Computational Systems Biology

Advanced Technologies in Bioscience 2008–2009

Chalmers Graduate School in Bioscience

T. M. Murali

August 18, 2008

Course Structure

Discuss state-of-the-art research papers.

Course Structure

Discuss state-of-the-art research papers.

- Reading assignments

Course Structure

Discuss state-of-the-art research papers.

- Reading assignments
- Lectures

Course Structure

Discuss state-of-the-art research papers.

- Reading assignments
- Lectures
- Exercises

Course Structure

Discuss state-of-the-art research papers.

- Reading assignments
- Lectures
- Exercises
- Class participation

Suggestions on Reading

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.

Suggestions on Reading

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.
- Algorithmic/computational papers:
  - Are the biological assumptions valid?
  - Is the algorithm good and computational efficient? Can you improve the technique?
  - Can you mathematically describe the output of the algorithm?

Suggestions on Reading

- Be sceptical/critical: even papers in Nature, Science, or PNAS have errors or invalid thinking.
- Algorithmic/computational papers:
  - Are the biological assumptions valid?
  - Is the algorithm good and computational efficient? Can you improve the technique?
  - Can you mathematically describe the output of the algorithm?
- Read supplementary information. Often has details about the assumptions, the techniques, and the results.

Sources of Information

- **There is no textbook for the course.**
- Useful/related books:
  - *Networks: From Biology to Theory*, Jianfeng Feng, Jürgen Jost, and Minping Qian, Springer-Verlag.
  - *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*, Eric H. Davidson, Academic Press.
  - *Computational Modeling of Genetic and Biochemical Networks*, James M. Bower and Hamid Bolouri, MIT Press
  - *Microarrys for an Integrative Genomics*, Isaac S. Kohane, Atul J. Butte, and Alvin Kho, MIT Press.
  - *Systems Biology in Practice: Concepts, Implementation and Application*, Edda Klipp, Ralf Herwig, Axel Kowald, Christoph Wierling and Hans Lehrach, Wiley.

More Sources of Information

- Conferences: ICSB, RECOMB, ISMB, PSB, KDD, machine learning conferences, discrete algorithms conferences.
- Journals (CS-oriented): Bioinformatics, Journal of Computational Biology, BMC Bioinformatics, TCBB, TKDE.
- Journals (biology-oriented) Nature, Science, Molecular Systems Biology, Nature Reviews Drug Discovery, Nature Biotechnology, Nature Reviews Cancer, Drug Discovery Today, PNAS, NAR, Genome Biology, Genome Research.

Rewind to 1953

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

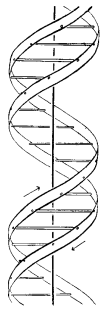
MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. Watson
F. H. C. Crick
Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

[1] Pauling, L., and Corey, R. B., *Nature*, **171**, 346 (1953); *Proc. U.S. Nat. Acad. Sci.*, **39**, 84 (1953).
[2] Furberg, S., *Acta Chem. Scand.*, **6**, 634 (1952).
[3] Chargaff, E., for references see Zamenhof, S., Brawerman, G., and Chargaff, E., *Biochim. et Biophys. Acta*, **9**, 402 (1952).
[4] Wyatt, G. R., *J. Gen. Physiol.*, **36**, 201 (1952).
[5] Astbury, W. T., Symp. Soc. Exp. Biol. 1, Nucleic Acid, 66 (Camb. Univ. Press, 1947).
[6] Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, **10**, 192 (1953).

Molecular Structure of Deoxypentose Nucleic Acids

While the biological properties of deoxypentose nucleic acid suggest a molecular structure containing great complexity, X-ray diffraction studies described here (cf. Astbury[1]) show the basic molecular configuration has great simplicity. The purpose of this communication is to describe, in a preliminary way, some of the experimental evidence for the polynucleotide chain configuration being helical, and existing in this form when in the natural state. A fuller account of the work will be published shortly.

The structure of deoxypentose nucleic acid is the same in all species (although the nitrogen base ratios alter considerably) in nucleoprotein, extracted or in cells, and in purified nucleate. The same linear group of polynucleotide chains may pack together parallel in different ways to give crystalline[1–3], semi-crystalline or paracrystalline material. In all cases the X-ray diffraction photograph consists of two regions, one determined largely by the regular spacing of nucleotides along the chain, and the other by the longer spacings of the chain configuration. The sequence of different nitrogen bases along the chain is not made visible.

Oriented paracrystalline deoxypentose nucleic acid ('structure B' in the following communication by Franklin and Gosling) gives a fibre diagram as shown in Fig. 1 (cf. ref. 4). Astbury suggested that the strong 3·4-A. reflexion corresponded to the internucleotide repeat along the fibre axis. The ∼ 34 A. layer lines, however, are not due to a repeat of a polynucleotide composition, but to the chain configuration repeat, which causes strong diffraction as the nucleotide chains have higher density than the interstitial water. The absence of reflexions on or near the meridian immediately suggests a helical structure with axis parallel to fibre length.

Diffraction by Helices

It may be shown[5] (also Stokes, unpublished) that the intensity distribution in the diffraction pattern of a series of points equally spaced along a helix is given by the squares of Bessel functions. A uniform continuous helix gives a series of layer lines of spacing corresponding to the helix pitch, the intensity distribution along the nth layer line being proportional to the square of J_n, the nth order Bessel function. A straight line may be drawn approximately through

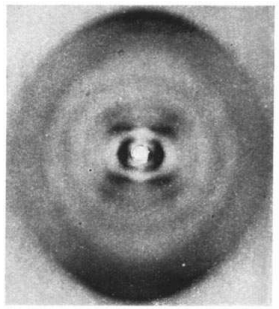


Fig. 1. Fibre diagram of deoxypentose nucleic acid from *B. coli*. Fibre axis vertical

the innermost maxima of each Bessel function and the origin. The angle this line makes with the equator is roughly equal to the angle between an element of the helix and the helix axis. If a unit repeats n times along the helix there will be a meridional reflexion (J_n^2) on the nth layer line. The helical configuration produces side-bands on this fundamental frequency, the effect[5] being to reproduce the intensity distribution about the origin around the new origin, on the nth layer line, corresponding to C in Fig. 2.

We will now briefly analyse in physical terms some of the effects of the shape and size of the repeat unit or nucleotide on the diffraction pattern. First, if the nucleotide consists of a unit having circular symmetry about an axis parallel to the helix axis, the whole diffraction pattern is modified by the form factor of the nucleotide. Second, if the nucleotide consists of a series of points on a radius at right-angles to the helix axis, the phases of radiation scattered by the helices of different diameter passing through each point are the same. Summation of the corresponding Bessel functions gives reinforcement for the inner-

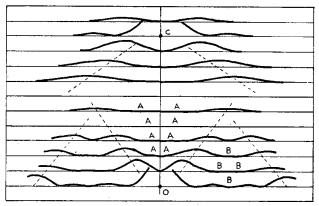


Fig. 2. Diffraction pattern of system of helices corresponding to structure of deoxypentose nucleic acid. The squares of Bessel functions are plotted about 0 on the equator and on the first, second, third and fifth layer lines for half of the nucleotide mass at 20 A. diameter and remainder distributed along a radius, the mass at a given radius being proportional to the radius. About C on the tenth layer line similar functions are plotted for an outer diameter of 12 A.

The Human Genome Project

The Human Genome Project

Before: human genome has about 100,000 genes.

The Human Genome Project

Before: human genome has about 100,000 genes.

After: human genome has about 30,000 genes.

Shock and Dismay

- The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.

Shock and Dismay

- The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.
- Washington Post: It also raises new and difficult questions, such as how human beings—with all their passions and fears, their capacity for art, music, culture and war—can be all that they are with just 30,000 or so genes, only five times as many as in baker's yeast.

Shock and Dismay

- The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.
- Washington Post: It also raises new and difficult questions, such as how human beings—with all their passions and fears, their capacity for art, music, culture and war—can be all that they are with just 30,000 or so genes, only five times as many as in baker's yeast.
- USA TODAY: Perhaps the biggest surprise since the code was deciphered in June is that it takes just 30,000 to 40,000 genes to make, maintain and repair a human. . . . "*If you're judging the complexity of an organism by the number of genes it has, we've just taken a big hit in the pride department*," says the National Genome Research Institute's director, Francis Collins, who also heads the U.S. arm of the International Human Genome Project.

Shock and Dismay

- The New York Times: **Genome Analysis Shows Humans Survive on Low Number of Genes** The two teams report that there are far fewer human genes than thought—probably a mere 30,000 or so—only a third more than those found in the roundworm. . . . The impact on human pride is another matter.
- Washington Post: It also raises new and difficult questions, such as how human beings—with all their passions and fears, their capacity for art, music, culture and war—can be all that they are with just 30,000 or so genes, only five times as many as in baker's yeast.
- USA TODAY: Perhaps the biggest surprise since the code was deciphered in June is that it takes just 30,000 to 40,000 genes to make, maintain and repair a human. . . . "*If you're judging the complexity of an organism by the number of genes it has, we've just taken a big hit in the pride department*," says the National Genome Research Institute's director, Francis Collins, who also heads the U.S. arm of the International Human Genome Project.
- The New York Times (Aug 24, 2001): **Human Genome Now Appears More Complicated After All** After a humiliating deflation this February, human dignity is on the recovery path, at least as measured by the number of genes in the human genome.

Relative Genome Sizes

Chimps vs. Humans

Chimps vs. Humans

Chimps vs. Humans

Chimps vs. Humans

Chimp and chump genomes are only about 1.2% different!

What Factors Differentiate Various Species?

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).
- Patterns of gene activity (gene expression) are different.

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).
- Patterns of gene activity (gene expression) are different.
- Ways in which proteins interact with and regulate each other and other molecules are different.

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).
- Patterns of gene activity (gene expression) are different.
- Ways in which proteins interact with and regulate each other and other molecules are different.

- *"It is the evolution of the regulatory networks and not the genes themselves that play the critical role in making organisms different from one another,"* The Digital Code of DNA, Hood and Galas, *Nature*, vol 421, 2003.

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).
- Patterns of gene activity (gene expression) are different.
- Ways in which proteins interact with and regulate each other and other molecules are different.

- *"It is the evolution of the regulatory networks and not the genes themselves that play the critical role in making organisms different from one another,"* The Digital Code of DNA, Hood and Galas, *Nature*, vol 421, 2003.
- We need to understand how genes, proteins, and other molecules interact with other in different cell states, different tissues, and under different external conditions.

What Factors Differentiate Various Species?

- Genes are different (only dogs have the *submaxillary mucin* genes).
- Patterns of gene activity (gene expression) are different.
- Ways in which proteins interact with and regulate each other and other molecules are different.

- *"It is the evolution of the regulatory networks and not the genes themselves that play the critical role in making organisms different from one another,"* The Digital Code of DNA, Hood and Galas, *Nature*, vol 421, 2003.
- We need to understand how genes, proteins, and other molecules interact with other in different cell states, different tissues, and under different external conditions.
- Study only of individual elements is unlikely to reveal higher-order principles.

Enter Systems Biology

- Systems Biology is the study of the parts of the cell, their properties, and their relationships.

Enter Systems Biology

- Systems Biology is the study of the parts of the cell, their properties, and their relationships.
- What are the structures and modules that make up cellular networks?

Enter Systems Biology

- Systems Biology is the study of the parts of the cell, their properties, and their relationships.
- What are the structures and modules that make up cellular networks?
- How do these modules interact with each other over time and in different situations?

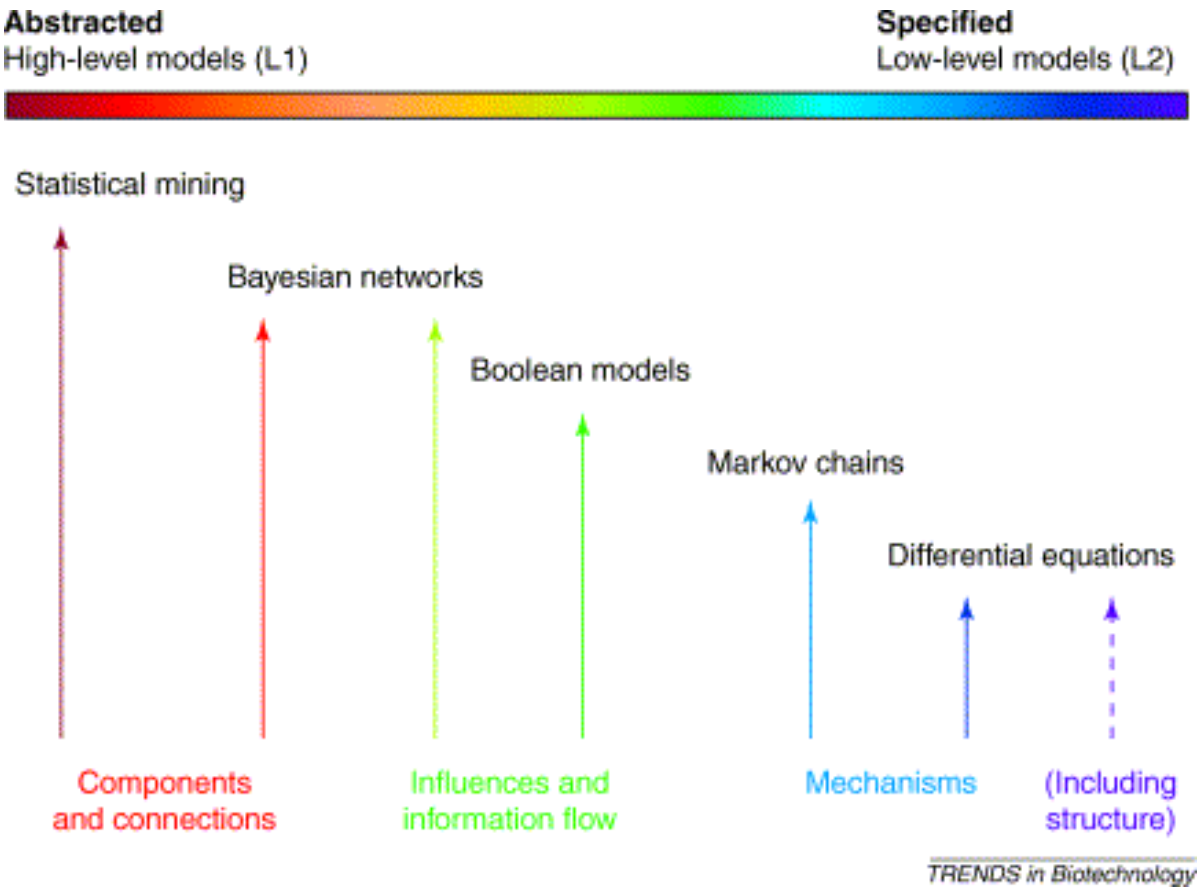
Enter Systems Biology

- Systems Biology is the study of the parts of the cell, their properties, and their relationships.
- What are the structures and modules that make up cellular networks?
- How do these modules interact with each other over time and in different situations?
- How can we interrogate the cell and iteratively refine our models of the cell?

Characteristics of Systems Biology

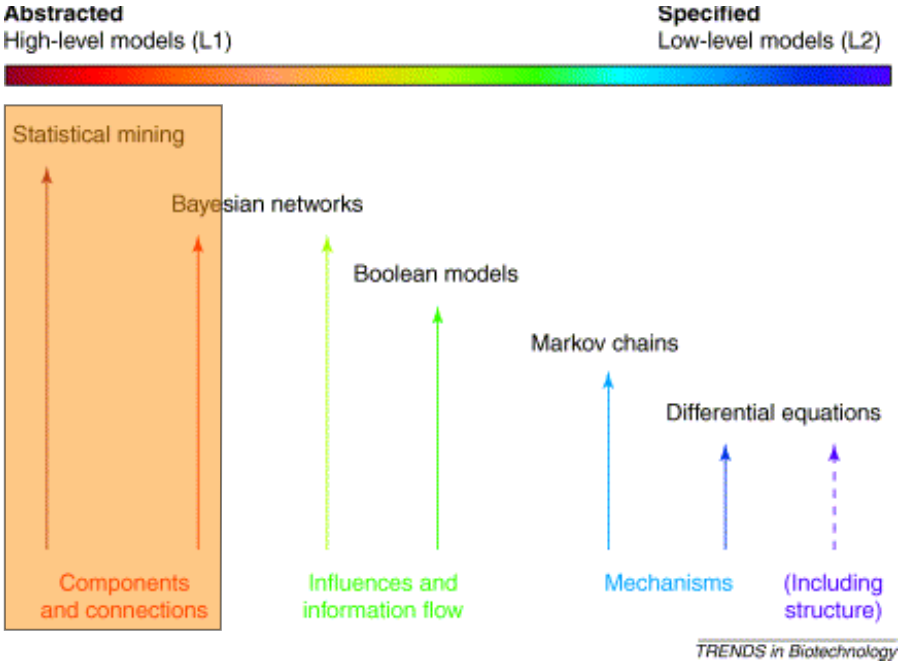
- Modular cell biology (rather than molecular).
- Discovery-driven *and* hypothesis-driven.
- Driven by high-throughput and accurate biological measurements.
- Uses and needs sophisticated computational, mathematical, and statistical ideas.
- Requires close collaboration between biologists and quantitative scientists.

Continuum of Models in Systems Biology



From *Building with a scaffold: emerging strategies for high- to low-level cellular modeling*, Ideker and Lauffenburger, Trends in Biotechnology Volume 21, Issue 6 , June 2003, Pages 255-262.

Goals of the Course



- We will cover "high-level" models.
- Emphasise a data-driven approach to systems biology.
- Focus on large-scale properties of biological systems.
- Integrate massive quantities of different types of data
- Learn techniques from clustering, data mining, and graph theory and apply them to solve specific biological questions.

Sources of Data

Sources of Data

- Gene expression data
  - Gene knockouts and external perturbations such as drugs.
  - Samples belonging to various classes
  - Time-series data.
  - GEO, SGD, the Whitehead institute.

Sources of Data

- Gene expression data
  - Gene knockouts and external perturbations such as drugs.
  - Samples belonging to various classes
  - Time-series data.
  - GEO, SGD, the Whitehead institute.
- Protein-protein interaction data
  - Large-scale Yeast 2-hybrid assays (yeast, worm, fruitfly).
  - Affinity precipitation + mass spectometry (yeast).
  - Literature (HPRD).

Sources of Data

- Gene expression data
  - Gene knockouts and external perturbations such as drugs.
  - Samples belonging to various classes
  - Time-series data.
  - GEO, SGD, the Whitehead institute.
- Protein-protein interaction data
  - Large-scale Yeast 2-hybrid assays (yeast, worm, fruitfly).
  - Affinity precipitation + mass spectometry (yeast).
  - Literature (HPRD).
- Transcriptional regulation
  - Protein-DNA binding data (yeast, human liver TFs).
  - Binding profiles for known TFs (SCPD, TRANSFAC).

Sources of Data

- Gene expression data
  - Gene knockouts and external perturbations such as drugs.
  - Samples belonging to various classes
  - Time-series data.
  - GEO, SGD, the Whitehead institute.
- Protein-protein interaction data
  - Large-scale Yeast 2-hybrid assays (yeast, worm, fruitfly).
  - Affinity precipitation + mass spectometry (yeast).
  - Literature (HPRD).
- Transcriptional regulation
  - Protein-DNA binding data (yeast, human liver TFs).
  - Binding profiles for known TFs (SCPD, TRANSFAC).
- Protein abundance and activity

Sources of Data

- Gene expression data
  - Gene knockouts and external perturbations such as drugs.
  - Samples belonging to various classes
  - Time-series data.
  - GEO, SGD, the Whitehead institute.
- Protein-protein interaction data
  - Large-scale Yeast 2-hybrid assays (yeast, worm, fruitfly).
  - Affinity precipitation + mass spectometry (yeast).
  - Literature (HPRD).
- Transcriptional regulation
  - Protein-DNA binding data (yeast, human liver TFs).
  - Binding profiles for known TFs (SCPD, TRANSFAC).
- Protein abundance and activity
- Metabonomics

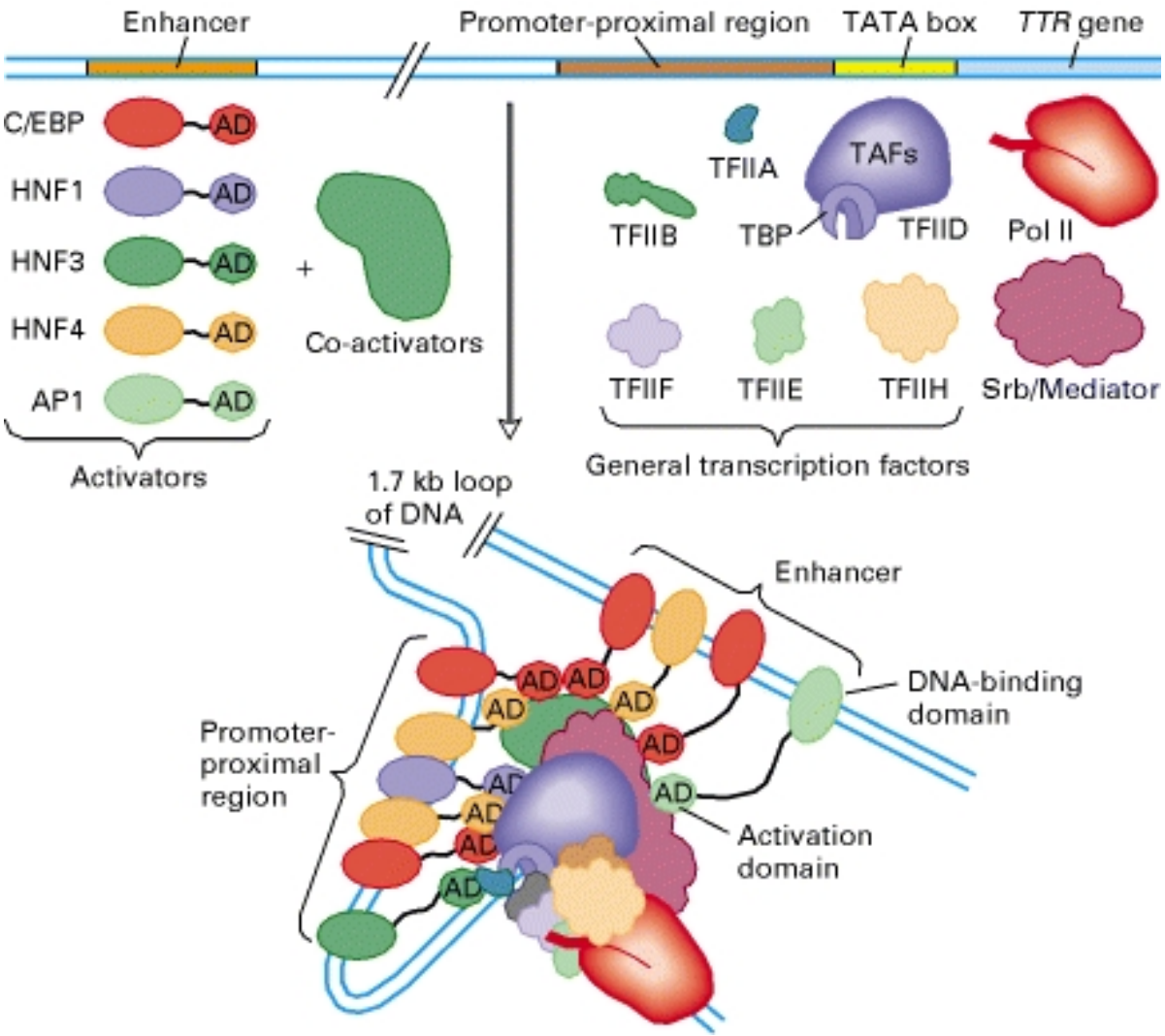
Sources of Data

- Literature, Computation, Databases
  - Transcriptional regulators (TRANSFAC)
  - Protein-protein interactions (DIP, GRID, Predictome, MIPS)
  - Metabolic networks (KEGG, EcoCyC, BioCarta, GenMAPP)
  - Functional annotations (GO, MIPS, species-specific databases)
  - Genetic Associations with Disease (GAD, MEDGENE, i-HOP).

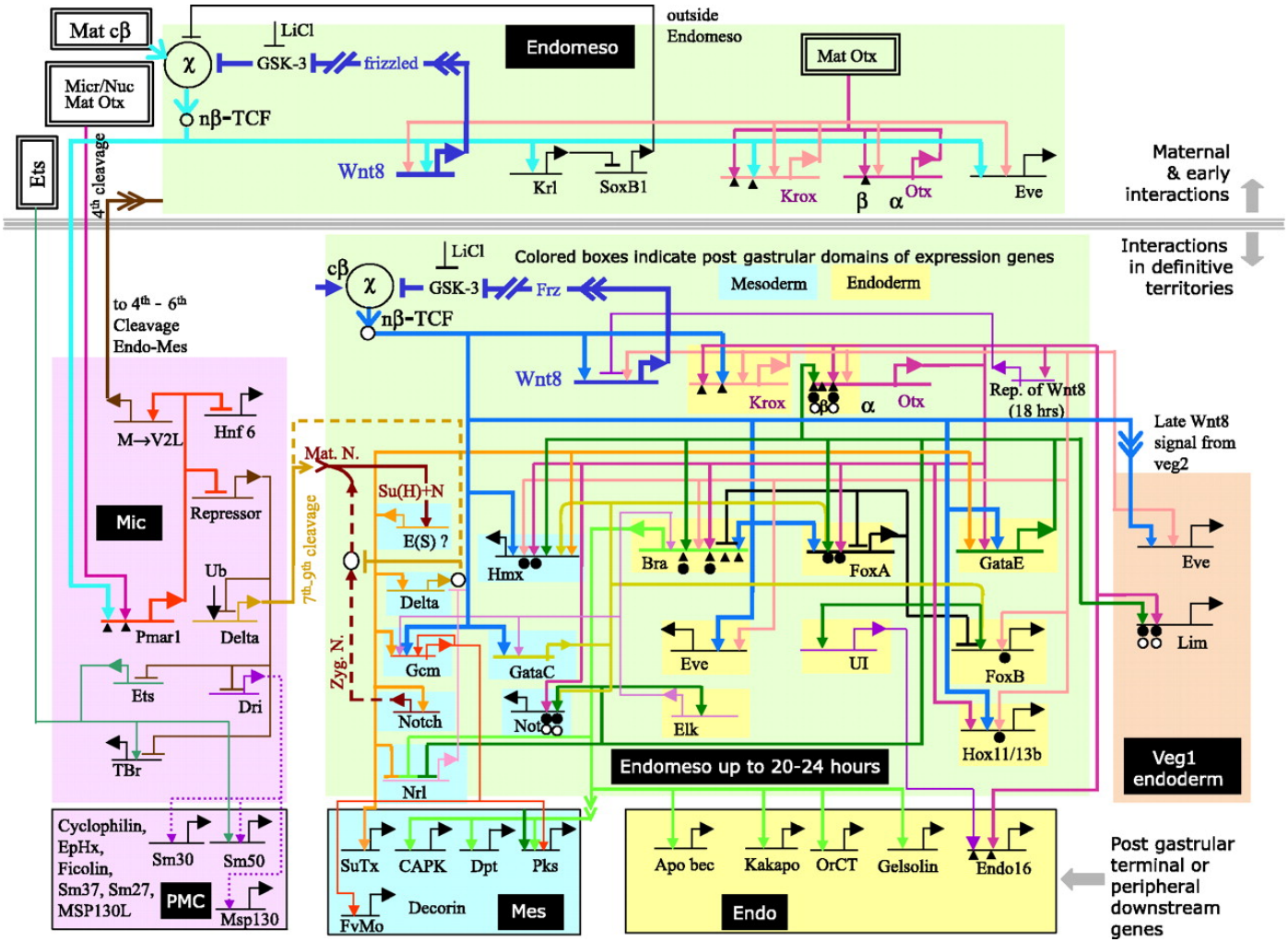
Specific Topics

Monday	Clustering gene expression data; application to find cancer gene modules.
Tuesday	Biclustering, application to data integration in *S. cerevisiae*.
Wednesday	Response networks and network legos.
Thursday	Gene function prediction.
Friday	Host-pathogen interaction networks (ICSB tutorial).

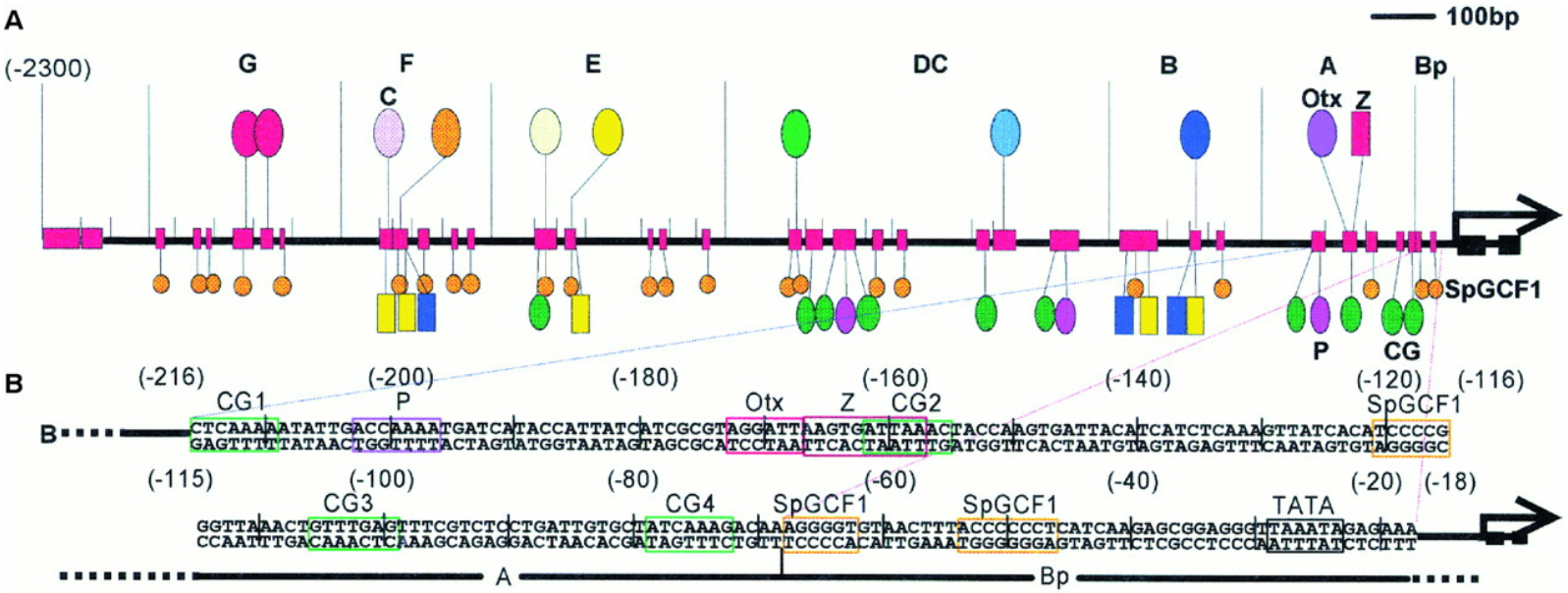
Gene Regulation



Regulatory Networks



Regulatory Networks



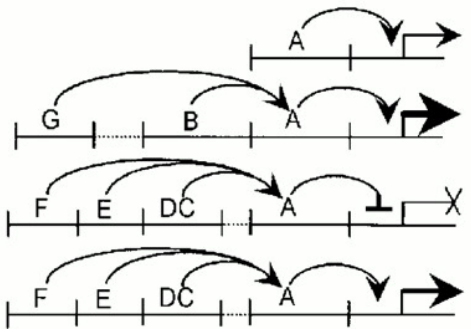
C Module A functions:

Vegetal plate expression in early development:

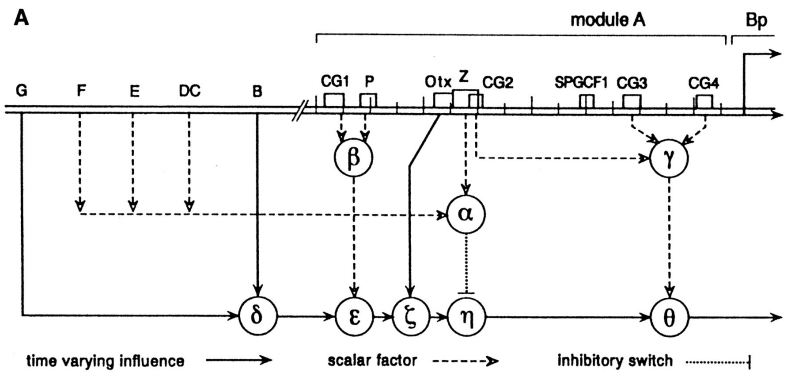
Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:



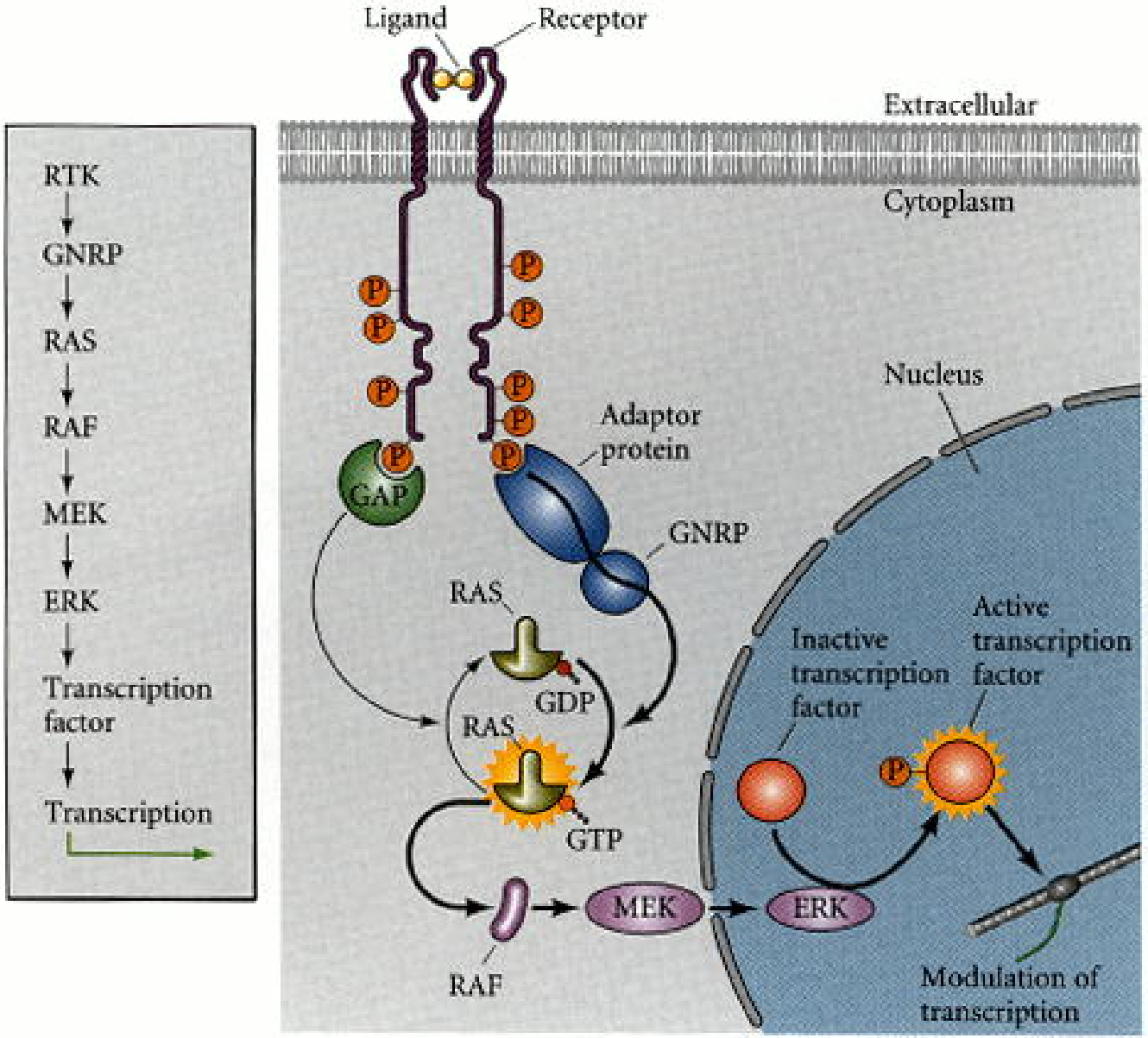
Regulatory Networks



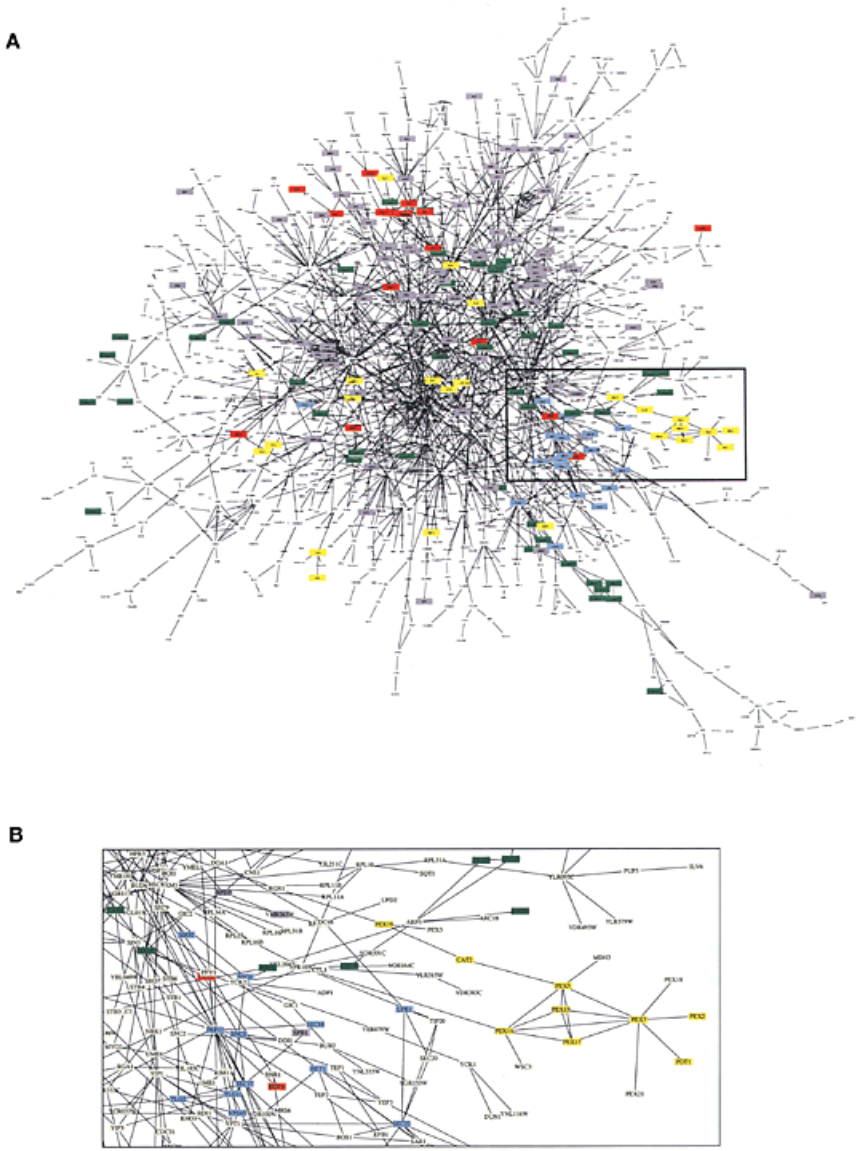
B

if (F = 1 or E = 1 or CD = 1) and (Z = 1)	Repression functions of modules F, E, and DC mediated by Z site
α = 1	
else α = 0	
if (P = 1 and CG_1 = 1)	Both P and CG_1 needed for synergistic link with module B
β = 2	
else β = 0	
if (CG_2 = 1 and CG_3 = 1 and CG_4 = 1)	Final step up of system output
γ = 2	
else γ = 1	
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta * \delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
$\xi(t) = Otx(t)$	
else $\xi(t) = \varepsilon(t)$	
if (α = 1)	Repression function inoperative in endoderm but blocks activity elsewhere
$\eta(t) = 0$	
else $\eta(t) = \xi(t)$	
$\Theta(t) = \gamma * \eta(t)$	Final output communicated to BTA

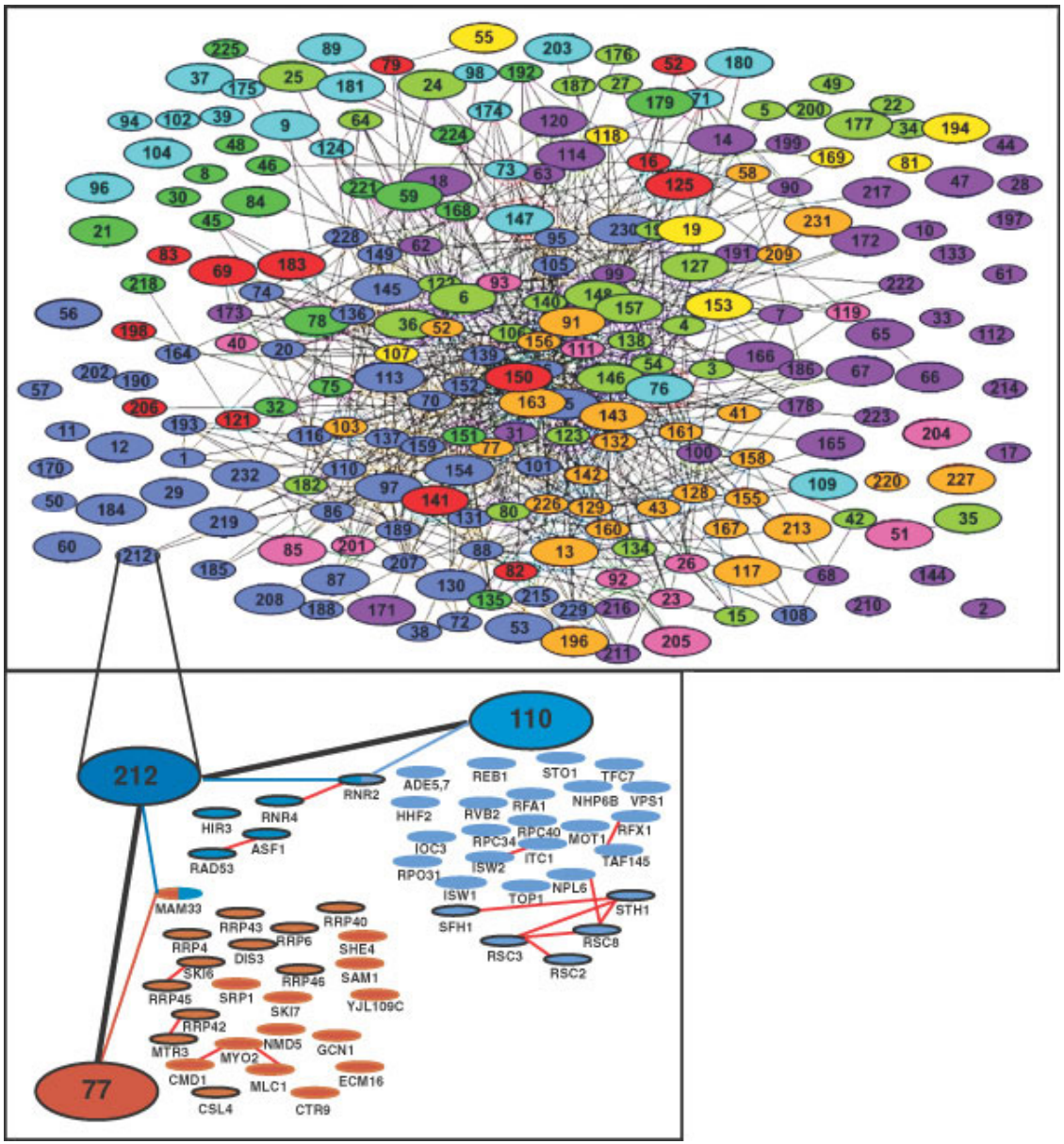
Signal Transduction Cascades



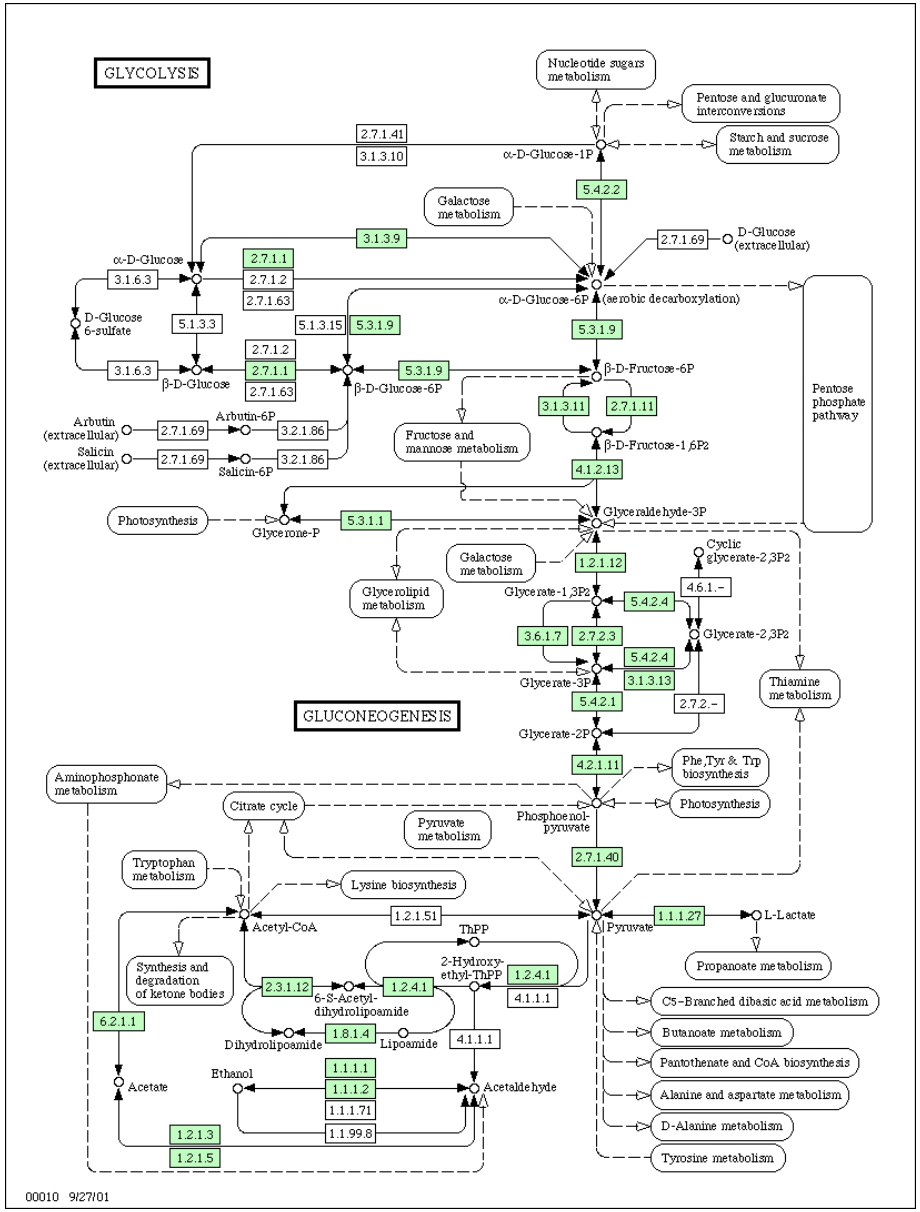
Protein-Protein Interaction Networks



Protein-Protein Interaction Networks



Metabolic Networks



Host-Pathogen Interactions