MÉTODOS BAYESIANOS
para la estimación de redes reguladoras de genes y de perfiles de proteínas a partir de
MICROARRAYS DE EXPRESIÓN GENÉTICA

Tesis doctoral

Doctorando:

Manuel SÁNCHEZ CASTILLO

Directores:
Mª del Carmen CARRIÓN PÉREZ
Isabel Mª TIENDA LUNA
David BLANCO NAVARRO

Departamento de Física Aplicada
Universidad de Granada

Department of Applied Physics
University of Granada

Bayesian methods for inferring gene regulatory networks and protein profiles from gene expression microarrays data

Thesis for obtaining the PhD mention at the Universisty of Granada
within the official program titled *Doctorado en Física y Ciencias del Espacio*

Author:
Manuel Sánchez Castillo

Supervisors:
María del Carmen Carrión Pérez[†]
Isabel MaríaTienda Luna[††]
David Blanco Navarro[†]

[†] Department of Applied Physisc, University of Granada
[††] Department of Electronics and Computer Technology, University of Granada

Índice general

Capítulo 1

Introducción

La herencia genética que un ser vivo transmite a su descendencia está almacenada en macromoléculas de ácidos nucleicos en el interior de las células. En organismos procariotas, esta información se almacena en moléculas de ácido desoxirribonucleico (ADN), que codifican la síntesis de proteínas que regulan el metabolismo celular. La parte del código que controla por completo la síntesis de una proteína se denomina gen y constituye la unidad de almacenamiento de información hereditaria [22].

El ADN almacena la información genética pero no es el responsable directo de la síntesis proteica. La información codificada en un gen se transcribe a una molécula funcional de ácido ribonucleico (ARN) que participa activamente en el metabolismo celular. Cuando la información almacenada en un gen es transcrita y finalmente traducida a una proteína, se dice que el gen se ha expresado. Este mecanismo de codificación, transmisión y traducción de la herencia genética se conoce como dogma central de la Biología Molecular.

Todas las células de un organismo contienen la misma información genética. La diferenciación y el metabolismo celular quedan controlados por la actuación conjunta de los genes expresados. La expresión genética es un proceso complejo en el que, a través de distintos mecanismos reguladores, una gran diversidad de biomoléculas se asocian e interaccionan entre sí para producir respuestas diferentes. Entre estos mecanismos reguladores se encuentran: la iniciación de la transcripción del ADN, la maduración del ARN, las modificaciones postranslacionales y la degradación proteica.

Durante la última década del siglo XX, la Genética ha avanzado enormemente gracias al desarrollo de la técnica del microarray. Un microarray de expresión es un ensayo experimental que permite cuantificar simultáneamente la expresión de un gran número de genes [55]. Mientras que otros

procedimientos experimentales clásicos sólo permiten cuantificar el estado de expresión para un numero limitado de genes, los microarrays son capaces de obtener el perfil de expresión de un genóma completo. A este gran avance se suman las técnicas de secuenciación de nueva generación, que han permitido secuenciar el ADN al completo de organismos tan complejos como el ser humano, con más de 20000 genes. Estas tecnologías abren nuevas perspectivas en el estudio del proceso de regulación genética [2].

Inicialmente, el análisis de micorarrays de expresión se limitaba al cálculo de estadísticos descriptivos y a la clasificación de genes con patrones de expresión similares [8]. Este tipo de estudios resulta útil para identificar qué genes comparten funciones y controlan el metabolismo celular. Otro tipo de análisis más ambicioso trata de modelar el mecanismo de regulación a diferentes escalas y estimar sus propiedades a partir de los datos [90] [87]. Para ello, es necesario el desarrollo de modelos con una base matemática robusta y la implementación de métodos de baja complejidad computacional capaces de procesar grandes cantidades de información. Este tipo de análisis exige un trabajo multidisciplinar que integra la rama de las ciencias biomédicas con las ciencias de la información.

1.1 Intereses y motivaciones

El potencial de la técnica del microarray reside en su capacidad para cuantificar simultáneamente el estado de expresión de un gran número de genes. Los microarrays de expresión constituyen una firma molecular que caracteriza el estado de una célula. Este tipo de información resulta de especial interés por que permite comparar diferentes perfiles moleculares y analizar sus diferencias a nivel genético, lo que ha favorecido su difusión y su implantación como herramienta estándar en la investigación biomolecular [99] [45].

El análisis de los datos de microarray difiere en función del enfoque del estudio. Por ejemplo, en Biología de Sistemas resulta interesante estudiar la regulación genética desde un punto de vista fenomenológico, en el que los genes son los responsables directos de los procesos celulares. Un modelo muy común usado en este tipo de descripciones son las redes de regulación genéticas [101]. Una red reguladora de genes es un modelo abstracto en el que sus elementos, los genes, interaccionan para producir diferentes respuestas. Esta descripción permite a los investigadores agrupar los genes atendiendo a sus funciones metabólicas en los distintos procesos celulares y estudiar los orígenes de la diferenciación celular a nivel genético.

Por otro lado, desde el punto de vista de la Biología Molecular es interesante conocer el mecanismo de regulación a un nivel más detallado. En lugar de analizar las funciones de los genes durante el desarrollo celular, este

enfoque trata de explicar la expresión genética a niveles moleculares, donde los genes son un elemento más de un mecanismo de regulación que integra diferente tipo de biomoléculas [60]. Desde esta perspectiva, se estudia cómo la célula es capaz de integrar diferentes señales químicas y el modo en el que estas interaccionan con los genes para producir diferentes respuestas. Por ejemplo, la transcripción genética está controlada por la actuación conjunta de una serie de proteínas con funciones reguladoras, conocidas como factores de transcripción.

Conocer la maquinaria celular a diferentes niveles funcionales y moleculares resulta de especial interés en la investigación médica y en la industria farmacológica [104] [83]. En concreto, este tipo de información es útil en el diagnóstico y pronóstico de enfermedades con una firma molecular característica. Además, permite analizar las respuestas moleculares a diferentes medicamentos y es útil en el diseño de tratamientos personalizados.

En esta tesis se aborda el análisis de datos de microarray mediante estos dos enfoques. En primer lugar, se propone un análisis de los datos de expresión a nivel genético en el que se modela la red reguladora de genes y se propone un método para su aprendizaje a partir de los datos. En segundo lugar, se trata el problema de la regulación genética a nivel transcripcional, en cual se modela cómo interaccionan los genes con los factores de transcripción y se propone un método para aprender esta estructura a partir de datos de microarray.

1.2 Antecedentes

Inicialmente, la técnica del microarray era un ensayo experimental exclusivo de grandes laboratorios equipados con un instrumental específico y los análisis realizados se limitaban a una mera descripción de los datos. El desarrollo de protocolos experimentales estándares y su comercialización ha favorecido la rápida difusión de esta técnica y la aparición de numerosas bases de datos de expresión genética.

El análisis de datos de microarray ha avanzado relativamente poco en comparación con las mejoras tecnológicas conseguidas en esta técnica experimental [45]. En la Figura 1.1 se muestra el número de publicaciones anuales indexadas que incluyen el término *microarray* en el área de Bioquímica, Genética y Biología Molecular y en Ciencias Computacionales durante la última década, donde puede comprobarse que el número de publicaciones en las ramas experimentales es mucho mayor que en las del análisis de los datos.

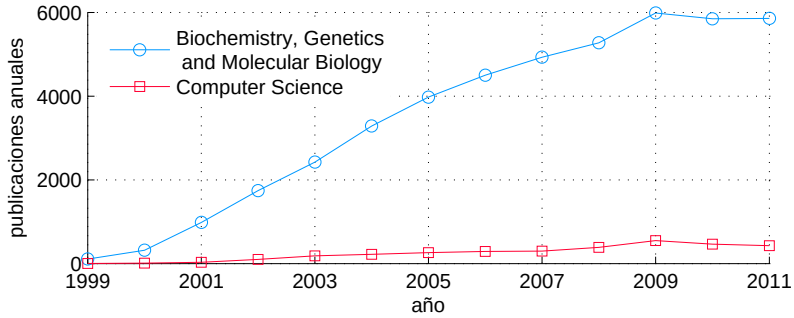

Figura 1.1: Número de publicaciones anuales indexadas que incluyen el término *microarray* en el área de Bioquímica, Genética y Biología Molecular y en Ciencias Computacionales.

El análisis de los datos de microarray se suele enmarcar en el formalismo de la Inferencia Estadística, en el cual los datos se consideran observaciones de un proceso en el que existen variable no observables. Dependiendo del análisis, estas variables ocultas describen diferentes propiedades de interés del mecanismo de regulación genética.

1.2.1 Análisis de microarrays para la inferencia de redes reguladoras de genes

Entre las primeras aplicaciones del análisis de datos de microarrays destacan los problemas de ingeniería inversa para la estimación de las redes de regulación genética [90]. Estos modelos proyectan los distintos mecanismos de regulación de la expresión al espacio de la interacción genética. En una red reguladora se considera que la expresión de un gen determinado depende directamente de la expresión de otros, entre los que se establecen una serie de relaciones causales. En los problemas de ingeniería inversa se plantean métodos de inferencia que permiten estimar las redes reguladoras a partir de datos de microarrays expresión.

El primer método de ingeniería inversa del que se tiene constancia se basa en el modelo de redes Booleanas [48]. En una red Booleana, la expresión genética se cuantifica con dos estados posibles, activación e inhibición, y la interacción genética se describe mediante un conjunto de reglas lógicas [93]. Este modelo permite describir gráficamente la red y posee un formalismo matemático sencillo que simplifica los cálculos.

Las estimaciones basadas en el modelo de red Booleana limitan los datos de expresión a un espacio binario, cuando en realidad estos toman valores continuos que indican diferentes niveles de expresión [36]. Las redes Bayesianas pueden considerarse una extensión de las redes Boolenas al espacio real [42]. En una red Bayesiana, el estado de expresión de un gen se adapta a las medidas proporcionadas en los experimentos de microarray y permiten describir las interacciones entre genes mediante funciones de probabilidad. Los métodos de inferencia basados en redes Bayesianas gozan del formalismo Bayesiano, que permite modelar a priori otras fuentes de información biológica relevante e incluirlas junto a los datos de expresión [99]. Además de la descripción gráfica de la red reguladora, las redes Bayesianas tienen la ventaja de poder explicar la generación de los datos de microarrays mediante modelos lineales simples.

1.2.2 Análisis de microarrays para la inferencia de perfiles de proteínas

Con la aparición de bases de datos con distintos tipos de información biológica, surge la necesidad de modelar esta información y combinarla con el análisis de datos de microarrays de expresión. En concreto, las bases de datos de interacciones entre genes y proteínas permiten modelar la regulación genética a un nivel molecular más detallado, desde un enfoque transcripcional [59].

La transcripción es uno de los primeros procesos regulación en la expresión genética. Este fenómeno está mediado por una serie de proteínas funcionales que interaccionan directamente con el ADN, favoreciendo o inhibiendo el ensamblaje del complejo proteico encargado de transcribir el ADN a ARN. De manera similar a las redes genéticas, las redes de regulación transcripcional describen cómo interaccionan los genes y las proteínas durante el proceso de transcripción. Los modelos para la descripción de redes reguladoras transcripcionales comparten algunas propiedades de las redes genéticas, pero posee una formulación propia e independiente. La estructura de esta red puede estimarse mediante algoritmos de predicción de alineamiento entre las proteínas y los genes [49]. Una vez aprendida, la red transcripcional puede incorporarse al modelado de microarrays para estimar otro tipo de información biológica de interés, como la abundancia de las proteínas resultantes de la expresión genética.

La mayoría de los análisis de datos de microarray que integran redes transcripcionales consideran modelos de factores latentes. El análisis factorial encuentra una aplicación inmediata en este tipo de problemas en los que se desea proyectar los datos de expresión genética al espacio de los factores de transcripción. Las diferencias principales entre los métodos de inferencia

basados en el análisis de factores reside en cómo se modela a priori la red transcripcional y en cómo se resuelven los problemas de indeterminación propios del análisis de factores.

Las técnicas clásicas de inferencia que consideran modelos factoriales, como el análisis de componentes independientes [54] y el análisis de componentes principales [40], permiten descomponer los datos de expresión en un espacio condensado con diferentes contribuciones a la regulación genética. Sin embargo, estos métodos no tienen en cuenta las propiedades estructurales de las redes transcripcionales y no se ajustan adecuadamente a la realidad del problema [14].

En otro tipo de técnicas, como el análisis de componentes de redes [53] y la factorización no negativa [27], se asumen una serie de restricciones en la red transcripcional que permiten encontrar una solución al problema. Aunque estos métodos encuentran una aplicación directa en algunos casos concretos, estas técnicas están limitadas a conjuntos de datos que cumplen unas condiciones muy estrictas y no pueden aplicarse a problemas reales.

1.3 Dificultades y objetivos

Las técnicas clásicas de Procesado de Señal no pueden aplicarse directamente al análisis de datos de microarrays y demandan un formalismo nuevo, adaptado a las características de las variables que se modelan. Uno de los problemas que hay que abordar en el análisis de microarrays es el alto nivel de ruido que afecta a los datos. Además del ruido experimental, existen otras fuentes de errores debidos a la variabilidad inherente de los procesos biológicos [77]. Esta característica exige un modelo que se ajuste a los datos sin sobredimensionar el ruido.

El modelado de las redes genéticas y las redes transcripcionales debe capturar correctamente sus propiedades estructurales [79]. En las redes genéticas, se considera que el número de genes que regulan la expresión de otro es muy limitado. Del mismo modo, en las redes transcripcionales se sabe que las proteínas interaccionan con el ADN de manera muy específica. Por tanto, una característica común a ambos modelos es el elevado número de elementos inconexos en las redes, propiedad conocida como dispersión [87].

Por otro lado, los microarrays proporcionan datos de un gran número de variables, mientras que el número de muestras disponibles suele ser muy limitado en comparación al número de genes [101]. La dimensionalidad de los datos también afecta a la implementación del método de inferencia, que suele encontrar problemas derivados de las limitaciones del cálculo computacional. Estos problemas afectan especialmente a métodos basados en técnicas de simulación como el muestreo de Gibbs [87].

Además, hay que tener en cuenta que los datos de microarray proporcionan una visión parcial del fenómeno de regulación genética completo y que existe un límite en la cantidad de información que se puede aprender de estos datos. En este sentido, la metodología Bayesiana ofrece un formalismo adecuado para el análisis de datos de microarrays que permite incorporar otro tipo de conocimiento biológico a priori y así completar la información proporcionada por los datos de expresión [45].

Dadas la dificultades que actualmente se encuentran en el modelado de datos de microarray, en esta tesis se proponen una serie de métodos nuevos para su análisis. En primer lugar, se propone un método para resolver el problema de ingeniería inversa de la red reguladora de genes mediante el análisis de series temporales de microarrays. A continuación, se enumeran las mejoras conseguidas con este método respecto a aproximaciones anteriores:

- Se propone un modelo observacional que permite capturar eficientemente la naturaleza de los datos sin sobrestimar el ruido.
- Se desarrolla un marco de trabajo Bayesiano específico que permite modelar a priori información adicional que enriquezca los datos de expresión.
- Se implementa un método de inferencia variacional Bayesiano de bajo coste computacional basado en el algoritmo de esperanza y maximización.
- Se demuestran las mejoras introducidas con el método nuevo en la inferencia de la red reguladora de genes con datos sintéticos.
- Se aplica el método nuevo a datos reales para la inferencia de la red reguladora de genes de la levadura como sistema biológico modelo.

Por otro lado, se propone un método para resolver el problema del modelado de la red transcripcional y la inferencia de la concentración de proteínas que actúan como factores de la transcripción a partir de datos de microarray. A continuación, se enumeran las mejoras consiguidas con este método respecto a aproximaciones anteriores:

- Se propone un modelo observacional que, apoyandose en el formalismo de las transformadas wavelets, reduce el número de incógnitas del problema.
- Se propone un modelado probabilístico de la red transcripcional que permite describir eficientemente su estructura dispersa.

- Se desarrolla un marco de trabajo Bayesiano específico, que incorpora información a priori sobre la estructura dispersa de la red trasncripcional mediante el aprendizaje de bases de datos de interacciones entre genes y proteínas.

- Se implementa un método de inferencia Bayesiano basado en el muestreo de Gibbs.

- Se demuestra la capacidad del método nuevo para estimar la red transcripcional y la concentración de las proteínas con datos sintéticos y datos reales.

- Se aplica la metodología nueva a datos reales para clasificar tumores de mama en diferentes subtipos.

1.4 Estructura de la memoria

En el Capítulo 2 se presentan a modo descriptivo una serie de conceptos fundamentales de Genética y Biología Molecular, imprescindibles para comprender los mecanismos de codificación, interacción y cuantificación del material genético que condicionan el modelado que se hace en capítulos posteriores. Para detalles más específicos y definiciones formales se reseña la bibliografía.

En el Capítulo 3 se introducen los fundamentos de la metodología Bayesiana y algunas de las técnicas de la Inferencia Estadística con el fin de repasar algunos conceptos importantes, los más básicos se suponen conocidos y se pueden encontrar en la literatura, además de presentar al lector la notación usada en el resto del texto.

En el Capítulo 4 se desarrolla un método de inferencia de la red reguladora de genes para el análisis de series de microarrays temporales. La metodología que se presenta considera un modelo autoregresivo de media móvil y hace uso del método variacional Bayesiano de esperanza y maximización para la inferencia de la red de regulación genética. El método desarrollado se valida analizando los errores cometidos en la inferencia de la estructura de la red genética con datos sintéticos simulados mediante el propio modelo observacional y con datos sintéticos generados con un modelo alternativo.

En el Capítulo 5 se aplica el método variacional Bayesiano desarrollado en el Capítulo 4 a un conjunto de datos reales. En concreto, se consideran datos de un organismo modelo como la levadura, para el que se conoce la red de regulación genética en diferentes procesos metabólicos. Las estimaciones que se obtienen muestran resultados satisfactorios.

En el capítulo 6 se desarrolla un método Bayesiano basado en el modelo de factores latentes expandidos para el análisis de datos de microarrays y la inferencia de la concentración de proteínas que actúan como factores de transcripción. La metodología que se presenta hace uso de la técnica del muestreo de Gibbs para estimar la abundancia de las proteínas. El método Bayesiano que se propone incorpora información a priori sobre la red transcripcional, que se obteniene a partir del análisis de bases de datos de interacciones entre genes y proteínas. La validez del método desarrollado se comprueba con datos sintéticos simulados mediante el propio modelo observacional.

En el Capítulo 7 se aplica el método Bayesiano de factores expandidos desarrollado en el Capítulo 6 a un conjunto de datos reales. En concreto, se consideran datos de microarrays de tumores de mama para estimar la abundancia de un conjunto de proteínas relevantes en esta enfermedad. Se comprueba que el método propuesto proporciona resultados satisfactorios en la clasificación de pacientes en base a los perfiles proteicos estimados y se proponen una serie de proteínas que actúan como biomarcadores del cáncer de mama.

Finalmente, en el Capítulo 8 se discuten los resultados obtenidos, se analizan las principales aportaciones y las líneas de trabajo futuro y se presentan las principales contribuciones al campo en revistas y congresos.

Chapter 8

Disccussion of the results

The microarray technique is a versatile tool that allows to quantify gene expression. The design of biological models learned from microarray data constitutes a valuable source of biological knowledge in many fields as clinical assays and pharmacology research. We address in this thesis the problems of (*i*) modeling the gene regulatory network (GRN) and its inference and (*ii*) the modeling of the transcriptional regulatory network (TRN) and the inference of protein activity profiles from microarray data.

Microarray time series hold information about the regulatory mechanisms conducted within the cell during its metabolic development. Uncovering the GRN form microarray data is very important in Biology Systems since it allows to identify gene functions and how they interact to produce different responses. We propose in Chapter 4 the AR1MA1-VBEM method for GRN reverse engineering from microarray time series data. The AR1MA1-VBEM method considers a novel first order autoregressive and first order moving average (AR1MA1) model for microarray time series fitting. We propose a Bayesian framework endowed with a conjugate model allowing to develop a low computational cost variational Bayesian expectation-maximization (VBEM) method for the GRN inference. We compare the AR1MA1-VBEM method with other approaches using synthetic and real data and we show that our method has a better performance in the inference of the GRN topology.

Gene expression is the result of complex regulatory mechanisms where different chemical signals are integrated to produce different responses. The transcription of the genes is the earliest stage of gene expression and it is controlled by a set of functional proteins known as transcription factors (TF). The TRN can be partially learned from gene-protein interaction databases

and this valuable information can be used in the inference of the TF activity profiles during the gene transcription. We propose in Chapter 6 the BFE method for the inference of protein profiles from microarray data, using prior information from gene-protein interaction databases. The BFE method considers a Bayesian expansion factor model to fit microarray data and also includes a novel probabilistic model of the TRN. We validate the BFE method with synthetic data sets and we perform an analysis with real breast cancer data that undercover the key roles of some TF in the classification of this disease.

8.1 Conclusions

We now summaryze the main contributions of the present work in the analysis of microarray data:

- We present a formal description of the GRN and the microarray time series for the problem of GRN reverse engineering.
- We propose a novel first order autoregressive and first order moving average (AR1MA1) model that differentiates between the real expression data and its noisy observations for fitting microarray time series.
- We develop a Bayesian framework for the prior modeling of the GRN and we develop a variational Bayesian expectation and maximization method based on the AR1MA1 model (AR1MA1-VBEM) for the inference of the network.
- We infer the GRN of the yeast as a biological model during its cell cycle and we compare our results with a GRN model designed by experts.
- The AR1MA1-VBEM method shows a better performance than other methods based on linear and non-linear models, providing less number of errors and enhancing the areas under the ROC and the PR curves.
- We propose an original probabilistic model for the description of sparse variables in the TRN that allows to capture the sparse nature of the network and to incorporate prior knowledge about this structure.
- We present a formal description of the TRN and the microarray data for the problem of inferring the TFs activity profiles.
- We propose a novel expansion factor model based on the wavelets formalism that reduces the number of unknowns in the problem of modeling the TRN.

- We present an original probabilistic model based on a functional prior induced Gaussian (FIG) distribution for the description of sparse variables.

- We develop a Bayesian framework for the prior modeling of the TRN that allows to capture the sparse nature of the network and to incorporate prior knowledge about this structure.

- We develop a Bayesian Expansion Factor (BFE) method for the inference of the TFs activity profiles from microarray data based on the Gibss sampling technique.

- We demonstrate the ability of the BFE method in the classification of samples based on the estimated protein profiles.

- The BFE method shows a good performance in the inference of the activity profiles with synthetic and with real breast cancer data sets.

8.2 Publications

The work in this thesis has been accepted and published in different journals and conferences.

National conferences

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modificacion del metodo EM variacional Bayesiano para el estudio de microarrays temporales. In *XXIV Simposio nacional de la Union Cientifica Internacional de Radio, URSI 2009*, Santander, pages 136-137, 2009.

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Analisis Bayesiano de microarrays temporales para la estimacion de la red de regulacion genetica. In *XXV Simposio nacional de la Union Cientifica Internacional de Radio, URSI 2010*, Bilbao, 2010.

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modelo Factorial Bayesiano para el Aprendizaje de la Red de Regulacion Transcripcional. In *XXVII Simposio nacional de la Union Cientifica Internacional de Radio, URSI 2012*, Elche, 2012.

International conferences

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Modified variational method for genes regulatory network learning. In *2010 IEEE 10th International Conference on Signal Processing*, Beijing, China, pages 1781-1784, 2010.

- J. Meng, M. Sanchez-Castillo, I.M. Tienda-Luna, and Y. Huang. Prediction of cancer subtypes using Bayesian factor network model. In *Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2011*, California, EE.UU. , pages 995-996, 2011.

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco, M.C. Carrion-Perez. Revision of the variational Bayesian method for uncovering gene regulatory networks. In: *International Workshop on Genomic Signal Processing and Statistics, GENSIPS 2011*, Texas, USA, 2011.

- M. Sanchez-Castillo, J. Meng, I.M. Tienda-Luna, and Y. Huang. Basis expansion factor models for uncovering transcription factor regulatory networks. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, Michigan, EE.UU, 2012.

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Microarray Time Series Modeling and Variational Bayesian Method for Reverse Engineering Gene Regulatory Networks. In *International Conference on Advances in Signal and Image Processing*, Dubai, UAE, 2012.

Papers and lecture notes

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Methods and recent patents for modeling and uncovering gene regulatory networks. *Recent Patents on Signal Processing*, 2:88-95, 2012, ISSN: 1867-8211.

- M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Microarray Time Series Modeling and Variational Bayesian Method for Reverse Engineering Gene Regulatory Networks. In *Signal Processing and Information Technology, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 71-76. Springer, ISSN: 1877-6124,2012.

8.3 Future work

The main problem in the design of biological models and its learning from microarrays is that expression data holds information of just a part of the whole regulatory process. The continuously development of new experimental assays for measuring different biological features, as the protein degradation by microRNA after the gene transcription, adds new sources of biological information that may be included in this kind of analysis. As a general future line of research, we propose to exploit the potential of the Bayesian formalism to model prior biological knowledge for performing an enrichment analysis, where microarray data is fused with other kind of biological information.

Regarding the GRN reverse engineering problem, we suggest to continue working with autoregressive moving average models of higher orders, that will capture with high efficiency the regulatory mechanism from microarray time series. Additionally, we propose to modify the fixed point approximations taken during the inference of the AR1MA1-VBEM method for obtaining exact results.

For the BFE method, we propose to explore alternative analytical approaches, as the VBEM method, to avoid high computational cost inference method. Moreover, we propose to outperform the prior modeling of the TRN to incorporate prior information about the complete structure of the network.

Bibliografía

[1] Genome browser. 178

[2] P. Baldi and G. W. Hatfield. *DNA microarray and gene expression*. Cambridge University Press, 2002. 2, 11

[3] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. 103

[4] V. Barnett. *Comparative Statistical Inference*. John Wiley & Sons, 1973. 32, 34, 35, 37

[5] T. Barret. Ncbi geo: archive for functional genomics data sets. *Nucleic acids research*, 39:1005–1010, 2011. 84, 137

[6] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. 32

[7] M. J. Beal and Z. Ghahramani. The variational bayesian EM algorithm for incomplete data with application to scoring graphical model structures. *Bayesian Statistics*, 7, 2003. 33, 38

[8] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–583, 2004. 2

[9] J. M. Bernardo. *Bioestadística*. Barcelona: Vicens-Vives, 1981. 27

[10] J. M. Bernardo and A. F. Smith. *Bayesian Theory*. John Wiley & Sons, 1994. 32

[11] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006. 37, 38, 42, 172

[12] J. U. Blohmer, M. Rezai, S. Kummel, and W. Eiermann. Using the 21-gene assay to guide adjuvant chemotherapy decision-making in early-stage breast cancer: a cost-effectiveness evaluation in the german setting. *Journal on Medical Economics*, 2012. 136

[13] BreastCancer.Org. Understanding breast cancer. 132

[14] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, 101:4164–4169, 2004. 6, 95

[15] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25:21–30, 2008. 99, 100

[16] C. M. Carvalho, J. Chang, J. E Lucas, J. R. Nevins, Q. Wang, , and M. West. High-dimensional sparse factor modelling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008. 93, 96

[17] K. C. Chen, L. Calzone, A. Csikasz-Nagyan, R. Cross, B. Novak, and J. J. Tyson. Integrative analysis of cell cycle control in budding yeast. *Moelcular Biology of the Cell*, 15:3841–3862, 2004. 82

[18] S. K. Chia, V. H. Bramwell, D. Tu, and T. O. Nielsen. A 50 gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. *Clinical Cancer Research*, 18:4465–4472, 2012. 137

[19] P. Collas. The current state of chromatin immunoprecipitation. *Nucleic Acids Research*, 45:87–100, 2010. 23

[20] S. Das, D. Caragea, W. H. Hsu, and S. M. Welch, editors. *Computational Methodologies in Gene Regulatory Networks*. IGI Global, 2008. 24, 48

[21] Saccharomyces Genome Database. Saccharomyces genome database (sgd). 76, 81

[22] A. Datta and E. R. Dougherty. *Introducction to genomic signal processing with control*. CRC Press, 2007. 1, 11

[23] A. Datta, R. Pal, A. Choudhary, and E. R. Dougherty. Control approaches for probabilistic gene regulatory networks. *IEEE Signal Processing Society Magazine*, 24:54–64, 2007. 48

[24] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing, 2011. 99, 100

[25] A. de la Fuente, P. Brazhnik, and P. Mendes. Linking the genes: inferring quantitative gene networks from microarray data. *Trends in Genetics*, 18:395–398, 2002. 24

[26] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997. 19

[27] K. Devarajan. Non-negative matrix factorization: An analytical and interpretive tool in computational biology. *PLOS Computational Biology*, 4:1–12, 2008. 6, 92, 97

[28] P. Diacones and D. Ylvisaker. Conjugate pripors for exponential families. *The annals of statistics*, 7:269–281, 1979. 37

[29] M. Djordjevic. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*, 24:179–189, 2007. 26

[30] L. Elsgoltz. *Ecuaciones diferenciales y cálculo variacional*. MIR, 1969. 43, 152

[31] P. Eroles, A. Bosch, J. A. Pérez-Fidalgo, and A. Lluch. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Journal on Medical Economics*, 38:698–707, 2012. 133

[32] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 2001. 103

[33] T. Fawcett. An introduction to roc analysis. *Patterns Recognition Letters*, 27:861–874, 2005. 67, 87

[34] MATLAB file exchange. DNA microrray image procesing case study. 21

[35] Union for International Cancer Control. *TNM clasiffication of Malignant tumors*. John Wiley & Sons, 2009. 132

[36] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000. 5, 49

[37] E. Gadaleta and N. R. Lemoine C. Chealala. Online resources of cancer data: barriers, benefits and lessons. *Briefs in Bioinformatics*, 12:52–63, 2010. 177

[38] A. Gelman. *Bayesian Data Analysis*. Chapman & Hall, 2003. 28, 37, 56, 102, 103, 111, 112, 114, 118, 125, 139, 158

[39] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:1–19, 2006. 158

[40] Y. Guan and J. G. Dy. Sparse probabilistic principal component analysis. *Journal of Machine Learning Research*, 5:185–192, 2009. 6

[41] E. Gutiérrez-Peĩa, A. Smith, and J. Bernardo. Exponential and bayesian conjugate families: Review and extensions. *TEST*, 6:1–90, 1997. 37

[42] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, 6:422–433, 2001. 5, 48

[43] M. Harva and A. Kabán. Variational learning for rectified factor analysis. *Signal Processing*, 87(3):509 – 527, 2007. 96, 97

[44] R. Henao and O. Winther. Sparse linear identifiable multivariate modeling. *Journal of Machine Learning Research*, 12:863–905, 2011. 92, 96

[45] Y. Huang, I. M. Tienda-Luna, and Y. Wang. A survey of statistical models for reverse engineering gene regulatory networks. *IEEE Signal Process Magazine*, 26:76–97, 2009. 2, 3, 7, 24, 48, 51

[46] Y. Huang, J. Wang, J Zhang, M. Sanchez, and Y. Wang. Bayesian inference of genetic regulatory networks from time series microarray data using dynamic bayesian networks. *Journal of multimedia*, 2:46–56, 2007. 49, 51, 52

[47] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33:245–254, 2003. 24

[48] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969. 4

[49] A. E. Kel, E. Gobling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH : a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31:3576–3579, 2003. 5, 26, 93, 99, 137, 177

[50] D. P. Kroese, T. Taimre, and Z. I. Botev. *Handbook of Monte Carlo methods*. Wiley, 2011. 31, 38, 39

[51] H. Lahdesmaki, S. Hautaniemi, I. Shmulevich, and O. Yli-Harjaa. Relationships between probabilistic boolean networks and dynamic bayesian networks as models of gene regulatory networks. *Signal Processing*, 86:814–834, 2006. 48

[52] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101:4781–4786, 2004. 24, 82, 83, 84, 85, 87, 88, 193

[53] J. C. Liao, R. Boscolo, Y. L. Yang, L. M. Tran, C. Sabatti, and V.P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS*, 100:15522–15527, 2003. 6, 92, 93, 95

[54] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60, 2002. 6, 92

[55] M. López, P.Mallorquín, and M. Vega. Microarrays y biochips de ADN. Technical report, Fundación Genoma España, 2002. 1, 11, 18

[56] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2002. 33, 38, 102, 103, 111

[57] S. Mallat. *A Wavelet Tour of Signal Processing*. ELSEVIER, 2009. 100, 101

[58] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 27:2263–2270, 2011. 77

[59] V. Matys and O. V. Kel-Margoulis. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34:108–110, 2006. 5, 23, 93, 99, 137, 175, 177

[60] J. Meng, M. Sanchez-Castillo, I.M. Tienda-Luna, and Y. Huang. Prediction of cancer subtypes using bayesian factor network model. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 995–996, 2011. 3

[61] J. Meng, J. Zhang, Y. Qi, Y. Chen, and Y. Huang. Uncovering transcriptional regulatory networks by sparse bayesian factor model. *EURASIP Journal on Advances in Signal Processing*, 2010:1–18, 2010. 94, 96, 97, 118

[62] E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi. A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, 4:759–813, 2010. 141

[63] K. P. Murphy. Conjugate bayesian analysis of the gaussian distribution. Technical report, 2007. 37

[64] S. Narasimhan, R. Rengaswamy, and R. Vadigepalli. Structural properties of gene regulatory networks: Definitions and connections. *BMC Bioinformatics*, 6:158–170, 2009. 93, 95

[65] KEGG: Kyoto Encyclopedia of Genes and Genomes. Kegg pathway database. 24

[66] Yeast Cell Cycle Analysis Project of Stanford University. Download data. 84

[67] D. S. Oh, M. A. Troester, J. Usary, and C. M. Perou. Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *JOURNAL OF CLINICAL ONCOLOGY*, 24:1656–1664, 2006. 133

[68] M. Palaiologou, J. Koskinas, M. Karanikolas, and E. Fatourou D. G. Tiniakos. E2f-1 is overexpressed and pro-apoptotic in human hepatocellular carcinoma. *VIRCHOWS ARCHIV*, 460:439–446, 2012. 141

[69] A. Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2006. 27

[70] J. S. Parker. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinicical Oncology*, 27:1160–1167, 2009. 132, 135, 139

[71] P. Z. Peebles. *Principios de probabilidad, variables aleatorias y señales aleatorias*. Mc Graw Hilll, 2008. 27, 28

[72] D. Peña. *Análisis de datos multivariantes*. McGraw Hill, 2008. 30, 32, 37, 53

[73] D. Peña. *Fundamentos de estadística*. Alianza Editorial, 2008. 27, 29, 31, 32, 34

[74] D. Peña. *Regresión y Diseño de Experimentos*. Alianza Editorial, 2010. 28

[75] C. M. Perou and T. Sorlie. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000. 131, 134

[76] N. Radde and L. Kaderali. Bayesian inference of gene regulatory networks using gene expression time series data. *Lecture Notes in Computer Science*, 2:46–56, 2007. 49

[77] J. M. Raser. Noise in gene expression origins, consequences and control. *Science*, 309:2010–2013, 2005. 6, 52

[78] J. E Reid, K. J. Evans, N. Dyer, L. Wernisch, and S. Ott. Variable structure motifs for transcription factor binding sites. *BMC Genomics*, 11:1–18, 2010. 23, 26, 175

[79] A. Ribeiro, R. Zhu, and S. A. Kauffman. A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology*, 9:1603–1609, 2006. 6, 24, 49

[80] V. Rodriguez-Galiano, E. Pardo-Iguzquiza, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. Downscaling landsat 7 etm+ thermal imagery using land surface temperature and ndvi images. *International Journal of Applied Earth Observation and Geoinformation*, 18:515–527, 2012. 123

[81] M. Ronen, R. Rosenberg, and B. I. Shraimanand U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99:10555–10560, 2002. 94, 95

[82] S. M. Ross. *Simulation*. Pearson, 1997. 39

[83] R. Rouzier and W. F. Symmans C. M. Perou. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research*, 11:5678–5685, 2005. 3

[84] J. J. K. Ruanaidh and W. J. Fitzgerald. *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996. 34

[85] C. Sabatti and G. M. James. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22:739–746, 2006. 92, 93, 95, 96, 97, 105, 118, 123

[86] C. Salon, G. Merdzhanova, C. Brambilla, E. Brambilla, S. Gazzeri, and B. Eymin. E2f-1, skp2 and cyclin e oncoproteins are upregulated and directly correlated in high-grade neuroendocrine lung tumors. *Oncogene*, 26:6927–6936, 2007. 141

[87] M. Sanchez-Castillo, J. Meng, I.M. Tienda-Luna, and Y. Huang. Basis-expansion factor models for uncovering transcription factor regulatory networks. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, Ann Arbor, Michigan, 2012. 2, 6

[88] M. Sanchez-Castillo, I. M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Microarray time series modeling and variational bayesian method for reverse engineering gene regulatory networks. In *Signal Processing and Information Technology*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 71–76. Springer, 2012. 52

[89] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. *2010 IEEE 10th International Conference on Signal Processing*, chapter Modified variational method for genes regulatory network learning, pages 1781–1784. Institute of Electrical and Electronics Engineers, Inc., 2010. 52

[90] M. Sanchez-Castillo, I.M. Tienda-Luna, D. Blanco-Navarro, and M. C. Carrion-Perez. Methods and recent patents for modeling and uncovering gene regulatory networks. *Recent Patents on Signal Processing*, 2:88–95, 2012. 2, 4, 49, 56

[91] T. Schaffter, D. Marbach, and D. Floreano. Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 7:1–17, 2006. 75

[92] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:1–11, 2008. 97

[93] I. Shmulevich and E. R. Dougherty. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18:261–274, 2002. 4, 48

[94] I. Shmulevich, E. R. Dougherty, and W. Zhang. From boolean to probabilistic boolean networks as models of genetic regulatory networks. In *Proceedings of the IEEE*, pages 1778–1792, 2002. 48

[95] American Cancer Society. American cancer society annual report 2011. 131

[96] T. Sorlie and C. M. Perou. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, 98:10869–10874, 2001. 132, 134

[97] P. T. Spellman. Comprehensive identification of cell cycle-regulated genes of the yeast SCE by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998. 24, 82, 84, 85

[98] M. R. Spiegel. *Fórmulas y Tablas de Matemática aplicada*. McGraw Hill, 2000. 97, 101

[99] I. M. Tienda-Luna, Y. Huang, and Y. Yin. Uncovering gene regulatory networks from time-series microarray data with variational bayesian structural expectation maximization. *EURASIP Journal on Bioinformatic and and Systems Biology*, 1, 2007. 2, 5, 49, 52, 56, 62, 68, 86

[100] I. M. Tienda-Luna, Y. Yin, M. C. Carrion-Perez, Y. Huang, M. Sanchez H. Cai, and Yufeng Wang. Inferring the skeleton cell cycle regulatory network of malaria parasite using comparative genomic and variational bayesian approaches. *Genetica*, 132:131–142, 2008. 49, 52

[101] I. M. Tienda-Luna, Y. Yin, and Y. Huang. Constructing gene networks using variational bayesian variable selection. *Artificial life*, 14:65–79, 2008. 2, 6, 47, 50, 52

[102] J. V. Vilar-Fernández. *Modelos estadísticos aplicados*. Univerisdad de da Coruña, 2003. 27

[103] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. 32, 33

[104] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98:11462–11467, 2001. 3, 91

[105] J. Xiang, X. Pan, J. Xu, and Q. Wei. Human epidermal growth factor receptor 2 protein expression between primary breast cancer and paired

asynchronous local-regional recurrences. *Experimental and Therapeutic Medicine*, 2:1187–1191, 2011. 133

[106] J. Xie and P. S. Crooke. A computational model of quantitative chromatin immunoprecipitation (ChIP) analysis. *Cancer Informatics*, 6:138–176, 2008. 21

[107] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001. 92