

Statistical issues for combining replicates and nearby species data and different omics

P. Liò^{1*}, C. Angelini², I. De Feis², V-A. Nguyen¹, L. Cutillo³, &
R. van der Wath¹

¹Computer Laboratory, University of Cambridge, United Kingdom

²Istituto per le Applicazioni del Calcolo “Mauro Picone” (CNR), Napoli, Italy

³Telethon Institute of Genetics and Medicine, Napoli, Italy

1 Introduction

Statistical genetics and Bioinformatics are experiencing a period of great capability in providing the methodologies to support wet laboratory research and they are keeping the pace with the growing availability of a large variety of molecular biology high through-put data. Molecular biologists are now pressing with more challenging requests. There are two important issues. The first is how to integrate the different types of high through-put data (omics). The other is to integrate replicates (how many?) with nearby species (how many, how evolutionary close?) data .

The problem with the first point is that it is not so obvious how one "omic" data (say transcriptomics) could support findings on other omics, say genomics or proteomics. Clearly all the omics are interdependent because related to the cell machinery and to the process of fitness and survival of the organism. In theory, we need all them but their contribution is not simply additive but more complex and related to the cell circuitries. For example it is possible to produce a quantity of a certain protein by having basal levels of transcription and translation of several copies of a gene or having one gene being highly transcribed or highly translated or simply slowly degraded (for example because of the attachment of a sugar chain - here glycomics may enter, or a lipid chain, lipidomics). In summary due to the untangling of the biological processes it is important to use all the evidences but we do not have models and methods and theory good enough. At the time when only few DNA sequences were known in databases, there was the general belief that the future availability of massive amount of sequences could dispel all statistical genomics uncertainties. The arrival of high through-put data has instead cleared many, not all, the uncertainties and has generated new ones. In many cases, the same is happening with replicates and nearby species data. A problem with gene expression data is the noise; when the experimental data is variable, the most obvious way to improve the signal to noise ratio is to perform experimental replicates. This can be performed by first averaging gene expression measurements observed at different experimental replicates since the standard deviation of such average levels will decrease with the square root of the number of replicates it is likely that the weaker associations will be detected than in the case of a single replicate is performed. When experiments are costly, particularly in high throughput biology, replicates are usually in a number to assure ‘just above’ the threshold statistical reliability for disseminating and publishing results; funding constraints sometimes result in seriously hampering statistical robustness.

The biological reasoning behind replicating experiments is that each organism has homeostatic mechanisms that maintain the genetic information and its expression and functions. This allows to replicate experiments with decent accuracy if the conditions remain the same and no

instrumental errors are present. In molecular biology it is often difficult to retain the exact experimental setup but close conditions are widely accepted. In a field where technology changes constantly and at great pace, results that are few months distant may be based on slightly different technologies or manufacturing aspects of some components leading to slightly different accuracy. The biological samples used in the experiments, i.e. culture cell lines, for examples bacterial or yeast cells which have short generation times, although kept in the same medium, may slightly change because of mutations and contamination during their periodical proliferation and in vitro amplification. Cells may have minimal different concentration of constituents such as nucleic acid, proteic, lipid or sugar factors, ions, giving them different fitness with respect other colonies. However, a large body of experimental evidences in comparative genomics is showing that recently diverged species may retain similarities in gene sequence, expression and genome organization. These considerations suggest that, in absence of experimental replicates, or in addition to these, statistical support to experimental evidences may also be searched by analyzing close variants of the species under examination or phylogenetic nearby species, i.e. species which have recently diverged. Of course the validity of the assumption of replicates through phylogenetical relatedness should be proved by copious experimental evidences or literature.

Our motivation is that a phylogenetic metric may allow to combine replicated and cross-species data. In a phylogenetic tree different species may be located on different leaves of the tree; internal nodes are ancestors; the distance between leaves provides an estimate of the effectiveness (weight) of different species in providing statistical support to the nearby species. Replicates may be located on the same leaf (but their contribution is weighted by the quality of experimental data). Quasi-species (for example rapidly mutating viral species in some how potentially speciating), mutants, strains or different variants which cannot be distinguished as completely diverged species may be accommodate on the tree at very small distances from the tip (leaf) under investigation. A situation in which we have different omics, several replicates and also nearby species data, and a proper metric, can be depicted in such a way that our original sample data is central to a basin of available data which would provide statistical support towards the central of the basin.

Recently we have inched our methodology towards both combining replicates and nearby species and also towards combining different omics, precisely, genomics and transcriptomics. A most challenging step in annotating a genome is to find the transcription factor binding sites (TFBS). These sites, located at various distances upstream each gene, directly influence the amount of gene transcripts. Our approach is to use transcriptomics to identify binding sites affecting transcription. For sake of space, details of methodologies and results for such approaches can be found in , Angelini *et al.* (2008), Angelini *et al.* (2007), Tadesse *et al.* (2004), Lió *et al.* (in preparation). Several computational methods for the discovery of transcription binding sites (TFBSs) have been described, see for example Tompa *et al.*, (2005) for some references. Most of them uses only sequence alignments or the presence of unusual patterns in sequences.

In order to identify TFBS, Conlon *et al.* (2003) applied linear regression with stepwise selection after getting a list of candidate TFBS motifs using MDScan, an algorithm that makes use of word-enumeration and position-specific probability matrix updating techniques. The candidate motifs were scored in terms of number of sites and degree of matching with each gene. Inspired by this approach, we proposed using Bayesian variable selection techniques instead of stepwise methods, see Tadesse *et al.* 2004. Our motivation was that Bayesian model selection methods perform a more thorough search of the model space and hence might potentially pick up motifs that can be missed by stepwise methods. We indeed showed that.

Then we have extended the Bayesian variable selection method to take into account the different and multiple sources information available, pool together results of several experiments (replicates) and allow the users to select the motifs that best explain and predict the changes in expression level in a group of co-regulated genes.

Our main result was that considering more replicates or data from more species, the marginal probabilities become much higher (about 3 fold) than those obtained using single replicates and one species, Angelini *et al.* (2008). We got both confirmation of known results and new findings (motifs) which have high marginal probability values. Next section presents a summary of the methodology.

2 Brief notes on methodology

Our methodology requires the following steps: 1) selection of a group of coregulated genes; in Angelini *et al.* (2008) and Angelini *et al.* (2007) we focused on the ENG1 cluster, a set of very strongly cell cycle-regulated genes of *S. pombe* and *S. japonicus*; the sequences up to 1000 bp upstream were extracted, shortening them, if necessary, to avoid any overlap with adjacent ORF's. For genes with a negative orientation, this was done taking the reverse complement of the sequences. Nucleotide patterns of length 5 to 12 bp have been searched and up to 30 distinct candidates for each width have been considered; 2) determination of the nearby species by means of a phylogenetic tree generated on the basis of the information of the selected genes. Phylogenetic inference can be conducted with several methodologies among which Bayesian and maximum likelihood approaches seem to be the ones with strong statistical basis. We first assessed the distance between fungi based on the ENG1 protein tree; we used the JTT amino acid substitution model of evolution due to the fact that the ENG1 protein family are globular cytoplasmic proteins. Likelihood maximization and maximum likelihood parameter estimation were performed by numerical optimization routines using a replacement matrix for all sites. 3) identification of the homologous genes in species under examination using BLAST; 4) choice of a set of about 200 biologically independent genes of the nearby species after a phylogenetic analysis; 5) determination and scoring a set of about 150 candidate motifs using the software MDSCAN; 6) selection of few 'best motifs' using Bayesian variable selection. We have run 10 parallel MCMC chains of length 100.000. We pooled together the sets of patterns visited by the MCMC chains and computed the normalized posterior probabilities of each distinct visited set. We also computed the marginal posterior probabilities for the inclusion of single nucleotide patterns.

3 Space for improvements

TFBS evolve quickly, but related sites (i.e. controlling the same gene in different species, may share a pattern because of evolutionary dependencies (vertical dependencies), protein binding selectivity constraints (horizontal dependencies) and noise. Having defined a metric to use nearby species tree, there are several questions we are able to answer and others we could simply attempt to. If we are able to estimate the quality of an experimental replicate (the quality may be even different for different genes) and to estimate the quality of an experiment with a nearby species data we can decide if would be better to add a species or a not-so-great quality replicate. The most important assumption in our model is that we are able to generate a perfect phylogenetic tree. The statistical method of maximum likelihood chooses amongst hypothesis

by selecting the one which maximizes the likelihood. Note that the likelihood of a tree is equal to the probability of observing the data (the alignment) if the hypotheses (branch lengths, topology, models of evolution) are correct. It would of course be more reasonable to say that the sequences are what have been observed and the alignment should then be inferred along with the phylogeny.

Noteworthy, we use models of evolution of coding sequences and not of TFBS. The currently available models for sequence evolution are based on fixed rate matrices generated from globular (see for example Dayhoff, JTT) or mitochondrial proteins (for example MTrev). Note that regulatory regions, although under purifying selection, i.e. they evolve slower than surrounding sequences, are nevertheless diverging at higher rate than many coding regions. An inverse correlation between the rate of evolution of transcription factors and the number of genes that they regulate has been found in most of cases. Therefore, for small gene networks, distant species may not provide adequate support and in general the distance may depend on the size of the genetic network. Even close species may have evolved a different regulation for some genes due to adaptation to environment conditions. TFBS are very short so there is a size limit effect. More important, while the discrepancy between coding region models and the actual patterns of changes depends on not considering the three dimensional building of a protein, the discrepancy between a model of TFBS and the true pattern depends on the binding energies of the interaction between the transcription factor and the binding sites. So coding regions and TFBS are inherently different. This error can lead to assign a small distance on the tree due to the slow accumulation of mutations within orthologous genes while the binding factors might have been under stronger or more relaxed evolution. Therefore our metric should be considered simply as a first step of inching towards reality.

A major issue is the choice of a good genome background model. Third and fourth order Markov chains provide good solutions for bacteria and fungi. Higher eukaryotes have longer and different patterns of interdispersed repeats which require complex background models. Our procedure should allow to investigate the relationship between replicates and the strength of the regulatory network in a species, i.e. small variations in gene expression may mean that the gene network is finely regulated, therefore under strong purifying selection and does not allow for so much natural variation. The inclusion of phylogenetic information may tell us how much that the gene network has changed in different species in terms of number of co-regulated genes, motif patterns, changes in expression. Therefore an important result of including additional species is that it provides insights on the evolution the gene regulatory network under investigation. Finally knowing TFBS will allow to understand fluctuations and extreme values of gene expression, potentially revealing anomalous behaviors of gene networks.

References

- Angelini, C., Cutillo, L., De Feis, I., van der Wath, R., and Liò, P. (2007), Identifying regulatory sites using neighborhood species, in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics: 5th European Conference, EvoBIO 2007, Valencia, Spain, April 11-13, 2007, Proceedings*, edited by E. M. et al., Series: Lecture Notes in Computer Science 4447, pp. 1–10.
- Conlon, E.M., Liu, X.S., Lieb, J.D., and Liu, J.S. (2003), Integrating regulatory motif discovery and genome-wide expression analysis, *Proc. Natl. Acad. Sci. USA*, **100**: 3339–3344

- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburttty, K., Simon, J., Bard, M., Friend, S.H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, **102**:109-26
- Tadesse, M.G. , Vannucci, M., and Liò, P. (2004). Identification of DNA regulatory motifs using Bayesian variable selection, *Bioinformatics*, **20**: 2553–2561
- Tompa, M., *et al.* (2005). Assessing computational tools for the discovery of transcription factor ing sites. *Nat Biotechnol.*, **23**: 137–44