BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(16)

Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures

Ulykbek Kairov[1, 2], Tatyana Karpenyuk[1], Erlan Ramanculov[2] & Andrei Zinovyev[3, 4, 5]*

[1]Kazakh National University after Al-Farabi, Almaty, Kazakhstan; [2]National Center for Biotechnology of the Republic of Kazakhstan, Astana, Kazakhstan; [3]Institute Curie, Paris, France; [4]INSERM U900, Paris, France; [5]Mines ParisTech, Fontainebleau, France; Zinovyev A – Email: andrei.zinovyev@curie.fr; *Corresponding author

Received August 19, 2012; Accepted August 20, 2012; Published August 24, 2012

Abstract:
Many genome-scale studies in molecular biology deliver results in the form of a ranked list of gene names, accordingly to some scoring method. There is always the question how many top-ranked genes to consider for further analysis, for example, in order creating a diagnostic or predictive gene signature for a disease. This question is usually approached from a statistical point of view, without considering any biological properties of top-ranked genes or how they are related to each other functionally. Here we suggest a new method for selecting a number of genes in a ranked gene list such that this set forms the Optimally Functionally Enriched Network (OFTEN), formed by known physical interactions between genes or their products. The method allows associating a network with the gene list, providing easier interpretation of the results and classifying the genes or proteins accordingly to their position in the resulting network. We demonstrate the method on four breast cancer datasets and show that 1) the resulting gene signatures are more reproducible from one dataset to another compared to standard statistical procedures and 2) the overlap of these signatures has significant prognostic potential. The method is implemented in BiNoM Cytoscape plugin (http://binom.curie.fr).

Background:

The most common result of analysis of high-throughput data in molecular biology represents a global list of genes, ranked accordingly to a certain score. The score can be a measure of differential expression, distance from a cluster center, contribution to a classifier or any other score. Many methods were developed for estimating a statistically justified threshold for the score value used to select a number of top-scored genes, using purely statistical approach without taking into account the functional relations between genes, such as physical interactions between their products. The derived in this way gene signatures are used, for example, for predicting outcome of treatment in cancer therapies **[1-4]**. A predictive signature is capable to give a prognosis on whether a patient will develop metastases or not after the surgery and chemio-, radio- or other forms of adjuvant therapies. One of the major problems with computing predictive gene signatures is in observation that various signatures obtained on different cohorts of patients studied in similar conditions, but in different hospitals, have little overlap **[5]**. This reduces their cognitive value since one can not claim that the genes selected for the signature represent molecules driving the disease.

Several efforts have been made in attempt to take into account physical interactions between gene products at the top of the ranked list of genes. For example, in **[6]** network signatures of breast cancer metastases were derived using protein-protein interaction (PPI) database in combination with differential gene expression values. Various machine learning frameworks were developed in order to include network information into the analysis of gene expression data **[7-9]**. Several attempts of meta-analysis of gene signatures were made for multiple cancer studies **[10, 11]** and for breast cancer in particular **[12, 13]**, finding recurrent patterns appearing in them (for example, the role of proliferation, RNA splicing, immune response genes). Here we suggest a new method, called Optimally Functionnaly Enriched Network (OFTEN) for associating a ranked list of genes with a network of protein-protein interactions. The genes forming this network are not necessary the most top-ranked genes in the list, but they represent a compact functionally related group of genes having relatively high scores.

OFTEN-analysis is inspired by the idea of percolation in graph theory. Given a connected graph and k randomly selected nodes, one may estimate the expected size of the largest connected component formed by these genes. For many type of graphs, the typical behavior is the following: at some critical k_{crit} number of nodes, most of them start to be connected in a large connected component. Our own estimation of the critical value k_{crit} for the graph of protein-protein interactions of the Human Protein Reference Database (HPRD) **[14]** approximately equals 1500 nodes. This means that if the first $k << k_{crit}$ top-ranke genes form a relatively large connected component (compared to the 1) randomly expected), their distribution on the graph of protein-protein interactions is highly non-random and they form a tightly connected functional group. We estimate the statistical significance of appearance of such a connected component, using proper random sampling strategy which conserves the degree distribution of the selected genes in the PPI network.

Finding OFTEN network associated with a ranked list of genes allows solving two important problems: (1) Detect the optimal number of genes to select based on their distribution in the global PPI network; (2) Detect the functional "core" at the top of the ranked list of genes which, however, not necessarily formed by the most ranked genes (which can be located at the very top because of pecularities of the statistical method or irreproducible features of biological sample collection). As a result, we expect OFTEN networks obtained for datase representing independent cohorts of patients to be more reproducible than the gene signatures obtained by naive 2) selection of the most top-ranked genes. We show that this is the 3) case using four independent breast cancer datasets and computing ranked lists of differentially expressed genes between the therapeutic success (absence of metastases and death from cancer in 5 years after the treatment) and the therapeutic failure (appearance of metastases and/or death from cancer in the first 5 years after the treament).

Our meta-analysis is based on finding the overlap between OFTEN networks found in independent datasets and provides a highly reproducible network which contains many known cancer driver genes involved in developing metastases and also new genes which are "guilty by association" in malignant tumorigenesis. We believe that this META-OFTEN network is a valuable tool for interpreting predictive gene signatures of breast cancer treatment.

Methodology:

Four publicly available microarray gene expression datasets (GSE1456, GSE2034, GSE2990, GSE3494) were used to compute the ranked lists of differentially expressed genes between those tumours that developed metastases and those that did not in five years following the treatment. In one dataset (GSE3494), only the survival clinical data were available. In this case, we assumed that the death caused by cancer was the result of appearance of metastases: an assumption which is justified by high co-occurence between 5-years survival and 5-years appearance of metastases: an assumption which is justified by high co-occurence between 5-years survival and 5-years apperance of metastases in two other datasets where boht data were available. We used HPRD version 9 database as a source of protein-protein interactions in human cells. For constructing the interaction graph, we used all binary protein interaction part of the database. In addition, the protein relations inside protein complexes were used. A complex was represented as a full clique of interactions between its components which added additional 9% of connections to the graph. The largest complex with id=COM_2971 was excluded from the database because of its anomalously big size. The whole interaction graph was prepared as a file which can be imported into Cytoscape software **[15].**

OFTEN analysis

For k top-ranked genes, we map them on the interaction graph. Let us assume that k' of them are found in the interaction graph. All connections between them are extracted, forming a subnetwork. (1) The largest connected component is extracted from the subnetwork and its size $C(k')$ is recorded; (2) k' genes are randomly sampled from HPRD preserving the connectivity distribution of the k' genes from the ranked list. $R(k')$ is the size of the largest connected component; (3) Step 3 is repeated 10000 times. As a result, the following percolation score is computed, $S = (1/k')(C(k')\text{-Mean}(R(k')))$, where Mean($R(k')$) is the mean value of the randomly formed largest connected component size; (4) Steps 1-3 are repeated for a range of values of k, and k_{opt} is estimated, which corresponds to the end of the plateau after which the score goes down **(Figure 1A);** (5) OFTEN network is the largest connected component of a subnetwork formed by k_{opt} top-ranked genes; (6) OFTEN analysis is implemented as a part of BiNoM Cytoscape plugin **[16]**. Example of the $S(k)$ dependence for one of the dataset is shown in **(Figure 1)**.

Four OFTEN networks were constructed for the datasets described earlier. The META-OFTEN network is formed by those nodes which appear at least in two OFTEN networks out of four **(Figure 1B)**. Genes of the META-OFTEN network were used to derive a simple score to predict development of metastases in unsupervised manner, as in **[17]**. We computed the first principal component of gene expression data matrix, using the META-OFTEN genes. The risk of metastases risk score for a sample is the contribution of this sample to the first principal component. The patients were separated in three groups, with low, intermediate and high score values. Kaplan-Meyer survival curves for these three groups for the GSE2034 dataset are shown in **(Figure 1C).**

Discussion:

OFTEN networks represent the functional "cores" found in the ranked lists of differentially expressed genes. We can say this, because interaction between gene products or their participation in the same complex is a natural indicator of functional proximity. Unlike the standard enrichment analysis, the OFTEN network does not use pre-defined borders of pathways and ontology gene sets but uses the whole global protein-protein interaction network. OFTEN-analysis automatically detects the optimal number of genes to select from the ranked gene list.

Four OFTEN networks computed for four independent breast cancer datasets show significant overlap: in average, 65% of nodes of each OFTEN network are found in at least one another

OFTEN and 25% of nodes are found in at least two others OFTENs. Gene signatures extracted from the lists of differentially expressed genes using the standard statistical approach usually show much more modest overlap (not more than few percents, see **[5]**). For example, in our analysis there are only 2 genes (RACGAP1 and RRM2) found in common between the top 100 (average size of the extracted OFTEN network) differentially expressed genes, and another 4 genes (DTL, NEK2, UBE2S and ZWINT) are found in at least three datasets.

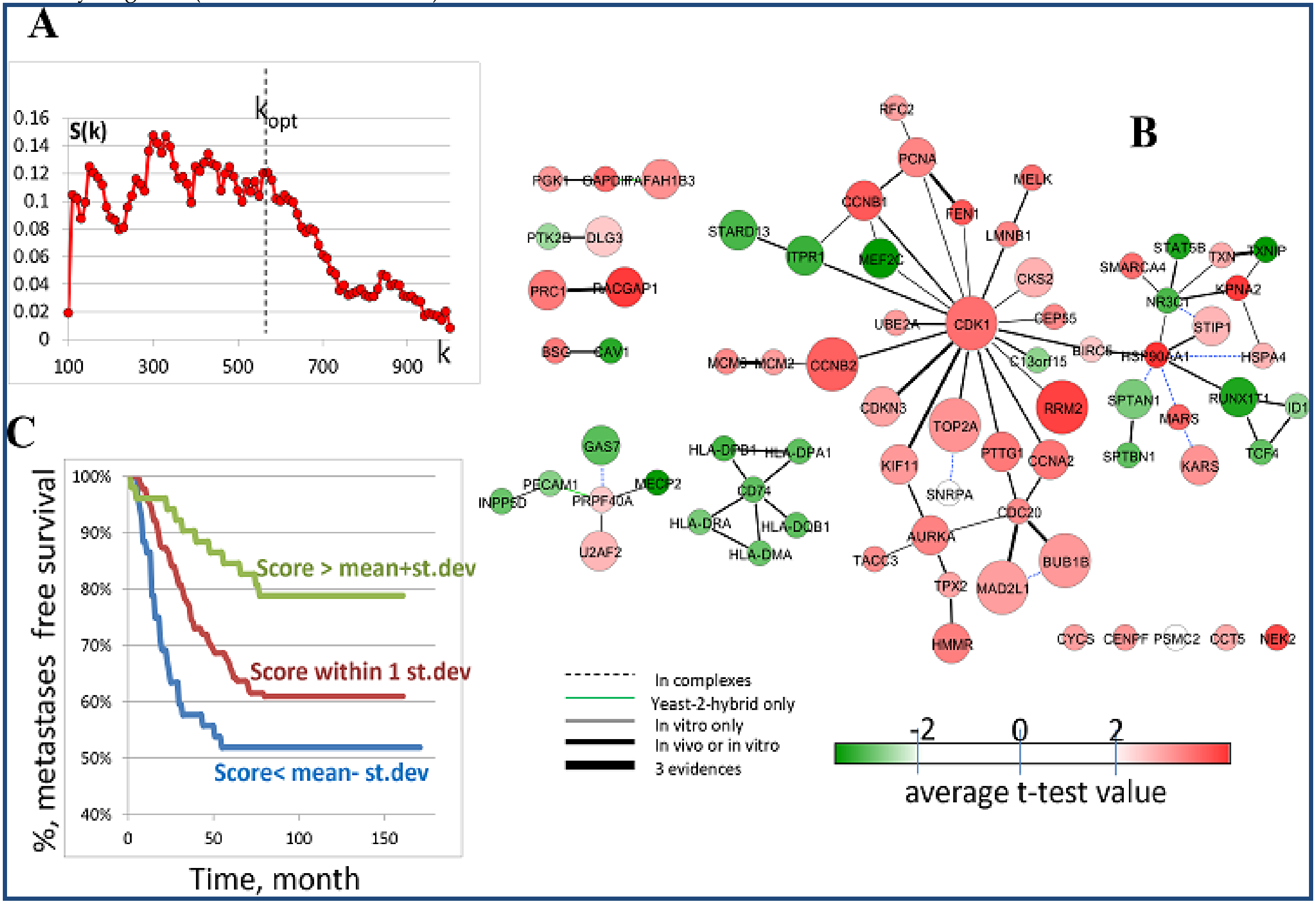

Figure 1: OFTEN analysis and META-OFTEN network of genes differentially expressed between metastatic and non-metastatic patient samples. **A)** Example of the percolation score S behavior with respect to the number of chosen top-ranked differentially expressed genes (GSE2034 dataset). **B)** META-OFTEN network constructed for four breast cancer datasets. Different edge types correspond to different evidences of protein-protein interactions as described in HPRD. Color shows average t-test values over those datasets where the gene is included in OFTEN network. Size of the node signifies the number of OFTEN networks in which the gene is found: small circles correspond to two datasets, average – to three, big nodes appear in all four datasets. **C)** Survival analysis made on the genes of the network, using unsupervised scoring strategy. The plot shows percentage of metastases-free survival for three groups of patients: with high, intermediate (within one standard deviation around the mean value) and low score values.

Nodes of the META-OFTEN network are organized into several functional subgroups, represented by network modules and containing many genes known to be implicated in tumorigenesis. The most evident component, as expected, is related to the regulation of cell cycle, especially in its G2 and M phases: CDK1, CCNB1/2, CCNA2, CDKN3, CDC20, AURKA are the classical cell cycle genes mentioned as frequent components of prognostic breast cancer signatures **[12]**. In the same component one can find various genes involved in cell cycle checkpoints and DNA replication such as MAD2L1, BUB1B, TOP2A, RRM2, PTTG1, MCM2/6. Most of the genes in the central cell-cycle related component are upregulated in metastatic tumours, indicating at more intense proliferation. At the same time, few genes are downregulated such as potential tumour suppressor and cell motility regulator STARD13 and two genes MEF2C and ITPR1 whose role in breast cancer is not yet well-established (however, they have known associations with other genetically transmitted disorders).

Another component, connected to the central one through BIRC5 protein (regulator of apoptosis) contains many stress-response genes, in particular, related to heat-chock response (HSP90AA1, HSPA4, STIP1), nuclear receptor signaling (NR3C1, KPNA2, SMARCA4), redox reactions (TXN, TXNIP), various regulators of transcription and epigenetic regulation (RUNX1T1, STAT5B, TCF4, ID1). The role of many of these genes in tumorigenesis is not yet clearly established though few are known to be associated with breast cancer (such as ID1 and STAT5B). Another component of the META-OFTEN network contains only downregulated components of the major histocompatibility complex (CD74 and various HLA-genes) responsible for presenting antigens to the immune system. Apparently, this indicates an importance of escaping immune response by tumor cells during distant tissue invasion: however, this mechnism seems not to be well characterized. Recently, CD74 was suggested as a promising therapeutic target for cancer treatment **[18]**.

A rather misterious component of META-OFTEN contains upregulated genes PRPF40A and U2AF2, and downregulated GAS7, PECAM1, MECP2 and INPP5D. Some of these genes are expressed only in hematopoetic cells, and many are involved in blood cell differentiation and migration. The role of this component in metastasation is obscure. The remaining components are small ones and represent some single reproducible interactions and individual genes. Among the most reproducible, there is involved in cytokinesis interaction between PRC1 and RACGAP1. Nodes of the META-OFTEN network can be ranked with respect to their role in forming the structure of the graph. For example, genes CCNB2, TPX2, CD74 have the highest relative connectivity (ratio of the connectivity in the network and the global connectivity in the global PPI network **[19]**), while CDK1, HSP90AA1, BIRC5 have the highest inbetweenness values (they can be classified as "routers" or "bottlenecks" as in **[9]**).

The survival analysis we have performed, using the META-OFTEN genes and the score from unsupervised analysis, shows that the set of genes in the network has significant prognostic potential. For example, if the same score as ours is derived from the set of 70-gene signature by van't Veer **[1]**, then the META-OFTEN set shows clear superiority. Comparison with other survival curves for multiple prognostic signatures **[12]** shows that it is also competitive with more elaborated supervised procedures.

Conclusion:

The main contribution of this paper is a method allowing to associate a network of molecular interactions with a ranked list of genes, called OFTEN-analysis. The analysis allows finding a functional core of a set of genes located at the top of the list but not necessarily formed by the most top-ranked genes. An advantage of the method is in that it does not use any pre-defined pathway or ontology bordersin the global PPI network. This is why OFTEN is more informative than the standard enrichment analysis and can lead to the discovery of not yet known molecular mechanisms. OFTEN-analysis can become a standard tool in high-throughput data analysis in molecular biology: therefore, we have implemented it in BiNoM Cytoscape plugin **[16]**. OFTEN-analysis can be applied to a ranked list of genes which can be produced by any type of statistical methods (for example, from Principal Component Analysis or regression).

We applied the method to four ranked gene lists produced from analysis of differential gene expression between metastatic and non-metastatic breast tumours. We show that the sets of genes forming OFTENs are characterized by larger overlap than the sets of the most differentially expressed genes. We derive a META-OFTEN network representing the most reproducible part of four OFTENs and show that it contains known cancer and metastases driver genes as well as new mechanisms whose role in metastatic development is to be understood. We show that the genes of the META-OFTEN network compose a gene signature with significant prognostic value.

Acknowledgement:
AZ is a member of the team "Computational Systems Biology of Cancer," Equipe labellisee by the Ligue Nationale Contre le Cancer. UK was supported by governmental grant from the Ministry of education and science, Republic of Kazakhstan.

References:

[1] van't Veer LJ *et al. Nature*. 2002 **415**: 530 [PMID: 11823860]
[2] van de Vijver MJ *et al. N Engl J Med*. 2002 **347**: 1999 [PMID: 12490681]
[3] Wang Y *et al. Lancet*. 2005 **365**: 671 [PMID: 15721472]
[4] Cobleigh MA *et al. Clin Cancer Res*. 2005 **11**: 8623 [PMID: 16361546]
[5] Ein-Dor L *et al. Bioinformatics*. 2005 **21**: 171 [PMID: 15308542]
[6] Chuang HY *et al. Mol Syst Biol*. 2007 **3**: 140 [PMID: 17940530]
[7] Rapaport F *et al. BMC Bioinformatics*. 2007 **8**: 35 [PMID: 17270037]
[8] Foekens JA *et al. J Clin Oncol*. 2006 **24**: 1665 [PMID: 16505412]
[9] Chen J *et al. J Biomed Inform*. 2010 **43**: 385 [PMID: 20350617]
[10] Finocchiaro G *et al. Nucleic Acids Res*. 2007 **35**: 2343 [PMID: 17389643]
[11] Daves MH *et al. BMC Med Genomics*. 2011 **4**: 56 [PMID: 21736749]
[12] Reyal F *et al. Breast Cancer Res*. 2008 **10**: R93 [PMID: 19014521]
[13] Yao C *et al. BMC Syst Biol*. 2010 **4**: 151 [PMID: 21059271]
[14] Keshava Prasad TS *et al. Nucleic Acids Res*. 2009 **37**: D767 [PMID: 18988627]
[15] Smoot ME *et al. Bioinformatics*. 2011 **27:** 431 [PMID: 21149340]
[16] Zinovyev A *et al. Bioinformatics*. 2008 **24:** 876 [PMID: 18024474]
[17] Bild AH *et al. Nature*. 2006 **439**: 353 [PMID: 16273092]
[18] Borghese F & Clanchy FI, *Expert Opin Ther Targets*. 2011 **15**: 237 [PMID: 21208136]
[19] Pinna G *et al. Math Model Nat Phenom*. 2012 **7**: 337

Edited by P Kangueane
Citation: Kairov *et al.* Bioinformation 8(16): 773-776 (2012)