Statistical Relational Learning for Gene Regulatory Network Inference

F. d'Alché-Buc[1]
Joint work with C. Brouard[1], J. Dubois[1,2], C. Vrain[2],M-A. Debily[3]

[1] IBISC, Université d'Evry Val d'Essonne
[2] LIFO, Université d'Orleans, [3] CEA, Evry

Email : florence.dalche@ibisc.fr

Fundings : ANR GD2GS (Genomic Data to Graph Structure)

How to learn biological networks from data ?

- **Pre-processing:**
  - Dimension reduction and clustering approaches
- **Modeling approaches** : modeling the (dynamical) behavior of the network and identifying it; once estimated, the model can be used to simulate and predict the behaviour of the network as a system
- **Predictive approaches** : complete a partially known matrix, approximate the relation symbolized by the edges only in a supervised way from static or dynamical data; once predicted, the adjacency matrix will serve as prior knowledge to modeling approaches

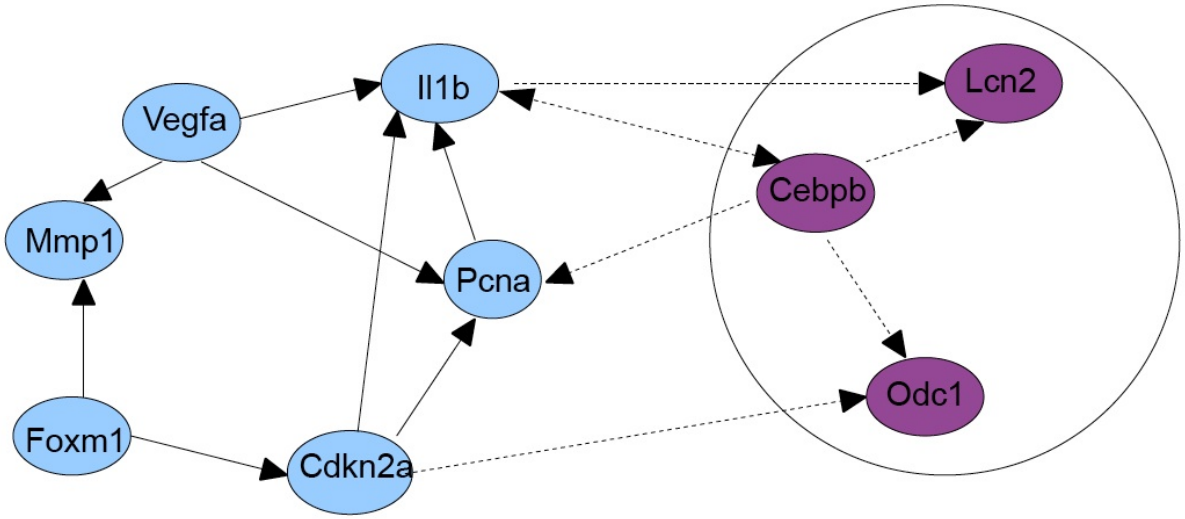
Our approaches to Network Inference

- **Predictive approaches**
  - Output Kernel Regression for semi-supervised link prediction in a protein-protein interaction network (Brouard et al. 2010)
  - Markov Logic Network for supervised link prediction in a gene regulatory network (Brouard et al. 2010)
- **Modeling approaches**
  - Model the evolution of the state of a gene regulatory network with a dynamical probabilistic model encapsulating an ODE (Quach et al. 2007, Fouchet, work in progress)
  - Unsupervised structure learning in a gene regulatory network (Auliac et al. 2008),with qualitative constraints Shenbabaoglu et al. work in progress)
  - Mixture of dynamical probabilistic models evolving in time for nonstationary networks (Bedo and d'Alché, work in progress)

Learning in a probabilistic setting

- Probabilistic Dynamical Models (Previous works)
  - $P(X_1, ..., X_T) = P_{\theta_1}(X_1) \prod_{t=1}^{T-1} P_\theta(X_{t+1}|X_t)$
  - X_t: state vector describing the network at time t
- Link Prediction Models (this talk)
  - $g_{W,\tau}(x_i, x_j) = sgn(P_W(y_{ij}|x_i, x_j) - \tau)$
  - Goal: estimate the posterior probability of a regulation relation given the description of two genes

Supervised link prediction for a gene regulatory network



- Available information
  - a set of known regulations between a regulator and a regulee
  - a corpus of knowledge about genes and their properties
- **Goal**: learning a classifier that is able to predict if given a couple of genes (gene A, gene B), gene A regulates gene B.
- We want to explore a new paradigm for supervised learning called Statistical Relational learning

Outline

Instanciation on a biological problem

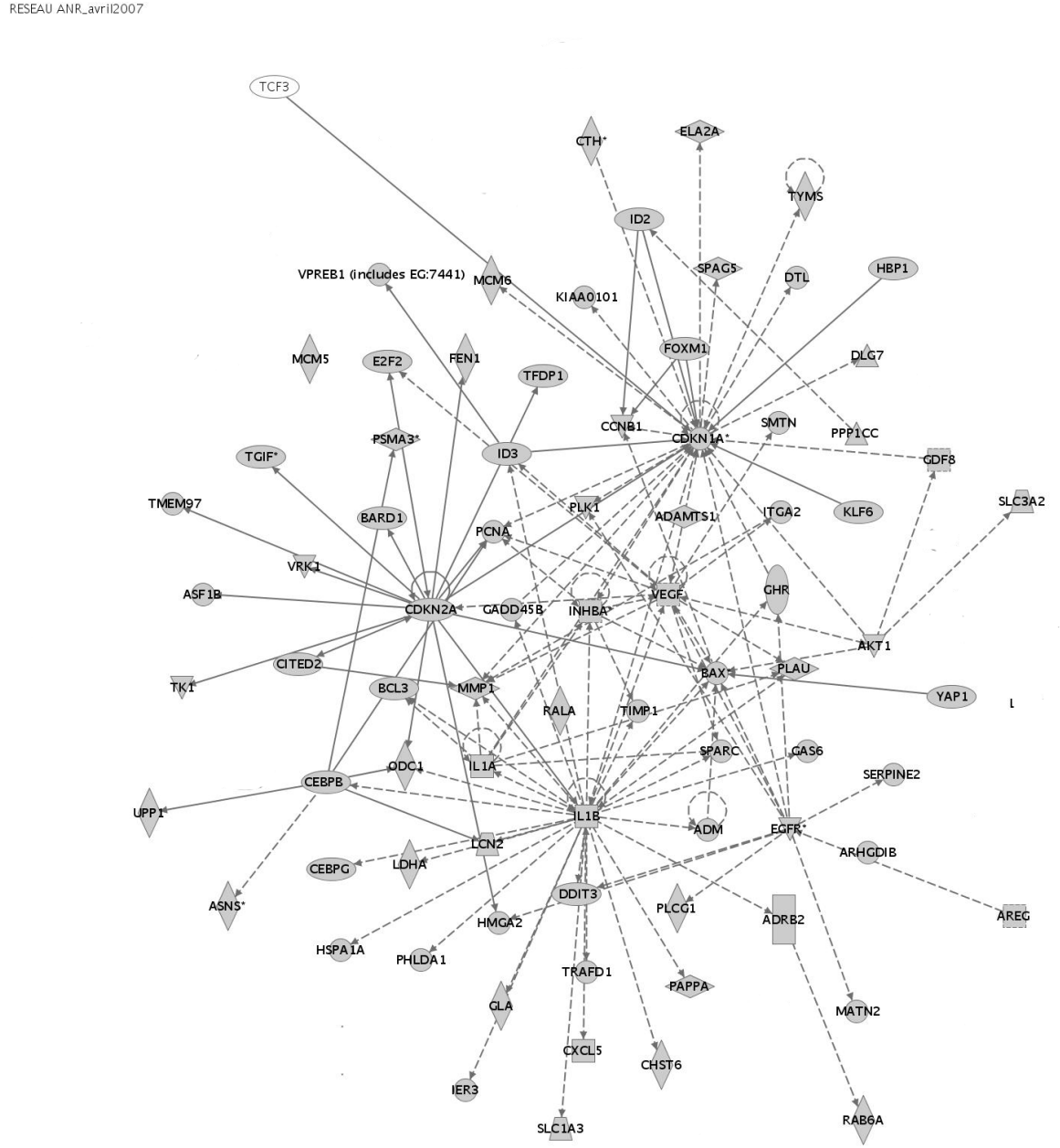
Switch proliferation/diffenciation of skin primary cells (human keratinocytes)

- Collaboration with a biologist of CEA : Marie-Anne Debily
- The laboratory of Xavier Gidrol has identified protein ID2 as a major component in this switch
- Transcriptomic analysis by microarray experiments of HaCaT cells presenting stable overexpression or transient knock-down achieved by RNA interference of ID2 expression.
- Selection of a subset of 63 differentially expressed genes
- **No kinetics here (unfortunately...)**

Goal :

Given a gene regulatory network provided by Ingenuity (text-mining), use experimental data and background knowledge to build a classifier devoted to link prediction

Gene regulatory network given by (Ingenuity)



Data

- 157 existing regulations (positive examples)

Experimental data for the 63 genes :

- expression level without a modification of ID2 expression level
- expression level after increasing or decreasing the expression level of ID2

Additional information

- genes position on chromosome
- cellular localization
- biological processes
- protein-protein interactions

Goal

- Build a classifier based on a set of weighted first oder logic rules that concludes on the target predicate **regulates**

If prop(a,C) and prop2(b) and prop3(b,a) then regulates(a,b)

- Advantages of relational learning or Inductive Logic Programming (ILP) :
  - interpretability of results
  - encoding heterogenous data into a single framework
- Drawback :
  - Learning is NP-difficult
  - Do not deal with noise

Statistical relational learning potentially allows one to combine advantages of ILP with powerful statistical inference methods

Markov Logic Network (MLN)

(introduced by Domingos et al., 2005)

Markov Logic Network (MLN)

- Let $\mathcal{X}$ be the set of all propositions describing a world (i.e. the et of all ground atoms)
- Let $\mathcal{F}$ be the set of all clauses in the MLN
- w_i is the weight (positive or negative) associated with the clause f_i, and $\mathcal{Z}$, the normalizing constant
- Then, the probability of a particular truth assignment x of variables in $\mathcal{X}$ is given by the formula:

$$P(\mathcal{X} = x) = \frac{1}{\mathcal{Z}} \exp(\sum_{f_i \in \mathcal{F}} w_i n_i(x))$$

Examples of predicates that encode experimental data and prior knowledge

- Expression data :
  - Expwt(gene, level), Expsiid2(gene,level), Expprcid2(gene,level)
  - For instance, Expsiid2(G,L) states that the level of expression of gene G is L when the level of expression of ID2 has been increased
- Position on chromosomes :
  - Memechro(gene, gene), Memebande(gene, gene)
- Physical interaction between proteins :
  - Interprot(gene, gene)
- Cellular localization of proteins
  - Loccell(gene, loc)
- Biological processes to which genes are contributing :
  - Processbio(gene, processus)

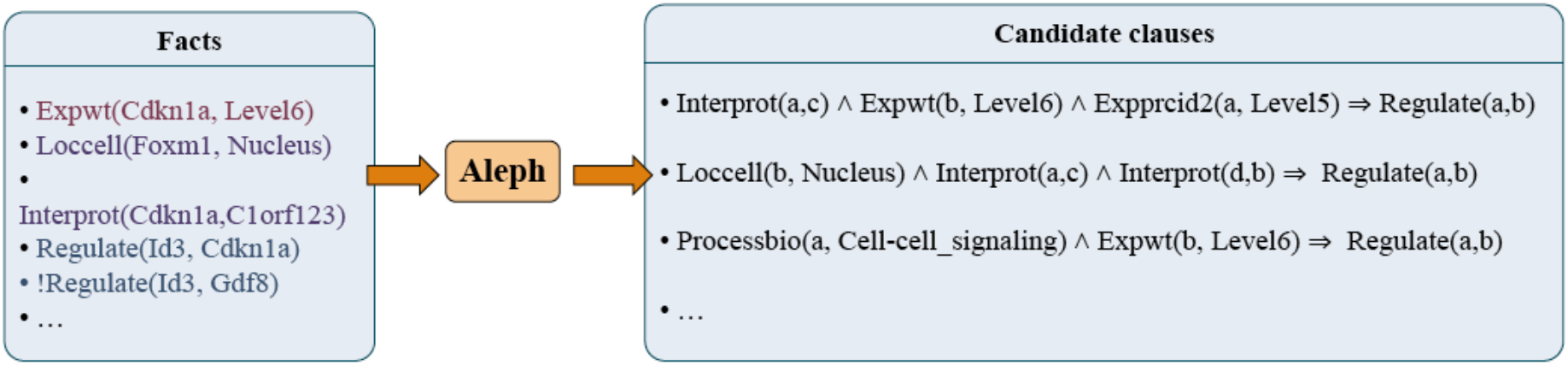
Discriminative learning of a MLN

A two-stage approach

- **Structure** learning: identify a set of candidate rules logiques
- Weight learning: given a set of candidate rules (the graph structure), determine the weights

Discriminative learning of the structure

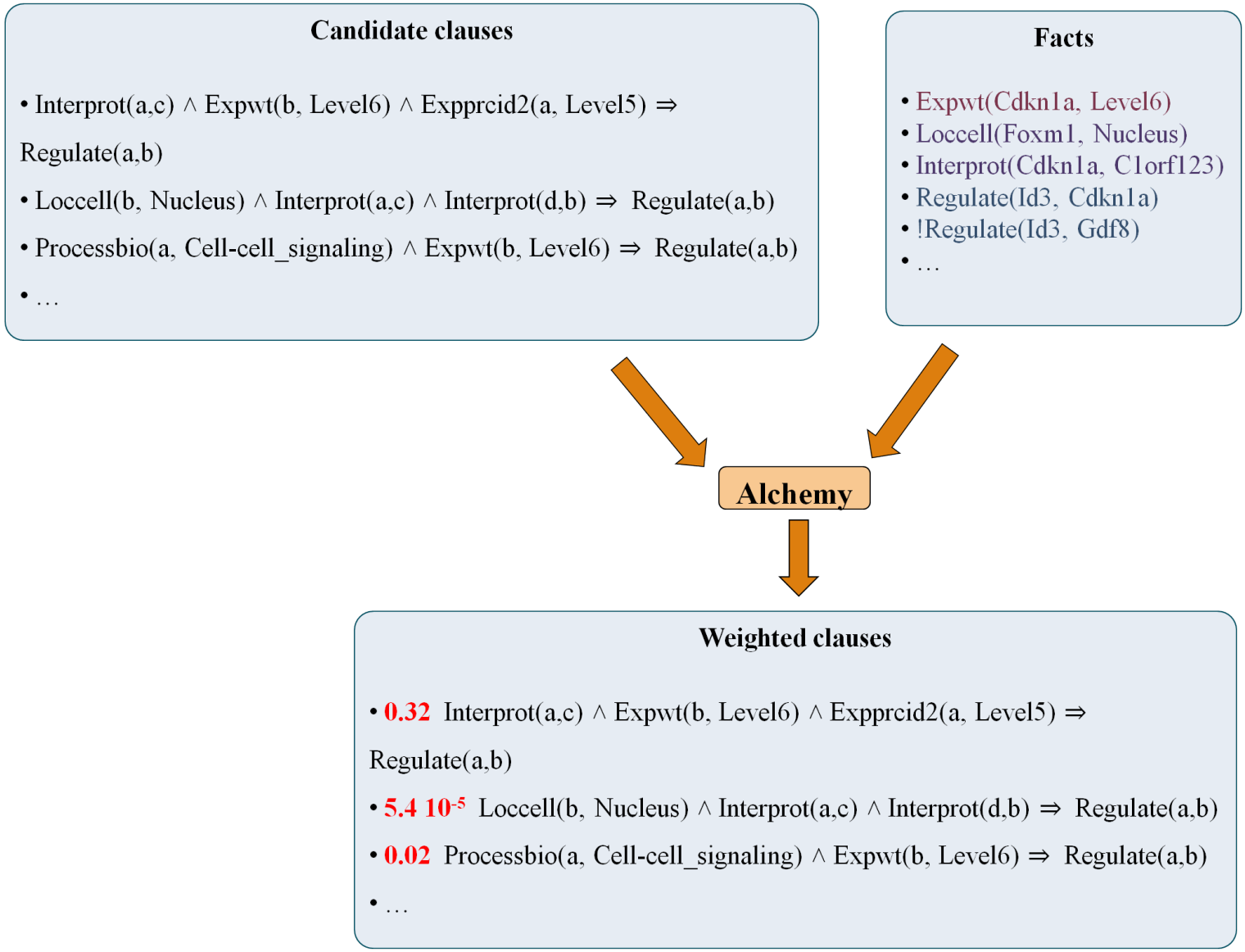
- Used tool: : **Aleph** (Srinivasan, 2001)
- Inductive Logic Programming



Learning procedure in Aleph

- **Aleph** (Srinivasan, 2001)
  - Selection of a positive example
  - Construction of the most specific rule satisfied by this example
  - Generalization of this rule by a top-down search
  - The process is iterated until all the positive examples be covered

Discriminative learning



Discriminative learning of weights (1)

Notations

- Let $\mathcal{Y}$ the set of query atoms (regulate predicate)
- $y = (y_{11}, \ldots, y_{nn})$ where y_{ij} correspond to the instanciated predicates **Regulate(**G_i**,** G_j**)**
  and thus to the labeled data.
- x correspond to all the other instanciated predicates

Maximization of the penalized conditional log-likelihood

$$\mathcal{L}(w) = \log P(\mathcal{Y} = y | X = x, w) + \log P(w) \quad (1)$$

$$= \sum_{i,j=1}^{n} \log P(\mathcal{Y}_{ij} = y_{ij} | X = x, w) + \log P(w) \quad (2)$$

Discriminative learning of weights (2)

$$P(y_{ij}|x,w) = \frac{\exp(\sum_{k \in \mathcal{F}_{y_{ij}}} w_k n_k(x, y_{ij}))}{\sum_{t=0,1} \exp(\sum_{k \in \mathcal{F}_{y_{ij}}} w_k n_k(x, y_{|Y_{ij}=t}))}$$

- ℓ_2 norm : $P(w) \propto \exp(-\lambda \parallel w \parallel^2)$
- Implementation with Alchemy (Kok et al.)

Experiments

Results 1

Positive examples

S^+ : 157 examples of regulation among 63 selected genes (Source: Ingenuity) Ingenuity

Negative examples

S^- : set of all the no-regulation links (3749)

- Unbalanced dataset:
- Given S^+, S^- is subsampled to provide 30 subsets of negative training examples S_i^- with $|S_i^-| = |S^+|$
- AUC-ROC is estimated by a10 fold-Cross-Validation for each $S^+ \cup S_i^-, i = 1...30$

Results 1 - Average behaviour on balanced datasets

- AUC-ROC and AUC-PR
- Different values tested for the regularization hyperparameter λ
- AUCs :

λ :	**AUC-ROC**	**AUC-PR**
20	0.803 ± 0.027	0.820 ± 0.030
50	0.821 ± 0.025	0.839 ± 0.025
100	$\mathbf{0.825 \pm 0.028}$	$\mathbf{0.847 \pm 0.027}$
500	0.822 ± 0.032	0.845 ± 0.031
750	0.818 ± 0.034	0.843 ± 0.032

Result 2: test with subbagging on an update of the network

- $\mathcal{G} = S^+ \cup S^-$: 2007 dataset
- S^+_{test} : test set with 51 new regulations: december 2009 dataset

Subbagging

- for $b = 1 \dots B$:
  - S^-_b = bootstrap subsample S^- with $|S^-_b| = |S^+|$
  - h_b = b-th classifier trained on (S^-_b, S^+)
- $H = \frac{1}{B} \sum_b h_b$

- Threshold τ selected τ by maximizing the F_1-measure (i.e. $F_1 = 2\frac{Pr \times Rec}{Pr + Rec}$)
- **Result** : 98% of good predictions

Results 3: Prediction between a test set of genes and the training set

- Selection of a new gene set by the biologist (M.-A. Debily)
  $\longrightarrow$ 24 genes obtained by a strict filtering process (genes differentially expressed and that can be described in same GO terms)
- Subbagging on $\mathcal{G}$
- Target task: completion of the links between the 63 training genes and the 24 new genes

AUC:	ROC
$\lambda = 50$	0.728
$\lambda = 100$	0.731
$\lambda = 500$	0.732
$\lambda = 750$	0.734

Which rules ?

- In general, rules are disappointing in the sense that some of the rules do not include properties on both genes (too general rules)
- Weights that can be negative of positive make the intepretation harder
- Example of a rule with positive weight:
  - $w = 0.19$ if $loccell(G2, plasma_membrane)$ and $expsiid2(G2, level3)$ and $expsiid2(G1, level3)$ then $regulates(G1, G2)$
- Need to get a better encoding of some of the properties (GO) to get more specific rules

Conclusion

- Markov Logic Network: provides a way to combine first order logic with statistical inference (here: MAP approaches)
- First order logic (FOL): a framework to encode heterogeneous information
- Unbalanced datasets can be handled with subbagging algorithms
- Literature data are confronted to experiments

Perspectives

- To benefit from FOL, better encoding of biological properties is needed
- First Order Logic alone is not sufficient: numerical constraints are needed
- Are there alternative and simpler models without FOL: need to compare to pure quantitative models (we solved this task for non oriented graphs with Output Kernel Regression)
- Mid-term Goal: Combine in a unified probabilistic framework supervised approaches and unsupervised ones.