The impact of bioinformatics on microbiology

Bioinformatics can be described as a merging of information technology and molecular biology, where the former is used to solve problems in biology (Altman, 1998) which involves the analysis and organisation of biological data (Perez-Iratxeta et al., 2007). It is a relatively recent discipline with it's roots in the construction of molecular sequence databases between the late 1960's and early 1970's on early computers by organisations such as the National Institutes of Health (NIH) in the United States (Smith, 1990). With the foundation of large integrated databases such as GenBank in 1982 (Smith, 1990) along with major advances in computer technology and the development of a variety of biochemical wet-lab (laboratory bench-work) techniques that allow rapid generation and analysis of genomic and proteomic data (Bansal, 2005), bioinformatics has become an important recognised field of its own in the last twenty-odd years in particular. It has had a major impact on all fields of biology, and has revolutionised some of the manners in which microbiological research is carried out.

As the discipline of bioinformatics has evolved, the areas of research in which it is used have split into a number of fields including genomics, proteomics, systematics (Bull et al., 2000). Various methods of modelling cell behaviour and using data to research and develop new types of anti-microbial drugs and other agents are also a significant topic (Bansal, 2005). In the next sections each of these fields along with their impact on microbiology will be discussed.

Genomics involves the analysis of all the expressed and non-expressed genes otherwise known as the genome, of an organism. Genomics data is generated via sequencing of genomes. Aspects of this data can then be

analysed via bioinformatic methods allowing insights into which genes are expressed and prediction of gene location and function (Perez-Iratxeta et al., 2007), some applications of this knowledge include the development of antimicrobial agents and/or drugs and optimising production by microbes that are used in industry. Comparative genomics is where two genomes are sequenced and compared with each other whereas metagenomics involves the comparison of the genomes of a community of bacteria and thus is of use in microbial ecology studies. An example of the manner in which bioinformatics has affected microbiological research in particular, can be seen in the method known as shotgun sequencing that was invented to carry out the first whole genome sequencing of a bacterial strain, namely H. influenzae Rd (Fleischmann et al., 1995). In brief, this method involves random fragmentation of the chromosome in to small sections of DNA that are then sequenced and assembled. The assembly of the contiguous DNA fragments is carried out via the use of various software programs such as “ Autoassembler” (Fleischmann et al., 1995). This method was much more rapid than previous sequencing methods which lacked this semi-automation. The ability of techniques such as this to be partially carried out in silico has allowed the sequencing of 1049 more bacterial genomes since 1995 according to the Genomes online database (GOLD). The further integration of computational methods and genomics has enabled the development of new high throughput methods such as pyrosequencing (Tettelin & Feldblyum, 2009), which serve to increase the speed and volume in which new genomes are sequenced. Informatics is then used to carry out the task of analysing this vast amount of data. Nucleotide sequences are uploaded onto databases such as EMBL, DDBJ or GenBank which now had over ten billion nucleotides

of sequence data in 2001, (Roos, 2001) and has still been growing at an exponential rate. Programs that enable analysis of this data include those that are based on Hidden Markov Model statistics such as “ GLIMMER”(Gene Locator and Interpolated Markov ModellER),(Tettelin & Feldblyum, 2009). These programs have the ability to predict open reading frames (ORF’s) in nucleotide sequences, i. e. protein coding regions on mRNA, by locating conserved regions of sequences. Automated search programs generally search for features such as a start and a triplet of stop codons, as well as accounting for codon bias-where in a particular organism there will be a bias for a certain codon when coding for certain amino acids- Guanine-Cytosine content is also a considered factor since a GC content of more that 50% on a sequence can indicate an ORF large enough to potentially encode a functional gene (Zavala et al., 2005). Comparative genomics is a method that allows confirmation of functionality of predicted ORF’s (Chakravarti et al., 2000). It involves carrying out a search for similarities between the predicted ORF and other sequenced and annotated genes on an online database, if a result showing high similarity is attained it is likely that the two sequences are homologous, meaning they are evolutionarily linked and potentially have a similar function. Software tools such as BLAST (Basic Local Alignment Search Tool) and FASTA allow rapid searches of these online databases to be carried out (Chakravarti et al., 2000). These programs can be used to search for protein-protein, nucleotide-nucleotide, protein-translated nucleotide as well as various other alignments. Alignments that can be searched for can be classified as local or global, which are short sections between sequences that are highly similar or the best alignment between entire sequences, these programmes can also accommodate

insertions, deletions, substitutions and deletions in sequences when aligning them. However there are also various drawbacks involved with these methods; including the fact that genes can be incorrectly annotated on databases, or homologous genes may simply have not been sequenced and uploaded yet. In these cases wet-lab analysis must be carried out for identification and annotation of potential genes. These methods can include inactivation of a predicted gene and testing whether there is any change in the phenotype of the cell.

An example of the use of genomics in the analysis of pathogenic bacteria is the comparative analysis that was carried out of the genome sequences of three Bordetella strains, namely; B. pertussis, B. parapertussis and B. bronchiseptica (Parkhill et al., 2003).

In this study, the genomes of the three pathogens were sequenced and compared. When comparing the operons of the three strains it was found that only the operon of B. bronchiseptica -the most virulent of the three strains- was fully operational and not containing and pseudogenes or mutations.

Proteomics involves the study of proteins and involves aspects such as modelling, visualisation and comparison of proteins to determine their structures, interactions functions and investigate the levels of protein synthesis and gene expression (Cash, 2000) The area of proteomics is key in the research of microbial pathogenesis (Cash, 2003) which is enabled by a range of powerful analysis and protein modelling software as well as expansive proteomic databases. The proteome is all the proteins encoded by

the genome of a particular strain (Cash, 2000). Similarly to genomics, there are a variety of proteome databases that all have slight differences, however Prosite, Swiss-Prot and TrEMBL are three of the largest ones (Biron et al., 2006), also, the universal protein database UniProt is an attempt to combine various databases in one (Bairoch et al., 2004). These databases include basic data on the proteins such as their sequence and taxonomic (their source organism) information, as well as details of their function, their various domains, sites (binding sites etc.), of any modifications they undergo post-translation, sequence homology to other proteins and their 3D structure (Bairoch & Apweiler, 2000). A proteins structure can be useful for predicting its function. One example where protein structure was used to produce vaccines was the study carried out by Bian et al. where a modelling program known as “ TEPITOPE” was used to identify antigenic epitopes which need to be recognised by T-cells in order to carry out immune response (Bian et al., 2003).

Bacterial systematics is another area on which computational techniques have had a significant impact. It has allowed analysis of bacterial evolution, interaction and development within a community or ecosystem (Dawyndt & Dedeurwaerdere, 2007). This knowledge can then be applied to areas such as ecological and industrial research. An example of where computer assisted bacterial systematics has been used in industrial microbiology is referred to by Zhu and others, where various methods of improving the productivity of lactic acid bacteria (LAB) were explored (Zhu et al., 2009). One particular study involved the study of the interactions between two LAB strains: S. thermophilus and L. bulgaricus with the use of various

bioinformatic methods. This study revealed that the presence of one strain in a medium would be advantageous for the other strain due to the gaining of amino acids and purine via various interactions.

The examples given here represent only a small sample of the major impact computational/bioinformatic methods have had on all areas of microbiological research. It is likely that bioinformatics will continue to grow in importance and relevance to the field of microbiology in the future with the development of better software tools and improvement and growth of online databases.

- Altman, R. (1998). Bioinformatics in support of molecular medicine. Proc AMIA Symp, 53-61.
- Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28, 45-48.
- Bairoch, A., Boeckmann, B., Ferro, S. & Gasteiger, E. (2004). Swiss-Prot: juggling between evolution and stability. Brief Bioinform 5, 39-55.
- Bansal, A. (2005). Bioinformatics in microbial biotechnology–a mini review. Microb Cell Fact 4, 19.
- Bian, H., Reidhaar-Olson, J. & Hammer, J. (2003). The use of bioinformatics for identifying class II-restricted T-cell epitopes. Methods 29, 299-309.
- Biron, D., Brun, C., Lefevre, T., Lebarbenchon, C., Loxdale, H., Chevenet, F., Brizard, J. & Thomas, F. (2006). The pitfalls of proteomics experiments without the correct use of bioinformatics tools. Proteomics 6, 5577-5596.

- Bull, A., Ward, A. & Goodfellow, M. (2000). Search and discovery strategies for biotechnology: the paradigm shift. Microbiol Mol Biol Rev 64, 573-606.
- Cash, P. (2000). Proteomics in medical microbiology. Electrophoresis 21, 1187-1201.
- Cash, P. (2003). Proteomics of bacterial pathogens. Adv Biochem Eng Biotechnol 83, 93-115.
- Chakravarti, D. N., Fiske, M. J., Fletcher, L. D. & Zagursky, R. J. (2000). Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. Vaccine 19, 601-612.
- Dawyndt, P. & Dedeurwaerdere, T. (2007). Exploring and exploiting microbiological commons: contributions of bioinformatics and intellectual property rights in sharing biological information. Int Soc Sci J.
- Fleischmann, R., Adams, M., White, O. & other authors (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496-512.
- Parkhill, J., Sebaihia, M., Preston, A. & other authors (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet 35, 32-40.
- Perez-Iratxeta, C., Andrade-Navarro, M. A. & Wren, J. D. (2007). Evolving research trends in bioinformatics. Brief Bioinform 8, 88-95.
- Roos, D. (2001). Computational biology. Bioinformatics–trying to swim in a sea of data. Science 291, 1260-1261.

- Smith, T. (1990). The history of the genetic sequence databases. Genomics 6, 701-707.
- Tettelin, H. & Feldblyum, T. (2009). Bacterial genome sequencing. Methods Mol Biol 551, 231-247.
- Zavala, A., Naya, H., Romero, H., Sabbia, V., Piovani, R. & Musto, H. (2005). Genomic GC content prediction in prokaryotes from a sample of genes. Gene 357, 137-143.
- Zhu, Y., Zhang, Y. & Li, Y. (2009). Understanding the industrial application potential of lactic acid bacteria through genomics. Appl Microbiol Biotechnol 83, 597-610.