Principles of Genome Analysis

基因分析原理

Second Edition

A Guide to Mapping and Sequencing
DNA from Different Organisms

B.PRIMROSE

Chinese Edition

世界图书出版公司

Principles of Genome Analysis

A Guide to Mapping and Sequencing DNA from Different Organisms

S. B. PRIMROSE

SECOND EDITION

Blackwell
Science

世界图书出版公司
西安·北京·广州·上海

(陕)新登字 014 号

陕版出图字　著作权合同登记　25—1999—044 号

Principles of Genome Analysis
基因分析原理　第 2 版
by S. B. Primrose

任卫军　重印责任编辑
世界图书出版西安公司重印发行
(西安市南大街 17 号　邮编:710001)
西安七二二六印刷厂印刷
787×1092 毫米　开本 1/32　印张:6.5　字数:208 千字
1999 年 7 月第 1 次印刷
ISBN　7-5062-2253-1/R·391
Wx2253　　定价:30.00 元

Preface

Since the beginning of this century, a central problem in genetics has been the creation of maps of whole chromosomes. These maps are crucial for the understanding of the structure of genes, their function and their evolution. Until recently these maps were created by genetic means, i.e. as a result of sexual crosses. In the last 10 years the widespread use of recombinant DNA technology has permitted the generation of molecular or physical maps defined here as the ordering of distinguishable DNA fragments by their position along the chromosome. In some instances these physical maps can be as detailed as the complete DNA sequence of entire chromosomes.

Physical mapping of a wide range of genomes has occurred with incredible speed. Already a wide range of manipulative and analytical repertoires exist along with specialist journals and a specialist language. This means that it is very difficult for the experienced geneticist or molecular biologist entering the field to comprehend the latest developments or even what has been achieved. The aim of this text is to provide a general overview of the methodology and rationale employed. In such a fast-moving field it is difficult to be completely up-to-date but for those who worry about such a thing the literature available as late as September 1997 has been surveyed.

A number of individuals provided much appreciated assistance during the preparation of the manuscript and must be acknowledged here. In particular, John Armour and Jay Lewington read the draft text and made many suggestions for improving it. Most of their recommendations were incorporated but any errors or omissions which remain are entirely my responsibility. Thanks also are due to Vera Butterworth and Lynne Goodman for cheerfully finding and checking references at short notice. Without their help the preparation of this edition would have been much more difficult. This list would not be complete without a mention of my family, and my wife in particular, who were expected to provide support and sustenance during the long hours I spent hunched over a keyboard. Hopefully, the final product justifies their efforts and sacrifices!

Abbreviations

ACR	ancient conserved region
AFLP	amplified fragment length polymorphism
AFM	atomic force microscopy
APP	amyloid precursor protein
ARS	autonomously replicating sequence
BAC	bacterial artificial chromosome
BCR	bacterial conserved region
bp	base pair
CAPS	cleaved amplified polymorphic sequences
CCD	charged couple device
CEPH	Centre d'Etude du Polymorphisme Humain
CHEF	contour-clamped homogenous electrical field
cM	centimorgan
ct	chloroplast
DIRVISH	direct visual hybridization
DMD	Duchenne muscular dystrophy
EMC	enzyme mismatch cleavage
ERIC	enterobacterial repetitive intergenic consensus (sequence)
ES	embryonic stem (cells)
EST	expressed sequence tag
FACS	fluorescence-activated cell sorting
FIGE	field-inversion gel electrophoresis
FISH	fluorescence *in situ* hybridization
G6PD	glucose-6-phosphate dehydrogenase
GDRDA	genetically directed representational difference analysis
GMS	genome mismatch scanning
GSS	genome sequence sampling
HAC	human artificial chromosome
HAEC	human artificial episomal chromosome
HPRT	hypoxanthine phosphoribosyl transferase
HTF	*HPa*II Tiny Fragment
kb	kilobase

试读结束：需要全本请在线购买：

www.ertongboc

LINE	long interspersed nuclear element
LOD	logarithm$_{10}$ of odds
LTR	long terminal repeat
Mb	megabase
mt	mitochondrial
OFAGE	orthogonal-field-alteration gel electrophoresis
ORF	open-reading frame
PAC	P1-derived artificial chromosome
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
QTL	quantitative trait loci
RAPD	randomly amplified polymorphic DNA
RARE	RecA-assisted restriction endonuclease
RC	recombinant congenic (strains)
RDA	representational difference analysis
rDNA/RNA	ribosomal DNA/RNA
REP	repeated extragenic palindrome
RFLP	restriction fragment length polymorphism
RI	recombinant inbread (strains)
SINE	short interspersed nuclear element
STC	sequence-tagged connector
STM	scanning tunnelling microscopy
STS	sequence-tagged site
TAFE	transversely alternating-field electrophoresis
THC	tentative human consensus (sequence)
UTR	untranslated region
YAC	yeast artificial chromosome

Contents

Preface vii

Abbreviations ix

1 Rationale for mapping and sequencing genomes 1

2 The organization and structure of genomes 17

3 Subdividing the genome 45

4 Assembling a physical map of the genome 63

5 Sequencing methods and strategies 95

6 Analysing sequence data 126

7 Finding genes in large genomes 139

References 167

Index 189

1 Rationale for mapping and sequencing genomes

Introduction

Currently a whole series of international efforts is underway to construct genetic and physical maps and to sequence the genomes of a diversity of organisms ranging from the bacterium *Escherichia coli* to humans. The task is a Herculean one. On a global basis it is occupying thousands of scientists and is costing hundreds of millions of dollars annually. Given the effort being expended, the task clearly is a complex one. So, how is the task being approached and why is it being undertaken in the first place? This chapter will focus mainly on the second question, i.e. the rationale for mapping and sequencing genomes. The remainder of the book will concentrate on the methodology for mapping and sequencing genomes and the application of the knowledge gained.

Mapping genes and genomes

Mapping genes and the creation of genetic maps is a fundamental part of the science of genetics. This is because much of genetics is concerned with the understanding and manipulation of the inheritance of particular traits. For plants and animals of agronomic importance this means selective breeding and the identification of those offspring with the desired combination of characteristics. In the case of humans the objective is to predict whether the fetus carries genes for important inherited disorders, i.e. prenatal detection of genetic diseases. Where traits are associated with particular genes the task is relatively easy. Unfortunately there are many times more traits than identified genes and so geneticists make do with marker genes. These are genes which can easily be identified and which are genetically linked to the gene for the trait of interest.

In order to map the locus of a trait by genetic linkage, a panel of markers is tested in turn for evidence of co-segregation with the trait at meiosis. Two loci, A and B, are genetically linked if the alleles present at those loci on a particular chromosome tend to be transmitted together through meiosis. To be linked it is necessary, but not sufficient, for loci to be syntenic, i.e. on the same chromosome. The combination of alleles at linked loci is called a haplotype; for example, haplotype A1B1 means a single chromosome carrying allele A1 at locus A and allele B1 at locus B. During meiosis, each pair of homologous chromosomes undergoes at least one recombination (cross-over) between non-sister chromatids. To show genetic linkage, loci must be located in close physical proximity on a chromosome. Thus, to map a new gene it is necessary to have a large number of different markers, ideally evenly spaced along each chromosome. In micro-organisms such as *E. coli* and yeast, such markers can be generated quite easily by classical mutation techniques. Moving up the evolutionary tree, the generation of markers becomes increasingly difficult and in humans is ethically unacceptable. Instead, naturally occurring polymorphisms are used. By definition, all polymorphisms occur at the level of DNA. Unfortunately few of them are readily scorable like the phenotypes used by Mendel in his classical work on peas. The advent of recombinant DNA technology suggested a completely new approach to defining potentially large numbers of marker loci: the use of DNA probes to identify polymorphic DNA sequences (Botstein *et al.* 1980). The first such DNA polymorphisms to be detected were differences in the length of DNA fragments after digestion with sequence-specific restriction endonucleases, i.e. restriction fragment length polymorphisms (RFLPs) (Fig. 1.1).

To generate an RFLP map the probes must be highly informative. This means that the locus must not only be polymorphic, it must be *very* polymorphic. If enough individuals are studied, any randomly selected probe will eventually discover a polymorphism. However, a polymorphism in which one allele exists in 99.9% of the population and the other in 0.1% is of little utility since it seldom will be informative. Thus, as a general rule, the RFLPs used to construct the genetic map should have two, or perhaps three, alleles with equivalent frequencies.

The first RFLP map of an entire genome (Fig. 1.2) was that described for the human genome by Donis-Keller *et al.* (1987). They tested 1680 clones from a phage library of human genomic DNA to see whether they detected RFLPs by hybridization to Southern blots of DNA from five unrelated individuals. DNA from each individual was digested with 6–9 restriction enzymes. Over 500 probes were identified that detected variable banding patterns indicative of polymorphism. From this collection, a subset of 180 probes detecting the highest degree of polymorphism was selected for inheritance studies in 21 CEPH families (see below). Additional probes were generated from chromosome-specific libraries such that ultimately 393 RFLPs were selected. The various loci were arranged into linkage groups representing the 23 human chromosomes by a combination of mathematical linkage analysis and physical location of selected clones. The latter was achieved by hybridizing probes to panels of rodent–human hybrid cells containing varying human chromosomal complements (see p. 53). RFLP maps have not been restricted to the human genome. For example, RFLP maps have been published for most of the major crops (see for example Moore *et al.*, 1995).

The human genome map produced by Donis-Keller *et al.* (1987) was a landmark publication. However, it identified RFLP loci with an average spacing of 10 centimorgans (cM). That is, the loci had a 10% chance of recombining at meiosis. Given that the human genome is 4000 cM in length, the distance between the RFLPs is 10 Mb on average. This is too great to be of use for gene isolation. However, if the methodology of Donis-Keller *et al.* (1987) was used to construct a 1 cM map, then 100 times the effort would be required! This is because ten times as many probes would be required and ten times more families studied. The solution has been to use more informative polymorphic markers and other mapping techniques and these are described in detail in Chapter 4. Use of these techniques has led to the generation of a

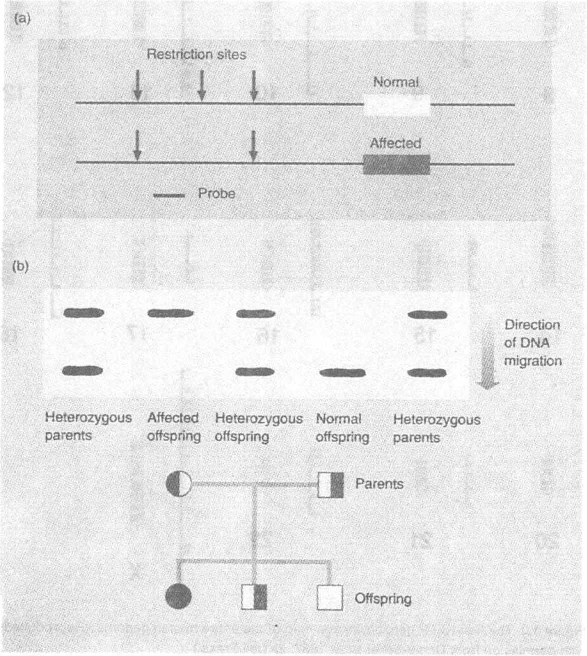
(a)
Restriction sites
Normal
Affected
Probe
(b)
Direction of DNA migration
Heterozygous parents
Affected offspring
Heterozygous offspring
Normal offspring
Heterozygous parents
Parents
Offspring

Figure 1.1 Example of a RFLP and its use for gene mapping. (a) A polymorphic restriction site is present in the DNA close to the gene of interest. In the example shown, the polymorphic site is present in normal individuals but absent in affected individuals. (b) Use of the probe shown in Southern blotting experiments with DNA from parents and progeny for the detection of affected offspring.

human map with a marker every 0.7 cM on average (Dib *et al.* 1996) and a mouse map with markers every 0.2 cM (Dietrich *et al.* 1996). More importantly, these advances in gene mapping have led to increased emphasis in developing representative genetic maps for a wide range of species, particularly domestic plants and

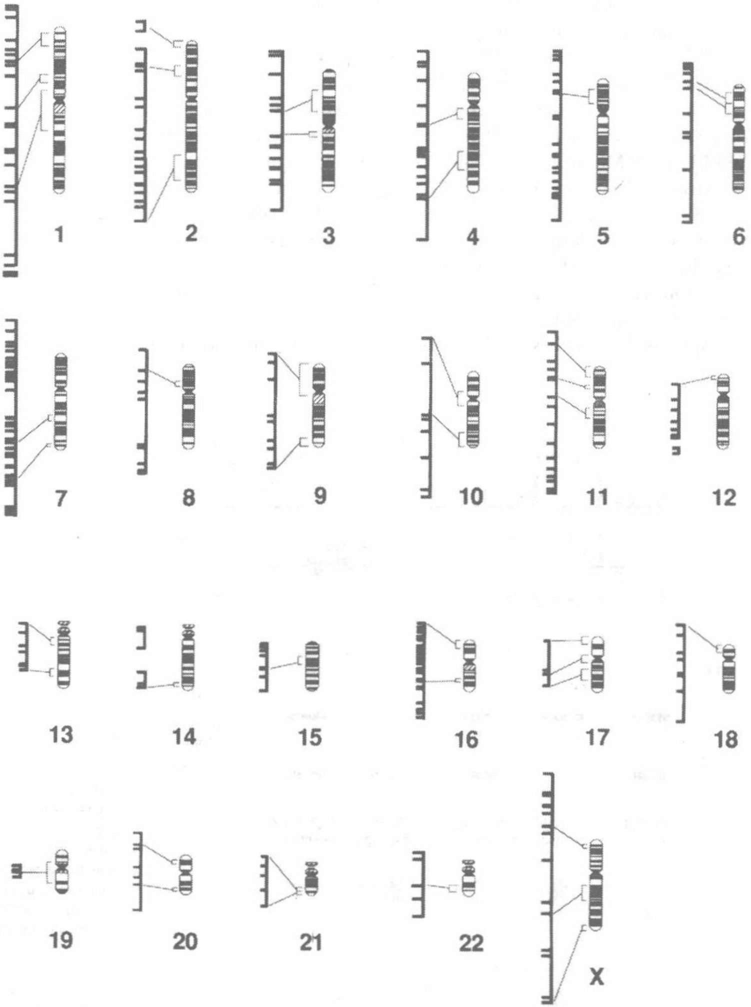

Figure 1.2 The first RFLP genetic linkage map of the entire human genome. (Reproduced with permission from Donis-Keller *et al.* 1987, © Cell Press.)

试读结束：需要全本请在线购买： www.ertongboc

animals, e.g. rice (Kurata *et al.* 1994) and pigs (Archibald 1994b).

It should be noted that humans represent an extreme case of difficulty in creating a genetic map. Not only are directed matings not possible, but the length of the breeding cycle (minimum 15–16 years) makes conventional analysis impractical. Consequently, the Centre d'Etude du Polymorphisme Humain (CEPH) was organized in Paris in 1984. The CEPH maintains cell lines from three-generation human families, consisting in most cases of four grandparents, two parents and an average of eight children (Dausset *et al.* 1990). Originally cell lines from 40 families were kept but the number now is much larger. Such families are ideal for genetic mapping because it is possible to infer which allele was inherited from which parent (Fig. 1.3). The CEPH distributes DNAs from these families to collaborating investigators around the world.

Understanding the phenotype

Using classical genetics it is relatively easy to show the mode of inheritance of a particular Mendelian trait, i.e. dominant vs. recessive, autosomal vs. sex linked. A full understanding of the biological basis of the phenotype requires a detailed knowledge of the appropriate gene(s), the genetic control of the gene(s) and identification of the gene product and its role in the life of the host. In some cases the jump from phenotype to gene analysis and biochemical explanation is relatively simple; for example, amino-acid auxotrophy and antibiotic resistance in micro-organisms or phenylketonuria and glucose-6-phosphate dehydrogenase (G6PD) deficiency in humans are phenotypes which are easy to analyse biochemically. However, for the vast majority of traits, including over 4000 in humans, no biochemical information is available. Recent developments in mapping and cloning have provided a solution, originally known as *reverse genetics* but now called

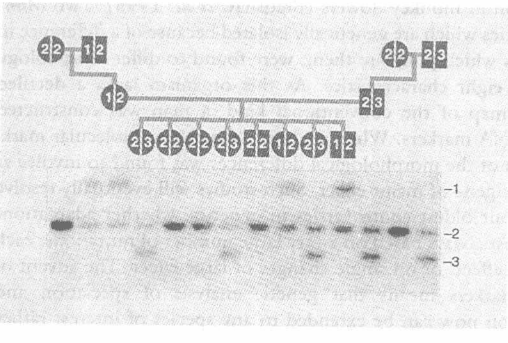

Figure 1.3 Inheritance of a RFLP in a CEPH family. The RFLP probe used detects a single locus on human chromosome 5. In the family shown, three alleles are detected on Southern blotting after digestion with *Taq*I. For each of the parents it can be inferred which allele was inherited from the grandmother and which from the grandfather. For each child the grandparental origin of the two alleles can then be inferred. (Redrawn with permission from Donis-Keller *et al.* 1987, © Cell Press.)

positional cloning. In essence, one maps the gene, clones it, sequences it and then deduces function. The methodology for doing this is described in Chapters 6 and 7. Among the successes achieved with the use of positional cloning, the most significant must be the determination of the biochemical deficiency responsible for cystic fibrosis. The genetics of the disease were well understood (autosomal recessive) and the phenotype (viscid secretions, chest infections, limited lifespan) was easily recognizable. However, until the gene was mapped, cloning and sequencing were not possible. Once sequence analysis could be undertaken it was relatively easy to show that cystic fibrosis is due to deletion of a phenylalanine codon in a gene encoding a chloride transport protein. Increasingly, the function of a gene in one organism, e.g. humans, is elucidated by comparison with a homologous gene in a completely unrelated organism such as yeast.

Comparative genome mapping

As noted above, recent advances in gene mapping have led to the development of representative genetic maps for a wide range of species. Initially these maps were compiled to provide a resource for genetic analysis in the species selected. More recently it has been realized that there is considerable conservation of gene order, not only in related species such as cereals, but in distantly related organisms such as the puffer fish and humans (see Chapter 7). This has two benefits. First, it is possible to transfer linkage information from 'map-rich' species to 'map-poor' species, thereby speeding up map construction. As more and more mapping takes place the increase in our databases should occur exponentially. Second, comparison of the maps should help dissect evolution of genome organization and provide clues about the adaptive rationale, if any, behind particular structural arrangements.

A good example of how molecular mapping helps the understanding of evolutionary processes is provided by an analysis of speciation in monkey flowers (Bradshaw *et al*. 1995). Two *Mimulus* species which are genetically isolated because of a difference in the birds which pollinate them, were found to differ morphologically in eight characteristics. As this organism lacks a detailed genetic map of the conventional kind, a map was constructed using DNA markers. When analysed using these molecular markers, each of the morphological differences was found to involve at least one gene of major effect. Such studies will eventually resolve one of our oldest controversies in genetics: whether adaptations are almost always based on a very large number of mutations, each of small effect, or on single changes of large effect. The advent of DNA markers means that genetic analysis of speciation and adaptation now can be extended to any species of interest rather than being restricted to those which have been intensively studied in the laboratory.

Why sequence genomes?

DNA is the genetic material in all cells. As such, it governs every facet of their existence. The information carried in the DNA determines when and how and where cells grow and divide; for example, why some yeast cells divide by budding and others by binary fission. But what are the differences at the DNA level? And why are some cells prokaryotic and others eukaryotic and what is the difference at the DNA level between simple and complex eukaryotes? Is there a core set of genes for all organisms? Can any organism be considered a 'model organism'. DNA has provided the basis for the evolutionary process that has generated the millions of different life-forms that exist on earth. We are all familiar with the story of Darwin's finches, but what happened at the level of the DNA? The same question can be asked for the monkey flower speciation example given above. Does evolution proceed in the ancestor–descendant pattern that Darwin saw for multicellular plants and animals, or are there other patterns, such as those of lateral gene transfer, obscuring the ancient relationships between domains of life? The answers to these questions, and many others like them, can be inferred from the vast knowledge we have of cellular and molecular biochemistry. However, they can only be answered definitively by studying the sequence of DNA in different organisms.

Although much of our knowledge of cellular and molecular biology was elucidated without DNA sequence information, this was cause rather than effect. Techniques for sequencing DNA did not exist. However, a complete understanding of many biological phenomena was dependent on the development of DNA sequencing techniques. Furthermore, recent experience has shown that the analysis of sequence data is a cost-effective way to generate answers to fundamental questions like those raised above, i.e. sequencing at the beginning of an investigation can be just as worthwhile, if not more so, than at the end. But is it necessary to sequence entire genomes? The answer to this must be in the positive.

Detailed understanding of an organism will only be achieved when every gene has been identified and its transcript and the timing of transcript synthesis known. As a minimum this demands knowledge of the complete gene sequence. In this context it is worth noting that when this was first available for a chromosome, that of yeast chromosome III (Oliver *et al.* 1992), the gene density was much higher than expected. An understanding of evolution will require comparative analysis of entire genomes rather than individual genes; for example, many bacterial genes are organized into operons which are transcribed as polycistronic mRNAs. By contrast, eukaryotic genes were thought to be regulated individually and transcribed as monocistronic mRNAs. Now analysis of 2 megabases (Mb) of DNA sequence from the nematode *Caenorhabditis elegans* has shown that it too has operons, i.e. it uses both the prokaryotic and eukaryotic patterns of gene organization (Zorio *et al.* 1994). Is the *C. elegans* genome becoming more compact to achieve bacterial status or is it expanding towards the eukaryotic monocistronic design?

Now that the complete genomic sequences are available for a number of bacteria and at least one eukaryote (see Chapter 5), it will soon be possible to identify at the DNA level what changes are necessary to go from being a prokaryote to being a eukaryote. Already an analysis of the sequences of the different bacteria is providing information on their physiology and biochemistry as well as their cultural characteristics and eco-pathology. Given that the complete sequence of one of these bacteria (*Mycoplasma genitalium*) was obtained for an estimated cost of only $200 000 then this information was obtained for a fraction of the cost of more conventional studies. The *M. genitalium* genome comprises only about 470 genes and is a big step towards what is the minimal gene complement compatible with a cellular existence. Based on comparisons with other genomic sequences it is possible that an organism might exist with only 200 genes (Koonin *et al.* 1996). With the knowledge acquired to date it might be possible to design experiments specifically focused on the discovery of bacteria with such a tiny genome that they might have escaped detection because of their inability to grow outside a host organism.

Already DNA sequence information is being used to great effect in disciplines such as palaeontology and archaeology (von Haeseler *et al.* 1995); for example, if the demographic history of a population is not known it can be reconstructed from the patterns of nucleotide substitutions in the genome. DNA sequences from the mitochondrial genome and those from the majority of the Y chromosome are particularly useful as they are passed on without recombination from mother to daughter and from father to son. Consequently, these sequences can be traced back directly to the genealogical maternal or paternal 'most recent common ancestor'. Molecular analysis of such sequences suggests that the human species originated in Africa only 100 000 to 200 000 years ago. From there it went on to colonize the world, replacing other human forms such as Neanderthals in the process.

Although the mapping of disease genes in humans, and useful traits in plants and animals, has been undertaken successfully in the past in the absence of sequence information, it now is generally recognized that a genome-wide effort will be more efficient in the long run. This is particularly true when one is seeking to control complex or quantitative traits (see Chapter 7) or to separate the genetic and environmental components of particular diseases. Pharmaceutical companies have taken a great interest in the human genome project for they see the data generated providing an understanding of multifactorial disease. This will enable them to design drugs that will treat the *causes* of disease rather than the *symptoms*, e.g. drugs that act at the level of transcription as opposed to the protein product or, alternatively, oligonucleotides for gene therapy. At the end of the day, the most detailed map available would be one in which every base pair had been identified. This being said, one must question the need to sequence the entire 3000 Mb of human DNA when there is a good chance that at least 2500 Mb of it will be uninformative (but see p. 104). After all, most of the single human gene disorders of any real frequency or medical importance have already been isolated. The real issue is that how much mapping and sequencing is of value depends on the type of genome and the questions one wants the genome map or sequence to answer. For example, it probably would not be worthwhile sequencing the sheep genome which contains a very high proportion of non-coding DNA, homologous with other ruminant genomes, if all we wished to study were a few quantitative traits of commercial importance. By contrast, the yeast genome has been completely sequenced and this was worthwhile: it is very gene dense and it can act as a model for the basic eukaryotic genome.

Genome sequencing projects

A large number of different genomes are being, or have been, sequenced and the most important are listed in Table 1.1. Each was selected for a different reason. There should be no surprise over the choice of *Escherichia coli*. Of all organisms it probably is the best characterized, both genetically and biochemically. *Bacillus subtilis* is of interest, partly because it is Gram-positive whereas *E. coli* is Gram-negative, and partly because it undergoes differentiation during the process of sporulation. The first cellular organism whose genome was completely sequenced was *Haemophilus influenzae* (Fleischmann *et al*. 1995) and it was selected simply because relatively little was known about it, i.e. it acted as a test case. This was followed fairly quickly by the complete sequences of *Mycoplasma genitalium* (Fraser *et al*. 1995), with its near 'minimal' genome, and *Methanococcus jannaschii* (Bult *et al*. 1996) which is a member of the Archaeae. Complete genomic sequencing of many other prokaryotes is being undertaken and the organisms selected are from all the major lineages.

One genome project is focusing on the smallest eukaryotic genomes identified so far, the nucleomorph genomes from the cryptophytic and chlorarachniophytic algae (McFadden *et al.* 1997). Nucleomorphs are remnant nuclei of former free-living eukaryotic algae that have been engulfed by another eukaryotic cell. The resulting chimaeric cell contains four genomes: the plastid, mitochondrion, host nucleus and nucleomorph (endosymbiotic nucleus). Coevolution between the two partner cells has resulted in the nucleus of the newest host assuming master control of plastid gene expression and the endosymbiotic nucleus has undergone significant reduction in genome size. All the nucleomorphs examined so far contain three small linear chromosomes. In chlorarachniophytes the nucleomorph genome is 380 kb in length and in cryptophytes it is 600 kb in length.

The first eukaryotic genome to be completely sequenced was that of *Saccharomyces cerevisiae* (Goffeau *et al.* 1996). It was selected because, like *E. coli*, it was well characterized genetically. In addition, because it has a nucleus and chromosomes, undergoes meiosis and mitosis etc., analysis of its genome sequence should provide much information on what constitutes a eukaryote. Already this analysis has begun (Oliver 1996a,b, 1997). Unlike most cells, *S. cerevisiae* multiplies by budding, hence the interest in *Schizosaccharomyces pombe* which divides by binary fission. Another reason for interest in *S. pombe* is that it has a genome of similar size

Table 1.1. Organisms whose genomes have been or are being sequenced. Up-to-date information can be found from the TIGR Database on the worldwide web (http://www.tigr.org.org/tdb/)

Genomes completely sequenced
Haemophilus influenzae (Fleischmann *et al.* 1995)
Mycoplasma genitalium (Fraser *et al.* 1995)
Mycoplasma pneumoniae (Himmelreich *et al.* 1996)
Methanococcus jannaschii (Bult *et al.* 1996)
Saccharomyces cerevisiae (Goffeau *et al.* 1996)
Synechocystis sp. (Kaneko *et al.* 1996)
Helicobacter pylori (Tomb *et al.* 1997)
Escherichia coli (Blattner *et al.* 1997)
Bacillus subtilis (in press)
Archaeglobus fulgidus (in press)
Borrelia burgdorferi (in press)
Methanobacterium thermoautotrophicum
Aquifex aeolicus
Treponema pallidum

Genomes being sequenced
Plasmodium falciparum (malarial parasite)
Schizosaccharomyces pombe (fission yeast)
Caenorhabditis elegans (nematode)
Drosophila melanogaster (fruit fly)
Danio rerio (zebra fish)
Fugu rubripes (puffer fish)
Arabidopsis thaliana (thale cress)
Oryza sativa (rice)
Mus musculis (mouse)
Rattus norvegicus (rat)
Homo sapiens (man)

试读结束：需要全本请在线购买： www.ertongboc