A novel pooling design for protein-protein interaction mapping

Nicolas Thierry-Mieg[(1)], Laurent Trilling[(1)], and Jean-Louis Roch [(2)]

[(1)] Laboratoire LSR-IMAG, BP 53, 38041 Grenoble Cedex 9, France
[(2)] Laboratoire ID-IMAG, 51, avenue Jean Kuntzmann, 38330 Montbonnot-Saint-Martin, France
{Nicolas.Thierry-Mieg,Laurent.Trilling,Jean-Louis.Roch}@imag.fr

Keywords. Group testing, protein-protein interactions, high-throughput screening

Introduction

Protein-protein interactions are critical in a wide range of biological processes, from the formation of macromolecular complexes to the transduction of signals in biological pathways. As such, they are the focus of several high-throughput functional genomics projects worldwide, relying on yeast two-hybrid systems ([1], [2], [3]), or more recently on mass spectrometry of purified complexes ([4], [5]). We work in close collaboration with Marc Vidal's group at Dana Farber Cancer Institute ([6], [7]), whose ultimate goal is to identify all protein-protein interactions in *Caenorhabditis elegans*, using a high-throughput version of the two-hybrid system [8].

As of today, most protein interactions from this lab have been identified by screening selected genes used as "baits" against a cDNA library. This method is reliable and quickly productive, but it requires the often redundant sequencing of all positives, therefore becoming costly and time-consuming at the genome scale. In order to attempt to identify the complete interactome efficiently, it might be advantageous to favor the so-called matrix approach, where pairs of known proteins are tested for interaction. Such a method requires the initial cloning of all nematode coding regions, a goal largely achieved in the Vidal lab by way of the *C. elegans* ORFeome project ([9], [10]).

With this resource, a genome-scale two-hybrid matrix approach becomes possible, but remains a daunting task. However, the proportion of expected positive protein pairs is small, of the order of 0.02% according to the data collected up to now [11]. What's more, the two-hybrid system can easily be applied in a semi-automated setting to a set of bait and fish proteins, in order to examine in a single test (i.e., a single spot on an agar plate) whether at least one bait-fish pair interacts. Therefore, it could prove highly efficient to use a pooling system, as is described below.

The group testing problem can be described as follows. Consider a set of K events which can be true or false, represented by the Boolean variables $\{A_1, A_2, \ldots, A_K\}$. For example, the event "bait B interacts with prey P_i" can be represented by the Boolean variable A_i. Let us call "pool" a disjunction of variables. The goal is to build a set of p pools with p minimal, such that by testing the values of the p pools, one can unambiguously determine the values of the K variables. Note that if all variables can be true simultaneously, one cannot hope to build less than K pools. This corresponds to the trivial "canonical" pooling system, where each pool is a singleton and each variable is present in exactly one pool. However, in many biological settings the total number of positive variables is small. In this situation, by building well-chosen redundant pools, it is possible to build sets of pools where p is much smaller than K. What's more, the redundancy inherent to pooling designs allows the detection and correction of some experimental non-systematic artifacts. Several pooling designs have been previously described and used in high-throughput biology (e.g. [12], [13], [14]).

In this work we present initial results concerning a novel pooling design, which displays improved performance and/or greater adaptability compared to previously described systems. This pooling design has been developed to allow the efficient and robust identification of protein-protein interactions by the yeast two-hybrid system on a genomic scale.

The shuffled multi-partition pooling design

In this section, a novel pooling design method, which we call "shuffled multi-partition", is described. We show that if at most τ variables are true among K variables (of unknown values), the values of the K variables can all be deduced by observing the values of a small number of pools generated by this method.

The pooling design: Let q be a prime number, and K be an integer such that K>q. We define the compression value of q relative to K, noted *comp(q,K)*, as the smallest integer c such that $q^{c+1} = K$. We will simply write *comp* for *comp(q,K)* whenever possible. Consider K Boolean variables $A_0, \ldots, A_{K-1}$ of unknown values. Let us call *pool* a disjunction of variables, represented as a subset of $\{A_0, \ldots, A_{K-1}\}$. A set P(L) of pools is built as follows. Successive *layers* of pools of variables are built, where each layer is a partition of $\{A_0, \ldots, A_{K-1}\}$. P(L) is composed of all the pools in the first L layers. This pooling design is therefore a so-called transversal design [13].

In this poster presentation, a procedure to build each layer will be described, and proofs for the following theorems will be presented.

Theorem: Consider the K Boolean variables $A_0, \ldots, A_{K-1}$, and let q be a prime integer with K>q. For $L \in \{0,..,q-1\}$, let P(L) be the set of pools defined by our shuffled multi-partition pool generation procedure. We will note *pools(i)* the set of pools that contain variable A_i: $\forall\ i \in \{0,\ldots,K-1\}$, *pools(i)* = $\{p \in P(L_{max}) \mid A_i \in p\}$. The following property holds: $\forall\ i_1,i_2 \in \{0,\ldots,K-1\}$, $[i_1 \neq i_2] \Rightarrow [\mathrm{Card}(pools(i_1) \cap pools(i_2)) \leq comp(q,K)]$.

Corollary: Let τ be an integer with $(\tau \times comp) < q$. Let $P(\tau \times comp)$ be the set of $(\tau \times comp + 1) \times q$ pools generated as described above. Then, if there are at most τ positive (i.e., true) variables among $A_0,\ldots,A_{K-1}$, the value of each variable can be identified unambiguously by examining the value of each pool from $P(\tau \times comp)$.

Example: Consider the K=9 variables named $\{A_0, A_1, \ldots, A_8\}$, and let q=3 and τ=1. Consider the pools from layers 0 and 1 as generated by our system, i.e. P(1) = $\{\{A_0,A_3,A_6\}, \{A_1,A_4,A_7\}, \{A_2,A_5,A_8\}, \{A_0,A_5,A_7\}, \{A_1,A_3,A_8\}, \{A_2,A_4,A_6\}\}$.

For reasons of symmetry, assuming a single variable is positive, the name of that variable is inconsequent: all are equivalent. Let us suppose that the only positive variable is A_8. Then pools $\{A_0,A_3,A_6\}$, $\{A_1,A_4,A_7\}$, $\{A_0,A_5,A_7\}$, and $\{A_2,A_4,A_6\}$ are negative, which shows that variables $A_0, A_1, \ldots, A_7\}$ are negative; and pools $\{A_2,A_5,A_8\}$ and $\{A_1,A_3,A_8\}$ are positive, which each prove that A_8 is positive (given that A_2, A_5, A_1 and A_3 have been shown to be negative).

Extensions

In this section, we present two extensions of the strict group testing problem that reflect the real-world requirements of high-throughput protein-protein interaction mapping. We show how the shuffled multi-partition pool generation procedure can be applied to tackle both of these extensions.

σ-satisfactory sets of pools: In the previous section we described a method for building a set $P(\tau \times comp)$ of pools of Boolean variables, such that if at most τ variables are positive, it is guaranteed that the value of each variable can be deduced from the values of the pools. In practice, this very strong property might not be required: one may consider sufficient that the values of a large proportion of the variables be deduced from the values of the pools.

More specifically, let $\{A_0, \ldots, A_{K-1}\}$ be a set of Boolean variables of unknown values, and let P be a set of pools on these variables. In general, observing the values of the pools leads after analysis to three classes of variables: variables that are false; variables that are true; variables that are ambiguous. Let us call *ambi* the number of ambiguous variables. Let $\sigma \in [0..1]$. We will say that P is *σ-satisfactory* if on average (over all possible values of $(A_0, \ldots, A_{K-1})$ such that at most τ are positive), $ambi \leq \sigma \times K$.

Note that the strict group testing problem corresponds to the case where no ambiguous variables are allowed, i.e. $\sigma = 0$. It is possible to apply our shuffled multi-partition pool generation method to obtain a σ-satisfactory set of pools P(L), where $L < (\tau \times comp)$. Given σ, the problem is then to determine the smallest value for L such that P(L) is σ-satisfactory. This problem can be addressed by performing simulations. In practice, if σ is chosen small enough, there are few ambiguous variables, and they can either be examined individually or discarded as probable negatives. However, simulations that we have performed suggest that tolerating even very small numbers of ambiguous variables can appreciably reduce the number of required pools.

Double pooling: The second extension that we propose, which we call "double pooling", stems from the following observation. In the case of protein-protein interactions, the interaction between a bait protein b and a prey protein p can be considered as a Boolean variable A(b,p). However, the pooling systems as defined above cannot be used directly, except for finding prey proteins that interact with a single given bait protein. Indeed, in the context of the 2-hybrid system, the value of a pool P of variables can only be evaluated if it is of the form $\{A(i,j) \mid (i,j) \in I \times J\}$, where I is a set of baits and J is a set of preys.

For example, the pool {A(1,2),A(3,4)} cannot be tested by 2-hybrid, by nature of the 2-hybrid system. Indeed, if baits 1 and 3 and preys 2 and 4 are mixed together, the test will be positive if at least one of the pairs (1,2), (3,4), (1,4), (3,2) interacts. Therefore this test yields the value of the pool of Boolean variables {A(1,2),A(3,4),A(1,4),A(3,2)}.

The double pooling strategy that we propose for dealing with this issue is the following. Build pools of baits as described in section 2, replacing K with K_b, the total number of baits under scrutiny, and L with the parameter L_b (number of layers of bait pools). Similarly, build pools of preys, replacing K with K_p (number of preys) and using another parameter L_p instead of L. Finally, pair each pool of baits with each pool of preys. The result is a set of pools that can each be tested in a single 2-hybrid elementary test. Selecting good values for L_b and L_p can again be addressed by performing simulations.

Discussion

We have identified one "strict" pooling problem and two extensions, namely: σ-satisfactory sets of pools, and double pooling. We have proposed a method for building pools, namely the shuffled multi-partition pool generation procedure. We have shown that this procedure satisfies the strict pooling problem, and considerably reduces the number of tests. We have also described how this procedure can be applied to generate σ-satisfactory sets of pools. Furthermore, we have described how it can be extended to tackle the double pooling problem.

To validate the multi-partition pool generation procedure and identify the best values for the procedure parameters L (single bait), L_b and L_p, we performed computer simulations. Most simulations were conducted in the context of double pooling and σ-satisfactoriness using $\sigma = 10^{-4}$, i.e. less than 0.01% of all examined pairs of proteins could be ambiguous. We used various experimental settings by using different values for K_b and K_p. The gains of the generated sets of pools, which we define as the ratio between the total number of pairs of proteins examined and the number of pools tested, varied between 70 in 90. For example, when searching for interactions between 1369 bait proteins and 1369 fish proteins, the gain is 85.6 when using the optimal values for L_b and L_p, which are 3 and 3. Using these values, on average out of 100 simulations, 933.1 of the 933.9 simulated interactions are identified, and 88 pairs of proteins are ambiguous. Note that almost every ambiguous pair is actually a negative pair, which suggests that individually retesting ambiguous variables might not be worth the effort, at least in this experimental setting. In this example, 1.87 million pairs of proteins can be examined by testing just 21904 pools, i.e. 229 96-well plates.

Experimental validation in a pilot project and further refinement of the method are planned. If current results are confirmed and technical wet lab difficulties are resolved, a pooling strategy such as the one described here could greatly increase the throughput of the *C. elegans* interactome project.

References

[1] Uetz et al (2000). *Nature*, 403, 623-627.
[2] Ito T et al (2000). *PNAS* 97(3), 1143-1147.
[3] Ito T et al (2001). *PNAS* 98(8), 4569-4574.
[4] Ho Y et al (2002). *Nature* 415(6868), 180-3.
[5] Gavin A-C et al (2002). *Nature* 415 (6868), 141-7.
[6] Walhout A et al (2000). *Science* 287, 116-22.
[7] Davy A et al (2001). *EMBO Reports* 2(9): 821-8.
[8] Walhout A, Vidal M (2001). *Methods* 24(3), 297-306.
[9] Reboul J et al (2001). *Nature Genetics* 27(3), 332-6.
[10] Reboul J et al (2003). *Nature Genetics* 34, 35-41.
[11] Walhout A et al (2002). *Current Biology* 12, 1952-8.
[12] Balding D and Torney D (1997). *Fungal Genetics and Biology* 21, 302-7.
[13] Ngo H and Du D (2000). In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Amer. Math. Soc.
[14] Cai WW et al (2001). *Genome Research* 11(10):1619-23.