C H A P T E R I

INTRODUCTION

1.01 Reassociation of DNA-overview

The complementary structure of the two strands of a DNA molecule is not a mere example of symmetry in nature. Though one cannot deny the aesthetic beauty of the double helical DNA, to a molecular biologist the complementary relation between the nucleotide pairs is important as it plays a pivotal role in primary processes of replication, transcription and translation. Matching of complementary nucleotide sequence is also involved in secondary processes, such as genetic recombination and repair, as well as in other events of recognition and control within the cell.

That dissociated complementary strands of DNA recognise each other and under suitable conditions do reassociate to form stable and biologically active hybrids has been a remarkable observation[1]. This has led to the building up of "reassociation kinetics", a technique which has now become a foremost tool in understanding sequence organisation, and genetic relatedness among various species.

Here we talk about the reassociation of DNA with special reference to eukaryotes. This is because it is here that reassociation has unravelled various information about sequences which were so far beyond the reach and untouched by any other physical, biochemical and genetic techniques.

The whole business of reassociation of eukaryotic DNA started when Britten and his colleagues faced the paradoxical result, that a part of eukaryotic DNA reassociated rapidly, and all the more was the striking observation that 10% of the DNA of mouse reassociated even faster than that of the smallest phages[2]. Now the haploid DNA content per cell in higher organisms is about thousand times more than in a bacteria. So it was logically assumed that the enormous dilution of the individual nucleotide sequence in eukaryotic DNA would make its reassociation so slow that months would be required for its completion at workable concentration. The only explanation to this queer observation was that certain sequences in DNA were reiterated again and again. Reiteration would make the concentration of these sequences much higher than normally expected and would thus account for the rapid reassociation.

This was in 1964, and in another four years Britten and his co-workers had investigated a wide range of eukaryotes (primates, other mammals, amphibians, bony and cartilageous fish, amphioxus, echinoderms, brachiopods, insects, other anthropods, mollusks, coelenterates, sponges, protozoans and plants ranging form algea to wheat) and conclusively found that the presence of repeated sequence was ubiquitous[3]. It was also found that the amount of repeated sequence and the frequency of reiteration had no direct relation with the complexity or the class of the species concerned. The amount

of repeated sequences varied from 20% to 80% while the frequency of reiteration could be in some cases over one million.

During the last decade the major thrust had been on the elucidation of sequence organisation in higher organisms. How are these repeated sequences arranged in different organisms? Most of the organisms tested, tend to follow a pattern known as the "Xenopus pattern"(4). Here the majority of the repeated sequences are short, about 300 nucleotides long and are interspersed by unique sequences of about thousand nucleotidesin length. The other pattern which has also been observed but not so common is the "Drosophilla pattern"(5). Here the repeated sequences mainly occur in long stretches and are not interspersed by unique sequences.

One question that has been asked from the beginning was about role of repeated sequences. The relatively precise pattern of interspersion of repeated and non-repeated sequences, the similarity in pattern obtained from an echinoderm, an amphibian and a mammal suggested that at the root of this common order, there should be a common cause. Britten, Davidson and their collaborators have proposed models which ascribe regulatory functions to middle repetitive DNA and evolutionary advantage in terms of adaptability to quantitative and qualitative changes in the middle repetitive DNA content observed even between closely related species(6-9).

Recently the very idea of searching roles for every portion of the DNA have been criticized[10,11]. This has been mainly done from an epistemological point of view. Based on the fact that the immediate role of a major part of the genome in cellular processes has yet not been found, it has been argued that the search for such roles (direct or indirect) for a major portion of the genome may prove if not intellectually sterile, ultimately futile. This is because, this DNA, which has been christened by the catch phrase 'selfish DNA', does not play any role in cellular processes, but just site idle, its only "motive" being self preservation. Prokaryotic transposable elements and eukaryotic middle repetitive sequence have been labelled as this kind of DNA. Though one cannot deny the immediate novelty of such an idea, as it explains such stumbling blocks as the C-value paradox (the relationship between genetic content and physical content of DNA), the role of a sequence reiterated thousands of times, the large difference of DNA content per cell in closely related species - the idea of useless pieces of DNA surviving the onslaught of evolution is too anarchic. Though one is not at present sufficiently armed to enter seriously into this controversy, it could modestly be said that as more and more positive roles of the repeated sequences are being uncovered[12], the idea that bulk of this type of DNA is junk may not find much ground.

As with any discovery the advent of repeated sequences have ushered in a new era. The technique of reassociation

kinetics first substantiated by Britten and Kohne[13] to detect the presence of repeated sequence has since then become a household tool of molecular biology. It has now been shown that this method can also be used in advantage for a more detailed study of the structure of the genome, for example - the mutual arrangement of repeated and unique sequences[14] and of inverted repeats[15], to detect injuries, mainly base modifications[16] and to determine the presence of short thermolabile segments of the same kind[17]. Comparative investigation of reassociation kinetics of normal and tumoral cell[18,19] together with the experiments on RNA hybridization[20,21] have yielded results which show that there is an increase in the content of middle repetitive sequence in the case of tumoral cells.

1.02 Experimental background of DNA reassociation

It is a well known fact that separated complementary strands of DNA recognise each other and reassociate under appropriate conditions. Reassociation is generally measured in two ways, each depending on some easily detectable difference between the single and double stranded DNA.

(a) Single strand DNA absorbs more ultraviolet (UV) radiation (i.e. O.D. at 260 n.m.) than double stranded DNA. So when single strands reassociate to form a hydrogen bonded duplex there is a fall in UV absorption. Thus one can measure the

changes in UV absorption with a spectrophotometer and monitor the amount of reassociation.

(b) Hydroxyapatite (HAP) crystals, a crystalline form of calcium phosphate, have different affinities for single and double stranded DNA. Thus reassociation can be followed by passing reassociated samples through HAP columns and estimating the amount of double stranded DNA bound to the column by first washing out the single strand DNA by an appropriate phosphate buffer (0.12 M, pH - 6.8 at 60°C) and then eluting out the bound duplex by higher molarity (0.48 M) or by raising the column to a high temperature (95-99°C) and eluting out with 0.12 M phosphate buffer (P.b.).

Condition of reassociation

The conditions of reassociation that have been used routinely by various laboratories are the ones originally standardized by Britten and his co-workers. The condition is 0.12 M P.b. (pH 6.8) and 60°C. This criteria has emerged out of an optimisation regarding the choice of various experimental parameters. Firstly it would be advantageous if the molarity of incubating buffer would be such that single stranded DNA could be eluted in the same buffer in one passage through the HAP column. It was also known that reassociation was maximum at a temperature of 20-30°C below the melting temperature[22]. Thus at 0.12 M P.b. reassociation of most

of the DNA types (GC content about 40-50%) should be optimum at about 60°C. Now it is at this temperature that single strand DNA fragments are eluted at the said molarity. If one goes in for a higher molarity of the incubating buffer (which is done in special cases of incubating at high C_0t value) then one has to, before passing the sample through the HAP column dilute it to an appropriate molarity in order to ensure efficient double strand binding to HAP. Incubation at high molarity should go with high temperature, thus changing the reassociation criteria which makes it difficult to compare various results. This is because we are now entering the field of thermal stability. Due to the sequence divergence, the reassociated duplexes have imprecise pairing i.e. not all bases are hydrogen bonded in the Watson-Crick format. This leads to a lowering of the thermal stability. Thus the incubating temperature becomes extremely important. At low incubating temperature complementary strands with a large degree of mismatch are selected as matched duplex whereas at high temperature precise matching is required. Thus one could vary the temperature of incubation to select duplexes of various thermal stability[23]. The conditions of 60°C and 0.12 M P.b. thus seems to be a golden mean.

Fragmentation of DNA

The presence of repeated sequence[5] was observed only when the DNA had been fragmented to a suitable length. That even after a decade, the DNA is fragmented to a length of 400-500 nucleotides for reassociation experiments as was originally done may be a case of chance coincidence. Before the inception of the idea of interspersion - that repeated segments are in fact about 300 nucleotides long as observed by Electron Microscope[24] or after digesting the single strand tails of the reassociated duplex by S_1 nuclease[25] and measuring its length - reassociation of eurakyotic DNA was done by fragmenting the DNA to 400-500 nucleotides. Now it is very well justified that this should be the length of the fragment if one has to quantitate the content, class and frequency of repeated sequence[26]. However, fragmentation to higher length is necessary when one is dealing with a Drosophilla pattern or when sequence interspersion is being estimated.

The conception of C_0t

The term C_0t is now a well established word in scientific lexicon. The design of most reassociation experiments is strongly influenced by the time required to complete the process. The reassociation of a pair of complementary sequences is a result of bimolecular collision and hence also depends on the concentration of the single strands. Thus the

product of DNA concentration and the time of incubation is the controlling parameters for estimating the state of the reaction. For convenience and simplification of language $C_o t$ is an useful parameter which denotes the product of concentration of nucleic acids (C_o) and the time of incubation (t). The units are second times moles of nucleotide per liter. A $C_o t$ of 1 mole x second/liter results if DNA is incubated for one hour at a concentration of 83 mg/ml which corresponds to an optical density of about 2.0 at 260 n.m.[3]

1.03 The reassociation kinetics of DNA

During the study of a reaction, two principal questions which comes into our mind are, how far the reaction will go and, how fast it will attain that stage? The answer to the first question deals with the equilibrium study and the answer to the second question is treated in the kinetic study of the reaction. Here we will be discussing about the kinetics of reassociation and its various ramifications.

Studies of hybridization and renaturation of nucleic acids dates back to the early sixties. Pioneering work in this field was by Ross and Sturtevant[27] on the formation of helices with RNA homopolymers, Marmur and Doty[28] on formation of reassociated DNA duplexes, and Nygaard and Hall[29] on the kinetics of DNA-RNA hybridization. A few years later Britten

and Kohne[3], and Wetmur and Davidson[30] systematized the approach, thus facilitating the interpretation of experimental results in a more coordinated fashion. This paved the way of an on flow research papers, culminating in its widespread use in molecular biology.

In the kinetic study the common practice is to correlate the rate of change in substrate concentration with either the concentration of the product or the substrate. Such a correlation when mathematically formulated is called the rate equation of the reaction. This forms the principal basis of the kinetic study.

The rate equation in case of DNA reassociation is dependent on the technique by which reassociation is monitored. The methodology initiated by Wetmur and Davidson describes reassociation as monitored optically. While the description by Britten et al holds for reassociation as monitored by HAP chromatography. The basic difference in these two techniques lies in the fact that while the former measures the fraction of total nucleotides that are paired, the latter measures the concentration of strand pairs that contain duplex regions. The implicit assumption that these two parameters are different is true because due to random fragmentation the majority of the reassociated duplexes contain single strand tails.

Monitoring reassociation by HAP chromatography has some practical advantages over optical reassociation :

1) At very low C_ot ($10^{-4} \sim 10^{-2}$) optical measurement cannot be done at standard criterion (60°C and 0.12 M P.b). This is because use of low concentration of DNA makes measurement insensitive while low time is not practically feasible.

2) At very high C_ot($10^2 \sim 10^4$) optical monitoring becomes cumbersome. On the one hand one has to handle high concentration of DNA, while in the other keeping the spectrophotometer idle for a long range of time is uneconomical. This feature along with the reproducible binding of DNA to HAP in a wide range of concentration (10^{-4} µg to 2 x 10^3 µg) has popularised the use of HAP chromatography as the principal tool for monitoring reassociation.

There are also some difficulties of interpretation of HAP based measurement.

1) Short repeated segment being unstable at the reassociation condition will go undetected. In bacteria and viruses operater site and sequences in promoter responsible for the formation of tight binary complex with RNA polymerase are about 20 bases or less[30].

2) Unreassociated sequences contiguous to the reassociated ones will be retained by HAP as duplex reassociated DNA, thus altering the shape of the kinetic curve. This will lead

to an overestimate to the amount of repeated DNA. This effect is stricking and is a function of fragment length (31-33).

Before discussing the rate equation for reassociation of DNA it will not be out of place to mention that order of a chemical reaction is the number of powers of the concentration of the substrate, that must appear in describing the rate of change of their concentration (or the formation of the products as a function of their concentrations). Since the process of reassociation involves two complementary strands of nucleic acids, one would expect it to be a second order process in terms of the concentration of the unpaired nucleotides. It is implicit that as reassociation is second order, it is limited by a slower step which may be preceeded or succeeded by one or more faster steps. It can be intuitively argued on the basis of DNA structure that the rate limiting step is the event of nucleation, i.e. the formation of one or few correct base pairs at some in register point along the two strands. The subsequent faster process involves the zippering reaction in which the rest of the base pairs are sequentially formed.

The rate equation

The general case of the reaction between two complementary strands A and B to form a helical duplex H is considered first

$$A + B \longrightarrow H$$

Since the concentration of A equal to B in DNA reassociation, the rate of the disappearance of single strand nucleotides could be written as

$$\frac{dC}{dt} = - KC^2 \qquad (1)$$

where C = total concentration of nucleotides in single strands and K is the rate constant.

Integration of equation (1) with the condition that at t = 0 C = C_o we have

$$\frac{1}{C} - \frac{1}{C_o} = Kt \qquad (2)$$

Equation 2 can be arranged in two ways each of which are followed in the reassociation study. The preference for a particular way of representation depends on the type of experiment concerned.

(1) We could write

$$\frac{C_o}{C} = \frac{1}{f} = K C_o t + 1$$

A plot of $\frac{1}{f}$ (where f is the fraction of nucleic acid in the duplex form) against t or $C_o t$ should give a straight line whose slope gives K value.

(2) In an alternate form

$$f = \frac{1}{1 + K C_o t}$$

A semi log plot of f against C_0t should give a sigmoid curve. This method evolved by the Britten and Kohne is of more practical importance. It has the advantage of presenting eukaryotic reassociation (eukaryotic reassociation is a multicomponent second order reaction) data which runs for several C_0t decades (generally from 10^{-4} to 10^4). It is evident that the rate constant can easily be evaluated from the C_0t value at which $f = 0.5$ i.e. $K = C_0t_{1/2}$.

Reassociation in higher organism

Reassociation is higher organisms is complicated by the presence of repeated sequences. Due to a spectrum in the degree of repetition, various fragments reassociate with rates that are proportional to the frequency of repetition. That is sequences having high frequency of repetition reassociate earlier than the rest, due to the availability of its complementary sequences being more. Thus reassociation in eukaryotes can be thought to be a multicomponent second order reaction. Britten et al have derived[26] the equation describing this reassociation. It is given by

$$\frac{C}{C_o} = \frac{(P.N_u/G)}{1 + K'C_0t} + \sum_{i=1}^{n} \frac{(P\ N_i/G)}{1 + K'N_i\ C_0t}$$

where P = fragment length of the DNA pieces

G = Genome length of the DNA

K = Rate constant ($k = \frac{k'}{a}$)

N_u = Number of unique fragments per genome

N_i = Frequency of repetition of the ith class

n = Total number of classes

1.04 The rate constant

The rate constant of reassociation is dependent on various parameters. These will now be discussed.

Complexicity of the species : The larger the genome length of a particular species the lesser is the probability of a particular DNA fragment to find its complementary strand. This is due to the fact that at a particular concentration of DNA the number of complementary fragments is proportionately decreased with the increase in genome length. Thus in the absence of repetition or undetected polyploidy the rate constant varies inversely as the genome length.

Temperature : Renaturation is generally carried at 60°C and 0.12 M P.b. This is because the rate of renaturation is maximum at about 20° - 30°C below the melting temperature. The rate falls below and above this temperature range. Since formation of a duplex is due to bimolecular collision, at low temperature the number of collision per unit time decreases and

hence the rate decreases. At high temperature the duplex (which generally has mismatch) is unstable. Thus at high temperature though the number of collisions increases the effective number of duplexes formed is decreased and hence the rate. Several authors[34-36] have discussed theoretical models to account for this temperature dependence on the rate of renaturation.

Fragment length

Assuming that there are a large number of nucleation sites along a strand homogeneously distributed, let D denote the average density of a nucleation site.

Conc. of any one nucleation site = $DP/2N$

where P = denatured phosphate concentration

N = total number of base pairs in non repeating sequences, i.e. = Genome length in case of prokaryotes.

$\therefore$ for all sites rate = $K_N \left(\frac{DP}{2N}\right)^2 DN$

where K_N = average rate constant for nucleation at one site, and DN = total number of sites. Thus rate of base pair forma- $K_N D^3 \frac{L}{4N} P^2$. This is because following one nucleation the zippering leads to a formation of L base pair. Of course it

is tacitly assumed that fragments of length L have end to end complementarity.

Now second order rate constant K is given by $v = K P^2/4$

where v = rate of base pair formation. Thus

$$K = D^3 K_N \frac{L}{N}$$

But it has been experimentally found that rate constant $K = 3.5 \times 10^5 \frac{L^{0.5}}{N}$. Wetmur and Davidson[30] have argued on the basis of excluded volume effect that K_N is not independent of the fragment length L. They have found that the fraction f of the second strand which can interpenetrate into the neighbourhood of first strand (i.e. its complementary) and thus be readily available for reaction varies as $L^{-0.5}$. This is believed to cause K_N vary as $L^{-0.5}$, thus making the 2nd order rate const. K vary as $L^{0.5}$, as is experimentally observed.

Wetmur[37] have also discussed the reassociation of fragments of unequal length and found that the rate constant varies as the square root of length of the shorter strand.

Ionic environment

Nucleic acids are poly anions, hence one would expect a dependence of the equilibrium constant for the formation of a nucleation complex on the counterion concentration. The studies of renaturation in high and low salt concentration reveal the following pertinent features :

1) In low salt concentration the rate is proportional to the cube of counterion concentration.

2) The degree of variation of rate with salt concentration is less at high salt concentration.

3) The intrinsic rate varies in a linear fashion between 0.4 M and 2.0 M with log of cation concentration.

4) The rate variations for cations are found to be in the following order :

$$Li^+ > Na^+ > K^+ > (CH_3)_4\ N^+ > (C_2H_5)_4\ N^+ > Cs^+$$

Orosz[38] has ascribed the electrostatic screening ability of cations being responsible in part for such rate variation.

5) Anionic variation has little effect unless it tells on the thermal stability of the DNA.

6) The effect of pH on renaturation is very little within the pH range 6 to 8.

Viscosity

The salient features regarding the effect of viscosity on renaturation can be summarised as :

1) There is a decrease in the renaturation rate with increasing sucrose concentration.

2) Substances which increase macroscopic viscosity (sodium poly acrylate and native T_4 DNA) has little or no effect on the rate.

The practical problem with viscosity arises when one works at high C_0t values. The factor by which the rate is decreased can be monitored by introducing internal standards such as radioactive E. coli DNA and measuring its renaturation.

Nucleic acid composition

The dependence of the nucleation parameter on the nucleic acid composition can be broadly classified into two types :

1) Dependence on base composition : Here the data accumulated is not self consistent. There are conflicting data on dependence of rate constant (general trend being an increase in K with GC content) on the base composition.

2) Dependence on the sugar either ribose or deoxyribose : Due to the unavailability of sufficient number of data on RNA-RNA hybridization kinetics (except in case of homopolymers) the study of the dependence of the nucleation rate on the sugar ring has been limited to mostly to the comparison of DNA-DNA renaturation and DNA-RNA hybridization rate. Such studies have shown that DNA-RNA hybridization is about 0.4 times as fast as DNA-DNA renaturation[39].

Why this study

The study of reassociation kinetics has two principle direction. One is the experimental side which deals with the uncovering of sequence organisation in various eukaryotes. The development of various techniques auxiliary to reassociation kinetics also fall in this category. The other is the theoretical part of it. The mathematical description of the kinetics of reassociation is still intuitive. The effect of various experimental parameters on reassociation is tackled, due to its fluid theoretical basis from an emperical standpoint. It is thus necessary to set a broad based footing of reassociation and incorporate the effect of various experimental variables. This is essential for a more accurate quantitation of sequence organisation and comparison of results under different circumstances. Described below is some of the salient topics where we think, studies could be made for a better and generalised understanding of reassociation kinetics.

a) DNA-DNA hybridization

Though the kinetics of reassociation of DNA is an extremely popular technique, DNA-DNA hybridization in higher organisms has still remained a sparsely touched field. This is because, the kinetics of DNA-DNA hybridization is thought to be complicated by the presence of repeated sequence. The kinetics

of DNA-RNA hybridization with a moderate[40] or large excess[39] of one component have been described. Much information about the precise nature and degree of homology among various species could be known if the kinetic equation for interspecies hybridization is known.

The effect of real parameters

In a practical reassociation experiment we deal with a distribution of fragment length. The distribution is in built as DNA sheared by any mechanical process should lead to a near Gaussian distribution curve. Iyenger and Quave[41] have discussed the details of shearing in relation to the distribution with simulated model for hydrodynamic shearing. It is very well known that in a reaction between two complementary strands, the rate is governed by the shorter strands[37]. That is, with a length distribution we have a multitude of rates, complicating the reaction. Thus it seems natural that the description of a real world reassociation with a unique rate constant (i.e. the rate const. proportional to the average length of the distribution) should be erroneous. This should lead to errors in evaluation of the class and frequency of sequence. The Russian group[42,43] have discussed the effect of random breakage (with fragments of uniform length) in the evaluation of sequence organisation. Thus it seems all the important to describe the kinetics taking into consideration concerted the effect of the real parameters (random breakage and length distribution).

Thermal stability of reassociated DNA

The thermal stability of reassociated DNA is an important issue. The generally accepted procedure is to correlate the degree mismatch with the fall in melting temperature of the reassociated DNA[44,45]. Further analysis into the optical melting data of reassociated DNA should give more information about the precise nature, order, and distribution of mismatch among families of repeated sequences.

Hydroxyapaptite binding to DNA

Differential binding of s.s. and d.s. DNA to HAP crystals is the principle basis of monitoring DNA reassociation. Also thermal chromatography as popularised by Miyazawa and Thomas is anextremely versatile tool for measuring thermal stability of reassociated DNA. With the nature of binding of DNA to HAP fairly well known investigation into the molecular mechanisms of DNA binding and elution raises interesting question about the structure and mode of binding of DNA to HAP. Of particular interest should be the question. How does a duplex DNA while bound to HAP uncoil and form single strands at high temperature?