INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
34100 TRIESTE (ITALY) - P.O.B. 586 - MIRAMARE - STRADA COSTIERA 11 - TELEPHONES: 224281/2/3/4/5/6
CABLE: CENTRATOM - TELEX 460392-1

SMR/111 - 14

SECOND SUMMER COLLEGE IN BIOPHYSICS

30 July - 7 September 1984

Design of immobile nucleic acid junctions.

N.R. KALLENBACH
Department of Biology
Faculty of Arts and Sciences
University of Pennsylvania
Philadelphia, PA. 19104
U.S.A.

DESIGN OF IMMOBILE NUCLEIC ACID JUNCTIONS

NADRIAN C. SEEMAN
Center for Biological Macromolecules, State University of New York at Albany, Albany, New York 12222

NEVILLE R. KALLENBACH
Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

ABSTRACT Nucleic acids that interact to generate structures in which three or more double helices emanate from a single point are said to form a junction. Such structures arise naturally as intermediates in DNA replication and recombination. It has been proposed that stable junctions can be created by synthesizing sets of oligonucleotides of defined sequence that can associate by maximizing Watson-Crick complementarity (Seeman N. C., 1981, Biomolecular Stereodynamics. Adenine Press, New York. 1: 269–278; Seeman, N. C., 1982, *J. Theor. Biol.* 99:237–247.) To make it possible to design molecules that will form junctions of specific architecture, we present here an efficient algorithm for generating nucleic acid sequences that optimize two fundamental properties: fidelity and stability. Fidelity refers to the relative probability of forming the junction complex relative to all alternative paired structures. Calculations are described that permit approximate prediction of the melting curves for junction complexes.

INTRODUCTION

The existence of DNA as a stable extended double helix is by now a concept that is familiar to all. Base-paired duplexes involving oligonucleotide model systems have provided a major source of detailed conformational information (Seeman, 1980; Kallenbach and Berman, 1977) on the state of the bases and backbones in various forms of double helical structure. While it is known that triply and even quadruply branched structures of DNA have a transient existence as intermediates in the replication or recombination of DNA molecules (Broker and Lehman, 1971; Kim et al., 1972), it has not been possible to investigate these forms structurally in short chain molecules, where the region of chain at the junction provides a significant component of the signal. Forked replicative intermediates or four-stranded recombinational structures of the type proposed by Holliday (1964) provide examples of what we define as nucleic acid junctions, i.e., structures in which three or more double helices emanate from a single point. Both replicative and recombinational intermediates are normally unstable due to internal sequence symmetries, which allow their resolution to double helices, via the process of branch point migration (Thompson et al., 1976; Warner et al., 1979; Nilsen and Baglioni, 1979; Seeman and Robinson, 1981). Since this is a very rapid process (Thompson et al., 1976; Warner et al., 1979), these forms have not been tractable to physical characterization at the oligonucleotide level.

It has recently been suggested that the range of migration available to junctions can be severely restricted to form semi-mobile junctions, or eliminated altogether to form immobile junctions from oligonucleotides (Seeman, 1981; 1982). The idea is that oligonucleotides can be constructed that will preferentially associate to form junctions via Watson-Crick base pairing, while the sequences of these molecules do not possess the symmetry necessary to permit branch point migration. Semi-mobile junctions have recently been constructed by P.-L. Hsu and A. Landy (Nash, 1981). Mobile junctions naturally arise whenever cruciform structures fold out from negatively supercoiled DNA circles (Gellert et al., 1978; Lilley, 1980; Panayotatos and Wells, 1981). An example of an immobile junction is illustrated in Fig. 1. A set of rules has been formulated (Seeman, 1981; 1982) that will minimize the sequence symmetry of oligonucleotides. Adherence to these rules will permit oligonucleotides to form stabilized junction structures. These conditions, however, must be supplemented by thermodynamic criteria to assure the stability of a given designated junction.

Here we show how free-energy criteria should be included in sequence design. As an example, we apply literature thermodynamic values appropriate to RNA duplexes in designing an immobile junction composed of hexadecameric strands, with an architecture analogous to the Holliday (1964) genetic recombination intermediate. Data on actual junctions resulting from using the procedure presented here (N. Kallenbach, R.-I. Ma, and N. Seeman, manuscript in preparation) suggest that the DNA sequences are substantially less stable than their RNA counterparts; in fact, we recommend an approximate shift

Dr. N. Seeman is a National Institutes of Health Research Career Development Award recipient.

BIOPHYS. J. © Biophysical Society · 0006-3495/83/11/201/09 $1.00
Volume 44 November 1983 201–209

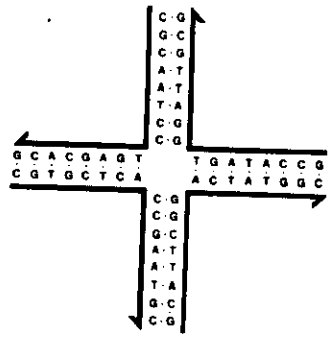


FIGURE 1 A fourth rank junction composed of four hexadecameric fragments. This fragment has four arms, each designated as being composed of eight base pairs. Since $Nc = 4$ (see text), the fidelity of this junction is very high. It contains no repeating G sequence longer than two, and it has a uniphasic melting profile. The melting temperature is estimated to be 65°C, at 1 M salt, with concentrations of 0.1 mM for each strand. This sequence was from those derived from the optimization procedure (see text) for an experimental test of the junction concept. This sequence does indeed form a junction in solution (N. Kallenbach, R.-I. Ma, and N. Seeman, manuscript in preparation).

in *Tm* values (via $\Delta S°$ changes) of 45°C instead of the 20°C used in the figures. The utility of the criteria presented rests on the fact that G-C pairs are significantly more stable than A-T or A-U pairs in either helix. When a complete data set becomes available that describes the stability of deoxynucleotide sequences, we will be better able to optimize design of sequences for any desired architecture. Nevertheless, our present experimental results indicate that the fidelity and stability considerations described here produce stable structures. Given the formidable chemical effort needed to synthesize oligonucleotides in amounts needed for high resolution physical characterization, this procedure is an extremely important practical step.

The rules indicated in the earlier publications (Seeman, 1981; 1982) can be outlined as follows. The construction of immobile and semi-mobile junctions relies on unique base-pairing patterns. These, in turn, are a function of the free energy of association of the individual strands involved. Each strand that is chosen to participate in the formation of an immobile junction may be considered to be composed of a series of overlapping segments of a given criterion length, *Nc*. For example, each hexadecameric strand in the immobile junction shown in Fig. 1 is a series of 13 overlapping segments of length 4. Each of these segments is termed a "criton." A given value of *Nc* implies a diversity of 4^{Nc} critons available with which to construct a given junction. Watson-Crick pairing arrangements that compete with the desired pairing must be considered from a thermodynamic point of view for lengths $<Nc$. However, if the rules indicated below are obeyed, there will be no competing Watson-Crick pairing interactions for segments of length *Nc* or longer. Clearly, *Nc* is a number to be minimized, since this in turn minimizes the strengths of competing interactions by shortening the lengths involved. The generation of junctions containing more and more bases implies that more and more critons are necessary to supply the necessary sequences. However, while the lengths of strands necessary to generate longer and more stable arms grows arithmetically, the diversity of sequences available with each increase in criton length grows geometrically.

Two further terms needed to be defined. A bend is a phosphodiester linkage that is flanked by bases paired to different strands; for semi-mobile junctions, the bend includes the mobile nucleotides. The rank, *R*, of a junction is the number of double helices that directly abut it. Thus, the junction shown in Fig. 1 has $R = 4$. To generate uniquely paired structures with nonmigratory junctions (for length *Nc* or greater), the following rules must be obeyed within the designated pairing regions: (*a*) each criton in the individual strands forming the junction must be unique throughout all strands; (*b*) the complement to any criton that spans a bend in a strand must not be present in any strand; (*c*) self-complementary critons are not permitted; if *Nc* is an odd number, this holds for all critons of size $(Nc + 1)$; (*d*) the same base pair can only abut the junction twice. If it is present twice, those two occurrences must be on adjacent arms.

The practical problem of choosing specific sequences for synthesis as model junctions has been alluded to above. It demands a procedure to optimize the sequence of junctions with a defined architecture, subject to thermodynamic criteria for both stability and fidelity, as well as any additional constraints that may be imposed by the investigators. The previous suggestion (Seeman, 1981; 1982) that immobile and perhaps semi-mobile junctions might be used as building blocks for the construction of rigid geometric figures makes optimization of these attributes particularly desirable. A rapid algorithm is necessary, because each independent base in the junction otherwise increases the extent of calculation by a factor of 4. It is desirable to be able to include potentially more complex investigator-imposed constraints on the generation of junction sequences. For example, these constraints may be used to eliminate certain sequence possibilities that give rise to non–Watson-Crick alternative pairing structures (such as G-G pairs) that have not yet been adequately characterized with thermodynamic data. Similarly, end-fraying may be minimized by requiring the 3′ and 5′ terminal bases to be G's or C's.

METHODS AND STRATEGY FOR OPTIMIZED JUNCTION SEQUENCE GENERATION

Fig. 2 indicates the fundamental logic associated with the junction sequence design algorithm we have developed.

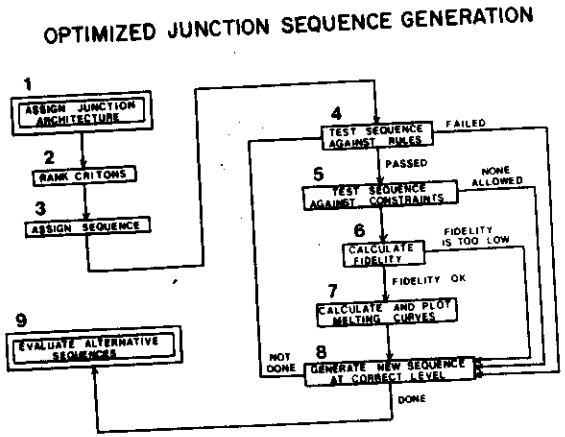


FIGURE 2 A flowchart for optimizing the sequence of a given junction. The nine logical steps in this procedure are indicated schematically. The two steps in double boxes must be done by the investigator, while the other steps may be done automatically by programs. In the first step, the covalent connectivity and desired base pairing are selected by the investigator. Furthermore, specific constraints can be imposed at this stage. In the second step, the critons are ranked by the order of the most rapidly changing base that they contain (see text). After that, an initial numerical sequence must be assigned (Step 3). This numerical sequence is tested against the junction rules, and if it fails, a new sequence is generated by the fast algorithm. If the sequence obeys the rules, its base permutations are then tested (step 5) against investigator selected constraints. If any of the eight sequences implied by the numerical procedure are acceptable, their fidelities are calculated (Step 6), and if these are acceptably high, melting curves are calculated and plotted (Step 7). New sequences are then generated (Step 8) and tested iteratively until all possibilities have been exhausted. The investigator may then evaluate the alternatives presented by the programs.

The double boxes represent the steps done by the investigator in this procedure, while the rest of the logic is readily programmable. We describe here those steps that are critical to understanding the logic of the procedure.

First, note that it is possible to establish within the computer a numerical sequence, in base 4, that obeys the same qualitative rules as Watson-Crick base-pairing complementarity. The sequence is conveniently represented in this numerical fashion. For example, a complementarity relationship of the sort $c = k - i$ is possible, where the independent base is represented by the number i, the complementary base by the number c, and k is a constant. There exist eight different permutations of bases corresponding to this numerical encryption. Sequences may be screened for adherence to the rules (Fig. 2, Step 4) at the numerical level before proceeding to thermodynamic calculations that involve specific base identities. This treatment of the problem leads to an eightfold reduction in computer time. We follow in the next sections the steps outlined in Fig. 2.

Step 1

Assign Junction Architecture. Choice of architecture means specifying how many strands are involved, the number of bases in each, the location of bends, and which bases are independent. The value of the minimum fidelity (Step 6) must also be given. Additional input information required is the size of a criton. This depends in turn on the size of the junction desired, as described above.

The architecture of a junction requires the specification of both covalent connectivity and base pairing relationships. Because of the complementary nature of the Watson-Crick double helices that constitute the junction structure, only half of the bases must be treated as independent variables; those bases complementary to them may be treated as dependent variables. In the case of semi-mobile junctions, only one out of four of the mobile bases is independent. Within the computer, new sequences may be generated simply by the process of counting in base 4. If all of the arms of a junction have the same length, it is possible to fix one independent base at the numerical level, thereby decreasing the number of independent variables by one. (This is analogous to specifying the origin in crystallographic phasing procedures [Hauptman and Karle, 1956].) Even when these considerations have been taken into account, a large number of tasks must still be done by the computer, since N independent bases imply 4^N individual sequences to be tested. It is possible to decrease this number by use of the following procedures.

Step 2

Rank Critons. Next, independent bases are ordered in terms of how fast the digits representing their identities change within the program. This concept is important, because it will be shown below that it permits

one to carry out an exhaustive search without having to test 4^I sequences, where I is the number of independent bases.

Order is an inverse measure of the rate at which the digit representing the base is incremented. Thus, the lowest ordered base will be changed on every pass, the next lowest ordered base will change on every fourth pass, the next lowest ordered base on every sixteenth pass, and so on. The critons themselves may be ordered according to the lowest ordered base within the criton. The critons are then tested for adherence to the rules sequentially, from highest to lowest order. Thus, if a given criton violates one of the rules, the base corresponding to the order of that criton is advanced, rather than the base of lowest order. Until a base at the highest order of violation has been changed, no changes at lower orders would correct the existence of the violation. When a base of any order is incremented, those bases of order lower than that of the incremented base are, of course, set to their lowest value.

This algorithm will be easier to understand if we note that the procedure is analogous to the generation of configurations of numbers, with defined properties, using an odometer or crowd counter, as indicated in Fig. 3. In that figure, the uniqueness of each digit is the specific property required for the numerical configurations. This property for digits is analogous to the first criton rule for junction formation. If we start, at the top of the figure, with six zeros as our initial configuration, and increment the most rapidly changing digit, sequentially, it will take 12,345 steps to get the first successful numerical configuration. On the other hand, if we correct the highest ordered digit that is violating the uniqueness rule, that indicated in the 10,000's place, and then proceed accordingly, it will only take 15 steps to reach the same point.

For example, consider generating the junction shown in

ALGORITHM	UNORDERED	ORDERED
START	0 0 0 0 0 0	0 0 0 0 0 0
FIRST TRY	0 0 0 0 0 1	0 1 0 0 0 0
FIRST SUCCESS	0 1 2 3 4 5	0 1 2 3 4 5
TOTAL STEPS	12,345	15

FIGURE 3 The odometer analogy to the rapid junction algorithm. The object in this example is to generate configurations of numbers with an odometer, in which each digit is unique. This is similar to the first criton rule for junction formation. Two alternative pathways are indicated. On the left, the odometer is incremented in the ordinary fashion, from right to left, until the first number which fulfills the criteria, namely, 012345, is discovered. On the right, we start at the same point, but the digits are ranked. The highest ranking digit that violates the uniqueness rule, that in the 10,000's place, is incremented in the first step. The next step will increment the digit in the 1,000's place from 0 to 1. It will still violate uniqueness relative to the digit in the 10'000's place, so a second incrementation of the 1,000's place will take place to yield 012000. In like fashion, the 100's place will be incremented three times, the 10's place four times, and the 1's place five times. The 15 steps are a great saving in time over the 12,345 steps needed by always incrementing the 1's place, as shown on the left.

Fig. 1. Here there are 64 bases, 32 of which are independent, the other 32 dependent, by complementarity. We arbitrarily designate the 5' octamers as being independent, the 3' octamers as dependent. If, for clarity, we proceed at the base level, rather than the numerical level, all independent bases can be initially set as G's. The top-ranked criton of $Nc = 4$ would be the tetramer at the 5' end of the first strand, 5' HO—GpGpGpG. . . 3'. Bases 2 through 5 of the same strand constitute the second criton, and this is also initially all G's. Since this violates the first rule, the second criton is changed, say, to 5' . . .GpGpGpA. . . 3', by changing the identity of the fifth base and its complement. To do this is much faster than changing the identity of the 32nd independent base four times, the 31st four times, etc., which involves 4^{27} useless operations, until the violation involving the fifth base is reached.

Step 3

Assign Sequence. The initial sequence must be numerically assigned to the independent bases, either by the investigator or by some simple default.

Step 4

Test Sequence Against Rules. The sequence is tested for adherence to the four rules stated above. If a violation occurs, a new numerical sequence is generated by changing the appropriate base in the criton of highest order that violates the rule. Thus, in the example described in Step 2, the second criton was changed, rather than some lower ranking criton. This is in essence a tree-pruning procedure, in which removal of dead branches closest to the trunk efficiently removes dead twigs without having to test each twig individually.

Step 5

Test Sequence Against Constraints. If no violation occurs, the sequences are tested for adherence to the investigator-specified constraints, for each of the eight possible permutations.

Step 6

Calculate Fidelity. Thermodynamic criteria must be applied to all sequences of length $<Nc$. The first question to consider is the pairing fidelity. Is the desired base pairing configuration the most probable configuration in which these particular sequences are to be found in solution? If so, what is its probability relative to other pairing configurations? We have treated this problem in a pairwise fashion; the program routinely considers all alternative binary base-pairing configurations for lengths $<Nc$.

The stability of an oligonucleotide duplex depends on its chain length, sequence and concentration, as well as on environmental variables, such as pH, ionic strength, and temperature (Kallenbach and Berman, 1977). Data on the

relative stabilities of oligoribonucleic acids in conditions equivalent to 1 M NaCl, pH 7, have been accumulated by Tinoco and his coworkers (Borer et al., 1974). The effects of sequence can be evaluated in terms of units representing adjacent sets of two base pairs; the equilibrium constants corresponding to the association within each unit are available, as is the nucleation constant for initial strand interaction. This is denoted by β, with units M^{-1}. A given sequence will then be paired with its complement by a weighting factor that depends on the product of a set of numbers

$$K_{AB} = \beta K_1 K_2 K_3 \ldots K_{N-1}. \quad (1)$$

where N is the chain length of complementary sequences between chains A and B, and β is the nucleation constant. The values of the K_i are tabulated at 25°C by Borer et al. (1974) as $K_i = \exp(-\Delta G^\circ/RT)$. To illustrate the use of Eq. 1, consider the tetramer

(5') AGCU (3')

(3') UCGA (5')

to be decomposed into the three subunit "pairs,"

(5') AG GC CU (3')

(3') UC CG GA (5'),

each of which has an approximate equilibrium constant assigned (Borer et al., 1974). The fidelity is then computed as the ratio, $p = Z^{-1} \exp(-\Delta G_A^\circ/RT)$, where ΔG_A° is the free energy of the desired architecture, and the partition function, Z, includes the ΔG's for all competing pairings of size $Nc - 1$ or lower (the rules exclude all competing interactions for segments $\geq Nc$).

In this way, the maximum concentration of paired molecules of a duplex of arbitrary sequence can be predicted. The situation for oligodeoxynucleotides is unfortunately not so completely defined as for oligoribonucleotides. However, thermodynamic data are available from which primitive sets of K_i's can be created, together with rough values of the ΔH_i°'s (Marky and Breslauer, 1980). These uncertainties do not prevent one from estimating rough relative contributions of different sequences, particularly if appropriately scaled values from oligoribonucleotides can be used. For scaling, we alter the Tm values of Borer et al. (1974), so as to lower the stability of the corresponding oligodeoxynucleotide by 20°C. Only strong reversals in stability of a given sequence from RNA to DNA will really invalidate this approach (Kallenbach, 1977). Comparison of the values of K_{AB} for each set of interactions below the criton length then permits us to estimate the relative contributions of the base pairing in each case. Junctions of maximum fidelity will be those that contain subcriton pairing sequences of minimal stability, relative to the interactions in the complete arms. All binary Watson-Crick alternatives are checked by the program, and their stabilities are compared with those calculated for the double helices chosen for the architecture of the junction. The highest probability junctions above a selected fidelity minimum are retained for further processing. Note that fidelity is a function of temperature. Clearly, sequences must be compared for relative fidelity at a standard temperature, for which we use 25°C. We have found that when pairing lengths are 4 or 5 residues greater than Nc, fidelity approaches unity very closely (0.999).

Step 7

Calculate and Plot Melting Curves. Junction sequences whose fidelities are sufficiently high must next be considered for stability in solution over a range of temperatures. High fidelity is necessary for a monodisperse junction complex, but this criterion is not sufficient. For example, Fig. 4 shows a pair of 4th rank junctions assembled from hexanucleotides. These junctions were generated with the constraint that the two base pairs furthest from the designed junction be G-C pairs, while GpG sequences were prohibited. The sequence in Fig. 4 *a* has very interesting symmetry properties, but its fidelity is not particularly high. The fidelity of the sequence in Fig. 4 *b* is certainly acceptable. However, it is necessary to consider the melting curves for these materials to make sure that they are likely to exist in intact form under the conditions of interest. Therefore, it is necessary to be able to estimate the thermal transition profile in solution. It is clear from the thermal transition profiles shown in Fig. 4 that neither of the junctions shown is likely to be a stable structure in solution.

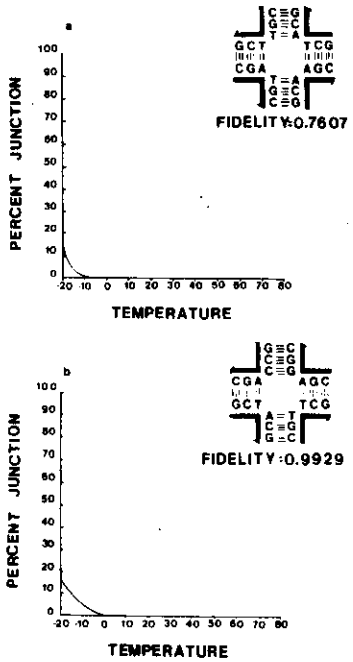


FIGURE 4 Hexadecamers forming junctions with arms 3 base pairs long. Both junctions were generated with the constraint that the two base pairs most distal from the junction be G-C, while GpC sequences were prohibited from this part of the structure. (*a*) A junction with interesting symmetry properties. The junction and estimated melting curve for pH 7, 1 M NaCl is shown. The fidelity is unacceptable, as is the melting temperature which is well below 0°C. (*b*) The hexadecameric junction with the highest fidelity. The fidelity is acceptable, but the melting curve is insufficiently improved to indicate the usefulness of trying to form a junction with these molecules.

The information contained in the estimated equilibrium constants for pairing specific sequences can be used to predict approximate transition profiles for junctions. To do this, enthalpy values, ΔH_i° corresponding to the equilibrium constants K_i, used to assess fidelity, are required. These are considerably less certain for oligodeoxynucleotides than for oligoribonucleotides, but nonetheless reasonable estimates are available, and missing values can be filled in by scaling the corresponding RNA data, as described.

In the case of pairing between sequences on two nonidentical strands, A and B, the value of K_{AB} and the starting concentrations of the two species uniquely characterizes the equilibrium; for starting concentrations, C_A and C_B, and paired complex concentration C_{AB} (moles per liter)

$$K_{AB} = \frac{C_{AB}}{(C_A - C_{AB})(C_B - C_{AB})}. \quad (2)$$

We have discussed how to approximate K_{AB}; thus, C_{AB} can be calculated (Zimm, 1960). This can be done at any temperature if the ΔH_i° for each K_i is known.

Consider next the interaction of four oligomers, A, B, C, and D, which contain uniquely complementing half sequences that can lead to formation of a 4th rank junction complex. Since at equilibrium the concentrations must be independent of reaction pathway, it is sufficient to calculate the junction concentration resulting from any one pathway. For example, one might select (*a*) A + B = AB, (*b*) C + D = CD, and (*c*) AB + CD = ABCD.

From the values of K_{AB}, K_{CD}, and introducing a new factor, σ_R to describe the statistical weight of the central junction loop structure of rank R, the concentrations of junction can be expressed in terms of known quantities. That is, C_{AB} and C_{CD} can be calculated by solving eq. 2, and these values can be introduced into reaction *c* given above to yield $K_{ABCD} = C_{ABCD}/(C_{AB} - C_{ABCD})(C_{CD} - C_{ABCD})$. The value of K_{ABCD} is estimated as $K_{ABCD} = \beta^{-1}[(\sigma_R K_{BC} K_{DA}) + K_{BC} + K_{DA}]$. This is very nearly equal to $\beta^{-1} \sigma_R K_{BC} K_{DA}$, since σ_R is not expected to be very different from unity, while the K's are very large in most cases. It is expected that $\beta^{-1}\sigma_R \gg 1$. If $\beta^{-1}\sigma_R < 1$, only negligible concentrations of the complete junction will be detectable, as discussed more fully below. If a junction entails no strain, we anticipate that only a Jacobson-Stockmayer term (Jacobson and Stockmayer, 1950) is involved, $\sigma_R \sim 1/[R(U+1)]^{3/2}$, where U is the number of unpaired bases abutting the junction on each arm. More generally, we can write $\sigma_R = \sigma_0[R(U+1)]^{-3/2}$, where the factor σ_0 reflects the difficulty of forming the junction.

Concentrations of the ternary and higher (for $R > 4$) intermediates can also be calculated, using stepwise paths such as AB + C = ABC, and ABC + D = ABCD. Thus, the equilibrium concentration of each intermediate, as well as the junction itself can be calculated; a series of relations exists among these intermediates of the form $C_{AB} + C_{BC} = C_{ABC} + C_B$, which simplifies the problem considerably for this approximate treatment.

For values of $\beta^{-1}\sigma_R$ much greater than unity, intermediate forms are much less favored than the junction. Therefore, the following relationship holds approximately for junction of any rank

$$K_J = C_J \left[\prod_{i=1}^{R} (C_i - C_J) \right]^{-1}.$$

In this equation, K_J is the equilibrium constant for the system, C_J is the concentration of junction, and C_i is the initial concentration of each component.

Step 8

Generate New Sequence at Correct Level. Either by failure or success in meeting all the criteria in Steps 4, 5, and 6, a sequence has been evaluated. Next, a new sequence must be tested. The new sequence is generated by the ordering rules described in Step 2.

Step 9

Evaluate Alternative Sequences. Final comparison of the surviving successful sequences is made by the investigator, based on the criteria most relevant to the experiment. These include stability, structure of the thermal transition profiles, Tm values, base composition, or simply ease of synthesis.

RESULTS

The concept of fidelity of synthetic junctions has been dealt with before (Seeman, 1982). Here, we have focused on estimating values for the fidelity, as well as on calculating melting curves for these structures. Short duplex oligonucleotides tend to denature in all-or-none fashion, and significant populations of intermediates arise only in longer chains (Kallenbach and Berman, 1977). The transition behavior of very short chains is such that (*a*) $1/Tm$ is found to be a linear function of log C_0, where C_0 is the strand concentration of each species of interacting mole-

cule, $C_0 = C_A = C_B$; and (*b*), for homogeneous sequences, $1/Tm$ is expected to be an approximately linear function of $1/(N-1)$, where N is the chain length.

Because junctions are inherently inhomogeneous in sequence, only the first of these relations applies. To investigate the behavior of the model, as well as the method of calculation, a computation was carried out. The concentration of all four strands of the junction of rank 4 shown in Fig. 1 was varied over a 10,000-fold range; the resultant Tm values were graphed, as shown in Fig. 5, as $1/Tm$ vs. log C_0, where now $C_0 = C_A = C_B = C_C = C_D$. As can be seen, even for junctions with $N = 16$, with arms of length 8, the reciprocal plot is linear. Thus, the nucleation characteristic of short duplexes is preserved in the junctions, with a strong tendency to favor an all-or-none transition.

The conclusion from this calculation is that the quaternary nucleation process required for forming a stable junction is not innately different from a duplex in its concentration dependence. If one arbitrarily sets all the K_i values equal, it can be shown that the linearity of $1/Tm$ vs. $1/(N-1)$ is also preserved in the quaternary complexes.

A final problem to consider is proper closure of the junction. Given favorable results in the foregoing considerations, two alternatives still exist: (*a*) pairing of the 5′ end of the R′th strand to the 3′ end of the first strand (proper closure) and (*b*) pairing of the 5′ end of the *R*′th strand to the 3′ end of another molecule of the same species as strand 1 (concatenation). The second alternative would lead to large aggregates, since the system is not closed. If $\beta^{-1}\sigma_R \gg 1$, the first alternative will be favored. Electrophoretic analysis of actual fourth rank junctions shows no tendency to concatenate at the concentrations tested.

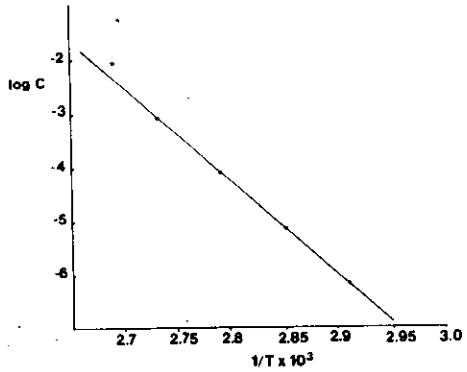


FIGURE 5 The concentration dependence of the melting temperature. The logarithm of the strand concentration has been plotted against the reciprocal of the melting temperature. Note that the linear relationship expected for duplex formation also holds for junction formation under the theory propounded in the text.

DISCUSSION

The purpose of this paper has been to present and clarify the procedures for choosing the sequences from which to construct immobile and semi-mobile junctions. It may be useful to review the steps by which an actual sequence for the immobile junction shown in Fig. 1 was chosen. This case is particularly germane, since we have shown that this sequence does indeed form a junction in solution. Furthermore, since there are 31 independent bases in this junction, limitations on computing time render blind application of the above procedure impossible; no matter how efficient the algorithm, it is impossible to scan all 4^{31} possible sequences that the design of this junction implies. As with most procedures of this sort, this sequence was generated in a stepwise fashion, optimizing the sequence at each step along the way.

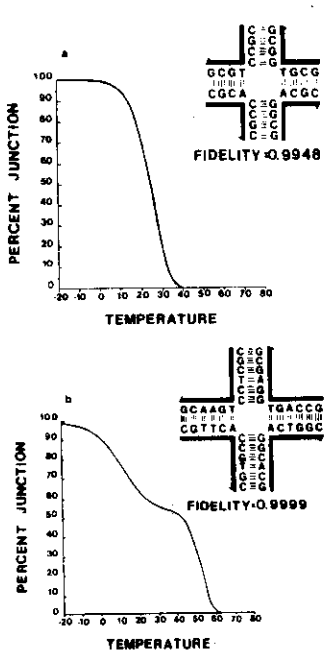


FIGURE 6 Octameric and dodecameric intermediate structures generated in the course of generating the best hexadecameric structure. (*a*) The octameric structure with the highest fidelity. This structure had the same constraints as the structures indicated in Fig. 4, except that no GpGpG sequences were permitted. Both the fidelity and melting temperature are vastly improved. (*b*) The dodecameric structure with the highest fidelity. This structure was generated by inserting two bases into the middle of each arm of the structure shown in *a*. The same constraints were applied. The biphasic melting curve results from the fact that there are more A-T base pairs in the horizontal arms than in the vertical arms.

The first step in this procedure involved using the program in an automatic fashion to generate a junction composed of four octameric strands. These strands were generated in accord with the following constraints. Only GpC or CpG sequences could form the ends of the double helices furthest from the junction, and no GpGpG sequences were permitted. These constraints ensured the stability of the ends of the double helices, while excluding the possibility of G-G non–Watson-Crick pairing which could interfere with junction formation. Application of the program generated the junction shown in Fig. 6 *a* as the one with the highest fidelity. The next pass involved the insertion of two bases in the middle of each tetrameric double helical segment, to yield a junction composed of four dodecameric strands. Again, the constraints against G-G pairing were applied. The sequences with the highest fidelities were considered optimal at both the octameric and dodecameric stages.

The sequence shown in Fig. 6 *b* had the highest fidelity of those sequences generated, although it clearly showed a biphasic melting curve. This feature of the junction composed of dodecameric strands was not a serious impediment in using it as a base for generating a junction composed of hexadecameric fragments. The same constraints were applied, and again the two new bases in each double helical fragment were inserted in the middle of the arm. However, the criteria for selecting the final sequence were different at this stage. This was because all sequences generated by the program had fidelities >0.999. Here, we selected the sequence shown in Fig. 1 as optimized, on the basis of both its sharp uniphasic melting curve (Fig. 7) and the fact that it had the highest melting temperature we encountered while scanning the 126 sequences generated by the last pass of the program. If we calculated the melting curve without the bases nearest the junction being paired, we still got a sharp uniphasic melting curve, this time with $Tm = 48°C$. For initial experimental studies, these features are of paramount importance. The fidelity of this junction is not the highest found by the program, but it is well above 99.9% at room temperature.

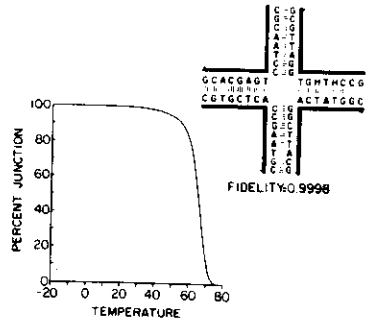


FIGURE 7 The best hexadecameric structure. The sequence shown here has (*a*) a uniphasic thermal transition profile and (*b*) the highest Tm of the sequences generated. The estimated melting curve shown here has a Tm 25° greater than has been observed (N. Kallenbach, R.-I. Ma, and N. Seeman, manuscript in preparation), but it is qualitatively similar.

Note that the initial step in this process, generating the octameric sequences, with 8 independent variables, took much more computer time than the subsequent steps. This is because of the nature of the application of the constraints in the two subsequent steps. The two extensions of the initial octameric sequence took only a few seconds on a Univac 1100/82 computer (Sperry Computer Corp., Blue Bell, PA). Thus, it can be seen that it is possible to generate a sequence for a junction with arms of moderate length, with only a nominal investment in computer time. This is due to the application of both the rapid algorithm and of constraints that are based on optimization of physical parameters at each stage in junction sequence generation. At the same time, however, it has to be recognized that in terms of either fidelity or stability, the resulting structure does not necessarily represent a global optimum. A similar strategy should be applicable to junctions of any architecture.

We would like to thank Leonard Lerman for valuable discussions. We thank Sue Fitzsimmons and Robert Speck for help with the figures, and Linda P. Welch for preparation of the manuscript. Computational facilities were generously provided by the State University of New York at Albany computer center.

This research has been supported by grants GM-26467, GM-29554, CA-31027, and ES-00117 from the National Institutes of Health.

Received for publication 4 January 1983 and in final form 8 June 1983.

REFERENCES

Borer, P. N., B. Dengler, I. Tinoco, and O. C. Uhlenbeck. 1974. Stability of ribonucleic acid double helices. *J. Mol. Biol.* 86:843–853.

Broker, T. R., and I. R. Lehman. 1971. Branched DNA molecules: intermediates in T4 recombination. *J. Mol. Biol.* 60:131–149.

Gellert, M., K. Mizuuchi, M. H. O'Dean, H. Ohmori, and J. Tomizawa. 1978. DNA gyrase and DNA supercoiling. *Cold Spring Harbor Symp. Quant. Biol.* 43:35–40.

Hauptman, H., and J. Karle. 1956. Structure invariants and seminvariants for non-centrosymmetric space groups. *Acta Crystallogr.* 9:45–55.

Holliday, R. 1964. A mechanism for gene conversion in fungi. *Genet. Res.* 5:282–304.

Jacobson, H., and W. Stockmayer. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* 18:1600–1606.

Kallenbach, N. R. 1977. On the secondary structure in mRNA. *Biosystems.* 9:201–210.

Kallenbach, N. R., and H. M. Berman. 1977. RNA structure. *Q. Rev. Biophys.* 10:138–236.

Kim, J., P. A. Sharp, and N. Davidson. 1972. Electron microscopic studies of heteroduplex DNA from a deletion mutant of bacteriophage X 174. *Proc. Natl. Acad. Sci. USA.* 69:1948–1952.

Lilley, D. M. J. 1980. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc. Natl. Acad. Sci. USA.* 77:6468–6472.

Marky, L. A., and K. J. Breslauer. 1980. Calorimetric and spectroscopic investigations of the helix-to-coil transition of the self-complementary d(G-G-A-A-T-T-C-C) duplex. *Fed. Proc.* 39:1880.

Nash, H. 1981. Integration and excision of bacteriophage λ. *Annu. Rev. Genet.* 15:143–167.
Nilsen, T., and C. Baglioni. 1979. Unusual base pairing of newly synthesized DNA in HeLa cells. *J. Mol. Biol.* 133:319–338.
Panayotatos, N., and R. D. Wells. 1981. Cruciform structures in supercoiled DNA. *Nature (Lond.).* 289:466–470.
Seeman, N. C. 1980. Crystallographic investigation of oligonucleotide structure. *In* Nucleic Acid Geometry and Dynamics. R. H. Sarma, editor. Pergamon Press, New York. 109–148.
Seeman, N. C. 1981. Nucleic acid junctions: building blocks for genetic engineering in three dimensions. *In* Biomolecular Stereodynamics. R. H. Sarma, editor. Adenine Press, New York. 1:269–278.
Seeman, N. C. 1982. Nucleic acid junctions and lattices. *J. Theor. Biol.* 99:237–247.
Seeman, N. C., and B. H. Robinson. 1981. Simulation of double stranded branch point migration. *In* Biomolecular Stereodynamics. R. H. Sarma, editor. Adenine Press, New York. 1:279–300.
Thompson, B. J., M. N. Camien, and R. C. Warner. 1976. Kinetics of branch migration in double stranded DNA. *Proc. Natl Acad. Sci. USA.* 73:2299–2303.
Warner, R. C., R. Fishel, and F. Wheeler. 1979. Branch migration in recombination. *Cold Spring Harbor Symp. Quant. Biol.* 43:957–968.
Zimm, B. H. 1960. Theory of melting of the helical form in double chains of the DNA type. *J. Chem. Phys.* 33:1349–1356.

Reprinted from Nature, Vol. 305, No. 5937, pp. 829-831, 27 October 1983

An immobile nucleic acid junction constructed from oligonucleotides

Neville R. Kallenbach*, Rong-Ine Ma* & Nadrian C. Seeman†

* Department of Biology, Leidy Laboratories, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
† Department of Biological Sciences, Center for Biological Macromolecules, State University of New York at Albany, Albany, New York 12222, USA

Base-paired DNA duplexes involving oligonucleotide model systems have provided the major source of detailed structural and dynamic information about double helical structure[1]. Triple- and quadruple-branched 'junction' structures of DNA have a transient existence as intermediates in the replication or recombination of DNA molecules[2–5] while cruciforms may be inducible by negatively supercoiling closed circular DNA[6–11]. However, it has not been possible to investigate these forms structurally at high resolution in short-chain molecules, where the junction will yield a significant component of the signal, because these naturally occurring intermediates are inherently unstable, due to internal sequence symmetry, which permits their resolution to double helices, via branchpoint migration[12–15]. We have recently proposed that migration can be eliminated to yield immobile junctions from oligonucleotides[16–19] by combining sequence symmetry constraints with equilibrium calculations. We present here electrophoretic and UV optical absorbance experiments which indicate that four hexadecadeoxynucleotides (Fig. 1) indeed do form a stable tetrameric junction complex in solution.

The electrophoretic mobility of a nucleic acid oligomer in non-denaturing conditions is a function of its size, shape and extent of base pairing[20,21]. When the individual strands in Fig. 1, or equimolar mixtures of pairs, triplets and the tetrad corresponding to the complete junction, are subjected to electrophoresis, the patterns shown in Fig. 2 result: lanes *c–f* show the mobility of the individual strands. Figure 2, lanes *m–r* contain each of the six possible pairs of strands in equimolar mixtures. The first four of these binary mixtures are combinations which should form an arm of the junction. The last two mixtures do not correspond to an arm, but rather represent the diagonal combinations which should not associate. The mixtures which correspond to an arm of the junction (Fig. 2 *m–p*) migrate as single bands, with mobilities markedly less than those of any of the single strands. The two diagonal mixtures in lanes *q* and *r* migrate as single strands, with mobilities that correspond to their components.

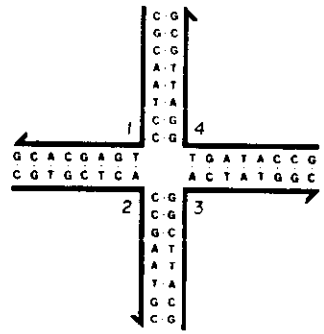


Fig. 1 An immobile nucleic acid junction composed of four hexadecanucleotides. This sequence has been designed by using the sequence symmetry minimization rules, supplemented by equilibrium distribution optimization[16–19]. The strand numbering indicated is used throughout the text. Note the lack of 2-fold symmetry around the centre, so that migration is not possible. This sequence also contains no repeating GpG sequence longer than two, in order to minimize this form of non-Watson–Crick pairing also. These sequences were commercially synthesized by phosphotriester techniques.

Figure 2 *g–j* contains the four possible equimolar triplet complexes which can be formed by omitting each of the four strands from the tetrad in turn. The major component of each of these lanes is again a single band, travelling slower than the paired bands. Lane *k* contains an equimolar mixture of all four strands. This mixture travels as a single band, with a mobility lower than any of the other oligomeric mixtures. The presence of a single band with appreciable mobility in the lane corresponding to the tetrameric complex indicates that a molecular species with a well defined stoichiometry predominates. Higher unclosed complexes (1:2:3:4:1:2:3:4:1 . . .) thus cannot represent a significant fraction of the total material present. This finding is in accord with earlier predictions[18,19].

The stoichiometry of the strands in the complex involving all four strands can be estimated from the gel electrophoresis experiment shown in Fig. 3. In this experiment, an equimolar mixture of strands 1, 2 and 4 [component (i)] was titrated with strand 3 [component (ii)]. Figure 3*f* shows the mobility of component (i) alone, while lane *h* contains only component (ii). Figure 3*c* shows that mixing components (i) and (ii) in equimolar ratios leads to a complex with a single major component. No excess free single strands of (ii) occur when (i) is in excess (Fig. 3*d,e*). By the same token, free single strands do accumulate when component (ii) is in one-half molar excess, as seen in Fig. 3*b*. In conjunction with ^{1}H-NMR data (to be published elsewhere) indicating that strands 3 and 4 form a 1:1 complex, we conclude that the stoichiometry of the junction formed is indeed 1:1:1:1.

The slope of the Ferguson plot of electrophoretic mobility versus acrylamide concentration is a means of estimating the frictional constant of any molecular species[22]. In Fig. 4*a*, the complex is compared with 36- and 72-bp restriction fragments; from the slopes shown, the four-stranded structure has a friction

Fig. 2 Polyacrylamide gel electrophoresis of oligodeoxynucleotide strands and mixtures. Each well contained 2.5 μg of each strand, alone or in combination with others. Lanes *a*, *b* and *l* contain restriction digests of ΦX174RFDNA; *a* is the *Hin*fI digest, *b* the *Hae*III digest and *l*, is the *Hin*cII digest. The lowest molecular weight fragments in these digests are: 42, 48, 66 and 82 (*Hin*fI), 72 and 118 (*Hae*III) and 79 and 162 (*Hin*cII). Lanes *c–f* contain strands 1, 2, 3 and 4 respectively. Lane *g* contains an equimolar mixture (based on extinction coefficients derived from dry weights) of strands 1, 2 and 3; lane *h*, 1, 2 and 4; lane *i*, 1, 3 and 4; lane *j*, 2, 3 and 4. Lane *k* contains an equimolar mix of strands 1, 2, 3 and 4. Lanes *m–r* contain equimolar mixtures of pairs of strands: *m* contains 1 and 2; *n*, 3 and 4; *o*, 1 and 4; *p*, 2 and 3; *q*, 1 and 3; and *r*, 2 and 4.

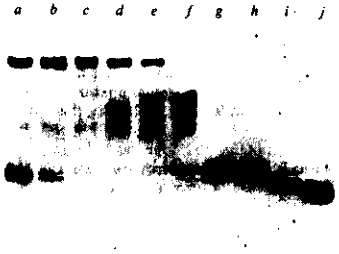


Fig. 3 Stoichiometry of mixing in the four-component complex. In this experiment, electrophoresis of mixtures containing different ratios of two components was carried out: component (i) consists of an equimolar mixture of strands 1, 2 and 4, while component (ii) consists of strand 3 alone. Lanes *g–j* contain 8 μg of free strands 4, 3, 2 and 1, respectively. Lane *f* contains 6 μg of component (i). Lanes *a–e* each contain 6 μg of component (i), plus component (ii) in the following amounts (μg): *a*, 4; *b*, 3; *c*, 2; *d*, 1; and *e*, 0.5.

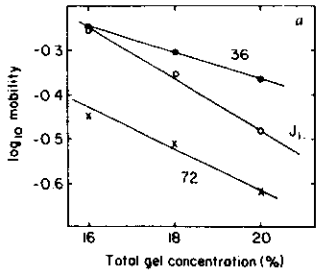
constant quite different from linear duplex DNA of approximately the same size, or indeed of any size. Thus, the junction is a structure distinct from linear duplex DNA.

Next, the relative stability of the different complexes of strands 1 to 4 was assessed by thermal denaturation studies monitored by hyperchromism at 260 nm[23]. Figure 4*b* shows the thermal transition profiles of the individual strand 3 and an equimolar mixture of strands 1 and 2, compared with that of the quaternary complex at half the total strand concentration. The fact that the hyperchromism in the junction is twice as great as in the same concentration of pairs strongly suggests that the complex is closed, with four arms nearly intact. The increase in junction stability (seen in the higher melting temperature, T_m) further strengthens this argument.

The present level of characterization does not allow us to conclude that the junction is completely immobile, despite the presence of non-complementary sequences flanking the junction. Further experimentation is needed to establish that a limited amount of non-Watson–Crick mobility is not occurring in the vicinity of the junction.

The above experiments clearly show that it is possible to design and synthesize immobile nucleic acid junctions by using optimization procedures[16–19]. We have a second synthetic junction, composed of dodecameric fragments (N.R.K., R.-I.M., A. J. Wand, G. H. Veeneman, J. H. van Boom and N.C.S., in preparation), which we have compared with the hexadecameric junction reported here. Gel electrophoresis analysis indicates that while this material also forms a junction, alternative pairings may occur unless optimization statistics are as favourable as those of the hexadecamers discussed here[18,19].

This work was supported by NIH grants GM-29554, ES-00117 and CA-31027. N.C.S. is the recipient of a NIH Research Career Development Award. We thank Dr P. Lu for his 36-bp restriction fragment.



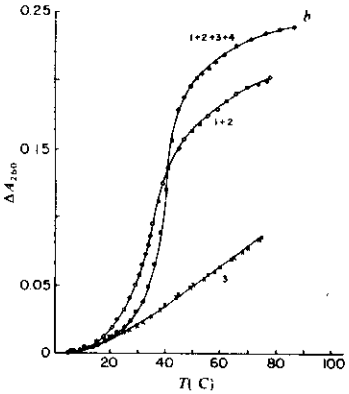


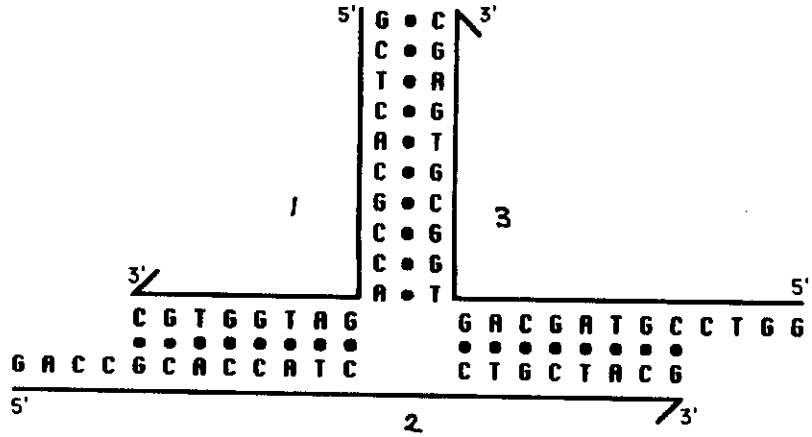
Fig. 4 Physical characterization of the junction. *a*, Ferguson plot of the equimolar complex of strands 1, 2, 3 and 4, indicated as J_1, compared with linear DNA duplexes containing 36 and 72 base pairs. Gels were cast with different compositions of total acrylamide, and the mobility of the four-component complex measured with respect to that of a reference dye, xylene cyanol FF, and reference duplexes of 36 and 72 bp. The smaller reference is a cloned fragment of *Escherichia coli lac* operator, the larger is the lowest molecular weight duplex from a *Hin*f digest of ΦX174 RF DNA. *b*, Thermal transition profiles of the quaternary complex (25 μM total strands), an equimolar mixture of strands 1+2 (49 μM total strands) and individual strand 3 (98 μM total strand) at 260 nm. The UV absorbance of the samples was measured at each of a series of temperatures (T), and the results expressed as $A_{260} = A_{260}(T)/A_{260}$ (10 °C) − 1. Typical A_{260} values for high-molecular weight DNA duplexes approach 30% (ref. 24); actual values depend on base composition, length and solvent. Fragment 3 alone exhibits a typical non-cooperative transition characteristic of nucleic acids in the absence of base pairing[24]. Note that the concentration of 4-fold complex is 6.25 μM, while that of the paired arm is 24.5 μM. Hence, the hyperchromicity in the junction is actually more than four times that of a single arm.

Received 6 June; accepted 11 August 1983.

1. *Biomolecular Stereodynamics*, Vol. 1 (ed. Sarma, R. H.) 1–343 (Adenine, New York, 1981).
2. Dressler, D. & Potter, H. *A. Rev. Biochem.* **51**, 727–761 (1982).
3. Holliday, R. *Genet. Res.* **5**, 282–304 (1964).
4. Sigal, N. & Alberts, B. *J. molec. Biol.* **71**, 789–791 (1972).
5. Nash, H. *A. Rev. Genet* **15**, 143–167 (1981).
6. Platt, J. R. *Proc. natn. Acad. Sci. U.S.A.* **41**, 181–183 (1955).
7. Gierer, A. *Nature* **212**, 1460–1461 (1966).
8. Hsieh, T. & Wang, J. C. *Biochemistry* **14**, 527–535 (1975).
9. Gellert, M., Mizuuchi, K., O'Dean, M. H., Ohmori, H. & Tomizawa, J. *Cold Spring Harb. Symp. quant. Biol.* **43**, 33–40 (1979).
10. Lilley, D. M. J. *Proc. natn. Acad. Sci. U.S.A.* **77**, 6468–6472 (1980).
11. Panayotatis, N. & Wells, R. D. *Nature* **289**, 466–470 (1981).
12. Thompson, B. J., Camien, M. N. & Warner, R. C. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2299–2303 (1976).
13. Warner, R. C., Fishel, R. & Wheeler, F. *Cold Spring Harb. Symp. quant. Biol.* **43**, 957–968 (1979).
14. Meselson, M. *J. molec. Biol.* **71**, 795–798 (1972).
15. Seeman, N. C. & Robinson, B. H. in *Biomolecular Stereodynamics* Vol. 1 (ed. Sarma, R. H.) 279–300 (Adenine, New York, 1981).
16. Seeman, N. C. in *Biomolecular Stereodynamics* (ed. Sarma, R. H.) 269–277 (Adenine, New York, 1981).
17. Seeman, N. C. *J. theor. Biol.* **99**, 237–247 (1982).
18. Seeman, N. C. & Kallenbach, N. R. in *Nucleic Acids: The Vectors of Life* (ed. Pullman, B.) (Reidel, Dordrecht, in the press).
19. Seeman, N. C. & Kallenbach, N. R. *Biophys. J.* (in the press).
20. Fangman, W. L. *Nucleic Acids Res.* **5**, 653–665 (1978).
21. Sealey, P. G. & Southern, E. M. in *Gel Electrophoresis of Nucleic Acids* (eds. Rickwood, D. & Hames, B. D.) 39–76 (IRL, Oxford, 1982).
22. Rodbard, D. & Chrambach, A. *Anafyt. Biochem.* **40**, 95–134 (1971).
23. Freifelder, D. M. *Physical Biochemistry*, 377–393 (Freeman, San Francisco, 1976).
24. Van Holde, K. E. *Physical Biochemistry*, 168–169 (Prentice-Hall, Englewood Cliffs, New Jersey, 1971).

Printed in Great Britain by Macmillan Production Ltd, Basingstoke, Hampshire

(12)



R=3 Junction for assembly experiments

ASSEMBLY OF COVALENTLY-CLOSED STRUCTURES FROM A JUNCTION.

① Allow assembly (C high). ② Ligate w T4 DNA ligase
③ Digest w exo III.