Sensitivity Analysis of Protein Role Prediction Methods: Which are the Relevant Data?

Liliana López-Kleine[1,*], Alain Trubuil[2] and Véronique Monnet[3]

[1]*Statistic Department at the Universidad Nacional de Colombia*

[2]*INRA, UR341 MIA, F-78350 Jouy-en-Josas, France*

[3]*INRA, UMR1319 Micalis, F-78350 Jouy-en-Josas, France*

Abstract: Genome sequencing has allowed the generation of genomic and high-throughput post-genomic data. The availability of huge amounts of this data has, in turn, led to the development of protein role inference methods. Some of these methods allow the use of heterogeneous data of varying quality which are more or less informative. However, only limited research has been devoted to finding relevant data in terms of the inference of protein roles. In this study, we identified relevant subsets of data for the prediction of protein roles within the framework of a kernel method (KCCA) used to predict the role of a bacterial protein. We carried out a sensitivity analysis based on a fractional factorial design in order to study the influence of different microarray experiments, as well as of bacterial orders (groups of families) used to construct the phylogenetic profiles, on the prediction of a protein role. The results of this analysis showed to be useful for interpreting biological predictions highlighting specific data that should be investigated. The method is not restricted to KCCA, nor to the organism or to the data we used here.

Keywords: Protein prediction algorithms, data quality, sensitivity analysis.

1. INTRODUCTION

The sequencing of more and more genomes and their annotation has produced a huge amount of genomic data. On the other hand it has identified several proteins whose roles are unknown. Despite the availability of modern statistical and computational tools, assigning a role to a protein remains a difficult task. Several methods have been proposed in order to infer protein roles [1-3]. These methods consist in the determination of relationships between proteins and are thus used to reconstruct protein networks. They use genomic and post-genomic data to infer relationships between proteins. Most of the published methods have been validated by the reconstruction of a known network in a model organism. Nevertheless, predictions can be made for all proteins (with known and unknown functions) of the organism and, consequently, completely novel roles can be predicted using an inference method.

In a previous study we applied a kernel-based protein role prediction method proposed by [3] to predict the role of the peptidase PepF of lacococci, and validated several predictions experimentally. We used all, at that moment, available genomic and post-genomic data of this organism to predict the role of proteins from the lactic acid bacterium *Lactococcus lactis* IL1403 by obtaining distances between all proteins [4]. For these predictions we used Kernel Canonical Correlation Analysis (KCCA) [3], a supervised inference method that allows the integration of different data types. We integrated protein metabolic network data, phylogenetic profiles and microarray data to carry out the KCCA. Through this analysis we predicted which lactococcal proteins were close to PepF and several predicted partners were validated experimentally. By carrying out these two steps (Prediction and Validation), we validated its implication in protein secretion, pyruvate metabolism and peptidoglycan synthesis. Since most of the predictions were confirmed, we concluded that KCCA is a valuable tool for the prediction of protein roles [4].

As in most studies on protein prediction [1-4] we did not look for the relevance of the specific data used for the predictions. Nevertheless, highlighting relevant subsets of data would inform about quality and importance of data subsets on the predictions.

In order to determine relevant data (or data subsets) four steps need to be undertaken and are proposed in the present study:

Step 1: Apply an experimental design indicating how to sample original data in order to change input variables in a controlled manner.

Step 2: Conduct the protein role prediction (here KCCA) with all possible subsets chosen in step 1. The different runs will be called simulations in this article.

Step 3: Construct a response or output variable that summarizes the results of the protein role prediction algo-

*Address correspondence to this author at the Statistic Department at the Universidad Nacional de Colombia; Tel: 57 1 3165000/13213;
Fax: 57 1 3165000/13210; E-mail: llopezk@unal.edu.co

1875-693X/11

rithm. Here the predictions for PepF protein compared to the reference predictions (using all available data) that were experimentally validated.

Step 4: Construct a model to explain the response variable in function of the experimental matrix *X* in order to measure the importance of each data. The response variable is related to the function predictions for PepF (relationship with proteins of known functions) and will change when different data are used for the predictions.

We address the identification of relevant genomic data in protein role prediction using an approach based on sensitivity analysis. It consists in inferring protein role predictions by sampling the data space and the changes in role predictions of one protein. The approach we propose here is general and applicable for other protein prediction methods and for other proteins, as well as for other organisms for which genomic data are available. To illustrate the method we chose one particular protein which does not make part of a known metabolic network but for which predictions were validated experimentally in our previous work [4].

Sensitivity analysis focuses on how perturbations on the inputs *X* of a model *M* generate perturbations in the output *Y*. It makes it possible to decompose uncertainty in the output between the input variables (factors) when regression and correlation measures are computed on the output [5]. The sampling method used to generate the different simulations is the key to the sensitivity analysis. The best solution would be to test all the possible combinations of input variables and to analyze the output. Nevertheless, taking into account that the number of input variables can be very high, the number of possibilities to be tested increases exponentially and this type of screening becomes impossible. Sensitivity analysis methods based on Monte Carlo Sampling have been used, for example, for the investigation of complex models. The model in this case is treated as a black box and the distribution of inputs and outputs is analyzed on a global basis [6]. Other sampling methods can be used that lead to a better resolution and ensure the determination of factor importance. Herein we chose a fractional factorial design [7] to sample the factor space.

We restricted the present study to two multivariate datasets: the microarray experiments and the phylogenetic profiles. The protein metabolic network is the dataset used as reference for the KCCA. By varying the inputs of the two datasets through the sensitivity analysis, we were able to determine which data are more useful for the prediction of the role of PepF (Fig. **1**).

The results of the sensitivity analysis and the search for a simplified model allowed us: (i) to construct acceptable linear models with low residual error; (ii) to identify the most relevant data or factors in terms of importance in the model and in terms of predictability (i.e., to identify a subset of factors that allows us to obtain the same results as the reference result that were obtained using all variables), and; (iii) to better understand the role prediction we had obtained for the target protein, assessing the robustness of our predictions.

This article is organized as follows: we explain how we applied the four steps needed to determine relevant data subsets for the protein prediction algorithm used to determine partners of the lactococcal protein PepF. Then we explain a heuristic method we propose to choose the most important subsets from all possible proposed by the sensitivity analysis. Finally, we assess the robustness of our approach by investigating the effect of artificial data.

2. METHODS

Step1: Experimental Design

The reference used for the protein prediction algorithm was the protein metabolic network represented as an adjacency matrix of the pathways of *L.lactis* in KEGG (http://www.genome.jp/kegg/).

The first sampled data were the microarray data representing the expression of genes in 115 hybridizations belonging to 40 groups of conditions. The groups of conditions were constructed based on repetitions of the same hybridization and comparisons of the same mutants. The microarray experiments come from three different sources: (1) GEO on NCBI (http://www.ncbi.nlm.nih.gov/geo/), (2) Eric Guedon (INRA, Jouy-en-Josas), and (3) the European project Express Fingerprint directed by Pierre Renault (INRA, Jouy-en-Josas) and available at http://genome.jouy.inra.fr/efp/base/www.

The second data came from phylogenetic profiles constructed for 276 bacteria belonging to 80 different bacterial

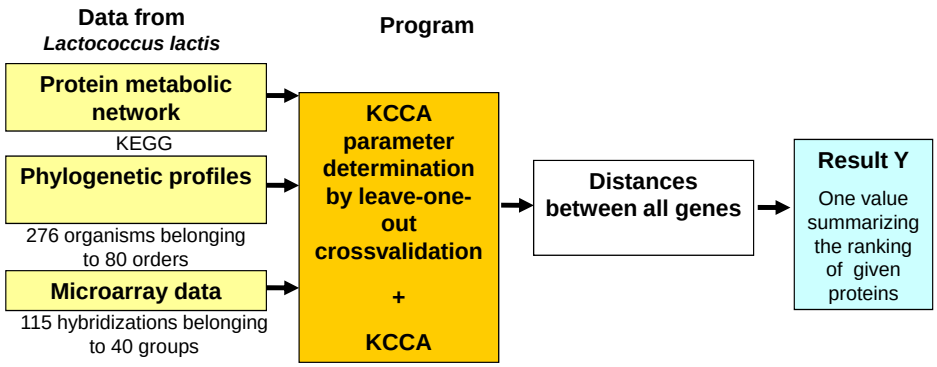

Fig. (1). Simplified representation of the KCCA method used for the prediction of protein roles in this study.

orders and the phylogenetic profiles were constructed by BLAST done against 276 bacteria using protein sequences of *L. lactis*. If the result of the BLAST search of a *L. lactis* protein against a protein of one of the 276 bacteria indicated an E-value below 10-5 the protein was declared present (1). In any other case the protein was declared absent.

We considered any of the microarray experiment conditions (40) or any of the bacterial orders (80) as variables or factors. (Fig. (**1**). In each simulation of the algorithm a factor (*p*) can take only two values (absence (-1), presence (1)). This means that some data are chosen or not for each step within the *p* microarray experiments (or within the organisms). The reference simulation is the result obtained when all factors (all microarray experiments and all organisms of the phylogenetic profiles) are present. This simulation predicts some proteins to be close to PepF that were validated experimentally and is therefore refered to as the reference result.

Several possibilities exist in order to choose the factors for each algorithm simulation as was already mentioned in the introduction. Taking into account that our interest was to control as much as possible the chosen factors and to explain the importance of each of them on the predictions, we decided to use a fractional factorial design [7]. This design allows the independent estimation of main and particular interaction effects between factors. A least square estimator of these coefficients is straightforward and given by $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, once the response variable *Y* is defined properly (step 3).

The fractional factorial design is very helpful as it solves the NP-hard problem proposing a reduced number of simulations of the algorithm to obtain changes in the results and still allowing to obtain importance of each factor on the result. Nevertheless, depending on the design matrix *X* the importance of some interaction factors will be confounded with others. This is illustrated in Fig. (**2**) with a resolution-4 design matrix, *X*, for which main effects are not confounded with two-factor interactions but some two-factor interactions are confounded with others. Four binary factors *x1*,…, *x4* and eight experiments or simulations are considered. The column corresponding to x4 is the product of columns x1, x2 and x3. So, two-factor interaction effect of x1 and x4 are confounded (or aliased) with the interaction effect of x2 and x3.

We used the FACTEX procedure in SAS (http://www.sas.com) to construct the design matrix *X*. This procedure builds orthogonal factorial experimental design matrices and indicates the groups of aliased two-factor interactions (the aliasing structure).

Step 2: The Inference Algorithm

Any inference algorithm that delivers ranked predictions (relationships to proteins of known function in order of importance) could be used. Nevertheless, our entire validation of the framework and results presented in Section 3 was obtained with the data and the algorithm illustrated in Fig. (**1**) and briefly detailed below.

This algorithm is based on KCCA [3], an extension of Canonical Correlation Analysis (CCA) but based on kernels. Considering two sets of variables associated with a collection of objects (genes, in this case), CCA consists of looking for low-dimensional referentials in which object projections are similar or correlated. This data analysis method has been extended to data objects represented by kernels (one for each type of data) or similarity matrices (see [3, 4, 8]). The algorithm outputs distances between the objects (genes). We thus assume that, the smaller the distance between two genes, the higher the confidence that the two genes could participate in a common function. In [4] we show that this turned out to be a correct assumption as most of the proteins predicted to be close do PepF participated in common functions. This was shown by experimental validations we conducted. A set of parameters is associated with the kernels and the KCCA method. Leave-one-out cross-validation was used to estimate those parameters and makes also part of the entire algorithm.

$$\mathbf{X} = \begin{array}{ccccc} \mathbf{x1} & \mathbf{x2} & \mathbf{x3} & \mathbf{x4} & \\ 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 \end{array}$$

Fig. (2). Design matrix X for 4 factors.

Step 3: Construction of a Response Variable Y

To evaluate the different results of each simulation a response variable *Y* was calculated on the basis of the ranking of six proteins SipL, SecA, FtsA, MurB, Glk and AldB. These proteins were chosen because they represent three of the cellular functions in which it was predicted and experimentally validated that PepF was involved: protein secretion (SipL and SecA), peptidoglycan synthesis (MurB) and pyruvate metabolism (Glk and AldB). We also included one predicted but not experimentally validated function: cell division (FtsA) [4]. Using the complete set of microarray data and phylogenetic profiles (reference KCCA), the ranking of these proteins with respect to PepF were: 1 for SipL, 24 for SecA, 8 for FtsA, 59 for MurB, 9 for Glk, 62 for AldB [4]. On the basis of these reference rankings, we built a response variable *Y* whose value is zero for these reference rankings. So, *Y=0* is the reference value assumed to indicate a correct ranking of the six chosen partners of PepF. If the ranking in these 6 proteins changes, the response variable will augment. If any other combination of data with less factors conducts to same protein ranking the output of this result will be *Y=0*, indicating that the same predictions could be obtained with less data. For each simulation, *Y* was calcu-

lated as a measure of change in the ranking of these six proteins as follows: $Y = 1/2 \sum_i (i = 1)^N [w_i \, erfc((S_i - A_i))/\sqrt{2}\sigma)]$, where σ plays the role of a smoothing parameter, $N=6$ is the number of proteins of interest, A_i, $i=1,\ldots,N$ is the absolute value of the differences in ranking between reference and simulations for the N genes, S_i, $i=1,\ldots,N$ are the tolerance thresholds with respect to ranking and w_i, $i=1,\ldots,N$ are weights chosen a priori on the basis of the confidence of the predictions for each predicted partner. The value Y can take only positive values. The complementary error function *erfc* is used to smooth the response function around the tolerance value S_i.

At this point it is worth mentioning that more than one protein role predictions could be addressed in parallel, using the same design matrix but increasing the dimension of Y.

Step 4: Analysis of the Output Modelling Y in Function of X

The sensitivity analysis itself is the analysis of the different Y values obtained by the simulations done based on matrix X. Due to the orthogonality of X and the resolution-4, a handful of models corresponding to combinations of main effects and interactions groups can be estimated independently by computing $\hat{\beta} = \frac{1}{n} X^T Y$. In that way, 40 main effects and 62 interactions were calculated for the microarray data. For the phylogenetic profiles, 80 main effects and 175 interactions were calculated.

We also used the results of the simulations to determine which of the investigated protein partners changed their ranking in each of the simulations. This result provides a stronger prediction for the partners that change less, indicating if several groups of variables predict the same result.

Step 5: Model Selection and Construction of Relevant Subsets

We propose a heuristic approach that was straightforward in our case. Our aim was the selection of models with good predictive properties, especially in the neighbourhood of $Y=0$. Moreover, we also looked for relevant subsets of factors of minimal size. This means subsets with as less microarray experiments as possible and phylogenetic profiles with as less organisms as possible. With respect to these two goals, our strategy consisted first in choosing a parsimonious model that fitted well the response variable based on the factor weights in the linear model, and then verifying that when those factors are used, the obtained value was indeed close to $Y=0$. In practice, we ranked the coefficients β according to their weight in the linear model and considered models with an increasing number of terms or coefficients. We then computed the standardized error and adjusted R^2. Models that provided an acceptable R^2 (at least 0.7) were selected for further investigation. The next stage of our procedure consisted in trying to reduce the number of factors involved in these models through a further evaluation and identification of subsets of factors of minimal size. To achieve this, we used the interaction structure for selecting models with a minimal number of factors. This selection turns out to be an NP-hard problem as a great number of combinations need to be tested. Therefore, we propose a heuristic approach of low complexity described in the following section, providing an approximate solution.

Heuristic Approach to Choose Minimal Subset of Important Factors

For one model that was previously chosen because of a good determination coefficient ($R^2>0.7$) we looked for factors belonging to main and interaction effects. One possibility would have been to propose all factors contained in the model (as main or as interaction effects) as important factors. Nevertheless, the number of factors can be very high because several interaction effects are confounded with others due to the way the sample matrix X was constructed. As our aim was to simplify the model and extract minimal data subsets (with as less factors as possible) we proceeded as follows:

(a) Selection of factors present in main and interaction effects.

(b) Association of different tags to the chose factors taking into account their importance (i.e. presence in main effects and more than one interaction factors counts more than present only in one interaction factor).

(c) Representation of the interaction structure in a matrix S of size $f \times f$ where f is the number of factors involved in main effects and interactions $S(i,i) = -i$.

Taking into account if:

- Factor x_i is involved in a main effect for the model $S(i,j) = m$
- (x_i, x_j) belongs to two-factor interaction aliased group G_m

(d) Sweeping of each line of matrix S and

1. Calculation of NI, the number of interactions covered considering x1 and factors already selected.
2. Calculation of NF, the number of extra factors needed to cover the interactions.

(e) Selection of the line with the highest NI/NF. If two lines have the same value, we take the one with the highest value for NI

(f) Tagging of factors again until then untagged factors involved in interactions considered on the selected line.

(g) Suppression of tags associated with covered interactions in matrix S.

(h) Iteration of this process until all two-factor interactions are covered using the originally selected factors.

To understand this algorithm, we proposed the following example. A selected model M is described by the following equation:

$$Y_p = \beta_0 + \beta_4 x_4 + \beta_6 x_6 + \beta_7 x_7 + \beta_{23}(x_2 x_3 = x_1 x_4 = x_5 x_7) + \beta_{36}(x_3 x_6 = x_4 x_5)$$

This model has two interaction terms. The equation takes the fact into account that two-factor interactions are confounded in both terms β23 and β36. We represented M in the matrix S where the main factors had negative tags (-1, -2 and -3) and interaction terms received positive tags. At this point we knew that our final model would contain the main factors. We then explored the other factors. The aim was to have a member of each interaction group. For each line of

matrix *S*, we calculated NI and NF. The highest value of NI/NF was 2, and so we selected line 5. We therefore retained factor 5, in addition to 4, 6 and 7 which came from the main effects. In this example, there is only one iteration because all the interactions are covered considering factors x4, x5, x6 and x7 (see Fig. **3** for example illustration).

Once the model and the factors were selected, a second selection of factors on the basis of the output value Y_P predicted by the model was made. This was done by using all the selected factors that conduced to a low Y_P. These subsets of factors were then considered as an input to the algorithm and the response variable was computed. Either the response variable was significantly non zero and the potential minimal subset was rejected, or the predictions were close to the reference predictions (*Y=0)* and the minimal relevant subset was retained. The procedure does not assure that, when repeated with another matrix *X*, the same subset is found. Nevertheless, it is assured that for a given matrix *X*, the factor subset is minimal and gives prediction results as close as possible to the reference (*Y=0).*

Robustness of the Minimal Subset

Once a relevant subset was retained, it is useful to know if the addition of factors (microarray experiments or organisms in our case), changes the prediction results. To investigate robustness in this sense, we added 3, 6 and 9 factors at random to the minimal subset 15 times and repeated this procedure five times, which gave us 225 simulations to test.

Construction and Analysis of Artificial Data

In order to assess the repeatability and robustness of the proposed method, we decided to include 10, 20 and 50 percent of artificial microarray experiments in our data and to apply the whole algorithm. If the algorithm is reliable, none of these artificial experiments should be selected to make part of the relevant subset. The artificial microarray experiments were constructed taking one of the measured expression values at random for each gene as explained below.

First, we independently permuted the existing matrix of 115 microarray experiments along the lines (genes). Second, we added the 4, 8 or 20 first columns to the already existing matrix of microarray experiments and created 4, 8 or 20 new hybridization groups. In other words, we added new hybridization groups composed of one artificial experiment to the existing 40 hybridization groups. We then ran the whole algorithm to determine the relevant subsets.

3. RESULTS

The framework was applied to identify which subsets of data seemed to be relevant for PepF's role predictions. We calculated the response value *Y* for each simulation (Fig. **4**) proposed by the factorial experimental design matrix X. *Y=0* is the reference value for the last simulation in which the complete dataset was used. One simulation corresponds to one line of *X*. None of the other simulations (leaving out the factors indicated by *-1* in the matrix *X*) lead to *Y=0*. We obtained models with a low global residual error (0.0116 for the microarray data and 0.0078 for the phylogenetic profiles). Interactions seem important, as weights for them are different from zero. The residual errors for a model that does not consider them were higher (0.1083) but still acceptable in comparison with the residual error of a model including interactions (0.0938). Several two-factor interactions were of the same order as main effects for both the model on microarray data and the model on phylogenetic profiles. The three most important main factors of the linear models constructed for both types of data are shown in the left-hand column of Tables **1** and **2**.

Search of a Simplified Model with a Heuristic Approach

Ranking of model coefficients and selecting factors while considering the highest coefficients, even with the complete aliasing structure, provided very poor results when used for prediction (e.g. output *Y*-value of 0.7073 with 33 of 44 factors of the microarray data!). As we were interested in minimizing *Y* we used the heuristic approach mentioned before to find a minimal subset of data. We obtained subsets of factors predicted to give the lowest possible output of *Y*. We tested the first 10 subsets retained after the selection procedure for seven linear models of 10, 15, 20, 25, 30, 35 and 40 terms or coefficients to see if the predicted output of *Y* was in fact low. The 7×10 values of Y are plotted in (Fig. **5**, left for microarray data and right for phylogenetic profiles). These figures show that several subsets obtained from models with 30 coefficients give a result of *Y=0* for the microarray data (20 coefficients for the phylogenetic profiles).

	1	2	3	4	5	6	7
1				1			
2			1				
3		1				2	
4	1			-1	2		
5				2			1
6						-2	
7					1		-3

	1	2	3	4	5	6	7	NI/NF
1				1				(1,1),1
2			1					(1,2), 0.5
3		1				2		(2,2),1
4	1			-1	2			(2,2),1
5				2			1	(2,1),2
6						-2		(1,1),1
7					1		-3	(1,1),1

Fig. (3). Construction of matrix S (left) and selection of factors (right). The factors in red are initially selected because they belong to main effects (tagged with -), the factors in blue are selected in a second step because they cover interaction effects.

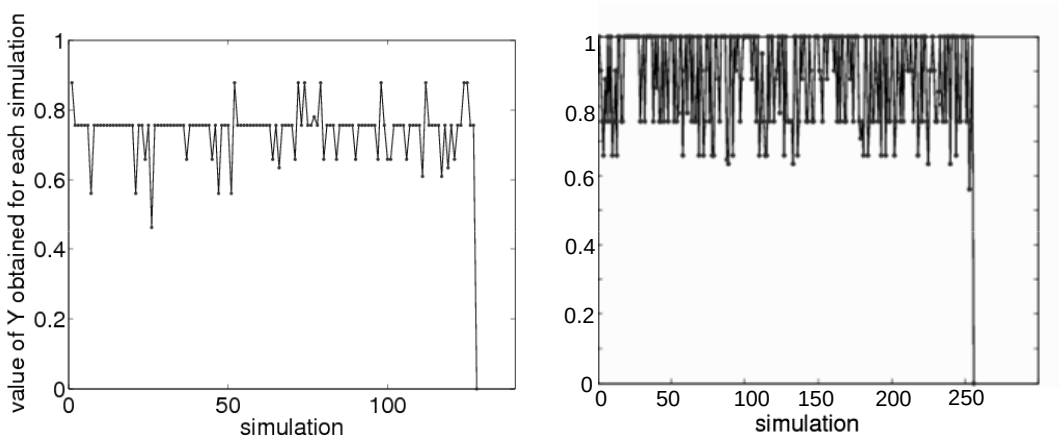

Fig. (4). Output Y of the simulations for microarray data (left) and phylogenetic profiles (right) indicated by points. Lines between points are added for visibility.

Table 1. Microarray Data: Columns on the Left Show Factor Numbers and Experiments Selected by the Linear Model (LM). Coloumns on the Right Show Factor Numbers and Experiments after Relevant Subset Identification

Factor (LM)	Experiments	Relevant Factors	Experiments
25	Sup pyrF	25	Sup pyrF
22	Natural strain	3	codY
36	Pip mutant	7	Pip mutant

Keeping only the factors that contributed to an output of *Y=0* allowed us to reduce the number of factors to 18 for the microarrays and 15 for the phylogenetic profiles.The first column of Table **4** shows the 18 factors retained for the microarray experiments. The solution is not unique but several factors are present in the majority of the subsets. Three of the factors (microarray experiments or orders of organisms) present in most of the minimal subsets are listed in Table **1** for the microarray data and Table **2** for the phylogenetic profiles. It shoud be observed that both types of data behave differently since more factors are necessary for the microarray data to obtain an output value near to *Y = 0* than for the phylogenetic profiles.

Table 2. Phylogenetic Profiles: Factors in Order of Importance (Ordered Coefficients of the Linear Model (LM)) and three Factors Making Part of the Relevant Subsets (not Ordered). E: Enterobacteriales, H: Halobacteriales, L: Lactobacillales, S: Sphingomonadales

Factor (LM)	Order	Factors Subset	Order
16	E	16	E
41	H	41	H
31	L	62	S

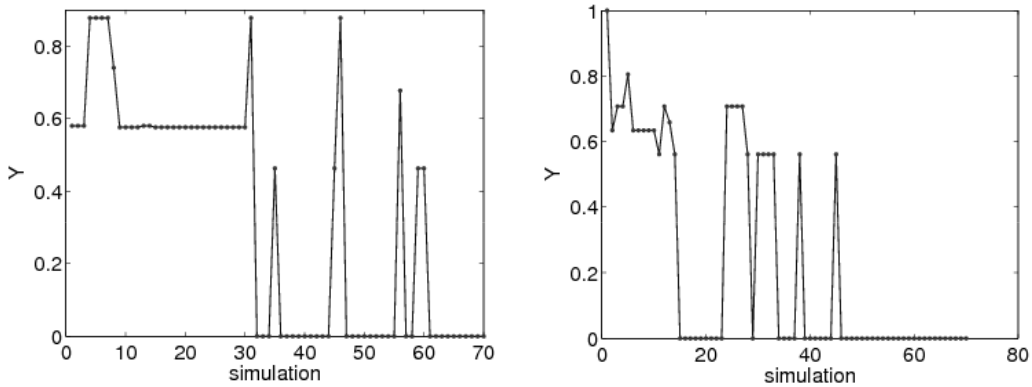

Fig. (5). Output of Y of the simulations with the subsets obtained using the heuristic approach for microarray data (left) and phylogenetic profiles (right) indicated by the points. Lines between points are added for visibility.

Table 3. Number of Original Simulations (Planned by *X*) in Which the Ranking of the Protein Changed (Sim) and Ranking Changes in the Tolerance test (T). MD: Microarray Data, PP: Phylogenetic Profiles. Column S_i Indicates Tolerance with Respect to Ranking

	SimMD	SimPP	S_i	T225MD	T225PP
	128 Total Simulations	256 Total Simulations			
SipL	47	68	5	0	7
SecA	63	138	10	4	16
FtsA	120	240	20	8	20
MurB	62	184	5	0	11
AldB	64	190	15	4	12
Glk	68	97	10	10	29

Robustness of the Relationships of PepF

For the six predicted partner proteins of PepF we chose to work on here (Rank 1 for SipL, rank 24 for SecA, rank 8 for FtsA, rank 59 for MurB, rank 9 for Glk, rank 62 for AldB in the reference prediction), we chose to evaluate their ranking across the different simulations to determine the robustness of the predictions. The number of simulations in which the ranking changes are greater than the tolerance thresholds S_i are shown in the first two columns of Table **3**. These results indicate that some predictions (i.e., SipL) are verified in more simulations than others (i.e., FtsA).

Perturbation of the Relevant Subsets of Factors

We added 3, 6 and 9 microarray experiments chosen at random from the complete set of experiments to the 15 simulations identified to give $Y=0$. This was done five times for each factor, giving us 75=15×5 new simulations to test. The ranking change of the six partners of PepF we investigated here, are shown in the last two columns of Table **3**. They show that the addition of new experiments does not destabilize the results and does therefore not add noise to the predictions.

Artificial Experiments do not Make Part of the Relevant Subsets

We determined the relevant subsets using simulated experiments added to the real data. In the case of 10% of simulated data, we found exactly the same relevant subsets as when using only the original data (Table **4**). This subset changed when we used 20% but did not include the simulated subsets. In the case in which 50% was used, we found one of the simulated experiments, number 51, between the relevant subsets (Table **4**).

4. DISCUSSION

The identification of relevant data subsets for protein role prediction raises many challenges in terms of (i) the complexity of the methods and high dimensions of the data used in prediction methods, (ii) the identification of relevant subsets, (iii) the robustness and reliability of the subset identification, and (iv) the interpretation of the relevant data pieces.

Table 4. Comparison of the Relevant Subsets Determined Using the Original Microarray Dataset (Original) and with 10 Percent (10simul), 20 Percent (20simul) or 50 Percent (50simul) Simulated Microarray Experiments

original	10simul	20simul	50simul
2	2	2	2
3	3	3	3
5	5	5	5
7	7	7	7
8	8	9	9
11	11	11	11
12	12	12	12
16	16	13	13
21	21	16	16
24	24	21	21
25	25	25	23
27	27	27	25
29	29	29	27
30	30	30	29
32	32	32	30
35	35	35	32
37	37	37	35
40	40	38	36
-	-	40	37
-	-	-	38
-	-	-	51

Simplification is Necessary

The analyzed model is very complex because many variables and parameters are brought into play and the output is a strong simplification of the results. The method we

propose addresses this difficulty. First of all, we simplified the process by investigating only the response variable *Y* concerning one protein, PepF, and not the whole distance matrix obtained by KCCA. This could be done, for example, investigating the determinant of this matrix. Nevertheless, the overall changes are not as strong as to highlight differences through the simulations. For this reason we propose taking a close look at one protein and their partners we do know well.

We simplified obtained models in a second step by taking the ranking of only some proteins in this output vector into account. A further simplification is done by the construction of one scalar that represents the ranking of the partners of PepF in the output vector with respect to the reference ranking. The decision to do this ranking, i.e., the tolerance to declare a ranking change as being important and the weight we give to the change, is arbitrary and, if chosen in a different manner, can change the obtained model. Moreover, the results we obtained are obviously dependent on the target protein. This means that the important and minimal subset obtained for another protein could be different (other microarray experiments and other organisms).

Taking this simplification into account, it is not surprising that the models obtained are very complex and contain many interactions, and that it is very difficult to assign one or two main factors that explain the results and to discard the rest. Since microarray experiments are done under different conditions that affect only a small group of genes, it is not surprising to find a high importance of interactions.

Identification of Relevant Subsets

The algorithm we present for this task is one of many possible solutions. It was useful since it allowed us to identify subsets that gave the same results as if the whole data set was used. Nevertheless, in more complex studies, this may not be the case and more sophisticated model selection methods might be necessary. The subsets of factors turned out to be more easily reduced for phylogenetic profiles than for the microarray data. This is not surprising since microarray experiments are more heterogeneous than the co-evolution of proteins among the different orders of bacteria.

Robustness and Reliability of the Method

The fact that the chosen partners for PepF did not drastically change their ranking in most of the simulations reinforces the obtained predictions. It is not surprising that the protein that changes its ranking most of the time is FtsA, implicated in cell division and for which the predicted relationship was not experimentally confirmed. On the other hand, the proteins that only change their ranking in few simulations are proteins implicated in secretion, which was experimentally suggested to be the main function of PepF [4]. It was also interesting to observe that the stability of the prediction is not related to ranking in the reference prediction, since SipL, FtsA and Glk, all having ranks in the first 10 positions, behave differently.

Two additional tests indicate that the proposed method is robust and of general applicability. First, we confirmed the robustness of the model by the fact that adding experiments to a group of experiments that gives a low value of *Y*, near the reference, does not affect the prediction. This means that a subgroup of experiments is sufficient to obtain the results, but that the use of more experiments does not change the predictions. Second, we confirmed the general applicability of the method since it is still able to reveal relevant subsets even when artificial data is added to the real data. This means that the subsets found by the proposed method are not an artefact but real relevant subsets.

Interpretation of the Relevant Pieces of Data

Regarding main effects without interactions in the case of microarray data, we found that factors representing experiments with highest coefficients mainly belong to the European project, Express Fingerprints. The most relevant experiment group turned out to be the comparison of a pyrF mutant to a genetically modified organism (GMO) that contains a suppressor of this mutation on a plasmid [9]. For this and the other relevant microarray experiments, it is very difficult to interpret their meaning in the relation to prediction of PepF function. Since we are analyzing a pleiotropic enzyme, it is not surprising to find that many different conditions are needed to explain its relationships with proteins belonging to different experiments. Nevertheless, knowing which experiments are relevant can help to highlight interesting experiments in other cases. If only one partner, and thus one function is investigated, the model becomes simpler and smaller subsets of factors are necessary to obtain reference predictions. To analyze a group of partners one by one can also help to determine which factor is responsible for each prediction. We know, for example, that the microarray data making up part of factors 25 (sup pyrF) and 22 (natural strain) is responsible for the prediction of SipL as a partner of PepF. In the case of the phylogenetic profiles, the most important orders belong, in fact, to the evolutionary more similar orders of the Lactobacillales. This would indicate that the use of evolutionary far organisms to construct phylogenetic profiles is not very informative. Nevertheless, the first order is not the order of *L. lactis* but of a close order, Enterococcales.

5. CONCLUSIONS

In this paper, we present an innovative and general method for the identification of minimal data subsets for protein role prediction algorithms. The prediction method can be complex but the approach is simple since it is based on sensitivity analysis. We found it useful to determine a minimal subset of data that still allows us to obtain the reference predictions, both simplifying the model and the interpretation of the results. Furthermore, we demonstrated that the obtained subsets are stable and that the method is robust even if simulated data is included. In our case, the determination of subsets took place after the experimental validation of predicted results. Nevertheless, this analysis could have been useful at the time biological hypotheses were proposed to guide the wet lab experimentations. Moreover, this approach is applicable for any organism for which genomic data is available and can be conducted on any protein of interest, for which reference predictions exist.

ACKNOWLEDGMENTS

The authors are grateful to Jean-Pierre Gauchi and Hervé Monod of INRA, UR 341 for the construction of the simulation matrix and helpful discussions on sensitivity analysis. Funding: INRA (French National Institute for Agricultural Research).

REFERENCES

[1] Friedman N, Linial M, Nachman L, Peer D. Using Bayesian networks to analyze expression data. J Comput Biol 2000; 7: 601-620.

[2] Werhli A, Husmeier D. Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. J Bioinform Comput Biol 2007; 6: 543-572.

[3] Yamanishi Y, Vert JP, Nakaya A, Kaneisha M. Extraction of correlated clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics 2003; 19: (Suppl. 1), i323-i330.

[4] Lopez Kleine L, Monnet V, Pechoux C, Trubuil A. Role of the bacterial peptidase F inferred by statistical analysis and further experimental validation. HFSP J 2008; 2: 29-41.

[5] Saltelli A, Chan K, Scott EM, Sensitivity Analysis: Wiley, 2000.

[6] Helton JC, Breeding RJ. Calculation of reactor accident safety goals. Reliab Eng Syst Saf 1993; 39: 129-158.

[7] Box GEP, Hunter WG, and Hunter JS. Statistics for Experimenters, an Introduction to Design, Data Analysis and Model Building: Wiley 1978.

[8] Vert JP, Yamanishi Y. Supervised graph inference. In: Lawarnce KS, Yasir W, Leon B, Eds. Advances in Neural Information Proven System. MIT Press 2005; Vol. 17: pp. 1443-40.

[9] Ueda K, Yamamoto Y, Ogawa K. Bacterial SsrA system plays a role in coping with unwanted translational readthrough caused by suppressor tRNAs. Genes Cells 2006; 7: 509-519.

Received: December 11, 2010 Revised: March 02, 2011 Accepted: March 07, 2011