Reconstructing networks of pathways via significance analysis of their intersections

Embedding biological knowledge in genomic statistical analysis

Mirko Francesconi, Daniel Remondini, Nicola Neretti, John Sedivy, Ettore Verondini, Luciano Milanesi, Leon N Cooper, Gastone Castellani

Collaboration Bologna-Brown

Brown University

Brain Research Center

Genomic Protemic Center

Theoretical Physics

Molecular Biology

Bologna University

CIG-BBB Biophysics

BioComplexity Bioinformatics

Systems Biophysics

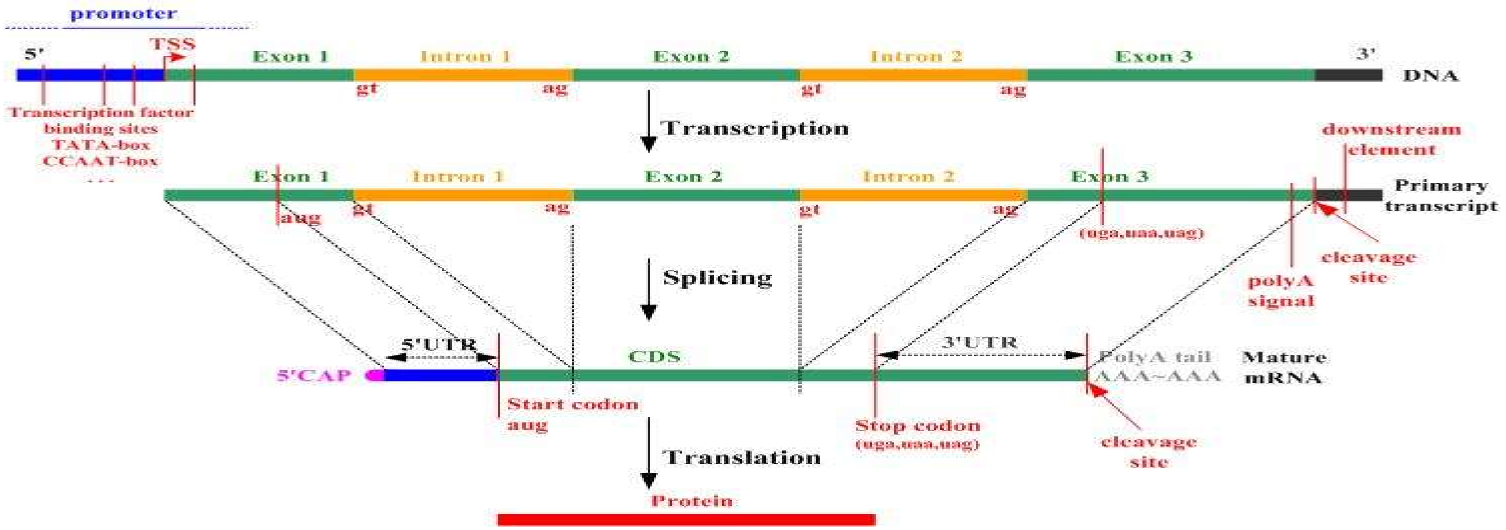
Unilever Research Center

ITB CNR Milano

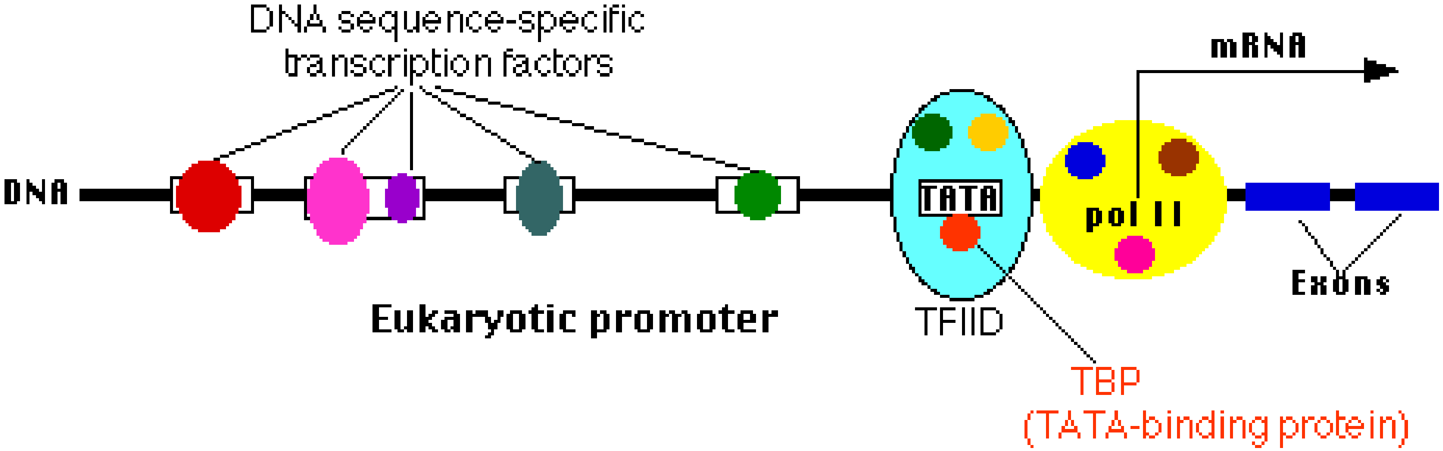
Atlantic ocean

C A N A D A
U N I T E D S T A T E S
MEXICO
Pacific Ocean
Atlantic Ocean
Gulf of Mexico
Los Angeles
San Diego
Phoenix
Dallas
Houston
San Antonio
Monterrey
Havana
Nassau
Santo Domingo
Chicago
Toronto
Ottawa
Montréal
New Castle
New York
Washington, D.C.
ICELAND
Reykjavik
NORWAY
SWEDEN
FINLAND
Oslo
Stockholm
Helsinki
St. Petersburg
Copenhagen
North Sea
UNITED KINGDOM
IRELAND
Dublin
London
Amsterdam
Brussels
Paris
FRANCE
GERMANY
Berlin
POLAND
Warsaw
BELARUS
Minsk
UKRAINE
Kiev
Prague
Vienna
Budapest
ROMANIA
Bucharest
Milan
Turin
ITALY
Rome
Naples
Palermo
SPAIN
Madrid
Barcelona
PORTUGAL
Lisbon
MOROCCO
Rabat
Algiers
Tunis
TUNISIA
Tripoli
Mediterranean Sea
GREECE
Athens
Sofia
BULGARIA
Belgrade
Istanbul
Cairo

Gene expression



- Regulation of transcription



We have generated and analyzed/ing several datasets

1) c-myc dataset (enginered rat fibroblasts)
2) TAC dataset (mouse)
3) Ewing sarcoma dataset (human)
4) Aging dataset (human time series & monozygotic twins)
5) c-myc exon array dataset (enginered rat fibroblasts)

Probe selection

- Time series (myc on and myc off data sets, cardiac hypertrophy dataset)
- Linear model with empirical bayes shrinkage of variance (limma, Bioconductor).
- Contrasts of any time point with respect to zero time point

Significance analysis:

ANOVA-MULTIPLE TEST COMPARISON

- Preprocessing for **“dimensionality reduction”** of the probeset number
- Identify genes with **significative expression levels difference** between the two conditions (perturbed and unperturbed)
- Differences are analyzed over all times
- Significance analysis applied to all probesets and **eventual correction with FDR**

c-Myc-triggered gene expression

- C-Myc encode for **transcriptional regulators** whose inappropriate expression is correlated with a wide array of human malignancies.

- Up-regulation of Myc enforces growth, antagonizes cell cycle withdrawal and differentiation, and in some situations promotes apoptosis.

- c-*myc*-/- cells reconstituted with the conditionally active, tamoxifen-specific c-Myc-estrogen receptor fusion protein (MycER) allows the fine and selective change of of c-Myc activity by Tamoxifen .

Time series experiment with **5 time points** in triplicate and 9000 probes

From the J.M. Sedivy lab O’Connel et al JBC 2004

Evaluation of global gene expression of left ventricular tissue in animal model of left ventricular hypertrophy (LVH) induced by transverse aortic constriction (TAC).

- Time series experimental design

- Measurements were done by 15 Affymetric chips at T1=0, T2=2,or T3=4 weeks after TAC.

- Each time point have been repeated with 5 replica

Genomic analysis drawbacks

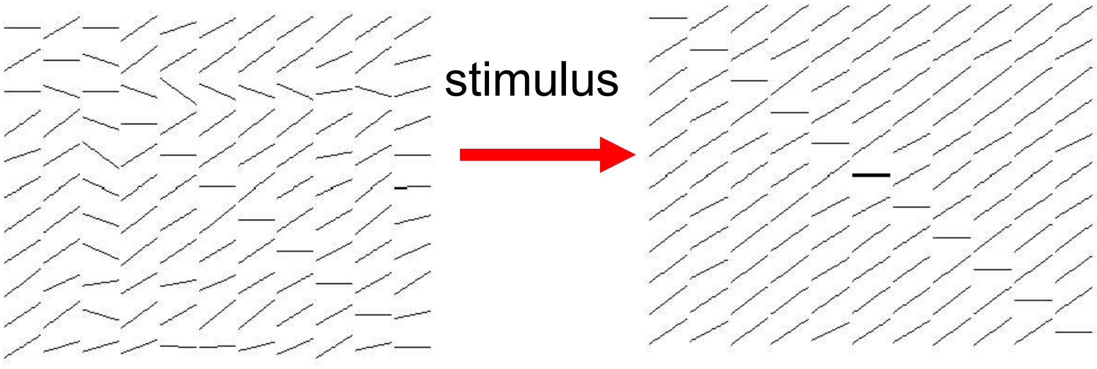
- single gene analysis is not sufficient to understand cell mechanisms undergoing experimental conditions
- cell behaviour is a complex phenomenon: several elements (e.g. genes) act together in order to generate it

Perturbation approach

- These experiments can be conceptualized as **“perturbation”** of a “basal state” (cell growth, metabolism, young phenotype, cancer phenotype etc)

- “External perturbations” like temperature in physical systems are realized by gene activation via transcription factor triggering (c-myc, dfoxo-nutrition, aging)

- Emergent properties arising in the context of perturbation theory are the so called **“phase transitions”** (superconductivity, superfluidity,etc) and **“condensation”** phenomena.



Increased order and cooperation

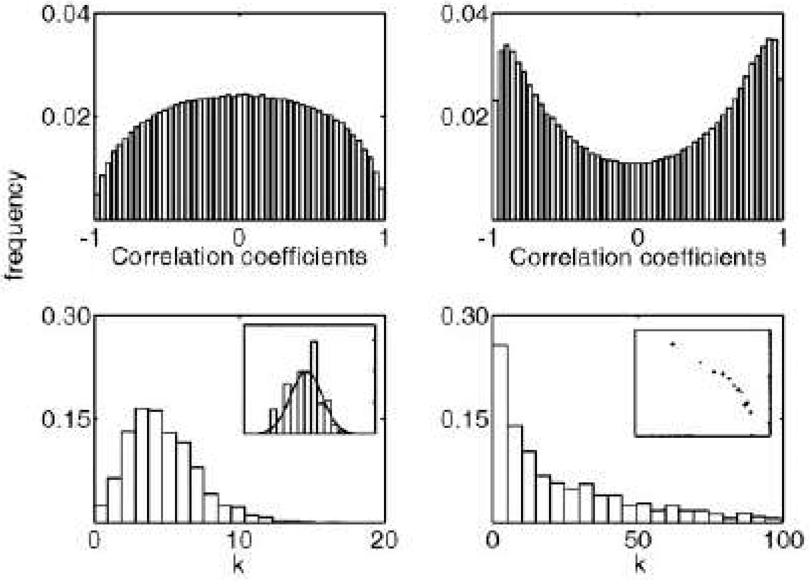
Targeting c-Myc-activated genes with a correlation method: Detection of global changes in large gene expression network dynamics

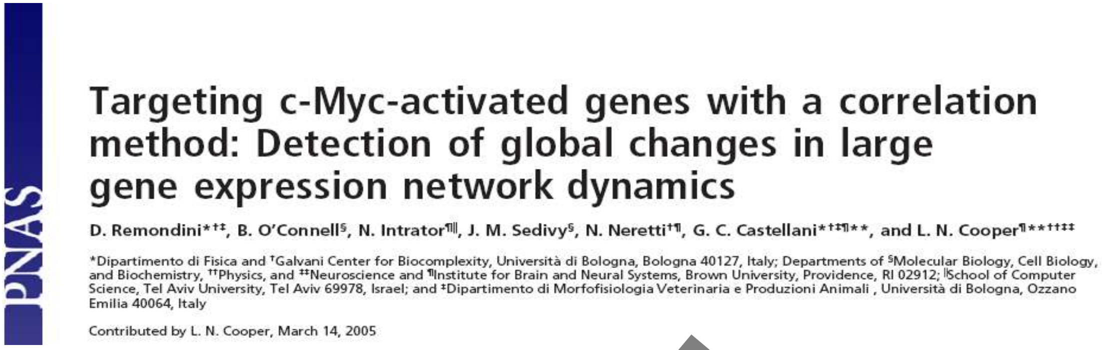
D. Remondini*[†‡], B. O'Connell[§], N. Intrator[¶‖], J. M. Sedivy[§], N. Neretti[†¶], G. C. Castellani*[†‡¶]**, and L. N. Cooper[¶]**[††‡‡]

*Dipartimento di Fisica and [†]Galvani Center for Biocomplexity, Università di Bologna, Bologna 40127, Italy; Departments of [§]Molecular Biology, Cell Biology, and Biochemistry, [††]Physics, and [‡‡]Neuroscience and [¶]Institute for Brain and Neural Systems, Brown University, Providence, RI 02912; [‖]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel; and [‡]Dipartimento di Morfofisiologia Veterinaria e Produzioni Animali , Università di Bologna, Ozzano Emilia 40064, Italy

Contributed by L. N. Cooper, March 14, 2005

This work studies the dynamics of a gene expression time series network. The network, which is obtained from the correlation of gene expressions, exhibits global dynamic properties that emerge after a cell state perturbation. The main features of this network appear to be more robust when compared with those obtained with a network obtained from a linear Markov model. In particular, the network properties strongly depend on the exact time sequence relationships between genes and are destroyed by random temporal data shuffling. We discuss in detail the problem of finding targets of the *c-myc* protooncogene, which encodes a transcriptional regulator whose inappropriate expression has been correlated with a wide array of malignancies. The data used for network construction are a time series of gene expression, collected by microarray analysis of a rat fibroblast cell line expressing a conditional Myc-estrogen receptor oncoprotein. We show that the correlation-based model can establish a clear relationship between network structure and the cascade of c-myc-activated genes.





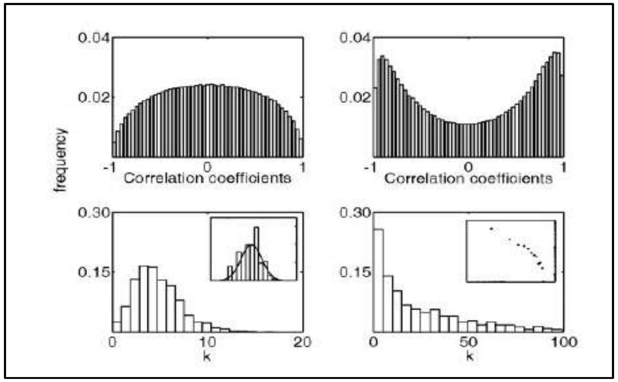
PNAS

Targeting c-Myc-activated genes with a correlation method: Detection of global changes in large gene expression network dynamics

D. Remondini[*††], B. O'Connell[§], N. Intrator[¶‖], J. M. Sedivy[§], N. Neretti[†¶], G. C. Castellani[*†‡¶**], and L. N. Cooper[¶**††‡‡]

[*]Dipartimento di Fisica and [†]Galvani Center for Biocomplexity, Università di Bologna, Bologna 40127, Italy; Departments of [§]Molecular Biology, Cell Biology, and Biochemistry, [††]Physics, and [‡‡]Neuroscience and [¶]Institute for Brain and Neural Systems, Brown University, Providence, RI 02912; [‖]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel; and [‡]Dipartimento di Morfofisiologia Veterinaria e Produzioni Animali , Università di Bologna, Ozzano Emilia 40064, Italy

Contributed by L. N. Cooper, March 14, 2005



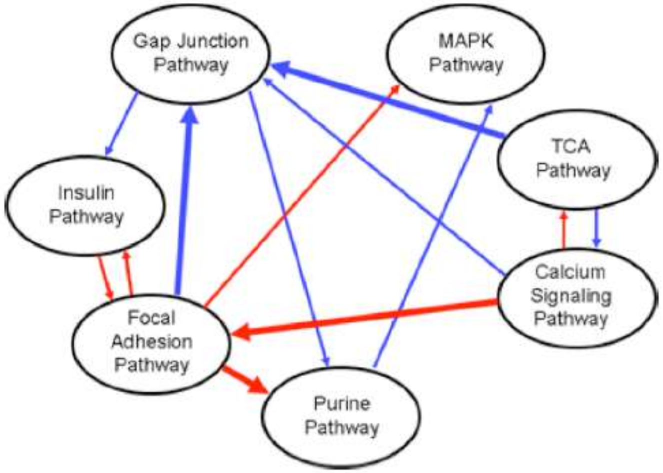


Fig. 3. Network of selected Myc-influenced pathways showing positive and negative correlations. The red and blue arrows denote positive and negative co-regulation, respectively. The thickness of the arrows is proportional to the magnitude, or absolute value, of the co-regulation. A network with these properties is called a weighted directed graph.

Multiscale correlation for co-regulation detection

•Capture **correlation profile changes at several scales** (whole array, gene family and pathways) and is informative of significative activity

•pathways synthesis into single functional forms (**Fluxes**) or index such as Subgraph Conductance.

•assessment of co-regulation between and within several pathways

•When the perturbation is conditionally switched on, the correlation between genes with a significant change in their expression level is altered on a genomic scale

We have strong indications that **a similar transition is conserved on different scales** and is indicative of **co-regulation changes**

To reduce the **dimensionality** of the problem and introduce "**a-priori biological knowledge**", we will extend this method by mapping the gene arrays data onto gene pathways and ontologies.

Castellani et al, PNAS 2001

Castellani et al, Learning and Memory 2005, BMC Bioinformatics 2007, IJCB 2007

12772–12777 | PNAS | October 23, 2001 | vol. 98 | no. 22 www.pnas.org/cgi/doi/10.1073/pnas.201404598

A biophysical model of bidirectional synaptic plasticity: Dependence on AMPA and NMDA receptors

Gastone C. Castellani*, Elizabeth M. Quinlan†, Leon N Cooper‡§¶, and Harel Z. Shouval‡‖

*Physics Department, CIG and Dimorfipa Bologna University, Bologna 40121, Italy; †Department of Biology, University of Maryland, College Park, MD 20742; and ‡Institute for Brain and Neural Systems, §Department of Neuroscience, and ¶Department of Physics, Brown University, Providence, RI 02912

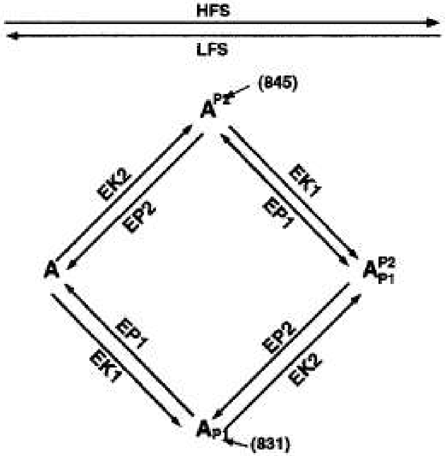


Fig. 1. An idealized model for the cycle of GluR1 phosphorylation/dephosphorylation at two sites. The model assumes two specific kinases (EK1, EK2) and two opposing specific phosphatases (EP1, EP2). It is assumed that high-frequency stimulation preferentially stimulates the activity of protein kinases, resulting in GluR1 phosphorylation, whereas low-frequency stimulation preferentially stimulates the activity of protein phosphatases, resulting in GluR1 dephosphorylation.

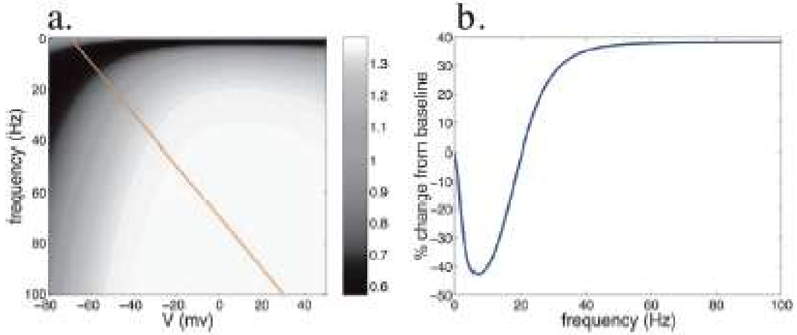


Fig. 3. Synaptic strength, measured as AMPAR conductance depicted as a function of presynaptic stimulation frequency (f) and postsynaptic membrane voltage (V). (*a*) A two-dimensional plot depicting postsynaptic membrane potential as a function of presynaptic stimulation frequency. The grey scale indicates the conductance level of the AMPAR. At low stimulation frequencies and postsynaptic voltages, the conductance is below baseline, defined as $f = 0, V = -100$. The diagonal line indicates a linear $f - V$ relation, which we assume to extract the results in *b*. (*b*) AMPAR conductance as a function of presynaptic stimulation frequency, where a linear dependence of V on f is assumed (as shown in *a*). Low-frequency stimulation induces LTD, whereas high-frequency stimulation induces LTP.

Research

A model of bidirectional synaptic plasticity: From signaling network to channel conductance

Gastone C. Castellani,[1,2,6] Elizabeth M. Quinlan,[4] Ferdinando Bersani,[1]
Leon N. Cooper,[2,3] and Harel Z. Shouval[2,5]

[1]Physics Department, DIMORFIPA, CIG, Bologna University, Bologna 40137, Italy; [2]Institute for Brain and Neural Systems and [3]Physics and Neuroscience Department, Brown University, Providence, Rhode Island 02912, USA; [4]Neuroscience and Cognitive Sciences Program, University of Maryland, College Park, Maryland 20742, USA; [5]Department of Neurobiology and Anatomy, University of Texas Medical School at Houston, Houston, Texas 77030, USA

In many regions of the brain, including the mammalian cortex, the strength of synaptic transmission can be bidirectionally regulated by cortical activity (synaptic plasticity). One line of evidence indicates that long-term synaptic potentiation (LTP) and long-term synaptic depression (LTD), correlate with the phosphorylation/dephosphorylation of sites on the α-Amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptor subunit protein GluR1. Bidirectional synaptic plasticity can be induced by different frequencies of presynaptic stimulation, but there is considerable evidence indicating that the key variable is calcium influx through postsynaptic *N*-methyl-D-aspartate (NMDA) receptors. Here, we present a biophysical model of bidirectional synaptic plasticity based on $[Ca^{2+}]$-dependent phospho/dephosphorylation of the GluR1 subunit of the AMPA receptor. The primary assumption of the model, for which there is wide experimental support, is that the postsynaptic calcium concentration, and consequent activation of calcium-dependent protein kinases and phosphatases, is the trigger for phosphorylation/dephosphorylation at GluR1 and consequent induction of LTP/LTD. We explore several different mathematical approaches, all of them based on mass-action assumptions. First, we use a first order approach, in which transition rates are functions of an activator, in this case calcium. Second, we adopt the Michaelis-Menten approach with different assumptions about the signal transduction cascades, ranging from abstract to more detailed and biologically plausible models. Despite the different assumptions made in each model, in each case, LTD is induced by a moderate increase in postsynaptic calcium and LTP is induced by high Ca^{2+} concentration.

12:000–000 ©2005 by Cold Spring Harbor Laboratory Press ISSN 1072-0502/05; www.learnmem.org

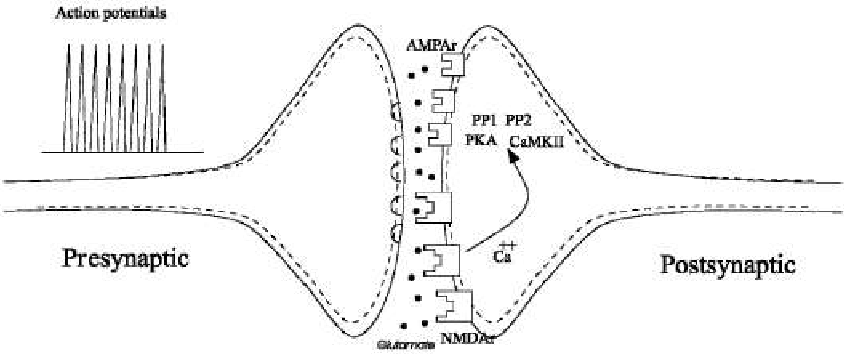
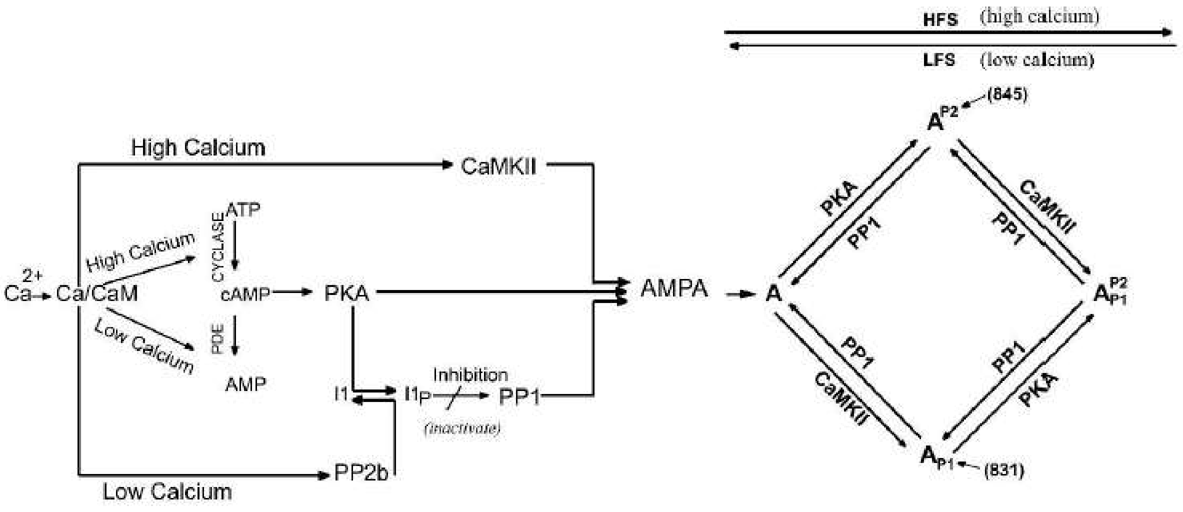


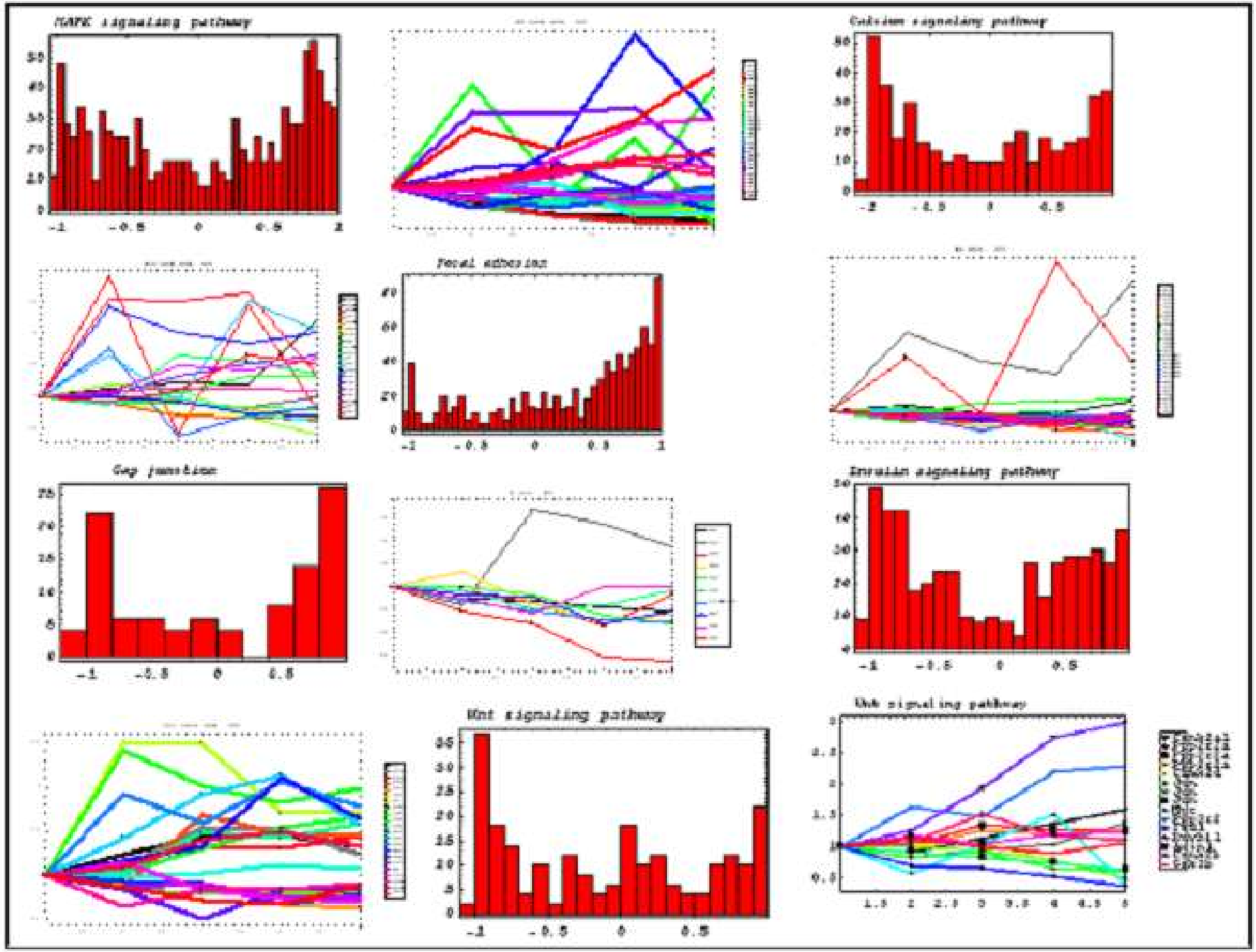
Figure 1. A schematic of an excitatory glutamatergic synapse. Action

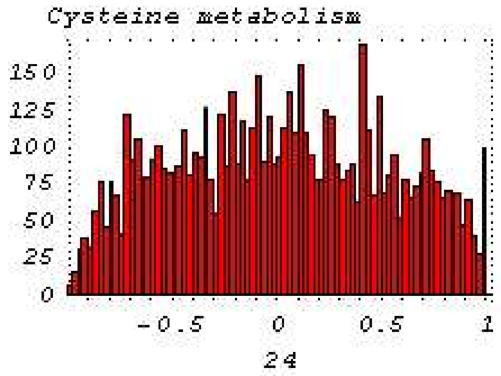


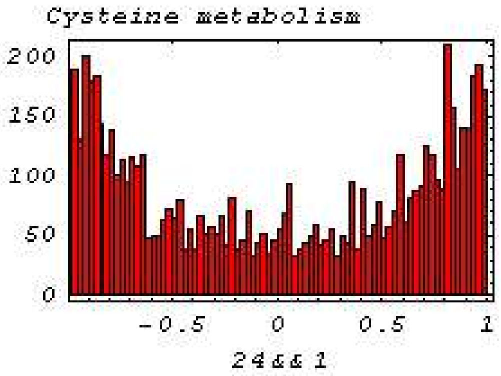
(Adapted from Lisman 89)

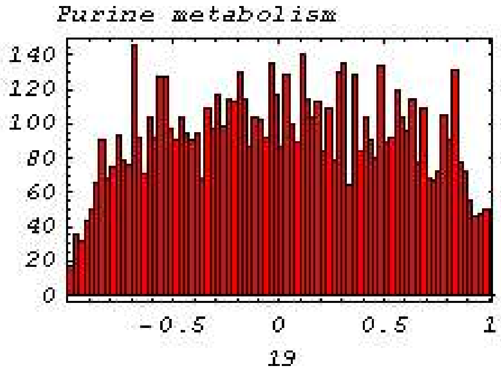
Figure 6. Calcium-dependent activity of the kinase/phosphatase network. *(Left)* Introduction of

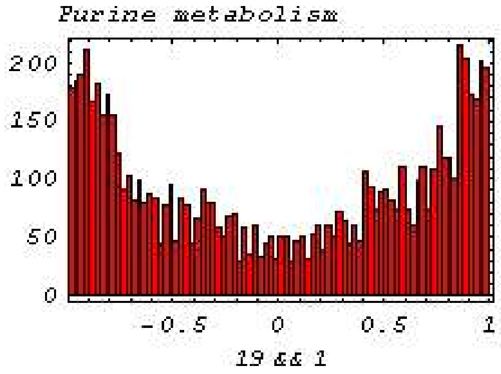
Multiscale Correlation Model:
c-Myc results

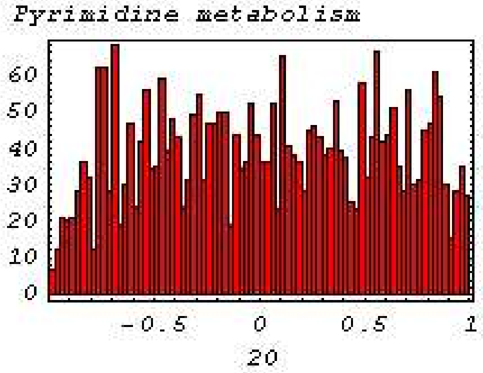


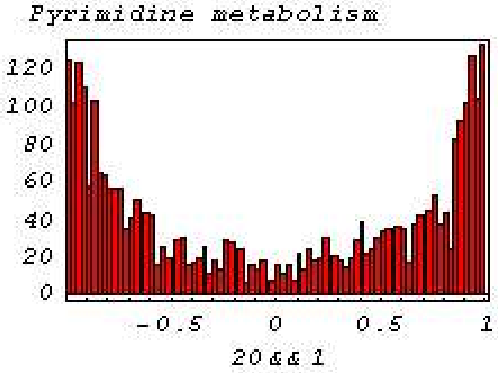

Cysteine metabolism
150
125
100
75
50
25
0
-0.5
0
0.5
1
24


Cysteine metabolism
200
150
100
50
0
-0.5
0
0.5
1
24&& 1


Purine metabolism
140
120
100
80
60
40
20
0
-0.5
0
0.5
1
19

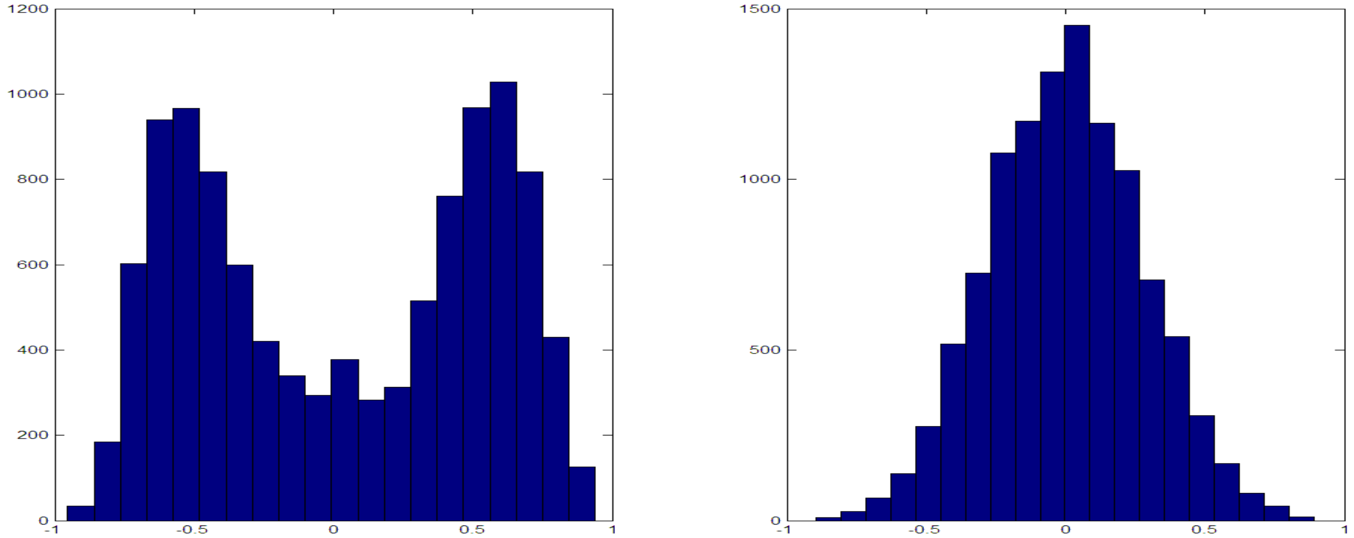

Purine metabolism
200
150
100
50
0
-0.5
0
0.5
1
19 && 1


Pyrimidine metabolism
60
50
40
30
20
10
0
-0.5
0
0.5
1
20

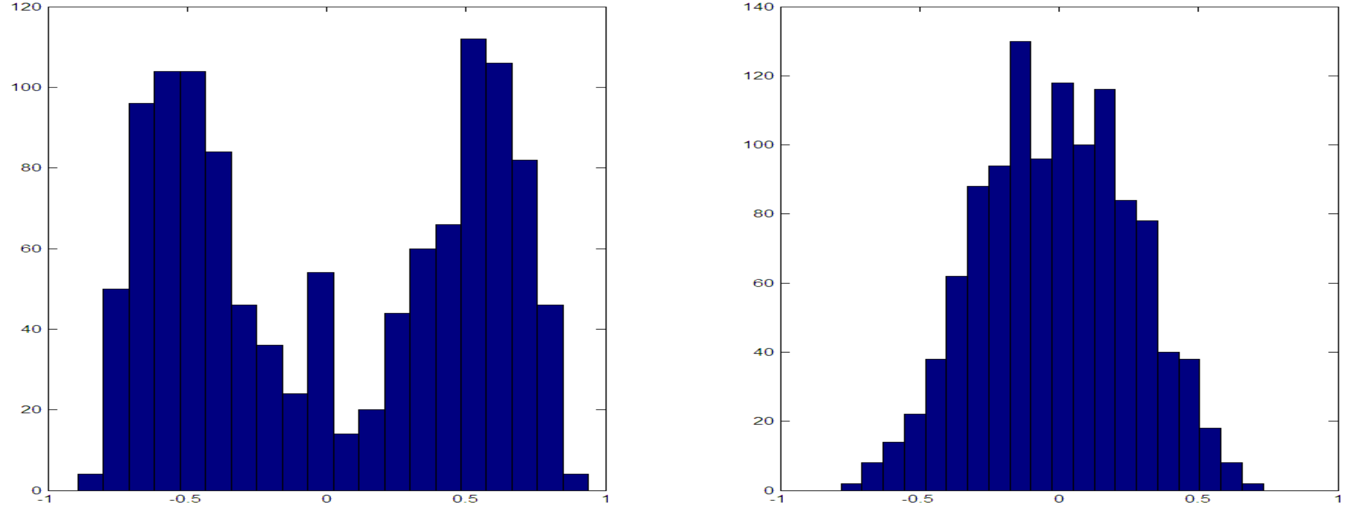

Pyrimidine metabolism
120
100
80
60
40
20
0
-0.5
0
0.5
1
20&& 1

Multiscale Correlation Model: human aging results

Protein Binding



Plasma Membrane



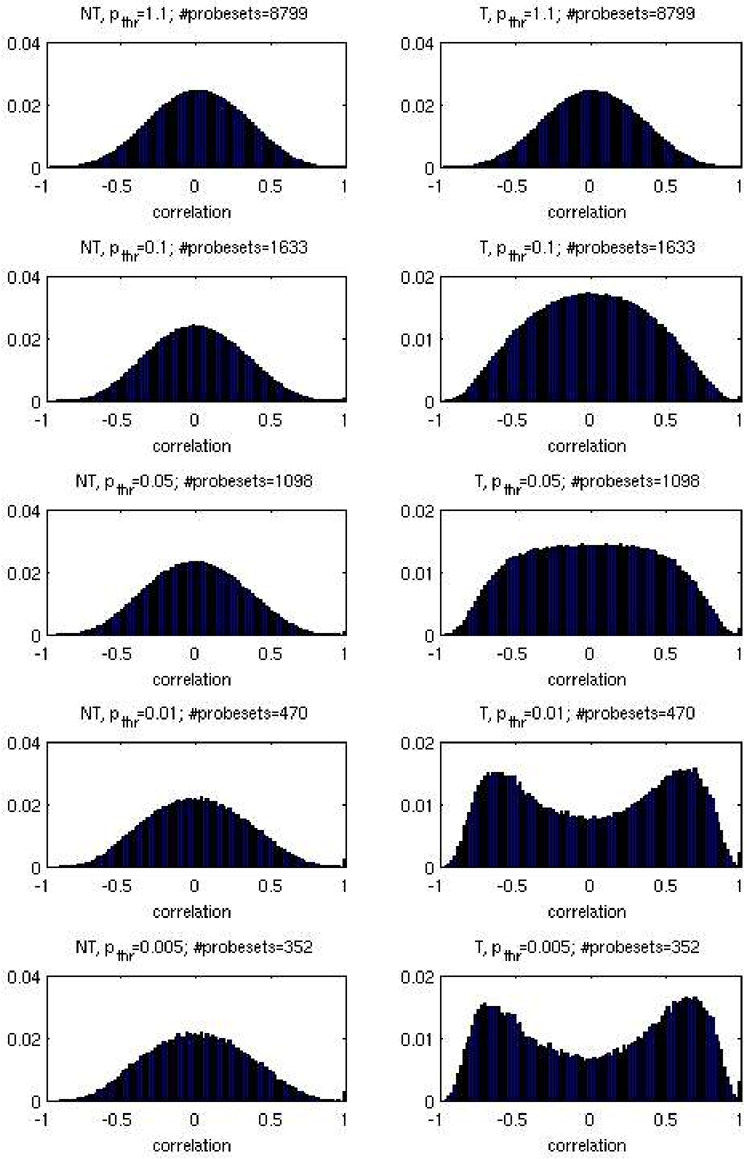
Castellani et al

International

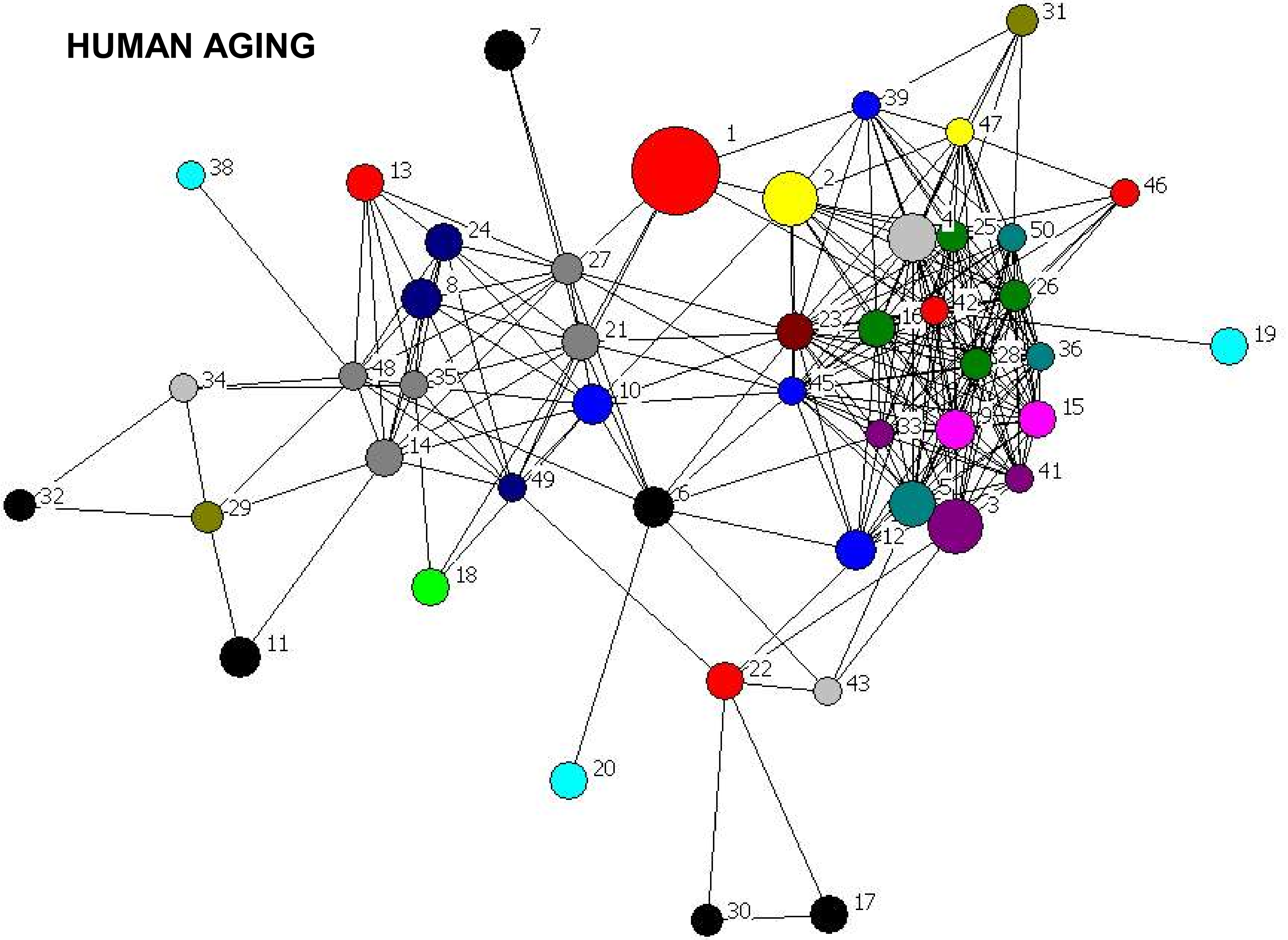
Journal of Chaos

and Bifurcation

2007



HUMAN AGING


31
7
39
47
1
38
13
2
46
24
4
25
50
27
26
8
42
16
23
19
21
28
36
48
34
35
45
10
15
9
33
14
41
49
6
5
3
32
29
12
18
11
22
43
20
17
30

1 PPAR SigPath

2 Adipocytokine SigPath

3 Inositol phosphate Met

4 Jak-STAT SigPath

5 Phosphatidylinositol SigSyst

6 Purine metabolism

7 Glyo and Dicarbo xylate Met

8 Cysteine metabolism

9 B cell receptor SigPath

10 Glycolysis-Gluconeogenesis

11 Styrene degradation

12 Long-term depression

26 Apoptosis

27 Carbon fixation

28 Colorectal cancer

29 Glutathione metabolism

30 γ-ExaCloCE Degr

31 Antigen ProcAndPres

32 Cyanoamino Ac Met

33 Gap junction

34 Taur HypoTaur Met

35 ALA-ASP Met

36 Leuk tr-e migration

37 Atrazine Deg

13 Alkaloid Bios I

14 Tyrosine Met

15 mTOR SigPath

16 Fc ε RI SigPath

17 Bisphenol A Degr

18 Val Leu ILeu Bios

19 Complement and Coag

20 Pyrimidine metabolism

21 Pyruvate metabolism

22 Benzoate degradation

13 Type II Diab Mell

14 PhenylAla Met

15 T cell Reec SigPath

38 Nitrogen metabolism

39 Hematopoietic cell lineage

40 Glycan STR-Bios 1

41 VEGF SigPath

42 Focal adhesion

43 Nicotinate and nicotinamide metabolism

44 Ribosome

45 Insulin SigPath

46 Cell cycle

47 Cytk-Citk RecInt

48 Glutamate Met

49 Propanoate Met

50 Toll-like Rec SigPath

“Databases” like KEGG have also an interesting **network structure**, it is possible that biologically relevant informations can be retrieved from the **topological structure** of **nodes** (pathways) and **edges** (common genes between two pathways)

The most relevant edges can be **focal areas** from which biological messages are spread throughout the network (like the **hubs** for the nodes)

Pathway network analysis

Given significant nodes and edges, the **pathway network** can be **reconstructed**.

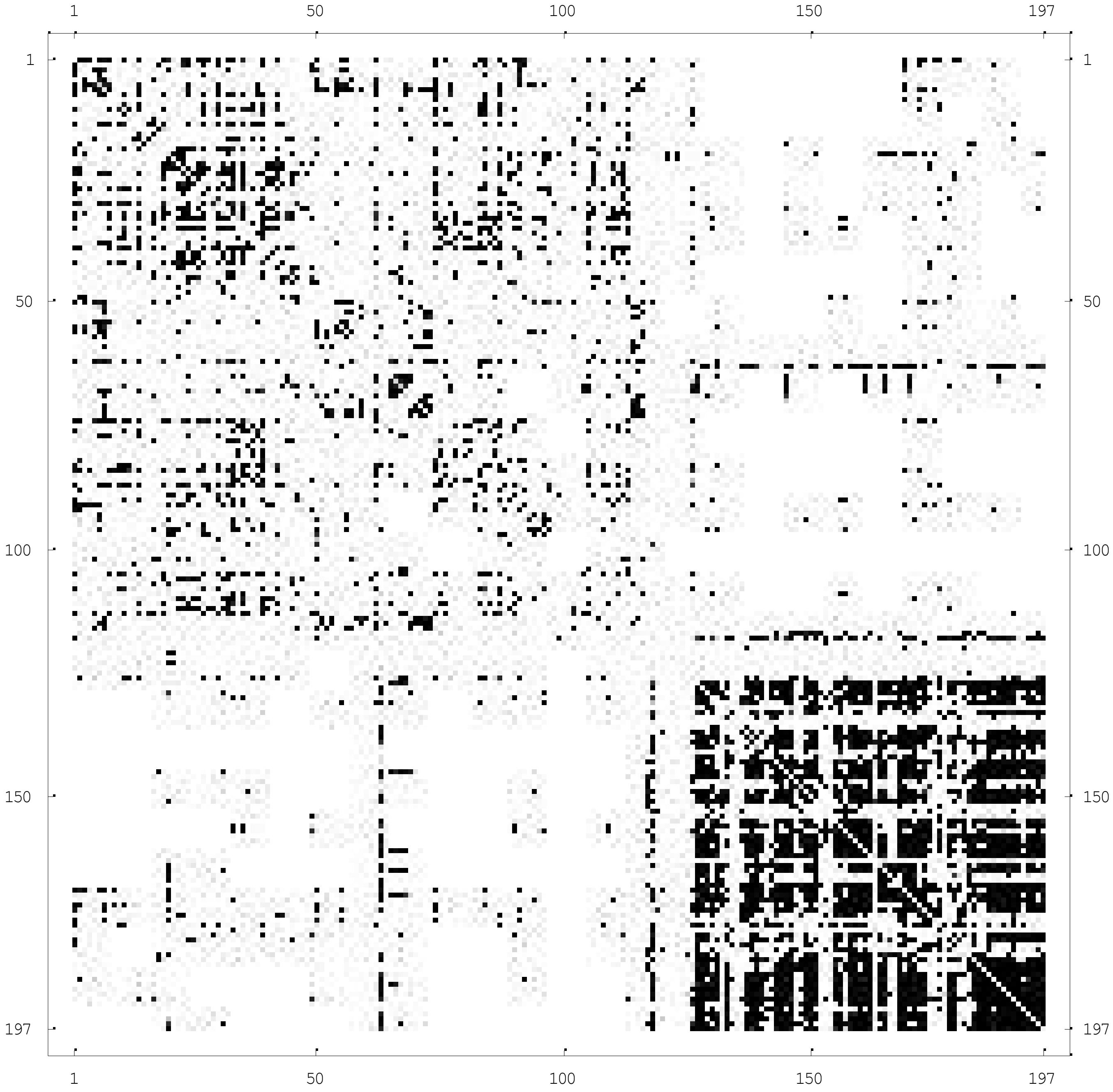
Edges and nodes can be **ranked** based on their **centrality** in the network (connectivity degree and betweenness)

Betweenness centrality

Betweenness centrality is a very interesting parameter because:

- it can be calculated both for nodes and edges

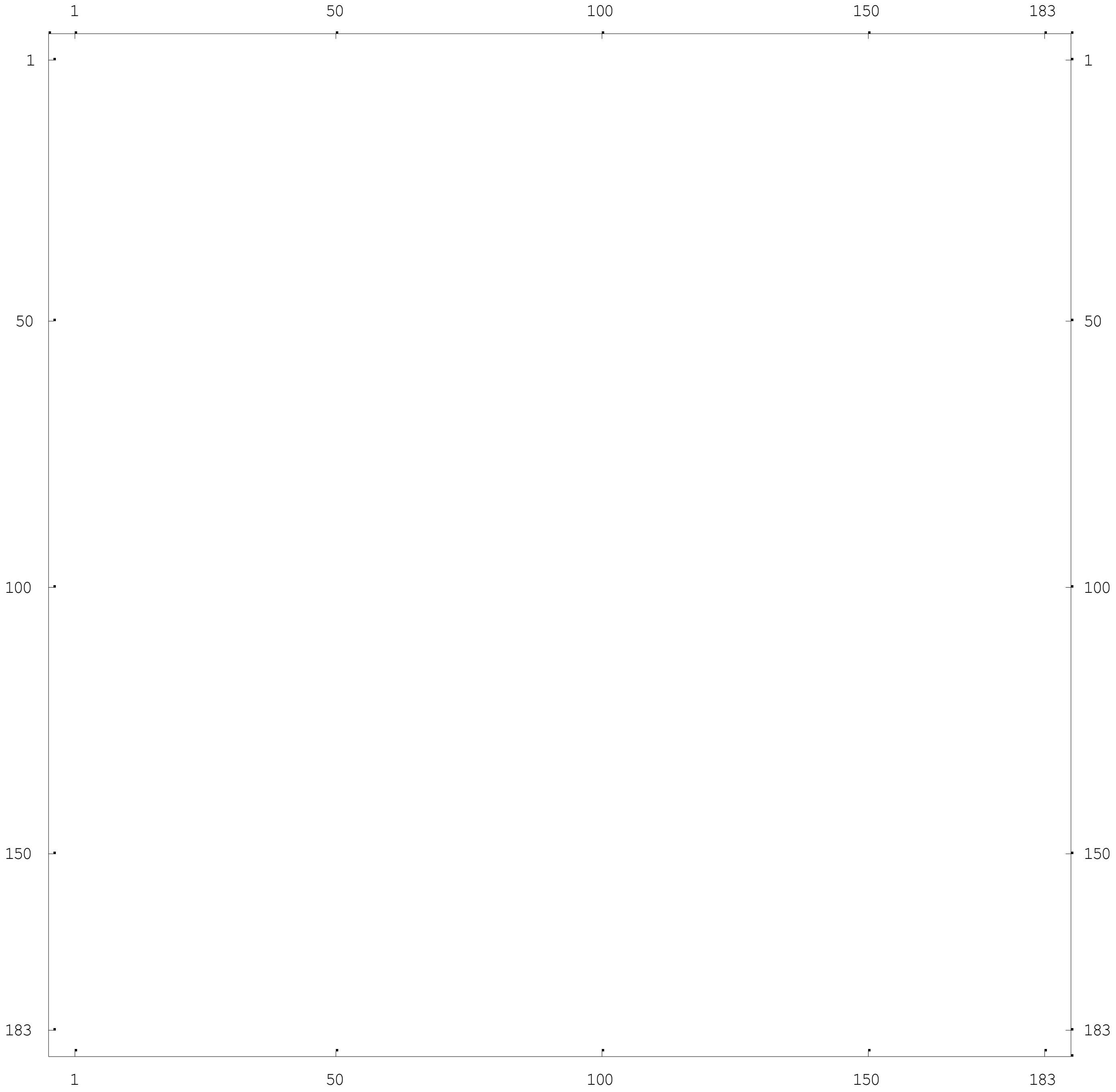
- it is a measure of the possible **information flow** through that element, thus if it is affected by experimental conditions it is very likely that such **perturbation** can **spread** to the whole system more easily



Hsa

KEGG

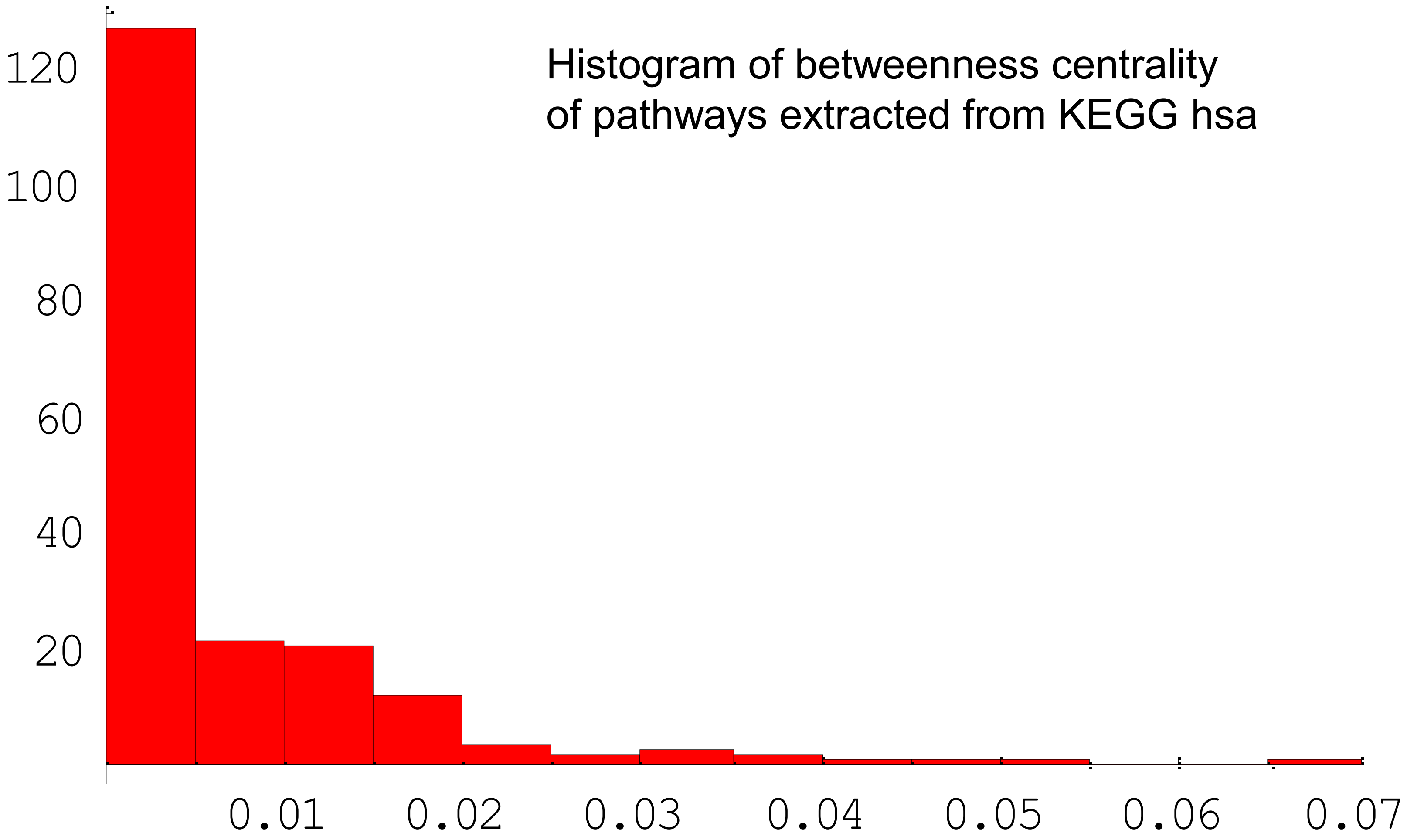
complete



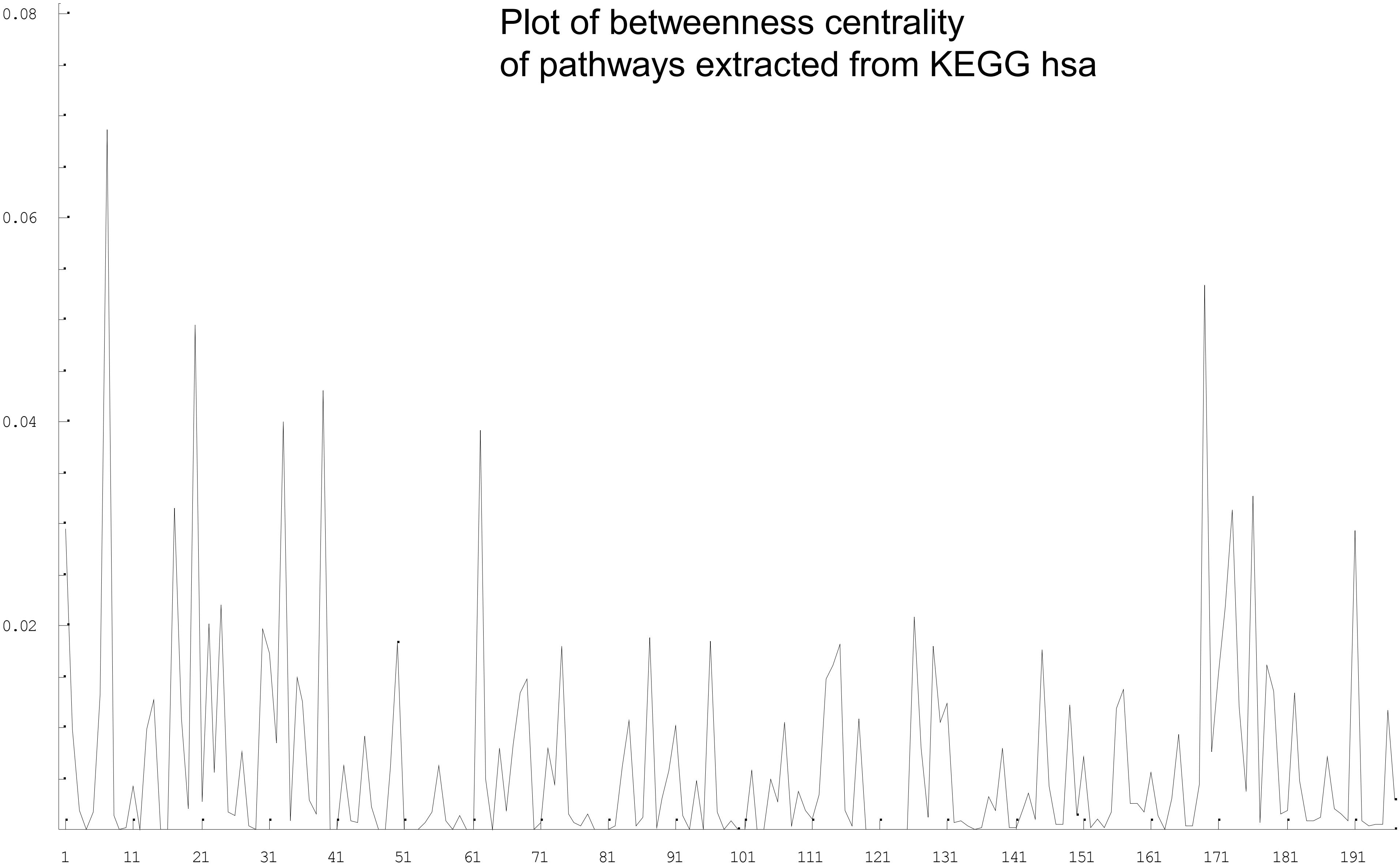
rno

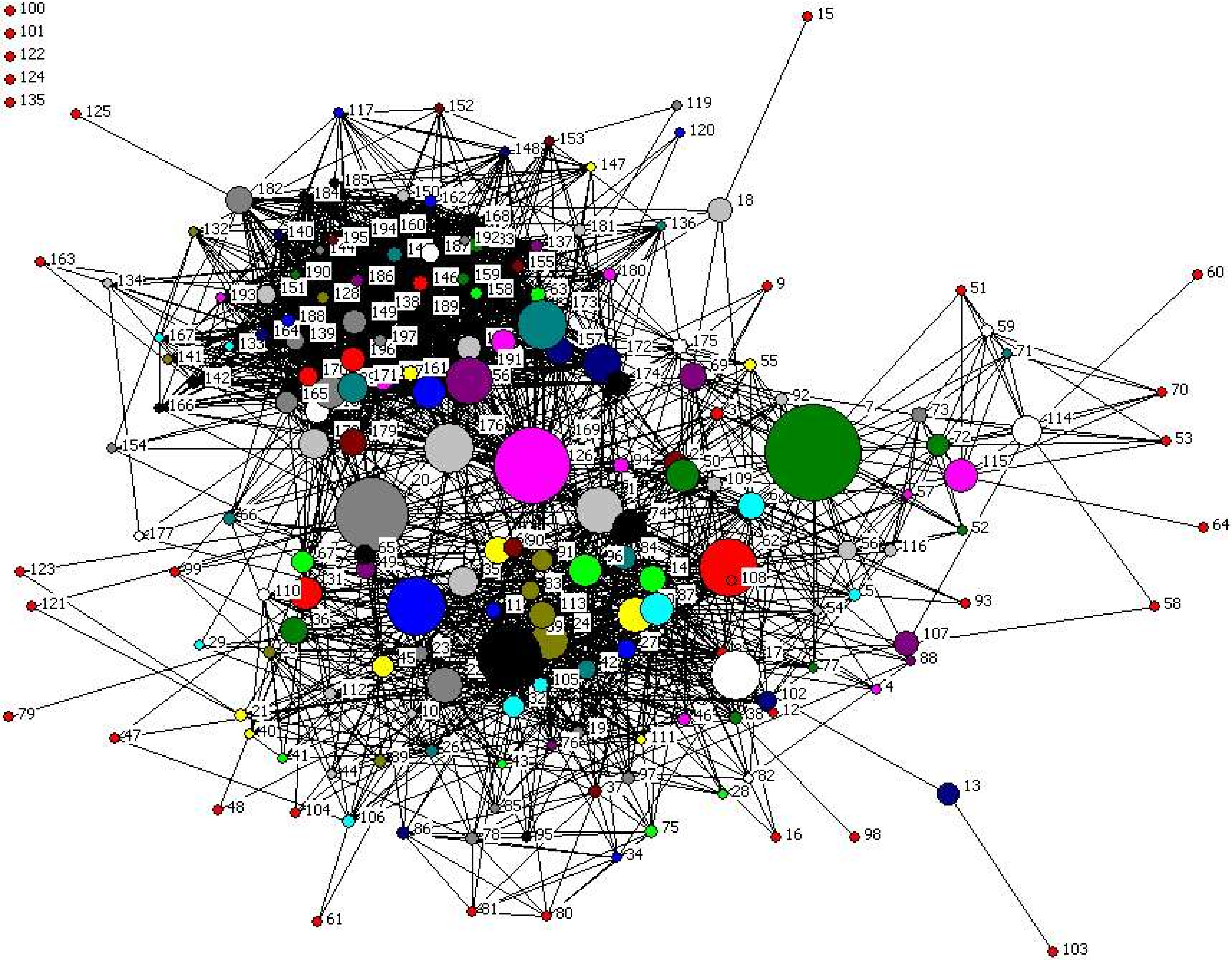
KEGG

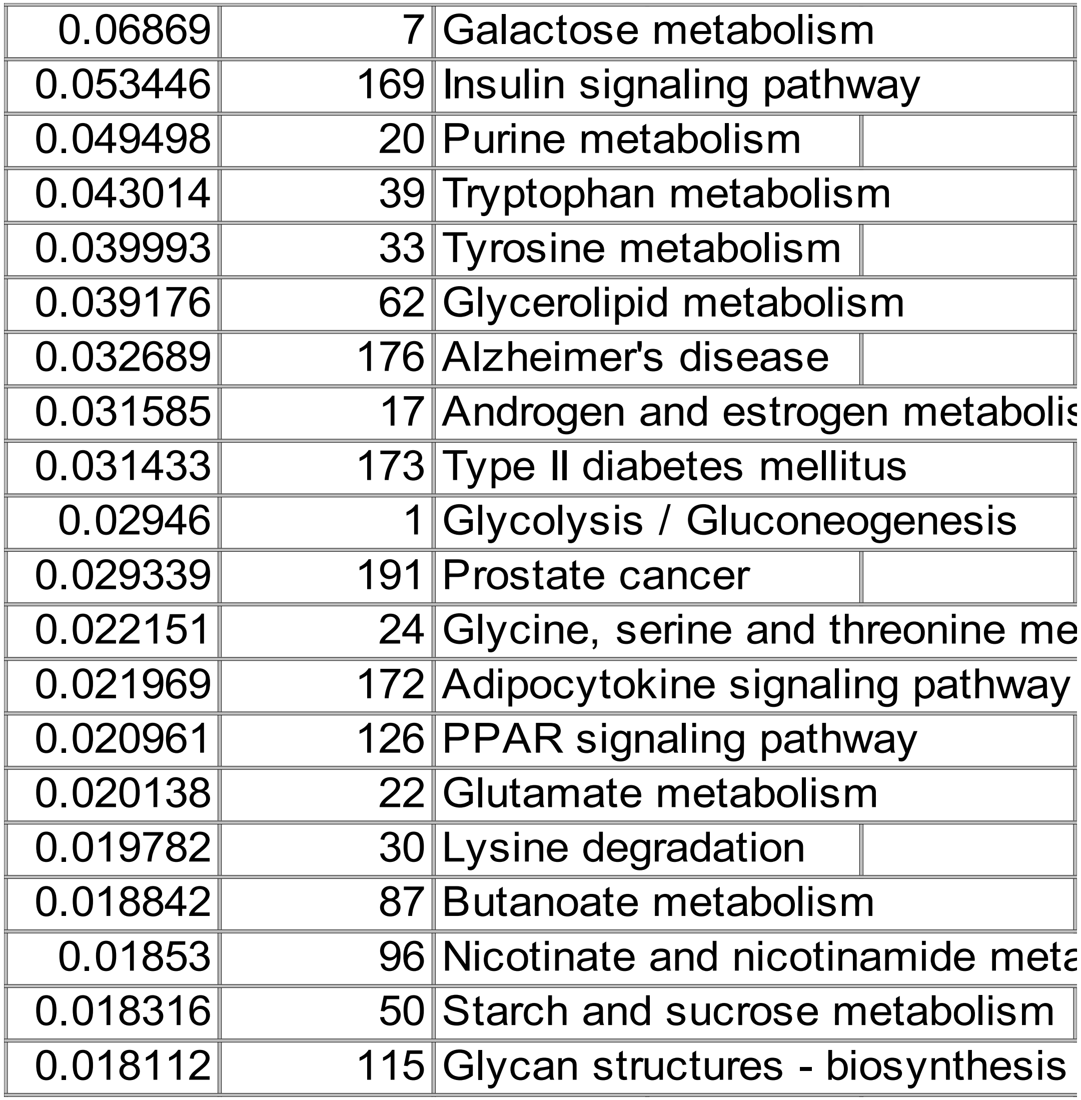
complete


Histogram of betweenness centrality
of pathways extracted from KEGG hsa
120
100
80
60
40
20
0.01
0.02
0.03
0.04
0.05
0.06
0.07

Plot of betweenness centrality
of pathways extracted from KEGG hsa




100
101
122
124
135
125
15
117
152
119
120
153
148
147
185
182
184
150
162
18
168
181
136
132
140
195
194
160
192
133
137
144
187
163
134
190
186
146
159
155
180
193
151
128
158
63
173
9
138
189
149
188
167
164
139
197
157
175
51
59
60
133
196
191
172
55
71
141
170
171
161
56
174
69
70
142
166
165
92
7
73
114
154
179
176
169
72
53
126
94
50
115
20
109
57
64
66
177
74
52
67
65
90
91
96
84
62
56
116
123
99
49
14
108
121
110
31
35
83
113
87
93
58
11
54
36
99
24
107
29
25
45
23
27
17
88
42
77
4
105
102
112
21
79
12
47
40
32
46
38
10
19
111
41
89
26
43
76
97
13
44
82
28
37
48
104
85
75
16
98
106
86
78
95
34
81
80
61
103



0.06869	7	Galactose metabolism
0.053446	169	Insulin signaling pathway
0.049498	20	Purine metabolism
0.043014	39	Tryptophan metabolism
0.039993	33	Tyrosine metabolism
0.039176	62	Glycerolipid metabolism
0.032689	176	Alzheimer's disease
0.031585	17	Androgen and estrogen metabolis
0.031433	173	Type II diabetes mellitus
0.02946	1	Glycolysis / Gluconeogenesis
0.029339	191	Prostate cancer
0.022151	24	Glycine, serine and threonine me
0.021969	172	Adipocytokine signaling pathway
0.020961	126	PPAR signaling pathway
0.020138	22	Glutamate metabolism
0.019782	30	Lysine degradation
0.018842	87	Butanoate metabolism
0.01853	96	Nicotinate and nicotinamide meta
0.018316	50	Starch and sucrose metabolism
0.018112	115	Glycan structures - biosynthesis

Top 20 pathways
extracted
from KEGG Database
ranked for their
betwennes centrality

Pathway significance analysis

Node (pathway) or edge (intersection) significance analysis can be performed by considering the total number of genes represented in KEGG and the total number of statistically significant genes, compared with the significant genes found in a node or edge and their total number of elements (e.g. by a test based on the **hypergeometric distribution**)

2.4 Fisher's exact test (Draghici *et al.*, 2003)

We consider that there are N single-symbol-annotated genes on the microarray (replicates were averaged by calculating the mean), which are either significantly differentially expressed (S) or not (F), and either belong to a pre-defined pathway list (P) or not (NP), see Table 2. If we pick randomly P genes, we would like to estimate the probability of having exactly α genes in S. The p-value of having α genes or fewer in S can be calculated by summing the probabilities of a random list of K genes having $1, 2, \ldots, \alpha$ genes in S:

$$p = 1 - \sum_{i=0}^{\alpha} \frac{\binom{S}{i}\binom{F}{P-i}}{\binom{N}{P}} \tag{1}$$

This is a one-sided test in which the P values correspond to over-represented lists of genes.

A review about similar current tools used for group testing on the level of Gene Ontology (GO) terms was given by Khatri and Draghici (2005).

	0	**1**	**Totals**
1	a	b	a+b
0	c	d	c+d
Totals	a+c	b+d	n

$$\mu_{ij} = \frac{T_{R_i} \times T_{C_j}}{T_G}$$

Null table is constructed

by the multinomial

distribution and then

tested by a χ^2 test

Fisher exact test for a 2x2 contingency table

	0	**1**	**Totals**
1	a	b	a+b
0	c	d	c+d
Totals	a+c	b+d	n

The probability
Is due by the
Hypergeometric
distribution

$$\frac{\dfrac{(a+c)!}{a!c!} \times \dfrac{(b+d)!}{b!d!}}{\dfrac{n!}{(a+b)!(c+d)!}}$$

Pathways and their intersections significance analysis

- calculated considering the hypergeometric distribution:

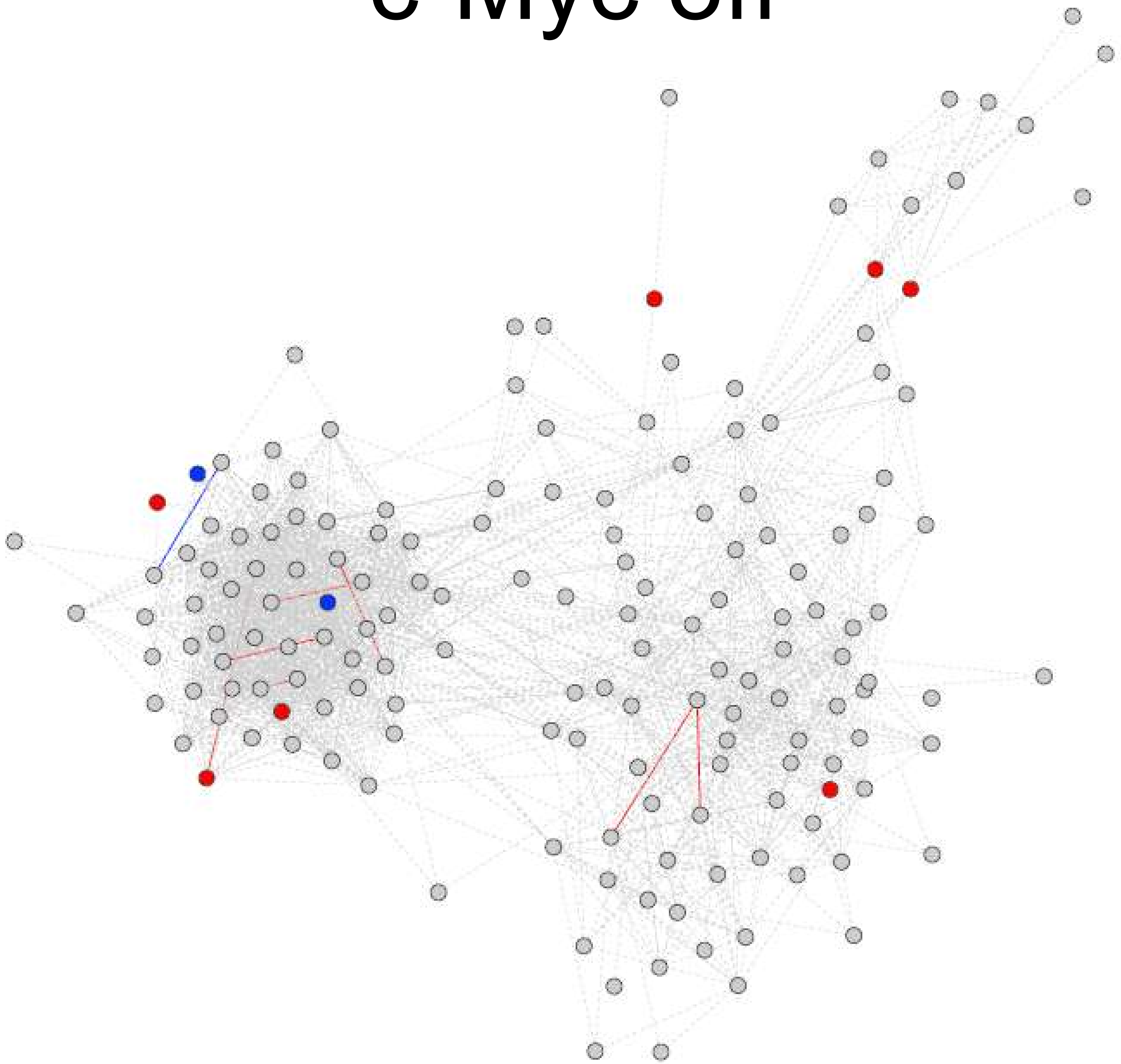
$$p(x) = choose(m, x)\ choose(n, k-x) / choose(m+n, k)$$

- where
  - p= probability.
  - x = number of significant probes in a pathway (or intersection)
  - m = total number of significant probes.
  - n = total number of non significant probes.
  - k = number of probes in a pathway.
- P <0.05 was considered as significant

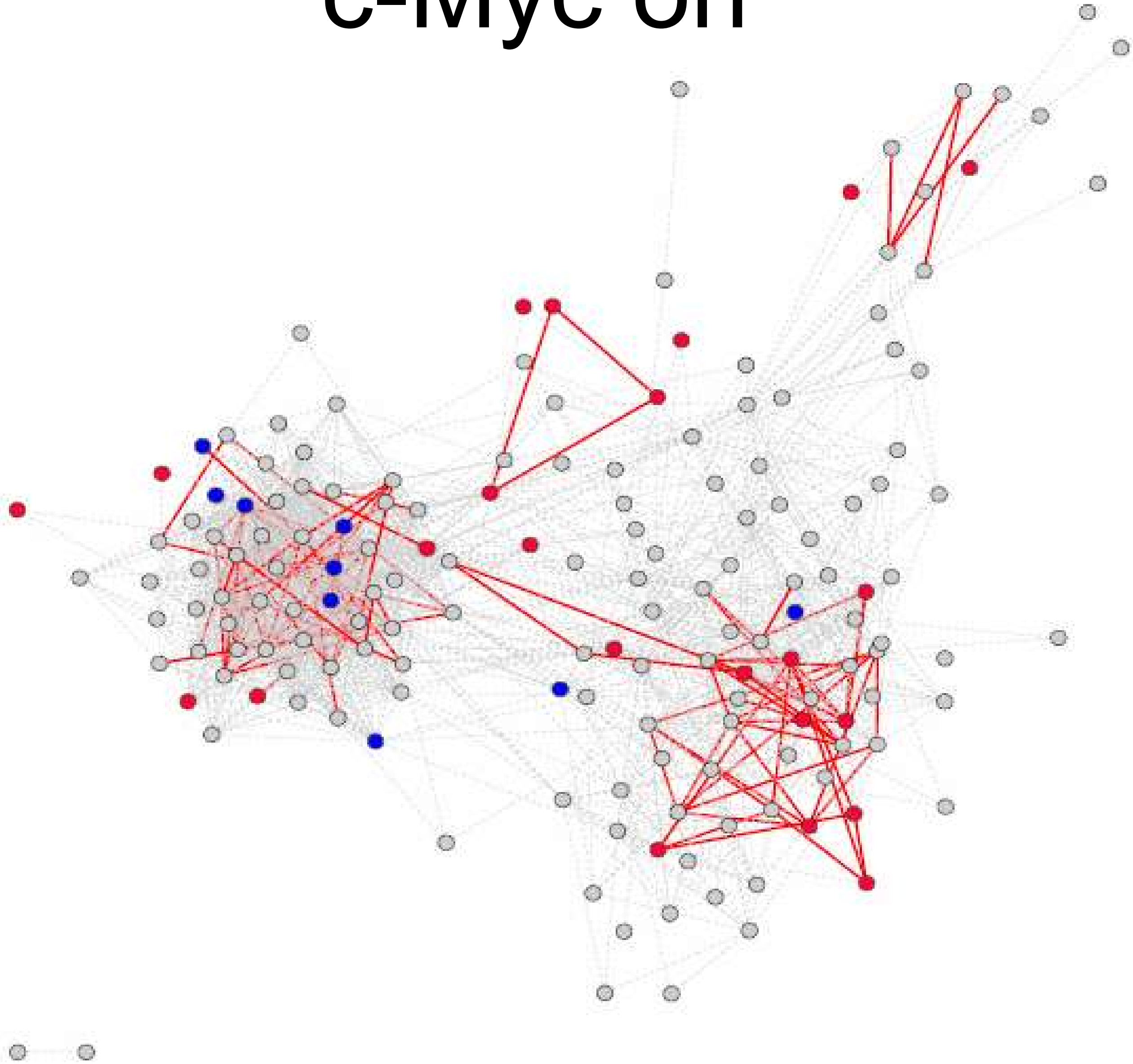
Network representation

- Significantly underrepresented: (-1) 🟦
- Significantly overrepresented: 1 🟥
- Not significant: 0 ⬜

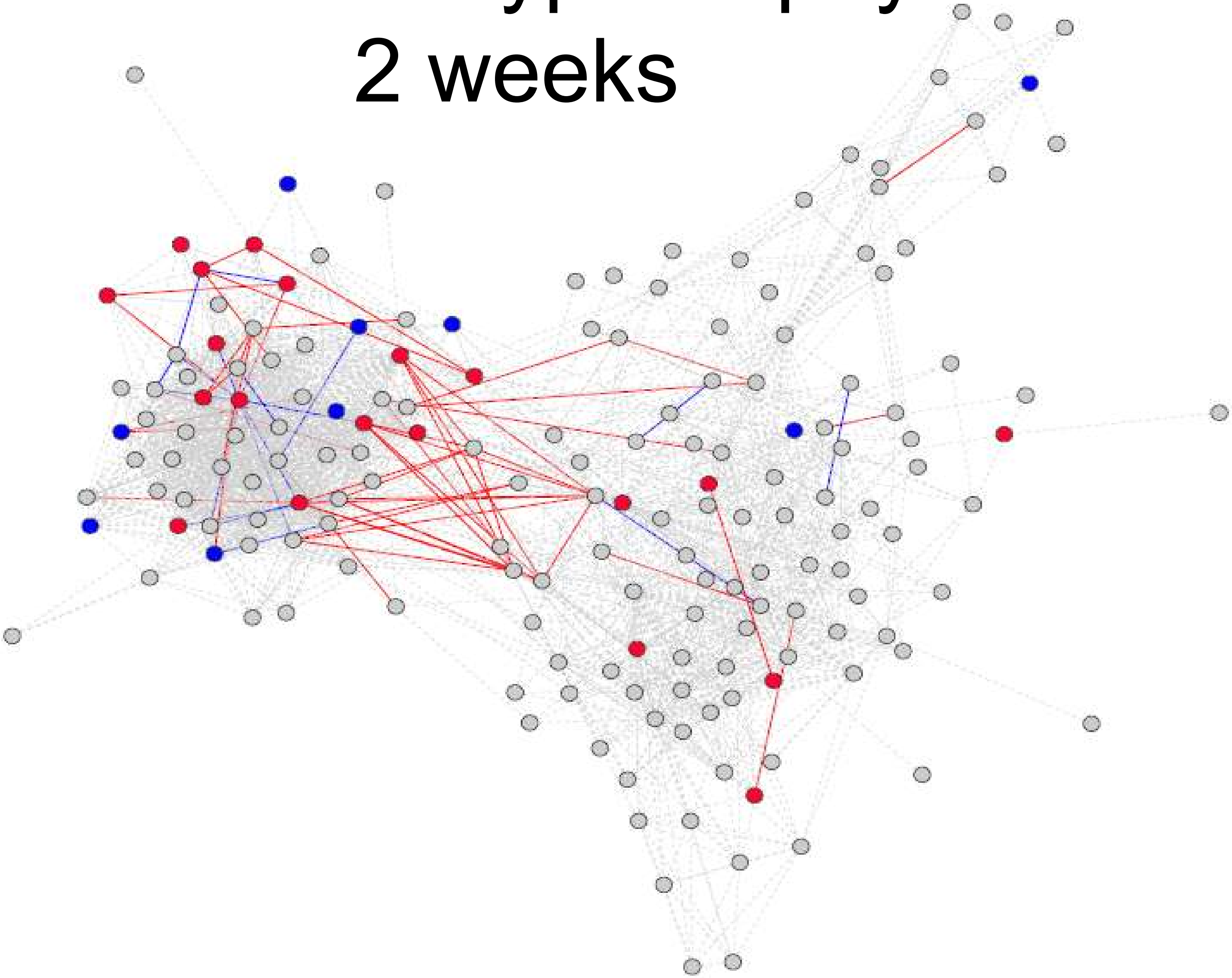
c-Myc off



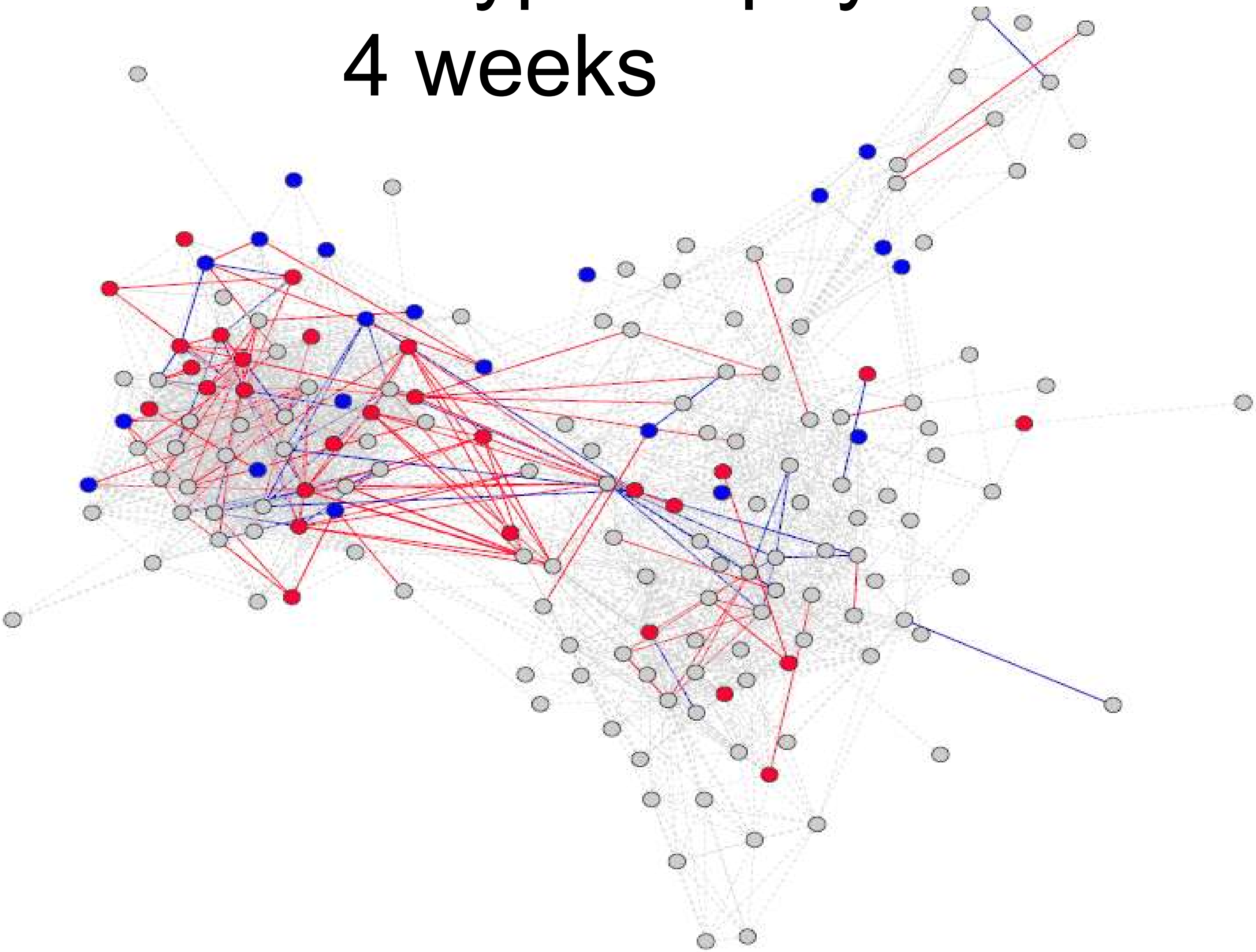
c-Myc on



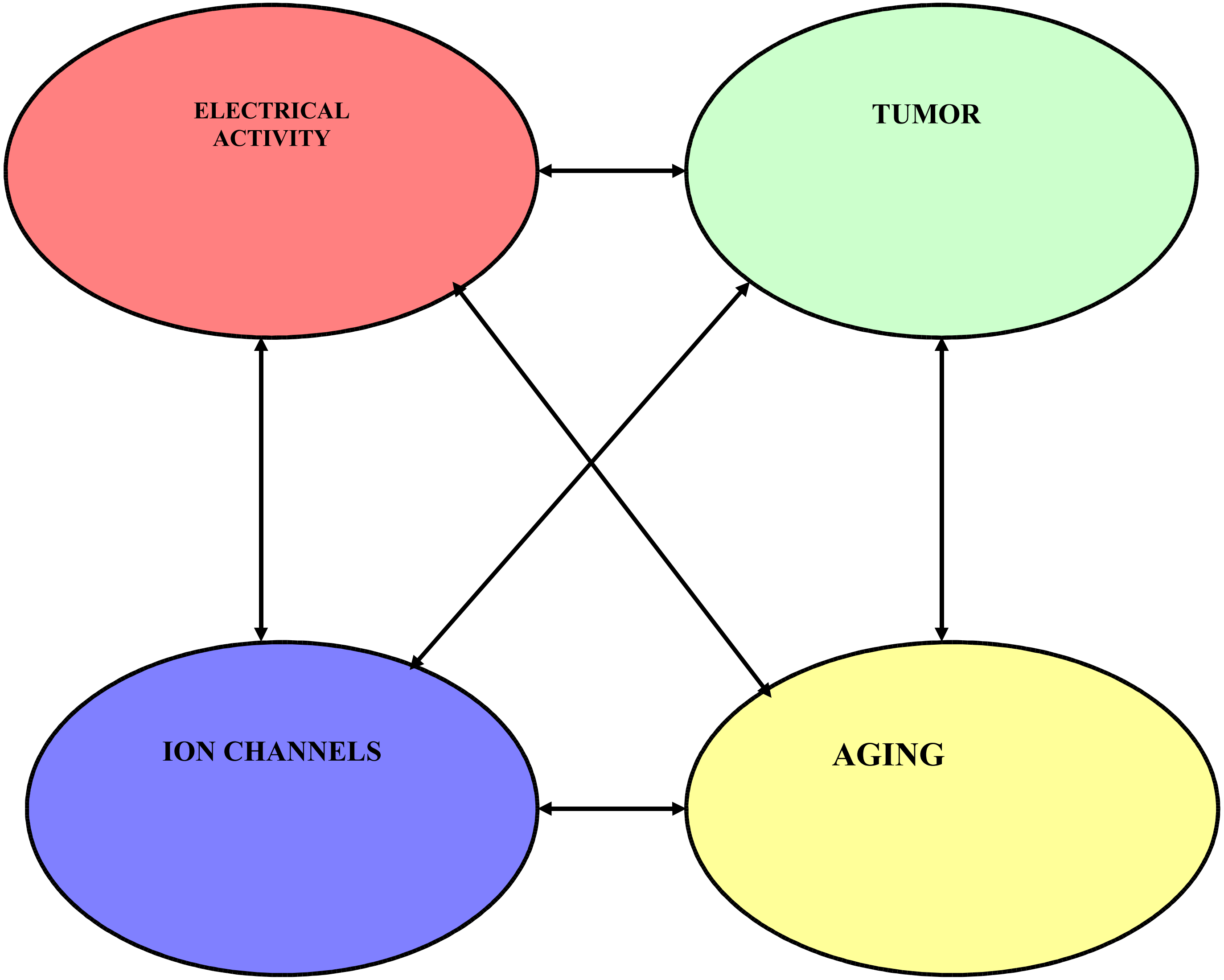
cardiac hypertrophy 2 weeks



cardiac hypertrophy
4 weeks



CONCLUSIONS AND PERSPECTIVES

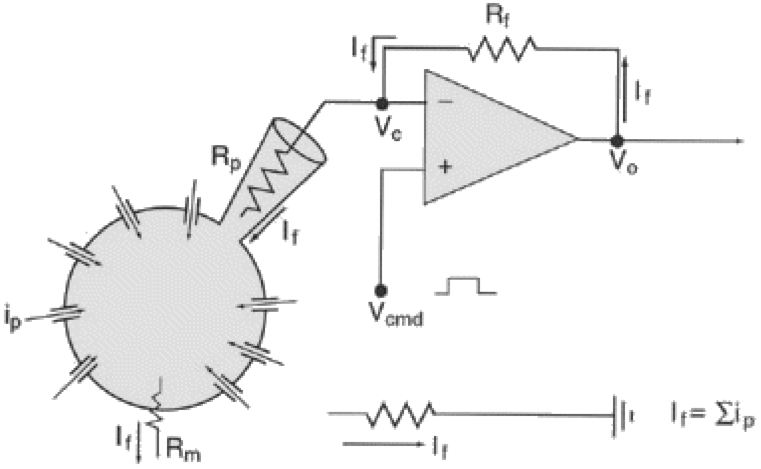


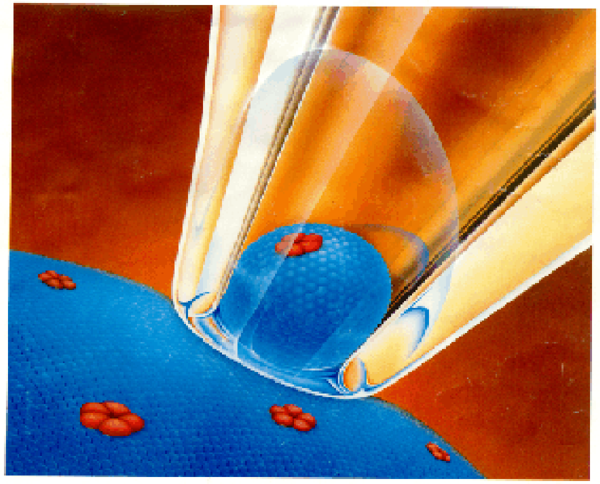
ELECTRICAL ACTIVITY

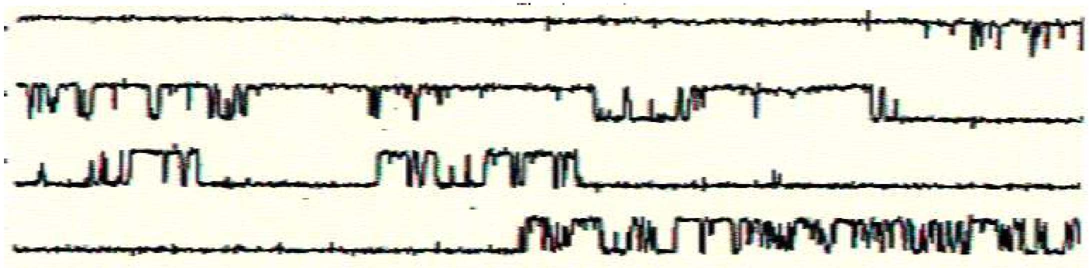
TUMOR

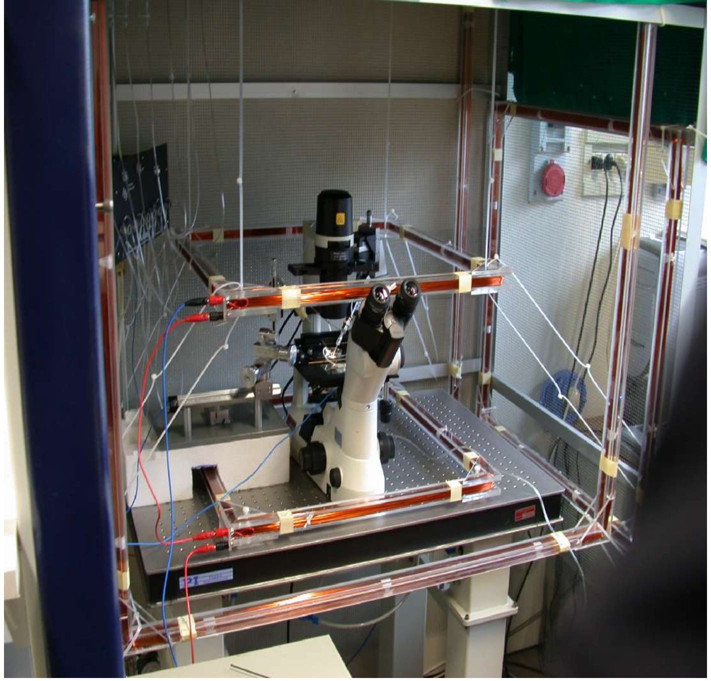
ION CHANNELS

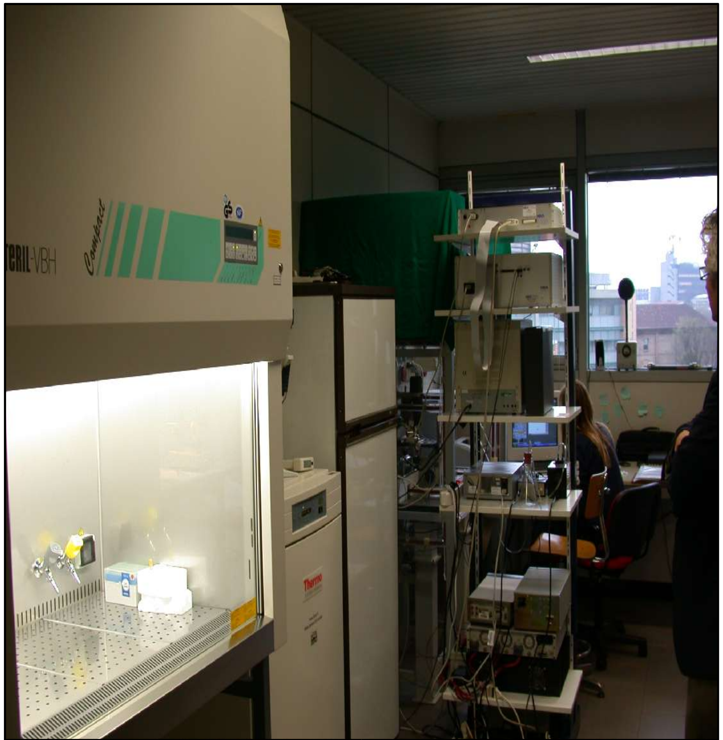
AGING


R_f
I_f
I_f
V_c
V_o
R_p
I_f
i_p
V_{cmd}
I_f
R_m
$I_f = \Sigma i_p$
I_f

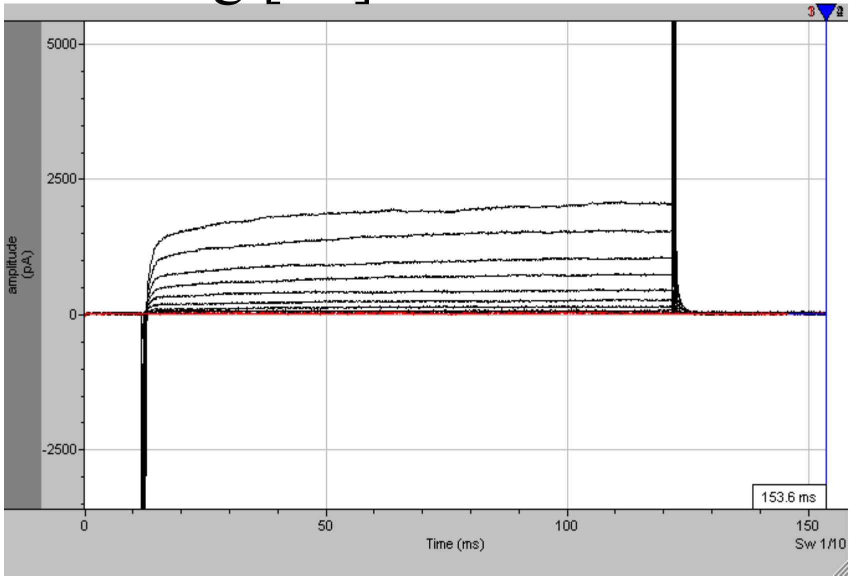




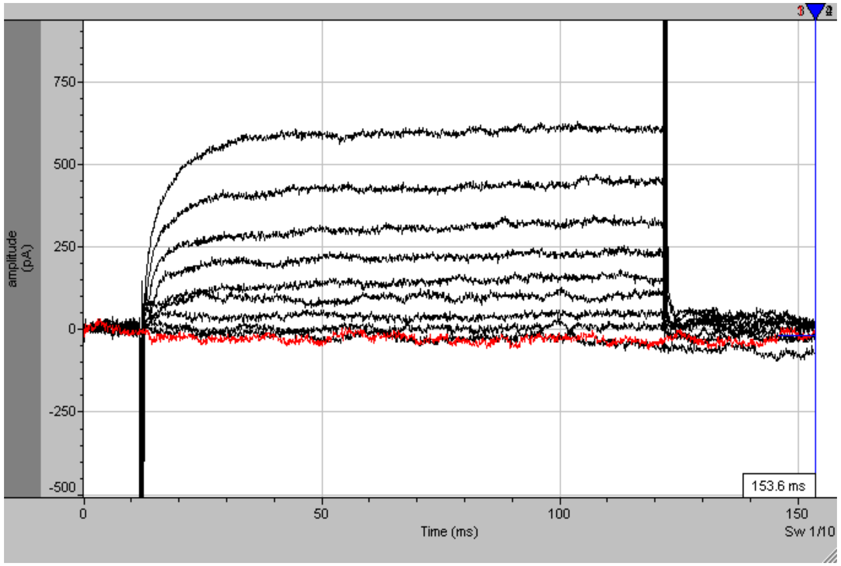



ERIL-VBH
Compact
Thermo

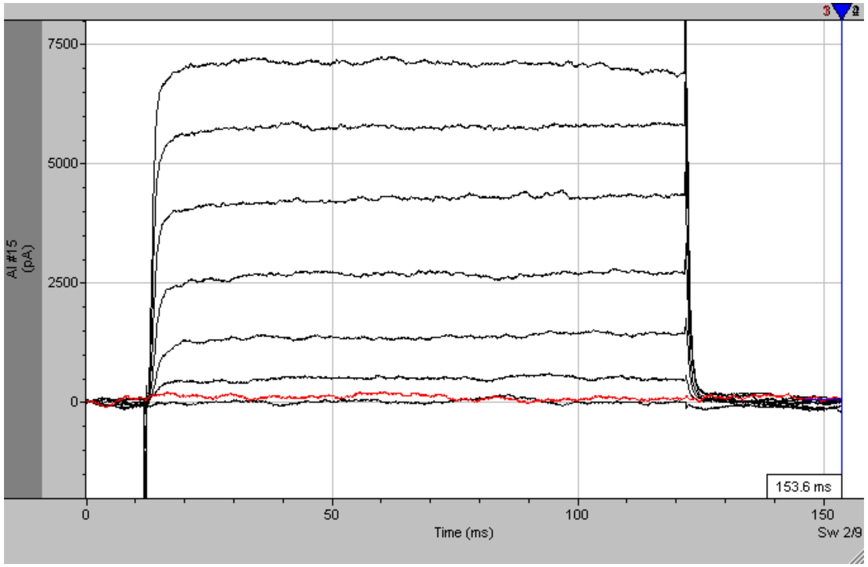
Young [Ca] 1 mm



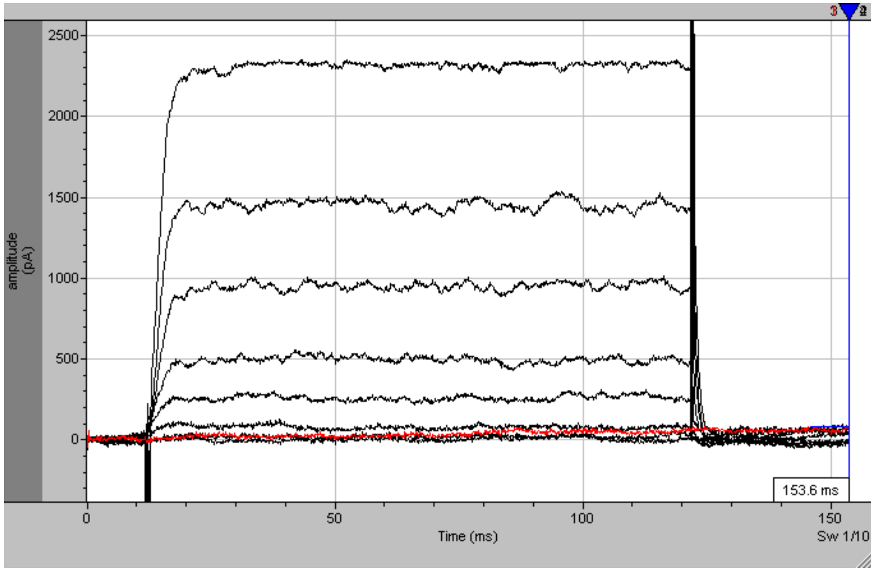
Centenarian [Ca] 1 mm



Young [Ca] 10 mm



Centenarians [Ca] 10 mm


amplitude (pA)
Time (ms)
153.6 ms
Sw 1/10