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


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 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

Direct interaction
Catalytic reaction
Diseases
Pathways

Semantic and linguistic rules for knowledge extraction (created by hand)
Over 10,000 rules

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

Integration of extracted knowledge into the semantic network
The knowledge base contains more than 30 million relationships between objects, confirmed by facts

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

Stage 1
A interacts with B  \( A \rightarrow B \)
C catalyzes D into E  \( C + D \rightarrow E \)

Stage 2

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

Training neural networks and extracting new knowledge using them
## Object dictionaries and interaction types used in ANDSSystem

<table>
<thead>
<tr>
<th>Physiology</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisms</td>
<td>21,982</td>
</tr>
<tr>
<td>Cells/tissues</td>
<td>644,304</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phenotype objects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases</td>
<td>15,478</td>
</tr>
<tr>
<td>Phenotypic traits</td>
<td>23,224</td>
</tr>
<tr>
<td>Drug side effects</td>
<td>5,226</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecular genetic objects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes</td>
<td>7,946,479</td>
</tr>
<tr>
<td>Proteins</td>
<td>550,657</td>
</tr>
<tr>
<td>Metabolites</td>
<td>42,594</td>
</tr>
<tr>
<td>MicroRNAs</td>
<td>28,512</td>
</tr>
<tr>
<td>Drugs</td>
<td>2,430</td>
</tr>
<tr>
<td>Biological processes/Pathways</td>
<td>122,297</td>
</tr>
<tr>
<td>Cell components</td>
<td>3,680</td>
</tr>
</tbody>
</table>

### Direct interaction

- Cleavage
- Co-expression
- Treatment

### Catalytic reaction

- Upregulation
- Regulation
- Downregulation

- Function/activity
- Expression
- Degradation/stability
- Transport/release
- Pathways
- Diseases

### Conversion

- Association
## ANDSystem knowledge base statistics

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>association</td>
<td>7,767,711</td>
<td>activity regulation</td>
<td>80,576</td>
</tr>
<tr>
<td>involvement</td>
<td>4,217,997</td>
<td>activity downregulation</td>
<td>295,160</td>
</tr>
<tr>
<td>interaction</td>
<td>625,594</td>
<td>activity upregulation</td>
<td>227,483</td>
</tr>
<tr>
<td>expression regulation</td>
<td>559,599</td>
<td>degradation regulation</td>
<td>3,222</td>
</tr>
<tr>
<td>expression upregulation</td>
<td>157,533</td>
<td>degradation downregulation</td>
<td>55,215</td>
</tr>
<tr>
<td>expression downregulation</td>
<td>129,273</td>
<td>degradation upregulation</td>
<td>34,004</td>
</tr>
<tr>
<td>pathway regulation</td>
<td>529,997</td>
<td>miRNA regulation</td>
<td>23,576</td>
</tr>
<tr>
<td>pathway upregulation</td>
<td>375,514</td>
<td>coexpression</td>
<td>6,617</td>
</tr>
<tr>
<td>pathway downregulation</td>
<td>374,856</td>
<td>cleavage</td>
<td>7,141</td>
</tr>
<tr>
<td>transport regulation</td>
<td>382,373</td>
<td>catalyze modification</td>
<td>3,089</td>
</tr>
<tr>
<td>treatment</td>
<td>188,264</td>
<td>conversion</td>
<td>27,078</td>
</tr>
<tr>
<td>catalyze</td>
<td>518,114</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation of the accuracy of the extraction of facts for different types of interactions using semantic and linguistic templates.
Comparison of different text-mining systems

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.
Preeclampsia transcription factor IRF3

IRF3 inhibits the expression of IL33. The level is lowered with preeclampsia.

IL33 activates NOS2. The level is lowered with preeclampsia.

With preeclampsia, nitric oxide levels are reduced. Increasing its concentration has a therapeutic effect.

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)

### ANDSysm templates for virus-host interaction pathways

<table>
<thead>
<tr>
<th>Template name</th>
<th>Template description*</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 – PPI 1</td>
<td>Vp – PPI → Kp</td>
</tr>
<tr>
<td>P2 – PPI 2</td>
<td>Vp – PPI → Hp – PPI → Kp</td>
</tr>
<tr>
<td>P3 – Activity/Stability regulation 1</td>
<td>Vp – PPI → Hp – Act/Stab/Pr/PPM/Tr → Kp</td>
</tr>
</tbody>
</table>

*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 – gene expression (protein production).
### Statistical significance of associations between virus-host regulatory pathways and KEGG metabolic processes

<table>
<thead>
<tr>
<th>Virus-host pathway template (Pi)</th>
<th>Aminoacyl-tRNA biosynthesis</th>
<th>Nitrogen metabolism</th>
<th>Phenylalanine metabolism</th>
<th>Citrate cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P value</td>
<td>FDR</td>
<td>P value</td>
<td>FDR</td>
</tr>
<tr>
<td>(P_1)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(P_2)</td>
<td>1.00e-08</td>
<td>2.50e-06</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>(P_3)</td>
<td>0.44</td>
<td>0.55</td>
<td>0.56</td>
<td>0.63</td>
</tr>
<tr>
<td>(P_4)</td>
<td>0.96</td>
<td>0.96</td>
<td>0.15</td>
<td>0.28</td>
</tr>
<tr>
<td>(P_5)</td>
<td>0.72</td>
<td>0.75</td>
<td>0.24</td>
<td>0.37</td>
</tr>
<tr>
<td>(P_6)</td>
<td>0.66</td>
<td>0.71</td>
<td>0.2</td>
<td>0.33</td>
</tr>
<tr>
<td>(P_7)</td>
<td>1.00e-07</td>
<td>1.25e-05</td>
<td>0.085</td>
<td>0.17</td>
</tr>
</tbody>
</table>
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

All the pathways found using template 7

Guanine nucleotide-binding protein

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

Phosphoenolpyruvate carboxykinase (EC:4.1.1.32)

Pathways starting with STAT3 protein

hev core protein

hevNS3

GNAS

EPCR

PIM1

FUMH

CADH2

PYC

GNAS2

PCKGC
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


Thank you for the attention