MODELING TRANSCRIPTIONAL REGULATION WITH EQUILIBRIUM MOLECULAR COMPLEX COMPOSITION

Mjolsness E.
Institute for Genomics and Bioinformatics, and Departments of Computer Science and Mathematics
University of California, Irvine, CA, USA
Corresponding author: e-mail: emj@uci.edu

Key words: equilibrium, composition, EMCC, cooperative activation, statistical mechanics, partition function, allosteric enzyme, transcriptional regulation, systems biology

SUMMARY

Motivation: Regulation of transcription has been modeled in a variety of ways in cellular and developmental systems.

Results: Here we apply a method for creating equilibrium models of hierarchical statistical systems, the Equilibrium Molecular Complex Composition (EMCC) family of models, to the problem of modeling the rate of initiation of transcription in the presence of overlapping binding sites, synergistic binding interactions in one dimension, and modular activation of a transcription complex.

INTRODUCTION

The essential steps for modeling a hierarchical system in equilibrium using the Equilibrium Molecular Complex Composition (EMCC) family of models are to (1) identify the hierarchical levels; (2) model each level with a partition function Z for a Boltzmann distribution, as a function of fugacity parameters $\mathbf{z}$ for constituent molecules or subcomplexes; (3) perform any possible model reduction (including justifiable approximations) on the resulting partition functions $Z(\mathbf{z})$; (4) compose the partition functions, substituting partition functions Z from a finer scale for fugacities $\mathbf{z}$ at a coarser scale. The validity of this procedure follows from an EMCC "Composition Theorem". We will illustrate this procedure in the case of the Monod Wyman Changeaux model of allosteric enzymes, and then apply it to the case of a hierarchical model of transcriptional regulation (Mjolsness, 2001) here generalized to the case of transcription factor binding sites with optional overlaps with their nearest neighbors in one dimension, and optional interaction energies with their second nearest neighbors, and hierarchical activation in terms of transcriptional regulatory modules.

Assume we have a molecular complex defined at each level by a set of binary occupancy variables $s_i \in \{0,1\}$, related through a high-order Ising model. For each slot there is a fugacity variable z_i. We can define a multidimensional array J of interaction energies, whose elements are indexed by the ordered set of indices $\rho(\sigma)$:

$$J_{\rho(\sigma)} = J_{(i(1)<i(2)<\ldots i(l))} \in \mathbb{R}$$

with the convention that any other values of J are 0. Defining $0^0 = 1$, the partition function for equilibrium statistical mechanics is

$$Z(z \mid J) = \sum_{\{s \mid s_i \in \{0,1\}\}} (\prod_i z_i^{s_i}) \prod_{\{\sigma \mid \sigma_i \in \{0,1\}\}} \exp[-\beta J_{\rho(\sigma)} \prod_j (s_j)^{\sigma_j}] \tag{1}$$

Considered as a function of the fugacities **z**, $Z(\mathbf{z})$ is a high-order polynomial and it is a generating function for the (unnormalized) probabilities of all configurations s. However, many J's can tend towards ∞ in such a way as to prohibit particular combinations of values of s_i by giving them zero probability. Also many J's can be exactly zero, so that particular interactions are absent. These possibilities can be encoded by the predicates $P(s)$ and $Q(\sigma)$, respectively, in the following expression for the partition function:

$$\begin{aligned} Z(z \mid J) &= \sum_{\{s \mid P(s)\}} (\prod_i z_i^{s_i}) \prod_{\{\sigma \mid Q(\sigma)\}} \exp[-\beta J_{\rho(\sigma)} \prod_j (s_j)^{\sigma_j}] \\ &\equiv \sum_{\{s \mid P(s)\}} (\prod_i z_i^{s_i}) \prod_{\{\sigma \mid Q(\sigma) \wedge (\wedge_i (\sigma_i \Rightarrow s_i))\}} (\omega)_{\rho(\sigma)} \end{aligned} \tag{2}$$

As a trivial example, a heterodimer of species 1 and 2 with no internal states would have $Z(z_1, z_2) = \omega_{1,2} z_1 z_2$. A protein with a single binding site that can be empty or occupied by species 1 or 2 would have $Z(z_1, z_2) = 1+\omega_1 z_1+\omega_2 z_2$. If the protein is itself regarded as one of the species that can be present or absent, with fugacity z_0, then it must be present and the partition function is $Z(z_1, z_2) = z_0(1+\omega_1 z_1+\omega_2 z_2)$. In each case, as for any probability generating function, the coefficients can be normalized to give the probabilities of each possible configuration of bindings.

Such partition functions can be put into a form with homogeneous degree by introducing the complementary fugacity variables $z_i = z_i^+ z_i^-$: $Z^{\text{homog}}(z^+, z^- \mid \omega) = Z^{\text{homog}}(z^+ / z^- \mid \omega)(\prod_i z_i^-)$. No information is lost since $Z^{\text{homog}}(z \mid \omega) = Z^{\text{homog}}(z^+ = z, z^- = 1 \mid \omega)$.

METHODS AND ALGORITHMS

Composition Theorem. Suppose we have a two-level hierarchical system, with a top level (coarse-scale) partition function Z_0 and a set of lower-level (finer-scale) partition functions. Given partition top-level internal state variables $\{s_0\}$ that can interact with lower-level systems, and lower-level activation variables p_i that can interact with higher-level systems, we can define lower-level partition functions $Z_i^{([s_{0j}], p_i)}(z, \omega)$. Without the indices s_0 and p_i, generating functions for discrete-time branching processes (birth-and-death processes) are obtained by function composition from the generating functions at each succeeding generation, with the first generation as the outermost composition (Athreyea, Ney, 1972). A similar result holds in the present situation.

A "Composition Theorem" gives conditions under which partition functions $Z_0(\mathbf{z})$ at the top level and $\{Z_i(\mathbf{z}_i) \mid i \geq 1\}$ at the next lower level in a scale hierarchy, all of which are in the form of (Equation 2), may be *composed* to give the partition function $Z_{2\text{-level}}([\zeta_i Z_i(\mathbf{z}_i) \mid i \geq 1])$, also in the form of (Equation 2), for the composite molecular complex. Optionally some of the ζ_i may be set to 1 if we do not need to differentiate with respect to them. For example, if $Z_i(\mathbf{z}_i) = (z_{1i})^2$ then there is a model level corresponding to obligatory homodimerization in binding to the top-level complex at position *i*. Likewise if $Z_i(\mathbf{z}_i) = (1 + \omega_1 z_{1i} + \omega_2 z_{2i})$, then there is a binding site which can be empty, or occupied by just one of two competing factors. The composition theorem may be used recursively to model many levels of complex composition.

IMPLEMENTATION AND RESULTS

MWC Example. A simple example is given by the Monod-Wyman-Changeaux model of allosteric enzymes.

Level 1 (top): global activation/inactivation: $Z^1 = \zeta_i \omega_0 Z^{2+} + Z^{2-}$.

Level 2: Independent identical subunits: $Z^{2\pm} = (Z^{3\pm})^n$.

Note: levels 1 and 2 are ordinarily combined.

Level 3: Independent binding heterogeneous sites within each subunit: $Z^{3\pm} = (\prod_{\alpha=1}^{A} Z_\alpha^{4\pm})^n$. The simplest case is $\alpha \in \{1, 2, 3\}$ for substrate/product, activator, and inhibitor respectively.

Level 4: Mutual exclusion (MutEx) for occupation: $Z_\alpha^{4\pm} = \omega_\alpha^\pm + \sum_{i=1}^{n} \omega_{\alpha i}^\pm Z_{\alpha i}^{5\pm}$.

Without loss of generality, take $\omega_\alpha^\pm$ =1 since empty binding sites are never prohibited.

Level 5: Convergence through sharing of fugacity variables, each of which is (for a dilute well-stirred solution in a fixed macroscopic volume) proportional to the number of molecules present and therefore to concentration: $Z_\alpha^{5\pm} = z_i$.

Composition of all levels: $Z = z_0 \omega_0 \prod_{\alpha=1}^{A} (1 + \sum_{i=1}^{n} \omega_{\alpha i}^{+} z_i)^n + \prod_{\alpha=1}^{A} (1 + \sum_{i=1}^{n} \omega_{\alpha i}^{-} z_i)^n$.

The original MWC model has $\omega_{\alpha i}^\pm = 0$ unless $I = \alpha$ and the following condition: $(s = +1 \wedge (\alpha = 1 \vee \alpha = 2)) \vee (s = -1 \wedge (\alpha = 1 \vee \alpha = 3))$ where $\alpha = (1, 2, 3)$ for substrate, activator, and inhibitor respectively. In that case we recover the original MWC model:

$$\begin{aligned} Z = {} & L(1 + \textstyle\sum_{i=1}^{n} c(S / K_S))^n (1 + \sum_{i=1}^{n} (A / K_A))^n \\ & + (1 + \textstyle\sum_{i=1}^{n} c(S / K_S))^n (1 + \sum_{i=1}^{n} (I / K_I))^n \end{aligned} \tag{3}$$

Clearly this model can be generalized to multiple substrates, activators and inhibitors on each subunit, as demonstrated and applied in (Tarek *et al.*, 2006) to amino acid synthesis pathways.

EMCC application to transcriptional regulation. With this apparatus we can rederive and extend a model similar to Hierarchical Cooperative Activation (Mjolsness, 2001) for transcriptional regulation. Transcription factors bind, alone or in multimers such as homodimers or heterodimers, to DNA binding sites that can overlap with their one-dimensional neighbors (in which case they can't be occupied simultaneously) or be sufficient close to their nearest nonoverlapping neighboring sites in one dimension that energetic interactions occur. These possibilities are summarized by allowing overlap with nearest neighbors to either side, interaction with next nearest neighbors to either side, and missing sites that break chains of overlap and/or interaction. At a coarser level, activation occurs in modules or cassettes (such as the Drosophila even-skipped minimal stripe three element) which contribute to overall activation of transcriptional initiation. Within these limitations, we can formulate an equilibrium complex model similar to MWC at several levels. The novel part of this model compared to HCA is the one-dimensional interactions through site overlap and synergy: second nearest neighbors (odd or even) interact energetically with factor ω. Therefore each successive pair of sites has three possible states. The model can be solved using 3 × 3 transfer matrices on site pairs:

$$Z = (1,1,1)\cdot\{\prod_{i=k\searrow 1}\begin{pmatrix} 1 & 1 & 1 \\ z_{2i+1} & z_{2i+1}\omega_{2i-1,2i+1} & 0 \\ z_{2i+2} & z_{2i+2} & z_{2i+2}\omega_{2i,2i+2} \end{pmatrix}\}\cdot\begin{pmatrix} 1 \\ z_1 \\ z_2 \end{pmatrix}$$

Any site can be omitted (removing its overlap constraints and interaction energies) by setting its z_i to be 1 and $\omega_{i*}= \omega_{*I} =1$.

ACKNOWLEDGEMENTS

Thanks to Vitali Likhoshvai, Bruce Shapiro, Tarek Najdi, and Chin-Ran Yang for valuable discussions. This work was supported by Biomedical Information Science and Technology Initiative (BISTI) grant (No. 4 R33 GM069013) from the National Institute of General Medical Sciences, USA, and by US National Science Foundation Frontiers in Biological Research (FIBR) grant No. EF-0330786.

REFERENCES

Athreyea K.B., Ney P.E. (1972) Branching Processes. Springer-Verlag; Dover.

Mjolsness E.D. (2001) Gene Regulation Networks for Modeling Drosophila Development. In Bower J.M., Bolouri H., (eds), *Computational Methods in Molecular Biology*, MIT Press.

Tarek S.N., Chin-Ran Yang, Shapiro B.E., Hatfield G.W., Mjolsness E.D. (2006) Application of a Generalized MWC Model for the Mathematical Simulation of Metabolic Pathways Regulated by Allosteric Enzymes. *J. of Bioinformatics and Computat. Biol.* to appear.