Stochastic motif finders

1. Stochastic motif finding using a greedy algorithm
2. Introduction to Monte Carlo Markov Chains
3. Metropolis-Algorithm
4. Gibbs sampler
5. Gibbs sampler application

Literature

This exposition is based on the following sources, which are all recommended reading:

1. Gibbs Sampler: Lawrence *et al*: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 1993
2. Markov Chain Monte Carlo: Andrieu *et al*. An Introduction to MCMC for Machine Learning. 2008

Objective and notation

Given t sequences $s_1, ..., s_t$ of length n, and an integer i, the goal is to find an l-mer in each of the sequences such that the similarity between the l-mers is maximized.

Let $(a_1, ..., a_t)$ be a list of l-mers contained in $s_1, ..., s_t$.
These form a $t \times l$ alignment matrix. Let $X(a) = (x_{ij})$ denote the corresponding $4 \times l$ profile, where x_{ij} denotes the frequency with which we observe nucleotide i at position j.
Usually, we add pseudo-counts to ensure that X does not contain zeros (Laplace correction).

Greedy profile search

For a given l-mer a, consider

$$P(a|X) = \prod_{i=1}^{l} x_{a_i}$$

the probability that a was generated by X.
l-mers similar to the consensus string of X have a high probability, dissimilars have low probabilities.

Greedy profile search (2)

Given a profile X, we can evaluate the probability of every l-mer a in a sequence s to find the X- most probable l-mer in a, defined as

$$a = \arg\max P(a|X)$$

.

Greedy profile search (3)

Greedy profile search algorithm.

- Given sequences $s_1, ..., s_t$ of length n, randomly select one l-mer a_i for each sequence s_i
- Construct an initial profile X.
- For each sequence s_i determine the X-most probable l-mer $a_i^{'}$.
- Set X equal to the profile obtained from $a_1^{'}, ..., a_t^{'}$ and repeat.

This naive approach starts with a random seed profile and then attempts to improve on it using a greedy strategy.

Markov Chain Monte Carlo

Describing a high-dimensional density samples $p(x)$ in a space $\mathcal{X}$ for computation by generating samples ${x^i}_{i=1}^N$.

The N samples can be used to describe an empirical point-mass function

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}(x)$$

This allows for calculating the integral over X with

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^i) \longrightarrow_{N\to\infty} \int_{\mathcal{X}} f(x)p(x)dx$$

.

The advantage of Monte Carlo integration over deterministic integration arises from the fact that the former positions the integration grid (samples) in regions of high probability.

How to sample

This can be used to find global optima in complex, high-dimensional settings.

To explore the state space $\mathcal{X}$ it was found advantageous to spend the sampling time in the important regions.

MCMC example

Stability

For any starting point, the chain will converge to the invariant distribution $p(x)$ if the transition matrix T is

1. *Irreducable* Positive probabailty to visit any state from any other state.
2. *Aperiodic* T contains no cycles.

How should this work in the continous case? Use a proposal distribution $q(x)$

Metropolis-Hastings Algorithm

1. Initialize x^0
2. For $i = 0$ to $N - 1$
   - Sample $u \sim \mathcal{U}_{[0,1]}$
   - Sample $x^{'} \sim q(x^{'}|x^i)$
   - Calculate acceptance ratio $A(x^i, x^{'} = \min\{1, \frac{p(x^{'})q(x^i|x^{'})}{p(x^i)q(x^{'}|x^i)}\}$
   - If $u < A$ then $x^{(i+1)} = x^{'}$ else $x^{i+1} = x^i$

Gibbs sampling

- Variety of the Metropolis algorithm with a constant acceptance rate A of 1
- In multivariate scenarios, only one parameter is changed in each step

Gibbs sampler

Greedy profile search can change all l-mers in every run but stay close to the profile, getting stuck in local maxima.

The Gibbs sampling strategy by Lawerence *et al* (1993) uses a random seed profile, changes only one l-mer per iteration over a larger search space.

Gibbs sampler (4)

- Choose an l-mer a_i at random in each sequence s_i
- Choose one of the sequences s_h at random
- Create a frequency matrix q from the remaining $t - 1$ sequences and the profile X.
- For each position i in s_h calculate the probability of $p_i = \prod_{j=0}^{w} q_{x^r_{i+j}}$ that the substring started at this position is generated from profile q_{ij}

- For each l-mer a in s_h, compute $w(a) = \frac{P(a|X)}{P(a|Q)}$
- Set $a_h = a$, for some $a \in s_h$ chosen randomly with probability $\frac{w(a)}{\sum_{a' \in s_h} w(a')}$.
- Repeat until convergence of X