INGENUITY®
SYSTEMS

Ingenuity Upstream Regulator Analysis in IPA®

TABLE OF CONTENTS

INTRODUCTION

The goal of the IPA Upstream Regulator analytic is to identify the cascade of upstream transcriptional regulators that can explain the observed gene expression changes in a user's dataset, which can help illuminate the biological activities occurring in the tissues or cells being studied. IPA makes it easy to take this result even further by examining what biological processes, pathways, and diseases the transcriptional regulators and their targets may control, and how these upstream molecules may regulate one another.

The upstream regulator analysis is based on prior knowledge of expected effects between transcriptional regulators and their target genes stored in the Ingenuity® Knowledge Base. The analysis examines how many known targets of each transcription regulator are present in the user's dataset, and also compares their direction of change (i.e. expression in the experimental sample(s) relative to control) to what is expected from the literature in order to predict likely relevant transcriptional regulators. If the observed direction of change is mostly consistent with a particular activation state of the transcriptional regulator ("activated" or "inhibited"), then a prediction is made about that activation state. IPA's definition of upstream transcriptional regulator is quite broad – any molecule that can affect the expression of other molecules, which means that upstream regulators (or "transcriptional regulators" as they are referred to in this document) can be almost any type of molecule, from transcription factor, to microRNA, kinase, compound or drug.

For each potential transcriptional regulator ("TR") two statistical measures, an overlap p-value and an activation *z*-score are computed. The overlap p-value calls likely upstream regulators based on significant overlap between dataset genes and known targets regulated by a TR. The activation *z*-score is used to infer likely activation states of upstream regulators 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 upstream regulators independently from the overlap p-value, based on significant pattern match of up/down regulation.

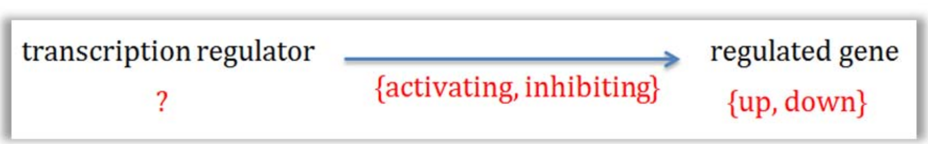
OVERLAP P-VALUE

The purpose of the overlap p-value is to identify transcriptional regulators that are able to explain observed gene expression changes. The overlap p-value measures whether there is a statistically significant overlap between the dataset genes and the genes that are regulated by a TR. It is calculated using Fisher's Exact Test, and significance is generally attributed to p-values < 0.01. Since the regulation direction ("activating" or "inhibiting") of an edge is not taken into account for the computation of overlap p-values the underlying network also includes findings without associated directional attribute, such as protein-DNA (promoter) binding.

ACTIVATION Z-SCORE

The primary purpose of the activation *z*-score is to infer the activation states of predicted transcriptional regulators[1].

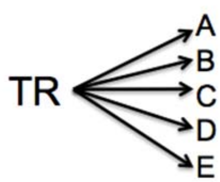
The basis for inference are edges (relationships) in the molecular network that represent experimentally observed gene expression or transcription events, and are associated with a literature-derived regulation direction which can be either "activating" or "inhibiting":



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

Regulation direction associated with edge	**Observed gene regulation**	**Predicted activation state of transcription regulator**
activating	Up	activated
activating	Down	inhibited
inhibiting	Up	inhibited
Inhibiting	Down	activated

In general a transcription regulator (TR) connects to several downstream genes in the data set:



For each of those regulated genes we can make a prediction about the activation state of the transcription regulator TR. In order to make the claim that activation (or inhibition) of TR is in fact responsible for the observed up/down-regulation pattern we could require the predicted activation state of TR 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. Also, genes are often modulated by several upstream regulators with possibly opposing effects, and it is not known which will dominate in a particular system. Therefore, we take a statistical approach by defining a quantity (z-score) that determines whether an upstream transcription regulator has significantly more "activated" predictions than "inhibited" 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 activated state and -1 to an inhibited 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 regulated by the regulator. Let:

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

where $N_{+/-}$ are the number of "activated"/"inhibited" 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 "activated" and "inhibited" 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 regulator (<0: inhibited, >0: activated). In practice, *z*-scores greater than 2 or smaller than -2 can be considered significant.

Note that in the case where the transcriptional regulator is encoded by a gene (for example is not a compound), the observed expression direction of the regulator itself (or whether it was even detectable on the expression platform) is ignored in the analysis. This is an advantage as there are many cases where negative feedback regulation reduces the expression of a transcriptional regulator although its protein product is still active.

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 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 the transcriptional regulator.

USING THE ACTIVATION Z-SCORE TO INDEPENDENTLY CALL UPSTREAM REGULATORS

Using the activation *z*-score to identify upstream regulators independent of the overlap 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 regulator with only activating downstream 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 regulators 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 downstream of the regulator.

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 an upstream regulator TR 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 transcription regulator 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 upstream regulators. For practical purposes we require $|\mu| < 0.25$, and flag all transcriptional regulators 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 transcriptional regulator 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 upstream regulators.

EXPLICIT EXAMPLE CALCULATION OF THE ACTIVATION Z-SCORE

A transcription factor TR regulates 5 downstream genes A, B, C, D, and E. A, C, and D are up-regulated in the dataset and B, E are down-regulated. The edges from TR are as follows: (TR,A): activating, weight=0.5, (TR,B): inhibiting, weight=0.9, (T,C): inhibiting, weight=0.8, (TR,D): activating, weight=0.8, and (TR,E): inhibiting, weight=1. We use the following formulas:

$$x = \sum_i w_i x_i$$

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

$$z = \frac{x}{\sigma_x}$$

Then predicted activation states of TR (x_i in the equations above) are obtained from the following table:

Regulated gene	Gene regulation	Regulation direction	Weight	Predicted state of Upstream Regulator (based on single gene)	
A	up	activating	$w_1 = 0.5$	activated	$x_1 = 1$
B	down	inhibiting	$w_2 = 0.9$	activated	$x_2 = 1$
C	up	inhibiting	$w_3 = 0.8$	inhibited	$x_3 = -1$
D	up	activating	$w_4 = 0.8$	activated	$x_4 = 1$
E	down	inhibiting	$w_5 = 1$	activated	$x_5 = 1$

With that we have:

$$x = 0.5 + 0.9 - 0.8 + 0.8 + 1 = 2.4$$

$$\sigma_x = \sqrt{0.5^2 + 0.9^2 + 0.8^2 + 0.8^2 + 1^2} = 1.8276$$

and:

$$z = \frac{2.4}{1.8276} = 1.3132$$

predicting an "active" state of TR. However, this would not be considered statistically significant. In general statistically significant hits are only expected for upstream regulators that are connected to a sufficient number of genes in the data set. For example, consider the case of a "perfect match" (i.e. all x_i are 1) with all weights set to 1 and $\mu = 0$ for simplicity. We then have $z = \sqrt{N}$ with *N* being the number of regulated genes. For *N* = 2, 3, 5, 10 we then have *z* = 1.41, 1.73, 2.24, 3.16.

EXPLICIT EXAMPLE CALCULATION OF THE BIAS TERM

Let us assume that 60% of the genes in the dataset are up-, and 40% down-regulated. Let us also assume that 35% of the edges downstream of TR are activating, and 65% inhibiting. Using the formulas:

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

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

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

we get:

$$\mu_{data} = 0.6 - 0.4 = 0.2$$

$$\mu_{TR} = 0.35 - 0.65 = -0.3$$

$$\text{"bias"} = \ \mu = 0.2 \cdot (-0.3) = -0.06$$

This is considered an "unbiased" situation (bias < 0.25) and the activation *z*-score can be used in a significance call for TR.

ADVANCED FEATURE: BIAS-CORRECTED Z-SCORE

This "bias" term introduced above can also be used to modify the original *z*-score such that significance calls for upstream regulators 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.

It turns out that bias-corrected z-scores are indeed approximately normally distributed with zero mean and standard deviation one for random data sets. We prepared random sets of 100 genes with randomly chosen up- or down-regulation and computed the empirical distribution function from pooled z-scores for regulators that connect to at least 5 downstream genes. This was done for parameters $\mu_{data} = 0, 0.5, 0.8$ (red, blue, and green data points in the diagram below). Results are indeed in good agreement with the CDF of a Gaussian distribution with zero mean and standard deviation one (straight line).

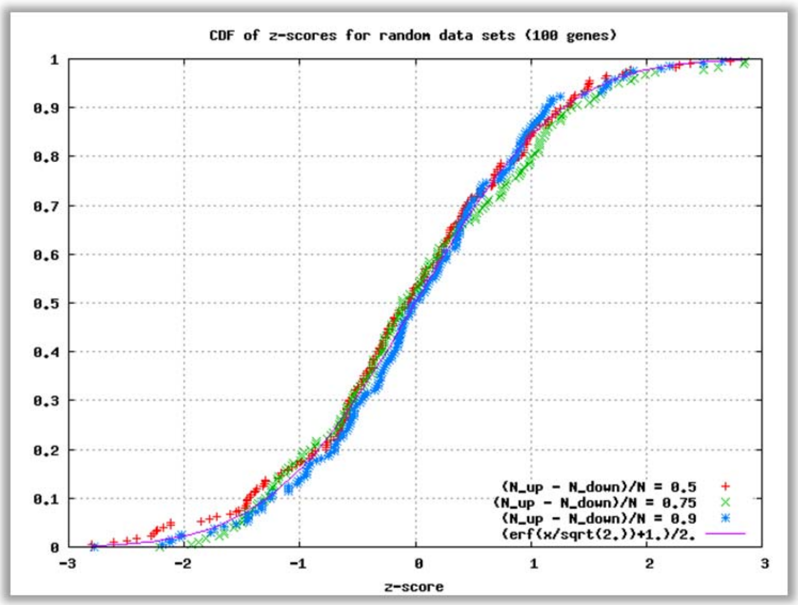


Figure 1: CDF of randomly sampled data.

CONCLUSION

Upstream Regulator Analysis in IPA is a remarkable step forward in the field of transcription regulator prediction. Unlike other tools, IPA will predict which transcriptional regulators are involved and whether they are likely *activated* or *inhibited*. IPA can then visualize this network of regulators and targets to explain how the regulators interact with one another and their targets to provide testable hypothesis for gene regulatory networks.

ENDNOTES

[1] Note that in the case of small molecules (drugs and compounds) or microRNAs predicted as upstream regulators, the terms “activated” or “inhibited” can be a bit confusing. In these cases, activated can be thought of as meaning that the expression pattern in the dataset is consistent with the upstream molecule having more activity (“activated”) or less activity (“inhibited”) in the experiment vs. the control.