Chapter 15

Microbial Genomics

Genomics

- study of molecular organization of genomes, their information content, and gene products they encode
- divided into three areas
  - structural genomics
    - physical nature of genomes
  - functional genomics
    - how genome functions
  - comparative genomics
    - compares genomes of different organisms

Determining DNA Sequences

- Sanger Method
  - uses dideoxynucleoside triphosphates (ddNTP)
- automated systems
  - use dideoxynucleotides labeled with fluorescent dyes

Dideoxyadenosine triphosphate

used as substrate by DNA polymerase

terminates DNA synthesis

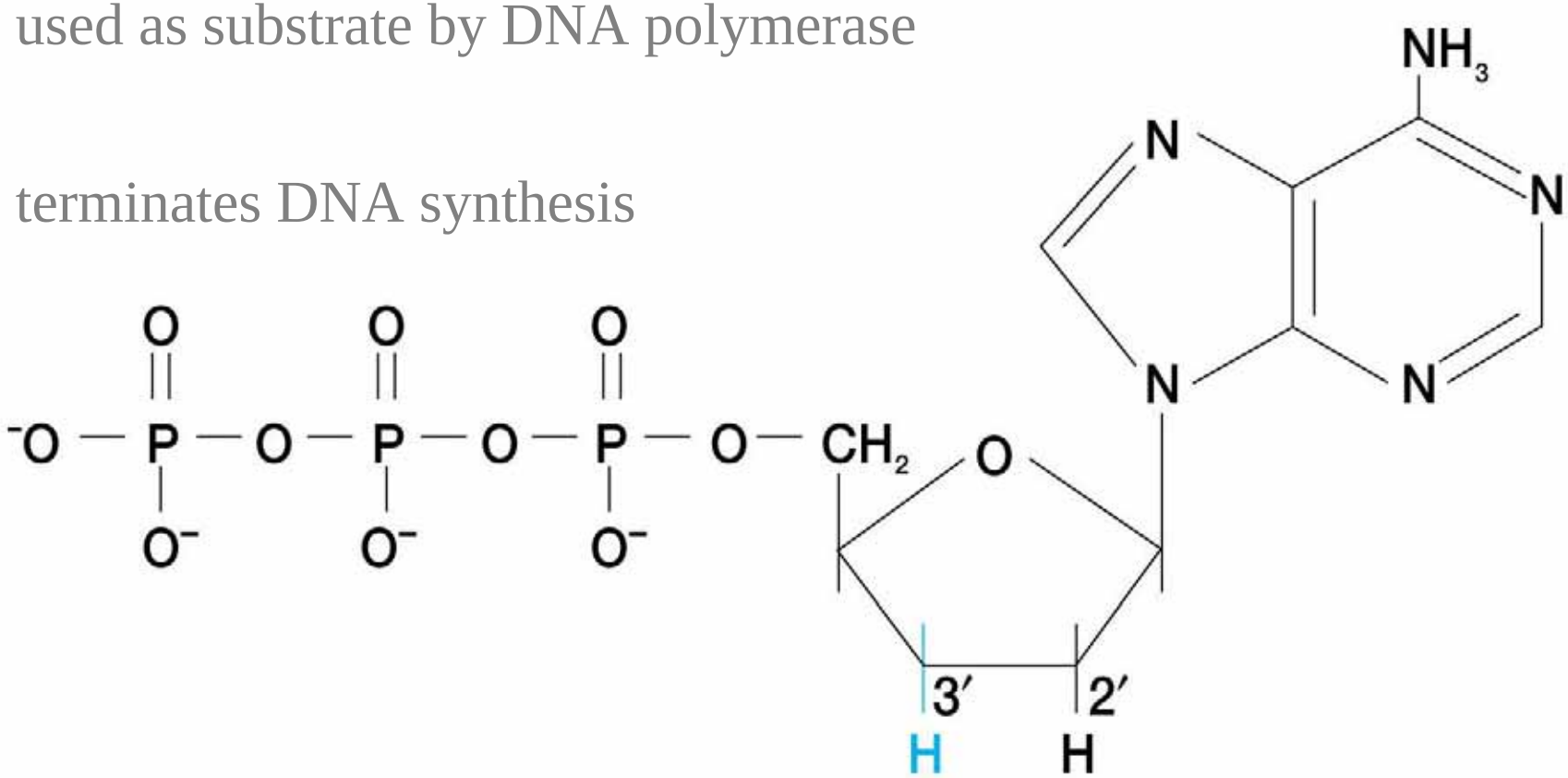

Figure 15.1

Sanger method

- mix single strands of DNA with primer, DNA polymerase I, 4 deoxynucleotides (one of which is radiolabeled), small amount of one ddNTP
- DNA synthesis occurs; random insertion of ddNTP generates DNA fragments of different lengths
- four reactions carried out; each with different ddNTP
- fragments in each reaction mixture separated electrophoretically
- gel autoradiographed and sequence read

Automated methods

work on same principle as Sanger method but use fluorescent labels

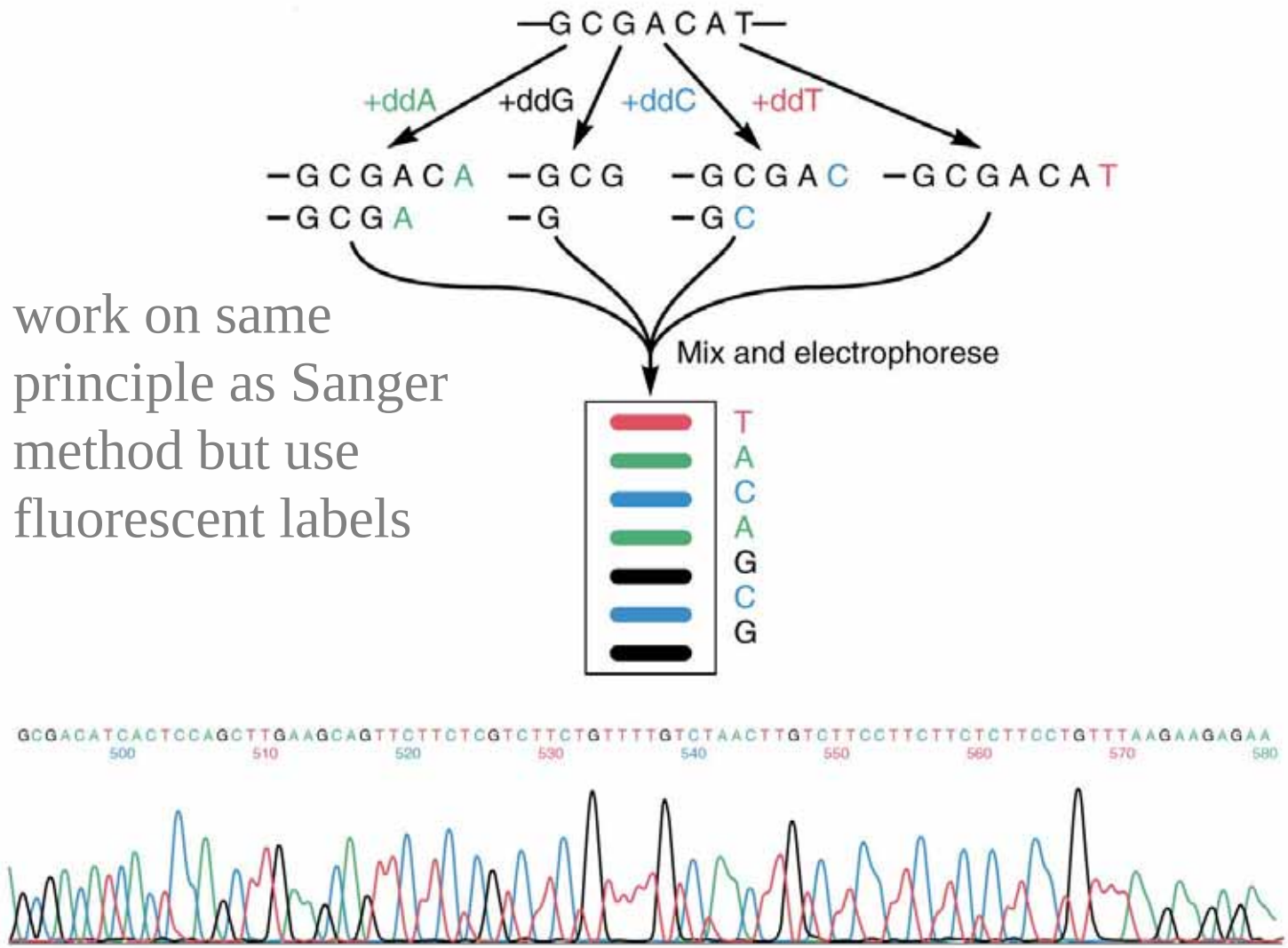

Figure 15.2

Whole-Genome Shotgun Sequencing

- developed in 1995 by J. Craig Venter and Hamilton Smith
- four stage process
  - library construction
    - generates clones of portions of genome
  - random sequencing
    - determines sequences of clones
  - fragment alignment and gap closure
  - editing

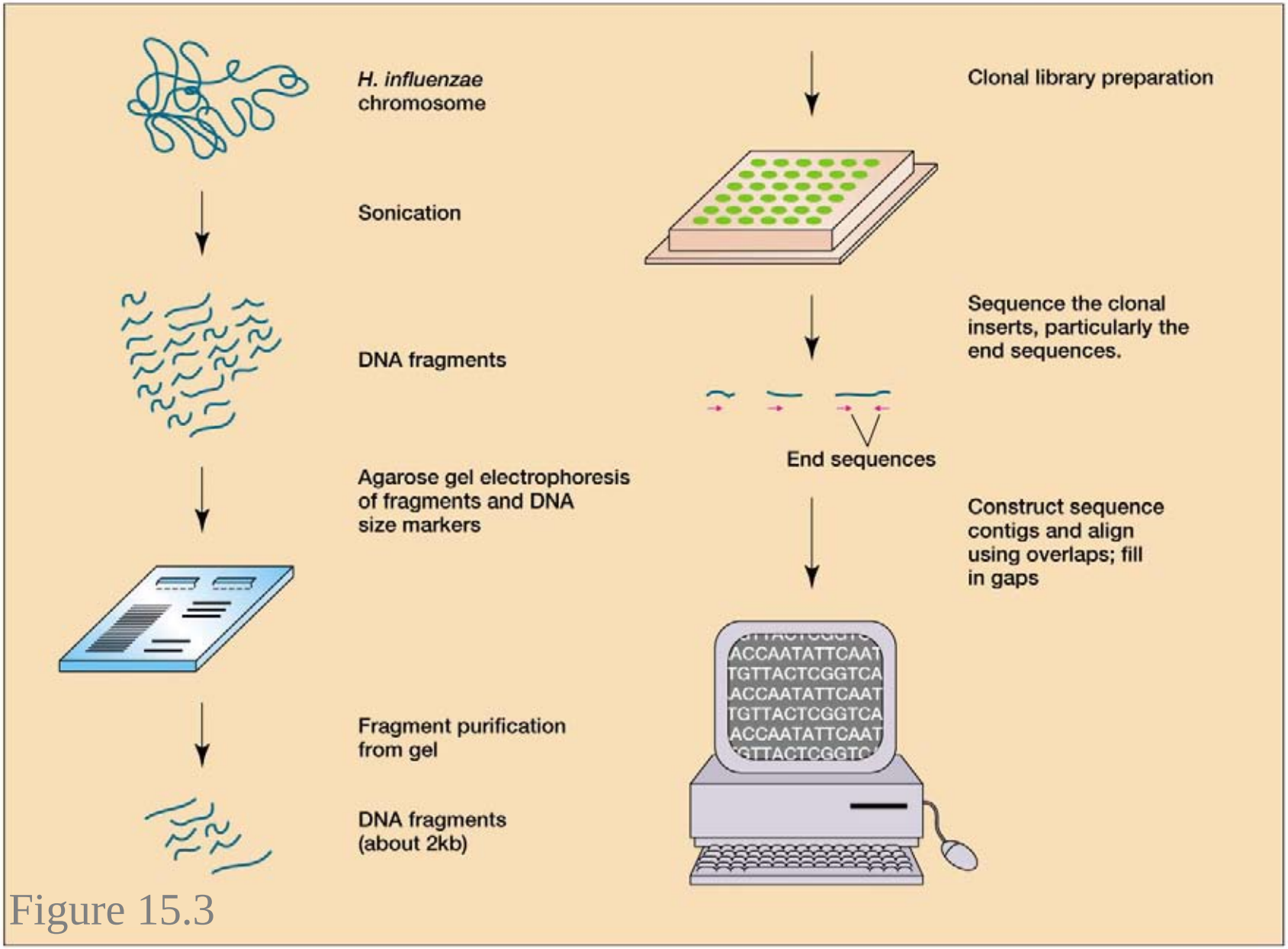
H. influenzae chromosome
Sonication
DNA fragments
Agarose gel electrophoresis of fragments and DNA size markers
Fragment purification from gel
DNA fragments (about 2kb)
Clonal library preparation
Sequence the clonal inserts, particularly the end sequences.
End sequences
Construct sequence contigs and align using overlaps; fill in gaps
ACCAATATTCAAT
TGTTACTCGGTCA
ACCAATATTCAAT
TGTTACTCGGTCA
ACCAATATTCAAT

Figure 15.3

Annotation

- process that locates genes in the genome map
- identifies each open reading frame in genome
  - a reading frame > 100 codons that is not interrupted by a stop codon
- uses databases to assign tentative function of gene

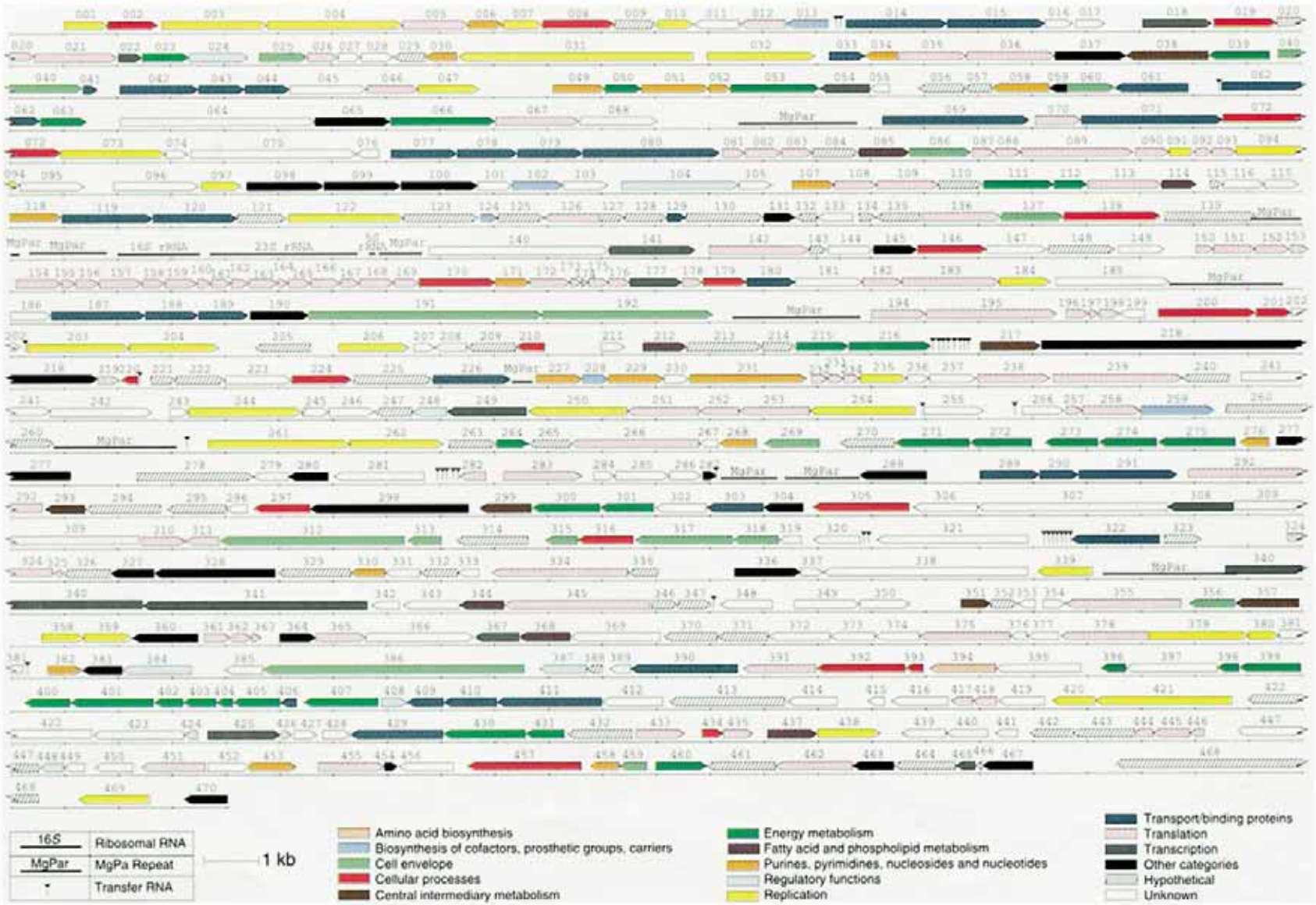
16S
Ribosomal RNA
MgPar
MgPa Repeat
Transfer RNA
1 kb
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid and phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Transport/binding proteins
Translation
Transcription
Other categories
Hypothetical
Unknown

Figure 15.5

Bioinformatics

- field concerned with management and analysis of biological data
- uses computers to do so
- DNA sequence data stored in large databases
  - e.g., International Nucleic Acid Sequence Data Library (GenBank)

General Characteristics of Microbial Genomes

- numerous genomes have been completed
- examination has led to formulation of numerous hypotheses about microbial genomes

Table 15.1
Examples of Complete Published Microbial Genomes

Genome	Domain*	Size (Mb)	% G + C
Agrobacterium tumefaciens	B	5.67†	57–63
Aquifex aeolicus	B	1.50	43
Archaeoglobus fulgidus	A	2.18	48
Bacillus subtilis	B	4.20	43
Borrelia burgdorferi	B	1.44	28
Campylobacter jejuni	B	1.64	31
Caulobacter crescentus	B	4.02	62–67
Chlamydia pneumoniae	B	1.23	40
Chlamydia trachomatis	B	1.05	41
Chlorobium tepidum	B	2.15	57
Clostridium perfringens	B	3.03	29
Corynebacterium glutamicum	B	3.3	55–58
Deinococcus radiodurans	B	3.28	67
Escherichia coli	B	4.60	50
Haemophilus influenzae Rd	B	1.83	39
Halobacterium sp. NRC-1	A	2.57	68
Helicobacter pylori	B	1.66	39
Listeria monocytogenes	B	2.9	37–39
Methanobacterium thermoautotrophicum	A	1.75	49
Methanococcus jannaschii	A	1.66	31
Mycobacterium leprae	B	3.27	58
Mycobacterium tuberculosis	B	4.40	65
Mycoplasma genitalium	B	0.58	31
Mycoplasma pneumoniae	B	0.81	40
Neisseria meningitidis	B	2.27	51
Nostoc sp. PCC 7120	B	6.41	41
Pseudomonas aeruginosa	B	6.3	67
Pyrococcus horikoshii	A	1.80	42
Rickettsia prowazekii	B	1.10	29
Saccharomyces cerevisiae	E	13	38
Salmonella typhimurium	B	4.86	50–53
Staphylococcus aureus	B	2.8	33
Streptococcus mutans	B	2.03	37
Streptococcus pneumoniae	B	2.16	40
Streptococcus pyogenes	B	1.9	39
Streptomyces coelicolor	B	8.67	72
Sulfolobus tokodaii	A	2.69	33
Synechocystis sp.	B	3.57	47
Thermoplasma acidophilum	A	1.56	46
Thermotoga maritima	B	1.80	46
Treponema pallidum	B	1.14	52
Vibrio cholerae	B	4.0	48
Yersinia pestis	B	4.65	48

*The following abbreviations are used: A, Archaea; B, Bacteria; E, Eucarya.
†The genome contains a circular chromosome, a linear chromosome, and two plasmids (AT and Ti).

Some interesting findings

- minimal genome size
  - based on analysis of *Mycoplasma genitalium* genome
    - smallest procaryotic genome sequenced
    - ~108-121 genes not required for growth in laboratory
    - ~265-350 genes required for growth in laboratory

More findings…

- many identified genes have unknown function
  - e.g., *Mycoplasma genitalium*
    - 22% have unknown function
  - e.g., *Haemophilus influenzae*
    - > 40% have unknown function
  - e.g., *Methanococcus jannaschii*
    - a member of Archaea
    - 66% have unknown function
  - e.g., *E. coli*
    - ~2500 of 4288 genes have unknown function

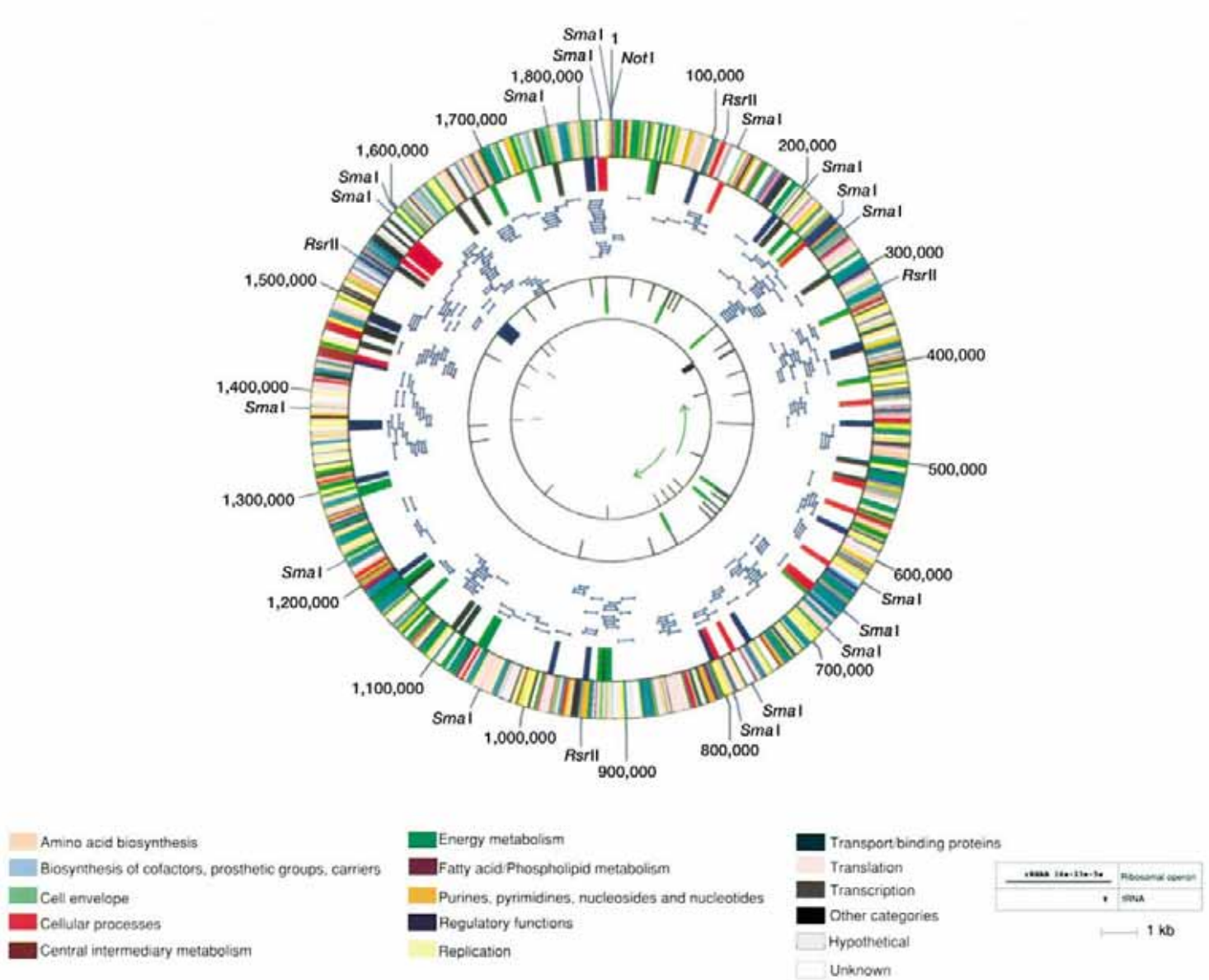
SmaI
1
SmaI
NotI
1,800,000
100,000
SmaI
RsrII
1,700,000
SmaI
200,000
1,600,000
SmaI
SmaI
SmaI
SmaI
RsrII
300,000
1,500,000
RsrII
400,000
1,400,000
SmaI
500,000
1,300,000
SmaI
600,000
1,200,000
SmaI
SmaI
SmaI
700,000
1,100,000
SmaI
SmaI
SmaI
1,000,000
800,000
RsrII
900,000
Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid/Phospholipid metabolism
Purines, pyrimidines, nucleosides and nucleotides
Regulatory functions
Replication
Transport/binding proteins
Translation
Transcription
Other categories
Hypothetical
Unknown
Ribosomal operon
tRNA
1 kb

Figure 15.6

More findings…

- evolutionary relationships

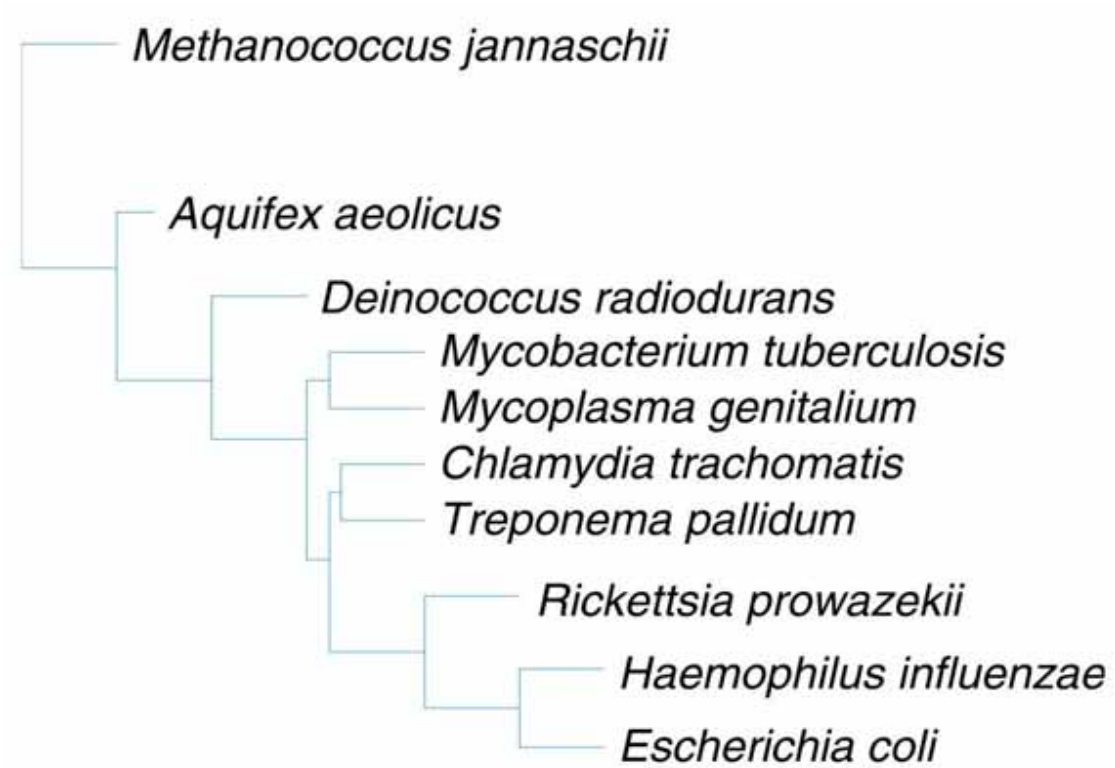

Figure 15.4

More findings…

- deinococci
  - radiation resistant bacteria
  - have large number of DNA repair genes
- *Rickettsia prowazekii*
  - causes typhus fever
  - closely related to ancient bacterium that gave rise to mitochondria by endosymbiosis

More findings…

- chlamydiae
  - despite unusual life cycles, lack of peptidoglycan, and parasitic life styles, have genomes similar to many other bacteria
    - e.g., genes for ATP synthesis
    - e.g., genes for peptidoglycan synthesis
  - other surprises
    - lack gene thought to be required for cell division
    - at least 20 genes obtained from eucaryotic hosts, including plant genes

More findings: *Treponema pallidum*…

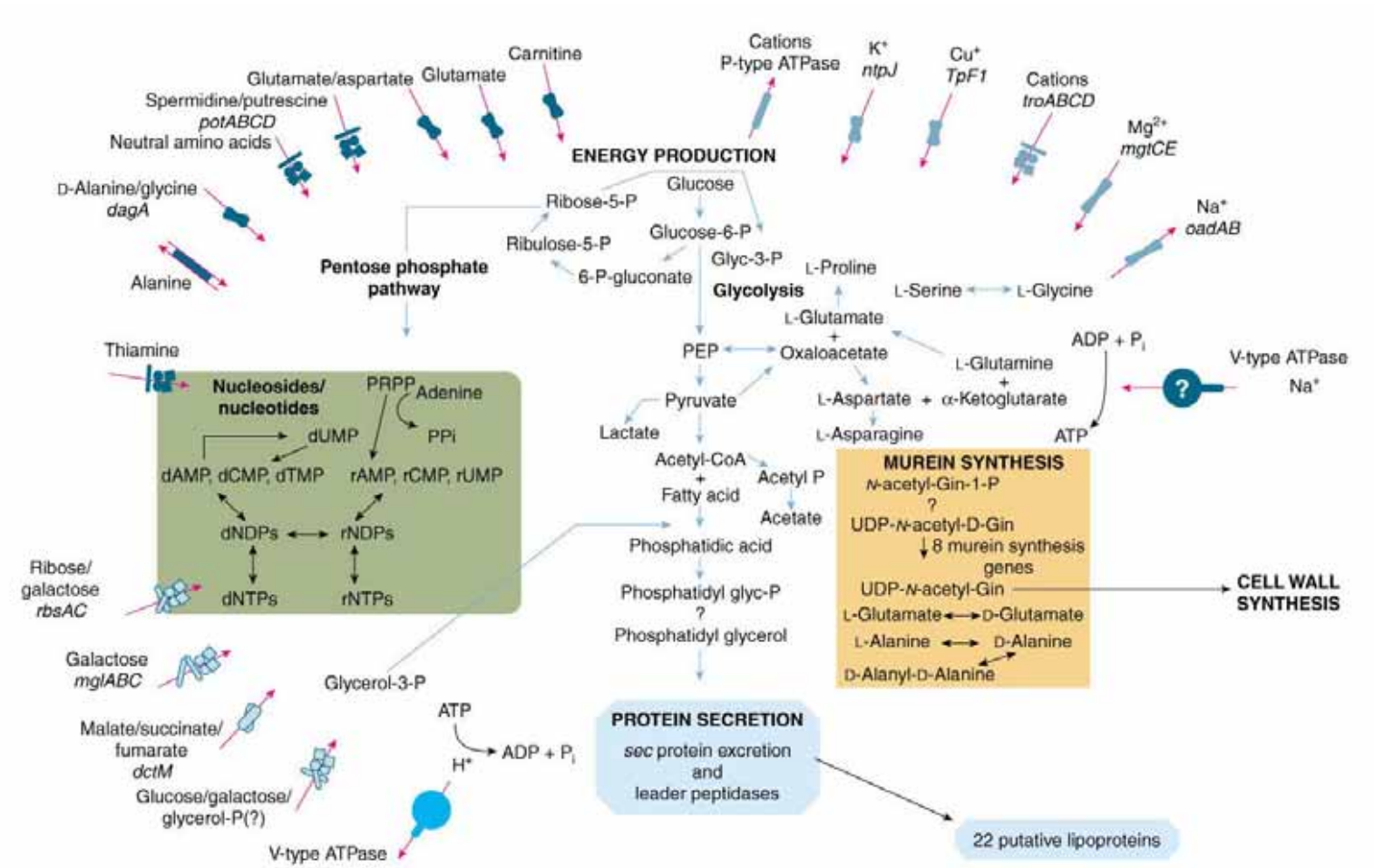

Figure 15.7

metabolically crippled

More findings…

- *Mycobacterium tuberculosis* and *Mycobacterium leprae*
  - though closely related have very different sized genomes
    - *M. tuberculosis* – large genome
      - >250 genes devoted to lipid synthesis
      - large number of regulatory genes
    - *M. leprae* – much smaller genome
      - ~ half of genome devoid of functional genes

More findings…

- *Staphylococcus aureus*
  - several antibiotic resistance genes on plasmids and transposons
  - virulence genes on genomic islands
- *Streptococcus pyogenes*
  - prophages contain genes for virulence factors
- *Streptococcus pneumoniae*
  - genes for enzymes that destroy host cell membranes
  - ~5% of genome consists of insertion sequences

General patterns

- despite conservation of protein sequences, genome organization is very variable
- considerable horizontal gene transfer has occurred in evolution of bacteria and archaea

Functional Genomics

- determination of how genome works
- three common approaches
  - genome annotation
  - study of RNA-level gene expression
  - study of protein-level gene expression

Genome Annotation

Table 15.2 Estimated Number of Genes Involved in Various Cell Functions[a]

Gene Function	*Escherichia coli* K12	*Bacillus subtilis*	*Mycoplasma genitalium*	*Treponema pallidum*	*Rickettsia prowazekii*	*Chlamydia trachomatis*	*Mycobacterium tuberculosis*	*Methanococcus jannaschii*	*Pyrococcus abyssi*
Approximate total number of genes[b]	2,933	2,232	477	757	523	847	2,095	1,271	1,345
Cellular processes[c]	179	123	6	77	27	43	65	26	44
Cell envelope components	146	86	29	53	36	42	50	25	25
Transport and binding proteins	304	223	33	59	18	57	87	56	67
DNA metabolism	97	80	29	51	39	53	57	53	33
Transcription	38	45	13	25	23	23	26	21	19
Protein synthesis	121	105	90	97	87	100	90	117	99
Regulatory functions	159	163	5	22	6	15	77	18	19
Energy metabolism[d]	351	230	33	54	48	61	211	158	116
Central intermediary metabolism[e]	64	61	7	6	6	12	57	18	25
Amino acid biosynthesis	89	97	0	7	9	13	72	64	51
Fatty acid and phospholipid metabolism	67	53	8	11	11	25	78	9	8
Purines, pyrimidines, nucleosides, and nucleotides	75	68	19	21	12	15	48	37	40
Biosynthesis of cofactors and prosthetic groups	97	79	4	15	17	31	84	49	31

[a]Data adapted from TIGR (The Institute for Genomic Research) databases.

[b]The number of genes with known or hypothetical functions.

[c]Genes involved in cell division, chemotaxis and motility, detoxification, transformation, toxin production and resistance, pathogenesis, adaptations to atypical conditions, etc.

[d]Genes involved in amino acid and sugar catabolism, polysaccharide degradation and biosynthesis, electron transport and oxidative phosphorylation, fermentation, glycolysis/gluconeogenesis, pentose phosphate pathway, Entner-Doudoroff, pyruvate dehydrogenase, TCA cycle, photosynthesis, chemoautotrophy, etc.

[e]Amino sugars, phosphorus compounds, polyamine biosynthesis, sulfur metabolism, nitrogen fixation, nitrogen metabolism, etc.

Evaluation of RNA-Level Gene Expression

- DNA microarrays (DNA chips)
  - solid supports (e.g., glass) that have DNA attached in highly organized arrays
    - expressed sequence tag (EST) – partial gene sequence unique to gene
  - incubated with labeled mRNA or cDNA (targets), which hybridize to complementary “spot” on chip
  - expressed genes “light up” on chip

construction
of DNA chip
by synthesizing
oligonucleotides
directly on glass
support

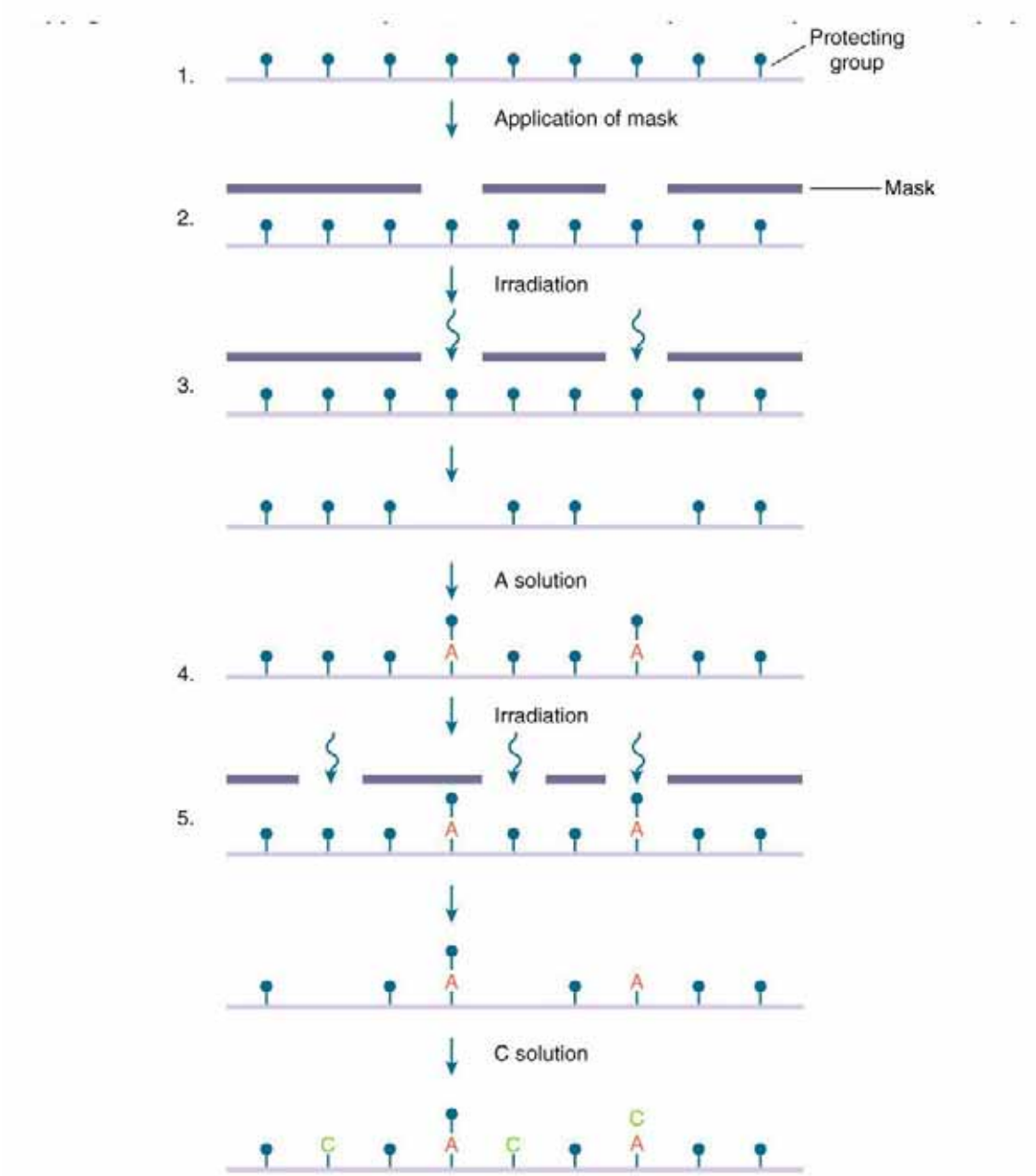

Figure 15.8

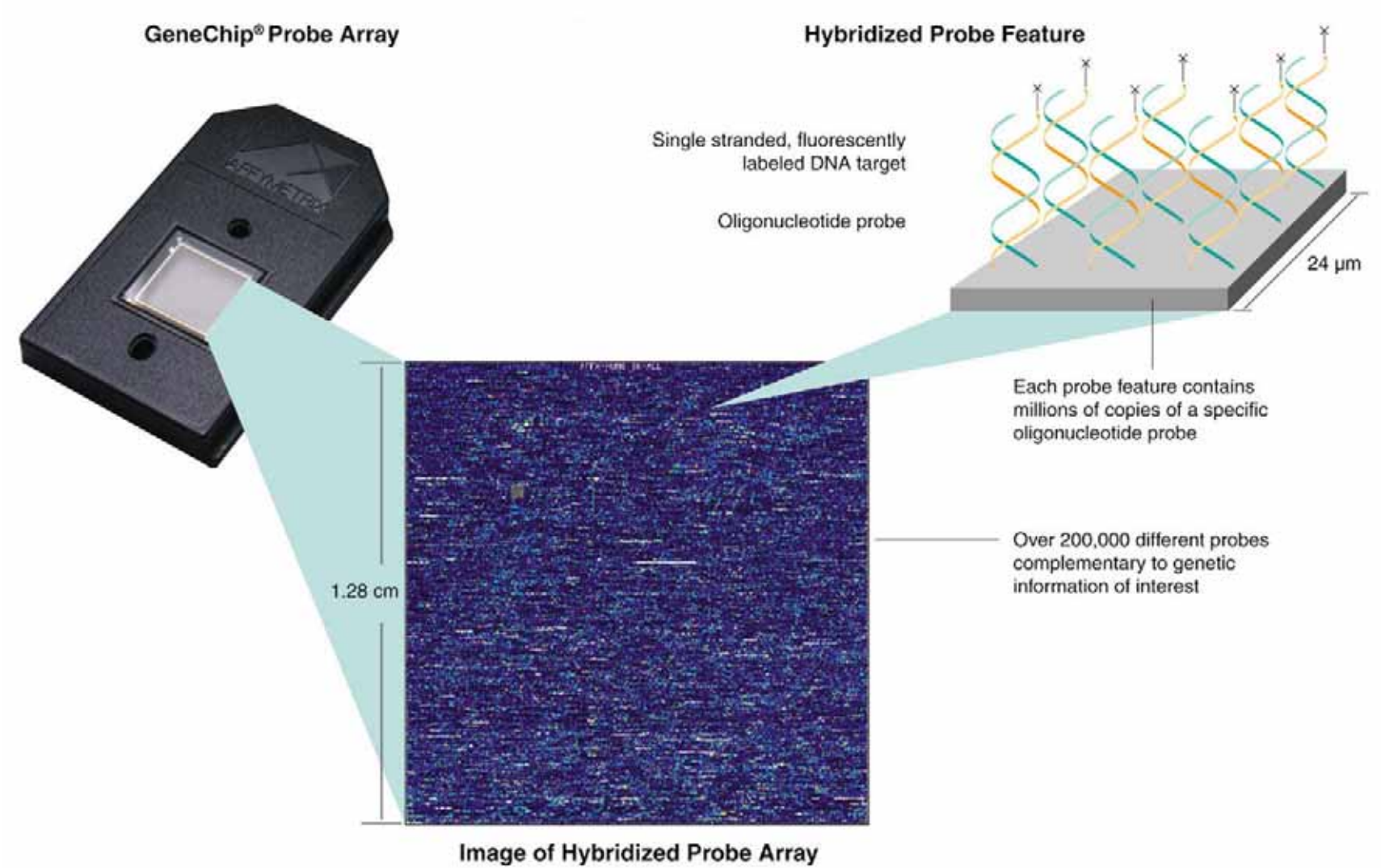
GeneChip® Probe Array
Hybridized Probe Feature
Single stranded, fluorescently labeled DNA target
Oligonucleotide probe
24 µm
Each probe feature contains millions of copies of a specific oligonucleotide probe
1.28 cm
Over 200,000 different probes complementary to genetic information of interest
Image of Hybridized Probe Array

Figure 15.9

Evaluation of Protein-Level Gene Expression

- proteome
  - entire collection of proteins that an organism produces
- proteomics
  - study of the proteome

Types of proteomic analysis

- functional proteomics
  - study of function of cellular proteins, how they interact, and ways they are regulated
- structural proteomics
  - using known protein structures to predict three-dimensional structures of other proteins and protein complexes

Evaluation of Protein-Level Gene Expression…

- usual approach is two-dimensional gel electrophoresis
- can also couple two-dimensional gel electrophoresis with mass spectrometry

Two-dimensional gel electrophoresis

- separation of a mixture of proteins along two dimensions
  - first dimension – isoelectric focusing
    - each protein moves through pH gradient until reaches pH that equals its isoelectric point
  - second dimension – SDS gel electrophoresis
    - proteins separated based on size

Two-dimensional gel electrophoresis

- when coupled with mass spectrometry, used to:
  - determine mass of each protein
  - determine amino acid composition or sequence of each protein
  - identify protein

The Future of Genomics

- many challenges and opportunities
  - e.g., development of new methods for large-scale analysis of genes and proteins
  - e.g., integration of information to understand workings of cell
  - e.g., insights into pathogenicity
  - e.g., development of new drugs to treat disease
  - e.g., identification of novel enzymes, biopesticides, and other molecules of industrial or agricultural importance