13. J. G. Tully, R. F. Whitcomb, H F. Clark, D. L. Williamson, *Science* **195**, 892 (1977).
14. H-2 medium consists of two parts TNM-FH medium to one part DCCM medium to one part M1D medium.
15. J. G. Tully and R. F. Whitcomb, in *The Prokaryotes*, M. P. Starr *et al.*, Eds. (Springer-Verlag, New York, 1981), vol. 2, p. 2278.
16. D. E. Lynn, S. G. Miller, H. Oberlander, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2589 (1982).
17. Homologous antiserum was prepared against WSRO's collected from the hemolymph of infected *Drosophila* flies (*3*).
18. D. L. Williamson, J. G. Tully, R. F. Whitcomb, *Int. J. Syst. Bacteriol.* **29**, 345 (1979).
19. D. L. Williamson, personal communication.
20. C. E. Yunker, J. G. Tully, J. Cory, in preparation.
21. Supported by Binational Agricultural Research and Development agreement 58-32R6-3-157. We thank G. McGarrity, H. Neimark, and J. Werren for suggestions in the preparation of the manuscript, and R. Henegar and V. Bray for care of the cultures.

13 December 1985; accepted 27 March 1986

# Non–Watson-Crick G · C and A · T Base Pairs in a DNA-Antibiotic Complex

GARY J. QUIGLEY, GIOVANNI UGHETTO, GIJS A. VAN DER MAREL, JACQUES H. VAN BOOM, ANDREW H.-J. WANG, ALEXANDER RICH

**The structure of a DNA octamer d(GCGTACGC) cocrystallized with the bis-intercalator antibiotic triostin A has been solved. The DNA forms an unwound right-handed double helix. Four base pairs are of the Watson-Crick type while four are Hoogsteen base pairs, including two A · T and two G · C base pairs. This is the first observation in an oligonucleotide of Hoogsteen G · C base pairs where the cystosine is protonated. It is likely that these also occur in solutions of DNA complexed to this antibiotic.**

IN RECENT YEARS WE HAVE BECOME aware of the large number of conformations that the DNA double helix can adopt. It is now clear that specialized nucleotide sequences in DNA may facilitate alternative DNA conformations. In addition to right-handed conformations, a left-handed double-helical conformation has been found, especially in sequences with alternating purines and pyrimidines (*1*). Furthermore, DNA sequences containing long stretches of purines on one strand and pyrimidines on the other may also adopt alternative conformations (*2*). An additional mode of conformational variability is found in the hydrogen bonding interaction between the bases. It has been known for some time that types of base pairing other than the Watson-Crick type are also possible. Recent work suggested that when quinoxaline antibiotics interact with DNA they can alter the type of hydrogen bonding found in the base pairs (*3, 4*). We now show that both A · T and G · C base pairs can exist in the same segment of DNA in both Watson-Crick and non–Watson-Crick conformations when the DNA is bound to antibiotics of this type.

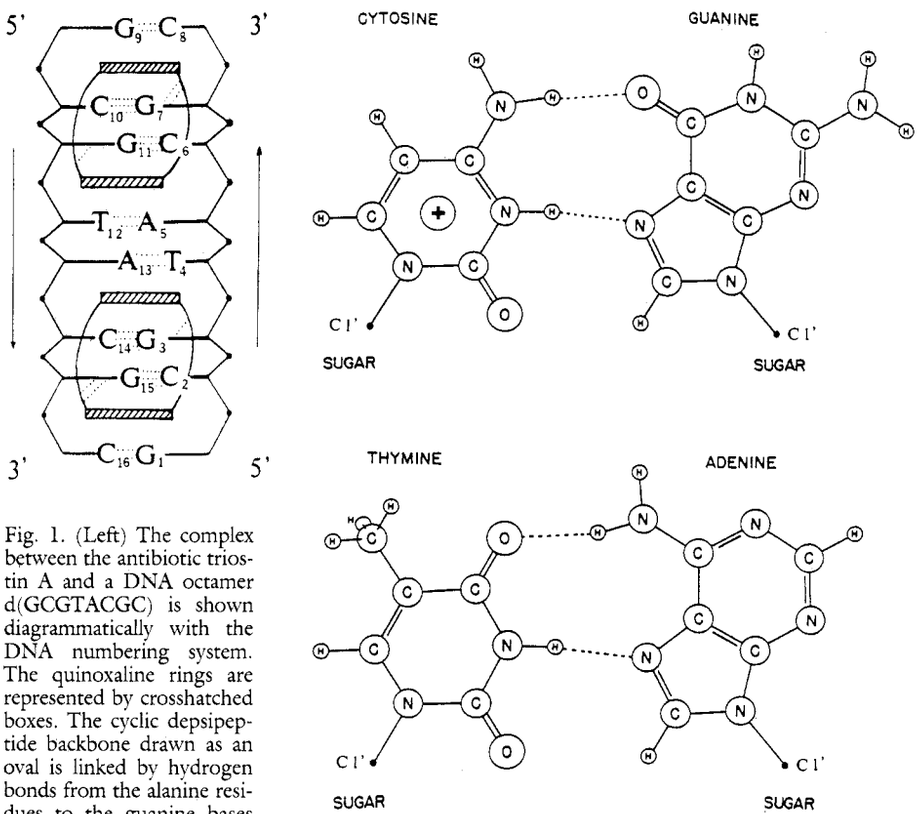


Fig. 1. (Left) The complex between the antibiotic triostin A and a DNA octamer d(GCGTACGC) is shown diagrammatically with the DNA numbering system. The quinoxaline rings are represented by crosshatched boxes. The cyclic depsipeptide backbone drawn as an oval is linked by hydrogen bonds from the alanine residues to the guanine bases (dotted lines). (Right) Schematic drawings show the G · C as well as the A · T Hoogsteen base pairs. The G · C Hoogsteen base pairing requires a protonated cytosine.

Triostin A, a cyclic octadepsipeptide antitumor antibiotic containing two quinoxaline rings, has been cocrystallized with the DNA octamer d(GCGTACGC) and its structure has been solved at near-atomic resolution by x-ray analysis. Two triostin A molecules bind to the DNA octamer with the quinoxaline rings intercalating as shown (Fig. 1, left). The central A · T base pairs in the complex are held together by Hoogsteen rather than Watson-Crick hydrogen bonds, as when triostin A was complexed with the DNA hexamer d(CGTACG) (*3, 4*). The cyclic depsipeptide of the drug molecule lies in the minor groove of the double helix, where it hydrogen bonds in a sequence-specific manner to two CpG base pairs that are surrounded by the bis-intercalating quinoxaline rings. In the octamer complex the outer G · C base pairs are also held together by Hoogsteen hydrogen bonds even though these require protonation of the cytosine residues (Fig. 1, right). Nevertheless, these crystals form readily at *p*H 6.5. This structure indicates that DNA is able to accommodate both Watson-Crick and Hoogsteen base pairs in a right-handed double helical structure at the same time. Conformational changes of this type, in which purine residues are in the *syn* conformation, may occur in a wide variety of alternative DNA forms.

Footprinting experiments of triostin A and its close relative echinomycin have been carried out on plasmid DNA fragments (*5–7*). These experiments showed that the

G. J. Quigley, G. Ughetto, A. H.-J. Wang, A. Rich, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.
G. A. van der Marel and J. H. van Boom, Gorlaeus Laboratories, Leiden State University, Leiden, 2300RA, The Netherlands.

tightest binding occurs with tetranucleotides with the sequence $\overset{A}{T}$ CG $\overset{A}{T}$; that is, preferentially with a central CpG sequence and either A · T or T · A base pairs flanking it. However, some sites were also found with sequences that had G or C in the first or fourth position of the sequence. Since the triostin A–DNA as well as the echinomycin-DNA hexamer complexes showed Hoogsteen base pairing in the central A · T base pairs, it was inferred that the binding specificity reflected the fact that flanking A · T base pairs on either side of the CpG observed in the footprinting experiments were associated with Hoogsteen base pairs. A · T can form Hoogsteen base pairs readily since purines can adopt the *syn* conformation easily and no protonation is required. However, G · C base pairs in the Hoogsteen conformation are less stable than Watson-Crick pairs since protonation of cytosine would be required and there is one fewer hydrogen bond (Fig. 1, right). For this reason, a DNA octamer was synthesized that has flanking terminal C · G base pairs at either end of the complex to see whether this base pair was capable of adopting a Hoogsteen conformation. Attempts at cocrystallization were carried out at a variety of *p*H conditions. Crystals could be formed at all *p*H values examined between *p*H 3.5 and 6.5. The difference due to the varying *p*H conditions was that the crystals seemed to form more consistently at the lower *p*H, which may reflect the need for protonation.

The best crystals were formed by vapor-phase equilibration with a 30 percent 2-methyl-2,4-pentanediol (2-MPD) solution from a solution containing 1.2 m*M* DNA octamer, 20 m*M* sodium acetate (*p*H 4.5), 10 m*M* NaCl, 8 m*M* $MgCl_2$, 1 m*M* spermine tetrachloride, 1.1 m*M* triostin A (in 1:1 methanol and chloroform), and 12 percent 2-MPD with a total volume of 40 μl. Clusters of hexagonal rodlike crystals grew slowly in the droplet and the largest of them attained a length of 0.5 mm and a diameter of 0.3 mm. The molecules crystallized in a hexagonal unit cell in the space group $P6_322$ with the dimensions $a = b = 40.9$ Å, $c = 80.7$ Å. The crystals diffracted to 2.0 Å resolution along the *c*-axis direction and slightly less along the *a* and *b* directions. The data were collected on a Nicolet P3 x-ray diffractometer at 15°C with CuKα radiation in the omega scan mode. The crystal diffracted x-rays strongly to 2.5 Å resolution with about 70 percent of the reflections collected being greater than the 1.5 σ intensity ($I$) level. Beyond that resolution the intensity dropped off rapidly. The strong 0, 0, 24 reflection suggested that there were two octamer complexes with 24 planar groups stacked 3.36 Å apart along the *c* axis. On the basis of the symmetry considerations, we inferred that the DNA octamer complex lies on a crystallographic twofold axis with the asymmetric unit consisting of four base pairs and one triostin A molecule. There were two possible alternatives for this position in the unit cell. A model was constructed with the coordinates from the previously solved hexamer structure (*3*). The model was built into the lattice and a rotation-translation search was then carried out. The result showed that the twofold axis at $z = 0$ is slightly favored and this solution was confirmed by the refinement. The structure was refined by the Konnert-Hendrickson constrained refinement procedure (*8*) and the current R factor is 20.0 percent at a resolution of 2.25 Å with 1130 reflections [$I > 1.5\ \sigma(I)$]. To discriminate which base-pairing scheme was adopted for the terminal G · C base pairs, a difference Fourier map was calculated by deleting the terminal deoxyguanosine and deoxycytidine groups as well as any solvent molecules within 5 Å of the base pairs from the model. This unbiased map showed that the Hoogsteen G · C base pair fitted the difference electron density map.

At this stage of the refinement 91 water molecules are seen in the asymmetric unit and no ion has been identified. In the hexagonal lattice, the molecules are stacked in an end-to-end manner along the twofold screw axis in the *c*-axis direction and they associate laterally in a fashion that leaves large solvent channels parallel to the *c* axis. There are two types of channels, one channel around the sixfold axis at the origin having a diameter of 30 Å and a smaller channel around the threefold axis with a diameter near 10 Å.

The quinoxaline rings intercalate between the outer GpC and between the inner GpT sequences (Fig. 1, left). The cyclic peptide drawn diagrammatically as an oval is linked by hydrogen bonds from the alanine residues to the guanine bases in the minor groove (dotted lines).

A stereo diagram showing the two antibiotics bound to the right-handed DNA octamer (Fig. 2) shows the cyclic peptide of triostin A filling most of the minor groove at the bottom of the diagram. Further details are shown in a skeletal drawing of the bottom portion of the DNA-antibiotic complex as viewed from the major groove of the DNA double helix (Fig. 3). Triostin A with solid bonds is sitting behind the DNA duplex, which is drawn with open bonds. It can be seen that the molecule is a true bis-intercalator with both quinoxaline rings inserted between base pairs. The upper quinoxaline ring is stacked largely on the *syn* A13 of the A13-T4 Hoogsteen base pair as well as the Watson-Crick base pair G3-C14. The details of this interaction are somewhat simi-

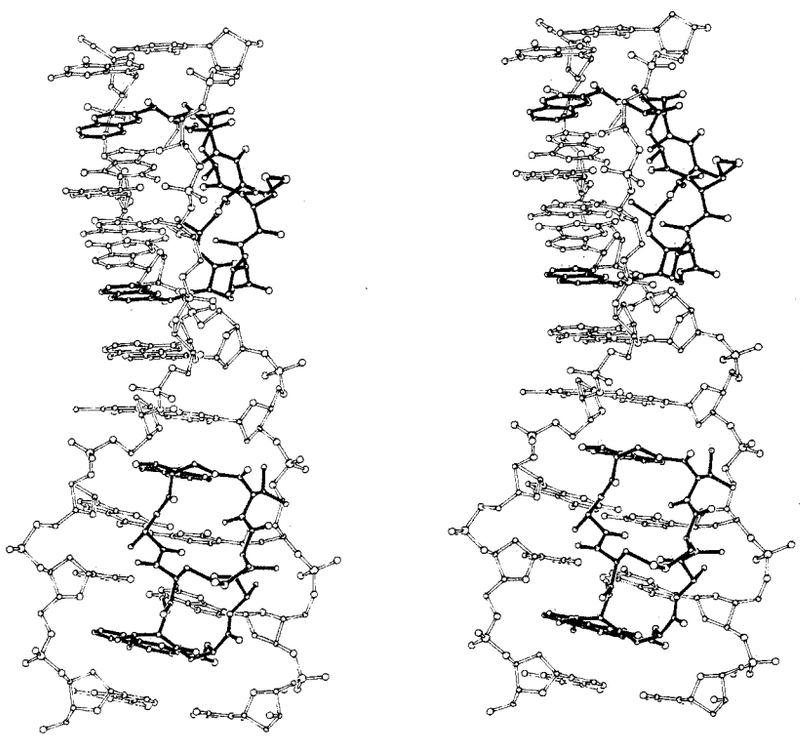

Fig. 2. A stereo diagram showing triostin A (solid bonds) complexed to the right-handed DNA octamer (open bonds). The minor groove at the bottom of the diagram illustrates the extent to which the cyclic peptide fills the groove. The extent of intercalation by the quinoxaline rings is seen in the upper part of the diagram. The helix shows no apparent discontinuity when it changes from Watson-Crick to Hoogsteen pairing between the bases.

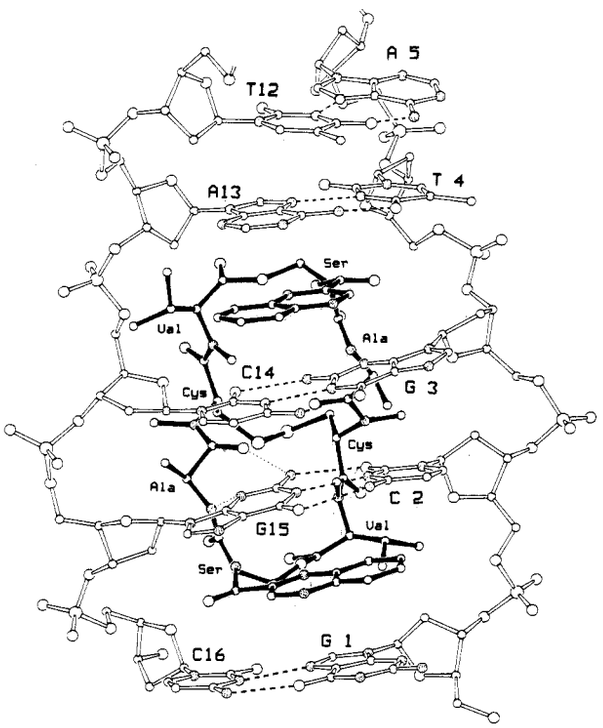


Fig. 3. A skeletal drawing of slightly more than one-half of the DNA octamer–triostin A complex is shown as viewed from the major groove of the DNA double helix which is considerably unwound. The G · C and A · T base pairs flanking the quinoxaline rings of the drug molecule are in Hoogsteen conformation. The amino acids in the cyclic backbone of triostin A are valine (Val), cysteine (Cys), serine (Ser), and alanine (Ala).

lar to those observed in the DNA hexamer complex (*3*). The lower quinoxaline ring (Fig. 2) is stacked over the bottom Hoogsteen base pair G1-C16. The G1 is in the *syn* conformation and it makes two hydrogen bonds with C16 (shown diagrammatically in Fig. 1, right). The quinoxaline ring is stacked largely over both six- and five-membered rings of guanine G1, in contrast to the stacking of the upper quinoxaline ring, which is stacked only over the six-membered ring of adenine A13. The lower quinoxaline ring also interacts with the Watson-Crick base pair C2-G15 above it.

The specificity of the interaction between the triostin A and DNA is similar to that seen in the DNA hexamer complex (*3*) in that there are two hydrogen bonds between alanine (NH and carbonyl oxygen) on the left and guanine G15 (N3 and N2), while the alanine on the right has a single hydrogen bond with N3 of guanine G3. In addition to hydrogen bonding and intercalating interactions, there are 27 van der Waals interactions (<3.5 Å) between the cyclic peptide and the DNA molecule to further stabilize the structure. Although the conformation of the triostin A is generally similar to that seen in the DNA hexamer complex, there are differences in details. For example, the orientation of the carbonyl groups of the D-serine residues at either end of the peptide differ somewhat from that which is seen in the hexamer complex. This suggests that when the antibiotic is forming a true bis-intercalator complex there are small adjustments in the peptide backbone associated with it. The depsipeptide backbone of triostin A contains two ester linkages involving D-serine residues at both the top and the bottom of the rectangular complex shown in Fig. 3. The longer vertical sides are made of peptide backbones that are in an antiparallel orientation. The direction of the peptide backbone from $NH_2$ to COOH is opposite to the 5′ to 3′ direction of the adjacent DNA backbone.

This structure resolves the issue of the manner in which C · G base pairs interact with quinoxaline antibiotics when they are next to the intercalating quinoxaline rings. They adopt a Hoogsteen pairing even though it requires protonation of the cytosine base. This probably occurs because of the considerable stability from the van der Waals interactions between the lower quinoxaline rings and the sugar-phosphate backbones associated with the G1-C16 base pair. Hoogsteen base pairs have a C1′-C1′ distance (Fig. 1, right) of 8.6 Å, in contrast to the distance of 10.5 Å associated with Watson-Crick base pairs. Thus the sugar phosphate backbones are almost 2 Å closer when Hoogsteen base pairing is used in a double helix compared to the distance found with Watson-Crick base pairs. The present structure shows that this is accommodated even in guanine-cytosine base pairs with a protonated cytosine. The $pK_a$ of cytosine is 4.6 (*9*). In polynucleotide structures in which cytosine is protonated, there is usually a shift in the $pK_a$ associated with the stable structure. For example, in the double helical structure of poly(ribocytidylic) acid in solution, it has been shown that the cytosines pair with each other so that one cytosine is protonated (*10*). In that case, the $pK_a$ shifts from 4.6 to 5.7, a change of 1.1 *p*K units. In our structure, it is possible that the van der Waals interactions around the quinoxaline ring also stabilize the formation of Hoogsteen base pairs so that there might be a comparable rise in the $pK_a$ of this cytosine. The fact that the crystals grow at *p*H 6.5, and the fact that quinoxaline antibiotics complexes of DNA have been footprinted at neutral *p*H, imply that a stable structure is formed. It should be interesting to determine the change in the $pK_a$ of cytosine associated with this complex formation. The contribution of the imino tautomer of cytosine, particularly at higher *p*H, cannot be totally ruled out. A substantially higher resolution structure, neutron diffraction studies, or both would help clarify this question.

The protonation of cytosine is selective and structurally dependent. The cystosines C8 and C16 are protonated but the other four cytosines in the octamer duplex adopt a normal Watson-Crick base pairing, implying they are not protonated. In this structure, four of the eight base pairs are in Watson-Crick and four are in Hoogsteen geometries. This shows how protonation is influenced by the structural environment of the nucleic acid. This structure raises several questions. It is of interest to ask how many Hoogsteen base pairs are formed near the binding site when quinoxaline antibiotics bind to a DNA polynucleotide. For example, in a complex between a DNA decamer d(GCGTATACGC) and triostin A, what would be the conformation of the central TATA sequence? Further experiments will be needed to answer such questions. As noted above the C1′-C1′ distance (Fig. 1, right) in Watson-Crick base pairs is near 10.5 Å as opposed to 8.6 Å for a Hoogsteen base pair (*3, 4*). However, it is clear from the footprinting experiments carried out on quinoxaline antibiotics bound to macromolecular DNA (*5–7*) that there is likely to be at some point along the DNA a conversion from Hoogsteen base pairs to Watson-Crick base pairs. The nature of that interface would also be of considerable interest in view of the differences cited above.

This complex illustrates the extent to which DNA is polymorphic. In the present study, four of the purines are in the *syn* conformation and four in the *anti* conformation. Purines are known to form *syn* conformations readily; this is one of the major structural features underlying the formation of the left-handed Z-DNA conformation in which every other residue (usually but not always a purine) is in the *syn* conformation (*1, 11*). DNA models have been proposed in which some purines are in the *syn* conformation and Hoogsteen base pair-

ing is found (*2*, *12*). This study shows that Hoogsteen base pairs can coexist with Watson-Crick base pairs and it reinforces the possibility that such DNA polymorphism may be involved in various biological phenomena.


REFERENCES AND NOTES

1. A. H.-J. Wang *et al.*, *Nature (London)* **282**, 680 (1979).
2. D. E. Pulleyblank, D. B. Haniford, A. R. Morgan, *Cell* **42**, 271 (1985).
3. A. H.-J. Wang *et al.*, *Science* **225**, 1115 (1984).
4. G. Ughetto *et al.*, *Nucleic Acids Res.* **13**, 2305 (1985).
5. M. M. Van Dyke and P. B. Dervan, *Science* **225**, 1122 (1984).
6. C. M. L. Low, H. R. Drew, M. J. Waring, *Nucleic Acids Res.* **12**, 4865 (1984).
7. C. M. L. Low, R. K. Olsen, M. J. Waring, *FEBS Lett.* **176**, 414 (1984).
8. W. A. Hendrickson and J. Konnert, in *Biomolecular Structure: Conformation, Function and Evolution*, R. Srinivasan, Ed. (Pergamon, Oxford, 1979), p. 43.
9. M. Windholz, Ed., *The Merck Index* (Merck and Co., Inc., Rahway, NJ, ed. 10, 1983).
10. K. A. Hartman, Jr., and A. Rich, *J. Am. Chem. Soc.* **87**, 2033 (1965).
11. A. Rich, A. Nordheim, A. H.-J. Wang, *Annu. Rev. Biochem.* **53**, 791 (1984).
12. H. R. Drew and R. E. Dickerson, *EMBO J.* **6**, 663 (1982).


13. Supported by grants from the National Institutes of Health, National Science Foundation, American Cancer Society, Office of Naval Research, National Aeronautics and Space Administration, and the Netherlands Organization for the Advancement of Pure Research. G.U. acknowledges support from NATO and Instituto Strutturistica Chimica, Consiglio Nazionale Delle Richerche (Italy). We thank T. Yoshida of Shionogi Co., Osaka, Japan, for the triostin A.

2 December 1985; accepted 18 March 1986


# Pertussis Toxin Gene: Nucleotide Sequence and Genetic Organization


CAMILLE LOCHT AND JERRY M. KEITH



**The current pertussis vaccines, although efficacious, in some instances produce undesirable side effects. Molecular engineering of pertussis toxin, the major protective antigen, could provide a safer, new generation of vaccines against whooping cough. As a first critical step in the development of such a vaccine, the complete nucleotide sequence of the pertussis toxin gene was determined and the amino acid sequences of the individual subunits were deduced. All five subunits are coded by closely linked cistrons. A promoter-like structure was found in the 5′-flanking region, suggesting that the toxin is expressed through a polycistronic messenger RNA. The order of the cistrons is S1, S2, S4, S5, and S3. All subunits contain signal peptides of variable length. The calculated molecular weights of the mature subunits are 26,024 for S1, 21,924 for S2, 21,873 for S3, 12,058 for S4, and 11,013 for S5. Subunits S2 and S3 share 70% amino acid homology and 75% nucleotide homology. Subunit S1 contains two regions of eight amino acids homologous to analogous regions in the A subunit of both cholera and *Escherichia coli* heat labile toxins.**


PERTUSSIS TOXIN (*1*) IS ONE OF THE various toxic components produced by virulent *Bordetella pertussis*, the microorganism that causes whooping cough. A wide variety of biological activities, such as histamine sensitization, insulin secretion, lymphocytosis promotion, and immunopotentiating effects can be attributed to this toxin (*2*). In addition, the toxin protects mice challenged intracerebrally (*3*, *4*) or by aerosol (*4*) with virulent *B. pertussis*. Pertussis toxin is, therefore, an important constituent in the vaccine against whooping cough and is included in the acellular component vaccines being tested and used in several countries (*5*). However, the toxin may also be the cause of the harmful side effects associated with the current vaccines (*6*). It may be possible to develop a new vaccine with reduced side effects by genetic manipulation of the toxin gene.

The toxin is structured in a hexamer composed of five dissimilar subunits, designated S1 through S5 relative to their electrophoretic migration in denaturing gels (*7*). Subunit S1 contains an enzymatic adenosine diphosphate (ADP)–ribosylation activity (*8*) and subunits S2 through S5 contain a target cell receptor binding activity (*9*). Thus, by analogy to other bacterial ADP-ribosylating toxins, pertussis toxin is structured in an A-B model (*7*, *10*) in which the A moiety contains the enzymatic activity and the B moiety the binding activity (*11*).

The isolation of a 4.5-kb DNA fragment containing the pertussis toxin S4 subunit gene was described previously (*12*). Because of the polycistronic nature of prokaryotic genes, and by analogy to other A-B structured bacterial toxins that have similar enzymatic activities (*13*, *14*), it seemed likely that the cistrons coding for all the pertussis toxin subunits would be linked. We therefore determined the nucleotide sequence of the 4.5-kb cloned fragment. Here, we present this sequence and demonstrate that it indeed codes for the entire pertussis toxin structural gene.

Digests with a variety of six base pair (bp)–specific restriction enzymes and DNA sequence analysis were used to establish the restriction map shown in Fig. 1a. DNA was sequenced by the base-specific chemical cleavage method (*15*) starting from the restriction sites Eco RI, Sal I, Sma I, and Bgl II and by the dideoxy chain termination method (*16*) after subcloning overlapping regions into vectors M13 mp18 and M13 mp19 (*17*). Because of the high C+G content of *B. pertussis* DNA, it was necessary to use both of these methods with a combination of 8% and 20% polyacrylamide–8*M* urea gels for sequence analysis. Each nucleotide has been sequenced in both directions an average of 4.13 times. The final consensus sequence of the sense strand is presented in Fig. 2. In agreement with the overall high C+G content of *B. pertussis* DNA (*18*), the entire sequence contains 62.2% C+G with 19.6% A, 33.8% C, 28.4% G, and 18.2% T in the sense strand.

The DNA sequence was translated in all six reading frames. On the basis of our previous data (*12*), the open reading frame (ORF) corresponding to the S4 subunit was identified and is shown in Fig. 1b. The assignment of the other subunits to their respective ORF's, as shown in Fig. 1, b and c, is based on the following evidence: size of ORF's, high coding probability, deduced amino acid composition, predicted molecular weights, ratios of acidic to basic amino acids, amino acid homology to other bacterial toxins, mapping of Tn*5*-induced mutations, and partial amino acid sequence.

Those ORF's long enough to code for any of the five toxin subunits were analyzed by the statistical TESTCODE algorithm designed to differentiate between real protein coding sequences and fortuitous open reading frames (*19*). The amino acid composition of each ORF with a high protein coding probability was calculated, starting from either the predicted amino terminus of the mature proteins or from the first amino acid of the mature protein determined by amino acid sequencing of subunits purified by high-performance liquid chromatography. These data were then compared with


Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute of Allergy and Infectious Diseases, Laboratory of Pathobiology, Molecular Pathobiology Section, Rocky Mountain Laboratories, Hamilton, MT 59840.