Non-Disjoint Clustered Representation for Distributions over a Population of Cells

Matthieu Pichené[1], Sucheendra Palaniappan[2], Eric Fabre[1], and Blaise Genest[3]

[1] Inria, Team SUMO, Rennes, France
[2] The Systems Biology Institute, Tokyo, Japan
[3] CNRS, IRISA, Rennes, France

1 Motivation

We consider a large homogenous population of cells, where each cell is governed by the same complex biological pathway. A good modeling of the inherent variability of biological species is of crucial importance to the understanding of how the population evolves. In this work, we handle this variability by considering multivariate distributions, where each species is a random variable. Usually, the number of species in a pathway -and thus the number of variables- is high. This appealing approach thus quickly faces the curse of dimensionality: representing *exactly* the distribution of a large number of variables is intractable.

To make this approach tractable, we explore different techniques to *approximate* the original joint distribution by meaningful and tractable ones. The idea is to consider families of joint probability distributions on large sets of random variables that admit a compact representation, and then select within this family the one that best approximates the desired intractable one. Natural measures of approximation accuracy can be derived from information theory. We compare several representations over distributions of populations of cells obtained from several *fine-grained* models of pathways (e.g. ODEs). We also explore the interest of such approximate distributions for approximate inference algorithms [1, 2] for *coarse-grained* abstractions of biological pathways [3].

2 Results

Our approximation scheme is to drop most correlations between variables. Indeed, when many variables are conditionally independent, the multivariate distribution can be compactly represented. The key is to keep the most relevant correlations, evaluated using the *mutual information (MI)* between two variables.

The simplest approximation is called *fully factored (FF)*, and assumes that all the variables are independent. It leads to very compact representation and fast computations, but it also leads to fairly inaccurate results as correlations between variables are entirely lost, even for highly correlated species (MI = 0.6).

Alternately, one can preserve a few of the strongest correlations, selected using MI, giving rise to a set of *disjoint clusters* of variables. For efficiency reason, we used clusters of size two. This model was able to capture some of the most significant correlations between pairs of variables (representing around 30% of the total MI), but dropped significant ones (MI = 0.2).

A better trade-off between accuracy and tractability was obtained by using *non-disjoint* clusters of two variables, structured as a tree, called the *tree-clustered approximation (TCA)*. The approximated joint distribution is fully determined by the marginals over each selected cluster of 2 variables. This gives a compact representation (< 800 values in our experiments). Further, any marginal over k out of n total variables can be computed with time complexity $O(nv^{k+1})$, where each variable can take v possible values. Last, a tractable algorithm [4] allows to compute the best approximation of any distribution by a tree of clusters. TCA succeeded in capturing most correlations between pairs of variables (representing around 70% of the total MI), losing no significant ones (MI < 0.1).

Regarding inference, FF, disjoint clusters and TCA were compared to *Hybrid FF (HFF)* [2]. In short, HFF preserves a small number of joint probabilities of high value (called spikes), plus an FF representation of the remaining of the distribution. The more spikes, the more accurate the approximation, and the slower HFF inference. Overall, TCA is very accurate, while HFF generates sizable errors, even with numerous spikes (32k). Further, TCA is faster than HFF, even with few spikes (3k). FF and disjoint-clusters are even faster (1 to 2 order of magnitudes) than TCA, but the accuracy of both remains problematic.

3 Perspectives

We now aim at modeling and studying a tissue, made of tens of thousands of cells. In this context, capturing the inherent variability of the population of cells is crucial. In order to study multi-scale systems in a tractable way, we advocate a two-step approach: Firstly, abstract the low level model of the pathway of a single cell into a stochastic discrete abstraction, e.g. using [3]. Secondly, use a model of the tissue, which does not explicitly represent every cell but qualitatively explains how the *population* evolves. In this way, one need not explicitly represent the concentration of each of the tens of thousands of cells, but rather only keep one probability distribution.

Acknowledgement : This work was partially supported by ANR-13-BS02-0011-01 STOCH-MC.

References

1. Murphy, K.; Weiss, Y. (2001). The factored frontier algorithm for approximate inference in DBNs. In *UAI'01*, p.378-385, Morgan Kaufmann.
2. Palaniappan, S. K.; Akshay, S.; Liu, B.; Genest, B.; Thiagarajan, P. S. A Hybrid Factored Frontier Algorithm for Dynamic Bayesian Networks with a Biopathways Application. In *TCBB 9(5)*: 1352-1365, IEEE/ACM.
3. Palaniappan, S. K.; Bertaux, F.; Pichené, M.; Fabre, E.; Batt, G.; Genest, B. (2017). Abstracting the dynamics of biological pathways using information theory: a case study of apoptosis pathway. *Bioinformatics* 33 (13): 1980-1986, OUP.
4. Chow, C. K.; Liu, C.N. (1968). Approximating discrete probability distributions with dependence tree. In *IEEE ToIT 14 (3)*: 462-467, IEEE.