제 5 장 Sequencing genomes

**학습목표**

1. sequencing 방법 이해

2. 자동염기서열 분석의 이해 및 genome 연구에 있어서 의미

3. shotgun 방법, whole-genome shotgun, clone contig 방법의 강점과 한계점

4. shotgun법에 의한 세균 genome sequencing 의 이해

5. clone contig 만드는 법

6. whole-genome shotgun 법에 의한 genome sequencing 이해

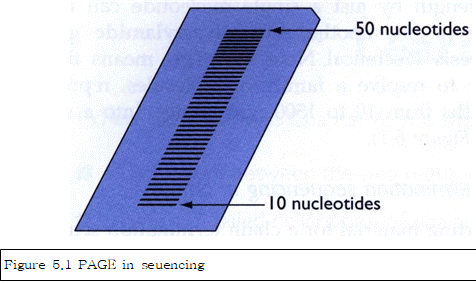
**5.1 DNA sequencing 법**

ㆍ Genome project의 궁극적인 목표는 대상 genome의 완전한 sequencing이다.

ㆍ따라서, sequencing 기술은 가장 중요한 부분이다.

ㆍ두가지 기본적인 방법이 개발되었다.

- chain termination method : 단선 DNA 분자의 염기서열을 상보적인 DNA의 효소적 합성할 때 특정 염기에서 반응이 중지되게 하는 방법으로 결정

- chemical degradation method : 특정한 염기에서 DNA 가닥을 자를 수 있는 chemical을 사용

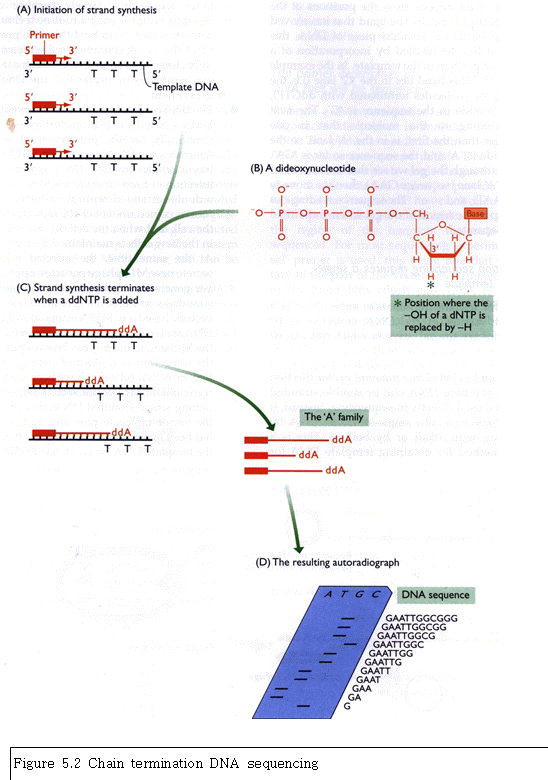
5.1.1 Chain termination

ㆍ단선DNA는 한 염기의 길이만 달라도 전기영동에 의해 polyacrylamide gel에서 서로 분리될 수 있다(10-1500 nt 범위).

■Chain termination 의 개요

ㆍ단선 DNA 가닥을 만드는 것으로 시작함

ㆍ단선 DNA에 oligonucleotide(합성한 것)을 anealing하여 primer로 사용되게 함.



ㆍDNA polymerase로 DNA 합성을 시행함. 4가 종류의 dNTP가 필요함.

ㆍ4가지 종류의 dNTP와 같이 ddNTP를 넣어서 반응을 시키면 ddNTP가 합성되어 들어감으로서 합성이 중지됨

예) dATP+dCTP+dGTP+dTTP+ddATP를 섞은 반응에서는 dATP가 들어가야 할 위치에 일정한 비율로 ddATP가 들어가서 반응이 중지되며, 합성된 DNA 가닥의 길이는 A의 위치에 해당하는 길이와 같아짐. 동일한 방법으로 ddCTP, ddGTP, ddTTP를 이용하면 각 반응에서 생성된 DNA의 길이는 각기 서로 다르게 되며, 전기영동에 의해 합성된 DNA 길이를 비교하면 A,C,G,T의 순서를 알 수 있음.

ㆍSequencing 에 사용하는 주형의 양은 매우 적은 양이며, 그 주형에 상보적으로 합성된 DNA 의 양도 작다. 따라서 전기영동한 후 band를 관찰하기가 쉽지 않다. band의 위치를 쉽게 알 수 있는 방법은 사용한 oligonucleotide의 끝을 방사성동위원소로 표지하여 사용하면 이 primer에서 합성된 DNA band는 X-ray film에 노출시킴으로서 쉽게 확인할 수 있다.

ㆍ다른 방법으로는 primer로 사용한 oligonucleotide를 방사성동위원소로 표지하는 것이 아니라, dNTP 중의 하나를 방사성동위원소로 표지하여 사용하는 것이다.

■단선 DNA 주형 만들기

ㆍplasmid vector에 cloning 한 경우 : alkali 혹은 끓여서 변성 후 사용

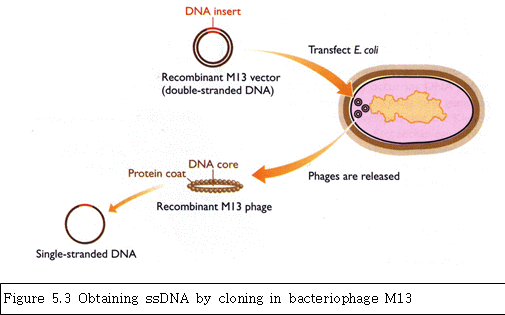
ㆍbacteriophage M13 vector에 cloning 한 경우 : phage particle을 분리하고 single stranded DNA를 분리하여 사용

※공부합시다

single stranded bacteriophage M13 의 생활사를 공부해봅시다.

ㆍphagemid에 cloning 한 경우 : phagemid 란 p;asmid 복제원점외이 single-stranded bacteriophage(M13 등)의 복제원점을 갖고 있는 plasmid이다. 이러한 plasmid를 갖고 있는 대장균을 helper phage(M13의 돌연변이체)를 감염하면 helper phage particle과 동시에 plasmid를 포함하는 particle도 분비하게 된다. 이 particle을 분리하여 DNA를 순수분리하면 단선 DNA를 얻을 수 있다.

ㆍPCR을 이용하는 방법 : PCR primer 중 한 종류의 primer 끝에 magnetic baed를 부착하여 증폭한 후에 자석으로 분리



■sequencing 할 부분의 위치에 맞게 primer 위치를 부착 위치를 선택한다.

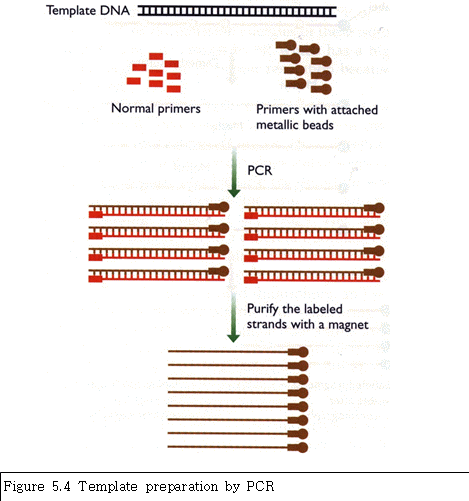
ㆍUniversal primer : cloning 에 사용하는 vector에 부착하는 primer로서 multiple cloning site 주위에 부착할 수 있는 primer임. multiple cloning site에 cloning 된 insert DNA를 양쪽에서 sequencing 할 때 사용될 수 있음. primer site 로부터 750 nt 정도는 읽을 수 있음

ㆍinternal primer : 크기가 너무 큰 insert DNA는 universal primer 로 sequencing을 끝내기가 곤란하므로, insert 내부에 부착하는 primer를 사용해야 함. universal primer로 sequencing 하여 내부의 염기배열을 알아낸 다음 그 염기배열을 이용하여 internal primer를 합성하여 사용

■PCR을 이용한 sequencing

ㆍ내열성 DNA polymerase를 이용한 PCR 반응을 이용하여 소위 thermal cycle sequencing을 개발하였음.

ㆍdouble stranded DNA를 주형으로 사용



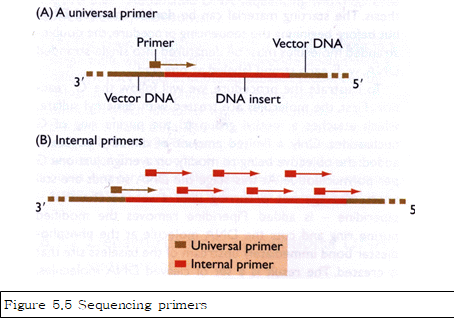
ㆍ소량의 주형 DNA가 필요하므로 반드시 cloning 하지 않아도 됨

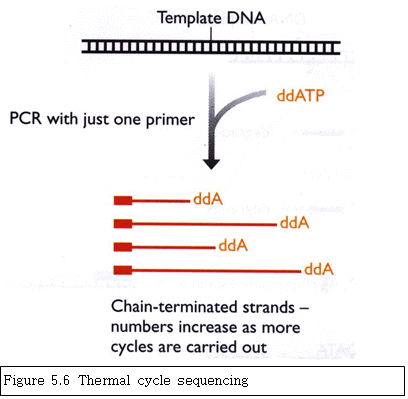
ㆍ하나의 primer 만을 이용하므로 한 쪽 방향으로만 합성된 DNA가 축적됨

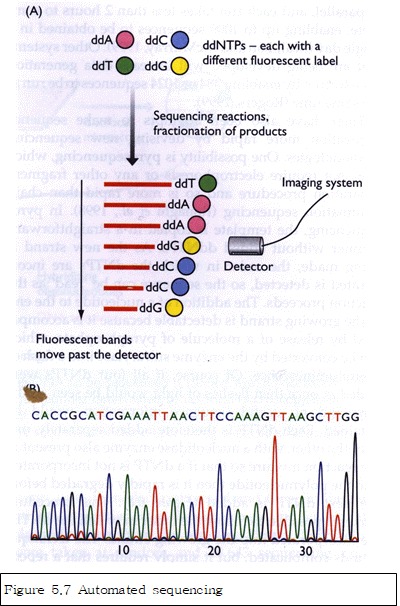
ㆍpolyacrylamide gel에서 전기영동하여 sequence 확인

■형광표지 primer를 사용하여 자동염기서열분석이 가능함

ㆍ일반적으로 chain termination 방법에서는 방사성동위원소(32P 혹은 35S)로 primer를 표지하여 사용하며 autoradiography에 의해 band를 볼 수 있음

ㆍ서로 다른 형광표지를 ddA, ddC, ddG, ddT에 각기 표지하여 사용하면, 한 tube에서 4가지 reaction을 반응할 수 있고, 형관 탐지기를 이용하여 한 lane에 전기 영동하여 sequence를 읽을 수 있다. 전기 영동하여 band가 이동하여 탐지기를 지날 때 형광의 종류를 탐지하여 sequence를 읽게 됨



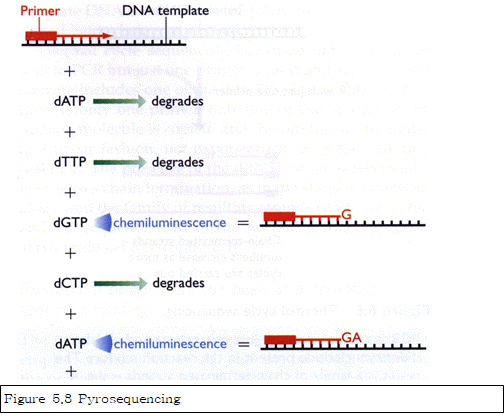


■특이한 sequencing 방법의 개발

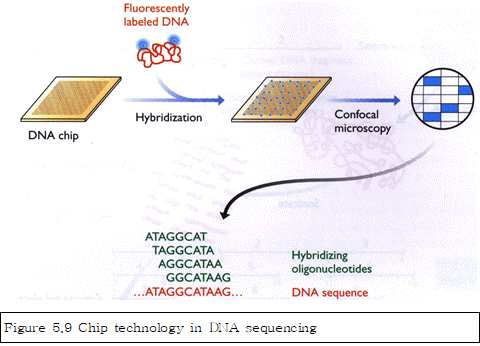
ㆍ자동염기서열분석을 사용하더라도 한 primer 로부터 읽을 수 있는 염기의 개수는 수백 개에 불과함

ㆍ단시간에 많은 수의 squencing을 하기해 개발된 방법 중의 하나가 capillary를 이용한 자동염기서열 분석 장치이다. 알반적인 polyacrylamide gel이 아닌 capillary gel을 이용.

ㆍpyrosequencing : 전기영동이 필요 없음. ddNTP를 사용하지 않고 주형에 따라 DNA 합성. 한번에 한 종류의 dNTP를 첨가하고 첨가한 dNTP가 DNA로 합성되어 들어가면 PP(pyrophosphate)가 방출됨. pyrophosphate는 sulfurylase 에 의해 빛을 내게 됨. 반응하고 남은 dNTP는 반응물에 들어 있는 nucleotidase 에 의해 파괴됨



ㆍDNA chip을 이용한 방법의 개발 : 서로 다른 oligonucletide의 배열(array)을 갖고 있는 DNA chip에 sequencing 대상 DNA 분자를 hybridization 시킴. hybridization 된 oligonucleotide를 찾아내어 그 sequence를 서로 overlapping 하여 연결 하면 대상 DNA의 sequence를 알 수 있음. 문제점은 많은 array가 필요하다는 것임. 8nt 의 oilgonucleotide를 이용하면 65536 가지의 oligonucleotide로 된 chip을 사용하여야 하고 256 bp의 DNA를 sequencing 할 수 있다. 10 mer를 이용하면 1048576개의 array가 필요하며, 1 kb를 sequencing 할 수 있다.



**5.2 연속적인 DNA sequence의 조립**

ㆍ작은 DNA 조각들의 sequence를 이용하여 어떻게 수십 Mb의 염색체의 sequence로 만들 수 있을까?

ㆍ비교적 작은 게놈을 갖고 있는 세균은 whole-genome shotgun sequencing 으로 가능하나, 큰 게놈을 갖고 있는 진핵세포의 경우는 에러가 발생함.

ㆍ그러나 whole-genome shotgun 과 mapping 자료를 이용하면 DNA 조각으로부터 나온 sequence를 조립할 수 있음.

ㆍ게놈 map 상에서 그 위치가 알려진 DNA 조각을 sequencing 하여 조립하여 clone contig를 만드는 방법을 사용하면 더 정확함.

5.2.1 Shotgun 에 의한 sequence assembly

가장 쉽게 생각할 수 있는 방법은 작은 DNA 조각을 sequencing 하여 그 염기서열을 분석하고, 겹치는 부분을 이용하여 잇는 것이다.

■shotgun 법이 hemophilus influenza sequencing 에 위력을 발휘함

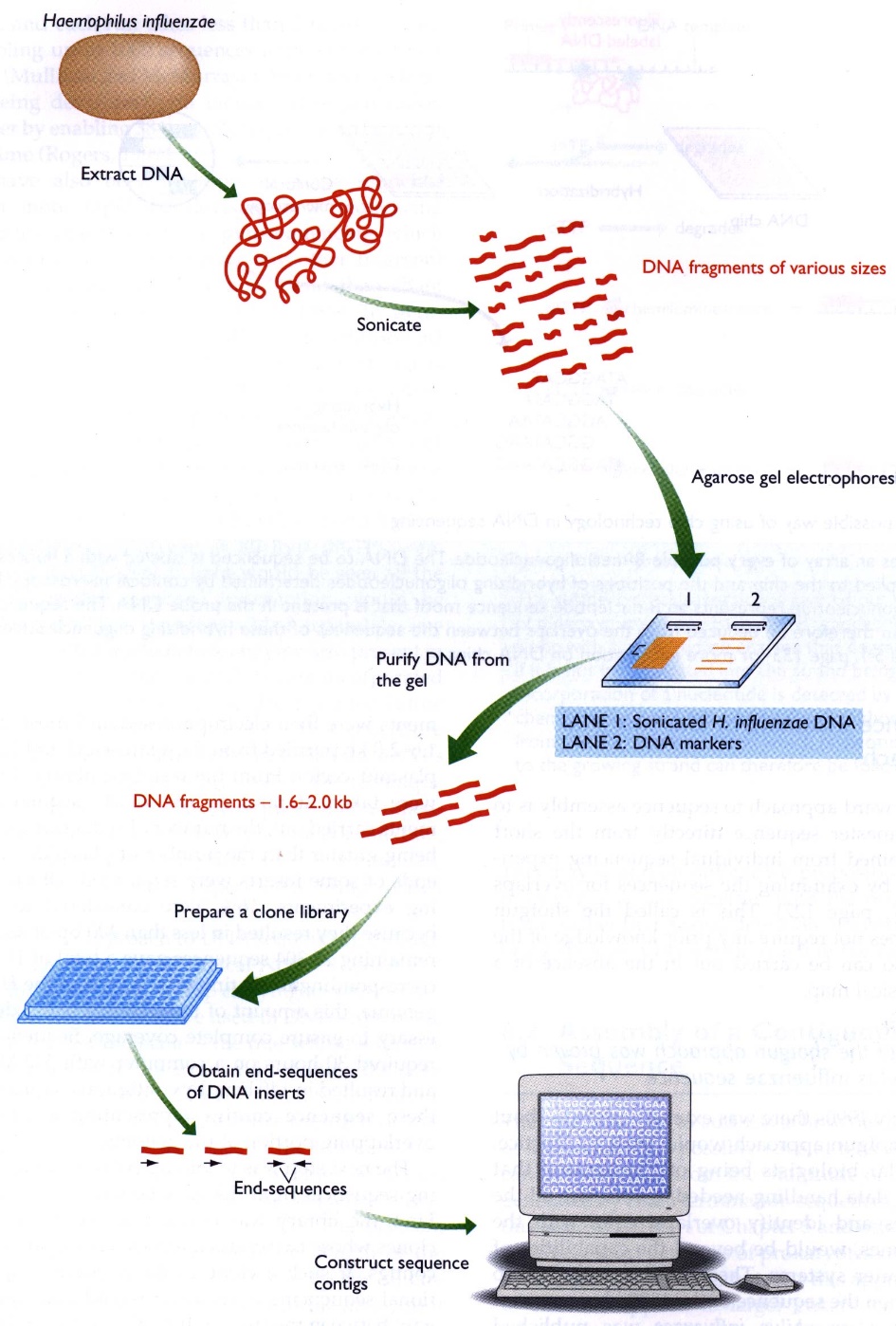
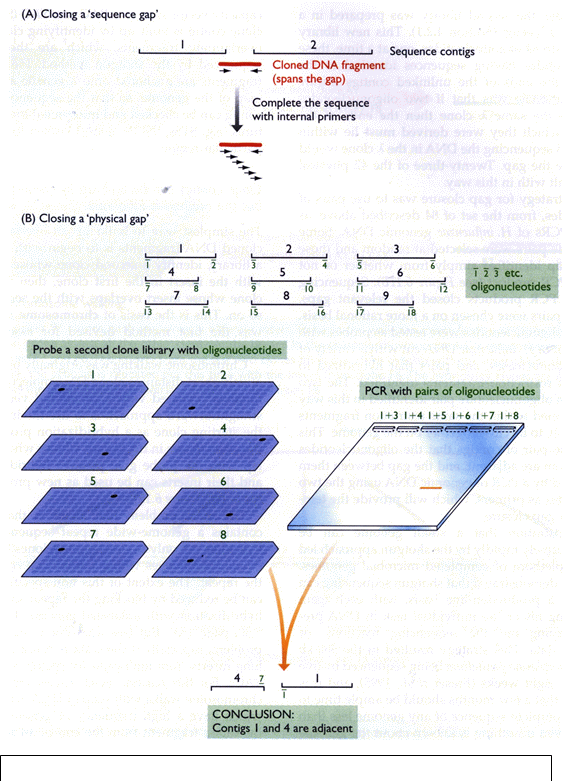


Figure 5.10 Shotgun approach Iin H. influenza genome project

ㆍ 가징 작은 genome이라 할지라도 작은 DNA 조각의 sequence 들을 비교하고 겹치는 부분울 이용하여 연결하는 것은 computer의 능력을 넘어서는 것이라는 의견이 있었으나, 1830 kb 의 H. influenza genome의 sequence가 shotgun에 의해 밝혀짐.

ㆍ 게놈을 sonication 에 의해 조각내고, 전기영동에 의해 1.6-2.0 kb를 분리하여 library를 만듬(19687 clones). 그중 28643개의 sequencing reaction을 수행하였다. 그 결가 11631485 bp(genome의 6배)가 결정되었고, computer로 30시간동안 분석하여 140개의 sequence contig를 만듬.

※Sequence contig와 clone contig의 차이점은?



ㆍ다음 각 sequence contig 사이의 gap에 해당하는 부분을 갖고있는 clone 들(99 clones)을 찾아내어 sequence를 조사하여 99 개의 gap을 매꿈. 나머지 42개의 gap은 library에 없었으며, 다른 종류의 vector(bacteriophage λ)에 만든 library로부터 clone들을 얻어 23개의 gap을 메꾸었음.

ㆍ나머지는 sequence contig의 말단부분에 해당하는 PCR primer를 design 하여 random하게 선택된 primer set 별로 PCR을 수행하여 PCR 산물이 나오는지 확인하고 PCR 산물을 sequencing하여 메꿈.

ㆍprimer set를 결정하는 방법으로 제한효소로 자른 게놈 DNA를 전기영동하고 oligonucleotide probe들을 이용한 southern blot analysis를 실시하고 같은 DNA 조각에 hybridization 하는 primer들을 한 set로하여 PCR을 수행함.

5.2.2 clone contig를 이용한 sequence assembly

ㆍ진핵세포 genome sequence를 얻는 전형적인 방법임

ㆍgenome을 제한효소 부분소화를 통하여 1.5Mb 정도로 자르고 BAC 혹은 YAC 와 같은 high capacity vector에 cloning함.

ㆍ겹치는 clone들을 이용해 clone contig을 만듬.

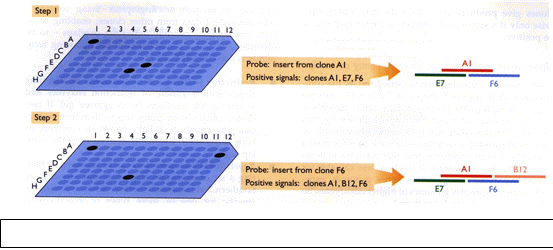
ㆍ그 다음 각각의 clone들을 shotgun법에 의해 sequencing함.

ㆍ 각 clone 들을 genetic 및 physical map에 맞추어 배열하고 sequence data를 이미 존재한다고 알려진 STS, SSLP gene 등으로 확인함

■ chromosome walking 에 의한 clone contig 형성 : 어려움

ㆍ chromosome walking는 원래 λ 혹은 cosmid library 등을 이용해서 짧은 거리를 염색체상에서 이동할 때 사용한 방법임

ㆍ한 clone의 insert DNA를 hybridization probe로 사용하여 겹치는 다른 clone과 연결함.



ㆍ주요문제점은 게놈 전체에 늘려있는 반복서열임. 반복서열이 probe에 있으면 겹치지 않는 clone에 hybridization 함

ㆍ반복서열을 이용한 prehybridization으로 어느정도 문제를 해결

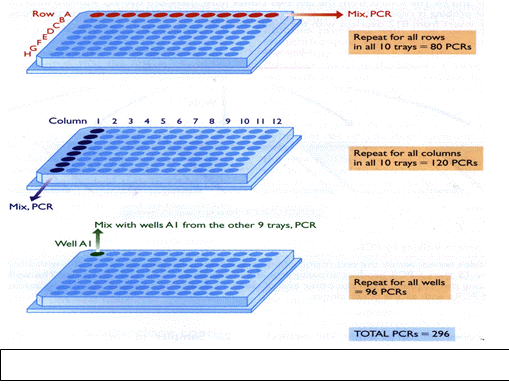
ㆍ큰 insert 전체를 probe로 사용하지 않고 그 끝부분의 DNA 조각을 probe로 사용면 반복서열이 포함되는 기회를 줄일 수 있다.

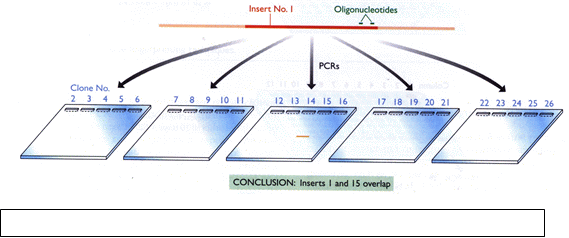
ㆍ만약 그 끝부분이 sequencing 되어 있다면 hybridization 대신 PCR로 겹치는 clone을 찾을 수 있으며 기대되는 PCR 산물이 나오는 clone을 찾음(Fig. 5.13)

ㆍ그 외에도 clone contig를 만드는 방법으로 각각의 clone을 PCR 하는 것보다 여러개의 clone들을 섞어서 PCR하는 소위 combinatorial PCR을 수행하면 된다(Fig.5.14).

■ clone contig를 만드는 더 새로운 방법들

ㆍ이상의 방법들로서는 15-20 clone들 정도만 조립할 수 있으며, 가까운 거리(수 Mb)에 존재하는 유전자를 찾아갈 때 사용하는 positional cloning등에 사용되고 있다.

ㆍgenome 전체적으로 clone contig를 조립하기 위해서는 다음과 같이 다른 방법이 필요



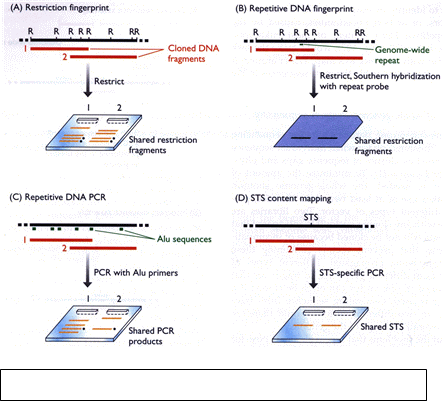
ㆍ그 중 가장 주된 것은 clone fingerprinting이며, cloned 된 DNA의 물리적 구조를 포함하는 것이다. 이러한 물리적 구조적 특징을 비교하여 clone contig를 조립한다. clone fingerprinting으로는 다음과 같은 것이 사용될 수 있다

- restriction pattern : 공통적인 제한효소조각 존재 유무

- repetitive DNA fingerprints :특이한 반복서열의 존재 유무

- repetitive DNA PCR or interspersed repeat element PCR(IRE-PCR) : 반복서열을 PCR로 확인함

- STS content mapping : single copy인 STS를 PCR로 확인함



5.2.3 Whole-genome shotgun sequencing

ㆍ전체 genome의 6.5-8배에 해당하는 sequence를 얻으면 전체게놈의 99.8%를 cover할 수 있다

ㆍ소수의 gap은 H. influenza에서 사용한 방법을 사용하여 메꿀 수 있다.

ㆍ500 bp clone 7천만 개이면 35000Mb에 해당하며 human genome해독에 충분함.

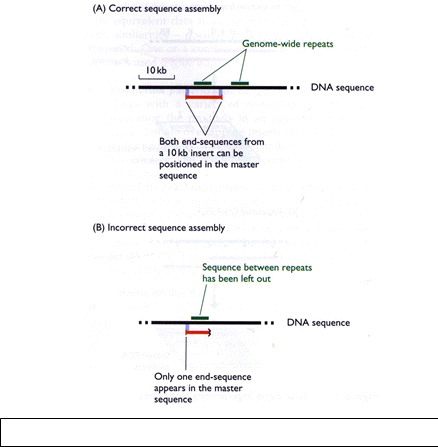
ㆍ하루에 1000개의 clone을 sequencing할 수 있는 자동염기서열분석기 75대를 3년간 작동하면 가능

ㆍ문제는 7천만 sequence를 조립할 수 있는가 하는 것이다. computing time이 많이 걸리고, 반복서열에의한 에러등을 고려하면 대답은 불가능다. 그러나 map에 있는 marker들을 사용하면 가능하다.

■ whole-genome sequencing 의 특징

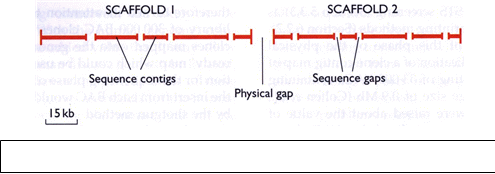
ㆍshotgun 법에서 가장 시간이 걸리는 부분은 각 sequence contig들을 조립하고 gap을 메꾸는 것이다.

ㆍ효과적으로 시간을 줄이기 위하여 두 가지 서로 다른 vector에 만든 library를 이용한다.

ㆍ반복서열의 문제해결은 여러 가지가 있지만 한 library는 가장 긴 반복서열보다 큰 DNA 조각들로 만든다는 것이다. 이렇게 함으로서 반복서열의 양쪽 끝을 probe로 이용하면 반복서열에 의한 실수를 줄일 수 있다(Fig. 5.16).

ㆍsequence를 조립하는 중간단계에 scaffold를 만들게 된다. scaffold는 sequence gap(clone을 sequencing하여 메꿀 수 있는 gap을 말함)에 의해 틈이 있는 여러개의 sequence contig로 구성되며, 각 scaffold는 physical gap(다른 library에 의해 메꿀 수 있는 gap을 말함)에 의해 분리되어져 있다.

ㆍ각 scaffold 가 염색체의 어느 부분에 해당하는지는 알기위해 어떤 STS가 scaffold에 있는지 알아내면 된다.



ㆍwhole-genome shotgun에 의해 얻어진 sequence의 정확도에 대한 신뢰가 문제가 될 수 있다. genome의 어떤 부분은 다른 부분보다 더 자주 sequencing 된다는 것이다. 따라서 어떤 부분은 한번 혹은 두 번 정도만 sequencing 된다는 것이다.

**5.3 인간게놈 프로젝트**

ㆍ지금까지 설명한 방법들을 인간 게놈프로젝트에 어떻게 이용하는가 하는 것을 설명.

5.3.1 Mapping of human genome

ㆍ인간게놈의 polymorphic marker의 수가 적게 알려져 있었으므로 mapping이 힘들 것으로 생각했지만 RFLP marker들이 밝혀지면서 가능성이 보임.

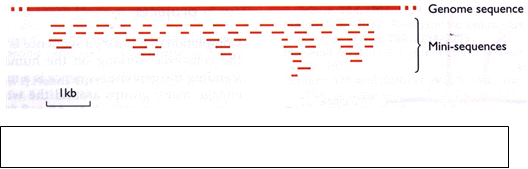
ㆍ1987년 393개의 RFLP marker로된 RFLP marker가 발표되었고, 약 10Mb간격의 RFLP mapping이 21번 염색체의 mapping에 사용됨

ㆍ인간 genome project를 위해서는 최대 2-5Mb 간격의 mapping이 이루어져야 하므로 1980년대 후반에 1Mb 간격의 mapping을 위해 국제협력이 시작되었고, 1994년 국제컨소시움은 SSLP marker와 The Centre d'Etudes du Polymorphisme Humain (CEPH) collection의 덕분에 이 보다 더 상세한 mapping 이 만들어짐. 이 map은 5800 marker를 포함하고 있으며, 그 중 4000개는 SSLP로서 0.7Mb에 하나의 marker를 갖고 있다.

ㆍ한편 physical mapping 도 STS, fingerprint 등을 이용하여 clone contig를 많이 만듬(33000 YAC(평균 0.9Mb fragments)).

5.3.2 인간게놈의 sequencing

ㆍ초기에는 YAC clone이 비교적 큰 DNA 조각을 포함하기 때문에 YAC clone으로 주로 clone contig을 만들었지만, YAC clone은 불연속적인 DNA 조각으로 되어 있는 경우가 있는 관계로 후에 BAC clone을 만들었음. 300000BAC clone을 만들어 염색체상에 이 clone들의 위치를 정하고 BAC clone을 shotgun에 의해 sequencing 함.



ㆍ이 때쯤 인간의 게놈을 sequencing 하기위해 힘든 clone contig 방법보다 whole-genome shotgun을 도입하자는 움직임이 있었음.

ㆍ1999년도 22번 염색체의 sequence draft가 완성됨. 그 후 2000년 21번의 sequence

draft가 완성됨

ㆍ미 대통령과 human genome project 대표가 2001년에 인간게놈프로젝트가 완성될 것이라고 발표함

ㆍ 2003년 인간게놈 project 완성

