

Predicted Biological Score (PBS®)

Introduction

The general strategy to identify potentially biologically significant interactions from yeast two-hybrid screens is to compute a predictive score for every single interaction based on experimental data. This score is referred to as the PBS®, or Predictive Biological Score. It is a statistical score relying on two different levels of analysis: firstly a local score taking into account the detailed results of a screen is computed for each interaction, and secondly, a global score is computed from the local scores by integrating results from all screens performed with the same library. As a consequence a local score is fixed when the corresponding screen had been completed, whereas global scores are computed incrementally as screening results are added to our knowledge database. Prior to the PBS computations, the distribution of fragments in the library and of overlapping prey fragments within screens must be determined:

Prior analysis #1: library fragment distribution

The distribution of fragments in prey libraries can be correctly determined only by randomly picking and sequencing a number of clones from the library allowing several times the coverage of the genome or, alternatively, by normalizing the library. Regarding the library calibration, this is intractable in practice for all but the smallest genomes (such as viruses).

Two different approximations are used depending on the prey library type, 'genomic' (derived from genomic DNA of viruses, prokaryotes or lower eukaryotes like yeast) or 'cDNA' (derived from RT-PCR of higher eukaryote mRNA):

i. Genomic prey libraries

Genomic prey libraries are built through mechanical, theoretically unbiased, breakage of genomic DNA. They are characterized by one number and two distributions :

- N_{ind} the number of independent fragments in the library: this number is approximated during the library construction process ; basically for Hybrigenics' library it is around 2 millions ;
- f_{size} the distribution of fragment size: this is experimentally determined by random sequencing of 100 to 200 prey fragments ;
- f_{start} the distribution of fragment start positions in the genome. This distribution is approximated by a uniform distribution:

$$f_{start}(x) = \frac{1}{s_g}$$

where s_g is the size of the genome from which the library is built from, and x is a nucleotide position in the genome. Regarding the way genomic libraries are built, this uniform f_{start} distribution is theoretically exact.

ii. cDNA prey libraries

These prey libraries are built from oligo(dT)- or random-primed reverse transcribed mRNA. If one can estimate and use the previously defined N_{ind} and f_{size} parameters to characterize the cDNA fragment distribution in such libraries, the f_{start} distribution is however almost impossible to correctly approximate. Indeed, the same distribution (using an approximation of the transcriptome size – basically the sum of lengths of all transcripts – for s_g) is false because of significant differences in mRNA representation in tissues and biases in cDNA synthesis.

We therefore prefer to use a lower grain approximation and characterize the fragment distribution in cDNA prey libraries only by a single distribution, $f_{presence}$, which represents the distribution of transcript occurrences in the library. This simpler distribution is still hard to define exactly because of the aforementioned variations of mRNA expression levels. As a complete characterization of transcript distribution would require random picking, sequencing and unambiguous identification of a high number of prey fragments, we approximate $f_{presence}$ by an uniform distribution:

$$f_{presence}(T) = \frac{1}{N_T}$$

where T is a given transcript and N_T is the total number of different transcripts in the library. This last parameter is initially approximated from literature data and eventually refined as additional transcripts are identified during library two-hybrid screening.

This hypothesis holds true for about 80% of transcripts and is used for simplicity's sake as well as to avoid the prohibitive cost of a full library calibration. But it is obviously a false approximation for very rare as well very abundant mRNA. This must be kept in mind for the results analysis (see the paragraph 'PBS interpretation' below).

Prior analysis #2: fragment distribution and gene identification

For each yeast two-hybrid screen, 5' and 3' sequences of all positive clones are determined and filtered by using PHRED (Ewing and Green 1998) and by masking ALU repeats. Sequence contigs are then built using CAP3 (Huang and Madan 1999) and compared to one or several reference database(s) using BLASTN (Altschul et al. 1997). A reference database can be an organism-specific database, such as GadFly for *Drosophila* or SGD for yeast, or a generic database such as the latest release of GenBank. If entries corresponding to the complete mRNA are found, the best annotated entry is assigned to every overlapping prey fragment family. The SID[®] (Selected Interacting Domain) is the common part of all overlapping fragments in a family. This region contains the domain that interacts with the bait. Each two-hybrid screening experiment generates a list of bait-SID interactions. Cardinalities for one screen are:

- 1 bait
- n SIDs assigned to m genes, with $n \geq m$ (several distinct fragment families may be assigned to the same gene).

Local PBS computation

Each fragment family is first analyzed in terms of coding capabilities (antisense and out-of-frame fusion fragments). The prey fragment families which have no or very improbable biological coding capability are then discarded:

- families containing only antisense fragments ;
 - families containing only out-of-frame fusion fragments selected in a single frame.
- Families containing several fragments in different frames (in-frame and out-of-frame or out-of-frame in the two non-coding frames) are kept for further analysis because we consider they encode valid biological polypeptides thanks to translational frameshift events which occur in yeast.

The local score is computed as an E-value for each remaining fragment family, by comparing observed results to a theoretical random background: the fragment distribution is compared to an expected distribution based on a statistical model of the two-hybrid experiment mechanism and calibrated by the specific prey library characteristics. This step is performed in two sub-steps.

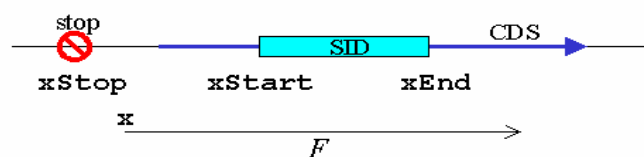
* First, for each overlapping fragment family, that is for each SID, we calculate the probability p that a fragment F randomly drawn from the library contains the SID.

For genomic libraries, p is calculated as follows:

$$p = \int_{x=xStop}^{xStart} \int_{s=xStart-x+1}^{+\infty} f_{start}(x) f_{size}(s) dx ds$$

where:

- $xStart$ is the start position of the SID ;
- $xStop$ is the position of the first STOP codon upstream and in frame with the CDS containing the SID ;
- f_{start} and f_{size} are the genomic prey library distributions defined above ;
- variables x and s represent the start and the size of the fragment F , respectively:



For cDNA libraries, p is simply approximated as follows:

$$p = f_{presence}(T)$$

where:

- $f_{presence}$ is the cDNA distribution defined above;
- T is the transcript containing the SID.

* Second, for each fragment family or SID, we calculate the PBS as the probability to pick randomly from the library at least the number n of overlapping fragments containing the SID (each having the probability p defined above) out of the N fragments identified in the screen:

$$PBS = \sum_{X=n}^N C_N^X p^X (1-p)^{N-X}$$

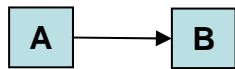
So this local PBS value represents the probability of a bait-SID interaction being non-specific. It measures how significantly different the profile of overlapping fragments is from a theoretical background noise representing a completely non-specific selection (simulated by the random drawing of fragments from the library).

Global PBS computation

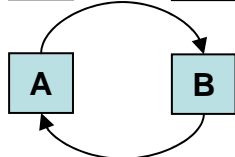
The PIM[®] (Protein Interaction Map) is then built by pooling all the bait-SID interactions from several screens. The result is a graph where vertices (nodes) are proteins and edges are protein-protein interactions, defined by one or several different bait-SID interactions.

The global PBS is computed from local PBS in two steps:

- the global connectivity of the interaction map is analyzed to mark highly connected prey polypeptides (SIDs which are found as prey with frequency above a fixed threshold). Note that this step tags domains, rather than whole proteins, as potentially “sticky”. In that case, the local PBS of the bait-sticky_SID interaction is set to 1.
- local E-values are combined when the same protein pair is involved, assuming that bait and prey fragments overlap for both proteins:

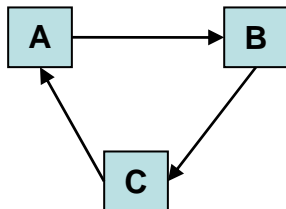


$$GlobalPBS_{A-B} = LocalPBS_{A \rightarrow B}$$

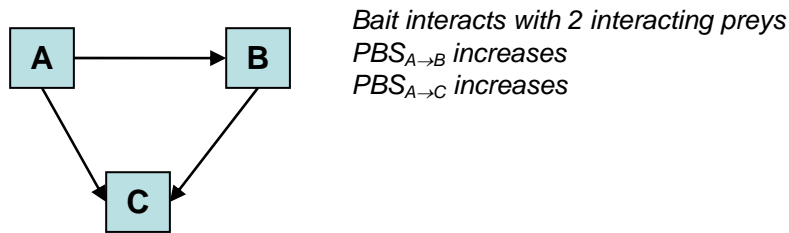


$$GlobalPBS_{A-B} = LocalPBS_{A \rightarrow B} \cdot LocalPBS_{B \rightarrow A}$$

The interaction map is searched for 3-cycles (i.e. small connectivity patterns linking 3 proteins) and PBS may be increased if they mimic probable biologically relevant networks. For examples:



Indirect Rebound :
 $PBS_{A \rightarrow B}$ increases
 $PBS_{B \rightarrow C}$ increases
 $PBS_{C \rightarrow A}$ increases



In conclusion, the PBS score reflects the probability that an interaction is found by chance. It ranges from 0 to 1, but is grouped in five categories (A, B, C, D, and E) for user convenience. Inter-category thresholds are chosen manually with respect to a training data set containing known true-positive and false-positive interactions (not shown): $A < 1e-10 < B < 1e-5 < C < 1e-2.5 < D < 1$. The E category gathers scores equal to 1.

PBS interpretation

PBS category	Interpretation
A B C	These interactions are technically very reliable. They correspond to interactions found in two reciprocal and independent screens (A->B and B->A) or interactions found in a single screen with many overlapping prey fragments.
D	Basically these interactions are defined by a single bait-SID interaction, the SID being defined by a singleton fragment instead of a family of several overlapping fragments. This category is the hardest to analyze because it mixes two classes of interactions: <ul style="list-style-type: none"> - false-positive interactions (background noise): one singleton fragment has been selected by chance by the bait (non-specific selection) - interactions hardly detectable by two-hybrid systems because of conformation, toxicity in yeast, very low representation of the mRNA in tissue (rare mRNAs, see above) etc... This class of interactions is potentially very interesting because Hybrigenics' PIM technology is the only one able to detect them.
E	These interactions involve SID that have non-specifically been found as prey in many independent screens. They are likely to be false-positives of the two-hybrid system.