VARIATION AND EVOLUTION OF CAULIFLOWER MOSAIC VIRUS ISOLATES

By

KELLY DAWN CHENAULT

Bachelor of Science

Oklahoma State University

Stillwater, Oklahoma

1987

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 1992

VARIATION AND EVOLUTION OF CAULIFLOWER

MOSAIC VIRUS ISOLATES

Thesis Approved:

Ulrich Melcher
Thesis Adviser

J. L. Sherwood

Franklin R. Leach

Richard C. Essenberg

Thomas C. Collins
Dean of the Graduate College

PREFACE

The focus of my doctoral research has been to obtain a better understanding of virus evolution. I chose to approach this subject by studying variability and phylogenetic relationships among different isolates of cauliflower mosaic virus (CaMV). Thus, there were essentially two objectives to my research project. First, I would examine variation among CaMV isolates. To complete this objective, I sequenced the complete genome of three isolates of CaMV: NY8153, CMV-1, and BBC. These sequences were then aligned with those of previously sequenced isolates. A CaMV consensus sequence was constructed and used to examine variability among CaMV isolate genomes. Specifically, I identified and characterized isolate-specific base substitutions, deletions, and insertions. These data were used to examine how and when mutations occur in the CaMV life cycle. The second objective of my research was to determine the phylogenetic relationships among CaMV isolates. I accomplished this task by using the CaMV nucleotide sequence alignment to construct phylogenetic trees. Species and gene trees were constructed by three different methods: parsimony, maximum likelihood, and distance. These phylogenetic trees were used to infer a certain genetic relationship between the CaMV

isolates and give probable explanations of how this relationship arose.

The results in this thesis are the components of four separate manuscripts (authored by myself and Dr. Ulrich Melcher) to be submitted for publication. Therefore, the results for each manuscript are represented as four separate parts of the Results section. Part 1 refers to the nucleotide sequence of CaMV isolate NY8153. Before, I began my doctoral research, David Steffens had already sequenced parts of the NY8153 isolate. Thus he is included as an author on the NY8153 manuscript, and I acknowledge his contribution to that work. Part 3 of the results section includes the nucleotide sequence of CaMV isolate CMV-1. A decision was made to submit this sequence for publication as part of a manuscript, written mainly by Ulrich Melcher, that contains the results of a separate project.

I wish to express my sincere gratitude to the Department of Biochemistry and the Robert Glenn Rapp Foundation for providing me with the financial support necessary to complete my graduate studies. I want to thank Dr. Franklin Leach who took me into his laboratory as an undergraduate and greatly influenced my career goals. I am grateful to the other members of my committee, Dr. Richard Essenberg and Dr. John Sherwood, for their advice and patience. In particular, I wish to thank my major adviser, Dr. Ulrich Melcher whose experience and wisdom has helped me to mature both as a scientist and as a person.

I would like to thank Bruce Roe from Oklahoma University for his help with the computer-aided sequence analysis described in this thesis.

Special thanks go to Sue Ann Hudiburg and Dr. George Odell for their kindness and friendship. Thanks also to Ann Williams and Dr. Robert Lartey for their support and advice. I especially wish to thank Dr. Rod Pennington and Dr. Steve Hartson, my former lab mates and fellow graduate students, for all of their friendship and helpful discussions.

On a more personal note, I want to thank my wonderful husband, Paul Chenault, for his love, dedication, and understanding. I also wish to thank my sister, Kristie Newby, for all of her love and support. Special thanks go to my mother, Beverly Hooper, for her never-ending, unconditional love. Finally, I wish to thank my father, the late Dr. Robert C. Hooper, who is largely responsible for my independence, motivation, and perseverance. He is truly my hero, and I dedicate this work to him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The rapid accumulation of viral nucleotide sequence data has lead to the development of detailed viral phylogenies based on objective criteria. Analysis of the genomic sequences of RNA viruses has resulted in numerous reports and several reviews concerning RNA virus evolution (41, 96). One conclusion of these studies is that RNA viruses mutate and evolve at a much higher rate than do DNA viruses, because RNA viruses lack the proof-reading enzymes that assure fidelity of DNA replication. Not all viruses fit cleanly into the category of an "RNA" or "DNA" virus. Retroviruses, such as human immunodeficiency virus 1 (HIV-1), use reverse transcription to replicate their RNA genomes via a DNA intermediate, and thus have an added error-prone step in their replication cycle. Retroviruses have an elevated mutation rate relative to other RNA viruses (39). Pararetroviruses contain DNA as their genetic material in the mature virion, but replicate through an RNA intermediate by employing reverse transcriptase. Pararetroviruses include vertebrate hepadnaviruses, bacilliform plant badnaviruses, and icosahedral plant caulimoviruses. Although

pararetroviruses use the same mechanisms as retroviruses to replicate their genomes, they have a mutation rate one to two orders of magnitude lower than that of retroviruses (39, 78). To further investigate pararetrovirus mutation and evolution, I examined the inter-isolate relationships of the type member of caulimoviruses, cauliflower mosaic virus.

CHAPTER II

LITERATURE REVIEW

CaMV Background

General

The caulimovirus group has eight certain members: carnation etched ring virus (CERV), dahlia mosaic virus (DMV), figwort mosaic virus (FMV), mirablis mosaic virus (MMV), strawberry vein-banding virus (SVBV), soybean chlorotic mottle virus (SCMV), peanut chlorotic streak virus (PCSV), and the type member, cauliflower mosaic virus (CaMV) (47, 91). CaMV virions are isometric particles about 50 nm in diameter. Approximately 80% of the virion is protein. The virion shell consists of a single protein with a molecular weight of 42Kd. The virus genome is double-stranded circular DNA about 8 kbp in size and is sandwiched between two layers of the protein shell, leaving the virion core empty. The host range of CaMV is limited to the Cruciferae and some Solanaceae. Virus transmission may occur mechanically (via inoculation), but is normally carried out by aphids in a semi-persistent manner. Most likely due to mutation, some CaMV isolates are aphid non-transmissible.

Following inoculation of susceptible plants with virus or viral DNA, systemic infection usually occurs and virions are produced in subsequently formed leaves. CaMV symptoms (usually isolate specific) may include chlorotic spots, necrotic flecks, mosaic and mottling, vein-clearing, vein-banding, stunting, crinkling, and paling of leaves.

Genome Organization

The DNA of CaMV virions has single-stranded interruptions at specific locations on the molecule. In general, caulimovirus DNA has one gap in one strand and 1-3 gaps in the other. DNA sequencing has shown that these 'gaps' are triple-stranded regions (overlaps) (44). The strand with a single gap is termed the minus (-) strand and eventually becomes the template for transcription. Ribonucleotides are associated with CaMV DNA and are believed to be remnants of primers of DNA synthesis. The minus strand of CaMV DNA serves as a template for two major transcripts, the 19S and 35S RNAs. Six major and two minor open reading frames (ORFs) are present in the 35S RNA. Probable functions for the gene products of ORFs 1, 2, 4, 5, and 6 are known. The genomic positions of these ORFs and their possible functions are shown in Table I. The basic structure of all retro- and pararetroviruses includes genes coding for a (1) structural protein (gag), (2) enzymatic functions (pol), and (3) an envelope (env) protein. CaMV genes have been suggested to

TABLE I

CAULIFLOWER MOSAIC VIRUS OPEN READING FRAME POSITIONS AND PROPOSED FUNCTIONS

Open Reading Frame	Nucleotide Position*	Proposed Function(s)
1	364-1344	movement
2	1349-1825	aphid transmission
4	2201-3667	capsid
5	3633-5669	reverse transcriptase
6	5776-7335	inclusion body matrix; transactivator of translation; host-range determinant

*According to the numbering used for the Cabbage S isolate (32).

correspond to these retroviral regions: ORF 4 is 'gag', ORF 5 is 'pol', and ORF 6 is 'env'(50).

Replication Cycle of CaMV

CaMV uses a replication strategy very reminiscent of that of the retroviruses (11, 50). As previously mentioned, CaMV may enter a host cell either by aphid transmission or mechanical inoculation. After uncoating of the virus, the gaps in the genome are repaired in the nucleus using host enzymes. The resulting DNA molecule is transcribed by host RNA polymerase II producing the two major transcripts. Both transcripts are polyadenylated and are transferred to the cytoplasm where they are translated by host machinery. The smaller transcript (19S RNA) codes for the inclusion body matrix protein. The 35S RNA contains the complete viral coding information and also serves as a template for reverse transcription, which produces the minus strand of the double-stranded DNA genome. Reverse transcription is the replication step which identifies CaMV as a pararetrovirus. The CaMV 35S RNA is similar to that of the retroviruses in that it possesses a direct terminal repeat. Also, near the 5' terminus of the CaMV 35S RNA, there is a 13 nucleotide sequence complementary to the 3' terminus of $tRNA_{met}$. Reverse transcription is thought to occur in replication complexes which are found in the same cell fraction as the cytoplasmic inclusion bodies (69). Minus strand synthesis is

initiated when a host $tRNA_{met}$ primer binds to the 35S RNA. The CaMV ORF 5 product, reverse transcriptase, then copies the RNA template to its 5' end where the enzyme stops, producing a small DNA molecule (sa DNA). An obligatory switch in template strands occurs as the reverse transcriptase jumps to the 3' end of the 35S RNA and resumes production of the DNA minus strand. As minus strand synthesis occurs, the RNase H activity of the reverse transcriptase rapidly degrades the already reverse transcribed 35S RNA. Polypurine patches of RNA that aren't degraded by this activity serve as primers for the synthesis of the plus strand. After the plus strand is made, these primers are displaced and trimmed producing the gaps that are present in the encapsidated DNA.

Some features of retro- and pararetrovirus replication may cause it to be an error-prone process, thus leading to accumulation of mutations and possible rapid virus evolution. First, reverse transcriptase and RNA polymerase II lack proofreading functions. Another factor that contributes to mutation in retro- and pararetroviruses is the template switch involved in the reverse transcription phase of their life cycle. If this template switch occurs abnormally, viral recombinants may arise. Evidence for this mechanism of recombination does exist for CaMV (10, 37, 54, 65, 105, 107). Retroviruses possess a characteristic which increases the chance that a replication error will occur. Retroviruses encapsidate two copies of their RNA genome, which has been

shown to result in high rates of recombination (57). Recombination between these two genomic RNAs has been shown to occur during DNA minus strand synthesis (as with the pararetroviruses), and also during DNA plus strand synthesis via a mechanism termed strand displacement-assimilation (57). Strand displacement-assimilation occurs when two DNA minus strands are made in the same virion. Since plus strand synthesis is initially discontinuous, a (+) strand fragment from one minus strand may hybridize with the alternate minus strand and be incorporated into that DNA molecule.

Because of the reverse transcription step in their life cycles, retro- and pararetroviruses may be evolving faster than those viruses without these steps. Rates of evolution for RNA genomes are much higher that those of DNA genomes, mainly due to the error-prone nature of RNA polymerases compared to DNA polymerases. DNA genomes have an estimated mutation rate between 10^{-7} and 10^{-11} substitutions per site per year. Some RNA viruses mutate rapidly while others do not. Gojobori and Yokoyama (40) estimated the mutation rate for the v-mos gene of Maloney murine sarcoma virus to be 1.31×10^{-3} substitutions per site per year, a rate that is a million-fold higher than c-mos, its cellular homolog. The human immunodeficiency virus (HIV-1) mutation rate has been estimated at 10^{-2} to 10^{-3} substitutions per site per year (45). One plant RNA virus, turnip yellow mosaic virus, has an estimated mutation rate of only 1.3×10^{-7} substitutions per site per year (7). The mutation rate and evolution of RNA

viruses (including retroviruses) have been extensively studied (16, 41, 51, 96). Less has been said about the pararetroviruses. Pennington and Melcher (78) estimated the mutation rate of CaMV to be 6 x 10^{-4} substitutions per site per year. In order to learn more about caulimovirus mutation and evolution, we constructed a CaMV base substitution profile and inferred phylogenetic relationships among different CaMV isolates.

Mechanisms of Mutation

There are several types of DNA sequence change and different mechanisms by which these changes can occur. These processes deserve consideration here since nucleotide sequence changes are used in studies of molecular evolution both for estimating the rate of evolution and for reconstructing evolutionary relationships.

Base substitutions occur at about 5% of the nucleotide positions in CaMV DNA when pairs of isolates were compared (3). Substitutions are usually classified into transitions or transversions. Transitions, which are more common, involve the substitution of one pyrimidine for another, or of one purine by another; thus a G-C pair is exchanged for an A-T pair or vice versa. Transversions require the replacement of a purine by a pyrimidine or vice versa, so that an A-T pair becomes a T-A or C-G pair. One source of transitions is the chemical conversion of one base to another. For example, deamination of cytosine converts it to uracil, which pairs

with adenine, resulting in a C-to-T transition in the next round of DNA replication. Base mispairing, the pairing of bases in defiance of Watson-Crick rules (104), may also result in transitions or in the less common transversions. Some base substitutions in retroviruses may occur by misincorporation due to transient template misalignment by reverse transcriptases (5, 63, 77). Although pararetroviruses such as CaMV also use reverse transcriptase, no evidence for this mechanism of base substitution has been found for this virus group. Another pattern of substitution, hypermutation, is characterized by extensive yet monotonous nucleotide substitution within a specific sequence. For example, in a given sequence, all A's may be converted to G's. Hypermutation has been shown to occur for several viruses (8, 106). Mispairing of A and I forms a 'wobble' base pair (6) that results in an A -> G transition. Bass et al. (4) attribute A ->G hypermutation to the RNA unwinding/modifying activity present in most eukaryotic cells. This activity introduces A-to-I changes in duplex RNA. The I residues would then result in the incorporation of C residues in one strand, giving rise to A-to-G changes in the other. Hypermutation is not known as a mechanism of substitution for CaMV DNA.

Another type of sequence change is the deletion of single or stretches of nucleotides. Some deletions in CaMV DNA have been attributed to RNA splicing. Following S-Japan isolate infection, 1/3 of the isolated progeny contain DNA

that lacks 856 nucleotides in ORF 1 (48). The missing region resembles an intron in that the ends are similar to splice donor and splice acceptor sequences. When point mutations disrupting these sequences were introduced, deletion of the region between them no longer occurred. Hohn et al., (49) inserted an intron into ORF 2 of CaMV and reported that upon several passages in host plants, progeny virions accumulated which had lost the intron due to apparent splicing at splice signals. Pennington and Melcher (78) observed deletion of an intron-like sequence in CaMV which did not occur when the splice donor site was mutated. Vaden and Melcher (105) also reported the deletion of sequences that resembled introns from CaMV DNA.

Most of the CaMV genome is necessary for infection (56). However, deletion of parts of CaMV DNA may result in virions that are still viable. CaMV isolate CM4-184 lacks ORF 2 (53) which in other isolates is required for aphid transmission (1, 110). Despite the ORF 2 deletion, CM4-184 will produce systemic infection if mechanically inoculated on susceptible leaves. The mechanism behind the CM4-184 deletion and some other deletions in CaMV DNAs is most likely template switching during reverse transcription. These template switches may be intra- or intermolecular. There are two stretches of 9 nucleotides at each end of the ORF 2 deletion that are imperfect direct repeats (16/18 nucleotides identical) (15). These nearly identical regions provide a potential site for an intramolecular template switch during

reverse transcription that would lead to the CM4-184 deletion.

There have been few reports of natural insertions resulting in virus that was still viable. Penswick (79) reported a natural duplication of part of the ORF 4-5 region in one CaMV isolate. Restriction fragment length polymorphisms (RFLPs) between different CaMV isolates have been used to show variation in their nucleotide sequences (35, 58). Hull (58) reported differences between CaMV isolate restriction patterns that suggested possible insertions in the DNAs of the Bari 4 and Australian isolates. Many CaMV isolate genomes have now been sequenced. Comparisons of these sequences with each other can serve as another method to distinguish insertion from deletion events. In addition, sequence comparison can also aid in identifying point mutation events.

Recombination between DNA sequences has played a role in the generation of CaMV variants. In the earliest report of recombination in CaMV, Howell et al.(56) reported successful infection of hosts by co-inoculating turnips with non-infectious parent CaMV DNAs. Based on restriction data, progeny DNAs did not contain the mutations present in parental DNAs, suggesting recombination had occurred. Chimeric progeny DNAs (recombinant DNAs that have sequences from each parent DNA) have often been recovered as a result of host inoculation with pairs of mutant non-infectious CaMV DNAs (10, 37, 56, 65, 105, 107).

Inoculation of a susceptible host with greater than full genome length CaMV clones has been shown to result in infection (65, 108). Some of the clones used for inoculation were constructed in a manner which allowed possible production of a full-length 35S RNA. Other clones contained sequences that disrupted the transcription template, suggesting some of the infectious progeny resulted from recombination. Grimsley et al. (42) analyzed progeny obtained from infection with a hybrid plasmid containing segments of CaMV DNA (full length genome of CM4-184 and a fragment of Cabbage S) and the T-DNA of *Agrobacterium tumefaciens*. Some of the chimeric viral progeny may have resulted from recombination, while the majority of the progeny were likely the result of chimeric 35S RNA production. Chimeras may occur naturally (15, 105). Isolate CM4-184 is one example of such a chimera. The CM4-184 genome is identical to that of isolate CM1841, except for the large intergenic region which is closely related to that of isolate Cabbage S (15). Vaden and Melcher (105) also reported a natural chimera, W, that seemed to be produced by recombination between an unidentified CaMV isolate and Cabbage B-JI.

Some of the observed recombination between CaMV DNAs may have resulted from double-stranded homologous crossover (33, 34, 37, 56, 65, 105, 108). Gene conversion has been suggested to occur for CaMV DNA. Choe et al. (10) reported restriction-fragment based evidence consistent with the formation and repair of heteroduplexes in CaMV DNA, but Vaden

and Melcher later examined these findings along with new evidence and concluded that a misinterpretation had occurred (105). Melcher et al. (75) suggested that gene conversion contributed to the recovery of only one type of progeny upon mixed infection with mutant and wild-type CaMV CabbS DNAs. Zhang and Melcher (111) later showed that this recovery of only one type of progeny was instead due to strong dominance of one isolate over another. However, Zhang and Melcher (111) also reported evidence of intergenomic genetic exchange at extensive regions of homology between CaMV DNAs, suggesting either gene conversion or a double homologous crossover may have occurred. Moreover, Melcher et al. (75) suggested that gene coversion may have contributed to interference when host plants were inoculated with mixtures of mutant and wild-type CaMV DNAs. Still, no substantial evidence exists of gene conversion occurring for CaMV DNA.

When the reverse transcription model of replication was suggested for CaMV (44, 55, 59) another mechanism of recombination between CaMV DNAs was uncovered. As discussed in Chapter 2, abnormal template switches that may occur during reverse transcription can result in intra- or intermolecular recombination. Recombination between two homologous sequences of different isolates creates a junction that marks the region in the recombinant DNA where the event took place. The mapping of recombinant sequence juctions to sites of normal CaMV template switches or the start site of reverse transcription suggests that recombination between

CaMV DNAs occurs during reverse transcription (15, 43, 105). There are now many reports of recombination of CaMV RNAs via template switches during reverse transcription. Repeats in sequence, such as those at each end of the 35S RNA, facilitate template switching by reverse trancriptase. These template switches may occur at regions of extensive homology during reverse transcription resulting in legitimate recombination. Illegitimate recombination can result from template switches at short stretches of similar sequence. Both legitimate (15, 43, 98, 105) and illegitimate (42, 53, 79) template switches have been well documented for CaMV.

CHAPTER III

RESULTS

The Complete Nucleotide Sequence of Cauliflower Mosaic Virus Isolate NY8153

Cauliflower mosaic virus (CaMV) is the type member of the caulimoviruses, a group of plant viruses with double-stranded DNA as their genetic material. Caulimoviruses have a restricted host range, usually one or two families. CaMV mainly infects members of the crucifereae and solanaceae. The details of CaMV molecular biology have been extensively reviewed (11). The double-stranded genome of CaMV contains three discontinuities (gaps), one in the minus (transcribed) strand, and two in the plus strand. There are two major transcripts of CaMV (Table II). The larger transcript (35S) has eight tightly packed potential reading frames (ORFs) and a non-coding region of approximately 700 bp. The known functions of five genes are shown in Table II.

Several CaMV isolates are known and the genomes of some have been sequenced completely. Here, we report the nucleotide sequence of CaMV isolate NY8153 (Figure 1). Disease symptoms induced on turnip by NY8153 have been described (72). NY8153 DNA was cloned into pBR322 (1), and

TABLE II

CHARACTERISTICS OF CAULIFLOWER MOSAIC VIRUS ISOLATE NY8153

Virus Group: Caulimoviridae
Particle Type: Isometric
Genome Type and Size: Double-stranded DNA; 8 kbp
Structural Features: 8 Potential open reading frames:

ORF	Start*§	End*	MW £	Function
1	364	1347	37	Movement
2	1349	1828	18	Aphid transmission
3	1830	2219	14	?
4	2201	3667	57	Coat protein precursor
5	3627	5669	79	Reverse transcriptase
6	5773	7332	58	Inclusion body protein/ Tranlation trans-activator
7	13	303	11	?
8	3259	3583	12	?

Two major transcripts:

RNA	Start*
19S	5761
35S	7432

Polyadenylation signal*: 7604-7609
$tRNA_{met}$ primer binding site*: 8028-13

Techniques: Restriction, ligation, cloning, nucleotide sequencing (73).
Accession No.: M90541

*Arabic numerals indicate nucleotide position where position 1 is equivalent to that of the DNA of the Cabbage S isolate (23).
§"Start" indicates first ATG
£Molecular weights of proteins in KDa, based upon calculation by MacVector™

Figure 1. The complete nucleotide sequence of CaMV isolate NY8153. The derived amino acid sequences of the six major CaMV ORFs are shown in one letter code below the nucleotide sequence. This figure spans pages 19-29.

```
  1  GGTATCAGAGCCATGAATCGGTTTAAAGACCAAACTCAAGAGGGTAAAACCTCATCAAAA   60

 61  TACGAAAGAGTTCTTAACTCTAAAGATAAAAGATCTTTCAAGATTAAAACTAGTTCCCTC  120

121  ACACCGGTGACCGACAGGTTTACCACCGTAAGGTTTCAGAACAACATCGAATGCGTTTAC  180

181  GCCAACTTCGACTCTCAGCTCAAGTCGTCGTACGATGGTAGATCTAAAAAGATCAAGAAT  240

241  CTAAGCCTTAAAAATCTTAGATGTCACGAAGCCTTCCTCAGGAAGTACCTTCTGGAACAA  300

301  TAAATCTCTCTGAGAATAGTACTCTATTGAGTATCCACAGATAAAATAATCTTCTGTGTT  360

361  GAGATGGATTTGTATCCAGAAGAAAAGACCCAAAGCAAGCAATCGCATAATTCTGAAAAT  420
        M  D  L  Y  P  E  E  K  T  Q  S  K  Q  S  H  N  S  E  N

421  AATATGCAAATATTTAAATCAGAAAATTCGGATGGATTCTCCTCCGATCTAATGATCTCA  480
     N  M  Q  I  F  K  S  E  N  S  D  G  F  S  S  D  L  M  I  S

481  AACGATCAATTAAAAAATATCTCTAAAACCCAATTAACTTTGGAAAAAGAAAAGATATTT  540
     N  D  Q  L  K  N  I  S  K  T  Q  L  T  L  E  K  E  K  I  F

541  AAAATGCCTAACGTTTTATCTCAAGTTATGAAAAAAGCGTTTAGCAGGAAAAACGAGATT  600
     K  M  P  N  V  L  S  Q  V  M  K  K  A  F  S  R  K  N  E  I

601  CTCTACTGCGTCTCGACAAAAGAATTATCAGTGGACATTCACGATGCCACAGGTAAGGTA  660
     L  Y  C  V  S  T  K  E  L  S  V  D  I  H  D  A  T  G  K  V

661  TATCTTCCTTTAATCACTAAAGAGGAGATAAATAAAAGACTTTCCAGTTTAAAACCTGAA  720
     Y  L  P  L  I  T  K  E  E  I  N  K  R  L  S  S  L  K  P  E

721  GTCAGAAAGACCATGTCCATGGTTCATCTTGGAGCGGTCAAAATATTGCTTAAAGCTCAA  780
     V  R  K  T  M  S  M  V  H  L  G  A  V  K  I  L  L  K  A  Q
```

```
781  TTTCGAAATGGGATTGATACCCCAATCAAAATTGCTTTAATCGATGATAGAATTAATTCT  840
     F  R  N  G  I  D  T  P  I  K  I  A  L  I  D  D  R  I  N  S

841  AGAAGAGATTGCCTTCTCGGTGCAGCCAAAGGTAATCTAGCATACGGTAAGTTTATGTTT  900
     R  R  D  C  L  L  G  A  A  K  G  N  L  A  Y  G  K  F  M  F

901  ACTGTATACCCCAAGTTTGGAATAAGCCTTAATACCCAAAGACTTAACCAAACCCTAAGC  960
     T  V  Y  P  K  F  G  I  S  L  N  T  Q  R  L  N  Q  T  L  S

961  CTTATTCATGATTTTGAAAATAAAAATCTTATGAATAAAGGTGATAAAGTTATGACCATA  1020
     L  I  H  D  F  E  N  K  N  L  M  N  K  G  D  K  V  M  T  I

1021 ACCTATATCGTAGGATATGCATTAACTAATAGTCATCATAGCATAGATTATCAATCGAAT  1080
     T  Y  I  V  G  Y  A  L  T  N  S  H  H  S  I  D  Y  Q  S  N

1081 GCTACAATTGAACTAGAAGACGTATTTCAAGAAATTGGAAATGTCCAGCAATGTGATTTC  1140
     A  T  I  E  L  E  D  V  F  Q  E  I  G  N  V  Q  Q  C  D  F

1141 TGTACAATACAGAATGACGAATGTAATTGGGCCATTGATATAGCCCAAAACAAAGCCTTA  1200
     C  T  I  Q  N  D  E  C  N  W  A  I  D  I  A  Q  N  K  A  L

1201 TTAGGAGCTAAAACCCAATCCCAAATTGGTAATAGTCTTCAAATAGGAAACAGTGCTTCA  1260
     L  G  A  K  T  Q  S  Q  I  G  N  S  L  Q  I  G  N  S  A  S

1261 TCCTCTAATACTGAAAATGAATTAGCTAGGGTAAGCCAAAACATAGATCTTTTAAAGAAT  1320
     S  S  N  T  E  N  E  L  A  R  V  S  Q  N  I  D  L  L  K  N

1321 AAATTAAAAGAAATCTGTGGAGAATAAAATGAGCATTACGGGTCAACCGCATGTTTATAA  1380
     K  L  K  E  I  C  G  E  *  M  S  I  T  G  Q  P  H  V  Y  K

1381 AAAGGATACTATTATTAGACTAAAACCATTGTCTCTTAATAGTAATAATAGAAGTTATGT  1440
      K  D  T  I  I  R  L  K  P  L  S  L  N  S  N  N  R  S  Y  V

1441 TTTTAGTTCCTCAAAAGGGAACATTCAAAATATAATTAATCATCTTAACAACCTCAATGA  1500
      F  S  S  S  K  G  N  I  Q  N  I  I  N  H  L  N  N  L  N  E
```

```
1501  GATTGTAGGAAGAAGCTTACTCGGAATATGGAAGATCAACTCATACTTCGGACTAAGCAA  1560
        I  V  G  R  S  L  L  G  I  W  K  I  N  S  Y  F  G  L  S  K

1561  AGACCCTTCGGAGTCCAAATCAAAAAACCCGTCAGTTTTTAATACTGCAAAAACCATTTT  1620
        D  P  S  E  S  K  S  K  N  P  S  V  F  N  T  A  K  T  I  F

1621  TAAGAGTGGGGGGGTTGATTACTCGAGCCAATTAAAGGAAATAAAATCCCTTTTAGAAGC  1680
        K  S  G  G  V  D  Y  S  S  Q  L  K  E  I  K  S  L  L  E  A

1681  TCAAAACACTAGAATTAAAAGTCTAGAAAATGCAATTCAATCCTTAGATAATAAGATTGA  1740
        Q  N  T  R  I  K  S  L  E  N  A  I  Q  S  L  D  N  K  I  E

1741  ACCAGAGCCCTTAACTAAAGAAGAAGTTAAAGAGCTAAAAGAATCGATTAACTCGATCAA  1800
        P  E  P  L  T  K  E  E  V  K  E  L  K  E  S  I  N  S  I  K

1801  AGAAGGATTAAAGAATATTATTGGCTGAAATGGCTAATCTTAATCAAATCCAAAAAGAAG  1860
        E  G  L  K  N  I  I  G  *  M  A  N  L  N  Q  I  Q  K  E  V

1861  TCTCTGAAATCCTCAGTGACCAAAAATCCATGAAATCGGATATAAAAGCTATCTTAGAAA  1920
          S  E  I  L  S  D  Q  K  S  M  K  S  D  I  K  A  I  L  E  M

1921  TGCTAGGATCCCAAAATCCTATTAAAGAAAGCTTAGAAGCCGTTGCAGCGAAAATCGTTA  1980
          L  G  S  Q  N  P  I  K  E  S  L  E  A  V  A  A  K  I  V  N

1981  ATGACTTAACCAAGCTCATCAATGATTGTCCTTGTAACAAAGAAATATTAGAAGCCTTAG  2040
          D  L  T  K  L  I  N  D  C  P  C  N  K  E  I  L  E  A  L  G

2041  GCAATCAGCCTAAAGAGCAACTAATAGAACAACCTAAAGAAAAAGGCAAAGGTCTTAATC  2100
          N  Q  P  K  E  Q  L  I  E  Q  P  K  E  K  G  K  G  L  N  L

2101  TAGGAAAATACTCTTACCCCAATTACGGTGTAGGAAATGAAGAATTAGGATCCTCTGGAA  2160
          G  K  Y  S  Y  P  N  Y  G  V  G  N  E  E  L  G  S  S  G  N

2161  ACCCTAAAGCTTTAACCTGGCCCTTCAAAGCTCCAGCAGGATGGCCGAATCAATTTTAGA  2220
          P  K  A  L  T  W  P  F  K  A  P  A  G  W  P  N  Q  F  *
                                                    M  A  E  S  I  L  D
```

```
2221  CAGAACCATTAATAGGTTTTGGTATAATCTGGGAGAAGATTGTCTCTCAGAAAGTCAATT  2280
       R  T  I  N  R  F  W  Y  N  L  G  E  D  C  L  S  E  S  Q  F

2281  TGACCTTATGATAAGGTTAATGGAAGAGTCCTTGAGCGGGGACCAAATTATTGATCTAAC  2340
       D  L  M  I  R  L  M  E  E  S  L  S  G  D  Q  I  I  D  L  T

2341  CTCTCTACCTAGTGATAATTTGCAGGTCGAACAGGTTATGACAACTACCGAAGACTCGAT  2400
       S  L  P  S  D  N  L  Q  V  E  Q  V  M  T  T  T  E  D  S  I

2401  CTCGGAAGAATCAGAATTCCTTCTAGCAATAGGAGAAACATCTGAAGACGAAAGCGATTC  2460
       S  E  E  S  E  F  L  L  A  I  G  E  T  S  E  D  E  S  D  S

2461  AGGAGAAGAACCTGAATTCGAACAAGTTCGAATGGATCGAACAGGAGGAACGGAGATTCC  2520
       G  E  E  P  E  F  E  Q  V  R  M  D  R  T  G  G  T  E  I  P

2521  CAAAGAAGAAGATGGTGAACCATCTAGATACAATGAGAGAAAGAGAAAGACCACGGAGGA  2580
       K  E  E  D  G  E  P  S  R  Y  N  E  R  K  R  K  T  T  E  D

2581  CCGGTACTTTCCAACTCAACCAAAGACCATTCCAAGACAAAAGCAAACGTCTATGGGAAT  2640
       R  Y  F  P  T  Q  P  K  T  I  P  R  Q  K  Q  T  S  M  G  M

2641  GCTCAACATTGACTGCCAAACCAATCGAAGAACCTTAATCGATGATTGGGCAGCAGAAAT  2700
       L  N  I  D  C  Q  T  N  R  R  T  L  I  D  D  W  A  A  E  I

2701  CGGACTGATAGTCAAGACCAATAGAGAAGACTATCTGAATCCAGAAACAATACTACTCTT  2760
       G  L  I  V  K  T  N  R  E  D  Y  L  N  P  E  T  I  L  L  L

2761  GATGGAACACAAAACATCAGGAATAGCCAAGGAGTTAATCCGAAATACAAGATGGAACCG  2820
       M  E  H  K  T  S  G  I  A  K  E  L  I  R  N  T  R  W  N  R

2821  TACTACCGGCGATATCATAGAACAGGTGATCGATCGGATGTACACCATGTTCTTAGGACT  2880
       T  T  G  D  I  I  E  Q  V  I  D  R  M  Y  T  M  F  L  G  L

2881  TAACTACTCCGACAACAAGGTTGCTGAAAAGATAGACGAGCAAGAGAAGGCCAAGATCAG  2940
       N  Y  S  D  N  K  V  A  E  K  I  D  E  Q  E  K  A  K  I  R
```

```
2941  AATGACCAAACTCCAGCTCTGCGACATCTGCTACCTTGAAGAATTTACATGTGATTATGA  3000
        M  T  K  L  Q  L  C  D  I  C  Y  L  E  E  F  T  C  D  Y  E

3001  AAAGAACATGTACAAGACGGAACTGGCGGATTTCCCAGGATATATCAACCAGTACCTGTC  3060
        K  N  M  Y  K  T  E  L  A  D  F  P  G  Y  I  N  Q  Y  L  S

3061  AAAAATCCCCATCATAGGAGAAAAAGCGCTAACACGCTTTAGGCATGAAGCCAACGGAAC  3120
        K  I  P  I  I  G  E  K  A  L  T  R  F  R  H  E  A  N  G  T

3121  CAGCATCTACAGCTTAGGTTTCGAGCGAAAGATATGCAAAGAAGAACTATCTAAAATTCG  3180
        S  I  Y  S  L  G  F  E  R  K  I  C  K  E  E  L  S  K  I  R

3181  CGACTTATCCAAGAACGAGAAGAAGTTGAAGAAATTCAACAAGAAGTGCTGCAGCATCGA  3240
        D  L  S  K  N  E  K  K  L  K  K  F  N  K  K  C  C  S  I  E

3241  AGAAGCTTCAGCAGAATATGGATGTAAGAAGACATCTACCAAAAAGTATCACAAGAAGCG  3300
        E  A  S  A  E  Y  G  C  K  K  T  S  T  K  K  Y  H  K  K  R

3301  ATACAAGAAAAAATATAAGGCTTATAAACCTTATAAGAAGAAGAAGAAATTCCGATCCGG  3360
        Y  K  K  K  Y  K  A  Y  K  P  Y  K  K  K  K  K  F  R  S  G

3361  AAAATACTTCAAGCCCAAAGAGAAGAAGGGCTCAAAGCAAAAGTATTGCCCAAAAGGCAA  3420
        K  Y  F  K  P  K  E  K  K  G  S  K  Q  K  Y  C  P  K  G  K

3421  GAAAGACTGCAGGTGTTGGATCTGCAATATCGAAGGTCATTACGCCAACGAATGTCCTAA  3480
        K  D  C  R  C  W  I  C  N  I  E  G  H  Y  A  N  E  C  P  N

3481  TCGACAAAGCTCGGAAAAGGCTCACATCCTTCAACAAGCAGAAAAAGTTGGCCTCCAGCC  3540
        R  Q  S  S  E  K  A  H  I  L  Q  Q  A  E  K  V  G  L  Q  P

3541  CATTGAAGCTCCCTATGAAGGAGTTCAAGAAGTATTCATCTTAGAATACAAAGAAGAGGA  3600
        I  E  A  P  Y  E  G  V  Q  E  V  F  I  L  E  Y  K  E  E  E
```

```
3601  AGAAGAAACCTCTACAGAAGAAAGCGATGATGAATCATCTACTTCTGAAGACTCAGACTC   3660
                                  M  M  N  H  L  L  L  K  T  Q  T  Q
       E  E  T  S  T  E  E  S  D  D  E  S  S  T  S  E  D  S  D  S

3661  AGACTGAGCAGGTGATGAACGTCACCAATCCCAATTCGATCTACATCAAGGGCAGACTCT   3720
        T  E  Q  V  M  N  V  T  N  P  N  S  I  Y  I  K  G  R  L  Y
       D  *

3721  ACTTCAAGGGATACAAGAAGATAGAGCTTCACTGTTTTGTAGACACGGGAGCAAGCTTAT   3780
        F  K  G  Y  K  K  I  E  L  H  C  F  V  D  T  G  A  S  L  C

3781  GCATAGCATCCAAGTTCGTCATTCCAGAAGAACATTGGGTCAATGCAGAAAGACCAATAA   3840
        I  A  S  K  F  V  I  P  E  E  H  W  V  N  A  E  R  P  I  M

3841  TGGTCAAAATAGCAGATGGAAGCTCAATCACCATCAGCAAAGTCTGCAAAGACATAGACT   3900
        V  K  I  A  D  G  S  S  I  T  I  S  K  V  C  K  D  I  D  L

3901  TGATCATAGTCGGCGTGATATTCAAAATTCCCACCGTCTATCAGCAAGAAAGTGGCATCG   3960
        I  I  V  G  V  I  F  K  I  P  T  V  Y  Q  Q  E  S  G  I  D

3961  ATTTCATAATCGGCAACAACTTCTGTCAGCTATATGAACCATTCATACAGTTTACGGATA   4020
        F  I  I  G  N  N  F  C  Q  L  Y  E  P  F  I  Q  F  T  D  R

4021  GAGTTATCTTCACAAAGAACAAGTCTTATCCTGTTCATATTGCGAAGCTAACCAGAGCAG   4080
        V  I  F  T  K  N  K  S  Y  P  V  H  I  A  K  L  T  R  A  V

4081  TGCGAGTAGGCACCGAAGGATTTCTTGAATCAATGAAGAAACGTTCAAAGACTCAACAAC   4140
        R  V  G  T  E  G  F  L  E  S  M  K  K  R  S  K  T  Q  Q  P

4141  CTGAGCCGGTGAACATTTCGACAAACAAGATAGAAAATCCGCTAGAAGAAATTGCTATTC   4200
        E  P  V  N  I  S  T  N  K  I  E  N  P  L  E  E  I  A  I  L

4201  TTTCAGAGGGGAGGAGGTTATCAGAAGAAAAACTCTTCATCACTCAACAAAGAATGCAAA   4260
        S  E  G  R  R  L  S  E  E  K  L  F  I  T  Q  Q  R  M  Q  K
```

```
4261  AAACCGAAGAACTACTTGAGAAAGTATGTTCAGAAAATCCATTAGATCCTAACAAGACTA  4320
        T   E   E   L   L   E   K   V   C   S   E   N   P   L   D   P   N   K   T   K

4321  AGCAATGGATGAAAGCTTCAATCAAGCTCAGCGACCCAAGCAAAGCTATCAAGGTTAAAC  4380
        Q   W   M   K   A   S   I   K   L   S   D   P   S   K   A   I   K   V   K   P

4381  CCATGAAGTATAGCCCAATGGATCGTGAAGAATTTGACAAGCAAATCAAAGAGTTACTGG  4440
        M   K   Y   S   P   M   D   R   E   E   F   D   K   Q   I   K   E   L   L   D

4441  ACCTTAAAGTCATTAAGCCCAGTAAAAGCCCTCACATGGCACCAGCCTTCTTGGTCAACA  4500
        L   K   V   I   K   P   S   K   S   P   H   M   A   P   A   F   L   V   N   N

4501  ATGAAGCCGAGAACGGAAGAGGAAACAAACGTATGGTAGTGAACTACAAAGCTATGAATA  4560
        E   A   E   N   G   R   G   N   K   R   M   V   V   N   Y   K   A   M   N   K

4561  AAGCCACCGTAGGAGACGCATACAATCTTCCCAACAAAGACGAGTTACTTACACTCATTC  4620
        A   T   V   G   D   A   Y   N   L   P   N   K   D   E   L   L   T   L   I   R

4621  GAGGAAAGAAGATCTTTTCTTCCTTCGACTGTAAGTCAGGATTCTGGCAAGTTCTGCTTG  4680
        G   K   K   I   F   S   S   F   D   C   K   S   G   F   W   Q   V   L   L   D

4681  ATCAAGAATCAAGACCTCTAACGGCGTTCACATGTCCACAAGGTCACTACGAATGGAATG  4740
        Q   E   S   R   P   L   T   A   F   T   C   P   Q   G   H   Y   E   W   N   V

4741  TGGTCCCTTTCGGCCTAAAGCAGGCACCATCCATATTCCAGAGACACATGGACGAAGCAT  4800
        V   P   F   G   L   K   Q   A   P   S   I   F   Q   R   H   M   D   E   A   F

4801  TTCGTGTGTTCAGAAAGTTCTGTTGCGTTTATGTCGACGACATTGTCGTATTCAGTAACA  4860
        R   V   F   R   K   F   C   C   V   Y   V   D   D   I   V   V   F   S   N   N

4861  ACGAAGAAGATCATCTACTTCACGTAGCAATGATCTTACAAAAGTGCAATCAGCATGGAA  4920
        E   E   D   H   L   L   H   V   A   M   I   L   Q   K   C   N   Q   H   G   I

4921  TTATCCTTTCCAAGAAGAAAGCACAACTCTTCAAGAAGAAGATAAACTTCCTTGGTCTAG  4980
        I   L   S   K   K   K   A   Q   L   F   K   K   K   I   N   F   L   G   L   E
```

```
4981  AAATAGATGAAGGAACACATAAGCCTCAAGGACATATTTTGGAACATATCAACAAGTTCC  5040
        I  D  E  G  T  H  K  P  Q  G  H  I  L  E  H  I  N  K  F  P

5041  CAGATACCCTTGAAGACAAGAAGCAACTTCAGAGATTCTTAGGCATCCTAACATATGCCT  5100
        D  T  L  E  D  K  K  Q  L  Q  R  F  L  G  I  L  T  Y  A  S

5101  CTGATTATATCCCGAATCTAGCTCAAATGAGACAGCCTCTGCAAGCCAAGCTTAAAGAAA  5160
        D  Y  I  P  N  L  A  Q  M  R  Q  P  L  Q  A  K  L  K  E  N

5161  ATGTTCCATGGAAATGGACAAAAGAGGACACCCTCTACATGCAAAAGGTGAAGAAAAATC  5220
        V  P  W  K  W  T  K  E  D  T  L  Y  M  Q  K  V  K  K  N  L

5221  TGCAAGGATTTCCTCCACTACATCATCCCTTACCAGAAGAGAAGCTGATCATCGAAACCG  5280
        Q  G  F  P  P  L  H  H  P  L  P  E  E  K  L  I  I  E  T  D

5281  ATGCATCAGACGACTACTGGGGAGGTATGTTAAAAGCTATCAAAATTAACGAAGGTACTA  5340
        A  S  D  D  Y  W  G  G  M  L  K  A  I  K  I  N  E  G  T  N

5341  ATACTGAGTTAATTTGCAGATACCGATCTGGAAGCTTTAAGGCTGCAGAAAGGAATTACC  5400
        T  E  L  I  C  R  Y  R  S  G  S  F  K  A  A  E  R  N  Y  H

5401  ACAGCAATGACAAAGAGACATTGGCGGTAATAAATACTATAAAGAAATTCAGTATTTATC  5460
        S  N  D  K  E  T  L  A  V  I  N  T  I  K  K  F  S  I  Y  L

5461  TAACTCCTGTTCATTTTCTGATCAGGACAGATAATACTCATTTCAAGAGTTTTGTTAATC  5520
        T  P  V  H  F  L  I  R  T  D  N  T  H  F  K  S  F  V  N  L

5521  TCAATTACAAAGGTGATTCAAAACTTGGAAGAAACATCAGATGGCAAGCATGGCTTAGCC  5580
        N  Y  K  G  D  S  K  L  G  R  N  I  R  W  Q  A  W  L  S  H

5581  ACTATTCATTTGATGTTGAACATATTAAAGGAACCGACAACCACTTTGCGGACTTCCTTT  5640
        Y  S  F  D  V  E  H  I  K  G  T  D  N  H  F  A  D  F  L  S

5641  CAAGAGAATTCAATAAGGTTAATTCCTAATTGAAATCCGAAGATAAGATTCCCACACACT  5700
        R  E  F  N  K  V  N  S  *
```

```
5701  TGTGGCTGATATCAAAAGGCTACTGCCTATATAAACACATCTCTGGAGACTGAGAAAATC  5760

5761  AGACCTCCAAGCATGGAGAACATAGAAAAACTCCTCATGCAAGAGAAAATACTAATGCTA  5820
                    M  E  N  I  E  K  L  L  M  Q  E  K  I  L  M  L

5821  GAGCTCGATCTAGTAAGAGCAAAAATAAGCTTAGCAAGAGCTAACGGCTCTTCGCAACAA  5880
      E  L  D  L  V  R  A  K  I  S  L  A  R  A  N  G  S  S  Q  Q

5881  GGAGACCTCCCTCTCCACCGTGAAACACCGGAAAAAGAAGAAGCAGTTCATTCTGCACTG  5940
      G  D  L  P  L  H  R  E  T  P  E  K  E  E  A  V  H  S  A  L

5941  GCCACTTTTACGCCAACTCAAGTAAAAGCTATTCCAGAGCAAACGGCTCCTGGTAAAGAA  6000
      A  T  F  T  P  T  Q  V  K  A  I  P  E  Q  T  A  P  G  K  E

6001  TCAACAAATCCGTTGATGGCTAGTATCTTGCCAAAAGATATGAACCCAGTTCAAACTGGG  6060
      S  T  N  P  L  M  A  S  I  L  P  K  D  M  N  P  V  Q  T  G

6061  ATAAGGCTTGCAGTGCCAGGGGACTTTTTACGTCCTCATCAGGGAATTCCAATCCCACAA  6120
      I  R  L  A  V  P  G  D  F  L  R  P  H  Q  G  I  P  I  P  Q

6121  AAATCTGAGCTTAGCAGCACAGTTGCTCCTCTCAGAGCAGAATCGGGTATTCAACACCCT  6180
      K  S  E  L  S  S  T  V  A  P  L  R  A  E  S  G  I  Q  H  P

6181  CATATCAACTACTACGTTGTGTATAACGGTCCACACGCCGGTATATACGATGACTGGGGT  6240
      H  I  N  Y  Y  V  V  Y  N  G  P  H  A  G  I  Y  D  D  W  G

6241  TGTACAAAGGCGGCAACAAACGGCGTTCCCGGAGTTGCACACAAGAAGTTTGCCACTATT  6300
      C  T  K  A  A  T  N  G  V  P  G  V  A  H  K  K  F  A  T  I

6301  ACAGAGGCAAGAGCAGCAGCTGACGCGTACACAACAAGTACGCAAACAGACAGGTTGAAC  6360
      T  E  A  R  A  A  A  D  A  Y  T  T  S  T  Q  T  D  R  L  N

6361  TTCATCCCCAAAGGAGAAGCTCAACTCAAGCCCAAGAGCTTTGCAGAGGCCTTAACCAGC  6420
      F  I  P  K  G  E  A  Q  L  K  P  K  S  F  A  E  A  L  T  S
```

```
6421  CCACCAAAGCAAAAAGCCCACTGGCTCACGCTAGGAACCAAAAGGCCCAGCAGTGATCCA  6480
      P  P  K  Q  K  A  H  W  L  T  L  G  T  K  R  P  S  S  D  P

6481  GCCCCAAAAGAGATCTCCTTTGCCCCGGAGATCACCATGGACGATTTCCTCTATCTCTAC  6540
      A  P  K  E  I  S  F  A  P  E  I  T  M  D  D  F  L  Y  L  Y

6541  CATCTAGGAAGAAAGTTCGACGGAGAAGGTGACGATACCATCTTCACCACTGATAATGAG  6600
      H  L  G  R  K  F  D  G  E  G  D  D  T  I  F  T  T  D  N  E

6601  AAGATTAGCCTCTTCAATTTCAGAAAGAATGCTGACCCACAGATGGTTAGAGAGGCCTAC  6660
      K  I  S  L  F  N  F  R  K  N  A  D  P  Q  M  V  R  E  A  Y

6661  GCAGCAGGTCTCATCAAGACGATCTACCCGAGTAATAATCTCCAGGAGATCAAATACCTT  6720
      A  A  G  L  I  K  T  I  Y  P  S  N  N  L  Q  E  I  K  Y  L

6721  CCCAAGAAGGTTAAAGATGCAGTAAAAGCATTAGGACCTAACTGCATCAAGAACACAGAG  6780
      P  K  K  V  K  D  A  V  K  A  L  G  P  N  C  I  K  N  T  E

6781  AAAGATATATTTCTCAAGATCAGAAGTCATATCCCAGTATGCACGATTCAAGGCCTCGTT  6840
      K  D  I  F  L  K  I  R  S  H  I  P  V  C  T  I  Q  G  L  V

6841  CATAAACCAAGGCAAGTAATAGAGATTGGAGTCTCTAAGAAAGTAGTTCCTACTGAATCA  6900
      H  K  P  R  Q  V  I  E  I  G  V  S  K  K  V  V  P  T  E  S

6901  AAGGCCATGGAGTCAAAAATTCAGATCGAGGATCTAACAGAACTCGCCGTGAAGACTGGC  6960
      K  A  M  E  S  K  I  Q  I  E  D  L  T  E  L  A  V  K  T  G

6961  GGACAGTTCATACAGAGTCTTTTACGACTCAATGACAAGAAGAAAATCTTCGTCAACATG  7020
      G  Q  F  I  Q  S  L  L  R  L  N  D  K  K  K  I  F  V  N  M

7021  GTGGAGCACGACACTCTCGTCTACTCCAAGAATATCAAAGATACAGTCTCAGAAGACCAA  7080
      V  E  H  D  T  L  V  Y  S  K  N  I  K  D  T  V  S  E  D  Q

7081  AGGGCTATTGAGACTTTTCAACAAAGGGTAATATCAGGAAACCTCCTCGGATTCCATTGC  7140
      R  A  I  E  T  F  Q  Q  R  V  I  S  G  N  L  L  G  F  H  C
```

```
7141  CCATCTATCTGTCACTTCATGGAAAGGACAGTAGAAAAGGAAGGTGGCTCCTACAAAGTC  7200
      P  S  I  C  H  F  M  E  R  T  V  E  K  E  G  G  S  Y  K  V

7201  CATCATTGCGATAAAGGAAAGGCTATCGTTCAAGATGCCTCTGCCGACAGTGGTCCCAAA  7260
      H  H  C  D  K  G  K  A  I  V  Q  D  A  S  A  D  S  G  P  K

7261  GATGGACCCCCACCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCAACCACGTCTTCA  7320
      D  G  P  P  P  T  R  S  I  V  E  K  E  D  V  P  T  T  S  S

7321  AAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGACGCACAATCCCACTAT  7380
      K  Q  V  D  *

7381  CCTTCGCAAGACTCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGACACGCTGA  7440

7441  AATCACCAGTCTCTCTCTACAAATCTATCTCTCTCTATTTTCTCCATAATAATGTGTGAG  7500

7501  TAGTTCCCAGATAAGGGAATTAGGATTCTTATAGGGTTTCGCTGATGTGTTGAGCATATA  7560

7561  AGAAACCCTTAGTATGTATTAGTATTAGTAAGATACTTCTATCAATAAAATTTCTAATTC  7620

7621  CTAAAACCAAAATCCAGTACTAAAATCCAGATCTCCTAAAGTCCCTATAGATCTATGTCG  7680

7681  AGAATATAAACCAGACACGAGACGACTAAACCTGGAGCCCAGACGCCGATTGAAGCTAGA  7740

7741  AGTACCGCTTAGGCAGGAGGCCGTTAGGGAAAAGATGCTAAGGCAGGGTTGGTTACGTTG  7800

7801  ACTCCCCCGTAGGTTTGGTTTAAATATGATGAAGTGGACGGAAGGAAGGAGGAAGACAAG  7860

7861  GAAGGATAAGGTTGCAGGCCCTGTGCAAGGTAAGAAGATGGAAATTTGATAGAGGTACGT  7920

7921  TACTATACCTATACTATAAGCTAAGGGAATGCTTGTATTTACCCTATATACCCTAATAAC  7980

7981  CCCTTATCGATTTAAAGAAATAATCCGCATAAGCCCCCGCTTAAAAAATT  8030
```

the resulting plasmid (pCMS31) was used for nucleotide sequencing (86). These results confirm and extend earlier work (105) which showed that NY8153 is a unique CaMV isolate. The ORFs in NY8153 correspond in length and genomic position to those of other sequenced isolates.

The Complete Nucleotide Sequence of Cauliflower Mosaic Virus Isolate BBC

Cauliflower mosaic virus (CaMV) is the type member of the caulimovirus group of plant viruses. Caulimovirus members have a double-stranded DNA genome of about 8 kbp. Caulimoviruses are classified as pararetroviruses (12) because they replicate via an RNA intermediate using a viral encoded reverse transcriptase. Transcription of the CaMV genome produces two major transcripts: the 19S and 35S RNAs. Six major open reading frames (ORFs) can be found tightly packed in the CaMV genome. The functions of five of these ORFs are known. Details of CaMV molecular biology have been reviewed (11). CaMV mainly infects members of the crucifereae and solanaceae. DNA was isolated from the BBC isolate of CaMV from infected Pak Choi plants obtained in 1988 from California (Melcher, unpublished, 1988). Symptoms induced by the BBC isolate on turnip include necrotic flecks, chlorotic mottle, mosaic, mid-rib curling, and pale green leaves. The cloned BBC genome was completely sequenced using the di-deoxy chain-termination method. The complete nucleotide sequence of the BBC isolate is shown in Figure 2.

Figure 2. The complete nucleotide sequence of CaMV isolate BBC. The derived amino acid sequences of the six major CaMV ORFs are shown in one letter code below the nucleotide sequence. This figure spans pages 32-42.

```
   1  GGTATCAGAGCCATGAATCGGTTTAAAGACCAAATTCAAGAGGGTAAAACCTCACCAATA   60

  61  AACAAAAGAGTTCTTAACTCTAATGATAAAAGATCTTTCAAGATCAACAATAGTTCCCTC  120

 121  ACACCGGTGACCGACAGGTTTACGACCGTAAGGTTTCAGAACAACATCGAAAGCGTTTAC  180

 181  GCCAACTTCGACTCTCGACTAAAGTCGTCGTACGATGGTAGATCTAAAAAGATCAAGAAT  240

 241  CTAAGCCTTAAAAATCTTAGATGTTACGAAGCCTTCCTCAGGAAGTACCTTCTGGAACAA  300

 301  TAAATCTCTCTGAGAATAGTACTCTATTGAGTATCCACAGAAAAAATAATCTTCTGTGTT  360

 361  GAGATGGATTTGTATCCAGAAGAAAACACCCAAAGCGAGCAATCGCATAATTCTGAAAAT  420
          M  D  L  Y  P  E  E  N  T  Q  S  E  Q  S  H  N  S  E  N

 421  AATATGCAAATATTTAAGTCAGAAAATTCGGATGGATTCTCCTCCGATCTAATGATCTCA  480
      N  M  Q  I  F  K  S  E  N  S  D  G  F  S  S  D  L  M  I  S

 481  AACGATCAATTAAAAAATATCTCAAAAACCCAATTAACTTTGGAAAAAGAAAAGATATTT  540
      N  D  Q  L  K  N  I  S  K  T  Q  L  T  L  E  K  E  K  I  F

 541  AAAATGCCTAACGTTTTGTCTCAAGTTATGAAAAGAGCGTTTAGCAGGAAAAACGAGATT  600
      K  M  P  N  V  L  S  Q  V  M  K  R  A  F  S  R  K  N  E  I

 601  CTTTACTGCGTCTCGACAAAAGAGTTATCAGTGGACATTCACGATGCCACAGGTAAGGTA  660
      L  Y  C  V  S  T  K  E  L  S  V  D  I  H  D  A  T  G  K  V

 661  TATCTTCCCTTAATCACTAGAGAGGAGATAAATAAAAGACTTTCAAGCTTAAAACCTGAA  720
      Y  L  P  L  I  T  R  E  E  I  N  K  R  L  S  S  L  K  P  E

 721  GTCAGAAAGACCATGTCCATGGTTCATCTTGGAGCGGTCAAAATATTGCTTAAAGCTCAA  780
      V  R  K  T  M  S  M  V  H  L  G  A  V  K  I  L  L  K  A  Q
```

```
781   TTTCGAAATGGGATTGATACCCCAATCAAAATTGCTTTAATCGATGATAGAATTAATTCT   840
      F  R  N  G  I  D  T  P  I  K  I  A  L  I  D  D  R  I  N  S

841   AGAAGAGATTGCCTTCTCGGTGCAGCCAAAGGTAATCTAGCATACGGTAAGTTTATGTTT   900
      R  R  D  C  L  L  G  A  A  K  G  N  L  A  Y  G  K  F  M  F

901   ACTGTATACCCCAAGTTTGGAATAAGCCTTAATACCCAAAGACTTAACCAAACCCTAAGC   960
      T  V  Y  P  K  F  G  I  S  L  N  T  Q  R  L  N  Q  T  L  S

961   CTTATTCATGATTTTGAAAATAAAAATCTTATGAATAAAGGTGATAAAGTTATGACCATA   1020
      L  I  H  D  F  E  N  K  N  L  M  N  K  G  D  K  V  M  T  I

1021  ACCTATATGGTAGGATATGCATTAACTAATAGTCATCATAGCATAGATTATCAATCGAAT   1080
      T  Y  M  V  G  Y  A  L  T  N  S  H  H  S  I  D  Y  Q  S  N

1081  GCTACAATTGAACTAGAAGACGTATTTCAAGAAATTGGAAATGTCCACGAGTCTGATTTT   1140
      A  T  I  E  L  E  D  V  F  Q  E  I  G  N  V  H  E  S  D  F

1141  TGTACAATACAAAATGACGAATGCAATTGGGCCATTGATATAGCCCAAAACAAAGCCTTA   1200
      C  T  I  Q  N  D  E  C  N  W  A  I  D  I  A  Q  N  K  A  L

1201  TTAGGAGCTAAAACCAAATCCCAAATTGGTAATAATCTTCAAATAGGAAACAGTGCTTCA   1260
      L  G  A  K  T  K  S  Q  I  G  N  N  L  Q  I  G  N  S  A  S

1261  TCCTCTAATACTGAAAATGAATTAGCTAGGGTAAGCCAGAACATAGATCTTTTAAAGAAT   1320
      S  S  N  T  E  N  E  L  A  R  V  S  Q  N  I  D  L  L  K  N

1321  AAATTAAAAGAAATCTGTGGAGAATAAAATGAGGATTACGGGTCAACCGCATGTTTATAA   1380
      K  L  K  E  I  C  G  E  *  M  R  I  T  G  Q  P  H  V  Y  K

1381  AAAAGATACTATTATTAGACTAAAACCATTGTCTCTTAATAGTAATAATAGAAGTTATGT   1440
       K  D  T  I  I  R  L  K  P  L  S  L  N  S  N  N  R  S  Y  V

1441  TTTTAGTTCCTCAAAAGGGAACATTCAAAATATAATTAATCATCTTAACAACCTCAATGA   1500
       F  S  S  S  K  G  N  I  Q  N  I  I  N  H  L  N  N  L  N  E
```

```
1501  GATTGTAGGAAGAAGCTTACTCGGAATATGGAAGATCAATTCATACTTCGGCTTAAGCAA  1560
        I  V  G  R  S  L  L  G  I  W  K  I  N  S  Y  F  G  L  S  K

1561  AGACCCTTCGGAGTCCAAATCAAAAAACCCGTCAGTTTTTAATACTGCAAAAACCATTTT  1620
        D  P  S  E  S  K  S  K  N  P  S  V  F  N  T  A  K  T  I  F

1621  TAAGAGTGGGGGGGTTGATTACTCGAGCCAACTAAAAGAAATAAAATCTCTTTTAGAAGC  1680
        K  S  G  G  V  D  Y  S  S  Q  L  K  E  I  K  S  L  L  E  A

1681  TCAAAATACTAGAATTAAAAATCTAGAAAAAGCAATTCAATCCTTAGATAATAAGATTGA  1740
        Q  N  T  R  I  K  N  L  E  K  A  I  Q  S  L  D  N  K  I  E

1741  ACCAGAGCCCTTAACTAAAAAAGAAGTTAAAGAGCTAAAAGAATCGATTAACTCGATCAA  1800
        P  E  P  L  T  K  K  E  V  K  E  L  K  E  S  I  N  S  I  K

1801  AGAAGGATTAAAGAATATTATTGGCTGAAATGGCTAATCTTAATCAAATCCAAAAAGAAG  1860
        E  G  L  K  N  I  I  G  *  M  A  N  L  N  Q  I  Q  K  E  V

1861  TCTCTGAAATCCTCAGTGACCAAAAATCCATGAAATCGGATATAAAAGCTATCTTAGAAT  1920
          S  E  I  L  S  D  Q  K  S  M  K  S  D  I  K  A  I  L  E  L

1921  TACTAGGATCCCAAAATCCTACTAAAGAAAGCTTAGAAGCCGTTGCAGCGAAAATCGTTA  1980
          L  G  S  Q  N  P  T  K  E  S  L  E  A  V  A  A  K  I  V  N

1981  ATGACTTAACCAAGCTCATCAATGATTGTCCTTGTAACAAAGAGATATTAGAAGCCTTAG  2040
          D  L  T  K  L  I  N  D  C  P  C  N  K  E  I  L  E  A  L  G

2041  GTAATCAACCTAAAGAGCAACTAATAGAACAACCTAAAGAAAAAGGCAAAGGCCTTAATC  2100
          N  Q  P  K  E  Q  L  I  E  Q  P  K  E  K  G  K  G  L  N  L

2101  TAGGAAAATATTCTTACCCTAATTACGGTGTAGGAAATGAAGAATTAGGATCCTCTGGAA  2160
          G  K  Y  S  Y  P  N  Y  G  V  G  N  E  E  L  G  S  S  G  N

2161  ACCCTAAAGCTTTAACTTGGCCCTTCAAAGCTCCAGCAGGATGGCCGAATCAATTTTAGA  2220
          P  K  A  L  T  W  P  F  K  A  P  A  G  W  P  N  Q  F  *
                                                    M  A  E  S  I  L  D
```

```
2221  CAGGACCATTAACCGGTTCTGGTATAATCTGGGAGAAGATTGTCTCTCGGAAAGTCAATT  2280
       R  T  I  N  R  F  W  Y  N  L  G  E  D  C  L  S  E  S  Q  F

2281  TGACCTTATGATAAGGTTAATGGAAGAGTCCCTTGACGGGGACCAAATTATTGATCTAAC  2340
       D  L  M  I  R  L  M  E  E  S  L  D  G  D  Q  I  I  D  L  T

2341  CTCTCTACCTAGTGATAATTTGCAGGTCGAACAGGTTATGACAACTACCGACGACTCGAT  2400
       S  L  P  S  D  N  L  Q  V  E  Q  V  M  T  T  T  D  D  S  I

2401  CTCGGAAGAATCAGAATTCCTTCTAGCAATAGGAGAAACATCTGAAGACGAAAGCGATTC  2460
       S  E  E  S  E  F  L  L  A  I  G  E  T  S  E  D  E  S  D  S

2461  AGGAGAAGAACCTGAATTCGAACAAGTTCGAATGGATCGAACAGGAGGAACGGAGATTCC  2520
       G  E  E  P  E  F  E  Q  V  R  M  D  R  T  G  G  T  E  I  P

2521  CAAAAAAGAAGATGGTGCAGAACCATCTAGATATAATGAGAGAAAGAGAAAGACCACGGA  2580
       K  K  E  D  G  A  E  P  S  R  Y  N  E  R  K  R  K  T  T  E

2581  GGACCGGTACTTTCCAACTCAACCAAAGACCATTCCAGGACAAAAACAAACGTCTATGGG  2640
       D  R  Y  F  P  T  Q  P  K  T  I  P  G  Q  K  Q  T  S  M  G

2641  AATACTCAACATTGACTGCCAAACCAATCGAAGAACCTTAATCGATGACTGGGCAGCAGA  2700
       I  L  N  I  D  C  Q  T  N  R  R  T  L  I  D  D  W  A  A  E

2701  AATCGGATTGATAGTCAAAACCAACAGAGAAGACTATCTTGATCCAGAAACAATACTACT  2760
       I  G  L  I  V  K  T  N  R  E  D  Y  L  D  P  E  T  I  L  L

2761  CCTGATGGAACACAAAACATCAGGAATAGCCAAGGAGTTAATCCGAAATACAAGATGGAA  2820
       L  M  E  H  K  T  S  G  I  A  K  E  L  I  R  N  T  R  W  N

2821  CCGCACTACCGGAGATATCATAGAACAGGTGATCGATGCGATGTACACCATGTTCTTAGG  2880
       R  T  T  G  D  I  I  E  Q  V  I  D  A  M  Y  T  M  F  L  G

2881  ACTAAACTACTCCGACAACAAGGTTGCTGAAAAGATAGACGAGCAAGAGAAGGCCAAGAT  2940
       L  N  Y  S  D  N  K  V  A  E  K  I  D  E  Q  E  K  A  K  I
```

```
2941  CAGAATGACCAAGCTCCAGCTCTGCGACATCTGCTACCTTGAAGAATTTACATGTGATTA  3000
        R  M  T  K  L  Q  L  C  D  I  C  Y  L  E  E  F  T  C  D  Y

3001  TGAGAAGAACATGTACAAAACGGAACTGGCGGATTTCCCAGGATATATCAACCAGTACCT  3060
        E  K  N  M  Y  K  T  E  L  A  D  F  P  G  Y  I  N  Q  Y  L

3061  GTCAAAAATCCCCATCATTGGAGAAAAAGCGCTAACACGCTTTAGGCATGAAGCTAACGG  3120
        S  K  I  P  I  I  G  E  K  A  L  T  R  F  R  H  E  A  N  G

3121  AACCAGCATCTACAGCTTAGGTTTCGCGGCAAAGATAGTAAAAGAAGAACTATCTAAAAT  3180
        T  S  I  Y  S  L  G  F  A  A  K  I  V  K  E  E  L  S  K  I

3181  CTGCGCATTATCCAAGAAGCAGAAGAAGTTGAAGAAATTCAACAAGAAATGCTGCAGCAT  3240
        C  A  L  S  K  K  Q  K  K  L  K  K  F  N  K  K  C  C  S  I

3241  CGGCGAAGCTTCAGTAGAATATGGATGCAAGAAAACATCCAAGAAGAAGTATCATAATAA  3300
        G  E  A  S  V  E  Y  G  C  K  K  T  S  K  K  K  Y  H  N  K

3301  GCGATACAAGAAAAAATATAAGGTCTATAAACCTTATAAGAAGAAGAAGAAATTCCGATC  3360
        R  Y  K  K  K  Y  K  V  Y  K  P  Y  K  K  K  K  K  F  R  S

3361  CGGAAAATACTTCAAGCCCAAGGAGAAGAAGGGCTCAAAGCAAAAGTATTGCCCAAAAGG  3420
        G  K  Y  F  K  P  K  E  K  K  G  S  K  Q  K  Y  C  P  K  G

3421  CAAGAAAGACTGCAGATGTTGGATCTCGAACATTGAAGGCCATTACGCCAACGAATGTCC  3480
        K  K  D  C  R  C  W  I  S  N  I  E  G  H  Y  A  N  E  C  P

3481  TAATCGACAAAGCTCGGAGAAGGCTCACATCCTTCAACAAGCAGAGAAATTGGGTCTCCA  3540
        N  R  Q  S  S  E  K  A  H  I  L  Q  Q  A  E  K  L  G  L  Q

3541  GCCCATTGAAGAACCCTATGAAGGAGTTCAAGAAGTATTCATCTTAGAATACAAAGAAGA  3600
        P  I  E  E  P  Y  E  G  V  Q  E  V  F  I  L  E  Y  K  E  E
```

```
3601  GGAAGAAGAAACCTCTACAGAAGAAAGTGATGGATCATCTACTTCTGAAGACTCAGACTC   3660
                                        M  D  H  L  L  L  K  T  Q  T  Q
      E  E  E  T  S  T  E  E  S  D  G  S  S  T  S  E  D  S  D  S

3661  AGACTGAGCAGGTGATGAACGTCACCAATCCCAATTCGATTTACATCAAGGGAAGACTCT   3720
        T  E  Q  V  M  N  V  T  N  P  N  S  I  Y  I  K  G  R  L  Y
      D  *

3721  ACTTCAAGGGATACAAGAAGATAGAGCTTCACTGTTTTGTAGACACGGGAGCAAGCTTAT   3780
        F  K  G  Y  K  K  I  E  L  H  C  F  V  D  T  G  A  S  L  C

3781  GCATAGCATCCAAGTTCGTCATTCCAGAAGAACATTGGGTCAATGCAGAAAGACCAATAA   3840
        I  A  S  K  F  V  I  P  E  E  H  W  V  N  A  E  R  P  I  M

3841  TGGTCAAAATAGCAGATGGAAGTTCAATCACCATCAGCAAAGTCTGCAAAGACATAGACT   3900
        V  K  I  A  D  G  S  S  I  T  I  S  K  V  C  K  D  I  D  L

3901  TGATCATAGCGCGCGAGATATTCAAAATTCCCACCGTCTATCAGCAAGAAAGTGGCATCG   3960
        I  I  A  R  E  I  F  K  I  P  T  V  Y  Q  Q  E  S  G  I  D

3961  ATTTCATAATCGGCAACAACTTCTGTCAGCTATATGAACCATTCATACAGTTTACGGACA   4020
        F  I  I  G  N  N  F  C  Q  L  Y  E  P  F  I  Q  F  T  D  R

4021  GAGTTATCTTCACAAAGAACAAGTCTTATCCTGTTCATATTGCGAAGCTAACAAGAGCAG   4080
        V  I  F  T  K  N  K  S  Y  P  V  H  I  A  K  L  T  R  A  V

4081  TGCGAGTAGGCACCGAAGGATTTCTTGAATCAATGAAGAAACGTTCAAAGACTCAACAAC   4140
        R  V  G  T  E  G  F  L  E  S  M  K  K  R  S  K  T  Q  Q  P

4141  CTGAGCCGGTGAACATTTCGACAAACAAGATAGAAAATCCACTAAAAGAAATTGCTATTC   4200
        E  P  V  N  I  S  T  N  K  I  E  N  P  L  K  E  I  A  I  L

4201  TTTCAGAGGGGAGGAGGTTATCAGAAGAAAAACTCTTCATCACTCAACAAAGAATGCAAA   4260
        S  E  G  R  R  L  S  E  E  K  L  F  I  T  Q  Q  R  M  Q  K
```

```
4261  AAATCGAAGAACTACTTGAGAAAGTATGTTCAGAAAATCCATTAGATCCTAACAAGACTA  4320
        I   E   E   L   L   E   K   V   C   S   E   N   P   L   D   P   N   K   T   K

4321  AGCAATGGATGAAAGCTTCAATCAAGCTCAGCGACCCAAGCAAAGCTATCAAGGTTAAAC  4380
        Q   W   M   K   A   S   I   K   L   S   D   P   S   K   A   I   K   V   K   P

4381  CCATGAAGTATAGCCCAATGGATCGTGAAGAATTTGACAAGCAAATCAAAGAGTTACTGG  4440
        M   K   Y   S   P   M   D   R   E   E   F   D   K   Q   I   K   E   L   L   D

4441  ACCTTAAAGTCATTAAGCCCAGTAAAAGCCCTCACATGGCACCAGCCTTCTTGGTCAACA  4500
        L   K   V   I   K   P   S   K   S   P   H   M   A   P   A   F   L   V   N   N

4501  ATGAAGCCGAGAAGCGAAGAGGAAAGAAGCGTATGGTAGTTAACTACAAGGCTATGAACA  4560
        E   A   E   K   R   R   G   K   K   R   M   V   V   N   Y   K   A   M   N   K

4561  AAGCCACCATAGGAGACGCATACAATCTTCCCAATAAAGACGAGTTACTGACACTTATTC  4620
        A   T   I   G   D   A   Y   N   L   P   N   K   D   E   L   L   T   L   I   R

4621  GAGGAAAGAAGATCTTCTCTTCCTTCGACTGCAAGTCAGGATTCTGGCAGGTTCTGCTAG  4680
        G   K   K   I   F   S   S   F   D   C   K   S   G   F   W   Q   V   L   L   D

4681  ATCAAGAATCAAGACCTCTAACGGCATTCACATGTCCCCAAGGTCACTACGAATGGAATG  4740
        Q   E   S   R   P   L   T   A   F   T   C   P   Q   G   H   Y   E   W   N   V

4741  TGGTCCCTTTCGGCTTAAAGCAGGCACCATCCATATTCCAAAGACACATGGACGAAGCAT  4800
        V   P   F   G   L   K   Q   A   P   S   I   F   Q   R   H   M   D   E   A   F

4801  TTCGTGTGTTCAGAAAGTTCTGTTGCGTTTATGTCGACGACATTCTCGTATTCAGTAACA  4860
        R   V   F   R   K   F   C   C   V   Y   V   D   D   I   L   V   F   S   N   N

4861  ATGAGGAAGATCACCTACTTCACGTAGCAATGATCTTACAAAAGTGCAATCAACATGGAA  4920
        E   E   D   H   L   L   H   V   A   M   I   L   Q   K   C   N   Q   H   G   I

4921  TCATCCTTTCCAAGAAGAAAGCACAACTCTTCAAAAAGAAGATAAACTTCCTTGGTCTAG  4980
        I   L   S   K   K   K   A   Q   L   F   K   K   K   I   N   F   L   G   L   E
```

```
4981  AAATAGATGAAGGAACACATAAGCCTCAAGGACATATCTTGGAACATATCAACAAATTCC  5040
        I  D  E  G  T  H  K  P  Q  G  H  I  L  E  H  I  N  K  F  P

5041  CAGATACCCTTGAAGACAAGAAGCAACTTCAGAGATTCTTAGGCATCCTAACATATGCCT  5100
        D  T  L  E  D  K  K  Q  L  Q  R  F  L  G  I  L  T  Y  A  S

5101  CCGATTATATCCCGAAGCTAGCTCAAATTAGAAAGCCTCTGCAAGCCAAGCTTAAAGAAA  5160
        D  Y  I  P  K  L  A  Q  I  R  K  P  L  Q  A  K  L  K  E  N

5161  ATGTTCCATGGAAATGGACAAAAGAGGACACCCTCTACATGCAAAAGGTGAAGAAAAATC  5220
        V  P  W  K  W  T  K  E  D  T  L  Y  M  Q  K  V  K  K  N  L

5221  TGCAAGGATTTCCTCCACTACATCATCCCTTACCAGAGGAAAAGCTGATCATCGAGACCG  5280
        Q  G  F  P  P  L  H  H  P  L  P  E  E  K  L  I  I  E  T  D

5281  ACGCATCAGACGACTACTGGGGAGGTATGTTAAAAGCTATCAAAATTAACGAAGGAACTA  5340
        A  S  D  D  Y  W  G  G  M  L  K  A  I  K  I  N  E  G  T  N

5341  ATACTGAGTTAATTTGCAGATACGCATCTGGAAGCTTTAAAGCTGCAGAAAGGAATTACC  5400
        T  E  L  I  C  R  Y  A  S  G  S  F  K  A  A  E  R  N  Y  H

5401  ACAGCAATGACAAAGAGACATTGGCGGTAATAAATACTATAAAGAAATTCAGTATTTATC  5460
        S  N  D  K  E  T  L  A  V  I  N  T  I  K  K  F  S  I  Y  L

5461  TAACTCCTGTTCATTTTCTGATTAGGACAGATAATACTCATTTCAAGAGTTTTGTTAATC  5520
        T  P  V  H  F  L  I  R  T  D  N  T  H  F  K  S  F  V  N  L

5521  TTAATTACAAAGGAGATTCAAAACTTGGAAGAAACATCAGATGGCAAGCATGGCTTAGCC  5580
        N  Y  K  G  D  S  K  L  G  R  N  I  R  W  Q  A  W  L  S  H

5581  ACTATTCGTTTGATGTTGAACATATTAAAGGAACCGACAACCACTTTGCGGACTTCCTTT  5640
        Y  S  F  D  V  E  H  I  K  G  T  D  N  H  F  A  D  F  L  S

5641  CAAGAGAATTCAACAAGGTTAATTCCTAATTGAAATCCGAAGATAAGATTCCCACACACT  5700
        R  E  F  N  K  V  N  S  *
```

```
5701  TGTGGCTGATATCAAAAGGCTACTGCCTATATAAACACATCTCTGGAGACTGAGAAAATC  5760

5761  AGACCTCCAAGCATGGAGAACATAGAAAAACTCCTCATGCAAGAGAAAATACTAATGCTA  5820
                  M  E  N  I  E  K  L  L  M  Q  E  K  I  L  M  L

5821  GAGCTCGATCTAGTAAGAGCAAAAATAAGCTTAGCAAGAGCTAACGGCTCTTCGCAACAA  5880
      E  L  D  L  V  R  A  K  I  S  L  A  R  A  N  G  S  S  Q  Q

5881  GGAGACCTCTCTCTCCACCGTGAAACACCGGTAAAAGAAGAAGCAGTTCATTCTGCACTG  5940
      G  D  L  S  L  H  R  E  T  P  V  K  E  E  A  V  H  S  A  L

5941  GCCACTTTTACGCCAACTCAAGTAAAGGCTATTCCAGAGCAAACGGCTCCTGGTAAAGAA  6000
      A  T  F  T  P  T  Q  V  K  A  I  P  E  Q  T  A  P  G  K  E

6001  TCAACAAATCCGTTGATGGCTAGTATCTTGCCAAAAGATATGAACCCAGTTCAAACTGGG  6060
      S  T  N  P  L  M  A  S  I  L  P  K  D  M  N  P  V  Q  T  G

6061  ATAAGGCTTGCAGTGCCAGGGGACTTTTTACGTCCTCATCAGGGAATTCCAATCCCACAA  6120
      I  R  L  A  V  P  G  D  F  L  R  P  H  Q  G  I  P  I  P  Q

6121  AAATCTGAGCTTAGCAGCACAGTTGTTCCTCTCAGAGACGAATCGGGTATTCAACACCCT  6180
      K  S  E  L  S  S  T  V  V  P  L  R  D  E  S  G  I  Q  H  P

6181  CATATCAACTACTACGTTGTGTATAACGGTCCACACGCCGGTATATACGATGACTGGGGT  6240
      H  I  N  Y  Y  V  V  Y  N  G  P  H  A  G  I  Y  D  D  W  G

6241  TGTACAAAGGCGGCAACAAACGGCGTTCCCGGAGTTGCACACAAGAAGTTTGCCACTATT  6300
      C  T  K  A  A  T  N  G  V  P  G  V  A  H  K  K  F  A  T  I

6301  ACAGAGGCAAGAGCAGCAGCTGACGCGTACACAACAAGTCAGCAAACAGACAGGTTGAAC  6360
      T  E  A  R  A  A  A  D  A  Y  T  T  S  Q  Q  T  D  R  L  N

6361  TTCATCCCCAAAGGAGAAGCTCAACTCAAGCCCAAGAGCTTTCGAGAGGCCTTAACCAGC  6420
      F  I  P  K  G  E  A  Q  L  K  P  K  S  F  R  E  A  L  T  S
```

```
6421  CCACCAAAGCAAAAAGCCCACTGGCTCACGCTAGGAACCAAAAGGCCCAGCAGTGATCCA  6480
      P  P  K  Q  K  A  H  W  L  T  L  G  T  K  R  P  S  S  D  P

6481  GCCCCAAAAGAGATCTCTTTTGCCCCGGAGATCACCATGGACGACTTTCTCTATCTCTAC  6540
      A  P  K  E  I  S  F  A  P  E  I  T  M  D  D  F  L  Y  L  Y

6541  GATCTAGGAAGAAAGTTCGACGGAGAAGGTGACGATACCATGTTCACCACTGATAATGAG  6600
      D  L  G  R  K  F  D  G  E  G  D  D  T  M  F  T  T  D  N  E

6601  AAGATTAGCCTCTTCAATTTCAGAAAGAATGCTGACCCACAGATGGTTAGAGAGGCCTAC  6660
      K  I  S  L  F  N  F  R  K  N  A  D  P  Q  M  V  R  E  A  Y

6661  GCAGCAGGTCTCATCAAGACGATCTACCCGAGTAATAATCTCCAGGAGATCAAATACCTT  6720
      A  A  G  L  I  K  T  I  Y  P  S  N  N  L  Q  E  I  K  Y  L

6721  CCCAAGAAGGTTAAAGATGCAGTCAAAAGATTCAGGACTAACTGCATCAAGAACACAGAG  6780
      P  K  K  V  K  D  A  V  K  R  F  R  T  N  C  I  K  N  T  E

6781  AAAGATATATTTCTCAAGATCAGAAGTACTATCCCAGTATGGACGATTCAAGGCTTGCTT  6840
      K  D  I  F  L  K  I  R  S  T  I  P  V  W  T  I  Q  G  L  L

6841  CATAAACCAAGGCAAGTAATAGAGATTGGAGTCTCTAAGAAAGTAGTTCCTACTGAATCA  6900
      H  K  P  R  Q  V  I  E  I  G  V  S  K  K  V  V  P  T  E  S

6901  AAGGCCATGGAGTCAAAAATTCAGATCGAGGATCTAACAGAACTCGCCGTGAAGACTGGC  6960
      K  A  M  E  S  K  I  Q  I  E  D  L  T  E  L  A  V  K  T  G

6961  GAACAGTTCATACAGAGTCTTCTACGACTCAATGACAAGAAGAAAATCTTCGTCAACATG  7020
      E  Q  F  I  Q  S  L  L  R  L  N  D  K  K  K  I  F  V  N  M

7021  GTGGAAGATGACACTCTCGTCTACTCCAAGAATATCAAAGATACAGTCTCAGAAGACCAA  7080
      V  E  D  D  T  L  V  Y  S  K  N  I  K  D  T  V  S  E  D  Q

7081  AGGGCTATTGAGACTTTTCAACAAAGGGTAATATCAGGAAACCTCCTCGGATTCCATTGC  7140
      R  A  I  E  T  F  Q  Q  R  V  I  S  G  N  L  L  G  F  H  C
```

```
7141  CCAGCTATCTGTCACTTCATCGAAAGGACAGTAGAAAAGGAAGGTGGCTCCTACAAAGTC  7200
      P  A  I  C  H  F  I  E  R  T  V  E  K  E  G  G  S  Y  K  V

7201  CATCATTGCGATAAAGGAAAGGCTATCGTTCAAGATGCCTCTGCCGACAGTGGTCCTAAA  7260
      H  H  C  D  K  G  K  A  I  V  Q  D  A  S  A  D  S  G  P  K

7261  GATGGACCCCCACCCACGAGGAGCATCGTGGAAAAAGAAGACGTTCCAACCACGTCTTCA  7320
      D  G  P  P  P  T  R  S  I  V  E  K  E  D  V  P  T  T  S  S

7321  AAGCAAGTGGATTGATGTGATATCTCCACTGACTGAAGGGATGACGCACAATCCCACTAT  7380
      K  Q  V  D  *

7381  CCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGACACGCTGA  7440

7441  AATCACCAGTCTCTCTCTACAAATCTATCTCTCTCTATTTTCTCCATAATAATGTGTGAG  7500

7501  TAGTTCCCAGATAAGGGAATTAGGGTTCTTATAGGGTTTCGCTCATGTGTTGAGCATATA  7560

7561  AGAAACTCTTAGTATGTATTTGAATTTGTAAAATACTTCTATCAATAAAATTTCTAATTC  7620

7621  CTAAAACCAAAATCCAGTACTAAAAGCCAGATCTCCTAAAGTCCCTATAGATCTTTGTGG  7680

7681  TGAATATAAACCAGACACGAGACGACTAAACCTGGAGCCCAGATGCCGTTTGAAGCTAGA  7740

7741  AGTACCGCTTAGGCAGGAGGCCGTTAGGGAAAAGATGCTAAGGCAGGGTTGGTTACGTTG  7800

7801  ACTCCCCCGTAGGGTTGGTTTAAATATCATGAAGTGGACTGAAGAAAGAAGGAAGACATG  7860

7861  GAAGGATAAGGTTGCAGGCCCTGTGCAAGGTAAGAAGATGGAAATTTGATAGAGGTACGC  7920

7921  TACTATACTTATACTATACGCTAAGAGAATGCTTGTATTTATACCCTATACCCCCTAATA  7980

7981  ACCCCTTATCAATTTAAAGAAATAATCCGCATAAGCCCCCGCTTAAAAAATT  8032
```

Although the nucleotide sequence of the BBC isolate varies in sequence by 5% when compared with isolate Cabbage S, its open reading frames correspond in approximate genomic position and length to those of all known CaMV isolates.

Fonts for the Display of Nucleotide and Amino Acid Sequences:Application to Cauliflower Mosaic Virus

The sequence of amino acid residues in proteins is usually represented by an N-terminal to C-terminal string of three-letter or one-letter abbreviations. Similarly, the sequence of nucleotides in nucleic acids is usually represented by a string of the letters A, G, C, T, and U. The visual appearance of the characters of the Roman alphabet used for these codes bears no relation to the structures or chemical properties of the residues they represent. One-letter abbreviations can, in some fonts, be confused for other characters (eg. G for C, V for Y, and uppercase I for lower case l). Alternate representations of nucleotide (46, 71, 80) and amino acid (2, 80, 81, 97) sequences have been proposed.

Puppy is an informative and space-efficient representation of nucleotide sequences (71). In the Puppy representation, named for purines and pyrimidines, nucleotides are represented by three vertically aligned spaces (Figure 3A: A, T, G, C). An occupied lowest space denotes a pyrimidine, an occupied uppermost space a purine;

occupation of the middle position indicates a guanine or cytosine base. The representation is efficient in its use of space and allows visual recognition of many patterns important to the biological functions of the nucleic acid. We modified Puppy to allow depiction of ambiguous bases. In this version, characters are composed of open circles rather than filled squares. Ambiguous residues have been encoded with three characters: one for any of four or more possible bases (Figure 3A: N); a second to represent three possible bases (Figure 3A: B, D, H, V); and the third to represent two possible bases (Figure 3A: R, Y, K, M, S, W).

To accompany Puppy, we devised Kitty (109), a representation of amino acid sequences of proteins that suggests the chemical structures and properties of the individual residues (Figure 3B). As with Puppy, the symbols for each amino acid are made up of one or more circles. The arrangement of circles for each residue type closely approximates the number and connectivity of carbon, oxygen, nitrogen and sulfur atoms in the residues. Hydrophobic and basic residues extend upward from the sequence line and hydrophilic residues extend downward. Wherever possible, heteroatoms were placed to the left or right of center. To distinguish serine from cysteine the circle for oxygen was placed to the left for the former and to the right for the latter. To distinguish acids from amides, the two oxygen circles of acids were placed at the same horizontal level, but the nitrogen circle of amides was placed one position

Figure 3. Symbols used in the Puppy (A) and Kitty (B) representations. Conventional one-letter symbols are used to identify the nucleotides and amino acids, respectively.

A

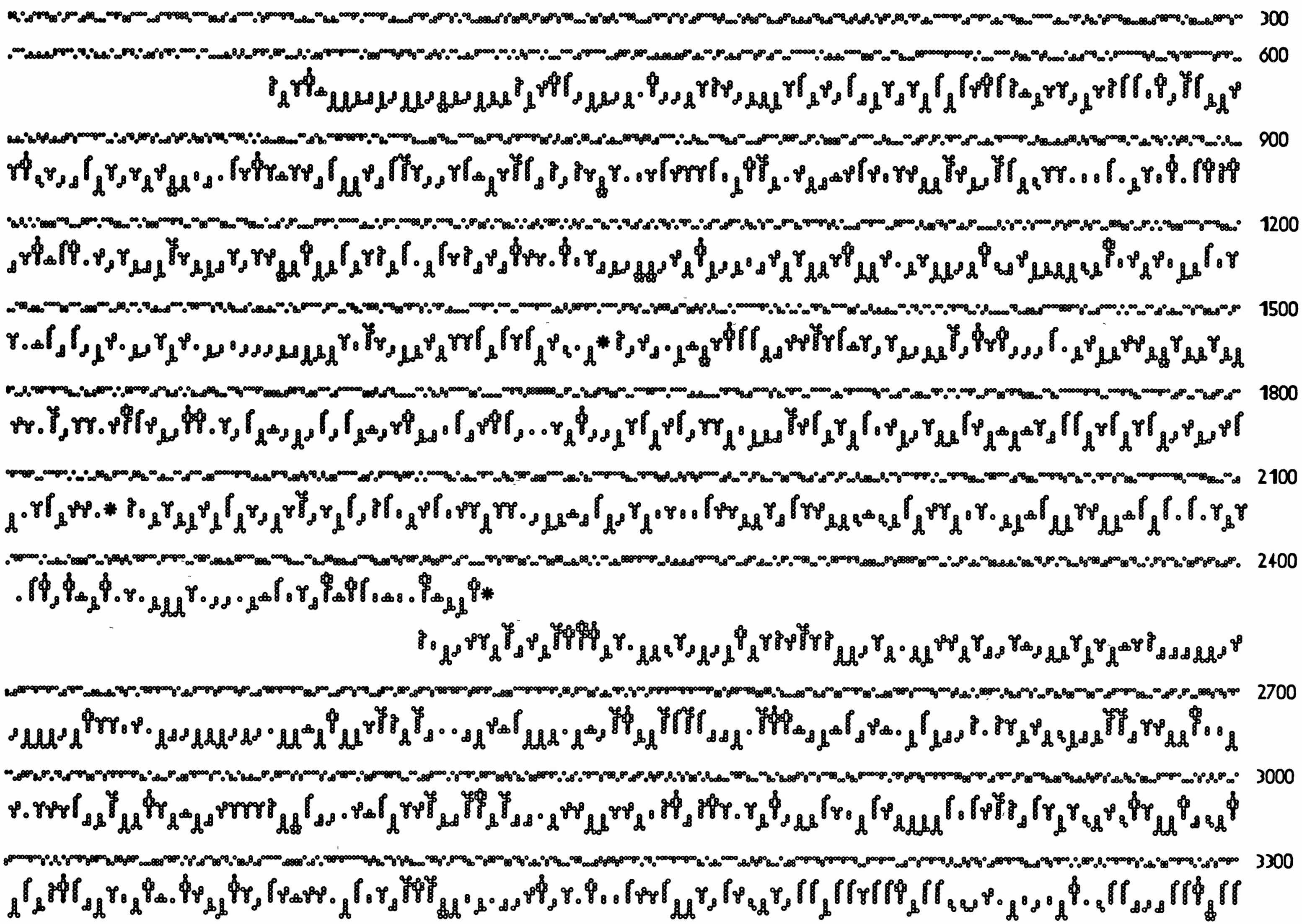
B

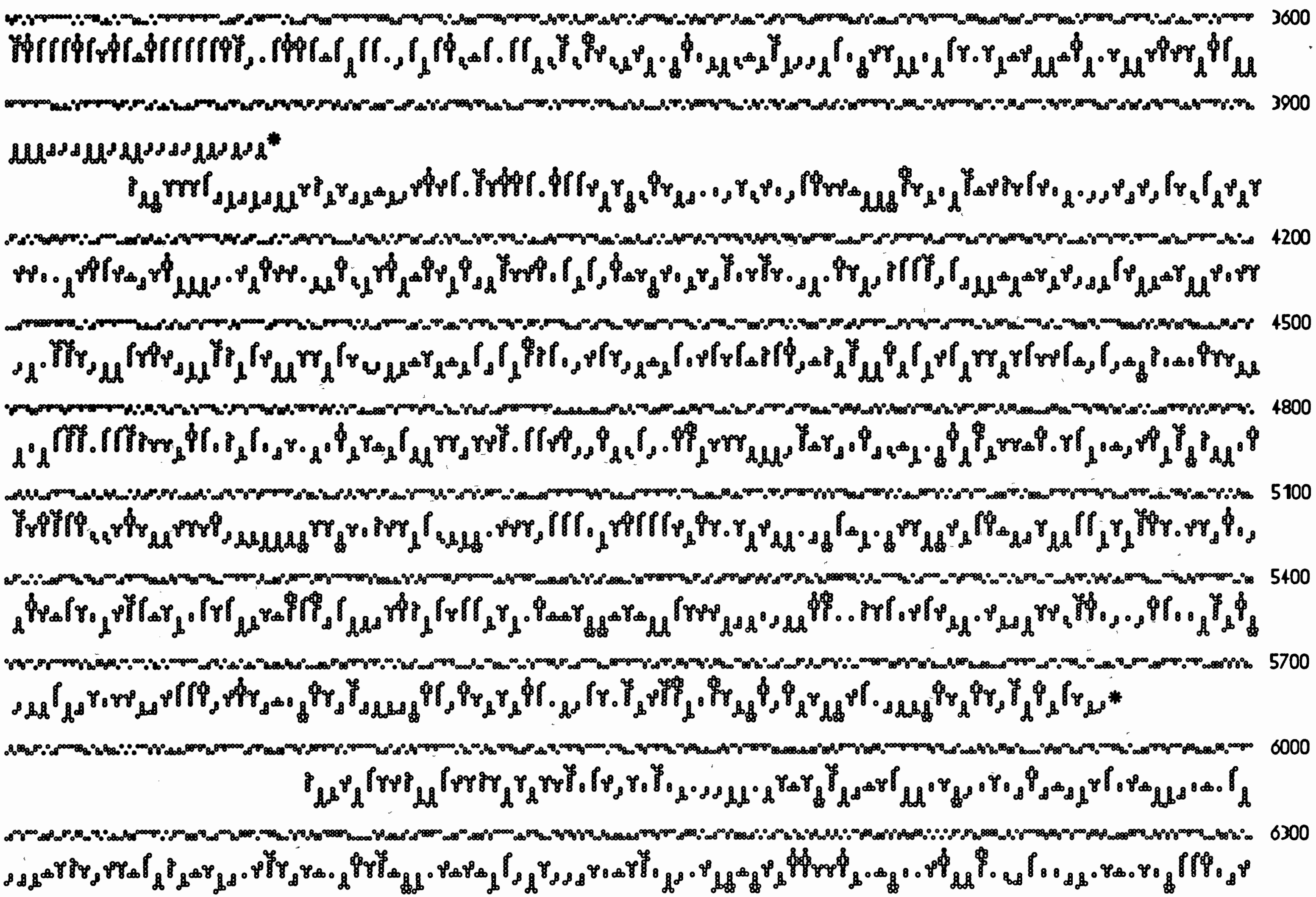
closer to the α-carbon row. Proline was arbitrarily represented as three consecutive circles in the α-carbon row with one circle centered in the row above. For simplicity, a bond closing the five-membered ring in tryptophan was omitted.

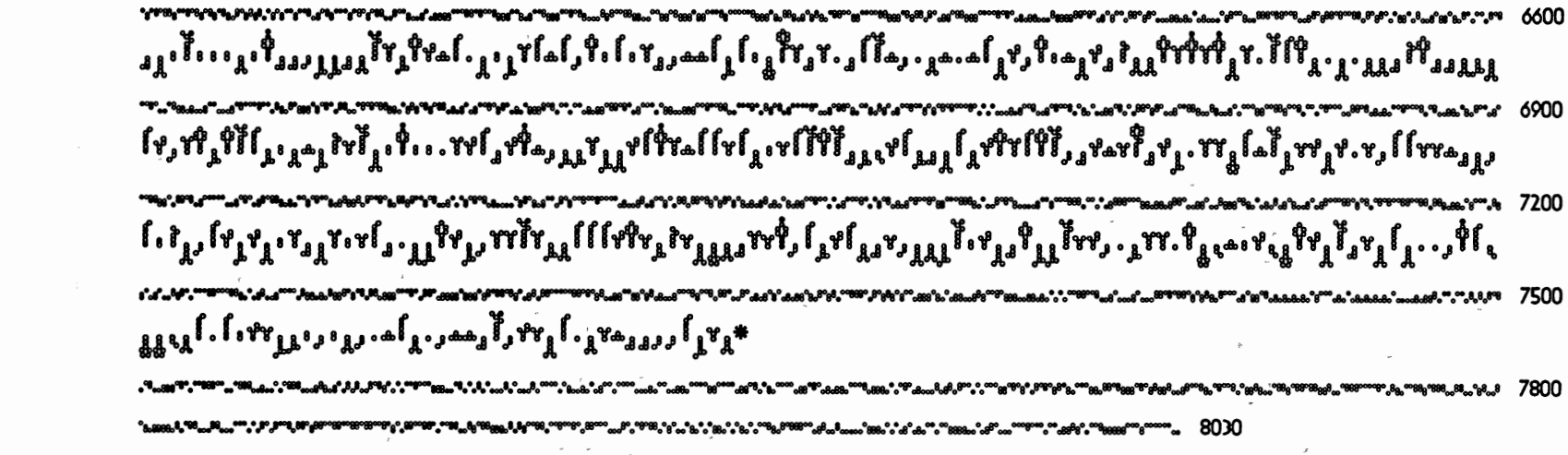
To implement Puppy and Kitty representations of nucleotide and amino acid sequences we designed two fonts for use with Macintosh computers. One font contains Puppy symbols. A combined font in which the lower case keys give Puppy symbols and the upper case keys give Kitty symbols was also created. The Kitty symbols are the width of three Puppy characters, allowing the presentation of nucleotide and amino acid sequences in adjacent rows. Both fonts were made in Postscript type 1 and Truetype formats. The fonts are available from the EMBL software server. The files PUPKIT_PS.HQX and PUPKIT_TT.HQX contain binhex-encoded, compressed files. The first contains Postscript type 1 fonts, suitable for use with Macintosh operating system 6. The second contains the same fonts but in True Type format and is suitable for system 7.

To illustrate the joint use of the Puppy and Kitty representations, we present the nucleotide and predicted amino acid sequences of CMV-1 (Figure 4). CMV-1 is the cauliflower mosaic virus (CaMV) DNA cloned in the plasmid pCaMV-1 (97). The nucleotide sequence was determined by enzymatic chain termination reactions using oligonucleotide primers specific to selected sequences of known CaMV DNAs

Figure 4. The nucleotide and derived amino acid sequences of DNA of cauliflower mosaic virus isolate CMV-1 in combined Puppy and Kitty representations. This figure spans pages 49-51.


300
600
900
1200
1500
1800
2100
2400
2700
3000
3300


3600
3900
4200
4500
4800
5100
5400
5700
6000
6300


6600
6900
7200
7500
7800
8030

(3, 32, 36, 85). The predicted open reading frames do not differ significantly in length or position from those of previously reported isolates. The CMV-1 nucleotide and predicted amino acid sequences deviate from those of the Cabbage S isolate (32) by about 3%. The nucleotide sequence has been deposited in GenBank/EMBL as accession number M90543.

In Figure 4, 16,060 nucleotides are represented (an inversion of the diagram displays the complementary strand) along with 2,303 amino acids at a higher information density per page than is usual for representations using the Roman alphabet representations. Further, visual scanning of the sequences for characteristic features is easier than with representations using letters of the Roman alphabet. For example, the region of the coat protein precursor (open reading frame 4) that contains a lysine rich stretch followed by an acidic rich C-terminus is clearly visible in the row from 3301 to 3600.

Sequence Analysis

Methods

The names and sources of the virus isolates analyzed in this study are shown in Table III. An alignment of these CaMV isolate genomes was developed using the program UMalign (73) which is described in Appendix A. This alignment was used to locate variable regions in the CaMV genome using the

TABLE III

GEOGRAPHIC AND PLANT SOURCES OF CAULIFLOWER MOSAIC VIRUS ISOLATES

ISOLATE	GEOGRAPHIC SOURCE	PLANT SOURCE	REFERENCE	ACCESSION NUMBER
Bari 1	Bari, Italy	*Diplotaxis tenuifolia*	(58)	D00335
*BBC	California, USA	Brassica rapa	This thesis	M90542
*Cabbage B-JI	Wisconsin, USA	Brassica sp.	(58)	-
*Cabbage S	Bari, Italy	Brassica ruvo	(32)	J02048
Campbell	California, USA	Brassica oleracea	(110)	M17415
*CM4-184	California, USA	Brassica sp.	(15)	M10385
*CM1841	California, USA	Brassica campestris	(87)	J02046
*CMV-1	California, USA	-	(97)	M90543
D-4	California, USA	Brassica campestris	(89)	M23620
*D/H	Budapest, Hungary	Brassica oleracea	(87)	J02047
*NY8153	New York, USA	Brassica sp.	(68)	M90541
PV147	Wisconsin, USA	Brassica rapa	(92)	X53860
S-Japan	Yokohama, Japan	Armoracia rusticana	(74)	X14911
W	California, USA	-	(10,108)	M32811
*XinJing	XinJiang, China	Brassica oleracae	(87)	-

*Complete genomic sequence is known

MalSig program (74). The CM4-184 isolate was not included in this analysis because its ORF2 deletion makes the ORF2 region appear hypervariable. The MalSig program compares residues at each position in the alignment to each other and calculates a similarity score for that position using a nucleic acid scoring table (identical = 2, transition = 1, transversion = 0). The similarity scores for a specified number (window size) of positions are then summed to give a similarity score for that window. A window size of 50 residues was specified, and a data point was collected once every 50 residues. Similarity scores were calulated for each window within the data set (160 windows total).

The CaMV genome alignment was also used to construct a CaMV consensus sequence. The consensus sequence was constructed one residue at a time by visual inspection. The nucleotide present in the majority of the sequences was chosen for the consensus sequence. If no majority nucleotide was found, isolate CM4-184 was excluded due to its similarity to isolate CM1841. The CaMV consensus sequence was used as a reference by which to identify and characterize isolate-specific base substitutions, insertions, and deletions.

In order to observe the phylogenetic relationships among CaMV isolates, I chose another caulimovirus as the tree outgroup. Based on comparisons of sequences of three caulimovirus members (83), I concluded that carnation etched ring virus (CERV) was more closely related to CaMV than to figwort mosaic virus (FMV). Thus, CERV was chosen as the

outgroup for the construction of CaMV phylogenetic trees. CERV was first aligned to CaMV isolate CMV-1 and then added to the alignment of other CaMV isolate sequences using UMalign and MacVector™. Phylogenetic trees were constructed by three different methods available in the PHYLIP package for phylogenetic inference (28). A brief description of each method used may be found in Appendix A. When necessary, program constants were adjusted to accommodate the input file. Parsimony trees were constructed using DNAPARS. Parsimony trees were shown because it was convenient to determine the significance of the branching order for these trees. A bootstrap value for each node in parsimony trees was calculated (using DNABOOT) by determining the number of times that node was present out of 500 randomized replicates. Minimum mutation distances between the isolates were calculated by DNADIST using the Kimura 2-parameter option (61). Distance trees were constructed from the resulting distance matrices using FITCH. The application of the molecular clock model to distance trees was attempted using KITSCH. Maximum likelihood trees were constructed using DNAML. All PHYLIP programs were executed either on a Macintosh IIsi or through use of the Oklahoma University Computer Group resource. To ensure that the best phylogenetic tree was obtained, each program was executed at least three times and, where possible, the input order of data was randomized using the Jumble option. Global rearrangement of each tree was also performed. Testing for

probable recombination between isolate genomes was performed using the VTDIST program (88) executed on an IBM-compatible personal computer. For this analysis, a fragment is defined as a stretch of sequence that is identical in two sequences. Fragment length is measured in total residues (uncondensed fragment) or number of polymorphic loci (condensed fragment). The algorithm searches for fragments that are significantly larger than expected based on random distributions of polymorphic sites. The P-value for each fragment represents the fraction of permuted fragments greater than or equal (in length) to the observed fragment. For these tests I considered a fragment significant if its P-value was 0.05 or lower. Options were invoked to test for outer recombination (between a sequence in the sample and one from outside the sample) and inner recombination (between pairs of sequences within the sample).

Results

A similarity plot for CaMV isolate nucleotide sequences is shown in Figure 5. Open reading frames (ORFs) 1, 2, 3 and 5 along with the intergenic region appear to be the least variable genomic regions. ORF 4 is slightly more variable while ORF6 is the most variable, possessing two hypervariable regions.

The base composition of the positive strand of the consensus sequence was 37% A, 19% G, 23% T, and 21% C. The consensus sequence was used as a reference by which to

Figure 5. Similarity plot for the genomes of eight sequenced CaMV isolates. Numbers above the plot indicate ORF regions; IGR = large intergenic region. A window of 50 residues was specified, and data points were taken every 50 residues.

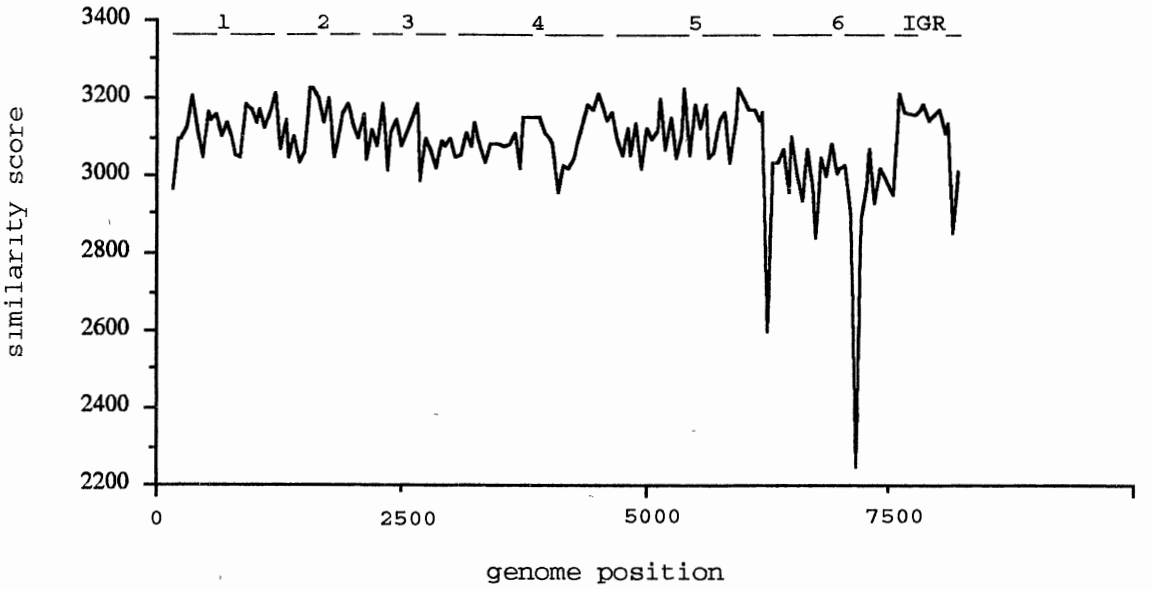

1
2
3
4
5
6
IGR
3400
3200
3000
2800
2600
2400
2200
similarity score
0
2500
5000
7500
genome position

TABLE IV

CAMV BASE SUBSTITUTION PROFILE

Nucleotide in Consensus	Nucleotide in Isolates A	G	C	T
A		25±7	11±4	12±6
G	26±11		4±3	5±2
C	9±6	4±3		38±15
T	10±7	5±3	31±12	

± Indicates standard deviation.

TABLE V

MEAN PERCENT SILENT SUBSTITUTIONS PER CAMV OPEN READING FRAME

Open Reading Frame	Mean % Silent Substitutions (± standard deviation)	Mean Number of Changes (± standard deviation)
1	75±14	18±5
2	69±18	7±2
3	79±10	7±2
4	75±12	42±15
5	90±6	45±16
6	54±11	41±22

categorize isolate-specific base substitutions (Tables IV and V). Base substitutions were found at 1077 positions out of 8110 possible sites. Transitions dominated over transversions by 2:1 (Table IV). Also, transversions involving A dominated over transversions involving G 2:1. Substitutions were also classified as either silent or expressed (Table V). The majorities of substitutions in each ORF were silent. ORFs 1-4 have approximately the same percentage of silent substitutions, while that of ORF5 was significantly higher, and that of ORF6 was considerably lower. Neighboring nucleotides of isolate-specific base substitutions (relative to the consensus sequence) were examined for evidence of mis-incorporation due to transient template misalignment. For substitutions resulting from transient template misalignment, the 3' neighboring nucleotide is identical to the base resulting from the substitution (ie: the sequence ATTGC would become ATTCC (63)). I examined all substitution sites for CaMV isolates (on the plus and minus DNA strands) for evidence of transient template misalignment. Of the possible substitution sites, an average of 28.5% of the base substitutions occurred next to identical neighboring nucleotides. The distribution of nucleotides in the consensus sequence results in a 27% chance of two neighboring nucleotides being identical. Therefore, no significant evidence of transient template misalignment was found for CaMV.

TABLE VI

CAMV ISOLATE-SPECIFIC INSERTIONS AND DELETIONS

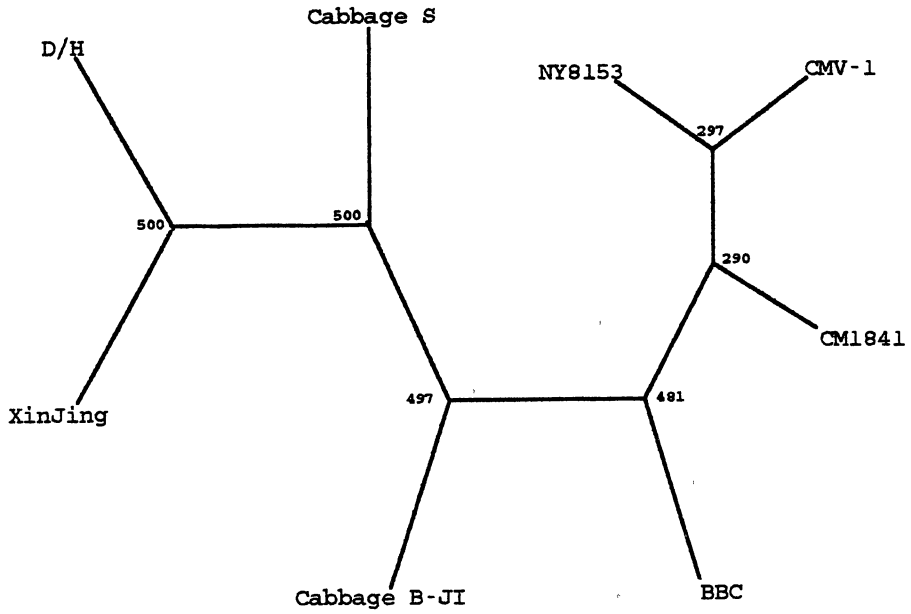
Position*	Isolate(s)	Insertion (I)/ Deletion (D)@	No. Nucleotides
306	D/H, XinJing	I	1
595	Cabbage B-JI	I	1
1347	Cabbage B-JI	I	1
1348	D/H, XinJing	D	5
1390	CM4-184	D	422
2411	D/H, Cabbage S	I	3
2442	XinJing	I	41
2588	NY8153, CMV-1, Cabbage B-JI	D	3
3347	NY8153, CMV-1, BBC, Cabbage S	I	3
3680	NY8153, BBC, CM1841, CM4-184	I	3
3717	D/H, XinJing	I	6
4226	D/H, XinJing	D	21
5777	CM1841, CM4-184	I	1
7321	D/H, XinJing	I	6
7365	XinJing	I	3
7373	XinJing	D	3
7381	XinJing	D	3
7434	CM4-184	D	1
7439	XinJing	I	1
7541	D/H	D	1
7550	XinJing	I	2
7555	Cabbage S	D	9
7557	XinJing	D	1
7558	D/H	D	1
7566	Cabbage B-JI	I	1
7583	Cabbage B-JI	I	1
7870	XinJing	D	1
8055	Cabbage B-JI, BBC	I	2
8079	Cabbage B-JI	D	1
8108	Cabbage B-JI	D	1

*According to CaMV isolate/consensus alignment (Figure 30,Appendix B)
@Relative to consensus sequence

An alignment of CaMV sequences with the consensus sequence was used to identify isolate-specific insertions and deletions (Table VI). Both insertion and deletion events were found in every sequenced CaMV isolate, with the exception of isolate CM1841, which only had insertions. An alignment gap shared by more than one isolate was considered as one event. I observed a slight excess of insertions (17 events) over deletions (13 events). Insertion events ranged from 1 to 41 nucleotides in length, averaging 2 nucleotides in length. Deletion events varied in length from 1 to 422 nucleotides, averaging 5 nucleotides. In considering all CaMV genomic regions, 43% of insertion/deletion events were in the large intergenic region. Of all CaMV ORFs, ORF4 contained the most insertion/deletion events (38%). Of all CaMV isolates, the nucleotide sequence of isolate XinJing contained the most insertion/deletion events. Also, 17% of all insertion/deletion events were shared between isolates XinJing and D/H.

The frequency and position of insertion and deletion events in CaMV isolate DNAs were examined (relative to the consensus sequence). The majority (56%) of insertion/deletion events may be attributed to transient template misalignment by the polymerase either at stretches of the same nucleotide (ie: an oligo(A) stretch), or at regions of direct repeats. Of the remaining events, four could possibly be deletions consistent with transient

Figure 6. Phylogenetic species tree for eight CaMV isolates obtained by the bootstrapped parsimony method. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.


Cabbage S
D/H
NY8153
CMV-1
297
500
500
290
CM1841
497
481
XinJing
Cabbage B-JI
BBC

template misalignment. Of the nine unexplained events, four involved isolate XinJing.

The alignment of CaMV sequences to the CERV nucleotide sequence was used as input for phylogenetic analysis. Because the placement of CERV varied extensively in individual trees, it was excluded from the figures in this thesis. The phylogenetic tree shown in Figure 6 depicts the inferred relationship for sequenced CaMV isolate genomes. Isolate CM4-184 was excluded from this tree due to its ORF 2 deletion and similarity to isolate CM1841. The 'species tree' (a tree constructed using each isolate's complete genomic sequence) in Figure 6 was the most parsimonious tree constructed after completion of 500 replicates by the bootstrapped DNA parsimony. The cluster of isolates on one side of Cabbage B-JI (XinJing, D/H, Cabbage S) were isolated from the Old World. New World isolates (Cabbage B-JI, BBC, NY8153, CMV-1, CM1841) clustered separately. All but two of the nodes in the species tree shown in Figure 6 were present in greater than 95% of the bootstrap replicates. Bootstrap values of the nodes within the New World cluster are lower that those in the Old World cluster, suggesting that the exact branching pattern within the New World group is uncertain. Members of the Old and New World isolate clusters were the same in species trees constructed by the parsimony, distance and maximum likelihood methods (see Appendix B). The placement of isolates within the Old World cluster was the same regardless of the method used. However, the

placement of isolates within the New World cluster was not consistent among all species trees constructed. Isolate CMV-1 was placed on the same branch as NY8153 using the parsimony and maximum likelihood methods, but branched with isolate CM1841 when the distance method was used. I attempted to apply a molecular clock to the distance matrix so as to estimate a CaMV mutation rate and the time of divergence. I used the F-test (25) to compare the KITSCH and FITCH distance trees. The calculated F-value suggested the trees were significantly different. Thus I rejected the validity of the molecular clock for these data.

Phylogenetic trees that are constructed using the same gene from different species are termed 'gene trees' (76). Separate phylogenetic trees were constructed for each of the six major CaMV ORFs and for the large intergenic region. Again all three methods of construction were used. Isolates used for these comparisons include those found in the species tree (Figure 6) and also those isolates for which a complete nucleotide sequence for that gene was available. Figures 7 and 8 depict the most parsimonious bootstrapped trees for CaMV ORF2 and ORF6, respectively. In these gene trees, only two exceptions to the Old and New World branching pattern were found. For the ORF 2 tree, isolate Cabbage B-JI branched with the Old World isolates while isolate S-Japan branched with the New World cluster Old and New World isolate With these two exceptions, partially sequenced isolates included in the gene trees branched according to their place

Figure 7. Bootstrapped parsimony gene tree for ORF2 of ten CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

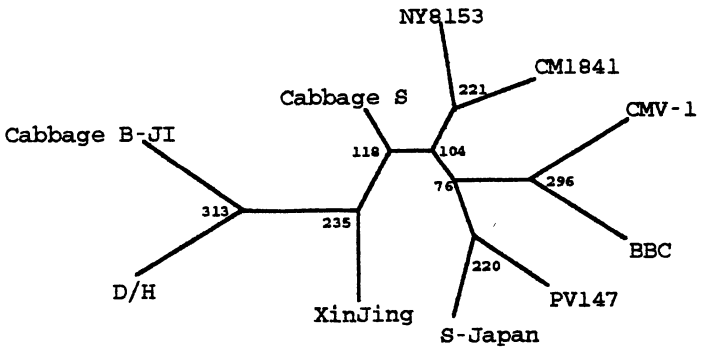
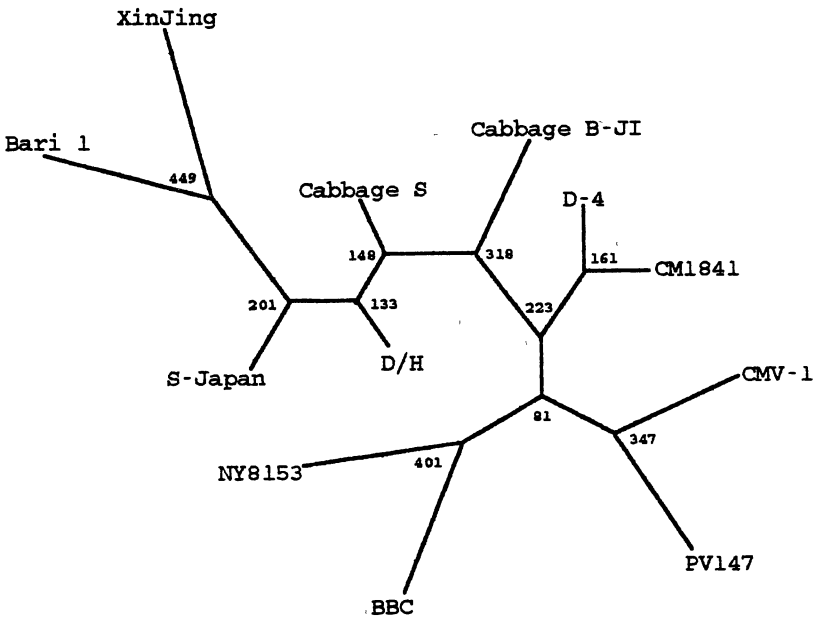

NY8153
CM1841
Cabbage S
CMV-1
Cabbage B-JI
221
118
104
76
296
313
235
BBC
220
D/H
PV147
XinJing
S-Japan

Figure 8. Bootstrapped parsimony gene tree for ORF6 of twelve CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.


XinJing
Bari 1
449
Cabbage B-JI
Cabbage S
D-4
148
318
161
CM1841
201
133
223
S-Japan
D/H
CMV-1
91
347
NY8153
401
PV147
BBC

of collection. Isolate PV147 branched with the New World isolates in trees for both ORF2 and ORF6. Isolate D-4 branched with the New World isolates in the ORF6 tree. The Bari 1 isolate branched with the Old World isolates in the ORF6 tree.

The exact placement of isolates within the New World cluster was not consistent between several of the gene trees and the species tree. For New World isolates, the ORF2 tree differed from the species tree in the placement of CMV-1 on the branch with BBC rather than between CM1841 and NY8153. The ORF6 tree differed from the species tree only in the placement of BBC between CMV-1 and NY8153 rather than between Cabbage B-JI and CM1841.

The ORF6 trees constructed by other methods differed from the ORF6 parsimony tree only in the exact placement of the Cabbage S isolate relative to D/H. ORF2 trees constructed by other methods agreed with the parsimony tree in branching order.

The Old and New World isolate lineages were present in all gene trees constructed for other ORFs (with the exception of S-Japan in the ORF1 tree) and for the large intergenic region (Appendix B) regardless of the method used. Isolate S-Japan was an exception to the lineage pattern by branching with the New World isolates for the ORF1 trees. Exact placement of isolates within each cluster was not consistent. In general, the bootstrap values for parsimony tree nodes

were lower in the gene trees than in the species tree, due to the reduced size of the data sets.

Thus, with two exceptions, the Old World and New World virus clusters were found in all trees constructed. However, the exact placement of isolates within each lineage was not consistent. Variation in the exact placement of *E. coli* strains among phylogenetic trees has been attributed to gentic exchanges between tree members (18). The CaMV DNA sequence alignment was examined in regions where the gene tree was not congruent with the species tree. For example, Cabbage B-JI branched with the Old World isolates in the ORF2 tree, but with the New World isolates for all other trees. Examination of the Cabbage B-JI and Old World isolate sequences in the ORF2 region revealed a stretch of 400 nucleotides where Cabbage B-JI is more like the Old World isolates than the New World isolates. Thus, a recombination event between Cabbage B-JI and an Old World isolate may have occurred in this region to produce the observed branching pattern in the ORF2 tree. Similar investigations were conducted for other isolates with inconsistent branching patterns. Isolate BBC branched closer to isolate CMV-1 in the ORF2 tree than in other trees. Examination of these isolate sequences in the ORF2 region showed a region of 120-180 nucleotides in length where BBC and CMV-1 were very similar. CM1841 branched closer to CMV-1 in the ORF5 tree relative to trees for ORFs 4 and 6. A 200 nucleotide stretch of similarity between CM1841 and CMV-1 in the ORF5 region may

account for this change. BBC branched closer to Cabbage B-JI in gene trees for ORFs 4 and 5 and in the intergenic region tree (relative to all other trees). Examination of the BBC and Cabbage B-JI sequences in these regions revealed stretches of similarity 100-200 nucleotides in length in these three regions. The placement of NY8153 was close to Cabbage B-JI in gene trees for ORFs 1, 2, and 3, but not in all other gene trees constructed. However, no convincing stretches of sequence similarity between Cabbage B-JI and NY8153 were found in ORFs 1 through 3.

The method of Sawyer (88) was used to further test for recombination between pairs of sequences within the CaMV alignment (inner-recombination). This test can also detect recombination between an aligned sequence and one not included in the alignment (outer-recombination). CaMV isolates used for this analysis are the same as those used to construct the species tree (Figure 6). No uncondensed fragments were significantly longer than expected from a random distribution of polymorphic sites. The significant (P-value of 0.05 or less) outer- and inner-condensed fragments are listed in Table VII , along with their genomic location. Inner-condensed fragments varied in length from 115 to 246 nucleotides. With one exception (between Cabbage S and D/H), inner fragments were found only in ORF6. Predicted inner-fragments were confined to isolates within the same lineage with the exception of fragments predicted for Old

World isolate Cabbage S and New World isolates NY8153 and CM1841.

Outer condensed fragments were 20 to 50 nucleotides in length. All outer fragments were found for the XinJing isolate in the ORF6 region, suggesting that XinJing is unlike other CaMV isolates in several regions of ORF6. One of the predicted fragments was within the ORF6 3' hypervariable region. The position of outer-fragments in ORF6 overlaps with all inner-fragments located in ORF6. Thus, it is likely that the outer-fragments for XinJing in ORF6 increased the statisical significance of inner-fragments in that region. Thus, the only statistically significant inner-fragment detected was shared between Cabbage S and D/H in large intergenic region.

TABLE VII

RESULTS FROM THE SAWYER TEST FOR RECOMBINATION

Isolate(s)*	Nucleotide Position[£]	Fragment Length[§¶]	# Polymorphic sites	P-Value
CMV-1/BBC	6554	246	63	0.0001
CMV-1/CM1841	6947	224	55	0.0008
D/H/Cabbage S	7484	400	38	0.0032
CM1841/Cabbage S	7224	128	38	0.0032
CMV-1/Cabbage B-JI	7221	115	37	0.0047
NY8153/Cabbage B-JI	7221	115	37	0.0047
D/H/Cabbage S	6678	210	46	0.0069
BBC/CM1841	6815	168	42	0.0222
NY8153/Cabbage S	7224	165	42	0.0222
NY8153/CM1841	7196	160	42	0.0222
XinJing	6997	43	9	0.0013
XinJing	6638	28	9	0.0013
XinJing	7262	20	9	0.0013
XinJing	6686	50	7	0.0220
XinJing	7283	41	7	0.0222

*Two isolates indicate recombination between those two isolates. One isolate indicates recombination between that isolate and a sequence not considered in this test.

[£]Numbering is the same as that used for the Cabbage S isolate (32).

[§]Only fragments with a P-value of 0.05 or less are reported.

[¶]Represents uncondensed fragment length.

CHAPTER IV

DISCUSSION

The results indicate that the majority of the CaMV genome is well conserved among CaMV isolates both in nucleotide and predicted amino acid sequence. Although the number of base substitutions in ORF5 is approximately equal to that of ORFs 4 and 6, the density of coding base substitutions per kilobase is lowest for ORF5 (relative to all other ORFs). Thus, ORF5 is the most stringently conserved of all CaMV ORFs, suggesting that the preservation of the amino acid sequence of the viral reverse transcriptase is important for CaMV propagation. The nucleotide sequence of ORF6 contains two hypervariable regions when compared to the rest of the CaMV genome. These two hypervariable regions in the nucleotide sequence correspond in position with those noted for the amino acid sequences of CaMV ORF6 by Sanger et al. (87). The product of ORF6 has been suggested to be a host-range determinant for CaMV (13, 89, 90). Although most of the CaMV isolates used in this study were isolated from the same host genus, host ranges vary among CaMV isolates (13, 89, 90). Thus, the variation in ORF6 of isolates collected from the same host genus may reflect differing

abilities to infect other, as yet untested, hosts. For example, mutants of isolate D-4 with point mutations specific to the two hypervariable regions in ORF6 were shown be altered in host interactions relative to wild-type D-4 (13). Therefore, ORF6 variation directed by host-imposed selection may lead to evolution during adaptation to a new host. Variation in the HIV-1 envelope gene (which may correspond to ORF6 of CaMV (50)) might be responsible for the great immunological diversity of the virus (93), suggesting evolutionary pressures may favor mutation in the HIV-1 envelope gene. Host-range related adaptive pressures may act on CaMV ORF6. Alternatively, evolutionary constraints may not be as stringent for the ORF6 region, relative to the remainder of the CaMV genome.

The retrovirus HIV-1, like CaMV, uses reverse transcription as a mechanism by which to replicate its genome. The retroviral encoded reverse transcriptase, due to its lack of proofreading functions, might account for the high retrovirus mutation rate of 10^{-2} to 10^{-3} substitutions per site per year (39). Since both pararetroviruses and retroviruses employ reverse transcription in their life cycles, a mutation rate similar to that of retroviruses would be expected for pararetroviruses. However, the estimated mutation rates for pararetroviruses are one to two orders of magnitude lower than those of retroviruses (38, 78).

A base substitution profile for CaMV isolates was constructed (Table IV) and compared to those of retroviruses

in order to gain perspective on how and when mutations in the CaMV genome occur during the virus replication cycle. Excesses of one type of base substitution (asymmetries) have been been found in the base substitution profiles for retroviruses (5, 84, 93). Asymmetries were noted in the CaMV base substitution profile. First, transitions dominated over transversions 2:1, an asymmetry also observed in HIV-1 base substitution profiles (84, 93). Second, transversions involving A dominated over transversions involving G 2:1. CaMV transversion frequencies involving each base correlated with the base composition of the positive strand of the CaMV consensus sequence. An excess of G -> T transversions has been found when testing the fidelity of HIV-1(84), avian myoblastosis virus (AMV), and Moloney murine lukemia virus (MMLV) reverse transcriptases (5). The excess of G -> T transversions did not reflect the base composition of the nucleic acid being polymerized (84). Roberts(5) and Bebenek(84) suggested transient template misalignment as a possible mechanism to account for the excess of G -> T transversions in the retrovirus base substitution profiles. I did not observe significant evidence of transient template misalignment for CaMV based upon the base substitution profile. Shimizu et al.(93) reported a large excess of G <-> A transitions in a base substitution profile constructed for HIV-1, and attributed the excess to the error-prone nature of the HIV-1 reverse transcriptase. Vartanian et al. (106) observed an excess of G -> A transitions for HIV-1, and

attributed this excess to transient template misalignment by the HIV-1 reverse transcriptase. I did not find an excess of A <-> G transitions for CaMV. Instead, for CaMV the number of G <-> A transitions was comparable to that of C <-> T transitions, a result similar to that found for influenza virus (93).

Thus, the base substitution profile for CaMV DNA is unlike those examined for HIV-1 and other retroviruses, except for the domination of transitions over transversions 2:1. I suggest two possible explanations for the differing base substitution profiles of CaMV and retroviruses. First, the base substitution profile for CaMV DNA provides no evidence that CaMV DNA is prone to errors characteristic of retrovirus reverse transcription. Thus, the reverse transcriptase of CaMV may not be as error-prone or may commit different errors when compared with that of retroviruses. Alternatively, the majority of CaMV spread through the plant may occur via amplification of the minichromosome by DNA replication, not reverse-transcription. CaMV has been shown to spread through the plant via the phloem tissue (66). Once in the phloem tissue of the plant, CaMV may reach the actively dividing cells of young leaves. Once inside an actively dividing cell, CaMV could be spread throughout the plant by simple cell division, requiring only the amplification of the minichromosome in the host nucleus. If minichromosome amplification occurs via DNA replication instead of reverse transcription, the importance of reverse

transcription for the spread of CaMV infection would be reduced. Both explanations could account for the observed CaMV base substitution profile and the lower estimated CaMV mutation rate (6×10^{-4} substitutions per site per passage) (78) relative to that of retroviruses (10^{-2} to 10^{-3} substitutions per site per year) (39).

The results of examination of the sequences surrounding insertion and deletion events in CaMV isolate DNAs indicate that most of these events may be attributed to transient template misalignment by the polymerase either at stretches of the same nucleotide (ie: an oligo(A) stretch), or at regions of direct repeats. Of the unexplained events, 44% involve isolate XinJing. Thus, XinJing may mutate differently or more often relative to other CaMV isolates. Alternatively, XinJing may be more diverged from the CaMV consensus sequence than other isolates.

In addition to examining the variability of the CaMV genome, I have attempted to determine the phylogenetic relationships among different isolates of CaMV in order to better understand CaMV evolution. Species and gene trees were constructed, each by three different methods, parsimony, distance, and maximum likelihood. Two discrete virus lineages were present in the majority of trees constructed, regardless of the method used. One lineage consisted of CaMV isolates collected in Old World countries of Europe and Asia, while the other lineage was composed of New World isolates. The branching of partially sequenced isolates in gene trees also

suggests the two lineage branching pattern, with the exception of isolate S-Japan in gene trees for ORFs 1 and 2. A more detailed history of the origination of crucifers in Japan may offer a possible explanation for the branching pattern of isolate S-Japan.

Sanger et al. (87) attempted to infer evolutionary relationships among CaMV isolates, based on comparisons of ORF6 predicted amino acid sequences. Evolutionary relationships were suggested for the following groups of isolates: Bari 1/XinJing, CM1841/D/H, and D-4/CM1841/S-Japan. Our results for the ORF6 nucleotide sequence support the relationships suggested by Sanger for Bari 1/XinJing and for D-4/CM1841, but not for CM1841/D/H or for isolates D-4/CM1841/S-Japan.

Insertion and deletion events noted among CaMV isolates were reflected in corresponding gene trees. For example, insertion/deletion events were shared between isolates D/H and XinJing in ORFs 4, 5, and 6. The corresponding parsimony gene trees show that D/H and XinJing branch together. Another example is the insertion event shared between BBC and Cabbage B-JI in the large intergenic region. The intergenic region tree (Appendix B) reflects this event by the branching patterns of BBC and Cabbage B-JI.

The Old and New world isolates may have evolved as separate lineages from a hypothetical CaMV common ancestor. Alternatively, one lineage may have evolved from the other. The latter explanation seems more plausible considering two

pieces of evidence. First, although cultivated in Europe for over 4000 years turnips (and possibly other cultivated cruciferae) were not introduced to the New World until around 1600 (82). Thus, if CaMV was transported to the New World via one of its hosts, the New World lineage may have evolved from an isolate of the Old World. Second, a molecular clock was applied to the distance trees (Appendix B) using the KITSCH program. The resulting trees were then tested for significance using the F-test (26). Although Felsenstein has expressed reservations in using the F-test for sequence data (27), the validity of the molecular clock for these data was rejected based upon the results of the F-test. Thus, no CaMV mutation rate or point of possible divergence between the two lineages was estimated. However, when considering only the topology of the KITSCH trees, the 2-lineage branching pattern was found, with the common ancestor of the Old World isolates being less diverged from the hypothetical caulimovirus common ancestor than that of the New World isolates. Thus, it seems likely that one branch of the Old World lineage gave rise to the New World isolates when they were separated geographically by the introduction of the crucifers to the New World.

Plant virus evolution may be influenced by various different factors, including both virus-vector (52, 70) and virus-host interactions (14, 52, 70). No CaMV isolates clustered according to whether they are aphid transmissible or non-transmissible. The majority of CaMV isolates used in

this study were isolated from Brassica species. No branching pattern specific to host source was found for CaMV isolates differing in host genus. Instead, my results suggest that the major factor contributing to CaMV evolution is CaMV-host geographic distribution. An evolutionary influence by host geographic distribution has been suggested for other plant viruses (7, 52, 70). Based upon hybridization tests, Blok et al.(7) suggested that turnip yellow mosaic virus (TYMV) isolates separate into two distinct lineages, one of Australian origin and the other of European origin. Howarth et al.(52) noted that geminivirus isolates clustered in phylogenetic trees according to their geographic origin. The effect of host geographic distribution on viral evolution has also been well documented for animal viruses (17, 67).

The species tree derived from comparisons of complete genomic sequences best represents the phylogenetic relationship among CaMV isolates. When comparing the CaMV gene trees, the Old and New World lineages are consistently found (with the two exceptions noted earlier) but the exact placement of isolates within the New World lineage was less consistent than that of the Old World lineage. Exact placement of strains also varied among trees for different *E. coli* genes (18). Dykhuizen and Green suggested that recombination events among the different *E. coli* strains were an important parameter influencing the placement of strains in phylogenetic trees. Li et al. (67) suggested that recombination had occurred between isolates of HIV-1, based

upon variation among gene tree branching patterns. Isolate sequences were examined in regions where their branching pattern in gene trees was inconsistent. In most cases considered, regions of possible recombination were found between CaMV isolates that could account for their inconsistent branching patterns.

The Sawyer test (88) was used to further examine whether recombination could be responsible for the inconsistent placement of isolates within the two lineages of CaMV phylogenetic trees. The test detects stretches of similar sequence between two isolates. Sawyer's method automatically controls for variable mutation rates and does not depend on potentially monophyletic subsets of the sample. One statistically significant inner fragment was found for Old World isolates D/H and Cabbage S and was located in the large intergenic region between the 35S RNA transcription start site and the gap in the DNA (-) strand. This fragment may have been produced via a reverse-transcription mediated template switch from the 5' end of one 35S RNA to the 3' end of another. This type of template switch was previously suggested to have occurred between CaMV isolates CM4-184/Cabbage S(15) and between W/Cabbage B-JI(105).

Outer-condensed fragments for XinJing were located in ORF6 between the two CaMV RNA transcription start sites. Five outer fragments for XinJing were inferred throughout this region, separated by small stretches of nucleotides where the sequence of XinJing is similar to other CaMV

isolates. The Sawyer test limits outer-fragment length to the region of polymorphism unique to one isolate. Considering this limitation of the Sawyer test, it is possible that these fragments are part of one recombination event which resulted from reverse-transcription mediated template switches from the 5' end of the 35S RNA to the 3' end of the 19S RNA and then back to the original 35S RNA. Recombinant junctions consistent with this type of template switch have been previously documented by Vaden and Melcher (105).

Recombination between two CaMV isolates would require the presence of both genomes in the same cell. Thus, an inter-isolate recombination event would dictate the same geographic location. Cross protection, the prevention of host super-infection by strains of the same virus, has been shown to occur between isolates of CaMV (103, 111). Therefore, simultaneous infection by both CaMV isolates would also be required to produce inter-isolate recombination. The one inner-fragment detected by the Sawyer test was for isolates within the same lineage (Cabbage S and D/H). The predicted recombination event for Cabbage S and D/H was not reflected in the phylogenetic tree for the large intergenic region, possibly due to the inclusion of isolate CM4-184 which has been shown to be similar to Cabbage S in the intergenic region (15). Other inconsistencies were noted between the results of the Sawyer test and those of the phylogenetic analysis. For example, no recombination was predicted for isolate Cabbage B-JI and any Old World isolate

in the ORF2 region. However, Cabbage B-JI clusters with the Old World isolates in the ORF2 gene tree, and inspection of Cabbage B-JI and Old World isolate sequences in ORF2 supports a possible recombination event for this region. Other comparisons of the gene trees and specific isolate sequences also suggest that recombination may be influencing CaMV evolution. With the exception mentioned earlier, the Sawyer test does not predict significant recombination between any of the CaMV isolates considered in this study. Thus, for detecting recombination events, the Sawyer test appears less sensitive than gene tree phylogenetic analysis. The Sawyer test searches only for similar stretches of sequence between two isolates, not specific recombinant junctions. Since CaMV isolate sequences vary at only about 5% (3) of their nucleotide positions in pair-wise comparisons, the inferred recombination may only reflect the similarity between the isolates, not true recombination events. Therefore, further studies may be necessary to determine if recombination is in fact influencing CaMV isolate phylogenetic distribution.

The quasispecies concept developed by Eigen and shown to occur in RNA phage QB by Weissmann (20), suggests that the result of self-replication competition over long periods of propagation is the eventual conservation of the master species. Evidence supporting the quasispecies concept has been suggested for several RNA viruses, including HIV-1 (8, 45, 96). The genetic relationship between CaMV isolates

predicted by the tree model does not support the quasispecies concept. Phylogenetic analysis results support the existence of two separate CaMV lineages separated geographically for almost 400 years. Within these two lineages, individual isolates continue to evolve. These lineages were found in the majority of phylogenetic trees that were constructed, regardless of the method used. Thus, no evidence of a conserved master sequence was found. Therefore, isolates of CaMV do not constitute a quasispecies.

REFERENCES

1. Armour, S. L., Melcher, U., Pirone, T. P., Lyttle, D. J., Essenberg, R. C. (1983). Helper component for aphid transmission encoded by region II of cauliflower mosaic virus DNA. Virology, 129, 25-30.

2. Attwood, T. K., Eliopoulos, E. E., Findlay, J. (1991). Multiple sequence alignment of protein families showing low sequence homology: A methodological approach using database pattern-matching discriminators for G-protein-linked receptors. Gene, 98, 153-159.

3. Balazs, E., Guilley, H., Jonard, G., Richards, K. (1982). Nucleotide sequence of DNA from an altered-virulence isolate D/H of the cauliflower mosaic virus. Gene, 19, 239-249.

4. Bass, B., Weintraub, H., Cattaneo, R., Billeter, M. (1989). Biased hypermutation of viral RNA genomes could be due to unwinding/modification of double-stranded RNA. Cell, 56, 331.

5. Bebenek, K., Abbotts, J., Roberts, J., Wilson, S., Kunkel, T. (1989). Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. Journal of Biological Chemistry, 264, 16948-16956.

6. Blackburn, G. M., Gait, M. J. (1990). Nucleic Acids in Chemistry and Biology. New York: IRL Press.

7. Blok, J., Mackenzie, A., Guy, P., Gibbs, A. (1987). Nucleotide sequence comparisons of turnip yellow mosaic virus isolates from Australia and Europe. Archives of Virology, 97, 283-295.

8. Cattaneo, R., Schmid, A., Eschle, D., Baczko, K., Meulen, V., Billeter, M. (1988). Biased hypermutation and other genetic changes in defective measles viruses in human brain infections. Cell, 55, 255-265.

9. Cavalli-Sforza, L., Edwards, A. (1964) Analysis of Human Evolution. 11th International Conference of Genetics.

10. Choe, I. S., Melcher, U., Richards, K., Lebeurier, G., Essenberg, R. C. (1985). Recombination between mutant cauliflower mosaic virus DNAs. Plant Molecular Biology, 5, 281-289.

11. Covey, S. N. (1985). Organization and expression of the cauliflower mosaic virus genome. In: Molecular Plant Virology. (121-159) CRC Press, Boca Raton, Fla.

12. Covey, S. N. (1991). Pathogenesis of a plant pararetrovirus: CaMV. Seminars in Virology, 2, 151-159.

13. Daubert, S., Routh, G. (1990). Point mutations in cauliflower mosaic virus gene VI confer host-specific symptom changes. Molecular Plant-Microbe Interactions, 3, 341-345.

14. Dawson, W. (1992). Tobamovirus-plant interactions. Virology, 186, 359-367.

15. Dixon, L., Nyffenegger, T., Delley, G., Martinez-Izquierdo, J., Hohn, T. (1986). Evidence for replicative recombination in cauliflower mosaic virus. Virology, 150, 463-468.

16. Domingo, E., Holland, J. (1987). High error rates, population equilibrium and evolution of RNA replication systems. In: Domingo, E., Ahlquist, P., Holland, J.(Eds.) RNA Genetics. Boca Raton, Fla.: CRC Press .

17. Donnis, R., Bean, W., Kawaoka, Y., Webster, R. (1989). Distinct lineages of influenza virus H4 hemagglutinin genes in different regions of the world. Virology, 169, 408-417.

18. Dykhuizen, D., Green, L. (1991). Recombination in Escherichia coli and the definition of biological species. Journal of Bacteriology, 173, 7257-7268.

19. Eck, R., Dayhoff, M. (1966). Atlas of Protein Sequence and Structure. Silver Springs, MD: National Biomedical Research Foundation.

20. Eigen, M., Gardiner, W., Schuster, P., Winkler-Oswatitsch, R. (1981). The origin of genetic information. Scientific American, 244, 88-118.

21. Farris, J. (1972). Estimating phylogenetic trees from distance matrices. American Naturalist, 106, 645-668.

22. Farris, J. (1977). On the phenetic approach to vertebrate classification. In: Hecht, M., Goody, P., Hecht, B. (Eds.) Major Patterns in Vertebrate Evolution. (823-850) New York: Plenum Press.

23. Felsenstein, J. (1973). Maximum likelihood and minimum steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology, 22, 240-249.

24. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution, 17, 368-376.

25. Felsenstein, J. (1983). Parsimony in systematics: biological and statistical issues. Annual Review of Ecology and Systematics, 14, 313-333.

26. Felsenstein, J. (1984). Distance methods for inferring phylogenies: a justification. Evolution, 38, 16-24.

27. Felsenstein, J. (1988). Phylogenies from molecular sequences: inferences and reliability. Annual Review of Genetics, 22, 521-565.

28. Felsenstein, J. PHYLIP. 1991 (unpublished).

29. Fitch, W. (1971). Toward defining the course of evolution: minimizing change for a specific tree topology. Systematic Zoology, 20, 406-416.

30. Fitch, W. (1977). On the problem of discovering the most parsimonious tree. American Naturalist, 3, 233-257.

31. Fitch, W., Margoliash, E. (1967). Construction of phylogenetic trees. Science, 155, 279-284.

32. Franck, A., Guilley, H., Jonard, G., Richards, K., Hirth, L. (1980). Nucleotide sequence of cauliflower mosaic virus DNA. Cell, 21, 285-294.

33. Gal, S., Pisan, B., Hohn, T., Grimsley, N., Hohn, B. (1991). Genomic homologous recombination in planta. Journal of the European Molecular Biology Organization, 10, 1571-1578.

34. Gal, S., Pisan, B., Hohn, T., Grimsley, N., Hohn, B. (1992). Agroinfection of transgenic plants leads to viable cauliflower mosaic virus by intermolecular recombination. Virology, 187, 525-533.

35. Gardner, C. O., Jr., Melcher, U., Shockey, M. W., Essenberg, R. C. (1980). Restriction enzyme cleavage maps of the DNA of two cauliflower mosaic virus isolates. Virology, 103, 250-254.

36. Gardner, R. C., Howarth, A. J., Hahn, P., Brown-Luedi, M., Shepherd, R. J., Messing, J. (1981). The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic Acids Research, 9, 2871-2888.

37. Geldreich, A., Lebeurier, G., Hirth, L. (1986). In vivo dimerization of cauliflower mosaic virus DNA can explain recombination. Gene, 48, 277-286.

38. Girones, R., Miller, R. (1989). Mutation rate of the hepadnavirus genome. Virology, 170, 595-597.

39. Gojobori, T., (1990). Molecular clock of viral evolution, and the neutral theory. Proceedings of the National Academy of Sciences, USA, 87, 10015-10018.

40. Gojobori, T., Yokoyama, S. (1985). Rates of evolution of the retroviral oncogene of Maloney murine sarcoma virus and of its cellular homologues. Proceedings of the National Academy of Sciences, USA, 82, 4198-4201.

41. Goldbach, R. W. (1986). Molecular evolution of plant RNA viruses. Annual Review of Phytopathology, 24, 289-310.

42. Grimsley, N., Hohn, B., Hohn, T., Walden, R. (1986). "Agroinfection", an alternative route for viral infection of plants by using the Ti plasmid. Proceedings of the National Academy of Sciences, USA, 83, 3283-3286.

43. Grimsley, N., Hohn, T., Hohn, B. (1986). Recombination in a plant virus: template-switching in cauliflower mosaic virus. Journal of the European Molecular Biology Organization, 5, 641-646.

44. Guilley, H., Richards, K. E., Jonard, G. (1983). Observations concerning the discontinuous DNAs of cauliflower mosaic virus. Journal of the European Molecular Biology Organization, 2, 277-282.

45. Hahn, B., Shaw, G., Taylor, M., Redfield, R., Markham, P. (1986). Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science, 232, 1548-1553.

46. Hamori, E. (1989). Graphic representation of long DNA-sequences by the method of H-curves: current results and future aspects. BioTechniques, 7, 710-720.

47. Hasegawa, A., Verver, J., Shimada, A., Saito, M., Goldbach, R., Van Kammen, A., Miki, K., Kameya-Iwaki, M., Hibi, T. (1989). The complete sequence of soybean chlorotic mottle virus DNA and the identification of a novel promoter. Nucleic Acids Research, 17, 9993-10013.

48. Hirochika, H., Takatsuji, H., Ubasawa, A., Ikeda, J.-E. (1985). Site-specific deletion in cauliflower mosaic virus DNA: possible involvement of RNA splicing and reverse transcription. Journal of the European Molecular Biology Organization, 4, 1673-1680.

49. Hohn, B., Balazs, E., Ruegg, D., Hohn, T. (1986). Splicing of an intervening sequence from hybrid cauliflower mosaic viral RNA. Journal of the European Molecular Biology Organization, 5, 2759-2762.

50. Hohn, T., Futterer, J. (1991). Pararetroviruses and retroviruses: a comparison of expression strategies. Seminars in Virology, 2, 55-69.

51. Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., Vandepol, S. (1982). Rapid evolution of RNA genomes. Science, **215**, 1577-1585.

52. Howarth, A., Vandemark, G. (1989). Phylogeny of geminiviruses. Journal of General Virology, 70, 2717-2727.

53. Howarth, A. J., Gardner, R. C., Messing, J., Shepherd, R. J. (1981). Nucleotide sequence of naturally occurring deletion mutants of cauliflower mosaic virus. Virology, 112, 678-685.

54. Howell, S. H. (1981). Ultraviolet mapping of RNA transcripts encoded by the cauliflower mosaic virus genome. Virology, 112, 488-495.

55. Howell, S. H., Walden, R. M., Marco, Y. (1983). Recombination and replication of cauliflower mosaic virus DNA. In R.B. Goldberg (Ed.):Plant Molecular Biology, 137-146. New York: A. R. Liss.

56. Howell, S. H., Walker, L. L., Walden, R. M. (1981). Rescue of in vitro generated mutants of cloned cauliflower mosaic virus genomes in infected plants. Nature, 293, 483-486.

57. Hu, W. S., Temin, H. M. (1990). Retroviral recombination and reverse transcription. Science, 250, 1227-1233.

58. Hull, R. (1980). Structure of the cauliflower mosaic virus genome III. Restriction endonuclease mapping of thirty-three isolates. Virology, 100, 76-90.

59. Hull, R., Covey, S. N. (1983). Does cauliflower mosaic virus replicate by reverse transcription? Trends in Biochemical Sciences, 8, 119-121.

60. Jukes, T., Cantor, C. (1969) Evolution of protein molecules. In: Munro, H. (Ed.). Mammalian Protein Metabolism. (21-132) New York: Academic Press.

61. Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16, 111-120.

62. Kishino, H., Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data and the branching order in Hominoidea. Journal of Molecular Evolution, 29, 170-179.

63. Kunkel, T., Soni, A. (1988). Mutagenesis by transient misalignment. Journal of Biological Chemistry, 263, 14784-14789.

64. Langley, C., Fitch, W. (1974). An examination of the constancy of the rate of molecular evolution. Journal of Molecular Evolution, 3, 161-177.

65. Lebeurier, G., Hirth, L., Hohn, B., Hohn, T. (1982). In vivo recombination of cauliflower mosaic virus DNA. Proceedings of the National Academy of Sciences , USA, 79, 2932-2936.

66. Leisner, S. M., Turgeon, R., Howell, S. H. (1992). Long distance movement of cauliflower mosaic virus in infected turnip plants. Molecular Plant-Microbe Interactions, 5, 41-47.

67. Li, W., Tanimura, M., Sharp, P. (1988). Rates and dates of divergence between AIDS virus nucleotide sequences. Molecular Biology and Evolution, 5, 313-330.

68. Lung, M. C. Y., Pirone, T. P. (1973). Datura stramonium, a local lesion host for certain isolates of cauliflower mosaic virus. Phytopathology, 62, 1473-1474.

69. Marsh, L., Kuzj, A., Guilfoyle, T. (1985). Identification and characterization of cauliflower mosaic virus replication complexes--analogy to hepatitis B viruses. Virology, 143, 212-223.

70. Matthews, R. (1991). Plant Virology. (third ed.). New York: Academic Press.

71. Melcher, U. (1988). A readable and space-efficient DNA sequence representation: application to caulimoviral DNAs. Computer Applications in the Biosciences, 4, 93-96.

72. Melcher, U. (1989). Symptoms of cauliflower mosaic virus infection in Arabidopsis thaliana and turnip. Botanical Gazette, 150, 139-147.

73. Melcher, U. (1990). Similarities between putative transport proteins of plant viruses. Journal of General Virology, 71, 1009-1018.

74. Melcher, U. MalSig 1992 (unpublished).

75. Melcher, U., Choe, I. S., Lebeurier, G., Richards, K., Essenberg, R. C. (1986). Selective allele loss and interference between cauliflower mosaic virus DNAs. Molecular and General Genetics, 203, 230-236.

76. Nei, M. (1987). Molecular Evolutionary Genetics. New York: Columbia University Press.

77. Pathak, K., Temin, H. (1992). 5-Azacytidine and RNA secondary structure increase the retrovirus mutation rate. Journal of Virology, 66, 3093-3100.

78. Pennington, R. (1991) In Planta Deletion of DNA Inserts from the Large Intergenic Region of Cauliflower Mosaic Virus DNA. Doctoral thesis. Oklahoma State University.

79. Penswick, J., Huebler, R., Hohn, T. (1988). A viable mutation in cauliflower mosaic virus, a retroviruslike plant virus, separates its capsid protein and polymerase genes. Journal of Virology, 62, 1460-1463.

80. Pickover, C. A. (1992). DNA and protein tetragrams: Biological sequences as tetrahedral movements. Journal of Molecular Graphics, 10, 2-6.

81. Poch, O., de Marcillac, G. D., Exinge, F., Roy, A., Losson, R. (1988). Functional domains of the regulatory protein PPR1: use of the V. R. P. computer program. Yeast, 4, S416.

82. Purseglove, J. (1969). Tropical Crops: Dicotyledons. London: Longman Group Limited.

83. Richins, R. D., Scholthof, H. B., Shepherd, R. J. (1987). Sequence of figwort mosaic virus DNA (caulimovirus group). Nucleic Acids Research, 15, 8451-8466.

84. Roberts, J., Preston, B., Johnston, L., Soni, A., Loeb, L., Kunkel, T. (1989). Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. Molecular and Cellular Biology, 9, 469-476.

85. Rongxiang, F., Xiaojun, W., Ming, B., Yingchuan, T., Faxing, C., Kequiang, M. (1985). Complete nucleotide sequence of cauliflower mosaic virus (Xinjing isolate) genomic DNA. Chinese Journal of Virology, 1, 247-256.

86. Sanger, F., Nicklen, S., Coulson, R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, USA, 74, 5463-5467.

87. Sanger, M., Daubert, S., Goodman, R. M. (1991). The regions of sequence variation in caulimovirus gene VI. Virology, 182, 830-834.

88. Sawyer, S. (1989). Statistical tests for detecting gene conversion. Molecular Biology and Evolution, 6, 526-538.

89. Schoelz, J., Shepherd, R. J., Daubert, S. (1986). Region VI of cauliflower mosaic virus encodes a host range determinant. Molecular and Cellular Biology, 6, 2632-2637.

90. Schoelz, J. E., Shepherd, R. J., Daubert, S. D. (1987) Host response to cauliflower mosaic virus (CaMV) in solanaceous plants is determined by a 496 bp DNA sequence within gene VI. In: Molecular Strategies for Crop Protection. (253-265) Alan R. Liss .

91. Shepherd, R. J. (1989) Biochemistry of DNA Plant Viruses. In A. Marcus (Ed.): The Biochemistry of Plants. (563-616) New York: Academic Press, Inc.

92. Shepherd, R. J., Bruening, G. E., Wakeman, R. J. (1970). Double-stranded DNA from cauliflower mosaic virus. Virology, 41, 339-347.

93. Shimizu, N., Okamoto, T., Moriyama, E., Takeuchi, Y., Gojobori, T., Hoshino, H. (1989). Patterns of nucleotide substitutions and implications for the immunological diversity of human immunodeficiency virus. FEBS Letters, 250, 591-595.

94. Sober, E. (1983). A likelihood justification of parsimony. Cladistics, 1, 209-233.

95. Sokal, R., Sneath, P. (1963). Principles of Numerical Taxonomy. San Francisco: Freeman.

96. Steinhauer, D. A., Holland, J. J. (1987). Rapid evolution of RNA viruses. Annual Review of Microbiology, 41, 409-433.

97. Stenger, D. C., Morris, T. J., Mullin, R. H. (1986). Molecular cloning and analysis of strawberry vein banding virus DNA. Phytopathology, 76, 154-159.

98. Stratford, R., Covey, S. N. (1989). Segregation of cauliflower mosaic virus symptom genetic determinants. Virology, 172, 451-459.

99. Tateno, Y., Nei, M., Tajima, F. (1982). Accuracy of estimated phylogenetic trees from molecular data I. Distantly related species. Journal of Molecular Evolution, 18, 387-404.

100. Thompson, E. (1975). Human Evolutionary Trees. Cambridge, Mass.: Cambridge University Press.

101. Thorne, J., Kishino, H., Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. Journal of Molecular Evolution, 33, 114-124.

102. Thorne, J., Kishino, H., Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. Journal of Molecular Evolution, 34, 3-16.

103. Tomlinson, J. A., Shepherd, R. J. (1978). Studies on mutagenesis and cross protection of cauliflower mosaic virus. Annals of Applied Biology, 90, 223-231.

104. Topal, M., Fresco, J. (1976). Complementary base pairing and the origin of substitution mutations. Nature, 263, 285-289.

105. Vaden, V. R., Melcher, U. (1990). Recombination sites in cauliflower mosaic virus DNAs: implications for mechanisms of recombination. Virology, 177, 717-726.

106. Vartanian, J. P., Meyerhans, A., Åsjö, B., Wain, H. S. (1991). Selection, recombination, and G-->A hypermutation of human immunodeficiency virus type 1 genomes. Journal of Virology, 65, 1779-1788.

107. Walden, R., Howell, S. (1982). Intergenomic recombination events among pairs of defective cauliflower mosaic virus genomes. Journal of Molecular and Applied Genetics, 1, 447-456.

108. Walden, R. M., Howell, S. H. (1983). Uncut recombinant plasmids bearing nested cauliflower mosaic virus genomes infect plants by intragenomic recombination. Plant Molecular Biology, 2, 27-31.

109. Williams, A., Chenault, K. D., Melcher, U. M. (in press). Kitty-a space-efficient representation of amino acid sequences of proteins. In C.A. Pickover (Ed.): The Visual Display of Biological Information.Teaneck, New Jersey: World Scientific.

110. Woolston, C. J., Covey, S. N., Penswick, J. R., Davies, J. W. (1983). Aphid transmission and a polypeptide are specified by a defined region of the cauliflower mosaic virus genome. Gene, 23, 15-21.

111. Zhang, X. S., Melcher, U. (1989). Competition between isolates and variants of cauliflower mosaic virus in infected turnip plants. Journal of General Virology, 70, 3427-3437.

112. Zuckerkandl, E., Pauling, L. (1962). Horizons in Biochemistry. New York: Academic Press.

APPENDIXES

APPENDIX A

METHODS OF INFERRING AND CONSTRUCTING PHYLOGENETIC TREES

The field of molecular evolution was dramatically changed by the onset of extensive sequencing of nucleic acids and proteins. Sequences of homologous molecules from different organisms provide useful data for examination of the relationships between these organisms. The amount and accessibility of this type of data is rising rapidly. Such an abundance of molecular data enables both the elucidation of an evolutionary history of a set of organisms and the inference of the mechanisms behind evolution. One important event in the study of molecular evolution was the suggestion of approximate constancy of the rate of nucleic acid substitution. Zuckerkandl and Pauling (112) first introduced this 'molecular clock' concept, which significantly reduces the number of variables to be considered when comparing data from diverse organisms. Although it is now known that rates of change in different genes and lineages may vary (70), the assumption of independent but constant evolutionary change is central to most methods developed for constructing phylogenetic trees (28, 76).

Evolutionists are interested in a phylogenetic tree which depicts the evolutionary pathway of a certain group of organisms. Several types of data may be used to construct phylogenetic trees, including gene frequencies, restriction enzyme sites, and molecular sequences (nucleotide or amino acid). When using molecular sequence data, a method may require the whole sequence or only the informative sites within that sequence. A site is informative only when there are at least two different kinds of residues, each represented at least two times.

Most computer programs that can be used to construct phylogenetic trees require that the sequences being analyzed are aligned in a reliable manner. The program UMalign written by Melcher (73) was used in the work described in this thesis to align both nucleic acid and amino acid sequences. This program allows the insertion of 'gaps' in individual or sets of sequences in order to achieve alignment. Insertion of gaps at the proper location by visual inspection is possible and easily done for CaMV DNAs since the isolates vary only in 5% of their residues. Gap translation is also possible in UMalign. Using this option, a gap is inserted before the region where it is expected to belong and then a residue comparison matrix is used to calculate a similarity value. The similarity value is adjusted as the gap is moved one position at a time for a specified distance. The gap is finally positioned in the alignment at the location which gave the highest similarity

value. The MacVector™ program for sequence analysis was also used to align sequences for the work in this thesis.

Each species considered in the construction of a tree is termed an operational taxonomic unit (OTU). One type of tree is termed a 'species' or 'population' tree, and the data from which it was constructed represent the entire genomes of the species involved. The species tree represents the amount of change that has occurred between the OTUs since the time they were considered the same species. Another type of phylogenetic tree may be constructed using the same gene from each OTU. Gene trees (76, 99), as they are termed, may differ in branching order from a corresponding species tree, especially if recombination between genomes has occurred.

The branching pattern of a tree is called its 'topology'. Trees may be constructed as 'rooted', which implies a known common ancestor, or 'unrooted' where that ancestor is unclear. The number of possible trees for a given set of OTUs varies, depending on the size of the data set. It is a very difficult task to find the best phylogenetic tree from observed sequence data. Several different methods have been developed to accomplish this task. There are three major classes of methods for inferring phylogenetic trees: (1) parsimony, (2) distance, and (3) maximum likelihood.

The parsimony method was first introduced by Edwards and Cavalli-Sforza (9) who called it the 'method of minimum net

evolution'. Eck and Dayhoff (19) first described the method's application to molecular sequences of nucleic acids and the method was adapted for nucleic acid sequences by Fitch (29, 30). The principle of this method is to infer the nucleic acid sequence of the ancestral species and then choose a tree that requires the minimum number of mutational changes. This tree would then be termed the 'most parsimonious tree'. The parsimony method is generally used to infer the topology of a tree, not branch length. When using the parsimony method, only the informative sites in the OTU sequences are needed. The assumptions of the parsimony method have been extensively reviewed by Felsenstein (23, 24, 25, 26, 27).
Taken from the PHYLIP manual (28), these assumptions are:

1. Each site evolves independently.

2. Different lineages evolve independently.

3. The probability of a base substitution at a given site is small over the lengths of time involved in a branch of the phylogeny.

4. The expected amounts of change in different branches of the phylogeny do not vary by so much that two changes in a high-rate branch are more probable that one change in a low-rate branch.

5. The expected amounts of change do not vary enough among sites that two changes in one site are more probable that one change in another.

The first step in the parsimony algorithm involves finding a particular topology for a group of OTUs and inferring the ancestral sequence for that topology. The minimum number of changes required for that tree topology is then counted. The process continues for all reasonable topologies, and the one which requires the smallest number of changes is chosen as the final 'most parsimonious' tree. For a more detailed discussion of parsimony methods, see Sober (94) or Felsenstein (25). The parsimony computer program DNAPARS was used for the work in this thesis and was developed as part of the PHYLIP package for sequence analysis by Felsenstein (28).

The recently developed statistical method known as the 'bootstrap' can be used to place confidence intervals on phylogenies. It involves sampling points from observed data to create a series of 'bootstrap' samples of the same size as the original data. Some of the residue positions may be duplicated and some may be omitted. Each time this is done (one replicate) a tree is made for the bootstrap sample. The process continues until the number of specified replicates have been completed. At this point, a tree is drawn with numbers on each node, representing the number of times that node occurred during bootstrap sampling. When considering

the significance of evidence for the monophyly of a pre-conceived group of OTUs, a group is significant if it occurs in 95% or more of the samplings. If a group of OTUs is considered due to the fact that it arises during tree construction, Felsenstein recommends a more conservative estimate of considering a group significant if it occurs in 100-5/(N-2) % of the bootstrap replicates, where N specifies the total number of species being considered. The computer programs DNABOOT and SEQBOOT in the PHYLIP package use a random number generator to draw bootstrap samples from the data. Felsenstein recommends that at least 100 replicates are carried out on a given set of data (28).

Distance matrix methods use the computation of a genetic distance value for all pairs of OTUs. A phylogenetic tree is constructed by considering the relationships among these distance values. Branch lengths are estimated from the distance values which are calculated by methods based on one of three models of nucleotide substitution. All three of these models are available for use with the DNADIST program which is part of the PHYLIP package. The Jukes and Cantor (60) model assumes that there is independent change at all sites with equal probability. Whether a base changes or not is independent of identity, and the probability of changing to each of the other three bases is equal. These assumptions are unrealistic in most cases, since in general transitions are more frequent than transversions. Kimura (61) proposed a model to take this fact into account. In his model,

transitions are allowed to occur at a different rate than transversions. A third model incorporates different rates of transition and transversion and also allows for different frequencies of change for the four nucleotides (62). The DNADIST program generates a matrix of distance values (D) using a specified model. This data set can then be used to generate a phylogenetic tree using a distance matrix program. According to the PHYLIP manual(28), the assumptions made by these programs are:

1. Each distance is measured independently from the others: no item of data contributes to more than one distance.

2. The distance between each pair of taxa is drawn from a distribution with an expectation which is the sum of values along the tree from one tip to the other.

The simplest distance matrix method is the unweighted pair group method with arithmetic mean (UPGMA). Originally developed by Sokal and Michener (95), UPGMA examines the distance matrix to find the smallest distance between two OTUs, and clusters them together on a tree, with a branch point located at D/2, making the branch length leading to these two OTUs equal. Those two OTUs are then considered as one and the process continues by calculating a new distance between the combined OTU and the others. In computer

simulation, UPGMA reliably gives the true species tree, even when the substitution rate between OTUs varies slightly (76). However, when the substitution rate varies extensively between OTUs, UPGMA is likely to give an incorrect topology.

Fitch and Margoliash (31) developed a method which allows for this variability in substitution rate. Tree topology construction is similar to UPGMA, but Fitch and Margoliash consider three OTUs at one time. When there are more than three OTUs, the third OTU represents a composite of all other OTUs. Fitch and Margoliash's method allows for varying substitution rates between tree members.

Both UPGMA and Fitch and Margoliash's methods are available in the PHYLIP package using the NEIGHBOR and FITCH programs respectively. Other variations of distance matrix methods exist such as the transformed distance method (22) and the Wagner method (21).

Distance methods which infer evolutionary clocks have been developed (26, 27). The KITSCH program in the PHYLIP package applies a molecular clock to the Fitch and Margoliash method. This method assumes that all OTUs are contemporaneous and thus that their distances from a hypothetical common ancestor are equal. To estimate phylogeny under the assumption of a clock, one would try to find that phylogeny, having all tree tips contemporaneous, which minimizes the measure of goodness of fit.

The goodness of fit parameter may vary among methods. The distance matrix programs in PHYLIP produce two measures

of error for a tree: the sum of squares (SSQ) and the average percent standard deviation (APSD). The SSQ calculation is shown in equation (4) where D is the observed distance between species i and j, and d is the expected distance, computed as the sum of lengths of the segments of the tree between species i and j. The best tree will be the one with the least SSQ.

$$(4) \quad SSQ = \Sigma_i \Sigma_j (D_{obs} - d_{exp})^2 / D_{obs}^2$$

$$(5) \quad ASPD = (SSQ/N-2)^{1/2} \times 100$$

The calculation of APSD is shown in equation (5) where SSQ is the sum of squares and N is the number of OTUs. More information about distance matrix methods may be obtained from Nei (76).

The maximum likelihood method of tree making was first studied by Cavalli-Sforza and Edwards (9). Later, Felsenstein (23) and also Thompson (100) both developed algorithms for constructing a maximum likelihood tree by using and extending Cavalli-Sforza and Edward's approach. These methods were based on using gene frequency data, but Felsenstein (23, 24, 100) and also Langley and Fitch (64) modified the procedure to construct trees based on molecular sequence data. The algorithm used in the maximum likelihood method is intended to obtain both topology and branch lengths. In this method, the likelihood of obtaining the

observed nucleotide sequence for a group of OTUs is calculated for many different topologies, and the one which shows the highest ('maximum') likelihood is chosen as the best tree. The DNAML program in PHYLIP uses a maximum likelihood algorithm under the following assumptions stated in the PHYLIP manual(28):

1. Each site in the sequence evolves independently.

2. Different lineages evolve independently.

3. Each site undergoes substitution at an expected rate which may be specified.

4. All relevant sites are included in the sequence, not just informative sites.

The DNAML program estimates its own error. That is for each branch, an attempt is made to estimate its significance by placing an approximate confidence interval on the branch length. This is only a rough estimate, but indicates regions in the tree of definite uncertainty. More information on the maximum likelihood method may be obtained from Nei (76) or Thorne (101, 102).

APPENDIX B

ADDITIONAL FIGURES

Figure 9. Phylogenetic species tree constructed for eight CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

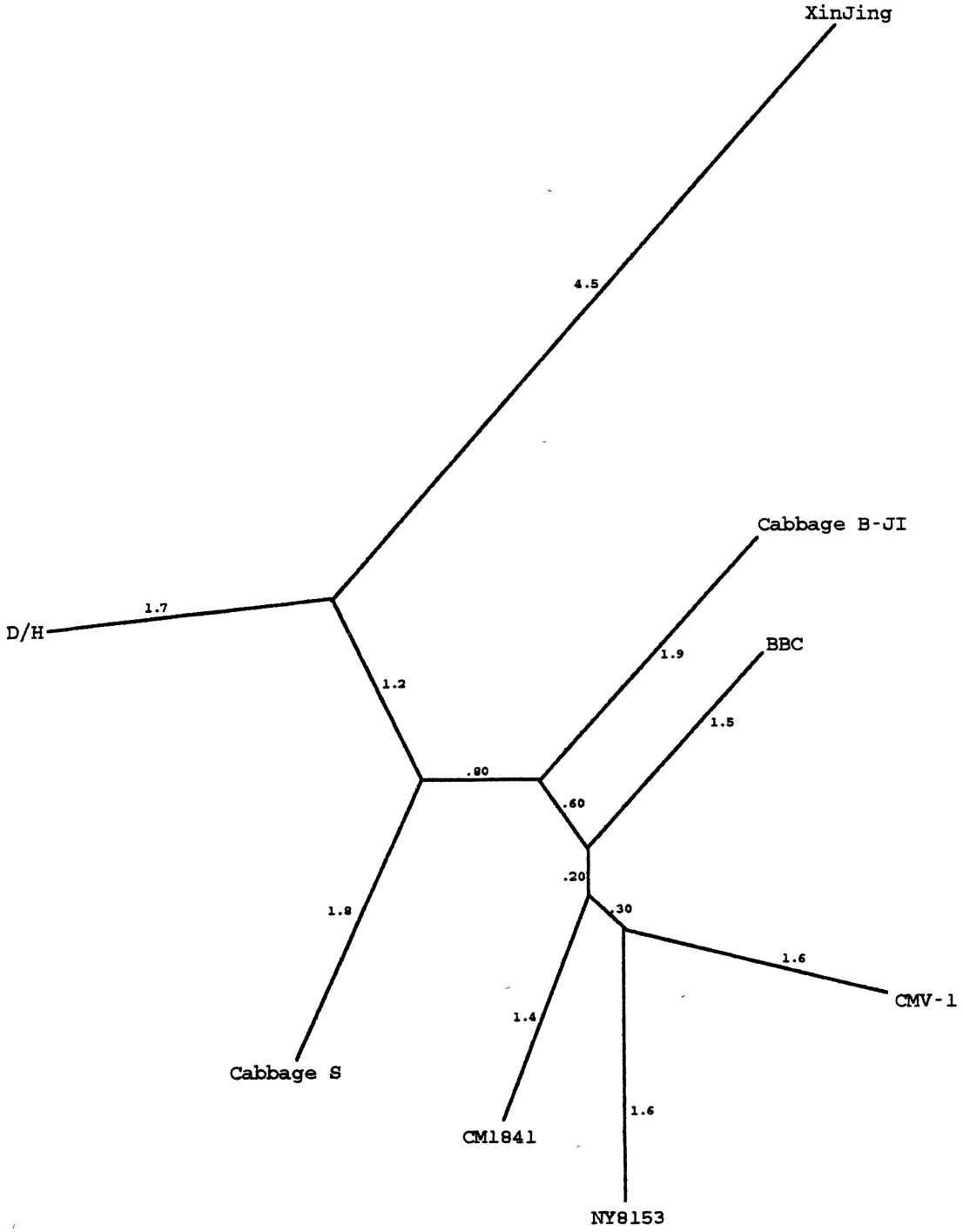

XinJing
4.5
Cabbage B-JI
1.9
D/H
1.7
1.2
BBC
1.5
.80
.60
.20
.30
1.8
1.6
CMV-1
1.4
Cabbage S
CM1841
1.6
NY8153

Figure 10. Phylogenetic species tree constructed for eight CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

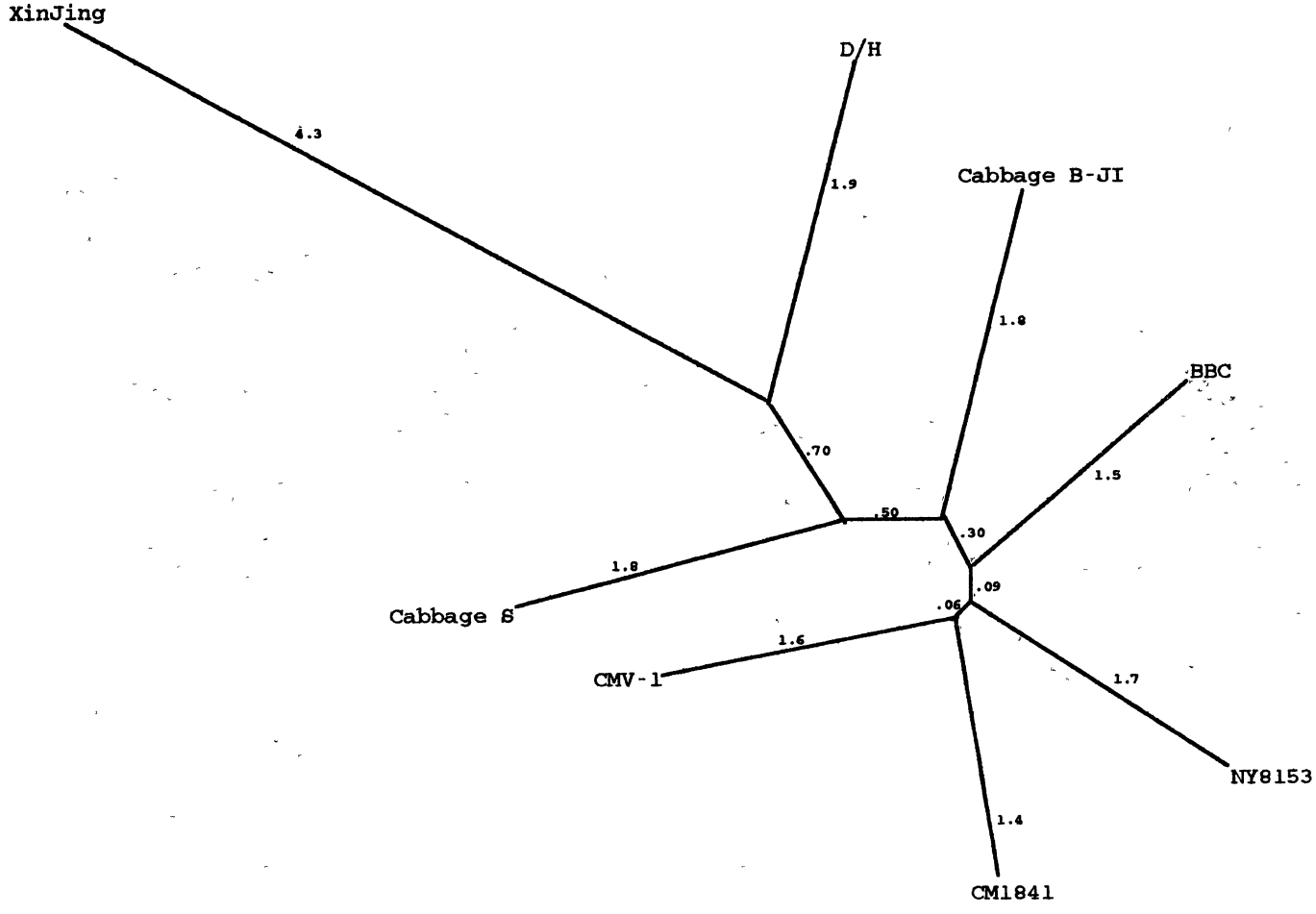

XinJing
4.3
D/H
1.9
Cabbage B-JI
1.8
BBC
1.5
.70
.50
.30
1.8
Cabbage S
.09
.06
1.6
CMV-1
1.7
NY8153
1.4
CM1841

Figure 11. Bootstrapped parsimony gene tree for ORF1 of nine CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

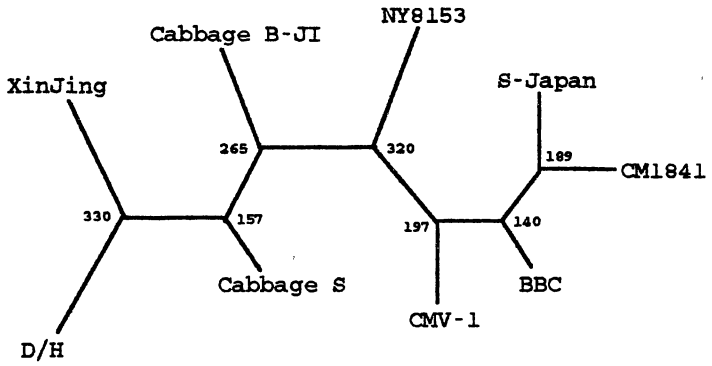

NY8153
Cabbage B-JI
XinJing
S-Japan
265
320
189
CM1841
330
157
197
140
Cabbage S
BBC
CMV-1
D/H

Figure 12. Phylogenetic gene tree for CaMV ORF1 constructed for nine CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

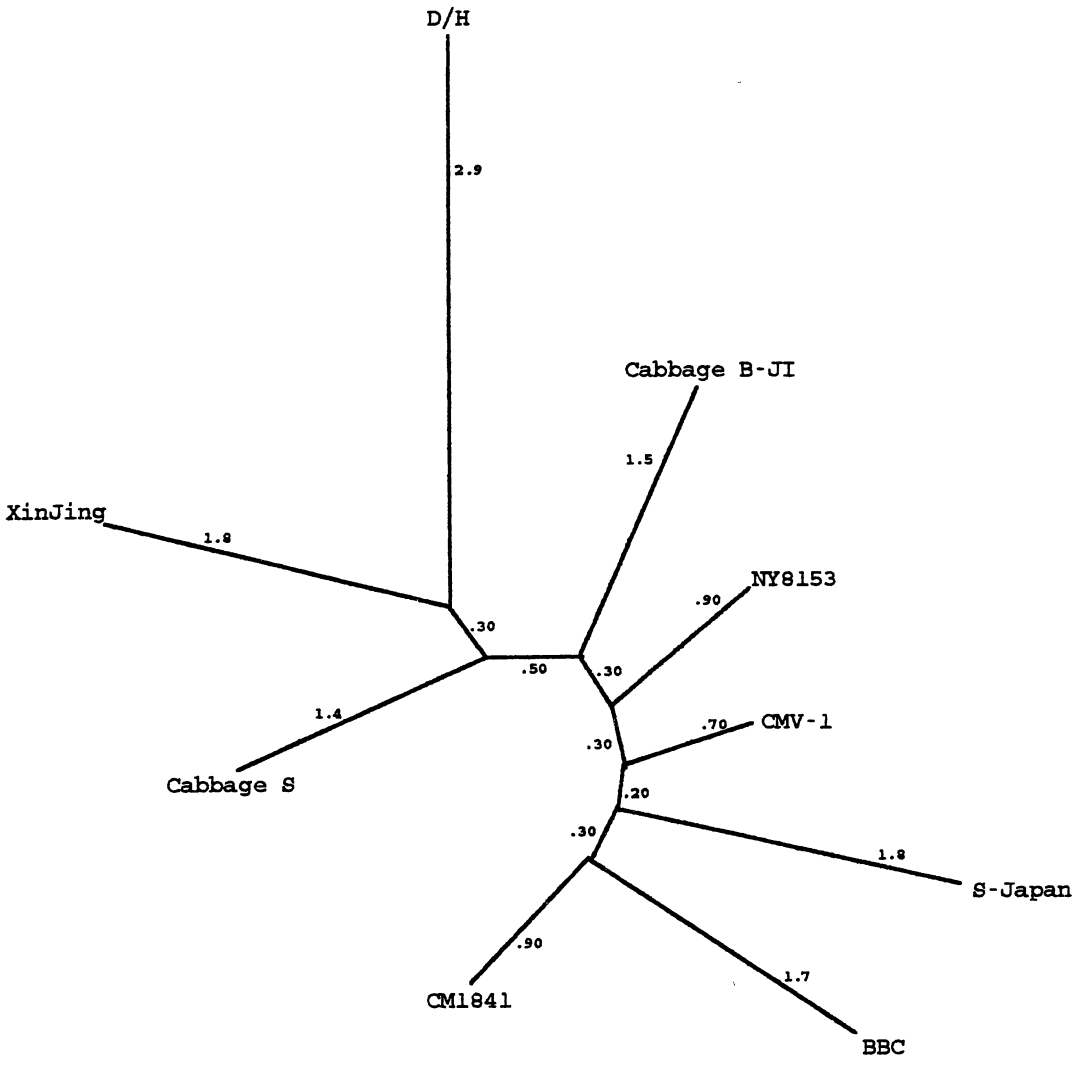

D/H
2.9
Cabbage B-JI
1.5
XinJing
1.8
NY8153
.90
.30
.50
.30
1.4
.70
CMV-1
.30
Cabbage S
.20
.30
1.8
S-Japan
.90
1.7
CM1841
BBC

Figure 13. Phylogenetic gene tree for CaMV ORF1 constructed for nine CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

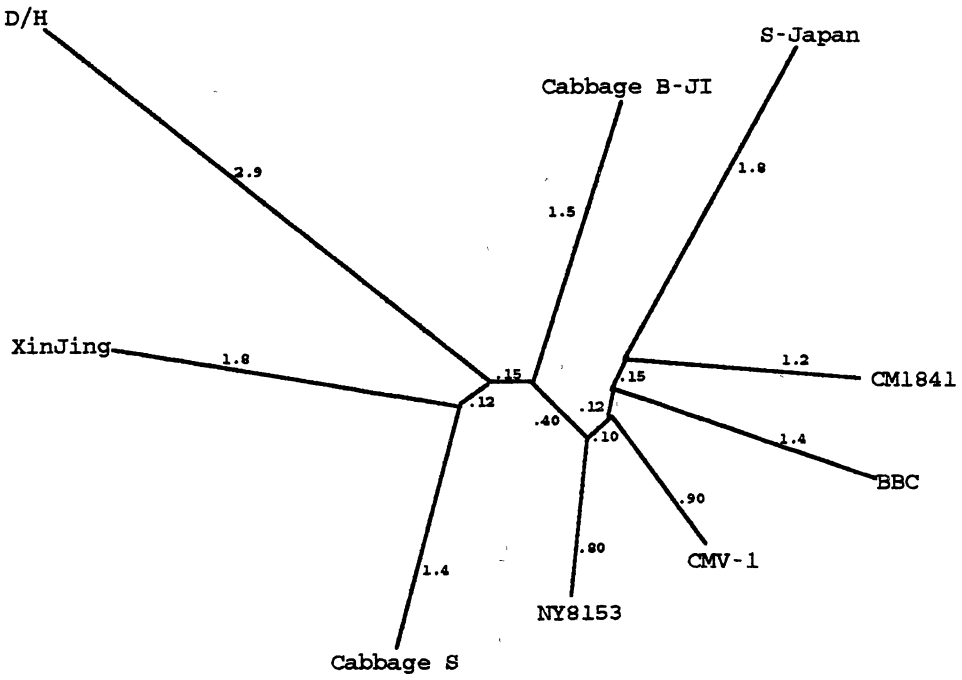

D/H
S-Japan
Cabbage B-JI
2.9
1.5
1.8
XinJing
1.8
.15
1.2
CM1841
.15
.12
.12
.40
.10
1.4
BBC
.90
.80
CMV-1
1.4
NY8153
Cabbage S

Figure 14. Phylogenetic gene tree for CaMV ORF2 constructed for ten CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

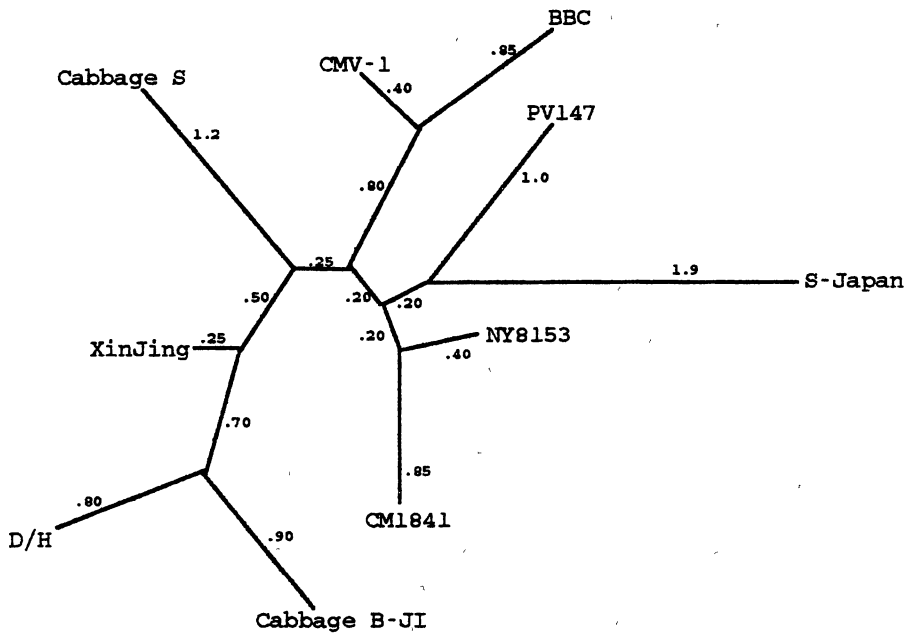

BBC
.85
CMV-1
.40
Cabbage S
1.2
PV147
1.0
.80
.25
1.9
S-Japan
.50
.20
.20
.20
NY8153
XinJing
.25
.40
.70
.85
.80
CM1841
D/H
.90
Cabbage B-JI

Figure 15. Phylogenetic gene tree for CaMV ORF2 constructed for ten CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

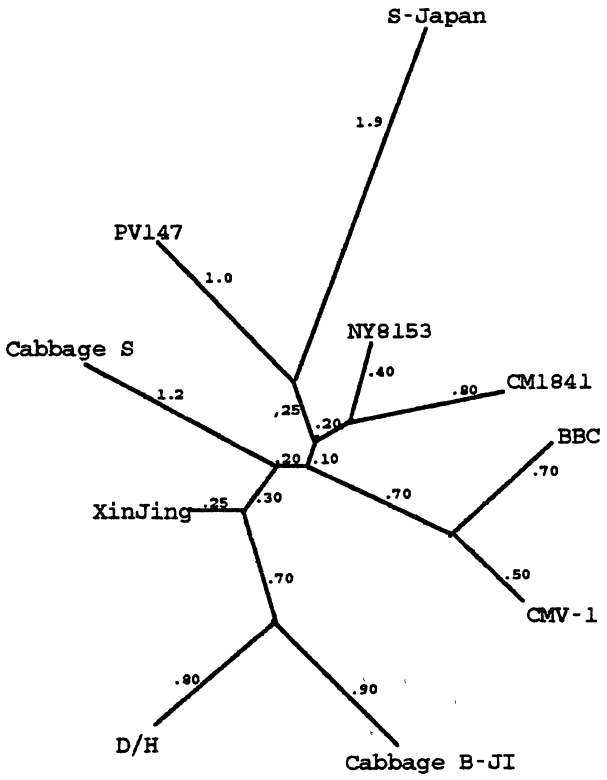

S-Japan
1.9
PV147
1.0
NY8153
.40
Cabbage S
1.2
.90
CM1841
.25
.20
BBC
.20
.10
.70
.70
XinJing
.25
.30
.50
CMV-1
.70
.80
.90
D/H
Cabbage B-JI

Figure 16. Bootstrapped parsimony gene tree for ORF3 of eight CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

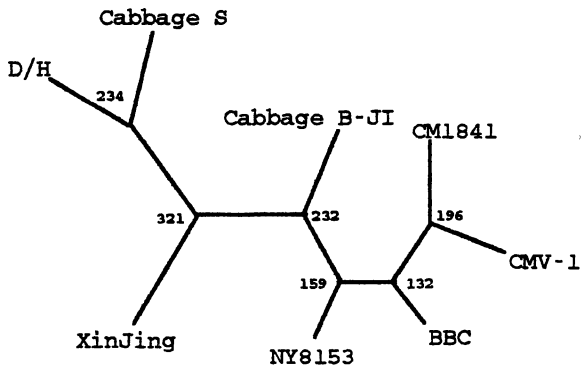

Cabbage S
D/H
234
Cabbage B-JI
CM1841
321
232
196
CMV-1
159
132
XinJing
NY8153
BBC

Figure 17. Phylogenetic gene tree for CaMV ORF3 constructed for eight CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

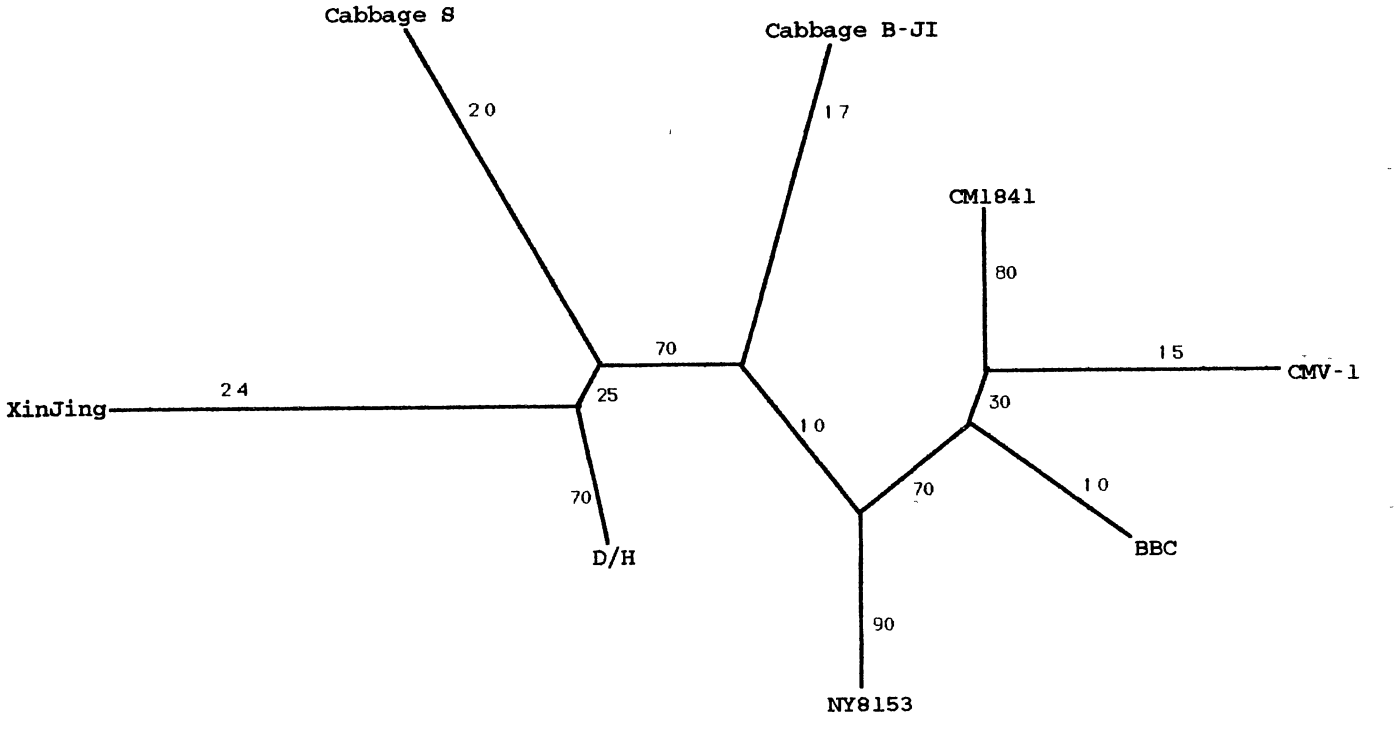

Cabbage S
Cabbage B-JI
20
17
CM1841
80
70
15
CMV-1
XinJing
24
25
30
10
70
10
70
D/H
BBC
90
NY8153

Figure 18. Phylogenetic gene tree for CaMV ORF3 constructed for eight CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

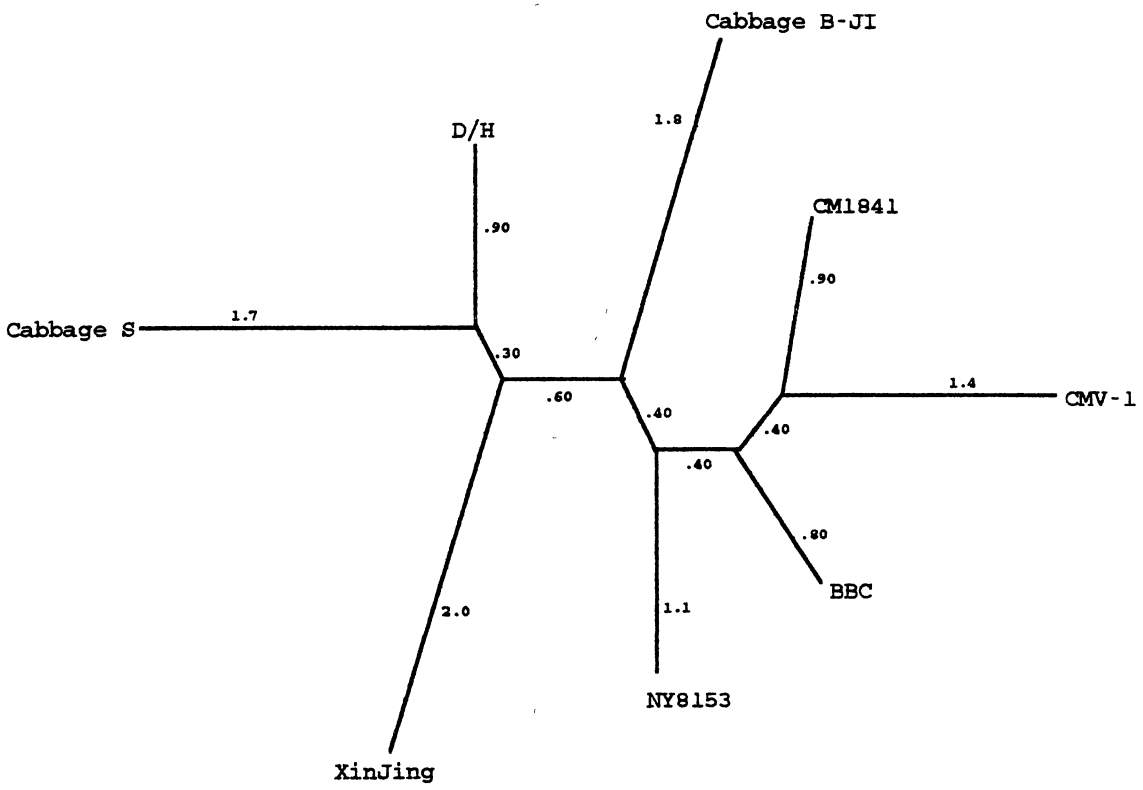

Cabbage B-JI
1.8
D/H
.90
CM1841
.90
Cabbage S
1.7
.30
1.4
CMV-1
.60
.40
.40
.40
.80
BBC
2.0
1.1
NY8153
XinJing

Figure 19. Bootstrapped parsimony gene tree for ORF4 of eight CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

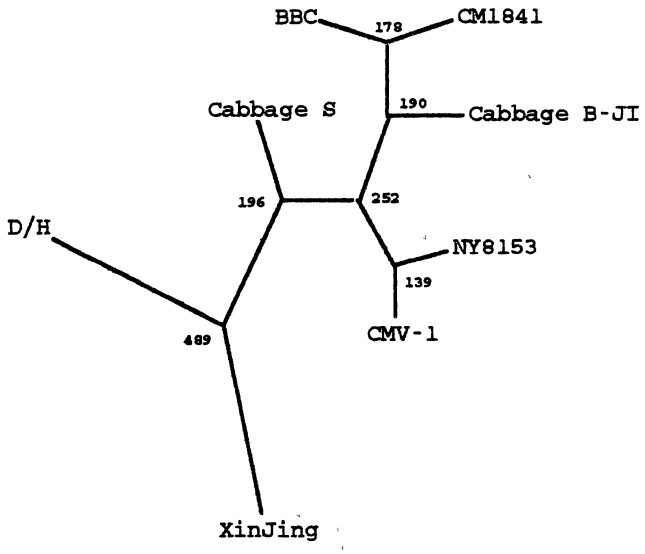

BBC
178
CM1841
Cabbage S
190
Cabbage B-JI
196
252
D/H
NY8153
139
CMV-1
489
XinJing

Figure 20. Phylogenetic gene tree for CaMV ORF4 constructed for eight CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

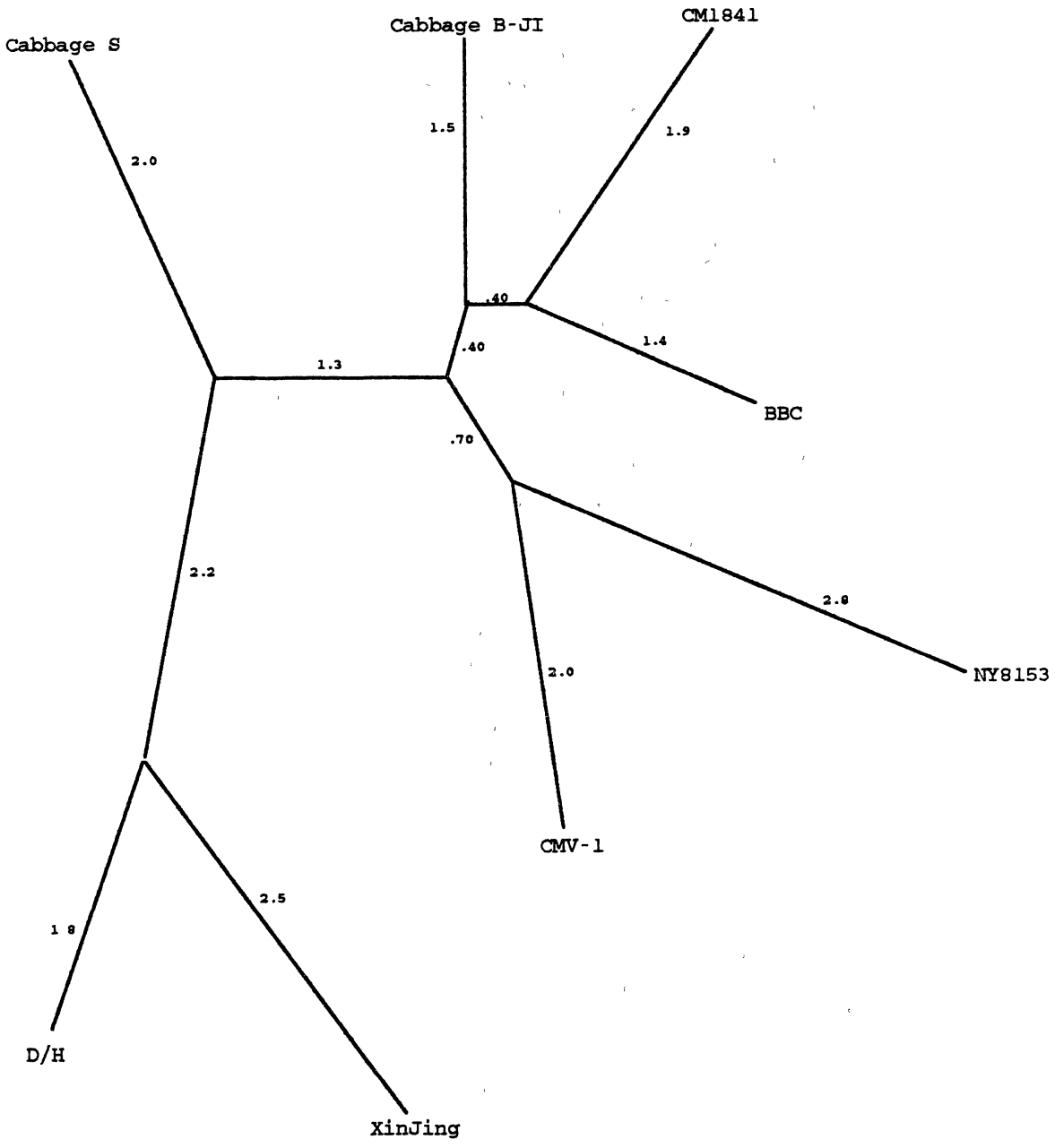

Cabbage S
Cabbage B-JI
CM1841
BBC
NY8153
CMV-1
D/H
XinJing
2.0
1.5
1.9
.40
1.4
.40
1.3
.70
2.2
2.8
2.0
2.5
1.8

Figure 21. Phylogenetic gene tree for CaMV ORF4 constructed for eight CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates

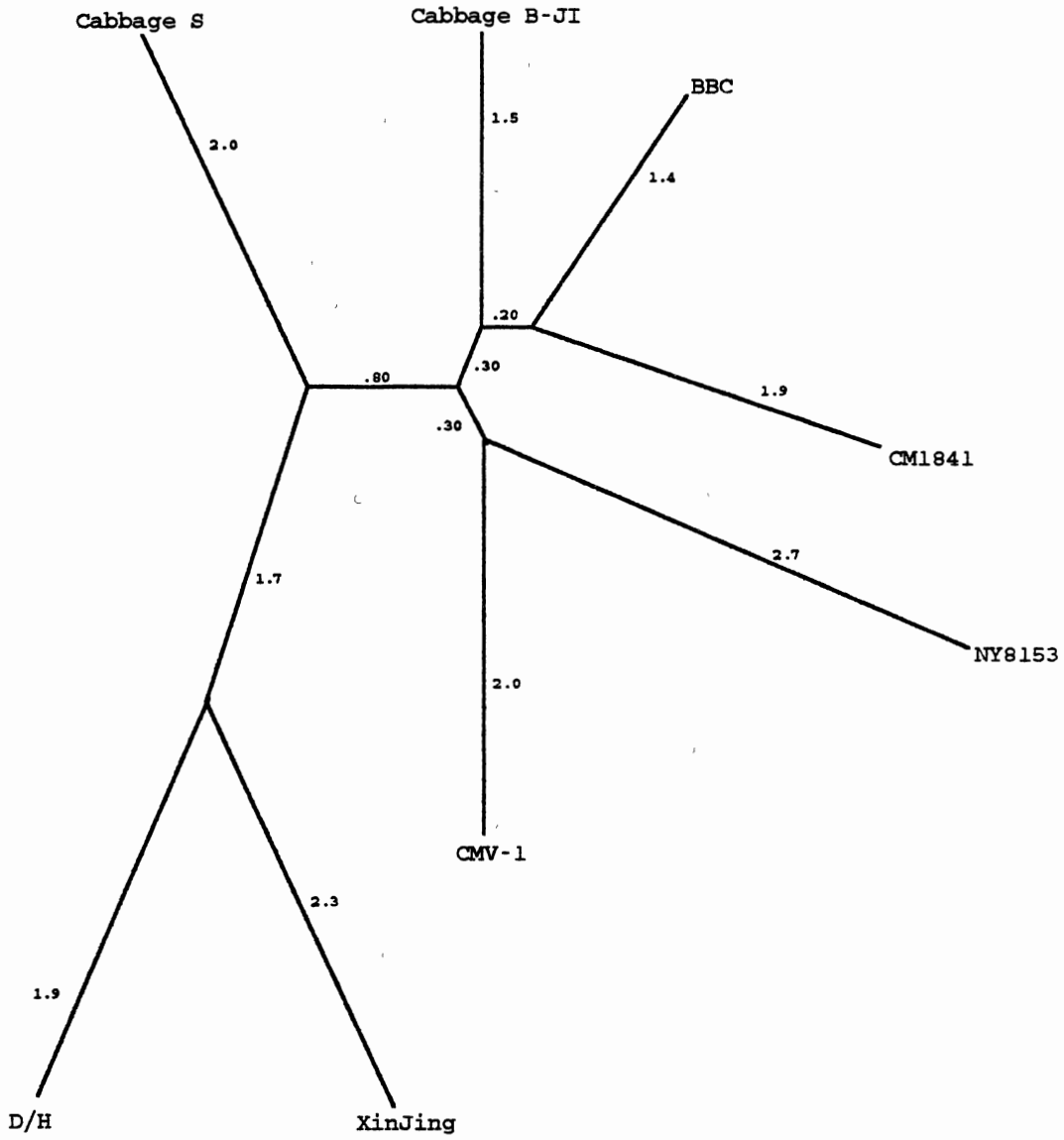

Cabbage S
Cabbage B-JI
BBC
CM1841
NY8153
CMV-1
D/H
XinJing
2.0
1.5
1.4
.20
.80
.30
1.9
.30
2.7
1.7
2.0
2.3
1.9

Figure 22. Bootstrapped parsimony gene tree for ORF5 of eight CaMV isolates. Numbers at each node indicate the bootstrap value for that node. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

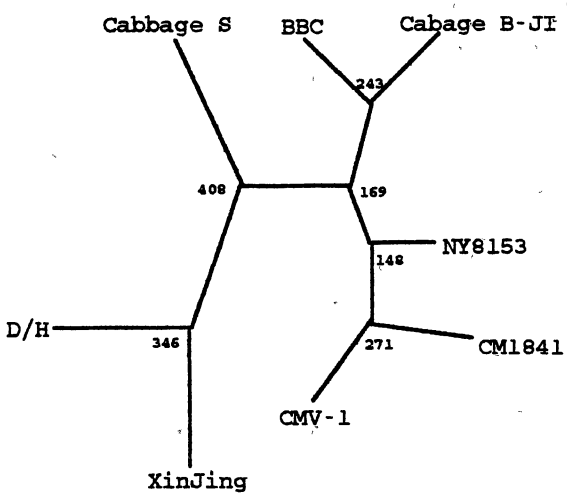

Cabbage S
BBC
Cabage B-JI
243
408
169
NY8153
148
D/H
346
271
CM1841
CMV-1
XinJing

Figure 23. Phylogenetic gene tree for CaMV ORF5 constructed for eight CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

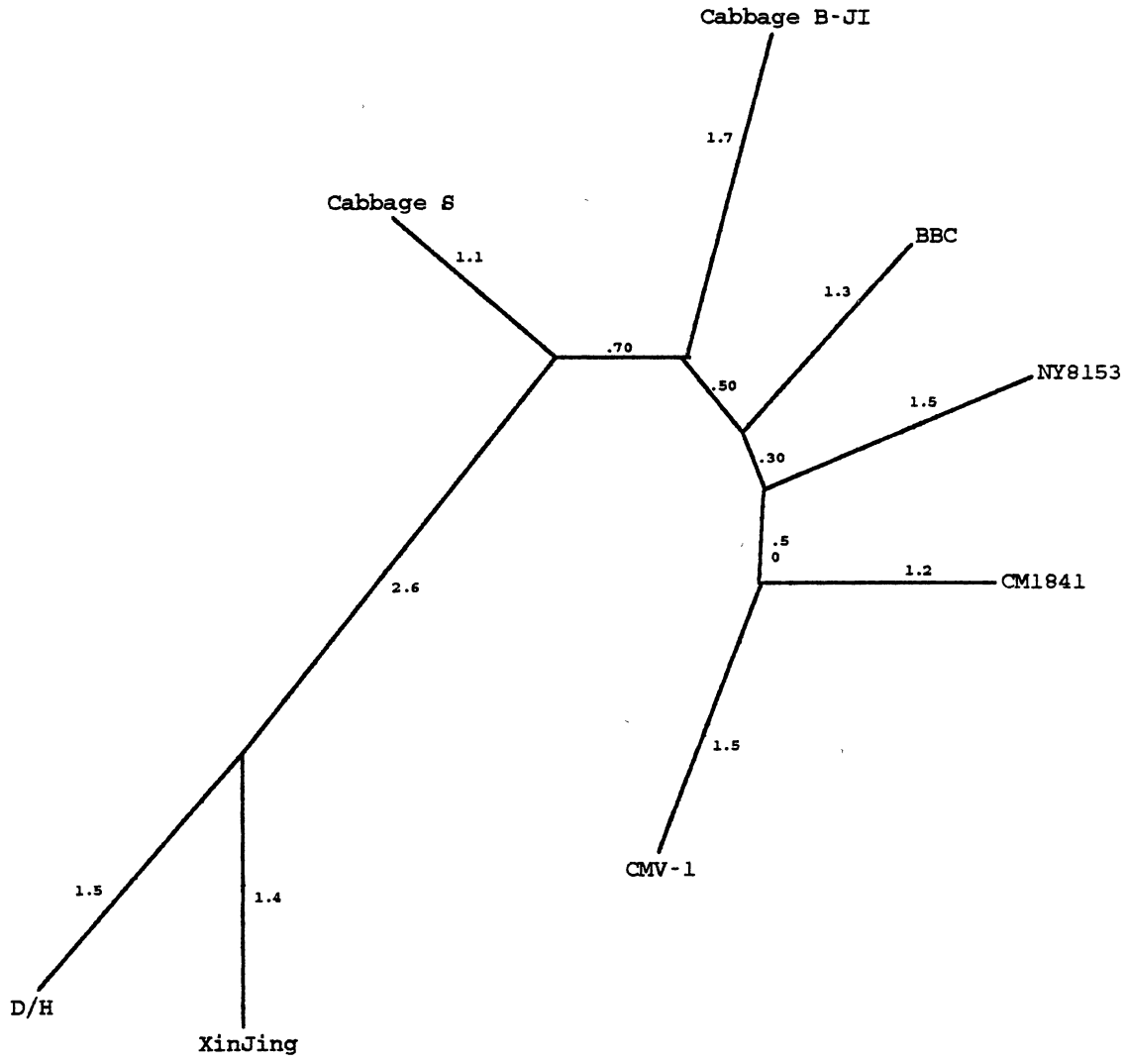

Cabbage B-JI
1.7
Cabbage S
1.1
BBC
1.3
.70
.50
NY8153
1.5
.30
.5
0
1.2
CM1841
2.6
1.5
CMV-1
1.5
1.4
D/H
XinJing

Figure 24. Phylogenetic gene tree for CaMV ORF5 constructed for eight CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

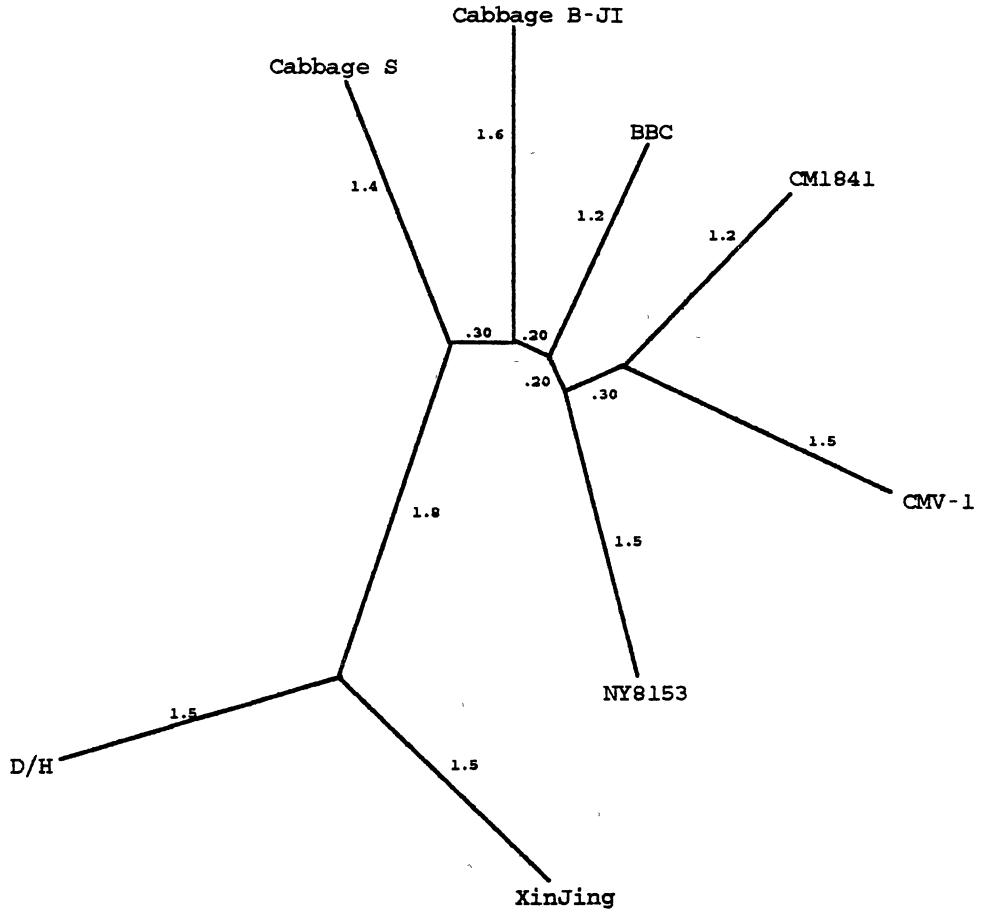

Cabbage B-JI
Cabbage S
BBC
CM1841
CMV-1
NY8153
XinJing
D/H
1.6
1.4
1.2
1.2
.30
.20
.20
.30
1.5
1.8
1.5
1.5
1.5

Figure 25. Phylogenetic gene tree for CaMV ORF6 constructed for eleven CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates. Branch lengths written as x10 are not drawn to scale.

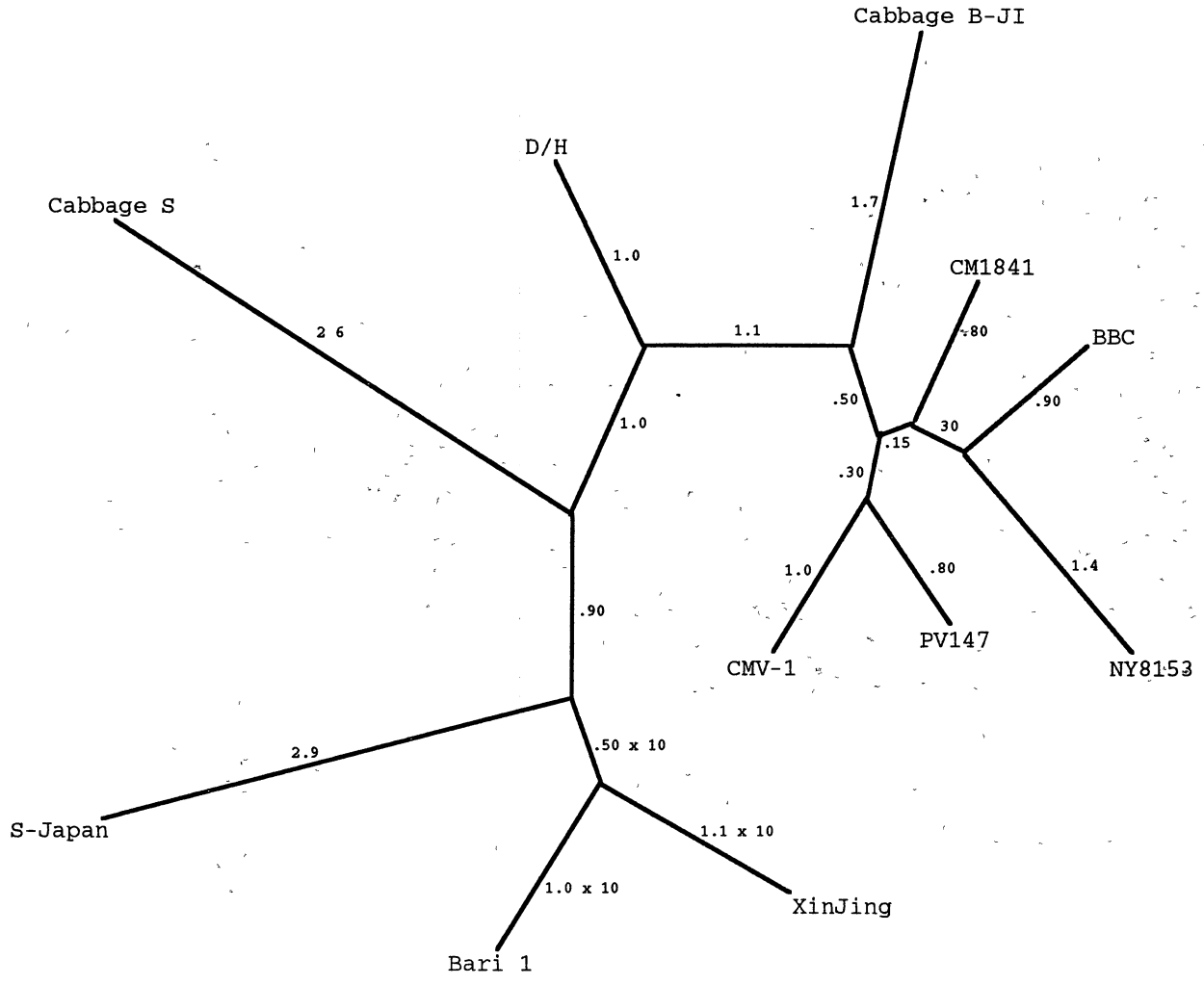

Cabbage B-JI
D/H
Cabbage S
1.7
1.0
CM1841
2 6
1.1
.80
BBC
.50
.90
1.0
30
.15
.30
1.0
.80
1.4
.90
PV147
CMV-1
NY8153
.50 x 10
2.9
S-Japan
1.1 x 10
1.0 x 10
XinJing
Bari 1

Figure 26. Phylogenetic gene tree for CaMV ORF6 constructed for eleven CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates. Branch lengths written as x10 are not drawn to scale.

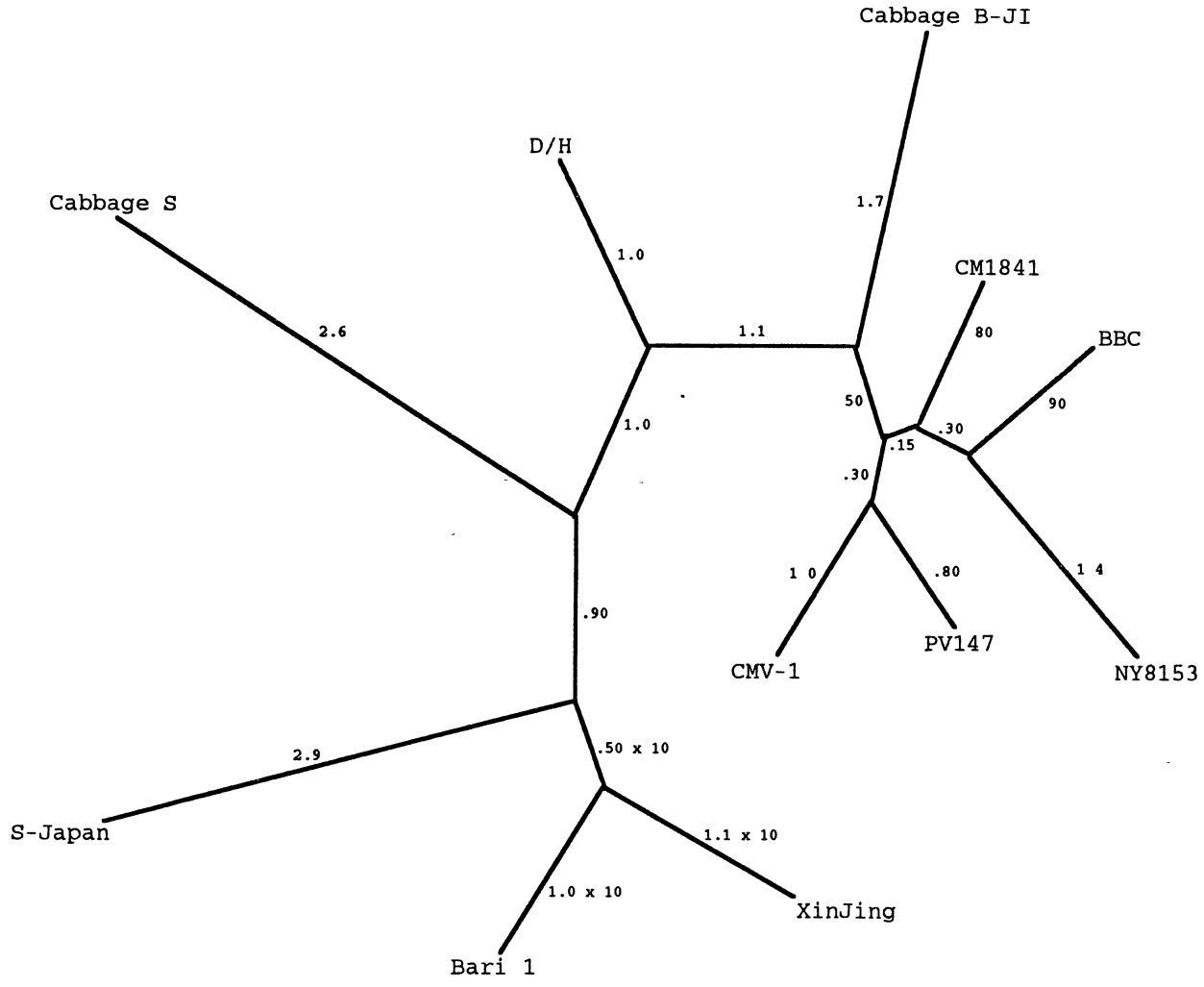

Cabbage B-JI
D/H
Cabbage S
1.7
1.0
CM1841
2.6
1.1
80
BBC
50
1.0
.30
90
.15
.30
1 0
.80
1 4
.90
CMV-1
PV147
NY8153
.50 x 10
2.9
S-Japan
1.1 x 10
1.0 x 10
XinJing
Bari 1

Figure 27. Bootstrapped parsimony tree for the large intergenic region of eleven CaMV isolates. Numbers at each node indicate the number of bootstrap replicates in which the corresponding node occurred. Branch lengths are proportionate to the sum of corresponding node bootstrap values and do not imply distance.

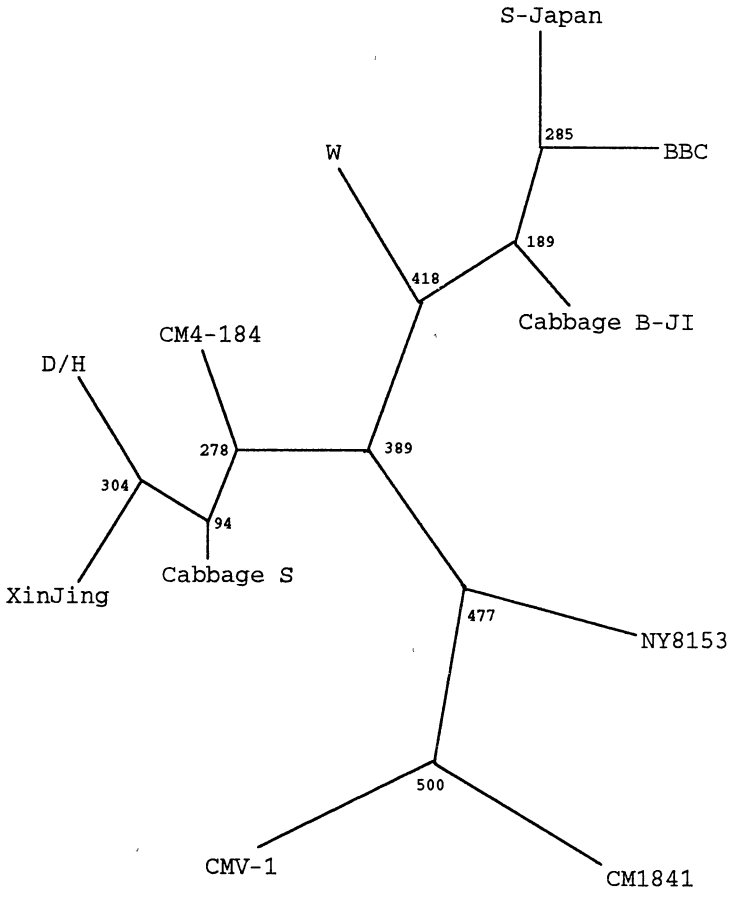

S-Japan
285
BBC
W
189
418
Cabbage B-JI
CM4-184
D/H
278
389
304
94
Cabbage S
XinJing
477
NY8153
500
CMV-1
CM1841

Figure 28. Phylogenetic tree for the large intergenic region of CaMV constructed for eleven CaMV isolates by the maximum likelihood method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

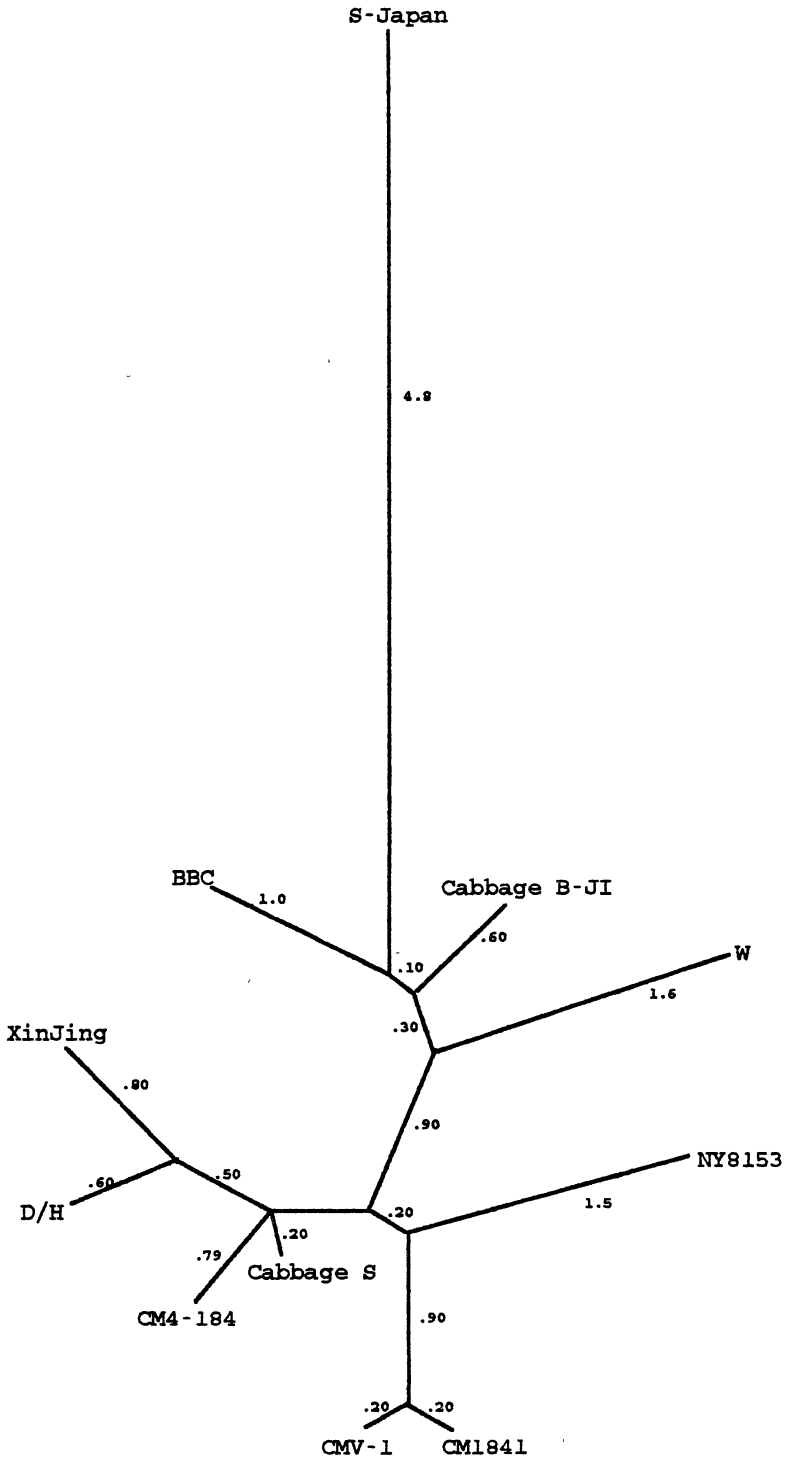

S-Japan
4.8
BBC
1.0
Cabbage B-JI
.60
.10
W
1.6
.30
XinJing
.80
.90
NY8153
.60
.50
D/H
1.5
.20
.20
.79
Cabbage S
CM4-184
.90
.20
.20
CMV-1
CM1841

Figure 29. Phylogenetic tree for the large intergenic region of CaMV constructed for eleven CaMV isolates by the distance method. Numbers indicate branch lengths and are proportionate to sequence divergence among CaMV isolates.

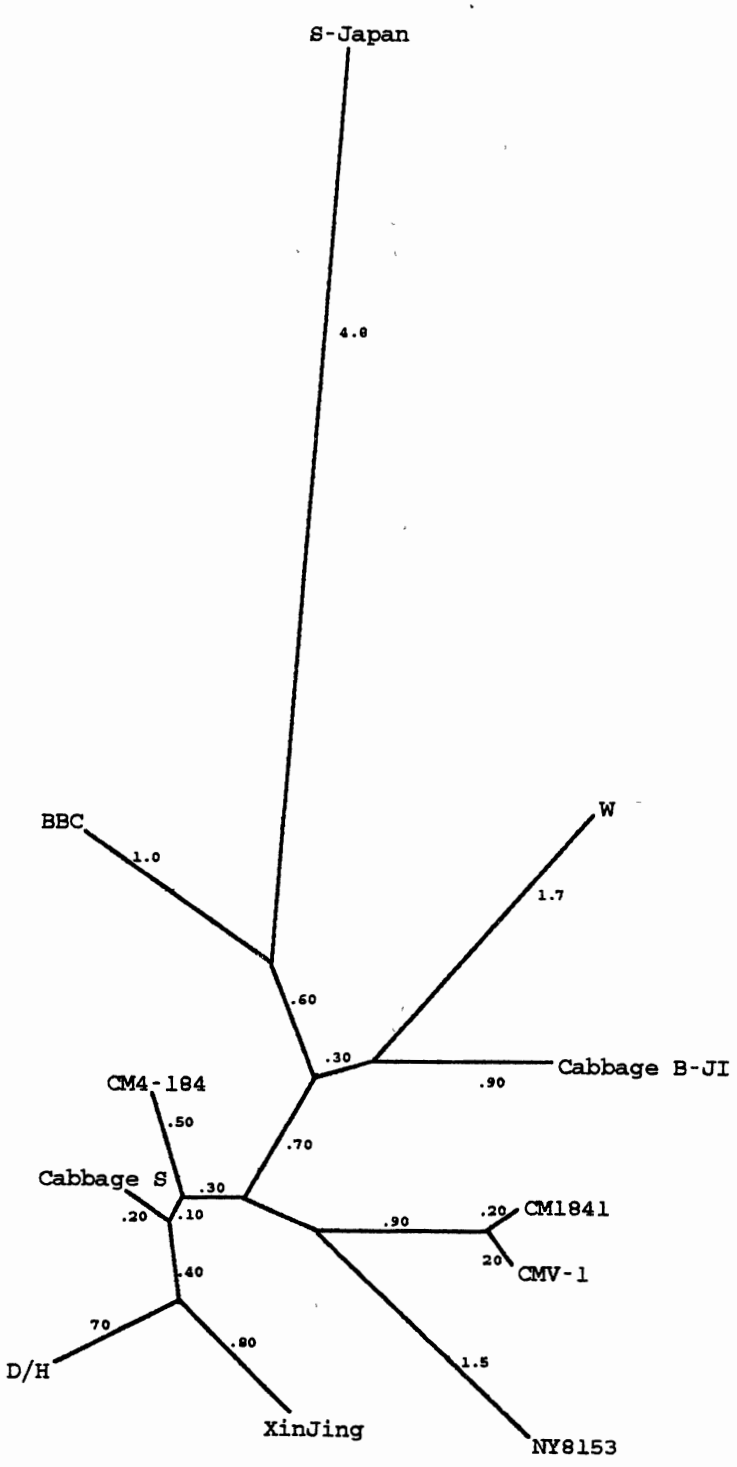
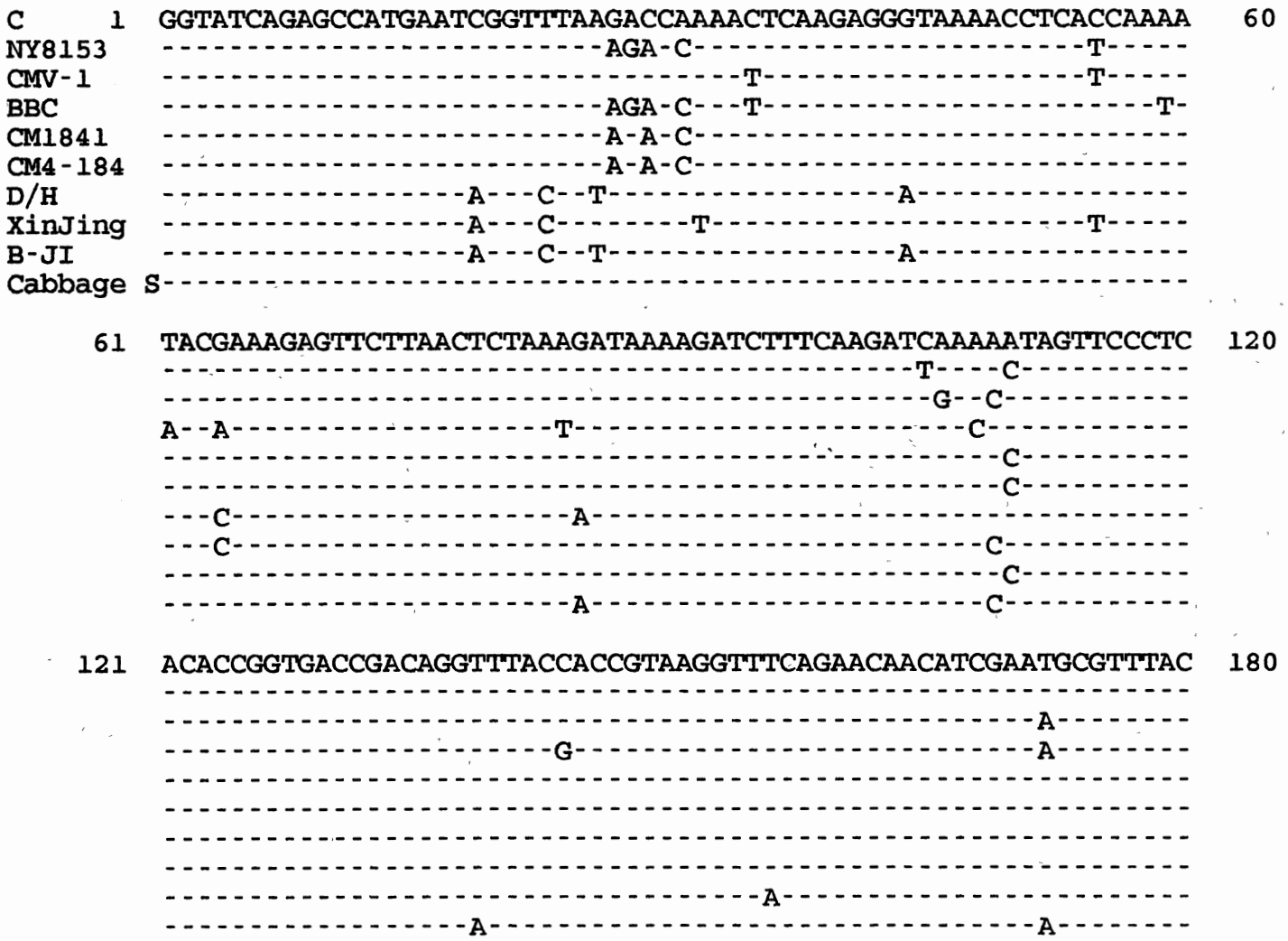

S-Japan
4.8
BBC
1.0
W
1.7
.60
.30
Cabbage B-JI
.90
CM4-184
.50
.70
Cabbage S
.30
.20
.10
.90
.20
CM1841
20
CMV-1
.40
70
.80
1.5
D/H
XinJing
NY8153

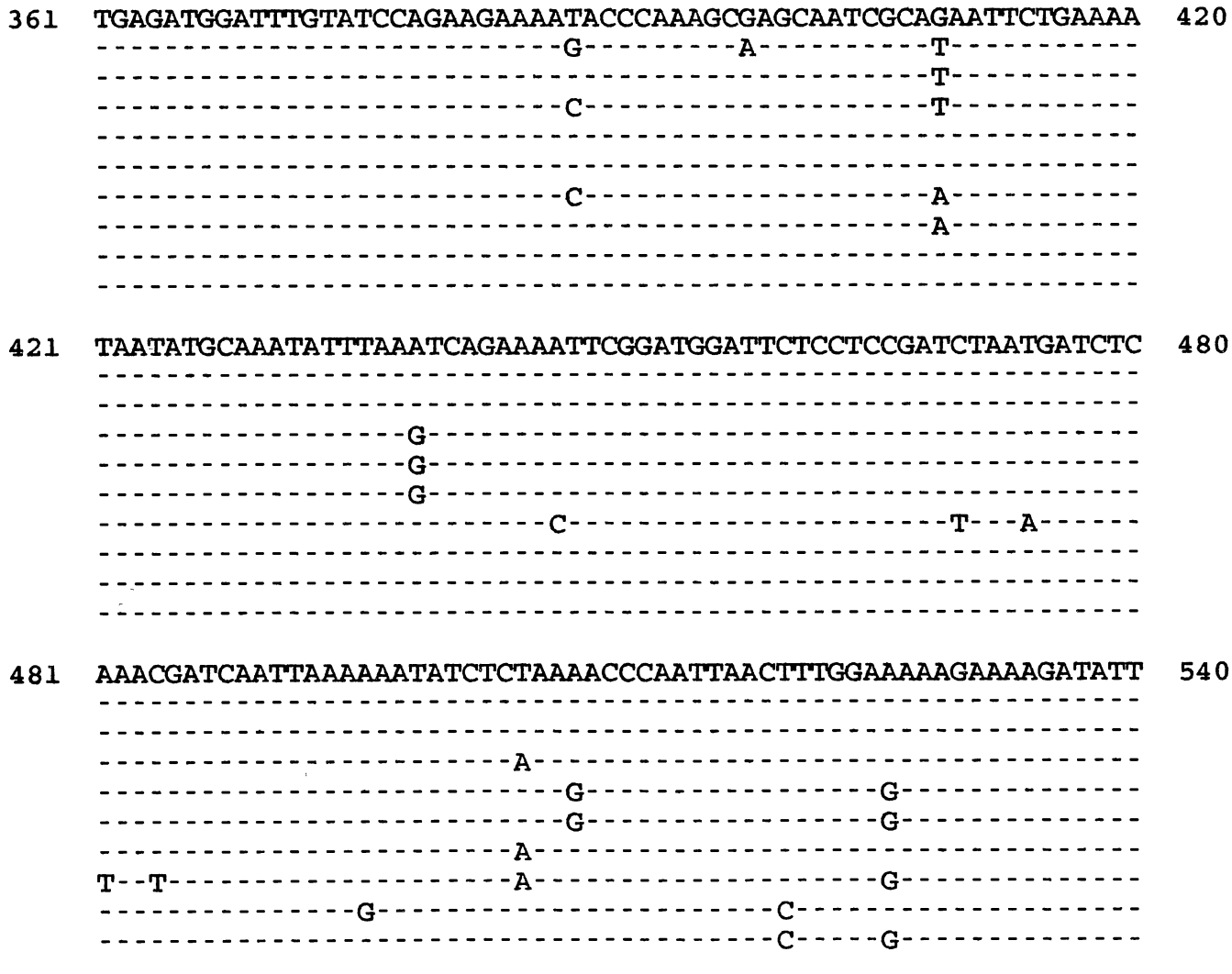
Figure 30. Alignment of CaMV consensus sequence (C) with the complete nucleotide sequences of nine CaMV isolates. Nucleotide position is indicated by numbers at the ends of consensus lines. Dashes represent residues that match the consensus. Dots represent regions where a residue is missing. This figure spans pages 155-200.

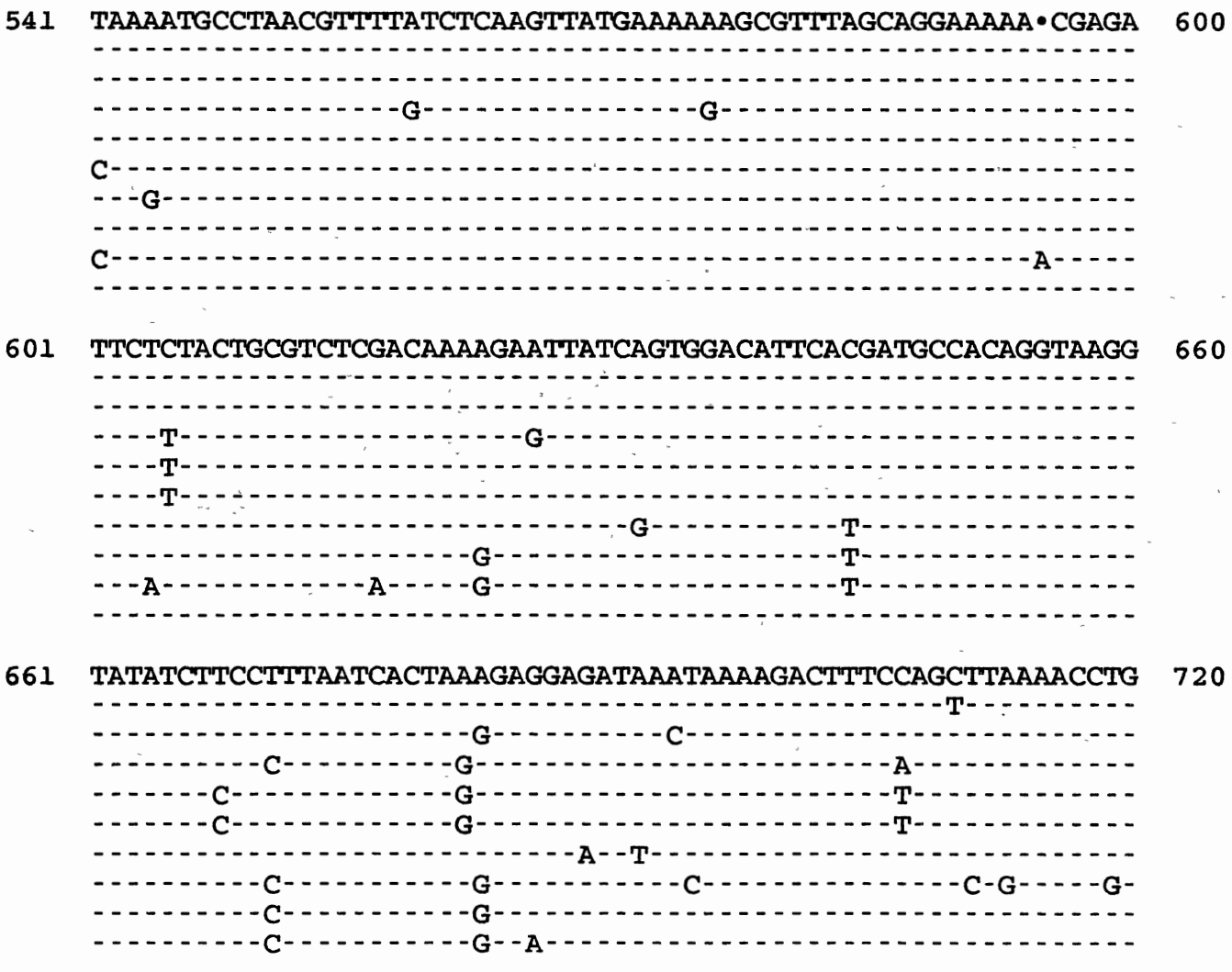

C 1 GGTATCAGAGCCATGAATCGGTTTAAGACCAAAACTCAAGAGGGTAAAACCTCACCAAAA 60
NY8153 ----------AGA-C----------T-----
CMV-1 ----------T----------T-----
BBC ----------AGA-C---T----------T-
CM1841 ----------A-A-C----------
CM4-184 ----------A-A-C----------
D/H ----------A---C--T----------A----------
XinJing ----------A---C---------T----------T-----
B-JI ----------A---C--T----------A----------
Cabbage S----------
61 TACGAAAGAGTTCTTAACTCTAAAGATAAAAGATCTTTCAAGATCAAAAATAGTTCCCTC 120
----------T----C----------
----------G--C----------
A--A----------T----------C----------
----------C----------
----------C----------
---C----------A----------
---C----------C----------
----------C----------
----------A----------C----------
121 ACACCGGTGACCGACAGGTTTACCACCGTAAGGTTTCAGAACAACATCGAATGCGTTTAC 180

----------A--------
----------G----------A--------

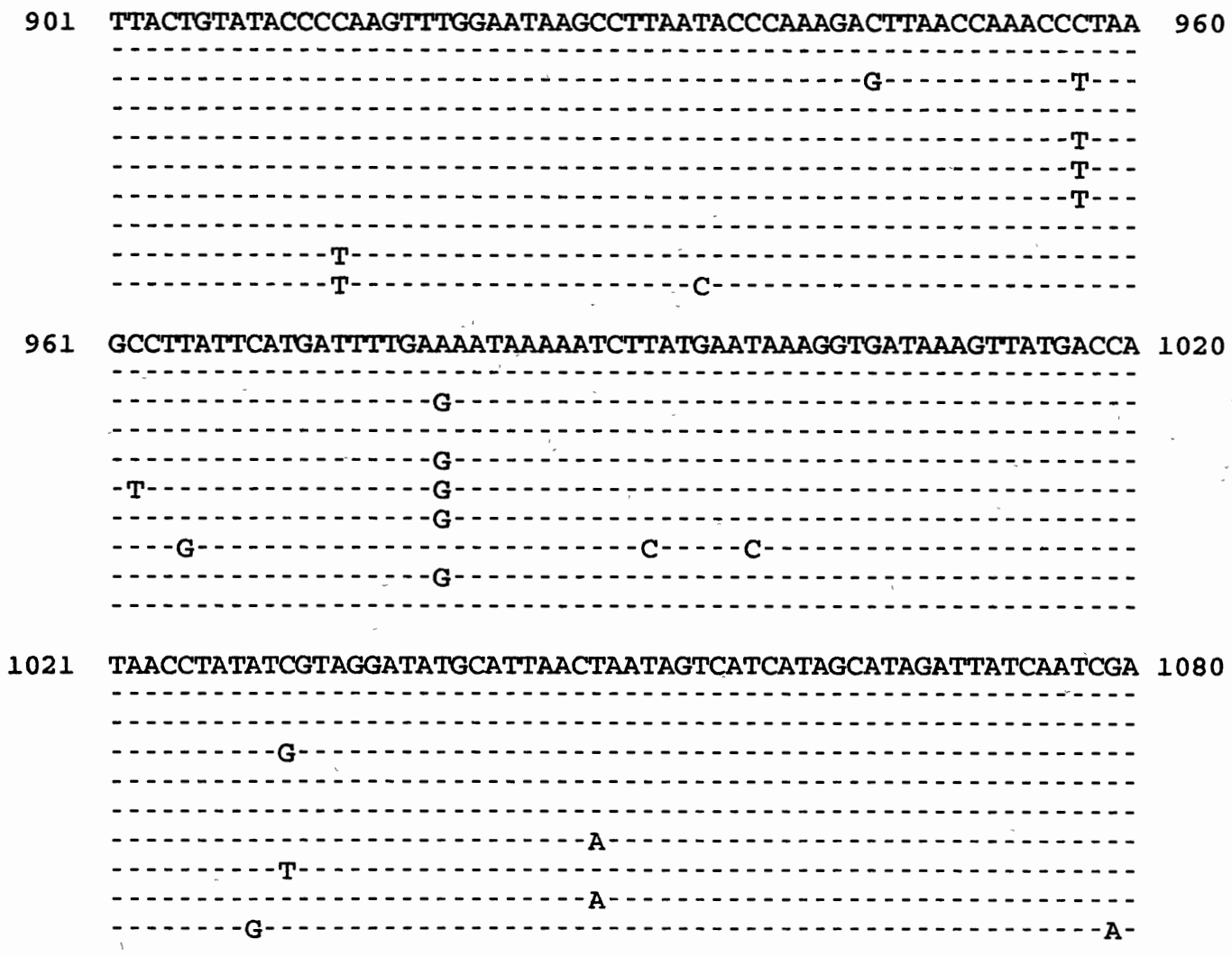
----------A----------
----------A----------A--------

```
181  GCCAACTTCGACTCTCAGCTCAAGTCGTCGTACGATGGTAGATCTAAAAAGATCAAGAAT  240
     ------------------------------------------------------------
     ------------------------------------------------------------
     ----------------GA--A--------------------------------------
     ------------------------------------------------------------
     ------------------------------------------------------------
     ------------------------------------------------------------C-
     ------------------------------------------------------------
     ----------------A-------------------------------------------
     ----------------A-------------------------------------------C-

241  CTAAGCCTTAAAAATCTTAGATGTTACGAAGCCTTCCTCAGGAAGTACCTTCTGGAACAA  300
     ----------------------C-------------------------------------
     ----------------------C-------------------------------------
     ------------------------------------------------------------
     ---------------------T--------------------------------------
     ---------------------T--------------------------------------
     ------------------------------A---------------------T--------
     ----------------------C------A---------------------T--------
     ------------------------------------------------------------
     ------------------------------------------------------------

301  TAAA•TCTCTCTGAGAATAGTACTCTATTGAGTATCCACAGAAAAAATAATCTTCTGTGT  360
     ------------------------------------------T-----------------
     ------------------------------------------T--------T--------
     ------------------------------------------------------------
     ---------------------AC------------G------------------------
     ---------------------AC------------G------------------------
     ----A-------------------------------------------------------
     ----A-------------------------------------------------------
     ---------------------AC----------------------C---------------
     ------------------------------------------G---------C---------
```

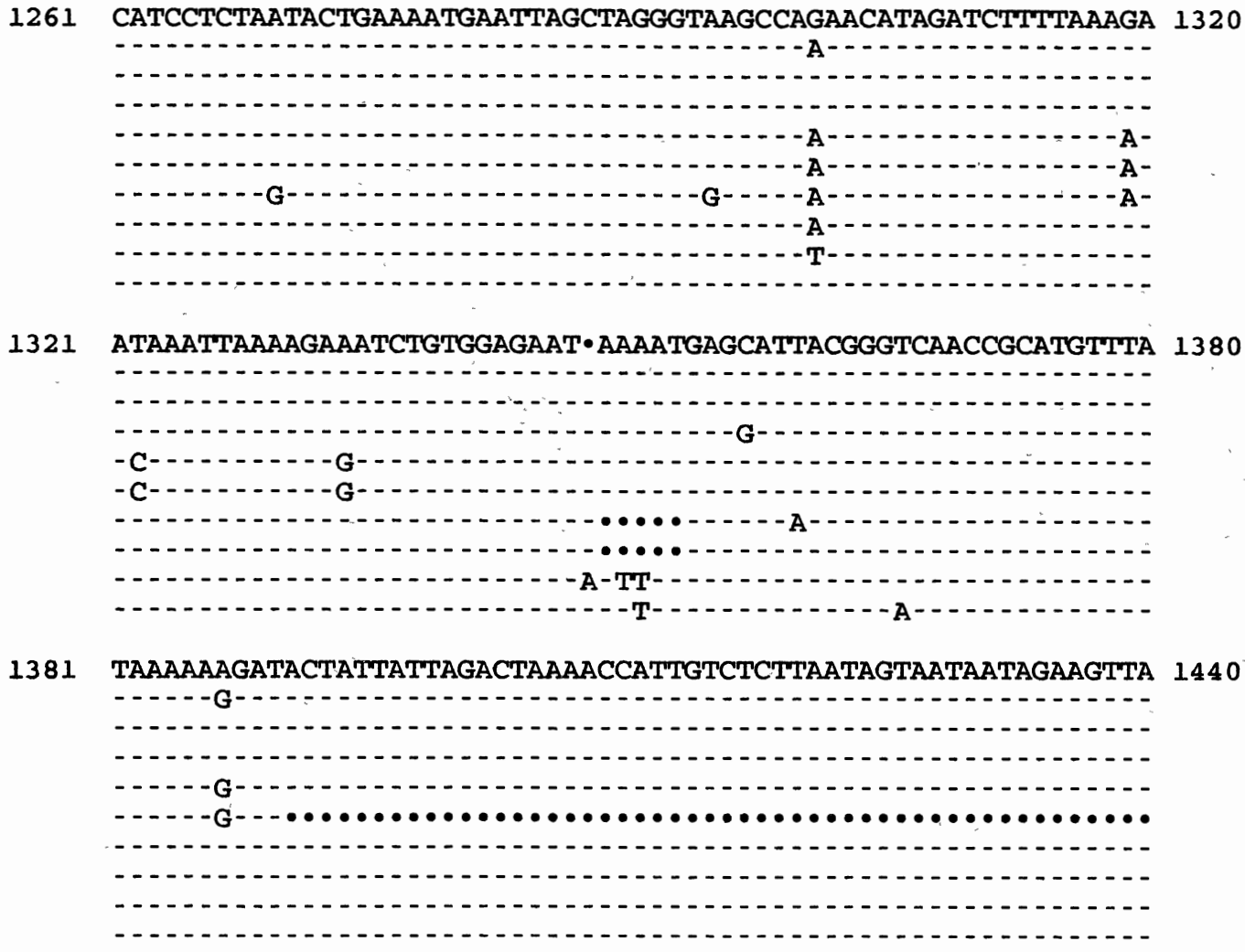

361 TGAGATGGATTTGTATCCAGAAGAAAATACCCAAAGCGAGCAATCGCAGAATTCTGAAAA 420
421 TAATATGCAAATATTTAAATCAGAAAATTCGGATGGATTCTCCTCCGATCTAATGATCTC 480
481 AAACGATCAATTAAAAAATATCTCTAAAACCCAATTAACTTTGGAAAAAGAAAAGATATT 540

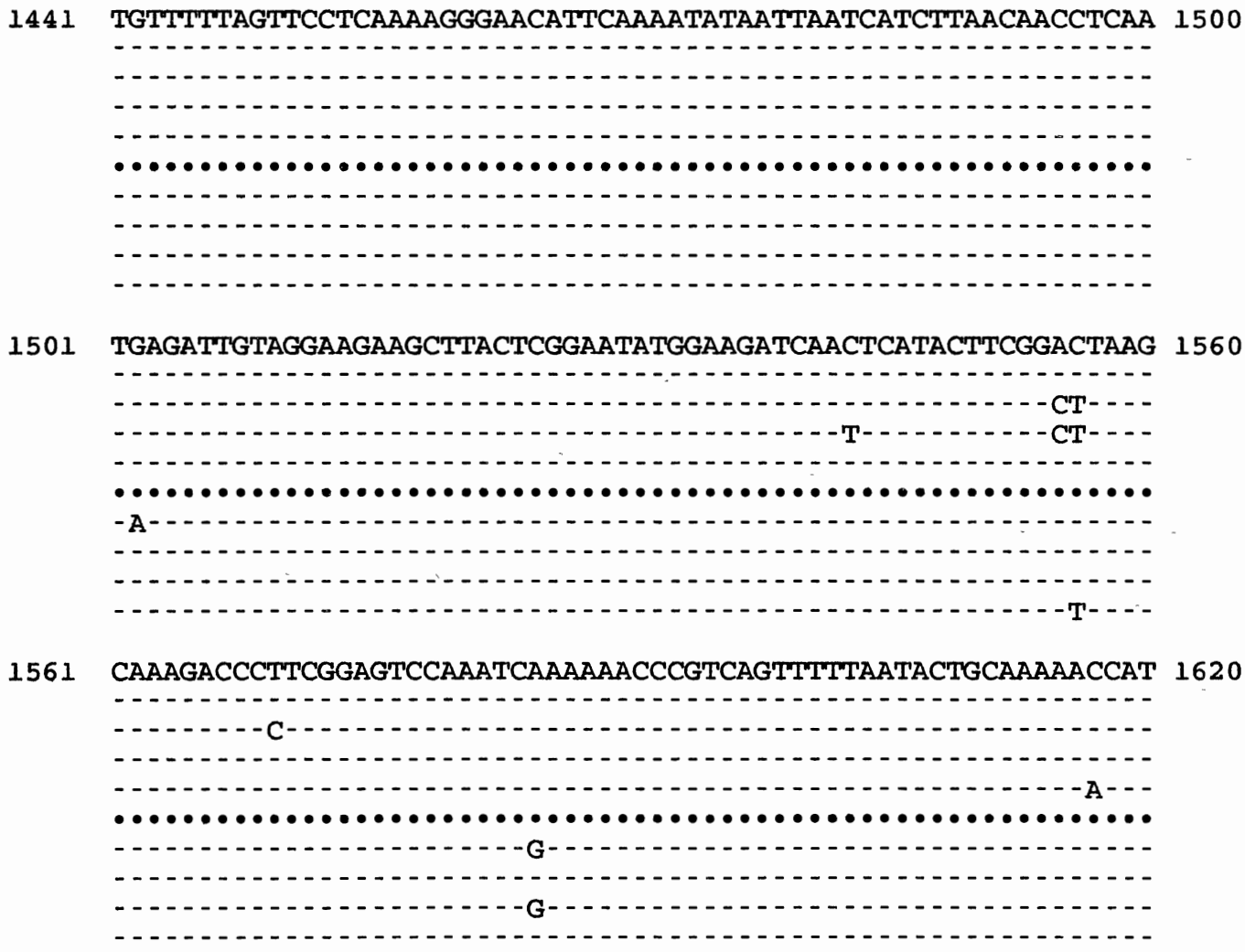

541 TAAAATGCCTAACGTTTTATCTCAAGTTATGAAAAAAGCGTTTAGCAGGAAAAA•CGAGA 600
601 TTCTCTACTGCGTCTCGACAAAAGAATTATCAGTGGACATTCACGATGCCACAGGTAAGG 660
661 TATATCTTCCTTTAATCACTAAAGAGGAGATAAATAAAAGACTTTCCAGCTTAAAACCTG 720

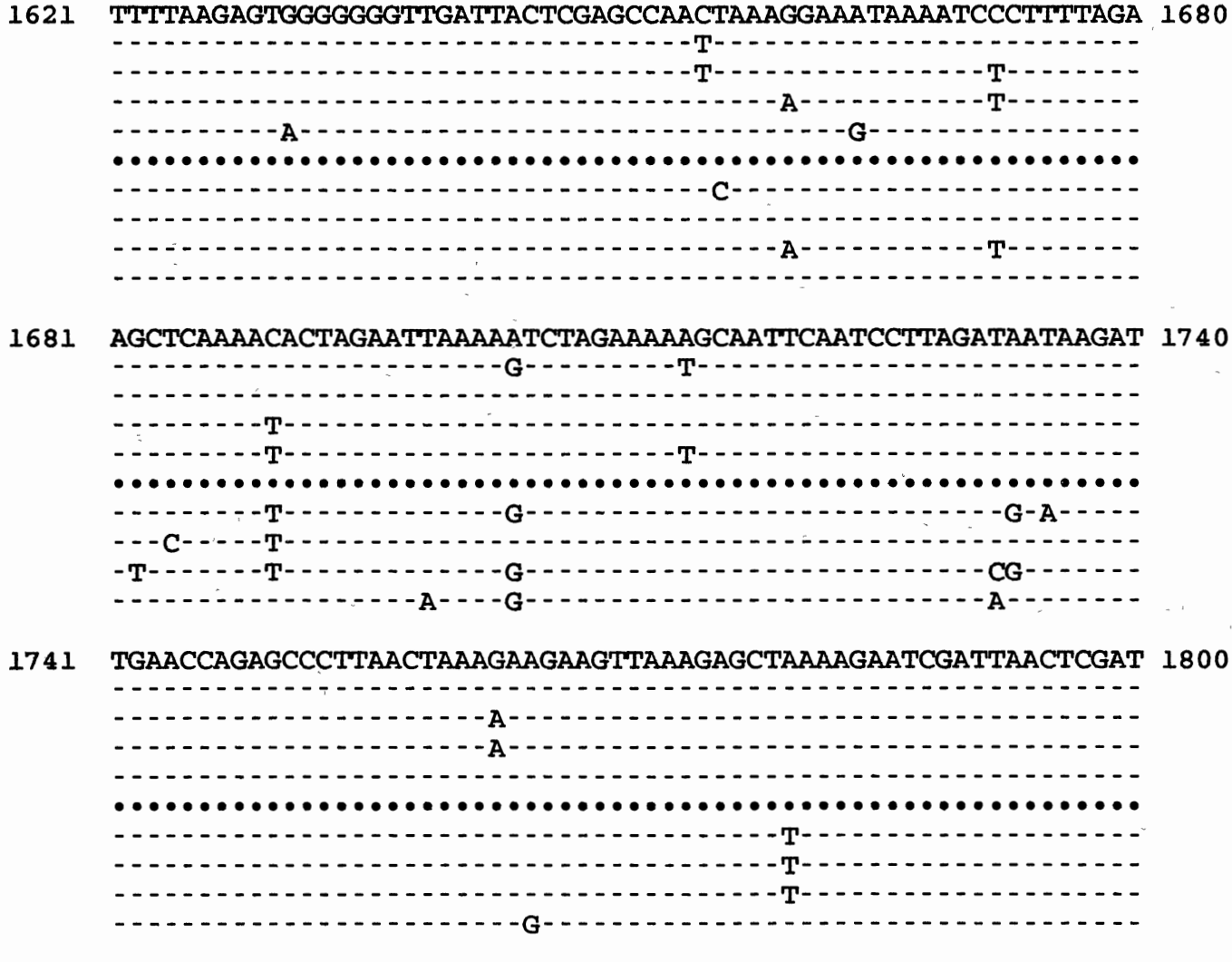
```
721   AAGTCAGAAAGACCATGTCCATGGTTCATCTTGGAGCGGTCAAAATATTGCTTAAAGCTC  780
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------T-------------------------------------------------
      ----------T-------------------------------------------------
      ---------GA--------------C---T-G--C-----------------------
      ----------------------A--A--------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------

781   AATTTCGAAATGGGATTGATACCCCAATCAAAATTGCTTTAATCGATGATAGAATTAATT  840
      ------------------------------------------------------------
      ------------------------------------------------------C-----
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      -----A------------------------------------------------C-----
      ------------------------------------------------------------
      ------------------------------------------------------C-----
      ------------------------------------------------------C-----

841   CTAGAAGAGATTGCCTTCTCGGTGCAGCCAAAGGTAATCTAGCATACGGTAAGTTTATGT  900
      ------------------------------------------------------------
      ------A-----------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------T-----T--------------------C--------------------
      ------A------T--A--T---------------------------------------
      ------------------------------------------------------------
      ----------T-----T---------------------------------------------
```

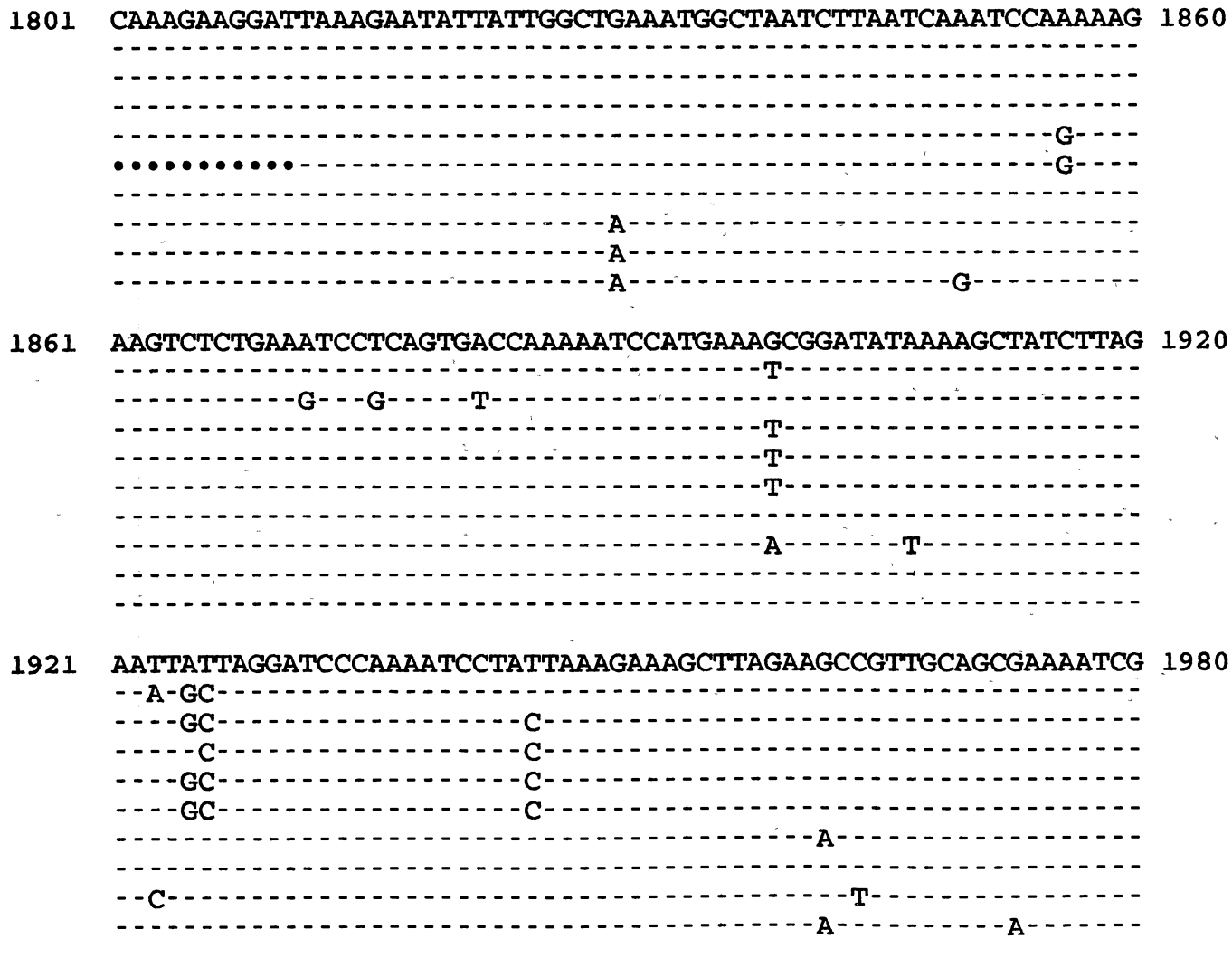

901 TTACTGTATACCCCAAGTTTGGAATAAGCCTTAATACCCAAAGACTTAACCAAACCCTAA 960
961 GCCTTATTCATGATTTTGAAAATAAAAATCTTATGAATAAAGGTGATAAAGTTATGACCA 1020
1021 TAACCTATATCGTAGGATATGCATTAACTAATAGTCATCATAGCATAGATTATCAATCGA 1080

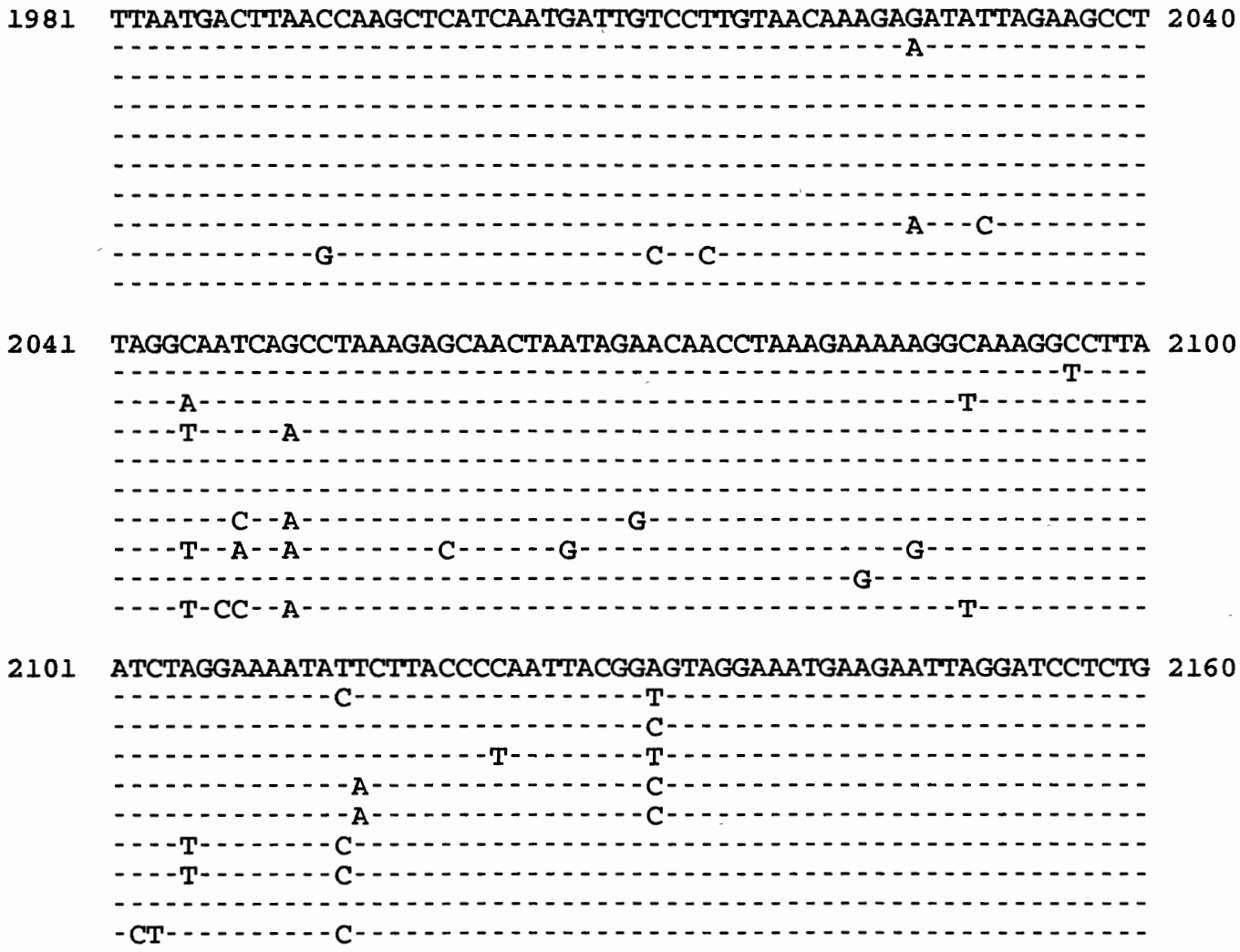
```
1081  ATGCTACAATTGAACTAGAAGACGTATTTCAAGAAATTGGAAATGTCCAGCAATCTGATT 1140
      ------------------------------------------------------G-----
      ------------------------------------------------------------
      ----------------------------------------------------CG-G-------
      ------------------------------------------------------------C-
      ------------------------------------------------------------C-
      ----------------------------------------------A-------------G-
      ------------------------------------------------------------G-
      ----------------------------------------------C--------G-------
      ------------------------------------------------------------G-

1141  TCTGTACAATACAGAATGACGAATGCAATTGGGCCATTGATATAGCCCAAAACAAAGCCT 1200
      --------------------------T-----------------------------------
      -T-----------A-------------------------------------------------
      -T-----------A-------------------------------------------------
      -T-----------A-------------------------------------------------
      -T-----------A-------------------------------------------------
      --------------------T-------------------------------------------
      --------------------T-------------------------------------------
      --------------------------T-----------------------------------
      --------------------T-------------------------------------------

1201  TATTAGGAGCTAAAACCAAATCCCAAATTGGTAATAGTCTTCAAATAGGAAACAGTGCTT 1260
      ------------------C-------------------------------------------
      -------TC----------------------------------------------------
      ------------------------------------A-----------------------
      ------------------C-------------------------------------------
      ------------------C-------------------------------------------
      ----------------------A-----------------------------T-T---A-
      ----------------------A-------------------------------------
      --------------G-----A-----------------------------------G------
      --------------------GA-T-------------AC-----------T----------
```

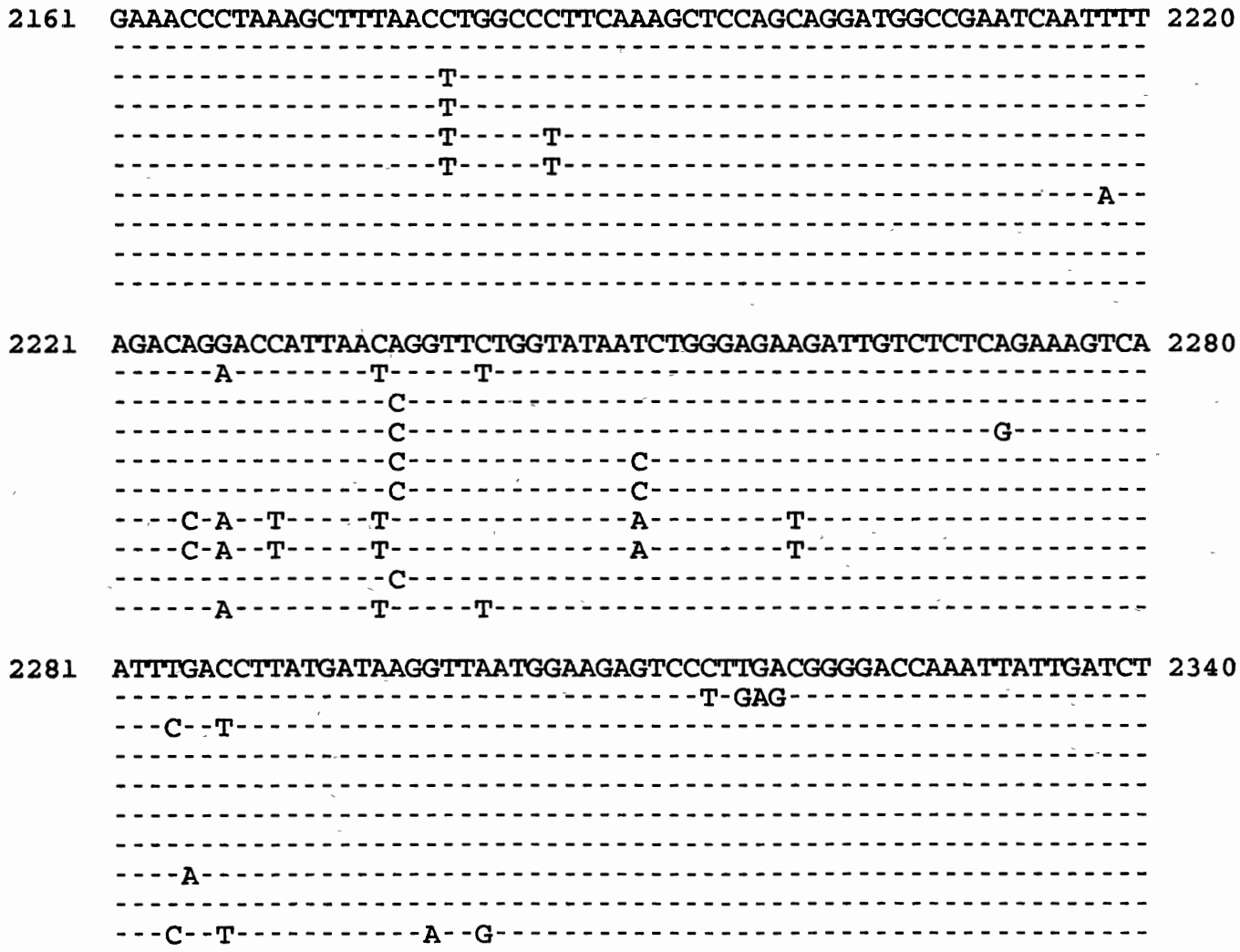

1261 CATCCTCTAATACTGAAAATGAATTAGCTAGGGTAAGCCAGAACATAGATCTTTTAAAGA 1320
1321 ATAAATTAAAAGAAATCTGTGGAGAAT•AAAATGAGCATTACGGGTCAACCGCATGTTTA 1380
1381 TAAAAAAGATACTATTATTAGACTAAAACCATTGTCTCTTAATAGTAATAATAGAAGTTA 1440


1441 TGTTTTTAGTTCCTCAAAAGGGAACATTCAAAATATAATTAATCATCTTAACAACCTCAA 1500
1501 TGAGATTGTAGGAAGAAGCTTACTCGGAATATGGAAGATCAACTCATACTTCGGACTAAG 1560
1561 CAAAGACCCTTCGGAGTCCAAATCAAAAAACCCGTCAGTTTTTAATACTGCAAAAACCAT 1620


1621 TTTTAAGAGTGGGGGGGTTGATTACTCGAGCCAACTAAAGGAAATAAAATCCCTTTTAGA 1680
1681 AGCTCAAAACACTAGAATTAAAAATCTAGAAAAAGCAATTCAATCCTTAGATAATAAGAT 1740
1741 TGAACCAGAGCCCTTAACTAAAGAAGAAGTTAAAGAGCTAAAAGAATCGATTAACTCGAT 1800


1801 CAAAGAAGGATTAAAGAATATTATTGGCTGAAATGGCTAATCTTAATCAAATCCAAAAAG 1860
1861 AAGTCTCTGAAATCCTCAGTGACCAAAAATCCATGAAAGCGGATATAAAAGCTATCTTAG 1920
1921 AATTATTAGGATCCCAAAATCCTATTAAAGAAAGCTTAGAAGCCGTTGCAGCGAAAATCG 1980

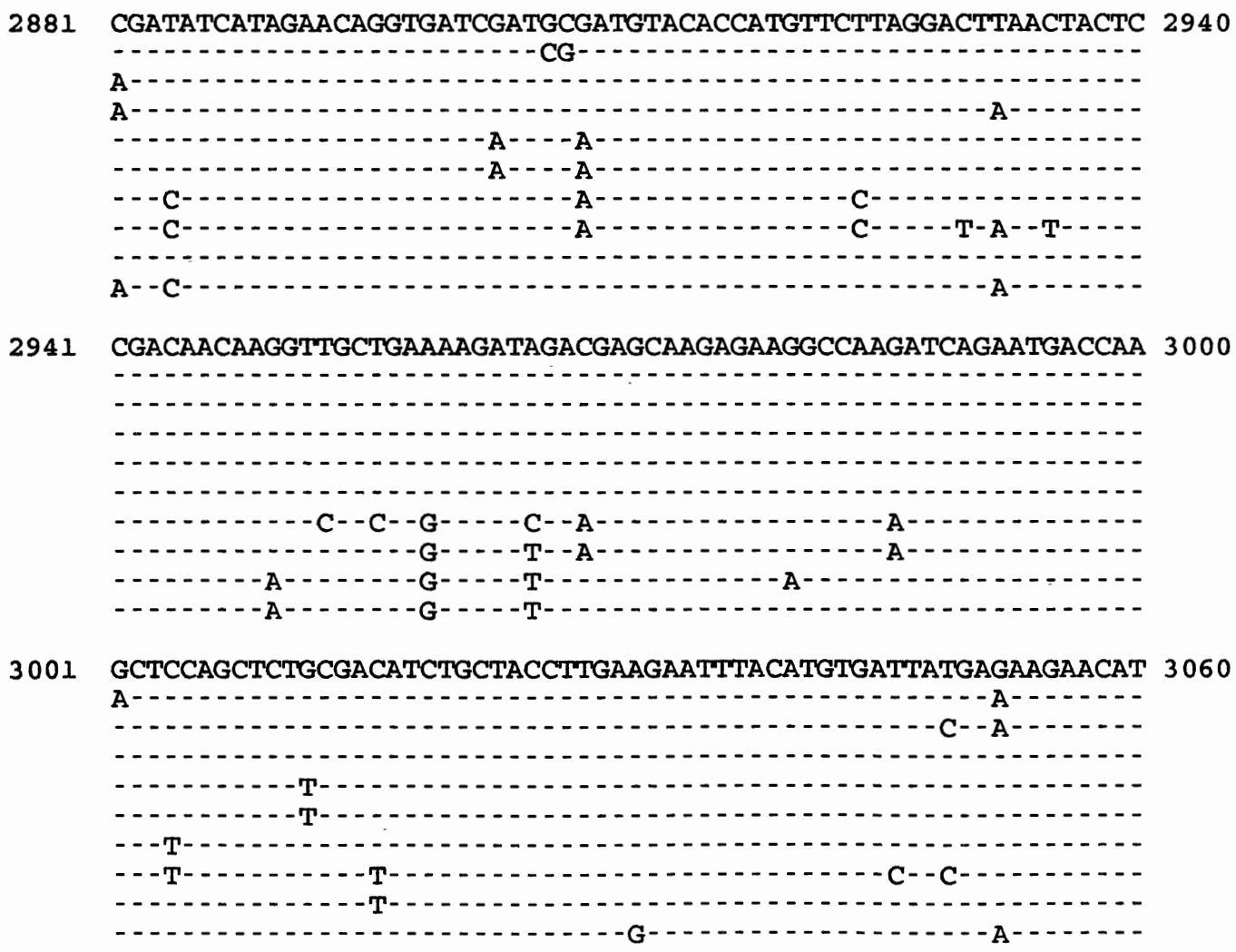

1981 TTAATGACTTAACCAAGCTCATCAATGATTGTCCTTGTAACAAAGAGATATTAGAAGCCT 2040
2041 TAGGCAATCAGCCTAAAGAGCAACTAATAGAACAACCTAAAGAAAAAGGCAAAGGCCTTA 2100
2101 ATCTAGGAAAATATTCTTACCCCAATTACGGAGTAGGAAATGAAGAATTAGGATCCTCTG 2160


2161 GAAACCCTAAAGCTTTAACCTGGCCCTTCAAAGCTCCAGCAGGATGGCCGAATCAATTTT 2220
2221 AGACAGGACCATTAACAGGTTCTGGTATAATCTGGGAGAAGATTGTCTCTCAGAAAGTCA 2280
2281 ATTTGACCTTATGATAAGGTTAATGGAAGAGTCCCTTGACGGGGACCAAATTATTGATCT 2340

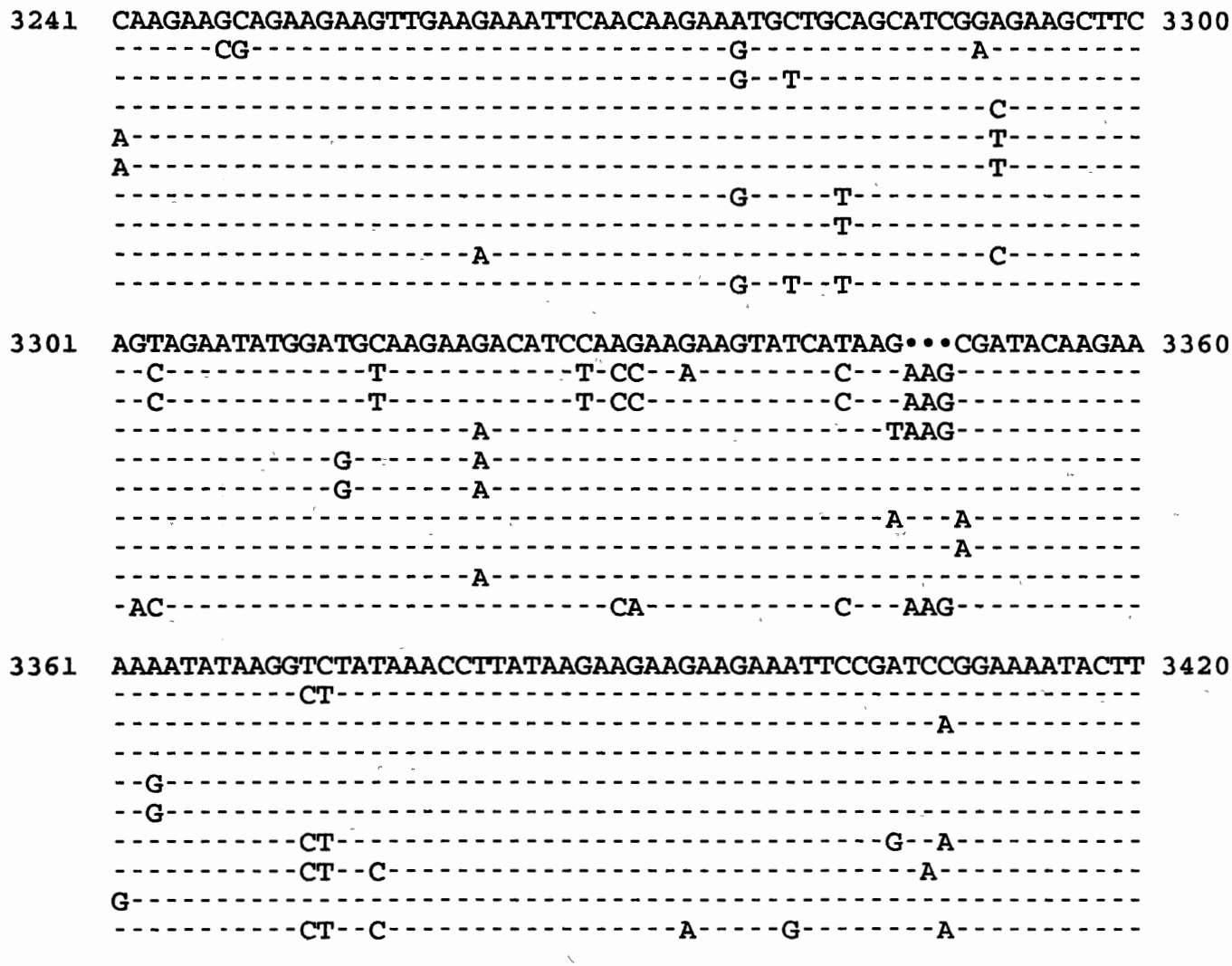

2341 AACCTCTCTACCTAGTGATAATTTGCAGGTCGAACAGGTTATGACAACTACCGAAGACTC 2400
2401 GATCTCGGAA•••GAATCAGAATTCCTTCTAGCAATAGGA•••••••••••••••••••• 2460
2461 ••••••••••••••••••••••GAAACATCTGAAGACGAAAGCGATTCAGGAGAAGAACC 2520


2521 TGAATTCGAACAAGTTCGAATGGATCGAACAGGAGGAACGGAGATTCCCAAAGAAGAAGA 2580
2581 TGGTGAAGAACCATCTAGATACAATGAGAGAAAGAGAAAGACCCCGGAGGACCGGTACTT 2640
2641 TCCAACTCAACCAAAGACCATTCCAGGACAAAAGCAAACGTCTATGGGAATGCTCAACAT 2700

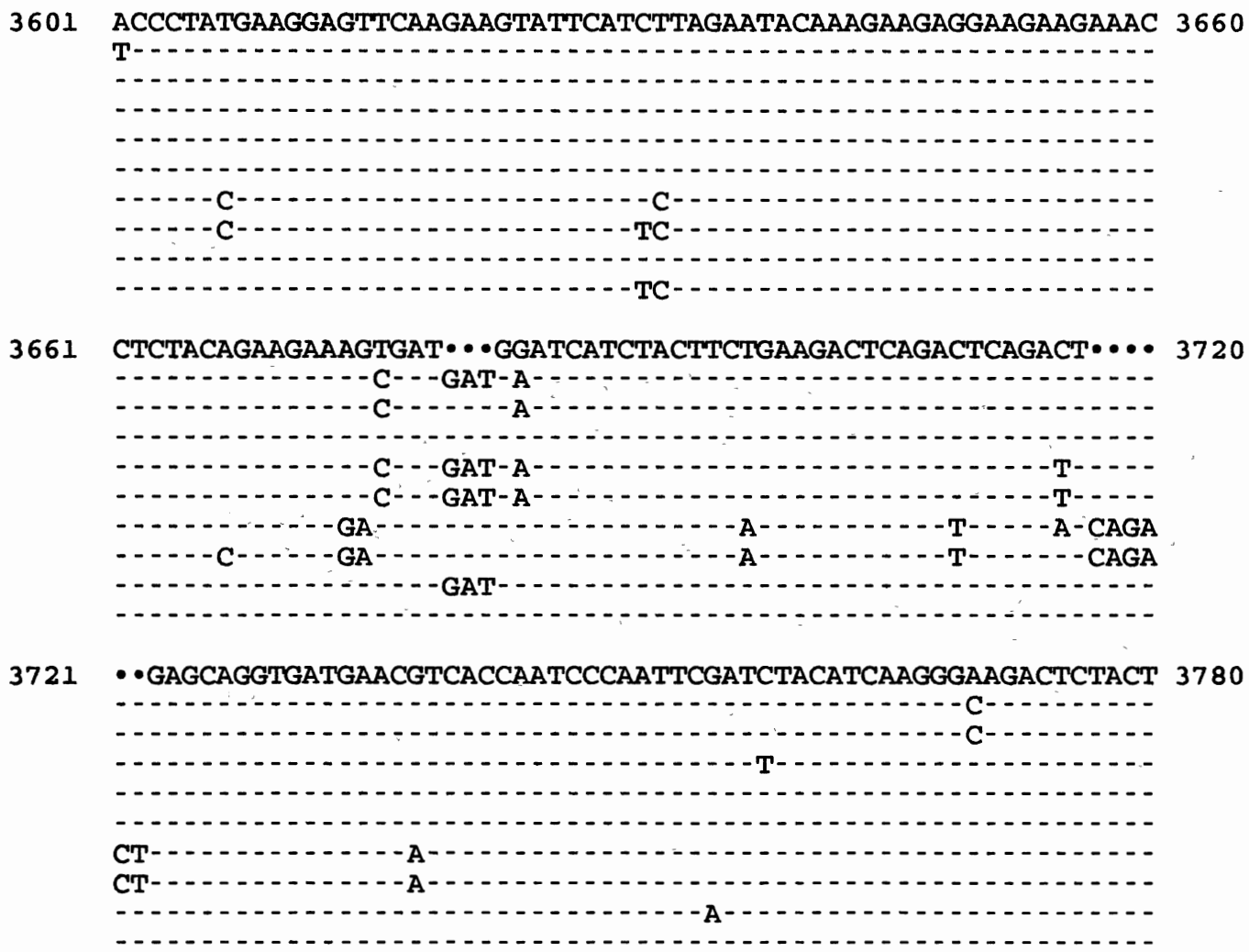

2701 TGACTGCCAAACCAATCGAAGAACTTTAATCGATGATTGGGCAGCAGAAATCGGATTGAT 2760
2761 AGTCAAGACCAACAGAGAAGACTATCTTGATCCAGAAACAATACTACTCTTGATGGAACA 2820
2821 CAAAACATCAGGAATAGCCAAGGAGTTAATCCGAAATACAAGATGGAACCGCACTACCGG 2880


2881 CGATATCATAGAACAGGTGATCGATGCGATGTACACCATGTTCTTAGGACTTAACTACTC 2940
2941 CGACAACAAGGTTGCTGAAAAGATAGACGAGCAAGAGAAGGCCAAGATCAGAATGACCAA 3000
3001 GCTCCAGCTCTGCGACATCTGCTACCTTGAAGAATTTACATGTGATTATGAGAAGAACAT 3060

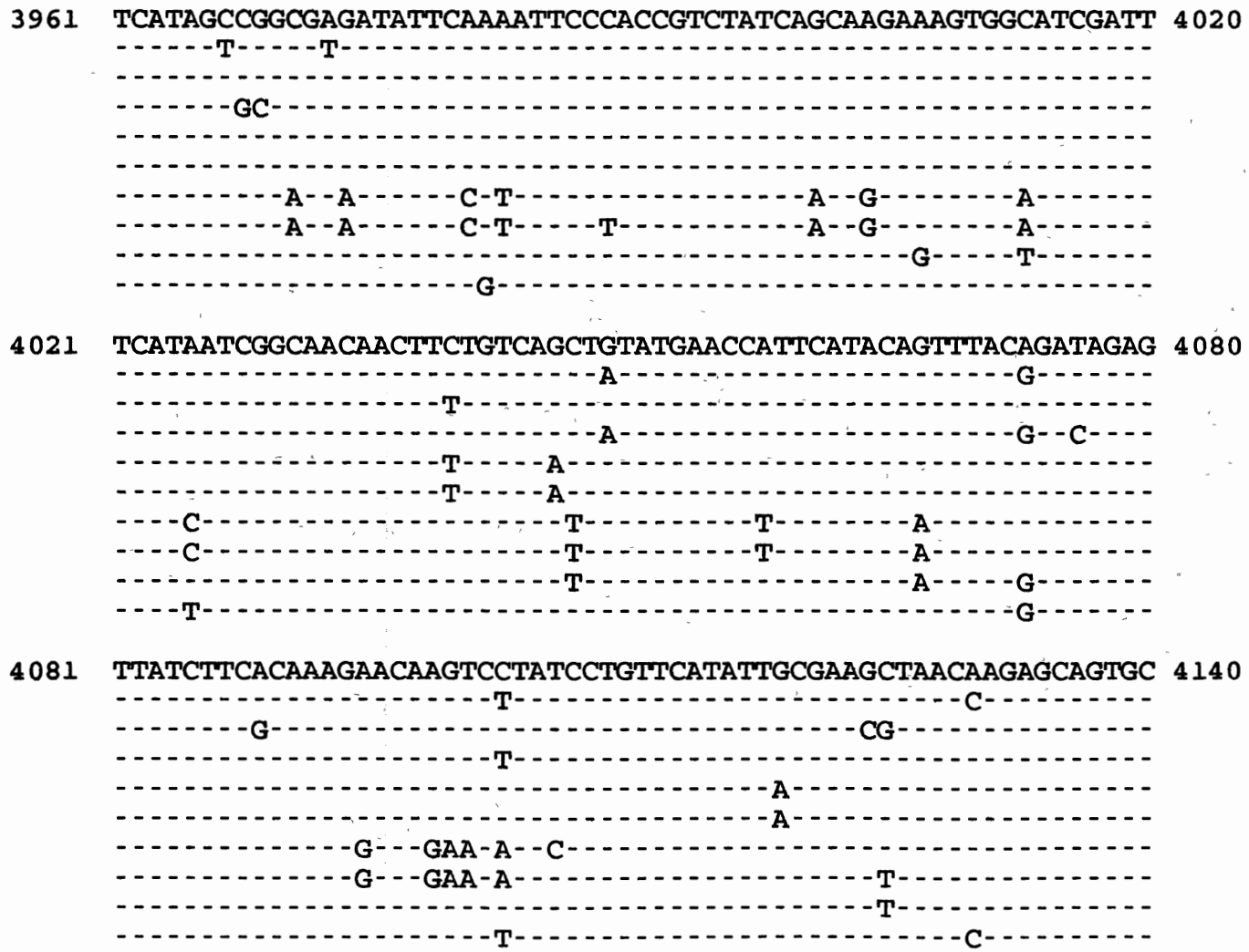
```
3061  GTACAAGACGGAACTGGCGGATTTCCCAGGATATATCAACCAGTACCTGTCAAAAATCCC 3120
      ------------------------------------------------------------
      ----------------------------------C-------------------------
      ------A-----------------------------------------------------
      -------------A-------------T-----C-------------------------
      -------------A-------------T-----C-------------------------
      ----------A-------------------------------------------------
      ---T-----A-------------------------------------------------
      ----------A-----------------------------------------A-------
      ---T-----A-------------------------------------------------

3121  CATCATTGGAGAAAAAGCGCTAACACGCTTTAGGCATGAAGCCAACGGAACCAGCATCTA 3180
      ------A-----------------------------------------------------
      ------A-----------------------------------T-----------------
      ----------------------------------------T-----------------
      ------A-----------------------------------------------------
      ------A-----------------------------------------------------
      --------------------T---------------------------------------
      ------------------------------------------------------------
      --------------------T----G--------A--------T-------------C---
      --------------------T-----------------------T-----------------

3181  CAGCTTAGGTTTCGCGGCAAAGATAGTAAAAGAAGAACTATCTAAAATCTGCGACTTATC 3240
      --------------A-CG-------TGC---------------------TC----------
      ------------------G-----------------------------------------
      ----------------------------------------------------CA-----
      ----------------------------------------A-----------------
      ----------------------------------------A-----------------
      ---T-------------------------------------------------------GA-
      ------------------------------------------------------------
      ------------------G-----------------------------------------
      ---T----------------------C-------------------------------
```

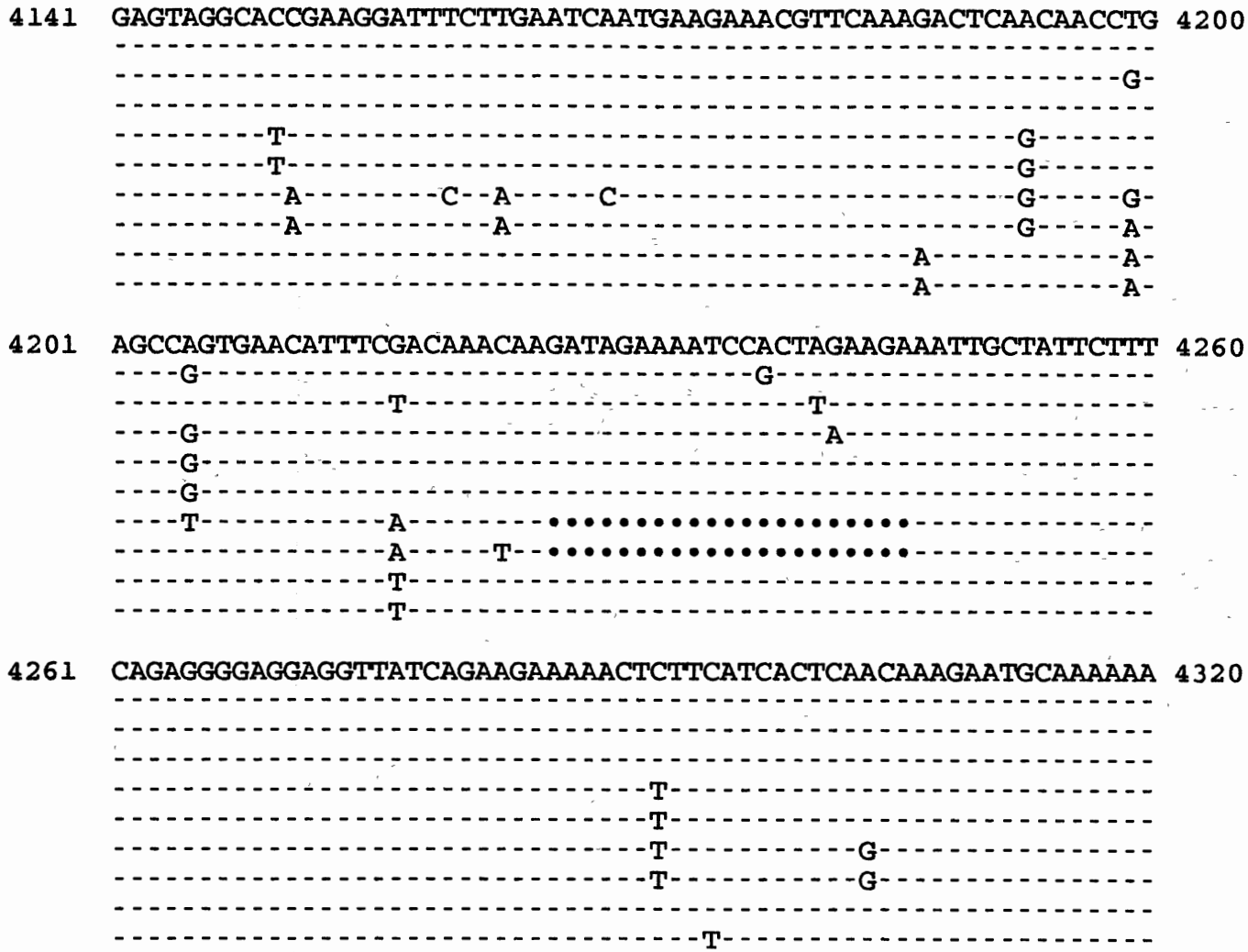

3241 CAAGAAGCAGAAGAAGTTGAAGAAATTCAACAAGAAATGCTGCAGCATCGGAGAAGCTTC 3300
3301 AGTAGAATATGGATGCAAGAAGACATCCAAGAAGAAGTATCATAAG•••CGATACAAGAA 3360
3361 AAAATATAAGGTCTATAAACCTTATAAGAAGAAGAAGAAATTCCGATCCGGAAAATACTT 3420

```
3421  CAAGCCCAAAGAAAAGAAGGGCTCAAAGCAAAAGTATTGCCCAAAAGGCAAGAAAGACTG 3480
      ------------G-------------------------------------------------
      ----------------------------------------------------------T--
      ---------G--G-------------------------------------------------
      ------------G----------------G-------------------------G-------
      ------------G----------------G-------------------------G-------
      -------------------------T----------------------G--------------
      -------------------------T-------------------------------T--
      ------T-------------------------------------------------------
      ----------------------------------------------------------T--

3481  CAGATGTTGGATCTGCAATATCGAAGGCCATTACGCCAACGAATGTCCTAATCGACAAAG 3540
      ---G------------------------T-----------------------------------
      ----------------------------------------------------------------
      --------------CG--C--T-----------------------------------------
      ----------------------------------------------------------------
      ----------------------------------------------------------------
      ----------------------------------------------------------------
      ---G---------------C--A-----------T-------------------------GT
      ---G------------------------------------------------------------
      ------------------C--T-----------------------------------------

3541  CTCGGAGAAGGCTCACATCCTTCAACAAGCAGAAAAATTGGGTCTCCAGCCCATTGAAGA 3600
      ------A-------------------------------G-T--C----------------C
      ------------------------------------G-----C----------------------
      ---------------------------------G--------------------------------
      ---------------------------------G--T---------------G--------------
      ---------------------------------G--T---------------G--------------
      ---A-------------------------------G---C----------------C--------
      ------------------------------------------C-----------A--------
      ------------------------------------G-----C----------------------
      ----------------------------------------------------------------
```


3601 ACCCTATGAAGGAGTTCAAGAAGTATTCATCTTAGAATACAAAGAAGAGGAAGAAGAAAC 3660
3661 CTCTACAGAAGAAAGTGAT•••GGATCATCTACTTCTGAAGACTCAGACTCAGACT•••• 3720
3721 ••GAGCAGGTGATGAACGTCACCAATCCCAATTCGATCTACATCAAGGGAAGACTCTACT 3780

```
3781  TCAAGGGATACAAGAAGATAGAGCTTCACTGTTTTGTAGACACGGGAGCAAGCTTATGCA 3840
      ------------------------------------------------------------
      ----A-------------------------------------T-------------------
      ------------------------------------------------------------
      ----A-------------------------------------------------------
      ----A-------------------------------------------------------
      ----------------------------------------------------------T-------
      ---------------------------------C--------------------T-------
      ------------------------------------C-------------------------
      ----------------------A-----------C------------------C------

3841  TAGCATCCAAGTTCGTCATACCAGAAGAACATTGGGTCAATGCAGAAAGACCAATAATGG 3900
      --------------------T-----------------------------------------
      ------------------------------------T-----------------------
      --------------------T-----------------------------------------
      ------------------------------------T-----------------------
      ------------------------------------T-----------------------
      ----------------------------------A-------------------C----
      ----------A-------------------------A-------------------C----
      --------------------T-----------------T-------------------C---
      ------------------------------------------------------------T----

3901  TCAAAATAGCAGATGGAAGTTCAATCACCATCAGCAAAGTCTGCAAAGACATAGACTTGA 3960
      --------------------C-----------------------------------------
      --------------------C-----------------------------------------
      ------------------------------------------------------------
      ---------------------------T--------------------------------
      ---------------------------T--------------------------------
      -------T---------------G-----------A-----------G------T---C---
      -------T-----------C--G-----------A-----------G------T---C---
      ------------------------------------------------------------T----
      --------------------C-----------------------------------------
```


3961 TCATAGCCGGCGAGATATTCAAAATTCCCACCGTCTATCAGCAAGAAAGTGGCATCGATT 4020
4021 TCATAATCGGCAACAACTTCTGTCAGCTGTATGAACCATTCATACAGTTTACAGATAGAG 4080
4081 TTATCTTCACAAAGAACAAGTCCTATCCTGTTCATATTGCGAAGCTAACAAGAGCAGTGC 4140


4141 GAGTAGGCACCGAAGGATTTCTTGAATCAATGAAGAAACGTTCAAAGACTCAACAACCTG 4200
4201 AGCCAGTGAACATTTCGACAAACAAGATAGAAAATCCACTAGAAGAAATTGCTATTCTTT 4260
4261 CAGAGGGGAGGAGGTTATCAGAAGAAAAACTCTTCATCACTCAACAAAGAATGCAAAAAA 4320

```
4321  TCGAAGAACTACTTGAGAAAGTATGTTCAGAAAATCCATTAGATCCTAACAAGACTAAGC 4380
      C------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------A-------------------------------------------------
      ----------A-------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------

4381  AATGGATGAAAGCTTCAATCAAGCTCAGCGACCCAAGCAAAGCTATCAAGGTTAAACCCA 4440
      ------------------------------------------------------------
      ------------------------T----------A-----------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------------T-------------------------------------------

4441  TGAAGTATAGCCCAATGGATCGTGAAGAATTTGACAAGCAAATCAAAGAGTTACTGGACC 4500
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----A--C------------------------------------------------------
      ----A--C------------------------------------------------------
      ----------------------------------T---------------A--------T-
      ------------------------------------------------AG----------
      ------------------------------------------------------------
      ----------------------C-------------------------A-----------
```

```
4501  TTAAAGTCATTAAGCCCAGTAAAAGCCCTCACATGGCACCAGCCTTCTTGGTCAACAATG 4560
      ------------------------------------------------------------
      ------------------------------------------------------------C-
      ------------------------------------------------------------
      --------------A-------------------------------------------------
      --------------A-------------------------------------------------
      -A--------C-----------------------------------------------------
      ------------------------------------------------------------C-
      ------------------------------------------------------------
      -A--------C-----------------------------------------------------

4561  AAGCCGAGAAGCGAAGAGGAAAGAAACGTATGGTAGTCAACTACAAAGCTATGAACAAAG 4620
      ----------CG----------C---------------G-----------------T----
      ------------------------------------------------------------
      -------------------------G-----------T--------G-------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----T-------------------------------------------------------
      -------------------------G--------------------G-------------
      ------------------------------------------------------------

4621  CCACCGTAGGAGACGCATACAATCTTCCCAACAAAGACGAGTTACTTACACTCATTCGAG 4680
      ------------------------------------------------------------
      ------------------------------------------------------------
      -----A--------------------------T--------------G-----T-------
      -----A-------------------------------------------------------
      -----A-------------------------------------------------------
      ----T-----------T-------C------------------------------------
      ----T-----------T--------------------------------------------
      -----A-----------------------------------------G-----T-------
      -T--T--------T--C--------------------------------------------
```

```
4681  GAAAGAAGATCTTCTCTTCCTTCGACTGTAAGTCAGGATTCTGGCAAGTTCTGCTAGATC 4740
      ----------T----------------------------------------------T----
      ---------------------------------C-----------------A--T----
      ------------------------------C-----------------G-------------
      ----------T-----------------C--C-----------------A--T----
      ----------T-----------------C--C-----------------A--T----
      ----------T-----------------C--------------G-------------
      ----A-----------------------------------------G-------C-----
      ------------------------------C-------------------------------
      ------------------------------------------------------------

4741  AAGAATCAAGACCTCTAACGGCATTCACATGTCCACAAGGTCACTACGAATGGAATGTGG 4800
      ----------------------G-------------------------------------
      ----------------------G---------C-----------------------------
      ---------------------------------C-----------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ---------------------------------C-----------T-------------
      ------------------------------------------------T-------------
      ---------------------------------C-----------------------------
      ------------------------------------------------------------

4801  TCCCTTTCGGCCTAAAGCAGGCACCATCCATATTCCAAAGACACATGGACGAAGCATTTC 4860
      ---------------------------------------G----------------------
      ---------------------------------------G----------------------
      -----------T------------------------------------------------
      ---------------------------------------G----------------------
      ---------------------------------------G----------------------
      -A---------T-------A--T-----------------------------------T--C-
      -T---------T-------A--------------------G-----------------C-
      ----------------------T-------------------------------------
      -----------T-------A--T-------------------------------------
```

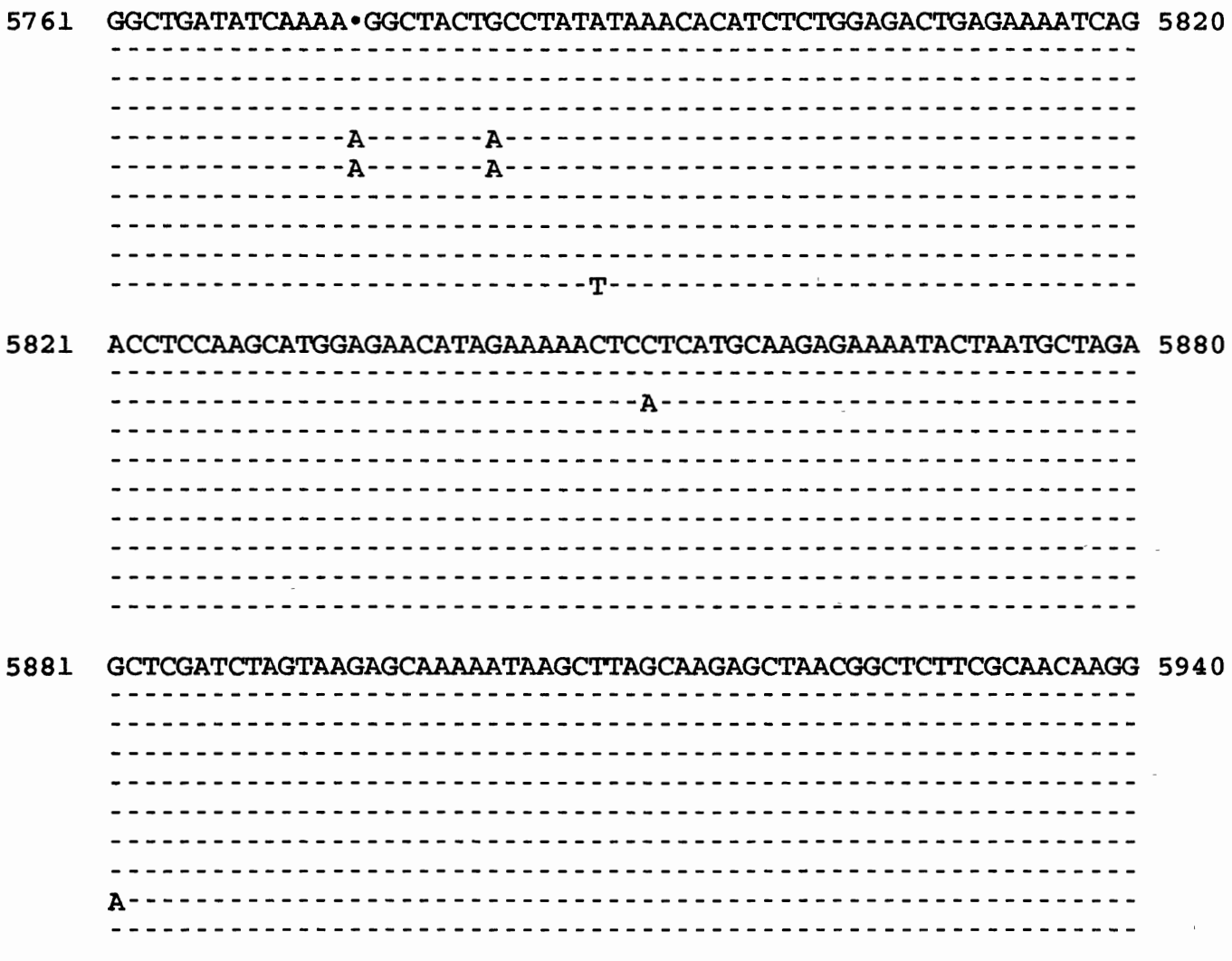
```
4861  GTGTGTTCAGAAAGTTCTGTTGCGTTTATGTCGACGACATTCTCGTATTCAGTAACAACG 4920
      ----------------------------------------G-------------------
      -----------------------------------T--------------------------
      ------------------------------------------------------------T-
      ----------A----------G--------------C---------------------
      ----------A----------G--------------C---------------------
      ------------------------------------------------------------T-
      ------------------------G-----------------------------------CT-
      ------------------------------------------------------------
      ------------------------------------------------------------

4921  AAGAAGATCACCTACTTCACGTAGCAATGATCTTACAAAAGTGCAATCAACATGGAATTA 4980
      ----------T----------------------------------------G----------
      ----------------T----------------------T----------------------
      -G-----------------------------------------------------------C-
      ------------------------------------------------------------C-
      ------------------------------------------------------------C-
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------------T-----G-----------------------------------C-
      ----------T----------------------------T----------------------

4981  TCCTTTCCAAGAAGAAAGCACAACTCTTCAAGAAGAAGATAAACTTCCTTGGTCTAGAAA 5040
      ------------------------------------------------------------
      ----------------G-------------------------------------------
      --------------------------A---------------------------------
      -T-----------------------------------------------------------
      -T-----------------------------------------------------------
      ------------------------------------------------------------
      ----------------------------------------T-------------------
      ----------A-------------------------------------------------
      ------------------------------------------------------------
```

```
5041  TAGATGAAGGAACACACAAGCCTCAAGGACATATCTTGGAACATATCAACAAATTCCCAG 5100
      ----------------T------------------T-----------------G-------
      ------------------------------------------------------------
      ----------------T-------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ---------------------------------C--------------------------
      ------------------------------------------C-----T-----------
      ----------------T--------------C-----------C---------G-------
      ----------------T----------------------------C---------G-----C-

5101  ATACCCTTGAAGACAAGAAGCAACTTCAGAGATTCTTAGGCATCCTAACATATGCCTCTG 5160
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------C-
      ------------------------------------------------------------
      ------------------------------------------------------------
      -------------T------------------------------A--C-----------A-
      ----------------A----------------------------A--------------G-
      ---------------------------------------------A--------------G-
      ---------------------------------------------A--------------G-

5161  ATTATATCCCGAAGCTAGCTCAAATCAGAAAGCCTCTGCAAGCCAAGCTTAAAGAAAATG 5220
      -------------T-----------G---C------------------------------
      -------T-------------------------------------------------G----
      ---------------------------T--------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      -------T-----------G---------------------------------G--G----
      ------------------------------------------------------------C-
      -C--------------C-------------------------------------------G----
      ----C-------------------------------------------------------C-
```

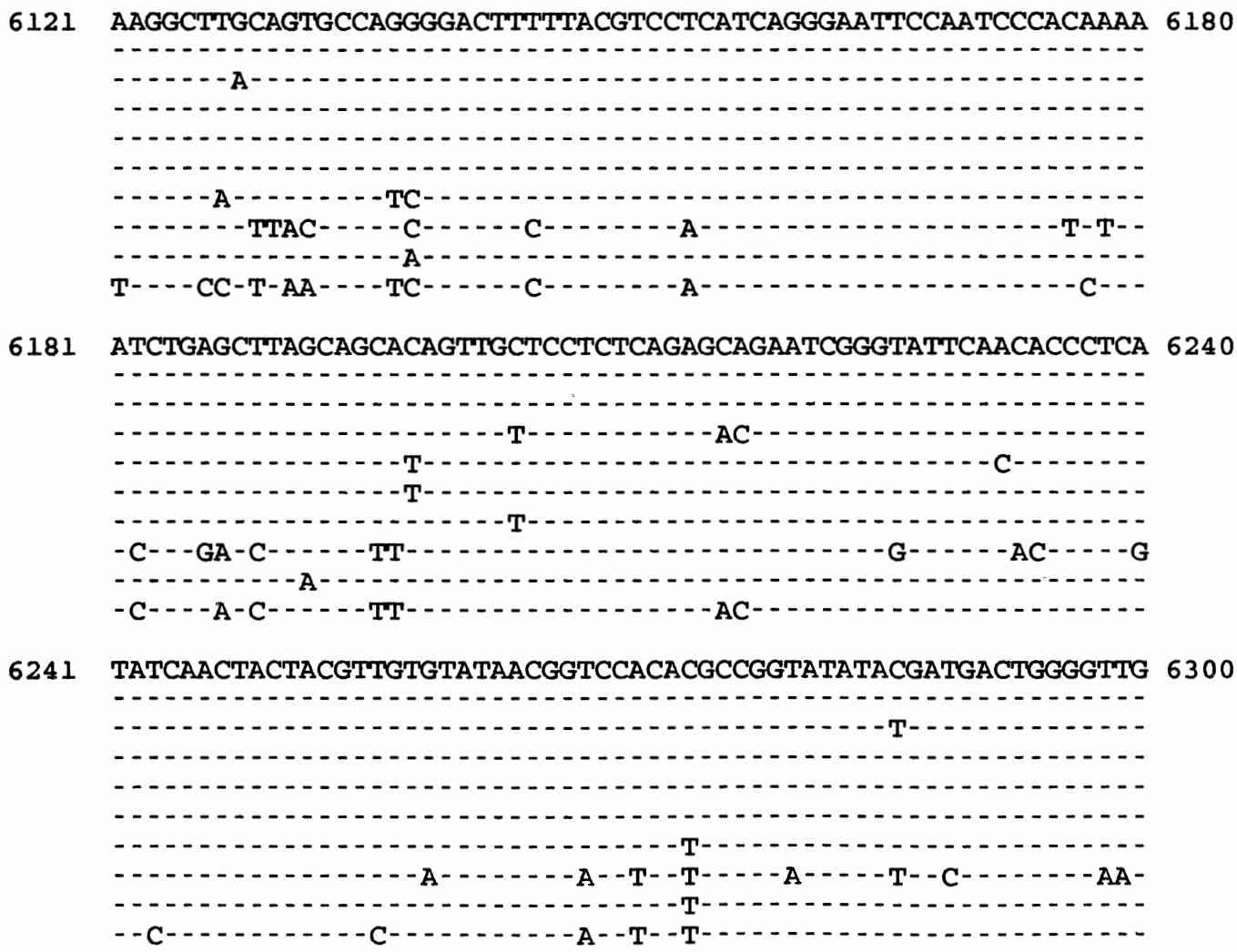
```
5221  TTCCATGGAAATGGACAAAAGAGGACACCCTCTACATGCAAAAGGTGAAGAAAAATCTGC 5280
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------G-----------------------------------------
      ------------------G-----------------------------------------
      ------------------------------------------------------------
      ------------------G-----------------------------------------
      ------------------G-----------------------------------------
      ---------G----------------T-------------------------------------

5281  AAGGATTTCCTCCACTACATCATCCCTTACCAGAGGAGAAGCTGATCATCGAGACCGACG 5340
      ----------------------------------A-----------------A-----T-
      ------------------------------------------------------------
      ---------------------------------A----------------------------
      ---------------------------------A----------------------------
      ---------------------------------A----------------------------
      ----------------------------------A------T----T----------------
      ---------------------G-----------A--A----------------------------
      ----------------------------------A----------------------------
      ------------------------------------------------------------T-

5341  CATCAGACGACTACTGGGGAGGTATGTTAAAAGCTATCAAAATTAACGAAGGTACTAATA 5400
      ------------------------------------------------------------
      -------T-----------------------------------------------T-----
      ------------------------------------------------A-------
      ------------------------------------------------------------
      ------------------------------------------------------------
      -------T-----------------------------------------------------
      -------T---------------------------------C--------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
```

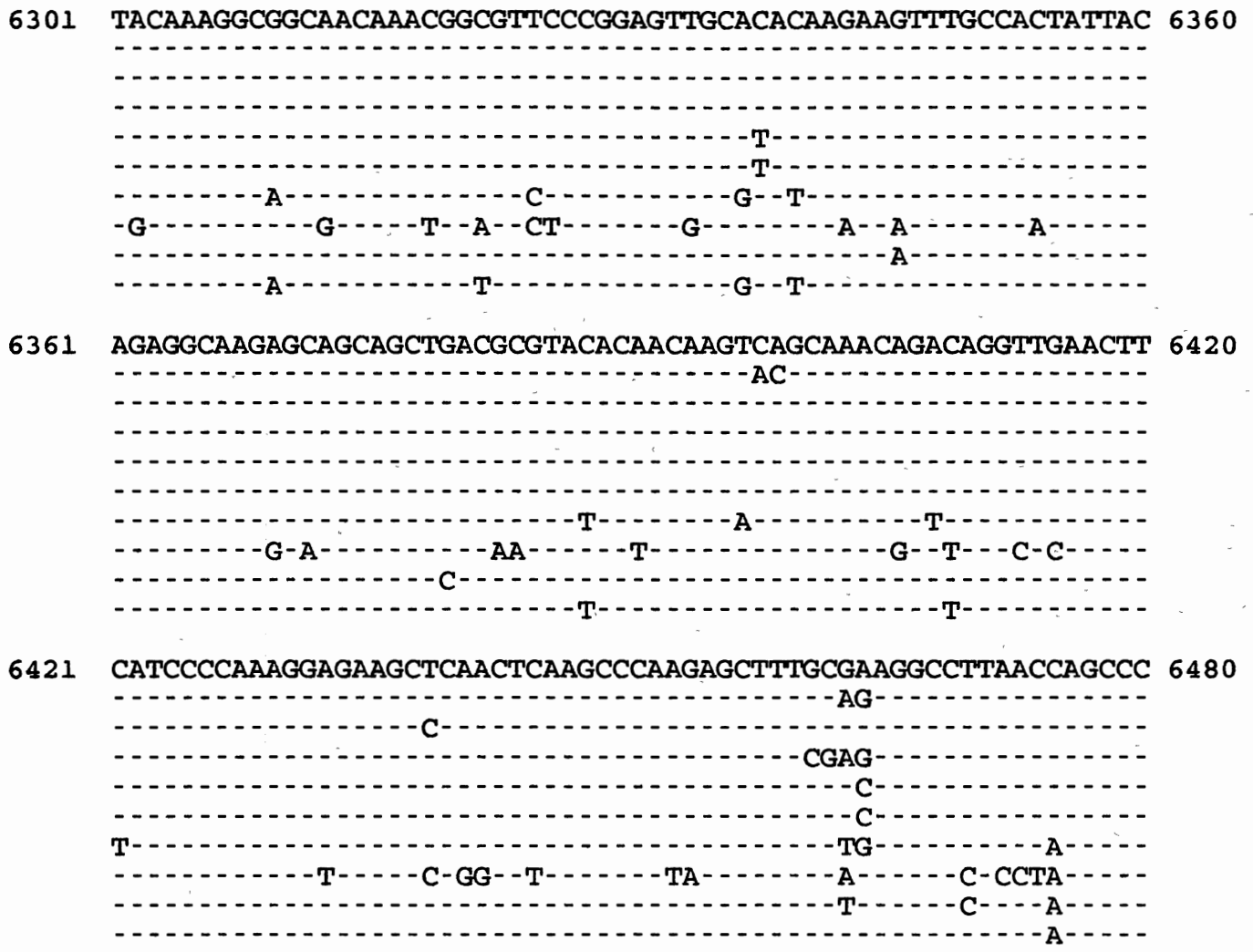
```
5401  CTGAGTTAATTTGCAGATACGCATCTGGAAGCTTTAAAGCTGCAGAAAAGAATTACCACA 5460
      --------------------CG---------------G----------G-----------
      ------------------------------------------------G-----------
      ------------------------------------------------G-----------
      -C----------------------------------------------G-----------
      -C----------------------------------------------G-----------
      ------------------------------------------------------------
      ----------------------------G-------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------

5461  GCAATGACAAAGAGACATTGGCGGTAATAAATACTATAAAGAAATTCAGTATTTATCTAA 5520
      ------------------------------------------------------------
      ----------------C-----------------------------------C-------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ----------------C-------------------------------T-------------
      ----------------------------------------------T-------------
      --------------------------------C---------------------------
      ----------------------------------------------T-------------

5521  CTCCTGTTCATTTTCTGATTAGGACAGATAATACTCATTTCAAGAGTTTTGTTAATCTCA 5580
      --------------------C-----------------------------------------
      ------C-------C----CC--A-----------------------------------C--T-
      ------------------------------------------------------------T-
      ------------------------------------------------------------C--T-
      ------------------------------------------------------------C--T-
      --------------------C--A----------------------------------------
      ------------------------------------------------C-----------
      -------------CT-A--C---------------G--------------------C--T-
      ------------------------------------------------C-----------
```

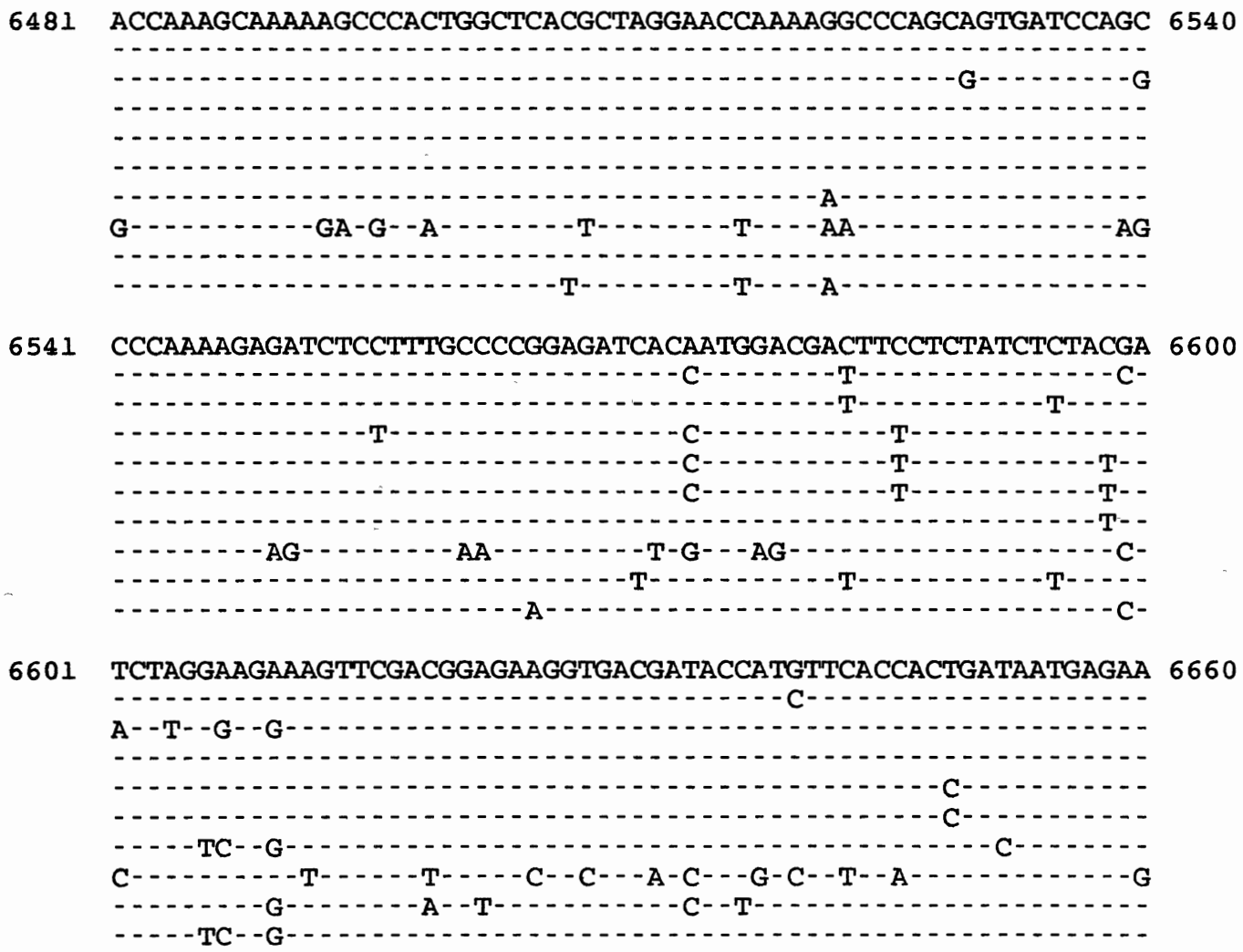
```
5581  ATTACAAAGGAGATTCAAAACTTGGAAGAAACATCAGATGGCAAGCATGGCTTAGCCACT 5640
      ----------T----------------------------------------------------
      ----------T-----------------------------------------------A-----
      ------------------------------------------------------------------
      ------------------------------------------------------------------
      ------------------------------------------------------------------
      ----------------G-----------------------------------------------T-
      ----------G-----G-------------------------------C-----------------
      ----------T------------------------------------------------------
      ----------------G-------------------------------------------------

5641  ATTCATTTGATGTTGAACATATTAAAGGAACCGACAACCACTTTGCGGACTTCCTTTCAA 5700
      ------------------------------------------------------------------
      ----------------------------------------------T-------------------
      ----G-------------------------------------------------------------
      ----G-------------------------------------------------------------
      ----G-------------------------------------------------------------
      ------------------C-----------------------------------------------
      ------------------C-------------------G---------------------------
      -C----------------C-----------------------------------------------
      ------------------C-----------------------------------------------

5701  GAGAATTCAATAAGGTTAATTCCTAATTGAAATCCGAAGATAAGATTCCCACACACTTGT 5760
      ------------------------------------------------------------------
      ------------------------G-----------------------------------------
      ----------C-------------------------------------------------------
      ------------------------------------------------------------------
      ------------------------------------------------------------------
      ------------G-----------------------------------------------------
      ------------G-----------------------A------------------------------
      ------------------------------------------------------------------
      ------------------------------------------------------------------
```

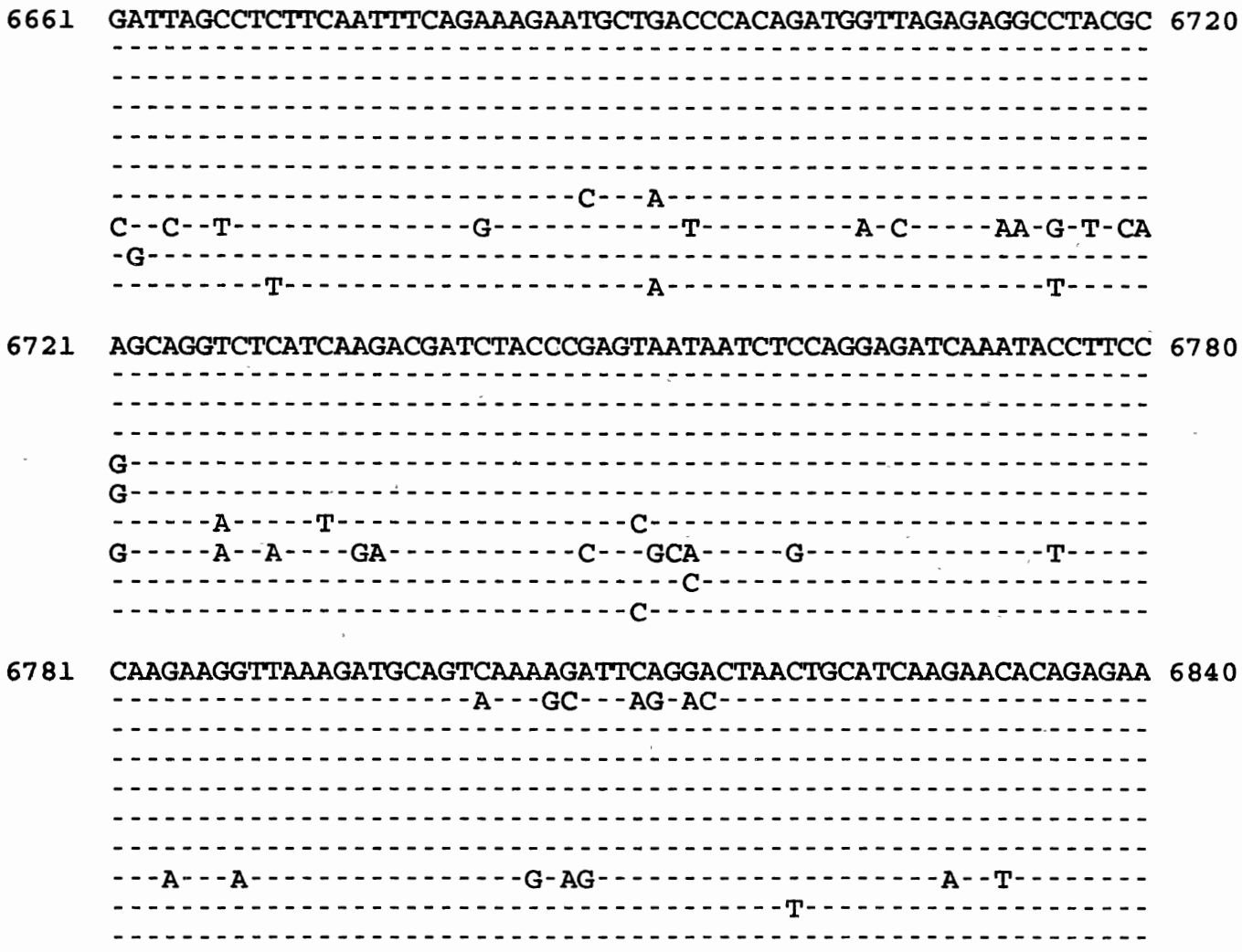

5761 GGCTGATATCAAAA•GGCTACTGCCTATATAAACACATCTCTGGAGACTGAGAAAATCAG 5820
5821 ACCTCCAAGCATGGAGAACATAGAAAAACTCCTCATGCAAGAGAAAATACTAATGCTAGA 5880
5881 GCTCGATCTAGTAAGAGCAAAAATAAGCTTAGCAAGAGCTAACGGCTCTTCGCAACAAGG 5940

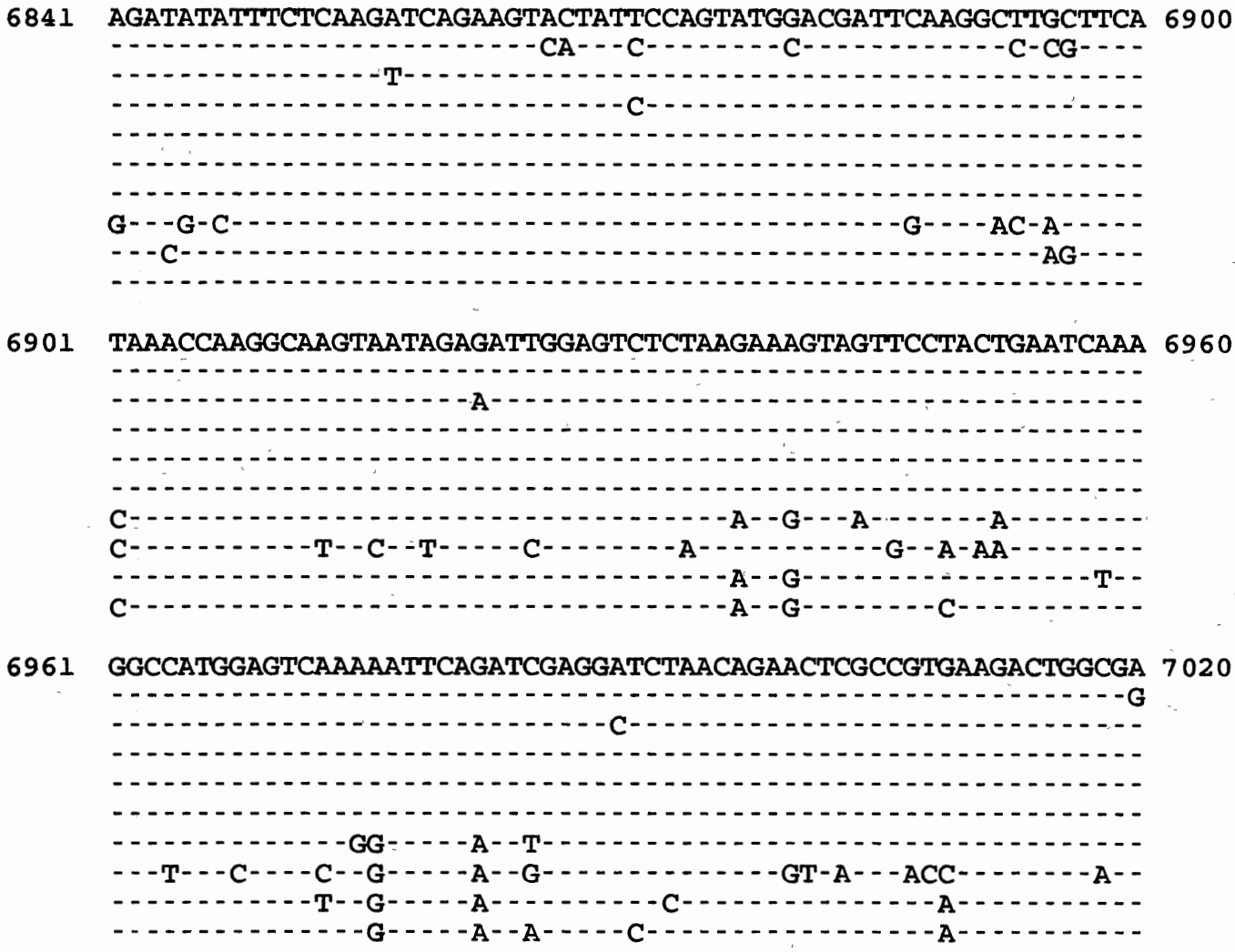
```
5941  AGACCTCTCTCTCCACCGTGAAACACCGGAAAAAGAAGAAGCAGTTCATTCTGCACTGGC 6000
      -------C-----------------------------------------------------
      -------C----------------------T-------------------------------
      ------------------------------T-------------------------------
      -------C----------------------T-------------------------------
      -------C----------------------T-------------------------------
      ---A-------------------------------------T---------------------T
      ---A-------------------------------------T---------------------
      -------C----------------------T-------------------------------
      -----------------------------------------------------------------

6001  CACTTTTACGCCAACTCAAGTAAAGGCTATTCCAGAGCAAACGGCTCCTGGTAAAGAATC 6060
      -------------------------A-------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      T----------------------A-------------------------------------
      -----------------------------------------------------------------
      T------------T----------A-------------------------------------

6061  AACAAATCCGTTGATGGCTAGTATCTTGCCAAAAGATATGAACCCAGTTCAAACTGGGAT 6120
      -----------------------------------------------------------------
      ----------------T------------------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      -----------------------------------------------------------------
      ------------------------------------------------G-------C
      ----------------T---------A--------------TT-G--A--G--C-AT-A
      --------------------A-------------------------------------------
      --------------------A---------------------TT-------G----AA--
```


6121 AAGGCTTGCAGTGCCAGGGGACTTTTTACGTCCTCATCAGGGAATTCCAATCCCACAAAA 6180
6181 ATCTGAGCTTAGCAGCACAGTTGCTCCTCTCAGAGCAGAATCGGGTATTCAACACCCTCA 6240
6241 TATCAACTACTACGTTGTGTATAACGGTCCACACGCCGGTATATACGATGACTGGGGTTG 6300

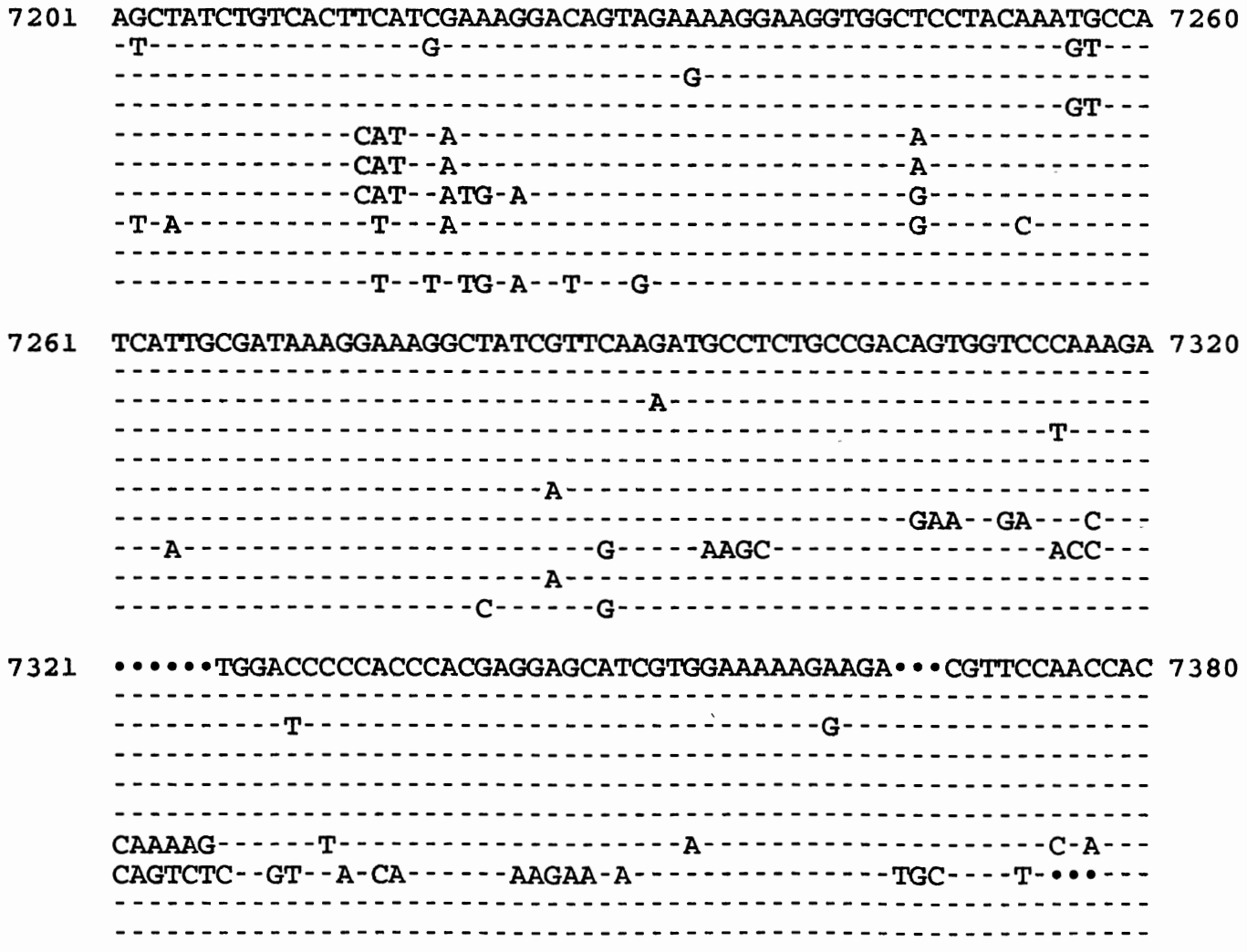

6301 TACAAAGGCGGCAACAAACGGCGTTCCCGGAGTTGCACACAAGAAGTTTGCCACTATTAC 6360
6361 AGAGGCAAGAGCAGCAGCTGACGCGTACACAACAAGTCAGCAAACAGACAGGTTGAACTT 6420
6421 CATCCCCAAAGGAGAAGCTCAACTCAAGCCCAAGAGCTTTGCGAAGGCCTTAACCAGCCC 6480


6481 ACCAAAGCAAAAAGCCCACTGGCTCACGCTAGGAACCAAAAGGCCCAGCAGTGATCCAGC 6540
6541 CCCAAAAGAGATCTCCTTTGCCCCGGAGATCACAATGGACGACTTCCTCTATCTCTACGA 6600
6601 TCTAGGAAGAAAGTTCGACGGAGAAGGTGACGATACCATGTTCACCACTGATAATGAGAA 6660


6661 GATTAGCCTCTTCAATTTCAGAAAGAATGCTGACCCACAGATGGTTAGAGAGGCCTACGC 6720
6721 AGCAGGTCTCATCAAGACGATCTACCCGAGTAATAATCTCCAGGAGATCAAATACCTTCC 6780
6781 CAAGAAGGTTAAAGATGCAGTCAAAAGATTCAGGACTAACTGCATCAAGAACACAGAGAA 6840


6841 AGATATATTTCTCAAGATCAGAAGTACTATTCCAGTATGGACGATTCAAGGCTTGCTTCA 6900
6901 TAAACCAAGGCAAGTAATAGAGATTGGAGTCTCTAAGAAAGTAGTTCCTACTGAATCAAA 6960
6961 GGCCATGGAGTCAAAAATTCAGATCGAGGATCTAACAGAACTCGCCGTGAAGACTGGCGA 7020

```
7021  ACAGTTCATACAGAGTCTTTTACGACTCAATGACAAGAAGAAAATCTTCGTCAACATGGT 7080
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------C-------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------C-------------------------------------------
      ---------------CT-GC-TAAG-----C--G-----T--G-------------------
      ------------------------------------------------------------
      ------------------C-------------------------------------------

7081  GGAGCACGACACTCTCGTCTACTCCAAGAATATCAAAGATACAGTCTCAGAAGACCAAAG 7140
      ------------------------------------------------------------
      ------------------------------------------------------------
      ---AG-T--------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------------------------------
      ------------------------------------G--A----A----------------
      A--A------GTGT-G--------A--A---C-A--G--A---ACTC-C-----T-----
      ---------------G-----------A---G----------------------------
      ------------G--T-----------A--------------------------------

7141  GGCTATTGAGACTTTTCAACAAAGGGTAATATCGGGAAACCTCCTCGGATTCCATTGCCC 7200
      ----------------------------------A---------------------------
      ------------------------------------------------------------
      ----------------------------------A---------------------------
      ------------------------------------------------------------
      ------------G-----------------------------------------------
      ---A-----------C--------------T-------------------------------
      ---A--------C-----GA----------T-------G---T--T---------------
      ------------------------------T-------------------------------
      ---A------------------------------C---------------------------
```


7201 AGCTATCTGTCACTTCATCGAAAGGACAGTAGAAAAGGAAGGTGGCTCCTACAAATGCCA 7260
-T----------G----------GT---
----------G----------
----------GT---
----------CAT--A----------A----------
----------CAT--A----------A----------
----------CAT--ATG-A----------G----------
-T-A----------T---A----------G-----C-------

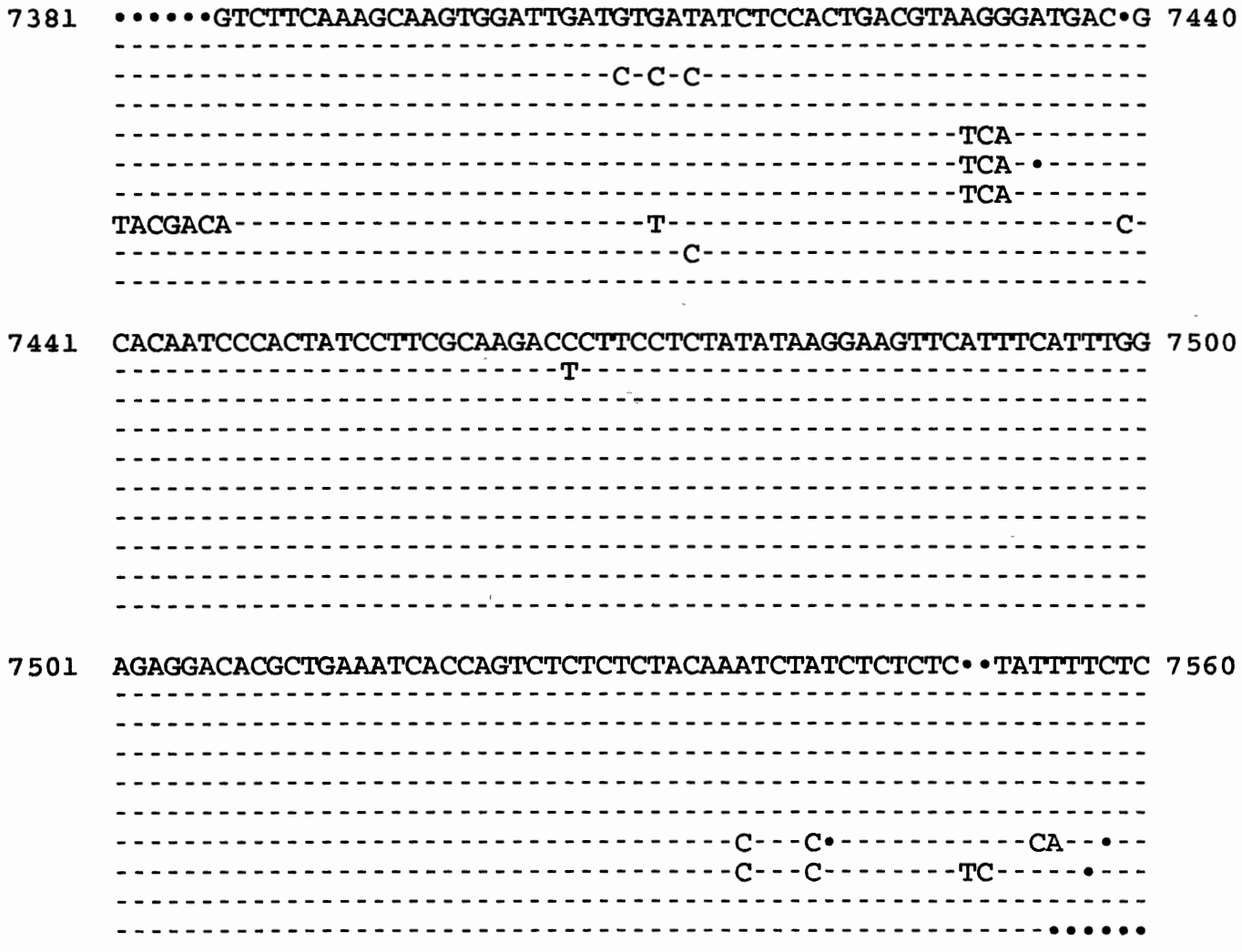
----------T--T-TG-A--T---G----------
7261 TCATTGCGATAAAGGAAAGGCTATCGTTCAAGATGCCTCTGCCGACAGTGGTCCCAAAGA 7320

----------A----------
----------T-----

----------A----------
----------GAA--GA---C---
---A----------G-----AAGC----------ACC---
----------A----------
----------C------G----------
7321 ••••••TGGACCCCCACCCACGAGGAGCATCGTGGAAAAAGAAGA•••CGTTCCAACCAC 7380

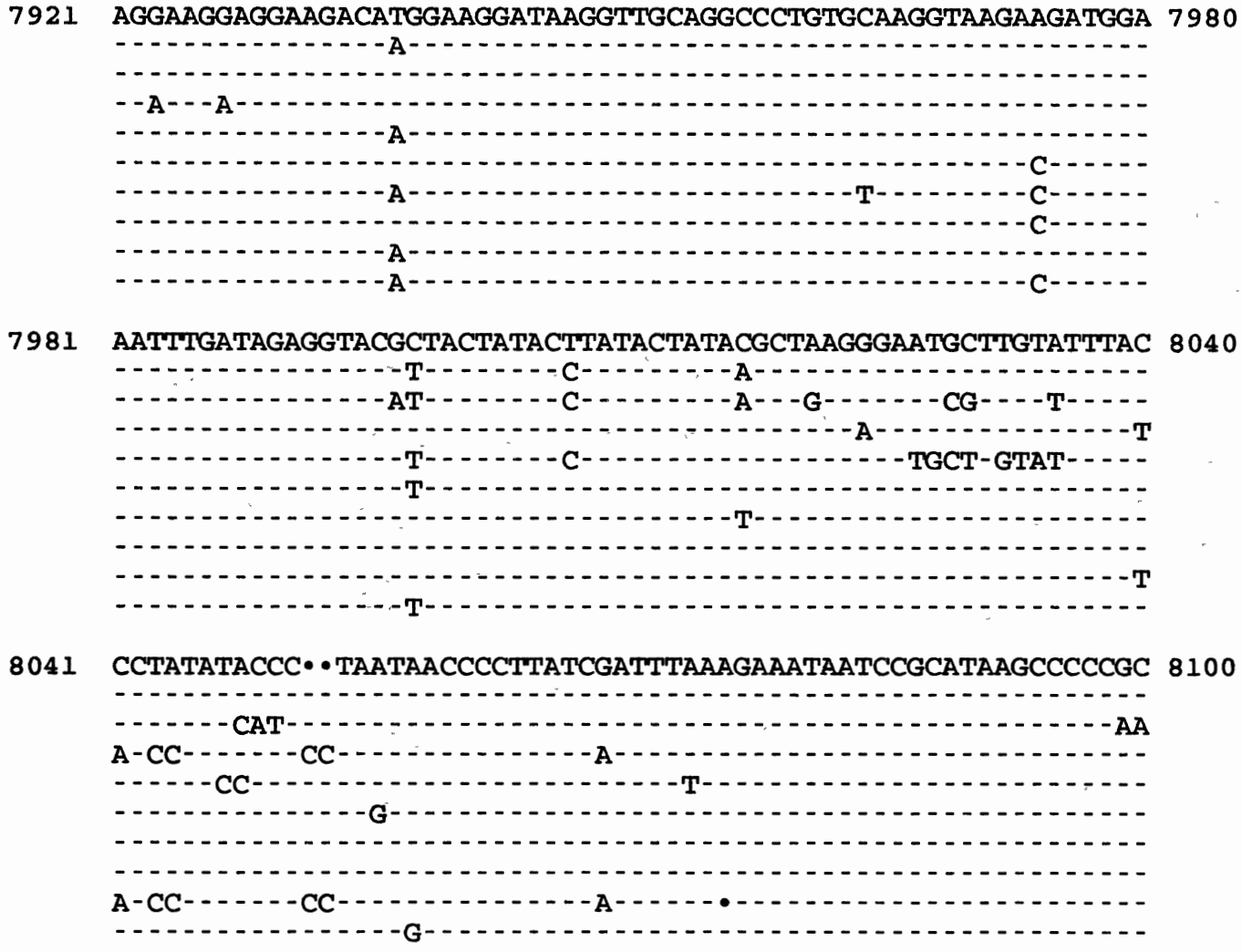
----------T----------G----------

CAAAAG------T----------A----------C-A---
CAGTCTC--GT--A-CA------AAGAA-A----------TGC----T-•••----


7381 ••••••GTCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGATGAC•G 7440
7441 CACAATCCCACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGG 7500
7501 AGAGGACACGCTGAAATCACCAGTCTCTCTCTACAAATCTATCTCTCTC••TATTTTCTC 7560


7561 CATAA•TAATGTGTGAGTAGTT•CCCAGATAAGGGAATTAGGGTTCTTATAGGGTTTCGC 7620
7621 TCATGTGTTGAGCATATAAGAAACCCTTAGTATGTATTTGTATTTGTAAAATACTTCTAT 7680
7681 CAATAAAATTTCTAATTCCTAAAACCAAAATCCAGTACTAAAATCCAGATCTCCTAAAGT 7740


7741 CCCTATAGATCTTTGTGGTGAATATAAACCAGACACGAGACGACTAAACCTGGAGCCCAG 7800
7801 ACGCCGTTTGAAGCTAGAAGTACCGCTTAGGCAGGAGGCCGTTAGGGAAAAGATGCTAAG 7860
7861 GCAGGGTTGGTTACGTTGACTCCCCCGTAGGTTTGGTTTAAATATCATGAAGTGGACGGA 7920


7921 AGGAAGGAGGAAGACATGGAAGGATAAGGTTGCAGGCCCTGTGCAAGGTAAGAAGATGGA 7980
7981 AATTTGATAGAGGTACGCTACTATACTTATACTATACGCTAAGGGAATGCTTGTATTTAC 8040
8041 CCTATATACCC••TAATAACCCCTTATCGATTTAAAGAAATAATCCGCATAAGCCCCCGC 8100

```
8101  TTAAAAAATT 8110
      ----------
      AC--------
      ----------
      ----------
      ----------
      ----------
      ----------
      --------•--
      ----------
      ----------
```

VITA

Kelly Dawn Chenault

Candidate for the Degree of

Doctor of Philosophy

Thesis: VARIATION AND EVOLUTION OF CAULIFLOWER MOSAIC VIRUS ISOLATES

Major Field: Biochemistry

Biographical:

Personal Data: Born in Stillwater, Oklahoma, May 10, 1965, the daughter of Dr. Robert C. and Beverly J. Hooper; married in 1991 to Paul D. Chenault.

Education: Graduated from Temple High School, Temple, Oklahoma, May, 1983; received Bachelor of Science Degree in Biochemistry from Oklahoma State University, Stillwater, Oklahoma, May 1987; completed requirements for Doctor of Philosophy Degree in Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma, July, 1992.

Professional Experience: undergraduate research assistant summers of 1985-1986, Noble Research Foundation, Ardmore, Oklahoma; undergraduate research assistant 1986-1987, Department of Biochemistry, Oklahoma State University, Stillwater, Oklahoma; graduate research and teaching assistant 1987-1992, Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma.