

## DNA, restriction enzymes

1. Small (viral) genomes may be composed of either RNA or DNA, but large (cellular) genomes are invariably composed of DNA. Why?

**Only DNA forms stable double-stranded structures, allowing complementary (redundant) information storage, which is required for high fidelity transmission of genetic information.**

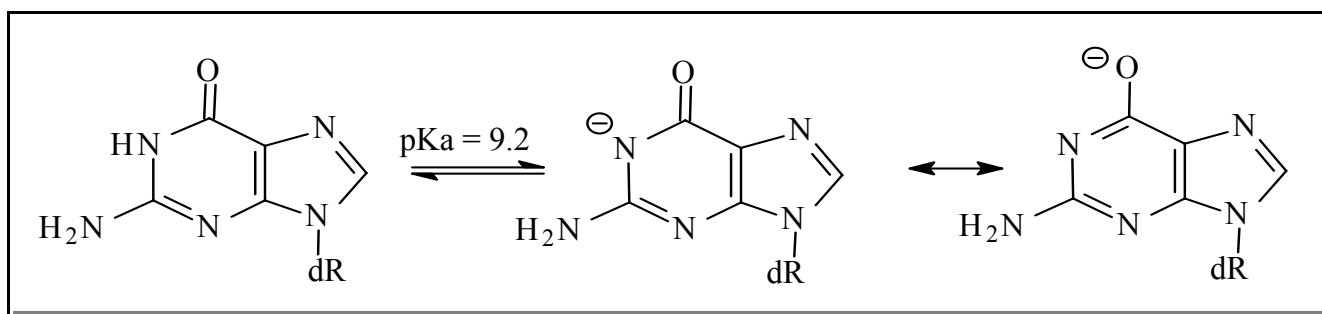
2. DNA is covalently stable in strongly alkaline solutions, but lipids, proteins, and RNA are not. Explain the action of base on those classes of molecules. Identify some biochemical techniques which exploit the stability of DNA under alkaline conditions.

The remarkable stability of DNA in strong base is exploited in several methods. For example, as we have seen, NaOH is used to denature DNA to single-stranded form in Southern blotting. DNA single-strand breaks can be detected by lysing cells under alkaline conditions (pH > 12) on sucrose gradients, followed by ultracentrifugation, as will be discussed later in the course (Yamada *et al.*, *Mutat. Res.* 364: 125-131, 1996). Electrophoresis of single cell nuclei to generate “comets”, following alkali treatment, is another popular technique for visualizing DNA damage (Tice, R.R., *Environ. Mol. Mutagen.* 35: 206-21, 2000).

Polypeptide amide bonds, like ester bonds, are labile to either acid- or base-catalyzed hydrolysis. Ester bonds in lipids are hydrolyzed, and so fats and phospholipids are saponified. RNA is hydrolysed by base, because deprotonation of the ribose 2'-OH group allows the 2' O atom to attack the P of the phosphodiester (this mechanism also occurs in the catalytic mechanism of the enzyme ribonuclease, which degrades RNA. In the enzymatic reaction, the enzyme acts as a general base catalyst). In contrast to RNA, DNA is stable to base, because the 2'-OH oxygen atom is not present. However, at pH > 9, many of the bases in DNA (and RNA, too, of course) ionize (see scheme).

The DNA double-strand is destabilized by the negative charge build-up on the bases, so linear dsDNA denatures, reversibly, under alkaline conditions.

(Note that, under acidic conditions, the glycosidic bonds of the purines in DNA are hydrolyzed, converting DNA into so-called “apurinic acid” - this was one of the techniques which Erwin Chargaff used to establish the base composition of DNA, back in the 1940s.)



3.	<i>NotI</i>	GCGGCCGC
	<i>HindIII</i>	AAGCTT
	<i>TaqI</i>	TCGA

a) How frequently does each of the above restriction enzymes cut DNA, on average, *i.e.*, what is the average length, in bp, of a DNA sample digested with each enzyme?

**There are 4 different bases, so the probability of finding a particular base at one location on a DNA strand = 1/4. So, the probability of finding the required base at each of n locations = (1/4)<sup>n</sup> (n = 4, 6, or 8). ∴ frequency of sites = 4<sup>-n</sup>**

$$4^4 = 256, 4^6 = 4,096, 4^8 = 65,536.$$

**If DNA strand length >> 1/ frequency, then: number of cuts ≈ bp × frequency  
Mean fragment size is 1/ frequency.**

b) The *Mycoplasma hominis* genome has 760 kbp of DNA; bacteriophage λ has 48.6 kbp; plasmid pBR322 has 4.3 kbp. Suppose we have isolated each of these DNA genomes. We are embarking on a “shotgun” sequencing program aimed at determining the sequence of the entire genome of each organism. Estimate the probable number of cuts and the average fragment size, following treatment of the sample with each enzyme.

c) How would you prepare DNA clones of suitable size (approx. 1-3 kbp) and homogeneity for use in DNA sequencing experiments, starting from each of these three genomes?

Total length of DNA (bp)		<i>Mycoplasma</i>	phage λ	plasmid pBR322
		$7.6 \times 10^5$	$4.86 \times 10^4$	$4.3 \times 10^3$
<i>number of cut sites expected (approximate)</i>				
<i>Not I</i>	1/ 65,536	12	0 or 1	none
<i>Hind III</i>	1/ 4,096	186	12	0 or 1
<i>Taq I</i>	1/ 256	2970	190	17

**Following cleavage with *TaqI*, plasmid pBR322 gives about a dozen fragments; each is about 250 bp, on average, and so each fragment is directly “ready to sequence”.**

**λ DNA would give too many fragments (>100) to separate easily on a gel, following digestion with *TaqI*. So, cleave λ DNA first with *HindIII*; separate these fragments (a dozen or so bands) on a gel. Each fragment can then be cut out and cleaved again with *TaqI*, to give pieces of an appropriate length. With *M. hominis* genomic DNA, a three-stage procedure is called for, first using *NotI*, then *HindIII*, and finally *TaqI*.**

4. Early efforts (c. 1994) in the human genome project were based on a systematic strategy of mapping clones along the chromosomes and sequencing them in an ordered fashion. But these

methods were overtaken by Celera Corporation's brute-force approach. In Celera's strategy, clones are isolated and sequenced willy-nilly, and then assembled into "contigs" by massive computational analysis. Suppose that we wish to isolate all of the genetic information of an organism, as a phage library of genomic DNA fragments. The length of the genome is  $G$  kbp and the genomic DNA fragments are of average length  $V$  kbp. The probability that a specific gene of interest is represented in any single clone is then  $p = V/G$ . (Obviously,  $p \ll 1$ , since the vector inserts are tiny compared to the size of the entire genome.)

(a) Derive a formula for the number ( $N$ ) of independent clones which must be sequenced, in order to be confident that 99% of the genes have been sequenced. (Hint: Write down the probability that a given clone does *not* contain the gene of interest. Raising this to the  $N^{\text{th}}$  power gives the probability that none of  $N$  independent clones contains the gene of interest. Finally, solve for  $N$ ).

$$P_{\text{yes}} = \frac{V}{G} \quad \therefore P_{\text{no}} = \left(1 - \frac{V}{G}\right) \quad \therefore \text{For } N \text{ clones, } P_{\text{not in } N} = \left(1 - \frac{V}{G}\right)^N$$

$$\therefore \log P_{\text{not in } N} = \log \left[\left(1 - \frac{V}{G}\right)^N\right] = N \log \left(1 - \frac{V}{G}\right)$$

$$N = \frac{\log P_{\text{not in } N}}{\log \left(1 - \frac{V}{G}\right)}$$

To  
simplify

the logarithm, we use the *Taylor's series expansion formula from elementary calculus*:

$$f(x) = f(a) + f'(a) \cdot (x-a) + (1/2!) \cdot f''(a) \cdot (x-a)^2 + \dots + (1/n!) \cdot f^{(n)}(a) \cdot (x-a)^n \dots$$

$$f(x + \alpha) = f(x) + \alpha f'(x) + \frac{\alpha^2}{2!} f''(x) + \dots$$

$$\therefore \ln(1 + \alpha) = \ln(1) + \alpha + \dots = 0 + \alpha + \dots$$

$$\text{And: } \log(a) = \frac{\ln(a)}{\ln(10)} = \frac{\ln(a)}{2.30}$$

$$\therefore N = \frac{2.30 \log P_{\text{not in } N}}{-\left(\frac{V}{G}\right)} = \frac{-G \cdot 2.30 \log P_{\text{not in } N}}{V}$$

(b) If the human genome (4 Gbp) is to be sequenced by cloning 5 kbp fragments, how many clones must be sequenced to be sure that we have obtained 99% of the genome?

For the human genome, 4 ≈ Gb and using 2 kb plasmid clones:

$$\therefore N = \frac{2.30 \log (.01)}{(-2 \times 10^3 / 4 \times 10^9)} = \frac{2.30 (-2)}{-5 \times 10^{-7}} \approx 10^7$$

or about 10 million clones.

More accurately, the genome is 3.1 Gbp, and the inserts used by Celera (see below) were actually about 500 bp long. These numbers give a denominator of  $500 / 3.1 \times 10^9 = 1.6 \times 10^{-7}$ ; so  $N = 4.6 / (1.6 \times 10^{-7}) = 2.85 \times 10^7 = 28.5$  million, which is very close to the number reported by Celera - see below.

From the Celera Corp. press release, June 26, 2000:

<http://www.pecorporation.com/press/prccorp062600.html>

“Celera Genomics today announced that it has completed the first assembly of the human genome, which has revealed a total of 3.12 billion base pairs in the human genome. ... Celera assembled the human genome using 26.4 million sequences of 550 base pairs long for a total of 14.5 billion base pairs sequenced, or  $4.6 \times$  sequence coverage. At  $4.6 \times$ , more than 99% of the genome is covered. Celera's whole genome shotgun sequencing technique involves sequencing from both ends of the double strands of DNA sequence. ... Celera shredded the data into 13.6 million segments 550 base pairs long for a total of 7.48 billion base pairs. ... The calculation to perform the assembly involved 500 million trillion base-to-base comparisons requiring over 20,000 CPU (central processor unit) hours on Celera's supercomputer. This believed to be the largest computational biology calculation to date.”

## DNA sequencing, PCR

1. A student isolates human genomic DNA, digests it with *HindIII*, runs the digested DNA on an agarose gel, and identifies a piece of DNA by the Southern blot procedure. She cuts the band out of the gel, extracts the DNA from the agarose, and tries to determine the sequence. Every band in every lane of the sequencing gel is labelled. What has gone wrong?

**The student assumed that the DNA was *pure*, just because there was only one visible band on the Southern blot. But the blotting techniques *detect*, they don't *purify* directly, and they are very selective: you only see DNA which is complementary to the probe. So, the DNA in the band that the student isolated did contain the DNA that was desired --- but a great deal of other DNA as well!**

2. The “wild-type” strain of the yeast *Pichia pastoris* is prototrophic (*i.e.*, it is able to grow on simple sugar/salts media). Beginning with this wild-type strain, we have isolated several auxotrophic mutants, each of which requires supplementation with histidine for growth. Enzyme assays demonstrate that each of these strains produces no detectable *histidinol dehydrogenase*, the 92 kDa enzyme which catalyzes the last step in histidine biosynthesis. We analyze the wild-type and mutant strains by a variety of molecular techniques. Interpret each of the following cases. That is, suggest the nature of the mutation (the DNA sequence change) which accounts for the observed phenotype, in each case.

a) In this case, a Southern blot of *Hind*III-digested genomic DNA is performed, using one entire exon of the wild-type gene as a labelled probe. The probe is hybridized to the nitrocellulose filter at a temperature of 41°C. The blot reveals a band at 4.3 kb for the wild-type strain but, for the mutant, no detectable hybridizing band is seen.

**This mutant is a deletion of (at least) the entire exon.**

b) In this case, the Southern blot (performed as described above) gives identical 4.3 kb bands for both the wild-type and mutant strains. However, a Northern blot of cellular mRNA, performed using the same probe, reveals an intense band at 2.6 kb for the wild-type strain but no detectable band for the mutant strain.

**This mutant lacks a functional promoter (perhaps due to a deletion or base-substitution upstream from the structural gene) , so no transcription occurs; but the structural gene (or, at least, one exon of it) remains.**

c) In this case, the Northern and Southern blots of the wild-type and mutant strains are identical. A Western blot (immunoblot) of the total cellular protein is undertaken. The blot is probed using immune serum raised in a rabbit by immunization with the purified wild-type histidinol dehydrogenase protein. The wild-type and mutant strains show a similar degree of staining of an approximately 92 kDa band. Next, a similar blot is probed using a *monoclonal antibody* to the enzyme, which we purchased from Rockland Immunochemicals, Inc. The 92 kDa band is again clearly visible for the wild-type strain, but it is hardly detectable in the mutant strain.

**This is a mis-sense mutation which alters the particular epitope recognized by the mAb, without much affecting the other epitopes on the protein. Perhaps the sequence of one peptide loop on the surface of the protein is altered.**

d) In this case, the Northern and Southern blots, and the Western blots, with both the polyclonal and the monoclonal antibodies, all give identical results for the wild-type and mutant strains.

**The only difference in the mutant enzyme is the change of a single amino acid residue - a critical one for catalysis - located in the active site of the enzyme and inaccessible to the antibodies.**

(Note: Other answers may be correct, too. These are some likely explanations.)

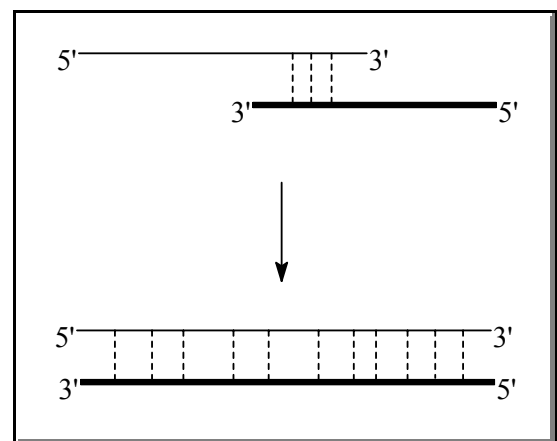
3. cDNA (complementary DNA) is DNA synthesized from an mRNA template by the action of *reverse transcriptase*. Two groups have isolated a cDNA from yeast, and wish to know if a similar gene is also present in mouse. Using the same radiolabelled cDNA probe, both groups do Southern blots on DNA from mice of the same strain. Group A reports that they have found the homologous gene, while group B finds nothing. Assuming that a related gene *is* actually present, and that no trivial mistakes were made by either group, what procedural difference might have caused the discrepancy?

**Even if the yeast and mouse genes code for homologous proteins, they will surely not be identical. The first lab must have used lower temperatures ( $T_M -15^\circ$  to  $T_M -10^\circ$ ) to obtain hybridization of the yeast cDNA to the Southern blot of mouse DNA. At this temperature, strand association will occur, even with substantial mismatches present. Caution is needed to avoid spurious matches. Lab number 2 must have incubated the Southern blot with yeast cDNA at higher temperatures ( $T_M -5^\circ$ ) where only exact matches will hybridize (stringent hybridization conditions).**

4. The Sanger dideoxy sequencing method can not be used to verify the sequence of a DNA oligonucleotide, such as a 20-mer. Why not?

**Sanger sequencing requires extension of a primer, and a 20-mer oligonucleotide is too short to allow the annealing of a primer. Short sequences like this must be analyzed by the Maxam-Gilbert (chemical-cleavage) sequencing protocol.**

5. Some simple rules can aid in the design of an oligonucleotide primer pair (*i.e.*, a set of one “forward” and one “reverse” primer) which give clean PCR (polymerase chain reaction) products. One of these rules is that one should avoid complementarity at the 3' ends of the primers, because complementarity promotes the formation of “primer-dimer” artifacts. What is “primer-dimer”, and why does complementarity at the 3' ends of the primers promote formation of primer-dimer? Explain.



**“Primer-dimer” is product formed by the extension of partially double-stranded molecules formed by the annealing of the two primers at their 3' ends (see right). Because primer is present at great molar excess relative to the target DNA, even a slight tendency of the primer pair towards primer-dimer formation can interfere with the PCR.**

6. The PCR amplifies the mass of DNA present in the reaction mix exponentially, approximately doubling with each cycle. This doubling process has to stop eventually, or else the mass of the DNA would soon exceed the mass of the solar system. :-) What factor, ultimately, limits the amount of DNA product that can be formed in a PCR reaction?

**Exhaustion of the supply of dNTPs.**

7. A bag contains 16 billiard balls, some black and the remainder white. Two balls are drawn at the same time. It is equally likely that the balls will be the same colour as different colours. How many balls of each colour are there?

**$b + w = 16$ . There are four possible draws:**

**black, black:  $p(bb) = (b/16) \cdot ((b-1)/15)$       white, white:  $p(ww) = (w/16) \cdot ((w-1)/15)$**

**black, white:  $p(bw) = (b/16) \cdot (w/15)$       white, black :  $p(wb) = p(bw)$**

$$\therefore (b/16) \cdot ((b-1)/15) + (w/16) \cdot ((w-1)/15) = 2 \cdot (b/16) \cdot (w/15)$$

**But  $w = 16-b$ ; substituting and simplifying gives the quadratic equation:**

$$b^2 - 16b + 60 = (b-6) \cdot (b-10) = 0 \qquad \qquad \qquad b = 6 \text{ (or } 10\text{).}$$