horse genome data on genome scale using bioinformatic methods. To facilitate the extraction of useful information among collected horse genome and biological data, we developed a user-friendly interface system, HorseDB; an Integrated Horse Resource and web Service. The database is categorized by the general horse information data, a sequence annotation data, and a world-wide web analysis program interface. The database also provides an easy access for user to find out the useful information within horse genomes and support analyzed information, such as sequence alignment and gene annotation results. HorseDB can be accessed at http://www.primate.or.kr./horse.

Key words – Horse, genome, bioinformatics

서 론

말은 대형 포유류의 하나이며, 발굽을 가진 대표적인 기제류이다. 분류학적으로 한 속이 존재하고 있으며 10개의 종이 현재 존재한다(Table1). 이들의 형태를 살펴보면 몸의 길이는 머리에서 엉덩이까지 약 2 m정도이며 몸무게 350~700 kg이고 꼬리길이 90 cm정도 이다. 체색은 흰색, 갈색, 검은색 등이 대부분지만, 간혹 다른 품종과의 교배를 통해서 얻어진 여러 색이 혼합된 품종도 존재한다. 수태 기간은 11개월 정도이며, 대부분 한배에 1마리를 낳는다. 그리고 평균수명은 25~35년 정도이다. 이들의 서식지는 온대림과 온대 초원지역이며 전세계에 걸쳐서 분포하고 있다. 현재 존재하는 말의 다양한 품종은 원산지에 따라 크게 5개의 품종 유럽 종, 아프리카 종, 아시아 종, 남북 아메리카 종, 대양주종으로 나눌 수 있는데, 이들의 분류에 대한 연구는 인간과 말의 관계가 가장 긴밀했던 19세기 후반부터 20세기 초기에 걸쳐 특히 독일, 프랑스를 중심으로 하여 활발히 이루어졌다. 오늘날에 지구상에 존재하는 말의 조상종에 관한 연구는 화석을 중심으로 활발히 이루어졌는데, 최초의 선조는 제 3기 시신세 초기에 북아메리카의 평원에서 최초로 출현된 Eohippus로서, 체구는 여우 정도 되었고, 발가락이 앞다리에 4개, 뒷다리에 3개가 있었다[3,5]. 그 뒤 다양한 진화 단계를 거친 말의 선조는 결국 현생의 말(Equus)이 되었다. 이렇게 진화된 말은 제4기 홍적세 초기에 베링해협을 통하여 아시아로 들어왔고, 이것이 유럽과 아프리카로 전파되었다[4,5]. 에쿠스(Equus)는 홍적세말부터 여러 종의 말로 분화되기 시작했는데, 대표적인 것이 프셰발스키말(Przewalski), 타르판말(Tarpan) 등이다. 프셰발스키말은 몽골 초원에서 발견된 대표적인 초원형 야생마의 일종으로 어깨높이 1.2 m, 머리가 크고, 다리는 짧으며, 갈기는 짧고 직립하며, 앞머리는 없고, 꼬리 기부의 털은 짧고, 몸은 회갈색인데 배와 입 끝은 백색으로 가축으로 이용되는 말과는 상당히 다르다. 이들은 몽골이나 고비사막 일대에 널리 분포하고 있었으나 현재는 몽골의 서쪽 끝의 사막에 국한되어 살고 있다. 타르판말은 동유럽에서 살았던 초원 형 야생마로서 앞머리가 있고 꼬리는 기부까지 긴 털로 덮여 있으며, 배가 백색이 아닌 점 등으로 미루어보아 가축으로 이용되는 말과 흡사하여 가축으로 이용되는 말의 조상종으로 생각되고 있다[4,5]. 말은 인간에게 중요한 가축의 하나로 전세계에서 널리 사육되고 있다. 말을 가축으로 키우기 전에는 인간의 식량을 위한 사냥의 대상물로 인식되었으나, 야생의 말을 길들여 키우기 시작하면서는 운송수단, 농업용, 군마 등으로 사용되었다. 그러나 현재의 말은 인간 노동력의 수단이 아닌 놀이수단으로 이용되기 시작하면서 승마나 경마용으로 중요하게 이용되고 있다.

말이 놀이수단(스포츠)에 많이 이용됨으로써 우수한 혈통을 가진 말을 생산하는 분야에서 많은 연구가 이루어지기 시작했다. 특히 말에 대한 유전학적인 연구는 우수한 말의 품종을 생산하고 육종하는 분야에서 고부가 가치를 창출할 수

*Corresponding author
Tel : +82-51-510-2259, Fax : +82-51-581-2962
E-mail : khs307@pusan.ac.kr

Table 1. Horse species and chromosome number

Genus	Chromosome number	Scientific Name	Common name
Equus	(33;Chromosome numbers)	*Equus przewalskii*	Przewalski's Horse(프제발스키 말)
		Equus caballus	Domestic Horse (말)
		Equus asinus	Donkey (당나귀)
		Equus africanus	African Wild Ass (아프리카 야생당나귀)
		Equus hemionus	Onager (야생당나귀, 오나거)
		Equus kiang	Kiang (캉당나귀)
		Equus quagga	Plains Zebra (초원 얼룩말)
		Equus zebra	Cape Mountain Zebra (산 얼룩말)
		Equus hartmannae	Hartmann's Mountain Zebra (하트만 산 얼룩만)
		Equus grevyi	Grevy's Zebra (그레비 얼룩말)

있는 중요한 분야로써 인식 되고 있다[4]. 이러한 유전학적인 연구는 다양한 품종의 말을 사용용도에 적합하도록 품종 개량을 가능하게 하여 우수한 능력을 가지는 말을 생산할 수 있게 함으로써 말을 이용한 스포츠 산업을 더욱 더 성장하게 하였다. 이러한 말 산업의 부흥에 힘입어 우수한 품종을 체계적으로 생산하고 계량하고, 유지하기 위한 일환으로 1995년에 국제 콘소시움이 형성되어 말의 지놈 프로젝트가 시작되었다[1,2,4]. 말의 지놈을 연구하는 중요한 목적은 첫째 말의 유전자의 지도를 완성하기 위한 것이고, 둘째 말 지놈 상에 분포하는 말의 유전자의 본질을 파악하는 것이며, 셋째 이들 유전자의 기능을 이해함으로써 말의 질병을 이해하기 위함이고, 마지막으로 다양한 포유류와의 비교분석을 통하여 이들의 유연관계를 이해하는 것이다[2]. 말의 지놈을 연구하게 됨으로써 얻을 수 있게 되는 상업적인 가치는 유전자의 분석을 통하여 우수한 혈통을 가지는 말 유전자의 다형성(single nucleotide polymorphisms, SNP)을 분석하여 우수한 말만이 가지는 특이적인 마커(marker) 유전자를 찾는 것이며, 또한 질병에 내성이 강한 말의 유전자를 비교 분석함으로써 말에서 발생하는 여러 질병을 진단 또는 예방할 수 있게 하는 것이다. 또한 우수한 유전자를 보유한 말의 계통도를 체계적으로 구축함으로써 용도에 적합한 우수한 말들을 선택적으로 생산 가능하게 하는 것이다.

그러므로 본 연구는 산재한 말에 대한 기초적인 정보들을 체계적으로 통합, 정리하여 말을 연구하고자 하는 연구자가 말에 대한 정보를 쉽고, 편리하게 얻을 수 있게 하고자 한다. 따라서 우리는 현존하는 말의 모든 정보(진화, 품종, 분포, 형태, 용도, expressed sequence tags (EST) 데이터, 지놈정보 등)를 통합하여 체계적으로 정리하여 데이터베이스를 구축하였다. 또한 본 데이터베이스는 계속해서 공개될 말 지놈서열의 공개에 대비하여 지놈서열과 유전자서열을 생물정보학적으로 분석하고 정리할 수 있는 기본적인 분석 환경을 제공하며 현재까지 공개되어있는 말의 모든 지놈서열을 미리 분석하여 이를 데이터로 제공함으로써 말을 연구하는데 필요한 중요한 기초적인 자료를 제공하고자 한다.

재료 및 방법

말 전사체 분석

본 연구에서 사용된 말의 지놈 데이터는 NCBI (http://www.ncbi.nlm.nih.gov/)의 GenBank Build 35에 포함 되어 있는 말의 전사체(expressed sequence tags) 서열이다. 말 전사체 서열을 분석하기 위해 TIGR (http://www.tigr.org/) 에서 전사체를 분석할 때 사용한 매뉴얼 분석작업을 응용하였고 이 때 TGICL (TIGR gene indices clustering tools)프로그램을 사용하였다(Fig. 1). 전사체 서열의 분석에 앞서 분석결과의 신뢰성과 정확성을 높이기 위해 100 bp 이하의 전사체 서열을 제거하였다. 이들 100 bp 이하의 전사체 서열은 이후 분석작업중 클러스터링(clustering)과 어셈블리(assembly) 작업할 때 오류를 발생시킬 수 있기 때문이다. 전처리 작업을 수행한 후 클러스터링(clustering) 작업을 수행하기 전에 전사체에 존재하는 반복서열을 제거하여(masking) 클러스터링(clustering)으로 생길 수 있는 오류를 줄였다. 클러스터링(clustering)에 사용한 프로그램은 BLAST (basic local alignment search tool) 알고리즘(algorism)을 이용하여 만든 프로그

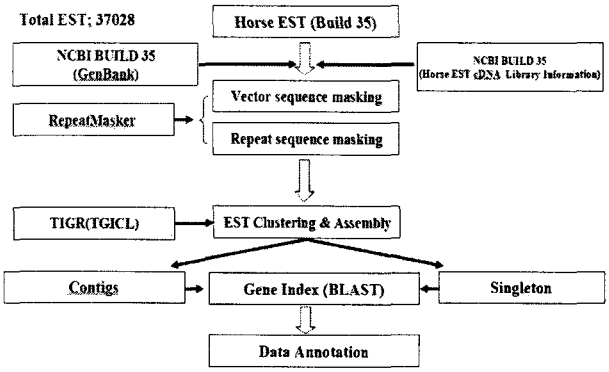

Fig. 1. Flow chart for analyzing horse ESTs. Horse ESTs were pre-processed and annotated by using both TGICL and BLAST program. The annotation of the ESTs was carried out by stand-alone BLAST programs and locally installed databases. The resulting data were stored to MySQL database.

램인 TGICL (TIGR gene indices clustering tools)의 클러스터링(clustering) 프로그램을 이용하였다. 클러스터링(clustering) 이후 어셈블리(assembly)과정에서는 CAP3라는 프로그램을 이용하였다. 어셈블리(assembly)된 말의 콘티그(contig) 데이터를 분석하기 위하여 BLAST (basic local alignment search tool)를 이용하여 분석된 결과를 유전자의 정보와 함께 서열이 정렬된 결과를 함께 제공하였다(Fig. 2).

말의 레트로엘레먼트 분석

말의 전사체에서 발현되는 레트로엘레먼트(retroelement)을 분석하기 위한 방법으로 말의 전사체서열을 RepeatMasker (http://www.repeatmasker.org/) 프로그램을 이용하여 분석하였다. 말에 대한 레트로엘레먼트의 정보가 많이 없었기 때문에 RepeatMasker 프로그램을 사용할 때 척추동물에 존재하는 모든 레트로엘레먼트를 라이브러리로 활용을 하였다.

말 데이터베이스 구축

말의 생물학적인 정보를 제공하며 말의 지놈서열을 분석할 수 있는 데이터베이스를 구축하여 말을 연구하는 연구자로 하여금 양질의 정보를 제공하게 하였다. 말 지놈 분석용 웹 BLAST를 구축하기 위해서 공개 데이터베이스인 NCBI GenBank에서 말의 지놈서열을 모두 이용하여 BLAST용 로컬데이터베이스를 구축하여 말 지놈분석을 용의하게 하였다. 또한 생물정보학적인 분석방법을 통하여 분석한 말의 유전자와 인간 유전자와의 비교분석 한 결과를 체계적으로 정리하여 분석데이터로 이용하였다. 또한 말의 분류정보는 NCBI의 taxonomy browser와 인터넷상에 공개된 데이터를 통합하여 사용하였으며, 말의 생물학적인 데이터는 인터넷상에 공개되어있는 정보들을 모두 이용하여 화석, 품종, 분포정보 등 체계적으로 데이터를 분류하여 데이터베이스를 구축하였다(Fig. 3).

말의 생물학적 정보 수집 및 웹 인터페이스

말의 용도가 농경이나 이동수단이 아닌 승마나 경마와 같은 레저수단으로 변화하여 우수한 말 품종을 육종하고 이를 보존하는데 모든 연구가 집중되어 있으며, 사업적으로 매우 가치 있기 때문에 다른 생물 종들과 달리 품종과 진화에 대한 연구가 많이 진행되어 말의 품종과 진화에 대한 많은 정보들이 공개되어 있다(Table 2). 이러한 정보들을 체계적으로 정리하기 위하여 품종에 대한 정보를 정리할 때 말의 원산지, 체격, 모색, 용도, 분포 등의 정보를 정리하였다. 또한 말의 분포도를 나눌 때에는 크게 유럽 종, 아프리카 종, 아시아 종, 남북 아메리카 종, 대양주 종 5개의 품종으로 구분하여 정리하였다. 말 지놈 데이터베이스는 MySQL을 이용하였으며 웹서버는 아파치(apache)를 사용하였다.

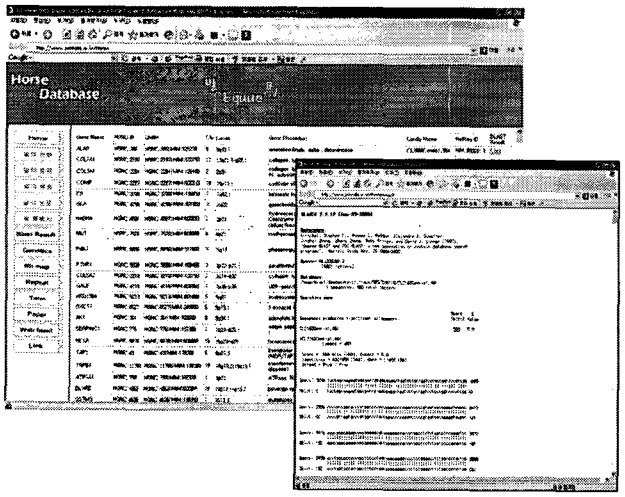

Fig. 2. Horse EST annotation and BLAST result. The individual high-quality ESTs were searched against the human RefSeq mRNA. We carried out a BLAST search with a cut-off identity of 85% and an E value of 1e-10.

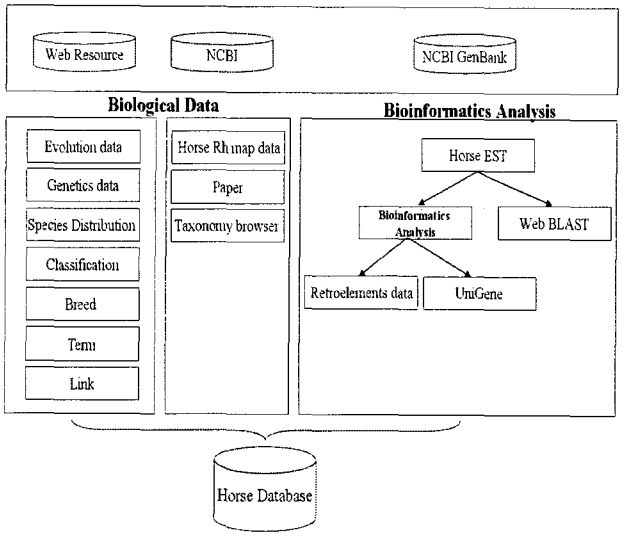

Fig. 3. System schema of horse database.

Table 2. Horse resource on the internet

Web site	URL	Discription
Breeds of Livestock	http://www.ansi.okstate.edu/breeds/	Department of Animal Science - Oklahoma State University
Horse Breeds of the World	http://www.imb.org/imh/bw/	The International Museum of the Horse
Horse Evolution	http://www.talkorigins.org/faqs/horses/horse_evol.html	Horse Evolution
NCBI Entrez Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi	NCBI taxonomy database
Fossil Horse Cybermuseum	http://www.flmnh.ufl.edu/natsci/vertpaleo/fhc/firstCM.htm	internet browsers gallery of fossil horses

결과 및 고찰

말의 유전자에서 발현되는 레트로엘레먼트 분석

현존하는 거의 모든 종은 진화 과정 동안 레트로엘레먼트의 전사체가 지놈 내로 유입되어 현재까지도 세포 내 유전자와 함께 발현되고 있다. 이렇게 발현되는 레트로엘리먼트는 지놈 분석을 위한 cDNA 제작시 이들 서열에 포함되어 실제 세포 내 유전자와 함께 분석되고 있다[9]. 이러한 레트로엘레먼트를 포함한 전사체는 직·간접적으로 질병을 발생시키기도 하고 또한 생물학적으로 중요한 기능을 담당하기도 한다. 그러므로, 말 전사체에서 레트로엘레먼트를 분석함으로써 말 지놈의 특성을 이해할 수 있으며 말의 유전자의 발현에 영향을 미치는 레트로엘레먼트를 분석함으로써 말에서 발생하는 특이적인 질병의 원인을 분석할 수도 있다(Fig. 4).

말 지놈 분석 프로그램

NCBI 데이터베이스에서 말 지놈 서열과 전사체 서열을 다운로드 받아, 말 지놈 DB를 구축하였다. 또한 이러한 말 지놈 DB를 사용자가 편리하게 이용하도록 하기 위해서 말 지놈분석용 웹 BLAST를 구축하였다. 말 지놈분석용 웹 BLAST를 통해 말 유전자 정보를 분석할 수 있으며, 현재까지 분석된 다양한 전사체 서열에 대한 검색이 가능하다. 이러한 말 DB는 말 지놈 분석을 원하는 연구자에게 어떤 다른 데이터 베이스보다 편리한 분석 환경을 제공할 수 있을 것이며, 빠른 정보 검색을 가능하게 할 것이다. 또한 말과 근연종인 소의 유전자 서열을 비교 분석하고 이를 인간의 유전자와 통합해서 연구할 수 있게끔 인간 Unigene과 소의 Unigene의 데이터도 또한 검색 가능하도록 웹 환경을 구축하였다(Fig. 4).

생물학적 정보의 통합 및 웹 환경 구축

현재 말의 생물학적인 연구는 인류가 정착생활을 하게 되고 경제활동이란 걸 시작하면서 말이 자연스럽게 농경에 사용되고 소와는 달리 전차를 끌거나 이동수단으로서 이용되어 다른 가축들 보다 더욱 인간에게 중요한 동물이었기 때문에 현존하는 동물 종에서 말만큼 그 진화과정이 잘 알려진 동물도 거의 없다. 인간에게 중요한 말의 진화정보를 제공함으로써 말을 연구하는 연구자들에게 기초적인 말의 진화정보를 제공하고, 말의 사용용도에 따른 우수한 말의 정보와 분포에 대한 정보를 제공함으로써 말의 연구에 도움이 되고자 하였다.

또한 이러한 다양한 정보를 말 지놈 데이터베이스로 통합하여 말의 정보를 쉽게 검색할 수 있게 하기 위하여 사용자의 편이의 데이터베이스를 구축하였으며 본 데이터베이스에는 말의 진화에 대한 정보, 말의 종류 정보, 말의 품종 정보, 말의 분포 정보, 말 전사체 분석결과 정보, 말의 유전학적인 연구 정보, 레트로엘레먼트 분석 정보 등 총 12개 정도의 말에 대한 자료와 웹 분석용 프로그램을 제공하였다(Fig. 5).

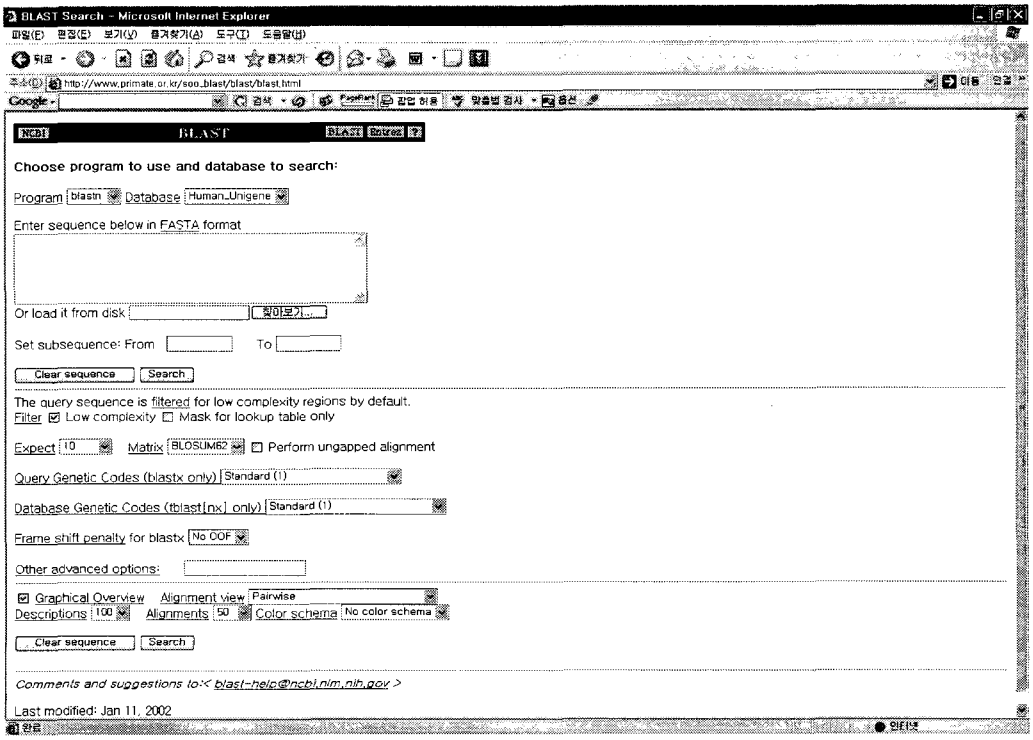

Fig. 4. Standalone web BLAST server. This viewer allows the BLAST searching of the user's query against horse EST, genome, human Unigene, and cow Unigene.

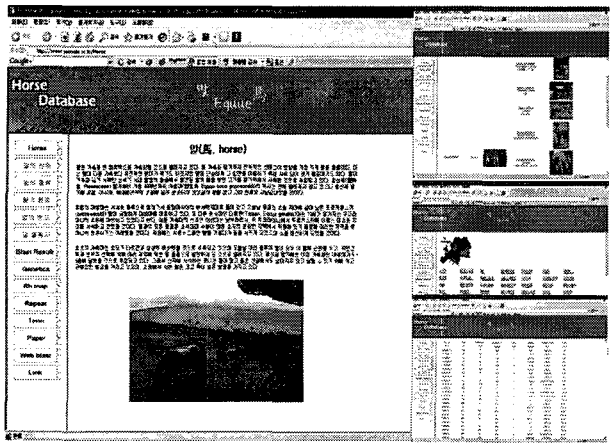

Fig. 5. HorseDB user interface. The HTML interface for HorseDB is made by using python and PHP. There are five web pages: The main page, biological data, retroelement analysis, EST analysis, and analysis tools. The user can easily search variable information about the horse including genome sequence, annotation information, retroelement information, biological information by using the database.

요 약

공개된 데이터베이스들에서 말에 대한 생물학적인 데이터와 지놈 데이터를 분석하여 말 데이터베이스를 구축하였다. 말 데이터베이스는 말의 생물학적인 데이터와 지놈 데이터를 생물정보학적인 분석방법으로 분석하고 이들 데이터를 통합하여 제공하는데 목적을 두고 있다. 본 데이터베이스는 말의 생물학적 데이터와 지놈 분석 데이터 그리고 생물정보학적 인 분석프로그램을 제공하는 인터페이스로 구성하였다. 또한 사용자의 편의를 돕기 위해서 쉽게 이용할 수 있도록 웹 메뉴를 구성 하였으며 말에 대한 다양한 정보를 제공할 수 있게 하였다. 말 데이터베이스를 이용할 수 있는 웹 주소는 http://www.primate.or.kr/horse이다.

감사의 글

본 연구는 부산대학교 연구비의 (2년과제) 지원으로 수행되었습니다.

참 고 문 헌

1. Bennett, D .K. 1986. The origins of breeds. *Equus* **110**:33, **11**:37, **112**:37.
2. Brinkmeyer-Langford, C., Raudsepp, T., Lee, E. J., Goh, G., Schaffer, A. A., Agarwala, R., Wagner, M. L., Tozaki, T., Skow, L. C., Womack, J. E., Mickelson, J. R., Chowdhary, B. P. 2005. A high-resolution physical map of equine homologs of HSA19 shows divergent evolution compared with other mammals. *Mamm. Genome.* **16** (8): 631-649.
3. Theodosius Dobzhansky. 1951. *Genetics and the Origin of Species,* 3rd Ed. Columbia University Press, New York.
4. Guérin, G., Bailey, E., Bernoco, D., Anderson, I., Antczak, D. F., Bell, K., Binns, M. M., Bowling, A. T., Brandon, R., Cholewinski, G., Cothran, E. G., Ellegren, H., Forster, M., Godard, S., Horin, P., Ketchum, M., Lindgren, G., McPartlan, H., Meriaux, J-C., Mickelson, J. R., Millon, L. V., Murray, J., Neau, A., Roed, K., Sandberg, K., Shiue, Y-L., Skow, L. C.,. 1999. Report of the International Equine Gene Mapping Workshop: male linkage map. *Animal Genetics.* **30**, 341 - 354.
5. Weinstock, J., Willerslev, E., Sher, A., Tong, W., Ho, Simon. Y. W., Rubenstein, D., Storer, J., Burns, Js., Martin, Ly., Bravi, C., Prieto, A., Froese, D., Scott, E., Xulong, L., Cooper, A. 2005. Evolution, Systematics, and Phylogeography of Pleistocene Horses in the New World: A Molecular Perspective. *PLoS. Biol.***3** (8): e241.
6. MacFadden, B. J. 1992. Fossil Horses: Systematics, Paleobiology, and Evolution of the Family Equidae. New York, Cambridge University Press.
7. Radinsky, L. 1983. Allometry and reorganization in horse skull proportions. *Science.* **221** (16):1189-1191.
8. Renders, E. 1984. The gait of *Hipparion* sp. from fossil footprints in Laetoli, Tanzania. *Nature.* **308**:179-181.
9. Sorek, R., Ast, G., Graur, D. 2002 .Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060-1067.