Evaluation of protein clustering of pancreatic cancer

Akram Safaei¹, Mostafa Rezaei Tavirani², Afsaneh Arefi Oskouei³, Mona Zamanian Azodi², Seyed Reza Mohebbi⁴

¹Students' Research Committee, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
²Proteomic research center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
³Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
⁴Basic and Molecular Epidemiology of Gastrointestinal Disorders Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Reprint or Correspondence: Mostafa Rezaei Tavirani; PhD Proteomic research center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
@ Tavirany@yahoo.com.

ABSTRACT

Pancreatic cancer (PC) is one of the most malignant tumors with a very poor prognosis. Late diagnosis and lack of effective therapeutic method is the weak point in pancreatic cancer therapy. For pancreatic cancer clustering analysis, thirty proteins were selected from Wiki pathway and sequence comparison was prepared by Needleman-Wunsch algorithm. Similarity scores were obtained by comparing of these proteins and the relevant dissimilarity matrix was constructed. Graphical illustrations of clustering data were performed by R software using package cluster to find subgraphs with maximal density. Finally the interactions of these proteins with their neighbors that was obtained from Mentha, Reactome-Fls Databases by the application of PSICQUIC source in Cytoscape software are presented. The result of this study showed that CREB-binding protein and epidermal growth factor receptor are reported in Wiki Pathway Androgen receptor and most significant hub that involved in pancreatic cancer, but not in the high close relationship. In conclusion cell cycle and signaling pathway have highlighted its importance in pancreatic cancer, as well as its potential as a therapeutic target in PC.

Keywords: Clustering, Pancreatic cancer, Protein-protein interaction, Cytoscape, Quick GO.

Received: 13 March 2016   Accepted: 19 May 2016

Introduction

Pancreatic cancer (PC) is one of the most mortal caners among human. Infrequent, it takes up 3 % of all reported cases of cancer it also eighth most common cause of cancer-related deaths in worldwide (1-3). For researches it's important to find effective methods to pancreatic cancer treatment because most of patients die in the first 12 months and only 4 percent survive 5 years after diagnosis (4). Late diagnosis and lack of effective therapeutic for pancreatic cancer are related to aggressive biological phenotype of this disease from the beginning of the formation and development and rapid metastatic extension and is relatively resistant to radiation therapy and/or
chemotherapy (5). To improve the treatment for patients with pancreatic cancer, the diagnosis should be made at an early stage which it requires to specific and sensitive biological markers (6). Accordingly it’s an urgent requirement to promote knowledge from molecular basic of pancreatic cancer to find appropriate treatment (7, 8). Recently, researches focus on molecular alteration in pancreatic cancer (9). It has been suggested the development of pancreatic cancer need special mutations in multi-step processes (10). Chromosomal instability, polyploidy and aneuploidy are common features of many tumors including pancreatic cancer (11).

Several pancreatic cancer genes have been established, including both low-penetrance genes such as the ABO blood group locus (12) and high-penetrance genes such as PALB2 (13), BRCA2 (14), STK11 (15) and p16/CDKN2 (16). Also, some protein alterations in pancreatic cancer has been reported in tissue (S100A11, galectin 3) (17), serum and plasma (APOC1, APOC2, PERK1/2, Pca-AMP, HSP27) (19) and in urine (annexin A2, gelsolin, CD59) (20). The importance of human molecular interaction networks not only summarization in reveals protein function interaction between two or more proteins but also it clear view of fundamental human biology as well as disease progression, diagnosis (21). Therefore, it is crucial to establish high sensitive algorithms to evaluate data to the better understanding of functional association of the disease protein (22). Data mining is the computational process of discovering patterns in large data sets which its goal is to make understandable structure of extracted information from a data set for further use (23). Clustering is the task of grouping a set of similar objects in the same group so that the grouped objects are more similar than to those in the others. In fact, it is the main task of discovering data mining (24, 25). This algorithm plays an essential role in the biological networks analysis, and it can be used to uncover functional modules and acquire concepts about the cellular organization (26). In previous study, it has been shown PPI network analysis confirm that the proteins with similarity in sequence and onthology are also in the close relationship (27).

The aim of present study was to evaluate the relationship between similarity in sequence and biological process via cluster analysis and also the assessment of their interactions to obtain a better understanding of the molecular basis of the pancreatic cancer to find sensitive points to introduce possible biomarker(s) and drug target discovering.

**Materials and Methods**

Proteins and their relationship to pancreatic cancer were identified and analyzed. Proteins were selected from Wiki pathway and then their human UniProt codes were retrieved for extracting protein sequences from UniProt database and then, sequence comparison was done by Needleman-Wunsch algorithm (pair-wise alignment) between each set of two proteins available in EBI Database. Similarity scores were obtained by comparing of these 30 proteins and the relevant dissimilarity matrix was constructed that use for clustering analysis. Clustering can be useful for understanding the annotation of protein molecules and their relationship to many diseases (28). Agglomerative hierarchical clustering analysis (bottom - up) was conducted.

Graphical illustrations of clusters of data were performed by R software using package cluster to find subgraphs with maximal density. R software is a free statistical software available on (www.r-project.org). In this study silhouette plots and
dendrogram are the graphical representations. The data was analyzed by correlation distance mode. UPGMA (Unweighted Pair Group Method using arithmetic Averages) is the method that used to define the distance between clusters as the average distance (29). It is probably the most popular algorithm for hierarchical data clustering, especially in computational biology (30).

Furthermore, the annotation terms based on (GO) features was obtained using Quik GO (https://www.ebi.ac.uk/QuickGO) (31). In this study sequence closed protein for hubs collected in table 4 and their biological process was brought from Quik GO and the origin pathway surveyed in Reactome database (http://www.reactome.org/PathwayBrowser). Cytoscape v:3.2.1 also was used for the visualization of predicted interactions of the proteins (32).

Results

Clustering, as an appropriate statistical procedure, is the essential task in automatic processing that used to natural groupings of data items. Hierarchical clustering methods aim to categorize data items into a hierarchical set of clusters organized in a tree structure (30). UPGMA is the most popular hierarchical clustering algorithm in use for protein sequence clustering (33) and gene expression (34).

Enrichment analysis of the interest proteins can reveal the mechanism of the disease pathology by classification of gene/proteins according the roles in cell and discovering the shared functions between genes/proteins (35, 36). Among enrichment analysis tools the Gene Ontology (GO) is the one of the most common use to underlying functions (37). Related proteins to pancreatic cancer that extracted from Wiki Pathway database has been shown in Table 1.

![Figure 1. Dendrogram clustering plot of global pair-wise alignment between each pair sets of investigated proteins of the pancreatic cancer. The agglomerative coefficient (AC) is 0.31 for this analysis.](image1)

![Figure 2. Silhouette plot presentation of global pair-wise alignment between each pair sets of investigated proteins of the pancreatic cancer, the agglomerative coefficient (AC) is 0.79 with k=3.](image2)

Pair-wise alignment among of pairs of these 30 proteins was performed and clustering method
was performed based on sequence comparison. Dendrogram clustering plot of protein sequences is presented as figure 1,2.

Agglomerative coefficient is 0.31. The similarity is the main concept for sequence evaluation. In figure 3, the interactions of 30 proteins with their neighbors are presented by the application of PSICQUIC source in Cytoscape software. To introduce the important key proteins in the network achieve by the use of Network Analyzer.

Protein-protein interactions are intrinsic to virtually every cellular process (38) and was applied for network relationship study and can make spatial proteins an attractive source of therapeutic targets (39). The integrated network

<table>
<thead>
<tr>
<th>row</th>
<th>UniProt Code</th>
<th>Gene Name</th>
<th>Protein Name</th>
<th>row</th>
<th>UniProt Code</th>
<th>Gene Name</th>
<th>Protein Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P62834</td>
<td>RAP1A</td>
<td>Ras-related protein Rap-1A</td>
<td>16</td>
<td>P19525</td>
<td>EIF2AK2</td>
<td>Interferon-induced, double-stranded RNA-activated protein kinase Apoptosis-stimulating of p53 protein 2 Androgen receptor</td>
</tr>
<tr>
<td>2</td>
<td>P42345</td>
<td>MTOR</td>
<td>Serine/threonine-protein kinase mTOR</td>
<td>17</td>
<td>Q13625</td>
<td>TP53BP2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>P25963</td>
<td>NFKBIA</td>
<td>NF-kappa-B inhibitor alpha</td>
<td>18</td>
<td>P10275</td>
<td>AR</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>O15111</td>
<td>CHUK</td>
<td>Inhibitor of nuclear factor kappa-B kinase subunit alpha</td>
<td>19</td>
<td>Q92793</td>
<td>CREBBP</td>
<td>CREB-binding protein</td>
</tr>
<tr>
<td>5</td>
<td>Q15078</td>
<td>CDK5R1</td>
<td>Cyclin-dependent kinase 5 activator 1</td>
<td>20</td>
<td>P06493</td>
<td>CDK1</td>
<td>Cyclin-dependent kinase 1</td>
</tr>
<tr>
<td>6</td>
<td>Q8N726</td>
<td>CDKN2A</td>
<td>Tumor suppressor ARF</td>
<td>21</td>
<td>P00533</td>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
</tr>
<tr>
<td>7</td>
<td>P38936</td>
<td>CDKN1A</td>
<td>Cyclin-dependent kinase inhibitor 1</td>
<td>22</td>
<td>Q01094</td>
<td>E2F1</td>
<td>Transcription factor E2F1</td>
</tr>
<tr>
<td>8</td>
<td>P08183</td>
<td>ABCB1</td>
<td>Multidrug resistance protein 1</td>
<td>23</td>
<td>Q03468</td>
<td>ERCC6</td>
<td>DNA excision repair protein ERCC-6 M-phase inducer phosphatase 3</td>
</tr>
<tr>
<td>9</td>
<td>P17936</td>
<td>IGFBP3</td>
<td>Insulin-like growth factor-binding protein 3</td>
<td>24</td>
<td>P30307</td>
<td>CDC25C</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>P10636</td>
<td>MAPT</td>
<td>Microtubule-associated protein tau</td>
<td>25</td>
<td>O14757</td>
<td>CHEK1</td>
<td>Serine/threonine-protein kinase Chk1</td>
</tr>
<tr>
<td>11</td>
<td>P01034</td>
<td>CST3</td>
<td>Cystatin-C</td>
<td>26</td>
<td>P52657</td>
<td>GTF2A2</td>
<td>Transcription initiation factor IIA subunit 2 Prostaglandin G/H synthase 2</td>
</tr>
<tr>
<td>12</td>
<td>Q6LD43</td>
<td>GA</td>
<td>GA protein</td>
<td>27</td>
<td>P35354</td>
<td>PTGS2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>P51948</td>
<td>MNAT1</td>
<td>CDK-activating kinase assembly factor MAT1</td>
<td>28</td>
<td>P24522</td>
<td>GADD45A</td>
<td>Growth arrest and DNA damage-inducible protein GADD45 alpha</td>
</tr>
<tr>
<td>14</td>
<td>P15531</td>
<td>NME1</td>
<td>Nucleoside diphosphate kinase A</td>
<td>29</td>
<td>P49959</td>
<td>MRE11A</td>
<td>Double-strand break repair protein MRE11A</td>
</tr>
<tr>
<td>15</td>
<td>Q8WUF5</td>
<td>PPP1R13L</td>
<td>RelA-associated inhibitor</td>
<td>30</td>
<td>O96017</td>
<td>CHEK2</td>
<td>Serine/threonine-protein kinase Chk2</td>
</tr>
</tbody>
</table>

Table 1. Proteins involved in pancreatic cancer achieving from Wiki pathway

Figure 3. Protein-protein interaction network analysis of 30 proteins using Cytoscape v.3.2.1 software derived from Mentha, Reactome-Fls public databases.

was obtained from Reactome-Fls, MINT by the application of Proteomics Standard Initiative Common QUery InterfAce (PSICQUIC) source. As it is shown in table 2, based on centrality parameters of network (Degree and Betweenness) Tumor suppressor ARF, Transcription factor E2F1, Inhibitor of nuclear factor kappa-B kinase subunit alpha, NF-kappa-B inhibitor alpha, Serine/threonine-protein kinase mTOR, Cyclin-dependent kinase 1, Cyclin-dependent kinase inhibitor 1, Androgen receptor, CREB-binding protein and Epidermal growth factor receptor are identified as hubs of network (table 2).

The extracted proteins are close distance with hubs. The biological process and pathway of closed proteins with hubs in the cluster was collected (table 3). Their biological process was brought from Quick GO and the origin pathway surveyed in Reactome database.

Table 2. Hub proteins with significant centrality values, based on two fundamental centrality properties Degree and Betweenness.

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Degree</th>
<th>Betweenness centrally</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor suppressor ARF</td>
<td>422</td>
<td>0.05268923</td>
</tr>
<tr>
<td>Transcription factor E2F1</td>
<td>519</td>
<td>0.05897625</td>
</tr>
<tr>
<td>Inhibitor of nuclear factor kappa-B kinase subunit alpha</td>
<td>568</td>
<td>0.05389607</td>
</tr>
<tr>
<td>NF-kappa-B inhibitor alpha</td>
<td>612</td>
<td>0.05496298</td>
</tr>
<tr>
<td>Serine/threonine-protein kinase mTOR</td>
<td>664</td>
<td>0.07942783</td>
</tr>
<tr>
<td>Cyclin-dependent kinase 1</td>
<td>771</td>
<td>0.17460146</td>
</tr>
<tr>
<td>Cyclin-dependent kinase inhibitor 1</td>
<td>938</td>
<td>0.11365093</td>
</tr>
<tr>
<td>Androgen receptor</td>
<td>1015</td>
<td>0.13463114</td>
</tr>
<tr>
<td>CREB-binding protein</td>
<td>1244</td>
<td>0.17327638</td>
</tr>
<tr>
<td>Epidermal growth factor receptor</td>
<td>2291</td>
<td>0.44109984</td>
</tr>
</tbody>
</table>

Discussion

Pancreatic cancer is one of the aggressive malignancy with a very poor prognosis (40). Late diagnosis and lack of effective therapeutic for pancreatic cancer are one of the essential concerns in medicine. Therefore introducing biomarkers in early diagnosis or drug targets can be helpful in predicting progression and in time therapeutic proceeding.

Diagnostic biomarkers and therapeutic targets can be accelerated by targeting the specific hubs (proteins with the high degree) (41). Protein Networks analysis provide a model that elevate systems-level understanding of the mechanisms of diseases (42, 43) and discovering novel network-based biomarkers (44) to analysis therapeutic drugs and their targets (45-47). In fact, The importance of PPI analysis has been reported in cancer-related protein (41, 48).

Enrichment analysis methods have been developed for the investigation in the set of disease-related genes and find common relations and can be helpful in understanding of
the disease pathology mechanism. The use of annotation methods for example mapping genes/proteins by gene ontology (GO) can be helpful in understanding and gain a better view of biological features of the interest sets of protein (49). Cluster analysis can classify...
proteins based on distinct similarities (28). Here related proteins are subdivided groups in via this method. In first, extracted proteins from Wiki Pathway in pancreatic cancer as shown in Table 1 were analyzed. Clustering analysis of the proteins according to sequence pair-wise alignment was done and PPI network was shown. According to proteins that involved in pancreatic cancer and reported in Wiki Pathway Androgen receptor, CREB-binding protein and epidermal growth factor receptor are most significant hubs but not in the high close relationship in sequence clustering. In this study, epidermal growth factor receptor has highest significant centrality values between hub proteins. Epidermal growth factor receptor (EGFR) is a member of the erbB/human epidermal growth factor receptor family of tyrosine kinases that transmit a growth-inducing signal to cells (50). Evidence suggests that over expression of EGFR is involved in the pathogenesis and progression of different carcinoma types and it able to induce cell transformation and metastasis, resistance to chemotherapy (51, 52). Overexpression of EGFR may be detected in up to 90% of pancreatic tumors (53). Another hub is Androgen Receptor that in previous studies have been reported the relationship of the androgen receptor (AR) with the pancreas carcinogenesis (54, 55). It has been suggested that The growth of pancreatic adenocarcinoma may be under the control of the sex steroid hormone testosterone there for this relationship may be acceptable (55, 56). The other hub is CREB protein that carries out its function by activating transcription via interaction with transcription factors. It plays critical roles in growth control, embryonic development and homeostasis by coupling chromatin remodeling (57). There have been relatively few studies addressing the pathobiology of pancreatic cancer and its association with CREB (58). In this study closed proteins in sequence for hubs was collected and their biological process information was extracted from Quick GO. The common pathways that pancreatic cancer hubs involved include of gene expression, cell cycle, signaling pathway (table 4). For example the most proteins that related with AR in sequence clustering, involved in signal transduction (table 4). Since, cell cycle and signal transduction are associated with control of cell growth and proliferation , they can play a significant role in tumorigenesis (59, 60).

The results show the closed proteins in sequence may participate in the same biological process. On the other hand, the most proteins with closed distance with hubs in clustering, regulate the events of cell cycle, apoptotic process and signal transduction. It also proven that regulation of cell cycle process (61, 62), signal transduction (63-65) and apoptotic process (66, 67) are related to pancreatic cancer. The knowledge about the complex pathways involved in cancer has strongly increased(68) and new therapeutic approaches based on this knowledge can develop(69). However, to unravel the possible role(s) of these proteins in pancreatic cancer, further investigations are needed.

Conclusion
The pathways that pancreatic cancer hubs involved in gene expression are included, cell cycle and signaling pathway. In this work, it has been visualized not only complex interactions
among the pancreatic cancer proteins involved but also comprehend their relative importance in the network based on sequence clustering. Cell cycle and signaling pathway have highlighted its importance in pancreatic cancer, as well as its potential as a therapeutic target in PC. The results show the closed proteins in sequence may participate in the same biological process. Such identification can promote a better understanding of the underlying disease process and also identify specific gene targets for therapy.

**Conflict of Interest**
The authors declare no conflict of interest.

**References**
20. Weeks ME, Hariharan D, Petronijevic L, Radon TP, Whiteman HJ, Kocher HM, et al. Analysis of the urine...
Evaluation of protein clustering of pancreatic cancer

47. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with...
application to personalized medicine. Mol Sys Biol 2011;7:496.


