

BioInformatica – Presentazione finale

***Una introduzione alle reti metaboliche e alla
loro analisi strutturale***

Saverio Delpriori, Luca Puccini

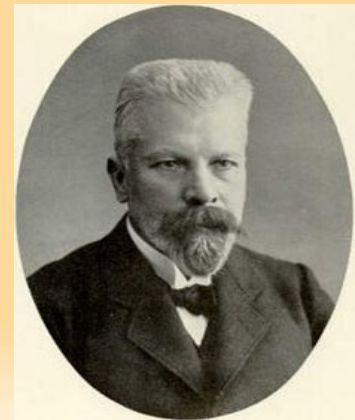
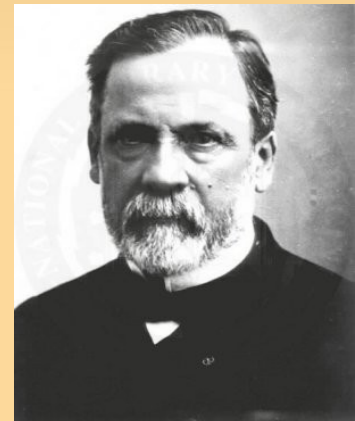
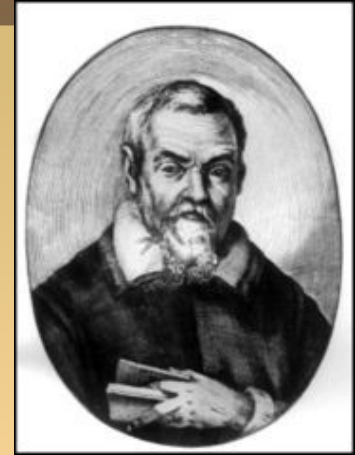
A.A. 2008 - 2009

Scaletta

- **Introduzione**
- Analisi strutturale
- Analisi topologica
- L'Evoluzione del Metabolismo

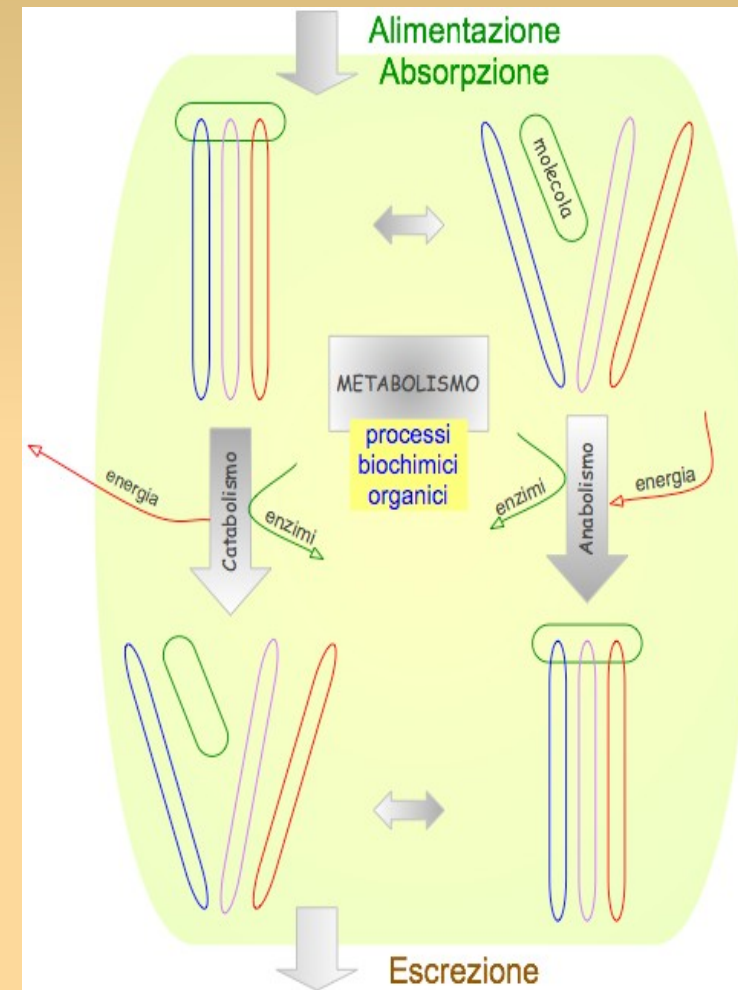
Cenni Storici

- La parola metabolismo deriva dal greco *metabole*, cambiamento.
- Santorio Santorio, medico italiano del XVII secolo, compie i primi esperimenti, comunque di scarsa valenza scientifica
- Due secoli dopo Louis Pasteur dimostra che le reazioni chimiche nella cellula non sono differenti da quelle di laboratorio
- All'inizio del XX secolo Eduard Buchner scopre gli enzimi, separando gli studi sul metabolismo di un organismo dagli studi sulla funzione biologica delle sue cellule



Il metabolismo: Anabolismo e Catabolismo

- Le reazioni chimiche in un metabolismo sono rappresentate come un "percorso metabolico":
 - Anabolico: se vengono sintetizzate molecole tramite l'uso di energia e il consumo di agenti riducenti
 - Catabolico: se vengono prodotte energia ed agenti riducenti degradando molecole



Scaletta

- Introduzione
- **Analisi strutturale**
 - Definizioni
 - Ricostruzione metabolica
 - Modellazione del metabolismo
- Analisi topologica
- L'Evoluzione del Metabolismo

Definizioni

- Una **rete metabolica** è una collezione di oggetti e relazioni tra loro
 - **Composti chimici o metaboliti**: piccole molecole sintetizzate o degradate in un organismo
 - **Reazioni biochimiche**: reazioni che a partire da un insieme di composti (*sostrati*) producono altri composti (*prodotti*). La direzione in cui avvengono queste reazioni dipendono dalle condizioni fisiologiche in cui si verificano
 - **Enzimi**: proteine che permettono il verificarsi di reazioni e ne aumentano sensibilmente la velocità
 - **Geni**: sequenze di acidi nucleici che servono a codificare gli enzimi
 - **Cofattori**: piccole molecole non proteiche che si associano agli enzimi per attivarli (*attivatori allosterici*) o disattivarli (*inibitori allosterici*)

Ricostruzione metabolica (1)

- La ricostruzione di una rete metabolica consiste nell'inferire relazioni tra geni, enzimi e reazioni in un dato sistema metabolico
- Per fare ciò vengono utilizzate tecniche di genomica comparativa confrontando DNA sconosciuti a DNA noti reperibili presso database pubblici
- In seguito i risultati sono raffinati utilizzando dati metabolomici
- Tecnica tanto più accurata quanto più è accurato il DNA di riferimento

Ricostruzione metabolica (2)

- La ricostruzione metabolica è composta da quattro fasi principali
 - Viene fornita un'annotazione funzionale dei geni metabolici mappando il genoma
 - Viene definita la relazione tra annotazioni funzionali e reazioni biochimiche
 - I dati vengono raffinati e gli errori corretti
 - Viene modellata la rete metabolica

Annotazione funzionale

- Assegnazione di una funzione ai geni in modo che non risulti ambigua
 - **Gene Ontology**: ha stabilito un vocabolario non ambiguo per geni omologhi
 - **Enzyme Commission**: ha formalizzato un metodo numerico per la classificazione gerarchica degli enzimi con 4 interi (*EC Number*)
A . B . C . D → *Tipo . Sottoclasse . Sotto-sottoclasse . Seriale*
- Il secondo metodo è molto utilizzato ma può condurre ad ambiguità.

Approcci al mappaggio (1)

- **Approccio** → **ortologia**: è il metodo più usato, si basa sul confronto di sequenze di proteine
 - *Enzimi ortologhi*: stessa attività, derivano da un comune gene ancestrale diviso in seguito ad un evento di speciazione
 - *Enzimi Paraloghi*: attività differenti, derivano da una duplicazione intragenomica
- L'approccio ricerca i geni ortologhi, cercando di riconoscere (ed evitare) i paraloghi
 - **BRH**: proteine simili sono classificate nello stesso gruppo
 - **Riconciliazione di alberi**: strutture ad albero per il confronto e il riconoscimento di geni ortologhi

Approcci al mappaggio (2)

- **Approccio** → **identificatore di sequenza**: raggruppa i geni confrontandone i "moduli" ovvero le più lunghe sottosequenze comuni. È soggetto ad errori su sequenze lunghe
- **Approccio** → **contesto genomico** : si basa sull'ipotesi che proteine funzionalmente legate si siano evolute in modo correlato
- **Altri approcci**: vi sono anche approcci
 - Basati su informazioni generali sulle proteine
 - Basati sull'unione dei precedenti

Definizione di relazioni tra annotazioni funzionali e reazioni biochimiche

- Ricerca dei collegamenti tra le annotazioni funzionali trovate nelle precedenti fasi e le reazioni chimiche corrispondenti
- Questa fase è essenzialmente automatica e si basa su ricerche all'interno di vari database
 - *ENZYME*: informazioni sulle proteine che hanno un EC Number
 - *Uniprot*: vasto catalogo di informazioni sulle proteine
 - *BRENDA*: contiene anche dettagli aggiuntivi sul sostrato
 - *BYOCYC*, *KEGG*: i database più comunemente utilizzati
 - ...

Raffinamento della ricostruzione metabolica

- Una volta completata la ricostruzione metabolica si possono riscontrare diversi errori e mancanze che possono essere dovute a
 - Bassa similitudine tra il gene esaminato e il gene di riferimento
 - Prodotti delle reazioni ottenibili tramite percorsi alternativi
 - ...

Per completare il percorso esistono vari metodi, tutti di tipo euristico

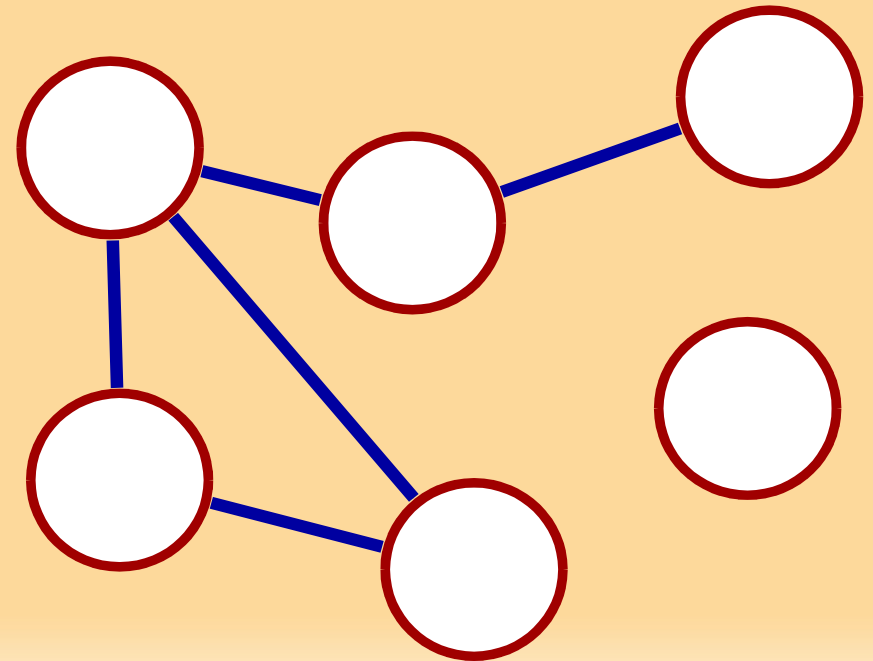
- In questa fase si ha poi:
 - Definizione della reversibilità della reazione
 - Inferenza di reazioni dai dati sui metaboliti
 - Utilizzo di dati sperimentali

Modellazione del metabolismo

- Un metabolismo viene studiato da due possibili punti di vista e ognuno di questi richiede un tipo differente di modellazione
- Punto di vista strutturale
 - Modello basato su grafi
 - Modello basato su vincoli (constraint based)
- Punto di vista dinamico
 - Equazioni differenziali
- Verrà posta l'enfasi principalmente sul primo punto di vista, e in particolare sui grafi.

I grafi

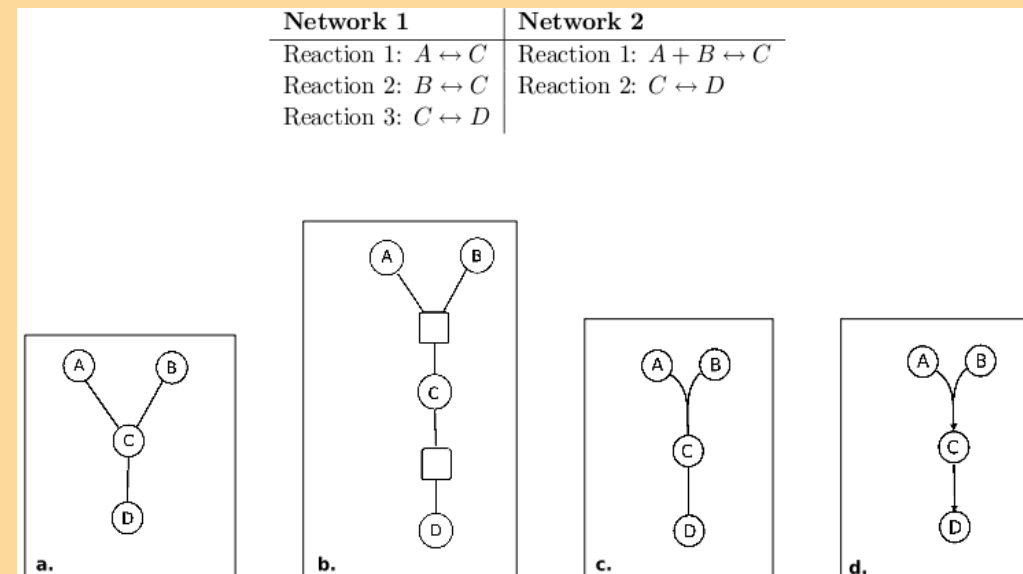
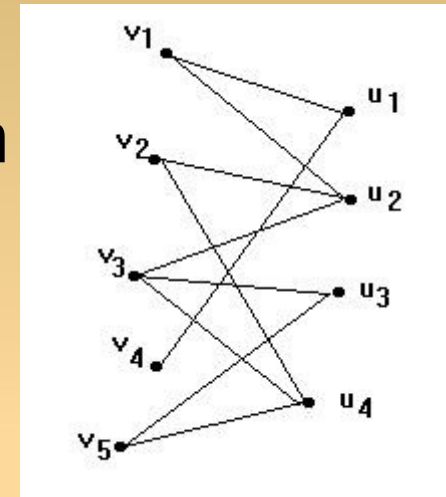
- Un grafo è definito da una coppia $G(N, A)$
- Utilizzare un grafo per modellare una rete metabolica significa scegliere quali entità biologiche associare a nodi e archi
 - Grafo di composti
 - Grafo di reazioni
 - Grafo degli enzimi



Modelli più espressivi

Utilizzati per risolvere problemi di ambiguità causati dai normali grafi

- *Grafi bipartiti*: sono grafi i cui nodi sono divisi in due sottoinsiemi disgiunti e in cui ogni arco deve avere un vertice in un insieme e l'altro nell'altro
- *Ipergrafi*: sono grafi in cui un arco può connettere più di due nodi



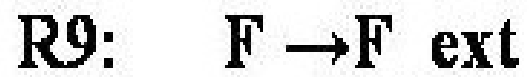
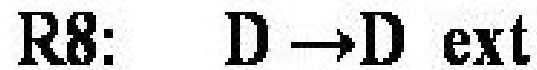
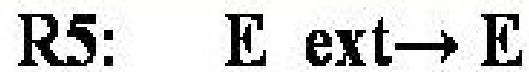
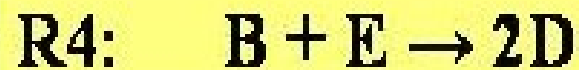
Possibili semplificazioni

- Nei precedenti modelli non si è tenuto in considerazione il fatto che alcuni composti e alcune reazioni abbiano un'importanza inferiore o superiore ad altri.
- Composti come ATP e NADH sono ad esempio coinvolti in una grande quantità di reazioni.
- In una rete metabolica parificarli ad altri composti aumenta di molto e in maniera artificiosa i legami tra le varie reazioni
- Per ovviare a questo problema si rimuovono i composti che compaiono in un grande numero di reazioni, anche se questo porta ad altri tipi di problemi
- Un altro metodo è quello di rimuovere composti quando compaiono come "secondari", qualora la distinzione tra composti si possibile

Modelli basati su vincoli

- Rete modellata come una matrice stechiometrica S
- Corrisponde ad un ipergrafo etichettato
- Le etichette sono i coefficienti stechiometrici dei composti nella reazione
- Il segno dei coefficienti indica se sono prodotti o consumati
- L'attenzione è incentrata sulla distribuzione delle masse in una reazione sotto certi vincoli, al fine di studiare agevolmente le capacità metaboliche di un organismo

Matrice stechiometrica



	R1	R2	R3	R4	R5	R6	R7	R8	R9
A	+1	-1	-1						
B		+1		-1		-2			
C			+1			+1	-1		
D				2			+1	-1	
E				-1	+1				
F						+1			-1

Vincoli e vettore di flussi

- Sia v un vettore di flusso, cioè un m -vettore in cui l' i -esimo elemento descrive il flusso nella reazione i
- Si considerano come vincoli:
 - Stato stabile: $S^*v = 0$
 - Vincoli termodinamici: $v_i > 0, \forall i \in Irrev$
- Questi vincoli definiscono una porzione nello spazio dei flussi rappresentata da un poliedro convesso che contiene tutti i flussi possibili

FBA

- Flux Balance Analysis: ricerca di un flusso ammissibile che ottimizzi la funzione obiettivo (produzione di ATP o studio della biomassa)
- Questo metodo è utilizzabile anche nel calcolo degli effetti sul fenotipo della cancellazione di geni o di altri tipi di perturbazioni del sistema
- Per eseguire questo tipo di predizioni basta porre dei vincoli che impongano l'uguaglianza a zero del flusso corrispondente al gene in questione
- Il metodo assume che le mutazioni si manifestino in un sistema metabolico ottimo

MoMA

- Questo metodo è simile al precedente ma non viene fatta l'assunzione del sistema ottimo
- Viene fornita invece una soluzione approssimata per un vettore di flusso sub-ottimale più simile a quello nello stato non perturbato
- Questo implica una differente ottimizzazione rispetto al metodo FBA, detta "*distance minimization in flux space*"

Altri metodi

- Esistono anche metodi per la ricerca di vettori in grado di generare tutti i vettori d'interesse, qualora ve ne sia più di uno
 - **Metodo elementare**: viene trovato un vettore che soddisfa i due vincoli visti in precedenza più un ulteriore vincolo necessario a rendere il problema la definizione formale di percorso metabolico
 - **Extreme pathways**: simile al precedente ma col vincolo che tutte le reazioni di scambio devono essere irreversibili
 - **Minimal T invariants**: concetto legato alle reti di petri, cioè grafi bipartiti diretti composti dai due insiemi di nodi "posti" e "transizioni". Questo metodo ammette solo reazioni irreversibili

Scaletta

- Introduzione
- Analisi strutturale
- **Analisi topologica**
 - Proprietà del grafo di una Rete Metabolica
 - Modularità
- L'Evoluzione del Metabolismo

Misure sintetiche del grafo di una Rete Metabolica (RM)

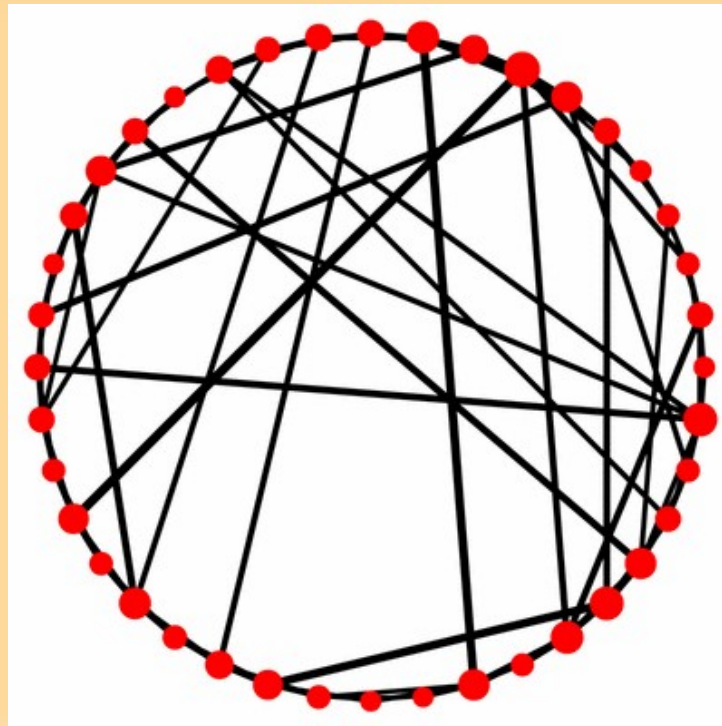
- Grado di un nodo
- Distanza Media fra i nodi
- Vicinanza dal centro
- Diametro
- Clustering Coefficients (distanza da una cricca)
- Assortativity Coefficients

$$ac = \frac{\sum_k (e_{kk} - a_k b_l)}{1 - \sum_k a_k b_l} \quad a_k = \sum_l e_{kl} \quad b_l = \sum_k e_{kl}$$

- Betweenness del nodo/dell'arco
- Synthetic Accessibility

Small - World Networks

- Esperimento dei “6 gradi di separazione” (Milgram)
- Il diametro cresce logaritmicamente (Watts e Strogatz)



Small – World nelle RM

- Fell e Wagner

Mostrano come la RM del batterio *Escherichia Coli* soddisfi le proprietà di SW networks

- Arita

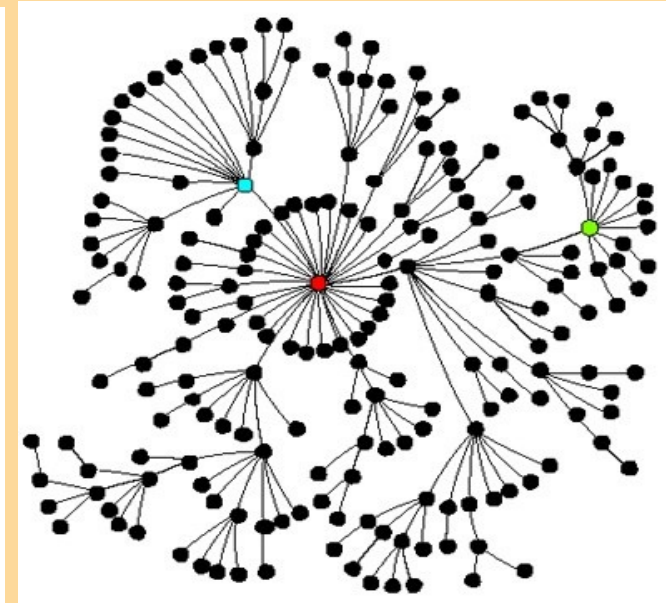
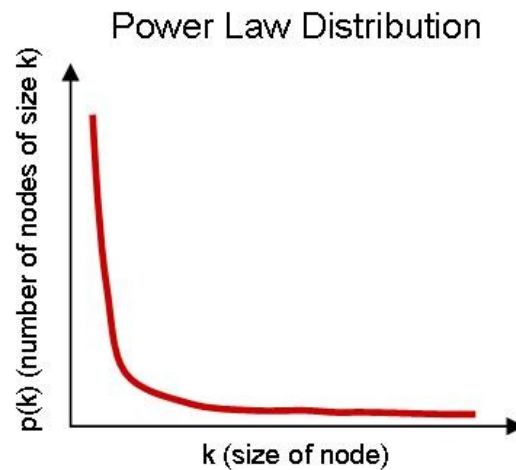
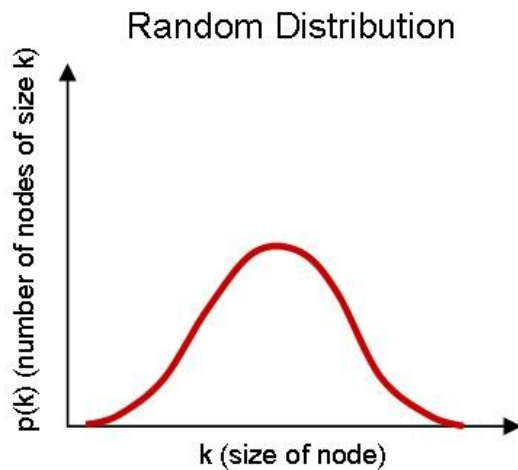
Riformula il modello di F&W fornendo una **nuova definizione strutturale** in cui il diametro è molto maggiore

- Alm e Arkin

Discutono sulla validità del modello di F&W che non tiene conto della **stechiometria** delle reazioni biochimiche

Scale – Free Networks

- Molti nodi di grado basso sono collegati a pochi nodi con grado alto
- una rete il cui grado di distribuzione può essere approssimato da una distribuzione power-law



Scale – Free Networks nelle RM

Barabasi et al mostrano come una RM sia una Small World Network

Critiche principali:

- Reti costruite su dati di scarsa qualità
- Validità metodologica
- Importanza dubbia del modello Scale-Free

Conclusioni sul grafo di una RM

- C'è il bisogno di trovare un **metodo di rappresentazione migliore** per poter esaminare e capire veramente le reti biologiche in genere
- Si necessita di **strumenti di misura specifici** per le RM che tengano maggiormente in conto gli aspetti biologici

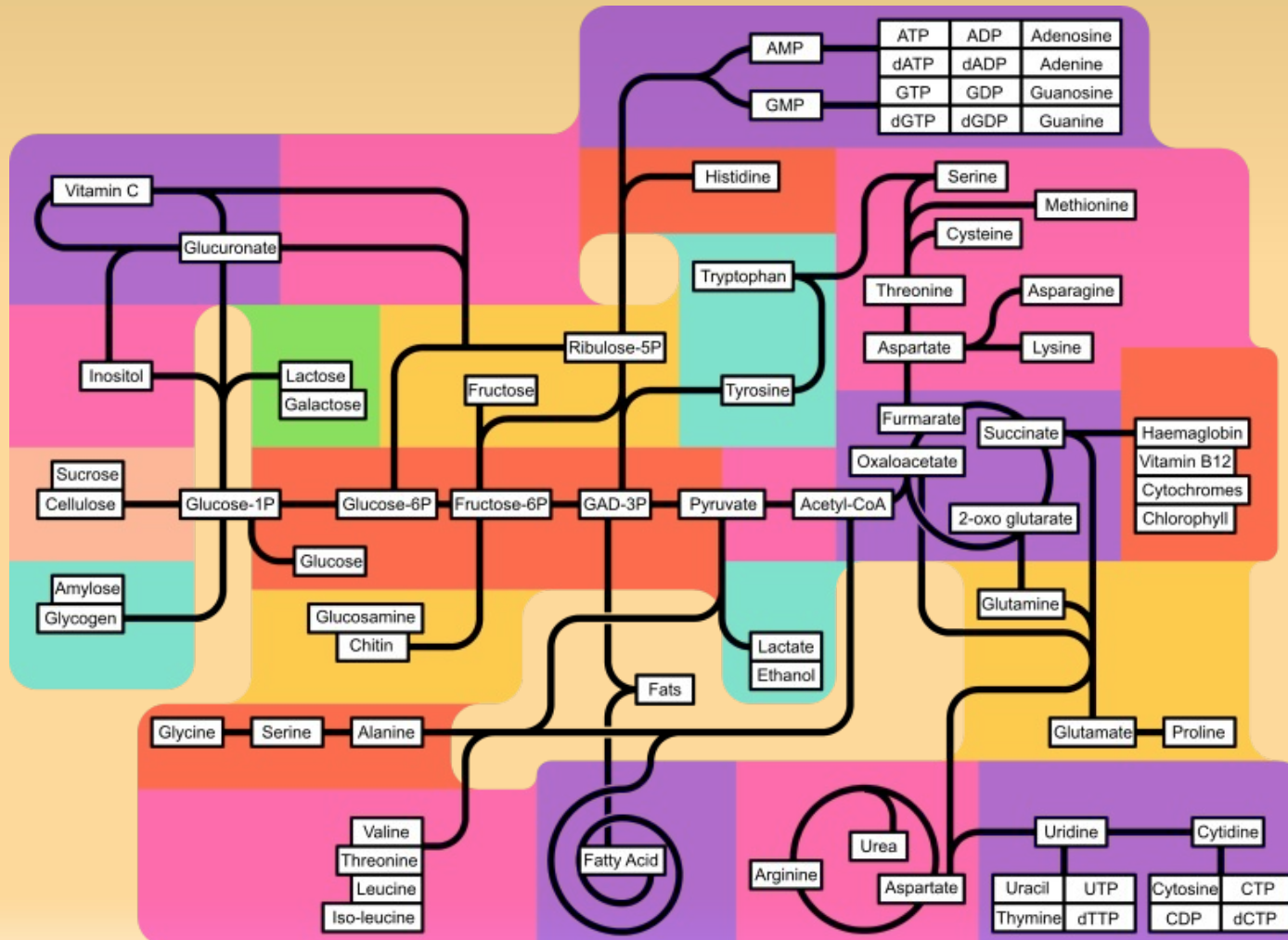
Modularità

Per le reti metaboliche sono stati utilizzati due metodi principali per identificare i moduli:

- Identificazione Top-Down
- Identificazione Bottom-Up

Identificazione Top Down (1)

- Pathways
 - E' il modo più naturale per definire i moduli



Identificazione Top Down (2)

- Il Concetto di Pathways

- Yamada et al

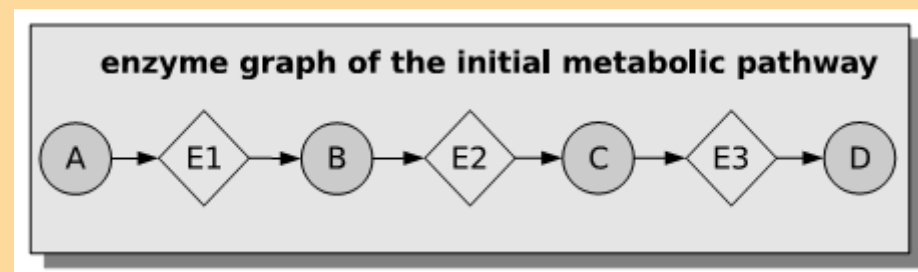
definiscono un "pathway module" come un insieme di enzimi aventi lo stesso profilo filogenetico

- Schuster e Hilgetag

formalizzano il concetto di "elementary mode" come l'insieme minimale di enzimi che operano insieme in condizione di equilibrio

Identificazione Top Down (3)

- E' l'insieme delle reazioni chimiche coinvolte in uno o più processi di anabolismo o catabolismo all'interno di una cellula
- I singoli passi della via metabolica sono reazioni catalizzate nella maggior parte dei casi da enzimi specifici che trasformano il substrato su cui agiscono, in un prodotto, utilizzato a sua volta come substrato dall'enzima del passo successivo.



Identificazione Top Down (4)

- Connectivity-Based Definitions

Identificazione dei moduli come gruppi di nodi altamente connessi tra loro e scarsamente connessi con gli altri moduli

Due due casi:

- Module Detections
- Network Decomposition

Module Detections

- Spirin e Mirny suggeriscono un metodo derivante delle "reti di interazione proteica"
 - Ogni modulo consiste in un sottografo di densità $Q(m, n) = 2m / (n(n-1))$
 - vengono usate: euristiche, tecniche di simulazione e approssimazione
 - Il problema di ricercare il sottografo più denso in una rete data è NP-Hard

Network Decomposition

Molti metodi restituiscono una **partizione del grafo** di partenza

- in generale il problema del partizionamento di un grafo è NP-Completo e il numero di sottografi in cui dividere la rete deve essere noto a priori
- Sono stati proposti altri metodi per i quali non è necessario conoscere il numero di sottografi
- Ottimizzano una "Funzione di Modularità"

Network Decomposition (1)

- *Funzione di Modularità* = somma per ogni modulo delle differenze tra il numero di archi interni al modulo osservati e quelli attesi in base ad un modello nullo
- Gli archi attesi sono $(k_i * k_j) / 2m$ dove k_i è il grado del nodo i e m il numero di archi nel grafo
- Sono stati presentati altri metodi basati sulle matrici stechiometriche (FBA, correlazione di Pearson)

Network Decomposition (2)

- Altri metodi "ad hoc" hanno rilevato che le RM presentano, in generale, una struttura modulare e gerarchica, nella quale i nuclei dei moduli base sono strettamente collegati tra loro, mentre i moduli che operano su funzioni più marginali o specializzate sono piuttosto isolati dal resto

Motif

- Nelle reti i motif possono essere visti come pattern di connessioni fra i nodi o come sottografi isomorfi nella rete
- La differenza tra i *motif* e i moduli è che i primi non necessitano di autonomia quanto di ripetizione nella rete
- Per le RM è stato introdotto il concetto di "motif colorato"

Identificazione Bottom-Up

- Snel e Huynem
 - discutono sulla convergenza tra i moduli funzionali e i moduli evolutivi
 - differenze tra moduli evolutivi e moduli funzionali
- Spirin et al
 - osservano che i pathways metabolici lineari sono favoriti dall'associazione genomica
 - vengono ereditati e regolati come singoli "mattoni" della RM

Conclusioni sulla Modularità

Concludendo: i Moduli strutturali, quelli funzionali e i moduli evolutivi sembrano essere tre concetti **scollegati tra loro**, dipendenti dal punto di vista con cui viene analizzato il metabolismo.

Scaletta

- Introduzione
- Analisi strutturale
- Analisi topologica
- **L'Evoluzione del Metabolismo**
 - Modelli di Evoluzione delle RM
 - Analisi comparativa per dedurre degli scenari evolutivi

Modelli di Evoluzione delle RM

- Modelli Biologici
- Modelli Computazionali

Modelli Biologici

- In quale ordine sono comparsi gli enzimi?
- Come si sono formati?

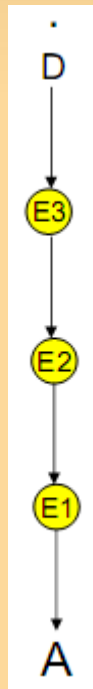
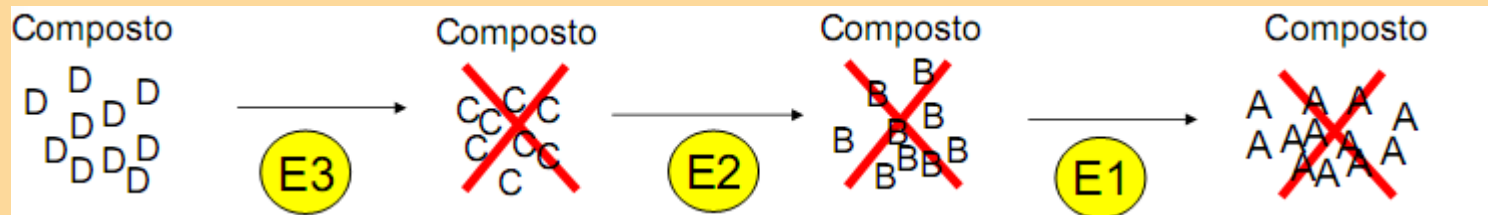
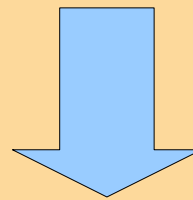
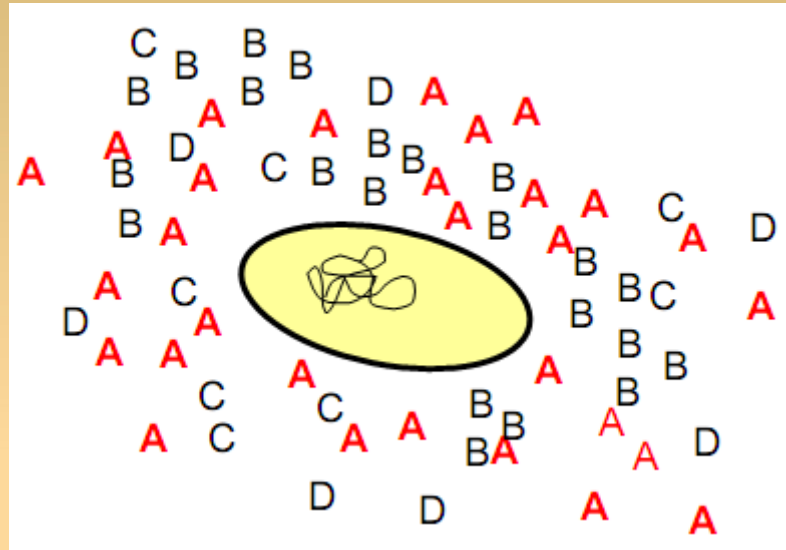
Soluzioni principali proposte:

- Horowitz (1945) - Ipotesi Retrograda
- Ycas (1974) e Jensen (1976) - Patchwork

Ipotesi Retrograda (1)

- Si basa sull'idea di “brodo primordiale”
- La via metabolica sarebbe stata costruita come **una scala**, per acquisizione successiva e sequenziale di enzimi in ordine inverso (retrograda)
- Tutti i geni di un operone si sarebbero originati per **uplicazione successiva** di un unico gene ancestrale comune.

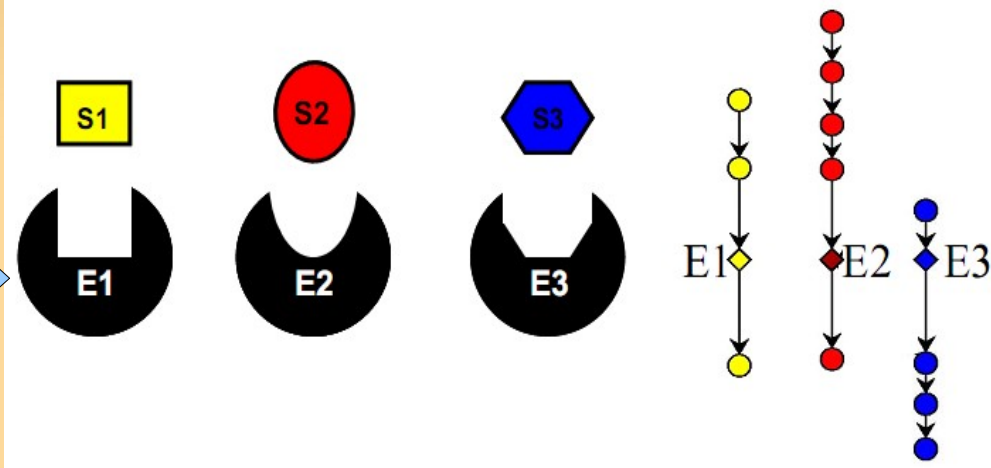
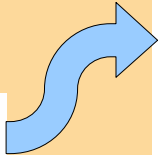
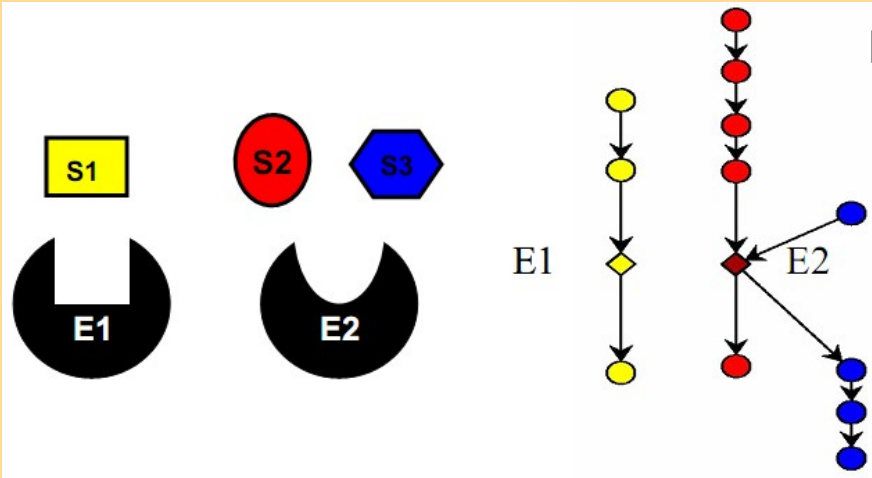
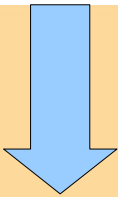
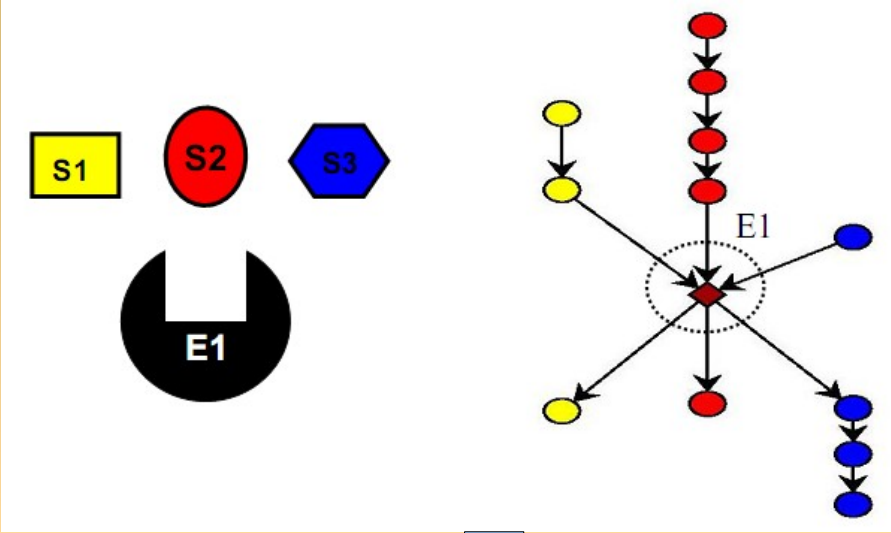
Ipotesi Retrograda (2)



Patchwork (1)

- Enzimi primitivi catalizzano molte reazioni enzimatiche, per quanto poco veloci ed efficienti
- Un enzima (E1) sarebbe stato capace di interagire con substrati diversi (S1, S2, S3)
- Duplicazioni dei geni codificanti avrebbero portato all'evoluzione di enzimi con **aumentata specificità di substrato**, entrati poi a far parte di **vie metaboliche diverse**

Patchwork (2)



Critiche a Horowitz e Jensen (1)

Differenze tra i due modelli (Dìaz-Mejìda):

- secondo Horowitz, gli enzimi prodotti catalizzerebbero **reazioni consecutive**
- per Jensen, gli enzimi specializzati generati tramite Patchwork catalizzerebbero **reazioni distanti** all'interno del pathway
- per il modello retrogrado gli enzimi avrebbero la proprietà di catalizzare **reazioni diverse** mantenendo substrati specifici
- per il patchwork enzimi diversi catalizzano **reazioni chimiche simili** pur partendo da substrati differenti

Critiche a Horowitz e Jensen (2)

Per Rison e Thornton l'evoluzione retrograda:

- non avrebbe potuto funzionare in un ambiente povero di metaboliti intermedi
- non spiega la presenza di **metaboliti instabili** nei pathways (Roy ipotizza la possibilità di una **evoluzione a salti** con enzimi multifunzionali)

Modelli Computazionali (1)

- Al contrario dei modelli biologici i modelli computazionali non devono avere dei meccanismi biologicamente fondati
- Sono considerati validi se portano alla costruzione di un grafo che abbia le proprietà osservate nelle reti biologiche

Modelli Computazionali (2)

Barabasi e Albert:

- definiscono le regole per un modello computazionale che porti alla costruzione di una **rete *power-law***
- Alcuni studi mettono in dubbio questo modello: la proprietà delle reti biologiche di essere scale-free non è ancora formalmente dimostrata

Modelli Computazionali (3)

Pfeiffer et al:

- Le loro simulazioni si basano sull'assunzione di enzimi multifunzionali con ampie, ma poco efficienti, possibilità di catalizzazione
- Le simulazioni hanno messo in evidenza:
 - **l'emergere di hub**
 - **perdita dei metaboliti intermedi** e di alcune reazioni bio-chimiche

Analisi comparativa per dedurre scenari evolutivi

- Analisi comparativa per dedurre degli scenari evolutivi
 - Allineamenti dei pathway
 - Metodi di deduzione degli scenari evolutivi
- Per ricostruire gli scenari evolutivi dei pathway e delle RM vengono spesso utilizzate le stesse tecniche utili al **confronto fra sequenze**

Distanza fra pathway

Proposte varie tecniche:

- **Enzymatic Classification** come base per la misura di pathway con struttura lineare
- Decomposizione dei grafi in **cammini**
- Distanza di **Hamming** fra le matrici di adiacenza dei grafi
- Sono stati proposti anche metodi che **trascurano la topologia** della rete e che considerano le RM come insiemi di enzimi e metaboliti o con vari indici strutturali

Metodi di deduzione degli scenari evolutivi

- Ricostruire l'ordine di apparizione dei pathway metabolici, modellandoli come delle **sequenze di assenze/presenze di enzimi** senza considerarne la struttura
- Modellare i pathways come **insiemi di enzimi** costruendo un albero filogenetico sulla base di una distanza che tiene conto sia della similarità fra due enzimi che della similarità fra gli *enzimi vicini* a quelli presi in considerazione
- Calcolare la distanza introducendo penalità nel caso in cui non ci fossero **enzimi omologhi** tra pathway differenti

Conclusioni (1)

- La struttura delle reti metaboliche spesso fornisce **informazioni funzionali ed evolutive** ma non sempre
- Misure semplici basate solo sulla struttura della rete, come il suo diametro, forniscono informazioni molto limitate
- Concetti più complessi come i **Motif** o i **Moduli** rappresentano meglio le caratteristiche biologiche ma spesso **mancano di una validazione biologica**

Conclusioni (2)

- Lo sviluppo di un modello di **Random Graph** che sia **realistico** per il metabolismo è ancora in atto
- La ricerca in questo campo dovrebbe dirigersi verso:
 - la realizzazione di un modello realistico per le RM
 - lo sviluppo di tecniche evolute di analisi sugli ipergrafi

“ put the subject by "the rest next time-" “

“ "Explain all that," said the Mock Turtle

"No, no! The adventures first," said the Gryphon “

(Alice nel paese delle meraviglie, Lewis Carroll)

Bibliografia (1)

- "An introduction to metabolic networks and their structural analysis" (Lacroix et al.)
- it.wikipedia.org/wiki/Santorio_Santorio
- en.wikipedia.org/wiki/Louis_Pasteur
- en.wikipedia.org/wiki/Eduard_Buchner
- "Algorithms for graph partitioning: A survey" (P. Fjallstrom)
- www.scienceandsociety.co.uk/results.asp
- it.wikipedia.org/wiki/Riducente
- [it.wikipedia.org/wiki/Cofattore_\(biologia\)](http://it.wikipedia.org/wiki/Cofattore_(biologia))
- it.wikipedia.org/wiki/Genomica
- it.wikipedia.org/wiki/Metabolomica

Bibliografia (2)

- www.geneontology.org/
- www.chem.qmul.ac.uk/iubmb/enzyme/
- www.chem.qmul.ac.uk/iubmb/enzyme/rules.html
- it.wikipedia.org/wiki/Grafo
- it.wikipedia.org/wiki/Grafo_bipartito
- en.wikipedia.org/wiki/Hypergraph
- "How scale-free are biological networks" (R. Khanin and E. Wit)
- Representing Reconstructed Networks Mathematically: The Stoichiometric Matrix da systemsbiology.ucsd.edu
- it.wikipedia.org/wiki/Carl_Adam_Petri
- it.wikipedia.org/wiki/Rete_di_Petri

Bibliografia (3)

- "Origine ed evoluzione delle vie metaboliche" di Renato Fani e Matteo Brilli
- www.anisnfirenze.it/Metabolismo_Primordiale
- www.researchgate.net/publication/Use_of_coefficients_of_enzymatic_activity
- www.cnr.it/istituti/ArticoliJCR.html?cds=Enzymatic_Coefficient
- "Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions" (M. Kotera et al.)
- it.wikipedia.org/wiki/Indice_di_correlazione_di_Pearson
- www.disi.unige.it/person/DellamicoM/RetiComplesse
- en.wikipedia.org/wiki/Orthology#Orthology
- www.genome.jp/kegg/

Bibliografia (4)

- "A bioinformatician's view of the metabolome" (I. Nobeli and J. M. Thornton)
- biocyc.org/