XML기반의 유전자 예측결과 분석도구

김 진 홍[†] · 변 상 희[†] · 이 명 준[††] · 박 양 수[††]

요 약

생명체의 주된 기능 요소인 유전자를 모두 식별하는 작업의 중요성이 증가함에 따라, 최근에 유전자 예측도구들이 활발히 개발되고 있다. 그러나 유전자 예측 프로그램들은 예측 결과를 그들 고유의 형식으로 제공하여 사용자가 그 결과를 이해하기 위해서는 상당히 많은 추가적인 노력이 필요하다. 따라서 유전자 예측결과에 대한 표준화된 표현과 유전자 데이터 집합에 대한 예측결과를 자동으로 계산하는 방법을 지원하는 것이 바람직하다.

본 논문에서는 다양한 유전자 예측 정보에 대한 효과적인 XML 표현과 이를 바탕으로 예측된 유전자 결과를 자동으로 분석하는 XML 기반 분석 도구에 대하여 기술한다. 개발된 도구는 유전자 예측도구를 사용하는 사용자들이 편리하게 예측결과를 분석하고 예측결과에 대한 통계결과를 자동으로 산출할 수 있도록 지원한다. 도구의 유용성을 보여주기 위하여 널리 사용되는 유전자 예측 도구인 GenScan과 GeneID의 처리 결과를 개발된 도구에 적용시켜 보았다.

키워드 : 유전자 예측, 유전자 분석, GenStructML, GenPredML, XML

An XML-Based Analysis Tool for Gene Prediction Results

Jin-Hong Kim[†] · Sang-Hee Byun[†] · Myung-Joon Lee[††] · Yang-Su Park[††]

ABSTRACT

Recently, as it is considered more important to identify the function of all unknown genes in living things, many tools for gene prediction have been developed to identify genes in the DNA sequences. Unfortunately, most of those tools use their own schemes to represent their programs results, requiring researchers to make additional efforts to understand the result generated by them. So, it is desirable to provide a standardized method of representing predicted gene information, which makes it possible to automatically produce the predicted results for a given set of gene data.

In this paper, we describe an effective XML representation for various predicted gene information, and, present an XML-based analysis tool for gene predication results based on this representation. The developed system helps users of gene prediction tools to conveniently analyze the predicted results and to automatically produce the statistical results of the prediction. To show the usefulness of the tool, we applied our programs to the results generated by GenScan and GeneID, which are widely used gene prediction systems.

Key Words : Gene Prediction, Gene Analysis, GenStructML, GenPredML, XML

1. 서 론

인간유전체사업(Human Genome Project)의 성공은 인간의 전체 염기 서열정보를 제공하였고, 이를 바탕으로 인간 유전체의 모든 유전자를 파악하기 위한 연구가 활발히 진행되고 있다. 특히 인간 유전체의 모든 유전자를 파악하기 위해서, 전체 염기 서열을 바탕으로 총체적인 서열 분석 방법과 새로운 유전자를 예측하는 도구의 개발이 요구되고 있다.

현재 유전체 염기서열을 분석하여 정확한 유전자의 위치를 파악하기 위한 GenScan[1, 2, 3]과 GeneID[4] 등의 다양한 도구들이 개발되었다. 이러한 유전자 예측 도구들은 전체 염기서열에서 단백질로 번역되는 유전자들을 예측하기 위하여 유전자의 엑손, 인트론, 시작코돈, 종료코돈 등의 정보를 예측하여 제공한다. 그러나 각 도구를 통하여 예측된 정보는 서로 다른 형식으로 제공되고 있으며, 더욱이 단순 텍스트 기반으로 제공되고 있다. 따라서 사용자는 각각의 예측 도구를 사용할 때마다 산출되는 결과물을 분석하기 쉽지 않으며, 단순 텍스트 파일로 제공되는 데이터를 바탕으로 통계정보를 계산할 때 보다 많은 노력이 요구된다. 그리고 각 도구들은 자신들의 도구에 대한 성능 평가를 기술하고 있어, 사용자는 이러한 도구들의 실제적인 예측 정확도를 비교할 수 없다. 이러한 문제를 해결하기 위하여 서로 다른 유전자 예측 도구의 결과를 정형적으로 표현할 수 있는 표준화된 문법 명세가 요구되며, 이를 바탕으로 유전자

※ 이 논문은 2004년 울산대학교의 연구비에 의하여 연구되었음.
† 준 회 원 : 울산대학교 컴퓨터 · 정보통신공학부
†† 정 회 원 : 울산대학교 컴퓨터 · 정보통신공학부
교신저자 : 박양수(yspk56@ulsan.ac.kr)
논문접수 : 2005년 4월 21일, 심사완료 : 2005년 7월 27일

예측 도구의 결과를 표현하고 예측 결과의 정확성을 자동으로 분석할 수 있는 도구가 요구된다.

본 논문에서는 다양한 형식으로 표현되는 유전자 예측 도구의 결과와 이에 대한 분석 데이터를 정형화하여 표현하는 XML 스키마[5]와 이를 기반으로 기존의 유전자 예측 도구의 예측 결과를 자동으로 분석하는 XML 기반의 도구에 대하여 소개한다. 개발된 XML 스키마는 *GenStructML(Gene Structure Markup Language)*, *GenPredML(Gene Prediction Markup Language)*, 그리고 *PredAccuracyML(Prediction Accuracy Structure Markup Language)*이다. 이들은 각각 XML 스키마를 이용하여 기존 도구의 유전자 예측 결과 및 GenBank에서 제공하는 주석정보, 하나의 유전자에 대한 예측정확도, 그리고 유전자 데이터 집합에 대한 예측 정확도 및 다양한 통계정보를 정형화하여 기술한다. 그리고 개발된 XML 기반의 유전자 예측결과 분석도구는 *GenStructML 변환기*와 *GenPredML 변환기*, 그리고 *PredAccuracyML 변환기*로 구성되어 있다.

GenStructML 변환기는 텍스트 기반의 유전자 예측 도구의 결과와 GenBank[6, 7, 8]에서 제공하는 주석(annotation)을 파싱하여 GenStructML 문서를 생성한다. 생성된 GenStructML 문서는 유전자 예측 정보를 정형화된 방법으로 기술하여 자동화된 분석 도구를 개발하는데 효과적으로 이용될 수 있다. 또한 사용자는 기존의 유전자 예측도구만을 사용하여 결과를 분석하는 과정보다 함께 제공되는 주석정보를 활용하여 보다 편리하게 결과를 이해할 수 있다. GenPredML 변환기는 GenStructML 문서에서 제공하는 예측정보(prediction)와 주석정보를 바탕으로 하나의 유전자에 대한 예측정확도 및 다양한 통계정보를 기술하는 GenPredML 문서를 생성한다. 그리고 PredAccuracyML 변환기는 GenPredML 문서의 통계정보를 바탕으로 유전자 데이터 집합(gene dataset)에 대한 예측정확도 및 통계정보를 제공한다. 3가지 도구로 구성된 유전자 예측결과 분석도구는 기존 유전자 예측 도구의 유전자 예측 결과에 대한 예측정확도 및 통계정보를 자동으로 생성하며, 기존의 도구에서 제공하지 않는 뉴클리오티드, 엑손 그리고 시그널 수준의 다양한 예측정확도를 제공한다.

본 논문에서 개발한 XML 기반 유전자 예측결과 분석도구는 다양한 데이터 셋을 이용하여 DNA 서열에서 단백질로 변역되는 구조를 예측하는 다수의 도구들의 예측 정확도를 효과적으로 비교하고 분석하는데 활용될 수 있다. 현재 개발된 각각의 유전자 예측 도구들은 자신들의 도구에 대한 성능 평가를 기술하고 있지만, 서로 다른 데이터 셋을 가지고 서로 다른 항목을 기준으로 평가하고 있어 특정 데이터 셋에 대한 예측 정확도를 서로 비교할 수 없다. 개발된 유전자 예측결과 분석도구를 이용할 경우, 사용자는 특정 데이터 셋에 대한 예측 정확도를 비교할 수 있으며, 예측 도구 개발자는 특정 부분의 예측 정확도를 분석할 수 있어 새로운 예측 도구를 개발하는데 있어서 중요한 정보를 제공받을 수 있다. 그리고 개발한 유전자 예측결과 분석도구를 이용하여 여러 예측 도구의 정확도를 자동으로 계산함으로써 주어진 서열의 종류에 따른 예측 도구의 파라미터를 보다 빠르게 결정할 수 있다. 그리고 생성된 XML 기반 문서들은 예측된 유전자의 정보를 이용한 다양한 연구 분야에 용이하게 활용될 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 대표적인 유전자 예측도구와 개발도구에서 사용되는 데이터베이스에 대하여 기술한다. 3장에서는 유전자 예측결과를 정형화하여 표현하고 통계결과를 효과적으로 분석하기 위한 XML 스키마에 대하여 설명한다. 4장에서는 XML 기반의 개발된 유전자 예측결과 분석도구에 대하여 소개한다. 5장에서는 개발된 도구를 이용하여 실험한 결과에 대하여 설명하고, 마지막으로 6장에서는 결론 및 향후과제에 대하여 기술한다.

2. 관련 연구

기존의 유전자 예측도구는 그들 고유의 형식으로 유전자 예측결과를 텍스트 형태로 제공하고 있다. 대부분의 도구들은 유전자 예측 결과를 정형화된 형식으로 표현하지 못하고 있으며, 더욱이 예측한 실행 결과의 정확도를 자동으로 계산하는 기능을 제공하지 않고 있다.

현재 개발된 대표적인 유전자 예측도구와 유전자의 주석정보를 제공하는 데이터베이스는 다음과 같다.

2.1 GenBank

GenBank(Genetic Sequence Data Bank)는 일반적인 관계형 데이터베이스와 다르게 텍스트 파일 형식으로 실제로 발현된 유전자의 정보를 저장한다. GenBank 레코드의 파일은 식별자(accession number)와 유전자 이름(LOCUS), 계통 발생학적 분류(division), 참고 문헌과 서열 데이터, 그리고 FEATURES 테이블 형식으로 다양한 정보를 제공한다. FEATURES 테이블은 조절 영역(regulatory region)의 위치, 단백질 번역(translation), 그리고 엑손(exon)과 인트론(intron) 같은 DNA 서열에 대한 정보를 자세하게 제공한다.

GenBank의 텍스트 파일 형식은 사람이 쉽게 읽을 수 있으며, 1982년 이래로 14개월 마다 배로 빠르게 증가하는 정보를 효과적으로 관리하기 위하여 GenBank 레코드의 파일 형식은 안에 있는 정보의 유연성을 허용한다. 이는 중요하다고 생각하는 것을 데이터의 주석에 입력할 수 있는 장점을 가진다.

2.2 GenScan

GenScan은 현재 가장 많이 사용되는 유전자 예측도구로써 주로 진핵세포(eucaryote) 유전자의 엑손 및 인트론의 구조(structure)와 위치(location)를 HMM(Hidden Markov Model)을 이용하여 예측한다. GenScan을 로컬컴퓨터에 설치하여 실행할 수 있으며 웹을 통하여 예측서비스를 이용할 수도 있다. GenScan은 파스타(fasta)형식이나 GenBank파일 형식의 유전자 서열을 입력으로 유전자 명, 염기서열의 길이, 그리고 하나의 단백질로 번역되는 CDS (Coding Sequence 또는 Coding Segment)의 엑손정보(타입(type)정보, 위치정보, 가닥(strand)정보, 정확도) 등의 유전자 예측정보를 텍스트 형태의 파일로 저

장한다. 결과가 비구조적인 문서로 제공되기 때문에, 사용자는 원하는 정보를 얻고자 할 때 추가적인 노력이 필요하다.

2.3 GeneID

GeneID는 1992년 Guigo에 의해 제한된 유전자 예측프로그램으로 2000년에 처음으로 공개된 이후 GenScan과 함께 많이 사용되고 있다. 현재 GeneID는 로컬에서 실행할 수 있으며, 또한 웹을 이용하여 서비스를 사용할 수 있다.

GeneID를 이용하여 유전자의 구조를 예측하는 방법은 다음과 같다.

(1) splice site와 개시/종료 코돈(start/stop codon)의 기본 인자의 위치를 예측한다. PWM(Position Weight Matrix)을 사용하여 예측된 기본 인자들의 점수(score)를 계산한다.

(2) (1)에서 예측된 기본 인자들을 이용하여 엑손을 예측하고 기본 인자들의 점수를 합하여 예측된 엑손의 점수를 계산한다.

(3) (1)과 (2)를 반복하여 예측된 엑손의 집합에서 유전자 구조를 예측한다. 예측된 유전자 구조는 엑손의 집합에서 점수가 가장 높은 엑손으로 구성된다.

GeneID 결과는 엑손의 종류(First, Internal, Terminal 그리고 Single), 엑손의 위치, 프레임 및 가닥정보, 그리고 예측된 유전자의 식별자 및 엑손의 점수 등의 정보를 포함하고 있다. 파스타 형식을 입력받아서 예측된 결과는 일반 텍스트 형식으로 저장되기 때문에 통계결과를 산출하는데 많은 노력이 요구된다.

3. 유전자 예측결과를 표현하는 XML 스키마

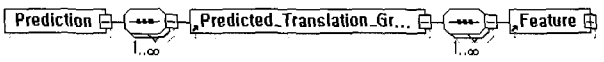
기존의 유전자 예측도구의 결과를 XML을 이용하여 표현하는 GenStructML과 이를 활용하여 사용된 도구의 예측 정확도를 기술하기 위한 GenPredML을 개발하였다. GenStructML은 유전자 예측 도구에서 제공하지 않는 유전자 주석정보와 각 유전자 예측 도구의 결과를 정형화된 방법으로 기술할 수 있는 장점을 제공한다. GenPredML은 하나의 유전자에 대한 예측정확도 및 통계정보를 제공하는 GenPredML 문서와 유전자 데이터 집합에 대한 예측정확도 및 통계정보를 제공하는 PredAccuracyML 문서를 생성하는데 사용된다. 특히, PredAccuracyML 문서는 기존의 유전자 예측도구의 다양한 측면의 예측 정확도를 제공하여 도구의 사용자 및 개발자에게 모두 유용한 정보를 제공한다.

3.1 GenStructML XML 스키마

GenStructML 스키마는 사용된 유전자 예측도구의 결과와 GenBank의 주석정보를 기술할 수 있는 Prediction 요소와 Annotation 요소를 가진다. Prediction 요소는 GFF(Gene-Finding Format, General Feature Format)[9]를 바탕으로 설계되었으며, Annotation 요소는 GenBank에서 제공하는 유전자의 기본정보와 FEATURES 테이블에 기술된 주석정보를 제공하기 위한 요소를 가진다.

3.1.1 Prediction 요소

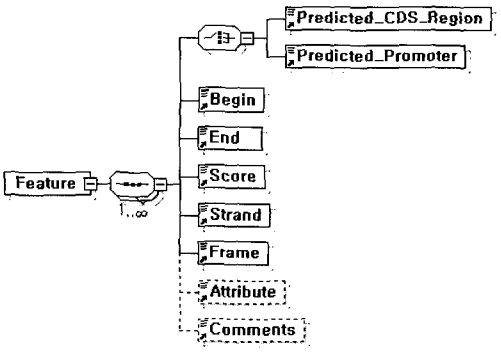
Prediction 요소는 GenScan과 GeneID의 결과파일에서 제공되는 예측된 유전자의 종류, 유전자의 시작과 끝 위치, 길이 등의 정보를 기술하기 위한 스키마이다(그림 1). Prediction 요소의 속성(attribute)은 사용된 유전자 예측도구의 종류를 나타낸다. Predicted_Translation_Group 요소는 예측된 하나의 CDS(coding sequence) 정보로써 유전자 서열에서 단백질로 번역(translation)되는 부분을 의미하며, 엑손(exon)과 프로모토(promoter) 정보를 기술하는 Feature 요소로 구성된다.



(그림 1) Prediction 요소의 스키마

(그림 2)는 Feature 요소의 하위 구성 요소를 보여준다. *Predicted_CDS_Region* 요소의 속성은 예측된 유전자의 종류(initial exon, internal exon, single exon, terminal exon)를 나타내며, Predicted_Promoter 요소는 예측된 프로모토의 시그널 정보(CAAT, TATA)를 나타낸다. Begin과 End, 그리고 Strand 요소는 엑손의 시작과 끝 위치, 가닥(strand)에 대하여 기술한다. Frame 요소는 ORF(Open Reading Frame)와 관련된 codon의 위치 정보('0', '1', '2', '.' 중 하나)의 값을 갖는다. 옵션으로 제공되는 Attribute와 Comments 요소는 각각 엑손의 구조와 예측 결과파일의 주석부분의 정보를 나타낸다.

<표 1>은 GenStructML 스키마의 *Prediction* 요소를 산출하기 위하여 참조한 GenScan의 결과파일 정보를 보여주고 있다.



(그림 2) Feature 요소의 스키마

〈표 1〉 GenScan결과와 GenStructML 스키마 요소의 관계

GenScan 결과	XML 요소	XML 속성
Sequence Name	<Gene_Name>	
Gene/Exon number	<Predicted_Translation_Group>, <Feature>	id
Type	<Predicted_CDS_Region>	type
Type	<Predicted_Promoter>	type
Strand, Begin, End, Length, Frame, P	<Strand>, <Begin>, <End>, <Length>, <Frame>, <Score>	

3.1.2 Annotation 요소

Annotation 요소는 GenBank에서 제공하는 주석 정보로써 실제로 발현된 유전자의 정보를 나타낸다. Annotation 요소는 사용자가 유전자 예측결과를 분석하거나 통계정보 산출하는데 활용된다.

다음은 주석 정보를 기술하기 위한 Annotation 요소의 하위 요소들이다.(그림 3, 4)

- Length 요소 : 전체 뉴클리오티드 염기쌍(nucleotide base pairs)의 수
- MType 요소 : 서열의 분자타입(Type of Molecule)
- Division 요소 : 계통 발생학적 분류(Division)
- Annotated_Translation_Group 요소 : 하나의 단백질로 번역되는 CDS
- DNASequence 요소 : DNA 서열(DNA Sequence)

특히, Annotated_Translation_Group 요소는 프로모토, 5'UTR(Untranslated Region), CDS, 그리고 3'UTR등의 하위 요소를 이용하여 하나의 단백질로 번역되는 유전자 정보를 기술한다. CDS 요소는 전체 유전자 서열에서 CDS의 시작과 끝 위치 및 가닥 정보를 나타내며, 하위 CDS_Resigns 요소를 이용하여 자신의 CDS를 구성하는 여러 엑손들에 대한 정보를 기술한다.

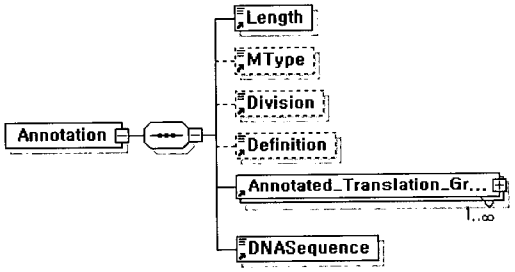
<표 2>는 GenStructML의 Annotation 요소를 생성할 때 사용되는 GenBank 파일의 정보를 보여주고 있다.

<표 2> GenBank 파일과 GenStructML 스키마 요소의 관계

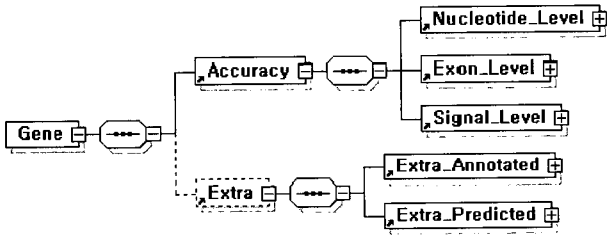
GenBank 파일	XML 요소	속성
LOCUS Name	<Gene_Name>	
ACCESSION, VERSION	<Accession>	version
Sequence Length	<Length>	
Molecule Type	<MType>	
DEFINITION	<Definition>	
GenBank Division	<Division>	
FEATURES	<Annotation>	
CDS, exon, gene, 5'UTR, 3'UTR	<Annotated_CDS_Region>, <CDS>, <_5UTR>, <_3UTR>	
promoter, CG_signal, TATA_signal, CAAT_signal, -10_signal, -35_signal	<Annotated_Promoter>	type
/protein_id	<Protein_Accession>	
/translation	<Translation>	
ORIGIN	<DNASequence>	

3.2 GenPredML XML 스키마

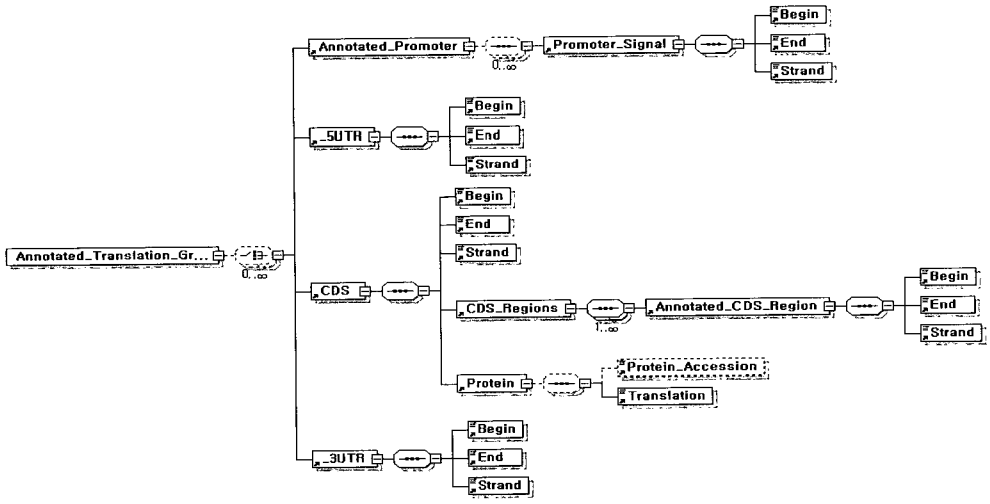
GenPredML은 사용된 도구의 유전자 예측정확도 및 통계정보를 기술하는 GenPredML 문서와 PredAccuracyML 문서에 대한 데이터 모델을 제공한다(그림 5). GenPredML 문서의 Accuracy 요소는 유전자에 대한 예측정확도를 뉴클리오티드, 엑손, 그리고 시그널 수준에서의 민감도(Sensitivity)와 특이도(Specificity)로 나타낸다. 민감도는 실제로 뉴클리오티드(엑손)로 알려진 뉴클리오티드 중에서 예측도구를 사



(그림 3) Annotation 요소의 스키마



(그림 5) GenPredML 요소의 스키마



(그림 4) Annotated__Translation__Group 요소의 스키마

용하여 제대로 예측한 확률을 의미하며, 특이도는 예측된 유전자 중에서 실제 유전자를 제대로 예측한 확률을 의미한다. 민감도와 특이도가 모두 높을수록 예측정확도의 측정 방법의 신뢰성이 높다고 볼 수 있다. 그리고 PredAccuracyML 문서는 유전자 데이터 집합에 대한 예측정확도를 기술하기 위한 Accuracy 요소를 가진다.

3.2.1 Accuracy 요소

GenPredML의 Accuracy 요소에서 제공하는 3가지(뉴클리오티드, 엑손, 시그널) 수준의 예측정확도는 GenStructML 문서의 Prediction 요소와 Annotation 요소에서 제공하는 정보를 이용하여 계산할 수 있다. 그리고 PredAccuracyML 문서의 Accuracy 요소에서 제공하는 유전자 데이터 집합에 대한 예측정확도는 유전자 데이터 집합에 포함되어 있는 GenPredML 문서들의 Accuracy 요소의 하위 요소인 Nucleotide_Level, Exon_Level, 그리고 Signal_Level들을 이용하여 계산할 수 있다[10].

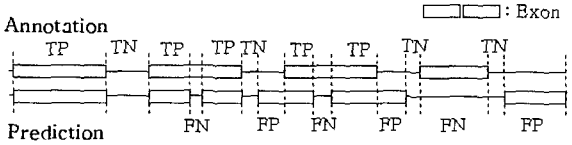
하나의 유전자에 대한 3가지 수준의 예측정확도를 정의한 스키마 및 각 요소의 값을 결정하는 방법은 다음과 같다.

① 뉴클리오티드 수준에서의 민감도와 특이도

유전자 예측 도구의 예측정확도를 뉴클리오티드 수준에서 평가하기 위해서는 (그림 6)과 같은 3가지 요소(True positive, False positive, False negative)의 값을 결정하여야 한다. 이와 같은 3가지 요소의 값은 GenStructML 문서에서 제공하는 Prediction 요소의 Feature 요소와 Annotation 요소의 CDS_Region 요소에서 제공하는 각각의 Begin과 End 요소의 값을 비교하여 계산된다.

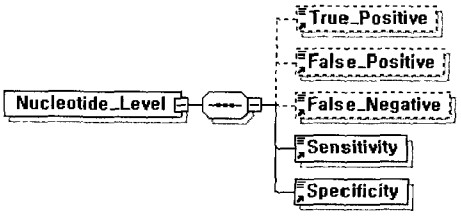
3가지 요소의 의미는 다음과 같다.

- True_Positive 요소 : 실제로 발현된 엑손의 위치와 예측된 엑손의 위치가 정확히 일치하는 범위 내에 존재하는 뉴클리오티드의 수
- False_Positive 요소 : 예측된 엑손의 위치가 실제 엑손의 위치를 벗어난 범위에 있는 뉴클리오티드의 수
- False_Negative 요소 : 실제 엑손의 위치가 예측된 엑손의 위치에 포함되어 있지 않은 범위에 존재하는 뉴클리오티드의 수



(그림 6) Nucleotide_Level의 기본요소 정의

계산되어진 3가지 요소의 값과 수식 (1)과 (2)를 이용하여 뉴클리오티드 수준에서의 민감도와 특이도를 계산할 수 있다. (그림 7)은 계산되어진 뉴클리오티드 수준에서의 민감도와 특이도를 정의하기 위한 스키마를 보여주고 있다.



(그림 7) Nucleotide_Level 요소의 스키마

$$Sensitivity = \frac{< True\,Positive > 의\ 수}{< True\,Positive > 의\ 수 + < False\,Negative > 의\ 수} \quad (1)$$

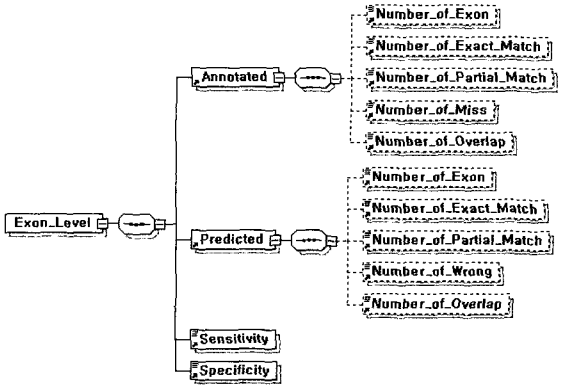
$$Specificity = \frac{< True\,Positive > 의\ 수}{< True\,Positive > 의\ 수 + < False\,Positive > 의\ 수} \quad (2)$$

② 엑손 수준에서의 민감도와 특이도

(그림 8)은 엑손 수준에서의 민감도와 특이도를 정의하는 스키마를 보여주고 있다. Annotated 요소는 GenStructML 문서의 Annotation 요소에 기술된 엑손정보를 기준으로 Prediction 요소의 엑손정보와 비교하여 산출된 결과를 정의한다. 그리고 Predicted 요소는 GenStructML 문서의 Prediction 요소에 기술된 엑손정보를 기준으로 Annotation 요소에 기술된 엑손정보와 비교하여 산출된 결과를 정의한다.

각 요소의 의미는 다음과 같다.

- Number_of_Exon 요소 : 실제 엑손의 수와 예측된 엑손의 수
- Number_of_Exact_Match 요소 : 엑손의 시작과 끝위치를 비교한 결과가 모두 일치하는 경우의 수
- Number_of_Partial_Match 요소 : 시작과 끝위치 중 한 부분만 일치한 경우의 엑손의 수
- Number_of_Overlap 요소 : 시작과 끝위치가 일치하지 않은 경우 중에서 일부분이 비교하는 엑손의 일부분에 속하게 되는 경우의 수
- Number_of_Miss 요소 : 예측하지 못한 실제 엑손의 수
- Number_of_Wrong 요소 : 예측된 엑손이 실제 엑손이 아닌 경우의 수
- Sensitivity 요소 : 실제 엑손의 수에서 엑손의 시작과 끝이 정확히 예측된 엑손의 수(3)



(그림 8) Exon_Level 요소의 스키마

· Specificity 요소 : 예측된 엑손의 수에서 엑손의 시작과 끝이 정확히 예측된 엑손의 수(4)

$$Sensitivity = \frac{< Annotated > 의 < Number of Exact Match > 의 수}{< Annotated > 의 < Number of Exon > 의 수} \quad (3)$$

$$Specificity = \frac{< Predicted > 의 < Number of Exact Match > 의 수}{< Predicted > 의 < Number of Exon > 의 수} \quad (4)$$

③ 시그널 수준의 민감도와 특이도

(그림 9)는 시그널 수준의 예측정확도를 정의하는 스키마를 보여주고 있다. 시그널 수준은

다음과 같은 4가지 요소 각각에 대한 민감도와 특이도를 정의한다.

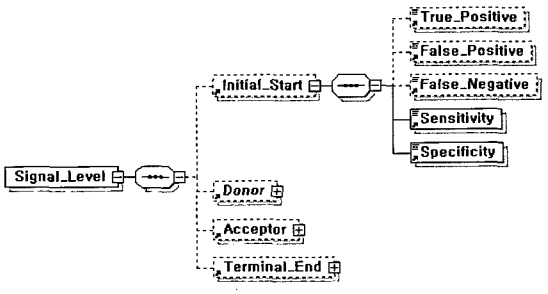
· Initial_Start 요소 : 시작엑손(initial exon)의 시작위치

· Donor 요소 : 엑손의 끝위치

· Acceptor 요소 : 엑손의 시작위치

· Terminal_End 요소 : 종료엑손(stop exon)의 끝위치

4가지 요소 각각은 True_Positive, False_Positive, False_Negative, Sensitivity, 그리고 Specificity 요소로 구성된다. True_Positive 요소는 유전자 예측도구가 정확히 엑손의 시작(또는 끝)위치를 예측한 경우를 의미하고 False_Positive 요소는 실제 엑손의 시작(끝)위치가 예측된 엑손의 시작(끝)위치에 속하지 않는 경우를 나타내며, False_Negative 요소는 실제 엑손의 시작(끝)위치가 예측된 엑손의 시작과 끝위치 사이에 존재하는 경우를 나타낸다. 이 요소들을 이용하여 뉴클리오티드 수준에서의 예측정확도를 계산하는 방법(1, 2)과 동일하게 Sensitivity와 Specificity 요소의 값을 산출한다.



(그림 9) Signal_Level 요소의 스키마

3.2.2 Extra 요소

GenPredML 문서의 Extra 요소는 하나의 유전자 정보에 대한 엑손의 시작과 끝, 엑손 수준의 결과정보(exactly correct, partially correct, miss, wrong, overlap)를 제공한다(그림 10). Extra 요소는 GenBank의 주석정보를 기준으로 예측정보를 비교하여 산출된 엑손 수준의 결과정보를 기술한 Extra_Annotated 요소와 예측정보를 기준으로 산출된 결과정보를 나타내는 Extra_Predicted 요소로 구성된다. Translation_Group 요소는 하나의 CDS를 구성하는 엑손의 정보를 기술하는 CDS_Region 요소로 구성된다. Result 요소는 각 엑손의 비교결과 정보를 정의하는 부분으로 예측된 엑손과 실제 엑손의 위치의 일치 정도를 기술한다.

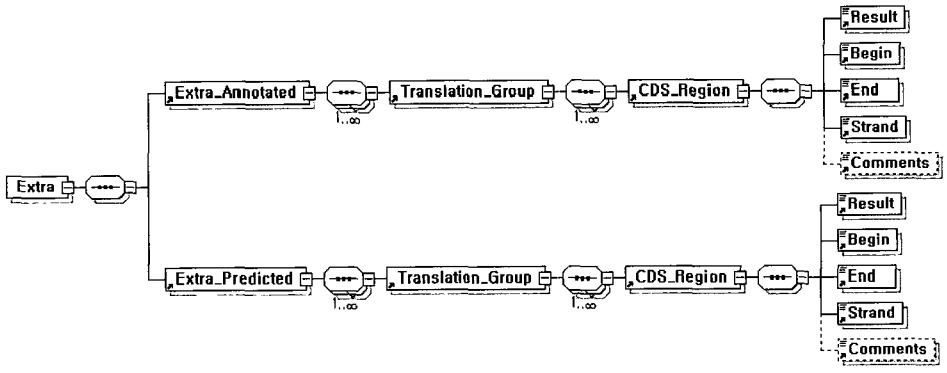
4. XML기반의 유전자 예측결과 분석도구

XML 기반의 유전자 예측결과 분석도구는 GenBank에서 제공하는 유전자 주석정보와 유전자 예측도구의 결과 정보를 이용하여 3장에서 개발된 각 XML 문서를 자동으로 생성한다. 생성된 XML 문서는 하나의 유전자뿐만 아니라 유전자 데이터 집합에 대한 예측정확도 및 통계결과를 계산하여 제공한다. 개발된 도구는 JAVA로 구현되었으며, XML 문서 파일의 파싱은 Apache XML 프로젝트에서 제공하는 XML 파서인 Xerces[11]를 이용하였다.

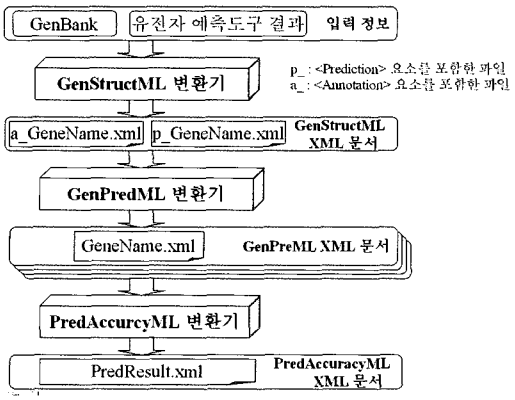
개발된 분석도구는 GenStructML 변환기, GenPredML 변환기, 그리고 PredAccuracyML 변환기로 구성된다(그림 11). GenStructML 변환기는 유전자 예측결과를 분석하여 정형화된 형식의 문서를 생성한다. GenPredML 변환기는 하나의 유전자의 예측정확도 및 통계결과를 산출하며, PredAccuracyML 변환기는 유전자 데이터 집합에 대한 예측정확도 및 통계결과를 생성한다.

4.1 GenStructML 변환기

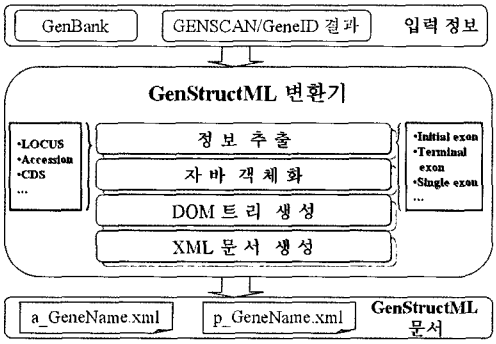
GenStructML 변환기는 유전자 예측도구의 비구조적인 텍스트 결과파일과 GenBank 파일을 입력으로 하여 실제 밝혀진 유전자 정보를 기술하는 GenStructML 문서(a_GeneName.xml)와 유전자 예측정보를 표현하는 GenStructML 문



(그림 10) Extra 요소의 스키마



(그림 11) 개발된 분석도구 구성



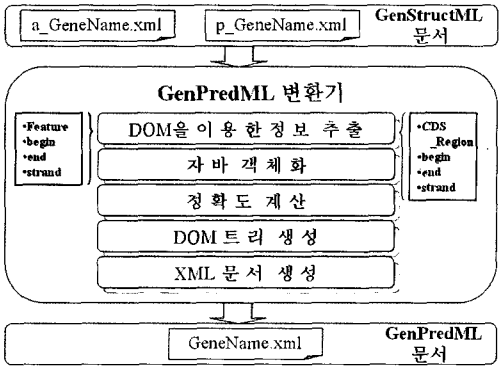
(그림 12) GenStructML 변환기

서(p_GeneName.xml)를 생성한다(그림 12). 현재 개발된 GenStructML 변환기는 GenScan과 GeneID의 예측결과를 GenStructML 문서로 변환한다. 그리고 추가적인 유전자 예측정확도 및 통계결과를 산출하기 위해서 GenStructML 변환기의 정보추출 과정만을 각 도구의 결과에 맞게 수정하면 가능하다.

(그림 12)는 GenStructML 변환기의 동작과정 및 데이터의 흐름을 보여준다. 우선 GenStructML 변환기는 유전자 예측도구의 결과파일을 분석하여 유전자로 예측된 정보들을 추출한 후 GenStructML 스키마에서 정의된 형식으로 객체화한다. 객체화된 데이터들은 DOM(Document Object Model) API[12]를 이용하여 Prediction 요소를 포함한 GenStructML 문서로 변환된다. 그리고 입력된 GenBank 파일의 유전자명, 전체서열의 길이 및 엑손 등의 정보(<표 2>)를 추출하고 Prediction 요소를 포함한 GenStructML 문서를 생성한 방법과 동일한 방법으로 Annotation 요소를 포함한 GenStructML 문서를 생성한다. GenStructML 스키마의 Prediction과 Annotation 요소는 각각 다른 파일로 생성된다.

4.2 GenPredML 변환기

GenPredML 변환기는 GenStructML 문서를 입력으로 하나의 유전자에 대한 예측정확도와 통계정보를 포함한 GenPredML 문서를 생성한다. (그림 13)은 GenPredML 문서를 생성하는 GenPredML 변환기의 전체구성 및 데이터의

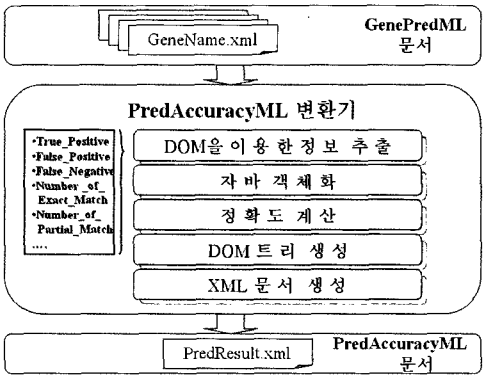


(그림 13) GenPredML 변환기

흐름을 보여준다. GenPredML 변환기는 DOM을 이용하여 GenStructML 문서의 Annotation 요소와 Prediction 요소의 값을 추출하여 자바 객체화한다. 이 데이터를 이용하여 뉴클리오티드, 엑손, 시그널 수준에서의 예측정확도 및 통계정보, 그리고 엑손 수준에서 보다 자세한 정보를 계산하고 계산된 정보를 GenPredML 문서로 저장한다.

4.3 PredAccuracyML 변환기

PredAccuracyML 변환기는 유전자 데이터 집합에 포함된 각 유전자의 예측정보인 GenPredML 문서들을 입력으로 PredAccuracyML 문서를 생성한다. PredAccuracyML 문서는 GenPredML의 Accuracy 요소만을 포함하여 생성된 XML 문서이다. PredAccuracyML 변환기는 GenPredML 문서를 DOM을 이용하여 자바 객체화하고, 각 수준별 통계정보를 통합하여 유전자 데이터 집합에 대한 통계정보를 산출한다. (그림 14)는 PredAccuracyML 문서를 생성하는 PredAccuracyML 변환기의 구성 및 데이터의 흐름을 보여준다.



(그림 14) PredAccuracyML 변환기

5. 실험 결과

XML 기반의 유전자 예측분석 도구를 이용하여 4개의 실험 데이터 집합에 대한 GenScan과 GeneID의 예측결과를 개발한 도구를 이용하여 분석하였다. 사용된 실험 데이터

집합은 Celeg76[13], GenBank에서 추출한 570개의 서열[14], 인간 유전자 서열인 H178[4], 그리고 HMR195[15]이다. GenScan과 GeneID의 유전자 예측 정확도를 분석하기 위하여, 각 실험 데이터 집합에 대한 GenScan과 GeneID의 실행 결과와 GenBank 데이터를 이용하여 GenStructML 문서를 생성하였다. 생성된 GenStructML 문서를 바탕으로 GenPredML 문서를 생성한 후, GenScan과 GeneID의 예측정확도 및 통계결과를 포함하는 PredAccuracyML 문서를 생성하였다. 분석 실험은 Intel(R) XEON(TM) 프로세서 2.0GHz와 1G 메모리를 탑재한 리눅스 운영체제에서 수행되었다.

5.1 실험 데이터 집합(the sequence test set)

실험에 사용된 데이터 집합에 포함된 모든 유전자 서열은 단백질로 변역 가능한 부분 서열은 ATG로 시작하는 시작 코돈과 TTA, TAG, 또는 TGA로 끝나는 종단 코돈을 가진다. 그리고 유전자 서열에 포함된 모든 엑손(exon)의 acceptor site에는 AG를 포함하고 있으며, donor site에는 GT를 가진다.

다음은 실험에 사용된 각 서열 데이터 집합의 특징들이다.

① Celeg76 : Caenorhabditis Elegans에서 추출된 서열로써 12,637 bp의 길이를 가진다. 유전자 하나에 포함된 엑손의 평균 개수는 6.96이다.

② 570Seq(서열 570개) : Burset과 Guigo가 실험한 570개의 척추동물 유전자 서열(vertebrate gene sequences)은 2,892,149bp 길이이며, 2,649개의 엑손(exon)을 가진다.

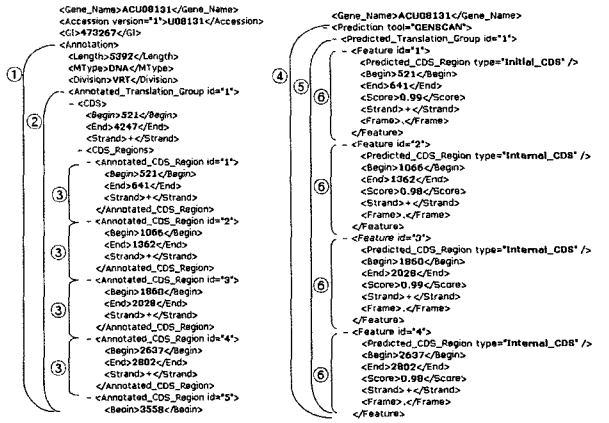
③ H178 : EMBL 데이터베이스에서 추출된 인간 유전자 서열이다. H178에 포함된 서열은 50%의 G+C를 포함하고 있으며, 평균 7169 bp의 길이를 가진다. 그리고 하나의 서열은 오직 하나의 유전자 정보를 가지며 평균 5.1개의 엑손(exon)으로 구성되어 있다.

④ HMR195 : 1997년 GenBank에 포함된 서열로써 인간(human-103개)과 쥐(mouse-82개, rat-10개)에서 추출되었다. 평균 서열의 길이는 7,096 bp이며 하나의 유전자에 4.86개의 엑손(exon)이 포함되어 있다.

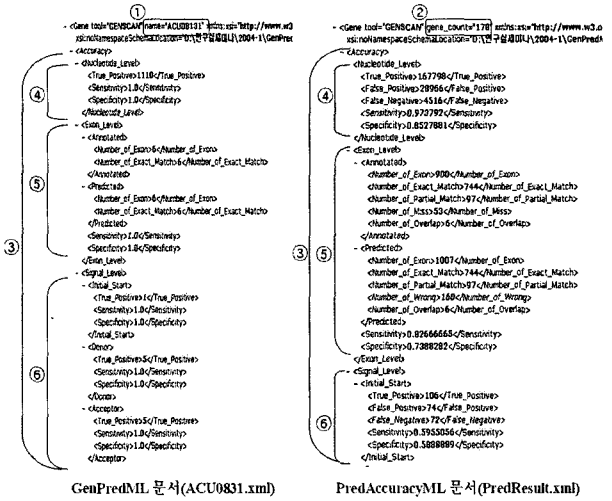
5.2 유전자 예측 결과 분석을 내포하는 XML 문서

(그림 15)는 유전자 ACU08131(h178)에 대한 GenStructML 변환기의 결과문서를 보여준다. GenStructML 문서의 (그림 15-①)은 실제 유전자 정보를 표현하며, (그림 15-④)는 유전자 예측도구의 결과로 예측된 유전자 정보를 나타낸다. (그림 15-②)와 (그림 15-⑤)는 하나의 단백질로 번역되는 CDS의 정보를 기술하며, (그림 15-③)과 (그림 15-⑥)은 하나의 CDS를 구성하는 엑손들의 정보를 정의하고 있다.

(그림 16)은 GenPredML 변환기를 통해 생성된 GenPredML 문서와 PredAccuracyML 변환기를 통해서 생성된 PredAccuracyML 문서를 보여준다. GenPredML 문서는 (그림 15)의 결과를 입력으로 하여 유전자 하나에 대한 예측정확도 및 통계정보를 나타낸다. PredAccuracyML 문서(PredResult.xml)는 하나의 데이터 집합에 대한 각 도구의



(그림 15) GenStructML 변환기의 결과문서



(그림 16) GenPredML 변환기와 PredAccuracyML 변환기의 결과문서

예측 정확도에 대한 정보를 제공한다. 데이터 집합 H178의 경우, 178의 GenPredML 문서가 생성되었고, 이들에 대한 유전자 예측 정확도는 1개의 PredAccuracyML 문서에서 제공된다. GenPredML 문서와 PredAccuracyML문서의 주요 요소에 대하여 살펴보면, GenPredML 문서와 PredAccuracyML 문서의 루트 요소의 속성 값으로 (그림 16-①)은 유전자 명을 정의하고 (그림 16-②)는 데이터 집합의 유전자의 전체 수를 의미한다. (그림 16-③)은 유전자의 예측 정확도를 기술하는 요소이며, (그림 16-④), (그림 16-⑤) 그리고 (그림 16-⑥)은 뉴클리오티드 수준의 예측정확도, 엑손 수준의 예측정확도 그리고 시그널 수준의 예측정확도를 각각 표현한다.

5.3 유전자 예측 도구의 예측정확도 분석

4개의 실험 데이터 집합을 대상으로 GenScan과 GeneID을 이용하여 유전자 예측 결과를 생성하였다. 개발된 XML 기반 유전자 예측분석 도구는 각 도구로부터 생성된 예측 정

〈표 4〉 유전자 예측 도구의 예측정확도 분석 결과

분석 단위	데이터 집합	예측정확도											
		GenScan						GeneID					
		TP(EE)	FP(PE)	FN(WE)	(OE)	SN	SP	TP(EE)	FP(PE)	FN(WE)	(OE)	SN	SP
뉴클리오티드	Celeg76	39673	28115	23706		63%	59%	37887	21633	26434		59%	64%
	570Seq	406854	50210	23156		95%	89%	367714	33578	63028		85%	92%
	H178	167798	28966	4516		97%	85%	153482	18775	18832		89%	89%
	HMR195	167452	25070	7172		96%	87%	157596	17319	17664		90%	90%
	평균					88%	80%					81%	84%
엑손	Celeg76	(290)	(95)	(126)	(21)	30%	33%	(76)	(48)	(69)	(26)	23%	35%
	570Seq	(1994)	(247)	(264)	(85)	78%	77%	(1624)	(298)	(140)	(93)	67%	75%
	H178	(744)	(97)	(160)	(6)	83%	74%	(615)	(125)	(90)	(5)	68%	74%
	HMR195	(682)	(111)	(135)	(5)	80%	73%	(606)	(103)	(71)	(8)	70%	77%
	평균					68%	64%					56%	65%
시그널 Initial	Celeg76	11	30	29		28%	27%	11	29	30		27%	28%
		359	179	192		65%	67%	271	214	282		49%	56%
		106	74	72		60%	59%	73	84	105		41%	46%
		103	63	61		63%	62%	91	63	75		55%	59%
	평균					54%	54%					43%	47%
시그널 Donor	570Seq	115	121	162		42%	49%	97	81	186		34%	54%
		1752	317	238		88%	85%	1489	264	505		75%	85%
		696	130	66		91%	84%	607	96	155		80%	86%
		663	123	63		91%	84%	583	89	148		80%	87%
	평균					78%	76%					67%	78%
시그널 Accepto	H178	103	146	174		37%	41%	83	96	200		29%	46%
		1705	347	285		86%	83%	1470	200	524		74%	88%
		671	156	51		93%	81%	586	92	136		81%	86%
		614	153	79		89%	80%	554	80	144		79%	87%
	평균					76%	71%					66%	77%
시그널 Terminal	HMR195	11	43	29		28%	20%	11	30	30		27%	27%
		422	99	129		77%	81%	329	73	224		59%	82%
		112	69	26		81%	62%	91	41	47		66%	69%
		96	51	35		73%	65%	90	26	43		68%	78%
	평균					65%	57%					55%	64%

* TP : True Positive, FP : False Positive, FN : False Negative, SN : Sensitivity, SP : Specificity
* EE : Exact Exon, PE : Partial Exon, WE : Wrong Exon, OE : Overlap Exon

보를 분석하여 8개의 PredAccuracyML 문서를 생성하였다.

<표 4>는 생성된 XML 문서를 바탕으로 작성되었으며, 실험 데이터 집합과 분석 수준별로 GenScan과 GeneID의 유전자 예측정확도를 보여준다. 결과적으로 GenScan은 GeneID 보다 뉴클리오티드, 엑손, 시그널 수준에서 모두 약 10% 정도 정확한 예측 결과를 보여주고 있다. GenScan은 뉴클리오티드 수준에서 평균적으로 88%의 민감도(sensitivity)와 80%의 특이도(specificity)를 나타내고 있다. 그러나 Celeg76을 제외하고 모두 95%가 넘는 민감도를 나타내어 포유동물의 서열에서 유전자를 예측하는 프로그램으로 충분한 만족도를 나타내고 있다. Celeg76의 경우, 각 분석단위에서 두 예측 도구 모두 낮은 수치의 유전자 예측정확도를 나타내고 있다. 개발된 XML 기반 유전자 예측분석 도구는 예측 오류로 나타난 엑손 및 시그널에 대한 정보를 포함하는 XML 문서를 제공한다. 사용자는 개발된 유전자 예측 분석 도구를 이용하여 자동으로 유전자 예측 도구의 예측정확도 및 통계정보를 산출할 수 있으며, 더욱이 각 도구에서 제공하지 않는 GenBank 정보를 함께 활용할 수 있어 보다 편리하게 예측결과를 분석할 수 있다. 또한, 같은 데이터 집합에 대하여 PredAccuracyML 문서에서 제공하는 결과를 조사하여 유전자 예측에 사용한 도구의 예측정확도를 <표 4>와 같이 비교해 볼 수 있다.

6. 결론 및 향후 연구과제

유전체 프로젝트가 진행되면서 유전체 전체의 염기서열 분석이 유전자에 대한 연구를 가능하게 할 수 있다는 것이 가시화됨으로써 유전체의 염기서열을 분석하여 정확한 유전자의 위치를 밝혀내기 위해 유전체 염기서열분석 연구가 활발히 진행되었다. 현재 개발되어 있는 대부분의 유전자 예측도구들은 고유의 단순 텍스트 파일 형식으로 결과를 제공하고 있어 결과를 분석하고 통계정보를 산출하는데 많은 노력이 필요한 형편이다. 그러므로 다수의 유전자 예측도구들의 예측 정확도를 자동으로 비교하고 평가할 수 있는 도구의 개발이 필요하다.

본 논문에서 기술된 GenStructML, GenPredML, 그리고 PredAccuracyML과 XML 기반 유전자 예측결과 분석도구는 유전자 예측결과를 XML 스키마를 이용하여 표현하고 자동으로 유전자의 예측정확도 및 통계정보를 생성한다. 유전자 예측결과 분석도구를 이용하여 생성된 XML 문서들은 유전자 예측결과를 정형화된 표현 방식으로 기술하며 관련된 연구 분야에 기반 자료로 활용될 수 있다. 특히 GenStructML 문서는 유전자 예측결과와 함께 GenBank에서 제공하는 유전자 주석 정보를 제공함으로써 사용자는 기존의 예측도구만 사용하였을 때보다 쉽게 결과를 이해할 수 있다.

그리고 기존의 도구만을 이용하여 얻을 수 있는 정보 외에도 뉴클리오티드, 엑손 그리고 시그널 수준의 다양한 예측 정확도를 제공한다. 개발된 도구를 이용하는 사용자는 기존의 유전자 예측도구들의 결과물을 분석하는 수작업을 거치지 않고 하나의 유전자에 대한 정보뿐만 아니라 유전자 데이터 집합에 대한 통계정보를 비교해 볼 수 있다.

향후에는 현재의 시스템을 다양한 유전자 예측도구들의 결과를 분석할 수 있는 보다 일반된 예측결과 분석도구로 발전시킬 예정이며, GenPredML을 이용한 유전자 구조 예측 뷰어를 개발할 예정이다.

참 고 문 헌

[1] Burge, C. and Karlin, S., "Prediction of complete gene structures in human genomic DNA," J Mol Biol, Vol.266, pp. 78-95, 1997.

[2] Burge, C., "Identification of genes in human genomic DNA," PhD thesis, Stanford University, Stanford, CA., 1997.

[3] Burge C. and Karlin, S., "Finding the genes in genomics DNA," Current Opinion in Structural Biology, Vol.8, pp. 346-354, 1998.

[4] Guigó, R., Agarwal, P., Abril, J.F., Burset, M. and Fickett, J.W., "An Assessment of Gene Prediction Accuracy in Large DNA Sequences," Genome Research Vol.10, No.10, pp. 1631-1642, 2000.

[5] Yergeau, F., Bray, T. and Paoli, J., Sperberg-McQueen CM, Maler E : Extensible Markup Language (XML) 1.0, 3rd Ed., W3C, 2004.

[6] Dennis, B., Ilene Karsch-Mizachi, David, L., James, O., Barbara R. and David, W., "GenBank. Nucleic Acids Research," Vol.28, No.1, pp.15-18, 2000.

[7] Dennis, B., Ilene Karsch-Mizachi, David, L., James, O., Barbara, R. and David, W., "GenBank. Nucleic Acids Research," Vol.32, No.1, pp.23-26, 2004.

[8] DDBJ, EMBL and GenBank, The DDBJ/EMBL/GenBank Feature Table : Definition, Ver. 6.0, 2003.

[9] GFF, GFF (General Feature Format) Specifications Document, WWW document (http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml), 2004.

[10] Burset, M. and Guigó, R., "Evaluation of gene structure prediction programs," Genomics, Vol.35, pp.353-367, 1996.

[11] The Apache Software Foundation, Xerces : XML parsers in Java, Apache XML Project, WWW document (http://xml.apache.org/), 2004.

[12] W3C, Document Object Model(DOM) Level 1 Specification, Ver. 1.0, WWW document (http://www.w3.org/TR/REC-DOM-Level-1/), 1998.

[13] Ana, P., Pedro T., "DECIDE - Gene Finding Evaluation Tool", WWW document (http://decide.inesc-id.pt/index.php), 2005.

[14] Burset M, Guigo R, "Evaluation of gene structure prediction programs," Genomics, Vol.35, pp.353-367, 1996.

[15] Rogic, S., Mackworth, AK., Ouellette, FB., "Evalution of gene-finding programs on mammalian sequences," Genome Research, Vol.11, No.5, pp.817-832, 2001.

김 진 홍

e-mail : avenue@ulsan.ac.kr

1999년 울산대학교 전자계산학과(공학사)

2001년 울산대학교 대학원 컴퓨터·정보통신 공학부(공학석사)

2005년 울산대학교 대학원 컴퓨터·정보통신공학부(공학박사)

2005년~현재 울산대학교 컴퓨터·정보통신공학부 객원교수

관심분야 : 생물정보학, 협업지원시스템, 분산시스템, 프로그래밍 언어 등

변 상 희

e-mail : heeya@mail.ulsan.ac.kr

2003년 울산대학교 컴퓨터·정보통신 공학과(학사)

2005년 울산대학교 대학원 컴퓨터·정보통신공학부(석사)

2005년~현재 온넷기술 주식회사 연구원

관심분야 : 생물정보학, 협업지원 시스템, 웹 프로그래밍, 유.무선 통신프로토콜 등

이 명 준

e-mail : mjlee@ulsan.ac.kr

1980년 서울대학교 수학과(학사)

1982년 한국과학기술원 전산학과(석사)

1991년 한국과학기술원 전산학과(박사)

1982년~현재 울산대학교 컴퓨터·정보통신공학부 교수

1993년~1994년 미국 버지니아대학 교환교수

2005년~현재 미국 캘리포니아 주립대학 교환교수

관심분야 : 프로그래밍언어, 분산 객체 프로그래밍 시스템, 병행 실시간 컴퓨팅, 인터넷 프로그래밍시스템, 생물정보학 등

박 양 수

e-mail : yspk56@ulsan.ac.kr

1978년 울산대학교 전자계산학과(학사)

1981년 서울대학교 계산통계학과(석사)

1986년~현재 서울대학교 계산통계학과 박사과정

1980년~현재 울산대학교 전자계산학과 교수

관심분야 : 분산처리, 컴퓨터알고리즘, 생물정보학 등