Study of the Stability of the Protein Interaction Network using Gene Expression data inspired on Simulated Annealing

KRISTINA IBÁÑEZ GARIKANO

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA. FACULTAD DE INFORMÁTICA
UNIVERSIDAD COMPLUTESNE DE MADRID

Trabajo Fin de Máster en Sistemas Inteligentes

Junio 2012

Calificación: SOBRESALIENTE (10)

Directores:

Gonzalo Pajares Martinsanz
María Guijarro Mata-García
Alfonso Valencia Herrera

Autorización de difusión

Kristina Ibáñez Garikano

21 de Junio de 2012

La abajo firmante, matriculada en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "Study of the Stability of the Protein Interaction Network using Gene Expression data inspired on Simulated Annealing", realizado durante el curso académico 2011-2012 bajo la dirección de Gonzalo Pajares Martinsanz y María Guijarro Mata-García [y con la colaboración externa de dirección de Alfonso Valencia Herrera] en el Departamento de Ingeniería del Software e Inteligencia Artificial , y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Resumen

Este trabajo describe un método basado en la filosofía del enfriamiento simulado, para el caso del estudio de la estabilidad de la red de interacción de proteínas en tejidos tumorales y normales, a partir de datos de expresión de genes.

Presentamos una propuesta original para estudiar la inestabilidad que se da en las redes de interacción de proteínas, basándonos en la conocida estrategia del Enfriamiento Simulado como técnica de optimización adaptado para nuestro propósito.

La estrategia propuesta se ha verificado con datos de expresión de genes para diferentes tejidos en estado normal y tumoral, sobre una red de interacción de proteínas. Con ella se ha demostrado la hipótesis y motivación inicial con la que partimos: la expresión de genes en tejidos tumorales desestabiliza más la red de interacción de proteínas frente a la expresión génica en tejidos normales.

Palabras clave

Enfriamiento Simulado, Inteligencia Artificial, Biología de Sistemas, Expresión Génica

Abstract

The aim of this project is to study the stability on the protein interaction network from gene expression data on tumor and normal tissue, implementing for that several algorithms based on the physical idea of Simulated Annealing.

We present a novel approach to identify the degree of destabilization based on the evolution of the well-known Simulated Annealing as an optimization approach customized for such purpose.

The proposed strategy is verified with gene expression data from different tissue in normal and tumor cases, as well as a protein interaction network. With this approach we demonstrate our hypothesis and initial motivation: gene expression in tumoral tissue destabilizes more the protein interaction network comparing to the result normal tissue do.

Keywords

Simulated Annealing, Systems Biology, Artificial Intelligence, Gene Expression

Contents

Acknowledgements

A Alfonso por sus ideas, críticas, por exigirme y por darme la oportunidad de realizar la tesis en su laboratorio. A Gonzalo por su entrega, ayuda y paciencia. A María por su ánimo siempre vivo. A mis compañeros de laboratorio que me han sufrido, a mis amigos y a mi familia que me han apoyado. En especial a Clara, que me has entendido y animado en todo momento y a Black Foot por alegrarme cada día.

Eskerrik asko!

Chapter 1

Introduction

The availability of high-throughput experimental data has allowed construction of increasingly comprehensive and accurate protein-protein interaction networks (interactome) and its study is more frequently providing valuable information on biological systems. The structure or topology of such networks sheds light on the complex cellular mechanisms and processes as well as gives insight into evolutionary aspects of the proteins.

1.1 Motivation

Johnsson and Bates (2006) show that known cancer genes exhibit a network topology that is different from that of protein not documented as being mutated in cancer. In particular, cancer proteins show an increase in the number of proteins they interact with. They also appear to participate in central hubs rather than peripheral ones, mirroring their greater centrality and participation in networks that form the backbone of the proteome. In the same way, Wachi *et al.* (2005) emphasize the high centrality of genes differentially expressed in lung cancer.

Systems biology studies shed light on the relationships between genes, the complex cellular mechanisms and processes as well as evolutionary aspects of the proteins. We can combine different type of biological data in order to simulate several biological behaviours.

The structure of the protein-protein interaction network varies depending on which protein is activated or expressed and also depending on the interaction protein partners. Microarray gene expression studies provide us the probabilities of proteins of being expressed in certain conditions. For the entire protein interaction network we could simulate how the topology changes in different conditions, from gene expression data.

1.2 Principle Objectives

The main idea is to simulate the dynamic structure of the protein interaction network in tumor and normal tissues using gene expression data from the biological point of view.

The structure and topology of the protein network emphasize the different interactions that proteins have. In turn, gene expression data provides somehow the probability of proteins of being expressed. Combining these biological data, our intention is to simulate the protein interaction network according to the probability of proteins to being expressed (gene expression data) and compare these simulations in normal and tumoral studies.

The simulation consists in re-create different protein interaction networks according to gene level expression data for different conditions; applying an approach based on the philosophy of the simulated annealing principle, which has been applied to computational problems (Duda *et al.* 2001), where optimization is a key issue.

In this work we map protein interactions in a network of nodes, where each node is characterized by its state and its relations among the other nodes, allowing the application of the simulated annealing process. With such purpose we define a new approach inspired on the Deterministic Simulated Annealing algorithm proposed in Duda *et al.* (2001). This explains the name given to our method, *i.e.* Pseudo-Deterministic Simulated Annealing.

1.3 Organization of this work

This work is organized as follows: In the chapter 2 we introduce basic concepts in molecular biology in order to better understand the biological problem presented in our study. The biological data used in this study goes on the chapter 3 as well as the preprocessing and normalization methods used on them. In the chapter 4 we present the new approach that identifies the degree of destabilization based on the evolution of the well-known Simulated Annealing (SA) approach followed by the results in chapter 5. The last chapter includes either conclusions or future ideas.

Chapter 2

Biological Concepts

2.1 Introduction to Molecular Biology

Cells are fundamental building blocks of living organisms. Cells contain a nucleus, mitochondria and chloroplasts, endoplasmatic reticulum, ribosomes, vacuoles, etc. The nucleus is important organelle because it houses chromosomes which include the DNA. The DNA is in essence a blueprint of the organism as it encodes information needed to synthesize proteins.

DNA is an extremely long molecule that forms a double-helix. The double-helix backbone of the molecule consists of sugars and phosphates, and there is one base attached to each sugar. There are four types of bases: cytosine (C), guanine (G), adenine (A) and thymine (T). The DNA consists of two strands, and each base attached to one strand forms a bond with a corresponding base on the other strand.

Protein is a sequence of amino acids, and the functional subunit of DNA that encodes a protein is called a gene. Protein is a class of macromolecules that carries out most of the activities in the cell. Cells are largely made up of proteins: structural proteins that give the cell rigidity and mobility, proteins that form pores in the cell membrane to control the traffic of small molecules into and out of the cell, and receptor proteins that regulate cellular activities in response to molecular signals from the growth medium or from other

cells. Proteins are also responsible for most of the metabolic activities of cells. They are essential for the synthesis and breakdown of organic molecules and for generating the chemical energy needed for cellular activities.

In molecular and cell biology, the central *dogma* is the passage of information from DNA to RNA to protein. Fig.2.1 displays a scheme of the central *dogma*: DNA sequence coding for the first seven amino acids in a polypeptide chain. The DNA sequence specifies the amino acid sequence through a molecule of RNA that serves as an intermediary "messenger". Although the decoding process is indirect, the net result is that each amino acid in the polypeptide chain is specified by a group of three adjacent bases (codon) in the DNA (translation). The term *dogma* means "set of beliefs"; it dates from the time the idea was put forward first as theory. Since then the "dogma" has been confirmed experimentally, but the term persists and it is not as simple as it seems. The central *dogma* is shown in Fig.2.1.

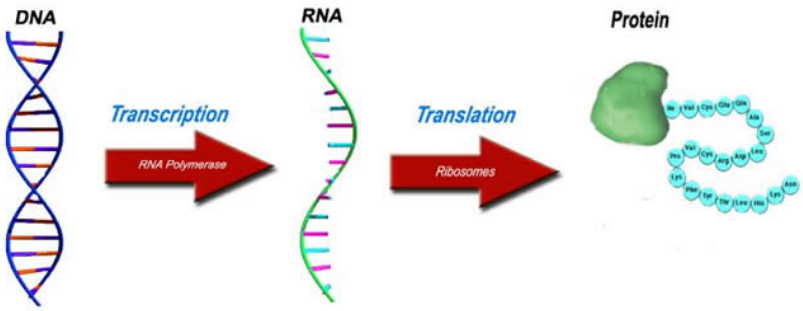

Figure 2.1: *Central* Dogma *in Molecular Biology*

The central *dogma* is the fundamental principle of molecular genetics because it summarizes how the genetic information in DNA becomes expressed in the amino acid sequence in a polypeptide chain.

The **transcription** is the process of making a RNA copy from a sequence of DNA (a gene). Transcription is the first step leading to gene expression. The stretch of DNA transcribed into and RNA molecule (*transcription unit*) encodes at least one gene. If the gene transcribed encodes a protein, the result of transcription is a messenger RNA (mRNA), which will then be used to create that protein via the process of **translation**. Alternatively, the transcribed gene may encode for either non-coding RNA genes (such as microRNA) or ribosomal RNA (rRNA) or transfer RNA (tRNA), other components of the protein-assembly process.

Gene expression is a two-step process in which DNA is converted into a protein it encodes. The first step is DNA transcription. In this step, the information from the archival copy of DNA is imprinted into short-lived mRNA. The structure of RNA is a little different, it contains ribose instead of deoxyrybose, and the four bases that bind to it are cytosine (C), guanine (G), adenine (A) and uracil (U). During transcription, DNA unfolds, and mRNA is created by pairing mRNA bases with the bases of RNA. In this process C in DNA translates to G, G to C, A to U, and T to A. After mRNA is translated, it is transported to the ribosome. The second step, protein translation occurs at the ribosome. During translation, the sequence of codons (triplets of bases) of mRNA is, with the help of tRNA, translated into a sequence of aminoacids.

The control of gene expression causes most phenotypic differences in organisms. Since many diseases result from complex changes on the molecular level, we need to observe and model these processes on the system level. Gene regulatory circuits are an example of machinery that allows us to depict gene expression graphically.

A measurement of the amount of gene product is sometimes used to infer how active a gene is. Abnormal amount of gene product can be correlated with a deregulation of the transcription that can directly cause anomalous behaviours related with a disease. An

example of this can be seen in the study published by Nagel *et al.* (2012): a low expression levels of ZHX2 is associated with a poor prognosis, indicating tumor suppresor activity in B-cell malignancies. Płuciennika *et al.* (2006) also show that reduced WWOX expression commonly observed in various neoplasias in cases of breast cancer is associated with markers of bad prognosis; high level expression of WWOX is associated with better disease free survival.

2.2 Gene Expression Analysis

Recent advances in biology provide different and diverse techniques and tools in order to measure gene expression. The data used in this study comes from experimental studies done with microarray technology, Affymetrix Human Genome U133 Plus2.0 Array [1].

A DNA microarray is a collection of microscopic DNA spots attached to a solid surface. Scientist use DNA microarrays to measure the expression levels of a large numbers of genes simultaneously or to genotype multiple regions of a genome.

The basic concept of microarray analysis is simple (Fig.2.2). RNA is harvested from a cell type or tissue of interest and labeled to generate the *target*. This is hybridized to the *probe* DNA sequences corresponding to specific genes that have been affixed, in a known configuration, onto a solid matrix. Hybridization, based on Watson and Crick base pairing, between probe and target provides a quantitative measure of the abundance of a particular sequence in the target population. This information is captured digitally and subjected to various analyses to extract biological information. Comparisons of hybridization patterns enables the identification of mRNAs that differ in abundance in two or more target samples. Thus microarrays provide a powerful tool with which to screen biological specimens for alterations in the expression of mRNAs that accompany, and may regulate physiological and pathological change.

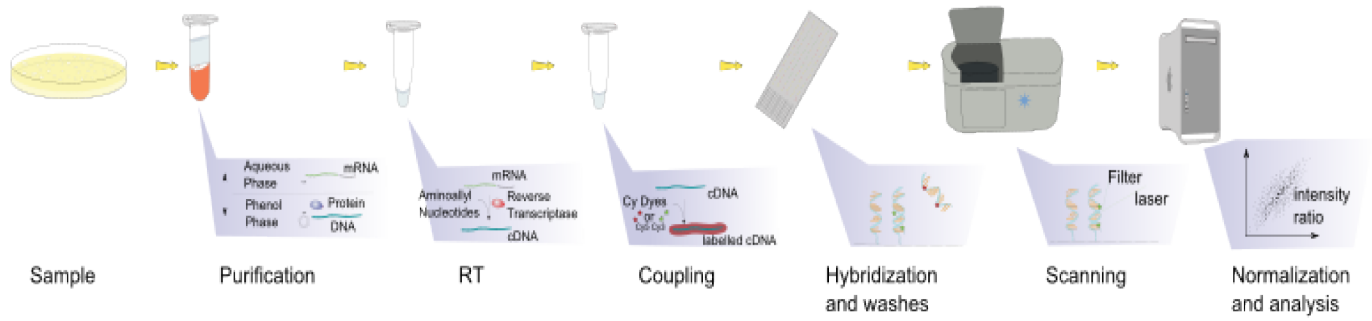

Figure 2.2: *Steps in a Microarray Experiment*

Microarray techniques are finding utility in **gene discovery** – the global description of genes potentially involved in developmental, physiological and pathological processes; **gene regulation** – the description of regulatory networks, based on the assumption that genes regulated in parallel share common control mechanisms; **diagnosis** – identification of patterns of gene expression that define disease states and that may represent prognostic indicators and **drug discovert** and **toxicology**.

Gene expression techniques have also several limitations: the output from the analysis of a microarray experiment is usually a large data spreadsheet filled with numbers related to the signal intensity for each gene on the chip. Further analysis is required to identify groups of genes that are similarly regulated across the biological samples under study. It is not a quantitative method, and so it does not reflect perfectly the reality.

In our study we extracted gene expression data from the GEO database (Wheeler *et al.*, 2001) as it is described on the following section3.1.

2.3 Protein-Protein Interactions

Proteins play crucial roles in living cells and in life; they perform their functions under constant motions, adopting different shapes, flexibility and interactions with other biological molecules. Typically, within cells, proteins interact with other proteins, metabolites and nucleic acids. Protein-protein interactions occur when two or more proteins bind together, often to carry out their biological function.

Proteins, as macromolecules they are, are composed of long complex chains of molecules (polymers) made up of simpler, smaller subunits (monomers). They are joined together in a process known as dehydration synthesis, in which a covalent bond is formed between two monomers by releasing a water molecule. Proteins are ready to interact under physic laws that are related with the superface and flexibility of the proteins.

Interactions between proteins are important for the majority of biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein-protein interactions of the signalling molecules. Proteins might interact for a long time to form part of a protein complex (group of two or more associated polypeptide chains), a protein may be carrying another protein or a protein may interact briefly with another protein just to modify it. In conclusion, protein-protein interactions are important for **virtually** every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches.

Protein-protein interactions have been measured using a variety of assays, such as immunoprecipitations and the yeast two-hybrid approach. These techniques have been scaled up to measure interactions on a genome-wide level. High-throughput techniques have also been developed to systematically identify protein complexes using affinity purification techniques followed by mass spectrometry to sequence proteins.

One of the grand challenges for molecular biology is to reconstruct the complete network of protein interactions within cells. This so-called interactome sheds unprecedented light on the inner workings of cellular machinery. Analysis of the network should also permit scientist to select protein targets for therapeutic intervention by understanding the underlying mechanisms of action. Eventually, protein networks may also be used to construct comprehensive dynamic models of molecular interactions within cells, allowing scientist to quantitatively predict the outcome of experiments.

High-throughput sequencing has facilitated the prediction of proteins coded within a genome, thus providing a list of the interactome's constituents, constructing such powerful models on a genome-wide scale.

Along with experimental approaches to detect protein interactions, computational methods have also been developed. These methods search for pairs of proteins that have co-evolved, implying that they are likely to be interacting within the cell. Although computationally derived interactions are generally not as reliable as experimentally measured ones, they provide a more complete and accurate understanding of protein interactions in combination with experimental data.

In summary, proteins do not act alone, and they often form complexes with other proteins to perform a specific task. The existing interactions between them have been studied and we use this information in order to see how the interactome evolves depending on the expression of the proteins.

As we presented in the section1.1, Johnsson and Bates (2006) show that known cancer genes exhibit a network topology that is different from that of protein not documented as being mutated in cancer. In particular, cancer proteins show an increase in the number of proteins they interact with.

In this study we used a collection of six protein interactions databases PINA (Wu *et al.*, 2009) : IntAct, MINT, BioGRID, DIP, HPRD and Mpact.

Chapter 3

Biological Datasets

Gene expression data is available in public sources due to the availability of high-throughput experimental data as well as protein interaction data.

3.1 Gene Expression Data

Gene Expression Omnibus (GEO)[3] is a public functional genomics data repository offered by the NCBI[15] (National Centre for Biotechnology Information). They provide array- and sequence-based data, as experiments and curated gene expression profiles.

There are thousands of gene expression samples available related with different types of diseases. A sample describes the conditions under which an individual sample was handled, the manipulations it underwent and the abundance measurement of each element derived from it; *e.g.* for a gene expression DNA microarray it would provide a list of (already pre-processed) expression values for each gene. Each sample record or identification is assigned a unique and stable GEO accession number of type GSMxxx (Appendix A). To know which samples choose, we used Barcode, the human transcriptome repository, so as to select the samples related with gene expression experiments related with different tissues in normal and tumoral conditions.

The Gene Expression Barcode (Zilliox and Irizarry, 2012) is a public repository that provides reliable absolute measures of expression for most annotated genes for 131 human and 89 mouse tissue types, including diseased tissue. They also implement an algorithm that leverages information from the GEO and ArrayExpress public repositories to build statistical models that permit converting data from a single microarray into expressed/unexpressed calls for each gene.

As their approach provides the original expression data for each tissue; in our study we used gene expression experimental samples in normal and tumoral cases of the following tissues: ovary, breast, colon, lung, liver and kidney. The identifications of the samples (GSM id's) are presented on Appendix A.

A platform defines the list of elements that may be detected and quantified in an experiment; *e.g.* for a gene expression DNA microarray it would provide a list of the cDNA spots or oligonucleotide probesets on one array, each time with an annotation of the gene. The platform of the technology used in all these samples was HGU133Plus2 (Affymetrix Human Genome U133 Plus 2.0 Array Annotation Data) [1]. This platform contains information measures for 10750 genes.

For microarrays to be used in a clinical setting to aid in diagnosis or treatment, one needs to be able to gain useful information from a single microarray hybridization: one needs to be able to:

- Preprocess a single microarray.
- Estimate the expression of each gene on the array.

A recent work by McCall *et al.* (2010) provide a method of single array preprocessing called Frozen Robus Microarray Analysis (fRMA). Previous work by Zilliox and Irizarry (2007) and, more recently, by McCall *et al.* (2009) show how to obtain gene expression estimates from the preprocessed data. Specifically, Zilliox and Irizarry (2007) develop a method to map gene intensities into a vector of ones and zeros denoting which genes are expressed (ones) and unexpressed (zeros) in a given example. They call to this sequence of ones and zeros a gene expression barcode.

In this study we implemented several scripts in order to follow the steps briefly shown in fig.3.1, then we explain some important facts of each stage.

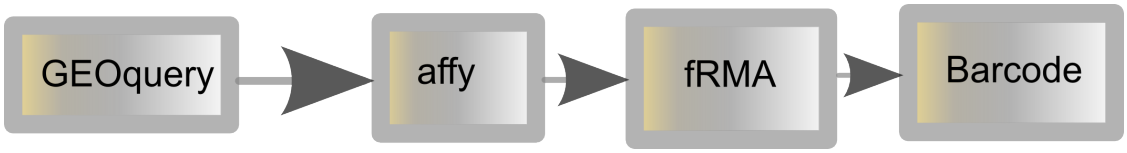

Figure 3.1: *Analysis Stages of the Gene Expression Data*

3.1.1 Preprocessing Single Affymetrix Microarrays

Before preprocess the raw data we extracted the list of samples (GSM files) that are shown on the Appendix A, from the GEO repository. For that, we used the GEOquery package. With the function, getGEO('gsm-id'), we got all the CEL files corresponding to all the studies. (Each GSM contains a CEL files).

For Affymetrix microarray data it is customary to view CEL files as the starting point for preprocessing and analysis. One or more CEL files can be read into R using the ReadAffy function from the affy package to produce an AffyBatch object (Gautier *et al.*, 2004). We adopted this convention and referred to the probe-level intensities from a single CEL file stored in an AffyBatch object as the raw data.

This is a fundamental step in order to do the fRMA analysis.

3.1.2 Frozen Robust Multiarray Analysis (fRMA)

The goal of fRMA is to obtain reliable gene-level intensities from a single microarray hybridization. This amounts to converting raw probe-level intensities into background-corrected, normalized gene-level intensities. The fRMA package contains a function fRMA which takes an AffyBatch as input and produces an object containing gene-level expression values.

To preprocessed single arrays using fRMA, one needs a number of frozen parameter vectors. Among these are the reference distribution to which the data are normalized and the probe-effect estimates. McCall *et al.* (2009) computed these frozen parameters for three popular Affymetrix platforms, and one of them is the one used in this study: hgu133plus2 [1].

3.1.3 Gene expression barcode

The creation of a gene expression barcode is designed to convert gene-level intensities into gene expression estimates. The algorithm by default takes the output from frma and creates a gene expression barcode: we get a z-score for each gene and each sample.

In conclusion, for every tissue and every sample we got 10740 genes with its corresponding z-score, denoting the intensity of the expression of the gene. Z-score is a test of statistical significance that helps decide whether or not to reject the null hypothesis. Z-scores are measures of standard deviation. Calculating the distribution function of each z-score, with the pnorm function, we also get the probability of the expression of each gene (values between 0 and 1) of being expressed.

The presenting approach, based on the philosophy of SA, uses as information of nodes the probability of each gene of being expressed, if it does, and the probability of not being expressed, if it does not.

3.2 Protein-protein Interaction Data

For the protein-protein interaction we used data from PINA (Wu et al., 2009). This database includes a quaterly updated, nonredundant data based on integration of data from six public PPI databases: IntAct, MINT, BioGRID, DIP, HPRD and MIPS Mpact.

These databases provide us protein-protein interactions data using different kind of methodology: IntAct provides all interactions derived from literature curation or direct user submissions. MINT stores information about molecular interactions by extracting experimental details from work published in peer-reviewed journals. BioGRID is a curated protein-protein and genetic interactions database from primary literature and compiled by in house large-scale curation efforts. DIP catalogs experimentally determined interactions between proteins, combining information from a variety of sources. HPRD contains manually curated scientific information pertaining to the biology of most human proteins. And MIPS offers a collection of manually curated high-quality protein-protein interaction data collected from the scientific literature by expert curators.

3.2.1 PPI data extraction

Once we extracted the information about the protein interactions we made an enrichment analysis on protein aliases. Each protein might have more than one identification. So as to consider all the combinations, we made the expansion on the network's aliases.

For all the existing genes on the platform used for the gene expression analysis there is its corresponding gene or alias on the protein network.

The resulting network has 64001 protein-protein interactions.

Chapter 4

Pseudo Deterministic Simulated Annealing Approach

Our aim is to see different dynamic structures on the protein-protein interactions network, mapping into the network expression of the genes in normal and tumoral tissues. Following the idea of Johnsson and Bates (2006), our hypothesis is that there is more destabilization on the network related with tumoral data, there are more interactions between proteins and a mutation or change in any of them causes a major destabilization on the network because it is more connected.

The novel method we are presenting in this study is the approach based on the main idea of Simulated Annealing (SA) method. SA technique permits to map interactions between every node, represent each node, as well as calculate the energy of the overall network. There is nothing done yet in this field to calculate and weight a network based on its "energy". This is our contribution in this work.

4.1 General Comments

SA is a probabilistic method proposed by Kirkpatrick et al. (1983) and Cerny (1985) in combinatorial optimization for finding the global minimum of a cost function that may possess several local minima. It works by emulating the physical process whereby a solid is slowly cooled so that when eventually its structure is "frozen", this happens at a minimum energy configuration.

This technique was first described by Metropolis *et al.* (1953) in the thermodynamic field. The main idea based on traditional processing of metals, a standard method to improve the quality of the metal is to heat it up to high temperatures, then slowly cool it down. Sometimes this is done in repetitive cycles. This method is called annealing.

The idea is to introduce a variable E (which may not have anything at all to do with any thermodynamic energy nor linked with a temperature). Then a generalized Metropolis scheme is used to simulate the system at the energy E under a given temperature, while E is gradually cooled down from some high starting temperature.

This technique has been demonstrated to be a successful approach in order to solve a variety of optimization problems (Laarhoven and Aarts, 1988). Duda *et al.* (2001) proposed a modified version of the original SA approach called deterministic against the stochastic version, which is essentially computationally slow, in part because of the discrete nature of the search along the space of all configurations.

4.2 Energy Minimization based on SA

In this study we propose a new approach based on the idea of simulated annealing as we mentioned previously. In the Appendix B is included the whole description of the deterministic SA (DSA) according to Haykin (1994) and Duda *et al.* (2001) which also follows the physical analogy. It is based on a set of nodes interconnected among them and each one with its associated state. These states evolve toward stable states based on the forces exerted by the nodes interconnected and also through the temperature according to the scheduled cooling process. Some nodes have influence over other nodes across these interactions and also the weights defined in these relationships.

This structure can be seen as a network of nodes with their states being dynamically updated toward a global stable state for the network. The network stability is measured by an energy function, which is conveniently defined according to the nature of the problem to solve. Again, this energy is inspired on the physical simulated annealing process, where the metal achieves more stable states as it cools, *i.e.* achieves minimum energy. Minimum energy values represent network stability. The term deterministic is because it is possible to determine deterministically the evolution of the states as the temperature decreases. The energy function involves both: 1) all states of the nodes and 2) the strength of the interconnections. The algorithm reaches the convergence criterion when this energy achieves its minimum value or a constant value, which indicates no more changes updating the states occur.

Our proposed strategy is inspired on the definition of an energy function, where as in the general deterministic approach it measures the stability of the network, where less energy is related to more stability. We cannot say that we are applying the DSA method to its fullest extent. This justifies the name of the proposed approach as PseudoDSA (hereinafter PDSA). In our case, the fact of using an energy function that decreases along interactions or time does not make sense due to the characteristics of the biological problem described,

which is to be solved. For this reason, as we ignore the descent of the energy during the time we cannot talk about a direct and a proper application of a DSA method, but based to that. This fact justifies the previous statements we already described about the inspiration of our approach on SA, not making use of all the methodology. Basically we are inspired on the idea and we implemented a new approach using the philosophy extracted from DSA methodology.

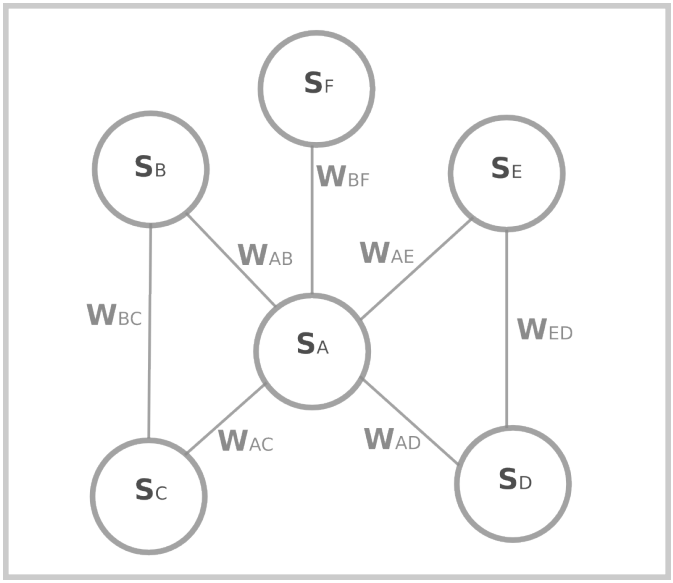

Figure 4.1: *Representation of states for the PseudoDSA approach*

Our system is represented by a protein-protein interaction network. Nodes represent proteins. As described in Fig.4.1, each S_i represents the probability of being expressed the protein i, if it is expressed; as well as the probability of not being expressed if it is not expressed. Specifically all S_i represent the states of the nodes in the original DSA approach. Edges show existing interactions between proteins if they both are expressed. Each W_{ij} represents the weight described in equation (1), we have defined this parameter in order to define the PDSA method.

As mentioned before, we based mainly on the idea of using an energy function that minimizes the energy in the system as the network reaches more stable states. The weights (W_{ij}) of the connections or interactions are defined in equation 1.

$$W_{ij} = \begin{cases} -1 \text{ if } S_i\, expressed, S_j\, expressed \\ -1 \text{ if } S_i\, expressed, S_j\, not\, expressed, prob(S_i) > prob(S_j) \\ +1 \text{ if } S_i\, expressed, S_j\, not\, expressed, prob(S_i) < prob(S_j) \\ +1 \text{ if } S_i\, not\, expressed, S_j\, not\, expressed \end{cases} \quad (1)$$

Following the main idea of the SA algorithm, the energy function E is defined as the sum of the energy from all the influenced nodes. These influences are calculated multiplying the value of each node (S_i) by the associated weights with all nodes interconnected with the first one (W_{ij}). Equation (2) summarizes this computation:

$$E = -\sum_i \sum_j \mathrm{W}_{ij} S_i S_j \qquad (2)$$

An interaction between two proteins happens when both of them are expressed. In our case, the energy decreases for those node connections that involve an interaction between proteins that are both expressed. Indeed, under such situation the three terms involved in equation (2) achieves their maximum values and the product also, because of the negative sign in equation (2) the energy decreases, *i.e.* this represents a favorable situation. Favorable situations also occur when the probability (S_i,S_j) of the protein expressed (to be expressed) is higher than the probability of the other one of not being expressed.

Similarly, the energy achieves minimum values when the products $W_{ij}S_iS_j$ are maxima, according to the definition of equation (2) this always occurs when those nodes connections involving an interaction between proteins that are not expressed or when the probability of the protein unexpressed for not being expressed is higher than the probability of the other one of being expressed.

Briefly, as mentioned before an overall low energy value will be related with a more stability of the network. It means that there are not so many interactions between proteins that are both expressed or at least the number of interaction between unexpressed genes is higher than the expressed ones. The more expressed protein interactions the network have, more will affect to the stability of the network if a mutation or change exists.

Chapter 5

Results

In this chapter we present different results obtained with the proposed PDSA approach, described on section 4.2. As mentioned before this study is based on the values provided by the energy function defined in equation (2), which simulates the network stability.

For each tissue, we studied the stability of the protein interaction network in tumoral and normal samples. As we have described on chapter 4, we implemented an original methodology based on the evolution of the well-known simulated annealing approach customized for study the energy and simulate the stability of a protein network. Fig.5.1 describes the steps we did during the different tests carried out during our experiments.

We simulated in two different ways the stability of the protein network: 1) mapping all the genes from the gene expression analysis stage (around 10750 genes) and 2) mapping only the overexpressed genes.

When we studied the overall expression we got significant results for all the tissues but not for breast. We realized that there were a large amount of genes in tumoral and normal cases in breast with a similar gene level expression. Including all of them in the study could be considered adding noise.

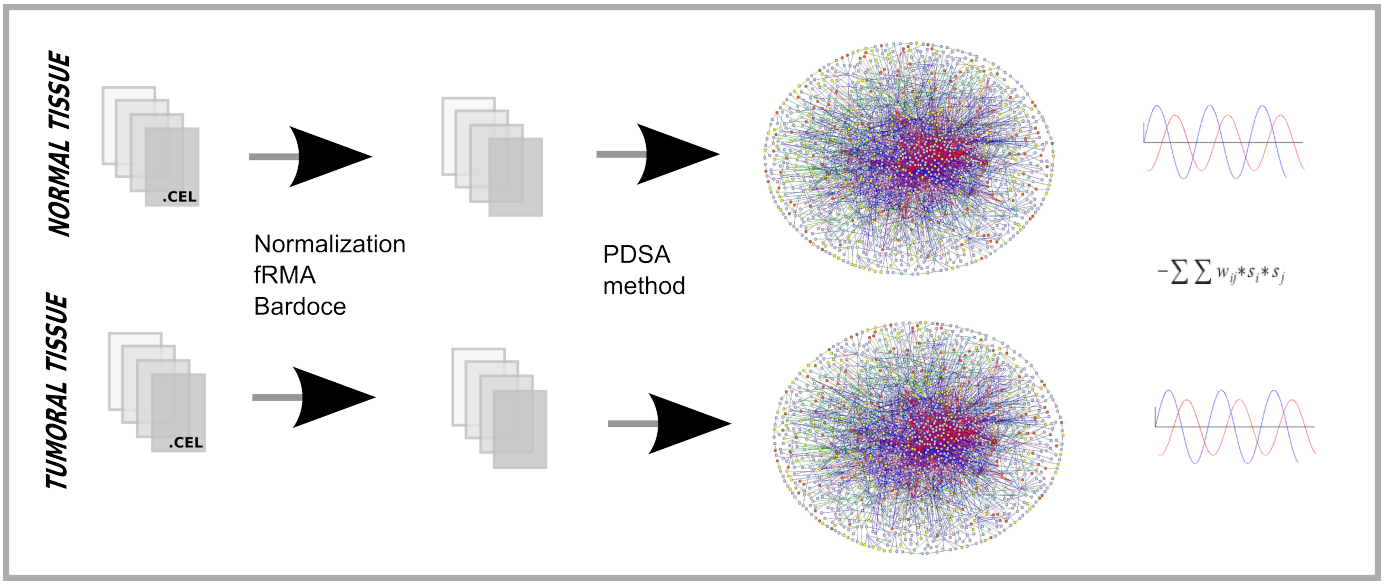

Figure 5.1: *Main schema of the overall study mapping all the genes*

This is the reason why we simulated the network with those genes overexpressed in tumoral but not in normal and *viceversa*, in order to get even more significant results in the simulation, not just for the breast tissue; we also did for the rest of tissues.

We used the algorithm for the tumor and normal pairs available for various tissues. For each interaction, we mapped into the protein-protein interaction network the gene expression analysis values and obtained the corresponding energy value, E, from the simulated network. In other words, we simulated an interaction network for every sample and studied the energy, E, of the corresponding network. The global energy in one normal tissue is the sum of all the energy values resulting after the simulations of all the normal samples; as well as the global energy in one tumor tissue is the sum of all the energy values resulting after the simulations of all the tumor samples in that tissue.

Boxplots represented by figures from Fig.5.3 to Fig.5.7 describe the distribution of the energy of the network simulated, in other words, the stability of the network (as described in the section 4.2), in normal tissues represented in blue and tumoral tissues, in red.

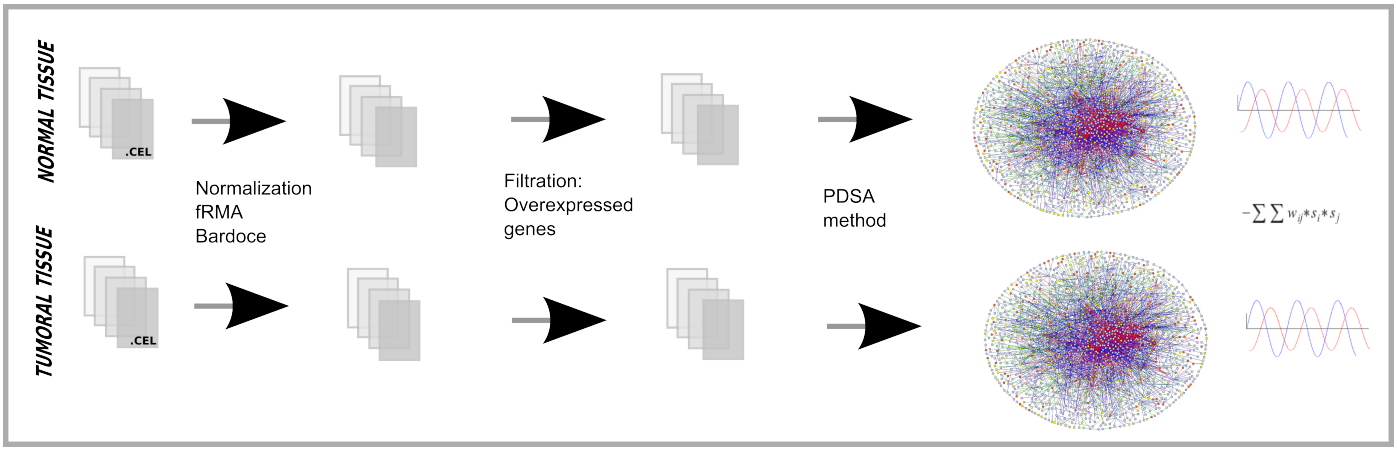

Figure 5.2: *Main schema of the overall study mapping genes overexpressed*

On one hand, the left subsection of each figure shows the resulting energy value of the simulated network in normal and tumor tissues, considering the information of all the genes. On the other hand, the subsection on the right represents the resulting energy value of the simulated network in normal and tumor tissues taking into account only those genes overexpressed in normal but not in tumor and those overexpressed in tumor and not in normal.

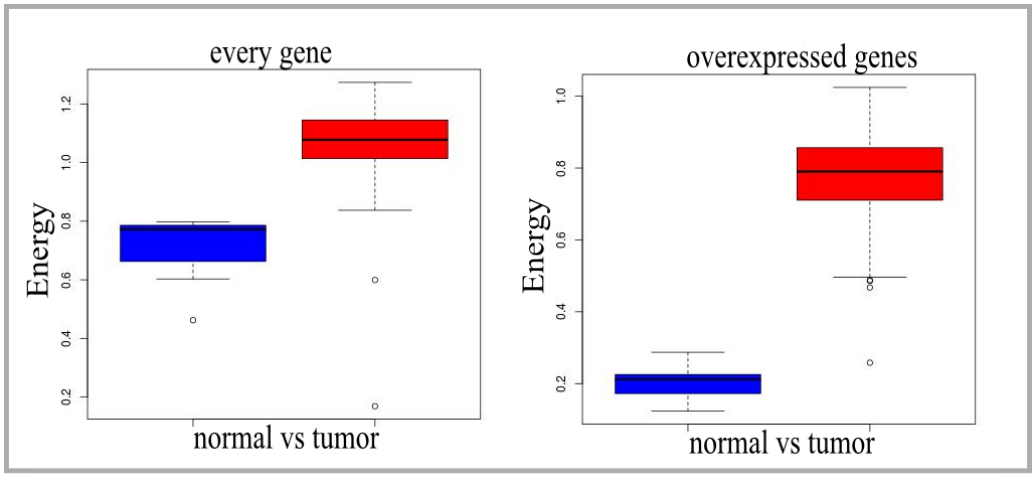

Figure 5.3: *Energy distribution mapping every genes and overexpressed genes in ovary*

For the ovary cancer data set we observe a clearly different distribution of normal and tumor tissues.

When the expression of all genes is considered, the median of the energy for the normal samples is of 0.7 while in the tumors is more than 1.0. This difference is more significant when only genes that are overexpressed are considered with a median for the normal samples is 0.2 and of 0.9 for the tumor samples.

The Mann-Whitney test evaluates whether the medians on a test variable differ significantly between two groups. We calculated a Mann-Whitney test in order to prove that these two energy distributions (in normal and tumor tissues) are significantly different: we get a p-value of 4.97e-07 when we compare the energy distributions in normal and tumor samples taking into account all the genes and a p-value of 2.8e-09 comparing both energy distributions, mapping only genes overexpressed.

The same tendency can be observed in the colon cancer data set. In this case we register a lower network energy when we map the expression of the genes in normal tissues. In this case the difference between normal and tumor tissues is even more significant. Mapping only overexpressed genes makes the differences tumor/normal even larger.

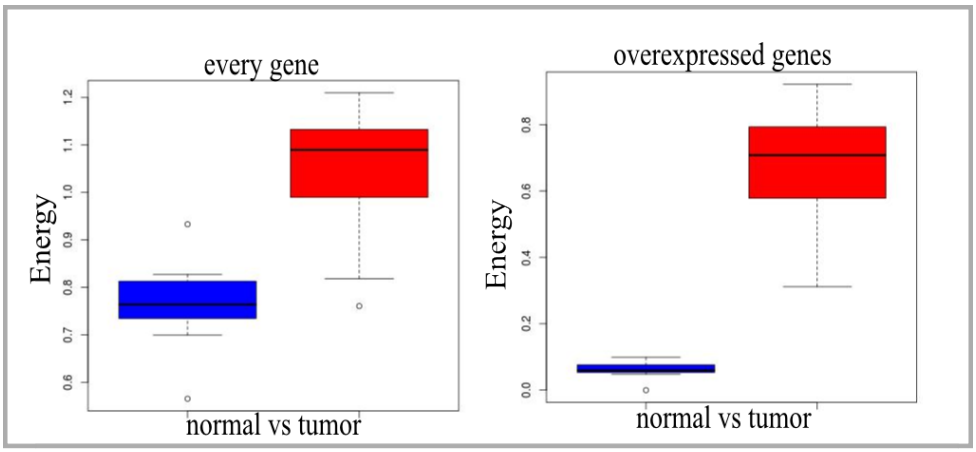

Figure 5.4: *Energy distribution mapping every genes and overexpressed genes in colon*

In summary, the energy distribution in both cases, normal and tumor, are significantly different. Not only the median is different between them, even the third quartile in the normal samples do not reach the first quartile in the tumoral samples; meaning that the energy distribution considering gene expression on normal tissues is more directed to lower energy values. We demonstrate this calculating a Mann-Whitney test in both energy distributions: a p-value of 2.33e-06 considering all the genes and a p-value of 5.11e-06 considering overexpressed genes, prove that both distributions are significantly different.

Fig.5.5 shows the energy distribution of the network on normal and tumoral cases in the kidney dataset. We can extract almost the same conclusions as we did with ovary and colon tissues. The energy distribution is different, even more considering overexpressed genes. Mapping all the genes on the network, we can observe that the median is different in normal and tumoral samples; but the third quartile in normal overlaps with the first quartile in tumoral samples. Calculating the Mann-Whitney test we obtain a p-value of 7.03e-05 taking into account all the genes and a p-value of 2.67e-15 mapping only genes overexpressed.

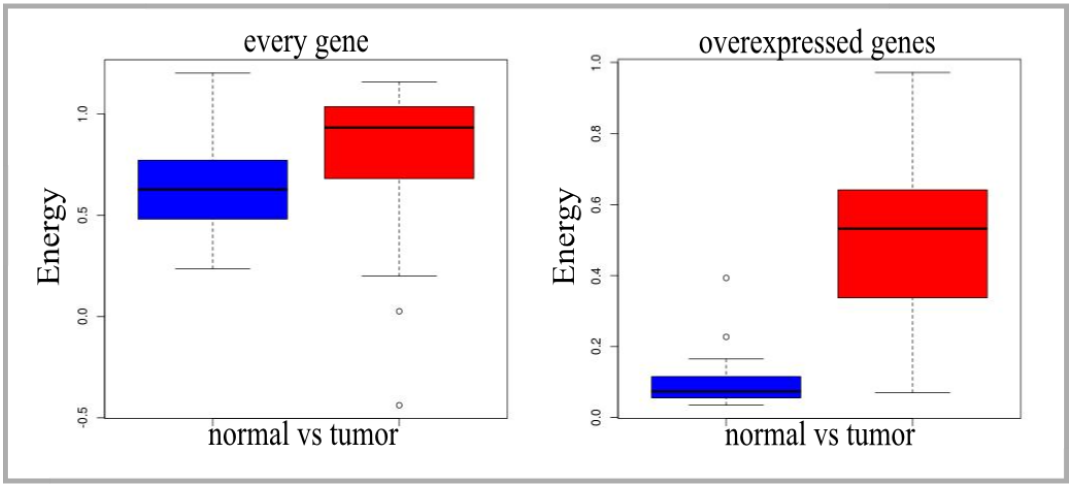

Figure 5.5: *Energy distribution mapping every genes and overexpressed genes in kidney*

The energy distribution between tumor and normal samples in the liver dataset, is significantly different as we can observe in Fig.5.6. Mapping all the genes as well as only the overexpressed genes show a total different configuration of the energy distribution. Calculating the Mann-Whitney test returns p-values of 1.9e-7 and 2.44e-14 respectivelly.

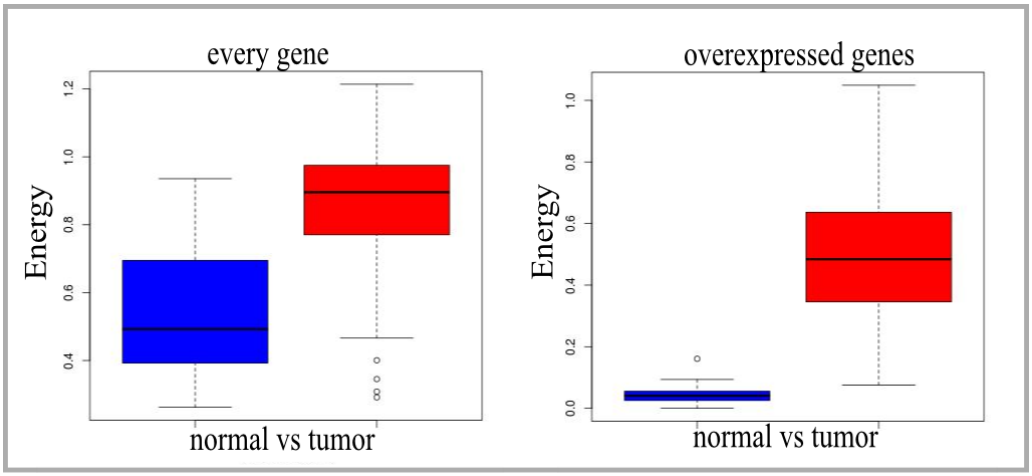

Figure 5.6: *Energy distribution mapping every genes and overexpressed genes in liver*

Fig.5.7 describes the last study. We cannot see any significant difference between normal and tumoral distributions, when we map all genes into the network. The Mann-Whitney

test calculates a p-value of 0.45, that means that both distributions are not significantly different. But we do taking into account only genes that are overexpressed in tumor but not in normal, and viceversa. The median is significantly different, as well as the entire distribution and the p-value provided by the Mann-Whitney test is around 3.81e-07.

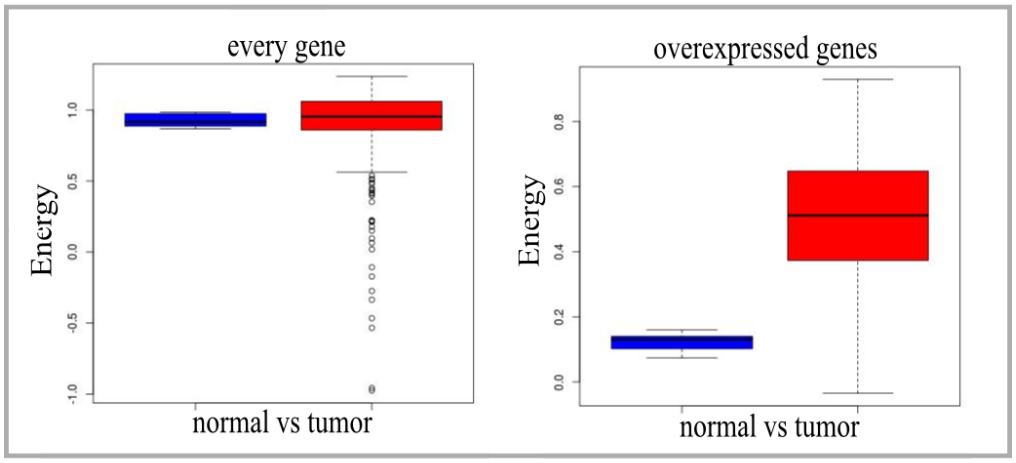

Figure 5.7: *Energy distribution mapping every genes and overexpressed genes in breast*

This different behaviour in the breast cancer dataset comparing to the previous ones can be due to the noise when we take into account the level of expression of all the genes. This might be due to a large amount of genes that have a similar gene expression level in tumor and normal breast samples. And so, when we only take into account those genes that are overexpressed in tumor and not normal samples and viceversa, we can observe a significant differences between both simulated networks' energy distribution.

Chapter 6

Conclusions and future work

6.1 Conclusions

Our main idea was to prove that proteins related with cancer exhibit a different network dynamic structure than proteins not related with cancer. We used gene expression data in order to see how the expression of each gene and its interactions along the network have influence on the total interactions system.

The main contribution of this work consists in the design of an approach based on the simulated annealing theory, representing the protein interaction network as a system of nodes with their states being dynamically updated toward a global stable state for the network. Our proposed strategy is inspired on the definition of an energy function, whereas in the general deterministic approach measures the stability of the network: less energy is related to more stability.

The network stability in tumor and normal samples in ovary, breast, colon, kidney and liver experimental data sets is significantly different. The distribution of the energy of the simulated network in the samples is significantly different, emphasizing lower energy values and so more stability in normal tissues than in cancer related samples.

A reason that could explain this behaviour would be the chaos that any tumor causes: the lack of control on the cell growth. Cancer is fundamentally a disease of failure of regulation of tissue growth. In order for a normal cell to transform into a cancer cell, the genes which regulate cell growth and differentiation must be altered. So, tumor samples represent a major destabilization because of an increased number of proteins that are expressed and are interacting with others that are also expressed.

For every tissue we have studied, except breast, we have seen a significantly different energy distribution in normal and tumor samples, taking into account the level of expression of all the genes. Along the same line, when we map the level of expression of genes that are overexpressed we have seen in every tissue even a more significant difference between tumor and normal energy distributions. An interpretation of this could be that lots of genes have a similar expression profile in normal and tumor samples, but those genes that are differentially overexpressed between normal and tumor are the responsible to represent a different dynamic structure of the protein interaction network.

An interpretation of this could be that a large amount of genes have a similar expression profile in normal and tumor samples, but those genes that are differentially overexpressed between normal and not tumor and *viceversa*, are the responsible to represent a different dynamic structure of the protein interaction network.

We propose this new methodology to simulate a dynamical protein structure and so study the stability of the protein interaction network based on gene expression.

6.2 Future Work

Based on the results obtained we propose the following considerations in order to continue this lines of the research.

We have seen that the energy between normal and tumor data samples is different after simulating a dynamical structure of the protein interaction network for each sample. But we do not know what happens in the network, whether this result is derived of an increase number of hubs (proteins with many interaction partners). Maslov *et al.* (2011) show that if a given number of proteins and distributed interactions among them are taken randomly, it is hardly find any particular protein that would have a lot of interactions. Proteins would all "talk" randomly with each other in such a network. So, hubs of highly-interacting proteins are not something that would be expect to happen by pure chance.

In this work we have considered as statics the relations between proteins. Another interesting idea is to implement a methodology that considers also the influence of the protein interactions. Vandin *et al.* (2012) propose something similar in order to find subnetworks of genes in an gene interaction network where the mutational status of these genes in the subnetwork are significantly associated with a phenotype. For that, they design an algorithm, HotNet, that finds groups of mutated genes using a heat diffusion model and two-stage statistical test. The idea would be to design a heat diffusion model so as to calculate the overall interactions between genes, instead of considering only the direct connections between them.

A straightaway idea in which we are currently working on is to apply also copy number variation (CNV) data in our PDSA approach. Chromosomal DNA copy number is the number of copies of genomic DNA. Normal somatic cells have two copies of the autosomal chromosomes; the copy number is two. In addition, cells of normal males have one copy of

the X and of the Y chromosome. Nuclei of Down syndrome patients show an extra copy of chromosome 21, whereas in cancer tissue the copy number may vary considerably over the genome. Chromosomal aberrations are a key event in the developmental and progression of cancer (Lengauer *et al.* 1998). And array comparative genomic hybridization (aCGH) is a high resolution method to detect these DNA copy numbers (Pinkel and Albertson, 2005).

Analyze this kind of data and remodelate our PDSA methodology is our most immediate proposal. We are going to study how protein interaction network's energy changes based on the gene expression data and the number of copies of each gene.

Bibliography

[1] Bioconductor - AnnotationData Packages:
http://www.bioconductor.org/packages/release/data/annotation/

[2] Duda, R.E., Hart, P.E. y Stork, D.G.: Pattern Classification. John Willey and Sons, New York (2001)

[3] Gene Expression Omnibus (GEO):http://www.ncbi.nlm.nih.gov/geo/

[4] Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)

[5] Jonsson P.F. and Paul A. Bates: Global topological features of cancer proteins in the human interactome. Bioinformatics (2006)

[6] Laarhoven van, P.J.M. y Aarts, E.H.L.: Simulated Annealing: theory and applications. Kluwer Academic Publishers, Boston, MA (1988)

[7] engauer C, Kinzler K. and Vogelstein B. : Genetic instabilities in human cancers. Nature. 396:623-27

[8] Lin J. et al.: A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Research (2007)

[9] Maslov S., Heo M. and Shakhnovich E.: Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci U S A. 2011 March 8; 108(10): 4258–4263. 3

[10] McCall et al.: The Gene Expression Barcode: leveraging public data repositores to begin cataloging the human and murine transcriptomes. Nucleic Acids Research. 2011 Jan;39(suppl 1):D1011-D1015

[11] McCall Matthew N.: Preprocessing and Barcoding of Single Microarrays and Microarray Batches (frma). Tutorial document (2011)

[12] McCall Matthew N., Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry: The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Research (2010)

[13] urphy D.: Gene Expression Studies Using Microarrays: Principles, Problems and Prospects Advan in Physiol Edu 26:256-270, 2002. doi:10.1152/advan.00043.2002

[14] Nagel S.,Schneider B.,Meyer C.,Kaufmann M.,Drexler H.G., MacLeod R.A.F: Transcriptional deregulation of homeobox gene ZHX2 in Hodgkin lymphoma Elsevier (2012)

[15] NCBI: http://www.ncbi.nlm.nih.gov/

[16] inkel D. and Albertson D: Array comparative genomic hybridization and its application in cancer. Nature Genetics, 20:207-11 (2005)

[17] łuciennika E.,Kusińskab R.,Potemskic P.,Kubiakb R.,Kordekb R.,Bednarek A.K.: WWOX—the FRA16D cancer gene: Expression correlation with breast cancer progression and prognosis Elsevier Volume 32, Issue 2, March 2006, Pages 153–157

[18] Vandin F, Clay P, Upfal E, Raphael BJ.: Discovery of mutated subnetworks associated with clinical data in cancer. Pac Symp Biocomput (2012)

[19] Wachi S. et al.: Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics (2005)

[20] Wu J, Vallenius T, Ovaska K, Westermack J, Mäkelä TP, Hautaniemi S.: Integrated Network Analysis Platform for protein-protein interactions. Nature methods (2009)

[21] Pellegrini M., Haynor D., Johnson J.M.: Protein Interaction Network. Expert Rev. Proteomics (2004)

[22] Zilliox Michael J. and Rafael A. Irizarry: A Gene Expression Barcode for Microarray Data. Nat Methods (2007) http://rafalab.jhsph.edu/barcode (available online)

Appendix A

Appendix A - Gene Expression Samples

The following samples have been used in the study:

- **Ovary**
  - **Normal samples**: GSM80757, GSM80758, GSM80759, GSM80780, GSM175789, GSM176131, GSM176136, GSM176237, GSM176318.
  - **Tumoral samples**: GSM38064, GSM38065, GSM38066, GSM38070, GSM38071, GSM38088, GSM38095, GSM46815, GSM46821, GSM46830, GSM46831, GSM46839, GSM46853, GSM46886, GSM46897, GSM46898, GSM46910, GSM46911, GSM46918, GSM46925, GSM249675, GSM249676, GSM249677, GSM249678, GSM249714, GSM249715, GSM249716, GSM249717, GSM249718, GSM249719, GSM249720, GSM249721, GSM249722, GSM249723, GSM249724, GSM249725, GSM249726, GSM249727, GSM249728, GSM249729, GSM249730, GSM249731, GSM249732, GSM249733, GSM249734, GSM249735, GSM249736, GSM249737, GSM249738, GSM249739, GSM249740, GSM249741, GSM249742, GSM249743, GSM249744, GSM249745, GSM249746, GSM249747, GSM249749, GSM249750, GSM249751, GSM249752, GSM249753, GSM249754, GSM249755, GSM249756, GSM249757, GSM249758, GSM249759, GSM249760, GSM249762, GSM249763, GSM249766, GSM249767, GSM249769, GSM249770, GSM249773, GSM249774, GSM249775, GSM249776, GSM249777, GSM249778, GSM249779, GSM249780, GSM249781,

GSM249782, GSM249783, GSM249784, GSM249785, GSM249786, GSM249788, GSM249789, GSM249790, GSM249791, GSM249792, GSM249793, GSM249794, GSM249795, GSM249796, GSM249797, GSM249798, GSM249799, GSM249801, GSM249802, GSM249803, GSM249804, GSM249805, GSM249807, GSM249808, GSM249809, GSM249811, GSM249812, GSM249815, GSM249816, GSM249817, GSM249818, GSM249819, GSM249820, GSM249821, GSM249822, GSM249824, GSM249825, GSM249826, GSM249827, GSM249830, GSM249832, GSM249833, GSM249835, GSM249836, GSM249837, GSM249838, GSM249839, GSM249840, GSM249841, GSM249842, GSM249844, GSM249845, GSM249846, GSM249847, GSM249848, GSM249849, GSM249850, GSM249851, GSM249852, GSM249853, GSM249854, GSM249855, GSM249856, GSM249857, GSM249858, GSM249859, GSM249860, GSM249861, GSM249862, GSM249863, GSM249866, GSM249868, GSM249869, GSM249870, GSM249871, GSM249872, GSM249874, GSM249875, GSM249876, GSM249877, GSM249878, GSM249879, GSM249880, GSM249881, GSM249882, GSM249883, GSM249884, GSM249885, GSM249886, GSM249888, GSM249889, GSM249890, GSM249892, GSM249893, GSM249894, GSM249895, GSM249897, GSM249898, GSM249899, GSM249900, GSM249901, GSM249902, GSM249903, GSM249904, GSM249905, GSM249906, GSM249907, GSM249908, GSM249909, GSM249910, GSM249912, GSM249913, GSM249915, GSM249916, GSM249917, GSM249918, GSM249919, GSM249920, GSM249922, GSM249923, GSM249924, GSM249925, GSM249926, GSM249927, GSM249928, GSM249929, GSM249930, GSM249932, GSM249933, GSM249934, GSM249936, GSM249939, GSM249940, GSM249941, GSM249942, GSM249943, GSM249946, GSM249948, GSM249950, GSM249951, GSM249953, GSM249954, GSM249955, GSM249956, GSM249957, GSM249958, GSM249959, GSM249960, GSM249961, GSM249962, GSM249963, GSM249965, GSM249966, GSM249967, GSM249968, GSM249969, GSM249971, GSM249972, GSM249973, GSM249974, GSM249975, GSM249976,

GSM249977, GSM249978, GSM249980, GSM249981, GSM249983, GSM249984, GSM249985, GSM249986, GSM249987, GSM249988, GSM249990, GSM249992, GSM249993, GSM249994, GSM249995, GSM249996, GSM249998, GSM249999, GSM250000, GSM250001.

- **Breast**
  - **Normal samples**: GSM85513, GSM85514, GSM85515, GSM85516, GSM85517, GSM85518, GSM85519, GSM175792, GSM175795.
  - **Tumoral samples**: GSM278156, GSM278157, GSM278158, GSM278159, GSM278160, GSM278161, GSM278162, GSM278163, GSM278164, GSM278165, GSM278166, GSM278167, GSM278168, GSM278169, GSM278170, GSM278171, GSM278172, GSM278173, GSM278174, GSM278175, GSM278176, GSM278177, GSM278178, GSM278179, GSM278180, GSM278181, GSM278182, GSM278183, GSM278184, GSM278185, GSM38051, GSM38054, GSM38057, GSM38059, GSM38062, GSM38063, GSM38080, GSM38081, GSM38082, GSM38083, GSM38086, GSM38090, GSM38091, GSM38092, GSM38094, GSM38099, GSM38102, GSM38106, GSM38109, GSM38110, GSM46816, GSM46820, GSM46827, GSM46836, GSM46846, GSM46849, GSM46852, GSM46855, GSM46859, GSM46862, GSM46863, GSM46869, GSM46870, GSM46871, GSM46873, GSM46874, GSM46880, GSM46883, GSM46885, GSM46890, GSM46891, GSM46893, GSM46894, GSM46900, GSM46905, GSM46908, GSM89102, GSM85473, GSM85474, GSM85475, GSM85476, GSM85477, GSM85478, GSM85479, GSM85480, GSM85481, GSM85482, GSM85483, GSM85484, GSM85485, GSM85486, GSM85487, GSM85488, GSM85489, GSM85490, GSM85491, GSM85492, GSM85493, GSM85494, GSM85495, GSM85496, GSM85497, GSM85498, GSM85499, GSM85500, GSM85501, GSM85502, GSM85503, GSM85504, GSM85505, GSM85506, GSM85507, GSM85508, GSM85509, GSM85510, GSM85511, GSM85512, GSM124994, GSM124995, GSM124996, GSM124997, GSM124998, GSM124999, GSM125000, GSM125001, GSM125002,

GSM125003, GSM125004, GSM125005, GSM125006, GSM125007, GSM125008, GSM125009, GSM125010, GSM125011, GSM125012, GSM125013, GSM125014, GSM125015, GSM125016, GSM125017, GSM125018, GSM125019, GSM125020, GSM125021, GSM125022, GSM125023, GSM125024, GSM125025, GSM125026, GSM125027, GSM125028, GSM125029, GSM125030, GSM125031, GSM125032, GSM125033, GSM125034, GSM125035, GSM125036, GSM125037, GSM125038, GSM125039, GSM125040, GSM125041, GSM125042, GSM125043, GSM125044, GSM125045, GSM125046, GSM125047, GSM125048, GSM125049, GSM125050, GSM125051, GSM125052, GSM125053, GSM125054, GSM125055, GSM125056, GSM125057, GSM125058, GSM125059, GSM125060, GSM125061, GSM125062, GSM125063, GSM125064, GSM125065, GSM125066, GSM125067, GSM125068, GSM125069, GSM125070, GSM125071, GSM125072, GSM125073, GSM125074, GSM125075, GSM125076, GSM125077, GSM125078, GSM125079, GSM125080, GSM125081, GSM125082, GSM125083, GSM125084, GSM125085, GSM125086, GSM125087, GSM125088, GSM125089, GSM125090, GSM125091, GSM125092, GSM125093, GSM125094, GSM125095, GSM125096, GSM125097, GSM125098, GSM125099, GSM125100, GSM125101, GSM125102, GSM125103, GSM125104, GSM125105, GSM125106, GSM125107, GSM125108, GSM125109, GSM125110, GSM125111, GSM125112, GSM125113, GSM125114, GSM125115, GSM125116, GSM125117, GSM125118, GSM125121, GSM125122, GSM151259, GSM151260, GSM151261, GSM151262, GSM151263, GSM151264, GSM151265, GSM151266, GSM151267, GSM151268, GSM151269, GSM151270, GSM151271, GSM151272, GSM151273, GSM151274, GSM151275, GSM151276, GSM151277, GSM151278, GSM151279, GSM151280, GSM151281, GSM151282, GSM151283, GSM151284, GSM151285, GSM151286, GSM151287, GSM151288, GSM151289, GSM151290, GSM151291, GSM151292, GSM151293, GSM151294, GSM151295, GSM151296, GSM151297, GSM151298, GSM151299, GSM151300, GSM151301, GSM151302,

GSM151303, GSM151304, GSM151305, GSM151306, GSM151307, GSM151308, GSM151309, GSM151310, GSM151311, GSM151312, GSM151313, GSM151314, GSM151315, GSM151316, GSM151317, GSM151318, GSM151319, GSM151320, GSM151321, GSM151322, GSM151323, GSM151324, GSM151325, GSM151326, GSM151327, GSM151328, GSM151329, GSM151330, GSM151331, GSM151332, GSM151333, GSM151334, GSM151335, GSM151336, GSM151337, GSM151338, GSM151339, GSM151340, GSM151341, GSM151342, GSM151343, GSM151344, GSM151345, GSM232194, GSM232195, GSM232196, GSM232197, GSM232198, GSM232199, GSM232200, GSM232201, GSM232202, GSM232203, GSM232204, GSM232205, GSM232206, GSM232207, GSM232208, GSM232209, GSM232210, GSM232211, GSM232212, GSM232213, GSM232214, GSM232215, GSM232216, GSM232217, GSM232218, GSM232219, GSM232220, GSM232221, GSM232222, GSM232223, GSM232224, GSM232225, GSM232226, GSM232227, GSM232228, GSM232229, GSM232230, GSM232231, GSM232232, GSM232233, GSM232234, GSM232235, GSM232236, GSM232237, GSM232238, GSM232239, GSM232240, GSM232241, GSM232242, GSM232243, GSM232244, GSM232245, GSM232246, GSM232247, GSM232248, GSM232249, GSM232250, GSM232251, GSM232252, GSM232253, GSM232254, GSM232255, GSM232256, GSM232257, GSM232258, GSM232259, GSM232260, GSM232261, GSM232262, GSM232263, GSM232264, GSM232265, GSM232266, GSM232267, GSM232268, GSM232269, GSM232270.

- **Colon**
  - **Normal samples**: GSM95473, GSM95474, GSM95475, GSM95476, GSM95477, GSM95478, GSM95479, GSM95480, GSM175905.
  - **Tumoral samples**: GSM38055, GSM38061, GSM38074, GSM38075, GSM38089, GSM38105, GSM38107, GSM46823, GSM46832, GSM46841, GSM46845, GSM46857, GSM46861, GSM46864, GSM46865, GSM46877, GSM46878, GSM46887, GSM46895,

GSM46899, GSM46915, GSM46921, GSM46924, GSM89103.

- **Liver**
  - **Normal samples**: GSM279063, GSM279064, GSM279065, GSM80728, GSM80729, GSM80730, GSM80739, GSM138595, GSM138596, GSM155919, GSM155926, GSM155927, GSM155928, GSM155947, GSM155948, GSM155961, GSM155964, GSM155988, GSM155989, GSM176332, GSM176333, GSM176334, GSM176335.
  - **Tumoral samples**: GSM38078, GSM38108, GSM46848, GSM139131, GSM143545, GSM143546, GSM143547, GSM143548, GSM143549, GSM143550, GSM143551, GSM143552, GSM143553, GSM248688, GSM248689, GSM248690, GSM248691, GSM248692, GSM248693, GSM248694, GSM248695, GSM248696, GSM248697, GSM248698, GSM248699, GSM248700, GSM248701, GSM248702, GSM248703, GSM248704, GSM248705, GSM248706, GSM248707, GSM248708, GSM248709, GSM248710, GSM248711, GSM248712, GSM248713, GSM248714, GSM248715, GSM248716, GSM248717, GSM248718, GSM248719, GSM248720, GSM248721, GSM248722, GSM248723, GSM248724, GSM248725, GSM248726, GSM248727, GSM248728, GSM248729, GSM248730, GSM248731, GSM248732, GSM248733, GSM248734, GSM248735, GSM248736, GSM248737, GSM248738, GSM248739, GSM248740, GSM248741, GSM248742, GSM248743, GSM248744, GSM248745, GSM248746, GSM248747, GSM248748, GSM248749, GSM248750, GSM248751, GSM248752, GSM248753, GSM248754, GSM248755, GSM248756, GSM248757, GSM248758, GSM248759, GSM248760, GSM248761, GSM248762, GSM248763, GSM248764, GSM248765, GSM248766, GSM248767, GSM248768, GSM248769, GSM248770, GSM248771, GSM248772, GSM248773, GSM248774, GSM248775, GSM248776, GSM248777, GSM248778.

- **Kidney**
  - **Normal samples**: GSM279060, GSM279061, GSM279062, GSM281311, GSM281312,

GSM281314, GSM281315, GSM281316, GSM175911, GSM198783, GSM198785, GSM240832, GSM240833, GSM240834, GSM240835, GSM240836, GSM240837, GSM240838, GSM240839, GSM240840, GSM240841, GSM240842, GSM240843, GSM240844, GSM240947, GSM240948.

- **Tumoral samples**: GSM281278, GSM281279, GSM281280, GSM281281, GSM281282, GSM281283, GSM281284, GSM281285, GSM281286, GSM281287, GSM281288, GSM281289, GSM281290, GSM281291, GSM281292, GSM281293, GSM281294, GSM281295, GSM281296, GSM281297, GSM281298, GSM281299, GSM281300, GSM281301, GSM281302, GSM281303, GSM281304, GSM281305, GSM281306, GSM281307, GSM281308, GSM281309, GSM281310, GSM281313, GSM281317, GSM281318, GSM281319, GSM281320, GSM281321, GSM281322, GSM281323, GSM281324, GSM281325, GSM281326, GSM281327, GSM281328, GSM281329, GSM281330, GSM281331, GSM281332, GSM281333, GSM281334, GSM281335, GSM281336, GSM281337, GSM281338, GSM281339, GSM281340, GSM281341, GSM281342, GSM281343, GSM281344, GSM305099, GSM305100, GSM305101, GSM305102, GSM305103, GSM305104, GSM305105, GSM305106, GSM305107, GSM305108, GSM305109, GSM305110, GSM305111, GSM305112, GSM305113, GSM305114, GSM305115, GSM305116, GSM38073, GSM46825, GSM46826, GSM46847, GSM46858, GSM46875, GSM46881, GSM46882, GSM46892, GSM89104.

Appendix B

Pseudocode for the Deterministic Simulated Annealing Method

Description of the deterministic simulated annealing:

begin

$initialize\ E, W_{ij}, S_i, S_j, i, j = 1..N$

$t = 0$

while $t <= tmax$

$$E = -\sum_{i}^{N} \sum_{j}^{N} \mathrm{W}_{ij} S_i S_j$$

$t = t + 1$

end

end

whereas:

E is the system's final energy. t corresponds to a gene expression sample in each tissue and type (normal or tumor). W_{ij} describes the weight explained in Equation 1, representing the existing influence between nodes S_i and S_j. In our approach S_i is the probability of each gene of being expressed.