한국산학기술학회논문지
Vol. 11, No. 2, pp. 720-726, 2010

공개용 리소스를 활용한 Haplotype 재조합 시스템 개발

김기봉[1*]
[1]상명대학교 의생명공학과

Development of Haplotype Reconstruction System Using Public Resources

Ki-Bong Kim[1*]
[1]**Department of Medical Biotechnology, Sangmyung University**

요 약 Haplotype은 연관성을 띠면서 함께 유전하는 SNP (Single Nucleotide Polymorphism) 집단을 반영하고 있기 때문에 맞춤의학 분야에서 haplotype기반의 연구 중요성이 지속적으로 급증하고 있다. *in silico* 방법을 바탕으로 Haplotype 재조합을 위해 현재 가장 널리 사용되는 공개용 리소스 응용소프트웨어로는 PL-EM, Haplotyper, PHASE 및 HAP 등이 있다. PL-EM, Haplotyper 및 PHASE 등은 리눅스와 유닉스 시스템에서 구동되는 명령라인 응용 소프트웨어이고, HAP는 클라이언트-서버 환경에서 웹기반으로 구동되는 소프트웨어이다. 본 논문에서는 실험적으로 검증된 데이터들을 이용하여 공개용 리소스 소프트웨어들의 정확성을 검증하고, 그러한 검증결과를 토대로 선별된 Haplotyper와 PL-EM 등으로 개발한 통합 haplotye 재조합 시스템에 대해 소개하고자 한다. 개발된 통합 시스템은 사용자 친화적 웹 인터페이스를 갖는 클라이언트-서버 시스템으로 최종 사용자들에게 양질의 haplotype 분석 결과를 제공할 수 있다. Haplotyper의 경우 5명의 개체로부터 얻은 길이가 5인 SNP 유전자형 데이터를 가지고 결과를 분석하였고, PL-EM의 경우 15명의 개체로부터 얻은 길이가 13인 SNP 유전자형 데이터를 가지고 결과를 분석하였다. 그 결과 본 시스템은 두 부분으로 나누어 개개인의 haplotype 정보와 haplotype 집단 정보를 이해하기 쉽게 체계적으로 제공하는 것을 확인하였다. 이러한 측면에서 본 시스템은 haplotype 지도 작성을 통한 질병 유전자 발굴 및 맞춤의약 개발 연구에 매우 유용한 도구로 사용될 수 있으리라 여겨진다.

Abstract Haplotype-based research has become increasingly important in the field of personalized medicine since the haplotype reflects a set of SNPs (Single Nucleotide Polymorphisms) that are genetically associated and inherited together. Currently, the most widely used application softwares available for haplotype reconstruction, based on *in silico* method, include PL-EM, Haplotyper, PHASE and HAP. PL-EM, Haplotyper and PHASE are command-line application running on LINUX or Unix system and HAP is a web-based client-server application. This paper deals with an integrated haplotype reconstruction system that have been developed with PL-EM and Haplotyper selected from the accuracy test with experimentally verified data on public application softwares. This integrated system is a kind of client-sever one with user friendly web interface and can provide end-users with a high quality of haplotype analysis. SNPs genotype data with a length of 5 derived from 5 people and SNPs genotype data with a length of 13 derived from 15 people were used to test the analysis results of Haplotyper and PL-EM respectively. As a result, this system has been confirmed to provide the systematic and easy-to-understand analysis results that consist of two main parts, i.e. individual haplotype information and haplotype pool information. In this respect, the integration system will be utilized as a useful tool for the discovery of disease related genes and the development of personalized drugs through facilitating the reconstruction of haplotype maps.

Key Words : SNP, Haplotype Reconstruction, Personalized Medicine, PL-EM, Haplotyper, PHASE, HAP

[*]교신저자: 김기봉 (kbkim@smu.ac.kr)
접수일 09년 10월 31일 수정일 (1차 09년 12월 15일, 2차 10년 01월 22일) 게재확정일 10년 02월 24일

1. 서론

21세기 포스트게놈 시대에 생명과학이 당면한 중요한 과제는 이미 밝혀진 유전체 서열들에 대한 수많은 정보들을 어떻게 활용하느냐에 달려 있다. 이러한 측면에서 많은 실험연구자들이 인류의 건강과 복지에 초점을 두고 다양한 각도에서 유전자 정보를 이용한 연구를 수행하고 있다. 활발한 유전체 연구가 이뤄지면서 각 개인의 유전자 변형의 종류와 빈도에 대한 관심이 높아지면서 대량의 SNP (Single Nucleotide Polymorphism) 발굴을 위하여 1999년 12개의 다국적 제약회사, 영국의 Wellcome Trust, 인간유전체사업을 수행했던 4개 기관이 참여한 The SNP consortium(TSC)이 설립되었고 2001년 연구가 완료되었을 때 총 180만여 개의 SNP를 발굴하는 성과를 거두었다. 각 개인은 피부의 색을 비롯하여 눈의 색, 머리카락의 형태, 약물에 대한 반응 등에 많은 다양성을 지니고 있다. 인간의 유전체 안에는 이런 다양성에 영향을 주는 SNP가 약 1,000만개 정도 있을 것으로 추정된다[1]. 이러한 SNP는 질병의 진단, 예후, 치료와 예방에 이용될 수 있다[2, 3]. 질병의 민감성을 유도하는 유전자의 발현조절 지역이나 유전좌위에 위치하는 SNP는 질병에 대한 민감성에 영향을 줄 수 있다. 따라서 SNP는 질병에 대한 민감성을 예측하고 그 질병에 대한 환경과 유전의 역할을 이해하는데 도움을 준다. 뿐만 아니라 SNP를 통해 태아나 아직 증상이 나타나지 않은 환자에서 특정 질병에 대한 진단이 가능하며, 나아가서는 질병의 예후 및 치료, 예방에까지 이용될 수 있다. 또 SNP는 맞춤의약 개발에 필요한 기초자료로 이용될 수 있다[4]. 개개인이 동일 약물에 대해 상이한 반응성, 효과, 부작용, 독성 등을 나타내는 원인은 개개인의 SNP 차이 때문이다. 개개인의 SNP 정보를 바탕으로 할 경우 약물의 효능과 부작용을 미리 예측하여 처방할 수 있으므로, 개개인의 유전적 배경에 근거한 맞춤의약 개발로 이어질 수 있다. 그러나 SNP는 질병 유전자를 식별해내는 표식자(Marker)로 작용하고, 이를 이용하여 질병 유전자를 찾고 개개인에 맞는 맞춤의약을 개발하는데 도움을 주지만 염색체 상의 엄청난 SNP들을 추적하는 것은 매우 어려운 일이며 그 비용도 엄청 날 수밖에 없다. 이러한 문제점을 극복하기 위해 연관성을 띠면서 동일한 유전경향을 띠는 염색체 상의 인접한 SNP들을 연결한 haplotype에 연구의 초점을 맞추는 것이 현실적이고 실효성이 높다고 할 수 있다. 즉, 현실적으로 개별 SNP 보다는 haplotype으로 얻은 정보가 보다 더 효율적이고 비용과 시간 측면에서도 효과적이기 때문에 haplotype에 대한 연구의 중요성이 매우 부각되었다[5]. 이러한 측면에서 International HapMap Project가 착수되었고, 이 프로젝트의 목표는 서로 다른 민족간 haplotype block의 유전체 지도 (genome map)를 완성하고, 모든 유전체 상의 SNP를 확인하고, 서로의 연관성을 확립하여 정상군과 질병군의 유전적 표식자들을 비교하는 것이었다[6]. 이 연구는 네 개의 서로 다른 민족을 대상으로 269 DNA 표본을 연구하여 1000만개의 SNP를 발표하였다. 그래서 질병 유전자 발굴과 개개인에 맞는 맞춤의약 개발의 중요성이 그 어느 때보다도 부각되고 있다. 앞에서 언급한 바와 같이, SNP는 질병 유전자를 걸러내는 표식자로 작용하지만 약 30억의 염기로 구성된 전체 인간 유전체상에서 개개인의 단일염기를 추적해야하기 때문에 연구가 어렵고 비용이 많이 드는 단점이 있다. 이에 반해 함께 유전되는 경향이 있는 염색체 상의 인접한 SNP군을 포함하는 haplotype은 이런 어려움을 일부 해결해 줄 수 있을 것이다[4]. 즉 약 1000만개의 SNP들을 그룹화 함으로써 유전체상의 변이를 밝혀내는 일이 용이해지는 것이다. 그렇기 때문에 인간의 질병연구에서 haplotype 분석의 중요성이 증가하고 있다[5]. 이배체로 구성된 인간의 염색체에서 SNP들을 찾아내기 위해 유전자형화 하고 이때 생성된 이배체 데이터에서 SNP들을 일렬로 구분하여 정렬하는 과정을 haplotype 재조합이라 한다. 한 유전자형 데이터에서 가능한 haplotype 조합의 경우의 수가 매우 많기 때문에 haplotype 재조합에 *in silico* 기법을 활용하는 것이 시간, 비용 및 정확성 등의 측면에서 효과적이다[7]. *in silico* 기법으로 haplotype 재조합을 생성하는 것은 인간의 haplotype 블록 구조를 정의하고 후보 유전자 연구에 중요한 역할을 한다[8]. Haplotype 재조합을 위해 일반적으로 가장 널리 사용되는 공개용 리소스 응용소프트웨어로는 PL-EM[9], Haplotyper[10], PHASE[11] 및 HAP[12] 등이 있다. PL-EM, Haplotyper 및 PHASE 등은 리눅스와 유닉스 계열의 운영체제에서 명령라인 양식으로 구동되며, 사용하기 위해서 직접 다운로드 받아서 설치 환경을 설정해야 하기 때문에 유닉스나 리눅스 계열의 운영체제에 익숙하지 않은 일반 실험연구자들이 사용하기에는 현실적으로 많은 어려움이 따를 것이다. 반면에 HAP의 경우 웹 인터페이스 기반의 클라이언트-서버 구조를 갖고 있다. 그러나 웹 상에서 직접 사용하기에 편리하지만 결과를 보여주는데 있어서나 내부 알고리즘 측면에서 다소 미흡한 점들이 있다. PL-EM과 Haplotyper는 내부적으로 EM 알고리즘(Expectation-Maximization algorithm)을 사용하고 있는데, EM알고리즘은 해석 용이성과 안정성 때문에 통계학적 알고리즘에서 가장 많이 사용되고 있다[9]. 본 논문에서는 가장 널리 사용되는 기존 공개용 리소스 응용 소프트웨어를 대상으로 실험적으로 검증된 공개용 리소스 데이터로 성능평가를 수행하고,

그러한 결과를 토대로 성능이 상대적으로 우수한 것으로 선정된 PL-EM과 Haplotyper 애플리케이션을 이용하여 구축한 웹 기반의 haplotype 재조합 분석 시스템을 소개하고자 한다. 본 시스템은 haplotype 재조합 뿐만 아니라 haplotype 지도 작성을 통한 질병 유전자 발굴 및 맞춤의약 개발 연구에 매우 유용한 도구로 활용될 수 있을 것으로 판단된다.

2. 재료 및 방법

2.1 공개용 프로그램 성능 평가 및 검증 데이터

분석 시스템 개발에 사용할 공개용 프로그램을 선정하기 위해 본 연구에서 시행한 성능평가 대상 프로그램은 PL-EM, Haplotyper, PHASE, 및 HAP 등이며, 이들은 가장 널리 사용되는 대표적인 공개용 haplotype 재조합 프로그램들이다. 또한 이러한 공개용 리소스 프로그램들의 성능을 객관적으로 검증하기 위해 본 논문에서는 실험적으로 검증된 공개용 리소스 데이터인 Daly 데이터[13]와 Gabriel 데이터[14]를 사용하였다. Daly 데이터는 아버지, 어머니 및 자식으로 구성된 유럽인 129 가족의 가계정보를 포함하여 총 378명의 SNP 정보로 구성되어 있고, Gabriel 데이터는 유럽인 12 가족의 93명 가계도 정보와 SNP 정보로 이뤄진 popA, 서로 무관한 50명의 아프리카계 미국인 SNP 정보인 popB, 서로 무관한 42명의 일본계 및 중국계 SNP 정보인 popC, 그리고 요루바족 63명의 가계도 정보와 SNP 정보를 담고 있는 popD 등으로 구성되어 있다. 이중에서 가계도 정보를 제외하고 자식 부분만의 Daly 데이터와 Gabriel 데이터의 popB만을 실제 성능평가 데이터로 사용하였다. 성능평가 결과는 [표 1]에서 볼 수 있듯이 평균적으로 PL-EM와 Haplotyper가 상대적으로 뛰어난 것으로 판명되었다. 평가대상 프로그램들은 각 sample 단위로 2줄의 haplotype를 예측하여 출력하게 된다. 그런 후에 해당 sample에 대한 haplotype 데이터와 예측출력 값을 비교하게 된다. 원래 데이터와 예측 데이터를 비교할 때, 바로 직접 해당 위치별로 비교하는 경우에 의한 값 (S_F)과 예측 데이터의 2줄을 아래위로 뒤바꾼 데이터와 비교하는 경우의 값 (S_R), 즉 두 가지 경우를 다 반영하여 정확률을 계산한다. 원래 데이터가 두 가지 경우와 동시에 일치할 때 1점을 부여한다. 모든 SNP에 대해 일치성을 판단한 뒤 두 가지 경우의 점수 중 높은 점수를 기준삼아 획득한 점수를 SNP의 수로 나누어 sample 하나에 대한 정확률을 계산한다. 모든 sample에 대해 그렇게 구한 정확률의 평균값을 계산함으로써 프로그램의 정확률을 구한다. 이것을 식으로 표현하면 다음과 같다.

$$AC = \frac{\sum_{i=1}^{N_D} \frac{\max(S_F, S_R)}{N_{snp}}}{N_D} \quad (1)$$

위의 수식에서 N_D는 sample 수를 의미하고, N_{snp}는 SNP의 개수를 의미한다.

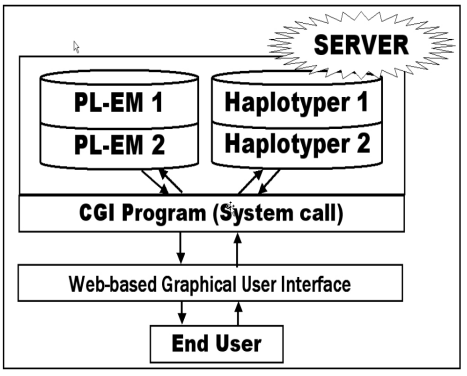
[표 1] 프로그램별 성능 평가결과

BLOCK	SNP 수	PL-EM	Haplotyper	PHASE	HAP
7b	13	0.83	0.81	0.76	0.78
37a	47	0.85	0.86	0.84	0.83
7a	55	0.85	0.86	0.83	0.83
39a	64	0.88	0.85	0.79	0.81
19a	72	0.87	0.88	0.85	0.83
40a	80	0.88	0.86	0.83	0.84
24a	92	0.82	0.82	0.79	0.79
block2	9	0.89	0.88	0.86	0.86
block4	8	0.91	0.91	0.92	0.92
block7	5	0.89	0.89	0.83	0.84
block11	6	0.86	0.82	0.79	0.79
block12	3	0.85	0.84	0.83	0.83
block13	7	0.88	0.86	0.84	0.86
평 균	35.46	0.87	0.86	0.83	0.83

2.2 분석 시스템 설계 및 구성

본 연구에서는 인간의 다양성 및 특정 질병관련 유전자와 관련된 연구를 수행하는 실험연구자들이 보다 편리하고 쉽게 haplotype 재조합 분석 시스템을 이용할 수 있도록 시스템을 설계하고 구현하였다. 앞의 공개용 리소스 프로그램들을 대상으로 행한 성능 평가 결과를 토대로 일반적으로 통계확률 분야에서 널리 사용되는 EM 알고리즘을 채택하고 있는 PL-EM과 Haplotyper를 리눅스 서버쪽 핵심 분석 프로그램으로 설치하고 웹 기반의 클라이언트 인터페이스를 구현하는 형태로 시스템을 고안하였다. 특히 PL-EM과 Haplotyper는 각각 2개의 버전을 갖고 있어 일반 사용자는 총 4가지의 방법으로 haplotype 재조합을 할 수 있도록 구성 하였고, HTML 태그와 PHP 스크립트 언어를 사용하여 재조합 결과를 일목요연하게 보여줄 수 있도록 웹 기반의 사용자 인터페이스를 개발하였다. 즉, Haplotype 재조합이 필요한 생명공학자들이 본 시스템을 통해 유전자형 데이터를 포함하는 파일이나 서열을 직접 입력하고, 다양한 옵션 파라미타를 적절히 적용하면 CGI(Common Gateway Interface) 프로그램에 의해 서버쪽 응용소프트웨어들이 구동되어 개별적인

haplotype 정보와 haplotype 집단 정보 결과를 웹으로 제공해 주도록 설계했다(그림 1 참조).



[그림 1] Haplotype 재조합 시스템의 전체 구성도

2.3 입력서열 데이터 처리

사용자가 웹 브라우저를 통해 유전자형 데이터를 포함하는 파일이나 서열을 직접 입력하고, 각 분석 프로그램이 요구하는 파라미터를 입력한 후 확인 버튼을 누르면 서버 쪽 구동 프로그램을 돌리기 전에 검색 조건의 정확성을 사전에 확인한 후 분석 조건이 성립하지 않을 때에는 에러로 처리하고 올바른 검색 조건과 정보로 판명되면 CGI 프로그램을 통해 해당 서비스를 요청하게 된다. 입력서열이나 입력서열 파일은 SNP 데이터 집합을 넣는데 입력서열은 유전자형(예> 012000010)을 가진다. 여기서 0은 이형접합체(Aa), 1은 야생형 동형접합체(AA), 2는 돌연변이형 동형접합체(aa), 3은 모두 손실(?/?), 4는 야생형과 손실(A/?) 그리고 5는 돌연변이형과 손실(a/?)을 의미한다(A는 야생형 대립유전자를, a는 돌연변이형 대립유전자를, ?는 손실을 의미함). 시스템 구축 후 시험 구동 시 Haplotyper의 경우 5명의 개체로부터 얻은 길이가 5인 유전자형 데이터를 입력서열로 사용했고, PL-EM의 경우 15명의 개체로부터 얻은 길이가 13인 유전자형 데이터를 입력서열로 사용하였다.

2.3.1 Haplotyper의 분석 파라미터

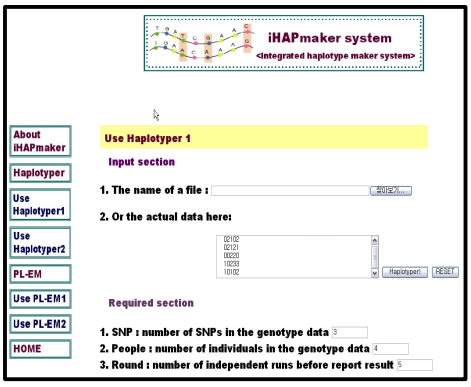
Haplotyper 1과 2에서 요구되는 파라미터에는 SNP, People, Round 등이 있는데, SNP 값은 입력파일의 유전자형 데이터에 있는 SNP의 수를 의미하고, People 값은 입력파일의 유전자형 데이터에 있는 개체수를 의미한다. 또 Round 값은 최적의 haplotype을 얻기 위한 소프트웨어내의 EM 알고리즘 반복 실행횟수를 의미한다. Haplotyper에 구현된 EM 알고리즘의 특성상 Haplotyper 1의 경우 SNP의 수가 1에서 256사이, People의 수가 1에서 100사이 일 때 선택하는 것이 좋고, Haplotyper 2의 경우 SNP의 수가 1에서 100사이, People의 수가 1에서 500사이 일 때 선택하는 것이 유리하다[10].

2.3.2 PL-EM의 분석 파라미터

PL-EM 1은 Top, Parsize, Buffer, Round 등의 파라미터가 존재하며, PL-EM 2는 Cut, Top, Buffer, Round 등의 파라미터가 있다. EM 알고리즘만으로 이루어졌다면 유전자형이 012000010000000000100010일 경우 218개의 가능한 모든 haplotype들을 고려해야하는데 PL-EM의 경우 여기에 분할 연결 전략(Partition-Ligation Strategy)이 첨가 되어 연관된 유전좌위로 분할을 시킨다[9]. 앞의 예의 경우 (012000),(010000),(000001),(00010)의 4개의 분할불가 단위체로 분할하여 각각 EM알고리즘을 실행시킨다. Cut은 사용자가 특별히 분할시키고 싶은 부분을 숫자로 넣어주거나 파일로 넣어주는 것이고 Parsize는 1에서 3 사이 숫자를 선택하는 것인데 1은 각 단편에서 3-4 유전좌위로, 2는 5-8 유전좌위로, 3은 9-16 유전좌위로 분할한다는 것이다. Top은 개개의 유전자형과 함께 양립할 수 있는 haplotype 쌍의 수를 의미하는데, 일반적인 분석을 위해서는 PL-EM 프로그램의 사용자 매뉴얼에서는 0을 설정하도록 권장하고 있다. Buffer는 분할시킨 데이터에서 각각 haplotype을 얻기 위해 알고리즘을 반복 실행시키는 횟수를 말하고, Round는 결론적으로 보여줄 haplotype을 얻기 위한 소프트웨어내의 EM알고리즘 반복 실행횟수를 의미한다.

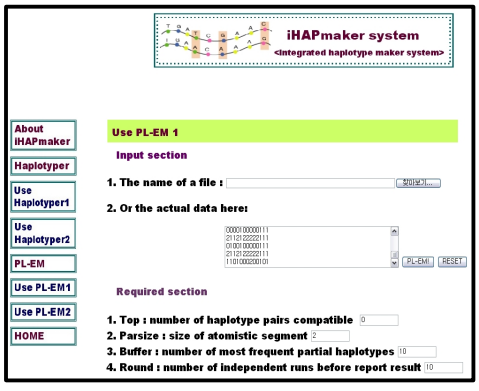
2.4 시스템 구동 방법

입력서열과 각 소프트웨어에 요구되는 파라미터가 제대로 입력되면 CGI 프로그램은 서버의 응용소프트웨어(PL-EM 1&2, Haplotyper 1&2)를 호출하여 시스템을 구동시키고 haplotype 재조합 결과를 다시 받아서 웹으로 디스플레이 시켜준다. 시스템 구동을 위한 각종 유틸리티, 파싱 엔진, CGI 프로그램 등은 PHP 스크립트와 HTML 태그를 사용하여 구현하였다. 각 분석 소프트웨어(PL-EM 1&2, Haplotyper 1&2)의 결과를 PHP 스크립트 언어와 HTML 태그를 사용해 데이터를 파싱 하고, 사용자가 알기 쉽게 재구성하여 웹상으로 보여준다. Haplotyper를 사용하고자 할 때는 Haplotyper 1이나 Haplotyper 2를 선택하여 그림 2과 같은 형식으로 입력 데이터를 편집 박스에 입력하거나 파일 형태로 업로드 시키고, PL-EM을 사용하고자 할 때는 PL-EM 1이나 PL-EM 2를 선택하여 그림 3과 같은 형식으로 데이터를 입력하면 된다.



[그림 2] Haplotyper 1의 웹 인터페이스

Haplotyper와 PL-EM은 입력서열 부분에서 유전자형 데이터를 포함한 파일을 업로드하거나 서열을 직접 입력하고 각각의 소프트웨어의 필수 입력 파라미타 부분에서 요구되는 파라미터가 다르므로 각각 적절히 설정해 줘야 한다. 입력서열 부분과 필수입력 파라미타 부분을 모두 제대로 설정해야 시스템을 구동시킬 수 있다. Haplotyper는 입력서열의 SNP 수와 People의 수에 따라 결과에 영향을 끼칠 수 있고[10], PL-EM 은 EM알고리즘에 분할 연결 전략이 첨가되어 너무 긴 입력서열의 경우 내부적으로 부분적인 haplotype들을 생성하므로 효율적인 결과를 얻을 수 있고 보다 다양한 파라미터들을 갖고 있는 장점이 있으므로[9] 사용자의 입력서열에 따라 적절한 소프트웨어를 선택하는 것이 유리하다.

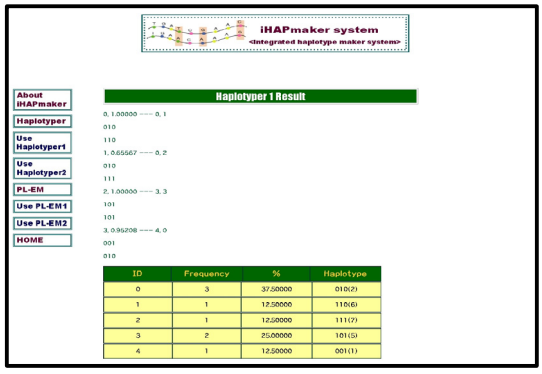


[그림 3] PL-EM 1의 웹 인터페이스

3. 시스템의 사용결과

3.1 Haplotyper 사용결과

임의의 5명의 개체로부터 얻은 길이가 5인 유전자형 데이터를 입력서열로 입력한 결과 그림 4에서 보는 바와 같이 Haplotyper 1과 2는 두 가지의 핵심 정보, 즉 개별적인 haplotype 정보와 haplotype 집단 정보를 제공한다. 첫 번째 부분은 Haplotyper에 의해서 개별적인 haplotype들이 예측이 되고, 주어진 개체군의 haplotype 집단에서 각각의 haplotype은 ID 숫자가 매겨져 표시된다. 뒷부분의 확률 값은 선택된 haplotype 쌍과 관련이 있는데 이것은 해당 haplotype 쌍의 예측이 얼마나 정확하다고 여겨지는지를 측정하는 것으로 사용되어질 수 있다[10]. 두 번째 부분은 해당 개체군의 haplotype 집단 정보를 테이블 형태로 보여주는데 각각의 haplotype이 개체군에서 나타날 빈도수, 퍼센트와 실제 haplotype이 무엇인지 보여준다. Haplotype은 이진수로 변환되어 보여 지는데 0은 야생형을 의미하고 1은 돌연변이형을 의미한다. 그러므로 각 값은 표식자 유전좌위의 고정된 숫자로써 유일한 것이다.



[그림 4] Haplotyper의 분석 결과 화면

3.2 PL-EM 사용결과

PL-EM에 대해서는 임의로 15명의 개체로부터 얻은 길이가 13인 유전자형 데이터를 입력서열로 입력하여 시험 구동해 보았다. PL-EM 1과 2의 결과도 두 부분으로 나누어 보여주게끔 출력 인터페이스를 고안하고 구현하였다. 첫 번째 부분은 PL-EM에 예측한 개별 haplotype들이 출력되고, 주어진 개체군의 haplotype 집단에서 각각의 haplotype은 ID 숫자가 매겨져 표시된다. 그 뒤에 따르는 확률 값은 선택된 haplotype 쌍과 관련이 있으며, 이것은 해당 haplotype 쌍의 예측이 얼마나 정확하다고 여겨질 수 있는지 가늠할 수 있는 측정치이다[9]. 또 PL-EM의 경우 가능한 haplotype 쌍 모두를 보여주고 이들 각각의 haplotype 쌍이 나타날 확률을 보여준다. 두 번째 부분은 전체적인 요약 부분으로 haplotype 집단 정보를 테이블 형식으로 보여주며, 각각의 haplotype이 개체군에서

나타날 빈도수, 표준 편차, 표준 오류율 등과 실제 haplotype이 무엇인지 보여준다. Haplotyper에서와 마찬가지로 haplotype은 이진수로 변환되어 나타나는데 0은 야생형을 의미하고 1은 돌연변이형을 의미한다. 그러므로 이 값은 표식자 유전좌위의 고정된 숫자로써 유일한 것이다.

4. 결론 및 고찰

본 논문에서는 실험적으로 검증된 공개용 리소스 데이터를 사용하여 haplotype 재조합에 활용될 수 있는 공개용 프로그램들의 성능을 평가하고, 그러한 평가결과를 토대로 상대적으로 정확성이 보다 높은 것으로 판명된 PL-EM와 Haplotyper를 활용하여 구축한 웹 기반의 haplotype 재조합 시스템에 대해 소개하였다. 성능평가에서는 표본수와 SNP수에 따라 다소 유동적이지만 대상 프로그램들은 대체로 79~88% 정도의 범위 안에서 정확률을 유지하고 있음을 알 수 있었다 (평균값으로 환산했을 경우에는 83~87%). 그러나 아마도 손실 데이터(missing data)를 제대로 고려하게 된다면 정확률에서 다소의 변화가 있을 것으로 여겨진다. 본 논문에서는 어디까지나 시스템 구성을 위한 대상 프로그램 선정 차원에서 성능평가가 이뤄졌기 때문에 굳이 손실 데이터를 고려하지 않더라도 동등한 입장에서 이뤄진 평가이므로 상대적인 우위를 파악하는데 전혀 문제가 없을 것으로 간주된다.

본 연구에서는 선행 성능평가를 통해 선정된 PL-EM과 Haplotyper를 통합하여 총 4 가지 방법으로 haplotype 재조합 분석을 행 할 수 있도록 시스템이 구축되었기에 일반 사용자들은 분석 대상 데이터에 보다 적합한 방법을 선택하여 맞춤 분석을 할 수 있을 뿐만 아니라 비교분석도 가능할 것이다. 특히 PL-EM와 Haplotyper는 리눅스나 유닉스 계열 운영체제 하에서 구동되는 프로그램들로 일반 실험연구자들이 쉽게 사용하기 어려운 면이 있기 때문에 이러한 측면을 극복하고 보완하기 위해 PHP 스크립트 언어와 HTML 태그를 사용하여 사용자 친화적인 웹 인터페이스 기반의 시스템을 구축하였다. 또한 소프트웨어의 내부 알고리즘도 생명정보학 분야에서 널리 사용됨과 동시에 신뢰성이 높은 것으로 간주되고 있는 EM 알고리즘 기반의 프로그램들을 시스템에 채택하여 일반 사용자들로 하여금 양질의 haplotype 재조합 정보를 얻을 수 있도록 하였다. 임의의 유전자형 데이터를 이용하여 시험 구동한 결과 개별적인 haplotype 정보와 haplotype 집단 정보로 구성된 결과를 일목요연하게 시스템이 제시해 주는 것을 확인할 수 있었다. 따라서 본 시스템을 통해 쉽게 웹상으로 양질의 haplotype 재조합을 얻을 수 있고, 또한 haplotype 지도 작성을 통한 질병 유전자 발굴 및 맞춤의약 개발 연구에 매우 유용한 도구로 사용될 수 있을 것으로 여겨진다. 향후 연구를 통해 본 시스템에서 구축된 내부 알고리즘들을 실질적으로 통합화할 수 있는 하이브리드 알고리즘 (hybrid algorithm)을 개발하고, 다른 분석프로그램들과 비교분석할 수 있는 표준 검증데이터를 확보하여 시스템의 성능을 향상시키고 객관적으로 입증하고자 한다.

참고문헌

[1] Venter *et al*. “The Sequence of the Human Genome”, *Science*, 291:1304-1351, 2001.

[2] Kwok, P. Y. and Z. Gu. “Single nucleotide polymorphism libraries: why and how are we building them?”, *Molecular Medicine Today,* 5:538-543, 1999.

[3] Kato, M., Y. Nakamura and T. Tsunodd. "An algorithm for inferring complex haplotypes in a region of copy-number variation", *The American Journal of Human Genetics*, 83:157-169, 2008.

[4] Niu, T., Z. S. Qin, X. Xu, and J. S. Liu. “Bayesian haplotype inference for multiple linked Single Nucleotide Polymorphisms”, *Am. J. Hum. Genet*., 70:157-169, 2002.

[5] Xing Wang. "HIT: a Haplotype inference Testbed" CAPSL technical report, University of Delaware, 2003.

[6] International HapMap Consortium. "The International HapMap Project", *Nature*, 18:426(6968):789-96, 2003.

[7] Tishkoff, S. A., A. J. Pakstis, G. Ruano, and K. K. Kidd, "The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus", *Am. J. Hum. Genet*., 67:518-522, 2000.

[8] Tabor, H. K., N. J. Risch, and R. M. Myers. "Candidate-gene approaches for studying complex genetic traits : practical considerations", *Nat. Rev. Genet,.* 3:391-397, 2002.

[9] Qin, Z. S., T. Niu, and J. S. Liu. "Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single Nucleotide Polymorphism.", *Am. J. Hum. Genet.,* 71:1242-1247, 2002.

[10] Niu, T., Z. S. Qin, X. Xu, and J. S. Liu. “*In silico* Haplotype determination of a vast set of Single Nucleotide Polymorphisms”, Technical report, Department of Statistics, Harvard University, 2001.

[11] Stephens, M., N. Smith, and P. Donnelly. "A new statistical method for haplotype reconstruction from population data", *American Journal of Human Genetics*, 68, 978-989, 2001.

[12] Halperin, E. and E. Eskin. "Haplotype Reconstruction from Genotype Data using Imperfect Phylogeny", *Bioinformatics*. 20(12):1842-9, 2004.

[13] Mark, J., D. John, D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander. "High-resolution haplotype structure in the human genome", *Nature Genetics*, 29(2):151-158, 2001.

[14] Gabriel *et al*. “The Structure of Haplotype Blocks in the Human Genome”, *Science*, 296(5576):2225-9. 2002.

김 기 봉(Ki-Bong Kim) **[정회원]**

- 1992년 2월 : 경북대학교 미생물학과 (이학사)
- 1997년 2월 : 경북대학교 미생물학과 (이학석사)
- 2003년 3월 : 충남대학교 컴퓨터공학과 (공학박사)
- 1999년 4월 ~ 2003년 8월 : (주) 스몰소프트 연구소장/기술이사
- 2003년 9월 ~ 현재 : 상명대학교 의생명공학과 부교수

<관심분야>
바이오데이터 마이닝, 유전체 정보학, 기계학습, 단백질 상호작용 및 조절 네트워크