

# Validating Y2H Data from Intact

T Chiang

April 20, 2006

First we create a table of various general statistics to give a sense as to how similar each separate experiment is with respect to every other individual experiment:

The various out degree statistics for the y2h data:

```
> tab <- createTables(y2hSysGW)
> tab
```

	Total Baits	Total Interactions	Bait per Interaction
ito-2001-1	1522	4524	0.336427940
cagney-2001-1	19	54	0.351851852
tong-2002a-1	20	125	0.160000000
hazbun-2003-1	66	2524	0.026148970
zhao-2005-2	1	102	0.009803922
uetz-2000-1	508	952	0.533613445
uetz-2000-2	139	524	0.265267176

  

	Minimum Degree	Maximum Degree	Median Degree	Mean Degree
ito-2001-1	1	279	1	2.972405
cagney-2001-1	1	6	2	2.842105
tong-2002a-1	1	18	4	6.250000
hazbun-2003-1	1	606	14	38.242424
zhao-2005-2	102	102	102	102.000000
uetz-2000-1	1	24	1	1.874016
uetz-2000-2	1	26	2	3.769784

Now we describe some more non-trivially derived statistics on the y2h di-graphs.

We wanted to test one hypothesis that deals with the nature of the Y2H experiments - namely that because the environment of the experiment is nuclear, nuclear proteins will display considerable more interactions than non-nuclear proteins.

We tested this hypothesis in two ways. In the first method, we take the set of baits that belong (or annotated) to the nucleus, denoted as  $\{B^n\}$ , and compare this bait set to two disjoint prey sets: the first prey set consists of those prey that are annotated to the nucleus, denoted as  $P^n = \{p^n\}$  and the second are those that are not, denoted as  $P^{\bar{n}} = \{p^{\bar{n}}\}$ . For each bait  $b_k \in \{B^n\}$ , we find two sets of averages:

The first are averages on the nuclear prey protein -

$$\{b_k^1\}_{k \in |B^n|} = \frac{|\text{Hits of } b_k \in P^n|}{|P^n|} \quad (1)$$

The second are averages on the non-nuclear prey protein -

$$\{b_k^2\}_{k \in |B^n|} = \frac{|\text{Hits of } b_k \in P^{\bar{n}}|}{|P^{\bar{n}}|} \quad (2)$$

Our conjecture is that the nuclear environment plays a critical role in the participation of binary interactions between proteins. Central to this conjecture is the premise that binary interactions between two nuclear proteins will be more prevalent than interactions between nuclear baits and non-nuclear preys.

Statistical verification of this conjecture is done by comparing each  $b_k^1$  with  $b_k^2$ . The first column in the table below gives the number of times when  $b_k^1 > b_k^2$  while the second column gives  $b_k^1 < b_k^2$ .

> nStat

	h1	h2
ito-2001-1	0.0005588562	0.0003828227
cagney-2001-1	0.0117187500	0.0283203125
tong-2002a-1	0.0020790021	0.0268744961
hazbun-2003-1	0.0009916477	0.0004755863
zhao-2005-2	0.0000000000	0.0153808594
uetz-2000-1	0.0010797448	0.0009087589
uetz-2000-2	0.0027264982	0.0018497174

The second method to verify this conjecture involves some of the same workings of the first.

In this method, we wanted to ascertain the reciprocity of nuclear/nuclear interactions versus nuclear/non-nuclear interactions. Testing for symmetry, we needed use a different subset of the baits than merely the nuclear baits  $B^n$ . This time we made use of the set of baits  $B^b$  that found at least one other bait as a hit:

$$B^b = \{b \in \text{Baits} \mid b \rightarrow p \text{ for some } p \in \text{Baits}\}. \quad (3)$$

In addition, we have also restricted the prey set to only those which were sampled as baits as well (call these  $\ddot{P}$ ) since if any prey  $p$  was not sampled as a bait, we cannot test for reciprocity. Again we divide  $\ddot{P}$  into two disjoint sets:  $\{\ddot{p}^n\}$  and  $\{\ddot{p}^{\bar{n}}\}$  defined similarly to our last method. For each nuclear bait  $b_k \in B^n \cap B^b$ , we again find two sets of averages:

The first -

$$\{\ddot{b}_k^1\}_{k \in |B^n \cap B^b|} = \frac{|\text{Hits of } b_k \in \{\ddot{p}^n\} \text{ that also find } b_k \cdot|}{|\{\ddot{p}^n\}|} \quad (4)$$

The second -

$$\{\ddot{b}_k^2\}_{k \in |B^n \cap B^b|} = \frac{|\text{Hits of } b_k \in \{\ddot{p}^{\bar{n}}\} \text{ that also find } b_k \cdot|}{|\{\ddot{p}^{\bar{n}}\}|} \quad (5)$$

Again, this method helps in the ascertaining the veracity of our conjecture by comparing  $\ddot{b}_k^1$  with  $\ddot{b}_k^2$ . Again the first column in the table below gives the number of times when  $\ddot{b}_k^1 > \ddot{b}_k^2$  while the second column gives  $\ddot{b}_k^1 < \ddot{b}_k^2$ .

```

> nStat2

      u1 u2
[1,] 29 11
[2,]  0  2
[3,]  0  0
[4,]  6  1
[5,]  0  0
[6,] 10  1
[7,]  6  2

```

This second method verifies the completeness of the Y2H experiment in verifying a symmetric interaction whereas the first method simply verifies the cardinality of interactions.

While we have devised two different methods to test our hypothesis, we cannot conclude that nuclear/nuclear interactions are more prevalent nor more complete for any of the experiments with the exception of the *uetz* – 2000 – 1 experiment.

We conclude with a few more general statistics on the derived Y2H data sets.

The table below accounts for two statistics, *c1* gives the number of baits that had been apart of at least one symmetric interaction (denoted as  $B^s$ ) while *c2* simple details the total number of baits.

```

> c3

      c1  c2
ito-2001-1  194 1522
cagney-2001-1  6  19
tong-2002a-1  2  20
hazbun-2003-1 13  66
zhao-2005-2  0  1
uetz-2000-1  61 508
uetz-2000-2  18 139

```

The numbers in column *c1* are somewhat misleading since they account for the bait proteins which are involved in homodimers (which are theoretically symmetric relationships). The following table accounts for the number of non-trivial symmetric interactions (given by the column *fir* and the total possible number of symmetric interactions based on the total number of baits sampled. (given by *sec*). The third column is the ratio of the first two.

```

> c4

      fir      sec
ito-2001-1  75 1157481 6.479588e-05
cagney-2001-1  3   171 1.754386e-02
tong-2002a-1  1   190 5.263158e-03
hazbun-2003-1  4  2145 1.864802e-03
zhao-2005-2  0     0      NaN
uetz-2000-1  10 128778 7.765302e-05
uetz-2000-2  7  9591 7.298509e-04

```

We also tried to constrain the total bait population. Again *c1* represents the number of baits involved in a symmetric interaction while *rl* is the restriction of the total sampled bait set to those baits that were also experimentally assayed as a hit as well, denoted  $B^p$ ; in graph theoretic terms, the in-degree  $\forall b \in B^p \geq 1$ . This condition substantially reduces the number of baits by which we make statistical inference. It is clear that  $B^s \subset B^p$ .

```
> c5
```

	c1	rl
ito-2001-1	194	738
cagney-2001-1	6	11
tong-2002a-1	2	5
hazbun-2003-1	13	28
zhao-2005-2	0	0
uetz-2000-1	61	114
uetz-2000-2	18	35

The following table details the number of non-trivial symmetric interactions (*fir*), the number of all possible symmetric interactions for the set  $B^p$  (*thi*), and the ratio of the two.

```
> c6
```

	fir	thi	
ito-2001-1	75	271953	0.0002757829
cagney-2001-1	3	55	0.0545454545
tong-2002a-1	1	10	0.1000000000
hazbun-2003-1	4	378	0.0105820106
zhao-2005-2	0	0	NaN
uetz-2000-1	10	6441	0.0015525540
uetz-2000-2	7	595	0.0117647059

```
> inOutD <- list()
> for (i in 1:length(bpStat2)) {
+   outD <- rowSums(bpStat1[[i]])
+   inD <- colSums(bpStat1[[i]])
+   inOutD[[i]] <- list()
+   inOutD[[i]]$outD <- outD
+   inOutD[[i]]$inD <- inD
+ }
> deg <- lapply(inOutD, function(d) {
+   d$outD - d$inD
+ })
> diffStat1 <- lapply(deg, function(x) {
+   sum(x < 0)
+ })
> diffStat2 <- lapply(deg, function(x) {
```

```

+     sum(x > 0)
+ })
> absStat <- lapply(deg, function(x) {
+     sum(abs(x) <= 2)
+ })

> bpStat3 <- vector("list", length = length(y2hSysGW))
> for (i in 1:length(y2hSysGW)) {
+     bpStat3[[i]] <- genBPGraph(y2h[i], nonWeighted = TRUE, homodimer = FALSE)
+ }
> names(bpStat3) <- names(y2hSysGW)
> symTest = sapply(bpStat3, function(x) {
+     x %% x
+ })
> one = lapply(symTest, diag)
> two = lapply(bpStat3, rowSums)
> three = lapply(bpStat3, colSums)
> tableStuff <- list()
> for (i in 1:length(symTest)) {
+     tableStuff[[i]] <- cbind(cbind(one[[i]], two[[i]]), three[[i]])
+     tableStuff[[i]] <- tableStuff[[i]][-(which(rowSums(tableStuff[[i]]) ==
+         0)), , drop = FALSE]
+ }
> resTable <- list()
> for (j in 1:length(tableStuff)) {
+     resTable[[j]] <- tableStuff[[j]][which(tableStuff[[j]][,
+         2] >= 10), , drop = FALSE]
+ }
> lookAtPreys <- list()
> for (k in 1:length(resTable)) {
+     lookAtPreys[[k]] <- colSums(bpStat3[[k]][rownames(resTable[[k])],
+         , drop = FALSE])
+ }
> res2Table <- list()
> for (j in 1:length(tableStuff)) {
+     res2Table[[j]] <- tableStuff[[j]][which(tableStuff[[j]][,
+         3] >= 10), , drop = FALSE]
+ }
> lookAtBaits <- list()
> for (k in 1:length(resTable)) {
+     lookAtBaits[[k]] <- rowSums(bpStat3[[k]][, rownames(res2Table[[k])],
+         drop = FALSE])
+ }
> bpStat4 <- vector("list", length = length(y2hSysGW))
> for (i in 1:length(y2hSysGW)) {

```

```
+     bpStat3[[i]] <- genBPGraph(y2h[i], nonWeighted = FALSE, homodimer = FALSE)
+ }
> names(bpStat3) <- names(y2hSysGW)
```