기능 도메인 예측을 위한 유전자 서열 클러스터링

Gene Sequences Clustering for the Prediction of Functional Domain

한 상 일, 이 성 근, 허 보 경, 변 윤 섭, 황 규 석*
(Sang il Han, Sung Gun Lee, Bo Kyeng Hou, Yoon Sup Byun, and Kyu Suk Hwang)

Abstract : Multiple sequence alignment is a method to compare two or more DNA or protein sequences. Most of multiple sequence alignment tools rely on pairwise alignment and Smith-Waterman algorithm to generate an alignment hierarchy. Therefore, in the existing multiple alignment method as the number of sequences increases, the runtime increases exponentially. In order to remedy this problem, we adopted a parallel processing suffix tree algorithm that is able to search for common subsequences at one time without pairwise alignment. Also, the cross-matching subsequences triggering inexact-matching among the searched common subsequences might be produced. So, the cross-matching masking process was suggested in this paper. To identify the function of the clusters generated by suffix tree clustering, BLAST and CDD (Conserved Domain Database)search were combined with a clustering tool. Our clustering and annotating tool consists of constructing suffix tree, overlapping common subsequences, clustering gene sequences and annotating gene clusters by BLAST and CDD search. The system was successfully evaluated with 36 gene sequences in the pentose phosphate pathway, clustering 10 clusters, finding out representative common subsequences, and finally identifying functional domains by searching CDD database.

Keywords : clustering, suffix tree, gene, BLAST, domain

I. 서론

Human genome project 연구와 그에 관련된 주요 미생물에 대한 유전체 연구가 전세계적으로 활발하게 진행되고 있으며, 최근 컴퓨터를 비롯한 실험장비의 고성능화는 이러한 연구분야에 큰 기여를 하였다. 그 결과로 대사 조절 과정, 단백질의 기능, DNA와 RNA를 비롯한 수 많은 유전 정보들이 NCBI, GenBank나 Swissprot, Unigene 같은 public database에 축적 되었고, 이를 인터넷을 통해 쉽게 접근 할 수 있게 되었다[1]. 이러한 대량의 데이터를 활용하여 다양하고 의미있는 전산적 분석을 실시함으로써, 진화 과정, 단백질의 기능, 유전자 발현 양상 등을 예측할 수 있다[2].

Gene은 염색체의 일부로써, 네 개의 염기(A, T, G, C)로 구성되어 있고, specific functional product (protein or RNA molecule)로 번역되는 DNA의 조각이다. 일반적으로, 이러한 gene sequences는 string형태로 public database [3]에 저장된다. Gene은 functional product를 암호화하는 최소단위로써, 유사한 기능을 나타내는 product를 암호화한 gene은 진화단계에서 조상이 같을 가능성이 크기 때문에, 서로 다른 종이나 같은 종들에서 보존된 common subsequence들을 가진다[3]. 이러한 서열간의 유사성을 보이는 서열들을 homologous sequence라 하고, 같은 homologous sequences에 속하는 유사한 서열들을 그룹화하는 것을 clustering 이라 한다. 이러한 clustering을 통해서 기능이 밝혀지지 않은 유전자의 기능을 예측할 수 있다.

Suffix tree 알고리즘은 선형 시간과 선형 공간으로 구축되어 공통되는 영역을 빠르게 발견할 수 있기 때문에, genomic data와 같은 대용량 데이터를 다루는데 적절하다. Volfovsky [4]등은 genome sequences에서 repeats를 빠르게 찾아내어서 new rice repeat database를 만들기 위해 suffix tree를 적용하였다. Delcher [5,6]는 두 종들의 genome에서 공통되는 Maximal Unique Matching subsequence(MUMs)를 찾기 위해서 suffix tree를 이용하였다. Kalyanaraman [7] 등은 parallel EST clustering 프로그램을 구축하였다. Threshold value 이하의 길이를 보여주는 common substring을 공유하는 서열 쌍들은 pairwise alignment로 비교함으로써 클러스터링을 실행하였다. 그들은 common subsequences를 찾아내기 위해 suffix tree 알고리즘만을 적용하였지만, 본 연구에서는 선형시간 알고리즘을 사용하는 suffix tree를 이용하고, multiple alignment를 사용하여 gene clustering을 수행함으로써, clustering 속도의 개선을 달성하였으며, 각각의 클러스터 중에서 예상되는 functional domain을 찾아내어 그 기능을 예측하였다.

본 연구에서는 Zamir가 제안한 STC(Suffix Tree Clustering) 방법을 gene sequence에 맞게 수정하였다[8]. Zamir [9]등은 web 문서를 클러스터링하기 위해 suffix tree를 도입하였고, STC 방법이 기존의 클러스터링 방법에 비해서 정확성과 속도 면에서 우수하다는 것을 보였다. Gene sequence를 다루는 경우에 있어서, common subsequences는 순차적으로 일치하여야 하므로, cross-matching subsequences를 제거하는 과정을 추가 하였고, 최종적으로 클러스터링된 gene sequences의 기능을 확인하기 위해서 BLAST[10] 검색과 CDD 데이터 베이스 검색을 추가하여 기존의 생물학 데이터 베이스 중에서 기능 도메인을 검색하였다.

II. 방법

본 연구에서 개발된 기능 도메인 예측 시스템은, 1) 먼저, 대상으로 하는 gene sequence 들에 대하여 suffix tree 를 구축하고, 2) 구축된 suffix tree를 기반으로 여러 개의 gene sequence 서열들에서 공통으로 존재하는 common subsequence를 찾아내

* 책임저자(Corresponding Author)
논문접수 : 2006. 3. 23., 채택확정 : 2006. 6. 5.
한상일, 이성근, 변윤섭, 황규석 : 부산대학교 화학공학과
(sangilh@pusan.ac.kr/lee_73@pusan.ac.kr/b97037@kosha.net/kshwang@pusan.ac.kr)
허보경 : 한국생명공학연구원(bkher71@kribb.re.kr)
※ 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음.

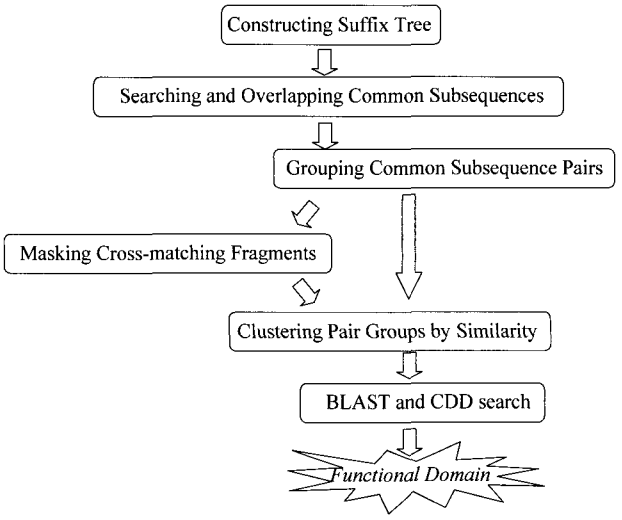


그림 1. 시스템 흐름도.

Fig. 1. System process flow.

고, 3) common subsequence 들의 위치 정보를 수치로 표현하여 gene sequence 들을 클러스터링 할 때 검색기준으로 사용한다. 4) 다음에, common subsequence 들의 유사도에 따라 gene을 몇 개의 그룹으로 clustering 한다. 5) 마지막으로 BLAST 검색과 CDD 검색을 통하여 기능 도메인을 확인한다. 그림 1은 본 연구에서 제안한 시스템의 흐름도를 나타낸다.

1. 공통서열(common subsequence)의 탐색과 클러스터 형성

Ukkonen은 suffix tree를 구축하는 선형 시간 알고리즘을 개발 하였다[11]. 이 알고리즘은 1973년 Weiner [12]에 의해 처음 제안된 방법보다 공간을 더 적게 차지하고 짧은 시간에 컴퓨터 프로그램화 (on-line) 될 수 있다는 장점을 가진다.

따라서 본 연구에서는 Ukkonen이 제안한 suffix tree 알고리즘을 도입하였다. Suffix tree형성 과정은 크게 update, test-and-split, canonize의 세 개의 모듈로 구성되고 다음의 알고리즘에 의해 구축된다[11].

1. create states root and ⊥;
2. for j ← 1, ··· ,m
   do create transition g(⊥,(-j, -j)) = root;
3. create suffix link f¡¯(root) = ⊥;
4. s ← root ; k ← 1 ; i ← 0;
5. while ti+1 ≠ # do
6. i ← i +1;
7. (s, k) ← update (s, (k, i));
8. (s, k) ← canonize (s, (k, i));

두 개 이상의 sequence를 나타내는 suffix tree를 형성하기 위해 각각의 sequence의 끝에 말단기호 '$'를 추가하여 구축된 tree를 GST (Generalized Suffix Tree) 라 한다[5]. 그림 2는 네 개의 sequences (1. ATGCA, 2. ACGCA, 3. TAATC, 4. TACTC)를 이용해 형성된 GST의 구조를 보여준다. 제시된 suffix tree는 root 노드를 포함해서 모두 10개의 노드로 구성되어있고, tree 말단의 네모 상자 안의 수치는 노드와 노드 사이의 string을 포함하는 서열의 번호, sequence에서 string의 시작위치, 끝 위치를 차례대로 나타낸다. 또한 말단기호 '$'는 네 개의 예제 sequences의 끝을 나타낸다.

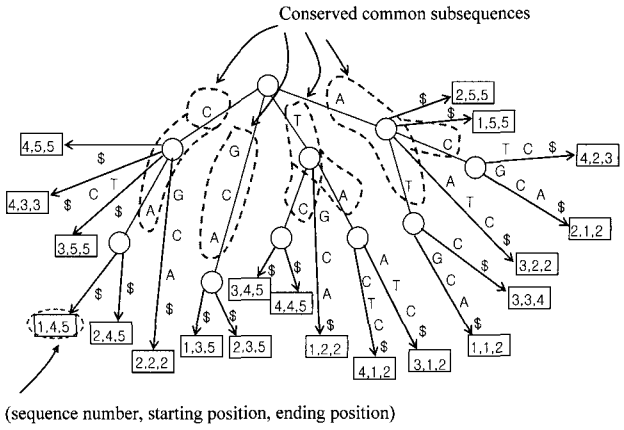


그림 2. 네 개의 서열들에 대한 서픽스트리.

Fig. 2. Suffix tree for example sequences.

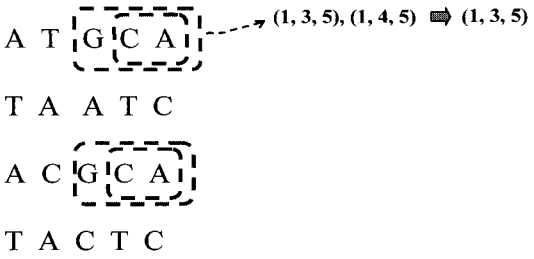


그림 3. 공통 서열 조각들의 결합.

Fig. 3. Overlapping common subsequences.

공통으로 존재하는 common subsequence는 그림 2에서 노드와 노드 사이의 label에 표시되고, 점선으로 common subsequence를 나타내었다. Multiple sequence alignment를 수행할 때, 이것을 기준으로 서열의 유사도를 결정하고 클러스터링을 실행한다. 서열의 종류와 길이에 따라 사용자가 임의로, 시스템이 인식 가능한 최소 서열 길이 (minimum block size)를 선택하게 함으로써, 검색 시간을 최소화하고자 하였다.

구축된 suffix tree는 서열들에서 존재하는 모든 common subsequence를 포함하므로, 서로 overlap되는 common subsequence가 생성될 수 있다. 이러한 overlapping string들은 검색 시간을 낭비하고 비효율적이므로, 서열들의 위치정보를 바탕으로 overlap 되는 common subsequence가 결합된다(그림 3).

컴퓨터가 인식할 수 있게 일반화하기 위해서, 위치 정보에서 각각의 서열 번호를 나타내는 첫 번째 숫자가 일치하는 것을 찾고, subsequence의 시작번호와 끝 번호를 기준으로 어느 하나가 다른 것을 포함하면, 포함되는 위치 정보를 제거하는 과정을 수행하였다(그림 3).

Overlapping을 수행하는 규칙은 다음과 같다. 여기에서 n_(number), a_(number), b_(number)는 각각 서열 숫자, common subsequence의 시작 지점, 종결 지점을 나타낸다.

Position information ; (n_1, a_1, b_1), (n_2, a_2, b_2)
If n_1 = n_2 and {(a_1 <= a_2 and b_1>=b_2) or
(a_1 >= a_2 and b_1<=b_2)}
then
Overlap

위의 과정에서 얻어진 common subsequence는 common subsequence를 나타내는 서열 번호, common subsequence의 시작 위치와 종결 위치로 구성된 위치정보에 의해 표현 된다. Common subsequence는 cross-matching 없이 순차적으로 매치되어야 하므로 본 연구에서는 cross-matching masking 절차가 추가되었다. 또한 검색 시간을 줄이기 위해 longest common subsequence의 길이가 threshold length보다 크다면, cross-matching masking 절차 없이 유사성을 판별 하였다. 이러한 과정을 통해 얻어진 common subsequence pairs를 바탕으로 각 서열들 간의 유사도가 결정되고 유사한 서열들로 구성된 클러스터가 형성된다 (상세한 내용은 참고문헌 [8]참조).

2. Domain 기능 확인

Functional domain을 비롯한 DNA나 단백질의 기능을 밝혀내는 것은 중요하다. 그러한 기능은 데이터베이스 검색을 통해 간접적으로 밝혀질 수 있다. 따라서 본 연구에서는 gene 클러스터링 시스템을 BLAST search와 결합하고 CDD검색을 수행하였다.

BLAST는 GenBank, EMBL, DDBS, Swissprot database에 대한 검색을 수행하고 그 결과 query 서열과 유사한 서열을 발견함으로써, 기능이 밝혀지지 않은 서열에 대한 기능 유추가 가능하다. CDD 는 도메인과 단백질에 대한 multiple sequence alignments 정보를 제공하는 데이터베이스이다. 따라서, CDD 검색을 통해 단백질 query 서열에 존재하는 보존된 domain을 발견할 수 있다.

'blastall' tool을 NCBI (National Center for Biotechnology Information) site에서 다운받아 설치하여 BLAST 검색 (blastn, blast, blastx)을 수행한다. (Download: ftp://ftp.ncbi.nih.gov/blast/executables/).

제안된 suffix tree clustering 방법을 통해 얻어진 각각의 cluster에서 가장 길이가 긴 common subsequence들을 이용하여 BLAST와 CDD database 검색을 수행하였다. BLAST 검색에서 사용자의 선택에 의해 DNA 나 protein database에 대한 검색을 할 수 있다. DNA database (GenBank, EMBL, DDBJ)와 protein database에 대한 검색 결과는 각각 'nucleotide_blast.out' 'protein_blast.out' 파일에 저장된다.

표 1. Pentose phosphate 대사경로의 36개 유전자.
Table 1. The 36 genes of pentose phosphate metabolic pathway.

Species	Entry
P.syringae	PSPTO2492,
P.fluorescens	PFL_4100
L.acidophilus	LBA1098
P.multocida	PM1248
A.tumefaciens_C	AGR_L_1647, Atu4029
M.musculus	14380, 14120
H.sapiens 1	2539, 5226, 8789
D.melanogaster	CG12529-PA,CG3724-PA, CG31692-PA, CG2827-PA
C.elegans	T25B9.9, K07A3.1
V.cholerae	VCA0898
E.coli_J	JW4346, JW3731
S.typhi	STY4920, STY3892
Y.pestis_Mediaevails	YP3743
V.fischeri	VF0506
S.cerevisiae	YLR354C
C.albicans	orf19.4371
C.neoformans	CNK03170
S.typhimurium	STM3885, STM3612
Y.pseudotuberculosis	YPTB0008
S.enterica_Paratyphi	SPA3468
S.flexneri_2457T	S4209
E.carotovora	ECA4360
B.bronchiseptica	BB1475
Burkholderia_383	Bcep18194_B1794
P.fluorescens_PfO1	Pfl_0086

III. 결과 및 토론

1. 서픽스트리 클러스터링 알고리즘에 의한 유전자 서열 클러스터링

KEGG(Kyoto Encyclopedia of Genes and Genomes) database 에 제시된 pentose phosphate pathway 와 관련된 36개의 gene sequence (표 1) 를 대상으로 본 연구에서 제안한 시스템을 적용하였다. 분석은 Intel's Pentium 2.4 GHz processor, 1GB memory 의 컴퓨터와 Linux OS 환경 하에서 수행되었다.

Fasta 파일 포멧의 gene sequence 데이터가 gene 클러스터링 프로그램에 입력된다. 프롬프트 윈도우에서 gene sequence들을 포함하는 파일 이름을 모니터상에 입력하고, minimum block size 를 설정한다. 본 연구에서는 minimum subsequence size를 10으로 설정하였다. 본 연구의 경우, minimum block size가 10보다 더 커진다면, 서열 유사성 비교에서 발견되는 공통 subsequence들의 개수가 줄어들어서 비교하는 서열들 간의 유사도를 떨어뜨리고, 따라서 부정확한 클러스터를 유발하는 것으로 나타났다. 마지막으로 BLAST 검색을 수행할 database (DNA or Protein) 형태를 선택하면 프로그램이 실행된다.

본 연구에서 시스템은 대상 예제 gene에 대해서 모두 10개의 cluster 그룹을 형성하였다.

36개의 gene sequence 중 몇몇 sequence를 제외하고 모두 클러스터링이 되었다. 클러스터링 결과(그림 4)는 KEGG database에서 ortholog 그룹과 일치하는 적절한 결과를 보여주

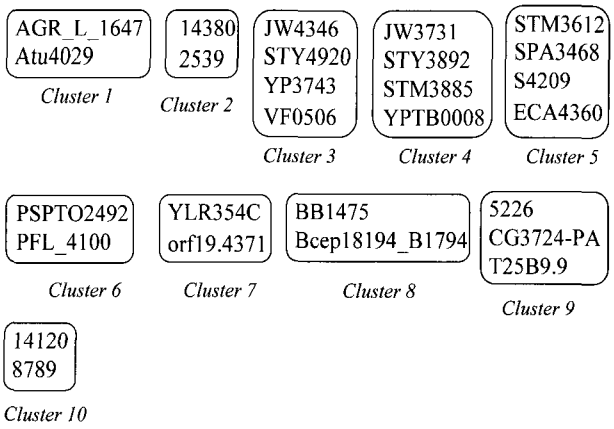


그림 4. 클러스터링 결과.
Fig. 4. Clustering result.

표 2. 그림 4의 클러스터에서 가장 긴 단백질 공통 서열.
Table 2. The longest protein subsequences in the clusters of Fig. 4.

Cluster	Common subsequence
1	MTTDKTSPYEIHDDRFRHLIVGNAELEELYSGCR WAEGPVWFADLNCLLFSDIPNERMLRWVPDGGIS VFRQPSNFTNGNTRDRQGRLVSCEHGGRRVTRTE VDGSITVLADSYGGKRLNSPNDVVVKSDGSVWF TDPTYGILSDYEGYKAEPEQKTRNVYRLDPATGKI DIAIDDFVQPNGLAFSPDETKLYVADSSYSHDISRP RHIRVFDVVDGVRLANGREFCNLDNGLPDGFRLD TAGNLWTSAGDGVHCFAPDGRLIGKIKVPQTVAN LTFGGPKKNRLFITATKSLYSVYVAATGAQTPX
2	IIMGASGDLAKKKI
3	MKRAFIMVLDSFGIGA
4	AGDTFNGAFITAILEETPLPEAIRFAHAAAAIAVTR KGAQPSVPWREEIEAFLHQQGX
5	MSKKIAVIGECMIELSQKGADVQRGFGGDTLNTS VYIARQVDSAALAVHYVTALGTDSFSQQMLEAW QHENVDTSLTQRMENRLPGLYYIETDDTGERTFY YWRNEAAAKFWLES

었다. 생성된 텍스트 파일 'cluster.out'에서 클러스터링된 gene sequence들의 결과를 볼 수 있다.

그림 5는 클러스터링 결과 중 Cluster 3, 4, 5를 보여준다. 사각의 박스는 각각의 sequence에서 보존된 부분을 나타낸다. 이러한 보존된 common subsequence를 바탕으로 sequence들 간의 유사도를 비교한 후에 클러스터링이 수행된다. Cross-matching common subsequence들은 부정확한 유사성을 유발할 수 있으므로, 본 연구에서는 cross-matching common subsequence들이 제거되었다. 또한, STC의 효율성을 보여주기 위하여, pyruvate metabolic pathway에서 22개의 유전자 서열들을 클러스터링 하여서 각 클러스터의 공통 서열을 발견하였으며, 그 결과는 웹페이지 (http://home.pusan.ac.kr/~sangilh/)에서 확인 할 수 있다. 한편, 클러스터링을 통해서 찾아지는 가장 긴 common subsequence는 클러스터링에 사용되는 유전자 셋에 따라 달라지는 것이 아니며, 클러스터링 되는 유전자들의 유사도에 따라 그 크기가 달라진다.

2. BLAST와 CDD 검색을 통한 기능 도메인 확인

Gene cluster들의 기능을 파악하기 위해, 각각의 클러스터에서 가장 긴 DNA common subsequence를 추출하여 유전암호에 따라 단백질서열로 번역하고, BLAST 검색을 위한 query sequence로 사용하였다 (표 2). 유전암호에 존재하지 않는 코돈은 'X'로 표시하였으며, BLAST 수행의 결과는 파일 'protein_blast.out'에서 텍스트 형태로 저장된다. Web 기반 BLAST는 불안정하고 느리기 때문에, 본 연구에서는 컴퓨터에 직접 local BLAST tool을 설치하고 사용하였다.

검색결과, cluster 2의 common subsequence에 대해 모두 13개의 유사한 sequence 들이 발견되었고, 13개 서열들의 기능은 'Glucose-6-phosphate 1-dehydrogenase (G6PD)' 임을 BLAST 검색을 통해 알 수 있다. 따라서 , cluster 2의 두 개의 서열은 Glucose-6-phosphate-1-dehydrogenase (G6PD) 기능과 관계가 있을 것이라고 판단된다. 한편 기능뿐만 아니라 G6PD 기능을 가지는 cluster 2의 서열들에서 어떠한 지역이 가장 많이 보존되었는지 도메인구조에 대한 정보를 얻기 위해, cluster 2의 가장 긴 common subsequence를 이용해 CDD 데이터베이스 검

그림 5. 파일 'cluster.out'의 일부분.
Fig. 5. A part of file 'cluster.out'.

색을 수행한다. 또한 cluster 1, 3, 4, 5의 common subsequence들에 대해서도 각각 'Precursor', 'Phosphopentomutase', 'Ribokinase', '2-dehydro-3-deoxygluconokinase'의 기능을 가지는 서열들이 발견되었다. 이것은 본래 데이터가 저장되어 있는 KEGG 에서 명명된 기능과 같다. 나머지 cluster의 common subsequence들은 짧은 길이 때문에 매치되는 서열들이 발견되지 않았다.

표 2에 나타난 가장 긴 common subsequence들은 각각 gene cluster들을 대표하는 매우 잘 보존된 의미 있는 지역으로 예상된다. 또한, nucleotide database들에 대한 검색 옵션을 시스템에 추가하여, DNA수준에서도 검색이 가능하다. Nucleotide database들에 대한 검색결과는 파일 'nucleotide_blast.out' 에서 제시된다.

3. 표 2의 common subsequence에 대한 생물학적 의미

클러스터링을 통해 밝혀진 common subsequence에 대한 생물학적 의미를 살펴보기 위해 NCBI의 CDD (Conserved Domain Database)를 검색하였다. 표 3은 각각의 subsequence에 대해 발견된 domain을 보여준다. Cluster 1의 common subsequence는 다섯 개의 도메인과 동시에 매치가 되었다. pfam03758은 간과 신장에서 중요한 역할을 하는 senescene marker protein-30 도메인, pfam03088은 alkaloid biosynthesis에서 중요한 효소작용을 하는 도메인, COG3386은 carbohydrate 전달과 대사작용에 관여하는 도메인, COG4257는 streptogramin lyase 작용을 하는 도메인임을 CDD 검색을 통해 알 수 있었다. COG3391 도메인의 기능은 아직 밝혀지지 않았다. 따라서, senescene marker protein-30의 기능, alkaloid biosynthesis 기능, carbohydrate 대사작용 등을 하는 도메인에서 구조적 · 기능적으로 유사한 지역이 존재한다고 판단된다.

Cluster 2에서는 어떠한 도메인과도 매치가 되지 않았는데, 이러한 점은 현재 개발된 suffix tree 방법만으로는, domain sequence가 완벽하게 일치하지 않은 경우, gap이나 mutation등의 경우에는 분석이 불가능하므로, pairwise 비교를 수행하는 local alignment를 도입한 정교한 시스템으로 추후 개선해야 할 점이다.

표 3. 표 2의 subsequences에 대한 CDD 검색 결과.
Table 3. CDD database search result for the subsequences in Table 2.

Cluster	Domain	E-value
1	pfam03758 (SMP-30)	2e-25
	pfam03088 (Str_synth)	5e-08
	COG3386	7e-71
	COG4257 (Vgb)	9e-04
	COG3391	0.006
2	Null	Null
3	COG1015 (DeoB)	0.001
4	cd01174 (ribokinase)	2e-06
	pfam00294 (PfkB)	3e-05
	COG0524 (RbsK)	8e-07
5	cd01166 (KdgK)	1e-26
	cd01167 (bac_FRK)	9e-12
	cd01174 (ribokinase)	3e-10
	cd01942 (ribokinase_A)	4e-04
	cd01945 (ribokinase_B)	0.002
	pfam00294 (PfkB)	3e-17
	COG0524 (RbsK)	2e-17

IV. 결론

본 연구에서는 선형시간과 선형공간으로 gene sequence 데이터를 정리하고 검색하기 위하여 suffix tree 알고리즘을 사용하였으며, 기존의 STC (Suffix Tree Clustering) 방법을 gene sequence 데이터 처리에 적합하도록 수정 보완하였다. 또한, 이를 BLAST 와 CDD 검색과 결합하여 gene sequence들을 빠르고 정확하게 클러스터링하고 functional domain을 예측할 수 있는 시스템을 구축하였다.

제안된 STC 방법을 pentose phosphate pathway의 실제 genomic 데이터에 적용하여 유효한 clustering 결과를 얻을 수 있음을 보여주었고, 각각의 gene들을 대표하는 공통된 common subsequence들도 찾을 수 있었다. 또, BLAST검색과 NCBI의 CDD (Conserved Domain Database) 데이터베이스 검색 결과로부터, common subsequence들에 대해서 기존에 밝혀진 여러가지 기능을 잘 예측할 수 있음을 확인하였고, 생물학적인 기능domain 들과도 잘 매치되었고, 서로 다른 종 사이에서 보존되는 중요한 영역을 잘 예측할 수 있음을 확인하였다.

참고문헌

[1] D. W. Mount, "Bioinformatics: Sequence and genome analysism," *Cold Spring Harbor Laboratory Press*, New York, pp. 3-5, 2001.

[2] J. Y. Chen and J. V. Carlis, "Genomic data modeling," *Information Systems*, vol. 28, pp. 287, 2003.

[3] J. M. Ostell, S. J. Wheelan, and J. A. Kans, "The NCBI data model.," *Methods Biochem. Anal,*. vol. 43, pp. 19, 2001.

[4] N. Volfovsky, B. J. Haas, and S. L. Salzberg, "A clustering method for repeat analysis in DNA sequences," *Genome Biol.*, vol. 2, pp. 1-11, 2001.

[5] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes," *Nucleic Acids Res.*, vol. 27(11), pp. 2369-2376, 1999.

[6] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg, "Fast algorithms for large-scale genome alignment and comparisonm," *Nucleic Acids Res.*, vol. 30(11), pp. 2478-2483, 2002.

[7] A. Kalyanaraman, S. Aluru, and S. Kothari, "Parallel EST clustering," *HICOMB,* 185, 2002.

[8] S. I. Han, S. G. Lee, B. K. Hou, S. H. Park, Y. H. Kim, and K. S. Hwang, "A gene clustering method with masking cross-matching fragments using modified suffix tree clustering method," *Korean J. Chem. Eng.*, vol. 22(3), pp. 345, 2005.

[9] O. Zamir, O. Etzioni, O. Madani and R. M. Karp, "Fast and intuitive clustering of web documents," *In Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining,* pp. 287-290, 1997.

[10] S. F. Altschul, W. Gish, W. Miller, E. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.,* vol. 215, pp. 403-410, 1990.

[11] E. Ukkonen, "On-line construction of suffix trees," *Algorithmica,* vol. 14, pp. 249-260, 1995.

[12] D. Gusfield, "Algorithms on strings, trees, and sequences: computer science and computational biology," Cambridge University Press, London, pp. 116, 1997.

한 상 일

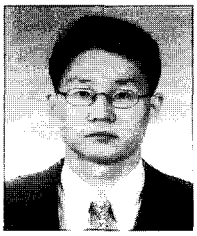
1978년 4월 4일생. 2003년 부산대학교 화학공학과 학사. 2005년 부산대학교 화학공학과 석사. 관심분야는 생물정보학, 시스템 미생물학.

허 보 경

1971년 5월 27일생. 1994년 부산대학교 화학공학과 학사. 2000년 부산대학교 화학공학과 박사. 현재 한국생명공학연구원 국가바이오정보센터 선임연구원. 관심분야는 생물정보학, 인공지능.

황 규 석

1955년 1월4일생. 1982년 부산대학교 화학공학과(학사). 1985년 일본 동경 공업대학 화학공학과(석사). 1988년 일본 동경공업대학 화학공학과 박사. 현재 부산대학교 화학공학과 교수. 관심분야는 화학공정전문가 시스템, 생물정보학.



이 성 근

1973년 3월 1일생. 1996년 부경대학교 화학공학과 학사. 1998년 부산대학교 화학공학과 석사. 2005년 부산대학교 화학공학과 박사. 관심분야는 생물정보학, 시스템 미생물학.

변 윤 섭

1968년 1월 12일생. 1992년 부산대학교 화학공학과 학사. 1995년 부산대학교 화학공학과 석사. 현재 부산대학교 화학공학과 박사과정. 관심분야는 공정시스템, 화학공정안전.