제 6 장 Genome sequence의 이해

**학습목표**

1. 게놈 sequence의 분석시 computer를 이용하는 것의 장단점 파악

2. ORF 분석의 원리를 알고, 유전자를 찾아내기 위해 사용하는 ORF scanning이 항상 etjdrhd적이지 못한 이유 파악

3. RNA로 전사되는 유전자의 부분을 찾아내는 여러 가지 실험방법 학습

4. homology의 의미 파악과 homology가 computer를 이용한 유전자 기능분석에 필요한 이유

5. homology 분석의 한계점

6. 유전자 기능분석을 위한 유전자 불활성화법 이해

7. 미지의 유전자에서 만들어진 단백질의 활성을 알아내기 위한 방법 학습

8. 전사체와 단백체 분석법

9. 단백질 상호작용 map 만드는 법

10. genome sequence를 이해하기 위한 비교 게놈학의 역할

**6.1 게놈상의 유전자 찾기**

ㆍ눈 혹은 computer를 이용하여 특별한 sequence를 찾아내는 방법과 실험적인 방법으로 유전자를 찾아내는 방법이 있음.

ㆍcomputer를 이용하는 방법은 Bioinformatics의 한 분야임.

6.1.1 Sequence를 조사하여 유전자 찾기

ㆍ유전자는 유전자만의 특징적 염기서열을 갖고 있으므로, 이것을 이용하여 유전자 여부를 판단함.

■암호부위는 ORF이다.

ㆍ단백질을 암호화하는 유전자는 ORF를 갖고 있다.

Open reading frame

Synonym(s)

* ORFs
* protein coding sequences

Definition(s)

Sequences of structural genes devoid of termination codons and therefore continuously "readable" by RNA polymerase.

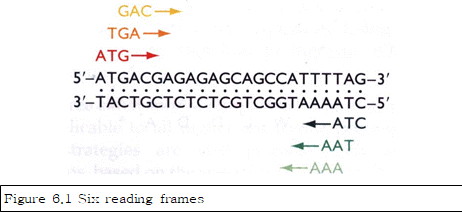
Definition from: [Unified Medical Language SystemEMB000036900be4](http://ghr.nlm.nih.gov/ghr/visit;jsessionid=FEA19E2B34E1AC7D1EF678890C72A31D?uid=7bH4sIAAAAAAAAAEvPyS8uTiyqjI-Pr8koKSmw0tcvzc0pzsxLy9fLy8nVy8vM0EvPL9MHAP0LL50oAAAA) (CRISP Thesaurus)   at the National Library of Medicine

A reading frame in a sequence of nucleotides in DNA that contains no termination codons and so can potentially translate as a polypeptide chain. (On-line Medical Dictionary)

Definition from: [Unified Medical Language SystemEMB000036900be5](http://ghr.nlm.nih.gov/ghr/visit;jsessionid=FEA19E2B34E1AC7D1EF678890C72A31D?uid=7bH4sIAAAAAAAAAEvPyS8uTiyqjI-Pr8koKSmw0tcvzc0pzsxLy9fLy8nVy8vM0EvPL9MHAP0LL50oAAAA) (NCI Thesaurus)   at the National Library of Medicine

A reading frame that does not contain a nucleotide triplet which stops translation before formation of a complete polypeptide -- abbreviation ORF.

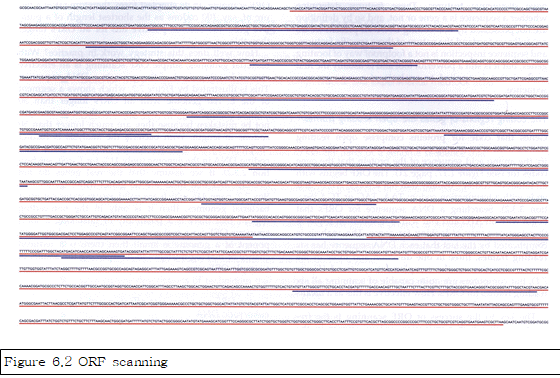
Definition from: [Merriam-Webster's Medical DictionaryEMB000036900be6](http://ghr.nlm.nih.gov/ghr/visit;jsessionid=FEA19E2B34E1AC7D1EF678890C72A31D?uid=7bH4sIAAAAAAAAAEvPyS8uTiyqjI-Pr8koKSmw0tcvLy.Xy8vJ1cvLzNBLzy.Tz01NycnMSy3IKS3WzwWRKZnJJZn5eUBdehkluTkAldcvi0MAAAA_) by Merriam-Webster Inc.



Lister Hill National Center for Biomedical Communicati

ㆍDNA sequence를 computer가 분석할 때 6개의 reading frame으로 분석

ㆍ대장균 유전자의 평균 codon의 길이는 317, 효모는 483, 사람은 450 임.

ㆍORF scanning 에 있어서 유전자로 볼 수 있는 ORF 크기는 100 codon 이상인 것을 유전자로 간주함.

■고등 진핵세포에서 ORF scanning은 비 효과적

ㆍ진핵세포 게놈의 62%가 intergenic 임.

ㆍ암호부위가 intron에 의해 분절되어 있음

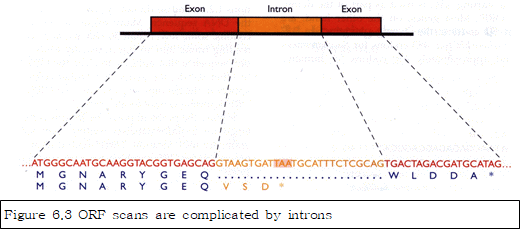
ㆍ이러한 문제점을 극복하기 위해서 ORF scanning program을 수정

- codon bias를 감안 : 유기체에 따라서 자주 사용하는 codon이 있으므로, intron의 경우는 다른 codon usage를 나타내므로 배제 가능.

- exon-intron 접합부위 탐색 : intron 부위는 GT로 시작하여 PyPyPyPyPyPyNCAG로 끝나므로 확인 가능.

- Upstream regulatory element : 유전자의 시작부위에 존재하는 조절부위를 확인. 조절부위의 종류가 다양하고 유전자 마다 서로 다른 조절부위를 가지므로 문제가 있음.

ㆍ위의 3가지 수정사항은 모든 진핵세포 게놈 분석에 적용되지만, 유기체에 따른 특징을 감안하여 다른 수정이 가능하다. 예를들면, 척추동믈은 많은 유전자의 상류에 1kb에 해당하는 CpG island가 있다는 점이다. 인간의 경우 40-50% 가 상류 CpG island가 있음.

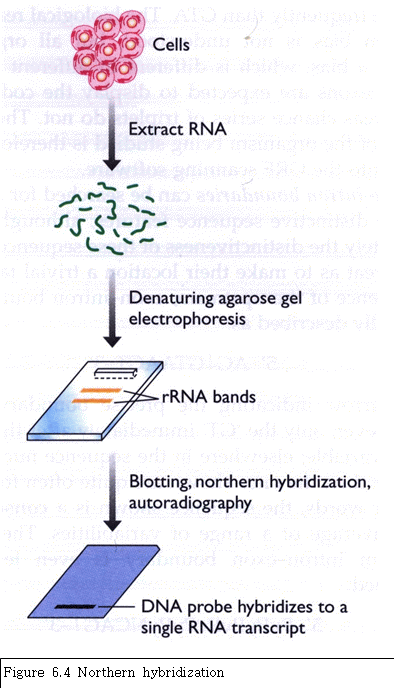


■Homology 탐색

ㆍ ORF 탐색 program은 95% 이상의 ORF를 찾아내지만 진정한 유전자(exon)인지 확인할 수 있는 방법으로 이미 알려진 유전자 database를 검색하여 homology가 있는지 조사함. homology가 있으면, 유전자일 가능성이 큼.

6.1.2유전자를 찾기위한 실험적 방법

ㆍ대부분의 실험적 방법은 DNA 분자 자체를 대상으로 하지 않고 전사된 RNA를 대상으로 하고 있다.

ㆍ모든 유전자는 RNA로 전사되어지고 전사된 RNA는 intron이 제거되고 exon으로만 구성되므로 RNA를 대상으로 조사를 하면 DNA 상의 exon의 위치를 찾을 수 있다.

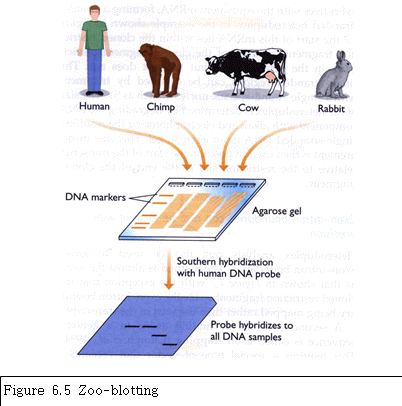
ㆍ문제는 전사된 RNA는 실제의 암호부위보다 길다는 점이다. 5'과 3'에 비 암호부위가 포함되어 있다. 따라서, RNA는 암호부위의 시작과 끝을 정확하게 말해주지 못한다.

■Hybridization test로 전사되어지는 DNA를 찾을 수 있다.

ㆍNorthern blot analysis : RNA를 전기영동하고 조사하고자 하는 DNA probe로 hybridization 하면, 그 DNA에 들어있는 유전자 수를 알 수 있다. 그러나, 다음과 같은 단점이 있다.

-한 유전자에서 서로 다른 크기의 RNA가 있을 수 있고, 혹은, 그 유전자가 mutigene family이면 에러가 발생함.

-어떤 조직 혹은 기관에서 분리한 RNA에는 모든 유전자의 RNA가 포함되어 있지 않으며, 설령 유기체 전부에서 RNA를 분리하더라도 발현이 적은 유전자는 탐지되지 않을 수도 있다.

ㆍ발현이 적게되거나 특정한 조직에서만 발현되는 유전자에 대해서는 다른 동물의 RNA를 이용한 소위 ZOO-blotting을 이용하면 어느 정도 위 문제점을 피할 수 있다.

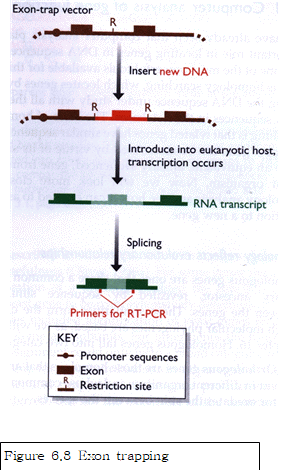
■cDNA sequencing으로 DNA 상에 유전자 mapping이 가능함

ㆍnorthern과 zoo-blotting은 DNA 상에 유전자의 존재 유무를 알 수 있으나, 위치에 관한 정보는 알려주지 않는다.

ㆍcDNA sequencing은 DNA 상의 유전자의 위치를 알 수 있게 해준다.

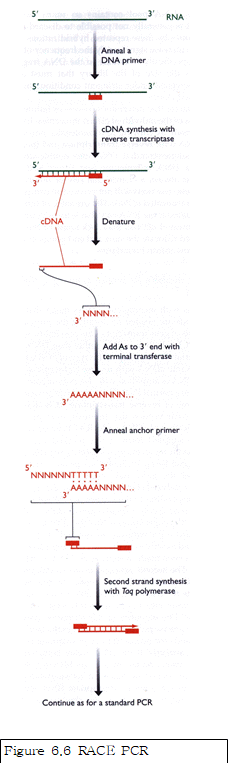
ㆍcDNA를 cloning 하기 위해서는 mRNA로부터 cDNA를 합성하고 cDNA library를 만들어 screening에 사용해야 한다.

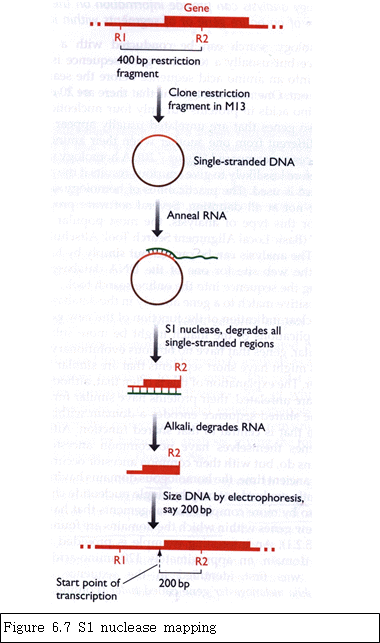
ㆍ문제점은 full-length cDNA 합성이 어렵다는 것이며, 그로 인하여 유전자의 양쪽 끝부분을 알아내는데 문제가 있다.



■ RNA 의 끝을 알아내는 법

ㆍRT-PCR의 일종으로 RACE PCR법을 이용(Fig 6.6)



ㆍheteroduplex analysis(Fig. 6.7)

■Exon-intron boundary 위치

ㆍheteroduplex analysis 이용가능

ㆍexon trapping(Fig. 6.8)

**6.2 각 유전자의 기능 결정**

ㆍ새로운 유전자를 찾은 다음 단계는 기능 연구임.

ㆍ실험적 방법과 computer를 이용한 방법이 있음

6.2.1 computer 에 의한 기능분석

ㆍdatabase에 있는 유전자와 homoogy 탐색을 통해 비슷한 유전자나 새로운 유전자를 찾아낼 수 있음

ㆍhomology 탐색으로 유전자 기능도 알 수 있음

■homology는 진화적 관련성을 나타낸다.

ㆍhomology가 있다면, 공동 조상 유전자에서 진화되었다고 볼 수 있다.

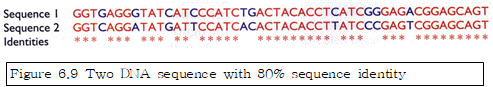
ㆍhomologous gene은 다음과 같이 두 가지 group 이 있다.

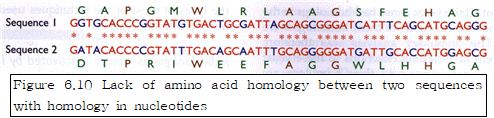
- orthologous gene : 다른 유기체에 있는 유전자로서 진화적으로 분리되기 이전의 조상에게서 유래된 유전자임

- paralogous gene : 같은 유기체에 존재하는 유전자로서 진화적 조상으로부터 유래되지 않을 수도 있음, 예를들면 다가족 유전자(multigene family)

ㆍhomologous gene은 정확하게 동일한 염기서열을 갖고 있지 않고, 돌연변이에 의해 다른 sequence를 갖고 있다. 이것은 진화적 관련성을 나타내며, 동일하거나, 같은 기능을 갖고 있음을 추정할 수 있다.

ㆍNucleotide sequence similarity와 homology의 차이 : 80% nucleotide identity(similarity)를 80%homologous 로 표시해서는 안됨(Fig. 6.9). 따라서 nucleotide sequence에 % homology는 별 의미가 없다.

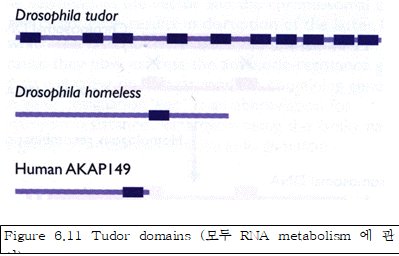




■Homology 탐색으로 유전자 기능에 관한 정보를 얻을 수 있다.

ㆍnucleotide sequence를 이용해서 homology search 하는 것 보다. 아미노산으로 번역한 다음 homology를 search 하는 것이 더 정확함.

ㆍdatabase 의 어떤 유전자와 homology가 있다면, 새 유전자의 기능이 동일 하거나 아니면 homology의 정도에 따라서 기능이 조금 관련성이 있을 수 있다. 진화적 관련성이 없더라도 짧은 부분에 있어서 유사할 수도 있다. 그 이유는 단백질의 domain이 같을 수 있으며, 그것은 비슷한 기능을 의미하는 것이다.



■효모 게놈 project에서 homology 분석

ㆍ효모 유전자는 약 6000개이었으며, 그 중 30%는 이미 고전적 유전학에 의해 알려진 것이고, 나머지 70%를 homology 분석하였다.

- homology search 결과 30 %는 유전자 기능이 밝혀졌고, 그 중 반은 database 유전자와 동일한 유전자이었으며, 나머지 반은 유사성이 떨어지거나 부분적으로 유사한 유전자로 밝혀졌다.

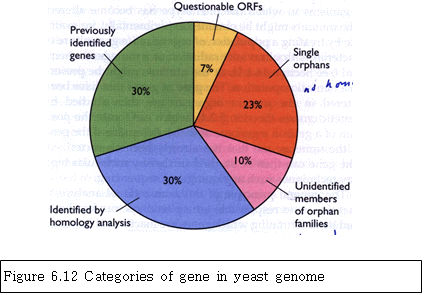
- 약 10%는 database에 상동성 유전자가 있으나 그 기능을 모르는 것으로 이 유전자들은 orphan family라고 함

- 나머지 30%는 database에 상동성이 없는 유전자로서 unique 한 것으로, single orphan이라함.

6.2.2 실험적 방법에 의한 기능조사

ㆍhomology search가 기능의 조사에 모든 것을 해결하는 것이 아님.

ㆍ실험적 방밥으로 보충이 되어야 하며, sequence로부터 기능을 밝혀야하므로, 표현형으로부터 유전자를 찾는 전통적인 방법이 아니라 gene으로부터 표현형을 찾아서 그 기능을 밝혀야 함.



■유전자 불활성화에 의한 기능분석

ㆍ고전적인 유전학적 분석에서는 표현형은 돌연변이체를 만들고 교배에 의해 그 유전자의 위치를 알아낸다

ㆍsequence 된 유전자가의 경우도 같은 원리로 그 유전자를 돌연변이시켜 그 표현형을 관찰함으로서 그 기능을 알 수 있다.

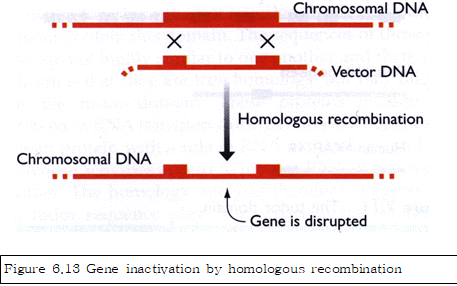
■각 유전자는 homologous recombination에 의해 불활성화 될 수 있다.(Fig. 6.13, Fig. 6.14)

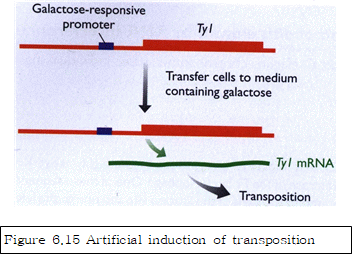
ㆍ유전자 clone의 한 부분을 항생제 유전자로 치환한 DAN를 이용한 homologous recombination을 수행하여 유전자 파괴함

ㆍES cell에서 homologous recombination를 수행하여 embryo에 주임하여 knock-out mouse 생산하여 표현형 관찰.

■homologous recombination를 이용하지 않은 유전자 불활성화

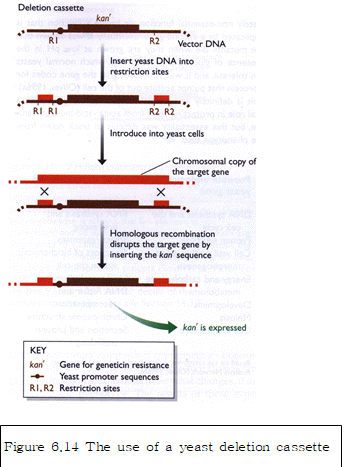
ㆍ유전자를 transposon으로 파괴할 수 있음 : recombinant technique를 이용하면 더 높은 빈도로 transposition을 유도할 수 있음(Fig. 6.15).





ㆍ문제는 transposon이 우리가 원하는 특정한 유전자를 파괴한다는 보장이 없으며 random 하게 유전자를 파괴한다는 것이다. 따라서 screening 과정을 거쳐서 필요한 돌연변이체를 골라야 한다.

ㆍtransposon에 의한 유전자 파괴는 전체적인 돌연변이 유도에 유용함



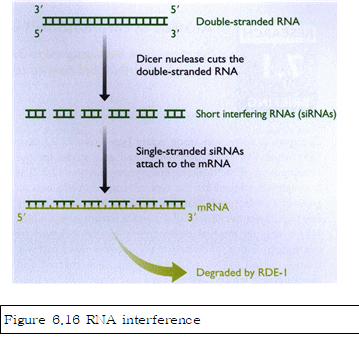
ㆍRNA interference라고 하는 전혀 새로운 방법을 사용가능함. 유전자의 mRNA의 한 부분과 동일한 염기배열을 갖는 21 bp의 작은 double stranded RNA 분자를 세포에 도입하면 그 유전자의 전사가 억제됨. 이 방법은 진핵세포에만 적용됨(Fig. 6.16)

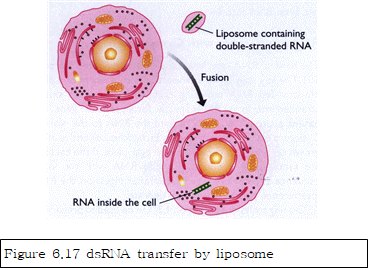
■Overexpression으로 유전자 기능 밝히기

ㆍ유전자 불활성화는 유전자 기능 손실(loss of gene function)을 이용하여 유전자 기능을 밝히는 방법

ㆍOverexpression은 정상보다 많은 양의 유전자 발현을 통하여 유전자의 기능 획득(gain of function)을 통하여 그 기능을 알아내는 것

ㆍOverexpress하는 법 : muticopy vector의 강력한 promoter 뒤에 암호부위를 cloning 하여 host에 도입.

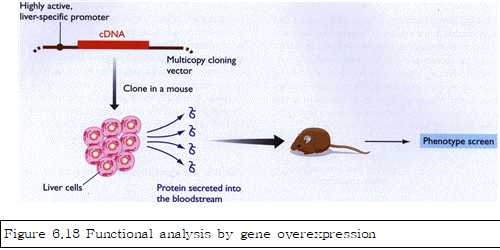
ㆍtransgenic mouse의 생산에 의한 유전자의 overexpression 에 의해서도 기능을 알 수 있다.



6.2.3 unknown gene에서 만들어진 단백질 기능의 상세한 연구

ㆍinactivation 과 overexpression은 새로은 유전자의 기능 결정을 위한 주된 기술

ㆍ유전자의 기능에 대한 부가적인 정보를 알 수 있는 방법이 있다.



■직접적 돌연변이에 의한 상세 기능 연구

ㆍinactivation 과 overexpression으로 유전자의 일반적인 기능은 알 수 있지만, 상세한 기능은 알 수 없다.

ㆍ단백질 중 어떤 특정한 아미노산의 기능을 알고 싶을 때는 그 codon만 돌연변이 시키면 됨.

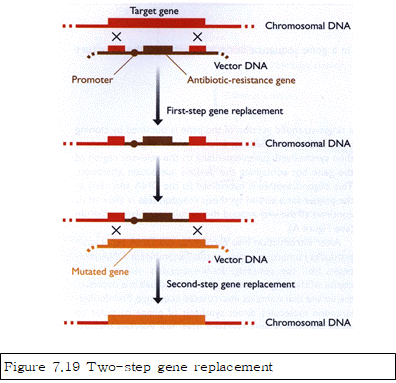
ㆍ이러한 목적을 위해서는 site-directed mutagenesis 기술을 사용할 수 있음

ㆍ이 방법은 또한 소위 protein engineering 이라는 분야에도 이용됨.

ㆍ돌연변이 후에는 돌연변이된 유전자를 homologous recombination에 의하여 이미 존재하는 유전자를 치환시켜서 그 돌연변이의 효과를 보아야 한다.

ㆍhomologous recombination 에 의해 유전자가 치환되었는지를 확인하기 위하여 marker gene(예, 항생제 내성 marker)을 돌연변이된 유전자옆에 붙여서 도입함.

ㆍmarker gene의 영향을 없애기 위하여 two-step gene replacement 법을 사용하여 marker gene 제거(Fig. 6.19)



■Reporter gene과 면역세포화학법으로 유전자 발현 탐지하기

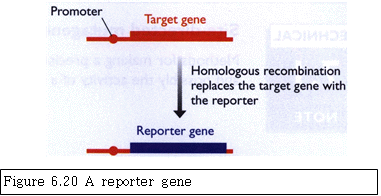
ㆍ유전자의 발현 장소에 대한 정보가 그 유전자의 기능을 밝히는 중요한 요소가 될 수 있다.

ㆍ한 유기체 내에서 유전자의 발현 pattern은 reporter gene을 통해 쉽게 파악할 수 있다. reporter gene은 색깔이나 형광등으로 표현될 수 있다.

ㆍreporter gene은 원래 test 유전자의 발현 조절부위를 파괴하지 않도록 삽입되어져야 하며, test 유전자의 ORF대신 치환시켜서 사용한다.

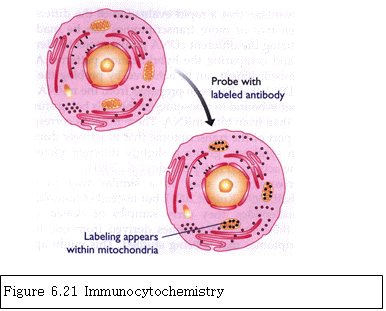
ㆍ때로는 test 유전자에서 만들어진 단백질이 어떤 세포에서 존재하는지 알아내는 것이 필요함

ㆍ또한 세포내에서 그 단백질이 어디 위치하는지도 그 단백질의 기능에 대한 중요한 정보를 얻는데 도움이 됨.



ㆍ이러한 경우 reporter gene은 별로 도움이 되지 않는다. 왜냐하면, reporter gene은 세포내 단백질의 targeting에 대한 필요한 신호는 갖고 있지 않기 때문이다.

ㆍ이러한 경우 단백질 그 자체를 탐지하는 것이 한 방법이며, immunocytochemistry를 이용하면 됨(Fig. 6.21)

ㆍ단백질에 대한 항체를 형광물질로 표지하여 사용함으로서 광학현미경으로 저 해상도에서 관찰가능하며, 고해상도의 전자현미경에서 관찰할려면 colloidal gold 등과 같이 electron dense label로 표지한 항체를 사용할 수 있다.

**6.3 게놈 활성을 전체적으로 연구하기**

ㆍ서로 다른 조직 혹은 서로 다른 발생단계, 혹은 병적상태 등의 transcriptome 혹은 proteome 등을 조사하여, 그 조직의 중요 특징을 이해함.

6.3.1 transcriptome 연구

ㆍtranscriptome은 특정한 시기에 특정한 세포에서의 mRNA 전체를 말함.

ㆍtranscriptome의 특징 연구는 mRNA 들의 동정과 그들의 상대적인 비율 등을 조사하는 것이다.

■SAGE 에 의한 transcriptome의 성분분석

ㆍtranscriptome을 조사하는 가장 직접적인 방법은 mRNA로부터 cDNA를 합성하여 library를 만들고 각 cDNA clone들을 sequencing하여 서로 그 빈도를 비교하는 것이다. 그러나 이 방법은 너무 힘든 방법이다.

ㆍ쉽게 transcriptome을 분석할 수 있는 방법으로 SAGE를 들 수 있다.

ㆍSAGE법에서는 cDNA의 전체서열을 사용하지 않고 cDNA 염기서열중 각 cDNA에 특이적으로 존재하는12bp길이의 염기서열을 이용하는 것이다.

ㆍ특정한 12bp의 염기서열이 나타날 확률은 412 bp에 한번이다. 이것은 진핵세포 mRNA의 평균 길이를 1500 nt라고 할 때, 11000 mRNA에 한번 존재한다는 것을 의미한다.

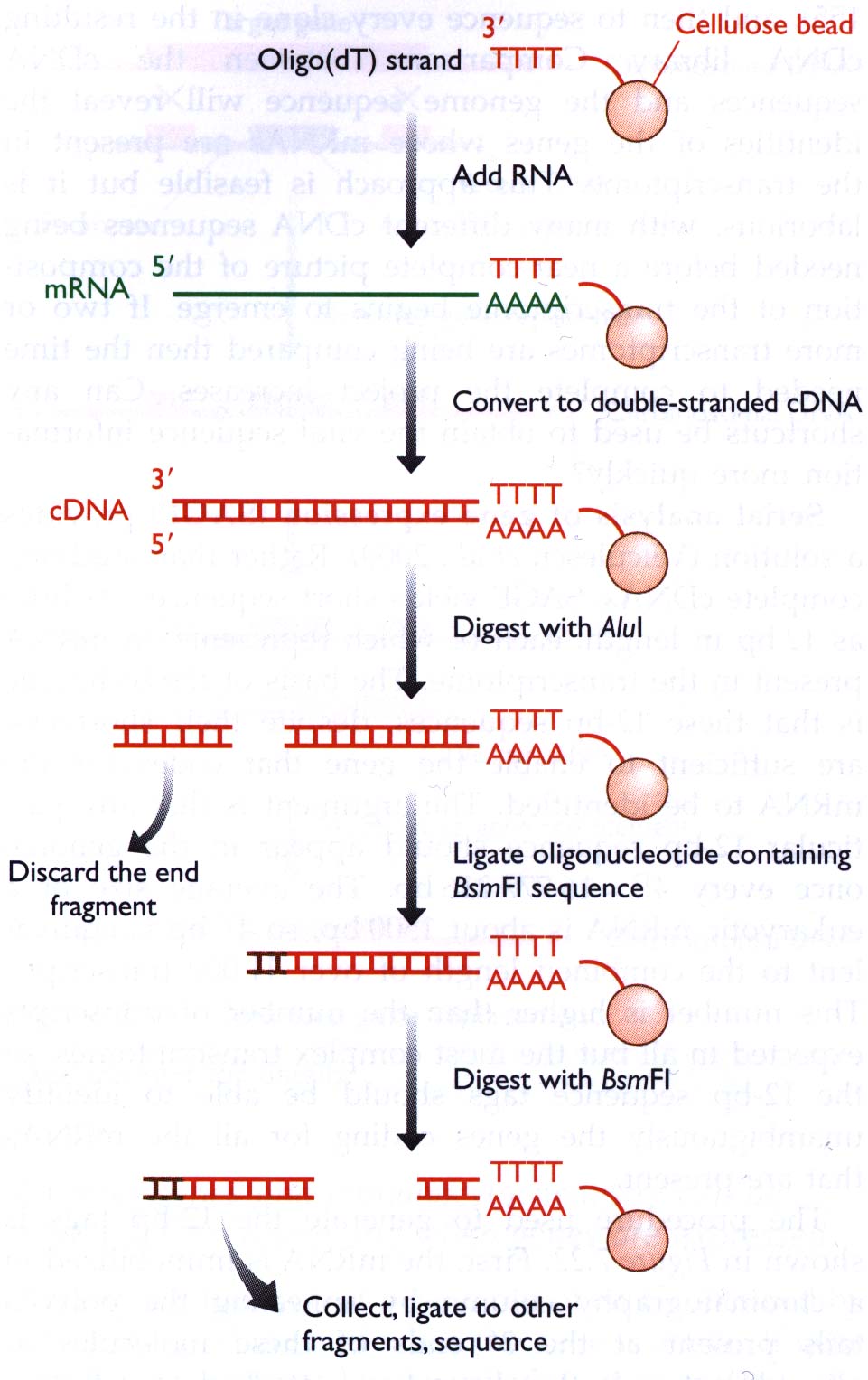
ㆍ기질에 고정된 oligo dT에 mRNA의 polyA tail을 anealing 시켜서 고정하고 dsDNA를 합성한다. 4-cut restriction enzyme으로 절단하여 oligo dT에 고정된 cDNA 부분만을 회수한다.

ㆍBsmF1 제한효소 site가 있는 oligonucleotide를 cDNA의 제한효소절단 부위에 ligation 하고, BsmF1으로 절단한다.

ㆍBsmF1은 그 인식부위로부터 downstream 쪽으로 10-14 bp 위치를 절단한다. 이때 각 cDNA끝에서 12 bp 씩 분리되어져 나오는 작은 cDNA 조각(tag이라고 함)을 head-to-tail로 길게 연결하여 cloning하고 sequencing 한다.

ㆍ각 tag은 한 종류의 mRNA로부터 나왔다고 가정할 수 있으며, 각 tag의 비는 그 mRNA의 비를 의미한다.

ㆍtag의 염기배열과 mRNA 혹은 genome의 염기배열을 비교하면 각 tag이 나타내는 mRNA를 동정할 수 있다.



■chip과 microarray를 이용한 transcriptome의 연구

ㆍ각 유전자를 대표할 수 있는 각 oligonucleotide로 구성된 chip을 이용할 수 있다.

ㆍmRNA로부터 합성한 cDNA를 표지하여 chip에 hybridization 시킴

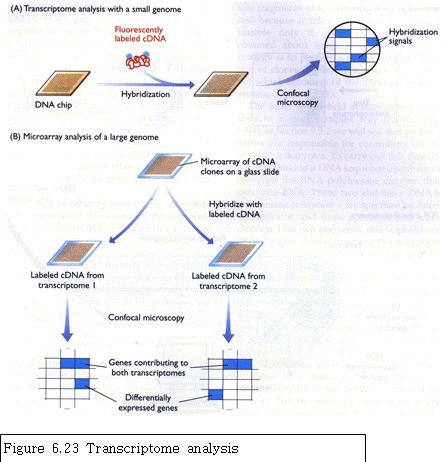
ㆍSAGE 보다 빠른 시간내에 서로다른 transcriptome을 비교분석할 수 있는

장점이 있음

Figure 6.22 SAGE

ㆍribosome에 bound된 mRNA 로부터 합성된 cDNA를 표지하여 hybridization 하면 좀더 다른 형태의 genome activity를 관찰할 수 있다.

ㆍmicroarray는 oligonucleotide 대신 cloned 된 DNA를 고정하여 사용



6.3.2 proteome 연구

ㆍproteome은 genome과 세포의 생화학적 능력을 연결시켜주는 것이라고 할 수 있으므로 proteome 연구는 중요함.

ㆍproteome의 차이를 알아내는 것은 genome 활성이 어떻게 기능을 발휘하는지 이해하는데 도움이 됨.

ㆍtranscriptome의 비교와 proteome의 비교는 항상 일치하는 것은 아님. 그 이유는

- 모든 mRNA가 어떤 특정 시기에 전부 활발하게 번역되어지는 것은 아니기 때문.

- 단백질의 양은 새로운 단백질의 합성과 이미 존재하는 단백질의 분해의 결과임

ㆍ따라서 proteome 연구는 genome 활성을 완전히 밝히기 위해서 필요함

■세포내 단백질양 측정법

ㆍproteome을 연구하는 방법을 총체적으로 proteomics라고 하며, 단백질 전기영동과 질량분석 등을 포함하고 있다.

ㆍ우선 순수한 단백질 시료를 준비

ㆍPAGE가 단백질을 분리하는 가장 기본적인 방법이지만 많은 단백질이 전기영동에의해 서로 분리되지 않는다.

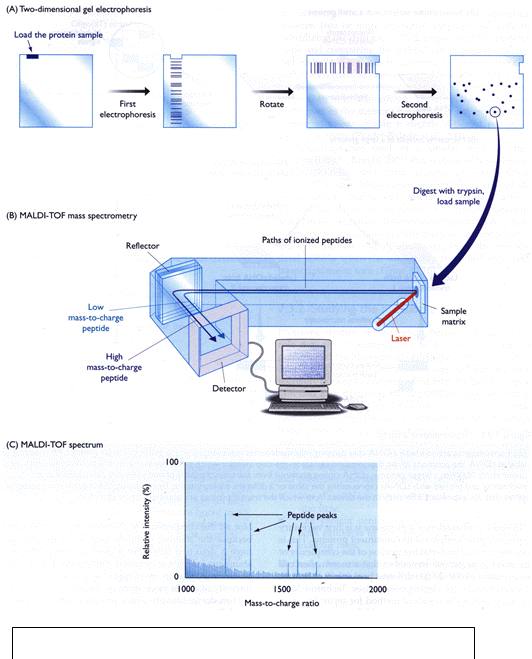
ㆍ이러한 점을 보완하기 위해서 2차원전기영동 사용함

ㆍ2차원전기영동에서 두 proteome의 차이는 spot의 위치와 진하기 정도에 따라 비교되어 진다.

ㆍspot의 단백질을 동정하기위해서 질량분석기를 사용함.

※질량분석기의 원리를 공부해 봅시다

ㆍ일반적인 질량분석기는 단백질과 같은 큰 분자의 분자량 측정이 어렵기 때문에 MALDI-TOF 라고하는 기술을 사용하면 50 a.a 까지의 peptide는 측정할 수 있음.(Matrix-assisted laser desorption/ionisation-time of flight mass spectrometry (MALDI-TOF MS)



ㆍ2차원 전기영동한 gel의 단백질 spot을 trypsin 등으로 소화 하면 짭은 peptide가 생성되며, MALDI-TOF spectrum을 genome sequence와 관련지을 수 있다.

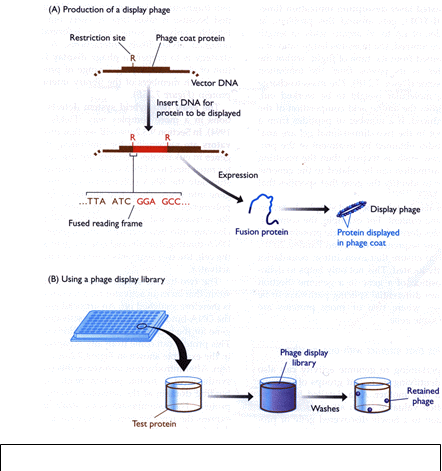
ㆍproteomics는 한 종류의 단백질내의 peptide 구성을 조사할 때도 유용하게 사용할 수 있다. 유전자의 exon-intron의 구성, splicing의 pattern 등도 알 수 있다.

■서로 상호작용하는 단백질 동정

ㆍgenome 활성과 관련된 중요한 자료로서 서로 상호작용하는 단백질 쌍 혹은 group을 알아내는 것을 들 수 있다. 이러한 자료는 새로이 발견된 단백질의 기능을 찾는데 긴요하다.

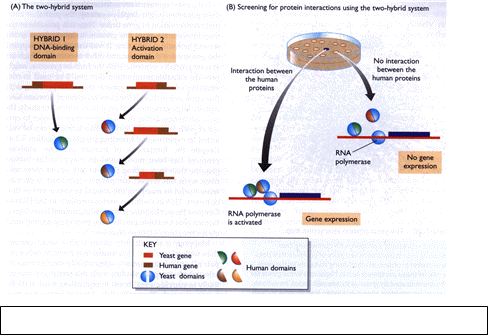
ㆍ단백질들간의 상호작용을 연결시키면 protein interaction map을 그릴 수 있다.

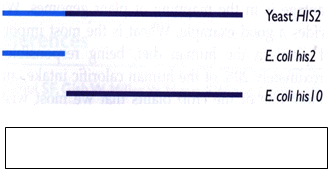
ㆍ단백질간의 상호작용을 밝힐 수 있는 몇 가지 방법이 있다. 그 중에서 phage display와 yeast two-hybrid system이 가장 중요한 방법이라고 할 수 있다.



ㆍbacteriophage λ 혹은 M13 등의 coat protein 유전자에 test protein 유전자를 융합시켜 발현 하면 phage의 표면에 test protein이 표출될 것임

ㆍ표출된 단백질과 다른 단백질과의 상호작용은 이미 일려진 단백질과의 상호작용을 조사할 수 도 있고, 혹은 phage display library를 만들면, library의 어떤 단백질이 test protein과 상호작용하는지도 조사할 수도 있다(Fig. 6.25A. 6.25B)

ㆍYeast two-hybrid system(Fig. 6.26)



■단백질상호작용지도

ㆍproteome 내에서의 모든 상호작용을 표시하는 것을 말함(Fig. 6.28.

ㆍ주로 yeast two-hybrid를 이용해 만듬.

ㆍ한 유기체에서 서로 다른 단백질이 다른 유기체에서는 그 두 단백질이 융합된 형태로 나타날 때, 그 두 단백질이 상호작용할 경우가 있다.

