The Abdus Salam International Centre for Theoretical Physics

2361-16

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

MACHINE LEARNING IN SYSTEMS BIOLOGY: Estimating the Size of the Transcriptome

Ole WINTHER

Technical University of Denmark DTU and University of Copenhagen KU
Denmark

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

Estimating the size of the transcriptome

Ole Winther

The Bioinformatics Centre/BRIC, University of Copenhagen (KU) and
Technical University of Denmark (DTU)

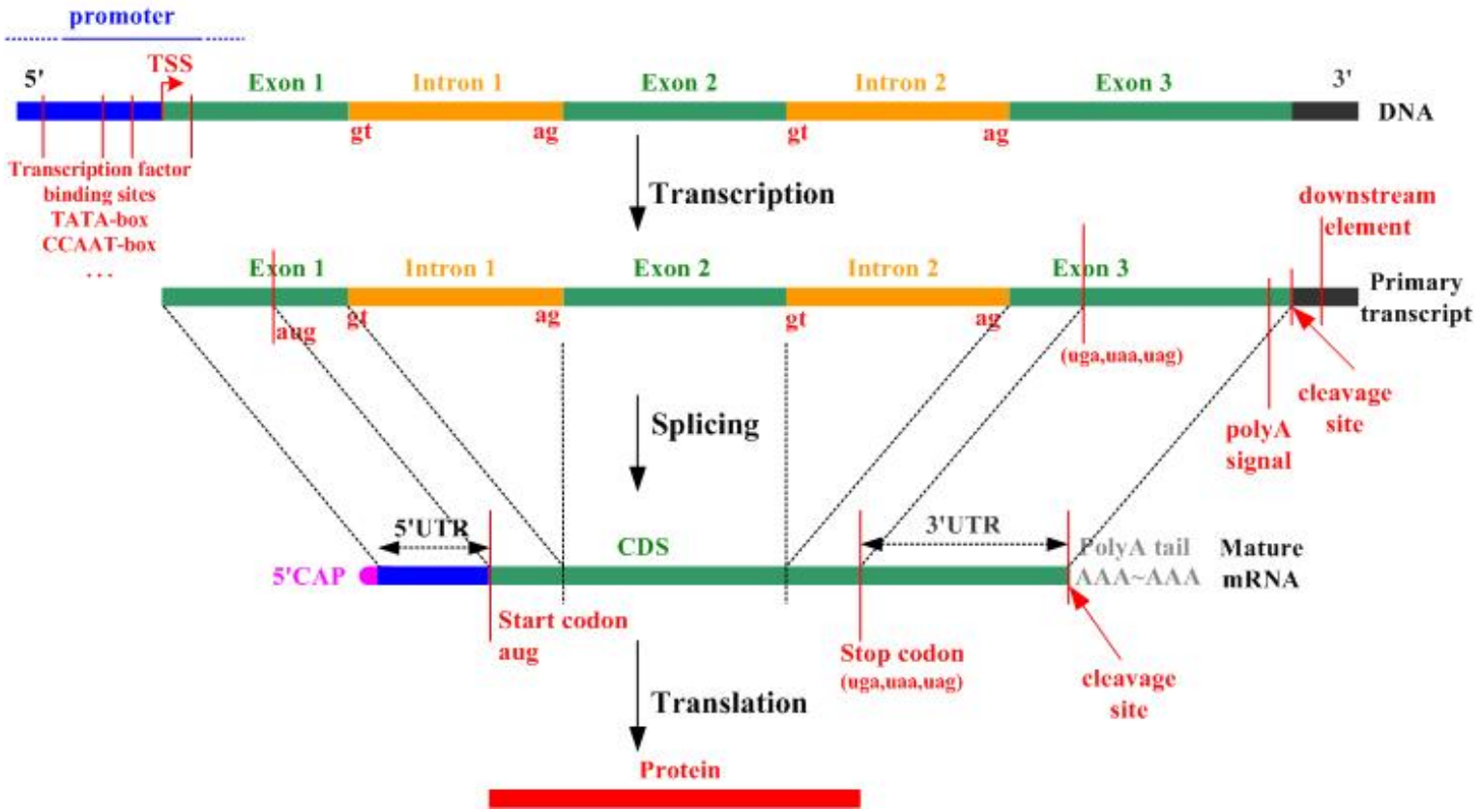
August 22, 2012

How many species?

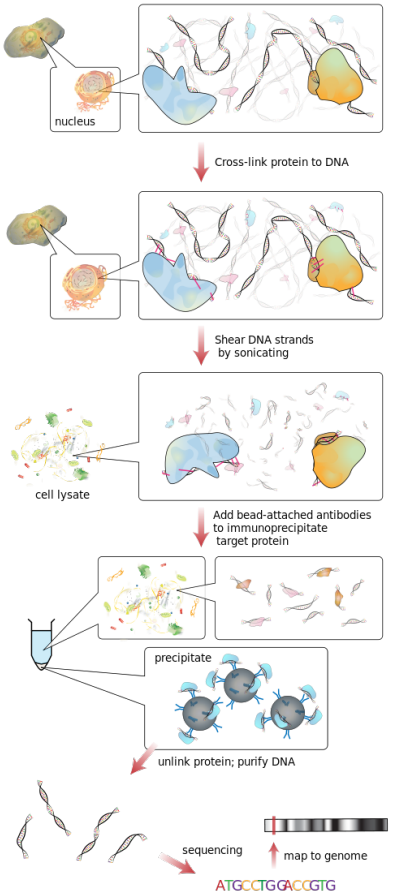
Solexa and Solid sequencing offer $10^6 - 10^8$ reads of length 20-60 nt at a price comparable to a micro-array.

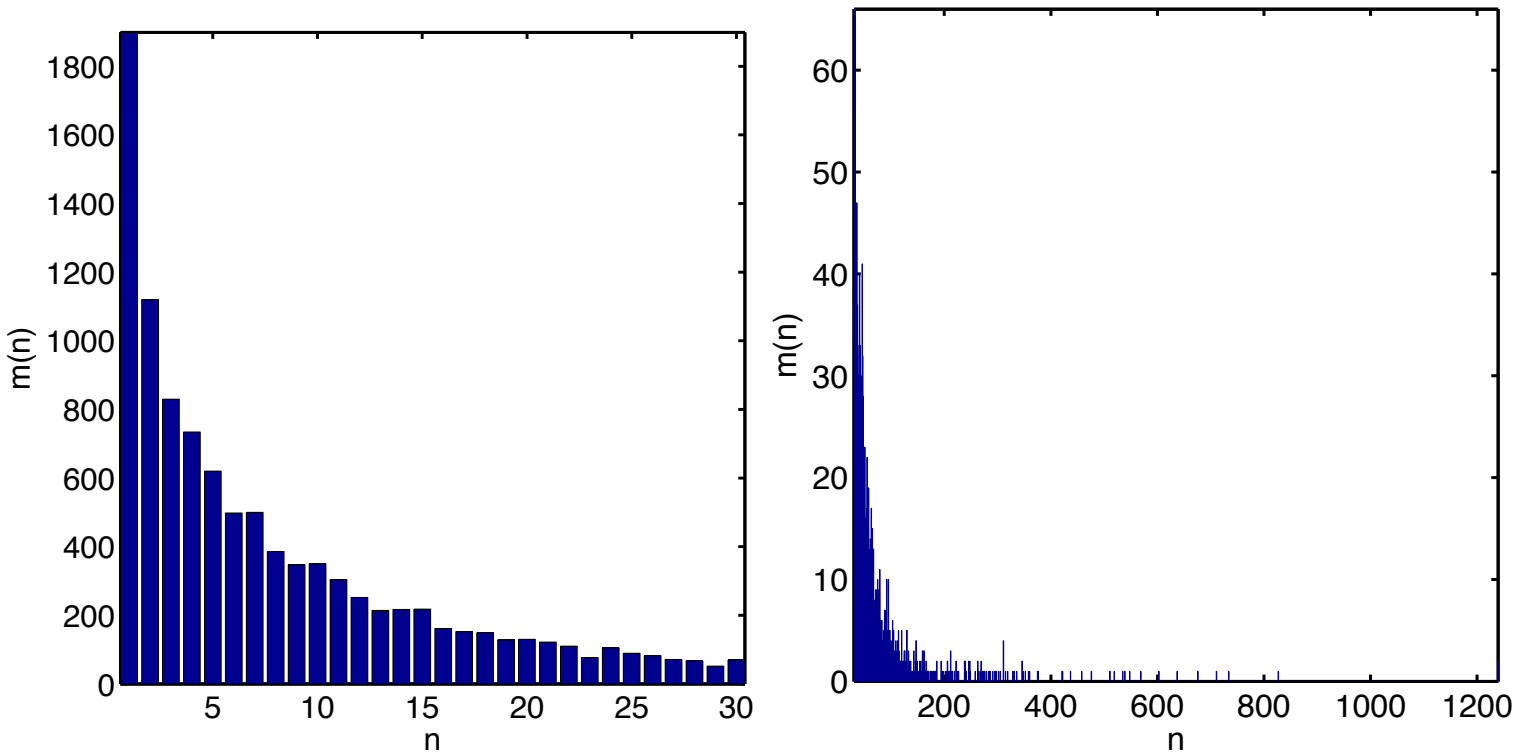
- CAGE = cap analysis gene expression. 5' end of mRNAs. Pinpoints transcription start sites (TSSs). High throughput.
- EST = expressed sequence tag. Relatively low throughput. Used for gene identification.
- SAGE = Serial analysis of gene expression. Medium throughput and longer reads.
- ChIP/DNA/RNA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

High-throughput sequencing technologies



- ChIP
- Chromatin ImmunoPrecipitation
- identifies protein-DNA interactions





Cerebellum library - frequency of frequency plot.

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:

$$c_{n+1} \in \{1, \ldots, k, k+1\} .$$

- Applied to this problem by Lijoi, Mena, Prünster, et. al.
- 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

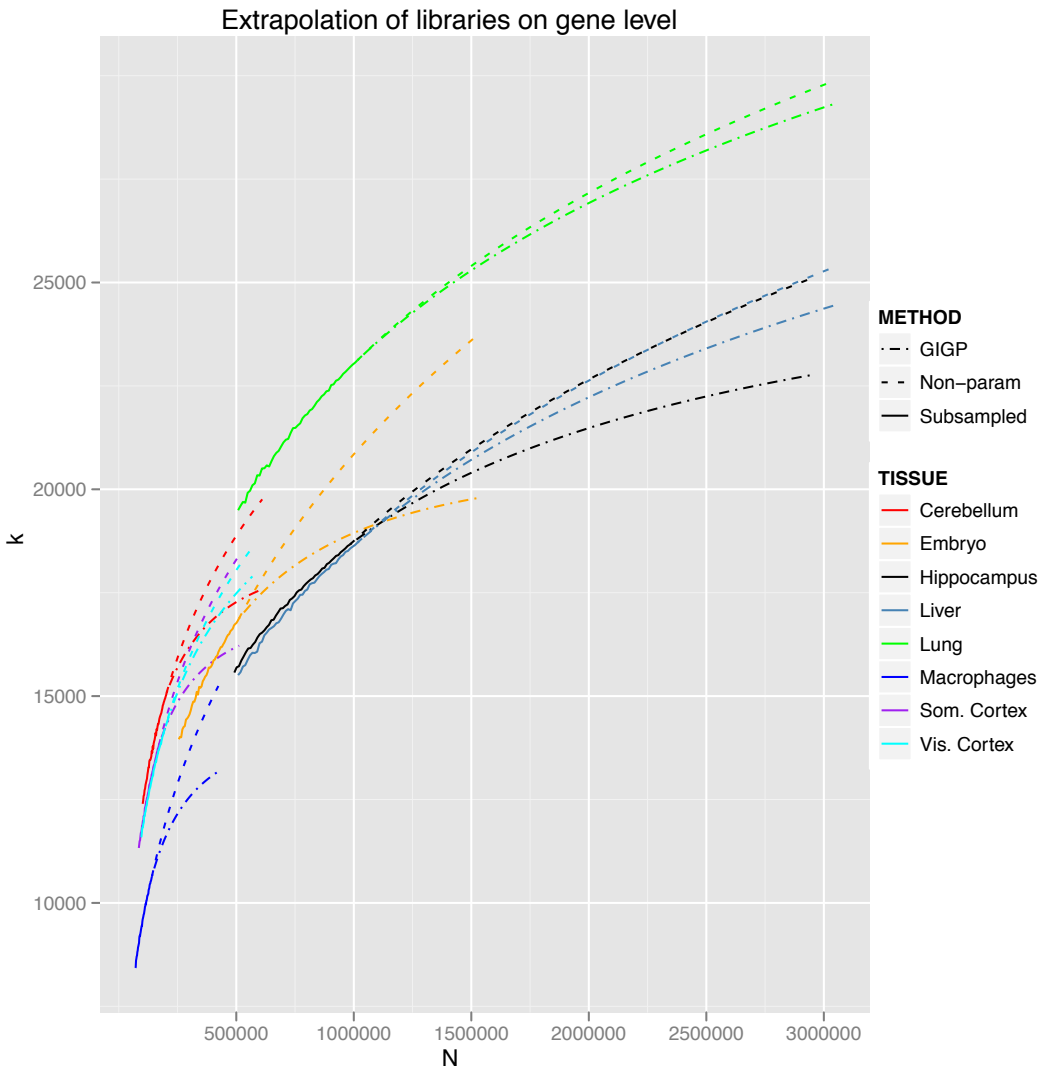
$$p(\mathbf{n}|\sigma,\theta) = \frac{\theta}{\theta}\frac{1-\sigma}{1+\theta}\frac{\theta+\sigma}{2+\theta}\frac{\theta+2\sigma}{3+\theta}\frac{1-\sigma}{4+\theta}$$

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

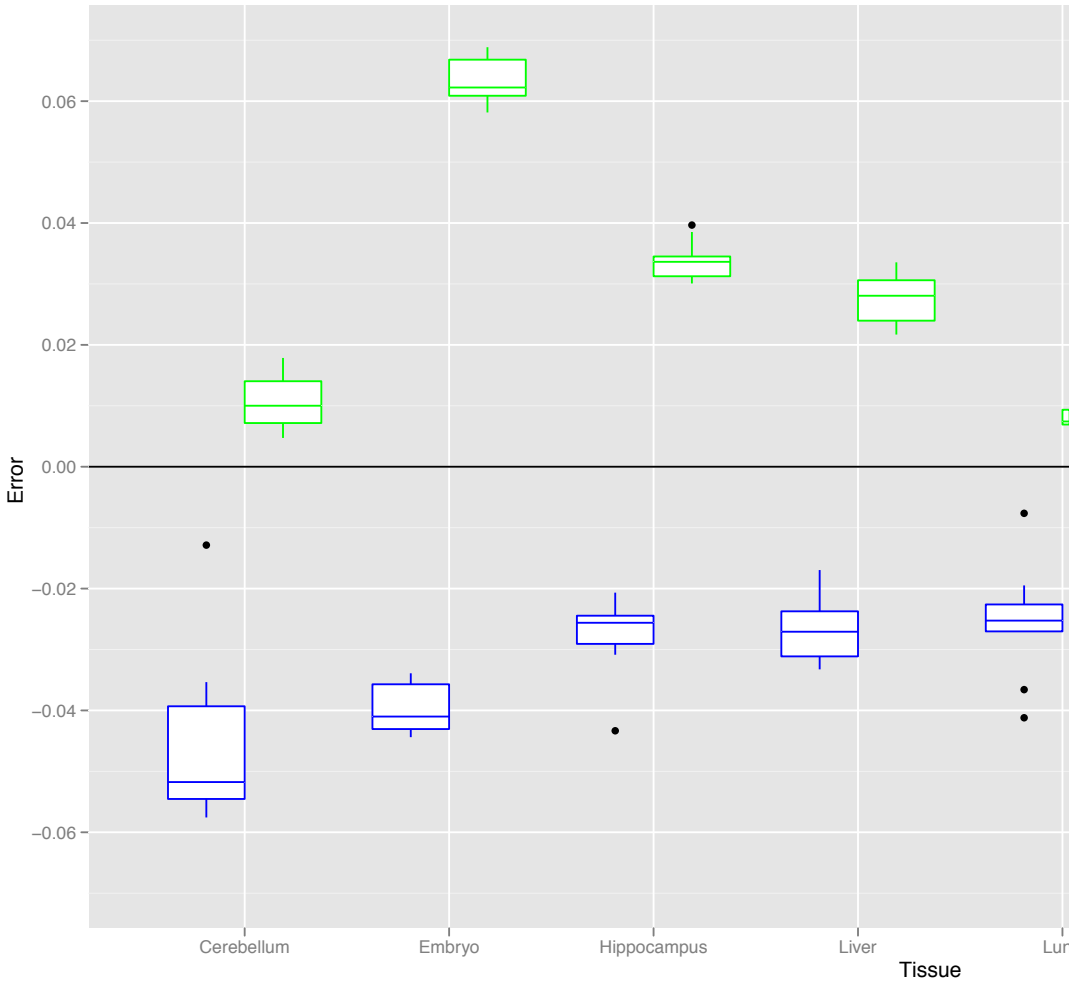
- Maximum likelihood (ML) inference or Gibbs sampling
- Predictions – simulate new sequence $c_{n+1}, c_{n+2}, \ldots, c_{n+n'}$ using the sampling formula iteratively:

$$p(c_{n+1}, \ldots, c_{n+n'}|\mathbf{n}, \sigma_{\mathrm{ML}}, \theta_{\mathrm{ML}})$$

Results


Extrapolation of libraries on gene level
k
N
25000
20000
15000
10000
500000
1000000
1500000
2000000
2500000
3000000
METHOD
GIGP
Non−param
Subsampled
TISSUE
Cerebellum
Embryo
Hippocampus
Liver
Lung
Macrophages
Som. Cortex
Vis. Cortex

Results



Cross-validated predictions from half to full size

- Model assigns a probability to each of the observed species, $j = 1, \ldots, k$:

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

- What is the probability to see something we have already seen?
- Coverage (weight species by their observation probabilities):

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

- Empirical predictions from 95%+ (genes) down to 60% (genomic positions).

Conclusion

- Experimental technologies develop fast, lot to learn!
- Species sampling models accurate but not completely accurate for real data.
- Link to code and data: `http://people.binf.ku.dk/albin/supplementary_data/tss_saturation/`