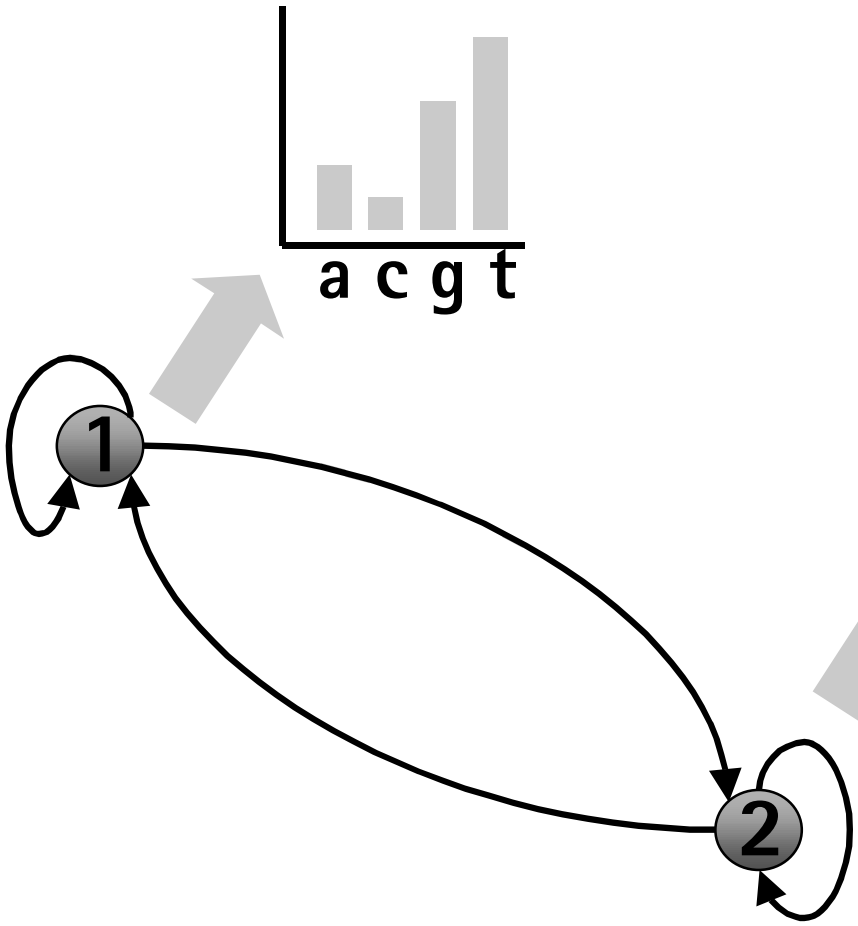
GHMM and HMMEd

Bernhard Knab *(Bayer AG, Leverkusen)*
Alexander Schliep *(Max Planck Inst for Molecular Genetics , Berlin)*
Barthel Steckemetz *(Science Factory GmbH, Köln)*
Bernd Wichern *(ifb AG, Köln)*
Achim Gädke, Peter Pipenbacher, Disa Thorarinsdottir *(ZAIK, U Köln)*

Modelling Sequences

- Biological sequences and associated features (DNA, RNA, Proteins)
- Time-series data
  - Gen-expression
  - Cellular processes (ion-channel)
  - Finance, Medicine, …
- General setting: Assume a statistical process generated data

Example: HMM



Emission probabilities

MC

Extensions

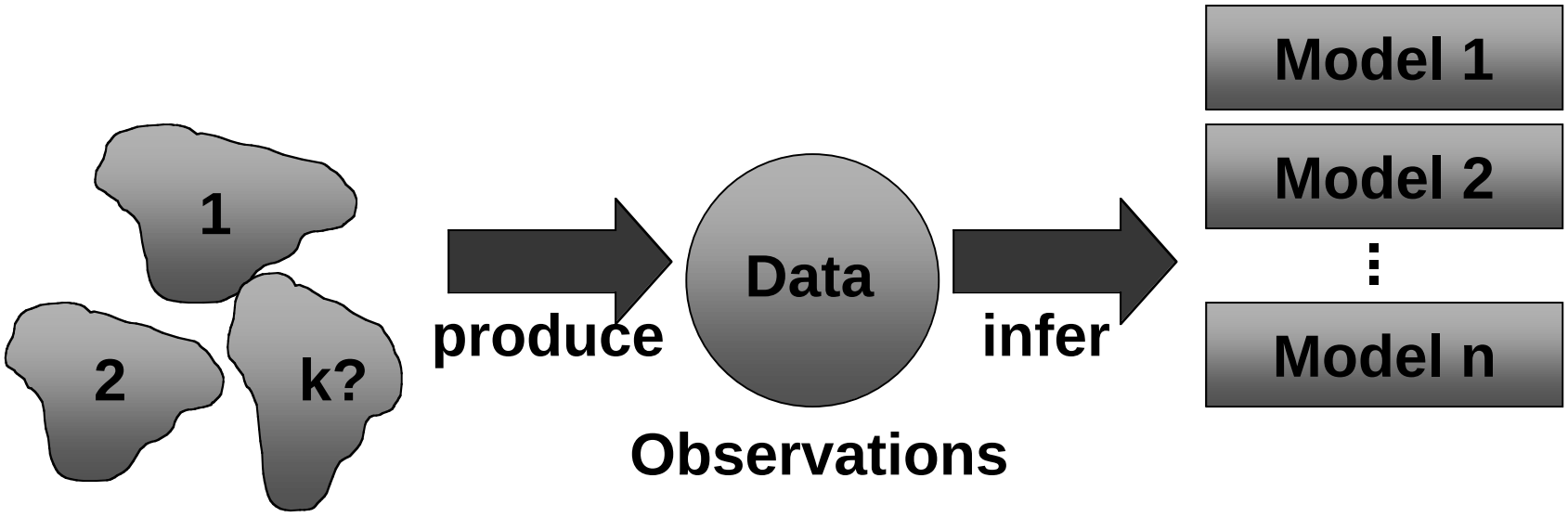
Non-Homogenous Markov chains:
Transition probabilities vary

- Either with time (cf. Simulated annealing)
- Based on observations:
  Cannot go to state X unless seen a total of Y observations

Class HMMs

- Create pairs of observations :
  - A nucleotid (A, C, G, T)
  - Its annotation (Exon, Intron)
- Adjustment of training required
  - Conditional Maximal Likelihood: Maximize the likelihood of the *correct* annotation

HMMs: Clusters and Mixtures



HMMs: Clusters and Mixtures

- Model-based clustering *(like k-means, only with HMMs instead of centroids)*
- Mixtures: Represent whole dataset as a combination of a collection of models

GHMM

- GNU (pending) Hidden Markov Model library
- C-library with C++ wrapper
- Weighted graph as main data structure
- GHMM is ***not*** a profile HMM library
- *Portable, autoconf, automake*

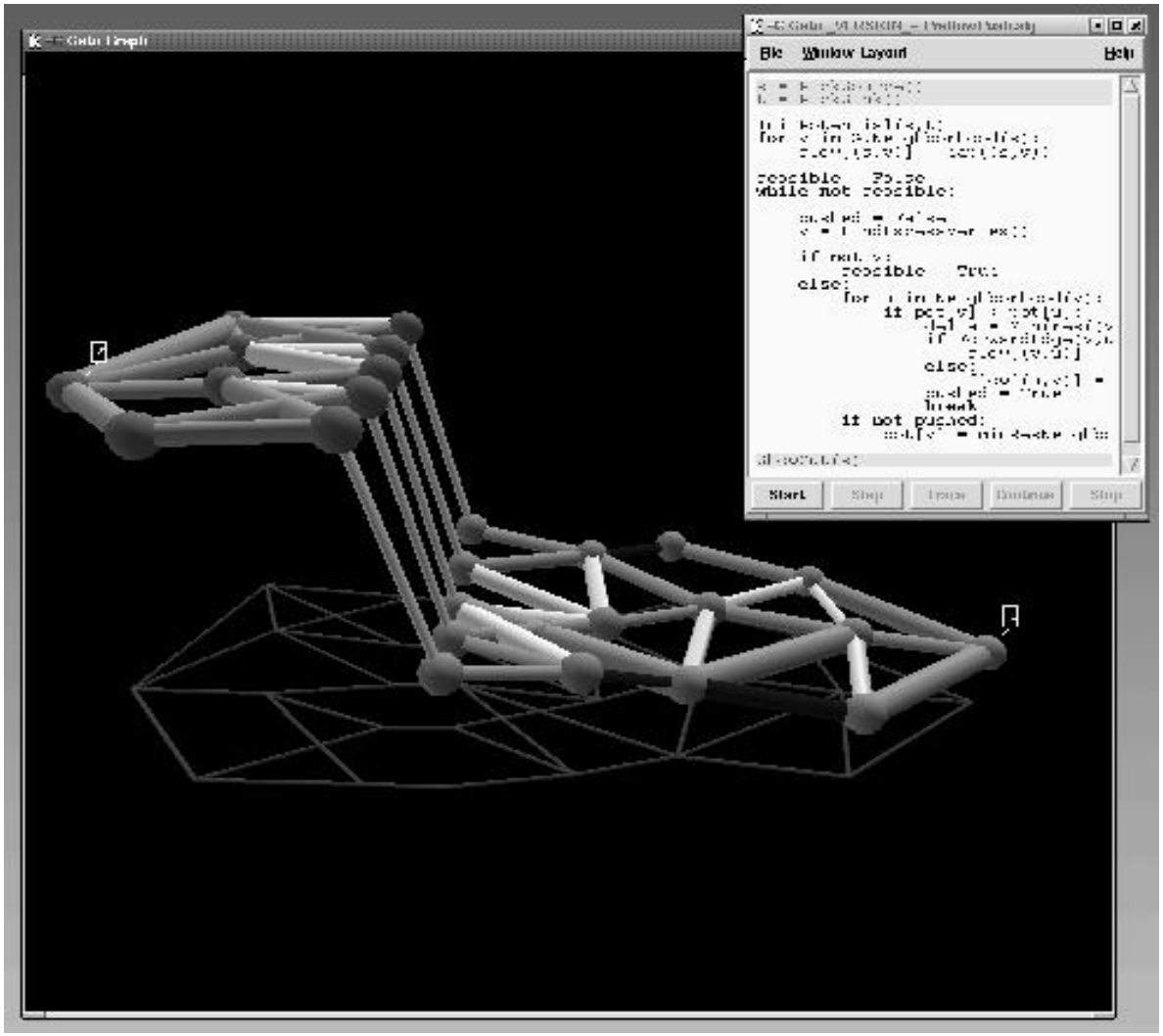
GHMM Model types

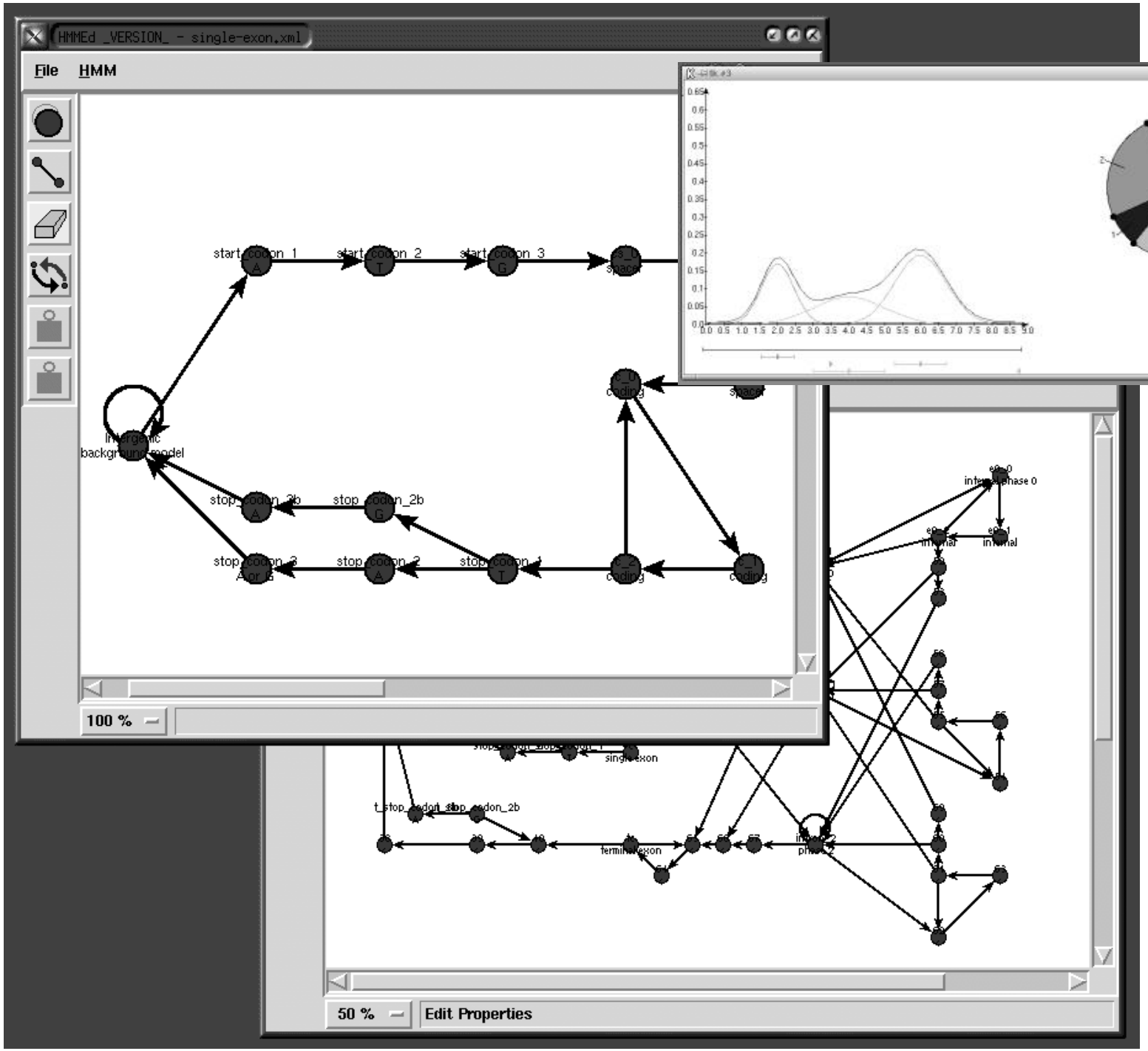
- Emissions:
  - Discrete
  - Continous: PDF, Mixtures of PDFs
  - *Multivariates*
  - *Discrete & Continous*

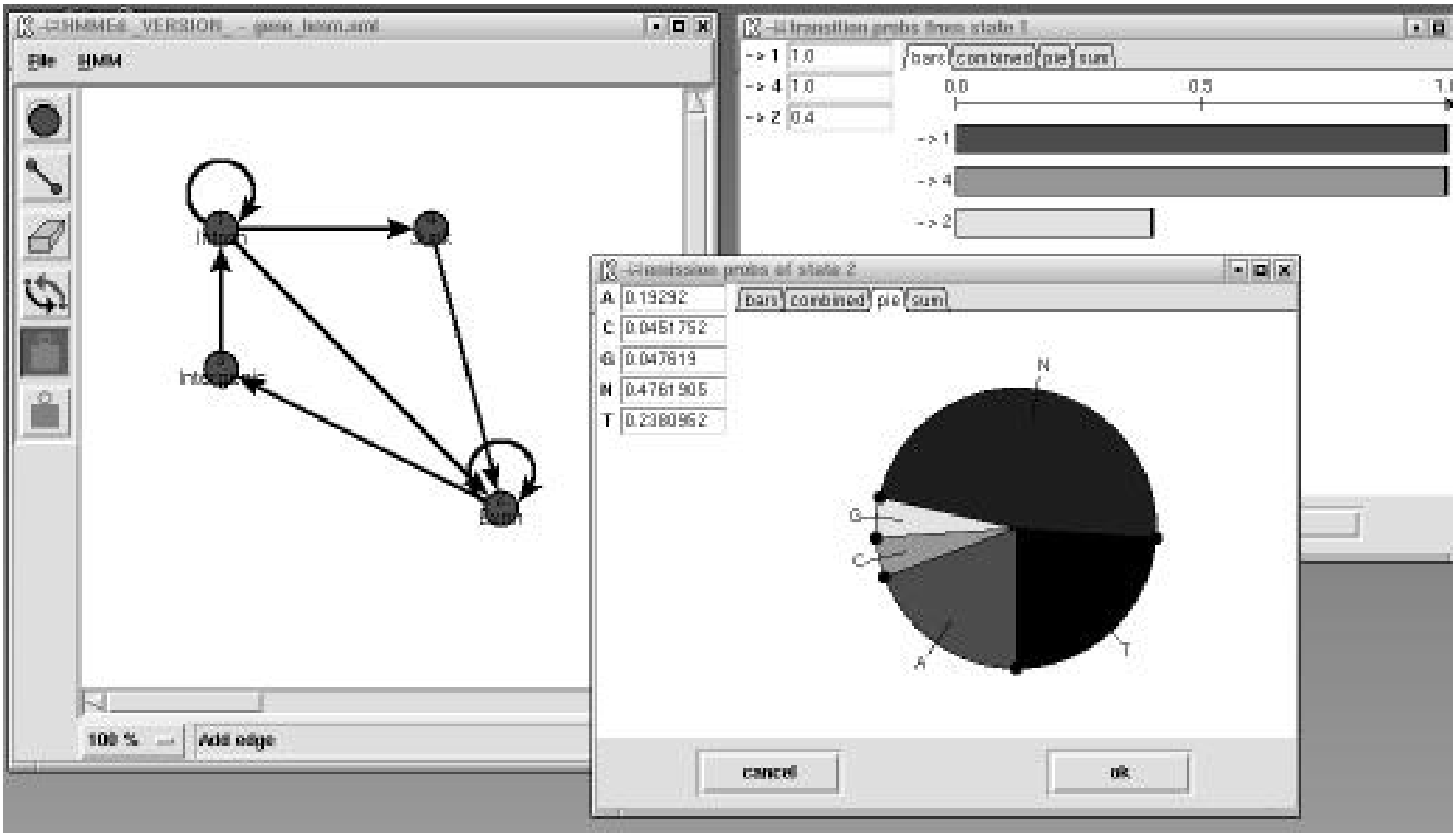
GHMM Functionality

- The basics:
  - Training: Baum-Welch (+noise injection)
  - Compute likelihoods etc.
  - Generate sequences
- Fun stuff
  - Simulated Annealing
  - Clustering
  - Mixtures

Gato: Software basis for HMMEd




HMMEd _VERSION_ - single-exon.xml
File
HMM
start_codon_1
start_codon_2
start_codon_3
intergenic background model
stop_codon_2b
stop_codon_2b
stop_codon_3
stop_codon_2
stop_codon_1
100 %
single_exon
terminal_exon
50 %
Edit Properties


File HMM
-> 1 1.0
-> 4 1.0
-> 2 0.4
bars
combined
pie
sum
0.0
0.5
-> 1
-> 4
-> 2
A 0.19292
C 0.0451792
G 0.047819
N 0.4761905
T 0.2380952
bars
combined
pie
sum
N
G
C
A
T
100 %
Add edge
cancel
ok

Hierarchical Models

- Gene Finding:
  - Codons
  - Exons, Introns, Promoters …
  - Transcription factor binding sites
- Finding remote homologs:
  - Models for individual domains

XML based file format

- Variation on GraphML
  www.graphdrawing.org/graphml/
- Quite rich format:
  - Hypergraphs
  - Hierarchical graphs (I.e. Graphs as Nodes)
- Human readable:
  - Default values for instance variables

In the pipeline …

- Finish ClassHMMs
- Finalize Python Bindings
- Documentation and more Documentation
- Finish HMMEd
- Finalize XML File Format
- Create a 1.0 release
- More Work on Training algorithms … (1.1)
- Length Modelling (1.2)

Acknowledgements

- German Landesbausparkassen for data and funding (unknowingly) original project
- Natascha Gayeva, Disa Thorarinsdottir, Achim Gädke, Peter Pipenbacher (ZAIK)
- Science Factory (partial funding of HMMEd)
- GHMM & HMMEd users: LANL, Northrop Grumman, Variagenics, MIT, U Köln, Science Factory, …

Thanks!

http://algorithmics.molgen.mpg.de