Transcriptional data: a new gateway to drug repositioning?

Francesco Iorio[1,4], Timothy Rittman[2,5], Hong Ge[3,5], Michael Menden[1] and Julio Saez-Rodriguez[1]

[1] EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
[2] Dept of Clinical Neurosciences, Herchel Smith Building, Forvie Site, Addenbrooke's Hospital, Robinson Way, Cambridge CB2 0SZ, UK
[3] Dept of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, UK
[4] Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

Recent advances in computational biology suggest that any perturbation to the transcriptional programme of the cell can be summarised by a proper 'signature': a set of genes combined with a pattern of expression. Therefore, it should be possible to generate proxies of clinicopathological phenotypes and drug effects through signatures acquired via DNA microarray technology.

Gene expression signatures have recently been assembled and compared through genome-wide metrics, unveiling unexpected drug–disease and drug–drug 'connections' by matching corresponding signatures. Consequently, novel applications for existing drugs have been predicted and experimentally validated.

Here, we describe related methods, case studies and resources while discussing challenges and benefits of exploiting existing repositories of microarray data that could serve as a search space for systematic drug repositioning.

Introduction

During past decades the main strategy of drug development has been high-throughput screening of different molecules to identify lead compounds showing activity against single therapeutic targets and pathways. However, the ratio of successfully identified drugs to screened molecules has decreased dramatically over the years [1]. Furthermore, targeting individual elements of pathogenic pathways is not always a successful approach for tackling the complexities of the disease state; even when a target pathway is identified, a suitable drug might not be found. For example, in Alzheimer's disease the 'amyloid hypothesis' has driven the search for drugs that stop aggregation of pathogenic beta-amyloid, which generates potentially toxic oligomers and plaques, but so far these efforts have not led to a successful disease-modifying treatment [2]. In addition, the cost of bringing an effective drug to the market is large and growing with a significant portion of investment needed in the research and development phase [3]. Many promising molecules never come into clinical use because they show unfavourable pharmacokinetic properties or cause adverse reactions in humans. As a consequence there is a pressing need to identify successful treatments for many diseases in innovative ways that could overcome these drawbacks.

Drug repositioning [4] is a potential alternative to new drug discovery that promises to address some of these issues by identifying new therapeutic applications for existing drugs. One of the advantages of reconsidering established drugs is that they have already been approved and, hence, they can potentially be re-marketed in a faster and more cost-efficient way – by skipping Phase I clinical trials [5]. Moreover, pharma company pipelines already include many drug candidates that have passed Phase I trials but were not successful in Phase II or III (i.e. being safe but not sufficiently effective in treating the condition they were originally designed for). This implies that the search basin for repositionable drugs is vast and much larger than the set of approved drugs [6].

Corresponding author:. Saez-Rodriguez, J. (saezrodriguez@ebi.ac.uk)
[5] These authors contributed equally to this work.

 1359-6446/06/$ - see front matter http://dx.doi.org/10.1016/j.drudis.2012.07.014

Most cases of successfully repositioned drugs can be linked to serendipity, such as the classic example of sildenafil which is used to treat erectile dysfunction but was originally developed as a cardiovascular drug [7]. However, systematic approaches have recently been proposed. Most of these are based on the principle that shared properties between compounds could hint at similar efficacy or commonality in their mode of action (MoA). Successful strategies based on this assumption have been devised and published in different areas of computational drug discovery: from chemoinformatics [8] and structural bioinformatics [9] to text mining and meta-data analyses [10] and, recently, genome-wide association studies [11]. Many of these strategies benefit from recent advances in data integration and systems biology [12] and among them a new trend has emerged over the past few years that is based solely on the analysis of gene expression data [13].

The traditional 'central dogma' of molecular biology is the principle of genes encoding mRNA that is translated into proteins. This defines a biological information flow that, moving through levels of increasing complexity and emerging properties, links the underlying genetic make-up of the cell to its clinicopathological state [13]. In such a context, transcriptional profiling enables the capture of a multidimensional view of this complexity at an intermediate level, reflecting genomic and environmental effects.

So far in computational drug discovery, drug response and disease phenotypes have been correlated with underlying pathological processes through 'back-tracking' approaches that can infer primary causes of transcriptional changes but require the integration of heterogeneous data sources and *a priori* known signalling and regulatory models [14–16]. Transcriptional profiles have also been used as a single data layer to dissect drug MoA through reverse-engineering techniques [17]. By contrast, recent studies suggest that purely data-driven approaches making use of gene expression data alone are well suited to identifying new drug repositioning opportunities. The leading idea is that comparing the expression profile of a cell before and after exposure can quantitatively assess the changes brought about by active compounds on the transcriptional programme. The corresponding signature of differential gene expression (SDE) can be considered as the summary of the compound's effect. Furthermore, a drug-induced SDE can then be compared with a disease-associated SDE similarly obtained through differential expression analysis of diseased versus healthy conditions. If they are sufficiently negatively correlated (i.e. the genes upregulated in the disease SDE are downregulated in the drug SDE and vice versa) then it is reasonable to hypothesise that the effect of the drug on transcription is opposite to the effect of the disease (Fig. 1a). As a consequence, the drug might be able to revert the disease SDE and hence the disease phenotype itself [18–20]. Alternatively, from a shared SDE it can be hypothesised that two drugs could share a therapeutic application, regardless of the similarity in their chemical structure and that they impinge on different intracellular targets or pathways [21–24] (Fig. 1b).

Despite the relative simplicity of these ideas, recent applications have shown that they could serve as the basis for identifying drug repositioning opportunities in different therapeutic areas to treat heterogeneous diseases from cancer [25,26] to Alzheimer's disease [24] and Crohn's disease [27].

In the following sections we examine how gene transcription profiles have been analysed in single case studies and we will describe several publicly available resources; finally we discuss challenges and future directions.

Matching gene expression signatures to 'connect' phenotypes

Pioneering studies have shown that collections of gene sets (i.e. groups of genes sharing a common biological function, chromosomal location or regulation) can be used to interpret and extract biological insights from genome-wide expression profiles, by using parametric [28] or non-parametric statistical methods [29].

A genetic signature is defined by associating a gene set with a specific pattern of expression [30]. Gene expression profiling has been widely used as phenotype proxies [31], to build phenotype taxonomies [30,32], for systematic functional discovery [33] and for classification and/or cataloguing purposes [30,34]. Most importantly, gene expression signatures have been effective in recovering 'connections' between genes, drugs and diseases involving (or involved in) the same biological process, by combining a large collection of gene expression data following drug treatment with a pattern-matching method [35]. A seminal example of this is given by the Connectivity Map (cMap) [18,35], which is the first large public database of genome-wide gene expression profiles from five different human cancer cell lines treated with more than 1000 bioactive small molecules.

The aim of the cMap project was to generate a 'map' that can be searched for 'connections' between gene expression profiles associated with disease states and those following treatment with a large set of existing drugs. To query this map, the authors devised a pattern-matching tool based on Gene Set Enrichment Analysis (GSEA) [29] through which these connections can be inferred and statistically assessed.

The effectiveness of this method for *in silico* drug discovery and drug repositioning has been demonstrated already by its very first applications [36,37], and it highlights the potential of gene transcription profiling to serve as the common language to link chemistry, biology and the clinic, by inferring genome-wide similarities or differences [35]. Numerous studies have been published using the cMap dataset and the cMap tool, with different aims (a comprehensive list is provided on the cMap website). This underscores the power of gene expression profiles and gene signatures in characterising biological states and acting as a surrogate phenotype, despite the difficulty in interpreting the meaning of predicted associations, let alone the precise part played by individual genes in these signatures [31]. Subsequent achievements have been to characterise the whole landscape of human gene expression [38], to establish large repositories of transcriptional data [39,40] and to make publically available a large amount of gene expression data that could be mined to compose drug and disease signatures (Fig. 2). Moreover, the robustness of these signatures has been shown across tissue types and experiments [41] and, during the past two years, the use of transcriptional data for drug repositioning has emerged as a useful and effective strategy [13,42], bringing about a new dawn for the vast quantities of DNA microarray data already in the public domain.

Although numerous approaches for *in silico* drug repositioning based on gene expression data have been published [19,20,22,24,