Drug Discovery Today • Volume 14, Numbers 5/6 • March 2009

The use of genomics in microbial vaccine development

Stefania Bambini and Rino Rappuoli

Novartis Vaccines, Via Fiorentina 1, 53100 Siena, Italy

Vaccination is one of the most effective tools for the prevention of infectious diseases. The availability of complete genome sequences, together with the progression of high-throughput technologies such as functional and structural genomics, has led to a new paradigm in vaccine development. Pan-genomic reverse vaccinology, with the comparison of sequence data from multiple isolates of the same species of a pathogen, increases the opportunity of the identification of novel vaccine candidates. Overall, the conventional empiric approach to vaccine development is being replaced by vaccine design. The recent development of synthetic genomics may provide a further opportunity to design vaccines.

'New powerful genomics technologies have increased the number of diseases that can be addressed by vaccination, and decreased the time for discovery research and vaccine development'

Infectious diseases, about half of which are caused by bacteria, remain the leading cause of death worldwide, despite antimicrobial therapies. Emerging infectious disease events are caused by strains that are resistant to antimicrobials, or by etiologic agents that have recently also become human pathogens increasing their spectrum, and by pathogens that have been present in humans previously but that have unexpectedly increased their incidence. Dozens of new infectious diseases are expected to emerge in the coming decades and the development of novel vaccines against these diseases appears an attractive approach for their control and prevention [1,2].

Vaccines are one of the most effective ways to prevent infectious diseases and to minimize their impact on the human population. The basic paradigm of vaccine development established by Louis Pasteur at the end of the 19th century (i.e. isolation, inactivation and injection of the causative agent) constitutes the foundation of classical vaccinology and led vaccine development throughout the 20th century [3]. Conventional approaches on the basis of these empiric principles have provided vaccines from three major categories: inactivated microorganisms, live-modified agents and subunit vaccines, composed by purified portions of the infectious agent [4]. Not all pathogens, however, can be grown in culture and some microorganisms may require specific, sometimes expensive, cell-cultures for growth. The application of safety procedures may be required for pathogen manipulation, and insufficient killing or attenuation may result in the presence of virulent organisms in the final vaccine. In addition, the conventional vaccine discovery approaches are time-consuming and could take decades. Recombinant DNA technologies have been used for the design of second-generation vaccines, to obtain rationally attenuated strains or highly purified antigenic components. Examples include bacterial toxins detoxified by molecular engineering, such as the pertussis toxin [5]. This approach, however, even if more refined, could require years [6]. In some cases, the conventional empiric approach was just insufficient to find appropriate solutions for the development of universal vaccines (i.e. meningococcus B).

With the recent development of genomics, the combination between the *in silico* analysis of genome sequences, a procedure known as 'genome mining', and the knowledge from functional and structural genomics provide novel strategies for a more rapid identification of antigens leading to a third generation of vaccines.

The genomics revolution

Although the history of genomics research can be traced back to the 1970s, with the development of DNA sequencing technology,

Corresponding author: Rappuoli, R. (rino.rappuoli@novartis.com)

1359-6446/06/$ - see front matter © 2009 Published by Elsevier Ltd. doi:10.1016/j.drudis.2008.12.007

the late 1990s marked the beginning of the so-called genomics era, with the first complete genome sequenced of the free-living organism *Haemophilus influenzae* in 1995 [7]. Since then, emerging technologies have allowed the sequencing of a genome to be completed very quickly and, during the past decade, sequencing of entire genomes has become a commonly used practice in research [8]. Besides bacterial genomes, several eukaryotic genomes, including the human genome, have been completely sequenced [9]. Genomics, studying the genome of organisms as a whole, and postgenomics technologies, investigating RNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics), have had a considerable impact in all areas of biological research [10], and the field of vaccinology is no exception. The rapid availability of complete and accurate pathogen genome sequences and the increase of analysis tools available for the mining of biological information included in genome sequences have remarkably decreased the time for vaccine discovery research and development (Fig. 1). Genome mining has made it possible to predict genes potentially encoding factors that promote pathogenesis, for example on the basis of sequence similarities to known pathogenic proteins already present in the database, to assign gene functions and to predict some features of the encoded proteins, such as cellular location, molecular weight, p*I* or solubility. In this regard, reverse vaccinology, combined with the knowledge from comparative and experimental genomics, can be considered a very attractive approach, providing conserved putative antigens in the shortest possible time. Moreover, the development of postgenomics approaches has accelerated the discovery of factors related to pathogenesis, which are key elements in the design of new vaccines. Genomics approaches are indeed of a revolutionary power in the understanding of the molecular mechanisms of disease, with the possibility of using genomic information for the discovery of potential vaccine candidates leading to a new paradigm in vaccine development (Fig. 2).

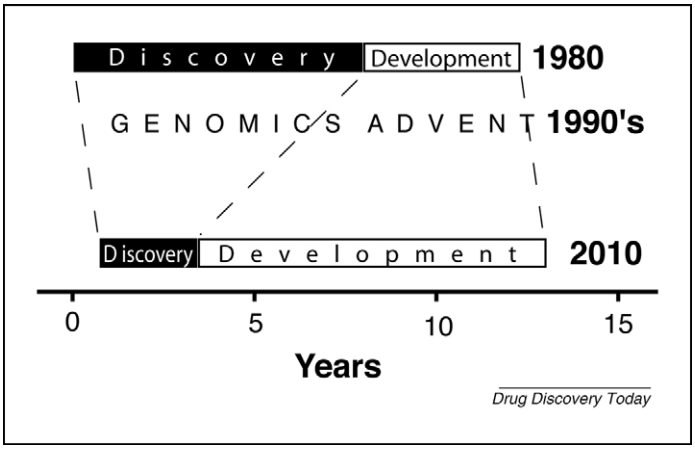

FIGURE 1

The impact of genomics in vaccine discovery research and development. In the past, the conventional vaccine discovery approaches were time-consuming and the identification of protective antigens took years or decades. The genomics advent in the 1990s, with the availability of whole-genome sequences and high-throughput technologies, and the progress in immunology have provided novel strategies for a more rapid (one to two years) identification of antigens, dramatically decreasing the time for discovery research and vaccine development. Following stages of development and manufacture until vaccine launch could, however, become more complex in the near future, resulting from the emergence of new infectious diseases and the largest request of more safe, effective and cheap vaccines. Major challenges still remain in optimizing the commercial aspect of vaccine development, due mainly to increasingly stringent regulatory requirements, to economic limits, to the need of increasing global priority attributed to vaccination and to other socioeconomic issues.

Reverse vaccinology: an *in silico* approach

The approach referred to as 'reverse vaccinology' uses the genome sequences of viral, bacterial or parasitic pathogens of interest rather than the cells as starting material for the identification of novel antigens, whose activity should be subsequently confirmed by experimental biology [11]. In general, the aim is the identification of genes potentially encoding pathogenicity factors and secreted or membrane-associated proteins. Specific algorithms suitable for the *in silico* identification of novel surface-exposed and, thus, antibody accessible proteins mediating a protective response are used.

The first example of a successful application of the reverse vaccinology approach was provided by Pizza and coworkers in collaboration with The Institute for Genomic Research (TIGR) [12,13]. They describe the identification of vaccine candidates against *Neisseria meningitidis* serogroup B or MenB, the major cause of sepsis and meningitis in children and young adults. Conventional approaches to obtain a MenB vaccine had failed for decades, mainly for two reasons. First, the capsular polysaccharide, successfully used to make conjugate vaccines against other serogroups, resembles components of human tissues resulting poorly immunogenic in humans and potentially able to induce an autoimmune response [14]. Second, protein-based vaccines focused on variable antigens that confer protection only against a limited number of strains.

The reverse vaccinology approach started from the determination of the complete genome sequence of a MenB pathogenic strain, MC58. Several computational tools were used to find in the genome, on the basis of sequence features, the presence of amino acid motifs responsible for targeting the mature protein to the outer membrane (signal peptides), to the lipid bilayer (lipoproteins), to the integral membrane (transmembrane domains) or for recognition and interaction with host structures. This analysis allowed researchers to identify 600 putative surface-exposed or secreted proteins. Three hundred and fifty proteins were expressed in a heterologous system, *Escherichia coli*, purified and used to immunize mice. Mice immune sera were tested for specificity by Western blot, accessibility on the surface of bacterium by flow cytometry and for the capacity to induce bactericidal antibodies by serum bactericidal assay. The last mentioned test estimates the ability of bactericidal antibodies to evoke a complement-mediated lysis of bacteria and, in the case of meningococcus, is internationally accepted as correlate of protection against the microorganism. Twenty-nine of these surface-exposed proteins were found to be bactericidal. The selected candidates were then checked for sequence conservation across a panel of strains representing the genetic diversity of meningococcus and including the clonal complexes most frequently disease-related. This analysis allowed the further selection of antigens able to elicit a cross-bactericidal response against most strains included in the panel. The results of this study were used for the setting up of 5CVMB, the recently presented universal vaccine against *Neisseria meningitidis* that is

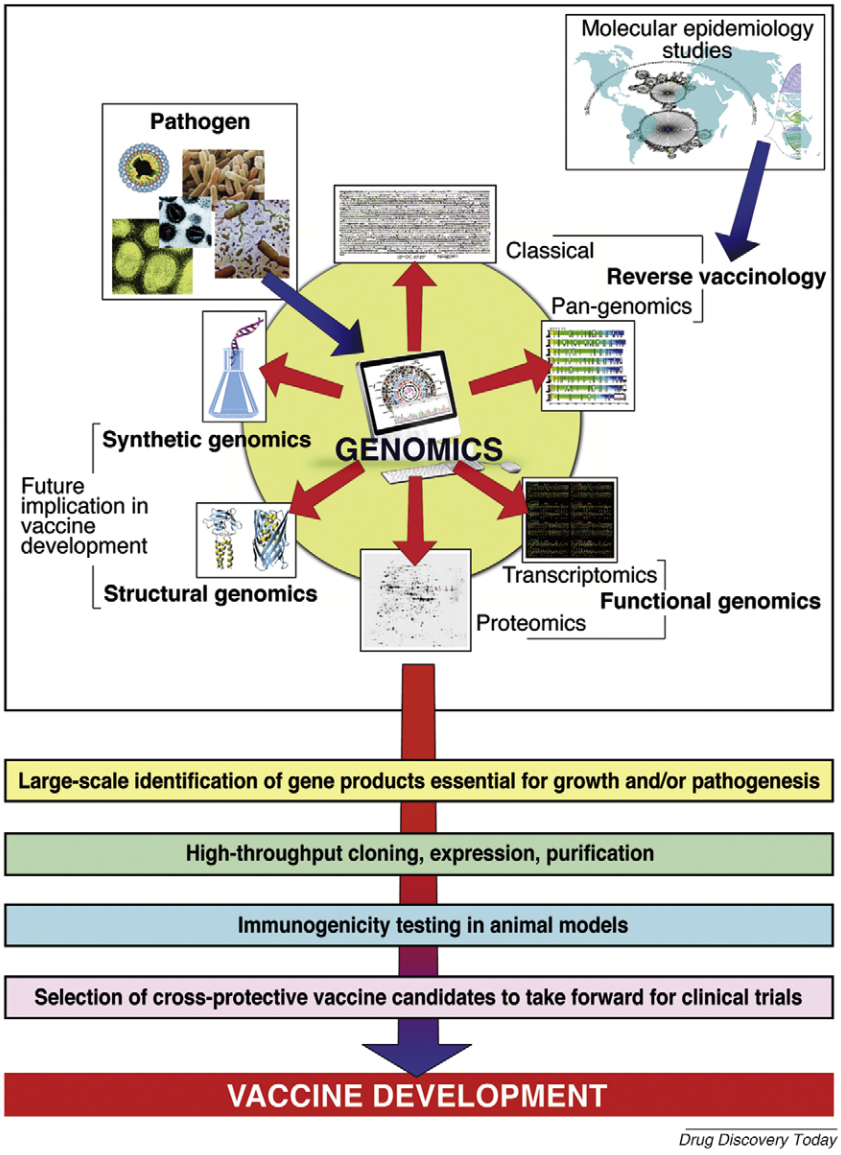

FIGURE 2

Schematic image of the main -omics fields and their applications in vaccine discovery research and development. The availability of complete pathogen genome sequences, with the contemporary development of bioinformatics tools, has led to a new paradigm of the vaccine development. The whole-genome sequence accessibility allowed the advent of the classical and, more recently, the pan-genomic reverse vaccinology, which are supported by molecular epidemiology studies for the selection of strains representative of a given pathogen. Functional (transcriptomics and proteomics) genomics are playing a central role in the understanding of host–pathogen interaction. Structural genomics, with the atomic resolution of the structure, could provide the basis for the rational engineering of potential antigens. Moreover, in very recent times, we are attending to the development of synthetic genomics. All these areas can contribute and/or have the potential to favor the development of the so-called third-generation vaccines, through the selection or design of promising vaccine candidates to take forward into clinical trials.

based on a 'cocktail' of five antigens identified by reverse vaccinology (fHBP, NadA, GNA2132, GNA1030 and GNA2091). This vaccine is currently tested in Phase II clinical trials [15]. The success of reverse vaccinology for meningococcus has led to the application of this approach to a variety of other human pathogens, such as *Streptococcus pneumoniae*, *Chlamydia pneumoniae*, *Bacillus anthracis*, *Porphyromonas gingivalis*, *Mycobacterium tuberculosis*, *Helicobacter pylori* and others [16–21] (Table 1). In most cases, some identified antigens have already entered the clinical or development phase [22].

Genomics-based studies were also recently started for the development of vaccines against viruses. The availability of the small genome of Hepatitis C virus (HCV), which until recently was not possible to cultivate *in vitro*, allowed the identification of genes encoding for the viral envelope proteins that were used for developing recombinant vaccines now in clinical trials [23]. After the recent emergence of severe acute respiratory syndrome (SARS), the genome sequence of the responsible agent, a coronavirus, was rapidly determined and made available, providing a rapid (less than a month) identification of protective vaccine candidates

TABLE 1

Examples of different postgenomics approaches in the development of vaccines against some bacterial pathogens, and the status of the corresponding vaccine development

Pathogen	Disease	Approach	Refs	Status of vaccine development
Neisseria meningitidis serogroup B	Bacterial meningitis and septicemia	Reverse vaccinology Microarray Proteomics	[12,13,15] [37,38] [49]	Phase II clinical trials
Streptococcus pneumoniae	Bacterial pneumonia, sepsis, sinusitis, otitis media and bacterial meningitis	Classical or comparative reverse vaccinology Proteomics	[16,50] [51]	Discovery/preclinical
Bacillus anthracis	Anthrax	Reverse vaccinology CGH microarray Microarray Proteomics and immunoproteomics	[18] [52] [53] [54]	Discovery/preclinical
Staphylococcus aureus	Variety of infections, including 'pelvic syndrome', rapidly progressive pneumonia, ocular infections, septic thrombophlebitis	CGH microarray Immunoproteomics	[55] [41]	Discovery/preclinical
Porphyromonas gingivalis	Periodontitis	Reverse vaccinology	[19]	Discovery/preclinical
Mycobacterium tuberculosis	Tuberculosis	Reverse vaccinology	[21]	Discovery/preclinical
Helicobacter pylori	Ulcer, atrophic gastritis, adenocarcinoma, lymphoma	Reverse vaccinology Immunoproteomics	[20] [56]	Discovery/preclinical
Streptococcus agalactiae (GBS)	Bacterial sepsis, pneumonia, meningitis	Classical or comparative reverse vaccinology	[36]	Discovery/preclinical
Streptococcus pyogenes (GAS)	Many systemic invasive infections including necrotizing fasciitis, myositis, pneumonia, sepsis, arthritis	Genome-wide analysis Proteomics (surface proteome)	[57] [40]	Discovery/preclinical
Chlamydia pneumoniae	Pneumonia, meningitis, middle era infections	Reverse vaccinology and proteomics	[17]	Discovery/preclinical

[24–26]. For some viruses, such as human immunodeficiency virus (HIV), the development of effective vaccines has been hampered by the high antigenic diversity of the pathogen. Computer-based methods have been used to identify sequences encoding short peptides representing the whole antigenic diversity of potential T-cell epitopes for the construction of proteins to be used as polyvalent vaccine antigens [27].

A further evolution in the *in silico* approaches has been provided by the availability of multiple complete genome sequences. Comparative, or pan-genomic, reverse vaccinology, which compares the genomes of related strains, could be considered as the progression of classical reverse vaccinology, on the basis of a single genome sequence to predict putative vaccine candidates, and could be used to develop universal vaccines, effective against different strains of a same species [28,29].

Comparative genomics and the concept of pan-genome

After more than 10 years since the first bacterial genome sequence was published, more than 700 complete microbial genomes are now available (Box 1). One of the most obvious conclusions from examining the sequences from bacterial genomes is the enormous amount of diversity, even in different genomes from the same species. This diversity is generated by a variety of mechanisms that determine horizontal gene transfer, including mobile genetic elements like transposable elements and bacteriophages [30]. The concept that genetic diversity intraspecies can be as significant as interspecies variation was revealed by subtractive hybridization and comparative genome hybridization (CGH) studies, where isolates belonging to the same bacterial species were analyzed by microarray using a sequenced strain as a reference (e.g. *Campylobacter jejuni*, *Escherichia coli* and *Shigella* [31,32]). These studies show that there are genes that are not conserved in all strains of the same species and that there is an extensive genetic diversity. CGH experiments, however, are not able to identify genes absent in the reference genome.

The rapid development of sequencing technologies, providing complete genome sequences for different isolates available for comparative analyses, has driven the use of genomics for the investigation of the variability within a single species (population genomics). While at the beginning of the genomic era genomes available belonged mainly to representative pathogenic strains of different species, the availability of a huge amount of sequence data stimulated the development of comparative genomics, which performs interspecies and intraspecies comparisons providing new insights into predictive biology [33,34]. In infectious diseases research, comparative genomics, on the basis of powerful tools

BOX 1

Bacterial genomes available

To date, 2063 microbial genomes are in available databases (e.g. http://www.sanger.ac.uk/pathogens or http://www.tigr.org/tdb/mdb/mdb.html), 774 are complete, 1289 are currently in progress (595: sequence available, 694: no sequence available).
Source: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html.

of bioinformatics and microarray technology, takes advantage of the comparison of genome sequences to identify virulence determinants, drugs targets and vaccine candidates. For instance, the comparison of genome sequences of closely related pathogenic and nonpathogenic bacteria can help in identifying virulence associated genetic patterns.

Tettelin *et al*. introduced the concept of pan-genome, which was defined as the global gene repertoire pertaining to a given species [35]. They underlined the intraspecies diversity by a study based on whole-genome sequence comparison of eight different strains of *Streptococcus agalactiae* (group B streptococcus or GBS) representative of species genetic diversity. The results showed that the pan-genome, consisting of core genes that are shared by all strains, and dispensable genes, which are absent in at least one strain, could be much larger than the genome of a single strain, given that the number of unique genes is vast. In fact, new genes continue to be added to the gene pool any time a new strain is sequenced. Besides the study of the diversity inside a species, one possible application of the pan-genome in vaccinology (pan-genomic reverse vaccinology) is the identification of novel vaccine candidates and targets for antimicrobials. Maione and colleagues performed the first application of the pan-genome to vaccines to design a universal vaccine against GBS [36]. By computational algorithms they predict 589 surface-associated proteins, 396 of which were core genes and those remaining were genes absent in at least one strain. Selected potential antigens were expressed as recombinant proteins, purified and tested for protection against GBS, and four were found to elicit protective immunity in an animal model. Among these antigens, only one was part of the core genome; however, it was not able to confer global protection, hence the final vaccine formulation should include a combination of the four antigens. The GBS example has demonstrated that multiple genome sequences of each species are important to cover the diversity of many pathogens. In the case of highly differentiated species, the genotypic variability of the pathogens could be a problem for the development of protein-based vaccines, which should be designed to cover a wide panel of strains (population vaccinology). For this purpose, the contribution of molecular epidemiology studies for the selection of worldwide representative strains clusters of a given microorganism could be fundamental (Box 2).

BOX 2

Relevance of molecular epidemiology in vaccine development

Molecular epidemiology studies concern the analysis of the microevolution of strains and the detection of new emerging clones. Molecular techniques, such as multilocus sequence typing (MLST), are suitable for genotyping, to identify specific clones and to study the population structure and the genetic diversity of a given microorganism [58]. Theses methods, which allow the identification of the most frequent disease-associated clonal complexes, are very important also for the surveillance of the disease over a period of years or for the analysis of relationships between invasive and carrier strains. For strain characterization, several web-accessible databases have been developed for a rapid assignment of sequence types and antigen variants, providing tools for molecular epidemiology and vaccine development (e.g. http://www.mlst.net/).

In the development of a universal vaccine, capable of inducing protection against a wide spectrum of strains and including a combination of antigens, molecular epidemiology studies are important to take into account the general trends and changes in the pathogen population structure, to indicate strain clusters, worldwide representative of a given microorganism diversity, and to support analyses on antigen diversification within and between the major clonal complexes. Moreover, molecular epidemiology studies could help to monitor the effects of the vaccine introduction, to evaluate its potency, the occurrence of vaccine escape variants or the appearance of novel pathogenic variants as a consequence of the strong selective pressure imposed.

Genome-based technologies: functional and structural genomics

The study of bacterial gene expression and function is essential for understanding pathogenesis and the interaction between pathogen and host. For this kind of investigation, there are two main subfields in genomics: functional genomics and structural genomics (Fig. 2).

Functional genomics has emerged as a scientific field from molecular biology characterized by the development of large-scale technologies such as transcriptomics and proteomics. They contribute to vaccinology in the selection of appropriate vaccine candidates not by examining directly the genetic content but by the transcription and expression profiles.

Transcriptomics provides an overview of the overall transcriptional activity of a given pathogen and allows the comparison of gene expression under different growth and environmental conditions. For vaccine antigen discovery, it is important to know what genes are upregulated *in vivo*, during infection, because they could represent protective vaccine candidates. DNA microarray technology is used to study the gene expression profile of genes. An example of the use of a microarray-based transcriptional profiling to identify potential MenB vaccine candidates was described by Grifantini *et al*. [37]. In this study, bacteria were incubated with human epithelial cells, cell-adhering bacteria were recovered and total RNA was purified at different times. In parallel, RNA was prepared from nonadherent bacteria grown in the absence of epithelial cells. The two RNA preparations were comparatively analyzed on DNA microarrays carrying the entire repertoire of PCR-amplified MenB genes. Twelve proteins whose transcription was found to be particularly activated during adhesion were expressed in *E. coli*, purified and used to produce antisera in mice. Five sera showed bactericidal activity against different strains. Subsequent transcriptome analyses showed that differentially regulated genes were there also during the later interaction of *N. meningitidis* with endothelial cells of the blood–brain barrier [38].

Proteomics is the study of proteins that are expressed in a cell. It has the advantage of defining proteins that are differentially expressed and differentially located, for example those situated outside the cell, the so-called 'surfaceome' [39], important in inducing immune responses. By this approach, the protein mixture is first resolved into individual components using separation procedures like 2D gel electrophoresis or liquid chromatography, digested with specific proteases and the molecular mass of each peptide fragment is then measured using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) and tandem mass spectrometry (MS/MS). The peptide-mass fingerprint is used

for a database search of predicted masses that result from the digestion of known proteins. Classical proteomics approaches provide an experimental support to the *in silico* prediction of protein localization and have already been used to identify additional proteins suitable as vaccine components. As an example, Rodriguez-Ortega *et al*. analyzed the surface proteome of *Streptococcus pyogenes* (Group A *Streptococcus*, GAS) by the surface digestion of live bacteria with different proteases to identify proteins expressed on the bacterial surface and thus accessible to antibodies [40]. Fragments generated after surface digestion of strain M1-SF370 were recovered, analyzed by MS/MS spectrometry, identified by comparison with public genome sequences and a novel possible vaccine target, Spy0416, was found. A proteomics approach could also be considered as complementary to reverse vaccinology, as shown by Montigiani and colleagues in the case of *Chlamydia pneumoniae*. After surface-exposed antigen identification by the reverse vaccinology approach, mass spectrometry analyses of 2DE maps of protein extracts were performed to confirm the presence of the FACS-positive antigens in the chlamydial cell. This approach allowed the identification of 28 surface-exposed proteins suitable as vaccine candidates [17]. Additionally, immunoproteomics approaches such as serological proteome analysis (SERPA), which is a combination of proteomics with serological analysis, have been used for identifying potential vaccine candidates. For example, this approach was used by Vytvytska and collaborators in the case of *Staphylococcus aureus*, by the resolution of the surface proteins by 2DE and the subsequent electrotransfer onto a membrane that was blotted with different serum pools from humans with infection. Spots corresponding to immunoreactive proteins were analyzed by MALDI-TOF MS. Fifteen novel and already known vaccine candidates were identified [41].

The increase of genome sequence data has also led to the development of structural genomics, a high-throughput application of the traditional structural biology, on the basis of experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy or molecular electron microscopy [42]. Structural genomics projects are likely to solve and provide high-quality 3D structures for the macromolecules encoded by complete genomes and to furnish a library of protein structures. The ultimate goal is to have a complete view of protein folds that may help in assigning function for hypothetical proteins, with the prediction of the protein structure by computational approaches such as homology modeling, founded on similarities with known protein structures for constructing atomic-resolution models from amino acid sequences.

A complete understanding of protein structure and function and the study of functional complexes between a given macromolecule and its effectors in the host will facilitate the rational design of drugs and vaccines. Structural genomics was used in the past two decades in the rational design of important chemotherapeutics, such as HIV [43] or influenza drugs [44], on the basis of the analysis of the complex structures from the target protein and the drug molecule. A focused structural genomics approach to determine the atomic-resolution crystal structures of virulence factors from high priority pathogens is used in the development of vaccines (structural vaccinology). Studies on the structural and antigenic characterization of the virus envelope of HIV provide some examples of the application of structural vaccinology [45]. To be effective, a vaccine must induce a strong protective immune response from B- and T-cells. Because antibodies, by specific recognition of antigen epitopes, are an effective line of defence in preventing infectious diseases, understanding the antibody/epitope interaction establishes a basis for the rational design of vaccines, leading to the development of epitope-driven vaccines, containing only selected epitopes that have been already described, for example, in the case of cancer [46]. Therefore, mapping of epitopes on the solved structure is a central point in the rational vaccine design. The atomic resolution of the epitope, however, requires a very high degree of sophistication and expertise. *In silico* prediction methods have the capacity to accelerate epitope discovery, and should assist in experimental design and interpretation of data. To this end, an important role is performed by immunoinformatics, which combines bioinformatics approaches with immunology (Box 3) and provides tools that are leading to a new understanding of the host immune response, with a positive impact in vaccine development [47].

BOX 3

The role of immunoinformatics: epitope prediction

Immunoinformatics is an emerging application of bioinformatics techniques that focuses upon the structure, function and interactions of the molecules involved in immunity. One of its principal goals is the *in silico* prediction of immunogenicity at the level of epitope. The set of pathogen epitopes that interface with the host immune system is known as the 'immunome'. Recently developed *in silico* tools and databases can be used to identify, characterize or predict antigen epitopes recognized by T- and B-lymphocytes, cells that play significant roles in infection and protective immunity. In particular, a central role is performed by T-lymphocytes, which activate B-cell growth and differentiation and are effectors of cell-mediated immunity. To carry out this role, they need to recognize peptide fragments of pathogen antigens displayed by major histocompatibility complex (MHC) proteins at the surface of antigen-presenting cells. Methods to model the MHC–peptide interface and predict immunogenic peptides sequences that account for T-cell responses, also known as T-cell epitopes, are essential in this approach to vaccine design. With the availability of entire pathogen genomes, the genome-wide application of immunogenetic approaches (immunogenomics), and immunoinformatics tools for epitope mapping, the discovery of T-cell epitopes has been accelerated, facilitating the development of vaccines [59]. Besides the cellular immune response of T-cells, the adaptive immune response is due to antibody-secreting B lymphocytes that are responsible for the humoral immune response. B-cell epitopes are defined by the discrete surface region of an antigen bound by the variable domain of an antibody. While T-cell epitopes are short linear peptides, B-cell epitopes can be linear, contiguous amino acids or they can be discontinuous (conformational) amino acids, separated within the sequence but brought together in the folded protein. The conformational aspects complicate the problem of B-cell epitope characterization and prediction. Several B-cell epitope prediction programs and analysis tools are available [60].

A global -omics approach to vaccine candidate identification

Starting from the knowledge of the genome sequence of a given pathogen, the combination of all strategies mentioned above

TABLE 2

Reverse vaccinology/functional/structural genomics approaches: features and limitations

Approach	Features	Limitations
Reverse vaccinology		
Classical	Fast	It cannot be used to develop vaccines on the basis of nonprotein-coding antigens, like lipopolysaccharides
	Comprehensive: it can virtually identify all potential antigens in a pathogen's genome, irrespective of their abundance, phase of expression and immunogenicity	It needs animal models, because there is a potential lack of method to measure *in vitro* efficacy
	It could be used against all pathogens, including those that cannot be grown *in vitro*	It lacks of information on gene expression
Pan-genomic	Very exhaustive	It requires the sequences of multiple isolates
	It performs interspecies and intraspecies comparisons	It needs a crucial selection of very representative strains of a given microorganism
	It could be useful to develop universal vaccines	
Functional genomics		
Transcriptomics	Very comprehensive	There is not a direct correlation between mRNA and protein expression level
	It provides indications on semiquantitative data of genes expressed during infection	It does not give information on protein localization and gene expression regulation at the transcriptional level
	It can identify pathogenicity factors	It requires a high number of bacteria
Proteomics	It provides qualitative and quantitative data on protein expression	It could identify only a fraction of all proteins
	It can identify membrane-associated proteins	It requires a large number of bacteria cells It is time-consuming and expensive
Structural genomics		
	It can provide insights into protein structure, create comparative models of the most similar proteins and assign a previously unknown molecular function to a protein, providing the opportunity to recognize homologies undetectable by sequence comparison	It is practically limited to comparative modeling for evolutionarily related proteins, with consequent problems for accurate protein model in case of low sequence similarity (less than 30%)
	It can provide a complete understanding of molecular interactions	It needs the implementing of *de novo* protein structure prediction for unique folds determination in the case of sequences that are divergent from those already in the Protein Data Bank
	It can help the rational design of target epitopes to be used as vaccine candidates and increase the understanding of immune recognition mechanisms	Structural genomics efforts often study individual protein domains rather than whole protein or complexes

can offer the highest chances of identifying the best genes/proteins of potential interest for vaccine development in a very brief period of time. The key to the development of effective vaccines is the identification of the pathogen components, often represented by protein antigens, eliciting protective immune responses. Genome sequencing and annotation provides the list of total proteins and their predicted function and localization. The *in silico* analysis represents the first important strategy for the selection of secreted and/or membrane-associated proteins that have a high probability of coming into contact with the human immune system. A second antigen selection criterion is given by functional genomics, which provides qualitative, semi-quantitative or quantitative data on protein expression and allows the comparison of gene expression under different growth/environmental conditions. A further support is furnished by structural genomics, in particular in the rational design of target epitopes to be used as vaccine candidates and in the understanding of immune recognition mechanisms. Each of these approaches can show some advantages and limitations (Table 2). The potential synergies offered by the combination of these genomics technologies, coupled to high-throughput protein expression and purification and appropriate immunogenicity assays, are very powerful and represent a concrete and efficient approach to vaccine development.

The future: synthetic genomics

Synthetic genomics is a new discipline related to the generation of organisms artificially using genetic material. It involves the design and assembly of genes, gene pathways, chromosomes and even whole genomes by the combination of methods for the chemical synthesis of DNA with computational techniques. The goal of synthetic genomics is to make extensive changes to the DNA of a chromosome, assemble it and insert it into an organism to obtain new genomes able to code for new types of cells with desired properties.

The era of synthetic genomics has officially begun and very recently the group of J. Craig Venter, in a work published by Gibson *et al.* [48], described a multistage process to construct the complete genome of *Mycoplasma genitalium*. This first construction of a synthetic genome encoding a living self-reproducing organism provides many potential positive commercial applications. In vaccine development, the chemical synthesis of genetic material could involve the creation of new proteins, the reduction in cost for protein engineering and structural analysis or the

possibility of rapidly generating recombinant vaccines against emerging microbial diseases (Fig. 2).

Conclusions

Vaccines still represent one of the most cost-effective interventions for preventing infectious diseases. Genomics provides an opportunity to vaccine development in particular in the case of pathogens for which the traditional approaches have failed. The genome approach, supported by advances in bioinformatics and high-throughput technologies arising from genomics studies, in which multiple antigens are screened simultaneously, is very powerful and implies that any future vaccine discovery project would strongly benefit from taking genome information into account, as confirmed by the promising results offered by reverse vaccinology. Moreover, comparative genomics is providing new insights into pathogen evolution and epidemiology, virulence mechanisms and host range specificity. In addition, the integration of pathogen and host genomics is likely to revolutionize the approach of developing safe and effective vaccines. Considering the evolution offered by structural genomics and by synthetic genomics, genes now have the potential to be the most important tools for the design of future vaccines.

Acknowledgements

We are grateful to Isabel Delany and Maurizio Comanducci for critical reading, to Giorgio Corsi for artwork and to Catherine Mallia for manuscript editing.

References

1 Jones, K.E. *et al.* (2008) Global trends in emerging infectious diseases. *Nature* 451, 990–993
2 Rappuoli, R. (2004) From Pasteur to genomics: progress and challenges in infectious diseases. *Nat. Med.* 10, 1177–1185
3 Serruto, D. and Rappuoli, R. (2006) Post-genomic vaccine development. *FEBS Lett.* 580, 2985–2992
4 Moylett, E.H. and Hanson, I.C. (2003) Immunization. *J. Allergy Clin. Immunol.* 111 (2 Suppl.), S754–765
5 Pizza, M. *et al.* (1988) Subunit S1 of pertussis toxin: mapping of the regions essential for ADP-ribosyltransferase activity. *Proc. Natl. Acad. Sci. U. S. A.* 85, 7521–7525
6 Fraser, C.M. and Rappuoli, R. (2005) Application of microbial genomic science to advanced therapeutics. *Annu. Rev. Med.* 56, 459–474
7 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512
8 Kaushik, D.K. and Sehgal, D. (2008) Developing antibacterial vaccines in genomics and proteomics era. *Scand. J. Immunol.* 67, 544–552
9 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
10 Hocquette, J.F. (2005) Where are we in genomics? *J. Physiol. Pharmacol.* 56 (Suppl. 3), 37–70
11 Rappuoli, R. (2001) Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 19, 2688–2691
12 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820
13 Tettelin, H. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287, 1809–1815
14 Diaz Romero, J. and Outschoorn, I.M. (1994) Current status of meningococcal group B vaccine candidates: capsular or noncapsular? *Clin. Microbiol. Rev.* 7, 559–575
15 Giuliani, M.M. *et al.* (2006) A universal vaccine for serogroup B meningococcus. *Proc. Natl. Acad. Sci. U. S. A.* 103, 10834–10839
16 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
17 Montigiani, S. *et al.* (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infect. Immun.* 70, 368–379
18 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
19 Ross, B.C. *et al.* (2001) Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 19, 4135–4142
20 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
21 Betts, J.C. (2002) Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis. *IUBMB Life* 53, 239–242
22 Muzzi, A. *et al.* (2007) The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discov. Today* 12, 429–439
23 Sarbah, S.A. and Younossi, Z.M. (2000) Hepatitis C: an update on the silent epidemic. *J. Clin. Gastroenterol.* 30, 125–143
24 Rota, P.A. *et al.* (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300, 1394–1399
25 Rappuoli, R. and Covacci, A. (2003) Reverse vaccinology and genomics. *Science* 302, 602
26 Yang, Z.Y. *et al.* (2004) A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice. *Nature* 428, 561–564
27 Fischer, W. *et al.* (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* 13, 100–106
28 Mora, M. *et al.* (2006) Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Curr. Opin. Microbiol.* 9, 532–536
29 Vernikos, G. (2008) Overtake in reverse gear. *Nat. Rev. Microbiol.* 6, 334–335
30 Binnewies, T.T. *et al.* (2006) Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics* 6, 165–185
31 Dorrell, N. *et al.* (2001) Whole genome comparison of Campylobacter jejuni human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* 11, 1706–1715
32 Fukiya, S. *et al.* (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* 186, 3911–3921
33 Zhang, R. and Zhang, C.T. (2006) The impact of comparative genomics on infectious disease research. *Microbes Infect.* 8, 1613–1622
34 Gay, C.G. *et al.* (2007) Genomics and vaccine development. *Rev. Sci. Technol.* 26, 49–67
35 Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955
36 Maione, D. *et al.* (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* 309, 148–150
37 Grifantini, R. *et al.* (2002) Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. *Nat. Biotechnol.* 20, 914–921
38 Dietrich, G. *et al.* (2003) Transcriptome analysis of *Neisseria meningitidis* during infection. *J. Bacteriol.* 185, 155–164
39 Cullen, P.A. *et al.* (2005) Surfaceome of *Leptospira* spp.. *Infect. Immun.* 73, 4853–4863
40 Rodriguez-Ortega, M.J. *et al.* (2006) Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nat. Biotechnol.* 24, 191–197
41 Vytvytska, O. *et al.* (2002) Identification of vaccine candidate antigens of *Staphylococcus aureus* by serological proteome analysis. *Proteomics* 2, 580–590
42 Chandonia, J.M. and Brenner, S.E. (2006) The impact of structural genomics: expectations and outcomes. *Science* 311, 347–351
43 Costin, J.M. *et al.* (2007) Viroporin potential of the lentivirus lytic peptide (LLP) domains of the HIV-1 gp41 protein. *Virol. J.* 4, 123
44 Yin, C. *et al.* (2007) Conserved surface features form the double-stranded RNA binding site of non-structural protein 1 (NS1) from influenza A and B viruses. *J. Biol. Chem.* 282, 20584–20592
45 Douek, D.C. *et al.* (2006) The rational design of an AIDS vaccine. *Cell* 124, 677–681
46 Mateo, L. *et al.* (1999) An HLA-A2 polyepitope vaccine for melanoma immunotherapy. *J. Immunol.* 163, 4058–4063
47 De Groot, A.S. and Rappuoli, R. (2004) Genome-derived vaccines. *Expert Rev. Vaccines* 3, 59–76
48 Gibson, D.G. *et al.* (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* 319, 1215–1220

49 Bernardini, G. *et al.* (2007) Postgenomics of *Neisseria meningitidis* for vaccines development. *Expert Rev. Proteomics* 4, 667–677
50 Hiller, N.L. *et al.* (2007) Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J. Bacteriol.* 189, 8186–8195
51 Morsczeck, C. *et al.* (2008) *Streptococcus pneumoniae*: proteomics of surface proteins for vaccine development. *Clin. Microbiol. Infect.* 14, 74–81
52 Read, T.D. *et al.* (2003) The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* 423, 81–86
53 Bergman, N.H. *et al.* (2007) Transcriptional profiling of *Bacillus anthracis* during infection of host macrophages. *Infect. Immun.* 75, 3434–3444
54 Chitlaru, T. *et al.* (2007) Identification of in vivo-expressed immunogenic proteins by serological proteome analysis of the *Bacillus anthracis* secretome. *Infect. Immun.* 75, 2841–2852
55 Jaing, C. *et al.* (2008) A functional gene array for detection of bacterial virulence elements. *PLoS ONE* 3, e2163
56 Utt, M. *et al.* (2002) Identification of novel immunogenic proteins of *Helicobacter pylori* by proteome technology. *J. Immunol. Methods* 259, 1–10
57 Beres, S.B. *et al.* (2004) Genome-wide molecular dissection of serotype M3 group A Streptococcus strains causing two epidemics of invasive infections. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11833–11838
58 Maiden, M.C. *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
59 De Groot, A.S. (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov. Today* 11, 203–209
60 Davies, M.N. and Flower, D.R. (2007) Harnessing bioinformatics to discover new vaccines. *Drug Discov. Today* 12, 389–395