Method for relating inter-patient gene copy numbers variations with gene expression via gene influence networks

Sylvain Blachon, Gautier Stoll, Carito Guziolowski, Andrei Zinovyev, Emmanuel Barillot, Anne Siegel and Ovidiu Radulescu

Abstract During tumorigenesis, genetic aberrations arise and may deeply affect the tumoral cell physiology. It has been partially demonstrated that an increase of genes copy numbers induces higher expression; but this effect is less clear for small genetic modifications. To study it, we propose a systems biology approach that enables the integration of CGH and expression data together with an influence graph derived from biological knowledge. This work is based on 3 key ideas. 1) Inter-individual variations in gene copy number and in expression allow to attack tumor varability and ultimately adresses the problem of individual-centered therapeutics. 2) Confronting post-genomic data to known regulations is a good way to check the soundness and limits of current knowledge. 3) The abstraction level of qualitative modeling allows integration of heterogeneous data sources. We tested this approach on Ewing tumor data. It allowed the definition of new biological hypotheses that were assessed by random permutation of the initial data sets.

Sylvain Blachon
IRISA, UMR 6074 CNRS/INRIA/Universite Rennes 1, campus de Beaulieu, F, 35042 Rennes Cedex, France e-mail: sylvain.blachon@irisa.fr

Gautier Stoll,
UMR 900 INSERM/Mines ParisTech/Institut Curie, 26 rue d'Ulm 75248 PARIS cedex 05, France

Carito Guziolowski
IRISA

Andrei Zinovyev
UMR 900 INSERM/Mines ParisTech/Institut Curie

Emmanuel Barillot
UMR 900 INSERM/Mines ParisTech/Institut Curie

Anne Siegel
IRISA

Ovidiu Radulescu
IRMAR , UMR 6625 CNRS/Universite Rennes 1, Campus de Beaulieu, 35042 Rennes cedex, France

Abreviations : *GCNv* = Gene Copy Number Variation ; *ELv* = Expression Level Variation; ES = Ewing Sarcoma

1 Introduction

Relating genomic instabilities to gene expression is a difficult challenge which is not yet completely resolved. The biological hypothesis is that a gene amplified genomically (in tumor cells for example) induces a higher expression.

Relating gene expression profiles to gene copy numbers was mostly performed using correlation analyzes, in order to find candidate genes serving as markers or as potential targets for therapy [12].

However, these correlation analyzes cannot explain all gene behaviours : in the best case, 50% of them can be explained [7, 10]. This proportion is much weaker for tumors that have less instabilities (like Ewing sarcoma) than more common tumors, like breast cancers. Hence, on those tumors, it appears difficult to extract relevant global properties to relate CGH data to tumor outcomes [8, 16, 11, 1] or to gene expression [3, 15].

We proposed new method for the study of genomic instabilities in tumors, based on the systems biology approach. In this approach, we include the biological processes that regulate transcription through the dynamics of one or several networks of interacting molecules. In such a model genes, transcripts and proteins are network components. The simple process one-gene-one-transcript-one-protein is replaced by a more global point of view involving all the connections among the network components.

In order to deal with small genetic modification, we adopt a more mechanistic approach to genetic variability via a network model. Genetic variability, having nowadays interesting perspectives in personalized medicine, has been addressed by various biologists since Darwin. The idea that interaction between genes can modulate the effects of this variability can be traced back to Conrad Waddington, whose chreods can be interpreted as representations of the “elastic” response of gene networks. Here, gene-gene interactions can stabilize the effect of genetic variability. However, this can be done only up to a certain extent, as some variability is necessarily persistent. The persistent variation is not entirely random, it bares information on the network.

Based on these ideas, a framework was conceived to address the following questions :

- how are gene copy number (*GCN*) and expression level (*EL*) variations related?
- is a (theoretical) gene regulation model consistent with (real life) observed variations in patient pairs? If so, what is the *GCN* contribution to consistency?

This framework is based on qualitative reasoning which formalizes biological interactions [13] and is efficient [1] on large scale regulatory networks [5, 17].

[1] see also the web interface : http://www.irisa.fr/symbiose/bioquali

We applied our methods to Ewing sarcoma: it is a pediatric bone tumors that originate from a translocation t(11;22)(q24;q12), producing a chimeric gene : *EWS-FLI1* [2]. This chimeric gene is thought to act as an aberant transcription factor. A set of target genes that can be either activated or repressed has been already discovered [14].

2 Available data and model for Ewing sarcoma

2.1 Data description and preprocessing

A home made Comparative Genomic Hybridization (CGH) array was built in-house at Institut Curie (3920 probes 60 bp long covering all the chromosomes). CGH data were produced on a set of 47 tumors, including 7 cell lines.

Among the 39 remaining tumors [2], 12 were diagnosed as metastatic tumors before analysis and therapy. After analysis and therapy, on the 27 remaining tumors, 10 evolved in metastasis while 17 remained localized tumors.

An Affymetrix U133A chip was used to measure expression levels on the biopsies of patient tumors. Microarray data were normalized by the GC-RMA technique.

Breakpoint detection on CGH data was performed using GLAD algorithm[6]. GLAD allows CGH level smoothing in a given genomic region flanked by two breakpoints.

2.2 Model description

A gene regulation model involving 130 genes, including $EWS-FLI1$, was designed within SITCON project [4]. The genes/pathways included in this network model were indentified by analysis of transcriptome time series on Ewing cell lines. The logical connections between genes are based on 1) scientific literature and 2) manually curation of TRANSPATH [9] database. Main tumor phetnypes are included in this network: cell cycle regulation, apoptosis and cell migration.

3 Systems biology method for analyzing genetic and expression variations

Our methodology aims at confronting pairwise variations with the ($EWS-FLI1$) gene network model described above.

[2] One was too noisy and was discarded from the analysis

In order to cope with genetic variability in gene networks, we represent differences between individuals as perturbations of the network. Biologically, the hypothesis is that data obtained on a cell population coming from a tumor biopsie reflect a molecular steady state. For a patient pair, the whole set of observed variations describes the qualitative differences between the two steady states. This can be coherently done in a framework that was first introduced in [13], based on interaction graphs and qualitative equations.

3.1 Qualitative equations

Consider a network of n interacting components. The interaction model is the digraph $G = (V, E)$, $V = \{1, \ldots, n\}$. There is an edge $j \rightarrow i \in E$ if j influences the production of i. Edges are labelled by a sign $\{+, -\}$ which indicates whether j activates or represses the production of i. Let us denote by $sign(\delta X_i) \in \{\mathbf{+}, \mathbf{-}, \mathbf{?}\}$ the sign of the variation of i between two conditions, and by $sign(j \rightarrow i) \in \{\mathbf{+}, \mathbf{-}\}$ the sign of the edge $j \rightarrow i$ in the interaction graph.

For every predecessor j of i, $sign(j \rightarrow i) * sign(X_j)$ provides the sign of the *influence variation* of j on the species i. Notice that this can be either positive (increased activation or decreased repression) or negative (decreased activation or increased repression). Then, the constraints that the network imposes on the variations can be expressed as qualitative equations:

$$sign(\delta X_i) \approx \sum_{j \rightarrow i} sign(j \rightarrow i) sign(\delta X_j). \tag{1}$$

The sign algebra is summarized in the following table.

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

$$\mathbf{+} \not\approx \mathbf{-} \,|\, \mathbf{?} \approx \mathbf{+} \,|\, \mathbf{?} \approx \mathbf{-}$$

3.2 Taking into account genetic variations

In order to take into account the genetic variability of the patients we introduced new qualitative variables representing, for a given pair of patients, the GCN variations. The corresponding nodes in the interaction digraph will be called “gene nodes”. There is one gene node for each gene considered and in our analysis we kept a set of 126 genes. The remaining nodes are either mRNA or protein nodes occurring in the ($EWS-FLI1$) network.

The **central hypothesis** here is that gene nodes act directly and positively on the mRNA nodes in the network.

To summarize, the interaction model contains:

1. gene nodes : the sign stems from *GCN* variation between two patients,
2. mRNA nodes : the sign stems from *EL* variation between two patients,
3. proteins : the sign stems from protein activity variation between two patients.

GCN variations and *EL* variations come from CGH and microarray data. The protein activity variations remain unknown but can be predicted thanks to our formalism.

3.3 Encoding variations

For each gene k, we define $GCNv_{i,j}^{k} = CGH(i,k) - CGH(j,k)$, where $CGH(i,k)$ is the CGH level of the gene k in the patient i smoothed by GLAD algorithm. When $\mid GCNv_{i,j}^{k} \mid > 0,2$ the variation is considered as significant [6].

Similarly, for gene expression variation $ELv_{i,j}^{k} = EL(i,k) - EL(j,k)$, where $EL(i,k)$ is the mean expression level measured by Affymetrix probes corresponding to the gene k in the patient i. To evaluate the significance of the variation, a Student test was used on the set of probesets measuring $EL(i,k)$ with an alpha risk of 5%.

Both for gene and mRNA nodes, significant variations are encoded **+** or **–**. The **?** sign is used for nodes that are undetermined at various steps of our calculations.

3.4 Consistency analysis

For each pair of patients, we solve the system of qualitative equations (1), augmented by the information on signs coming from data. If there are solutions, the system is declared compatible. In case of compatibility some nodes have the same unique sign in each one of the many possible solutions. The unique signs of these nodes (called hard components) are predictions of the model. By this, the signs on protein nodes are predicted.

If no solution can be found, a localization of the source of conflict is attempted by subsystem analysis. First, all local violations (meaning that at least one equation (1) is violated by data information) are declared “local inconsistencies”. All locally inconsistent patterns have the same structure : one node together with its predecessors. All the other situations are declared “global inconsistencies”. Globally inconsistent patterns are more complex (they contain at least two nodes with their respective predecessors).

Notice that testing the consistency and looking for sources of conflicts is actually a NP-hard question. It appears that the topology of the network allows to handle these questions. We used decision diagrams, a data structure meant to represent functions on finite domains; it is widely used for the verification of circuits or network protocols. Using such a compact representation of the set of solutions, we proposed efficient algorithms for computing solutions of the systems, predictions, and other properties of a qualitative system [17].

3.5 Monte Carlo estimates for statistical significance of consistency

Consistency could occur also by chance. In order to estimate the significance of consistency results, we used random perturbations and Monte Carlo estimates of the mean numbers of pairs of patients for which random data is consistent with the network.

For a pair of patients (i,j), let us note:

1. $C^+_{i,j}$ and $C^-_{i,j}$ the set of genes for which the gene copy numbers vary positively, resp. negatively, between the patients i and j.
2. $E^+_{i,j}$ and $E^-_{i,j}$ the set of genes for which the gene expressions vary positively, resp. negatively, between the patients i and j.

Straightforwardly, $C^+_{i,j} \cap C^-_{i,j} = \oslash$ and $E^+_{i,j} \cap E^-_{i,j} = \oslash$

The qualitative equations (1) were solved with $N = 1000$ data sets (each data set contains $P(P-1)/2$ patient comparisons, where P is the number of patients) produced by randomly permuting the elements contained in $C^+_{i,j}$, $C^-_{i,j}$, $E^+_{i,j}$ and $E^-_{i,j}$.

For each random dataset, consistency was tested. In case of consistency, predictions on network nodes were computed. Each random dataset r is consistent N^C_r times, locally inconsistent N^{LI}_r times and globally inconsistent N^{GI}_r times. Note that $N^C_r + N^{LI}_r + N^{GI}_r = P(P-1)/2$.

The distributions of N^C, N^{LI} and N^{GI} provide the estimates for the number of consistent and inconsistent pairs with random data we are looking for.

4 Results

In this section, we apply our method to Ewing sarcoma data. We show results for a couple of questions - the first concerning the relation between *GCNv* and *ELv* ; the second concerning the model consistency tests and the impact of *GCNv* on them.

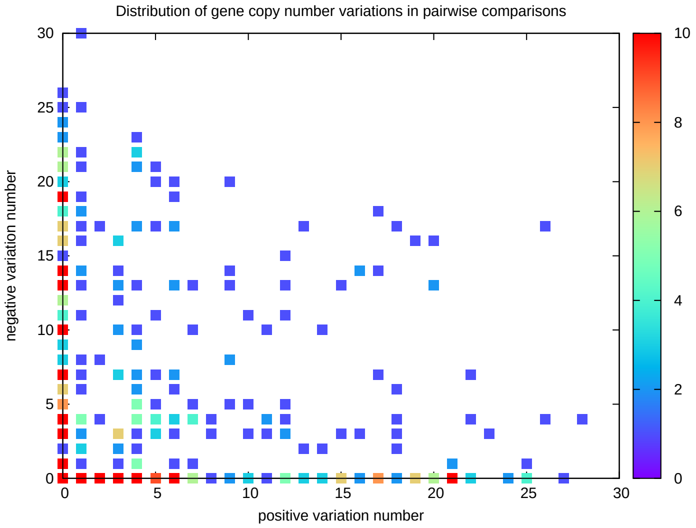

Fig. 1 Repartition of the $CGHv^+$ and $CGHv^-$ cardinals in the 741 patient pairs. Each point represents at least one patient pair - the number of patient pair is color-coded according to the palette on the right. Observe that a lot of patient pairs have few $CGHv$. The higher peak is the point (0,0) on which 164 patient pairs aggregate.

4.1 Discovering links between gene copy number variations and expression level variations

How are Gene Copy Number variations ($GCNv$) and Expression Level variations (ELv) related ?

$GCNv$ and ELv were evaluated for each gene and each patient pair. There are 39 patients, thus 741 patient pairs.

First, the *Figure 1* and *Figure 2* show the repartition of the variation numbers in the patient pairs.

It is striking to see the difference in variation repartitions. $GCNv$ are less frequent but also mainly distributed along the x and y axis. This is coherent with the relative genome stability of ES.

In spite of this general trend, some ES can exhibit a high number of $GCNA$s. When these unstable tumors are compared to rather stable tumors, imbalances are favored in one sense rather than the other, giving this picture with most of the pairs around the 0,0 point and distributed along the x and y axis.

A different picture can be observed with expression data. Variations are more frequent and distributed in a larger area, showing a rather homogeneous variability of ELv among the patient pairs.

From these figures, one can imagine that $GCNv$ and ELv are independent variables.

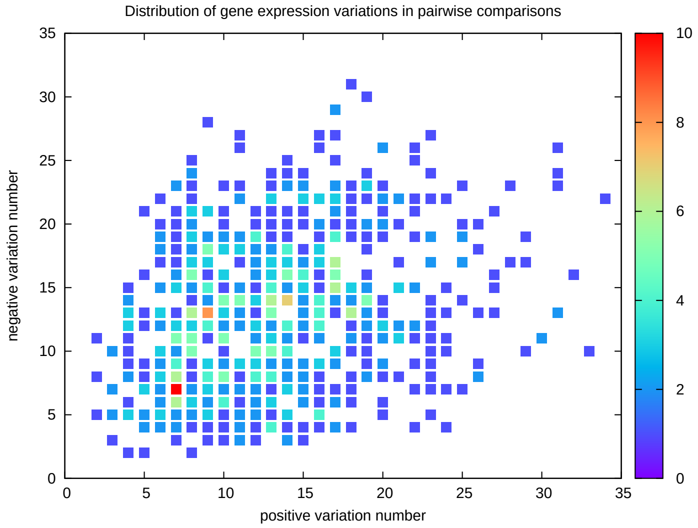

Fig. 2 Repartition of the ELv^+ and ELv^- cardinals in the 741 patient pairs. Each point represents at least one patient pair - the number of patient pair is color-coded according to the palette on the right. The distribution is much different than the *vCGH* one.

To verify this, a Pearson χ^2 independence test was performed under the null hypothesis that *GCNv* and *ELv* are two independent variables. The repartitions and contributions to the χ^2 score are shown in *Table 1*.

χ^2	*GCNv* / *ELv*	0	+	-	?
Observed	0	53117	2197	2598	14447
	+	6969	294	573	1948
	-	6781	550	283	2127
	?	1194	38	32	218
Expected	0	52547	2386	2701	14523
	+	7132	323	365	1964
	-	7101	321	364	1955
	?	1080	49	55	297
Contribution to *Chi*^{2}	0	2,59	15,0	3,98	0,40
	+	3,74	2,54	118,09	0,13
	-	14,4	162,91	17,9	15,10
	?	11,96	2,42	9,84	21,23

Table 1 Pearson χ^2 Independence test. $\chi^2 = 402,2 \gg 27,88$, the χ^2 value for 9 freedom degrees with an alpha risk of 0,001. The major contribution is given by situations where a gene changes in an anti-correlated way in *GCN* and in *EL*. The "?" sign corresponds to cases when a probe signal in at least one experiment is too noisy to assess the variation sign.

The χ^2 statistics equals 402, much greater than the value for an alpha risk of 0,1%., with 9 freedom degrees. The null hypothesis can be confidently rejected. This confirms that *GCNA*s can affect transcription in ways that can be investigated by comparing pairs of patients.

Moreover, the major contribution occurs when *GCNv* and *ELv* have opposite signs. This is even more striking on the whole gene set (χ^2 = 88761 ; contribution of *GCNv* and *ELv* having opposite sign = 83,7%). This was clearly unexpected and it is highly counter-intuitive.

We will show that our qualitative reasoning method allows to find explanations to this surprising phenomenon.

4.2 Checking the EWS-FLI1 regulation model

To address this issue, we used a systems biology approach based on the qualitative analysis to confront an interaction model with CGH and expression data.

The consistency analysis raw results are shown on *Figure 3*.

We were concerned with two main questions:

- In which proportion is the EWS-FLI1 network model consistent with real data? In cases of inconsistency, what does this tell about the model?
- Is information contained in CGH data useful to uncover regulations ?

4.2.1 Explaining inconsistencies

On real data including *GCNv* influences, the model is consistent with the data in 317 patient pairs (42,8% of the 741). Additionally, 314 (42,4%) local inconsistencies and 110 (14,8%) global inconsistencies were found.

Understanding the incompatibility sources may help to focus on the model weaknesses. First, it is necessary to analyze the sources of local inconsistencies, the most numerous ones.

All the local inconsistencies have the same origin : a patient pair (*i*,*j*) where there is at least one gene *k* for which : $GCNv(i,j,k) = -ELv(i,j,k)$. We call this an *anticorrelated variation*.

This is not a rare case: from the *Table 1*, there are 1123 cases in the 741 patient pairs. They are spread in 367 patient pairs and involve 67 genes.

Hence, 367 local inconsistencies were expected. This means that 50 pairs that were expected to be locally inconsistent were explained by the model.

More precisely, on the 67 genes that are involved in anticorrelated variations, 23 are never involved in local consistencies This is due to the presence in the network model of at least one transcription regulation on those genes. Those explained locally inconsistent influences appear 414 (36,9%).

In other words, local inconsistencies point to the lack of transcription regulations in the model. Adding them can potentially remove all local inconsistencies.

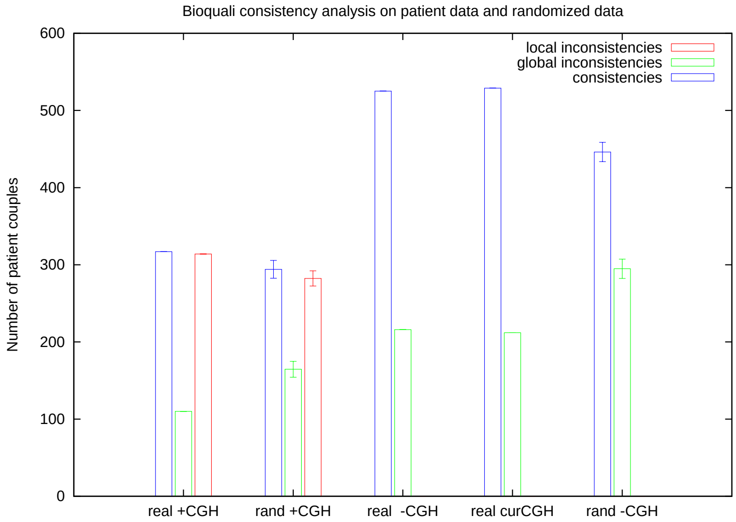

Fig. 3 Consistency global results on patient data and randomized data. *real* + *CGH* means that the model was confronted to patient data including all *GCNv* influences. *rand* + *CGH* means that the model was confronted to randomized data including all random *GCNv* influences. *real* − *CGH* means that the model was confronted to *ELv* observed on patient data without *GCNv* influences. *realcurCGH* means that the model was confronted to patient data without the unexplained anti-correlated *vCGH* / *ELv*. *rand* − *CGH* means that the model was confronted to randomized data without random *GCNv* influences.

The global inconsistencies point to other model weaknesses. Unfortunately, our solver is not able to localize the whole set of inconsistent subgraphs (see section 2)[3].

Only 33 influences are involved in the inconsistent set we obtained. All of themare implied in transcription regulation, including: $TP53$, $E2F1$ and $EWS-FLI1$.Therefore, our method allows to focus on subgraphs of a complex model that need refinement to become consistent with observations.

4.2.2 CGH data increase the consistency between the data and the model

To assess what relevant information is contained in CGH data, it can be useful to :

- reproduce the consistency analysis without taking into account the anticorrelated variations that remain unexplained by the model ;
- compare the result to the consistency analysis when *GCN* influences are removed (by taking into account only *ELv*, hence discarding CGH information).

This analysis gave the results shown in *Figure 3*. Without CGH information, the model is consistent with *ELv* alone in 525 (70,9%) patient pairs. Using CGH

[3] Notice that a more powerful implementation of constraints solver, with Answer Set Programming, will be soon available to overcome this technical problem.

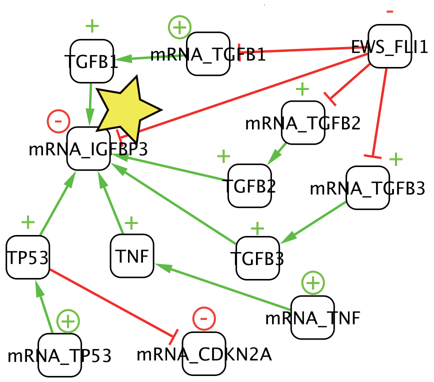

Fig. 4 Inconsistent subgraph without *GCNv* influences on the patient pair (EW57,EW58). Observed signs of gene expression variation are circled. The star points to the inconsistency localization, here the *mRNA_IGFBP3* node.

information on the genes having transcription regulators, the model is consistent with *ELv* and *GCNv* in 528 (71,3%) patient pairs.

In 3 cases, the model becomes consistent thanks to information from CGH data.

To understand better the impact of *GCNv* influences, the 3 inconsistent subgraphs were represented using Cytoscape software. The *Figures 4* and 5 show an example of this analysis.

On this example, the inconsistency is localized on the *mRNA_IGFBP3* node. *IGFBP3* is positively targeted by a set of regulators, except $EWS-FLI1$ that was shown to inhibit *IGFBP3*. However, between the two patients, given the observations, *mRNA_IGFBP3* should be activated. This is not the case, producing the inconsistency.

The *IGFBP3GCNv* proposes an explanation to this phenomenon : due to its negative variation between the two patients, the negative *IGFBP3ELv* can be understood. A biological interpretation for such a pattern can be: despite the positive signals arising from various regulators, the difference in gene copy number is sufficient to decrease *IGFBP3EL*.

Obviously, this hypothesis must be confirmed by experimental validation.

To conclude, CGH data bring information on local variations that have an influence on the $EWS-FLI1$ network model.

4.2.3 Assessing the statistical quality of consistency frequencies

In order to assess the quality of consistency tests, a randomization of input data was performed using 1000 random permutation on *GCNv* and *ELv* for each patient pair. The 741000 data sets with *GCNv* were confronted to the EWS-FLI1 model; the same analysis was repeated on the same datasets without considering *GCN* influences.

Consistency analyzes with and without *GCN* influences were performed to be compared to results on real datasets. Results are exposed in *Figure 3*. This shows that there are less consistencies and proportionally more inconsistencies on random data than on real data.

The distribution of consistency frequency distribution obtained for the 1000 datasets including *GCN* influences follows a normal distribution ($\mu = 279$, $\sigma = 10,1$ - Kolmogorov-Smirnov normality test value = 0,0289 $<$ 0,0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 317[4] equals 3,79%.

Similarly, the distribution of the consistency frequency distribution obtained on the 1000 datasets without *GCN* influences (see *Figure 6*). The distribution follows a normal distribution ($\mu = 446$, $\sigma = 12,5$ - Kolmogorov-Smirnov normality test value = 0,0251 $<$ 0,0386 , the bilateral value for an alpha risk of 5%).

Given such a distribution, the probability to obtain a consistency frequency equal to or greater than 525 [5] equals $1,31.10^{-10}$. This probability is even lower for the real data set using *GCN* explained by the model[6].

This proves that one can trust the consistency frequency obtained on real data sets.

However, it is surprising to observe such a high number of consistent cases on randomized data sets. We are currently investigating the reasons. Two hypotheses motivate us :

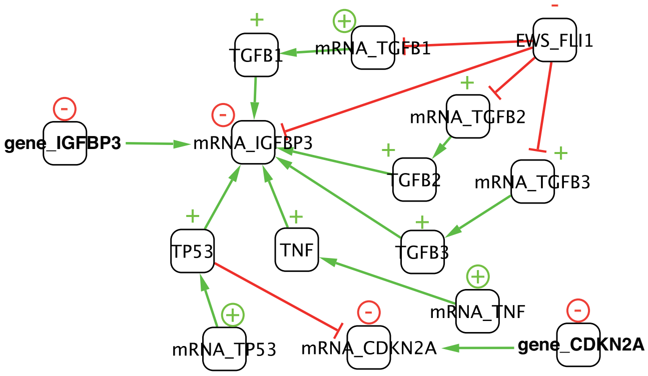

Fig. 5 Subgraph with *GCN* influences on patient pair (EW57,EW58). Observed signs of gene expression variation are circled. The *IGFBP3GCN* varies negatively between the two patients and resolves what was previously an inconsistency between the model and *ELv* alone.

[4] the consistency frequency obtained on real data set without *GCN*.

[5] The consistency frequency obtained on real data set without *GCN*.

[6] The consistency frequency obtained on real data set with the *GCN* influences explained by the model equals 528.

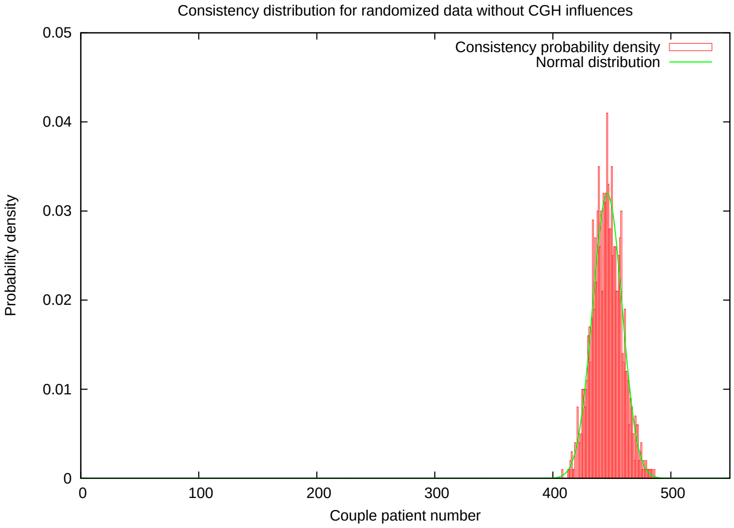

Fig. 6 Consistency frequency distribution for randomized data without ***vCGH*** **influences.** The distribution follows a normal distribution ($\mu = 446$, $\sigma = 12,5$). It must be compared to the consistency number obtained on real data without *GCNv* influences (525) and with *GCNv* explained by the model (528).

- the network topology may be robust to random variations;
- there is an effect of the number of constraints imposed by observations - as there is a variability in $\mid GCNv \mid$ and $\mid ELv \mid$ as shown on *Figure 1 and 2*.

A simple way is to test wether a significant correlation exists between $\mid GCNv(i,j) \mid$ and of $\mid GCNv(i,j) \mid$ and the consistency for each patient pair $(i$,$j)$.

5 Discussion

Given the difficulties to analyze CGH data on ES tumors, we propose to change of paradigm and propose a systems biology approach dedicated to invesitgate the inter-patient variability simultaneously at genomic and transcriptomic levels and their compatibility with a EWS-FLI1 gene regulation network.

This study addresses two main issues: 1) the link between CGH and expression pairwise variations in 39 ES; 2) the consistency between a $EWS-FLI1$ model and these variations.

To handle the first question, interesting representations of patient pairs as functions of *GCNv* and of *ELv* cardinalities were produced. It appears that the patient pair distribution following *GCNv* is highly different from its counterpart following *ELv*. This shows that patient pairwise comparisons exhibit different transcriptomic and genomic variability patterns. One could be mistaken in interpreting this as the result of an independence between the variables *GCNv* and *ELv*.

The χ^2 independence test states that these two variables are undoubtely related. This agrees with biological intuition: when a gene copy number increases, the gene expression level is expected to increase - and *vice versa*.

However, surprisingly, the major dependency contribution comes from anticorrelated variations. This is true for the whole set of measured genes. This suggests the existence of a feedback regulation of genes present in altered regions that counteracts *GCN* imbalances.

The $EWS - FLI1$ network model is able to deal in part with these anticorrelated variations : no gene having at least one transcriptional regulation appears in local inconsistencies. On the contrary, if a "deficient gene" does not have a transcription regulation, it will be involved in a local inconsistency if its *GCNv* and *ELv* appear anticorrelated. Thus, our method points to the model incompleteness. Adding missing transcription regulations can potentially remove all local inconsistencies.

To answer the second question, the compatibility of the EWS-FLI1 model with the pairwise genomic and transcriptomic variations was verified. It appears that the model is consistent with expression data in more than 70% of the cases after having silenced the "deficient gene" *GCNv* influence.

3 cases that were inconsistent using *ELv* alone become consistent. The analysis of these inconsistent subgraphs shows that some *ELv* that were unexplained by the model could be explained by local *GCN* variations. This suggests that local *GCN* variations carry valuable information that can propagate through the interactions. This result validates the capacity to investigate such local effects of *GCNv* by our approach.

Finally, we compared our results to 1000 randomized datasets. It appears that the consistency frequencies obtained on real datasets cannot be obtained by chance.

Another intriguing phenomenon appeared during this latter analysis: we did not expect so high consistency frequencies on randomized datasets. We are currently studying wether this is related to the genomic and transcriptomic variation number or whether this is a consequence of the intrinsic network robustness.

Biological system robustness may be the key to understand apparent contradictions in experimental data on ES. Let us consider the following paradox: the existence of a general trend that relates genetic instabilities to worst prognosis is opposed to the difficulty of finding repeated and specific genetic disorders linked to tumor outcomes.

If we consider that genetic instability acquisition is a stochastic process, stemming from a disturbed DNA repair machinery, it is likely that the largest part of genetic disorders have no individual effect on the cell physiology. This may result from a negative feedback control.

In the same time, it is also possible that, in exceptional cases, a specific gene disorder manages to overcome feedback and have visible effects on cell physiology. $EWS - FLI1$ itself is an extreme example.

As future work, we intend to use our method to detect these exceptional cases. We already proved that in a very limited number of cases (3 on 216 inconsistencies) the information on *GCNv* carried by specific genes can explain an unusual network behavior.

The novel hypothesis that outcomes from this work is that genetic disorder accumulation can have a global impact by increasing the probability that a specific gene disorder has consequences on a stabilized network. We expect that such events will be found more frequently in metastatic tumors than in non metastatic ones.

6 Acknowledgements

We acknowledge the financial support from ANR BIOSYS-2006 program (SITCON project) and the PIC Bioinformatique et Biostatistiques from Institut Curie. AZ, EB and GS are members of the team "Systems Biology of Cancer", équipe labellisée par la Ligue Nationale Contre le Cancer. We thank Philippe Hupé and Philippe La Rosa for help and disponibility for using Vamp. We also thank Michel Le Borgne and Jacques Nicolas for stimuling discussions and the bioinformatics GenOuest platform for their invaluable help to overcome technical problems and use the Bioquali application efficiently. SB's work was supported by a INRIA-French ministry postdoc grant.

References

1. S Brisset, G Schleiermacher, M Peter, A Mairal, O Oberlin, O Delattre, and A Aurias. Cgh analysis of secondary genetic changes in ewing tumors: correlation with metastatic disease in a series of 43 cases. *Cancer Genet Cytogenet*, 130(1):57–61, Oct 2001.
2. O Delattre, J Zucman, B Plougastel, C Desmaze, T Melot, M Peter, H Kovar, I Joubert, P de Jong, and G Rouleau. Gene fusion with an ets dna-binding domain caused by chromosome translocation in human tumours. *Nature*, 359(6391):162–5, Sep 1992.
3. B I Ferreira, J Alonso, J Carrillo, F Acquadro, C Largo, J Suela, M R Teixeira, N Cerveira, A Molares, G Goméz-López, A Pestaña, A Sastre, P Garcia-Miguel, and J C Cigudosa. Array cgh and gene-expression profiling reveals distinct genomic instability patterns associated with dna repair and cell-cycle checkpoint pathways in ewing's sarcoma. *Oncogene*, 27(14):2084–90, Mar 2008.
4. Stoll G., Zinovyev A., Tirode F., Laud-Duval K., Delattre O., and Barillot E. Model-based approach for analysis of transcriptome perturbation reveals ewing oncogene interaction network. In *International Conference on Systems Biology, poster*, Long Beach, CA, USA, 2007.
5. Carito Guziolowski, P Veber, Michel Le Borgne, Ovidiu Radulescu, and Anne Siegel. Checking consistency between expression data and large scale regulatory networks: a case study. *The Journal of Biological Physics and Chemistry*, 7(2):37–43, 2007.
6. Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–22, Dec 2004.
7. Elizabeth Hyman, Päivikki Kauraniemi, Sampsa Hautaniemi, Maija Wolf, Spyro Mousses, Ester Rozenblum, Markus Ringnér, Guido Sauter, Outi Monni, Abdel Elkahloun, Olli-P Kallion-

iemi, and Anne Kallioniemi. Impact of dna amplification on gene expression patterns in breast cancer. *Cancer Res*, 62(21):6240–5, Nov 2002.

8. S Knuutila, G Armengol, AM Bjorkqvist, W ElRifai, ML Larramendy, O Monni, and J Szymanska. Comparative genomic hybridization study on pooled dnas from tumors of one clinical-pathological entity. *Cancer Genetics and Cytogenetics*, 100(1):25–30, Jan 1998.
9. Mathias Krull, Susanne Pistor, Nico Voss, Alexander Kel, Ingmar Reuter, Deborah Kronenberg, Holger Michael, Knut Schwarzer, Anatolij Potapov, Claudia Choi, Olga Kel-Margoulis, and Edgar Wingender. Transpath: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Research*, 34(Database issue):D546–51, Jan 2006.
10. Hyunju Lee, Sek Won Kong, and Peter J Park. Integrative analysis reveals the direct and indirect interactions between dna copy number aberrations and gene expression changes. *Bioinformatics*, 24(7):889–96, Apr 2008.
11. T Ozaki, M Paulussen, C Poremba, C Brinkschmidt, J Rerin, S Ahrens, C Hoffmann, A Hillmann, D Wai, K L Schaefer, W Boecker, H Juergens, W Winkelmann, and B Dockhorn-Dworniczak. Genetic imbalances revealed by comparative genomic hybridization in ewing tumors. *Genes Chromosomes Cancer*, 32(2):164–71, Oct 2001.
12. Jonathan R Pollack, Therese Sørlie, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA*, 99(20):12963–8, Oct 2002.
13. O Radulescu, S Lagarrigue, A Siegel, P Veber, and M Le .... Topology and linear response of interaction networks in molecular biology. *Journal of The Royal Society Interface*, Jan 2006.
14. N Riggi and I Stamenkovic. The biology of ewing sarcoma. *Cancer Letters*, Jan 2007.
15. S Savola, F Nardi, K Scotlandi, P Picci, and S Knuutila. Microdeletions in 9p21.3 induce false negative results in cdkn2a fish analysis of ewing sarcoma. *Cytogenet Genome Res*, 119(1-2):21–6, Jan 2007.
16. M Tarkkanen, S Kiuru-Kuhlefelt, C Blomqvist, G Armengol, T Bohling, T Ekfors, M Virolainen, P Lindholm, O Monge, P Picci, S Knuutila, and I Elomaa. Clinical correlations of genetic changes by comparative genomic hybridization in ewing sarcoma and related tumors. *Cancer Genetics and Cytogenetics*, 114(1):35–41, Jan 1999.
17. Philippe Veber, Carito Guziolowski, Michel Le Borgne, Ovidiu Radulescu, and Anne Siegel. Inferring the role of transcription factors in regulatory networks. *BMC Bioinformatics*, 9:228, Jan 2008.