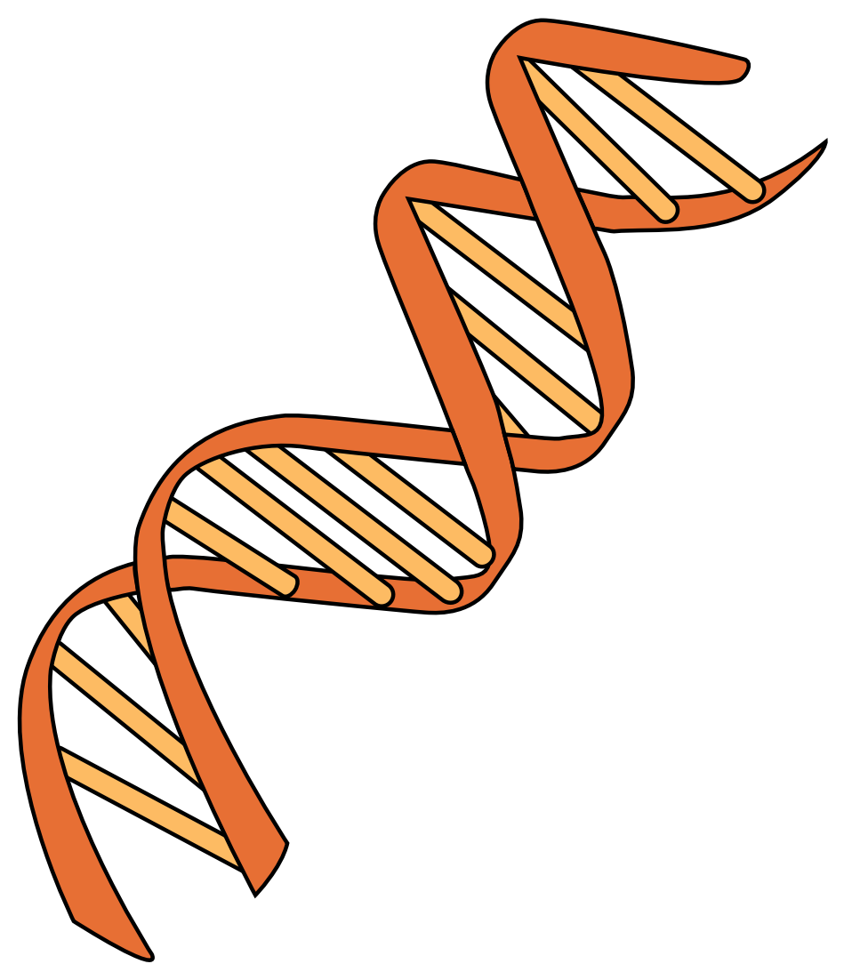
LOOKING FOR APPLICATIONS OF MIXTURES OF MARKOV TREES IN BIOINFORMATICS

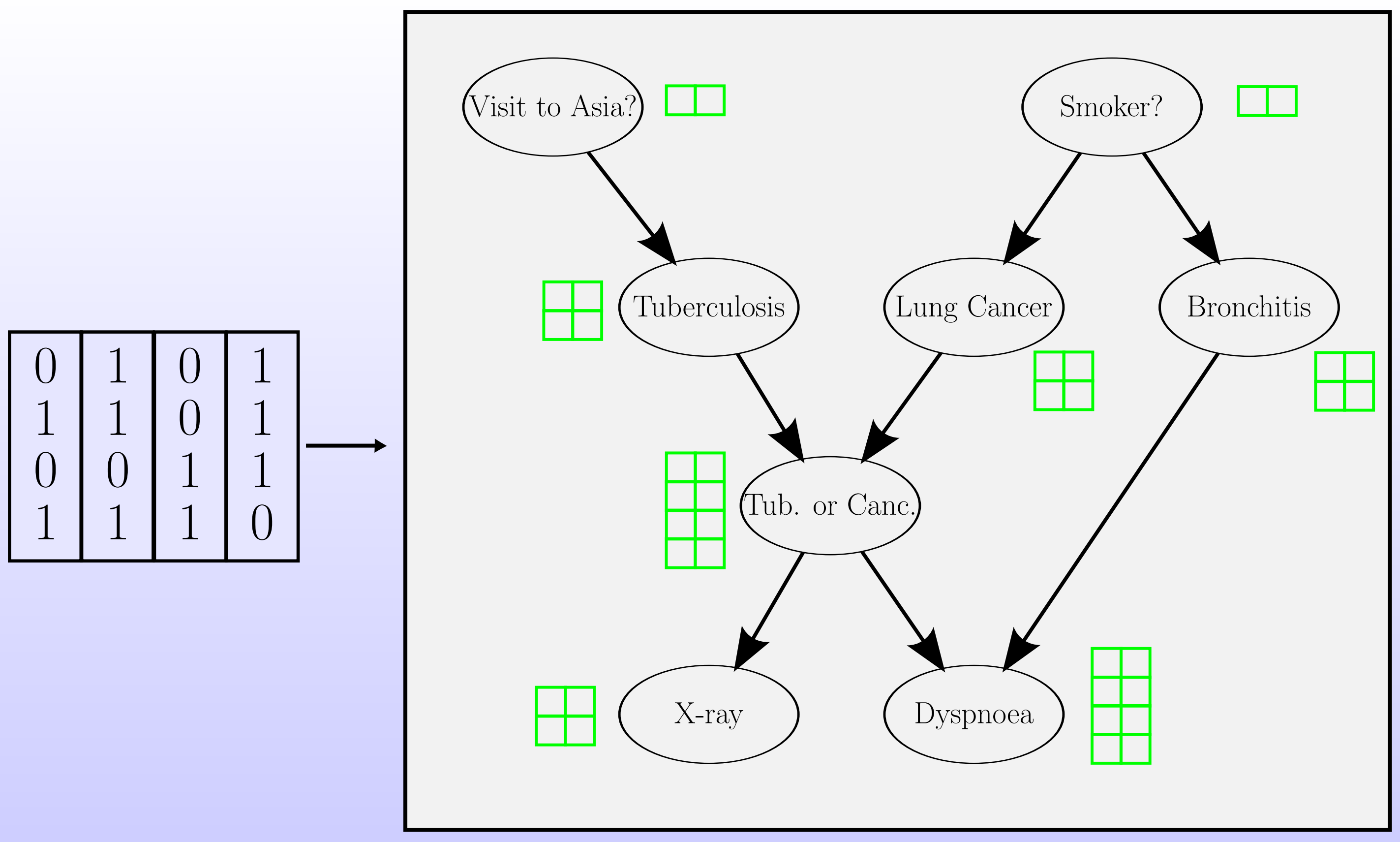
F. Schnitzler, P. Geurts, L. Wehenkel

Systems and Modeling & GIGA research, University of Liège, Belgium

fschnitzler@ulg.ac.be

Probabilistic graphical models (PGM) efficiently encode a probability distribution on a large set of variables. While they have already had several successful applications in biology, their **poor scaling** in terms of the number of variables may make them unfit to tackle problems of increasing size. **Mixtures of trees** however scale well by design. Experiments on synthetic data have shown the interest of our new learning methods for this model, and we now wish to apply them to relevant problems **in bioinformatics**.





PGM Example : In Bayesian networks, the product of conditional probability distributions is the joint probability distribution over all variables.

$$\mathbb{P}(X|S,\theta) = \prod_{i=1}^{n} P_\theta(X_i|Pa_S(X_i))$$

Interest :

- Inference of conditional distribution : $\mathbb{P}(A|B)$ where $A, B \in X$,
- Maximum a posteriori probability inference : $a = \arg\max_A \mathbb{P}(A|B)$ where $A, B \in X$,
- Visual representation of relationships between variables.

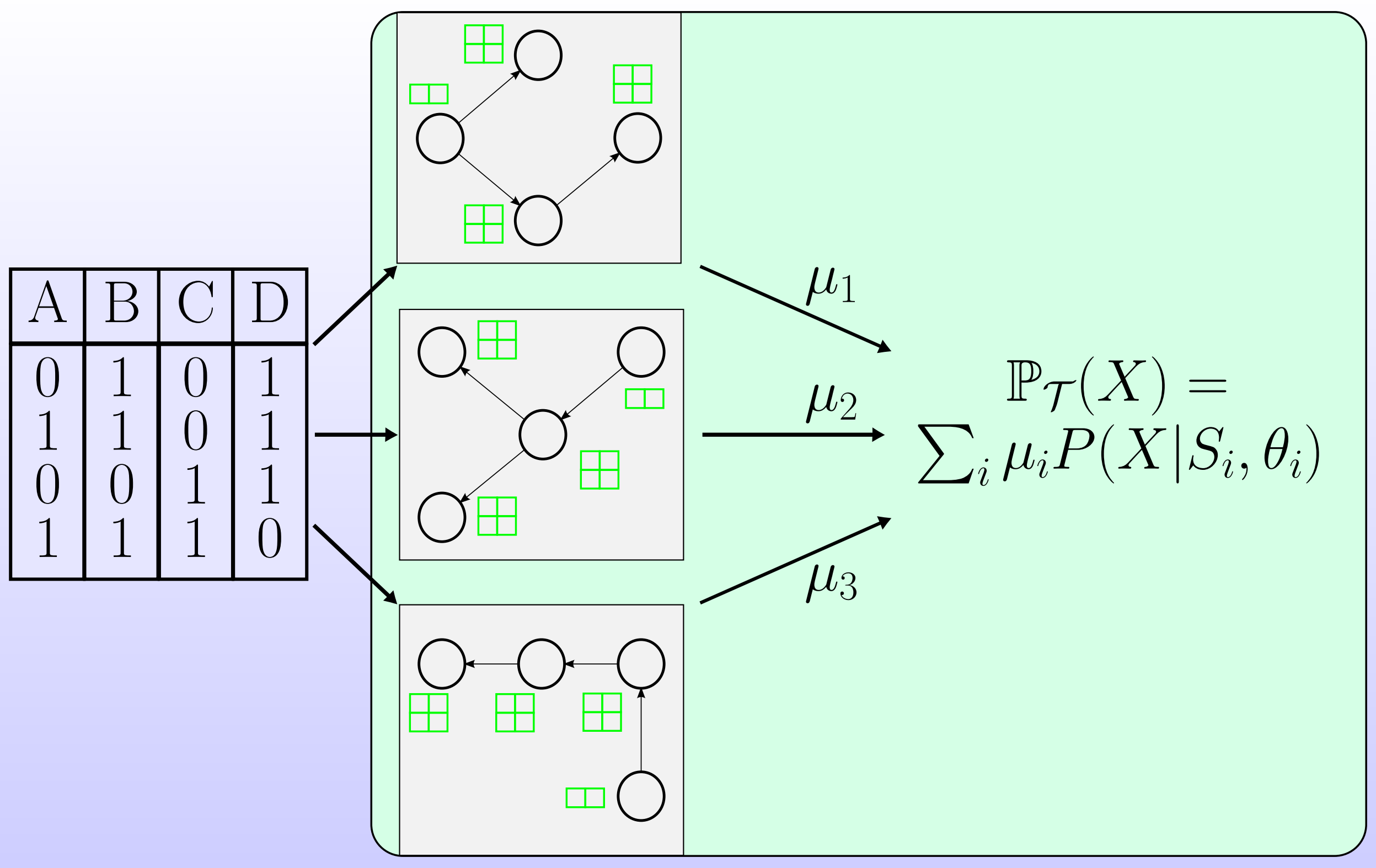
Problems : Intractable algorithms: Learning and inference are np-hard, and in practice are challenging on more than 1000s variables.

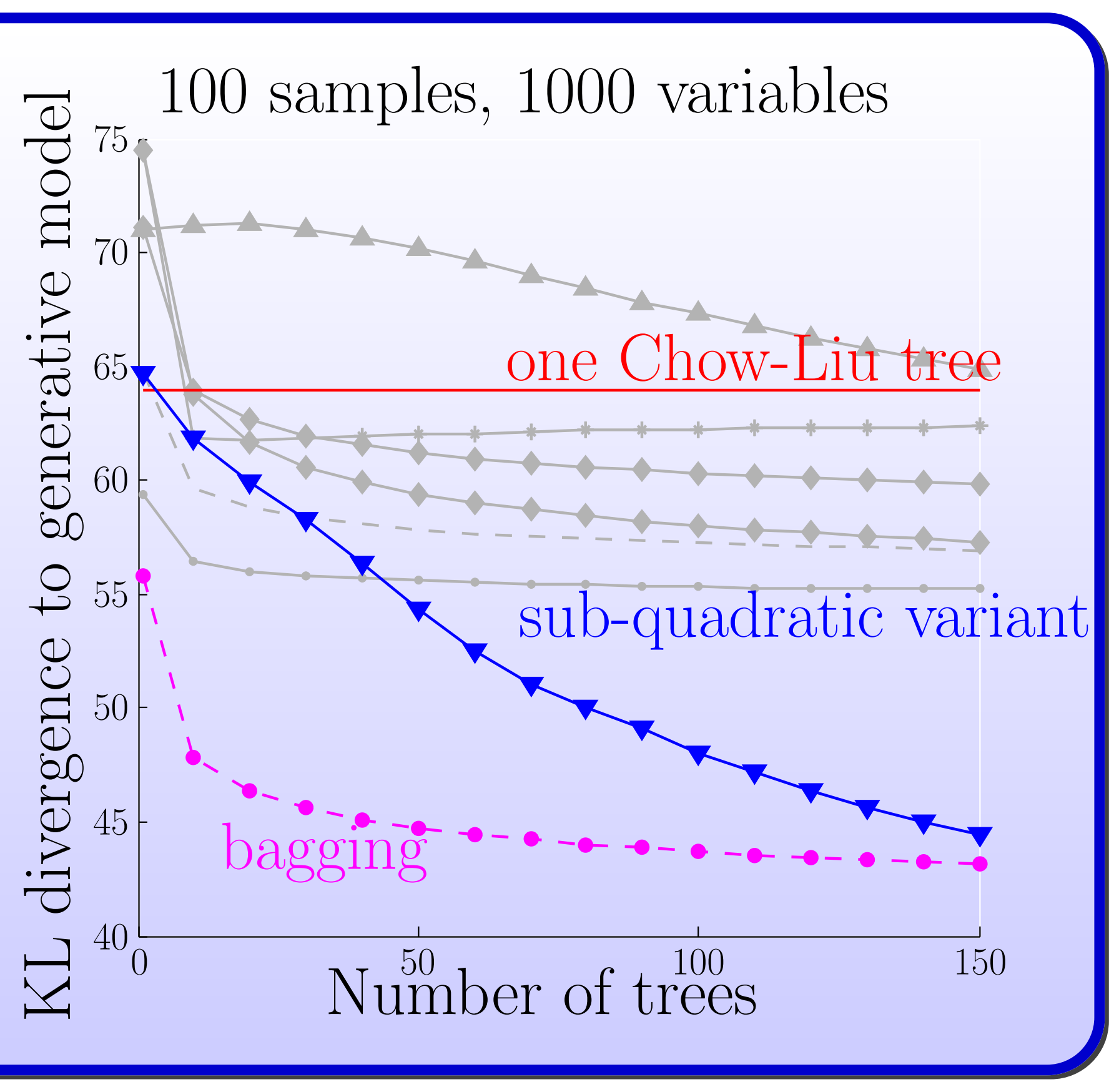
Mixtures of trees $\mathcal{T}$:

- Generate an ensemble of m tree structures $\{S_i\}$.
- Determine their parameter vectors $\{\theta_1^*, \ldots, \theta_m^*\}$.
- Average their predictions by using some appropriate weighting scheme $\{\mu_1, \ldots, \mu_m\}$.

Key points:

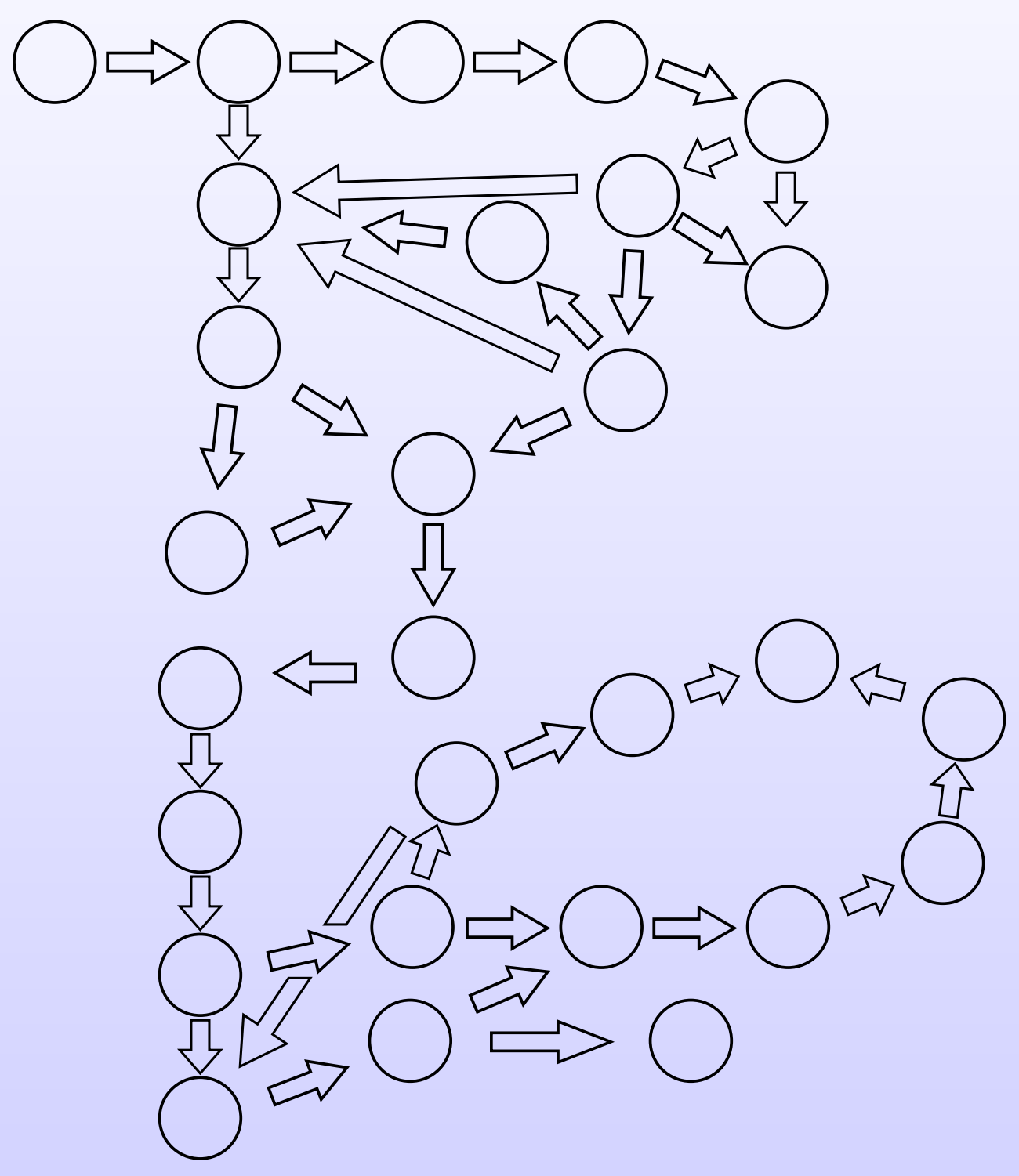
- Trees $\rightarrow$ efficient algorithms.
- Mixture $\rightarrow$ improved modeling power.



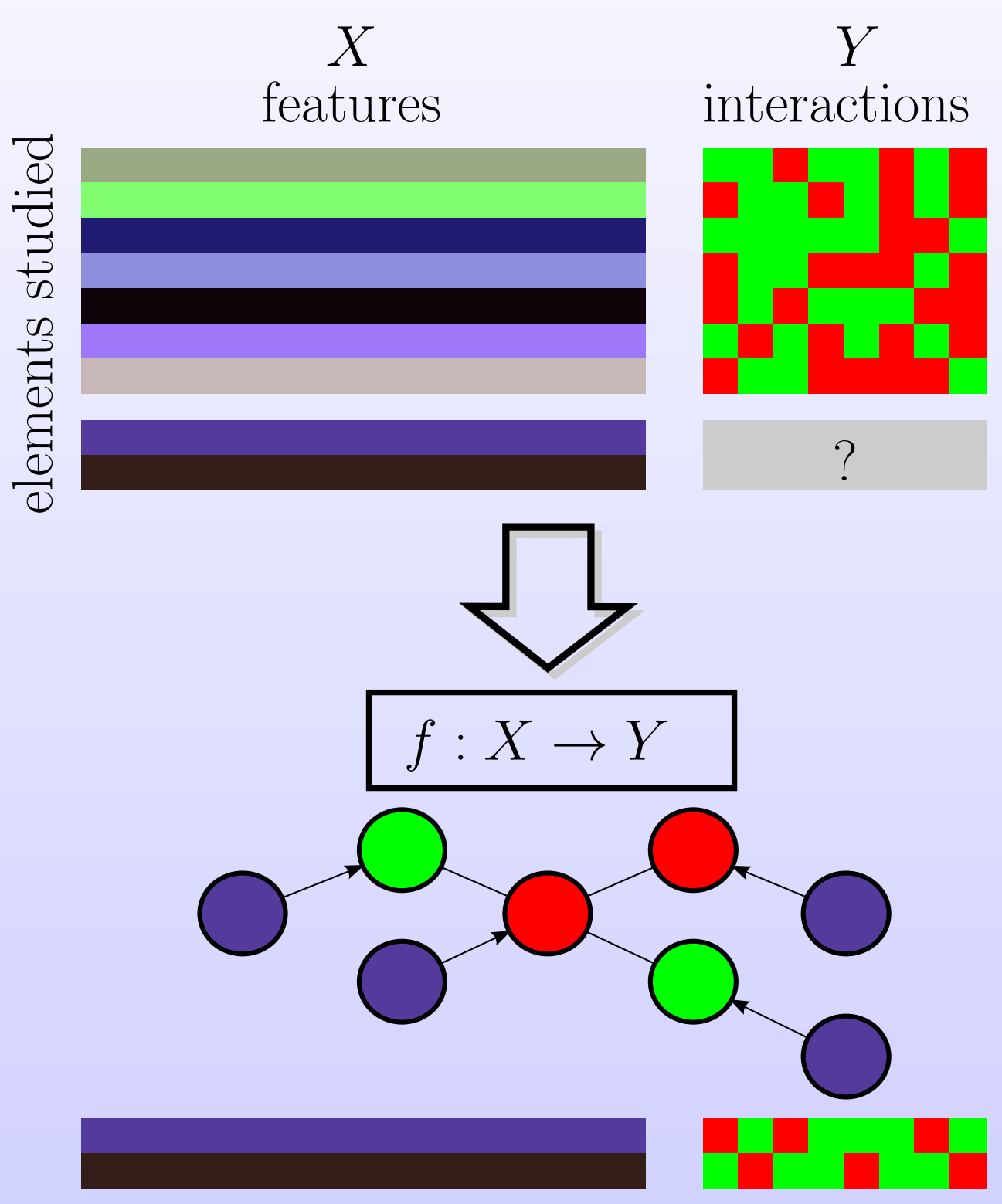


How could those models be useful in bioinformatics?

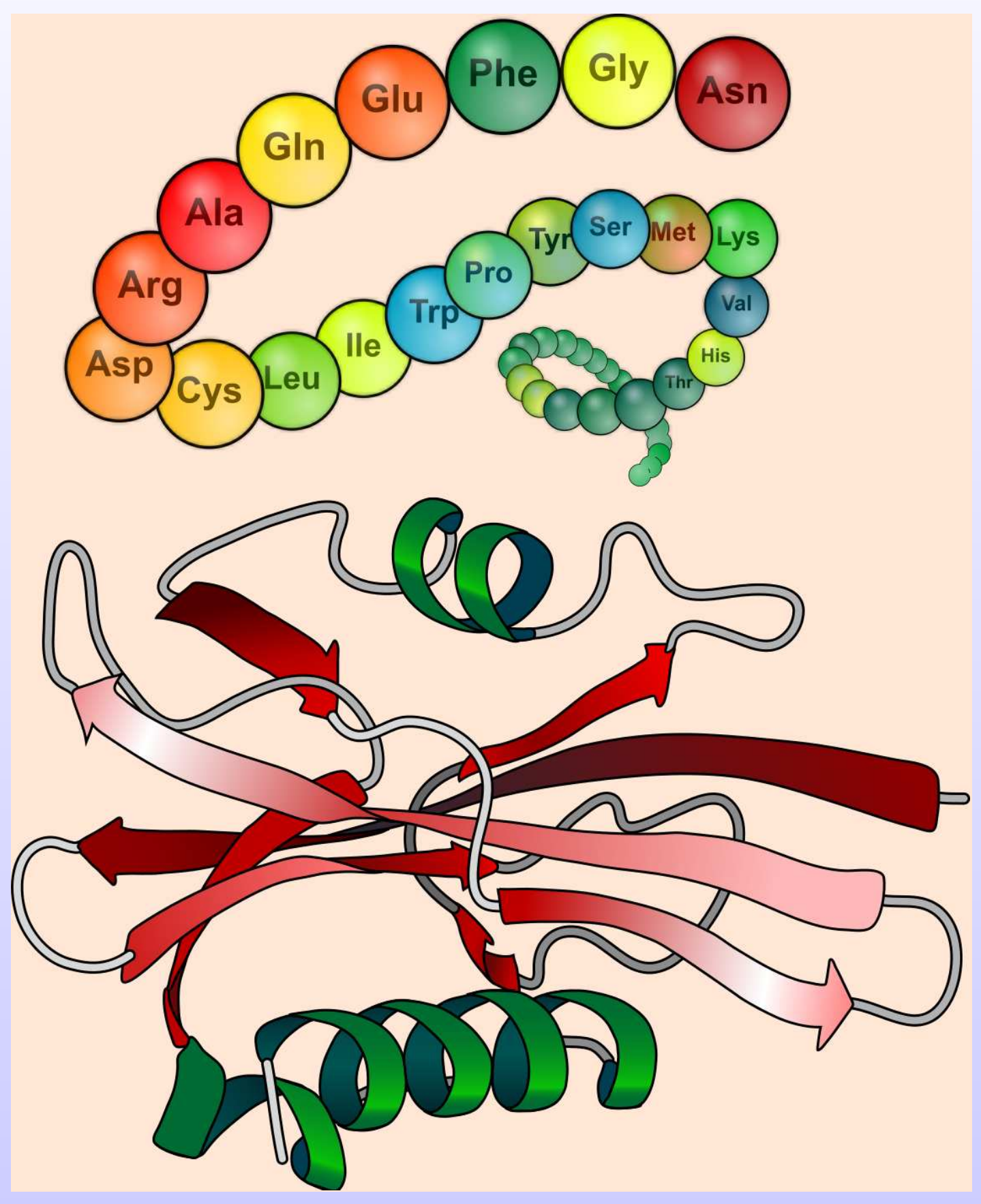
Microarray Analysis



Interactions prediction



Protein structure inference



?

References

Ammar, S., Leray, P., Schnitzler, F., Wehenkel, L. (2010). *Sub-quadratic Markov tree mixture learning based on randomizations of the Chow-Liu algorithm.* In Proceedings of the Fifth European Workshop on Probabilistic Graphical Models (PGM-2010) (pp. 17-24).

Needham, C., Bradford, J., Bulpitt, A., Westhead, D. (2007) *A Primer on Learning in Bayesian Networks for Computational Biology.* PLoS Comput Biol. 2007;3:e129

Acknowledgement

This work was funded by the Belgian Fund for Research in Industry and Agriculture (FRIA). This poster presents research results of the Belgian Network BIOMAGNET, funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author.