INTERNATIONAL JOURNAL OF LEPROSY Volume 62, Number 1
Printed in the U.S.A.

The Genome of *Mycobacterium leprae**

The behavior of all life forms is determined by a series of complex messages found in coded form in their genes, and the term "genome" is commonly used to refer to the complete set of generic material found within an organism. In bacteria, the genome generally consists of one, sometimes two chromosomes and these are usually circular, although linear chromosomes have been described in a few special cases. Extra-chromosomal elements, such as episomes, plasmids or phages, often provide an additional source of genetic information, and are valid constituents of the genome. The discovery of the universal genetic code in the 1960s, and the development of DNA cloning and sequencing techniques in the 1970s have "revolutionized" biological research by allowing genes to be isolated and their functions to be deduced from the coded message they contain.

The genome of *Mycobacterium leprae* can be considered as a book containing all the information necessary for the growth, survival, reproduction and especially the pathogenicity of the microbe. Thanks to the power of current molecular biological techniques, it has been possible to divide that book into chapters and to start to decipher the text contained therein through the identification and characterization of the various genes. In many cases, one can predict the function of the gene products, the proteins, as homologous counterparts with similar sequences described and studied in other organisms. These predictions can subsequently be tested by means of further genetic, biochemical and immunological techniques and, in the case of *M. leprae,* this may be facilitated by the use of surrogate genetics in which the genes are introduced into suitable heterologous hosts such as *M. smegmatis* or *M. bovis* BCG.

An integrated strategy was devised to tackle the problem of mapping and sequencing the genome of *M. leprae,* and this involved analysis of chromosomal DNA by pulsed field gel electrophoresis (PFGE) of large restriction fragments and the construction of ordered libraries in cosmid vectors. Owing to the difficulties of cultivating *M. leprae,* the former approach was unsuccessful and, in consequence, the cosmid approach was employed. Cosmids are small vectors, with a large capacity for foreign DNA (up to 45 kilobases), which facilitate the introduction of recombinant DNA molecules into *Escherichia coli* as they exploit the packaging and infection systems of phage lambda. Conventional cosmids can be used only with *E. coli* but are ideal substrates for DNA sequencing projects since they are large but not unmanageable. Shuttle cosmids can be used with both *E. coli* and mycobacteria and, thus, are valuable tools for studying biological functions since they enable *M. leprae* genes to be faithfully expressed in cultivable mycobacterial hosts.

There are at least five good reasons to justify the use of ordered libraries: 1) *M. leprae* is an obligately intracellular pathogen that cannot be cultivated *in vitro* so cloning its DNA, in the form of large pieces, generates a readily renewable source. 2) Ordered libraries are valuable tools that considerably facilitate genetic research and, on many occasions, it has been found that genes for related functions are linked. 3) Comparison of the maps obtained from the ordered clones of the chromosomes of related organisms, e.g., *M. tuberculosis* and *M. leprae,* could lead to the identification of loci that have been modified or truncated which may be associated with pathogenesis or slow growth. 4) Ordered shuttle libraries will allow us to do surrogate genetics and to identify the genes for virulence factors, protective antigens, etc. 5) They also represent an excellent starting point for a genome sequencing project.

A fingerprinting technique has been used to order the cosmids, and this involved generating a set of radio-labeled restriction fragments, characteristic of each clone, that were resolved by gel electrophoresis to generate typical patterns or "fingerprints." These were subjected to computer-assisted matching techniques to detect regions common to several clones and then assembled

* This State-of-the-Art Lecture in Microbiology was presented on 30 August 1993 at the XIV International Leprosy Congress in Orlando, Florida, U.S.A.

into "contigs" or blocks of contiguous DNA sequences. At present the *M. leprae* chromosome can be represented as four contigs and by summing their sizes one can estimate the chromosome to have a size of about 2.8 megabases. If *M. leprae* were a typical bacteria it would have a circular chromosome. Since we have four contigs, this suggests that we have four gaps in the collection, probably as a result of unclonable sequences. Nevertheless, this is a remarkable achievement since very few ordered libraries of bacterial chromosomes exist and those that do exist were obtained from organisms that are easy to cultivate.

All of the cloned *M. leprae* genes have been positioned on the contig map, by means of hybridization, together with a set of genes of a well-conserved sequence from other bacteria. There are over 80 loci on the present gene map, including the *M. leprae*-specific repetitive sequence RLEP.

To obtain more insight into the structure, function and organization of the *M. leprae* chromosome, a genome sequencing project was initiated with the aim of systematically sequencing large stretches of chromosomal DNA. If we return to the earlier analog in which the chromosome was likened to a book, then each cosmid that is sequenced can be considered as a chapter and the information that it contains will help us to unravel the mysteries of the life of *M. leprae.* The message obtained from the first chapter, deduced from the sequence of cosmid B1790, can be summarized. Two criteria were used to identify the genes. Firstly, their open reading frames, defined by the distance between two stop codons, should be at least 80 codons long. Secondly, the codons employed should conform to the pattern found in known *M. leprae* genes. In this way 12 candidate genes were found, and the functions of 11 of them were established by comparison of their sequences with those of all of the genes and proteins present in the biological databases.

Four of these genes encode proteins that are, or could be, potential drug targets: *rpoB* encodes the β subunit of RNA polymerase which is the target for rifampin, the backbone of the multidrug therapy currently used to control leprosy. The sequence information obtained has been very useful for developing a simple PCR-based diagnostic test for monitoring rifampin resistance (this will be described elsewhere during the Congress by Nadine Honoré). The *rpsL* gene encodes ribosomal protein S12, which is known to be the target for streptomycin in other bacteria, including *M. tuberculosis,* while *efg* and *tuf* code for two elongation factors, G and Tu, required for protein synthesis. Fusidic-acid resistance in bacteria is associated with mutations in *efg* and tetracycline resistance can result from mutations in *tuf.* Since *M. leprae* is susceptible to both fusidic acid and minocycline, knowledge of the sequence of the target genes will enable diagnostic tests for resistance to be developed, should these drugs find widespread clinical application in leprosy control programs.

One of the surprising features which emerged on reading the first chapter was the scarcity of the genes because only 40% of the potential coding capacity was used. In fast-growing bacteria, such as *E. coli,* more than 85% of the sequence is coding and there is a clear succession of genes. This does not appear to be the case in *M. leprae* and, to confirm this, it was decided to sequence many more cosmids. This was done in conjunction with Doug Smith at Collaborative Research in Boston, Massachusetts, U.S.A., who was sponsored by the NIH Human Genome program to test the feasibility of a new DNA sequencing technique. A further nine cosmids were sequenced completely and analyzed for genes and their functions predicted. Again, it was clear that genes were relatively scarce and that noncoding regions were extensive. Based on these findings one can predict that *M. leprae* contains only about 1000 protein-coding genes compared to the estimated 4000 present in *E. coli.* It is conceivable that this gene sparsity, and the consequent reduction in metabolic potential, account for the very slow growth of *M. leprae.*

However, as with all rules, there is a striking exception because one of the cosmids was found to carry a very large amount of coding information. On comparison of the sequences, it was apparent that the proteins encoded were involved in the synthesis of polyketides. These are interesting organic compounds produced by a variety of bacteria and fungi, and they include many pigments such as actinorhodin, antibiotics such as erythromycins, and toxins, like members

of the aflatoxin family. In mycobacteria, examples of polyketides could include cell-wall components, such as phenolic glycolipid or mycoserosic acid. Since these compounds are specific for mycobacteria, they represent attractive targets for chemotherapy and immunotherapy.

Mechanistically, polyketide biosynthesis is very similar to that of fatty acids as nascent chains are extended by the addition of two-carbon units in a step-wise manner catalyzed by a series of enzymatic reactions. In the first step, acetyl-CoA is bound to fatty-acid synthase (FAS), a large multifunctional enzyme, via a thiol group in the acyl carrier protein (ACP) domain, then transferred to a second thiol group by an acyl transferase activity (AT). Acyl transferase is again involved in the binding of malonyl-CoA to the freed thiol group, and this is condensed with the bound acetyl-CoA in a step catalyzed by ketoacyl synthase (KS), and subsequently subjected to a series of reactions involving ketoreductase (KR), dehydratase (DH), and enoylreductase (ER). In the terminal step, the product is released from fatty-acid synthase by a thioesterase, and the enzyme is free to catalyze further fatty-acid synthesis. Likewise, the reaction product can be used as a substrate by other FASs.

On inspection of the genes present on cosmid L518, it was clear that several of them encode huge proteins bearing strong resemblances to fatty-acid synthases. Three of them have ACP and AT motifs, 4 of them carry KS domains, 2 have KR motifs, and 1 an ER domain. The close proximity of the genes and the remarkable functional similarity of their products suggest that they probably interact and effect consecutive steps in the synthesis of a novel polyketide in *M. leprae.* The gene arrangement is typical of a polycistronic operon, and the putative transcript would be over 30,000 nucleotides long. At its distal end, this transcript carries the information required to produce two membrane-bound proteins and a third polypeptide capable of binding ATP. Since these proteins are all highly homologous to proteins involved in the resistance to, and the secretion of, a polyketide antibiotic in *Streptomyces,* it seems highly likely that they will be involved in the export of the putative *M. leprae* polyketide. Clearly, extensive biochemical studies are warranted.

To date, ten of the chapters in the leprosy book have been read and a further ten are being "browsed" and will be finished before the end of 1993. This means that we will soon have at our disposal over a third of the information required by *M. leprae* to infect, survive, and multiply in man. To date, 140 genes have been sequenced and putative functions attributed to 94 of them on the basis of their homologies with known gene products. Already we have learned a great deal about the biology of *M. leprae* from its genes, and we have been able to identify a variety of housekeeping functions, as well as pathways required for the biosynthesis of heme, purines, amino acids and peptidoglycan.

At this stage it should be apparent that the genome project is generating enormous amounts of information about the genetics, biochemistry and immunology of *M. leprae.* In order to make this readily accessible to the research community, we have developed a customized database, MycDB, using the object-orientated software ACeDB to ensure that efficient storage, annotation and dissemination of the data can occur. MycDB runs on a variety of computers and workstations that use the UNIX operating system and a WINDOWS system. Information is accessed by selecting (clicking) with the mouse button on different objects. Any one piece of information can lead to further relevant information since they are all interconnected.

On starting the program the first window displays the various "classes" stored in the database and any one of these, for instance the chromosome class, can be selected by clicking. Having selected the mycobacterial chromosome of interest, one can then select a particular gene, for instance the *groEL* gene of *M. leprae,* and access the information linked to it. As more windows are opened different classes of interrelated information can be obtained. Thus, one can go from the gene to its product, the 65-kDa protein antigen, and learn what monoclonal antibodies are available that recognize it and from whom they can be obtained. Likewise one can interrogate MycDB with respect to the location of the *groEL* gene and, subsequently, obtain its sequence and other relevant information. A literature service is also available which contains the abstracts of selected papers and in most cases pro-

vides information about the author, his address, and telephone and fax numbers. MycDB has been developed with the support of the World Health Organization and the Association Française Raoul Follereau, and is supplied free of charge on request or can be downloaded via the INTERNET computer network using anonymous FTP.

To conclude, I hope that I have convinced you that the *M. leprae* genome project is advancing at great pace, that it is uncovering many new leads for research in chemotherapy and rational drug design, that it is opening up new avenues for biochemical and immunological investigation and, above all, is making a significant contribution to the fight against leprosy.

Many thanks to my coworkers and to you for your attention.

—Stewart T. Cole, Ph.D.

Unite de Genetique Moleculaire Bacterienne
Institut Pasteur
28 Rue du Dr. Roux
75724 Paris 15, France