Summary probabilistic scores and data-specific metrics in FunCoup

Summary scores from Naïve Bayesian network

The naïve Bayesian network (NBN) is trained as follows. First, the discretisation algorithm finds the bins in each metric's range that produce the highest contrast in respect of functional coupling (FC). In our framework, we analyse the set ε of evidence features $E_i \in \varepsilon$ to estimate the integrated support for FC given all non-empty evidences. Starting with an individual evidence E_i that falls in the bin j, the probability that a particular gene pair is functionally coupled is defined by Bayes' rule:

$$P(FC \mid E_{ij}) = \frac{P(FC)P(E_{ij} \mid FC)}{P(E_{ij})} \qquad (1)$$

The probabilities corresponding to 4 different FC classes are here collapsed into one *FC* for brevity. We integrate over all evidences by summing the logarithms of the remaining ratio in order to obtain a simplified classifier called *final Bayesian score, FBS*:

$$FBS(\varepsilon) = \sum_{i=1}^{|\varepsilon|} \log \frac{P(E_{ij} \mid FC\ )}{P(E_{ij})} \qquad (2)$$

where $P(E_{ij} \mid FC\ )$ is estimated from occurrence of E_{ij} in the positive training set. The background probabilities $P(E)$ are estimated from the general population of gene pairs. Only couplings with FBS>3 are kept.

FBS is convenient to store and analyse evidence components. Some components may be negative despite an overall positive *FBS*. However, the *FBS* score does not have strict bounds, and is not intuitively interpretable. We therefore use an approximation of an alternative form of Bayes' theorem (MacKay, 2003) that gives an intuitive and user-friendly FunCoup confidence score between 0 and 1:

$$pfc(\varepsilon) = \frac{P(FC)\prod_{i=1}^{|\varepsilon|} P(E_{ij} \mid FC)}{P(FC)\prod_{i=1}^{|\varepsilon|} P(E_{ij} \mid FC) + \prod_{i=1}^{|\varepsilon|} P(E_{ij})} \qquad (3)$$

pfc is a probability estimate that the pair is functionally coupled.

Metrics of pairwise similarity

1.1. mRNA co-expression

Co-expression metrics
PLC stands for Pearson linear correlation coefficient

http://mathworld.wolfram.com/CorrelationCoefficient.html

1.2. Protein-Protein Interactions (PPI)

PPI-based evidence worked well across species and FC classes. In fact, the likelihood values were much stronger affected by the quality of training and test sets then by the evolutionary distance.

The protein interactions are often reported in the binary form, i.e. listing pairs of proteins. However, there have been multiple attempts to differentiate them in respect of confidence (Giot et al., 2003; von Mering et al., 2005; Suthram et al., 2006). Many approaches re-evaluated PPI confidence *post hoc*, matching them with independent evidence (equivalent to final score assignment from multiple data sources in FunCoup). Others employed "insider" information: number of tested baits / preys, their relative output, produced network topology etc. Meanwhile, times have changed and the analysis faced novel challenges:

1) it is feasible now to compile large multi-source datasets and evaluate them against gold standard sets in several organisms;

2) many PPIs have been reported more than once in peer-reviewed publications, and these independent confirmations became the major evidential factor.

The modern release of IntAct database keeps track of multiple replicates, experiments, methods, and publications that support a particular interaction. While even two independent experimental reports on an interaction drastically raise its confidence, many PPIs have been confirmed by 3 or more publications with in total >10 experiments.

Having compiled data from IntAct, HPRD, GRID, BIND databases plus some published datasets not yet there, we accumulated 183983 distinct protein-protein interactions in human, 38717 in *D. melanogaster*, 40732 in *M. musculus*, 183546 in *R. norvegicus*, 8887 in *C. elegans*, 173178 interactions in *S. cerevisiae*, and 3000 in *A. thaliana*. 38738 of these interactions were reported more than once. Moreover, one article might have reported validation of the same interaction in more than one assay/experiment. Normally, yeast-2-hybrid and some other techniques employ a binary approach and discover PPI as asymmetric interactions of a "bait" protein with a "prey" protein. A bait might be then tested positive against multiple preys. Although the latter are not observed in direct interactions with each other, e.g. IntAct database presents them as members of the same interaction. We accounted for this information and included "prey-prey" interactions into our compilation. To account for such cases as well as multiplicity of reports and experiments, we introduced a novel score. It favors support from *different* articles in PubMed. The details are parsed out of the current release of the IntAct database, stored in XML format PSI-MI 2.5. This data source (October 2007) became the principal part of our compilation. Interactions from HPRD, GRID, BIND, and the large list of yeast complex members reported by Krogan et al. (2006) were added when they were traced to other PubMed publication IDs, or simply unique.

The score was similar to the main formula of Bayesian probability of functional coupling (see *Methods*), and combined the probabilistic scores S_+ (for being coupled) and S_- (for *not* being coupled):

$$S_{PPI} = \frac{S_+}{S_+ + S_-},$$

where $S_+ = P(PPI)\prod_{p=1}^{|Papers|}\prod_{a=1}^{|Assays_p|}\frac{pc_+}{\sqrt{|Assays_p(A,B)|}\cdot\log_2|IP_a(A,B,..)|}$

and $S_- = P(PPI)\prod_{p=1}^{|Papers|}\prod_{a=1}^{|Assays_p|}pc_-$

Thus, the score grew with the number of individual experiments (|*Assays*|) reported the interaction between proteins *A* and *B* (pairwise or in a group/complex). The total number of partners |IP_a| in the interaction involving simultaneously *A* and *B* penalized multi-member interactions. On the other hand, an article might present positive results on the same interaction from multiple assays. While this was a useful feature, we also wanted to distinguish such reports from ones coming from different papers, as the latter are even more reliable. Hence, the number of assays in the same article *p* was square-rooted.

The score was not intended to deliver the exact probability of interaction. Hence, the probabilities:

- *P*(*PPI*), "an interaction exists between a pair of proteins", 0.001
- pc_+, "a single positive report is published given the interaction is true", 0.1
- pc_-, "a single positive report is published given the interaction is false" 0.001

were assigned roughly, equaling, and, respectively.

For example, two proteins reported to interact only in one paper, one assay, and without a third partner, received a score 0.091. If the assay had listed hundreds of interaction partners, the score fell down to 0.011. The highest score reached in the current IntAct dataset is 0.99999899. It was assigned when e.g. an interaction was tested in 10 assays reported in 2 papers or in 6 assays each published in a separate paper. After the discretisation, the range of score S_{PPI} was split into several bins with distinct LLR ranging between 1.5 and 8.5.

1.3. Phylogenetic profiling

As the number of distinct co-presence profiles is finite, they can be expressed as unique *phylogenetic signatures.* For example, {`mammal, fly, worm`} denotes presence of both genes in mammals, *D. melanogaster*, and *C. elegans*. Then likelihoods is assigned to such individual strings. However, the total number of distinct signatures should not be too large – this would lead to insufficient number of observations of each category in the training procedure. For this reason, the situations "both absent" and "one present, one absent" are collapsed together, and the number of species is reduced to that of major clades (*Fungi, Plantae, Animalia*) plus closer relatives.

The LL ratios range from –2 ({`human, worm`} in human FC-PI) to +3.5 ({`human, mouse, rat, fly, worm, yeast, plant`} in human FC-CM).

1.4. Sub-cellular co-localization

The common formula for mutual information has been modified:

$$WMI_{SLC} = \sum_{a=\{0,1\}}\sum_{b=\{0,1\}} P(i=a, j=b)\log\frac{P(i=a, j=b)}{P(i=a)P(j=b)}$$

where *a, b* are the "presence/absence" indicators of proteins *i* and *j* in locations *L*. The absence/presence observations in each sub-cellular location *l* are accounted as complement to the relative size of *l*:

$c(l) = 1 - n_l / N_L$.

Therefore

$$P(i = a) = \frac{\sum_{l=1}^{N_L} c(l)}{N_L},$$

where N_L is the total number of protein localizations mentioned for the organism. The fewer proteins assigned to *l*, the more informative is the co-localization.

1.5. Protein-DNA binding

A "TF -> target" pair is a link between two genes and can be considered as input for FunCoup *per se*. However, such links would not overlap with the training sets used in FunCoup (FC-ML, FC-PI, FC-CM), and only rarely would with FC-SL. Hence, a confident LLR could not be derived in the training procedure. We thus provide such couplings as additional information. On the other hand, the FC between genes – targets of the same (sets of) TFs can be tested with any of our training sets. The score for overlap is calculated as:

$$S_{TF-ctgt} = \frac{|\{BS_i\} \cap \{BS_j\}|^2}{|\{BS_i\} + \{BS_j\}|},$$

i.e. the shared fraction of the binding site sets $\{BS\}$ between genes *i* and *j* is multiplied by the cardinality of the shared subset. (identically to the S_{miRNA_ctgt} score).

1.6. miRNA-gene targeting

For each pair of genes *i* and *j* with sets of predicted miRNA target sites having the complementarity score > 15 (about 80% of the database content), a functional coupling score for overlap is calculated as:

$$S_{miRNA-ctgt} = \frac{|\{BS_i\} \cap \{BS_j\}|^2}{|\{BS_i\} + \{BS_j\}|},$$

i.e. the shared fraction of the binding site sets $\{BS\}$ between genes *i* and *j* times the cardinality of the shared subset (identically to the S_{TF_ctgt} score).

1.7. Protein co-expression

The Human Protein Atlas (Hober and Uhlen, 2008) provided data on staining 1400 cell line and tissue samples with antibodies to about 3000 human proteins. Each sample had been analyzed for dye intensity of the staining antibody and received a grade ("white": negative, "yellow": weak, "orange": moderate, and "red": strong). The protein co-expression score in FunCoup had thus to deal with quantitatively ordered coarse-resolution data. We tested a number of opportunities and found an optimal score.

The common formula for mutual information (http://mathworld.wolfram.com/MutualInformation.html) is:

$$WMI_{PEX} = \sum_{a=\{w,y,o,r\}} \sum_{b=\{w,y,o,r\}} P(i = a, j = b) \log \frac{P(i = a, j = b)}{P(i = a)P(j = b)}$$

where *a, b* are the color indicators of the staining of genes *i* and *j*. To account for the respective color abundance in each sample, the sample-specific weighting coefficients rather

then unities are summated along the staining profiles. A weight coefficient was the fraction of color grade *a* (one of {“w”, “y”, “o”, “r”}) in sample *s*:

$w(s,a) = n_{s,a} / N_s$,

where N_s is the total number of successfully stained genes in sample *s*, i.e. the sum of all “w”, “y”, “o”, and “r”′s. Hence,

$$P(i=a) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,a)}{N_{i\cap j}} \text{ , } P(j=b) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,b)}{N_{i\cap j}} \text{ , and } P(i=a, j=b) = \frac{\sum_{s=1}^{N_{i\cap j}} w(s,a)\cdot w(s,b)}{N_{i\cap j}}$$

rather than actual probabilities of *i*=*a* etc. in the common formula. $N_{i\cap j}$ is the number of cell/tissue samples which both *i* and *j* stained successfully.

The “*cell line*” and “*tissue*” subsets of HPA are processed as separate datasets and yield distinct likelihood values.