Overcoming Undersampling in Generative Models for Biological Sequences

Marion Chauveau[1,3], Yaakov Kleeorin[2], Ivan Junier[1], Olivier Rivoire[3]

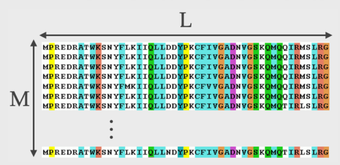
[1]TIMC, Univ. Grenoble Alpes, [2]University of Chicago, [3]Gulliver Laboratory, ESPCI Paris

Generative Models

Fundamental problems	Current approach
Undersampling: Inference relies on limited datasets [2]	L2-Regularization
Model Evaluation: How to score generative power ?	Comparison of empirical statistics with model statistics

Potts Model

Maximum entropy model trained to match the empirical **one and two-body frequencies** of amino acids [1]

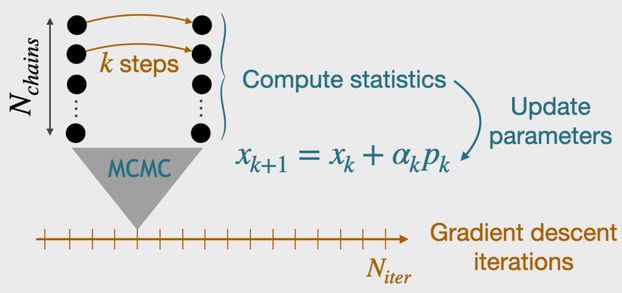


$$\underline{\text{Protein}}: \quad \boldsymbol{\sigma} = \{ \sigma_1, \ldots, \sigma_L \}$$

σ_i = a.a. at site i (q possibilities)

$$p(\{\sigma_i\}_{i=1,\ldots,L}) = \frac{1}{Z(\boldsymbol{h},\boldsymbol{J})} \prod_{i=1}^{L} e^{h_i(\sigma_i)} \prod_{i<j} e^{J_{ij}(\sigma_i,\sigma_j)}$$

Methods



Problems with BM & L2-regul. :

- Energy discrepancy between natural & artificial sequences
- Bias in the inference of heterogeneous couplings
- Computationally slow

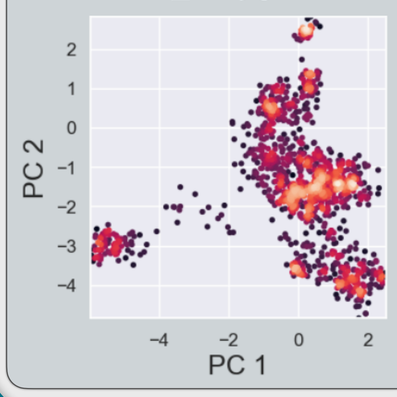
	Current: Boltzmann Machine (BM) [1]	New: Stochastic Boltzmann Machine (SBM)
Optimization Method	Steepest Gradient Descent $p_k = -\nabla f_k$ (Log-Likelihood)	L-BFGS Gradient Descent [4] $p_k = -B_k^{-1}\nabla f_k$ (Low rank approximation of the Hessian)
Hyperparameters	Nb of Gradient Descent iterations: N_{iter} Nb of MCMC steps and chains: k, N_{chains} L2-regularization: λ_h, λ_J	Rank of the approximated Hessian: m
Regularization	(λ_h, λ_J)	$(m, N_{chains}, N_{iter})$

Energy discrepancy between natural & artificial sequences

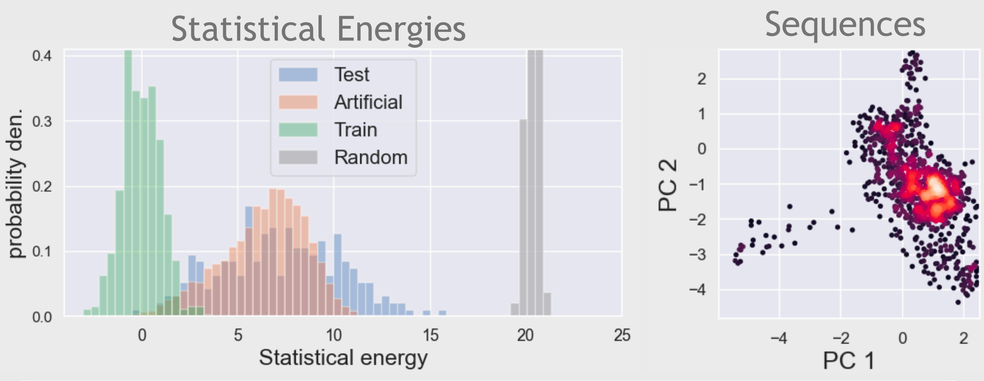
Chorismate Mutase family [3]

$M_{train} = 904$, $M_{test} = 226$

$L = 96$



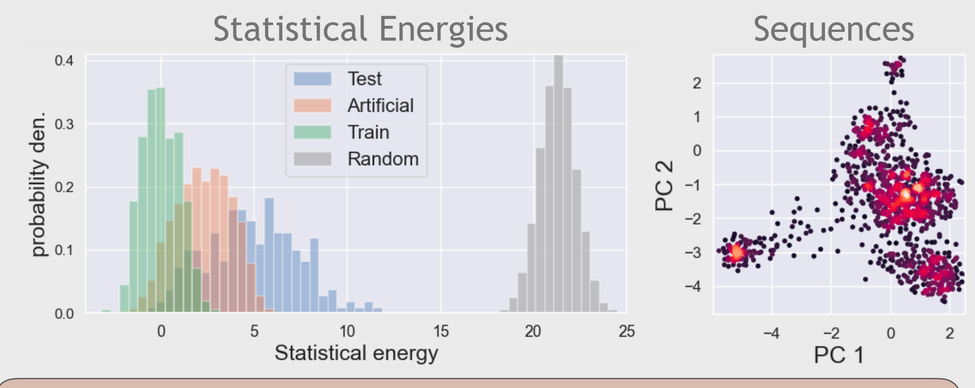
BM ($\lambda = 0.01$, $T = 1$)



→ L2-regularization of BM leads to an energy discrepancy between natural & artificial sequences

→ PCA shows a loss of diversity in the artificial sequences

SBM ($N_{chains} = 40$, $m = 1$, $N_{iter} = 500$, $T = 1$)



→ Energy discrepancy between natural and artificial sequences is smaller than for the BM method

→ PCA shows a greater diversity than for the BM method

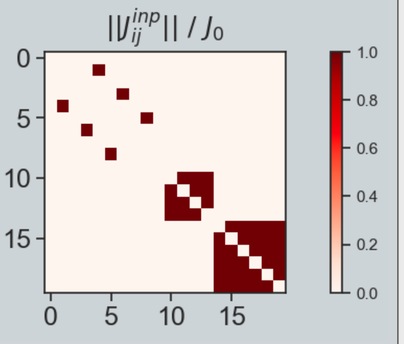
Bias in the inference of heterogeneous couplings

Toy Model [2]

$q = 10$, $L = 20$

$J_0 = 2$, $h_i = 0 \ \forall i$

$M = 300$



BM ($\lambda_h = 0.01$, $N_{chains} = 1000$)

$\lambda_J \sim 10^{-3}$: large collective features have lower magnitude

→ L2-regularization of BM leads to a bias in the coupling inference

SBM ($N_{iter} = 500$, $m = 1$)

$n_{states} \sim 200$: features well inferred with same magnitude

→ Implicit regularization of SBM does not lead to a bias in the coupling inference

Discussion & Future work

- Computation Time: SBM method converges much faster than the BM method
- Open question: How to score without comparing to the true model?

☞ Use this approach to infer Generative models for Bacterial & Archaeal genomes

References

[1] M. Figliuzzi, P. Barrat-Charlaix, and M. Weigt, 2018.

[2] Y. Kleeorin, W. P. Russ, O. Rivoire, and R. Ranganathan, 2021.

[3] W. P. Russ, M. Figliuzzi, C. Stocker, P. Barrat-Charlaix, M. Socolich, P. Kast, D. Hilvert, R. Monasson, S. Cocco, M. Weigt et al., 2020.

[4] D. C. Liu and J. Nocedal, 1989.

marion.chauveau@espci.fr