Clinical review

Science, medicine, and the future

Microbial genome sequencing—beyond the double helix

Peter J Jenks

Considerable advances in automated DNA sequencing and bioinformatics have made the sequencing of entire microbial genomes a reality (see fig 1). Researchers have been quick to recognise the scientific value of microbial sequence data, and there is now a frantic rush to sequence a wide range of important micro-organisms. The continuing release of data from these projects, combined with advances in techniques used to investigate pathogenesis, will lead to an unprecedented understanding of micro-organisms and how they cause disease.

In this article I discuss how this sequence information will affect the practice of clinical microbiology and the management of infectious diseases. In particular, I examine how the large scale study of the entire genetic complement of micro-organisms should result in the provision of more rapid and efficient diagnostic techniques, as well as the development of new antimicrobial strategies.

Microbial genome sequencing

In the past few years a large number of clinically important bacteria, parasites, and fungi have been sequenced, and genomic data for many human pathogens will be available by the turn of the century (see box). Because many of the early sequencing efforts were funded by private industry, concerns were raised about ownership and freedom of access to the data generated by these projects. As sequencing technology has become more refined and efficient, however, the cost of such projects has fallen, and funding is increasingly being provided by government agencies and medical charities. This has greatly improved access to microbial sequence information, and data from most projects are deposited, often on a daily basis, on websites on the internet.

A comprehensive catalogue of all genome sequencing projects is provided by the Magpie Genome Sequencing Project List (http://www.ncsc.anl.gov/home/gaasterl/magpie.html), and links to individual projects are also available at this site. Other useful resources include the website of the Institute for Genomic Research (http://www.tigr.org), which has performed much of the pioneering work on genomics, and the Pasteur Institute's home page (http://www.pasteur.fr), which has a version in English and provides cross referencing links to related sites in general microbiology, medicine, and molecular biology.

Towards a better understanding of microbial pathogenesis

In order to understand how micro-organisms cause disease, it is necessary to identify those microbial gene products specifically implicated in the infection process. By characterising the strategies used by different organisms, it is possible to appreciate why certain species or strains are associated with clinical disease, whereas others are merely harmless commensals. The genome sequence of a pathogenic micro-organism is a tremendous resource, but it is essentially only an inventory of all its genes. To unravel the complexities of pathogenesis, we must identify exactly which of these genes are associated with colonisation of the host, persistence, and disease causation.

Screening for virulence genes

With modern bioinformatic software, it is possible to screen out potential virulence genes by computer analysis of the genome sequence of a micro-organism. The power and simplicity of this approach became clear after the publication of the H influenzae genome.1 Before this genome had been published, researchers

Predicted developments

Comprehensive study of microbial pathogenesis and the interaction between pathogens and their hosts

Identification of sensitive and specific molecular targets suitable for microbial identification, typing, and for use as markers of antimicrobial resistance

Discovery of microbial molecular markers associated with substantial variance in the risk and severity of disease

Selection of potential candidates for the rational development of new therapeutic agents and vaccines

Improved understanding of normal human gene function and provision of experimental systems for the further analysis of human hereditary disease

1568 BMJ 1998;317:1568–71
had identified only seven genes involved in the biosynthesis of lipopolysaccharide, an important virulence factor. A search of the entire genome of *H. influenzae* for lipopolysaccharide genes known to exist in other organisms (a homology search) identified 25 new biosynthetic genes. Further characterisation of these may lead to the identification of a potential target for a vaccine against non-type b strains of *H. influenzae*. Similarly, it is possible to search for sequences that act as markers for virulence genes. In *H. influenzae* and other bacteria the expression of cell surface molecules (which allow the organism to adhere to host cells, adapt to the host’s microenvironment, and avoid the host’s immune defences) is regulated by multiple repeating sequences. These repeating sequences can therefore be used as markers for genes important in mucosal colonisation and the interaction between the microbe and its host.

Computer analysis can also be used to identify “pathogenicity islands,” hot spots within microbial genomes that contain blocks of important virulence genes. Because these distinct genetic elements usually have a different nucleotide composition to the rest of the bacterial chromosome, they can be identified with sophisticated computer software. By comparing the entire genome sequences of micro-organisms, we should be able to determine which factors are conserved for pathogenicity and which are unique to certain pathogens. Consequently, it will be possible to make predictions of clinical outcome on the basis of the genetic makeup of particular species or strains.

Biological and physiological studies

Although computer analysis of sequence data can rapidly generate new information on micro-organisms, this technique has its limitations and often only provides clues about pathogenic mechanisms. To complete the functional analysis of an organism and obtain meaningful information on how pathogens cause disease, it is necessary to perform biological and physiological studies. The traditional method of studying microbial pathogenesis has been to “knock out” or mutate an individual gene and then characterise the effect this has on the micro-organism. Normally, mutating a gene will result in the loss of a certain function (for example, loss of motility), and it is then possible to determine if loss of this function affects the ability of the micro-organism to cause disease in an animal model. This gene by gene approach is laborious and time consuming, particularly when even the smallest free living micro-organism (*Mycoplasma genitalium*) has 470 genes. Since the advent of genome sequencing the challenge facing researchers has been to develop techniques with high throughput that allow the simultaneous examination of all the genes of a micro-organism in a systematic fashion.

One system that has been developed for the large scale study of pathogenesis is signature tagged mutagenesis. In this technique, a large number of genes (usually about 100) are mutated simultaneously, so creating a pool of knockout mutants. As each gene is knocked out, a unique identifying “bar code” (the tag) is inserted into each knockout mutant, so making it possible to identify any of the 100 mutants within the pool. This means that the entire pool of knockout mutants can be studied together (rather than one by one), and any individual mutant that shows interesting features can be pulled out of the pool by means of the bar code identification system. This technique has proved particularly useful for identifying new virulence factors in animal models. A pool of bar coded knockout mutants is used to inoculate and cause disease in the animal. Individual mutants that are not subsequently recovered from the animal have lost the ability to establish infection and therefore harbour a mutation in a gene likely to be involved in causing disease. These particular knockout mutants can be picked out of the pool, and the role of the mutated gene in pathogenesis can be studied further. Signature tagged mutagenesis is an approach that can be used in a wide range of micro-organisms and has already identified important virulence genes in *Salmonella typhimurium*, *Staphylococcus aureus*, and *Vibrio cholerae*.

Other techniques that allow a more global approach to the study of microbial pathogenesis have also recently been developed. An important feature of these approaches is that they can be used either in animal models of infection or in conditions that mimic those encountered within a host cell. This allows specific identification of microbial factors preferentially expressed during the infection process and which therefore play a crucial role in the development of dis-

### Large scale projects for sequencing microbial genomes that are completed or in progress

<table>
<thead>
<tr>
<th>Bacteria and related organisms</th>
<th>Parasites</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Borrelia burgdorferi</em></td>
<td><em>Brugia malayi</em></td>
</tr>
<tr>
<td><em>Escherichia coli</em> O157</td>
<td><em>Cryptosporidium parvum</em></td>
</tr>
<tr>
<td><em>Haemophilus influenzae</em></td>
<td><em>Leishmania major</em></td>
</tr>
<tr>
<td><em>Helicobacter pylori</em></td>
<td><em>Plasmodium falciparum</em></td>
</tr>
<tr>
<td><em>Mycobacterium leprae</em> and <em>M tuberculosis</em></td>
<td><em>Pneumocystis carinii</em></td>
</tr>
<tr>
<td><em>Mycoplasma pneumoniae</em></td>
<td><em>Schistosoma mansoni</em></td>
</tr>
<tr>
<td><em>Neisseria gonorrhoeae</em> and <em>N meningitidis</em></td>
<td><em>Trypanosoma brucei</em> and <em>T cruzi</em></td>
</tr>
<tr>
<td><em>Rickettsia prowazekii</em></td>
<td><em>Staphylococcus aureus</em></td>
</tr>
<tr>
<td><em>Streptococcus pneumoniae</em> and <em>S pyogenes</em></td>
<td><em>Aspergillus nidulans</em></td>
</tr>
<tr>
<td><em>Treponema pallidum</em></td>
<td><em>Candida albicans</em></td>
</tr>
<tr>
<td><em>Vibrio cholerae</em></td>
<td><em>Saccharomyces cerevisiae</em></td>
</tr>
</tbody>
</table>
New approaches to diagnosis

While enormous changes have occurred elsewhere in clinical practice, the routine diagnosis of microbial infection by clinical microbiology laboratories has changed little in the past 50 years. Identification of micro-organisms continues to rely heavily on morphological characterisation, which is both time consuming and labour intensive. Despite considerable advances in molecular biology, molecular diagnostic techniques are generally restricted to the identification of fastidious, slow growing, or dangerous organisms such as Chlamydia trachomatis and Mycobacterium tuberculosis. As well as problems of cost and practicability, the paucity of truly sensitive and specific molecular targets and of cheap, easy to use methods of detection have been major obstacles to routine implementation of these techniques.

As the entire genome sequences of many common human pathogens become available, it will finally be possible to discover if specific genetic markers do exist for individual pathogens and if it is possible to distinguish between micro-organisms on the basis of these sequences. Moreover, it will also be possible to identify genetic markers that will assist in typing (strain specific markers), in predicting clinical outcome (virulence markers), and in predicting susceptibility to antimicrobial drugs (resistance markers). Once these targets have been identified, the challenge will be to incorporate them into standardised, automated assays that will allow the rapid and efficient examination of large numbers of clinical specimens.

Classic detection of genetic markers is achieved by hybridisation (in which the target sequence pairs up and binds to a complementary “probe” sequence) or by genetic amplification of the target sequence (such as by the polymerase chain reaction).

Gene amplification

Currently, gene amplification can be used to detect only a small number of human pathogens, including Mycoplasma pneumoniae and Bordetella pertussis. Recently, there has been interest in developing amplification techniques that might allow the detection of more than one target sequence (and hence more than one pathogen) in a sample. This can be achieved with multiplex polymerase chain reaction, a technique that uses two or more sets of primers to amplify multiple target DNA sequences in the same reaction. With such an approach, it would theoretically be possible to screen a clinical specimen for all commonly found pathogens (for example, a stool specimen for the microbial causes of diarrhoea). Unfortunately, the number of primer pairs that can be included in the same amplification reaction is limited, and other factors, such as the need to remove inhibitory substances from clinical specimens, also prevent the wide scale adoption of such techniques. Although gene amplification may prove useful for diagnosing specific microbial infections, it is unlikely to have a major role in the large scale examination of specimens in a clinical diagnostic laboratory.

Hybridisation

The recent development of high density oligonucleotide arrays has revolutionised hybridisation technology, and these “DNA chips” have exciting potential applications for the diagnosis of micro-organisms. DNA chips consist of thousands of specific oligonucleotide probes (short fragments of DNA) arranged on glass or nylon supports (the array) about the size of a postage stamp. They are constructed by means of light directed oligonucleotide synthesis, a modification of a technique used in the computer chip industry, which ensures that each probe has its own specific coordinates on the array. Because so many probes can be included on such a small area, these DNA chips can be used to screen for an enormous number of potential target DNA sequences and are therefore ideal for examining clinical samples for microbial pathogens. Target DNA would be extracted from the sample and labelled with fluorescent dye, before being hybridised to the miniaturised probe (see fig 2). The result of each hybridisation reaction is a distinctive pattern of fluorescence on the array, the coordinates of the positive fluorescent spots allowing identification of the target sequences in the sample.

Once information on microbial genomes is available, it will be possible to design DNA chips that contain representative markers for all potential pathogens that may be present in individual clinical samples. For example, a chip could contain genetic markers for the pathogens of the central nervous system. If the target sequence of Neisseria meningitidis was present in the cerebrospinal fluid of a patient with meningitis, this would hybridise to the appropriate probe on the DNA chip and be identified by the coordinates of the gener-
Drug and vaccine development

New approaches to the development of drugs and vaccines are desperately needed given the current problems of antimicrobial resistance and the fact that it has been almost 30 years since a new class of antibiotic was discovered. Once decoded, the entire genome sequence of a micro-organism can be examined for potential target sites for attack, opening new avenues for the rational development of antimicrobial strategies. Identification of those microbial factors essential for the viability of an organism or those preferentially expressed during the infection process would allow the selection of potential molecular targets for antimicrobial design.

The development of new vaccines is also likely to be greatly accelerated by the availability of data on microbial genomes. With sophisticated computer software, all the proteins of a micro-organism could be examined for those that may be protective. This is achieved by identifying proteins that are known to be major antigens in other species or predicting which proteins are expressed on the cell surface and therefore likely to be immunogenic. Once identified, these can be purified and studied for protective effect, either alone or in synergy with other antigens. A system that permits direct screening of the entire genome of a micro-organism for immunogenic antigens has also recently been described.15 Such a system has considerable potential for the development of both prophylactic and therapeutic vaccines.

Insights into human biology and disease

Microbial genome sequencing and the application of new molecular techniques will inevitably reveal more information on the complexities of microbial pathogenesis and the interaction between pathogens and their hosts. Further examination of how pathogens exploit mammalian host cell function will also assist our understanding of human cell biology.14

Analysis of microbial genome sequences has also revealed that many proteins implicated in human hereditary diseases are also found in micro-organisms.15 These conditions include hereditary non-polyposis colon cancer, Wilson’s disease, and adrenoleukodystrophy, and further homologues will probably be identified as other microbial sequencing projects are completed. Micro-organisms can therefore be used as convenient models for studying the function of genes implicated in human hereditary diseases. For example, the human gene implicated in adrenoleukodystrophy, a neurodegenerative disease characterised by defective long chain fatty acid metabolism, is similar to two genes identified in the yeast Saccharomyces cerevisiae. Further characterisation of these yeast genes has shown that they are involved in the transport of activated long chain fatty acids, suggesting that this is the underlying defect in patients with adrenoleukodystrophy.16

As well as predicting the biological function of genes mutated in human hereditary diseases, the study of conserved proteins in micro-organisms will ultimately lead to improvements in the diagnosis and treatment of these conditions. Microbial genome sequencing is likely to play an increasingly important role in the analysis of newly discovered human genes and provide further clues to the molecular basis of hereditary diseases.

I thank Dr Agnès Labigne for helpful comments during the preparation of this manuscript. I also thank the Station de Microscopie Electronique, Institut Pasteur, and Catherine Chevalier for providing images used in the figures.

Funding: PJ is supported by a research training fellowship in medical microbiology from the Wellcome Trust.

Conflict of interest: None.