

Mini project-examination

- *It is expected to be 3 days worth of work.*
- *You will be given this in week 8*
- *I would expect 7-10 pages*
- *You will be given 2-4 key references*
- *A set of guiding questions that might help you in your writing*
- *You can chose between a set of topics broadly covering the taught material*

"Where a topic is assessed by a mini-project, the mini-project should be designed to take a typical student about three days. You are not permitted to withdraw from being examined on a topic once you have submitted your mini-project to the Examination Schools."

The Cell, the Central Dogma and the Multicellular Organism

The Cell – ignoring shape and compartmentalisation (10^{-5} m):

DNA – string over 4 letters/nucleotides {A,C,G,T}

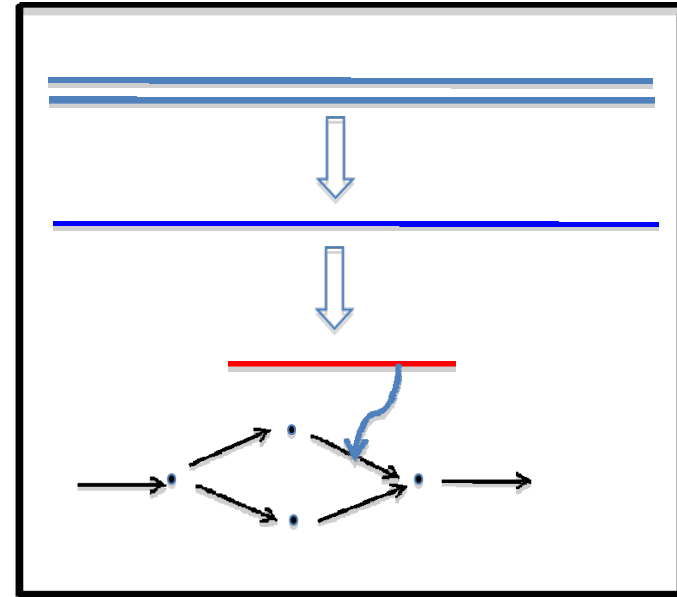
Transcribed by base pairing (A-T(U), C-G) into:

RNA – string over 4 letters/nucleotides {A,C,G,U}

Nucleotides in groups of 3 (codons) translated into amino acids:

Protein – string over 20 letters/amino acids

Proteins governs (among other things) Metabolism



Epigenetics – DNA and chromosome is modified as part of governing regulation.

Data: *highthroughput*-collected without reference to a hypothesis, *experiment* – data collected relative to hypothesis

The Cell creates the individual through ~40 duplications

Structure of Integrative Genomics

Classes

DNA

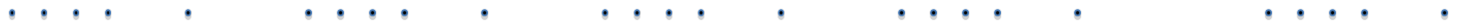
mRNA

Protein

Metabolite

Phenotype

Parts



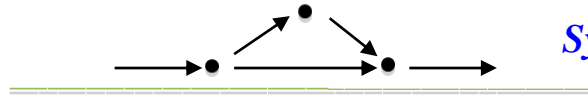
Concepts

G → F Mapping



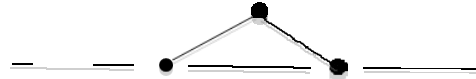
Models: Networks

Physical models:



Systems Biology

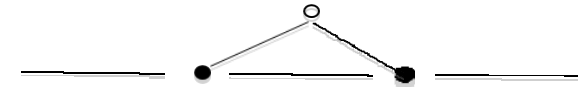
Phenomenological models:



Integrative Genomics

Hidden Structures/ Processes

○ Unobserved/unobservable



Knowledge:

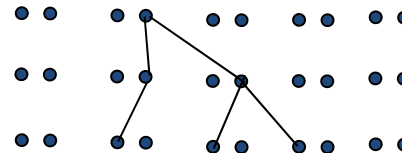
Externally Derived Constraints on which Models are acceptable

Evolution:

Cells in Ontogeny



Individuals/Sequences in a Population



Species



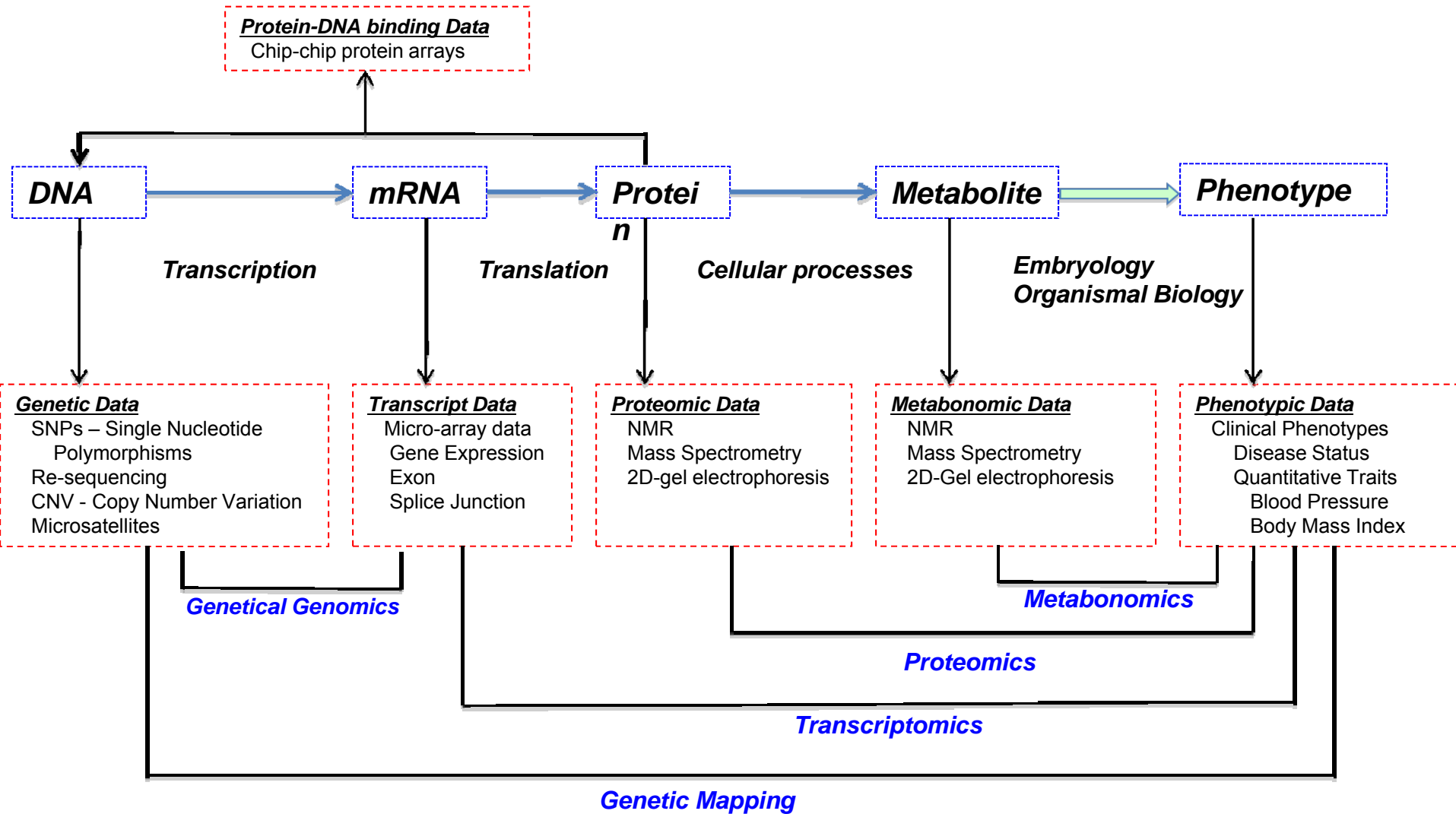
Analysis: *Data + Models + Inference*



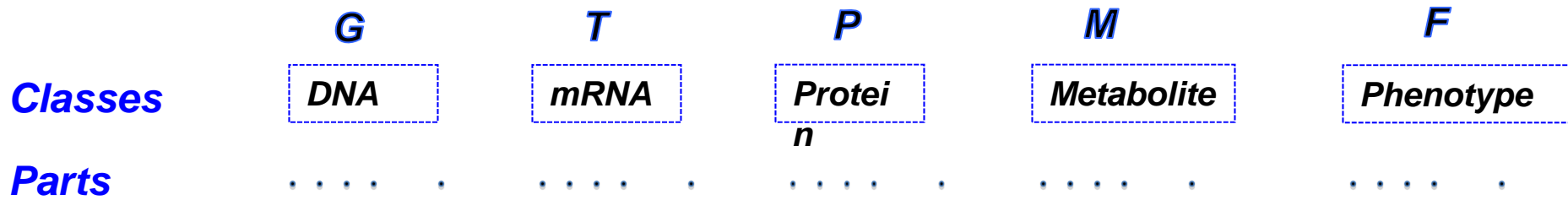
Model Selection

Functional Explanation

The Central Dogma & Data



The key questions for any data type(s)



- *What is the state space of a single of observable and its (unobservable) biological state ?*
- *What is the dimension of the observation vector at each level?*
- *What is the distribution of an individual observable*
- *Are there correlation **within** a level? Statistical? Mechanistic?*
- *Are there correlation **between** levels? Statistical? Mechanistic?*
- *Are there conditional independencies? Say T and M are conditionally independent given P ?*
- *How does a level evolve between species? How does it vary within a population?*
- *Does it vary between tissues or diseases states?*

Networks → A Cell → A Human

- *A cell has $\sim 10^{13}$ atoms.* 10^{13}
- *Describing atomic behavior needs $\sim 10^{15}$ time steps per second* 10^{28}
- *A human has $\sim 10^{13}$ cells.* 10^{41}
- *Large descriptive networks have 10^3 - 10^5 edges, nodes and labels* 10^5
- *What happened to the missing 36 orders of magnitude???*
- *Which approximations have been made?*
 - A** *Spatial homogeneity → 10^3 - 10^7 molecules can be represented by concentration* $\sim 10^4$
 - B** *One molecule (10^4), one action per second (10^{15})* $\sim 10^{19}$
 - C** *Little explicit description beyond the cell* $\sim 10^{13}$
- A** *Compartmentalisation can be added, some models (ie Turing) create spatial heterogeneity*
- B** *Hopefully valid, but hard to test*
- C** *Techniques (ie medical imaging) gather beyond cell data*

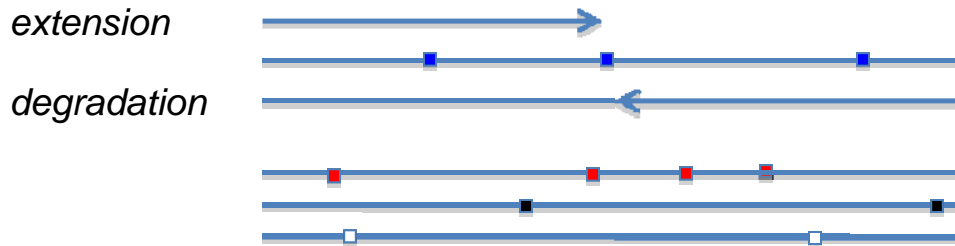
G: Genomes

A diploid genome:

Key challenge: Making a single molecule observable!!

Classical Solution (70s): Many

De Novo Sequencing: Halted extensions or degradation



80s: From one to many: PCR – Polymerase Chain Reaction

00s: Re-sequencing: Hybridisation to complete genomes

Future Solution: One is enough!!

Observing the behavior of the polymerase

Passing DNA through millipores registering changes in current

G: Assembly and Hybridisation

Target genome

3×10^9 bp

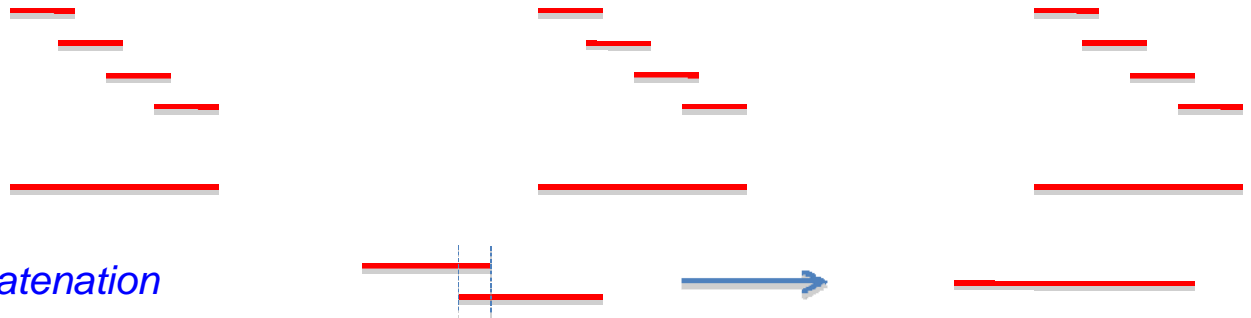
(unobservable)

Reads

3-400 bp

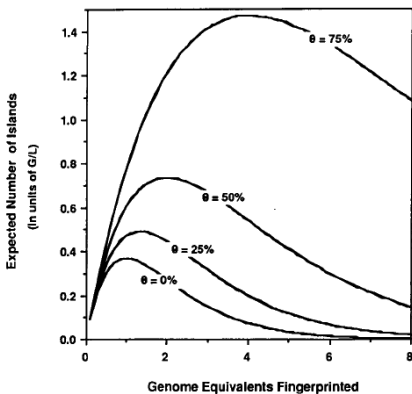
(observable)

Contigs



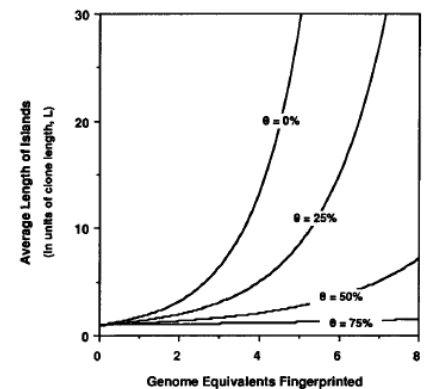
Sufficient overlap allows concatenation

Contigs and Contig Sizes as function of Genome Size (G), Read Size (L) and overlap (\emptyset):

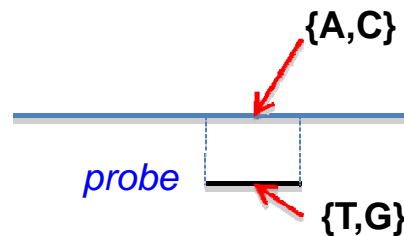


Approximate value of G/L

	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
<i>E. coli</i>	267	100	4
<i>S. cerevisiae</i>	1333	500	20
<i>C. elegans</i>	5,667	2,125	85
Human	200,000	75,000	3,000



Complementary or almost complementary strings allow interrogation.



T - Transcriptomics

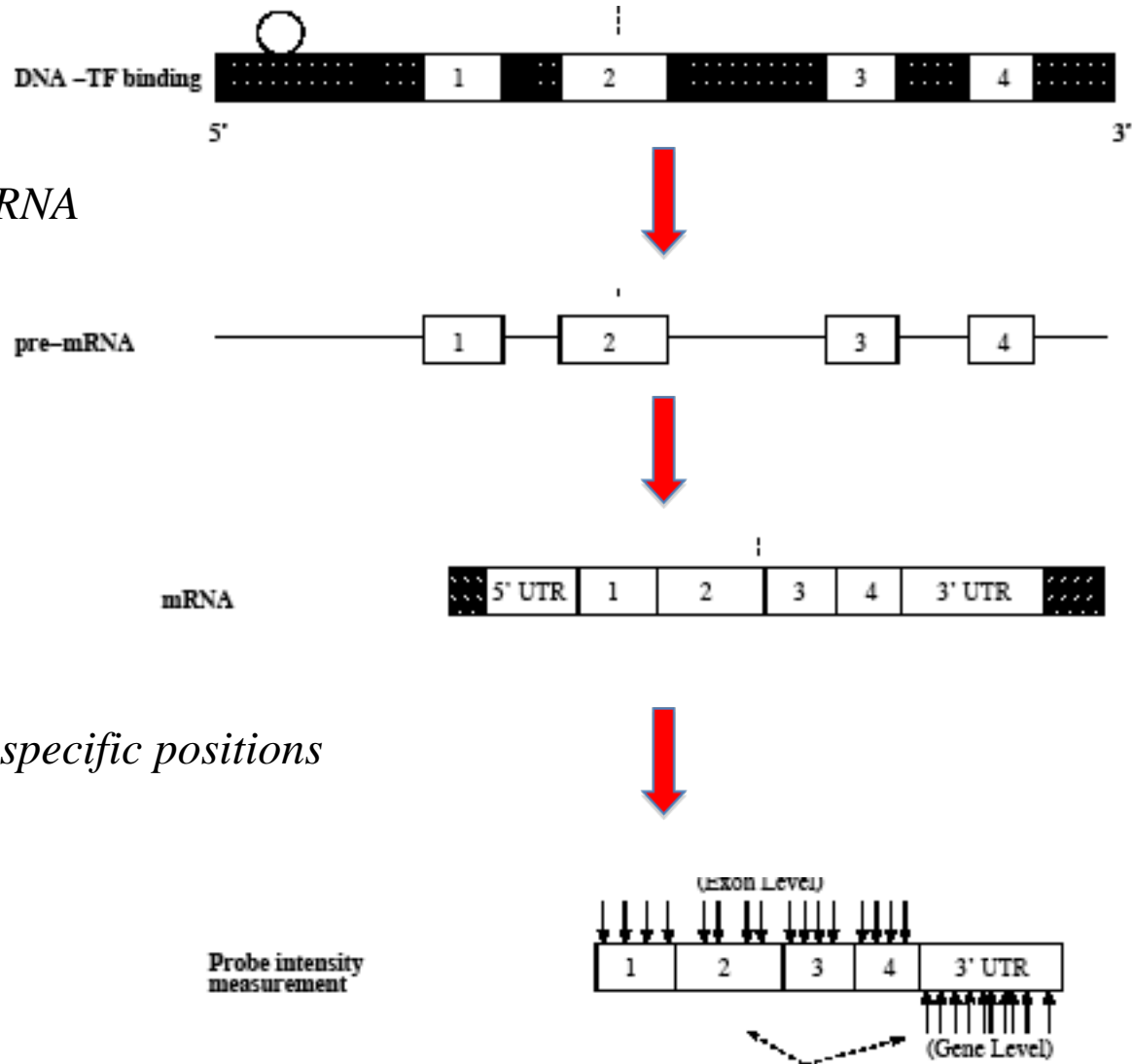
Classical Expression Experiment:

The Gene is transcribed into pre-mRNA

Pre-mRNA is processed into mRNA

Probes are designed hybridizing to specific positions

Measures transcript levels averaging of a set of cells.



T - Transcriptomics

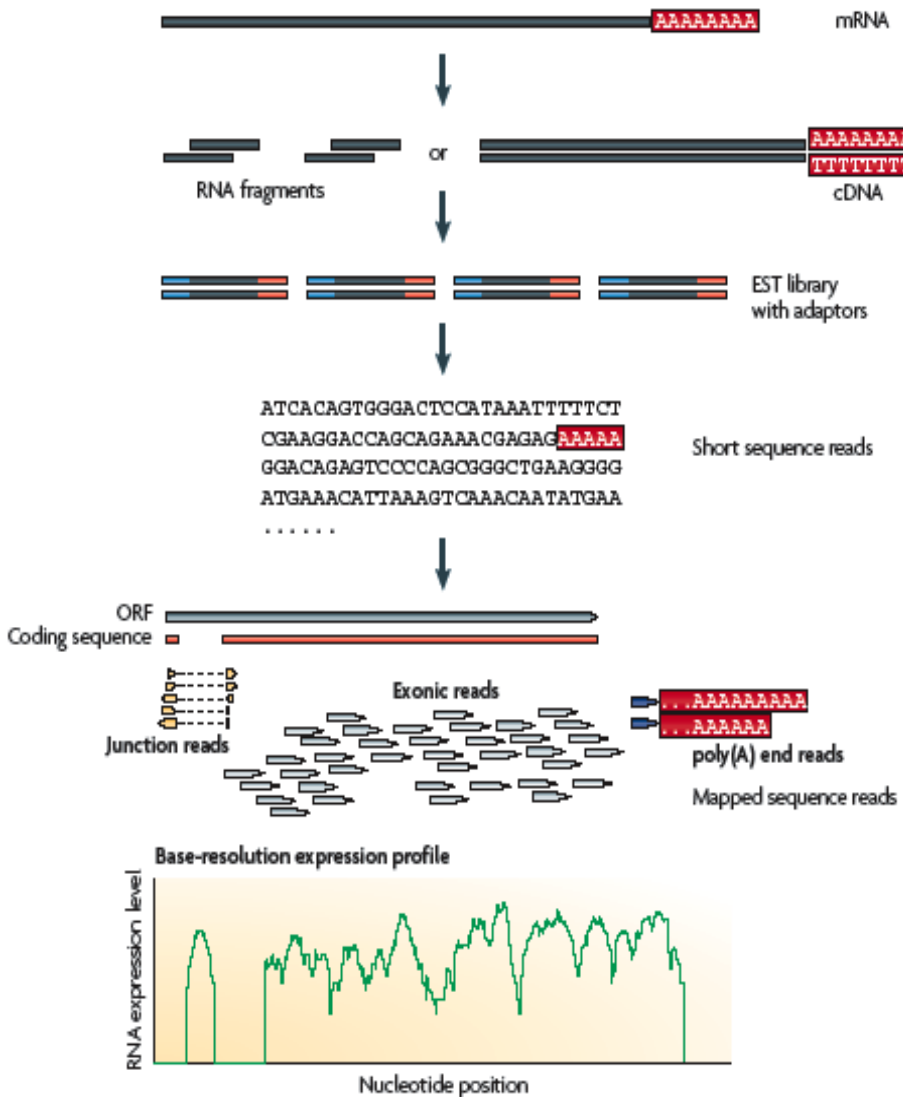
RNA-Seq Expression Experiment: Advantages - Discoveries

More quantitative in evaluating expression levels

More precise in positioning

Much more is transcribed than expected.

Transcription of genes very imprecise



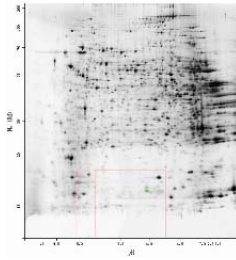
P – Proteomics

The Size of the Proteome:

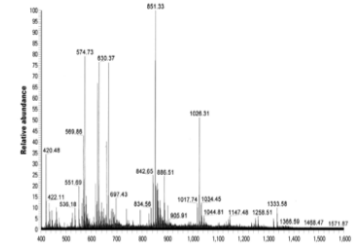
- *24.000 genes*
- *Alternative Splicing*
- *Post-translational modifications*
 - *Phosphorylation of especially serine and threonine*
 - *Glycolysation*
 - *Ubiquitination*

Experimental techniques:

- *2D electrophoresis*



- *Mass Spectroscopy*



Analysis Techniques:

Segments of proteins have known weights, modifications create known weight changes.

148.2 261.3 376.4 491.5 606.6 719.8 820.9 936.0 1051.1 1164.2 1311.4 1472.6 1571.7
Phe [Leu [Asp [Asp [Asp [Leu [Thr [Asp [Asp [Ile [Met [Cys [Val [Lys

Properties of Data:

- *Noisy*
- *Hard to make dynamic*
- *Quality improving quickly*
- *Qualitative*
- *Average over an ensemble of cells*

M – Metabonomics

The Size of the Metabolome:

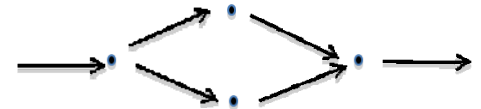
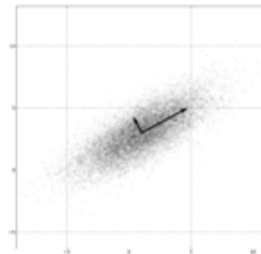
- *Set of small molecules*
 - *Combinatorial techniques allow exhaustive listing – extremely large numbers*
 - *Databases exists (eg Beilstein) with all empirically known – millions.*
 - *Standard textbook – maximally thousands. Observed tens of thousands*

Experimental techniques:

- *Gas chromatography*
- *Mass Spectroscopy*
- *Nuclear Magnetic Resonance (NMR)*

Analysis Techniques:

- *Principal Component Analysis*
- *Partial Least Squares, SIMCA*
- *Metabolic Network Analysis*



Properties of Data:

- *Noisy*
- *Hard to make dynamic*
- *Quality improving quickly*
- *Qualitative*
- *Average over an ensemble of cells*

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

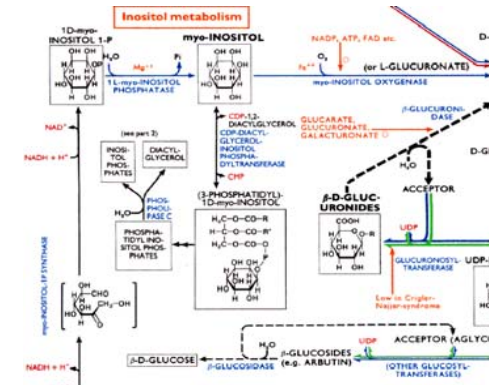
Cascade of Protein reactions that sends signal from receptor on cell surface to regulation of genes.

D. Protein Interaction Networks

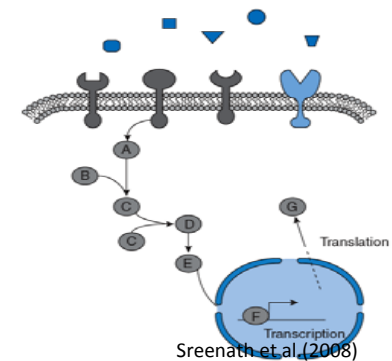
Some proteins stick together and appear together in complexes

E. Alternative Splicing Graph (ASG)

Determines which transcripts will be generated from a genes



Boehringer-Mannheim



A repertoire of Dynamic Network Models

To get to networks:

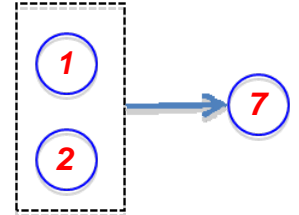
No space heterogeneity → molecules are represented by numbers/concentrations

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*

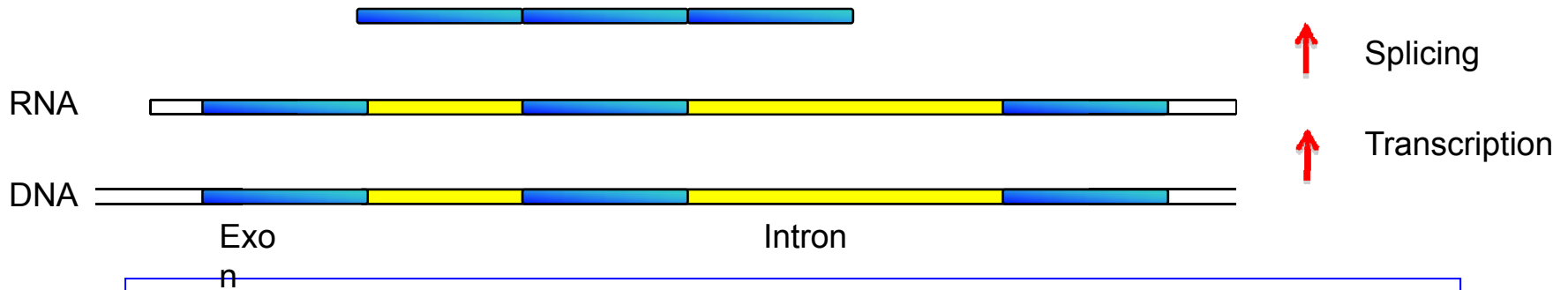
$$a_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} 2^{k(n-k)} a_{n-k}$$

- *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

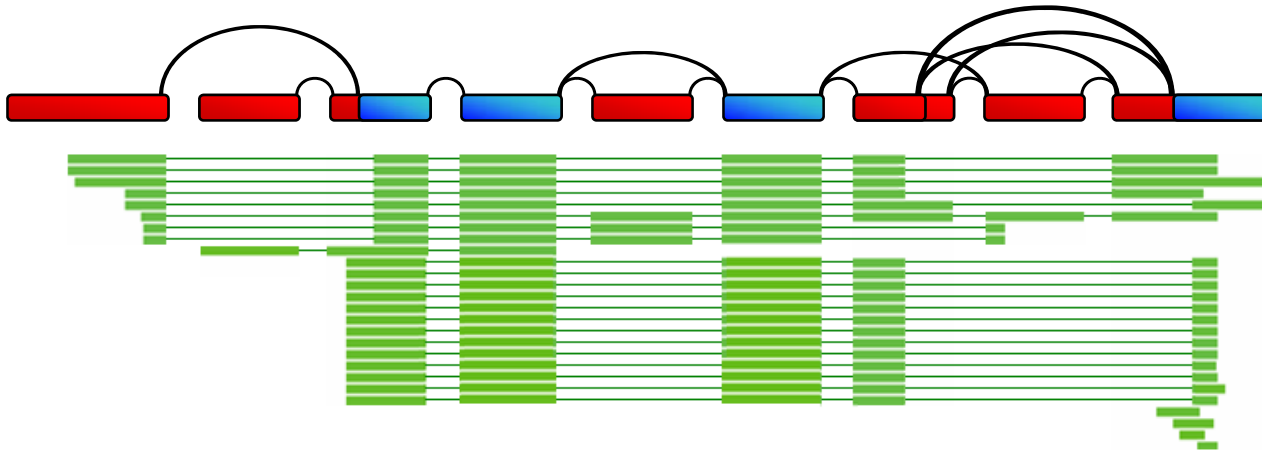
- AS: one genomic segment can create different transcripts by skipping exons (sequence intervals)



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



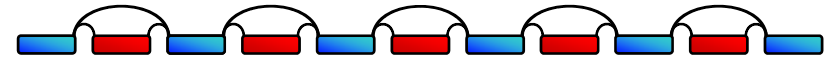
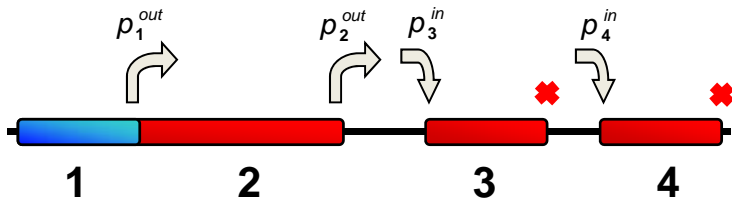
$G \rightarrow T$: Alternative Splicing

Problem: Inferring the ASG from transcripts

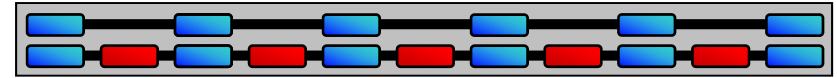
- *Maximally informative transcripts*
- *Minimally informative transcripts*
- *Random transcripts*

A Hierarchy of Models can be envisaged

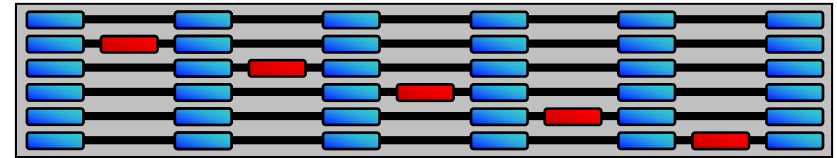
Simpler still: model 'donation' and 'acceptance' separately
Jump 'in' or 'out' of transcript with well-defined probabilities
Isolated exons are included independently, based only on the strength of its acceptor site



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



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

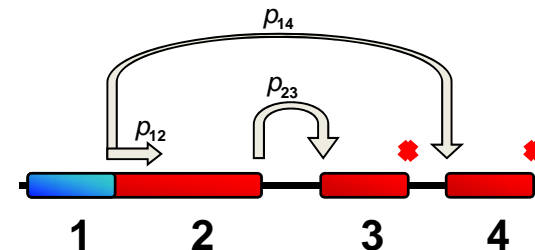


Enrich the ASG to a Markov chain

Pairwise probabilities

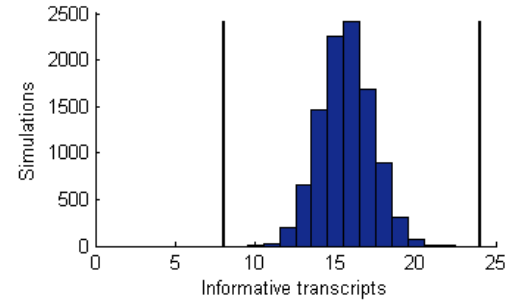
Transcripts generated by a 'walk' along the ASG

A natural model for dependencies between donors and acceptors

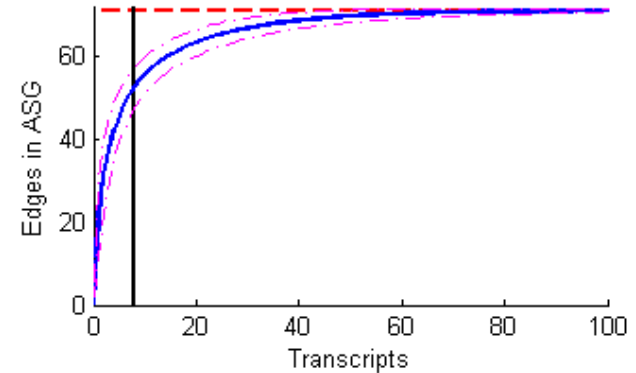


G → T: Alternative Splicing

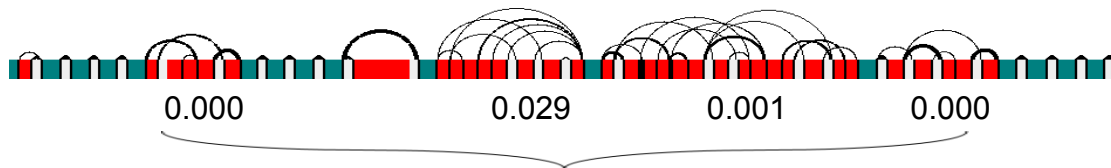
- *The distribution of necessary distinct transcripts*



- *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

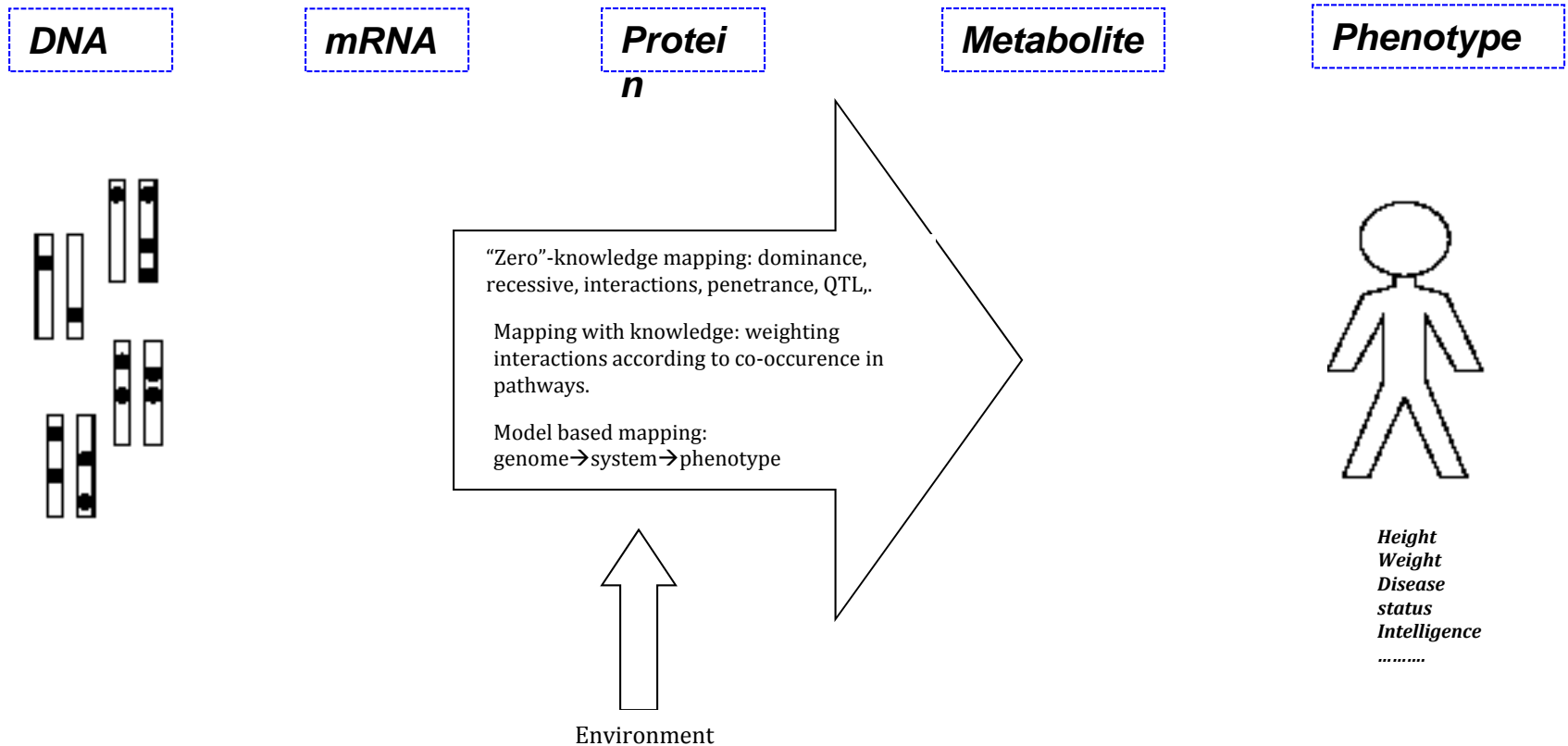
Hence, given sufficient observations, likelihood ratio tests can determine the most appropriate model for transcript generation

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*

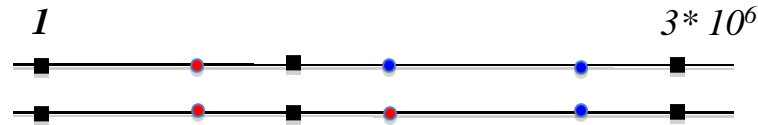
G is the most abundant data form - heritable and precise. F is of greatest interest.



The General Problem is Enormous

Set of Genotypes:

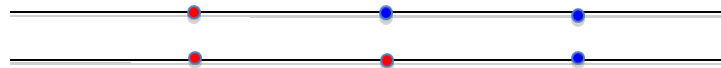
- Diploid Genome**



- 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**

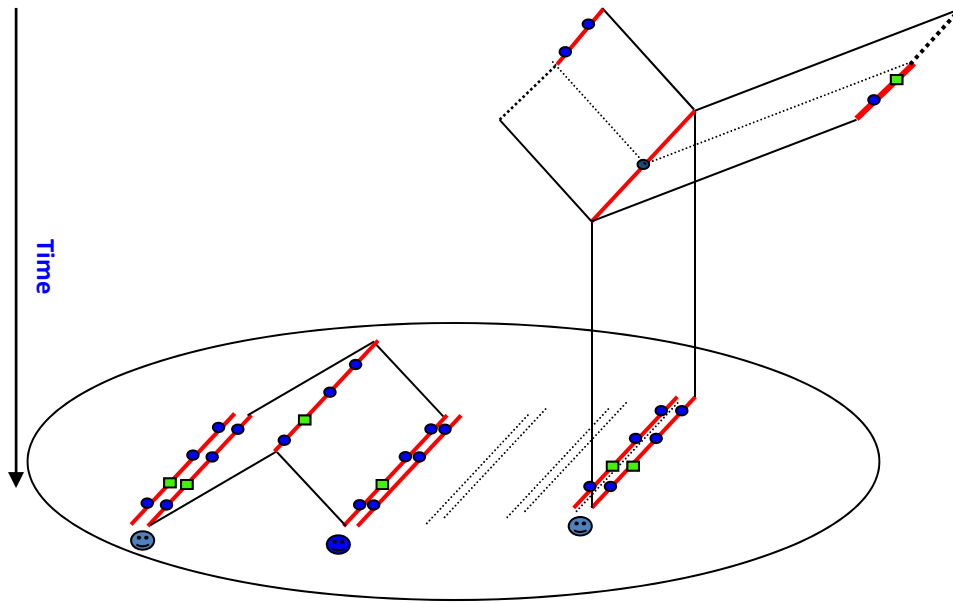
Epistasis: The effect of one locus depends on the state of another

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

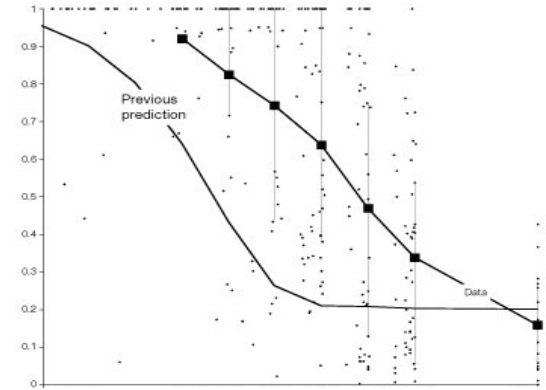
$$\sum_{i \text{ causative positions}} X_i(G_i) + \varepsilon$$

Genotype and Phenotype Co-variation: Gene Mapping

Sampling Genotypes and Phenotypes

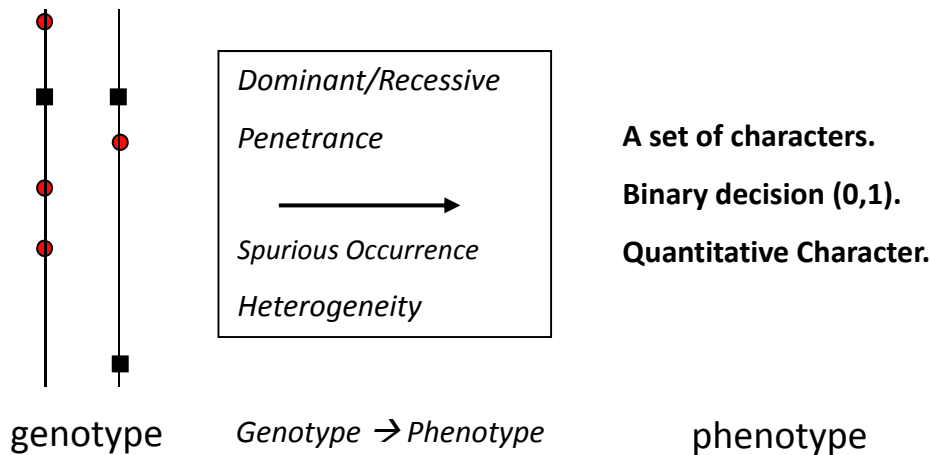


Decay of local dependency

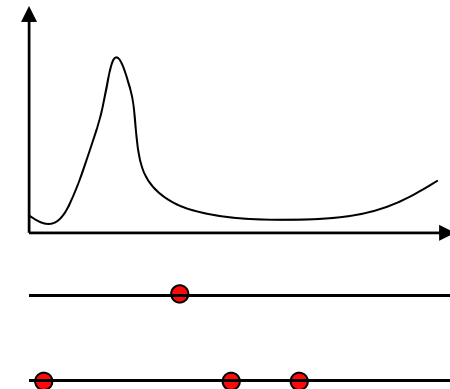


Reich *et al.* (2001)

Genotype --> Phenotype Function

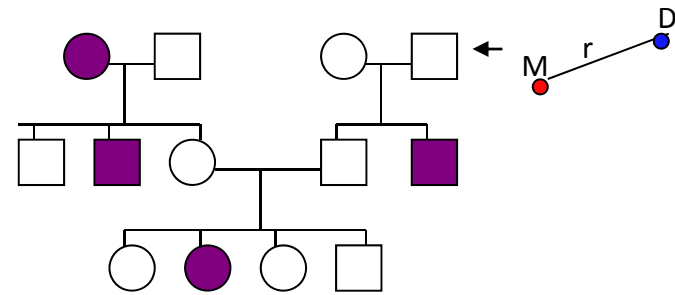


Result: The Mapping Function



Pedigree Analysis & Association Mapping

Pedigree Analysis:

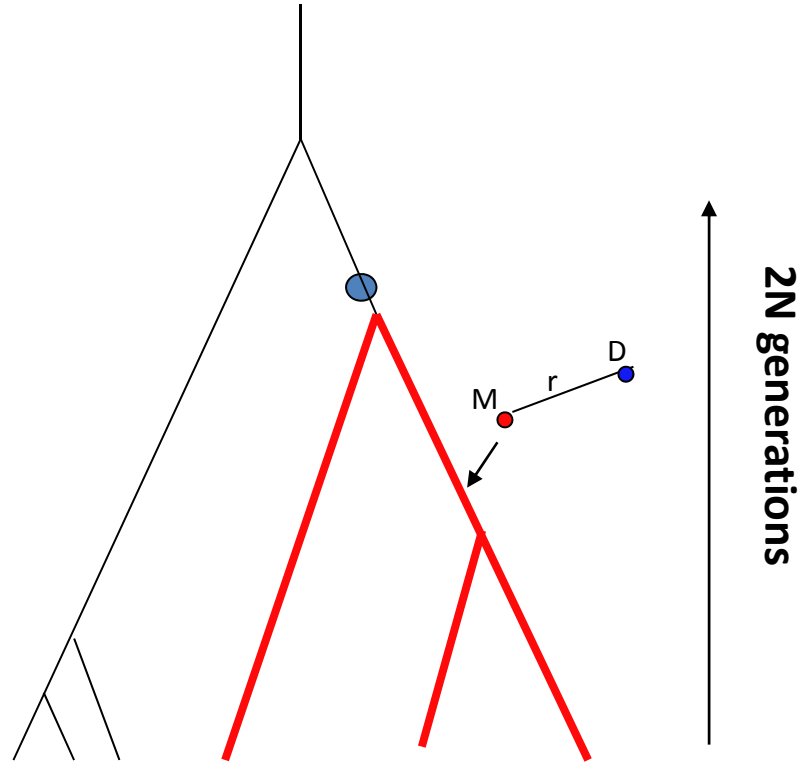


Pedigree known

Few meiosis (max 100s)

Resolution: cMorgans (Mbases)

Association Mapping:



Pedigree unknown

Many meiosis ($>10^4$)

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

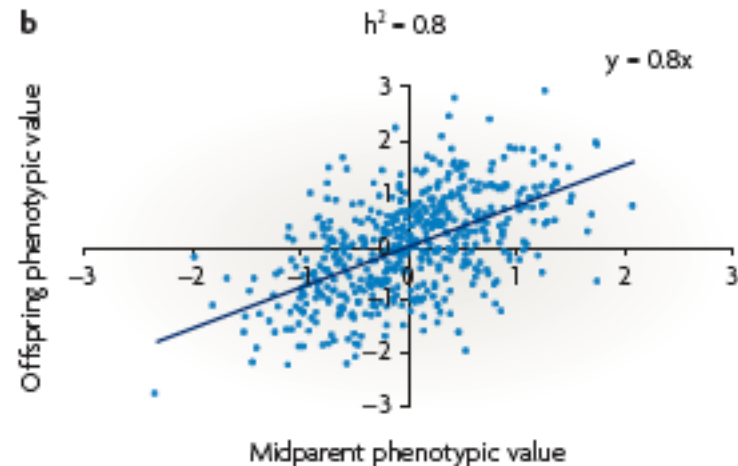
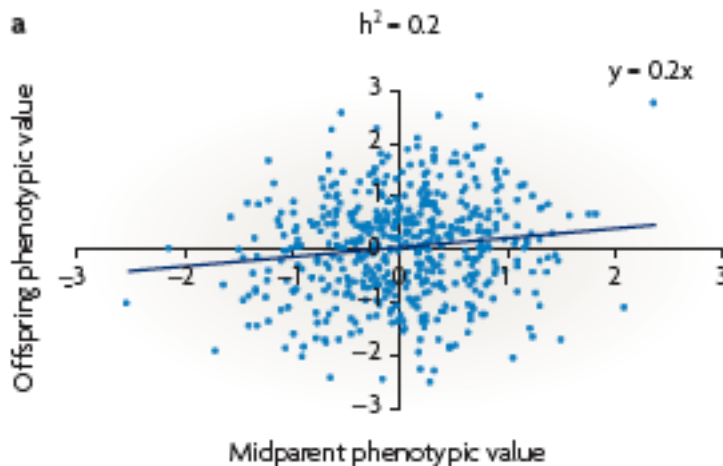
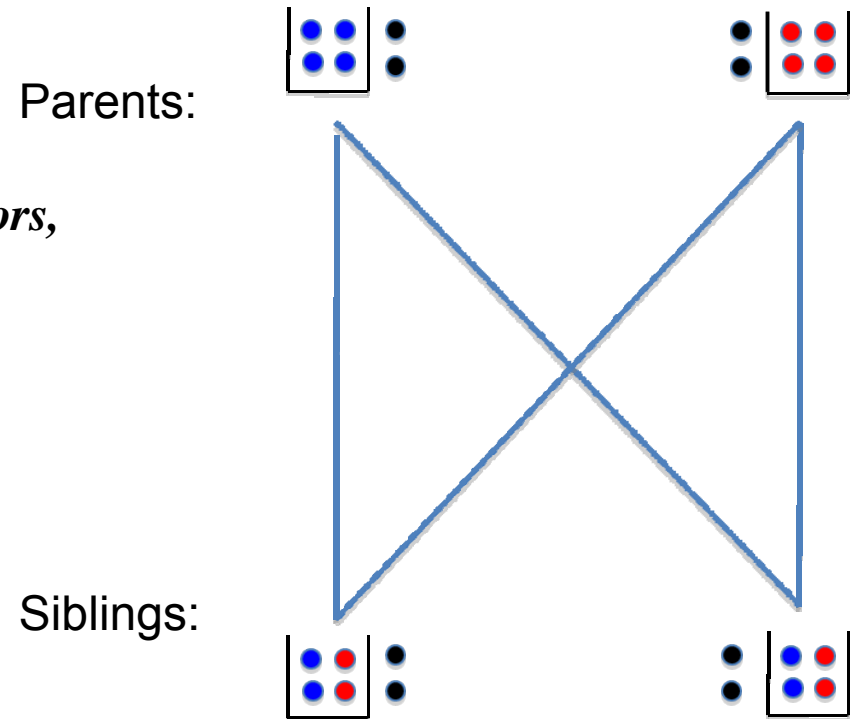
Heritability: Inheritance in bags, not strings.

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.

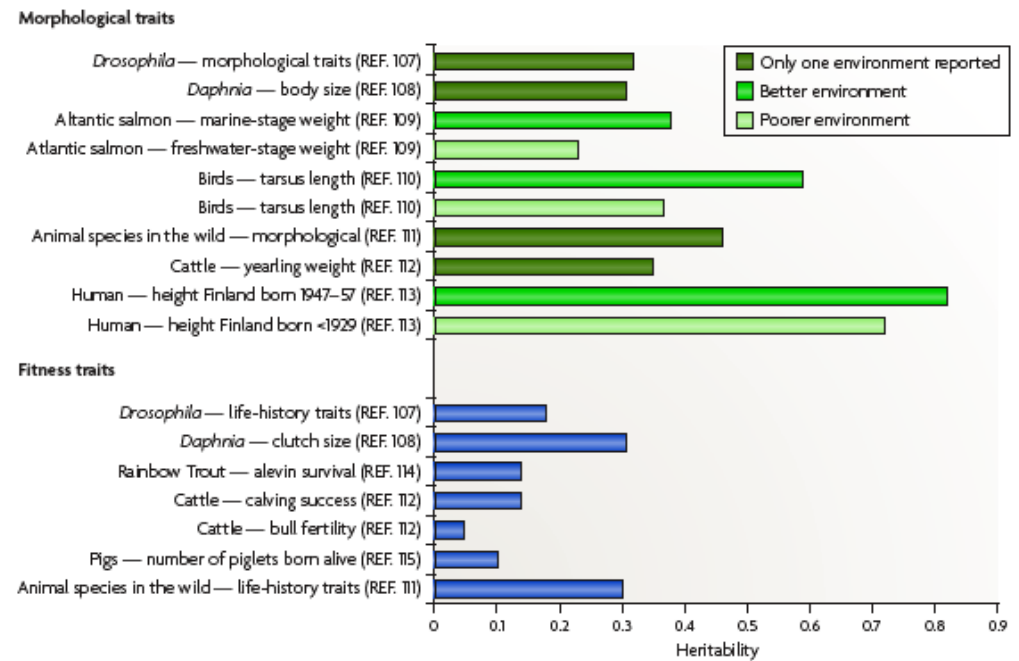
This allows calculation of relative effect of genetics and environment

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

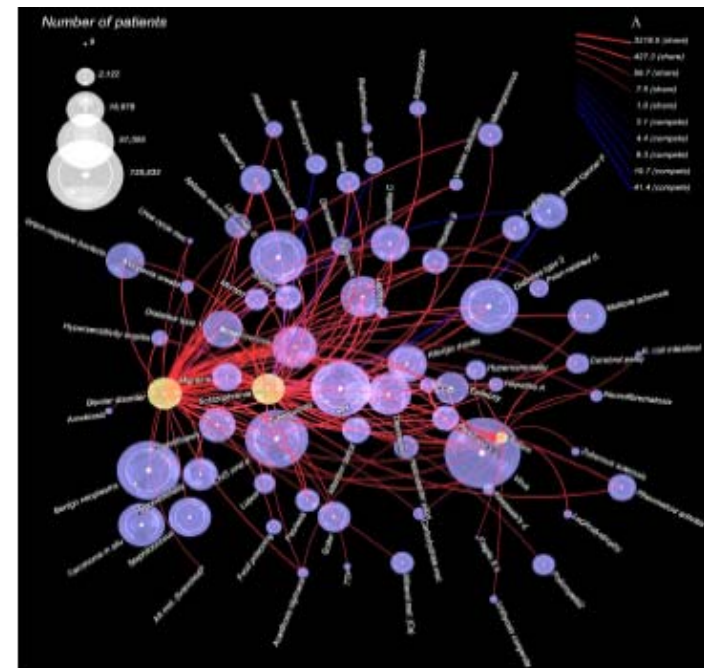
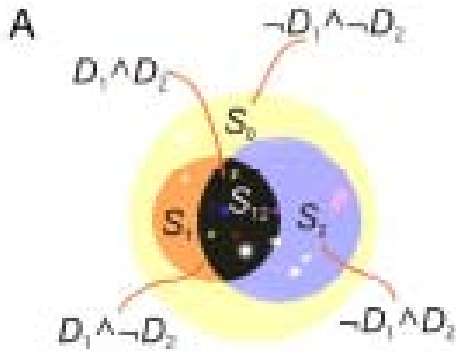


Heritability

Examples of heritability

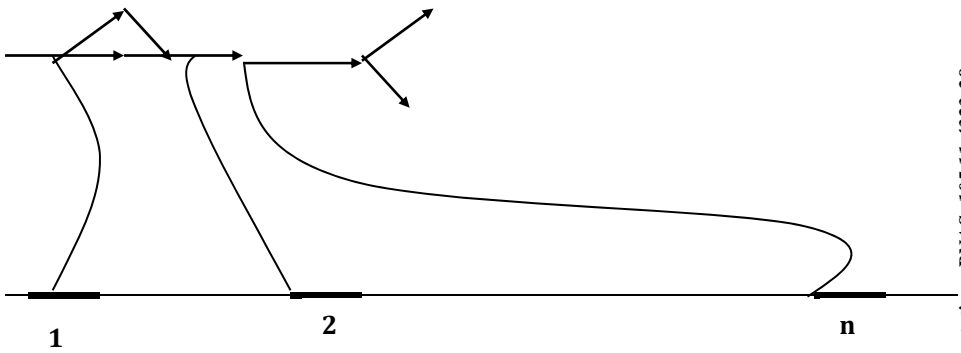
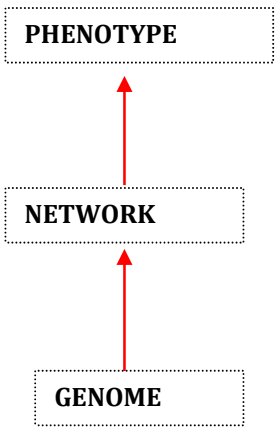


Heritability of multiple characters:

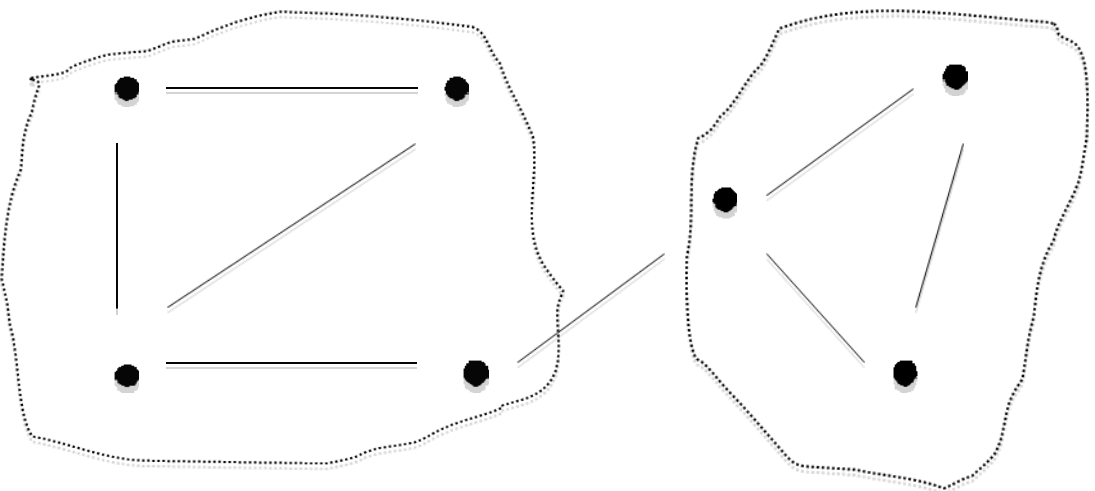


Protein Interaction Network based model of Interactions

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

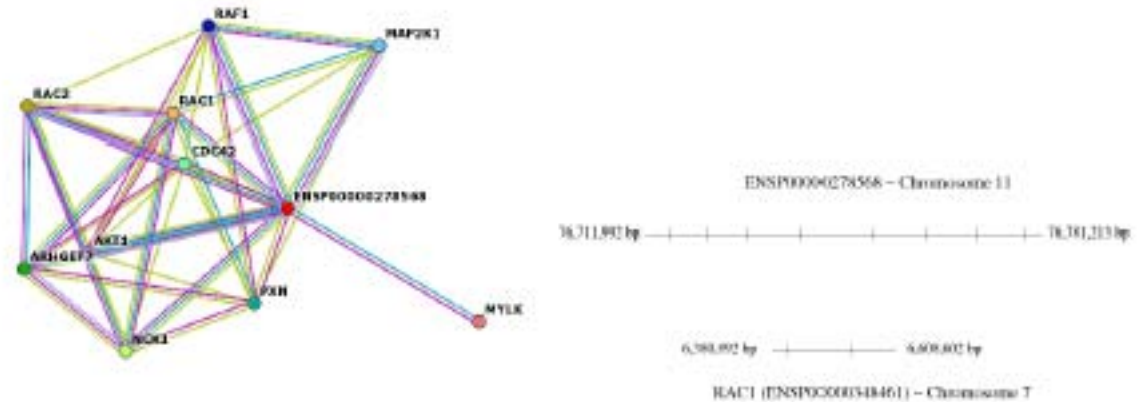
Emily et al, 2009

Single marker association

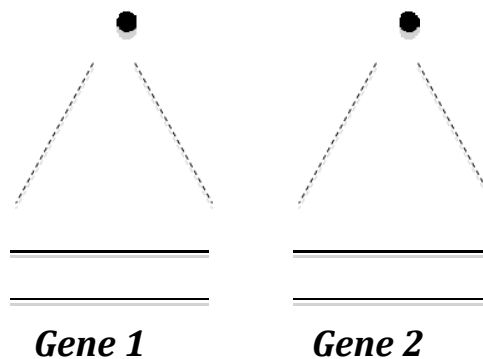


Single marker scan for T1 Diabetes in the WTCCC dataset

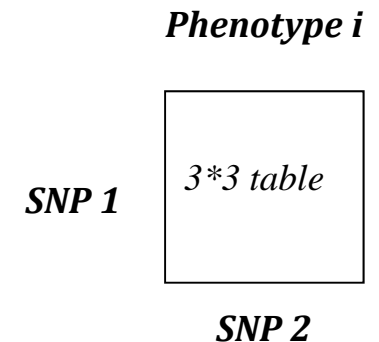
Protein Interaction Network



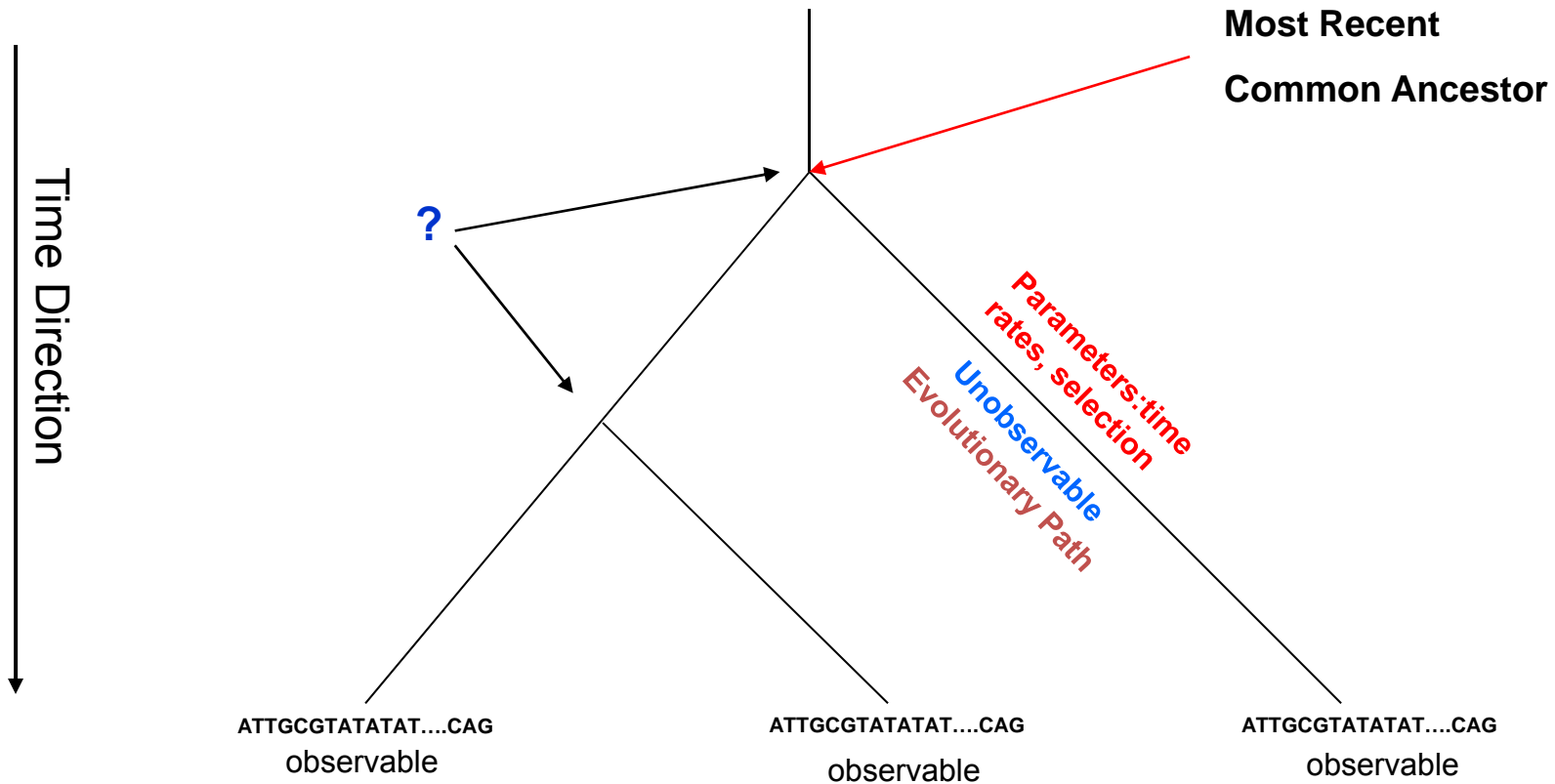
PIN gene pairs are allowed to interact



Interactions creates non-independence in combinations



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	-	