EuPA Open Proteomics 13 (2016) 1–13

Contents lists available at ScienceDirect

EuPA Open Proteomics

journal homepage: www.elsevier.com/locate/euprot

Multi-omics "upstream analysis" of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer

Alexander E. Kel[a,b,c,*], Philip Stegmaier[c], Tagir Valeev[b,d], Jeannette Koschmann[c], Vladimir Poroikov[e], Olga V. Kel-Margoulis[c], Edgar Wingender[c,f]

[a] Institute of Chemical Biology and Fundamental Medicine, SBRAS, Novosibirsk, Russia
[b] Biosoft.ru, Ltd, Novosibirsk, Russia
[c] geneXplain GmbH, D-38302 Wolfenbüttel, Germany
[d] A.P. Ershov Institute of Informatics Systems, SB RAS, Novosibirsk, Russia
[e] Institute of Biomedical Chemistry, Moscow, Russia
[f] Institute of Bioinformatics, University Medical Center Göttingen, D-37077 Göttingen, Germany

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form 5 September 2016
Accepted 8 September 2016
Available online 9 September 2016

Keywords:
Proteomics data
Microarray data
ChIP-seq
Upstream analysis
Promoter analysis
Pathway analysis

ABSTRACT

We present an "upstream analysis" strategy for causal analysis of multiple "-omics" data. It analyzes promoters using the TRANSFAC database, combines it with an analysis of the upstream signal transduction pathways and identifies master regulators as potential drug targets for a pathological process. We applied this approach to a complex multi-omics data set that contains transcriptomics, proteomics and epigenomics data. We identified the following potential drug targets against induced resistance of cancer cells towards chemotherapy by methotrexate (MTX): TGFalpha, IGFBP7, alpha9-integrin, and the following chemical compounds: zardaverine and divalproex as well as human metabolites such as nicotinamide N-oxide.

1. Introduction

Cancer cells are currently subject of very intense studies of the molecular mechanisms of cancerogenesis. Multiple "-omics" data are generated worldwide measuring expression of proteins, miRNAs and long non-coding RNAs of the cancer cells and, as prerequisite, the epigenomic signatures of DNA methylation and various modifications of chromatin. One of the most important problems is to decipher the mechanisms how cancer cells develop resistance against chemotherapy and search for possible ways to suppress such resistance by interacting with specific molecular targets. One of the important drugs currently widely used in cancer therapy is methotrexate (MTX). Emergence of resistance to MTX of various cancer cells is one of the most important problems in the long-term application of this drug. Several authors compared MTX resistant cells with sensitive cells and generated various sets of "-omics" data [1,2]. We focused our attention on the MTX resistant cells of the colon cancer cell line HT29.

According to the classical view on the mechanism of resistance to the chemotherapy, the resistant clones/lineages are already present in the tumor tissue ab-initio (due to some randomly occurring "favorite" mutations) and get proliferated during the drug treatment while other cells die. However, more recently, a different point of view gets more and more evidences that at least in some cases the cancer cell populations experiencing transitions from a sensitive state to the resistant state during and sometime as a result of the treatment using various chromatin reprogramming mechanisms [3,4]. In this paper we follow this novel point of view and search for such specific reprograming mechanisms in the cancer cells.

Methotrexate (MTX) is a folate antagonist, which kills the proliferating cell by binding tightly to the enzyme dihydrofolate reductase (DHFR). Due to this binding the pathway of de novo DNA synthesis is blocked [1]. But continued administration to patients often results in the emergence of drug-resistance [2]. The analysis of the molecular mechanisms of the resistance can help to identify the most promising targets to combat this resistance. Numerous "-omics" studies on the molecular mechanisms of resistance offer the possibility to mine these high-throughput data by applying

* Corresponding author at: Institute of Chemical Biology and Fundamental Medicine, SBRAS, Novosibirsk, Russia.
E-mail addresses: alexander.kel@genexplain.com, alexander.kel2@googlemail.com (A.E. Kel).

http://dx.doi.org/10.1016/j.euprot.2016.09.002

computational tools and analyzing functions and regulation of the involved genes. Such "-omics" data are often deposited in databases such as ArrayExpress [5] or Gene Expression Omnibus (GEO) [6], and derived sets of differentially expressed genes (DEG) (expression signatures) can be found in more specialized databases such as the Expression Atlas [7], the Mouse Expression Database (GXD) [8] and others. These signatures can be used directly for selection of potential drug targets using the mere statistical significance of the expression changes. For a more refined analysis of the molecular mechanisms a conventional approach of mapping the DEG sets to Gene Ontology (GO) categories or to KEGG pathways, for instance by GSEA (gene set enrichment analysis), is usually applied [9,10].

Since such approach provides only a very limited clue to the causes of the observed phenomena, we introduced earlier a novel strategy, the "upstream analysis" approach for causal interpretation of the expression changes [11–13,18]. This strategy comprises two major steps: (1) analysis of promoters and enhancers of identified DEGs to identify transcription factors (TFs) involved in the process under study; (2) reconstruction of signaling pathways that activate these TFs and identification of master-regulators on the top of such pathways. The first step is done with the help of the TRANSFAC database [14] and site identification algorithms, Match [15] and CMA [16]. The second step is done with the help of the TRANSPATH database [17], one of the first signaling pathway databases available, and special graph search algorithms implemented in the geneXplain platform [18].

In this paper, we introduce two enhancements to the upstream analysis approach. First, we add a new graph-weighting schema to the algorithm of master-regulator search that enables to incorporate proteomics data by adding a "context protein" list that pushes the graph search towards those nodes that are expressed in the cell. The second improvement of the approach is an adding the option to analyse TF binding sites in potential enhancer and silencer areas of the genome that are inferred from overlapping transcriptomics and epigenomics ChIP-seq data. These two enhancements of our "upstream analysis" approach at present open the possibility to perform multi-omics studies using the geneXplain platform.

Our study revealed that the novel multi-omics "upstream analysis" approach allows to identify a number of important master regulators of MTX resistance. Among them are some that are known to play essential roles as targets for anti-cancer drug therapy and our results suggest them for the use as anti-resistance targets. These targets were used in the final step of our analysis, i.e. the identification of chemical compounds that have the potential of inhibiting or activating these targets and consequently suppressing the MTX resistance mechanisms.

In silico discovery of chemical compounds that are able to inhibit or activate given molecular targets is one of the most important problems in chemoinformatics. Most often such drug discovery attempts involve the design of molecules that are complementary in shape and charge to the target with which they are supposed to interact. This usually relies on computational molecular modeling techniques. This type of modeling is often referred to as structure-based drug design [19]. In the current work we used an alternative method called ligand-based drug design, or (Q)SAR (Quantitative) Structure-Activity Relationships, which relies on the knowledge of other molecules that bind to the biological target of interest [20]. We are using one of the most powerful instruments in this field, the computer program PASS, which is based on Multilevel Neighborhoods of Atoms (MNA) descriptors to consider the chemical structures of the known ligands of the target of interest and Bayesian approach to estimate the probability that new ligands interact with the same target [21,22]. The PASS program was trained on more the 3500 different molecular targets and can be used now to scan thousands and millions of chemical compounds and find new potential ligands for those targets.

In the current work we applied PASS for the identification of chemical compounds that have the potential to be ligands for the selected targets to combat the MTX resistance mechanisms. Among the promising compounds we found some known drugs, such as zardaverine and divalproex as well as human metabolites such as nicotinamide N-oxide.

As a conclusion, we propose a novel combination of multi-omics bioinformatics analysis with a systems biology approach to the analysis of signaling networks for predicting drug targets and with an advanced chemoinformatics approach for the identification of potentially effective chemical compounds. This approach was successfully applied to the analysis of cancer drug resistance mechanisms.

The workflow of drug target identification is freely accessible online on the geneXplain platform [23].

2. Data and methods

2.1. Microarray data, differential expression analysis

For the analysis of gene expression changes in MTX resistant cells we took publicly available microarray data from Gene Expression Omnibus (NCBI, Bethesda, MD, USA), data entry GSE11440 [24]. The authors analyzed the transcriptome of the colon cancer HT29 cells that were MTX-sensitive and compared them to MTX-resistant cells generated from the same cell line. In total 6 Affymetrix microarray experiments were done, 3 biological replicates for the sensitive cells and 3 replicates for the resistant cells.

Raw microarray data of MTX-resistant and sensitive cells, the latter being used as control in our study, were normalized and background corrected using RMA (Robust Multi-array Average). The Limma (Linear Models for Microarray Data) method was applied to define fold changes of genes and to identify the statistically significantly expressed genes using a Benjamini-Hochberg adjusted p-value cutoff ($\leq$0.05) [25].

2.2. Proteomics data

Proteomics data of the HT29 colon cancer cell line were extracted from the PRIDE database (EBI, Hinxton, UK, http://www.ebi.ac.uk/pride), with the project accession number PRD000369 (http://www.ebi.ac.uk/pride/archive/projects/PRD000369). The data were generated and analyzed in the publication [26]. The authors extracted proteins from different regions of multicell tumour spheroids grown from HT29 colon carcinoma cells. They used trypsin digestion iTRAQ 4-plex labeling, 2D separation using OffGel (24 fractions) and RP nanoHPLC, MALDI TOF-TOF MS/MS instruments to determine changes in protein expression across the regions analysed. Authors identified proteins using Mascot software version 2.2 (Matrix Science, U.K.), which compared MS/MS generated data against the Swiss-Prot 2010 human protein database containing 20473 sequences. They set Mascot search parameters for Peptide mass tolerance at 100 ppm (ppm) and MS/MS tolerance at $\pm$0.7 Da. Trypsin proteolysis (cleavage to the C-terminal side of lysine and arginine except when proline is present) was selected allowing for one missed proteolytic cleavage. A 95% confidence threshold ($p < 0.05$) was used for searching the MS/MS data, which corresponded to a Mascot score threshold of $\geq$28. We took the list of proteins (with UniProt accession numbers) from PRIDE (1107 unique accession numbers) and converted them into Ensembl genes (1109 genes). No protein quantitative data were used in our further analysis.