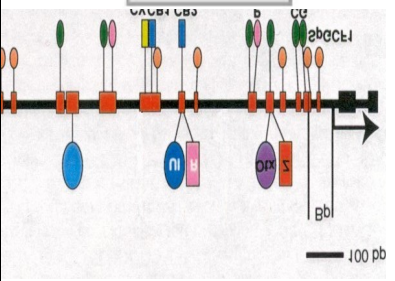
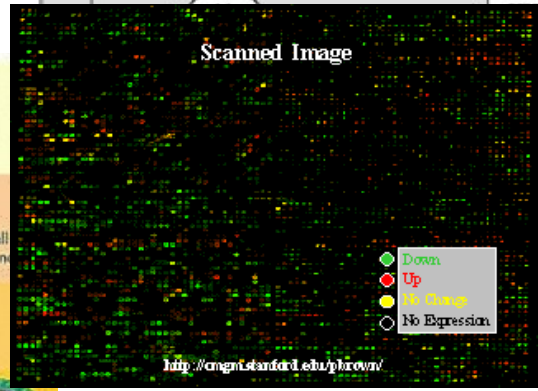
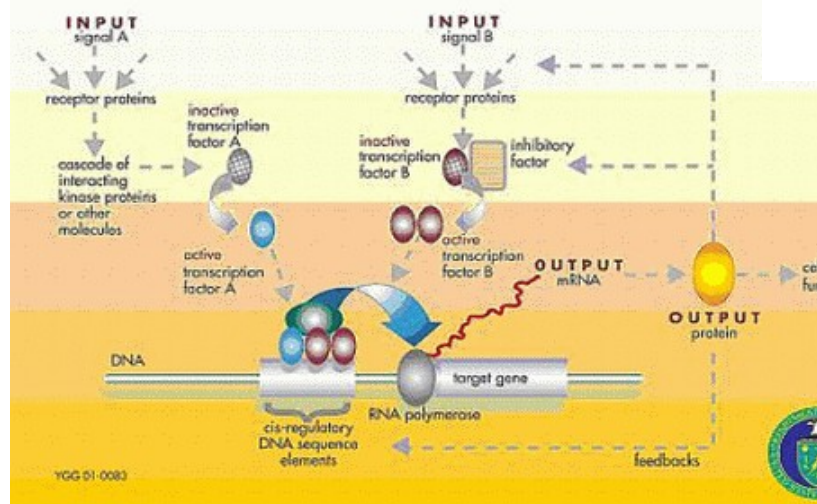
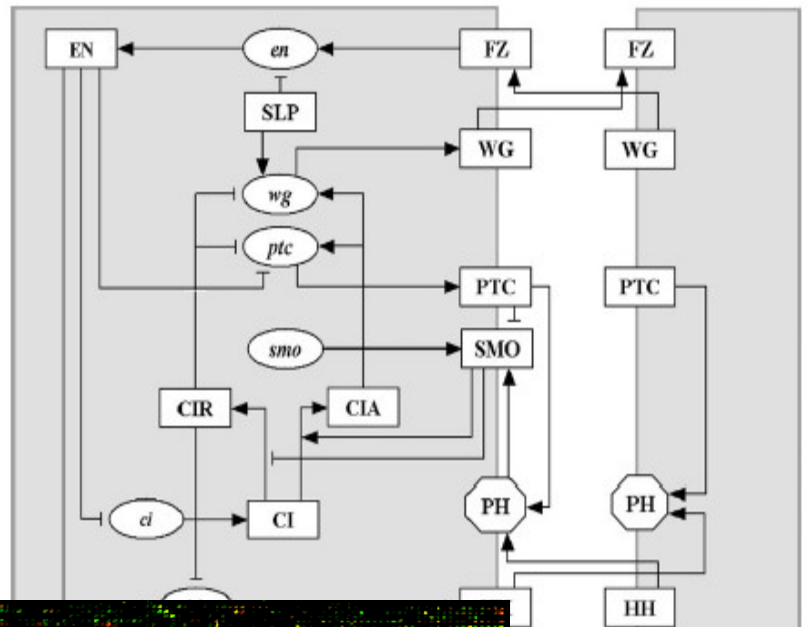
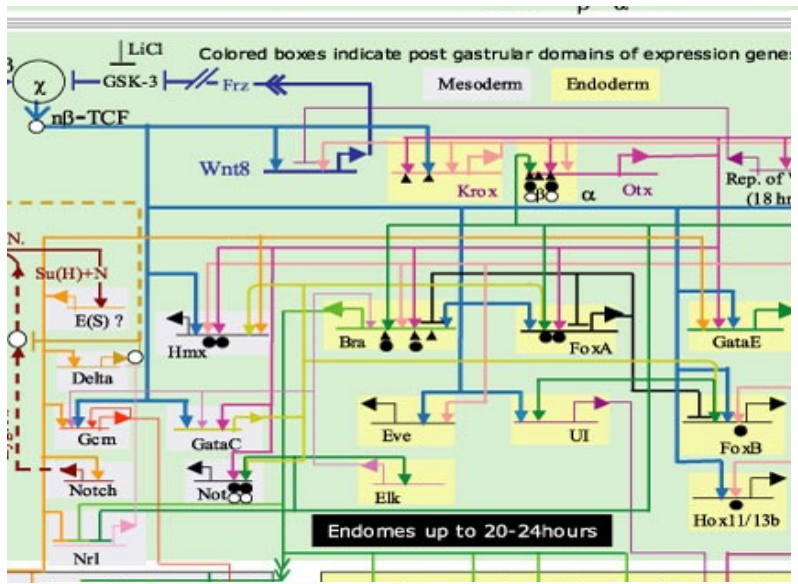


ECS 234: Introduction to Computational Functional Genomics



Administrativa

- Prof. Vladimir Filkov
3023 Kemper
filkov@cs.ucdavis.edu
- Appts:
 - Office Hours: Wednesday, 1:30-3p
 - Ask me or email me any time for appt

- ECS 234, 4 credits, CRN: 54135
- <http://www.cs.ucdavis.edu/~filkov/234/>
- No text required, papers and handouts
- Grading:
 - 60% project
 - 30% presentations
 - 10% class participation

Projects

- Groups should be formed by **Jan 16**
- Projects will ideally combine biological topics and computational methods
 - Project proposals are due **Jan 22**
 - Progress reports are due **Feb 17**
 - Project presentations **March 12**
 - Final reports are due **March 17**
- If you have an idea write to me and/or come and talk to me asap

Presentations

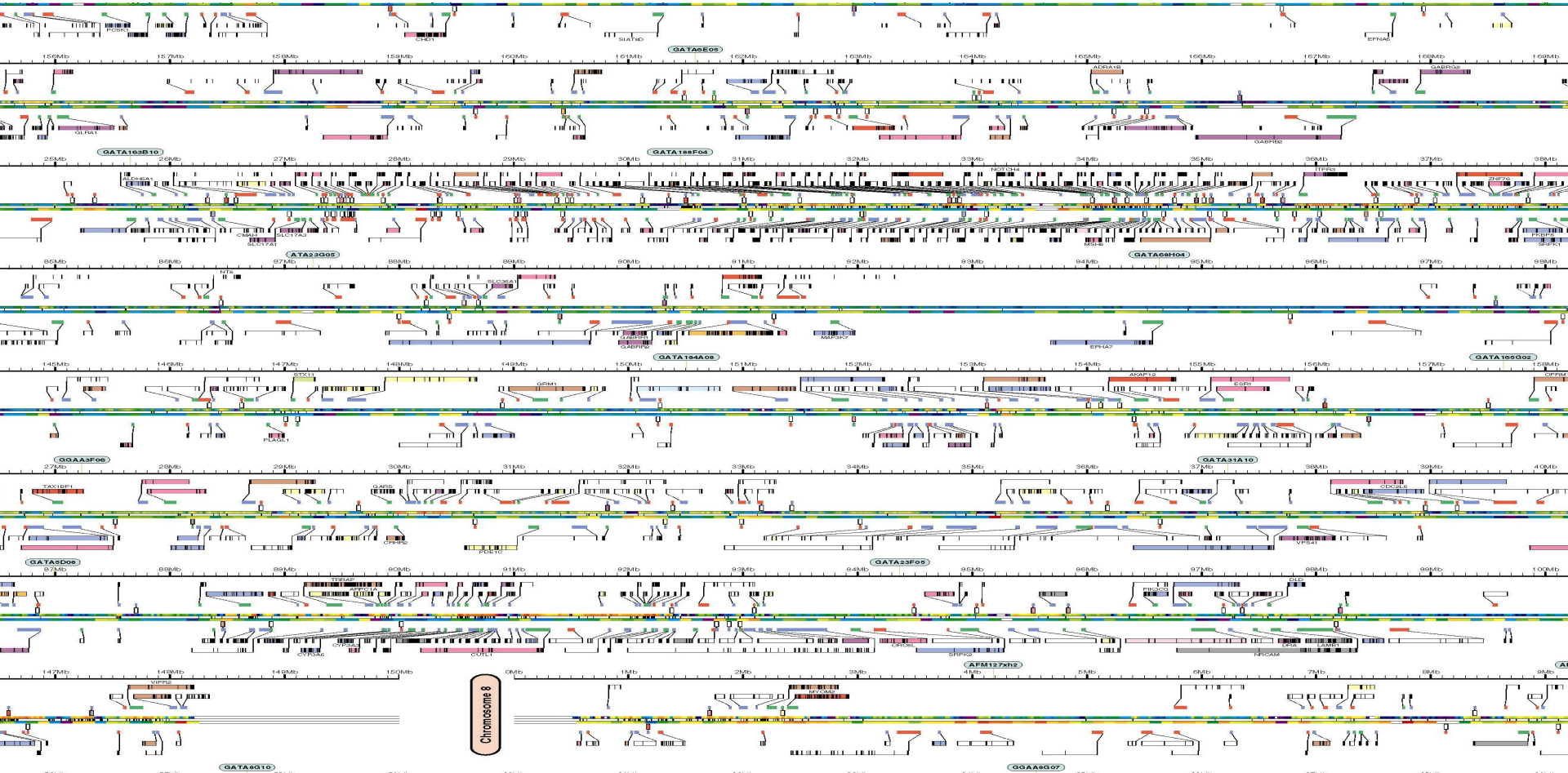
- Presentations: paper and/or software
- Each presentation will be during class time, 40-45 min long and 10-20min for discussion
- Need 2 volunteers for presentations on
 - Normalizing microarray data (different methods comparison),
 - Biclustering of microarray data,
 - or your choice.

Computational Biology and Bioinformatics

- Computational Biology: Computational scientists researching and developing methods for solving molecular biology problems
- Bioinformatics: More practical, deals with the appropriate use of tools to solve genomics problems
- Necessary background: molecular biology+computer science+statistics

- Sub-areas in Computational Biology/Bioinformatics:
 - Genomics
 - Functional Genomics
 - Proteomics
 - Phylogenetics
 - Bionetworks
 - etc.

What good is Comp. Bio?



This Course: Computational Functional Genomics

- **Genomics** = molecular biology + robotics + informatics
 - Depends on tools and techniques of recombinant DNA technology (e.g. DNA sequencing)
 - The technologies used are high-throughput (e.g. sequencing, microarraying, etc.)
 - Data processed and analyzed by computers
- **Functional** = exploration of gene and protein function in the cell processes (e.g. regulation, interactions, pathways, networks)
- **Computational** = using computational / algorithmic methods

How Is Function Discovered?

- Observe Nature, analyze data, make and test hypotheses, repeat
- Small scale: gene by gene
 - Manually curated databases (e.g. MIPS for yeast)
- Large scale: robotics + computational methods
 - Sequence
 - Gene Expression
 - Protein Interactions

The System and Its Parts



Rube Goldberg's Pencil Sharpener invention



Emergency knife (S) is always handy in case opossum or the woodpecker gets sick and can't work.

Examples of Genomic Data and Computational Tools for Functional Discovery in Genomics

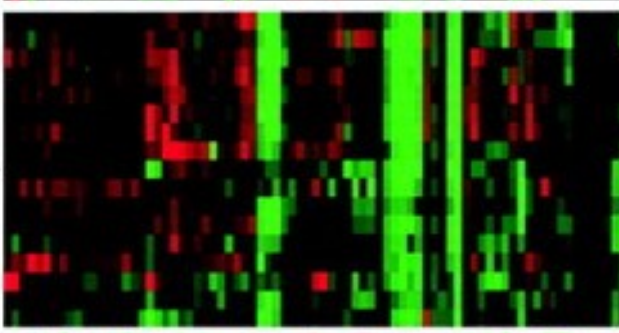
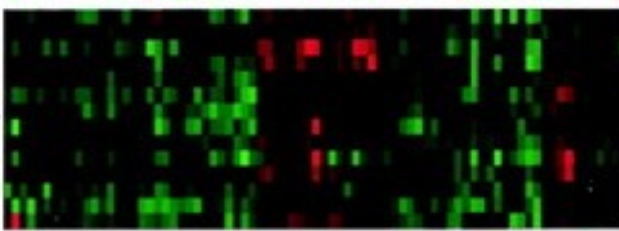
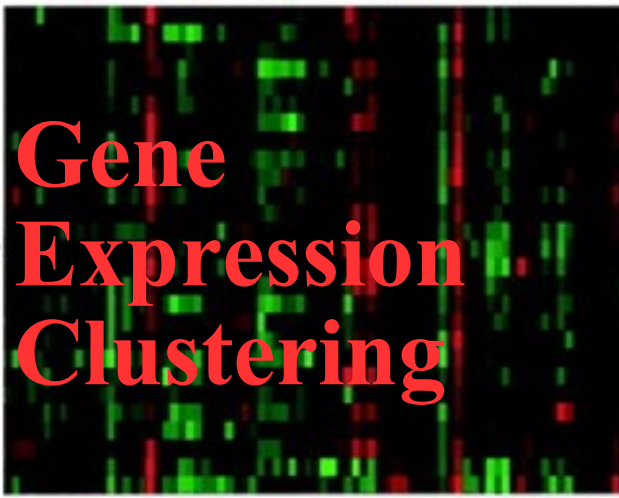
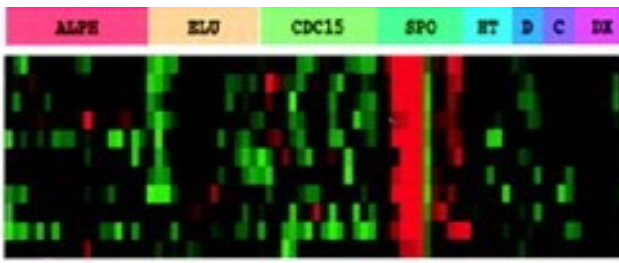
```

q9yl36 : -----YMKELLRQAKTAEIESFIDQKADPCNDFTAFSCGNYKRINSALS-----
q9vcu1 : -----MGDEKQQAANNFYACANWPRLHPARK-----
q9vcu2 : -----EQTVRRSKAQDMGSHVNSKIDPCVDFTAFYACGNKSTL-----
q9vcu3 : -----SRREQLVVRQKIATDVQKYHNLSADPCTDFTFYACCGQGRYHHRQL-----
q9vas4 : -----LQQIMRIAKSAEMCSFMDTSVSPCDDFTYCYACGNATINAATS-----
q9vb85 : -----LIGIQSNKECL-----RTAASLIYAMDDEQTDPCEDFTYQFTCGRMANEHPRPD-----
kell_human : -----PRPC-ETSV-----LDLRDHYLASGNTSVAFCPTDFTFSFACCGRAKETNN-----
q9ny95 : -----PR-C-PECC-----PERKAFARAAARFLAANDASIDPCQDFTYSFACCGMLRRHAIPD-----
q9jm10 : -----C-PECC-----PERKAFARAAARFLSANDASIDPCQDFTYSFACCGMLRRHAIPD-----
q9jh13 : -----C-PECC-----PERKAFARAAARFLSANDASIDPCQDFTYSFACCGMLRRHAIPD-----
o95672 : -----C-PECC-----PERKAFARAAARFLAANDASIDPCQDFTYSFACCGMLRRHAIPD-----
q9z192 : -----
q23684 : -----TLPC-TSREC-----VMTASRLKRVDSQVSPCNDFTYQFACCGMINQSVNLK-----
o16796 : TSTE-KPKKPEVC-STPCC-----VRAATHFLNANTSVDPCCDDFTFYACCGWINDQHPFIPD-----
q9w5y0 : -----KDCPSCNELPC-LNKHCI-----FASSEILKSLDVTVDPCDDFTYCYSCNQWIKNNPIPE-----
ecel_rat : -----TRTPPVC-LTEAC-----VSVTSSILNSMDPTVDPCCQDFTFYACCGWIKANPVPD-----
q28010 : -----TRTPSVC-LSEAC-----ISVTSSILSSMDPTVDPCCQDFTFYACCGWIKANPVPD-----
q28868 : -----TRTPSVC-LSEAC-----ISVTSSILSSMDPTVDPCCQDFTFYACCGWIKANPVPD-----
aaf98287 : -----ARPPAVC-LSEAC-----ISVTSSILSSLDRTVNPCCEDFTFSYACCGWVKANPLPD-----
q9upm4 : -----TRSPSVC-LSEAC-----VSVTSSILSSMDPTVDPCHDFTFSYACCGWIKANPVPD-----
ecel_cavpo : -----TRTPPVC-LSEAC-----VSVTSSILNSMNPVDPCCQDFTFYACCGWIKANPVPD-----
ecel_human : -----TRSPSVC-LSEAC-----VSVTSSILSSMDPTVDPCHDFTFSYACCGWIKANPVPD-----
ecel_bovin : -----TRTPSVC-LSEAC-----ISVTSSILSSMDPTVDPCCQDFTFYACCGWIKANPVPD-----
ece2_human : -----RDPSSHSTC-LTEAC-----IRVAGKILES LDRCVSPCEDFTYQFSCCGWIRRNPLPD-----
ece2_bovin : -----RDPSSHSTC-LTEAC-----IRVAGKILES LDRCVSPCEDFTYQFSCCGWIRRNPLPD-----
o44857 : -----VC-LSPCC-----IKTASVILSSMNSSVDPCCDDFTFYACCGWIKCHPFPD-----
nep_rabit : -----IC-KSDDC-----IKSAARLIQNMDATEPCTDFTFKYACCGWLKRNVIPE-----
nep_human : -----IC-KSDDC-----IKSAARLIQNMDATEPCTDFTFKYACCGWLKRNVIPE-----
nep_mouse : -----IC-KSDDC-----IKSAARLIQNMDATEPCTDFTFKYACCGWLKRNVIPE-----
nep_rat : -----IC-KSDDC-----IKSAARLIQNMDATEPCTDFTFKYACCGWLKRNVIPE-----
o93394 : -----IC-TIGDC-----TQASRLIENMDDSEVDPNDFTYQYACCGWLKKNPIPE-----
q9u9p2 : KGHSKQDLKDEVC-NTKEC-----VQIASKIIDVHDSVDPCKDFTFYACCGWLKSVFPVPD-----
cac07576 : RVLRDSSQKSDIC-TTPSC-----VIAAARLIQNMDSKPPCNDFTYQYACCGWLRRHVPE-----
q9j1i3 : RALRDSSLKSDIC-TTPSC-----VIAAARILENMDSRNPFCENFTYQYACCGWLRRHVPE-----
q9qzv6 : -ALRDSSLKSDIC-TTPSC-----VIAAARILENMDSRNPFCENFTYQYACCGWLRRHVPE-----
q9qzv7 : RALRDSSLKSDIC-TTPSC-----VIAAARILENMDSRNPFCENFTYQYACCGWLRRHVPE-----
q9ua44 : CGTTSADSDGKIC-VHECC-----VTAARIMSNLDKSVHPCDNFTYNYACANWEYDRDIPK-----
q9xz01 : SKCKSGDS-GDVC-LTQEC-----IHTASTVLRKHKPVEPCDNFTFYFACGTYLEENIPD-----
pex_mouse : GLLSFQA-KQEYC-LKPEC-----IEAAAAIMSKVNLVDPFCENFTFRFACDGMISNPIPE-----
o35812 : GLLSFQA-KQEYC-LKPEC-----IEAAAAIMSKVNLVDPFCENFTFRFACDGMISNPIPE-----
pex_human : GLLSLQA-KQEYC-LKPEC-----IEAAAAILSKVNLVDPFCDNFTFRFACDGMISNPIPE-----
q19831 : -HVNYGS-DNSTC-LTPEC-----IRLAASYLNNMNRDANPCEDFTFYFACGKYATRKVIAE-----
ycyl_caeel : -----DVVC-TSREC-----VRLAGFLAENLNSKINPCEDFTFYFACGNYLNRKNLPA-----
q9w435 : -----HDLNSDPCEDFTYQYACGTWKKHHPFIPD-----
q18673 : -----EKKTYTVGDSSEGYQEAASRLQLKSLNLSLDPCDDFTFYFACRAVVDSHPIPD-----
pepo_lache : -----MBRYLAVRCCAGDVAEPDLNAKFDQDNLFLAVNSEWLSKAEIPA-----
q91996 : -----MTRVQDDLFATVNADWLEKAQIPA-----
pepo_lacla : -----TRIQDDL FATVNAEWLENAEIPA-----
pepo_lac1c : -----TRIQDDL FATVNAEWLENAEIPA-----
ysc6_strgc : -----MTRLQDDFYDAINCEWAKTAVIPD-----
q9xd02 : -----MURLQDDFYDGVNCEWAEATAVIPD-----
cac14579 : -----MTRLQDDFYHAINCEWEKTAVIPD-----
q9pft1 : LASCNRMAPAPVAIPTPNPTPKTNTNDIDLSTLPPVIRFTASDLDPTGNEPCTDLHTGVNENWLNKANPVPD-----
o50642 : KFLCLAPAVIGALMLTGCNGNKGQNTDTRKREPVPAIDLSDMNTSVRFPQDDFTYRYENGNWMMKNNPLKP-----
o06075 : -----MNIETAIRSGIDLSTYVDANTRPQDDLFGHVNRRLYEYEIPA-----
o53649 : -----MTLAIPSGIDLSHDADARPQDDLFGHVNRRLAEHEIPA-----
o45569 (b) : IDTTTFCLPNVKTIDINVVPRAPITSTNVDMKNAYQMAVDYYAKSVNTAIDPCDDFTYSFACGNFNQSVSFFY-----
o45569 (a) : -----QPAKVSPCTDKYNSYQVVQLFKASVNLVDPFCNDFTYATTCGPKGDMSFDI-----
o76751 : -----DNVFC-----PNVGNANRSKEWKNAANTLLFLGLDESVDPCEDFTYQFTCNKFFIERIDLDEL-----
q22763 : -----STPCTVNTSPSYTQAANYLLNGLDPTVDPCCQDFTYAFTCNKFLQNTDLQKL-----
aag29103 : -----DPKYC-----PSYCEPDKKYAYQEAASVLLSGLDQTVDPCEDLFAFTCNTYLRNHNATD-----
o45131 : -----SSKYC-----PSYCEALFTPPWKAAASLLRNAINDESVDPCEDFTYQFTCCSYITQH-----
q25051 : -----SSKYC-----PSYCEALFTPPWKAAASLLQNAINDESVDPCEDLQFTCCSYITQH-----
o76750 : -----SSKYC-----PSYCEALNTPAWKEAANNLQNALDREDVNPCCDDFTYKFPSCGKYISHT-----

```

Sequence
Homology
(multiple
alignment)

A



Gene Expression Clustering

B

STG2	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
DHS1	DNA REPAIR	KINCHLASE; ALSO RECOMBINATION
NSR1	CYTOSKELETON	ACTIN FILAMENT ORGANIZATION
SPC42	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
CHM67	CYTOSKELETON	SPINDLE POLE BODY COMPONENT
CLB4	CELL CYCLE	G2/M CYCLIN
CDC10	CYTOKINESIS	GTP BINDING PROTEIN
CDC3	CYTOKINESIS	SEPTIN
CLB3	CELL CYCLE	G2/M CYCLIN
APC4	CELL CYCLE	ANAPHASE-PROMOTING COMPLEX SUBUNIT
CDC14	CELL CYCLE	ANAPHASE-PROMOTING COMPLEX SUBUNIT

C

PPW11	PROTEASOME DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
UPD1	PROTEIN DEGRADATION	UBIQUITIN FUSION DEGRADATION
PPW9	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
EPT1	PROTEIN DEGRADATION	14S PROTEASOME SUBUNIT
PPW6	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW4	PROTEIN DEGRADATION	PROTEASOME SUBUNIT, B TYPE
PPW5	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW4	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW7	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW3	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW2	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT ALPHA5
SC11	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT SC7ALPHA/YS
PPW5	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT ALPHA6
PPW3	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT T13 ALPHA3
PPW1	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT C11 BETA4
PPW2	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT BETA5
PPW3	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT BETA1
PPW10	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT C1 ALPHA7
PPW1	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT BETA2
PPW5	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT ALPHA4
PPW7	PROTEIN DEGRADATION	10S PROTEASOME SUBUNIT
PPW10	PROTEIN DEGRADATION	14S PROTEASOME SUBUNIT
PPW3	PROTEIN DEGRADATION	14S PROTEASOME SUBUNIT
PPW5	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW12	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT
PPW5	PROTEIN DEGRADATION	14S PROTEASOME SUBUNIT
PPW9	PROTEIN DEGRADATION	14S PROTEASOME REGULATORY SUBUNIT

D

POG4	TRNA PROCESSING	RNASE P AND RNASE MFP SUBUNIT
CAF16	TRANSPORT	ATP-BINDING CASSETTE ABC FAMILY
MSL1	MRNA SPLICING	UNKNOWN
MSL3	MRNA SPLICING	CORE SMN2 PROTEIN
TAP40	TRANSCRIPTION	TPSID 40 ED SUBUNIT
PPW39	RNA PROCESSING	CURL mRNA STABILITY
PPW1	TRANSCRIPTION	TPY1A
YHU1	MRNA 3'-5' END PROCESSING	CLEAVAGE/POLYADENYLATION FACTOR CF II COMPONENT
MSL1	MRNA STABILITY	UNKNOWN
PPW24	MRNA SPLICING	G4/U5 SMN2 PROTEIN
STO1	GLUCOSE REPRESSION	MODULATOR OF GLUCOSE REPRESSION
MSL1	MITO GENOME MAINT	DYNAMIS FAMILY PROTEIN
PPW19	MRNA SPLICING	NON-SMNS2 SPLICOSOME COMPONENT
SLM1	MITOCHONDRIAL METABOLISM	INTRINSAL MEMBRANE PROTEIN

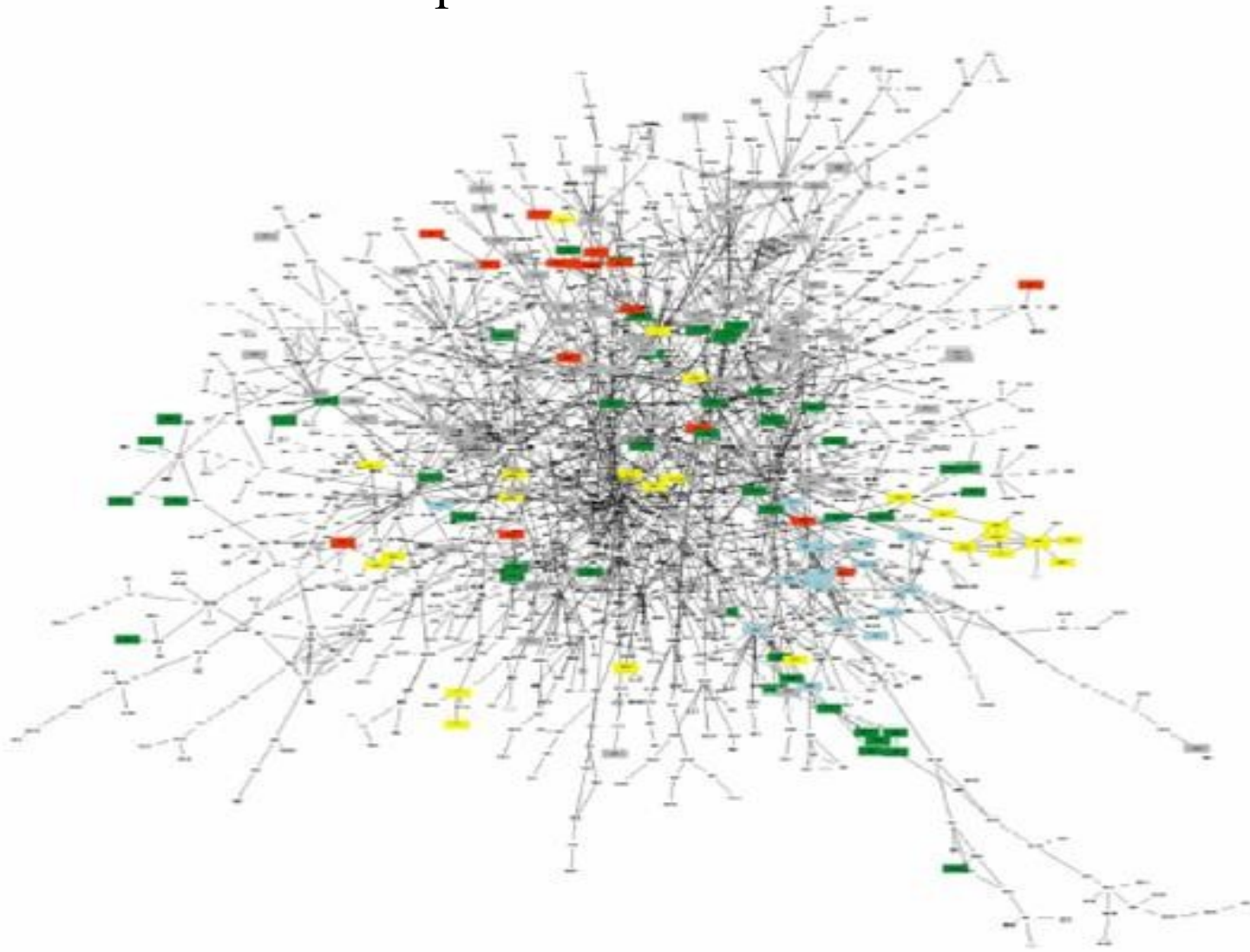
E

TP11	GLYCOLYSIS	TRIOSEPHOSPHATE ISOMERASE
GPW1	GLYCOLYSIS	PHOSPHOGLYCERATE MUTASE
PGW1	GLYCOLYSIS	PHOSPHOGLYCERATE KINASE
TDH3	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 3
TDH2	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 2
ENO2	GLYCOLYSIS	ENOLASE II
TDH1	GLYCOLYSIS	GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 1
PFAL	GLYCOLYSIS	ALDOOLASE
TKL1	PENTOSE PHOSPHATE CYCLE	TRANSKETOLASE
PFK5	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
PFK6	GLYCOLYSIS	PYRUVATE DECARBOXYLASE 3
PFK1	GLYCOLYSIS	PYRUVATE DECARBOXYLASE
CDC19	GLYCOLYSIS	PYRUVATE KINASE
MSK2	GLYCOLYSIS	HEXOKINASE II
TYE7	GLYCOLYSIS	BASIC H-L-N TRANSCRIPTION FACTOR
PPW1	GLYCOLYSIS	PHOSPHOGLYCOKINASE
ACS2	ACETYL-COA BIOSYNTHESIS	ACETYL-COENZYME A SYNTHETASE

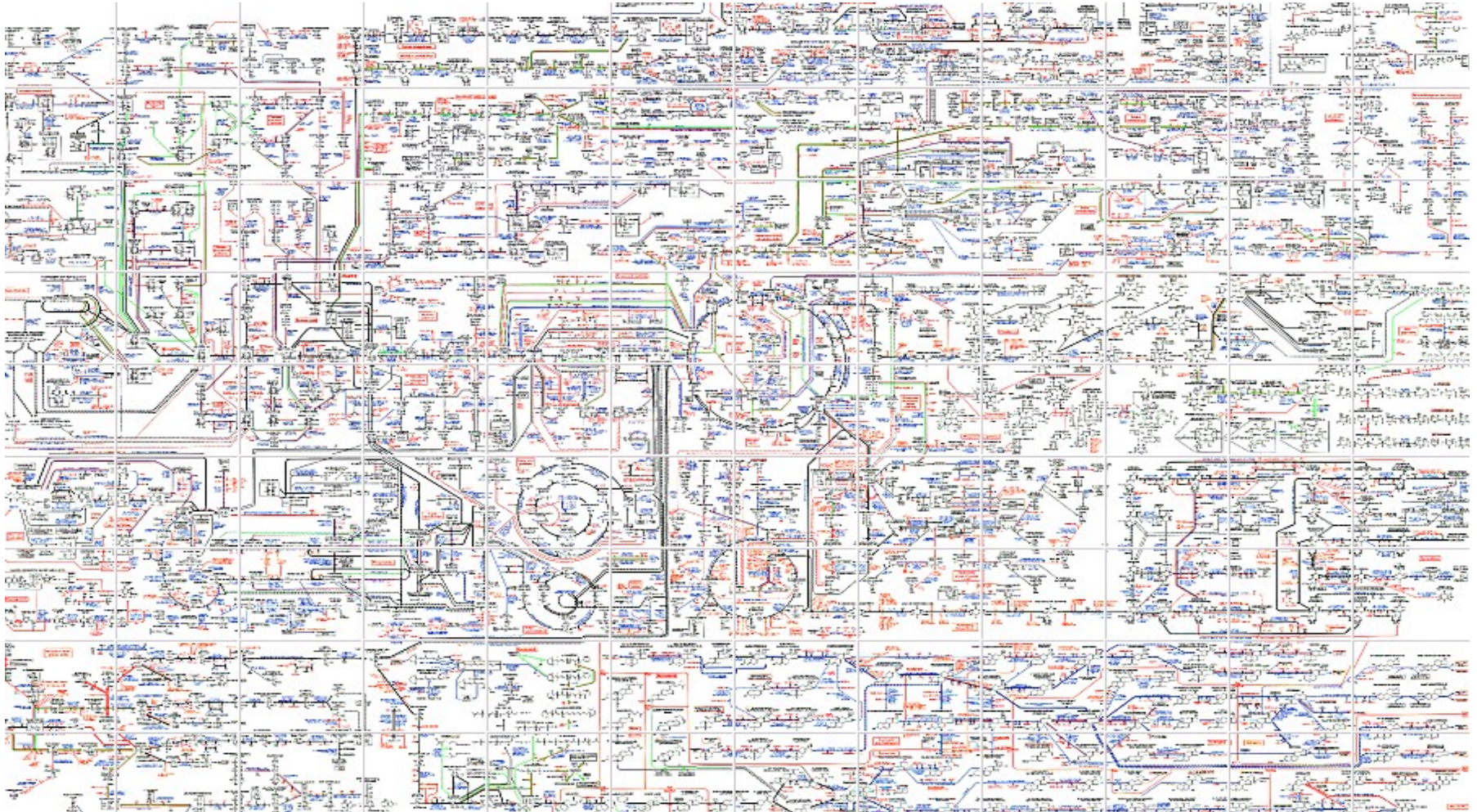
F

MS16	UBIQUITIN BIOSYNTHESIS	METHYLTRANSFERASE
MS11	MRNA SPLICING	UNKNOWN
MS1	PROTEIN SYNTHESIS	MITOCHONDRIAL METHYLALANYL-TRNA SYNTHETASE SUBUNIT
MS15	PROTEIN SYNTHESIS	EROSIONAL PROTEIN, MITOCHONDRIAL S5

PPI Network Maps



Data Integration



Course Overview

Intro: Biology, Biotechnologies, Computational Methods

- A. Central Dogma of Molecular Biology
- B. High-Throughput technologies: DNA sequencing, Gene Expression Arrays, Protein-DNA, and Protein-Protein Interactions
- C. CS for Biologists

Bioinformatics and Data Mining of Large-Scale Data

- A. Gene Expression analysis (statistics, classification, clustering)
- B. Sequence analysis (promoter region analysis)
- C. TF-DNA and Protein-Protein interactions analysis (topological properties and comparison)

Course Overview, contd.

Gene Network Inference

- A. Graph Models
- B. Boolean Networks
- C. Bayesian Networks
- D. Linear Additive Models

Combining Heterogeneous Data Sources

- A. Sequence + gene expression
- B. Gene Expression + protein-protein interactions
- C. Methods for general data integration