30

Whole Genome Annotation: In Silico Analysis

Vasco Azevedo et al.*
Federal University of Minas Gerais (UFMG) and Federal University of Pará (UFPA), Brazil

1. Introduction

After a genome is assembled, the next step is genomic annotation, which can generate data that will allow various types of research of the model organism. Complete DNA sequences of the organism are then mapped in areas pertinent to the research objectives. In this chapter, we explore relevant ongoing research on genes and consider the gene as a basic mapping unit. Gene prediction is the first hurdle we come across to begin the extensive and intensive work demonstrated in first item, which deals with assembly of the genome. Gene prediction can be made with computational techniques for recognizing gene sequences, including stop codons and the initial portions of nucleotide sequences; it involves empirical rules concerning minimum coding sequences (CDS's) and is limited due to overlapping sequences coding forward and reverse.

Finishing gene prediction step by a computer initiates the functional annotation stage. Functional annotation, item 3, can be done initially by computer, using similarity in sequence alignment. However, no software is capable of generating a functional annotation without many false positive results, since conserved protein domains with varied functions make gene sequence alignment difficult. In this case, after automatic annotation, the predicted genes need to be revised manually. In manual curation, item 4, an expert can more accurately locate frameshifts in the DNA strand. Depending on the number of errors found, genomic annotation may be postponed, requiring a return to the previous stage of genome assembly. In manual curation, the principal contributions are usually correction of the start codon position, gene name, gene product and, finally, identification of frameshifts.

When functional annotation is completed, the genome should subsequently be submitted. It occurs after the assembly and annotation steps making the data generated available in public-access databanks. Submission is a pre-requisite for publication in scientific journals. Another advantage of genome publication in public-access sites is that it permits use of various genome analysis tools. For example, searches for genomic plasticity, pangenomic study, exported antigens and evaluation of innate and adaptive immune responses. The pangenome approach, item 5, concepts of species can be used as a filter for targeting candidates for vaccines, diagnostic kits and drug development. For drug development, the

* Vinicius Abreu, Sintia Almeida, Anderson Santos, Siomar Soares, Amjad Ali, Anne Pinto, Aryane Magalhães, Eudes Barbosa, Rommel Ramos, Louise Cerdeira, Adriana Carneiro, Paula Schneider, Artur Silva and Anderson Miyoshi
Federal University of Minas Gerais (UFMG) and Federal University of Pará (UFPA), Brazil

core set of proteins is a more likely source of useful information, for developing both vaccines and diagnostic materials for a unique pangenome set of a species of interest.
Genomic plasticity, item 6, is the dynamic property of genomes, involving DNA gains, losses, and rearrangement; it allows bacteria to adapt to new hosts and environments. There are several mechanisms that can drive these changes, including point mutations, gene conversions, rearrangements (inversion or translocation), deletions and DNA insertions from other organisms (through plasmids, bacteriophages, transposons, insertion elements and genomic islands). Gene acquisition and loss by all these mechanisms influences bacterial lifestyles and physiological versatility. Analyses of HGT regions in silico has become feasible due to the introduction of next–generation sequencing technologies, which allows sequencing of prokaryotic genomes at a faster rate than the earlier Sanger method and at a considerably lower operational cost. Consequently, the number of complete genome sequences available for analysis has grown and continues to grow rapidly.
In post-genomics, study of Reverse Vaccinology (RV), item 7, can provide predictions of the sub cellular locations of an entire predicted proteome. Additionally, these previous annotations, prediction of peptides with high affinity for class I and II MHC proteins is another in silico analysis that increases the probability of selecting antigens that can promote immune responses in organisms infected by a pathogen. The field of research referred to as immunoinformatics, item 8, is giving us the opportunity to analyze antigens with greater selectivity and increase the likelihood of developing a successful vaccine.

2. Gene prediction

The development of modern sequencing technology has resulted in an exponential increase in the number of available genome sequences. To illustrate, in 1997 there were 10 complete genome sequences of bacteria available in the NCBI (Lukashin & Borodovsky, 1998); by 2011, this number had sharply increased to 1,538 http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. This enormous increase in the quantity of available information stimulated the development of tools for gene prediction. The development of these tools is a tremendous challenge, and it is a major contribution of Bioinformatics to the field of genomics.

2.1 Gene prediction strategies

Gene prediction programs can be divided into two categories: an empirical category, which relies on sequence similarity; and ab initio, which uses signal and content sensors. Empirical gene predictors search for similarity in the genome; they predict genes based on homologies with known databases, such as genomic DNA, cDNA, dbEST and proteins. This approach facilitates the identification of well–conserved exons. Ab initio gene finders use sequence information of signal and content sensors. Usually, these programs are based on Hidden Markov Models. Ab initio can be organized into categories based on the number of genome sequences used in gene analysis; it includes single, dual and multiple–genome predictors. Integrated approaches couple the extrinsic methodology of empirical gene–finders and intrinsic ab initio prediction. This technique significantly improves gene prediction protocols (Allen et al., 2004).

2.2 Eukaryotes

The complexity of the challenge faced by Bioinformatics is only completely understood when we look at the complexity of the eukaryotic genome. Within genomes, genes are not

organized in a continuous cluster. Instead, the coding regions (exons) are often widely interspersed with non–coding intervening sequences (introns). Furthermore, in many cases the intronic region is much larger than the exonic region. These low–density coding sequences are evident in the human genome, in which only approximately 3% of the DNA generates proteins. The exon and intron issue can be compared to trying to read a non–continuous article in a journal. In an analogy, one must first identify in which part of the journal (genome) the article (gene) of interest is; then, as the DNA sequences are read, it is necessary to identify which part is informative (exon) and which part contains random information (intron). Also, genes can be altered by alternative splicing, which is a process that generates multiple protein sequences from the same gene sequence template (Schellenberg et al., 2008).

Gene prediction methodology for eukaryotes involves two distinct aspects; the first focuses on the information utilized for gene recognition, basically recognizing signal functions in the DNA strand; the second uses algorithms implemented by prediction programs for accurate prediction of gene structure and organization. The signal function search can be divided into two mechanisms utilized for locating genes. One classifies the content of the DNA strand and the other searches for functional signals in the genome:

(i) The content sensor classifies the DNA regions into coding and non-coding segments (introns, intergenic regions and untranslated regions). This mechanism involves two approaches, intrinsic and extrinsic. The extrinsic approach relies on the assumption that coding regions are evolutionarily more conserved than non–coding regions. Consequently, this methodology employs local alignment tools, like BLAST (Johnson et al., 2008) ; this makes it possible to make comparisons within the genome and between closely-related species. However, one important flaw in this approach involves the necessity of identifying homologies within the database in order to extract results. If none is found, this methodology is unable to determine if a region "codes" for a protein (Sleator, 2010). (ii) The functional sensor approach searches the genome for consensus sequences. Consensus sequences are extracted from multiple alignments of functionally-related documented sequences. The functional signals involve transcription, translation and splice sites. Transcriptional signals includes the CAP signal at the transcriptional start site and the polyadenylation signal located 20 to 30 bp downstream of the coding region. Another important signal to identify is the translation initiation site, although this feature has limitations due to a lack of knowledge concerning initiation sites in eukaryotes (Mathé et al., 2002).

2.3 Prokaryotes

Unlike eukaryotes, the archaeal, bacterial and virus genomes are highly gene-dense. The protein coding regions usually represent more than 90% of the genome. Therefore the accuracy of gene predictors depends primarily on determining which of the six frames contains the real gene. The simplest approach in gene prediction is to look for Open Reading Frames (ORFs). An ORF is a DNA sequence that initiates at a start codon and ends at a stop codon, with no other intervening stop codon. One way to locate genes is to look for ORFs with the mean size of proteins (roughly 900 base pairs) (Allen et al., 2004). Therefore, long ORFs indicate possible genes, although this methodology fails to predict small genes.

The major problem in simply applying this technique is the possibility of ORF overlap in the different DNA strains. This approach must be used along with guidelines to avoid

overlapping, choosing the more likely candidates. Also, numerous false positives are found in non-coding regions. Due to the high gene density, it is difficult to confidently state that any gene predicted in a non–coding region is false. This problem can be minimized by searching for homologies in closely–related organisms. If we do not find a conserved sequence in related species, it is assumed that the prediction (of a gene) is false.
Another problem faced by prediction programs in prokaryotes is how to determine the start codon of a sequence. The first initiation site in a sequence is not necessarily the true one. To solve this problem, programs can employ ribosome binding sites (RBS), which provide a strong signal, indicating the position of the true start site. In conclusion, there is a drop in prediction accuracy in high–GC–content genomes. Rich GC genomes contain fewer stop codons and more spurious ORFs. These false ORFs are often chosen by prediction programs instead of the real ones in the same DNA region. Additionally, the longer ORFs in GC–rich genomes contain more potential start codons, leading to a drop in the accuracy of translation initiation site prediction (Hyatt et al., 2010).

2.4 Tools

2.4.1 Glimmer

The first version of Glimmer (Gene Locator and Interpolated Markov ModelER) was released in 1998 ; the 3.02 version was released in 2006. Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer uses interpolated Markov models (IMMs) to identify coding regions and distinguish them from noncoding DNA. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and it has been used to annotate the complete genomes of over 100 bacterial species from TIGR and other labs. Like other gene prediction programs, Glimmer can be installed and run locally and has a web-based platform (Salzberg et al., 1998). All one needs for online gene prediction of a genome is the fasta version of the sequence and access to the site:
http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi.

2.4.2 FgenesB

FgenesB is a package developed by Softberry Inc. for automatic annotation of bacterial genomes. The gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites. The package includes options to work on sets of sequences, such as scaffolds of bacterial genomes or short sequencing reads extracted from bacterial communities. For community sequence annotation, it includes ABsplit program, which separates archebacterial and eubacterial sequences. FGENESB was used in the first published bacterial community annotation project (Tyson et al., 2004).

2.4.3 Prodigal

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) is a microbial (bacterial and archaeal) gene finding program developed at Oak Ridge National Laboratory and the University of Tennessee. Prodigal focuses specifically on three goals: improved gene structure prediction, improved translation initiation site recognition, and reduced false positives (Hyatt et al., 2010). The source code is freely available under the General Public License and the program can be accessed at http://compbio.ornl.gov/prodigal/.

2.4.4 GeneMarkTM

GeneMark is a public access program for gene prediction in eukaryotes. It is a family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA. GeneMark can operate in two ways: the first one is online, where one can make predictions, using for comparison one of the many available models; the second option is for novel genomes, in this way one can install and run the program locally. The web–based version of GeneMark is available at http://exon.biology.gatech.edu/.

For gene prediction in eukaryotes, GeneMark combines two programs, GeneMark–E* and GeneMark.hmm–E. The GeneMark-E program determines the protein-coding potential of a DNA sequence (within a sliding window) by using species-specific parameters of the Markov models of coding and non-coding regions. This approach allows delineating local variations with coding potential. The GeneMark graph shows details of the protein-coding potential distribution along a sequence, while the GeneMark.hmm-E program predicts genes and intergenic regions in a sequence as a whole. The Hidden Markov models take advantage of the "grammar" of gene organization. The GeneMark.hmm programs identify the most likely parse of the whole DNA sequence into protein coding genes (with possible introns) and intergenic regions.

The statistical model employed in the GeneMark.hmm algorithm is a hidden Markov model. It includes hidden states for initial, internal and terminal exons, introns, intergenic regions and single exon genes. It also includes hidden states for start site (initiation site), stop site (termination site), and donor and acceptor splice sites. The protein-coding states (initial, internal, terminal exons and single exon genes) emit nucleotide sequences modeled by inhomogeneous 3–periodic fifth–order Markov chains. The non-coding states (intron and intergenic regions) emit sequences modeled by homogeneous Markov chains (Lukashin & Borodovsky, 1998).

3. Automated functional annotation

Automated functional annotation of genomes can be quite efficient because it is a computational process based on the alignment of ORF sequences of the organism with sequences from various other organisms (Kislyuk et al., 2010). Public domain databases contain full annotations of thousands of prokaryotic organisms (Benson et al., 2008). Automatic functional annotation takes advantage of knowledge concerning ORFs of homologous organisms, saving considerable time in manual curation (Li et al., 2010). However, care must be taken with fully automated functional annotation, since similarity of sequences can easily incur false positives (Lorenzi et al., 2010). In this section we discuss the advantages and dangers of using fully-automated functional annotation, and we explore some features of tools and services for this purpose.

3.1 Massive sequence alignments must be planned

Algorithms for alignment of biological sequences are intensively used in automatic functional annotation (Aparicio et al., 2006; Meyer et al., 2003). Alignments of ORFs from a newly assembled genome with counterpart ORFs can provide the first hints about the new genome. For an organism with about 2,000 ORFs, analysis of similar sequences against a database of non-redundant (NR) proteins from NCBI can consume several processing hours. For example, assuming that this analysis is done on a computer isolated from the internet, hardware with 24 Gb RAM and eight processors, totaling 24 GHz CPU, this task will consume approximately eight hours of processing time.

Though it is a completely automated computer process, the user has considerable responsibility to set conditions to be utilized in the computation in order to obtain good quality data. These conditions define the quality criteria that best fit the type of organism, for example, the cut–off value for a significant alignment with sequences of other organisms in the NCBI, the number of homologous sequences to be returned as a result and the file format of the output alignment. An additional parameter is required if the sequence search (query) and the targeted search sequences (subject) are in different formats (nucleotides versus amino acids). This parameter determines the most adequate table for translation of codons of the organism in question so that the alignment algorithm of sequences is able to interpret the correspondence between the query sequences and the subject. The number of parameters of an algorithm for aligning sequences can be quite large, justifying training with a heavy workload for optimal utilization. Our objective here is not to explore possible situations, but to alert users that the results of these algorithms can improve these alignments by reading the manual algorithm and consequently adjusting it to a particular situation concerning a query organism or subject. Thus, when beginning a massive alignment sequences project involving a novel genome, with an analysis that will take hours and create high expectations, it is advisable to use not just the basic configurations in these alignment algorithms. It would be useful to take time to weigh and incorporate options that will determine the success or failure of these alignments.

3.2 Knowledge reapplication and time saving

There has been significant growth in the number of DNA sequences available in public databases, because of new genome sequencing technologies, which have made it simpler, more efficient and cheaper to obtain complete genomes (Zhao & Grant, 2010). Fully assembled and annotated genomes of various forms of bacterial life are available to facilitate the processing and inclusion of a newly assembled genome. This wide range of genomes provides the opportunity for new research into large-scale SNPs, DNA methylation and mRNA expression profiles, and resequencing data (Datta et al., 2010). It also allows comparison of annotations from different research groups working with different organisms, some of which may be homologous to a newly-sequenced genome. Just as one can take advantage of knowledge about the function of genes from different organisms, it is also advisable to use the personal knowledge of a researcher on a specific organism in order to accelerate the process of automatic annotation. Based on evidence about a high degree of evolutionary proximity between a newly-assembled genome and a particular organism homolog that already has a fully-assembled and annotated genome, we can choose to use only the annotation of such an organism as a resource for a first automatic annotation.

The problems a researcher would normally encounter when utilizing annotations from various genomes could be resolved by comparison with the annotation of a homologous organism. This situation is common when one examines the pangenome of a species, as it is expected that most of the coding sequences of different strains of bacteria are not very different (Trost et al., 2010). In this case, it appears to be advantageous to identify a small set of target organisms (subject) in a sequence similarity search, with the objective of providing a first genome annotation (query); this may even be a set with only one organism.

3.3 Error propagation: Automated versus manual annotation

It is important to bear in mind that the GenBank is not a fully curated database (Benson et al., 2008); many genomes may have been deposited only as automatic annotations. With

current technology, it is not possible to dispense with manual curation of an automatic annotation, or even experimental evidence concerning gene prediction and annotation based on sequence similarities (Poptsova & Gogarten, 2010). Although it is not normally feasible to initially include experimental verification of gene prediction, it seems reasonable to take advantage of expert human annotation of genomes to help determine the outcome of automatic annotation. Assuming one is working on the pangenome of an organism, such a measure can not only reduce false positives in comparisons of sequence similarities, but also determination of homologous genomes based on a particular annotation. During automatic annotation, a measure that has the potential to minimize error propagation would be allocating different weights for the results of sequence similarity to genes from organisms for which there is evidence of expert manual curation.

3.4 Tools

The following are some tools for automatic annotation of entire genomes, with brief descriptions of their core functionality and instructions on how to use them.

3.4.1 GenDB

One of the reasons that GenDB is included among a select set of tools for automatic annotation of genomes is the fact that it was developed for the web platform (Meyer et al., 2003). Geographically dispersed research groups can benefit from web interfaces using standard tools and a centralized database. Version 2.4 of GenDB has three modules: core, web, and gui. The core module has programs written in perl that allow creation of an annotation project, importation of data in fasta / EMBL format, execution pipeline automatic annotation, display of circular genomic maps, data export and annotation project deletion. Implementation of the programs in the module allows a team of curators to work on the web and edit diverse features of various genes. The gui module has editing features that are more sophisticated than those of the web module, allowing execution of tasks performed by the core module, but with a graphical interface. The GenDB program performs sequence alignments using the program Blast (Altschul et al., 1997) and allows incorporation of predictions of conserved domains of protein families based on InterPro-Scan (Hunter et al., 2009),as well as transmembrane domains based on TMHMM (Krogh et al., 2001), and indications of export to the extracellular medium through SignalP (Bendtsen et al., 2004).

3.4.2 BLAST2GO (B2G)

This tool was designed as an interface for Gene Ontology (GO); additional features have transformed it into a more comprehensive annotation platform (Aparicio et al., 2006). The program menus include various steps initiating annotation, with an automatic alignment of genome sequences against a protein-based non-redundant (NR) NCBI database, through prediction of conserved domains (InterPro–Scan), GO annotation ratings against the enzymatic English Enzymatic Code (EC) and subsequent visualization of molecular interactions in a genome by means of maps in the format of the Kyoto Encyclopedia of Genes and Genomes (KEEGO). Being a visually oriented tool, it has graphical tools to help analyze the vast amount of data generated in the predictions. A user of B2G does not necessarily have to perform all the steps of analysis that are offered, but in order to advance to the next phase of analysis it is imperative that the previous phase be performed

beforehand. Processing of an entire genome with approximately two thousand ORFs can take several days, as the first step is always sequence alignment against the NCBI NR base. Fortunately, B2G is designed to be a modular analysis tool. If a B2G user has computational resources that are more efficient than the shared resources on the public server, the user can perform alignment of sequences on his own hardware to generate an output in HTML format and continue the alignment processes following annotation with B2G. Should the user be dissatisfied with the efficiency of processes of annotating GO terms of the server's common B2G, there is a version of B2G than he can run separately with his superior hardware. The results generated in the offline mode can be uploaded to the online tool to continue the review process using a variety of tools, including statistical comparisons between two genomes. B2G was developed with Sun Java technology, which can be run on any operating system; however, the B2G offline module is designed to run on the Linux platform.

3.4.3 CpDB relational schema: a practical example

This tutorial has approximately 100 steps, including software installation and configuration, edition of files by Linux commands or through interfaces with biological sequence manipulation programs. The tutorial presumes that the programs Artemis, Java (Sun) and Blast version 2.2.20 or previous were locally installed. Many editions of files are made with the "sed" program of Linux, which is included in most Linux versions. All of the steps in this manual can be automated in order to develop an automatic pipeline for annotation, allowing the *Corynebacterium pseudotuberculosis* DataBase (CpDB), a relational database schema and tools for bacterial genomes annotation and pos-genome research, to become another web-based automatic annotation environment. For now, this tutorial has an instructional character, to help make a student aware of the necessities and difficulties involved in the process of automatic annotation of genomes. In order to obtain the tutorial files, type the following command in Linux, Ubuntu 10.10 or later:

svn checkout svn://150.164.37.20/genomes/autoannotation --username=student --password=bioinfo

After finalizing the verification of all of the files, this tutorial continues in the document "Tutorial.pdf", which will be in the folder "autoannotation".

4. Manual curation

Genome annotation is a process that consists of adding analyses and biological interpretations to DNA sequence information. This process can be divided (Stein, 2001), into three main categories: annotation of nucleotides, proteins and processes. Annotation of nucleotides can be done when there is information about the complete genome (or DNA segments) of an organism. It involves looking for the physical location (position on the chromosome) of each part of the sequence and discovering the location of the genes, RNAs, repeat elements, etc. In the annotation of proteins, which is done when there is information about the genes (obtained by genome or cDNA sequencing) of an organism, there is a search for gene function. Besides general predictions about gene and protein function, other information can be found in an annotation, such as biochemical and structural properties of a protein, prediction of operons, gene ontology, evolutionary relationships and metabolic cycles (Stothard & Wishart, 2006). Consequently, functional annotation or manual curation is a fundamental part of the process of assembling and annotating a genome, in which the curator is the person responsible for validating the elements. In manual curation, all of the

predicted genes will be validated and their products named (Stein, 2001). A more detailed description of the gene or gene family product is obtained through similarity analyses using protein data banks that contain well-characterized and conserved proteins (Overbeek et al., 2005).

4.1 Technical terms used in manual annotation

In functional annotation done with Artemis, several fields should be filled out to increase knowledge about particular genome elements. It is necessary to use annotation terms, which involve an official nomenclature developed for this purpose. Some of these terms and respective examples are given below: "LOCUS-TAG" is the term used to identify all of the genome elements, except for the feature "misc". Generally, one uses an abbreviation to identify the particular species, followed by an underline (_) and numbers, for example: Cp1002_0001 (*Corynebacterium pseudotuberculosis*, strain 1002). For tRNAs, the nomenclature is the abbreviation, followed by underline, a "t" and numbers, with a specific count, which is not included in the total CDS count, among others; for example: Cp1002_t001. For rRNAs, the nomenclature is the symbol followed by underline, an "r" and numbers, with specific counts, not included in the total CDS count; for example: Cp1002_r001. "PROTEIN_ID" is used to designate all of the elements of the genome, except for the feature "misc". It is a standardized form for NCBI to identify e proteins; for example: gnl|gbufpa|Cp1002_0001. "GENE" is one of the most important topics to be informed in manual annotation, indicating the gene symbol of the protein; fore example: pld. The field "SIMILARITY" corresponds to information obtained from the best similarity search result – BLASTp. Various types of information should be entered into this field, such as similarity among organisms, size of the amino-acids sequence analyzed, e-value and also the percentage identity between its own protein and the protein found in the data bank; for example: similar to *Corynebacterium pseudotuberculosis* 1002, hypothetical protein Cp1002_00047 (345 aa), e–value: 0.0, 98% ID in 344 aa. In "PRODUCT", there is a description of the gene product, for which similarity was found in the public domain data bank; for example: Phospholipase D. The tag "PSEUDO" should be added whenever a protein presents one or various breaks, due to insertion of a premature stop codon. These are the famous proteins that have frameshifts or probable pseudogenes. Consequently, the manual annotation window has this pattern:

/gene="dnaA"
/product="Chromosomal replication initiation protein"
/locus_tag="Cp1002_0001"
/protein_id="gnl|ufmg|Cp1002_0001"
/colour=3
/similarity="Similar to *Corynebacterium pseudotuberculosis* FRC41,
Chromosomal replication initiation protein (603 aa), e value: 0.0, 98% id in 599s aa"

4.2 Steps for manual curation

Manual curation is a very complex task and is subject to errors for various reasons. One of these is a lack of padronization in the interpretation of BLAST results. Another problem is propagation of errors, which involves prediction of protein function based on proteins that were also predicted but could have imprecise or even incorrect annotation (Gilks et al., 2002). For these reasons, some criteria are suggested in order to obtain reliable functional

annotation. The fundamental step for doing this well is mining data obtained from similarity analyses of BLASTp data banks. It is recommended to give greater value to annotation of proteins of individuals of the same species or of species that are phylogenetically close to the organism under study, the protein of which one wants to infer the function of, decreasing in this way the possibility of annotation errors. Another parameter is to observe if there is any consensus among the first 10 hits (the same protein is identified among various). In this case, even if the best hit is not identified as such, it is preferable to identify the sequence as similar to that of an organism that appears various times in the BLASTp results and is within the consensus. In cases where there is no consensus or when the e-value of the best hit (first BLAST result and which corresponds to the best alignment within the data bank that is being researched) is significantly larger than that of the following sequences, it is preferable to transfer the annotation of the best hit (Prosdocimi, 2003), or if necessary, in cases of non-significant alignments, always also run a similarity search at the nucleotide level (BLASTn). Other criteria are also analyzed, such as percentage identity between the sequence being analyzed and the sequence in the data bank, score value and e-value, as well as pair-by-pair alignment evaluation. This evaluation consists of checking the texture of the alignment (evaluating the number of gaps, size of the gaps, and the number of conserved substitutions of amino acids). If doubts remain, research of domain data banks and protein classification are also commonly utilized.

4.3 Frame shifts (Pseudogenes)

Comparisons between non-coding regions of genomes of various prokaryotic species has aided in the identification and characterization of genome segments with regulatory roles (Pareja et al., 2006), contributing to the elucidation of genetic circuits of transcriptional regulation. These non-coding regions, known as pseudogenes, are DNA sequences that are highly similar to functional genes but do not express a functional protein, probably because of deleterious mutations. These degraded genes contain one or more inactivating mutations, such as a nonsense mutation that introduces a premature stop codon, resulting in an incomplete protein and a later change in the open reading frame (Lerat & Ochman, 2005). When found in the genome, the break region is checked with Artemis, and the quality of the bases in that region is also evaluated. Whenever possible, addition or removal of erroneous bases can restore the reading frame. If there is no data that justifies addition or removal of bases, the genes should be classified as pseudogenes (tag /pseudo).

4.4 Tools

4.4.1 Artemis

The program Artemis, (Berriman & Rutherford, 2003), available for download at http://www.sanger.ac.uk/Software/Artemis is a freely-distributed algorithm developed for visualization of genomes and for annotation and manual curation. Artemis allows the curator to visualize various characteristics of the genome sequences, such as: product coded by the predicted gene; presence of tRNAs and rRNAs; search for protein and nucleotide similarity in biological data banks; visualization of probable domains and conserved protein families; visualization of GC / AT content, and misplaced codon use; and various other functions. These data can be visualized in the six phases of translating DNA reads into proteins (Rutherford et al., 2000). Also, the program provides a visualization of BLAST visits between two complete genome sequences, allowing rapid analysis of the degree of synteny

(conservation at the level of genes), the main genomic rearrangements and integration of new genomic islands (Field et al., 2005). This algorithm is written in the Java language and is available for the following operating systems: UNIX, Macintosh and Windows. Artemis is capable of processing data in the formats EMBL and GENBANK, or even sequences in the format FASTA.

4.5 Sequence similarity searches

4.5.1 BLAST (Basic Local Alignment Search Tool)

BLAST (Altschul et al., 1990) is a tool that is widely used for the characterization of products coded by genes that are identified by gene prediction. It is able to identify a great majority of the alignments that attend the desired criteria, with a significant gain in performance (Gibas & Jambeck, 2001). This program is available on the NCBI - National Center for Biotechnology Information site `http://www.ncbi.nlm.nih.gov` (Stein, 2001), which is considered the central databank for genome information. As shown in the figure, BLAST has programs for alignment of protein and nucleotide sequences, among others, according to the needs of the work that is to be undertaken:

Program	Entry sequence	Type of sequence target
BLASTp	Protein	Protein
BLASTn	Nucleotide	Nucleotide
BLASTx	Translated nucleotide	Protein
TBLASTn	Protein	Translated nucleotide
TBLASTx	Translated nucleotide	Translated nucleotide

Table 1. Types of BLAST – NCBI programs.

Through this type of algorithm, we can compare any DNA sequence or protein (query) with all of the genome sequences in the public domain (subject) (Altschul et al., 1997). It is important to note that the program BLAST does not try to make a comparison of the full extension of the molecules that are being compared, but rather it identifies in the data bank a sequence that is sufficiently similar to that of the sequence that is being studied.

4.5.2 Interpreting blast results

In the manual annotation of genomes, analysis of BLAST parameters, such as the number of points obtained (score), gap opening/extension penalties, number of expected alignments in the case of scores equal to or superior to the alignment that is being investigated (expectation value), and the normalized score (bitscore), are indispensible for the interpretation of the results. The smaller the value of "E", the smaller the chance of such a comparison being found merely by chance, consequently inferring a greater homology between the sequence being investigated and the data base (Baxevanis & Ouellette, 2001). Among the sequences with identity above 50%, a general approach is to characterize the function of the known sequence and transfer this annotation to the new sequence. Though annotation transfer is a common practice, a high rate of error has been reported when this is done without due caution (Liberman, 2004). Based on this principle, we consider that for sequences with identity above 80%, a simple alignment or a comparison with a protein that has been experimentally characterized using BLAST can be sufficient to infer function, as long as the pair being compared has similar lengths and align end to end without large

deletions or insertions. For pairs with identity in the range of 50–80%, the general approach for attributing function includes evaluation of databanks with homologous protein and protein domain families.

4.5.3 PFAM

Proteins generally are composed of one or more functional regions, or domains. Different combinations of domains result in the large variety of proteins found in nature. Identification of the domains that are found in proteins can, therefore, provide insight about protein function (Sanger Institute, 2009). In sequences with an identity of less than 70%, without end to end similarity, the approach that is used is to evaluate the protein domains through a search of the Pfam database, which gives very extensive coverage (Mazumder & Vasudevan, 2008). The Pfam database is accessible via the Web http://pfam.sanger.ac.uk and is available in various formats for download. This databank is contains two complementary groupings; Pfam–A is composed of high–quality protein domains that have been manually verified, while Pfam–B contains data that has been generated automatically from the ProDom databank (Finn et al., 2010). Pfam–B is generally lower in quality, though it can suggest new domains that can be added to the manual annotation, if they are not available in Pfam–A. Basically, in Pfam, the sequences that are in full alignment are identified through a search for a hidden profile using the algorithm Hidden Markov Model (HMM), which is later generated using the software HMMER, based on the UniProt database (UniProt, 2007). These HMMs are statistical models that capture specific information about how much each alignment column is conserved and indicates the residuals in this evaluation.

5. Genomics

A genome is the complete set of DNA sequences of a living organism; it consists of coding and non–coding sequences. Genomics is a discipline of genetics that deals with genomes or DNA sequences. Simply put, genomics is the study of genomes. Computational genomics derives knowledge from genome sequences and related data, including both DNA and RNA sequences as well as experimental data. Computational biology mainly deals with whole genome analysis to understand the DNA mechanisms and molecular biology of a species. As biological datasets are extremely large, computational biology has become an important part of modern biology.

5.1 Pangenomics

The efficient and low cost sequencing technologies that are currently available provide complete genome sequences of pathogenic, industrially useful, and other economically-important organisms. Genome sequences, and information that is coded in these sequences, can help identify pathogenicity and other important genes.

Complete genomic sequences of various strains of a species are important to help us understand pathogenesis mechanisms and to determine how genetic variability affects pathogenesis; it would be difficult to extract such useful information from a single genome (Lefébure & Stanhope, 2007).

A pangenome consists of a "core genome", which contains the gene or sequences present in all strains. In other words, genes that are found in all the genomes in a species of bacteria are

called the core genome. A "dispensable genome or accessory genome" consists of genome sequences present in more than two strains but are not part of the core genome. "Unique genomic sequences" or "unique genes" are strain-specific genes. These genes are limited to single strain. The pangenome is important for identification and for designing effective vaccines and drug targets (Mira et al., 2010).

There are many web tools and softwares available to manage and efficiently extract data from genomes of various strains of the same species. These tools recognize the accession numbers allotted to complete genomes submitted to NCBI and to other databanks. Online tools developed by the Computational Genomics group of Bielefeld University, Germany, EDGAR – "Efficient Database framework for comparative Genome Analyses using BLAST score Ratios" http://edgar.cebitec.uni-bielefeld.de are efficient web tools to determine the core genome, along with dispensable and unique genes in the form of colored graphs and tables (Blom et al., 2009)

For example, we analyzed the core genome, dispensable genes and unique genes, using "EDGAR", of three different *Corynebacterium pseudotuberculosis* strains, *C. pseudotuberculosis* Cp–I19, *C. pseudotuberculosis* Cp1002 and *C. pseudotuberculosis* CpC231.

This core genome consists of 1,862 genes, with 48 dispensable genes between Cp–I19 and Cp1002, 52 dispensable genes between Cp-I19 and CpC231, and 103 dispensable genes between Cp1002 and CpC231. There were 208, 46 and 36 unique genes in strains Cp-I19, Cp1002 and CpC231, respectively.

6. Genome plasticity

The high degree of adaptability of bacteria to a wide range of environments and hosts is long known to be influenced by genome plasticity, a dynamic property that involves DNA gain, loss and rearrangement (Maurelli et al., 1998). Various mechanisms can drive these changes, including point mutations, gene conversions, rearrangements (inversion or translocation), deletions and DNA insertions from other organisms (plasmids, bacteriophages, transposons, insertion elements and genomic islands) (Schmidt & Hensel, 2004).

6.1 Plasmids

Plasmids contribute to genomic plasticity through their transfer capability. They are also able to mobilize co-resident plasmids and integrate into the chromosome. Plasmids may harbor antibiotic resistance genes and other genes associated with pathogenicity (Dobrindt & Hacker, 2001); e.g., Rhodococcus equi harbors a virulence plasmid that codes for surface-associated proteins (vap genes) that is absent in avirulent strains (Takai et al., 2000).

6.2 Bacteriophages

Bacteriophages are viruses that infect bacteria and which influence genome plasticity through transduction mechanisms. Functional phages inject DNA from one bacterium into another one without causing damage to the acceptor organism; the DNA can incorporate into the acceptor genome leading to adaptive changes. Additionally, prophages (viral DNA incorporated in the bacterial chromosome) confer protection against lytic infections and they can harbor virulence genes that may be acquired by the acceptor bacterium and directly affect its pathogenicity; this has been reported from various species, including Clostridium

botulinum, Streptococcus pyogenes, Staphylococcus aureus, Escherichia coli and C. diphtheriae (Brüssow et al., 2004).

6.3 Genomic islands

Genomic islands (GEIs) affect genome plasticity because of their mobility and their capability of carrying a large number of genes as a single block, including operons and groups of coding genes with related functions. These GEIs can cause dramatic changes that lead the acceptor bacterium to evolve very rapidly compared to wild-type counterparts. GEIs are characterized as large DNA regions acquired from other organisms. They vary in size (10-200 kb), and can harbor sequences derived from phages and/or plasmids, including integrase genes; GEIs are flanked by tRNA genes or direct repeats, which help produce their characteristic instability (Hacker & Carniel, 2001). The instability of GEIs is exemplified by rapid gene acquisition and/or loss and changes in gene composition, as seen in different strains of Burkholderia pseudomallei (Tumapa et al., 2008). Additionally, GEIs can be classified into several classes according to gene content. These include Symbiotic Islands, which are involved in the association of bacterium with Leguminosae hosts (Barcellos et al., 2007); Resistance Islands, which harbor genes related to antibiotic resistance (Krizova & Nemec, 2010); Metabolic Islands, which contain genes associated with secondary metabolite biosynthesis (Tumapa et al., 2008); and Pathogenicity Islands (PAIs), which have a high concentration of virulence genes. PAIs are associated with pathogenic bacteria and have been implicated in the reemergence of various pathogens as causes of serious disease problems (Dobrindt et al., 2000). The first description of a PAI was made in 1990, in vitro (Hacker et al., 1990),. The identification was based on the observation of a close relation between deletion of hemolysin and fimbrial adhesin coding regions and non pathogenic strains of E. coli. This was investigated by gene cloning technique, pulse field electrophoresis and Southern hybridization. Using these procedures, they showed that the hemolysin and fimbrial adhesin coding genes are located in the same chromosomal region in several wild-type strains of E. coli and that they go through deletion events both in vivo and in vitro (Hacker et al., 1990).

6.4 "Black Holes"

Additionally, it is important to keep in mind that gene deletion is just as important as gene acquirement in some organisms. One example of this event is the so called "Black Holes" or deletion events of "antivirulence" genes, i.e. genes whose expression in pathogenic organisms is incompatible with virulence. The concept of evolution through deletion of "antivirulence" genes is based on the premise that genes required for adaptation of one organism in a specific niche may inhibit adaptability in another niche, a potential host, for example (Maurelli, 2007). In E. coli, loss of cadA, the lysine decarboxylase (LDC) coding gene, and ompT, which synthesizes an outer membrane protein, may confer virulence (Suzuki & Sasakawa, 2001). The mechanism of action of cadaverine, produced by decarboxylation of lysine by LDC, is still unknown. However, there are two hypotheses: cadaverine inactivates the synthesized toxin, or cadaverine acts directly on the target cell to protect it. Maurelli et al. (1998) demonstrated that when rabbit mucous cells are pre-treated with cadaverine and then washed, they are protected from enteroxin effects. Absence of Omp-T in Shigella strains and enteroinvasive E. coli strains is crucial for maintaining VirG on the cell surface, a pre-requisite for mobility on mammal cells, including bacterial dispersion through epithelial cells (Suzuki & Sasakawa, 2001).

6.5 Software to identify horizontal gene transfer (HGT) events

Gene acquisition and loss through HGT influence bacterial lifestyles and their physiological versatility (Dobrindt & Hacker, 2001). The increasing number of complete genome sequences available for analysis has stimulated in silico research in an effort to identify HGT events. Horizontally–acquired regions can be identified based on observation G+C content and codon usage patterns, which differ among species and species groups. Sets of genes acquired by HGT events show deviations in these patterns that reflect the genomic signature of the donor genome (Langille et al., 2008). Various softwares can be used to identify HGT events based on base composition patterns (wavelet analysis of G+C content, cumulative GC profile, P–web, IVOM, IslandPath and PAI–IDA) and codon usage deviation (SIGI–HMM and PAI–IDA). However, due to adaptations in codon usage (Karlin et al., 1998), which tend towards homogenous base composition distributions (Hershberg & Petrov, 2009), identification of mobile regions based on genomic signature is only possible for regions that have recently been acquired from phylogenetically distant organisms, i.e. those that have a discrepant genomic signature when compared to the acceptor genome.

Additionally, identification of HGT events may be aided by concentrating on regions that are flanked by tRNA genes, which are "hot spots" for transfer elements since they possess 3'–terminal insertion sequences that are recognized by various integrases (Hou, 1999). The integration of PAIs into these insertion sequences is responsible for their instability, since a single integrase may cause excision of the entire region. Insertion/deletion events have been demonstrated in PAIs I and II of E. coli strain 536, which are flanked by selC and leuX tRNA genes (Blum et al., 1994), and in high pathogenicity islands (HPIs) of several *Yersinia pseudotuberculosis* and *Y. pestis* strains (Lesic et al., 2004), which frequently insert into ASN3 tRNA genes.

However, although efficient in the identification of HGT events, approaches based on genomic signature and flanking tRNAs are not aimed at classification of GEIs, since they do not consider the overall gene content of the region. Additionally, horizontally acquired regions may deviate only in G+C content or codon usage alone, which would be a problem for the identification process if only one of these features is used to identify the event. However, there are tools designed to identify a specific class of GEIs, pathogenicity islands, through a multi-pronged strategy that overcomes such constraints. These tools are named PredictBias (Pundhir et al., 2008), IslandViewer (Waack et al., 2006) and PIPS (unpublished); they perform analyses based on genomic signature deviations that are not found in closely-related organisms and finding of genes coding for virulence factors. Although all of these programs use similar strategies and are complementary, PIPS deserves special attention since it surpasses the others in accuracy and is easy to install.

In analysis of C. diphtheriae strain NCTC 13129, PIPS outperformed the other approaches, identifying 12 out of the 13 PAIs of the reference strain, compared to 10 by IslandViewer and six by PredictBias. In the identification of PAIs of uropathogenic E. coli strain CFT073, PIPS had an overall accuracy of 93.9% (unpublished) against 89.5% for IslandViewer and 88.1% for PredictBias.

7. Reverse vaccinology

Reverse Vaccinology (RV) (Rappuoli, 2000) starts from the genomic sequence of a pathogen, which is an expected coded sequence for all the possible genes expressed during the life cycle of the pathogen. All open reading frames (ORF's) derived from the genome sequence

can be evaluated with a computer program in order to determine their aptitude as vaccine candidates. Special attention is given to exported proteins because they are essential in host-pathogen interactions. Examples of this interaction can be cited: (i) adherence to host cells, (ii) invasion of compliant cells, (iii) damage to host tissues, (iv) resistance to environmental stress by the machinery defense of the cell being infected and finally, (v) mechanisms for subversion of the host immune response (Sibbald & van Dij, 2009). The word "Reverse" in RV can be explained by the reverse genetics (RG) technique. Before the dawn of genomics, there were attempts to discover the genes responsible for each phenotype. With Crick's central dogma (DNA > RNA > Protein) the research path was reversed. In possession of the likely gene sequence, several techniques have been developed to identify changes in the phenotype of an organism derived from sequence changes in genes. The principle of Crick's dogma is also used by RV; when a gene sequence is found, one can determine whether a probable protein encoded by this sequence can be an antigen capable of stimulating an immune response in a host organism.

Long before the creation of the term RV, a number of approaches had been considered to determine exported proteins in order to move to the next step of the production of a subunit vaccine (Diaz Romero & Outschoorn, 1994). For example, research on exported proteins was advanced as an alternative to subunit vaccines based on the polysaccharide capsule of meningococci. Vaccines produced with such antigens had a low capacity to induce a satisfactory immune response. This research effort on exported proteins includes almost two decades of work searching for a vaccine against meningococcal serogroup B, which now gives good results. This vaccine currently is the best RV alternative for the production of a subunit vaccine for Neisseria meningitidis serogroup B. Meningitis caused by serogroup B (Men B) is responsible for approximately half of the worldwide incidence of this disease (Diaz Romero & Outschoorn, 1994), and this research result for targeted vaccination is commonly used as a demonstration of the usefulness of RV, because of the excellent results.

Currently, a subunit vaccine against Men B created with antigens targeted by RV is being tested in phase-2 clinical trials (Bambini & Rappuoli, 2009). The advantages of RV continue to be attractive, enabling vaccine research for organisms whose cultivation in the laboratory is difficult or impossible. Reducing the time needed to select target proteins could allow investigation of different species or strains at the same time, for selecting vaccine candidates that can elicit adaptive immune responses. To achieve these benefits all we need is to have a sequenced genome, a personal computer and core software widely available to the scientific community. These conditions demonstrate another advantage of using RV, the low cost. What we call core software is a set of tools for identifying well-known motifs, such as, for example, SignalP, TMHMM, LipoP, and HMMSEARCH. There is still room for innovation in the use of core software; the choice of software strategies can be directed to the identification of vaccine candidates specific to an organism, such as in the case of gram-negative (bilayer) or gram positive (monolayer) bacteria, or also according to heuristics for selection of vaccine candidates with specific characteristics. For example, membrane or exported to the extracellular environment (Barinov et al., 2009).

The concept of RV was adapted to fit a new reality of widespread availability of genomic data (Rinaudo et al., 2009). Instead of researching vaccine targets for a single strain or subspecies of an organism, we can do it simultaneously in dozens of genomes, exploring potential joint antigens or those exclusive to multiple genomes (Lapierre & Gogarten, 2009). The possibility of having a large number of genomes available to implement RV leads to the

emergence of the concept of pangenomics RV (PGRV) (Bambini & Rappuoli, 2009). PGRV can also apply the concepts of core, extended, and character genomes. The core genome in PGRV is composed of exported genes (genes that transcribe exported proteins) that are common to all strains, genes that could be candidates for a universal vaccine, while the extended genome consists of genes that are absent in at least one of the strains of the studied species, while the character genome consists of genes that are specific to a strain (Lapierre & Gogarten, 2009). From the standpoint of vaccines, the core and character genomes would be good candidates to develop a vaccine that is suitable for all strains, without losing sight of the particularities of specific genes in each strain.

8. Immunoinformatics

The immune system has considerable diversity in its components, such as, for example, immunoglobulin receptors of lymphocytes, or cytokines, with the principle cell types being B- and T-cells, which have important roles in inflammation, infection and protection (Evans, 2008). Immunoinformatics is very complex and can be characterized as a combinatory science, since it has a great complexity of regulatory cycles and network type interactions, which allows the utilization of computational models to resolve problems that can be converted into biological significant responses (Brusic & Petrovsky, 2003). This leads us to immunoinformatics, which is the application of informatics techniques to immune system molecules, with the main objective of helping develop vaccines through the prediction of immunogenic epitopes (Flower & Doytchinova, 2002).

8.1 Immunological databases

The immunological databases are a source of data used to explore, refine and develop new tools and algorithms (Salimi et al., 2010). There is a large variety of databases that group information relevant to the immune system. The Nucleic Acids Research Molecular Biology Database Collection http://www3.oup.co.uk/nar/database/c/ included 29 immunological databases in March 2011. The International ImMunoGeneTics information system (IMGT), the world reference databank for immunogenetics and immunoinformatics, was created by Marie-Paule Lefranc in 1989 (Lefranc et al., 2009). This databank is specialized in immunoglobulins or antibodies, T-cell receptors (TCR), MHCs, and others. The IMGT is constituted of a variety of databanks, including: structure, monoclonal antibody, sequence and genome databanks. All of these databanks are curated manually and daily by a team that works fulltime, which helps maintain high-quality annotation and standardization of the information. Other databases that house information related to epitopes, such as AntiJen (Toseland et al., 2005) and FIMM (Schonbach et al., 2000), have not been maintained and their data has migrated to other websites. Among these, the most promising epitope database seems to be the Immune Epitope Database (IEDB) (Peters et al., 2005), which is a curated database that has information based on experimental data associated with the target epitope; consequently, it is hoped that all of the information in the various existing databanks also migrates to IEDB within the next few years.

8.2 Epitope prediction

The principal goal of immunoinformatics is the development of algorithms that can both help develop vaccines and analyze the gene products of pathogens, such as viruses and

bacteria. This is why it is very important to understand antigen-antibody interactions. Macallum et al. (1996) made a detailed analysis of 26 antigen–antibody complexes; they found that binding between molecules is very complex, and that there are different antibody–antigen classes for different types of molecules. A later study of 59 antigen–antibody interactions (Almagro, 2004) found results similar to those of Macallum. These studies show that tools that can identify molecules and predict their interactions with other molecules need to be very accurate and sensitive.

8.2.1 B cell epitope prediction

Epitopes of B cells are antigenic regions that are recognized by antibodies of the immune system, specifically those that interact with B cell receptors. These epitopes can be continuous or discontinuous (Kumagai & Tsumoto, 2001). B–cell epitopes can be used to design vaccines and new diagnostic tests (Larsen et al., 2006). As with T cells, there are also numerous methodologies to model and predict B–cell epitopes. The classic system to predict B–cell epitopes (Hopp & Woods, 1981) uses propensity scale methods (Parker et al., 1986; Levitt, 1978). This method attributes a propensity value to each amino acid, based on studies of the physical–chemical properties. A combination of various scales can improve the prediction results (Pellequer et al., 1991). This work used hydrophilicity scales (Parker et al., 1986), as well as secondary structure (Levitt, 1978; Chou & Fasman, 1978) and accessibility (Emini et al., 1985). The Immune Epitope Database and Analysis Resource, IEDB (Peters et al., 2005), utilizes parameters such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antigenic propensity of polypeptides chains, which have been correlated with the location of continuous epitopes. All of the prediction calculations are based on propensity scales. Another methodology that can be used to predict continuous B-cell epitopes combines hidden Markov model (HMM) and propensity scale methods (Parker et al., 1986; Levitt, 1978); it is called Bepipred http://www.cbs.dtu.dk/services/BepiPred/ (Larsen et al., 2006). This methodology has given increased prediction accuracy. Prediction of discontinuous B–cell epitopes has also improved, due to an increase in the number of three–dimensional (3D) structures of antibody–antigen complexes available in PDB and in IMGT/3Dstructure-DB (Kaas et al., 2004) and in the Epitome database (Schlessinger et al., 2006).

8.2.2 T cell epitope prediction

There are two classes of T cells: (1) CD8+ T cytotoxic (Tc) cells, which produce cytotoxins responsible for cell lysis, recognize peptides presented by class I MHCs and (2) CD4+ T helper (Th) cells, which recognize proteins associated with MHC class II. Interferon γ (IFN–γ) and tumor necrosis factor β (TNF–β) are produced by Th1 cells. Th2 cells produce interleukin 4 (IL–4), IL–5, IL–10 and IL–13. Eptitopes that bind to MHC de class I generally are 8–10 amino acids long, with a mean of nine amino acids (Reche et al.., 2002), while epitopes that bind to MHC class II are 13–17 amino acids long (Sercarz & Maverakis, 2003; Chicz et al., 1992). There are various online tools for predicting T–cell epitopes on the basis of MHC class I and class II binding. Prediction of MHC binding is based on motifs associated with epitopes or binders for specific alleles. SYFPEITHI is a tool that is widely used for prediction of T–cell epitopes and MHC binding; however, these predictions have been found to be of low quality (Ruppert et al., 1993). More sophisticated tools that use quantitative matrixes, artificial neural network decision trees, hidden Markov models

(HMM), support vector machines (SVM), homology modeling, protein threading and docking techniques have been developed. The NetMHC 3.2 server http://www.cbs.dtu.dk/services/NetMHC/ predicts binding of peptides to a series of different HLA alleles using artificial neural networks (ANNs) and weight matrixes. All of the previous versions are available online, for comparison and reference. ANNs were trained with 57 different human MHCs (HLA), representing all of the 12 HLA alleles, supertypes A and B (Lund et al., 2004). Also predictions are available for 22 animal alleles (monkey and rat). ANN prediction values are given in nM IC50 values. Weight prediction matrixes use an aptitude score, with a high aptitude score indicating strong binding. Predictions can be made for sizes from 8 to 11 for all of the alleles using an ANNs algorithm trained with 9mer peptides. Probably because of the limited quantity of 10mer data available, this method has better prediction value when an ANNs algorithm is trained with 10mer data. However, one should be careful with 8mer predictions, since some alleles do not link to 8mer to a significant degree. Binding peptides are indicated at output as strongly binding (SB) and weakly binding (WB). The allele for each HLA supertype is indicated in the selection window for HLA alleles (Lundegaard et al., 2008).

The NetMHCII 2.2 server `http://www.cbs.dtu.dk/services/NetMHCII/` predicts peptides that bind to MHC classe II alleles HLA–DR, HLA–DQ, HLA–DP and mouse alleles, using ANNs. Predictions can be obtained for the 14 HLA–DR alleles, including the nine HLA–DR, six HLA–DQ, and six HLA–DP supertypes and two H2 class II alleles in mice. The prediction values are given in nM IC50 values, and in %–Rank for a random set of 1,000,000 natural peptides. Strongly and weakly binding peptides are indicated in the output file (Nielsen et al., 2007).

Without a doubt, there is a great variety of predictors, which when they are combined can be quite precise in the prediction of T–cell epitopes; however, this is only possible when well–characterized alleles are available, which is true for some alleles that have been predicted as MHC class I alleles, but much less so for those predicted as MHC class II. This is even more of a problem in the prediction of B cell proteins, for which it is often necessary to have prior knowledge of the structure and sequence of the protein. Nevertheless, it is known that no method can go further than the data used to train it, and only through extensive compilation and by obtaining high quality data, will it be possible to create excellent models that will can be generally applied (Flower & Doytchinova, 2002).

9. References

Allen, J. E., Pertea, M. Salzberg, S. L. 2004. Computational gene prediction using multiple sources of evidence, *Genome Res* 14(1): 142–8.

Almagro, J. C. 2004. Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires, *J Mol Recognit* 17(2): 132–43.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. Lipman, D. J. 1990. Basic local alignment search tool, *J Mol Biol* 215(3): 403–10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. Lipman, D. J. 1997. Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic Acids Res* 25(17): 3389–402.

Aparicio, G., Götz,, S., Conesa, A., Segrelles, D., Blanquer, I., García, J. M., Hernandez, V., Robles, M. Talon, M. 2006. Blast2go goes grid: developing a grid-enabled prototype for functional genomics analysis, *Stud Health Technol Inform* 120: 194–204.

Bambini, S. Rappuoli, R. 2009. The use of genomics in microbial vaccine development, *Drug Discov Today* 14(5-6): 252–60.

Barcellos, F. G., Menna, P., da Silva Batista, J. S. Hungria, M. 2007. Evidence of horizontal transfer of symbiotic genes from a bradyrhizobium japonicum inoculant strain to indigenous diazotrophs sinorhizobium (ensifer) fredii and bradyrhizobium elkanii in a brazilian savannah soil, *Appl Environ Microbiol* 73(8): 2635–43.

Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., Maguin, E. van de Guchte, M. 2009. Prediction of surface exposed proteins in streptococcus pyogenes, with a potential application to other gram-positive bacteria, *Proteomics* 9(1): 61–73.

Baxevanis, A. D. Ouellette, F. F. 2001. A practical guide to the analysis of genes and proteins, *Wiley* (2): 260–2.

Bendtsen, J. D., Nielsen, H., von Heijne, G. Brunak, S. 2004. Improved prediction of signal peptides: Signalp 3.0, *J Mol Biol* 340(4): 783–95.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. Wheeler, D. L. 2008. Genbank, *Nucleic Acids Res* 36(Database issue): D25–30.

Berriman, M. Rutherford, K. 2003. Viewing and annotating sequence data with artemis, *Brief Bioinform* 4(2): 124–32.

Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M. Goesmann, A. 2009. Edgar: a software framework for the comparative analysis of prokaryotic genomes, *BMC Bioinformatics* 10: 154.

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschäpe, H. Hacker, J. 1994. Excision of large dna regions termed pathogenicity islands from trna-specific loci in the chromosome of an escherichia coli wild-type pathogen, *Infect Immun* 62(2): 606–14.

Brown, T. A. 1999. Genes e expressÃ£o gênica., *Genética – um enfoque molecular* 1(2): 124–132.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. Kahn, D. 2005. The prodom database of protein domain families: more emphasis on 3d, *Nucleic Acids Res* 33(Database issue): D212–5.

Brusic, V. Petrovsky, N. 2003. Immunoinformatics–the new kid in town, *Novartis Found Symp* 254: 3–13; discussion 13–22, 98–101, 250–2.

Brüssow, H., Canchaya, C. Hardt, W.-D. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion, *Microbiol Mol Biol Rev* 68(3): 560–602.

Chicz, R. M., Urban, R. G., Lane, W. S., Gorga, J. C., Stern, L. J., Vignali, D. A. Strominger, J. L. 1992. Predominant naturally processed peptides bound to hla-dr1 are derived from mhc-related molecules and are heterogeneous in size, *Nature* 358(6389): 764–8.

Choi, G.-E., Eom, S.-H., Jung, K.-H., Son, J.-W., Shin, A.-R., Shin, S.-J., Kim, K.-H., Chang, C. L. Kim, H.-J. 2010. Cysa2: A candidate serodiagnostic marker for mycobacterium tuberculosis infection, *Respirology* 15(4): 636–42.

Chou, P. Y. Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence, *Adv Enzymol Relat Areas Mol Biol* 47: 45–148.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Barrell, B. G. 2001. Massive gene decay in the leprosy bacillus, *Nature* 409(6823): 1007–11.

Datta, S., Datta, S., Kim, S., Chakraborty, S. Gill, R. S. 2010. Statistical analyses of next generation sequence data: A partial overview, *J Proteomics Bioinform* 3(6): 183–190.

Diaz Romero, J. Outschoorn, I. M. 1994. Current status of meningococcal group b vaccine candidates: capsular or noncapsular? , *Clin Microbiol Rev* 7(4): 559–75.

Dobrindt, U. Hacker, J. 2001. Whole genome plasticity in pathogenic bacteria, *Curr Opin Microbiol* 4(5): 550–7.

Dobrindt, U., Janke, B., Piechaczek, K., Nagy, G., Ziebuhr, W., Fischer, G., Schierhorn, A., Hecker, M., Blum-Oehler, G. Hacker, J. 2000. Toxin genes on pathogenicity islands: impact for microbial evolution, *Int J Med Microbiol* 290(4-5): 307–11.

Emini, E. A., Hughes, J. V., Perlow, D. S. Boger, J. 1985. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide, *J Virol* 55(3): 836–9.

Evans, M. C. 2008. Recent advances in immunoinformatics: application of in silico tools to drug development, *Curr Opin Drug Discov Devel* 11(2): 233–41.

Field, D., Feil, E. J. Wilson, G. A. 2005. Databases and software for the comparison of prokaryotic genomes, *Microbiology* 151(Pt 7): 2125–32.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. Bateman, A. 2006. Pfam: clans, web tools and services, *Nucleic Acids Res* 34(Database issue): D247–51.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. Bateman, A. 2010. The pfam protein families database, *Nucleic Acids Res* 38(Database issue): D211–22.

Flower, D. R. Doytchinova, I. A. 2002. Immunoinformatics and the prediction of immunogenicity, *Appl Bioinformatics* 1(4): 167–76.

Gibas, C. Jambeck, P. 2001. Developing bioinformatics computer skills, *O'Reilly* 1(1): 21–22.

Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S. Ouzounis, C. A. 2002. Modeling the percolation of annotation errors in a database of protein sequences, *Bioinformatics* 18(12): 1641–9.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. Goebel, W. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal escherichia coli isolates, *Microb Pathog* 8(3): 213–25.

Hacker, J. Carniel, E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. a darwinian view of the evolution of microbes, *EMBO Rep* 2(5): 376–81.

Hershberg, R. Petrov, D. A. 2009. General rules for optimal codon choice, *PLoS Genet* 5(7): e1000556.

Hopp, T. P. Woods, K. R. 1981. Prediction of protein antigenic determinants from amino acid sequences, *Proc Natl Acad Sci U S A* 78(6): 3824–8.

Hou, Y. M. 1999. Transfer rnas and pathogenicity islands, *Trends Biochem Sci* 24(8): 295–8.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Yeats, C. 2009. Interpro: the integrative protein signature database, *Nucleic Acids Res* 37(Database issue): D211–5.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics* 11: 119.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. Madden, T. L. 2008. Ncbi blast: a better web interface, *Nucleic Acids Res* 36(Web Server issue): W5–9.

Kaas, Q., Ruiz, M. Lefranc, M. P. 2004. Imgt/3dstructure-db and imgt/structuralquery, a database and a tool for immunoglobulin, t cell receptor and mhc structural data, *Nucleic Acids Res* 32(Database issue): D208–10.

Karlin, S., Mrázek, J. Campbell, A. M. 1998. Codon usages in different gene classes of the escherichia coli genome, *Mol Microbiol* 29(6): 1341–55.

Kendrew, J. 1999. In: The encyclopedia of molecular biology, *in* B. Science (ed.), *Gene*, Porto Alegre, pp. 343–401.

Kislyuk, A. O., Katz, L. S., Agrawal, S., Hagen, M. S., Conley, A. B., Jayaraman, P., Nelakuditi, V., Humphrey, J. C., Sammons, S. A., Govil, D., Mair, R. D., Tatti, K. M., Tondella, M. L., Harcourt, B. H., Mayer, L. W. Jordan, I. K. 2010. A computational genomics pipeline for prokaryotic sequencing projects, *Bioinformatics* 26(15): 1819–26.

Krizova, L. Nemec, A. 2010. A 63 kb genomic resistance island found in a multidrug-resistant acinetobacter baumannii isolate of european clone i from 1977, *J Antimicrob Chemother* 65(9): 1915–8.

Krogh, A., Larsson, B., von Heijne, G. Sonnhammer, E. L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, *J Mol Biol* 305(3): 567–80.

Kumagai, I. Tsumoto, K. 2001. Antigen-antibody binding, *Encyclopedia of Life Sciences - Nature Publishing Group* pp. 1–7.

Langille, M. G. I. Brinkman, F. S. L. 2009. Islandviewer: an integrated interface for computational identification and visualization of genomic islands, *Bioinformatics* 25(5): 664–5.

Langille, M. G. I., Hsiao, W. W. L. Brinkman, F. S. L. 2008. Evaluation of genomic island predictors using a comparative genomics approach, *BMC Bioinformatics* 9: 329.

Lapierre, P. Gogarten, J. P. 2009. Estimating the size of the bacterial pan-genome, *Trends Genet* 25(3): 107–10.

Larsen, J. E., Lund, O. Nielsen, M. 2006. Improved method for predicting linear b-cell epitopes, *Immunome Res* 2: 2.

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. Duroux, P. 2009. Imgt, the international immunogenetics information system, *Nucleic Acids Res* 37(Database issue): D1006–12.

Lefébure, T. Stanhope, M. J. 2007. Evolution of the core and pan-genome of streptococcus: positive selection, recombination, and genome composition, *Genome Biol* 8(5): R71.

Lerat, E. Ochman, H. 2005. Recognizing the pseudogenes in bacterial genomes, *Nucleic Acids Res* 33(10): 3125–32.

Lesic, B., Bach, S., Ghigo, J.-M., Dobrindt, U., Hacker, J. Carniel, E. 2004. Excision of the high-pathogenicity island of yersinia pseudotuberculosis requires the combined actions

of its cognate integrase and hef, a new recombination directionality factor, *Mol Microbiol* 52(5): 1337–48.

Levitt, M. 1978. Conformational preferences of amino acids in globular proteins, *Biochemistry* 17(20): 4277–85.

Li, L., Shiga, M., Ching, W.-K. Mamitsuka, H. 2010. Annotating gene functions with integrative spectral clustering on microarray expressions and sequences, *Genome Inform* 22: 95–120.

Liberman, F. 2004. *Análise dos fatores determinantes para a qualidade da anotação genÃ´mica automática*, Master's thesis, Universidade Católica de Brasília.

Lorenzi, H. A., Puiu, D., Miller, J. R., Brinkac, L. M., Amedeo, P., Hall, N. Caler, E. V. 2010. New assembly, reannotation and analysis of the entamoeba histolytica genome reveal new genomic features and protein content information, *PLoS Negl Trop Dis* 4(6): e716.

Lukashin, A. V. Borodovsky, M. 1998. Genemark.hmm: new solutions for gene finding, *Nucleic Acids Res* 26(4): 1107–15.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. Brunak, S. 2004. Definition of supertypes for hla molecules using clustering of specificity matrices, *Immunogenetics* 55(12): 797–810.

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O. Nielsen, M. 2008. Netmhc-3.0: accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8-11, *Nucleic Acids Res* 36(Web Server issue): W509–12.

Macallum, R. M., Martin, A. C. R. Thornton, J. M. 1996. Antibody-antigen interactions: Contact analysis and binding site topography, *Journal of Molecular Biology* 262: 732–45.

Mathé, C., Sagot, M.-F., Schiex, T. Rouzé, P. 2002. Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res* 30(19): 4103–17.

Maurelli, A. T. 2007. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens, *FEMS Microbiol Lett* 267(1): 1–8.

Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. Fasano, A. 1998. "black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of shigella spp. and enteroinvasive escherichia coli, *Proc Natl Acad Sci U S A* 95(7): 3943–8.

Mazumder, R. Vasudevan, S. 2008. Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function, *PLoS Comput Biol* 4(9): e1000151.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. Pühler, A. 2003. Gendb–an open source genome annotation system for prokaryote genomes, *Nucleic Acids Res* 31(8): 2187–95.

Mira, A., Martín-Cuadrado, A. B., D'Auria, G. Rodríguez-Valera, F. 2010. The bacterial pan-genome:a new paradigm in microbiology, *Int Microbiol* 13(2): 45–57.

Nielsen, M., Lundegaard, C. Lund, O. 2007. Prediction of mhc class ii binding affinity using smm-align, a novel stabilization matrix alignment method, *BMC Bioinformatics* 8: 238.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T. Edwards, e. a. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic Acids Res* 33(17): 5691–702.

Pareja, E., Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Bonal, J. Tobes, R. 2006. Extratrain: a database of extragenic regions and transcriptional information in prokaryotic organisms, *BMC Microbiol* 6: 29.

Parker, J. M., Guo, D. Hodges, R. S. 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites, *Biochemistry* 25(19): 5425–32.

Pearson, W. R. Lipman, D. J. 1988. Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A* 85(8): 2444–8.

Pellequer, J. L., Westhof, E. Van Regenmortel, M. H. 1991. Predicting location of continuous epitopes in proteins from their primary structures, *Methods Enzymol* 203: 176–201.

Peters, B., Sidney, J., Bourne, P., Bui, H. H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo, R., Lund, O., Nemazee, D., Ponomarenko, J. V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Wilson, S. Sette, A. 2005. The immune epitope database and analysis resource: from vision to blueprint, *PLoS Biol* 3(3): e91.

Poptsova, M. S. Gogarten, J. P. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes, *Microbiology* 156(Pt 7): 1909–17.

Prosdocimi, F. 2003. Bioinformática: manual do usuário., *Biotecnologia Ciência & Desenvolvimento* 2(29): 2.

Pundhir, S., Vijayvargiya, H. Kumar, A. 2008. Predictbias: a server for the identification of genomic and pathogenicity islands in prokaryotes, *In Silico Biol* 8(3-4): 223–34.

Rappuoli, R. 2000. Reverse vaccinology, *Curr Opin Microbiol* 3(5): 445–50.

Retter, I., Althaus, H. H., Munch, R. Muller, W. 2005. Vbase2, an integrative v gene database, *Nucleic Acids Res* 33(Database issue): D671–4.

Rinaudo, C. D., Telford, J. L., Rappuoli, R. Seib, K. L. 2009. Vaccinology in the genome era, *J Clin Invest* 119(9): 2515–25.

Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M. Sette, A. 1993. Prominent role of secondary anchor residues in peptide binding to hla-a2.1 molecules, *Cell* 74(5): 929–37.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. Barrell, B. 2000. Artemis: sequence visualization and annotation, *Bioinformatics* 16(10): 944–5.

Salimi, N., Fleri, W., Peters, B. Sette, A. 2010. Design and utilization of epitope-based databases and predictive tools, *Immunogenetics* 62(4): 185–96.

Salzberg, S. L., Delcher, A. L., Kasif, S. White, O. 1998. Microbial gene identification using interpolated markov models, *Nucleic Acids Res* 26(2): 544–8.

Schellenberg, M. J., Ritchie, D. B. MacMillan, A. M. 2008. Pre-mrna splicing: a complex picture in higher definition, *Trends Biochem Sci* 33(6): 243–6.

Schlessinger, A., Ofran, Y., Yachdav, G. Rost, B. 2006. Epitome: database of structure-inferred antigenic epitopes, *Nucleic Acids Res* 34(Database issue): D777–80.

Schmidt, H. Hensel, M. 2004. Pathogenicity islands in bacterial pathogenesis, *Clin Microbiol Rev* 17(1): 14–56.

Schonbach, C., Koh, J. L., Sheng, X., Wong, L. Brusic, V. 2000. Fimm, a database of functional molecular immunology, *Nucleic Acids Res* 28(1): 222–4.

Sercarz, E. E. Maverakis, E. 2003. Mhc-guided processing: binding of large antigen fragments, *Nat Rev Immunol* 3(8): 621–9.

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D. Kahn, D. 2002. Prodom: automated clustering of homologous domains, *Brief Bioinform* 3(3): 246–51.

Setúbal, J. Meidanis, J. 1997. *Introduction to Computational Molecular Biology*, Pacific Grove.

Sibbald, M. J. J. B. van Dij, J. M. l. 2009. Secretome mapping in gram-positive pathogens. in karl wooldridge (ed.), bacterial secreted protein: Secretory mechanisms and role in pathogenesis, *Caister Academic Press* pp. 193–225.

Sleator, R. D. 2010. An overview of the current status of eukaryote gene prediction strategies, *Gene* 461(1-2): 1–4.

Smith, T. F. Waterman, M. S. 1981. Identification of common molecular subsequences, *J Mol Biol* 147(1): 195–7.

Stein, L. 2001. Genome annotation: from sequence to biology, *Nat Rev Genet* 2(7): 493–503.

Stothard, P. Wishart, D. S. 2006. Automated bacterial genome analysis and annotation, *Curr Opin Microbiol* 9(5): 505–10.

Suzuki, T. Sasakawa, C. 2001. Molecular basis of the intracellular spreading of shigella, *Infect Immun* 69(10): 5959–66.

Takai, S., Hines, S. A., Sekizaki, T., Nicholson, V. M., Alperin, D. A., Osaki, M., Takamatsu, D., Nakamura, M., Suzuki, K., Ogino, N., Kakuda, T., Dan, H. Prescott, J. F. 2000. Dna sequence and comparison of virulence plasmids from rhodococcus equi atcc 33701 and 103, *Infect Immun* 68(12): 6840–7.

Trost, B., Haakensen, M., Pittet, V., Ziola, B. Kusalik, A. 2010. Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera, *BMC Microbiol* 10: 258.

Tumapa, S., Holden, M. T. G., Vesaratchavest, M., Wuthiekanun, V., Limmathurotsakul, D., Chierakul, W., Feil, E. J., Currie, B. J., Day, N. P. J., Nierman, W. C. Peacock, S. J. 2008. Burkholderia pseudomallei genome plasticity associated with genomic island variation, *BMC Genomics* 9: 190.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. Banfield, J. F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* 428(6978): 37–43.

UniProt 2007. The universal protein resource (uniprot), *Nucleic Acids Res* 35(Database issue): D193–7.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P. Merkl, R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models, *BMC Bioinformatics* 7: 142.

Zhao, J. Grant, S. F. A. 2010. Advances in whole genome sequencing technology, *Curr Pharm Biotechnol*.