StatSeq Systems Genetics Benchmark

Andrea Pinna[1], Nicola Soranzo[1], Ina Hoeschele[2,3], 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, USA
3. Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0477, USA

1 Introduction

The StatSeq benchmark dataset is meant to be used for training and evaluating algorithms and techniques for the inference of networks from systems genetics data. The goal is to comprehend which methodology has the best overall inferring performance, and which eventually performs better under particular conditions (i.e. population size, large or small marker distances, high or low heritability, network size).

This short document describes how the data have been generated through SysGenSIM. Detailed information is provided about the construction of the gene networks, the simulation of the genotype and of the gene expression, and the submission and evaluation of the predictions.

2 In Silico Networks

The whole compendium is a collection of 72 datasets computed on a total of 9 gene networks of different size. In particular, 3 networks have been generated, respectively, for each of the following sizes: $n = \{100, 1000, 5000\}$. The datasets related to networks of size $n = 100$ are intended for testing purposes only, and their predictions will not be evaluated.

The networks share the following characteristics:

- Exponential in- and power law out-degree distributions (EIPO).
- Average node degree[1] about $K = 6$.
- Largest strongly connected components of size at least 20% of the network nodes (for $n = 100$), at least 15% (for $n = 1000$), and at least 10% (for $n = 5000$).

[1]The average number of both ingoing and outgoing edges for a node: $K = K_{in} + K_{out}$.

Details are summarized in Table 1, where for each network is shown the number of edges, the size of the largest strongly connected component (LSCC), the size of the in- and of the out-components[2], the size of tendrils[3] and tubes[4]. Network 1000-2 has one isolated node.

Table 1: Topological information on the in silico networks.

Size	Network	Edges	LSCC	In	Out	Tendrils	Tubes
100	1	285	21	25	18	30	6
100	2	304	26	7	63	4	0
100	3	296	34	9	57	0	0
1000	1	3149	165	71	660	96	8
1000	2	2881	162	58	648	108	23
1000	3	3038	150	106	563	158	23
5000	1	14678	528	224	3498	657	93
5000	2	15672	526	233	3580	595	66
5000	3	15270	598	231	3660	480	31

3 Simulation of Datasets

Datasets have been generated by means of SysGenSIM 1.0.2, version released on May 8$^{\text{th}}$, 2012. More information are available in the SysGenSIM manual. Each network has been simulated with 8 different parameter settings, i.e. by combining marker distances d (small and large), median heritability H (high $\simeq 0.8$ and low $\simeq 0.4$) and population size m (small and large). Simulations have been run with the settings combined as described in Table 2. By keeping

Table 2: SysGenSIM settings to generate the datasets.

Network Topology	EIPO		Network Size	$\{100, 1000, 5000\}$
Sign Assignment	Node-wise		Sign Probability	0.5
Average Node Degree	6		Marker Positions	Generate
Gene Positions	at Markers		Mapping Function	Haldane
RIL Type	Selfing		Chromosome	$\{5, 25, 25\}$
Markers per Chromosome	N($\{20, 40, 200\}$,2)		Distances	$\{\text{N}(1, 0.2), \text{N}(5, 1)\}$
Cis-Effect %	25		Genotyping Error %	5
Z Lower	0.5		Z Upper	0.8
Basal Transcription	Constant		Parameters	1
Interaction Strength	Constant		Parameters	1
Cooperativity Coefficient	Gamma		Parameters	[1,1.67]
Basal Degradation	Constant		Parameters	1
Transcription Biological Variance	Gaussian		Parameters	$[1, \{0.1, 0.25\}]$
Degradation Biological Variance	Gaussian		Parameters	$[1, \{0.1, 0.25\}]$
Expression Measurement Noise	Gaussian		Parameters	$[1, 0.1]$
Phenotype Nodes	0		Genotype Matrix	✓
Gene Expression Matrix	✓		Edge List	✓
Node Degree Distributions	✓		Heritability Distribution	✓
Population	$\{300, 900\}$		Experiments	1

[2]Respectively, all the nodes from which the LSCC is reachable and that are not reachable from the LSCC, and all the nodes reachable from the LSCC but from which the LSCC cannot be reached.

[3]Nodes from which the LSCC cannot be reached, and that cannot be reached from the LSCC itself.

[4]Nodes connecting the in- to the out-component, and not belonging to the LSCC.

most of the parameters fixed, each dataset has been simulated with the setting configurations summarized in Table 3.

Researchers are provided, for each of the 72 datasets, with the following files:

Edge list. A signed list of network edges.

Gene expression matrix. A $n \times m$ matrix of gene expression measurements. Entry (i, j) is the simulated steady state expression value of gene i in individual j.

Genotype matrix. A $n \times m$ matrix of genotype values $\{0, 1\}$. Entry (i, j) is the genotype value of gene i in individual j.

File names are structured as:

```
Dataset_i_Network_n-j_Configuration_c_D.tsv
```

where `i` $= \{1, \ldots, 72\}$, `n` $= \{100, 1000, 5000\}$, `j` $= \{1, 2, 3\}$, `c` $= \{1, \ldots, 8\}$, `D` $=$ {edge_list, gene_expression_matrix, genotype_matrix}. As an example, the file named

```
Dataset_47_Network_1000-3_Configuration_7_genotype_matrix.tsv
```

contains the genotype matrix from dataset 47, which consists of a population of 300 individuals simulated from the third 1000-gene network with reduced marker distances and low heritability.

Table 3: Setting configurations for each network.

Configuration	Marker Distance	Biological Variance	Heritability	Population Size
1	N(5,1)	N(1,0.1)	High	300
2	N(5,1)	N(1,0.1)	High	900
3	N(5,1)	N(1,0.25)	Low	300
4	N(5,1)	N(1,0.25)	Low	900
5	N(1,0.1)	N(1,0.1)	High	300
6	N(1,0.1)	N(1,0.1)	High	900
7	N(1,0.1)	N(1,0.25)	Low	300
8	N(1,0.1)	N(1,0.25)	Low	900

Datasets are divided in archives by network size, and are available for download here:

```
http://resources.bioinformatica.crs4.it/sysgensim_dataset/...
        .../StatSeq_Datasets_Size100.zip
        .../StatSeq_Datasets_Size1000.zip
        .../StatSeq_Datasets_Size5000.zip
```

A file containing the median value of the heritability for all the datasets is available at:

```
http://sysgensim.sourceforge.net/StatSeq_Heritability.tsv
```

4 Evaluation of Predictions

An evaluation script is provided for assessing the goodness of the inference algorithms. Predictions can be saved in MAT-files or in tab-separated-value files, which must be called as:

```
Dataset_i_Predictions.{mat,tsv}
```

where `i` is the number identifying the dataset (i.e. `i` = 31). Predictions can be represented as:

- square $(n \times n)$ matrices,
- *source-target-score* lists, i.e. $(n^2 \times 3)$ matrices.

In case of MAT-file, predictions must be saved in the variable `predictions`. Else, in case of tab-separated-value files, the predictions must be entered with no headers nor comments.

In both cases, predictions can be formatted as square matrices or as edge-score lists. In case of square $(n \times n)$ matrices, each entry (i, j) corresponds to the confidence score assigned to edge (i, j), and the main diagonal entries (i, i) must be equal to zero.

In case of *source-target-score* lists, i.e. $(n^2 \times 3)$ matrices sorted in descending order by the third column values, for each row the first entry is the source node i, the second entry is the target node j, and the third entry is the confidence score associated to the edge (i, j). All scores must be in the $[0, 1]$ range. An example of prediction list is the following:

```
1289 3115 0.9374216
 741 4133 0.8512534
3701  991 0.7831446
.... .... .........
  29 4984 0.0000001
   1    1 0.0000000
.... .... .........
5000 5000 0.0000000
```

The only input for the evaluation script is the path to the folder containing the prediction files. The script will work even in the case of incomplete predictions, i.e. the researcher is willing to evaluate the prediction for a number of datasets smaller than 72.

The evaluation script (complete with gold standard networks) is available for download here:

```
http://resources.bioinformatica.crs4.it/sysgensim_dataset/...
        .../StatSeq_EvaluationScript.zip
```

Researchers are strongly encouraged to submit their predictions, for evaluation and comparison with other algorithms, to the FTP server:

```
ftp://ftp.bioinformatica.crs4.it
```

Please contact Alberto de la Fuente (`alf@crs4.it`) to receive username and password to upload the prediction files, preferably as a compressed archive.