frontiers in GENETICS

PERSPECTIVE ARTICLE
published: 25 March 2014
doi: 10.3389/fgene.2014.00055

# Oncofinder, a new method for the analysis of intracellular signaling pathway activation using transcriptomic data

**Anton A. Buzdin[1,2,3,4], Alex A. Zhavoronkov[1,2,4], Mikhail B. Korzinkin[1,5], Larisa S. Venkova[5], Alexander A. Zenin[5], Philip Yu. Smirnov[5] and Nikolay M. Borisov[1,4,5]***

[1] Pathway Pharmaceuticals, Limited, Wan Chai, Hong Kong, Hong Kong
[2] D. Rogachev Federal Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia
[3] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia
[4] Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[5] Burnasyan Federal Medical Biophysical Center, Moscow, Russia

**Edited by:**
*Fangqing Zhao, Beijing Institutes of Life Science, China*

**Reviewed by:**
*Ji Qi, Fudan University, China*
*Tapesh Santra, University College Dublin, Ireland*

***Correspondence:**
*Nikolay M. Borisov, Burnasyan Federal Medical Biophysical Center, 46, Zhivopisnaya St., Moscow, Russia*
*e-mail: nborisov@fmbcfmba.ru*

We propose a new biomathematical method, OncoFinder, for both quantitative and qualitative analysis of the intracellular signaling pathway activation (SPA). This method is universal and may be used for the analysis of any physiological, stress, malignancy and other perturbed conditions at the molecular level. In contrast to the other existing techniques for aggregation and generalization of the gene expression data for individual samples, we suggest to distinguish the positive/activator and negative/repressor role of every gene product in each pathway. We show that the relative importance of each gene product in a pathway can be assessed using kinetic models for "low-level" protein interactions. Although the importance factors for the pathway members cannot be so far established for most of the signaling pathways due to the lack of the required experimental data, we showed that ignoring these factors can be sometimes acceptable and that the simplified formula for SPA evaluation may be applied for many cases. We hope that due to its universal applicability, the method OncoFinder will be widely used by the researcher community.

**Keywords: mitogenic signaling pathways, microchip transcriptome investigation, targeted anti-cancer drugs, expression level profiling, signalome profiling, stochastic robustness analysis**

Intracellular signaling pathways (SPs) regulate numerous processes involved in normal and pathological conditions including development, growth, aging, and cancer. Many bioinformatic tools have been developed recently that analyze SPs. However, none of them makes it possible to efficiently do the high-throughput quantification of pathway activation scores for the individual biological samples. Here we propose a method for quick, informative and large-scale screening of changes in signaling pathway activation (SPA) in cells and tissues. These changes may reflect various differential conditions like differences in physiological state, aging, disease, treatment with drugs, infections, media composition, additives, etc. One of the potential applications of SPA studies may be in utilizing mathematical algorithms to identify and rank the medicines based on their predicted efficacy.

The information about SPA can be obtained from the massive proteomic or transcriptomic data. Although the proteomic level may be somewhat closer to the biological function of SPA, the transcriptomic level of studies today is far more feasible in terms of performing experimental tests and analyzing the data. The transcriptomic methods like Next-generation sequencing (NGS) or microarray analysis of RNA can routinely determine expression levels for all or virtually all human genes (Shirane et al., 2004). Transcriptome profiling may be performed for the minute amount of the tissue sample, not necessarily fresh, but also for the clinical formalin-fixed, paraffin-embedded (FFPE) tissue blocks. For the molecular analysis of cancer, gene expression can be interpreted in terms of abnormal SPA features of various pro- and antimitotic SPs. Such analysis may improve further decision-making process of treatment strategy selection by the clinician.

Pro- and antimitotic SPs that determine various stages of cell cycle progression remained in the spotlight of the computational biologists for more than a decade (Kholodenko et al., 1999; Borisov et al., 2009; Kuzmina and Borisov, 2011). Today, hundreds of SPs and related gene product interaction maps that show sophisticated relationships between the individual molecules, are cataloged in various databases like UniProt (The UniProt consortium, 2011), HPRD (Mathivanan et al., 2006), QIAGEN SABiosciences (SABiosciences), WikiPathways (Bauer-Mehren et al., 2009), Ariadne Pathway Studio (Nikitin et al., 2003), SPIKE (Elkon et al., 2008), Reactome (Haw and Stein, 2012), KEGG (Nakaya et al., 2013), etc. One group of bioinformatic approaches integrated the analysis of transcriptome-wide data with the models employing the mass action law and Michaelis-Meten kinetics (Yizhak et al., 2013). These methods which were developing during last 15 years, however, remained purely fundamental until recently, primarily, because of the multiplicity of interaction domains in the signal transducer proteins that enormously increase the interactome complexity (Conzelmann et al., 2006; Borisov et al., 2008). Secondly, a considerable number of unknown free parameters, such as kinetics constants and/or concentrations of protein molecules, significantly complicated the SPA analysis. Yizhak et al. (2013) suggested that the clinical efficiency of several drugs, e.g., geroprotectors, may be evaluated as the ability to induce the kinetic models of the pathways into the steady state. However, protein-protein interactions were quantitatively characterized in detail only for a tiny fraction of SPs. This approach is also time-consuming since to process each transcriptomic dataset it requires extensive calculations for the kinetic models (Yizhak et al., 2013).

However, all the contemporary bioinformatical methods that were proposed for digesting large-scale gene expression data followed by recognition and analysis of SPs, have an important disadvantage. They do not allow tracing the overall pathway activation signatures and quantitively estimate the extent of SPA (Kuzmina and Borisov, 2011; Hwang, 2012; Yizhak et al., 2013). This may be due to lack of the definition of the specific roles of the individual gene products in the overall signal transduction process, incorporated in the calculation matrix used to estimate SPA.

Here we propose a new method that, to our knowledge, for the first time makes it possible to quantitatively estimate SPA for individual samples basing on the large-scale gene expression data. The method was previously announced by our team here (Zhavoronkov et al., 2014). Theoretically, the signal transduction efficiency at every stage of the SP depends on the concentrations of the interacting gene products. The computational modeling of the signal transduction processes indicated that most of the interacting proteins can be found in the living cells at the concentrations significantly lower than the saturation levels for each transduction step (Birtwistle et al., 2007; Borisov et al., 2009). Our model is based on the correlation of the signal transducer concentrations and the overall SPA. We also determined the overall individual roles of certain gene products in the functioning of each individual SP. These roles can be either positive or negative signal transduction regulators; alternatively, for some proteins the roles may be undefined or neutral. Finally, these roles may be characterized quantitatively depending on the individual importances of the individual interactors in the overall SPA. The determination of these roles for each individual SP is a non-trivial task that has several uncertainties. Namely, protein interactions within each pathway may be competitive or independent, and therefore, belong to a sequential or parallel series of the nearby events (Borisov et al., 2006; Conzelmann et al., 2006). The overall graph for the protein interaction events may include both sequential (pathway-like) and parallel (network-like) edges (Conzelmann et al., 2006; Borisov et al., 2008). The role of each gene product in the signal transduction may depend on whether it works in a sequential or a parallel way. Alternatively, as the raw approximation of this situation, one may propose a simplified method that utilizes only the overall roles of each gene product in the SPA. In this case, each simplified signaling graph includes only two types of branches of protein interaction chain: one for sequential events that promote SPA, and another for repressor sequential events. Under these conditions, it can be presumed that all activator/repressor members have equal importances for the SPA, and come to the following formula for the overall signal outcome ($SO$) of a given pathway, $SO = \frac{\prod_{i=1}^{N}[AGEL]_i}{\prod_{j=1}^{M}[RGEL]_j}$. Here the multiplication is done over all possible activator and repressor proteins in the pathway, $[AGEL]_i$ and $[RGEL]_j$ are relative gene expression levels of activator ($i$) and repressor ($j$) members, respectively. To obtain an additive value, it is possible to take the logarithmic levels of gene expression, and thus come to a function of *pathway activation strength*, *PAS*, which operates with the experimental datasets obtained during comprehensive profiling of gene expression, for a pathway $p$, $PAS_p = \sum_n ARR_{np} \cdot \lg(CNR_n)$. Here the *case-to-norm ratio*, $CNR_n$, is the ratio of the expression levels of a gene $n$ in the sample (e.g., of a cancer patient) and in the control (e.g., average value for healthy group). The discrete value *ARR* (*activator/repressor role*) shows whether the gene product promotes SPA (1), inhibits it (−1) or plays an intermediate role (0.5, 0 or −0.5, respectively). Negative and positive overall *PAS* values correspond, respectively, to decreased or increased activity of SP in a sample, with the extent of this activity proportional to the absolute value of *PAS*.

However, the assumption of sequential protein-protein interaction in pathways may seem rather artificial. Although it is difficult to precisely estimate the importance of certain gene products that act in the pathway in a non-sequential mode, the solution may come from the kinetic models of SPA that use the "low-level" approach of mass action law describing each act of protein interactions. Some of these models were previously experimentally validated by us and others using Western blot analysis (Kholodenko et al., 1999; Kiyatkin et al., 2006; Birtwistle et al., 2007; Borisov et al., 2009; Kuzmina and Borisov, 2011). Our previous experience suggests that the two approaches can be used to estimate the importance of distinct genes/proteins in the pathways. One of them operates with the concept of sensitivity of the ordinary differential equation system with the free parameters (Kholodenko et al., 2003), which is generally applied to kinetic constants, but may be used for assperating with the protein concentrations in the kinetic model of a pathway (Kuzmina and Borisov, 2011), according to a formula,$w_j^{(1)} = \frac{1}{T}\int_0^T \left|\frac{\partial \ln[EFF(t)]}{\partial \ln_j C^{tot}}\right| dt$. Here $w$ is the importance factor, $[EFF(t)]$ is the time-dependent concentration of the active pathway effector protein (experimentally traced marker of a pathway activation), the upper integration limit $T$ is the time of reaching the steady-state, and $C_j^{tot}$ is the total concentration for the protein $j$.

Another way to calculate the importance factor for the gene products deals with the stiffness/sloppiness analysis of the effector activation (Daniels et al., 2008). This approach comprises analyzing the Hesse matrix, $H_{ij} = \frac{\partial^2}{\partial C_i^{tot} \partial C_j^{tot}} \sum_k \frac{\left([EFF(C^{tot}, t_k)] - [EFF]_k^{\exp}\right)^2}{\sigma_k^2}$, where $\mathbf{C}^{tot}$ is the vector of total concentrations for every protein in the pathway, $[EFF(C^{tot}, t_k)]$ is concentration of an active pathway effector protein at the time point $t_k$, $[EFF]_k^{\exp}$ is the experimentally measured (e.g., by Western blots) total concentration of the effector at the same time, and $\sigma_k$ is the experimental error for this measurement. The sloppiness/stiffness analysis looks for the eigenvalues, $\lambda_m$, and eigenvectors, $\boldsymbol{\xi}_m$, for the Hesse matrix, $\mathbf{H}\boldsymbol{\xi}_m = \lambda_m \cdot \boldsymbol{\xi}_m$. The higher is the absolute value of $\lambda_n$, the "stiffer" is the direction within the $n$-dimensional space of $\mathbf{C}^{tot}$ (where $n$ is the number of protein types in the pathway model). The eigenvector components along with the stiffest direction, $\boldsymbol{\xi}_s$, may be used for assessment of the importance factor $w$ of a certain gene products in a pathway according to the formula: $w_j^{(2)} = |\xi_{sj}|$.

Taking into account the above considerations, we come to the following final formula for assessing the SPA: $PAS_p^{(1,\ 2)} = \sum_n ARR_{np} \cdot BTIF_n \cdot w_n^{(1,2)} \cdot \lg(CNR_n)$. Here the Boolean flag *BTIF* (*beyond tolerance interval flag*) indicates that the expression level for the gene n for the given sample is different enough from the respective expression level in the reference sample or set of reference samples. For this demonstration of our method we applied two simultaneous restriction/inclusion criteria to the expression of each individual gene: (i) 50% expression level cutoff rate compared to the average for the reference set, and (ii) the sample expression level should differ stronger than two standard deviations from the average of the reference set.

We next explored the effect of the introduction of the importance factors $w$ in calculating *PAS* compared to the simplified model of *PAS* evaluation lacking $w$. Importance factors were calculated using either sensitivity-based, $w^{(1)}$, or stiffness-based, $w^{(2)}$, algorithms. We performed this verification for the EGFR pathway, for which we established and published this model previously (Kuzmina and Borisov, 2011). For these two sets of the importance factors, and for the $w$-free model, we performed a computational analysis of nine transcriptomes established using microarray hybridization technology for human glioblastoma samples from the published datasets (Supplementary dataset 1). The information on SP organization was taken from the Web-based SABiosciences database. The data on *ARR* were manually curated by analyzing the same database. Our findings suggest that the cloud of values for the ratio $\frac{PAS_{\mathrm{EGFR}}^{(1)}}{PAS_{\mathrm{EGFR}}}$ (where $PAS_{\mathrm{EGFR}}$ is the *PAS* value for the EGFR pathway in the simplified model, where all importance factors equal to 1) lies within the interval of (0.6 ± 0.8), whereas the ratio $\frac{PAS_{\mathrm{EGFR}}^{(2)}}{PAS_{\mathrm{EGFR}}}$ belonged to the interval (1.0 ± 0.8). Overall, we conclude that for such a complex SP like EGFR which includes >300 gene products, incorporation of the importance factors had only a moderate effect on the *PAS*. This suggests that, in principle, the simplified formula for *PAS* calculation may be applied for the pathway analysis.

For the overwhelming majority of the SPs, there is no experimental data available that makes impossible for them to calculate the importance factors using kinetic models. For them we performed the stochastic robustness analysis using the simplified formula for *PAS*. We introduced the additional random perturbation factor, $w_n$, which was used as the analog of importance factor for *PAS* evaluation. In our computational simulation, the distribution of $w_n$ was logarithmically normal and calculated as follows: $w_n = 2^{x_n}$, where $x_n$ were normally distributed random numbers with the expected value of $M = 0$ and standard deviation $\sigma = 0.5$. The random perturbation factors $w_n$ were applied to the glioblastoma transcriptional dataset GSM215422 (GSM215422 dataset). Importantly, although the perturbation was done independently 98 times with independent weighting factors $w_n$, for each gene, the values of standard deviation for the set of alternate *PAS* (*APAS*) were nor big enough to mask the proportional trend between the average perturbed *PAS* and unperturbed *PAS* for each of the 68 SPs analyzed in this study (**Figure 1**; Supplementary dataset 2).

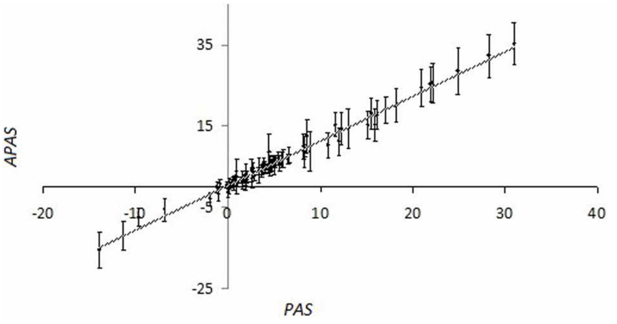

**FIGURE 1 | Values of pathway activation strength (*APAS*) that were calculated using the 98 random trials, each having random log-normally distributed weighting factors $w_n$, vs. non-perturbed *PAS* for the different SPs, calculated using OncoFinder method.** The pathway information was extracted from the SABiosciences database. Primary data are shown on the Supplementary dataset 3. For the perturbed values (*APAS*), both average values (points at the plot) and standard deviation bars are shown.

We propose here a new biomathematical method, OncoFinder, for both quantitative and qualitative analysis of the intracellular SP activation. It can be used for the analysis of any physiological, stress, malignancy and other perturbed conditions at the molecular level. The enclosed mathematical algorithm enables processing of high-throughput transcriptomic data, but there is no technical limitation to apply OncoFinder to the proteomic datasets as well, when the developments in proteomics allow generating proteome-wide expression datasets. We hope that due to its universal applicability, the method OncoFinder will be widely used by the biomedical researcher community and by all those interested in thorough characterization of the molecular events in the living cells. We also want to encourage building international scientific partnership aimed at the standardized experimental characterization of the importance factors for individual proteins, starting at least with the SPs most relevant to the major aspects of human physiology.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00055/abstract

## REFERENCES

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* 5, 290. doi: 10.1038/msb.2009.47

Birtwistle, M. R., Hatakeyama, M., Yumoto, N., Ogunnaike, B. A., Hoek, J. B., and Kholodenko, B. N. (2007). Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Mol. Syst. Biol.* 3:144. doi: 10.1038/msb4100188

Borisov, N., Aksamitiene, E., Kiyatkin, A., Legewie, S., Berkhout, J., Maiwald, T., et al. (2009). Systems-level interactions between insulin-EGF networks amplify mitogenic signaling. *Mol. Syst. Biol.* 5:256. doi: 10.1038/msb.2009.19

Borisov, N. M., Chistopolsky, A. S., Faeder, J. R., and Kholodenko, B. N. (2008). Domain-oriented reduction of rule-based network models. *IET Syst. Biol.* 2, 342–351. doi: 10.1049/iet-syb:20070081

Borisov, N. M., Markevich, N. I., Hoek, J. B., and Kholodenko, B. N. (2006). Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains. *Biosystems* 83, 152–166. doi: 10.1016/j.biosystems.2005.03.006

Conzelmann, H., Saez-Rodriguez, J., Sauter, T., Kholodenko, B. N., and Gilles, E. D. (2006). A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks. *BMC Bioinformatics* 7:4. doi: 10.1186/1471-2105-7-34

Daniels, B. C., Chen, Y. J., Sethna, J. P., Gutenkunst, R. N., and Myers, C. R. (2008). Sloppiness, robustness and evolvability in systems biology. *Curr. Opin. Biotechnol.* 19, 389–395. doi: 10.1016/j.copbio.2008.06.008

Elkon, R.,Vesterman, R., and Amit, N. (2008). SPIKE—a database, visualization and analysis tool of cellular signaling pathways. *BMC Bioinformatics* 9:110. doi: 10.1093/nar/gkq1167

GSM215422 dataset. Available online at: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM215422

Haw, R., and Stein, L. (2012). Using the Reactome database. *Curr. Protoc. Bioinformatics* ñhapter 8, unit 8:7. doi: 10.1002/0471250953.bi0807s38

Hwang, S. (2012). Comparison and evaluation of pathway-level aggregation methods of gene expression data. *BMC Genomics* 13(Suppl. 7):S26. doi: 10.1186/1471-2164-13-S7-S26

Kholodenko, B., Kiyatkin, A., Bruggeman, F., Sontag, E., Westerhof, H. V., and Hoek, J. B. (2003). Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. U.S.A.* 20, 12841–12846. doi: 10.1073/pnas.192442699

Kholodenko, B. N., Demin, O. V., Moehren, G., and Hoek, J. B. (1999). Quantification of short term signaling by the epidermal growth factor receptor. *J. Biol. Chem.* 274, 30169–30181. doi: 10.1074/jbc.274.42.30169

Kiyatkin, A., Aksamitiene, E., Markevich, N. I., Borisov, N. M., Heok, J. B., and Kholodenko, B. N. (2006). Scaffolding protein GAB1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops. *J. Biol. Chem.* 281, 19925–19938. doi: 10.1074/jbc.M600482200

Kuzmina, N. B., and Borisov, N. M. (2011). Handling complex rule-based models of mitogenic cell signaling (On the example of ERK activation upon EGF stimulation). *Int. Proc. Chem. Biol. Environ. Eng.* 5, 76–82.

Mathivanan, S., Periaswamy, B., Gandhi, T., Kandasamy, K., Suresh, S., Mohmood, R., et al. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* 7:S19. doi: 10.1186/1471-2105-7-S5-S19

Nakaya, A., Katayama, T., Itoh, M., Hiranuka, K., Kawashima, S., Moriya, Y., et al. (2013). KEGG OC: a large-scale automatic construction of taxonomy-based ortholog clusters. *Nucleic. Acids Res.* 41, D353–D357. doi: 10.1093/nar/gks1239

Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* 19, 2155–2157. doi: 10.1093/bioinformatics/btg290

SABiosciences, a Qiagen company. Available online at: http://www.sabiosciences.com/pathwaycentral.php

Shirane, D., Sugao, K., Namiki, S., Tanabe, M., Iino, M., and Hirose, K. (2004). Enzymatic production of RNAi libraries from cDNA. *Nat. Genet.* 36, 190–196. doi: 10.1038/ng1290

The UniProt consortium. (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic. Acids Res.* 39, D214–D219. doi: 10.1093/nar/gkq1020

Yizhak, K., Gabay, O., Cohen, H., and Rupin, E. (2013). Model-based identification of drug targets that, revert disrupted metabolism and its application to ageing. *Nat. Commun.* 4:2632–doi: 10.1038/ncomms3632

Zhavoronkov, A., Buzdin, A. A., Garazha, A. V., Borisov, N. M., and Moskalev, A. A. (2014). Signaling pathway cloud regulation for in silico screening and ranking of the potential geroprotective drugs. *Front. Genet.* 5:49. doi: 10.3389/fgene.2014.00049

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

*Received: 09 December 2013; accepted: 03 March 2014; published online: 25 March 2014.*

*Citation: Buzdin AA, Zhavoronkov AA, Korzinkin MB, Venkova LS, Zenin AA, Smirnov PY and Borisov NM (2014) Oncofinder, a new method for the analysis of intracellular signaling pathway activation using transcriptomic data. Front. Genet.* ***5****:55. doi: 10.3389/fgene.2014.00055*

*This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.*