XML 기반의 유전자 예측결과 분석도구

변상희[0*] 윤형석[**] 안건태[*] 박양수[*] 이명준[*]
[*]울산대학교 [**]국립 보건원 유전체 연구부 역학정보실
[*] [**]{heeya[0], miracle, java2u, yspark, mjlee}@ulsan.ac.kr

An XML-Based Analysis Tool for Gene Prediction Results

Sang-Hee Byun[0*] Hyeong-Seok Yoon[**] Geon-Tae Ahn[*] Yang-Su Park[*] Myung-Joon Lee[*]
[*]School of Computer Engineering & Information Technology, University of Ulsan
[**]Division of Epidemiology and Bioinfomatics,
National Genome Research Institute, Korea National Institute of Health

요 약

염기서열의 분석이 유전체에 대한 연구를 가능하게 해 줄 수 있다는 것이 밝혀짐에 따라 다양한 생명체에 대한 유전체 염기서열 분석 도구의 개발이 활발히 진행되었다. 이러한 유전자 예측 도구들은 고유의 단순 텍스트 형식으로 결과를 제공하므로 사용자는 결과를 분석하고 통계정보를 산출하는데 많은 노력이 필요하다.

본 논문에서는 유전자 예측결과를 보다 효율적으로 표현하고 분석하기위한 XML 기반의 분석도구를 개발하였다. 개발된 시스템은 유전자 예측결과를 효과적으로 표현하는 GenStructML, 이 정보를 분석한 GenPredML과 PredAccuracyML로 구성되어 있다. GenPredML과 PredAccuracyML은 GenStructML에 대하여 뉴클레오티드 수준(nucleotide level), 엑손 수준(exon level) 그리고 신호 수준(signal level)에서의 예측 정확도(Accuracy)를 계산하고 Genbank의 정보와 비교하여 통계정보를 산출함으로써 보다 자세한 정보를 제공한다.

1. 서 론

생명체가 지닌 유전 정보의 특성을 분석하고, 규명하려는 유전체 프로젝트가 시작된 이후, 염기서열 분석이 전체 유전자에 대한 연구를 가능하게 해 줄 수 있다는 것이 가시화 되면서 다양한 생명체의 유전체 염기서열 분석 프로젝트가 진행되었다. 그 결과 유전체 내의 정확한 유전자의 위치를 알아내기 위해 GENSCAN[1], Genie[2], GeneMark[3], Unveil[4] 등의 다양한 유전자 예측 도구들이 개발되었다.

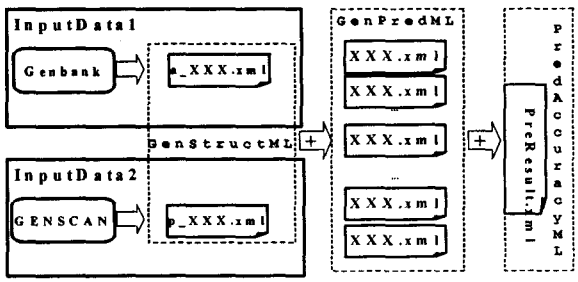
이러한 유전자 예측 도구들은 그들 고유의 방법으로 결과를 산출하고 있으며, 대부분이 단순 텍스트 형식으로 산출 결과를 제공하고 있다. 사용자는 각각의 도구를 사용할 때마다 다른 형식으로 제공되는 결과물 분석이 쉽지 않으며 단순 텍스트 파일로 표현된 데이터를 바탕으로 통계정보를 산출하고자 할 때 많은 노력이 필요하다.

본 논문에서는 다양한 형식으로 표현된 유전자 예측 도구의 결과를 정형화하여 표현하고 annotation 정보도 같이 제공한다. 이 정보들을 바탕으로 사용자가 보다 쉽게 이해할 수 있고 활용할 수 있도록 XML(eXtensible Markup Language)[5] 기반의 GenStructML을 개발하였다. 또한 prediction 정보와 annotation 정보를 바탕으로 통계자료를 보다 효과적으로 계산하기 위하여 GenPredML과 PredAccuracyML을 개발하였다.

본 논문의 구성은 다음과 같다. 2장에서는 개발된 분석도구의 전체 구성에 대해서 살펴본다. 그리고 3장에서는 GenStructML의 구성 요소들에 대하여 살펴보고 4장에서는 하나의 유전자에 대한 통계정보를 표현한 GenPredML의 구성요소와 유전자 데이터 세트[6] 전체의 통계정보를 나타내는 PredAccuracyML에 대해서 소개한다. 끝으로 5장에서는 결론 및 향후 연구방향에 대해서 기술한다.

2. XML 기반 유전자 예측 분석도구

본 장에서는 유전자 예측도구의 결과를 정형화하여 표현하고 통계결과를 효과적으로 분석하기 위하여 GenStructML, GenPredML 그리고 PredAccuracyML로 구성된 분석도구를 소개한다. GenStructML, GenPredML 그리고 PredAccuracyML은 구조적인 웹 문서 표준인 XML 기술을 이용하여 개발된 유전자 예측 분석도구의 결과를 기술하기 위한 마크업 언어이다.



(그림 1) 개발된 분석도구 구성

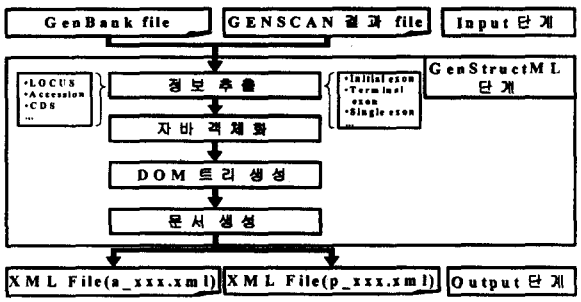
개발된 유전자 예측 결과 분석도구의 전체적인 구성은 (그림 1)과 같다. GenStructML은 Genbank 파일을 입력 데이터로 하여 구성된 <annotated> 요소(Element)와 GENSCAN의 결과파일로부터 계산된 <predicted> 요소로 구성된다. GenPredML은 하나의 유전자에 대한

annotation과 prediction 정보를 표현한 GenStructML을 바탕으로 GenStructML 보다 상세한 prediction 정보 외 통계정보를 표현하도록 설계되었다. 끝으로 유전자 데이터 세트에 대한 통계정보를 산출하기 위해 각각의 유전자에 대한 GenPredML을 입력으로 하여 PredAccuracyML 문서를 생성하였다.

3. 유전자 예측결과를 표현하기 위한 GenStructML

대부분의 유전자 예측 도구들은 각각 고유의 형식으로 기술된 단순 텍스트 파일로 결과를 제공한다. 이러한 비구조적인 문서를 구조적인 문서인 XML 문서로 변환하기 위해 GenStructML 변환기를 구현하였다.

GenStrcutML DTD(Document Type Definition)는 실제로 밝혀진 유전자 정보를 나타내는 <annotated> 요소와 예측된 유전자의 결과를 표현한 prediction 요소로 구성된다. GENSCAN의 결과 파일을 분석하여 유전자로 예측된 정보들을 추출한 후 자바 객체화한다. 이와 같이 필터링된 데이터들은 DOM (Document Object Model)을 이용하여 GenStrcutML 문서의 <predicted> 요소로 변환된다. 그리고 Genbank에서 제공하는 파일을 이용하여 실제 유전자로 발현된 부분의 정보를 추출하고 <predicted> 요소를 생성한 방법과 동일한 방법으로 GenStructML의 <annotated> 요소를 생성한다. 두 요소들은 각각 다른 파일로 생성된다. 변환도구의 구체적인 동작은 (그림 2)에서 보는 바와 같다.



(그림 2) GenStructML 변환기

3.1 GenStructML의 annotation 요소

GenStructML은 실제 밝혀진 유전자 정보를 나타내기 위해 Genbank 파일을 입력으로 <annotated> 요소를 포함한 XML 문서로 생성된다. Genbank 파일은 유전자 정보를 저장하는 주요한 저장소로 미국의 NCBI에서 제공하고 있다. 접근 번호와 유전자 이름, 계통 발생학적 분류를 비롯하여 조절 영역(regulatory region)의 위치나 단백질 번역(translation), 엑손(exon)과 인트론(intron) 같은 서열에 대한 정보를 구체적으로 집대성해 놓은 Features 외에도 다량의 정보를 보유하고 있다.

<annotated>의 주 구성요소는 하나의 CDS(coding sequence)와 관련된 정보를 나타내는 <translation_group>이다. <translation_group>의 CAAT_signal, TATA_signal 등의 정보로 이루어진 <Promoter>, <5'UTR>, <CDS>그리고 <3'UTR> 요소는 Genbank파일의 Features 정보로 분류된다. Features에 CDS의 수는 translation_group의 수를 의미한다. <annotated> 요소의 translation_group, LOCUS, 접근번호 등의 요소는 Genbank 파일에서 biojava[7]을 이용하여 파싱하며 DOM을 사용하여 GenStrcutML의 <annotated> 요소로 변환된다.

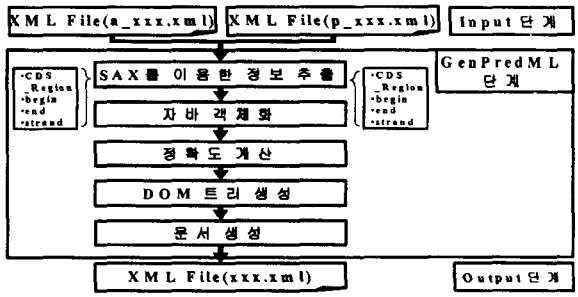
3.2 GenStructML의 prediction 요소

GenStructML의 <predicted> 요소는 GENSCAN의 결과 파일을 입력으로 XML 문서를 생성한다. GENSCAN은 통계에 관한 정보와 유전자 구조에 관한 확률 모델을 결합하여 유전자의 위치를 예측하는 프로그램이다. GENSCAN]의 결과에서 예측된 유전자의 종류(poly-A 부분을 제외한 시작 엑손(initial exon), 내부 엑손(internal exon), 단일 엑손(single exon), 종료 엑손(terminal exon)), 엑손의 시작, 끝 위치 등의 정보를 추출한다. 이 데이터는 GFF(Gene-Finding Format, General Feature Format)[8]을 기반으로 하는 GenStructML의 <predicted> 요소로 변환된다. Sanger Centre에서 제안된 GFF는 모든 유전자 예측 프로그램의 결과 파일에 기본이 되는 형식이다.

4. 유전자 예측결과를 분석하기 위한 GenPredML과 PredAccuracyML

GenPredML과 PredAccuracyML의 기본적인 XML 문서구조를 표현하기 위해 GenPredML DTD를 정의하고 있다. GenPredML DTD는 <Accuracy>와 <Extra>요소로 구성되어 있다. GenPredML은 하나의 유전자에 대한 통계정보와 유전자 위치의 보다 자세한 정보를 기술하기위한 마크업 언어이다. 그리고 PredAccuracyML은 GenPredML을 바탕으로 유전자 데이터 세트에 대한 통계정보를 표현한다.

GenPredML은 XML문서의 구조와 내용을 접근하기 위한 표준화된 API인 SAX(Simple API for XML)를 이용하여 GenStructML의 <annotated>와 <predicted> 요소 정보를 추출한다. 이 데이터를 자바 객체화하여 예측 정확도와 기타 정보를 계산하여 유전자 예측 통계정보를 표현하는 GenPredML 문서로 변환한다. 아래 (그림 3)은 GenPredML 변환기 시스템의 전체 구성 및 자료의 흐름을 보여준다.



(그림 3) GenPredML 변환기

PredAccuracyML은 GenPredML의 정보를 입력으로 하여 (그림 3)과 유사한 과정을 통하여 유전자 데이터 세트의 통계정보를 산출한다.

4.1 GenPredML의 Accuracy와 Extra 요소

GenPredML은 하나의 유전자에 대한 다양한 예측 정확도와 본 분석도구에서 제공하는 기타 정보로 표현된다. GenPredML의 root 요소는 <gene>이며 유전자의 이름을 gname으로 표현한다.

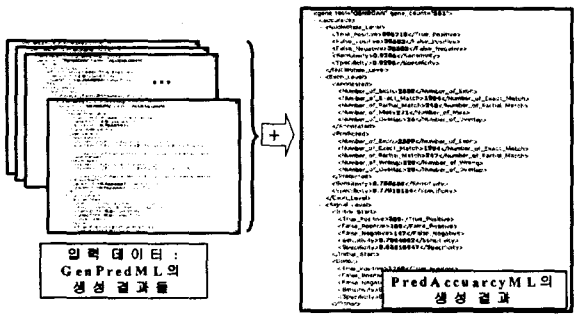
예측 정확도는 GenStructML의 <annotated>과 <predicted> 요소에서 CDS_Region의 시작, 끝 위치의 값을 비교하여 뉴클레오티드 수준(nucleotide level), 엑손 수준(exon level) 그리고 신호 수준(signal level) 에서의 민감도(Sensitivity)와 특이도(Specificity)를 계산한다.

개발된 분석도구에서는 GenPredML에서만 <Extra> 요소는 하나의 유전자 정보에 대한 엑손의 시작과 끝, 엑손 수준의 결과정보(Exact Matching, Partial Matching, Miss, Wrong, Overlap) 등을 제공한다.

4.2 데이터 세트의 통계를 위한 PredAccuracyML

GenPredML이 하나의 유전자에 대한 통계결과를 나타낸다면 PredAccuracyML은 유전자 데이터 세트에 대한 통계결과를 표현한다. PredAccuracyML은 root 요소는 <gene>이며 gene_counter 속성으로 데이터 세트에 포함된 유전자의 수를 나타내고 <Accuracy> 요소만 기술한다.

PredAccuracyML은 Burset과 Guigo가 실험한 570개의 척추동물 유전자 서열(vertebrate gene sequences)[6]을 바탕으로 통계정보를 산출하였다. GENSCAN 프로그램을 이용한 prediction 결과와 Genbank 파일 정보로 GenStructML을 표현하고 GenPredML 결과를 산출하여 전체 데이터 세트의 통계결과를 계산하였다. (그림 4)는 개발된 분석도구의 최종 결과물인 PredAccuracy.xml의 일부와 PredAccuracyML의 입력 데이터인 GenPredML의 결과물들을 나타낸 것이다. GenPredML의 각 수준의 예측 정확도의 데이터들을 바탕으로 하여 유전자 데이터 세트에 대한 통계결과를 산출하였다.



(그림 4) PredAccuarcyML의 입력 데이터 및 생성 결과

(그림 4)의 결과를 보면 총 570개의 실험 유전자 중에서 Genbank 데이터를 사용하여 551개의 annotation 정보를 표현한 GenStructML을 생성하였다. 그리고 GENSCAN 결과를 이용하여 568개의 GenStructML의 prediction 정보를 얻을 수 있었다. 이 두 정보를 바탕으로 551개의 GenPredML 파일이 생성되었으며 이를 바탕으로 PredAccuracyML 통계결과를 산출하였다. [표 1]은 PredAccuracyML의 각 수준의 예측 정확도에 대한 통계결과를 보여준다.

[표 1] PredAccuracy.xml 파일의 예측 정확도

Accuracy / 수준(Level)		민감도 (Sensitivity)	특이도 (Specificity)
뉴클레오티드 (nucleotide)		93 %	93 %
엑손 (exon)		79 %	77 %
신호 (signal)	Initial_Start	71 %	66 %
	Donor	87 %	84 %
	Acceptor	83 %	84 %
	Terminal_End	76 %	83 %

5. 결론 및 향후 연구과제

현재 개발되어 있는 대부분의 유전자 예측 도구들은 고유의 단순 텍스트 파일 형식으로 결과를 제공하고 있어 결과를 분석하고 통계정보를 산출하기가 쉽지 않다.

본 논문은 유전자 예측 결과를 보다 효과적으로 표현하고 분석하기 위해 XML 기반 분석도구를 개발하였다. 여러 도구들의 결과를 정형화하여 표현하고 Genbank에서 제공하는 정보도 함께 제공함으로써 기존의 예측 도구만을 사용했을 때보다 사용자가 이해하기 쉽게 정보를 제공할 수 있다. 또한 구조화된 XML기술을 이용하여 데이터를 표현하고 있어 보다 용이하게 통계정보를 계산할 수 있다. 하지만 GenPredML은 GenStructML의 annotation과 prediction 정보가 모두 존재 할 때만 통계정보를 계산할 수 있기 때문에 두 정보 중 하나라도 존재하지 않는 경우 GenPredML파일이 만들 수 없어서 그 유전자에 대한 통계정보는 산출할 수 없다.

향후 연구과제로 GENSCAN의 결과 뿐 만 아니라 다양한 유전자 예측 도구들의 결과도 분석하여 본 시스템을 보다 일반화시킬 것이며 GenPredML을 바탕으로 유전자 구조 예측 뷰어도 개발할 예정이다.

6. 참고 문헌

[1] Burge C, Karlin S "Prediction of complete gene strcutures in human genomic DNA" J Mol Biol 1997, 266:78-95

[2] Kulp D, Haussler D, Reese MG, Eeckman FH "A generalized Hidden Markov Model for the recognition of human genes in DNA" In Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press; 1996

[3] Borodovsky M, McIninch J "GeneMark: parallel gene recognition for both DNA strands" Comput Chem 1993, 17:123-133.

[4] William H. Majoros, Mihaela Pertea, Corina Antonescu, and Steven L. Salzberg "GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders" Nucl. Acids. Res. 2003 31: 3601-3604.

[5] Yergeau F, Bray T, Paoli J "Sperberg-McQueen CM, Maler E: Extensible Markup Language (XML) 1.0 (Third Edition)" W3C Recommendation, February 2004

[6] Burset M, Guigo R "Evaluation of gene structure prediction programs" Genomics 1996, 35:353-367

[7] BioJava, http://www.biojava.org/index.html

[8] http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml