Gene 659 (2018) 29–36

Contents lists available at ScienceDirect

Gene

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

Computational systems biology analysis of biomarkers in lung cancer; unravelling genomic regions which frequently encode biomarkers, enriched pathways, and new candidates

Ibrahim O. Alanazi[a], Sami A. AlYahya[b], Esmaeil Ebrahimie[c,d,e,f,*], Manijeh Mohammadi-Dehcheshmeh[g]

[a] National Center for Biotechnology, Life Science and Environment Research Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia
[b] National Center for Biotechnology, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia
[c] Adelaide Medical School, The University of Adelaide, Adelaide, South Australia, Australia
[d] School of Information Technology and Mathematical Sciences, Division of Information Technology, Engineering and the Environment, The University of South Australia, Adelaide, SA, Australia
[e] Institute of Biotechnology, Shiraz University, Shiraz, Iran
[f] School of Biological Sciences, Faculty of Science and Engineering, Flinders University, Adelaide, SA, Australia
[g] School of Animal and Veterinary Sciences, The University of Adelaide, South Australia, Australia

ARTICLE INFO

Keywords:
Biomarker
Lung Cancer
MicroRNA
Systems biology

ABSTRACT

Exponentially growing scientific knowledge in scientific publications has resulted in the emergence of a new interdisciplinary science of literature mining. In text mining, the machine reads the published literature and transfers the discovered knowledge to mathematical-like formulas. In an integrative approach in this study, we used text mining in combination with network discovery, pathway analysis, and enrichment analysis of genomic regions for better understanding of biomarkers in lung cancer. Particular attention was paid to non-coding biomarkers. In total, 60 MicroRNA biomarkers were reported for lung cancer, including some prognostic biomarkers. MIR21, MIR155, MALAT1, and MIR31 were the top non-coding RNA biomarkers of lung cancer. Text mining identified 447 proteins which have been studied as biomarkers in lung cancer. EGFR (receptor), TP53 (transcription factor), KRAS, CDKN2A, ENO2, KRT19, RASSF1, GRP (ligand), SHOX2 (transcription factor), and ERBB2 (receptor) were the most studied proteins. Within small molecules, thymosin-a1, oestrogen, and 8-OHdG have received more attention. We found some chromosomal bands, such as 7q32.2, 18q12.1, 6p12, 11p15.5, and 3p21.3 that are highly involved in deriving lung cancer biomarkers.

1. Introduction

Lung cancer is the leading cause of cancer death and a major concern worldwide. Despite many studies on biomarker discovery in lung cancer, it is still in its infancy and needs reliable prognostic and diagnostic biomarkers (Bakhtiarizadeh et al., 2011). MicroRNAs, as a class of non-coding regulatory RNAs, have opened a new vista in biomarker discovery because of their regulatory roles and their ability to match many target mRNAs (Alanazi and Ebrahimie, 2016). It is estimated that more than half of human genes are under microRNA regulation (Ryan et al., 2010). Reported crosstalk between microRNAs and transcription factors and their locations as network hubs demonstrate their high biomarker potential (Alanazi and Ebrahimie, 2016). Furthermore, microRNAs have been considered as a novel class of serum biomarkers because of the unexpected high stability of microRNAs in blood and their circulating characteristics (Mitchell et al., 2008; Cho, 2010; Chen et al., 2008). Considerable attention has been paid to microRNA biomarker discovery in lung cancer (Cho, 2010), providing a considerable data for meta-analysis and computational systems biology.

An integrative approach of text mining and systems biology has proven its high potential in biomarker knowledge discovery and finding new candidates (Alanazi and Ebrahimie, 2016; Jensen et al., 2006; Fruzangohar et al., 2013a; Pashaiasl et al., 2016a; Ebrahimie et al., 2015). Network analysis, network topology analysis, genomic hotspot discovery, Gene Ontology (GO), and enrichment analysis are the most employed systems biology tools in biomarker discovery (Alanazi and

*Abbreviations:*NLP, natural language processing

[*] Corresponding author at: Adelaide Medical School, Faculty of Health and Medical Sciences, The University of Adelaide, Adelaide, South Australia, Australia

E-mail addresses: ialenazi@kacst.edu.sa (I.O. Alanazi), salyahya@kacst.edu.sa (S.A. AlYahya), esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie), manijeh.mohammadidehcheshmeh@adelaide.edu.au (M. Mohammadi-Dehcheshmeh).

https://doi.org/10.1016/j.gene.2018.03.038
Received 24 August 2017; Received in revised form 7 February 2018; Accepted 15 March 2018
Available online 17 March 2018

Ebrahimie, 2016; Jensen et al., 2006; Fruzangohar et al., 2013a; Pashaiasl et al., 2016a; Ebrahimie et al., 2015). For example, we have recently argued that apoptotic microRNAs are not distributed randomly across the genome and tend to enrich or represent specific genomic locations (chromosomal bands) (Alanazi and Ebrahimie, 2016; Alisoltani et al., 2014). Literature mining includes some tools to automate the process of the extraction and storage of biological relationships as well as the analysis and visualisation of the extracts in a biologically meaningful way such as network topology or enrichment analysis (Fruzangohar et al., 2013b). PubMed is the key source of data in many literature mining tools in biomedical research (Lu, 2011; Becker et al., 2003). Cytoscape, Ingenuity, and Pathway Studio are the main tools in network analysis (Thomas and Bonchev, 2010). Among them, Pathway Studio (Elsevier) is enriched with the Mammalian + ChemEffect + DiseaseFx Database with 284,613 entities (including 142,270 proteins), 1911 pathways, and 6.5 million of relationships, including 49,685 relations specific for biomarkers, collected by Medscan, a natural language processing (NLP) engine (Nikitin et al., 2003; Ebrahimie et al., 2016; Hosseinpour et al., 2012). "Union Selected Pathways", "Common Targets", "Common Regulators", "Common Binding Partners" and "Direct Interaction" are the algorithms for network construction implemented in Pathway Studio (Nikitin et al., 2003; Ebrahimie et al., 2016; Hosseinpour et al., 2012). Enrichment analysis is a widely used approach in many researches which highlights the enriched target based on Fisher's exact test (Nikitin et al., 2003; Hedegaard et al., 2009).

Considering the importance of biomarkers in lung cancer and the necessity of a comprehensive meta-analysis of lung cancer biomarkers, this study was performed to mine the published biomarkers in lung cancer. The mined biomarkers were further analysed in terms of genomic location enrichment, examined tissue, upstream regulator, downstream targets, and network topology. Crosstalk between microRNA biomarkers and transcription factors was also studied. Ample effort was made to pin down the biomarkers located on the upstream of the other biomarkers. In-depth literature mining, enriched with system biology analysis, provided a new insight into lung cancer biomarker mechanisms and possible new candidates.

2. Methods

2.1. Text mining process

For literature mining, we used MedScan, a Natural Language Processing (NLP) to extract relations from biomedical texts (Novichkova et al., 2003). Text mining has the following steps: (1) reading sentences in literature, (2) recognition of entities (proteins, microRNAs, lung cancer, etc.) in the sentence, (3) utilising the grammar rules to find the described relationships between entities, (4) the identification of the interaction (relation) type, and (5) adding the extracted rule to database (Novichkova et al., 2003). Medscan also records the title of literature, authors, the year of publication, Medline (PubMed) reference number, etc. Each entity in database has also a range of information such as the subcellular location (from Gene Ontology Consortium) and its class (such as receptor, ligand, transcription factor, small RNA, small molecule, etc.).

For depositing the relations, we employed Mammalian + ChemEffect + DiseaseFx database (Elsevier), which is a comprehensive dataset of protein, small molecule, disease, GO, and function gathered by NLP tool (Novichkova et al., 2003; Yuryev et al., 2009). The relations were collected from PubMed, KEGG, Science Signalling, GO Consortium, and Prolexys HyNet protein-protein interaction databases as well as the full texts of both Elsevier and non-Elsevier journals (Supplementary 1). Mammalian + ChemEffect + DiseaseFx database is updated weekly using cloud technology. The statistics of Database are presented in Table 1. Pathway Studio (Elsevier) was used to build networks and pathways from relationships of Mammalian + ChemEffect + DiseaseFx Database as previously described (Ebrahimie et al., 2015; Pashaiasl et al., 2016b; Pashaei-Asl et al., 2017; Bakhtiarizadeh et al., 2013; Alanazi et al., 2013).

Table 1
Statistics of Mammalian + ChemEffect + DiseaseFx database used for literature mining in this study.

Entity	Sub-entity	Number
Pathways		19111
Gene ontologies		28922
Entities		284613
	Proteins	142270
	Cell process	8880
	Cells	764
	Clinical parameters	4387
	Complex	559
	Diseases	15911
	Functional class	5038
	Small molecules	106732
	Treatments	72
Relations		6654660
	Binding relations	191954
	Biomarker relations	49685
	Cell expression relations	2999898
	Chemical reaction relations	54967
	Clinical trial relations	77371
	Direct regulation relations	136018
	Expression relations	609493
	Functional associations	457128
	Genetic change relations	291450
	Molsynthesis	39996
	Moltransport	140926
	Promoter binding	32842
	Protein modification relations	52881
	Quantitative change relations	289655
	Regulation relations	3866670
	State change relations	25028
	MicroRNA effects	38698

In addition, the mentioned information in the text of publications regarding the direction of relation/interaction was deposited in the database. For example, if Protein A is a transcription factor which binds to the promoter region of Protein B, the relation was recorded as "promoter binding" and Protein A was recorded in database as upstream of Protein B as:

Protein A (PROMOTER BINDING) → Pr otein B.

Recording the direction of interaction between entities provides the chance of directional systems biology analysis such as: expression downstream target analysis and upstream regulator analysis (Alanazi and Ebrahimie, 2016).

2.2. Term definition in text mining

For text mining, it is necessary to clearly define the terms to be mined in literature. We used the following name/alias and definitions for "lung cancer" and "biomarker" terms:

2.2.1. Lung cancer (MeSH heading: lung cancer, MedScan ID: 9012750) term

The following terms were mined from literature as the alias for lung cancer: Cancer of the Lung; Cancers of the Lung; Lung Cancer; Lung Cancers; Malignant Lung Neoplasm; Malignant Lung Neoplasms; Malignant Lung Tumour; Malignant Lung Tumours; Malignant Lung Tumour; Malignant Lung Tumours; Malignant Neoplasm of the Lung; Malignant Tumour of the Lung; Malignant Tumour of the Lung; Pulmonary Cancer; Pulmonary Cancers; carcinogenesis of the lung; lung cancerogenesis; lung carcinogenesis; lung carcinogenesis; lung carcinogenesis; lung malignancies; lung malignancy; malignancies of lung; malignancy of lung; malignant neoplasms of the lung; malignant tumours of the lung; malignant tumours of the lung; pulmonary