

다중서열정렬에 기반한 종의 차이

권혁주* · 김상진** · 김근무***

Differences between Species Based on Multiple Sequence Alignment Analysis

Hyeok-Zu Kwon* · Sang-Jin Kim** · Geun-Mu Kim***

요 약

다중서열정렬(MSA : multiple sequence alignment)은 다양한 생명체에서 같은 기능을 하는 여러 개의 단백질 서열이나 핵산 서열을 한 번에 모아서 서로 정렬하는 방법이다. 바이오파이썬을 이용하여 인간이 다른 동물과 어떻게 다른지 조사하였다. 대표적인 다중서열정렬 알고리즘인 clustalW는 열의 위치별로 정렬된 정도를 비교한다. 또한, 웹로고와 계통수를 만들어서 보존서열을 가시화하여 이해도를 향상한다. 인간과 다른 동물의 차이점을 확인하는 예를 제시하고 바이오파이썬을 활용도를 제시한다.

ABSTRACT

Multiple sequence alignment (MSA) is a method of collecting and aligning multiple protein sequences or nucleic acid sequences that perform the same function in various organisms at once. clustalW, a representative multiple sequence alignment algorithm using BioPython, compares the degree of alignment by column position. In addition, a web logo and phylogenetic tree are created to visualize conserved sequences in order to improve understanding. An example was given to confirm the differences between humans and other species, and applications of BioPython are presented.

키워드

Python, Biopython, Library, Multiple Sequence Alignment, Phylogenetic Tree
파이썬, 바이오파이썬, 라이브러리, 다중서열정렬, 계통수

1. 서 론

현대인들의 컴퓨터 프로그래밍에 관한 높은 관심과 클라우드 컴퓨팅과 같은 연산 속도의 향상으로 다양한 분야의 정보 분석이 현실화하고 있다. 초등학생도 코딩을 배우고 인문학 분야에서도 프로그래밍의 활용도가 확장되고 있다. 인터프리터 프로그래밍 언어로 자연어(인간의 언어)에 가까운 프로그래밍 언어 중의

하나가 파이썬(Python)이다[1]. 파이썬은 매우 간결하고 어디서나 다운로드하여 사용할 수 있는 오픈 소스이다. 특정 목적을 해결하기 위해 사용자가 호출하여 사용할 수 있는 다양한 기능의 묶음을 규모에 따라 함수, 모듈, 패키지, 라이브러리라고 한다. 바이오파이썬(Biopython)은 파이썬의 라이브러리 중 하나로 생물정보학 프로그래밍을 위한 매우 유용한 라이브러리이다[2].

* 부경대학교 해양기상정보센터(hzuk0706@naver.com)

** 교신저자 : 부경대학교 지구환경시스템과학부(sjgreat373@gmail.com)

*** 부경대학교 지구환경시스템과학부(riarmsan1@nate.com)

• 접수일 : 2024. 02. 29

• 수정완료일 : 2024. 03. 21

• 게재확정일 : 2024. 04. 12

• Received : Feb. 29, 2024, Revised : Mar. 21, 2024, Accepted : Feb. 12, 2024

• Corresponding Author : Division of Earth and Environmental System Sciences, Pukyong National University,

Email : sjgreat373@gmail.com

생물정보학이 오늘날 생물학과 의학계의 혁신을 이끄는 핵심 분야이다. 생물정보학은 컴퓨터공학의 방법론을 이용해 DNA (: Deoxyribo Nucleic Acid) 서열 등 주요 생명 정보를 취득·저장하고 분석해 활용한다 [3-6]. ‘인간 게놈 프로젝트’가 1990년에 시작되어 2003년에 마무리되는 사이에 유전 정보를 취득하고 분석하는 기술이 급속도로 발전했다. 정보기술(IT : Information Technology)이 빠르게 발전하면서 유전 정보 분석 작업에 속도가 붙었고, 이렇게 얻은 유전체 정보는 다시 생물학과 의학의 발달을 가져왔다. 생명정보학은 생명 활동의 근원인 단백질의 구조와 역할을 밝히고, 개인 유전자 분석을 통한 맞춤형 정밀의료와 효율적인 신약 후보 물질 발굴 등을 가능케 하며 질병 치료의 새로운 길을 열어가고 있다.

본 연구의 목적은 맞춤형 헬스케어에 필요한 정보를 분석하고 생산하기 위한 바이오파이썬의 기본적인 사용법을 조사하고 활용 사례를 실습하는 것이다. 연구 방법에서 바이오파이썬의 구조를 설명하고 사용 방법을 조사한다. 결과에서는 생명체의 보존서열과 계통수를 산출하여 그 차이를 분석한다. 결론에서는 바이오파이썬의 유용성을 논의하고 향후 응용할 수 있는 주제를 제시한다.

II. 연구 방법

2.1 바이오파이썬

바이오파이썬은 생물학자들이 컴퓨터로 작업을 할 때 도움이 될 만한 파이썬의 라이브러리 모음이다. 바이오파이썬은 다양한 파일 형식으로부터 원하는 정보를 추출하고 웹 기반 프로그램을 처리하는 자동화로 연구 효율성을 높여준다. 바이오파이썬으로 할 수 있는 주요 작업은 첫째, 생물정보 파일에서 원하는 정보를 가져오는 파싱(parsing), 둘째, 유전체 서열정보의 문자열화, 셋째, 웹에 있는 정보를 브라우저를 통하지 않고 프로그래밍으로 수집, 그리고 넷째, 웹 기반의 프로그램을 사용하지 않고 생물정보학에서 사용하는 툴의 활용이다.

의학 및 생물학 분야에서는 주로 PubMed를 사용하여 네트워크로부터 정보를 수집한다. PubMed는 MEDLINE, 생명과학 저널 및 온라인 서적의 생의학

문헌에 대한 3,600만 건 이상의 인용으로 구성되어 있다. 바이오파이썬을 활용하여 PubMed 정보를 검색하고 파싱할 수 있다.

BLAST (: Basic Local Alignment Search Tool)는 염기서열 또는 단백질서열을 입력받아 데이터베이스를 검색하여 유사한 서열을 찾아 정보를 제공하는 tool이다. BLST의 결과를 XML 파일로 저장할 수 있으며 XML을 NCBIWWW 모듈로 파싱한다.

2.2 파일 포맷

필요한 정보를 얻기 위해서는 파일의 포맷을 이해해야 한다. FASTP 프로그램에 기반한 FASTA는 텍스트 기반 포맷으로 염기서열과 단백질 정보를 제공한다. 문자 ‘>’로 시작하는 첫째 줄의 헤더와 두 번째 줄부터 한 줄에 120자 이하의 서열이 이어진다. 한 종류의 서열만 있는 단일 FASTA뿐만 아니라 여러 종류의 서열로 구성된 멀티 FASTA 파일도 있다.

FASTQ는 텍스트 기반 포맷으로 염기서열과 quality 점수를 포함한다. 네 줄로 구성되며 네 줄이 하나의 read를 이루는데, read는 시퀀서가 샘플의 서열을 한 번에 읽는 길이이다. 첫 번째 줄은 @로 시작하는 헤더로 염기서열이 읽힌 시퀀싱에 대한 정보이다. 두 번째 줄은 시퀀서가 읽은 염기서열이다. 세 번째 줄은 구분 문자 + 기호이다. 네 번째 줄은 시퀀서가 읽은 염기서열의 품질정보이다, 염기서열의 정확도를 백분율로 산출할 수 있다.

SAM (: Suenice Alignment Map)은 텍스트 기반의 alignment of read이고, BAM (: Binary Alignment Map)는 SAM의 binary file이다. SAM/BAM는 헤더 @ 기호로 시작하는 헤더 부분과 염색체 정보가 나열되는 정렬 부분으로 나뉜다.

BED (: Browser Extensible Data) 포맷은 유전체를 구간별로 나누어 특징을 주석으로 표기한다. 각 항목은 탭으로 나뉜 텍스트이다. 필수적으로 3개의 열이 있고 추가로 열을 9개까지 확장할 수 있다.

VCF (: Variant Calling Format)는 variant를 만들기 위해 만든 포맷으로 메타데이터와 내용으로 나뉜다. 메타데이터는 #으로 시작하고, 2개의 #으로 시작하면 VCF에 대한 정보로 ‘key=value’ 관계로 표현한다. 내용 부분은 8개의 필수 열과 샘플에 따라 추가되는 열로 이루어진다. 각 열은 탭으로 나뉜다.

GenBank Flat File은 NCB (: National Center for Biotechnology Information)의 데이터 베이스에서 대중에게 제공하는 정보이다. 염기서열과 CBS (Coding Sequence)별로 번역된 아미노산 서열, 종의 정보, 관련 논문 저자, 제목, pumbed ID 등의 메타 정보를 담고 있다. 파일 내부에는 1개 이상의 GenBank 정보가 들어 있고, 끝맺음 구분은 2개의 슬래쉬 기호(//)로 마무리한다.

XML (: eXtensible Markup Language)은 구조화된 데이터를 표현하고 정보를 교환할 때 범용적으로 사용하는 언어로 태그를 이용하여 데이터 구조를 나타낸다. <태그>내용</태그>로 이루어진 텍스트 파일이고, JSON (: JavaScript Object Notation)은 구조화된 데이터를 표현하고 정보를 교환할 때 범용적으로 사용하는 파이 포맷이다. 기능과 목적이 XML과 많이 중복되지만 가독성이 떨어지는 XML 형식의 대안으로 사용된다. 대부분의 언어에서 parser를 지원하여 파일을 읽고 쓰기 쉽다. 파이썬의 dictionary와 형태가 같아서 'key-value' 쌍이 중괄호({})로 이루어진 구조이다.

롬프트 창에서 진행한다. 'muscle파일이름 -alignn (변환할 파일이름) -output (변환될 파일이름)'와 같은 순서로 입력하면, 정렬된 파일을 생성한다(그림 3).

| Entry | Entry Name |
|--|------------|
| <input checked="" type="checkbox"/> P01942 | HBA_MOUSE |
| <input checked="" type="checkbox"/> P01958 | HBA_HORSE |
| <input checked="" type="checkbox"/> P60529 | HBA_CANLF |
| <input checked="" type="checkbox"/> P69905 | HBA_HUMAN |
| <input checked="" type="checkbox"/> P18971 | HBA_BALAC |
| <input checked="" type="checkbox"/> P69907 | HBA_PANTR |
| <input checked="" type="checkbox"/> P01994 | HBA_CHICK |
| <input checked="" type="checkbox"/> P01948 | HBA_RABIT |

그림 1. www.uniprot.org에서 동물 8종 선택.

Fig. 1 Selection of 8 species at www.uniprot.org.

III. 연구 결과

3.1 다중서열정렬

인간과 다른 동물의 차이점을 파악하기 위해 다중서열정렬을 이용하여 웹로그와 계통수로 표현한다. 다중서열정렬은 세 단계로 준비한다. 우선 비교할 서열이 필요하므로 인간뿐만 아니라 다른 동물로부터 얻을 수 있는 정보를 선택한다. 체내 산소를 운반하는 단백질인 헤모글로빈의 사합체 중 하나인 HBA 서열을 단백질 정보 데이터베이스에서 받는다(그림 1)[7]. 단백질찾기에서 HBA를 입력하여 검색하고, 그림 1과 같이 인간을 비롯한 7종의 동물, 쥐(MOUSE), 말(HORSE), 개(CANLF), 인간(HUMAN), 멩크고래(BALAC), 침팬지(PANTR), 닭(CHICK), 그리고 토끼(RABIT)를 선택한다. FASTA 양식의 파일로 내려받으면 그림 2와 같이 정리되지 않은 자료를 확인할 수 있다.

다음 단계는 다중서열정렬 과정을 진행하는 것으로 MUSCLE이나 Clustal Omega 툴을 사용한다[8]. MUSCLE은 그래픽 인터페이스가 없으므로 명령 프

```

# example.fasta
1 >sp|P01942|HBA_MOUSE Hemoglobin subunit alpha OS=Mus musculus OX=10090 GN=Hba PE=1 SV=2
2 MVLSEEDKSNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
3 KKVADALASAGHLDLPGALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
4 AVHASLDFKFLASVSTLTSKYR
5 >sp|P01958|HBA_HORSE Hemoglobin subunit alpha OS=Equus caballus OX=9796 GN=HBA PE=1 SV=2
6 MVLSAADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
7 KKVGDALTLAVGHLDLPGALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
8 AVHASLDFKFLASVSTLTSKYR
9 >sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens OX=9606 GN=HBA1 PE=1 SV=2
10 MVLSPADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
11 KKVADALTNVAHVDPPHIALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
12 AVHASLDFKFLASVSTLTSKYR
13 >sp|P69907|HBA_PANTR Hemoglobin subunit alpha OS=Pan troglodytes OX=9598 GN=HBA1 PE=1 SV=2
14 MVLSPADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
15 KKVADALTNVAHVDPPHIALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
16 AVHASLDFKFLASVSTLTSKYR
17 >sp|P01948|HBA_RABIT Hemoglobin subunit alpha-1/2 OS=Oryctolagus cuniculus OX=9986 PE=1 SV=2
18 MVLSPADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
19 KKVSEALTKAVGHLDLPGALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
20 AVHASLDFKFLASVSTLTSKYR
21 >sp|P01994|HBA_CHICK Hemoglobin subunit alpha-A OS=Gallus gallus OX=9031 GN=HBA PE=1 SV=2
22 MVLSAADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
23 KKVAALEAHNHDIDTGLSKLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAAFTP
24 EVHASLDFKFLASVSTLTSKYR
25 >sp|P18971|HBA_BALAC Hemoglobin subunit alpha OS=Balaenoptera acutorostrata OX=9767 GN=HBA PE=1 SV=2
26 MVLSPDTSNKKVATKAGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
27 KKVADALTKAVGHLDLPGALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAEFTP
28 AVHASLDFKFLASVSTLTSKYR
29 >sp|P60529|HBA_CANLF Hemoglobin subunit alpha OS=Canis lupus familiaris OX=9615 GN=HBA PE=1 SV=1
30 MVLSPADKTNKAAAMKGGHGAEGYGAELERMFSPPTTKTYFPHDLSHSAQKWKHG
31 KKVADALTNVAHLDLPGALSLSLHAKLRLDPAVNFKLLSHCLLVLAHLPAAFTP
32 AVHASLDFKFLASVSTLTSKYR
    
```

그림 2. 8종에 관한 FASTA 양식의 HBA 서열.
Fig. 2 Sequences of 8 species by FASTA format.

```

1 >sp|P01994|HBA_CHECK Hemoglobin subunit alpha-A 05-Gallus gallus OX=9931 GN=HBA PE=1 SV=2
2 MVLSPADKTNVKAALWGVGAHAGEYGAELERMFSTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
3 LSKLSLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
4 >sp|P01942|HBA_MOUSE Hemoglobin subunit alpha 05-Mus musculus OX=10899 GN=Hba PE=1 SV=2
5 MVLSEDEKSNLKAEMKIGGHAGEYGAELERMFASFTTKTYFPHFDVSHGSAQVKGHKVVADALTNVAHVDMPNA
6 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
7 >sp|P69907|HBA_PANTR Hemoglobin subunit alpha 05-Pan troglodytes OX=9598 GN=HBA1 PE=1 SV=2
8 MVLSPADKTNVAAMKGVGAHAGEYGAELERMFSLPFTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
9 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
10 >sp|P69905|HBA_HUMAN Hemoglobin subunit alpha 05-Homo sapiens OX=9686 GN=HBA1 PE=1 SV=2
11 MVLSPADKTNVAAMKGVGAHAGEYGAELERMFSLPFTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
12 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
13 >sp|P01958|HBA_HORSE Hemoglobin subunit alpha 05-Equus caballus OX=9796 GN=HBA PE=1 SV=2
14 MVLSPADKTNVAAMKGVGAHAGEYGAELERMFSLPFTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
15 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
16 >sp|P01940|HBA_RABBIT Hemoglobin subunit alpha-1/2 05-Oryctolagus cuniculus OX=9986 PE=1 SV=2
17 MVLSPADKTNLKAEMKIGGHAGEYGAELERMFGLPFTTKTYFPHFDVSHGSAQVKGHKVVADALTNVAHVDMPNA
18 LSTLSLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
19 >sp|P18971|HBA_BALAC Hemoglobin subunit alpha 05-Balaenoptera acutorostrata OX=9767 GN=HBA PE=1 SV=2
20 MVLSPDKSNVKAEMKIGGHAGEYGAELERMFNFTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
21 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
22 >sp|P60529|HBA_CANLF Hemoglobin subunit alpha 05-Canis lupus familiaris OX=9615 GN=HBA PE=1 SV=1
23 -VLSADKTNLKAEMKIGGHAGEYGAELERMFSLPFTTKTYFPHFDLSHGSAQVKGHKVVADALTNVAHVDMPNA
24 LSALSLLHAKLRVDPVNFKLLSHCLLVTAALHPAEFTPAVHASLDFLAVSTVLSKYR
25
    
```

그림 3. MUSCLE로 정렬한 8종의 HBA 서열.
Fig. 3 Sequences of 8 species aligned by MUSCLE.

3.2 웹로고

웹로고(WebLogo)는 시퀀스 로고를 최대한 쉽게 생성할 수 있도록 설계된 웹 기반 응용 프로그램이다. 각 로고는 일련의 심볼로 구성되며, 시퀀스의 위치마다 하나씩 쌓인다. 쌓인 전체 높이는 해당 위치에서의 서열 보존을 나타내며, 각 심볼의 높이는 해당 위치에서의 아미노산 또는 핵산의 상대적인 빈도를 각각 나타낸다. 결국, 다중서열정렬은 여러 개의 서열을 각 위치에서 비교하여 정렬하는 것으로 각 위치의 서열이 어떻게 다른지 알아볼 수 있다.

그림 4는 인간의 서열을 나타내고, 그림 5는 인간을 포함한 8종의 서열을 동시에 나타낸다. 단일 알파벳의 크기가 클수록 공통된 서열을 나타낸다.

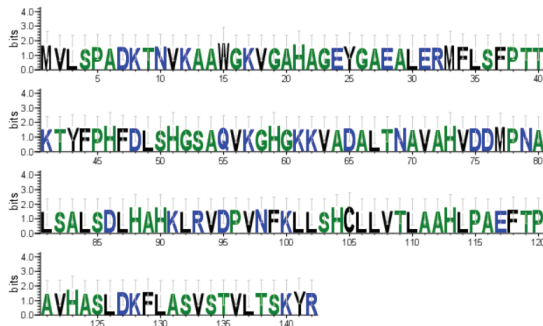


그림 4. 인간의 웹로고.
Fig. 4 WebLogo of human.

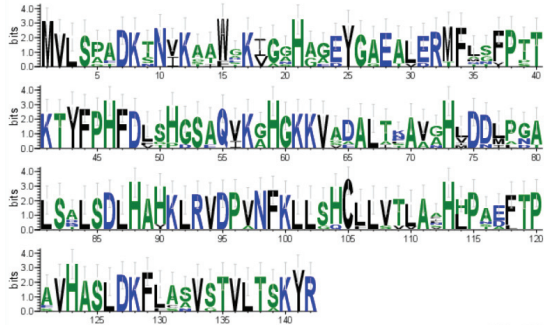


그림 5. 8종의 웹로고.
Fig. 5 WebLogo of 8 species.

인간이 쥐와 어느 서열에서 같고, 어느 서열에서 다른지를 확인하려면 두 서열만의 웹로고를 생성한다 (그림 6). 위치 5, 6, 9, 11, ..., 114, 117에서 인간과 쥐가 다르다는 것을 알 수 있다. 다른 종과도 마찬가지로 직접 비교할 수 있다.

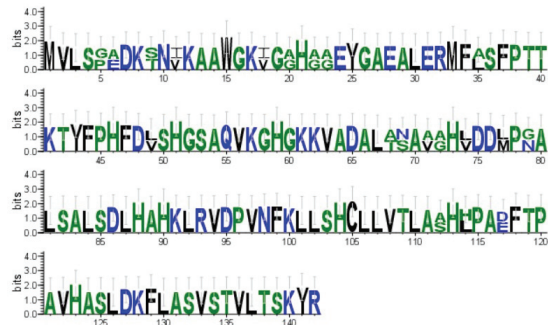


그림 6. 인간과 쥐의 웹로고.
Fig. 6 WebLogo of human and mouse.

3.2 계통수

계통수는 생물 사이의 관계의 멀고 가까운 정도의 유사도를 그림으로 나타낸다. 거리가 가까울수록 서열이 유사하고, 멀수록 서열의 차이가 크다. 생물학 분야의 전반적인 진화학적 연구 결과를 제시하고 다양한 생물학적 의미를 끌어내기 위해 사용된다. 좁게는 유전자 및 단백질의 진화부터 넓게는 생물 종 혹은 특정 분류군의 진화가 계통수로 표현된다. 계통수는 간단히 tree라고도 불리며, tree의 시작점을 계통수의 root, root에서 출발하여 뻗어 나가는 것을 branch, branch가 갈라지는 부분을 node, tree의 가장 끝부분을 tip 혹은 leaf라고 한다. 하나의 node 이후에 갈라져 나온 모든 계통을 하나로 묶어 clade라고 한다(그림 7). 계통수는 root 방향으로 갈수록 기원한 조상을 향하게 되며, tip으로 갈수록 후손을 향하여 가장 끝 tip에는 현재 관찰되는 분류군(A ~ E)이 위치한다. 이러한 구성요소와 기본 개념으로 진화 역사를 표현하면, tree의 형태는 다양하게 표현될 수 있다.

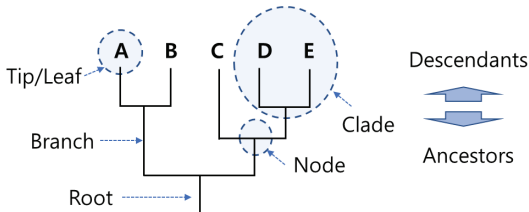


그림 7. 계통수의 기본구조.
Fig. 7 Basic structure of phylogenetic tree.

EMBL-EBI에서 제공하는 phylogenetic tree 포맷 생성 페이지의 입력란에 서열을 넣고 실행하면 그림 8의 계통수가 생성된다. 최근에 공통 조상을 공유하는 분류군일수록 더 가까운 계통 관계이므로 인간은 실험 대상 7종의 동물 중에서 침팬지와 가장 가까운 관계이다. 같은 이유로 쥐와 멧크고래도 가까운 관계로 이해된다.

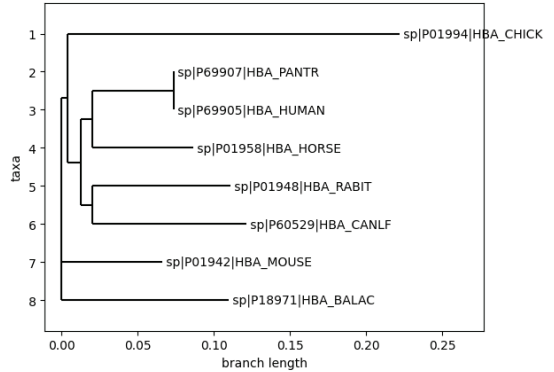


그림 8. 8종의 계통수.
Fig. 8 Phylogenetic tree of 8 species.

계통수에서 일부 tip을 가지치기하는 것은 계통 관계에 영향을 미치지 않으므로 tip끼리 가깝게 보인다고 가까운 관계로 판단할 수 없다[9,10]. 또한, node를 중심으로 회전하여 표현할 수 있으므로 단순히 가까운 위치로 서열의 유사성을 파악하지 않는다. 그림 8에서 말, 토끼 그리고 개를 가지치기하면 인간은 닭보다는 쥐나 멧크고래와 더 가깝다.

IV. 결론

3개 이상의 DNA 혹은 단백질의 염기서열을 정렬하는 다중서열정렬은 서열 사이의 진화 관계, 단백질의 구조와 기능에 관한 연구에 필수적인 도구이다. 파이썬 기반의 바이오파이썬을 이용하여 다중서열정렬을 분석하는 과정을 제시하였다. 가시화 방법으로 웹로그보다는 계통수가 종의 서열 이해에 효과적이라는 것을 보였다. 바이오파이썬은 원인 불명의 환자에서 나온 미지의 종을 찾는대거나 코로나 바이러스와 같은 전염병의 실체를 밝히는 유용한 도구로 활용된다.

References

- [1] L. Lubanovic and G. Choi, "Introducing Python", *Hanbit Media*, 2020, pp. 744.
- [2] J. Han, "Bioinformatics through BioPython", *BJ Public*, 2019, pp. 292.

- [3] G. Seok and S. Park, "Probabilistic model for bio-cells information extraction", *J. of the Korea Institute of Electronic Communication Sciences*, vol. 6, no. 5, 2011, pp. 649-656.
- [4] S. Lee and D. Suh, "A Novel Way of Diversifying Context Awareness Based on Limited Event Data of Sensors using Exon-Intron Theory in the Internet of Things Environment", *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 4, 2021, pp. 675-682.
- [5] E. Kim, J. Jeong, and B. Lee, "A Big Data Based Random Motif Frequency Method for Analyzing Human Proteins", *J. of the Korea Institute of Electronic Communication Sciences*, vol. 13, no. 6, 2018, pp. 1397-1404.
- [6] H. Ceong and C. Park, "Enzyme Metabolite Analysis Using Data Mining", *J. of the Korea Institute of Electronic Communication Sciences*, vol. 11, no. 10, 2016, pp. 969-982.
- [7] Find your protein, Uniport Consortium, 2024, <https://www.uniprot.org/>.
- [8] Clustal Omega, EMBL-EBI, 2023, <https://ebi.ac.uk/jdispatcher/msa/clustalo/>.
- [9] D. Baum and S. Smith, "Tree thinking: An Introduction to Phylogenetic Biology", *Roberts and Company Publishers*, 2013, pp. 496.
- [10] Understanding Evolution, University of California Museum of Paleontology, 2021, <https://evolution.berkeley.edu/>.

저자 소개



권혁주(Hyeok-Zu Kwon)

2024년 부산대학교 생명과학과
(이학사)
2024년~ 해양기상정보센터

※ 관심분야 : 생명과학, 생물학



김상진(Sang-Jin Kim)

2014년 부경대학교 환경대기과학
과(이학사)
2017년 부경대학교 지구환경시스
템과학부(이학석사)
2017년~ 부경대학교 지구환경시
스템과학부 환경대기과학 전공 박사과정
※ 관심분야 : 대기과학, 원격탐사



김근무(Geun-Mu Kim)

2019년~ 부경대학교 환경대기과
학과 학사과정
※ 관심분야 : 대기과학, 원격탐사