Way2Drug
XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

ANDSYSTEM AUTOMATED RECONSTRUCTION OF GENE NETWORKS FOR OMICS-DATA INTERPRETATION IN MEDICAL AND BIOLOGICAL RESEARCH

Ivanisenko V.A., Demenkov P.S.
Ivanisenko T.V., Ivanisenko N.V.,
Antropova E.A., Adamovskaya A.V.,
Revva P.M., Lavrik I.N., Kolchanov N.A.

ICG SB RAS

NSU

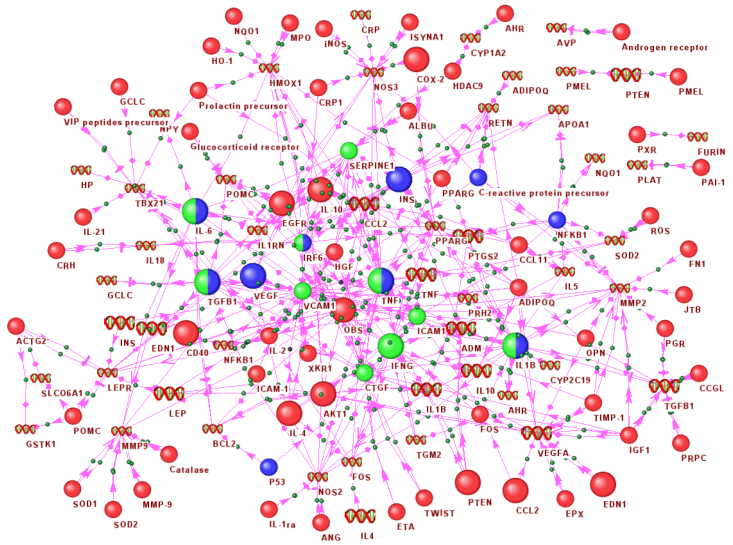
Novosibirsk 2022

Gene networks - a graph of knowledge about the molecular genetic mechanisms of diseases

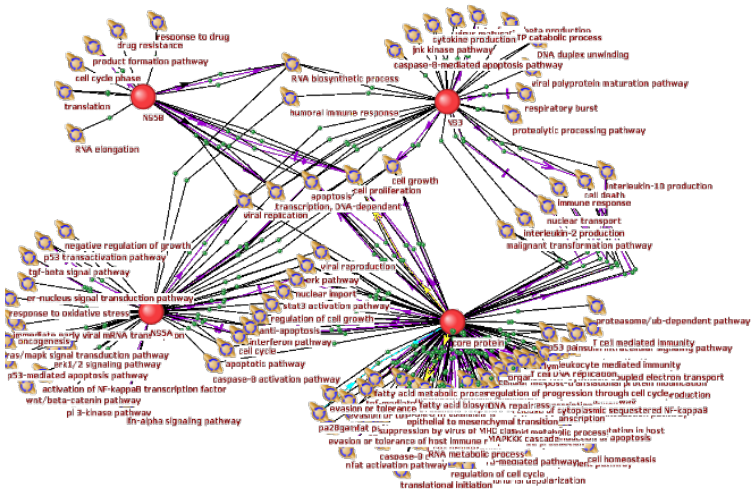
A gene network is a group of genes that function in a coordinated manner, controlling any phenotypic traits of an organism (Kolchanov et al, 2013)
The interactions between genes in the gene network are carried out through their primary and secondary products (RNA, proteins, metabolites).

An associative gene network is an extension of a **gene network**.
Additional network members:

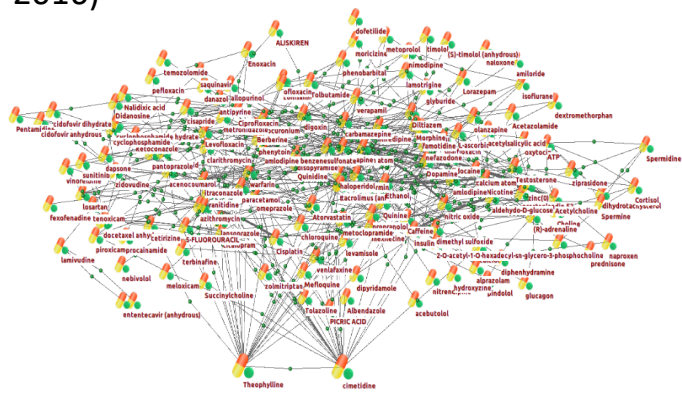
- environmental factors that are not a product of the functioning of genes (drugs, nutrition, etc)
- members of a higher level of organization than individual molecular compounds (diseases, biological processes, phenotypic traits, etc).



A gene network describing the genetic regulation of the expression of genes potentially involved in the comorbid state of asthma and hypertension (Saik et al, 2018)



An associative gene network of interactions of hepatitis C virus proteins with human biological processes (Saik et al, Virus Res 2016)



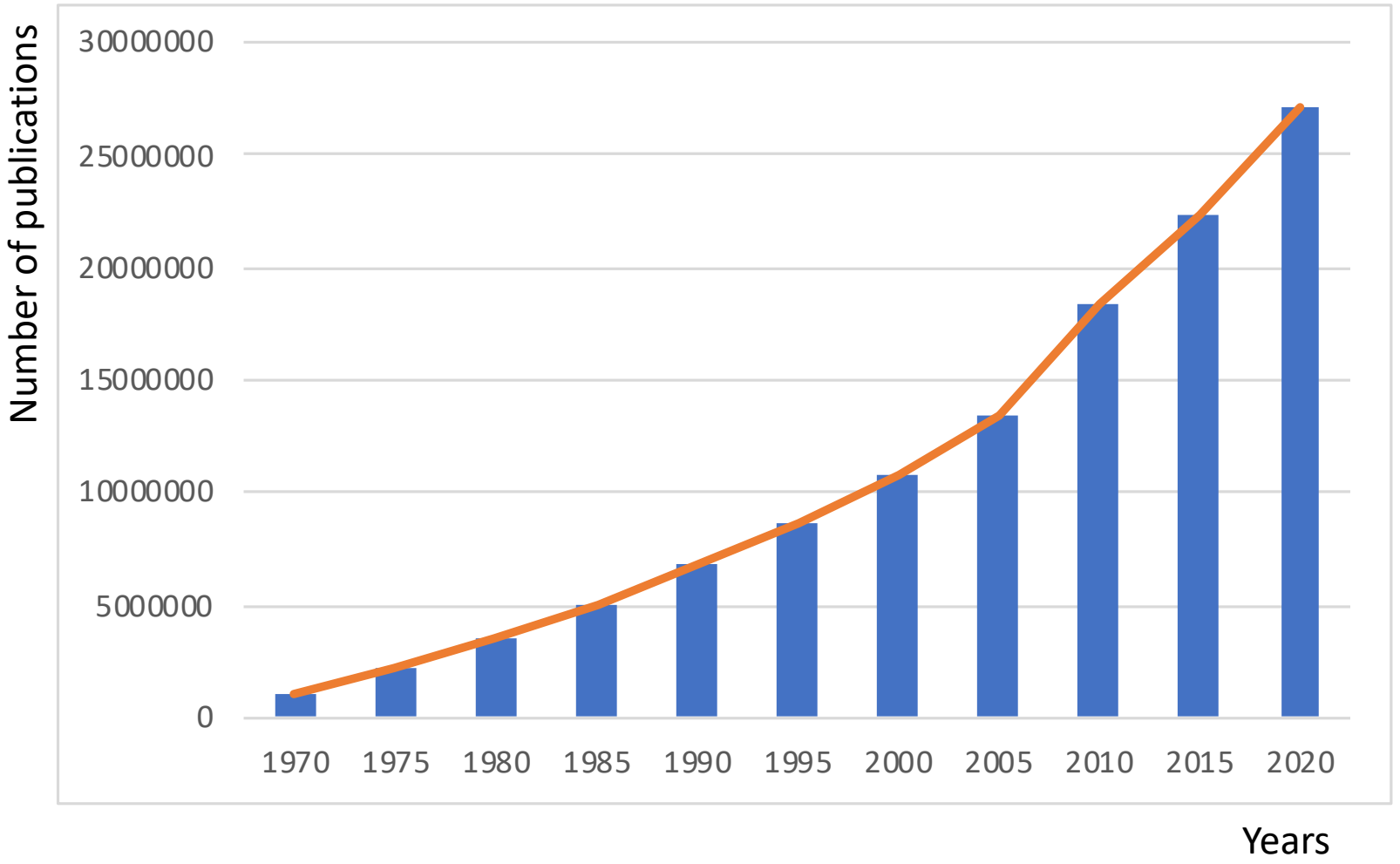
An associative gene network of drug-drug interactions

Gene networks are often used in:

- Interpretation of omix data (genomic, transcriptomic, proteomic, metabolomic and epigenomic)
- Identification of disease biomarkers
- Search for pharmacological targets
- Repurposing drugs
- etc.

PubMed Indexing Statistics

- PubMed contains about 30 million abstracts.
- The number of publications grows about 1 million over the years.
- To process such large amounts of information necessary to use methods of automated knowledge extraction.

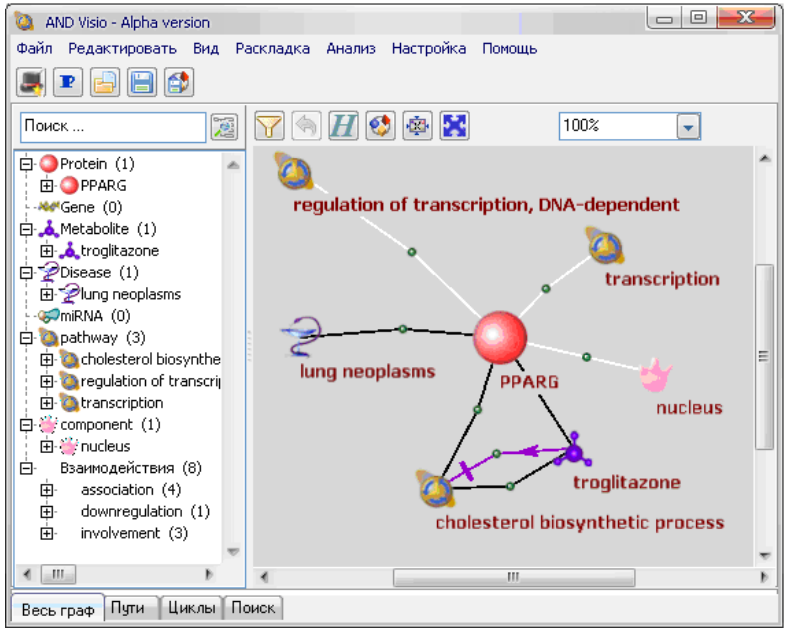


Cognitive software tools used for automated reconstruction of molecular-genetic networks

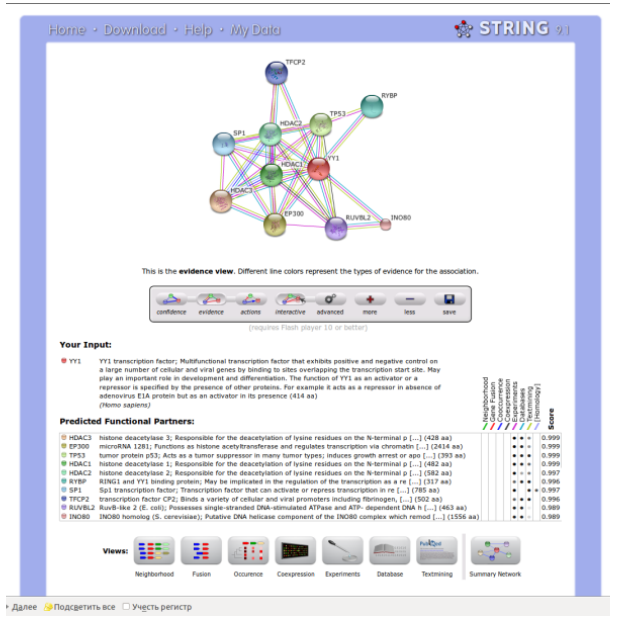
- Ontological description of the subject area (formalization of extracted knowledge)
- Automated knowledge extraction with text mining
- Integration of extracted knowledge and their presentation in the knowledge base
- User access to the knowledge base

ANDSystem

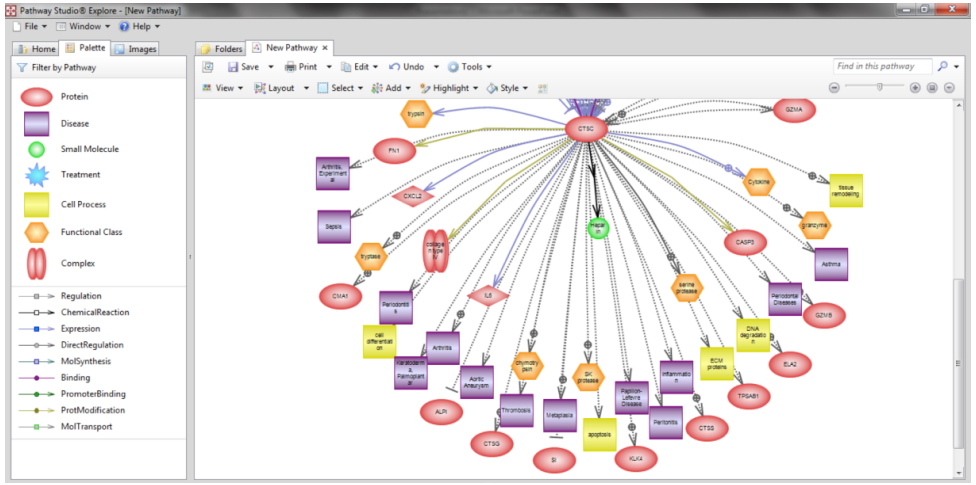
Ivanisenko et al, BMC Bioinformatics. 2019
and more than 15 additional publications



String

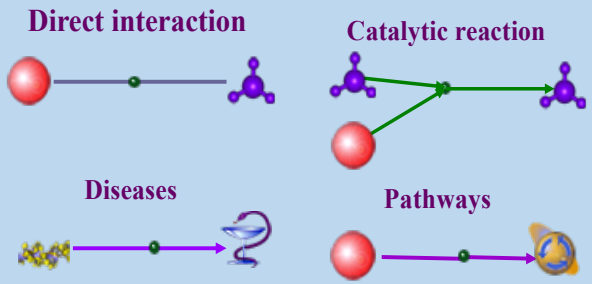


Pathway Studio

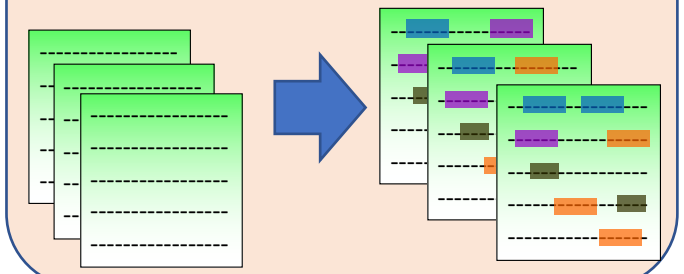


ANDSystem ontology

Object dictionaries, types of interaction between objects



Named entity recognition module

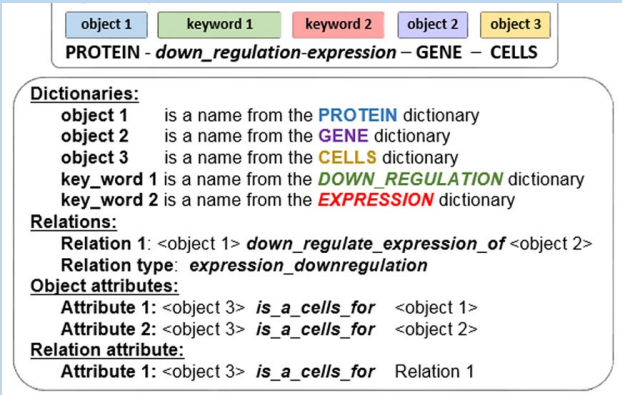


Factual Database Analysis Module

Extracting facts about interactions from structured data (IntAct, Mint, Gene Ontology и др.)

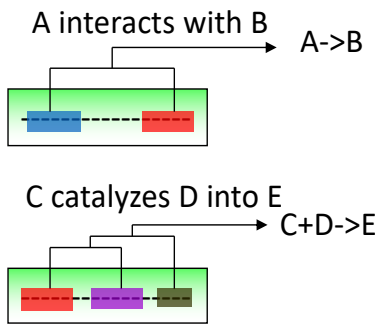
Semantic and linguistic rules for knowledge extraction (created by hand)

Over 10,000 rules



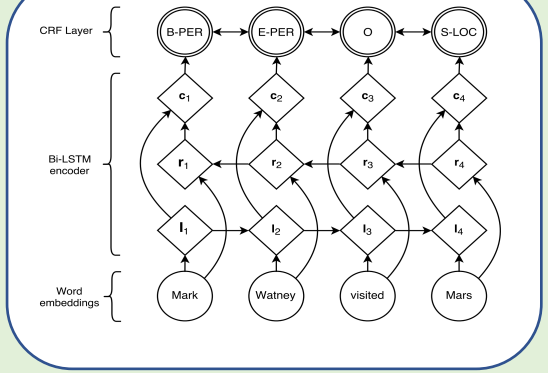
Module for extracting information about interactions from texts using artificial intelligence methods

Stage 1



Creation of a training set for AI methods using semantic linguistic rules

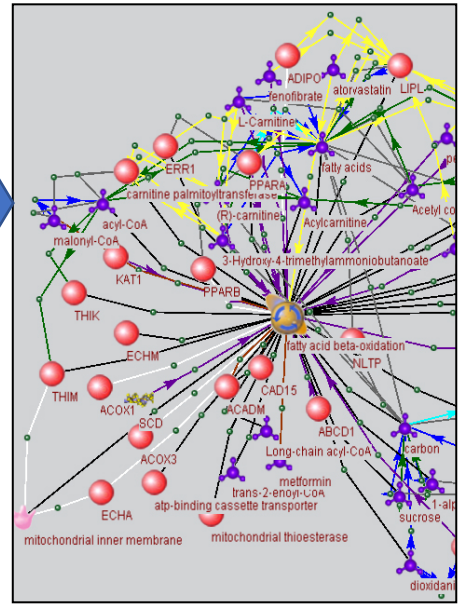
Stage 2



Training neural networks and extracting new knowledge using them

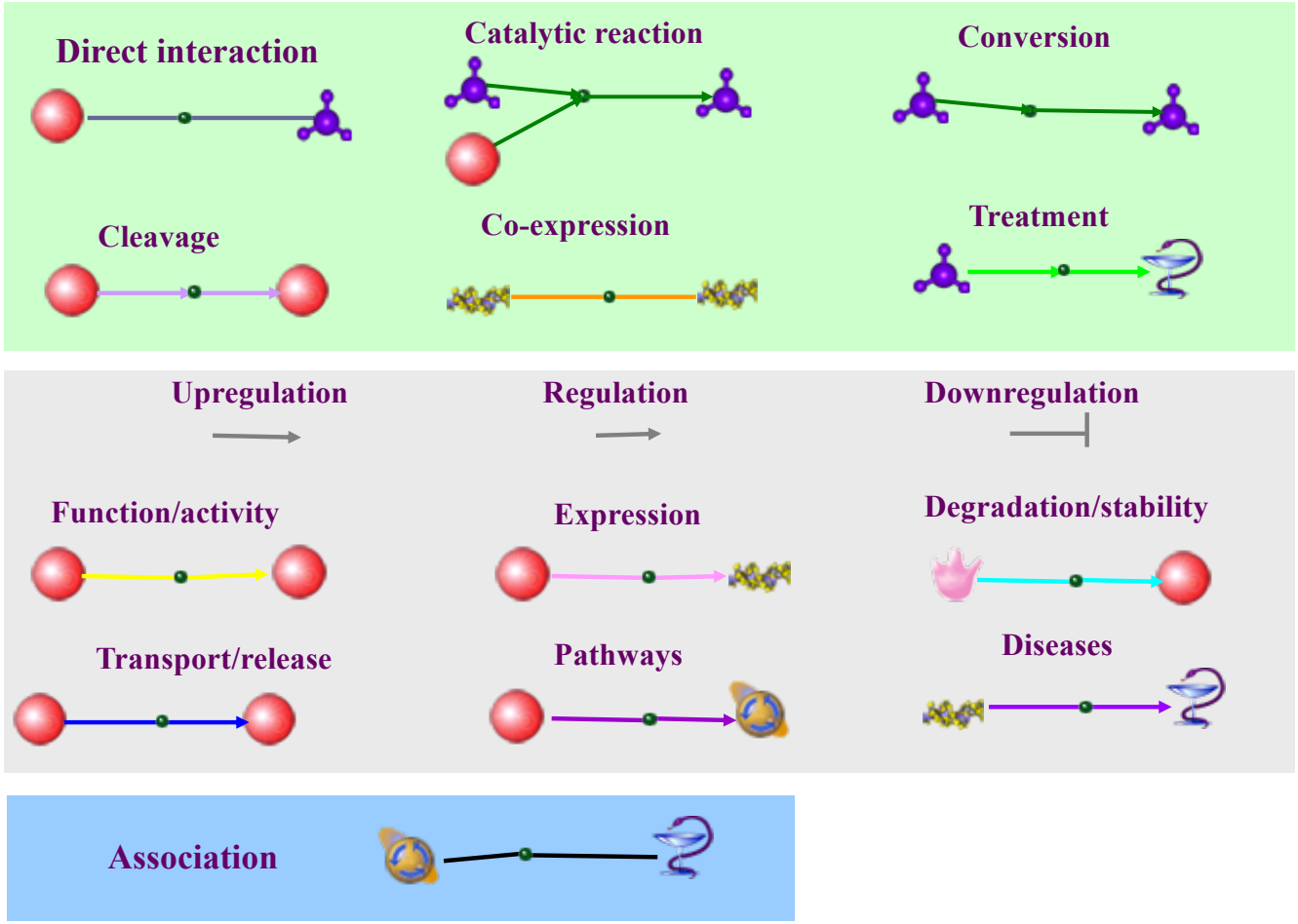
Integration of extracted knowledge into the semantic network

The knowledge base contains more than 30 million relationships between objects, confirmed by facts



Object dictionaries and interaction types used in ANDSystem

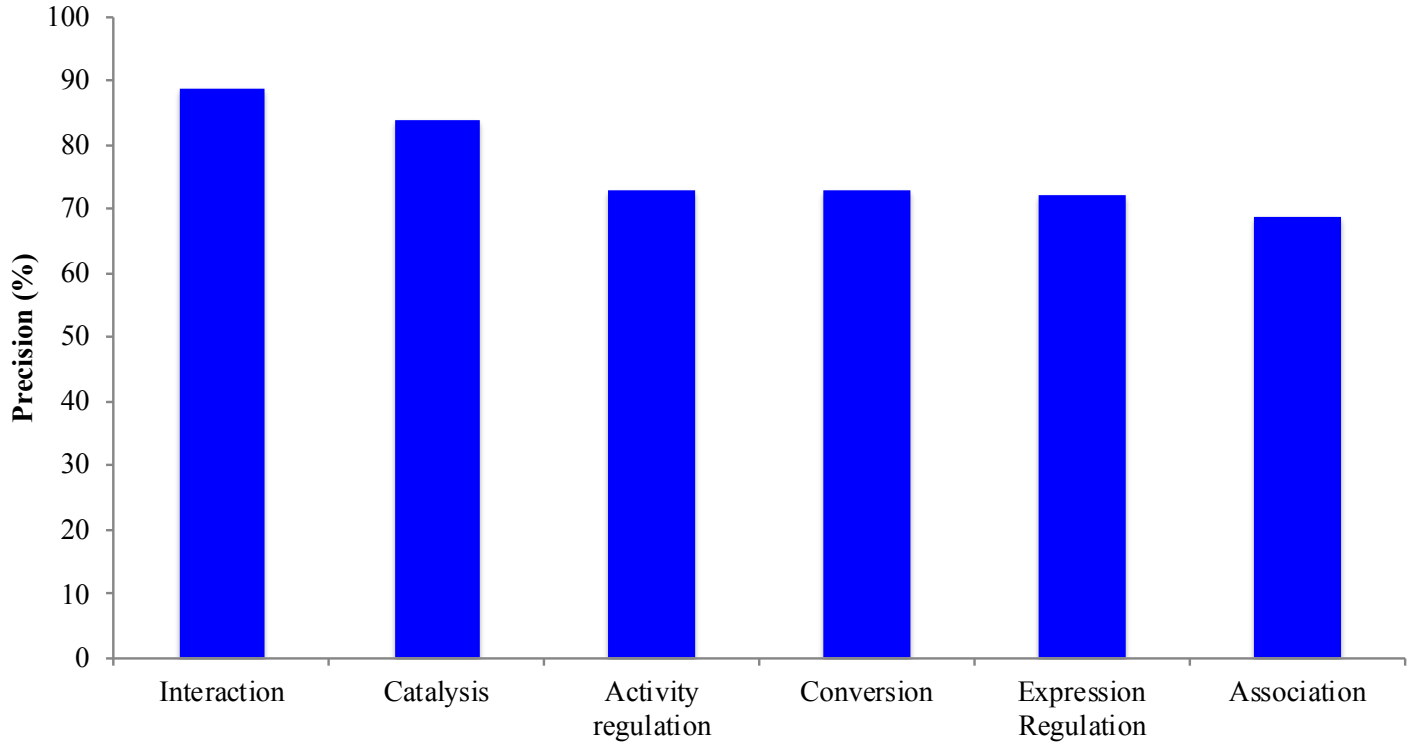
Physiology	
Organisms	21 982
Cells/tissues	644 304
Phenotype objects	
Diseases	15 478
Phenotypic traits	23 224
Drug side effects	5 226
Molecular genetic objects	
Genes	7 946 479
Proteins	550 657
Metabolites	42 594
MicroRNAs	28 512
Drugs	2 430
Biological processes/Pathways	122 297
Cell components	3 680



ANDSystem knowledge base statistics

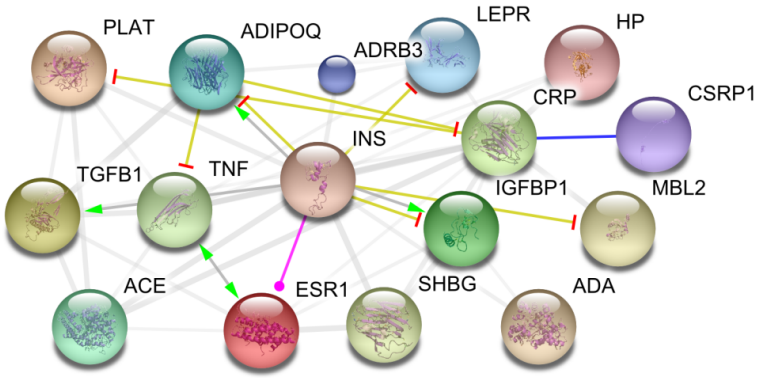
association	**7 767 711**	***activity regulation***	**80 576**
involvement	**4 217 997**	***activity downregulation***	**295 160**
interaction	**625 594**	***activity upregulation***	**227 483**
expression regulation	**559 599**	***degradation regulation***	**3 222**
expression upregulation	**157 533**	***degradation downregulation***	**55 215**
expression downregulation	**129 273**	***degradation upregulation***	**34 004**
pathway regulation	**529 997**	***miRNA regulation***	**23 576**
pathway upregulation	**375 514**	***coexpression***	**6 617**
pathway downregulation	**374 856**	***cleavage***	**7 141**
transport regulation	**382 373**	***catalyze modification***	**3 089**
treatment	**188 264**	***conversion***	**27 078**
catalyze	**518 114**		

Evaluation of the accuracy of the extraction of facts for different types of interactions using semantic and linguistic templates



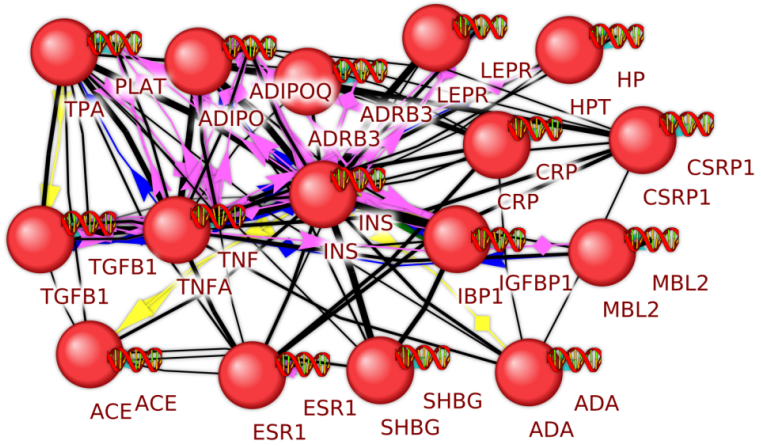
	PRESICION	RECALL
FULL-LENGTH ARTICLES	90.34%	78.9%
PUBMED ABSTRACTS	88.81%	73.8%

Comparison of different text-mining systems



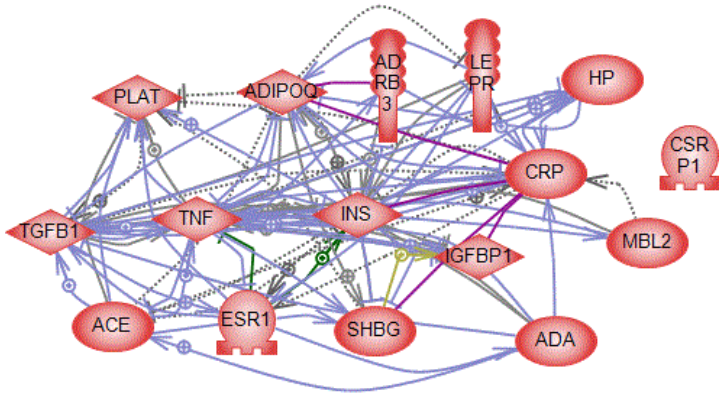
STRING

- Activation
- Fuctional link
- Binding
- Post-translational modification
- Expression with inhibition



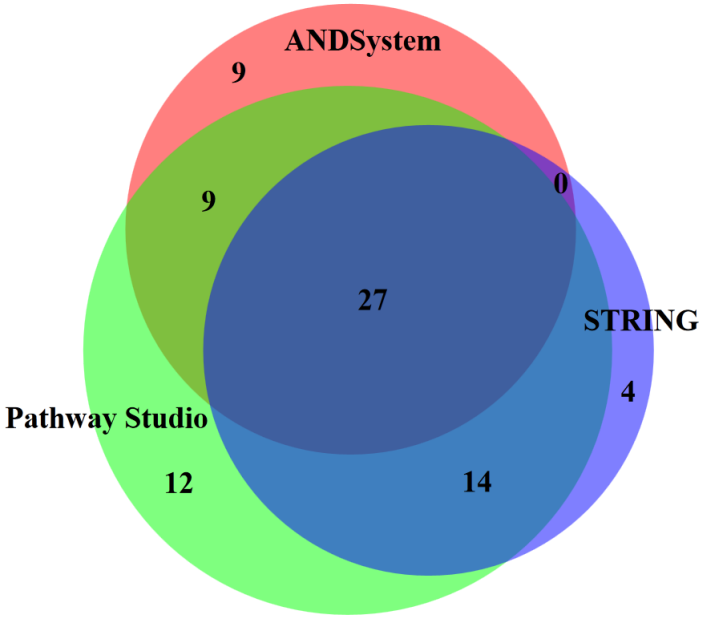
ANDSystem

- Activity downregulation
- Activity regulation
- Activity upregulation
- Catalysis
- Coexpression
- Cleavage
- Expression
- Degradation downregulation
- Degradation upregulation
- Expression downregulation
- Expression upregulation
- Expression regulation
- Transport regulation
- Association



Pathway Studio

- ProtModification
- MolTransport
- PromoterBinding
- Expression
- Regulation
- Binding



The intersection of interactions in the gene network reconstructed with ANDSystem, Pathway Studio and STRING.

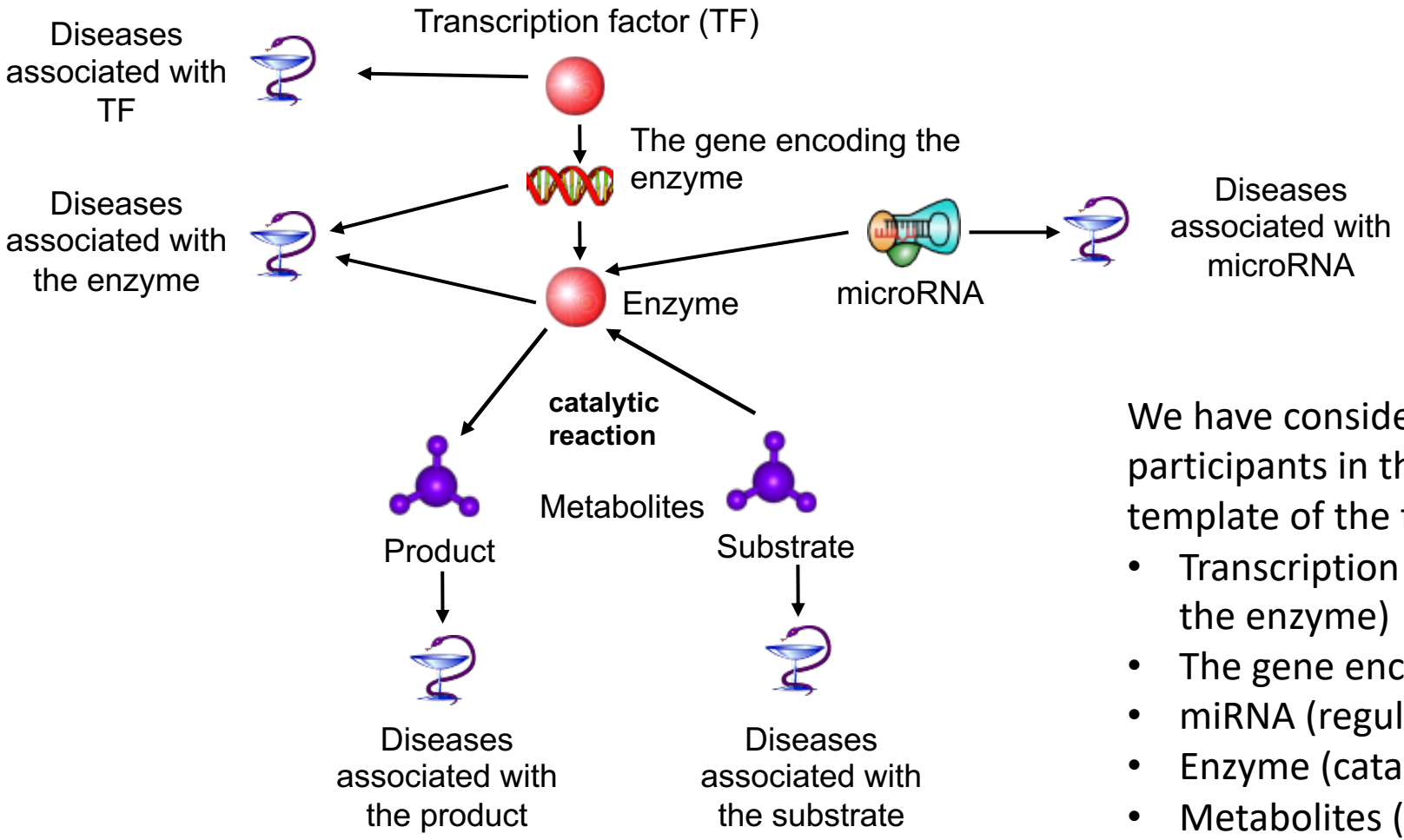
The use of ANDSystem for the interpretation of metabolomic data aimed at analysing the genetic regulation of metabolic processes

- Currently, metabolomic analysis is one of the widely used omics technologies in various areas of biomedical research
- However, the interpretation of metabolomic data in almost all studies is limited to the search for overrepresented metabolic processes. At the same time, such important processes as genetic regulation of metabolic processes are overlooked
- Knowledge of the genetic regulation of overrepresented processes may make it possible to clarify the relationship between the factors acting on the organism and the response observed in the form of a perturbated metabolic process.

To support genetic regulation analysis, ANDSystem provides a function that allows user to search for significant pathways of genetic regulation of metabolic processes using pathway templates.

Schematic representation of genetic regulatory pathways and associated diseases

Template of genetic regulation pathways

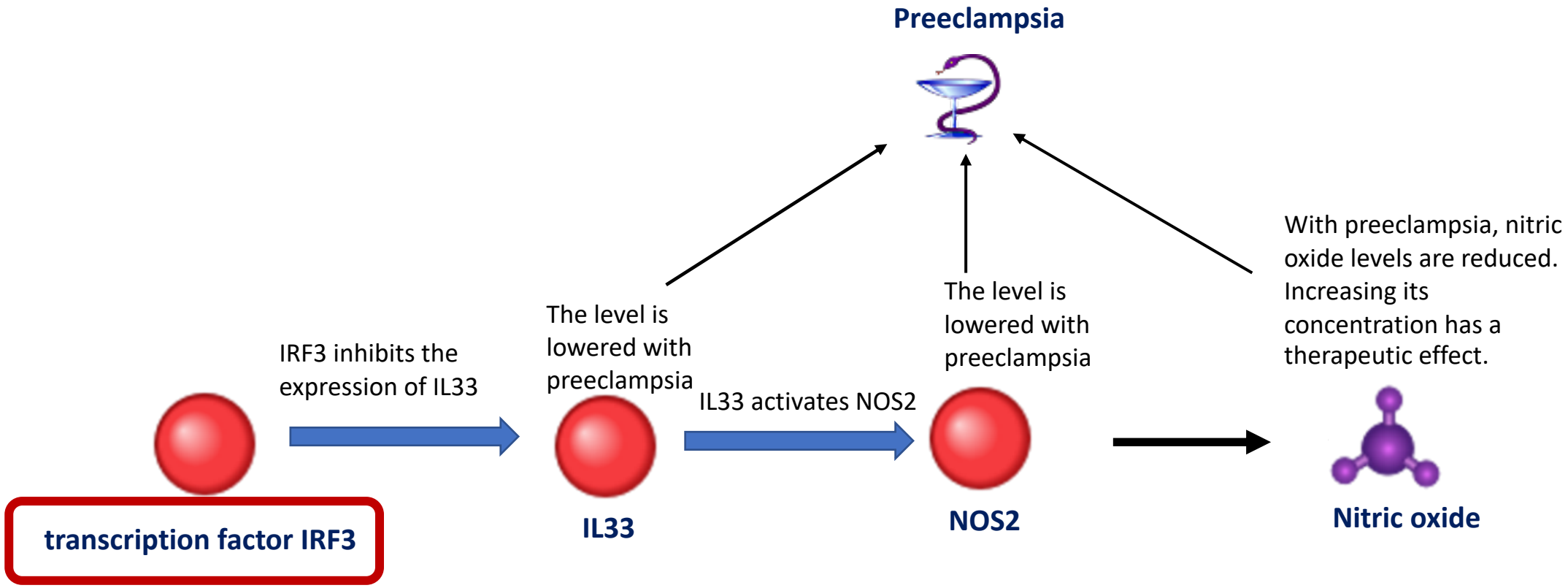


We have considered the distribution of diseases by participants in the regulatory pathways, given by the template of the following type:

- Transcription factor (regulates the gene encoding the enzyme)
- The gene encoding the enzyme
- miRNA (regulates enzyme production)
- Enzyme (catalyses a reaction)
- Metabolites (product and substrate)

We called such templates frame models of regulatory pathways.

An example of a pathway for preeclampsia



The pathway includes the transcription factor IRF3, the IL33 protein, the NOS2 enzyme, and the nitric oxide metabolite. The last three participants are associated with preeclampsia. The model suggests that the IRF3 transcription factor can also be associated with preeclampsia.

An example of using ANDSystem to interpret metabolomic data

- Zhang et al, 2013 published the results of a metabolomic analysis of the urine of hepatitis C virus (HCV) patients by high-throughput ultra-performance liquid chromatography−mass spectrometry (UPLC−MS). The authors identified 20 urinary differential metabolites and calculated KEGG overrepresented metabolic pathways.
- Using the ANDSystem, we reconstructed regulatory pathways leading from human proteins that HCV proteins interact with to genes involved in these KEGG pathways. Thus, the molecular genetic mechanism of the effect of viral proteins on the observed changes in the functioning of metabolic pathways can be described.

List of top 4 overrepresented KEGG pathways (Zhang et al, 2013):

1. Aminoacyl-tRNA biosynthesis
2. Nitrogen metabolism
3. Phenylalanine metabolism
4. Citrate cycle (TCA cycle)

Zhang AH, Sun H, Han Y, Yan GL, Yuan Y, Song GC, Yuan XX, Xie N, Wang XJ. Ultraperformance liquid chromatography–mass spectrometry based comprehensive metabolomics combined with pattern recognition and network analysis methods for characterization of metabolites and metabolic pathways from biological data sets. Analytical chemistry. 2013 Aug 6;85(15):7606-12.

ANDSystem templates for virus-host interaction pathways

Template name	Template description*
P1 – PPI 1	Vp – PPI –> Kp
P2 – PPI 2	Vp – PPI –> Hp – PPI –> Kp
P3 – Activity/Stability regulation 1	Vp – PPI –> Hp – Act/Stab/Pr/PPM/Tr –> Kp
P4 – Expression regulation 1	Vp – PPI –> Hp – Exp reg –> Kg – Exp –> Kp
P5 – Expression regulation 2	Vp – PPI –> Hp – Exp reg –> Hg – Exp –> Hp– Exp reg –> Kg – Exp –> Kp
P6 – Expression + Activity regulation	Vp – PPI –> Hp – Exp reg –> Hg – Exp –> Hp – Act/Stab/Pr/PPM/Tr –> Kp
P7 – Expression + PPI regulation	Vp – PPI –> Hp – Exp reg –> Hg – Exp –> Hp – PPI –> Kp

***Objects:**
Vp - HCV proteins
Kp - KEGG metabolic pathway proteins
Kg - KEGG metabolic pathway genes
Hp – any host proteins involved in the interactions
Hg – any host genes involved in the interactions
Interactions:
PPI – protein-protein interactions
Act/Stab/Pr/PPM/Tr – regulation of activity or stability, or proteolysis, or post translational modifications, or transport (release).
Exp reg – regulation of gene expression.
Exp – gene expression (protein production).

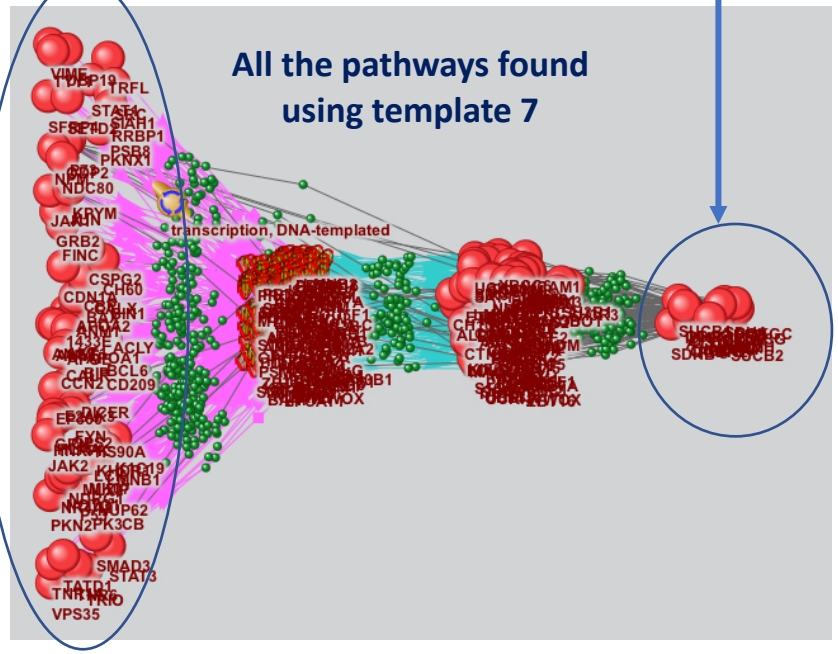
Statistical significance of associations between virus-host regulatory pathways and KEGG metabolic processes

Virus-host pathway template (Pi)	Aminoacyl-tRNA biosynthesis		Nitrogen metabolism		Phenylalanine metabolism		Citrate cycle	
	P value	FDR	P value	FDR	P value	FDR	P value	FDR
P_1	-	-	-	-	-	-	**0.004**	**0.015**
P_2	**10e-08**	**2.5e-06**	0.35	0.46	0.26	0.38	**0.0005**	**0.004**
P_3	0.44	0.55	0.56	0.63	**0.0013**	**0.008**	0.28	0.38
P_4	0.96	0.96	0.15	0.28	0.066	0.15	0.18	0.32
P_5	0.72	0.75	0.24	0.37	**0.0045**	**0.015**	0.049	0.12
P_6	0.66	0.71	0.2	0.33	**0.005**	**0.015**	0.033	0.09
P_7	**10e-07**	**1.25e-05**	0.085	0.17	0.53	0.63	**0.0027**	**0.013**

An example of regulatory pathways found using P7 template (Expression regulation + PPI regulation) targeted at Citrate cycle

Human proteins interacting with viral proteins

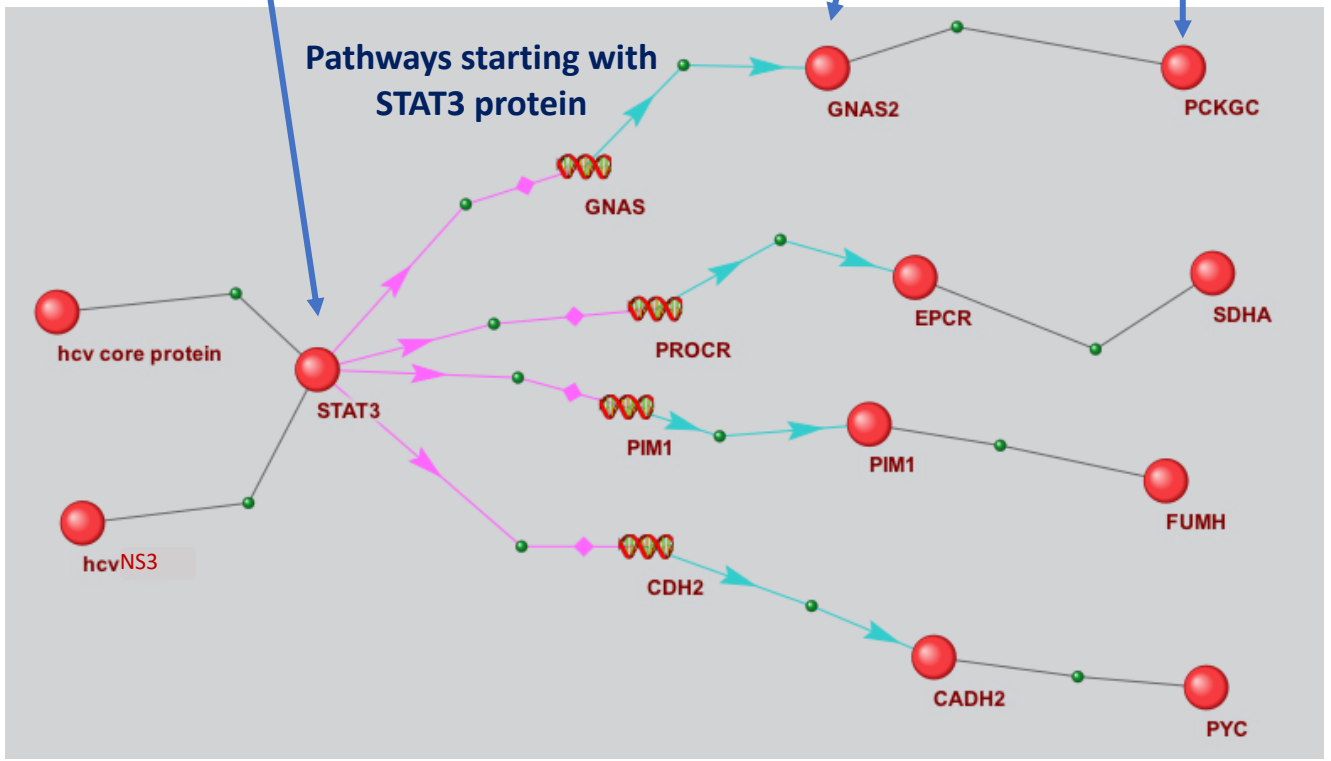
Citrate cycle proteins



Guanine nucleotide-binding protein

Signal transducer and activator of transcription 3
Well known an anti-apoptotic factor

Phosphoenolpyruvate carboxykinase (EC:4.1.1.32)



Conclusion

ANDSystem implements a complete cycle of knowledge engineering describing the relationship between genotype, phenotype and environment using artificial intelligence techniques

ANDSystem can be used to reconstruct the molecular mechanisms of diseases, interpret omix data, and other tasks in the medical and biological research

Selected publications on ANDSystem

1. Ivanisenko, T.V., Saik, O.V., Demenkov, P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. **BMC Bioinformatics** 21, 228 (**2020**). IF 3.24, Q1, https://doi.org/10.1186/s12859-020-03557-8.
2. Zolotareva O, Saik OV, Königs C, Bragina EY, Goncharova IA, Freidin MB, Dosenko VE, Ivanisenko VA, Hofestädt R. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. **Sci Rep**. **2019** Nov 8;9(1):16302. IF 3.998, Q1, doi: 10.1038/s41598-019-52762-w.
3. Ivanisenko VA, Demenkov PS, Ivanisenko TV, Mishchenko EL, Saik OV. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. **BMC Bioinformatics**. **2019** Feb 5;20(Suppl 1):34. IF 3.24, Q1, doi: 10.1186/s12859-018-2567-6.
4. Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. **BMC Systems Biology 2015**, 9(Suppl 2):S2. Q1, doi: 10.1186/1752-0509-9-S2-S2
5. Andrey S Glotov, Evgeny S Tiys, Elena S Vashukova, Vladimir S Pakin, Pavel S Demenkov, Olga V Saik, Timofey V Ivanisenko, Olga N Arzhanova, Elena V Mozgovaya, Marina S Zainulina, Nikolay A Kolchanov, Vladislav S Baranov, Vladimir A Ivanisenko Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). **BMC Systems Biology 2015**, 9(Suppl 2):S4. Q1, doi: 10.1186/1752-0509-9-S2-S

Thank you for the attention