Protease Evolution in *Streptomyces griseus*

DISCOVERY OF A NOVEL DIMERIC ENZYME*

Sachdev S. Sidhu‡, Gabriel B. Kalmar, Leslie G. Willis, and Thor J. Borgford§

From the Department of Chemistry and Institute of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada

This report describes the cloning and sequencing of a novel protease gene derived from *Streptomyces griseus*. Also described is the heterologous expression of the gene in *Bacillus subtilis* and characterization of the gene product. The *sprD* gene encodes a prepro mature protease of 392 amino acids tentatively named *S. griseus* protease D (SGPD). A significant component of the enzyme preregion was found to be homologous with the mitochondrial import signal of hsp60. The *sprD* gene was subcloned into an *Escherichia coli*/*B. subtilis* shuttle vector system such that the pro mature portion of SGPD was fused in frame with the promoter, ribosome binding site, and signal sequences of subtilisin. The gene fusion was subsequently expressed in *B. subtilis* DB104, and active protease was purified. SGPD has a high degree of sequence homology to previously described *S. griseus* proteases A, B, C, and E and the α-lytic protease of *Lysobacter enzymogenes*, but unlike all previously characterized members of the chymotrypsin superfamily, the recombinant SGPD forms a stable α_2 dimer. The amino acid sequence of the protein in the region of the specificity pocket is similar to that of *S. griseus* proteases A, B, and C. The purified enzyme was found to have a primary specificity for large aliphatic or aromatic amino acids. Nucleotide sequence data were used to construct a phylogenetic tree using a method of maximum parsimony which reflects the relationships and potentially the lineage of the chymotrypsin-like proteases of *S. griseus*.

Serine proteases catalyze the hydrolysis of amides and esters by a common catalytic mechanism involving a triad of the residues serine, histidine and aspartic acid. Beyond mechanism, this family of enzymes has two branches that are differentiated from one another by the type of protein fold. One branch is comprised of enzymes which have a subtilisin-like tertiary structure; the other branch has a chymotrypsin-like tertiary structure. It is commonly believed that the two branches of the family evolved independently and converged upon the same catalytic mechanism (1).

In terms of function, proteases of the chymotrypsin superfamily are an extraordinarily divergent group of enzymes. The

group encompasses enzymes involved in mammalian blood clotting cascades, digestive enzymes of the pancreas (2), enzymes involved in the regulation of the cell cycle (3), and enzymes involved in the maturation and secretion of other proteins (4). In a previous study (5), we isolated two genes of the organism *Streptomyces griseus* by virtue of their genetic homology to the chymotrypsin-like *S. griseus* protease B (SGPB).[1] In that study the sequence and preliminary characterization of one of two enzymes, designated *S. griseus* protease C (SGPC), was presented. SGPC was found to have a primary specificity for large aliphatic or aromatic amino acids and, remarkably, possessed a carboxyl-terminal domain with homology to chitin-binding domains of certain chitinases. We now present the sequence and preliminary characterization of the gene encoding a second enzyme, tentatively named *S. griseus* protease D (SGPD). This enzyme also has a primary specificity for substrates with large aliphatic and aromatic side chains, but has an exceptional quaternary structure. Unlike any known protease of the chymotrypsin superfamily, SGPD is a stable dimer and, unlike the known homologues found in *S. griseus* and *Lysobacter enzymogenes*, SGPD has an acidic isoelectric point.

MATERIALS AND METHODS

Restriction endonucleases and DNA modifying enzymes were purchased from either New England Biolabs or Life Technologies, Inc., with the exception of T7 DNA polymerase from Pharmacia Biotech Inc. and calf intestinal phosphatase (CIP) from Boehringer Mannheim. All chemicals and reagents were of the highest grade commercially available.

Cloning of the Gene Encoding SGPD—As described in a previous report (5), a library was prepared from a *Bam*HI digest of *S. griseus* genomic DNA. The library was probed with a fragment of the SGPB gene (B-mat probe-encoding amino acids 9–185 of the mature enzyme) (6) which had been radiolabeled with [^{32}P]ATP (7). Four protease genes were cloned on the basis of their homology to the probe. One plasmid derived from the genomic library contained a strongly hybridizing insert of 2.3 kbp. This plasmid, carrying the *S. griseus* protease D gene, was designated pDS-D.

DNA Sequencing—Restriction fragments of pDS-D were selected for sequencing on the basis of Southern blot hybridization with the B-mat probe. Subclones were sequenced by a combination of manual and automated methods as described elsewhere (7).

SGPD Expression in Bacillus subtilis—SGPD was expressed in *B. subtilis* using a secretion expression vector (pEB11) previously used to express the enzymes SGPB, SGPC (5), SGPE (7), SGPA, and α-lytic protease.[2] To amplify a fragment of *sprD* encoding the promature portion of SGPD, two oligonucleotides were synthesized based on the sequence of *sprD*: DF1, (5′ oligonucleotide), 5′-AGTACTAGTGACGAT-GTACCG-3′, DR1, (3′ oligonucleotide), 5′-GAATTCGCTCCGGCCGGT-3′. DF1 and DR1 were used as polymerase chain reaction primers with

[1] The abbreviations used: SGPA, SGPB, SGPC, SGPD, and SGPE, *Streptomyces griseus* proteases A, B, C, D, and E, respectively; SGT, *Streptomyces griseus* trypsin; CIP, calf intestinal phosphatase; kbp, kilobase pair(s); PAGE, polyacrylamide gel electrophoresis.

[2] S. S. Sidhu and T. J. Borgford, unpublished results.

pDS-D as template. The amplified product (a fragment of *sprD* encoding for promature SGPD) was digested with *Eco*RI and ligated into pUC18 (8) which had been digested with *Eco*RI and *Sma*I and treated with CIP. The polymerase chain reaction product (0.2 μg) and vector (0.2 μg) were ligated in a final volume of 10 μl, and the ligation mixture was used to transform *Escherichia coli* DH5α/P3. Vectors with inserts of the correct size were identified by restriction analysis. One such plasmid (designated pDS-D8) was partially sequenced to verify accurate amplification and ligation. pDS-D8 was then digested with *Eco*RI and *Sca*I and treated with T4 DNA polymerase to produce blunt-ended fragments (9). The fragment containing the pro mature portion of *sprD* was gel-purified and ligated into pEB11 (7) (digested with *Sma*I and treated with CIP). *E. coli* DH5α/P3 was transformed with the ligation mixture, and a vector containing the correct insert in the correct orientation was isolated on the basis of restriction enzyme analysis. This vector was designated pEB-D8. *B. subtilis* DB104 (10) was transformed with pEB-D8, and transformants containing the correct plasmid were identified using restriction enzyme digests. Expression and secretion of SGPD in *B. subtilis* DB104 transformants was verified using a skim milk clearing assay (11).

Media and Growth Conditions—*E. coli* and *B. subtilis* transformants were maintained on plates and in broth cultures as described previously (5, 7). For the purposes of SGPD expression and purification, *B. subtilis* DB104 transformants harboring the plasmid pEB-D8 were expressed in YTC medium (5) containing 20 μg/ml of kanamycin and 10 mM CaCl₂. Cultures were grown at 30 °C in a 20-liter fermentor according to protocols originally designed for the expression of the enzyme SGPC (5).

Purification of SGPD—Bacteria were removed from liquid cultures by ultrafiltration using a Millipore Pellicon apparatus equipped with a HVMP membrane cassette (0.45-μm cutoff). The filtrate (containing SGPD) was next concentrated to 1.0 liter with a PTGC membrane cassette (10,000 nominal molecular weight limit). The retentate was centrifuged for 30 min at 10,000 × g to remove any additional precipitate. Sodium acetate (3.0 M, pH 4.8) was added to the concentrated retentate (containing SGPD) to a final concentration of 100 mM.

Acetone was added to the retentate with stirring to a final concentration of 60% (v/v). After stirring for 10 min, the mixture was centrifuged at 4,000 × g for 15 min, and the pellet was discarded. Acetone was added to the supernatant to a final concentration of 75% (v/v), and the mixture was again stirred and centrifuged as above. The pellet from this second fractionation was resuspended in 150 ml of 100 mM sodium phosphate (pH 7.0). Proteolytic activity was monitored during all fractionations.

The sample was applied to a 60 × 3-cm S-Sepharose cation exchange column (Pharmacia) equilibrated with 10 mM sodium phosphate, pH 7.0 (buffer A), in order to remove cationic contaminants. The column was washed with the same buffer, and the flow-through was collected in 25-ml fractions. Fractions with activity toward *N*-succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide were pooled and dialyzed against 10 mM Tris, pH 8.0 (buffer B), overnight at 4 °C.

The dialyzed sample was applied to a Pharmacia Mono-Q anion exchange column using a Pharmacia fast protein liquid chromatography system. The column was washed with buffer B until A_{280} baselined. The enzyme was then eluted in a salt gradient from 0 to 0.25 M NaCl in buffer B in 60 min. The proteolytic activity of the recombinant enzyme was monitored during all purification steps with the chromogenic substrate *N*-succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide (12) as described previously (5). Fractions with activity toward *N*-succinyl-Ala-Ala-Pro-Phe-*p*-nitroanilide were analyzed by SDS-PAGE in 12% gels (13), and those fractions exhibiting a single 34-kDa band were pooled. Protein concentrations were determined using the method of Lowry (14). The amino-terminal sequence of the purified protein was determined using an Applied Biosystems model 473 protein sequencer at the Microsequencing Center of the University of Victoria, British Columbia, Canada.

Primary Specificity—The specific activities of SGPB and SGPD were compared in assays that made use of peptide substrates having the sequence *N*-succinyl-Ala-Ala-Pro-*X*-*p*-nitroanilide (where *X* was one of Phe, Met, Leu, Ala, Val, Ile, or Glu). The chromogenic substrates were purchased from Bachem and Sigma. Briefly, assays were performed in 50 mM Tris, pH 8.0, containing 5% methanol (v/v), and the release of *p*-nitroaniline was monitored spectrophotometrically at 412 nm. One unit of activity is defined as the amount of enzyme required to produce 1 mmol of *p*-nitroaniline in 1 h at 20 °C.

Gel Filtration Chromatography—Size exclusion chromatography was conducted on a Pharmacia fast protein liquid chromatography system equipped with a Superose-12 gel filtration column (Pharmacia). The column was equilibrated with 50 mM Tris, pH 8.0. Typically 100-μl samples (approximately 10 μg of protein) were loaded, and the column

was eluted with 50 mM Tris, pH 8.0, at a flow rate of 0.3 ml/min. Proteins eluting from the column were detected by absorbance at 280 nm.

SDS-PAGE—Protein samples for analysis by SDS-PAGE were prepared in one of two ways prior to electrophoresis, 1) "standard denaturing conditions" in which samples were heated for 10 min at 100 °C in 30 mM Tris, 10% glycerol, 4.4% β-mercaptoethanol, and 2% SDS, pH 6.8, and 2) "nondenaturing conditions" in which samples were prepared in 30 mM Tris, 10% glycerol, pH 6.8, without heating.

Homology Searches—Homologous DNA and protein sequences were searched for and identified using the BLAST network service at the National Center for Biotechnology Information (NCBI).

Sequence Alignment and Phylogeny—Nucleotide sequence data corresponding to the mature regions of the proteases were examined using the DNA parsimony program (DNApars) of Felsenstein (15). Amino acid sequences were first aligned, taking into consideration the known three-dimensional structures of SGPA (16), SGPB (17), SGPE (18), and the α-lytic protease (19) during the process of alignment. Subsequently, the nucleotide sequences were aligned according to the correspondence of amino acid sequences. Confidence limits were also estimated using the bootstrap program (DNAboot) of Felsenstein (15).

RESULTS

Cloning and Sequencing of sprD—In a previous study (5), we detected a 2.3-kbp *S. griseus* genomic DNA *Bam*HI fragment which hybridized strongly with a DNA fragment encoding amino acids 9–185 of mature SGPB (B-mat). In this report, the B-mat DNA probe was used to isolate plasmids containing the 2.3-kbp DNA fragment from a DNA library prepared from *S. griseus* genomic DNA digested to completion with *Bam*HI. Approximately 1×10^4 *E. coli* transformants were screened by colony hybridization. Strongly hybridizing clones were then selected for further analysis. Southern blot analysis of plasmid DNA isolated from these clones revealed nine that contained a 2.3-kbp insert which hybridized strongly with the B-mat probe. One of these plasmids was designated pDS-D and was chosen for sequencing.

The 2.3-kbp insert of pDS-D contains a gene, designated *sprD*, which encodes a polypeptide of 392 amino acids (Fig. 1). The organization of the open reading frame is analogous to that of the previously characterized *S. griseus* proteases A, B (6), E (7), and C (5) and the α-lytic protease of *L. enzymogenes* (20). On the basis of sequence alignments with the open reading frames of these protease genes, we concluded that *sprD* encodes a prepro mature form of an uncharacterized serine protease which we designated *S. griseus* protease D (SGPD).

The prepro and pro mature junctions shown in Fig. 1 were initially assigned on the basis of sequence alignments with the other *S. griseus* and *L. enzymogenes* serine proteases (amino-terminal analysis of the mature enzyme confirmed the location of the junction between the pro and mature portions of the polypeptide; see below). Mature SGPD, encompassing the final 188 amino acids of the open reading frame, is preceded by a leader peptide of 204 amino acids. The amino-terminal 64 residues of this leader peptide constitute a pre-peptide while the remaining 140 residues form the propeptide.

The 64-residue preregion of SGPD is significantly longer than the preregions of the other proteases which range in length from 29 to 40 amino acids. The 40-residue carboxyl-terminal segment of the SGPD preregion is characteristic of bacterial secretion signals (21) and shares significant homology with the preregion of SGPB (Fig. 2). The 24 amino-terminal residues form an amino-terminal extension not present in the other proteases. Interestingly, a computer search of the complete nonredundant DNA/protein data base revealed that this region shares significant homology with the mitochondrial signal sequence of hsp60 (22, 23) (Fig. 2). Moreover, the predictive method of Gavel and von Heijne (24) revealed that residues 1–44 comprise a sequence which is consistent with mitochondrial import signals, and residues 24–40 have the potential of
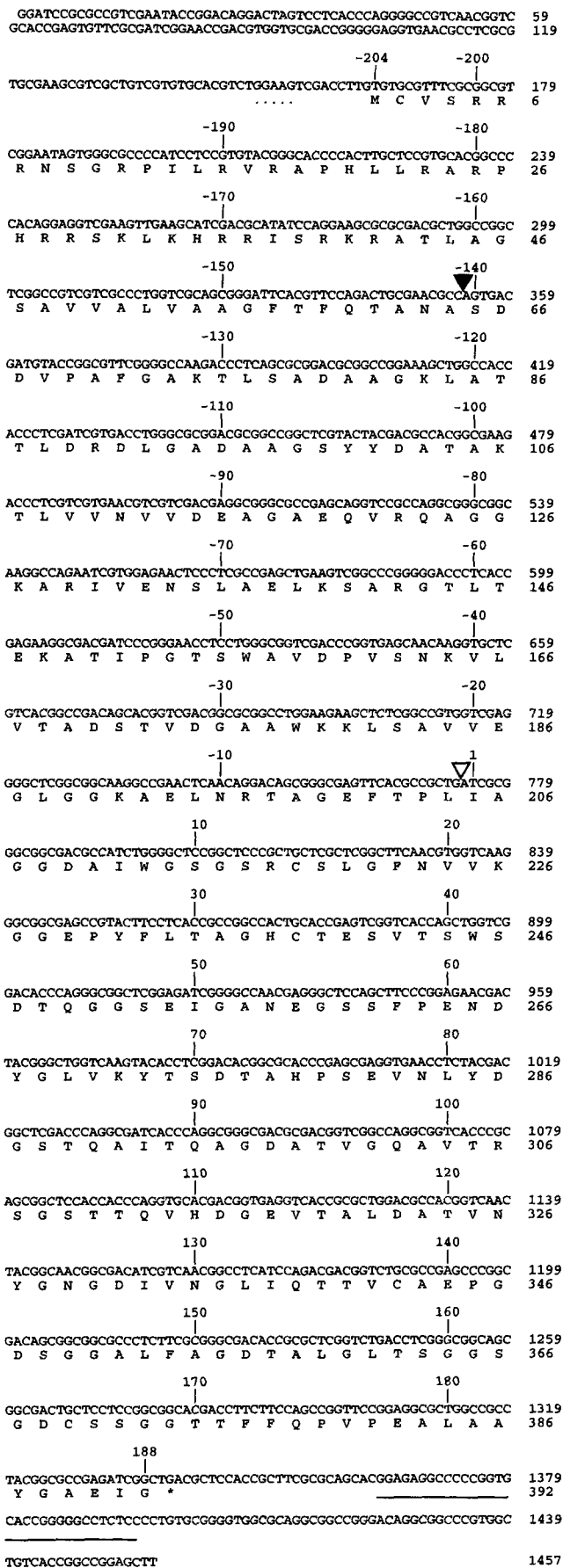
```
GGATCCGCGCCGTCGAATACCGGACAGGACTAGTCCTCACCCAGGGGCCGTCAACGGTC  59
GCACCGAGTGTTCGCGATCGGAACCGACGTGGTGCGACCGGGGGAGGTGAACGCCTCGCG  119

                         -204        -200
                          |           |
TGCGAAGCGTCGCTGTCGTGTGCACGTCTGGAAGTCGACCTTGTGTGCGTTTCGCGGCGT  179
                                   M   C   V   S   R   R       6

         -190                       -180
          |                          |
CGGAATAGTGGGCGCCCATCCTCCGTGTACGGGCACCCCACTTGCTCCGTGCACGGCCC  239
R   N   S   G   R   P   I   L   R   V   R   A   P   H   L   L   R   A   R   P   26

       -170                         -160
        |                            |
CACAGGAGGTCGAAGTTGAAGCATCGACGCATATCCAGGAAGCGCGCGACGCTGGCCGGC  299
H   R   R   S   K   L   K   H   R   R   I   S   R   K   R   A   T   L   A   G   46

         -150                  -140
          |                      ▼|
TCGGCCGTCGTCGCCCTGGTCGCAGCGGGATTCACGTTCCAGACTGCGAACGCCAGTGAC  359
S   A   V   V   A   L   V   A   A   G   F   T   F   Q   T   A   N   A   S   D   66

           -130                     -120
            |                         |
GATGTACCGGCGTTCGGGGCCAAGACCCTCAGCGCGGACGCGGCCGGAAAGCTGGCCACC  419
D   V   P   A   F   G   A   K   T   L   S   A   D   A   A   G   K   L   A   T   86

         -110                      -100
          |                          |
ACCCTCGATCGTGACCTGGGCGCGGACGCGGCCGGCTCGTACTACGACGCCACGGCGAAG  479
T   L   D   R   D   L   G   A   D   A   A   G   S   Y   Y   D   A   T   A   K   106

       -90                         -80
        |                           |
ACCCTCGTCGTGAACGTCGTCGACGAGGCGGGCGCCGAGCAGGTCCGCCAGGCGGGCGGC  539
T   L   V   V   N   V   V   D   E   A   G   A   E   Q   V   R   Q   A   G   G   126

       -70                         -60
        |                           |
AAGGCCAGAATCGTGGAGAACTCCCTCGCCGAGCTGAAGTCGGCCCGGGGGACCCTCACC  599
K   A   R   I   V   E   N   S   L   A   E   L   K   S   A   R   G   T   L   T   146

       -50                       -40
        |                         |
GAGAAGGCGACGATCCCGGGAACCTCCTGGGCGGTCGACCCGGTGAGCAACAAGGTGCTC  659
E   K   A   T   I   P   G   T   S   W   A   V   D   P   V   S   N   K   V   L   166

     -30                       -20
      |                         |
GTCACGGCCGACAGCACGGTCGACGGCGCGGCCTGGAAGAAGCTCTCGGCCGTGGTCGAG  719
V   T   A   D   S   T   V   D   G   A   A   W   K   K   L   S   A   V   V   E   186

       -10                      1
        |                        ▽|
GGGCTCGGCGGCAAGGCCGAACTCAACAGGACAGCGGGCGAGTTCACGCCGCTGATCGCG  779
G   L   G   G   K   A   E   L   N   R   T   A   G   E   F   T   P   L   I   A   206

         10                      20
          |                       |
GGCGGCGACGCCATCTGGGGCTCCGGCTCCCGCTGCTCGCTCGGCTTCAACGTGGTCAAG  839
G   G   D   A   I   W   G   S   G   S   R   C   S   L   G   F   N   V   V   K   226

         30                      40
          |                       |
GGCGGCGAGCCGTACTTCCTCACCGCCGGCCACTGCACCGAGTCGGTCACCAGCTGGTCG  899
G   G   E   P   Y   F   L   T   A   G   H   C   T   E   S   V   T   S   W   S   246

         50                      60
          |                       |
GACACCCAGGGCGGCTCGGAGATCGGGGCCAACGAGGGCTCCAGCTTCCCGGAGAACGAC  959
D   T   Q   G   G   S   E   I   G   A   N   E   G   S   S   F   P   E   N   D   266

         70                      80
          |                       |
TACGGGCTGGTCAAGTACACCTCGGACACGGCGCACCCGAGCGAGGTGAACCTCTACGAC  1019
Y   G   L   V   K   Y   T   S   D   T   A   H   P   S   E   V   N   L   Y   D   286

         90                      100
          |                       |
GGCTCGACCCAGGCGATCACCCAGGCGGGCGACGCGACGGTCGGCCAGGCGGTCACCCGC  1079
G   S   T   Q   A   I   T   Q   A   G   D   A   T   V   G   Q   A   V   T   R   306

         110                     120
          |                       |
AGCGGCTCCACCACCCAGGTGCACGACGGTGAGGTCACCGCGCTGGACGCCACGGTCAAC  1139
S   G   S   T   T   Q   V   H   D   G   E   V   T   A   L   D   A   T   V   N   326

         130                     140
          |                       |
TACGGCAACGGCGACATCGTCAACGGCCTCATCCAGACGACGGTCTGCGCCGAGCCCGGC  1199
Y   G   N   G   D   I   V   N   G   L   I   Q   T   T   V   C   A   E   P   G   346

         150                     160
          |                       |
GACAGCGGCGGCGCCCTCTTCGCGGGCGACACCGCGCTCGGTCTGACCTCGGGCGGCAGC  1259
D   S   G   G   A   L   F   A   G   D   T   A   L   G   L   T   S   G   G   S   366

         170                     180
          |                       |
GGCGACTGCTCCTCCGGCGGCACGACCTTCTTCCAGCCGGTTCCGGAGGCGCTGGCCGCC  1319
G   D   C   S   S   G   G   T   T   F   F   Q   P   V   P   E   A   L   A   A   386

         188
          |
TACGGCGCCGAGATCGGCTGACGCTCCACCGCTTCGCGCAGCACGGAGAGGCCCCCGGTG  1379
Y   G   A   E   I   G   *                                        392

CACCGGGGGCCTCTCCCCTGTGCGGGGTGGCGCAGGCGGCCGGGACAGGCGGCCCGTGGC  1439

TGTCACCGGCCGGAGCTT                                            1457
```

FIG. 1. **Nucleotide sequence of the *sprD* gene and the deduced amino acid sequence of SGPD.** The numbering to the *right* of the

forming an amphipathic α-helix with one face highly positively charged, a motif considered essential for translocation of proteins into mitochondria (25, 26).

The 5'-untranslated region of *sprD* contains a putative ribosome binding site that was identified by comparison with other *Streptomyces* gene sequences (6, 27). The translation stop codon is followed by an inverted repeated sequence capable of forming a stable hairpin loop (Fig. 1). Such structures, believed to be involved in transcription termination, have been identified in other *Streptomyces* genes (6, 27).

Protease Expression in B. subtilis—The expression vector pEB-D8 contains an open reading frame encoding a fusion protein composed of the preregion of subtilisin connected in-frame to pro mature SGPD (residues 65–392) by a dipeptide sequence (Pro-Thr). Transcription and translation of the open reading frame are initiated from the subtilisin BPN' promoter and ribosome binding site, respectively, and secretion of the translated polypeptide is directed by the subtilisin preregion (7). Secretion of active SGPD was evidenced by zones of clearing (due to milk protein hydrolysis) around transformants harboring pEB-D8.

Substrate Specificity—The final purified yield of recombinant SGPD expressed in *B. subtilis* was 5 mg/liter. Fig. 3 illustrates the relative activities of SGPB and SGPD toward a series of substrates that vary solely in the residue at the P1 position of the peptide. The two enzymes were found to have very similar substrate specificities and, notably, neither enzyme was able to hydrolyze a substrate containing glutamic acid in the P1 position.

Quanternary Structure of SGPD—Purified SGPE, SGPC (5), and SGPD were chromatographed separately on a size exclusion column; each sample produced a distinct single peak (data not shown), and the following retention times were determined: SGPE (54.5 min), SGPC (51 min), and SGPD (48 min). Since retention time is inversely proportional to molecular mass, these results suggest that native SGPE and SGPC exist as monomers with molecular masses of 18 and 26 kDa, respectively, while native SGPD exists as a homodimer with a molecular mass of 36 kDa.

SDS-polyacrylamide gel electrophoresis of SGPC and SGPD provided further evidence that SGPD exists as a very stable homodimer. Under standard denaturing conditions (see "Materials and Methods"), SGPC showed a single band corresponding to a molecular mass of 26 kDa, in good agreement with the monomeric molecular mass predicted from the DNA sequence of *sprC* (5). Under the same conditions, SGPD (which exhibited a single peak by gel filtration chromatography) resolved into two distinct bands with apparent molecular masses of approximately 17 and 34 kDa, approximating monomeric (α) and dimeric (α₂) molecular masses for the protease. Indeed, under these conditions, SGPD exists mainly in dimeric form (Fig. 4). A sample prepared in nondenatured condition was also included in the analysis to establish the position of the native form of the enzyme after electrophoresis. The predominantly negatively charged SGPD (Fig. 5) had a high electrophoretic mobility in its undenatured form but migrated to a distinct position relative to the monomeric or dimeric forms of the denatured enzyme. Amino-terminal analysis of the 34-kDa

sequences is relative to the first nucleotide in the known sequence and to the first amino acid coded by the gene. The numbering that appears *above* the sequence is relative to the first amino acid in the mature protease. A putative ribosome binding site is indicated by a series of *dots* preceding the initiation codon. Inverted repeated regions which follow the termination codon are *underlined*. Junctions between the pre- and proregions and pro and mature regions are indicated by a *closed* and an *open triangle*, respectively.
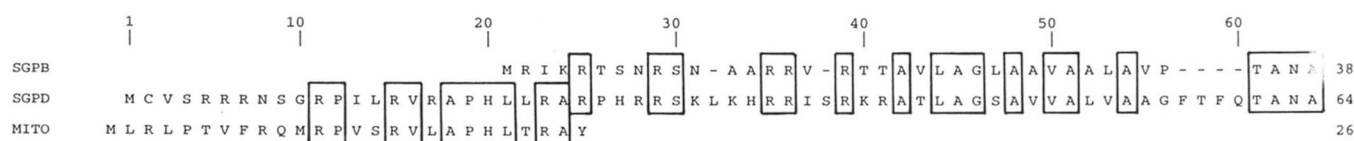
```
                     1            10              20            30              40            50              60
                     |            |               |             |               |             |               |
                     |            |               |             |               |             |               |
SGPB                                        M R I K R T S N R S N - A A R R V - R T T A V L A G L A A V A A L A V P - - - T A N A   38
SGPD       M C V S R R R N S G R P I L R V R A P H L L R A R P H R R S K L K H R R I S R K R A T L A G S A V V A L V A A G F T F Q T A N A   64
MITO       M L R L P T V F R Q M R P V S R V L A P H L T R A Y                                                                          26
```

FIG. 2. **Homology of the SGPD preregion with bacterial and mitochondrial signal sequences.** The preregion of SGPD is shown in alignment with the presequence of the protease SGPB and the mitochondrial import sequence of human hsp60 (22, 23). The unusually long preregion of SGPD can be divided into two domains on the basis of homologies with the prokaryotic secretion and mitochondrial import signals. The predictive method of Gavel and von Heijne (24) revealed that residues 1–44 comprise a sequence that is consistent with mitochondrial import signals. Residues 24–40 have the potential of forming an amphipathic α-helix with one face highly positively charged, a motif considered essential for translocation of proteins into mitochondria (25, 26). The carboxyl-terminal portion of the preregion (amino acids 45–64) is homologous with the preregion of SGPB.
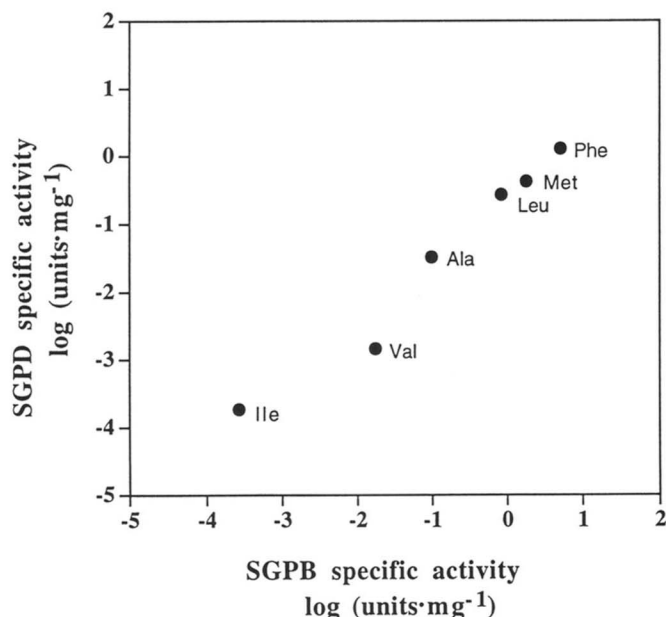

FIG. 3. **Substrate specificity of SGPD.** The specific activity of SGPD is shown relative to the protease SGPB which is known to be chymotrypsin-like in activity. Specific activities were examined for a series of substrates having the general sequence N-succinyl-Ala-Ala-Pro-X-p-nitroanilide; X, the amino acid at the P1 site of the substrate, was varied. The P1 amino acid is indicated to the *right* of each data point. The data points fall on line with a slope = 1, indicating that the two enzymes have approximately the same substrate specificities.

FIG. 4. **SDS-PAGE analysis of SGPD.** Polyacrylamide gel electrophoresis of the proteases SGPC and SGPD. *Lane 1*, denatured molecular weight standards; *lane 2*, SGPC prepared under "standard denaturing conditions" (see "Materials and Methods"); *lane 3*, SGPD prepared under standard denaturing conditions; *lane 4*, SGPD prepared under nondenaturing conditions. SGPC has an apparent molecular mass of 26 kDa (5), whereas, under standard denaturing conditions, SGPD resolves into two bands with apparent molecular masses of 17 and 34 kDa, approximating monomeric (α) and dimeric (α₂) masses for the protease. Electrophoresis of SGPD prepared in nondenaturing conditions produced a single band clearly distinguishable from SGPD prepared under standard denaturing conditions.

band confirmed the position of the pro mature junction, ruling out the possibility that this band represents an unprocessed pro mature form of the enzyme.

Alignment and Phylogenetic Analyses—As a first step in the generation of a phylogenetic tree for the *S. griseus* and *L. enzymogenes* proteases, mature segments were aligned as shown in Fig. 6. This alignment of amino acid sequences was then used to generate the corresponding alignment of the nucleotide sequences. Phylogenetic relationships were examined using a parsimony method which selects a tree that requires the minimum number of mutational changes in order to explain the data set. The DNApars program used in the analysis performed unrooted parsimony (analogous to Wagner trees) on the nucleotide sequences and calculated the number of changes for each base required for a given tree (28). The result was a single most parsimonious tree requiring 796 steps. Bootstrap analysis with DNAboot (15) was used to place confidence limits on the phylogeny presented in Fig. 7.

DISCUSSION

In a probe of *S. griseus* genomic DNA we detected five genes with significant homology to *S. griseus* protease B (5). We now know that these genes correspond to three well characterized *S. griseus* proteases (namely SGPA, SGPB itself (6) and SGPE (7)) and two novel proteases (SGPC (5) and SGPD). Hybridiza-
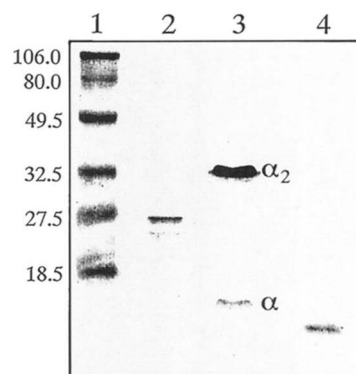
tion studies and sequence analyses indicated a very close relationship between the mature regions of the five enzymes. Nevertheless, the two newly discovered enzymes were found to be remarkably distinct in aspects of their structure. For example, SGPC has a carboxyl-terminal addition with a high degree of homology to chitin-binding domains and, as discussed below, the recombinant SGPD forms an extraordinarily stable dimer.

In prokaryotes, the preregion acts to signal translational secretion of extracellular enzymes. The preregion sequence of SGPD (Fig. 1) is significantly longer than that of other bacterial proteases (62 amino acids) and it can be divided into two parts, an amino-terminal segment with the characteristics of a mitochondrial import signal and a carboxyl-terminal segment characteristic of bacterial secretion signals. These characteristics are most evident when the preregion is aligned with the preregions of SGPB (6) and the mitochondrial heat shock protein hsp60 (22, 23) (Fig. 2). We are aware of no other prokaryotic enzyme with this type of signal sequence. The unusual organization of the preregion in SGPD suggests that the protease has a function distinct from that of other *S. griseus* proteases.

A recombinant *sprD* gene was constructed and expressed in *B. subtilis* in a system that we have used successfully to express the proteases SGPB, SGPC (5), SGPE (7), SGPA, and the α-lytic protease.[2] The purified enzyme showed a primary specificity toward large aliphatic and aromatic amino acids that was virtually identical to that of SGPB (Fig. 3). However, SGPD isolated from *B. subtilis* culture supernatants proved to be much larger than anticipated from the nucleotide sequence of *sprD*. SDS-PAGE gels gave two bands corresponding to pro-
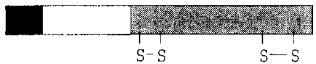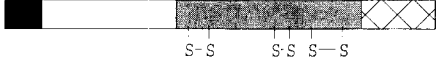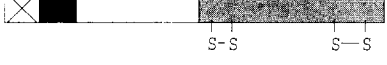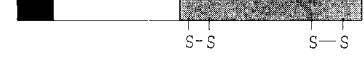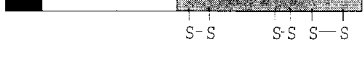
Enzyme	Organization of Zymogens	Quaternary Structure	pI	Amino Acid Specificity	Ref.
SGPA		α	7.10	large aliphatic/ aromatic	6,16
SGPB		α	7.01	large aliphatic/ aromatic	6,17
SGPC		α	7.59	large aliphatic/ aromatic	5
SGPD		α_2	3.55	large aliphatic/ aromatic	
SGPE		α	7.86	glutamic acid	7,18
α-Lytic		α	9.84	small aliphatic	19

FIG. 5. **Summary of the properties of bacterial chymotrypsin-like serine proteases.** The prepro mature organization of the six homologous proteases is illustrated. The homologous preregions are shown in *solid boxes*, proregions are in *open boxes*, and mature regions are *shaded*. The carboxyl- and amino-terminal extensions of the enzymes SGPC and SGPD are *cross-hatched*. To the *right* of each illustration is the quaternary structure of the mature protease, the isoelectric point (*pI*) deduced from sequence information using the computer program PC/Gene (IntelliGenetics, Inc., Mountain View, CA), and the primary amino acid specificity of the enzymes.

teins with molecular masses of approximately 34 and 17 kDa (Fig. 4). Gel filtration (size exclusion) chromatography subsequently established that the enzyme exists in the form of a stable homodimer. The molecular mass of a monomeric SGPD should be 18.7 kDa according to sequence data, and consequently, a homodimer should have a mass of roughly 36 kDa. The high mobility of the protein in SDS-PAGE is most likely due to the high negative charge on the protein, although it is possible that SGPD experiences some limited proteolysis during maturation causing a reduction in its mass. Amino-terminal analysis of the protein ruled out the possibility that the 36-kDa band corresponded to an unprocessed pro mature form of the protein.

It is remarkable that SGPD should have such a high degree of homology to its monomeric cousins and yet form such a stable dimer. Hence, the transition from monomer to dimer (or vice versa) involves relatively few residues in the protein. Given the high negative charge on SGPD (Fig. 5) one might expect the monomers to repel one another. Therefore, dimerization may involve metal chelation, the formation of intermolecular salt bridges, or both. We are currently examining the physical basis for the extraordinary stability of the SGPD dimer in denaturing conditions.

The unusual quaternary structure of SGPD combined with the unusual signal sequence suggests that the enzyme is targeted to a subcellular location. This notion is supported by the fact that SGPD has never been observed in *S. griseus* secretions even though the substrate specificity of the enzyme is similar to that of the well characterized proteases SGPA and SGPB (Fig. 3). Prokaryotes are not known to contain specific organelles, however subcellular compartments, or mesosomes, have been observed in *Streptomyces* (30) and other genera of bacteria (31–33). Although the significance of mesosomes is controversial, they may have functions similar to the periplasmic spaces of Gram-negative bacteria or even the organelles of eukaryotic cells (34). It is tempting to speculate that SGPD is directed to one of these structures.

The unusual signal sequence of SGPD has implications for the endosymbiont hypothesis which proposes that mitochon-

dria are derived from bacteria. According to this hypothesis the "mitochondrial" genes of a proto-eukaryotic cell were moved to the nucleus (35) where they had targeting sequences attached to them. Therefore, similarities between bacterial and mitochondrial targeting are to be expected. The fact that the preregion of SGPD contains features of both mitochondrial and prokaryotic signal sequences lends support to the endosymbiont hypothesis, but it also implies that so-called "mitochondrial" targeting sequences predate the existence of mitochondria.

Serine proteases are divided into two groups according to their structural (tertiary) similarity to either the enzyme chymotrypsin or subtilisin and the enzymes that are the subject of this study are all chymotrypsin-like in structure. Membership in the chymotrypsin branch of the family can be further divided according to the dimensions of the enzymes' proregions. With one exception, chymotrypsin-like enzymes derived from bacteria possess large propeptide regions. The related mammalian enzymes possess small propeptides, amounting in some cases to a few amino acids. Studies of the α-lytic protease have demonstrated the importance of bacterial proregions in catalyzing the proper folding and maturation of bacterial enzymes (36, 37). Contrasting the situation in bacteria, the function of the proregion in mammalian enzymes is to block the amino terminus of the protease, holding the enzyme in an inactive state until the propeptide is cleaved from the zymogen. Hence, the mammalian propeptides appear not to be involved in "catalyzing" the folding process.

There is considerable variation in the lengths of proregions even within the group of bacterial enzymes compared in Figs. 5 and 6. SGPA and SGPB appear to fall into one group according to the length of their proregions, SGPC and the α-lytic protease in another group, and SGPE and SGPD in a third. This arrangement is also reflected in the phylogeny produced using the parsimony (DNApars) and bootstrap (DNAboot) analyses of the mature regions of each protein (Fig. 7).

S. griseus trypsin (SGT) is the one exception to the distinction between bacterial and mammalian enzymes. It has a bacterial origin but in terms of sequence and structure it is more closely related to mammalian trypsin than other *Streptomyces*

A

```
sgpd pro                                          SDDVPAFGAKTLSADAAGKLATTLDRDLGADAAGSYYDATAKTLVVNVVDEAGAEQVR·········
sgpe pro                                          ADTPPAAPSAIPAPSAYALDAAVERQLGAATAGTYLDAKTGGLVVTVTTDRAEEQAR·········
sgpc pro  APPVSADSLSPGMLAALERDLGLDEDAARSRIANEYRAAAVAAGLEKSLGARYAGARVSGAKATL··TVATTDASEAARITEAGARAE
lyti pro        ADQVDPQLKFAMQRDLGIFPTQLPQYLQTEKLARTQAAAIEREFGAQFAGSWIERNEDGSFKLVAATSGARKSSTL·GGVEVR


sgpa pro  APEAESKATVSQLA·DASSAIL··AADVA·GTAWYTEASTGKIVLTADSTVSKAELAKVSNALAGSKAKLTVKRAEGKFTPL
sgpb pro      ETPRTFSANQLT·AASDAVL··GADIA·GTAWNIDPQSKRLVVTVDSTVSKAEINQIKKSAGANADALRIERTPGKFTKL
sgpd pro  QAGGKARIVENSLAELK·SARGTLTE·KATIP·GTSWAVDPVSNKVLVTADSTVDGAAWKKLSAVVEGLGGKAELNRTAGEFTPL
sgpe pro  AAGATVRRVARSAAQLD·AAMATLEA·EAKIT·GTSWGVDPRTNRVAVEADSSVSARDMARLEAVAERLGSAVDIKRVPGVFHRE
sgpc pro  VVGHSLDRFEGVKKSLDKAALDKAPKN······VPVWYVDVAANRVVV··NAASPAAGQA·FLKVAGVDRGLVTVARSAEQPRAL
lyti pro  NVRYSLKQLQSAMEQLDAGANARVKGVSKPLDGVQSWYVDPRSNAVVVKVDDGATDAGVD·FVALSGADSAQVRIESSPGKLQTT
```

B

```
sgpa mat  IAGGEAITTG·GSRCSLGFNVSVNGVAHALTAGHCTNISASWS········IGTRTGTSFPNNDYGIIRHSNPAAA·DGRVYLYNG
sgpb mat  ISGGDAIYSS·TGRCSLGFNVRSGSTYYFLTAGHCTDGATTWWANSARTTVLGTTSGSSFPNNDYGIVRYTNTTIPKDGTVG····
sgpd mat  IAGGDAIWGS·GSRCSLGFNVVKGGEPYFLTAGHCTESVTSWS·DTQGGSEIGANEGSSFPENDYGLVKYTSDTAH·PSEVNLYDG
sgpe mat  VLGGGAIYGG·GSRCSAAFNVTKGGARYFVTAGHCTNISANWS·ASSGGSVVGVREGTSFPTNDYGIVRYTDGSSP·AGTVDLYNG
sgpc mat  ADIRGGDAYYMNGSGRCSVGFSVTRGTQNGFATAGHCGRVGTTTN··GVNQQAQGTFQGSTFPGRDIAWVATNANWTP·RPLVNGYGR
lyti mat  ANIVGGIEYSINNASLCSVGFSVTRGATKGFVTAGHCGTVNATAR···IGGAVVGTFAARVFPGNDRAWVSLTSAQTL·LPRVA··NG


sgpa mat  S·YQDITTAGNAFVGQAVQRSGSTTGLRSGSVTGLNATVNYGSSGIVYGMIQTNVCAEPGDSGGSLFAGS·TALGLTSGGS·····GN
sgpb mat  ··GQDITSAANATVGMAVTRRGSTTGTHSGSVTALNATVNYGGGDVVYGMIRTNVCAEPGDSGGPLYSGT·RAIGLTSGGS·····GN
sgpd mat  S·TQAITQAGDATVGQAVTRSGSTTQVHDGEVTALDATVNYGNGDIVNGLIQTTVCAEPGDSGGALFAGD·TALGLTSGGS·····GD
sgpe mat  S·TQDISSAANAVVGQAIKKSGSTTKVTSGTVTAVNVTVNYGD·GPVYNMVRTTACSAGGDSGGAHFAGS·VALGIHSGSS·····G·
sgpc mat  G·DVTVAGSTASVVGASVCRSGSTTGWHCGTIQQLNTSVTYPE·GTISGVTRTSVCAEPGDSGGSYISGS·QAQGVTSGGS·····GN
lyti mat  SSFVTVRGSTEAAVGAAVCRSGRTTGYQCGTITAKNVTANYAE·GAVRGLTQGNACMGRGDSGGSWITSAGQAQGVMSGGNVQSNGNN


sgpa mat  CR···TGGTTFYQPVTEALSAYGATVL
sgpb mat  CS···SGGTTFFQPVTEALSAYGVSVY
sgpd mat  CS···SGGTTFFQPVPEALAAYGAEIG
sgpe mat  CS··GTAGSAIHQPVTEALSAYGVTVY
sgpc mat  CS···SGGTTYFQPINPLLQAYGLTLVTS
lyti mat  CGIPASQRSSLFERLQPILSQYGLSLVTG
```

FIG. 6. **Alignment of protease amino acid sequences.** The best alignment of *A*, the pro, and *B*, the mature regions of the proteases SGPA, B, C, D, E, and α-lytic protease are shown. Regions of significant homology are indicated in *bold* and correspond to identities in at least 2 of 2, 3 of 4, 4 of 5, or 4 of 6 of the aligned sequences.
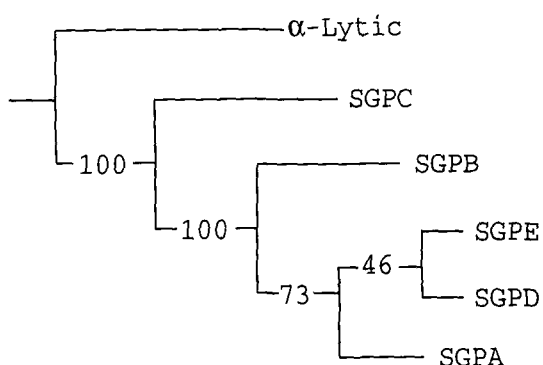

FIG. 7. **Phylogenetic tree of the bacterial proteases SGPA, B, C, D, and E and the α-lytic protease.** The tree was constructed using the nucleotide sequences of the mature regions of the respective proteases. Nucleotide sequence alignments correspond to the amino acid alignment shown in Fig. 6*B*. Pre- and proregion sequences and the sequence of the carboxyl-terminal domain of SGPC were not included in the analysis. The number at each of the forks represents the number of times that the particular grouping (consisting of the species to the right of the fork) was generated during the 100 bootstrap replicates.

enzymes. The propeptide of SGT is 4 residues in length (38), similar to the proregions of mammalian trypsins (4 residues) (39) and notably, SGT was not detected by hybridization with SGPB. The anomolous relationship of SGT to bacterial and mammalian enzymes has even led some authors to speculate that SGT was acquired from a mammalian source only recently

(40, 41). Perhaps a more satisfactory explanation is that the proregions of the bacterial proteases are becoming shorter through the course of evolution and SGT is simply furthest of the *S. griseus* enzymes from the ancestral.

The phylogenetic tree shown in Fig. 7 indicates that the enzymes SGPC and α-lytic protease have diverged the most from the other proteases in the analysis. The six proteases form a monophyletic group beginning with α-lytic and followed by SGPC, SGPB, SGPA, and finally SGPD and SGPE. DNA bootstrap analysis strongly supports the relationships, placing excellent confidence limits on the branches between α-lytic, SGPC, SGPB, and the tricotomy formed between SGPB and the remaining three proteases (SGPA, D, and E).

We believe that α-lytic protease and SGPC are the two most ancient proteases in our study, primarily because the two enzymes have the most extensive proregions. Notably, these are also the only two enzymes with three disulfide bonds instead of two (Fig. 5). It can be argued that the presence of the two homologous proteases in different genera of bacteria is evidence that they arose from a common ancestor before the organisms diverged.

REFERENCES

1. Fersht, A. R. (1985) *Enzyme Structure and Mechanism*, 2nd Ed., W. H. Freeman & Co., New York
2. Neurath, H. (1986) *J. Cell. Biochem.* **32,** 35–49
3. Prendergast, J. A., Pinkoski, M., Wolfenden, A., and Bleackley, R. C. (1991) *J. Mol. Biol.* **220,** 867–875
4. Bond, J. S., and Butler, P. E. (1987) *Annu. Rev. Biochem.* **56,** 333–364
5. Sidhu, S. S., Kalmar, G. B., Willis, L., and Borgford, T. J. (1994) *J. Biol. Chem.*

269, 20167–20171

6. Henderson, G., Krygsman, P., Liu, C. J., Davey, C. C., and Malek, L. T. (1987) *J. Bacteriol.* **169,** 3778–3784
7. Sidhu, S., Kalmar, G., and Borgford, T. (1993) *Biochem. Cell Biol.* **71,** 454–461
8. Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene (Amst.)* **33,** 103–119
9. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual,* 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
10. Kawamura, F., and Doi, R. H. (1984) *J. Bacteriol.* **160,** 442–444
11. Wells, J. A., Ferrari, E., Henner, D. J., Estell, D. A., and Chen, E. Y. (1983) *Nucleic Acids Res.* **11,** 7911–7925
12. Nakajima, K., Powers, J. C., Ashe, B. M., and Zimmerman, M. (1979) *J. Biol. Chem.* **254,** 4027–4032
13. Laemmli, U. K. (1970) *Nature* **227,** 680–685
14. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) *J. Biol. Chem.* **193,** 265–275
15. Felsenstein, J. (1989) *Cladistics* **5,** 164–166
16. James, M. N. G., Sielecki, A. R., Brayer, G. D., Delbaere, L. T. J., and Bauer, C. A. (1980) *J. Mol. Biol.* **144,** 43–88
17. Read, R. J., Fujinaga, M., Sielecki, A. R., and James, M. N. G. (1983) *Biochemistry* **22,** 4420–4433
18. Nienaber, V. L., Breddam, K., and Birktoft, J. J. (1993) *Biochemistry* **32,** 11469–11475
19. Fujinaga, M., Delbaera, L. T., Brayer, G. D., and James, M. N. G. (1985) *J. Mol. Biol.* **184,** 479–502
20. Silen, J. L., McGrath, C. N., Smith, K. R., and Agard, D. A. (1988) *Gene (Amst.)* **69,** 237–244
21. Randall, L. L., and Hardy, S. J. S. (1984) *Microbiol. Rev.* **48,** 290–298
22. Jindal, S., Dudani, A. K., Singh, B., Harley, C. B., and Gupta, R. S. (1989) *Mol. Cell. Biol.* **9,** 2279–2283
23. Venner, T. J., and Gupta, R. S. (1990) *Biochim. Biophys. Acta* **1087,** 336–338
24. Gavel, Y., and von Heijne, G. (1990) *Protein Eng.* **4,** 33–37
25. Hartl, F. U., and Neupert, W. (1990) *Science* **247,** 930–938
26. Lemire, B. D., Fankhauser, C., Baker, A., and Schatz, G. (1989) *J. Biol. Chem.* **264,** 20206–20215
27. Hutter, R., and Eckhardt, T. (1988) in *Actinomycetes in Bio/Technology* (Goodfellow, M., Williams, S. T., Mordarski, M., ed) pp. 89–184, Academic Press, London
28. Fitch, W. M. (1971) *Syst. Zool.* **20,** 406–416
29. Deleted in proof
30. Kurylowicz, W., Kurzatkowski, W., Williams, S. T., Woznicka, W., and Paszkiewicz, A. (1975) *Atlas of Ultrastructure of Streptomyces in Course of Biosynthesis of Antibiotics,* PZWL-Polish Medical Publishers, Warszawa, Poland
31. Cherepova, N. V., Baykousheva, S. P., and Ilieva, K. Z. (1986) *J. Gen. Microbiol.* **132,** 669–675
32. Horne, D., and Tomasz, A. (1985) *J. Bacteriol.* **161,** 18–24
33. Nakasone, N., Masuda, K., and Kawata, T. (1987) *Microbiol. Immunol.* **31,** 403–415
34. Greenwalt, J. W., and Whiteside, T. L. (1975) *Bacteriol. Rev.* **39,** 405–463
35. Ostermann, J. (1990) in *Bio/Technology & Genetic Engineering Reviews* (Tombs, M. P., ed) pp. 219–249, Intercept, Andover
36. Silen, J. L., and Agard, D. A. (1989) *Nature* **341,** 462–464
37. Silen, J. L., Frank, D., Fujishige, A., Bone, R., and Agard, D. A. (1989) *J. Bacteriol.* **171,** 1320–1325
38. Kim, J. C., Cha, S. H., Jeong, S. T., Oh, S. K., and Byun, S. M. (1991) *Biochem. Biophys. Res. Commun.* **181,** 707–713
39. Huber, R., and Bode, W. (1978) *Acc. Chem. Res.* **11,** 114–122
40. Hartley, B. S. (1979) *Proc. R. Soc. Lond. Ser. B* **205,** 443–452
41. Hartley (1970) *Philos. Trans. R. Soc. Ser. B* **257,** 77–87