Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study

Carito Guziolowski[1] & Philippe Veber & Michel Le Borgne[1] & Ovidiu Radulescu[1,2] & Anne Siegel[1]

[1]IRISA (Inria, UMR CNRS, Université de Rennes 1), Campus de Beaulieu, 35042 Rennes Cedex
[2]IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex

Abstract

We proposed in previous articles a qualitative approach to check the compatibility between a model of interactions and gene expression data. The purpose of the present work is to validate this methodology on a real-size setting. We study the response of *E.coli* regulatory network to nutritional stress, and compare it to publicly available DNA microarray experiments. We show how the incompatibilities we found reveal missing interactions in the network, as well as observations in contradiction with available literature.

1 Introduction

There exists a wide range of techniques for the analysis of gene expression data. Following a review by Slonim [**?**], we may classify them according to the particular output they compute: 1. list of significantly over/under-expressed genes under a particular condition, 2. dimension reduction of expression profiles for visualization, 3. clustering of co-expressed genes, 4. classification algorithms for protein function, tissue categorization, disease outcome, 5. inferred regulatory networks.

The last category may be extended to all model-based approaches, where experimental measurements are used to build, verify or refine a model of the system under study.
Following this line of research, we showed in previous papers (see [**?**], [**?**] and [**?**]) how to define and to check consistency between experimental measurements and a graphical regulatory model formalized as an interaction graph. The purpose of the present work is to validate this methodology on a real-size setting. More precisely, we show 1. that the algorithms we proposed in [**?**] are able to handle models with thousands of genes and reactions, 2. that our methodology is an effective strategy to extract biologically relevant information from gene expression data.
For this we built an interaction graph for the regulatory network of *E. coli* K12, mainly relying on the highly accurate database RegulonDB [**?**], [**?**]. Then we compared the predictions of our model with three independant microarray experiments. Incompatibilities between experimental data and our model revealed:

- either expression data that is not consistent with results showed in literature – *i.e.* there is at least one publication which contradicts the experimental measurement,
- either missing interactions in the model

We are not the first to address this issue. Actually, in the work of Gutierrez-Rios and co-workers [**?**], an evaluation of the consistency between literature and microarray experiments of *E. coli* K12 was presented. The authors designed on-purpose microarray experiments in order to measure gene expression profiles of the bacteria under different conditions. They evaluate the consistency of their experimental results first with those reported in the literature, second with a rule-based formalism they propose. Our main contribution is the use of algorithmic tools that allow inference/prediction of gene expression of a big percentage of the network, and diagnosis in the case of inconsistency between a model and expression data.

2 Mathematical framework

2.1 Introductory example

We choose as an illustration a model for the lactose metabolism in the bacterium E.Coli (lactose operon). The interaction graph corresponding to the model is presented in Fig.1. This is a common representation for biochemical systems where arrows show activation or inhibition. Basically, an arrow between A and B means that an increase of A tends to increase or decrease B depending on the shape of the arrow head. Common sense and simple biological intuition can be used to say that an increase of allolactose (node A on Figure 1) should result in a decrease of $LacI$ protein. However, if both $LacI$ and $cAMP-CRP$ increase, then nothing can be said about the variation of $LacY$.

The aim of this section is first, to provide a formal interpretation for the graphical notation used in Figure 1; second, to derive constraints on experimental measurements, which justify our small scale common sense reasoning; finally apply these constraints to the scale of data produced by high throughput experimental techniques. For this, we resort to qualitative modeling ([**?**]), which may be seen as a principled way to derive a discrete system from a continuous one.

2.2 Equilibrium shift of a differential system

Let us consider a network of n interacting cellular constituents (mRNA, protein, metabolite). We denote by X_i the concentration of the i^{th} species, and by $\mathbf{X}$ the vector of concentrations (whose components are X_i). We assume that the system can be adequately described by a system of differential equations of the form $\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X},\mathbf{P})$, where $\mathbf{P}$ denotes a set of control parameters (inputs to the system). A *steady state* of the system is a solution of the system of equations $\mathbf{F}(\mathbf{X},\mathbf{P}) = \mathbf{0}$ for fixed $\mathbf{P}$.

A typical experiment consists in applying a perturbation (change $\mathbf{P}$) to the system in a given initial steady state condition $eq1$, wait long enough for a new steady state $eq2$, and record the changes of X_i. Thus, we shall interpret the sign of DNA chips differential data as the sign of the variations $X_i^{eq2} - X_i^{eq1}$.

The particular form of vector function $\mathbf{F}$ is unknown in general, but this will not be needed as we are interested only in the signs of the variations. Indeed, the only information we need about $\mathbf{F}$ is the sign of its partial derivatives $\frac{\partial F_i}{\partial X_j}$. We call *interaction graph* the graph whose nodes are the constituents $\{1,\ldots,n\}$, and where there is an edge $j \to i$ iff $\frac{\partial F_i}{\partial X_j} \neq 0$ (an arrow $j \to i$ means that the rate of production of i depends on X_j). As soon as $\mathbf{F}$ is non linear, $\frac{\partial F_i}{\partial X_j}$ may depend on the actual state $\mathbf{X}$. In the following, we will assume that the *sign* of $\frac{\partial F_i}{\partial X_j}$ is constant, that is, that the interaction graph is independent of the state. This rather strong hypothesis, can be replaced by a milder one specified in [**?**, **?**] meaning essentially that the sign of the interactions do not change on a path of intermediate states connecting the initial and the final steady states.

2.3 Qualitative constraints

In the following, we introduce an equation that relates the sign of variation of a species to that of its predecessors in the interaction graph. To state this result with full rigor, we need to introduce the following algebra on signs.

We call sign algebra the set $\{+,-,\mathbf{?}\}$ (where **?** stands for indeterminate), endowed with addition, multiplication and qualitative equality, defined as:

$$\begin{array}{cccccc} + + - = \mathbf{?} & + + + = + & - + - = - & + \times - = - & + \times + = + & - \times - = + \\ \mathbf{?} + - = \mathbf{?} & \mathbf{?} + + = \mathbf{?} & \mathbf{?} + \mathbf{?} = \mathbf{?} & \mathbf{?} \times - = \mathbf{?} & \mathbf{?} \times + = \mathbf{?} & \mathbf{?} \times \mathbf{?} = \mathbf{?} \end{array}$$

≈	+	−	**?**
+	T	F	T
−	F	T	T
?	T	T	T

Some particularities of this algebra deserve to be mentioned:

- the sum of **+** and **−** is indeterminate, as is the sum of anything with indeterminate,
- qualitative equality is reflexive, symmetric but not transitive, because **?** is qualitatively equal to anything; this last property is an obstacle against the application of classical elimination methods for solving linear systems.

To summarize, we consider experiments that can be modelled as an equilibrium shift of a differential system under a change of its control parameters. In this setting, DNA chips provide the sign of variation in concentration of many (but not necessarily all) species in the network. We consider the signs $s(X_i^{eq2} - X_i^{eq1})$ of the variation of some species i between the initial state X^{eq1} and the final state X^{eq2}. Both states are stationary and unknown.

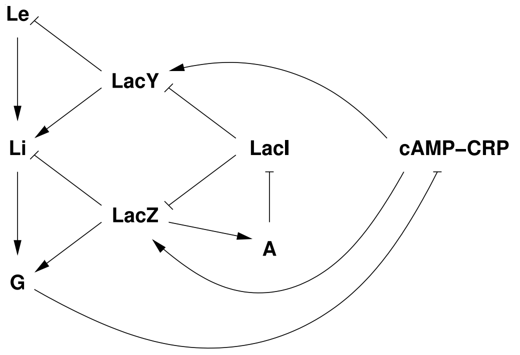
In [**?**], we proved that under some reasonable assumptions, in particular if the sign of $\frac{\partial F_i}{\partial X_j}$ is constant in states along a path connecting $eq1$ and $eq2$, then the following relation holds in sign algebra for all species i:

$$s(X_i^{eq2} - X_i^{eq1}) \approx \sum_{j \in pred(i)} s(\frac{\partial F_i}{\partial X_j}) s(X_j^{eq2} - X_j^{eq1}) \tag{1}$$

where $s : \mathbb{R} \rightarrow \{\mathbf{+}, \mathbf{-}\}$ is the sign function, and where $pred(i)$ stands for the set of predecessors of species i in the interaction graph. This relation is similar to a linearization of the system $\mathbf{F}(\mathbf{X}, \mathbf{P}) = \mathbf{0}$. Note however, that as we only consider signs and not quantities, this relation is valid even for large perturbations (see [**?**] for a complete proof).

2.4 Analyzing a network: a simple example

Let us now describe a practical use of these results. Given an interaction graph, say for instance the graph illustrated in Figure 1, we use Equation 1 at each node of the graph to build a qualitative system of constraints. The variables of this model are the signs of variation for each species. The qualitative system associated to our lactose operon model is proposed in the right side of Figure 1. In order to take into account observations, measured variables should be replaced by their sign values. A *solution* of the qualitative system is defined as a valuation of its variables, which does not contain any ”**?**” (otherwise, the constraints would have a trivial solution with all variables set to ”**?**”) and that, according to the qualitative equality algebra, will satisfy all qualitative constraints in the system. If the model is correct and if data is accurate, then the qualitative system must posses at least one solution.



$$\left\{ \begin{array}{rcll} LacI & \approx & -A & (1) \\ A & \approx & LacZ & (2) \\ LacZ & \approx & cAMP - LacI & (3) \\ Li & \approx & Le + LacY - LacZ & (4) \\ G & \approx & Li + LacZ & (5) \\ cAMP & \approx & -G & (6) \\ LacY & \approx & cAMP - LacI & (7) \end{array} \right.$$

Figure 1: Interaction graph for the lactose operon and its associated qualitative system. In the graph, arrows ending with ”>” or ”−|” imply that the initial product activates or represses the production of the product of arrival, respectively.

A first step then is to check the *self-consistency* of the graph, that is to find if the qualitative system without observations has at least one solution. *Checking consistency* between experimental measurements and an interaction graph boils down to instantiating the variables which are measured with their experimental value, and see if the resulting system still has a solution. If this is the case, then it is possible to determine if the model predicts some variations. Namely, it happens that a given variable has the same value in all solutions of the system. We call such variable a *hard component*. The values of the hard components are the predictions of the model.

Whenever the system has no solution, a simple strategy to *diagnose the problem* is to isolate a minimal set of inconsistent equations. In our experiments, a greedy approach was enough to solve all inconsistencies (see next section). Note that in our setting isolating a subset of the equations is equivalent to isolating a subgraph of the interaction graph. The combination of the diagnosis algorithm and a visualization tool is particularly useful for model refinement.

Finally, let us mention that we provided in [**?**] an efficient representation of qualitative systems, leading to effective algorithms, some of them could be used to get further insights into the model under study. We shall see in the next section, that these algorithms are able to deal with large scale networks.

3 Results

3.1 Construction of the Escherichia coli regulatory network

For building *E.coli* regulatory network we relied on the transcriptional regulation information provided by RegulonDB ([**?**], [**?**]) on March 2006. From the file containing *transcription factor to gene interactions* we have built the regulatory network of *E.coli* as a set of interactions of the form $A \rightarrow B\ sign$ where $sign$ denotes the value of the interaction: **+**, **−**, **?**(expressed, repressed, undetermined), and A and B can be considered as genes or proteins, depending on the following situations:

- The interaction $genA \rightarrow genB$ was created when both $genA$ and $genB$ are notified by RegulonDB, and when the protein A, synthesized by $genA$, is among the transcriptional factors that regulate $genB$. See Figure 2 A.
- The interaction $TF \rightarrow genB$ was created when we found TF as an heterodimer protein (protein-complex formed by the union of 2 proteins) that regulates $genB$. See Figure 2 B. In *E.coli* transcriptional network we have found 4 protein-complexes which are: IHF, HU, RcsB, and GatR.
- The interaction $genA \rightarrow TF$ was created when we found the transcriptional factor TF as an heterodimer protein and $genA$ synthesizes one of the proteins that form TF. See Figure 2 B.

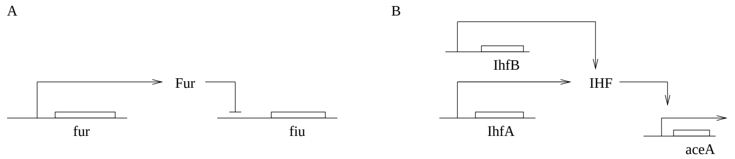


Figure 2: Representation of genetical interactions. **(A)** Negative regulation (repression) of gene fiu by the transcription factor Fur represented as $fur \rightarrow fiu\ -$. **(B)** Biological interaction of genes $ihfA$ and $ihfB$ forming the protein-complex IHF represented as $ihfA \rightarrow IHF\ +$ and $ihfB \rightarrow IHF\ +$, positive regulation of gene $aceA$ by the protein complex IHF represented by $IHF \rightarrow aceA\ +$

3.2 Adding sigma factors to obtain self-consistency

Using the methods and the algorithms described with detail in [**?**] we built a qualitative system of equations for the interaction graph obtained from *E.coli* network. For solving qualitative equations we have used our own tool, the PYTHON module PYQUALI. The system was not found to be self-consistent and we used a procedure available in PYQUALI library to isolate a minimal inconsistent subgraph (see Figure 3). A careful reading of the available literature led us to consider the regulations involving sigma factors which were initially absent from the network. Once added to complete the network, we obtained a network of 3883 interactions and 1529 components (genes, protein-complexes, and sigma-factors). This final network (global network) was found to be self-consistent.

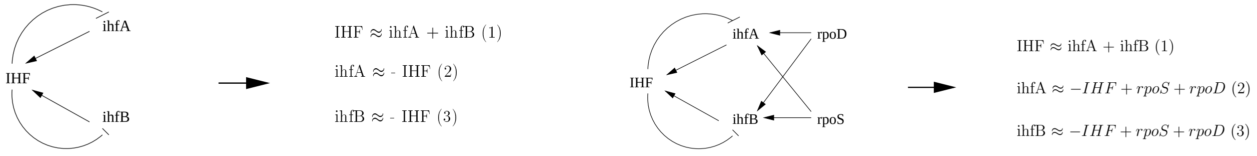


Figure 3: (Left) A minimal inconsistent subgraph, isolated from the whole *E.coli* regulatory network using PYQUALI. (Right) Correction proposed after careful reading of available literature on ihfA and ihfB regulation.

3.3 Compatibility of a network with a set of observations

A compatible network can be tested with different sets of observations of varied stresses: thermal, nutritional, hypoxic, *etc.* An observation is a pair of values of the form $gene = sign$ where $sign$ can be **+** or **−** indicating that the gene is expressed or respectively repressed under certain condition. To test the global network of *E. coli*, we have chosen a set of 40 observations for the stationary phase condition provided by RegulonDB (Table 1).

Table 1: Table of the 40 variations of products observed under stationary growth phase condition. Source: RegulonDB March 2006

gene	variation	gene	variation	gene	variation	gene	variation	gene	variation
acnA	$+$	csiE	$+$	gadC	$+$	osmB	$+$	recF	$+$
acrA	$+$	cspD	$+$	hmp	$+$	osmE	$+$	rob	$+$
adhE	$+$	dnaN	$+$	hns	$+$	osmY	$+$	sdaA	$-$
appB	$+$	dppA	$+$	hyaA	$+$	otsA	$+$	sohB	$-$
appC	$+$	fic	$+$	ihfA	$-$	otsB	$+$	treA	$+$
appY	$+$	gabP	$+$	ihfB	$-$	polA	$+$	yeiL	$+$
blc	$+$	gadA	$+$	lrp	$+$	proP	$+$	yfiD	$+$
bolA	$+$	gadB	$+$	mpl	$+$	proX	$+$	yihI	$-$

The set of 40 observations of the stationary phase was found to be inconsistent with the global network of *E. coli*. We found a direct inconsistency in the system of equations caused by the values fixed by the observations given to ihfA and ihfB: $\{ihfA = -, ihfB = -\}$, implying repression of these genes under stationary phase. This mathematical incompatibility agreed with the literature related to genes $ihfA$ and $ihfB$ expression under stationary growing phase. Studies [**?**],[**?**],[**?**],[**?**] agree that transcription of $ihfA$ and $ihfB$ increases during stationary phase. Supported by this information, we have modified the observations of $ihfA$ and $ihfB$ and the compatibility test of the global network of *E.coli* was successful.

3.4 Predictions over a compatible network from a set of observations

As mentioned earlier, a regulatory network is said to be consistent with a given set of observations when the associated qualitative system has at least one solution. If a variable is fixed to the same value in all solutions, then mathematically we are talking about a hard component, which is a *prediction or inference* for this set of observations.

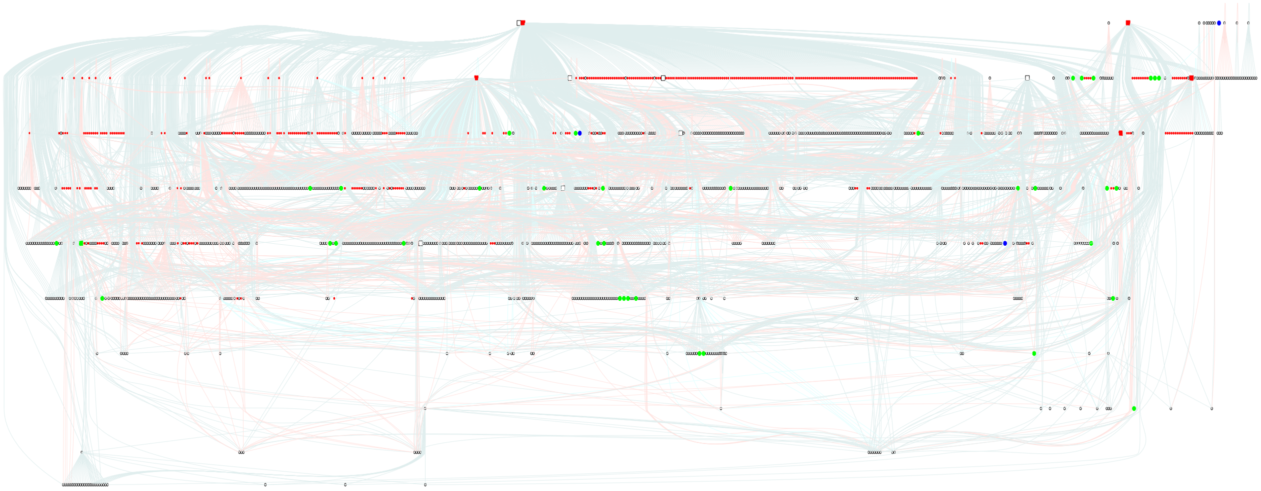


Figure 4: Global *E.coli* regulatory network with transcriptional and sigma-factors interactions (3883 interactions and 1529 products). Blue and red interactions represent activation or, respectively, repression. Green and blue nodes correspond to positive and negative observations (40). Red nodes (381) are the total inferred variations of products under stationary growth phase condition.

We have mentioned that the regulatory network including sigma factors is consistent with the set of 40 observations for stationary phase, after some correction. Actually there are about $2,66 \cdot 10^{16}$ solutions of the qualitative system which are consistent with the 40 observations of stationary phase. Furthermore, in all these solutions, 381 variables of the system have always the same value (they are hard components, see Figure 4). In other words, we were able to predict the variation: expressed (+) or repressed (−) of 381 components of our network (25% of the products of the network). We provide a subset of these predictions in Table 2.

Table 2: Table of 42 products inferred under stationary phase condition.

gene	variation
IHF	+
ada	+
agaR	+
alsR	+
araC	+
argP	+
argR	+
baeR	+
cadC	+
cpxR	+
crp	+
cusR	+
cynR	+
cysB	+
cytR	+
dnaA	+
dsdC	+
evgA	+
fucR	+
fur	+
galR	+
gcvA	+
glcC	+
gntR	+
ilvY	+
iscR	+
lexA	+
lysR	+
melR	+
mngR	+
oxyR	+
phoB	+
prpR	+
rbsR	+
rhaR	+
rpoD	+
rpoS	+
soxR	+
soxS	+
srlR	+
trpR	+
tyrR	+

3.5 Validation of the predicted genes

In order to verify whether the 381 predictions obtained from stationary phase data were valid, we have compared them with three sets of microarray data related to the expression of genes of

E.Coli during stationary phase. The result obtained is showed in Table 3. The number of compared genes corresponds to the common genes, the validated genes are those genes which variation in the prediction is the same as in the microarray data set.

Table 3: Validation of the prediction with microarray data sets

Source of microarray data	Compared genes	Validated genes (%)
Gutierrez-Rios and co-workers [**?**], stationary phase	249	34%
Gene Expression Omnibus ([**?**],[**?**]), stationary phase after 20 minutes	292	51.71%
Gene Expression Omnibus ([**?**],[**?**]), stationary phase after 60 minutes	281	51.2%

From the sets of microarray data provided by GEO (Gene Expression Omnibus) for stationary phase measured after 20 and 60 minutes, we have taken into account gene expressions whose absolute value is above a specific threshold and compared only these expression data with the 381 predictions. The percentage of validation obtained for different values of thresholds is illustrated in Figure 5. This percentage increases with the threshold, which is normal because stronger variations are more reliable.

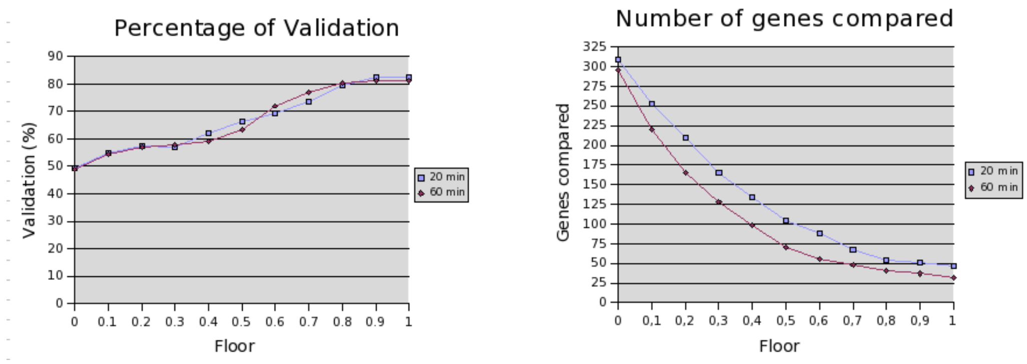


Figure 5: (Left) Percentage of validation of the 381 predicted variations of genes with microarray data sets from GEO (Gene Expression Omnibus) for stationary phase after 20 and 60 minutes. For both experiments we validate the 381 predictions with different sets of microarray observations considering only those genes which absolute value of expression is above certain value (threshold). (Right) Number of genes considered for the validation for the different used thresholds of both microarray data sets.

The percentage of our predictions that does not agree with the microarray results is due to:

- Erroneous microarray indications for certain genes. The genes $xthA$, cfa, $cpxA$, $cpxR$, gor are predicted as expressed by our model and as repressed by the microarray data [**?**]. Nevertheless, there is strong evidence that they are expressed during the stationary phase (see [**?**, **?**]).

- Incompleteness of our network model. Our model predicts that the gene $ilvC$ is expressed, which contradicts microarray data. More careful studies [**?**] document the decrease of the protein $IlvC$ due to an interaction with $clpP$ which is absent in our model. Indeed, under the introduction of a negative interaction between these species, $ilvC$ is no longer a hard component, which lifts the conflict with data.

4 Conclusions

Given an interaction graph of a thousand products, such as *E.coli* regulatory network, we were able to test its self-consistency and its consistency with respect to observations. We have used mathematical methods first exposed in [**?**, **?**, **?**].

We have found that the *E.coli* transcriptional regulatory network, obtained from RegulonDB site [**?**],[**?**] is not self consistent, but can be made self-consistent by adding to it sigma-factors which are transcription initiation factors. The self-consistent network (including sigma-factors) is not consistent with data provided by RegulonDB for the stationary growth phase of *E.coli*. Sources of inconsistency were mistaken observations.

Finally, a step of inference/prediction was achieved being able to infer 381 new variations of products (25% of the total products of the network) from *E.coli* global network (transcriptional plus sigma-factors interactions). This inference was validated with microarray results, obtaining in the best case that 40% of the inferred variations were consistent (37% were not consistent and 23% of them could not be associated to a microarray measure). We have used our approach to spot several imprecisions in the microarray data and missing interactions in our model.

This approach can be used in order to increase the consistency between network models and data, which is important for model refinement. Also, it may serve to increase the reliability of the data sets. We plan to use this approach to test different experimental conditions over *E.coli* network in order to complete its interaction network model. It should be also interesting to test it with different (signed and oriented) regulatory networks. All the tools provided to arrive to these results were packaged in a Python library called *PYQUALI* which will be soon publicly available. All scripts and data used in this article are available upon simple request to the authors.