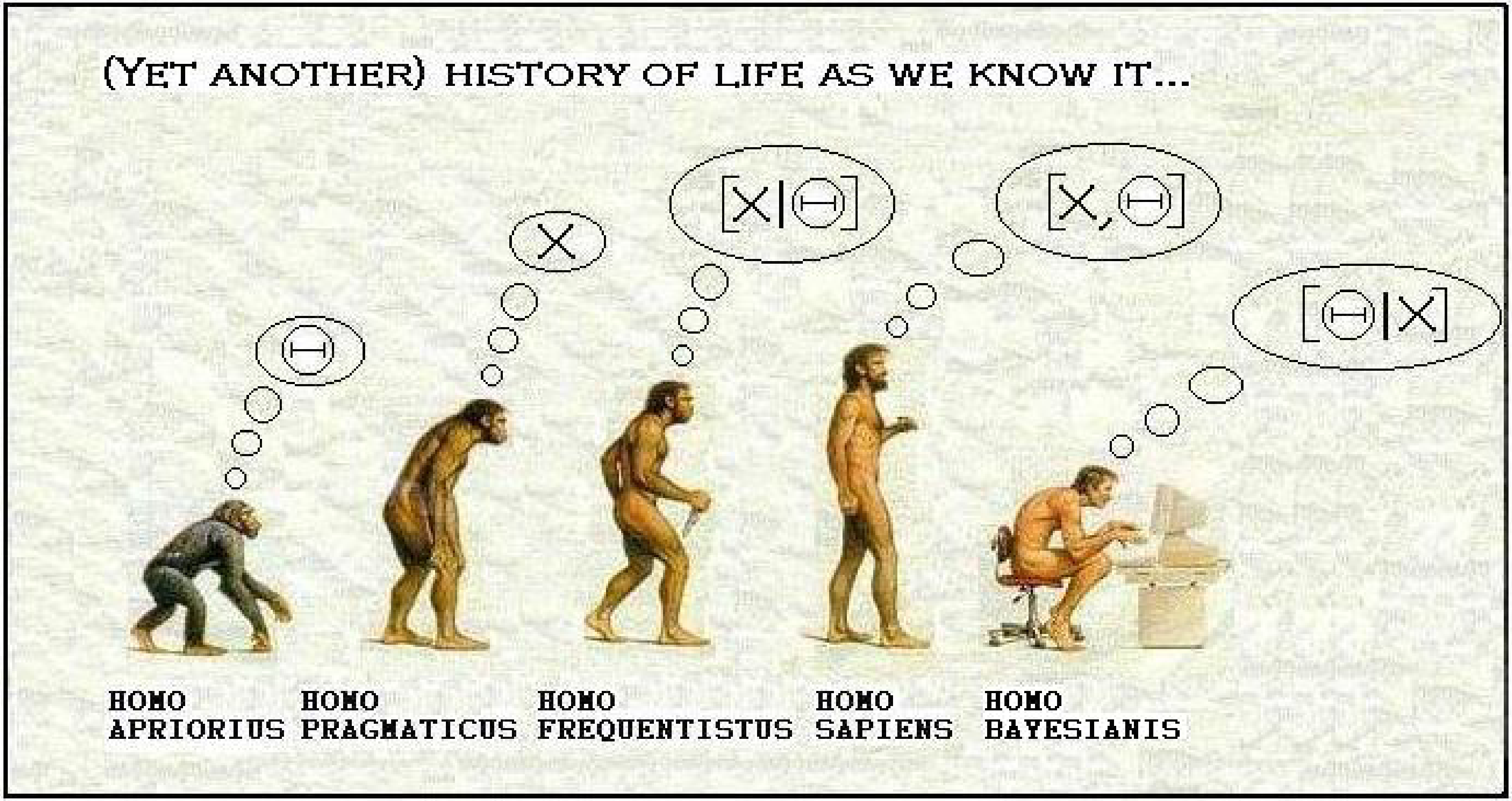
Graphical Models



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
$[X|\Theta]$
$[X,\Theta]$
$[\Theta|X]$
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

Stochastic Computation for Large-Scale Graphical Models

—

Exploratory Analysis of Gene Expression Data

Mike West, Duke University

Adrian Dobra
Carlos Carvalho

Beatrix Jones
Chris Hans

Institute of Statistics and Decision Sciences
&
Computational and Applied Genomics Program

Quanli Wang Joseph Nevins Guang Yao

High-Dimensional Graphical Models

- Exploratory data analysis & visualisation
- Observational data: Varying contexts
- Uses in prediction

- Sparsity

Exploring High-Dimensional Observational Data

Exploring & summarising associations:

- structure in p($x_{1},\ldots,x_{p}$)
- visualisation, clues

Predictive models:

- regressions and classification models
- compositional regressions for retrospective data

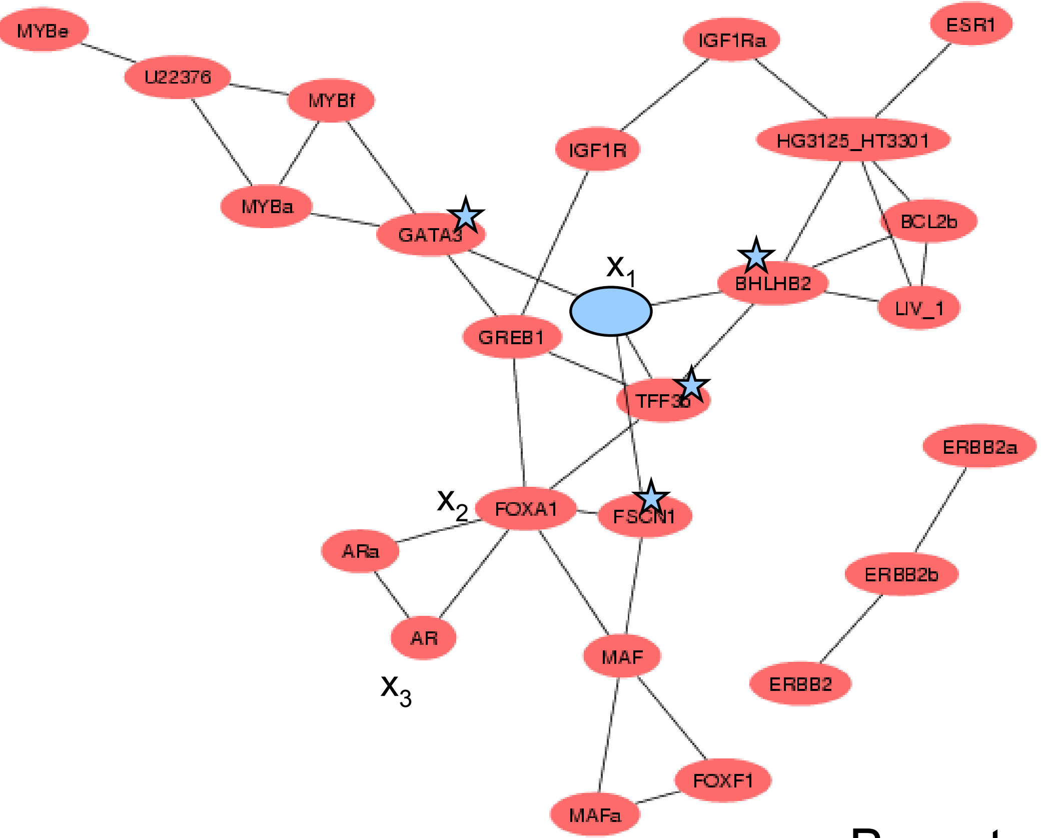
$$\mathrm{p}(y|x) \propto \mathrm{p}(x_1|y)\mathrm{p}(x_2|y, x_1)\mathrm{p}(x_3|y, x_1, x_2)$$

Graphical Models: p(x)

Nodes=p variables

Edges: ne(i)=neighbours

Conditional (in)dependencies

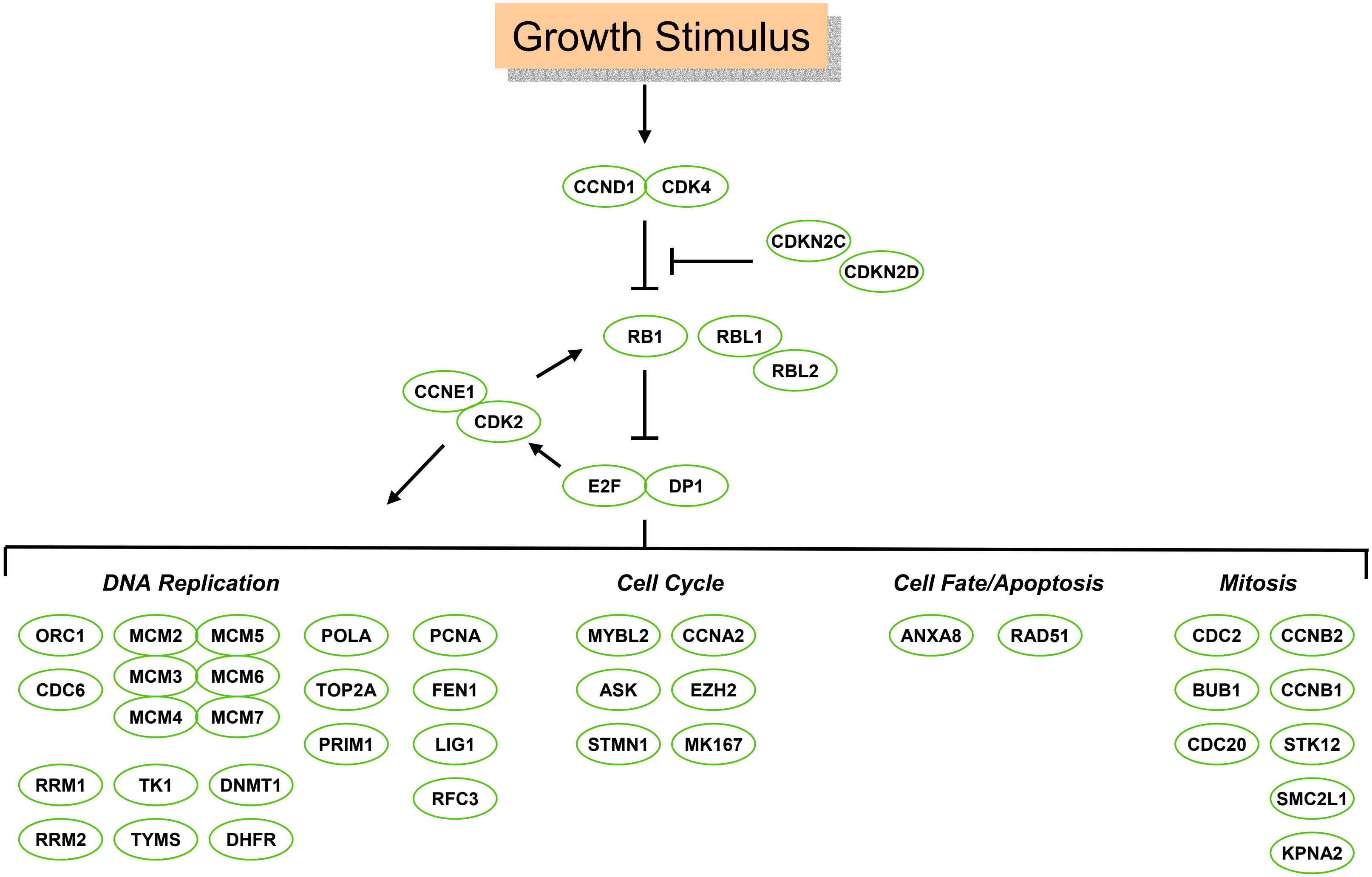


$$p(x_i|x_{-i}) = p(x_i|x_{ne(i)})$$

(prediction: $y=x_0$)

Breast cancer - ER subgraph: p=12558

The Rb-E2F Pathway



Growth Stimulus

CCND1 CDK4

CDKN2C CDKN2D

RB1 RBL1 RBL2

CCNE1 CDK2

E2F DP1

DNA Replication

ORC1 CDC6

MCM2 MCM5 MCM3 MCM6 MCM4 MCM7

POLA TOP2A PRIM1

PCNA FEN1 LIG1 RFC3

RRM1 TK1 DNMT1

RRM2 TYMS DHFR

Cell Cycle

MYBL2 CCNA2

ASK EZH2

STMN1 MK167

Cell Fate/Apoptosis

ANXA8 RAD51

Mitosis

CDC2 CCNB2

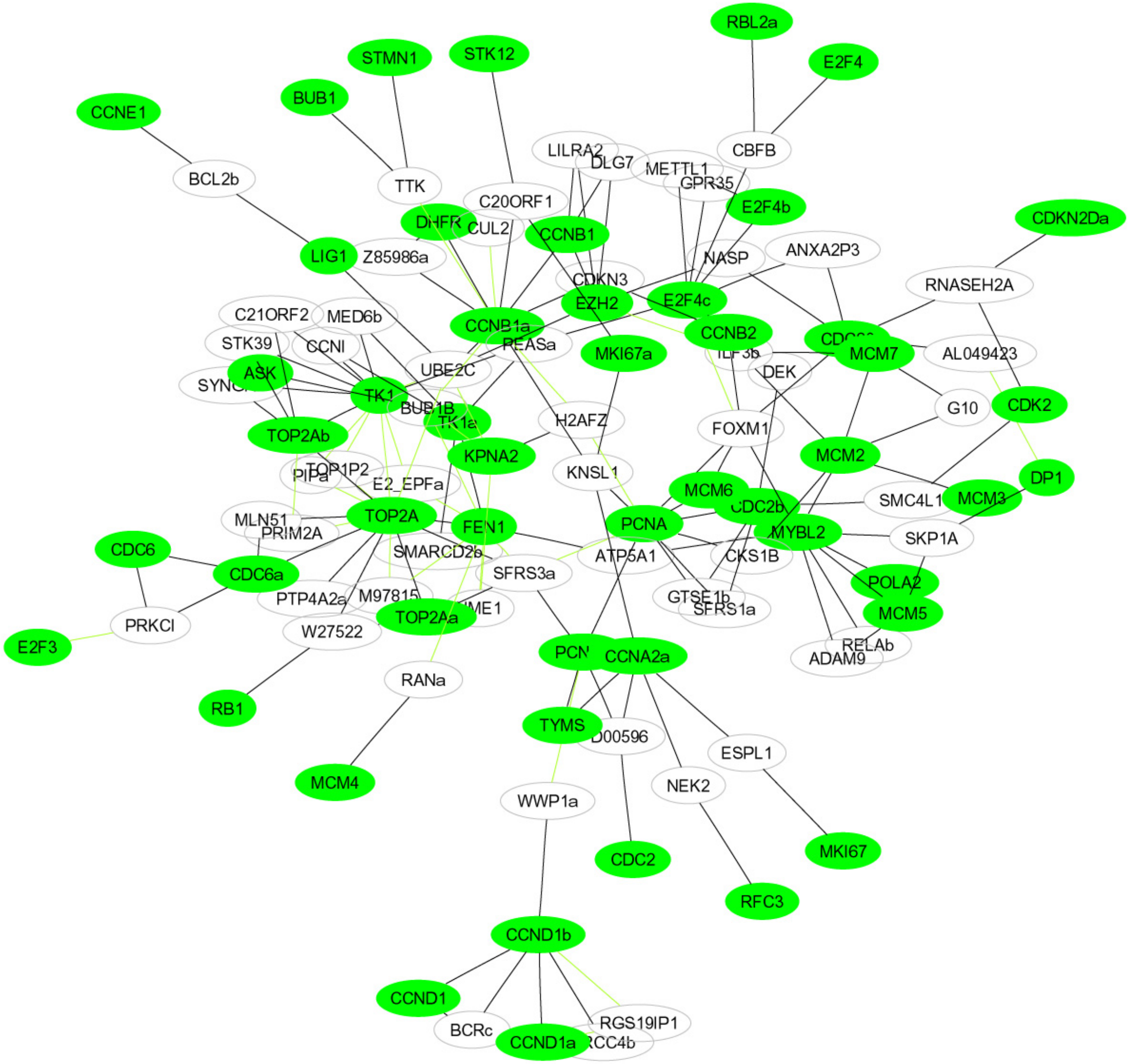
BUB1 CCNB1

CDC20 STK12

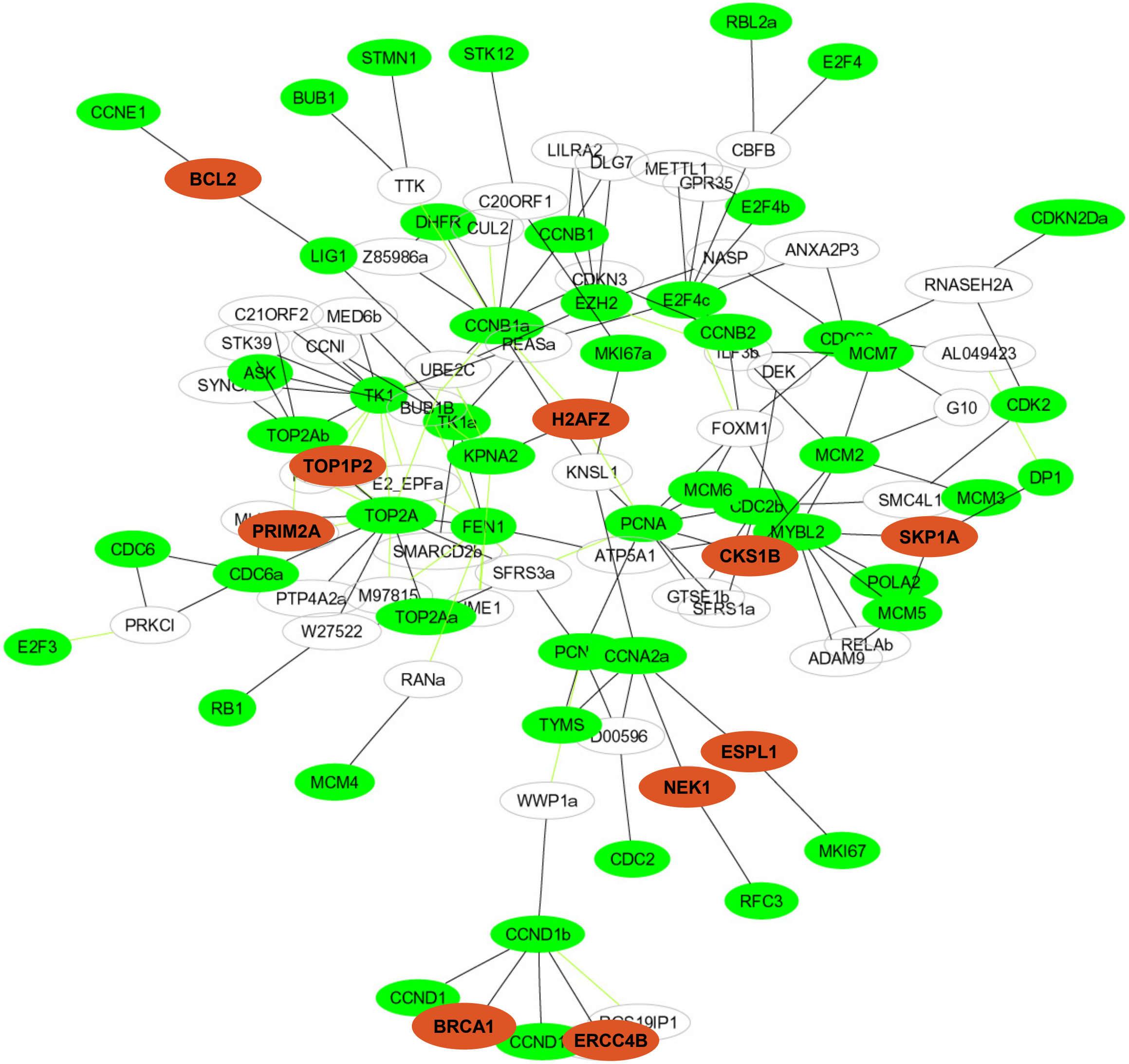
SMC2L1

KPNA2

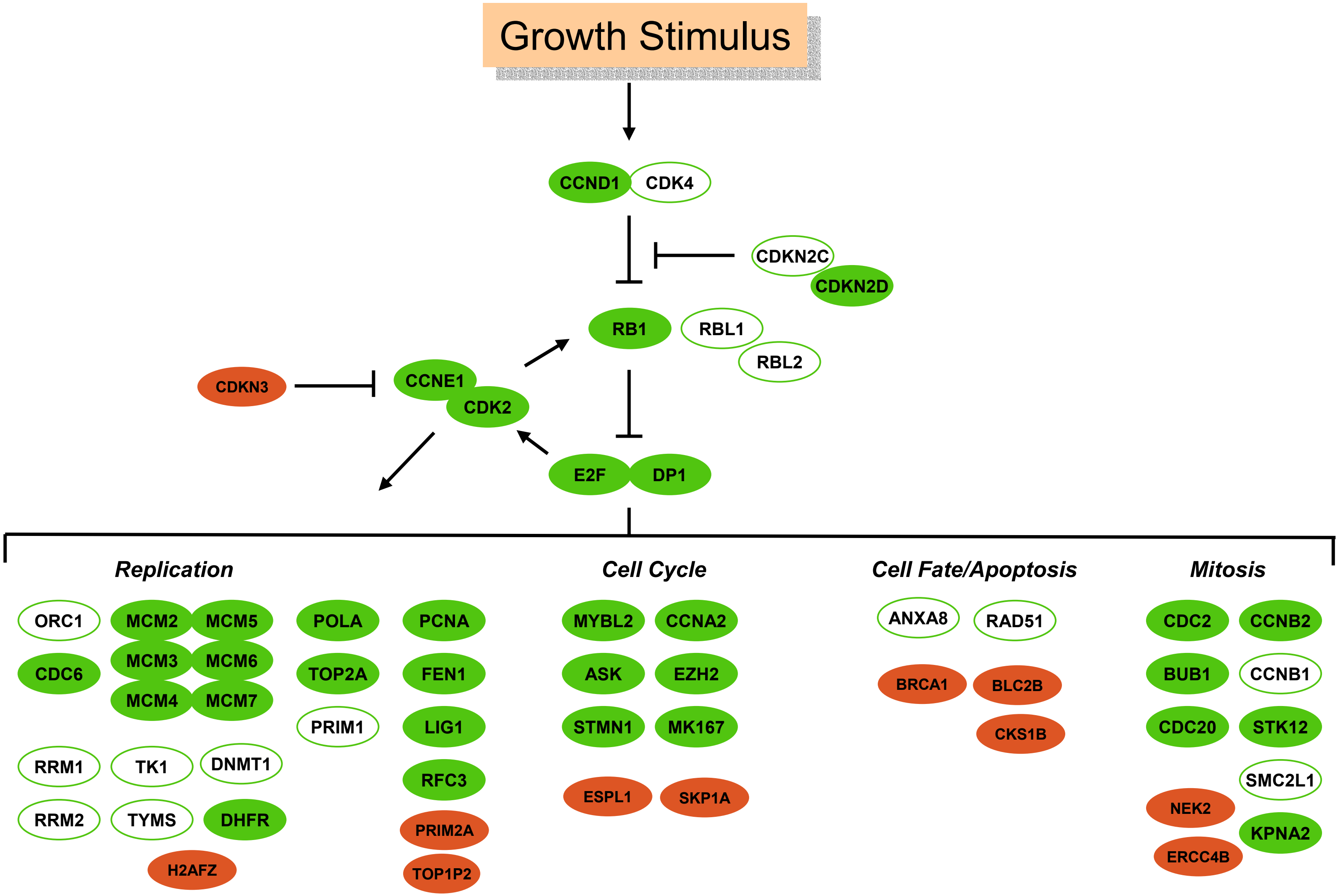
A Subgraph of an Rb-E2F Association Graph
- Pathway Exploration -



Gene Set Enrichment by Neighbour & Path Exploration



Enriching the Rb-E2F Pathway Picture



Growth Stimulus

CCND1 CDK4

CDKN2C CDKN2D

RB1 RBL1 RBL2

CDKN3

CCNE1 CDK2

E2F DP1

Replication

ORC1
MCM2 MCM5
CDC6
MCM3 MCM6
MCM4 MCM7
POLA
PCNA
TOP2A
FEN1
PRIM1
LIG1
RFC3
PRIM2A
TOP1P2
RRM1
TK1
DNMT1
RRM2
TYMS
DHFR
H2AFZ

Cell Cycle

MYBL2 CCNA2
ASK EZH2
STMN1 MK167
ESPL1 SKP1A

Cell Fate/Apoptosis

ANXA8 RAD51
BRCA1 BLC2B
CKS1B

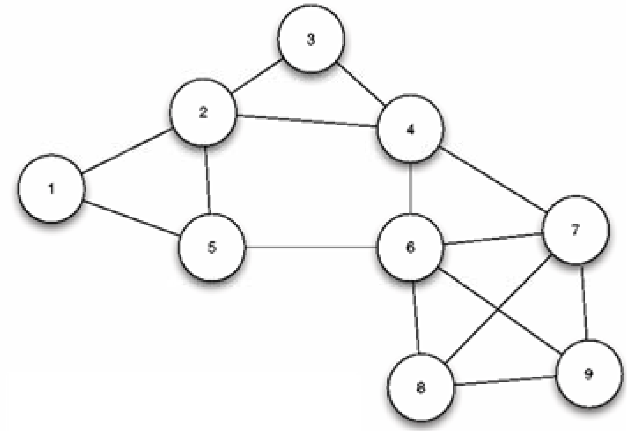
Mitosis

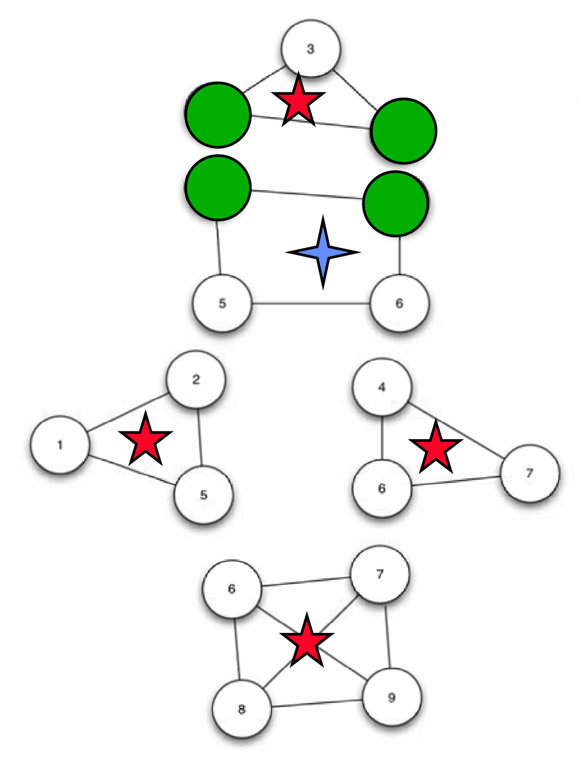
CDC2 CCNB2
BUB1 CCNB1
CDC20 STK12
SMC2L1
NEK2
KPNA2
ERCC4B

Graph Decompositions: The Key to Dealing with Dimension

Dimension 'reduces' via graph decompositions:

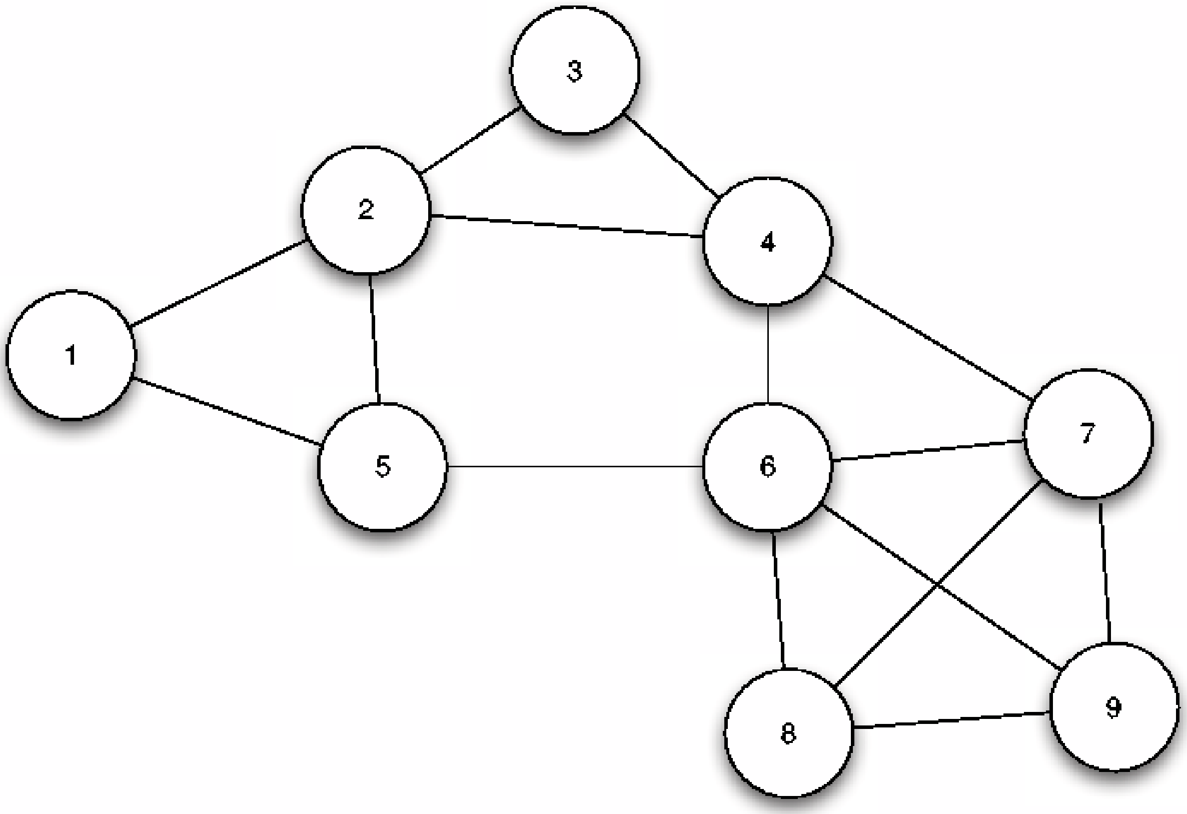
Intersecting or disconnected subgraphs



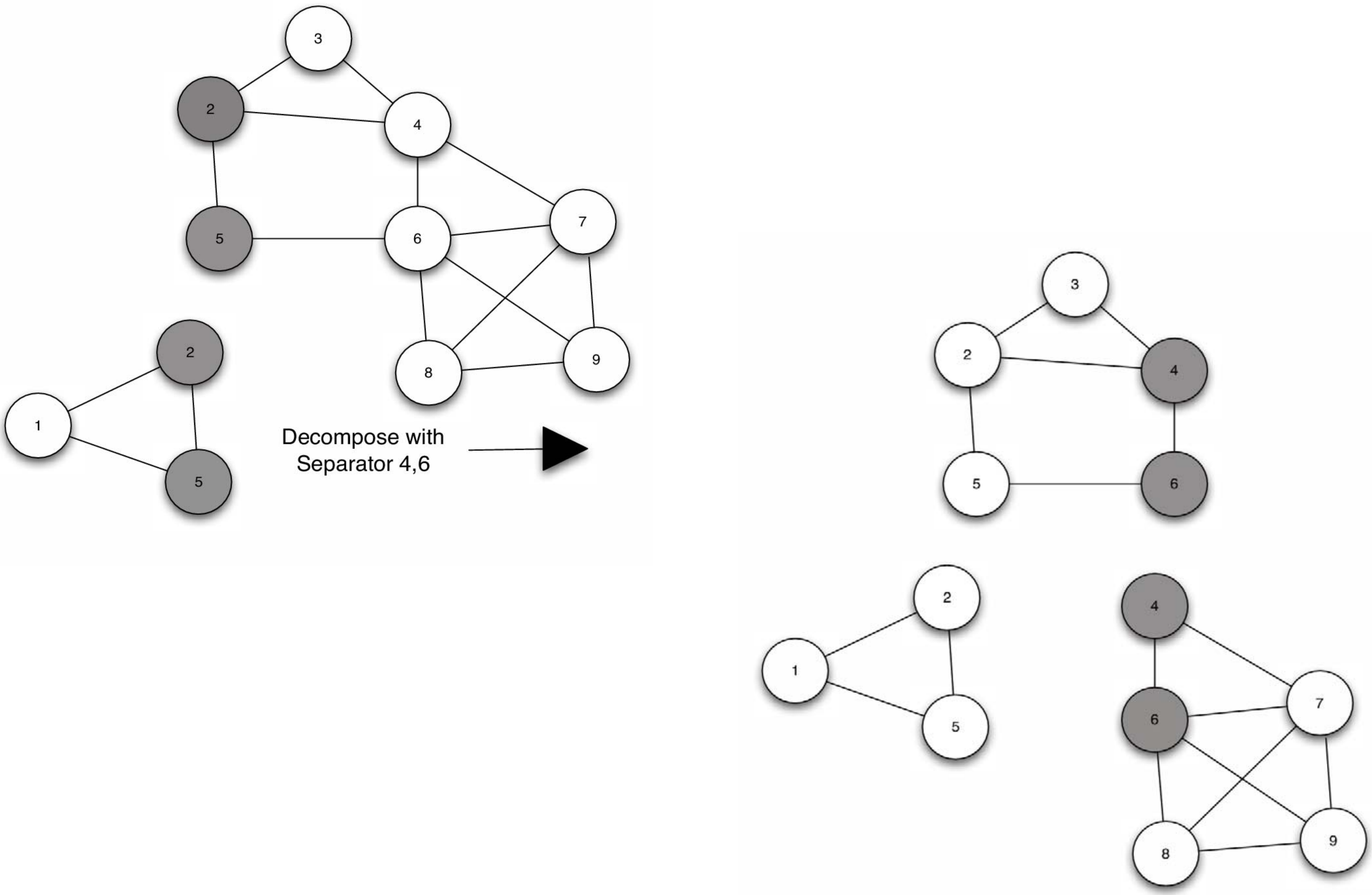


S : separator …
- complete subgraph that "separates" PCs

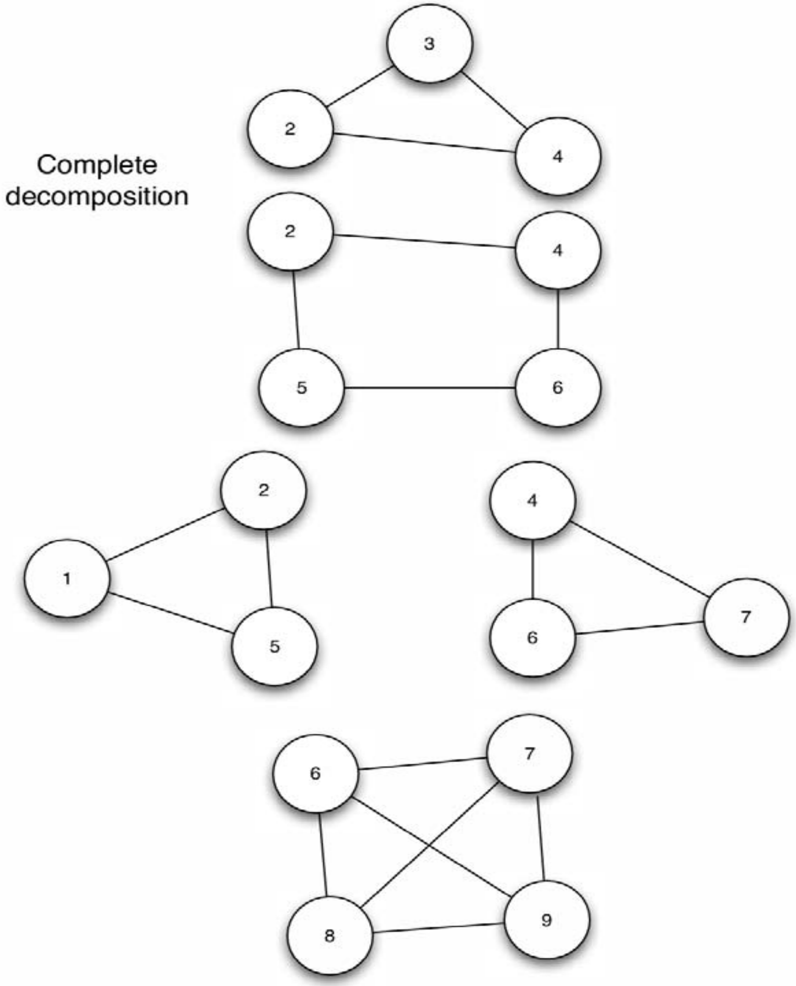
PC : prime component …
- either a complete subgraph, ★
- or cannot be separated ✦

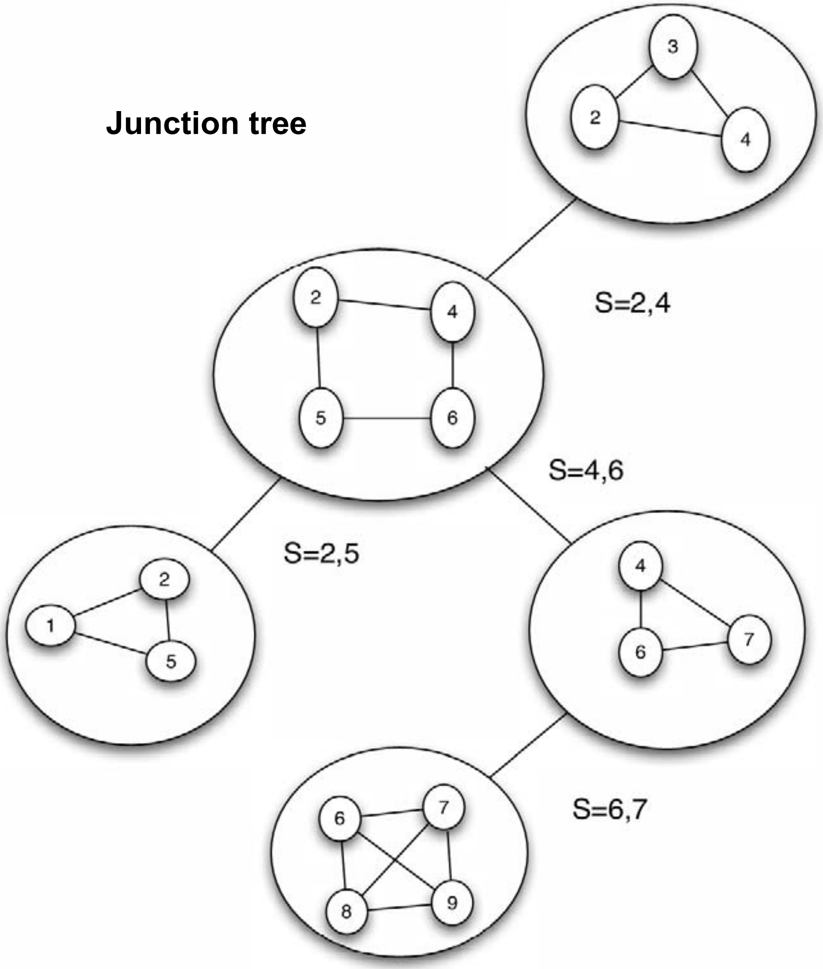


Decompose with
Separator 2,5



Decompose with Separator 4,6





Decomposable graphs:
PC=clique (maximal complete subgraph)

Non-decomposable:
PC=anything goes (e.g., big chains, cycles, …

Gaussian Graphical Models

Precision **$K=\Sigma^{-1}$** on a graph *G*

$$\mathrm{E}(x_i|x_{-i}) = \sum_j (-\mathrm{K}_{ij}/\mathrm{K}_{ii})x_j$$

Edges ~ $|K_{ij}|>0$

Inference:

"covariance selection"

Priors:

- over graphs *G* ...
- then non-zero elements of **K** on *G*
- **sparsity inducing**

Computation: finding graphs: MCMC, stochastic search, annealing

- Dimension!
- Distributed/cluster computation

Graph, Model and Prior Decompositions: Dealing with Dimension

Dimension 'reduces' via graph decompositions:

$$p(x \mid G, K) = \prod_{PC} p(x_{PC} \mid K_{PC}) / \prod_{S} p(x_S \mid K_S)$$

- PC : prime components
- S : separators – complete subgraphs separating PCs

Large, realistic sparse graphs: massive decomposition

Decomposable graphs: PC=clique

Non-decomposable: Anything goes (e.g., big chains)

General graphical model

Priors over Covariance/Precision Matrix

Likelihood:

$$\mathrm{p}(x \mid G, K) = \prod_{PC} \mathrm{p}(x_{PC} \mid K_{PC}) / \prod_{S} \mathrm{p}(x_S \mid K_S)$$

Conjugate prior: Hyper Wishart (Roverato 2002)

$$\mathrm{p}(K \mid G) \propto \prod_{PC} \mathrm{p}(K_{PC}) / \prod_{S} \mathrm{p}(K_S)$$

"Local Hyper-Wishart Prior"

Wishart on cliques

Constrained Wishart on non-complete prime components

Graph Decomposition
- Stochastic Computational Strategies -

Wander around G space: evaluate p(x|G), then p(G|x)

(marginalised over parameters **K**)

$$\mathrm{p}(x\,|\,G) = \prod\nolimits_{PC} \mathrm{m}_{PC} \Big/ \prod\nolimits_{S} \mathrm{m}_{S}$$

m_S: Wishart normalising constants

m_{PC}:

– PC complete: analytic (inverse Wishart normalizing constants)

(e.g., decomposable: Giudici & Green 99; Wong & Carter 02)

- PC incomplete: (hard) integral over constrained Wishart

Monte Carlo (Atay-Kayis & Massam 02 … 04)

(Roverato 2002, Scand J Stat)

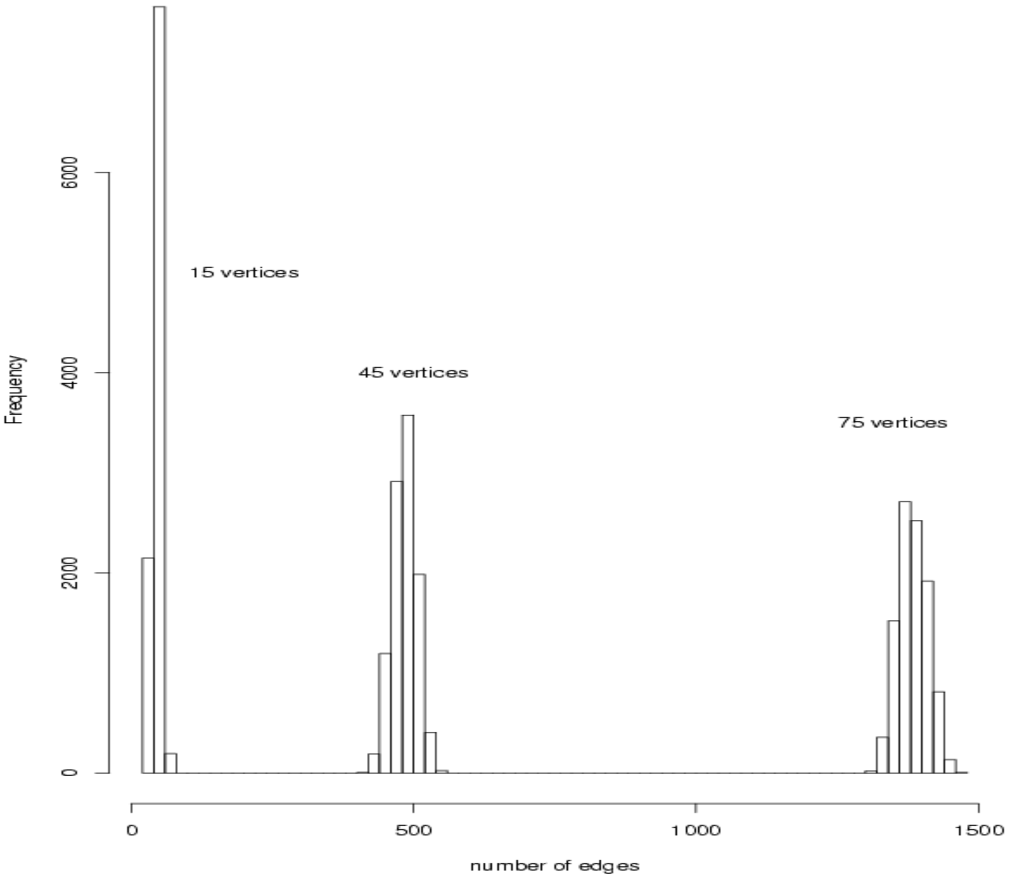
Priors on Graphs

Uniform prior ?

Unrestricted graphs:

$E(\#edges)=p(p-1)/4$

Decomposable cases



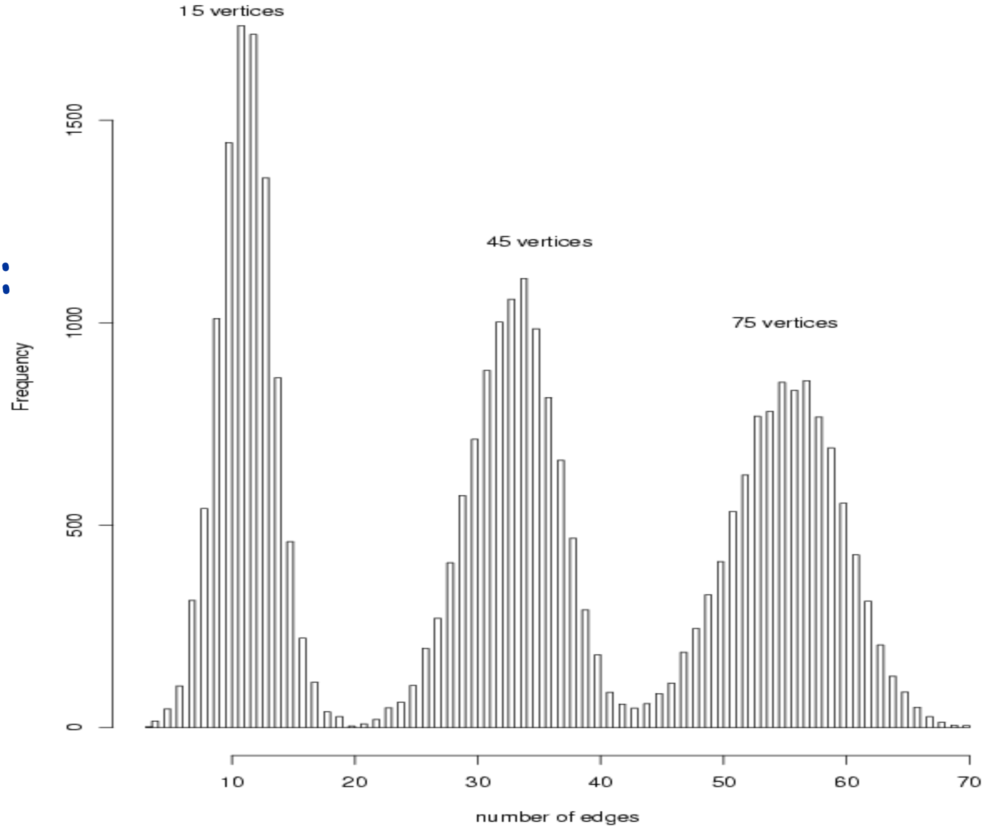
Sparsity Priors on Graphs

Edge inclusion prob.

β=2/(p-1)

Unrestricted graphs: mode of p edges

Decomposable cases →



Local Computations on Graph Space

Current graph G : $p(x|G)$

G 1-edge neighbours G^*: $p(x|G^*)$

Major efficiencies if decomposable:

- change 2 cliques at most
- efficiently check decomposability

(Giudici & Green 99)

Non-decomposable? *Anything can happen*

Practical relevance of decomposability?

Most published examples: $p=4-12$

Challenges: efficiency, dimension

MCMC on Graph Space

Gibbs: random 1-edge moves

Decomposable models: Giudici & Green 99, Wong & Carter 02

Conditional posterior for edge in/out

Metropolis Hasting

random choice of add/delete,
random 1-edge move (Jones et al, Duke team, 03)

Challenges: computations in non-decomposable cases

Monte Carlo (Atay-Kayis & Massam 02/04) hugely challenging

Issues: Horrible/impossible as p increases

Annealed Stochastic Search

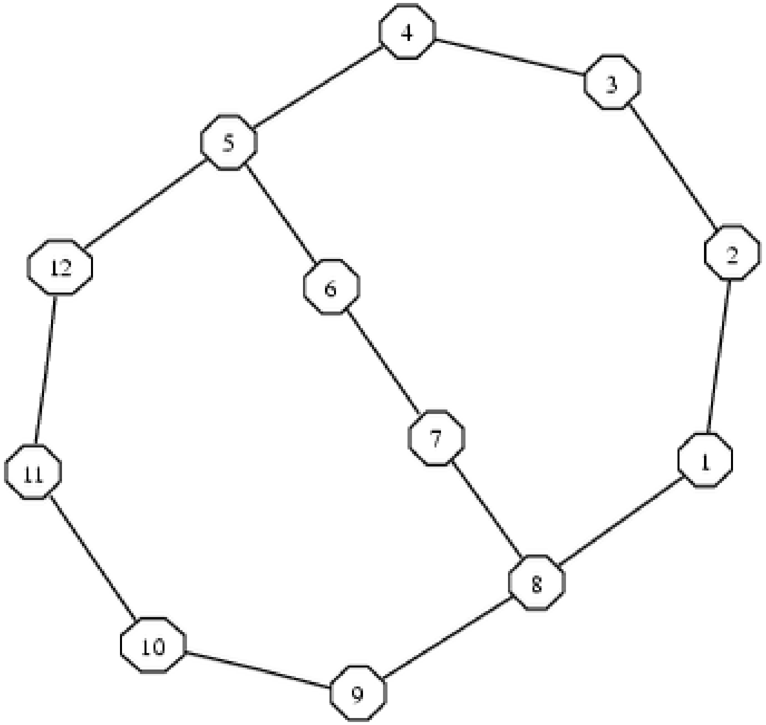
Wander around G space: find “high probability” graphs

- 1-edge different graphs (“neighbours”)
  - store top h
- Select “next” : $p(G|x)^a$
- repeat

Parallelisable

Multiple “interesting” regions of model space

12 Node Example



- Non-decomposable

  A non-complete prime component

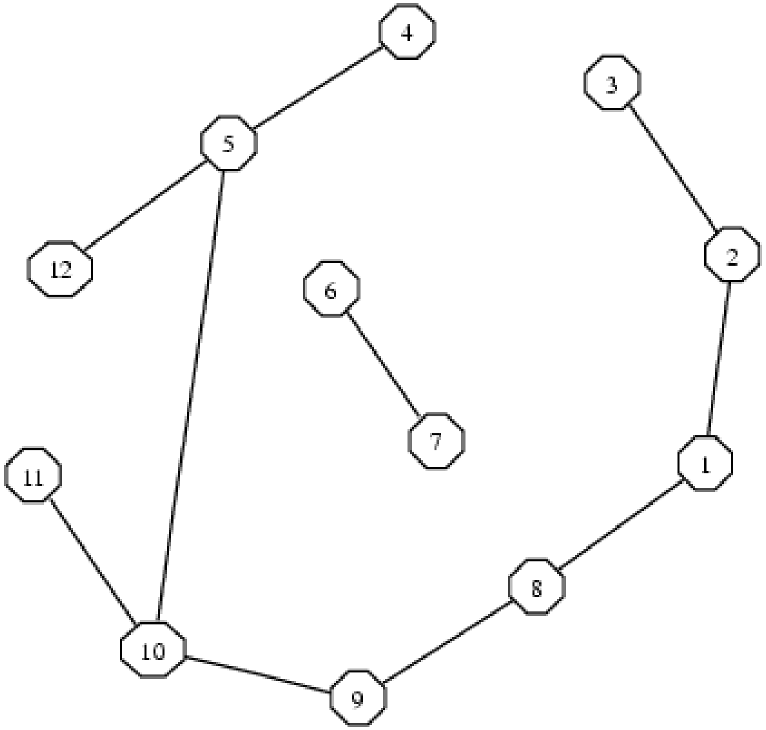
- n=250
- 10,000 SS steps
- 10,000x66 MH steps
- a=1

12 Node Example

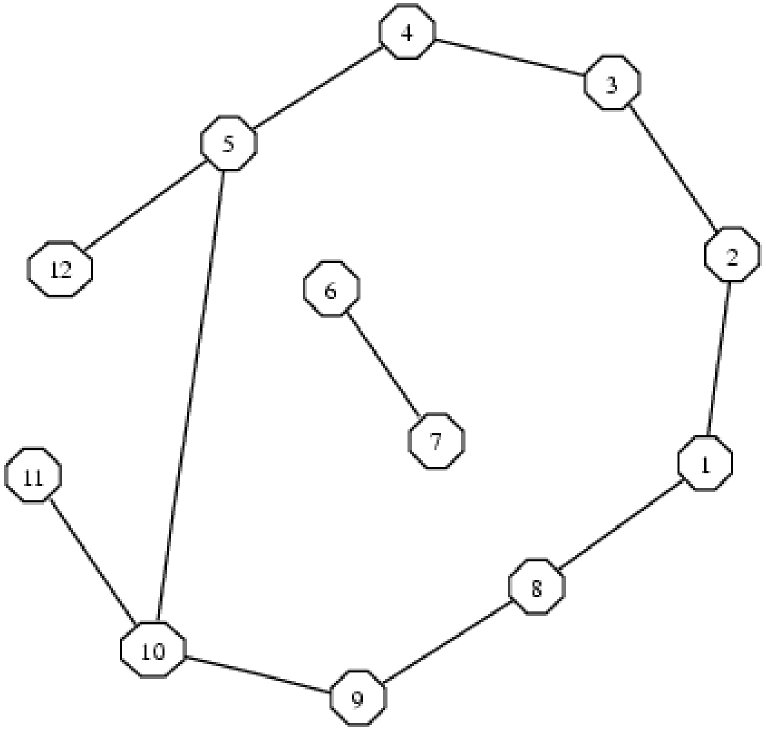
Method	Time (s)	Top log posterior	Evaluations to top grph	Time to top graph
MH decom-posable	36	-2591.18	912	1
SS decom-posable	183	-2591.18	792	2
MH un-restricted	15,220	-2590.94	415	2
SS un-restricted	2773	-2590.94	13266	5

True graph has lower marginal likelihood than either of these

12 Node Example "Top Graphs"

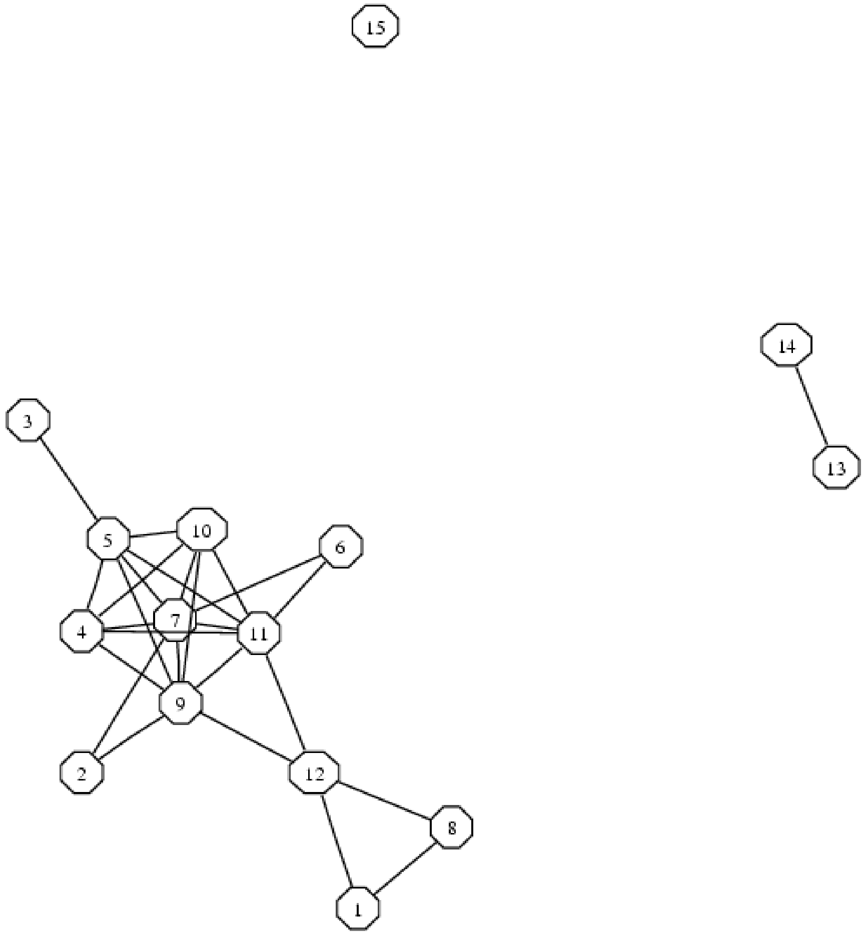


Decomposable



Unrestricted

15 Node Example



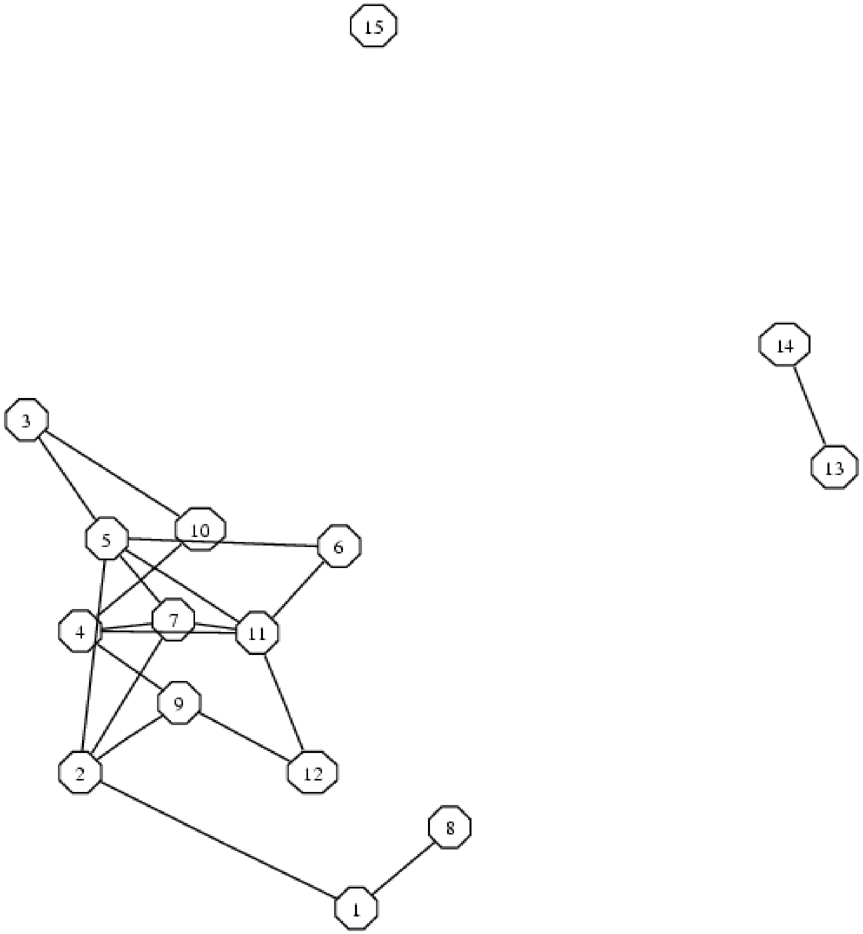
- Decomposable
- n=250
- 10,000 SS steps
- 10,000x105 MH steps
- a=1

15 Node Example

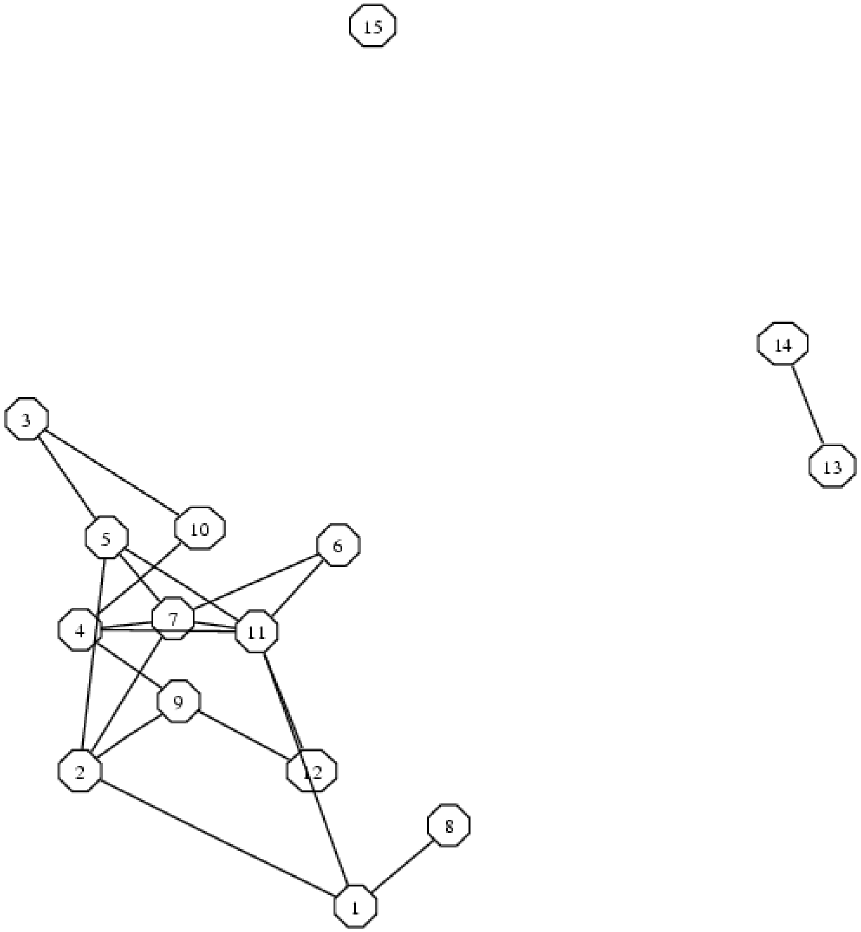
Method	Time (s)	Top log posterior	Evaluations to top grph	Time to top graph
MH decom-posable	93	15633.76	349,484	36
SS decom-posable	234	15633.76	33,495	9
MH un-restricted	513,077	15633.83	666, 425	309, 222
SS un-restricted	5930	15636.33	82845	112

True graph has lower marginal likelihood than any of these

15 Node Example "Top Graphs"



Decomposable



Unrestricted

150 Node Example

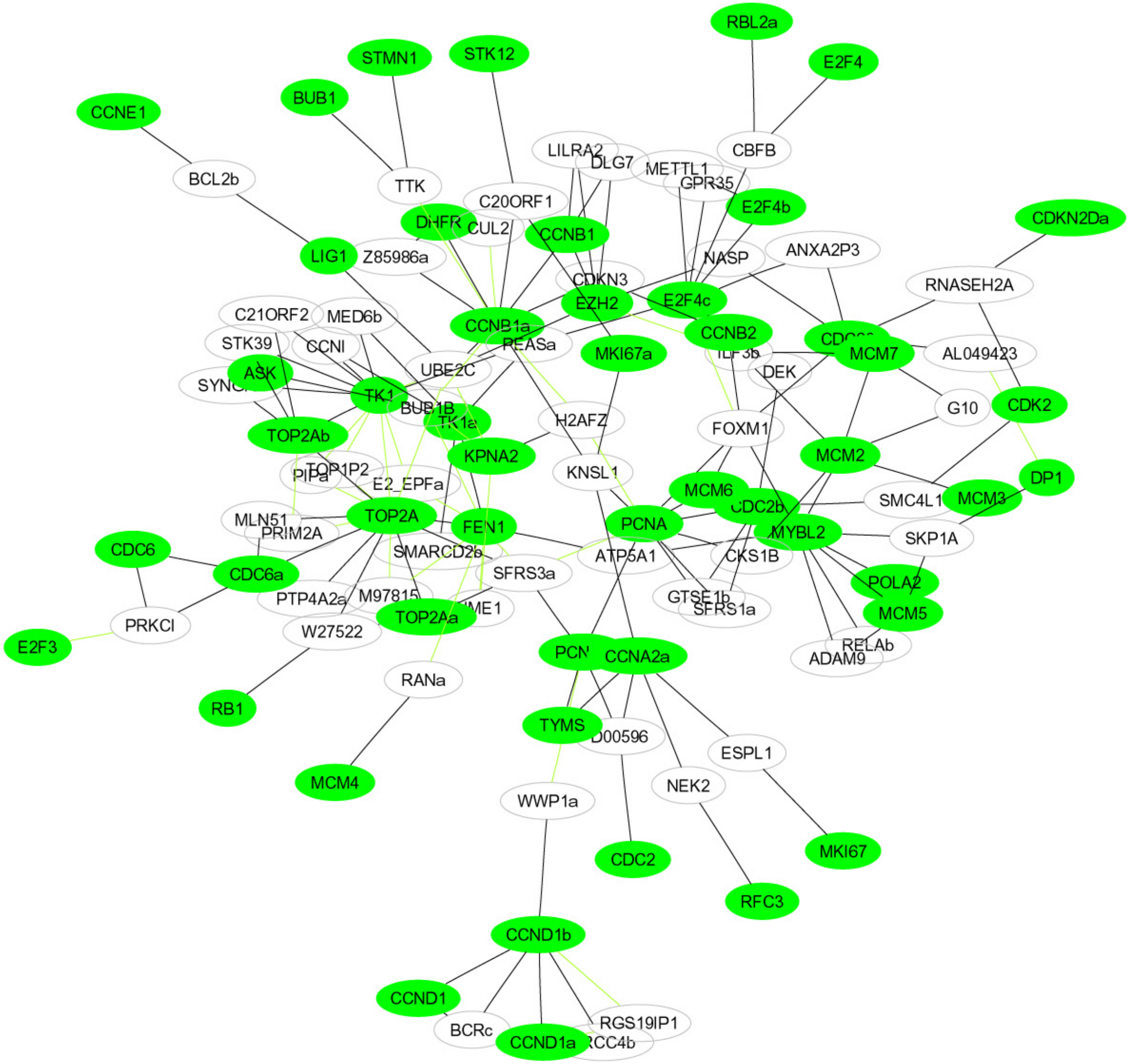
- gene expression – breast cancer : ER
- (O)Estrogen receptor pathway
- n=49
- 25,000 SS steps
- 25,000x11,175 MH steps
- annealing?
- unrestricted: accuracy of MC evaluations

150 Node Example

Method	Time (hrs)	Top log posterior	Evaluations to top grph	Time to top graph
MH decomposable	18.02	-9417.97	100,467k	6.51
SS decomposable	0.03	-9260.35	1699k	0.03
★ SS unrestricted	**6.29**	**-9227.68**	**44.7k**	**3.39**

★ **Starting from the ‘best’ decomposable graph, and is a (local) mode**

Scaling-Up?



RBL2a
E2F4
STMN1
STK12
BUB1
CCNE1
CBFB
LILRA2
DLG7
METTL1
GPR35
BCL2b
TTK
C20ORF1
E2F4b
CDKN2Da
DHFR
CUL2
CCNB1
LIG1
Z85986a
NASP
ANXA2P3
CDKN3
RNASEH2A
EZH2
E2F4c
C21ORF2
MED6b
CCNB1a
CCNB2
STK39
CCNI
REASa
MKI67a
CDC
MCM7
AL049423
ILF3b
SYNC
ASK
UBE2C
DEK
TK1
BUB1B
G10
CDK2
TK1a
H2AFZ
FOXM1
TOP2Ab
KPNA2
MCM2
TOP1P2
PIPa
E2_EPFa
KNSL1
DP1
MCM6
CDC2b
SMC4L1
MCM3
MLN51
TOP2A
FEN1
PCNA
PRIM2A
MYBL2
SKP1A
CDC6
SMARCD2b
ATP5A1
CKS1B
CDC6a
SFRS3a
POLA2
PTP4A2a
M97815
GTSE1b
SFRS1a
ME1
MCM5
TOP2Aa
PRKCI
W27522
RELAb
E2F3
ADAM9
PCN
CCNA2a
RANa
RB1
TYMS
D00596
ESPL1
MCM4
NEK2
WWP1a
CDC2
MKI67
RFC3
CCND1b
CCND1
RGS19IP1
BCRc
CCND1a
CC4b

Scaling Up: p>15
High-Dimensional Sparse Models

- Build directed graphical models
- Induce conditional independence graph
- Priors on directed (acyclic) graphs

Compositional Networks: Parallel Regressions

$$\mathrm{p}(x \mid K) = \prod_{i=1}^{p} \mathrm{p}(x_i \mid x_{cne(i)}, \theta_i), \qquad cne(i) \subseteq \{(i+1):p\}$$

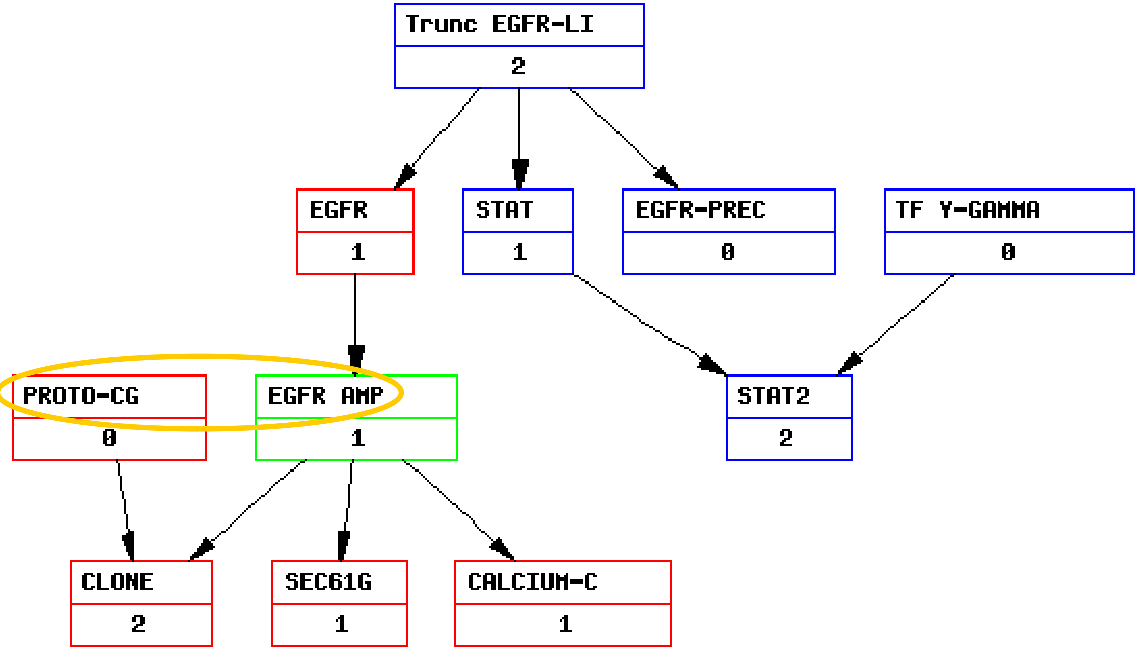
Regression parameters functions of **K**

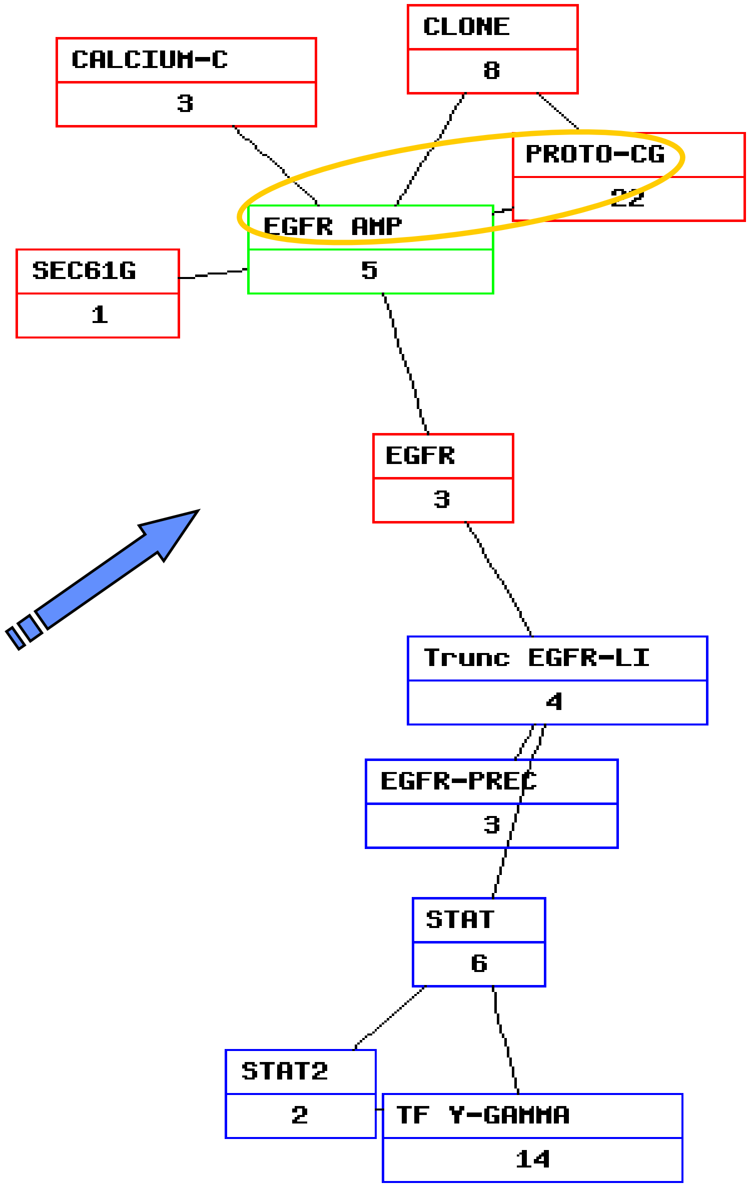
Triangular array: Order matters

EGFR

Brain cancer gene expression

Duke Keck Center for Neurooncogenomics





DAG to G: moralise

Compositional Nets: Priors & Computation

- Normal/inverse gamma priors ~ Wisharts

- Sparsity: Regression variable inclusion/exclusion priors

- Stochastic model search via sets of regressions:
    - parallel MCMC for regression search
    - select from multiple 'neighbouring' graphs
    - 'local' exploration near 'good' graphs
    - 'local' MCMC ?

- Evaluate relative posterior prob on directed graph

Shotgun Stochastic Search on DAGs

- Current DAG: ordering, edges, posterior probability
- Local shuffle of ordering: switch two neighbours
- Local recomputation of regression search, sampling
- Compute posterior prob on new DAG

- Optimisation steps to initialise and update orderings

  Concept: "explanatory" genes low in ordering

- Gibbs sampling for full conditional regressions
  - to initialise
  - and reduce candidate predictors for each x
- Shotgun search for regressions (forthcoming)

Code: C++/MPI Beowulf cluster implementation

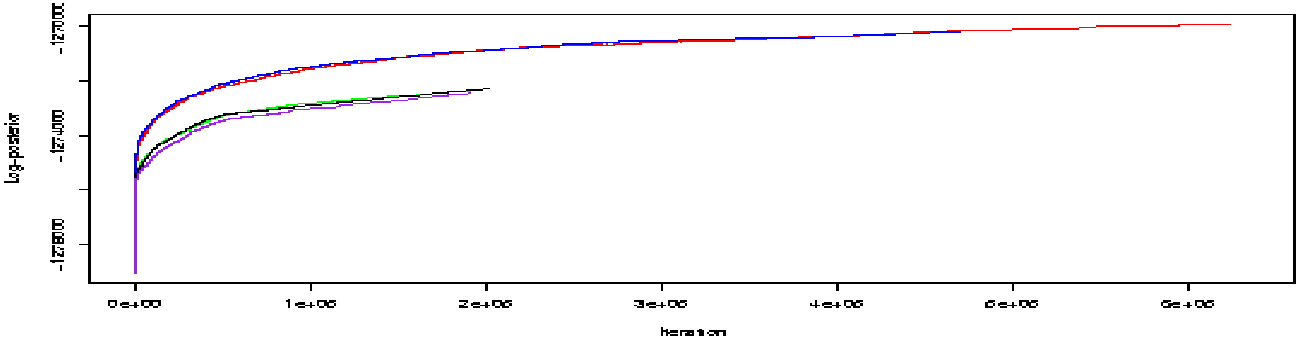
HdBCS (Dobra, Duke)

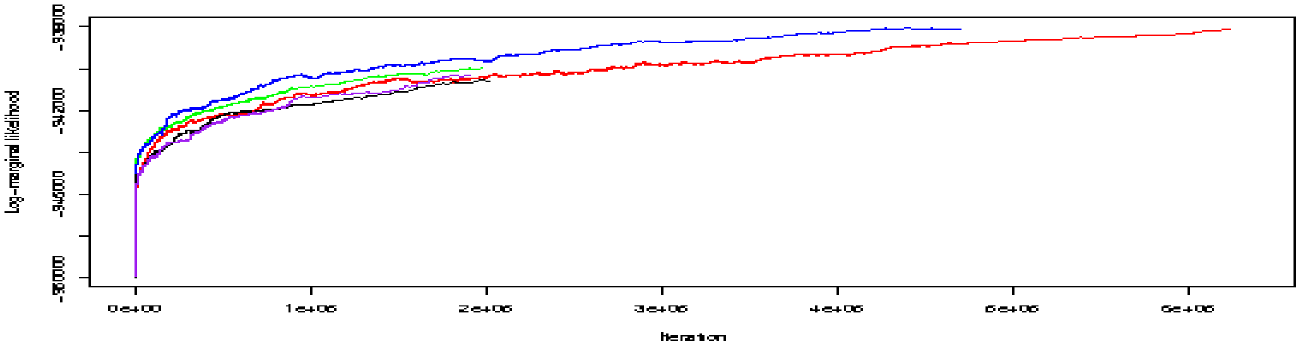
Breast Cancer Gene Expression

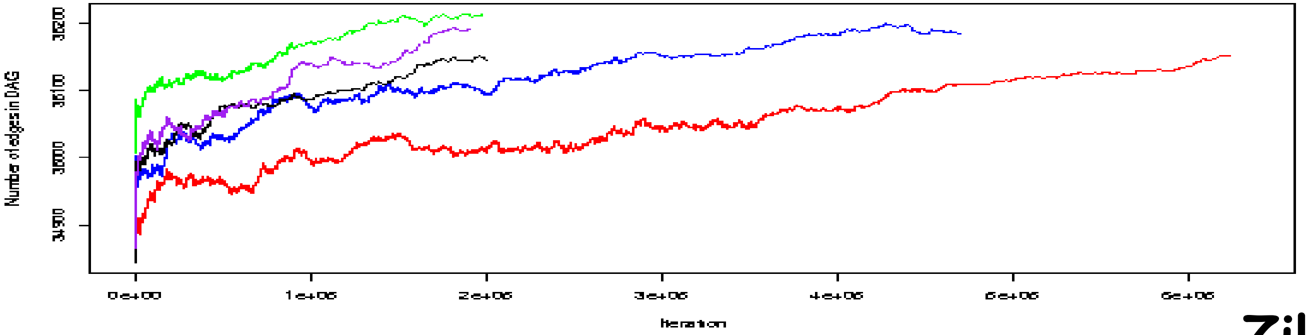
- n=158 tumour samples
- p=12558
- 48x2cpu cluster
- Summarisation of high prob graphs?

Code: C++/MPI Beowulf cluster implementation
HdBCS: Adrian Dobra, Duke

Beowulf cluster is needed

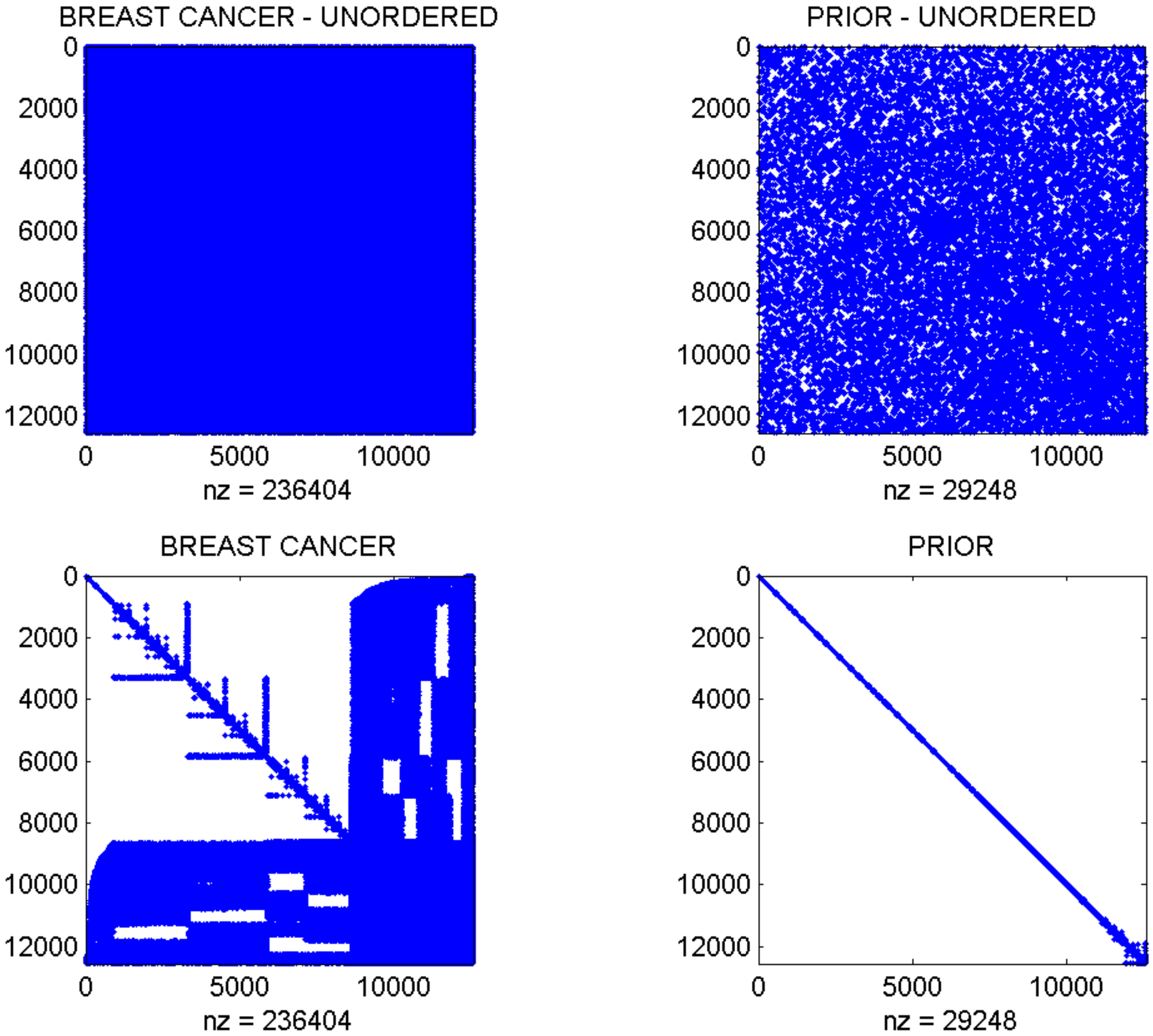




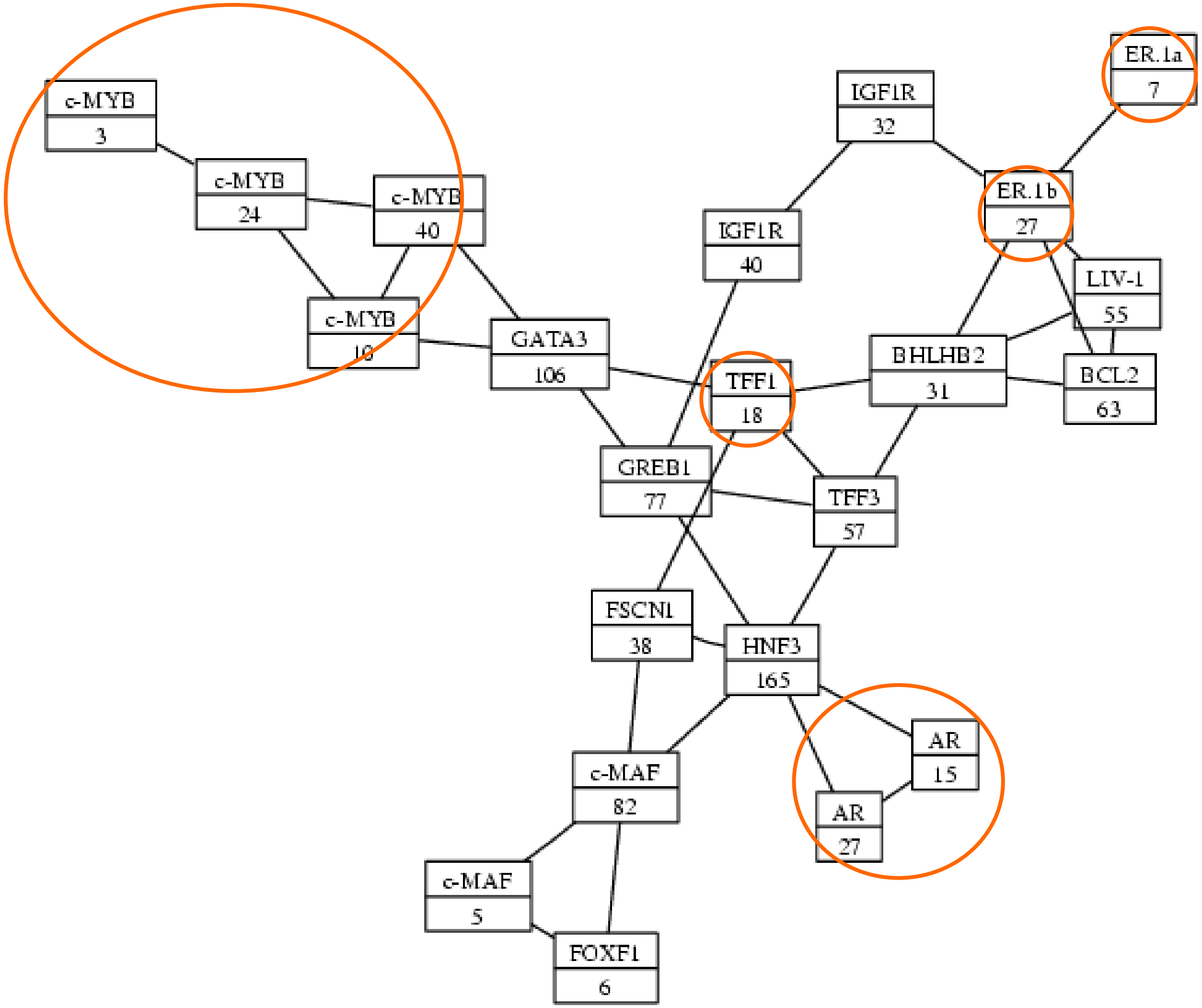


Zillions

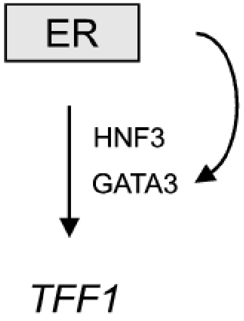
p=12558
Adjacency Matrix of An "Interesting" Graph

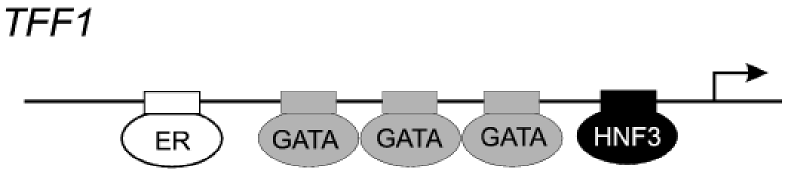


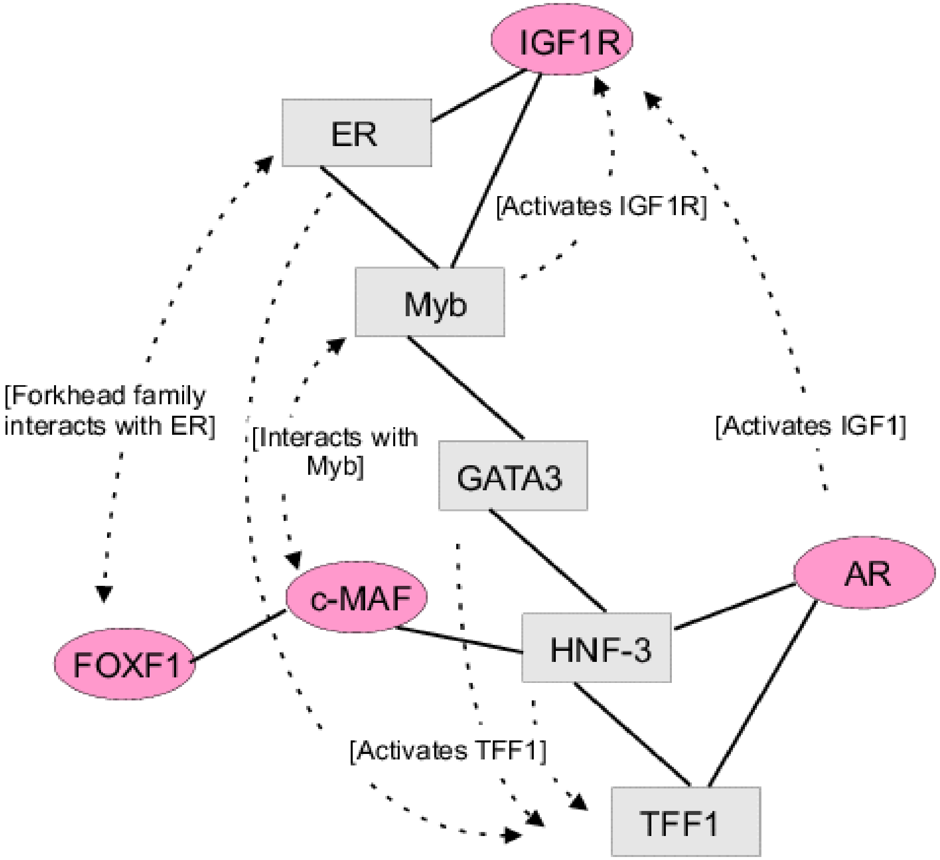
p=12558: ER Genes Subgraph



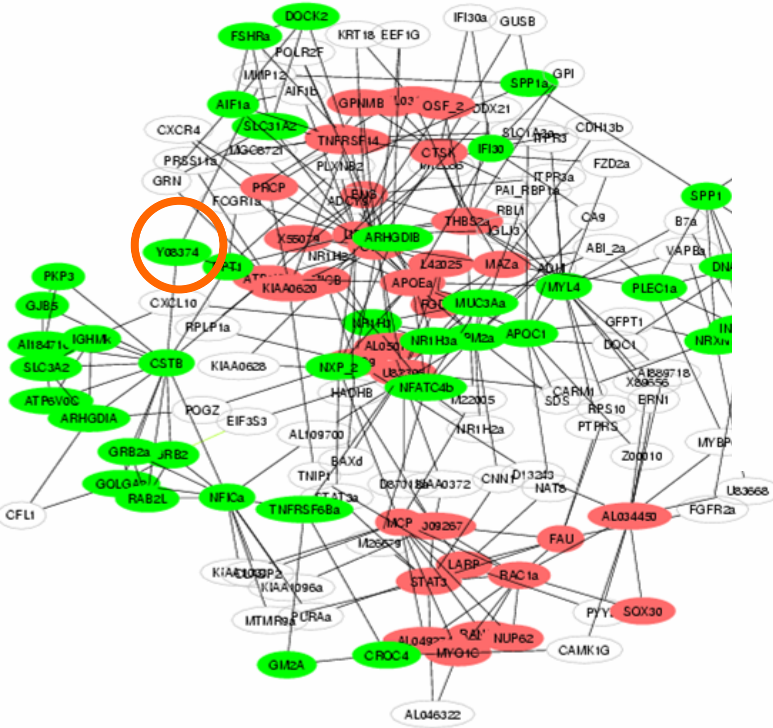
p=12558: Breast Cancer Gene Expression

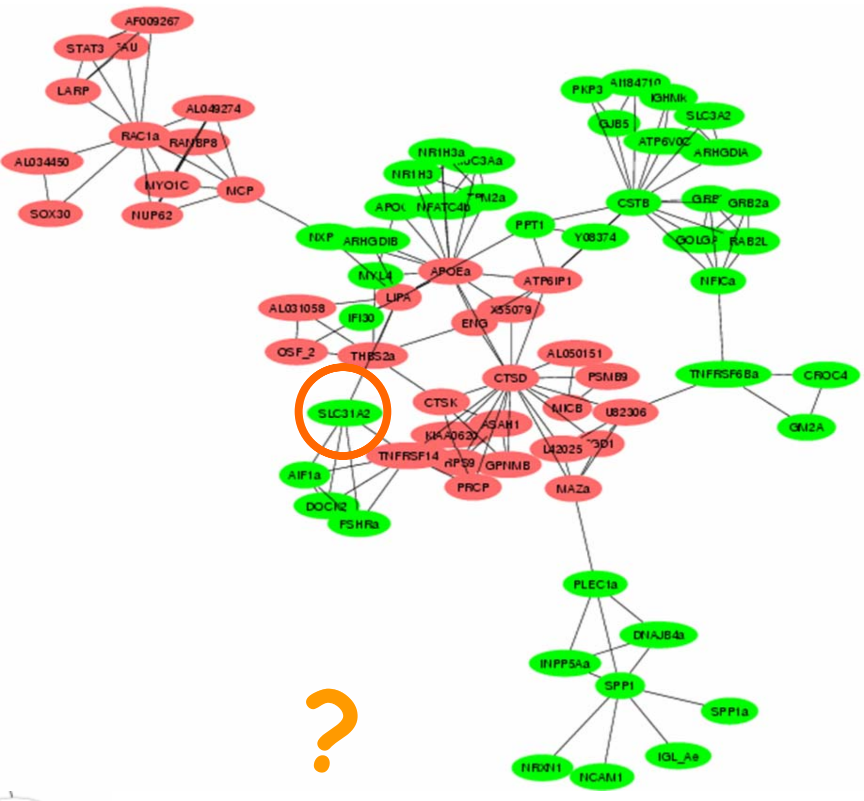






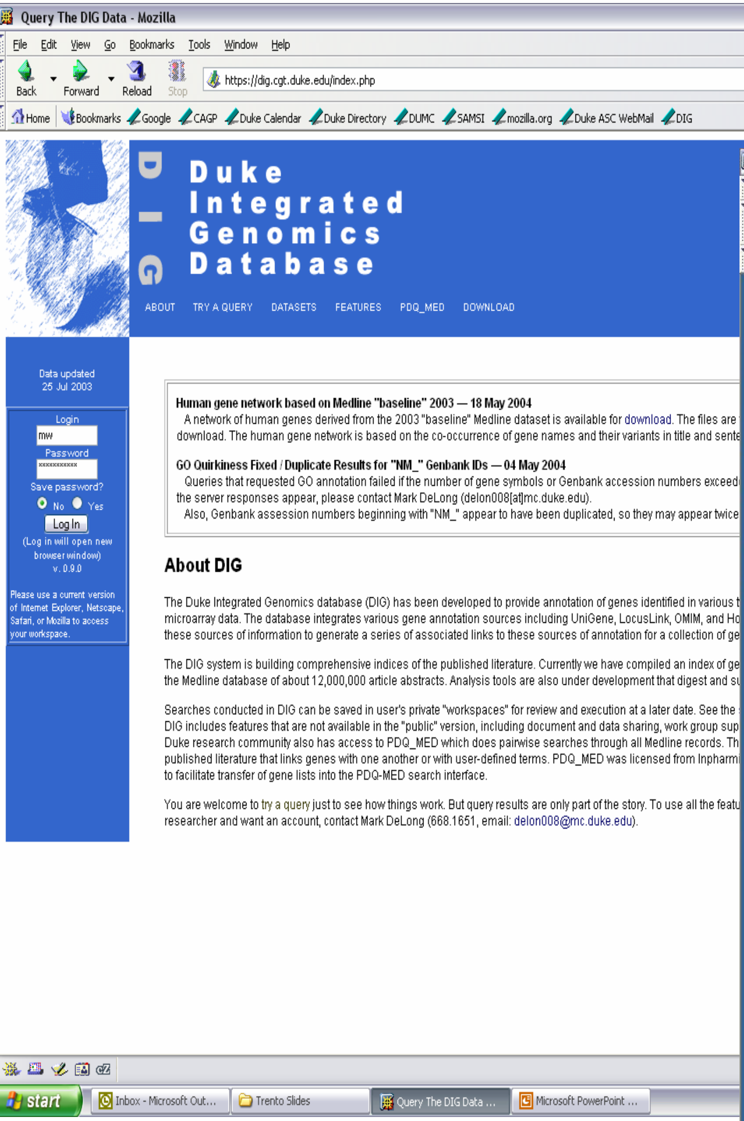
Neighbour & Path Exploration





?

Exploring Graphs: Gene Discovery & Annotation


Query The DIG Data - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
https://dig.cgt.duke.edu/index.php
Search Print
Home Bookmarks Google CAGP Duke Calendar Duke Directory DUMC SAMSI mozilla.org Duke ASC WebMail DIG
Duke Integrated Genomics Database
ABOUT TRY A QUERY DATASETS FEATURES PDQ_MED DOWNLOAD
Login
mw
Password
Save password?
○ No ○ Yes
Log In
start Inbox - Microsoft Out... Trento Slides Query The DIG Data ... Microsoft PowerPoint ...

Data updated
25 Jul 2003

(Log in will open new browser window)
v. 0.9.0

Please use a current version of Internet Explorer, Netscape, Safari, or Mozilla to access your workspace.

Human gene network based on Medline "baseline" 2003 — 18 May 2004

A network of human genes derived from the 2003 "baseline" Medline dataset is available for download. The files are
download. The human gene network is based on the co-occurrence of gene names and their variants in title and sente

GO Quirkiness Fixed / Duplicate Results for "NM_" Genbank IDs — 04 May 2004

Queries that requested GO annotation failed if the number of gene symbols or Genbank accession numbers exceed
the server responses appear, please contact Mark DeLong (delon008[at]mc.duke.edu).

Also, Genbank assession numbers beginning with "NM_" appear to have been duplicated, so they may appear twice

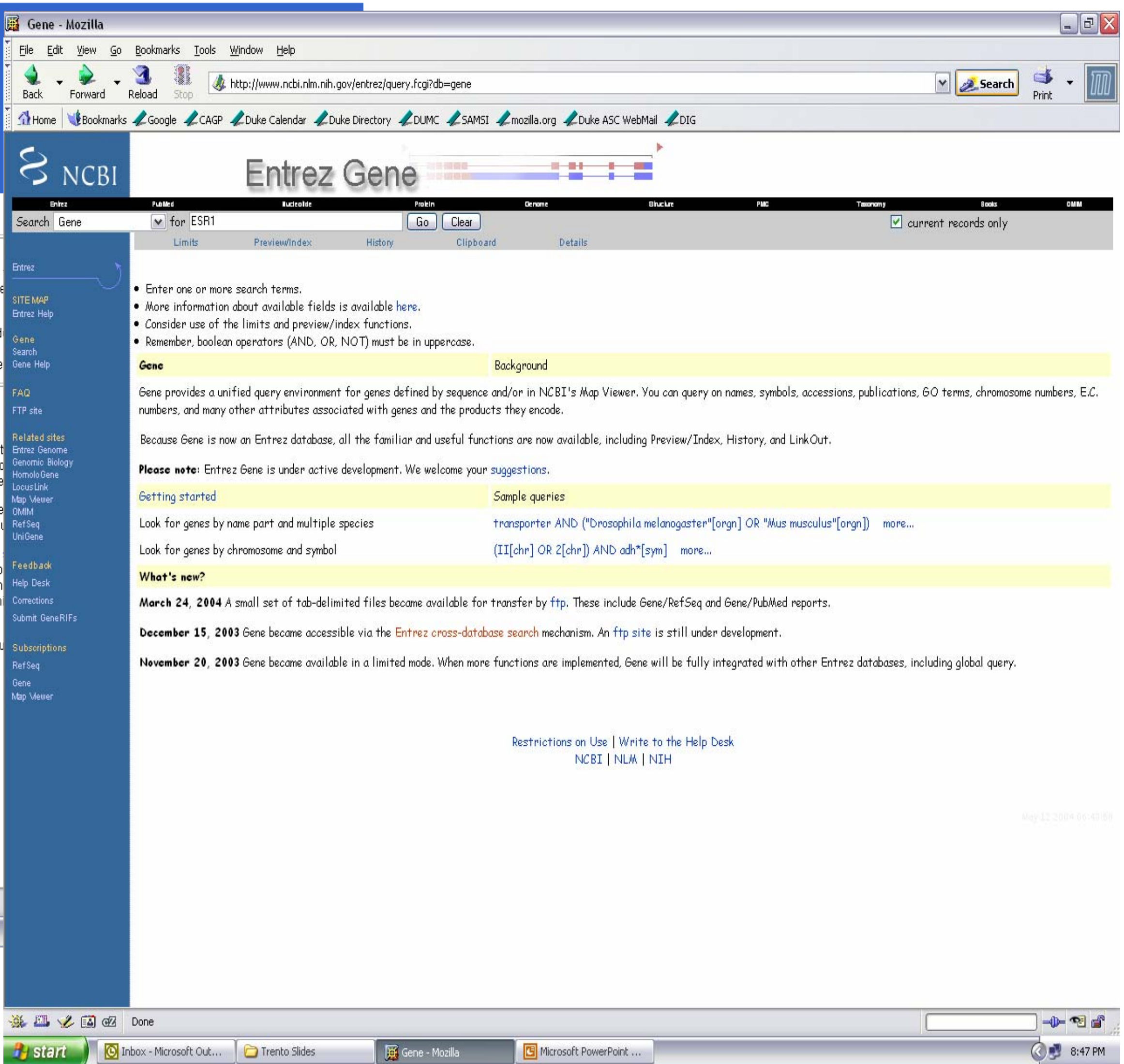
About DIG

The Duke Integrated Genomics database (DIG) has been developed to provide annotation of genes identified in various t
microarray data. The database integrates various gene annotation sources including UniGene, LocusLink, OMIM, and Ho
these sources of information to generate a series of associated links to these sources of annotation for a collection of ge

The DIG system is building comprehensive indices of the published literature. Currently we have compiled an index of ge
the Medline database of about 12,000,000 article abstracts. Analysis tools are also under development that digest and su

Searches conducted in DIG can be saved in user's private "workspaces" for review and execution at a later date. See the
DIG includes features that are not available in the "public" version, including document and data sharing, work group sup
Duke research community also has access to PDQ_MED which does pairwise searches through all Medline records. Th
published literature that links genes with one another or with user-defined terms. PDQ_MED was licensed from Inpharmi
to facilitate transfer of gene lists into the PDQ-MED search interface.

You are welcome to try a query just to see how things work. But query results are only part of the story. To use all the featu
researcher and want an account, contact Mark DeLong (668.1651, email: delon008@mc.duke.edu).


Gene - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
Search Print
Home Bookmarks Google CAGP Duke Calendar Duke Directory DUMC SAMSI mozilla.org Duke ASC WebMail DIG
NCBI
Entrez Gene
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM
Search Gene for ESR1 Go Clear
☑ current records only
Limits Preview/Index History Clipboard Details
Entrez
SITE MAP
Entrez Help
Gene
Search
Gene Help
FAQ
FTP site
Related sites
Entrez Genome
Genomic Biology
HomoloGene
LocusLink
Map Viewer
OMIM
RefSeq
UniGene
Feedback
Help Desk
Corrections
Submit GeneRIFs
Subscriptions
RefSeq
Gene
Map Viewer
Done
start Inbox - Microsoft Out... Trento Slides Gene - Mozilla Microsoft PowerPoint ...
8:47 PM

- Enter one or more search terms.
- More information about available fields is available here.
- Consider use of the limits and preview/index functions.
- Remember, boolean operators (AND, OR, NOT) must be in uppercase.

Gene	Background

Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. You can query on names, symbols, accessions, publications, GO terms, chromosome numbers, E.C. numbers, and many other attributes associated with genes and the products they encode.

Because Gene is now an Entrez database, all the familiar and useful functions are now available, including Preview/Index, History, and LinkOut.

Please note: Entrez Gene is under active development. We welcome your suggestions.

Getting started	Sample queries
Look for genes by name part and multiple species	transporter AND ("Drosophila melanogaster"[orgn] OR "Mus musculus"[orgn]) more...
Look for genes by chromosome and symbol	(II[chr] OR 2[chr]) AND adh*[sym] more...

What's new?

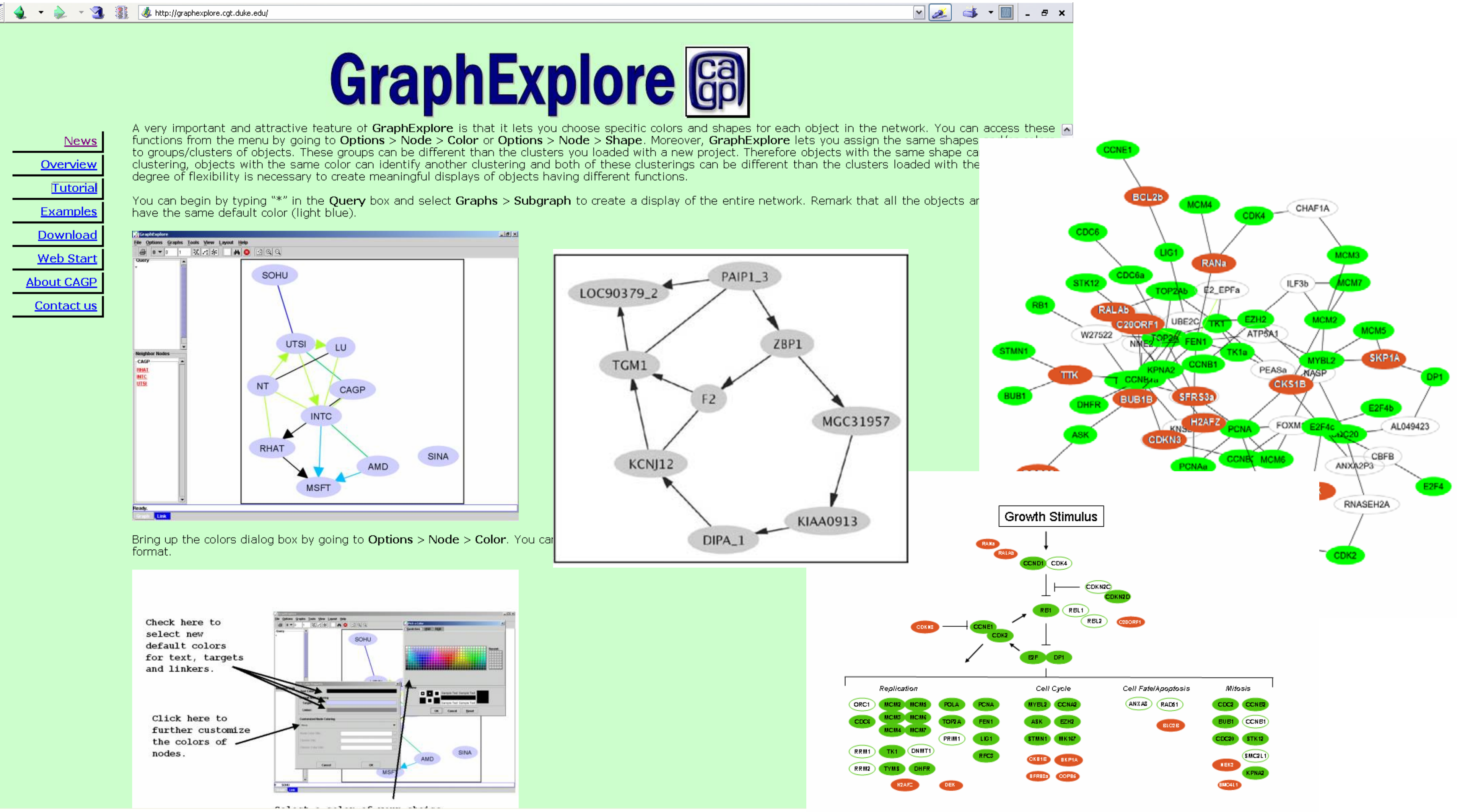
March 24, 2004 A small set of tab-delimited files became available for transfer by ftp. These include Gene/RefSeq and Gene/PubMed reports.

December 15, 2003 Gene became accessible via the Entrez cross-database search mechanism. An ftp site is still under development.

November 20, 2003 Gene became available in a limited mode. When more functions are implemented, Gene will be fully integrated with other Entrez databases, including global query.

Restrictions on Use | Write to the Help Desk
NCBI | NLM | NIH

http://dig.cgt.duke.edu



A very important and attractive feature of GraphExplore is that it lets you choose specific colors and shapes for each object in the network. You can access these functions from the menu by going to Options > Node > Color or Options > Node > Shape. Moreover, GraphExplore lets you assign the same shapes to groups/clusters of objects. These groups can be different than the clusters you loaded with a new project. Therefore objects with the same shape ca clustering, objects with the same color can identify another clustering and both of these clusterings can be different than the clusters loaded with the degree of flexibility is necessary to create meaningful displays of objects having different functions.

You can begin by typing "*" in the Query box and select Graphs > Subgraph to create a display of the entire network. Remark that all the objects ar have the same default color (light blue).

Bring up the colors dialog box by going to Options > Node > Color. You car format.

Check here to select new default colors for text, targets and linkers.

Click here to further customize the colors of nodes.

Large-scale graphical model search and evaluation

Inference on large, sparse inverse covariance matrix

(Dobra et al, JMVA 2004)   graphexplore.cagp.duke.edu

Some key refs:

- Giudici and Green 1999, Biometrika

  MCMC in decomposable graphical models
- Roverato 2002, Scand J Stat

  HIW priors on graphical models
- Wong and Carter 2002, tech report, Hong Kong Univ

  covariance selection
- Atay-Kayis and Massam 2002, tech report, York Univ

  Monte Carlo evaluation of marginal likelihoods
- Jones, West et al 2003, tech report, SAMSI & Duke Univ

  stochastic computation and search
- Dobra, West et al 2004, J Multivariate Anal

  initial DAG based approach, stochastic search & gene expression studies

Graphical Models – Some Current Foci

Weighting paths between two nodes
(Beatrix Jones & MW 2004)

SSS: "Shotgun" Stochastic Search – annealing & rapid search over Zillions of models in regression
(Chris Hans, Adrian Dobra & MW 2004)

Latent graphs: measurement error

Data exploration (Duke team, DIG paper)
- One pathway, multiple data sets
- Transcription factor binding sites
- Graphs from literature data

Software and visualization tools (see links)

Adrian Dobra
Carlos Carvalho

Beatrix Jones
Chris Hans

Institute of Statistics and Decision Sciences
&
Computational and Applied Genomics Program

Quanli Wang Joseph Nevins Guang Yao

Duke University