EST Clustering와 열람 Viewer의 개발에 관하여

EST에 대한 데이터는 많이 등록되어 있지만 서열의 중복, annotation의 부재 등 문제점이 있으므로, EST clustering을 하여 중복 없는 전사서열 database를 작성하고, annotation을 첨부하여 효율적으로 이용할 수 있는 Viewer 개발의 필요성이 대두되었다.
최근 Dragon Genomics는 송사리 EST 데이터 clustering과 열람 Viewer를 개발하여 제 8회 소형어류 연구회(2002년 8월)에서 발표하였다.

■ 개요

GenBank의 송사리 EST mRNA 서열 데이터(46,522 서열)를 토대로, PARACEL Inc.의 PCP(Paracel Clustering Package)를 이용하여 vector 서열 masking, clustering, assembly로 각 cluster의 consensus 서열을 작성하여 14,307 종으로 분류하였다.
이 cluster 서열에 annotation을 첨부하기 위하여 NCBI BLAST를 이용하여 GenBank non-redundant database로 BLASTX 검색하여 EST viewer에 결과 데이터를 정리하였다. 이 결과 SNP 후보부위 및 homology 후보를 검색하고, cluster 서열 기능분류도 하였다.

■ Algorithm

Clustering 해석에서 가장 중요한 것은 서열을 cluster로 분류(서열의 유사성을 기준)하는 것이다. 이 cluser 분류 결과에 따르면, NCBI database UniGene에서는 BLAST(megaBLAST) algorithm을 이용하고 있는 것에 반해 당 센터에서는 Haste algorithm을 이용하고 있다(그림 1, 2). 어떤 algorithm이든 일부 서열의 일치도가 높은 것만을 검출하여 분류 하는 것으로, 고속 분류를 하고 있으나, 일치하는 서열을 검출하기 위한 각각의 특징이 있으므로 일률적으로 좋고 나쁨을 평가할 수는 없다. Dragon에서는 보다 맞춤화된 Haste algorithm(PARACEL Inc.의 소프트웨어 사용)을 이용하고 있다(그림 5 참조).

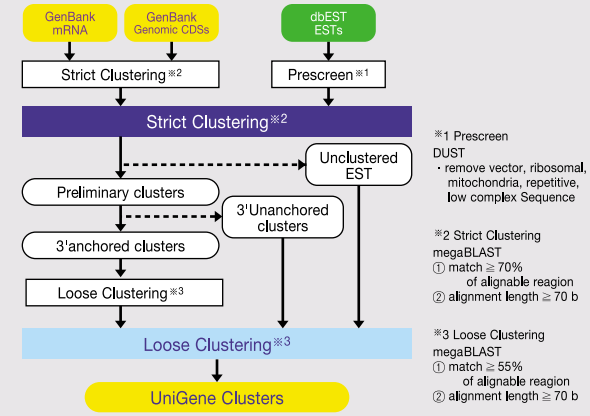

그림 1 UniGene에서 사용하고 있는 algorithm

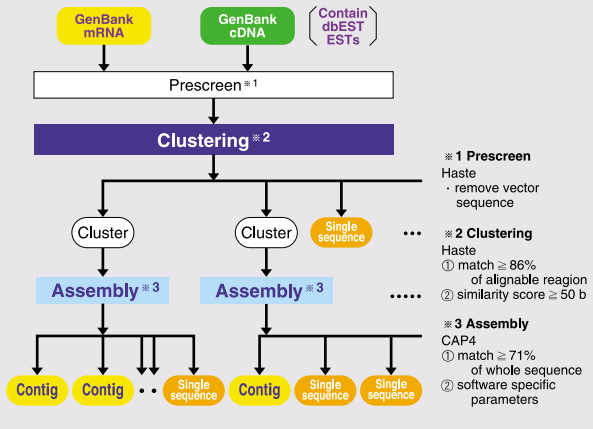

그림 2 Dragon Genomics에서 사용하고 있는 Algorithm

■ EST 해석 각 작업 단계

EST 해석의 각 작업 단계(그림 3)를 순서대로 설명한다.

NCBI의 검색 시스템 Entrez를 이용한 Keyword search를 통하여 송사리 전사서열 데이터를 얻었다(그림 4). 이 서열을 mask 처리(vector 서열을 제거), clustering 처리(서열의 그룹분류), assembly 처리(contig 형성)로 중복 없는(엄밀히 "중복이 거의 없는") database를 작성한다(그림 2, 그림 5, 그림 6). 이 결과 cluster, contig, consensus 서열, single sequence를 얻어(그림 7), 효과적으로 볼 수 있는 viewer를 개발하였다. Cluster 개요는 그림 8과 같은 표 형식으로, contig 개요는 그림 9와 같은 alignment로 표시된 viewer로 열람할 수 있다.

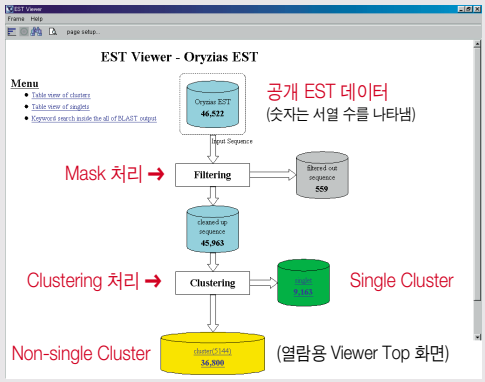

그림 6 Clustering 결과를 나타내는 Viewer의 Top 화면(Step 2)

그림 3 EST 해석 작업순서

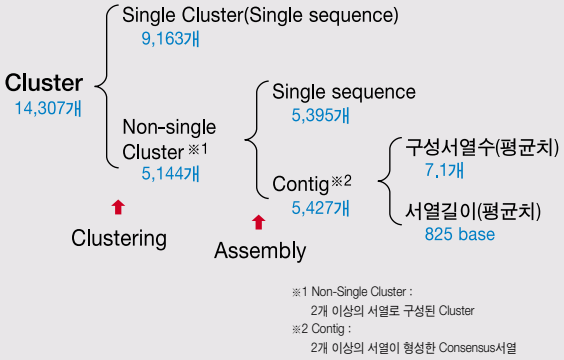

그림 7 Cluster 내용(Step 2)

● GenBank에서 EST에 해당하는 데이터

Entrez Search Keyword :

Orizias latipes [Organism]

AND (***mRNA*** [Title word]

OR ***cDNA*** [Title word]

NOT genome [Title word]

NOT chromosome [Title word]

Total sequence	: 46,522 seq
Average	: 545.0 base
Max. sequence length	: 5,845 base
Min. sequence length	: 30 base

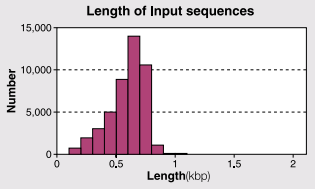

그림 4 데이터 획득 (Step 1)

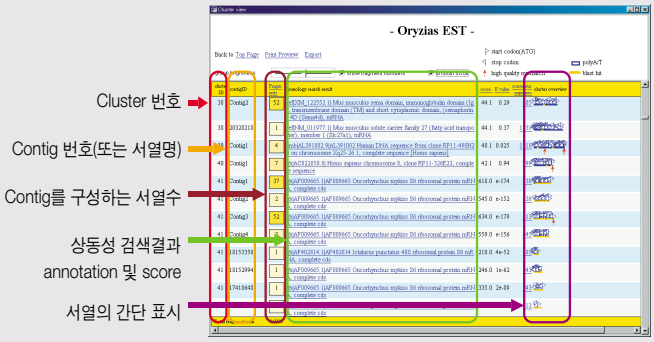

그림 8 Cluster의 열람 viewer(Step 2)

● Clustering 작업의 개요

Clustering에 사용되는 서열을 포함하는 vector 서열을 제거한다(Clustering 및 assembly 작업에 고려하지 않는 서열로 한다).

↓

Clustering

↓

Clustering된 서열을 각 cluster에서 assembly하여 consensus 서열을 작성한다

● 이용 소프트웨어

Process	UniGene Database	Dragon 전사서열 Database
prescreen	Dust	PFP(Haste)
1st Grouping	Strict Clustering (megaBLAST)	Clustering (Haste)
2nd Grouping	Loose Clustering (megaBLAST)	Assembly (CAP3)

※ 좌측 algorithm은 Paracel Clustering Package(PCP) 소프트웨어를 사용하였으며, version up되어 Paracel Transcript Assembler(PTA)으로 명칭이 변경되었다.
※ Haste는 Smith-Watermenn Algorithm을 근사계산 한 것이다.

그림 5 Clustering 작업의 개요(Step 2)

● 아래 그림은 Assembly에 의해 형성된 Contig 하나를 viewer로 열람한 예

Contig를 형성하는 서열의 alignment 개요도

확대그림

각 서열의 alignment

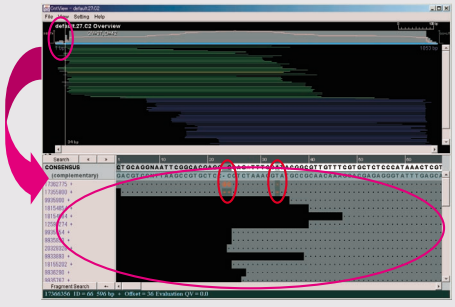

그림 9 Contig의 열람 Viewer(Step 2)

■ SNPs 후보 검색

이와 같이 작성된 열람용 viewer를 이용하여 SNPs 후보를 찾을 수 있다(그림 10~12). Contig 열람 viewer(그림 10)에서 alignment가 일치하지 않는 염기(즉, 서열을 2종류로 분류할 수 있는 경우)(그림11)는 SNPs일 가능성이 크다. 이런 염기는 엄격한 수치 조건을 적용하여 SNPs 후보 존재밀도를 잡은 결과 0.4개/ Contig(=0.5개/1 kb)가 된다.

이 수치는 human의 일반적인 SNPs의 존재밀도(0.5~2개/1 kb)와 비교해도 유효한 수치이다(그림 12)

Cluster 일람 viewer에서 일치하지 않는 염기 표시 (빨간 화살표)

Contig viewer에서 일치하지 않는 염기의 표시(빨간점) 이 부위는 consensus 서열값도 낮다

Alignment 표시에서 일치하지 않는 염기의 표시(빨간색)

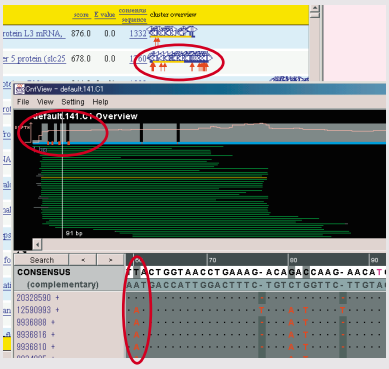

그림 10 일치하지 않는 염기의 열람 viewer(Step 2)

형성된 Contig의 내역

확대그림

각 서열의 alignment

SNP 후보부위

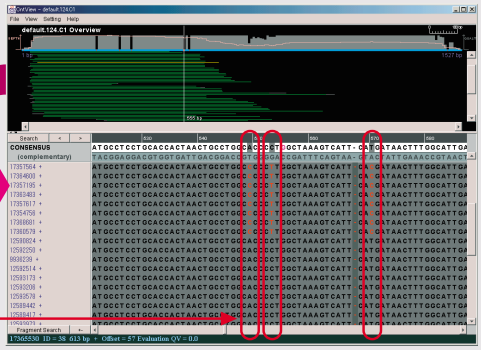

그림 11 SNP 후보 예 (Step 2)

SNP부위의 후보 검색

Clustering으로 만들어진 contig(일치하지 않는 염기의 수가 작은 경우의 consensus 서열)에서 다음의 두 가지 조건을 만족시키는 부위를 SNP 후보 부위로 계산하였다.

(1) Consensus 서열 값(QV)이 "1" 또는 "0"으로 된 경우(단, 각 EST 데이터의 모든 염기에 10을 할당한 후 alignment하여 consensus 서열을 형성할 때, 그 값를 CAP4 법으로 계산하였다.)

(2) (1)의 염기 중, 그 주변의 consensus 염기 값이 20 이상인 경우

SNP 후보총수	2,053개
SNP 후보의 존재밀도	0.4개/contig

그림 12 SNP 후보 검색(Step 2)

■ Annotation 첨부

Clustering 작업에서 얻은 contig, single sequence에 annotation을 첨부하기 위하여 GenBank nr(non-redundant protein database) database의 BLASTX를 이용하여 상동성 검색을 하였다. 그 결과 구성하고 있는 cluster 중, 하나의 서열로 이루어진 cluster(single cluster)는 상동성이 높은 것은 거의 등록되어 있지 않아 신규 유전자가 많은 것으로 생각되지만, 두 개 이상으로 이루어진 서열의 cluster(non-single cluster)는 이미 알려진 유전자가 많이 포함되어 있었다(그림 13).

이 결과를 열람하기 위한 viewer(그림 14)는 cluster 일람에 각 cluster의 구성서열 annotation을 첨부하고, 그 annotation을 첨부하기 위하여 실시한 상동성 검색 결과 페이지로 링크하였다.

- Cluster 서열에 annotation을 첨부하기 위하여 NCBI BLAST를 이용하여 GenBank non-redundant Database의 BLASTN을 이용하여 상동성 검색을 하여 그 결과를 viewer의 표에 부가하고 검색결과 E-value를 이용하여 유사서열의 존재를 아래 표에 나타내었다.

유사성 정도	Nonsingle Cluster	Single Cluster	Total	(%)
Strong (E-value 〈10^{-100})	2,386	167	2,553	13
Good (E-value : 10^{-100}~10^{-20})	2,987	930	3,917	20
Novel (E-value 〉10^{-20})	5,449	8,066	13,515	68
Total	10,822	9,163	19,985	100

※ BLAST 검색에 사용한 Query 서열은 non-single cluster내의 contig서열과 singlet서열이다.

그림 13 상동성 검색(Step 3)

Cluster 일람

Alignment 표시

Contig 서열의 BLAST 결과

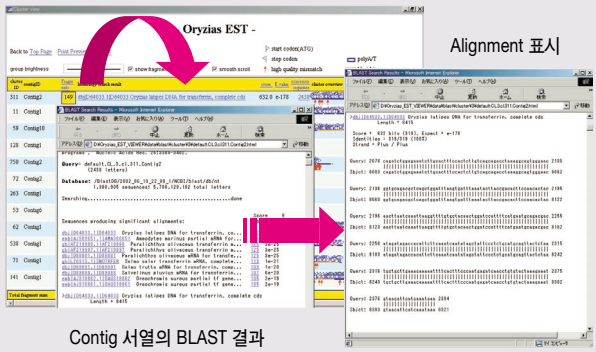

그림 14 상동성 검색결과 viewer 열람(Step 3)

■ Homology 후보의 검출

Annotation 일람을 이용하면 homology 후보 검출이 가능하다(그림 15). 작성한 database는 일부 영역만 상동성이 높은 서열이 같은 cluster로 분

● 같은 cluster로 분류된 서열이 두 종류 이상의 consensus 서열(두개 이상의 contig)을 형성하는 경우는 homology인 경우가 많다고 생각할 수 있다. 또한 annotation이 같은 경우는 splicing variation를 포함할 가능성이 있다. 따라서 이 데이터를 상세하게 해석하면 splicing variation 후보를 검색할 수 있다.

cluster ID	contigID	Fragments	homology search result	score	E value	consensus sequence
53	Contig1	23	dbj|D87740.1|D87740 Oryzias latipes mRNA for muscle actin OlMA1, complete cds	1953.0	0.0	1702
53	Contig2	61	dbj|AB036756.1|AB036756 Chrysophrys major mRNA for B-actin, complete cds	1340.0	0.0	1946
53	Contig3	4	dbj|AB037865.1|AB037865 Tilapia mossambica mRNA for beta-actin, complete cds	585.0	e-164	732
53	Contig4	4	dbj|D50029.1|CRASAA2 Goldfish mRNA for skeletal alpha-actin, complete cds	722.0	0.0	794
53	Contig5	6	dbj|D89627.1|D89627 Oryzias latipes mRNA for cytoplasmic actin OlCA1, complete cds	1203.0	0.0	633
53	Contig6	114	dbj|D89627.1|D89627 Oryzias latipes mRNA for cytoplasmic actin OlCA1, complete cds	3515.0	0.0	1883
53	Contig7	3	ref|NM_131591.1| Danio rerio actin, alpha 1, skeletal muscle (acta1), mRNA	628.0	e-177	650

그림 15 Homology 후보의 검출(Step 3)

류되며, assembly 처리를 통해 전체적으로 상동성이 높은 서열을 contig로 구성하므로 homology 유전자(같은 서열을 그 일부로 공유하는 유전자)는 동일한 cluster에서 다른 contig를 형성하는 경향이 있다. 이로 인해 homology 후보 중에서 splicing variation 후보를 검색할 수 있다.

■ 기능분류

각 cluster 서열의 기능을 분류하기 위해서 각 서열을 query로 KEGG(Kyoto Encyclopedia of Genes and Genomes)의 database에 BLASTX 검색을 다음과 같이 하였다.

Cluster 서열의 기능분류 방법(Step 4)

전체적인 pathway 관련유전자가 있는 KEGG(Kyoto Encyclopedia of Genes and Genomes) database를 이용하여 *Oryzias Latipes*의 cluster 서열의 기능을 분류하였다. KEGG는 각 유전자에 독특한 EC번호가 있으며 pathway는 이 EC 번호를 갖는 효소의 서열로 표현된다. 여기서는 query를 cluster 서열로 하여 BLASTX 검색을 하였다.

그 결과 KEGG의 database에 Hit(E-value≤10^{-5})한 cluster수는 1,946개로, 전체 cluster 서열의 1/10정도로 전체 cluster에 대한 기능분포정보를 얻기에는 불충분하다(이 정보는 각각의 기능을 갖고 있는 유전자가 어느 정도 알려져 있는지 반영한다).

그 결과를 그림 16에 나타낸다.

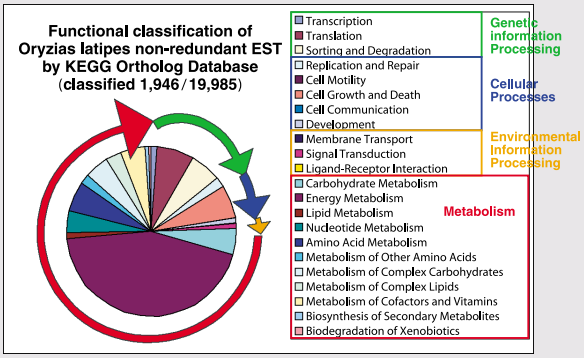

그림 16 Cluster 서열의 기능분류 결과(Step 4)

■ 결론

자체 개발한 열람용 viewer로, 위와 같이 공개된 EST 데이터를 이용하여 중복되지 않는 전사서열 database를 작성하여, 유용하게 이용할 수 있다. EST viewer의 기능은 다음과 같다.

- 해석한 서열 수나 clustering 개요표시(그림 6)
- Cluster의 일람표시(그림 8)
- Contig를 구성하는 서열수에서 서열전환기능(발현량이 많은 유전자의 추측)
- Consensus 서열을 구성하는 EST 서열의 alignment 표시(그림 9)
- Consensus 서열의 quality 표시
- Consensus 서열의 export 기능
- 상동성 검색결과의 상동성 영역표시(그림 14)
- 상동성 검색결과의 key word 검색기능

위 해석은 clustering 입력 데이터로 공개된 EST 데이터를 이용하였지만, 의뢰자가 갖고 있는 EST 데이터를 이용할 수 있다. EST 해석 서비스를 이용할 경우는 EST 열람 viewer 서비스를 제공한다. EST해석 서비스에 대한 자세한 내용은 아래로 연락하기 바란다.