Reverse vaccinology

Marirosa Mora, Daniele Veggi, Laura Santini, Mariagrazia Pizza and Rino Rappuoli

Whole-genome sequencing of bacteria and advances in bioinformatics have revolutionized the vaccinology field, leading to the identification of potential vaccine candidates without the need for cultivating the pathogen. This approach, termed 'reverse vaccinology', reduces the time and cost required for the identification of candidate vaccines and provides new solutions for those diseases for which conventional approaches have failed. The first example of the potential of reverse vaccinology has been the identification of novel antigens of meningococcus B as potential candidates for a novel and effective vaccine. The same approach has been successfully applied to other important human pathogens, demonstrating the feasibility to develop vaccines against any infectious disease. This review focuses on some recent advances in the identification of vaccine candidates by mining the genomic sequences of pathogenic bacteria.

Marirosa Mora
Daniele Veggi
Laura Santini
Mariagrazia Pizza
Rino Rappuoli*
IRIS
Chiron S.r.l.
via Fiorentina 1
53100 Siena
Italy
*e-mail: rino_rappuoli@chiron.it.

▼ Until recently, the development of vaccines for use in humans relied on biochemical, immunological and microbiological methods. These conventional approaches have been successful in many cases but they require the pathogen to be grown in laboratory conditions, are time-consuming and allow for the identification of only the most abundant antigens, which can be purified in quantities suitable for vaccine testing. Furthermore, when dealing with non-cultivatable microorganisms, there is no approach to vaccine development.

With the advent of whole-genome sequencing and advances in bioinformatics, the vaccinology field is radically changed, providing the opportunity for developing novel and improved vaccines. Although DNA-sequencing methods have changed little over the years, advances in automation and bioinformatics have enabled the determination of complete microbial genome sequences in a short period of time. One can now mine the sequences for potential surface targets using various algorithms, characterize these gene targets and choose primers for cloning, all before one enters the laboratory.

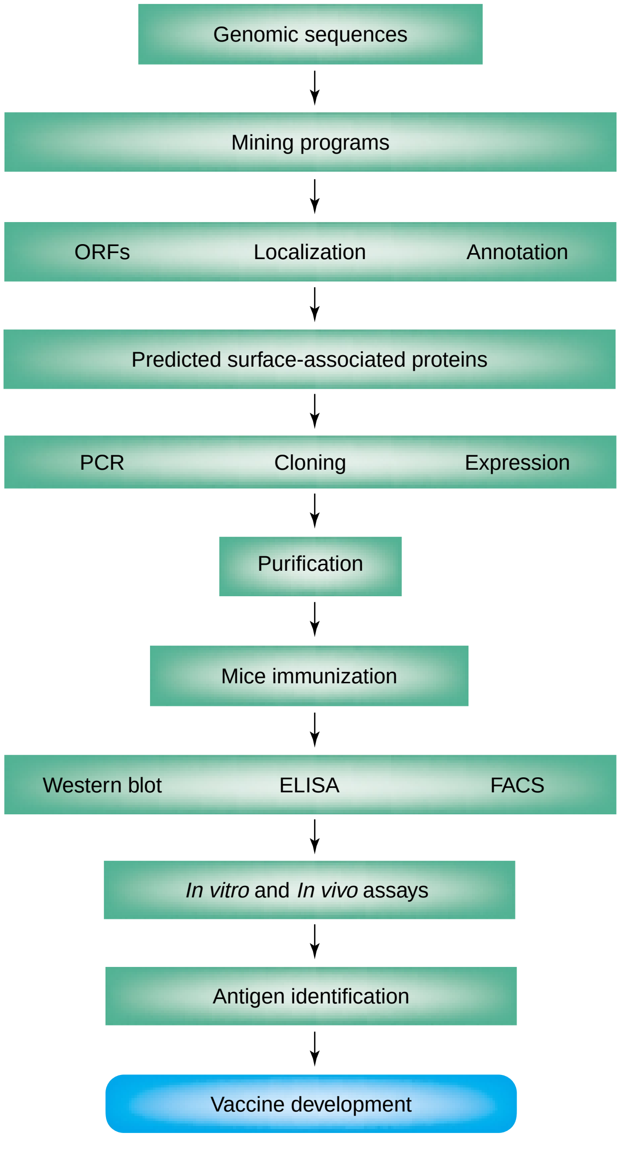
This new approach, termed 'reverse vaccinology' [1,2], in conjunction with advances in molecular biology technologies, enables systematic identification of all the potential antigens of a pathogen, making it theoretically possible to develop a safe and efficacious vaccine against any infectious disease.

Vaccines at the end of the 20th century: classical vaccinology

Conventional vaccines consist of live-attenuated microbes, killed, inactivated microorganisms, purified microbial components, polysaccharide-carrier protein conjugates or recombinant proteins. The two basic types of vaccine, based on the microorganism in an attenuated but live form or on the killed, inactivated microorganism, were among the first vaccines generated. A third type of vaccines, made from the diphtheria and tetanus toxoids, was developed in the 1920s and represents a much more sophisticated product. The two toxins, which were previously shown to be the essential for causing the disease, were chemically detoxified to yield the non-toxic toxoids. Although other purified single-component vaccines (e.g. surface polysaccharides of encapsulated bacteria) are immunogenic, they might not be able to induce immunological memory. The ability of a protein antigen to induce a T-cell immune response can be improved by conjugation to a polysaccharide. These glycoconjugate vaccines were first prepared and studied with *Haemophilus influenzae* type b and have proved to be a great success [3].

In the last quarter of the 20th century, the revolution of microbiology through recombinant DNA techniques provided new tools for vaccine research. In this approach, specific antigens, selected on the basis of immunological data from patients, are purified from

1359-6446/03/$ – see front matter ©2003 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(03)02689-8



Drug Discovery Today

Figure 1. Flow chart of the genome-based approach to vaccine development. This approach involves the *in silico* analysis of microbial genome sequences followed by the high-throughput expression of the genes of interest. The recombinant proteins are then used to immunize mice and the post-immunization sera are analyzed to assess the ability of the polypeptide to elicit a quantitative and qualitative immune response.

heterologous systems in which the corresponding genes have been cloned and tested for safety and efficacy. This approach generated the first recombinant vaccine for human use, the hepatitis B vaccine, in 1984 [4].

The impact of vaccinology in the 20th century has been enormous; however, all these approaches have several limitations, including the fact that proteins that are immunogenic *in vivo* are not necessarily protective antigens, are often variable in sequence and are difficult to express and/or purify in a large amount, leading to high production costs. In addition, the antigens identified are only the most abundant and only a few antigens are analyzed simultaneously. As a result, there are several infectious diseases for which these traditional approaches have failed and for which vaccines have not yet been discovered.

Vaccines at the beginning of the 21st century: reverse vaccinology

The technology for sequencing entire genomes of microorganisms is a recent development in the history of genetics [5]. The first complete genome sequence for any free-living organism (*H. influenzae*) was published by Venter and coworkers in 1995 [6], who employed a strategy of random whole-genome ‘shotgun’ sequencing. It therefore became possible to sequence the entire genomes of prokaryotes with great rapidity and efficiency.

The possibility of determining the whole sequence of a bacterial genome led to the idea of using the genomic information to discover novel antigens that had been missed by conventional vaccinology. This novel reverse vaccinology approach (Fig. 1) involved the *in silico* analysis of the microbial genome sequence [1,2] and was first used to identify antigens as potential candidates for a vaccine against serogroup B meningococcus [7].

In silico *analysis of genomes*

Secreted or extracellular proteins are more easily accessible to antibody than are intracellular proteins and therefore represent ideal vaccine candidates. Various algorithms currently exist to identify proteins with these features from data bank sequences. However, computer algorithms are not always able to correctly predict the cellular localization of proteins [8]. The success of genomic-based strategies for vaccine development is highly dependent on the criteria used for the *in silico* selection of the potential antigens. Several approaches can be used to mine genomic sequences, and the appropriate combination of the various algorithms and the critical evaluation of the information generated are essential for the proper selection of the antigens.

A primary screen for coding capacity is carried out on DNA segments or contigs using database and computer programs included in the Wisconsin package version 10.0 [Genetics Computer Group (GCG); http://www.accelrys.com]. As a second step, all the predicted open reading frames (ORFs) are used for homology searches against a database with BLASTX, BLASTN and TBLASTX programs [9,10] to identify DNA segments with potential coding regions. The ORFs coding for known cytoplasmic functions or known antigens are excluded, whereas the other coding regions are selected for further analysis. A third screening step, designed to identify putative proteins with a cellular localization spanning the inner membrane to outside the bacterium, is applied. BLAST, FASTA, MOTIFS, FINDPATTERNS and PSORT [11,12], in addition to the ProDom [13,14], Pfam [15] and Blocks [16,17] databases, are used to predict features typical of surface-associated proteins, such as transmembrane domains, leader peptides, homologies to known surface proteins, lipoprotein signatures, outer membrane anchoring motives and host-cell binding domains such as RGD [18].

The genome is also searched for proteins that are homologous to putative virulence factors previously characterized in other organisms. The sequences are examined for the presence of tandem repeats in or at the 5′ ends of genes that characterize certain virulence genes [19,20]. Owing to changes in the number of copies of tandem repeats, the expression of these genes is turned on or off (phase variation), leading to an attenuated or virulent phenotype. In addition, the G+C content of the genome is analyzed to establish the possible acquisition of genes by horizontal transfer. Finally, the bioinformatic analysis is expanded to include similarity searches against unfinished microbial genome sequence databases, in addition to the available sequences of related genomes for sequence conservation.

High-throughput expression

The *in silico* approach results in the selection of a large number of genes, covering as much as 25% of the total number of ORFs in the genome. Therefore, it is necessary to use simple procedures that permit large numbers of genes to be cloned and expressed. Fortunately, the development of robotics and PCR make this possible. Oligonucleotide primers for the PCR reaction are designed from the genomic sequences. Each pair of primers also contains sequences that correspond to appropriate restriction sites for cloning into prokaryotic expression vectors or recognition sites for recombinases where *in vitro* recombination is employed to clone genes. The product of each PCR reaction is then cloned and screened for expression in a heterologous system. Successful expression depends on the predicted localization of the protein. Integral membrane proteins have proven to be particularly difficult to produce by recombinant techniques in *Escherichia coli*.

The most commonly used methods to express genes are based on their fusion to a histidine-tag and/or to glutathione S-transferase to facilitate rapid purification of the recombinant proteins by simple affinity column chromatography. In the case of insoluble antigens, the purification with denaturing agents (e.g. urea) is followed by a renaturation procedure based on two-step dialysis in the presence of arginine, glycerol and, in the case of proteins with disulfide bridges, the addition of reduced and oxidized glutathione. Enhancement of solubility, yield and purity of the antigens is, in some cases, achieved by using detergents or changing the temperature at which the microbial strain is grown.

Immunogenicity testing

Once purified, the recombinant proteins are used to immunize mice and the post-immunization sera are analyzed to verify the computer-predicted surface localization of each polypeptide and their ability to elicit a quantitative and qualitative immune response.

First, the immune sera are tested using western blot analysis of the recombinant proteins, outer membrane vesicles (OMVs) and total extracts of the bacterium to determine if the antibodies are able to recognize both the recombinant and the bacterial protein, and to confirm the predicted localization of the protein. A limitation of immunoblotting is that it requires boiling of the samples, which results in disruption of the native structure of antigens, preventing antibody from binding to conformational epitopes. To further confirm the presence of the proteins on the bacterial surface and to assess their immunogenicity, sera are analyzed by ELISA and fluorescence–activated cell sorter analysis (FACS) to measure antibody titers and to determine the ability of antisera to bind to the surface of live bacteria.

The direct means to study the protective efficacy of candidate antigens is to test the immune sera in an animal model in which protection is dependent on the same effector mechanisms as in humans. The lack of reliable animal models has often hampered the development of vaccines, and alternative *in vitro* assays that are known to correlate with vaccine efficacy in humans have to be developed (e.g. assays that measure bactericidal activity [21] and opsonophagocytosis [22]).

A milestone: MenB

Neisseria meningitidis is a human pathogen that, despite available antibiotic therapy, is still a major cause of mortality as a result of sepsis and meningitis. Using traditional approaches, vaccines have been developed against serogroups A, C, Y, and W135, but for group B meningococcus (MenB), an efficacious vaccine is not yet available. Sequence variation of surface-exposed proteins and cross-reactivity of the serogroup B capsular polysaccharide with human tissues have hampered efforts to develop a successful vaccine. Therefore, a completely different approach was needed for this bacterium.

MenB represents the first example of the application of reverse vaccinology and the demonstration of the power of genomic approaches for target antigen identification [7]. While the *N. meningitidis* sequencing project was in progress, the incompletely assembled DNA fragments were screened by computer analysis to select proteins predicted to be on the bacterial surface or those with homologies to known bacterial factors involved in pathogenesis and virulence. After discarding cytoplasmic proteins and known *Neisseria* antigens, 570 genes predicted to code for surface-exposed or membrane-associated proteins were identified. Successful cloning and expression was achieved for 350 proteins, which were then purified and tested for localization, immunogenicity and protective efficacy. Of the 85 proteins found to be surface-exposed, 22 were able to induce complement-mediated bactericidal antibody response, providing a strong indication of proteins capable of inducing protective immunity. In addition, to test the suitability of these proteins as candidate antigens for conferring protection against heterologous strains, the proteins were evaluated for gene presence, phase variation and sequence conservation in a panel of genetically diverse MenB strains representative of the global diversity of the natural *N. meningitidis* population [23]. Most of the selected antigens were able to induce cross-protection against heterologous strains, demonstrating that the antigens, identified by *in silico* analysis, are good candidates for the clinical development of a vaccine against MenB [7–24].

A more general application: other pathogens

The availability of an increasing number of bacterial genome sequences, together with the MenB example, has prompted the application of the reverse vaccinology approach to other pathogens. The use of bioinformatics tools in combination with molecular biology techniques, enables the systematic investigation of the utility of potential genomic sequences to act as antigens for vaccine production. It is now possible to conceive of the development of new vaccines against a wide variety of pathogens for which classical vaccinology has failed so far and, in theory, this approach could be extended to parasites and viruses.

Streptococcus pneumoniae

Streptococcus pneumoniae is the leading cause of bacterial sepsis, pneumonia, meningitis and otitis media in young children in the USA. Vaccines in current use have poor efficacy in infants and there is little or no cross-protection between the different serotypes. To identify a more suitable vaccine candidate, the whole genome sequence of *S. pneumoniae* was scanned and 130 potential ORFs with significant homology to surface proteins and virulence factors of other bacteria were selected [25]. From this set, 108 ORFs were expressed and purified for mice immunization and evaluated as vaccine candidates. A subset of six novel antigens was able to induce protective antibodies against pneumococcal challenge in a mouse sepsis model. Furthermore, the six predictive targets showed a high degree of cross-reactivity against the majority of capsular antigens that are expressed *in vivo* and are immunogenic during human infection. Thus, this genomic approach to vaccine development led to the identification of new antigens, which could constitute parts of a more powerful vaccine.

Porphyromonas gingivalis

Another application of the reverse vaccinology strategy is represented by the attempt to develop a protein-based vaccine for the prevention of chronic adult periodontitis disease caused by *Porphyromonas gingivalis*, a gram-negative anaerobic bacterium that has been found in subgingival plaques. From an anticipated 2000 genes, 120 proteins were selected for screening using bioinformatics tools [26] and 107 of these were expressed in *E.coli* and analyzed by western blotting using sera from human periodontitis patients and animal antisera. These candidates were reduced to a group of 40 proteins, which were purified and used to immunize mice. The mice were then challenged with the bacteria in a subcutaneous abscess model. From this subset of candidates, two antigens demonstrated significant protection in this animal model. These proteins have also shown homology to OprF of *Pseudomonas aeruginosa*, which is part of the vaccine formulation that is currently in human clinical trial. Therefore, these proteins could represent potential vaccine antigen candidates.

Staphylococcus aureus

To identify vaccine candidates for *Staphylococcus aureus*, an approach based on genomic libraries was developed [27]. This represents an alternative way to use whole-genome sequence information. *S. aureus* peptides were displayed on the surface of *E. coli* via fusion to one or two outer membrane proteins (LamB and FhuA) and probed with sera selected for high antibody titers and opsonic activity. The exhaustive screening of the two different peptide expression libraries by the application of MACS technology (magnetic cell sorting) enabled the profile of antigens that are expressed *in vivo* and that are able to elicit an immune response in humans to be identified. A total of 60 antigenic proteins were identified, most of which were predicted to be secreted or located on the surface of the bacterium, and their antibody-binding sites were mapped. This approach, which makes use of whole-genome sequence information, can facilitate the development of novel vaccines and can be extended to other related bacteria.

Chlamydia pneumoniae

Chlamydia pneumoniae is an obligate intracellular bacterium and a human pathogen associated with respiratory infections, atherosclerotic and cardiovascular disease. *Chlamydia* has developed a biphasic life cycle, with two morphologically different forms: a spore-like infectious form, called elementary bodies (EBs), and an intracellular form, called reticulate bodies (RBs). The technical difficulty in working with *Chlamydia* and the absence of reliable tools for genetic manipulation have prevented the characterization of the EB form.

Using a systematic genomic–proteomic approach to identify EB surface proteins, 28 surface-exposed proteins were identified [28]. Through *in silico* selection, 157 putative proteins were identified. The recombinant form of these proteins was expressed in *E. coli*, purified and used to immunize mice. The antisera were used to detect cell surface localization of the putative proteins by FACS analysis. This methodology led to the identification of 53 proteins that were also screened by western blot analysis. To support FACS and western blot data, a proteomic analysis of EB total proteins was carried out using 2D gel electrophoresis combined with spot identification by mass spectrometry. Combining the results, 28 antigens were identified. This work represents the first successful attempt at a systematic genome–proteome analysis of protein organization on the cell surface in *C. pneumoniae* and a rational way to select new vaccine candidates.

Bacillus anthracis

Bacillus anthracis is the causative organism of the potentially fatal disease anthrax. Fully virulent forms of *B. anthracis* carry the two large plasmids pXO1 and pXO2, which are considered to be the major virulence determinants. pXO1 carries the protective antigen (PA), the common cell-binding domain capable of interacting with two different domains, the lethal and edema factors, which elicit cell damage. The PA is the major immunogenic component of the vaccines licensed for use in humans and requires multiple immunizations and occasionally provokes cases of reactogenicity. Identification of novel antigens is therefore essential for the development of second-generation *B. anthracis* vaccines. This search focused on secreted and surface-exposed proteins along with ORF products similar to proteins involved in bacterial pathogenesis [29]. Using functional genomic analysis, 11 candidates were selected. In this case, a simple method that relied on the *in vitro* translation of the linear full-length DNA of the selected ORFs was used. Polypeptides obtained *in vitro* were then evaluated for immunogenicity by analysis of their reactivity with hyperimmune anti-*B. anthracis* antisera. The combination of bioinformatic genomic analysis and an efficient and fast screening facilitated the identification of unknown antigenic proteins, three of which appear to be similar to immunogenic PA, the major constituent of the *B. anthracis* vaccine. These new proteins could represent parts of a second-generation anthrax vaccine.

Conclusions

The late 1990s marked the beginning of the era of genomics. Genomic sequencing and analysis are now in a period of ‘exponential growth.’ More than 90 eukaryotic and prokaryotic genomes have been completely sequenced, with analysis of more than 200 genome sequences currently under way.

The availability of the entire genetic content of several important human pathogens has opened new innovative and efficient research avenues to identify a wide array of new antigens for vaccines against any infectious disease. However, genome data alone cannot be used to accurately predict the *in vivo* efficacy of candidate antigens. Therefore, vaccine candidates selected on the basis of *in silico* criteria need to be validated using genomic, proteomic, genetic, biochemical and bioinformatic approaches, in addition to appropriate animal models.

The advantage of the reverse vaccinology approach is that it is not limited by growth conditions or detection methods for gene expression. This strategy identifies a small number of the most promising new vaccine candidates from a large number of antigens and enables more rapid development of these candidates at a reasonable cost compared with traditional strategies.

References

1 Rappuoli, R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol.* 3, 445–450

2 Rappuoli, R. (2001) Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 19, 2688–2691

3 Mäkelä P.H. *et al.* (1995) Vaccines against *Haemophilus influenzae* type B. In *Molecular and Clinical Aspects of Bacterial Vaccine Development* (Ala’Aldeen, D.A.A. and Hormaeche, C.E., eds), pp. 41–91, John Wiley and Sons

4 McAleer, W.J. *et al.* (1984) Human hepatitis B vaccine from recombinant yeast. *Nature* 307, 178–180

5 Broder, S. and Venter, J.C. (2000) Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.* 40, 97–132

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

7 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820

8 Chakravarti, D.N. *et al.* (2000) Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine* 19, 601–612

9 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

10 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

11 Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* 54, 277–344

12 Nakai, K. and Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36

13 Corpet, F. *et al.* (1998) The ProDom database of protein domain families. *Nucleic Acids Res.* 26, 323–326

14 Corpet, F. *et al.* (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.* 27, 263–267

15 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266

16 Henikoff, J.G. *et al.* (1999) New features of the blocks database servers. *Nucleic Acids Res.* 27, 226–228

17 Henikoff, S. and Henikoff, J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics* 19, 97–107

18 Brennan, M.J. and Shahin, R.D. (1996) Pertussis antigens that abrogate bacterial adherence and elicit immunity. *Am. J. Respir. Crit. Care Med.* 154, S145–149

19 Saunders, N.J. *et al.* (1998) Simple sequence repeats in the *Helicobacterpylori* genome. *Mol. Microbiol.* 27, 1091–1098

20 Hood, D.W. *et al.* (1996) DNA repeats identify novel virulence genes in *Haemophilus influenzae. Proc. Natl. Acad. Sci. U. S. A.* 93, 11121–11125

21 Goldschneider, I. *et al.* (1969) Human immunity to the meningococcus. I. The role of humoral antibodies. *J. Exp. Med.* 129, 1307–1326

22 Ross, S.C. *et al.* (1987) Killing of *Neisseria meningitidis* by human neutrophils: implications for normal and complement-deficient individuals. *J. Infect. Dis.* 155, 1266–1275

23 Maiden, M.C.J. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3140–3145

24 Jodar, L. *et al.* (2002) Development of vaccines against meningococcal disease. *Lancet* 359, 1499–1508

25 Wizemann, T.M. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* 69, 1593–1598

26 Ross, B.C. *et al.* (2001) Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis. Vaccine* 19, 4135–4142

27 Etz, H. *et al.* (2002) Identification of *in vivo* expressed vaccine candidate antigens from *Staphylococcus aureus. Proc. Natl. Acad. Sci. U. S. A.* 99, 6573–6578

28 Montigiani, S. *et al.* (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae. Infect. Immun.* 70, 368–379

29 Ariel, N. *et al.* (2002) Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: *in silico* and *in vitro* screening. *Infect. Immun.* 70, 6817–6827

Editor's choice

bmn.com/genetics

As a busy scientist, searching through the wealth of information on BioMedNet can be a bit daunting – the new gateway to **genetics** on BioMedNet is designed to help.

The new **genetics** gateway is updated weekly and features relevant articles selected by the editorial teams from *Drug Discovery Today, Trends in Biotechnology* and *Current Opinion in Biotechnology*.

The regular updates include:
News – our dedicated team of reporters from BioMedNet News provide a busy researcher with all the news to keep up-to-date on what's happening – right now.
Journal scan – learn about new reports and events in genetics every day, at a glance, without leafing through stacks of journals.
Conference reporter – daily updates on the most exciting developments revealed at key conferences in the life sciences – providing a quick but comprehensive report of what you missed by staying home.
Minireviews and Reviews – a selection of the best review and opinion articles from all *Trends* and *Current Opinion* journals and *Drug Discovery Today*.

Why not bookmark the gateway at http://bmn.com/genetics for access to all the news, reviews and informed opinion on the latest scientific advances in genetics.