INGENUITY®
SYSTEMS

Ingenuity Downstream Effects Analysis in IPA®

TABLE OF CONTENTS

INTRODUCTION

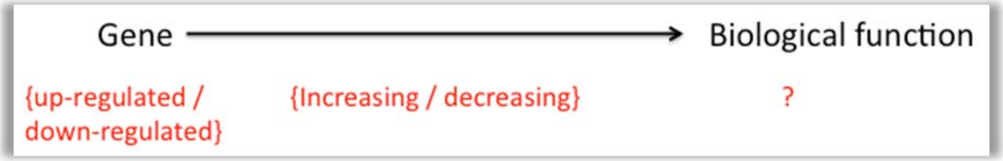
The goal of the IPA Downstream Effects analytic is to identify biological functions that are expected to be increased or decreased given the observed gene expression changes in a user's dataset.

The downstream effects analysis is based on prior knowledge of expected causal effects between genes and biological functions stored in the Ingenuity® Knowledge Base. The analysis examines genes in the user's dataset that are known to affect each biological function and compares their direction of change (i.e. expression in the experimental sample(s) relative to control) to what is expected from the literature. If the observed direction of change is mostly consistent with a particular activation state of a biological function ("increased" or "decreased"), then a prediction is made about that activation state.

For each biological function a statistical quantity is computed, called the activation z-score. The activation z-score is used to infer likely activation states of biological functions based on comparison with a model that assigns random regulation directions. Under ideal circumstances (the "un-biased" case described below) the activation z-score can also be used to predict implicated biological functions independently from their associated p-value, based on significant pattern match of up/down regulation.

ACTIVATION Z-SCORE

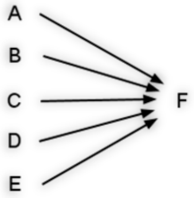
The primary purpose of the activation z-score is to infer the activation states ("increased" or "decreased") of implicated biological functions. The basis for this inference are edges (relationships) in the molecular network that represent experimentally observed causal relationships between genes and those functions. These edges are associated with a literature-derived direction of effect which can be either "increasing" or "decreasing":



Given the observed differential regulation of a gene ("up" or "down") in the dataset, the activation state of a biological function is determined by the direction of effect associated with the relationship from the gene to the function:

Effect associated with edge	**Observed gene regulation**	**Predicted activation state of biological function**
activating	up	increased
activating	down	decreased
inhibiting	up	decreased
inhibiting	down	increased

In general a biological function (F) connects to several upstream genes in the data set:



For each of those regulated genes we can make a prediction about the activation state of the biological function F. In order to make the claim that an increase (or decrease) of F is likely to occur for the observed up/down-regulation pattern we could require the predicted activation state of F to be consistent for all connected genes A, B, C, D, E. However, in practice we expect some inconsistent predictions, for a variety of reasons. For instance, it cannot be guaranteed that all of the relationships shown in the picture above are relevant (actually occur) in the given experimental context. Therefore, we take a statistical approach by defining a quantity (z-score) that determines whether a biological function has significantly more "increased" predictions than "decreased" predictions (z>0) or vice versa (z<0). Here, significance means that we reject the hypothesis that predictions are random with equal probability. The distribution underlying this null hypothesis is defined by a random variable:

$$x_i \in \{-1,1\}$$

where +1 corresponds to an "increased" state and -1 to an "decreased" state, and both values are chosen with probability 1/2. The index i runs from 1 to *N* with *N* being the number of genes affecting the function. Let:

$$x = \sum_i x_i = N_+ - N_-$$

where $N_{+/-}$ are the number of "increased"/"decreased" predictions and $N_+ + N_- = N$. The variance of x_i is $\sigma^2 = 1$, so the variance of x is given by:

$$\sigma_x^2 = N\sigma^2 = N$$

and the z-score statistic (with mean equal to zero and variance equal to 1) is defined by:

$$z = \frac{x}{\sigma_x} = \frac{\sum_i x_i}{\sqrt{N}} = \frac{N_+ - N_-}{\sqrt{N}}$$

The exact distribution of z is related to a binomial distribution, however, if *N* is large enough it can be approximated by a Gaussian. Since z is approximately normally distributed with zero mean and standard deviation one under the null hypothesis, we can use it to assess statistical significance of the observed number of "increased" and "decreased" predictions: If the absolute value of the z-score calculated from those numbers is large (i.e. falls into the "tail" of the Gaussian distribution) it would be unlikely to obtain that value of z by chance. Moreover, the sign of the calculated z-score will reflect the overall predicted activation state of the biological function (<0: decreased, >0: increased). In practice, z-scores greater than 2 or smaller than -2 can be considered significant.

So far it has been assumed that a direction of regulation (either activating or inhibiting) can be unambiguously assigned to an edge. This is not always the case because generally a single edge is associated with a number of findings that represent experimental observations reported in the literature. Since these observations have not necessarily been obtained under the same experimental conditions but could represent different contexts (e.g. organism, tissue, cell line, or more complex situations) it is not surprising that the direction of regulation can be different for different findings underlying the same edge. Ideally we would only consider findings that are applicable to the experimental context at hand (i.e. the biological context in which the data was observed), but since this context is unknown by the algorithm, we take the approach to put less weight on edges with fewer findings or ambiguous direction of regulation (a larger weight would in principle make it more likely that the given direction of regulation is applicable to the context at hand). Let $M_{activating}$ be the number of activating findings underlying an edge, and $M_{inhibiting}$ be the number of inhibiting findings. Then we define the weight of the edge leading to the *i*th gene as:

$$w_i = \frac{|M_{activating} - M_{inhibiting}|}{M_{activating} + M_{inhibiting} + 1}$$

(The choice of the constant in the denominator is arbitrary but reasonable – the weight of edges supported by a single finding will be 1/2, while the weight of edges with many consistent underlying findings will approach 1).

Let us use the same random variable x_i as defined above but now consider a weighted sum:

$$x = \sum_i w_i x_i$$

We then have:

$$\sigma_x^2 = \sum_i w_i^2$$

and the z-score statistic is given by:

$$z = \frac{x}{\sigma_x} = \frac{\sum_i w_i x_i}{\sqrt{\sum_i w_i^2}}$$

This latter formula is used in IPA to compute the activation z-score of a biological function.

USING THE ACTIVATION Z-SCORE TO INDEPENDENTLY CALL BIOLOGICAL FUNCTIONS

Using the activation z-score to identify biological functions independent their associated p-value is based on the idea that the z-score can detect “unlikely” matches between the patterns of observed up/down-regulation and activating/inhibiting edges downstream of the transcriptional regulator. However, this does not work in all situations. For instance a data set with only up-regulated genes would produce significant positive z-scores for any function with only activating upstream edges and sufficient overlap. In order to fix this, one needs to take a closer look at the underlying null model of the statistical test. While choosing the random variable x_i defined above to be 1 or -1 with equal probability works well for calling the activation state, a more appropriate null model for identifying biological functions should (at least approximately) be based on randomly permuting up/down-regulation labels of data set genes, and also (independently) permuting regulation directions associated with edges upstream of the function.

Let’s assume that the dataset has N_{up} genes that are up-regulated and N_{down} genes that are down-regulated. A random variable that models random choice of a gene from the dataset and which is set to +1 if that gene is up-regulated, and -1 if it is down-regulated has an expectation value of:

$$\mu_{data} = \frac{N_{up} - N_{down}}{N_{up} + N_{down}}$$

Likewise, let’s assume we have a biological function F that is connected to $N_{activating}$ genes through activating edges and to $N_{inhibiting}$ genes through inhibiting edges. A random variable that models random choice of a downstream edge and which is set to +1 if that edge was activating, and to -1 if that edge was inhibiting, then has an expectation value of:

$$\mu_{TR} = \frac{N_{activating} - N_{inhibiting}}{N_{activating} + N_{inhibiting}}$$

If both choices (random gene and random edge) are made independent of each other, the expectation value for a random variable representing random assignment of the predicted activation state of the biological function is:

$$\mu = \mu_{data} \cdot \mu_{TR}$$

However, for the purpose of the activation z-score this random variable (the x_i defined above) had been defined with $\mu = 0$. We can therefore use the actual value of μ (called “bias” here) as calculated above as a criterion for using the activation z-score to call significant biological functions. For practical purposes we require $|\mu|$ < 0.25, and flag all functions that do not meet that requirement as “biased” (the word “bias” appears in a new “Notes” column in the IPA user interface in these cases). This means that both gene regulation in the data set as well as regulation of edges for a given biological function is skewed towards a particular direction. For an “unbiased” data set with (approximately) equal number of up- and down-regulated genes, the activation z-score can always be used as an independent test to call significant biological functions.

ADVANCED FEATURE: BIAS-CORRECTED Z-SCORE DEVELOPMENT

The “bias” term introduced above can also be used to modify the original z-score such that significance calls for biological functions can also be made in the biased case. Note, however that this bias-corrected z-score should not be used to infer the activation state if the bias is strong. The bias-corrected z-score was called “regulation z-score” in versions of IPA prior to the IPA Summer 2012 release (June ‘12), and is now made available as an option.

Instead of choosing $x_i = +/-1$ with equal probability we now assume that x_i has an expectation value (mean) μ (the “bias” defined above) that is different from zero. The variance is then given by:

$$\sigma^2 = 1 - \mu^2$$

and the mean and variance of the random variable x are:

$$\mu_x = \sum_i w_i \mu$$

and:

$$\sigma_x^2 = \sum_i w_i^2 \sigma^2 = \sum_i w_i^2 - \sum_i w_i^2 \mu^2$$

The z-score (again approximately following a normal distribution with zero mean and standard deviation one) is then given by:

$$z = \frac{x - \mu_x}{\sigma_x} = \frac{\sum_i w_i (x_i - \mu)}{\sqrt{\sum_i w_i^2 - \sum_i w_i^2 \mu^2}} \approx \frac{\sum_i w_i (x_i - \mu)}{\sqrt{\sum_i w_i^2}}$$

The last step (approximation) above assumes that $|\mu| \ll 1$ (the second term in the denominator will only produce higher-order terms in μ). In practice we find this to be a valid assumption. The approximation also makes computation of z more robust since it avoids the singularity especially since values for μ are determined empirically from the dataset and network properties.

CONCLUSION

The Downstream Effects calculation underlies a new tool in IPA that enables the visualization of biological trends in an experiment in a whole new way. Other tools only indicate which biological processes might be *affected*, ignoring the actual direction of change of the genes that were measured. Downstream Effects Analysis takes you to faster insights by predicting whether there is an increase or decrease in the biological processes (function) that interest you.