DOI : 10.5392/JKCA.2011.11.3.065

DNA 서열 분석을 위한 통합 시스템

Unification System for Analysis of DNA Sequence

송영옥, 장덕진
우송대학교 컴퓨터정보학과

Young-Ohk Song(yysong@wsu.ac.kr), Duk-Jin Chang(djchang@wsu.ac.kr)

요약

첨단 과학기술의 등장으로 유전자 정보의 활용 방법과 다양한 분야에서의 융합 형태가 속출하고 있는 현실에 우리는 서있다. 바이오 데이터의 분석을 기반으로 많은 연구와 개발이 이루어지면서 새로운 연관성과 정보를 찾아내기 위한 바이오인포매틱스의 많은 목표들이 설정되고 있는 실정에서 데이터의 정확한 분석을 도울 수 있는 도구의 필요성이 더욱 더 대두되고 있다. 본 논문에서는 기존에 제공되는 바이오 데이터 분석을 위한 여러 가지 도구들의 단점들을 보완할 수 있는 시스템을 개발함으로써 사용자에게 보다 편리한 연구 도구를 제공하고자 한다. 바이오 데이터 분석을 위한 작업으로 ORF 축출, 바이오 서열 정보 검색 및 유사성 비교등의 작업을 분리된 환경이 아닌 통합된 환경에서 제공하고 기존 분석 시스템에서 부족한 연속성을 제공하도록 설계하였다.

■ 중심어 : | 바이오인포메틱스 | DNA 서열 | 단백질 서열 | ORF |

Abstract

We stand at real world that some practical use method of gene information appears in succession by entrance on the stage of advanced techonlogy. As a lot of studies and development are achieved based on analysis of bio data, necessity of a tool that can help correct interpretation of data is required more and more in a lot of targets of bioinformatics to search new relation and information are established..

In this paper, we are offered in existing I wish to offer user a more convenient study tool developing system that can supplement shortcomings of various tools for data analysis. So we've designed to offer in united environment that is not environment that is parted ORF driving out, bio information retrieval and work of similarity comparison lamp to work for bio data analysis and offers lacking consecutiveness in existing analysis system.

■ keyword : | Bioinformatics | DNA Sequence | Protein Sequence | ORF |

Ⅰ. 서 론

바이오인포매틱스라는 용어가 우리에서 보편화 되고 있는 현실에서 바이오인포메틱스의 목표는 바이오 데이터의 정확한 분석을 통해 현실세계에서 필요한 연관성을 찾고 정보를 이용하여 고부가가치를 얻을 수 있는 방법들을 모색하는 것으로 향하고 있다. 이와같은 바이오인포매틱스는 생명 과학 분야에 전산학, 통계학, 그리고 수학 등이 통합되면서 IT분야와 BT분야를 최적으로 융합한 학문이라 할 수 있다.

접수번호 : #101129-008
접수일자 : 2010년 11월 29일
심사완료일 : 2010년 12월 22일
교신저자 : 장덕진, e-mail : djchang@wsu.ac.kr

생명과학 분야에서 컴퓨터를 활용할 수 있는 대표적인 예로는 서열화, 서열화 분석, 비교, 진화, 돌연변이 추적, 약 설계를 위한 유사성 비교, 단백질 기능 예측, 그리고 세포 메커니즘과 질병 발생에서의 유전자 역할 예측 등 다양한 분야를 들 수 있다. 또한 데이터베이스를 구축함으로써 다른 데이터 연구에서 클로닝 작업을 하고자 할 때 가용성을 제공할 뿐만 아니라 비교 유전학을 위한 기반으로 사용될 수 있다. 바이오 데이터 분석의 가장 초기 과정으로 DNA와 단백질 서열에 대한 데이터 정보 검색을 들을 수 있으며, 이와 같은 생물학적 데이터 마이닝 작업을 하기위해 NCBI, EBI, GenomeNet 등에서 제공하는 데이터베이스를 활용하는 각종 도구들이 출시되고 있다. 이와같은 도구들의 작업으로 정보검색을 위한 스트링 검색과 서열이나 구조의 검색, 배열 및 비교를 위한 유사성 검색과 같은 작업을 수행한다[1-4][7].

Ⅱ. 연구배경

바이오 데이터 분석을 위해 현재까지 가장 일반적으로 이용되는 도구로 NCBI[6], GenemeNet[8] 등을 볼 수 있다. 이들 도구는 대부분 웹기반으로 제공되고 있으므로 공통적인 특징은 관련 분석도구들을 언제, 어디서나 인터넷이 이용한 환경이라면 이용할 수 있도록 하는 것을 목적으로 하고 있다. 그러므로 웹의 일반적인 특징처럼 정보제공을 목적으로 하는 위의 분석 사이트들은 개개인의 분석 데이터 정보 관리 부분은 고려하지 못하고 있다. 또한 각각의 분석 작업을 독립적으로 수행하도록 설계되어있다. 이로 인해 분석 작업을 위해 필요할 때마다 웹사이트에 접속하여 메뉴를 선택하고 분석하고자 하는 데이터의 텍스트를 입력하는 단순한 작업을 반복해야할 뿐 아니라 분석 작업 종료 이전에 웹 사이트 종료나 기타 컴퓨터 시스템의 종료가 이뤄진다면 같은 과정을 필요할 때마다 반복해야하는 불편함이 존재한다. 이는 기존 시스템의 개발 목표에서는 고려할 수 없는 한계점이지만 사용자편에서는 반드시 개선되어야할 사항이라 할 수 있다[1].

본 논문에서는 기존 시스템의 단점들을 보완하면서 데이터 분석을 효율적으로 할 수 있는 통합 환경을 설계[1]한 것을 바탕으로 하고 소프트웨어를 구현하기위한 알고리즘을 설계하여 소프트웨어 시스템을 완성하고자 한다. 본 논문은 연구배경에 이어 다음과 같이 구성하였다. 제 3절에서는 DNA 설열 분석을 위한 시스템 설계와 구현부분으로 분석 시스템에서 구현하고자하는 분석 작업들을 위한 세부적인 알고리즘 부분과 구현에 필요한 절차들을 소개한다. 제 4절에서는 3절에서 구현된 시스템을 이용하여 분석 작업을 테스트하고 성능평가를 거쳐 기존 시스템과의 차이점들을 비교 분석하였고 마지막으로 5절에서는 본 논문에서 구현된 시스템을 바탕으로 바이오 데이터 분석에서 활용됨으로써 사용자에게 주어질 장점들을 소개하고 향후 추가 연구 진행 방향 등을 제시하고자 한다.

Ⅲ. DNA 서열 분석 알고리즘 설계 및 구현

이번 장에서는 인터넷 기반 바이오 데이터 분석 시스템을 설계[1]를 바탕으로 분석 시스템에 포함되는 구성요소인 ORF 검색을 위한 알고리즘, 유전자 탐색을 위한 알고리즘을 설계 구현하고 이를 바탕으로 서열 유사성 비교를 위해 필요한 조건들을 자동 선택할 수 있는 시스템을 구현하고자 한다[5][9][10].

3.1 DNA 서열 분석 알고리즘 설계

유전자를 탐색하는 과정은 DNA전사, mRAN 번역 등의 과정을 처리하기 위한 알고리즘과 전체 염기서열에서 ORF부분을 찾아내는 과정을 처리하기 위한 알고리즘을 필요로 한다.

본 논문에서 이미 설계한 시스템[1]을 바탕으로 이루어질 수 있는 전반적인 분석 프로세스를 다음 [그림 1]과 같이 표현할 수 있다. 바이오 데이터 전체 서열정보에서 유용한 ORF 자료가 축출될 것이며 이를 바탕으로 DNA 복제, DNA 전사, mRNA 번역과정이 진행되고 관련 DB서버에 자동 연결되어 가장 유사한 서열정보를 얻게 될 것이다.

그림 1. DNA에서 단백질 정보 획득 과정

위의 그림에서 표기된 각 부분은 바이오 데이터 분석을 위한 필수과정으로 이를 위한 각 알고리즘을 추가하였으며 각 단계의 연관성을 적용함으로써 기존 제공된 분석도구들에서 문제시 되었던 분석 작업의 연속성, 수동적인 입력 과정의 과다로 인한 오류 발생률이 증가할 수밖에 없었던 사항에 대해서 본 시스템에서 개선시키고자 한다.

3.1.1 ORF 검색 알고리즘

리딩 프레임이란 연구 중인 세포의 DNA에서 실제로 어디에서부터 DNA가 단백질로 번역되는지를 이르는 말이다. 위의 [그림 1]에서 보인 진행과정 중에서 ①에 해당하는 단계이다.

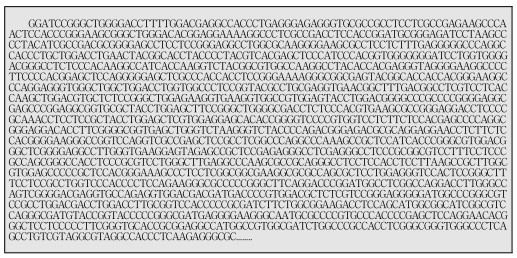

그림 2. DNA 서열 샘플 데이터

연구 분야에서 검색된 DAN를 텍스트로 변환한 입력값을 통해 조건에 맞는 ORF를 검색해야하는 과정을 예로 들면 [그림 2]의 DAN 서열 일부분에서 [그림 3]과 같이 ORF의 축출 결과를 찾을 수 있어야 한다. 이는 ORF 검색 조건에서 주어진 서열의 길이보다 작은 값들은 무시하게 된 것이다.

그림 3. 100dp 조건의 ORF 검색 결과

이와같은 DNA 또는 mRNA 서열에서 ORF를 찾기 위한 알고리즘을 다음 [그림 4]와 같이 설계한다.

1. constant process
- *1.1 Sequence minimum length of ORF for finding*
- *1.2 Overlapping data length*

2. Initialization of some variables
- *2.1 시작 코돈의 위치 저장*
- *2.2 종료 코돈의 위치 저장*
- *2.3 시작 코돈과 종료 코돈 사이에 존재하는 전체 서열길이 저장*
- *2.4 시작 코돈 처리를 위한 배열*
- *2.5 종료 코돈 처리를 위한 배열*

3. 입력된 DNA 또는 mRNA의 서열에서 6개 프레임 각각에 대한 3개의 뉴클리오타이드로 구성된 코돈 정의
- *3.1 입력된 서열의 타입을 판단하여 그에 따른 시작 코돈과 종료 코돈을 결정*
- *3.2 original sequence에 대한 6 frame 조사*
- *3.3 cpmpliment seqence에 대한 6 frame 조사*

4. 서열 중 시작 코돈과 일치하는 부분에서부터 종료 코돈을 만나는 지점까지 ORF로 간주
- *4.1 검색된 ORF 중간에 또 다른 시작 코돈이 존재하는 경우 이 시작 코돈에서 종료 코돈까지의 길이가 overlapping data length보다 작다면 새로운 ORF의 시작 코돈으로 간주하며, 이 길이보다 높다면 이전 ORF 코드만 취한다.*

5.검출된 각각의 ORF 정보에 대한 배열처리

그림 4. ORF 검색 알고리즘

3.1.2 DNA 전사 생성 알고리즘

다음은 [그림 2]의 ②부분에 해당하는 DNA를 RNA로 전사하는 과정에 대한 알고리즘 부분이다. DNA로부터 RNA가 만들어지는 과정을 전사과정이라고 하며 RNA는 구조적으로 DNA와 유사하다. 유전자 발현의

첫 단계인 전사과정은 유전자를 구성하는 DNA를 베낀 RNA분자를 합성하는 것으로 이와 같은 과정을 알고리즘 표현의 일부분을 다음 그림과 같이 구성하였다.

```
void ToRNA(KEY *Seq, KEY *RNA) {
    int i;
   for(i=0;i<RNA->size;i++)
    if(Seq->buf[i]=='T')
        RNA->buf[i]='U'
   else
        RNA->buf[i]=Seq->buf[i];
    .....
 }
```

그림 5. DNA replication 알고리즘

본 알고리즘은 크게 DNA 복제과정, DNA를 RNA로 전사하는 모듈과, RNA를 아미노산으로 번역하는 모듈, ORF Finder 모듈 그리고 염기서열이나 단백질 서열의 쌍에 대한 유사성 비교를 위한 모듈로 나눌 수 있다.

3.1.3 단백질 획득 알고리즘

다음 [그림 6]과 [그림 7]에서는 단백질 획득 과정을 위한 절차와 단백질 획득 알고리즘에서 주요 부분에 대한 코드부분을 보여주고 있다.

```
1. Initialization of variables
 1.1 Init. of Amino array
 1.2 Store values at Amino sequence
 1.3 Store of sequence of each kind of frame
2. Start of conversion for first frame
 2.1 Make a unit by three RNA sequence
     U->0, C->1, A->2, G->3
 2.2 얻어진 서열에 주어진 값을 순서에 맞게 조합
 2.3 아미노산 코드 배열에서 해당 값을 뽑아 RNA 3단
     위 서열에 할당
 2.4 첫 번째 프레임에 대해 반복 작업
3. 프레임을 하나 증가하여 코드 변환 시작
4. 각 프레임에 대해 위의 과정을 종료 시점까지 반복
```

그림 6. mRNA 번역 알고리즘

```
char *start_lit[]={"AUG","CUG,"GUG", "UUG", "AUU";
char *stop_lif[] = {"UAA","UAG","UGA"};

char symbol[][4][4]={{{'F','F','L','L'},
{'S','S','S','S'},
        {'Y','Y','*','*'},{'C','C','*','W'} },
    {{'L','L','L','L'}, {'P','P','P','P'},
    {'H','H','Q','Q'}, {'R','R','R','R'}},
    {{'I','I','I','M'}, {'T','T','T','T'},
    {'N','N','K','K'}, {'S','S','S','S'} },
    {{'V','V','V','V'}, {'A','A','A','A'},
    {'D','D','E','E'}, {'G','G','G','G'}},
    };.

void ToAMIINOI(char *Seq,char*Amino, int
aminosize)
{
    int idx[3], size=aninosize...
    char *three...
      for(i=0;i<size;i++)
        Anino[i]=symbol[idx[0]][idx[1]][idx[2]]; ...
}

void ToAMIINOI_multi(List *dna_list, KEY **amino)
{
   .....

*amino=(KEY*)malloc(sizeof(KEY)*lise_size(dna_list));
    ....

for(i=0;i<list_size(dna_list);i++,element=element->next)
    {
     KeyInit(&(*amino)[i],free);
     ...
     ToAMINO(dan_bufstart(element), (*amino)[i].buf,
              (*amino)[i].size);
    }
}
```

그림 7. 단백질 획득 알고리즘

3.2 DNA 서열 분석 통합 시스템 구현

3.1절에서 각각 설계한 알고리즘을 바탕으로 구현된 소프트웨어의 일부분을 다음 [그림 8], [그림 9]와 같이 나타냈다. 본 논문에서 구현하고자 한 주요 부분을 보여주고 있으며 파일로 저장된 DNA 서열에 대한 텍스트 파일을 바탕으로 유효한 ORF를 찾고 각각의 ORF에 대한 의미를 분석할 수 있도록 DNA DB 웹사이트 연결 부분을 자동 처리함으로써 사용자에게 편리한 분석 환경을 제공할 수 있다.

다음 [그림 8]은 본 논문에서 구현한 소프트웨어 시스템의 일부분으로 본 시스템을 통해 찾아진 ORF 개수와 검색된 부분을 한눈에 볼 수 있도록 각 ORF 부분을 간단한 그래픽으로 표현하여 간단히 계속되어야 하는 과정을 제어할 수 있도록 한 부분이다.

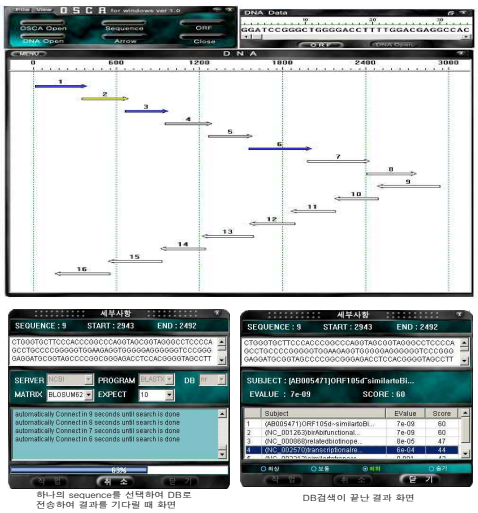

그림 8. ORF 검색 결과

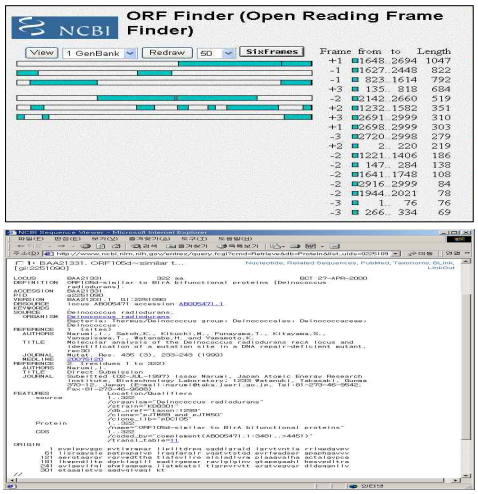

그림 9. ORF 분석 데이터

[그림 9]는 각각의 ORF 값의 의미를 NCBI를 통해 분석된 값을 보여주는 부분이다. 직접 NCBI 사이트를 통해 값을 복사하고 정해진 텍스트 필드에 붙여넣기를 한 후 필요한 데이터베이스 연결하는 등의 과정을 생략하고 [그림 8]의 각 ORF 값에 해당하는 그림을 선택함으로써 이와 같은 절차가 자동 실행될 수 있도록 구현하였다.

Ⅳ. 성능평가 및 기존 시스템 비교

4.1 기존 시스템과의 기능 비교

다음 [표 1]는 바이오 데이터 분석에 필요한 여러 가지 기능을 제공하는 분석 시스템인 NCBI, GeneNet, GeneWeb Ⅱ 그리고 본 논문에서 설계 및 구현한 분석 시스템간의 비교를 나타낸 결과이다.

시스템 비교에 이용된 기능들은 ORF 검색, mRNA 번역 기능, 서열 다중 정렬 기준 등 기본 바이오 데이터 분석에 필요한 사항과 검색 데이터의 보관 방법들을 비교할 수 있는 항목들을 추가하였다.

표 1. 분석시스템 간의 비교

분석시스템 / 테스트종류	NCBI 분석도구	Gene Net	GeneWeb II	본 논문
ORF Finder	○	×	×	○
mRNA 번역 및 역번역	○	×	×	○
서열 다중 정렬	○	×	○	×
기능 통합 환경 제공	△	×	○	○
작업 내용 저장 기능	×	×	×	○
데이터 편집 기능	×	×	○	○
서열 유사성 비교 기능	○	○	×	△
유사성 예측 기능	○	×	×	○
자체 데이터베이스 제공	○	×	○	○
다중 데이터베이스 검색	○	○	○	○
ORF 결과 제공 방법	이미지 제공	–	–	이미지 제공
유성성 비교 결과 제공	Web E-mail	E-mail	JAVA Applet	Web Application.
분석도구 제공환경	Web	Web	JAVA Applet	Web Application

시스템 비교한 표에서 볼 수 있는 결과를 기반으로 하여, 앞서 기존 시스템의 문제점 중에서 데이터 분석 과정의 연속성과 수동적인 작업에 따른 오류발생률 증가를 개선하기 위해 각 분석 과정을 연계시키고 기억시킬 수 있는 등의 특징을 파악하는데 기준으로 이용될

수 있다.

NCBI에서 얻은 결과화면은 GeneNet와 본 논문에서 설계한 시스템 모두 기본 5개로 설정되었으며 본 논문의 경우 보여주는 결과 리스트의 수를 사용자가 결정할 수 있으며 설정하지 않은 경우 기본값인 5개로 설정된다.

다음 [표 2]에서는 이들 분석 기능 중에서 가장 먼저 수행되는 ORF 검색 결과에 대한 부분만을 비교하였으며 비교 대상은 연구자들이 가장 많이 이용하고 있는 미국의 NCBI 센터에서 제공하는 ORF Finder의 결과와 본 논문의 분석 도구 결과를 나타낸 것이다.

표 2. NCBI와 본 시스템의 기능 비교

분석 도구 / 테스트 종류	NCBI 분석도구	본 논문의 분석도구
전체 찾아진 ORF 수	17	16
이용가능한 ORF 수	6	6
ORF의 이미지 제공	○	○
작업 저장 기능	×	○
	작업 결과의 저장기능의 존재하지 않아 작업 종료후 작업 재 개시 시점에서는 이전의 작업 결과를 알고자 할 경우 다시 DNA 데이터 입력부터 시작해야 하므로 연구자가 작업한 과정들을 재 반복해야 하는 불편을 겪어야 한다.	현재까지의 작업환경을 그대로 메모리에 저장할 수 있으므로 제공되므로 작업 종료 후 작업 재 개시 시점에서는 이전의 작업의 결과는 반복 실행 없이 볼 수 있으며 이전 작업을 이어 계속 연구할 수 있도록 제공한다.
데이터 편집 기능	×	○
사용 가능 DB	NCBI 제공 DB	NCBI 제공 DB, SWISSPROT, PROSITE, 본 논문에서 구축한 Local DB

위의 분석 결과에서와 같이 실제 전체 바이오 데이터에 대한 텍스트 파일 중 찾아진 유용한 의미의 가능성이 있는 ORF 개수나 이용 가능한 ORF 개수는 크게 차이가 없었으나 작업하는 과정에서 발생하는 데이터 보관이나 편집 기능들의 제공 유무에서 연구자의 편의성을 보다 개선했다는 것을 볼 수 있다.

4.2 알고리즘 성능평가

성능평가를 위하여 본 시스템의 분석 요소 중 ORF검색 모듈과 유전자 탐색 모듈들에 대한 성능평가를 실시한다.

ORF 검색을 위한 성능평가 기준은 유전체로부터 6개의 리딩프레임을 읽고 각 프레임에 대하여 단백질 변환 가능한 블록을 찾는 것으로 이는 찾고자 하는 기준에 따라서 결과가 달라질 수 있다. 또한 유전자 탐색에 대한 성능평가 기준은 유전자 탐색 정확성을 평가하기 위하여 예측된 엑손을 비교한다. 이러한 비교로부터 염기 레벨과 엑손 레벨의 정확성 측정을 계산한다. 다음은 염기 레벨의 정확성 측정을 위하여 사용될 파라미터이며 코딩으로 정확하게 예측되는 코딩된 염기의 비율인인 정확한 코딩으로 예측되는 염기의 예측 비율을 특이민감도(sensitivity:S_n), 특이성(specificity:S_p), 민감성과 특이성 사이의 상호관계율(Correlation Coefficient:CC)를 등을 이용하여 측정할 수 있다.

$$S_n = \frac{TP}{TP+FN} \quad \text{(식 1)}$$

$$S_p = \frac{TP}{TP+FP} \quad \text{(식 2)}$$

$$CC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}} \quad \text{(식 3)}$$

$$AC = (ACP-0.5)*2 \quad \text{(식 4)}$$

[표 3]과 [그림 10]은 기존 이용되고 있는 대표적인 분석 시스템 6가지와 본 논문에서 구현된 분석 시스템을 바탕으로 염기 레벨의 정확성을 비교한 결과를 나타내고 있다. 같은 서열수를 기반으로 하여 S_n, S_p 그리고 CC등의 수치화 된 값을 비교해 볼 수 있다.

표 3. 염기 레벨의 정확성 비교

시스템	서열수	염기 레벨 정확성			
		S_n	S_p	AC	CC
FGENES	195	0.86	0.88	0.84±0.19	0.83
GenMark	195	0.87	0.89	0.84±0.18	0.83
Genie	195	0.91	0.90	0.89±0.16	0.88
Genscan	195	0.95	0.94	0.91±0.12	0.91
Hmmgene	195	0.93	0.93	0.91±0.13	0.91
Paper	195	0.96	0.95	0.92±0.13	0.92

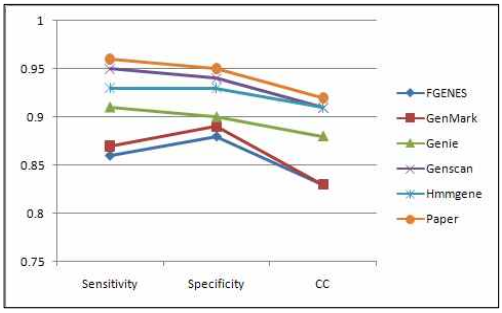

그림 10. 염기 레벨의 정확성 비교

표 4. 엑손 레벨의 정확성 비교

시스템	엑손 레벨의 정확성							
	ESn	ESp	(Esn+Esp)	ME	WE	PCa	PCp	OL
1	0.67	0.67	0.67±0.32	0.12	0.09	0.20	0.17	0.12
2	0.53	0.54	0.67±0.32	0.13	0.11	0.29	0.27	0.09
3	0.71	0.70	0.67±0.32	0.19	0.11	0.15	0.15	0.02
4	0.70	0.70	0.67±0.32	0.08	0.09	0.21	0.19	0.02
5	0.76	0.77	0.67±0.32	0.12	0.07	0.14	0.14	0.02
6	0.74	0.77	0.67±0.32	0.09	0.08	0.2	0.16	0.01

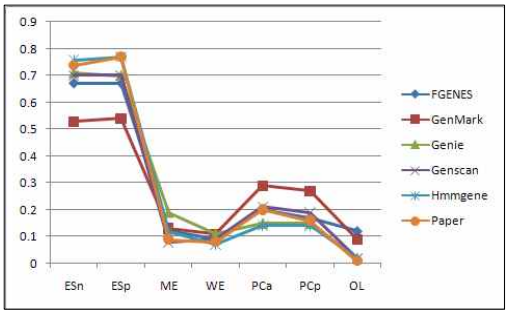

그림 11. 엑손 레벨의 정확성 비교

다음으로 [표 4]와 [그림 11]에서는 비교한 6개의 분석시스템과 본 논문의 분석 시스템간의 엑손 레벨의 정확성을 비교한 값과 그래프를 나타내고 있다.

V. 결론

본 논문에서 바이오 인포메틱스 분야의 많은 전산학적 접근 방법이 필요한 요소들을 분석하고 전산학적인 알고리즘들을 적용하여 바이오 데이터를 분석하고 있는데, 이를 바탕으로 현재 제공되고 있는 분석방법들과 데이터베이스들을 수용하여 바이오 데이터 분석을 하기 위한 시스템을 설계하였다.

분석 시스템에 포함된 데이터 분석 요소로는 기본적인 작업과정인 DNA전사, mRNA 번역, ORF 검색등과 이러한 작업과정을 통해 얻어진 결과들을 이용하여 데이터베이스로부터 서열 유사성 비교를 하고 전체 유전체로부터 유용한 유전자 탐색 등의 과정이 포함되었다.

기존 시스템의 개선점으로 제시하였던 분석 작업의 연속성, 수동적인 입력 과정의 과다로 인한 오류 발생률이 증가할 수 밖에 없었던 사항에 대해서는 본 시스템을 통해 크게 개선되었고, 해당 웹 서버의 부하 과중이나 네트워크 트래픽 증가로 인한 문제점을 해결할 수 있는 방법이 모색되었다. 또한 염기 서열 분석이나 엑손 분석과정에서 아주 작은 차이이지만 본 논문의 알고리즘을 통해 찾아진 정확성이 기존의 시스템보다 우수함을 보일 수 있었다. 따라서 본 시스템의 이용은 현재 바이오 데이터를 활용한 분야에서 초기 데이터 검색 및 분석하는 자동화 도구로 활용될 수 있을 것으로 전망해 본다. 이에 따라 많은 분야에서 연구자가 직접 유용한 데이터를 찾아야했던 작업을 대신할 수 있을 것이다.

그러나 여전히 다양한 분야에서의 연구 방법을 고려한다면 앞으로 추가될 항목이나 기존 방식에서의 개선점을 더 찾아 볼 수 있을 것이다. 예를 들어 서열의 개수가 많아질수록 서열 비교 분석 처리에 걸리는 시간이 기하급수적으로 증가하기 때문에 여전히 어느 분석 시스템을 이용하여도 어느 정도 불편함을 감소해야하는 부분이 남아있다. 이와같은 추가될 항목이나 개선되어야 할 사항들에 대해 앞으로 계속적인 연구가 이뤄져야 한다.

참 고 문 헌

[1] 송영옥, 김성영, 장덕진, “바이오 데이터 분석을 위한 시스템 및 알고리즘 설계”, 한국콘텐츠학회 논문지 10권, 2010(8).

[2] Cynthia Gibas and Per Jambeck, Developing

Bioinformatics Computer Skills, 2001.
[3] Des Higging and Willie Taylor, Bioinformatics: Sequence, structure and databanks, Oxford University Press, 2000.
[4] Andreas D. Baxevanis and B. F. Francis Ouellette, Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins, 2000.
[5] James Tisdall, Beginning Perl for Bioinformatics, 2001
[6] http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html
[7] Andreas D. Baxevanis and B. F. Francis Ouleeltte, Bioinformatics : A paractical Guide to the Analysis of Genes and Proteins, WILEY-INTERSCIENCE, 2001.
[8] http://www.genome.ad.jp/dbget/db_growth.html,
[9] Z. Galil and K. Park, Dynamic Programming with convexity, concavity and sparsity, theoretical computer science, Vol.92, No.49, 1992.
[10] Richard Neapolitan, Kumarss Naimipour, Jones and Bartlett Publishers, "Foundations of Algorithms using C++ pseudocode," 2nd,

저 자 소 개

송 영 옥(Young-Ohk Song) 종신회원

- 2003년 2월 : 충북대학교 컴퓨터공학과 공학박사 졸업
- 2000년 3월 ~ 2002년 2월 : 대전보건대학 멀티미디어학과 겸임교수
- 2002년 3월 ~ 2010년 : 우송대학교 컴퓨터정보학과 초빙교수
- 2008년 10월 ~ 현재 : AQUARITAS, INC. USA, CTO

<관심분야> : Software Engineering, Data Mining, Embedded System, SmartPhone App, Bioinformatics

장 덕 진(Duk-Jin Chang) 정회원

- 1983년 ~ 1986년 : 미국 텍사스 A&M 대학교 박사수료
- 1984년 ~ 1990년 : 텍사스 교통연구원 연구조원
- 1990년 ~ 1995 : 시스템공학연구소 선임연구원
- 1995 ~ 현재 : 우송대학교 컴퓨터정보학과 교수

<관심분야> : Software Engineeirng, Project Management, Intelligent Transportation System, IT&BT