RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF ENGINEERING &TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY

Dr. Simranjit Singh
Assistant Professor
Department of Biotechnology
Rama University, Kanpur

Microbial Genomics

Genetic engineering, DNA sequencing of bacterial genomes and the transcriptome.

Outline of lecture

- Traditional recombinant DNA methodology including; restriction enzymes, plasmids as cloning vectors, clones selection and Southern blotting.
- DNA sequencing
- Polymerase chain reaction
- High throughput DNA sequencing
- Prokaryote genome sequencing and annotation
- High throughput mRNA analysis
- Transcriptome

Restriction modification systems: the biological phenomena

- An *E. coli* strain, lets say K12 was infected with a bacteriophage grown on *E. coli* strain B and only a fraction of the expected plaques was observed !
- One plaque was picked and a new phage suspension prepared which in turn was used to infect *E. coli* strain K12. This time the correct number of plaques was observed. What’s going on ?

Restriction modification systems: the biological explanation

- E. coli K12 has a restriction modification system which is not found in E. coli B.
- The restriction modification system consists of two components: a site (sequence) specific endonuclease and a corresponding methylase.
- The methylase methylates the recognition site and the endonuclease only cuts the DNA at non-methylated sites. So the E. coli DNA is protected by methylation.
- The bacteriophage DNA is not methylated and when it is injected into the bacteria cell it is cut up into pieces by the endonuclease.
- The system is not 100% perfect and a few bacteriophage manage to compete a successful infection. These bacteriphage are methylated in the correct sites and the phage can now infect E. Coli K12 without being cut up.

The tools

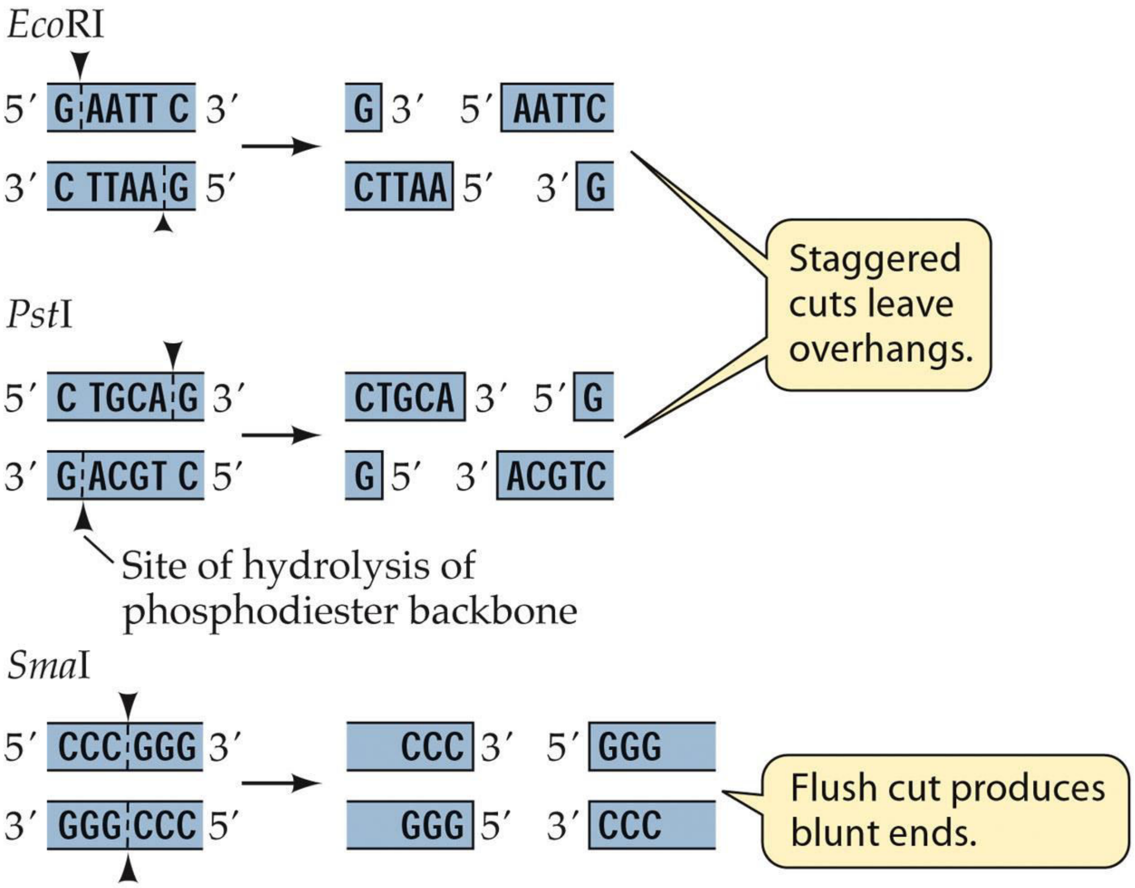
Restriction enzymes

Vectors

Cloning

Selection

Restriction enzymes



MICROBIAL LIFE, **Figure 16.1** © 2002 Sinauer Associates, Inc.

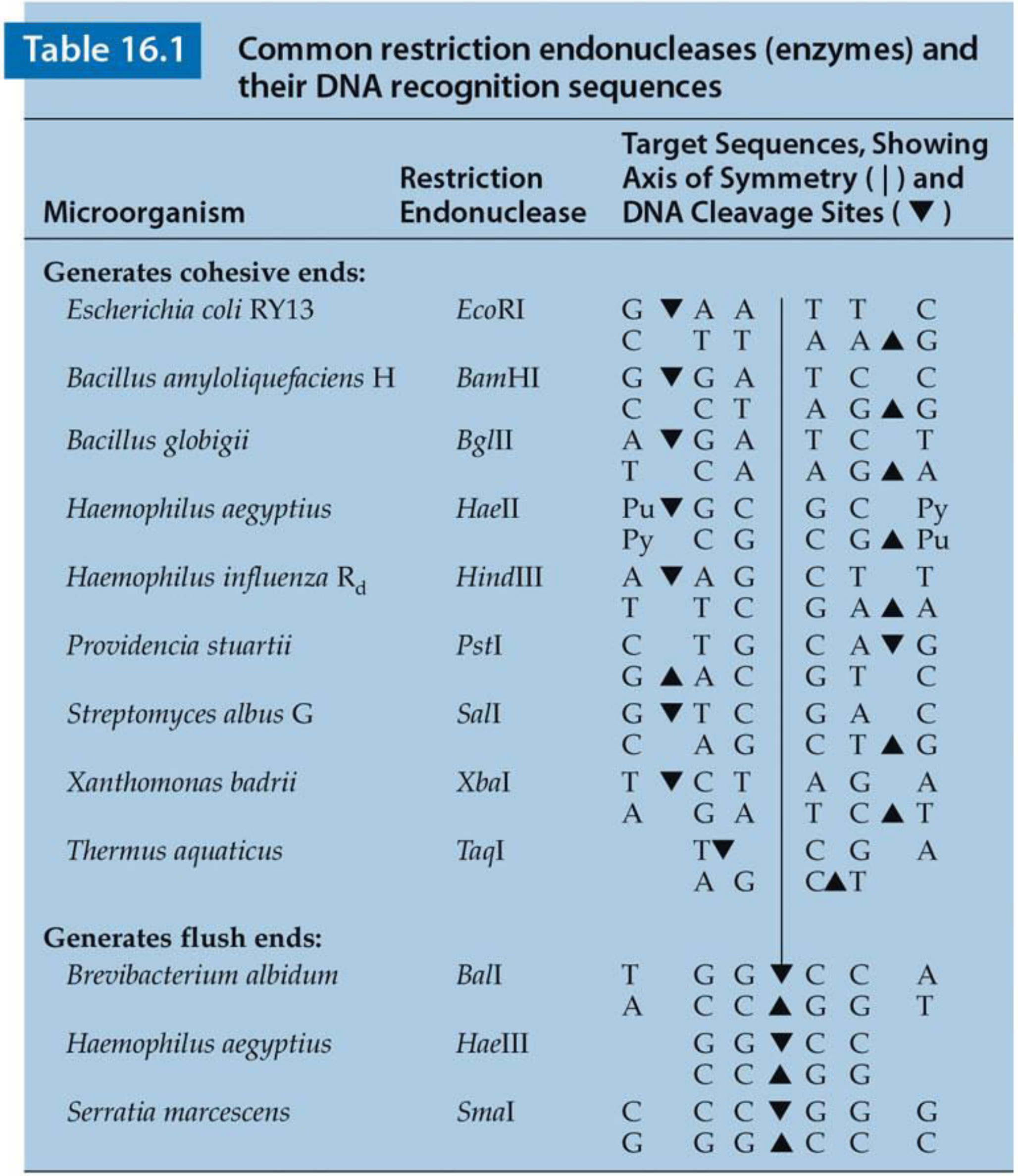


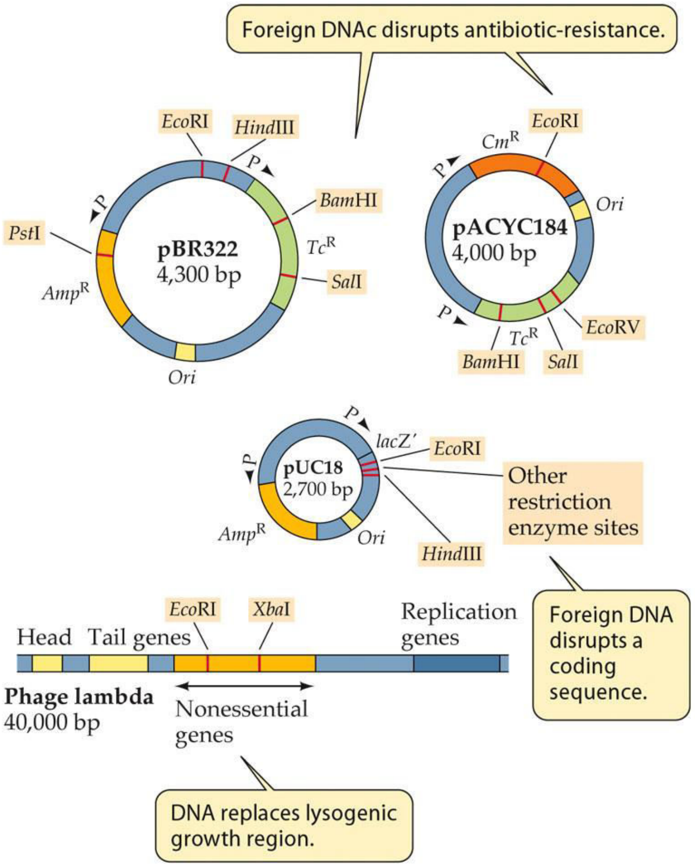
Table 16.1 Common restriction endonucleases (enzymes) and their DNA recognition sequences

Microorganism	Restriction Endonuclease	Target Sequences, Showing Axis of Symmetry (|) and DNA Cleavage Sites (▼)
Generates cohesive ends:		
Escherichia coli RY13	*Eco*RI	G ▼ A A | T T C C T T | A A ▲ G
Bacillus amyloliquefaciens H	*Bam*HI	G ▼ G A | T C C C C T | A G ▲ G
Bacillus globigii	*Bgl*II	A ▼ G A | T C T T C A | A G ▲ A
Haemophilus aegyptius	*Hae*II	Pu ▼ G C | G C Py Py C G | C G ▲ Pu
Haemophilus influenza R_d	*Hind*III	A ▼ A G | C T T T T C | G A ▲ A
Providencia stuartii	*Pst*I	C T G | C A ▼ G G ▲ A C | G T C
Streptomyces albus G	*Sal*I	G ▼ T C | G A C C A G | C T ▲ G
Xanthomonas badrii	*Xba*I	T ▼ C T | A G A A G A | T C ▲ T
Thermus aquaticus	*Taq*I	T ▼ | C G A A G | C ▲ T
Generates flush ends:		
Brevibacterium albidum	*Bal*I	T G G ▼ C C A A C C ▲ G G T
Haemophilus aegyptius	*Hae*III	G G ▼ C C C C ▲ G G
Serratia marcescens	*Sma*I	C C C ▼ G G G G G G ▲ C C C

MICROBIAL LIFE, **Table 16.1** © 2002 Sinauer Associates, Inc.

New England BioLabs has 240 different restriction endonucleases on sale

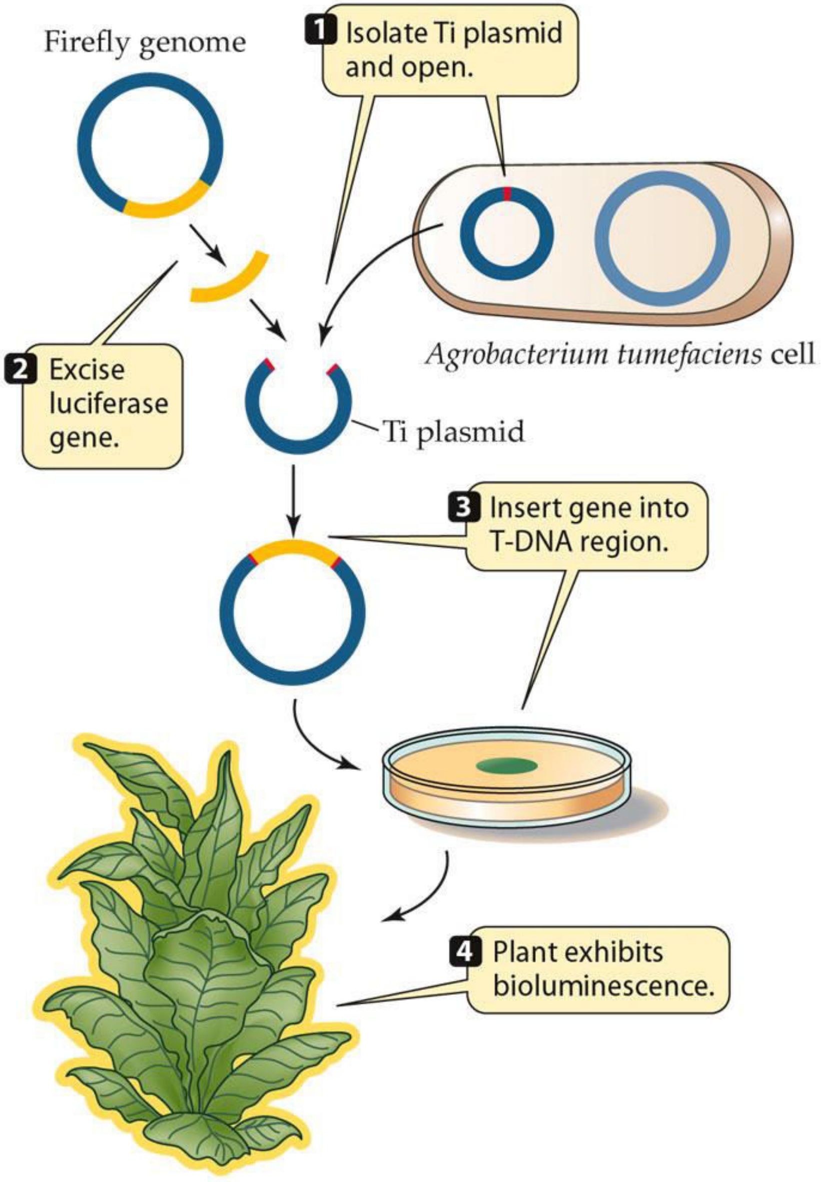
Plasmid vectors for construction of recombinant molecules



MICROBIAL LIFE , Figure 16.4 © 2002 Sinauer Associates, Inc.

Plasmids pBR322, pACYC184, and pUC18 contain several unique restriction enzyme recognition sequences (shaded) as well as resistance determinants for ampicillin (*Amp*R), tetracycline (*Tc*R), and chloramphenicol (*Cm*R). Plasmid pUC18 has a multiple cloning site with recognition sequence for 13 different restriction enzymes. In the case of pUC18, to screen for inserted DNA, the plasmids must be transformed into special *E. coli* strains that carry the gene for the non-*lac*Z' portion of β-galactosidase; the two portions of the protein form the active enzyme. Artificial formation of an active enzyme from two fragments (called complementation) allows screening for insertions by colony color in gel containing X-gal, the chromogenic indicator of β-galactosidase. Bacterial colonies with intact plasmid have a blue appearance; colonies with plasmids that carry inserts are white.

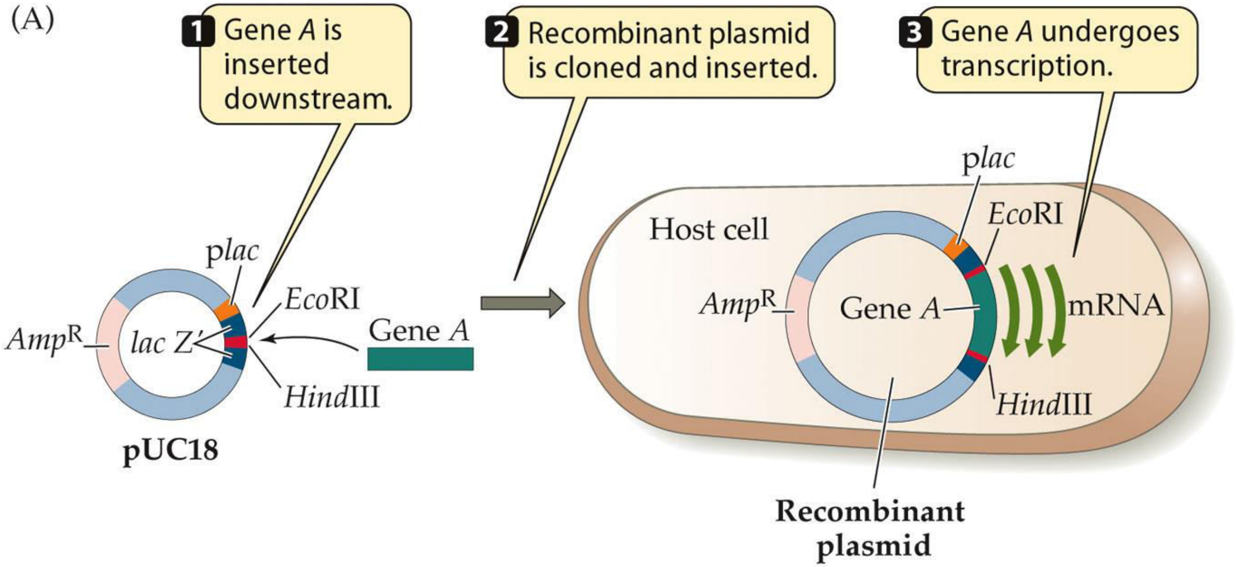
Genetic engineering of plants using *Agrobacterium tumefaciens* and Ti plasmids


Firefly genome
1 Isolate Ti plasmid and open.
Agrobacterium tumefaciens cell
2 Excise luciferase gene.
Ti plasmid
3 Insert gene into T-DNA region.
4 Plant exhibits bioluminescence.

MICROBIAL LIFE, Box 16.2 © 2002 Sinauer Associates, Inc.

Expression vectors

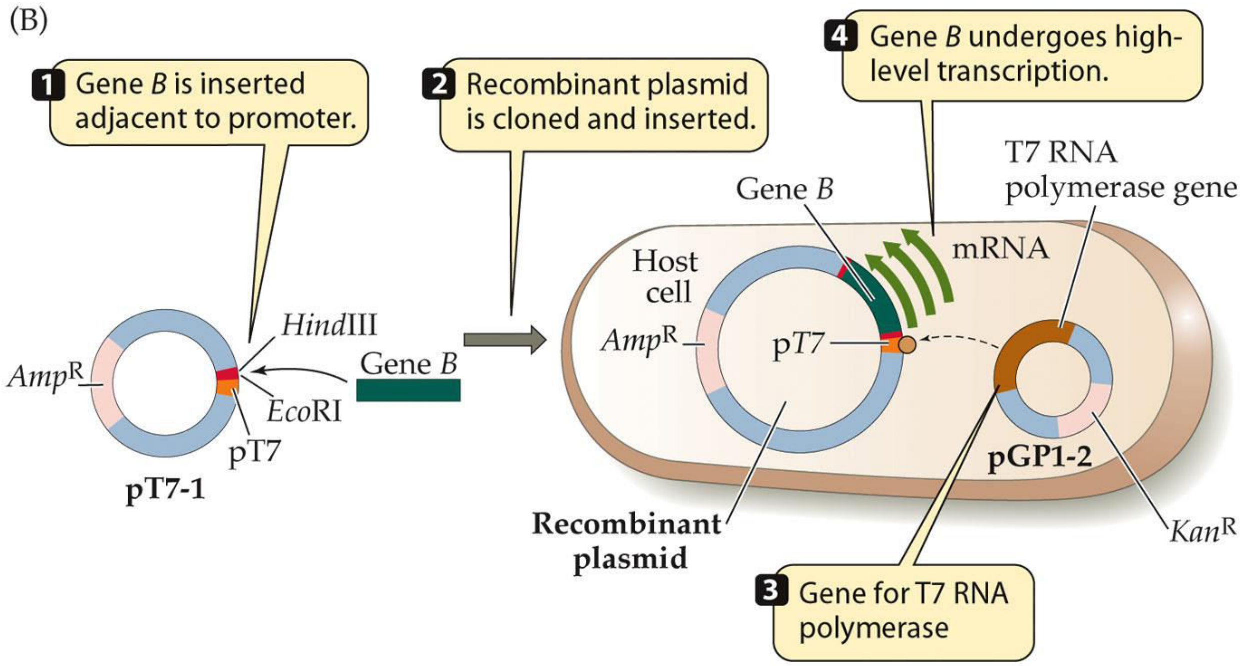
Plasmid pUC18 contains the promoter for the *lac* operon of *E. coli* (*plac*) adjacent to a cluster of restriction sites, which can be used to insert foreign genes. Transcription from *plac* results in high-level expression of the cloned gene.



MICROBIAL LIFE, Figure 16.7 (Part 1) © 2002 Sinauer Associates, Inc.

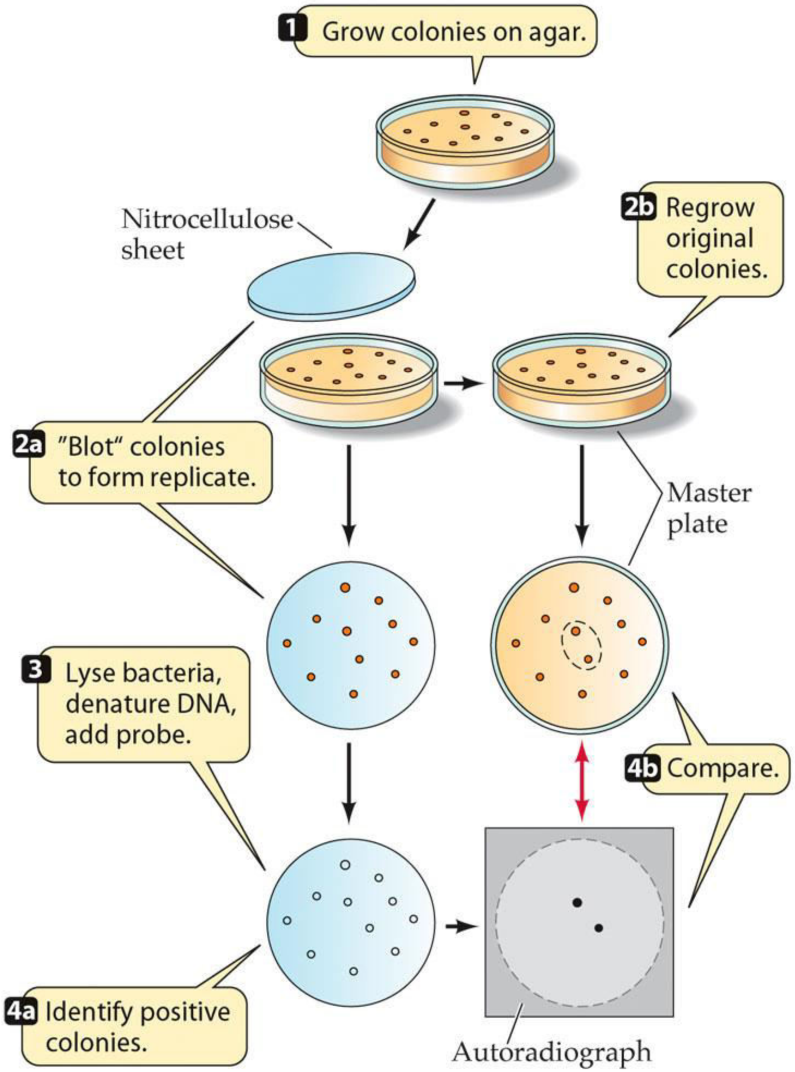
Expression vectors

An expression system based on components of a bacteriophage. The cloning vector pT7-1 contains a promoter from bacteriophage T7 (pT7), which is recognized by a specialized RNA polymerase from the same phage.


(B)
1 Gene *B* is inserted adjacent to promoter.
2 Recombinant plasmid is cloned and inserted.
4 Gene *B* undergoes high-level transcription.
3 Gene for T7 RNA polymerase
*Hind*III
Gene *B*
*Eco*RI
*Amp*R
pT7
pT7-1
Host cell
Gene *B*
mRNA
T7 RNA polymerase gene
*Amp*R
pT7
pGP1-2
*Kan*R
Recombinant plasmid

MICROBIAL LIFE , **Figure 16.7 (Part 2)** © 2002 Sinauer Associates, Inc.

Colony hybridization


1 Grow colonies on agar.
Nitrocellulose sheet
2b Regrow original colonies.
2a "Blot" colonies to form replicate.
Master plate
3 Lyse bacteria, denature DNA, add probe.
4b Compare.
4a Identify positive colonies.
Autoradiograph

MICROBIAL LIFE, Figure 16.8 © 2002 Sinauer Associates, Inc.

Individual bacterial colonies are immobilized on nitrocellulose. The nitrocellulose replica is then exposed to a radioactive DNA probe, and positive colonies are identified.

DNA sequencing

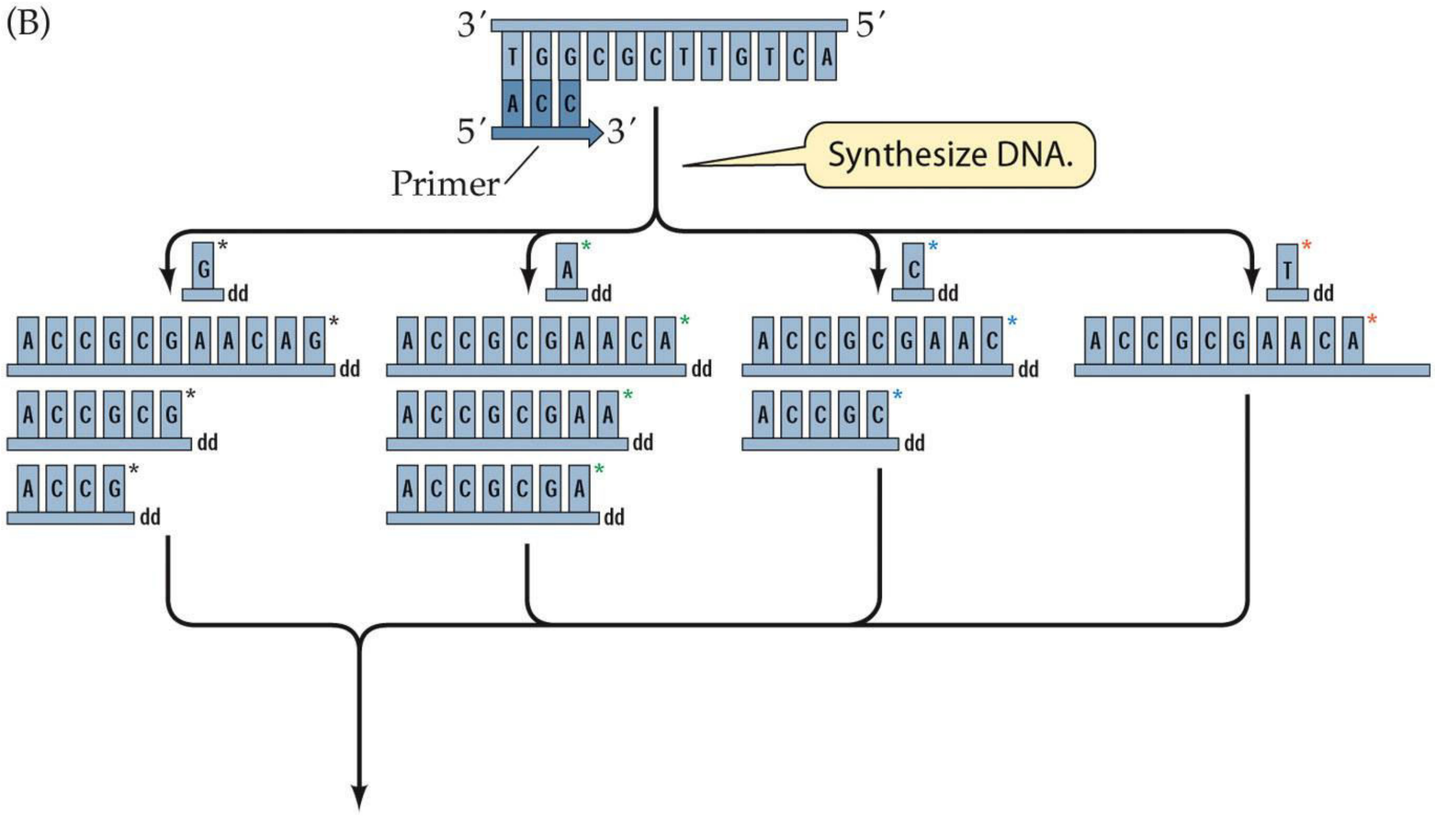
- What do we need ?
  A template; single strand DNA
  A primer which is a short synthetic DNA oligonucleotide
  A DNA polymerase
  Deoxynucleotide triphosphates
  Chain terminating Dideoxynucleotide triphosphates

  A separation method and a detection system

The basic principles have not changed but the DNA polymerases are better, the separation methods are better as are the detection methods.

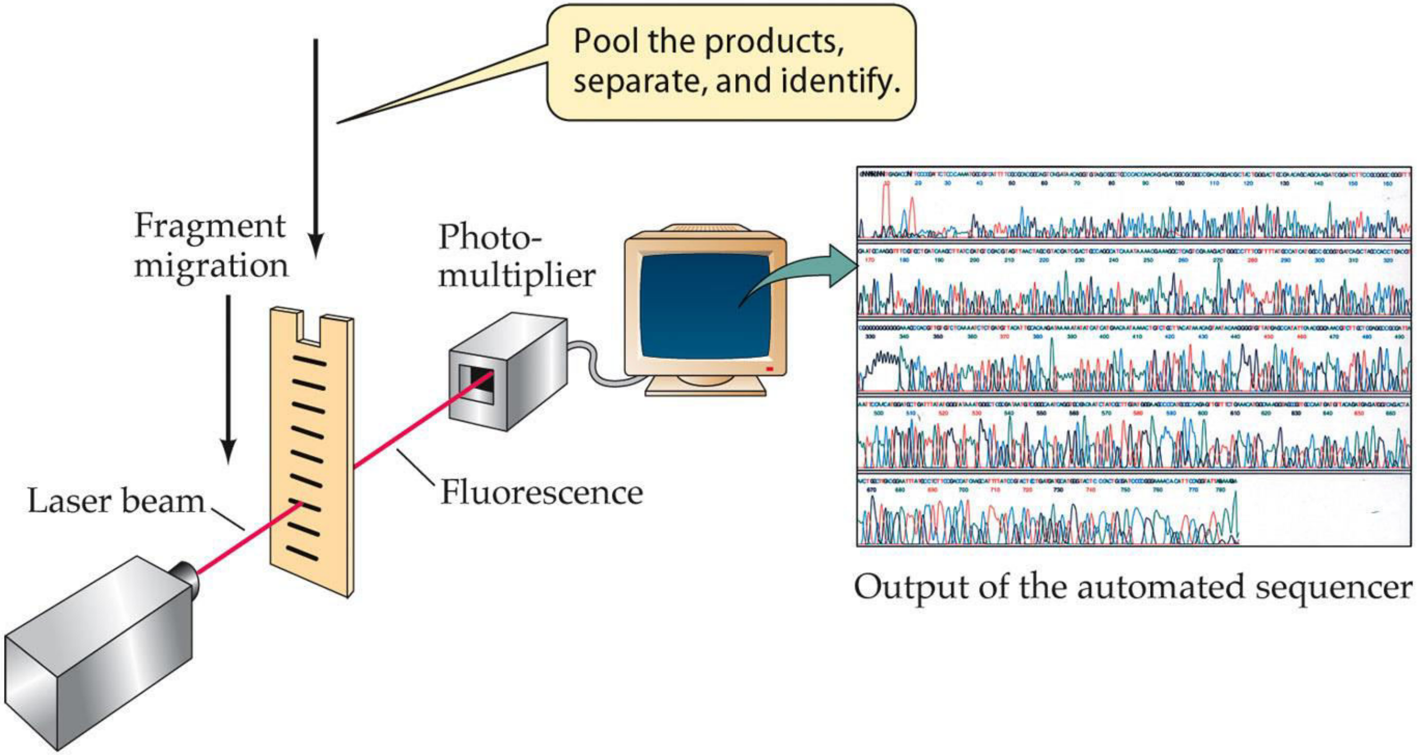
In 1983 we took several weeks to sequence 1 kb, today we can sequence 100 kb a day and the sequencing factories manage several hundred times more!

DNA sequence determination using the dideoxynucleotide (ddNTP) chain termination method.



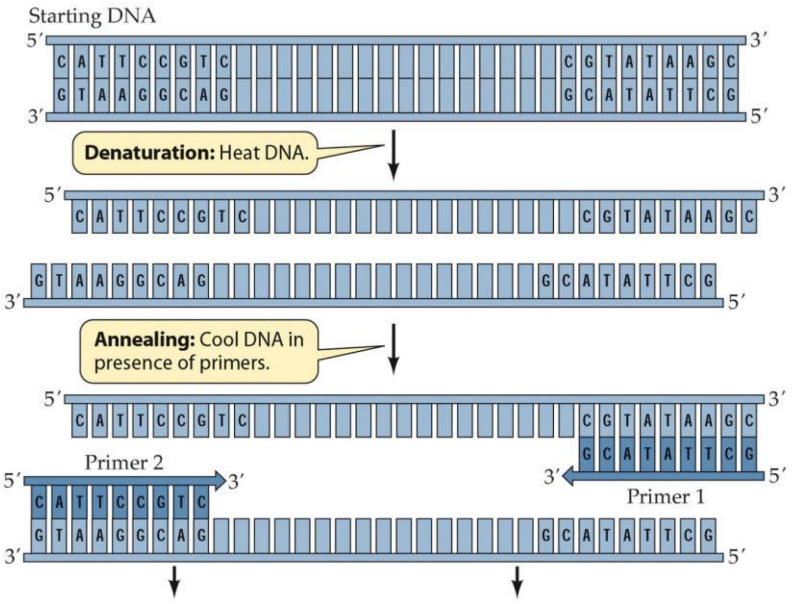
MICROBIAL LIFE, Box 16.3 (Part 2) © 2002 Sinauer Associates, Inc.

DNA sequence determination using the dideoxynucleotide (ddNTP) chain termination method.


Pool the products, separate, and identify.
Fragment migration
Photo-multiplier
Laser beam
Fluorescence
Output of the automated sequencer

MICROBIAL LIFE, Box 16.3 (Part 3) © 2002 Sinauer Associates, Inc.

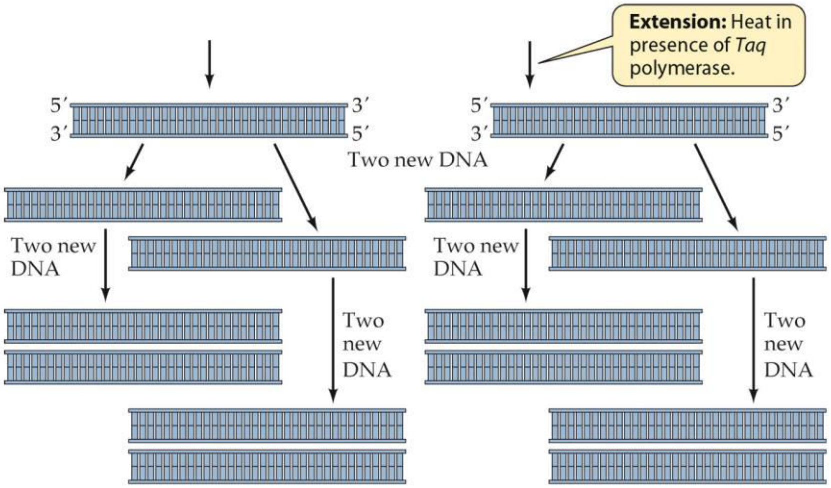
Polymerase Chain Reaction



Denaturation: Heat DNA.

Annealing: Cool DNA in presence of primers.

MICROBIAL LIFE, **Box 16.1 (Part 1)** © 2002 Sinauer Associates, Inc.



Extension: Heat in presence of *Taq* polymerase.

MICROBIAL LIFE, **Box 16.1 (Part 2)** © 2002 Sinauer Associates, Inc.

How are these methods used in microbiological research?

Bottom up approach:

To start of with we had an observable phenotype which we wanted to investigate.

The organism was mutated using transposon insertional mutagenesis and a very large number of mutants were screened for strains in which the observed phenotype was affected.

So the insertion of a transposon knocked out a gene which was important for the phenotype under investigation. Make a genomic library, find the clone containing the transposon, find the gene that was knocked out and sequence it.

Gene identified !

Repeat the procedure for all the mutants and isolate all the genes affecting the phenotype under investigation.

Carry out complementation analysis to verify that the correct gene has been isolated.

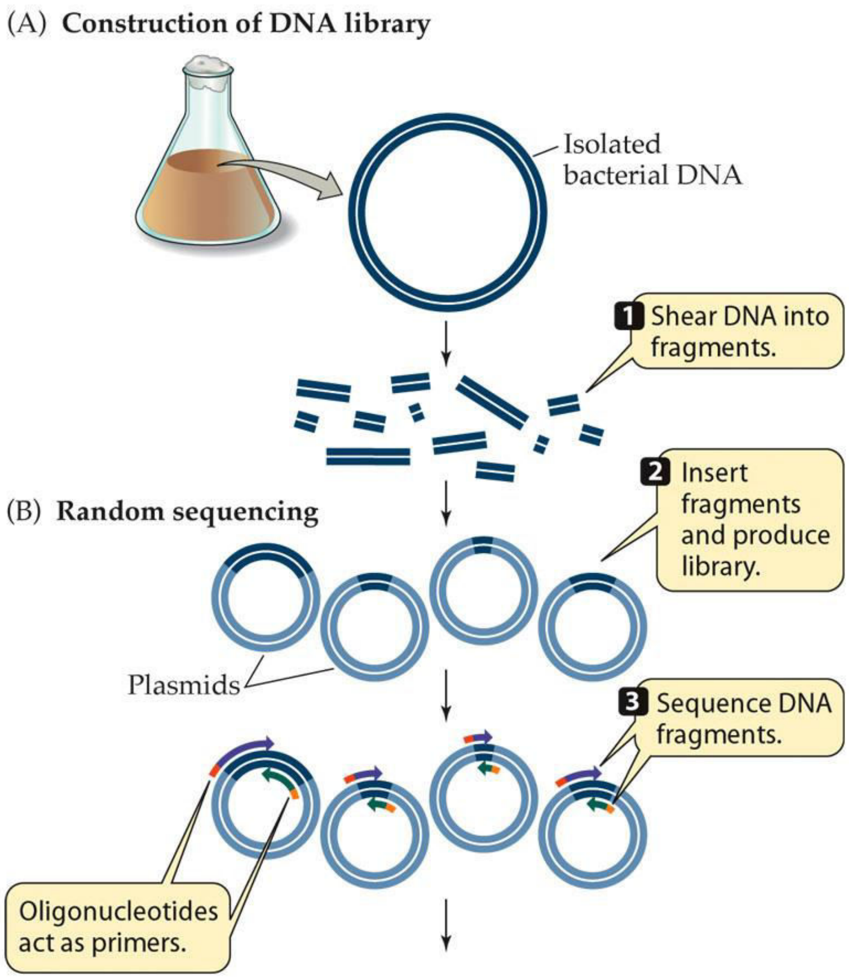
Time consuming but immensely productive and the basis of modern molecular biology.

Whole genome sequencing and microbial genomics

In the next part of this lecture we will see how easy it is to sequence complete microbial genomes and how this nucleotide sequence data is transformed into biological data.

This is called the top down approach.

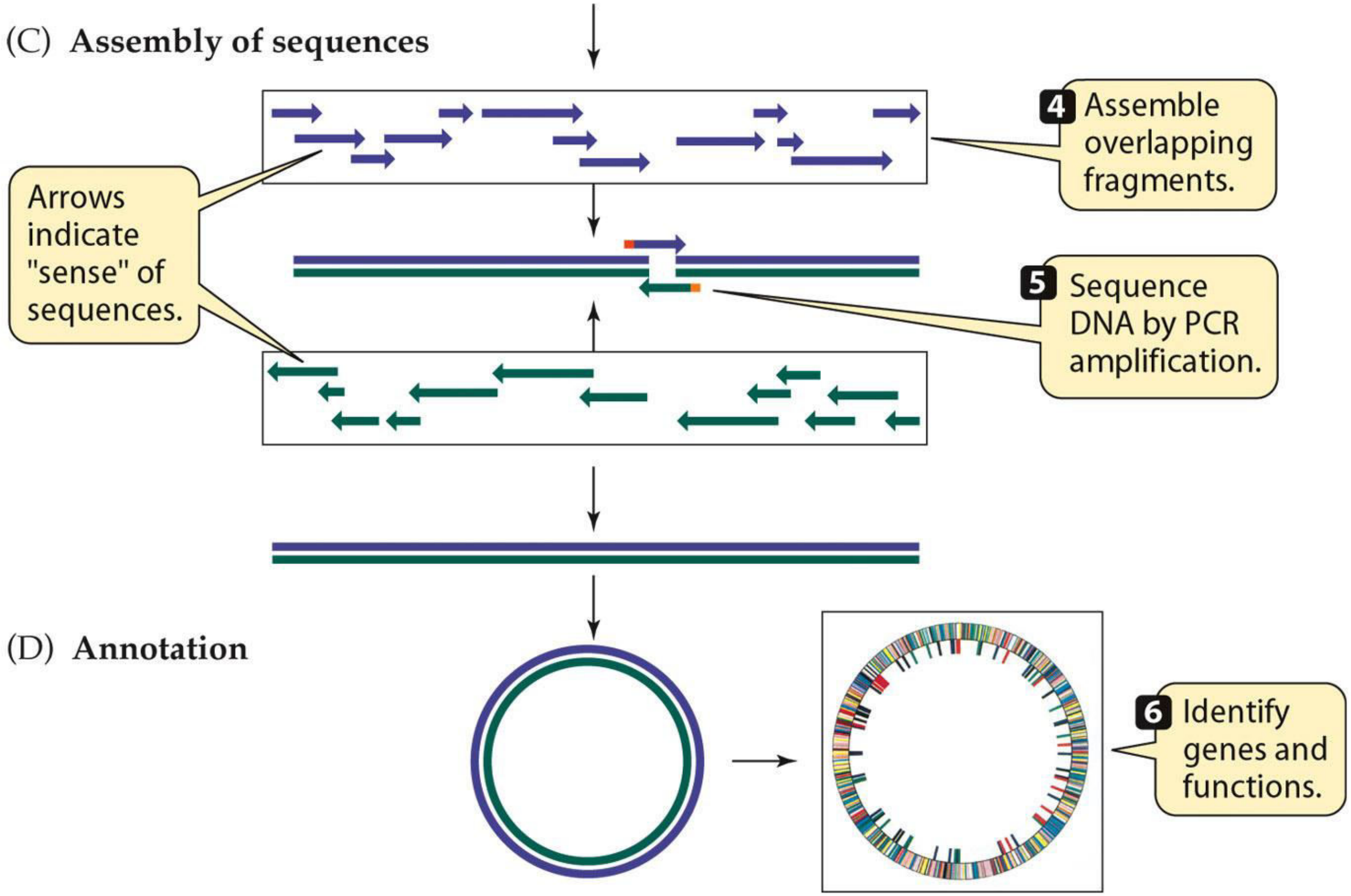
Whole-genome shotgun sequencing


(A) Construction of DNA library
Isolated bacterial DNA
1 Shear DNA into fragments.
(B) Random sequencing
2 Insert fragments and produce library.
Plasmids
3 Sequence DNA fragments.
Oligonucleotides act as primers.

MICROBIAL LIFE, Figure 16.9 (Part 1) © 2002 Sinauer Associates, Inc.

Determination of a microbial genome sequence. (A) Construction of a random library of DNA fragments in a cloning vector. (B) Random sequencing of clones. Short sequences are obtained from each end of the cloned DNA, and thousands of clones are sequenced

Whole-genome shotgun sequencing


(C) Assembly of sequences
4 Assemble overlapping fragments.
Arrows indicate "sense" of sequences.
5 Sequence DNA by PCR amplification.
(D) Annotation
6 Identify genes and functions.

MICROBIAL LIFE, **Figure 16.9 (Part 2)** © 2002 Sinauer Associates, Inc.

Identification of genes and elucidation of gene function

From nucleotide sequence to potential genes: bacterial genes have well defined features which makes it possible to identify genes be computational analysis. The features are; promotor regions 5’ to the gene, a DNA dependent RNA polymerase binding site, a start codon, an open reading frame in units of 3 nucleotides, a stop codon, a transcription termination sequence.

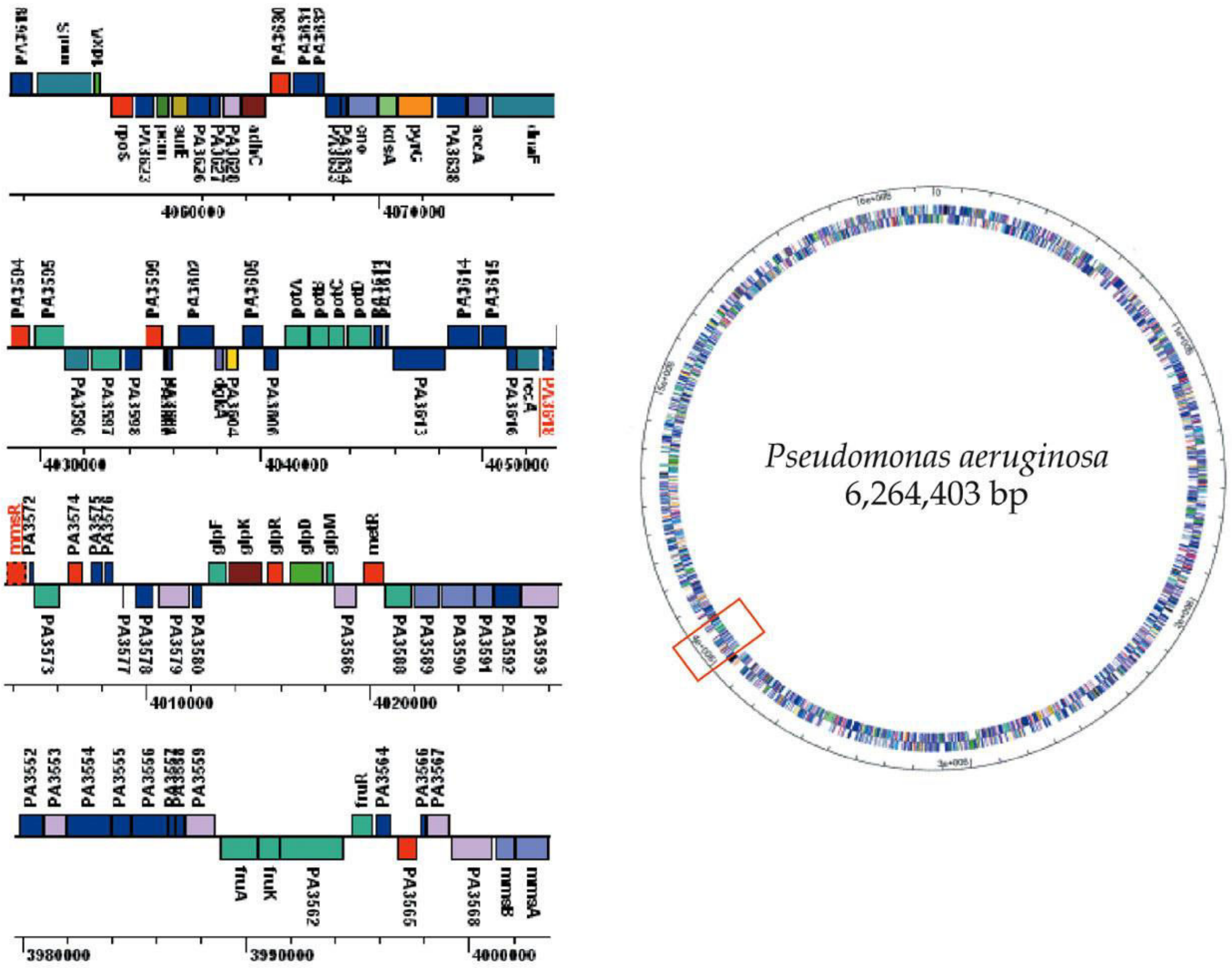
From open reading frames to amino acid sequences: translate codons to corresponding amino acid sequences.

From amino acid sequences to protein function: this is the tricky part. Compare the amino acid sequence for sequence similarity to all known prokaryote proteins. Computational analysis using Basic local alignment tool = BLAST.

Interpretation of the results: At the one extreme, the function of the protein is well known in other species and the sequences are very similar = function assignment. At the other extreme, there are no proteins in the databases which resemble our sequence.

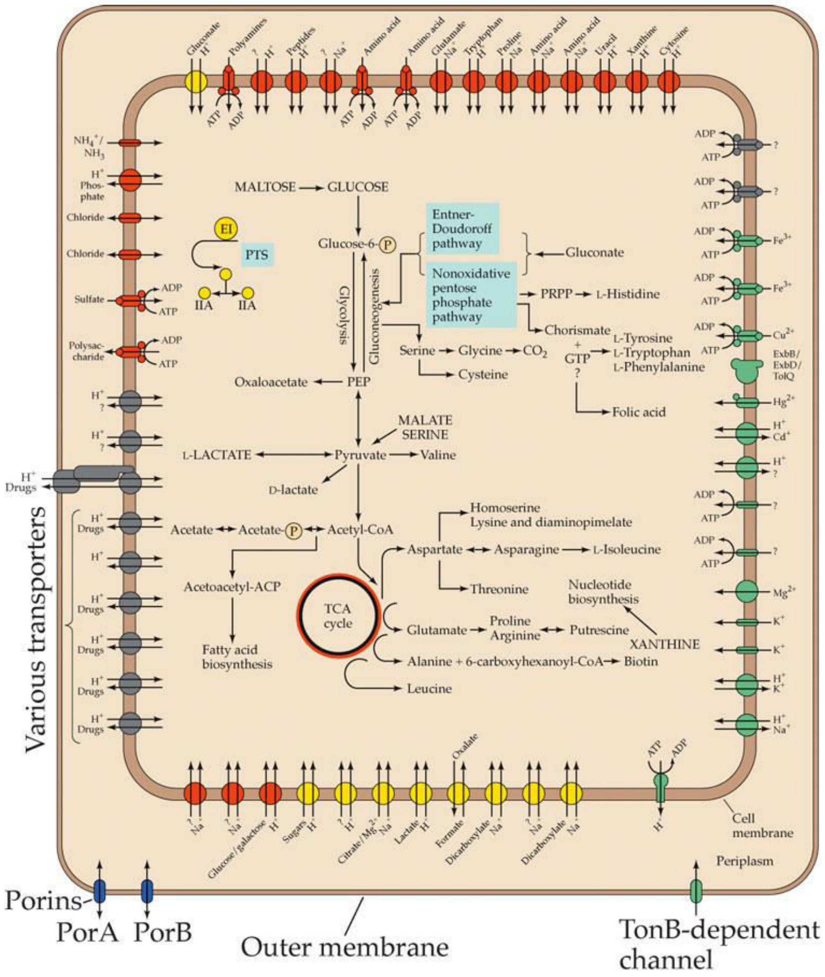
Reality: The function of about 40% of the proteins in most bacteria is unknown.

Genes in a portion of a bacterial genome



MICROBIAL LIFE , Figure 16.10 © 2002 Sinauer Associates, Inc.

Cellular functions based on an annotated genome



MICROBIAL LIFE, **Figure 16.11** © 2002 Sinauer Associates, Inc.

Reconstruction of transport and metabolism of *Neisseria meningitidis*, based on the annotated genome. The reconstruction shows the potential pathways for the generation of energy and metabolism of organic compounds. Question marks indicate that a transporter's substrates are unknown, and functional assignment is based on its overall similarity to other transporters. After Nelson, K. T., I. T. Paulson and C. M. Fraser. 2001. *ASM News* 67: 310Ð317.

Table 16.2 Comparison of regulatory genes in selected bacterial genomes

Microorganism	# Genes in the Genome	# Regulatory Proteins	% of Total
Pseudomonas aeruginosa	5570	468	8.4
Escherichia coli	4289	250	5.8
Bacillus subtilis	4100	217	5.3
Mycobacterium tuberculosis	3918	117	3.0
Helicobacter pylori	1566	18	1.1

Table 16.3 Distribution of genes of unknown function among selected bacterial genomes *(Part 1)*

Organism	Genome Size (Mbp)	No. of ORFs (% coding)	Unknown Function	Unique ORFs
Aeropyrum pernix K1	1.67	1,885 (89%)		
A. aeolicus VF5	1.50	1,749 (93%)	663 (44%)	407 (27%)
A. fulgidus	2.18	2,437 (92%)	1,315 (54%)	641 (26%)
B. subtilis	4.20	4,779 (87%)	1,722 (42%)	1,053 (26%)
B. burgdorferi	1.44	1,738 (88%)	1,132 (65%)	682 (39%)
Chlamydia pneumoniae AR39	1.23	1,134 (90%)	543 (48%)	262 (23%)
Chlamydia trachomatis MoP_n	1.07	936 (91%)	353 (38%)	77 (8%)
C. trachomatis serovar D	1.04	928 (92%)	290 (32%)	255 (29%)
Deinococcus radiodurans	3.28	3,187 (91%)	1,715 (54%)	1,001 (31%)
E. coli K-12-MG1655	4.60	5,295 (88%)	1,632 (38%)	1,114 (26%)
H. influenzae	1.83	1,738 (88%)	595 (35%)	237 (14%)
H. pylori 26695	1.66	1,589 (91%)	744 (45%)	539 (33%)
Methanobacterium thermotautotrophicum	1.75	2,008 (90%)	1,010 (54%)	496 (27%)

Table 16.3 Distribution of genes of unknown function among selected bacterial genomes *(Part 2)*

Organism	Genome Size (Mbp)	No. of ORFs	(% coding)	Unknown Function		Unique ORFs	
Methanococcus jannaschii	1.66	1,783	(87%)	1,076	(62%)	525	(30%)
M. tuberculosis CSU#93	4.41	4,275	(92%)	1,521	(39%)	606	(15%)
M. genitalium	0.58	483	(91%)	173	(37%)	7	(2%)
M. pneumoniae	0.81	680	(89%)	248	(37%)	67	(10%)
N. meningitidis MC58	2.24	2,155	(83%)	856	(40%)	517	(24%)
Pyrococcus horikoshii OT3	1.74	1,994	(91%)	589	(42%)	453	(22%)
Rickettsia prowazekii Madrid E	1.11	878	(75%)	311	(37%)	209	(25%)
Synechocystis sp.	3.57	4,003	(87%)	2,384	(75%)	1,426	(45%)
T. maritma MSB8	1.86	1,879	(95%)	863	(46%)	373	(26%)
T. pallidum	1.14	1,039	(93%)	461	(44%)	280	(27%)
Vibrio cholerae El Tor N1696	4.03	3,890	(88%)	1,806	(46%)	934	(24%)
	50.60	52,462	(89%)	22.358	(43%)	12,161	(23%)

From Fraser et al., *Nature* 2000, vol. 406. p. 800.

Status of genome sequencing projects 6 october 2004

http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html
http://www.ncbi.nlm.nih.gov/RefSeq/
http://www.genomesonline.org/

191 Prokaryote genomes fully sequenced and annotated (172 Bacteria, 19 Archaea)

521 Ongoing prokaryote genome projects: (494 Bacteria, 27 Archaea)

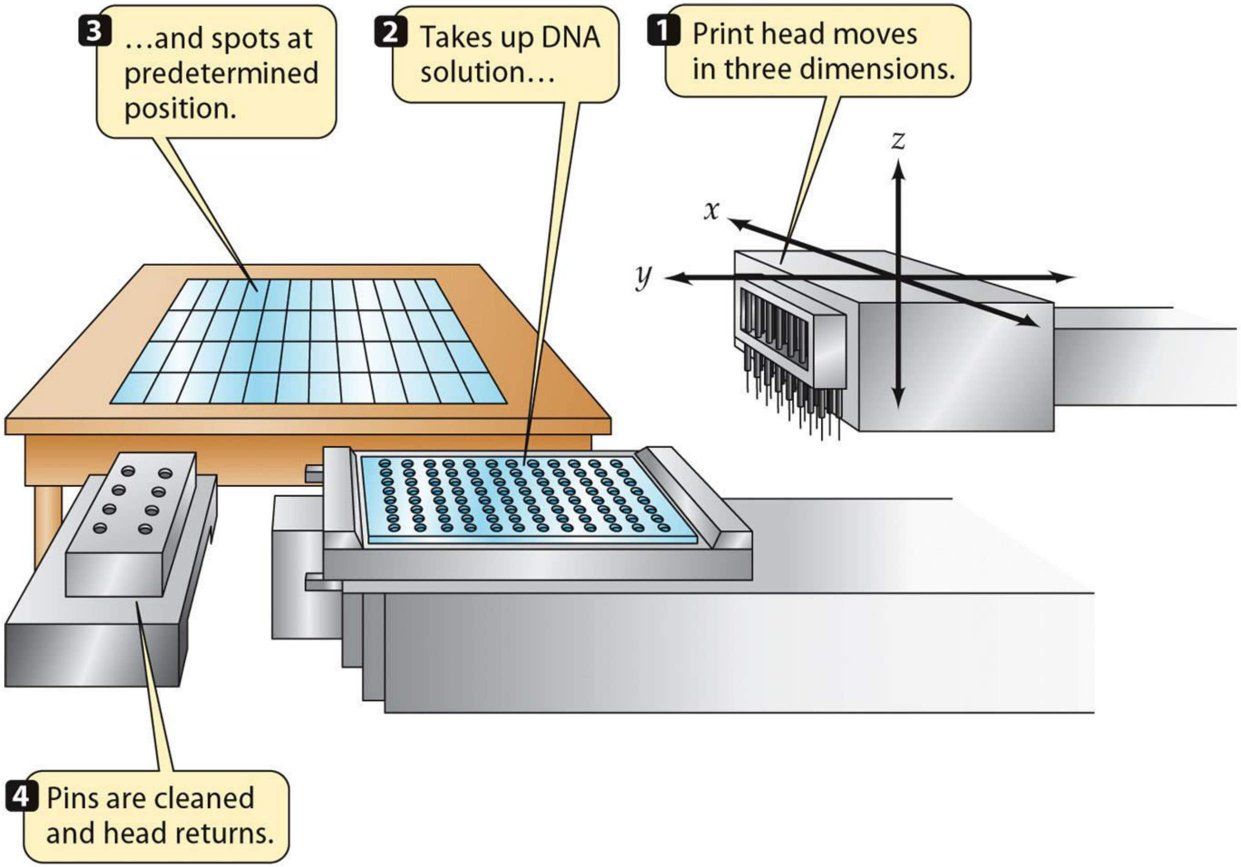
28 Eukaryote genomes fully sequenced and annotated. 453 ongoing eukaryote projects. 7 vertebrates, 9 invertebrates, 7 fungii, 10 protozoa and 2 flys.

1674 virus genomes have been sequenced and annotated

Transcription analysis

- DNA > RNA > protein
- How do we measure the levels of all the different mRNA transcripts ?
- When the genome sequence is known a set of DNA molecules which correspond to each gene can be synthesized.
- These a spotted on a glass slide and this is called a DNA micro-array because the spots are very small and can only be seen under a microscope.
- Total RNA is isolated and reverse transcribed into cDNA. The cDNA is labeled using fluorochromes and hybridize to the array. Positive signals means that the gene is expressed.
- Change the growth conditions and isolate a new RNA. Repeat as above and compare.
- In practice the two different RNA preparations (cDNA) are labeled with two different fluorochromes and the comparative levels of transcription from the two different growth conditions are compared. Not as easy as it sounds.

A robotic DNA spotting (printing) device



1 Print head moves in three dimensions.

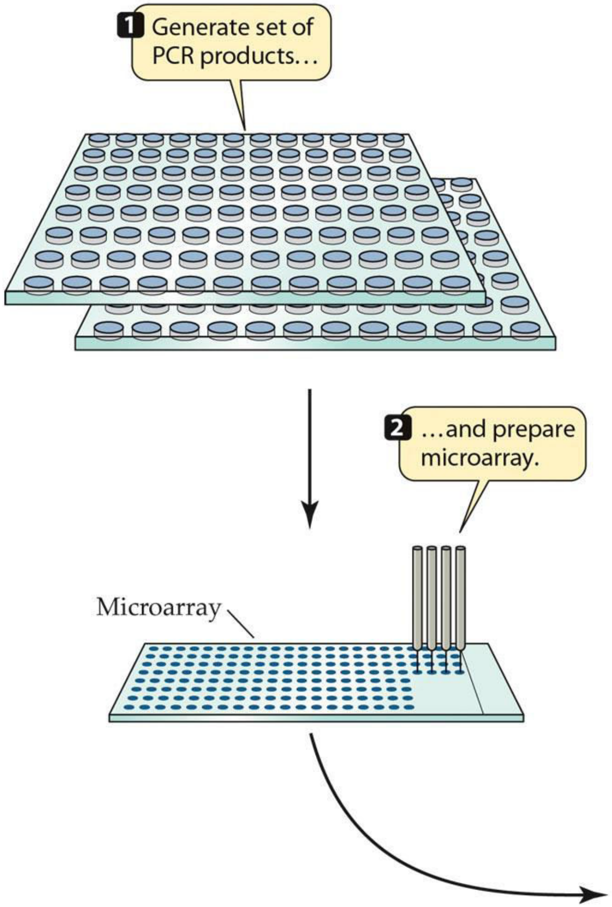
2 Takes up DNA solution…

3 …and spots at predetermined position.

4 Pins are cleaned and head returns.

MICROBIAL LIFE , Figure 16.12 © 2002 Sinauer Associates, Inc.

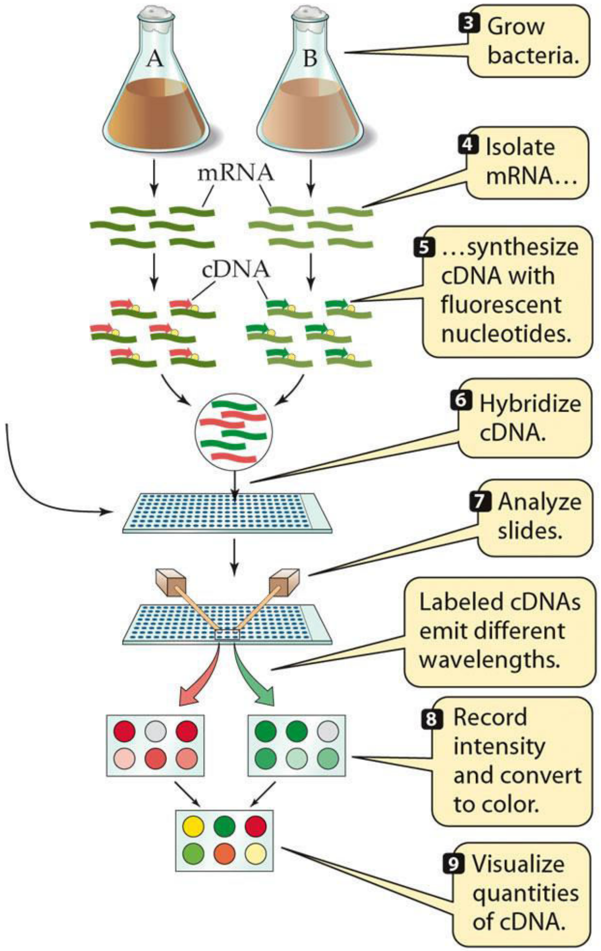
Microarray analysis



MICROBIAL LIFE, Figure 16.13 (Part 1) © 2002 Sinauer Associates, Inc.

Steps in the microarray analysis of all genes expressed in a particular bacterium grown under two different laboratory conditions, designated A and B. These could be variations in physical environment or nutritional content, or comparisons of a mutant to a wild-type cell

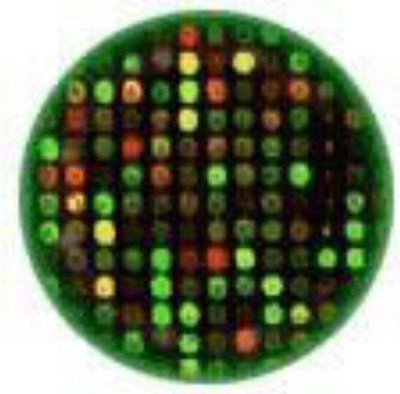
Microarray analysis


A
B
3 Grow bacteria.
mRNA
4 Isolate mRNA…
cDNA
5 …synthesize cDNA with fluorescent nucleotides.
6 Hybridize cDNA.
7 Analyze slides.
Labeled cDNAs emit different wavelengths.
8 Record intensity and convert to color.
9 Visualize quantities of cDNA.

MICROBIAL LIFE, **Figure 16.13 (Part 2)** © 2002 Sinauer Associates, Inc.

Steps in the microarray analysis of all genes expressed in a particular bacterium grown under two different laboratory conditions, designated A and B. These could be variations in physical environment or nutritional content, or comparisons of a mutant to a wild-type cell.

The Colors of a Microarray



Reproduced with permission from the Office of Science Education, the National Institutes of Health

In this schematic:

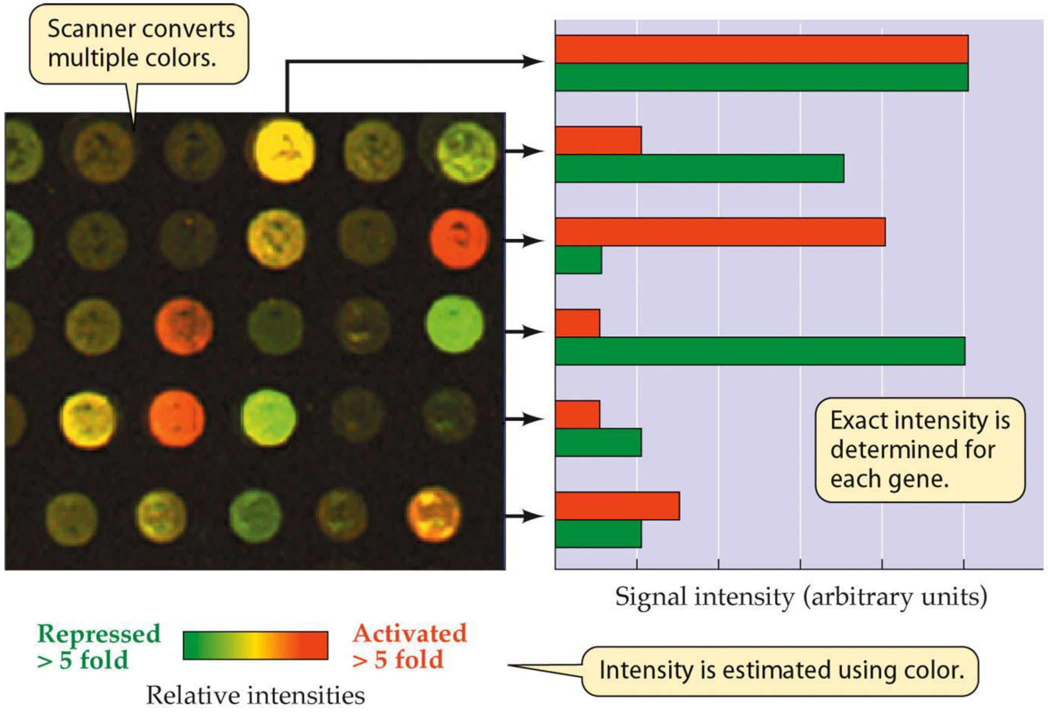
GREEN represents **Control DNA** where either DNA or cDNA derived from normal tissue is hybridized to the target DNA.

RED represents **Sample DNA** where either DNA or cDNA is derived from diseased tissue hybridized to the target DNA.

YELLOW represents **a combination of Control and Sample DNA** where both hybridized equally to the target DNA.

BLACK represents areas where **neither the Control nor Sample DNA** hybridized to the target DNA.

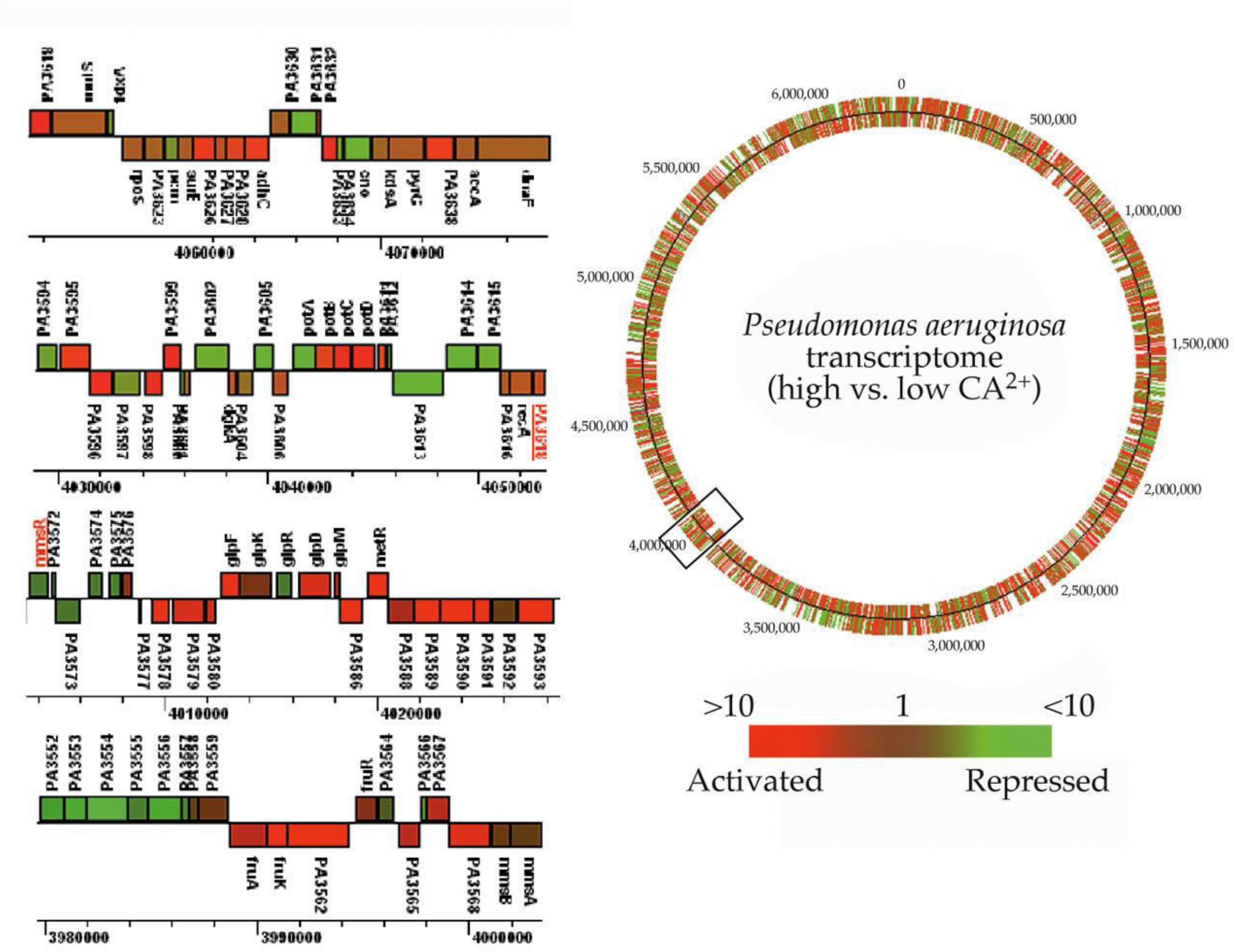
Analysis of data from a microarray experiment



Scanner converts multiple colors.
Exact intensity is determined for each gene.
Signal intensity (arbitrary units)
Repressed > 5 fold
Activated > 5 fold
Relative intensities
Intensity is estimated using color.

MICROBIAL LIFE, **Figure 16.14** © 2002 Sinauer Associates, Inc.

Transcriptome



MICROBIAL LIFE , **Figure 16.15** © 2002 Sinauer Associates, Inc.

Where to from here?

The top down approach has not and will not, in the foreseeable future, replace the bottom up approach. The different strategies are complementary.

High through-put technologies are changing the face of biology in general and microbiology in particular. There is a tremendous amount of information which can only be accessed computationally.

The computer analysis of genome sequences give us predictions and definitive gene function can only be assigned by laboratory experimentation.

Computer based predictions provide a firm basis for the design of new experiments.