Network Based Models in Bioinformatics and Biocomputing

Instructors: Dr. Jeff Solka and Dr. Jennifer Weller

Agenda

- Course guidelines
- Discussion of class meeting time
- Part 1 Lecture: Chapter 1 from Davidson encompasses two main goals
  - Introduction to the specific domain to be discussed: cis-regulatory elements in development, rather than the many other types of gene regulatory networks that might be considered
  - Some of the vocabulary that will be used in the chapters
  - Overview of the topics of the chapters

Course Business

- Homework assignments will generally be due two weeks after assignment, in class, as hard copy. Late assignments will not be accepted.
- Quizzes will be primarily multiple choice, closed book
- If you must miss a presentation time you must provide the instructors with **advance** notice and arrange a time to make it up.

Work Expected

- Students are encouraged to discuss problems and assignments with one another but work turned in must reflect your individual efforts.
- Homework assignments and projects must reflect the individual efforts of students.
  - Also note that 70% or more of an assignment must reflect the words and synthesis of a student: a mass of material 'glued together' from Web resources does not constitute original work even if properly cited, and will be graded accordingly.
- Paper presentation (Apr 23): each student must select a paper (from the list or another you choose but first vetted by an instructor) and summarize the method and findings in a presentation to the class.
- Projects (Apr 30): The student must present the project to the rest of the class. A formal project report will also be expected.
  - take a method developed during the course and expand or alter it and then apply it to a dataset presented in one of the papers and compare the results of the two methods
  - apply one of the methods learned in class to a new dataset and discuss how the outcome compares to the previous pathway or to expectations based on other types of knowledge.

Citation policy

- Students are encouraged to read and cite the literature to support their solutions to homework problems.
- When part of a solution for a homework assignment or the project has been found in a reference, whether Web, journal or book, the student must properly cite that source.
- Failure to properly cite the work of another constitutes plagiarism; **both cheating and plagiarism will be grounds for referral to the GMU Honor Council.**
  - Our department has a one-strike policy on cheating and plagiarism: faculty MUST report students accused of either to the Honor Council, this is not at the discretion of the instructor.

Assignments

- Chapter 1 in The Regulatory Genome
- Chapter 1 in Graph Theory
- Homework 1 (Weller section) is posted and also described at the end of the notes here.

Professor Availability

- My drop-in office hours are 1-3 on Mondays.
- Although I am mostly here, I am managing 14 students research efforts this term, so if you drop in without an appointment please do not be offended if I ask you to come back another time. On the other hand I am pretty responsive to e-mail.

Questions?

Development Paradigm

- Development is mediated by regulation of genes encoding proteins that catalyze the creation/turnover of all other cell constituents
  - Spatial component
    - What part of the cell or what subset of cells
  - Temporal component
    - The timing of the expression, the kinetics of the expression of the genes, the time step between states

Gene regulatory states

- A state of the cell is defined by the existing concentration and activation state of regulatory proteins
  - Regulatory states exist as part of a dynamic progression, which in development are not reversible
    - The reason why stem cells are so important
- The regulatory proteins that matter in development are the transcription factors
  - These are DNA-recognizing proteins that modulate the transcription of distant genes.

Cis-regulatory elements

- The DNA sequences recognized by the proteins provide the lock into which the tf key must fit in order to have an effect
  - Individual recognition sites are quite short, so a number of such elements are always combined and must be activated together before a module is effective
    - Some enhance expression and others repress it

Progression in Development

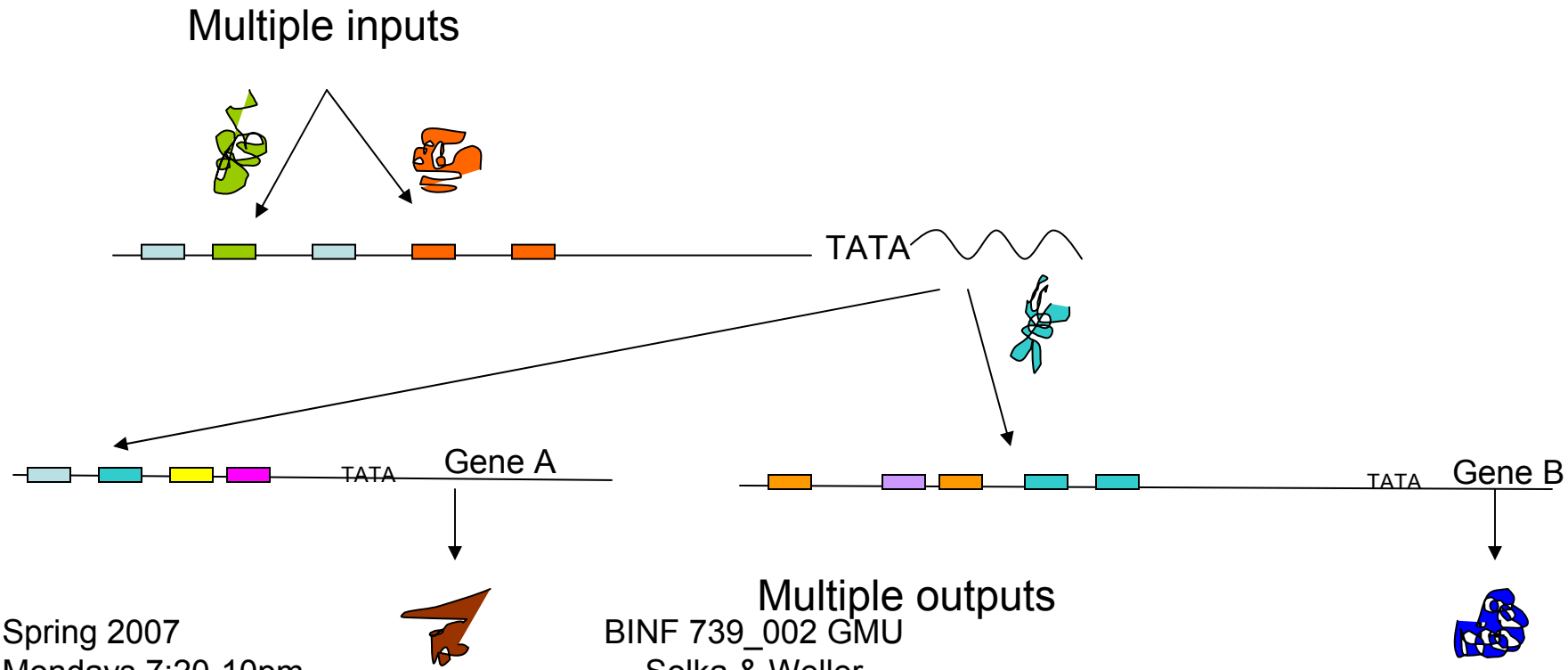
- Alterations in development are caused by changes in how these cis-regulatory modules (or elements, abbreviated CRE hereafter) are organized and combined.
  - The order, distance between them and number of each as well as total all change the effect on modulation.
- All cells in an organism have the same genomic sequence
  - The CREs are the only hard-wired component directing the development sequence
    - They must contain the total of instructions for all conditions requiring a response

Transcription factors

- Transcription factors (abbreviated tf) themselves are regulated by CREs
- Observation 1: signal transduction causes changes in the spatial patterning of gene expression
  - This means that the final termini of these pathways must lead to the regulatory elements controlling tf genes
- Observation 2: Every tf binds to sites at multiple target genes (one to many relationship).
- Observation 3: Genes important in development have multiple tf binding sites
- 2 and 3 together mean that developmental control systems have properties that can be well-represented by graphs, since they are networks having multiple inputs and outputs at a single node.

Network representation

- A logical representation of the CRE is as a node in a network
  - The node is unique in its set of {inputs, outputs}
  - Regulatory genes are invoked at different times in development and at each time perform different functions.
    - A modification of the inputs as well as the available outputs is what leads to the different outcomes



Regulatory sequences

- In terms of numbers and complexity of the regulatory sequences the genome size is irrelevant, but the mRNA complexity is highly correlated
  - For a wide range of bilaterans the complexity of mRNA expressed in the oocyte is comparable
    - Caveat: mature mRNA complexity, there is wide divergence in the intron length, which correlates with genome size, so the pre-mRNA complexity is not as correlated.
- The minimal genome for bilaterans seems to include ~15,000 genes and it goes up to 30/45,000 (so far).
- Two notes of caution:
  - The state of knowledge about genomes: we have only a few completed genomes and they are clustered in a few clades (for eukaryotes)
  - The definition of a ‘gene’ is somewhat changeable since alternate start sites and alternate splicing give internal differences
    - So far the maximum number of alternate start sites is 3 (for isoforms it is 30 and counting).

Regulatory genes

- Some regulatory genes are organism (or at least clade) specific
  - Chemosensors, immune functions
- Many regulatory genes are highly conserved
  - Cell differentiation mechanisms, cytological structure mechanisms, and of course enzymatic functions such as cell cycle and mitosis
  - The differences lie in how many are made, as well as protein coding modifications that alter activity and specificity
- If you focus on tf genes and the signaling systems required in development there are few differences across clades of bilaterans
  - Differences are in the number expressed, their spatial distribution and temporal program rather than whether they exist.

CRE sequence conservation

- Genes are identified by having a sequence-dependent constraint in the rate of change through generations (evolution)
  - The protein-coding and structural RNAs are best studied
- Background, or selectively neutral, rates of change are usually benchmarked on sequences such as introns, wobble bases and intergenic DNA
- Non-coding DNA in the genome does not all show the same rate of change.
  - If you assume that the constraint on CREs is at least as strict as for protein sequences, then there appears to be at least as large a fraction of non-coding DNA whose sequence is strictly conserved as conventional genes
    - An assumption is that a large fraction of this is regulatory sequences
- True intergenic DNA has several characteristics
  - In many genomes there is a significant component of repeated elements, from microsatellites to transposable elements
  - Other regions that are sequence unconstrained appear to fulfill a topology requirement – giving enough space for distance elements to be brought into proximity by looping, for example.
    - The looping is not on a tight tether, but rather allows reconfigurable and dynamic alignments to occur – this allows combinatorics to apply.
    - Another element is chromatic structure, the higher-order organization of elements.

Using tf Combinations

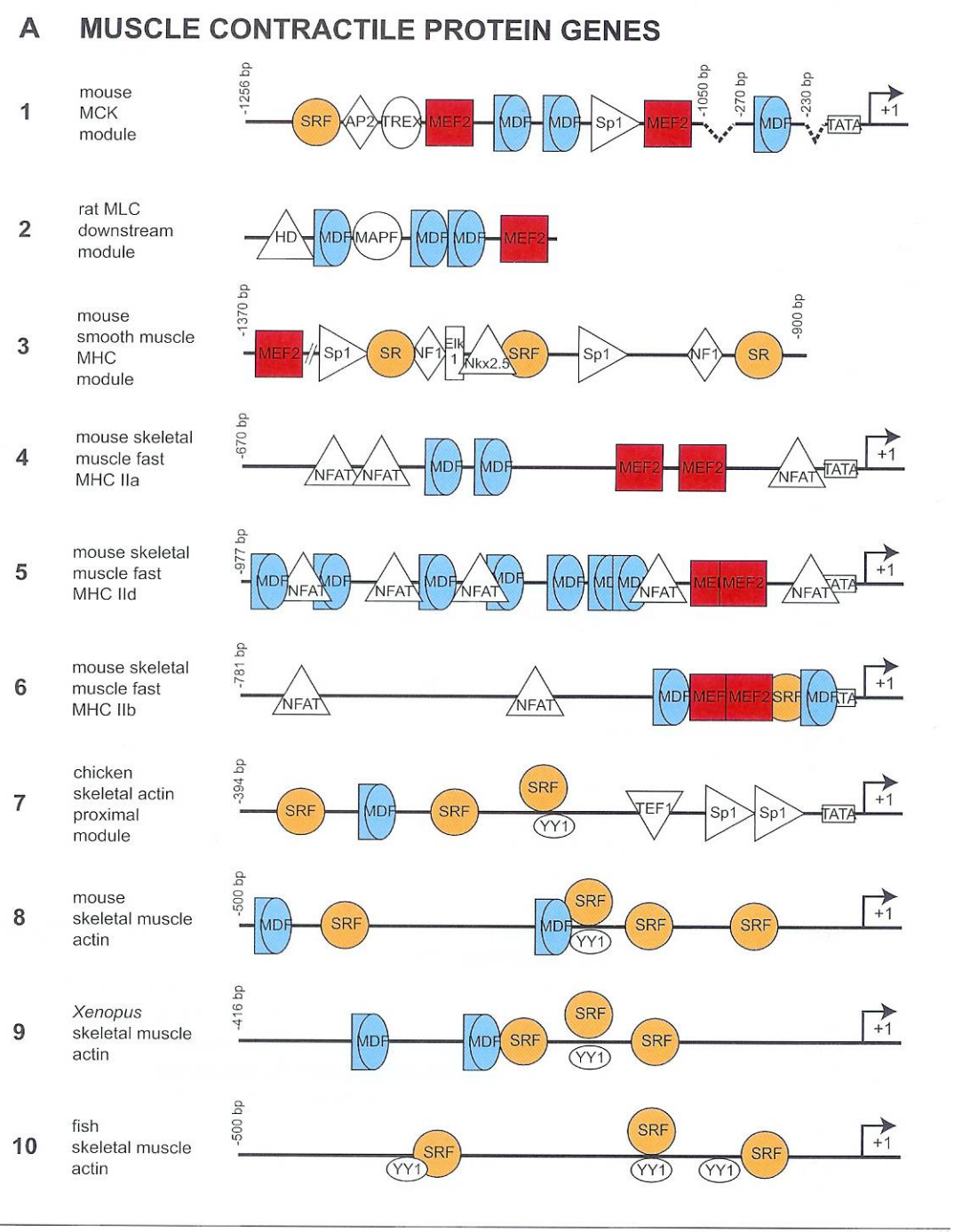
- Why is it harder to find CREs than gene sequences?
  - Length, spacing and order
- Each tf recognizes quite a short length of DNA.
  - Accidental occurrences of short sequences are quite likely, which would be a false positive signal for an event
  - The insurance against this is to require dense clusters of tf binding site units, in non-random order.

Cis-and trans-components of regulatory modules

- For the genes active in developmental processes, cis-regulatory modules are usually located on the same DNA molecule as the gene controlled (cis).
  - There are usually 5-10 binding sites (for multiple tfs, but one can be represented more than once) per module.
- Transcription factors are proteins, that can diffuse, so they are often made on the other DNA strand from the gene, at a great distance (other chromosomes, for example).
- There are functional classes of each component, which we will be studying in more detail in this class.
  - Enhancers are modules that cause up-regulation of gene transcription. They are often many kbp away from the promoter of the gene regulated.
    - Looping must occur to achieve proximity AND directionality is optional since looping can occur in either orientation.
  - Many CRE work indirectly, but affecting something that then affects the transcriptional apparatus

Multiplicity of inputs

- A cis-regulatory element has (as stated) 5-10 tf binding sites, on average.
  - The binding of any one set of TFs causes a particular instruction to be given to the basal transcriptional apparatus
    - The result is either gene silencing or gene transcription at a particular rate
  - Changes in the occupancy of the set of binding sites results in a different set of instructions.
    - On average 4-8 tfs bind a given module at one time.
- Some of the changes are multiplicative: if the outcome is that a new tf is made, then this output has become multiple inputs for new genes
  - The network grows.
  - The output is never recursive: input ≠ output for the same gene



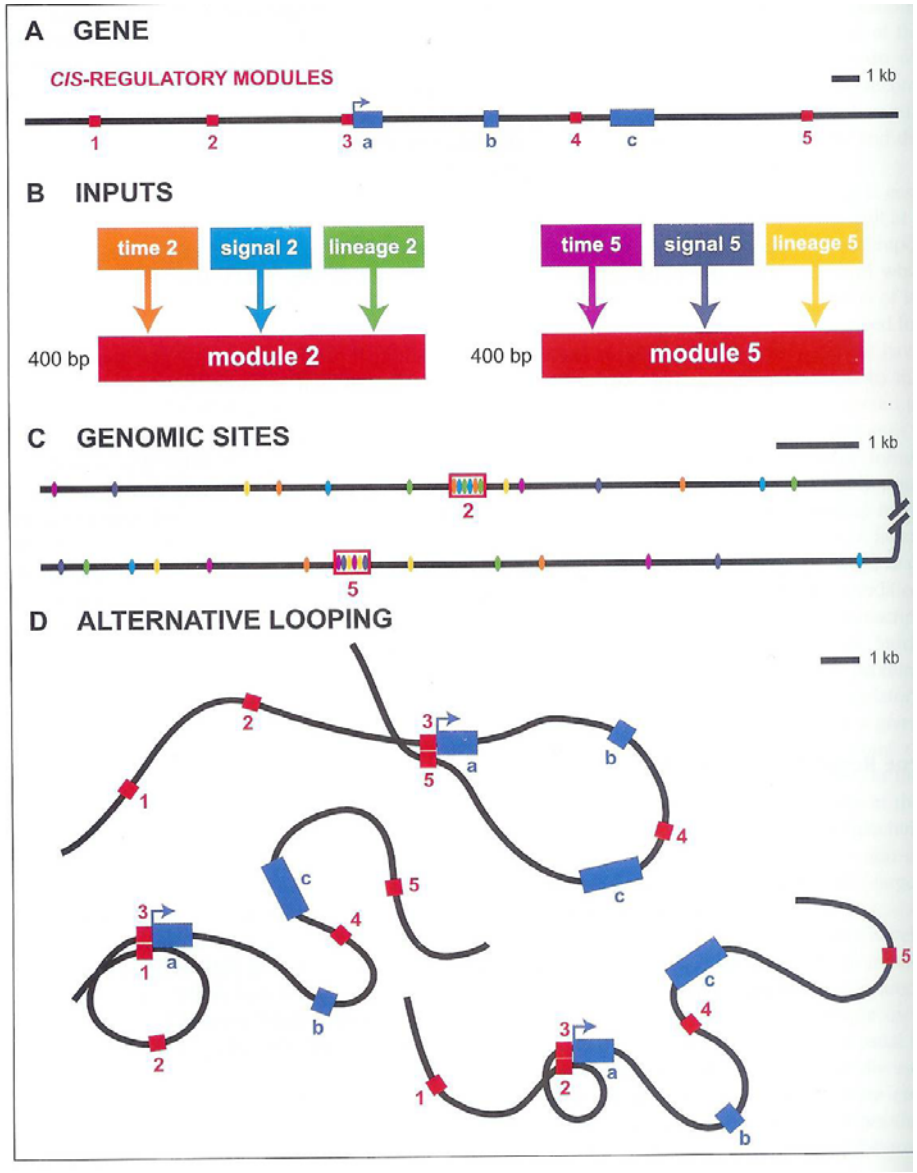
- The use of TFs for similar functions in different organisms, or slightly different functions in the same organism.

Regulatory State Dynamics

- A regulatory state is given by the concentration and activity of the tfs present.
  - This provides the cell with all it can ‘know’ about its own status and that of its neighbors, and sets the limits to what it can do next
    - The concentrations reflect the stage just past
    - The activities determine (predict) the next stage
  - The activities reflect the signaling pathway inputs, since these frequently cause modification of the tfs, or provide an essential cofactor.

tf spatial distribution

- A tf may be made in cells in only one part of an organism, or activated in only one part by signaling molecule inputs
- Once a tf binds to a CRE, it often initiates the formation of a large protein complex with complex responses to cell conditions
  - The other components of the complex are secondary to the hard-wired control of the developmental process (even though required for successful progression)
- A gene often has several CREs, one for each of the times it is expressed in development
  - The average CRE is several hundred bp in length and they may be separated by several thousand basepairs. They may occur at either the 5' or 3' end or even in introns.
- Another level of control that must be exerted is looping control – if a loop has to form with one CRE and not the other then proteins must regulate this process by recognizing some sequence.
  - These are almost completely uncharacterized



- Note that the CREs are not all on the 5’ side of the gene
- Different looping modes are shown to illustrate the ways that two CREs can be brought into proximity

Developmental Outcomes

- Developmental processes frequently result in morphologically distinct structures (gives a phenotype to measure)
- Network architecture in this evo-devo context means the following the topology of functional linkages.
  - Early stages have the multiplicative effect where an output becomes the input to several other genes
  - The periphery of the network is recognized because the gene products do something physical, rather than affect more gene transcription
  - Sub-networks in the spatial domain are linked by signaling inputs.
- An important aspect of development is the end-stage or terminal differentiation: how do you use CREs to set up stable, long-term expression of the genes defining a cell type after the initial, transient event that begins the process?
  - You want to minimize the maintenance load

Developmental processes

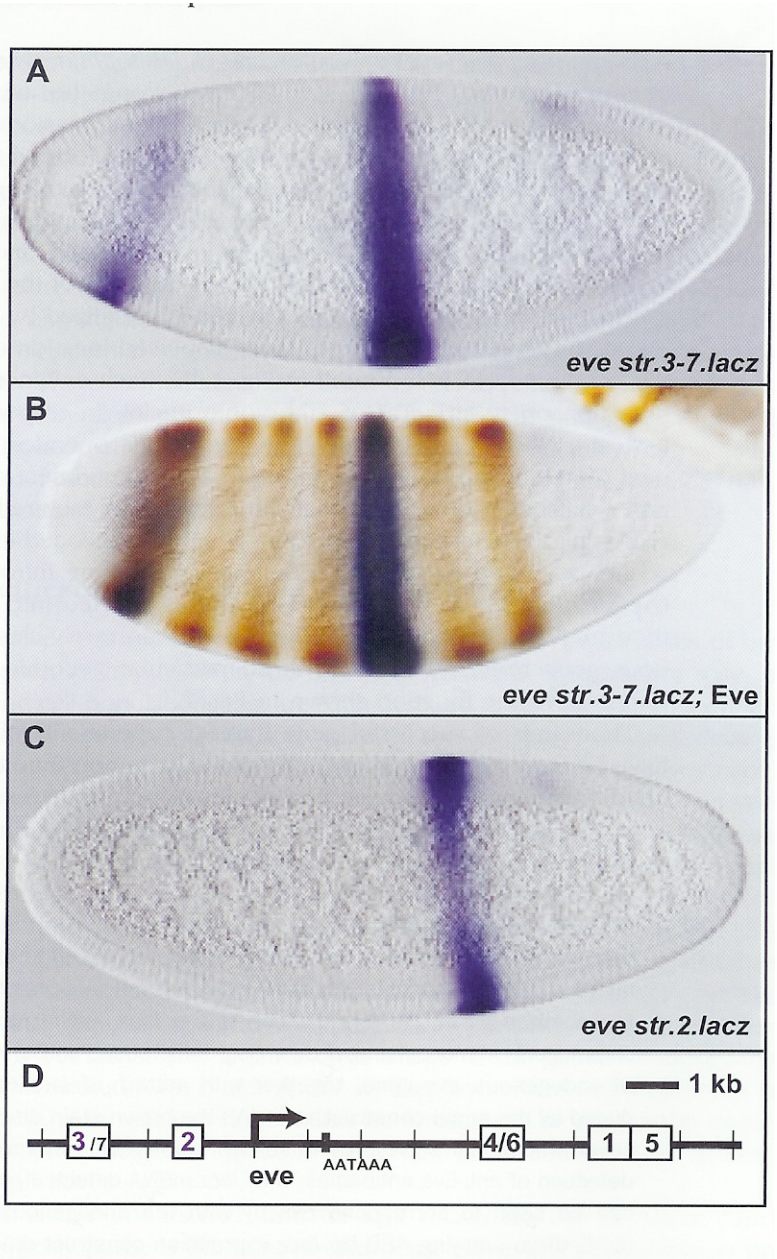
- Development proceeds with an increasing complexity in the number and types of cells
  - The cells communicate and to some extent co-regulate events.
- The stages of development are self-referential: the subsequent stage only occurs when signal are received that the previous stage has been successful
  - Unlike a building, you do not build a skull and a hand in isolation and connect them up later.
  - The inputs available are sufficient to define the next regulatory state and the availability of the inputs is supplied internally by tfs acting in the nucleus.

Cell specification

- The cell identity (specification) is set by the CREs plus the genes expressed at their direction, which carry out the prescribed functions
- The CRE gives the mechanism to integrate the spatial signaling inputs and intracellular inputs to evoke a ‘decision’, or progression to a new cell state.
- A transient change initiates a specification, but the new state is irreversible, so it in turn progresses to a new state.
  - The rate of that progression is organism-specific
    - Drosophila: a state may last only minutes
    - Sea urchin: a state lasts several hours
  - Transience: the signaling molecules have a finite life (this is temporal control)
    - You must be able to silence as well as turn on genes
    - The gene products (mRNA or protein) must be subject to turnover (this gives the kinetics of the dynamical system).

Pattern formation

- In development, pattern formation is the prevailing phenotype.
  - That is, there is control of spatial gene expression so that the organization of cells is controlled, and the further organization of bits of those in the right place.
    - This can be within a cell as well as between adjacent cells and out to a whole body plan
    - Further specialization occurs within bounded regions that are called pattern elements.



- Pattern formation in Drosophila embryos – the mRNA is shown in blue and the protein in brown. The gene expressed is the evenskipped protein.
- The CREs function properly even when introduced into different parts of the genome.

Terminal Differentiation

- At the end of development, cells have specialized and any that they produce are clones of themselves, with the same CRE settings and the same gene product outputs.
  - E.g. red blood cells make hemoglobin
- A battery of differentiation genes are coordinately expressed because the CREs share target sites for the same tfs.
  - Genes of a given battery (the most common form of peripheral gene regulatory networks) respond to the same tfs but never have identical CREs.
    - Different subsets of tfs bind and the binding motifs in the CREs are never in the same order, number or spacing, so uniqueness is preseved.

Locking down differentiated states

- There are ways of locking CREs into a configuration that ensures that the status of genes needed for expression (or to not be expressed) remains correct in terminal differentiated states.
  - Mechanisms include methylation of CpG islands, methylation and acetylation of histones to create condensed chromatin states and protein complexes that inhibit (polycomb) or induce (Trithorax) expression.
  - Some of these can be transmitted to the next generation.

Mechanisms of later development

- Alternate splicing and microRNA induction are used to modulate function and thus output in developmental stages farther down the process.
  - microRNAs act somewhat like tfs in that a gene is turned on to affect another gene
    - They act at the gene product level rather than the CRE of the gene
    - They however are regulated by CREs

CREs and Evolution

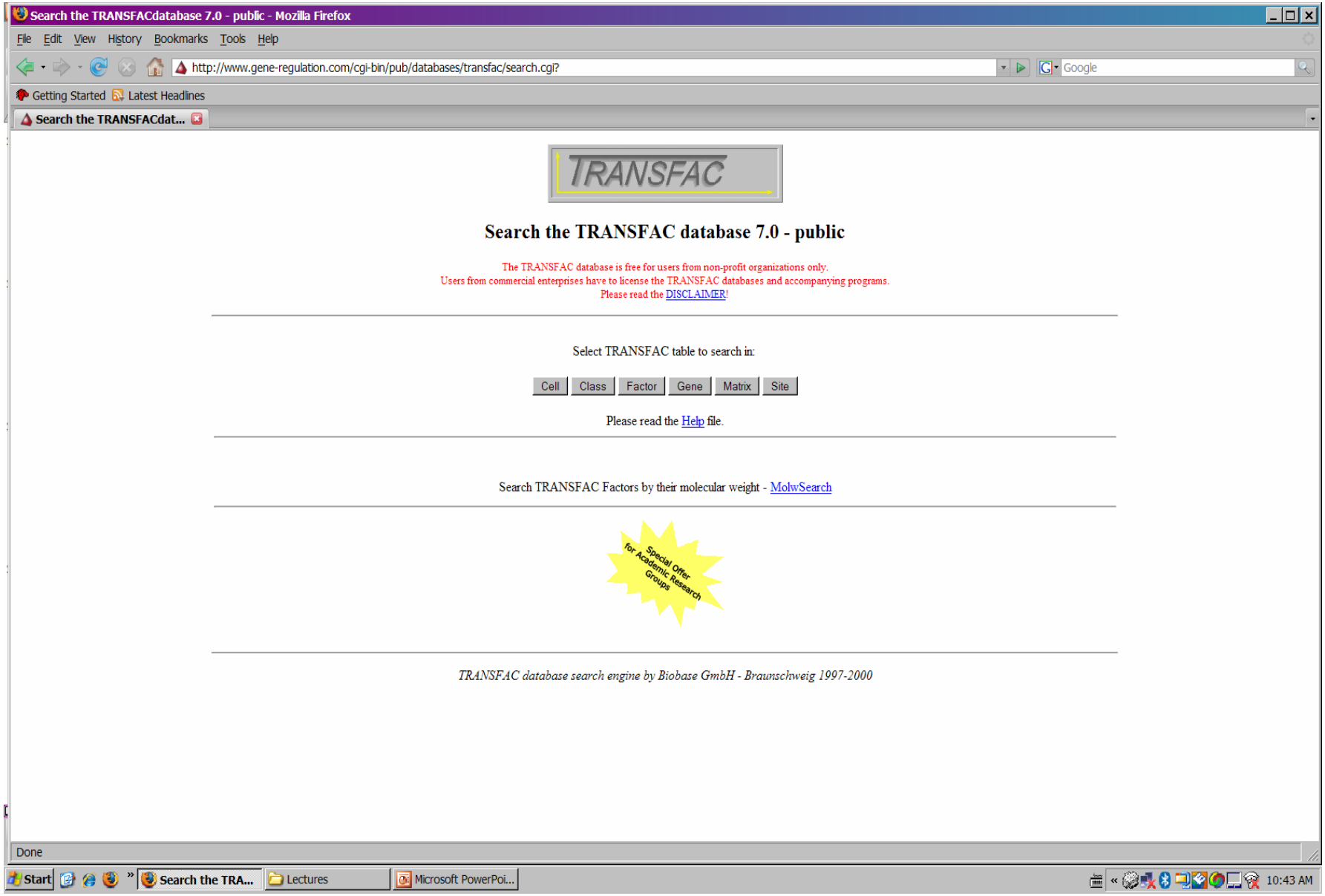
- Since the basic units are so conserved, a compelling question is how the features are altered, reused and recombined to get new structures, and what are the consequences of specific changes.
  - Paleontology and phylogeny show what is possible and give some idea of how near- and distant- relatives arrange similar material for different outcomes but only genomics studies can give the true mechanisms.
  - Near relatives usually differ in size and the products of terminal differentiation gene batteries
    - The number of rounds of cell division of particular organs, slight change sin properties of a protein that allow new wavelengths of light to be perceived, or types of cellulose linkages to be degraded.
  - Far relatives have more fundamentally different morphology and life habits, and in this case it is clear that CREs from earlier developmental stages must be coming into play.
    - Comparative genomics at this level is the next great leap in understanding developmental rules.

Acknowledgments

- The images on pages 21, 24 and 29 are from the course textbook by Eric Davidson.

Homework 1

- There is a database of eukaryotic transcription factors and their binding sites called TRANSFAC. It is described in the publication: Heinemeyer et al., 1998, Nucleic Acids Res. 26:364-370.
- It can be accessed at: http://www.gene-regulation.com/
  - You will have to register for an account but it is free
- You can visualize the items of the feature list (FACTOR table) as well as element positions within regulatory regions (GENE table).
- These are experimentally derived data: with 2285 FACTOR and 4602 SITE entries.
-


Search the TRANSFACdatabase 7.0 - public - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi?
Google
Getting Started Latest Headlines
Search the TRANSFACdat...
TRANSFAC
Search the TRANSFAC database 7.0 - public
The TRANSFAC database is free for users from non-profit organizations only.
Users from commercial enterprises have to license the TRANSFAC databases and accompanying programs.
Please read the DISCLAIMER!
Select TRANSFAC table to search in:
Cell Class Factor Gene Matrix Site
Please read the Help file.
Search TRANSFAC Factors by their molecular weight - MolwSearch
Special Offer for Academic Research Groups
TRANSFAC database search engine by Biobase GmbH - Braunschweig 1997-2000
Done
Start
Search the TRA...
Lectures
Microsoft PowerPoi...
10:43 AM

HW 1 questions

- Make an account. Read the paper and the Help information.
- Question 1: how is the determination made to accept a transcription factor and the binding site into the database?
- Question 2: for the factors NFAT, SRF, MDF, Sp1 and MEF2, find the information for two species.
- Question 3: Compare the binding sites for each transcription factor between the species.
- Question 4: How conserved are the protein sequences between the two species?