제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

복수 서열 정렬을 위한 시스템 개발에 관한 연구

김동회,김진

한림대학교 컴퓨터공학과

e-mail: dhkim@center.cie.hallym.ac.kr
jinkim@hallym.ac.kr

A study of system development for multiple sequence alignment

Dong-Hoi Kim, Jin Kim
Dept of Computer Science, Hallym University

요 약

유전체 서열결정이 폭발적으로 증가해 가고 있다. 인간 유전체사업(Human genome project)의 궁극적인 목적은 인간 염색체에 있는 30억개의 뉴클레오티드와 10만개의 유전자를 밝혀내는 것이고 생의학에서 새로운 발견이나 응용을 위한 정보로 이용하는 것이다. 이 사업은 1980년대 후반에 시작되었고 현재 서열의 결정이 완료된 상태이다. 본 논문에서는 인간 유전체 사업에서 파생된 가장 중요한 문제 중의 하나인 복수 염기서열 정렬 문제와 복수 염기서열 정렬 시스템의 구현에 대하여 논한다.

1. 서론

DNA,RNA 및 단백질들은 특정한 생물학적 기능을 수행하도록 잘 정렬된 선형 중합체(linear polyner)이다. 생물학적 고분자를 형성하기 위해서 선형적으로 정렬되어 있는 뉴클레오티드와 아미노산의 서열 정보는 실험에 의해 쉽게 얻어진다. 따라서 컴퓨터를 이용한 서열분석의 목적은 뉴클레오티드 또는 아미노산 서열로 쓰여진 고차원의 구조와 기능적인 정보를 밝혀내기 위한 것이다. 두 분자가 유사한 서열을 가지면, 진화적인 관계나 물리 화학적인 제약 때문에 유사한 3차원 구조와 비슷한 생물학적 기능을 갖기 쉽다. 따라서 서열 분석의 주요 작업은 구조적 기능적 속성으로 확장 될 수 있는 서열특징을 찾는 것이다.염기 서열 정렬 기법에는 2개의 핵산-핵산 또는 단백질-단백질을 대상으로하여 주로 상동성(homology)을 나타내기 위해 사용되는 쌍정렬(pairwise alignment)과 3개 이상의 핵산 또는 단백질 서열을 하나의 정렬로 나타내는 다중정렬(multiple alignment)이 있다. 쌍정렬은 동적 프로그래밍(dynamic programming) 기법을 사용하여 최적의 해를 찾을 수 있다.두 개의 서열만을 비교할 때 서열간의 유사성이 낮게 나왔더라고, 여러개의 서열을 모아

동시에 비교하면 높은 유사성으로 바뀔 수 있다.다중 서열정렬은 서열들의 그룹을 동시에 비교하는 것이며, 기능적/구조적 중요성을 가지는 지역적으로 보존된 영역을 식별할 때 매우 유용하다. 이는 패밀리 분석, 계통관계분석, 도메인분석 등의 기능분석연구에 매우 다양하게 사용된다. DNA 염기 서열의 결정이 급속히 진전되면서 같은 그룹 내에 여러 개의 핵산이나 단백질 서열에 대한 비교가 필요하게 되었고, 단백질 패밀리나 관련 핵산의 모티프(motif)서열 등에 대한 자료가 증가 하고 기능 분석 연구가 활발해 지고 있으므로, 효율적인 복수 염기서열 정렬의 필요성이 지속적으로 제기되고 있다. 현재 대부분의 서열정렬 프로그램은 리눅스 환경에서 개발되어 있으며 따라서 실제 시스템의 사용자인 생물학자에게 있어서 리눅스 시스템에 대한 지식을 요구하므로 사용에 제한이 있으며, 이들 프로그램은 효율적인 분석을 위한 시각화 부분을 제공하고 있지 않고 서열정렬에만 그치고 있기 때문에 생물학자들은 이 정렬된 서열을 가지고 분석을 위한 또 다른 시도를 해야만 한다. 본 논문에서는 윈도우 환경에서 Comalign 프로그램을 이용한 서열정렬시스템을 개발하였다.

2. 복수 서열정렬 시스템의 개요

본 논문에 구현된 복수 서열정렬 시스템은 서열정렬만이 아닌 생물학자들이 서열의 정렬된 서열을 쉽게 분석할 수 있도록 4부분의 시각화 부분을 제공한다. 각각은 서열의 정보를 나타내는 서열 리스트뷰와 서열 사이의 유사도에 따른 진화관계를 나타내는 진화트리뷰, 서열간의 지역적 유사도를 볼 수 있는 유사도 그래프뷰, 서열의 부분 편집과 정렬된 서열, 서열내의 특정 문자열의 위치를 볼 수 있는 서열뷰 부분으로 구성하여 서열정보를 효율적으로 분석할 수 있도록 한다.

3. 구현환경

본 복수 서열정렬 시스템은 프로그램의 수행 속도와 확장성을 고려하여 Visual C++을 사용하여 윈도우즈 운영체제 환경에서 사용할 수 있도록 개발하였다.

4. 기능

이 시스템의 주요 구성은 서열의 정보를 나타내는 서열 리스트뷰와 서열 사이의 진화관계를 나타내는 진화관계를 나타내는 진화트리뷰, 서열간의 지역적 유사도를 볼수 있는 유사도 그래프뷰, 서열의 부분 편집과 정렬된 서열, 서열내의 특정 문자열의 위치를 볼 수 있는 서열뷰 부분으로 구성 되어있다.. 그림 1은 본 논문에서 개발한 복수 서열정렬 시스템의 화면이다.

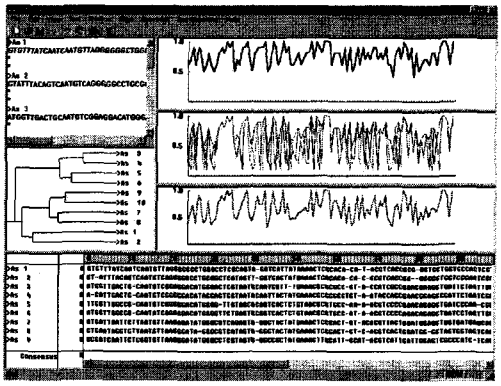

그림 1 복수 서열정렬 시스템

4.1 서열 리스트

그림 2은 서열 리스트 부분으로 각 서열들에 대하여 정렬이전의 서열을 나타낸다.본 시스템에서는 대표적인 서열포맷인 fesaA형식을 취하고 있다.

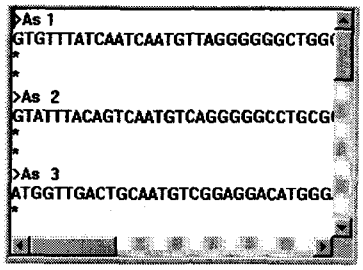

그림 2 서열 정보

4.2 유사도 그래프

그림 3은 유사도를 그래프로 나타내는 부분이다. 각각은 전체 서열간의 유사도 부분,전체에 대한 특정 서열에 대한 유사도그래프, 특정 두 서열간의 유사도를 나타내는 유사도 그래프를 나타내며 그래프의 특정부분을 확대, 축소 할 수 있다. 사용자는 이 유사도 그래프부분을 통하여 특정 부분에 대한 유사도 정도를 파악할 수 있으며, 이를 통해 서열간의 보전적인 지역을 보다 쉽게 파악할수 있다.

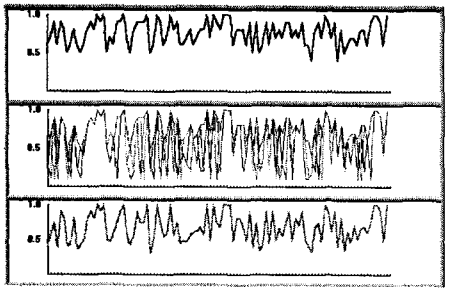

그림 3 유사도 그래프

4.3 진화관계 트리

진화트리뷰는 각 서열들의 진화관계를 나타내는 부분으로 복수염기서열 정렬의 주요 과제중 하나이다. 진화트리를 재구성하는데 사용되는 방법에는 거리행렬을 이용한 점급 방법과 절약 접근방법이 있는데 본 시스템에서는 거리행렬을 사용하여 가장 짧은 거리가 되는 서열들을 결집하는 UPGMA(*unweighted pair-group method by arithmetric* averaging)을 사용하였다. 그림 4는 본 논문에서 구현한 UPGMA를 이용한 진화트리이다. 각 서열에 대한 진화트리를 구성함으로서 각 서열간의 진화관계를 쉽게 파악할 수 있다.

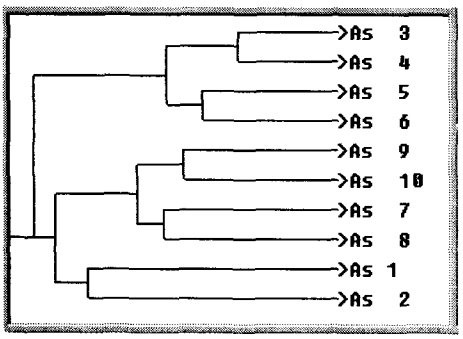

그림 4 진화트리

4.4 정렬 서열

그림 5는 정렬 서열부분으로 각 서열을 정렬된 텍스트와 서열내의 문자열의 위치를 나타내고 서열의 일부분에 대해 유사도 그래프의 위치를 파악할수 있도록 되어 있다. 이 서열 부분에서 특정위치의 서열을 부분 편집할수 있다. 그림 6은 서열의 편집 화면이다.

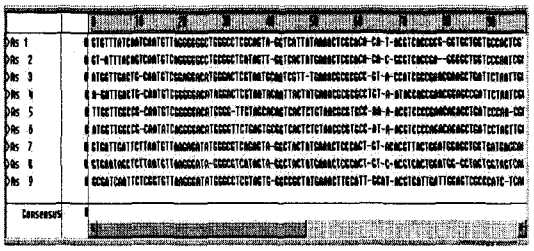

그림 5 정렬 서열

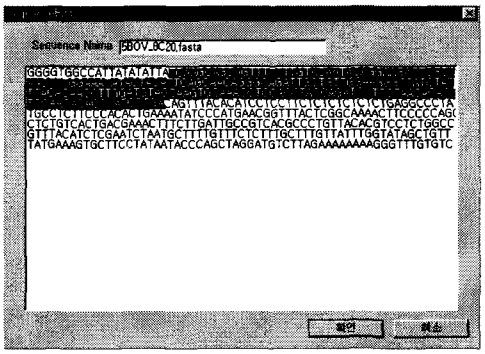

그림 6 서열 부분편집

6. 결론 및 향후연구

본 논문에서는 Comalign 알고리즘을 이용한 복수 염기서열 정렬을 위한 시스템 개발에 대하여 논하였다. 본 시스템은 생물학에서의 가장 주된 관심사인 염기서열간의 정렬을 윈도우 환경에서 구현함으로서 별도의 전산지식 없이 쉽게 서열간의 진화관계 및 유사성을 쉽게 파악할 수 있는 인터페이스를 제공한다는 장점이 있다. 본 복수염기서열 정렬 시스템을 사용함으로서 각 서열간의 전체 혹은 지역적 유사성을 쉽게 판단할 수 있으며,진화트리를 통해 서열간의 진화관계를 알 수 있다. 또한 서열의 특정 스트링의 위치를 쉽게 찾을 수 있으며 유사도 그래프를 통해 서열간의 지역적 유사도를 측정할 수 있다. 본 논문에서 구현된 복수 서열정렬 시스템은 현재 Comalign 알고리즘을 통해 서열정렬을 하고 있으나 이 Comalign도 휴리스틱 알고리즘이기 때문에 항상 다른 알고리즘보다 좋은결과를 가진다고는 볼 수 없다 따라서 향후 서열정렬에 대하여 여러 가지 알고리즘을 선택적으로 적용할수 있도록 하고자 한다. 또한 입력 데이터 형식을 fastA형식을 사용하고 있으나 향후 모든 서열데이터 형식을 취하고자 한다. 현재 정렬하고자 하는 서열에 대하여 파일처리

로 사용하고 있으나 차후 웹과의 연동으로 Genbank 와 같은 유전자 데이터베이스 서버의 유전자 서열을 입력 자료로 사용할 수 있도록 수정 보완하고자 한다.

참고문헌

[1] Biological sequence analysis , Cambridge univ press

[2] Genbank http://www.ncbi.nlm.nih .gov 1998

[3] Klaus Bucka-Lassen ,Ole Caprani and Jotun Hein. "Combining many multiple alignment in one imroved alignment."

[4] Bacon, D,G and W. F Anderson. "Multiple sequence alignment". J.Mol.Biol 1986