

Computational Biology of Networks

JH – Jotun Hein hein@stats.ox.ac.uk, RL – Rune Lyngsø lyngsøe@, BT – Bhalchandran Thatte thatte@

12.10	JH	Integrative Biology, Networks and the Structure of Biology I	
13.10	JH	Integrative Biology, Networks and the Structure of Biology II	
19.10	RL	Network Algorithms – Paths	
20.10	RL	Network Algorithms – Connectivity	
26.10	RL	Network Algorithms – Flow	
27.10	RL	Combinatorial Methods of Network Comparison	
02.11	BT	Probability Theory of Networks	<i>Biology</i>
03.11	BT	Reconstruction Problems in Networks	<i>Algorithmics</i>
09.11	BT	Enumeration of Networks	<i>Combinatorics</i>
10.11	BT	Sampling Networks	<i>Probability Theory</i>
16.11	RL	Network Inference	<i>Modelling</i>
17.11	RL	Network Robustness	<i>Statistics</i>
23.11	JH	Evolution of Networks I	
24.11	JH	Evolution of Networks II	
30.11	JH	Biological Networks and their Temporal Dynamics: Deterministic Models	
01.12	JH	Biological Networks and their Temporal Dynamics: Stochastic Models	

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}

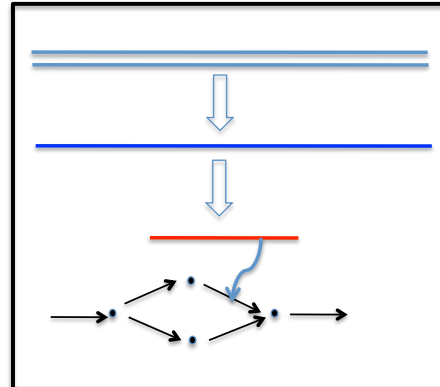
Transcribed by base pairing (A-T(U), C-G) into:

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

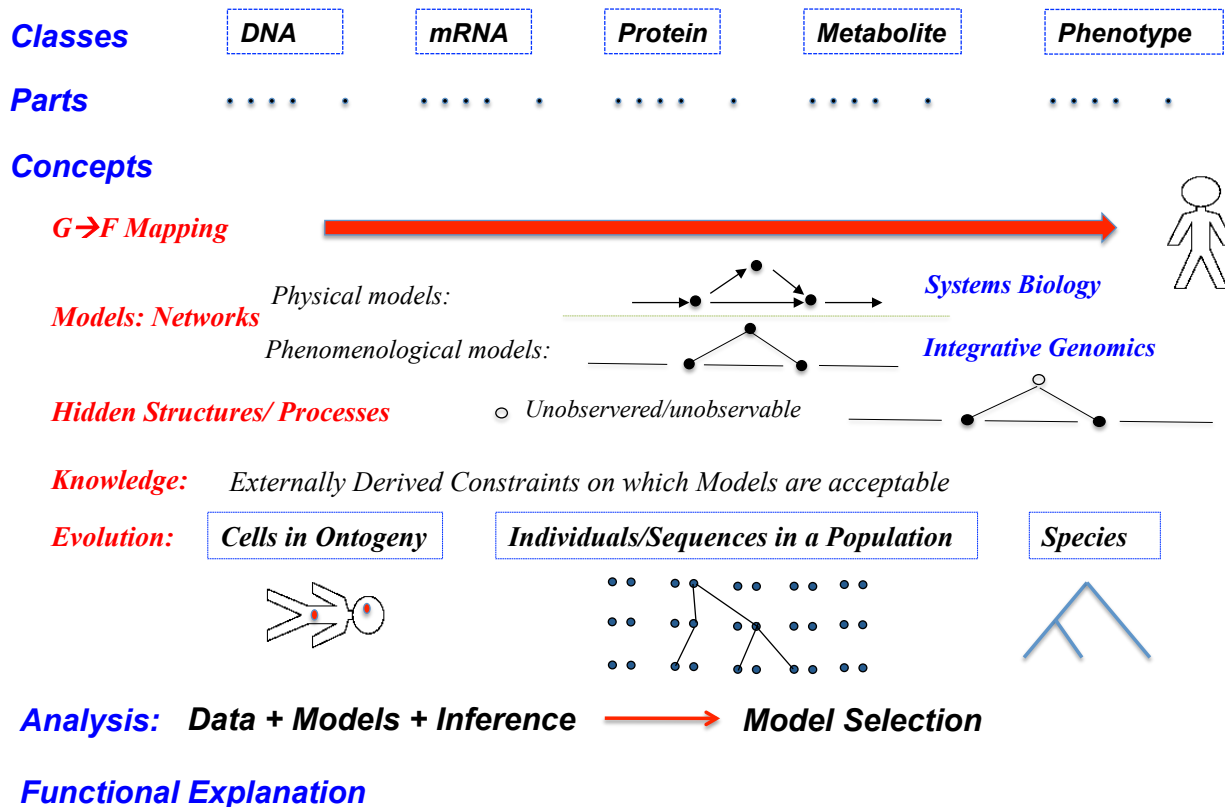


Epigenetics – DNA and chromosome is modified as part of governing regulation.

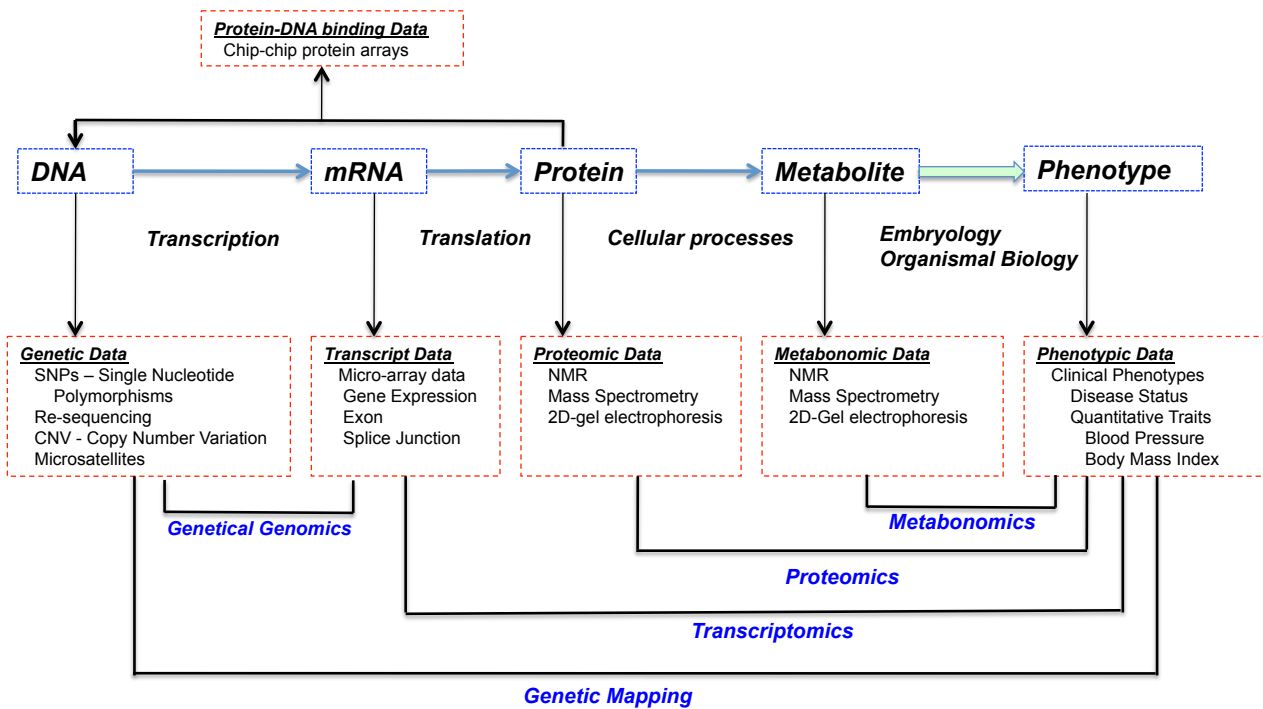
Data: *highthroughput*-collected without reference to a hypothesis, *experiment* – data collected relative to hypothesis

The Cell creates the individual through ~40 duplications

Structure of Integrative Genomics



The Central Dogma & Data



The key questions for any data type(s)

	G	T	P	M	F
Classes	DNA	mRNA	Protein	Metabolite	Phenotype
Parts

- 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

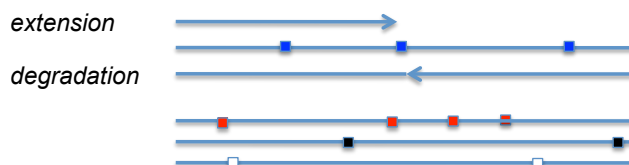
A diploid genome:



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)

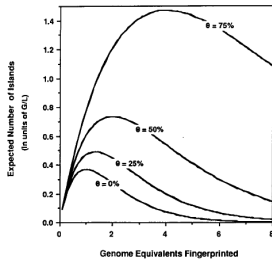
Reads

3-400 bp
(observable)

Contigs

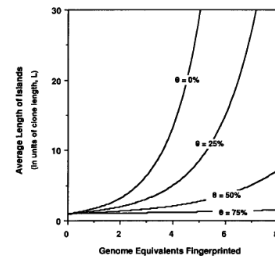
Sufficient overlap allows concatenation

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



Approximate value of G/L

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



Complementary or almost complementary strings allow interrogation.



Lander & Waterman, (1988) Statistical Analysis of Random Clone Fingerprinting

T - Transcriptomics

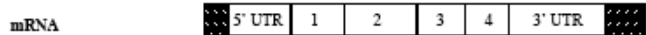
Classical Expression Experiment:



The Gene is transcribed into pre-mRNA

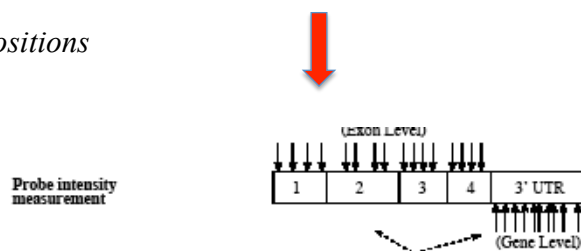


Pre-mRNA is processed into mRNA



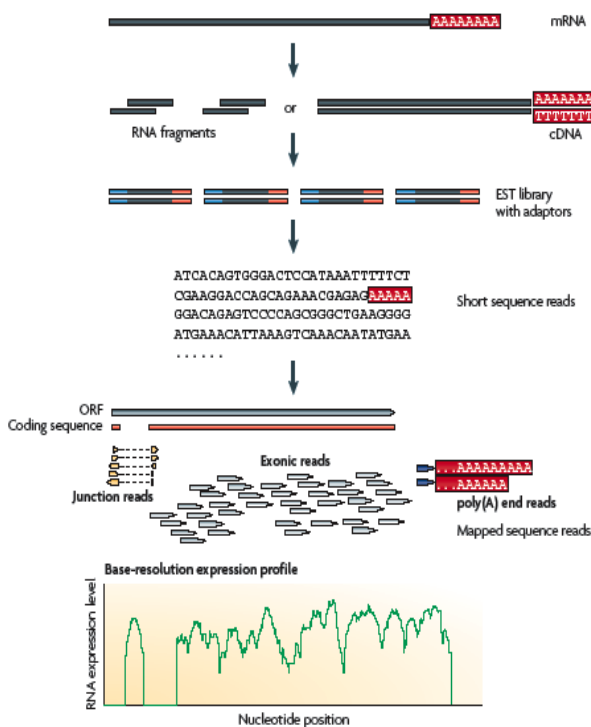
Probes are designed hybridizing to specific positions

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*

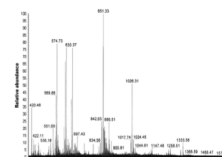
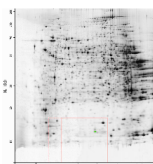
P – Proteomics

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.

148.2 261.3 326.4 491.5 606.6 719.8 820.9 936.0 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:

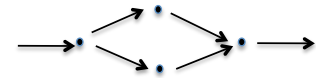
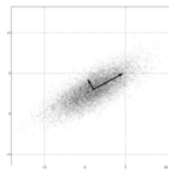
- *Set of small molecules*
 - *Combinatorial techniques allow exhaustive listing – extremely large numbers*
 - *Databases exists (eg Beilstein) with all empirically known – millions.*
 - *Standard textbook – maximally thousands. Observed tens of thousands*

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*