Algorithmic Annotation of Functional Roles for Components of 3,044 Human Molecular Pathways

Maxim Sorokin¹,²,³, Nicolas Borisov¹,³, Denis Kuzmin³, Alexander Gudkov², Marianna Zolotovskaia³, Andrew Garazha¹ and Anton Buzdin¹,³,⁴,⁵*

¹Omicsway Corp., Walnut, CA, United States, ²Laboratory of Clinical Genomic Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia, ³Laboratory for Translational Bioinformatics, Moscow Institute of Physics and Technology, Moscow, Russia, ⁴Laboratory of Systems Biology, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, ⁵World-Class Research Center “Digital Biodesign and Personalized Healthcare”, Sechenov First Moscow State Medical University, Moscow, Russia

Current methods of high-throughput molecular and genomic analyses enabled to reconstruct thousands of human molecular pathways. Knowledge of molecular pathways structure and architecture taken along with the gene expression data can help interrogating the pathway activation levels (PALs) using different bioinformatic algorithms. In turn, the pathway activation profiles can characterize molecular processes, which are differentially regulated and give numeric characteristics of the extent of their activation or inhibition. However, different pathway nodes may have different functions toward overall pathway regulation, and calculation of PAL requires knowledge of molecular function of every node in the pathway in terms of its activator or inhibitory role. Thus, high-throughput annotation of functional roles of pathway nodes is required for the comprehensive analysis of the pathway activation profiles. We proposed an algorithm that identifies functional roles of the pathway components and applied it to annotate 3,044 human molecular pathways extracted from the Biocarta, Reactome, KEGG, Qiagen Pathway Central, NCI, and HumanCYC databases and including 9,022 gene products. The resulting knowledgebase can be applied for the direct calculation of the PALs and establishing large scale profiles of the signaling, metabolic, and DNA repair pathway regulation using high throughput gene expression data. We also provide a bioinformatic tool for PAL data calculations using the current pathway knowledgebase.

Keywords: functional algorithmic annotation, signaling pathways, DNA repair pathways, metabolic pathways, transcriptomics, proteomics, human molecular pathway regulation

INTRODUCTION

Intracellular molecular pathways are specific networks of interacting molecules that are involved in certain molecular functions (Junaid et al., 2020; Ma and Liao, 2020; Zheng et al., 2020). Knowledge of molecular pathways regulation is important for understanding intracellular processes related to all major events, including cell survival, growth, differentiation, motility, proliferation, senescence, malignization, and death (Buzdin et al., 2018). Molecular pathways are affected during organism growth and development, aging and disease progression (Parkhitko et al., 2020). Current methods of large-scale molecular and genomic analyses enabled to catalogue thousands of human molecular pathways (Wishart et al., 2020). In turn, high-throughput gene expression analyses like RNA sequencing (Sorokin et al., 2020a), expression
Functional Algorithmic Molecular Pathway Annotation

We proposed an algorithm that identifies functional roles of the pathway components based on the pathway topology and applied it here to annotate 3,044 human molecular pathways extracted from the Biocarta, Reactome, KEGG, NCI, and HumanCYC databases, collectively covering 9,022 gene products. The resulting knowledgebase can be applied for the direct calculation of the PALs and establishing large scale profiles of the signaling, metabolic, and DNA repair pathway regulation using high throughput gene expression data.

RESULTS AND METHODS

Extraction of Molecular Pathway Data

We extracted structures of molecular pathways from the National Cancer Institute (NCI; Schaefer et al., 2009), Biocarta (Nishimura, 2001), Qiagen Pathway Central,1 HumanCyC (Romero et al., 2004), Reactome (Croft et al., 2014), and Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2010) databases (Table 1). For all the databases but Qiagen Pathway Central, the data on the pathway architecture, nodes and pairwise activation/inhibition interactions were extracted in biopax format. In the case of Qiagen Pathway Central database, no machine-readable format of data was available, and we manually curated data from the available graphical pathway representations (Table 1).

In addition to the extracted full-size pathways, we also generated a number of subsequent “micropathways” that were derivatives of the complete pathways (Figures 1A,B). Micropathway is a sub-graph, which contains “molecular function” node and nodes from all possible paths of length 3 including terminal “molecular function” node. Many full-size pathways have two or more terminal branches that may have different functional impact(s). We, therefore, introduced micropathways to characterize molecular processes in more detail by separately analyzing different terminal branches of the pathways. Totally, we processed 3,044 pathways including 2018 full-size, or “core” pathways, and 1,026 micropathways that covered collectively products of 9,022 human genes (Table 1).

1https://www.qiagen.com/gb/resources/resourcedetail?id=5869c38a-5033-4ccb-a281-d869893acfe8&lang=en

### TABLE 1 | Statistics of the curated pathway databases.

| Database      | References                          | Number of     | Data curation format |
|---------------|-------------------------------------|---------------|----------------------|
|               |                                     | core pathways | all pathways         |                      |
| Biocarta      | Nishimura, 2001                     | 198           | 337                  | 1,082                |
| Reactome      | Croft et al., 2014                  | 945           | 945                  | 6,105                |
| KEGG          | Kanehisa et al., 2010               | 288           | 288                  | 5,593                |
| Qiagen        | Pathway Map Reference Guide–QIAGEN, 2014 | 57          | 380                  | 2,493                |
| NCI           | Schaefer et al., 2009               | 211           | 775                  | 2,214                |
| HumanCYC      | Romero et al., 2004                 | 319           | 319                  | 1,038                |
| Total number  |                                     | 2,018         | 3,044                | 9,022                |

Number of all pathways includes core pathways and micropathways. Number of unique gene products covered by pathways from the respective database. For total number, the amount of unique gene products for all pathways is shown.
Note that number of pathway nodes was smaller than the number of genes involved in a pathway because one node could correspond to several gene products.

For several pathway components alternative gene names were used and we then converted all gene names according to the Human Genome Organization HGNC nomenclature (Povey et al., 2001).

### Algorithmic Annotation of Molecular Pathways

For most of the published PAL applications, maximum five types of functional roles for gene products were comprised. These roles, described by an Activator/Repressor Role (ARR) parameter can be formulated as follows: pathway activator (ARR = 1), rather activator (ARR = 0.5), repressor (ARR = −1), rather repressor (ARR = −0.5), and gene product with uncertain or inconsistent role (ARR = 0). In the previous studies, ARR values were obtained by manually curating pathway graphs. This is however not feasible for annotating thousands of molecular pathways. We developed an original algorithm that automatically assigns ARR score values to gene products that participate in a molecular pathway.

The ARR annotation algorithm is based on the machine reading of gene product interaction graph within each pathway. Nodes correspond to gene products, and the ribs between every pair of nodes represent molecular interaction between the corresponding gene products. Each rib on the graph has a direction and is characterized by an activator or inhibitory nature of the molecular interaction it represents. For the correct calculation of ARR values, the pathway graph must be connected, wherein a weak connectivity is acceptable.

If the pathway molecular interaction graph meets these criteria, then ARR coefficients can be algorithmically assigned to the participating gene products. For the biochemical pathways, we put enzyme gene names on the pathway nodes, and the interaction ribs represented directions of the catalyzed reactions.

The algorithm used consisted of the following major steps.

1. **Initialization.** At this stage, a major node is algorithmically identified to be the “central” node of the pathway graph (Figure 1C). The major node will be used as the standard of pathway function with ARR = 1. To identify the central node, for every pathway node (V) two parameters N and M are calculated where N is the number of other nodes, which can be reached when moving from the node V, and M is the number of other nodes from which the node V can be reached. N+M, therefore, is the number of other nodes that are directly connected with the node V. The central node will be the node V_max, for which N+M reaches the maximum value. The central node identified is then assigned with ARR = 1 value. It serves as the starting point for further recursive assignment of ARR values to the other nodes. If multiple nodes have the same maximal N+M, then V-node for a pathway is selected randomly among those “maximal” nodes. Therefore, the algorithm is suitable also for circular-organized pathways, where all nodes will have equal N+M.

2. **Recursion.** For every node V, all connected nodes P, under ARR annotation may have ribs either directed toward V (Pi → V) or outward V (Pi ← V) on the graph. During recursion, each rib can be considered only once in order to prevent endless recursion in case of cyclic interactions on the graph. If the rib has an “activator” characteristic, temporary ARRtemp = 1 is assigned to the node Pi. In contrast, if the rib has an “inhibitor” characteristic, Pi is assigned with ARRtemp = −1. Conversely, all the gene products included in the node Pi receive the same ARRtemp characteristics.

   Let gene product GP∈ belongs to node Pi. If GP∈ was never previously considered in the graph traversal, ARR = ARRtemp(Pi) for the node P, would be assigned for GP∈. In the case when GP∈ was previously considered in the graph traversal and the previously assigned ARR of it node is equal to the current ARRtemp(Pi) then ARR = ARRtemp would be assigned to the node Pi. If GP∈ was previously considered in the graph traversal but its previously assigned ARR is not equal to ARRtemp then ARR is assigned to the gene product GP∈, according to the following conflict resolution rule.

   If a gene product GP∈ with previously specified ARR or ARRs is currently considered in the graph traversal but its previously assigned ARR(s) contradict(s) with the ARRtemp(Pi) then the conflict(s) should be resolved as follows:

   1. If the signs of the previous ARR coefficient(s) and ARRtemp(Pi) are different, then the resulting ARRfinal(Pi) = 0;
   2. If the difference between ARRs(Pi) and any of the previous ARRs(PG∈) does not exceed 0.5 and at least one of the ARRs is positive, the resulting ARRfinal(Pi) = 0.5;
   3. If the difference between ARRs(Pi) and any of the previous ARRs(PG∈) does not exceed 0.5 and at least one of the ARRs is negative, the resulting ARRfinal(Pi) = −0.5.

   Then the recursion R is initiated for every node Pi, all of its gene products starting from the nodes proximate to the central node V. As a result, the algorithm will assign ARR values to all the connected graph nodes and the encoded gene products.

   After the recursion finalization pathway activators will have ARR = 1, rather activators – ARR = 0.5, inhibitors – ARR = −1, rather inhibitors – ARR = −0.5, and genes with inconsistent role – ARR = 0. The recursion is stopped when a vertex with 0, 0.5, or −0.5 ARR is encountered during the traversal of the graph. This rule is needed because otherwise all vertices will have ARR 0, 0.5, or −0.5 in case of the only one ARR inconsistency found. However, this rule also may lead to exclusion of some genes described in the original source.

   Therefore, the gene products included in the molecular pathway database will have the assigned ARR values representing their functional significances in the given molecular pathway. These values can be used for further calculations of the PALs according to any algorithm of PAL calculation.

### Annotated Pathways Knowledgebase

We report here an ARR-curated database of 3,044 molecular pathways including 2,018 core pathways and 1,026 micropathways
(Supplementary Dataset 1). The current pathway name reflects its source database and its name in the source database. For every pathway, there is a separate .csv file including the following three worksheets: (i) genes, (ii) edges, and (iii) nodes. The worksheet (i) genes include gene names according to HUGO Gene Nomenclature Committee (HGNC) nomenclature and the corresponding ARRs for the gene products participating in the pathway under consideration. The worksheet (ii) edges include information about molecular interactions between every pair of the interacting pathway nodes. Every node is defined by the names of gene products or physiological outcome(s) that form this node. The interaction type is specified as “activation,” “inhibition,” or “undefined,” where appropriate. The worksheet (iii) nodes include node names and gene names corresponding to every node on the pathway graph.

It should be noted that annotation of similar pathways may be different between the source databases. For example, EGFR signaling pathway is presented in Qiagen database as “EGF_ Pathway,” in Reactome as “reactome_Signaling_by_EGFR_Main_ Pathway” and in Biocarta as “biocarta_egf_signaling_Main_Pathway.” Yet conceptually similar, all three pathways have different gene and edge compositions. In this study, we did not aim to identify inconsistencies between different source databases and annotated all the pathways under their original names.

We made freely accessible software for PAL calculation using the annotated pathway database accessible following the link: https://pypi.org/project/oncoboxlib/. Algorithm is implemented as a Python library. It takes normalized (by DESeq2, quantile normalization or other) gene expression data as an input. Gene symbols should be provided in HGNC format accessible through the web-site genenames.org. At least two groups of samples are required: cases and controls, each group represented by at least one sample. Sample names should contain “Norm_” (for controls) or “Tumour_” (for cases). Output will contain PAL values for each pathway in each sample. All annotated pathway datasets mentioned in this paper alternatively can be downloaded and used for PAL calculation using the same link.2

2https://pypi.org/project/oncoboxlib/
We also provide here an example of PAL calculation for real-world data. We extracted gene expression data for gastric cancer samples \( n = 16 \); Sorokin et al., 2020b) together with gene expression profiles of healthy stomach \( n = 7 \) samples of patients who died in road accidents (Suntsova et al., 2019), that were sequenced using the same equipment and protocols. Cancerous samples were marked as “Tumour_” and normal samples – as “Norm_.” Then we calculated PAL values (3,044 for each sample) for all molecular profiles using the above software, which produced an output file “pal.csv” (Supplementary Dataset 1).

**DISCUSSION**

We propose here the recursive algorithm for functional annotation of the molecular pathway nodes, and its application to annotation of 3,044 human molecular pathways, including signaling, metabolic, and DNA repair pathways extracted from six major pathway hubs (Table 1). The ARR-annotated pathways can be used for further calculations of PALs using high-throughput gene expression data, e.g., RNA sequencing or proteomic profiles (Buzdin et al., 2018; Figure 2). To this end, several previously published bioinformatic methods can be employed (Buzdin et al., 2014a; Ozerov et al., 2016; Borisov et al., 2020), and the PAL values returned can be applied for a variety of applications including fundamental research (Panteunig-Vuhman et al., 2017), drug development (Aliper et al., 2017a; Ravi et al., 2018; Bakula et al., 2019), and personalized medicine (Poddubskaya et al., 2019a; Moisseev et al., 2020). Technically, PAL values can be used as the next-generation molecular biomarkers (Aliper et al., 2017b; Borisov et al., 2017; Sorokin et al., 2020c) or as the substrates for various machine learning applications (Borisov et al., 2018; Tkachev et al., 2018).

The proposed algorithm is suitable for the analysis of pathways with already established gene content and known topology of its molecular components. The algorithm can be used for agnostic...
objective characterization of interacting gene networks. The underlying rationale allows reducing operator’s errors and subjectivity in annotating the molecular roles of pathway components, which are inevitable in case of manual curation of the pathway graphs including hundreds of nodes. Another advantage is the pathway-centric approach during annotation, when gene product role in one pathway can be different from its role in another pathway.

The major limitations deal with the algorithm applicability only for the tasks of further calculations of pathway activation scores/ranks. Such an approach also does not address the issue of crosstalk between different molecular pathways, because all pathways are analyzed separately.

In this study, we annotated a collection of previously published human molecular pathways (Supplementary Dataset 1). We plan to update the current human knowledgebase annually with new releases of already included datasets and addition of new pathway collections, e.g., recently published by Wishart et al. (2020). However, the method proposed here can be used to characterize any new set of molecular pathways with the connectivity and pairwise nodes activation/inhibition information not only for the human interactome, but also for the other biological objects under investigation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

REFERENCES

Alexandrova, E., Naga, G., Corleone, G., Buzdin, A., Aliper, A. M., Terekhanova, N., et al. (2016). Large-scale profiling of signalling pathways reveals an asthma specific signature in bronchial smooth muscle cells. Oncotarget 7, 25150–25161. doi: 10.18632/oncotarget.7209

Aliper, A., Belikov, A. V., Garazha, A., Jellen, L., Artemov, A., Sunstova, M., et al. (2016). In search for geroprotectors: in silico screening and in vitro validation of signalome-level mimetics of young healthy state. Aging 8, 2127–2152. doi: 10.18632/aging.101047

Aliper, A. M., Caska, A. B., Buzdin, A., Jetka, T., Roumiantsev, S., Moskalev, A., et al. (2015). Signaling pathway activation drift during aging: Hutchinson-Gilford Progeria Syndrome fibroblasts are comparable to normal middle-age and old-age cells. Aging 7, 26–37. doi: 10.18632/aging.100717

Aliper, A., Jellen, L., Cortese, F., Artemov, A., Karpinsky-Semper, D., Moskalev, A., et al. (2017a). Towards natural mimetics of metformin and rapamycin. Aging 9, 2245–2268. doi: 10.18632/aging.101319

Aliper, A. M., Korzinkin, M. B., Kuzmina, N. B., Zinin, A. A., Venkova, L. S., Smirnov, P. Y., et al. (2017b). Mathematical justification of expression-based pathway activation scoring (PAS). Methods Mol. Biol. 1613, 31–51. doi: 10.1007/978-1-4939-7027-8_3

Artemov, A., Aliper, A., Korzinkin, M., Lezhmina, K., Jellen, L., Zhukov, N., et al. (2015). A method for predicting target drug efficiency in cancer based on the analysis of signaling pathway activation. Oncotarget 6, 29347–29356. doi: 10.18632/oncotarget.5119

Bakula, D., Ablasser, A., Aguzzi, A., Antebi, A., Barzilai, N., Bittner, M. I., et al. (2019). Latest advances in aging research and drug discovery. Aging 11, 9971–9981. doi: 10.18632/aging.102487

Borisyov, N., Sorokin, M., Garazha, A., and Buzdin, A. (2020). Quantitation of molecular pathway activation using RNA sequencing data. Methods Mol. Biol. 2063, 189–206. doi: 10.1007/978-1-0716-0138-9_15

AUTHOR CONTRIBUTIONS

MS, NB, DK, and AB contributed to conception and design of the study. MS developed recursive pathway annotation algorithm. DK, AGu, MZ, and AGa manually curated the pathways and performed recursive algorithm implementations. AB, MZ, and MS wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Amazon and Microsoft Azure grants for cloud-based computational facilities. Financial support was provided by the Russian Foundation for basic research grant 19-29-01108.

ACKNOWLEDGMENTS

We thank Oncobox/OmicsWay research program in machine learning and digital oncology for software and starting pathway databases for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617059/full#supplementary-material

Borisyov, N., Sunstova, M., Sorokin, M., Garazha, A., Kovalchuk, O., and Aliper, A., et al. (2017). Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. Cell Cycle 16, 1810–1823. doi: 10.1080/15384101.2017.1361068

Borisyov, N., Tkachev, V., Sunstova, M., Kovalchuk, O., Zhavoronkov, A., Muchnik, L., et al. (2018). A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. Cell Cycle 17, 486–491. doi: 10.1080/15384101.2017.1417706

Borisyov, N. M., Terekhanova, N. V., Aliper, A. M., Venkova, L. S., Smirnov, P. Y., Roumiantsev, S. A., et al. (2014). Signaling pathways activation profiles make better markers of cancer than expression of individual genes. Oncotarget 5, 10198–10205. doi: 10.18632/oncotarget.2548

Buzdin, A. A., Artcibasova, A. V., Fedorova, N. F., Sunstova, M. V., Garazha, A. V., Sorokin, M. I., et al. (2016). Early stage of cytomegalovirus infection suppresses host microRNA expression regulation in human fibroblasts. Cell Cycle 15, 3378–3389. doi: 10.1080/15384101.2016.1241928

Buzdin, A., Sorokin, M., Garazha, A., Glusker, A., Aleshin, I., Poddubskaya, E., et al. (2019). RNA sequencing for research and diagnostics in clinical oncology. Semin. Cancer Biol. 60, 311–323. doi: 10.1016/j.semcancer.2019.07.010

Buzdin, A., Sorokin, M., Garazha, A., Sekacheva, M., Kim, E., Zhukov, N., et al. (2018). Molecular pathway activation - new type of biomarkers for tumor morphology and personalized selection of target drugs. Semin. Cancer Biol. 53, 110–124. doi: 10.1016/j.semcancer.2018.06.003

Buzdin, A. A., Zhavoronkov, A. A., Korzinkin, M. B., Roumiantsev, S. A., Aliper, A. M., Venkova, L. S., et al. (2014a). The OncoFinder algorithm for minimizing the errors introduced by the high-throughput methods of transcriptome analysis. Front. Mol. Biosci. 1:8. doi: 10.3389/fmolb.2014.00008

Buzdin, A. A., Zhavoronkov, A. A., Korzinkin, M. B., Venkova, L. S., Zinin, A. A., Smirnov, P. Y., et al. (2014b). 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
Zhu, Q., Izumchenko, E., Aliper, A. M., Makarev, E., Paz, K., Buzdin, A. A., et al. (2015). Pathway activation strength is a novel independent prognostic biomarker for cetuximab sensitivity in colorectal cancer patients. *Hum. Genome Var.* 2:15009. doi: 10.1038/hgv.2015.9

**Conflict of Interest:** MS, AGa, and AB have a financial relationship with OmicsWay Corp.

The remaining 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.

The handling editor declared a shared affiliation with the authors AB, MS, and AGu at the time of review.

Copyright © 2021 Sorokin, Borisov, Kuzmin, Gudkov, Zolotovskaia, Garazha and Buzdin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.