Total synthesis of a gene

Total Synthesis of a Gene

*H G Khorana**

Summary. The method developed for the total synthesis of a given DNA containing biologically specific sequences consists of the following. The DNA in the double-stranded form is carefully divided into short single-stranded segments with suitable overlaps in the complementary strands. All the segments are chemically synthesized starting with protected nucleosides and mono-nucleotides. The 5'-OH ends of the appropriate oligonucleotides are then phosphorylated with the use of [γ-^{32}P]ATP and polynucleotide kinase. A few to several neighboring oligonucleotides are then allowed to form bihelical complexes in aqueous solution, and the latter are joined end to end by polynucleotide ligase to form covalently linked duplexes. Subsequent head-to-tail joining of the short duplexes leads to the total DNA. The methods are described for the construction of a biologically functional suppressor transfer RNA gene. The total work involved (i) the synthesis of a 126-nucleotide-long bihelical DNA corresponding to a known precursor to the tyrosine suppressor transfer RNA, (ii) the sequencing of the promoter region and the distal region adjoining the C-C-A end, which contained a signal for the processing of the RNA transcript, (iii) total synthesis of the 207 base-pair-long DNA, which included the control elements, as well as the Eco R1 restriction endonuclease specific sequences at the two ends, and (iv) full characterization by transcription in vitro and amber suppressor activity in vivo of the synthetic gene.

Organo-chemical methods for the synthesis of oligonucleotides began to be developed (*1*) soon after the elucidation of the structures of the nucleic acids (*2, 3*). While considerable advances were made in the 1950's and 1960's in constructing polydeoxyribonucleotides of defined nucleotide sequences (*4*), there continued to be severe practical limits on the size of the polynucleotide chains that could be assembled

* The author is Alfred P. Sloan, Professor of Biology and Chemistry, Massachusetts Institute of Technology, Cambridge 02139.

Reprinted with permission from AAAS, H G Khorana, Total Synthesis of a Gene, *Science*, Vol.203, pp.614–625, 16 February 1979.

unambiguously by purely chemical methods. However, for biological studies, completely defined polynucleotides in the size range often much higher than those accessible by chemical methods are required. The formidable tasks confronting organic chemistry were recognized early (*5*) and, therefore, attempts were made to couple chemical methods, which alone offer oligonucleotides of controlled sequences, to other concepts, which together would afford high-molecular-weight nucleic acids of defined structures. Thus, in the 1960's, it was possible to use short synthetic deoxyribo-oligonucleotides with repeating sequences as templates for the nucleic acid-polymerizing enzymes, and this approach enabled the preparation of a variety of double-stranded DNA-like polymers of high molecular weight and messenger RNA's (mRNA) of defined sequences (*4, 6, 7*). The latter proved to be very useful in studies of the genetic code. However, the large objective of the synthesis of macromolecular DNA's having nonrepeating and biologically specific sequences (*5*) required a different approach. Toward this goal, the central concept was the inherent ability of polynucleotide chains to form ordered bihelical complexes by virtue of base pairing. Thus, the goal was to join, end to end, chemically synthesized polydeoxyribonucleotides while these were held together in properly aligned bihelical complexes. The discoveries of the enzymes polynucleotide ligase (*8*) and polynucleotide kinase (*9*) proved crucial in these studies and, fortunately, the average size of oligonucleotides, which was demanded of chemical synthesis, proved to be quite short, being well within the range of chemical techniques (*10*). Synthesis of a double-stranded DNA corresponding to the major yeast alanine transfer RNA (tRNA), which was the first one to be sequenced, became the first objective, and by 1970 this objective had been accomplished (*11*). From various considerations, which are described later, an *Escherichia coli* tRNA gene offered much greater opportunities for biochemical and biological studies; and, already in 1968, the *E. coli* tyrosine suppressor tRNA gene had been chosen as the target for synthesis. The total project involved (i) the synthesis of a 126-nucleotide-long bihelical DNA corresponding to a precursor to the tRNA, (ii) the sequencing of the adjacent promoter region and the distal region, adjoining the C-C-A end, which contained a signal for processing the RNA transcript, (iii) syntheses of the DNA's corresponding to these control regions, and (iv) biochemical and biological studies of the totally synthetic gene. This article presents a brief overview of the different phases of the above work, from its start in the late 1960's until its completion recently.

Chemical Synthesis of Deoxyribo-oligonucleotides

While a detailed account of the chemical methodology is beyond the scope of this article,

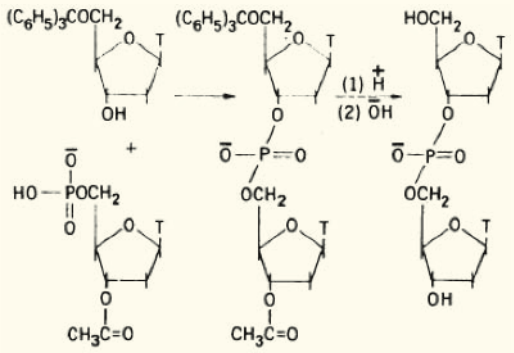


Fig. 1. Condensation of two protected components to form a dinucleotide. On top left is the 5'-protected thymidine (T) and in bottom left is the 3'-protected thymidine 5'-phosphate. Condensation to form the inter-nucleotide bond (middle) is effected by activation of the phosphate group. The protecting groups are removed by mild acid and alkali to give the unprotected dinucleotide (right). Mild alkaline treatment alone would free the 3'-OH end group for further chain elongation by condensation with protected mono- or oligonucleotides carrying 5'-phosphate groups. [From (*4*); reproduced by permission of *Pure and Applied Chemistry*]

a brief outline of the principles used in the syntheses of the very large number of oligonucleotides required in the work presented here is in order. Chemical synthesis, which creates most (>90 percent) of the required internucleotide bonds present in both strands of the intended DNA, continues to occupy easily more than 50 percent of the effort required in the total synthesis of DNA, even though notable advances have been made, both in methods of synthesis and in rapid separation of required intermediates and products.

The scheme of synthesis of the simplest dinucleotide, thymidylylthymidine (TpT), is shown in Fig. 1. Three concepts are worthy of note. (i) One of the starting components is a nucleoside with a free 3'-hydroxyl group, the 5'-hydroxyl group being blocked by the bulky trityl group, an acid-sensitive group. (ii) The second component in the condensation is a mononucleotide that has a 3'-hydroxyl group blocked by a simple alkali-labile group. (iii) The 5'-phosphomonoester group of a nucleotide is directly activated by a reagent so as to effect condensation with the 3'-OH group of the nucleoside to form an internucleotide bond. This example focuses on the problems of suitable protecting groups in dinucleotide synthesis as well as in the stepwise synthesis of higher oligonucleotides and, then, on the question of activating agents (*4*). Figure 2 gives a summary of the protecting groups, all readily and safely removable, and of the condensing agents

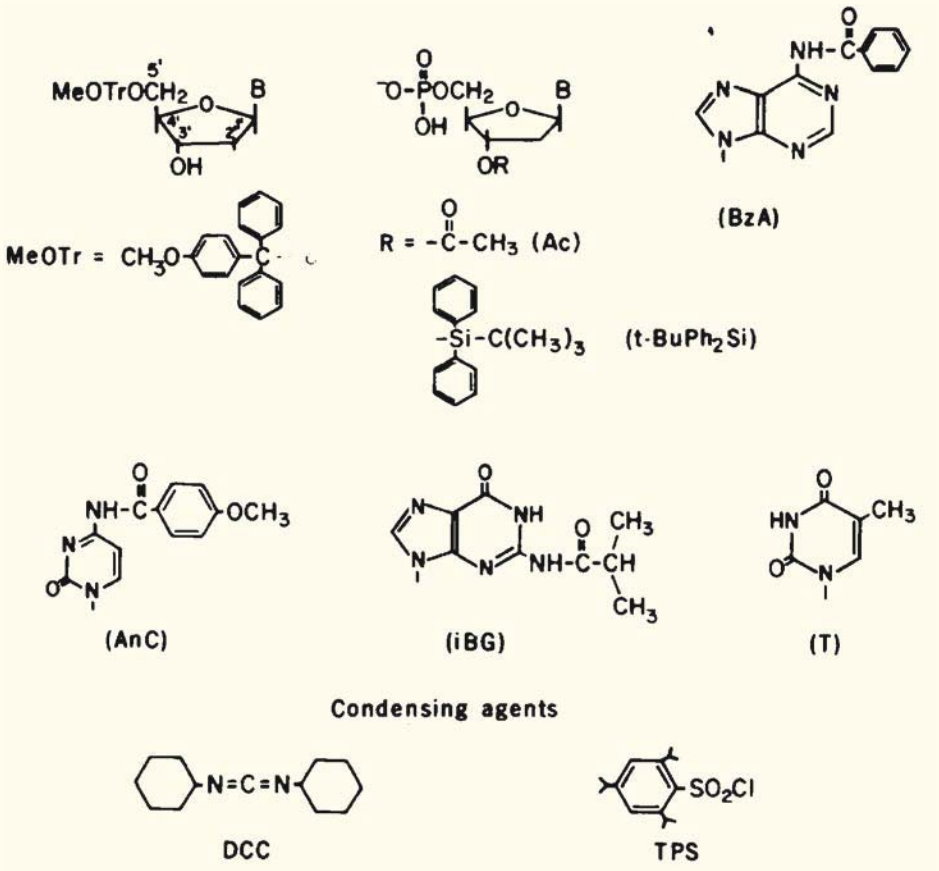


Fig. 2. Groups used for the protection of deoxyribonucleosides and deoxyribonucleotides and the condensing agents used in the chemical synthesis of polynucleotides. The abbreviations used for the different groups are also shown. [Reproduced by permission of Academic Press, New York]

currently used. All the groups shown in Fig. 2 can be removed either under very mildly acidic conditions (methoxytrityl), or with ammonia (acyl protecting groups for amino functions), or with fluoride ion (silyl groups). A practically important feature is the selective unblocking of the terminal 3'-OH group in order to elongate the chain by condensation with a 5'-mononucleotide or oligonucleotide. The example shown in Fig. 3 illustrates the principles in the stepwise synthesis of an undecanucleotide. Unambiguous synthesis requires the separation and characterization of the required products at every step of the synthesis. Currently used procedures, which are much more rapid than the older anion exchange chromatography, include extraction methods and high pressure

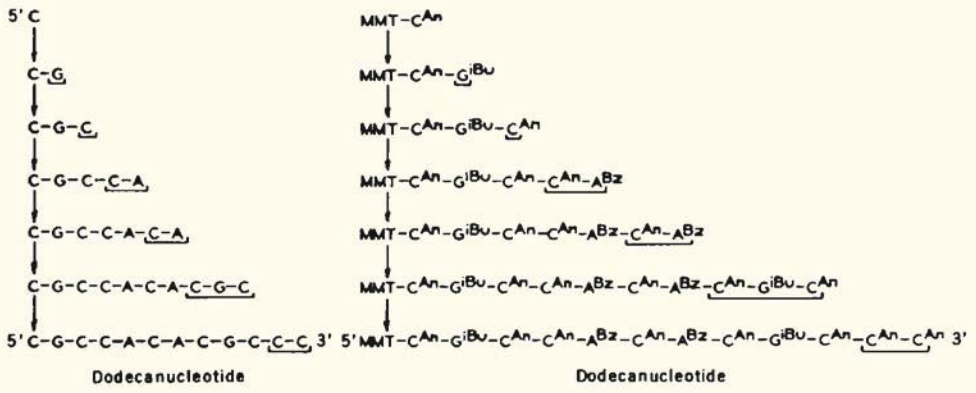


Fig. 3. Steps in the chemical synthesis of a dodecanucleotide. The incoming mono- and oligonucleotides are shown by horizonal brackets under the oligonucleotide blocks. The single letters C, G, and A stand for deoxyribonucleosides. The protecting groups (Fig. 2) used to protect the amino groups in different bases are shown above the nucleoside initials. The phosphate groups connecting the individual nucleosides are shown simply by hyphens between the nucleoside initials. The 5'-OH end group is at the left, and the chain grows in the 3' direction. MMT, monomethoxytrityl. [Reproduced by permission of the Academic Press, New York]

liquid chromatography (*12, 13*). Syntheses of as many as 40 oligonucleotide segments described below were all performed according to the above principles.

Strategy for the Synthesis of Double-Stranded DNA

Early experiments on the stabilities of short, synthetic, double-stranded complexes with overlapping sequences and on the utilization of such duplexes by polynucleotide ligase for joining reactions was encouraging in that deoxyoligonucleotide segments within the practical range of organic synthesis were adequate (*10*). These results led to the three-phase strategy shown in Fig. 4 for the total synthesis of a bihelical DNA (*11*). Thus, the

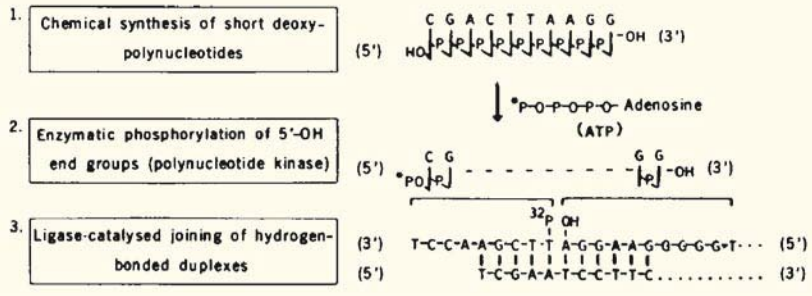


Fig. 4. Three-step strategy for the total synthesis of double-stranded DNA. [Reproduced by permission of Academic Press, New York]

first phase requires the chemical synthesis of short polynucleotides of chain lengths in the range of 10 to 12 nucleotides: these would correspond to the entire two strands. A hypothetical sequence showing the 3'- and 5'-OH ends is shown at top right in the figure, with the Fischer projection method. In the second step, the 5'-OH ends of the segments, except those which are to be at the 5' termini of the duplexes, are phosphorylated by using (γ-^{32}P)-labeled adenosine triphosphate (ATP) and the polynucleotide kinase. This enzyme is admirably suitable for the purpose because of its absolute specificity for phosphorylating the 5'-OH end group of a short or a long polynucleotide chain without any regard to the sequence.

Since the main task in the next phase (phase III) is the accurate and specific end-to-end joining of all of the synthetic segments, the presence of suitable radioactive phosphate groups at 5' ends of the synthetic segments, as introduced above, is crucial. Three, four, or as many as six or seven segments with overlapping complementary sequences are brought together under suitable conditions of ionic strength, divalent ions, and temperature, and the polynucleotide ligase is then used to bring about covalent joinings to form covalently linked duplexes. The joinings can be monitored by analyses of the joined products in various ways. For example, the double-stranded product and its constituent strands, after separation, may be sized quite precisely by electrophoresis in polyacrylamide gels. Further, they may be degraded to 5'- or 3'-mononucleotides, and the distribution of radioactivity in different mononucleotides will immediately show the accuracy in joining. Characterization is further aided by manipulation of the specific activity of [γ-^{32}P]ATP used in phosphorylation reactions. Thus, specific activity at the internal linkages in the short duplexes formed first (see below) may be much lower than that at the termini which are to connect these short duplexes to form the total DNA. Different joinings may also be further distinguished by concomitant use of ^{32}P and ^{33}P isotopes. Indeed, both ^{32}P- and ^{33}P-labeled phosphoryl groups may be used at two levels of specific activity at different sites, allowing completely error-free characterization.

Transfer RNA Genes as Synthetic Targets

In the middle 1960's, a number of genes specifying polypeptides or small proteins, whose amino acid sequences were known, could have been considered as targets for synthesis since at least the nucleotide sequences of the regions specifying the amino acids could be inferred from the knowledge of the genetic code. However, tRNA's were proving to be uniquely interesting molecules. They were shown to be at the center of specificity in the

biological synthesis of proteins. Because of this vital biological function, their intriguing structural features and extensive biochemistry connected with them, they were chosen as the goals of synthetic work. Finally, an equally important argument in their favor was the progress that was being made in the determination of their nucleotide sequence, the major yeast alanine tRNA being the first one to be sequenced (*14*). Work on the total synthesis of the DNA corresponding to this tRNA sequence was started soon after and successfully completed some years later (*11*).

While the above work strengthened confidence in the methodology for DNA synthesis, from the biochemical standpoint, limitations were apparent for further work with it. The bacterial tRNA gene, that for tyrosine tRNA suppressor, was selected as the next target. The main arguments, which fortunately became even more compelling as we proceeded with this project, for this choice were as follows: (i) firm knowledge of the primary nucleotide sequence by at least two groups of workers (*15*), (ii) the advantages which this tRNA offered in studying the biochemistry of the amino acid charging reaction, (iii) the dramatic progress in the biochemistry of protein synthesis in vitro in the cell-free *E. coli* system and in study of suppression in vitro of amber mutation by suppressor tRNA's, (iv) the insertion of this tRNA gene into the transducing bacteriophage ϕ80 by Brenner and Smith and their coworkers (*16*) and the convenience of working with ϕ80 psu^{+}_{III} and related derivatives for probing the nucleotide sequences of the regions adjacent to and controlling the tyrosine tRNA structural gene, (v) extensive work on the structure-function relationships in the same gene by the above group (*17*) using the genetic approach, and (vi) the discovery of a precursor for this tRNA (Fig. 5) (*18*). This structure

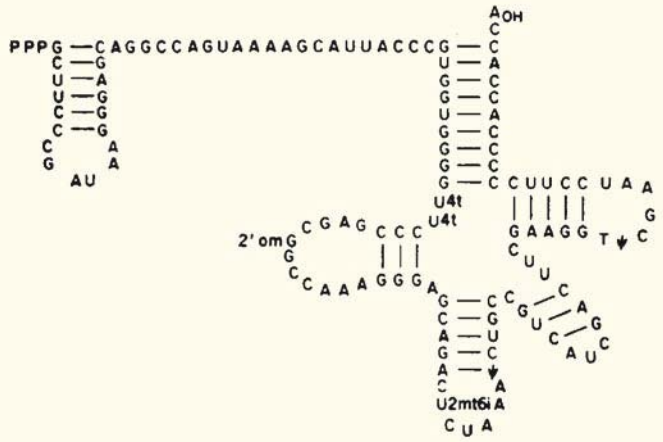


Fig. 5. The primary nucleotide sequence of an *E. coli* tyrosine tRNA precursor. The sequence is written with the standard clover-leaf structure for the tRNA portion and the possible hairpin at the 5' end. [From (*20*); reproduced by permission of the *Journal of Biological Chemistry*]

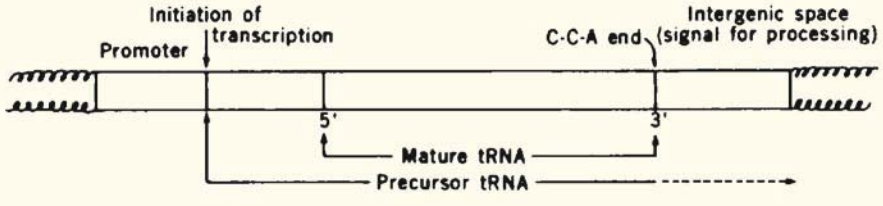


Fig. 6. Linear arrangement of the different parts of the tyrosine suppressor tRNA gene. The gene, which is one of the two duplicate genes in tandem, is located at 27 minutes on the *E. coli* genetic map. The suppressor gene arises by a single nucleotide change in the anticodon of one of the two tyrosine genes in tandem. U, C, A, and G stand for uridine, cytidine, adenosine, and guanosine, respectively; U4t for 4-thiouridine; 2'omG for 2'-O-methylguanosine; 2mt6iA for 2-methylthio-6-isopentyladenosine, and ψ for pseudouridine. [Reproduced by permission of Academic Press, New York]

containing a 5'-triphosphate end group unambiguously defined the site of initiation of transcription and, consequently, attention could be focused precisely on the promoter (preinitiation) region and on the process of transcription of this gene. In fact, one could envisage a linear arrangement for the different regions including the control elements of the gene as shown in Fig. 6. There is the region of the structural gene, 126 base pairs in length, that corresponds to the precursor RNA (Fig. 5). The primary RNA product is then cleaved at the point shown to give the tRNA length. The starting point of transcription must be at the point shown by the arrow and the DNA that precedes it must by definition be the promoter region, the region recognized by the RNA polymerase. Transcription begins at the indicated point and continues, presumably, until there is a termination signal. Until recently the view was held that the signal for termination in this gene was probably soon after the C-C-A end, but it has now been shown that rho-dependent termination of transcription in fact occurs about 225 nucleotides downstream from the structural gene (*19*). As is shown later, an early event in the processing and maturation of the tRNA is an endonucleolytic cleavage of the primary transcript seven nucleotides downstream from the C-C-A sequence.

Total Synthesis of the 126-Base-Pair DNA

The plan adopted for the synthesis of this DNA, which corresponds to the tyrosine tRNA precursor, consisted of a total of 26 chemically synthesized deoxypolynucleotide segments (*20*). These segments, when subgrouped into the four duplexes (duplex I to duplex IV), are shown in Fig. 7. How was this plan deduced? Obviously, there is an enormously

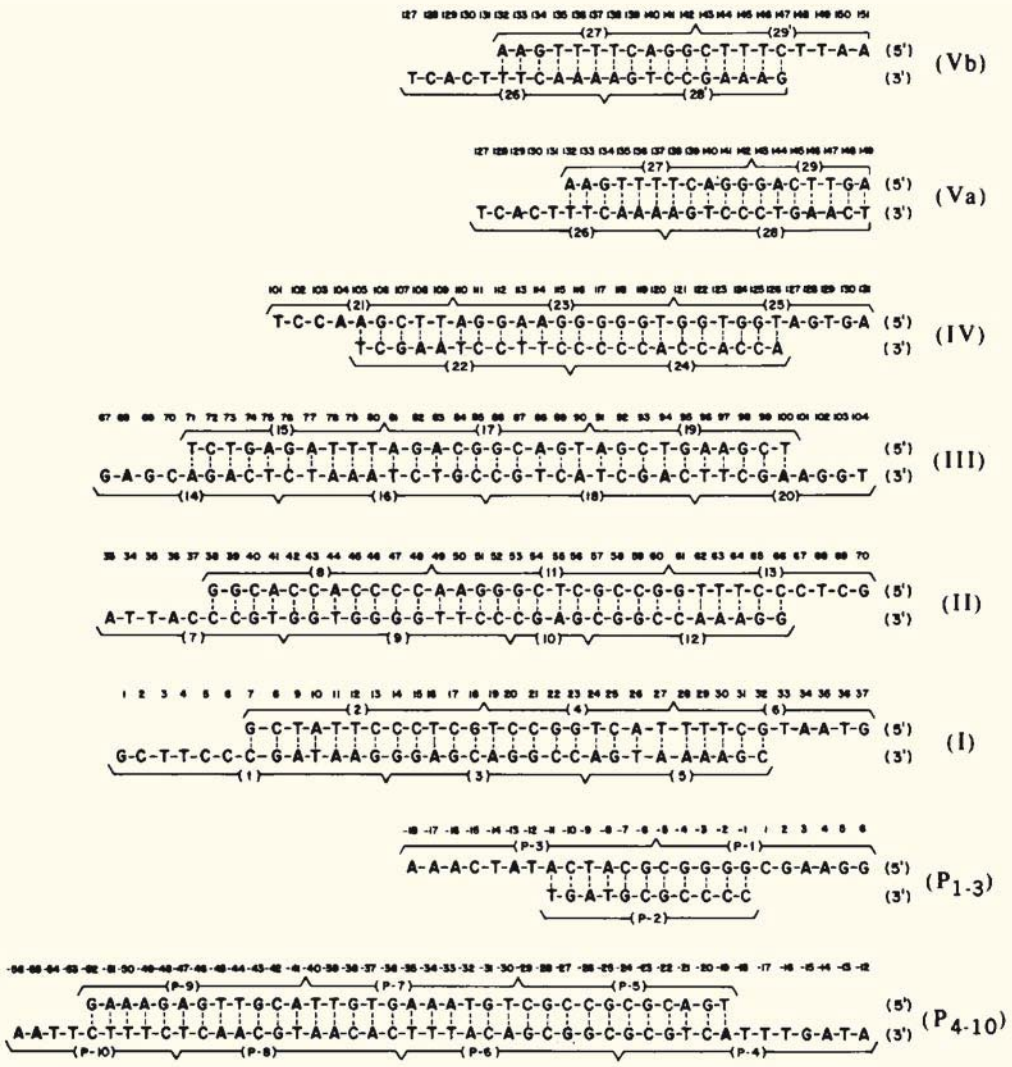


Fig.7. Plan for the total synthesis of the *E. coli* tyrosine suppressor tRNA gene including the promoter region and the region containing the processing signal adjoining the C-C-A end. Duplexes I, II, III, and IV contain a total of 25 chemically synthesized segments and represent the entire tRNA precursor sequence shown in Fig.5. The 5' –pppG terminus of the precursor sequence corresponds to nucleotide 1 (segment 1, numbers of segments in parentheses) of duplex I. The 3' end of the precursor sequence is contained in segment 25 (duplex IV) (nucleotide 126). The grouping of the total segments to the four duplexes shown was arrived at following a large amount of systematic enzymatic work. For details regarding duplexes P_{4-10} , Va, and Vb, see text. [Reproduced by permission of Academic Press, New York]

large number of possibilities for dividing a DNA of this size into segments having a chain length of 10 to 12 nucleotides. The main rules that were followed arise from the demands of chemical and enzymatic work. Thus, there is the first basic requirement of the overlap of about five base pairs at the protruding ends. Second, it is important from the standpoint of enzymatic complementarity (self-structure) within the single-stranded segments as well as self-complementarity at 5' protruding ends be avoided as far as possible (*21*). Another important consideration is economy in the chemical work. Multiple use of synthetic oligonucleotide blocks, short or long, should be maximized. Thus, in the plan adopted (duplex I to duplex IV, Fig. 7), a systematic search (by computer program) revealed, for example, that the nonanucleotide sequence C-C-C-C-A-C-C-A-C occurs twice (segments 8 and 24), the octanucleotide sequence, G-C-T-C-C-C-T-T occurs twice (segments 2 and 13), the hexanucleotide sequence, T-T-C-G-A-A (self-complementarity) occurs four times. Arguments of this kind can be made for parts of a large number of other segments. Finally, at least some synthetic intermediates or complete segments were available from previous work with the yeast alanine tRNA system (*11*).

When syntheses of all of the required segments have been accomplished, sub-grouping with a view to optimal and error-free enzymatic joinings is the next objective. Inevitably, there would be a large number of alternatives for grouping the segments. The formation of perfectly ordered double-helical complexes could easily be predicted to be favored, resulting in quantitative joinings to form covalently linked duplexes. Paradoxically, the extent of joinings in different systems has varied widely, the yields, in general, being less or much less than quantitative. The protocols for optimal joinings, the rates of reactions, and the yields of the required duplexes from single-stranded oligonucleotides require detailed investigation and may vary considerably (*11, 22*).

Similarly, ultimate choices in grouping of segments continue to require a large amount of empirical work. A desirable feature to be borne in mind while grouping the single-stranded segments is that, as far as possible, the termini bearing 5'-OH groups should be protruding. Polynucleotide kinase-catalyzed phosphorylations of the terminal 5' hydroxyl groups for the subsequent purpose of joining the duplexes goes to completion much more rapidly in the above configuration. The four groups that were finally arrived at in the present work are shown in Fig. 7, duplex I to duplex IV. Thus, each duplex consisted of five to seven chemically synthesized segments. While the formation of each one of these duplexes had unique characteristics, only one example, that of the synthesis of duplex I, will suffice to illustrate the procedures (Fig. 8) (*23*).

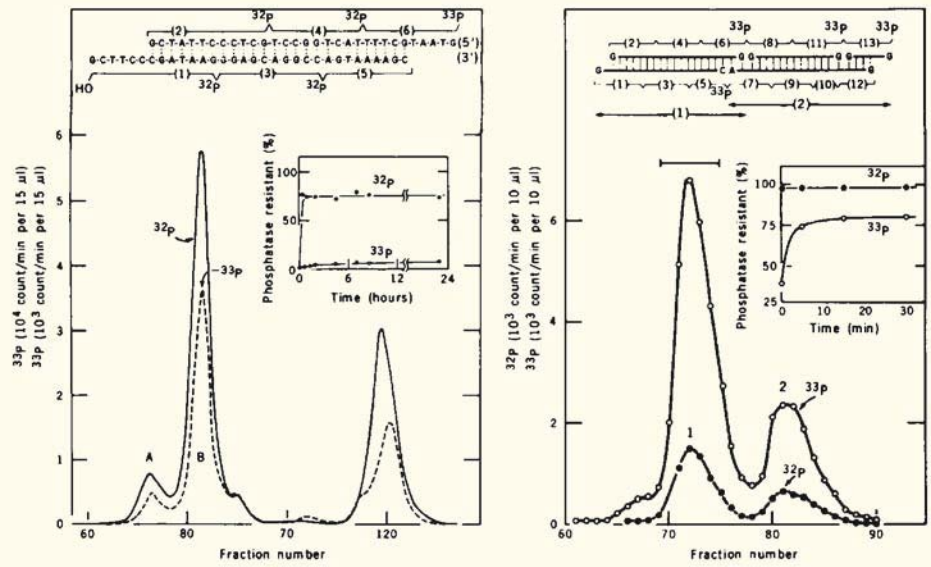


Fig. 8 (left). The synthesis and purification of duplex I. Radioactive (^{32}P or ^{33}P) labeling pattern of the 5' ends in different segments is shown in the inset, as are the kinetics of joining. The latter were followed by the formation of phosphatase-resistant radioactivity. Separation of the products was accomplished on a Bio-Gel (0.5 m) column. Peak B contained the pure duplex I. Peak A contained a dimer, the structure of which has been discussed (*23*). Unreacted segments appeared last. [Reproduced by permission of the *Journal of Biological Chemistry*]

Fig. 9 (right). Joining of duplex I to duplex II. The ^{33}P radioactive labeling sites are shown in the duplexes. At the remaining sites of enzymatic joining. there was weak ^{32}P radioactivity. The kinetics of joining are shown in the inset. Separation was on a Bio-Gel (0.5 m) column. Peak I contained the joined product. Unjoined products were in peak 2. Further details are in (*24*). [From (*24*); reproduced by permission of the *Journal of Biological Chemistry*]

In the last phase, the duplexes corresponding to different parts of the total sequence are joined through their protruding single-stranded ends to form the total duplex. Fortunately, joinings at this stage are efficient and rapid. An example, shown in Fig. 9, is the joining of duplex I to duplex II. Both duplexes carried distinctively labeled 5'-phosphate groups to facilitate analysis (*24*). Finally, Fig. 10 shows the resulting total synthetic duplex corresponding to the tRNA precursor, and it also depicts the synthetic chemical and enzymatic steps used.

Promoter Region and That Next to C-C-A End of Tyrosine tRNA Suppressor Gene

The promoter region lies to the bottom left of the structural gene (Fig. 6), and the signal

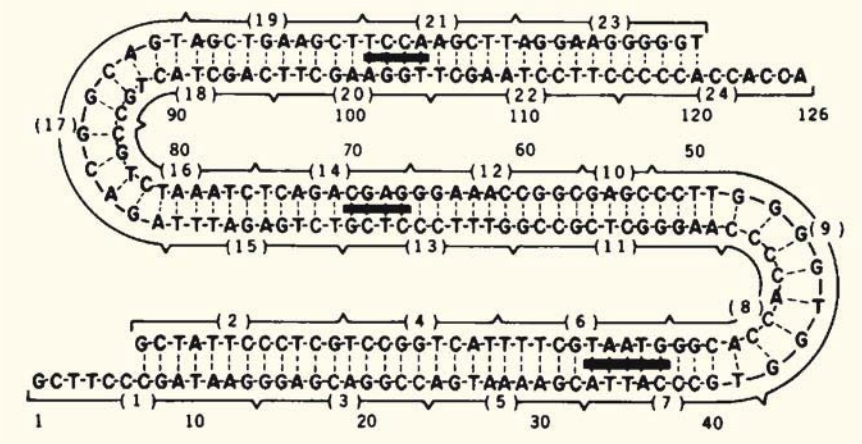


Fig. 10. The totally synthetic DNA duplex corresponding to the entire sequence (126 nucleotides) of the tRNA precursor shown in Fig. 5. The number and distances between the carets show the oligonucleotides that were synthesized chemically. The carets indicate the sites where joining was accomplished by the use of polynucleotide ligase. The three thick strips placed between the strands of the duplex at different positions indicate the sites where the preformed duplexes (I to IV) were joined to each other. The sequence of their joinings was as follows: Duplex I was joined to duplex II (Fig. 9). Separately, duplex III was joined to duplex IV. Duplex (I + II) was finally joined to duplex (III + IV). [From (*24*); reproduced by permission of the *Journal of Biological Chemistry*]

for the processing of the primary transcript of the gene is at the distal end. As was mentioned above, the *E. coli* tyrosine suppressor tRNA gene can be integrated into the temperate bacteriophage ϕ80. The resulting ϕ80 psu^{+}_{III} provides a much more convenient starting material for work on the required sequences of the suppressor gene. However, the bacteriophage DNA still is much too large (the single-strand molecular weight is $\sim 15 \times 10^6$), and it is necessary to focus specifically on the very short regions of interest that adjoin the DNA shown in Fig. 10. Because of the fortunate circumstance that, from the synthetic work described above, deoxyribopolynucleotides with a large selection in chain lengths and corresponding to both strands of the structural gene (Fig. 10) are available, the following general approach to the desired sequences becomes possible. The two strands of the bacteriophage ϕ80 psu^{+}_{III} DNA containing the above gene may be separated, and synthetic deoxyribopolynucleotides of suitable size may be specifically hybridized to the separated strands at sites adjoining the regions whose sequences are to be determined. Primer-template relationships can thus be established with proper polarities, and the 3' ends of the primers may then be extended into the two regions of unknown

sequence by means of DNA polymerases. The latter would bring about the nucleotide incorporations according to the nucleotide sequence in the template strand, and the nucleotide sequence of the unknown region can then be deduced from the pattern of nucleotide incorporation (*25, 26*).

Only one set of experiments (Fig. 11) concerned with the determination of a part of the sequence in the promoter region will be outlined to illustrate the concepts (*26*). The primers DNA-I and, preferably, DNA-II are hybridized to the 1-strand of ϕ80 psu$^{+}_{III}$ DNA, which had been shown earlier (*27*) to have the sequence such that DNA-I would be expected to hybridize to it. In the figure are shown the starting point and direction of transcription (to the right) and the start of the promoter region (to the left). DNA-II, when extended by five nucleotides, will be at the start of the promoter region. The addition of the five expected nucleotides will provide further assurance that hybridization of the primer occurred at the correct site. To illustrate the methods that were used to keep the newly growing chains within manageable size, the following points may be made. In the first step (experiment A in Fig. 11), three deoxynucleoside triphosphates were provided in the polymerase reaction mixture. The absence of dTTP (deoxythymidine triphosphate) limited the chain growth to the sequence shown in the dotted box in A. The sequence of

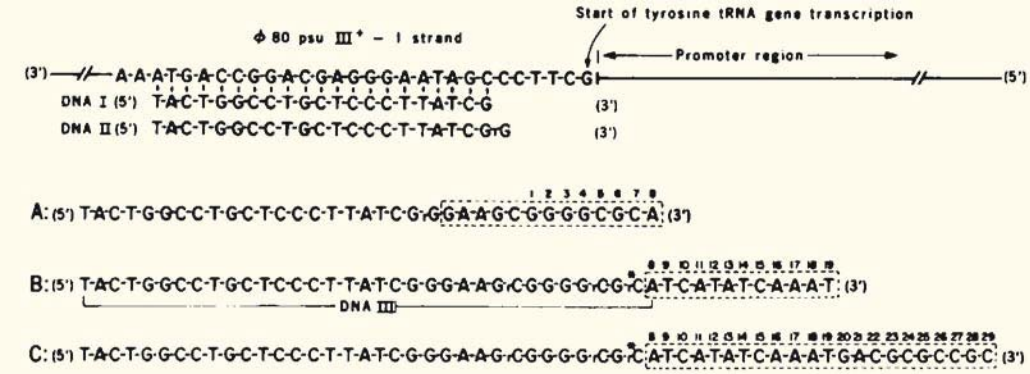


Fig. 11. Experimental design for sequencing and a part of the nucleotide sequence determined in the promoter region of the tyrosine tRNA gene. The primer-template complexes prepared at different stages were obtained by hybridizing DNA I, DNA II, or DNA III to the I-strand of the ϕ80 psu$^{+}_{III}$, DNA. The DNA polymerase-catalyzed elongations were carried out by using three nucleoside triphosphates at a time and then substituting rCTP for dCTP. The new nucleotide sequence discovered after each elongation, subsequent alkaline cleavage, and analysis of the new fragment is shown in the appropriate dashed box. [Reproduced by permission of Elsevier Scientific Publishers, Amsterdam]

the latter having been determined, the elongated chain was used again as a primer with a different set of three triphosphates (experiment B in Fig. 11). A third device used for selective fragmentation of the newly grown chain was the use of rCTP (ribocytidine triphosphate) in place of dCTP (deoxyribocytidine triphosphate). This conferred alkaline lability at specific sites in the chain. Finally, in experiment C of Fig. 11, another principle was used, namely, to add three nucleoside triphosphates at standard concentrations, but the fourth triphosphate (in this case dTTP) in a rate-limiting amount. In this way successively longer chains useful for sequencing were obtained. This work was carried out several years ago, and it is of interest that these fundamental concepts are widely used in the current methodology for sequencing of DNA fragments of much longer chain lengths.

The total promoter sequence as far as determined is shown in Fig. 12A. The arrow at the top right shows the direction of transcription, nucleotide No. 1 in-dicating the starting point. The sequence thus far determined is from nucleotide –1 to –59. The sequence has interesting elements of palindromic symmetry that are indicated by dashed and solid boxes, and their relation to each other is indicated by direction of the arrows. Unfortunately, the significance of this remarkable symmetry in the sequence remains unclear, especially since its presence or extent is variable in the different promoters that have been sequenced. How could one tell that the sequence determined was adequate to constitute a functional promoter? Work from our own laboratory *(28)* and cumulative

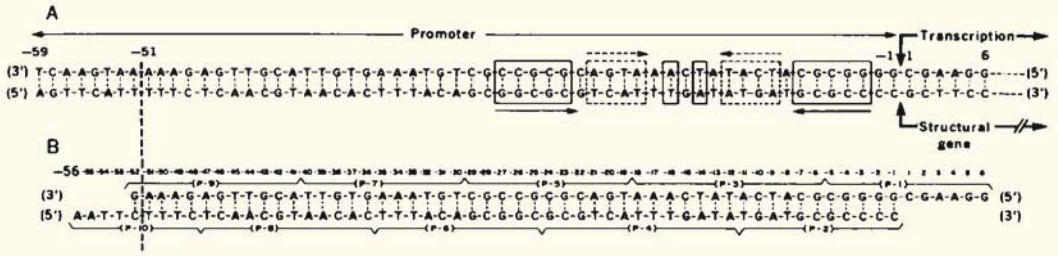


Fig. 12. (A) The nucleotide sequence in the promoter region of the tyrosine suppressor tRNA gene. The point of initiation of transcription and its direction into the structural gene are shown. Elements of twofold symmetry in the sequence are shown in the boxes, their correspondence being indicated by arrows. **(B)** Plan for the total synthesis of the promoter region of the tyrosine suppressor tRNA gene. Included are the 51 nucleotide base pairs in the promoter region plus one CoG base pair plus the single-stranded A-A-T-T sequence at the 5' end. Together the A-A-T-T-C- sequence at the 5' end corresponds to one-half of the self-complementary duplex which serves as the recognition sequence for the Eco R1 restriction enzyme. The ten segments (P-1 to P-1O) to be synthesized are indicated by horizontal brackets, [Reproduced by permission of Academic Press, New York]

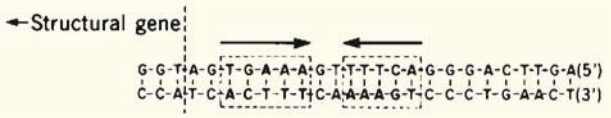


Fig. 13. Palindromic symmetry in the nucleotide sequence found in the tyrosine tRNA gene beyond the 3' end sequence of the tRNA. The sequence, which is shown in the double-stranded configuration, results in the formation of a loop in the RNA transcript. [From Loewen, Sekiya, and Khorana in (*25*); reproduced by permission of the *Journal of Biological Chemistry*]

evidence forthcoming from different laboratories did strengthen the notion that this was so. The minimal functional length was estimated to be between 40 and 50 nucleotides.

Turning to the part of the gene beyond the C-C-A sequence, that is, the amino acid acceptor end of the tRNA, the question again was: How much sequencing will have to be done to cover the signal for the processing of the gene transcript to the tRNA? It may be noted that processing to generate the C-C-A terminus evidently had occurred in vivo in the tRNA precursor isolated (Fig. 5). An arbitrarily short sequence of 23 nucleotides adjoining the amino acid acceptor end was determined (*25*), and the corresponding duplex was synthesized (*29*). The sequence is shown in Fig. 13. The palindromic sequence turned out to be of special significance. Thus, by virtue of this symmetry, the RNA transcript can assume a hairpin structure which, as seen later, is evidently recognized by a specific endonuclease. The signal having thus been located within the sequence of Fig. 13, it was concluded to be unnecessary to pursue further the sequence work in this region.

Total Synthesis of the Gene

In sum, the work reviewed above required the synthesis and joining of a total of eight duplexes (Fig. 7). Of these, the synthesis of duplexes I to IV and their joining to form the "structural gene" has already been reviewed (Fig. 10). Duplexes P_{1-3} and P_{4-10} (bottom of Fig. 10) correspond to the promoter sequence described above (Fig. 12A); however, in synthetic work, modification was introduced at the terminus distal to the structural gene (Fig. 12B). Thus, the natural sequence was retained up to nucleotide 51, and the terminal sequence was changed to include the restriction enzyme, Eco R1, recognition sequence. This would enable the insertion of the synthetic gene into a suitable vector. The total duplex corresponding to the promoter region involved the chemical synthesis of ten segments shown in Fig. 12B. Enzymatic joining of these segments required special strategy because of the common sequences within segments forming parts of the

palindromes. Hence, the synthesis of the promoter was accomplished in two parts, duplexes P_{1-3} and P_{4-10}.

Analogous to the above considerations for modification of the promoter terminus, the duplex at the distal end, which has been introduced above (Fig. 13, also duplex Va in Fig. 7), was also modified to the duplex Vb (Fig. 7). Thus, the naturally occurring sequence adjoining the CC-A end was shortened to only 16 nucleotides. The latter completely retained the palindrome and evidently, because of this, there was no deleterious effect on processing of the transcript by the specific endonuclease (see below). Following the 16 nucleotides, a short artificial sequence, identical to that used at the promoter end, was added. Thus, the Eco R1 restriction enzyme sequence was present at this terminus as well.

The joining of the eight duplexes (Fig. 7) can be approached and was, in fact, performed in a number of alternative ways. Only one final stage experiment leading to the totally synthetic gene is shown in Fig. 14. As was mentioned above, the joining of preformed

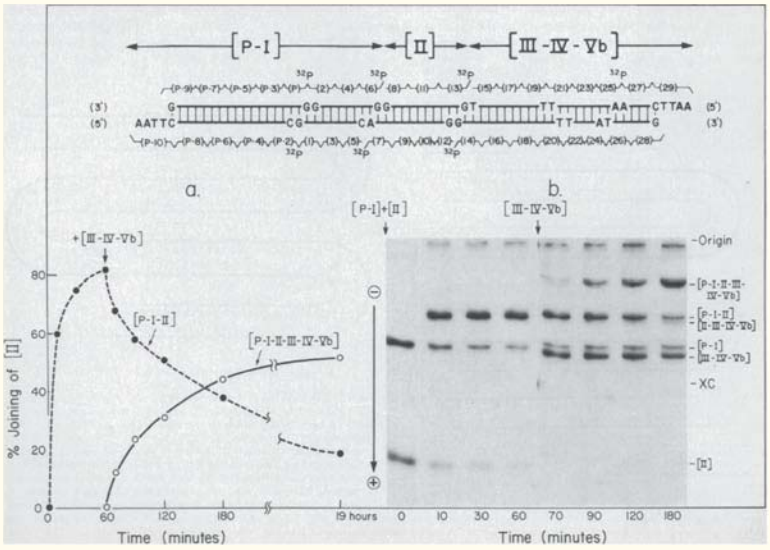


Fig.14. Stepwise synthesis of the synthetic gene P-I-II-III-IV-Vb. The promoter duplex was first joined to duplex I to form P + I, To the latter was added duplex II, the joining to form P-I-II being rapid as seen in the first curve. Preformed III-IV-Vb was then added, and the kinetics of formation of the total gene was followed by electrophoresis on a polyacrylamide gel. [Reproduced by permission of Academic Press, New York]

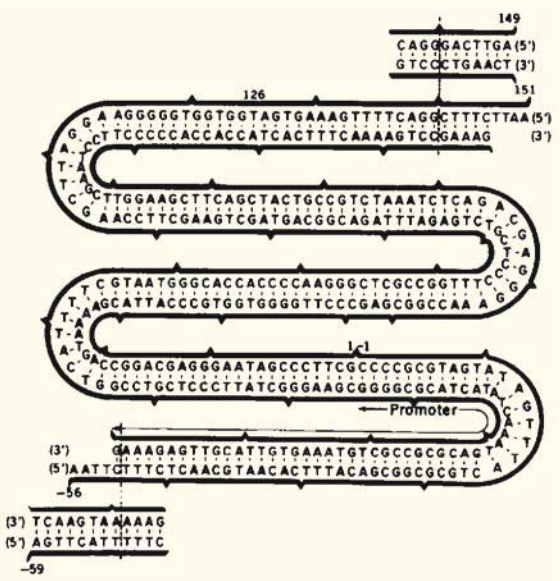
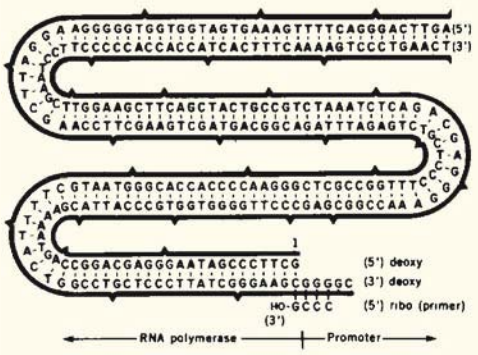


Fig. 15. Synthetic tyrosine suppressor tRNA gene. From bottom left, following the terminal Eco R1 endonuclease-specific sequence, the DNA contains a 51-nucleotide-long promoter region, a 126-nucleotide-long DNA corresponding to the precursor RNA and, finally, a 25-nucleotide-long duplex. Of the latter, 16 nucleotides belong to the natural sequence adjoining the C-C-A end, and the remainder is a modified sequence including the Eco R1 endonuclease-specific sequence. Additional duplex segments shown on top right and bottom left are the natural sequences as they occur in these regions and continue in the two directions. [Reproduced by permission of Academic Press, New York]

duplexes through their complementary protruding oligonucleotide sequence usually proceeds rapidly and in high yield. The totally synthetic gene, isolated and characterized in a variety of ways, is shown in Fig. 15. The additional sequences shown at top right and at bottom left are the natural ones as they would continue in the two directions in the *E. coli* genome.

Transcription in vitro of the Synthetic Gene

Two of the first aims after total synthesis were (i) the in vitro transcription and subsequent processing of the primary transcript and base modifications to form the mature tyrosine tRNA and (ii) demonstration, in vivo, of suppression of the amber mutation after insertion of the gene into suitable vectors used to transform suitable strains of *E. coli.* Transcription was first studied in detail with the synthetic DNA shown in Fig. 16. The structure of the promoter was largely unknown at this time, and only a segment containing a single-stranded sequence corresponding to the first five nucleotides of the promoter region was linked to the structural gene. The DNA in Fig. 16 also contained the duplex Va (Fig. 7) at the distal (C-C-A) end. After earlier studies of primer-dependent transcription of single- and double-stranded synthetic deoxyribopolynucleotides *(30),* the ribotetranucleotide (5')-C-C-C-G-(3'), complementary to the 3' terminus sequence



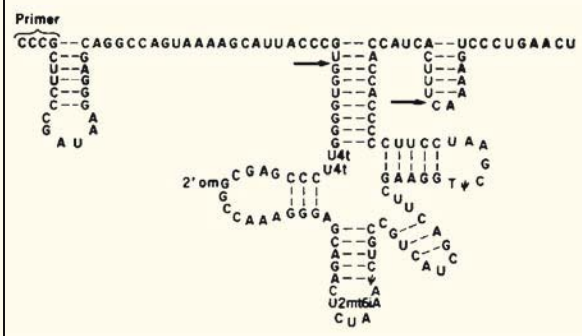


Fig. 16 (left). Primer-dependent transcription of the synthetic gene lacking the promoter but containing an additional 23-nucleotide-long duplex corresponding to the sequence adjoining the C-C-A end. There is only one synthetic oligonucleotide segment of the promoter end, which has the first five nucleotides of the promoter sequence protruding into the promoter region (bottom right). The tetranucleotide rCCCG was used as the primer for transcription. The direction of transcription by RNA polymerase is shown. Transcription was specific to the required strand, the one with the single-stranded pentanucleotide sequence. [Reproduced by permission of Academic Press, New York]

Fig. 17 (right). Structure of the major primary transcript with the use of primer-dependent transcription of synthetic DNA as shown in Fig. 16. Completion of the different modifications of bases as shown is not certain. Probably modifications were incomplete. The two crucial endonucleolytic cleavages (arrows) for processing to functional tRNA are shown. The arrow on the left is the previously determined site of cleavage by endonuclease P. The arrow on the right is the site now determined as the first step in processing of the 3' end of the transcript. [Reproduced by permission of Academic Press, New York]

was used as the primer (Fig. 16). Under suitable conditions, transcription was essentially restricted to that of the desired strand, and the primary transcript was shown to consist largely of the RNA with the structure shown in Fig. 17. When the latter was exposed to an *E. coli* extract, the supernatant resulting from centrifugation at 30,000g, processing and maturation occurred, as expected. Thus, the first two endonucleolytic cleavages occurred at the sites indicated by the two arrows. The nuclease generating the 5' terminus of the tRNA had been described (*18, 31*), but the nuclease acting beyond the C-C-A sequence was new and it evidently recognized the hairpin structure shown. The seven nucleotides remaining at the 3' end were demonstrated to be removed stepwise by another exonuclease (*32*). Base modifications, at least partial, were shown to occur in the resulting tRNA. The latter, as well as the primary transcript, were thoroughly characterized by their two dimensional chromatographic patterns and by comprehensive nearest-

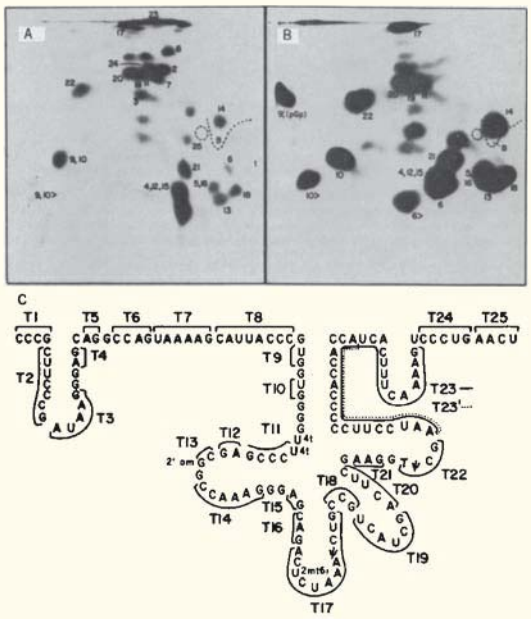


Fig. 18 (left). Two-dimensional oligonucleotide patterns, (A) and (B), of the products obtained on T1 ribonuclease digestion of the primary transcript (transcription experiment performed as in Fig. 16) and of the tRNA obtained by processing with *E. coli* extracts. The oligonucleotides to be expected in the primary transcript and in the processed transcript are shown systematically in (C). In the sequence are shown modified nucleosides, as present in mature $tRNA_1^{Tyr}$. In the primary transcript, the modifications are, of course, absent and the parent nucleosides are present. [Reproduced by permission of Academic Press, New York]

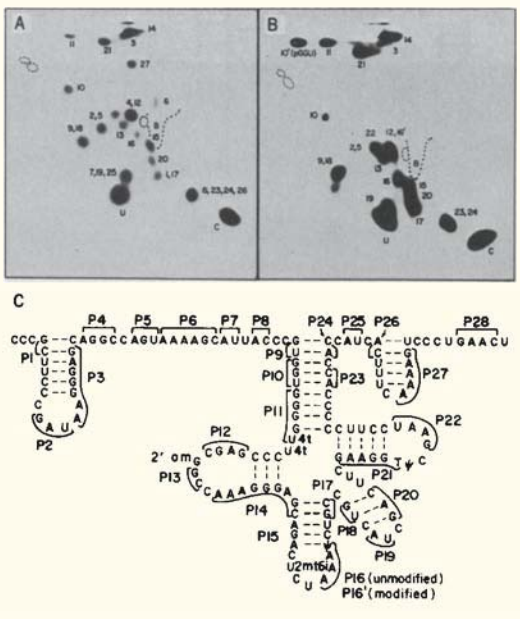


Fig. 19 (right). Two-dimensional oligonucleotide patterns, (A) and (B) respectively, of the products obtained by pancreatic ribonuclease digestion of the primary transcript (the transcription experiment was performed as described for Fig. 16) and of the tRNA obtained by processing with *E. coli* extracts. The oligonucleotides to be expected are shown schematically in (C). As in Fig. 18, the sequence contains modified nucleosides as present in mature $tRNA_1^{Tyr}$. The modifications are absent in the primary transcript. [Reproduced by permission of Academic Press, New York]

neighbor analyses of the oligonucleotides produced by digestion with T1 and pancreatic ribonucleases (Figs. 18 and 19). All the data were completely consistent with those expected for the total sequence. Full analyses of the transcripts provided sensitive means of confirming the accuracy of the synthetic work, at least in the transcribed part.

Identical transcription experiments were next performed to study transcription of the synthetic gene containing the promoter region (Fig. 15). Transcription was indeed found to be determined by the promoter in that it was initiated with pppG as the 5' end nucleotide

and it was followed by the sequence which is found naturally at the 5' end. Furthermore, the transcription was strand-specific.

Biological Activity of the Synthetic Gene (Suppression of Amber Mutations in vivo)

Suppression of amber (non-sense) mutations was a decisive experiment to be performed with the synthetic gene. Indeed, the demonstration of suppression and the clear-cut demonstration that this involved a single nucleotide change in the anticodon of the tyrosine tRNA (change in anticodon from 5'---G*UA---3' to 5'-----CUA---3') to enable recognition of the amber codon, UAG, was a strong argument for the choice of tyrosine suppressor tRNA as the synthetic target.

A number of experiments performed with the synthetic gene demonstrated its ability to suppress amber mutations in bacteria and in bacteriophages *(33)*. The protocol for a typical experiment is shown in Fig. 20. Bacteriophage λ DNA carrying *A* am32 and *B*aml mutations (*34*) was digested with Eco R1 endonuclease. A nonessential segment was thus excised (Fig. 20). The synthetic gene was then covalently inserted into the linear bacteriophage DNA by using T4 polynucleotide ligase (the insertion must have occurred randomly in regard to direction). The resulting circular molecules (Fig. 20) were used to transform *E. coli* bacterial cultures. As seen in Fig. 21, suppression of the two amber mutations in the phage DNA and consequent formation of phage plaques were observed only with the phage DNA carrying the synthetic suppressor gene.

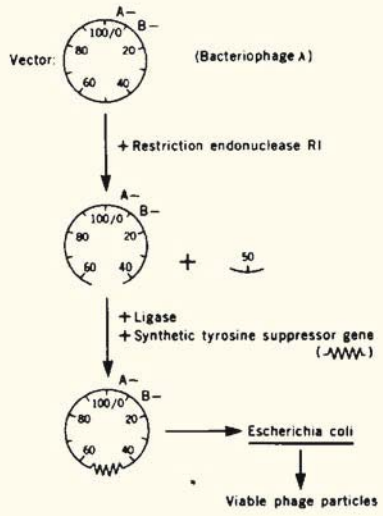


Fig. 20. Cloning of the synthetic gene for the tyrosine suppressor tRNA. The vector used was a derivative of bacteriophage λ with two amber mutations (A^- and B^-). Digestion with restriction endonuclease R1 excised a nonessential piece from the middle of the total genome. Subsequent insertion of the synthetic gene (WWW) with the use of polynucleotide ligase gave the circular phage containing the synthetic gene. This, on transfection of *E. coli*, produced viable phage particles. The results are shown in Fig. 21. [Reproduced by permission of Academic Press, New York]

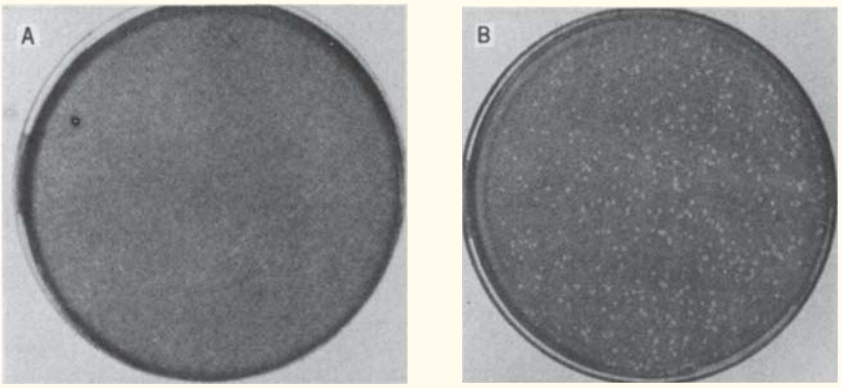


Fig. 21. Experiment showing phage growth in *E. coli* following insertion of the synthetic amber suppressor gene. The two culture dishes show (A) bacteriophage λ with amber mutation used as the vector for the synthetic gene and (B) the same vector with inserted suppressor tRNA gene, prepared as in Fig. 20. As is seen, phage plaques are present only in (B). [Reproduced by permission of Academic Press, New York]

Conclusion

How can a gene be properly defined for the purpose of a total synthesis? Genes comprise parts of a continuum of DNA and most often individual genes are present as clusters. Thus, one set of control elements may govern transcription, or translation, of one such group. In the specific case of the synthetic tyrosine suppressor tRNA gene, at least two comments may be made: (i) The synthetic promoter gene initiates transcription as is found in vivo and, therefore, is "functional." However, the possibility cannot be excluded that there might be an additional regulatory region which modulates the activity of the promoter. (ii) Downstream from the C-C-A end of the tRNA structural gene, a long region is now known (*19*) to precede the signal for termination of transcription. The work described here has not been taken as far as the end of the transcriptional unit; the discovery of the processing signal soon after the C-C-A end made this unnecessary. For every conceivable aim in structure-function studies of the tRNA, the synthetic gene appears to be completely satisfactory. Thus, initiation of transcription is controlled, the transcript contains the two sites where processing occurs at the two ends, and the product undergoes maturation by extensive base modifications.

Total synthesis of a DNA containing specific nucleotide sequences differs conceptually from the in vivo or in vitro DNA replication catalyzed by enzymes that polymerize nucleic acids. The latter require a preformed template strand to bring about

polynucleotide synthesis; and, therefore, the important biological function served is the faithful replication of information rather than the creation of new information. In total synthesis, as is illustrated by the work described here, both strands are constructed in response to a predetermined sequence. Since the sequence is generated by chemical synthesis, there is full choice on the subssquent manipulation of the sequence information by designed chemical change. This ability is the essence of the chemical approach to the study of biological specificity in DNA and RNA. Thus, the work described here provides a new and systematic approach to the study of structure-functional relations in tRNA molecules. Other structural genes may be studied in the same way. Further, new opportunities open up for studies of protein-nucleic acid interactions involved in a myriad of regulatory functions, such as RNA interactions between polymerase and promoter, and between operator and repressor (*35*). Until the recent dramatic progress in the determination of DNA sequences, the degrees of freedom in applying synthesis to such problems were too great. However, now that the framework in protein-nucleic acid interactions is being narrowed down, synthesis can provide systematic approaches to the precise studies of the control regions in different systems.

The synthetic approach is now being used increasingly in such studies (*36*) and may often be the only one that enables a deeper understanding of biological controls at molecular level. A further important area with large potential is the understanding of transcription in *E. coli*. All the promoters in the genome of this organism and at least a few of the promoters in the infecting viruses are evidently recognized by one RNA polymerase. Despite the fact that some 25 promoters have been sequenced, the structural features important in (i) the recognition of the promoter region by the RNA polymerase, (ii) the formation of the stable complex, and (iii) the point of initiation of transcription remain poorly understood. Systematic synthetic modifications in one promoter may prove to be the only approach toward dissecting the role of different sections comprising the promoter region.

Finally, in the general field of work involving recombinant DNA, the present methodology may complement the genetic approaches. For example, it may not be necessary to synthesize full lengths of the genes with their own control elements as was done in our work. Instead, shorter, more easily manageable segments corresponding to the necessary gene products—for example, peptide hormones—may be synthesized and inserted into the plasmids or vectors (*37*).

Work on the chemical-enzymatic synthesis of bihelical DNA began in the middle 1960's and its progress has been made possible by enthusiastic collaboration between a large number of organic chemists and biochemists (*38*).

References and Notes

[1] H. G. Khorana, G. M. Tener, J. G. Moffatt, E. H. Pol, *Chem. Ind. (London)* (1956), p.1523.

[2] D. M. Brown and A. R. Todd, in *Nucleic Acids,* E. Chargaff and J. N. Davidson, Eds. (Academic Press, New York, 1955), vol. 1.

[3] J. D. Watson and F. H. C. Crick, *Nature (London)* 171, 737 (1953).

[4] H. G. Khorana, *Pure Appl. Chem.* 17, 349 (1968).

[5] ——, *Fed. Proc. Fed. Soc. Exp. Biol.* 19, 931 (1960).

[6] ——, in *Harvey Lect., Ser.* 62 (1968), p. 79.

[7] ——, *Biochem. J.* 109, 709 (1968).

[8] M. Gellert, *Proc. Natl. Acad. Sci. U.S.A.* 57, 148 (1967); S. B. Zimmerman, J. W. Little, C.K. Oshinsky, M. Gellert, *ibid.,* p. 1841; B. Weiss and C. C. Richardson, *ibid.,* p. 1021; B.M. Olivera and I. R. Lehman, *ibid.,* p. 1426; M.L.Gefter, A. Becker, J. Hurwitz, *ibid.* 58, 240 (1967); N. R. Cozzarelli, N. R. Melechen, T. M. Jovin, A. Kornberg, *Biochem. Biophys. Res. Commun.* 28, 578 (1967).

[9] C. C. Richardson, *Proc. Natl. Acad. Sci. U.S.A.* 54, 158 (1965).

[10] N. K. Gupta, E. Ohtsuka, H. Weber, S. H. Chang, H. G. Khorana, *ibid.* 60, 285 (1968); N. K. Gupta, E. Ohtsuka, V. Sgaramella, H. Büchi, A. Kumar, H. Weber, H. G. Khorana, *ibid.,* p. 1338.

[11] K. L. Agarwal *et al., Nature (London)* 227, 27 (1970); H. G. Khorana *et al., J. Mol. Biol.* 72, 209 (1972), and accompanying papers.

[12] H.-J. Fritz, R. Belagaje, E. L. Brown, R. H. Fritz, R. A. Jones, R. G. Lees, H. G. Khorana, *Biochemistry,* in press.

[13] R. A. Jones, H.-J. Fritz, H. G. Khorana, *ibid.,* in press.

[14] R. W. Holley *et al., Science* 147, 1462 (1965).

[15] H. M. Goodman, J. N. Abelson, A. Landy, S. Brenner, J. D. Smith, *Nature (London)* 217, 1019 (1968); U. L. RajBhandary, S. H. Chang, H. J. Gross, F. Harada, F. Kimura, S. Nishimura, *Fed. Proc. Fed. Soc. Exp. Biol.* 28, 409 (1969).

[16] J. D. Smith, J. N. Abelson, B. F. C. Clark, H.M. Goodman, S. Brenner, *Cold Spring Harbor Symp. Quant. Biol.* 31, 479 (1966); R. L. Russell, J. N. Abelson, A. Landy, M. L. Gefter, S. Brenner, J. D. Smith, *J. Mol. Biol. 47,* 1 (1970).

[17] J. D. Smith, *Br. Med. Bull.* 92, 220 (1973).

[18] S. Altman, *Nature (London) New Biol.* 229, 19 (1971); —- and J. D. Smith, *ibid.* 233, 35 (1971).

[19] H. Küpper, T. Sekiya, M. Rosenberg, J. Egan, A. Landy, *Nature (London),* in press.

[20] H. G. Khorana *et al., J. Biol. Chem.* 251, 565 (1976), and accompanying papers.

[21] V. Sgaramella and H. G. Khorana, *J. Mol. Biol.* 72, 427 (1972); J. H. van de Sande, M. H. Caruthers, V. Sgaramella, T. Yamada, H. G. Khorana, *ibid.,* p. 457; P. C. Loewen, R. C. Miller, A. Panet, T. Sekiya, H. G. Khorana, *J. Biol. Chem.* 251, 642 (1976).

[22] T. Sekiya, P. Besmer, T. Takeya, H. G. Khorana, *J. Biol. Chem.* 251, 634 (1976), and succeeding papers.

[23] M. H. Caruthers, R. Kleppe, K. Kleppe, H. G. Khorana, *ibid., p.* 568.

[24] R. Kleppe *et al., ibid.,* p. 667.

[25] P. C. Loewen and H. G. Khorana, *ibid.* 248, 3489 (1973); P. C. Loewen, T. Sekiya, H. G. Khorana, *ibid.* 249, 217 (1974).

[26] T. Sekiya and H. G. Khorana, *Proc. Natl. Acad. Sci. U.S.A.* 71, 2978 (1974); T. Sekiya, M.J. Gait, K. Norris, B. Ramamoorthy, H. G. Khorana, *J. Biol. Chem.* 251, 4481 (1976); T. Sekiya, R. Contreras, H. Küpper, A. Landy, H. G.Khorana, *ibid.,* p. 5124.

[27] R. C. Miller, P. Besmer, H. G. Khorana, M. Fiandt, W. Szybalski, *J. Mol. Biol.* 56, 363 (1971).

[28] T. Sekiya, T. Takeya, R. Contreras, H. Küpper, H.G. Khorana, A. Landy, *in RNA Polymerase,* R. Losick and M. Chamberlin, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1976), pp. 455-472.

[29] B. Ramamoorthy, R. G. Lees, D. G. Kleid, H. G.Khorana, *J. Biol. Chem.* 251, 676 (1976).

[30] T. Terao, J. E. Dahlberg, H. G. Khorana, *ibid.* 247, 6157 (1972); R. Contreras, unpublished results.

[31] H.D. Robertson, S. Altman, J. D. Smith, *J. Biol. Chem.* 247, 5243 (1972).

[32] E.K. Bikoff and M. L. Gefter, *ibid.* 251, 6240 (1976).

[33] M. J. Ryan *et al., Fed. Proc. Fed. Am. Soc. Exp. Biol.* 36 (No. 3), 732 (abstr.) (1977).

[34] F.R. Blattner *et al., Science* 196, 161 (1977).

[35] W. Gilbert, *in RNA Polymerase,* R. Losick and M. Chamberlin, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1976), pp. 193-205; R. Ogata and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* 74, 4973 (1977); C. P. Bahl, R. Wu, J. Stawinsky, S. A. Narang, *ibid.,* p. 966; D. V. Goeddel, D. G. Yansura, M. H. Caruthers, *ibid.,* p. 3292.

[36] D. V. Goeddel, D. G. Yansura, M. H. Caruthers, *Nucleic Acids Res.* 4, 4049 (1977).

[37] K. Itakura, T. Hirose, R. Crea, A. D. Riggs, H. L. Heyneker, F. Bolivar, H. W. Boyer, *Science* 198, 1056 (1977).

[38] Colleagues participating in earlier published work have been mentioned in the literature citations given above. For more recent work, it has been the most rewarding experience for me to have been associated with Ramamoorthy Belagaje, Eugene L. Brown, Roland Contreras, Hans-Joachim Fritz, Michael J. Gait, Robert G. Lees, Kjeld Norris, Michael J. Ryan, Takao Sekiya, and Tatsuo Takeya.

[39] Supported by grants from the National Institutes of Health, Public Health Service (CA11981); National Science Foundation (PCM73-06757), and American Cancer Society (NP-140).