Estimating the size of the transcriptome

Statistical Modeling and Machine Learning in Computational Systems Biology June 22-26, 2009, Tampere, Finland

Ole Winther

Technical University of Denmark (DTU) & University of Copenhagen (KU)

June 24, 2009

All lectures

1. Introduction to graphical models and Bayesian networks
2. Estimating the size of the transcriptome
3. Using biological prior information in motif discovery
4. Learning linear Bayes networks with sparse Bayesian models

Common theme:

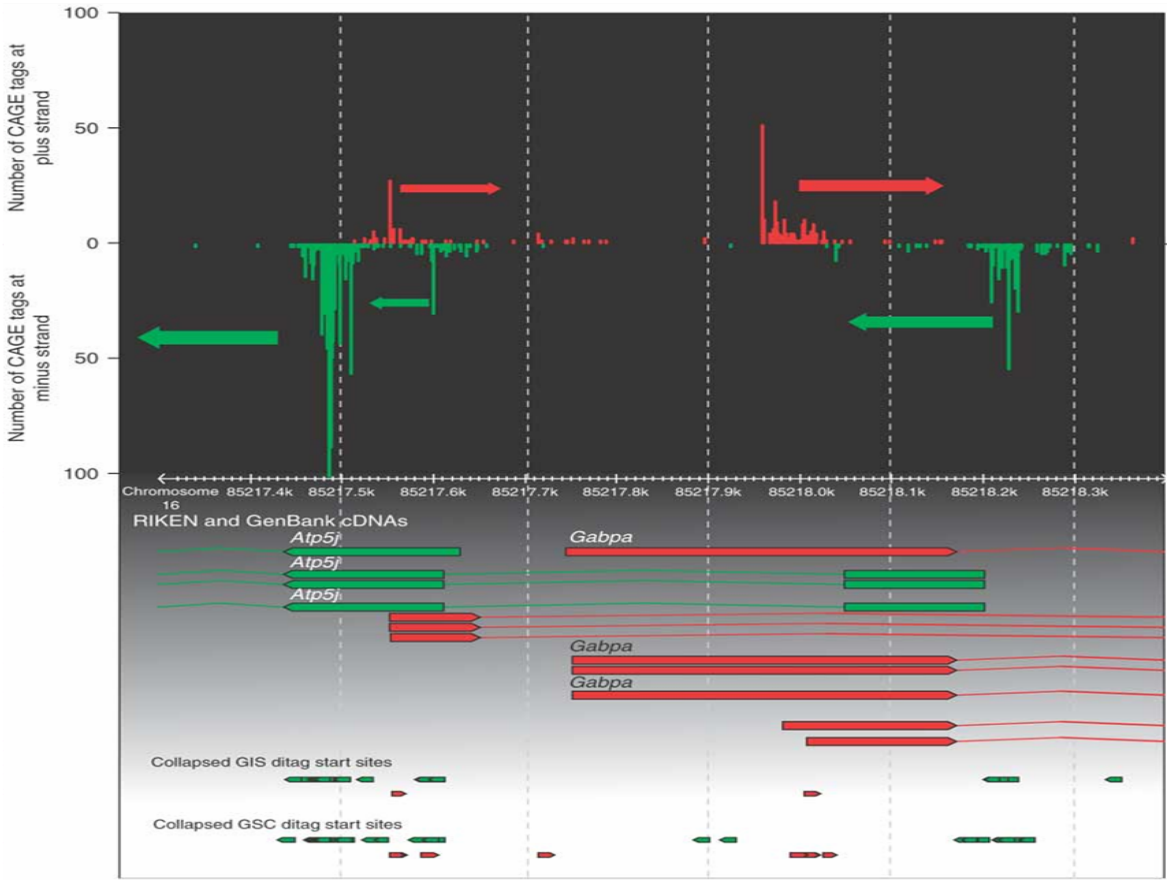
- Complex Bayesian model building possible and advantageous
- Model checking – prediction, marginal- and test-likelihood

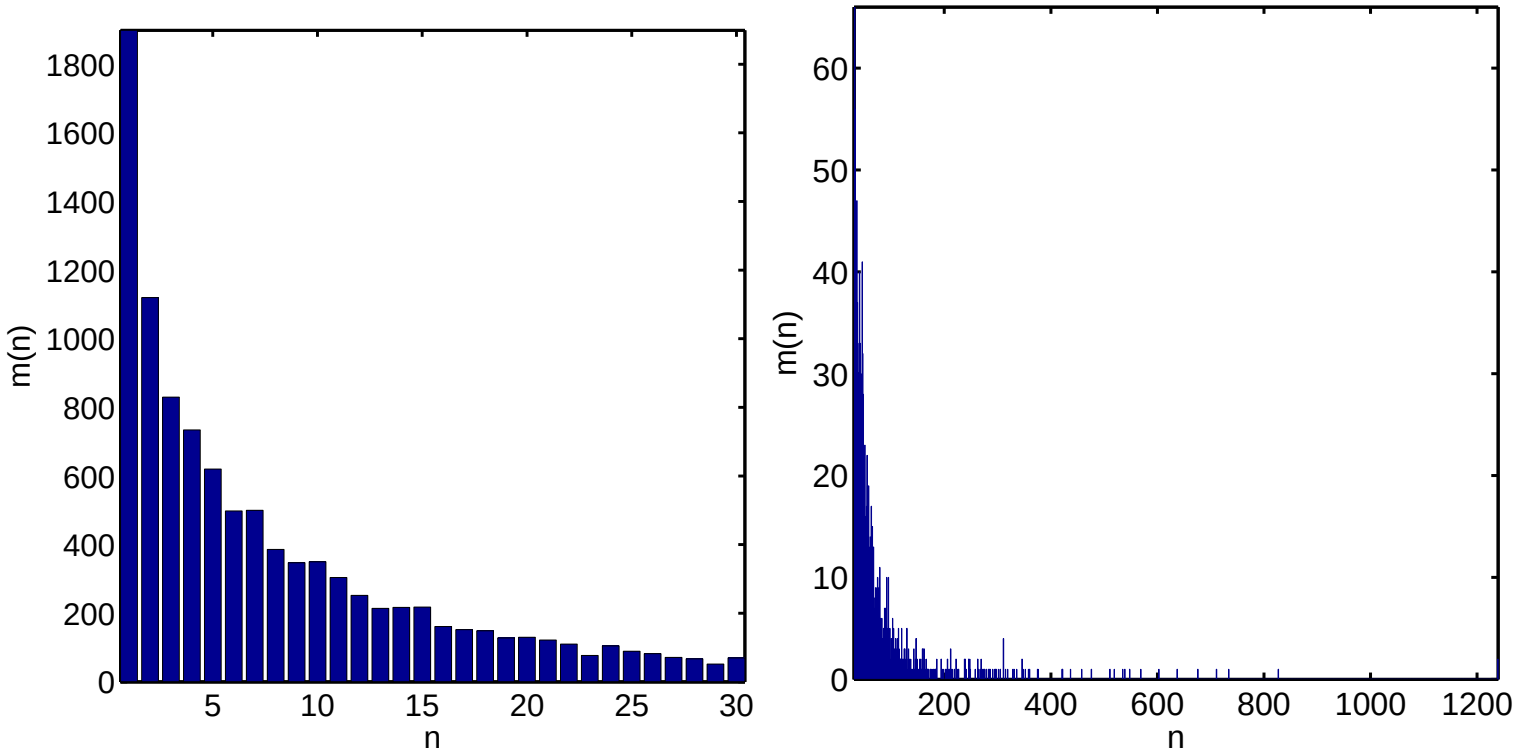
Lecture 2

- How many species? or in a genomic content
- Estimating the size of the transcriptome
- High throughput sequencing
- Non-parametric Bayesian model
- The model is always wrong (and Bayes can't tell)
- Model checking with cross-validation.

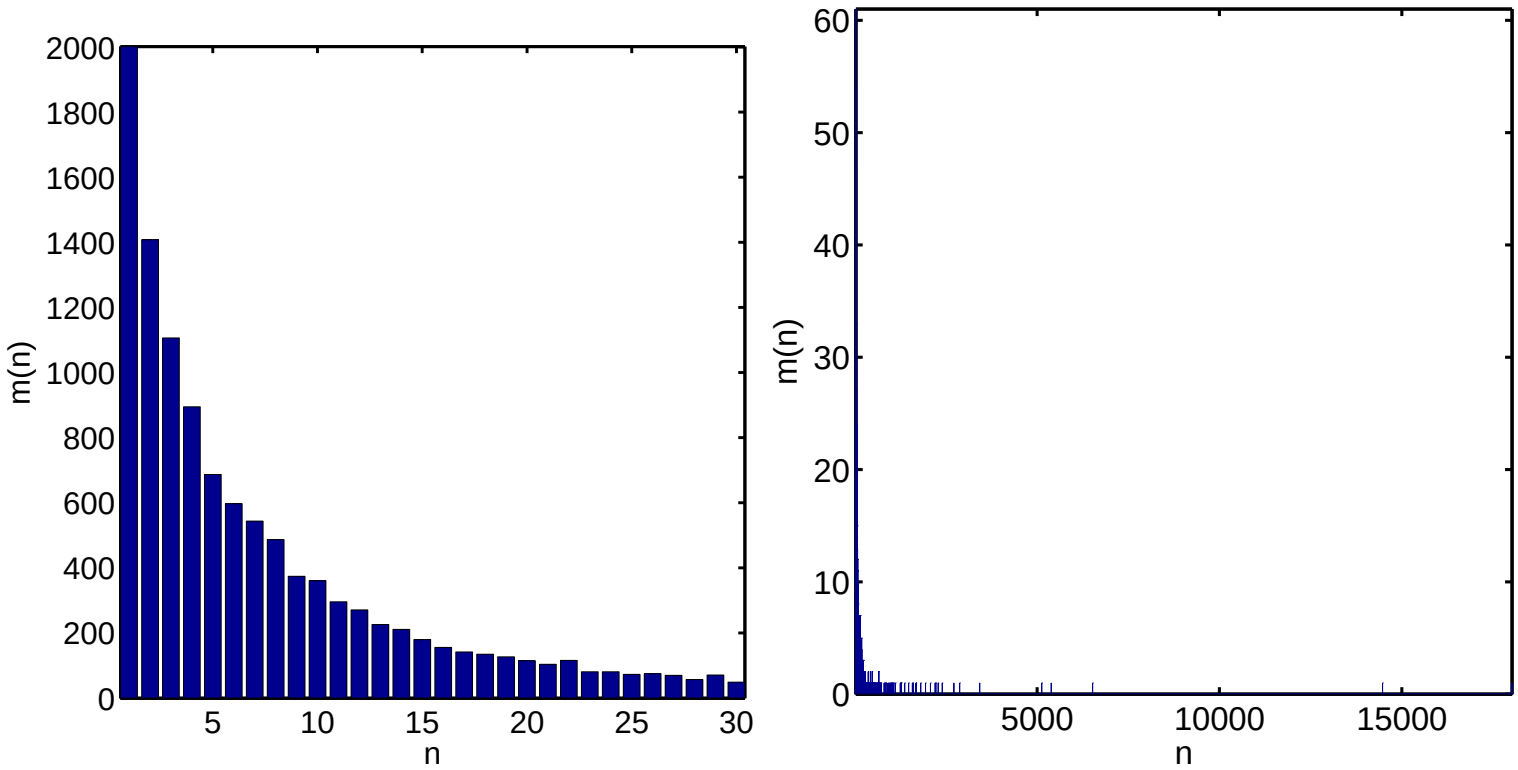
How many species?

Carninci et. al. Nat. Gen. 2006





Cerebellum library - frequency of frequency plot.



Embryo library

Solexa and Solid sequencing offer $10^6 - 10^8$ reads of length 20-60 nt at a price comparable to a micro-array.

- CAGE = cap analysis gene expression. 5' end of mRNAs. Pinpoints transcription start sites (TSSs). High throughput.
- EST = expressed sequence tag. Relatively low throughput. Used for gene identification.
- SAG = Serial analysis of gene expression. Medium throughput and longer reads.
- RNA seq or whole transcriptome shotgun sequencing. High throughput and longer reads.

- Non-parametric statistics from Wikipedia:

  "distribution free methods which do not rely on assumptions that the data are drawn from a given probability distribution. As such it is the opposite of parametric statistics. It includes non-parametric statistical models, inference and statistical tests. . . ."
- Non-parametric models have parameters that are learned in the same way as in parametric statistics.
- Lijoi, Mena and Prünster, in a series of papers, recently applied the Poisson-Dirichlet process (chinese restaturant) in the context of genomic tag data.

Setting up the problem

- Library of n tags (reads)
- A sequence of genomic coordinates $(c_1, c_2, \ldots, c_n)$.
- Contains k unique TSSs with counts $\mathbf{n} = (n_1, \ldots, n_k)$, $n = \sum_{j=1}^{k} n_j$.
- Label the tags in order of their arrival such that $c_i \in \{1, \ldots, k\}$.
- The $n+1$th tag may either be one of the k previously seen TSSs or a new one. to belong to one: $c_{n+1} \in \{1, \ldots, k+1\}$.

Observing new species given counts $\mathbf{n} = n_1, \ldots, n_k$ in k bins:

$$p(c_{n+1} = k+1|\mathbf{n}, \sigma, \theta) = \frac{\theta + k\sigma}{n+\theta} \quad \text{with} \quad \sum_{i=1}^{k} n_i = n$$

Re-observing j:

$$P(c_{n+1} = j|\mathbf{n}, \sigma, \theta) = \frac{n_j - \sigma}{n+\theta}$$

Exchangeability – invariant to re-ordering

$$E, E, M, T, T: \quad p_1 = \frac{\theta}{\theta}\frac{1-\sigma}{1+\theta}\frac{\theta+\sigma}{2+\theta}\frac{\theta+2\sigma}{3+\theta}\frac{1-\sigma}{4+\theta}$$

$$M, E, T, T, E: \quad p_2 = \frac{\theta}{\theta}\frac{\theta+\sigma}{1+\theta}\frac{\theta+2\sigma}{2+\theta}\frac{1-\sigma}{3+\theta}\frac{1-\sigma}{4+\theta} = \ldots = p1$$

- Likelihood function, e.g. E, E, M, T, T

$$p(\mathbf{n}|\sigma,\theta) = \frac{\theta}{\theta}\frac{1-\sigma}{1+\theta}\frac{\theta+\sigma}{2+\theta}\frac{\theta+2\sigma}{3+\theta}\frac{1-\sigma}{4+\theta}$$

$$= \frac{1}{\prod_{i=1}^{n-1}(i+\theta)}\prod_{j=1}^{k-1}(\theta+j\sigma)\prod_{i'=1}^{k}\prod_{j'=1}^{n_{i'}-1}(j'-\sigma)$$

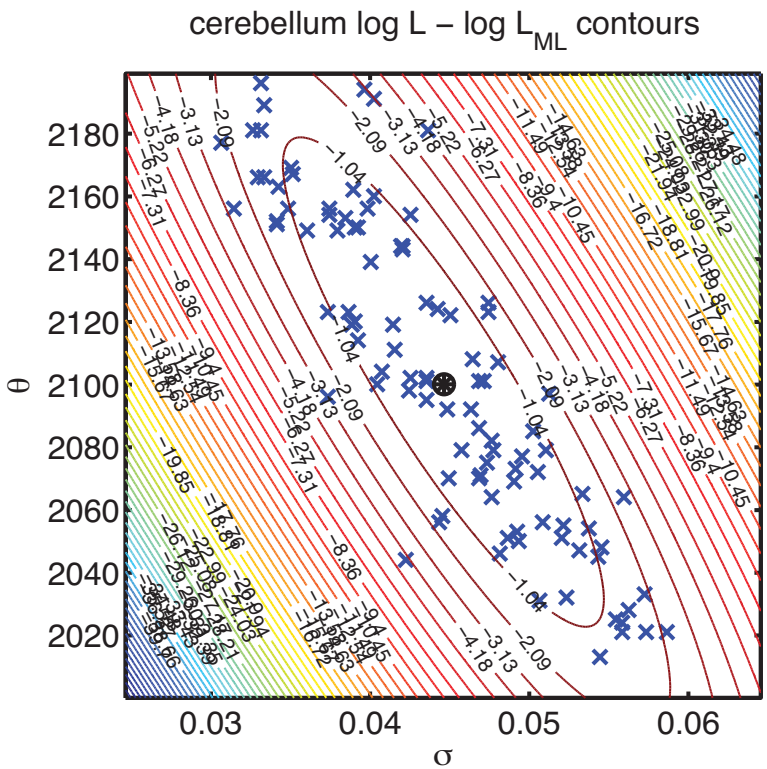
- Flat prior for $\sigma \in [-\theta/k, 1]$ and $\theta \geq 0$ pseudo-count parameter.
- Predictions – simulate new sequence $c_{n+1}, c_{n+2}, \ldots, c_{n+n'}$:

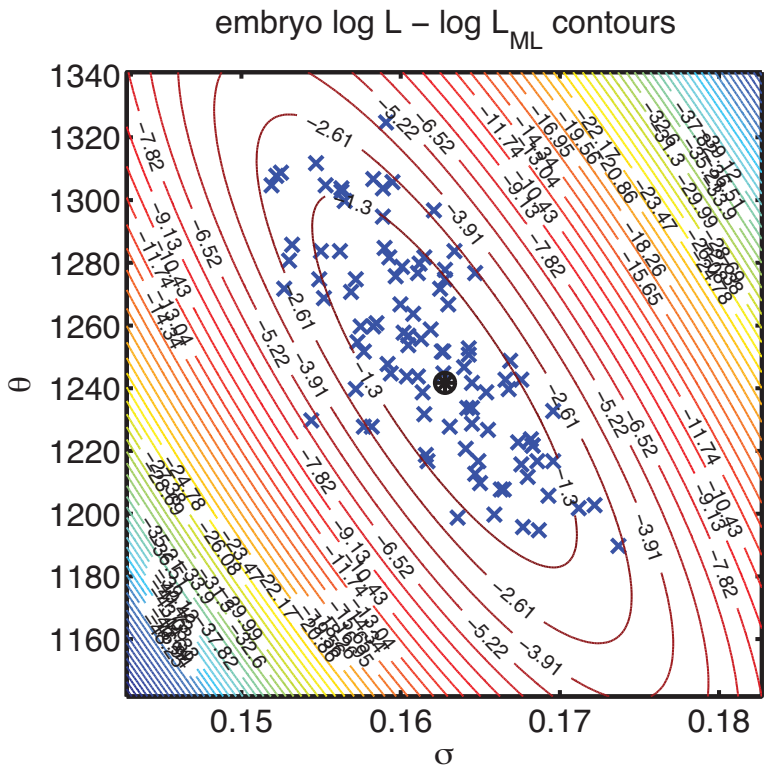
$$p(c_{n+1}, \ldots, c_{n+n'}|\mathbf{n}) = \int p(c_{n+1}, \ldots, c_{n+n'}|\mathbf{n}, \sigma, \theta)\, p(\sigma, \theta)\, d\sigma d\theta$$

with Gibbs sampling (σ, θ) and Yor-Pitman sampling for $c_{n+1}, \ldots$.

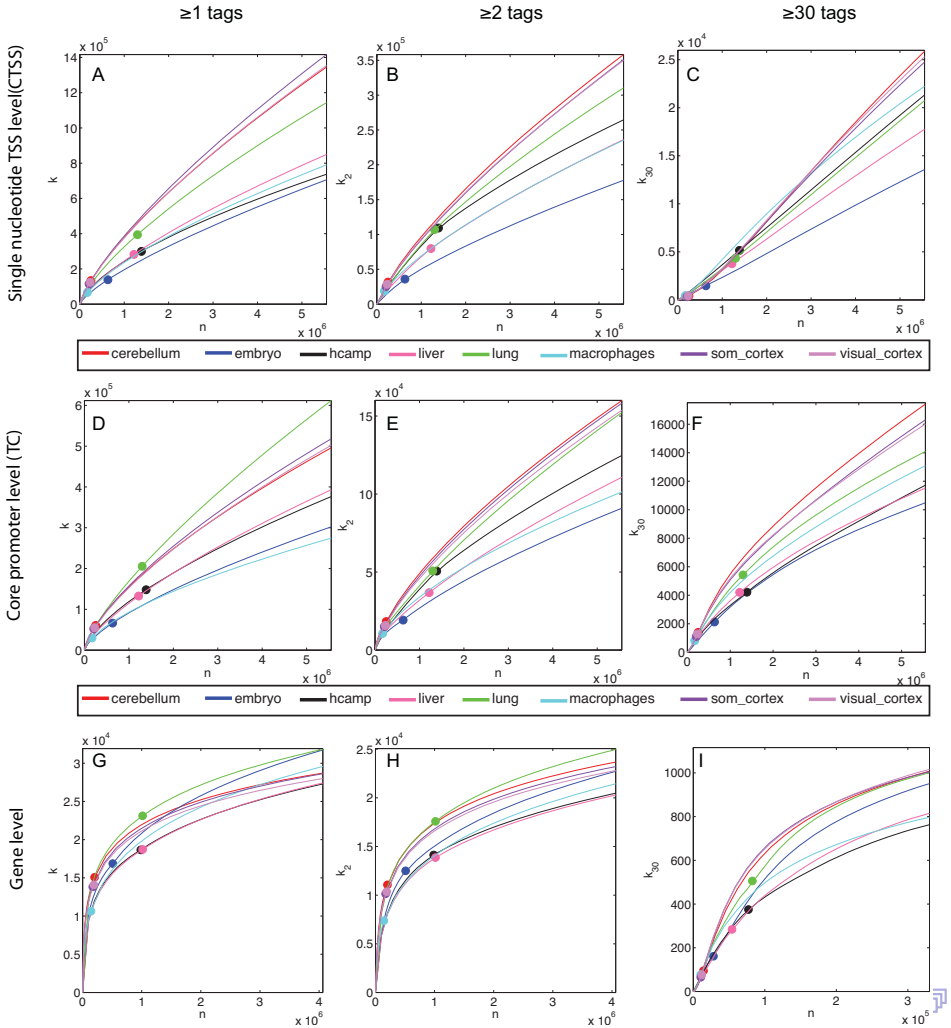
- 95% highest posterior density (HPD) intervals.

Averaging and maximum likelihood

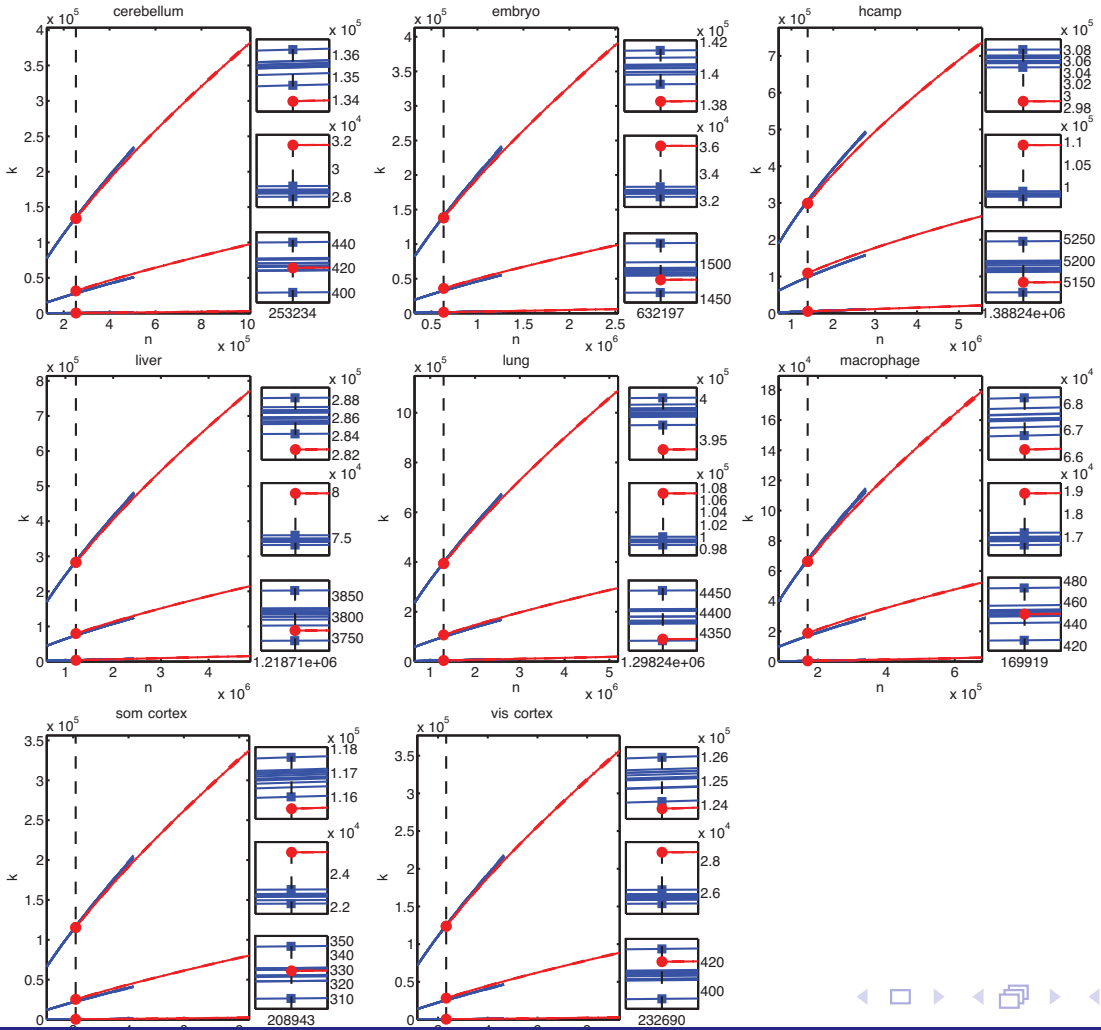




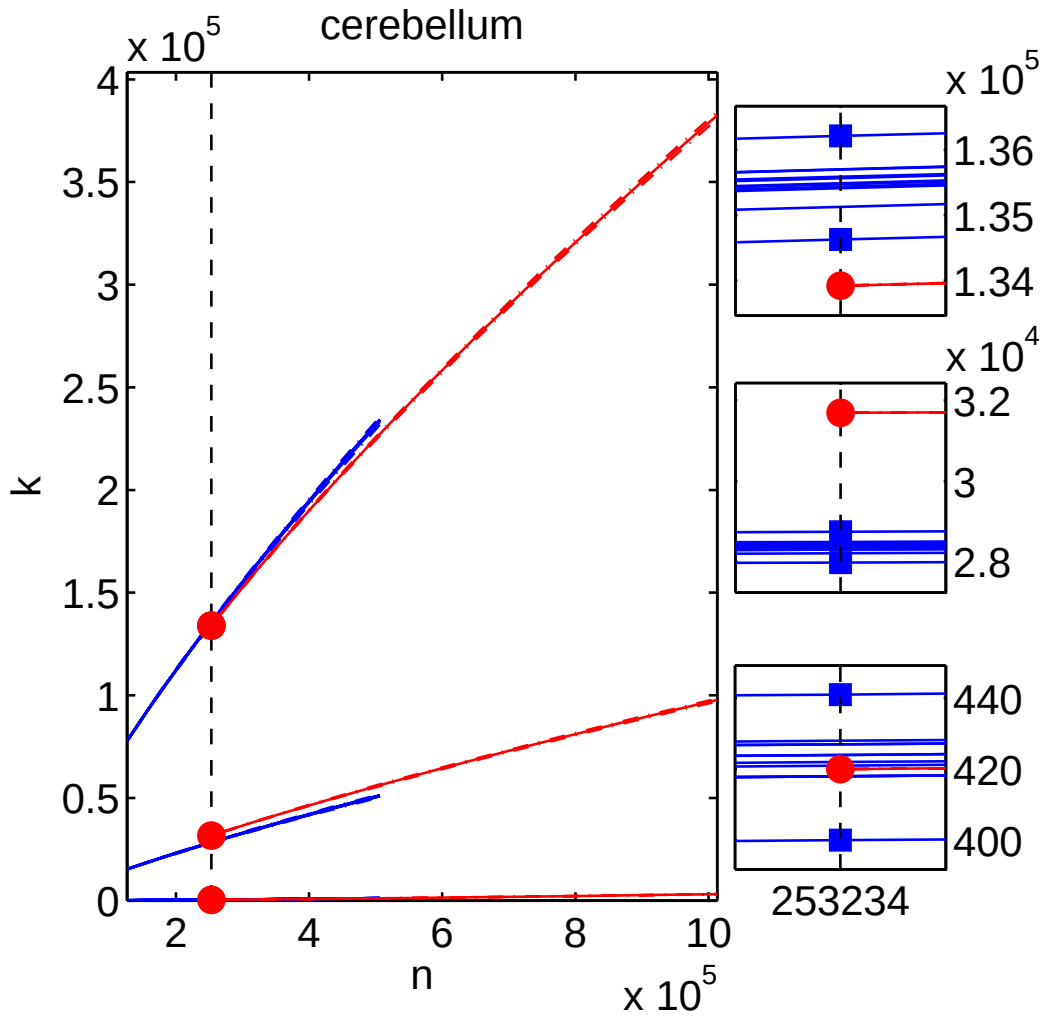
Predictions



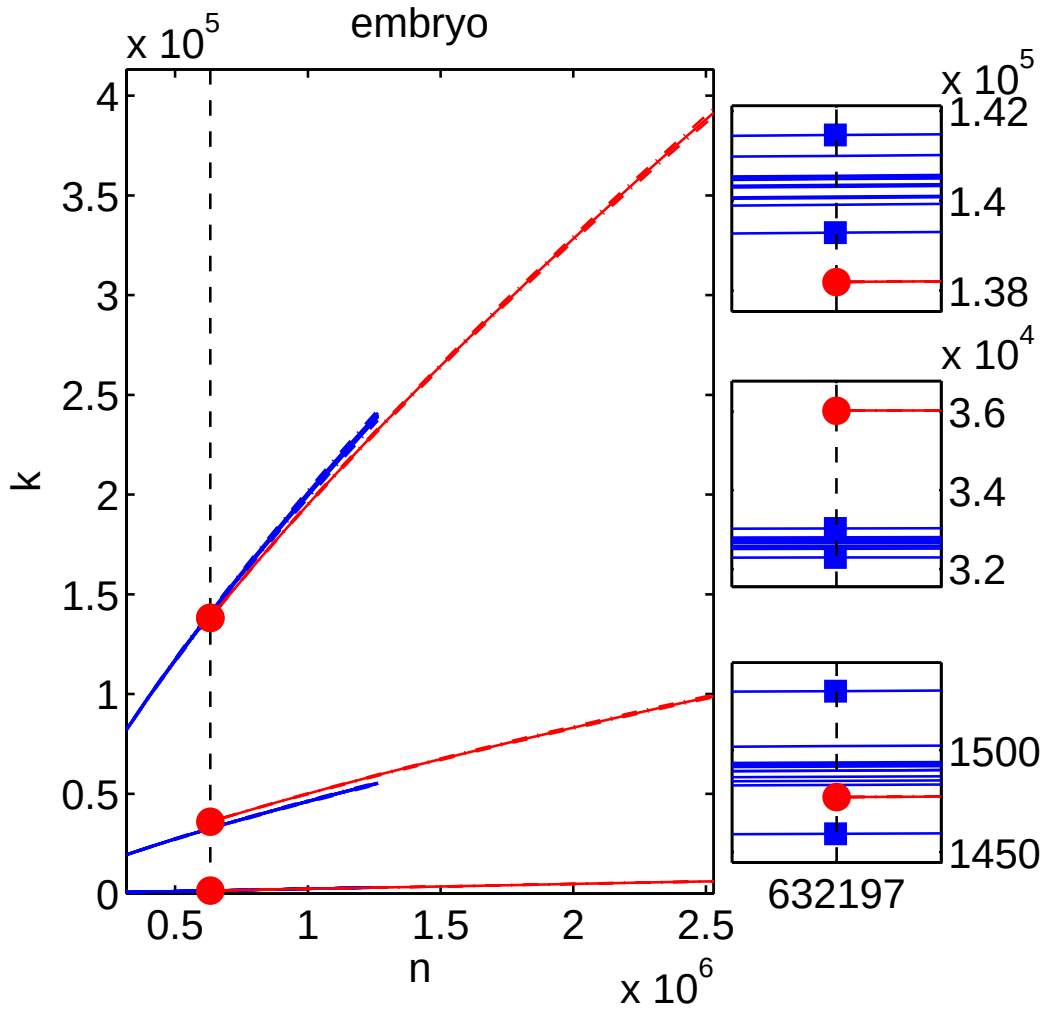
Cross validation – notice anything funny?



Cross validation – cerebellum



Cross validation – embryo



- We can actually predict more than just k.
- Each observed species in the library has a certain frequency.
- We estimate a probability for each species based upon model and observed frequency

$$\frac{n_j - \sigma}{n + \theta}$$

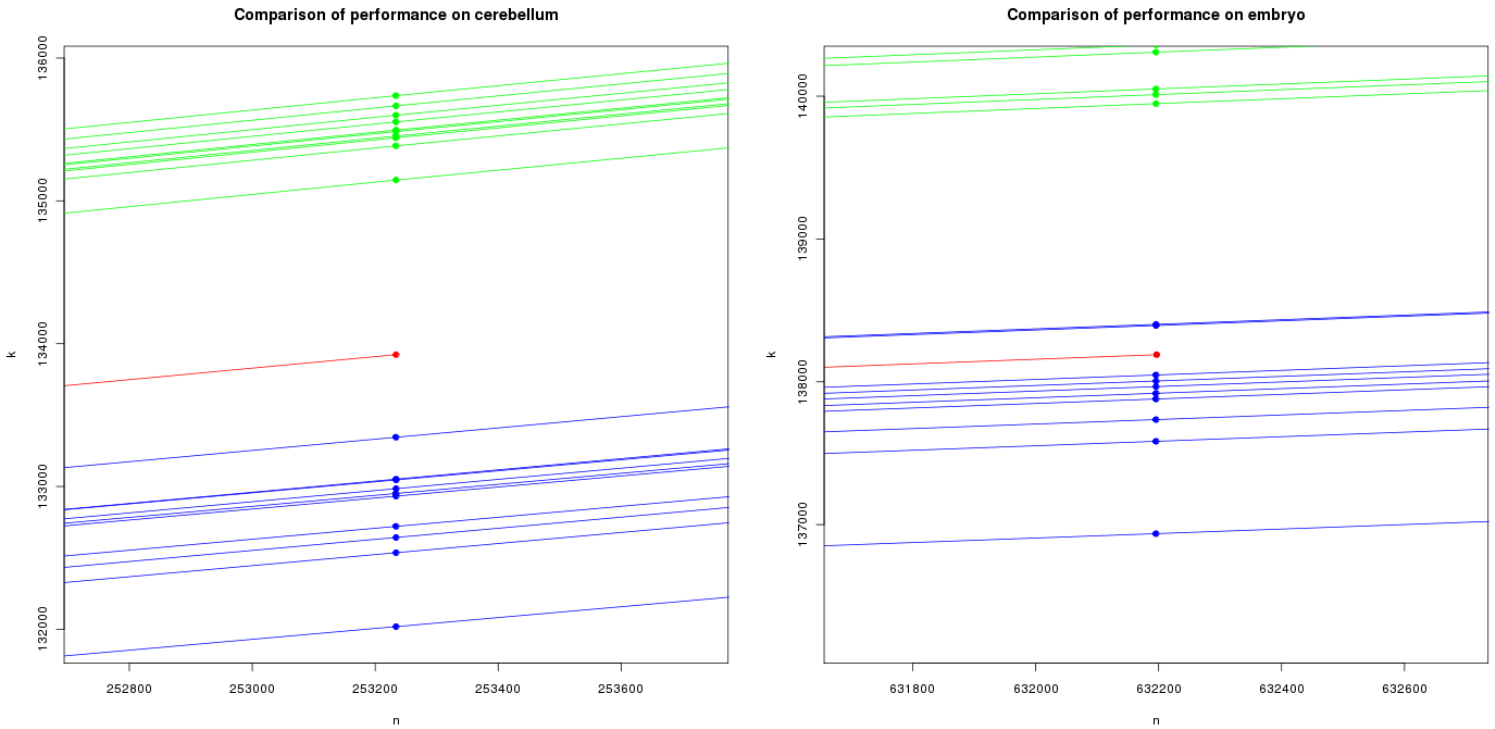
  Probabilities don't add up to one.
- The coverage says how we have already seen.
- Coverage (weight species by their observation probabilities):

$$\text{Coverage} = \sum_{j=1}^{k} \frac{n_j - \sigma}{n + \theta} = 1 - \frac{\theta + k\sigma}{n + \theta} .$$

Coverage predictions

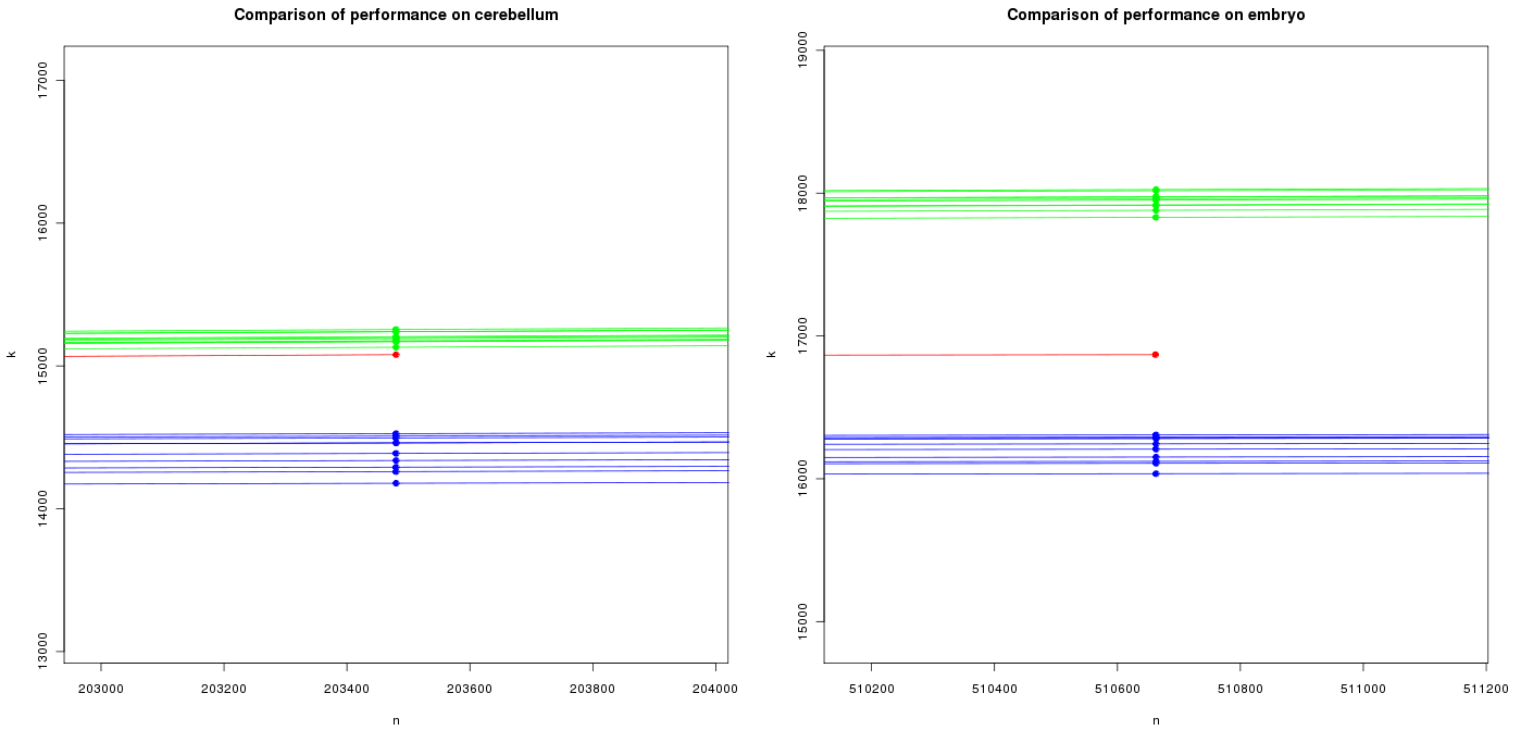
Library CTSS	sample stats n	k	coverage	parameters σ_{ML}	θ_{ML}	90.0 % coverage predictions coverage	n	k
cerebellum	253234	133923	0.592	0.728	9714	0.849	10000000	2068648
embryo	632197	138190	0.836	0.749	475	0.900	4576200	610448
hcamp	1388237	299038	0.859	0.639	6217	0.900	3633400	560111
liver	1218713	282555	0.831	0.722	2027	0.900	8032400	1109792
lung	1298243	393792	0.776	0.727	5441	0.872	10000000	1756587
macrophages	169919	66257	0.716	0.699	2620	0.900	5529000	787295
som cortex	208943	115480	0.565	0.748	7980	0.834	10000000	2206260
visual cortex	232690	123847	0.587	0.736	8431	0.845	10000000	2090184
						99.9 % coverage predictions		
gene	n	k	coverage	σ_{ML}	θ_{ML}	coverage	n	k
cerebellum	203481	15079	0.980	0.069	3090	0.990	436800	18312
embryo	510662	16869	0.989	0.254	1161	0.990	554400	17320
hcamp	983532	18665	0.995	0.210	1292		already reached	
liver	1015553	18711	0.995	0.223	1192		already reached	
lung	1015700	23121	0.994	0.142	2418		already reached	
macrophages	141236	10643	0.974	0.233	1223	0.990	495400	16020
som cortex	170823	13813	0.977	0.093	2720	0.990	434000	17675
visual cortex	189515	14080	0.979	0.086	2744	0.990	420000	17302

Parametric approach



Comparison TSSs: True, non-parametric and parametric Zhu et. al., PLoS ONE, 2008.

Parametric approach



Comparison genes: True, non-parametric and parametric.

Summary

- Non-parametric Poisson Dirichlet (PD) process for complex data.
- The model is always wrong! (as revealed in model checking with sufficient data).
- We need more complex model. With so much data only underfitting is a problem. Go beyond PD.
- Joint w Albin Sandelin, Eivind Valen and Anders Krogh, KU building upon Lijoi, Mena and Prünster.