SUPPLEMENTS

7.32.1 DNA reassociation kinetics

Key Terms

A **Cot curve** is a plot of the extent of renaturation of DNA against time.

Complexity is the total length of different sequences of DNA present in a given preparation.

The **fast component** of a reassociation reaction is the first to renature and contains highly repetitive DNA.

Intermediate component(s) of a reassociation reaction are those reacting between the fast (satellite DNA) and slow (nonrepetitive DNA) components; contain moderately repetitive DNA.

The **slow component** of a reassociation reaction is the last to reassociate; usually consists of nonrepetitive DNA.

Nonrepetitive DNA shows reassociation kinetics expected of unique sequences.

Repetitive DNA behaves in a reassociation reaction as though many (related or identical) sequences are present in a component, allowing any pair of complementary sequences to reassociate.

The **repetition frequency** is the (integral) number of copies of a given sequence present in the haploid genome; equals 1 for nonrepetitive DNA, >2 for repetitive DNA.

Highly repetitive DNA (Simple sequence DNA) is the first component to reassociate and is equated with satellite DNA.

The **stringency** of a hybridization describes describes the effect of conditions on the degree of complementarity that is required for reaction. At the most stringent conditions, only exact complements can hybridize. As the stringency is lowered, an increasing number of mismatches can be tolerated between the two strands that are hybridizing.

A **tracer** is a radioactively labeled nucleic acid component included in a reassociation reaction in amounts too small to influence the progress of reaction.

The general nature of the eukaryotic genome can be assessed by the kinetics of reassociation of denatured DNA. Reassociation between complementary sequences of DNA occurs by base pairing. This reverses the process of denaturation by which they were separated (see **Figure 1.17**). The kinetics of the reassociation reaction reflect the variety of sequences that are present; so the reaction can be used to quantitate genes and their RNA products.

Figure S 1 describes the reaction. Renaturation of DNA depends on random collision of the complementary strands, and follows second-order kinetics. The reaction for any particular DNA can be characterized by conditions required for half-completion. This is the product of $C_0 \times t_{½}$ and is called the **Cot½**;. It is inversely proportional to the rate constant. Since the $Cot_{½}$ is the product of the concentration and time required

to proceed halfway, a greater $Cot_{½}$ implies a slower reaction.

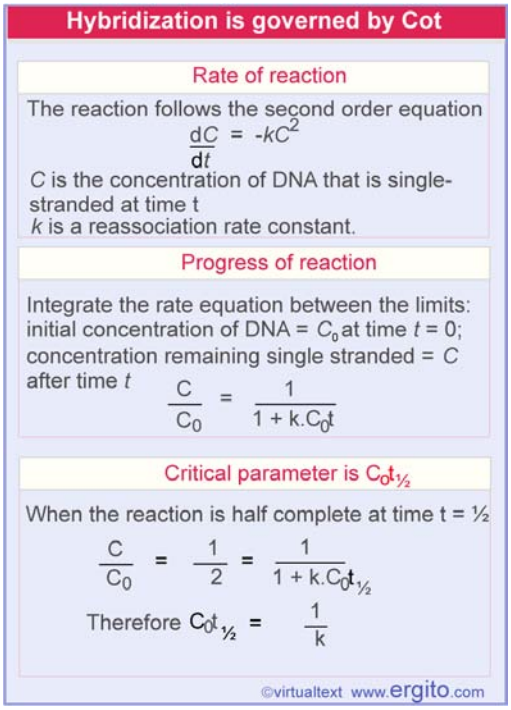

Figure S 1 A DNA reassociation reaction is described by the $Cot_{½}$.

The reassociation of DNA usually is followed in the form of a **Cot curve**, which plots the fraction of DNA that has reassociated ($1 - C/C_0$) against the log of the Cot. **Figure S 2** gives Cot curves for several simple genomes. The form of each curve is similar, with renaturation occurring over an ~100-fold range of Cot values between the points of 10% reaction and 90% reaction. But the $Cot_{½}$ for each curve is different.

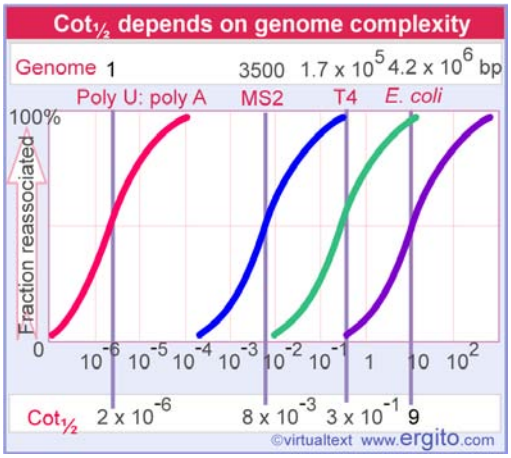

Figure S 2 Rate of reassociation is inversely proportional to the length of the reassociating DNA.

The genomes in **Figure S 2** represent a series of DNAs. Each is unique in sequence,

and they become progressively longer. *The $Cot_{½}$ is directly related to the amount of DNA in the genome.* This reflects a situation in which, as the genome becomes more complex, there are fewer copies of any particular sequence within a given mass of DNA. For example, if the C_0 of DNA is 12 pg, it will contain 3000 copies of each sequence in a bacterial genome whose size is 0.004 pg, but will contain only 4 copies of each sequence present in a eukaryotic genome of size 3 pg. So the same *absolute* concentration of DNA measured in moles of nucleotides per liter (the C_0) will provide a concentration of each eukaryotic sequence that is 3000/4 = 750× less than that of each bacterial sequence.

Since the rate of reassociation depends on the concentration of complementary sequences, for the eukaryotic sequences to be present at the same *relative* concentration as the bacterial sequences, it is necessary to have 750× more DNA (or to incubate the same amount of DNA for 750 times longer). So the $Cot_{½}$ of the eukaryotic reaction is 750× the $Cot_{½}$ of the bacterial reaction.

The $Cot_{½}$ of a reaction therefore indicates the *total length of different sequences* that are present. This is described as the **complexity**, usually given in base pairs. The Cot½ for the renaturation of the DNA of any genome (or part of a genome) is proportional to its complexity. The complexity of any DNA can be determined by comparing its $Cot_{½}$ with that of a standard DNA of known complexity. Usually *E. coli* DNA is used as a standard. Assuming that the *E. coli* genome of 4.2×10^6 bp consists of unique sequences:

When the DNA of a eukaryotic genome is characterized by reassociation kinetics, usually the reaction occurs over a range of Cot values spanning up to eight orders of magnitude. This is much broader than the 100-fold range expected from the examples of **Figure S 2**. The reason is that each of these curves follows the equation that describes the kinetics of reassociation for a single component. *A eukaryotic genome actually includes several such components, each reassociating with its own characteristic kinetics. The Cot curve reveals a crucial difference between bacterial and eukaryotic genomes: bacterial genomes essentially consist of a single kinetic component, but eukaryotic genomes are much more complex.*

Figure S 3 shows the reassociation of a (hypothetical) eukaryotic genome, starting at a Cot of 10^{-4} and terminating at a Cot of 10^4. The reaction falls into three distinct phases, outlined by the shaded boxes. Each of these phases represents a different kinetic component of the genome:

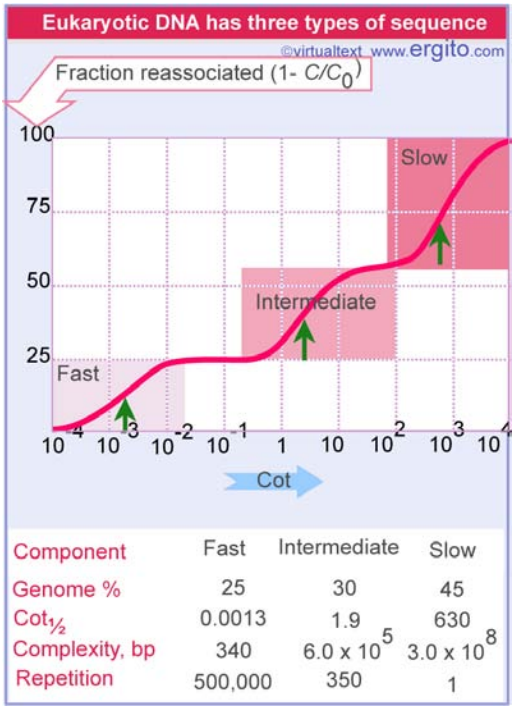

Figure S 3 The reassociation kinetics of eukaryotic DNA show three types of component (indicated by the shaded areas). The arrows identify the $Cot_{½}$ values for each component.

- The **fast component** is the first fraction to reassociate. In this case, it represents 25% of the total DNA, renaturing between Cot values of 10^{-4} and ~2×10^{-2}, with a $Cot_{½}$ value of 0.0013.
- The next fraction is called the **intermediate component**. This represents 30% of the DNA. It renatures between Cot values of ~0.2 and 100, with a $Cot_{½}$ value of 1.9.
- The **slow component** is the last fraction to renature. This is 45% of the total DNA; it extends over a Cot range from ~100 to ~10,000, with a $Cot_{½}$ of 630.

To calculate the complexities of these fractions, each must be treated as an independent kinetic component whose reassociation is compared with a standard DNA. The slow component represents 45% of the total DNA, so its concentration in the reassociation reaction is 0.45 of the measured C_0 (which refers to the total amount of DNA present). The $Cot_{½}$ applying to the slow fraction alone is 0.45 × 630 = 283.

Suppose that under these conditions, *E. coli* DNA reassociates with a $Cot_{½}$ of 4.0. This corresponds to a complexity for the slow fraction of 3.0×10^8 bp (= $4.2 \times 10^6 \times$ 283 /4). Treating the other components in a similar way shows that the intermediate component has a complexity of 6×10^5 bp, and the fast component has a complexity of only 340 bp. This provides a quantitative basis for our statement that, the faster a component reassociates, the lower is its complexity.

Reversing the argument, suppose we took three DNA preparations, each containing a unique sequence of the appropriate length (340 bp, 6×10^5 bp, and 3×10^8 bp, respectively) and mixed them in the proportions of mass 25:30:45. Each would renature as though it were a single component. Together the mixture would display the same kinetics as those determined for the whole genome of **Figure S 3**.

The complexity of the slow component corresponds with its physical size. Suppose that the genome reassociating in **Figure S 3** has a haploid DNA content of 7.0×10^8 bp, determined by chemical analysis. Then 45% of it is 3.15×10^8 bp, which is the same (within experimental error) as the value of 3.0×10^8 bp measured by the kinetics of reassociation. The complexity of the slow component corresponds to its physical length.

The slow component comprises sequences that are unique in the genome: on denaturation, *each single-stranded sequence is able to renature only with the corresponding complementary sequence.* This part of the genome is the sole component of prokaryotic DNA and is usually a major component in eukaryotes. It is called **nonrepetitive DNA**.

What is the nature of the components that renature more rapidly than the nonrepetitive (slow) DNA? In the example of **Figure S 2**, the intermediate component occupies 30% of the genome. Its chemical complexity is $0.3 \times 7 \times 10^8 = 2.1 \times 10^8$ bp. But its kinetic complexity is only 6×10^5 bp.

The unique length of DNA that corresponds to the $Cot_{½}$ for reassociation is much shorter than the total length of the DNA chemically occupied by this component in the genome. In other words, the intermediate component behaves as though consisting of a sequence of 6×10^5 bp that is present in 350 copies in every genome (because $350 \times 6 \times 10^5 = 2.1 \times 10^8$). Following denaturation, *the single strands generated from any one of these copies are able to renature with their complements from any one of the 350 copies.* This effectively raises the concentration of reacting sequences in the reassociation reaction, explaining why the component renatures at a lower $Cot_{½}$.

Sequences that are present in more than one copy in each genome are called **repetitive DNA**. The number of copies present per genome is called the **repetition frequency** (f).

Repetitive DNA is often classed into two general types, corresponding approximately to the intermediate and fast components of **Figure S 3**:

- *Moderately repetitive DNA* occupies the intermediate fraction, usually reassociating in a range between a Cot of 10^{-2} and that of nonrepetitive DNA.
- **Highly repetitive DNA** occupies the fast fraction, reassociating before a Cot of 10^{-2} is reached.

The behavior of a repetitive DNA component represents only an average that is useful for describing its sequences. The relevant parameters do not necessarily represent the properties of any particular sequence.

The moderately repetitive component of **Figure S 3** includes a total length of 6×10^5 bp of DNA, repeated ~350× per genome. But this does not correspond to a single, identifiable, continuous length of DNA. *Instead, it is made up of a variety of individual sequences, each much shorter,* whose total length together comes to 6×10^5 bp. These individual sequences are dispersed about the genome. Their average repetition is 350, but some will be present in more copies than this and some in fewer.

When a eukaryotic genome is analyzed by reassociation kinetics, the individual sequence components are rarely so well separated as shown in **Figure S 3**. In fact, they often overlap extensively, so that in reality there is probably a continuum of repetitive components, reassociating over a range from >10× to >20,000× that of the nonrepetitive component.

The different components of eukaryotic DNA can be isolated in the form of the DNA that becomes double-stranded after renaturation to a particular Cot value. The properties of renatured nonrepetitive and repetitive DNA differ significantly.

Nonrepetitive DNA forms duplex material that behaves very much like the original preparation of DNA before its denaturation. When denatured again, the duplex molecules melt sharply at a T_m only slightly below that of the original native DNA. This shows that strand reassociation has been accurate: each unique sequence has annealed with its exact complement.

Different behavior is shown by renatured repetitive DNA. The reassociated double strands tend to melt gradually over rather a wide temperature range, as shown in **Figure S 4**. This means that they do not consist of exactly paired molecules. Instead, they must contain appreciable mispairing. The more mispairing in a particular molecule, the fewer hydrogen bonds need be broken to melt it, and thus the lower the T_m.

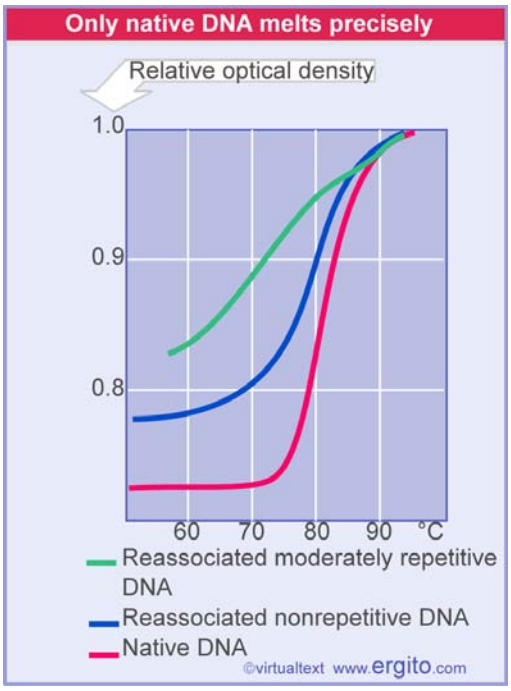

Figure S 4 The denaturation of reassociated nonrepetitive DNA takes place over a narrow temperature range close to that of native DNA, but reassociated repetitive DNA melts over a wide temperature range.

The breadth of the melting curve shows that renatured repetitive DNA contains a spectrum of sequences, ranging from those that have been formed by reassociation between sequences that are only partially complementary, to those formed by reassociation between sequences that are very nearly or even exactly complementary. How can this happen?

Repetitive DNA components consist of families of sequences that are not exactly the same, but are related. The members of each family consist of a set of nucleotide sequences that are sufficiently similar to renature with one another. The differences between the individual members are the result of base substitutions, insertions, and deletions, all creating points within the related sequences at which the complementary strands cannot base pair. The proportion of these changes establishes the relationship between any two sequences. When two closely related members of the family renature, they form a duplex with high T_m. When two more distantly related members associate, they form a duplex with a lower T_m. Overall, we see the broad range represented in the figure.

The ability of related but not identical complementary sequences to recognize each other can be controlled by the **stringency** of the conditions imposed for reassociation. A higher stringency is imposed by (for example) an increase in temperature, which requires a greater degree of complementarity to allow base pairing. So by performing the hybridization reaction at high temperatures, reassociation is restricted to rather closely related members of a family; at lower temperatures, more distantly related members may anneal. The measured size of a repetitive family is arbitrary, since it is determined by the hybridization conditions.

Moderately repetitive DNA is dispersed throughout the genome, usually in the form of relatively short individual sequences. It is responsible for the high degree of secondary structure formation in pre-mRNA, when (inverted) repeats in the introns pair to form duplex regions. Highly repetitive DNA often forms discrete clusters (see *Molecular Biology 1.4 Clusters and repeats*). Neither class represents protein.

The genome sequence components represented in mRNA can be determined by using the RNA as a **tracer** in a reassociation experiment. A very small amount of radioactively labeled RNA (or cDNA) is included together with a much larger amount of cellular DNA. The tracer RNA (or cDNA) participates in the reaction as though it were just another member of the sequence component from which it was transcribed. The Cot values at which the labeled RNA hybridizes identify the repetition frequencies of the corresponding genomic sequences.

Figure S 5 shows a typical result for a population of mRNAs. A small proportion of the RNA, generally 10% or less, hybridizes with a $Cot_{½}$ corresponding to moderately repetitive sequences. The major component hybridizes with nonrepetitive DNA.

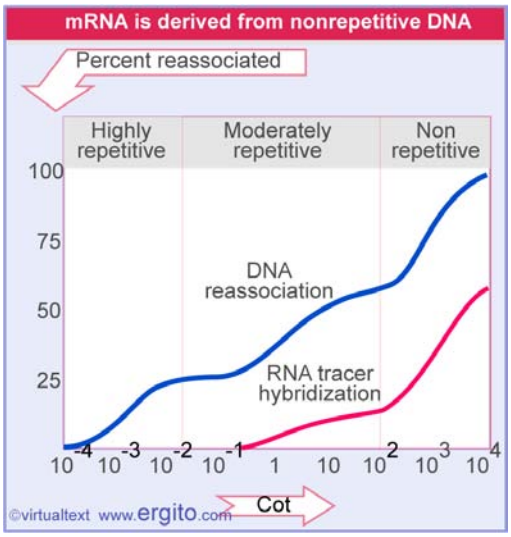

Figure S 5 The hybridization of an mRNA tracer preparation in a reassociation curve shows that most mRNA sequences are derived from nonrepetitive DNA, the remainder from moderately repetitive DNA, and none from highly repetitive DNA.

Reassociation analysis can be also used to measure the complexity of an RNA population. One method is to hybridize nonrepetitive DNA with an excess of RNA; the proportion of the DNA that is bound at saturation identifies the complexity of the RNA population. Another method is to follow the kinetics of hybridization between a excess of an RNA population and a DNA copy prepared from it. This is exactly analogous to reassociation analysis of genomic DNA. The reaction is described in terms of the $Rot_{½}$ (where R_0 is the starting concentrationof RNA).

In looking at the DNA that hybridizes with mRNA, we are basically examining the exons in the genome. The conclusion therefore is that most exons are present at low repetition frequency – depending on the stringency of hybridization, they may be

unique or present in a small number of copies.

SUPPLEMENTS

7.32.2 Mendel's laws and dominance

Key Terms

An **allele** is one of several alternative forms of a gene occupying a given locus on a chromosome.

An individual is said to be **homozygous** when it has identical alleles of a given gene.

An individual is said to be **heterozygous** when it has different alleles of a given gene on each of its homologous chromosomes.

Complete dominance is the state in which the phenotype is the same when the dominant allele is homozygous or heterozygous.

A **dominant** allele determines the phenotype displayed in a heterozygote with another (recessive) allele.

A **recessive** allele is obscured in the phenotype of a heterozygote by the dominant allele, often due to inactivity or absence of the product of the recessive allele.

Incomplete dominance is a state in which the heterozygote has a phenotype in between that of each of the homozygotes.

Two alleles are said to be **codominant** when they are each equally evident in the phenotype of the heterozygote.

Mendel's law of **independent assortment** states that the assortment of one gene does not influence the assortment of another.

A **parental genotype** is one that is identical to the genotype of one of the contributing parents.

Recombinant progeny have a different genotype from that of either parent.

The essential attributes of the gene were defined by Mendel more than a century ago. As he concluded in his analysis of pea genetics in 1865: "The law of combination of different characters, which governs the development of the hybrids, finds therefore its explanation in the principle enunciated, that the hybrids produce egg cells and pollen cells, which in equal numbers, represent all constant forms which result from combinations of the characters brought together in fertilization." *Summarized in his two laws, the gene was recognized as a "particulate factor" that passes unchanged from parent to progeny.*

A gene may exist in alternative forms that determine the expression of some particular characteristic. For example, the color of a flower may be red or white. The forms of the gene are called **alleles**. Mendel's first law describes the *segregation of alleles*: *alleles have no permanent effect on one another when present in the same plant, but segregate unchanged by passing into different gametes.*

When an organism has two identical alleles of a gene, it is said to be **homozygous** (or true-breeding) for the trait conveyed by that gene. If the alleles are different, the organism is **heterozygous** (or hybrid). *The phenotype of a homozygote directly*

reflects the genotype of the (single type of) allele, but the phenotype of a heterozygote depends on the relationship between the types of alleles that are present.

Mendel's first law recognizes that the genotype of a heterozygote includes both alleles, irrespective of the phenotype that is displayed. When a homozygote for one allele is crossed with a homozygote for another allele, all the progeny in the first (F1) generation are heterozygotes with the same phenotype. But when the heterozygotes are crossed with one another to generate a second (F2) generation, the genotypes of the original parents reappear. The critical point is that the alleles must consist of discrete physical entities that contribute independently (or fail to contribute) to the phenotype.

Figure S 6 shows how the results of such crosses differ according to the type of relationship between alleles:

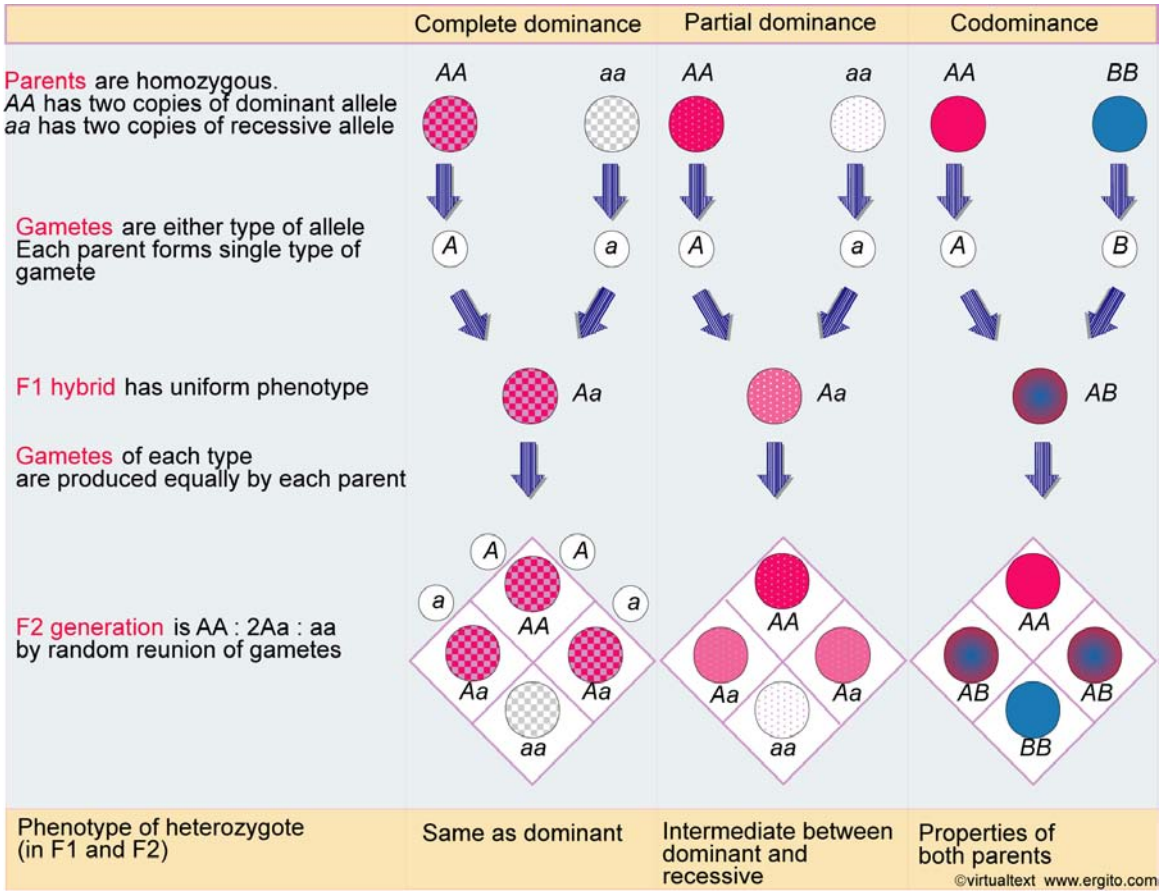

Figure S 6 Mendel#s first law: alleles segregate each generation.

- The case analyzed by Mendel corresponds to **complete dominance**, and is shown in the first column of results. When one allele is **dominant** and the other is **recessive**, the phenotype of a heterozygote is determined by the dominant allele. The recessive allele makes no contribution. The single dominant allele produces the same phenotype that is seen in a wild-type homozygote. The appearance of the heterozygote is indistinguishable from that of the homozygous

dominant parent. The F1 resembles the dominant parent. Complete dominance generates the classic 3:1 ratio of dominant to recessive phenotypes in the F2.

- Some alleles exhibit **incomplete dominance** (or partial dominance), as shown in the middle column. The phenotype of the heterozygote is intermediate between that of the two homozygotes. In the snapdragon, for example, a cross between red and white generates heterozygotes with pink flowers. However, the same rule is observed that the first hybrid (F1) generation is uniform in phenotype; and the same ratios are generated in the second (F2) generation, except that three phenotypes can be distinguished instead of two (*AA* is red, 2 *Aa* are pink, *aa* is white). This type of situation arises through quantitative effects; the single red allele in the heterozygote produces half as much pigment as the two red alleles in a homozygote.

- Alleles are said to be **codominant** when they contribute equally to the phenotype, as shown in the final column. In human blood groups, the *AA* and *BB* combinations are homozygous, and *AB* is a codominant heterozygote in which the *A* and *B* groups are equally expressed. The result is that each genotypic class produces a different phenotype. The F1 has the properties of both parents, and the F2 has the phenotypes A: 2 AB: B.

Mendel's second law summarizes the **independent assortment** of different genes. When a homozygote that is dominant for *two different characters* is crossed with a homozygote that is recessive for both characters, as before the F1 consists of plants whose phenotype is the same as the dominant parent. But in the next (F2) generation, two general classes of progeny are found:

- One class consists of the two **parental genotypes**.

- The other class consists of *new* phenotypes, representing plants with the dominant feature of one parent and the recessive feature of the other. These are called **recombinant** types; and they occur in both possible (*reciprocal*) combinations.

Figure S 7 shows that the ratios of the four phenotypes comprising the F2 can be explained by supposing that gamete formation involves an entirely random association between one of the two alleles for the first character and one of the two alleles for the second character. All four possible types of gamete are formed in equal proportion; and then they associate at random to form the zygotes of the next generation. Once again, the phenotypes conceal a greater variety of genotypes.

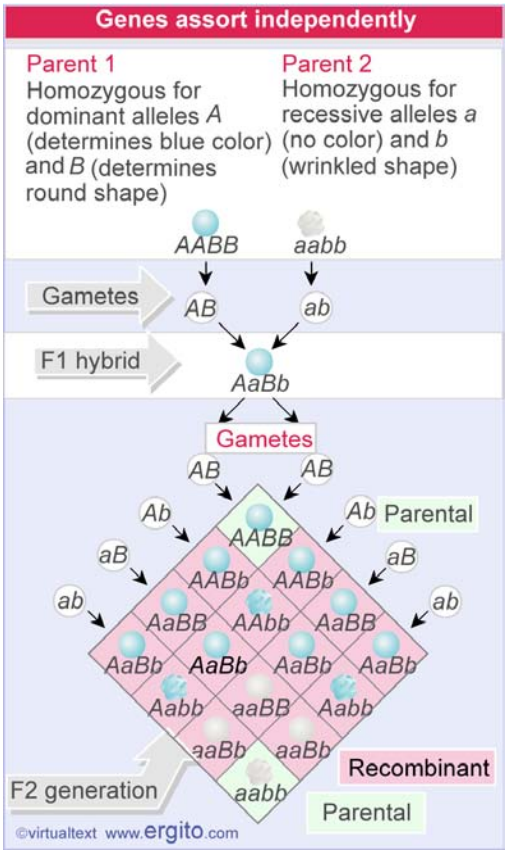

Figure S 7 Mendel#s second law: different genes assort independently in genetic crosses.

The law of independent assortment establishes the principle that the behavior of any pair (or greater number) of genes can be predicted overall by the rules of mathematical combination. *The assortment of one gene does not influence the assortment of another*. Implicit in this concept is the view that assortment is a matter of *statistical probability* and not an exact result. The ratio of progeny types will approximate increasingly closely to the predicted proportions as the number of crosses is increased.

Appreciation of Mendel's discoveries was inhibited by the lack of any known physical basis for the postulated factors (genes). When the chromosomal theory of inheritance was subsequently proposed, however, it was realized that the behavior of chromosomes at meiosis and fertilization corresponds precisely with the properties of Mendel's particulate units of inheritance.

There is an exact parallel between the behavior of chromosomes and Mendel's units of inheritance:

- Genes occur in allelic pairs. One member of each pair is contributed by each parent; so the diploid set of chromosomes results from the contribution of a haploid set by each parent.

- The assortment of nonallelic genes into gametes is independent of (parental) origin; correspondingly, nonhomologous chromosomes undergo independent segregation at meiosis.

The critical proviso is that each gamete obtains a complete haploid set, and this condition is fulfilled whether viewed in terms of Mendel's factors or chromosomes.

SUPPLEMENTS

7.32.3 Linkage and mapping

Key Terms

A **recombinant genotype** is one that consists of a new combination of genes produced by crossing over.

A **parental genotype** is one that is identical to the genotype of one of the contributing parents.

Linkage describes the tendency of genes to be inherited together as a result of their location on the same chromosome; measured by percent recombination between loci.

A **linkage map** is a map showing the linear order of genes on a chromosome and the relative distances between them in recombinational units.

An **allele** is one of several alternative forms of a gene occupying a given locus on a chromosome.

A **backcross** describes a genetic cross in which a hybrid strain is crossed to one of its two parental strains.

Crossing-over describes the reciprocal exchange of material between chromosomes that occurs during prophase I of meiosis and is responsible for genetic recombination.

A **chiasma** (*pl.* chiasmata) is a site at which two homologous chromosomes appear to have exchanged material during meiosis.

Breakage and reunion describes the mode of genetic recombination, in which two DNA duplex molecules are broken at corresponding points and then rejoined crosswise (involving formation of a length of heteroduplex DNA around the site of joining).

Map distance is measured as cM (centimorgans) = percent recombination (sometimes subject to adjustments).

A **map unit** is the distance between two genes that recombine with a frequency of 1%.

A **locus** is the position on a chromosome at which the gene for a particular trait resides; a locus may be occupied by any one of the alleles for the gene.

A **linkage group** includes all loci that can be connected (directly or indirectly) by linkage relationships; equivalent to a chromosome.

A **marker** is an identifiable and inheritable difference that can be mapped to a location on a chromosome. A genetic marker is an allele that is identified with its genetic trait. A molecular marker is a DNA sequence difference that can be identified by molecular methods.

Phage T4 is a virus that infects E. coli causing lysis of the bacterium.

Rapid lysis (*r*) mutants display a change in the pattern of lysis of *E. coli* at the end of an infection by a T-even phage.

Mendel's laws predict that genes carried on different chromosomes will segregate independently (for additional description see *Molecular Biology Supplement 32.2 Mendel's laws and dominance*). However, genes that are on the same chromosome show *linked inheritance*. The basic observation is that genes on different chromosomes *recombine* at random from one generation to the next, whereas genes that are linked show a *reduction in recombination*, that is, they tend to stay together.

The results of a genetic cross are analyzed by determining determining the proportions of **recombinant genotypes** (where an allele of one parent is found with an allele of the other parent) and **parental genotypes** (which have the same combination of alleles as either parent). Genes on different chromosomes segregate independently, as predicted by Mendel, to give 50% parental and 50% recombinant progeny. Genes on the same chromosome behave differently, because they are present on the same (very long) molecule of DNA. Instead of generating the proportions depicted by independent assortment, the proportion of parental genotypes is greater than expected, *because there is a reduction in the formation of recombinant genotypes.* The propensity of some characters to remain associated instead of assorting independently is called **linkage**.

Linkage is measured by the per cent recombination between two loci (in formal terms a *map distance of 1 centimorgan = 1% recombination*). When pairwise combinations of loci on the same chromosome are tested in genetic crosses, loci close to one another are linked, as defined by a map distance <50 cM. Loci that are farther apart recombine at the limit of 50%. But a **linkage map** corresponding to the chromosome can be generated by extending a series of genetic crosses in which in effect two loci >50 cM apart are connected because they show linkage to a locus between them. This genetic map corresponds to the physical existence of the chromosome.

A crucial concept in the construction of a genetic map is that the distance between genes does not depend on the particular **alleles** that are used, but only on the genetic *loci*. *The **locus** defines the position occupied on the chromosome by the gene representing a particular trait. The various alternative forms of a gene – that is, the alleles used in mapping – all reside at the same location on its particular chromosome.* So genetic mapping is concerned with identifying the positions of genetic loci, which are fixed and lie in a linear order. In a mapping experiment, the same result is obtained irrespective of the particular combination of alleles.

Figure S 8 shows how a **backcross** to a recessive homozygote is used to measure linkage. The alleles of the recessive parent make no contribution to the phenotype of the progeny. As a result, the backcross essentially makes it possible to examine directly the genotype of the organism being investigated. In each cross the progeny show an increase in the proportion of parental types (70%) and a decrease in the proportion of recombinant types (30%), compared with the 50% of each type that is expected from independent assortment. The linkage between A and B is measured as 30%.

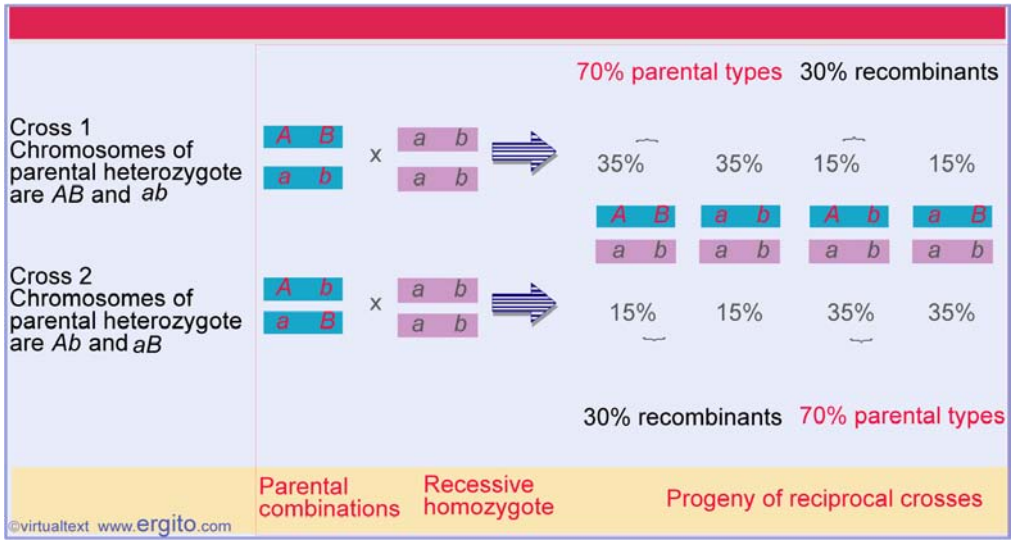

Figure S 8 Linkage can be measured by a backcross with a double recessive homozygote.

The smaller the proportion of recombinants in the progeny, the tighter the linkage. A crucial characteristic is that the *same proportion of recombinants is obtained irrespective of the arrangement of parental alleles (AB/ab or Ab/aB).* And in each case, both of the (reciprocal) recombinant types are present in the same proportions..

Morgan proposed that genetic linkage is the "simple mechanical result of the location of the (genes) in the chromosomes." He suggested that the production of recombinant classes can be equated with the process of **crossing-over** that is visible during meiosis. Early in meiosis, at the stage when all four copies of each chromosome are organized in a bivalent, pairwise exchanges of material occur between the closely associated (synapsed) chromatids.

The visible result of a crossing-over event is called a **chiasma**, and is illustrated diagrammatically in **Figure S 9**. A chiasma represents a site at which two of the chromatids in a bivalent have been broken at corresponding points. The broken ends have been rejoined crosswise, generating new chromatids. Each new chromatid consists of material derived from one chromatid on one side of the junction point, with material from the other chromatid on the opposite side. The two recombinant chromatids have reciprocal structures. The event is described as a **breakage and reunion**.

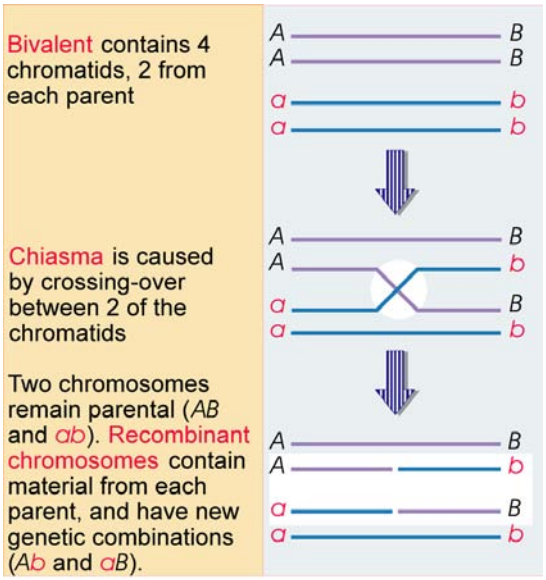

Figure S 9 Chiasma formation is responsible for generating recombinants.

If the likelihood that a chiasma will form between two points on a chromosome depends on their distance apart, genes located near each other will tend to remain together. As the distance decreases, the probability of crossing-over between them will decrease. If crossing-over is responsible for recombination, the closer genes lie to one another, the more tightly they will be linked. Reversing the argument, genetic linkage can be taken to be a measure of physical distance.

The extent of recombination between two genes on the same chromosome can be used as a **map distance** to measure their relative locations. The formula to measure genetic distance is:

$$\text{Map distance} = \frac{\text{Number of recombinants} \times 100}{\text{Total number of progeny}}$$

Map units are defined as 1 unit (or centiMorgan, abbreviated cM) equals 1% crossover. For short distances (<10%), map units are given directly by the percent recombinants. However, when two crossovers occur near one another, they restore the parental arrangement of the loci on either side. This reduces the number of recombinants, so that recombination frequency underestimates the map distance.

A critical feature is observed when multiple characters are followed together. For genes carried on the same chromosome, *individual map distances are (approximately) additive.* If two genes *A* and *B* are 10 units apart, and gene *C* lies a further 5 units beyond *B*, the direct measure of distance between *A* and *C* will be close to 15 units. The genes can therefore be placed in a *linear order*.

A crucial concept in the construction of a genetic map is that the distance between genes does not depend on the particular *alleles* that are used, but only on the genetic

loci>. *The **locus** defines the position occupied on the chromosome by the gene representing a particular trait. The various alternative forms of a gene* – that is, the alleles used in mapping – *all reside at the same location on its particular chromosome.*

So genetic mapping is concerned with identifying the positions of genetic loci, which are fixed and lie in a linear order. In a mapping experiment, the same result is obtained irrespective of the particular combination of alleles (see **Figure S 8**, where in either combination, there are 70% parental and 30% recombinant types).

Linkage is not displayed between all pairs of genes located on a single chromosome. The maximum recombination between two loci is the 50% corresponding to the independent segregation predicted by Mendel's second law. (Although there is a high probability that recombination will occur between two genes lying far apart on a chromosome, each individual recombination event involves only two of the four associated chromatids, so there is a limit of 50% recombination between the genes.)

In spite of their presence on the same chromosome, genes that are far apart therefore assort independently. But although they show no direct linkage, each can be linked to genes that lie between them. This allows the genetic map to be extended beyond the limit of 50% recombination that can be measured directly between any pair of genes. A genetic map is usually based on measurements involving genes that are fairly close together (and is subject to corrections from the simple percent recombination).

A **linkage group** includes all those genes that can be connected either directly or indirectly by linkage relationships. Genes lying close together show direct linkage; those >50 cM apart assort independently. As linkage relationships are extended, the genes of any organism fall into a discrete number of linkage groups. Each gene identified in the organism can be placed into one of the linkage groups. Genes in one linkage group always show independent assortment with regard to genes located in other linkage groups.

The number of linkage groups is the same as the (haploid) number of chromosomes. The relative lengths of the linkage groups are similar to the relative sizes of the chromosomes.

Mendel's concept of the gene as a discrete particulate factor can therefore be extended into the concept that *the chromosome constitutes a linkage group, divided into many genes, whose physical arrangement underlies their genetic behavior.* We sometimes use the term genetic marker to describe a gene of interest, for example, one being used in a mapping experiment or identifying a particular region. Thus a chromosome may be said to carry a particular set of **markers**, that is, alleles.

On the genetic maps of higher organisms established during the first half of this century, the genes are arranged like beads on a string. They occur in a fixed order, and genetic recombination involves transfer of corresponding portions of the string between homologous chromosomes. The gene is to all intents and purposes a mysterious object (the bead), whose relationship to its surroundings (the string) is unclear.

The resolution of the recombination map of a higher eukaryote is restricted by the

small number of progeny that can be obtained from each mating. Recombination occurs so infrequently between nearby points that it is rarely observed between different mutations in the same gene. This forces the questions: does recombination occur within a gene; and can its frequency at these close quarters be used to arrange sites of mutation in a linear order?

To answer these questions by conventional genetic means requires a microbial system in which a very large number of progeny can be obtained from each genetic cross. A suitable system is provided by **phage T4**, a virus that infects the bacterium *E. coli*. Infection of a single bacterium leads to the production of ~100 progeny phages in less than 30 minutes.

The constitution of an individual locus was investigated by Benzer in a series of intensive studies of the *rII* genes of the phage, which are responsible for a change in the pattern of bacterial killing known as **rapid lysis**. When two different *rII* mutant phages are used to infect a bacterium simultaneously, the conditions can be arranged so that progeny phages will be produced *only* if recombination has occurred between the two mutations to generate a wild-type recombinant. The frequency of recombination depends on the distance between sites, just as in the eukaryotic chromosome.

The selective power of this technique in distinguishing recombinants of the desired type allows even the rarest recombination events to be quantitated, so that the map distance between *any* pair of mutations can be measured. About 2400 mutations fall into 304 different mutant sites. (When two mutations fail to recombine, they are assumed to represent independent and spontaneous occurrences at the *same* genetic site.)

The mutations can be arranged into a linear order, showing that *the gene itself has the same linear construction as the array of genes on a chromosome.* So the genetic map is linear within as well as between loci: it consists of an unbroken sequence within which the genes reside. This conclusion has of course now been extended in molecular terms to all known genetic systems.

SUPPLEMENTS

7.32.4 Protein folding

Key Terms

A **cofactor** is a small inorganic component (often a metal ion) that is required for the proper structure or function of an enzyme.

Chaperones are a class of proteins which bind to incompletely folded or assembled proteins in order to assist their folding or prevent them from aggregating.

A **domain** of a protein is a discrete continuous part of the amino acid sequence that can be equated with a particular function.

We can consider two principles that might control the folding of a protein into the correct higher-order structure.

- *Folding is an intrinsic feature of the primary sequence.* In this case, the final structure must always be the most stable thermodynamically and can be generated at any time after synthesis of the polypeptide chain is complete.
- *The correct structure can be generated only during the synthesis of the polypeptide.* Then it becomes possible that an intrinsically less stable structure could prevail because the protein becomes "trapped" in it during synthesis.

The relationship between higher-order structures and the primary structure may be revealed when a protein is *denatured* by heating or by chemical treatments that disrupt protein conformation. Most denaturing events involve the breakage of hydrogen and other noncovalent bonds. An exception is the disruption of S–S bridges that results from treatment with reducing agents. However, all of these changes affect the *conformation*; the primary sequence of amino acids in the polypeptide chain remains unaltered.

In some cases, the higher-order structure follows ineluctably from the primary sequence. The enzyme ribonuclease is the classic example (1115). After the protein has been denatured, its active conformation can be regained by reversing the denaturing procedure. *All the information necessary to form the secondary structure resides in the primary sequence.* Thus the production of active ribonuclease is an inevitable event whenever the intact primary chain is placed in the appropriate conditions.

In other cases, proteins can be irreversibly denatured. Thus under certain (nonphysiological) conditions, a protein may have alternative stable conformations. In some cases the correct conformation probably can be attained *only* during synthesis of the protein. The conformation could depend on specific interactions between regions of the protein that can occur only in the absence of other regions (that is, those that have not yet been synthesized). This is probably the more common situation.

In some instances, a **cofactor** that is part of the active protein (such as the iron-binding heme group of the cytochromes) must be present in order for the polypeptide chain to take up its proper conformation. In the case of multimeric proteins, it may be necessary for one subunit to be present in order for another to acquire the proper conformation.

Protein folding is usually rapid *in vivo*, occurring within seconds or less. It begins even before a protein has been completely synthesized. Probably it involves a *sequential folding* mechanism, in which the reaction passes through discrete (although highly transient) intermediates. The process is initiated by the collapse of hydrophobic side chains into the "core" of the protein; this occurs within milliseconds. Units of secondary structure, largely α-helices and β-sheets, form on the same time scale. The transition from this structure to the final tertiary structure is slower. The process appears to be cooperative, so that formation of one region of secondary structure enhances formation of the next region, and so on (for review see 2389)

The acquisition of structure when a protein is synthesized is not a spontaneous process, but may require assistance. More precisely, we should say that spontaneous folding is a slow reaction, which under normal cellular conditions is a rate-limiting step. The rate is significantly increased by several types of additional functions. These are summarized in **Figure S 10**. They fall into two groups: enzymes that catalyze specific isomerization steps; and factors that act stoichiometrically to influence folding directly.

Protein Activity	Families	Function	Effect
Catalytic			
Protein disulfide isomerase	Thioredoxin	Formation of S-S bonds	Increases rate of folding
Peptidyl prolyl isomerase	Cyclophilin PPI FKBP PPI	*cis/trans* bond conversion "	
Stoichiometric			
Chaperones	Hsp-70 (DnaK) Hsp-60 (GroEL) Hsp-90	Binds improperly folded protein	Inhibits wrong folding pathway

Figure S 10 Both catalytic and stoichiometric functions are required to assist protein folding

The formation of disulfide bonds is shown in **Figure S 11**. The animation shows that formation of a disulfide bond may have a major effect on the conformation of the protein. It is influenced by both environment and specific accessory proteins. Disulfide bonds are rare in cytoplasmic proteins, but common in exported proteins. This is related to a difference in the thiol/disulfide redox state between internal and external conditions. It may help to prevent bond formation in a bacterium, but helps to drive it in the periplasm (the layer surrounding the bacterial cell).

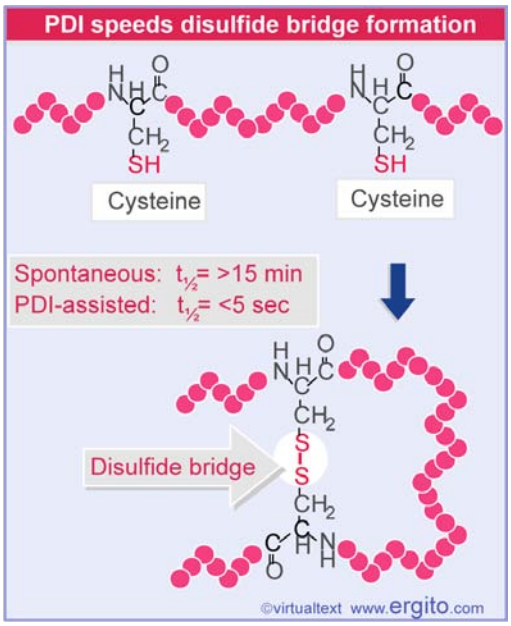

Figure S 11 Formation of a disulfide bridge between the sulfhydryl groups of two cysteines may connect different parts of a polypeptide chain.

Disulfide bond formation can occur spontaneously *in vitro*, but the rate is slow. It has a $t_{½}$ >15 min, compared with the ability to form disulfide bonds correctly within a few seconds *in vivo*. The process is catalyzed *in vivo* by an enzyme, protein disulfide isomerase (PDI). This is a curious protein, which participates in a variety of functions concerned with protein-modification, in addition to its sponsorship of disulfide bridge formation. It is not entirely clear whether it simply helps the initial formation of disulfide bonds or whether it also catalyzes rearrangement of disulfide bonds that have formed incorrectly (for review see 3443)

Proline has a major effect upon protein structure because of the restrictions imposed by its ring structure. Proline introduces a bend in a polypeptide chain, because the nitrogen atom is restrained by the ring structure. The existence (and interconversion) of two stereochemical forms of the peptidyl-proline link is an important feature of protein structure. The direction of the bend is determined by whether the proline is in the *cis* or *trans* configuration, as shown in **Figure S 12**.

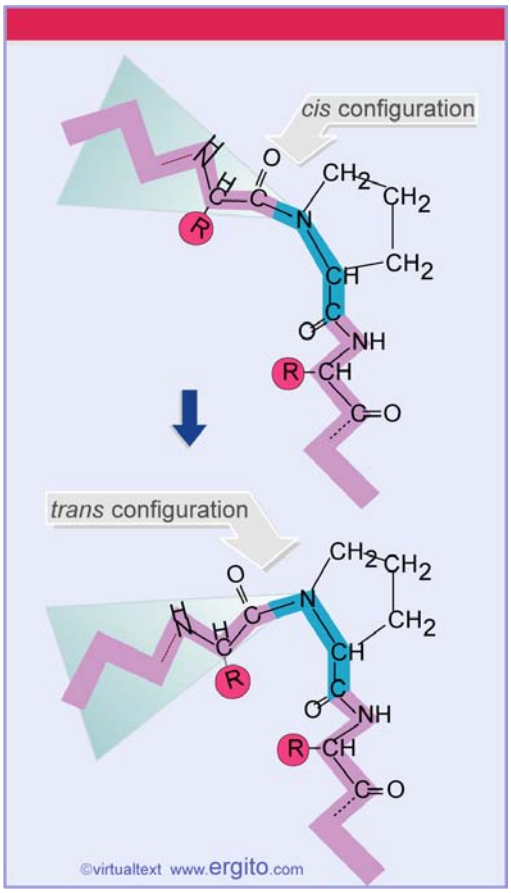

Figure S 12 The configuration of proline has an important effect on protein conformation.

Proteins containing proline fold slowly because the peptidyl-proline link does not necessarily form in correct stereochemical conformation. The enzyme peptidyl-prolyl isomerase (PPI) catalyzes the *cis-trans* conversion, and by this means significantly accelerates the folding reaction. Enzymes with PPI activity fall into two major groups, named for their abilities to bind certain drugs: cyclophilin PPI binds the drug cyclosporin A, and FKBP PPI binds the drug FK506. Members of the cyclophilin class are better characterized, and they vary in their specificity of action from those that appear to be generic (able to act on any protein) to those that appear to work only with specific proteins. This makes the point that, although control of proline isomerization is a general feature of many proteins, it can also be used to control specifically the maturation of an individual protein.

Proteins that act stoichiometrically on the folding of other proteins are called molecular **chaperones**. A chaperone forms a complex with a protein during folding, *but is required only during assembly, and is not part of the mature structure.* The major role of a chaperone is to prevent the formation of incorrectly folded structures, in which the substrate protein might otherwise become trapped during folding (see **Figure 8.8** in *Molecular Biology 2.8.4 Chaperones may be required for protein folding*).

Protein folding is an intricate process. The primary sequence of a protein is a crucial

determinant of its higher-order structures. Sometimes it is the sole determinant, but in most cases additional interactions are involved in acquiring the final conformation. In each case, however, if the primary sequence is synthesized within the appropriate environment, it will acquire the proper higher-order structures.

Although higher-order structure follows from primary sequence, the same general tertiary structure can be determined by different primary sequences. For example, the globin (red blood cell) proteins of different species vary substantially in sequence, but have the same general tertiary structure.

An important concept is that a protein may consist of **domains**. A domain is a (relatively) independent region of the protein. In some cases its conformation can be acquired independently by the relevant fragment of the polypeptide chain. Some globular proteins consist of discrete domains connected by "clefts." Sometimes a substrate binds to the cleft between domains.

A domain may represent a functional unit that is identified with a particular activity of the protein, for example, its ability to perform a certain catalytic activity, to bind a certain ligand, or to interact specifically with other types of domains. The lengths of recognized domains vary from 30–300 amino acid residues. Certain types of domains may be found in proteins with particular locations; for example, on the exterior of the cell. A domain may represent an evolutionary unit. It may have arisen as a functional polypeptide or region of a polypeptide and later have associated with other domains to generate a new protein with additional abilities. The occurrence of closely related domains in different proteins is common; **Figure S 13** compares the use of domains in two blood cell proteins.

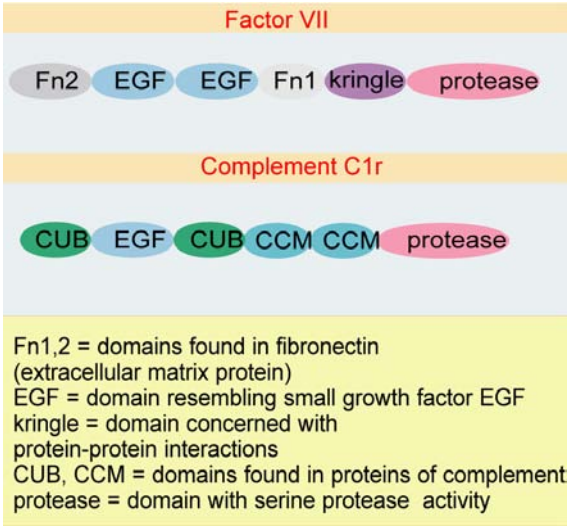

Figure S 13 Overlapping arrangements of discrete domains are found in proteins.

Last updated on 3-21-2002

Reviews

2389. Fersht, A. R. and Daggett, V. (2002). *Protein folding and unfolding at atomic resolution.* Cell 108, 573-582.

3443. Sevier, C. S. and Kaiser, C. A. (2002). *Formation and transfer of disulphide bonds in living cells.* Nat. Rev. Mol. Cell Biol. 3, 836-847.

References

1115. Anfinsen, C. B. (1973). *Principles that govern the folding of protein chains.* Science 181, 223-230.

SUPPLEMENTS

7.32.5 Membranes and membrane proteins

Key Terms

Amphipathic structures have two surfaces, one hydrophilic and one hydrophobic. Lipids are amphipathic; and some protein regions may form amphipathic helices, with one charged face and one neutral face.

A **saturated** fatty acid only has single carbon-carbon bonds in its backbone.

An **unsaturated** fatty acid has some double carbon-carbon bonds in its backbone.

A **phospholipid** is a lipid that has a positively charged head that is linked by a phosphate group to the fatty acid tails.

A **glycolipid** has a head consisting of an oligosaccharide, linked to a fatty acid tail.

A **sterol** is a compound containing a planar steroid ring.

A **lipid bilayer** is a structure formed by phospholipids in an aqueous solution. The structure consists of two sheets of phospholipids, in which the hydrophilic phosphate groups face the aqueous solution and the hydrophobic tails face each other.

Fluidity is a property of membranes; it indicates the ability of lipids to move laterally within their particular monolayer.

A **transmembrane protein (Integral membrane protein)** extends across a lipid bilayer. A hydrophobic region (typically consisting of a stretch of 20-25 hydrophobic and/or uncharged aminoa acids) or regions of the protein resides in the membrane. Hydrophilic regions are exposed on one or both sides of the membrane.

The **transmembrane region (transmembrane domain)** is the part of a protein that spans the membrane bilayer. It is hydrophobic and in many cases contains approximately 20 amino acids that form an α -helix. It is also called the transmembrane domain.

A **hydropathy plot** is a measure of the hydrophobicity of a protein region and therefore of the likelihood that it will reside in a membrane.

The side of the plasma membrane, or of the membrane of an organelle, which faces the cytoplasm is its **cytoplasmic face**.

The **extracellular matrix (ECM)** is a relatively rigid layer of insoluble glycoproteins that fill the spaces between cells in multicellular organisms. These glycoproteins connect to plasma membrane proteins.

The characteristic properties of membranes result from their high contents of lipids. A crucial feature of lipids creates the membranous environment: they are **amphipathic**. One end of the molecule consists of a polar "head," while the other end consists of a hydrophobic "tail."

The major bulk of a lipid is provided by its hydrophobic tails, which differ in overall

length and in the nature of the carbon-carbon bonds. One type of fatty acid tail is **saturated**: all the carbon-carbon links are single bonds. The other type of tail is **unsaturated**: one or more carbon-carbon links consist of double bonds. Because rotation is restricted around the double bond, the unsaturated tail has a bend, while the saturated tail can extend freely. Fatty acid tails are usually ~20 residues long. Distinguished by their polar heads, membranes contain the three principal types of lipids illustrated in **Figure S 14**:

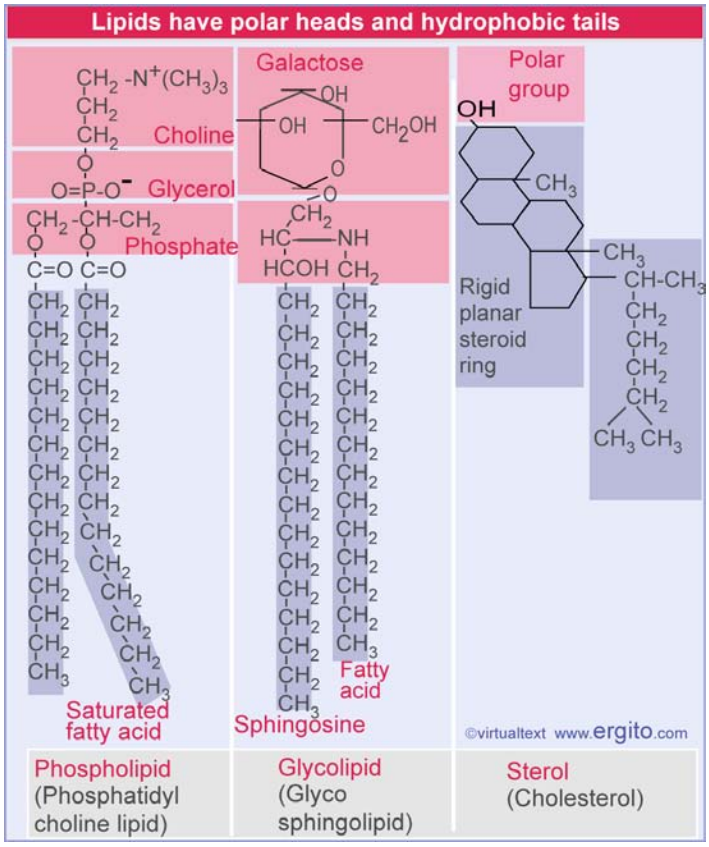

Figure S 14 A lipid has a polar head and a hydrophobic tail.

- In **phospholipids** the head has a positively charged group linked via a negatively charged phosphate group to the rest of the molecule. The example of **Figure S 14** has a head consisting of choline-phosphate-glycerol, attached to two hydrophobic tails. Lipids based on glycerol have one saturated and one unsaturated fatty acid tail.

- **Glycolipids** are characterized by the presence of oligosaccharide. The chain of sugars typically consists of 1-15 residues. In animal cells, the connection between the saccharide head and the fatty acid tail is sphingosine (a long amino alcohol). Lipids based on sphingosine have a fatty acid chain in addition to the long hydrocarbon chain of sphingosine itself. In plants and bacteria, glycerol connects the head and tail.

- **Sterols** contain a steroid ring. They lend rigidity to a membrane because the steroid ring is planar. Cholesterol, a prominent component of animal cell

membranes, has a polar hydroxyl group at the terminus.

In an aqueous environment, a lipid is happy to have its polar head exposed, but tries to bury its hydrophobic tail away from the water. **Figure S 15** illustrates how this is accomplished in the cell. Two lipid monolayers are juxtaposed to form a **lipid bilayer**, a sheet in which the polar heads of the lipids face out toward the aqueous environment on either side, while the hydrophobic tails face in to create a hydrophobic environment.

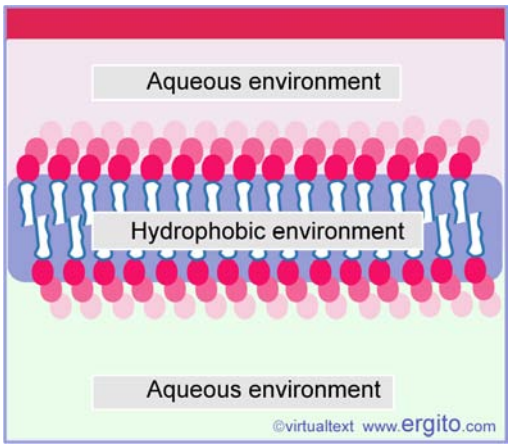

Figure S 15 A lipid bilayer forms in an aqueous environment when the the polar heads are immersed in the water and the hydrophobic tails of the lipids segregate away from water.

Although a membrane consists of a specific type of structure, the lipid bilayer, there is variety in the constitution of different membranes. Overall lipid compositions of membranes vary considerably, with regard to both the ratio of protein to lipid and the types of lipids. These differences mean that different membranes have different biophysical properties.

One of the important properties of a membrane is the ability of the constituent lipids to move within it. Lipid molecules rarely move from one monolayer to the other in a bilayer, but frequently move laterally to exchange places with their neighbors within the monolayer. The property of movement is called **fluidity**; and a membrane is often regarded as a "two dimensional fluid" (1023).

The more readily the tails of adjacent lipids can pack together, the more crystalline the membrane structure can become, and the less fluid. The major determinants of membrane fluidity are therefore the types and lengths of the lipid tails. The proportion of saturated versus unsaturated residues in the tails has a major effect on fluidity; unsaturated chains are more difficult to pack, and therefore give a more fluid structure.

A protein that resides in a membrane is called a **transmembrane protein**. The structure of such a protein is illustrated in **Figure S 16**. A typical transmembrane protein resides in the membrane by means of a specific region called the **transmembrane domain**. This consists of a stretch of ~21 amino acids with

sufficient hydrophobicity to be comfortable in the environment of the lipid bilayer. It forms an α-helix that just spans the membrane. The transmembrane domain is flanked by regions that protrude into the interior of the cell or out into the surrounding environment. These regions are generally hydrophilic, like any protein that resides in the cytosol. A protein may have several transmembrane domains, in which case the parts of the polypeptide chain connecting them can be viewed as "looping out" into the cytoplasm or the exterior.

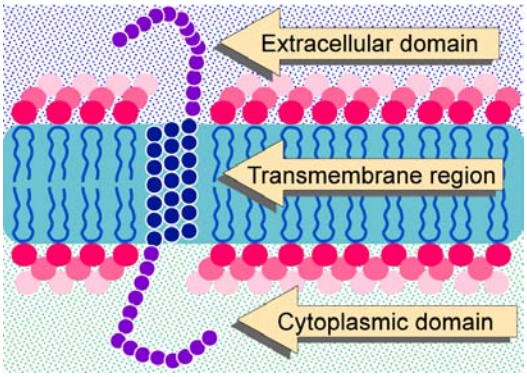

Figure S 16 A transmembrane protein crosses the lipid bilayer. The hydrophobic transmembrane region spans the bilayer, and hydrophilic regions are exposed on either side.

Within the lipid bilayer, water is effectively excluded. As a result, the regions of the proteins located within the bilayer are not subjected to the aqueous environment of the cytosol. The lipid "solvent" does not form hydrogen bonds with the protein, and therefore solvates neither the groups of the peptide backbone nor polar side chains. Hydrogen bonding occurs solely between groups within the protein itself, and functions to form α-helices and (to a lesser degree) ß-sheets. This allows the protein to acquire a different conformation from what would be reached in an aqueous environment. Indeed, it is possible that membrane proteins can attain their natural conformation *only* in the hydrophobic environment.

The hydrophobicity of a sequence of amino acids can be used to predict (although not perfectly) whether it is likely to reside in a membrane. A **hydropathy plot** shows the sequence of a protein in terms of the hydrophobicity of overlapping segments. There are various means of measuring hydrophobicity, but whichever scale is used, it is conventional to assess hydrophobicity as a positive score and hydrophilicity as a negative score. Most of the scales utilize the energy required in kcal/mol to transfer from a hydrophobic to a hydrophilic phase.

In the example of **Figure S 17**, the hydrophobicity is calculated for each position in the protein by summing the scores of the individual amino acids in the next 21 positions. A region with a positive score is therefore a candidate to provide a transmembrane domain that resides in a membrane.

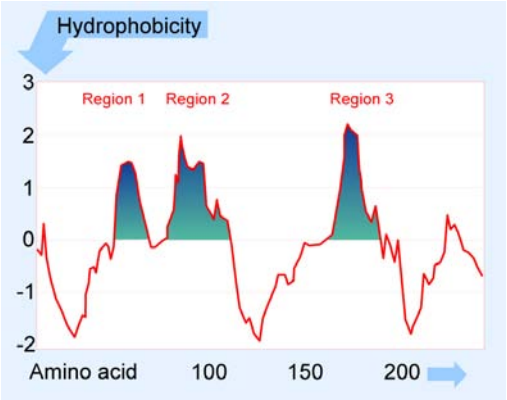

Figure S 17 A hydropathy plot identifies potential membrane-spanning regions as the most hydrophobic sequences of a protein.

Proteins may also be associated with a membrane by means of covalent linkage to a fatty acid that is incorporated into the lipid bilayer. **Figure S 18** depicts four forms of such association. In each case, a fatty acid or lipid is attached to an amino acid near to or at one terminus of the protein, with the result that the entire polypeptide chain resides on one side of the membrane, but is attached to it.

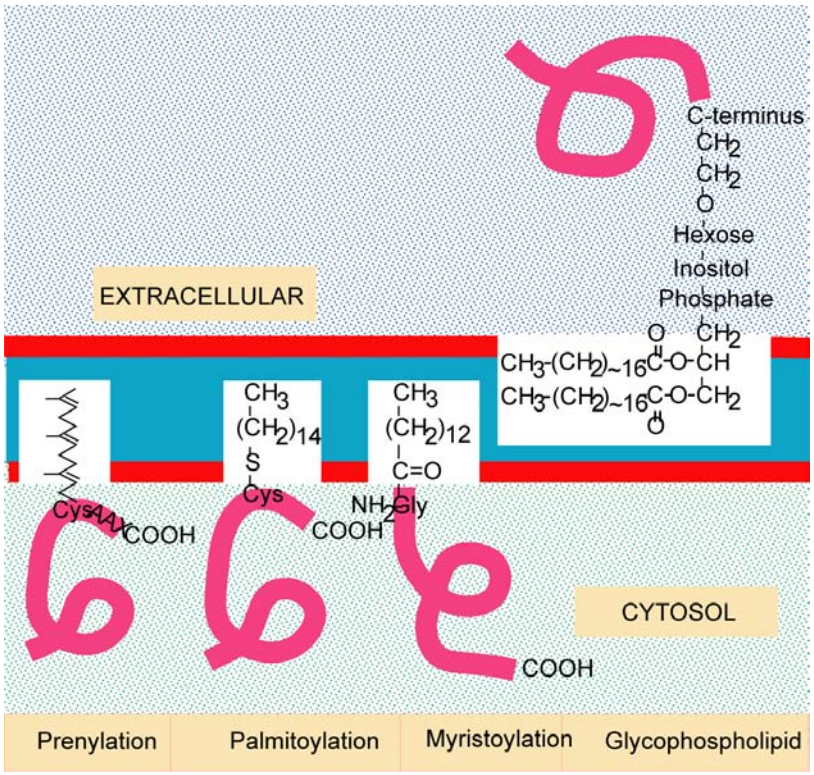

Figure S 18 Proteins may be associated with one face of a membrane by acyl linkages to fatty acids.

Prenylation is used to attach proteins to both the plasma membrane and internal membranes. Two types of prenyl groups have been identified: farnesyl is a 15 carbon

isoprenoid (shown in the figure), and geranylgeranyl is a 20 carbon chain. They are added to cysteine residues by a thioester linkage; the cysteine is always located at the fourth position from the C terminus, as part of the sequence CAAX, where A represents aliphatic amino acids and X is methionine or serine for farnesylation, and leucine for geranylgeranylation. The usefulness of the prenyl groups for assisting attachment to the membrane is obvious, but we do not yet know how specificity is conferred with regard to the choice of membrane.

Two fatty acids are used to anchor proteins on the cytoplasmic side of the plasma membrane. Palmitic acid, a 16 carbon-chain saturated fatty acid, is linked through a sulfide bond to a cysteine residue located close to the terminus (usually the C-terminus, but sometimes the N-terminus). Myristic acid, a 14 carbon-chain saturated fatty acid, is linked to the amino group of N-terminal glycine. Myristoylated proteins are often, but not always, associated with a membrane.

The more complex structure of a glycosyl-phosphatidyl-inositol (GPI) anchor is linked to the carboxyl group of the C-terminal amino acid of protein exposed on the extracellular side of the membrane. Addition of the GPI anchor actually involves cleavage of the original polypeptide chain near the C-terminus, generating a new C-terminus that is linked to the anchor. Enzymes exist that can cleave the GPI anchor from the protein, releasing the protein into the extracellular medium.

The membranes bounding different cellular compartments are different not only in their overall composition, but also in the particular proteins that reside within them. The proteins in each type of membrane serve, for example, to control transport into or out of the particular compartment, and are therefore designed to recognize the particular molecules or macromolecules that travel this route.

Each membrane has two "faces," as indicated in **Figure S 19**. In all membranes, the **cytoplasmic face** is defined as the surface that contacts the general cytosol. The noncytoplasmic face is given various names, depending on the membrane. On a plasma membrane, it provides the outside surface of the cell. In a membrane within the cell, it provides a limit for an interior compartment, comprising the surface that separates the lumen of the compartment from the cytosol.

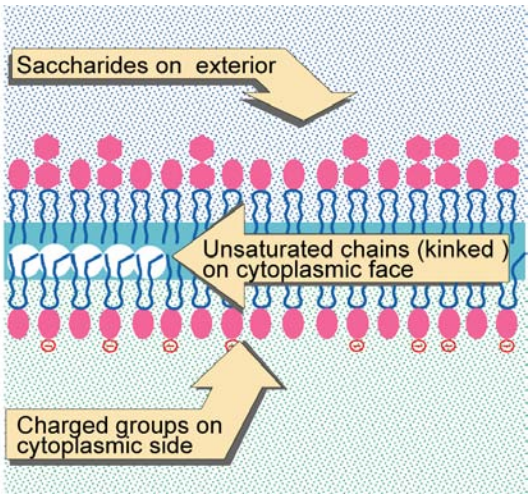

Figure S 19 The asymmetry of the membrane bilayer distinguishes the cytoplasmic face from the exterior face.

A major feature of the lipid bilayer is an asymmetry created by biochemical differences between the two faces of the membrane. Three components of the plasma membrane are unevenly distributed:

- Different lipids are concentrated in the cytoplasmic and extracellular monolayers. This affects both the polar heads and the hydrophobic tails. Lipids on the cytoplasmic face are more highly charged and tend to be unsaturated. How is the difference between the monolayers established? Lipids are synthesized within the cell, and initially inserted into the cytoplasmic surface of the membrane. A specific protein, a "flippase," may be responsible for transporting a lipid from one monolayer to the other. The existence of specific flippases for different lipids could be responsible for creating some of the asymmetries in lipid distribution between the bilayers.

- Proteins are oriented so that different sequences (or even entire proteins) are present on each face. The location of a protein is determined by its sequence, which contains signals that cause it to be inserted in the membrane in a particular orientation. This is discussed in detail in *Molecular Biology 2.8.15 Anchor sequences determine protein orientation*.

- Carbohydrate groups (on glycolipids or glycoproteins) are found exclusively on the extracellular face. The consequence of this organization for the plasma membrane is that the exterior of the cell has a surface rich in oligosaccharides.

The plasma membrane circumscribes a cell. It marks the boundary between the cellular milieu inside and the environment outside. In the case of a unicellular organism, the surroundings constitute the environment in which the organism lives. In a multicellular organism, the environment for any one cell is created by other cells. In some cases, the plasma membrane is extended by the presence of additional glycoproteins, connected to those actually included in the membrane. This type of arrangement may form a cell coat. In some cases, the cell coat is extended into an **extracellular matrix**, rich in glycoproteins, and providing a thicker layer at the cell surface.

The role of the plasma membrane extends beyond providing a mere barrier to the outside. It controls ingress and egress for molecules both small and large. Within the membrane reside specific transport systems that pump ions in or out, that allow proteins to be secreted from the cell into the environment (see *Molecular Biology 6.27 Protein trafficking*), and that recognize molecules outside and as a result transmit messages to the interior (see *Molecular Biology 6.28 Signal transduction*).

Plasma membranes of animal cells contain relatively large amounts of cholesterol, which increases mechanical stability because of the steroid rings near its polar head. (Plant cells lack cholesterol, but have other sterols instead.) Plasma membranes also are the only membranes to contain significant amounts of glycolipids.

A plasma membrane contains about equal masses of lipid and protein. Internal membranes, such as those surrounding mitochondria, have a greater proportion of protein. The mass of an individual protein molecule is much larger than any lipid, so there are 10–100× more lipid molecules than protein molecules. We might view the basic structure of a membrane as consisting of a lipid bilayer that provides a residence for relatively large protein molecules.

Protein components of the membrane can move laterally within the lipid bilayer, although they diffuse much more slowly than the lipid components. As the result of a stimulus, proteins can be "internalized," when they are removed to the interior of the cell. Other proteins are secreted from the interior of the cell to the exterior by passing through the membrane. The lipid bilayer itself, and the proteins associated with it, therefore comprises a dynamic structure.

References

1023. Singer, S. J. and Nicolson, G. L. (1972). *The fluid mosaic model of the structure of cell membranes.* Science 175, 720-731.

SUPPLEMENTS

7.32.6 ER and Golgi

Key Terms

The **endoplasmic reticulum (ER)** is an organelle involved in the synthesis of lipids, membrane proteins, and secretory proteins. It is a single compartment that extends from the outer layer of the nuclear envelope into the cytoplasm. It has subdomains, such as the rough ER and smooth ER.

The **lumen** describes the interior of a compartment bounded by a membrane, usually the endoplasmic reticulum or the Golgi apparatus.

Rough endoplasmic reticulum (rough ER) refers to the region of the endoplasmic reticulum to which ribosomes are bound. It is the site of synthesis of membrane proteins and secretory proteins.

Smooth ER consists of a regions of endoplasmic reticulum devoid of ribosomes.

The **ribosome** is a large assembly of RNA and proteins that synthesizes proteins under direction from an mRNA template. Bacterial ribosomes sediment at 70S, eukaryotic ribosomes at 80S. A ribosome can be dissociated into two subunits.

The **Golgi apparatus** is an organelle that receives newly-synthesized proteins from the endoplasmic reticulum and processes them for subsequent delivery to other destinations. It is composed of several flattened membrane disks arranged in a stack.

The **cisternae** of the Golgi apparatus are the successive stacks, each bounded by a membrane, that make up individual compartments.

The ***cis* face** of the Golgi is the side juxtaposed to the nucleus.

The ***trans* face** of the Golgi is juxtaposed to the plasma membrane.

Lipid trafficking is the movement of lipids among the various membranes of a eukaryotic cell.

An **endosome** is an organelle that functions to sort endocytosed molecules and molecules delivered from the trans-Golgi network and deliver them to other compartments, such as lysosomes. It consists of membrane-bounded tubules and vesicles.

A **lysosome** is an organelle that contains hydrolytic enzymes and has an acidic lumen (pH as low as 4.5). Its primary function is the degradation of endocytosed material.

The **peroxisome** is an organelle in the cytoplasm enclosed by a single membrane. It contains oxidizing enzymes.

Membranes occupy a major part of the eukaryotic cell. In addition to surrounding individual organelles, large sheets of membranes are a prominent feature of many cells (especially those involved in secreting proteins). The electron micrograph in **Figure 8.19** shows an extensive sheet of membranes that extends from the nucleus. This is the **endoplasmic reticulum**, a highly convoluted sheet of membranes

representing 30–60% of total membrane. The interior space comprises the **lumen**.

Visualized in the electron micrograph, the endoplasmic reticulum can be divided into two types: **rough ER** and **smooth ER**. They are part of the same membrane sheet. The characteristic appearance of the rough ER results from the presence of **ribosomes** on its cytoplasmic surface. The ribosomes are small particles concerned with the synthesis of proteins. Their presence is an indication that proteins are being synthesized at the cytoplasmic surface of the endoplasmic reticulum, which then processes them for assignment to various cell compartments.

Between the endoplasmic reticulum and the plasma membrane lies the **Golgi apparatus**. The electron micrograph in **Figure S 20** shows the Golgi as a series of individual membrane sheets, tightly packed together.

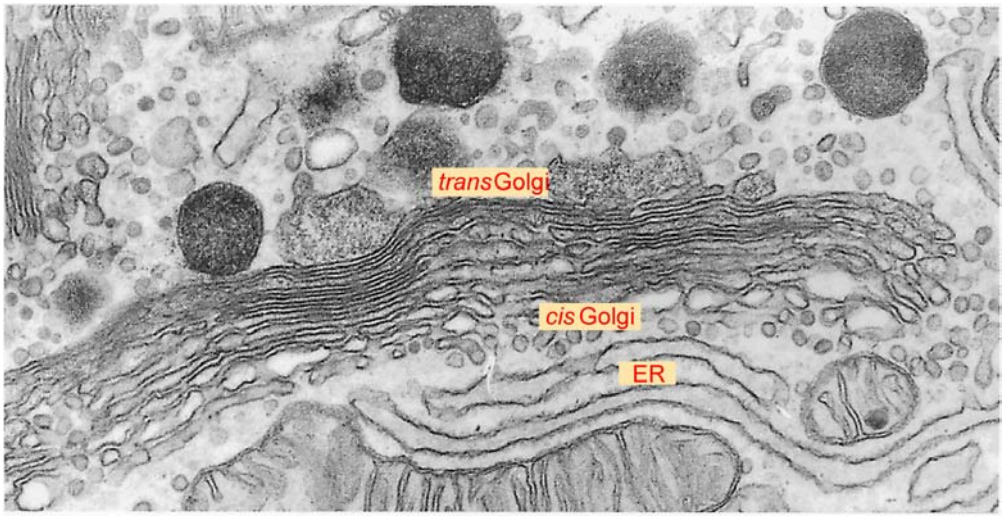

Figure S 20 The Golgi apparatus consists of a series of individual membrane stacks. Photograph kindly provided by Alain Rambourg.

The relationship between the ER and Golgi is depicted diagrammatically in **Figure S 21**. The ER is shown as a sheet made from the folding of a single lipid bilayer that extends from the outer membrane surrounding the nucleus. The Golgi consists of a "stack" of flat **cisternae**, like a pile of discs. Each cisterna consists of a closed structure bounded by a single continuous membrane. A stack usually consists of <10 cisternae, but the number of stacks varies considerably among different types of cell. The cisternae appear to be separate structures, each contained by its own membrane.

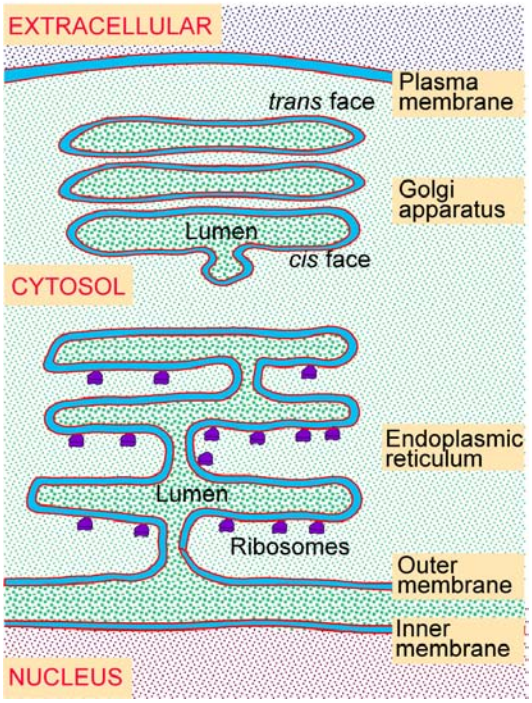

Figure S 21 The endoplasmic reticulum consists of a continuous sheet of highly folded membranes extending from the outer nuclear membrane. The Golgi consists of stacks of separate cisternae.

The Golgi apparatus is polarized. In secretory cells, the ***cis* face** is associated with the endoplasmic reticulum. The ***trans* face** lies toward the plasma membrane. Proteins that are moving through the system are transported from the ER to enter the Golgi at the *cis* face, pass through the stacks as they are transported to the *trans* face, and then exit at the TGN (*trans* Golgi network). The Golgi stacks have different biochemical constitutions, and can be fractionated on this basis to give preparations that represent the different enzymatic activities associated with progression across the Golgi.

Most lipids are synthesized in the endoplasmic reticulum. Other membranes presumably gain their lipid components by transport from the ER, although little is known about this process of **lipid trafficking**. The lipid transport system must be responsible for the differences in lipid composition between different cellular membranes.

Some small membrane-bound organelles are associated with the ER-Golgi system. In particular, **endosomes** are also involved in protein trafficking. They are small bodies located in the vicinity of the *trans*-Golgi and plasma membrane, and they undergo a process of active interchange with the Golgi and plasma membrane.

Lysosomes are rather small, spherical membrane-enclosed bodies that contain hydrolytic enzymes. They are formed by budding these bags of enzymes from the

Golgi apparatus. Lysosomes are heterogeneous; different vesicles contain different enzymes. Their properties depend on the particular hydrolytic activities of the particular enzymes.

An organelle of rather similar size is the **peroxisome**, another membrane-enclosed body, which contains the enzyme catalase. Together with other enzymes, it is responsible for a series of oxidizing reactions.

SUPPLEMENTS

7.32.7 Allostery

Key Terms

Allosteric regulation describes the ability of a protein to change its conformation (and therefore activity) at one site as the result of binding a small molecule to a second site located elsewhere on the protein.

The **effector site** is the site that is bound by a small molecule on an allosteric protein. The result of binding is to change the activity of the active site, which is located elsewhere on the protein.

An **active site** is the restricted part of an enzyme to which a substrate binds.

Feedback inhibition describes the ability of a small molecule product of a metabolic pathway to inhibit the activity of an enzyme that catalyzes an earlier step in the pathway.

Allosteric proteins exist in alternative conformations and have different biological properties in each conformation. The transition between conformations is influenced by the interaction of the protein with a cofactor or with another protein.

Figure S 22 illustrates the consequences of an allosteric transition. The crucial feature in allostery is the ability of a small molecule to bind to one site on the protein (the **effector site**) to trigger a change in conformation that alters the structure and activity of another site (the substrate-binding or **active site**). Allosteric transitions affect *only* the conformation – they do not change the primary sequence – and they rely heavily on making and breaking hydrogen bonds.

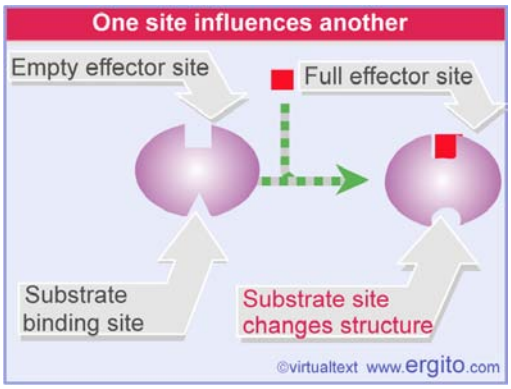

Figure S 22 When a small molecule binds to an effector site, an allosteric protein changes conformation so that the function of the substrate-binding site is altered.
This is a static version of an interactive figure; see http://www.ergito.com/main.jsp?bcs=MBIO.7.32.7 to view properly.

Allosteric proteins play an important role in both metabolic and genetic regulation. Often the product of a metabolic pathway can bind to an enzyme catalyzing an early step in the pathway to prevent it from sending further small molecules through the pathway. This interaction is called **feedback inhibition**. It is illustrated in **Figure S 23**. The equivalent interaction in genetic regulation occurs when the product of a pathway binds to a protein that turns off expression of the genes coding for the enzymes of the pathway.

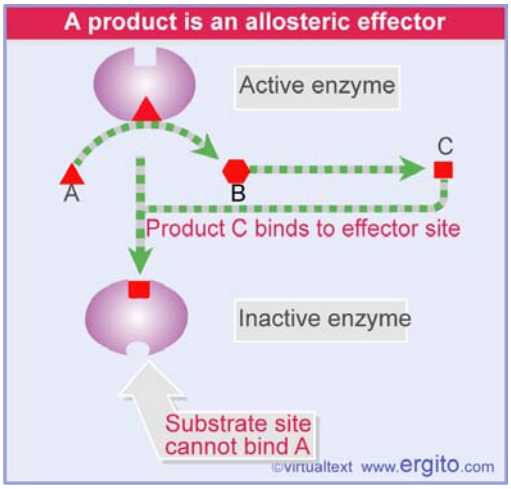

Figure S 23 Feedback inhibition results when an effector that inhibits an enzyme#s activity is itself the product of the pathway in which the enzyme functions.

Allosteric proteins are often multimeric. In such cases, the conformation of one subunit can influence the conformation of the other subunit(s). One common form of interaction occurs when the protein consists of two different types of subunit, one bearing the regulatory (effector) site, and the other bearing the catalytic (substrate-binding) site. When we redraw the interaction of **Figure S 22** in terms of a dimer, we see the result illustrated in **Figure S 24**: binding of the effector molecule to the regulatory subunit causes a change in the structure of the substrate site on the catalytic subunit. Such a change may in principle be responsible either for activating or for inhibiting the catalytic activity.

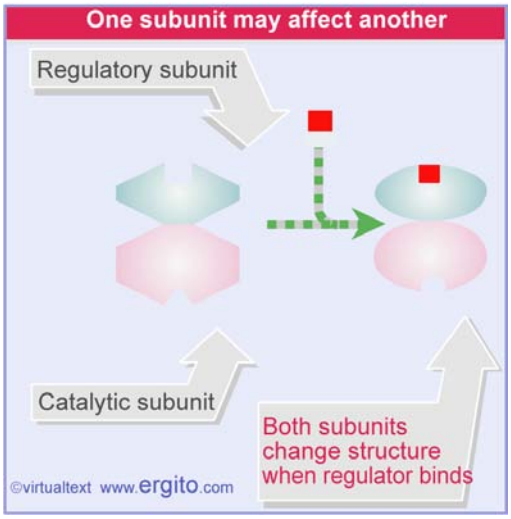

Figure S 24 Effector binding to a regulatory subunit can control the activity of the catalytic subunit of an allosteric protein.

Another interesting interaction occurs when an allosteric protein consists of identical subunits. Binding of an effector molecule to one subunit makes it become much easier for the other subunits to bind the effector. As illustrated in **Figure S 25**, this effect amplifies the effect of binding the first small molecule in such a way that the protein characteristically flips very rapidly from one state to another. This is an important ability in a regulatory protein.

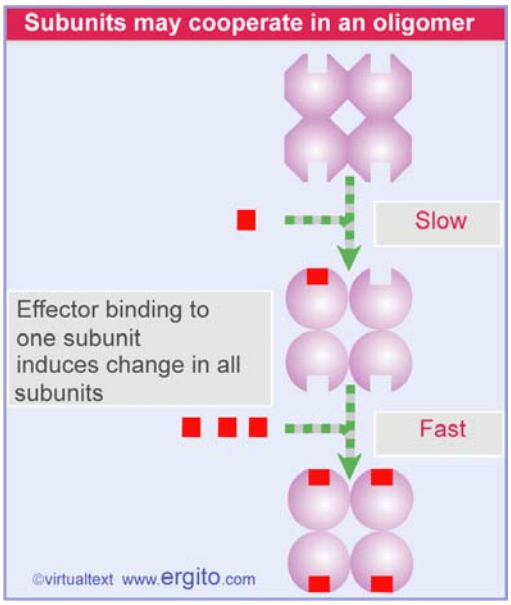

Figure S 25 Effector binding to one subunit of an allosteric protein may cause a structural change that enhances binding at the other subunits.
This is a static version of an interactive figure; see http://www.ergito.com/main.jsp?bcs=MBIO.7.32.7 to view properly.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.7.32.7

SUPPLEMENTS

7.32.8 tRNA sequences

Key Terms

The **extra arm** of tRNA lies between the T ψC and anticodon arms. It is the most variable in length in tRNA, from 3-21 bases. tRNAs are called class 1 if they lack it, and class 2 if they have it.

Figure S 26 shows the numbering of tRNA base positions in more detail.

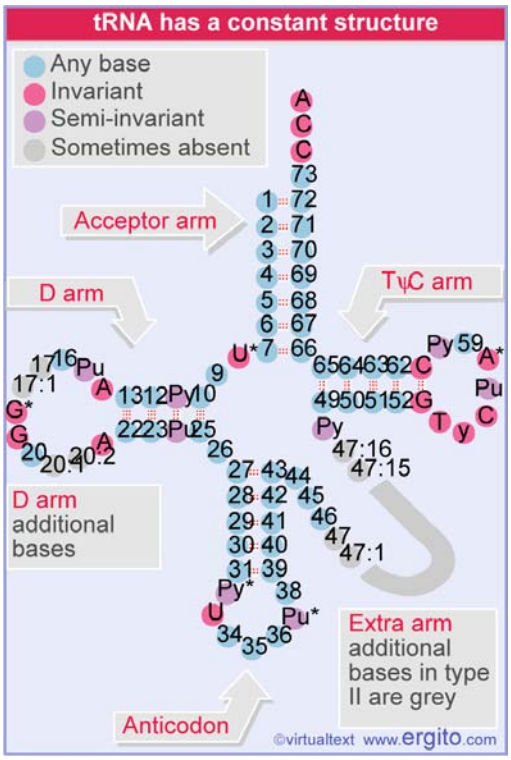

Figure S 26 tRNAs vary in length according to the sizes of the D and T Ψ C arms.

As tRNAs were sequenced, positions that seemed entirely invariant did display occasional exceptions. So for practical purposes, the description of any position as invariant means that the specified base is present in >90-95% of tRNAs. Sometimes the exceptions are individual; sometimes they fall into groups representing some peculiarity of a particular cell.

The length of the D loop varies by up to 4 residues. The extra nucleotides relative to the most common structure are denoted 17:1 (lying between 17 and 18) and 20:1 and 20:2 (lying between 20 and 21). However, in the smallest D loops, residue 17 as well as these three is absent.

The most variable feature of tRNA is the so-called **extra arm**. Depending on the nature of the extra arm, tRNAs can be divided into two classes. *Class 1 tRNA*s have a small extra arm, consisting of only 3-5 bases. They represent ~75% of all tRNAs. *Class 2 tRNA*s have a large extra arm – it may even be the longest in the tRNA – with 13–21 bases, and ~5 base pairs in the stem. The additional bases are numbered from 47:1 through 47:18. The functional significance of the extra arm is unknown.

SUPPLEMENTS

7.32.9 Complementation

Key Terms

Interallelic complementation (intragenic complementation) describes the change in the properties of a heteromultimeric protein brought about by the interaction of subunits coded by two different mutant alleles; the mixed protein may be more or less active than the protein consisting of subunits only of one or the other type.

Negative complementation occurs when interallelic complementation allows a mutant subunit to suppress the activity of a wild-type subunit in a multimeric protein.

A **dominant negative** mutation results in a mutant gene product that prevents the function of the wild-type gene product, causing loss or reduction of gene activity in cells containing both the mutant and wild-type alleles. The effect may result from the titration of another factor that interacts with the gene product or by an inhibiting interaction of the mutant subunit on the multimer.

Complementation was originally developed as a test to determine whether two mutations lie in different genes. It consists of comparing the phenotypes when both mutations are the same piece of DNA (called the *cis* configuration) and when they are on different pieces of DNA (called the *trans* configuration).

Figure S 27 shows that, if the mutations are in the same gene, the *trans* configuration has a mutant phenotype. Both copies of the gene are mutated (each copy has one of the mutations). The *cis* configuration has wild phenotype, however, because one copy of the gene has both mutations, and the other has no mutations.

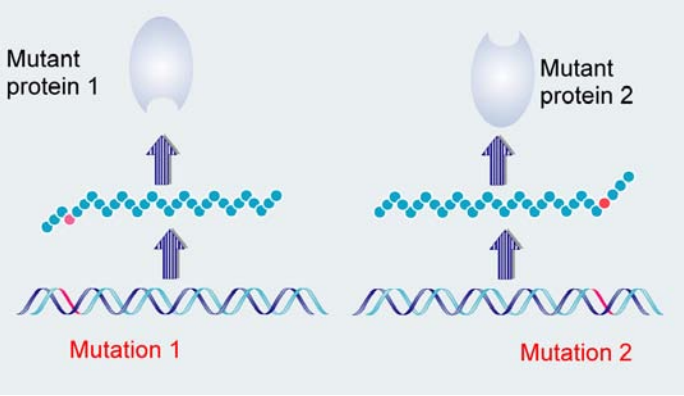

Figure S 27 When mutations are in the same gene, each allele has a different mutation, so only mutant protein is produced in the *trans* configuration.

Figure S 28 shows that, if the mutations are in different genes, the configuration does not matter. There is always one mutant copy of each gene and one wild-type copy of each gene.

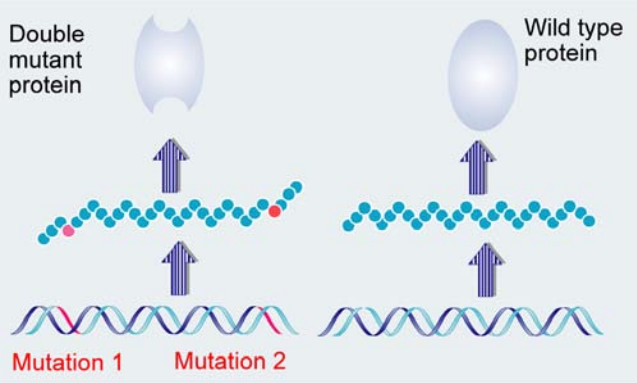

Figure S 28 When mutations are in different genes, one allele has two mutations, but the other has none, and so produces wild-type protein.

The practical form of the test, therefore, is to use the *cis* configuration as a control (it is always wild-type) and to determine whether the *trans* configuration is mutant (mutations are in the same gene) or wild-type (mutations are in different genes).

The complementation test applies to genes whose products function as independent proteins. When the gene product is a subunit of a multimeric protein, interactions between the subunits can either allow complementation between alleles or cause one allele to suppress the effect of the other.

An exception to the rule that only different genes can complement is sometimes found when a gene represents a polypeptide that is the subunit of a homomultimeric protein. In the wild-type cell, the active protein consists of several *identical* subunits. In a cell containing two mutant alleles, however, their products can mix to form multimeric proteins that contain *both types* of subunit. **Figure S 29** shows that if the two mutations compensate, the mixed-subunit protein is active, even though the proteins consisting solely of either type of mutant subunit are inactive. This effect is called **interallelic complementation**.

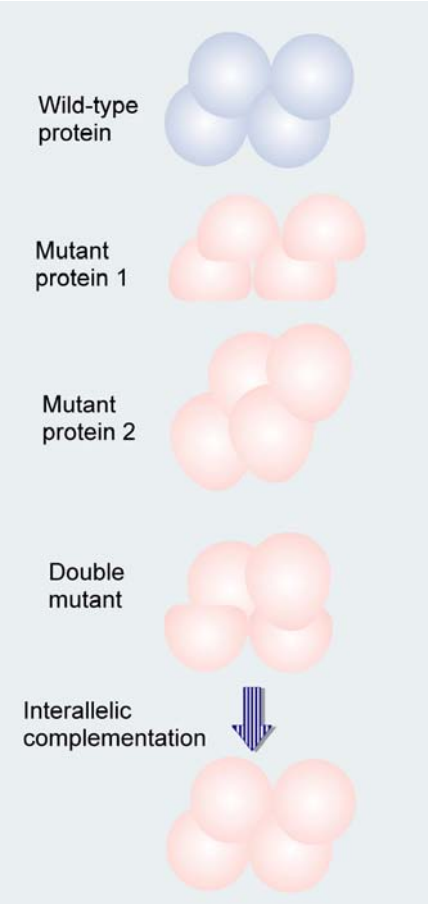

Figure S 29 Interallelic complementation occurs when different mutations in the subunits of a multimeric protein can compensate to make an active protein even though each separate subunit can only form inactive multimers.

In the reverse type of interaction, a defective subunit produced by one allele inhibits the active subunits produced by another allele. This is called **negative complementation**. **Figure S 30** shows that in effect the "bad" subunit poisons the multimeric protein so that the "good" subunits cannot function. An allele that is able to prevent other alleles from functioning is called a **dominant negative**.

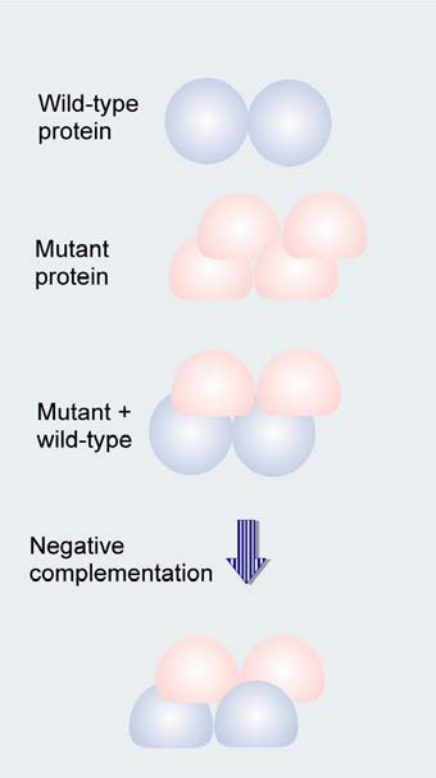

Figure S 30 Negative complementation occurs when a mutant subunit prevents the wild-type subunit from functioning.

Dominant negatives can be constructed by targeting mutagenesis at an active site in the protein. **Figure S 31** shows that one common use for this technique is to delete the DNA-binding site from a protein that is a subunit in a DNA-binding factor. When the gene for the mutant protein is introduced into the cell, the defective subunit overwhelms the normal subunits, and prevents them from forming multimers with sufficient activity to bind DNA. The same principle can be applied to any situation in which a protein forms a subunit of a multimeric protein. The subunits do not have to be identical. In **Figure S 31**, one subunit could bind DNA, while the other might interact with other proteins. The function of the second protein would be prevented by a dominant-negative subunit without a DNA-binding site.

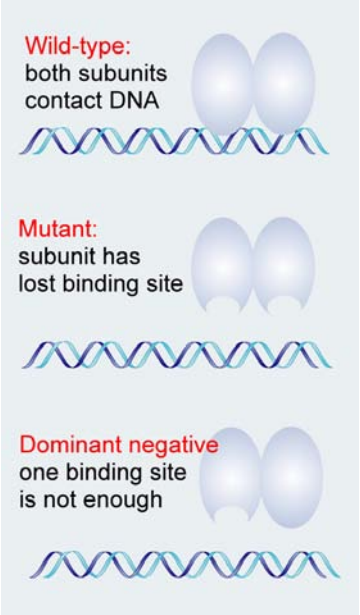

Figure S 31 A dominant negative mutant can be constructed by removing the DNA-binding site from a subunit of a DNA-binding protein.

The same technique can be used to target any active site, for example, the kinase site of a multimeric protein kinase enzyme. It is enormously useful because it can be used in circumstances where it is impossible to mutate the endogenous protein. This allows the function of the protein to be tested directly by introducing the mutant allele.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.7.32.9

SUPPLEMENTS

7.32.10 G proteins

Key Terms

A **serpentine** receptor has 7 transmembrane segments. Typically it activates a trimeric G protein.

A **second messenger** is a small molecule that is generated when a signal transduction pathway is activated. The classic second messenger is cyclic AMP, which is generated when adenylate cyclase is activated by a G protein (when the G protein itself was activated by a transmembrane receptor).

There are two types of G proteins. The name reflects the ability to bind a guanine nucleotide. The guanine nucleotide can alternate between GDP and GTP, and controls the activity of the protein. Both types of G protein work on the same principle that the GDP-bound form is inactive, and the GTP-bound form is active.

Trimeric G proteins are associated with the cytosolic face of the membrane. They are involved in the initial stages of signal transduction. They are activated by transmembrane receptors, most typically by **serpentine** receptors (7-membrane pass proteins). The three subunits are called α, β, and γ. The α subunit binds the guanine nucleotide.

The inactive form of the G protein is bound to GDP. In this form, the G protein is constitutively associated with a membrane receptor. When the receptor is activated (usually by binding ligand) it causes GDP to be displaced from the G protein. Because the concentration of GTP in the cytosol is much greater than that of GDP, the vacant nucleotide binding site is filled with GTP.

Figure S 32 shows how trimeric G proteins are activated. Binding of GTP causes the G protein to dissociate into a free α subunit and free β γ dimer. Depending on the individual G protein, it can be either the α subunit or the β γ dimer that transmits the signal to the next stage in the pathway. Whichever is the active component (and sometimes both are active) may either activate or repress the activity of a target protein.

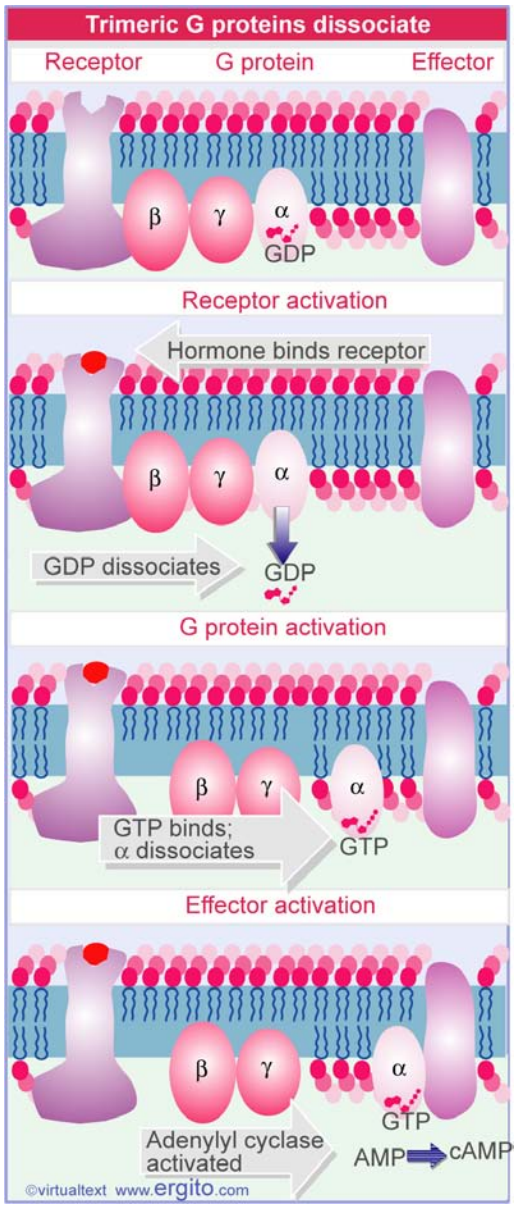

Figure S 32 When a receptor is activated by hormone binding, it causes GTP to replace GDP on a G_α subunit. The G_α subunit dissociates from the β γ dimer, and activates an effector such as adenylate cyclase.
This is a static version of an interactive figure; see http://www.ergito.com/main.jsp?bcs=MBIO.7.32.10 to view properly.

The target protein often is also associated with the membrane. This chain of events often stimulates the production of **second messengers**. In one classic example, when the protein G_s is activated, the α subunit then activates adenylate cyclase, which generates cyclic AMP.

How long the G protein remains active is controlled by the α subunit. All α subunits are GTPases. When the GTP is hydrolyzed to GDP, the α subunit reassociates with the β γ dimer to reconstitute the trimeric G protein. By removing the individual subunits, the hydrolysis of GTP terminates the physiological response.

Each α subunit hydrolyzes its GTP *in vitro* at a characteristic (slow) rate, typically with a half-life ~15 secs. But some of the physiological reactions are much shorter lived. For example, in the classic system of vision, a light response terminates in ~100 msec. The rate of GTP hydrolysis can be accelerated *in vivo* by interaction with another component of the system. This type of interaction was originally discovered for monomeric G proteins (see below), where the relevant component is a called a GAP. A common type of protein with GAP function for the α subunits of trimeric G proteins is the RGS (G protein signaling) class (for review see 2276). An RGS acts indirectly by affecting teh conformation of the α subunit so that it becomes a more effective GTPase.

Monomeric G proteins are cytosolic and are often used as binary switches in signalling or other pathways. They work on the same principle as the α subunit of a trimeric G protein. A monomeric G protein is a GTPase that hydrolyzes its bound GTP. This converts it from an active state to an inactive state.

Figure S 33 shows that three types of ancillary proteins influence the balance between the GDP- and GTP-bound forms of a monomeric G protein.

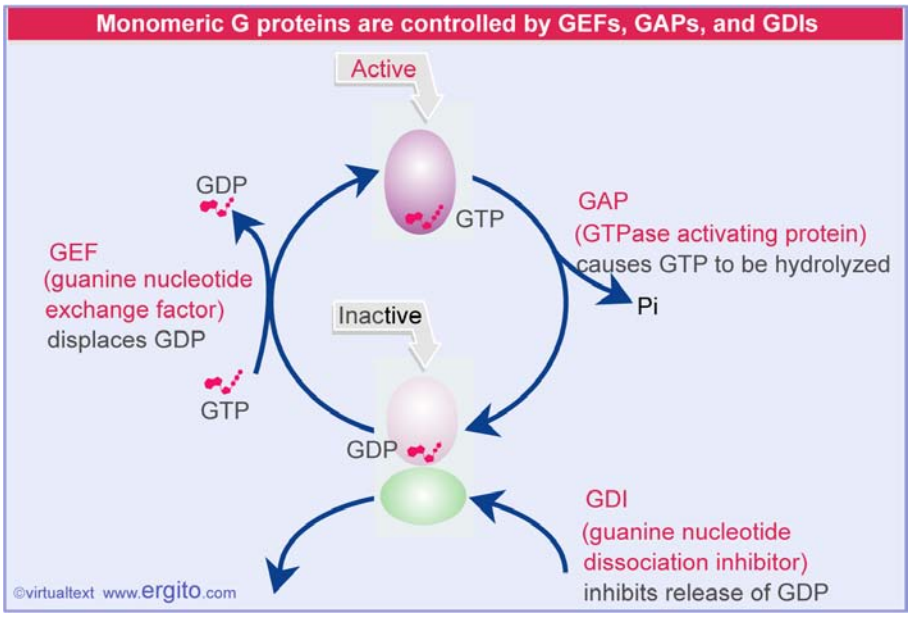

Figure S 33 Monomeric G proteins are active when bound to GTP and inactive when bound to GDP. Their activity is controlled by other proteins.
This is a static version of an interactive figure; see http://www.ergito.com/main.jsp?bcs=MBIO.7.32.10 to view properly.

- A GAP (GTPase activating protein) stimulates the GTPase activity. This is needed for a fast reaction time, because the intrinsic rate of GTP hydrolysis is slow. Thus GAP activity inactivates the G protein. Different GAPs have

specificities for different GTP-binding proteins; they are typically named as *Protein*-GAP, where Protein is the monomeric G-protein on which they act.

- A GEF (guanine nucleotide exchange factor) displaces the GDP bound to an inactive G protein. The principle of replacement is the same as for the trimeric α subunit. Release of the GDP creates an empty site. The concentration of GTP in the cytosol is greater than that of GDP, so the site is then filled with GTP. This activates the protein. GEFs have the same sort of specificity as GAPs, and similarly are named in the form *Protein*-GEF (for review see 3217).

- A GDI (guanine nucleotide dissociation inhibitor) can block the displacement reaction. This maintains the G protein in the inactive state.

Examples of specific monomeric G proteins are EF-Tu (see *Molecular Biology 2.6.10 Elongation factor Tu loads aminoacyl-tRNA into the A site*), Ran (see *Molecular Biology 2.8.28 Transport receptors carry cargo proteins through the pore*), ARF and Rab (see *Molecular Biology 6.27.7 Vesicles can bud and fuse with membranes*), and Ras and the Rho family (see *Molecular Biology 6.28.15 The activation of Ras is controlled by GTP*).

Last updated on 1-22-2002

Reviews

2276. Ross, E. M. and Wilkie, T. M. (2000). *GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins.* Annu. Rev. Biochem. 69, 795-827.

3217. Schmidt, A. and Hall, A. (2002). *Guanine nucleotide exchange factors for Rho GTPases: turning on the switch.* Genes Dev. 16, 1587-1609.

SUPPLEMENTS

7.32.11 Restriction mapping

Key Terms

End labeling describes the addition of a radioactively labeled group to one end (5′ or 3′) of a DNA strand.

The principle of restriction mapping is to break a piece of DNA into several fragments that overlap, and then to place these fragments in order by making use of the overlap. One way of generating overlapping fragments is to use enzymes that cleave at different target sites. Another is to use partial cleavage, so that an enzyme attacks any individual target site with less then 100% efficiency.

Figure S 34 shows an example of cleavage by different enzymes. A DNA molecule of length 5000 bp is incubated separately with two restriction enzymes, A and B. After cleavage the DNA is electrophoresed. The sizes of the individual fragments generated by enzyme A (left) or enzyme B (right) are determined by comparison with the positions of fragments of known size, such as the control shown in the center. This demonstrates that enzyme A has cut the substrate DNA into four fragments (lengths 2100, 1400, 1000, and 500 bp), while enzyme B has generated three fragments (lengths 2500, 1300, and 1200 bp).

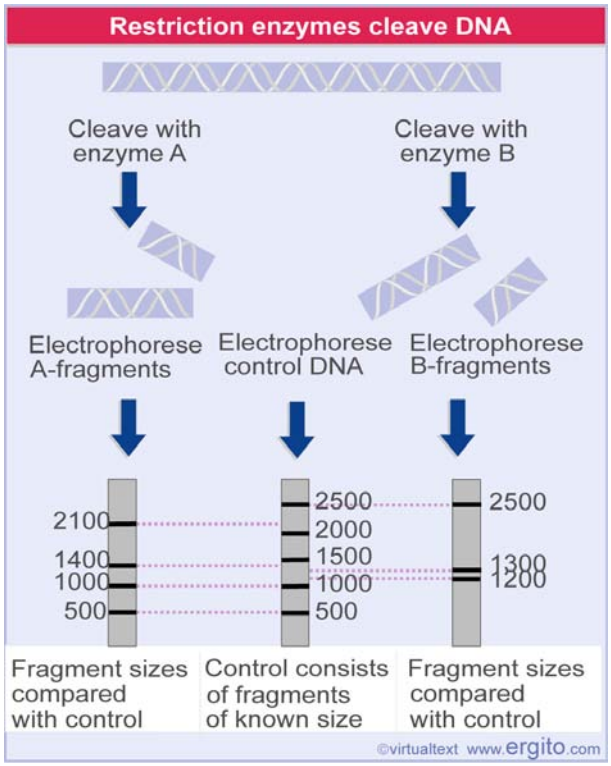

Figure S 34 DNA can be cleaved by restriction enzymes into fragments that can be separated by gel electrophoresis.

The patterns of cutting by the two enzymes can be related by several means. **Figure S 35** illustrates the principle of analysis by *double digestion*. In this technique, the DNA is cleaved simultaneously with two enzymes as well as with either one by itself. The most decisive way to use this technique is to extract each fragment produced in the individual digests with either enzyme A or enzyme B and then to cleave it with the other enzyme. The products of cleavage are analyzed again by electrophoresis.

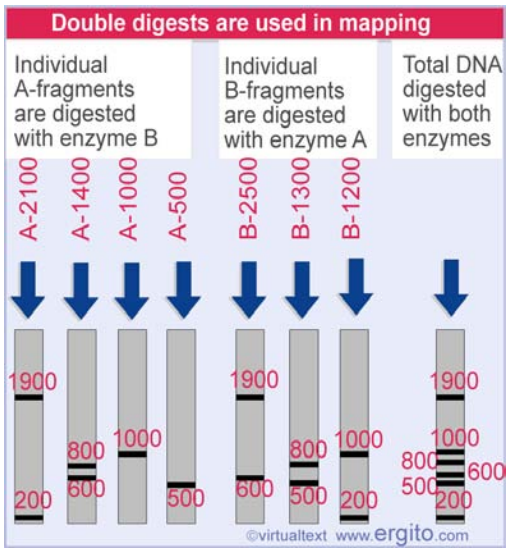

Figure S 35 Double digests define the cleavage positions of one enzyme with regard to the other.

We can use these data to construct a map of the original 5000 bp molecule of DNA, as illustrated by the stages of **Figure S 36**.

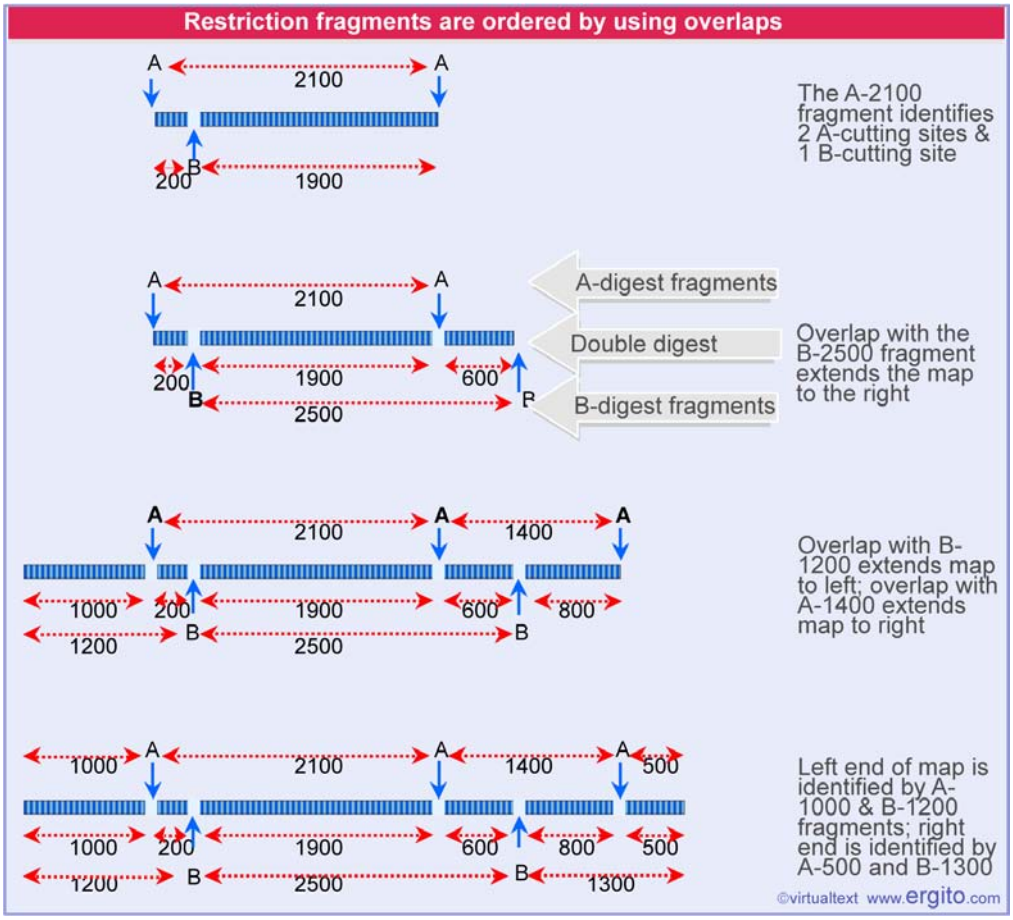

Figure S 36 A restriction map can be constructed by relating the A-fragments and B-fragments through the overlaps seen with double digest fragments.

Each gel in **Figure S 35** is labeled according to the fragment that was isolated from the gel in **Figure S 34**. A-2100 identifies the fragment of 2100 bp produced by degrading the original DNA molecule with enzyme A. When this fragment is retrieved and subjected to enzyme B, it is cut into fragments of 1900 and 200 bp. So one of the cuts made by enzyme B lies 200 bp from the nearest site cut by enzyme A on one side, and is 1900 bp from the site cut by enzyme A on the other side. This situation is described by the top map in **Figure S 36**.

A related pattern of cuts is seen when we examine the susceptibility of fragment B-2500 to enzyme A. It is cut into fragments of 1900 and 600 bp. So the 1900 bp fragment is generated by double cuts, with an A site at one end and a B site at the other end. It can be released from either of the single-cut fragments (A-2100 or B-2500) that contain it. These single-cut fragments must therefore *overlap* in the region of the 1900 bp of the common fragment that can be generated from them. This is described in the second map of **Figure S 36**, which extends our map to the right to add a cleavage site for enzyme B.

The key to restriction mapping is the use of overlapping fragments. Because of the overlap of A-2100 and B-2500 in the central region of 1900 bp, we can relate the A site 200 bp to the left of the 1900 bp region with the B site 600 bp to the right. In the

same way, we can now extend the map farther on either side. The 200 bp fragment at the left is also produced by cutting B-1200 with enzyme A, so the next B site must lie 1000 bp to the left. The 600 bp fragment at the right is also produced by cutting A-1400 with enzyme B, so the next A site must lie 800 bp to the right. This gives the third map in **Figure S 36**.

We can now complete the map by identifying the source of the two fragments at each end. At the left end, the 1000 bp fragment arises from B-1200 or in the form of A-1000, which is not cut by enzyme B. So A-1000 lies at the end of the map. Proceeding from the left end of the complete 5000 bp region, it is 1000 bp to the first A site and 1200 bp to the first B site. (This is why a B cut is not shown at the left end of the map above, although formally we treated the end as a B-cutting site in the analysis.)

At the right end of the map, the 800 bp double-cut fragment is generated by cutting B-1300 with enzyme A, so we must add a fragment of 500 bp to the right. This is the terminal fragment, as seen by its presence as A-500 in the single-cut A digest. So our completed map takes the form of the bottom map in **Figure S 36**.

The actual construction of a restriction map usually requires recourse to several enzymes, so it becomes necessary to resolve quite a complex pattern of the overlapping fragments generated by the various enzymes. Several other techniques are used in conjunction with comparison of fragments, including **end labeling**, in which the ends of the DNA molecule are labeled with a radioactive phosphate (certain enzymes can add phosphate moieties specifically to 5′ or to 3′ ends). **Figure S 37** shows that this allows the fragments containing the ends to be identified directly by their radioactive label. So in the fragment A preparation, A-1000 and A-500 would be placed immediately at opposite ends of the map; similarly, fragments B-1200 and B-1300 would be identified as ends.

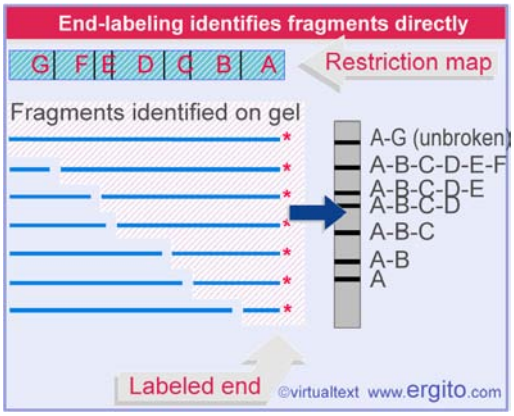

Figure S 37 When restriction fragments are identified by their possession of a labeled end, each fragment directly shows the distance of a cutting site from the end. Successive fragments increase in length by the distance between adjacent restriction sites.

A complex set of overlapping fragments can be generated directly by a single enzyme by using conditions of partial cleavage, as illustrated in **Figure S 38**. Of course, for mapping purposes then it is necessary to distinguish different fragments

that have the same size, and to determine overlaps. However, this is a useful technique when we want essentially to introduce random breaks into a large DNA (for example a whole genome) in order to obtain cloned fragments (see *Molecular Biology Supplement 32.12 Genome mapping*).

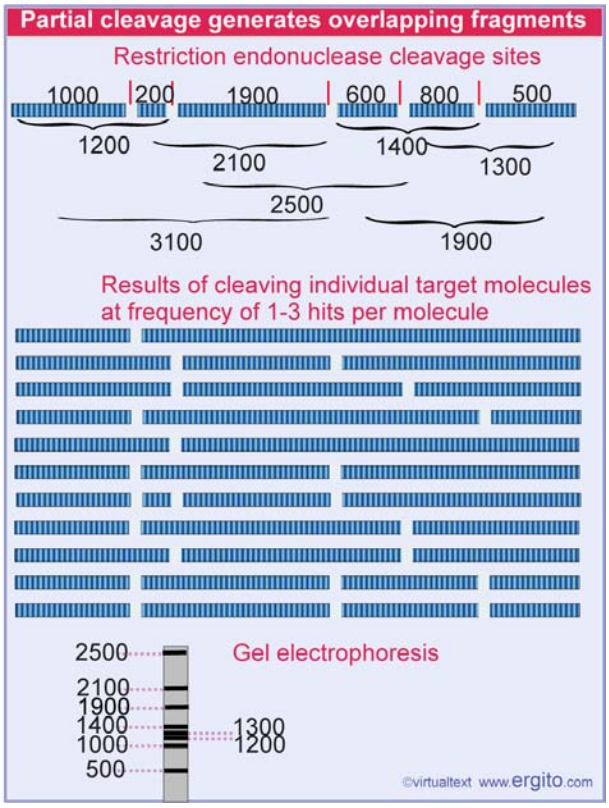

Figure S 38 Partial cleavage by a restriction endonuclease generates a series of overlapping fragments.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.7.32.11

SUPPLEMENTS

7.32.12 Genome mapping

Key Terms

A **library** is a set of cloned fragments together representing the entire genome (genomic library) or all the expressed genes (cDNA library).

A **yeast artificial chromosome (YAC)** is a synthetic DNA molecule that contains an origin for replication, a centromere to support segregation, and telomeres to seal the ends. It is used as a means to propagate whatever genes it carries in yeast cells.

A **bacterial artificial chromosome (BAC)** is a synthetic DNA molecule that contains the sequences needed for replication and segregation in bacteria. This is used in genomic cloning to amplify sequences typically 100-200 kb long. They are usually derived from derived from the naturally-occurring F factor episome.

A **contig** is a continuous stretch of genomic DNA generated by assembling cloned fragments by means of their overlaps.

Shotgun cloning analyzes an entire genome in the form of randomly generated fragments.

An **expressed sequence tag (EST)** is a short sequence of DNA taken from a cDNA copy of an mRNA. The EST is complementary to the mRNA and can be used to identify genes corresponding to the mRNA.

The principle of genome mapping is the same as restriction mapping (see *Molecular Biology Supplement 32.11 Restriction mapping*): to break a large DNA molecule into smaller molecules that are assembled in order by the principle of overlaps between their ends. Of course, there is a difference of scale when we start with a whole genome, and special tricks have to be used to assemble the fragments.

The starting point is usually the generation of a **library** of cloned fragments that represent the whole genome. This can be achieved by cleaving DNA at random, for example, by using partial cleavage conditions with a restriction endonuclease of low specificity (see **Figure S 38**). We can calculate statistically how many clone are needed for the whole genome to be represented in fragments of a given size (1441). Over the past 20 years, it has been possible to move from cloning relatively small fragments of DNA in individual plasmid or phage vectors to using synthetic chromosomes, either **YACs** (yeast artificial chromosomes) or **BACs** (bacterial artificial chromosomes). The basic principle is that the artificial chromosome consists of a long length of DNA (100-200 kb)which is able to propagate in the host cell because it has the features required for replication, segregation, and stability.

Cloned fragments are joined together by the principle of overlap, in which their ends are compared. When the end of one fragment is identical to the end of another fragment, we may conclude that the two fragments represent overlapping stretches of the genome. A series of fragments may be joined together into a **contig**, which is effectively a continuous stretch of the genome generated from overlapping

fragments. **Figure S 39** shows an example of a chromosome "walk", in which successive fragments were joined together by identifying overlapping clones from a library. A subfragment from one end of the first clone is used to isolate clones that extend farther along the chromosome. These clones in turn are used to isolate the next set. In each cycle, a new clone is selected because its restriction map coincides at one end with the end of the previous clone, but at the other end has new material. It is possible to walk for hundreds of kb, typically at a rate of >100 kb per month. Chromosome walking allows large contiguous regions of the chromosome to be represented in a library of clones.

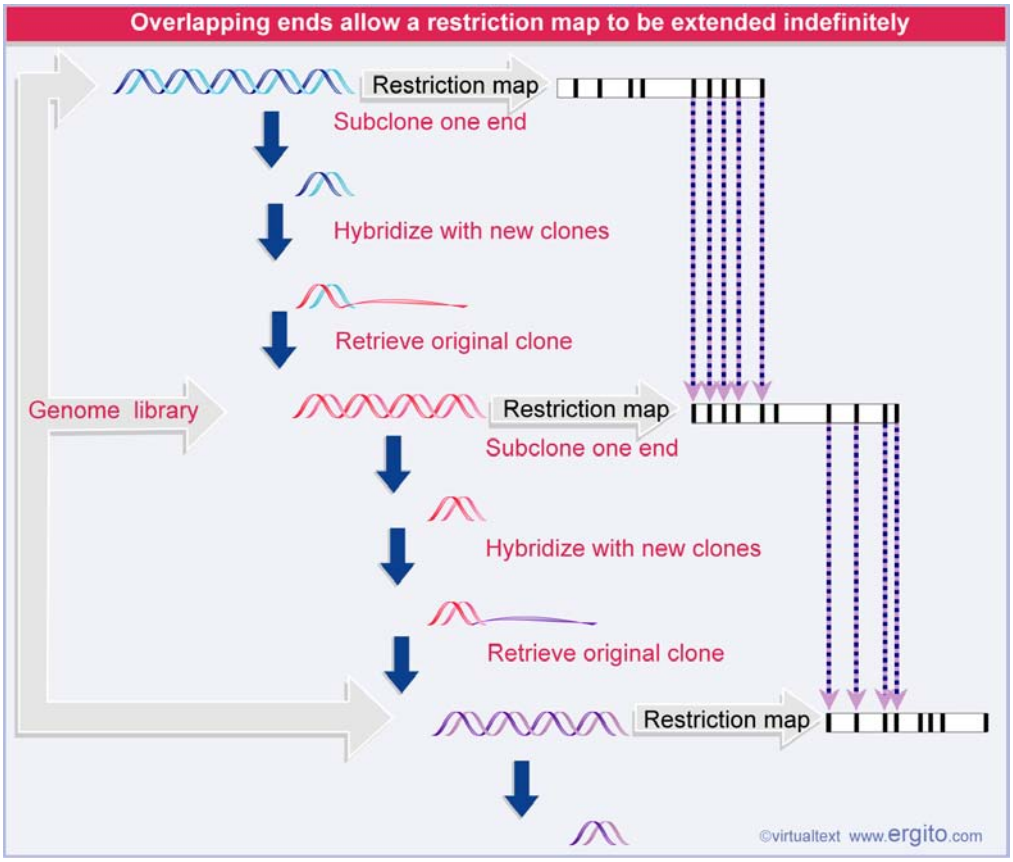

Figure S 39 Chromosome walking is accomplished by successive hybridizations between overlapping genomic clones.

Two approaches have been taken to large-scale mapping of genomes. One is to break the genome first into a series of large contigs, as illustrated in **Figure S 40**. With the contigs in hand, each contig is then sequenced, and we know that these sequences together will add up to the entire genome. This approach was used by one of the groups that sequenced the human genome (1440). In this case, the raw material was a total of ~29,000 BACs that were assembled into ~1250 contigs. The contigs were mapped to chromosomal locations and merged.

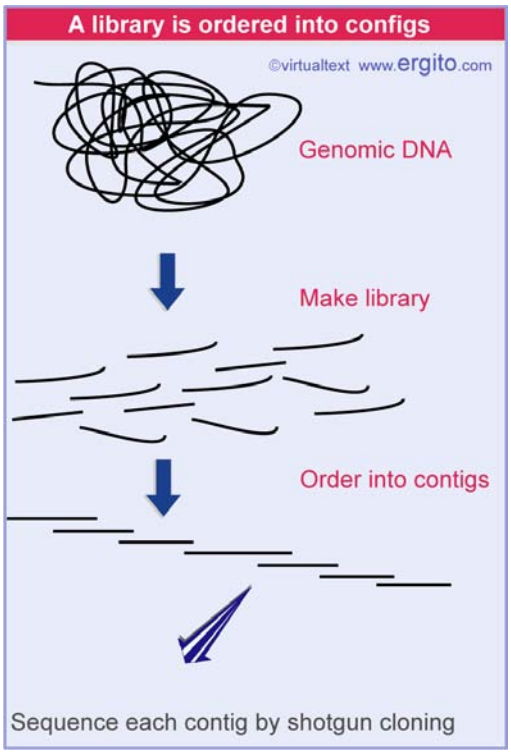

Figure S 40 A genome may be organized into a set of contigs before it is sequenced.

Another approach is to use **shotgun** cloning, as illustrated in **Figure S 41**. Here the entire genome is broken into random fragments that are sequenced, and then the sequences are joined by the usual principle of identifying overlaps. This requires very sophisticated computer processing to identify all the overlaps. This was also used to analyze the human genome (1439).

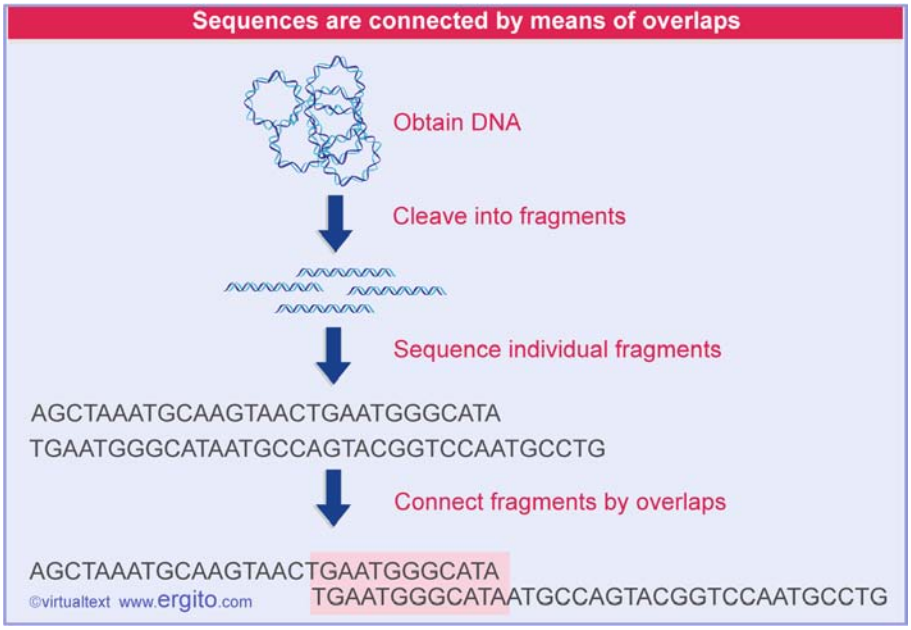

Figure S 41 The principle of shotgun cloning is to break a DNA at random into fragments that are sequenced. Overlapping fragments are identified by sequence comparisons and connected into a continuous sequence.

Analysis of fragments generated by shotgun cloning is usually involved at some stage. Identifying the overlaps becomes easier the smaller the size of the starting fragment. So when a contig is analyzed by shotgun cloning, the process is simpler than when a genome is analyzed, by virtue of the reduced complexity of the DNA sequences.

Joining sequences on the basis of overlaps would be relatively straightforward if the genome consisted only of unique (nonrepetitive) sequences. The major problem in practice is caused by the presence of repetitive sequences, where several stretches of DNA may be so similar that it is impossible to distinguish among them. **Figure S 42** illustrates the difficulty. In its simplest form, it is difficult to line up two overlapping fragments because we do not know how many copies of the repeating unit there are. **Figure S 43** shows the more complex condition in which the same sequences occur in multiple locations in the genome, we do not know which of any two overlapping fragments should be joined.

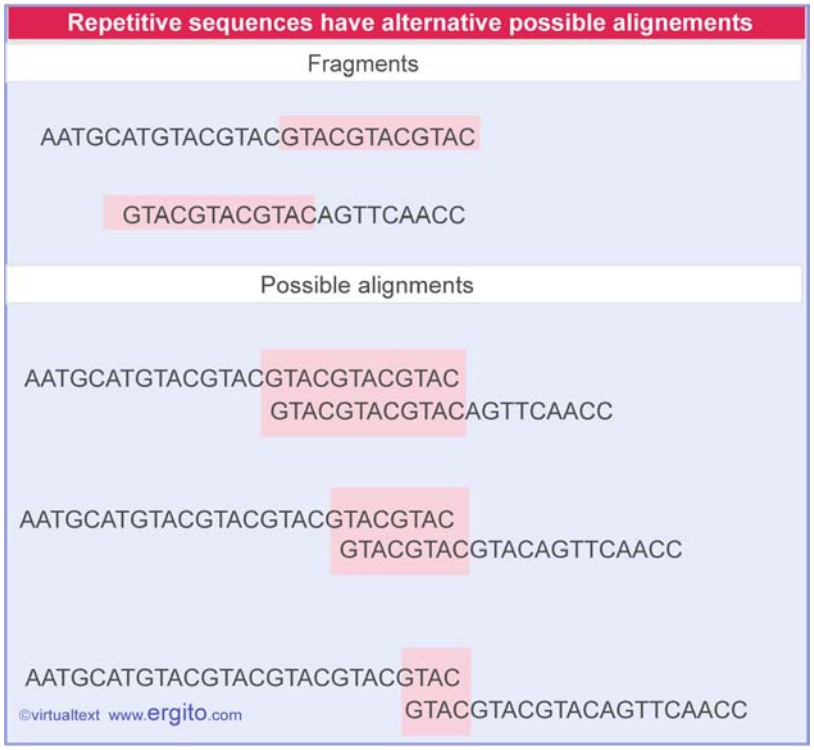

Figure S 42 Fragments that overlap in a repeated sequence may have multiple possible alignments.

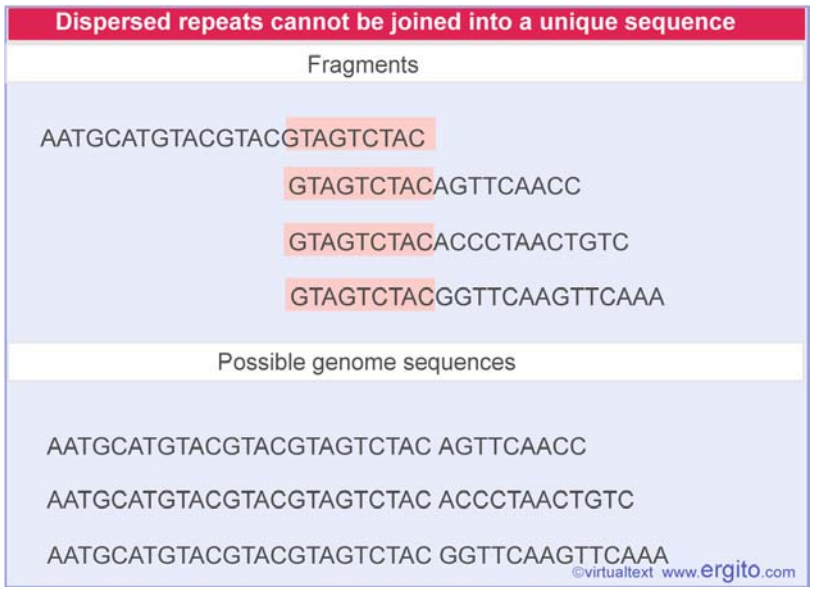

Figure S 43 Dispersed duplicated sequences may have multiple possible alignments.

In the same way that genomic libraries can be prepared from genomicDNA, cDNA libraries can be prepared by reverse transcribing populations of mRNA. This is an important tool in identifying and mapping expressed genes. In principle, each RNA molecule provides the template for synthesizing a single-stranded cDNA that is

complementary to it. The cDNA is then converted to a double-stranded form and cloned in the usual way. Any individual RNA should be represented in the cloned population in proportion to its abundance in the original RNA population (although the actual proportions are distorted by the idosyncracies of reverse transcription).

In its simplest form, a cDNA library consists of a population of cloned DNA molecules that represent all of the expressed sequences in the source cell type. However, using cDNA clones as such turns out to be an efficient strategy for identifying and mapping only the shorter genes.

More sophisticated variations of the procedure have been introduced to make it possible to automate data collection. **Expressed sequence tags** (**EST**) are obtained by sequencing the extremities of cDNA clones, so typically they correspond to the 5′ leader and 3′ trailer. They can be used to provide large data sets that allow expressed genes to be identified (2220). The principle that an EST is located on the genome by its sequence identity with a cloned sequence of genomic DNA.

The most recent development is the use of a strategy qhere sequences are produced from within the transcript, instead of from the ends. This is called ORF EST (meaning that the EST represents an open reading frame). This provides a better data set, with less confusion from overlapping or repeated sequences, and allows large libraries to be made that accurately represent the expresed set of genes (2219).

Last updated on 12-13-2001

References

1439. Venter, J. C. et al. (2001). *The sequence of the human genome*. Science 291, 1304-1350.

1440. International Human Genome Sequencing Consortium. (2001). *Initial sequencing and analysis of the human genome.* Nature 409, 860-921.

1441. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978). *The isolation of structural genes from libraries of eucaryotic DNA.* Cell 15, 687-701.

2219. Carmargo, A. A. et al. (2001). *The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome.* Proc. Natl. Acad. Sci. USA 98, 12103-12108.

2220. Adams, M D. et al. (1991). *Complementary DNA sequencing: expressed sequence tags and human genome project*. Science 252, 1651-1656.

SUPPLEMENTS

7.32.13 Two-component signal transduction

Two-component signaling systems were discovered in bacteria, where they provide the most common form of signaling pathway that responds to extracellular events (2307). There are ~30 such pathways in *E. coli*. **Figure S 44** illustrates a generic system. They are also found in plant cells (a typical plant may have 10-15 such pathways), occasionally in yeast (~5 pathways), but are relatively rare in animal cells (there are none in mammals), where kinase cascades involving Ser/Thr and Tyr kinases are more common.

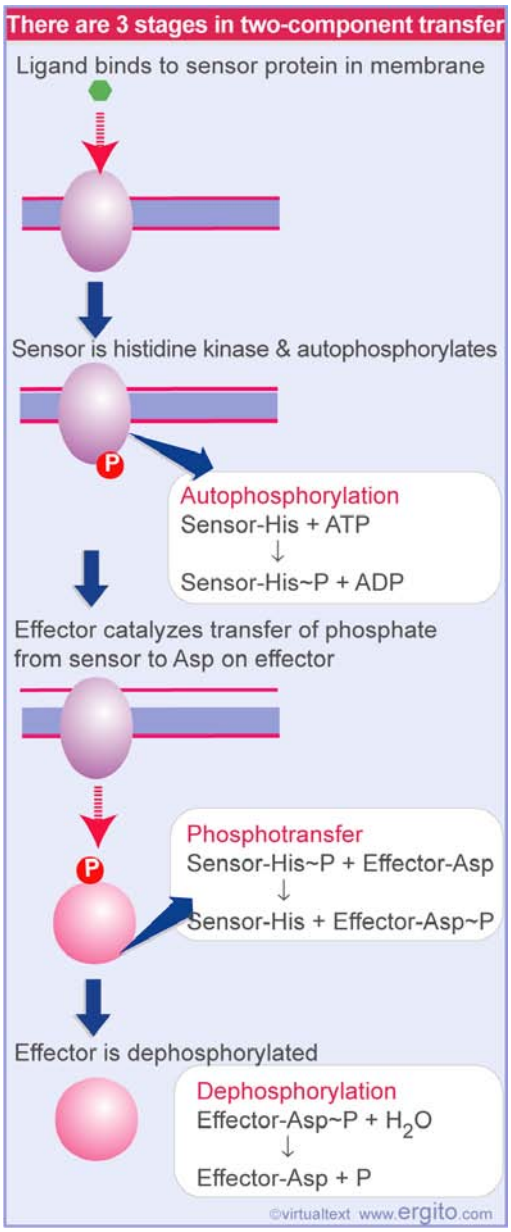

Figure S 44 A two component system consists of a sensor that is an autophosphorylating histidine kinase and an effector that can catalyze transfer of phosphate from the sensor to itself.

The sensor protein is a histidine kinase that is located in the bacterial membrane. It can be activated by binding a ligand that is in the extracellular medium. Activation causes the kinase to autophosphorylate (that is, to phosphorylate itself). The reaction transfers the γ phosphate from ATP on to a histidine residue in the kinase.

The sensor interacts with an effector protein (also called a "response regulator"). The effector catalyzes transfer of the phosphate group from the histidine on the sensor to an aspartic acid residue in its own regulatory domain. This activates the effector. It is later deactivated by dephosphorylation.

The chemistry of phosphorylation by a histidine kinase is different from that of the eukaryotic serine/threonine or tyrosine kinases. **Figure S 45** shows that the phosphate is transferred on to a nitrogen atom in the histidine ring, creating a high-energy phosphoramidate bond. This type of high energy bond is used for phosphoryl transfer in many proteins. When the phosphate is transferred to aspartic acid, it generates a high energy acyl bond. By contrast, the bonds formed with serine, threonine, or tyrosine are between phosphate and hydroxyl group, generating a low-energy ester bond.

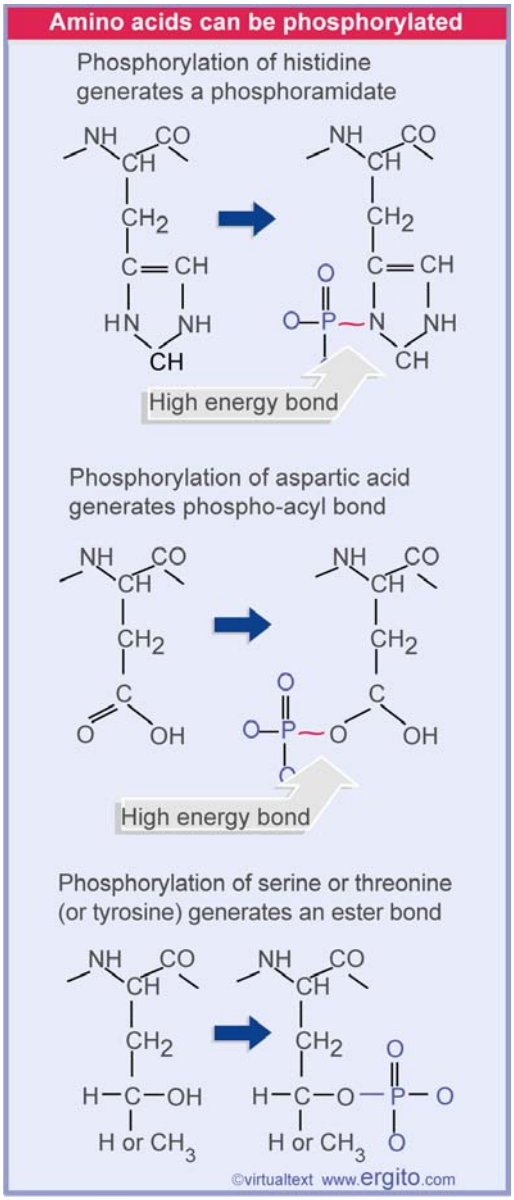

Figure S 45 Phosphorylation of histidine or aspartic acid occurs through a high energy bond, but phosphorylation of serine, threonine, or tyrosine occurs as a low energy ester.

More elaborate versions of these systems are called phosphorelays. In such cases, the

phosphate group is transferred several times. In a typical case, the pathway starts just like a two-component system, but the second protein in turn transfers the phosphate group to another, and the process may continue (for review see 2308; 2263). The high energy bonds tend to be short-lived, which makes the response rapid and transient.

The usual end target of a two-component pathway is the regulation of gene transcription, as summarized in**Figure S 46**. In the typical bacterial pathway, the effector is the terminal component. It has two domains. The regulatory domain catalyzes transfer on to itself of the phosphate from the sensor histidine kinase (for review see 2263). When it is phosphorylated, it activates the effector domain, which most commonly binds DNA to activate or to repress transcription .

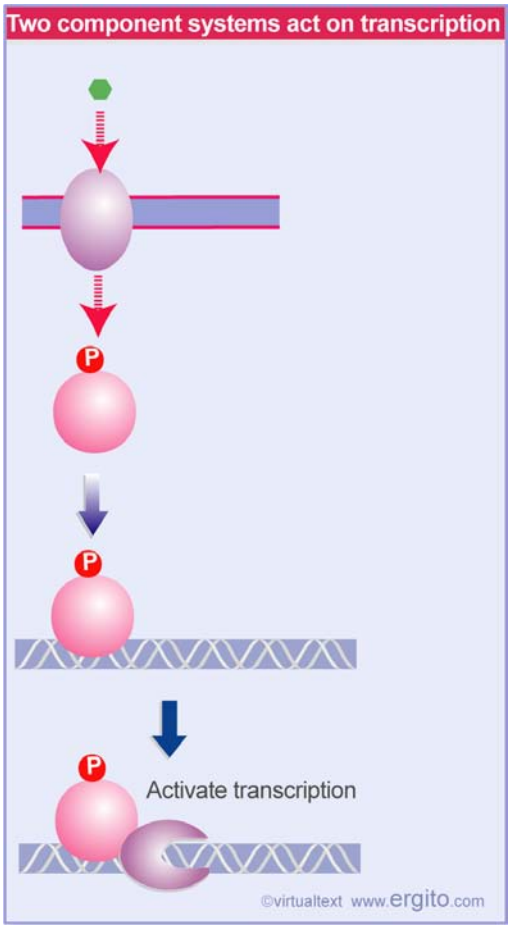

Figure S 46 The effector protein of a two-component system has two domains. When the regulator domain is phosphorylated, the effector domain binds to DNA. This may activate transcription (as shown) or may repress it.

Several basic responses to the environment are mediated by two-component systems, in *E. coli* including the response to osmotic pressure, redox control, and chemotaxis. A two-component system is used by Agrobacterium when it infects a plant cell (see *Molecular Biology 4.18.14 T-DNA carries genes required for infection*). Sporulation

of *B. subtilis* is initiated by a phosphorelay system (see *Molecular Biology 3.9.19 Sporulation is controlled by sigma factors*).

Last updated on 1-21-2002

Reviews

2263. Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). *Two-component signal transduction.* Annu. Rev. Biochem. 69, 183-215.

2308. Parkinson, J. S. (1993). *Signal transduction schemes of bacteria.* Cell 73, 857-871.

References

2307. Nixon, B. T., Ronson, C. W., and Ausubel, F. M. (1986). *Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes ntrB and ntrC.* Proc. Natl. Acad. Sci. USA 83, 7850-7854.