Exact Chemical Potentials for Prokaryotic Transcription Factors

Patrick K. O'Neill[1], Ivan Erill[2]

Much theoretical work on transcription factor dynamics relies on the tenet that the occupancy probabilities for transcription factor binding sites are Fermi-Dirac distributed. While the literature grants the assumption that the chemical potential of such a distribution is well-modeled by an ideal gas approximation, it is not difficult to solve for it numerically. We do so for 27 transcription factors in *Escherichia coli* and show that in many cases the ideal gas approximation can diverge dramatically from the exact solution. We conclude that researchers may observe qualitatively different behaviors in models of transcriptional regulation which admit approximate versus exact chemical potentials.

***Keywords* — Transcriptional regulation, chemical potential, Fermi-Dirac statistics**

I. Introduction

Transcriptional regulation is the most ubiquitous mechanism for control of gene expression. Although several specific systems have been studied in great detail within bacterial models, the overwhelming number of distinct transcription factors (TFs) present within a typical bacterial cell motivates the search for models of transcriptional regulation with general applicability.

Early on, researchers in transcription factor dynamics made use of an analogy between steric exclusion and Pauli exclusion to conclude that the genomic distribution of a TF should obey Fermi-Dirac statistics. The Fermi-Dirac distribution is parametrized in part by a chemical potential, conjugate in this case to TF copy number. It was further granted that in the dilute limit the chemical potential could be estimated by an ideal gas approximation which is logarithmic in the copy number [1-3].

II. Model and Methods

In this study, we extend earlier models of TF dynamics by computing the chemical potential to arbitrary precision over the entire genomic sequence in order to derive a stationary occupancy distribution. Binding sites were collected from the PRODORIC database [4] for every TF in *E. coli* containing at least ten experimentally verified sites, and the resulting count matrices were used to fit a site-specific binding energy function for each TF using the TRAP model [5]. These functions were used to compute free energy landscapes over the genome. Chemical potentials were recovered from the resulting equations for conservation of copy number via standard numerical methods. For each TF, we contrasted the *E. coli* genomic sequence to random controls of equivalent length and mononucleotide composition. Additionally, the use of continuous, deterministic methods in modeling dilute reaction systems was validated by comparing the resulting occupancy distributions to those obtained through exact stochastic simulation [6].

III. Conclusion

We find that despite good agreement in the dilute limit, the ideal gas approximation significantly underestimates the exact chemical potential for many TFs in the physiologically relevant regime, leading to underestimation of occupancy. Moreover, the exact chemical potentials for the *E. coli* genome and randomized controls converge as copy number grows large, suggesting a method for estimating the number of functional sites harbored by the genome; the ideal gas approximation cannot capture this effect. We examine the occupancies of experimentally verified sites and show that many regulons are effectively occupied when the chemical potential is computed exactly but empty when approximated; the copy numbers required by the ideal gas scheme can exceed physiological plausibility by orders of magnitude.

Quantitative models of transcriptional regulation demand trade-offs between full attention to the subtleties of TF specificity, regulon size, genome size, and copy number on one hand, and analytic tractability on the other. For many bacterial TFs, though, the ideal gas approximation for the chemical potential leads to qualitatively different model behaviors. Numerical solution is straightforward and relatively inexpensive, and is hence recommended when binding site occupancies are required.

References

[1] Gerland U, Moroz JD, Hwa T (2002) *PNAS* **99**, 12015-12020.
[2] Djordjevic M, Sengupta AM, Shraiman BI (2003) *Genome Res* **13**, 2381-2190.
[3] Zhao Y, Granas D, Stormo G (2009) *PLOS Comput Biol* **9**, e1000590.
[4] Münch R *et al.* (2003) *Nucleic Acids Res.* **31**, 266-269
[5] Roider H *et al.* (2006) *Bioinformatics* **23**, 134-141.
[6] Gillespie D (1977) *J. Phys. Chem.* **81**, 2340–2361.

[1]Department of Biological Sciences, University of Maryland Baltimore County. E-mail: pon2@umbc.edu
[2] Department of Biological Sciences. University of Maryland Baltimore County. E-mail: erill@umbc.eudu