Methodology article

THE PIPELINE FOR *IN SILICO* PREDICTION OF TARGET GENES AND TARGET METABOLIC PATHWAYS FOR SMALL MOLECULES, USING CHELIDONIC ACID AS AN EXAMPLE

T. F. Nasibov[1], A. V. Gorokhova[1], K. S. Brazovsky[2]
E. D. Porokhova[1], E. Yu. Avdeeva[3]
M. V. Belousov[3], I. A. Khlusov[1, 4,*]

[1] Department of Morphology and General Pathology, Siberian State Medical University, 634050, Tomsk, Russia
[2] Department of Medical and Biological Cybernetics, Siberian State Medical University, 634050, Tomsk, Russia
[3] Department of Pharmaceutical Analysis, Siberian State Medical University, 634050, Tomsk, Russia
[4] Laboratory of Cellular and Microfluidic Technologies, Siberian State Medical University, 634050, Tomsk, Russia

* Correspondence: khlusov.ia@ssmu.ru

Abstract: Modern pharmaceutical science is undergoing significant changes due to the active integration of computer technologies into the drug development process. Prior to the conventional stages of synthesis and experimental testing, researchers are increasingly turning to *in silico* methods, which enable the simulation of the structure and pharmacological effects of chemical compounds with a high degree of predictive accuracy. This approach significantly saves time and financial resources, optimizing subsequent experimental studies. In parallel, there is an active development of epigenome-targeted therapy, a new approach that allows for the modulation of gene expression involved in pathological processes without directly altering the primary DNA structure. Using chelidonic acid as a case study, we present a pipeline for computer-based assessment of small molecule effects on gene expression and metabolic pathways. The algorithm is based on the use of the DIGEP-Pred 2.0 web service, followed by multilevel bioinformatics analysis, including: (1) gene over-representation analysis; (2) assessing the involvement of metabolic pathways; (3) functional clustering of pathways. The proposed approach can help solve problems in both fundamental and applied research on the molecular mechanisms of drug action.

Keywords: bioinformatics, computer simulation, intracellular events, molecular targets, methodic aspects

To cite this article:

Nasibov T. F., Gorokhova A. V., Brazovsky K. S., Porokhova E. D., Avdeeva E. Yu., Belousov M. V., Khlusov I. A. The pipeline for *in silico* prediction of target genes and target metabolic pathways for small molecules, using chelidonic acid as an example. *Advanced targets in Biomedicine*. 2025;1(1):06—19. https://doi.org/10.5922/ATB-2025-1-1-1

Received
23.04.2025
Revised
04.06.2025
Accepted
05.06.2025
Published
15.07.2025

© Nasibov T. F., Gorokhova A. V., Brazovsky K. S., Porokhova E. D., Avdeeva E. Yu., Belousov M. V., Khlusov I. A., 2025

Методологическая статья

АЛГОРИТМ *IN SILICO* ПРОГНОЗИРОВАНИЯ ГЕНОВ-МИШЕНЕЙ И ЦЕЛЕВЫХ МЕТАБОЛИЧЕСКИХ ПУТЕЙ ДЛЯ МАЛЫХ БИОМОЛЕКУЛ НА ПРИМЕРЕ ХЕЛИДОНОВОЙ КИСЛОТЫ

**Т. Ф. Насибов[1], А. В. Горохова[1], К. С. Бразовский[2]
Е. Д. Порохова[1], Е. Ю. Авдеева[3],
М. В. Белоусов[3], И. А. Хлусов[1, 4,*]**

[1] Кафедра морфологии и общей патологии,
Сибирский государственный медицинский университет,
634050, Россия, Томск
[2] Кафедра медицинской и биологической кибернетики,
Сибирский государственный медицинский университет,
634050, Россия, Томск
[3] Кафедра фармацевтического анализа,
Сибирский государственный медицинский университет,
634050, Россия, Томск
[4] Лаборатория клеточных и микрофлюидных технологий,
Сибирский государственный медицинский университет,
634050, Россия, Томск

* Автор-корреспондент: khlusov.ia@ssmu.ru

Резюме: Современная фармацевтическая наука претерпевает значительные изменения благодаря активному внедрению компьютерных технологий в процесс разработки лекарственных средств. Перед традиционными этапами синтеза и экспериментальной проверки исследователи все чаще применяют *in silico* подходы, позволяющие с высокой точностью моделировать структуру и фармакологические эффекты химических соединений. Такой подход обеспечивает значительную экономию временных и финансовых ресурсов, оптимизируя последующие экспериментальные исследования. Параллельно с этим наблюдается активное развитие эпигеном-направленной терапии — нового подхода, позволяющего модулировать экспрессию генов, вовлеченных в патологические процессы, без прямого воздействия на первичную структуру ДНК. В данной работе на примере хелидоновой кислоты представлен комплексный алгоритм оценки влияния малых молекул на экспрессию генов и метаболических путей. Методика основана на использовании веб-сервиса DIGEP-Pred 2.0 с последующим многоуровневым биоинформатическим анализом, включающим: (1) анализ избыточной репрезентативности генов; (2) оценку вовлеченности метаболических путей; (3) их функциональную кластеризацию. Предлагаемый подход способствует решению задач как для фундаментальных, так и для прикладных исследований механизмов действия лекарственных веществ.

Ключевые слова: биоинформатика, компьютерное моделирование, внутриклеточные процессы, молекулярные мишени, методические аспекты

Для цитирования:

Насибов Т. Ф., Горохова А. В., Бразовский К. С., Порохова Е. Д., Авдеева Е. Ю., Белоусов М. В., Хлусов И. А. Алгоритм *in silico* прогнозирования генов-мишеней и целевых метаболических путей для малых биомолекул на примере хелидоновой кислоты. *Современные направления в биомедицине*. 2025;1(1):06—19. https://doi.org/10.5922/ATB-2025-1-1-1

Поступила
23.04.2025 г.
Прошла рецензирование
04.06.2025 г.
Принята к печати
05.06.2025 г.
Опубликована
15.07.2025 г.

© Насибов Т. Ф., Горохова А. В., Бразовский К. С., Порохова Е. Д., Авдеева Е. Ю., Белоусов М. В., Хлусов И. А., 2025

Background and the Theory

Pharmaceutical research and development are expensive, time-consuming, and risky processes [1]. The process of targeted drug synthesis is challenging due to limited knowledge of many molecular targets, as well as the high cost of researching their structures and ligands [2]. *In silico* technologies have become essential in the modern pharmaceutical industry, as they can significantly reduce the time and resources required for drug development [3]. These technologies include predictive data analysis, building models with specified properties, and virtual screening of large chemical libraries in order to find effective drug candidates, making it possible to solve many problems of modern pharmacy [4]. Thus, more and more scholarly attention is now focused on the strategy of computer-aided drug design (CADD), which includes computational identification of potential drug targets [5], virtual screening of large chemical libraries to identify effective drug candidates [6], and further optimization of candidate compounds and computational assessment of their potential toxicity [7]. Following these computational processes, candidate compounds undergo *in vitro*/*vivo* experiments for confirmation [8]. Therefore, in contrast to the traditional approach, which involves the expensive synthesis of numerous compounds followed by their experimental testing for biological activity, selectivity, and toxicity [9], CADD methods can significantly reduce the number of compounds tested and increase the chances of success by eliminating ineffective and toxic substances first. Virtual screening, for instance, has been shown to increase efficiency (defined as the number of compounds that bind at a given concentration divided by the number of compounds tested experimentally) by approximately 100—1000 times when compared to random screening [10].

In the modern pharmaceutical industry, there is a growing interest in developing drugs targeting genes involved in pathological metabolic pathways [11]. This approach, known as epigenome-directed or epigenetic therapy [12; 13], aims to modulate gene expression without altering the DNA sequence. Epigenetic therapies involve the use of drugs that selectively affect the activity of specific genes, leading to changes in their expression [14]. These drugs can be used to treat a variety of diseases, including cancer and metabolic disorders. Currently, predicting changes in gene expression under the influence of a particular compound without the use of computer technology is impossible [15; 16]. An important tool in the creation of new drugs has become the assessment of structural similarity based on the principle that similar molecules have similar biological activity [17]. The first tool to predict the biological activity of compounds based on their structure is the PASS (Prediction of Activity Spectra for Substances) computer program [18]. The training set for this program consists of information on the structures and biological activities of compounds obtained from large experimental databases such as ChEMBL, PubChem and PASS, which together contain more than 300 million records. The prediction accuracy of PASS exceeds 0.96 for over 5,000 different types of biological activities [19]. In order to predict gene expression profiles, the DIGEP-Pred (Prediction of drug-induced changes in gene expression profile) program was developed [20].

The DIGEP-Pred 2.0 web service allows the users to model *in silico* the effect of substances on gene expression profiles based on their molecular structure [21]. To make predictions, the service uses literary data from the Comparative Toxicogenomics Database (CTD) as well as experimental data obtained from microarray analysis of the MCF7, PC3, and HL60 cell lines. The accuracy of the predictions reaches 86.5 % for models trained on the CTD and over 87 % for those based on experimental data [21]. The prediction results are presented as three key parameters for each gene: the probability of being active (Pa), reflecting the similarity of the molecule to known active compounds in the PASS training set; the probability of being inactive (Pi), showing the structure resemblance to inactive compounds in the same set; and the invariant

prediction accuracy (IAP), obtained through the leave-one-out cross-validation (LOO CV) procedure and numerically equal to the AUC value of the ROC curve. LOO CV is used to assess the robustness of the model on the full PASS training set.

Technical requirements

The DIGEP-Pred 2.0 web service does not require local installation and is available online (https://www.way2drug.com/digep-pred/). The R programming language (version 4.5.0), distributed under the GNU GPL v2 license (https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html), was used for data processing. The minimum system requirements and installation instructions for working with R can be found on the official product page (https://www.r-project.org/). To ensure proper operation of the used packages, it is recommended to use the 64-bit version of R. Bioinformatics analysis was performed using the following R packages: "clusterProfiler" [22—25)], "DOSE" [26], "AnnotationDbi" [27], "org.Hs.eg.db" [28], "KEGGREST" [29], "GO.db" [30], "rWikiPathways" [31]. All packages were installed via Bioconductor version 3.21.0. Basic R "graphics" and "ggplot2" were used for general data visualization, as well as the specialized "RCy3" package [32] for more advanced graphics.

Description of the Pipeline

The first step is to obtain a small molecule structure file. There are several restrictions to consider for the molecule, including electrical neutrality, consisting of a single component, the presence of covalent bonds only, and a maximum molecular weight of 1,250 atomic mass units; for a complete list of requirements, please see publication [19]. If there is *a priori* information about the compound under study, it is recommended to search in the open chemical databases such as PubChem (https://pubchem.ncbi.nlm.nih.gov/) and CheMBL (https://www.ebi.ac.uk/chembl/). From these databases, you can download the 2D structure in SDF format or a SMILES (simplified molecular input line entry system) line. If the substance is new or the information is not available from the previously mentioned databases, researchers need to depict the structure themselves using the DIGEP-Pred 2.0 web service. Since chelidonic acid (ChA) has a known chemical structure, a search in PubChem revealed the SMILES string C1 = C(OC(= CC1 = O)C(= O)O)C(= O)O.

The next step is to use DIGEP-Pred 2.0, allowing the researcher to obtain a list of genes that may change their expression under the effect of the studied compound. However, it is important to remember that DIGEP-Pred 2.0 does not provide information about the causes of these changes. In addition to the compound being studied, DIGEP-Pred 2.0 requires other input parameters such as (1) the dataset used to train the model, (2) the threshold of Pa values, and (3) the direction of expression changes (Up/Down). The choice of the first parameter (dataset) is crucial and depends on the study design and the intended methods for verifying the results *in vitro*/*vivo*. The model creators provide the option to use both long-term literature data from the CTD or experimental data from three cell lines: MCF7, PC3, and HL60. When studying the general effects of a test substance, the CTD_mRNA or CTD_protein training sets from the literature are recommended, as they offer extensive gene output data. However, experimental validation may be challenging given the data's generalization across diverse cell lines. At the same time, using experimental training data can solve this problem, but the resulting list of genes will be significantly smaller compared to the literature sets. It is also worth noting that experimental sets are divided by logarithmic (base 2) fold change in gene expression (LFC). To identify all genes with modulated expression, run the simulation for each LFC value (0.5, 0.7, 1, 1.5, and 2). For clarity, *Table 1* provides an example.

Table 1

Comparison of simulation results using different training data sets

Dataset name	Number of genes obtained by modeling the effects of ChA
CTD_mRNA	9,151
CTD_protein	2,133
cMAP_HL60 (LFC 0.5 + 0.7 + 1 + 1.5 + 2)	3,871 (1,809 + 1,194 + 589 + 193 + 86)
cMAP_MCF7 (LFC 0.5 + 0.7 + 1 + 1.5 + 2)	1,603 (762 + 497 + 237 + 72 + 35)
cMAP_PC3 (LFC 0.5 + 0.7 + 1 + 1.5 + 2)	895 (396 + 275 + 133 + 54 + 37)

Note: The remaining input parameters were kept constant for all simulations: the lower limit of the values was Pa > Pi, and the direction of regulation was Up.

The choice of the lower limit for the Pa value directly affects the number of results obtained. A higher threshold increases the likelihood of detecting the test substance's activity in relation to the identified genes, but it also reduces the number of genes selected. The authors of the model propose two approaches for choosing the lower limit value [18]. For initial modeling, it is reasonable to use the value Pa > Pi. For further analysis, either of the two proposed approaches can be applied. The direction of expression changes determines which genes are included in the list: the expression of those increases or decreases u

nder the influence of the substance. For a complete analysis, both lists are needed. In this study, the input parameters as follows were used: training dataset CTD_mRNA, thresholds Pa > Pi and Pa > 0.4 [18], and both directions of expression change — Up and Down. As a result of the *in silico* prediction of the gene expression assessment under the influence of ChA, a list of genes with corresponding expression directions (over- or under-expression), activity values (Pa), inactivity values (Pi), and the prediction accuracy score (IAP value) was obtained. This resulted in 9,151 over-expressed genes and 9,207 under-expressed genes, with approximately 5,000 genes present in both lists. The authors of the model do not describe the process for dealing with the bidirectional influence of a substance on the expression of the same gene, so we introduced a weighted coefficient (CA) calculated using the formula: CA = Pa * IAP. The CAs for the same genes in each list were compared, and the gene with the lowest CA value was removed from the corresponding list. This resulted in the reduced up-list of 6,309 genes and the down-list of 7,023 genes. Due to the large number of genes involved in the analysis and the difficulty in assessing the effects of over- or under-expression of a single gene in isolation from others, we did not perform a detailed analysis of each individual gene. However, a detailed analysis of individual genes can be carried out if required by the specific objectives of the study.

Gene set enrichment analysis (GSEA) and over-representation analysis (ORA) are two commonly used methods for grouping genes into metabolic pathways. GSEA uses *a priori* sets of genes that are grouped based on their involvement in the same biological process. It then determines whether genes from this functional set are clustered at the top (over-expression) or bottom (under-expression) of a ranked gene list. Meanwhile, ORA tests whether genes from a certain set (such as a biological pathway) are represented in this list more often than expected by chance.

Owing to the lack of information on the degree of gene expression change, we opted for ORA. In ORA, genes were grouped according to a metabolic pathway or common gene ontology in which their products are involved. The statistical significance of their association of a particular group of genes was then assessed using Fisher's exact test. The list of possible metabolic pathways was obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and WikiPathways databases. The gene ontology was obtained from the Gene Ontology database (GO). A simple and easy-

to-use ORA is presented in the "clusterProfiler" package [22—25], which contains several functions for performing ORA in each database ("enrichKEGG()", "enrichWP()", "enrichGO()"). The researchers can find a full description of these functions in the documentation at https://www.bioconductor.org/packages/release/bioc/manuals/enrichplot/man/enrichplot.pdf. As a result, a table was generated for each database containing the following information: the metabolic pathway identifier, its short name, a list of gene products involved in the pathway, their number, and significance estimates — p.adjust and q-value. The p.adjust represents a p-value adjusted using the Benjamini—Hochberg procedure, which determines the expected proportion of false positive results (FDR) among all positive (significant) ones. In turn, the q-value determines the minimum FDR when a result is considered significant. While the choice of the significance estimation ultimately rests with the researcher, we would like to note that p.adjust is a more rigorous estimate, and q-value has higher statistical power. Our study was a pilot one, and therefore, it was preferable to obtain as many candidate pathways with the best metrics as possible for further investigation. With this in mind, we chose to use the q-value. The number of significant (q-value < 0.05) pathways and ontologies are given in Table 2.

Table 2

The number of significant (q-value < 0.05) metabolic pathways / ontologies identified during ORA in different databases

Gene expression directions	Database		
	GO	KEGG	WikiPathways
Up	1,016	53	2
Down	1,258	86	48

One can notice a large variability in the number of pathways / ontologies obtained. For researchers without relevant experience, the gold standard is KEGG, as it is regularly updated and has strict data moderation. We recommend using GO when functional annotation of genes is required, and WikiPathways when more detailed or new information not present in KEGG is needed. Thus, manual assessment of pathways from KEGG and WikiPathways seems quite feasible under conditions of limited human and time resources, without considering gene ontologies.

The next step could be to explore possible metabolic pathways associated with the *a priori* target effects, if such effects are present. In our case, such an effect may be the stimulation of osteogenesis. This possibility was previously demonstrated by our team in [33; 34]. As a result of ORA, the up-regulated genes were combined into eight significant (q-value < 0.05) metabolic pathways related to osteogenesis, while the down-regulated genes were combined into four metabolic pathways (Table 3).

Table 3

Metabolic pathways obtained by ORA for ChA target genes

Pathway / Ontology ID	Name	Genes count	Pa values of genes included in the pathway (Me (Q_1;Q_3)
The overexpressed genes (Up-regulation)			
GO:0097553	Ca^{2+} ion transmembrane import into cytosol	66	0.72 (0.65; 0.82)
GO:0030282	Bone mineralization	47	0.72 (0.65; 0.79)

The end of Table 3

Pathway / Ontology ID	Name	Genes count	Pa values of genes included in the pathway (Me (Q_1;Q_3)
GO:0045667	Regulation of osteoblast differentiation	55	0.75 (0.67; 0.82)
GO:0009612	Response to mechanical stimulus	77	0.70 (0.63; 0.75)
GO:0030198	Extracellular matrix organization	116	0.70 (0.62; 0.78)
hsa04064	NF-kappa B signaling pathway	42	0.71 (0.64; 0.81)
hsa04310	Wnt signaling pathway	63	0.77 (0.70; 0.82)
WP4141	Vitamin D3 signaling	16	0.75 (0.62; 0.83)
The underexpressed genes (Down-regulation)			
GO:0030178	Negative regulation of Wnt signaling pathway	75	0.78 (0.68; 0.83)
hsa04010	MAPK signaling pathway	132	0.75 (0.66; 0.84)
hsa04668	TNF signaling pathway	61	0.74 (0.64; 0.83)
WP4787	Osteoblast differentiation and related diseases	59	0.80 (0.68; 0.86)

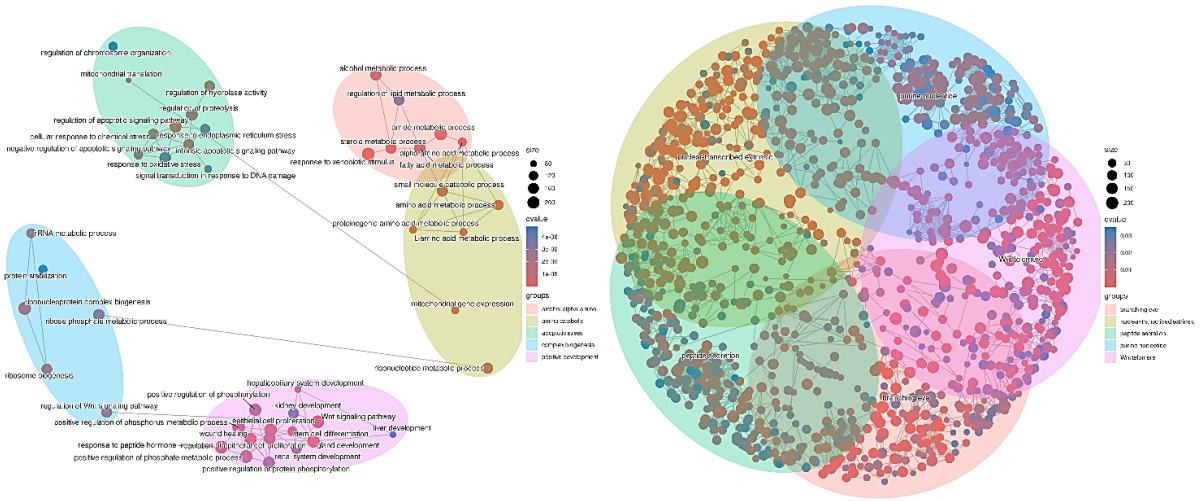
Therefore, a number of metabolic pathways and ontologies containing candidate genes were obtained for further experimental verification *in vitro* / *vivo*. The presented algorithm is effective in the targeted search for metabolic pathways associated with specific biological processes.

In cases when the researcher does not know in advance what processes the substance under study may affect, we propose to first assess the overall direction of the effects by clustering all the identified metabolic pathways into larger modules. This procedure is particularly relevant when dealing with a large number of identified pathways and ontologies. The set of tools necessary to solve problems like this is included in the "clusterProfiler" package [22—25] for the R programming language. To illustrate this approach, we used the previously obtained gene ontology sets. The ontologies in each list were grouped using the "pairwise_termsim()" function (a detailed description of the parameters can be found at https://www.bioconductor.org/packages/release/bioc/manuals/enrichplot/man/enrichplot.pdf), based on the determination of pairwise semantic similarity to find the most informative common ancestor in the ontology hierarchy for each pair of terms.

Clustering was performed using the Partitioning Around Medoids (PAM) algorithm, based on the definition of k-medoids. This classical clustering method partitions a dataset into k clusters, where k is assumed to be known *a priori*. The cluster median is the object in the cluster with the smallest sum of distances to all other objects in the same cluster, making it the most central point within the cluster [35]. Unlike the k-means, the k-medoids only selects existing data points as centers, providing a clearer interpretation of cluster centers compared to k-means, where the cluster center may not be one of the original data points, but rather the average of the points within the cluster. Additionally, the k-medoids algorithm can be used with arbitrary dissimilarity measures, whereas the k-means algorithm typically requires Euclidean distance [36]. The k-medoids approach minimizes the sum of pairwise differences between data points instead of the sum of squared distances, making it more robust to noise and outliers than the k-means approach [37].

Clustering and visualization of the results is implemented in the "emapplot()" function (detailed description can be found at https://www.bioconductor.org/packages/release/bioc/manuals/enrichplot/man/enrichplot.pdf), which also creates a network enrichment map. This graph combines ontologies, represented as nodes, into a

network with edges connecting overlapping sets of genes. Thus, intersecting sets of genes tend to be grouped together, making it easier to identify functional modules. The color of a node reflects the statistical significance of the enrichment, and the size of the node represents the number of genes associated with each term. This allows us to quickly identify important or highly enriched terms. Figure 1 presents the clustering results for the 45 most significant (q-value < 0.05) gene ontologies of the up-list (Fig. 1, *a*) and all 1,016 significant (q-value < 0.05) ontologies (Fig, 1, *b*) from this list. A closer examination of the first 45 ontologies (Fig, 1, *a*) reveals that the "ossification" ontology, which means bone tissue formation, bone matrix mineralization and etc. in GO terms, is included in the cluster of the canonical NF-kB metabolic pathway (blue oval, Figure 1, *a*). This indirectly confirms our selection in Table 3. The study of all significant (q-value < 0.05) ontologies (Fig. 1, *b*) is primarily necessary for identifying the general directions of the molecule's action. Therefore, listing all ontologies in this case is unnecessary.



a *b*

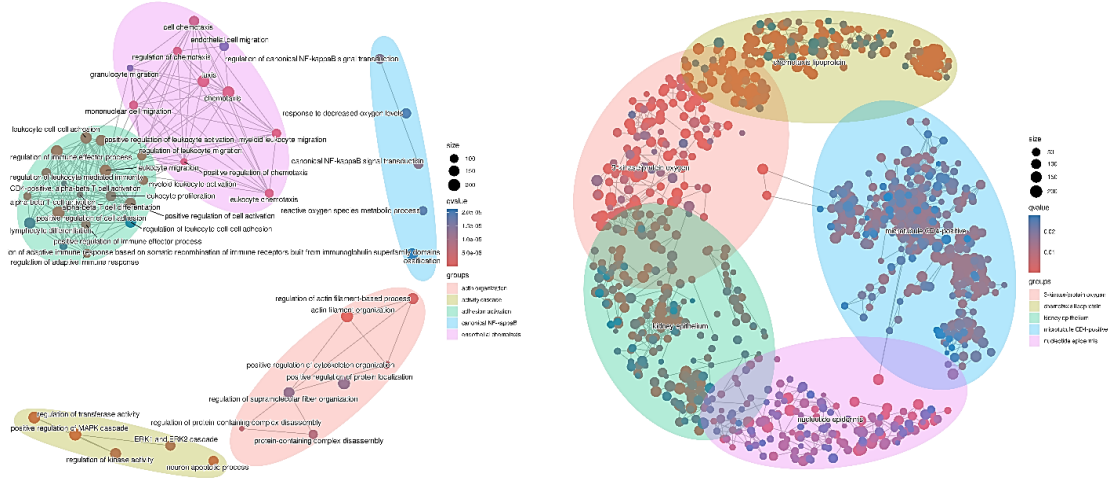
Fig. 1. The example of clustering of significant (q-value < 0.05) up-list gene ontologies: the first 45 (*a*) and all 1,016 (*b*)

Similarly, Figure 2 shows the clusters for the 45 most significant (q-value < 0.05) ontologies of the down-list (Fig. 2, *a*) and all 1,258 significant (q-value < 0.05) ontologies of the corresponding list (Fig. 2, *b*).

At the same time, it should be noted that as the number of ontologies considered for the two lists increases, the patterns of cluster arrangement in the space of semantic similarity coefficient values are preserved. This could indicate the appropriate use of the "emapplot()" function's parameters.

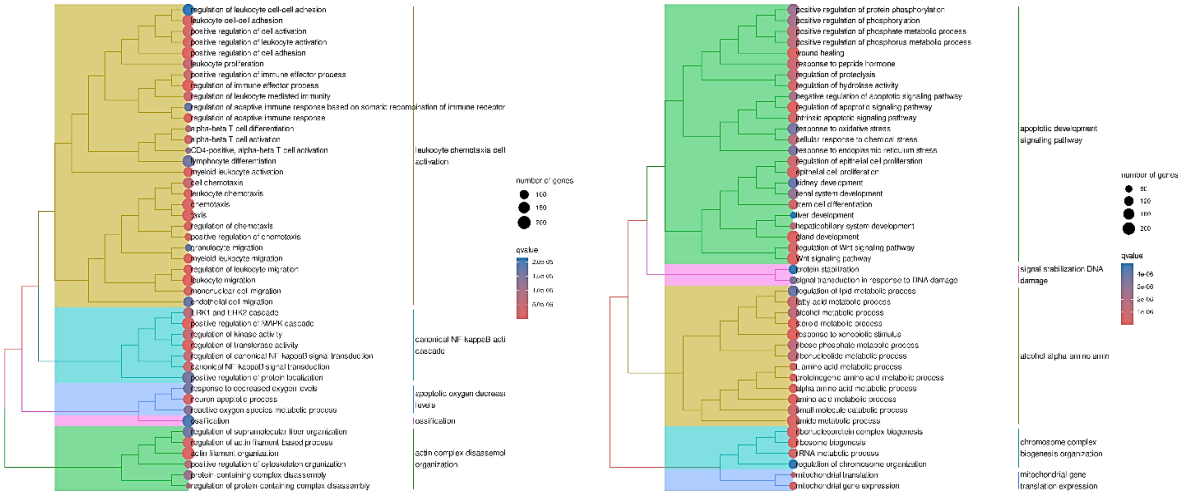
In addition to clustering, the "clusterProfiler" package allows you to explore the possible hierarchy of the obtained ontologies. For this purpose, the "treeplot()" function is provided, which enables the creation of a tree-like graph of semantic similarity (Fig. 3). Researchers can find a detailed description of this function at (https://www.bioconductor.org/packages/release/bioc/manuals/enrichplot/man/enrichplot.pdf). As a result, "treeplot()" creates a graph where each path is represented by a node, and parent-child relationships between terms are indicated by connecting lines. Similar to a network map, the size of each node depends on the number of genes included in the ontology, while the color indicates the statistical significance of gene enrichment within the ontology. In our example, the output of "treeplot()" for the 45 most

significant (q-value < 0.05) ontologies of the up-list (Fig. 3, *a*) and down-list (Fig. 3, *b*) identifies 5 groups, whose ontologies are related to each other through a common parent term.



a *b*

Fig. 2. The example of clustering of significant (q-value < 0.05) down-list gene ontologies: the first 45 (*a*) and all 1,258 (*b*)



a *b*

Fig. 3. The example of the possible hierarchy of gene ontologies of the up-list (*a*) and down-list (*b*)

Conclusions

The developed pipeline demonstrates complex capabilities. On the one hand, it allows analyzing the effect of chemical compounds on *a priori* known genes and metabolic pathways; on the other hand, it helps to identify potentially new pharmacological effects of the compounds under study. This toolkit is of particular value for preclinical studies, where rapid data processing and cost-effective methods are essential. As modern studies show, the use of *in silico* approaches significantly

optimizes the screening of biologically active compounds [38]. In particular, similar methods can save 30—40 % of the time compared to traditional approaches [39] and significantly reduce financial costs during the initial stages of new drug development [3; 40]. In addition, the application of this approach in personalized medicine, where rapid analysis of individual genetic and metabolic profiles of patients is crucial, is of particular practical interest. These results open up prospects for further development of this area of research. The authors hope that the proposed methodology will become a useful tool for the scientific community, combining three key advantages: (1) high analytical efficiency, (2) simplicity of use, and (3) cost-effectiveness. The presented approach can be widely applied in both fundamental studies of the mechanisms of medicinal substance effects and in practical applications across molecular biology, pharmacology, and bioinformatics.

Funding. The research was conducted with the financial and organizational support of BioGlobus LLC (Moscow) and Innovative Communications LLC (Tomsk).

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

Финансирование. Исследование проводилось при финансовой и организационной поддержке ООО «БиоГлобус» (Москва) и ООО «Инновационные коммуникации» (Томск).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

References

1. Paul S. M., Mytelka D. S., Dunwiddie C. T., Persinger C. C., Munos B. H., Lindborg S. R., et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat. Rev. Drug Discov. 2010, 9(3), 203—214, DOI: 10.1038/nrd3078.

2. Swinney D. C., Anthony J. How were new medicines discovered? Nat. Rev. Drug Discov. 2011, 10(7), 507—19, DOI: 10.1038/nrd3480.

3. Ekins S., Puhl A. C., Zorn K. M., Lane T. R., Russo D. P., Klein J. J., et al. Exploiting machine learning for end-to-end drug discovery and development. Nat. Mater. 2019, 18(5), 435—441, DOI: 10.1038/s41563-019-0338-z.

4. Yang K., Swanson K., Jin W., Coley C., Eiden P., Gao H., et al. Correction to Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model. 2019, 59(12), 5304—5305, DOI: 10.1021/acs.jcim.9b01076.

5. Anighoro A., Bajorath J., Rastelli G. Polypharmacology: Challenges and Opportunities in Drug Discovery. J. Med. Chem. 2014, 57(19), 7874—7887, DOI: 10.1021/jm5006463.

6. Lyu J., Wang S., Balius T. E., Singh I., Levit A., Moroz Y. S., et al. Ultra-large library docking for discovering new chemotypes. Nature. 2019, 566(7743), 224—229, DOI: 10.1038/s41586-019-0917-9.

7. Shaker B., Ahmad S., Lee J., Jung C., Na D. In silico methods and tools for drug discovery. Computers in Biology and Medicine. 2021, 137, 104851, DOI: 10.1016/j.compbiomed.2021.104851.

8. Macarron R., Banks M. N., Bojanic D., Burns D. J., Cirovic D. A., Garyantes T., et al. Impact of high-throughput screening in biomedical research. Nat. Rev. Drug. Discov. 2011, 10(3), 188—195, DOI: 10.1038/nrd3368.

9. Drews J. Drug discovery: a historical perspective. Science. 2000, 287(5460), 1960—1964, DOI: 10.1126/science.287.5460.1960.

10. Tang Y., Zhu W., Chen K., Jiang H. New technologies in computer-aided drug design: Toward target identification and new chemical entity discovery. Drug Discovery Today: Technologies. 2006, 3(3), 307—313, DOI: 10.1016/j.ddtec.2006.09.004.

11. Dawson M. A., Kouzarides T. Cancer Epigenetics: From Mechanism to Therapy. Cell. 2012, 150(1), 12—27, DOI: 10.1016/j.cell.2012.06.013.

12. Omarov M. A., Mulyukov A. R., Khalitov R. V., Safarov S. I., Ayupova G. U., Demianenko O. N., et al. Epigenetic modulation in medicine: Regulation of gene expression in the context of pathogenesis and therapy. Acta Biomedica Scientifica. 2024, 9(6), 22—33, DOI: 10.29413/ABS.2024-9.6.3.

13. Allis C. D., Jenuwein T. The molecular hallmarks of epigenetic control. Nat. Rev. Genet. 2016, 17(8), 487—500, DOI: 10.1038/nrg.2016.59.

14. Kelly T. K., De Carvalho D. D., Jones P. A. Epigenetic modifications as therapeutic targets. Nat. Biotechnol. 2010, 28(10), 1069—1078, DOI: 10.1038/nbt.1678.

15. Huang G., Yan F., Tan D. A Review of Computational Methods for Predicting Drug Targets. Current Protein & Peptide Science. 2018, 19(6), 562—572, DOI: 10.2174/1389203718666161114113212.

16. Sun X., Hu B. Mathematical modeling and computational prediction of cancer drug resistance. Briefings in Bioinformatics. 2018, 19(6), 1382—1399, DOI: 10.1093/bib/bbx065.

17. Bender A., Glen R. C. Molecular similarity: a key technique in molecular informatics. Org. Biomol. Chem. 2004, 2(22), 3204—3218, DOI: 10.1039/B409813G.

18. Poroikov V. V. Search for new pharmacological substances based on computer prediction of biological activity spectra. Laboratory and production. 2021, 1(16), 72—80, DOI: 10.32757/2619-0923.2021.1.16.72.80.

19. Filimonov D. A., Druzhilovskiy D. S., Lagunin A. A., Gloriozova T. A., Rudik A. V., Dmitriev A. V., et al. Computer-aided Prediction of Biological Activity Spectra for Chemical Compounds: Opportunities and Limitations. Biomedical Chemistry: Research and Methods. 2018, 1(1), e00004—e00004, DOI: 10.18097/BMCRM00004.

20. Lagunin A., Ivanov S., Rudik A., Filimonov D., Poroikov V. DIGEP-Pred: web service for in silico prediction of drug-induced gene expression profiles based on structural formula. Bioinformatics. 2013, 29(16), 2062—2063, DOI: 10.1093/bioinformatics/btt322.

21. Ivanov S. M., Rudik A. V., Lagunin A. A., Filimonov D. A., Poroikov V. V. DIGEP-Pred 2.0: A web application for predicting drug-induced cell signaling and gene expression changes. Mol. Inform. 2024, 43(12), e202400032, DOI: 10.1002/minf.202400032.

22. Yu G. Thirteen years of clusterProfiler. Innovation (Camb). 2024, 5(6), 100722, DOI: 10.1016/j.xinn.2024.100722.

23. Xu S., Hu E., Cai Y., Xie Z., Luo X., Zhan L., et al. Using clusterProfiler to characterize multiomics data. Nat. Protoc. 2024, 19(11), 3292—3320, DOI: 10.1038/s41596-024-01020-z.

24. Wu T., Hu E., Xu S., Chen M., Guo P., Dai Z., et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021, 2(3), 100141, DOI: 10.1016/j.xinn.2021.100141.

25. Yu G., Wang L. G., Han Y., He Q. Y. ClusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology. 2012, 16(5), 284—287, DOI: 10.1089/omi.2011.0118.

26. Yu G., Wang L. G., Yan G. R., He Q. Y. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015, 31(4), 608—609, DOI: 10.1093/bioinformatics/btu684.

27. Pagès H., Carlson M., Falcon S., Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. AnnotationDbi, R package version 1.70.0. 2025, DOI: 10.18129/B9.bioc.

28. Carlson M. org.Hs.eg.db: Genome wide annotation for Human, DOI: 10.18129/B9.bioc.org.Hs.eg.db.

29. Tenenbaum D., Maintainer B. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.48.0. 2025, DOI: 10.18129/B9.bioc.KEGGREST.

30. Carlson M. GO.db: A set of annotation maps describing the entire Gene Ontology. R package version 3.21.0, DOI: 10.18129/B9.bioc.GO.db.

31. Slenter D. N., Kutmon M., Hanspers K., Riutta A., Windsor J., Nunes N., et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2018, 46(D1), D661—667, DOI: 10.1093/nar/gkx1064.

32. Gustavsen J. A., Pai S., Isserlin R., Demchak B., Pico A. R. RCy3: Network biology using Cytoscape from within R. F1000Res. 2019, 8, 1774, DOI: 10.12688/f1000research.20887.3.

33. Avdeeva E., Shults E., Rybalova T., Reshetov Y., Porokhova E., Sukhodolo I., et al. Chelidonic Acid and Its Derivatives from Saussurea Controversa: Isolation, Structural Elucidation and Influence on the Osteogenic Differentiation of Multipotent Mesenchymal Stromal Cells In Vitro. Biomolecules. 2019, 9(5), 189, DOI: 10.3390/biom9050189.

34. Avdeeva E., Porokhova E., Khlusov I., Rybalova T., Shults E., Litvinova L., et al. Calcium Chelidonate: Semi-Synthesis, Crystallography, and Osteoinductive Activity In Vitro and In Vivo. Pharmaceuticals. 2021, 14(6), 579, DOI: 10.3390/ph14060579.

35. Kaufman L., Rousseeuw P. J. Partitioning Around Medoids (Program PAM). John Wiley & Sons, Inc. 2008, 68—125, DOI: 10.1002/9780470316801.ch2.

36. Hennig C., Meila M., Murtagh F., Murtagh F., Rocci R., Eds. Handbook of cluster analysis. CRC Press. CRC Press: Boca Raton. 2015, DOI: 10.1201/B19706.

37. Schubert E., Rousseeuw P. J. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In: Amato G., Gennaro C., Oria V., Radovanović M., Eds. Similarity Search and Applications. Springer International Publishing: Cham. 2019, 171—187, DOI: 10.1007/978-3-030-32047-8_16.

38. Stokes J. M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., Donghia N. M., et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020, 180(4), 688—702, DOI: 10.1016/j.cell.2020.01.021.

39. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today. 2015, 20(3), 318—331, DOI: 10.1016/j.drudis.2014.10.012.

40. Elton D. C., Boukouvalas Z., Fuge M. D., Chung P. W. Deep learning for molecular design—a review of the state of the art. Mol. Syst. Des. Eng. 2019, 4(4), 828—849, DOI: 10.1039/C9ME00039A.

The authors

Temur F. Nasibov, Laboratory research assistant, Division of Morphology and General Pathology, Siberian State Medical University, Russia.

ORCID: 0000-0002-8056-3967

Anna V. Gorokhova, Laboratory research assistant, Division of Morphology and General Pathology, Siberian State Medical University, Russia.

ORCID: 0000-0001-8401-7181

Konstantin S. Brazovsky, Professor, D.Sc. (Eng.), Division of Medical and Biological Cybernetics, Siberian State Medical University, Russia.

ORCID: 0000-0002-4779-9820

Ekaterina D. Porokhova, Assistant lecturer, Division of Morphology and General Pathology, Siberian State Medical University, Russia.

ORCID: 0000-0002-7317-2036

Elena Yu. Avdeeva, Professor, D.Sc. (Pharm.), Division of Pharmaceutical Analysis, Siberian State Medical University, Russia.

ORCID: 000-0001-7061-9843

Mikhail V. Belousov, Head, D.Sc. (Pharm.) / Professorship, Division of Pharmaceutical Analysis, Siberian State Medical University, Russia.

ORCID: 0000-0002-2153-7945

Igor A. Khlusov, Head, D.Sc. (Med.) / Professorship, Laboratory of Cellular and Microfluidic Technologies, Siberian State Medical University, Russia.

ORCID: 0000-0003-3465-8452

Об авторах

Темур Фируддин оглы Насибов, лаборант-исследователь, кафедра морфологии и общей патологии, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0002-8056-3967

Анна Владимировна Горохова, лаборант-исследователь, кафедра морфологии и общей патологии, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0001-8401-7181

Константин Станиславович Бразовский, профессор, доктор технических наук, кафедра медицинской и биологической кибернетики, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0002-4779-9820

Екатерина Даниловна Порохова, ассистент, кафедра морфологии и общей патологии, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0002-7317-2036

Елена Юрьевна Авдеева, профессор, доктор фармацевтических наук, доцент, кафедра фармацевтического анализа, Сибирский государственный медицинский университет, Россия.

ORCID: 000-0001-7061-9843

Михаил Валерьевич Белоусов, заведующий кафедрой, доктор фармацевтических наук, профессор, кафедра фармацевтического анализа, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0002-2153-7945

Игорь Альбертович Хлусов, руководитель, доктор медицинских наук, профессор, лаборатория клеточных и микрофлюидных технологий, Сибирский государственный медицинский университет, Россия.

ORCID: 0000-0003-3465-8452