KSBB Journal 2009. 24: 410~414 KSBB

SNPchaser : DNA서열의 SNPs 치환 및 Heterozygosity 확인 프로그램

장진우[1] · 이현철[2] · 이명훈[2] · 최연식[3] · 추동원[1] · 박기정[4] · 이대상[1]*

[1]한국폴리텍 바이오대학 바이오생명정보과, [2](주)디앤피바이오텍,
[3]한국폴리텍 바이오대학 의생명동물과, [4](주)스몰소프트 정보기술연구소

SNPchaser : A Web-based Program for Detecting SNPs Substitution and Heterozygosity Existence

Jin-Woo Jang[1], Hyun-Chul Lee[2], Myung-Hoon Lee[2], YeonShik Choi[3], Dongwon Choo[1], Kiejung Park[4], and Daesang Lee[1]*

[1]Department of Bioinformatics, Korea Bio Polytechnic College, [2]D&P BioTech, Ltd., [3]Department of Laboratory Animals, Korea Bio Polytechnic College, [4]Information Technology Institute, SmallSoft Co. Ltd.

Abstract Single-nucleotide polymorphisms (SNPs) are the DNA sequences difference among the same species in the level of nucleic acids and are widely applied in clinical fields such as personalized medicine. The routine and labor-intensive methods to determine SNPs are performing the sequence homology search by using BLAST and navigating the trace of chromatogram files generated by high-throughput DNA sequencing machine by using Chromas program. In this paper, we developed SNPchaser, a web-based program for detecting SNPs substitution and heterozygosity existence, to improve the labor-intensive method in determining SNPs. SNPchaser performed sequence alignment and visualized the suspected region of SNPs by using user's reference sequence, AB1 files, and positional information of SNPs. It simultaneously provided the results of sequences alignment and chromatogram of relevant area of SNPs to user. In addition, SNPchaser can easily determine existence of heterozygosity in SNPs area. SNPchaser is freely accessible via the web site http://www.bioinformatics.ac.kr/SNPchaser and the source codes are available for academic research purpose.

Keywords: AB1, Alignment, Chromatogram, Heterozygosity, Sequence, SNPs

서 론

SNPs는 같은 종의 생명체 개체별 편차를 나타내는 한 개 또는 수십 개의 염기변이를 일컫는 말이다(1). 인간의 경우, 개체별 SNPs의 차이 때문에 인종의 피부, 머리카락, 체질, 질병, 약물에 대한 감수성 등의 특성이 서로 다르게 나타난다고 알려져 있다(2). 이러한 SNPs의 차이를 조기에 발견함으로써 개인의 다양한 생리작용과 체질의 변화 뿐 아니라 발병 가능성에 대해 제노타이핑 (genotyping) 검사를 통해 발병 가능성을 예측할 수 있으며, 나아가 환자 개개인의 특성에 맞게 약을 진단, 처방할 수 있을 것으로 예견된다(3, 4).

인간의 SNPs의 유무를 결정함에 있어서 고려해야 할 중요한 요소들 가운데 하나는 SNPs서열의 이형접합 (heterozygosity)여부 확인이다. 단일 배체만 존재하는 미생물과는 달리 사람의 경우, 모계와 부계로부터 각각 유전자를 물려받아 2배체 (2 n)의 형태로 염색체가 일반 세포에 존재하는데, 물려받은 유전자의 서열이 똑 같은 경우도 있고 그렇지 않은 경우도 있다. 부모로부터 받은 유전자가 서로 같을 경우를 동형접합 (homozygosity), 차이가 있을 경우 이를 heterozygosity라고 한다. 이러한 heterozygosity가 존재하는 위치에 대해 사람의 염색체를 주형 (template)으로 이용하여 sequencing반응을 수행하였을 경우, 그 결과물인 chromatogram에는 해당 위치의 염기의 chromatogram에는 한 개의 염기

***Corresponding author**
Tel: +82-41-746-7374, Fax: +82-41-746-7370
e-mail: gencia@gmail.com

가 아닌 두 개의 염기가 비슷한 높이를 가진 peak가 존재하는 경우가 발생한다.

이러한 heterozygosity에 해당되는 부분에 대한 1차적인 검정은 ABI 3700 automated DNA sequencer (Perkin-Elmer, Foster City, CA, USA)와 같은 대용량 염기서열 결정기계의 산출물인 AB1 파일을 Chromas (Technelysium Pty Ltd, Helens-vale, Queensland)와 같은 프로그램을 이용하여 해당 지역의 염기에 대한 chromatogram을 일일이 육안으로 확인하는 것이다. 이렇게 heterozygosity를 보일 것으로 추정되는 서열을 입력서열로 사용하여 NCBI BLAST(5)와 같은 상동성 검색을 통해 기존 서열들과의 SNPs의 차이 유무를 최종적으로 검증하는 방법이다(6). 또한 SNP분석에 사용되고 있는 GENETYX라는 프로그램은 FASTA 양식으로 입력한 서열끼리만 비교가 가능하므로 SNP로 추정되는 부분의 chromatogram을 직관적으로 보여주는 기능이 없다는 단점을 지니고 있다(7).

이러한 기존 분석 방법들은 heterozygosity의 여부와 SNPs유무를 정확하게 판단할 수 있는 장점이 있으나, 많은 시간과 노동력이 수반되는 단점을 지니고 있다. 일반적으로 사용되고 있는 Chromas프로그램의 경우 상동성 검색을 지원하는 기능이 없고, BLAST 검색은 FASTA 양식의 서열과 ASCII로 된 DNA 서열만을 입력 서열로 사용할 수 있으므로, chromatogram파일을 읽어 들여 사용자가 관심을 가지고 있는 부분에 대한 특정 SNPs의 heterozygosity를 확인 할 수 없다는 단점을 가지고 있다.

SNPs분석을 위해 주로 사용되는 프로그램인 Chromas와 BLAST가 가지고 있는 이러한 각각의 단점을 보완하고자 SNPchaser를 개발하였다. SNPchaser는 서열정렬과 chromatogram을 보여주는 작업을 동시에 수행하여 사용자가 SNPs 및 heterozygosity유무를 손쉽고 신속하게 확인할 수 있는 장점을 가지도록 구현하였다.

재료 및 방법

개발환경

프로그래밍 언어로 Perl을 사용하여 SNPchaser를 개발하였으며, 분자생물학 사용자에게 익숙한 웹 기반으로 운영되도록 하였다. SNPchaser는 대용량 sequencer로부터 산출되는 바이러니 파일인 AB1파일을 입력파일로 사용하여 A, T, G, C 각각의 chromatogram, sequence, base call, 사용자가 수정한 서열과 base calling 값을 얻어 처리하도록 프로그래밍 하였다.

주요기능

SNPchaser의 전체 데이터 흐름도와 처리 절차는 Fig. 1과 같으며, SNPchaser의 주요 기능은 다음과 같다.

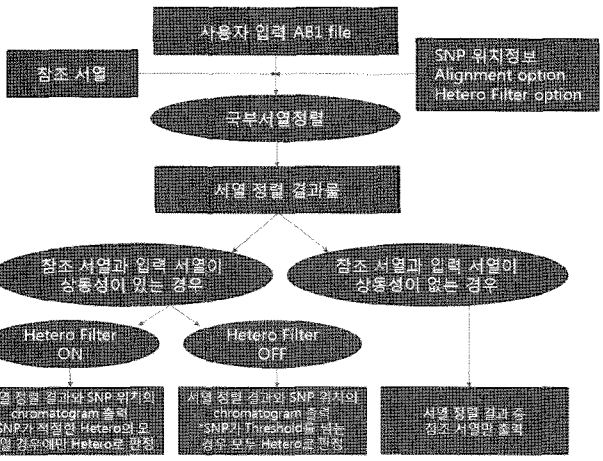

Fig. 1. Overview and schematic data flow process in SNPchaser

첫 째로, 사용자로부터 입력 받은 참조 서열 (reference sequence)과 SNPs의 위치 정보를 이용하여 AB1파일의 DNA sequence와 서열 정렬을 수행하여 출력하는 방식을 택하였다. 서열 정렬은 국부정렬 알고리즘 (local alignment algorithm)으로 널리 사용되고 있는 Smith-Waterman algorithm을 구현하여 사용하였다(8).

사용자가 reference 서열과 비교를 통해 SNPs유무를 확인하고자 입력한 DNA 서열 (target sequence)에서 SNPs를 포함하는 주변 서열을 seed로 사용하여 AB1파일로부터 추출한 DNA sequence를 찾아낸다. 서열정렬을 위한 word-size는 SNPchaser의 입력화면에 존재하는 옵션을 통해 변경할 수 있도록 하였다. SNPchaser의 메인화면과 옵션의 screen shot은 Fig. 2(A)와 같다.

두 번째로, target 서열의 SNPs부분의 A, T, G, C의 chromatogram의 trace를 웹 화면을 통해 사용자에게 제공토록 하였다. 한 개의 SNPs부분만 보여 줄 것인지, 주변의 trace도 같이 보여 줄 것인지는 옵션을 통해 변경할 수 있도록 하였다. 이것의 장점은, heterozygosity의 유무를 확인 할 때, 사용자가 Chromas와 같은 프로그램을 이용하여 AB1파일을 불러들여 특정 SNPs 위치의 chromatogram peak를 확인하기 위해, 서열 전체의 trace를 일일이 육안으로 확인하는 단순 반복 작업의 불편을 덜 수 있다는 것이다. 사용자가 입력한 reference 서열과 target 서열간의 정렬결과와 chromatogram의 trace 화면은 Fig. 2(B)와 같다.

세 번째로, 사용자가 지정한 위치에서 heterozygosity가 발견되었다고 학계에 보고되었을 경우, SNPs의 여부를 판단함에 있어 편리하도록 하였다. 우선, 사용자로부터 heterozygosity를 판단하는 여부인 최대 높이 피크에 대한 두 번째 높이의 피크의 제한 값 (threshold value)을 조절할 수 있게 하였다. Target 서열내의 SNPs에 해당되는 염기 주변의 A, T, G, C trace의 증가와 감소를 기준으로 heterozygosity를 확인 할 수 있도록 하였다. 이 방법을 통해 heterozygosity로 판명이 될 경우, 국제 표준 IUB/IUPAC 핵산 표기 기준에 입각하여 해당 염기를 A 또는 T는 W로,

C 또는 G는 S와 같이 표기하여 사용자에게 제공하였다.

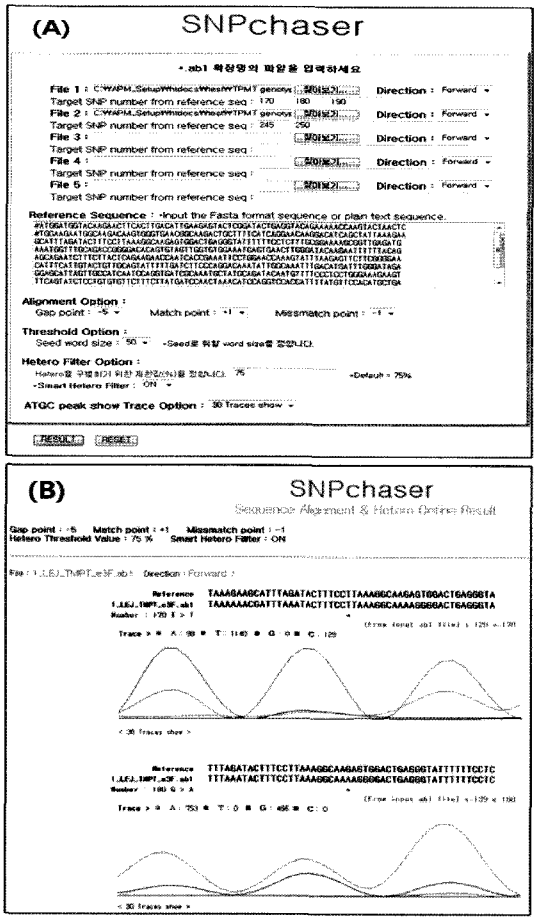

Fig. 2. Screen shots of SNPchaser Users can insert their reference sequence, AB1 files, and suspected position information about the SNPs, and also can modify alignment options through the main page of SNPchaser (A). Results of sequence alignment and visualized chromatogram of relevant region of SNPs (B).

Heterozygosity의 판별과 관련하여 사용자의 편의성을 돕기 위해 SNPchaser에서 별도로 구현한 기능이 heterozygosity filter옵션이다. 이 기능은 target SNPs의 peak가 최고 높이를 기록한 염기를 기준으로 나머지 염기의 trace가 threshold value를 넘더라도, 해당 염기들의 chromatogram상 peak의 모양이 적절치 않으면 heterozygosity로 판단하지 않도록 하였다. AB1 파일은 A, T, G, C 4종류 염기의 trace를 기준으로 10개의 trace당 하나의 염기를 호출 (calling)한다. heterozygosity filter를 사용 한 경우, target SNPs를 중심으로 9개의 trace를 불러와 최고 높이를 가진 염기 대비 두 번째로 높은 peak의 염기가 threshold (default: 75%) 값을 상회하는지의 여부와 9개의 trace값의 증가와 감소를 기준으로 하여 heterozygosity의 여부를 판단하도록 하였다. 반면에 heterozygosity filter를 사용하지 않은 경우에는, 염기의 trace값 증가와 감소는 고려치 않고 오직 최고 높이의 염기를 기준으로 두 번째 높이를 가진 염기가 threshold 값을 상회하는지 여부만 가지고 heterozygosity를 판단토록 하였다. Heterozygosity filter 옵션의 사용 유무에 따른 결과 예시 화면은 Fig. 3과 같다.

마지막으로, SNPchaser는 사용자로부터 입력받은 target 서열에 대해 forward 혹은 reverse에 해당되는 서열로 변환하여 검색이 수행 가능하게 하였다. 최대 5개의 AB1 파일과 각각의 AB1 파일 당 6개의 SNPs위치 정보를 입력 받을 수 있어 최대 30개의 SNPs위치에 대해 SNPs와 heterozygosity 여부를 판별 가능하도록 하였다.

서열정렬과 관련된 옵션은 다양한 서열을 이용하여 시험해 본 결과 국부정렬의 gap point는 -5, match point는 +1, mis-match point는 -1로 사용하였을 때 이상적인 결과를 산출하여 default gap, match, mis-match point를 각각 -5, +1, -1로 할당하였으며, 이들의 point는 SNPchaser의 메인 화면의 옵션을 통해 사용자가 변경 가능하도록 하였다.

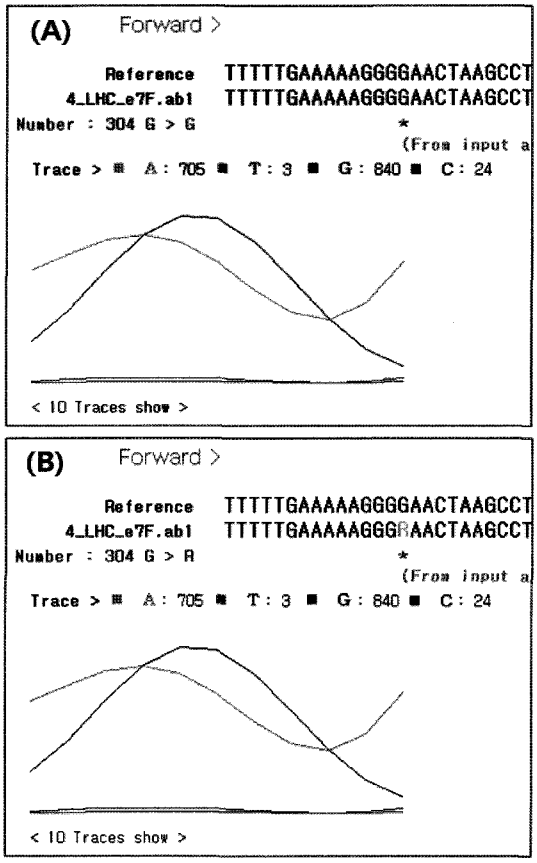

Fig. 3. The effect of heterozygosity filter option in displaying results If the heterozygosity filter option is turn off (B), SNPchaser decides heterozygosity by only considering the peak height difference (default threshold is 75%) between first and second trace. In case of heterozygosity filter option is turn on (A), SNPchaser decides heterozygosity by considering threshold option as well as peak tendency such as increment and decrement of peak in the first and second height trace.

결과 및 고찰

사례연구

SNPchaser를 이용하여 N-acetyltransferase (NAT2)와 thiopurine S-methyltransferase (TPMT)유전자에 대한 SNPs를 각각 분석하였다. 병원에서 항 결핵제로 처방되고 있는 isoniazid는 간에 손상을 주는 독성이 있어 환자에게 문제를 일으키기도 하는데, isoniazid의 간독성은 NAT2의 유전형과 관련이 있는 것으로 밝혀졌으며 NAT2의 acetylation 활성이 떨어지는 유전형이 간독성을 많이 일으키는 것으로 학계에 보고되었다(9). NAT2에 대한 SNPs검사는 NAT2의 유전형을 분석하여 isoniazid를 처방하게 될 환자의 간독성문제를 사전에 대비하는데 일반적으로 이용되고 있다.

TPMT는 Thiopurine 계열 약물 (azathiopurine, 6-mercaptopurine, 6-thioguanine)의 대사에 관여하는 효소로 TPMT유전자형에 따라 효소 활성도가 결정된다. TPMT 돌연변이 allele 환자에게 일상적인 용량의 6-MP/AZA를 투여하면 대사산물의 농도가 높게 축적되어 골수기능 억제 등에 의한 부작용의 위험도가 높으므로 약 용량의 조정이 필요하다. TPMT효소 활성도는 TPMT유전자의 유전자형에 따라 정상 이상, 중등도, 매우 감소 형으로 구분되므로, 이 유전자형을 검사하여 TPMT의 활성도를 추정할 수 있다(10).

3개의 파일로 구성된 AB1 파일을 사용하여 개발시스템 (Pentium IV 3.0 GHz CPU, 512MB RAM, Linux Fedora Core 10)에서 SNPchaser로 분석을 수행하였다. NAT2 참조서열의 길이는 1,030 bp이고, 각각 3개의 파일에 18개의 SNPs의 유무를 검색하는데 소요된 시간은 4초가 걸렸다. 소요시간의 경우, 분석할 AB1파일을 1개, 2개, 3개로 순차적으로 개수를 증가시키자 각각 2.2초, 3.2초, 4.1초가 소요되는 것으로 보아 AB1파일 한 개가 증가함에 따라 소요시간이 약 1초씩 증가된다는 것을 알 수 있었다.

NAT2의 경우, 3개의 파일이 각각 forward, middle, reverse로 구성되어 있으며 이를 연결하면 한 개의 유전자를 구성하지만 TPMT의 경우, exon 중간에 intron이 있어서 5개의 파일로 구성되어 있다. 위의 같은 개발 환경에서 SNPchaser를 실행한 결과, 참조 서열의 길이가 777 bp이고 5개의 AB1파일에 13개의 SNPs의 유무를 검색하는데 소요시간은 3.7초가 걸렸다. 이러한 분석 속도는 여러 환자들로부터 산출되는 수십 개 이상의 유전자에 대해 한꺼번에 SNP 분석을 하는데 있어 무리가 없는 속도라고 판단된다.

특히, TPMT는 첫 번째 파일과 네 번째 파일에 연속적으로 염기 T가 반복되는 영역이 존재하여 연속적인 T영역 후반부에 각 염기가 나타내는 trace의 baseline이 높게 잡혀 있는 특징이 있다. Heterozygosity filter 옵션을 사용하지 않을 경우, heterozygosity가 아닌 경우에도 heterozygosity로 잘못 판정하여 원하지 않는 결과가 나오는 경우가 많았다. 3명의 sample 파일을 분석한 결과, 첫 번째 파일과 네 번째 파일의 G-C contents의 평균은 각각 34.95%, 32.7%로 전체 파일의 평균 G-C contents 값인 33.54%에 비하여 크게 낮지 않은 것을 보아, A나 T의 염기비율보다는 연속적인 A나 T가 baseline에 영향을 주는 것으로 추정된다.

향후계획

현재는 정렬하고자 하는 참조 서열에 관계없이 매번 모든 값을 새롭게 입력하여 SNPchaser를 실행 시킬 수 있다. 임상 실험실과 같은 바이오산업 현장에서 똑 같은 참조서열에 대해 서로 다른 사람으로부터 채취한 DNA sample의 서열과의 비교를 통해, 특정 유전자의 SNPs에 대해 반복적인 검사를 수행하고 그 결과를 살펴보는 경우가 많이 존재 한다. 이러한 상황에서의 SNPs genotyping검사에 대한 사용자의 편의성을 도모하기 위해, 자주 사용되고 있는 참조 서열에 대해 데이터베이스를 만들 계획이다. 또한 이러한 데이터베이스작업을 통해, 사용자 별로 계정을 만들어 참조 서열과 SNPs 검색 옵션들을 자동으로 읽어 들여, 사용자가 AB1파일만 입력하면 결과가 출력되도록 SNPchaser를 추가로 개선할 계획이다. 또한 A, T, G, C의 chromatogram peak에 대한 출력화면에 대해 확대, 축소 기능을 추가하여 사용자의 편의성을 높일 계획이다.

SNPchaser를 단독실행 (stand alone)으로 설치 및 운영에 필요한 소스코드는 학문적인 목적으로만 사용할 경우 저자에게 요청 시 제공 가능하다.

요 약

단염기 다양성 (Single-Nucleotide Polymorphisms, SNPs)은 핵산수준에서의 개개인의 유전 서열간의 차이를 나타내는 말로 최근 맞춤의약 분야에서 각광 받고 있다. 일반적으로 SNPs존재 유무를 확인하는데 주로 사용되는 방법은 ABI automated DNA sequencer와 같은 대용량 염기서열 결정 기계에서 산출되는 결과물 파일로부터 DNA서열을 추출하여 BLAST와 같은 상동성 검색을 수행하는 것이다. 본 논문에서는 사용자로부터 참조서열, AB1파일, SNPs 존재 가능성을 가진 염기의 위치 정보를 입력 값으로 받아 해당 위치에 존재하는 염기의 SNPs 치환 및 heterozygosity 여부를 확인 할 수 있는 프로그램인 SNPchaser를 개발하였다. 특정 유전자 서열 내에서 SNPs를 보이는 염기의 위치에 대한 정보를 사용자가 알고 있는 경우, 전체 유전자 서열에 대해 SNPs유무를 조사할 필요 없이 SNPs를 보인다고 보고된 위치의 염기를 조사하여 SNPs유무를 판단하고, 해당지역의 염기의 chromatogram정보를 사용자에게 제공하는 기능을 가지고 있다. 또한 SNPchaser는 사람과 같은 2배체의 염색체를 가진 생명체에 존재 하는 SNPs지역의 염기에 대한 heterozygosity여부를 사용자가 손쉽게 판별

할 수 있도록 하였다. 본 논문에서 개발한 SNPchaser는 http://www.bioinformatics.ac.kr/SNPchaser에서 사용 가능하다.

감 사

본 연구는 한국폴리텍 바이오대학의 Factory Learning System의 일환인 현장실습, 프로젝트 실습 과제와 지식경제부 지방기술혁신사업 (RTI04-01-01) 지원으로 수행되었습니다.

접수 : 2009년 6월 1일, 게재승인 : 2009년 8월 27일

REFERENCES

1. Kenneth, M. W. and Joseph, D. T. (2000), How many diseases does it take to map a gene with SNPs, *Nature Genetics* **26**, 151-157.
2. Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton (2008), Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, *Nature Genetics* **40**, 1253-1260.
3. 조영훈, 박기정, 이대상 (2008), 핵산증폭용 특정 길이의 Primer 검색 프로그램, *한국미생물학회지* **44**, 164-167.
4. Matsuzaki, H., S. Dong, H. Loi, X. Di, G. Liu, E. Hubbell, J. Law, T. Berntsen, M. Chadha, H. Hui, G. Yang, G. C. Kennedy, T. A. Webster, S. Cawley, P. S. Walsh, K. W. Jones, S. P. Fodor, and R. Mei (2004), Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays, *Nature Methods.* **1**, 109-111.
5. http://blast.ncbi.nlm.nih.gov/.
6. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997), Gapped BLAST and PSI-BLAST : A new generation of protein database search programs, *Nucleic Acids Res.* **25**, 3389-3402.
7. Yamada, K., M. Hirota, Y. Niimi, H. Nguyen, Y. Takahara, Y. Kami, and J. Kanek (2006), Nucleotide sequences and organization of the genes for carotovoricin (Ctv) from *Erwinia carotovora* indicate that ctv evolved from the same ancestor as *Salmonella typhi* prophage. *Bioscience, Biotechnology, and Biochemistry* **70**, 2236-2247.
8. Smith, T. F. and M. S. Waterman (1981), Identification of common molecular subsequence, *J. Mol. Biol.* **48**, 443-453.
9. Cho, H., W. Koh, Y. Ryu, C. Ki, M. Nam, J. Kim, and S. Lee (2007), Genetic polymorphisms of NAT2 and CYP2E1 associated with antituberculosis drug-induced hepatotoxicity in Korean patients with pulmonary tuberculosis, *Tuberculosis* **87**, 551-556.
10. Payne, K., W. Newman, E. Fargher, K. Tricker, I. N. Bruce, and W. E. R. Ollier (2007), TPMT testing in rheumatology : any better than routine monitoring, *Rheumatology* **46**, 727-729.