Pacific Symposium on Biocomputing 13:178-189(2008)

A BAYESIAN FRAMEWORK FOR DATA AND HYPOTHESES DRIVEN FUSION OF HIGH THROUGHPUT DATA: APPLICATION TO MOUSE ORGANOGENESIS

MADHUCHHANDA BHATTACHARJEE

School of Mathematics and Statistics, University of St Andrews
St Andrews, Fife, Scotland, KY16 9SS, UK

COLIN PRITCHARD & PETER NELSON

Division of Human Biology, Fred Hutchinson Cancer Research Center
Seattle, WA 98109-1024, USA

In this paper we present a framework for integrating diverse data sets under a coherent probabilistic setup. The necessity of a probabilistic modeling arises from the fact that data integration does not restrict to compiling information from data bases with data that are typically thought to be non-random. Currently wide range of experimental data is also available however rarely these data sets can be summarized in simple output data, e.g. in categorical form. Moreover it may not even be appropriate to do so. The proposed setup allows modeling not only the observed data and parameters of interest but most importantly to incorporate prior knowledge. Additionally the setup easily extends to facilitate more popular data-driven analysis.

1. Introduction

1.1. *Challenges in Data Integration*

It has been realized that in order to address biological questions more fully and to extract more knowledge from the wealth of data, researchers require tools that will allow them to integrate different datasets in a dynamic, hypothesis-driven fashion and to analyze them within a biologically meaningful framework[13].

However integration is often mistaken as making vast amount of data available to the researcher by warehousing or other methods. It is often seen that integration of large number of data sets in such a manner results in a messy incomprehensible scenario. Such output might contain vast amount of biological information however fails to generate testable hypotheses and theories that with proper validation may add to our knowledge. It is also not uncommon that a painstaking effort to integrate information sources has produced rather trivial observations on the system.

Advancement in computing abilities makes it plausible to deal with large amount of data; unfortunately it is often done in an adhoc manner. On the other

hand currently more systematic approaches are also confined to amalgamating specific databases to a single experiment. There is a growing support to the idea that a more hypotheses driven choice of data sources should be made which then needs to be carefully analyzed[7].

1.2. *Challenges in Statistics Inference*

High throughput molecular biology techniques have posed new challenges for statistics, and with this we have seen some more adventurous use of statistics. Although data sets are typically large, the number of features is also large. This added to the fact that the features frequently depart from the i.i.d. set-up, makes a credible feature level inference near impossible. Moreover if this unstructured and unknown dependence is not accounted population level inference also becomes inaccurate.

An additional level of complication is introduced by the fact that the process of obtaining meaningful results involves numerous decision-making steps. Apart from a few attempts[8,12] this aspect is generally not discussed and consequently not accounted in the analysis and inference.

1.3. *Objective*

The initial findings from an analysis of a high-throughput data are often messy due to several reasons. The level of specificity of the phenotype is an important factor behind this. For most diseases there would be multiple known factors affecting the overall variability. For practical reasons it is difficult to design experiments where one would be able to account for all these factors. The hence uncontrolled factors would contribute to the observed variability. Additionally although we target to study a specific aspect of the system, as cells continue to live independent of the experiment or disease under study, changes in normal life functions also affect the results[16].

Evidently some additional information is needed to identify relevant quantities from such an inference. Most experiments are designed and carried out using prior knowledge of the conditions, diseases or treatments under study and can be used critically to control variation. We would use such experimental data, which are complete, meaningful and possibly complementary to each other. This can be thought as the hypothesis-driven part of the data fusion.

For most phenotype/disease there will be limited number of such experimental data available which are useful. The knowledge from existing databases can then be augmented to obtain a better understanding of our findings, which is the purely data driven aspect of the investigation.

We will derive an integrated modeling and inference procedure on such data, which are high dimensional data sets of varied data types from various sources. Integration of these sources utilizes existing knowledge of the phenotype of interest. We will carry out population as well as individual level inference, in presence of dependence, combining multiple decision-making steps that allows for propagation of error.

2. Data

The experimental data consists of one time-course data and three other data sets. The biological objective behind the choice of these experiments is to study developmental behavior of prostate, preferably at cell-type level, highlighting behavior of key genes like androgen regulated/responsive ones. These data sets were collected as a first part of a two-part study of Prostate cancer.

2.1. *Molecular Characteristics of Developing Mouse Prostate*

To identify genes potentially involved with prostate development, temporal expression changes were determined by measuring transcript abundance levels in cDNA libraries constructed from distinct stages of maturation. A purpose-built cDNA microarray[4,5] enriched for genes in the developing mouse prostate, which serves as a unique resource for molecular studies of prostate development were utilized.

Table 1. Summary of biological samples used for timecourse microarray experiment, where UGS: Uro-genital-sinus, DLP: dorsolateral prostate, VP: ventral prostate AP: anterior prostate.

Time point	Tissue	Developmental Process	Androgen level [1]
E15.5	♂UGS	Undifferentiated	2 days exposure
E16.5	♂UGS	Undifferentiated	3 days exposure
E17.5	♂UGS	Prostate buds	High
E18.5	♂UGS	Branching morphogenesis	High
Day 7	DLP,VP,AP	Branching morphogenesis (Peak)	Low
Day 30	DLP,VP,AP	Puberty	Very High
Day 90	DLP,VP,AP	Adult, fully differentiated	Very High

The transcriptional program of prostate development was characterized by profiling gene expression at seven time points corresponding to critical stages of prostate differentiation (Table 1). For each time point three biological samples were generated and each sample was hybridized twice for a total of 42 microarrays. The samples were hybridized with a common reference RNA consisting of embryonic age 14.5 days (E14.5) male UGS. At E14.5, UGS is undifferentiated and has not been exposed to significant levels of androgens, therefore is thought as 'start' for prostate development. Thus, the microarray ratios depict the unfolding program of prostate development in relationship to the most undifferentiated state.

2.2. *Characterizing cell-type specific expression heterogeneity*

The interpretation of temporal changes in gene expression from whole tissue is complicated by cell heterogeneity. The developing prostate can be roughly broken down into two major cell compartments: epithelium and mesenchyme/ stroma. Reciprocal signaling between mesenchyme and epithelium is critical for prostate differentiation. Mesenchyme is the precursor to the adult stroma, Urogenital sinus mesenchyme (UGM) induces epithelial budding and the epithelium, in turn, stimulates mesenchymal differentiation. In subsequent tissue recombination experiments it was shown that several different types of epithelium of endodermal origin can form prostate in combination with an inducing mesenchyme. Thus, there is some inductive promiscuity in both epithelial and mesenchymal compartments.

2.3. *Androgen Response Program of the Developing Mouse Prostate*

Androgens act via the UGS mesenchyme to induce prostatic epithelial development, presumably through a paracrine mechanism. Yet no androgen-regulated genes have been identified in the UGS mesenchyme.

Using a custom cDNA microarray enriched for genes expressed in the developing mouse prostate, three in vivo strategies were adopted to identify androgen-regulated genes at the time of prostate induction. We compared (1) male UGS to female UGS, (2) female UGS dosed with testosterone *in vivo* to female dosed with placebo, and (3) wild-type male UGS to androgen receptor-deficient (*tfm*) male UGS. Each comparison is a distinct way of assessing androgen-regulated genes in the UGS. Three biological replications and two array replications were performed for each comparison at both E16.5 and E17.5 for thirty-six total microarray experiments.

2.4. *Gene-Ontology and KEGG pathways data*

The Gene Ontology consortium (GO)3 and Kyoto Encyclopedia of Genes and Genomes (KEGG)2 databases enable statistical analysis of biological processes or pathways that may be enriched or depleted in a certain experiments.

We considered 208 Biological process, 64 Cell components and 151 Molecular functions from the Gene Ontology database. Data for the present analysis was from all nodes up to level 6 that were represented by at least ten genes to ensure that functional conclusions were not drawn from very few genes. As GO terms have multiple parents, the completed trees based on these nodes consisted of 462 nodes for Biological processes, 84 nodes for Cellular Components and 185 nodes for Molecular Functions. From the KEGG database

20 specialized pathways were chosen based on their relevance to the biological problem undertaken here.

2.5. *Overall data summary*

In the overall analysis the data comes from a varied source and is schematically presented in Figure 1. The current practice of handling multiple experimental data would be to work with very crude summaries, e.g. list if differential (or otherwise interesting) genes, ignoring the fact that almost always such lists were outcome of decision-making and decisions were not taken with 100 percent confidence. Some associated measure of confidence should be utilized.

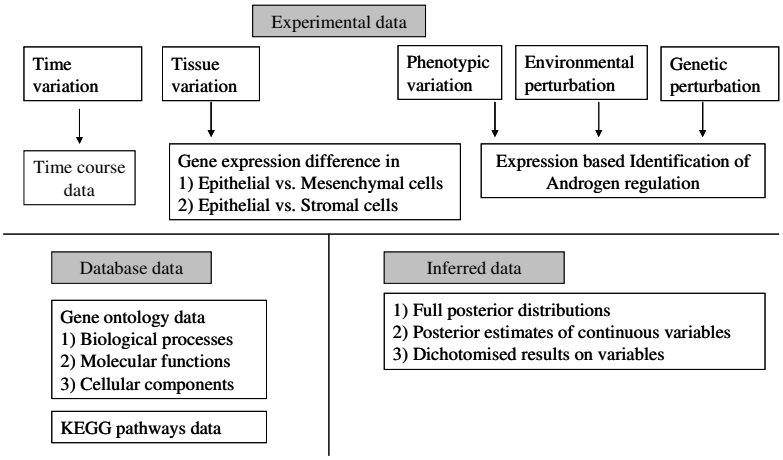

Figure 1. Schematic Presentation of all data sources used for integrated modeling

Moreover in some aspects of analysis such summarization is potentially misleading. For example, many have experienced that normalization affects subsequent biological conclusions from a microarray data analysis. One solution would be carry out a model based normalization and use the joint distribution of the (distinct) genes for further analysis[6,8]. Similarly while integrating gene characteristics from one experiment to another the joint distribution of genes with respect to the (conditioning) characteristic could be used in the subsequent experiment.

Such probabilistic summarization is denoted as inferred data and it reflects our knowledge and confidence on the data. The best possible usage of one source of data would be using the whole joint distribution, if that is too large or complex then a summary (possibly first two marginal moments for each feature) can be used. In some situations we might have to use the categorical (typically binary) summarization of a data.

3. Model

We will describe the model for the time course data. For the three additional experiments the models are similar in principle with appropriate parameterization for eliciting the desired characteristics. We will derive the model in a stepwise manner aiming to describe the parameters along with their utilities. Note that in order to exploit conjugacy we parameterized the Normal distributions using mean and precision (i.e. inverse-variance).

3.1. *Normalization of individual data set*

Normalization is carried out at block level for each array using constrained piecewise linear models (in Bayesian framework). The range of values of the log-intensities from the reference sample was divided into three windows, with breakpoints chosen to be at 5 and 7. The normalized data thus produced is highly comparable with standard loess type normalization[8], however has the advantage of being model based hence allows for propagation of error to next stage of analysis. The linear model parameters for the normalization are denoted (and described) as follows:

$\beta_{ljkm} \sim N(1 , 0.1)$ and $\alpha_{ljk2} \sim N(0 , 0.1)$ where
$\alpha_{ljk1} = \alpha_{ljk2} +(\beta_{ljk2} - \beta_{ljk1})*5$ and $\alpha_{ljk3} = \alpha_{ljk2} +(\beta_{ljk2} - \beta_{ljk3})*7$, l:tissue/time-point, j:array, k:blocks on each array and m:1, 2 and 3 (number of windows).

Let LIR and LIE denote the Log intensities from reference and experimental samples respectively. In the rest of model description, following notation will be used, l: tissue, i: spot, j: array within l-th tissue, b(i): print tip/ block number of i-th spot, w(lij): window number for LIR_{lij} for the l-th tissue, i-th sopt on j-th array, d(i): distinct gene corresponding to i-th spot on the array. Note the arrays contain multiple spots/probes for several genes; however this was not designed in a balanced manner. The (incomplete) models for LIR and LIE experimental samples were:

$LIR_{lij} \sim N(\mu_{d(i)} , 0.1)$ and $LIE_{lij} \sim N((\alpha_{ljb(i)w(lij)} + \beta_{ljb(i)w(lij)} *LIR_{lij}) ,)$.

3.2. *Characterizing gene-expression behavior within an experiment:*

Assume that, a priori each gene has its own expression ratio, say θ^0_k, where $\theta^0_k \sim N(0, 0.1)$, with k = 1,…, no. of distinct genes. Let the expression ratio for the k-th gene from the l-th tissue be θ^1_{lk}, where θ^1_{lk} are assumed to be drawn from a Normal distribution with mean θ^0_k, i.e.

$\theta^1_{lk} \sim$ Normal(θ^0_k, τ^0_k) and let $\tau^0_k \sim$Gamma(1 , 1).

The completed model for LIE_{lij} is as follows:

$LIE_{lij} \sim$ Normal(($\alpha_{ljb(i)w(lij)} + \beta_{ljb(i)w(lij)}$ ($LIR_{lij} + \theta^1_{ld(i)}$)), $\tau^1_{ld(i)}$).

The available knowledge indicated several possible profiles for cohorts of genes during mouse prostate differentiation that would be of interest. Such profiles could be explored based on posterior behavior of different functions of the parameters (θ^1_{1k} , τ^1_{1k}) and also possibly other parameters. For example for any time point upregulated genes can be identified by the posterior probability of (studentized) θ^1_{1k} exceeding a pre-specified cut-off (e.g. Normal-distribution percentile) and such probability estimates have been noted[8] to be monotonically related to the d-scores obtained using well-known SAM.

3.3. *Characterizing gene-expression behavior across experiments:*

For k-th gene let μ^1_k, μ^2_k be the expression parameter describing epithelium & mesenchyme specific behavior. Similarly ν^j_k be androgen response of gene k in platform j, j=1, 2, 3. To infer whether a gene exhibits branching morphogenesis profile while being epithelium specific and androgen regulated we use posterior distribution of the variable $I(\theta^1_{1k}<\theta^1_{2k}<\theta^1_{3k}<\theta^1_{4k}<\theta^1_{5k}$ and $\theta^1_{7k}<\theta^1_{6k}$, $\mu^1_k>\varphi_{0.95}$, $\nu^*_k>\varphi_{0.95})$, where ν^* is function of ν^j's and φ is Normal percentile. If the data sets and genes exhibit varied precision then standardized parameters are used for the indicators. These may still be biased by the size (or variability) of any particular data set. Hence as a cautionary measure the information flow was allowed only from individual experimental data to the integrated analysis and not otherwise. This is easily implemented in WinBUGS framework using "cut"-function appropriately.

3.4. *Probabilistic assessment of biological processes enrichment*

For expression profile of interest the basic functional enrichment analysis is as described in (8). Apart from enrichment testing we utilized several summary assessment of functional enrichment. For example, the log-ratio of proportions of genes with certain functionality in S and Sc can be thought as similar to "expression" ratio for that functionality. These can then be visualized as heat map. These we have further analyzed using standard clustering techniques which have provided useful insight into functional pattern over time.

3.5. *Overall analysis setup*

The overall setup allows us to analyze each microarray based experiment in quite extensive manner. For each of these data sets can be analyzed individually, additionally our setup allows for modeling these data sets where desired output from one analysis along with it the uncertainty involved in decision making/inference is carried to analysis of another data.

One major achievement of this data analysis is the magnitude of the statistical problem that we succeeded in implementing using freely available software WinBUGS without having to write custom MCMC codes. This opens up numerous modeling possibilities to address complex biological questions.

To give an idea of the number of parameters being jointly monitored in this implementation, consider the following: normalization parameters 12 000+, hyperparameters 40 000+, expression parameters 220 000+, for each of 10 000+ genes 13 expression profiles from individual dataset analyses and 30 expression profiles using two datasets at a time and 20 profiles for all three data sets, 750+ GO-KEGG processes enrichment for each of these profiles. Approximately posterior distributions of 1 million variables of interest are jointly monitored under this setup. Additionally all missing data points are augmented which typically increases with size of experiments.

4. Results

The analyses of individual experimental data sets and different combination of their integration brought out many interesting results. This is due to the nature of these experiments and also due to the flexibility of the model parameterization. In the following we describe a few such outputs from the analyses.

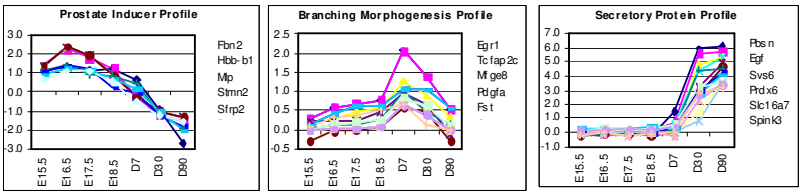

Figure 2. The critical profiles (known from pathological information) were translated into functions of expression parameters. The figures present (log) expression change of top genes with high posterior probability of having these profiles during mouse prostate differentiation.

4.1. *Analysis of time course data*

By deriving the posterior distribution of different constrained combinations of the parameters we can identify genes having specific expression profiles over time. In Figure-2 we present some of the genes estimated to have high probability of having some known profiles. Some of the genes thus identified were already known and some new ones have been verified subsequently.

In the joint analysis of the Gene Ontology information and the time course data, several functionalities clearly depict the distinctly different behavior in the two phases of life, namely embryonic and otherwise. However our model treats all time points were equally, which gives us more confidence in our findings.

By analyzing the heat map of functional enrichment on time course, we were able to identify very interesting clusters of functions whose profile correspond to the known prostate development profiles (see Figure 3).

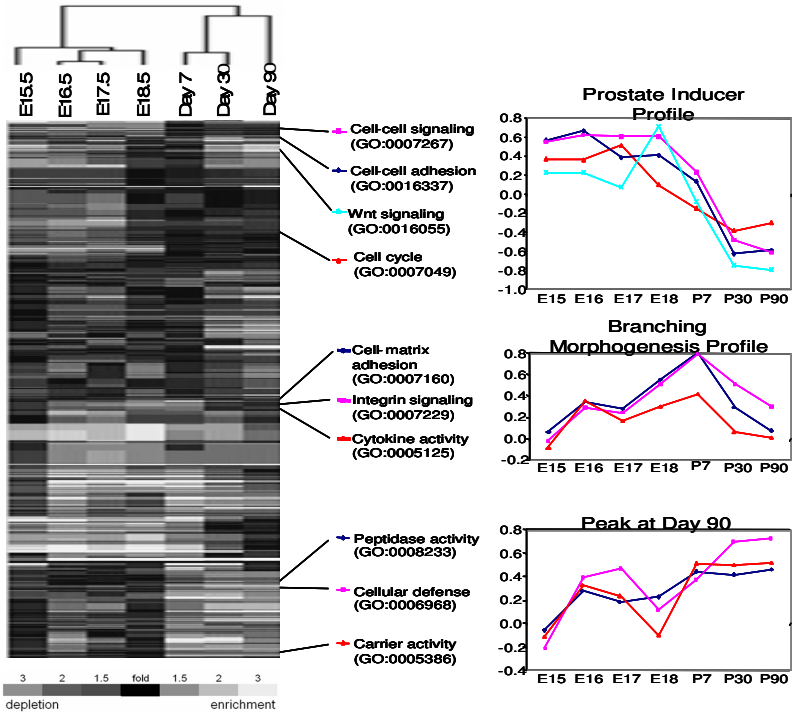

Figure 3. Examples of GO terms of interest are listed according to the time-point of peak representation and position in the heat-map, with a detailed plot of their behavior over the time-course. The data for the heatmap was generated by the methods described in Section 3.4.

4.2. *Integrated analyses of multiple experimental data sources*

The integrated analysis of the cell-type specific and the time course experiments yielded fascinating expression diversity within the developing tissues. In Table-2 we present a GO-based interpretation of these expression behaviors. The resulting findings of this integrated analysis were in high concordance with existing knowledge and our hypothesis. The distinctive nature of the two cell types is clearly visible along with their time stochastic nature. The tree structure of GO provides additional information on change in pattern over time.

By jointly analyzing all the experiments we are able to explore in each cell-type the expression profiles over time of the Androgen responsive genes. The major Androgen responsive genes showed two distinct expression profiles over

time, one with highest expression observed at E16.5 and another where highest expression occurs in adult state (D90). The cell-type specific expression data indicated that the early expressing gene was in Messenchymal/Stromal cells (e.g. Sfrp2) where as the adult ones were expressed in Epithelium (e.g. Agr2 and Mmp7).

Table 2: Functional analysis of cell-type specific genes up-regulated at a certain time point. The entries represent estimated (log) change in enrichment. Upregulated functions have been highlighted in white against dark-grey-background and those downregulated in black against light grey-background ("###" indicates high negative value).

GO	Epithelium							Mesenchyme						
	E15.5	E16.5	E17.5	E18.5	D7	D30	D90	E15.5	E16.5	E17.5	E18.5	D7	D30	D90
Development	0.5	0.5	0.6	0.7	0.5	0.4	0.2	1.1	1.3	1.2	1.4	1.3	1.2	1.3
Morphogenesis	0.3	0.4	0.5	0.6	0.4	0.2	0.1	1.3	1.5	1.5	1.6	1.5	1.5	1.5
Organogenesis	0.2	0.3	0.4	0.4	0.1	-0.1	-0.3	1.4	1.6	1.6	1.7	1.7	1.6	1.7
Neurogenesis	-1.1	-0.5	0.2	-0.6	0.3	-0.1	-0.4	1.4	1.3	1.1	0.9	0.8	0.0	-1.7
axonogenesis	-1.8	-2.5	-1.8	-1.9	-2.3	-3.2	-5.3	1.7	1.5	0.9	0.6	0.3	-3.4	-3.0
morphogenesis of an epithelium	0.5	0.9	1.1	1.5	1.5	1.7	1.4	-3.9	-6.1	-7.5	-4.4	-3.3	###	-4.8
epithelial cell differentiation	-2.8	-0.1	1.3	1.7	1.7	1.4	1.4	-4.5	###	###	-5.3	-4.1	###	-4.1
epithelial to mesenchymal transition	-1.8	0.9	2.3	2.7	2.7	2.4	2.4	###	###	###	###	###	###	###
morphogenesis of embryonic epithelium	1.3	1.5	1.1	1.5	0.6	1.7	1.7	###	###	###	###	###	###	###
Growth	0.1	-0.2	0.1	0.1	0.1	0.4	0.2	1.7	1.2	1.4	1.6	1.5	1.8	1.6
cell growth	0.3	0.0	0.5	0.4	0.4	0.7	0.6	2.0	1.6	1.7	1.9	1.9	2.1	1.9
regulation of cell growth	0.5	0.2	0.7	0.6	0.6	0.9	0.8	2.1	1.6	1.7	2.0	1.9	2.1	1.8

Amongst the Androgen-responsive genes that are epithelium specific we noticed three major expression patterns (see figure 4), 1) higher expression in late embryonic state (e.g. Anxa1, Psca) 2) higher expression at infant stage (e.g. Itgb4, Sox9) and 3) high expression in adult stage (e.g. Aldh1a1, Agr2, Cldn8).

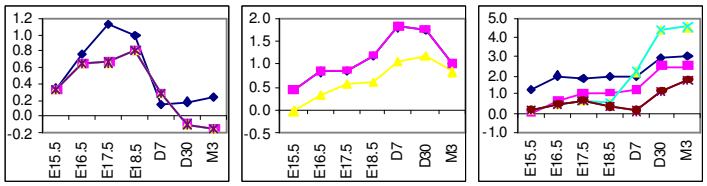

Figure 4. Developmental expression profile of Androgen-responsive genes expressed in Epithelium.

4.3. *Experimental and literature based cross-validation*

For Sfrp2 real-time RT-PCR at each of the seven time points were performed and results were highly concordant with the microarray measurements. In situ hybridization confirmed mesenchymal expression of Sfrp2. Quantitative PCR confirmed that Sfrp2 is up regulated with androgen in the UGS. Agr2 is a gene that is thought to be involved in breast cancer metastasis[14]. Subsequent analysis with prostate cancer related data has shown this gene to be significant. Mmp7 is a gene that has been previously shown to influence cancer progression[15].

Psca is a known to be androgen regulated and is associated with prostate cancer. A further experiment using whole mount in situ hybridization showed Psca was highly expressed in epithelium. Sox9 have been shown to be directly relevant for tumor suppression in Prostate[10] and in some organs domain of expression of Sox9 protein is normally known to be the distal epithelial compartment[9,11].

5. Discussion

The composite data we consider for this analysis comprises of data from several experiments, which are meaningful and complete by themselves. The hypothesis driven fusion of these enabled us to reduce the experiment size from 7x2x3 to 7+2+3 experiment for time points, cell-types and Androgen-platforms. Even if we had the resource to do the full experiment some of these may not have been biologically feasible to do in reality.

The biological hypotheses were translated in statistical framework as functions of parameters (e.g. expression) and were assessed a-posteriori. The different functional combinations of the model parameters cover a wide range of biological characteristics to be studied. This is another aspect of our modeling setup that is not easily available in the commonly used statistical tools, simply because complex hypotheses would require specialized testing procedure which may not be available readily. Complexity of individual modeling units was kept moderate to optimize computation time and parameter space of interest.

While analyzing using existing models quite often we observe that even moderate change in analysis technique for one data set & for any single step of analysis can influence overall biological conclusions. It is well-known that this happens due to not propagating uncertainties in these analysis steps to subsequent steps. The Bayesian setup proposed in (8) enables us to avoid this problem and was extended here to a much larger problem. Analytic intractability is a common consequence of such complex models. In this respect a mentionable achievement here is being able to implement this integrated model using available software, opening up varied modeling and input data-type possibilities.

Our objective was to be able to balance between quantity of data and quality of inference. Although one would be tempted to use as much data as possible we need to remember a few aspects of these data. Most experimental data come with a lot of error/noise. Using only a summary from each of these ignoring the noise potentially can (and often does) lead to non-reproducible results. This is where robust inference method is crucially needed and is provided by our method.

References

1. C. Corpechot, E.E. Baulieu and P. Robel, *Acta Endocrinol (Copenh)* 96, 127-35 (1981).
2. M. Nakao, H. Bono, S. Kawashima, T. Kamiya, K. Sato, S. Goto, and M. Kanehisa, *Genome Inform Ser Workshop Genome Inform 10*, 94-103 (1999).
3. M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, *Nat Genet* 25, 25-9 (2000).
4. P.S. Nelson, C.C. Pritchard, D. Abbott and N. Clegg, N. *Nucleic Acids Res* 30, 218-20 (2002).
5. D.E. Abbott, C.C. Pritchard, N.J. Clegg, C. Ferguson, R. Dumpit, R.A. Sikes and P.S. Nelson, *Genome Biol* 4, R79 (2003).
6. M. Bhattacharjee, C.C. Pritchard, M.J. Sillanpää and E. Arjas, *Proceedings of the CAMDA02 Conference*, Ed. Johnson K. and Lin, S., Kluwer-Academic Publishers (2003).
7. J.H.G.M. van Beek, Comp Funct Genom; 5: 201–204, (2004).
8. M. Bhattacharjee, C.C. Pritchard, P.S. Nelson and E. Arjas, *Bioinformatics*, 20, 2943-2953 (2004).
9. P. Blache, M. van de Wetering, I. Duluc, C. Domon, P. Berta, J.N. Freund, H. Clevers, P. Jay, *J Cell Biol*, 166 (1): 37-47, (2004).
10. R. Drivdahl, K.H. Haugk, C.C. Sprenger, P.S. Nelson, M.K. Tennant, S.R. Plymate, *Oncogene*, 23 (26): 4584-93, (2004).
11. T. Okubo, P.S. Knoepfler, R.N. Eisenman, B.L. Hogan, *Development*, 132 (6): 1363-74, (2005).
12. M. Bhattacharjee and M.J. Sillanpää, (To appear in) *Proceedings of CAMDA 2006*.
13. M. Mirona and R. Nadon, *Trends in Genetics*, 22 (2), 84-89, (2006).
14. C.L.Wilson, A.H. Sims, A. Howell, C.J. Miller, R.B. Clarke. *Endocr Relat Cancer*, 13 (2): 617-28, (2006).
15. D. Bianchi-Frias, C.C. Pritchard, B.H. Mecham, I.M. Coleman, P.S. Nelson, *Genome Biol*, 8 (6): R117, (2007).
16. T. Werner, *Mechanisms of Ageing and Development*, 128, 168–172, (2007).