ELSEVIER

Physica A 313 (2002) 211–237

PHYSICA A

www.elsevier.com/locate/physa

Physics of protein–DNA interaction

Robijn F. Bruinsma[a,b,*]

[a] *Department of Physics and Astronomy, University of California, Los Angeles, CA, 90024, USA*
[b] *Instituut-Lorentz for Theoretical Physics, Universiteit Leiden, Postbus 9506, 2300 RA, Leiden, The Netherlands*

1. Introduction

1.1. The central dogma and bacterial gene expression

1.1.1. Two families

Life is based on a symbiotic relationship between two families of biopolymers: DNA and RNA, constituted of nucleic acids, and proteins, constituted of amino acids (for a general introduction to gene expression, see Ref. [1]). Proteins are the active agents of the cell. As *enzymes*, they control the *rates* of biochemical reactions taking place inside the cell. They are responsible for the *transcription* of the genetic code, i.e., the production of copies of short segments of the genetic code that are used as blueprints for the production of new proteins, and for the *duplication* of the genetic code, i.e., the production of a full copy of the genetic code during cell division. Synthesis of other macromolecules, such as lipids and sugars, is carried out by proteins, the mechanical force of our muscles is generated by specialized proteins adept at "mechano-chemistry", they detect light, sound, and smell, and maintain the structural integrity of cells.

If we view the cell as a miniature chemical factory that simultaneously runs many chemical processes, then the proteins form the control system of the factory, turning reactions on and off. The control system obeys orders from the central office: the cell nucleus. The DNA inside the nucleus can be considered as the *memory* of the computer system of the central office: it is the *information storage system* of the cell. Blueprints for the synthesis of proteins are stored in the form of DNA base-pair sequences, much like strings of zeros and ones store information in digital computers. A *gene* is the data string required for the production of one protein (actually,

* Corresponding author. Department of Physics and Astronomy, University of California, Los Angeles, CA, 90024, USA.
E-mail address: bruinsma@physics.ucla.edu (R.F. Bruinsma).

0378-4371/02/$ - see front matter
PII: S0378-4371(02)01038-5

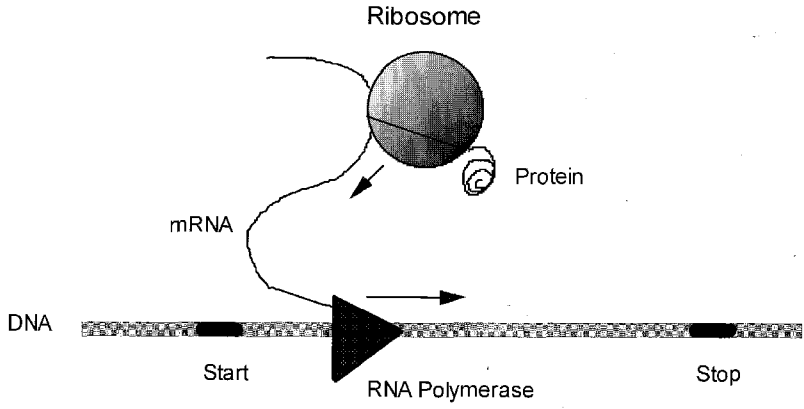

Fig. 1. Gene transcription.

multiple variants of a protein can be produced from the same gene). The beginning and end points of a gene are marked by special "start" and "stop" signals. When a protein has to be synthesized, a specialized copying protein, *RNA polymerase*, transcribes a copy of a gene beginning at the start signal and ending at the stop signal (see Fig. 1).

This copy is in the form of an *RNA* strand known as *mRNA* (or "messenger" RNA). A huge molecular machine, the *ribosome*, synthesizes the protein from the mRNA blueprint. Interestingly, these ribosomes are compound constructs of RNA strands (known as *rRNA*) and proteins, with the active biochemistry carried out not by the protein part, as you might have expected, but by the RNA part. Indeed, unlike DNA, RNA strands are in fact capable to act as enzymes.

The information stream is strictly *one way*: DNA contains the information required for the synthesis of proteins. The genetic code is not altered by the transcription, and RNA strands do not insert their code into DNA. We call this basic principle of biochemical information flow the "*central dogma*". We know next to nothing about how this elaborate relationship between the nucleic and amino acids developed. The basic chemical structure of the two families is quite different. The molecular biology of living organisms is all highly similar and based on the central dogma and we do not know of the existence of more primitive molecular information and control systems from which we could somehow infer a developmental history (though we suspect that once upon a time both information storage and enzymatic activity were based purely on RNA since RNA is able to carry out enzymatic activity as we saw). The central dogma applies to *living* organisms. Retroviruses are able to insert their RNA code into host DNA, using a special enzyme called "reverse transcriptase". This looks like an exception to the central dogma but viruses are not considered living organisms since they are not able to reproduce themselves independently nor do they carry out metabolic activity, the two defining requirements of a living organism.

It is reasonable to ask why *bio*polymers should be of special interest to physicists. The physics of polymers—particular synthetic polymers—has been studied for a number of decades and an elegant, general theoretical framework is available. The motivation behind a study of the interaction between DNA and proteins is quite different from that of a study of synthetic polymers. In polymer physics, we want to compute the free energy and correlation functions of a *typical* polymer in a solution or melt, with results that are as much as possible independent of the detailed molecular structure of the polymers. That philosophy does not apply to biopolymers where we are dealing with highly *atypical* molecules that carry out certain *functions*. Their structure presumably evolved under the adaptive pressures exerted on micro-organisms that relied for their survival on efficient performance of the functions these molecules are involved with. A molecular biophysicist tries to shed light on how functional molecular devices work and how their design constraints are met. These are of course very complex systems, so it is a good strategy to focus as much as possible on basic principles of physics of general validity and relying as little as possible on assumptions concerning the detailed molecular structure. The hope is that this will provide us with constraints on the design and operation of functional biopolymers in the way that the second law of thermodynamics constrains the maximum efficiency of steam engines.

In order to illustrate this approach, we will focus on two special cases that have been particularly important in development of our understanding of protein–DNA interaction, the *lac repressor* and the *nucleosome complex*. These two systems have been studied in such detail that we may hope to understand how they "work" as molecular devices. In these lectures, we will see what insights thermodynamics, statistical mechanics and elasticity theory can provide us into their design. We start with a very brief introduction to the physico-chemical structure of the nucleic and amino acid biopolymer families (for an excellent introduction to DNA from the viewpoint of physics: see Ref. [2]).

1.2. Procaryote gene expression

How does an organism "know" when to turn gene transcription on and when to turn it off? We divide cells in two groups: eucaryotes and procaryotes. The cells of animals and plants—the eucaryotes—have their DNA sequestered inside a nucleus and the cell has a complex set of internal "organs" called organelles. Gene expression of eucaryotic cells, the focus of much current research, is a complex affair, which we will discuss in a later section. Bacteria, procaryotes, lack a nucleus and organelles and their gene expression is much better understood (for a very readable introduction see Ref. [3]). We will discuss a simple example: the expression of the "lac" gene of the bacterium *Escirichia coli* (*E. Coli* for short) [4].

Large numbers of the *E. coli* parasitic bacteria live inside your intestines ("colon"). When you drink a glass of milk, part of it will be metabolized not by you but by your *E. coli* bacteria. The first step is the breakdown of *lactose*, sugar molecules consisting of two linked molecular rings. Lactose is broken down into two single-ring glucose molecules. This chemical reaction requires an enzyme, called "β galactosidase", to

proceed because lactose does not dissociate spontaneously (an enzyme speeds up a reaction by lowering the energy barrier). First though, the lactose molecules must be transferred from the exterior of the bacterium to the cell interior (or "cytoplasm") across the membrane that surrounds *E. coli*. This is done by another protein, called "permease". Finally, a third protein, called "transacetylase", is required for chemical modification of the sugar molecules.

The DNA of *E. coli* carries three separate genes for the production of the three enzymes: lacY, lacZ, and lacA. Expression of these three genes starts when the environmental lactose concentration rises, and it stops when the lactose concentration drops (to avoid wasteful use of precious macromolecular material). The three genes are located right behind each other on the DNA, and—sensibly—they are transcribed collectively. Such a cluster of functionally connected genes is called an "operon". The lac operon also contains three *regulatory* sequences:

(a) *Promoter sequence*. This sequence is "recognized" by RNA polymerase. By that we mean that RNA polymerase molecules in solution bind to promoter sequences on the DNA but not to other sequences. From this start site, RNA polymerase can transcribe RNA *in either direction*. In one direction, "downstream", it produces the RNA code of our three enzymes. In the other direction, "upstream", it transcribes the neighboring "regulator" sequence.

(b) *Regulator sequence*. The regulator sequence is the code of a fourth protein: *lac repressor*. The lac repressor, which is *not* involved in the metabolic of lactose, plays a key regulatory role in turning the gene "on" or "off".

(c) *Operator sequence*. The operator sequence is a DNA sequence that is recognized by lac repressor. If lac repressor is bound to the operator sequence, then downstream gene expression is blocked. Fig. 2 shows how this "genetic switch" works.

First, assume that the concentration of lactose in the environment is high. Lactose molecules bind reversibly to the repressor protein. For high lactose concentrations, the lactose-bound form is favored under conditions of chemical equilibrium. In the lactose-bound (or "induced") form, the repressor has a different structure in which it does *not* bind to the operator sequence. RNA polymerase proteins binding to the promoter sequence are free to transcribe in the downstream (and upstream) direction. Along the downstream direction, it will produce an RNA copy of the genes of the three enzymes required for lactose breakdown. Transcription along the upstream direction produces an RNA copy of the lac repressor gene. Production of repressor proteins at a low level is necessary to maintain their concentration since proteins have a finite lifetime (after a certain period, a protein receives a molecular "tag" targeting it for future breakdown as part of the scheduled maintenance program of the cell).

Next, assume that the lactose concentration has dropped. The chemical equilibrium now favors the lactose-free conformation of the repressor. Lac repressor binds to the operator sequence and downstream gene transcription is blocked. Genetic switches of this type are used by *E. coli* (and other bacteria) to respond to changes in temperature, salinity, acidity, and the oxygen level. Efficiency of these switches clearly is a matter of life and death for the bacterium, so we should expect that the structure of proteins like

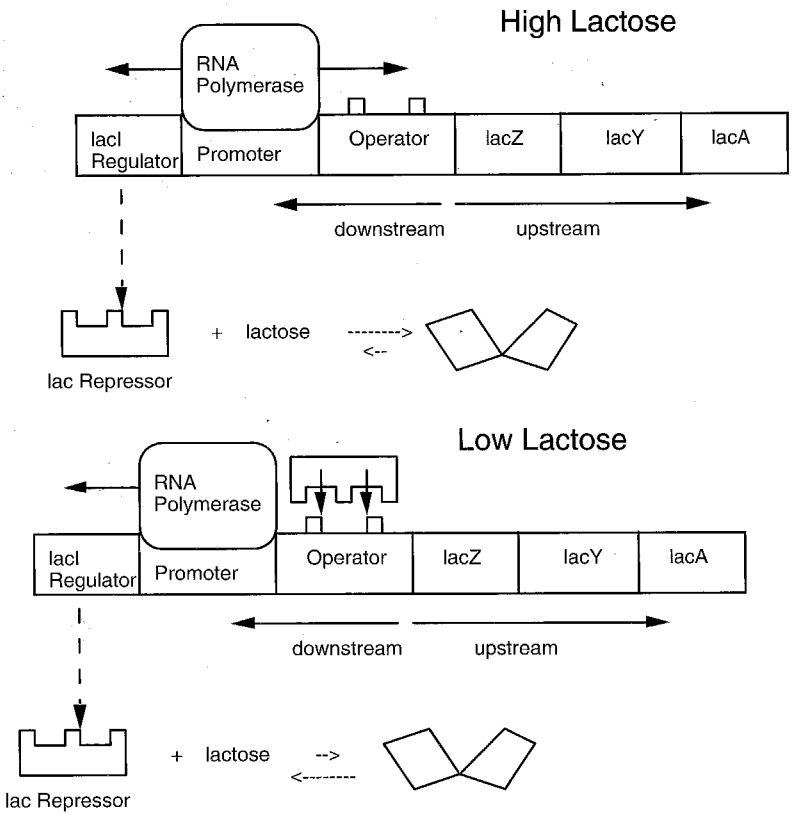

Fig. 2. The lac operon and gene regulation.

the lac repressor has been "sharpened" by natural selection for optimal performance. If you would put yourself the task of designing a lac repressor protein some obvious minimum engineering requirements would be:

Specificity. The lac repressor must be able to recognize the operator sequence. Repressor proteins must be able to efficiently "read" the DNA code.

Reversibility. The lac repressor must bind reversibly to lactose or else gene expression could not be turned off. Similarly, it must bind reversibly to DNA or else gene expression could not be turned on.

Reactivity. The lac repressor must locate the operator sequence within minutes after the lactose concentration drops. If it takes too long to turn a genetic switch, then the bacterium could be dead before it had the chance to respond to the changing environment.

In the next sections, we will see what thermodynamics, statistical mechanics, and elasticity theory have to say on these requirements. First, we have to learn more about the molecular structure of the two biopolymer families.

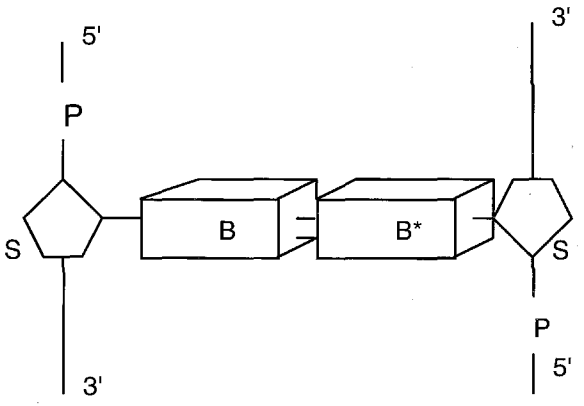

Fig. 3. Double-stranded DNA repeat unit.

1.3. Molecular structure

1.3.1. Chemical structure of DNA

The basic monomer unit—the polymer repeat unit—of double-stranded DNA is shown in Fig. 3.

The parts marked B and B* are large, planar organic groups consisting of one or two 5-atom *aromatic rings*. They resemble benzene and, like benzene, these groups do not dissolve very easily in water. The symbols B and B* stand either for the smaller single-ring cytosine and thymine (the "pyrimidines"), or the larger two-ring guanine and adenine (the "purines"). We will use the notation G, T, C, and A for short. The four groups all have the chemical character of a *base* (i.e., they are proton acceptors).

Not every combination of bases is permitted: in particular only B–B* pairs of purines and pyrimidines are possible. The Watson–Crick base pairing consists of combining A with T and G with C. An A–T pair is connected by two *hydrogen bonds* and a G–C pair by three hydrogen bonds, so they have a higher binding energy. Other purine–pyrimidine pairings (like G with T) are possible, but they do have a lower binding energy. The genetic code of an organism, or "genome", is simply a listing of the different base pairings along the DNA sequence of that organism. Note that if you know the sequence of bases of one strand, you always can reconstruct the other "complementary" strand, assuming that Watson–Crick base pairing is valid.

The bases are connected to sugar groups (indicated by S in the figure). Sugars have the general formula $(CH_2O)_n$ and usually are water soluble. The particular sugar of DNA belong to the group of *pentoses*, 5-atom sugar rings, and is known as *deoxyribose*. The deoxyriboses of two adjacent bases are connected together by tetrahedral *phosphate groups* (PO_4^-) to form together the sugar–phosphate "backbone". Adjacent sugar groups are separated by 6 Å. The backbone strands have a *directionality*: they start with a deoxyribose at the 3′ end and end with a phosphate at the 5′ end. The backbone has two important physical characteristics for our purposes: it is *highly flexible* and, in

water at room temperature, it is *highly charged*. The negative charge of the backbone is due to the fact that the phosphate groups in water at physiological acidity levels are fully dissociated. Charged molecular groups are usually soluble in water and the sugar–phosphate backbone is indeed highly soluble in water. The flexibility is due to the fact that the covalent P–O bonds can freely rotate around, so adjacent PO_4^- tetrahedra and ribose rings along the backbone can rotate around their joining axis. We can describe the backbone as a *charged, freely jointed chain*.

RNA molecules are similar to single-stranded DNA molecules with two differences. First, the base thymine is replaced by another base, uracil, and second the sugar group has an extra OH group and is called a *ribose*.

Hydrogen bonding and the hydrophobic force: Hydrogen bonding provides the binding mechanism between complementary bases. Hydrogen bonding plays in general a central role whenever macromolecules are dissolved in water. The hydrogen bond is an electrostatic bond with a positively charged proton from one molecular group associating with a negatively charged atom of another molecular group, usually an oxygen (O^-), carbon (C^-) or nitrogen (N^-) atom. The cohesion of water is due to hydrogen bonding between water molecules, with the proton of one water molecule binding to the oxygen of another water molecule. The characteristic energy scale of the hydrogen bond is of the order of the thermal energy k_BT, so it is a relatively weak bond. At room temperature, a thermally fluctuating network of hydrogen bonds connects the water molecules.

Molecules such as alcohol that are easy to dissolve in water are called "hydrophilic" while molecules, such as hydrocarbons, that are not soluble in water are called "hydrophobic" [5]. Hydrophobic molecules cannot be incorporated in the thermally fluctuating network of the hydrogen bonds. They are surrounded by a shell of water molecules that have a reduced entropy, since they have fewer potential partners for the formation of a hydrogen bonding network. As far as the water molecules are concerned, the surface of a large hydrophobic molecule resembles the air–water surface, which has a surface energy γ of about 70 dyn/cm. We thus can estimate the solvation free energy—the free energy cost of inserting a molecule in a solvent—as the surface area of the hydrophobic molecule times γ. If we wanted to dissolve a number of hydrophobic molecules, then we should reduce the total exposed surface area in order to minimize the free energy cost. This is done by collecting the hydrophobic molecules in dense clusters. This effect is known as the "*hydrophobic interaction*", though it obviously is not a pair-wise interaction between molecules. Ultimately, the clustering leads to phase separation, which you can observe when you try to mix oil with water. An important thermodynamic characteristic of the hydrophobic interaction is that it is predominantly entropic in nature.

1.3.2. Physical structure of DNA

The physical structure of double-stranded DNA is determined by the fact it is neither hydrophobic nor hydrophilic. It belongs to a special intermediate group called the "*amphiphiles*" that share properties from both classes: one part of DNA—the backbone —is hydrophilic and another part of DNA—the bases—is hydrophobic. This frustrated, amphiphilic character of DNA, plus the flexibility of the backbone, produces the

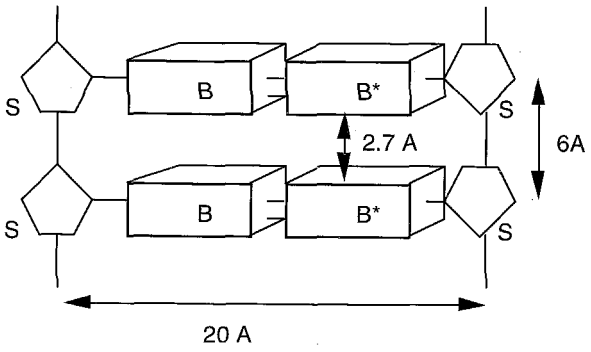

Fig. 4. Geometry of stretched DNA.

famous double-helical structure of DNA. To see how, imagine DNA stretched out like a (straight) ladder (see Fig. 4).

It turns out that the gaps between the rungs of the ladder, 2.7 Å, are wide enough to allow water molecules to slip in between the bases. Under the action of the hydrophobic force, the bases attract each other. The fixed 6 Å spacing between the sugar groups prevents a contraction of the ladder, but there is another way to bring the bases in contact. Imagine that we gradually *twist* the ladder, thereby forming a double spiral. This *is* possible because of the flexibility of the backbone. As we increase the twisting, the bases are brought into closer contact and the water molecules are squeezed out. For a twist angle T of about 32° between adjacent bases, the gap is completely closed. This produces the classical double helix shown in Fig. 5. The repeat length is $360/T$ bases, or about 11 bases. The repeat distance along the helix, or pitch, is about 35 Å (using the previous figure, compute T yourself).

The DNA double helix is thus held together by the hydrophobic attraction between bases, sometimes called the stacking interaction, and the hydrogen bonding between complementary bases. The double helix is not very stable. If you heat DNA, the two strands start to fall apart for temperatures in the range of $70–80^\circ$C. In addition, a number of different variants of the double helix can be realized by modest changes in the environmental conditions. Under conditions relevant for the life of cells, the dominant structure is the B form, a right-handed helix with a 24 Å diameter. Increasing the salt concentration somewhat weakens the electrostatic repulsion between the two backbones. A new structure, known as "A DNA", appears, with a smaller 18 Å diameter for the double helix and larger pitch of about 45 Å. This structural flexibility of DNA is actually essential for its function: in order for the genetic code to be "read" by RNA polymerase and other proteins, you must be able to "open-up" the double helix. Storing the genetic code in an overly rigid and stable storage device would be like having a library with no doors.

1.3.3. Chemical structure of proteins

There are 20 different monomer units, or "residues", that can be used to construct protein biopolymers. These are the naturally occurring *amino acids*. Amino acids have

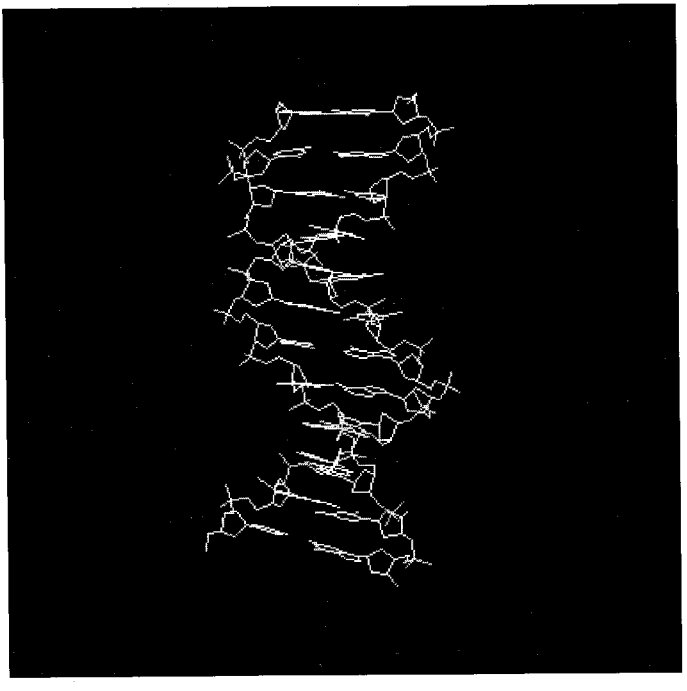

Fig. 5. B DNA.

the form of a tetrahedron with a carbon atom at the center, denoted by C_α. Recall that carbon has four electrons available to make chemical bonds. For the central C_α atom, these four electrons occupy four electronic orbitals (the "sp^4" orbitals) directed toward the vertices of a tetrahedron (as in diamond). At the four vertices are placed, respectively, a hydrogen atom, an NH_2 "amino" group, an acidic COOH "carboxyl" group, and finally one of 20 different side groups denoted by "R". There are two distinct ways of distributing the four molecular groups over the tetrahedron. The one shown in Fig. 6 is known as the *left-handed* or L form. If the amino and hydroxyl groups are exchanged, we obtain the right-handed or R form. In proteins, only the L form is encountered. The side group determines the chemical character of an amino acid. It can be neutral or charged, hydrophobic or hydrophilic, acidic or basic.

The simplest case is glycine, with R a hydrogen atom. There are two abbreviations for glycine: Gly and G. Each amino acid has two of these abbreviations (which biochemists know by heart). Lysine for instance is a positively charged amino acid with R equal to $(CH_2)_4NH_2^+$ having the abbreviations Lys and K.

Nature uses its own abbreviation for the 20 amino acids when it stores the information required to produce a protein. Three adjacent DNA bases code for one amino acid. For instance, the triplet AAA is the code for the amino acid "phenylalanine" while TTT is the code for lysine. We call such a triplet a "codon". You can construct $4^3=64$ different codons from such a triplet, more than enough for the 20 natural amino

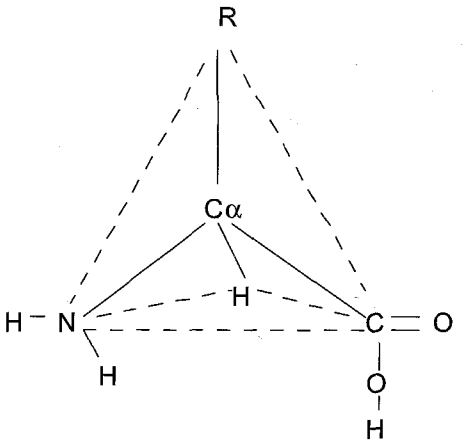

Fig. 6. Amino acid.

acids. Finding the complete set of codons of all the amino acids was one of the great landmark achievements of molecular biology.

To construct a protein, we must hook together these amino acids by a polymerization reaction. This takes place between the amino group of one amino acid and the hydroxyl group of another amino acid creating a covalent bond—known as a *peptide bond*—between a carbon and a nitrogen atom under release of a water molecule: ($NH_2 + COOH \rightarrow C\text{-}N + H_2O$). Two amino acid complexes are not rigid: the peptide bond allows considerable freedom of motion. When repeated over and over, this reaction produces a flexible string of amino acids—a polypeptide—that starts with an amino group, the so-called "N-terminal" and that ends with a hydroxyl group, the "C-terminal".

We saw that the base-pair sequence of the DNA of an organism is a code for the production of amino acid strings with three adjacent base pairs coding for one amino acid. The DNA sequence shown in Fig. 7 for instance will produce a simple polypeptide of four amino acids, starting with an N terminal and ending with a C terminal. (In the figure "Ala" stands for "alanine", a hydrophobic amino acid.) Note that only one of the two DNA strands is actually used for the production of proteins, the "coding" strand.

1.3.4. Physical structure of proteins

The physical structure of protein is determined by two physical mechanisms. On the one hand, proteins are again amphiphiles. Among the string of residues making up a protein, certain will have side chains that are hydrophobic, like Ala, and certain that are hydrophilic, like Lys. When dissolved in water, the string will try to fold up into a ball, with the hydrophobic residues hidden in the interior and the hydrophilic residues on the exterior surface. Such a ball is called a "globule", with a radius of the order of 2–3 nm.

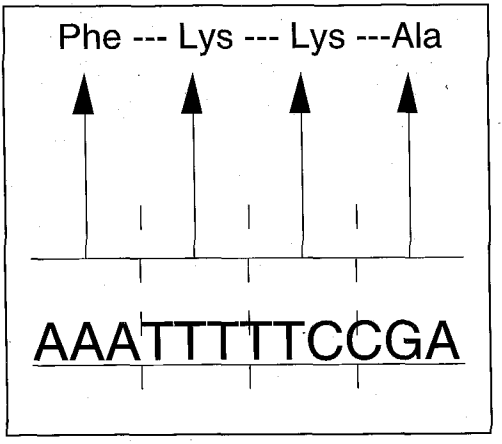

Fig. 7. Codons and amino acids.

The second important effect is the ability of amino acids to establish hydrogen bonds. The oxygen atom of the C=O group at one of the corners of the amino acid tetrahedron of one residue can act as a proton receptor for the C–H or N–H group of another residue. Linus Pauling first proposed that for a *helical* polypeptide string having the right pitch and diameter, known as the α *helix*, every residue can establish a hydrogen bond with a residue further up (or down) the helix. Not every residue of a protein "likes" to be part of an α helix because certain side groups may interfere with each other by steric hindering. Hydrogen bonding also can be used to link two straight polypeptide strands that run either parallel or anti-parallel, which is known as a β *sheet*.

The actual structure of a protein is determined by the combined effects of α-helix/β-sheet formation and the requirement to keep hydrophobic residues inside the protein interior. Drawings of protein structures show the α-helical regions as spirals and the β sheets as arrows. Fig. 8 is an example of a very simple protein with two α helices and a β sheet.

Note that not all amino acids are part of α helices or β sheets. What is remarkable about natural proteins compared with a random polypeptide chain is that over a certain range of temperatures, the minimum free energy state is a unique, folded structure with most of the atoms of the protein occupying well-defined positions. Only in this folded state can proteins act as molecular machines. This functional, folded state of a protein is not very stable. Formation of the folded structure involves a significant loss of entropy. Heating indeed unfolds, or "denatures" proteins. The "folding energy"—the difference between the properly folded state and the denatured state—is only of order $10k_BT$ or so. Moderately raising (or lowering!) temperature, changing salt concentration or acidity level can produces unfolding. However, it is precisely the fragility of folded proteins that allows them to adopt multiple configurations, which permits their use as switching devices, catalysts, and detection devices. For a molecule to act as a molecular machine, it must have "moving parts".

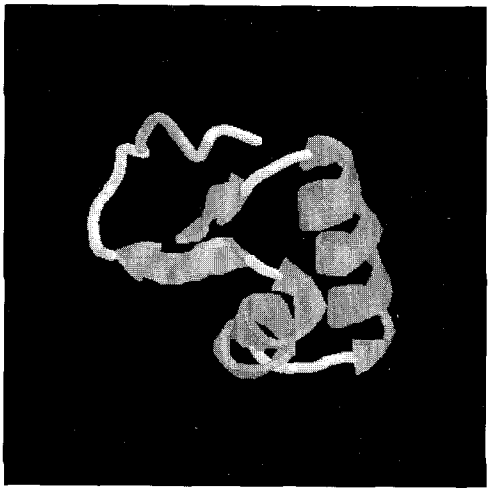

Fig. 8. A small protein with two α helices and a β sheet.

The following website has a nice tutorial on the chemical structure of DNA and proteins: http://www.clunet.edu/BioDev/omm/exhibits.htm#displays.

2. Thermodynamics and kinetics of repressor–DNA interaction

2.1. *Thermodynamics and the lac repressor*

The first branch of physics that we will bring to bear on the design of a repressor protein is thermodynamics/statistical mechanics. We will apply the principles of thermodynamics to understand how the *specificity* and *reversibility* requirements are met for the interaction between the lac repressor and DNA.

2.1.1. *The law of mass action*

Prepare an aqueous solution containing a certain low concentration of short, identical DNA strands and the repressor proteins. The base-pair sequence of the DNA strands may or may not contain the operator sequence. We can describe the reversible binding of the repressor to the DNA as an associative chemical reaction:

$$\mathrm{R} + \mathrm{DNA} \rightleftharpoons \mathrm{R|DNA}\,.$$

R | DNA stands for a repressor–DNA complex. The concentration of DNA strands with no repressor is indicated by [DNA], that of free repressor by [R], and that of the complexes by [R|DNA]. Concentrations can be measured by filtration methods and the results are expressed in "molar", or moles per liter (symbol M). Salt water has, for instance, a salt concentration of about 0.1 M while one molecule per micrometer^3 (volume of a bacterium) equals 10^{-9} M.

Thermodynamic processes inside cells normally take place under conditions of (nearly) fixed temperature and pressure. Under these conditions, the Gibbs free energy G must be minimized according to the second law of thermodynamics. The Gibbs free energy can be expressed as

$$G = N_{\mathrm{DNA}}\mu_{\mathrm{DNA}} + N_{\mathrm{R}}\mu_{\mathrm{R}} + N_{\mathrm{R|DNA}}\mu_{\mathrm{R|DNA}} . \tag{2.1}$$

Here, N and μ are the number of molecules and the chemical potential of each of the three species (actually, we also should add a term for the water molecules). At low concentrations, the chemical potential $\mu([\mathrm{C}])$ of "solute" molecules in a solvent like water has the following general form:

$$\mu([\mathrm{C}]) = k_{\mathrm{B}}T \ln([\mathrm{C}]v_{\mathrm{C}}) + \mu_{\mathrm{C}} . \tag{2.2}$$

The first term, with v_{C} the volume of the molecule, is very similar to the free energy per particle of an ideal gas and it is indeed due to the translational degrees of freedom of the solute particles. The second term, the "standard chemical potential", can be viewed as the intrinsic free energy per solute particle meaning that it depends on the type of solute molecule, and on temperature and pressure, but *not* on concentration.

Assume that there is a very small variation δN in the number of R|DNA complexes. According to the reaction scheme $\mathrm{R} + \mathrm{DNA} \rightleftharpoons \mathrm{R|DNA}$, there must be corresponding variation of $-\delta N$ in the number of uncomplexed DNA and repressor molecules. The change in G equals

$$\delta G = [\mu([\mathrm{R|DNA}]) - \mu([\mathrm{DNa}]) - \mu([\mathrm{R}])]\delta N . \tag{2.3}$$

The second law of thermodynamics demands that $\delta G=0$, so $\mu([\mathrm{R|DNA}])=\mu([\mathrm{DNA}])+\mu([\mathrm{R}])$. Using Eq. (2.2) and this condition gives

$$\frac{[\mathrm{R}][\mathrm{DNA}]}{[\mathrm{R|DNA}]} = \frac{1}{v}\exp^{-\Delta G_0/k_{\mathrm{B}}T} , \tag{2.4}$$

where we introduced the following two quantities:

$$\Delta G_0 = \mu_{\mathrm{R}} + \mu_{\mathrm{DNA}} - \mu_{\mathrm{R|DNA}} ,$$

$$v = \frac{v_{\mathrm{R}}v_{\mathrm{DNA}}}{v_{\mathrm{R|DNA}}} . \tag{2.5}$$

The energy scale ΔG_0 is called the "standard free energy change" of the reaction. At the intuitive level, you can think of it as the free energy gain when a repressor combines with a DNA strand, the *binding energy* in other words. The quantity v has dimensions of volume. You can think of it as the "*reaction volume*": if the repressor is located inside this volume, then it can bind to DNA.

Eq. (2.4) is a special case of a fundamental principle of chemical thermodynamics: the *law of mass action*. The law of mass action is such an important principle that the right-hand side of Eq. (2.4) has its own name and symbol: the *equilibrium constant* K_{eq},

$$K_{\mathrm{eq}} = \frac{1}{v}\exp^{-\Delta G_0/k_{\mathrm{B}}T} . \tag{2.6}$$

Equilibrium constants of associative reactions have dimensions of concentration, so they are expressed in molar. Using the law of mass action, the equilibrium constant can be obtained by measuring the concentrations, and hence the standard free energy change. The beauty is that we obtain this way an important microscopic quantity, the standard free energy change, by measuring purely macroscopic quantities.

When such an experiment is performed in a test tube ("in vitro") on a DNA/repressor solution [6], one finds that the result is very sensitive to the absence or presence of the operator sequence on the DNA:

$$K_{\mathrm{eq}} \approx \begin{cases} 10^{-10} & \text{operator DNA}, \\ 10^{-4} & \text{non-operator DNA}. \end{cases} \tag{2.7}$$

This large difference between the specific and non-specific equilibrium constants is the thermodynamic signature of the ability of repressor proteins to read DNA sequences.

We call the interaction between lac repressor and operator DNA the "specific" protein–DNA interaction and that with non-operator DNA the "non-specific" interaction. You might expect the equilibrium constant for the non-specific interaction to be independent of the DNA sequence but it actually can vary over two orders of magnitude when the non-operator sequence is varied. Later, this will turn out to be a quite important effect. From Eqs. (2.6) and (2.7), one finds that the standard free energy change for the operator case $\Delta G_0(\text{specific})$ is of the order of $20–25k_{\mathrm{B}}T$ while for the non-operator case $\Delta G_0(\text{non-specific})$ is of the order of $5–10k_{\mathrm{B}}T$.

What happens if the law of mass action is applied to conditions relevant to the crowded interior of *E. coli* (rather than test tubes)? The genome of *E. coli* contains about 10^7 base pairs (or bp) restricted to a volume of the order of $1\ \mu\mathrm{m}^3$. Let us approximate the *non-operator* part of the bacterial genome as a fairly concentrated solution of short (10 bp) DNA sequences having a concentration of the order of $10^6/\mu\mathrm{m}^3$ (of the order of 10 mM). First suppose that the lac repressors all are bound to lactose molecules, so they will not recognize the operator sequence. Let F be the fraction of unbound lac repressors. This "free fraction" can be related to the equilibrium constant through the law of mass action:

$$\begin{aligned} F &= \frac{[\mathrm{R}]}{[\mathrm{R}] + [\mathrm{R}|\mathrm{DNA}]} \\ &= \frac{[\mathrm{R}][\mathrm{DNA}]/[\mathrm{R}|\mathrm{DNA}]}{[\mathrm{R}][\mathrm{DNA}]/[\mathrm{R}|\mathrm{DNA}] + [\mathrm{DNA}]} \\ &= \frac{K_{\mathrm{eq}}}{K_{\mathrm{eq}} + [\mathrm{DNA}]}. \end{aligned} \tag{2.8}$$

The law of mass action was used in the last step. If we insert the measured value of the equilibrium constant (for the non-specific interaction) and our estimated value for [DNA], we find that F is of the order of 10^{-2} (the non-specific equilibrium constant K_{eq} is measured in the absence of lactose, so we are assuming here that lactose binding does not affect the non-specific interaction). That is an interesting result! Induced lac

repressors in *E. coli* still "live" most of the time on DNA even though they do not recognize the operator sequence.

Energy scales in molecular biochemistry: This $25k_BT$ value for ΔG_0(specific) is a typical energy scale for the complexation of biological macromolecules. On the one hand, this energy scale must be sufficiently *high* compared with the thermal energy scale k_BT, so thermal fluctuations do not break up the complex. On the other hand, the energy scale must be sufficiently low, so the binding is reversible and can be easily disrupted when required for the signaling process. In molecular biology, the universal "energy currency" for driving thermodynamically unfavorable processes is the hydrolysis of an ATP molecule: $ATP + H_2O \rightarrow ADP + Pi + H$, which delivers about $10k_BT$ in free energy. A $25k_BT$ value for the binding energy is thus quite reasonable. Protein complexes are in general maintained by multiple "weak bonds", such as the van der Waals attraction, hydrogen bonds, and the "polar" interaction (i.e., screened electrostatic interaction), all of the order of k_BT. Spatial patterns of these weak links provide a basis for highly specific "lock-and-key" type recognition between proteins. This must be contrasted with the covalent "strong bonds" (of the order of a hundred k_BT) that maintain the structural integrity of the macromolecules themselves.

2.1.2. *Statistical mechanics and operator occupancy*

Now assume that the lactose concentration has dropped, so the lac repressor proteins can bind to the operator sequence. Efficient design requires a high probability for the operator site to be occupied (to avoid unwanted gene transcription). We will compute the operator occupancy probability P using elementary statistical mechanics. Let there be M copies of the lac repressor distributed over N possible sites of the bacterial genome (with N, of the order of 10^7, large compared to M). We will neglect the small fraction of free repressors. There are then $A(N,M) = N(N-1)\cdots(N-M)$ ways to distribute the M proteins over the N non-operator sites and there are $C(N,M) = M[N(N-1)\cdots(N-(M-1))]$ ways of choosing one of the M proteins to occupy the operator site and distribute the remaining $M-1$ proteins over the non-operator sites, treating the proteins as classical, distinguishable objects. Let the Boltzmann factor of a protein occupying an operator site, respectively, a non-operator site, be $B_{s,ns} \equiv \exp^{+\Delta G_0(\text{specific,non-specific})/k_BT}$. The occupation probability is then

$$P = \frac{C(N,M)B_s(B_{ns})^{M-1}}{C(N,M)B_s(B_{ns})^{M-1} + A(N,M)(B_{ns})^M} \,. \tag{2.9}$$

This simplifies to

$$P = \frac{1}{1 + (N/M)\exp^{-\Delta\Delta G_0/k_BT}} \,. \tag{2.10}$$

The quantity $\Delta\Delta G_0 = \Delta G_0(\text{specific}) - \Delta G_0(\text{non-specific})$ is the difference between the specific and non-specific binding energies. [1]

[1] For an adsorbing surface exposed to a gas atmosphere, a similar relation between surface coverage and gas pressure is known as the "Langmuir isotherm".

When we put in the "numbers" for the binding energy obtained earlier, something interesting shows up: the very large number N and the very small number $\exp(-\Delta\Delta G_0/k_BT)$ nearly cancel each other ($N\exp(-\Delta\Delta G_0/k_BT)$ is about 10). Suppose we wanted to make sure that the operator is at least 99% of the time occupied. According to Eq. (2.10), that requires the number of copies M of the lac repressor to exceed 10^3. The actual number of lac repressors of an *E. coli* bacterium is maintained at a comparable value (about 10^2). There is thus a "design connection" between the values of the specific and non-specific binding energies on the one hand and the number of repressor copies maintained by the cell on the other hand. Simple statistical mechanics arguments provide us with an insight as to how the "working parameters" are set for bacterial gene expression. The most important lesson is that the value of quantities such as ΔG_0(specific), ΔG_0(non-specific), N and M must be understood in the light of the functioning of the bacterium as an *integrated system.*

What is puzzling at this stage is why we need the non-specific interaction in the first place. According to Eq. (2.10), if we turned off the non-specific interaction, we would only need about 10 repressor copies. We will return to that question in the discussion of the kinetics.

2.1.3. Entropy, enthalpy, and direct read-out

The Gibbs free energy is defined as

$$G = H - TS\,. \tag{2.11}$$

It is the sum of an "energetic" term: the enthalpy $H=E+PV$ (E is the internal energy) and an "entropic" term. The change in Gibbs free energy ΔG_0 that takes place when a repressor molecules binds to DNA can be obtained from the equilibrium constant. Can we obtain the separate enthalpic and entropic contributions as well? Under conditions of fixed pressure and temperature, the change in enthalpy equals

$$\Delta H = \Delta E + P\Delta V = \Delta Q \tag{2.12}$$

with ΔQ the heat released (which is why we call the enthalpy also the "heat function"). We call chemical reactions *exothermic* if $\Delta Q > 0$ and *endothermic* if $\Delta Q < 0$. Endothermic reactions are interesting because the driving mechanism is entropy increase rather than reduction of the potential energy of interaction between molecules. The heat released by a reaction can be measured by calorimetry, so the change in enthalpy can be found. Since the total change of the Gibbs free energy is known, we can also deduce the change in entropy.

When the enthalpic and entropic contributions ΔH and $-T\Delta S$ are determined in this manner for the interaction between the lac repressor and DNA, one finds the following results [7].

Specific interaction. The dominant contribution to ΔG_0 is entropic. As a function of temperature, $-T\Delta S$ decreases significantly with T. ΔH is negative, so the reaction is endothermic.

Non-specific interaction. The dominant contribution to ΔG_0 is again entropic, but $-T\Delta S$ now does not depend significantly on temperature. The enthalpic contribution is again negative.

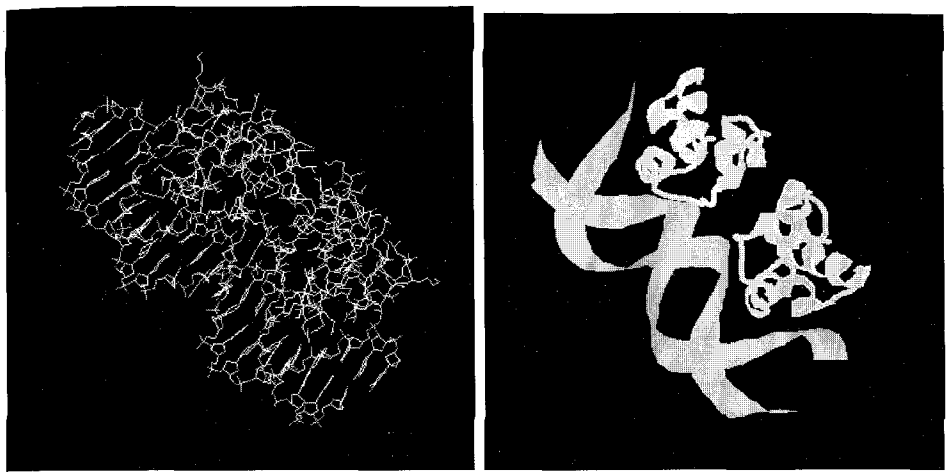

Fig. 9. Cro-repressor/DNA complex. First panel: Chemical bonds. Second panel: Cartoon showing reading heads.

Both are surprising results. To see why, we turn to the results of structure determinations of protein–DNA complexes. It is possible to grow crystals of repressor proteins complexed with short bit of DNA, known as "co-crystals". X-ray diffraction experiments on these crystals allow us to determine atomic positions with a resolution of 2 Å, and sometimes even better than that.[2] In Fig. 9 we show the result of such an experiment for the case of "cro", a very simple bacterial repressor (unlike the lac repressor).

The first panel in Fig. 9 shows the pattern of chemical bonds. Note the C_2 rotation symmetry. This symmetry is a characteristic of many procaryote repressor proteins. The DNA operator sequence has a corresponding (approximate) rotation symmetry. Simple repressor proteins like cro address the DNA with "reading heads". A reading head is an α helix that can be inserted into the major or minor groove of the DNA double helix (usually the major groove). The second panel is a cartoon of the cro repressor/DNA complex showing the α helices of the protein. There are two reading heads visible, one near the top and one near the bottom. The ends of certain side chains of the reading head can establish specific links with certain DNA bases. An example is the interaction between the amino acid arginine and the base guanine shown in Fig. 10.

The Arg side chain terminates with two N–H pairs. The two hydrogen atoms are positively charged and they "fit" exactly with negatively charged nitrogen and oxygen atoms of the guanine base. The nitrogen and oxygen atoms act as proton acceptors, so hydrogen bonds can be established, indicated in the figure by the two ovals. Base pairs are surrounded by a unique combination of proton donors and proton acceptors that can be read by specific amino acids. For instance, the amino acid glutamine "recognizes"

[2] High-resolution studies are done at synchrotron facilities such as at Argonne National Laboratory.

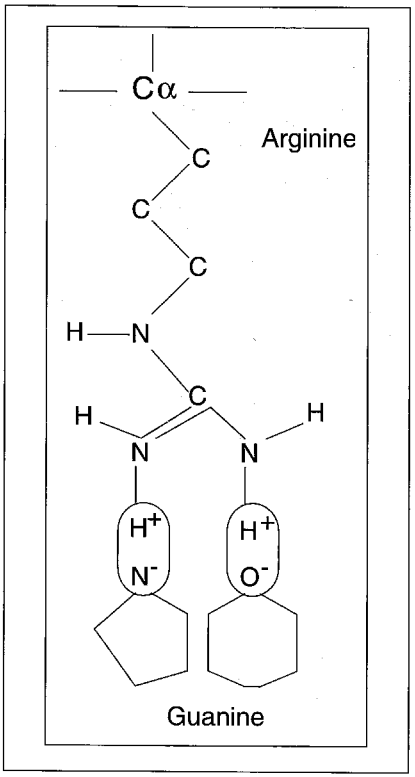

Fig. 10. Direct read-out.

an A–T pair in the major groove of DNA, just as arginine recognizes a G–C pair in the major groove, while asparagine recognizes a G–C pair in the minor groove.

We call this the "direct read-out" mechanism [8] and it is based on hydrogen bonding between amino acids and nucleic acids.

The second code: Molecular biologists have established long lists detailing contacts between the amino acids of DNA-binding proteins and DNA base pairs [9]. They hoped they could determine a "second code". By this they mean a one-to-one relation between amino acids and base pairs, so they could predict to which base-pair sequence a given repressor protein would bind. That would enable design of highly specific drugs turning on or off particular genes. Unfortunately, there appears to be no universal second code. DNA-associating proteins come in different design forms. The same amino acids interact differently in different types of proteins.

There is an obvious discrepancy between the direct read-out model and the results obtained from thermodynamics. If hydrogen bonding between the reading heads and DNA really was the dominant binding mechanism, then DNA/repressor binding should have been enthalpic in nature and formation of the complex would be associated with a loss of entropy. The puzzle is that there can be little doubt that

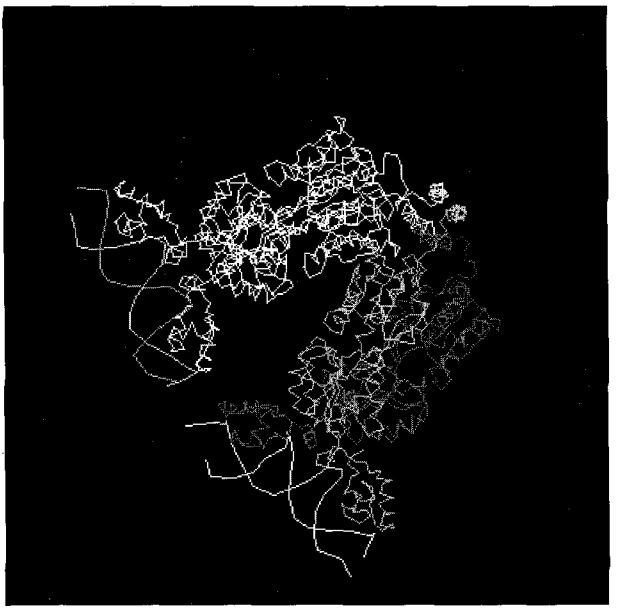

Fig. 11. The lac repressor complex.

direct-read out is an important mechanism for the reading of DNA sequence by proteins.

2.1.4. The lac repressor complex: a molecular machine

The resolution of this paradox comes from X-ray structural studies of lac repressor/DNA co-crystals [10] shown in Fig. 11.

The actual structure responsible for the repression of gene transcription is a complex consisting of two lac repressor protein dimers, so four copies in all. The four proteins are shown in Fig. 11. They bind pairwise to two separate operator sequences; note the four reading heads. The four reading heads are pairwise attached to the body of the complex by a linker unit that undergoes an *order–disorder transition* upon lactose binding. In the presence of lactose, the complex adopts a structure in which the linker unit is disordered and the reading heads cannot be inserted into the DNA major groove. Release of the lactose produces a ordering of the linkers and allows insertion of the reading heads into DNA. In addition, the transition brings two *hydrophobic surfaces*, belonging to the two dimers, into close contact. It seems reasonable to assume that if lactose-free repressor monomers or dimers move along non-operator DNA, locate the operator sequence, and form the full four-protein repressor complex, then the hydrophobic attraction plays a central role as well, so we can understand at least qualitatively why the specific binding of the lac repressor has an entropic character. The intervening DNA sequence between the two operator sequences loops around as shown in the next figure. Interestingly, another protein, known as CAP, binds to the

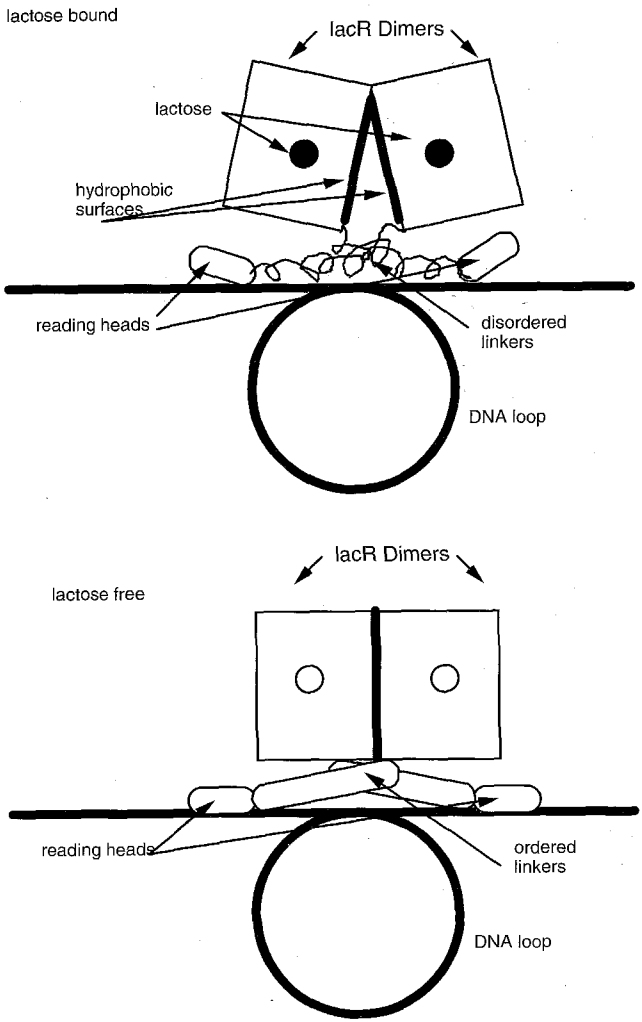

Fig. 12. Order–disorder transition of the lac repressor complex.

DNA sequence *inside* the loop. This stabilizes the loop but once the loop opens, it also stimulates gene expression! See Fig. 12.

The reading heads thus are only a small part of the lac repressor complex. We could view the complex as a molecular *detector* and *amplifier*. The binding of lactose to the repressor complex triggers a large structural transition that breaks up the complex and opens the loop. Release of the lactose closes the loop and restores the complex. Note that there is an analogy between the operation of the lac repressor complex and the molecular motors discussed in J. Howards' lectures where the binding/release of ATP and ADP also drives a cyclical structural transition that performs work.

At this point, you should go to the following website where you will find an elegant tutorial on the structural changes of the lac repressor tetramer and its interaction with DNA: http://www.worthpublishers.com/lehninger3D/index_title.html.

2.2. Kinetics of repressor–DNA interaction

We now turn to the third engineering requirement: *reactivity*. How quickly does a lac repressor respond to environmental changes, such as a reduction in lactose concentration? We start again with a discussion of in vitro experiments.

2.2.1. Reaction kinetics

The rate of change with time of the concentration of a repressor–DNA complex is the sum of two terms. A positive contribution due to complex formation between a previously unbound DNA molecule and a previously free repressor, and a negative contribution due to complex break-up. At sufficiently low concentrations, the first term must be proportional to the probability of finding a free DNA molecule and a free repressor molecule at the same site, and the second term must be proportional to the concentration of the complex:

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathrm{R}|\mathrm{DNA}] = k_{\mathrm{a}}[\mathrm{R}][\mathrm{DNA}] - k_{\mathrm{d}}[\mathrm{R}|\mathrm{DNA}]\,. \tag{2.13}$$

The proportionality constants k_{a} and k_{d} are called, respectively, the "*on-rate*" and the "*off-rate*". These constants are supposed not to depend on concentration though they can be quite strongly temperature dependent. The off-rate really does have dimensions of a rate but the (so-called) on-rate has dimensions of volume/time (chemists and biologists have a free-and-easy attitude to units). The on-rate and the off-rate have a surprising connection. Under conditions of thermodynamic equilibrium, the concentrations of the reactants obviously must be constant, so the left-hand side must be zero. That means that in equilibrium the following relation must hold:

$$\frac{[\mathrm{R}][\mathrm{DNA}]}{[\mathrm{R}|\mathrm{DNA}]} = \frac{k_{\mathrm{d}}}{k_{\mathrm{a}}}\,. \tag{2.14}$$

This is just the law of mass action, so the right-hand side must equal the equilibrium constant:

$$\frac{k_{\mathrm{d}}}{k_{\mathrm{a}}} = K_{\mathrm{eq}}\,. \tag{2.15}$$

Because the on-rate and off-rate do not depend on concentration, *this relation must hold also away from thermal equilibrium*! That means that we only need to determine one of the two rates, the other rate follows from Eq. (2.15). In vitro experiments on repressor–DNA solutions (containing the operator sequence) report that for the lac repressor k_{a} is of order $10^{10}\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$ under standard conditions.

Using this information, let us apply Eq. (2.13) to a colony of *E. coli* bacteria. Suppose that at times $t < 0$ there are no complexes because the environmental concentration of lactose is high. At time $t = 0$, the lactose concentration drops to zero. How long will it take the activated lac repressors to locate the operator sequence and switch-off

gene expression? There are only a few operator sequences per *E. coli*. Assuming a volume of 1 μm^3, the (initial) concentration of unoccupied operator sequences is of order $1/\mu m^3$ or about 10^{-9} M. According to Eq. (2.13), for early times t, the concentration of occupied operator sequences in the colony will grow linearly in time as

$$\frac{d}{dt}[R|DNA] \approx (k_a[DNA])[R] \tag{2.16}$$

keeping in mind that at $t = 0$, $[R|DNA] = 0$. It follows that we can identify $\tau = 1/(k_a[DNA])$ as the characteristic time scale for a free repressor to locate the operator sequence, the *switching time* in other words. For the measured value of k_a, this switching time is of order 0.1 s. This is a sensible result from the viewpoint of design: the actual switching time should be less than a minute or so for genetic switching to be a relevant response to a changing environment. Our estimate of the switching time must be viewed as a *lower bound*, because the cell environment is quite crowded. The actual on-rate inside a cell must be significantly less than this in vitro value. This means that the in vitro on-rate *must* be of order 10^{10} M^{-1} s^{-1} (or higher) for reasonable in vivo repressor reactivity.

2.2.2. Debye–Smoluchowski theory

Let us try to compute this on-rate. The classical theory of the on-rate of diffusion-limited chemical reactions is due to Debye and Smoluchowski (DS). Assume a spherical container (the cell) of radius R. Place the operator sequence at the center of the container. Let $C(\vec{r}, t)$ be the concentration of free repressors. The concentration field obeys the diffusion equation

$$\frac{\partial C}{\partial t} = D_3 \nabla^2 C \tag{2.17}$$

with D_3 the diffusion constant of the lac repressor in water. It is about 3×10^{-7} cm^2/s in vitro though under the crowded conditions of the bacterial interior, the effective diffusion constant is likely to be smaller.

We now want to know when the operator is occupied *for the first time* by a repressor. Assume that this will happen when a diffusing repressor enters for the first time a small sphere, of radius $b \ll R$, at the origin (b is a molecular length scale). You can view this sphere as the reaction volume of the law of mass action.

Remark: You can estimate the diffusion constant for proteins using the formula $D = k_B T/6\pi\eta r$ for the diffusion constant of a sphere of radius r of order a few nanometer in a fluid with viscosity η (for water $\eta = 10^{-2}$ P).

We actually will solve an easier problem by assuming that the small sphere at the origin acts as an *absorber*, so whenever a diffusing particle hits the small sphere, it disappears. The concentration at the outer radius R is kept at a fixed value $c(\bullet)$. This is an easier problem because under these conditions, a time-independent steady-state current I is established of repressor molecules diffusing from the outer to the inner sphere. To obtain this current, we must solve Laplace's law:

$$\nabla^2 c = 0 \tag{2.18}$$

with the boundary conditions $c(R) = c(\bullet)$ and $c(b) = 0$ (because diffusing particles disappear at $r = b$). The only solution of Laplace's law with spherical symmetry is the monopole field. Assuming $b \ll R$, and imposing the boundary conditions:

$$c(r) = c(\infty)\left[1 - \frac{b}{r}\right] . \tag{2.19}$$

The diffusion current density $\vec{J} = -D_3\vec{\nabla}c$ is along the radial inward direction, so the diffusion current I equals $J(r)$ times the surface area $4\pi r^2$:

$$I = -4\pi D_3 b c(\infty) . \tag{2.20}$$

Now compare this result with Eq. (2.13) for the case $k_d = 0$. The left-hand side of Eq. (2.13) is the number of complexes forming per second. That must equal (minus) the incoming current I. On the right-hand side we can identify $c(\bullet)$ with the repressor concentration [R] far from the operator. This leads to

$$k_a = 4\pi D_3 b \tag{2.21}$$

known as the Debye–Smoluchowski (DS) rate. If we use for the "target radius" b the typical size of a protein, say 4 nm, we find that on-rate is of order $10^9\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$.

It turns out that this is a "hard" upper bound. Actual on-rates are nearly always smaller than the DS rate because it takes a certain time for the protein to line up with the target. Associative reactions involving proteins able to achieve on-rates approaching a limiting value of $10^9\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$ are said to have reached "*kinetic perfection*". Now recall that for repressor/DNA association, an on-rate of $10^{10}\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$ was obtained, *an order of magnitude larger than kinetic perfection.* We also saw that this high on-rate was quite essential to support a reasonable response time of the bacterial gene-transcription system. If the bacterium had to make do with a typical protein–protein association on-rate it would be living a live on the razor's edge.

How does the lac repressor manage this phenomenally high rate? It was suggested by M. Eigen in the 1970s that the non-specific protein–DNA interaction may provide the answer. If inactive repressors are mostly located on the DNA, then diffusion is a predominantly *one-dimensional* process, not three dimensional as assumed in the DS theory. This ought to speed up the on-rate since less time is wasted searching empty space. Eigen's idea can be tested. If we could somehow reduce the non-specific repressor–DNA interaction, we should find that the on-rate decreases and approaches the DS value, since one-dimensional diffusion is replaced by three-dimensional diffusion. This is actually possible: increasing the *salt concentration* reduces the non-specific binding energy ΔG_0(non-specific) since this interaction is predominantly electrostatic. Experimentally, one finds the following dependence of the non-specific equilibrium constant on salt concentration:

$$-\log_{10} K_{eq}(\text{non-specific}) \cong -10\log_{10}[\mathrm{KCl}] - 2.5 . \tag{2.22}$$

It follows from the definition of the equilibrium constant that $-\log_{10}(vK_{eq}) = 0.43\Delta G_0/k_B T$, so the non-specific binding energy decreases monotonically with the salt concentration [KCl]. If Eigen's idea is correct, we would expect that the on-rate decreases monotonically as well. Actually, the dependence of the on-rate on [KCl] in laboratory

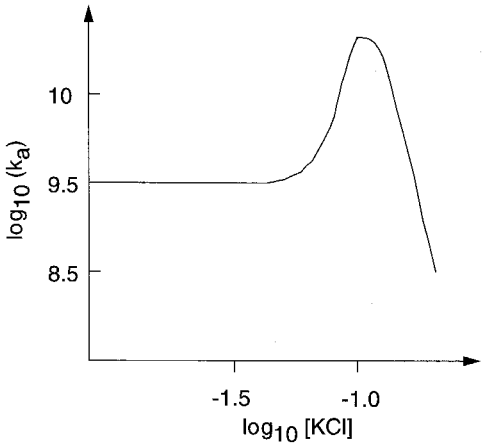

Fig. 13. Dependence of the on-rate on salt concentration.

experiments is highly non-monotonic. As shown in Fig. 13 there is sharp *maximum* for salt concentrations around 0.1 M (which happens to be the physiological salt concentration).

There are also functional objections against Eigen's idea. DNA does not really provide a nice one-dimensional "track" for a repressor. Under realistic conditions, the repressor will encounter many obstacles such as other repressor proteins bound to their respective operator sites or structural proteins that keep the DNA properly folded. These obstacles would quickly terminate a one-dimensional search. We could not hope to have free "runs" for one-dimensional diffusion of more than a few hundred bp.

2.2.3. *BWH theory*

Our present understanding of the on-rate for protein–DNA interaction is based on the work of Berg, Winter, and von Hippel [11] (BWH). Assume that at time $t = 0$ a repressor protein, located somewhere on the highly convoluted genome of an *E. coli* bacterium, is activated due to the release of its bound lactose molecule. How long will it take for the repressor to locate the operator site, assuming that there are no other repressors? Let $L(t)$ be the length of DNA searched by the repressor at time t. The characteristic time T^* for the protein to locate the operator sequence is obtained by equating $L(T^*)$ with the total length L_{tot} of the genome.

To find $L(t)$, recall that we learned earlier that a non-specifically bound repressor spends most of its time on the DNA, say 99%. *Most* of the time the repressor motion is thus restricted to the DNA. When, however, the repressor is released from the DNA, it starts a three-dimensional random walk—as in the classical DS theory—that terminates when the protein comes once again in contact with bacterial DNA. The key idea of the BWH theory is that even though the *Euclidean Distance* between the start and end of the three-dimensional random walk is likely to be short—since the interior of a bacterium is densely crowded with DNA—the *base-pair distance* is likely to be very large since the bacterial genome is highly convoluted (Fig. 14).

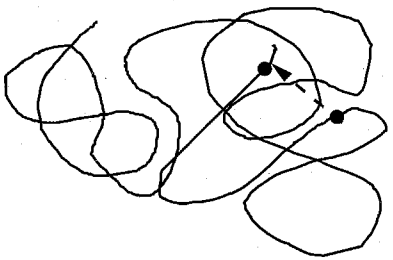

Fig. 14. Two points close in space but distant along the DNA.

This means that, as long as $L(t) \ll L_{\mathrm{tot}}$, it is very likely that the new section of DNA that will be explored by the repressor following a three-dimensional jump has not yet been explored after the repressor was activated.

Assume that at time t_0, the repressor is just starting a new one-dimensional random walk. At time t, it has explored a length of DNA equal to

$$\Delta L(t) \approx \sqrt{D_1(t - t_0)} \tag{2.23}$$

with D_1 the one-dimensional diffusion constant. Let k_{d}^* be the repressor dissociation rate for *non-operator* DNA, which can be measured experimentally, just as for the case of operator DNA. The typical duration of the one-dimensional random walk is thus about $1/k_{\mathrm{d}}^*$ seconds, so the length of DNA section searched equals

$$\Delta L \approx \sqrt{D_1/k_{\mathrm{d}}^*}\,. \tag{2.24}$$

Since repressors spend only a small fraction of their time away from DNA, the duration of the three-dimensional random walk must be short compared to that of the one-dimensional random walk. This means that after $T \gg 1/k_{\mathrm{d}}^*$ seconds, there have been of order $k_{\mathrm{d}}^* T$ one-dimensional random walks. The total DNA length visited equals ΔL times $k_{\mathrm{d}}^* T$ or

$$L(T) \approx T\sqrt{D_1 k_{\mathrm{d}}^*} \tag{2.25}$$

valid as long as $L(T)$ is less than L_{tot}.

That is a surprising prediction. Even though the molecule is performing a random walk, the length of searched DNA grows *linearly* in time. The total search time T^* required to locate the operator is found by equating $L(T^*)$ with the total length L_{tot}:

$$T^*(L_{\mathrm{tot}}) \approx \frac{L_{\mathrm{tot}}}{\sqrt{D_1 k_{\mathrm{d}}^*}} \tag{2.26}$$

so the search time is proportional to the length of the genome.

Let us put in the numbers. The one-dimensional diffusion constant for a sphere in water confined to a cylindrical surface with the appropriate dimensions is about 10^{-9} cm^2/s (considerably less than the three-dimensional diffusion constant). In vitro measurements of the non-specific dissociation rate show that k_{d}^* is quite sensitive to the salt concentration. In the physiological range, it is of the order of 10 s^{-1} for lac repressor. The distance ΔL of DNA searched per jump (Eq. (2.24)) is then of the

order of 1000 Å, and the total search time for a genome of 10 μm is about 10 s. For a purely one-dimensional search, the corresponding search time $T^*(L_{\rm tot}) \approx L_{\rm tot}^2/D_1$ would have been about 1000 s. Note that the search time would be proportional to the *square* of the total DNA length for purely one-dimensional diffusion.

These results are quite encouraging. The one-dimensional part of the search process extends only over stretches of the order of 1000 Å, a few hundred bp, as it should and the total *single*-protein search time of 10 s is reasonable. Keep in mind that there could be of order 100 copies of the repressor searching at the same time. "Mixed diffusion" works much better as a search strategy than either purely one-dimensional diffusion or three-dimensional diffusion. The "on-rate" can be calculated as a function of $k_{\rm d}^*$ and, using the measured dependence of $k_{\rm d}^*$ on salt concentration, one indeed finds that the on-rate of the lac repressor has a maximum around the physiological value of 0.1 M representing the cross-over from one-dimensional diffusion to three-dimensional diffusion.

We now begin to appreciate the biological role of the non-specific protein–DNA interaction: it significantly speeds up the search kinetics. Recall that when we discussed the equilibrium properties, the non-specific interaction only had a "nuisance value" since it required the bacterium to maintain an extra number of lac repressor copies to assure high operator occupancy. We can speculate that for bacteria the adaptive value of rapid genetic switching outweighs the metabolic cost of maintaining extra copies of the repressor.

2.2.4. *Indirect read-out and induced fit*

Apart from the direct read-out mechanism, there actually is a *second* mechanism enabling repressors to read the DNA sequence [12]. Recall that the non-specific equilibrium constant $K_{\rm eq}$ (non-specific) depends on the bp sequence: it can vary by two orders of magnitude if the bp sequence is varied. It turns out that the non-specific binding energy is maximized when the DNA sequence approaches the operator sequence.

How is the lac repressor able to identify the operator sequence, at least in a crude way, without addressing directly the bp's? The geometrical parameters characterizing the DNA double helix and the local *deformability* depend on the base-pair sequence. When a protein binds to DNA, the DNA structure is deformed. If you look carefully at the structure of the cro–DNA complex shown in Fig. 9, you will see that the DNA is *bent*. Transcription regulation proteins indeed usually induce a local bend or kink in the DNA structure [13]. As a result, certain sequences allow a better structural fit between the repressor and DNA than others (even if they do not contain the precise operator sequence). The idea that, apart from direct read-out, the local structural and elastic properties of the DNA operator sequence must present a good fit for the repressor is known as the "induced-fit" model [14].

What is the point of a second read-out mechanism? The indirect read-out mechanism is much less sensitive than direct read-out (for specific recognition, the equilibrium constant of the operator sequence is a factor 10^6 smaller than that of a random sequence while for the non-specific part the variation is only a factor 10^2). Consider how much time the lac repressor has available to make sure that it is or is not at the operator site. The lac repressor should be within a bp of the target site for the reading heads

to be able to swing in place. The time τ spent by lac repressor on one bp is of the order $\tau \approx a^2/D_1$ with a the distance between bp (say 3 Å) and D_1 the one-dimensional diffusion constant. This is of the order of a micro-second, taking our earlier value for D_1. Now recall that we know from the structural and thermodynamics studies that full recognition of the operator by lac repressor involves a significant structural change. The characteristic time scale for large structural changes of a protein is in the micro-to-milli second range. It does not look very likely that there is enough time to "test" each and every DNA site by continually swinging the reading heads in and out of position all the time.

We thus can speculate that the lac repressor is *slowed down*, by induced fitting, on DNA sequences that structurally resemble the operator sequence. The extra time available provides the opportunity for the full direct read-out mechanism to test whether the sequence actually is the operator sequence. If correct, this would mean making the engineering design of the lac repressor even more impressive.

Acknowledgements

I would like to thank the Altenberg Center for its hospitality. I have enormously benefited in my work in this area from interaction with numerous colleagues, in particular William Gelbart, Itamar Borukhov, Andrea Liu, Jay Mashl, Stella Park, Fyl Pincus, Joseph Rudnick, Cyrus Safinya, and Helmut Schiessel. The lecture notes formed part of an earlier course given in the 2001 Les Houches Summer School, and are reproduced with the kind permission of the editors.

References

[1] Alberts, et al., Molecular Biology of the Cell, Garland, New York, 1994.
[2] C. Calladine, H. Drew, Understanding DNA, 2nd Edition, Academic Press, New York, 1997.
[3] M. Ptashne, A Genetic Switch, Blackwell Scientific Publishing, Cambridge MA, 1992.
[4] F. Jacob, J. Monod, J. Mol. Biol. 3 (1967) 318.
[5] J.N. Israelachvili, Intermolecular and Surface Force, Academic Press Inc., San Diego, 1985.
[6] P. von Hippel, et al., Proc. Natl. Acad. Sci. 71 (1974) 4808.
[7] R. Spolar, T. Record, Science 263 (1994) 777.
[8] A. Travers, DNA–Protein Interactions, Chapman & Hall, London, 1994 (Chapter 3).
[9] D.M.J. Lilley (Ed.), DNA–Protein Structural Interactions, IRL Press, Oxford, 1995.
[10] Lewis et al., Science, 271 (1996) 1247.
[11] Berg, Winter, von Hippel, Biochemistry, 20 (1981) 6929.
[12] Z. Otwinowski, et al., Nature 355 (1988) 321.
[13] M. Werner, A. Gronenborn, M. Clore, Science 271 (1996) 778.
[14] R. Spolar, T. Record, Science 263 (1994) 777.