Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA

John R. Riordan, Johanna M. Rommens, Bat-sheva Kerem, Noa Alon, Richard Rozmahel,Zbyszko Grzelczak, Julian Zielenski, Si Lok, Natasa Plavsic, Jia-Ling Chou, Mitchell L. Drumm, Michael C. Iannuzzi, Francis S. Collins, Lap-Chee Tsui

Overlapping complementary DNA clones were isolated from epithelial cell libraries with a genomic DNA segment containing a portion of the putative cystic fibrosis (CF) locus, which is on chromosome 7. Transcripts, approximately 6500 nucleotides in size, were detectable in the tissues affected in patients with CF. The predicted protein consists of two similar motifs, each with (i) a domain having properties consistent with membrane association and (ii) a domain believed to be involved in ATP (adenosine triphosphate) binding. A deletion of three base pairs that results in the omission of a phenylalanine residue at the center of the first predicted nucleotide-binding domain was detected in CF patients.

Cystic fibrosis (CF) is an autosomal recessive genetic disorder affecting a number of organs, including the lung airways, pancreas, and sweat glands (*1*). Abnormally high electrical potential differences have been detected across the epithelial surfaces of the CF respiratory tract, including the trachea and nasal polyps, as well as across the walls of CF sweat gland secretory coils and reabsorptive ducts (*2*). The basic defect has been associated with decreased chloride ion conductance across the apical membrane of the epithelial cells (*3*). That the defect also appeared to persist in cultured cells derived from several epithelial tissues suggested that the CF gene is expressed in these cells (*4*). More recently, patch clamp studies showed that this defect is probably due to a failure of an outwardly rectifying anion channel to respond to phosphorylation by cyclic AMP–dependent protein kinase (protein kinase A) or protein kinase C (*5*). Although progress has been made in the isolation of polypeptide components of an epithelial chloride channel that mediates conductance (*6*), their relation to the kinase-activated pathway and CF has yet to be established, and the basic biochemical defect in CF remains unknown.

Molecular cloning experiments have permitted the isolation of a large, contiguous segment of DNA spanning at least four transcribed sequences from a region thought to contain the CF locus (*7*). These sequences were initially identified on the basis of their ability to detect conserved sequences in other animal species by DNA hybridization and were subsequently characterized by RNA hybridization experiments, cDNA isolation, and direct DNA sequence analysis (*7*). Three of the transcribed regions were excluded from being the CF locus by earlier genetic or DNA sequence analyses (*7, 8*). The fourth one, as shown by genetic analysis (*9*) and DNA sequencing analysis presented below, corresponds to a portion of the CF gene locus.

Isolation of cDNA clones. Two DNA segments (E4.3 and H1.6) that detected cross-species hybridization signals (*7*) were used as probes to screen cDNA libraries made from several tissues and cell types (*10*). After screening seven different libraries, one single clone (10-1) was isolated with H1.6 from a cDNA library made from the cultured epithelial cells of the sweat glands of an unaffected (non-CF) individual (*10*).

DNA sequencing showed that 10-1 contained an insert of 920 base pairs (bp) in size and one potential, long open reading frame (ORF). Since one end of the sequence shared perfect sequence identity with H1.6, it was concluded that the cDNA clone was probably derived from this region. The DNA sequence in common was, however, only 113 bp long (Figs. 1 and 2). This sequence in fact corresponded to the first axon of the putative CF gene. The short sequence overlap thus explained the weak hybridization signals in library screening and our inability to detect transcripts in RNA gel-blot analysis. In addition, the orientation of the transcription unit was tentatively established on the basis of alignment of the genomic DNA sequence with the presumptive ORF of 10-1.

Since the corresponding transcript was estimated to be about 6500 nucleotides in length by RNA gel-blot hybridization experiments, further cDNA library screening was required in order to clone the remainder of the coding region. As a result of several successive screenings with cDNA libraries generated from the colon carcinoma cell line T84, normal and CF sweat gland cells, pancreas,

J. R. Riordan, N. Alon, and Z. Grzelczak are in the Department of Biochemistry, Research Institute, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. J. M. Rommens, B. Kerem, R. Rozmahel, J. Zielenski, S. Lok, N. Plavsic, J.-L. Chou, and L.-C. Tsui are at the Department of Genetics, Research Institute, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. M. L. Drumm, M. C. Iannuzzi, and F. S. Collins are at the Howard Hughes Medical Institute and Departments of Internal Medicine and Human Genetics, University of Michigan, Ann Arbor, MI 48109. J. R. Riordan is also a member of the Departments of Biochemistry and Clinical Biochemistry, and L.-C. Tsui is also a member of the Departments of Medical Genetics and Medical Biophysics, University of Toronto, Toronto, Ontario M5S 1A8, Canada.

and adult lungs, 18 additional clones were isolated (Fig. 1). DNA sequence analysis revealed that none of these cDNA clones corresponded to the length of the observed transcript, but it was possible to derive a consensus sequence based on overlapping regions. Further cDNA clones corresponding to the 5′ and 3′ ends of the transcript were derived from 5′ and 3′ primer-extension experiments (Fig. 1). Together, these clones span about 6.1 kb and contain an ORF capable of encoding a protein of 1480 amino acids (Fig. 2).

It was unusual that most of the cDNA clones isolated here contained sequence insertions at various locations (Fig. 1). While many of these extra sequences corresponded to intron regions reverse-transcribed during the construction of the cDNA, as revealed on alignment with genomic DNA sequences, the identities of several others were uncertain because they did not align with sequences at the corresponding exon-intron junctions, namely, the sequences at the 5′ ends of clones 13a and T16-1 and at the 5′ and 3′ ends of T11, and the insertions between exons 3 and 4 in 13a and between exons 10 and 11 in T16-4.5 (legend to Fig. 1). More puzzling were the sequences corresponding to the reverse complement of exon 6 at the 5′ end of 11a and the insertion of a segment of a bacterial transposon in clone C16-1; none of these could be explained by mRNA processing errors.

In that the number of recombinant cDNA clones for the putative CF gene detected in the library screening was much less than would have been expected from the abundance of transcripts estimated from RNA hybridization experiments, it seemed probable that the clones that contained aberrant structures were preferentially retained while the proper clones were lost during propagation. Consistent with this interpretation, poor growth was observed for most of our recombinant clones isolated, regardless of the vector used.

RNA analysis. To visualize the transcript of the putative CF gene, we used RNA gel-blot hybridization with the 10-1 cDNA as the probe (Fig. 3). The analysis revealed a prominent band, about 6.5 kb in size, in T84 cells. Identical results were obtained with other cDNA clones as probes. Similar, strong hybridization signals were also detected in pancreas and primary cultures of cells from nasal polyps, suggesting that the mature mRNA of the putative CF gene is about 6.5 kb. Minor hybridization signals, probably representing degradation products, were detected at the lower size ranges, but they varied between different experiments. On the basis of the hybridization band intensity and comparison with those detected for other transcripts under identical experimental conditions, it was estimated that the putative CF gene transcripts constituted about 0.01 percent of total mRNA in T84 cells.

Additional tissues were analyzed by RNA gel-blot hybridization in an attempt to correlate the expression pattern of the putative CF gene and the pathology of CF. Transcripts, all of identical size, were found in lung, colon, sweat glands (cultured epithelial cells), placenta, liver, and parotid gland, but the signal in these tissues was generally weaker than that detected in the pancreas and nasal polyps (Fig. 3). Intensity varied among different preparations; for example, hybridization in kidney was not detectable in the preparation shown in Fig. 3 but was clearly discernible subsequently. Transcripts were not detected in the brain or adrenal gland, nor in skin fibroblast and lymphoblast cell lines.

Thus, expression of the putative CF gene appeared to occur in many of the tissues examined, with higher levels in those tissues severely affected in CF. While this epithelial tissue–specific expression pattern is in good agreement with the disease pathology, no significant difference was detected in the amount or size of transcripts from CF and control tissues (Fig. 3), consistent with the assumption that CF mutations are subtle changes at the nucleotide level.

Characterization of cDNA clones. As indicated above, a contig-

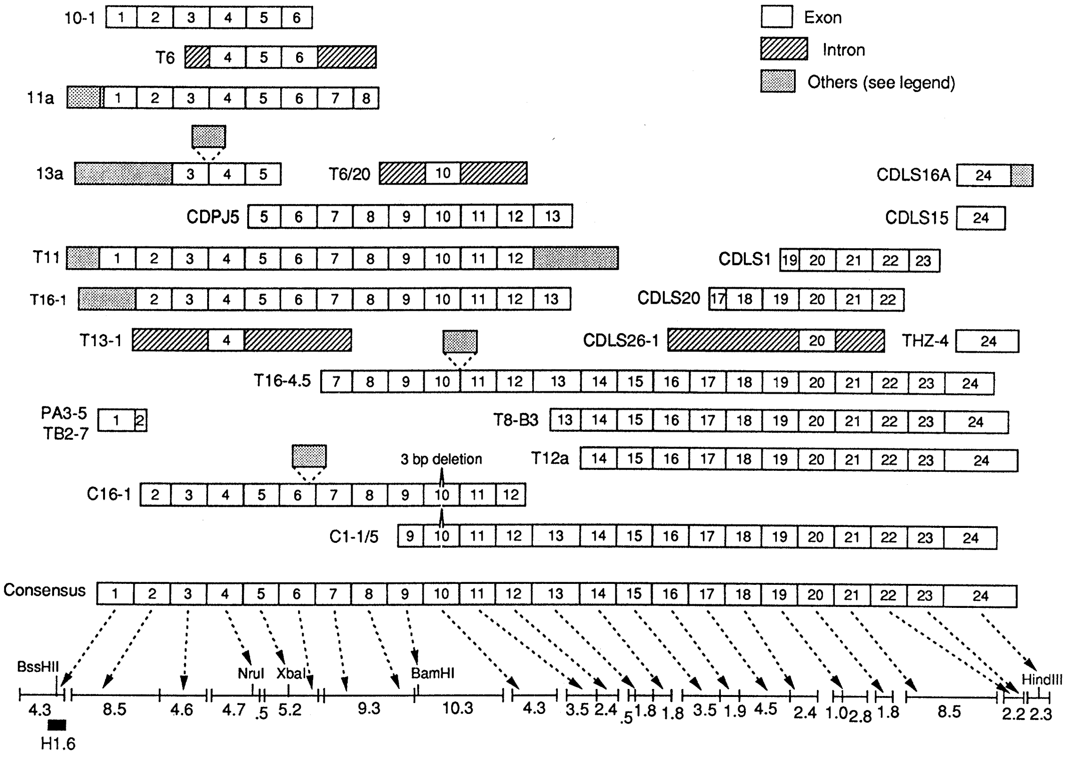
10-1 1 2 3 4 5 6
T6 4 5 6
11a 1 2 3 4 5 6 7 8
13a 3 4 5
T6/20 10
CDPJ5 5 6 7 8 9 10 11 12 13
T11 1 2 3 4 5 6 7 8 9 10 11 12
T16-1 2 3 4 5 6 7 8 9 10 11 12 13
T13-1 4
T16-4.5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
PA3-5
TB2-7 1 2
T8-B3 13 14 15 16 17 18 19 20 21 22 23 24
3 bp deletion
T12a 14 15 16 17 18 19 20 21 22 23 24
C16-1 2 3 4 5 6 7 8 9 10 11 12
C1-1/5 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Consensus 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Exon
Intron
Others (see legend)
CDLS16A 24
CDLS15 24
CDLS1 19 20 21 22 23
CDLS20 17 18 19 20 21 22
CDLS26-1 20
THZ-4 24
BssHII
NruI
XbaI
BamHI
HindIII
4.3 8.5 4.6 4.7 .5 5.2 9.3 10.3 4.3 3.5 2.4 .5 1.8 1.8 3.5 1.9 4.5 2.4 1.0 2.8 1.8 8.5 2.2 2.3
H1.6

Fig. 1. Overlapping cDNA clones aligned with genomic DNA fragments. The cDNA clones are represented by open boxes with exons indicated. The corresponding genomic Eco RI fragments are schematically presented on the bottom, with lengths in kilobases. The hatched boxes denote intron sequences, and stippled boxes represent other sequences as outlined below. The filled box in the lower left is the position of the clone H1.6, which was used to isolate the first cDNA clone 10-1 from a normal (N) sweat gland library (*10*). The definitive restriction sites used for the alignment of cDNA and genomic fragments are indicated. Clones T6, T6/20, T11, T16-1, T13-1, T16-4.5, T8-B3, and T12a were isolated sequentially from the T84 cell library (*10*). Clones isolated from the human lung cDNA library (*10*) are designated with the prefix CDL. CDPJ5 is derived from a pancreas library (*10*). The CF sweat gland cDNA clones, C16-1 and C1-1/5, together cover all but exon 1 and a portion of the 3′ untranslated region. Both clones revealed a 3-bp deletion in exon 10. Clones that contain intron sequences are CDLS26-1, T6/20, and T13-1. Clones T11, T16-4.5, CDLS16A, 11a, and 13a contain extraneous sequences of unknown origin at positions indicated. Clone C16-1 contains a short insertion corresponding to a portion of the γ transposon of *E. coli*. Both PA3-5 and TB2-7 are 5′ extension clones generated from pancreas and T84 RNA by the anchored PCR technique (*12*), respectively. THZ-4 is a 3′ extension clone obtained from T84 RNA. Both T12a and THZ-4 contain a polyadenylation signal and a poly(A)$^+$ tail.

```
   1  AATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCA

  61  GAGTAGTAGGTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGC

                   M  Q  R  S  P  L  E  K  A  S  V  V  S  K  L  F   16
 121  GCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACTTTTT

       F  S  W  T  R  P  I  L  R  K  G  Y  R  Q  R  L  E  L  S  D   36
 181  TTCAGCTGGACCAGACCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGAC

       I  Y  Q  I  P  S  V  D  S  A  D  N  L  S  E  K  L  E  R  E   56
 241  ATATACCAAATCCCTTCTGTTGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAA

       W  D  R  E  L  A  S  K  K  N  P  K  L  I  N  A  L  R  R  C   76
 301  TGGGATAGAGAGCTGGCTTCAAAGAAAAATCCTAAACTCATTAATGCCCTTCGGCGATGT

       F  F  W  R  F  M  F  Y  G  I  F  L  Y  L  G  E  V  T  K  A   96
 361  TTTTTCTGGAGATTTATGTTCTATGGAATCTTTTTATATTTAGGGGAAGTCACCAAAGCA

       V  Q  P  L  L  L  G  R  I  I  A  S  Y  D  P  D  N  K  E  E  116
 421  GTACAGCCTCTCTTACTGGGAAGAATCATAGCTTCCTATGACCCGGATAACAAGGAGGAA

       R  S  I  A  I  Y  L  G  I  G  L  C  L  L  F  I  V  R  T  L  136
 481  CGCTCTATCGCGATTTATCTAGGCATAGGCTTATGCCTTCTCTTTATTGTGAGGACACTG

       L  L  H  P  A  I  F  G  L  H  E  I  G  M  Q  M  R  I  A  M  156
 541  CTCCTACACCCAGCCATTTTTGGCCTTCATCACATTGGAATGCAGATGAGAATAGCTATG

       F  S  L  I  Y  K  K  T  L  K  L  S  S  R  V  L  D  K  I  S  176
 601  TTTAGTTTGATTTATAAGAAGACTTTAAAGCTGTCAAGCCGTGTTCTAGATAAAATAAGT

       I  G  Q  L  V  S  L  L  S  N  N  L  N  K  F  D  E  G  L  A  196
 661  ATTGGACAACTTGTTAGTCTCCTTTCCAACAACCTGAACAAATTTGATGAAGGACTTGCA

       L  A  H  F  V  W  I  A  P  L  Q  V  A  L  L  M  G  L  I  W  216
 721  TTGGCACATTTCGTGTGGATCGCTCCTTTGCAAGTGGCACTCCTCATGGGGCTAATCTGG

       E  L  L  Q  A  S  A  F  C  G  L  G  F  L  I  V  L  A  L  F  236
 781  GAGTTGTTACAGGCGTCTGCCTTCTGTGGACTTGGTTTCCTGATAGTCCTTGCCCTTTTT

       Q  A  G  L  G  R  M  M  M  K  Y  R  D  Q  R  A  G  K  I  S  256
 841  CAGGCTGGGCTAGGGAGAATGATGATGAAGTACAGAGATCAGAGAGCTGGGAAGATCAGT

       E  R  L  V  I  T  S  E  M  I  E  N  I  Q  S  V  K  A  Y  C  276
 901  GAAAGACTTGTGATTACCTCAGAAATGATTGAAAATATCCAATCTGTTAAGGCATACTGC

       W  E  E  A  M  E  K  M  I  E  N  L  R  Q  T  E  L  K  L  T  296
 961  TGGGAAGAAGCAATGGAAAAAATGATTGAAAACTTAAGACAAACAGAACTGAAACTGACT

       R  K  A  A  Y  V  R  Y  F  N  S  S  A  F  F  F  S  G  F  F  316
1021  CGGAAGGCAGCCTATGTGAGATACTTCAATAGCTCAGCCTTCTTCTTCTCAGGGTTCTTT

       V  V  F  L  S  V  L  P  Y  A  L  I  K  G  I  I  L  R  K  I  336
1081  GTGGTGTTTTTATCTGTGCTTCCCTATGCACTAATCAAAGGAATCATCCTCCGGAAAATA

       F  T  T  I  S  F  C  I  V  L  R  M  A  V  T  R  Q  F  P  W  356
1141  TTCACCACCATCTCATTCTGCATTGTTCTGCGCATGGCGGTCACTCGGCAATTTCCCTGG

       A  V  Q  T  W  Y  D  S  L  G  A  I  N  K  I  Q  D  F  L  Q  376
1201  GCTGTACAAACATGGTATGACTCTCTTGGAGCAATAAACAAAATACAGGATTTCTTACAA

       K  Q  E  Y  K  T  L  E  Y  N  L  T  T  T  E  V  V  M  E  N  396
1261  AAGCAAGAATATAAGACATTGGAATATAACTTAACGACTACAGAAGTAGTGATGGAGAAT

       V  T  A  F  W  E  E  G  F  G  E  L  F  E  K  A  K  Q  N  N  416
1321  GTAACAGCCTTCTGGGAGGAGGGATTTGGGGAATTATTTGAGAAAGCAAAACAAAACAAT

       N  N  R  K  T  S  N  G  D  D  S  L  F  F  S  N  F  S  L  L  436
1381  AACAATAGAAAAACTTCTAATGGTGATGACAGCCTCTTCTTCAGTAATTTCTCACTTCTT

       G  T  P  V  L  K  D  I  N  F  K  I  E  R  G  Q  L  L  A  V  456
1441  GGTACTCCTGTCCTGAAAGATATTAATTTCAAGATAGAAAGAGGACAGTTGTTGGCGGTT

       A  G  S  T  G  A  G  K  T  S  L  L  M  M  I  M  G  E  L  E  476
1501  GCTGGATCCACTGGAGCAGGCAAGACTTCACTTCTAATGATGATTATGGGAGAACTGGAG

       P  S  E  G  K  I  K  H  S  G  R  I  S  F  C  S  Q  F  S  W  496
1561  CCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGG

       I  M  P  G  T  I  K  E  N  I  I  F  G  V  S  Y  D  E  Y  R  516
1621  ATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGA

       Y  R  S  V  I  K  A  C  Q  L  E  E  D  I  S  K  F  A  E  K  536
1681  TACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGGACATCTCCAAGTTTGCAGAGAAA

       D  N  I  V  L  G  E  G  G  I  T  L  S  G  G  Q  R  A  R  I  556
1741  GACAATATAGTTCTTGGAGAAGGTGGAATCACACTGAGTGGAGGTCAACGAGCAAGAATT

       S  L  A  R  A  V  Y  K  D  A  D  L  Y  L  L  D  S  P  F  G  576
1801  TCTTTAGCAAGAGCAGTATACAAAGATGCTGATTTGTATTTATTAGACTCTCCTTTTGGA

       Y  L  D  V  L  T  E  K  E  I  F  E  S  C  V  C  K  L  M  A  596
1861  TACCTAGATGTTTTAACAGAAAAAGAAATATTTGAAAGCTGTGTCTGTAAACTGATGGCT

       N  K  T  R  I  L  V  T  S  K  M  E  H  L  K  K  A  D  K  I  616
1921  AACAAAACTAGGATTTTGGTCACTTCTAAAATGGAACATTTAAAGAAAGCTGACAAAATA

       L  I  L  N  E  G  S  S  Y  F  Y  G  T  F  S  E  L  Q  N  L  636
1981  TTAATTTTGAATGAAGGTAGCAGCTATTTTTATGGGACATTTTCAGAACTCCAAAATCTA

       Q  P  D  F  S  S  K  L  M  G  C  D  S  F  D  Q  F  S  A  E  656
2041  CAGCCAGACTTTAGCTCAAAACTCATGGGATGTGATTCTTTCGACCAATTTAGTGCAGAA

       R  R  N  S  I  L  T  E  T  L  H  R  F  S  L  E  G  D  A  P  676
2101  AGAAGAAATTCAATCCTAACTGAGACCTTACACCGTTTCTCATTAGAAGGAGATGCTCCT

       V  S  W  T  E  T  K  K  Q  S  F  K  Q  T  G  E  F  G  E  K  696
2161  GTCTCCTGGACAGAAACAAAAAAACAATCTTTTAAACAGACTGGAGAGTTTGGGGAAAAA

       R  K  N  S  I  L  N  P  I  N  S  I  R  K  F  S  I  V  Q  K  716
2221  AGGAAGAATTCTATTCTCAATCCAATCAACTCTATACGAAAATTTTCCATTGTGCAAAAG

       T  P  L  Q  M  N  G  I  E  E  D  S  D  E  P  L  E  R  R  L  736
2281  ACTCCCTTACAAATGAATGGCATCGAAGAGGATTCTGATGAGCCTTTAGAGAGAAGGCTG

       S  L  V  P  D  S  E  Q  G  E  A  I  L  P  R  I  S  V  I  S  756
2341  TCCTTAGTACCAGATTCTGAGCAGGGAGAGGCGATACTGCCTCGCATCAGCGTGATCAGC

       T  G  P  T  L  Q  A  R  R  R  Q  S  V  L  N  L  M  T  H  S  776
2401  ACTGGCCCCACGCTTCAGGCACGAAGGAGGCAGTCTGTCCTGAACCTGATGACACACTCA

       V  N  Q  G  Q  N  I  H  R  K  T  T  A  S  T  R  K  V  S  L  796
2461  GTTAACCAAGGTCAGAACATTCACCGAAAGACAACAGCATCCACACGAAAAGTGTCACTG

       A  P  Q  A  N  L  T  E  L  D  I  Y  S  R  R  L  S  Q  E  T  816
2521  GCCCCTCAGGCAAACTTGACTGAACTGGATATATATTCAAGAAGGTTATCTCAAGAAACT

       G  L  E  I  S  E  E  I  N  E  E  D  L  K  E  C  L  F  D  D  836
2581  GGCTTGGAAATAAGTGAAGAAATTAACGAAGAAGACTTAAAGGAGTGCCTTTTTGATGAT

       M  E  S  I  P  A  V  T  T  W  N  T  Y  L  R  Y  I  T  V  H  856
2641  ATGGAGAGCATACCAGCAGTGACTACATGGAACACATACCTTCGATATATTACTGTCCAC

       K  S  L  I  F  V  L  I  W  C  L  V  I  F  L  A  E  V  A  A  876
2701  AAGAGCTTAATTTTTGTGCTAATTTGGTGCTTAGTAATTTTTCTGGCAGAGGTGGCTGCT

       S  L  V  V  L  W  L  L  G  N  T  P  L  Q  D  K  G  N  S  T  896
2761  TCTTTGGTTGTGCTGTGGCTCCTTGGAAACACTCCTCTTCAAGACAAAGGGAATAGTACT

       H  S  R  N  N  S  Y  A  V  I  I  T  S  T  S  S  Y  Y  V  F  916
2821  CATAGTAGAAATAACAGCTATGCAGTGATTATCACCAGCACCAGTTCGTATTATGTGTTT

       Y  I  Y  V  G  V  A  D  T  L  L  A  M  G  F  F  R  G  L  P  936
2881  TACATTTACGTGGGAGTAGCCGACACTTTGCTTGCTATGGGATTCTTCAGAGGTCTACCA

       L  V  H  T  L  I  T  V  S  K  I  L  H  H  K  M  L  H  S  V  956
2941  CTGGTGCATACTCTAATCACAGTGTCGAAAATTTTACACCACAAAATGTTACATTCTGTT

       L  Q  A  P  M  S  T  L  N  T  L  K  A  G  G  I  L  N  R  F  976
3001  CTTCAAGCACCTATGTCAACCCTCAACACGTTGAAAGCAGGTGGGATTCTTAATAGATTC

       S  K  D  I  A  I  L  D  D  L  L  P  L  T  I  F  D  F  I  Q  996
3061  TCCAAAGATATAGCAATTTTGGATGACCTTCTGCCTCTTACCATATTTGACTTCATCCAG

       L  L  L  I  V  I  G  A  I  A  V  V  A  V  L  Q  P  Y  I  F 1016
3121  TTGTTATTAATTGTGATTGGAGCTATAGCAGTTGTCGCAGTTTTACAACCCTACATCTTT

       V  A  T  V  P  V  I  V  A  F  I  M  L  R  A  Y  F  L  Q  T 1036
3181  GTTGCAACAGTGCCAGTGATAGTGGCTTTTATTATGTTGAGAGCATATTTCCTCCAAACC

       S  Q  Q  L  K  Q  L  E  S  E  G  R  S  P  I  F  T  H  L  V 1056
3241  TCACAGCAACTCAAACAACTGGAATCTGAAGGCAGGAGTCCAATTTTCACTCATCTTGTT

       T  S  L  K  G  L  W  T  L  R  A  F  G  R  Q  P  Y  F  E  T 1076
3301  ACAAGCTTAAAAGGACTATGGACACTTCGTGCCTTCGGACGGCAGCCTTACTTTGAAACT

       L  F  H  K  A  L  N  L  H  T  A  N  W  F  L  Y  L  S  T  L 1096
3361  CTGTTCCACAAAGCTCTGAATTTACATACTGCCAACTGGTTCTTGTACCTGTCAACACTG

       R  W  F  Q  M  R  I  E  M  I  F  V  I  F  F  I  A  V  T  F 1116
3421  CGCTGGTTCCAAATGAGAATAGAAATGATTTTTGTCATCTTCTTCATTGCTGTTACCTTC

       I  S  I  L  T  T  G  E  G  E  G  R  V  G  I  I  L  T  L  A 1136
3481  ATTTCCATTTTAACAACAGGAGAAGGAGAAGGAAGAGTTGGTATTATCCTGACTTTAGCC

       M  N  I  M  S  T  L  Q  W  A  V  N  S  I  I  D  V  D  S  L 1156
3541  ATGAATATCATGAGTACATTGCAGTGGGCTGTAAACTCCAGCATAGATGTGGATAGCTTG

       M  R  S  V  S  R  V  F  K  F  I  D  M  P  T  E  G  K  P  T 1176
3601  ATGCGATCTGTGAGCCGAGTCTTTAAGTTCATTGACATGCCAACAGAAGGTAAACCTACC

       K  S  T  K  P  Y  K  N  G  Q  L  S  K  V  M  I  I  E  N  S 1196
3661  AAGTCAACCAAACCATACAAGAATGGCCAACTCTCGAAAGTTATGATTATTGAGAATTCA

       H  V  K  K  D  D  I  W  P  S  G  G  Q  M  T  V  K  D  L  T 1216
3721  CACGTGAAGAAAGATGACATCTGGCCCTCAGGGGGCCAAATGACTGTCAAAGATCTCACA

       A  K  Y  T  E  G  G  N  A  I  L  E  N  I  S  F  S  I  S  P 1236
3781  GCAAAATACACAGAAGGTGGAAATGCCATATTAGAGAACATTTCCTTCTCAATAAGTCCT

       G  Q  R  V  G  L  L  G  R  T  G  S  G  K  S  T  L  L  S  A 1256
3841  GGCCAGAGGGTGGGCCTCTTGGGAAGAACTGGATCAGGGAAGAGTACTTTGTTATCAGCT

       F  L  R  L  L  N  T  E  G  E  I  Q  I  D  G  V  S  W  D  S 1276
3901  TTTTTGAGACTACTGAACACTGAAGGAGAAATCCAGATCGATGGTGTGTCTTGGGATTCA

       I  T  L  Q  Q  W  R  K  A  F  G  V  I  P  Q  K  V  F  I  F 1296
3961  ATAACTTTGCAACAGTGGAGGAAAGCCTTTGGAGTGATACCACAGAAAGTATTTATTTTT

       S  G  T  F  R  K  N  L  D  P  Y  E  Q  W  S  D  Q  E  I  W 1316
4021  TCTGGAACATTTAGAAAAAACTTGGATCCCTATGAACAGTGGAGTGATCAAGAAATATGG

       K  V  A  D  E  V  G  L  R  S  V  I  E  Q  F  P  G  K  L  D 1336
4081  AAAGTTGCAGATGAGGTTGGGCTCAGATCTGTGATAGAACAGTTTCCTGGGAAGCTTGAC

       F  V  L  V  D  G  G  C  V  L  S  H  G  H  K  Q  L  M  C  L 1356
4141  TTTGTCCTTGTGGATGGGGGCTGTGTCCTAAGCCATGGCCACAAGCAGTTGATGTGCTTG

       A  R  S  V  L  S  K  A  K  I  L  L  L  D  E  P  S  A  H  L 1376
4201  GCTAGATCTGTTCTCAGTAAGGCGAAGATCTTGCTGCTTGATGAACCCAGTGCTCATTTG

       D  P  V  T  Y  Q  I  I  R  R  T  L  K  Q  A  F  A  D  C  T 1396
4261  GATCCAGTAACATACCAAATAATTAGAAGAACTCTAAAACAAGCATTTGCTGATTGCACA

       V  I  L  C  E  H  R  I  E  A  M  L  E  C  Q  Q  F  L  V  I 1416
4321  GTAATTCTCTGTGAACACAGGATAGAAGCAATGCTGGAATGCCAACAATTTTTGGTCATA

       E  E  N  K  V  R  Q  Y  D  S  I  Q  K  L  L  N  E  R  S  L 1436
4381  GAAGAGAACAAAGTGCGGCAGTACGATTCCATCCAGAAACTGCTGAACGAGAGGAGCCTC

       F  R  Q  A  I  S  P  S  D  R  V  K  L  F  P  H  R  N  S  S 1456
4441  TTCCGGCAAGCCATCAGCCCCTCCGACAGGGTGAAGCTCTTTCCCCACCGGAACTCAAGC

       K  C  K  S  K  P  Q  I  A  A  L  K  E  E  T  E  E  E  V  Q 1476
4501  AAGTGCAAGTCTAAGCCCCAGATTGCTGCTCTGAAAGAGGAGACAGAAGAAGAGGTGCAA
```

```
        D  T  R  L  -                                          1480
4561  GATACAAGGCTTTAGAGAGCAGCATAAATGTTGACATGGACATTTGCTCATGGAATTGG
4621  AGCTCGTGGACAGTCACCTCATGGAATTGGAGCTCGTGGAACAGTTACCTCTGCCTCAG
4681  AAAACAAGGATGAATTAAGTTTTTTTTTAAAAAAGAAACATTTGGTAAGGGGAATTGAGG
4741  ACACTGATATGGGTCTTGATAAATGGCTTCCTGGCAATAGTCAAATTGTGTGAAAGGTAC
4801  TTCAAATCCTTGAAGATTTACCACTTGTGTTTTGCAAGCCAGATTTTCCTGAAAACCCTT
4861  GCCATGTGCTAGTAATTGGAAAGGCAGCTCTAAATGTCAATCAGCCTAGTTGATCAGCTT
4921  ATTGTCTAGTGAAACTCGTTAATTTGTAGTGTTGGAGAAGAACTGAAATCATACTTCTTA
4981  GGGTTATGATTAAGTAATGATAACTGGAAACTTCAGCGGTTTATATAAGCTTGTATTCCT
5041  TTTTCTCTCCTCTCCCCATGATGTTTAGAAACACAACTATATTGTTTGCTAAGCATTCCA
5101  ACTATCTCATTTCCAAGCAAGTATTAGAATACCACAGGAACCACAAGACTGCACATCAAA
5161  ATATGCCCCATTCAACATCTAGTGAGCAGTCAGGAAAGAGAACTTCCAGATCCTGGAAAT
5221  CAGGGTTAGTATTGTCCAGGTCTACCAAAAATCTCAATATTTCAGATAATCACAATACAT
5281  CCCTTACCTGGGAAAGGGCTGTTATAATCTTTCACAGGGGACAGGATGGTTCCCTTGATG
5341  AAGAAGTTGATATGCCTTTTCCCAACTCCAGAAAGTGACAAGCTCACAGACCTTTGAACT
5401  AGAGTTTAGCTGGAAAAGTATGTTAGTGCAAATTGTCACAGGACAGCCCTTCTTTCCACA
5461  GAAGCTCCAGGTAGAGGGTGTGTAAGTAGATAGGCCATGGGCACTGTGGGTAGACACACA
5521  TGAAGTCCAAGCATTTAGATGTATAGGTTGATGGTGGTATGTTTTCAGGCTAGATGTATG
5581  TACTTCATGCTGTCTACACTAAGAGAGAATGAGAGACACACTGAAGAAGCACCAATCATG
5641  AATTAGTTTTATATGCTTCTGTTTTATAATTTTGTGAAGCAAAATTTTTTCTCTAGGAAA
5701  TATTTATTTTAATAATGTTTCAAACATATATTACAATGCTGTATTTTAAAAGAATGATTA
5761  TGAATTACATTTGTATAAAATAATTTTTATATTTGAAATATTGACTTTTTATGGCACTAG
5821  TATTTTTATGAAATATTATGTTAAAACTGGGACAGGGGAGAACCTAGGGTGATATTAACC
5881  AGGGGCCATGAATCACCTTTTGGTCTGGAGGGAAGCCTTGGGGCTGATCGAGTTGTTGCC
5941  CACAGCTGTATGATTCCCAGCCAGACACAGCCTCTTAGATGCAGTTCTGAAGAAGATGGT
6001  ACCACCAGTCTGACTGTTTCCATCAAGGGTACACTGCCTTCTCAACTCCAAACTGACTCT
6061  TAAGAAGACTGCATTATATTTATTACTGTAAGAAAATATCACTTGTCAATAAAATCCATA
6121  CATTTGTGT(A)n
```

Fig. 2. Nucleotide sequence of cDNA encoding the CF transmembrane conductance regulator together with the deduced amino acid sequence. DNA sequencing was performed by the dideoxy chain termination method (*34*) with ^{35}S-labeled nucleotides or by the Dupont Genesis2000 automatic DNA sequencer. Numbers on the left of columns indicate base positions and numbers on the right amino acid residue positions. The first base position corresponds to the first nucleotide in the 5′ extension clone PA3-5, which is one nucleotide longer than TB2-7 (*12*). The 3′ end and the noncoding sequence are shown above [nucleotides 4561 to 6129 plus the poly(A)$^+$ tail]. Arrows indicate position of transcription initiation site by primer extention analysis (*11*). Nucleotide 6129 is followed by a poly(A) tract. Positions of exon junctions are indicated by vertical lines. Potential membrane-spanning segments ascertained with the use of the algorithm of Eisenberg *et al.* (*35*) are enclosed in boxes. Amino acids comprising putative ATP-binding folds are underlined. Possible sites of phosphorylation (*21*) by protein kinases A or C are indicated by open and closed circles, respectively. The open triangle indicates the position at which 3 bp are deleted in CF. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

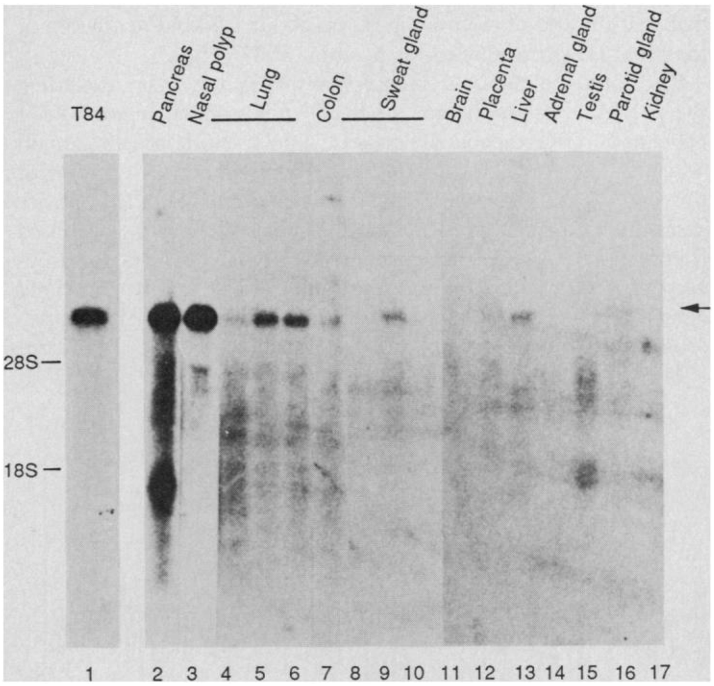

Fig. 3. RNA gel-blot analysis. Hybridization by the cDNA clone 10-1 to a 6.5-kb transcript is shown in the tissues indicated. RNA samples were prepared from cells and tissue samples obtained from surgical pathology or at autopsy according to the methods described in (*10*). Total RNA (10 μg) from each tissue and 1 μg of poly(A)$^+$ RNA from T84 cells were separated on formaldehyde gels and transferred onto nylon membranes (Zetaprobe, Bio-Rad), which were hybridized with DNA probes labeled to high specific activity by the random priming method (*36, 37*). The positions of the 28*S* and 18*S* rRNA bands are indicated.

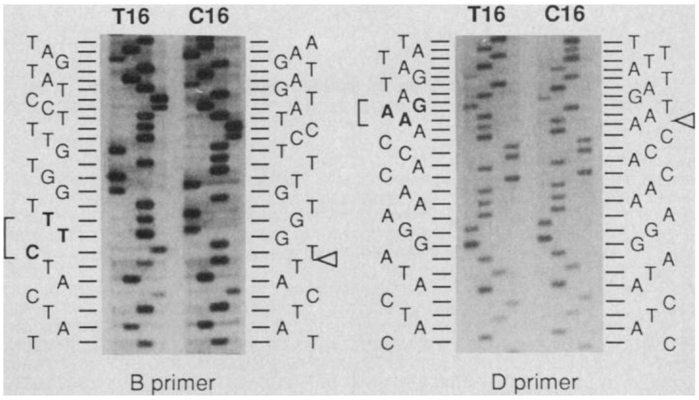

Fig. 4. DNA sequence around the ΔF_{508} deletion. The normal sequence from base position 1627 to 1651 (from cDNA T16-1) is shown beside the CF sequence (from cDNA C16-1). The left panel shows the sequences from the coding strands obtained with the B primer (5′-GTTTTCCTGGATTATGCCTGGGCAC-3′) and the right panel those from the opposite strand with the D primer (5′-GTTGGCATGCTTTGATGACGCTTC-3′). The brackets indicate the three nucleotides in the normal that are absent in CF (arrowheads). Sequencing was performed as described in (*34*).

uous coding region of the CF locus could be deduced from overlapping cDNA clones. Since most of the cDNA clones were apparently derived from unprocessed transcripts, further studies were performed to ensure the authenticity of the consensus sequence. Each cDNA clone was first tested for chromosome localization by hybridization analysis with a human-hamster somatic cell hybrid containing a single human chromosome 7 and by pulsed field gel electrophoresis (7). The ones that did not map to the correct region on chromosome 7 were not pursued. Fine restriction enzyme mapping was then performed for each clone. While overlapping regions were clearly identified for most of the clones, many contained single copy, additional regions not readily recognizable by restriction enzyme analysis.

The cDNA was further characterized in gel hybridization experiments with genomic DNA. Five to six different restriction fragments could be detected with the 10-1 cDNA in Eco RI– or Hind III–digested total human DNA and a similar number of fragments with several other cDNA clones, suggesting the presence of multiple exons for the putative CF gene. The hybridization studies also identified the cDNA clones with unprocessed intron sequences when they showed preferential hybridization to a smaller subset of genomic DNA fragments with relatively greater intensities. For the confirmed cDNA clones, their corresponding genomic DNA segments were isolated (7) and the exons and exon-intron boundaries were sequenced. In all, 24 exons were identified (Fig. 2). Physical mapping experiments (7) showed that the gene locus spanned about 250 kb.

The 5′ terminus of the transcript was determined by primer extension (*11*). A modified polymerase chain reaction, anchored PCR (*12*), was also used to facilitate cloning of the 5′ end sequences. Two independent 5′ extension clones, one from pancreas and the other from T84 RNA, were characterized by DNA sequencing and differed by only 1 base in length, thus establishing the most probable initiation site for the transcript (Fig. 2). Since the initial cDNA clones did not contain a poly(A)$^+$ tail indicative of the end of a mRNA, anchored PCR was also applied to the 3′ end of the transcript (*12*). The results derived from the use of several different 3′-extending oligonucleotides were consistent with the interpretation that the end of the transcript was about 1.2 kb downstream of the Hind III site at nucleotide position 5027 (Fig. 2).

The complete cDNA sequence spans 6129 base pairs excluding the poly(A)$^+$ tail at the end of the 3′ untranslated region and it contains an ORF capable of encoding a polypeptide of 1480 amino acids (Fig. 2). An ATG (AUG) triplet is present at the beginning of this ORF (base position 133–135). Since the nucleotide sequence surrounding this codon (5′-AGACC**AUG**CA-3′) has the proposed features of the consensus sequence (CC)$^{A}_{G}$CC**AUG**G(G) of a eukaryotic translation initiation site (*13*), with a highly conserved A at the −3 position, it is highly probable that this AUG corresponds to the first methionine codon for the putative polypeptide.

Detection of mutation. A comparison between the cDNA sequences derived from CF and unaffected (N) individuals was next conducted. Two clones, C16-1 and C1-1/5, were derived from a CF sweat gland cDNA library and together they spanned almost the entire coding region. The most striking difference between CF and N sequences was a 3-bp deletion (Fig. 4), which would result in a loss of a phenylalanine residue (position 508) in the predicted CF polypeptide. This deletion (ΔF_{508}) was detected in both CF clones. To exclude the possibility that this difference was due to a cloning artifact, sequence-specific oligonucleotides were used to screen DNA samples from CF families. Specific hybridization could be observed for each oligonucleotide probe with genomic DNA amplified by PCR, confirming the presence of corresponding genomic DNA sequences (*9*). Furthermore, the oligonucleotide specific for the 3-bp deletion hybridized to 68 percent of chromosomes carrying a CF mutation but not to any of the normal chromosomes (0/198), an indication that a silent sequence polymorphism was unlikely. Sequence differences found elsewhere among the different cDNA clones probably represented sequence polymorphisms or cDNA cloning artifacts (*14*).

Predicted protein structure. Analysis of the sequence of the overlapping cDNA clones (Fig. 2) predicted a polypeptide of 1480 amino acids with a molecular mass of 168,138 daltons. The most characteristic feature of the predicted protein is the presence of two repeated motifs, each of which consists of a domain capable of spanning the membrane several times and sequences resembling consensus nucleotide (ATP)-binding folds (NBF's) (*15*) (Figs. 5 and 6). These characteristics are remarkably similar to those of the mammalian multidrug resistance P-glycoprotein (*16*) and a number of other membrane-associated proteins (as discussed below), suggesting that the predicted CF gene product is likely to be involved in the transport of substances (ions) across the membrane and is probably a member of a membrane protein superfamily (*17*). For the convenience of future discussion and to avoid confusion with the previously named CF protein and CF factor (*18*), we will call the

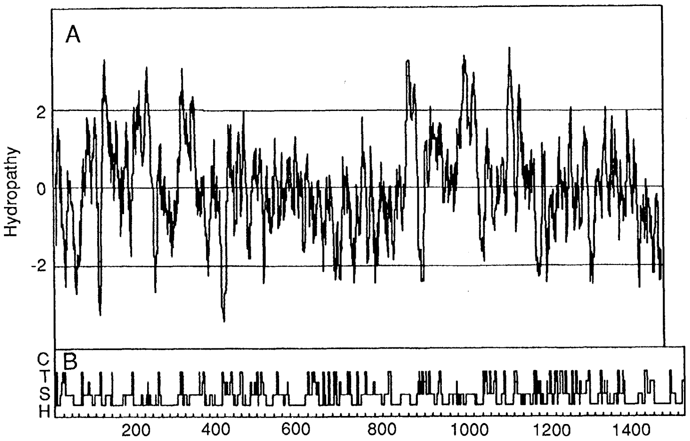

Fig. 5. Hydropathy profile and predicted secondary structures of the CFTR. (**A**) The mean hydropathy index determined according to Kyte and Doolittle (*19*) of nine-residue peptides is plotted against the amino acid number. (**B**) The corresponding positions of features of secondary structure predicted according to Garnier *et al.* (*19*). C, coil; T, turn; S, sheet; H, helix.

		*	
CFTR (N)	FSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMG	ISFCSQFSWIMPGTIK-ENIIFGVSYD	GEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEK
CFTR (C)	YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLR	DSITLQQWRKAFGVIPQKVFIFSGTFR	VDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQ
hmdr1 (N)	PSRKEVKILKGLNLKVQSGQTVALVGNSGCGKSTTVQLMQR	IGVVSQEPVLFATTI-AENIRYGRENV	GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
hmdr1 (C)	PTRPDIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER	LGIVSQEPILFDCSI-AENIAYGDNSR	GDKGTLLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr1 (N)	PSRSEVQILKGLNLKVKSGQTVALVGNSGCGKSTTVQLMQR	IGVVSQEPVLFATTI-AENIRYGREDV	GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr1 (C)	PTRPNIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER	LGEVSQEPILFDCSI-AENIAYGDNSR	GDKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr2 (N)	PSRANIKILKGLNLKVKSGQTVALVGNSGCGKSTTVQLLQR	IGVVSQEPVLSFTTI-AENIRYGRGNV	GDRGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr2 (C)	PTRANVPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER	LGIVSQEPILFDCSI-AENIAYGDNSR	GDKGTQLSGGQKQRIAIARALIRQPRVLLLDEATSALDTESEK
pfmdr (N)	DTRKDVEIYKDLSFTLLKEGKTYAFVGESGCGKSTILKLIE	IGVVSQDPLLFSNSI-KNNIKYSLYSL	GSNASKLSGGQKQRISIARAIMRNPKILILDEATSSLDNKSEY
pfmdr (C)	ISRPNVPIYKNLSFTCDSKKTTAIVGETGSGKSTFMNLLLR	FSIVSQEPMLFNMSI-YENIKFGREDA	PYGKS-LSGGQKQRIAIARALLREPKILLLDEATSSLDSNSEK
STE6 (N)	PSRPSEAVLKNVSLNFSAGQFTFIVGKSGSGKSTLSNLLLR	ITVVEQRCTLFNDTL-RKNILLGSTDS	GTGGVTLSGGQQQRVAIARAFIRDTPILFLDEAVSALDIVHRN
STE6 (C)	PSAPTAFVYKNMNFDMFCGQTLGIIGESGTGKSTLVLLLTK	ISVVEQKPLLFNGTI-RDNLTYGLQDE	RIDTTLLSGGQAQRLCIARALLRKSKILILDECTSALDSVSSS
hlyB	YKPDSPVILDNINISIKQGEVIGIVGRSGSGKSTLIKLIQR	VGVVLQDNVLLNRSI-IDNISLAPGMS	GEQGAGLSGGQRQRIAIARALVNNPKILIFDEATSALDYASEH
White	IPAPRKHLLKNVCGVAYPGELLAVMGSSGAGKTTLLNALAF	RCAYVQQDDLFIGLIAREHLIFQAMVR	PGRVKGLSGGERKRLAFASEALTDPPLLICDEPTSGLDSFTAH
MbpX	KSLGNLKILDRVSLYVPKFSLIALLGPSGSGKSSLLRILAG	MSFVFQHYALFKHMTVYENISFGLRLR	FEYPAQLSGGQKQRVALARSLAIQPDLLL-DEPFGALDGELRR
BtuD	QDVAESTRLGPLSGEVRAGRILHLVGPNGAGKSTLLARIAG	YLSQQQTPPFATPVWHYLTLHQHDKTR	GRSTNQLSGGEWQRVRLAAVVLQITLLLLDEPMNSLDVAQQSA
PstB	FYYGKFHALKNINLDTAKNQVTAFIGPSGCGKSTLLRTFNK	VGMVFQKPTPFPMSI-YDNIAFGVRLF	HQSGYSLSGGQQQRLCIARGIAIRPEVLLLDEPCSALDPISTG
hisP	RRYGGHEVLKGVSLQARAGDVISIIGSSGSGKSTFLRCINF	GIMVFQHFNLWSHMTVLENVMEAPIQV	GKYPVHLSGGQQQRVSIARALAMEPDVLLFDEPTSALDPELVG
malK	KAWGEVVVSKDINIDIHEGEFVVFVGPSGCGKSTLLRMIAG	VGMVFQSYALYPHLSVAENMSFGLKPA	DRKPKALSGGQRQRVAIGRTLVAEPSVFLLDEPLSNLDAALRV
oppD	TPDGDVTAVNDLNFTLRAGETLGIVGESGSGKSQTAFALMG	ISMIFQDPMTSLNPYMRVGEQLMEVLM	KMYPHEFSGGMRQRVMIAMALLCRPKLLIADEPTTALDVTVQA
oppF	QPPKTLKAVDGVTLRLYEGETLGVVGESGCGKSTFARAIIG	IQMIFQDPLASLNPRMTIGEIIAEPLR	NRYPHEFSGGQCQRIGIARALILEPKLIICDDAVSALDVSIQA
RbsA (N)	KAVPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTG	AGIIHQELNLIPQLTIAENIFLGREFV	DKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTCALIDTETE
RbsA (C)	VDNLCGPGVNDVSFTLRKGEILGVSGLMGAGRTELMKVLYG	ISEDRKRDGLVLGMSVKENMSLTALRY	EQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPTPGVDVGAKK
UvrA	LTGARGNNLKDVTLTLPVGLFTCITGVSGSGKSTLINDTLF	TYTGVFTPVRELFAGVPESRARGYTPG	GQSATTLSGGEAQRVKLARELSKRGLYILDEPTTGLHFADIQQ
NodI	KSYGGKIVVNDLSFTIAAGECFGLLGPNGAGKSTIIRMILG	IGIVSQEDNLDLEFTVRENLLVYGRYF	NTRVADLSGGMKRRLTLAGALINDPQLLILDEPTTGLDPHARH
FtsE	AYLGGRQALQGVTFHMQPGEMAFLTGHSGAGKSTLLKLICG	IGMIFQDHHLLMDRTVYDNVAIPLIIA	KNFPIQLSGGEQQRVGIARAVVNKPAVLLADEPTGNLDDALSE

Fig. 6. Alignment of the three most conserved segments of the amino acid sequences (single letter code) of the extended NBF's of CFTR with comparable regions of other proteins. These three segments consist of residues 433 to 473, 488 to 513, and 542 to 584 of the amino-terminal (N) half and 1219 to 1259, 1277 to 1302, and 1340 to 1382 of the carboxyl-terminal (C) half of CFTR. The heavy overlining points out the regions of greatest similarity. The star indicates the position corresponding to the phenylalanine that is deleted in CF. Additional general homology can be seen even with the introduction of very few gaps. The other sequences are of proteins involved in multidrug resistance in human (hmdr1), mouse (mmdr 1 and 2) (*16*), and *Plasmodium falciparum* (pfmdr) (*38*); the α-factor pheromone export system of yeast (STE6) (*39*); the hemolysin (hlyB) system of *E. coli* (*22*); screening of eye pigments in *Drosophila* (White) (*23*); an unknown liverwort chloroplast function (Mbpx) (*25*); vitamin B12 transport in *E. coli* (BtuD) (*24*); phosphate transport in *E. coli* (PstB) (*40*); histidine transport in *Salmonella typhimurium* (hisP) (*41*); maltose transport in *E. coli* (malK) (*42*); oligopeptide transport in *S. typhimurium* (oppD and oppF) (*43*); ribose transport in *E. coli* (RbsA) (*44*). UvrA is one component of an *E. coli* DNA repair system (*45*); NodI is a gene product involved in nodulation in *Rhizobium* (*46*); FtsE is a protein that contributes to the regulation of cell division (*47*). In addition to these proteins that contain this long NBF, there is a large number of others that contain the two short nucleotide binding motifs A and B initially pointed out by Walker *et al.* (*48*). Further, there are other proteins containing only motif A or B (*49*).

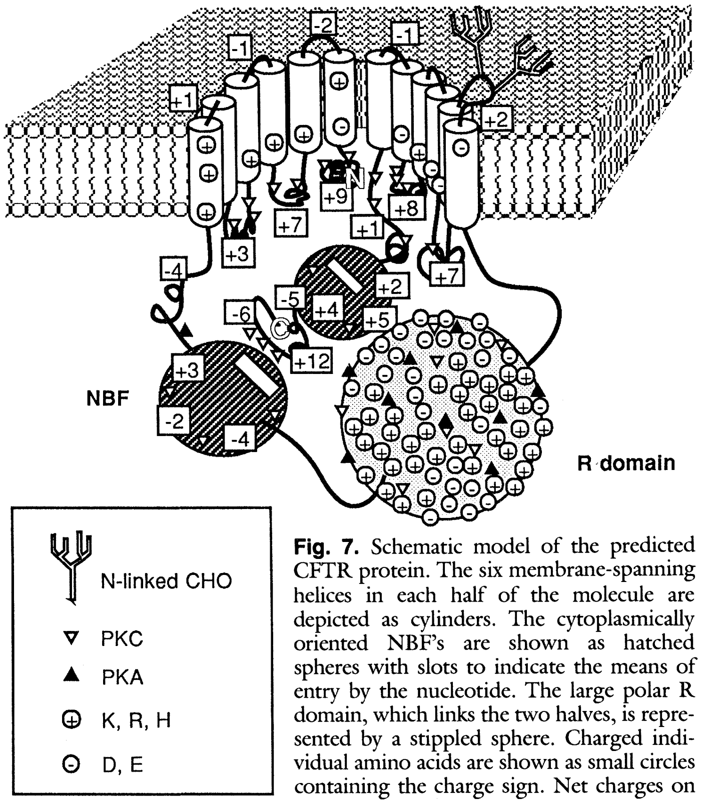

Fig. 7. Schematic model of the predicted CFTR protein. The six membrane-spanning helices in each half of the molecule are depicted as cylinders. The cytoplasmically oriented NBF's are shown as hatched spheres with slots to indicate the means of entry by the nucleotide. The large polar R domain, which links the two halves, is represented by a stippled sphere. Charged individual amino acids are shown as small circles containing the charge sign. Net charges on the internal and external loops joining the membrane cylinders and on regions of the NBF's are contained in open squares. Potential sites for phosphorylation by protein kinases A or C (PKA or PKC) and N-glycosylation (N-linked CHO) are as indicated. K, Lys; R, Arg; H, His; D, Asp; and E, Glu.

putative CF gene product the cystic fibrosis transmembrane conductance regulator (CFTR).

Each of the predicted membrane-associated regions of CFTR consists of six hydrophobic segments capable of spanning a lipid bilayer (*19*), which are followed by a large hydrophilic region containing the NBF's (Fig. 5). On the basis of sequence alignment with other nucleotide-binding proteins, each of the putative NBF's in CFTR comprises at least 150 residues (Fig. 6). The single residue deletion (ΔF_{508}) detected in most of the CF patients is in the first NBF, between the two most highly conserved segments within this sequence. The amino acid sequence identity between the region surrounding the ΔF_{508} mutation and the corresponding regions of several other proteins suggests that this region is of functional importance (Fig. 6). A hydrophobic amino acid, usually one with an aromatic side chain, is present in most of these proteins at the position corresponding to Phe^{508} of CFTR.

Despite the overall symmetry in the two-motif structure of the protein and the sequence conservation of the NBF's, sequence identity between the two motifs of the predicted CFTR protein is modest. The strongest identity is between sequences at the carboxyl ends of the NBF's. Of the 66 residues aligned within these regions, 27 percent are identical and 11 percent are functionally similar. The overall, weak internal sequence identity is in contrast to the much higher degree (>70 percent) in P-glycoprotein for which a sequence duplication hypothesis has been proposed (*16*). The lack of conservation in the relative positions of the exon-intron boundaries in the CF gene also argues against recent exon duplication as a mechanism in the evolution of this gene (Fig. 2).

Since there is apparently no signal-peptide sequence at the amino terminus of CFTR (Fig. 7), the highly charged hydrophilic segment preceding the first transmembrane sequence is probably oriented in the cytoplasm. Each of the two sets of hydrophobic helices are expected to form three traversing loops across the membrane and little of the sequence of the entire protein is expected to be exposed to the exterior surface, except the region between transmembrane segments 7 and 8. It is of interest that the latter region contains two potential sites for N-linked glycosylation (*20*).

A highly charged cytoplasmic domain can be identified in the middle of the predicted CFTR polypeptide, linking the two halves of the protein. This domain, named the R domain, is operationally defined by a single large exon in which 69 of the 241 amino acids are polar residues arranged in alternating clusters of positive and negative charges. Moreover, nine of the ten sites at which there are consensus sequences for phosphorylation by protein kinase A and seven of the potential substrate sites for protein kinase C found in CFTR are located in this exon (*21*).

Properties of CFTR could be further derived from comparison to other membrane-associated proteins (Fig. 6). In addition to the overall structural similarity with P-glycoproteins, each of the two predicted motifs in CFTR shows resemblance to the single motif structure of hemolysin B of *Escherichia coli* (*22*) and the product of the White gene of *Drosophila* (*23*). These proteins are involved in the transport of the lytic peptide of the hemolysin system and of eye pigment molecules, respectively. The vitamin B12 transport system of *E. coli*, BtuD (*24*), and MbpX (*25*), which is a liverwort chloroplast gene product whose function is unknown, also have a similar structural motif. Further, CFTR shares structural similarity with several of the periplasmic solute transport systems of Gram-negative bacteria, where the transmembrane region and the ATP-binding folds are contained in separate proteins that function in concert with a third substrate-binding polypeptide (*26*).

The overall structural arrangement of the transmembrane domains in CFTR is similar to several cation channel proteins (*27*) and some cation-translocating adenosine triphosphatases (ATPases) (*28*) as well as the recently described adenylate cyclase of bovine brain (*29*). Short regions of sequence identity have also been detected between the putative transmembrane regions of CFTR and other membrane-spanning proteins (*30*). In addition, a sequence of 18 amino acids situated approximately 50 residues from the carboxyl terminus of CFTR shows some identity (12/18) with the raf serine-threonine kinase proto-oncogene product of *Xenopus laevis* (*31*).

Finally, a sequence identity (10 of 13 amino acid residues) has been noted between a hydrophilic segment (position 701 to 713) within the highly charged R domain of CFTR and a region immediately preceding the first transmembrane loop of the sodium channels in both rat brain and eel (*32*). This feature of CFTR is not shared with the topologically closely related P-glycoprotein; the 241–amino acid linking peptide is apparently the major difference between the two proteins.

Relevance to the CF anion transport defect. In view of the genetic data of Kerem *et al.* (*9*) and the tissue specificity and predicted properties of the CFTR protein, it is reasonable to conclude that CFTR is directly responsible for CF. It remains unclear, however, how CFTR is involved in the regulation of ion conductance across the apical membrane of epithelial cells.

It is possible that CFTR serves as an ion channel itself. For example, 10 of the 12 putative transmembrane regions contain one or more amino acids with charged side chains (Fig. 7), a property similar to that of the brain sodium channel and the γ-aminobutyric acid (GABA) receptor chloride channel subunits, where charged residues are present in four of the six, and three of the four, respective membrane-associated domains per subunit or repeat unit (*32, 33*). The amphipathic nature of these transmembrane segments is believed to contribute to the channel-forming capacity of these molecules. In contrast, the closely related P-glycoprotein, which is not believed to conduct ions, has only two charged residues in all 12 transmembrane domains. Alternatively, CFTR may not be an ion channel but instead it may serve to regulate ion channel activities. In support of the latter possibility, none of the recently purified polypeptides (from trachea and kidney) that are capable of reconstituting chloride channels in lipid membranes (*6*) appear to be CFTR, judged on the basis of molecular mass.

In any case, the presence of ATP-binding domains in CFTR suggests that ATP hydrolysis is directly involved and required for the transport function. The high density of phosphorylation sites for protein kinases A and C and the clusters of charged residues in the R domain may both serve to regulate this activity. The deletion of Phe^{508} in the NBF may prevent proper binding of ATP or the conformational change required for normal CFTR activity, consequently resulting in the observed insensitivity to activation by protein kinase A– or protein kinase C–mediated phosphorylation of the CF apical chloride conductance pathway (*5*). Since the predicted structure of CFTR contains several conserved domains and belongs to a family of proteins, most of which function as parts of multicomponent molecular systems (*15*), the CFTR protein may also participate in epithelial cell functions not related to ion transport.

To understand the basic defect in CF, it is necessary to determine the precise role of Phe^{508} in the regulation of ion transport and to understand the mechanism that leads to the pathophysiology of the disease. With the CF gene (that is, the cDNA) now isolated, it should be possible to elucidate the control of ion transport pathways in epithelial cells in general. Knowledge gained from study of the CF gene product (CFTR), both the normal and mutant forms, will provide a molecular basis for the development of improved means of treatment of the disease.

REFERENCES AND NOTES

1. T. F. Boat, M. J. Welsh, A. L. Beaudet, in *The Metabolic Basis of Inherited Disease*, C. L. Scriver, A. L. Beaudet, W. S. Sly, D. Valle, Eds. (McGraw-Hill, New York, ed. 6, 1989), pp. 2649–2680.
2. M. R. Knowles *et al.*, *Science* **221**, 1067 (1983); M. R. Knowles, J. Gatzy, R. Boucher, *J. Clin. Invest.* **71**, 1410 (1983); P. M. Quinton, *Nature* **301**, 421 (1983); K. Sato, *Am. J. Physiol.* **247**, R646 (1984).
3. J. H. Widdicombe, M. J. Welsh, W. E. Finkbeiner, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6167 (1985); P. M. Quinton and J. Bijman, *N. Engl. J. Med.* **308**, 1185 (1983).
4. J. R. Yankaskas, C. U. Cotton, M. R. Knowles, J. T. Gatzy, R. C. Boucher, *Am. Rev. Resp. Dis.* **132**, 1281 (1985); P. S. Pedersen, E. H. Larsen, B. Hainau, N. J. Brandt, *Med. Sci. Res.* **15**, 1009 (1987); T. J. Jensen and J. R. Riordan, unpublished results.
5. M. J. Welsh, *Science* **232**, 1648 (1986); R. A. Frizzell, G. Rechkemmer, R. L. Shoemaker, *ibid.* **233**, 558 (1986); M. J. Welsh and C. M. Liedtke, *Nature* **322**, 467 (1986); R. A. Schoumacher *et al.*, *ibid.* **330**, 752 (1987); M. Li *et al.*, *ibid.* **331**, 358 (1988); T.-C. Hwang *et al.*, *Science* **244**, 1351 (1989); M. Li *et al.*, *ibid.*, p. 1353.
6. D. W. Landry *et al.*, *Science* **244**, 1469 (1989).
7. J. M. Rommens *et al.*, *ibid.* **245**, 1059 (1989).
8. M. Farrall *et al.*, *Am. J. Hum. Genet.* **43**, 471 (1988).
9. B. Kerem *et al.*, *Science* **245**, 1073 (1989).
10. The cDNA libraries from cultured epithelial cells were prepared as follows: sweat gland cells derived from a non-CF individual and from a CF patient were grown to first passage as described [G. Collie, M. Buchwald, P. Harper, J. R. Riordan, *In Vitro Cell. Dev. Biol.* **21**, 592 (1985)]. The presence in these cells of an outwardly rectifying Cl$^-$ channel was confirmed (J. A. Tabcharani, T. J. Jensen, J. R. Riordan, J. W. Hanrahan, *J. Membrane Biol.*, in press), but the CF cells were insensitive to activation by cyclic AMP [T. J. Jensen, J. W. Hanrahan, J. A. Tabcharani, M. Buchwald, J. R. Riordan, *Pediatric Pulmonol. Suppl.* **2**, 100 (1988)]. Polyadenylated RNA was isolated [J. M. Chirgwin, A. E. Przybyla, R. J. Macdonald, W. J. Rutter, *Biochemistry* **18**, 5294 (1979); H. Aviv and P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 1408 (1972)] and used as template for the synthesis of cDNA according to U. Gubler and B. Hoffman [*Gene* **25**, 263 (1983)]. After methylation of internal Eco RI sites, ends were made flush with T4 DNA polymerase, and phosphorylated Eco RI linkers were added to the cDNA. After digestion with Eco RI and removal of excess linkers, the cDNA products were ligated into the Eco RI site of λ ZAP (Stratagene, San Diego, CA). The same procedures were used to construct a library from RNA isolated from preconfluent cultures of the T84 colon carcinoma cell line [K. Dharmsathaphorn, J. A. McRoberts, K. G. Mandel, L. D. Tisdale, H. Masui, *Am. J. Physiol.* **246**, G204 (1984)]. The numbers of independent recombinants in the three libraries were: 2.0 × 10^6 for the non-CF sweat gland cells, 4.5 × 10^6 for the CF sweat gland cells, and 3.2 × 10^6 from T84 cells. Standard procedures were used for screening [T. Maniatis, E. F. Fritsch, J. Sambrook, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982)]. Bluescript plasmids were rescued from plaque-purified clones by excision with M13 helper phage (Stratagene). The lung and pancreas libraries were purchased from Clontech Lab Inc. (Catalog Nos. HL1066b and HL1069h, respectively).
11. The start point of the CF gene transcript was derived by primer extension procedures [F. J. Calzone, R. J. Britten, E. H. Davidson, *Methods Enzymol.* **152**, 611 (1987)]. The oligonucleotide primer [positioned 157 nucleotides (nt) from the 5′ end of the 10-1 clone] was end-labeled with [γ-^{32}P]ATP (Amersham, 5000 Ci/mmole) and T4 polynucleotide kinase, purified by gel filtration, and annealed with ~5 μg of T84 poly(A)$^+$ RNA for 2 hours at 60°C. The extension reaction was performed at 41°C for 1 hour with avian myeloblastosis virus (AMV) reverse transcriptase (Life Sciences, Inc.) and terminated by addition of NaOH to 0.4*M* and EDTA to 20 m*M*, with subsequent neutralization with ammonium acetate (*p*H 4.6). The products were treated with phenol, precipitated with ethanol, redissolved in buffer with formamide, and analyzed on a polyacrylamide sequencing gel.
12. The anchored PCR procedure [M. A. Frohman, M. K. Dush, G. R. Martin, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 8998 (1988)] was used to synthesize cDNA corresponding to the 5′ and 3′ ends of the transcript. For the 5′ end clones, poly(A)$^+$ RNA from pancreas and T84 cells were subjected to reverse-transcription with the use of an exon 2–specific primer (*11*). The first strand cDNA products were fractionated on an agarose column and the fractions containing large species were identified by gel electrophoresis after the polymerase chain reaction [R. K. Saiki *et al.*, *Science* **230**, 1350 (1985)] with a pair of oligonucleotide primers (145 nt apart within the 10-1 sequence) just 5′ of the extension primer. These products were pooled, concentrated, and treated with terminal deoxynucleotidyl transferase (BRL) and dATP, as recommended by the supplier. Second strand synthesis was performed with Taq Polymerase (Cetus, AmpliTaq) and an oligonucleotide containing a linker sequence, 5′-CGGAATTCTCGAGATC(T)$_{12}$-3′. This linker, together with another primer (internal to the extension primer) with an Eco RI restriction site at its 5′ end, was then used for PCR. After digestion with Eco RI and Bgl II, products were purified and cloned in Bluescript KS (Stratagene) by standard procedures. All the recovered clones contained inserts of more than 350 nt. The 3′ end clones were generated with the use of similar procedures. PCR amplification was carried out with the linker described above and an oligonucleotide with the sequence 5′-ATGAAGTCCAAGGATTTAG-3′, which is ~70 nt upstream of the Hind III site at position 5027 (Fig. 2). The products were digested with Hind III and Xho I and cloned in the Bluescript vector. Candidate clones were identified by hybridization with the 3′ end of cDNA T16-4.5. All PCR's were performed for 30 cycles as described by the enzyme supplier.
13. M. Kozak, *Nucleic Acids Res.* **12**, 857 (1984); *Cell* **44**, 283 (1986).
14. Other sequence differences were noted between the normal (T16-4.5) and CF (C1-1/5) cDNA clones. At position 2629 (Fig. 2), T16-4.5 contained a C and C1-1/5 a T, resulting in a change of Leu to Phe. At position 4555, the base was G in T16-4.5 but A in C1-1/5 (Val to Met). The differences may be results of cDNA cloning artifacts or may represent sequence polymorphisms. Specific oligonucleotide hybridization analysis of patient or family DNA should distinguish these possibilities. Since these changes are conserved amino acid substitutions, they are unlikely to be causative mutations. Additional nucleotide differences were observed in the 3′ untranslated region between different cDNA clones and the corresponding genomic DNA sequence.
15. C. F. Higgins, M. P. Gallagher, M. L. Mimmack, S. R. Pearce, *BioEssays* **8**, 111 (1988); C. F. Higgins *et al.*, *Nature* **323**, 448 (1986).
16. P. Gros, J. Croop, D. E. Housman, *Cell* **47**, 371 (1986); C. Chen *et al.*, *ibid.*, p. 381; J. H. Gerlach *et al.*, *Nature* **324**, 485 (1986); P. Gros, M. Raymond, J. Bell, D. Housman, *Mol. Cell. Biol.* **8**, 2770 (1988).
17. Several large families of integral membrane proteins are known, including: (i) A number of ligand-gated ion channels of which the nicotinic acetylcholine receptor [R. M. Stroud and J. Finer-Moore, *Annu. Rev. Cell Biol.* **1**, 317 (1985)] is the prototype. Receptors for the inhibitory neurotransmitters GABA (*33*) and glycine are included in this family. (ii) A family of ion channels with a totally different structural motif are the voltage-gated, sodium, calcium, and potassium channels (*27*). (iii) Involved in the translocation of ions are the structurally related cation pumps such as the Ca^{2+}-ATPase [C. J. Brandl, N. M. Green, B. Korczak, D. H. MacLennan, *Cell* **44**, 597 (1986)], the Na$^+$,K$^+$-ATPase [G. E. Shull and J. B. Lingrel, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4039 (1987)], and the H$^+$,K$^+$-ATPase [G. E. Shull and J. B. Lingrel, *J. Biol. Chem.* **261**, 16788 (1986)]. These are but examples.
18. G. B. Wilson, T. L. Jahn, J. R. Fonseca, *Clin. Chim. Acta* **49**, 79 (1973); V. van Heyningen, C. Hayward, J. Fletcher, C. McAuley, *Nature* **315**, 513 (1985).
19. J. Garnier, D. J. Osguthorpe, B. Robson, *J. Mol. Biol.* **120**, 97 (1978); J. Kyte and R. F. Doolittle, *ibid.* **157**, 105 (1982).
20. D. D. Pless and W. J. Lennarz, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 134 (1977).
21. P. J. Blacksshear, A. C. Nairn, J. F. Kuo, *FASEB J.* **2**, 2957 (1989).
22. J. Hess, W. Wels, M. Vogel, W. Goebel, *FEMS Microbiol. Lett.* **34**, 1 (1986).
23. K. O'Hare, C. Murphy, R. Levis, G. M. Rubin, *J. Mol. Biol.* **180**, 437 (1984).
24. M. J. Friedrich, L. C. DeVeaux, R. J. Kadner, *J. Bacteriol.* **167**, 928 (1986).
25. K. Ohyama *et al.*, *Nature* **322**, 572 (1986).
26. G. F.-L. Ames, *Annu. Rev. Biochem.* **55**, 397 (1986).
27. M. Noda *et al.*, *Nature* **320**, 188 (1986); T. Tanabe *et al.*, *ibid.* **328**, 313 (1987); A. Baumann *et al.*, *EMBO J.* **6**, 3419 (1987).
28. C.-M. Chen, T. K. Misra, S. Silver, B. P. Rosen, *J. Biol. Chem.* **261**, 15030 (1986).
29. J. Krupinski *et al.*, *Science* **244**, 1558 (1989).
30. In addition to the major NBF homologies, searches of the PIR and SWISSPROT data bases detected shorter stretches of sequence homology with other proteins including the following:

```
CFTR 122-136      Y L G I G L C L L F I V R T L
                  : :   :   :   :   : : . :   :
GLNP 198-212      Y L I I T L V L S F I L R R L

CFTR 307-319      S S A F F F S G F F V V F
                  :     : : : . : : : .   :
COX   89-101      S E V F F F A G F F W A F

CFTR 701-713      I L N P I N S I R K F S I
                  : :   : . :   : : : . . :
NaCh 111-123      I L T P F N P I R K L A I

CFTR 1425-1442    D S I Q K L L N E R S L F R Q A I S
                  : : :   : :     : :   : :   :   . :
raf   578-595     D S I K K L R D E R P L F P Q I L S
```

GLNP, glutamine permease of *E. coli* [T. Nohno, T. Saito, J. Hŏng, *Mol. Gen. Genet.* **205**, 260 (1986)]; COX, human cytochrome c oxidase polypeptide III [S. Anderson *et al.*, *Nature* **290**, 457 (1981)]; NaCh, rat brain sodium channel III (*32*); raf, the serine-threonine kinase proto-oncogene of *Xenopus laevis* (*31*). The first two sequences are within membrane spanning segments and probably reflect only coincidental arrangements of the hydrophobic residues suited to this function. In contrast, the latter two sequences are both in polar hydrophilic regions of the proteins. The large extent of amino acid conservation (11 of 13 residues) implies some functional relation between these short segments of the primary structure of the Na^+ channel and CFTR. Similarities between sequences at the same relative locations with respect to the COOH-termini of the raf kinase and CFTR suggest that they may also share at least a small facet of their structures and functions.

31. R. LeGuellec, K. LeGuellec, J. Paris, M. Philippe, *Nucleic Acids Res.* **16**, 10357 (1988).
32. M. Noda *et al.*, *Nature* **312**, 121 (1984); L. Salkoff *et al.*, *Science* **237**, 744 (1987).
33. P. R. Schofield *et al.*, *Nature* **328**, 221 (1987).
34. F. Sanger, S. Nicklen, A. R. Coulsen, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
35. D. Eisenberg, E. Schwarz, M. Komaramy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984).
36. A. P. Feinberg and B. Vogelstein, *Anal. Biochem.* **132**, 6 (1983).
37. J. Rommens *et al.*, *Am. J. Hum. Genet.* **43**, 645 (1988).
38. S. J. Foote *et al.*, *Cell* **57**, 921 (1989).
39. J. P. McGrath and A. Varshavsky, *Nature* **340**, 400 (1989).
40. B. P. Surin, H. Rosenberg, G. B. Cox, *J. Bacteriol.* **161**, 189 (1985).
41. C. F. Higgins *et al.*, *Nature* **298**, 723 (1982).
42. E. Gilson, H. Nikaido, M. Hofnung, *Nucleic Acids Res.* **10**, 7449 (1982).
43. I. D. Hiles, M. P. Gallagher, D. H. Jamieson, C. F. Higgins, *J. Mol. Biol.* **195**, 125 (1987).
44. A. W. Bell *et al.*, *J. Biol. Chem.* **261**, 7652 (1986).
45. R. F. Doolittle *et al.*, *Nature* **323**, 451 (1986).
46. I. J. Evans and J. A. Downie, *Gene* **43**, 95 (1986).
47. D. R. Gill, G. H. Hatfull, G. P. C. Salmond, *Mol. Gen. Genet.* **205**, 134 (1986).
48. J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay, *EMBO J.* **1**, 945 (1982).
49. D. C. Fry, S. A. Kuby, A. S. Mildvan, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 907 (1986).

50. We thank O. Augustinas for the collection of tissues; T. Jensen and R. Baird for the culturing of epithelial cells; L. Naismith for the isolation of RNA; D. Kennedy and D. Markiewicz for technical assistance and M. Buchwald and M. Dean for discussions. Supported by NIH grants DK39690 (F.S.C.) and DK34944 (L.C.T.), the Cystic Fibrosis Foundation (U.S.A.), the Canadian Cystic Fibrosis Foundation, and the Sellers Fund. J.M.R. holds a postdoctoral fellowship from the Medical Research Council (MRC) of Canada and F.S.C. is an Associate Investigator of the Howard Hughes Medical Institute.

7 August 1989; accepted 18 August 1989

Identification of the Cystic Fibrosis Gene: Genetic Analysis

Bat-sheva Kerem, Johanna M. Rommens, Janet A. Buchanan, Danuta Markiewicz, Tara K. Cox, Aravinda Chakravarti, Manuel Buchwald, Lap-Chee Tsui

Approximately 70 percent of the mutations in cystic fibrosis patients correspond to a specific deletion of three base pairs, which results in the loss of a phenylalanine residue at amino acid position 508 of the putative product of the cystic fibrosis gene. Extended haplotype data based on DNA markers closely linked to the putative disease gene locus suggest that the remainder of the cystic fibrosis mutant gene pool consists of multiple, different mutations. A small set of these latter mutant alleles (about 8 percent) may confer residual pancreatic exocrine function in a subgroup of patients who are pancreatic sufficient. The ability to detect mutations in the cystic fibrosis gene at the DNA level has important implications for genetic diagnosis.

Although the frequency of cystic fibrosis (CF) is not uniformly high among all Caucasian populations, a consensus estimate is that it occurs once in 2000 live births (*1*). On the basis of the autosomal recessive mode of inheritance for this disease, a mutant allele frequency of 0.022 may be derived. Several different mechanisms, including high mutation rate (*2*), heterozygote advantage (*3*), genetic drift (*4*), multiple loci (*5*), and reproductive compensation (*6*), have been proposed in attempts to explain the high incidence and, indirectly, the nature of the CF mutations. Although some of these hypotheses could not be further addressed because of the lack of knowledge about the basic defect in CF, several important observations have been made during the past few years through genetic analysis of the families of affected individuals (*7–20*).

Extensive linkage analysis provides evidence for the existence of a single CF locus on human chromosome 7 (region q31) (*7–10, 21*). The detection of allelic and haplotype association between the CF

B. Kerem, J. M. Rommens, J. A. Buchanan, D. Markiewicz, M. Buchwald and L.-C. Tsui are in the Department of Genetics, Research Institute, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada. T. K. Cox and A. Chakravarti are in the Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA 15261. M. Buchwald and L.-C. Tsui are also members of the Departments of Medical Genetics and Medical Biophysics, University of Toronto, Toronto, Ontario M5S 1A8, Canada.