Systems Biology and Evolutionary Computation

GECCO Tutorial
July 8, 2006

Stefan Bleuler, Philip Zimmermann, Eckart Zitzler

ETH Zurich, Switzerland
Goals and Schedule

Questions
• What is Systems Biology?
• What are the basic types of biological experiments and measurements?
• What are the computational issues in Systems Biology?
• What are examples of successful application of EC in Systems Biology?

Schedule
• 50 min  Part 1: Introduction to Systems Biology.
• 10 min  break
• 50 min  Part 2: Computational Issues in Systems Biology.
1. Introduction to Systems Biology
Reverse Engineering Problem
Reverse Engineering Problem
Reverse Engineering Problem II
A Definition: Understanding of network behavior using computational approaches tightly linked to experiments. (M. Cassman)

Goals:
- system level understanding
- simulators for cells and organisms
- personalized, predictive and preventive medicine

Methods:
- experiment: mostly high throughput
- models
- computational analysis

Key Idea: capture emergent properties
Compared to Classical Biology

**Classical Biology**

- focus on single elements (gene, protein, pathway)
- focus on building blocks
- bottom up
- interpretation

- Examples:
  - structure determination of a protein
  - study effects of knockout on one pathway

**Systems Biology**

- focus on all elements (genome, proteome, metabolome)
- focus on interactions
- top down
- simulation

- Examples:
  - module identification
  - robustness analysis of genetic networks
Some cell biology…
Cellular Compartments of a Plant Cell

NUCLEUS
- Chromatin
- Nucleus
- Nuclear envelope

CENTROSOME

ROUGH ENDOPLASMIC RETICULUM

SMOOTH ENDOPLASMIC RETICULUM

RIBOSOMES

GOLGI APPARATUS

CENTRAL VACUOLE
- Tonoplast

MICROFILAMENTS
- Intermediate filaments
- Microtubules

CYTOSKELETON

MITOCHONDRION

PEROXISOME

PLASMA MEMBRANE

CELL WALL

WALL OF ADJACENT CELL

CHLOROPLAST

PLASMODESMATA

NOT IN PLANT CELLS:
- Lysosomes
- Centrioles
- Flagella (in some plant sperm)
Cellular Compartments of a Human Cell
Central Dogma of Molecular Biology

DNA → Transcription → mRNA → Translation → Protein
Central Dogma of Molecular Biology

![Diagram of the Central Dogma](image)
Chromosome – chromatin – nucleosome – gene

Dividing cell

Non-dividing cell
Genes

DNA

RNA

PROTEIN

Sugar-phosphate “backbone”
Hydrogen bonds between nitrogenous bases

OH
3’ end

5’ end

Phosphodiester bond
Transcription

- The process by which a molecule of DNA is copied into a complementary strand of RNA.
- 1 Strand DNA → 2 Strands RNA
- RNA Polymerase
Messenger RNA

- DNA
- RNA
- PROTEIN

- Ribosome binding signal
- Initiation codon
- Start of transcript
- CAU
- AGGAGGU
- AUG

- Hairpin loop
- Stop codon
- UAA
- UUUUU

- End of message
Transcription
Translation

DNA → RNA → PROTEIN

Polypeptide: Amino acid 1 → Amino acid 2 → Amino acid 3
Adaptors
Nucleic acid: Codon 1, Codon 2, Codon 3
Translation

DNA

RNA

PROTEIN

Growing polypeptide chain

Amino acid

tRNA

Ribosome

mRNA

5' 3'
In the cell, proteins can:

- work as channels
- span the membrane for transport, signalling, …
- perform enzymatic reactions
- serve as ligands
- or can have many other functions
Protein complexes are like factories: efficiency is increased by proximity, interaction, and chain-like production setup.
high throughput technologies
Types of Experiments

What to target.
- different developmental stages
- different organs

How to perturb an organism.
- knock-out / knock-down
- over-expressions
- stimulus
- disease

} change expression of specific genes

How to measure.
- steady state measurement
- time course measurements
Typical Workflow of an -omics Experiment

1. Destroy cells from experiment sample
2. Collect cell content
3. Filter and/or purify cell content
4. Perform high-throughput measurement
5. Process and store data
6. Analyze data and interpret results
DNA Microarrays

cDNA or GST Arrays

Oligonucleotide Arrays

Dual dye (red/green)
(Many companies)

Single dye (Affymetrix, Nimblegen)
Dual or single dye (Agilent)
DNA Microarrays: Oligonucleotide Arrays

GeneChip Probe Array

Single stranded, labeled RNA target

1.28 cm

18 µm

Scanned image

10^6-10^7 copies of a specific oligonucleotide probe per feature

>1,000,000 different probes

Oligonucleotide probe
DNA microarrays: cDNA arrays

Treatment

Control
Protein Microarrays

**a**
- Antibody
- Antigen
- Aptamer
- Allergen

Serum probes
Cell lysates
Living cells

Protein expression level
Protein profiling
Diagnostics

**b**
- Protein
- Peptide

Protein probes
Nucleic acid probes
Drug probes
Enzymes

Protein binding properties
Pathway building
Drug discovery
Post-translational modification
Proteomics

Proteomics = study of the protein repertoire expressed in the cell

Measurements

• protein expression levels (quantitative and qualitative)
• localization
• protein interaction

Protein interactions elucidate…

• pair-wise interactions
• protein complexes
Shotgun Proteomics

lyse cells

Mixture of 1000’s of peptides

digest with trypsin

2-D LC-MS/MS

RPLC-MS/MS

LC-IMS-MS/MS

Database searching - matching MS/MS data with peptide sequence or bioinformatics de novo sequencing of proteins
Proteomics: Mass-Spectrometry Analysis
Tandem Affinity Purification

This molecule binds the ‘tag’.  

Protein of interest (e.g. HA/GST/His)
Tandem Affinity Purification (TAP)

Untagged proteins go through fastest (flow-through).

Tagged proteins bind to beads.
TAP
Yeast two-hybrid

Source: Griffiths et. al. Modern Genetic Analysis.
Metabolomics

• Metabolomic methods:
  
  – Chromatography
  
  – Mass spectrometry (MS)
  
  – Nuclear magnetic resonance (NMR)
Synthetic Lethal Interactions
OMICS…

Gene - Genome - Genomics
Protein - Proteome - Proteomics
Metabolite - Metabolome - Metabolomics
Complex Systems
Formalized Biological Knowledge

- scientific literature
- functional annotation
  - Gene Ontology
  - swissprot
- pathway databases
  - KEGG
- phenotype and patient information
2. Computational Issues in Systems Biology
Computational Challenges

- Experimental design
- Data preprocessing
- Data visualization
- Simulators
- Network inference
- Module identification
- Classification
- Structure prediction
- Text mining
Classification of Tumor Samples – Problem

Goals
- discrimination between classes
- feature extraction

Data
- mostly gene expression
- proteomics
- known outcome

Challenges
- noisy data
- few samples, high dimensionality
- overfitting
- multiple testing
Classification of Tumor Samples – General Approach

**Ingredients**
- gene set selection
- classifier
- objective function
- optimizer

**Fighting Overfitting**
- cross validation in objective function
- keep models small
Classification of Tumor Sample – EC approaches

Optimization Approaches

- genetic programming (GP) [1, 6, 7]
- simulated annealing [4]
- multiobjective evolutionary algorithm (including size) [3, 5]

Classification of Tumor Samples - Moore et al. [1]

Individual
- real valued function \( f \) of gene expression
- represented as GP tree

Classifier
- \( f > \) median of all function values

Objective Function
- classification error

Optimizer
- parallel GP

Results
- data: AML/ALL (Golub et al.) two class problem
- two nearly perfect predictors:
  - \( X_{1835} + X_{2546} \)
  - \( X_{2555} \times (X_{1153} + X_{2289} + X_{3193}) \)
Classification of Tumor Samples - Deutsch [4]

**Individual**
- set of predictive genes
- represented as list

**Classifier**
- k-nearest neighbor (k = 1)

**Objective Function**
- weighted sum of LOOCV accuracy and clustering score

**Optimizer**
- variant of simulated annealing (replication algorithm)
- mutation: add or remove one gene

**Results**
- data: multiple data sets (incl. one with more than 2 classes)
- results: smaller gene sets and good classification
Classification of Tumor Samples – Liu et al. [3]

**Individual**
- set of predictive genes
- represented as bit string

**Classifier**
- normalized distance to class mean

**Objective Function**
1. gene set size
2. missclassification rate
3. difference of missclassification rates

**Optimizer**
- multiobjective EA
- called replication algorithm

**Results**
- data: Leukemia, Lymphoma and Colon cancer data sets
- results: many diverse and small gene sets
Classification of Tumor Samples – Langdon et al. [7]

Individual
- sum S of 5 real valued function of expression values
- represented as 5 GP trees

Classifier
- $S > 0$

Objective Function
- mean of LOOCV accuracies for both classes

Optimizer
- GP

Results
- data: Central nervous system embryonal tumors [Pomeroy et al. 2002]
- results: good classification, surprisingly small gene sets
Module Identification

Goal

high throughput data  functional gene groups

Approaches

• guilt by association
• clustering, biclustering
• integration with additional data, e.g., promoter elements

Challenges

• huge search space
• data integration
Clustering of Gene Expression Data

Inputs
- gene expression data
- number of clusters

Clustering algorithms ...
- group similar things.
- partition the matrix.
- use all measurements.
Clustering with EC – Falkenauer et al. [9]

**Individual**
- clustering = partitioning of input matrix
- specific representation

**Objective Function**
- total variance within clusters

**Optimizer**
- Grouping Genetic Algorithm

**Results**
- data: 3 different gene expression data sets
- results: much better than k-means (which uses the same objective function)

Individual

- clustering = partitioning of input matrix
- locus based adjacency representation

Objective Function

1. total deviation from cluster means
2. total connectivity (high if neighbors are not in the same cluster)

Optimizer

- multiobjective EA (PESA-II)

Results

- good performance compared to k-means and average-linkage hierarchical clustering algorithms
- automatic determination of the number of clusters

Drawbacks of Standard Clustering

- A gene cannot be in two clusters
- Each gene is assigned to a cluster
- Local patterns are missed

New problem formulation needed.
Biclustering

Goal

- find subsets of genes – subsets of conditions
- may overlap
- two objectives: size and similarity

Existing Algorithms

- definition of similarity
- number of biclusters
- search strategy
EC for Biclustering – Bleuler et al. [11]

Approach
- existing algorithms as local search
- EA as global search
- systematic sampling of the search space
- applicable to many existing algorithms

Basics
- individual = submatrix
- binary encoding
- independent bit mutation
- uniform crossover
- local search
- tournament selection ($t \in \{3, 20\}$)
- 100 individuals, 50 generations

EC for Biclustering – Bleuler et al. [11]

**Initialization**
- set each bit to 1 with \( p = 0.5 \)?
  - biclusters have similar size
- better: distribute bicluster sizes

**After the local search...**
- update (Lamarckian evolution)
- don’t update (Baldwinian evolution)

**Diversity Maintenance**
- \( N \) biclusters in one run
- optimize coverage
- select individual with most new area
EC for Biclustering – Bleuler et al. [11]
Gene Expression is not Enough!

- DNA
- RNA
- Protein

AT4G16750
AT5G21150
AT5G19920
AT4G25660
AT4G10480
AT4G03060
AT3G22750
AT2G38460
AT2G35000
AT4G16750
AT5G21150
AT5G19920
AT3G02670
AT3G05530
EC for Data Integration - Speer et al. [12]

Goal
• clustering of gene expression data and Gene Ontology graph

Individual
• clustering = partitioning of input matrix
• representation based on minimum spanning tree
• represented as n-1 bits determining whether to cut the MST at edge i.

Objective Function
• weighted sum of distance on gene expression and distance on Gene Ontology graph

Optimizer
• EA with local search

Results
• data: gene expression data from human fibroblast and GO
• results: some clusters more gene expression oriented others more GO oriented

[12] N. Speer et al., A Memetic Co-Clustering Algorithm for Gene Expression Profiles and Biological Annotation, CEC, 2004
EC for Data Integration - Bleuler et al. (work in progress)

gene expression

protein-protein interactions

distance PPI

distance GE
EC for Data Integration - Bleuler et al. (work in progress)

multiobjective evolutionary algorithm

distance PPI

distance GE
EC for Data Integration - Bleuler et al. (work in progress)
EC for Data Integration - Bleuler et al. (work in progress)

Similarity of the Clusters on the Front

gene expression – gene expression
gene expression - PPI
Network Inference

Goal

High throughput data

Approaches

- network analysis (structure, robustness, etc.)
- inference of network topology (Bayesian networks, Gaussian Graphical Models, etc.)
- inference of network function (Boolean networks, differential equations, etc.)

Challenges

- underdetermined problem
- noisy data
- experiment design
EC for Network Inference

Network Models

- S-systems [3, 6, 7]
- Petri nets [4]
- electronic circuit [1]
- differential equations [3]
- real valued matrix [5]

References:

[18] S. Kimura et al., Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm, Bioinformatics, 2005
EC for Network Inference – Koza et al. [13]

Individual
- chemical reaction network
- modeled as electronic circuit
- represented as GP tree

Objective Function
- comparing predicted and measured concentration of end product
- sum of absolute differences for test cases
- evaluated using SPICE

Optimizer
- GP
- popsize 100,000

Results
- input: E-cell simulation of phospholipid cycle (4 reactions) and synthesis of ketone bodies (3 reactions)
- good recovery of network topology and reaction rates
EC for Network Inference – Cho et al. [19]

**Individual**
- biochemical reaction network or gene regulatory network
- modeled as S-tree
- represented as GP tree

**Objective Function**
- comparing prediction to measurement on all time points and all substances
- sum of relative squared errors

**Optimizer**
- GP
- local hill climbing

**Results**
- 1. input: simulation of artificial networks modeled as S-systems
- 1. results: good recovery of network topology and parameters
- 2. input: gene expression from SOS DNA repair in *E. coli* (6 genes)
- 2. results: all but one known interaction recovered (in 35 h).
3. Status Quo and Future Trends
Status Quo

Advantages of EC Approach
- flexible
- global search method
- multiobjective

Open Problems
- benchmark problems missing
- little comparison with non-EA methods
- no common methodology
Future Trends

Biology and Measurements

- more data (more genomes, transcriptomes and proteomes)
- more data types (tiling arrays, synthetic lethal, etc.)
- more specific measurements (towards single cell analytics)
- more formalized information about experiments

Computational

Data Integration of ...

- different qualities (accuracy)
- different data types (proteomics, metabolomics, etc.)
- different scales
- different precision (qualitative vs. quantitative)