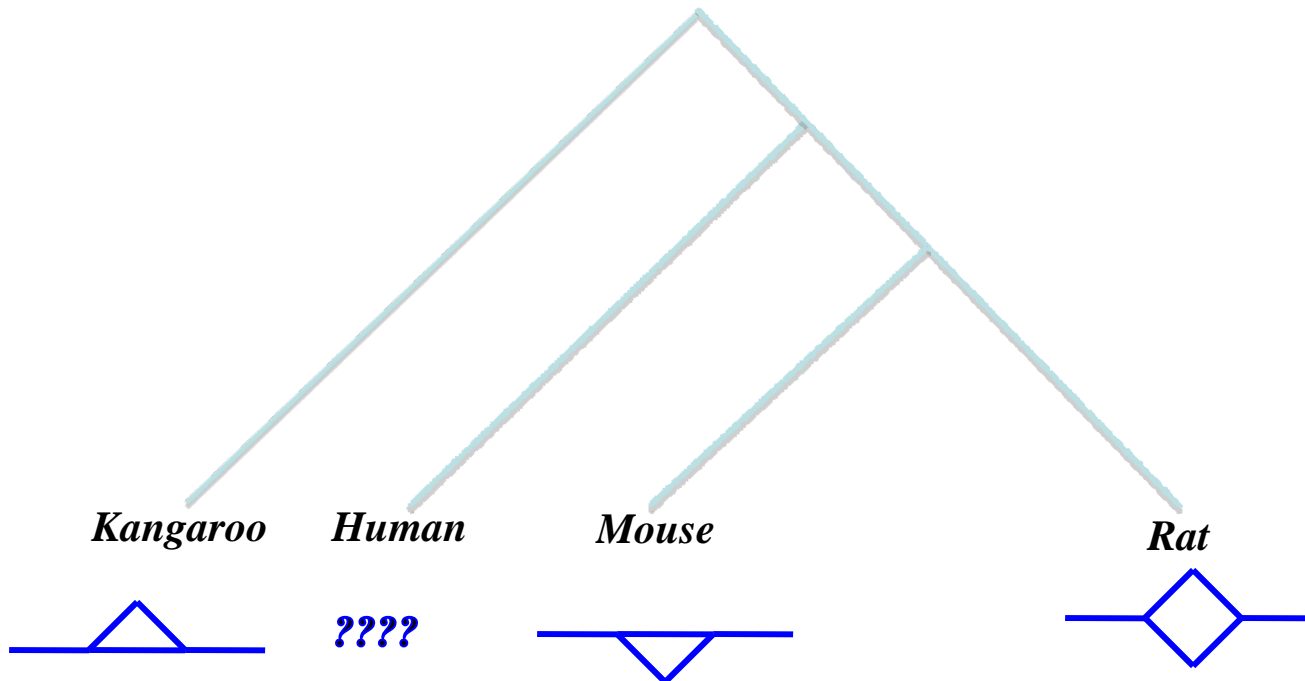
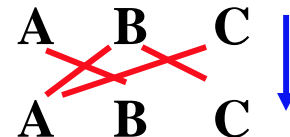
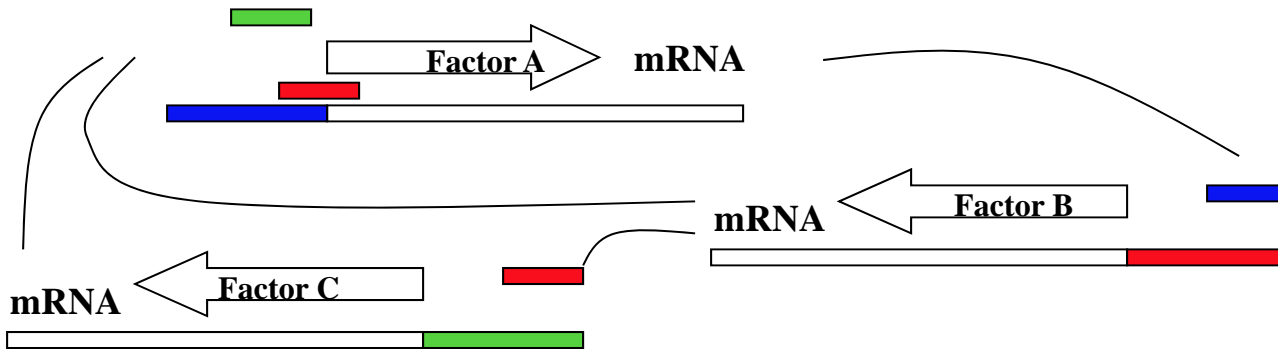
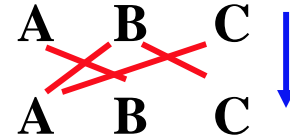
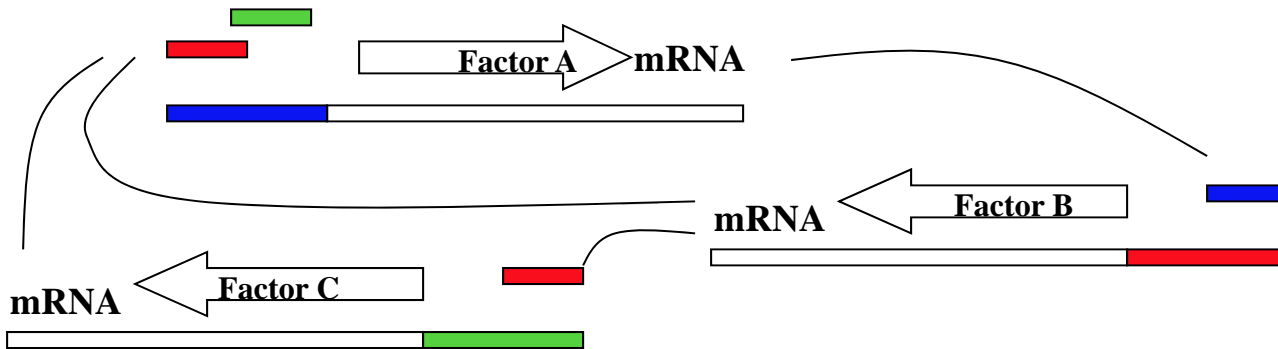
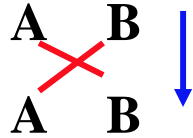
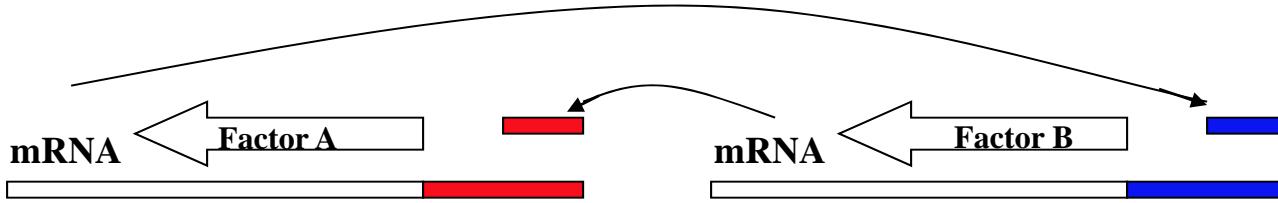


Network Evolution

- *Understanding Evolution*
- *Comparative Annotation*
- *Knowledge and Model Organisms*

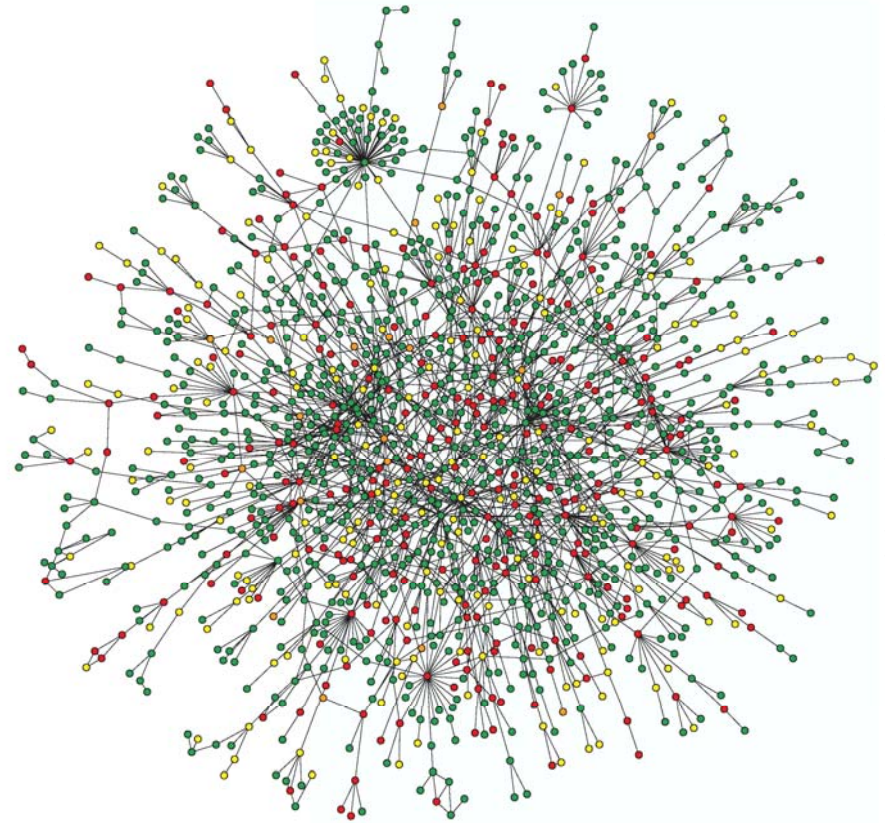


Gene Regulatory Network



Protein Interaction Network

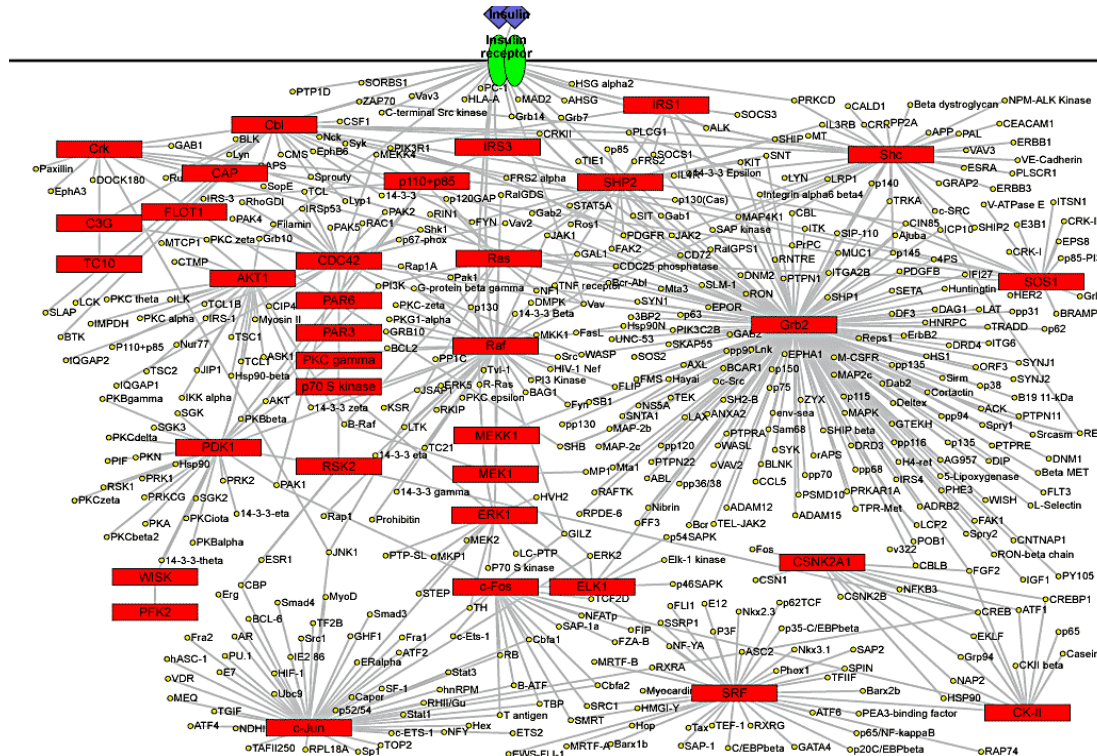
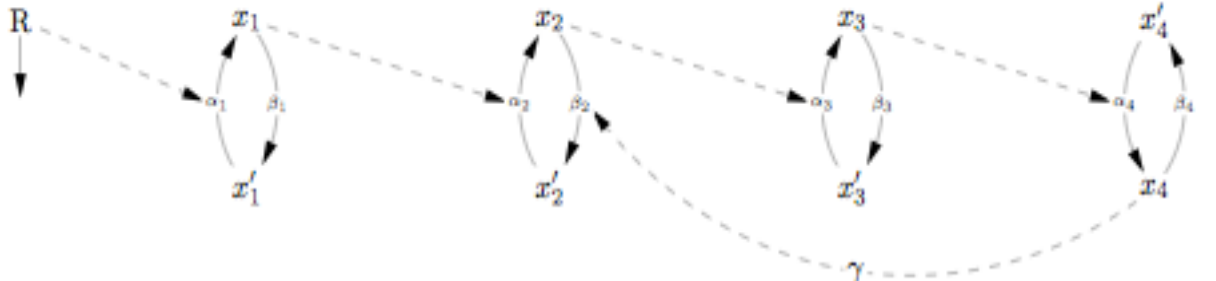
- *The sticking together of different protein is measured by mass spectroscopy.*
- *The nodes will be all known proteins.*
- *Two nodes are connected if they stick together. This can be indicator of being part of a functional protein complex, but can also occur for other reasons.*



Signaling Pathways

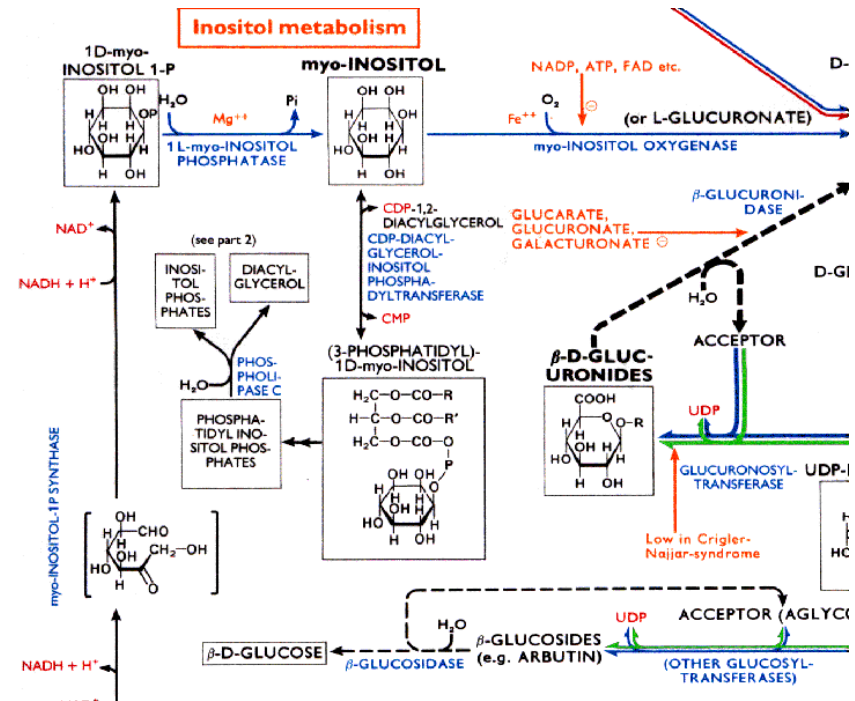
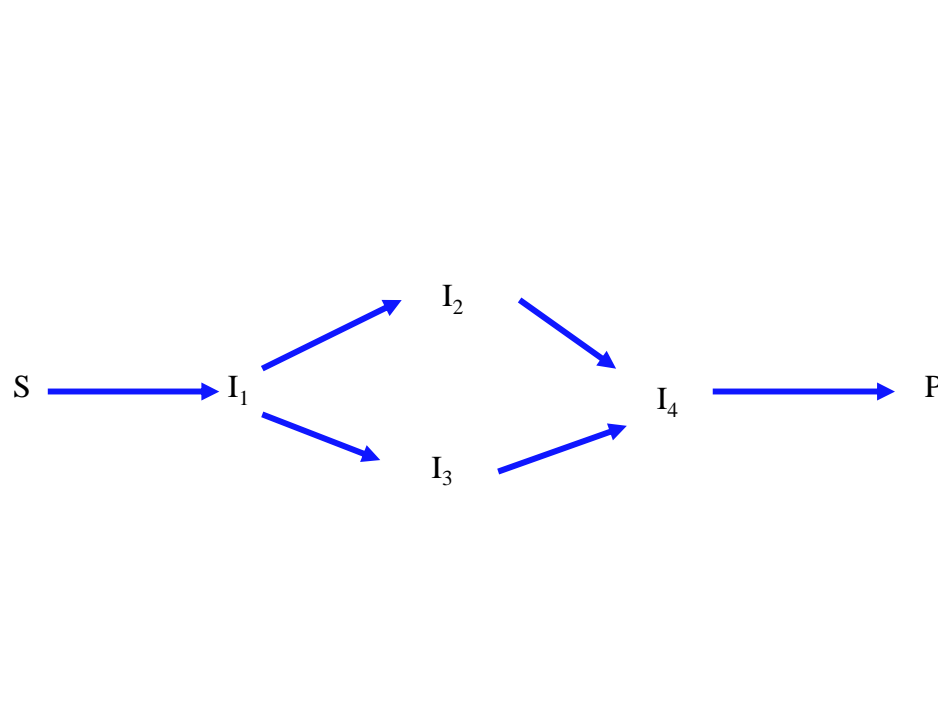
- Transmits signals from membrane to gene regulation.

- Its function is enigmatic as some of the molecules involved are common to different functions and how cross-interaction is avoided is unknown.



Metabolic Pathways

- Flux Analysis
- Metabolic Control Theory
- Biochemical Systems Theory
- Kinetic Modeling



RAFs – Reflexive Autocatalytic Foodsets

Kauffman, 1986; Steel, 2001, Hordijk and Steel, 2004; Mossel and Steel, 2005

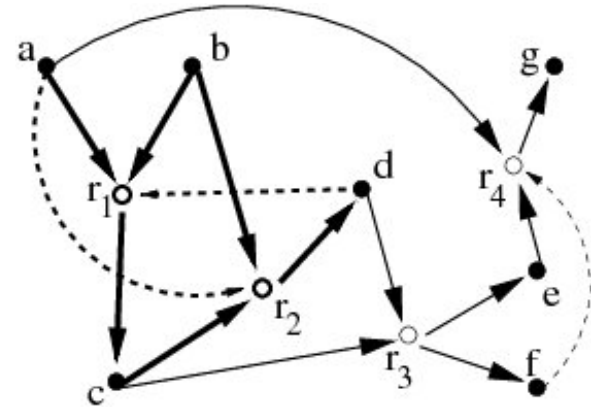
- *a set of molecule types, X ;*
- *a set of reactions where each reaction converts one set of molecules (reactants) into another set (products), R ;*
- *a set of catalysations: molecules that accelerate a reaction (or set of reactions), C ;*
 - *a food set: a small set of molecules assumed to be freely available and constantly replenished, F .*

$X = \{a, b, c, d, e, f, g\}$

$R = \{r_1, r_2, r_3, r_4\}$

$C = \{(d, r_1), (a, r_2), (f, r_4)\}$

$F = \{a, b\}$



RAFs – Reflexive Autocatalytic Foodsets

Kauffman, 1986; Steel, 2001, Hordijk and Steel, 2004; Mossel and Steel, 2005

Key achievements:

- *The probability of existence;*
- *Algorithms to find them*

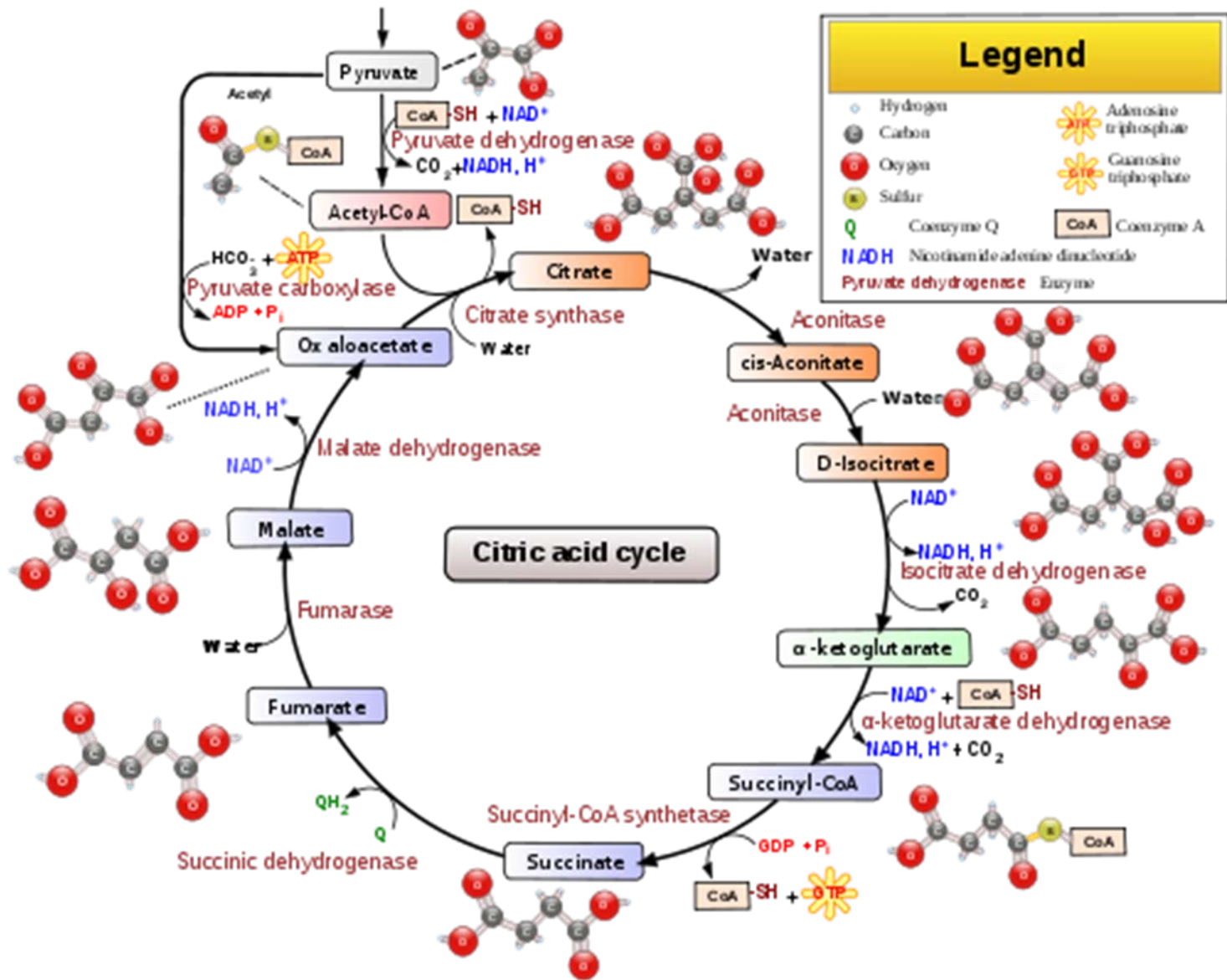
Key problems:

- *Realism;*
- *Predicting catalysis;*

Natural Extensions:

- *Let RNA be the molecules, concatenation by base-pairing*
- *Kinetic version: concentrations and rates*
- *Evolving version*
- *RAFs based on real molecules*
 - *combinatorially defined*
 - *empirically defined – observed molecules (Beilstein)*
 - *empirically defined – observed **life** molecules (Metabolism)*

Real Molecule Example of RAF: Citric Acid Cycles

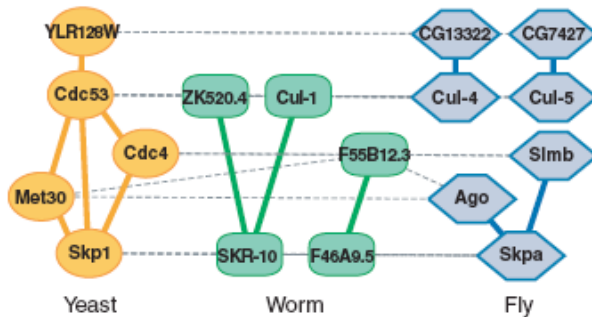


Network Alignment & Motifs

Barabasi & Oltvai, 2004, Sharan & Ideker, 2006

1. Are nodes/edges labelled?
2. Which operations are allowed?
3. Pair/Multiple?

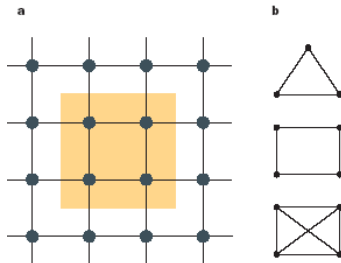
•Network Alignment



•Network Search

Find (approximately) a network within a set of others.

•Motifs



•Network integration

Combine a set of networks to one large network.

Metabolic Random Fields

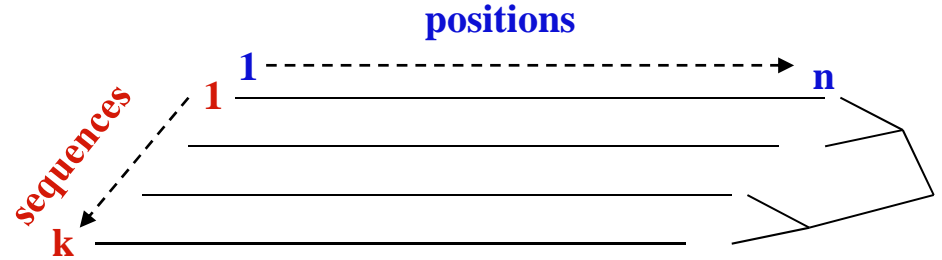
Artemisa Labi, Chris Campbell, Istvan Miklos,

Sequence Analysis:

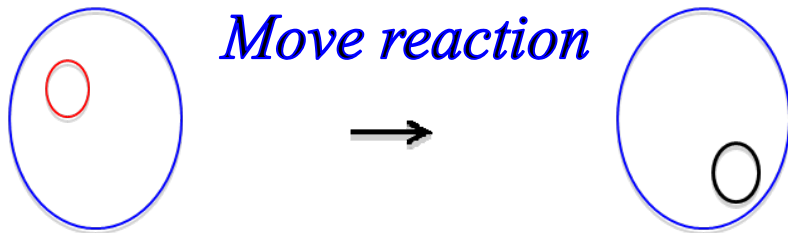
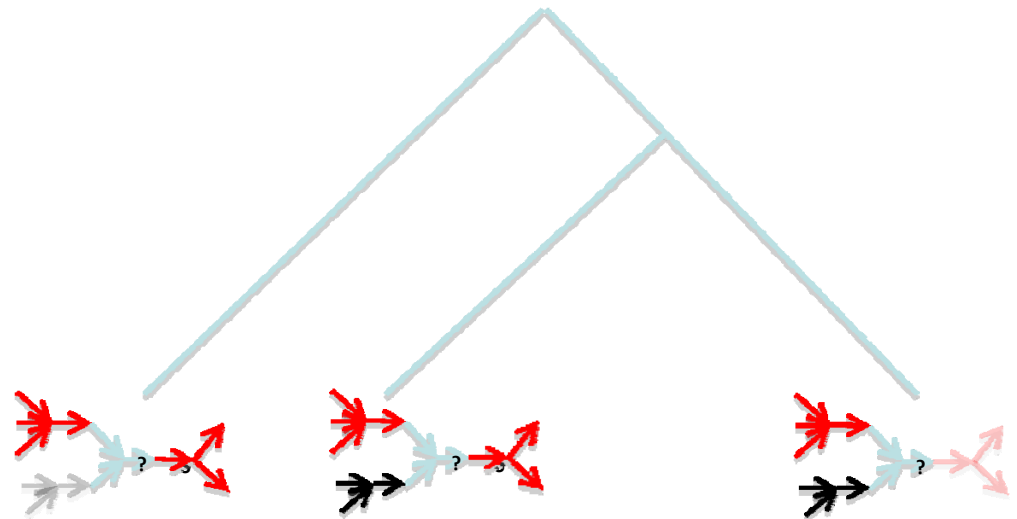
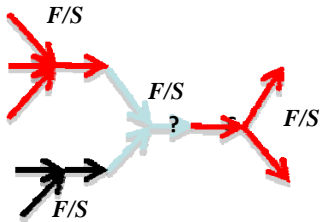


slow - r_s
fast - r_f

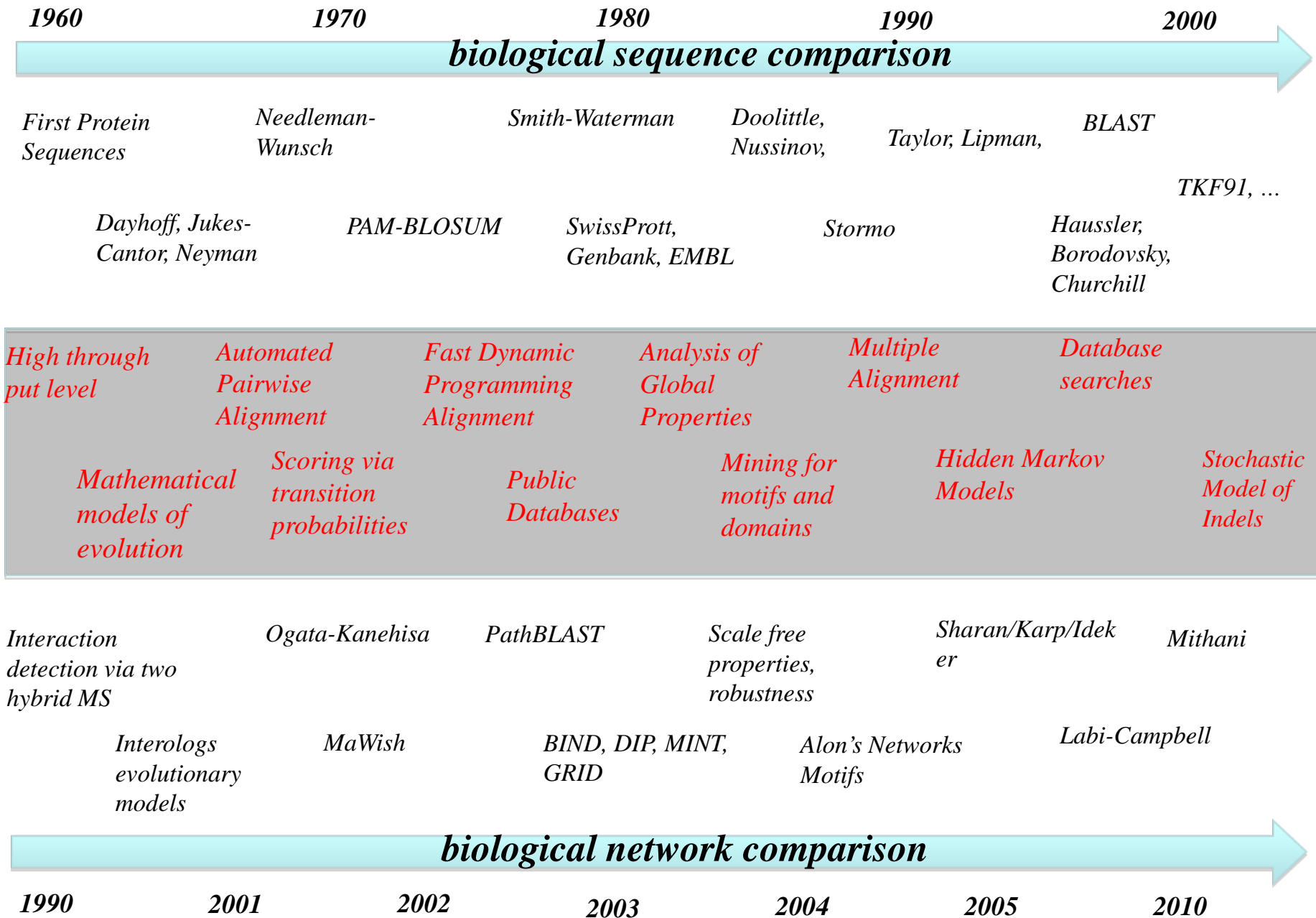
HMM



Network Analysis:



Development of Network/Sequence Analysis



Network Description and Statistics I

Barabasi & Oltvai, 2004

- Degree/Indegree/Outdegree
- Shortest Path $Dist(i, j)$
- Mean Path Length
- Diameter: $Max_{i, j} \{Dist(i, j)\}$
- Clustering Coefficient - $C_I = 2T_I / n_I(n_I - 1)$

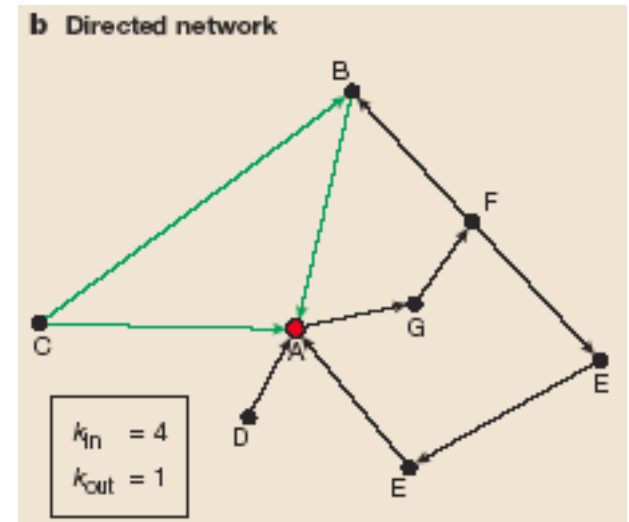
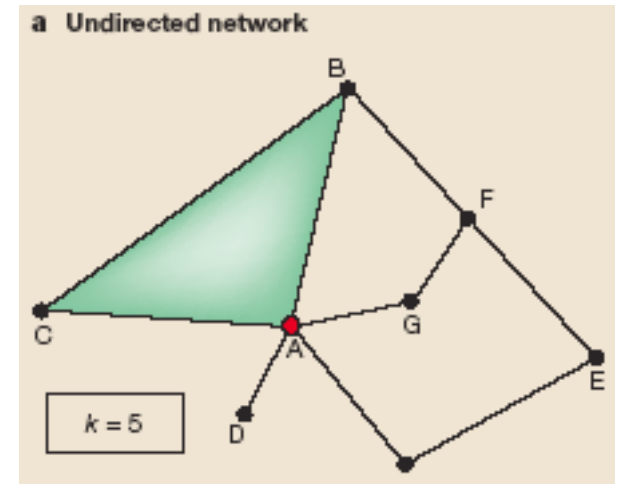
$$C_A = 1/10$$

- Degree Distribution - $P(k)$
- Scale Free Networks $P(k) \sim k^{-\gamma}$ $\gamma > 2$
- Hubs: multiply connected nodes

The lower γ , the more hubs.

Small World Property:

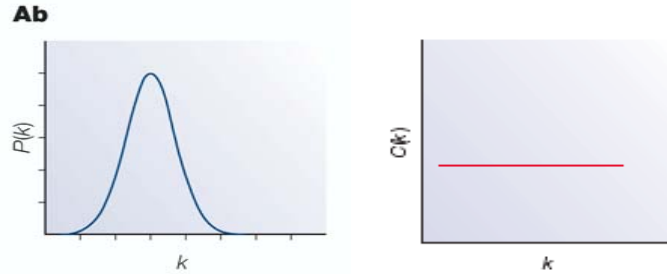
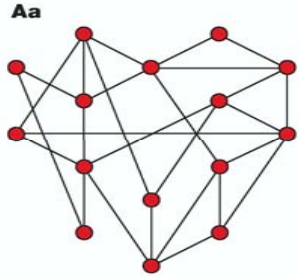
Graph connected and path lengths small



Network Description and Statistics II

Barabasi & Oltvai, 2004

A. Random Networks [Erdos and Rényi (1959, 1960)]



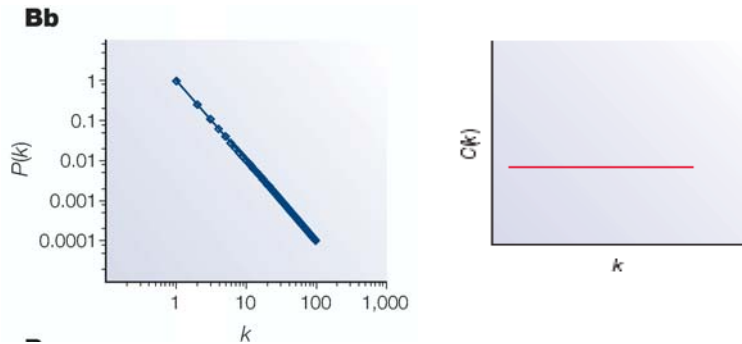
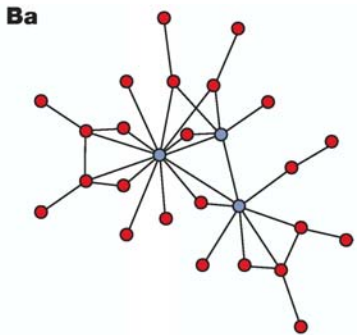
$$P(k) = \frac{e^{-\bar{k}} \bar{k}^k}{k!}$$

Mean path length $\sim \ln(k)$

Phase transition:

Connected if: $p \geq \ln(k)/k$

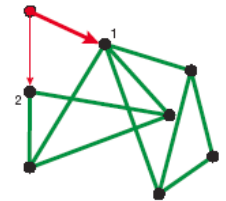
B. Scale Free [Price,1965 & Barabasi,1999]



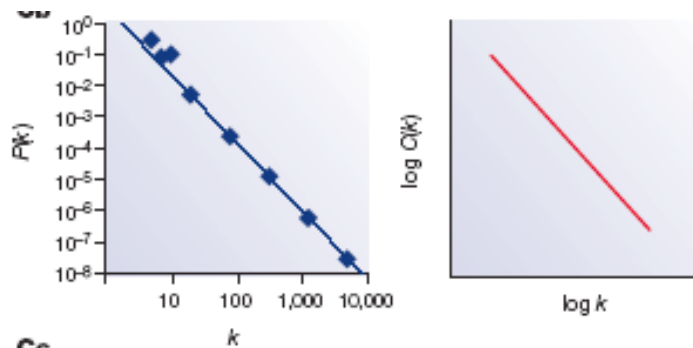
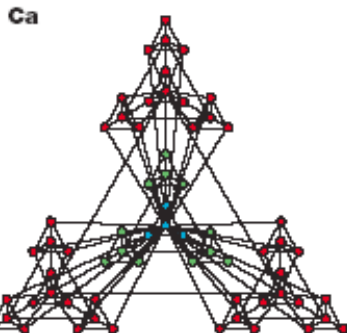
$$P(k) \sim k^{-\gamma}, k \gg 1, 2 < \gamma$$

Mean path length $\sim \ln \ln(k)$

Preferential attachment. Add proportionally to connectedness



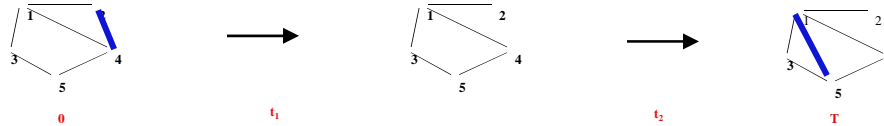
C. Hierarchical



Copy smaller graphs and let them keep their connections.

Stochastic Modeling of Network Evolution

Only topology of networks will be considered. I.e. dynamics and continuous parameters often ignored.



What do models of network evolution do?:

Test models

Estimate Parameters in the Evolutionary Process

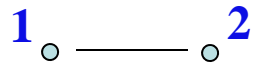
Ancestral Analysis

Framework for Knowledge Transfer

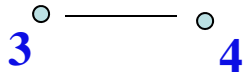
Likelihood of Homologous Pathways

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

Number of Metabolisms:



+ 2 symmetrical versions



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



Approaches:

Continuous Time Markov Chains with computational tricks.

MCMC

Importance Sampling



A Model for the Evolution of Metabolisms

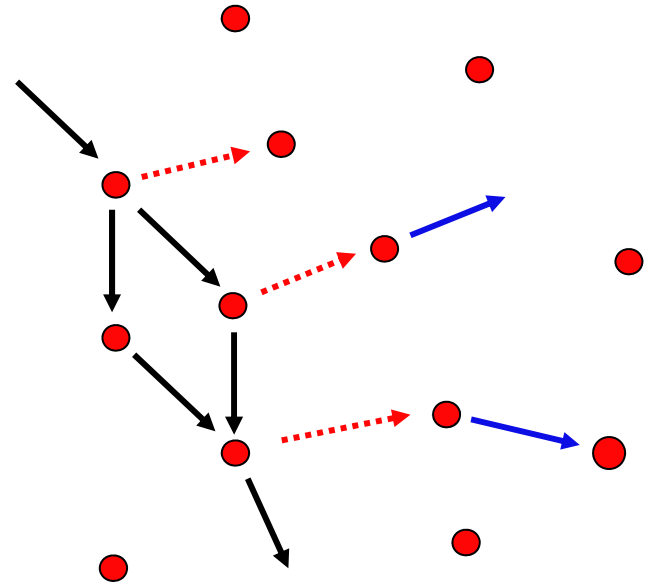
- A given set of metabolites: 
- A given set of possible reactions -
arrows not shown.
- A core metabolism: 
- A set of present reactions - **M**
black and **red** arrows

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

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

$P(N_1 \rightarrow N_2)$ and Corner Cutting

- How many networks could be visited on “almost shortest” paths?

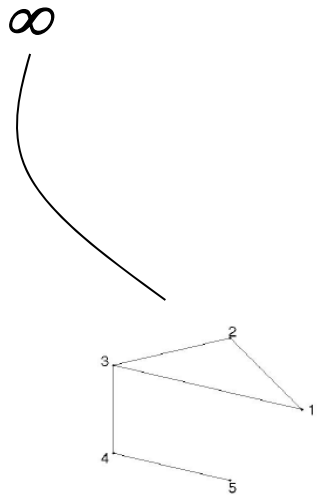


If $d(N_1, N_2) = k$, then there are 2^k networks are visitable on shortest paths. If 2ϵ additional steps are allowed, then $2^k (L + L(L-1)/2 + (L(L-1) \dots (L-\epsilon+1)/\epsilon!)$ are visitable.

Example. 15 nodes, $L=105$, $\lambda t = \mu t = 0.05$, $\epsilon = 2$, $d=4$. $P(4) = e^{-.5} \cdot .5^4 / 4! \sim .003$ $P(6) = e^{-.5} \cdot .5^6 / 6! < 10^{-4}$

How can $P(\infty)$ be evaluated?

Can be found in $P(\infty)$ at appropriate rows.
 In general not very useful (number of metabolisms).



Simulations

Forward with symmetries could be used in specific cases.

Backward (coupling from the past)

A Toy Example

(by Aziz Mithani)

Equilibrium Probability

Metabolic Universe

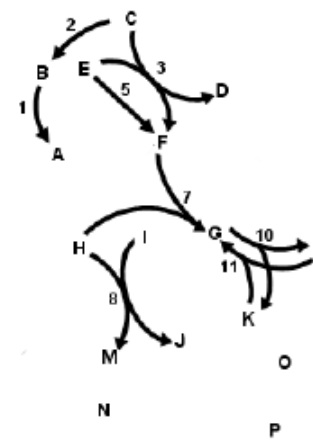
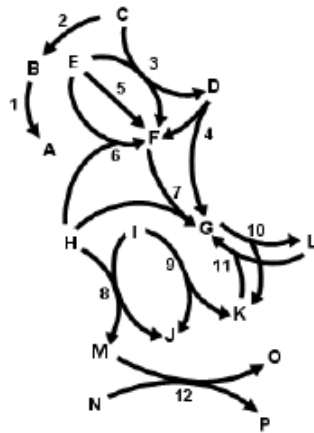
12 possible edges

1i 1u 3

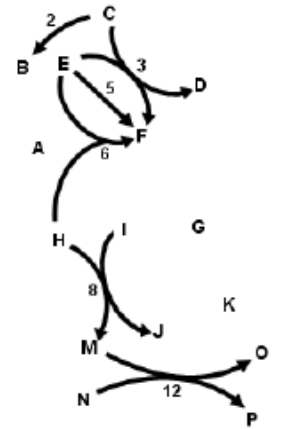
1i 2u 3

2u 1i 3

2i 2u 3



Transition Probability



dist=6

Transition Probability:

Full Exponentiation (2^{12} states 4096)

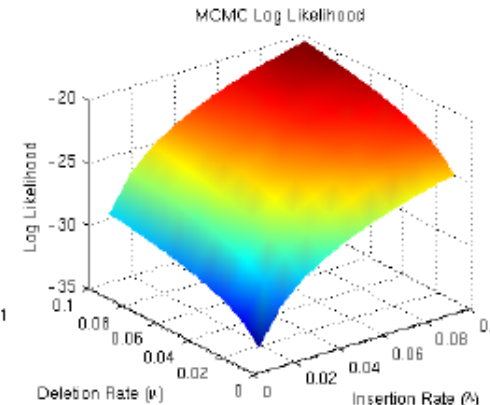
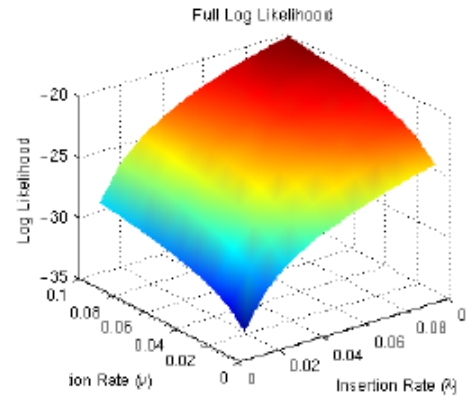
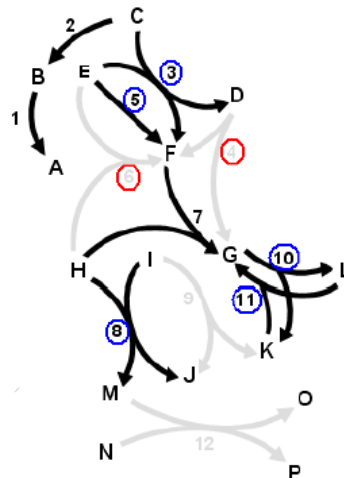
Exponentiation with corner cutting

$2^6 - 64, 384, 960, 1280, 960, 384, 64$

MCMC Integration

Adding Connectedness

Favouring insertions connecting

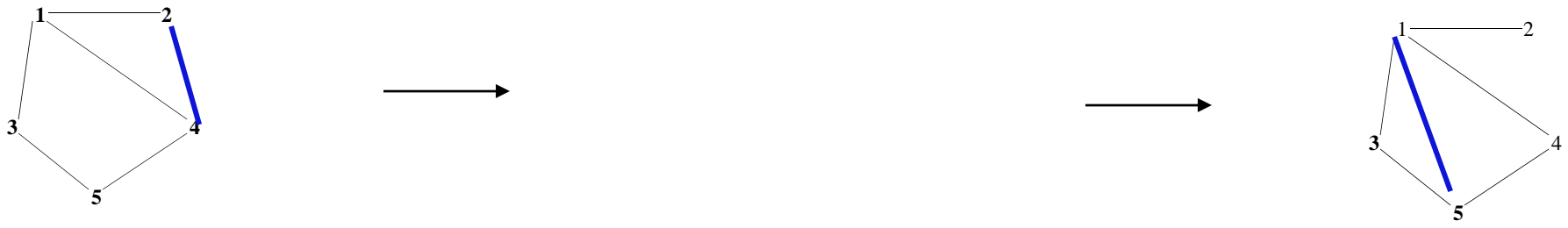


The proportion present:

$$\frac{5}{7} = 0.714$$

$P(N_1 \rightarrow N_2)$ and Corner Cutting

- How many networks could be visited on “almost shortest” paths?

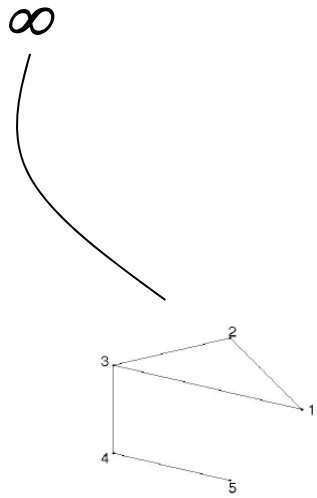


If $d(N_1, N_2) = k$, then there are 2^k networks are visitable on shortest paths. If 2ϵ additional steps are allowed, then $2^k (L + L(L-1)/2 + (L(L-1)..(L-\epsilon+1)/\epsilon!)$ are visitable.

Example. 15 nodes, $L=105$, $\lambda t = \mu t = 0.05$, $\epsilon = 2$, $d=4$. $P(4) = e^{-.5} \cdot .5^4 / 4! \sim .003$ $P(6) = e^{-.5} \cdot .5^6 / 6! < 10^{-4}$

How can $P(\infty)$ be evaluated?

Can be found in $P(\infty)$ at appropriate rows.
 In general not very useful (number of metabolisms).

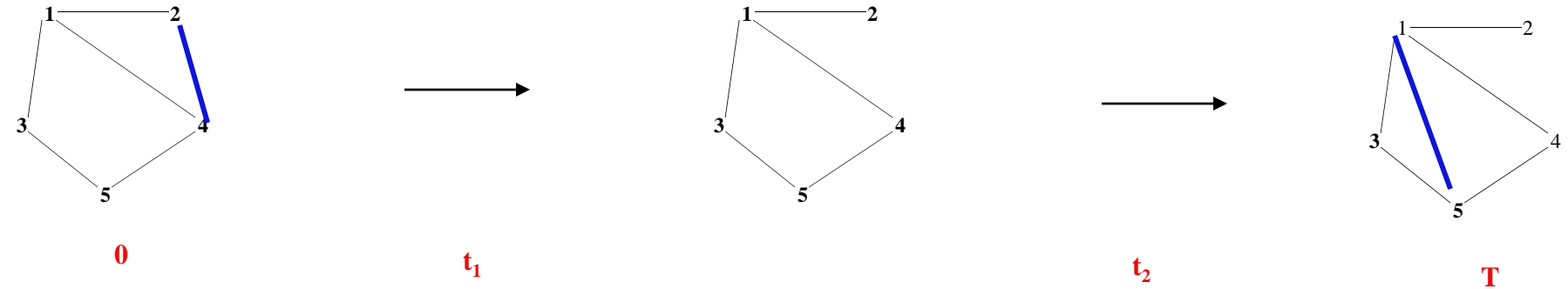


Simulations

Forward with symmetries could be used in specific cases.

Backward (coupling from the past)

Evolving Networks: Integration



- *Integrate of all waiting times (t_1, \dots, t_i) and state assignments of length i gives probability of specific trajectory*

$$P(N \rightarrow \dots N_{i-1} \rightarrow N') = \iiint_{t_1, \dots, t_i} P(N \rightarrow \dots N_{i-1} \rightarrow N'; t_1, \dots, t_i) d\bar{t}$$

- *The above expression can be shown to be of the form*
And recursions $O(N^2)$ exists to calculate coefficients.

$$\prod_{n=1}^N \sum_{n=0}^M e^{-q_{i0}T} \sum_{k=0}^{d_n} c_n^k T^k$$

- *Sum over i state assignments gives probability of paths of length i .*

$$P(N \rightarrow N'; i \text{ steps}) = \sum_{N_1, N_2, \dots, N_i} P(N \rightarrow \dots N_{i-1} \rightarrow N')$$

- *Sum over all path lengths gives probability of N turning into N'*

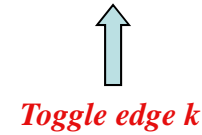
$$P(N \rightarrow N') = \sum_i P(N \rightarrow N'; i \text{ steps})$$

Evolving Networks: MCMC

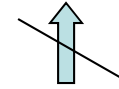
Present pathway:



• *Insertion of an edge pair*



• *Deletion of an edge pair*

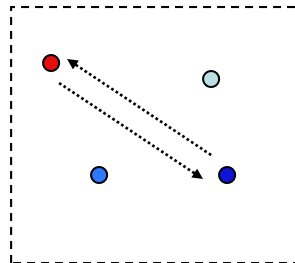


• *Moving of a pair or singles*



• *Metropolis-Hasting integrating of all paths - Green (1995) version:*

Set of paths:



Likelihood - $L(\bullet)$

Probability of going from \bullet to \bullet - $q(\bullet, \bullet)$

J - Jacobian

Acceptance ratio

$$\frac{L(\bullet)q(\bullet, \bullet)}{L(\bullet)q(\bullet, \bullet)} J$$

A Toy Example

(by Aziz Mithani)

• Metabolic Universe

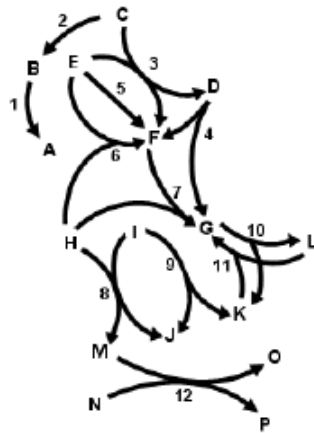
• 12 possible edges

1i 1u 3

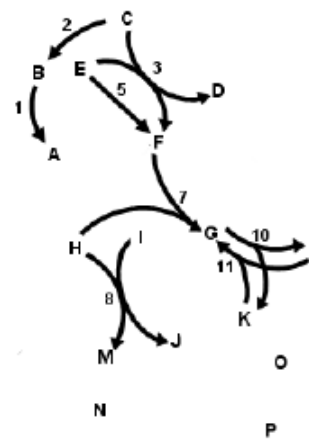
1i 2u 3

2u 1i 3

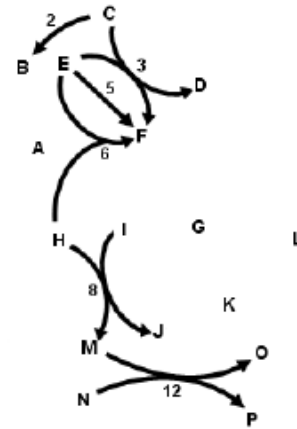
2i 2u 3



Equilibrium Probability



Transition Probability



dist=6

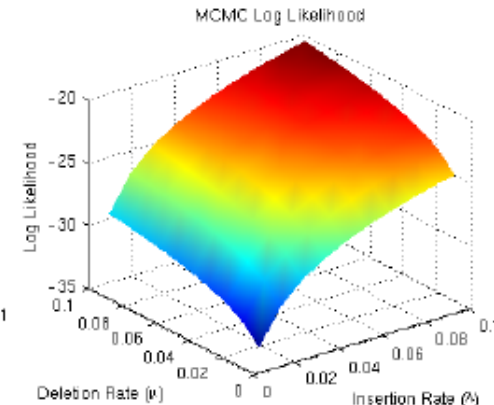
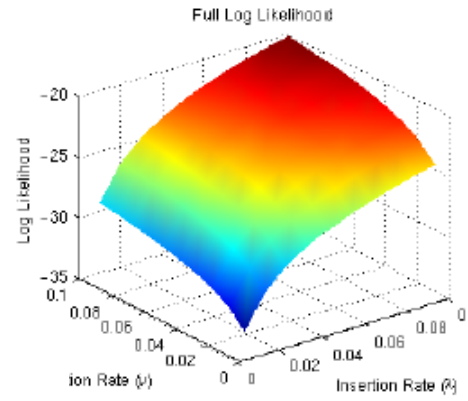
Transition Probability:

Full Exponentiation (2^{12} states 4096)

Exponentiation with corner cutting

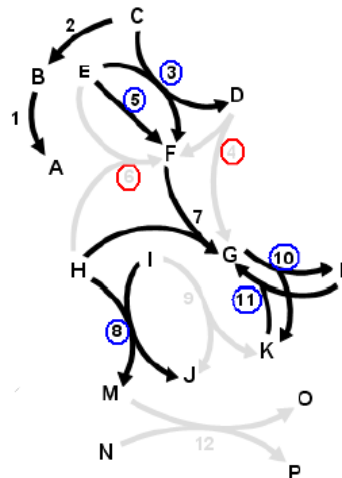
$2^6 - 64, 384, 960, 1280, 960, 384, 64$

MCMC Integration



Adding Connectedness

Favouring insertions connecting

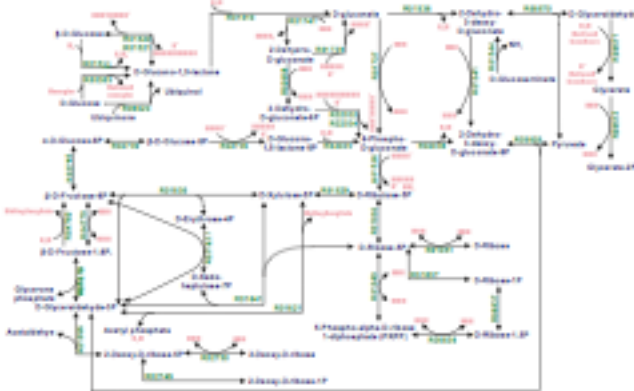


The proportion present: $\frac{5}{7} = 0.714$

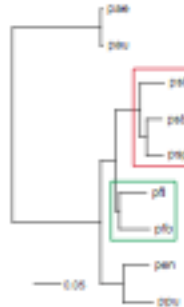
Application and extension to Phylogeny

One metabolism

Pentose Phosphate Pathway



Pseudomonas



Pairwise analysis

Pathway Map	Start Organism	End Organism	Differences*	CnI ^{***}	λ_t	STD	μ_t	STD	λ/μ	
Pentose phosphate pathway (MAP00030)	<i>P. aeruginosa</i> PAO1	<i>P. syringae</i> DC3000	9 (I:8, D:1)	-	0.31694	0.00853	0.06168	0.00656	5.13849	
	<i>P. fluorescens</i> PF-5	<i>P. fluorescens</i> Pfl0-1	2 (I:2, D:0)	+	1.91912	0.26728	0.32402	0.12122	5.92285	
					-	0.09724	0.00176	0.01852	0.00186	5.25147
Lysine degradation (MAP00310)	<i>P. aeruginosa</i> PAO1	<i>P. syringae</i> DC3000	1 (I:1, D:0)	-	0.02283	0.00058	0.06976	0.00576	0.32732	
	<i>P. fluorescens</i> PF-5	<i>P. fluorescens</i> Pfl0-1	2 (I:0, D:2)	+	0.38110	0.22654	3.25666	3.28194	0.11702	
					-	0.00787	0.00074	0.29881	0.00534	0.02635
Phenylalanine metabolism (MAP00360)	<i>P. aeruginosa</i> PAO1	<i>P. syringae</i> DC3000	6 (I:4, D:2)	-	0.07867	0.00759	0.48296	0.08009	0.16289	
	<i>P. fluorescens</i> PF-5	<i>P. fluorescens</i> Pfl0-1	7 (I:2, D:5)	+	0.52749	0.13457	1.95439	0.71231	0.26990	
					-	0.04486	0.00528	0.78573	0.06140	0.05709
					+	0.34472	0.10428	2.05643	0.46145	0.16763

Multiple analysis with parameter for neighbor dependence

Pathway Map	Phylogeny	$E(\delta)$	$\text{var}(\delta)$	$E(\lambda)$	$\text{var}(\lambda)$	$E(\mu)$	$\text{var}(\mu)$	λ/μ
Pentose phosphate pathway (MAP00030)	(pae,(pfl,pfo))	0.3345	0.0072	2.2492	1.8556	1.0969	0.1865	2.0505
	(pae,((psb,psp),pst))	0.3368	0.0073	1.7887	1.3182	0.8588	0.0998	2.0829
	<i>Pseudomonas</i>	0.2508	0.0044	1.2345	0.0580	0.9431	0.0178	1.3089
Lysine degradation (MAP00310)	(pae,(pfl,pfo))	0.0861	0.0037	0.6244	0.2013	2.6961	4.7734	0.2316
	(pae,((psb,psp),pst))	0.0707	0.0024	0.4732	0.0433	4.0998	6.1328	0.1154
	<i>Pseudomonas</i>	0.0660	0.0022	0.4955	0.0235	2.2045	2.3058	0.2248
Phenylalanine metabolism (MAP00360)	(pae,(pfl,pfo))	0.0634	0.0020	0.9337	0.1257	3.2754	2.9057	0.2851
	(pae,((psb,psp),pst))	0.0552	0.0019	0.9180	0.0825	3.0896	2.6660	0.2971
	<i>Pseudomonas</i>	0.0479	0.0014	0.8943	0.0417	1.9824	0.5205	0.4511

Main results

Evolution rates

- Insertion rate higher for pathway maps involved in central metabolism and amino acid biosynthesis than those involved in secondary metabolism and amino acid degradation
- Rates higher in *Pseudomonas syringae* compared to *Pseudomonas fluorescens*
 - Supports experimental findings – High number of deletions in *P. syringae* lineage
- Low insertion to deletion ratio (λ/μ) for pathway maps related to amino acids which are poor nutrient sources

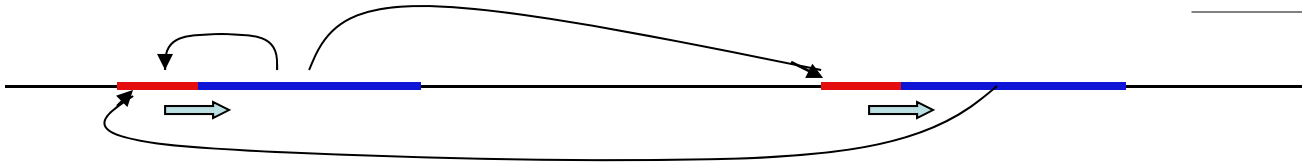
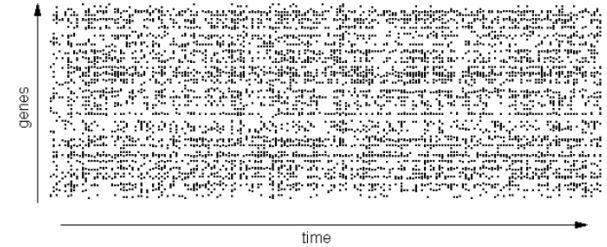
Neighbourhood structure

- Pathway maps involved in central metabolism and metabolism of essential amino acid have strong neighbourhood structure.

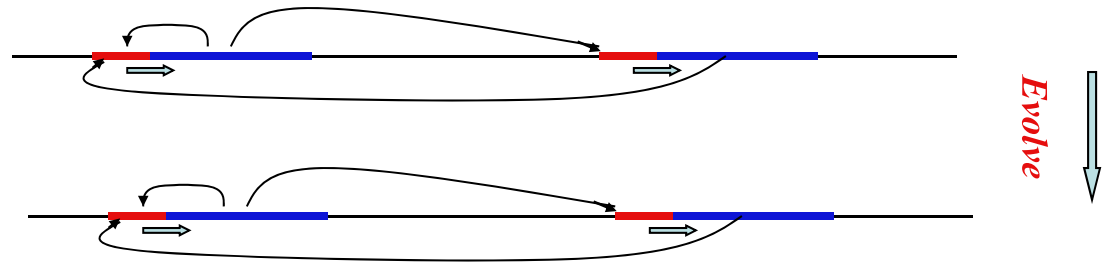
Regulatory Network Evolution

Artificial Genome
Riel, 1999:

- Regulatory control according to rules
- Proteins can bind the regulatory regions



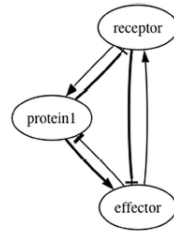
Evolving Artificial Genome
Quant & Bullocks, 2007:



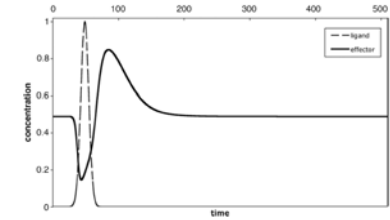
- Selection will influence final dynamics

Networks: Signal Transduction Pathways

- Dynamics**



	receptor	protein 1	effector
receptor	0.000	-0.986	0.007
protein 1	0.020	0.000	-0.040
effector	-0.733	0.726	0.000



*One protein is receptor, one effector.
Activating receptor creates cascade effect
described by simple equation system.*

$$\frac{d[P_i]}{dt} = [P_i^* \sum_j l_{ij} [P_j^*]] - [[P_j] (\delta_{il} [L] \sum_j k_{ij} [P_j^*])]]$$

- Mutational Process:** *recruitment/loss + change of interactions*

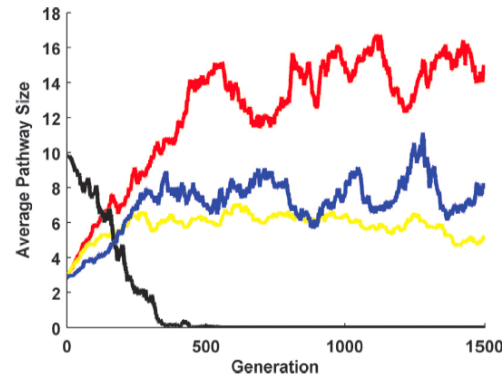
- Fitness**

$$F = 1 - nc \quad \text{if } \alpha = 1$$

$$F = 0 \quad \text{if } \alpha = 0$$

n - number of proteins, c - fitness cost per protein, α - functionality criteria

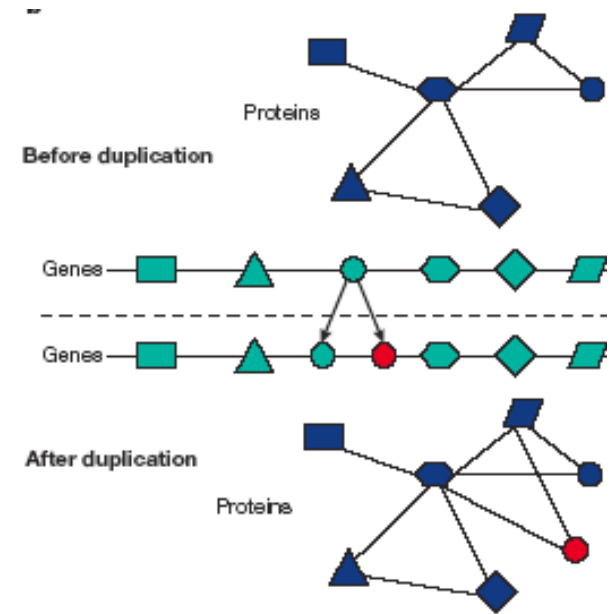
- Evolution**



Models of Protein Interaction Networks Evolution

Barabasi & Oltvai, 2004 & Berg et al. ,2004; Wiuf et al., 2006

- A gene duplicates
- Inherits its connections
- The connections can change



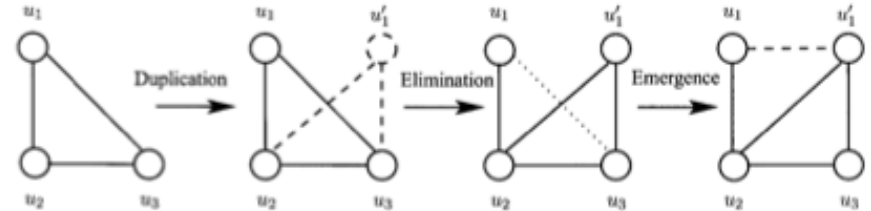
Berg et al. ,2004:

- Gene duplication slow $\sim 10^{-9}$ /year
- Connection evolution fast $\sim 10^{-6}$ /year
- Observed networks can be modeled as if node number was fixed.

Network Alignment

Basic Operations:

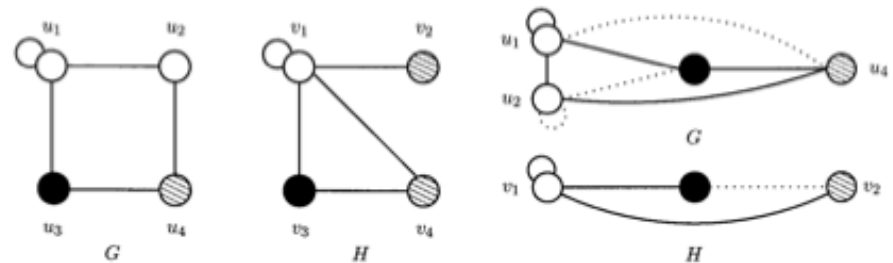
- **Duplication**
- **Elimination** (loss of interaction – edge)
- **Emergence** (gain of interaction – edge)
- **Deletion** (loss of node with edges)



Associated cost function for each operation and cut-off Δ (matching disallowed)

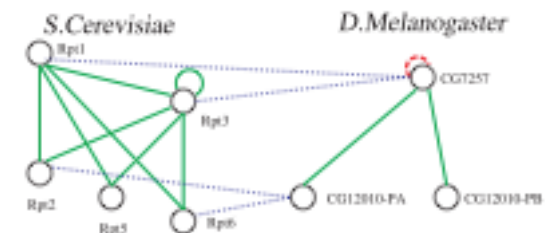
Heuristic Search of optimal matching subnetworks

Example of small matched networks



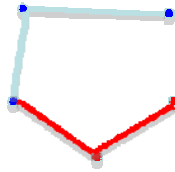
Application to Yeast, C.elegans, D.melanogastor,

Organism pair	# Nodes	# Matched nodes		# Matches		# Mismatches	# Duplications	
		$\bar{\Delta} = 1$	$\bar{\Delta} = 2$	$\bar{\Delta} = 1$	$\bar{\Delta} = 2$	$\bar{\Delta} = 1$	Org. 1	Org. 2
SC vs CE	2746	312	1230	412	3007	40262	6107	6886
SC vs DM	15884	1730	8622	2061	42781	1054241	6107	32670
CE vs DM	11805	491	3391	455	6626	205593	6886	32670

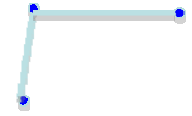


The size of the human interactome

True network – $G_N=(V_N, E_N)$



Sampled network $G_S=(V_S, E_S)$



$$P_{p,\theta}(G_S) = \sum_{G_N \supseteq G_S} P_p(G_S|G_N) P_\theta(G_N)$$

$$P_{p,\theta}(G_S) = Q_p(N_S) \sum_{G_N \supseteq G_S} q(G_S, G_N) P_\theta(G_N)$$

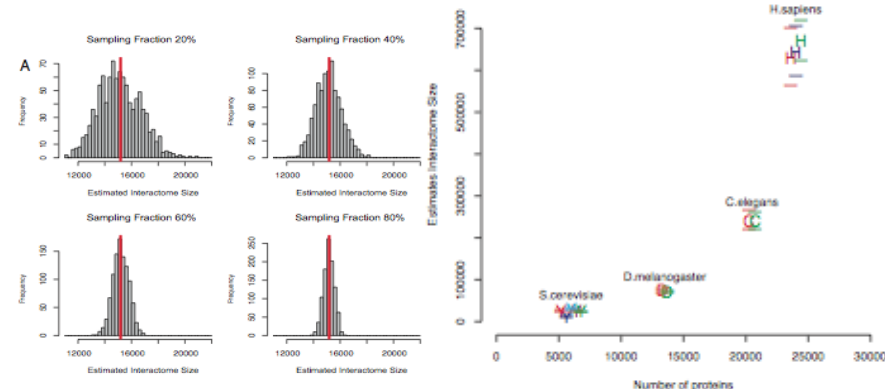
$$\hat{p} = N_S / N_N$$

$$P_\theta(G_N^* | G_S) = q(G_S, G_N^*) P_\theta(G_N^*) / \sum_{G_N \supseteq G_S} q(G_S, G_N) P_\theta(G_N)$$

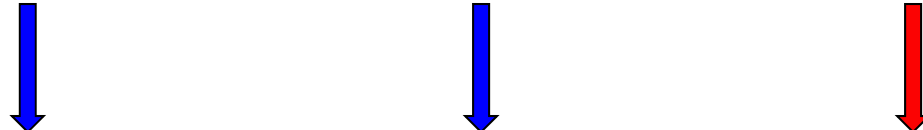
$$\hat{\pi} = \frac{2M_S}{N_S(N_S - 1)}$$

$$\hat{M}_N = M_S \frac{N_N(M_N - 1)}{N_S(N_S - 1)}$$

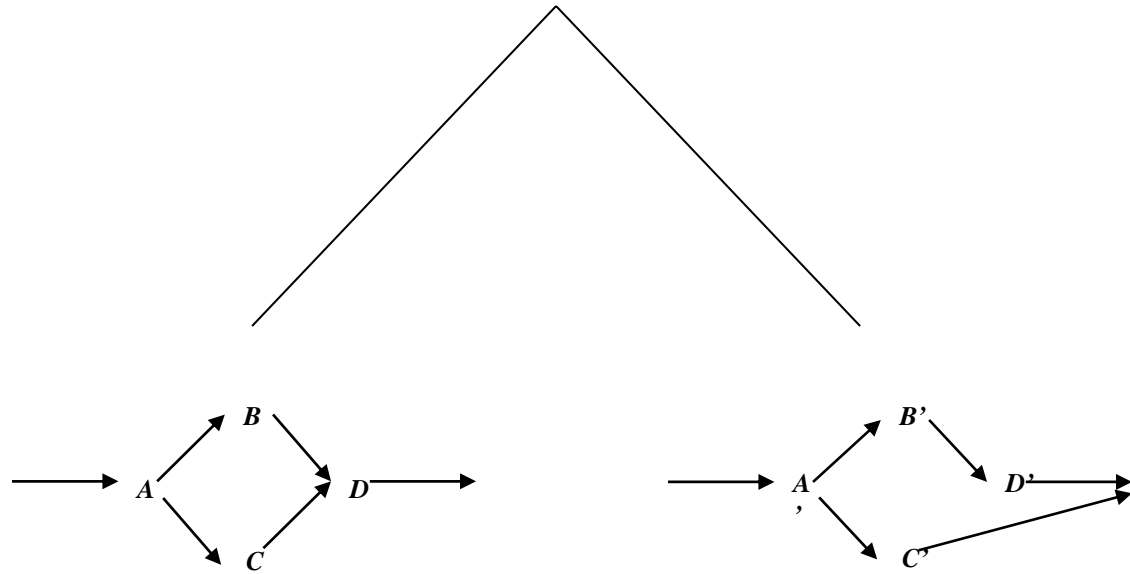
Species	Nodes	Edges	MN	95% CI
Yeast	4,959	17,226	25,229	24,100- 26,440
Drosophila	7,451	22,636	74,336	71,700- 77,100
C.Elegans	2,638	3,970	240,544	220,030-263,270
H.Sapiens	1,085	1,346	672,918	625,170-722,670



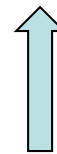
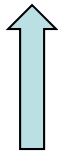
Inference and Evolution

$$P(D_{mouse}, D_{human}) = \sum_{N_1, N_2} P(D_{human} | N_{human}) P(D_{mouse} | N_{mouse}) P(N_{human}, N_{mouse})$$


Evolve



Infer network



Observe (data)

Human

Mouse

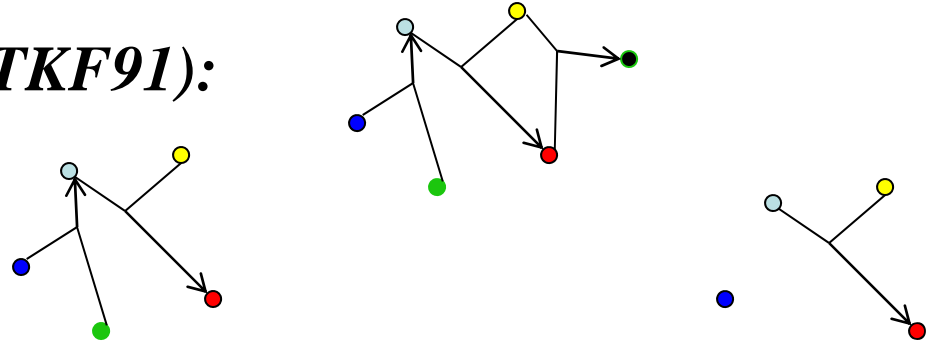
Suggestion: Evolving Dynamical Systems

- *Goal: a time reversible model with sparse mass action system of order three!!*

Adding/Deleting components (TKF91):

Add rate: $(k+1)\lambda$

Delete rate: $k\mu$

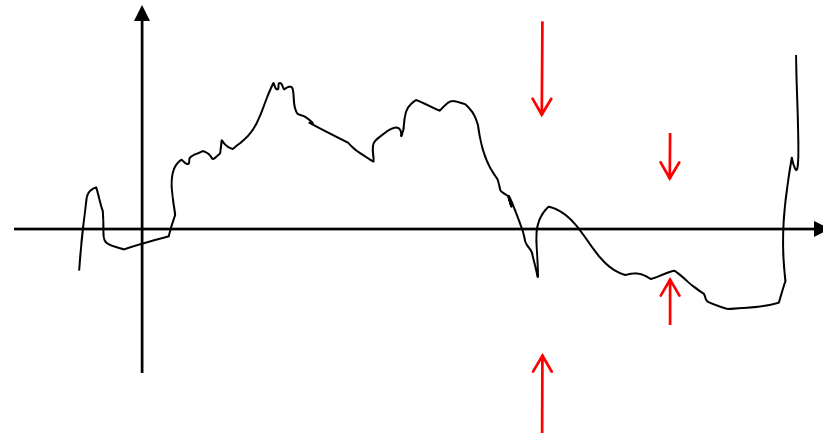


Adding reactions with birth of component:

There are $3k(k-1)$ possible reactions involving a new-born

Reaction Coefficients:

- *Continuous Time Continuous States Markov Process - specifically Diffusion.*
- *For instance Ornstein-Uhlenbeck, which has Gaussian equilibrium distribution*



Summary

The importance of modelling

The main classes of networks

The development of sequence and networks models

Integration of paths

Basic Models

Modeling Dynamic Systems Evolution