

Preview: Some illustrations of graphs in Integrative Genomics

- *Biological Graphs and their models/combinatorics*
- *Genomics → Transcriptomics: Alternative Splicing*
- *Genomics → Phenotype: Genetic Mapping*
- *Comparative Biology: Evolution of Networks*

Networks in Cellular Biology

Dynamics - *Inference* - *Evolution*

A. Metabolic Pathways

Enzyme catalyzed set of reactions controlling concentrations of metabolites

B. Regulatory Networks

Network of {Genes \rightarrow RNA \rightarrow Proteins}, that regulates each other transcription.

C. Signaling Pathways

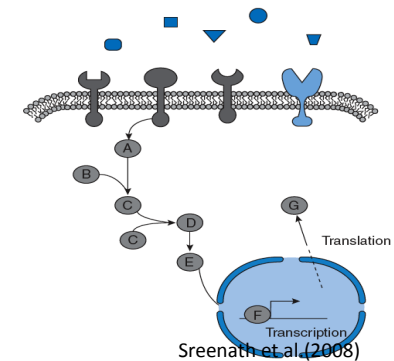
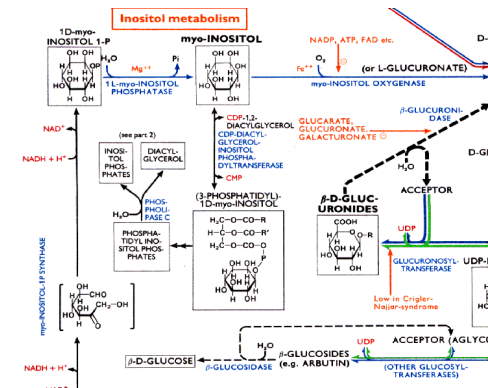
Cascade of Protein reactions that sends signal from receptor on cell surface to regulation of genes.

D. Protein Interaction Networks

Some proteins stick together and appear together in complexes

E. Alternative Splicing Graph (ASG)

Determines which transcripts will be generated from a genes



A repertoire of Dynamic Network Models

To get to networks:

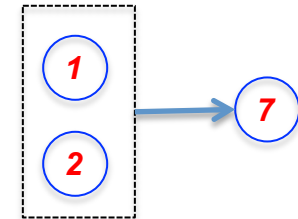
No space heterogeneity → molecules are represented by numbers/concentrations

Definition of Biochemical Network:

- *A set of k nodes (chemical species) labelled by kind and possibly concentrations, X_k*



- *A set of reactions/conservation laws (edges/hyperedges) is a set of nodes. Nodes can be labelled by numbers in reactions. If directed reactions, then an inset and an outset.*



- *Description of dynamics for each rule.*

ODEs – ordinary differential equations $\frac{dX_7}{dt} = f(X_1, X_2)$

Mass Action $\frac{dX_7}{dt} = cX_1X_2$

Time Delay $\frac{d\bar{X}(t)}{dt} = f(\bar{X}(t - \tau))$

Discrete Deterministic – the reactions are applied.

Boolean – only 0/1 values.

Stochastic

Discrete: the reaction fires after exponential with some intensity $I(X_1, X_2)$ updating the number of molecules

Continuous: the concentrations fluctuate according to a diffusion process.

Number of Networks

- *undirected graphs*

$$\alpha_n = 2^{\frac{n(n-1)}{2}}$$

- *Connected undirected graphs*

$$c_n = \alpha_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} c_k \alpha_{n-k}$$

- *Directed Acyclic Graphs - DAGs*

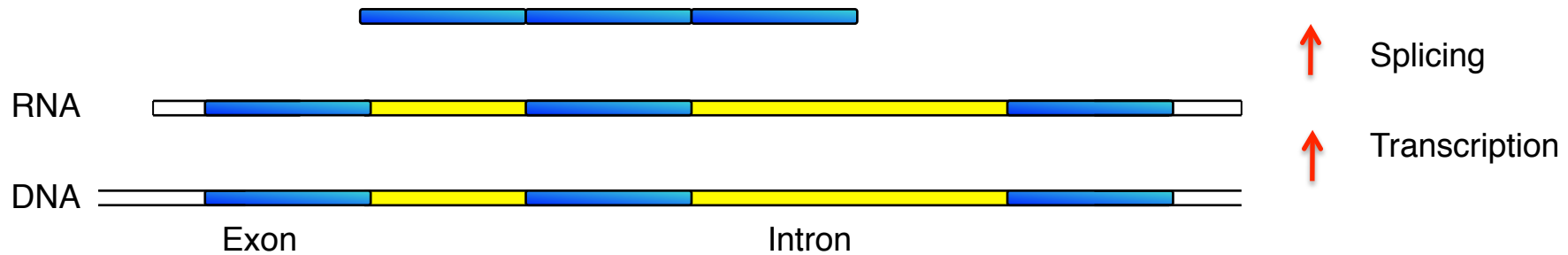
$$a_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

- *Interesting Problems to consider:*

- *The size of neighborhood of a graph?*
- *Given a set of subgraphs, how many graphs have them as subgraphs?*

Genomics → Transcriptomics: Alternative Splicing

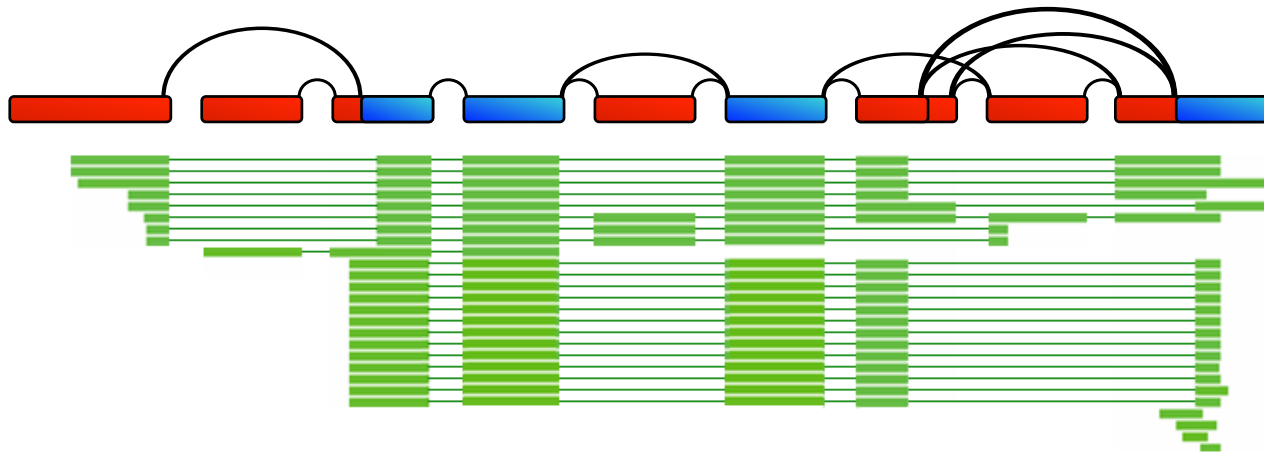
- *AS: one genomic segment can create different transcripts by skipping exons (sequence intervals)*



Problem: Describe the set of possible transcripts and their probabilities.

Define the alternative splicing graph (ASG) –

- ❑ *Vertices are exon fragments*
- ❑ *Edges connect exon fragments observed to be consecutive in at least one transcript*
- ❑ *This defines a directed, acyclic graph*
- ❑ *A putative transcript is any path through the graph*



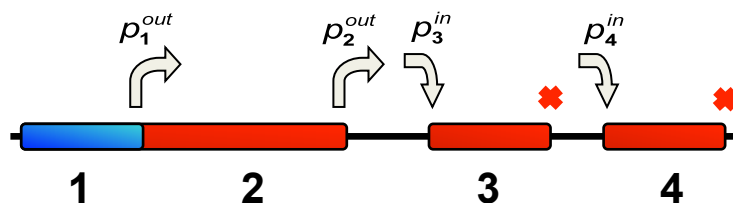
G→T: Alternative Splicing

Problem: Inferring the ASG from transcripts

- Maximally informative transcripts
- Minimally informative transcripts
- Random transcripts

A Hierarchy of Models can be envisaged

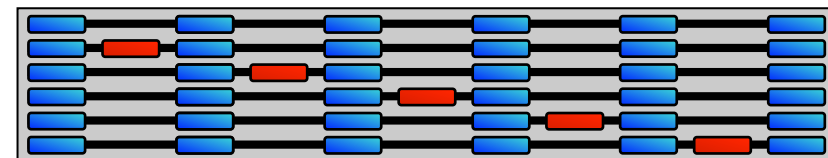
*Simpler still: model 'donation' and 'acceptance' separately
 Jump 'in' or 'out' of transcript with well-defined probabilities
 Isolated exons are included independently, based only on the
 strength of its acceptor site*



This ASG could have been obtained from as few as two 'informative' transcripts...



...or as many as six. There are 32 putative transcripts.

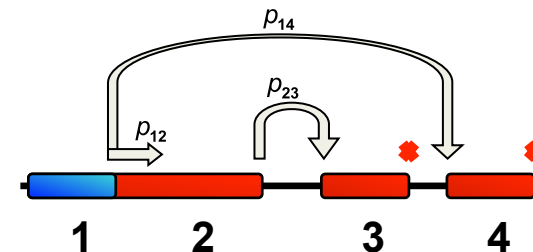


Enrich the ASG to a Markov chain

Pairwise probabilities

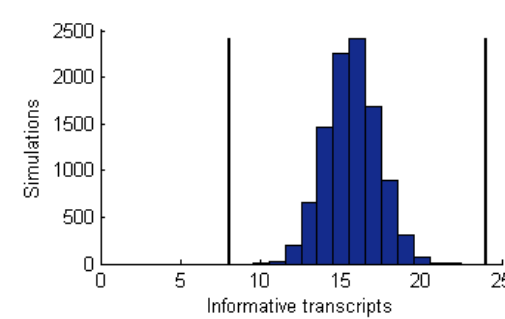
Transcripts generated by a 'walk' along the ASG

A natural model for dependencies between donors and acceptors



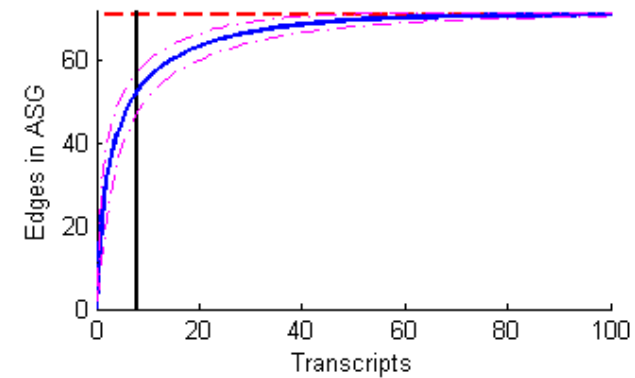
G → T: Alternative Splicing

- *The distribution of necessary distinct transcripts*

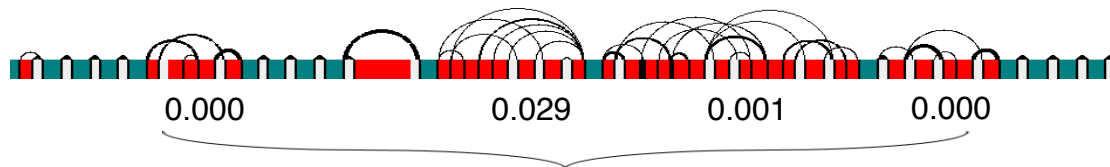


Human gene ABCB5
Paul Jenkins from Leipzig et al. (2004) "The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome"

- *The size of the inferred ASG*



- *Testing nested ASG modes*



Pairwise model: V^2 parameters

In-out model: V parameters

Models can be nested:

In-out \subseteq pairwise \subseteq non-parametric

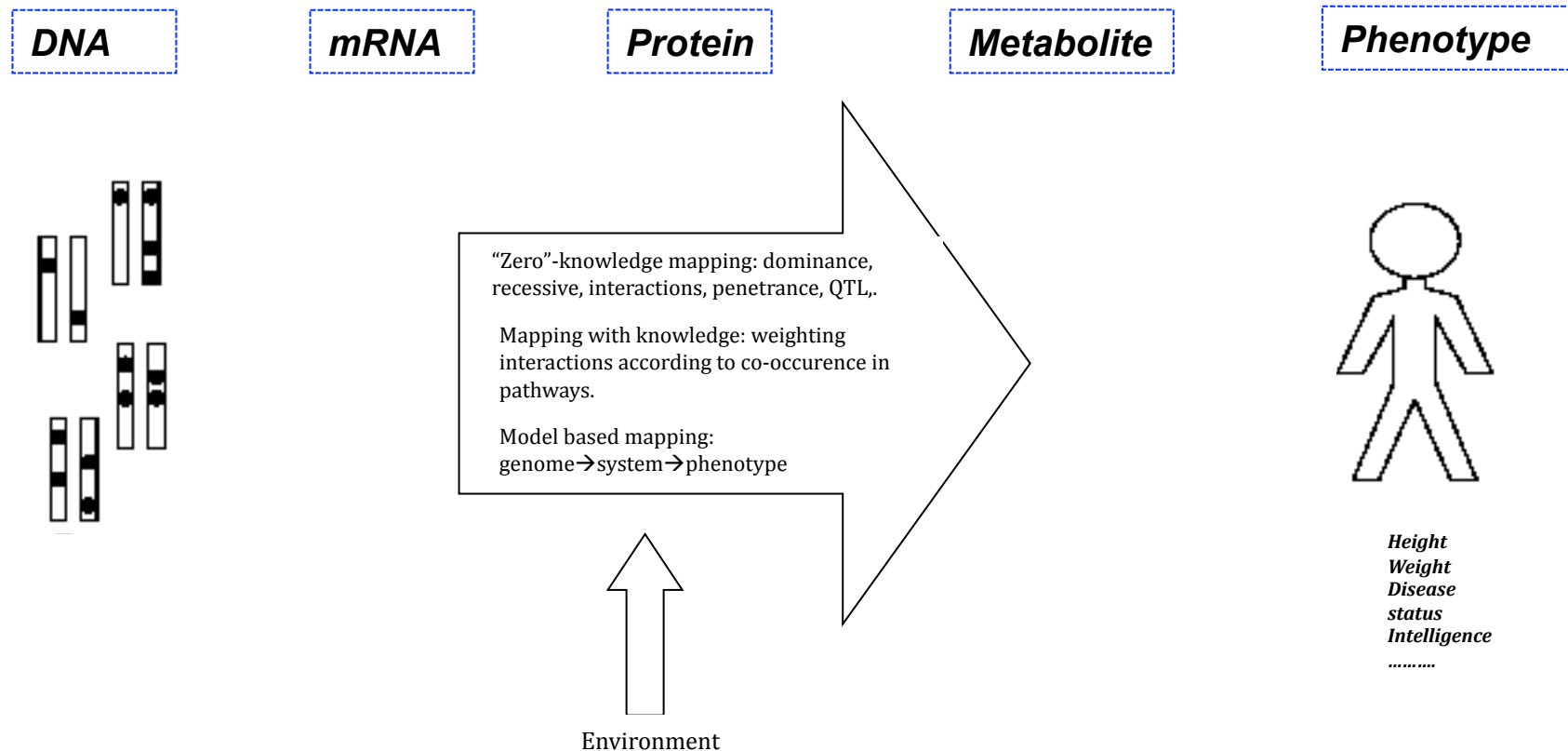
Hence, given sufficient observations, likelihood ratio tests can determine the most appropriate model for transcript generation

The pairwise model was accepted, In-Out rejected

$G \rightarrow F$

- *Mechanistically predicting relationships between different data types is very difficult*
- *Empirical mappings are important*
- *Functions from Genome to Phenotype stands out in importance*

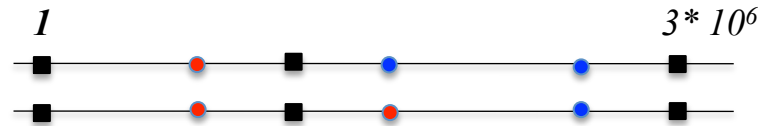
G is the most abundant data form - heritable and precise. F is of greatest interest.



The General Problem is Enormous

Set of Genotypes:

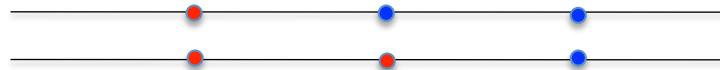
- Diploid Genome**



- In 1 individual, $3 * 10^6$ positions could segregate*
- In the complete human population $2 * 10^8$ might segregate*
- Thus there could be $2^{200.000.000}$ possible genotypes*

Partial Solution: Only consider functions dependent on few positions

- Causative for the trait**



Classical Definitions:

- Single Locus**

Dominance

Recessive

Additive

Heterotic

- Multiple Loci**

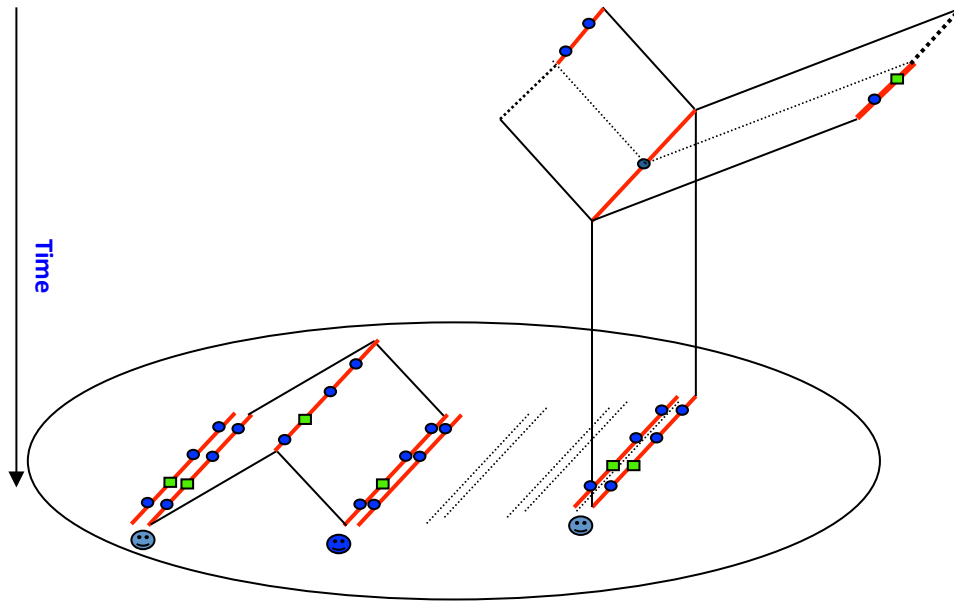
Epistasis: The effect of one locus depends on the state of another

Quantitative Trait Loci (QTL). For instance sum of functions for positions plus error term.

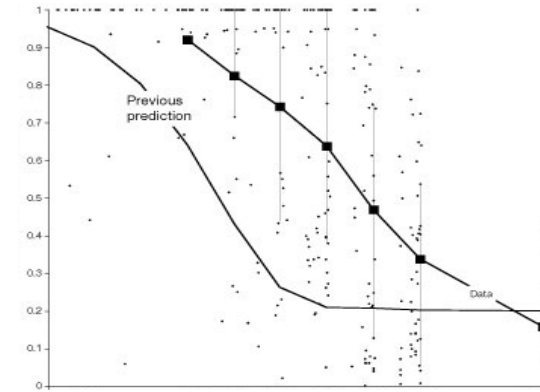
$$\sum_{i \text{ causative positions}} X_i(G_i) + \varepsilon$$

Genotype and Phenotype Co-variation: Gene Mapping

Sampling Genotypes and Phenotypes

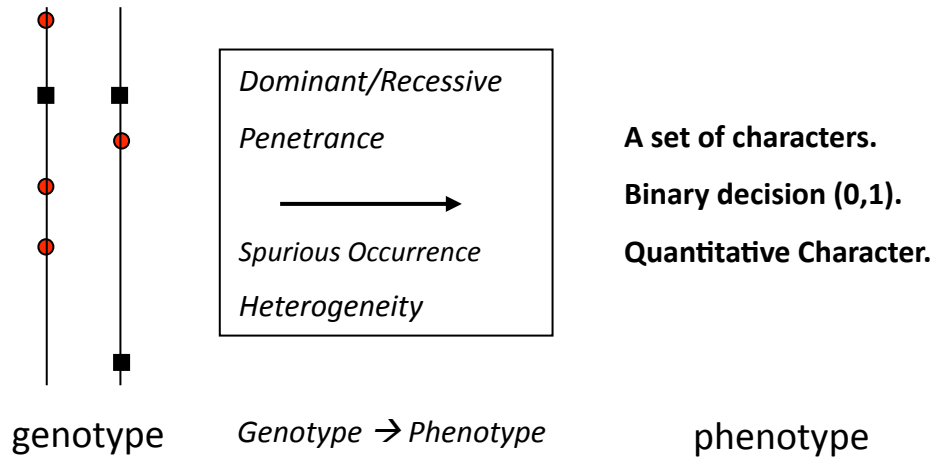


Decay of local dependency

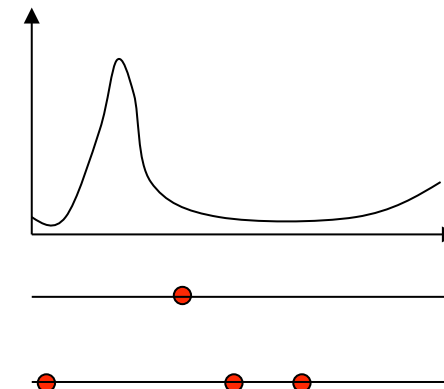


Reich et al. (2001)

Genotype -->Phenotype Function

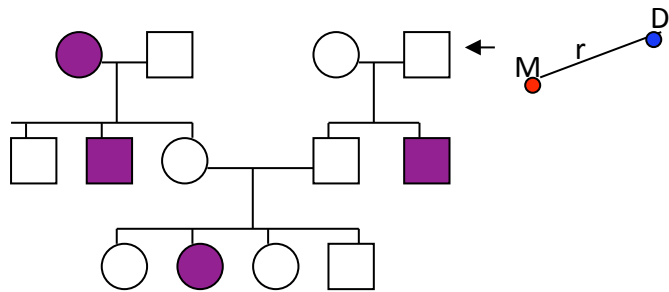


Result: The Mapping Function



Pedigree Analysis & Association Mapping

Pedigree Analysis:

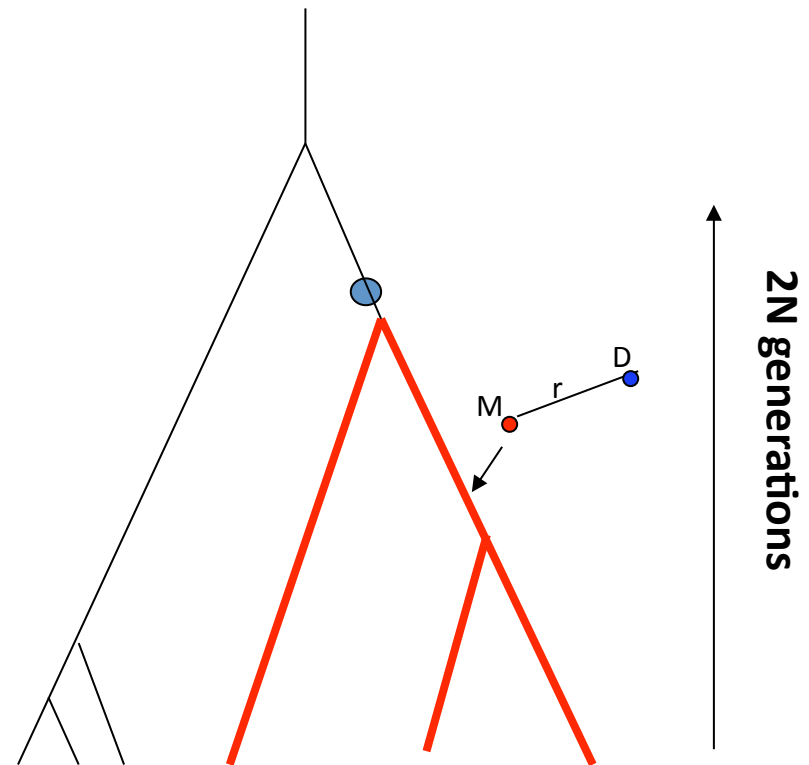


Pedigree known

Few meiosis (max 100s)

Resolution: cMorgans (Mbases)

Association Mapping:



Pedigree unknown

Many meiosis ($>10^4$)

Resolution: 10^{-5} Morgans (Kbases)

Adapted from McVean and others

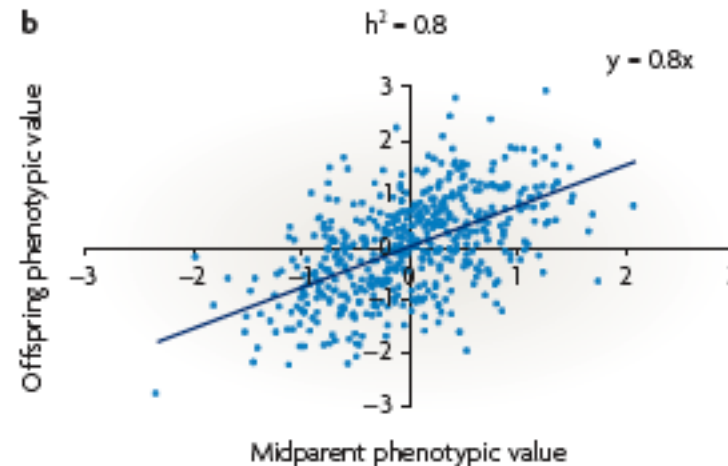
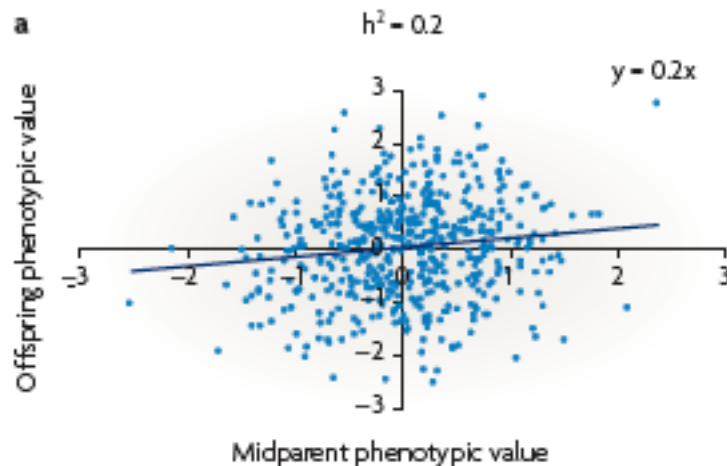
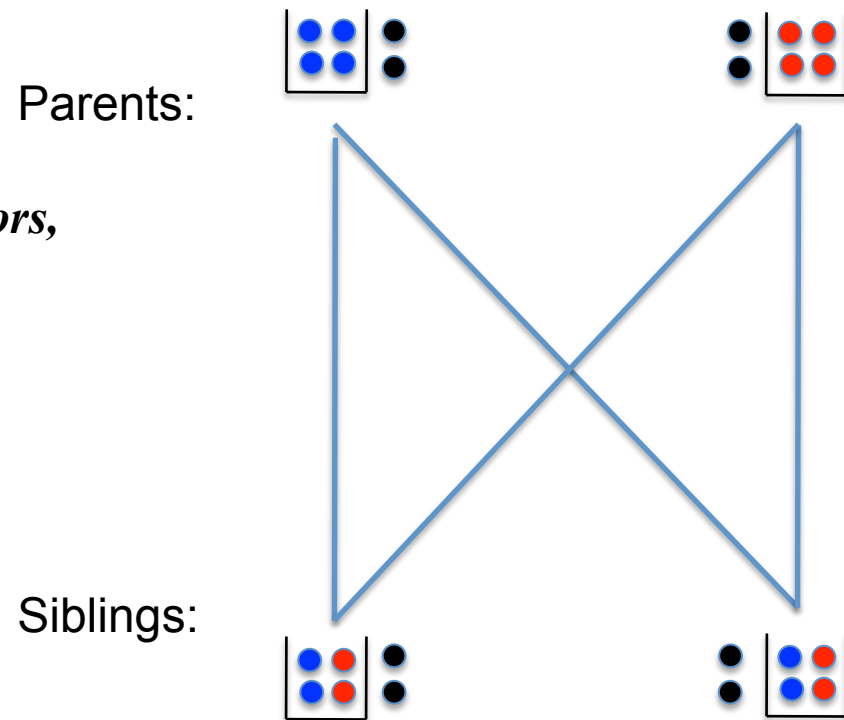
Heritability: Inheritance in bags, not strings.

The Phenotype is the sum of a series of factors, simplest independently genetic and environmental factors: $F = G + E$

Relatives share a calculatable fraction of factors, the rest is drawn from the background population.

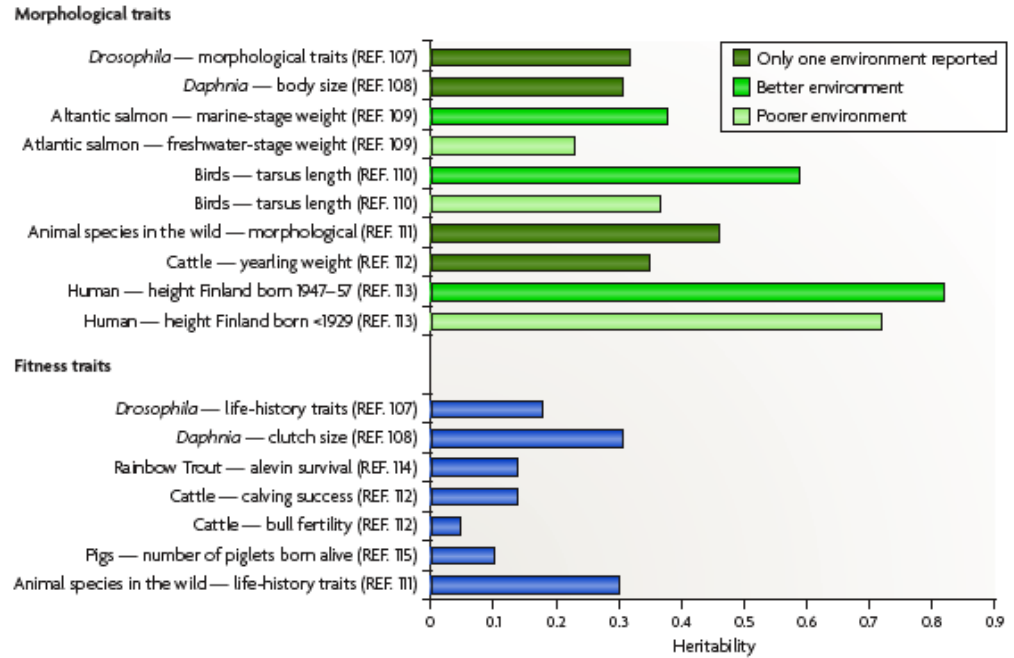
This allows calculation of relative effect of genetics and environment

Heritability is defined as the relative contribution to the variance of the genetic factors: σ_G^2 / σ_F^2

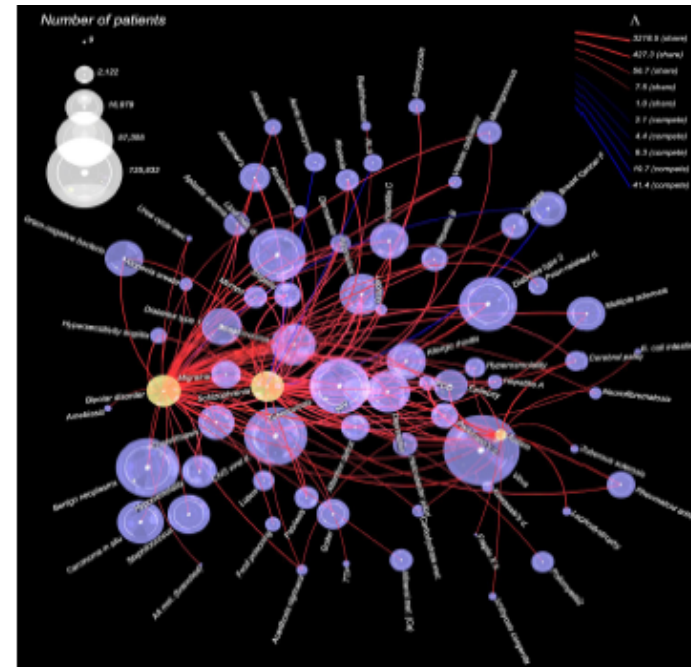
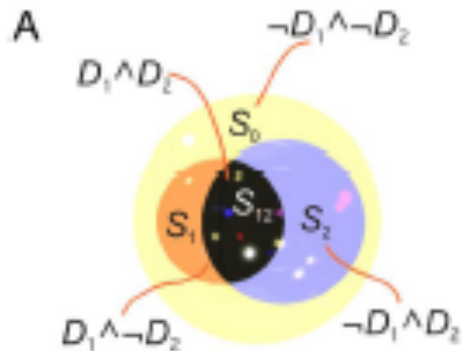


Heritability

Examples of heritability



Heritability of multiple characters:

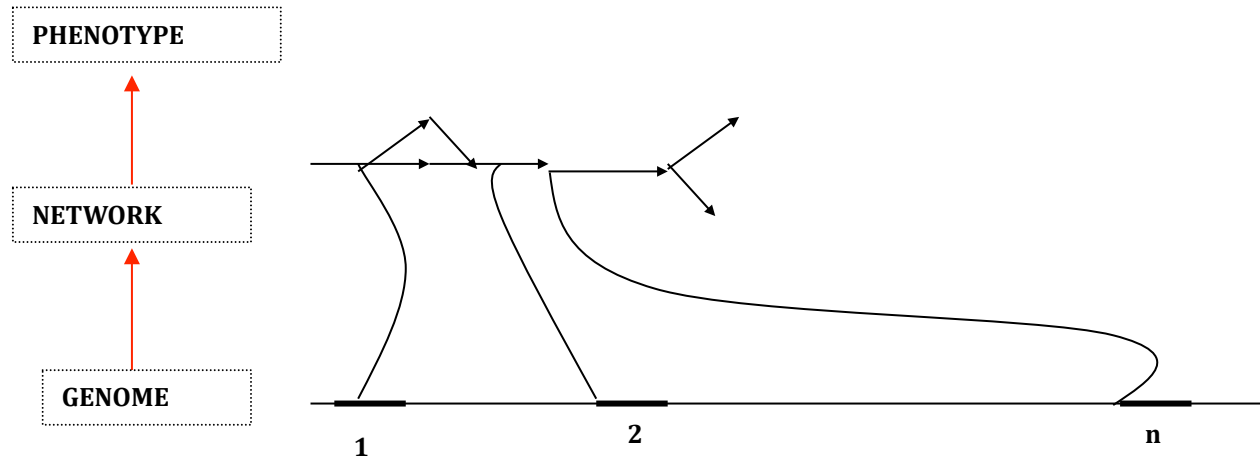


Rzhetsky et al. (2006) Probing genetic overlap among complex human phenotypes PNAS vol. 104 no. 28 11694–11699

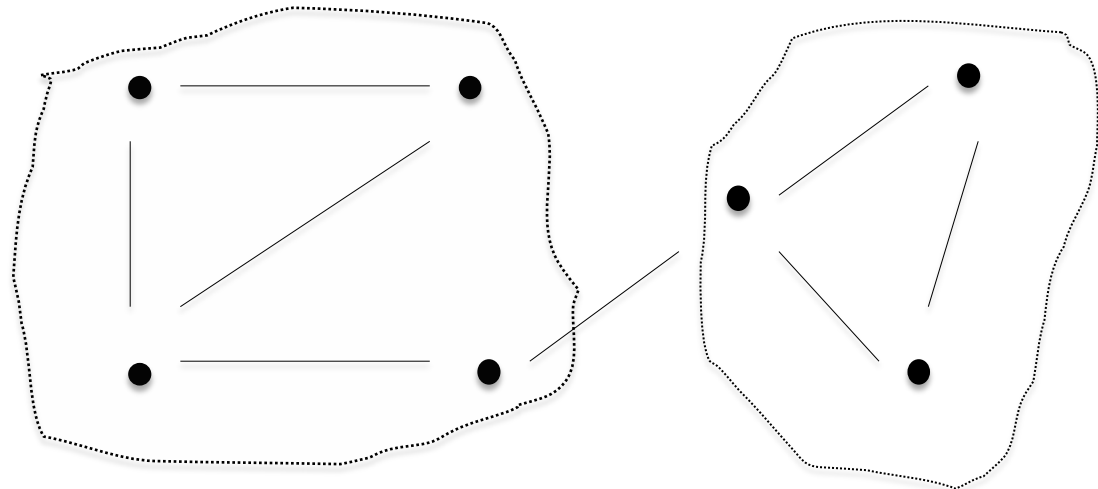
Visscher, Hill and Wray (2008) Heritability in the genomics era — concepts and misconceptions nATurE rEVIEWEWS | genetics volume 9.255-66

Protein Interaction Network based model of Interactions

The path from genotype to phenotype could go through a network and this knowledge can be exploited



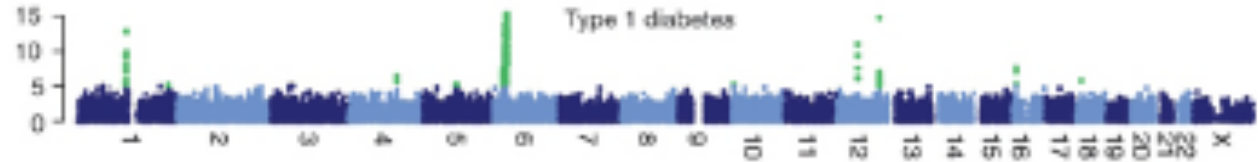
Groups of connected genes can be grouped in a supergene and disease dominance assumed: a mutation in any allele will cause the disease.



PIN based model of Interactions

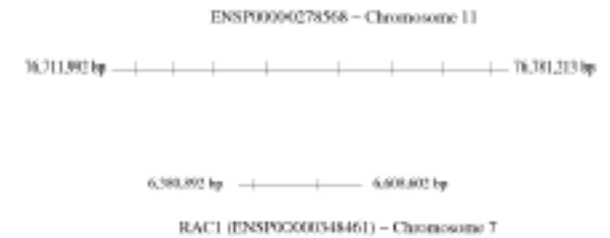
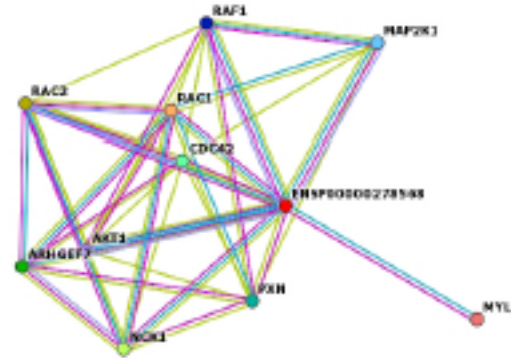
Emily et al, 2009

Single marker association



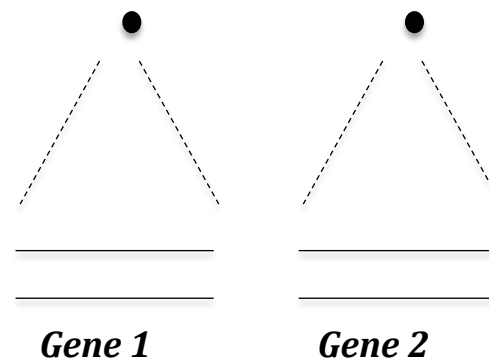
Single marker scan for T1 Diabetes in the WTCCC dataset

Protein Interaction Network



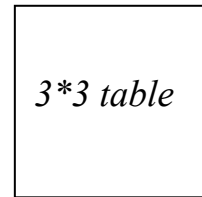
PIN gene pairs are allowed to interact

Interactions creates non-independence in combinations



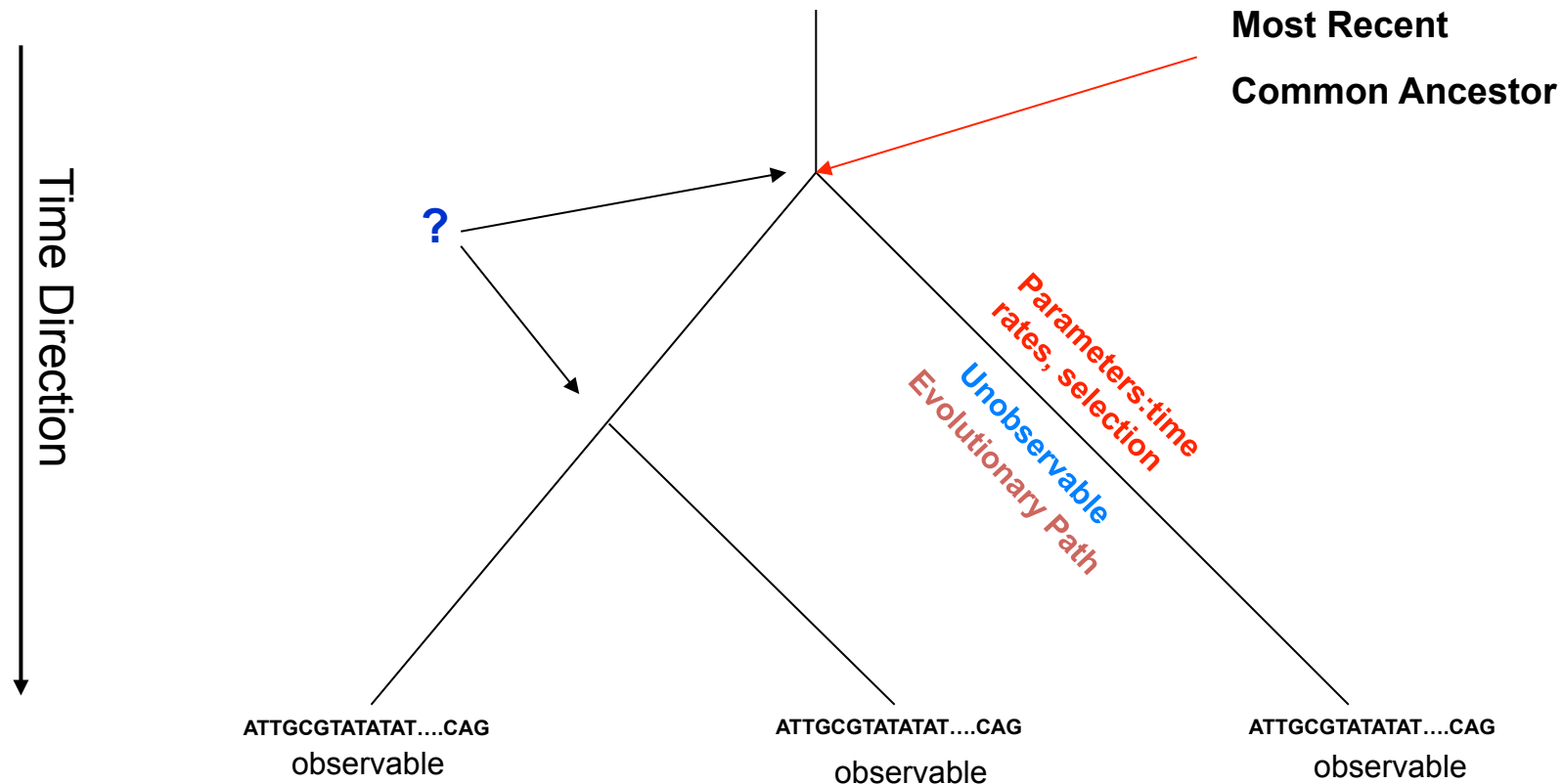
Phenotype i

SNP 1



SNP 2

Comparative Biology



Key Questions:

- Which phylogeny?
- Which ancestral states?
- Which process?

Key Generalisations:

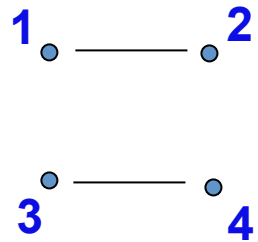
- Homologous objects
- Co-modelling
- Genealogical Structures?

Comparative Biology: Evolutionary Models

<u>Object</u>	<u>Type</u>	<u>Reference</u>
Nucleotides/Amino Acids/codons	CTFS continuous time finite states	Jukes-Cantor 69 +500 others
Continuous Quantities	CTCS continuous time countable states	Felsenstein 68 + 50 others
Sequences	CTCS	Thorne, Kishino Felsenstein,91 + 40others
Gene Structure	Matching	DeGroot, 07
Genome Structure	CTCS MM	Miklos,
Structure		
RNA	SCFG-model like	Holmes, I. 06 + few others
Protein	non-evolutionary: extreme variety	Lesk, A; Taylor, W.
Networks	CTCS	Snijder, T (sociological networks)
Metabolic Pathways	?	
Protein Interaction	CTCS	Stumpf, Wiuf, Ideker
Regulatory Pathways	CTCS	Quayle and Bullock, 06
Signal Transduction	CTCS	Soyer et al.,06
Macromolecular Assemblies	?	
Motors	?	
Shape	- (non-evolutionary models)	Dryden and Mardia, 1998
Patterns	- (non-evolutionary models)	Turing, 52;
Tissue/Organs/Skeleton/....	- (non-evolutionary models)	Grenander,
Dynamics		
MD movements of proteins	-	
Locomotion	-	
Culture	analogues to genetic models	Cavalli-Sforza & Feldman, 83
Language		
Vocabulary	“Infinite Allele Model” (CTCS)	Swadesh,52, Sankoff,72, Gray & Aitkinson, 2003
Grammar		Dunn 05
Phonetics		Bouchard-Côté 2007
Semantics		Sankoff,70
Phenotype	Brownian Motion/Diffusion	
Dynamical Systems	-	

Likelihood of Homologous Pathways

Number of Metabolisms:



+ 2 symmetrical versions

n	Number of all graphs with n nodes	Number of states
1	1	1
2	2	2
3	8	8
4	64	61
5	1024	969
6	32768	31738
7	2097152	2069964
8	268435456	267270033
9	68719476736	68629753641
10	35184372088832	35171000942698

$$P_{\Theta}(\text{graph}_1, \text{graph}_2) = P_{\Theta}(\text{graph}_1) P_{\Theta}(\text{graph}_1 \rightarrow \text{graph}_2)$$



Approaches:

Continuous Time Markov Chains with computational tricks.

MCMC



Importance Sampling

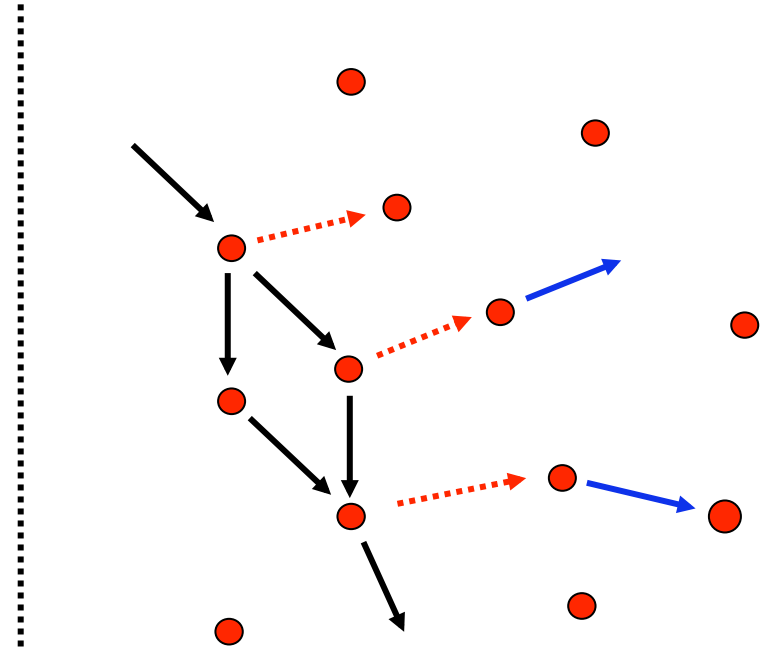
A Model for the Evolution of Metabolisms

- A given set of metabolites: 
- A given set of possible reactions -
arrows not shown.
- A core metabolism: 
- A set of present reactions - **M**
black and **red** arrows

Restriction R:

A metabolism must define a connected graph
M + **R** defines

1. a set of deletable (dashed) edges **D(M)**: 
2. and a set of addable edges **A(M)**: 



Let μ be the rate of deletion
 λ the rate of insertion

Then

$$\frac{dP(M)}{dt} = \lambda \sum_{M' \in D(M)} P(M') + \mu \sum_{M'' \in A(M)} P(M'') - P(M)[\lambda|D(M)| + \mu|A(M)|]$$