Development of Primer and Probe Design System for Microbial Identification

미생물 동정을 위한 프로브와 프라이머 고안 시스템의 개발

Junhyung Park[1], Byeongchul Kang[2], Heekyung Park[1], Hyunjung Jang[3], Eunsil Song[3], Seungwon Lee[1], HyunJin Kim[1], Cheolmin Kim[1*]

[1] Busan Genome Center, College of Medicine, Pusan National University, Busan, South Korea
[2] Devision of Applied Bioengineering, Dongseo University, Busan, South Korea
[3] Institute for Genomics Medicine , GeneIn. Co., Ltd., Busan, South Korea

*To whom correspondence should be addressed. E-mail: kimcm@pusan.ac.kr

Abstract

모든 생명체의 genetic information에는 보존적 염기서열과 다형적 염기서열이 존재한다. 다형적 염기서열과 보존적 염기서열은 하나의 종(species)을 감별하거나, 여러 종류의 종을 동시에 감별할 수 있는 genotyping의 표지자로 각각 이용될 수 있다. 본 논문은 병원성 감염질환 세균, 식중독 유발 세균, 생물의약품 오염 유발 세균 및 환경오염 세균 등 세균의 존재 유무와 속과 종 감별을 위해 대부분 세균 종의 보존적 염기서열과 다형적인 염기서열을 포함하고 있는 23S rDNA 유전자의 표적 염기 서열로부터 고안된 세균 특이적(bacterial-specific), 속 특이적(genus-specific), 종 특이적(species-specific) 올리고 뉴클레오티드 프로브와 프라이머를 디자인하는 시스템을 소개한다. 시스템을 통해서 얻어진 프로브와 프라이머들은 PCR을 통한 검증단계를 거쳐서 디자인 결과의 정확성을 확인하였다. 본 시스템의 이용으로 프로브와 프라이머를 디자인하는데 몇 주가 소요되는 시간을 몇 일 내로 줄일 수 있었으며, 체계적인 데이터의 관리로 결과의 정확성을 높일 수 있었다.

Introduction

세균을 동정하기 위해 널리 사용되고 있는 유전자는 세균에서 매우 보존적인 공통서열로 나타나는 16S rDNA를 기초하고 있다. 하지만 염기 서열 변이 영역이 작아서 일부 특정 세균의 감별이 어려운 한계가 있다. 최근에는 과변이 영역(hypervariable region)을 보유한 ITS(internal transcribed spacer region, 내부전사지역)을 이용하거나, 아직 염기서열 정보가 많이 밝혀지지 않았지만, 16S rDNA 보다 서열 길이가 평균적으로 1K 이상 길고, 서열 변이 영역이 보다 많은 것으로 알려진 23S rDNA

본 연구는 과학기술부 지역기술개발용역사업(부산-0102) 지원으로 수행되었음.

유전자를 기초로 세균을 검출하는 방법이 있으며, 본 연구에서도 23S rDNA 유전자를 이용하여 세균을 검출 및 genotyping에 필수적인 프로브와 프라이머를 디자인하는 시스템을 개발하고자 하였다. 세균을 검출하기 위해 PCR과 DNA chip을 많이 사용하고 있으며, 이 두 가지 방법을 사용하기 위해서는 프라이머와 프로브 디자인 작업이 선행되어야 한다.

많은 연구실에서 공개 툴인 Primer3 또는 Oligo와 같은 소프트웨어를 사용하여 프로브, 프라이머 세트를 디자인하고 있다. 이와 같은 소프트웨어는 주어진 시퀀스와 프로브, 프라이머의 조건에 맞는 파라메터 입력에 의해서 프로브, 프라이머를 디자인하지만, 주어진 시퀀스가 종 특이적인지 속 특이적인지 확인해야 하는 BLAST 검색과는 연계되어 있지 않으므로 일일이 BLAST 검색 툴을 이용하여 확인해야 하는 불편함이 있다.

병원성 세균의 경우에는 세균의 존재 유무를 먼저 1차 스크리닝하여 세균의 존재를 확인한 후 각 원인균 속의 정확한 감별을 위해 2차 스크리닝을 실시함으로써 진단 비용의 절감과 함께 항생제 오남용의 예방을 위한 적절한 치료가 이루어질 수 있다. 이처럼 세균의 존재 유무와 원인균 속 특이적, 종 특이적 감별을 위해서는 각각의 프로브 또는 프라이머 세트가 필요하다. 하지만 대부분의 프로브, 프라이머 디자인 소프트웨어들이 한 균종에 대하여 특이적 프로브와 프라이머만을 디자인하는데 초점이 맞추어져 있으며, 속 특이적, 세균 특이적 프로브와 프라이머를 포함하여 디자인하는 소프트웨어는 없다. 또한 유전자 데이터의 수집부터 프로브, 프라이머 디자인의 결과물까지 상당히 많은 데이터가 발생하므로, 이를 체계적으로 관리하고 추후 업데이트하는 시스템이 요구되지만 그러한 시스템은 찾아볼 수 없다.

따라서 본 시스템은 세균, 속, 종 특이적 프로브와 프라이머를 디자인하기 위해 GenBank에서 얻어진 23S rDNA 유전자와 Web-Lab에서 직접 시퀀싱을 수행한 데이터를 수집한다. 그리고 종(species) 또는 속(genus)에 따라서 다중서열정렬 수행하며, 다중서열정렬 후 얻어지는 각각의 일치서열(consensus sequence)을 관리한다. 또한 윈도우 슬라이딩과 욕심쟁이 알고리즘(greedy algorithm) 및 디자인에 필요한 파라메터들을 이용하여 프로브와 프라이머 후보군을 추출한다. 최종적으로 BLAST의 자동검색을 통해 후보군들의 세균, 속, 종 특이적 프로브와 프라이머 유무를 확인할 수 있다. BLAST 검색 결과는 Taxonomy Browser형태의 이미지로 표현하여 판단을 보다 용이하게 할 수 있도록 하였으며, 그 결과는 데이터베이스에 업로드되도록 설계되었다.

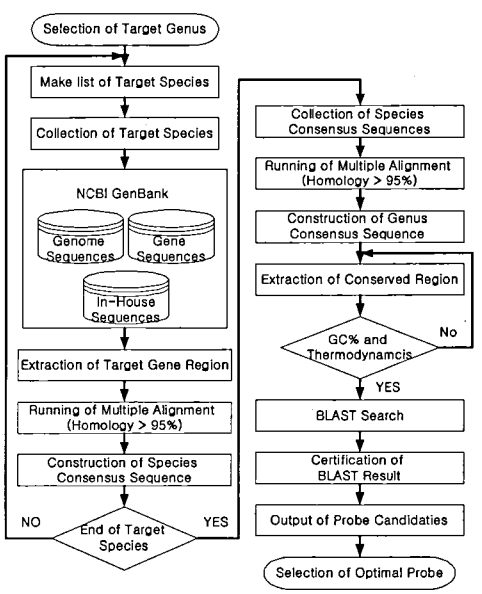


그림 1. 속 특이적 프로브와 프라이머 디자인 흐름도

그림 1은 속 특이적인 프로브와 프라이머를 디자인하기 위한 흐름을 나타낸 것이며, 세균 특이적, 종 특이적 프로브와 프라이머를 디자인하기 위해서도 이와 유사한 단계의 절차를 거쳐야 된다.

위와 같은 각 과정을 수작업으로 수행한다는 것은 상당한 시간이 소요될 뿐만 아니라 처리하고 관리해야 될 데이터의 양이 기하급수적으로 늘어나 체계적인 관리가 거의 불가능하며, 추후 데이터의 업데이트에서도 문제점이 발생할 수 있다.

따라서 이와 같은 흐름을 데이터 수집 단계, 다중서열정열을 통한 일치서열 구축 단계, 프로브와 프라이머를 디자인하는 단계, 마지막으로 디자인된 프로브와 프라이머의 블라스트 검증 단계로 구분하여 설계를 하였으며, 아래에서 그 내용을 자세히 다루고자 한다.

System Development

본 시스템은 리눅스 시스템을 기본 사양으로 구성하였으며, Perl과 Bioperl을 이용하여 프로그램하였다. 또한 웹형태로 구현하기 위해 Apache 서버를 기반으로 한 HTML과 CGI 형태로 프로그램하였으며, 데이터베이스는 MySQL을 사용하였다.

Data Collection

데이터의 수집은 크게 두 부분으로 나누어진다. 첫 번째는 GenBank의 Entrez 검색을 통해서 얻어지는 시퀀스 데이터이며, 두 번째는 Web-Lab에서 직접 시퀀싱을 통해서 얻어지는 시퀀스 데이터이다.

GenBank에서 얻어지는 데이터는 Complete Genome 데이터와 23S rDNA 유전자 데이터로 나누어지며, 수많은 유전자 서열을 포함하고 있는 Complete Genome 데이터에서는 다시 23S rDNA 유전자를 추출하여야 한다. 최종적으로 23S rDNA 시퀀스 데이터만을 각 속과 종에 따라서 데이터를 수집하여 데이터베이스 형태로 구축한다. 잘못된 데이터가 포함되면 이후의 모든 결과가 틀릴 수 밖에 없기 때문에 데이터 수집은 가장 기본적이면서 가장 중요한 단계이다.

따라서 검색 로봇을 이용하여 완전 자동으로 데이터를 수집하는 방법보다 연구자가 좀 더 쉽게 데이터를 검색하고, 연구자의 선별과 확인에 의해서 데이터가 수집되어 데이터베이스에 구축되도록 하는 시스템의 설계가 요구된다. 본 시스템은 이와 같은 반자동 형태의 데이터 수집 방법으로 설계하였으며, NCBI의 nt 데이터를 로컬화하여 자체 개발한 필터링을 거쳐서 23S rDNA 유전자를 검색할 수 있도록 하였다.

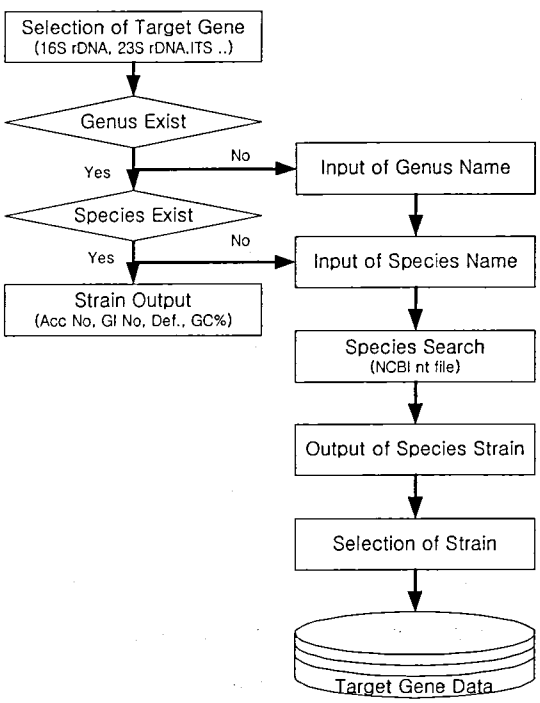


그림 2. 데이터 수집 흐름도

그림 2는 웹으로 데이터를 수집하는 방

법을 나타내고 있다. 가장 먼저 대상 유전자(23S rDNA)를 선정하면 로컬화된 유전자 데이터베이스를 검색해서 데이터가 존재하는지를 살펴본 후 데이터가 없으면 로컬화된 NCBI의 nt파일에서 Definition을 검사하여 찾고자 하는 데이터의 Definition을 보여준다. 제시된 데이터에서 연구자는 체크박스를 선택한 후 클릭하여 데이터베이스로 업데이트를 수행하도록 설계하였다.

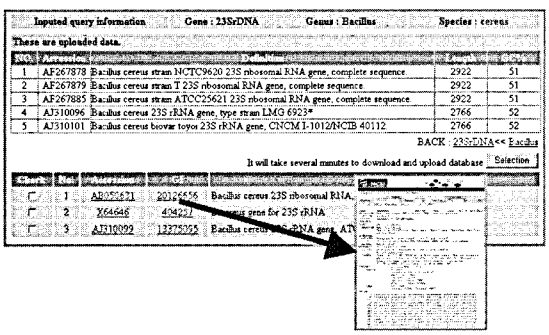


그림 3. 데이터 수집 화면

그림 3은 Bacillus cereus의 23S rDNA 유전자를 수집하는 인터페이스를 보이고 있다. 이미지의 상단은 이미 로컬화된 유전자 데이터베이스에 구축된 데이터이고, 하단은 아직 데이터베이스에 구축되지 않은 데이터이다. GenBank와 링크되어 있는 Accession 또는 GenBank ID를 클릭하여 찾고자하는 데이터가 맞는지 확인 후 체크박스를 선택하도록 하며, 최종적으로 Selection 버튼을 클릭하면 데이터는 유전자 데이터베이스에 구축된다.

Construction of Consensus Sequence

일치서열(consensus sequence)의 생성은 속(genus)과 종(species)의 분류에 따라 수집된 데이터를 이용하여 수행된다. 하나의 종은 다양한 strain으로 구성되어 있으며, 속은 다양한 종으로 구성되어 있다. 따라서, 종 일치서열은 strain 시퀀스들의 다중서열정렬을 통해서 얻을 수 있으며, 속 일치서열은 종 일치서열들의 다중서열정렬을 통해서 얻을 수 있다. 이와 같이 일치서열구성은 세균, 속과 종의 계층적인 구조형태를 가지고 있으며, 23S rDNA 유전자의 경우에는 새롭게 밝혀질 부분이 많이 있으므로, 수집된 데이터가 업데이트됨에 따라서 일치서열구축 단계도 업데이트 가능하도록 설계하였다.

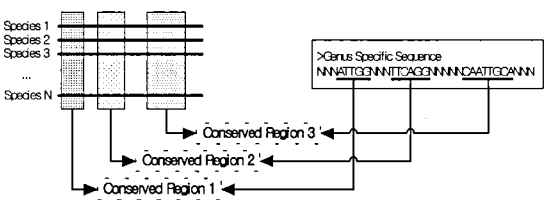


그림 4. 일치서열 생성 및 보존영역의 선택

그림 4는 속(genus) 일치서열을 구성하기 위한 다중서열정렬을 보여주고 있으며, 일치서열에서 프로브와 프라이머를 디자인하기 위한 보존적인 영역의 선택을 나타내고 있다.

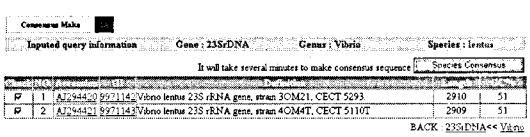


그림 5. 종(specie) 일치서열 구축 화면

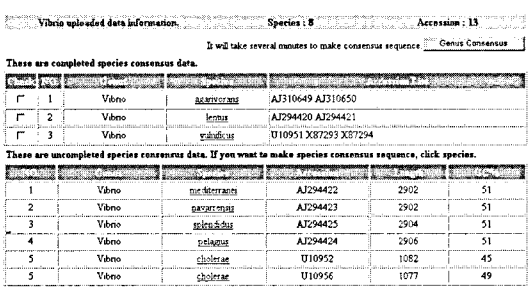


그림 6. 속(genus) 일치서열 구축 화면

그림 5는 종(species) 일치서열을 구축하기 위한 인터페이스를 보여주고 있다. 그림 6은 속(genus) 일치서열을 구축하기 위한 인터페이스로써, 그림 5에서 생성된 일치서열의 업데이트와 기존에 생성된 다른 종(species)의 일치서열 및 아직 일치서열을 구축하지 않은 종 데이터를 보여주고 있다.

Construction of Candidate Probe and Primer

프로브와 프라이머의 후보군은 이전 단계에서 구축된 일치서열을 이용하여 생성된다.

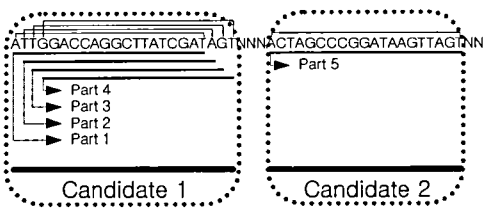


그림 7. 윈도우 슬라이딩과 욕심쟁이 알고리즘에 의한 후보군의 선별

그림 7은 일치서열에서 윈도우 슬라이딩과 욕심쟁이(greedy) 알고리즘에 의해서 후보군을 찾는 방법을 보이고 있다. 프로브, 프라이머의 최소, 최대 길이가 설정되면 일치서열의 보존적인 영역에서 설정된 길이에 맞게 후보 영역들을 검색해 나간다. 이렇게 검색된 후보 영역들은 프로브, 프라이머를 디자인하는데 요구되는 파라메터인 서열상의 G와 C의 함량, 열역학적 조건 등을 검토하여 후보군으로 설정한다. 설정된 후보군들은 데이터베이스에 업로드되어 BLAST를 통한 검증 단계를 거치게 된다.

그림 8은 데이터베이스로 구축된 프로브, 파라이머의 후보군들을 테이블과 이미지형태로 나타낸 것이며, 이미지를 통해서 프로브, 프라이머 후보군이 많이 존재하는 위치를 파악할 수 있다.

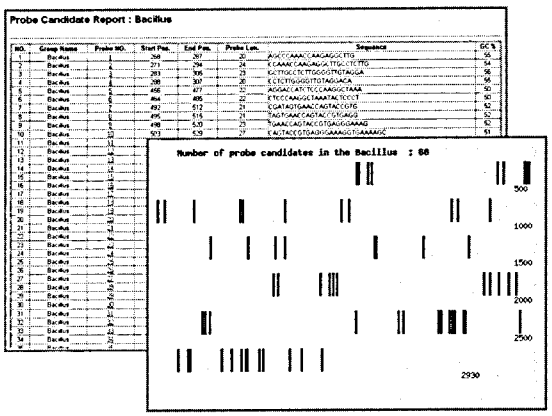


그림 8. 프로브, 프라이머 후보군의 화면

Validation of Candidate Probe and Primer

프로브, 프라이머의 후보군들은 BLAST 검색을 통해서 적절성을 확인하게 된다. 본 시스템은 데이터베이스에 업로드되어 있는 각 프로브, 프라이머 후보군들을 호출하여 자동으로 NCBI의 nr(non-redundant) 데이터베이스에 BLAST를 수행하고 결과를 데이터베이스에 구축하였다. 또한 결과를 Taxonomy Browser 형태의 이미지로 표현하여 좀 더 쉽고 정확하게 프로브, 프라이머 세트의 특이성을 확인하도록 하였다.

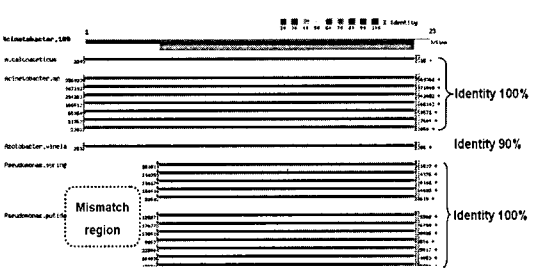


그림 9. Taxonomy Browser 형태의 BLAST 검색 결과 화면

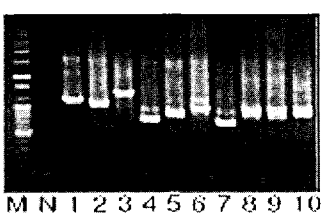
그림 9는 Taxonomy Browser 형태로 BLAST 결과를 이미지로 표현한 것이다. 검출하고자 하는 대상 세균과 100% 일치하고 다른

세균과 일치하지 않는 영역이 중간부위에 있는 후보군들이 프로브의 후보로써 적당하고, 일치하지 않는 영역이 3' 부위에 있는 후보군들이 프라이머의 후보로써 적당하므로 그림 9와 같은 Taxonomy Browser 형태의 이미지는 프로브와 프라이머의 후보군을 선별하는데에도 쉽게 이용될 수 있다.

Experiments

본 시스템을 통해서 41개의 속과 133개의 종에서 1349개의 프로브, 프라이머 후보군을 디자인하였다. 디자인된 프로브, 프라이머의 후보군들 중 PCR을 통해서 결과의 정확성을 검증하였다.

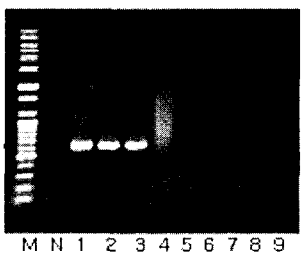
그림 10은 디자인된 프라이머 중 한 세트를 선정하여 PCR기법으로 검증한 예이며, 10가지 속에 대해서 모두 반응을 보이고 있는 세균 특이적 프라이머의 PCR결과이다.



M: 100bp분자량표시 N: 증류수(음성대조군) 1: *Acinetobacter baumannii* 2: *Aeromonas salmonicida* 3: *Bacteroides forsythus* 4: *Clostridium difficile* 5: *Legionella pneumophilia* 6: *Prophyromonas asaccharolytica* 8: *Proteus mirabilis* 9: *Mycobacterium tuberculosis* 10: *Mycoplasma pneumoniae*

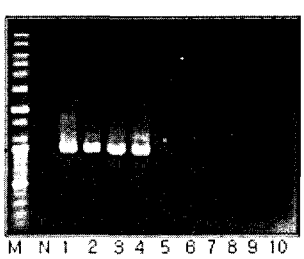
그림 10. 세균 특이적 프라이머의 PCR 결과

그림 11은 *Enterococcus* 속 특이적 프라이머의 PCR 결과이다. *Enterococcus*, *Aeromonas*, *Mycobacterium* , *Streptococcus*, Hepetitis B virus DNA 등 여러가지 속이 포함된 실험 균주 중에서 *Enterococcus* 속 (*Enterococcus faecalis*, *Enterococcus faedium*, *Enterococcus hirae*)특이적인 결과들이 정확하게 구분하여 동정하는 것을 확인할 수 있었다. 그림 12와 13은 *Mycobacteria* 속 특이적, *Streptococcus* 속 특이적 프라이머 PCR 결과이다.



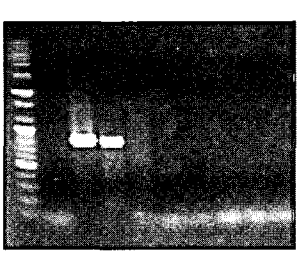
M: 100bp분자량표시 N: 증류수(음성대조군) 1: *Enterococcus faecalis* 2: *Enterococcus faedium* 3: *Enterococcus hirae* 4: *Aeromonas hydrophila* 5: *Mycobacterium xenopii* 6: *Mycobacterium falconis* 7: *Streptococcus anginosus* 8: Human blood DNA 9: Hepetitis B virus DNA

그림 11. *Enterococcus* 속 특이적 프라이머의 PCR 결과



M: 100bp분자량표시 N: 증류수(음성대조군) 1: *Mycobacterium xenopi* 2: *Mycobacterium flavescence* 3: *Mycobacterium simiae* 4: *Mycobacterium tuberculosis* 5: *Aeromonas hydrophila* 6: *Mycobacterium falconis* 7: *Streptococcus anginosus* 8: *Enterococcus faecalis* 9: Human blood DNA 10: Hepatitis B virus DNA

그림 12. *Mycobacteria* 속 특이적 프라이머의 PCR 결과



M: 100bp분자량표시 N: 증류수(음성대조군) 1: *Streptococcus anginosus* 2: *Streptococcus bovis* 3: *Aeromonas hydrophila* 4: *Mycobacterium falconis* 5: *Mycobacterium xenopi* 6: *Enterococcus faecalis* 7: Human blood DNA 8: Hepatitis B virus DNA

그림 13. *Streptococcus* 속 특이적 프라이머의 PCR 결과

Conclusion

본 연구는 23S rDNA 유전자를 이용하여 세균의 동정을 위한 프로브와 프라이머를 디자인하는 시스템 개발이다. 병원성 세균의 경우에는 세균의 존재 유무를 먼저 1차 스크리닝한 후 세균의 존재가 밝혀지면 각 원인균 속의 정확한 감별을 위해 2차 스크리닝을 실시함으로써 진단 비용의 절감과 함께 항생제 오남용의 예방 및 적절한 치료가 이루어 질 수 있는 신속하고 민감한 진단 방법이 제공되어야 한다. 그러나 세균의 존재 유무와 원인균 속 특이적, 종 특이적 감별을 위해서는 각각의 프로브 또는 프라이머 세트가 필요하지만 기존의 소프트웨어들은 종 특이적이거나 속 특이적인 프로브, 프라이머만을 디자인하는 한계가 있다. 따라서 본 연구에서는 각 단계의 특이적인 프로브, 프라이머를 디자인 할 뿐만 아니라 데이터의 업데이트 및 관리를 수행할 수 있는 통합시스템을 개발하였다. 또한 프로브, 프라이머를 디자인하기 위해 필요한 데이터의 수집과 BLAST를 통한 검증도 웹으로 자동으로 구현할 수 있도록 하였으며, 모든 결과는 데이터베이스에 구축되도록 설계하였다.

본 시스템을 개발하기 전 수작업으로 각 단계별 프로브, 프라이머를 디자인 하였을 때, 수 개월이라는 시간이 소요되었으며, 엄청나게 많은 데이터를 관리해야 하는 번거로움이 있었다. 하지만 본 시스템을 사용하여 동일한 과정을 반복한 결과 하루, 이틀 만에 동일한 결과를 얻을 수 있었으며, 예전의 잘못된 디자인 결과물도 검출할 수 있었다.

따라서 본 시스템은 미생물의 동정을 위한 PCR과 DNA chip 개발에 많은 이바지할 것으로 생각된다. 향후 23S rDNA 유전자 뿐만 아니라 미생물들이 보유하고 있는 필수 유전자를 이용하여 이와 같은 프로브, 프라이머를 디자인하고자 한다.

본 연구에서 개발된 시스템의 활용이나 고안된 프로브와 프라이머의 자세한 내용은 교신저자(부산대학교 의과대학 김철민교수 kimcm@pusan.ac.kr)와 협의하여 활용할 수 있습니다.

Acknowledgements

본 연구는 과학기술부 지역기술개발용역사업의 지원에 의하여 이루어진 것임.(부산-0102)

References

[1] Alexander E Pozhitkov and Diethard Tautz, An algorithm and program for finding sequence specific oligo-nucleotide probes for species identification, *BMC Bioinformatics*, 3:9(2002)

[2] Alexander Loy, Matthias Horn and Michael Wagner, probeBase: an online resource for rRNA-targeted oligonucleotide probes, *Nucleic Acids Research*, Vol.31, No. 1,pp.514-515(2003)

[3] Henrik Bjorn Nielsen, Rasmus Wernersson and Steen Knudsen, Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays, *Nucleic Acids Research*, Vol. 31, No. 13, pp 3491-3496(2003)

[4] Dong Xu, Guangshan Li, Liyou Wu, Jizhong Zhou and Ying Xu, PRIMEGENS: robust and efficient design of gene-specific probes

for microarray analysis, *Bioinformatics*, Vol. 18 no. 11, pp 1432-1437(2002)

[5] Fugen Li and Gary D. Stormo, Selection of optimal DNA oligos for gene expression arrays, *Bioinformatics,* Vol.17 no.11 pp1067-1076(2001)

[6] James Borneman, Marek Chrobak, Gianluca Della Vedova, Andres Figueroa and Tao Jiang, Probe selection algorithms with applications in the analysis of microbial communities, *Bioinformatics,* Vol.17 Suppl. 1, pp S39-S48(2001)

[7] T.Z.DeSantis, I. Dubosarskiy, S. R. Murray and G. L. Andersen, Comprehensive aligned sequence construction for automated design of effective probes(CASCADE-P) using 16S rDNA, *Bioinformatics,* Vol.19 no. 12, pp 1461-1468(2003)

[8] Xiaowei Wang and Brian Seed, Selection of oligonucleotide probes for protein coding sequences, *Bioinformatics,* Vol.19 no. 7, pp 796-802(2003)

[9] Vincent Thareau, Patrice Dehais, Carine Serizet, Pierre Hilson, Pierre Rouze and Sebastien Aubourg, Automatic design of gene-specific sequence tags for genome-wide functional studies, *Bioinformatics,* Vol.19 no. 17, pp 2191-2198(2003)

[10] Simon N. Jarman, Amplicon:software for designing PCR primers on aligned DNA sequences, *Bioinformatics Application Note*, Vol.20 no.10:1644-1645(2004)