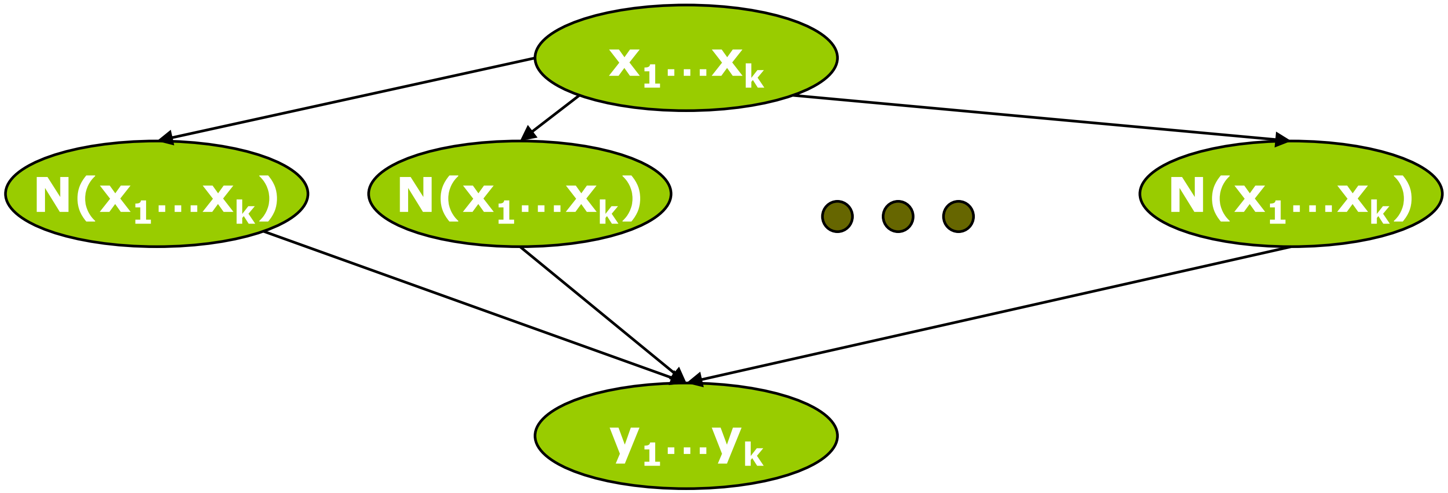
Global Alignment of Molecular Sequences via Ancestral State Reconstruction

ICS 2010

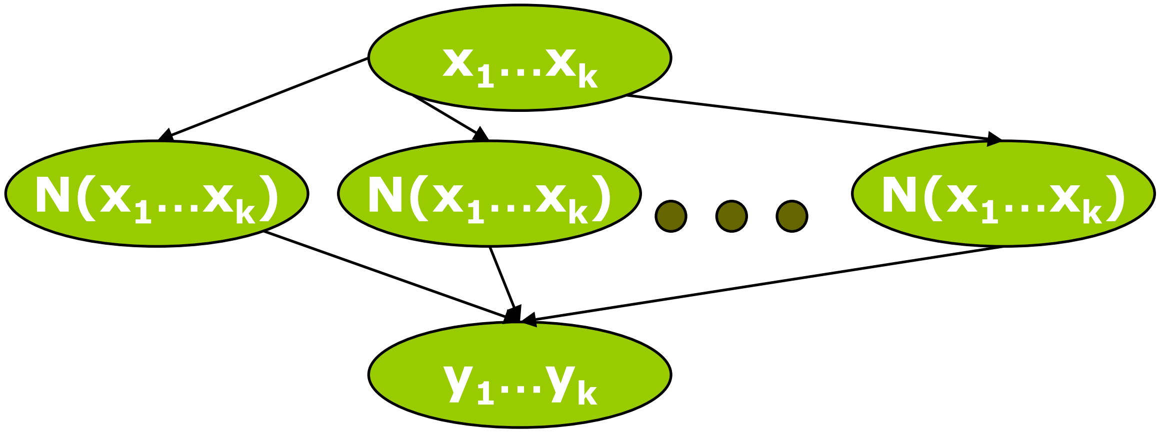
Alex Andoni, Costis Daskalakis, Avinatan Hassidim and Sebastien Roch

Capacity of noisy channels or Trace Reconstruction on a star



- Choose k random bits $x_1 \ldots x_k$
- N – some noisy channel
- Goal: Given many applications of $N(x_1 \ldots x_k)$ reconstruct $y_1 \ldots y_k$ s.t. $\Pr(x_i = y_i) \geq 0.99$
- How many channel uses do we need?

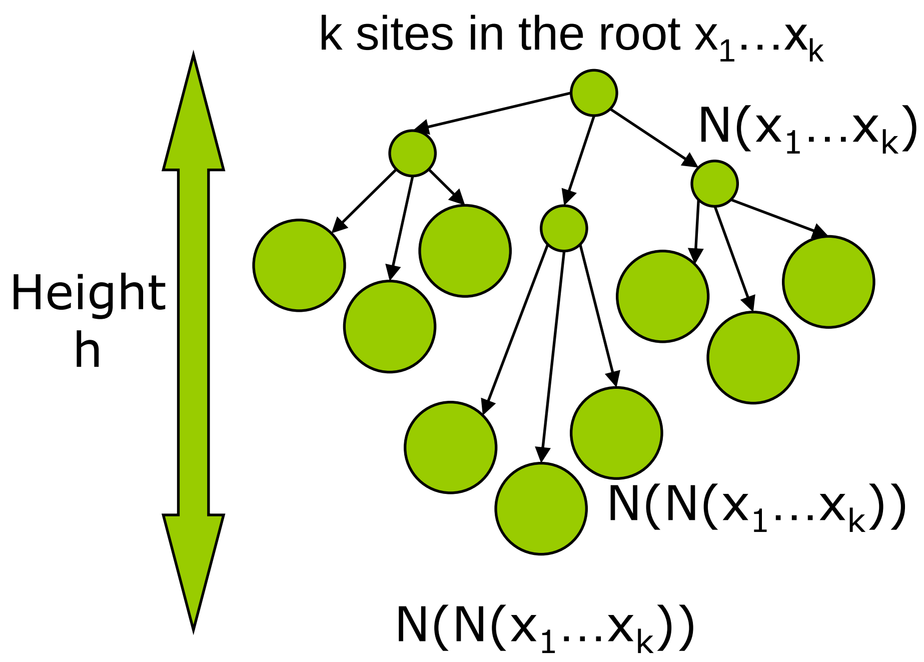
Number of channel uses



- N applies i.i.d substitutions: constant number of uses (bit wise majority)
- N applies i.i.d. deletions, with constant probability – poly(k) uses [HMPW08]
- Both insertions and deletions, more general channels, subconstant probabilities – many open questions

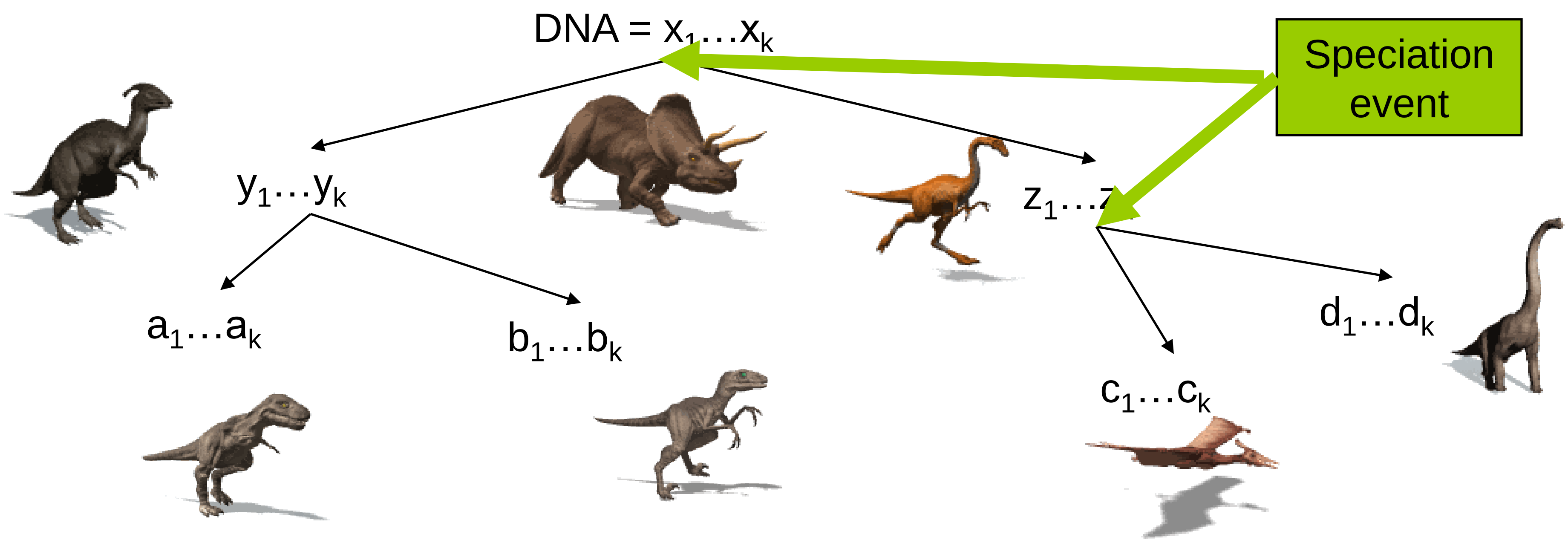
Trace reconstruction on a tree

- A recursive variant of trace reconstruction on a star
- On each edge, there is a probability for insertions, deletions and substitutions
- We are interested at a constant expansion ratio d



Motivation

- Study of more general noisy channels
- Phylogenetic reconstruction



- Statistical Physics

Main result

- Consider a d-ary tree, and a channel N:
  - N applies i.i.d substitutions with probability p_s, s.t.

$$(1-2p_s)^2 > O\left(\frac{\log d}{d}\right)$$

  - N applies i.i.d insertions with probability at most $O(1/k^{2/3}h)$
  - N applies i.i.d insertions with probability at most $O(1/k^{2/3}h)$
- Then one can "reconstruct" $x_1 \ldots x_k$ from the leaves of the tree: Find $y_1 \ldots y_k$ s.t. $\Pr(x_i=y_i)>0.99$
- Some lower bounds:
  - Maximum substitution probability (without indels)

$$(1-2p_s)^2 > 1/d$$

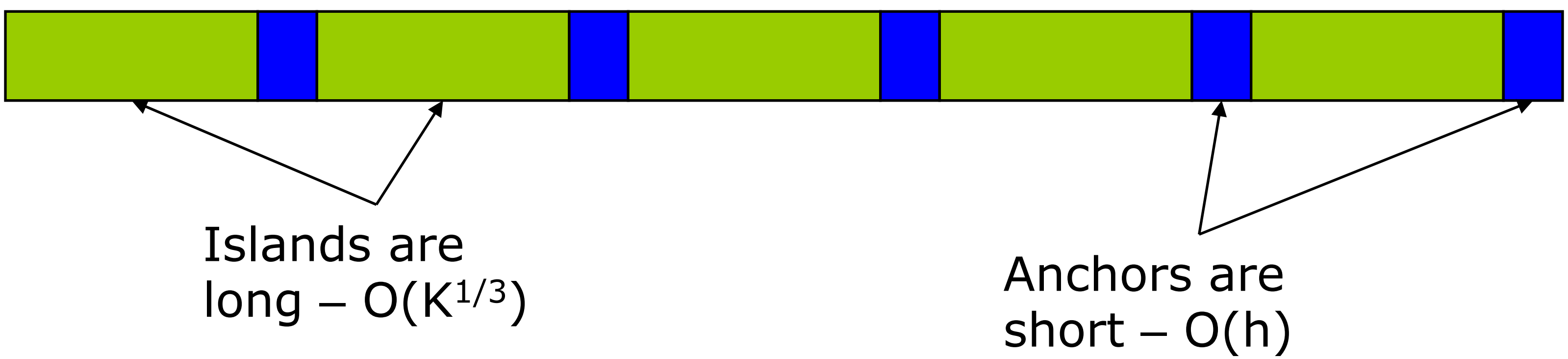
  - 1/h dependency corresponds to a constant fraction of deletions in the star case

Recursive reconstruction (Mos98)

- Reconstruct the tree layer by layer
- Given d input vertices, reconstruct their father
- Continue recursively until the root is reconstructed
- Challenges:
  - Sometimes the reconstruction fails
  - Even when the reconstruction succeeds, and the children are perfect, the father is reconstructed up to some noise

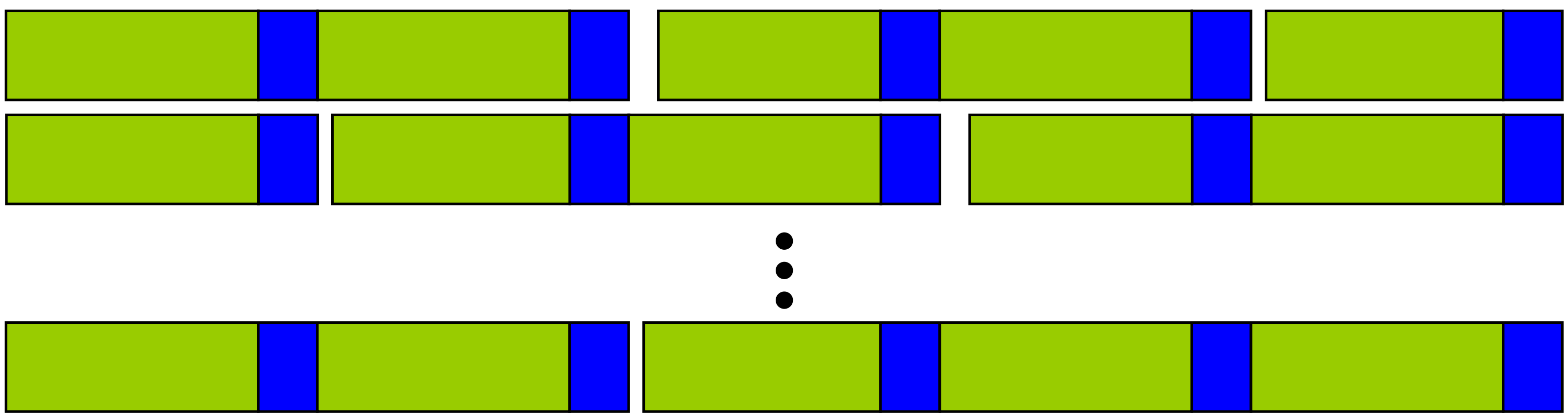
Reconstructing the father from the children

- Each Child is divided into anchors and islands



Reconstruction cntd.

$x_i = \mathrm{Maj}(a_i, b_i, .., e_i)$



- Align the children according to the anchors
- Do a place-wise majority on the children

Where can we succeed?

- We can not reconstruct all the vertices correctly
  - Suppose the first bit gets deleted going from the father to all the children

- Call a vertex v good if all three hold:
  - There are no indel operations in the anchor, when going from v to its children
  - In each island of v, there is at most one indel operation
  - It has at least d-2 good children

- The algorithm reconstructs all good vertices

Correctness of the reconstruction

- Main Result follows from two theorems
- Thm 1: With high probability, the root is good
- Thm 2: The algorithm reconstructs all good vertices correctly

Thm 1 – w.h.p the root is good

- Proof sketch: Show by induction on the height that most vertices are good.
- When is a vertex not good?
  - When there are indels in the anchors: Improbable event, as anchors are short
  - When there are two deletions in the same island: Improbable event, the islands have length $k^{1/3}$, and the indel probability is $1/k^{2/3}$
  - When two children are not good: improbable event by induction hypothesis
- Probability that the root is good >0.99

Thm 2: reconstruction of good vertices

- Proof is by induction.
- Suppose v is good. All good children (> d-1) reconstructed "correctly"
- Reconstructed children + No indels in anchors → Alignment of the anchors is "correct"
- Correct alignment + each island suffers at most one deletion → y_i is a majority of d values, such that d-2 of them are his true descendents
- Given that the majority is on the right descendents, we do not need to worry about the indels. Thm 2 holds because majority does error correction.

Open questions

- Adversarial root
- Can we use these techniques to say something about the star case?
- Improving the parameters. In particular, a weaker definition of reconstruction, with higher deletion probabilities.
- Can we do something even without the tree?
  - Follow up work shows how to reconstruct the topology of the tree (ABH'09)

Thank you