Quantitative Sequence Activity Models with a little help from quantum field theory

Deborah Striegel[1] and Vipul Periwal[2]

***Short Abstract* — The Gamma function grows faster than the exponential function. This basic mathematical fact implies that the exponential increases in processing power and data storage associated with Moore's Law are no match for the combinatorial complexity of modeling problems associated with high-throughput biological data. Systems of partially known structure such as position specific scoring matrices for transcriptional regulatory codes will remain beyond our ken if we cannot induce a model directly from the data without going through optimization bottlenecks associated with explicitly introducing parameters for each possible unknown interaction between nucleotides at distinct positions. We present here a physical heuristic from quantum field theory for the direct computation of a model from cumulants of data, in a parameter independent manner. Our theory is associated with the physics of a new kind of phase transition, driven by the degree of belief in the model prior. Above the phase transition, the model prior dominates and the only data that contributes is that which agrees best with the model prior maximum. Below the phase transition, we identify collective excitations that are the Goldstone bosons of model symmetry breaking. We give explicit applications of our theory to DNA enhancer motif optimization and to global allosteric modes and contact maps in orthologous protein families.**

***Keywords* — quantitative sequence activity model (QSAM), transcriptional regulation, protein structure, allosteric modes**

I. Introduction

Physical heuristics have a long history in computation. Simulated annealing is the most illustrious recent exemplar of this tradition. When faced with the copious amounts of data engendered by Moore's Law, often dubbed `big data', machine learning approaches meet the even bigger hurdles posed by the factorial Gamma function. As the number of potential factors influencing the data increases, the number of possible interactions between these factors increases combinatorially. It is imperative, therefore, to use some form of model selection to extract information from the data. Eukaryotic gene regulation is combinatorial but it is impractical to test each possible model, for each model comes with its own optimization problem. The use of the rubric of Bayesian model selection presupposes some understanding of the information in the data, accurate enough to determine the model prior. It is usually supposed that using equal probabilities for all models considered or some form of maximum entropy is the appropriate approach, based on analogies with statistical mechanics and conserved quantities or symmetry principles. What should we use if we have protein sequences for orthologous proteins or possible DNA transcription factor enhancer sequences?

II. Methods and Results

We use an exact computation of the effective action associated with a dataset and a model prior at the maximum likelihood model to approximate the phenotype associated with a sequence. Treating the model prior width as a temperature, we find phase transitions at finite temperature defining the lower limits of model uncertainty. The order parameter is the maximal weight of a sequence in the ensemble.

A. Transcriptional regulation

Applying our method to published massively parallel reporter assays[1], our QSAMs with three parameters do as well or significantly better than published QSAMs with hundreds of parameters.

B. Protein Structure

We find contact maps and global allosteric modes using multiple sequence alignments of several protein ortholog families[2]. These contact maps are in agreement with known three-dimensional structures. The flat directions in the energy landscape are global allosteric modes, the Goldstone bosons of model symmetry breaking

III. Conclusion

Quantum field theory methods let us (almost) completely eliminate optimization from the deduction of QSAMs directly from data. This eliminates optimization as a bottleneck from the analysis of high-throughput quantitative phenotype assays. Many problems can be formulated in sequence-phenotype form so this is a protean solution.

References

[1] Melnikov A, Murugan A et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotech* **30**, 271-277

[2] R. Ranganathan and K. Reynolds shared their MSAs. See Halabi N, Rivoire O, Leibler S, Ranganathan R. (2009) *Cell* **138**, 774-786.

Acknowledgements: This work was supported by the Intramural Research Program of the NIH, NIDDK.

[1,2] Laboratory of Biological Modeling, NIDDK, NIH, 9000 Rockville Pike, Bethesda MD 20892. E-mail: deborah.striegel@nih.gov

[2] E-mail: vipulp@mail.nih.gov