Making contacts on a nucleic acid polymer

Karsten Rippe

The interaction of proteins bound at distant sites on a nucleic acid chain plays an important role in many molecular biological processes. Contact between the proteins is established by looping of the intervening polymer, which can comprise either double- or single-stranded DNA or RNA, or interphase or metaphase chromatin. The effectiveness of this process, as well as the optimal separation distance, is highly dependent on the flexibility and conformation of the linker. This article reviews how the probability of looping-mediated interactions is calculated for different nucleic acid polymers. In addition, the application of the equations to the analysis of experimental data is illustrated.

Karsten Rippe
Deutsches Krebsforschungszentrum, Organisation komplexer Genome (H0700), Im Neuenheimer Feld 280; and Kirchhoff-Institut für Physik, Physik Molekularbiologischer Prozesse, Universität Heidelberg, Schröderstr. 90, D-69120 Heidelberg, Germany.
e-mail: Karsten.Rippe@dkfz.de

Binding of proteins to nucleic acids is an essential feature of central processes in molecular biology. Transcription, replication, recombination and DNA repair all involve many different protein–nucleic acid complexes. Sometimes, the components of these complexes have to interact with each other despite being located at distant binding sites. The formation of this type of contact requires the looping of the intervening nucleic acid chain. For example, the looping of dsDNA plays an important role in the regulation of prokaryotic and eukaryotic gene expression [1], facilitating the formation of contacts between transcription activator proteins bound to so-called enhancer sequences and other factors within the transcription machinery at the promoter.

Bringing protein factors together by looping depends on the flexibility and conformation of the nucleic acid linker involved. This process can be described quantitatively by applying polymer models and concepts developed several decades ago [2–4]. The probability of interaction between two proteins that are bound on the same polymer molecule is expressed by the local concentration j_M in moles per liter of one binding site in the proximity of the other. The value of j_M is equivalent to the concentration of one protein that would be required free in solution (*in trans*) to obtain the same contact frequency. A productive interaction between two proteins usually requires that their concentration is higher than the value of the dissociation constant K_d of the interaction. Thus, if the nucleic acid tether between the two proteins leads to $j_M > K_d$, associations are promoted that would not take place if the proteins were free in solution at a concentration below K_d. Typical protein concentrations in a eukaryotic nucleus are in the order of 10^{-7} to 10^{-9} M, with K_d values for specific interactions both in the same range and smaller. For two proteins, A and B, present at 1 nM, <1% of the proteins would associate into the AB complex according to the mass equation law (Eqn 1), if $K_d = 10^{-7}$ M is assumed.

$$A + B \rightleftarrows AB \qquad K_d = \frac{[A]\times[B]}{[AB]} \qquad [1]$$

However, if the two protein factors bind to the same interphase chromatin fiber at binding sites separated by 10 kb, this could lead to a local concentration of j_M as high as 6×10^{-7} M of protein A in the proximity of B as estimated below. This would lead to a shift of the equilibrium so that about two-thirds of each protein is in the AB complex.

Nucleic acid looping has been studied in detail with dsDNA chains. Much of the theoretical description of DNA looping has been derived for analyzing the DNA cyclization reaction, in which the single-stranded ends of a linear DNA fragment are linked into a circle [5–9]. However, interactions at a distance are not restricted to the looping of linear DNA molecules. For example, protein–protein contacts have been characterized that are promoted by a ssRNA linker [10–12]. Furthermore, the DNA in the eukaryotic cell is compacted by histone proteins into a chromatin fiber with a flexibility and length per base pair that is very different from that of histone-free dsDNA [13]. Here, it is reviewed how the local concentration, j_M, for the interaction of bound proteins, can be calculated from their binding-site separation distance, and from the specific polymer contour length and stiffness. The approximations used are applicable to flexible (relaxed) circles or linear nucleic acid chains, consisting of either double- or single-stranded DNA or RNA, or interphase or metaphase chromatin. The effect of the size of interacting proteins, as well as the presence of intrinsically curved regions within the chain, is also taken into account.

The following text does not distinguish between the contact of linear polymer ends such as, for example, in DNA cyclization, or the interaction of two separated sites on a longer linear or circular polymer [14], and the equations do not cover topologically constrained circles in which the polymer adopts a (super)helical structure. In addition, excluded volume effects that might be relevant for very long polymers are neglected. Finally, the requirement of a certain torsional orientation of the interacting sites to each other is not considered.

For the case of dsDNA, the corresponding twisting energy is given by a harmonic potential and can be calculated as described previously [6,7,9,15]. All of the above mentioned effects can be included if the local concentration is determined from Monte-Carlo or

 0968-0004/01/$ – see front matter PII: S0968-0004(01)01978-8

Brownian dynamics computer simulations that model the behavior of a specific polymer molecule [8,9,14,16–18]. However, these simulations require sophisticated custom-made computer programs and expert knowledge in polymer physics and/or chemistry. By contrast, many experimental data can be rationalized with sufficient accuracy by approximation formulas. The equations described below have been developed as simple tools to avoid complicated computer simulations. It is hoped that they will facilitate the quantitative analysis of molecular biological processes, in which looping of nucleic acids is involved.

Calculating the local concentration for circular and linear polymers

In many aspects, sufficiently long polymers behave similar to an idealized chain of n segments of length l, where the chain segments are not restricted in their torsional movement with respect to one another. Such a chain is termed a Gaussian or freely jointed chain (FJC) [2,19]. The parameter l is called the statistical segment length or Kuhn length after the Swiss scientist Werner Kuhn who developed the concept and much of the theoretical description of the FJC model in the 1930s [2]. The numerical value of l increases with the stiffness of the polymer. An equivalent parameter that is frequently used to describe the polymer stiffness is the persistence length a. The Kuhn length, and the persistence length, can be converted into each other according to the relation $l = 2a$. For the linear FJC the local molar concentration j_M of one end in the proximity of the other end can be calculated as reviewed in Ref. [1]. The resulting values for j_M are only accurate if the polymer length is >5–6 Kuhn segments. For shorter polymer linkers, an expression has been derived by Shimada and Yamakawa (SY) [7] using the Kratky–Porod (KP) model [3] that treats the polymer as an elastic rod or worm-like chain.

For looping of linear dsDNA, the FJC model and the SY approximation have been combined into a single equation that showed an excellent agreement with the measured frequency of site-specific recombination by FLP recombinase [20]. In a more general form, a relaxed circular polymer can also be described with the same approach when applying relations presented in Chapter five of Ref. [21]. This leads to Eqn 2 for calculating j_M between two sites that have a distance n on a circle of total size N.

$$j_M(n) = 0.53 \times \left(n - \frac{n^2}{N}\right)^{-3/2} \times \exp\left(\frac{d-2}{\left(n - \frac{n^2}{N}\right)^2 + d}\right) \times l^{-3} \frac{\text{mol nm}^3}{\text{liter}} \qquad [2]$$

The first part of the equation that includes the term to the power of –3/2 describes the FJC behavior. The second exponential term reduces the value of j_M at short distances so that the polymer behaves as predicted by the KP model. The local molar concentration j_M of one site in the proximity of the other site is given per l^3 in mol × liter^{-1} × nm^3. Inserting the value of l in nm for a specific polymer will yield a molar concentration.

In Eqn 2, the separation distance n between the two sites and the total size or length of the circle N are given in terms of the dimensionless-reduced separation distance or reduced contour length. This is simply the number of Kuhn segments with length l that corresponds to the site-separation distance or to the length of a given circle. Neither n nor N needs to be an integer. It should also be noted that it does not matter in which direction along the circle the separation distance n is determined; that is, whether the short or the long separation distance on the circle is measured (Fig. 1a). Equation 2, as well as Eqn 3, can be applied approximately in the range of $0.5 \leq n < \approx 100$, with $n < N$. The upper limit results from neglecting the volume occupied by the polymer, which can effect the value of j_M for very long chains. The magnitude of this effect will be dependent on the solution conditions. Below a separation distance of 0.5 Kuhn segments or one persistence length, the value of j_M becomes very small and deviations from the behavior predicted by the SY approximation increase. It is also important to realize that treating a nucleic acid chain as a homogenous polymer for $n < 0.5$ is questionable because the local structure can have large effects on the relation between j_M and n in this regime.

An additional parameter d is incorporated into Eqn 2, which reduces the contribution of the exponential term if $d > 0$. This leads to an increase of j_M at short separation distances ($n < 4$). The parameter d can be used to describe two effects that facilitate contacts at small values of n as illustrated in Fig. 1b and 1c. First, the interaction between two nucleic acid bound proteins or protein complexes occurs once the binding sites on the nucleic acid have approached each to the average diameter of the proteins. This is referred to as the reaction distance r (Fig. 1b). An average value of $r = 10$ nm is frequently assumed for the contact distance of two proteins [1,14,20]. By contrast, for chemical cross-linking of two sites, or for ligating the ends of a linear polymer, a much closer approach is required, and $r \approx 0$ nm (Fig. 1d). Second, intrinsically curved regions within the polymer can also significantly increase the interaction probability at short site-separation distances [1,9,14] (Fig. 1c). Both a reaction distance of $r > 0$ and polymer curvature can be accounted for by inserting appropriate values for d into Eqn 2. This is demonstrated later in the text for the looping of dsDNA [22]. If the interaction occurs at a reaction distance of $r \approx 0$ nm, and the nucleic acid polymer is homogenous; that is, without non-random curvature, a value of $d = 0$ is to be used.

Although Eqn 2 looks somewhat complicated, it has the advantage that it is not restricted to a specific polymer, and simpler expressions can be directly derived from it. For example, a linear chain can be

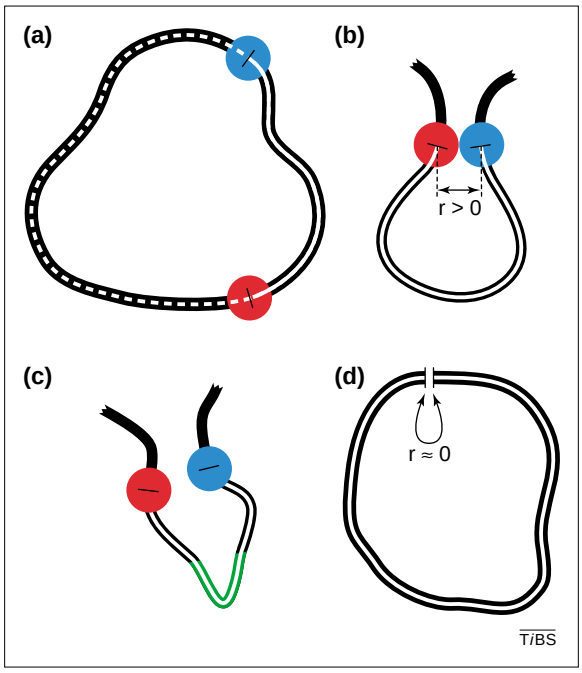

Fig. 1. Interactions mediated by nucleic acid looping. The polymer chain is represented by a black line and the site-separation distance n along the chain is drawn as a thinner white line. Interacting proteins are represented as red or blue spheres. (a) Circular polymer. For the use of Eqn 2, it makes no difference whether n is determined by measuring the short (continuous white line) or the long site-separation (dashed white line) distance. The total circle size or contour length is the sum of the continuous and the dashed white lines, and is referred to as N. Both n and N are expressed in Eqns 2 and 3 by the corresponding number of Kuhn segments with length l, which does not have to be an integer value. (b) Facilitated contacts that are relevant only at short separation distances ($n < 4$) can be described in Eqns 2 and 3 with the parameter d. Contact between the proteins occurs once the binding sites have approached each other to the average diameter of the proteins. Thus, the reaction distance $r > 0$. A value of $r = 10$ nm has been frequently used in the calculations of the local concentration j_M for protein–protein interactions [1,14,20]. (c) The presence of an intrinsically curved region within the polymer (highlighted in green) can increase the local concentration at short separation distances as well. This can also be accounted for by introducing an appropriate value for d as discussed in the text. (d) For the cyclization of a linear polymer, the two ends that are to be linked have to approach each other very closely. In this case $r \approx 0$, and as a consequence $d = 0$ if no curvature is present.

treated as a relaxed circle that is very large compared with the separation distance (i.e. $N >> n$), so that the n^2/N term approaches zero. The direction of the chain is not correlated between regions that are separated by a large distance. Thus, if the site separation is small when compared with the total circle size, the interacting sites will not 'feel' whether or not the adjacent regions of the chain are joined in a circle. In this case, one obtains Eqn 3 for a linear polymer, in which j_M is the local concentration for two sites separated by a distance of n Kuhn segments:

$$j_M(n) = 0.53 \times n^{-3/2} \times \exp\left(\frac{d-2}{n^2+d}\right) \times l^{-3} \frac{\text{mol nm}^3}{\text{liter}} \qquad [3]$$

The cyclization efficiency of a linear polymer as, for example, the ligation of dsDNA into a circle, can be described with Eqn 3 or similar expressions [5–9]. In this case, the reaction distance is $r \approx 0$ nm so that a value of $d = 0$ is inserted unless intrinsically curved regions are present.

Using an example of a circle with a total size of $N = 33$ Kuhn segments, Fig. 2 shows how the local concentration is dependent on the site-separation distance n according to Eqn 2 with $d = 0$. The function for the circular FJC chain according to Ref. [21] yields almost the same value of j_M within $6 \le n \le (N-6)$, whereas at shorter or larger separation distances, the values given by the FJC model are too high. For small n values, the circular chain behaves similar to a linear polymer that is given by Eqn 3 as discussed above. The maximum value for the local concentration is reached at a separation distance of $n \approx 1.7$ segments. This is the optimal separation distance for loop formation. Equation 2 cannot be applied if the molecule is in a topological constrained conformation so that the polymer adopts an interwound (super)helical structure as observed for DNA plasmids. In this case, the value of j_M also becomes dependent on the degree of under- or over-twisting, and can increase by more than an order of magnitude [16–18].

Applying the approximations to a specific polymer

The expressions in Eqns 2 and 3 are independent of the characteristics of a specific polymer. To calculate j_M for a certain circular or linear nucleic acid chain, the flexibility of the polymer as given by the Kuhn length l needs to be known. Values for l that have been determined experimentally are summarized in Table 1. In addition, the site-separation distance n, as well as the circle length N, have to be expressed by a corresponding number of nucleic acid monomer units, such as, for example, base pairs in dsDNA. Literature values for the length L_m per monomer unit of the chain are also given in Table 1. With L_m and l, the

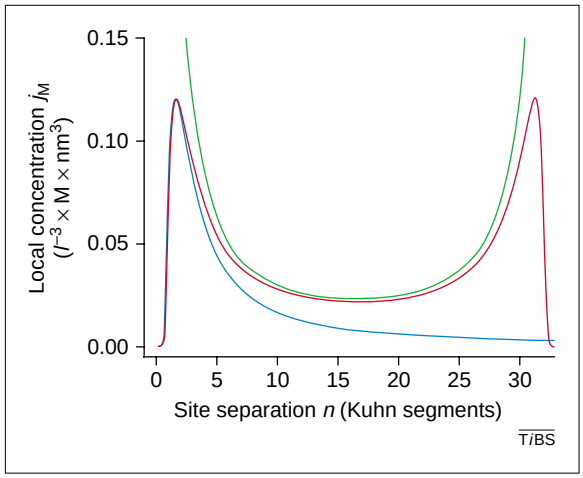

Fig. 2. Dependence of the local concentration j_M on the site separation distance given by the number of Kuhn segments n. Shown is the approximation given by Eqn 2 for a relaxed circle with a total contour length of $N = 33$ Kuhn segments (red) in comparison with the expression for the freely jointed chain model calculated according to Ref. [21] (green) and the behavior of a linear polymer (blue) as described by Eqn 3. The curves have been calculated with $d = 0$.

Table 1. Length and flexibility of nucleic acid polymers[a,b]

Nucleic acid chain	Length L_m of monomer unit	Kuhn length l (nm)	Monomers per Kuhn length	Refs
dsDNA	0.34 nm b^{-1}	100	290 b	[38]
dsRNA	0.27 nm b^{-1}	70–80	260–300 b	[39]
ssDNA	0.50–0.60 nm nt^{-1}	2–6	4–12 nt	[40–43]
Single-stranded poly(rU)	0.65 nm nt^{-1}	4	6 nt	[44]
Single chromatin fiber	8.60 nm kb^{-1}[c]	60[d]	7 kb	[36]
Chromatin fiber	9.60 nm kb^{-1}[c]	137–440[e]	14–46 kb	[33,34]
Metaphase chromosome	34.00 nm Mb^{-1}	300–5400[f]	9–160 Mb	[45,46]

[a]Abbreviations: b, base pair; nt, nucleotide; kb, kilobase pairs; Mb, megabase pairs.
[b]The contour length of the nucleic acid chain is described by the length L_m of one monomer unit. The flexibility or stiffness is given by the value of the Kuhn length, which is equal to two times the persistence length.
[c]The mass density of isolated 30 nm fibers from chicken erythrocytes was determined to be approximately six nucleosomes per 11 nm fiber in solution at physiological salt concentrations [35]. The values given for L_m were derived with this mass density and an average nucleosome repeat length of 212 base pairs for chicken erythrocytes and of ~190 base pairs as found in mammalian cell lines [13].
[d]Determined from the analysis of single chicken erythrocyte chromatin fibers *in vitro*.
[e]Determined by fluorescence *in situ* hybridization experiments with human fibroblast cells [32].
[f]From experiments in *Xenopus*, newt, grasshopper and *Drosophila* cells. The values are likely to reflect differences between the organisms and the method of analysis, as well as varying stages of mitosis.

separation distance n can be expressed in terms of the number of monomer units m according to Eqn 4:

$$n = \frac{m \times L_m}{l} \quad [4]$$

The same relation is applied to convert the circle size N. In this case, m gives the total number of monomers constituting the circle. These relations are then substituted into Eqn 2 or 3 to obtain the equations described in the following text.

Double-stranded DNA

Protein–protein contacts between distant binding sites are often mediated by looping of dsDNA (for reviews see Refs [1,15,23–28]). The theoretical framework outlined above has been used successfully for a quantitative analysis in various systems, including the interaction between lac repressor complexes [28,29], the *in vitro* and *in vivo* frequency of site-specific recombination by FLP recombinase [20], and transcription activation of *E. coli* RNA polymerase·σ^{54} holoenzyme by the enhancer-binding protein NtrC [22,30].

The general description given in Eqn 3 is adapted for the specific case of linear double-stranded B-DNA looping as follows: the monomer unit of B-DNA is one base pair with a length of $L_m = 0.34$ nm and the Kuhn length is $l = 100$ nm at physiological salt concentrations (Table 1). Thus, the contour length of the DNA linker between the two interacting sites is given by the number of base pairs b multiplied by 0.34 nm. According to Eqn 4, for a site-separation distance of n_{DNA} Kuhn segments, Eqn 5 is obtained:

$$n_{DNA} = \frac{b \times 0.34\ \text{nm}}{100\ \text{nm}} \quad [5]$$

Substituting n with this expression and l =100 nm in Eqn 3, yields Eqn 6 for linear dsDNA:

$$j_M(b) = 2.7 \times 10^{-3} \times b^{-3/2} \times \exp\left(\frac{d-2}{1.2 \times 10^{-5} \times b^2 + d}\right) \frac{\text{mol}}{\text{liter}} \quad [6]$$

From this equation, the local concentration j_M in mol × liter^{-1} as a function of the DNA linker length in base pairs can be calculated. For $d = 0$ ($r \approx 0$ nm) and $d = 0.13$ ($r = 10$ nm) expressions equivalent to those developed in Ref. [20] are derived. It is illustrated in Fig. 3 that the data from Brownian dynamics simulations [14] can be described with Eqn 6 using different values of d. For $d = 0$ the cyclization efficiency of a linear DNA fragment is obtained. With a value of $d = 0.13$ the interaction probability calculated with Eqn 6 is also in good agreement with the simulation data for a reaction radius of $r = 10$ nm as shown previously [20]. Furthermore, it is demonstrated that the computed j_M values for DNAs with a central kink of 30–120° can also be fitted with the appropriate values of d. Thus, the use of the parameter $d > 0$ in Eqns 2 and 3 modifies the equations so that a polymer conformation is described, in which contacts at short separation distances are facilitated as discussed above in the context of Fig. 1.

Single-stranded RNA

Various protein–protein contacts have been reported that are mediated by looping of a ssRNA chain (for examples see Refs [10–12]). One example refers to the mechanism by which RNA splicing enhancers operate. These RNA elements are usually located within 100 nucleotides of the 3′ splice position. They are thought to constitute binding sites for protein factors that interact with the general splicing machinery at the nearby intron [10,31]. An experimental analysis of the effect of varying the enhancer–intron distance revealed that the interaction was in good agreement with the predicted interaction probability between the two sites as expressed by the calculated j_M values [10].

Another well-studied case for RNA-looping is the process of 'antitermination' by the protein *N* from phage λ that binds to a specific hairpin structure

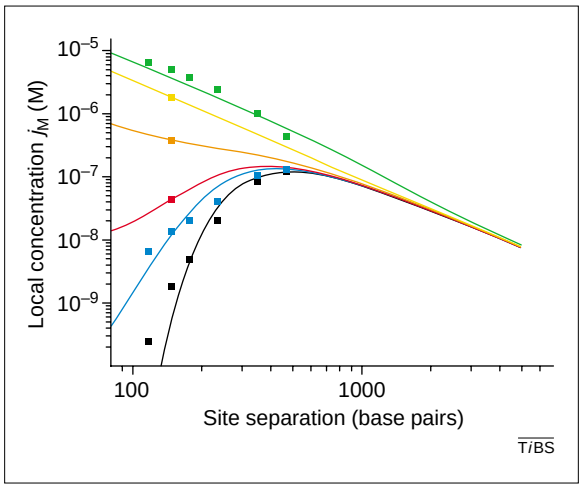

Fig. 3. Dependence of local concentration j_M on the site separation distance for dsDNA. The data points correspond to the Brownian dynamics simulations described in Ref. [14]. The curves are derived from a fit of these data to Eqn 6. The parameter *d* accounts for facilitated contacts at short separation distances. With the exception of the black data points/curve a reaction radius of $r = 10$ nm is assumed, which would represent the interaction of two proteins (or protein complexes) with a diameter of 10 nm (see Fig. 1b); black: no curvature, reaction radius $r = 0$ nm, $d = 0$; blue: no curvature, $d = 0.130$; red: 30° curvature, $d = 0.239$; orange: 60° curvature, $d = 0.652$; yellow: 90° curvature, $d = 2.610$; green: 120° curvature, $d = 21.70$. The curved DNA region was always located in the center between the two sites.

(boxB) within the *nut* site on the nascent RNA transcript. It interacts with the elongating *E. coli* RNA polymerase via looping of the intervening RNA, and induces a termination-resistant conformation of the polymerase (reviewed in Ref. [12]). *N*-dependent antitermination is observed in the absence of additional proteins when the boxB hairpin is located 100–200 nucleotides from the elongating RNA polymerase (estimated value of $j_M = 10^{-4}$ M). However, if the RNA linker is an order of magnitude longer (estimated value of $j_M = 10^{-6}$–10^{-7} M), *N* can no longer promote elongation through terminator sequences. This suggests a dissociation constant K_d for the interaction between *N* and the transcription complex in the order of 10^{-5} M just in between the two j_M values [12]. As discussed above in relation to Eqn 1, the value of j_M has to be larger than K_d if the equilibrium is to favor complex formation between the two bound proteins. For $K_d = 10^{-5}$ M, this will be the case if $j_M = 10^{-4}$ M, whereas with the longer RNA linker j_M would be much smaller than K_d.

The appropriate expression for calculating j_M for a ssRNA linker can be derived with $l = 4$ nm and $L_m = 0.65$ nm as the average internucleotide distance (Table 1). For $d = 0$, one obtains an expression from Eqn 3 for the looping of a linear RNA chain where *c* is the separation distance in nucleotides (Eqn 7):

$$j_M(c) = 0.13 \times c^{-3/2} \times \exp\left(\frac{-76}{c^2}\right) \frac{\text{mol}}{\text{liter}} \qquad [7]$$

It should be noted that the values given in Table 1 for RNA have been determined with poly(rU), which has little tendency to form secondary structures. By contrast, mixed mRNA sequences found *in vivo* usually form partial duplex regions and additional tertiary interactions under physiological conditions. These will affect both the apparent contour length per nucleotide L_m of the RNA chain as well as the average stiffness reflected by the value of *l*. So far, these parameters have not been determined for random RNA sequences so that the values derived from poly(rU) are used as an approximation.

An interesting feature of the *N* protein antitermination system is the observation that activation also occurs in the absence of the boxB-binding site under conditions where *N* protein can bind unspecifically to the RNA transcript [11,12]. What is the local concentration of *N* in this case? For an RNA transcript of length *t* there are approximately *t* binding sites, because every nucleotide can serve as the start of a new site. Assuming that only one *N* protein is bound per transcript, the average local concentration $\overline{j_M}(t)$ for the interaction between unspecifically bound *N* and the RNA polymerase can be estimated from Eqn 7 to be as shown in Eqn 8:

$$\overline{j_M}(t) \approx \frac{1}{t}\sum_{c=1}^{t} j_M(c) \quad \text{or} \quad \overline{j_M}(t) \approx \frac{1}{t}\int_1^t j_M(c)\,dc \qquad [8]$$

This function is plotted in Fig. 4 together with the curve for specific binding. It is evident that the unspecific binding of *N* protein leads to a rather high local concentration of *N* in the proximity of the RNA polymerase. The calculated values of $\overline{j_M}$ are well above the putative dissociation constant of 10^{-5} M for the interaction with RNA polymerase (see above and Ref. [12]), even with a transcript length in the order of 1000 nucleotides. Thus, the predicted interaction probabilities are in good agreement with the experimental observations showing that antitermination by *N* also occurs if the protein binds unspecifically to the RNA transcript.

It should be noted that the actual size of the interacting protein complexes can limit the maximum value of j_M that can be reached for single-stranded RNA or DNA linkers under optimal conditions; that is, separation distances of 10–20 nucleotides. For the permanent contact of two protein complexes with a diameter of 10 nm, one binding site would always be present in a 10 nm sphere around the other site (Fig. 1c). This corresponds to a local concentration of $j_M = 4 \times 10^{-4}$ M, and constitutes the upper limit for the value of j_M in this scenario. The theoretical maximum of j_M increases if the volume of closest approach becomes smaller and, for a 5 nm reaction radius, a value of $j_M = 3 \times 10^{-3}$ M is calculated.

Interphase chromatin fibers

In eukaryotes, the DNA is structured by histone proteins into a chain of nucleosomes, in which ~146 base pairs of DNA are wrapped around a histone octamer complex. This nucleosome chain associates under physiological conditions into a condensed fiber with a diameter of ~30 nm [13]. The 30 nm fiber adopts a complex and dynamic structure that is modified by a large number of protein complexes. To estimate the local concentration of one site in the proximity of another site for chromosomal DNA, the assumption of a simple free DNA linkage as described by Eqn 6 will, in general, not be correct. However, the relation between genomic distance and three-dimensional position of two sites on the same interphase chromosome can be quantitated at least

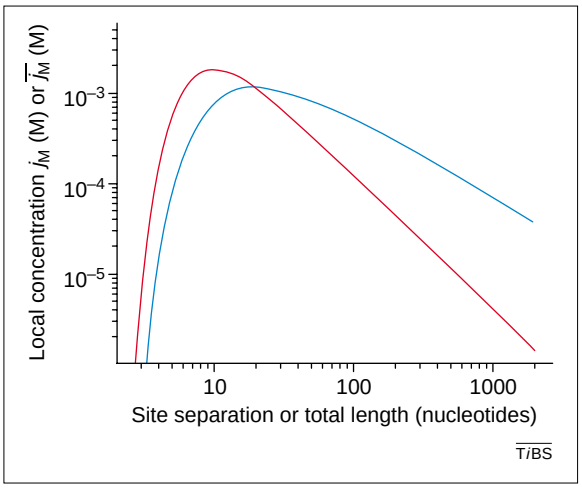

Fig. 4. Local concentration of interactions mediated by looping of a ssRNA linker with specifically or unspecifically bound proteins. The red line shows the local concentration j_M for specific RNA binding in dependence of the site separation distance in nucleotides, and has been calculated according to Eqn 7. The blue line is the plot of Eqn 8 and describes the average local concentration $\overline{j_M}$ for the interaction of a specifically bound protein at the beginning of the RNA and a second protein that is bound unspecifically to the RNA chain of the given length.

partly by modeling the 30 nm fiber as a polymer [32–34]. This can be done by using Eqns 2 and 3 with the appropriate values for the contour length L_m of the 30 nm fiber and its stiffness given by the Kuhn length l. Different measurements in solution agree that the fiber has a mass density of approximately six nucleosomes per 11 nm fiber at physiological salt concentrations (Ref. [35] and references therein). With this mass density, the corresponding contour length L_m per kb DNA will inversely depend on the nucleosome repeat length, which varies between organisms from ~165 base pairs in yeast to 212 base pairs in chicken erythrocytes [13]. If this is taken into account, L_m values of 11.1 nm kb^{-1} (yeast), 9.6 nm kb^{-1} (human fibroblast cells) and 8.6 nm kb^{-1} (chicken erythrocytes) are derived (Table 1).

By fluorescence *in situ* hybridization (FISH) experiments, the distance of dye-labeled probes in fixed human fibroblast cells was measured by confocal laser scanning microscopy and compared with the genomic distance [32]. Analysis of the data provided values of $l = 137$–440 nm [33] and $l = 196$–272 nm [34], with an average value of $l \approx 250$ nm from this type of work. The analysis of single chromatin fibers *in vitro*

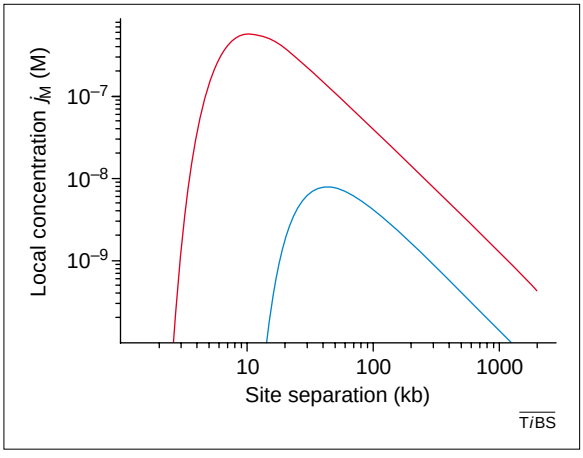

Fig. 5. Dependence of the local concentration j_M on the site-separation distance for linear chromatin fibers. The curves are calculated with a fiber contour length of 9.6 nm kb^{-1} for a Kuhn length of $l = 60$ nm (red) and of $l = 250$ nm (blue) according to Eqns 9 and 10.

yielded a higher fiber flexibility of $l = 60$ nm [36]. These rather large variations of l from 60 to 440 nm (Table 1) might reflect, in part, the technical difficulties in studying and analyzing the flexibility of chromatin fibers as well as 'real' differences in l. These could be a result of the source of the fiber (isolated fibers from chicken erythrocytes versus fixed human fibroblast cells) and the conditions of the experiments (e.g. ionic strength, fixation method). The value of $l = 60$ nm from single molecule experiments [36] was determined in a low salt buffer containing 10 mM Tris, 5 mM NaCl and 2 mM EDTA. Under these conditions, the chromatin fiber is likely to be in a somewhat extended and more open conformation. A statistical segment length of $l = 60$ nm might appear surprisingly small for a fiber that has a diameter of 30 nm. However, it should be noted that one Kuhn segment of 60 nm would consist of 7 kb DNA in a condensed chain of nucleosomes (Table 1). The stiffness of this fiber is likely to be determined by interactions between the nucleosomes rather than by the stiffness of the DNA, which would behave as a random coil or FJC polymer on this length scale (7 kb of free dsDNA would be equivalent to almost 24 Kuhn segments, Table 1). Thus, if the attractive forces between nucleosomes are relatively weak, the fiber will be soft and bend easily.

It is not clear at this point what value of l should be used for an accurate estimate of the interaction probability of two separated sites located on the same fiber. As an example, Eqn 9 ($l = 60$ nm) and Eqn 10 ($l = 250$ nm) have been derived from Eqn 3 with $d = 0$ for a linear fiber with $L_m = 9.6$ nm/kb. The site separation distance s is given in kilobases.

$$j_M(s) = 3.9 \times 10^{-5} \times s^{-3/2} \times \exp\left(-\frac{80}{s^2}\right) \frac{\text{mol}}{\text{liter}} \qquad [9]$$

$$j_M(s) = 4.6 \times 10^{-6} \times s^{-3/2} \times \exp\left(-\frac{1400}{s^2}\right) \frac{\text{mol}}{\text{liter}} \qquad [10]$$

As discussed previously (for example in Ref. [34]), the 30 nm chromatin fiber might be constrained to form circular domains. In this case, j_M should be calculated from Eqn 2 within these regions, if no additional torsional stress is present. The functions in Eqns 9 and 10 are plotted in Fig. 5. The different values of 60 and 250 nm for l have a large effect on the local concentration with a maximum j_M value of 6×10^{-7} M ($l = 60$ nm) as compared to 8×10^{-9} M ($l = 250$ nm). The same is true for the separation distance that is optimal for an interaction (10 kb versus 40 kb). This demonstrates that an accurate knowledge of the fiber flexibility and contour length is essential for a reliable estimate of the local concentration for contacts between distant sites on the 30 nm fiber.

Conclusions

The optimal separation distances for looping-mediated interactions with their respective j_M values are summarized for the various nucleic acid polymers

Table 2. Optimal separation distances for interaction by looping[a]

Nucleic acid	Separation distance	j_M (mol liter^{-1})
dsDNA	500 base pairs	1×10^{-7}
Relaxed 2.5 kb DNA circle	500 base pairs	1×10^{-7}
Superhelical 2.5 kb DNA circle[b]	200–2300 base pairs	5×10^{-6}
Superhelical 9.9 kb DNA circle[b]	200–500 base pairs	3×10^{-6}
dsRNA	460 base pairs	3×10^{-7}
ssDNA[c]	10–20 nucleotides	2×10^{-3}
Single-stranded poly rU[c]	15–20 nucleotides	1×10^{-3}
Chromatin fiber l = 60 nm	10 kb	6×10^{-7}
Chromatin fiber l = 250 nm	40 kb	8×10^{-9}
Drosophila metaphase chromosome[d]	15 Mb	5×10^{-9}

[a]Unless noted otherwise, the values were determined from the calculated maximum of the local concentration j_M at a separation distance of $n \approx 1.7$ Kuhn segments applying Eqns 2 and 3 with the values for L_m and l summarized in Table 1.
[b]Determined by Monte-Carlo simulations for a superhelical density of $\sigma = -0.05$ and a contact distance r = 10 nm [18]. Similar values have been determined by Vologodskii and co-workers as partly reviewed in Ref. [16].
[c]The real maximum value of j_M might be somewhat lower for protein–protein interactions because of excluded volume effects.
[d]Values calculated with L_m = 34 nm Mb^{-1} and l = 300 nm [46]. This is likely to constitute the upper limit of j_M, as the other reported values for l are much larger (Table 1).

in Table 2. It is evident that single-stranded DNA or RNA are much more effective in promoting protein–protein interactions ($j_M = 10^{-4}$ to 10^{-3} M, 10–20 nucleotides separation distance) than their double-stranded forms ($j_M \approx 10^{-7}$ M at ~500 base pairs). This is because of the highly increased flexibility of the single-stranded nucleic acids. In fact, double-stranded DNA and RNA are rather stiff polymers, which is reflected by the relatively high values of the Kuhn length being equivalent to ~300 base pairs (Table 1). As a consequence, an intrinsically curved region favors interactions on double-stranded DNA and RNA at short separation distances (<500 base pairs) [1,14,22]. Polymer curvature, as well as the reaction distance between the proteins, can be included in the equations for estimating the local concentration as demonstrated above for DNA (Fig. 3). Another possibility for facilitating contacts is the organization of the nucleic acid chain into a superhelical domain, which increases j_M by more than an order of magnitude for DNA, and is effective also at large separation distances (Table 2). At present, no simple analytical method exists to determine the interaction probability for circular superhelical polymers of a given statistical segment length, circle length and separation distance. In general, Monte-Carlo or Brownian dynamics simulation are required to determine j_M for this case [16–18].

When the DNA is condensed into an interphase chromatin fiber, the length scale for an optimal separation distance becomes completely different and interactions between sites are promoted that are separated by 10–40 kb. The exact value of the maximum of j_M will depend on the actual flexibility of the fiber and its contour length. So far, no 'consensus' literature data are available and the values are likely to differ between organisms. If the DNA is further compacted into a metaphase chromosome that, for example, contains 50–263 Mb of DNA in humans, interactions between distant sites become rather unlikely except for sites that are brought into proximity by the condensation process itself. This effect might also occur within the 30 nm chromatin fiber. For example, it has been shown that by wrapping the DNA around a nucleosome particle, short-range interactions between a 160 base pair separated enhancer and promoter in the gene encoding *Xenopus* vitellogenin B1 are facilitated [37]. Also, for the *in vivo* recombination frequency of FLP recombinase, a maximum ~200 base pairs separation distance was observed [20]. On such short separation distances, the interaction probability no longer reflects the flexibility of the chromatin fiber but rather its structural organization. It is likely to resemble more that of free dsDNA with an intrinsically curved region and a different apparent contour length. By contrast, for separation distances from 0.5–100 Kuhn segments, a quantitative agreement between experimental results and the calculated local concentration is observed in various systems with the approach described here. Thus, a simple polymer model can be applied successfully for the analysis of interactions that involve looping of very different nucleic acid linkers such as dsDNA, ssRNA and interphase chromatin fibers.

Acknowledgements
I thank John Schellman, Peter von Hippel, Konstantin Klenin, Job Dekker, Malte Wachsmuth and Jörg Langowski for helpful discussions and comments to the manuscript. The review was written at the division 'Biophysik der Makromoleküle' of the German Cancer Research Center. Financial support from the DFG (grant Ri-828/1) and the Volkswagen Foundation in the programme 'Junior Research Groups at German Universities' is gratefully acknowledged.

References

1 Rippe, K. *et al.* (1995) Action at a distance: DNA-looping and initiation of transcription. *Trends Biochem. Sci.* 20, 500–506
2 Kuhn, W. (1934) Über die Gestalt fadenförmiger Moleküle in Lösungen. *Koll. Z.* 68, 2–15
3 Kratky, O. and Porod, G. (1949) Röntgenuntersuchung gelöster Fadenmoleküle. *Rec. Trav. Chim.* 68, 1106–1113
4 Jacobson, H. and Stockmayer, W.H. (1950) Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* 18, 1600–1606
5 Wang, J.C. and Davidson, N. (1966) On the probability of ring closure of λ DNA. *J. Mol. Biol.* 19, 469–482
6 Shore, D. and Baldwin, R.L. (1983) Energetics of DNA twisting. I. Relation between twist and cyclization probability. *J. Mol. Biol.* 170, 957–981
7 Shimada, J. and Yamakawa, H. (1984) Ring-closure probabilities of twisted wormlike chains. Application to DNA. *Macromolecules* 17, 689–698
8 Hagerman, P.J. and Ramadevi, V.A. (1990) Application of the method of phage T4 DNA ligase catalyzed ring-closure to the study of DNA structure. I. Computational analysis. *J. Mol. Biol.* 212, 351–362
9 Crothers, D.M. *et al.* (1992) DNA bending, flexibility, and helical repeat by cyclization kinetics. *Methods Enzymol.* 212, 3–29
10 Graveley, B.R. *et al.* (1998) A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. *EMBO J.* 17, 6747–6756
11 Van Gilst, M.R. *et al.* (1997) Complexes of N antitermination protein of phage λ with specific and nonspecific RNA target sites on the nascent transcript. *Biochemistry* 36, 1514–1524
12 Van Gilst, M.R. and von Hippel, P.H. (2000) Quantitative dissection of transcriptional control system: N-dependent antitermination complex of phage λ as regulatory paradigm. *Methods Enzymol.* 323, 1–31
13 van Holde, K.E. (1989) *Chromatin*, Springer
14 Merlitz, H. *et al.* (1998) Looping dynamics of linear DNA molecules and the effect of DNA curvature: a study by Brownian dynamics simulation. *Biophys. J.* 74, 773–779
15 Wang, J.C. and Giaever, G.N. (1988) Action at a distance along a DNA. *Science* 240, 300–304
16 Vologodskii, A.V. and Cozzarelli, N.R. (1994) Conformational and thermodynamic properties of supercoiled DNA. *Annu. Rev. Biophys. Biomol. Struct.* 23, 609–643
17 Klenin, K.V. *et al.* (1995) Modulation of intramolecular interactions in superhelical DNA

by curved sequences: a Monte Carlo simulation study. *Biophys. J.* 68, 81–88
18 Klenin, K.V. and Langowski, J. (2001) Kinetics of intrachain reactions of supercoiled DNA: theory and numerical modeling. *J. Chem. Phys.* 114, 5049–5060
19 Flory, P.J. (1969) *Statistical mechanics of chain molecules*, Wiley
20 Ringrose, L. *et al.* (1999) Quantitative comparison of DNA looping *in vitro* and *in vivo*: chromatin increases effective DNA flexibility at short distances. *EMBO J.* 18, 6630–6641
21 Bloomfield, V.A. *et al.* (1974) *Physical chemistry of nucleic acids*, Harper & Row
22 Schulz, A. *et al.* (2000) The effect of the DNA conformation on the rate of NtrC activated transcription of *E. coli* RNA polymerase σ^{54} holoenzyme. *J. Mol. Biol.* 300, 709–725
23 Blackwood, E.M. and Kadonaga, J.T. (1998) Going the distance: a current view of enhancer action. *Science* 281, 61–63
24 Müller-Hill, B. (1998) The function of auxiliary operators. *Mol. Microbiol.* 29, 13–18
25 Schleif, R. (2000) Regulation of the L-arabinose operon of *Escherichia coli*. *Trends Genet.* 16, 559–565
26 Hochschild, A. (1990) Protein–protein interactions and DNA loop formation. In *DNA Topology and Its Biological Effects* (Cozzarelli, N.R. and Wang, J.C., eds), pp. 107–138, Cold Spring Harbor Laboratory Press
27 Zaman, Z. *et al.* (1998) Gene transcription by recruitment. *Cold Spring Harbor Symp. Quant. Biol.* 63, 167–171
28 Bellomy, G.R. and Record, M.T., Jr (1990) Stable DNA loops *in vivo* and *in vitro*: roles in gene regulation at a distance and in biophysical characterization of DNA. *Progr. Nucleic Acids Res. Mol. Biol.* 39, 81–128
29 Mossing, M.C. and Record, M.T., Jr (1986) Upstream operators enhance repression of the lac-promoter. *Science* 233, 889–892
30 Rippe, K. *et al.* (1997) Transcriptional activation via DNA-looping: visualization of intermediates in the activation pathway of *E. coli* RNA polymerase-σ 54 holoenzyme by scanning force microscopy. *J. Mol. Biol.* 270, 125–138
31 Blencowe, B.J. (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* 25, 106–110
32 van den Engh, G. *et al.* (1992) Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* 257, 1410–1412
33 Hahnfeldt, P. *et al.* (1993) Polymer models for interphase chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* 90, 7854–7858
34 Ostashevsky, J.Y. and Lange, C.S. (1994) The 30 nm chromatin fiber as a flexible polymer. *J. Biomol. Struct. Dyn.* 11, 813–820
35 Gerchman, S.E. and Ramakrishnan, V. (1987) Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proc. Natl. Acad. Sci. U. S. A.* 84, 7802–7806
36 Cui, Y. and Bustamante, C. (2000) Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl. Acad. Sci. U. S. A.* 97, 127–132
37 Schild, C. *et al.* (1993) A nucleosome-dependent static loop potentiates estrogen-regulated transcription from the *Xenopus* vitellogenin B1 promoter *in vitro*. *EMBO J.* 12, 423–433
38 Hagerman, P.J. (1988) Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.* 17, 265–286
39 Hagerman, P.J. (1997) Flexibility of RNA. *Annu. Rev. Biophys. Biomol. Struct.* 26, 139–156
40 Smith, S.B. *et al.* (1996) Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* 271, 795–799
41 Rivetti, C. *et al.* (1998) Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J. Mol. Biol.* 280, 41–59
42 Tinland, B. *et al.* (1997) Persistence length of single-stranded DNA. *Macromolecules* 30, 5763–5765
43 Mills, J.B. *et al.* (1999) Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of Poly(dT) and Poly(dA). *J. Mol. Biol.* 285, 245–257
44 Inners, L.D. and Felsenfeld, G. (1970) Conformation of polyuridylic acid in solution. *J. Mol. Biol.* 50, 373–389
45 Houchmandzadeh, B. and Dimitrov, S. (1999) Elasticity measurements show the existence of thin rigid cores inside mitotic chromosomes. *J. Cell Biol.* 145, 215–223
46 Marshall, W.F. *et al.* (2001) Chromosome elasticity and mitotic polar ejection force measured in living *Drosophila* embryos by four-dimensional microscopy-based motion analysis. *Curr. Biol.* 11, 569–578

Evolution of functional diversity in the cupin superfamily

Jim M. Dunwell, Alastair Culham, Carol E. Carter, Carlos R. Sosa-Aguirre and Peter W. Goodenough

The cupin superfamily of proteins is among the most functionally diverse of any described to date. It was named on the basis of the conserved β-barrel fold ('*cupa*' is the Latin term for a small barrel), and comprises both enzymatic and non-enzymatic members, which have either one or two cupin domains. Within the conserved tertiary structure, the variety of biochemical function is provided by minor variation of the residues in the active site and the identity of the bound metal ion. This review discusses the advantages of this particular scaffold and provides an evolutionary analysis of 18 different subclasses within the cupin superfamily.

As has been previously reported [1], members of protein superfamilies can often be detected through pairwise or multiple sequence alignments and, perhaps more convincingly, through similarities in their three-dimensional structures. This review takes the three-dimensional (3D) approach and summarizes the latest information on a recently identified superfamily of proteins, the cupins [2–4]. The cupins, along with the triosephosphate isomerase (TIM) barrel superfamily [5], have possibly the widest range of biochemical function of any superfamily described to date. These cupin functions, varying from isomerase and epimerase activities involved in the modification of cell wall carbohydrates in bacteria, to non-enzymatic storage proteins in plant seeds, and transcription factors linked to congenital baldness in mammals [6], are summarized in this article. More importantly, these functions are considered in relation to the three-dimensional structures of the individual proteins (Fig. 1).

The cupin superfamily was originally identified following the realization that the wheat protein

 0968-0004/01/$ – see front matter PII: S0968-0004(01)01981-8