Genomics and computational molecular biology
Douglas L Brutlag

There has been a dramatic increase in the number of completely sequenced bacterial genomes during the past two years as a result of the efforts both of public genome agencies and the pharmaceutical industry. The availability of completely sequenced genomes permits more systematic analyses of genes, evolution and genome function than was otherwise possible. Using computational methods—which are used to identify genes and their functions including statistics, sequence similarity, motifs, profiles, protein folds and probabilistic models—it is possible to develop characteristic genome signatures, assign functions to genes, identify pathogenic genes, identify metabolic pathways, develop diagnostic probes and discover potential drug-binding sites. All of these directions are critical to understanding bacterial growth, pathogenicity and host–pathogen interactions.

Introduction
During 1995 and 1996, five complete microbial genomes were determined [1–5]. During 1997, seven more genomes were completely sequenced, including yeast [6–13]. The Institute for Genomic Research (TIGR) microbial database (http://www.tigr.org/tdb/mdb/mdb.html) now lists 41 additional bacterial genomes being sequenced in various laboratories around the world. (A list of Web sites to these and other major genomic resources is given in Table 1.) This explosion of genome sequences is the direct result of public and private investment in genome efforts.

The goals of sequencing bacterial genomes are manifold [14–19]. For instance, the complete sequence of many pathogenic species are being determined with the hope of understanding the disease process. Along with such understanding comes the ability to develop molecular diagnostic probes (both nucleic acid probes and antigenic determinants) and the ability to define new drug targets and vaccines to treat infections caused by these organisms.

The identification of pathogenic genes [20–24], drug targets and potential antigenic sites requires the combined use of laboratory and computational approaches. Genomics, the science of genome analysis and the mapping of genes to specific traits and phenotypes, can specify unique molecular probes that define both the organisms and the antibiotic sensitivity of those organisms. Bioinformatics, sometimes referred to as functional genomics, helps identify gene function and can be used to understand the disease process and identify drug targets unique to the infectious agent.

Limitations on the length of this article permit me to mention only the most recent and major advances in techniques for gene identification but there are a number of other reviews and compendiums that cover this area in more depth [25,26,27,28,29–33]. In addition, Table 1 includes a list of Web sites of most of the bioinformatics methods that are presented in this review.

Computational methods for gene identification
The first step in gene identification is the location of coding regions or open reading frames (ORFs). This task is simplified in bacteria because of the absence of splicing. Sequencing errors and translational frameshifting [34] can lead to partial protein sequences or interrupted ORFs but these are often resolved during the early steps of gene identification by sequence similarity with proteins from other organisms [35–40]. In the absence of homologous sequences in other organisms—and especially with short bacterial genes—probabilistic gene models (hidden Markov models) can often identify biologically significant coding regions [41,42].

Pairwise sequence homology
Given a database of potential ORFs, many methods can be used to define the biological function of the putative proteins. The most commonly applied methods search for sequence similarity of the translated ORFs with a database of known protein sequences [43,44,45,46]. The search for gene function is usually carried out at the protein level to eliminate the redundancy of the genetic code. In addition, the use of amino acid substitution matrices that describe the acceptable replacements permits the discovery of even distantly related protein homologies [47,48].

One of the most sensitive methods for comparing two sequences, which uses amino acid scoring matrices and allows penalties for the presence of gaps in the alignment, is the original subsequence alignment algorithm of Smith and Waterman [44]. Because of the computational complexity of this approach, several approximations have been developed such as FASTA [43], BLAST [49] as well as parallelized versions of the algorithm and implementations of the algorithm in specialized hardware [50–53]. The primary advantage of the original Smith and
Table 1

Web sites of major genome resources and critical bioinformatics methods.

<table>
<thead>
<tr>
<th>Web site</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic resources</td>
<td></td>
</tr>
<tr>
<td>Genome Sequence Database</td>
<td><a href="http://www.ncgr.org/gsdb/">http://www.ncgr.org/gsdb/</a></td>
</tr>
<tr>
<td>Genome Sequencing Projects</td>
<td><a href="http://www.mcs.anl.gov/home/gaasterl/genomes.html">http://www.mcs.anl.gov/home/gaasterl/genomes.html</a></td>
</tr>
<tr>
<td>Microbial Genomes Mail List</td>
<td><a href="http://www.mailbase.ac.uk/lists/microbial-genomes/">http://www.mailbase.ac.uk/lists/microbial-genomes/</a></td>
</tr>
<tr>
<td>National Human Genome Research Institute</td>
<td><a href="http://www.nhgri.nih.gov/">http://www.nhgri.nih.gov/</a></td>
</tr>
<tr>
<td>The Sanger Centre</td>
<td><a href="http://www.sanger.ac.uk/">http://www.sanger.ac.uk/</a></td>
</tr>
<tr>
<td>TIGR Microbial Database</td>
<td><a href="http://www.tigr.org/tdb/mdb/mdb.html">http://www.tigr.org/tdb/mdb/mdb.html</a></td>
</tr>
<tr>
<td>UK Genome Mapping Project</td>
<td><a href="http://www.hgmp.mrc.ac.uk/">http://www.hgmp.mrc.ac.uk/</a></td>
</tr>
<tr>
<td>Bioinformatics resources</td>
<td></td>
</tr>
<tr>
<td>BLOCKS database</td>
<td><a href="http://www.blocks.fhcrc.org/">http://www.blocks.fhcrc.org/</a></td>
</tr>
<tr>
<td>CATH Protein Structure Classification</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/cath/">http://www.biochem.ucl.ac.uk/bsm/cath/</a></td>
</tr>
<tr>
<td>DSSP Database of Secondary Structure</td>
<td><a href="http://www.sander.ebi.ac.uk/dssp/">http://www.sander.ebi.ac.uk/dssp/</a></td>
</tr>
<tr>
<td>FSSP Fold Structure Database</td>
<td><a href="http://www2.ebi.ac.uk/dali/fssp/">http://www2.ebi.ac.uk/dali/fssp/</a></td>
</tr>
<tr>
<td>GeneQuiz</td>
<td><a href="http://columbia.ebi.ac.uk:B765/ext-genequiz/">http://columbia.ebi.ac.uk:B765/ext-genequiz/</a></td>
</tr>
<tr>
<td>IDENTIFY Database</td>
<td><a href="http://motif.stanford.edu/identify/">http://motif.stanford.edu/identify/</a></td>
</tr>
<tr>
<td>Pfam HMM Database</td>
<td><a href="http://genome.wustl.edu/Pfam/">http://genome.wustl.edu/Pfam/</a></td>
</tr>
<tr>
<td>PRINTS Database</td>
<td><a href="http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/">http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/</a> PRINTS.html</td>
</tr>
<tr>
<td>Sequence Alignment and Modeling UCSC</td>
<td><a href="http://www.cse.ucsc.edu/research/compbio/sam.html">http://www.cse.ucsc.edu/research/compbio/sam.html</a></td>
</tr>
<tr>
<td>Structural Classification of Proteins</td>
<td><a href="http://scop.mrc-lmb.cam.ac.uk/scop/">http://scop.mrc-lmb.cam.ac.uk/scop/</a></td>
</tr>
</tbody>
</table>

*This table does not contain direct pointers to genome databases but these are to be found in many of the following links.

Waterman algorithm is the ability to allow gaps in the alignments, which is particularly important for detecting distantly related protein sequences [54**,55]. A version of BLAST was recently developed that now allows gaps (Gapped-BLAST) [45**], however, a detailed comparison of the sensitivity of this gapped-BLAST method versus Smith-Waterman method has not yet been published.

One of the most important features of the pairwise sequence similarity methods is that each of them calculates the likelihood that the similarity score would be observed by chance using an extreme value distribution [56–58]. This approach has been shown to give a reliable estimate of the probability of such similarities occurring by chance. Such estimates are especially important when thousands of similarity searches are being carried out in the analysis of an entire genome.

Comparing coding regions with protein families

If no significant similarity is detected using pairwise sequence comparison methods, then other approaches that compare the potential coding regions to entire families of sequences are applied. These methods include profiles, templates, hidden Markov models or the more general position-specific scoring matrices [59,60,61*,62,63,64**]. Comparing a protein sequence to a profile or a position-specific scoring matrix is more sensitive than pairwise comparisons because the matrix represents a mathematical average of an entire protein family. This averaging tends to emphasize the conserved features of the sequences in the family and places less emphasis on the features of any single sequence.

A method has been developed recently for rapidly forming families of aligned sequences and developing a profile from them based on Gapped-BLAST. This method, called Psi-BLAST, can rapidly build a sequence profile from the most significant sequence similarities in a BLAST search [45**]. This profile can then be used to re-search the protein databases for even more distantly related members of the family. What is needed now is a complete compendium of all the Psi-BLAST families that can be derived from the entire protein database and which can be updated routinely as more genomes are sequenced.

Searching for shorter conserved regions

If no significant homology is observed for a particular ORF using all of the similarity methods mentioned above, then that ORF is scanned by other methods which search for shorter regions of conservation such as those that represent conserved sequence motifs (consensus sequences), BLOCKS, PRINTS or domains [65,66,67**–70**].

A major problem with using many of the protein sequence motifs in the literature is that most of them have been constructed to be as sensitive as possible (encompassing all known examples) which leads to a marked decrease in specificity; that is, they often include a significant number of ‘false hits’ [69,71**]. Such sequence motifs are useful for testing individual coding regions for the presence of potential functional site but they are generally not useful for searching thousands of proteins in an entire genome automatically.

Methods have now been developed that can generate very specific protein sequence motifs (called EMOTIFs)
which make less than one false prediction per $10^8$ to $10^{10}$ tests and yet maintain a high level of sensitivity [66]. The specificity of EMOTIFs are adequate for searching entire proteomes (all the proteins encoded by a genome) with a tolerable level of false predictions. A database of >50,000 EMOTIFs has been developed which can be used to identify >7000 different biological functions in protein sequences [67**]. This approach can assign function to ORFs in the absence of any extensive similarity to an existing sequence.

The BLOCKS and PRINTS databases of aligned protein families can also be used to search entire genomes if the expectation threshold is set appropriately high. The advantages of searching for short conserved regions using EMOTIFs, BLOCKS or PRINTS is that the likelihood of a false positive result is associated with every finding, permitting the user to decide the significance of the finding. Secondly, a protein’s function can often be determined without any known homologs, as long as it shares one or more motifs with a known superfamily. A final, and most important, use of short motifs is that they can often serve as potential drug targets. The conserved regions in protein families often represent active sites, binding sites, or parts of such biologically critical sites.

Protein threading and fold recognition
If all of the above methods fail to identify a gene, then one must usually resort to ‘protein threading’ methods such as three-dimensional profiles or compatibility with known protein folds [72–76]. Several databases of common protein folds also exist [77**,78,79,80**,81–83]. The problem with these methods is that knowing the fold of a protein by itself tells one very little about its biological function. Nevertheless, when used in conjunction with other data—marginal sequence similarities, motifs with marginally significant hits—structural similarities and threading approaches can help confirm gene identification. One of the most popular approaches to gene identification is the ‘GeneQuiz’ system which uses multiple sequence alignments of homologous protein families and structural information in addition to homology and motifs to identify genes [84,85**].

Global analyses of bacterial genomes
One of the most exciting results of having the complete sequence of a bacterial genome is the possibility of examining the entire metabolism, regulation and organization of genes and sequences therein [86,87**,88–90]. One can also know with measurable assurance that specific functions are absent from the genome. With the availability of multiple genomes, comparative and evolutionary studies are also possible [86,87**,89]. One of the more elusive goals is to attempt to discover the minimal gene set at which cellular life can exist [91–94]. The wide variety of environmental conditions under which bacteria can grow makes defining a minimal gene set a difficult task. The complete sequence of a genome also permits one to mutate each and every gene and examine such mutations functionally and genetically too [95,96**].

Global analysis of genome sequences has also led to novel findings. Examination of dinucleotide, trinucleotide and higher frequencies of oligonucleotides in bacterial genomes has resulted in the definition of a genomic signature that is characteristic for bacteria and can classify bacteria into biologically related classes [97,98**,99,100]. Examining dinucleotide frequencies in 50 kb long windows has shown much more homogeniety within a single species and striking differences between species. The frequencies of dinucleotides in bacterial genomes constitute a genomic signature that can be used to track bacterial DNA segments. It also raises the problem of the origin and maintenance of the sequence biases in the first place.

Conclusions
The availability of complete genomes sequences of many bacterial species is, for the first time, enabling many novel experimental approaches. The complete definition of all the gene products by gene identification techniques described here is just the first step. Mapping all gene products to functions and all functions to metabolism will confer the ability to predict the phenotype of an organism with highly increased certainty. We will be able to locate critical pathways and steps in pathogenesis; to target these steps with new drugs; and to target the infectious agents with new vaccines. We will be able to follow protein structure and function through evolution with confidence. In addition, we will be able to engineer bacteria for specific purposes (drug production, toxic waste removal, etc.) and minimize possible adverse consequences.

Acknowledgements
The author would like to acknowledge support from the National Library of Medicine (grant LM 05716), the Howard Hughes Medical Institute and SmithKline Beecham for their generous support of the research in his laboratory.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest

•• of outstanding interest


28. Gusfield, D: Algorithms on Strings, Trees and Sequences. ** Cambridge, UK: Cambridge University Press; 1997. This is an excellent introductory book to the field of sequence computing written for the mathematician and computer scientist. Because of the extensive explanatory text, it is quite accessible to molecular biologists too.


This new method for finding sequence similarities and building up sequence profiles of related sequences has been the primary sequence similarity method used by molecular biologists.


54. Shepaer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkappler T: Sensitivity and selectivity in protein similarity searches: a

This is one of the most systematic and objective comparisons of three sequence similarity search methods. Using an entire database of queries and optimizing for the amino acid substitution matrices shows that the rigorous Smith and Waterman method is preferable for the most sensitive similarity search.


85. GeneQuiz is an intelligent system for combining partial or marginal data from a number of sequence and structure similarity methods to either confirm or deny putative gene assignments. By combining data in this way, GeneQuiz can discover functional assignments that must other methods miss.


88. This is the latest publication from the combined groups of Peter Karp and Monica Riley on their encyclopedic work concerning the metabolism of E. coli. Their database contains substrates, reactions, enzymes, pathways, genetic maps and pointers to the protein sequence for the bulk of the E. coli gene products. This database will gain increased value as a standard for comparison with other bacterial genomes as evidenced in this paper.


96. Lashkari DA, McCusker JH, Davis RW: Whole genome analysis: experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR. Proc Natl Acad Sci USA 1997, 94:8945-8947.

Given a complete sequence of every gene in an organism, the methods described in this paper permit one to PCR amplify each gene independently. In addition, similar methods will permit one to mutate every gene and study its function under different selective conditions. This paper marks the beginning of massively parallel genome analysis.


This paper is the latest in a series from the Karlin group analyzing the dinucleotide genome signatures that characterize the bacterial genome. This signatures permit the classification of bacteria into evolutionary related classes and it will be interesting to determine the mechanism of the formation and maintenance of these dinucleotide biases.
