Several factors including globalization and sanitation conditions, have been shaping the world landscape of infectious diseases over the years. The World Health Organization has estimated that each year, 1.3 million people die of tuberculosis, 0.2 million die of pertussis and 0.1 million die of syphilis. Diarrheal diseases, many of which are bacterial, are the second leading cause of death in the world (after cardiovascular diseases), killing 2.5 million people annually. This scenario evidences that even today, infectious diseases are a permanent threat for human health over the world. Understanding the biology of the causative agents of these diseases has been a permanent challenge since the beginning of bacteriology. Nowadays, the mechanisms involved in the virulence (defined as the relative capacity of a microbe to cause damage in a host) of pathogenic bacteria are widely studied in clinical bacteriology, but the advent of new technologies has enabled their study from different perspectives. The development of ‘next-generation’ sequencing technologies have significantly reduced the cost of sequencing and have simultaneously yielded an increase in DNA sequencing speed. In this sense, the great majority of organisms whose genomes have been sequenced so far are bacteria, with 1505 complete published and 6037 ongoing projects. Comparative genomics, including comparison at DNA, transcriptome, and proteome levels, have emerged as a key to give a biological sense to all this massive information. Focused on improving the knowledge on pathogenicity determinants two bioinformatic approaches have been used, based on two complementary explanations for bacterial pathogenesis. On one hand, pathogenicity have been related aminoacid substitutions, in shared genes by pathogens and non-pathogens, which lead to modify protein structure, and hence its function.

In this work we exploit the idea that genes causing pathogenicity would be present in pathogenic species but absent in non-pathogenic ones or viceversa. The most spread wide approach to evaluate this is the pairwise comparison between genomes of pathogenic and non-pathogenic bacteria or even multiple comparisons between different strains of the same species. This kind of approaches can give information regarding the presence or absence of genes involved in pathogenicity of a particular species or even a genus, but it is difficult to extrapolate this information to higher taxonomic levels, impeding to draw conclusions about general features that are determining bacterial pathogenicity. For this reason, our motivation was: i) try to identify presence/absence patterns of virulence-related genes which could explain the pathogenic phenotype of bacteria at higher taxonomic levels than species or genus, ii) discuss the biological significance of those genes giving an integrative view of genetic determinants of bacterial pathogenicity, iii) use this information to develop a machine learning model to classify bacterial genomes in human pathogens and non-pathogens and iv) implement this model in a software that can be used to predict pathogenicity in incoming sequenced bacterial genomes.

In this work the idea that bacterial species can be effectively grouped in human pathogens and non-pathogens based on their virulence-genes composition, arises from preliminary results that indicated differential patterns in presence or absence of this kind of genes among both classes (human pathogens and non-pathogens). All finished and annotated genomes of pathogenic and non-pathogenic bacteria were used to perform a presence/absence analysis over 814 groups of orthologous genes belonging to 8 functional categories (toxins, two-component systems, ABC transporters, motility, flagellar assembly, LPS biosynthesis, secretion systems and chemotaxis), in order to determine which ones are strongly related to pathogenicity in different bacterial phyla (Actinobacteria, Alphaproteobacteria, Betaproteobacteria, Bacteroidetes/Chlorobi, Chlamydiae/Verrucomicrobia, Deltaproteobacteria, Epsilonproteobacteria, Firmicutes, Gammaproteobacteria and Spirochaetes). This was accomplished calculating the frequency of genes belonging to each functional category in pathogenic and non-pathogenic species of each phyla. In most cases genes belonging to the 8 functional categories presented clear frequency bias to pathogenic or non-pathogenic species. These findings supported the idea that presence/absence patterns of virulence-related genes are enough informative to discriminate between human pathogenic bacterial species and non-pathogenic species, so this data can be used to construct a classification model based on highly significant biological information.

In this work a machine learning approach based on a linear Support Vector Machine (SVM) model is proposed. Preliminary models were constructed using the whole 814 set of genes, but number of
genes was systematically reduced by means of a feature selection process until a good compromise between the number kept and classification was obtained. The definitive model included the first 120 genes ranked by their significance for classification. In order to avoid the overfitting phenomena, the model was evaluated using a y-randomization test and a 10-fold cross validation scheme. The number of correctly/incorrectly classified genomes in the complete set is 621/30, obtaining a 95.4% of the organisms correctly classified. The model performance is also preserved across the whole taxonomy, ranging from 91.86% in Gammaproteobacteria, up to 100% in Bacteroidetes/Chlorobi and Epsilonproteobacteria. Betaproteobacteria, Actinobacteria and Alphaproteobacteria show a success prediction rate similar or better than the general performance rate. Finally the Firmicutes, the shows an excellent classification level of 96.37%.

All available genomes belonging to Deltaproteobacteria (28) are non-pathogenic and those from Chlamydiae/Verrucomicrobia (14) are all human pathogens. These two phyla were not included in the dataset used to construct the classification model because they do not have genomes comprising both classes, but these were then used to investigate the classification performance of the SVM. This was implemented in order to simulate the situation in which a new sequenced genome, belonging to a different taxonomic group than those used to build the model, is used to challenge the model. The 28 non-pathogenic genomes of Deltaproteobacteria were in fact classified as non-pathogenic, the same performance was observed for Chlamydiae/Verrucomicrobia which were all classified as pathogenic. This evidences the robustness of our model, that keeps a superlative performance with genomes that contain "new" information, originally not included in the dataset used for construction.

The subset of 120 genes selected by the classifier model represents the 8 pathogenicity-related functional categories investigated in this work. Forty genes belong to ABC transporters, 41 are two-component systems and chemotaxis proteins, 11 are toxins, 6 belong to LPS biosynthesis pathway and 22 are flagellar assembly proteins, motility proteins and secretion systems. From each group we selected the most distinctive genes and discussed their biological meaning considering their implications in bacterial pathogenesis. For example, within the 40 ABC transporters genes the majority were related to iron/metallic cations transport and aminoacids transport, highlighting the importance of this kind of transporters for pathogenicity.

A Java software, BacFier, was implemented in order to facilitate the use of the classifier. A simple interface allows the user to upload the genome (finished or unfinished) sequence of the organism of interest. The genome is used as query to perform BLAST against the 120 genes creating a presence/absence vector for the genome. The vector is evaluated with the classifier, and an outcome pathogen/non-pathogen is produced with an associated probability. Moreover, a sensitivity analysis can be automatically performed with the software, this is assessed by selectively "turn off" or "turn on” desired genes in the presence/absence vector and re classifying the result. This might indicate genes that are likely to change the label of the organism, so that one can pay more attention to those. The constructed SVM model can classify bacterial genomes in human pathogens and non pathogens with 95% of average accuracy. To the best of our knowledge, this is the best performance of a statistical model with this purpose reported so far. Moreover, our method can classify bacterial genomes independently of their taxonomic context, in contrast with other similar approaches that only take into account a certain part of bacterial diversity, being useful only to classify specific taxa. Our in silico approach is also supported by results obtained in the wet lab, indicating the importance of selected genes and supporting that bacterial pathogenicity can be deeply explained by the presence or absence of a set of specific genes that code for virulence determinants.

The application of BacFier may be useful for clinical purpose, for example to determine if a new sequenced strain could be pathogenic for humans. Furthermore, it can be used to check if a certain species is harmful or not for humans, which may result particularly interesting for food industry.