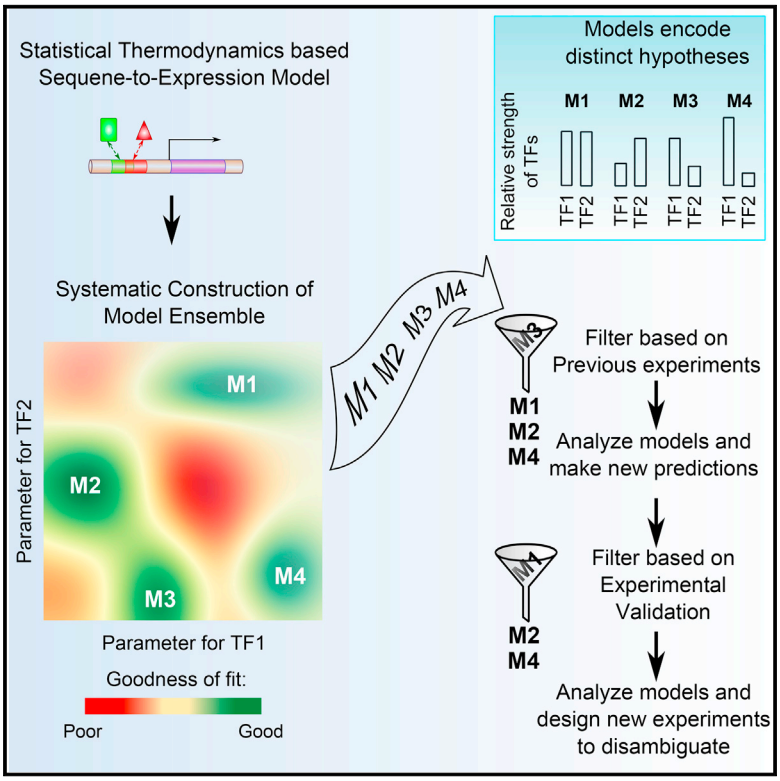
Article

Cell Systems

A Systematic Ensemble Approach to Thermodynamic Modeling of Gene Expression from Sequence Data

Graphical Abstract



Highlights

- Conventional modeling of enhancer readouts may lead to incorrect conclusions
- A model ensemble can capture many distinct hypotheses plausible with data
- Filtering and analysis of a model ensemble can reveal new testable hypotheses
- Ensemble modeling improved current understanding of *ind* regulation in *Drosophila*

Authors

Md. Abul Hassan Samee, Bomyi Lim, Núria Samper, ..., Gerardo Jiménez, Stanislav Y. Shvartsman, Saurabh Sinha

Correspondence

sinhas@illinois.edu

In Brief

A systematically constructed ensemble of sequence-to-expression models captures many distinct, biologically plausible hypotheses. Systematic analysis and filtering of the models, followed by in vivo experiments, improve our understanding of combinatorial regulation of the *Drosophila ind* gene.

Samee et al., 2015, Cell Systems *1*, 396–407
December 23, 2015
http://dx.doi.org/10.1016/j.cels.2015.12.002

CellPress

Cell Systems
Article

A Systematic Ensemble Approach to Thermodynamic Modeling of Gene Expression from Sequence Data

Md. Abul Hassan Samee,[1,8] Bomyi Lim,[2] Núria Samper,[3] Hang Lu,[4] Christine A. Rushlow,[5] Gerardo Jiménez,[3,6] Stanislav Y. Shvartsman,[2] and Saurabh Sinha[1,7,*]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[2]Department of Chemical and Biological Engineering and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[3]Department of Developmental Biology, Instituto de Biología Molecular de Barcelona, Consejo Superior de Investigaciones Científicas (CSIC), Barcelona 08208, Spain
[4]School of Chemical and Biomolecular Engineering and Parker H. Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332, USA
[5]Department of Biology, New York University, New York, NY 10003, USA
[6]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
[7]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[8]Present address: The Gladstone Institutes, University of California, San Francisco, San Francisco, CA 94158, USA
*Correspondence: sinhas@illinois.edu
http://dx.doi.org/10.1016/j.cels.2015.12.002

SUMMARY

To understand the relationship between an enhancer DNA sequence and quantitative gene expression, thermodynamics-driven mathematical models of transcription are often employed. These "sequence-to-expression" models can describe an incomplete or even incorrect set of regulatory relationships if the parameter space is not searched systematically. Here, we focus on an enhancer of the *Drosophila* gene *ind* and demonstrate how a systematic search of parameter space can reveal a more comprehensive picture of a gene's regulatory mechanisms, resolve outstanding ambiguities, and suggest testable hypotheses. We describe an approach that generates an ensemble of *ind* models; all of these models are technically acceptable solutions to the sequence-to-expression problem in light of wild-type data, and some represent mechanistically distinct hypotheses about the regulation of *ind*. This ensemble can be restricted to biologically plausible models using requirements gleaned from in vivo perturbation experiments. Biologically plausible models make unique predictions about how specific *ind* enhancer sequences affect *ind* expression; we validate these predictions in vivo through site mutagenesis in transgenic *Drosophila* embryos.

INTRODUCTION

Transcription factors (TFs) work in concert with other DNA-binding molecules to regulate gene expression. These molecules act as inputs at enhancers, distinct genomic regions that contain binding sites for TFs and can regulate the transcription of target genes (Shlyueva et al., 2014). Maintaining a quantitative relationship between input and transcriptional output is key to the precise patterning of gene expression. Accordingly, as the levels of inputs vary across different cell types, the enhancer-controlled levels of gene expression (also termed as the "readout" of the enhancer) also vary (Yáñez-Cuna et al., 2013). These relationships are a direct function of the enhancer's DNA sequence. However, a detailed understanding of how enhancer sequence affects a gene's expression level remains elusive (Yáñez-Cuna et al., 2013). Such understanding may be achieved by interrogating a mathematical model that explains the available experimental results about the gene both qualitatively and quantitatively, suggests experiments to improve upon the current model, and is capable of predicting the gene's expression pattern upon *cis* or *trans* perturbations. Here, we refer to such models as "sequence-to-expression" models, and we show how they can form the basis of a systematic, unbiased enquiry into gene regulation by multiple TFs.

A common paradigm of sequence-to-expression modeling is based on equilibrium thermodynamics (Shea and Ackers, 1985). This approach models the rate of transcription initiation based on quantitative descriptions of variable site affinities ("motifs") (Stormo, 2000) and expression levels of TFs. Because they can incorporate the DNA-sequence-dependent characteristics of TF binding, sequence-to-expression thermodynamic models of this genre are arguably more realistic than thermodynamic models where all TF-binding sites are assumed to have the same affinity (Cohen et al., 2014; Fakhouri et al., 2010; Papatsenko and Levine, 2008; Zinzen and Papatsenko, 2007) or only classified as "strong" versus "weak" (Bintu et al., 2005; Gertz et al., 2009; Parker et al., 2011; White et al., 2012). We previously reported one such sequence-to-expression model called GEMSTAT (Gene Expression Modeling based on Statistical Thermodynamics) and used it