Parallel Session

Genetics and Genomics III

SHEDDING NEW LIGHT ON RANDOM CHROMOSOME SEGREGATION

Qi Zheng

qzheng@sph.tamhsc.edu

Texas A&M School of Public Health, College Station, Texas 77843, USA

Keywords: Polyploidy, Mutation, Homozygosity, Gene conversion, Agent-based model.

A cell's ploidy value is the number of chromosomes (genomes) residing in that cell. Biologists now believe that bacterial polyploidy is more common than they thought before. For a given mutation, a cell is called homozygous if all genomes in that cell carry the same mutation of interest. A long-standing challenge is to explain how homozygosity arises in a highly polyploid cell population. Let a cell having c genomes carry just one mutated genome; simulation suggests, if chromosome segregation is random, one would wait on average for c cell generations to see a homozygous cell arising. With $c = 100$, one would then witness about 10^{30} cell divisions, which is about the total number of microbial cell divisions occurring annually on the earth. To circumvent this conundrum, biologists in 1980 proposed a model of nonrandom segregation. Recent biologists favor the gene conversion hypothesis. The random segregation model, despite its conceptual simplicity and intuitive appeal, fell out of favor before its elementary properties were ever understood. The advent of high performance computing technology has somewhat untethered the random segregation model from intractable mathematics. Using an agent-based simulation approach, I have caught glimpses into the joint effects of mutation and selection on the formation of homozygosity. For example, if cells carrying relatively more mutated genomes are selected very 20 generations, a succession of 3 rounds of selection is sufficient to produce a large number of homozygous cells having a ploidy value of 100. I shall discuss important evolutionary implications of these findings.

Acknowledgements: Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

Parallel Session

Genetics and Genomics III

MODELING AND ESTIMATION OF SUBSTITUTION RATES ALONG PHYLOGENETIC TREES BY STOCHASTIC BRIDGES

Nicolas Privault

nprivault@ntu.edu.sg

Nanyang Technological University

Keywords: Evolutionary rates, Geometric Brownian bridge, Molecular clocks, Phylogenetics.

This talk reviews several techniques for the accurate estimation of the probability distribution of substitution rates along phylogenetic trees. The approach relies on the modeling of the rate of molecular evolution using geometric Brownian motion or other diffusion processes. This includes the implementation of moment matching techniques that can be applied more generally to the estimation of path integrals. We also present numerical simulations and error bounds, in agreement with other approximations proposed in e.g. [1] for small values of the autocorrelation parameter.

References

[1] S. Guindon. From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages. *Syst. Biol.*, 62(1):22–34, 2013.

[2] N. Privault and S. Guindon. Closed form modeling of evolutionary rates by exponential Brownian functionals. *J. Math. Biol.*, 71(6-7):1387–1409, 2015.

Parallel Session

Genetics and Genomics III

REVEALING INDIVIDUAL-LEVEL HETEROGENEITY IN INFECTIVITY FROM PATHOGEN PHYLOGENY

Yun Jun Zhang

zhang@math.su.se

Department of Mathematics, Stockholm University, Stockholm, Sweden and Department of Health Dataology, Peking University Health Science Center, Beijing, China

Joint work with Tom Brittion (Department of Mathematics, Stockholm University, Stockholm, Sweden), Jan Albert (Department of Microbiology, Tumor and Cell Biology, Karolinska Institute and Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden) and Thomas Leitner (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America).

Keywords: Epidemic model, Heterogeneity, Phylodynamics, Epidemic control.

Individual level heterogeneity in infectivity are frequently observed in real epidemic outbreaks, which will not only affect the spread of the disease but also disturb the control strategies designed for a homogeneous system. From an evolutionary point of view, this heterogeneity contributes to shaping the genetic diversity of pathogens and hence leaves fingerprints in the reconstructed pathogen phylogenies from genetic data. Neglecting this heterogeneity and instead assuming equal infectivity for all hosts will bias parameter estimates in the analysis of pathogen phylogeny.

In this study, we introduce a heterogeneous birth-death model to extract information on the heterogeneity of infectivity from the pathogen phylogeny. In this model, each host draws a random infection rate from a continuous probability distribution that encodes all variation in infectivity of individuals. By studying the relationship between neighboring internal branches in a phylogeny, it is possible to recover the infectious information of individuals and furthermore to estimate the population-level heterogeneity in infectivity (i.e., the standard deviation of the random infection rate).

With numerical simulations, the method yields the accurate estimation of population level mean and heterogeneity in infectivity. Also, the new method has been applied to the real dataset of HIV outbreak and reveals the heterogeneity in HIV transmission. Finally, based on the estimated heterogeneity in infectivity, we study the optimization of intervention strategy to control the outbreak under the heterogeneous situation.

Parallel Session

Genetics and Genomics III

STOCHASTIC ACTIVATION IN A GENETIC SWITCH

Joanna Tyrcha

joanna@math.su.se

Dept. of Mathematics, Stockholm University

Joint work with John Hertz (University of Copenhagen and Nordita, Stockholm) and Alvaro Correales (UA Madrid and Dept. of Mathematics, Stockholm University).

Keywords: Gene regulation, Stochastic dynamics, Path integrals.

Biological networks generically exhibit discontinuous transitions between different states, characterised by different gene expression patterns. Sometimes these transitions can be stochastic, driven by random fluctuations of molecular concentrations. In this work, we analyse such transitions in a minimal model of a self-regulating gene. We employ generalisations of theoretical treatments of simpler problems ([1, 2]) to calculate the activation rate in the limit where the transitions are rare.

Our simple model contains a single mRNA species and its associated protein. The mRNA is assumed to be transcribed at a rate proportional to a Hill function of protein concentration, and the rate of protein production is taken proportional to the mRNA concentration. An important parameter of the problem is the ratio of the lifetime of the protein to that of the mRNA. We denote it by γ. The large-γ limit has been studied extensively in the steady-state limit ([3, 4]), where proteins are produced in exponentially (geometrically) distributed bursts. In this limit, the mRNAs are slaved to the proteins and we have an effective one-species problem.

It is well-known that a Fokker-Planck approach (or, equivalently, adding Gaussian noise to the chemical rate equations) is not sufficient to describe the large fluctuations involved in the rare transitions of interest here. In our calculations, we use both a path-integral formulation and Chapman-Kolmogorov equations to describe the stochastic chemical kinetics. We study the problem for a range of values of γ, with particular attention to the bursting limit ($\gamma \to \infty$). The activation rate can be written in the form $r_0 \exp(-S_0)$, where S_0 is the action of the optimal path (history of the molecular concentrations) from a locally stable state to the transition state through which the system passes en route to the other locally stable state. It is inversely proportional to the size of the molecular concentration fluctuations. The prefactor r_0 comes from fluctuations around the optimal path. It does not depend on the size of the fluctuations.

In the bursting limit we are able to calculate both S_0 and r_0 exactly. For general γ, numerical calculations are required. We can also calculate S_0 analytically for large γ by expanding

in $1/\gamma$ around the bursting limit. This approximation agrees well with the numerical calculations for γ as small as 2-3. We have also performed numerical simulations of the model. These agree well with the theoretical calculations.

References

[1] H. A. Kramers. (1940). *Brownian motion in a field of force and the diffusion model of chemical reactions*. Physica 7, 284-304.

[2] S. Glasstone, K. J. Laidler, H. Eyring. *The Theory of Rate Processes*. McGraw-Hill, New York, 1941.

[3] N. Friedman, L. Cai, X.S. Xie. (2006). *Linking stochastic dynamics to population distribution: an analytical framework of gene expression*, Phys. Rev. Lett. 97, 168302.

[4] M. C. Mackey, M. Tyran-Kamińska, R. Yvinec. (2011). *Molecular distributions in gene regulatory dynamics*. J. Theor. Biol. 274, 84-96.

Parallel Session

Genetics and Genomics III

PROBABILISTIC GRAPHICAL MODELS FOR LOSS-OF-FUNCTION GENOMIC SCREENS ANALYSIS

Ksenia Nikitina

nikitina@mpi-cbg.de

Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany

Joint work with Yannis Kalaidzidis (MPI-CBG), Marino Zerial (MPI-CBG).

Keywords: High-throughput screen, Bayesian network, Machine Learning.

High content (HC) and -throughput screens (HTS) based on light microscopy imaging and quantitative analysis give rise to a large amount of multi-parametric phenotypic data. Such data-sets hold great promise for the understanding of cellular mechanisms by systems biology. The current approaches of clustering and enrichment analysis [1, 2] generate mostly a catalog of genes. However, only limited success is typically achieved when one tries to extract functional information, such as novel links between cellular processes (e.g. signaling, metabolism, etc.) and new functions for unknown genes. The limitation of HTS analysis results from the complex interdependence of functional cellular modules. Furthermore, the technology used for the (genetic or chemical) perturbation itself can be susceptible to a number of experimental biases (e.g. off-target, siRNA seed effects, etc.), which cannot be fully compensated by classical statistical analysis.

An alternative approach to genomic HTS data analysis is based on Probabilistic Graphical Models (PGM) [3, 4, 5]. A Bayesian Network (BN) can be represented as a directed acyclic graph, where nodes correspond to measured phenotypic parameters and edges correspond to dependencies between them. The process of BN learning applied to screen datasets results in reconstruction of relationships between measured features and, in some cases, provides the underlying causal interpretation. Since PGM encodes a join probability distribution of parameters, the law of probabilistic inference is applicable to such model, which allows predicting a gene function with some probability by its phenotype in the screen. However, the current results of PGM analysis of HTS data are limited and new methods and algorithms need to be developed and generalized for multiple applications.

In this study, we developed a PGM-based Machine Learning method for the analysis of genomic HTSs and applied it to HC image-derived HTS for endocytosis in HeLa cells [4]. The changes of phenotypic characteristics of treated cells relative to untreated ones were quantified and combined in multi-parametric vectors (profiles). In total the HTS dataset consists of profiles of 6000 different genes after knockdown via several siRNAs.

Since endocytosis has a strong crosstalk with signaling, metabolism and several other pathways, downregulation of genes from those pathways may score in the screen. Therefore,

reconstruction of PGM based on such data promises to identify connections between cellular modules revealed by measuring endocytosis.

As proof-of-principle, we selected a subset of phenotypic data of well-annotated genes of basic cellular pathways based on information from KEGG. This dataset was then used for training and testing of Machine Learning models and validation of their prediction quality. Our POP demonstrated that even the simplest Naïve BN over-performs the best combination of clustering and enrichment analysis (precision of pathway prediction increases 3.5 folds).

The key advantage that favorably distinguishes PGM from other Machine Learning technics is the possibility to interpret the structure learned from the data. For instance, the use of general BN revealed causal relationships between cell morphological and endocytic parameters. Moreover, the BN framework provides a natural way to implement compensation of siRNA seed-based off-target effects [6] by imposing prior in network learning procedure.

References

[1] M. Boutros, A.A. Kiger, S. Armknecht, K. Kerr, M. Hild, B. Koch, S.A. Haas, R. Paro, N. Perrimon, Heidelberg Fly Array Consortium. (2004). *Genome-wide RNAi analysis of growth and viability in Drosophila cells*. Science, 303(5659), 832-835.

[2] A. Birmingham, L.M. Selfors, T. Forster, D. Wrobel, C.J. Kennedy, E. Shanks, J. Santoyo-Lopez, D.J. Dunican, A. Long, D. Kelleher, Q. Smith. (2009). *Statistical methods for analysis of high-throughput RNA interference screens*. Nature methods, 6(8), 569-575.

[3] N. Friedman, M. Linial, I. Nachman, D. Pe'er. (2000). *Using Bayesian networks to analyze expression data*. Journal of computational biology, 7(3-4), 601-620.

[4] C. Collinet, M. Stöter, C.R. Bradshaw, N. Samusik, J.C. Rink, D. Kenski, B. Habermann, F. Buchholz, R. Henschel, M.S. Mueller, W.E. Nagel, E. Fava, Y. Kalaidzidis, M. Zerial. (2010). *Systems survey of endocytosis by multiparametric image analysis*. Nature, 464(7286), 243-250.

[5] B. Snijder, R. Sacher, P. Rämö, E.M. Damm, P. Liberali, L. Pelkmans. (2009). *Population context determines cell-to-cell variability in endocytosis and virus infection*. Nature, 461(7263), 520-523.

[6] A.L. Jackson, J. Burchard, J. Schelter, B.N. Chau, M. Cleary, L. Lim, P.S. Linsley. (2006). *Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity*. Rna, 12(7), 1179-1187.

Parallel Session

Genetics and Genomics III

THE RIGHT WORD IN THE RIGHT PLACE: OPTIMIZING CODON USAGE FOR PROTEIN TRANSLATION

Christel Kamp

christel.kamp@pei.de

Paul-Ehrlich-Institut, Langen, Germany

Joint work with Jan-Hendrik-Trösemeier (Paul-Ehrlich-Institut, Langen, Germany), Sophia Rudorf (Max Planck Institute of Colloids and Interfaces, Potsdam-Golm, Germany), Holger Loessner (Paul-Ehrlich-Institut, Langen, Germany), Benjamin Hofner (Paul-Ehrlich-Institut, Langen, Germany), Andreas Reuter (Paul-Ehrlich-Institut, Langen, Germany), Thomas Schulenborg (Paul-Ehrlich-Institut, Langen, Germany), Ina Koch (Goethe University Frankfurt, Germany), Isabelle Bekeredjian-Ding (Paul-Ehrlich-Institut, Langen, Germany) and Reinhard Lipowsky (Max Planck Institute of Colloids and Interfaces, Potsdam-Golm, Germany).

Keywords: Codon-specific elongation model, Ribosome dynamics, Protein translation, Codon optimization, Optimized Codon Translation fOr PrOtein Synthesis.

Predicting and optimizing protein expression levels of synthetic genes is a complex task. One important aspect is the adjustment of codon usage which is largely done by adaptation of codon choices to those seen in highly expressed genes of a given organism. In view of frequently encountered suboptimal outcomes we introduce the codon-specific elongation model (COSEM) as a mechanistic approach to study translational control by codon choice. COSEM allows for stochastic simulations of ribosome dynamics based on codon-specific translation speed and accuracy and as such for predictions on protein translation rates per mRNA. We refine this predictor for protein synthesis rates into a protein expression score by considering additional mRNA sequence features whose impact on translation we estimate by model-based boosting methods. We validate the protein expression score by comparing predicted with observed protein levels from genes found in *E. coli*, *S. cerevisiae* and human HEK293 cell lines. Choosing codons that maximize the protein expression score further allows for inference of optimal codons, i.e. those that maximize protein synthesis rates as often desired in heterologously expressed genes. In contrast to standard, heuristic procedures for codon adaptation, our systems view of translation allows for fine tuning of features such as translation speed or accuracy in a context dependent manner to achieve improved results particularly where standard procedures do not deliver satisfactory results. To demonstrate this, we optimized and tested heterologous expression of two genes, *manA* and *ova*, in *Salmonella enterica* serovar Typhimurim, which showed a threefold increase in protein yield compared to wild type and commercially optimized sequences. Our multi-parameter algorithm for codon-adaptation is implemented in the software OCTOPOS (Optimized Codon

Translation fOr PrOtein Synthesis), which will be briefly introduced in combination with potential applications including tailor-made protein synthesis or pathogen attenuation.