IPA: Overview and Networks Generation Algorithm

Dr. Michael Shmoish
<mshmoish@cs.technion.ac.il>

Bioinformatics Knowledge Unit
The Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering
Technion - Israel Institute of Technology

22 January 2014
Outline

• What is IPA
• What are IKB and GMN
• 3 basic principles behind the design of the network generation algorithm
• 5 core design goals
• 6-steps network generation algorithm
• Conclusion
The Challenge

Data Analysis Across Multiple Dimensions of Biology

Systematic Generation of Novel Biological and Therapeutic Insights

Find genes implicated in disease

Identify related cellular processes, pathways

Generate hypothesis of molecular mechanism

Informed in vivo, in vitro assays

- Expression Arrays
- Proteomics
- Traditional Assays

Disease Processes

Cellular Processes

Molecules

Experimental Platforms

- Cancer
- Apoptosis
- Angiogenesis
- Fas
- Vegf
What is IPA

• Ingenuity Pathway Analysis (IPA) is a system that transforms a list of genes into a set of relevant networks based on extensive records maintained in the Ingenuity Knowledge Base.

• Helps to understand complex “omics” data at multiple levels by integrating data from a variety of experimental platforms and providing insight into the molecular and chemical interactions, cellular phenotypes, and disease processes of your system.
Ingenuity® Knowledge Base

Content Acquisition

Ingenuity Ontology

- Patient Phenotypes (e.g., Docetaxel)
- Disease Mechanisms (e.g., Prostate Cancer)
- Cellular Mechanisms (e.g., Apoptosis, Angiogenesis)
- Molecular Mechanisms (e.g., Fas, Vegf)
- Sequence Mechanisms (e.g., DNA, RNA)
IPA citation trend over time

IPA has been broadly adopted by the life science research community and is cited in thousands of peer-reviewed journal articles.
IPA (summary)
IPA (networks)

The analysis is composed of 25 networks. To view a network, select the appropriate network(s) and click View Networks. To merge selected networks, click Merge Networks.

<table>
<thead>
<tr>
<th>Molecules in Network</th>
<th>Score</th>
<th>Focus Molecules</th>
<th>Top Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>29</td>
<td>Cellular Development, Cellular Growth and Proliferation, Endocrine System Development and Function</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>29</td>
<td>Genetic Disorder, Cellular Assembly and Organization, Cellular Function and Maintenance</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>27</td>
<td>Genetic Disorder, Cellular Assembly and Organization, Developmental Disorder</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>27</td>
<td>Developmental Disorder, Genetic Disorder, Neurological Disease</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>26</td>
<td>Cell Death, Free Radical Scavenging, Carbohydrate Metabolism</td>
</tr>
<tr>
<td>6</td>
<td>34</td>
<td>26</td>
<td>Molecular Transport, RNA Trafficking, Genetic Disorder</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>24</td>
<td>Antimicrobial Response, Cell-to-Cell Signaling, Inflammatory Response</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>23</td>
<td>Genetic Disorder, Cell Death, Cardiovascular System Development and Function</td>
</tr>
<tr>
<td>9</td>
<td>23</td>
<td>20</td>
<td>Cell-To-Cell Signaling and Interaction, Neurological Response</td>
</tr>
</tbody>
</table>
IPA (a network)
IPA (overlay)
IPA (overlapping networks)
IPA (pathways)
Ingenuity® Knowledge Base

Content Acquisition

Ingenuity Ontology

PATIENT PHENOTYPES (e.g., Docetaxel)
DISEASE MECHANISMS (e.g., Prostate Cancer)
CELLULAR MECHANISMS (e.g., Apoptosis, Angiogenesis)
MOLECULAR MECHANISMS (e.g., Fas, Vegf)
SEQUENCE MECHANISMS (e.g., DNA, RNA)
The Knowledge Base is distinctive because of the breadth of biology and chemical knowledge, accuracy and structure of the content for relationship and computation using the Ingenuity Ontology.

• **Not Just Simple “A to B” Relationships**

  Context: allow to ask questions such as “Does transcription factor X increase expression of a gene Y in a specific cell type?”

• **Supports Computation - structured**

  The Ingenuity Ontology makes information computationally accessible to infer novel insights from the data or find likely paths between molecular concepts (gene to disease, drug to gene, etc.)

• **Provides Synonym Resolution**

• **Content Acquisition**
The information is organized into the following categories:

- **Expert Findings** - from ~300 top journals, curated from the full text, including tables and figures; weekly updates

- **ExpertAssist Findings** - are manually reviewed, automatically extracted Findings from the abstracts

- **Expert Knowledge** - Ingenuity-modeled knowledge such as signaling and metabolic pathways, drug/target/disease relationships, toxicity lists, etc.

- **Supported Third Party Information**
IKB-supported 3rd-Party Information

- EntrezGene, RefSeq, OMIM (NCBI)
- Targets of FDA approved and clinical trial drugs
- Clinical biomarkers
- Gene Ontology annotations
- Normal gene expression body atlas for over 30 tissues
- COSMIC
- TargetScan
- BIND, DIP, MINT, MIPS, BIOGRID, INTACT, COGNIA protein-protein interactions
- More...
• Altogether there are more than 5M findings at the moment (December 2013 release). Each one is categorized as a study of either human, mouse or rat genes (based on Entrez Gene annotation), and supported by a literature citation and a link to the PUBMED abstract.

• All information (both manually curated relationship from the literature as well as a relationship described in a third party database) in the Ingenuity Knowledge Base is **manually** reviewed by experts to ensure that the content is accurate and detailed.
The above knowledge base has been abstracted into a large network, called the Global Molecular Network, composed of thousands of genes and gene products that interact with each other.

Term “gene” is used to describe both genes and gene products. A relationship between two genes in this network is called an “edge”.

Only a single edge is formed between two genes regardless of whether they have one or more associations/findings that connect them (e.g., if CCND1 binds to CDK4 and also activates CDK4, in the GMN there exists just one edge between CCND1 and CDK4).

All edges are treated equally for network generation.
Three basic principles behind the network generation algorithm

Given a list of genes from the biological experiment:

1. IPA is intended to return networks that help biologists to understand how their genes of interest are biologically related, i.e., how the genes in a given dataset appear to work together at the molecular level.

2. Highly interconnected networks are believed to represent significant molecular biology, so IPA optimizes for “triangular” relationships between genes, in essence favoring denser networks over more sparsely-connected ones.

3. IPA is designed to give a maximally useful amount of information on a single screen while still producing interpretable networks.
Algorithm overview: user’s perspective

Experimental Data → List of Focus Genes, optionally with expression values or other attributes

GFA / LFA → Gene-By-Gene

Network generation algorithm → Networks

Global Molecular

Ingenuity Knowledge Base

Drill down
IPA network generation algorithm

The selection of high scoring networks, typically based on gene expression data, has been well studied in the literature, and several authors have solved one or more of the above goals, often using objective functions.

- **Ideker** looked for subnetworks in protein-protein and protein-DNA networks that maximized an overall subnetwork function based on expression data [2].
- **Rajagopalan** integrated three distinct networks and improved on Ideker’s optimization function [3].
- **Pradines** combined metabolic and regulatory pathways from two databases to compute the inverse problem: what genes have significantly upregulated neighborhoods, where neighborhoods are defined using the networks [4].
- **Bader** addresses gene network interconnectivity using a pre-existing interaction network [5].

Unlike these algorithms, IPA [1] constructs networks that optimize for both interconnectivity and number of Focus Genes under the constraint of a maximal network size.

Focus Genes (or Focus Molecules) are those from the uploaded list that pass filters and are eligible for generating networks, i.e., potentially could be linked to some other genes.
Core design goals

The IPA network generation algorithm has five core design goals:

1) Analyze user-specified lists of genes, some of which may have expression values or ranks associated with them;

2) Highlight the molecular biology implied in a dataset by identifying how user-specified genes interact with each other or with neighboring genes;

3) Prefer highly-interconnected molecular networks over sparsely connected ones because the former tend to reflect significant biology;

4) Generate non-redundant networks that are manageable in size and can be easily visualized and understood;

5) Analyze datasets quickly.
IPA network generation

• Ingenuity’s approach is based on a multi-stage, heuristic algorithm that executes six steps to try to satisfy all of the above goals.

• The IPA network generation algorithm iteratively constructs networks that greedily optimize for both interconnectivity and number of Focus Genes under the constraint of a maximal network size.
IPA network generation: 6 steps

1. Sort Focus Genes with respect to their interconnectivity. Highly interconnected Focus Genes which optimize triangle connectivity will be processed first.

2. Construct networks from Focus Genes. Optimizes number of Focus Genes and number of specific edges (measured using specific connectivity) within each network.


5. Form all edges between genes.

6. Compute the p-score and rank the networks.
Definitions

a) Let $F$ be the set of user genes $U$ that the user considers to be interesting (e.g. all genes that are above a certain expression value or below a p-value cut off).

b) Let the graph $G = (V, E)$ represent the Global Molecular Network, where $V$ is the set of nodes representing genes and $E$ is the set of edges representing interactions. $G$ contains no self edges.

c) Let $F_G = F \cap V$ be the set of focus genes, i.e. the set of genes in $F$ that are also contained in the Global Molecular Network.

d) Let $N_G(v)$ denote the neighborhood of a node $v$ in $G$.

$N_G(v) \subseteq V$ is the set of all nodes adjacent to $v \in V$. Similarly let $N_G(S)$ denote the neighborhood of a set of nodes $S \subseteq V$ in $G$. 

e) Define the **triangle connectivity** \( tr_G(v) \) of a node \( v \in V \) as the number of triangles in \( G \) that contain \( v \).
f) Define the **specific connectivity** of a node \( v \in V \) with respect to a set of nodes \( S \subset V \) as
\[
s_{c_G}(v, S) = \frac{|[N_G(v) \cup \{v\}] \cap S|}{|[N_G(v) \cup \{v\}] \cap S|}.
g) Define the **edge connectivity** of a node \( v \in V \) with respect to a set of nodes \( S \subset V \) as
\[
e_{c_G}(v, S) = |N_G(v) \cap S|.
h) Define the **user-score** \( u_{s_G}(v) \) of a node \( v \in V \) as the average of the user-provided scores of 
\([N_G(v) \cup \{v\}] \cap U\). Scores such as expression values, p-values, etc., are part of the user’s input dataset. Scores are changed so the most significant scores have larger values.
i) Define two sets of node sets in \( G \), the display list \( D \) and the candidate list \( C \).
j) Let \( N_{\text{max}} \) be the maximum number of nodes in displayed networks. In the actual implementation of the algorithm we set \( N_{\text{max}} = 35 \).
Triangle connectivity measures the number of triangles, or pairs of connected genes, to which a gene is connected. It is a generalization of ...???
Triangle connectivity measures the number of triangles, or pairs of connected genes, to which a gene is connected.

It is a generalization of the concept of node degree, which measures the number of single genes to which a gene is connected.
Triangle connectivity reflects interconnectivity

- 0 triangles
- 1 triangle
- 2 triangles
- 3 triangles
Step 1

- Genes selected as Focus Genes (Network Eligible molecules) are sorted in decreasing order of triangle connectivity.
Step 1: Sort Focus Genes by triangle connectivity

1.1. Initialize $D$ and $C$ to be empty sets, $D = \emptyset$, $C = \emptyset$.
1.2. Initialize $F_G' = F_G$ to be the set of unused focus genes.
1.3. Pick the node $v \in F_G'$ with the largest triangle connectivity $tr_G(v)$ in $F_G'$. Update $F_G' = F_G' \setminus \{v\}$ and initialize a set of nodes $S = \{v\}$. 
Step 2

The top ranked (most connected) gene is removed from the sorted Focus Gene list prepared in step 1. Focus Genes are added to this seed gene until we reach the maximum network size (set at 35 by default).

As the seed grows in size to 2, 3, 4, ... genes, several Focus Genes are found that could be added. How to choose which one to add?
Step 2 (cont.)

Under working hypothesis that biological function involves locally dense interactions the gene is picked up that is most connected to the growing network.

As a network grows, one compares how much a new gene’s neighborhood overlaps the current network and pick the gene that overlaps the most. The metric is called “specific connectivity”
Step 2: Specific connectivity

\[
\text{Specific connectivity} = \frac{\text{Number of genes in intersection of the neighborhood and network}}{\text{Number of genes in union of neighborhood and network}}
\]
Step 2: Example
Step 2: Example – Q.

Which of 3 candidate gene would be connected first?
### Step 2: example – A.

<table>
<thead>
<tr>
<th>Color</th>
<th>Genes in neighborhood</th>
<th>Genes in network</th>
<th>Specific Connectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>8</td>
<td>3</td>
<td>0.18</td>
</tr>
<tr>
<td>Green</td>
<td>7</td>
<td>2</td>
<td>0.12</td>
</tr>
<tr>
<td>Blue</td>
<td>7</td>
<td>3</td>
<td>0.19</td>
</tr>
</tbody>
</table>

The blue node has the highest specific connectivity, so it would be selected as the next node to add to the growing network.
Step 2: Greedily construct networks of Focus Genes up to a maximum pre-defined size.

2.1. Make a set \( N = N_G(S) \cap F_G \) of all neighboring focus genes of \( S \). Pick the node \( w \in N \) with the largest specific connectivity \( sc_G(w, S) \) with respect to \( S \). Update \( S = S \cup \{w\} \).

2.2. Repeat step 2.1 until \( |S| = N_{\text{max}} \) or \( N = \emptyset \).

2.3. If \( |S| = N_{\text{max}} \) update display list \( D = D \cup \{S\} \), else update candidate list \( C = C \cup \{S\} \).

2.4. Repeat steps 1.3 – 2.3 until \( F_G = \emptyset \).
Step 3

• Combine small connected Focus Gene networks together to make larger networks through a single additional gene from GMN.

• The step starts with the smallest network, identifies “linker” genes that “link” it to other small networks and chooses such a gene with the most edges to both networks. If possible, this process continues until the merged networks become large enough to display. If not, the networks are enlarged in step 4 “Linker” might be occasionally Focus Genes but probability is low.
Step 3 (algorithm)

Step 3: Merge small networks using linker genes

3.1. Make a set $C'$ of all small networks $S$ in $C$ for which $|S| < \frac{N_{\text{max}}}{2}$. Update $C = C \setminus C'$.

3.2. Select the smallest network $S_s \in C'$.

3.3. Make a set $L$ of all linker nodes of $S_s$. $L$ is the set of all nodes $l \in V$ that are in the neighborhood of $S_s$ and at least one other small network in $C'$. If $L = \emptyset$ update $C' = C' \setminus \{S_s\}$, $C = C \cup \{S_s\}$ and go to 3.7.

3.4. For each linker node $l \in L$ pick the smallest set $S_i \in C'$ that links to $l$.

3.5. Select the $l \in L$ that maximizes the combined edge connectivity with respect to both $S_s$ and $S_i$: $ec_{C'}(l, S_s) + ec_{C'}(l, S_i)$. If more than one $l \in L$ have the same combined edge connectivity, choose the one with the largest triangle connectivity $tr_{C'}(l)$. Form the merged set $M = S_s \cup S_i \cup \{l\}$. Update $C' = C' \setminus \{S_s, S_i\}$.

3.6. If $|M| < \frac{N_{\text{max}}}{2}$ update $C' = C' \cup \{M\}$ else update $C = C \cup \{M\}$.

3.7. Repeat steps 3.2 - 3.6 until $C' = \emptyset$. 
Step 4

For those networks that still have less than 35 genes, other genes are added to the periphery of the network to provide additional biological context to those focus genes. One starts by adding genes that have two or more edges connecting them to any gene in the network. If there are no genes with two or more edges then new genes are added with only a single edge to some gene in the network. 

As in step 3, the added genes can be Focus Genes but ordinarily the probability is low.
Step 4: Grow remaining small networks using available neighborhood genes

4.1. Update $D = D \cup C$.
4.2. For each $S \in D$, if $|S| < N_{\text{max}}$ make the set $N = N_G(S)$ of all neighboring nodes of $S$.
4.3. If $|S| > 1$ remove from $N$ all nodes with only a single edge to $S$.
4.4. If $|S| + |N| \leq N_{\text{max}}$ update $S = S \cup N$, else make the set $N' \subseteq N$ containing the $N_{\text{max}} - |S|$ nodes $v$ in $N$ with the largest value of the product $us_G(v) \cdot sc_G(v, S)$ with respect to $S$. Update $S = S \cup N'$.
Step 5

• Network genes and all edges between them are pulled together into a single network. These networks are intended to contain approximately 35 genes. However, occasionally despite attempts to merge and grow networks with non-Focus Genes, small networks are the only ones that appear in the Global Molecular Network. In this case the small networks are reported.

• Assumption: each Focus Gene should be at least in one network; even if it is only connected to one other gene in the GMN, this network is of value and should be displayed.
Step 5: Display subgraph defined by network nodes

5.1. For each node set $S \in D$ display the corresponding network, i.e. the subgraph $G_S$ of $G$ induced by $S$. 
Step 6

The final step is the calculation of p-scores used to rank networks on the IPA “Results” page.

The p-scores are derived from hypergeometric p-values (Fisher’s exact test):

\[ p\text{-score} = -\log_{10}(p\text{-value}) \]
Step 6: P-score calculation

• Say there are $n$ genes in the network and $f$ of them are Focus Genes. The p-value is the probability of finding $f$ or more Focus Genes in a set of $n$ genes randomly selected from $N$ genes of the Global Molecular Network. It is calculated using Fisher’s exact test (based on the hypergeometric distribution).
Step 6 (algorithm)

**Step 6: Rank the networks**

6.1. For a subgraph $G_s$ with node set $S \in D$ and using the $2 \times 2$ contingency table shown below, compute the p-value as the probability of finding $a$ or more Focus Genes in $|S|$ genes randomly selected from the Global Molecular Network.

| Focus Gene | In network | $-a$ in network | $|F_G|$ |
|------------|------------|-----------------|-------|
| $a$        |            | $|F_G| - a$      |       |
| $|S| - a$   | $-a$       | $|V| - |S| - |F_G|$ | $|V| - |F_G|$ |

The p-value is the right-tailed sum of the hypergeometric distribution, also known as Fisher’s exact test:

$$p-value = \sum_{i = a}^{\min(|F_G|, |S|)} \binom{|F_G|}{i} \binom{|V| - |F_G|}{i} \binom{|V|}{|S| - i}$$

The p-score = $-\log_{10}(p$-value)
Conclusions

We have learned:

• IPA is an important software tool for the biomedical research

We enumerate

• basic principles behind the design of the network generation algorithm
• core design goals
• steps of the network generation algorithm
References


