The 65-Kilodalton Antigen of *Mycobacterium tuberculosis*[†]

THOMAS M. SHINNICK[‡]

Department of Molecular Biology, Research Institute of Scripps Clinic, La Jolla, California 92037

Received 16 June 1986/Accepted 8 December 1986

The immune response of the host to the antigens of *Mycobacterium tuberculosis* plays the key role in determining immunity from infection with as well as the pathogenicity of this organism. A 65-kilodalton (kDa) protein has been identified as one of the medically important antigens of *M. tuberculosis*. The gene encoding this antigen was isolated from a λ gt11-*M. tuberculosis* recombinant DNA library using monoclonal antibodies directed against the 65-kDa antigen as the specific probes. The nucleotide sequence of this gene was determined, and a 540-amino-acid sequence was deduced. This sequence was shown to correspond to that of the 65-kDa antigen by constructing a plasmid in which this open reading frame was fused to the *lacZ* gene. The resulting fusion protein reacted specifically with the anti-65-kDa protein antibodies. A second long open reading frame was found downstream of the 65-kDa antigen gene which could encode a protein of 517 amino acids. This putative protein contained 29 tandemly arranged partial or complete matches to a pentapeptide sequence.

The mycobacteria are a diverse collection of acid-fast, gram-positive bacteria, some of which cause important human and animal diseases (reviewed in references 4 and 6). In humans, the two most common mycobacteria-caused diseases are tuberculosis and leprosy, which result from infection with *Mycobacterium tuberculosis* and *Mycobacterium leprae*, respectively. These two diseases afflict more than 35 million individuals worldwide and result in over 4 million deaths annually (4). The pathogenicity of these mycobacterial infections is closely tied to the immune response of the host to the invading mycobacterium (6, 10, 12, 18). Not only does *M. tuberculosis* infect and grow within cells of the host immune system, primarily the alveolar macrophage, it is also the host cellular immune response that plays the key role in immunity from infection, containment of the infection at the initial focus of infection, progression or regression of the infection, and tissue damage or destruction at the foci of infection (6, 10, 12, 18). In addition, the standard method of detecting an *M. tuberculosis* infection, the tuberculin skin test, actually measures the host cellular immune response to the mycobacterium (37). The mycobacterial components that are important in eliciting the cellular immune response are not yet well defined.

A number of studies have attempted to define the mycobacterial antigens by standard biochemical and immunological techniques, including analysis of the target antigens of monoclonal hybridoma antibodies directed against mycobacteria (6, 11, 14, 15, 18, 23, 38). One particular antigen, a 65-kilodalton (kDa) protein, is present in a wide range of mycobacterial species and has been most intensively studied as an antigen of *M. leprae* (13-17, 45). This antigen has been designated the 65K antigen or the cell wall protein a (CWP-a) antigen, since it appears to copurify with cell walls in some isolation procedures (12). In Western blot assays, monoclonal antibodies directed against this antigen react with two major components in an *M. leprae* extract that migrate with apparent sizes of 55 and 65 kDa and react occasionally with smaller components as well (14, 17). It is not known if these species represent discrete proteins or precursors and prod-

ucts or result from chemical or enzymatic cleavage during isolation. In other species, such as *M. gordonae*, only a single species of about 65 kDa is detected with monoclonal antibodies (17).

The 65K antigen is one of the major immunoreactive proteins of the mycobacteria. This antigen contains epitopes that are unique to a given mycobacterial species as well as epitopes that are common to various species of mycobacteria (14-17). Purified 65K antigen can elicit a strong delayed-type hypersensitivity reaction in experimental animals infected with *M. tuberculosis* (unpublished results). Antibodies directed against this protein can be detected in the sera of patients with tuberculosis or leprosy, and T-cells reactive with this antigen can be isolated from patients with leprosy or tuberculosis as well as from BCG-vaccinated persons (13-15, 30, 39; T. Godal and S. H. E. Kaufmann, personal communications). Overall, the 65K antigen appears to be a major, medically important B- and T-cell immunogen and antigen in humans.

In this paper, I report the isolation of the gene encoding the *M. tuberculosis* 65K antigen and determination of its nucleotide sequence. The sequence contains an open reading frame (ORF) encoding 540 amino acids, or about 60 kDa, which corresponds to the 65K antigen. A second long ORF which could encode a protein of 517 amino acids was also found on the mycobacterial DNA fragment containing the 65K antigen gene. Interestingly, the central region of the deduced amino acid sequence of the 517-amino-acid protein contains up to 29 tandemly arranged, perfect and imperfect repeats of a 5-amino-acid sequence. This feature is reminiscent of the features of the sequence of the major T-cell antigen of the sporozoite stage of the human malarial parasite (reviewed in reference 32).

MATERIALS AND METHODS

Bacteria, phage, and plasmids. The *Escherichia coli* strains used in this work were BNN97 (44), JM83 (42), JM101 (42), Y1089 (44), and Y1090 (44). Plasmids pUC19 (42) and pMC1871 (35) were purchased from Pharmacia. The recombinant DNA library of *M. tuberculosis* genomic DNA fragments in the λ gt11 vector was constructed by Young et al. (43) and made available through the World Health Organization's (WHO) Program for Research in the Immunology of Tuberculosis. Recombinant phages λ RY3143 and λ RY3146

[†] Publication 4451-MB from the Research Institute of Scripps Clinic.

[‡] Present address: Hansen's Disease Laboratory, Division of Bacterial Diseases, Centers for Disease Control, Atlanta, GA 30333.

were generously provided by R. A. Young (Whitehead Institute) (43). Subclones of the mycobacterial DNA inserts in these recombinant phage were constructed in pUC19 or M13mp9 (27) vectors by standard recombinant DNA techniques (25).

Antisera. Monoclonal antibodies specific for the 65K antigen were obtained from the Immunology of Tuberculosis Scientific Working Group under a grant from the WHO/World Bank/UNDP Special Programme for Vaccine Development. These antibodies included IT-13 (WTB-78 [9]), IT-31 (SA2D5H4; T. Buchanan, unpublished), and IT-33 (MLIIH9 [16]). Anti- β -galactosidase antibodies were purchased from Cooper Biomedical. Polyclonal rabbit antisera directed against a sonicate of *M. tuberculosis* H37Rv were elicited as described previously (29).

Immunoscreening of λ gt11-*M. tuberculosis* library. Clones reactive with the monoclonal antibodies specific for the 65K antigen were isolated essentially as described by Young et al. (43). Briefly, for each 150-mm LB plate, 0.6 ml of a fresh overnight culture of Y1090 was infected with 1×10^5 to 2×10^5 PFU of the library. After 3.5 to 4 h of growth at 42°C, the plaques were overlaid with a dry nitrocellulose filter which had been saturated with 10 mM isopropyl- β -D-thiogalactopyranoside (IPTG). The plates were incubated an additional 3.5 to 4 h at 37°C and then removed to room temperature, and the position of the filters was marked. The filters were washed briefly in TBST (50 mM Tris hydrochloride, pH 8, 150 mM NaCl, 0.05% Tween 20) and then incubated in TBST with 20% fetal calf serum. After 30 min at room temperature, the filters were transferred to TBST plus antibody. For the initial screen, the antibody mix contained a 1:1,000 dilution of IT-13, IT-31, and IT-33. The filters were incubated with the antibody solution overnight at 4°C with gentle agitation, washed in TBST, and reacted with biotinylated goat anti-mouse immunoglobulin, the Vectastain ABC reagent, and developer as described by the manufacturer (Vector Laboratories). After the color had developed, the filters were washed with several changes of water and air dried. Phage corresponding to positive signals were plaque purified twice. To determine which monoclonal antibodies reacted with which of the recombinant phage, about 100 PFU of a purified phage stock were inoculated in a small spot on a lawn of Y1090 bacteria on an LB plate. The phage were allowed to grow and induced to synthesize the foreign proteins as described above. The filters were then reacted with a 1:1,000 dilution of one of the monoclonal antibodies. The subsequent steps were the same as for the initial screen.

Western blot assays. Cells containing phage or plasmids in which expression of the foreign sequences was under the control of the *E. coli lac* gene regulatory sequences were induced to synthesize the foreign proteins by incubating the cells in the presence of 2.5 mM IPTG for 2 h. Crude lysates of cells expressing λ gt11 recombinants were made as described by Huynh et al. (21). Crude lysates of cells expressing plasmid-encoded proteins were made by harvesting cells from overnight cultures and suspending the cells in 10 mM Tris (pH 7.5)–10 mM EDTA containing 100 μ g of lysozyme per ml. After 10 min at room temperature, sodium dodecyl sulfate (SDS) was added to a final concentration of 0.5%. A protease inhibitor (Trasylol; Boehringer Mannheim) was added to all crude lysates at a final concentration of 0.3%. The crude protein preparations were electrophoresed on 10% polyacrylamide–SDS Laemmli gels (24), and the separated proteins were electrophoretically transferred to nitrocellulose (40). The immobilized proteins were reacted with a 1:1,000 dilution of monoclonal antibody IT-13 in TBST

overnight at 4°C. The nitrocellulose filters were then washed, reacted with peroxidase-conjugated goat anti-mouse immunoglobulin, and developed as previously described (31).

Nucleic acid sequencing. The sequences of 5'-end-labeled restriction fragments of the mycobacterial DNA were determined by a modification of the partial chemical degradation technique of Maxam and Gilbert (5, 26). For the M13 dideoxy sequencing studies, *Sau*3AI fragments from the mycobacterial DNA inserts were subcloned into the *Bam*HI site of M13mp9. Phage DNA was isolated from the M13 recombinants and subjected to the dideoxy chain termination sequencing reactions (3, 34). The products of the sequencing reactions were electrophoresed on 6% acrylamide–7 M urea–0.5 to 2.5 $\times$ TBE gradient sequencing gels (3). The gels were dried under vacuum and exposed to Kodak XRP-1 film. The nucleotide sequences were determined independently for both strands of the mycobacterial DNA.

Computer-aided analyses of the nucleic acid sequences and deduced protein sequences were performed with the databases and programs provided by the Nucleic Acid and Protein Identification Resources of the National Institutes of Health as well as the programs of Chou and Fasman (7) and Hopp and Woods (20).

β -Galactosidase assays. Cells were grown in B broth or B broth plus 2.5 mM IPTG to an OD₆₀₀ of about 0.3. Crude lysates were made, and β -galactosidase activity was assayed as described by Miller (28).

RESULTS AND DISCUSSION

Isolation and analysis of recombinants expressing the 65K antigen. To isolate the gene that encodes the 65K antigen, monoclonal antibodies directed against this antigen were used to screen a protein expression library constructed with mycobacterial DNA. An expression library was chosen since it was not known a priori whether the *M. tuberculosis* genes would be expressed in *E. coli*. Such a recombinant DNA library has been constructed by Young et al. (43) and contains genomic DNA fragments of *M. tuberculosis* inserted into the expression site of the λ gt11 vector. In this system the inserted coding sequences can be expressed as a fusion protein with β -galactosidase. The 65K antigen-specific monoclonal antibodies used in these studies were generated in the laboratories of T. M. Buchanan (Pacific Medical Center, University of Washington, Seattle) and J. Ivanyi (MRC Tuberculosis Unit, Hammersmith Hospital, London) and obtained from the Steering Committee on the Immunology of Tuberculosis of WHO.

As the initial antibody probe, a pool containing three monoclonal antibodies directed against the 65K antigen was used (IT-13, IT-31, and IT-33). Thirty-eight positive signals were detected in a screen of about 8×10^5 recombinant phage. The phage corresponding to the positive signals were plaque purified twice and then assayed for reactivity with the individual antibodies (Table 1). Twenty-eight of the recombinants produced antigens that reacted with all three antibodies, while 10 recombinants produced antigens that reacted with one or two of the antibodies. Overall, the 38 recombinants displayed five different patterns of reactivity with the three antibodies. The results indicate that although the three antibodies react with the same mycobacterial antigen, each recognizes a different epitope on that antigen. Richard A. Young (Whitehead Institute) has also screened this λ gt11-*M. tuberculosis* library with one of these antibod-

FIG. 1. Restriction map of the DNA inserts in the recombinants expressing the 65K antigen. DNA was isolated from phage stocks as described by Helms et al. (19), and a restriction enzyme cleavage site map was constructed. The thick line represents a composite deduced restriction map of this region of the mycobacterial genome. The thin lines represent the individual recombinants, which are identified to the right. The short vertical lines at the ends of the inserts represent *EcoRI* cleavage sites which are expected to flank the inserted DNA. The dashed horizontal lines indicate that that end of the insert did not contain an *EcoRI* cleavage site and thus was not precisely mapped. Abbreviations: A, *SacI*; B, *BglII*; K, *KpnI*; M, *BamHI*; P, *PstI*; R, *EcoRI*; S, *Sall*; V, *PvuII*; X, *XhoI*.

ies (IT-13) and detected 10 additional recombinants (43; personal communication). These recombinants were not assayed for reactivity with the other antibodies.

DNA was isolated from 20 of the recombinants expressing the 65 K antigen, and a restriction enzyme cleavage site map was deduced for this region of the mycobacterial genome (Fig. 1). In most of the recombinants, the mycobacterial DNA insert was flanked by *EcoRI* sites, as expected from the way in which the library was constructed. However, in 6 of the 20 recombinants studied, only one of the expected *EcoRI* sites was present. This observation raises the possibility that a significant fraction of the recombinant phage in this library might have arisen from the insertion of a fragment containing only one functional *EcoRI* site into the λ gt11 *EcoRI* site or that some clones might have undergone some sort of recombination, rearrangement, or deletion

event during propagation which removed one of the *EcoRI* sites. The deduced restriction map is in good agreement with the published map of the gene for the *M. bovis* 65K antigen (39) except for the presence of two additional *SmaI* sites in the *M. tuberculosis* gene. The map does not match that of the *M. leprae* 65K antigen gene (45). This is not unexpected because, based on DNA homology studies, *M. tuberculosis* is at least 90% homologous with *M. bovis* and only about 20 to 30% homologous with *M. leprae* (1, 22).

To determine the nucleotide sequence of this region of the mycobacterial genome, several fragments from the λ gt11 recombinants were subcloned into the plasmid vector pUC19. The majority of the sequence of this region was determined from a subclone (pTB7) of the 1.4-kilobase-pair (kbp) *EcoRI* fragment of λ SK7 and a subclone (pTB9) of the 2.6-kbp *EcoRI* fragment of λ RY3143. The sequence across the *EcoRI* site at the junction of these two fragments was determined from a fragment isolated from a subclone (pTB11) of the 2.8-kbp *KpnI* fragment of λ SK119. The sequence of the region 5' to the 2.6-kbp *EcoRI* fragment was determined from a subclone (pTB12) of the 2.4-kbp *KpnI* fragment of λ SK119. In all, the nucleotide sequence of 4,380 base pairs of the mycobacterial DNA was determined by a combination of the Sanger dideoxy chain termination (34) and Maxam-Gilbert chemical degradation (26) sequencing techniques. The sequence is shown in Fig. 2. As expected for *M. tuberculosis* genomic DNA (41), the base composition

TABLE 1. Patterns of antibody reactivity

No. of clones	Reactivity with antibody:		
	IT-13	IT-31	IT-33
28	+	+	+
3	+	+	—
3	—	+	+
2	—	+	—
2	—	—	+

of this fragment was about 66% G+C. The high G+C content increased the chances of sequencing artifacts due to compression and made it imperative that the sequences be determined for both strands in all regions.

ORFs. The sequence contained five ORFs that began with an ATG triplet and encoded more than 120 amino acids. Two of these exceeded 200 amino acids in length. One encoded 517 amino acids and the other 540 amino acids. There were an additional three ORFs encoding 140 to 190 amino acids that did not contain an initiation ATG triplet but did contain a GTG triplet. It is not known whether a GTG triplet can function as a translation initiation triplet in mycobacteria. The locations of these eight ORFs are shown schematically in Fig. 3. No portions of the deduced amino acid sequences of any of these ORFs displayed any significant homologies with sequences in the protein sequence database of the Protein Identification Resource as determined by a computer-aided search of the database with the homology search program of the Protein Identification Resource.

It should be noted that although an ORF encoding more than 100 amino acids would be considered to have a high probability of being expressed in most bacteria, this may not be true for the mycobacteria. That is, given that the G+C content of the insert is about 66%, a translation termination triplet (TAA, TAG, or TGA) would be expected to occur on average about once every 41 codons, compared with about once every 21 codons in a genome with a G+C content of 50%. Perhaps then an ORF for as many as 150 to 200 amino acids might be due to the random distribution of termination triplets rather than signifying possible biologic importance. As such, I describe in detail here only the two very long ORFs which could encode proteins of 517 and 540 amino acids, respectively.

The 540-amino-acid ORF corresponds to the 65K antigen. One of the long ORFs began with an ATG triplet at positions 252 to 254 of the sequence and extended to a TGA triplet at positions 1872 to 1874 and encoded 540 amino acids. To determine whether this ORF corresponded to the gene for the 65K antigen, the 1,511-base-pair (bp) *Bam*HI-*Kpn*I fragment from pTB12 (residues 438 to 1948 of the sequence presented in Fig. 2), which contains the majority of this ORF, was inserted into *Bam*HI-*Kpn*I-cleaved pUC19. In this construct (pTB22), the ORF is expressed from the *lacZ* transcription and translation initiation signals present in the pUC19 vector, resulting in the production of a fusion protein containing 15 amino acids at the amino terminus encoded by the *lacZ* gene of pUC19 followed by 478 amino acids of the mycobacterial ORF. Crude extracts were prepared from cells containing this plasmid and tested for reactivity with 65K antigen-specific antibodies in a Western blot analysis. The reactivities with monoclonal antibody IT-13 are shown in Fig. 4A. In all, five different monoclonal antibodies specific for the 65K antigen reacted with a species in the crude extract that migrated with an apparent molecular weight of about 55,000 (lane 3). No reactivity was seen in extracts of *E. coli* lacking the plasmid (lane 1). Furthermore, expression of this fusion protein was inducible with IPTG (compare lanes 2 and 3). Therefore, I conclude that this long ORF encompassing codons for residues 251 to 1870 encodes the *M. tuberculosis* 65K antigen.

65K antigen gene expression in *E. coli*. Because previous studies had shown that most mycobacterial genes were not expressed in *E. coli* by the mycobacterial transcription and translation signal sequences (8, 39), a protein expression library was used in the cloning experiments. In the λ gt11-*M. tuberculosis* library, the inserted mycobacterial coding se-

quences should be expressed as fusion proteins with β -galactosidase (43). Hence, it was somewhat surprising to find that the ORF encoding the 65K antigen did not extend to the 5' end of the mycobacterial DNA insert in λ SK119. This suggested that the 65K antigen was being expressed from the mycobacterial transcription and translation signal sequences. With respect to the previously described *E. coli* consensus signal sequences, the mycobacterial sequences 180 to 230 bp upstream of the presumed initiator ATG codon did display reasonable matches with the consensus sequences for the -35 (3 of 3 match with the highly conserved TTG) and -10 (4 of 6 match with TATAAT) regions of *E. coli* promoters (33). There was also a 5 of 5 match with the Shine-Dalgarno sequence (36) for a procaryotic ribosome-binding site (GGAGG) 13 bp upstream of the presumed initiator triplet for the 65K antigen ORF. Although the precise location of the mycobacterial regulatory sequences has not been determined experimentally, the results of the two experiments described below suggest that the mycobacterial sequences are indeed functional in *E. coli*.

The size of the anti-65K-reactive material produced by the recombinants was determined in a Western blot assay. To do this, crude lysates of cells expressing recombinant plasmids or phage which had been shown to contain the entire 65K antigen gene (λ SK116 and pTB12) as well as those which had been shown to contain the 65K antigen ORF fused to β -galactosidase (λ RY3146) were prepared as described in Materials and Methods. The lysates were electrophoresed on 10% SDS-polyacrylamide gels, and the separated proteins were electrophoretically transferred to nitrocellulose. The immobilized proteins were then reacted with monoclonal antibodies specific for the 65K antigen. The results with antibody IT-13 are shown in Fig. 4B. In cells expressing recombinants carrying the fused ORF, the monoclonal antibodies detected a single strongly reactive species migrating with an apparent molecular weight of about 160,000, as well as occasionally detecting smaller species (Fig. 4B, lane 3). In the extracts of the cells expressing recombinants that contained the entire 65K antigen gene, the monoclonal antibodies detected a single strongly reactive species that migrated with an apparent molecular weight of about 64,000 (Fig. 4B, lanes 1 and 2). Smaller reacting species (~40 to 55 kDa) were observed when large amounts of the extracts were loaded (lane 5) or when the protease inhibitor was omitted from the lysis buffer. Occasionally, a minor reacting species was also observed migrating with an apparent molecular weight of about 67,000 (not shown). Given the sizes of the anti-65K-reactive material, these data indicate that the 65K antigen can be expressed by using the mycobacterial translation initiation signals present in the 65K antigen gene. Also, since the vector contribution to the recombinant plasmids does not contain any known sequences that are properly located and oriented to promote the transcription of the inserted DNA, these data suggest that the mycobacterial transcription initiation signals function in *E. coli* to allow expression of the 65K antigen.

To obtain an approximate measure of the efficiency of utilization of the mycobacterial transcription and translation initiation signals in *E. coli*, two plasmids were constructed that placed the expression of enzymatically active β -galactosidase under the control of either the mycobacterial signal sequences or the *lac* gene signal sequences present in the plasmid pUC19. First, the 3,000-bp *Bam*HI fragment from pMC1871, which contains the coding sequences for amino acids 8 to 1021 of β -galactosidase (35), was inserted into the *Bam*HI site of pTB12 (residues 437 to 442 of the

FIG. 2. Nucleotide sequence of the region containing the *M. tuberculosis* 65K antigen gene. The deduced amino acid sequences of the proteins encoded by the two very long ORFs are presented in the one-letter code over (540-amino-acid protein) or under (517-amino-acid protein) the appropriate triplets. The first ATG triplet (the presumed initiator triplet) in each of these ORFs frames is marked (+). The sequences of the other four long ORFs are not shown. These ORFs are shown schematically in Fig. 3 and occupy residues 769 to 245 (F175), 1670 to 1197 (D158), 2368 to 1943 (E142), 3085 to 3645 (A187), 3395 to 2961 (E145), and 3464 to 3991 (B176).

The diagram shows a horizontal line representing a 4000-foot runway, with markers at 1000, 2000, 3000, and 4000 feet. Aircraft are represented by arrows indicating their direction of travel, with their names and distances from the runway ends in parentheses. The aircraft and their positions are as follows:

- C-540** (1000 ft from left end): Moving right.
- D-517** (3000 ft from left end): Moving left.
- F-175** (500 ft from left end): Moving left.
- E-145** (3500 ft from left end): Moving left.
- B-176** (3800 ft from left end): Moving right.
- D-158** (1500 ft from left end): Moving left.
- E-142** (2000 ft from left end): Moving left.
- A-187** (3200 ft from left end): Moving right.

FIG. 3. Schematic representation of the ORFs found in the mycobacterial DNA sequence that encode more than 140 amino acids. The putative initiation triplet is identified as m (AUG) or v (GUG) at the 5' end of each ORF.

sequence presented in Fig. 2). The resulting 8.1-kbp plasmid (pTB27) contained an ORF that encoded a fusion protein with 63 amino acids derived from the 65K antigen gene followed by 1,014 amino acids of β -galactosidase and whose expression was under the control of the transcription and translation signal sequences present in the mycobacterial DNA. As expected, this construct expressed a protein of about 120 kDa that reacted with anti- β -galactosidase antibodies in a Western blot assay (data not shown). Second, the 3,000-bp *Bam*HI fragment from pMC1871 was inserted into the *Bam*HI site in the polylinker of pTB9, which contains a 2.4-kb fragment of the 65K antigen gene inserted in the *Eco*RI site of pUC19. The resulting 8.1-kbp plasmid (pTB28) contained an ORF that encoded a fusion protein with 15 amino acids derived from the pUC19 *lacZ* gene and polylinker sequences followed by the 1,014 amino acids of β -galactosidase and whose expression was under the control of the *lac* gene signal sequences present in pUC19. Crude extracts of cells containing these plasmids were assayed for β -galactosidase activity as described previously. In cells containing pTB27, β -galactosidase activity ($\sim 2,800$ U/ μ g of protein) was about one-fourth of that ($\sim 11,000$ U/ μ g of protein) found in IPTG-induced cells containing pTB28. Given the unknowns inherent in this experiment (e.g., the specific activities and relative stabilities of the two fusion proteins), one cannot make a precise quantitative statement about the relative strengths of the mycobacterial signal sequences and the *E. coli lac* gene signal sequences based on the relative enzymatic activities found in the two cell extracts. However, the data indicate that these mycobacterial transcription and translation signal sequences are efficiently recognized in *E. coli*.

65K antigen sequence. Several interesting features of this long ORF have been revealed by a computer-aided analysis of the sequence. The overall base composition of this ORF is 65.5% G+C. However, the G+C content varies considerably within the codons so that the G+C content of the bases occupying the first two residues of the codons was 55%, while it was 87% for the bases found in the third position of the codons, thereby producing a bias towards codons that have a G or C in the third position. For example, 50 of the 51 leucine codons (CTX) had a G or C in the third position. Interestingly, the essentially random occurrence of any of the four bases in the first two positions of a codon plus the preference for G or C in the third position of a codon is one strategy that allows an organism to have a high G+C content without limiting access to the amino acids whose codons contain A or T residues in the first two positions.

Although the deduced amino acid sequence of the 65K antigen was particularly rich in alanine, glycine, leucine, and valine residues, the overall amino acid composition contained 52% hydrophobic and 48% hydrophilic residues. Computer-aided analysis of the alpha-helical content (7) and hydrophobicity (20) of the amino acid sequence revealed numerous regions that could participate in alpha-helical structures and no extended regions of high hydrophobicity. These data suggest that the 65K antigen is not an integral membrane protein but rather its sequence resembles a soluble protein.

As discussed above, the 65K antigen appeared to be a major T-cell immunogen and antigen in humans. It has been suggested that immunodominant T-cell epitopes are short stretches of amino acids that can form amphiphilic helices in which one side of the helix is hydrophobic and the other side is hydrophilic (2). Based on computer modeling, I have identified seven stretches of amino acids within the sequence

FIG. 4. Immunoblot analysis of products of the 540-amino-acid ORF. Cells were grown and induced (except for the sample in lane 2, panel A) and crude extracts were prepared as described in Materials and Methods. For each lane except lane 5, 200 μ g of protein was electrophoresed on a 10% Laemmli gel and transferred to nitrocellulose. For lane 5, 500 μ g of protein was loaded. The immobilized proteins were reacted with the IT-13 antibodies. (A) Lane 1, JM83; lane 2, JM83(pTB22), uninduced; lane 3, JM83(pTB22), induced with IPTG. (B) Lane 1, JM83(pTB12); lane 2, Y1089(λ SK116); lane 3, Y1089(λ RY3146); lane 4, BNN97[=C600(λ gt11)]; lane 5, JM83(pTB12).

of the 65K antigen that might form such amphiphilic helices. It is not known yet whether any of these regions actually interact with T cells.

The 517-amino-acid ORF. A second long ORF began with an ATG codon at positions 3948 to 3946 and extended to a TAA triplet at positions 2397 to 2395 on the DNA strand complementary to the DNA strand encoding the 65K antigen. Given that the two long ORFs are located downstream from each other on the complementary strands, one might expect that transcription of one gene might interfere with transcription of the other unless there were transcription termination signals within the intergenic region. Indeed, there were several short sequences (e.g., 2134 to 2160) within the 520-bp intergenic region that had features reminiscent of the transcription termination signals of gram-negative bacteria (33), i.e., regions containing short, G+C-rich inverted repeats capable of forming stem and loop structures followed by a stretch of three or more T residues about 20 bases from the center of dyad symmetry. Perhaps these inverted repeats function as transcription termination signals to allow the independent expression of each of these mycobacterial genes.

To determine whether the 517-amino-acid-coding ORF was expressed into protein in *E. coli*, extracts of cells containing a plasmid (pTB11) carrying the complete ORF were probed with polyclonal rabbit antisera elicited with a sonicated extract of *M. tuberculosis* bacteria in a Western blot assay. In these recombinants, the putative protein product of the 517-amino-acid ORF would have to be expressed from the mycobacterial regulatory sequences. The polyclonal antiserum detected more than 100 species in an extract of *M. tuberculosis* cells as well as the 65K antigen in extracts of *E. coli* cells carrying the appropriate plasmid (pTB12), but did not detect any novel proteins in extracts of *E. coli* cells containing plasmids carrying the 517-amino-acid ORF (data not shown). Hence, either this ORF is not expressed in *E. coli* from the mycobacterial regulatory sequences or the particular antiserum used in the immunoblots did not contain antibodies directed against this protein. It is not surprising that this ORF is not expressed in *E. coli*, since previous studies have suggested that most mycobacterial genes are not expressed in *E. coli* (8, 39).

Also, this ORF does not contain any impressive matches to the *E. coli* consensus promoter sequences within the 400 bases upstream of the ATG triplet, although it does contain a 3 of 5 match with the Shine-Dalgarno consensus sequence for ribosome-binding sites 12 bases upstream of the initiator ATG triplet. Nonetheless, given the size of this ORF and its unique structural features (discussed below), it may indeed be expressed into protein in *M. tuberculosis*.

The second long ORF could encode 517 amino acids or a protein of about 51 kDa (calculated M_r , 50,561). The deduced amino acid sequence was rich in alanine, asparagine, glycine, and serine and overall composed of 54% hydrophobic residues and 46% hydrophilic residues. The amino acid sequence of this protein did not display significant homologies with any of the protein sequences in the protein database (see above). The most striking features of this sequence occurred between amino acid residues 200 and 350. This region contained many repeats of short stretches of amino acids. For example, the 5-amino-acid sequence Asp-Asp-Ile-Gly was repeated three times consecutively. There was a direct repeat of a 14-amino-acid sequence with only one mismatch (residues 295 to 308 and 315 to 328). But perhaps the most interesting feature was a 5-residue sequence which displayed at least partial matches with more than 25 sequences in this region. The consensus sequence of this repeat appeared to be X-Gly-Asp-Y-Gly. For the 15 sequences matching this consensus sequence, X was most often phenylalanine, serine, or threonine (12 of 15) and Y was most often isoleucine, serine, or threonine (12 of 15). Fourteen additional sequences displayed partial matches with the consensus sequence (i.e., match two of the three core residues). These sequences were arranged in two sets of tandemly repeated sequences. The first set contained 16 consecutively repeated 5-residue sequences, while the second set contained 13 consecutively repeated matches to the consensus sequence. As expected, since the amino acid repeats were not exact, the nucleotide sequences in this region were not exact repeats. This observation suggests that recombinational processes such as unequal crossing-over may not play a role in causing rapid evolutionary changes in this region, as is often observed for highly repeated nucleotide sequences. The remainder of this protein sequence does not display any other particularly striking features.

The highly repetitious nature of the 517-residue protein is reminiscent of the repeated structures found in the major coat proteins of the sporozoite stage of the malaria parasite (reviewed in reference 32). These circumsporozoite (CS) proteins are 40- to 60-kDa proteins located on the membrane of the infectious sporozoite and contain a strongly immunodominant epitope which reacts with most of the antisporozoite antibodies found in polyclonal antisera as well as all of the monoclonal antibodies raised against the sporozoite stage. The central region of these proteins contains 20 to 40 tandemly arranged repeats of an 11- to 12-amino-acid sequence. In *Plasmodium falciparum*, the immunodominant epitope is contained within three consecutive repeats of the sequence Asn-Ala-Asn-Pro (which is repeated 37 times in one isolate), and antibodies directed against this 12-residue repeat can provide immunologic protection against infection with the malaria parasite. The sequence of the repeat differs in the various species of this parasite, and the number of repeats can vary within different isolates of the same species. The similarity of the repeated nature of the CS protein and that of the 517-amino-acid *M. tuberculosis* protein raises the interesting possibility that the

repeated sequences in the 517-residue protein play some role in the immune response to mycobacteria.

The availability of the recombinant DNA clones expressing these mycobacterial antigens, the nucleotide sequences of the mycobacterial genes, and the deduced amino acid sequences of the mycobacterial antigens may open up new approaches to the immunotherapy and immunodiagnosis of tuberculosis. One possibility is that 65K antigen produced in *E. coli* may form the basis of an antigen preparation to replace the crude preparations that are currently used in the immunodiagnosis of tuberculosis. Perhaps a chemically synthesized peptide corresponding to a portion of the 65K antigen might be useful in immunodiagnosis. These possibilities are being investigated by genetically engineering plasmids to express the 65K antigen or the 517-residue protein in *E. coli* as well as by chemically synthesizing portions of the proteins. Such antigen preparations will then be analyzed for their immunologic properties and potential clinical applications.

ACKNOWLEDGMENTS

I thank Cindy Krat and Sabine Schadow for excellent technical assistance and P. Minden, R. Ogata, J. G. Sutcliffe, and R. Young for helpful discussions and suggestions. I also thank my colleagues who made reagents available through the Committee for Research in the Immunology of Tuberculosis of the WHO, in particular R. A. Young for the recombinant DNA library and individual clones and T. M. Buchanan and J. Ivanyi for the anti-65K antibodies.

This work was supported in part by Public Health Service grant AI-22217 from the National Institute of Allergy and Infectious Diseases.

LITERATURE CITED

1. Athwal, R. S., S. S. Deo, and T. Imaeda. 1984. Deoxyribonucleic acid relatedness among *Mycobacterium leprae*, *Mycobacterium lepraemurium*, and selected bacteria by dot blot and spectrophotometric deoxyribonucleic acid hybridization assays. *Int. J. Syst. Bacteriol.* **34**:371-375.
2. Berzofsky, J. A. 1985. Intrinsic and extrinsic factors in protein antigenic structure. *Science* **229**:932-940.
3. Biggin, M. D., T. J. Gibson, and G. F. Hong. 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. *Proc. Natl. Acad. Sci. USA* **80**:3963-3965.
4. Bloom, B. R., and T. Godal. 1983. Selective primary health care: strategies for control of disease in the developing world. *V. Leprosy. Rev. Infect. Dis.* **5**:765-780.
5. Brow, M. A. D., R. Pesin, and J. G. Sutcliffe. 1985. The tetracycline repressor of pSC101. *Mol. Biol. Evol.* **2**:1-12.
6. Chaparas, S. D. 1982. The immunology of mycobacterial infections. *Crit. Rev. Microbiol.* **9**:139-197.
7. Chou, P. Y., and G. D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* **47**:45-148.
8. Clark-Curtiss, J. E., W. R. Jacobs, M. A. Docherty, L. R. Ritchie, and R. Curtiss. 1985. Molecular analysis of DNA and construction of genomic libraries of *Mycobacterium leprae*. *J. Bacteriol.* **161**:1093-1102.
9. Coates, A. R. M., J. Hewitt, B. W. Allen, J. Ivanyi, and D. A. Mitchison. 1981. Antigenic diversity of *Mycobacterium tuberculosis* and *Mycobacterium bovis* detected by means of monoclonal antibodies. *Lancet* **ii**:167-169.
10. Collins, F. M. 1982. The immunology of tuberculosis. *Am. Rev. Respir. Dis.* **125**:42-49.
11. Daniel, T. M., and B. W. Janicki. 1978. Mycobacterial antigens: a review of their isolation, chemistry, and immunological properties. *Microbiol. Rev.* **42**:84-113.
12. Dannenberg, A. M. 1982. Pathogenesis of pulmonary tuberculosis. *Am. Rev. Respir. Dis.* **125**:25-29.
13. Emmrich, F., J. Thole, J. van Embden, and S. H. E. Kaufmann.

1986. A recombinant 64 kilodalton protein of *Mycobacterium bovis* Bacillus Calmette-Guérin specifically stimulates human T4 clones reactive to mycobacterial antigens. *J. Exp. Med.* **163**: 1024-1029.
14. Engers, H. D., M. Abe, B. R. Bloom, V. Mehra, W. Britton, T. M. Buchanan, J. Ivanyi, A. H. J. Kolk, and C. C. Shepard. 1985. Results of a World Health Organization-sponsored workshop on monoclonal antibodies to *Mycobacterium leprae*. *Infect. Immun.* **48**:603-605.
15. Engers, H. D., V. Houbas, J. Bennedsen, T. M. Buchanan, S. D. Chaparas, G. Kadival, O. Closs, J. R. David, J. D. A. van Embden, T. Godal, S. A. Mustafa, J. Ivanyi, D. B. Young, S. H. E. Kaufmann, A. G. Khomenko, A. H. J. Kolk, M. Kubin, J. A. Louis, P. Minden, T. M. Shinnick, L. Trnka, and R. A. Young. 1986. Results of a World Health Organization-sponsored workshop to characterize antigens recognized by mycobacteria-specific monoclonal antibodies. *Infect. Immun.* **51**:718-720.
16. Gillis, T. P., and T. M. Buchanan. 1982. Production and partial characterization of monoclonal antibodies to *Mycobacterium leprae*. *Infect. Immun.* **37**:172-178.
17. Gillis, T. P., R. A. Miller, D. B. Young, S. R. Khanolkar, and T. M. Buchanan. 1985. Immunochemical characterization of a protein associated with *Mycobacterium leprae* cell wall. *Infect. Immun.* **49**:371-377.
18. Grange, J. M. 1984. The humoral immune response in tuberculosis: its nature, biological role and diagnostic usefulness. *Adv. Tuberc. Res.* **21**:1-78.
19. Helms, C., M. Y. Graham, J. E. Dutchik, and M. V. Olson. 1985. A new method for purifying lambda DNA from plate lysates. *DNA* **4**:39-49.
20. Hopp, T. P., and K. P. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78**:3824-3828.
21. Huynh, T. V., R. A. Young, and R. W. Davis. 1985. Constructing and screening cDNA libraries in λ gt10 and λ gt11, p. 49-78. *In* D. Glover (ed.), *DNA cloning techniques: a practical approach*, vol. 1. IRL Press, Oxford.
22. Imaeda, T. 1985. Deoxyribonucleic acid relatedness among selected strains of *Mycobacterium tuberculosis*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium microti* and *Mycobacterium africanum*. *Int. J. Syst. Bacteriol.* **35**:147-150.
23. Ivanyi, J., J. A. Morris, and M. Keen. 1985. Studies with monoclonal antibodies to mycobacteria, p. 59-90. *In* A. J. L. Macario and E. C. Macario (ed.), *Monoclonal antibodies against bacteria*. Academic Press, Inc., New York.
24. Laemmli, U. K. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature (London)* **227**:680-685.
25. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
26. Maxam, A. M., and W. Gilbert. 1976. A new method of sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**:560-564.
27. Messing, J., and J. Vieira. 1982. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. *Gene* **19**:269-276.
28. Miller, J. H. 1972. *Experiments in molecular genetics*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
29. Minden, P., P. J. Kelleher, J. H. Freed, L. D. Nielsen, P. J. Brennan, L. McPherson, and J. K. McClatchy. 1984. Immunological evaluation of a component isolated from *Mycobacterium bovis* BCG with a monoclonal antibody to *M. bovis* BCG. *Infect. Immun.* **46**:519-525.
30. Mustafa, A. S., H. K. Gill, A. Nerland, W. J. Britton, V. Mehra, B. R. Bloom, R. A. Young, and T. Godal. 1986. Human T-cell clones recognize a major *M. leprae* protein antigen expressed in *E. coli*. *Nature (London)* **319**:63-66.
31. Niman, H. L., R. A. Houghten, L. E. Walker, R. A. Reisfeld, I. A. Wilson, J. M. Hogle, and R. A. Lerner. 1983. Generation of protein-reactive antibodies by short peptides is an event of high frequency: implications for the structural basis of immune recognition. *Proc. Natl. Acad. Sci. USA* **80**:4949-4953.
32. Nussenzweig, V., and R. S. Nussenzweig. 1985. Circumsporozoite proteins of malaria parasites. *Cell* **42**:401-403.
33. Rosenberg, M., and D. Court. 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu. Rev. Genet.* **13**:319-353.
34. Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. Rose. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *J. Mol. Biol.* **143**:161-178.
35. Shapira, S. K., J. Chou, F. V. Richaud, and M. J. Casadaban. 1983. New versatile plasmid vectors for expression of hybrid proteins encoded by a cloned gene fused to *lacZ* gene sequences encoding an enzymatically active carboxy terminal portion of β -galactosidase. *Gene* **25**:71-82.
36. Shine, J., and L. Dalgarno. 1974. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to non-sense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA* **71**:1342-1346.
37. Snider, D. E. 1982. The tuberculin skin test. *Am. Rev. Respir. Dis.* **125**:108-118.
38. Stanford, J. L. 1983. Immunologically important constituents of mycobacteria: antigens, p. 85-127. *In* C. Ratledge and J. Stanford (ed.), *The biology of the mycobacteria*, vol. 2. Academic Press, Ltd., London.
39. Thole, J. E. R., H. G. Dauwerse, P. K. Das, D. G. Groothuis, L. M. Schouls, and J. D. A. van Embden. 1985. Cloning of *Mycobacterium bovis* BCG DNA and expression of antigens in *Escherichia coli*. *Infect. Immun.* **50**:800-806.
40. Towbin, H., T. Staehelin, and J. Gordon. 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. USA* **76**:4350-4354.
41. Wayne, L. G., and W. M. Gross. 1968. Base compositions of deoxyribonucleic acid isolated from mycobacteria. *J. Bacteriol.* **96**:1915-1919.
42. Yanisch-Perron, C., J. Vieira, and J. Messing. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**:103-119.
43. Young, R. A., B. R. Bloom, C. M. Grosskinsky, J. Ivanyi, D. Thomas, and R. W. Davis. 1985. Dissection of *Mycobacterium tuberculosis* antigens using recombinant DNA. *Proc. Natl. Acad. Sci. USA* **82**:2583-2587.
44. Young, R. A., and R. W. Davis. 1983. Yeast RNA polymerase II genes: isolation with antibody probes. *Science* **222**:778-782.
45. Young, R. A., V. Mehra, D. Sweetser, T. M. Buchanan, J. Clark-Curtiss, R. W. Davis, and B. R. Bloom. 1985. Genes for the major protein antigens of *Mycobacterium leprae*. *Nature (London)* **316**:450-452.