Drugs of Today 2009, 45(Suppl. B): 173-187

CCC: 1699-3993/2009

BIOINFORMATIC APPROACHES TO THE STUDY OF CHLAMYDIAL DISEASES

X. Rabasseda[1], S.A. Morré[2] and S. Ouburg[2]

[1]Medical Department, Life Sciences, Healthcare & Science, Thomson Reuters, Barcelona, Spain; [2]Laboratory of Immunogenetics, Department of Pathology, VU University Medical Center, Amsterdam, the Netherlands

CONTENTS

Correspondence: Xavier Rabasseda, MD, Medical Department, Life Sciences, Healthcare & Science, Thomson Reuters, Provença 388, 08025 Barcelona, Spain. E-mail: xavier.rabasseda@thomsonreuters.com

SUMMARY

The combination of bioinformatics, genomics and proteomics is a rapidly expanding research field that has now become essential for the investigation of biological, cellular and molecular phenomena, both in terms of established facts and in prediction models based on simulations, and therefore can greatly contribute to the development of potential new therapeutic approaches. This article provides numerous, well-known examples illustrating the importance of bioinformatics in modern biological and medical research in the field of infectious diseases, specifically Chlamydia trachomatis *infection. Specialized chlamydia databases and analysis programs and the information they contain provide the opportunity for understanding and identifying key processes and phenomena associated with the unique infection cycle of chlamydia, and are paving the way towards the development of novel diagnostic and therapeutic tools. Bioinformatics may provide the clues necessary for a true breakthrough in the fight against chlamydial diseases: the development of a vaccine.*

INTRODUCTION

"Bioinformatics" is the generic name for procedures jointly using applied mathematics, computer sciences, statistics, artificial intelligence, chemistry, biochemistry and related biochemical sciences to understand complex molecular and cellular biological phenomena. To do this, bioinformatics uses computer-run mathematical algorithms to study gene identification and sequencing, genomic assembly, prediction or determination of protein structures, interactions between proteins, and between proteins and other cell components. These are fields that are rapidly and significantly advancing with the use of high-throughput methods that require complex, multidimensional algorithms to assess the experimental results and generate new hypotheses and predictions to be tested and then reassessed with the use of bioinformatics. Strictly speaking, "bioinformatics" is the process by which computers are used to create algorithms for statistical analysis of biological results derived from testing hypotheses. The use of powerful computers to analyze experimental or simulated theoretical data has resulted in a better understanding of very complex biological processes. Bioinformatics has also been applied to test previously formulated hypotheses and to confirm theories on how molecules interact to create a wide range of biological phenomena. The U.S. National Institutes of Health (NIH) clearly distinguish between "bioinformatics" and "computer biology", the latter being reserved for hypothesis validation; however, in practice, they are very closely related and are based on the same concept of using highly complex computer-generated and computer-run algorithms for biological and medical research. Similarly, bioinformatics and computer biology have also been used in medical research for molecular design and testing of new therapies, thereby widening the classical in vitro and in vivo tests into a third in silico approach to research, which has been applied to many research fields and has had a large impact on the study of host–pathogen interactions (1).

Bioinformatics, computer biology and in silico biomedical research all use mathematical tools to analyze and condense information from biological data to generate and validate conclusions. These complex mathematical calculations are applied when studying very diverse processes: from genomic sequencing based on assembly of partially identified genomic sequences, to prediction of expression profiles based on data from restricted observations. To achieve this, bioinformatics uses artificial learning that allows computers to recognize patterns and accumulate experience that is used by the computer processes themselves to correct and redirect very complex analysis that reaches beyond the scope of the starting algorithms. Given the enormous volume of data currently generated using modern molecular technology, it is understandable that the use of not only preestablished analytical algorithms but of algorithms that are self-modified to better analyze all the data, can help in achieving more sound conclusions and developing better research approaches to obtain even more data on a specific phenomenon. This has been one of the bases for the study of gene expression in cancer and its association with prognosis and response to treatment; the same approach has been used in many other fields within molecular and cellular biology, including many related to pathogens and infectious diseases (1). Analysis of biomarkers, prediction of infectivity based on specific characteristics and markers contained in the pathogen and host, and development of selective and ultraselective therapies are just a few examples of how bioinformatics can help achieve a better understanding of the interplay between genetic predisposition factors, chronic inflammatory conditions, environmental factors, aging, lifestyle and many other individual patterns; of how these impact on the acquisition, evolution, prognosis, response to treatment (including host response and pathogen resistance issues) and outcome of infectious

diseases; and how all this information can contribute to the development of better treatments.

How bioinformatics can assist in understanding, treating and preventing infectious diseases will be discussed in further detail in this review using chlamydiae as the basic model. The accumulation of data related to genes, epigenetic phenomena, biomarkers, protein expression, enzymes, channels, receptors and all molecules that are involved in infectivity would be impossible to analyze without the use of bioinformatic approaches, just as bioinformatics, and especially semiautomatic bioinformatics and artificial intelligence, is currently the basis for all genomic studies and their application to gene sequencing, molecular analysis and susceptibility, diagnosis, prognostic and treatment of human diseases, leading to the concept of individualized disease prevention and therapy.

Before describing in detail the bioinformatics applied to infectious diseases, we will briefly discuss the main uses of bioinformatic techniques.

Sequence analysis

Intensive genomic research has resulted in sequencing the DNA of many organisms, and all such data have been indexed in databases and are the basis for studies aimed at identifying and comparing genes coding for peptides and regulatory sequences from different species, using molecular systematic rather than the classical phylogenetic trees to establish evolutionary patterns. However, the increasing volume of data limited the possibility of manually studying DNA sequences, and bioinformatics is the current standard, with software specifically designed to seek related, nonidentical sequences within the genomes of hundreds of organisms that contain billions of nucleotides, including polymorphic variants and mutations. Bioinformatics is also the approach for reconstructing partially identified gene sequences, analyzing the components of newly described genomes and translating from genome to proteome based on DNA sequences which may contain as yet unidentified portions. A clear example is the Human Genome Project, which has opened a brand-new era in the history of science through the combined use of rapid, effective sequencing techniques, identification of gene and gene–gene relationship patterns and detection of mutations, variants and polymorphisms, none of which would be possible without the use of bioinformatics. However, in infectious disease research, the use of systematic BLAST protein database screening to identify signature proteins unique to chlamydiae is a particularly fitting example. The researchers identified 59 proteins specific to *Chlamydiales*, 79 specific to *Chlamydiaceae*, 20 each specific for *Chlamydia* and *Chlamydophila*, and 445 open-reading frames specific to *Protochlamydia*, along with a number of possible gene losses and lateral transfers (2). In addition to the phylogenetic relevance of these observations, the results of these studies also have relevant therapeutic implications. Indeed, the high degree of conservation of many of these proteins among chlamydiae may suggest novel biomarkers to help in the diagnosis and monitoring of infections caused by these pathogens.

To distinguish *Chlamydia trachomatis* strains, called serovars, usually serovar typing is used to identify 19 different serovars. Besides this one gene analysis, multigene sequence techniques have also been described, of which the multilocus sequence typing (MLST) is the most known. Three different MLST systems have been described for chlamydiae (3-5). The MLST systems developed by Dean et al. (4) and Pannekoek et al. (3) are more appropriate for evolutionary studies, since they are based on selected housekeeping genes. In general, these two systems have also generated quite comparable results. The programs used to make the phylogenetic and evolutionary trees described in these articles included eBURST and MEGA4. The third MLST system (5) is more suitable for epidemiologic, network and transmission studies and the data are available at the data warehouse at http://mlstdb.bmc.uu.se/. This system using PHYLIP analyses software has a threefold higher discriminating capacity than conventional *ompA* sequencing, the gene which is used to assess the 19 different *C. trachomatis* serovars.

Gene identification

Bioinformatic software has been used for a number of years as the main tool for genetic annotation, as it is the process for identifying genes and other functional sequences within the DNA. It started with the process of decodification of the *Haemophilus influenzae* genome, which simply stated, was a process of establishing alignments between DNA sequences coding for proteins and the protein sequences themselves. Modern bioinformatic approaches follow the same concept, but have been largely improved through the processes of self-learning included within the algorithms themselves. These improvements have dramatically increased the capability of sequencing and annotating genes such that they

are able to, for instance, relate genes and diseases. This is the case of the GEO database from the NIH which was the basis for a cross-analysis of gene banks and PubMed-indexed genomic studies. Expansion of this approach to the Human Genome Project could result in identifying gene-related and gene-modulated diseases, but this requires very complex bioinformatics and biostatistical tools to identify the relationships between genes and diseases and initiate genomic prognostic studies. A good example of how bioinformatics can help identify genes and their products was the use of such an approach for determining the product of the CP0718 gene in *Chlamydia pneumoniae*, which was found to be a bioinformatically characterized hyperthermostable manganese-containing superoxide dismutase, and comparing it to a similar enzyme from *Propionibacterium shermanii* regarding the role of manganese ion on hyperthermostability. ClustalW software was used for sequence alignment, while Insight II was used to predict secondary structures (6). Another example is the characterization and functional validation of σ^{28} RNA polymerase-regulated genes in the *C. trachomatis* genome through a probability weight matrix bioinformatic method based on known gene promoters from other bacteria. Five active genes were identified and validated in in vitro transcription assays in addition to the only one previously known gene, which is only expressed late in the pathogen development cycle. Characterization of these genes, also following computer-assisted bioinformatic protocols, suggested that at least one of them, *dnaK*, was related to responses to cellular stress and, consequently, was susceptible to therapeutic intervention to help fight chlamydial infections (7).

Evolutionary biology

As mentioned above, bioinformatic techniques have been effectively used to track the origin and evolution of species and their changes over time by monitoring the parallel course of many organisms through changes in their DNA sequences, rather than physical or physiologic taxonomic observations. Studies based on genomes offer better insight into evolutionary phenomena such as genetic duplications, lateral transfer and specialization, resulting in population models useful for predicting success/failure and drawing vital trees including all coexisting species. To illustrate this, a genomic research program using publicly available protein and gene datasets and phylogenetic analysis with the computer-based neighbor-joining method PHYLIP identified plant-like genes in *Chlamydia* spp. that reflect an ancestral relationship between these important pathogens and cyanobacteria, specifically the chloroplast, suggesting an evolutionary relationship rather than horizontal gene transfer and acquisition. Note that this has a double implication: the evolutionary pattern, as well as the suggestion that *Chlamydia*, which are obligatory intracellular pathogens, do not likely exchange DNA with their hosts (8). Similarly, bioinformatic approaches based on selective constraint and coevolution analyses were used to document adaptive evolutionary changes of the GroEL heat shock protein (HSP) in *Chlamydia* (9).

Measuring biodiversity

If biodiversity can be measured as the total genomic component of a particular ecological niche, which is the sum of all species it contains, biostatistics is the tool for compiling all species, descriptions, distributions, genetic information, status and size of populations into databases and analyzing the needs and interactions between species. Specialized computer software can enable not just an analysis of the information but also the visualization of the results in simulated dynamic populations to establish the genetic health and the risks of extinction, so that DNA sequences from species at risk can be preserved and, perhaps eventually, extinct species can be recovered.

Regulation analysis

The control of each and every specific event that occurs upon stimulation or inhibition of a membrane sensor until the cell responds is grouped under the term "regulation", and only bioinformatic techniques are able to explore all steps involved in these cascading processes. Examples may include all relationships between gene promoters and DNA transcription (when considering gene regulation) or all phenomena that result from a contact between a host cell and an infective pathogen. Concepts such as coexpression and coregulation apply to both examples, and it is only due to the capability of algorithms used in bioinformatics that all these distinct events can be cross-analyzed in the context of signaling pathways, host and signal factors and resulting interplays.

An example of how bioinformatics facilitates complex regulatory analysis is how the infectivity of *Chlamydia* was determined using genetically resistant cells. *Chlamydia* is an intracellular pathogen that requires internalization to cause infection. Lack of knowledge of the processes by which the pathogen binds and enters

host cells for its survival, growth and pathogenesis prompted research using the most advanced computer-assisted methods to characterize the resistant cells. The results of algorithms correlating protein spectra with known proteins from databases showed a defect in the leader sequence of the gene coding for protein disulfide isomerase. Forced expression of the identified gene was found to restore binding and infectivity by *Chlamydia*, thus validating the bioinformatic results and identifying surface-expressed protein disulfide isomerase as key to chlamydial infectivity (10).

Protein expression analysis

Gene expression can be measured as the amount of messenger RNA, which can be estimated, although with a risk for uncertainty and bias. Sophisticated bioinformatic algorithms have been developed to better filter out background noise (denoising) and allow more reliable biological analyses for determining gene and protein expression in tissue samples. As an example of these approaches, the PDQuest 2-D analysis software was successfully applied to the study of abnormal protein expression patterns in *Chlamydia*. Through this integrated proteome-works interaction system, the impact of stimulated growth conditions with interferon (IFN)-γ on the expression of adenylate cyclase, thiol-specific antioxidant, 15-kD cysteine-rich protein, methionine aminopeptidase and a hypothetical Cpn0710 protein was studied. PDQuest 2-D was used to analyze and quantify protein expression patterns in electrophoresis gels of stimulated and nonstimulated samples through an integrated master image that compared profiles with a matchset profile. Grouping of spots (belonging to proteins) by fold increase or decrease was used to determine the differential protein expression pattern of the pathogen during growth, and should form the basis for future research into the involvement of each of the proteins. The protein expression is altered during growth on the infectivity and pathogenicity of the organisms and any novel pathways identified may eventually lead to new therapeutic targets (11).

Mutation analysis of infectious diseases

Mutations have an impact on a number of phenomena related to infectious diseases, from host susceptibility to infection, to pathogen resistance and treatment. A wealth of knowledge on these issues has been derived from simple observations. Using bioinformatics to complement all known facts may lead to automatic algorithm-based systems that are able to identify the most probable cause of infection in a particular patient; the best treatment for a particular infection based on a number of epidemiologic and host demographic characteristics; and the risk of complications associated with an individual pathogen in a specific patient. However, it could also identify novel mutations, polymorphisms and gene variants that could be related to how an individual acquires an infection, reacts to the pathogen and responds to treatment. The increased use of gene expression pattern assessment combined with DNA and oligonucleotide microarray-based test systems has been increasingly used in the context of viral infections, but the scope is also extending to bacterial and fungal pathogens. Bioinformatics is the tool that can be used to simultaneously monitor thousands of sites within the genome, resulting in terabytes of information that contain actual data along with variability and background noise. All of this can only be managed by the most advanced bioinformatic tools (EpiGenChlamydia Consortium).

Protein structure prediction

Prediction of protein structures from their primary structure (amino acid sequence) is a common use of bioinformatics, with the aim of calculating the tridimensional structure within the structures to which they belong. By using homology or other known genes and proteins, comparative algorithms can predict tertiary structures of proteins in their native location and determine which part(s) of a protein are essential for interacting with other cellular or extracellular components.

Comparative genomics

Orthological analysis, or comparative genomics, is used to establish concordance between genes and genomic characteristics in different species in order to draw intergenomic maps showing evolutionary patterns and genome splitting from specific mutations, duplications, gene transfers, inversions, transpositions, deletions or insertions. Thus, a metabolic evolution resource, metaTIGER, for eukaryotes and prokaryotes has recently been developed, with genomic information and sensitive sequence search techniques for predicting, comparing and following the evolutionary pattern of metabolic enzymes across species. It could be described as a complex genomic phylogenetic database that could also help identify lateral gene transfers by comparing and visualizing the metabolic networks of different organisms (12).

Modeling biological systems

Computer simulations of biological phenomena, including polycellular, single-cellular and intracellular components, to analyze and understand complex connections and processes and compare scenarios, commonly known as "artificial life", was only made possible through the use of bioinformatic approaches.

Image sequential analysis

Computer-assisted methods for speedy and automatic image analysis, interpretation and validation can be the only means of processing high amounts of image data as generated by modern technologies, while also improving reliability and objectivity. As another component of bioinformatics, this can help in diagnosing and monitoring diseases as well as in research into new therapies, and can be applied to many diverse areas such as screening, pathology, morphometry, imaging, real-time assessment of metabolic and biological processes, quantification of size and composition, among many others.

BIOINFORMATIC TOOLS IN THE STUDY OF INFECTIOUS DISEASES

Bioinformatic tools can range from single-line software to very complex graphic applications and Web platforms. A very good example is BLAST, an algorithm used for determining similarities between protein or DNA sequences. A comprehensive database from the National Center for Biotechnology Information (NCBI) is available that contains the full genome of many organisms, including human organisms, and offers nucleotide and protein search functions and BLAST (13). This is just an example of the many SOAP (short oligonucleotide alignment program)-based bioinformatic applications that have been developed for online remote server database use. A typical example is BioMoby, which integrates various remote servers for sharing and integrating data that can be used but not modified by users (14). Indeed, bioinformatics may now appear to be an extension of the Internet; however, it is not restricted to remote access and can also function in Intranet environments and on individual computers. The main issue for bioinformatics is the mathematical algorithms devised for the specific aims and purposes, and for the treatment of data.

Recent uses of bioinformatics in *Chlamydia*

Two very comprehensive programs, HHpred and MODELLER, to identify catalytic residues in *C. pneumoniae*-secreted protease CPAF (Chlamydial Protease/proteasome-like Activity Factor) are a fitting example of how bioinformatics has been applied in infectious disease research. *Chlamydia* spp. are obligate Gram-negative intracellular pathogens that cause respiratory and sexually transmitted infections and have potential for non-replicating, persistent infections that can resist eradication by antimicrobial treatments. Understanding the pathogenesis of infection caused by these pathogens has been one of the objectives of using bioinformatic approaches for the characterization of the host–pathogen profiles based on genomics, proteomics and cell biology. Knowledge of the essential interactions between *Chlamydia* and host cells can be the basis for developing innovative therapies against such infections, and indeed genomic, bioinformatic data have already been acquired that better characterize the molecular players in the inclusion interactions of the bacteria with the host cytosol (15). CPAF degrades host proteins, allowing the pathogen to avoid defenses and replicate, but the mechanism of action had long escaped identification. The use of HHpred to compare CPAF with possible structural homologs identified putative catalytic domains similar to the tricorn protease from *Thermoplasma acidophilum*. These were contained in the residues H746m S965 and D1023 of CPAF, and substitution of the fragments that were homologs to *T. acidophilum* protease resulted in loss of the ability of CPAF to degrade substrates, unlike what occurs with the wild-type factor or mutations in other sites, thus validating the theoretical results obtained by bioinformatics. HHpred, which is available on the Internet, follows an alignment strategy to compare test sequences with sequences contained in NCBI databases using a pairwise protocol with hidden Markov models. In the first step, multiple interactions of PSI-BLAST created a query sequence from CPAF against nonredundant sequences from the database, until a hidden Markov model could be generated containing statistical descriptions of the underlying alignments along with secondary structural information. Following calculation of the probability of each of the 20 amino acids and their respective probability of being inserted or deleted at a given position, the software was used to translate insertion/deletion probabilities into position-specific gaps for each of the subsequent known structures within the protein data bank. A library of hidden Markov models with which the model of CPAF could be compared was generated and scored for probabilities until HHpred could output an alignment of a sequence modeled on known sequences. This was

entered into a second program, MODELLER, used to automatically calculate a model of the query sequence containing nonhydrogen atoms according to spatial restraint patterns (16). Independent research programs also using bioinformatics for proteomic characterization of infected cells suggested a role for CPAF in the extensive degradation of cytoskeletal proteins, offering important new information on the pathogenicity and infectivity of *Chlamydia* (17).

Inclusion-associated proteins are also very important elements of the *Chlamydia* proteome. Computational studies were able to reveal 46 chlamydial open reading frames that contain a bi-lobed hydrophobic domain analogously contained by inclusion-associated proteins from other pathogens. Gene characterization and analysis of the resulting proteins from these open reading frames revealed Inc proteins (IncA in *Chlamydia psittaci*, and orthologs in *C. trachomatis* and *C. pneumoniae*) that are incorporated into the inclusion membrane and that resulted in aberrant inclusion profiles when the cells were treated with specific antibodies. Specifically in *C. pneumoniae*, analyses of the published genome led to the identification of a novel gene family that encodes for 11 polypeptides containing hydrophobic domains that are characteristic of proteins belonging to the inclusion membrane but are highly variable among individual pathogens, especially in relation to the length of polycysteine tracts. Thus, as in the case of CPAF, bioinformatics was indeed able to identify important pathogenic components of *Chlamydia* through the use of complex algorithms for sorting genomic sequences and denoising in order to identify homologies with other known sequences (15).

Experimental evidence that the major outer-membrane protein (MOMP) of *C. trachomatis* was a porin was corroborated using secondary structure prediction algorithms to calculate the mean hydrophobicity, predict the position of transmembrane segments and assign outer loops using an artificial neural network trained to predict the topology of outer membrane proteins. The results of this bioinformatic model showed consistency with experimental biochemical and immunologic evidence. The FORESST program, which is a series of algorithms to fit protein structures into possible models, predict secondary structures and determine the degree of homology with proteins of known structures, confirmed homology with the consensus porin model through comparison with 349 known porin structures. The results, or the theoretical demonstration that MOMP is indeed a porin, were likewise applied and confirmed in other chlamydial species (18). Further, independent refining of some of these observations used the XTOF software to analyze MALDI-TOF mass spectrometric data from fragments of the MOMP of *C. trachomatis* serovar F residues involved in disulfide bonds in the surface-exposed regions of the protein to create a two-dimensional model with functional implications (19).

Similar approaches have been applied to identify and characterize chlamydial cytotoxins, and specifically a *C. trachomatis* replication-independent cytotoxin that was found to be highly homologous to the large clostridial toxin B of *Clostridium difficile*. The identification of this cytotoxin was also made possible through the identification of open reading frames and analysis for homology with known sequences contained in large databases using computer programs similar to those described above. This cytotoxin should be the basis for further research into therapies for diseases such as trachoma. It should be noted that this was possible because chlamydial genomes had been sequenced first, which was also only possible through the use of bioinformatics, and this information was entered into databases accessed by bioinformatic tools (20).

Parallel independent studies also in *C. trachomatis* aimed to compare the genome of oculotropic and genitotropic strains and gain insight into the pathogenicity of these pathogens. Sequencing and comparing the genome of both types of strains, which required the use of bioinformatic approaches, revealed 99.6% identity, supporting the conclusion that tryptophan synthase enzyme and toxin contained in the genome mediate the virulence and organotropism of *C. trachomatis*. In addition, variable numbers of repeats of translocated actin-recruiting phosphoprotein were observed, and a correlation was established between serovars causing lymphogranuloma venereum and the number of *N*-terminal repeats. Furthermore, single nucleotide polymorphism (SNP) analysis using the SYFPEITHI algorithm showed a disproportionate number of variants in the polymorphic membrane protein autotransporter gene predicting T-cell epitopes binding to human leukocyte antigen (HLA) class I and II alleles (21). In addition to independent studies leading to the discovery of polymorphic membrane protein I as a CD8-positive T-cell epitope associated with the same pathogen – a research program that also used bioinformatic tools through the design of ranked photocleavable analogs (22) – these findings indicate the

potential of polymorphic membrane proteins as immune targets for future vaccine strategies.

Many more examples of how bioinformatics have helped in the research into the mechanisms by which *Chlamydia* infects and causes diseases could be described. The number increases dramatically when all pathogens, even just bacterial pathogens, are considered. For instance, bioinformatic manipulation of mass spectrometry data led to the elucidation of the structure and function of macrophage infectivity potentiator, a chlamydial lipoprotein contained in the outer membrane of elementary bodies that shares analogy with lipoproteins from other human pathogens, notably murein from *Escherichia coli* (23). Bioinformatics was also the means for in silico prediction of immunogenic antigens and development of ELISA tests for serodiagnosis of *C. trachomatis* infections (24), for immunoproteomic identification and characterization of *C. pneumoniae* antigens associated with persistent infection (25), and also for a proteomic analysis of cells infected by *C. pneumoniae*, which revealed extensive cytoskeletal protein degradation, possibly through the CPAF protease mentioned above (26). Bioinformatics have also been used recently to study the type III secretion system needle protein of *C. trachomatis* (27), and, in a very recent study, to identify the catalytic residues of CPAF and establish correlates with substrates and enzymatic products (16).

Approaches such as those described above can be used to characterize proteins and pathologic processes associated with pathogens in all the examples provided for the field of *Chlamydia* research; however, as suggested in the last example, the same or very similar research programs can be of further value by helping characterize protein candidates for vaccines. Combined use of bioinformatics with proteomics and DNA microarray technology was successfully used to describe and identify vaccine candidates in meningococci through high-throughput cloning and expression of antigens selected by in silico analysis of the genome sequence and transcriptome analysis of infecting bacteria. In *Chlamydia* this approach was applied to elucidate *C. pneumoniae* protein subproteomes and identify vaccine candidates. Computer-predicted surface protein structures that were homologous to known antigens from other bacterial pathogens were identified, cloned and tested for specificity, a process that requires complex computer software framed within the concept of bioinformatics, at least for the process of structure prediction and homology testing. Model testing based on computer prediction conducted through analogy to known data may eventually help develop vaccines requiring an investment of less time and fewer resources in laboratory and clinical research, thus adding *Chlamydia* to the list of "antipoverty" vaccines, the development of which depends largely on bioinformatics (28).

BIOMARKERS FOR INFECTIOUS DISEASES: *CHLAMYDIA* INFECTION

Insight into the biological mechanisms underlying the clinical course of *Chlamydia* infections will help to improve diagnostics and treatment, thus reducing the disease burden. BIOMARKERcenter™ recently launched by Thomson Reuters offers a novel database to support biomarker research and it may be useful to the study of *Chlamydia* and biological mechanisms involved in infection (www.biomarkercenter.com).

The study of both *Chlamydia* and host biomarkers will further the understanding of these biological mechanisms, and is thus an essential step in the reduction of *Chlamydia* infections and their effects worldwide.

Molecular biomarkers

Several *Chlamydia* proteins have been identified as potential biomarkers for *Chlamydia* infection and disease progression. The *omp*1 gene, encoding the MOMP, is well known, as it is also used to subdivide *Chlamydia* strains. Other well-studied proteins include HSP60 and HSP70 *Chlamydia* lipopolysaccharide (LPS). Other researchers have focused on the mechanisms of *Chlamydia* inclusion (including the inclusion proteins IncA and IncB), and factors related to *Chlamydia* infectivity and immune evasion (e.g., CPAF and the macrophage infectivity potentiator [MIP]).

Genetic biomarkers

It is now known that approximately 40% of the clinical variation observed in human *Chlamydia* infections is due to host genetic variation (29), making the study of host genetic variation not only relevant for the understanding of *Chlamydia* pathogenesis, but also for the identification of potential biomarkers. The genetic variations may be used as biomarkers for susceptibility to *Chlamydia*, but also for the risk of development of late complication. Currently, several host genetic variations have been linked to *Chlamydia* infection and pathogenesis. These

can be divided into variation in pathogen recognition receptors (e.g., Toll-like receptor [*TLR*]*2*, *TLR4*, *TLR9* and mannose-binding lectin [*MBL*] polymorphisms have been linked to the development of tubal pathology [30-33]), and immune system regulation (e.g., interleukin [*IL*]-*10*, tumor necrosis factor [*TNF*]-α, and *IFN*-γ polymorphisms have been associated with scarring trachoma [34-36]). Recently, Morré et al. have published a comprehensive overview of host genetic variation in relation to *Chlamydia* infections (37).

It is to be expected that the combination of both *Chlamydia* and host biomarkers will yield the most insight into a person's susceptibility to *Chlamydia* infection, and the subsequent disease progression, which is of potential high value for stratifying women presenting with subfertility without invasive laparoscopy in those with and without tubal pathology.

CHLAMYDIA WEB SITES: FOCUS ON GENETIC VARIATION

An overview of *Chlamydia*-related Web sites is shown in Table I. The best known *Chlamydia* Web site is www.chlamydiae.com, created by Emeritus Professor Michael Ward and receiving more than 65,000 hits per month. Prof. Ward is currently updating the Web site together with Servaas Morré and Sander Ouburg, who will maintain the Web site from the end of 2009 onwards. General sequence information on *Chlamydia* can be found on www.chlamydiaebd.org and in 2010 www.chlamydiadb.org will be operational. Both were presented at the Fourth *Chlamydia* Basic Research Society (CBRS) Meeting in 2009. Genotyping of *C. trachomatis* is limited by highly conserved genomes, but two different MLST systems have been described (3, 4) that are suitable for evolutionary studies. In contrast to the other MLST typing systems, a third

Table I. *Chlamydia-related Web sites and multilocus sequence typing (MLST) systems databases.*

Web site	URL	Information
Chlamydiae.com	www.chlamydiae.com	Resource on *Chlamydia*-related information, including health, research and epidemiology
STD sequence database (STDgen)	http://stdgen.northwestern.edu/	Information on compilation and analysis of molecular sequence information pertaining to sexually transmitted bacteria and viruses. Limited to: *Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Haemophilus ducreyi*, *Mycoplasma genitalium*, *Neisseria gonorrhoeae*, *Streptococcus agalactiae*, *Treponema pallidum* and *Ureaplasma urealyticum*
Chlamydiales MLST databases	http://pubmlst.org/chlamydiales/	MLST database of *Chlamydiales* housekeeping genes. Options to query allelic and sequence information, as well as sequence comparison
Chlamydia trachomatis MLST database	http://mlstdb.bmc.uu.se/	*C. trachomatis*-specific MLST Web site focusing on highly variable genes instead of housekeeping genes. Options to query allelic and sequence information, as well as sequence comparison
*Chlamydiae*db	http://www.chlamydiaedb.org	Allows the exploration of genome sequence data in a comparative manner. Contains automatically derived as well as manually curated annotations and literature links. Tools to comment and update the annotation of proteins at *Chlamydiae*DB are currently under development
Chlamydia research today	http://chlamydia.researchtoday.net/	Comprehensive database of *Chlamydia*-related literature
EpiGenChlamydia	www.EpiGenChlamydia.eu	Information of the EpiGenChlamydia Consortium, which aims to structure transnational research so that comparative genomics and genetic epidemiology can be performed in large numbers of unrelated individuals

MLST system (5) is more suitable for epidemiologic, network and transmission studies with one species such as *C. trachomatis*. Table I shows MLST databases.

PATHWAY ANALYSIS: *CHLAMYDIA* INFECTION

Pathway analyses enable the study of interaction of proteins within specific pathways, and the interaction of different pathways with each other, employing interrelated biological data. Sophisticated data analysis algorithms can search for interactions between genes, proteins, and other biologically relevant molecules, in both annotated databases and online available databases, using "natural language processing" techniques. These interactions can then be graphically represented in a pathway diagram on which experimental data (e.g., expression data and microarray data) can be projected.

Two major companies provide extensive programs for pathway analyses: Ingenuity Systems (Ingenuity Pathway Analysis) (Figs. 1, 2) and Ariadne Genomics (Pathway Studio). Both programs provide similar functionality, enabling the user to interpret high throughput data, to build, expand and analyze pathways, to find relationships between genes, proteins, cell processes and diseases, to build graphically appealing publication-quality path-

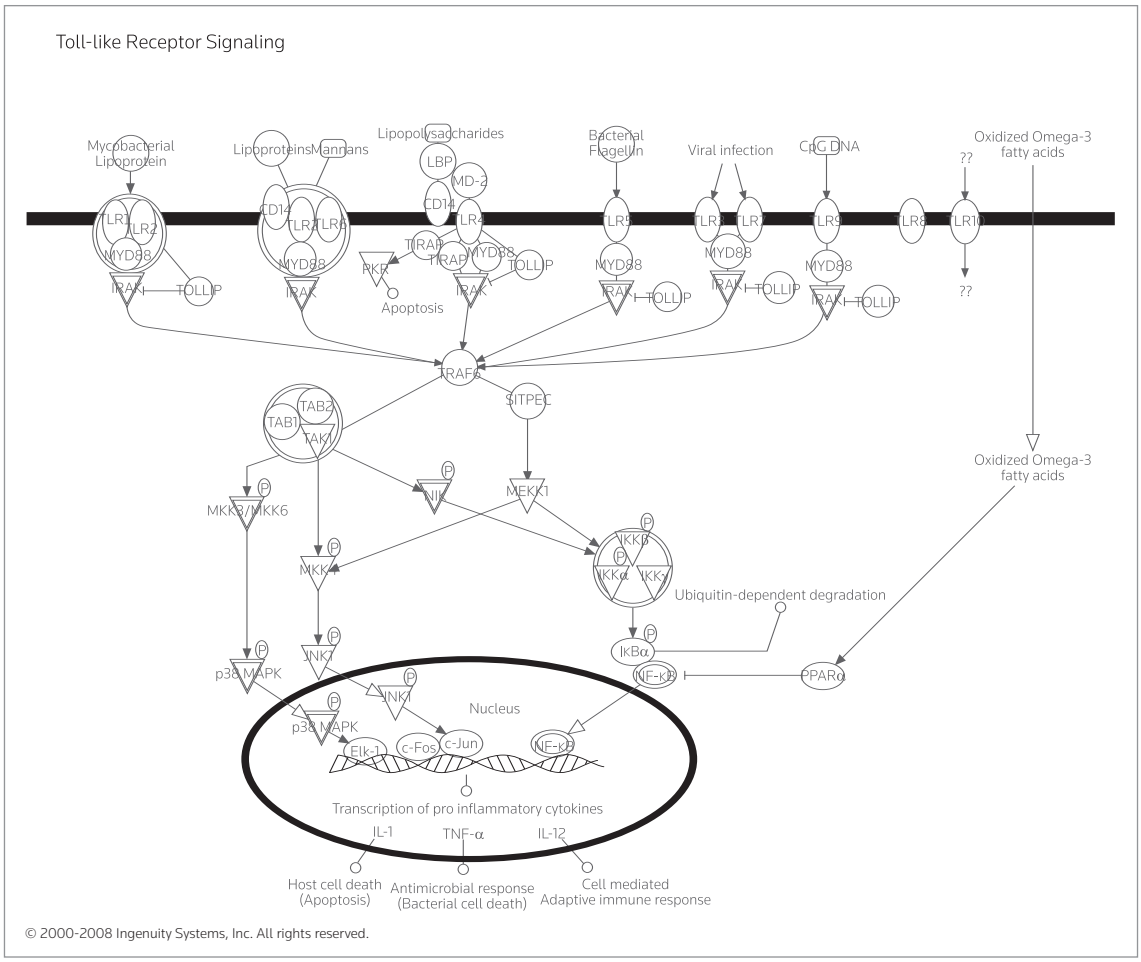

Figure 1. Ingenuity curated pathway (www.ingenuity.com). The curated pathways can be used in conjunction with user-built pathways. The pathway can be extended based on user-defined criteria, and expression and microarray data can be projected on the pathway.

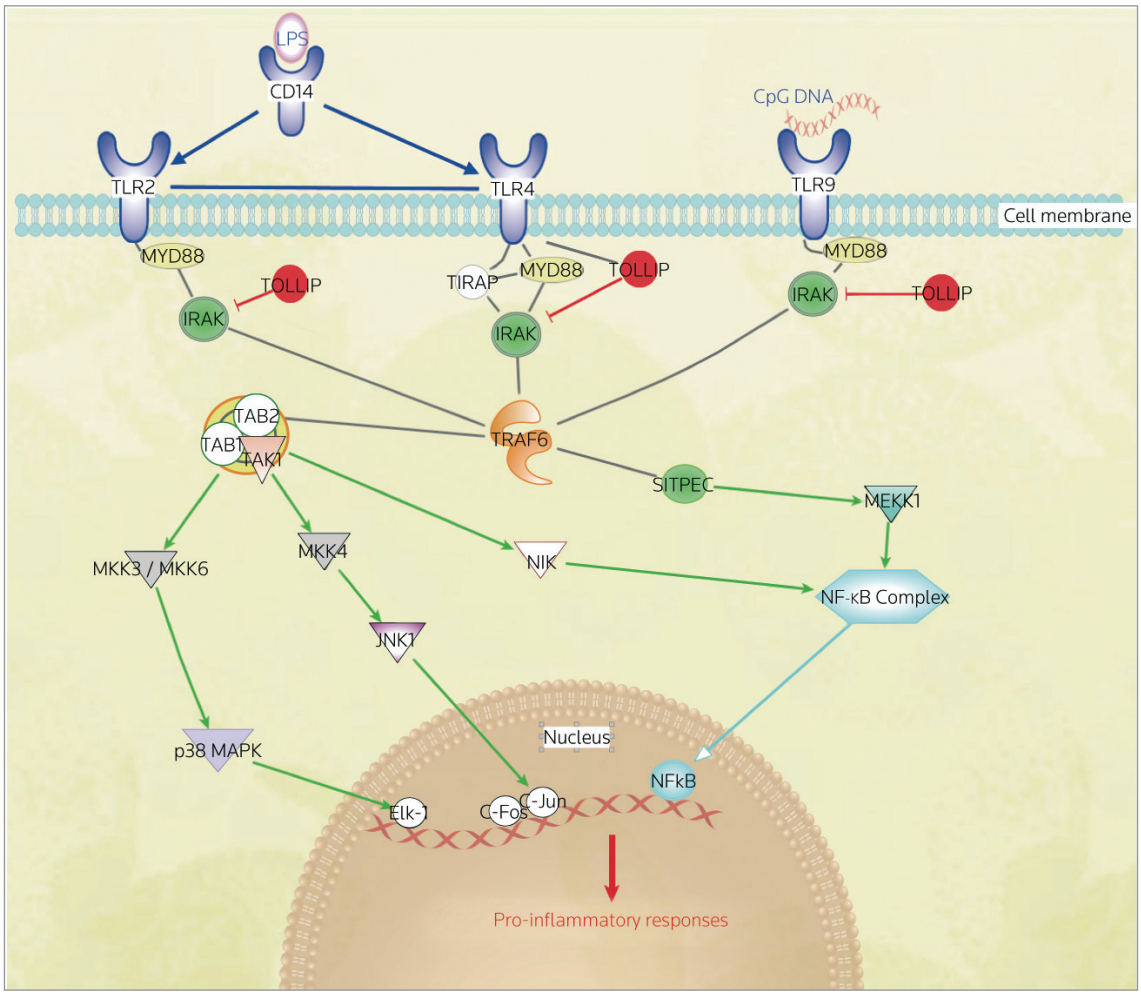

Figure 2. Ingenuity user-built pathway (www.ingenuity.com). Proteins and interactions were limited to user-specified criteria. The pathway can be extended to user-specified criteria and high throughput data may be projected on the pathway, allowing the analysis of which parts of the pathway(s) are up- or downregulated.

ways, and to search different databases (both annotated and publically accessible) for relevant information. Several free accessible online pathway analysis systems are available; however these are limited in scope. Web sites such as KEGG (www.genome.jp/kegg), Pathway-Explorer (https://pathwayexplorer.genome.tugraz.at), and InnateDB (www.innatedb.ca) provide curated static pathways (Fig. 3), while Web sites such as STRING (http://string.embl.de) (Fig. 4) and iHOP (www.ihop-net.org/UniPub/iHOP) (Fig. 5) allow users to interactively build their own pathways. A full list of pathway resources can be found at Pathguide (http://www.pathguide.org).

In general, these pathway resources help researchers gain insight into the diverse pathway interactions. Some resources allow the users to alter the pathway, while the commercial resources provide the most functionality and flexibility, with user-specified criteria to build pathways from curated databases and sophisticated literature search engines. Both Ingenuity Systems and Ariadne Genomics are working on incorporating genetic variation (polymorphisms) into their systems, meaning that in the future pathways may be analyzed by combining literature, expression assays and SNP assays in a fast, graphical representation.

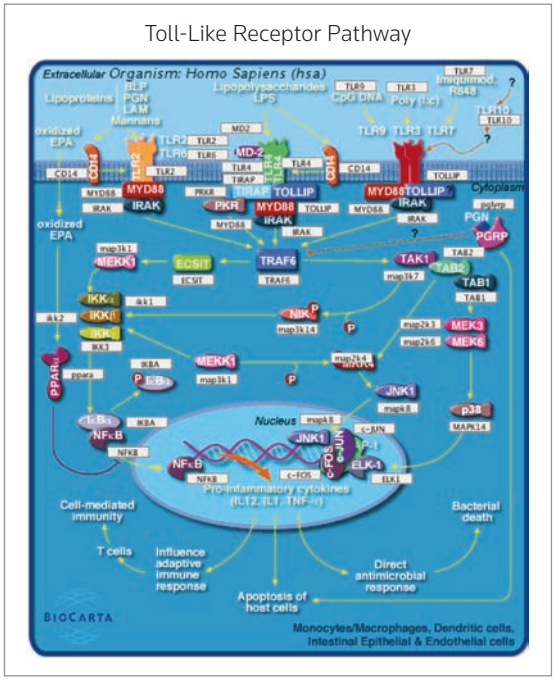

Figure 3. Example of a static (curated) pathway for Toll-like receptor. The layout and represented proteins are fixed and cannot be changed.

THE EPIGENCHLAMYDIA CONSORTIUM

The use of an integrated approach to the study of *C. trachomatis* infection of the female genital tract, resulted in the creation of the ICTI Consortium (Integrated approach to the study of *Chlamydia Trachomatis* Infection) in 2005 (38). The ICTI Consortium is based on a strong interaction and collaboration between basic scientists, clinicians, epidemiologists, and health-care policy makers. In the same year that consortium was established, a Dutch NWO grant obtained to write a Sixth European Framework Programme (FP6) grant, together with 20 groups in Europe, Africa and the USA. This FP6 consortium grant started in July 2007 and ends December 31, 2009 (39).

The EpiGenChlamydia Consortium coordinated by Servaas Morré (www.EpiGenChlamydia.eu) aims to structure transnational research to such a degree that comparative genomics and genetic epidemiology can be performed in large numbers of unrelated individuals. Biobanking and data warehouse building are the most central deliverables of this Coordination Action (CA) in functional genomics research. The study of sequence variation, mainly SNPs, is a technique employed by different consortium partners to gain insight into the differences in the clinical course of infection to identify gene-

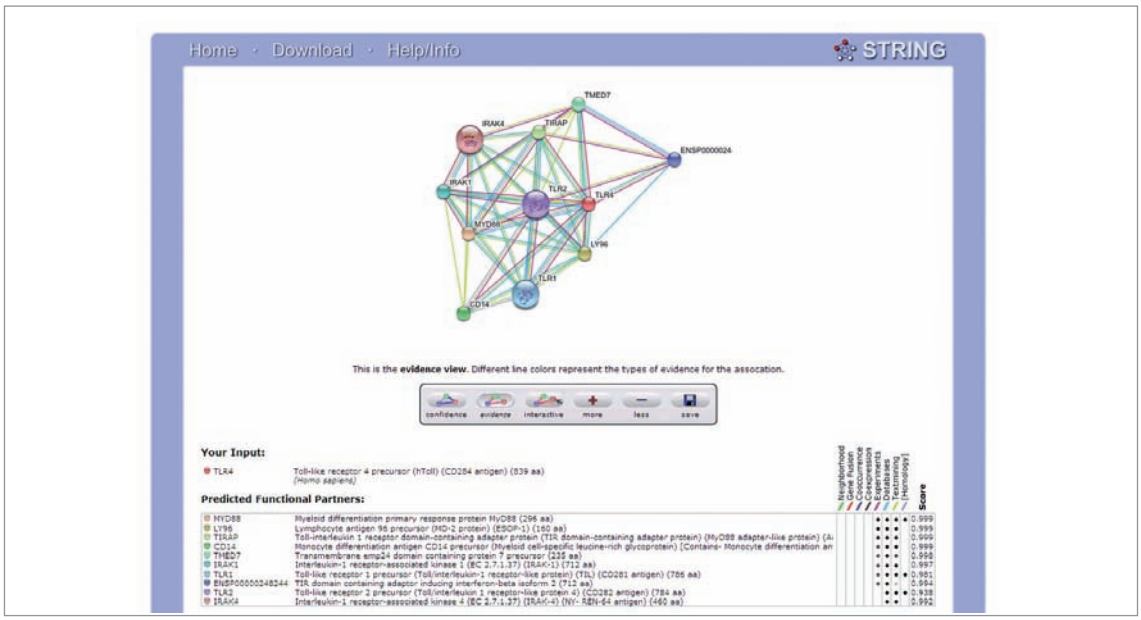

Figure 4. Example of an interactive pathway (STRING). The program uses colored lines to indicate the different interactions between the proteins. STRING is limited in its options for adding proteins or selecting interactions.

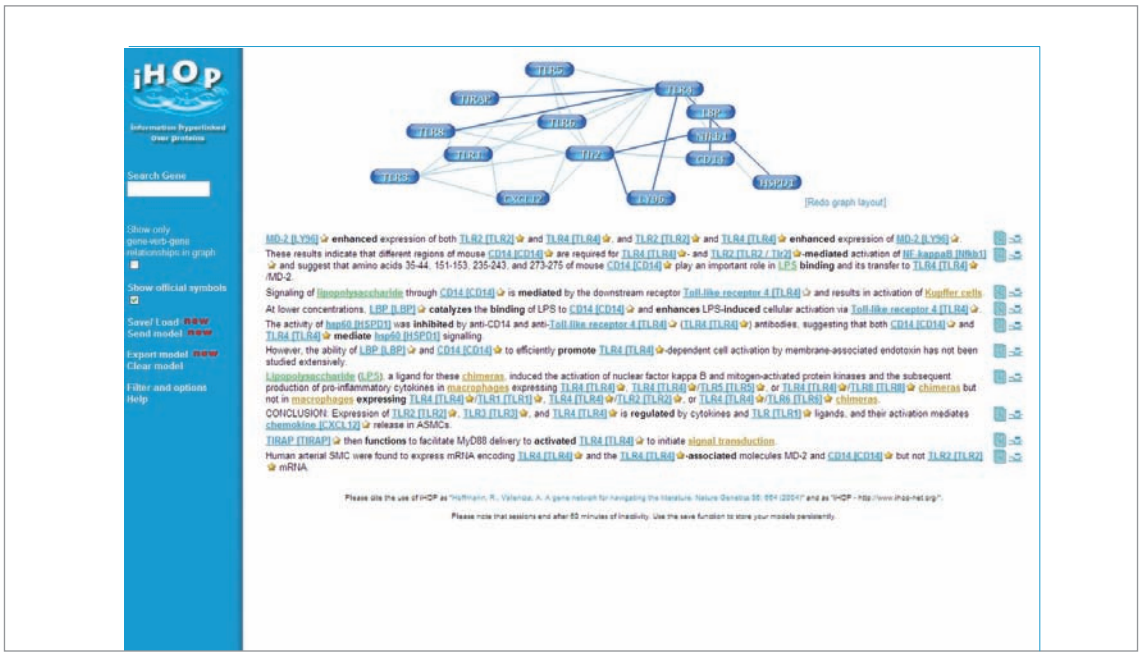

Figure 5. iHOP. With this pathway resource the user can select which interactions to add to the pathway; however the user has little influence on the layout. The user has to read the different interactions and for each decide which to add.

tic markers for disease susceptibility and progression, in part using the pathway programs described above.

DISCUSSION

Combined with bioinformatics, genomics and proteomics are rapidly expanding research fields that are already of enormous assistance when investigating biological, cellular and molecular phenomena, and are expected to be even more so in the future. This is in terms of actual data and also in predictions based on simulations. These fields can therefore greatly contribute to the development of potential new therapeutic approaches. The numerous examples described above illustrate the importance of bioinformatics in modern biological and medical research. As important virulent pathogens, chlamydial infections remain a challenge, largely due to their unique developmental cycle, the obligatory intracellular growth and the lack of a gene engineering system. To facilitate new research into markers, pathophysiology, virulence, infectivity and all phenomena by which *Chlamydia* and host cells interact, the *Chlamydia* Interactive Database (CIDB) has been established and contains genomic and proteomic information for these pathogens. The database contains reverse transcriptase-polymerase chain reaction (RT-PCR), microarray and proteomic data sets that can be used for cross-querying during new research, and is the result of many independent endeavors by different research teams. The database specifically contains quantitative RT-PCR data for 66 genes from two developmental time points under normal and IFN-γ-stimulated growth, microarray gene expression profiles for *C. trachomatis* serovar D, promoter data with a list of genes they regulate and proteomics data for 14 genes (40). Similarly, the Effective T3 program, available at www.chlamydiaedb.org, has been implemented as an in silico program for identifying and predicting type III secreted protein effectors and improving the understanding of the involvement of the type III secretion system in pathogen–host interactions (41).

The databases and programs described above and all the information they contain offer unique opportunities for understanding and identifying key processes and phenomena associated with *Chlamydia* infection, and are helping to pave the way towards novel diagnostic and therapeutic tools. Specifically, as seen in some of the examples dis-

cussed above, bioinformatics may provide the tools necessary for a true breakthrough in the fight against chlamydial diseases: the development of a vaccine.

ACKNOWLEDGMENTS

The aims of this work are in part in line with the European EpiGenChlamydia Consortium which is supported by the European Commission within the Sixth Framework Programme through contract no. LSHG-CT-2007-037637. See ww.EpiGenChlamydia.eu for more details about this Consortium.

DISCLOSURE

The authors have nothing to disclose.

REFERENCES

1. Kirschner, D.E., Linderman, J.J. *Mathematical and computational approaches can complement experimental studies of host-pathogen interactions.* Cell Microbiol 2009, 11: 531-9.
2. Griffiths, E., Ventresca, M.S., Gupta, R.S. *BLAST screening of chlamydial genomes to identify signature proteins that are unique for the Chlamydiales, Chlamydiaceae, Chlamydophila and Chlamydia groups of species.* BMC Genomics 2006, 7: 14.
3. Pannekoek, Y., Morelli, G., Kusecek, B. et al. *Multi locus sequence typing of Chlamydiales: Clonal groupings within the obligate intracellular bacteria Chlamydia trachomatis*. BMC Microbiol 2008, 8: 42.
4. Dean, D., Bruno, W.J., Wan, R. et al. *Predicting phenotype and emerging strains among Chlamydia trachomatis infections.* Emerg Infect Dis 2009, 15(9): 1385-94.
5. Klint, M., Fuxelius, H.H., Goldkuhl, R.R. et al. *High-resolution genotyping of Chlamydia trachomatis strains by multilocus sequence analysis*. J Clin Microbiol 2007, 45(5): 1410-4.
6. Yu, J., Yu, X., Liu, J. *A thermostable manganese-containing superoxide dismutase from pathogen Chlamydia pneumoniae.* FEBS Lett 2004, 562(1-3): 22-6.
7. Yu, H.H., Kibler, D., Tan, M. *In silico prediction and functional validation of sigma28-regulated genes in Chlamydia and Escherichia coli.* J Bacteriol 2006, 188(23): 8206-12.
8. Brinkman, F.S., Blanchard, J.L., Cherkasov, A. et al. *Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast.* Genome Res 2002, 12(8): 1159-67.
9. McNally, D., Fares, M.A. *In silico identification of functional divergence between the multiple groEL gene paralogs in Chlamydiae.* BMC Evol Biol 2007, 7: 81.
10. Conant, C.G., Stephens, R.S. *Chlamydia attachment to mammalian cells requires protein disulfide isomerase.* Cell Microbiol 2007, 9(1), 222-32.
11. Mukhopadhyay, S., Miller, R.D., Summersgill, J.T. *Analysis of altered protein expression patterns of Chlamydia pneumoniae by an integrated proteome-works system.* J Proteome Res 2004, 3(4): 878-83.
12. Whitaker, J.W., Letunic, I., McConkey, G.A., Westhead, D.R. *metaTIGER: A metabolic evolution resource.* Nucleic Acid Res 2009, 37, D531-8.
13. Sayers, E.W., Barrett, T., Benson, D.A. et al. *Database resources of the National Center for Biotechnology Information.* Nucleic Acid Res 2009, 37: D5-15.
14. Néron, B., Ménager, H., Maufrais, C. et al. *Mobyle: A new full web bioinformatics framework*. Bioinformatics 2009, Advance publication.
15. Bannantine, J.P., Griffiths, R.S., Viratyosin, W., Brown, W.J., Rockey, D.D. *A secondary structure motif predictive of protein localization to the chlamydial inclusion membrane.* Cell Microbiol 2000, 2(1), 35-47.
16. Chen, D., Chai, J., Hart, P.J., Zhong, G. *Identifying catalytic residues in CPAF, a Chlamydia-secreted protease.* Arch Biochem Biophys 2009, 485(1): 16-23.
17. Savijoki, K., Alvesalo, J., Vuorela, P., Leinonen, M., Kalkkinen, N. *Proteomic analysis of Chlamydia pneumoniae-infected HL cells reveals extensive degradation of cytoskeletal proteins.* FEMS Immunol Med Microbiol 2008, 54(3): 375-84.
18. Rodríguez-Marañón, M.J., Bush, R.M., Peterson, E.M., Schirmer, T., de la Maza, L.M. *Prediction of the membrane-spanning beta-strands of the major outer membrane protein of Chlamydia.* Protein Sci 2002, 11(7): 1854-61.
19. Wang, Y., Berg, E.A., Feng, X., Shen, L., Smith, T., Costello, C.E., Zhang, Y.X. *Identification of surface-exposed components of MOMP of Chlamydia trachomatis serovar F.* Protein Sci 2006, 15(1): 122-34.
20. Belland, R.J., Scidmore, M.A., Crane, D.D., Hogan, D.M., Whitmire, W., McClarty, G., Caldwell, H.D. *Chlamydia trachomatis cytotoxicity associated with complete and partial cytotoxin genes.* Proc Natl Acad Sci USA 2001, 98(24): 13984-9.
21. Carlson, J.H., Porcella, S.F., McClarty, G., Caldwell, H.D. *Comparative genomic analysis of Chlamydia trachomatis oculotropic and genitotropic strains.* Infect Immun 2005, 73(10): 6407-18.
22. Grimwood, J., Stephens, R.S. *Computational analysis of the polymorphic membrane protein superfamily of Chlamydia trachomatis and Chlamydia pneumoniae.* Microb Comp Genomics 1999, 4(3), 187-201.
23. Neff, L., Daher, S., Muzzin, P., Spenato, U., Gülaçar, F., Gabay, C., Bas, S. *Molecular characterization and subcellular localization of macrophage infectivity potentiator, a*

Chlamydia trachomatis lipoprotein. J Bacteriol 2007, 189(13): 4739-48.

24. Frikha-Gargouri, O., Gdoura, R., Znazen, A., Gargouri, B., Gargouri, J., Rebai, A., Hammami, A. *Evaluation of an in silico predicted specific and immunogenic antigen from the OmcB protein for the serodiagnosis of Chlamydia trachomatis infections.* BMC Microbiol 2008, 8: 217.
25. Bunk, S., Susnea, I., Rupp, J. et al. *Immunoproteomic identification and serological responses to novel Chlamydia pneumoniae antigens that are associated with persistent C. pneumoniae infections.* J Immunol 2008, 180(8): 5490-8.
26. Savijoki, K., Alvesalo, J., Vuorela, P., Leinonen, M., Kalkkinen, N. *Proteomic analysis of Chlamydia pneumoniae-infected HL cells reveals extensive degradation of cytoskeletal proteins.* FEBS Immunol Med Microbiol 2008, 54(3): 375-84.
27. Betts, H.J., Twiggs, L.E., Sal, M.S., Wyrick, P.B., Fields, K.A. *Bioinformatic and biochemical evidence for the identification of the type III secretion system needle protein of Chlamydia trachomatis.* J Bacteriol 2008, 190(5): 1680-90.
28. Grandi, G. *Rational antibacterial vaccine design through genomic technologies.* Int J Parasitol 2003, 33(5-6): 615-20.
29. Bailey, R.L., Natividad-Sancho, A., Fowler, A. Peeling, R.W.W., Mabey, D.C.W., Whittle H.C., Jepson, AP. *Host genetic contribution to the cellular immune response to Chlamydia trachomatis: Heritability estimate from a Gambian twin study*. Drugs Today (Barc) 2009, 45(Suppl. B): 45-50.
30. Karimi, O., Ouburg, S., de Vries, H.J.C., Land, J., Pleijster, J., Peña, A.S., Morré, S.A. *TLR2 haplotypes in the susceptibility to and severity of Chlamydia trachomatis infections in Dutch women.* Drugs Today (Barc) 2009, 45(Suppl. B): 67-74.
31. den Hartog, J.E., Lyons, J.M. Ouburg, S. et al. *TLR4 in Chlamydia trachomatis infections: Knockout mice, STD patients and women with tubal factor subfertility*. Drugs Today (Barc) 2009, 45(Suppl. B): 75-82.
32. Ouburg, S., Lyons, J.M., Land, J.A. et al. *TLR9 KO mice, haplotypes and CpG indexes in the susceptibility to and severity of Chlamydia trachomatis infections*. Drugs Today (Barc) 2009, 45(Suppl. B): 83-93.
33. Sziller, I., Babula, O., Ujházy, A. et al. *Chlamydia trachomatis infection, Fallopian tube damage and a mannose-binding lectin codon 54 gene polymorphism*. Hum Reprod 2007, 22: 1861-5.
34. Natividad, A., Wilson, J., Koch, O. et al. *Risk of trachomatous scarring and trichiasis in Gambians varies with SNP haplotypes at the interferon-gamma and interleukin-10 loci*. Genes Immun 2005, 6: 332-40.
35. Natividad, A., Hanchard, N., Holland, M.J. et al. *Genetic variation at the TNF locus and the risk of severe sequelae of ocular Chlamydia trachomatis infection in Gambians.* Genes Immun 2007, 8: 288-95.
36. Natividad, A., Holland, M.J., Rockett, K.A. et al. *Susceptibility to sequelae of human ocular chlamydial infection associated with allelic variation in IL10 cis-regulation*. Hum Mol Genet 2008, 17: 323-9.
37. Morré, S.A., Karimi, O., Ouburg, S. *Chlamydia trachomatis: Identification of susceptibility markers for ocular and sexually transmitted infection by immunogenetics*. FEMS Immunol Med Microbiol 2009, 55(2): 140-53.
38. Lyons, J.M., Ouburg, S., Morré, S.A. *An integrated approach to Chlamydia trachomatis infection: The ICTI Consortium, an update.* Drugs Today (Barc) 2009, 45(Suppl. B): 15-23.
39. Morré, S.A., Ouburg, S., Peña, A.S., Brand, A. *The EU FP6 EpiGenChlamydia Consortium: Contribution of molecular epidemiology and host–pathogen genomics to understanding chlamydia trachomatis-related disease.* Drugs Today (Barc) 2009, 45(Suppl. B): 7-13.
40. Chen, Y., Timms, P., Chen, Y.P. CIDB: *Chlamydia Interactive Database for cross-querying genomics, transcriptomics and proteomics data*. Biomol Eng 2007, 24(6): 603-8.
41. Arnold, R., Brandmaier, S., Kleine, F. et al. *Sequence-based prediction of type III secreted proteins.* PLoS Pathog 2009, 5(4): e1000376.