Suffix Tree algorithm을 이용한 gene sequence들의 유사성 비교에 관한 연구

한상일, 이성근, 안대명, 황규석*
부산대학교 화학공학과 공정시스템 연구실
(kshwang@pusan.ac.kr*)

A Study on Comparing with the Similarities of Gene Sequences

Sang il Han, Sung Gun Lee, Dae Myung An, Kyu Suk Hwang*
Department of Chemical Engineering, Pusan National University
(kshwang@pusan.ac.kr*)

서론

DNA sequence data들을 다루는 방법 중의 하나인 multiple sequence alignment는 세 개 이상의 단백질이나 DNA 서열들을 배열하여서 유사하거나 같은 부분을 찾아낸다. 그러나 기존의 SP-method, CLUSTALW, PILEUP 을 비롯한 multiple sequence alignment 방법들은 pairwise comparison 을 하므로 서열의 개수가 증가할수록 검색 시간이 크게 증가하는 단점이 있다. 따라서 본 연구에서는 탐색 시간을 줄이기 위해 pairwise comparison을 하지 않고 여러 개의 서열들을 동시에 비교하기 위해 Suffix Tree Clustering알고리즘을 구현하여 multiple sequence alignment에 적용하였다. . 우리는 gene clustering의 5단계를 제시하였다. i)Constructing suffix tree ii)Searching and overlapping common subsequences iii)Grouping subsequence pairs iv)Masking cross-matching pairs v)Clustering pair groups.

Suffix Tree Clustering 알고리즘에서 Suffix Tree Construction Algorithm은 서열길이에 비례하는 선형시간 알고리즘으로써 genomic data같은 대용량의 데이터를 다루기에 효율적이다. Perl language를 이용해 유전자들을 유사성에 따라 clustering하는 suffix tree clustering program 을 만들었고, 프로그램을 평가하기 위해, Mus musculus 종에서 가져온 23개의 유전자를 입력했을 경우와 다른 종들에서 가져온 22개의 유전자를 입력했을 때의 두 가지 경우로 나누어서 multiple sequence alignment를 수행하고 clustering 하였다.

본론

우리는 web document clustering에 주로 이용되는 STC[2](Suffix Tree Clustering)를 도입하여서 Gene을 클러스터링 하기 위해 적용하였다. STC는 common subsequence를 효율적으로 검색하기 위해 Suffix Tree를 이용하며, 검색된 염기서열을 토대로 클러스터들을 형성한다. Zamir et al. 은 collection size에 따른 수행속도를 비교하였고(Figure 1), 같은 클러스터를 형성하여 다른 클러스터링 방법들과 STC의 정확성을 비교하였다.(Figure2) 이러한 STC는 Single-pass, K-means, Buckshot, Fractionation, GAHC(Group-average Agglomerative Hierarchical Clustering)와 같은 클러스터링 방법들 보다 빠르고 정확한 결과를 산출한다.

우리는 suffix tree[1]를 이용한 gene clustering 프로그램을 만들기 위해 Perl language 를 사용하였다. Perl 은Practical Extraction and Report Language의 약어이며 프로그래머 Larry Wall에 의해 만들어졌다. 약어의 의미처럼 문자 data을 추출하고 구성하는 데에 강력하고 실용적인 언어이므로 시스템 관리와 world wide web에서 CGI 프로그래밍에 주로 사용되다가 근래에 genomic data 처리에도 사용되고 있다. 따라서 Perl 은 gene sequence를 다루고 공통부분을 찾아내는 우리의 목적에 부합하는 언어이

다.

본 연구에서 만든 STC algorithm을 이용한 gene clustering 프로그램을 평가하기 위해 NCBI의 homologene database에서, Mus musculus(house mouse)종에서 가져온 23개의 유전자 set와 다른 종들에서 가져온 22개의 유전자 set을 test sequences로 사용하였다.

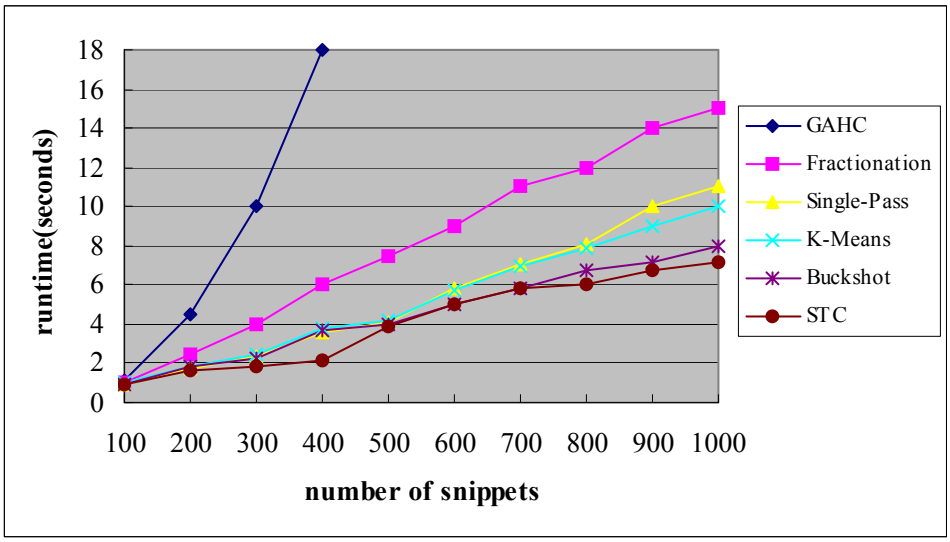

Figure 1. The runtime of the different clustering algorithms on snippet collections as a function of the collection size

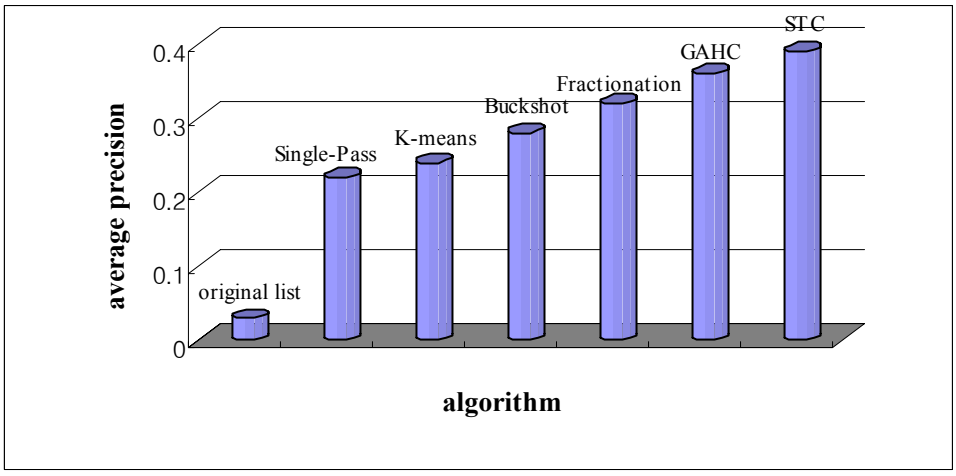

Figure 2. The average precision of the clustering algorithms and the original ranked list.

방법

STC(Suffix Tree Clustering)는 string의 공유된 조각을 바탕으로 clusters를 만드는 standard clustering methods 보다 더 빠른 incremental, linear time 알고리즘이고, Web documents를 clustering하는 STC 의 절차는 다음과 같다.

The procedure of STC(Suffix Tree Clustering) for web documents clustering;

-Step 1
Document "Cleaning" (the string of text representing each document is transformed)
-Step2
Identifying Base Clusters (searching for sets of document sharing common phrase)
-Step3
Combining Base Clusters (merging base clusters with a high overlap)

우리는 유전자를 클러스터링 하기 위해 STC를 도입하여서, document string을 변환하는 불필요한 Document "Cleaning"단계는 수행하지 않고, Step2와 Step3를 수행하였다. 그리고 매우 긴 길이를 가진 유전자들의 엇갈리는 공통부분을 없애고 common subsequences가 순차적으로 매치되도록 하기 위해 Step3-Combining Base Clusters를 수정하였고, step 3를 two steps (grouping the common subsequence pairs and clustering the common subsequence pair groups)으로 나누어서 유사한 DNA sequence들이 clustering 될 수 있도록 하였다.
Suffix Tree 알고리즘을 이용해 여러 개의 서열들에서 공통으로 존재하는 subsequences를 찾아내고 위치 정보를 테이블화 하여서 sequences을 클러스터링 하였다. 우리가 만든 프로그램의 흐름도는 Figure 3과 같다.

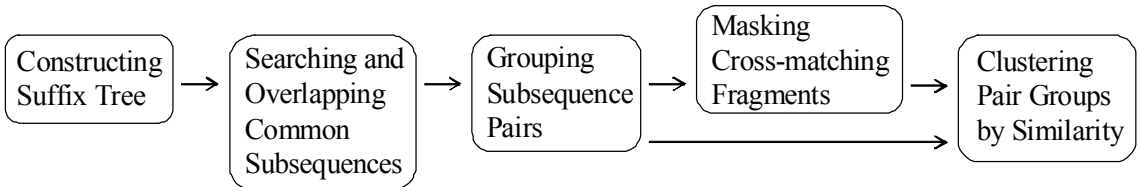

Figure 3. The organization of our gene clustering system

결과 및 토론

gene을 클러스터링 하기 위해 Suffix Tree를 도입하였고 NCBI의 Homologene database의 Mus musculus species에서 가져온 23개의 gene과 서로 다른 species에서 가져온 22개의 gene을 입력했을 때의, 두 가지의 경우에 대해서 Pentium 4-2.4, 1Ghz ram personal computer의 Linux OS가 탑재된 시스템에서 프로그램을 실행하였다.
Table 1과 2는 각각 같은 종에서 가져온 gene sequences와 다른 종에서 가져온 gene sequences를 나타낸다. 우리가 만든 gene clustering 프로그램을 실행하였을 때, Table 1의 데이터는 9개의 cluster를 형성하였고, Table 2의 데이터는 8개의 cluster를 형성하였다. 이는 Homologene database의 cluster들과 대부분 일치함을 보여주었다.

Species	Accession Number
Mus musculus	XM_204449, XM_289927, NM_027609, XM_355690, XM_203409, XM_356889, XM_357648, XM_357087, XM_356880, NM_021300, XM_356386, XM_109566, NM_013548, NM_175653, NM_178215, XM_358107, XM_358117, NM_027650, NM_173069, XM_355572, XM_357595, XM_142567, XM_355567

Table 1. 23 Genes in the Mus musculus species.

Species	Accession number
Homo sapiens	NM_001547, NM_003810, NM_133492, NM_052890, NM_000546, NM_173565 , NM_006926
Mus musculus	NM_008332, NM_009425, NM_175731, NM_011640, NM_009921, NM_023134, XM_205565
Rattus norvegicus	XM_220060, NM_145681, XM_236790, XM_234838, NM_030989, XM_236642, NM_017329, XM_213713

Table 2. 22 Genes in the several species.

현재의 알고리즘에서 gap penalty 혹은 other refinements를 고려하지 않았는데, high sensitivity를 위해서 앞으로 이러한 옵션들이 고려되어야 할 것이다. 그리고 optimization을 위해알고리즘의 중요한 부분을 lower level language를 이용해서 개선하고 parallel implementation을 중요한 부분에 적용하면, 더 빠른 수행속도를 가져와 대량의 데이터에 적용하는 것도 가능할 것이다.

요약하면, 우리는 선형시간 알고리즘인 Suffix Tree를 도입하여 gene clustering 프로그램을 만들었고 database에서 gene을 가져와 두 set의 gene을 클러스터링을 하여 적절한 결과를 보였다. 클러스터링된 data들은 gene의 기능을 유추하거나 진화관계를 파악하고 대량의 데이터를 체계화해서 구성하는 데에 도움을 줄 수 있을 것이다.

참고문헌

1. E. Ukkonen., “On-line construction of suffix trees”, Algorithmica, 14:249-260 (1995).
2. Oren Zamir and Oren Etzioni, "Web Document Clustering : A Feasibility Demonstration", Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (1998).