

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.

Addresses

Department of Biochemistry, Beckman Center B400, Stanford University, Stanford, California 94305-5307, USA;
e-mail: brutlag@stanford.edu

Current Opinion in Microbiology 1998, 1:340–345

<http://biomednet.com/elecref/1369527400100340>

© Current Biology Ltd ISSN 1369-5274

Abbreviation

ORF open reading frame

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.

Web site	URL*
Genomic resources	
EcoCyc Metabolic Database	http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html
Genome Sequence Database	http://www.ncgr.org/gsdb/
Genome Sequencing Projects	http://www.mcs.anl.gov/home/gaasterl/genomes.html
Microbial Genomes Mail List	http://www.mailbase.ac.uk/lists/microbial-genomes/
National Human Genome Research Institute	http://www.nhgri.nih.gov/
The Sanger Centre	http://www.sanger.ac.uk/
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdb.html
UK Genome Mapping Project	http://www.hgmp.mrc.ac.uk/
Bioinformatics resources	
BLOCKS database	http://www.blocks.fhcrc.org/
CATH Protein Structure Classification	http://www.biochem.ucl.ac.uk/bsm/cath/
DSSP Database of Secondary Structure	http://www.sander.ebi.ac.uk/dssp/
FSSP Fold Structure Database	http://www2.ebi.ac.uk/dali/fssp/
Gapped-BLAST and Psi-BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
GeneQuiz	http://columba.ebi.ac.uk:8765/ext-genequiz/
IDENTIFY Database	http://motif.stanford.edu/identify/
Pfam HMM Database	http://genome.wustl.edu/Pfam/
PRINTS Database	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html
PROSITE Database	http://www.expasy.ch/sprot/prosite.html
Sequence Alignment and Modeling UCSC	http://www.cse.ucsc.edu/research/compbio/sam.html
Structural Classification of Proteins	http://scop.mrc-lmb.cam.ac.uk/scop/

*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 ORFs 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 homogeneity 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 genome 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
1. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM *et al.*: **The minimal gene complement of *Mycoplasma genitalium***. *Science* 1995, **270**:397-403.
 2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**:496-512.
 3. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD *et al.*: **Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii***. *Science* 1996, **273**:1058-1073.
 4. Himmelreich R, Hilbert H, Plagens H, Pirkel E, Li BC, Herrmann R: **Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae***. *Nucleic Acids Res* 1996, **24**:4420-4449.

5. Kaneko T, Tanaka A, Sato S, Kotani H, Sazuka T, Miyajima N, Sugiura M, Tabata S: **Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64% to 92% of the genome (supplement).** *DNA Res* 1995, **2**:191-198.
6. Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers MMA, Alexandraki D *et al.*: **The yeast genome directory.** *Nature (suppl)* 1997, **387**.
7. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA *et al.*: **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388**:539-547.
8. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD *et al.*: **The complete genome sequence of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*.** *Nature* 1997, **390**:364-370.
9. Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K *et al.*: **Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics.** *J Bacteriol* 1997, **179**:7135-7155.
10. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277**:1453-1474.
11. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessières P, Bolotin A, Borchert S *et al.*: **The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390**:249-256.
12. Smith DR, Richterich P, Rubenfield M, Rice PW, Butler C, Lee HM, Kirst S, Gundersen K, Abendschan K, Xu Q *et al.*: **Multiplex sequencing of 1.5 Mb of the *Mycobacterium leprae* genome.** *Genome Res* 1997, **7**:802-819.
13. Fraser CM, Fleischmann RD: **Strategies for whole microbial genome sequencing and analysis.** *Electrophoresis* 1997, **18**:1207-1216.
14. Dujon B: **The yeast genome project: what did we learn?** *Trends Genet* 1996, **12**:263-270.
15. Danchin A: **Why sequence genomes? The *Escherichia coli* imbroglio [letter].** *Mol Microbiol* 1995, **18**:371-376.
16. Davies JE: **Redundant genome sequencing? [letter].** *Science* 1996, **273**:1155.
17. Coleb ST: **Why sequence the genome of *Mycobacterium tuberculosis*?** *Tuber Lung Dis* 1996, **77**:486-490.
18. Olson MV: **A time to sequence.** *Science* 1995, **270**:394-396.
19. Yamagishi A, Oshima T: **What we can learn from the whole genome sequence of an archaeon *Methanococcus jannaschii*.** *Tanpakushitsu Kakusan Koso* 1997, **42**:174-177.
20. Cabello F: **Pathogenicity islands: important but not unique factors contributing to *Salmonella* virulence [letter].** *Trends Microbiol* 1997, **5**:431-432.
21. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H: **Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution.** *Mol Microbiol* 1997, **23**:1089-1097.
22. Ochman H, Groisman EA: **Distribution of pathogenicity islands in *Salmonella* spp.** *Infect Immun* 1996, **64**:5410-5412.
23. Groisman EA, Ochman H: **Pathogenicity islands: bacterial evolution in quantum leaps.** *Cell* 1996, **87**:791-794.
24. Lee CA: **Pathogenicity islands and the evolution of bacterial pathogens.** *Infect Agents Dis* 1996, **5**:1-7.
25. Doolittle RF: **Protein sequence comparisons: searching databases and aligning sequences.** *Curr Opin Biotechnol* 1994, **5**:24-28.
26. Doolittle, RF: *Computer Methods for Macromolecular Sequence Analysis*. New York: Academic Press; 1996.
This volume contains some of the best descriptions of bioinformatics and gene identification methods written by the authors who developed them.
27. Brutlag DL, Sternberg MJE (Eds): **Sequences and Topology.** In *Current Opinion in Structural Biology*. Edited by Hendrickson W, Blundell TL. London: Current Biology Ltd; 1996:343-406.
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.
29. Adams MD, Fields C, Venter JC: *Automated DNA Sequencing and Analysis*. New York: Academic Press; 1994.
30. Bishop MJ: *Guide to Human Genome Computing*. London: Academic Press; 1994.
31. Waterman MS: **Genomic sequence databases.** *Genomics* 1990, **6**:700-701.
32. Waterman, MS: **Computer analysis of nucleic acid sequences.** *Methods Enzymol* 1988, **164**:765-793.
33. Lander ES, Waterman MS: *Calculating the Secrets of Life: Applications of the Mathematical Sciences in Molecular Biology*. Washington DC: National Academy Press; 1995.
34. Farabaugh PJ: **Programmed translational frameshifting.** *Annu Rev Genet* 1996, **30**:507-528.
35. Koonin EV, Tatusov RL, Rudd KE: **Protein sequence comparison at genome scale.** *Methods Enzymol* 1996, **266**:295-322.
36. Claverie JM: **Effective large-scale sequence similarity searches.** *Methods Enzymol* 1996, **266**:212-227.
37. Taylorb P: **GeneJockeyII. Translation and open reading frame analysis.** *Methods Mol Biol* 1997, **70**:221-225.
38. Guerdoux-Jamet P, Rislér JL: **Searching for a family of orphan sequences with SAMBA, a parallel hardware dedicated to biological applications.** *Biochimie* 1996, **78**:311-314.
39. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci USA* 1996, **93**:9061-9066.
40. Sze SH, Pevzner PA: **Las Vegas algorithms for gene recognition: suboptimal and error-tolerant spliced alignment.** *J Comput Biol* 1997, **4**:297-309.
41. Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Midigue C, Danchin A: **Detection of new genes in a bacterial genome using Markov models for three gene classes.** *Nucleic Acids Res* 1995, **23**:3554-3562.
42. Yada T, Hirosawa M: **Detection of short protein coding regions within the cyanobacterium genome: application of the hidden Markov model.** *DNA Res* 1996, **3**:355-361.
43. Lipman DJ, Pearson WR: **Rapid and sensitive protein similarity searches.** *Science* 1985, **227**:1435-1441.
44. Smith TF, Waterman M: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-197.
45. Altschul SF, Madden TL, Schdffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
•• 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.
46. Pearson WR: **Comparison of methods for searching protein sequence databases.** *Protein Sci* 1995, **4**:1145-1160.
47. Henikoff S: **Scores for sequence searches and alignments.** *Curr Opin Struct Biol* 1996, **6**:353-360.
48. Henikoff, JG, Henikoff, S: **BLOCKS database and its applications.** *Methods Enzymol* 1996 **266**:88-105.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
50. Hughey R: **Parallel hardware for sequence comparison and alignment.** *Comput Appl Biosci* 1996, **12**:473-479.
51. Brutlag DL, Dautricourt JP, Diaz R, Fier J, Moxon B, Stamm R: **BLAZE: an implementation of the Smith-Waterman Comparison Algorithm on a massively parallel computer.** *Comput Chem* 1993, **17**:203-207.
52. Julich A: **Implementations of BLAST for parallel computers.** *CABIOS* 1995, **11**:3-6.
53. Chen ES, Asano C, Davison DB: **Parallel alignment of DNA sequences on the connection machine CM-2.** *Comput Appl Biosci* 1993, **9**:375.
54. Shpaer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkapiller T: **Sensitivity and selectivity in protein similarity searches: a**

comparison of Smith-Waterman in hardware to BLAST and FASTA. *Genomics* 1996, **38**:179-191.

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.

55. Shpaer EG: **GeneAssist. Smith-Waterman and other database similarity searches and identification of motifs.** *Methods Mol Biol* 1997, **70**:173-187.
56. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87**:2264-2268.
57. Altschul SF, Gish W: **Local alignment statistics.** In *Methods in Enzymology*. Edited by Doolittle R. New York: Academic Press; 1996:460-480.
58. Altschul SF: **A protein alignment scoring system sensitive at all evolutionary distances.** *J Mol Evol* 1993, **36**:290-300.
59. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling.** *J Mol Biol* 1994, **235**:1501-1531.
60. Sonnhammer EL, Kahn D: **Modular arrangement of proteins as inferred from analysis of homology.** *Protein Sci* 1994, **3**:482-492.
61. Bucher P, Karplus K, Moeri N, Hofmann K: **A flexible motif search technique based on generalized profiles.** *Comput Chem* 1996, **20**:3-23.

The flexibility of this novel motif permits it to represent a wide variety of different protein structures and functions. It will become more popular in the future when there is sufficient sequence information for a large number of protein families so that these models can be accurately estimated.

62. Henikoff JG, Henikoff S: **Using substitution probabilities to improve position-specific-scoring matrices.** *Comput Appl Biosci* 1996, **12**:135-143.
63. Gribskov M, Veretnik S: **Identification of sequence patterns with profile analysis.** *Methods Enzymol* 1996, **266**:198-211.
64. Sonnhammer EL, Eddy SR, Durbin R: **Pfam: a comprehensive database of protein domain families based on seed alignments.** *Proteins* 1997, **28**:405-420.

This database of protein domains will grow with time. Because of its sensitivity, it will most likely be the primary tool used with pairwise sequence similarity fails.

65. Attwood TK, Avison H, Beck ME, Bewley M, Bleasby AJ, Brewster F, Cooper P, Degtyarenko K, Geddes AJ, Flower DR *et al.*: **The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology.** *J Chem Inf Comput Sci* 1997, **37**:417-424.
66. Nevill-Manning C, Sethi K, Wu TD, Brutlag DL: **Enumerating and ranking discrete motifs.** *ISMB-97* 1997, **4**:202-209.

67. Nevill-Manning CG, Wu TD, Brutlag DL: **Discovering function in genomic databases using highly specific sequence motifs.** *Proc Natl Acad Sci USA* 1998, **95**:in press.

This new database of highly specific sequence motifs are made from the sequence alignments in BLOCKS and PRINTS databases and are specific enough to search entire genomes without resulting in false predictions.

68. Henikoff JG, Pietrokovski S, Henikoff S: **Recent enhancements to the BLOCKS Database servers.** *Nucleic Acids Res* 1997, **25**:222-225.

The BLOCKS database was the first probabilistic motif database constructed completely automatically and can often discover protein function in the complete absence of a homologous sequence in the protein database.

69. Bairoch A, Bucher P, Hofmann K: **The PROSITE database, its status in 1997.** *Nucleic Acids Res* 1997, **25**:217-221.
70. Attwood TK, Beck ME, Bleasby AJ, Degtyarenko K, Michie AD, Parry-Smith DJ: **Novel developments with the PRINTS protein fingerprint database.** *Nucleic Acids Res* 1997, **25**:212-217.

Like the BLOCKS database, the PRINTS database provides an extremely sensitive motif search facility. PRINTS is also characterized by its excellent annotation.

71. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence data bank and its supplement TrEMBL.** *Nucleic Acids Res* 1997, **25**:31-36.

The SWISS-PROT database has long been the stalwart for protein comparison. SWISS-PROT is characterized by its highly regular syntax, numerous database hyperlinks and excellent annotation.

72. Rost B, Schneider R, Sander C: **Protein fold recognition by prediction-based threading.** *J Mol Biol* 1997, **270**:471-480.
73. Taylor WR: **Multiple sequence threading: an analysis of alignment quality and stability.** *J Mol Biol* 1997, **269**:902-943.
74. Bowie JU, Zhang K, Wilmanns M, Eisenberg D: **Three-dimensional profiles for measuring compatibility of amino acid sequence with three-dimensional structure.** In *Methods Enzymol*. Edited by Doolittle R. New York: Academic Press; 1996:598-616.
75. Fischer D, Rice D, Bowie JU, Eisenberg D: **Assigning amino acid sequences to 3-dimensional protein folds.** *FASEB J* 1996, **10**:126-136.
76. Lathrop RH, Smith TF: **Global optimum protein threading with gapped alignment and empirical pair score functions.** *J Mol Biol* 1996, **255**:641-665.
77. Hubbard TJP, Murzin AG, Brenner SE, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 1997, **25**:236-239.

Although not made automatically, the SCOP structural classification of proteins is very useful, particularly because of its well organized hierarchical structure.

78. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH - a hierarchic classification of protein domain structures.** *Structure* 1997, **5**:1093-1108.
79. Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins.** *Nucleic Acids Res* 1996, **24**:206-209.
80. Holm L, Sander C: **Dali/FSSP classification of three-dimensional protein folds.** *Nucleic Acids Res* 1997, **25**:231-234.
81. Chothia C, Hubbard T, Brenner S, Barns H, Murzin A: **Protein folds in the all-beta and all-alpha classes.** *Annual Rev Biophys Biomol Struct* 1997, **26**:597-627.
82. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**:536-540.
83. Brenner SE, Chothia C, Hubbard TJ, Murzin AG: **Understanding protein structure: using SCOP for fold interpretation.** *Methods Enzymol* 1996, **266**:635-642.
84. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C: **GeneQuiz: a workbench for sequence analysis.** *ISMB* 1994, **2**:348-353.
85. Casari G, Ouzounis C, Valencia ACS: **GeneQuiz II: automatic function assignment for genome sequence analysis.** In *First Annual Pacific Symposium on Biocomputing*. Hawaii, USA: World Scientific; 1996:707-709.

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 most other methods miss.

86. Karp PD, Ouzounis C, Paley S: **HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*.** *ISMB* 1996, **4**:116-124.
87. Karp PD, Riley M, Paley SM, Pellegrini-Toole A, Krummenacker M: **EcoCyc: encyclopedia of *E. coli* genes and metabolism.** *Nucleic Acids Res* 1998, **26**:50-53.

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.

88. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: **Prediction of enzyme classification from protein sequence without the use of sequence similarity.** *ISMB* 1997, **5**:92-99.
89. Tamames J, Casari G, Ouzounis C, Valencia A: **Conserved clusters of functionally related genes in two bacterial genomes.** *J Mol Evol* 1997, **44**:66-73.
90. Ouzounis C, Casari G, Sander C, Tamames J, Valencia A: **Computational comparisons of model genomes.** *Trends Biotechnol.* 1996, **14**:280-285.
91. Goffeau A: **Life with 482 genes.** *Science* 1995, **270**:445-446.

92. McFadden GI, Gilson PR, Douglas SE, Cavalier-Smith T, Hofmann CJ, Maier UG: **Bonsai genomics: sequencing the smallest eukaryotic genomes.** *Trends Genet* 1997, **13**:46-49.
93. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.
94. Koonin EV: **Big time for small genomes.** *Genome Res* 1997, **7**:418-421.
95. Bassett DE Jr, Connelly C, Hyland KM, Kitagawa K, Mayer ML, Morrow DM, Page AM, Resto VA, Skibbens RV, Hieter P: **Exploiting the complete yeast genome sequence.** *Curr Opin Genet Dev* 1996, **6**:763-766.
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.

97. Karlin S, Mrazek J: **Compositional differences within and between eukaryotic genomes.** *Proc Natl Acad Sci USA* 1997, **94**:10227-10232.
98. Karlin S, Mrazek J, Campbell AM: **Compositional biases of bacterial genomes and evolutionary implications.** *J Bacteriol* 1997, **179**:3899-3913.
- 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.
99. Karlin S, Burge C: **Dinucleotide relative abundance extremes: a genomic signature.** *Trends Genet* 1995, **11**:283-290.
100. Karlin S, Ladunga I, Blaisdell BE: **Heterogeneity of genomes: measures and values.** *Proc Natl Acad Sci USA* 1994, **91**:12837-12841.