Discovering Genetic Regulatory Networks

Bioinformatics Group, Delft University of Technology

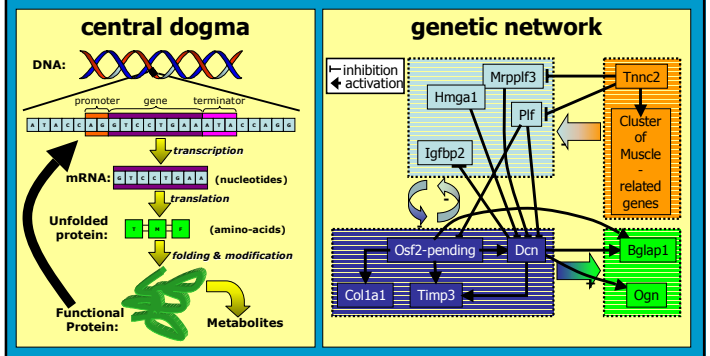
Marcel Reinders
Eugene van Someren
Rogier van Berlo
Lodewyk Wessels

Marcel Reinders
June 7-8, 2004

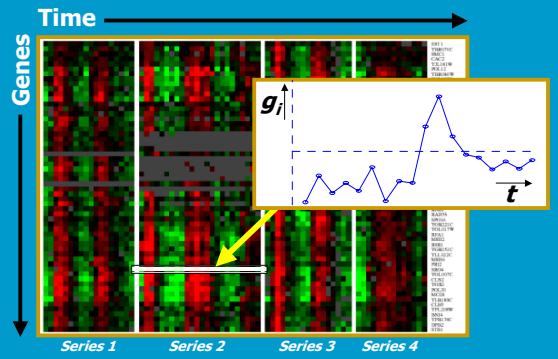
Outline

- Genetic network modeling
- Linear model
- Constrained modeling
  - Remove redundancy
  - Increase sparseness
  - Increase robustness
- Comparison using artificial example
- Preliminary study

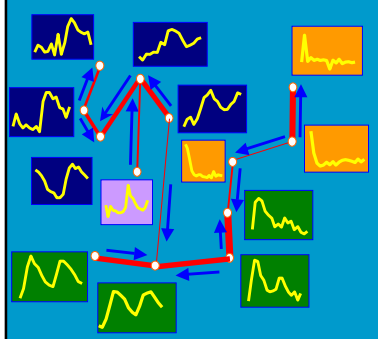
The central Dogma ↔ Genetic Network



(mRNA) Expression data set



Genetic Network Modeling



- **Clustering (MST)**
  - Group genes into functional units based on correlations in expression
  - Expresses *co-regulation* and *not causality*
- **System identification approach**
  - Build dynamic models for gene regulatory networks
  - Estimate model from genome-wide scale expression data
- **Infer "causal relationships" between genes from microarray data**

Regression networks

- **Mathematical model**
  - Allows to make prediction under different conditions
  - Estimate model parameters by fitting predicted and measured expression profiles
- **Current modeling techniques**
  - (Dynamic) Bayesian models (Murphy, Pe'er, …, *van Berlo*)
  - Non-linear models (Weaver, Wahde, …, *Au Yeung, van Roon*)
  - Linear models (*Someren*, …, D'haeseleer)
- **Linear models**
  - Continuous valued, analytical solutions exists
  - Allows for (math.) incorporation (biologically motivated) constraints
  - Allows to study small sample size problem
  - Gained knowledge re-usable for more complex models

Outline

- **Genetic network modeling**
- **Linear model**
- **Constrained modeling**
  - Remove redundancy
  - Increase sparseness
  - Increase robustness
- **Comparison using artificial example**
- **Preliminary study**

Linear model

model

$$\hat{g}_i(t+1) = \sum_{j=1}^{N} w_{ij} g_j(t) \qquad t = 1,\ldots,T-1$$

wiring diagram

w_{ij} : $w_{ij}=0$: no interaction between g_i and g_j
$w_{ij}>0$: g_j is activating g_i
$w_{ij}<0$: g_j is repressing g_i

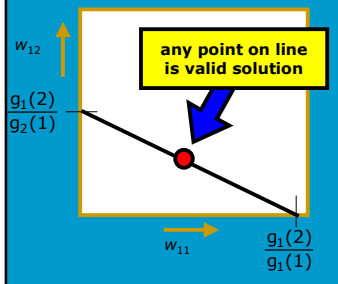
error model

$$E^S = \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(g_i(t+1) - \hat{g}_i(t+1)\right)^2$$

$$E^S = \sum_{i=1}^{N}\sum_{t=1}^{T-1}\left(g_i(t+1) - \sum_{j=1}^{N}\left(w_{ij} g_j(t)\right)\right)^2$$

finding w_{ij} by minimizing E^S wrt w_{ij}

Limited measurements



- **Thousands of genes (*N*) and tens of microarray measurements (*T*)**
- **Small sample size problem**
- **For example: 2 Genes, 2 arrays**
  For both genes only one equation
  Error can be made zero by:
  $$g_1(2) = \hat{g}_1(2) = w_{11}g_1(1) + w_{12}g_2(1)$$
  Solve weights w_{ij}
  $$w_{12} = \frac{g_1(2)}{g_2(1)} - w_{11}\frac{g_1(1)}{g_2(1)}$$
- **A SET of solutions gives PERFECT prediction!**

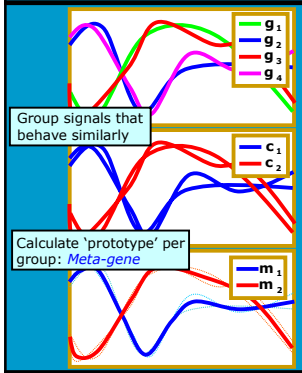
Dealing with limited measurements?

- **Remove redundancy (ambiguity)**
  Reducing number of genes by exploiting their co-regulation
- **Increase sparseness (Arnone: limited connectivity)**
  Reducing number of weights by allowing only a few non-zero weights (a few incoming connections)
- **Increase robustness (regularization)**
  Inference should be somewhat insensitive to small amounts of noise

Outline

- **Genetic network modeling**
- **Linear model**
- **Constrained modeling**
  - **Remove redundancy**
  - Increase sparseness
  - Increase robustness
- **Comparison using artificial example**
- **Preliminary study**

Remove redundancy



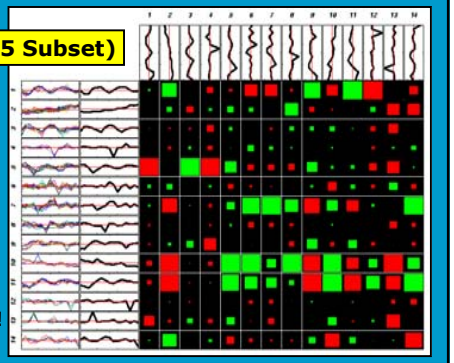
- **Redundancy**
  - Co-regulation
  - Ambiguity due to noise in the data
- **Remove redundancy**
  - Cluster genes
  - Construct "meta-genes"
- **Model similar (grouped) signals in same way**
  - Find regulation between "meta-genes"

Appeared in: ISMB 2000

Yeast Cell Cycle Data

(Spellman; CDC 15 Subset)

- **Threshold at 2**
  113 genes
- **Mean-Variance Normalization**
- **14 Prototypes !!!**
- **FULL CONNECTIVITY !**



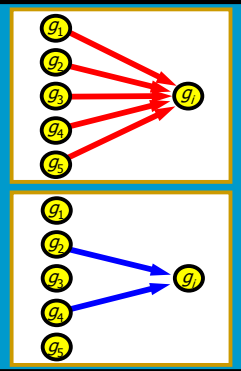
(I,C)T Appeared in: ISMB 2000 TU Delft

Outline

- **Genetic network modeling**
- **Linear model**
- **Constrained modeling**
  - Remove redundancy
  - **Increase sparseness**
  - Increase robustness
- **Comparison using artificial example**
- **Preliminary study**

(I,C)T TU Delft

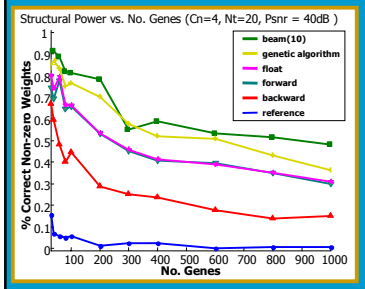
Increase sparseness



- **Current model**
  Gene is possibly influenced by **ALL** genes
- **Practice**
  Gene is influenced only by **limited** number of genes
  (6-8 regulatory sites (Arnone))
  But we don't know which ones!
- **Greedy search approach**
  - Find gene that best predicts g_j
  - Extend current set with that gene that when included gives the best prediction
  - Repeat until convergence

(I,C)T Appeared in: ICSB 2001 TU Delft

Other search techniques



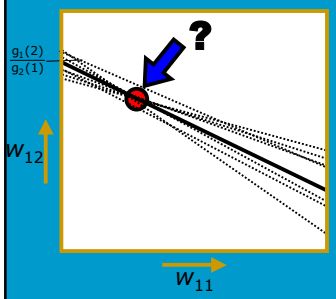
- **Greedy search**
  - Forward (previous method)
  - Backward
- **Beam search**
  - Expand only the *N* most promising solutions
  - *N*=1, equal to greedy search
- **Floating search**
  - Greedy expand but allow to withdraw previous made choices
- **Stochastic search**
  - Genetic algorithm
- **Comparison**
  - Beam search outperforms others

(I,C)T Appeared in: ICSB 2001 TU Delft

Outline

- Genetic network modeling
- Linear model
- Constrained modeling
  - Remove redundancy
  - Increase sparseness
  - **Increase robustness**
- Comparison using artificial example
- Preliminary study

Increase robustness



- Independent noise on each measurement

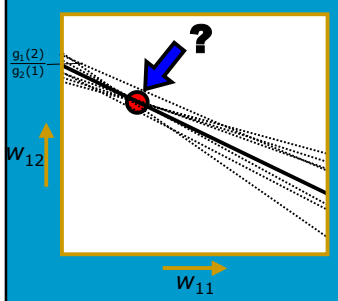
$$g_i^{\sigma}(t) = g_i(t) + \varepsilon_i(t)$$

- Now solving weights w_{ij}

$$w_{12} = \underbrace{\left(\frac{g_1(2)+\varepsilon_1(2)}{g_2(1)+\varepsilon_2(1)}\right)}_{\text{different offset}} - w_{11}\underbrace{\left(\frac{g_1(1)+\varepsilon_1(1)}{g_2(1)+\varepsilon_2(1)}\right)}_{\text{different direction}}$$

- Different sets of solutions !
- Solution that "fits" all solutions is most robust

Noise injection



- From initial dataset G_0 create K new datasets G_k in which each measurement is slightly perturbed

$$G_k = G_0 + \varepsilon \quad \text{with} \quad \varepsilon_{ij} = N(0,\sigma^2)$$

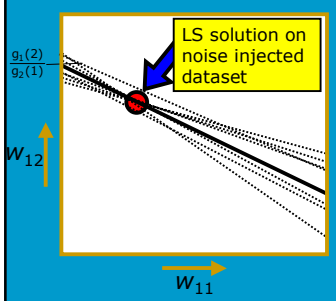
- Divide into input X and target Y

$$X = \{G_k(1:T-1), \forall k\}$$
$$Y = \{G_k(2:T), \forall k\}$$

- Problem (re)formulation

$$\hat{g}_i(t+1) = w_{i1}g_1(t) + \ldots + w_{iN}g_N(t)$$
$$\hat{Y} = XW$$

Noise injection



- Linear model

$$\hat{Y} = XW$$

- Find W by minimizing squared error (Least Squares solution)

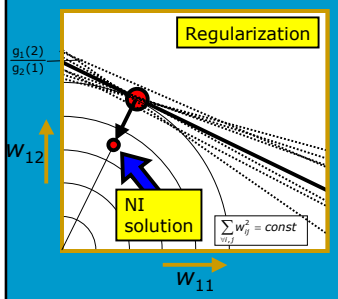
$$E^S = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\left(y_i(t) - \hat{y}_i(t)\right)^2$$

- If $K*T > N$ then LS solution exists

$$W^{LS} = (X^T X)^{-1} X^T Y$$

- Noise strength σ^2 determines robustness !

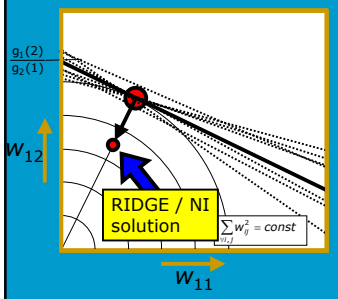
Noise injection



- **Bishop**
  Noise injection = Regularization
  $E^{NI} = E^S + \eta^2 E^{REG}$
- **Regularization term (Tikhonov)**
  $$E^{REG} \approx \frac{1}{2(T-1)} \sum_{t=1}^{T-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\frac{\partial y_i(t)}{\partial x_j(t)} \right)$$
- **Linear model**
  $$y_i(t) = \sum_j w_{ij} x_j(t) \quad \rightarrow \quad \frac{\partial y_i(t)}{\partial x_j(t)} = w_{ij}$$
  $$E^{REG} \approx \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}^2$$
- **Importance of regularization controlled by η^2 *same role as* σ^2**

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

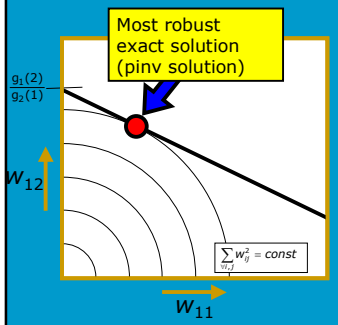
Ridge regression



- **Equivalent to ridge regression**
  $$E^{RIDGE} = E^S + \lambda \sum_{\forall i,j} w_{ij}^2$$
- **Analytical solution exists**
  $$\boldsymbol{W}^{RIDGE} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$
  Positive constant on diagonal of X^TX prevents singularity
- **Small λ: $W^{RIDGE} = W^{LS}$**

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

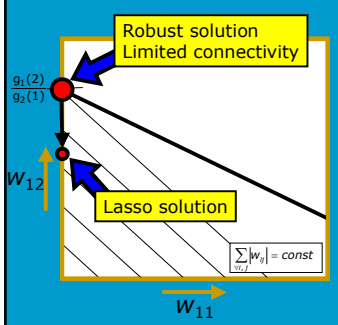
Moore-penrose solution



- **Moore-penrose solution**
  $$E^{PINV} = \left(E + \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}^2 \right)_{\lambda \to 0}$$
- **In other words: select exact solution which is most robust, ie**
  $$\sum_{\forall i,j} w_{ij}^2 \text{ is minimal}$$
- **Still an exact solution !!!**

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

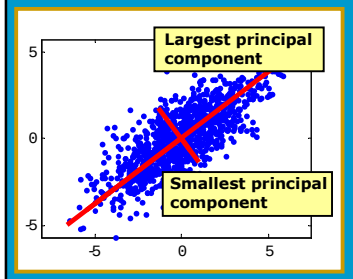
Lasso regression



- **Some solutions have limited connectivity**
- **Choose solution**
  - Minimal connectivity
  - Most robust against noise
- **Lasso regression**
  $$E^{LASSO} = E^{LS} + \mu \sum_{\forall i,j} |w_{ij}|$$
- **Small μ: $W^{LASSO} = W^{LS}$**
- **Large μ : Shrinks coefficients to zero!**

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

PCA equivalence



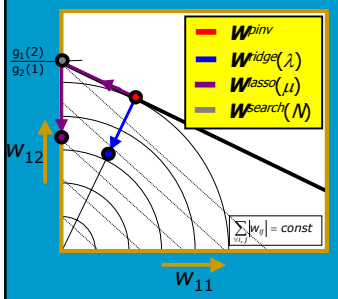
- **SVD analysis reveals**

$$\boldsymbol{X}\boldsymbol{w}^{RIDGE} = \sum_{k=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j^T \boldsymbol{y}$$

$\boldsymbol{u}_j$: principal components $\boldsymbol{X}$
d_j: amount of variance

- **Ridge regression shrinks directions with smallest variance most**

Comparison regularization



- **Methods**
  - **PINV**
    One solution
  - **LASSO/RIDGE**
    Change tuning parameter
  - **Search**
    Select best input gene
  - **PCA**
    Select inputs in PCA mapped space
- **RIDGE/PCA**
  Tend to behave similar
- **LASSO**
  "soft-thresholding" weights

Outline

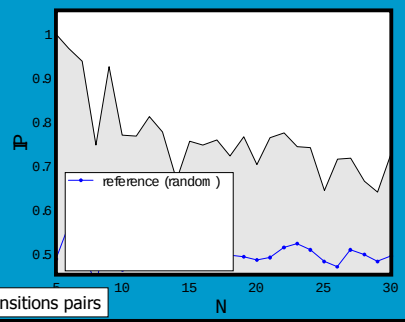
- **Genetic network modeling**
- **Linear model**
- **Constrained modeling**
  - Remove redundancy
  - Increase sparseness
  - Increase robustness
- **Comparison using artificial example**
- **Preliminary study**

Artificial example: Set-up

- **Generation of the data**
  - For varying number of genes (x-axis of the plot)
  - Fixed connectivity (C=4)
  - Random generation of W
  - For each gene T=17 time points generated from random initial state
  - Noise added to these time points (40dB PSNR)
- **Measured**
  - Inferential power: Correlation between true $\boldsymbol{W}$ and estimated $\boldsymbol{W}$
  - Averages of 40 repetitions of the experiment
- **Parameter setting methods done using leave-one-out-procedure**

Artificial example: Results

$IP = \frac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$



good

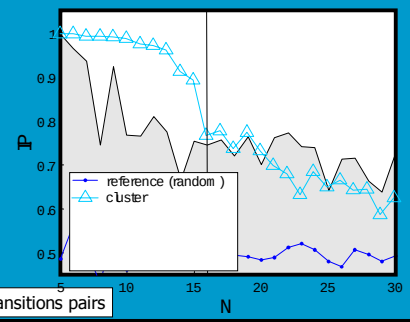
bad

Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Artificial example: Results

$IP = \frac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$



good

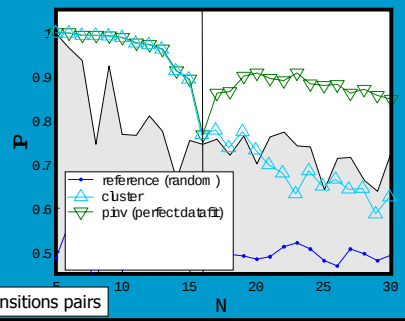
bad

Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Artificial example: Results

$IP = \frac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$



good

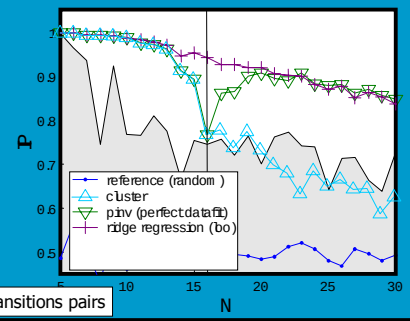
bad

Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Artificial example: Results

$IP = \frac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$



good

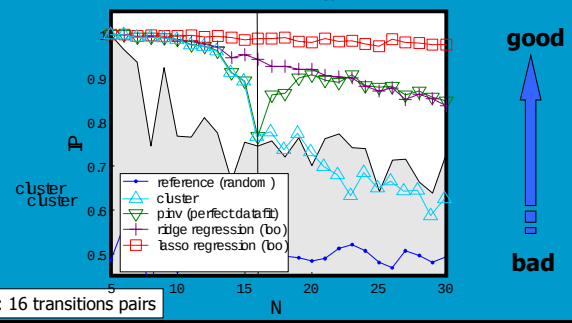
bad

Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Artificial example: Results

$$IP = \tfrac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$$

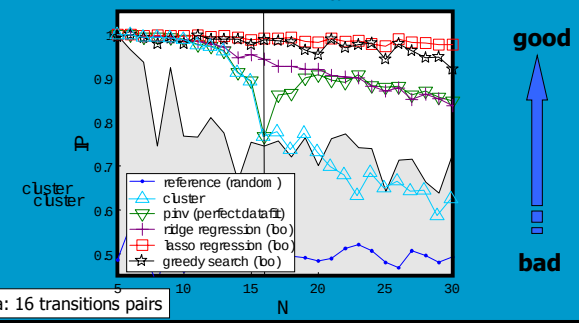


Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Artificial example: Results

$$IP = \tfrac{1}{2}(1 + \rho(\boldsymbol{W}_{real}, \hat{\boldsymbol{W}}))$$



Data: 16 transitions pairs

(I,C)T Appeared in: Computational and Statistical Approaches to Genomics, 2002 TU Delft

Outline

- **Genetic network modeling**
- **Linear model**
- **Constrained modeling**
  - Remove redundancy
  - Increase sparseness
  - Increase robustness
- **Comparison using artificial example**
- **Preliminary study**

(I,C)T TU Delft

Some general conclusions

- **As a result of the small sample size problem**
  - Studying models under noisy conditions is essential (PSB, BIOS 2001, results not shown)
  - Constraining models is necessary to be able to find 'sensible' solutions
  - Need to be careful when using more complex models (since they suffer more from the small sample size problem)
- **When properly constrained suggestions for new relationships between genes can be made**
- **Further improve modeling by exploiting additional constraints**

(I,C)T TU Delft

Future directions

- **Dealing with small sample size problem**
  - How to cope with pseudo structure
  - Experimental design: Predicting most valuable next experiment (significance analysis: Which predicted link should be examined first)
- **Integrative approach, i.e How to integrate:**
  - Different experiments: E.g. knock-out and time series, data generated in different labs
  - Different types of data: E.g. sequence data, protein-protein interaction, metabolite concentrations
  - Data bases information: How to bias solutions towards existing knowledge in databases
  - Data from different organisms: E.g. conserved pathways

(I,C)^T TU Delft