The *in silico* prediction of bacterial essential genes

I.A.M. Cooper¹ and M.L. Duffield¹

¹ Defence Science and Technology Laboratory, Department of Biomedical Sciences, Porton Down, Salisbury, Wiltshire, United Kingdom

An essential gene can be defined as a gene without which a cell is unable to survive and whose deletion or disruption results in the death of the organism. For this reason, essential genes could furnish novel drug targets for the therapy of bacterial infections. Several strategies have been utilised to identify essential genes and these generally fall into two groups; identification by laboratory methods and prediction by computational methods. This chapter will give an overview of some of the methodologies employed for the identification and prediction of essential genes, discussing the benefits and drawbacks of the methods, with particular focus on *in silico* strategies.

**Keywords** essential gene; bacteria; drug discovery; *in silico* identification

© Crown copyright 2011. Published with the permission of the Defence Science and Technology Laboratory on behalf of the Controller of HMSO.

1. Introduction

Bacterial diseases are a major contributor to global morbidity and mortality. The emergence of bacteria resistant to frontline antibiotics and the ability of bacteria to acquire and spread resistance are drivers for the development of novel antibiotics [1,2]. There are several articles and review papers in the literature that report on the steady decline over a number of decades in the number of new antibacterial agents making it through the regulatory processes to the end user (review [3]). Antibiotic discovery strategies have adapted to utilise the wealth of genetic sequence data recently available in publicly accessible databases. Over 1400 bacterial genomes have been completely sequenced [4], providing a substantial amount of data that requires the development of *in silico* techniques for the prediction, identification and selection of targets for investigation (Reviews [5,6]). Current antibiotics have often targeted proteins involved in cell maintenance or structure. However, more recently, essential genes and their protein products have become of particular interest, essentiality being one of the characteristics used for identification and prioritisation drug target by the TDR Targets database [7]. A full list of the criteria they used to prioritize targets is listed in Table 1.

<table>
<thead>
<tr>
<th>Target Feature</th>
<th>Benefit as drug target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essential for the survival of the organism</td>
<td>Disruption to the gene or the gene product results in bacterial death</td>
</tr>
<tr>
<td>Conserved in a number of different bacteria</td>
<td>Selects targets that are likely to be broad spectrum rather than species specific</td>
</tr>
<tr>
<td>Involvement in a crucial pathway</td>
<td>The pathway can be interrupted</td>
</tr>
<tr>
<td>‘Druggable’</td>
<td>Interacts and binds with a small molecule or ligand indicating there is a binding site that can be blocked</td>
</tr>
<tr>
<td>Low molecular weight and without transmembrane domains</td>
<td>Smaller proteins are more likely to be soluble and easier to purify, membrane proteins are especially difficult to purify.</td>
</tr>
<tr>
<td>Characterised protein</td>
<td>Including functional and structural characterisation further benefits molecule selection</td>
</tr>
<tr>
<td>No homology to human or other mammalian proteins</td>
<td>Since an eventual drug will be used in humans, a target specifically bacterial is preferred to avoid toxicity or detrimental interactions with human proteins resulting in side effects</td>
</tr>
</tbody>
</table>

Essential genes are by definition necessary for replication and viability, and therefore the deletion, interruption or blocking of the protein expressed by an essential gene, results in death of the organism, making them attractive targets for drug development. By identifying a protein essential to a bacteria and developing inhibitors to that protein, new antimicrobials can be identified. Essential genes that are conserved across bacterial genera have been proposed as promising candidates for broad spectrum drug targets, active against multiple bacterial species, [8–10].

©FORMATEX 2011

A. Méndez-Vilas (Ed.)
Essential genes have traditionally been identified experimentally, when researchers have repeatedly attempted to make mutants of certain genes and failed. However, as interest in essential genes has increased, a range of techniques have been developed to study them, both in silico and experimentally. There are advantages and disadvantages to each technique and this article will discuss some of these methods, and validation of the data produced.

1.1 Essential genes

The definition of essential genes varies between researchers. In its strictest sense it would refer to any gene that is Absolutely required by an organism in order to grow and survive. By this definition, essential proteins make interesting targets for the development of inhibitory compounds – blocking the function of the essential protein and resulting in bacterial cell death (rather than having a bacteriostatic effect, which is less desirable). When grown in a rich medium (one containing all nutrients required to support growth), essential genes confer a lethal phenotype regardless of the presence of the remaining genes in the organism upon deletion [11]. Culture in rich media identifies genes essential to the core of the cell’s functions [12] but may miss those genes essential for growth under specific environmental conditions, such as within insect vectors or host cells. In yeast, around 20% of genes were found to be essential under nutrient-rich conditions whilst under some environmental conditions, this number substantially increased [13].

In well characterised organisms, such as Escherichia coli, it has also been shown that the mutational technique employed can also result in wide variations when trying to determine essential genes. Two separate studies have investigated essential genes in E. coli with a degree of disparity between the numbers of essential genes identified. The studies used different strain of E. coli, MG1655 [14] and K12 [15] and different techniques for determining essential genes were also employed. This highlights that the strain and technique employed require careful consideration and the problem that essentiality may be a function of the conditions under which an organism is grown.

Computer-based in silico methods can be used to predict gene essentiality but the data set used for each prediction is a major contributor to the accuracy of the output. Publicly available databases are often unvalidated datasets where there are many putative proteins and genes that are assigned functions based on homology rather than experimentally confirmation and this can lead in inaccuracies in prediction [16].

1.2 The Database of Essential Genes (DEG)

The data from genomic screens for essential genes has been collated by the Database of Essential Genes (DEG). Originally set up with data from 7 organisms in 2004 (DEG 1.0) [17], the most current version (6.5, April 2011), now contains essential genes from 23 organisms, including eukaryotes (Table 2) and prokaryotes (Table 3) [18].

For each organism, the database holds the nucleotide and peptide sequence, a link to the Clusters of Orthologous Groups protein link at NCBI and a link to the relevant publication on essentiality. The collation of essential gene data into one database is an extremely useful research tool and has been employed in a number of studies. However, as discussed previously, the experimental conditions used by researchers can have an impact on determining whether a gene is essential under a particular set of conditions. Each dataset is determined under differing conditions and different criteria may be employed to describe an essential gene.

Table 2 Eukaryote essential gene studies from DEG [18]

<table>
<thead>
<tr>
<th>Organism</th>
<th>Method employed</th>
<th>No. of essential genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>RNAi</td>
<td>294</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Retroviral insertional mutagenesis</td>
<td>288</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>P-element insertion</td>
<td>339</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Comparative genomics(^a)</td>
<td>118</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Gene knockout(^b)</td>
<td>2114</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Mitotic recombination</td>
<td>1110</td>
</tr>
</tbody>
</table>

\(^a\) Comparison made with Mus musculus

\(^b\) Utilised only genes in the Mouse Genome Informatics database reported using “gene knockout” methods; random gene disruption, gene trap mutagenesis and targeted deletion
2. Experimental identification of essential genes

The minimal genome is the smallest set of genes an organism needs to live in a particular environment and by default each gene is considered essential [19]. Though not a new concept in life sciences this has been given more attention in recent years through the work of the J Craig Venter Institute (JCVI) with Mycoplasma [20]. Research in the field of the minimal genome concerns the question “how few genes can make a viable cell?” There are two approaches that can be taken when addressing this question: top-down and bottom-up [21]. The top-down approach involves the deletion of genes from an existing bacterial cell with the aim of reducing the genome to a simplified version and has been employed to investigate the essentiality of genes in E. coli [22]. The drawback of this method is that the order of deletion can affect the outcome of the screen and the deletion of two non-essential genes can also result in lethality. Such gene pairs are known as synthetic lethals and have been investigated in E. coli [23]. The bottom-up approach is the chemical construction of an artificial genome, as has recently been achieved by Gibson and colleagues with the construction of Mycoplasma genitalium JCVI-syn1.0 [24] and Mycoplasma mycoides JCVI-syn1.0 [20], the latter being the first self-replicating organism with a synthetic genome.

However, it is not necessary to determine the minimal genome to identify essential genes. Two main experimental methods employed for the identification of essential genes in bacteria are ‘targeted’ and ‘random’ mutagenesis. Studies in the literature utilising targeted methods vary in approach from investigating the essential status of individual/small numbers of genes to determining the total number of essential genes in an organism’s genome.

2.1 Targeted gene deletion

Targeted gene deletion is achieved using allelic replacement; some studies are by their nature small i.e. investigating the function of a single gene, whilst others are whole genome investigations where each gene is studied [15,25,26]. Targeted deletion systems fall into two categories dependent on the system used, creating marked or unmarked mutants. Unmarked mutants are desirable as the marker being used may obscure the mutant phenotype and in turn, may cause a mutant phenotype. However, marked systems are used predominantly and generally involve conferring antibiotic resistance to the mutant, replacing part, or all of the targeted gene by an antibiotic resistance gene. Failed repeated
attempts to ever recover a viable mutant can be taken as presumptive evidence that the gene in question is essential to bacterial survival [27], although cannot absolutely confirm that a gene is essential. This method has been used to identify potential essential genes across a whole genome. Baba et al [15] conducted a whole genome experimental screen in E. coli K-12 replacing all non-essential genes with a kanamycin cassette and constructing the Keio collection, thereby determining the essential genes in the bacterium. However, this method is generally not attractive for determining essential mutants as it is resource intensive requiring several rounds of attempted mutagenesis, under the screening conditions employed.

The difficulties that have been encountered by target gene deletion has been reviewed by Payne and colleagues [28], describing the seven year effort by GlaxoSmithKline in the field of antimicrobial drug discovery. From the 358 genes identified in the comparative genomics effort using Haemophilus influenza and Streptococcus pneumoiae, 127 genes were determined to be essential. However, high throughput screening against large libraries of compounds identified only 16 of these to have hits, and of these, only five resulted in leads. Further of the five, four failed as broad spectrum targets.

2.2 Conditional mutants

Confirmation of gene essentiality can be achieved through the construction of a conditional mutant, where the expression of the gene of interest is controlled through a regulatory promotor. Regulation of the promotor is used to either allow or suppress expression of the gene of interest. If bacterial growth is only achieved when gene expression is induced, then it is possible to say that under the experimental conditions used, the gene is essential. Various methods for achieving this have been described and include the use of an IPDT-inducible Poxp conditional expression system to induce, then it is possible to say that under the experimental conditions used, the gene is essential. Various methods for achieving this have been described and include the use of an IPDT-inducible Poxp conditional expression system to investigate gene inactiviation in Bacillus subtilis [29], a riboswitch system to control csrA in E. coli [30], temperature-sensitive mutants in Saccharomyces cerevisiae [31] and tetracycline regulation which has been used in both pro- and eukaryotic systems [32].

2.3 Random mutagenesis

An alternative method to study essential genes is through random gene deletion. This approach is high through-put allowing coverage of the whole genome. The experimental method most commonly utilised for bacterial random mutational studies is transposon mutagenesis. Transposons are mobile genetic segments that can disrupt gene function by their insertion into genes [33]. This approach aims to disrupt as many genes of the bacterial genome as possible and recover the resulting mutants. Through sequencing, and comparison to a completed genome sequence, the position of transposon in each mutation can be identified and any gene for which no mutant is recovered is potentially lethal. However, there are several possible drawbacks with this technique. As this technique is random, there is always the chance that some areas of DNA have no insertions and the smaller the gene, the greater this possibility. Some areas of DNA are known to be ‘hot-spots’, were a higher number than expected of insertions are recovered, so it is equally possible that some areas are corresponding ‘cold-spots’ [34]. Insertion at the N- or C-terminal of the gene may not always inactivate the gene, allowing a partial gene product to be expressed, sufficient to ensure survival, even though complete deletion of that product would prevent growth. Frameshift mutations can also occur resulting in mutations downstream of the insertion so more than one gene may be affected resulting in polar effects.

Despite these drawbacks, transposon mutagenesis has been used for mutational studies in wide a range of bacterial species [35-38] and modifications of this method, Signature Tagged Mutagenesis (STM) [39,40] and Transposon Directed Insertion-site Sequencing (TraDIS) [41] are also evident in the literature.

2.4 Identification of genes essential in pathogenesis

In order to be potential targets for the development of inhibitory compounds, genes can either be absolutely essential to the bacteria for survival under all conditions, or more specifically, essential for the pathogenesis of the bacteria and its survival within the host environment. Identification of essential genes in the laboratory under artificial culture conditions may not identify proteins required for survival during infection [12]. Other than traditional virulence factors such as toxins, essential genes, ergo essential proteins, involved in infection may be identified by using techniques such as STM. STM is a random mutagenesis technique whereby mutants are marked with uniquely marked transposons, pooled and then screened through a selective condition such as infection of an animal model. Unique tags that are present in the starting or “input” pool and not recovered in the “output” pool are therefore unable to survive the selection process (have succumbed in the animal model) and are then identified by hybridisation to arrays of signature tags. More recently the internal tag in the transposon has been replaced by using transcripts of the unique genome sequence adjacent to each transposon generated. STM combines the advantages of transposon mutagenesis with the ability to screen a larger number of mutants. A recent advancement in a study by Langridge and colleagues [41] investigated Salmonella using more than 1 million transposon mutants. Using this method, TraDIS, they were able to cover the entire chromosome of Salmonella enterica serovar Typhi with a transposon insertion occurring on average
every 13 base pairs throughout the chromosome. The technique allowed for the essentiality of every gene in the chromosome to be determined simultaneously and was validated under different culture conditions, including bile tolerance, an important mechanism in carriage of Salmonella.

One system that can specifically identify niche specific essential genes is in vivo expression technology (IVET) [42]. IVET is a promoter trap strategy employing a library of random genomic fragments ligated to a promoterless reporter gene. The library is then used to identify promoters that are active in vivo and not active in vitro through assaying the transcriptional activity of promoter–reporter fusions. Bacteria carrying promoter clones can then be expanded in vitro. Promoter characterisation is achieved through sequencing. Studies using IVET include the identification of genes from Enterococcus faecalis required for the in vivo persistence and pathogenic potential of this opportunistic pathogen [43] and the identification of genes in Lactobacillus reuteri that may be essential for growth in the gastrointestinal ecosystem [44]. A potential drawback of IVET is that genes that are transiently expressed or expressed at low levels may be missed during the IVET screen.

### 3. In silico identification of essential genes

The increasingly large data sets generated through genome and protein sequence, molecular biology, etc, are reliant on the continuing advancements in in silico techniques to make sense of and analyze this data. Several in silico analyses have also been developed to predict essential genes using a variety of criteria. Generally, larger numbers of genes are found in the leading strand, relative to the lagging strand of bacterial genomes [45] and essential genes have been found much more frequently in the leading strand than other genes, including highly-expressed non-essential genes [46,47]. However, using experimentally determined data from 10 bacteria, Lin and co-workers found that this bias was found only for essential genes with certain functionalities [47].

Several groups have used multiple criteria for prediction of essential genes including homology, functionality, GC content, protein length, codon adaptation measures, protein evolution rate, protein-interaction connectivity, gene-expression cooperativity and gene-duplication data [31, 48-50]. Whilst all predictors can provide additional information, the usefulness of this can be species dependent. Gustafson et al [51] looked at phyletic retention, gene size, codon bias and protein interactions in both S. cerevisiae and E. coli as predictors for essential genes and found phyletic retention and codon bias to be most the predictive features. Gene size was also indicative of essentiality but varied between the two species, smaller proteins being essential in E. coli and larger ones being essential in S. cerevisiae.

#### 3.1 Subtractive genomics

Subtractive genomics is a well-utilised approach to identify potential targets for antimicrobials and works by comparing all the genes or proteins from one organism to another. For human bacterial pathogens, this may be a comparison against a closely related non-pathogenic bacterium, to identify genes unique to the pathogen that may have a role in virulence, or a comparison against the human host. Any homologous genes or proteins are then removed and a further down-selection can then be employed to identify a set of genes or proteins that are likely to be essential in the pathogen (usually through comparison to DEG) [52]. The technique has been used in a number of studies to predict putative drug targets in number of bacterial species [53-62], although few of these have been tested experimentally.

Vetrivel and colleagues [63] incorporated a codon adaptation index to increase the stringency of the selection criteria for the identification of drug targets in S. pneumoniae and H. influenzae. Although novel targets were identified, many of these were hypothetical proteins which would rely on determining the mode of action/function to be determined before inhibitory compounds can be developed. While the exclusion of hypothetical proteins reduces the number of potential targets for investigation, it also reduces the burden of investigating targets where little is often known about the protein. However, although well characterised proteins with known functions are more desirable as targets, proteins of unknown function could still be investigated by cloning and expression of the protein, obtaining crystal structures and screening against a library of compounds for suitable leads [64].

#### 3.2 Metabolic networks

Re-constructing a genome scale metabolic network of an organism can be undertaken following the sequencing and annotation of the organism’s genome, providing another tool for increasing the understanding of the organism. Networks can be constructed in two ways; automated, by use of computer modelling software, or manually, by the researcher. Automated construction is far quicker than manual construction; however manual construction requires less post-construction corrective actions in comparison to those that are constructed by software. Metabolic pathways are available through sites such as the Integrated Microbial Genomes (IMG) system [65,66] and the Kyoto Encyclopaedia of Genes and Genomes [67,68]. The IMG system is a community resource for comparative analysis and annotation of publically available genomes and provides metabolic capabilities and metabolic pathways that can be analysed to identify missing enzymes in a pathway and identify candidate genes that may be compensating for the missing enzyme.
KEGG is an integrated database resource for genes and genomes and includes the KEGG pathway tool which is an online collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks.

Both automated and manual constructions rely on the accuracy of the data held in the databases being used to construct the network. As the method relies on the areas of experimental biochemistry, the technique can encounter problems when the use of standard nomenclature for annotation of reactions and metabolites is not consistent or incorrectly annotated, for example, genes that are incorrectly annotated can result in incomplete or missing pathways and software will be needed to predict gaps [69]. Computer software such as ScrumPy [70] and Cyclone [71] can be used for metabolic modelling but missing metabolites, co-factors or the incidence of enzymes catalysing more than one reaction in the organism can cause imbalances and this reduces the accuracy of the network [70]. Jamshidi and Palsson have used a network approach for the investigation of essential genes in *Mycobacterium tuberculosis* using predicted biomass as a determinant for gene essentiality [72]. Metabolic networks can also be used in conjunction with flux balance analyses to examine the effects of gene deletion on *in silico* growth of the organism for which the network has been constructed.

Flux balance analysis (FBA) is used to study biochemical networks constructed from an annotated genome of a particular organism and using the metabolic pathways described above. These pathways are then converted to constraint-based models, where the flow of metabolites through the network can be mathematically represented and controlled. FBA assumes that the metabolic network will reach a steady state that satisfies certain constraints (e.g.; mass balance and flux limitations) and maximizes biomass production. FBA is then used to measure the flow of metabolites through the network under different environmental conditions. It has a number of uses including looking at species-specific biology, identifying minimal media requirements, prediction of cellular phenotypes under different environmental conditions and exploration of metabolic weaknesses [73]. This technology also highlights gaps in networks, where genes appear to be missing, and missing networks, where no genes can be identified for known reactions [69]. By systematically altering the flow of metabolites through each pathway, the effect on other pathways and cell phenotype can be predicted [74].

By constraining the flux of each reaction in the metabolic network to zero one at a time (to simulate deletion or disruption of that network); FBA can also be used to predict essential genes. This has been reported for a number of organisms including *E. coli* [75,76,77], *Staphylococcus aureus* [78], yeast [79], *Plasmodium falciparum* [73] and most recently for the identification of target genes in *P. aeruginosa* [54]. Experimental transcriptomic and proteomic data from both wild type and adaptively-evolved *E. coli* strains have been shown to be consistent with FBA predictions and therefore supports the accuracy of these models [80]. Accuracy of the method has also been confirmed by del Rio and colleagues with up to 85% accuracy in *S. cerevisiae* [81].

4. Validation of *in silico* prediction

*In silico* approaches are typically validated by applying the *in silico* methodology to a genome with experimentally derived data on essential genes already available and comparing the predicted to the experimental output [82,49]. Few workers have tried to confirm their predictions experimentally in a species where lethal genes are not generally identified. However, in one study utilising DEG to identify essential proteins conserved across several bacterial genera to predict essential proteins in *Tversinia pseudotuberculosis*, 7 of the 8 genes tested were suggested to be essential due to the inability to create deletion mutants, demonstrating the usefulness of the predictive tools but highlighting the need to validate predictions experimentally [83].

5. Discussion

*H. influenzae*, a Gram negative opportunistic pathogen of man, was the first organism to have its entire genome sequenced [84]. Since this achievement in 1995, many bacterial species have had their entire genome elucidated and the advancement of scientific techniques in the field of genomics and the study of essential genes has expanded greatly during this time for both prokaryotic and eukaryotic organisms. However, the data that has been generated for essential genes is not uniform, with some workers interpreting essential as absolutely required for survival under all conditions, whilst others take it as meaning required under the conditions applied [85].

Whilst prediction tools for essentiality continue to be developed and improved, experimental validation of predicted essential genes is a necessity. Further, depending on the conditions used (mutagenesis, media, stress conditions, etc) and the analysis criteria applied (essential or conditionally essential), experimental datasets can also differ and may not reflect the organisms response to an environmental niche such as survival in a host cell. For drug discovery, it is important to understand bacterial survival within the host and identify genes or gene products essential under the conditions of infection.
Whilst there are many factors to be considered when determining essential genes and proteins, they continue to be of great interest to the research community as potential drug targets for bacterial pathogens and so effort into their discovery and validation is likely to continue.

Acknowledgements
This work was supported by the United Kingdom Ministry of Defence. The authors are very grateful to Petra C.F. Oyston for assistance in the preparation of this manuscript.

References


[69] Saha S & Heber S.


