BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 17 2011, pages 2459–2462
doi:10.1093/bioinformatics/btr407

Systems biology

Advance Access publication July 6, 2011

Simulating systems genetics data with SysGenSIM

Andrea Pinna[1], Nicola Soranzo[1], Ina Hoeschele[2,3] and Alberto de la Fuente[1,*]

[1]CRS4 Bioinformatica, 09010 Pula (CA), Italy, [2]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 and [3]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0477, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: SysGenSIM is a software package to simulate Systems Genetics (SG) experiments in model organisms, for the purpose of evaluating and comparing statistical and computational methods and their implementations for analyses of SG data [e.g. methods for expression quantitative trait loci (eQTL) mapping and network inference]. SysGenSIM allows the user to select a variety of network topologies, genetic and kinetic parameters to simulate SG data (genotyping, gene expression and phenotyping) with large gene networks with thousands of nodes. The software is encoded in MATLAB, and a user-friendly graphical user interface is provided.

Availability: The open-source software code and user manual can be downloaded at: http://sysgensim.sourceforge.net/

Contact: alf@crs4.it

Received on May 31, 2011; revised on July 1, 2011; accepted on July 4, 2011

1 INTRODUCTION

The central goal of systems biology is to gain a predictive, system-level understanding of biological networks. This entails inferring causal networks from observations on a perturbed biological system. An ideal experimental design for causal inference is randomized, multifactorial perturbation (Fisher, 1926). The recognition that the genetic variation in a segregating population represents randomized, multifactorial perturbation (Jansen, 2003; Jansen and Nap, 2001) gave rise to 'Genetical Genomics' and 'Systems Genetics' (SG), where a segregating or genetically randomized population is genotyped at (many) DNA variants, and is profiled for (disease) phenotypes of interest, genome-wide gene expression and potentially other omics variables (epigenomics, microRNA expression, proteomics, metabolomics, etc.). SG experiments and studies enable us to elucidate the genetic control of gene expression (and other omics variables) (Brem *et al.*, 2002; Keurentjes *et al.*, 2006; Schadt *et al.*, 2003), to annotate DNA polymorphisms implicated in previous genome-wide association studies (GWAS) for particular diseases and to infer key control genes and pathways causally underlying a disease or biomedical trait of interest (Rockman, 2008; Schadt, 2009).

Many statistical and computational methods are being developed for the analysis of SG data. An important component of any SG analysis is the quantitative trait locus (QTL) mapping of all expression traits (etraits) and other omics traits if available. It is well known that the etraits of groups of genes share common regulators (DNA variants), which are more easily identified when associated with a group of etraits rather than with individual etraits. Several approaches to associating DNA variants with groups of etraits have recently been proposed (e.g. Chun and Keles, 2009; Lee *et al.*, 2009, 2006; Parkhomenko *et al.*, 2007; Waaijenborg *et al.*, 2008; Zhang *et al.*, 2010).

A major goal of SG studies is to reconstruct a causal network whose nodes are the phenotypes, the etraits (and potentially other omics variables) and the DNA variants. Methods proposed to achieve this goal include Bayesian networks [BN; (Zhu *et al.*, 2004)], differential equation models (Bansal *et al.*, 2007; de la Fuente *et al.*, 2002), structural equation modeling [SEM; (Li *et al.*, 2006, 2008)] and undirected dependency graph or co-expression network with edge orientation using DNA variants as causal anchors (Aten *et al.*, 2008; Chaibub Neto *et al.*, 2008).

While multiple methods for QTL mapping of etraits (omics variables) and for causal network inference are available, at the present time not much is known about the strengths and weaknesses of all of these proposed methods and whether or when some methods perform better than others. However, researchers increasingly realize that thorough verification of algorithms in bioinformatics and (genetical) systems biology is required. In fact, several international competitions are organized on an annual basis to compare computational methods for systems biology and genetic analysis. These include the Dialogue for Reverse Engineering Assessments and Methods (DREAM) project with its Reverse-Engineering Challenges (Stolovitzky *et al.*, 2007, 2009; http://www.the-dream-project.org/), for which SysGenSIM has been used to produce the SG challenges in 2010, and the Genetic Analysis Workshops (Cupples *et al.*, 2009; http://www.gaworkshop.org), which compare analysis tools relevant for current analytical problems in genetic epidemiology, statistical genetics and genetical systems biology. The availability of realistically simulated (artificial) datasets, which are generated under a set of assumptions most relevant to real SG data, is of utmost importance for the verification of algorithms for SG data analysis. Several SG papers use simulations which are typically simplistic and not general (e.g. Liu *et al.*, 2008; Zhang *et al.*, 2010; Zhu *et al.*, 2004). Other more general software packages have been developed for simulating gene expression data with network models for gene network inference algorithm evaluation [e.g. ABIOCHEM (Mendes *et al.*, 2003), GeneNetWeaver (Marbach *et al.*, 2009; Schaffter *et al.*, 2011) and Ingeneue (Meir *et al.*, 2002)], but experimental designs are restricted to time-series and steady-state measurement after environmental or kinetic parameter perturbations, and single-gene perturbation experiments. These and

*To whom correspondence should be addressed.

other existing packages do not permit the simulation of SG data, in particular the integration of DNA variation, transcriptomics, epigenomics, etc. This is the reason why we have developed and continue to develop SysGenSIM to simulate SG data.

2 GENE EXPRESSION DYNAMICS

Steady-state gene expression traits are simulated for a population of individuals, based on a gene network topology and the individuals' genotypes at a set of genome-wide DNA variants, using non-linear ordinary differential equations (ODEs). The rate law used in SysGenSIM for transcription is not based on any explicit biochemical mechanism, but it displays two main features of biochemical kinetics: saturation and cooperativity (Mendes *et al.*, 2003). We assume that mRNA decay is a first-order process. The ODE for gene g is:

$$\frac{dG_g}{dt} = v_{\text{transcription}_{G_g}} - v_{\text{degradation}_{G_g}} = Z_g^c \cdot V_g \cdot \theta_g^{\text{syn}} \cdot \prod_k \left(1 + A_{kg} \frac{G_k^{h_{kg}}}{G_k^{h_{kg}} + (K_{kg}/Z_k^t)^{h_{kg}}} \right) - \lambda_g \cdot \theta_g^{\text{deg}} \cdot G_g \quad (1)$$

where G_g is the mRNA concentration of gene g, V_g is its basal transcription rate and λ_g is the degradation rate constant. The G_k are the mRNA concentrations of genes which have directed edges into node G_g. K_{kg} is a Michaelis constant (representing the concentrations of input gene k at which its effect on the transcription rate of gene g is half of its maximum effect), h_{kg} is a cooperativity coefficient and A_{kg} is an element of matrix **A** encoding the signed network structure ($A_{kg} = -1$ for inhibitor, $A_{kg} = 1$ for activator, $A_{kg} = 0$ for no effect). The parameters θ_g^{syn} and θ_g^{deg} represent non-genetic internal biological noise in the transcription and degradation rates, respectively; their values are sampled from normal distributions with mean 1 and user-specified SDs prior to the calculation of each steady state. Z_g^c and Z_k^t are parameters which incorporate effects of DNA variants (see Section 6 for details). After generating a network topology (Section 3), the non-linear equations are formulated according to this topology, encoded in matrix **A**. Kinetic parameters V_g, K_{kg}, h_{kg} and λ_g are initialized by sampling values from certain distributions [Uniform, (truncated) Gaussian or Gamma with default or user-specified parameter values] to generate a set of base parameter values, i.e. the 'genetic background' of the organism. The gene expression variability among individuals in the population results from different genotypes (values of the Z_g^c and Z_k^t parameters) and additional biological fluctuations (represented by the noise parameters θ_g^{syn} and θ_g^{deg}).

After setting the values of all parameters Z_g^c and Z_k^t according to the genotypes of an individual in the population, a value for the biological noise terms θ_g^{syn} and θ_g^{deg} is sampled, and the steady-state mRNA concentrations are calculated. This process is repeated for all individuals in the population. Finally, normally distributed multiplicative experimental noise is added to each mRNA concentration at a user-specified level, resulting in a set of expression values for all genes in the system and all individuals. The values for parameters θ_g^{syn} and θ_g^{deg}, and the experimental noise level can be chosen such that the distribution of estimated 'heritabilities' of the etraits [steady-state variances simulated without biological (θ_g^{syn} and θ_g^{deg}) and experimental noise divided by the steady-state variances simulated with these noise terms] is close to those found in real data. For example, in our previous work (Liu *et al.*, 2008), the simulated expression traits had an average heritability of 56%, close to what was observed in a yeast SG experiment (Brem and Kruglyak, 2005).

Due to a highly efficient implementation to solve for steady states, SysGenSIM is able to efficiently generate data with networks of 10 000 nodes with the non-linear dynamical model (~2 min per steady state using a single core of an AMD Opteron X2380 QuadCore, 2.5 GHz). This approach will be described in detail elsewhere, but essentially we solve for steady-state values of genes that are not involved in any cycle very quickly and analytically, while we only deal with the cyclic components of the network numerically by using the function *ode45* in MATLAB. The decomposition of the network in acyclic and cyclic components increases the computational efficiency substantially, because cyclic components usually make up a relatively small part of biological networks (Ma'ayan *et al.*, 2008; Ma and Zeng, 2003).

3 NETWORK TOPOLOGY

The precise topological structure of genotype–gene–phenotype networks is largely unknown. Multiple studies (protein interaction, metabolomic, transcriptomic, etc.) provide evidence for topologies that are scale free, hierarchical and modular (e.g. Barabasi and Oltvai, 2004; Hartwell *et al.*, 1999). Many algorithms to generate (or 'grow') networks *in silico* have been proposed, each reproducing particular characteristics observed in biomolecular networks (such as clustering, degree distributions, motif occurrences, etc.), but none can generate networks displaying all observed topological properties simultaneously. SysGenSIM is able to generate data under the current, standard topology models [Erdös-Renyi random graph (Erdös and Renyi, 1959) and scale-free network (Barabasi and Albert, 1999)]. Furthermore, SysGenSIM is capable of generating random modular networks and, most importantly, modular networks with exponential in-degree and power law out-degree distributions, as observed in real gene networks (Guelzim *et al.*, 2002). SysGenSIM also allows the user to input the network structure as inferred from an actual dataset in the form of a (signed) edge list. The signs of edges, representing activation versus inhibition, can be assigned randomly node wise or edge wise (see the user manual for more information).

4 GENETIC DATA

In terms of the type of the segregating population of individuals for which the SG data are generated, SysGenSIM is currently limited to an inbred line cross commonly employed in real SG experiments in model organisms (e.g. mouse) and plants: Recombinant Inbred Lines (RIL) created by selfing or brother–sister matings from two inbred parental lines. In an RIL population, each DNA variant has two genotypes. SysGenSIM simulates genotype data at all functional (gene) and measured (marker) DNA variants according to a randomly generated genetic map based on user-specified parameter values (e.g. chromosome number, number of genetic markers per chromosome with constant or normally distributed pair-wise distance among DNA variant locations in centi Morgan) or based on a (real) map provided by the user (see the user

manual for more information). The user can choose between Haldane's or Kosambi's mapping function to convert map distance to recombination rate in the generation of genotypes at linked loci. The user can choose between placing one marker in perfect linkage with each functional polymorphism (in this case the number of markers is equal to the number of genes, i.e. the network size) or generating a (sparser) marker map first and then placing functional variants randomly throughout the genome (at minimum distance of 100 kb; see Section 6 below).

5 PHENOTYPE DATA

The user can select one or more continuous macroscopic phenotypes which will be added as nodes to the gene network. As genes can be causal or reactive to the phenotype(s) (Schadt *et al.*, 2005), the user can select the number of genes which directly affect a phenotype and the number of genes which are directly affected by a phenotype. Inputs and outputs of the phenotype node are randomly selected from the gene network. Currently, a phenotype is modeled with Equation (1) where it non-linearly depends on its input genes and additional biological variability.

6 GENOTYPE EFFECTS ON EXPRESSION DYNAMICS

We currently assume that each gene in the network has a single functional DNA variant. The variant is located either in the gene's promoter region affecting its own transcription rate (*cis*-variant with, for example, $Z_g^c = 1$ for one genotype and $Z_g^c = 0.75$ for the other; reduced Z_g^c reflects a less efficient transcription process), or in the coding region of a regulatory gene altering the strength of its regulatory effect (*trans*-variant for which a reduced Z_t^k reflects a less potent inhibitor/activator). Promoter variants modify the kinetics of recruitment of the transcriptional machinery to the promoter sequence, which affects the efficiency of transcription, so a change in Z_g^c results in a change of the basal transcription rate of G_g. A *trans*-effect occurs through changes in the kinetic properties of the product of the gene containing the polymorphism in its coding region. Because we do not explicitly include proteins in our networks, we model these kinetic changes by their effect on the transcription rates of the target genes, by altering their Michaelis constant. The protein products of allelic variants of G_k may have reduced or increased strength through adjustment of Z_k^t. The probabilities of a locus acting in *cis* or in *trans* can be set by the user, as well as the allelic values of Z_g^c and Z_k^t.

7 FUTURE DEVELOPMENT

SysGenSIM is a work in progress with many possible future developments. Of highest priority are improvements to the simulation of the continuous phenotype nodes (e.g. realistic heritabilities, numbers and sizes of QTLs, numbers of causal and reactive modules), the inclusion of discrete (disease) phenotype data, and extensions of the simulation of genotype and steady-state data to other types of inbred line crosses and to human cohorts and case–control designs. To keep pace with recent and future real SG experiments and studies, we plan to extend the simulation of genotype data from bi-allelic DNA variants (single nucleotide polymorphisms) to copy number variation and to incorporate epigenomics data (e.g. DNA methylation sites) and microRNAs into the gene networks. Given the general SG simulation methodology described in this article and the existence of simulators for genome-wide association studies [HapSample (Wright *et al.*, 2007); genomeSIMLA (Dudek *et al.*, 2006)], these extensions are actually quite straightforward. Furthermore, to ensure that the simulated data display known characteristics of real SG data, such as distributions of means, variances and heritabilities of etraits and correlations among etraits, we will continue to estimate the values of such parameters from real SG data and utilize the results from similar studies in the literature. Finally, we continue to implement additional topology models for the generation of gene networks (with emphasis on hierarchical modularity and scale-free out-degree and exponential in-degree distributions).

Funding: Regional Authorities of Sardinia to A. de la Fuente and National Institutes of Health grant (1R01HG005254-01 to I.H.) in part.

Conflict of Interest: none declared.

REFERENCES

Aten,J.E. *et al.* (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst. Biol.*, **2**, 34.

Bansal,M. *et al.* (2007) How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, **3**, 78.

Barabasi,A.L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Brem,R.B. and Kruglyak,L. (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA*, **102**, 1572–1577.

Brem,R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Chaibub Neto,E. *et al.* (2008) Inferring causal phenotype networks from segregating populations. *Genetics*, **179**, 1089–1100.

Chun,H. and Keles,S. (2009) Expression quantitative trait Loci mapping with multivariate sparse partial least squares regression. *Genetics*, **182**, 79–90.

Cupples,L.A. *et al.* (2009) Genetic Analysis Workshop 16: Strategies for genome-wide association study analyses. *BMC Proc.*, **3** (Suppl. 7), S1.

de la Fuente,A. *et al.* (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.*, **18**, 395–398.

Dudek,S.M. *et al.* (2006) Data simulation software for whole-genome association and other studies in human genetics. *Pac. Symp. Biocomput.*, **11**, 499–510.

Erdös,P. and Renyi,A. (1959) On Random Graphs. *Publ. Math. Debrecen*, **6**, 290–297.

Fisher,R.A. (1926) The arrangement of field experiments. *J. Ministry Agric. Great Britain*, **33**, 503–511.

Guelzim,N. *et al.* (2002) Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.*, **31**, 60–63.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Jansen,R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat. Rev. Genet.*, **4**, 145–151.

Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Keurentjes,J.J. *et al.* (2006) The genetics of plant metabolism. *Nat. Genet.*, **38**, 842–849.

Lee,S.I. *et al.* (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl Acad. Sci. USA*, **103**, 14062–14067.

Lee,S.I. *et al.* (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genet.*, **5**, e1000358.

Li,R. *et al.* (2006) Structural model analysis of multiple quantitative traits. *PLoS Genet.*, **2**, e114.

Liu,B. *et al.* (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**, 1763–1776.

Ma'ayan,A. *et al.* (2008) Ordered cyclic motifs contribute to dynamic stability in biological and engineered networks. *Proc. Natl Acad. Sci. USA*, **105**, 19235–19240.

Ma,H.W. and Zeng,A.P. (2003) The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **19**, 1423–1430.

Marbach,D. *et al.* (2009) Generating realistic *in silico* gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol.*, **16**, 229–239.

Meir,E. *et al.* (2002) Ingeneue: a versatile tool for reconstituting genetic networks, with examples from the segment polarity network. *J. Exp. Zool.*, **294**, 216–251.

Mendes,P. *et al.* (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19** (Suppl. 2), ii122–ii129.

Parkhomenko,E. *et al.* (2007) Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.*, **1** (Suppl. 1), S119.

Rockman,M.V. (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature*, **456**, 738–744.

Schadt,E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature*, **461**, 218–223.

Schadt,E.E. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Schadt,E.E. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.

Schaffter,T. *et al.* (2011) GeneNetWeaver: *In silico* benchmark generation and performance profiling of network inference methods. *Bioinformatics* [Epub ahead of print June 22, 2011, doi:10.1093/bioinformatics/btr373].

Stolovitzky,G. *et al.* (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann. N Y Acad. Sci.*, **1115**, 1–22.

Stolovitzky,G. *et al.* (2009) Lessons from the DREAM2 Challenges. *Ann. N Y Acad. Sci.*, **1158**, 159–195.

Waaijenborg,S. *et al.* (2008) Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article3.

Wright,F.A. *et al.* (2007) Simulating association studies: a data-based resampling method for candidate regions or whole genome scans. *Bioinformatics*, **23**, 2581–2588.

Zhang,W. *et al.* (2010) A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.*, **6**, e1000642.

Zhu,J. *et al.* (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.*, **105**, 363–374.