J. Mol. Microbiol. Biotechnol. (2002) 4(4): 357–365.

JMMB Review

Bacterial Bioinformatics: Pathogenesis and the Genome

Kelly Paine* and Darren R. Flower

The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire, RG20 7NN, United Kingdom.

Abstract

As the number of completed microbial genome sequences continues to grow, there is a pressing need for the exploitation of this wealth of data through a synergistic interaction between the well-established science of bacteriology and the emergent discipline of bioinformatics. Antibiotic resistance and pathogenicity in virulent bacteria has become an increasing problem, with even the strongest drugs useless against some species, such as multi-drug resistant *Enterococcus faecium* and *Mycobacterium tuberculosis*. The global spread of Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome (AIDS) has contributed to the re-emergence of tuberculosis and the threat from new and emergent diseases. To address these problems, bacterial pathogenicity requires redefinition as Koch's postulates become obsolete. This review discusses how the use of bacterial genomic information, and the *in silico* tools available at present, may aid in determining the definition of a current pathogen. The combination of both fields should provide a rapid and efficient way of assisting in the future development of antimicrobial therapies.

Introduction

When antibiotics were first introduced in the 1940's, they were hailed as miracle drugs, and quickly provided effective therapy for many of the more dangerous pathogens then prevalent. However, resistance to these antimicrobials developed quickly. A recent World Health Organisation report into antimicrobial resistance published online [http://www.who.int/infectious-disease-report/2000/index.html], notes that formerly curable bacterial diseases are on the increase. For example, 98% of all South-East Asian gonorrhoea cases are presently multi-drug resistant, while up to 60% of nosocomial infections in the developed world are caused by drug-resistant and often opportunistic pathogens. Infections with rare virulent micro-organisms like *Acinetobacter* are also on the increase (Brown *et al.*, 1998), and opportunistic bacterial infections such as *Pseudomonas aeruginosa* and *Salmonella* spp. are becoming more common. Several factors contributing to this phenomenon during five decades of antibiotic mishandling have included: health workers misdiagnosing illness or providing the wrong prescription, patients failing to adhere to treatment, and the misuse of antimicrobials in animals with secondary effects observed in humans (Van den Bogaard and Stobberingh, 2000). The major factor causing a sharp rise in TB infections has been HIV infection, in that patients are immunocompromised and unable to respond to other microbial infections; one third of the world's HIV positive population are now infected with TB. This correlates with the spread of resistant *Mycobacterium tuberculosis*. An additional problem is that patients who suffer from persistent TB will often be a continuous transmission source for the rest of their community (Godfrey-Faussett *et al.*, 2000). It is evident that some current therapies are no longer effective, and accordingly, novel antimicrobials will need to be developed.

To help counteract these problems, advances in technology can be used to hasten the hunt for new drug and vaccine targets. One obvious advantage of using computer-based screening techniques to scan newly sequenced pathogen genomes is the speed at which identification of novel targets can be carried out (Rosamund and Allsop, 2000; James *et al.*, 2000). Weinstock (Weinstock, 2000) commented on the use of bioinformatics and bacterial genomic data to find new mechanisms of virulence, and eventually, targets for novel antimicrobials. Bioinformatics itself can be defined as utilising large databases of biological information with specific *in silico* tools to complement traditional wet laboratory-based biology (Murray-Rust, 1994). This review aims to define what identifies a "21st century" pathogen, and how molecular and computational biology can help this understanding. The use of bioinformatics to explore bacterial pathogenicity will be discussed, as well as the future for the combined uses of these two disciplines.

Genome Sequencing and Analysis

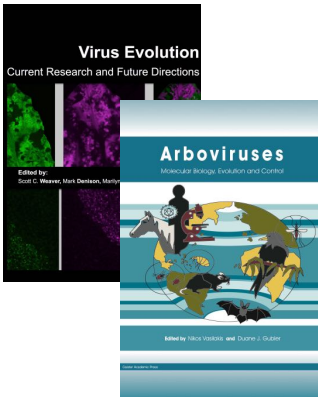
The first whole organism to be sequenced was that of bacteriophage ϕX174, by Sanger (Sanger *et al.*, 1977). His group also went on to complete the genome of bacteriophage λ, making use of cloned restriction endonuclease fragments. Whole genome sequencing is now commonplace; the first genome from a *free-living* organism, *Haemophilus influenzae* was published in 1995 (Fleischmann *et al.*, 1995), and others followed rapidly. At present (March 2002) there are 59 complete *bacterial* genomes in the TIGR microbial database [http://www.tigr.org/tdb/mdb/mdbcomplete.html], 35 of which

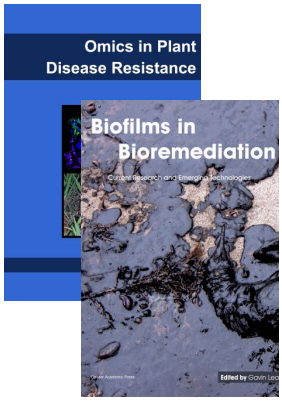
*For correspondence. Email kelly.paine@jenner.ac.uk; Tel. +44 (0) 1635 577968; Fax. +44 (0) 1635 577908.

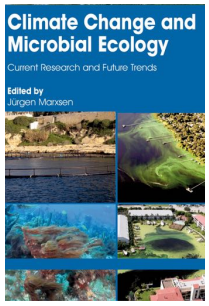
Further Reading

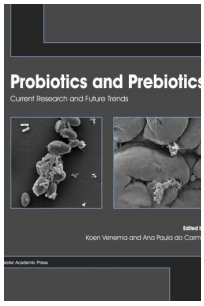
Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at **caister.com**

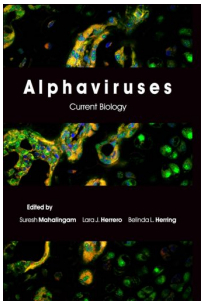
- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof
- ***Aspergillus* and *Penicillium* in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2
- **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins
- **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr
- **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles
- **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate
- **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem
- **Microalgae: Current Research and Applications**
  Edited by: MN Tsaloglou (2016)
  www.caister.com/microalgae
- **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma
- **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol
- **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo
- ***Shigella*: Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella
- **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romaní, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms
- **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha
- **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile

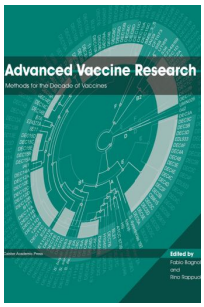


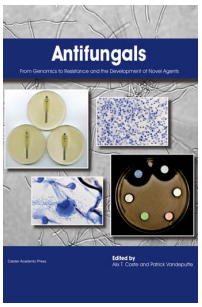












- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow
- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics
- **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015
- ***Corynebacterium glutamicum*: From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory2
- **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines
- **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals
- **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  Edited by: J Murillo, BA Vinatzer, RW Jackson, et al. (2015)
  www.caister.com/bacteria-plant
- ***Aeromonas***
  Edited by: J Graf (2015)
  www.caister.com/aeromonas
- **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics
- ***Leishmania*: Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2
- ***Acanthamoeba*: Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2
- **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2
- **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order

are pathogenic to humans. This is in addition to the 145 bacterial genomes in progress, of which 64 are pathogenic to humans. The food-borne pathogen *Escherichia coli* O157: H7 was one of the most recent bacterial genomes completed (Perna *et al.*, 2001). Such a flood of raw sequence information obviously requires refinement and further analysis. Both *in silico* and *in vitro* research can assist in this area.

In the field of raw genomic sequence annotation, *in silico* prediction of genes and open reading frames (ORFs) (Burge and Karlin, 1998; Delcher *et al.*, 1999) enables the rapid identification of coding regions in prokaryotic sequences. One of the difficulties in discovering microbial genes in a stretch of sequence is that the gene density is greatly increased in prokaryotes. At its simplest, one assumes that any ORF above a reasonable threshold (~300 bps) contains a coding sequence or gene, but in doing so, smaller moieties may be missed. Bacterial genomes may also contain overlapping coding regions due to frameshift mutations (Schmatkov *et al.*, 1999). Therefore, the preferred prediction software used in prokaryotic genome interpretation is GLIMMER (Delcher *et al.*, 1999), which caters for overlapping ORFs and high gene density using an interpolated Markov Model (IMM). GeneScan may also be used, although it is more suited to the analysis of eukaryotic genomes (Ramakrishna and Srinivasan, 1999). A recent review by (Fraser *et al.*, 2000) gives a good overview of the genome annotation process. Often, the information is displayed in an online format like the EcoGene database (Rudd, 2000), a re-annotated version of the original published genome that includes features such as additional literature references and corrections to the genomic data. Other databases are derived from comparisons of sequenced microbial genomes, and include the HOBACGEN system (Perriere *et al.*, 2000) and Paulsen *et al.*'s transporter classification (TC) scheme (Paulsen *et al.*, 2000). Comparative genomics allow the similarities between known sequences and a new dataset to be analysed, and may identify families of related proteins.

Another recent comparative genomic study, published in *Yeast*, detailed two major rearrangements in the BCG vaccine bacterium *Mycobacterium bovis* strain BCG Pasteur chromosome, when compared to the whole genome of *M. tuberculosis* H37Rv (Brosch *et al.*, 2000). This has implications for the attenuation and immunogenicity of the BCG vaccine, the most widely used globally. Using the proteome derived from the *M. tuberculosis* H37Rv genome, Tekaia *et al.* discovered that over half of the current ORFs result from gene duplication or domain shuffling events (Tekaia *et al.*, 1999). One sixth of the total number of ORFs were of unknown function, and so could be used as novel targets for further research. Considering that resistant strains are on the increase, any new and effective drugs developed against these targets will be very welcome. Similarly, an investigation into functional classification of unknown ORFs, using bioinformatics (King *et al.*, 2000), developed a method of estimating protein function from both the *M. tuberculosis* and *E. coli* genome sequences, with 60–80% accuracy. However, the true biochemical functions of these proteins will need to be elucidated experimentally. Indeed, this paper remarks, "The actual function of an ORF can only be determined by 'wet' experiment."

One problem of relying on genomic information and pair-wise sequence homology alone is that many genes of unknown function are evident i.e. approx. 40% in the case of *Treponema pallidum* (Wassenaar and Gaastra, 2001). Recently, bacteriology, proteomics, and bioinformatics were combined to produce complete protein analyses of two pathogens, *Helicobacter pylori*, the causative agent of gastritis, ulcers and stomach carcinoma, and *M. tuberculosi*s (Tekaia *et al.*, 1999; Jungblut *et al.*, 2000). Both structural and functional aspects of the species' proteins were predicted using *in silico* techniques, leading to an improved understanding of each bacterium's pathogenicity and evolution. These predictions could then be investigated experimentally to confirm the theoretical analysis. "Conventional" screening involves exhaustive laboratory testing of novel compounds against bacterial cells in *in vitro* cultures, and may take several months to produce a panel of suitable compounds and/or targets for *in vivo* research. By employing computational techniques to select a panel of targets *before* carrying out *in vitro* work, this process should be shortened considerably

Proteomics, Microarrays and Expression Profiling

With post-genomic research comes the rapid growth of proteomics, the study of a particular species' complete protein repertoire encoded for by its genome. Combining 2-D gel electrophoretic separation of complex protein mixtures with protein analytical methods like peptide mass spectrometry allows identification of expressed proteins at a set point in time under certain conditions (Grandi, 2001). Bioinformatics can play an important support role for proteomics, permitting recognition of known proteins through the use of peptide mass databases like MOWSE (MOlecular Weight SEarch) (Pappin *et al.*, 1993), and thus highlighting uncharacterised proteins for further research.

Chakravati *et al.* focused on the prediction of surface localised proteins (Chakravati *et al.*, 2001) while carrying out database mining of pathogenic genomes. Proteins from both *H. influenzae* and *H. pylori* were analysed using proteomics and suitable vaccine candidates were identified. A lipoprotein (P6) from *H. influenzae* and several *H. pylori* virulence factors were then selected for further immunogenicity studies. This method of screening provides good goals for potential vaccine studies, as surface localised or secreted proteins are the most usual targets for the humoral immune system. Gomez *et al.* devised a different procedure (Gomez *et al.*, 2000) using bioinformatics and *phoA*' fusion technology. Coding regions of the *M. tuberculosis* genome were scanned *in silico* for predicted secreted proteins, and these moieties then analysed *in vitro*. This fast track approach allowed

identification of a large number of targets with 90% accuracy.

Gene expression profiling using DNA microarrays allows researchers to monitor changes to an organism's mRNA expression either in response to a set of specified conditions, or compared to another genome (Schena *et al.*, 1995). Combining genomics, biotechnology and computational biology, this method is ideal for analysing bacterial species, as virtually all of a microbe's ORFs can be fitted onto one microarray chip. An important role for bioinformatics evolved when the large microarray data output required investigation. By using algorithms that can group together similar elements i.e. genes displaying a homologous function, and reject redundant or dissimilar entries in the data, the output can be reduced to a manageable size (Raychaudhuri *et al.*, 2001).

Two studies have utilised this technology to study *M. tuberculosis* (Behr *et al.*, 1999; Wilson *et al.*, 1999), and their findings are reviewed in (Barry and Schroeder, 2000). The first observed comparisons between the genome of *M. tuberculosis,* and that of the *M. bovis* BCG strain, used in current TB vaccination programmes. Variations among the two showed where novel vaccine targets could be directed to enhance host secondary immune responses towards the pathogens (Behr *et al.*, 1999). In contrast, Wilson *et al.*'s work (Wilson *et al.*, 1999) focused on differential gene expression. In response to treatment by the anti-tubercular drug isoniazid, the transcription of several important genes was altered. Future antimicrobials can thus be tested this way to measure their effectiveness. However, one drawback is that only processes subject to transcriptional regulation are identified and poorly expressed genes may be hard to detect; conventional clone screening is still required.

Koch's Postulates Redefined

Robert Koch first defined pathogenicity in 1890 (Domingue Sr. and Woody, 1997; Falkow, 1988), when he formulated his postulates about disease-causing micro-organisms. These stated that a specific disease-causing organism should be present in all cases of the disease, and after isolation, should grow in both pure culture and a healthy susceptible animal host, exhibiting the same disease symptoms as seen in the original case. Finally, the organism should be re-isolated in pure culture. It is clear that these postulates have become obsolete. For example, viral pathogens require living cells in which to reproduce and cannot grow on artificial media alone, thereby contradicting the most fundamental postulate. A similar case is found with the problematic *M. tuberculosis*; the bacterium has a doubling time of around 15 hours in batch culture (James *et al.*, 2000) and is not always culturable. Neither are the postulates relevant to pathogens that specifically infect humans and for which there is no suitable animal model, such as *Neisseria meningitidis*, the causative agent of meningococcal pneumonia.

By far the largest group of micro-organisms to which Koch's postulates do not apply are opportunistic pathogens. These fall into several categories, the first being a host's normal flora that has acquired virulence genes, or factors, through lateral transfer (see Figure 1). Commensal bacteria like intestinal *E. coli* are most likely to be affected by this. *E. coli* has the potential to produce severe diarrhoea in infants if it acquires certain virulence factors from other pathogenic forms, like Enteropathogenic *E. coli* (EPEC) and Enterohaemorrhagenic *E. coli* (EHEC). When infected with an opportunistic pathogen, some obligate intracellular bacterial species may not always be recoverable from the host. Hence, another postulate is not fulfilled.

Patients with severely compromised immune systems such as those suffering from HIV/AIDS or undergoing major surgery i.e. for third degree burns are far more susceptible to infection by pathogens that normally would be considered benign. *Serratia marcescens*, a free-living soil bacterium, has recently been shown to cause necrotizing fasciitis in patients receiving immunosuppressive therapy in hospital (Dorsey *et al.*, 2000). Another hugely successful opportunistic pathogen is *P. aeruginosa,* a ubiquitous bacterium that can adapt to its environment, be it in the soil or inside a host. When the genome of this organism was sequenced (Stover *et al.*, 2000), it revealed that *P. aeruginosa* contained nearly as many ORFs as the primitive eukaryote *Saccharomyces cerevisiae*, and showed remarkable intrinsic drug resistance.

Lastly, the same disease symptoms may be caused by different organisms, so contradicting the postulate that "the specific organism should be present in all disease cases". Bioinformatic analyses have shown that Enteroinvasive *E. coli* (EIEC) has 90% genomic homology to *Shigella* spp. and can produce a severe reaction in hosts very similar to dysentery, with identical symptoms (Rolland *et al.*, 1998). Another complicating factor is the *host's* involvement in pathogenesis of the bacterial infection. It has been suggested that human pathogens can be classified according to host damage as a function of the immune response (Casadevall and Pirofski, 1999). Pathogenicity is usually designated according to the actions of the micro-organism.

Lateral Transfer, Pathogenicity Islands and the Spread of Virulence

Some bacterial species are able to acquire new virulence genes through "lateral" or horizontal transfer, a natural process that promotes the spread of resistant and/or pathogenic cells (Ochman *et al.*, 2000). Resistance is able to spread quickly through the resident bacterial population in a competitive environment. When antibiotics kill most susceptible cells, the residual resistant bacteria quickly colonise the empty niche, and pathogenic species can obtain a significant amount of their genetic diversity in this way (see Figure 1). In line with the idea of lateral transfer between pathogens, Nguyen *et al* suggested, through computer-based phylogenetic analyses, that certain

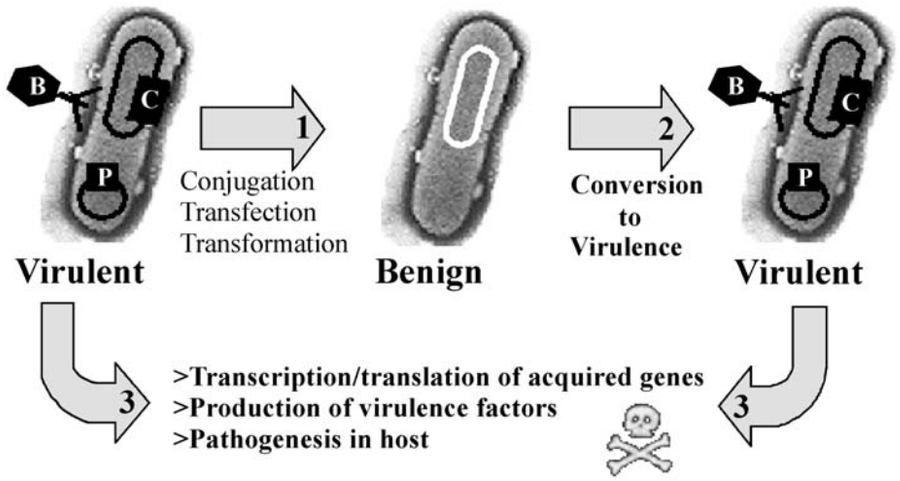


Figure 1. Lateral transfer of virulence/resistance genes from virulent to benign strain. Nucleic acid sequences passed between a resistant or virulent microbe and a benign/avirulent counterpart can take several forms. 1) Donation of virulence genes to avirulent strain via a variety of methods. Conjugation involves the physical transfer of specific DNA regions between donor and recipient, in the form of self-transmissible plasmids (i.e. LT1 enterotoxin from *E. coli*), transposons (i.e. *E. coli* haemolysin), or plasmids carrying chromosomal nucleic acids (i.e. *P. aeruginosa* exotoxin A). Transformation involves the uptake of naked DNA from the environment by the bacterium and allows prospective virulence genes to pass into very distinctly related organisms (i.e. *N. gonorrhoea* is particularly competent throughout its lifecycle). Transfection by bacteriophage can introduce large random fragments of foreign DNA into host chromosomes (i.e. *Clostridium botulinum* botulinum toxin, *C. diptheriae* diptheria toxin). 2) The benign strain is then converted to virulence by the presence of the acquired genes. 3) Transcription/translation of the virulence/resistance genes then leads to production of virulence proteins, and eventual pathogenesis in the host. Key: B – bacteriophage, C – Bacterial chromosome, P – Plasmid.

virulence clusters were always horizontally transmitted as units, without the formation of hybrids (Nguyen *et al.*, 2000).

A good example of lateral transfer can be found in Enteroinvasive *E. coli* (EIEC): it shows a 90% similarity to *S. dysenteriae* and other *Shigella* spp. in terms of *in silico* genome comparisons. As several enterotoxins from both species also share this high sequence similarity, some have speculated that they may have evolved from a distinct common ancestor. For example, the curli loci found in *Shigella* show homology to the same genes in EIEC, and are structurally different to the rest of the *Shigella* genome (Sakellaris *et al.*, 2000). Maurelli's group (Maurelli *et al.*, 1998) in contrast, suggested that in addition to virulence factor transfer, "a complementary but inverse pathway may exist to enable commensal bacteria to evolve towards a pathogenic lifestyle: the formation of 'black holes' i.e. deletions of genes that are detrimental to a pathogenic lifestyle". This is certainly an interesting hypothesis. Even horizontal transfer from the Archaea to Eubacteria has been proven; Olendzenski *et al.* used both molecular and computational biology to investigate the movement of archaeal A/V ATPases into members of the Deinococcacaea (Olendzenski *et al.*, 2000).

As most virulent bacteria rely on the presence of virulence genes/factors or mechanisms to provide them with the means to infect a human host, these moieties must be considered essential for pathogenicity. Such genes can be organised into distinct clusters in the pathogenic genome, and these "pathogenicity islands" (PAIs) provide the means for large-scale lateral transfer of virulence when passed from a donor strain to a recipient. They are also found only in pathogenic strains. Several features that define PAIs include: a different G + C content to the rest of the host genome, flanking by insertion loci i.e. tRNAS or IS elements, and mobility between distinctly related bacterial species. Often a PAI will encode a specific set of virulence factors that convey a specific virulent function. Table 1 gives several examples of this phenomenon. Methods for locating PAIs within a genome have been developed, for example, by comparing codon frequencies of genes to several gene classes. Highly expressed genes are most likely to belong to the host, while "alien" genes, clustered into operons, are predicted to be PAIs (Mrazek *et al.*, 2001). Analysing the global usage of dinucleotide signatures in prokaryotic genomes may also indicate novel pathogenic gene clusters (Karlin, 1998).

Virulence Factors

Virulence factors and mechanisms can be roughly divided into several classes (Finlay and Falkow, 1997), but there are some exceptions to this classification. A description of the main virulence factor sets can be found in Table 2, with examples of each. Several groups have explored the quandary of what *exactly* constitutes a virulence factor. For example, Wassenaar and Gaastra suggest that proteins thought of as necessary for pathogenicity fall into three categories: "true" virulence genes, those that are associated with virulence such as expression regulators of "true" factors, and lastly, virulence "life-style" genes that are required by the bacterium to enable colonisation of the host. In essence, a virulence factor is any moiety produced by a pathogen that is essential for causing

Table 1. Examples of chromosomally located pathogenicity islands from some of the well-known disease causing bacteria. Key; TTSS – Type three secretion system, TFSS – Type four secretion system.

Pathogen	PAI Name	Relevent genes/factors	Function	Size of PAI
Uropathogenic *E. coli*	PAI II	α-hemolysin, P-fimbriae genes	Toxic activity, receptor attachment	190 kb
Enteropathogenic *E. coli*	LEE	TTSS (esc) genes, intimin and translated intimin receptor (tir) genes	Bacterial entry into host cells, toxic and enterocyte effacement activity	35 kb
Vibrio chlorae	VPI	Toxin-coregulated-pilus (tcp) genes, transposase, integrase genes.	Colonization, virulence factor expression-regulation	39.5 kb
H. pylori	Cag PAI	CagA exotoxin, TFSS (cag) structural genes	Bacterial entry into host cells, cytoskeletal rearrangement	40 kb
S. typhimurium	SPI-I, SPI-II	TTSS (spa/inv and spi/ssa) genes	Bacterial entry into host cells, adherence and cytoskeletal rearrangement	40 kb (both)
Y. pestis	HPI	Yersiniabactin (ybt) genes	Siderophore production, iron-uptake	102 kb

disease in a host. Bacterial pathogens have devised ingenious methods to invade the human host successfully, but many seemingly diverse pathogens can share common virulence traits. This will be of great use when designing novel compounds to combat diseases. As our knowledge of microbial pathogenesis grows with each new genome sequenced, so too should the number of targets available for therapeutic research.

Bacterial metabolism components are also becoming the focus of research into novel targets. Primarily, this has concentrated on metabolic enzymes that catalyse important intracellular reactions, such as amino acid biosynthesis and DNA replication. Computational predictions using these proteins offer a good starting point for drug target discovery. Utilising tertiary structure prediction tools, O'Donoghue *et al.* were able to develop a preliminary molecular model of the *E. coli* imidazole glycerol phosphate synthase (IGPS) holoenzyme (O'Donoghue *et al.*, 2001) using bioinformatics. More recently, interest has shifted towards energy metabolism macromolecules, like the sodium ion (Na^+) cycle found in many microbes (Hase *et al.*, 2001). The bacterial outer membrane protein NqrA, involved in the Na^+ cycle, functions as a channel for sodium ions and elicits a highly immunogenic response in the host. NqrA has been suggested as a suitable broad spectrum vaccine candidate, as phylogenetic studies have revealed the moiety is conserved across a wide range of pathogens (Cruz *et al.*, 1996). With the advent of functional genomic databases such as the EcoCyc web server (Ouzounis and Karp, 2000), metabolic networks are becoming better elucidated, and key proteins involved in important steps can be targeted.

Many novel virulence factors have been discovered through the use of homology searches like BLAST and FASTA with bacterial genomic sequence data. However, around a third of all ORFS in each genome published so far have unknown function (Weinstock, 2000), more recently shown by the sequencing of the *E. coli* O157 genome. Further prediction using computational methods can clarify the function of these putative proteins, and where they lie in the bacterial cell. As prokaryotic organisms do not possess distinct subcellular compartments there are, at the maximum, five areas to which any ORF of unknown function could be targeted. If the bacterium is Gram-positive, there are three – secretion to the extracellular environment, anchorage in the cell membrane, or remaining in the cytoplasm. Because of the double-membrane cell envelope in Gram-negative organisms, an ORF could also be targeted to the periplasm or the second outer membrane in addition to those areas already mentioned (Nakai, 2000). Proteins that are expressed in virulent strains of bacteria and predicted to have either a) a secretion signal and be secreted to the exterior of the cell, or b) alpha helical transmembrane (TM) segments and/or beta barrel TM regions and anchored in the outer cell membrane, are often the focus of antimicrobial therapies.

Table 2. Virulence factor classes and some examples of each. Key; TTSS – Type three secretion system.

Class	Main Function(s)	Model bacterium	Example
Adherens	Colonisation of mucosal sites	*E. coli*	Intimin and tir
Invasins	Extracellular breakdown of local host defenses, effective internalisation of facultative intracellular pathogens	*Clostridium* spp., *Yersinia* spp.	Lecithinase, TTSS
Endotoxins	Interaction with host cell surfaces, immunomodulatory effects.	Gram-negative species	Lipopolysaccharide
Exotoxins	Specialised to each pathogen, many different activites	Many examples i.e. *Corynebacterium diptheriae*	Diphtheria toxin
Surface components	Protection from host immune response	*Streptococcus* spp. – antigenic mimicry	Protein M
Siderophores	Iron scavenging and uptake from host proteins	*E. coli*	Enterobactin

There are a number of online servers available to predict the tertiary structure of proteins, and most rely on multiple scoring methods to identify putative TM regions, secretion signals, the orientation of integral membrane proteins etc. For example, the PredictProtein server based at EBI in Heidelberg (Rost, 1996) generates a calculated tertiary structure for a query sequence based upon the predictions of twelve other algorithms. Conversely, kPROT is more specific, and more suited to the prediction of membrane proteins only (Pilpel *et al.*, 1999). However, one problem with Gram-negative prokaryotic membrane proteins is that a number of them contain beta barrel and not alpha helical TM segments (Koebnik *et al.*, 2000), and the current prediction methods available are either too inaccurate or not successful enough yet to deal with this. Another is that most of the prediction programs are ''trained'' using eukaryotic protein data sets. Further developments within the existing software should alleviate this problem.

One collection of pathogenic genes that have attracted great interest are the type III secretion systems (TTSS), encoded in a number of Gram-negative animal enteropathogenic PAIs (Cornelis and Van Gijsegem, 2000). They contain both structural and effector proteins; the former display a remarkable homology amongst different species, yet all have heterologous function. Although four pathogenic bacterial secretion systems have been described so far (China and Goffaux, 1999) the TTSS have attracted attention because, upon close contact with the target cell, the structural components assemble into a ''needle'' in the prokaryotic envelope, and deliver the effector proteins directly into the host cytoplasm. However, predicting which proteins are secreted by this system is difficult; all TTSS effectors discovered so far show no secretion signal at either their amino or carboxyl terminals. The same situation is found with the type IV system. In the type II or general secretory system, a conventional *sec* signal is found at the N terminus of exported proteins, and therefore these can be computationally predicted using, for example, the SignalP algorithm (Nielsen *et al.*, 1997). Type I secreted virulence determinants carry a signal present in their C terminus, but this is not cleaved upon export from the bacterial cell. Figure 2 shows the possible cellular locations of expressed prokaryotic virulence factors, and the methods by which these proteins can be predicted.

Protein Profiling

A recognised and powerful method of classifying new protein families is to use conserved regions within multiple alignments of related proteins. Each homologous

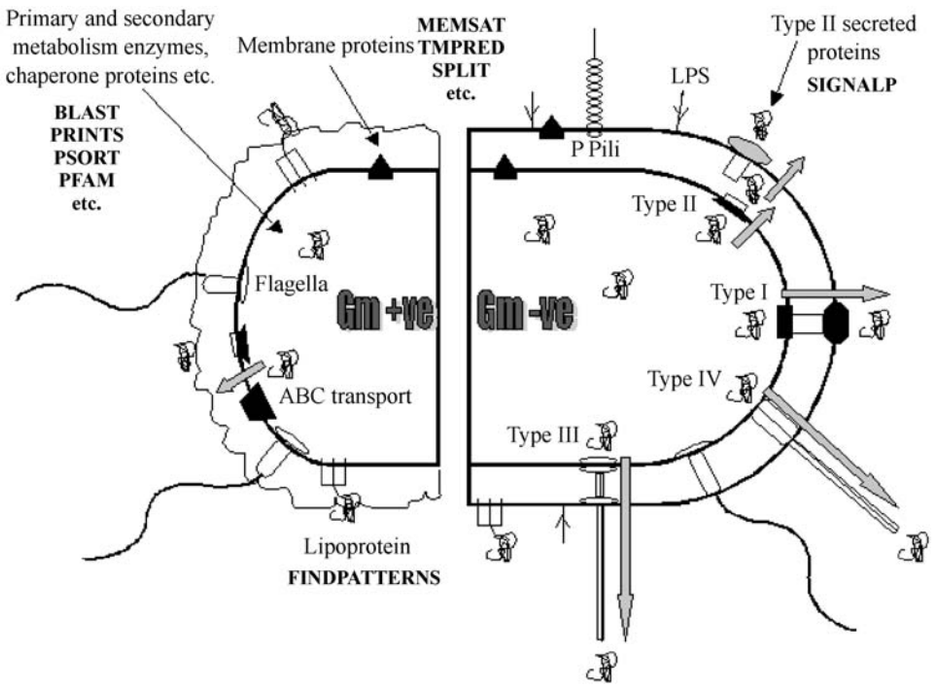


Figure 2. Prediction and secretion of virulence factors across the Gram-negative and -positive envelope. The subcellular location of expressed proteins can be calculated using a variety of methods as shown. In Gram-negative organisms, type II secreted toxins could be found using SIGNALP, while alpha helical TM regions in membrane proteins can be predicted using servers such as MEMSAT, kPROT, TMPRED and SPLIT. With all putative ORFS, homology searches with BLAST and/or protein profiling searches like PRINTS and PSORT can reveal function and structure as well as location. In addition, unusual patterns in coding sequences, for example those within lipoproteins, can be found with FINDPATTERNS. Also shown are the two other Gram-negative secretion systems, and various cell appendages i.e. flagella and P pili. Arrows denote the action and direction of secretion. Key: LPS – lipopolysaccharide, TYPE I – Type I secretion system, TYPE II – Type II secretion system, TYPE III – Type III secretion system, TYPE IV – Type IV secretion system, Gm + ve – Gram-positive, Gm – ve – Gram-negative.

region is a ''motif'', and sets of motifs provide a signature or fingerprint for unique identification. These motifs usually denote a common structure and/or function between individual family members. Currently, the most commonly used online databases include PROSITE, BLOCKS, Pfam, ProDom, and PRINTS (Attwood, 2000) and a compendium of these, InterPro, has also been released (Apweiler *et al.*, 2001). All are valuable tools in the characterisation and categorisation of novel protein families, although there are advantages and disadvantages in the use of each. For example, the annotated databases (PRINTS, Pfam and PROSITE) provide a more concise information resource for each entry than, for example, automatically generated databases like BLOCKS.

Our group noted that relatively few virulence factor protein families have been studied when compared with other entries in these databases. Therefore, in collaboration with the Bioinformatics group at the University of Manchester, UK (Attwood *et al.*, 2000), a number of pathogenic signatures have been researched and added to the PRINTS database. These have included the different protein components of the TTSS found in Gram-negative pathogens. Both the *Salmonella typhimurium* SPI-I and the *S. flexneri* mxi needle structures have been resolved through transmission electron microscopy (Kubori *et al.*, 1998; Tamano *et al.*, 2000), and more recently, the *E. coli* needle structure was also discovered (Sekiya *et al.*, 2001).

Although most TTSS protein structural components show sequence homology, the majority of the effectors do not. This is due to the fact that they affect the host cell in various ways in order to facilitate bacterial spread. For example, the six secreted Yop exotoxins of *Yersinia pestis* (Bliska, 2000) drastically affect the actin cytoskeleton, interfering with integrin-mediated phagocytosis and allowing uptake of the facultative intracellular bacterium. The Ipa proteins from *S. flexneri* contribute to the killing of neutrophils by necrosis, thus allowing the pathogen to enter host cells via disruption of the epithelial barrier (Francois *et al.*, 2000). Some characteristics that these effectors share include: a lack of the traditional *sec* signal as seen in some other secreted exotoxins, extensive chaperoning by accessory proteins (also encoded in TTSS pathogenicity islands) while in the bacterial cytoplasm, and a possible translocation signal within the mRNA encoding each toxin. While a consensus secretion signal does not exist for *all* secreted effectors, one group recently suggested that the amino termini of several *S. typhimurium* exotoxins could contain a specific species-related signal (Miao and Miller, 2000).

Therefore, working with the TTSS protein components and various other virulence factors, a working method to generate PRINTS entries was evolved. From an initial alignment of related protein sequences, motifs were manually selected from conserved regions. Once the motif set was analysed to convergence, any non-specific matches were removed and

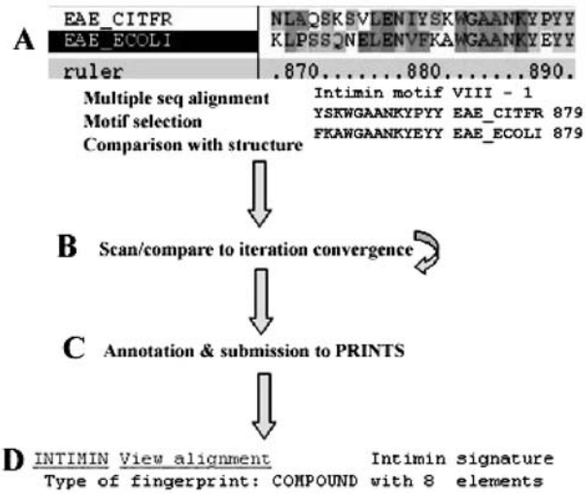


Figure 3. A diagram of how the *Escherichia coli* intimin (Attachment and Effacement protein) protein family was characterised for PRINTS. (A) The process began with a BLAST search of *E. coli* intimin to ascertain family members, before representatives were aligned. Motifs were then picked according to conserved areas, and also any that corresponded to important elements in the 3-D structure. For example, Motif 8 spanned the C-type lectin domain at the C-terminus, and this is the motif given. (B) The set of motifs was scanned/compared until a convergence of true family members was reached (10 sequences), and then, (C) various annotations were added, such as literature references, database links, etc. (D) Once submitted and placed online, the finished entry appears like this.

an annotation written for the specific protein family. As homology was concentrated on regions rather than the whole sequence, the background ''noise'' associated with conventional pair-wise matching was reduced. One advantage gained by the PRINTS methodology is that each family is thoroughly researched and then manually annotated before being placed in the database. Figure 3 shows how the PRINTS entry INTIMIN was generated.

The full list of PRINTS entries generated so far can be found at http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/.

Limitations of Functional Genomics and Bioinformatics

Although many new scientific discoveries have been achieved with the aid of bioinformatics, there are some limitations with this science, as there is with any other discipline. It has to be remembered that the field is a *theoretical* one, and was designed as a complement to the more conventional lab sciences like microbiology and molecular biology. Ultimately, any bioinformatic work is reliant on the validity of the biological sequence dataset that the *in silico* analysis originated from.

Certain molecular characteristics cannot, as yet, be accurately predicted using available *in silico* tools from raw sequence data. For example, *ab initio* gene structure, protein function, novel ''true'' bacterial virulence factors, and the tertiary structure of polypeptides still have to be fully resolved using *in vitro/vivo*

technology (Yaspo *et al.,* 2001). Current bioinformatical molecular tools may not accurately predict other aspects of sequence data. For example, bacterial transmembrane beta-barrel proteins, as already mentioned, are hard to predict using existing applications. Also, the external conditions under which certain pathogenic proteins may be expressed cannot be predicted by any current *in silico* method. Clearly, more sophisticated tools for annotation need to be developed for better analysis of raw sequence data, as well as the improvement of existing technology.

Bridging the Gap – Concluding Remarks

Bacterial pathogenicity is fast becoming an area of high priority, as the level of resistance in disease-causing micro-organisms rapidly increases and the prevalence of global HIV rises. A new definition of pathogenicity could be; "Pathogens can be distinguished from their avirulent counterparts by the presence of specific genes (virulence factors) and/or clusters of such genes (pathogenicity islands) within their genome, and must cause a disruption in the host immune response upon infection". By examination of a pathogen's repertoire of genes, those that favour pathogenesis are identified. These virulence factors are often organised into specific gene clusters (pathogenicity islands) on the bacterial chromosome, and play an important role in disease. The Type three secretion system is unique amongst some Gram negative pathogens like *S. typhimurium* and *Y. pestis*, and may provide a novel way of introducing live vaccines to a human host. Teaming bioinformatics with genome information and biotechnology, it could be possible to combat enteritis by introducing antigens to a specific mucosal area via this system.

Examples of how the two sciences can be combined include the use of genome annotation and sequence analysis. Protein profiling is also a good method of comparing proteins that may be very weakly homologous using other techniques. Arguably the key advantage of computer-based screening techniques is the speed at which the selection of novel targets can be made. However, further development of these technologies, both *in silico* and *in vitro/vivo*, is necessary to ensure the maximum amount of information is gleaned from new genomic sequences. The onus, therefore, is on scientists to ultimately move their research from the computer screen to the laboratory bench. Making a reality of the predictions on how a protein may act a certain way *in vivo*, or what sort of immune response will be elicited from a virulence factor carefully selected from database mining and gene expression profiling, ultimately falls to the more conventionally trained biologist. Of course, bringing two very distinct disciplines together is no easy task, but the marriage of "dry" and "wet" molecular biology and bacteriology promises to be a fruitful one.

Acknowledgements

We thank Professor P. C. L. Beverly for critical comments on this paper.

References

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., and Servant, F. 2001. The InterPro database, an integrated documentation resource for protein families, domains, and functional sites. Nucleic Acids Res. 29: 37–40.

Attwood, T.K. 2000. The quest to deduce protein function from sequence: the role of pattern databases. Int. J. Biochem. Cell. Biol. 32: 139–155.

Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., and Wright, W. 2000. PRINTS-S: the database formally known as PRINTS. Nucleic Acids Res. 28: 225–227.

Barry, C.E., and Schroeder, B.G. 2000. DNA microarrays: translational tools for understanding the biology of *Mycobacterium tuberculosis*. Trends Microbiol. 8: 209–210.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284: 1520–1523.

Bliska, J.B. 2000. Yop effectors of *Yersinia* spp. and actin rearrangements. Trends Microbiol. 8: 205–208.

Brosch, R., Gordon, S.V., Buchreiser, C., Pym, A.S., Garnier, T., and Cole, S.T. 2000. Comparative genomics uncovers large tandem chromosomal duplications in *Mycobacterium bovis* BCG Pasteur. Yeast 17: 111–123.

Brown, S., Banter, C., Young, H.K., and Amyes, S.G.B. 1998. Limitation of *Acinobacter baumannii* treatment by plasmid-mediated carbapenemase ARI-2. Lancet 351: 186–187.

Burge, C.B., and Karlin, S. 1998. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8: 346–354.

Casadevall, A., and Pirofski, L.A. 1999. Host-pathogen interactions: Redefining the basic concepts of virulence and pathogenicity. Infect. Immun. 67: 3703–3713.

Chakravarti, D.N., Fiske, M.J., Fletcher, L.D., and Zagursky, R.J. 2001. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. Vaccine 19: 601–612.

China, B., and Goffaux, F. 1999. Secretion of virulence factors by *Escherichia coli*. Vet. Res. 30: 181–202.

Cornelis, G.R., and Van Gijsegem, F. 2000. Assembly and function of type III secretory systems. Annu. Rev. Microbiol. 54: 735–774.

Cruz, W.T., Nedialkov, Y.A., Thacker, B.J., and Mulks, M.H. 1996. Molecular characterization of a common 48-kilodalton outer membrane protein of *Actinobacillus pleuropneumoniae*. Infect. Immun. 64: 83–90.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27: 4636–4641.

Domingue, G.J., Sr., and Woody. H.B. 1997. Bacterial persistence and expression of disease. Clin. Microbiol. Rev. 10: 320–344.

Dorsey, G., Borneo, H.T., Sun, S.J., Wells, J., Steele, L., Howland, K., Perdreau-Remington, F., and Bangsberg, D.R. 2000. A heterogeneous outbreak of *Enterobacter cloacae* and *Serratia marcescens* in a surgical intensive care unit. Infect. Control Hosp. Epidemiol. 21: 465–469.

Falkow, S. 1988. Molecular Koch's postulates applied to microbial pathogenicity. Rev. Infect. Dis. 10: S274–276.

Finlay, B.B., and Falkow, S. 1997. Common themes in microbial pathogenicity revisited. Microb. Mol. Biol. Rev. 16: 136–169.

Fleischmann, R.D., *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512.

Francois, M., Le Cabec, V., Dupont, M.A., Sansonetti, P.J., and Maridonneau-Parini, I. 2000. Induction of necrosis in human neutrophils by *Shigella flexneri* requires type III secretion, IpaB and IpaC invasins, and actin polymerization. Infect Immun 68: 1289–1296.

Fraser, C.M., Eisen, J.A., and Salzberg, S.L. 2000. Microbial genome sequencing. Nature 406: 799–803.

Godfrey-Faussett, P., Sonnenberg, P., Shearer, S.C., Bruce, M.C., Mee, C., Morris, L., and Murray, J. 2000. Tuberculosis control and molecular epidemiology in a South African gold-mining community. Lancet 356: 1066–1071.

Gomez, M., Johnson, S., and Gennaro, M.L. 2000. Identification of secreted proteins of *Mycobacterium tuberculosis* by a bioinformatic approach. Infect. Immun. 68: 2323–2327.

Grandi, G. 2001. Antibacterial vaccine design using genomics and proteomics. Trends Biotechnol. 19: 181–188.

Hase, C.C., Fedorova, N.D., Galperin, M.Y., and Dibrov, P.A. 2001. Sodium ion cycle in bacterial pathogens: evidence from cross-genome comparisons. Microb. Mol. Biol. Rev. 65: 353–370.

James, B.W., Williams, A., and Marsh, P.D. 2000. The physiology and pathogenicity of *Mycobacterium tuberculosis* grown under controlled conditions in a defined medium. J. Appl. Microbiol. 88: 669–677.

Jungblut, P.R., Bumann, D., Haas, G., Zimny-Arndt, U., Holland, P., Lamer, S., Siejak, F., Aebischer, A., and Meyer, T.F. 2000. Comparative proteome analysis of *Helicobacter pylori*. Mol. Microbiol. 36: 710–725.

Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1: 598–610.

King, R.D., Karwath, A., Clare, A., and Dehaspe, L. 2000. Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis* and *Escherichia coli* genomes using data mining. Yeast 17: 283–293.

Koebnik, R., Locher, K.P., and Van Gelder, P. 2000. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol. Microbiol. 37: 239–253.

Kubori, T., Matshushima, Y., Nakamura, D., Uralil, J., Lara-Tejero, M., Sukhan, A., Galan, J.E., and Aizawa, S.I. 1998. Supramolecular structure of the *Salmonella typhimurium* type III protein secretion system. Science 280: 602–605.

Maurelli, A.T., Fernandez, R.E., Bloch, C.A., Rode, C.K., and Fasano, K. 1998. "Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. Proc. Nat. Acad. Sci. USA 95: 3943–3948.

Miao, E.A., and Miller, S.I. 2000. A conserved amino acid sequence directing intracellular type III secretion by *Salmonella typhimurium*. Proc. Nat. Acad. Sci. USA 97: 7539–7544.

Mrazek, J., Bhaya, D., Grossman, A.R., and Karlin, S. 2001. Highly expressed and alien genes of the *Synechocystis* genome. Nucleic Acids Res. 29: 1590–1601.

Murray-Rust, P. 1994. Bioinformatics and drug discovery. Curr. Opin. Biotechnol. 5: 648–653.

Nakai, K. 2000. Protein sorting signals and prediction of subcellular localization. Adv. Prot. Chem. 54: 277–344.

Nguyen, L., Paulsen, I.T., Tchieu, J., Hueck, C.J., and Saier, M.H. Jr., 2000. Phylogenetic analyses of the constituents of type III protein secretion systems. J. Mol. Microbiol. Biotechnol. 2: 125–144.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10: 1–6.

Ochman, H, Lawrence, J.G., and Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

O'Donoghue, P., Amaro, R.E., and Luthey-Schulten, Z. 2001. On the structure of hisH: protein structure prediction in the context of structural and functional genomics. J. Struct. Biol. 134: 257–268.

Olendzenski, L., Liu, L., Zhaxybayeva, O., Murphey, R., Shin, D.G., and Gogarten, J. P. 2000. Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. J. Mol. Evol. 51: 587–599.

Ouzounis, C.A., and Karp, P.D. 2000. Global properties of the metabolic map of *Escherichia coli*. Genome Res. 10: 568–576.

Pappin, D.J.C., Hojrup, P., and Bleasby, A.J. 1993. Rapid identification of proteins by peptide-mass fingerprinting. Curr. Biol. 3: 327–332.

Paulsen, I.T., Nguyen, L., Sliwinski, M.K., Rabus, R., and Saier, M.H. Jr., 2000. Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J. Mol. Biol. 301: 75–100.

Perna, N.T., *et al.* 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157: H7. Nature 409: 529–533.

Perriere, G. Duret, L., and Gouy, M. 2000. HOBACGEN: Database system for comparative genomics in bacteria. Genome Res. 10: 379–385.

Pilpel, Y., Ben-Tal, N., and Lancet, D. 1999. kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. J. Mol. Biol. 294: 921–935.

Ramakrishna, R., and Srinivasan, R. 1999. Gene identification in bacterial and organellar genomes using GeneScan. Comput. Chem. 23: 165–174.

Raychaudhuri, S., Sutphin, P.D., Chang, J.T., and Altman, R.B. 2001. Basic microarray analysis: grouping and feature reduction. Trends Biotechnol. 19: 189–193.

Rolland, K., Lambert-Zechovsky, N., Picard, B., and Denamur, E. 1998. *Shigella* and enteroinvasive *Escherichia coli* strains are derived from distinct ancestral strains of *E. coli*. Microbiology 144: 2667–2672.

Rosamond, J., and Allsop, A. 2000. Harnessing the power of the genome in the search for new antibiotics. Science 287: 1973–1976.

Rost, B. 1996. PHD: predicting 1D protein structure by profile based neural networks. Meth. Enzym. 266: 525–539.

Rudd, K.E. 2000. EcoGene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res. 28: 60–64.

Sakellaris, H., Hannink, N.K., Rajakumar, K., Bulach, D., Hunt, M., Sasakawa, C., and Adler, B. 2000. Curli loci of *Shigella* spp. Infect. Immun. 68: 3780–3783.

Sanger, F., Nicklen, S., and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Nat. Acad. Sci. USA 74: 5463–5467.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–470.

Sekiya, K., Ohisi, M., Ogino, T., Tamano, K., Sasakawa, C., and Abe, A. 2001. Supermolecular structure of the enteropathogenic *Escherichia coli* type III secretion system and its direct interaction with the EspA-sheath-like structure. Proc. Nat. Acad. Sci. U.S.A. 98: 11638–11643.

Shmatkov, A.M., Melikyan, A.A., Chernousko, F.L., and Borodovsky, M. 1999. Finding prokaryotic genes by the "frame-by-frame" algorithm: targeting gene starts and overlapping genes. Bioinformatics 15: 874–886.

Stover, C.K., *et al.* 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406: 959–964.

Tamano, K., Aizawa, S., Katayama, E., Nonaka, T., Imajoh-Ohmi, S., Kuwae, A., Nagai, S., and Sasakawa, C. 2000. Supramolecular structure of the *Shigella* type III secretion machinery: the needle part is changeable in length and essential for delivery of effectors. EMBO J. 19: 3876–3887.

Tekaia, F., Gordon, S.V., Garnier, T., Brosch, R., Barrell, B.G., and Cole, S.T. 1999. Analysis of the proteome of *Mycobacterium tuberculosis in silico*. Tubercle Lung Dis. 79: 329–342.

Van den Bogaard, A.E., and Stobberingh, E.E. 2000. Epidemiology of resistance to antibiotics. Links between animals and humans. Int. J. Antimicrob. Agents 14: 327–335.

Wassenaar, T.M., and Gaastra, W. 2001. Bacterial virulence: can we draw the line? FEMS Microbiol. Lett. 201: 1–7.

Weinstock, G.M. 2000. Genomics and bacterial pathogenesis. Emer. Infect. Dis. 6: 496–504.

Wilson, M., DeRisi, J., Kristensen, H.H., Imboden, P., Rane, S., Brown, P.O., and Schoolnik, G.K. 1999. Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. Proc. Nat. Acad. Sci. USA 96: 12833–12838.

Yaspo, M.L. 2001. Taking a functional genomics approach in molecular medicine. Trends Mol. Med. 7: 494–501.