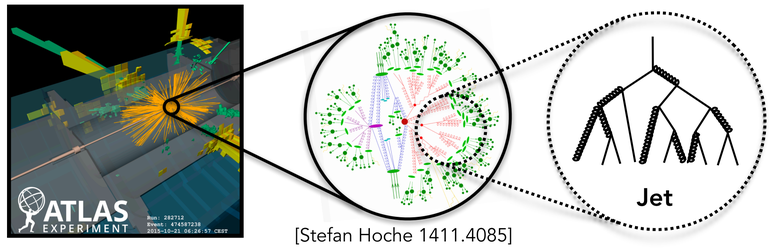
Exact Inference on Hierarchical Clustering in Particle Physics and Cancer Genomics

Craig S. Greenberg[1], Sebastian Macaluso[2], Nicholas Monath[1], Ji-Ah Lee[1], Patrick Flaherty[1], Kyle Cranmer[2], Andrew McGregor[1], Andrew McCallum[1]

[1] University of Massachusetts Amherst, [2] New York University

Jet Physics

- Essentially every collision at the Large Hadron Collider at CERN has one or more jets (sprays of particles from quarks & gluons).
- Jets originate from a *showering* process where an initial (unstable) particle goes through successive binary splittings, represented by a tree structure (hierarchical clustering).
- Inverting this showering process is a common task in data analysis, i.e. to infer the underlying tree structure of a jet.
- Current algorithms are greedy and based on heuristics.



[Stefan Hoche 1411.4085]

Cancer Genomics

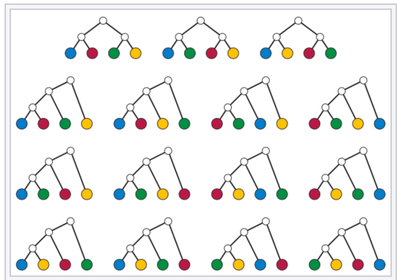
- Model subtypes of cancer to help determine prognosis and treatment plans.
- Hierarchical clustering is a common clustering approach for gene expression data.
- Standard approach uses uses a greedy agglomerative or divisive heuristic to build a tree.

Challenge

- Number of clustering histories for N elements grows super-exponentially

$$a(N) = (2N-3)!!$$

# of leaves	Approx. # of trees
4	15
5	100
7	10 k
9	2 M
11	600 M



The fifteen different rooted binary trees (with unordered children) on a set of four labeled leaves, illustrating $15 = (2 \times 4 - 3)!!$ (see article text).

Wikipedia-DoubleFactorial

Hierarchical Cluster Trellis Algorithm

- Directed acyclic graph data structure for exact inference in hierarchical clustering.
- Consider an energy-based probabilistic model.
- Measure the compatibility of each pair of sibling nodes for dataset X, described by a potential function $\Psi : 2^X \times 2^X \to \mathbb{R}^+$
- Probability of hierarchical clustering

$$P(\mathtt{H}|X) = \frac{\phi(X|\mathtt{H})}{Z(X)} \quad \text{with} \quad \phi(X|\mathtt{H}) = \prod_{X_L, X_R \in \text{siblings}(\mathtt{H})} \psi(X_L, X_R)$$

Partition Function

- Computed for every vertex in a bottom-up approach, memoizing the partial value at each vertex.

$$Z(X) = \sum_{\mathtt{H} \in \mathcal{H}(X)} \phi(X|\mathtt{H})$$

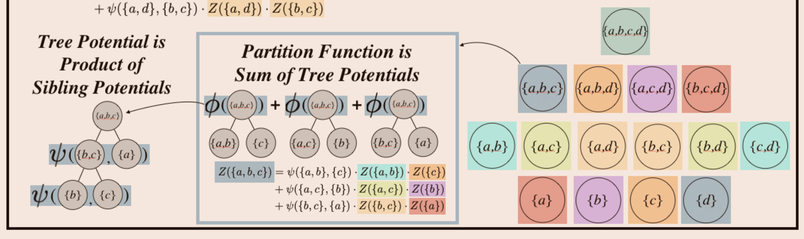
Exhaustive Computation of the Partition Function - *O((2N-3)!!)*

$$\begin{aligned} Z(\{a,b,c,d\}) = {} & \psi(\{a,b,c\},\{d\}) \cdot \psi(\{a,b\},\{c\}) \cdot \psi(\{a\},\{b\}) \\ & + \psi(\{a,b,c\},\{d\}) \cdot \psi(\{a,c\},\{b\}) \cdot \psi(\{a\},\{c\}) \\ & + \psi(\{a,b,c\},\{d\}) \cdot \psi(\{b,c\},\{a\}) \cdot \psi(\{b\},\{c\}) \\ & + \psi(\{a,c,d\},\{b\}) \cdot \psi(\{a,c\},\{d\}) \cdot \psi(\{a\},\{c\}) \\ & + \psi(\{a,c,d\},\{b\}) \cdot \psi(\{a,d\},\{c\}) \cdot \psi(\{a\},\{d\}) \\ & + \psi(\{a,c,d\},\{b\}) \cdot \psi(\{c,d\},\{a\}) \cdot \psi(\{c\},\{d\}) \\ & + \psi(\{a,b,d\},\{c\}) \cdot \psi(\{a,b\},\{d\}) \cdot \psi(\{a\},\{b\}) \\ & + \psi(\{a,b,d\},\{c\}) \cdot \psi(\{a,d\},\{b\}) \cdot \psi(\{a\},\{d\}) \\ & + \psi(\{a,b,d\},\{c\}) \cdot \psi(\{b,d\},\{a\}) \cdot \psi(\{b\},\{d\}) \\ & + \psi(\{b,c,d\},\{a\}) \cdot \psi(\{b,c\},\{d\}) \cdot \psi(\{b\},\{c\}) \\ & + \psi(\{b,c,d\},\{a\}) \cdot \psi(\{b,d\},\{c\}) \cdot \psi(\{b\},\{d\}) \\ & + \psi(\{b,c,d\},\{a\}) \cdot \psi(\{c,d\},\{d\}) \cdot \psi(\{c\},\{d\}) \\ & + \psi(\{a,b\},\{c,d\}) \cdot \psi(\{a\},\{b\}) \cdot \psi(\{c\},\{d\}) \\ & + \psi(\{a,c\},\{b,d\}) \cdot \psi(\{a\},\{c\}) \cdot \psi(\{b\},\{d\}) \\ & + \psi(\{a,d\},\{b,c\}) \cdot \psi(\{a\},\{d\}) \cdot \psi(\{b\},\{c\}) \end{aligned}$$

Computation using Trellis - *O(3^N) << O((2N-3)!!)*

$$\begin{aligned} Z(\{a,b,c,d\}) = {} & \psi(\{a,b,c\},\{d\}) \cdot Z(\{a,b,c\}) \cdot Z(\{d\}) + \psi(\{a,b,d\},\{c\}) \cdot Z(\{a,b,d\}) \cdot Z(\{c\}) \\ & + \psi(\{a,c,d\},\{b\}) \cdot Z(\{a,c,d\}) \cdot Z(\{b\}) + \psi(\{b,c,d\},\{a\}) \cdot Z(\{b,c,d\}) \cdot Z(\{a\}) \\ & + \psi(\{a,b\},\{c,d\}) \cdot Z(\{a,b\}) \cdot Z(\{c,d\}) + \psi(\{a,c\},\{b,d\}) \cdot Z(\{a,c\}) \cdot Z(\{b,d\}) \\ & + \psi(\{a,d\},\{b,c\}) \cdot Z(\{a,d\}) \cdot Z(\{b,c\}) \end{aligned}$$

Recursive Computation of Z using Memoization



Maximum Likelihood Hierarchical Clustering

- Also computed for every vertex in a bottom-up approach.

$$\mathtt{H}^\star(X) = \text{argmax}_{\mathtt{H} \in \mathcal{H}(X)} P(\mathtt{H}|X)$$

Experiments

Jet Physics

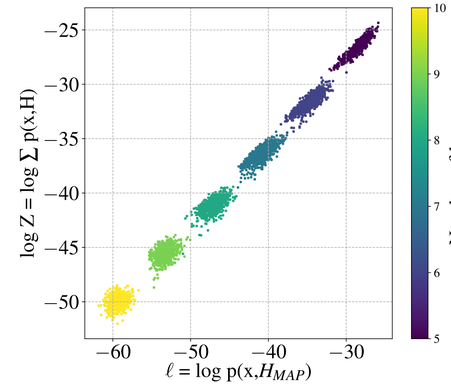
Cranmer, Macaluso & Pappadopulo

- We use simulated datasets from a toy generative model for jet physics, Ginkgo.
- Ginkgo captures essential ingredients of generative models in full physics simulations and has a tractable likelihood.

	Beam Search	Greedy
Trellis	0.4 ± 0.5	1.5 ± 1.1
Beam Search		1.1 ± 1.1

Average difference in maximum likelihood computation over 5000 datasets.

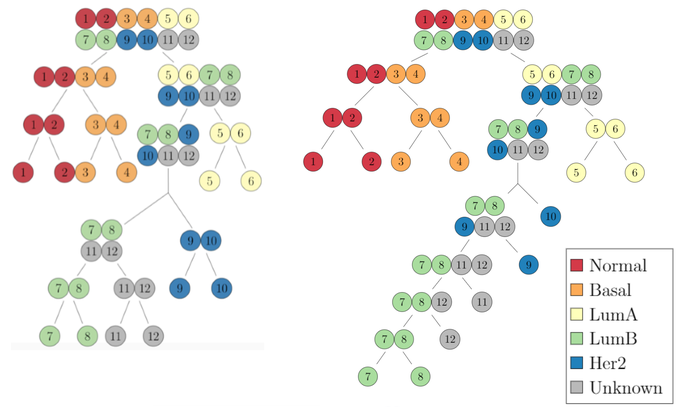
- The trellis allows us to find the exact maximum likelihood, which improves on greedy and beam search baselines.
- The trellis allows us to obtain the partition function.



Cancer Genomics

- Energy model as the sum of the positive edges from elements in one child to elements in the other one, minus the negative edges between two elements in the same child.
- Prediction Analysis of Microarray 50 (pam50) gene expression dataset.

Approximate Tree via Greedy | **MAP Tree via Trellis**



arXiv: 2002.11661