

University of Pavia
Dep. of Electrical, Computer and Biomedical Engineering

Analisi di reti biologiche basate sull'integrazione di dati e conoscenza

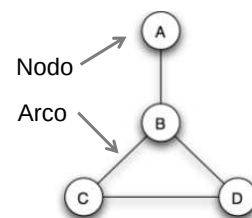
Francesca Vitali, Elisabetta Sauta, Riccardo Bellazzi



Corso di Bioinformatica e Biologia Sintetica, 11/12/15

Che cos'è una rete?

- Per rete si intende comunemente **una serie di componenti, sistemi o entità interconnessi tra di loro**
- Un grafo consiste di
 - Un insieme di **nodi** (vertici)
 - Un insieme di **archi** (collegamenti)
 - Un arco collega due nodi
- Due nodi sono **vicini** (neighbor) se sono collegati da un arco
 - I nodi C e D sono adiacenti all'arco (C, D)



Il ruolo delle reti

Dietro ogni sistema complesso c'è una **rete** che ne definisce le interazioni tra le componenti.

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

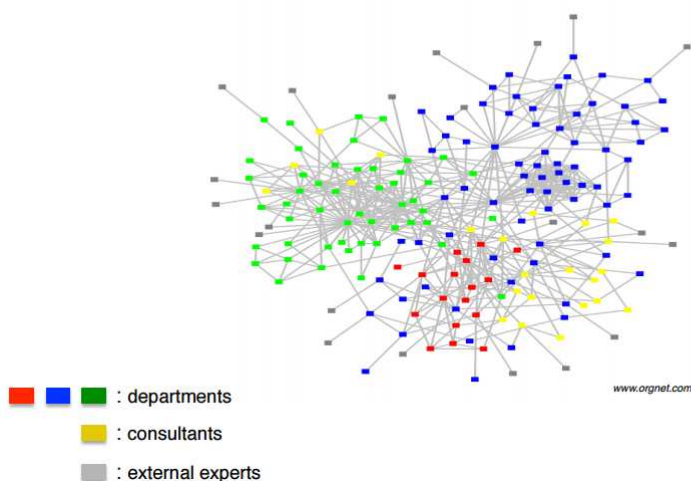


Facebook

Il grafo alla base di Facebook



Struttura di un'organizzazione

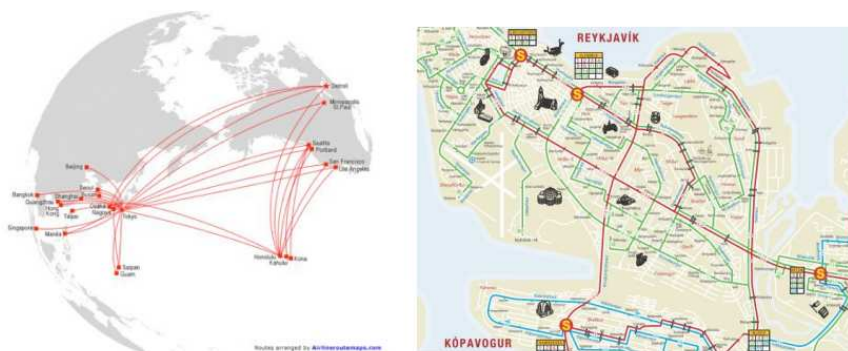


Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Reti di trasporto



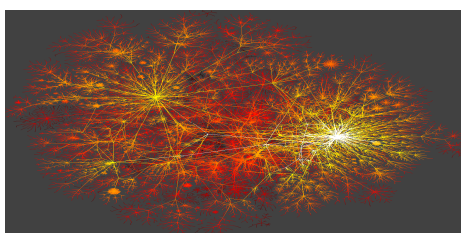
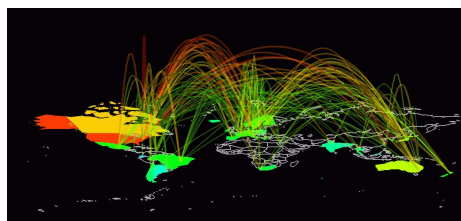
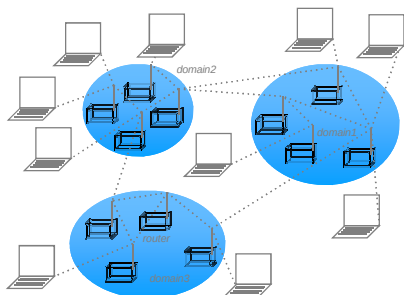
- La terminologia dei grafi è in gran parte derivata dal mondo dei trasporti
 - "cammino minimo", "diametro", "flusso"
- Questioni a cui siamo interessati
 - La struttura della rete è in grado di supportare le performance richieste? Quanto è robusta? Si presta a fallimenti a cascata?

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Internet



- Ideare motori di ricerca versatili
- Costruire algoritmi per il filtraggio automatico o la classificazione di documenti e dati

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Cervello

**Human Brain
has between
10-100 billion
neurons.**

Francesca Vitali & Elisabetta Sauta

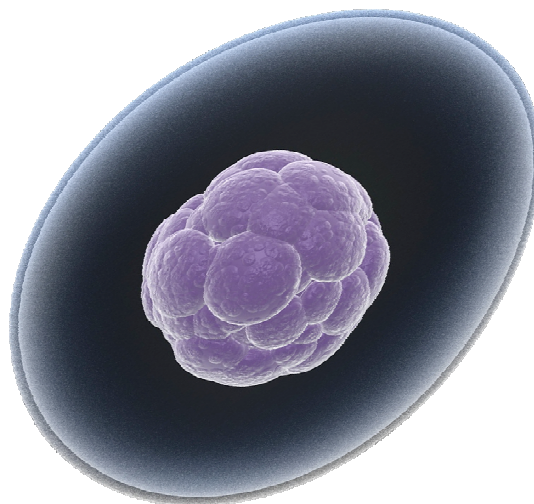
Corso di Bioinformatica, 11/12/2015



Genoma umano

Quanti geni ci sono
nel genoma umano?

~26 000

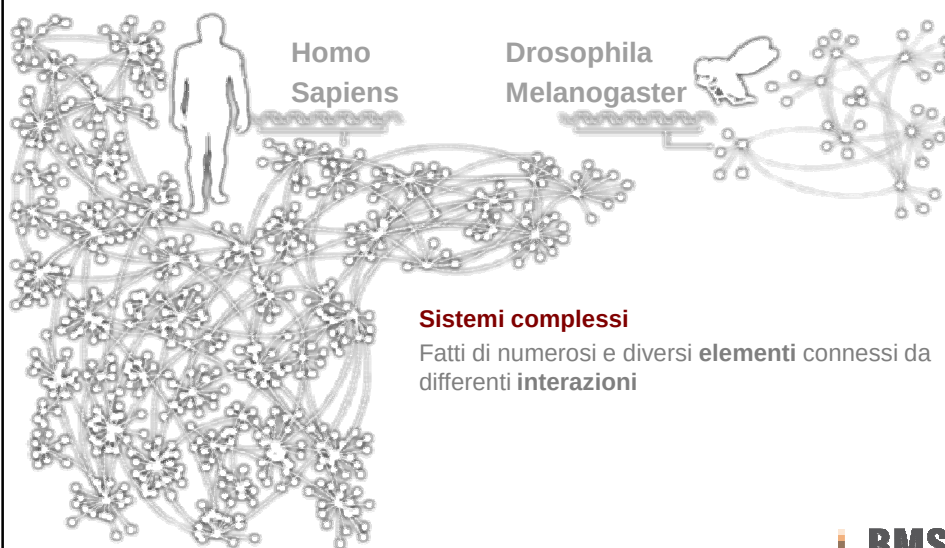


Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Reti Biologiche



Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Dietro ad ogni sistema complesso c'è un intricato diagramma , o una rete, che ne definisce le interazioni tra le componenti.

Non riusciremo mai a comprendere un sistema complesso senza analizzarne la reti che ne stanno alla base.

- Albert-László Barabási

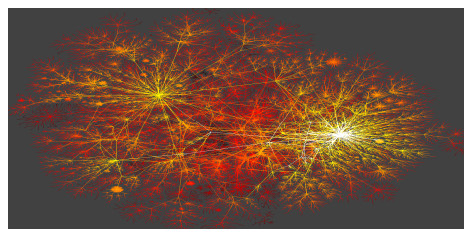
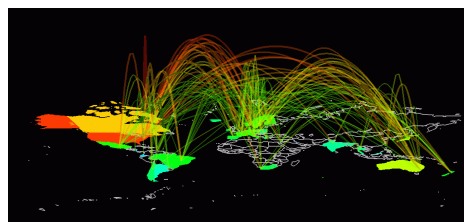
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Analizzare le reti

- Come si inizia ad analizzarle?
- Come si fa a capirle?
- Come si estraggono le loro proprietà?
- Quali sono le proprietà che ci interessano?



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



La storia delle reti

- *Eulero*: **TEORIA DEI GRAFI** 1735
- *Erdos-Renyi*: **RETI RANDOM** ~1960
- *Albert-László Barabási*: **SCALE-FREE NETWORK** ~1999

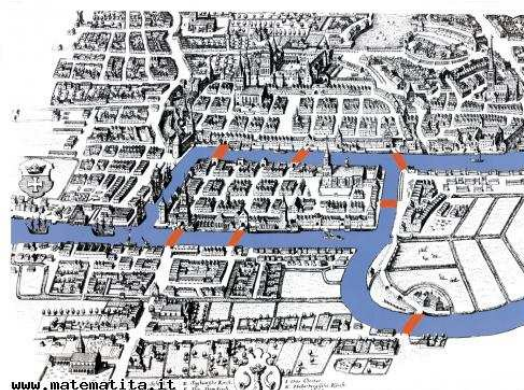
Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



TEORIA DEI GRAFI

- Lo studio delle reti fa parte della cosiddetta *matematica discreta* e si basa sulla **TEORIA DEI GRAFI** introdotta da **Eulero** nel 1736
 - **Problema dei ponti di Königsberg**



www.matematita.it

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015

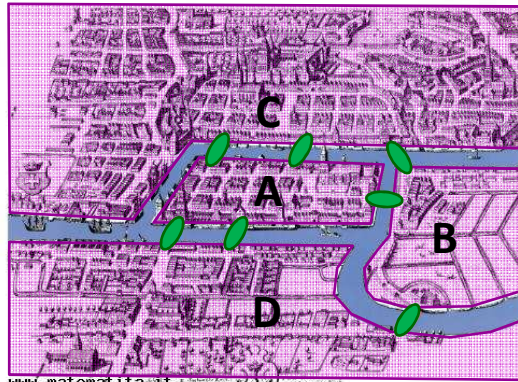


Teoria dei grafi

PROBLEMA

Come in ogni cittadina, era uso anche a Königsberg che alla domenica gli abitanti facessero la loro passeggiata per le vie della città;

E' possibile progettare un percorso che permetta a ciascuno di partire da casa e tornarvi dopo aver attraversato ciascun ponte una ed una sola volta?



www.matematika.it

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Teoria dei grafi

PROBLEMA

Come in ogni cittadina tedesca, era uso anche a Königsberg che alla domenica gli abitanti facessero la loro passeggiata per le vie della città;

è possibile progettare un percorso che permettesse a ciascuno di partire da casa e tornarvi dopo aver attraversato ciascun ponte una ed una sola volta?

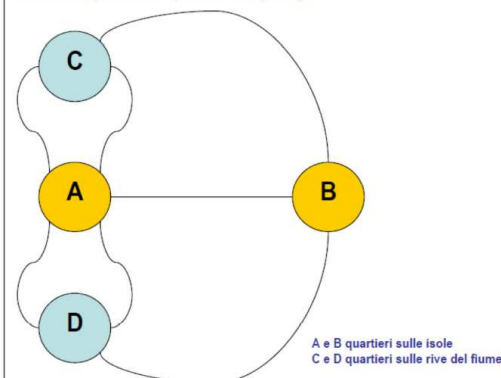
RAPPRESENTAZIONE ATTRAVERSO
STRUTTURA MATEMATICA

GRAFO

Condizione necessaria e sufficiente

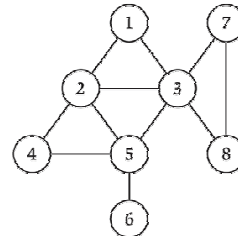
- DEVE ESISTERE UN PERCORSO FRA OGNI COPPIA DI NODI
- OGNI NODO DEVE ESSERE TOCCATO DA UN NUMERO PARI DI ARCHI

Schema del problema "I ponti di Königsberg"



Teoria dei grafi

- Un grafo è una coppia di insiemi
 - $G = (V, E)$
- Rappresentazioni più utilizzate
 - **Rappresentazione grafica**
 - **Matrice di adiacenza**
 - Matrice $n \times n$ (n numero dei vertici in V)
 - Elemento (i, j) vale 1 se esiste l'arco tra i e j
 - **Lista dei vertici V e degli archi E**
- La **Teoria dei Grafi** fornisce un **linguaggio unificante per descrivere la struttura di una network**



Matrice di adiacenza

$$V: \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$E: \{2,5\},$

$n: \{3\},$

$m: \{2,5\},$

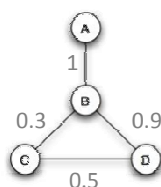
Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015

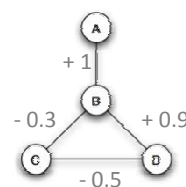


Grafi pesati e grafi segnati

- Possiamo associare agli archi delle informazioni aggiuntive
 - segno (amici-nemici)
 - forza del legame (forza della relazione sociale)
 - distanza (lunghezza del collegamento)
 - ritardo (tempo di trasmissione)
 - affidabilità (percentuale di errore nella trasmissione)
 - costo (costo di utilizzo del collegamento)



Grafo pesato



Grafo segnato

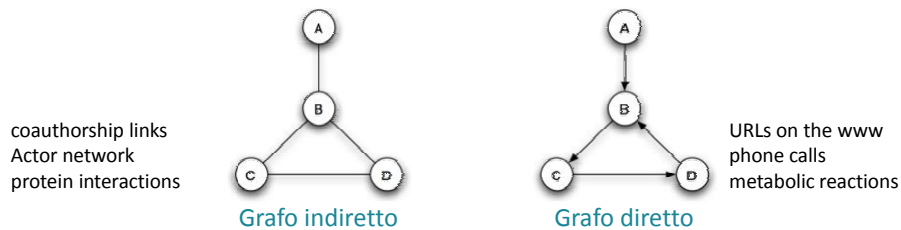
Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Grafi diretti e indiretti

- Gli archi possono avere un **orientamento**
 - Archi **indiretti**: relazione vale in tutte e due le direzioni
 - Archi **diretti**: relazione vale solo tra testa e coda dell'arco



- La differenza è sostanziale
 - Modelli di formazione e mantenimento della rete differenti
 - Algoritmi differenti

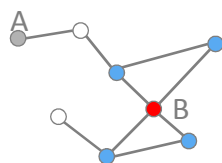


Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

NODE DEGREE - GRADO DEL NODO

Indiretti

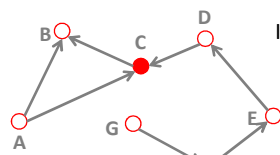


Node degree: Numero di link connessi ad un nodo.

$$k_A = 1$$

$$k_B = 4$$

Diretti

Nelle reti dirette possiamo un **in-degree** e un **out-degree**.

Il grado (totale) è la somma dell'in- e out- degree.

$$k_c^{in} = 2 \quad k_c^{out} = 1 \quad k_c = 3$$

Source: nodo con $k^{in} = 0$;**Sink:** nodo con $k^{out} = 0$.

Francesca Vitali & Elisabetta Sauta

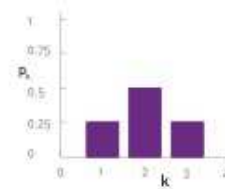
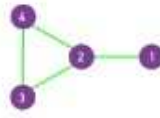
Corso di Bioinformatica, 11/12/2015

Distribuzione del grado

- Degree distribution**

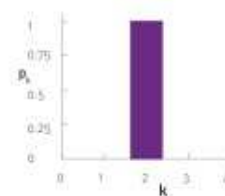
$P(k)$

Probabilità che un nodo scelto a caso abbia grado k



$N_k = \# \text{ nodi con grado } k$

$$P(k) = \frac{N_k}{N}$$



Francesca Vitali & Elisabetta Sauta

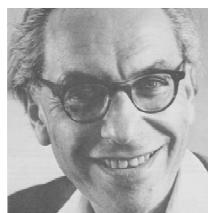
Corso di Bioinformatica, 11/12/2015



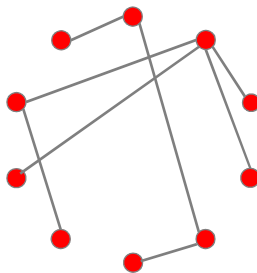
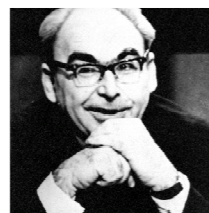
RANDOM NETWORK MODEL

COME SI
DESCRIVONO
RETI MOLTO
COMPLICATE
?

Pál Erdős
(1913-1996)



Alfréd Rényi
(1921-1970)



Erdős-Rényi model (1960)

Probabilità di connessione p

$P=1/6$ $N=10$

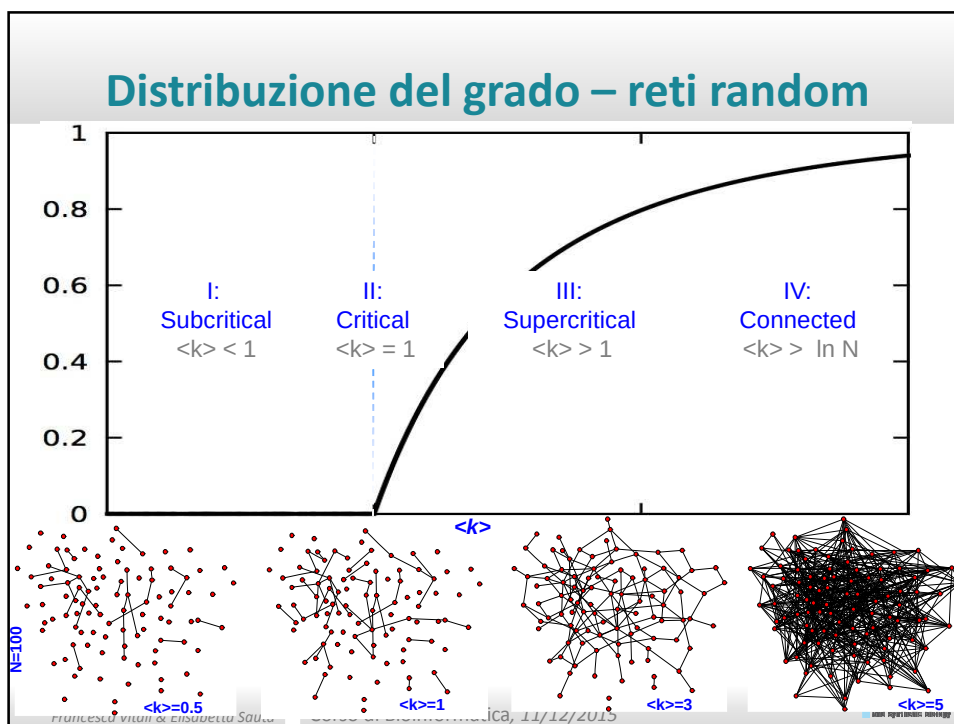
$\langle k \rangle \sim 1.5$

<https://www.youtube.com/watch?v=mpe44sTSof8>

Francesca Vitali & Elisabetta Sauta

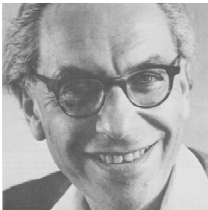
Corso di Bioinformatica, 11/12/2015





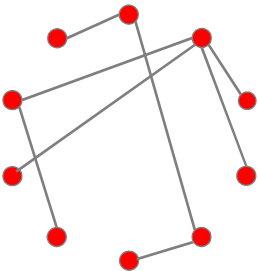
RANDOM NETWORK MODEL

Pál Erdős
(1913-1996)



Alfréd Rényi
(1921-1970)





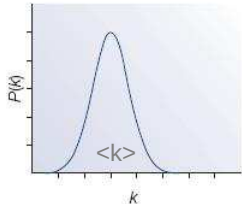
Erdős-Rényi model (1960)

Probabilità di connessione p

$P=1/6$ $N=10$

$\langle k \rangle \sim 1.5$

Distribuzione del grado
distribuzione di Poisson



Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015

Le reti reali sono random?

Le reti reali non sono random

Una società random sarebbe composta da un numero di individui con più o meno lo stesso numero di amici

- Non ci sarebbero outliers: persone più o meno popolari

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



SCALE - FREE NETWORK

Albert-László Barabási



- Porta l'attenzione a **ripensare le reti**
- Ripercorre le tappe nella storia della scienza delle reti
- Definisce un nuovo modello teorico:
 - RETI AD INVARIANZA DI SCALA
(*scale-free network*)

Many real world networks have a similar architecture: WWW, Internet (routers and domains), electronic circuits, computer software, movie actors, coauthorship networks, sexual web, instant messaging, email web, citations, phone calls, metabolic, protein interaction, protein domains, brain function web, linguistic networks, comic book characters, international trade, bank system, encryption trust net, energy landscapes, earthquakes, astrophysical network...

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



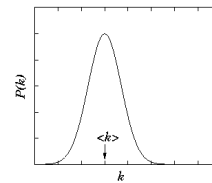
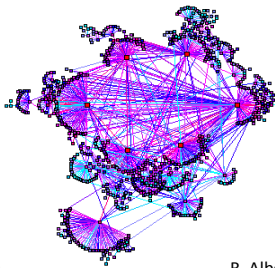
World Wide Web

Nodes: **WWW documents**

Links: **URL links**

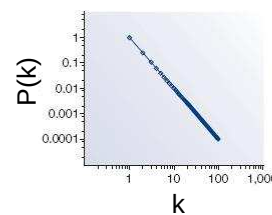
Over 3 billion documents

ROBOT: collects all URL's found in a document and follows them recursively



Expected

$$P(k) = \sim k^\gamma$$



Found

R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

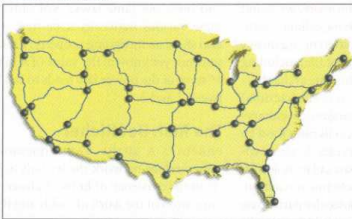
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

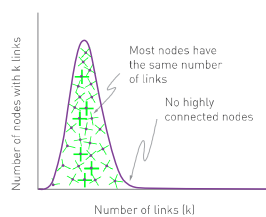


Altri esempi

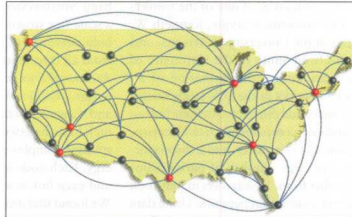
Random Network



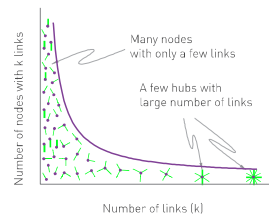
(a) POISSON



Scale-Free Network



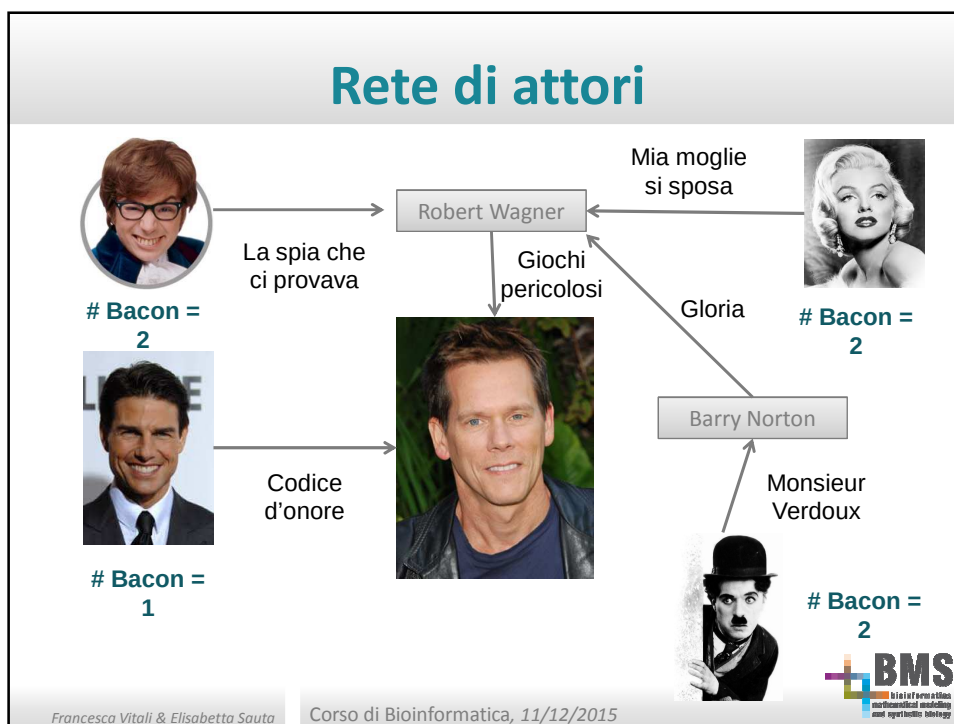
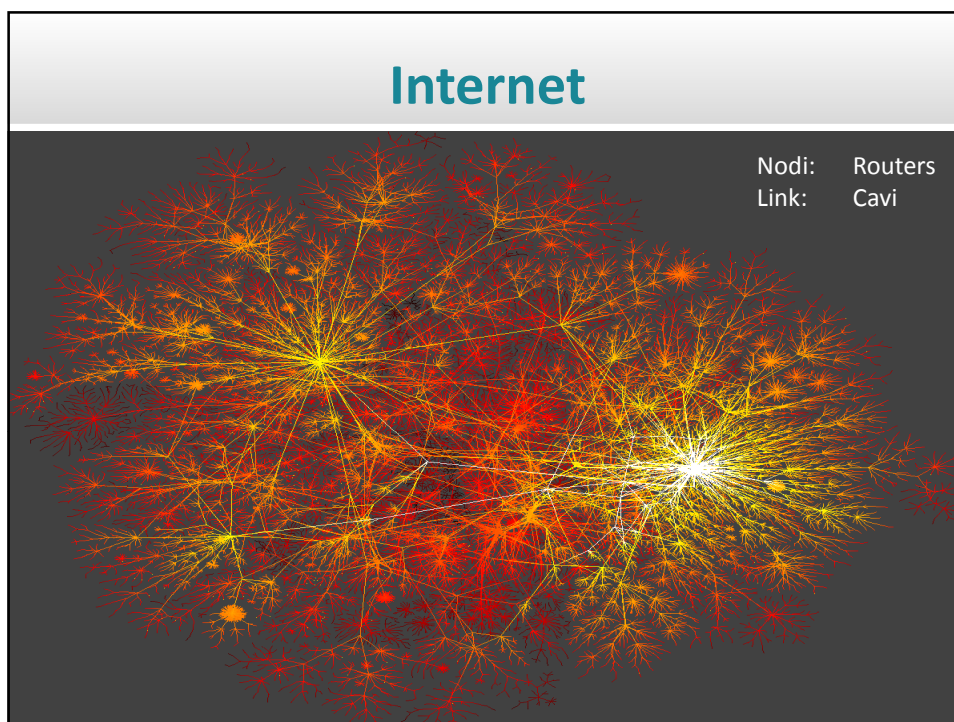
(c) POWER LAW



Francesca Vitali & Elisabetta Sauta

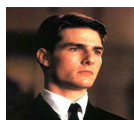
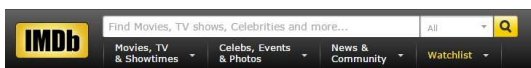
Corso di Bioinformatica, 11/12/2015





Rete di Hollywood

Nodes: **Attori**
Links: **se hanno recitato insieme**



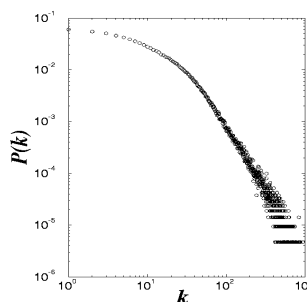
Days of Thunder (1990)
Far and Away (1992)
Eyes Wide Shut (1999)



N = 212.250
attori

$\langle k \rangle = 28,78$

$$P(k) = \sim k^\gamma$$

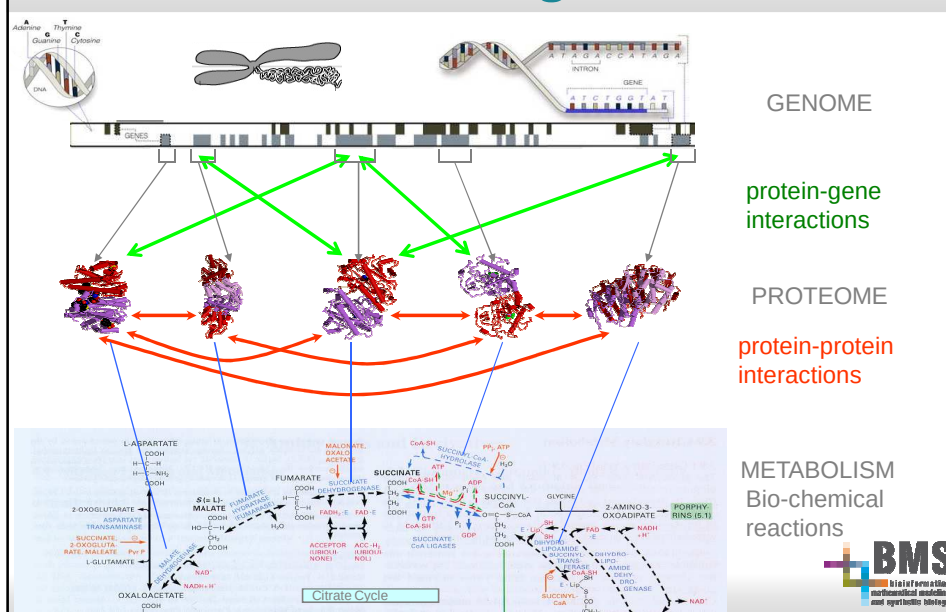


Francesca Vitali & Elisabetta Sautu

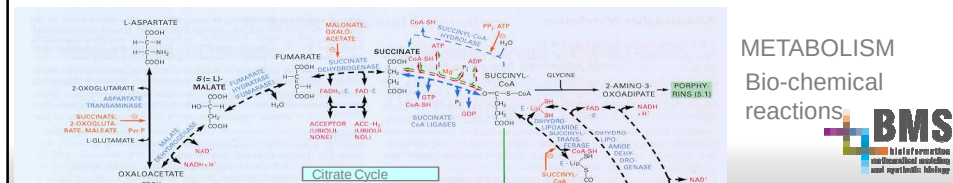
Corso di Bioinformatica



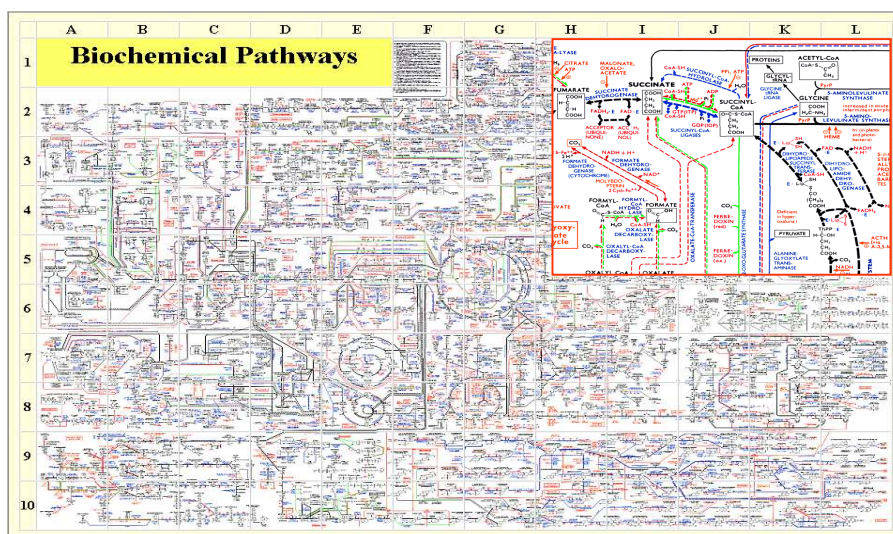
Reti biologiche



Reti biologiche

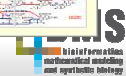


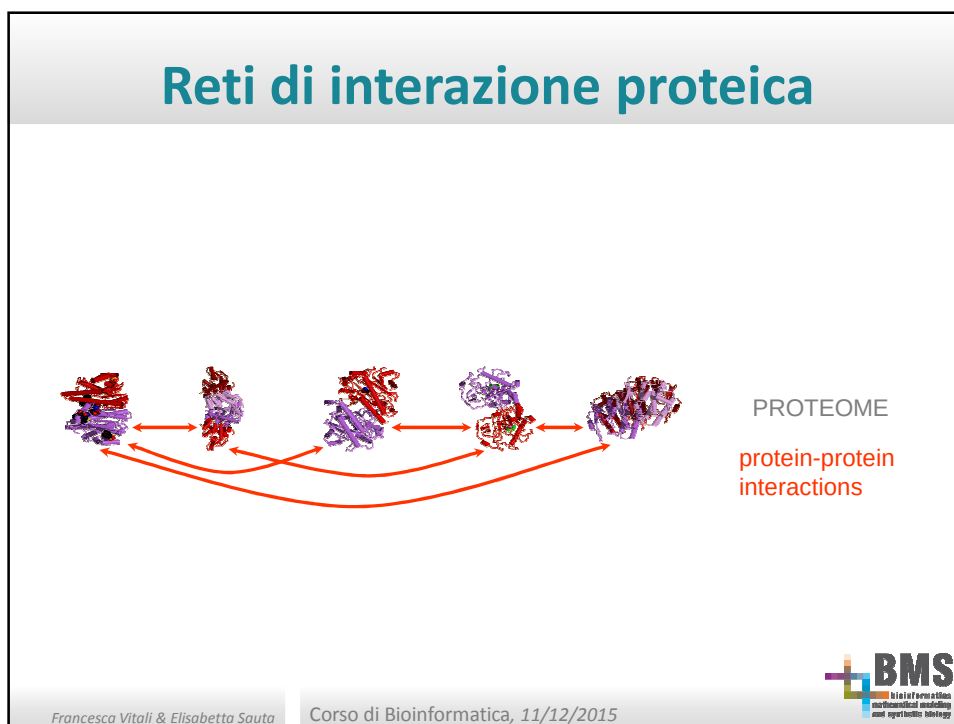
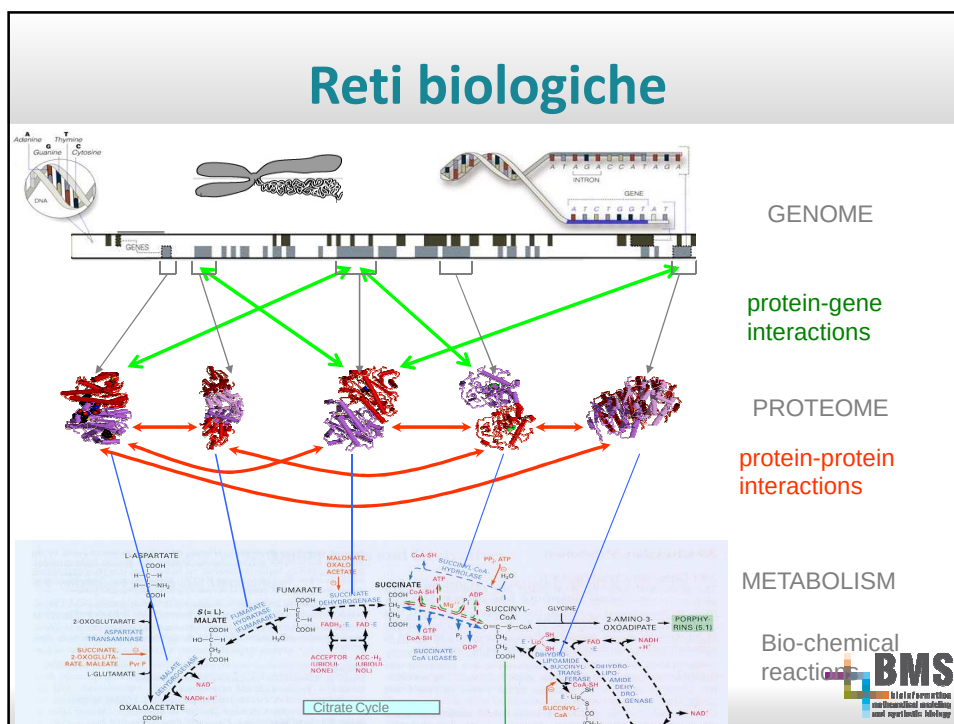
Reti Metaboliche



Francesca Vitali & Elisabetta Sauta

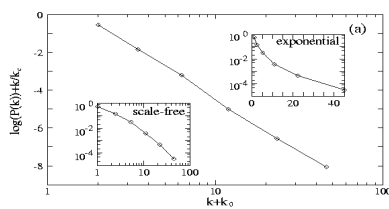
Corso di Bioinformatica, 11/12/2015



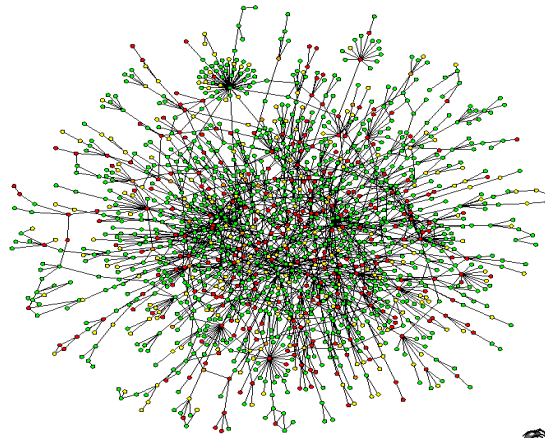


Reti di interazione proteica

Nodes: **Proteine**
Links: **interazioni fisiche di binding**



$$P(k) \sim (k + k_0)^{-\gamma} \exp\left(-\frac{k + k_0}{k_t}\right)$$



Flack

BMS
Biology of Metabolism
and Systems
and Systems Biology

H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Nature 411, 41-42 (2001)

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

Scale – Free Networks

Examples of Scale-Free Networks

NETWORK	NODES	LINKS
Cellular metabolism	Molecules involved in burning food for energy	Participation in the same biochemical reaction
Hollywood	Actors	Appearance in the same movie
Internet	Routers	Optical and other physical connections
Protein regulatory network	Proteins that help to regulate a cell's activities	Interactions among proteins
Research collaborations	Scientists	Co-authorship of papers
Sexual relationships	People	Sexual contact
World Wide Web	Web pages	URLs

HANNO TUTTE LA STESSA STRUTTURA

BMS
Biology of Metabolism
and Systems
and Systems Biology

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

Proprietà delle reti

- **Scale – Free:** Pochi nodi con grado piccolo connessi da hub.
- **Small – World:** distanza tra due nodi qualsiasi è piccola.
- **Evoluzione:** la presenza di hub è dovuta all'evoluzione ed alla proprietà detta di «attacco preferenziale».
- **Robustezza:** Capacità di recupero ad attacchi random; fragilità ad attacchi mirati (Tallone d'Achille)
- **Comunità:** Gruppi di nodi hanno comportamenti simili

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Proprietà delle reti

- **Scale – Free:** Pochi nodi con grado piccolo connessi da hub.
- **Small – World:** distanza tra due nodi qualsiasi è piccola.
- **Evoluzione:** la presenza di hub è dovuta all'evoluzione ed alla proprietà detta di «attacco preferenziale».
- **Robustezza:** Capacità di recupero ad attacchi random; fragilità ad attacchi mirati (Tallone d'Achille)
- **Comunità:** Gruppi di nodi hanno comportamenti simili

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Organizing principles of complex networks

- **Scale – Free:** Many small nodes held together by a few hubs.
- **Small – World:** Short paths between any two nodes.
- **Evolutions:** Hubs emerge through growth and preferential attachment.
- **Robustness:** Resilience against random errors: fragility to attacks (Achilles' Heel)
- **Communities:** Groups behave similarly

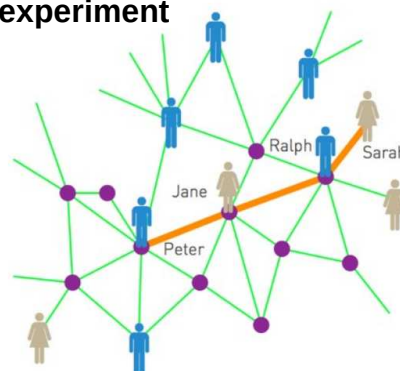
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Teoria del piccolo mondo

Milgram experiment



Ogni persona può essere collegata a qualunque altra persona o cosa attraverso una catena di conoscenze e relazioni con **non più di 5 intermediari**

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Le reti reali hanno la stessa struttura

Topologia SCALE - FREE

WWW, Internet (routers and domains), electronic circuits, computer software, movie actors, coauthorship networks, sexual web, instant messaging, email web, citations, phone calls, metabolic, protein interaction, protein domains, brain function web, linguistic networks, comic book characters, international trade, bank system, encryption trust net, energy landscapes, earthquakes, astrophysical network...

COME E' POSSIBILE?

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

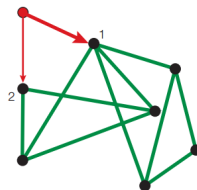


Origine delle reti reali

Reti random: Connessioni random con numero di nodi iniziali fisso

Reti reali:

1. Crescono → **nuovi nodi** si aggiungono
www: nuovi documenti
2. Nuovi nodi si connettono preferibilmente a nodi già molto connessi (*preferential attachment*)
www: connessioni a siti noti



Rich get richer

Scale – Free topology

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Organizing principles of complex networks

- **Scale – Free:** Many small nodes held together by a few hubs.
- **Small – World:** Short paths between any two nodes.
- **Evolutions:** Hubs emerge through growth and preferential attachment.
- **Robustness:** Resilience against random errors: fragility to attacks (Achilles' Heel)
- **Communities:** Groups behave similarly

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Proprietà delle reti

- **Scale – Free:** Pochi nodi con grado piccolo connessi da hub.
- **Small – World:** distanza tra due nodi qualsiasi è piccola.
- **Evoluzione:** la presenza di hub è dovuta all'evoluzione ed alla proprietà detta di «attacco preferenziale».
- **Robustezza:** Capacità di recupero ad attacchi random; fragilità ad attacchi mirati (Tallone d'Achille)
- **Comunità:** Gruppi di nodi hanno comportamenti simili

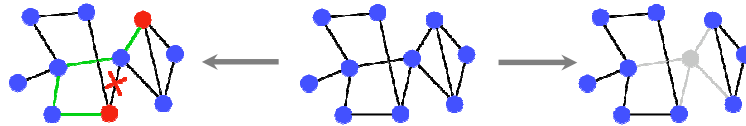
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Robustezza di una rete

Robustezza = Quanto una rete cambia (o non cambia) se vengono rimossi dei nodi o dei link



I sistemi complessi sono in grado di mantenere le loro funzioni base in caso di errori o difetti

Cellula → Mutazioni

Esistono numerose mutazioni ed altri errori nelle nostre cellule che non hanno alcuna conseguenza sul funzionamento delle stesse

Internet → router breakdowns

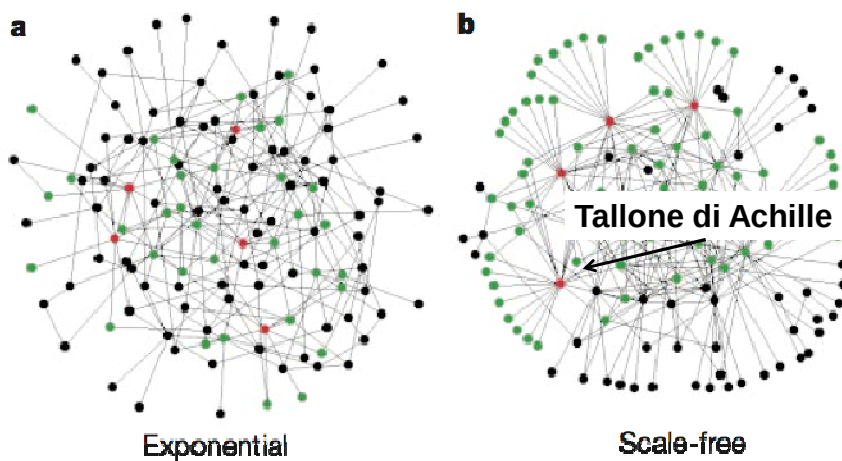
Anche se nello stesso momento si dovessero rompere centinaia di routers, Internet non smette di funzionare.

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Random vs Scale-free



Exponential

130 nodi, 215 link

I **top 5** nodi con più alto grado k sono connessi al...
... 27% della rete

Scale-free

... 60% della rete

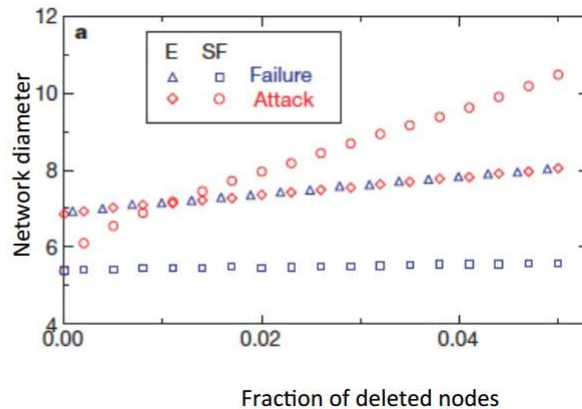
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Let's look at the blue lines

- Random graph: increasing monotonically
- SF: remains unchanged until at least 5%



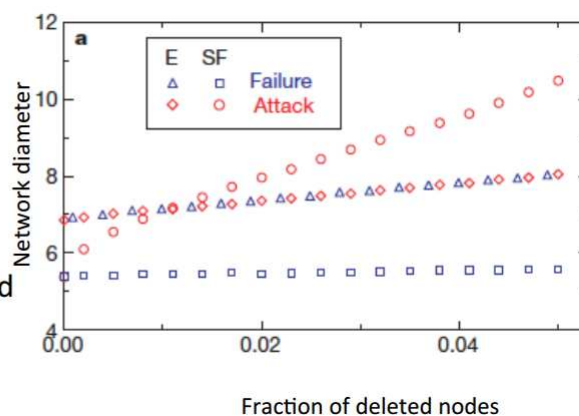
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Let's look at the red lines

- Random graph: same behaviour if nodes with most links are chosen first
- SF: with 5% nodes removed the diameter is doubled



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Organizing principles of complex networks

- **Scale – Free:** Many small nodes held together by a few hubs.
- **Small – World:** Short paths between any two nodes.
- **Evolutions:** Hubs emerge through growth and preferential attachment.
- **Robustness:** Resilience against random errors: fragility to attacks (Achilles' Heel)
- **Communities:** Groups behave similarly

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Proprietà delle reti

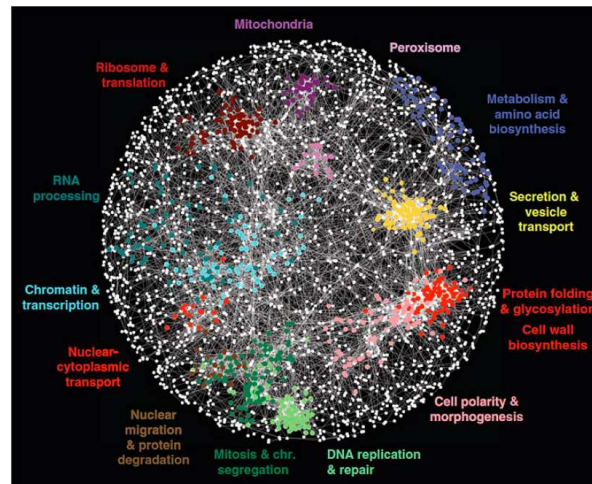
- **Scale – Free:** Pochi nodi con grado piccolo connessi da hub.
- **Small – World:** distanza tra due nodi qualsiasi è piccola.
- **Evoluzione:** la presenza di hub è dovuta all'evoluzione ed alla proprietà detta di «attacco preferenziale».
- **Robustezza:** Capacità di recupero ad attacchi random; fragilità ad attacchi mirati (Tallone d'Achille)
- **Comunità:** Gruppi di nodi hanno comportamenti simili

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Comunità



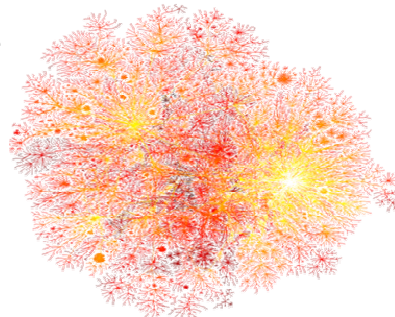
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Le reti: oggi

- Lo **sviluppo tecnologico** e la capacità di collezionare più dati ha portato alla comparsa di reti più grandi e complesse
- Tali reti mostrano caratteristiche non intuitive e possono arrivare ad essere costituite da **milioni di unità comunicanti fra loro**
- Il risultato è che non è possibile studiare le reti complesse mediante semplice ispezione visiva
- I metodi matematici basati sulla **teoria dei grafi** vengono dunque impiegati per estrarre le informazioni dalle reti complesse in maniera sintetica ed oggettiva



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Reti Biologiche

- Disponibilità di dati biologici → Esplosione
 - High-throughput data, microarrays, ...
- Goal:
 - **Comprensione** a livello di sistema dei sistemi biologici (organismi)
 - **Studio** delle interazioni tra componenti in modo da individuare **proprietà emergenti**
 - proprietà che esistono solo considerando insiemi di componenti e che non sono riscontrabili in nessun singolo componente
 - **Inferenza** in reti biologiche: estrazione di nuova conoscenza biologica dalle caratteristiche topologiche delle reti biologiche

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Esempi di reti biologiche

	Nodi	Link
Reti di interazione proteica	Proteine	Interazioni fisiche
Reti metaboliche	Metaboliti	Se partecipano allo stesso processo biochimico
Reti di regolazione	Fattori di trascrizione (TF) e geni	Relazioni dirette di regolazione tra TF e gene
Reti di co-espressione	Geni	Geni con pattern di coespressione simile sono collegati

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Human Disease - Gene Network

OMIM
Online Mendelian Inheritance in Man

1284 patologie
1777 geni malattia

HUMAN DISEASE NETWORK (HDN)

Nodes: **patologie**
Links: **se condividono
almeno un gene mutato**

Se le patologie umane avessero
tutte un'unica e distinta origine:
→ HDN sarebbe disconnessa

Francesca Vitali & Elisabetta Sautu

Corso Goh, Kwang-H et al., (2007) Proc. Natl. Acad. Sci. USA 104



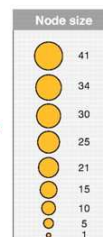
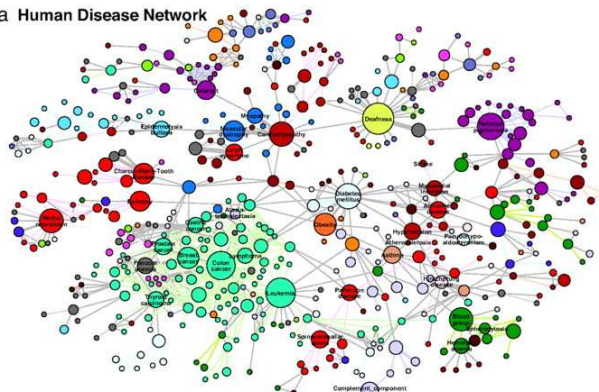
Human Disease Network

OMIM
Online Mendelian Inheritance in Man

Nodes: **patologie**
Links: **se condividono
almeno un gene mutato**



a Human Disease Network



Delle 1284
patologie

867 hanno
almeno un
link



Human Disease - Gene Network

OMIM
Online Mendelian Inheritance in Man

1284 patologie
1777 geni malattia

HUMAN DISEASE NETWORK (HDN)

Nodes: **patologie**
Links: **se condividono almeno un gene mutato**

DISEASE-GENE NETWORK (DGN)

Nodes: **Geni**
Links: **se associati con la stessa patologia**

Se le patologie umane avessero tutte un'unica e distinta origine:
→ HDN sarebbe disconnessa

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015

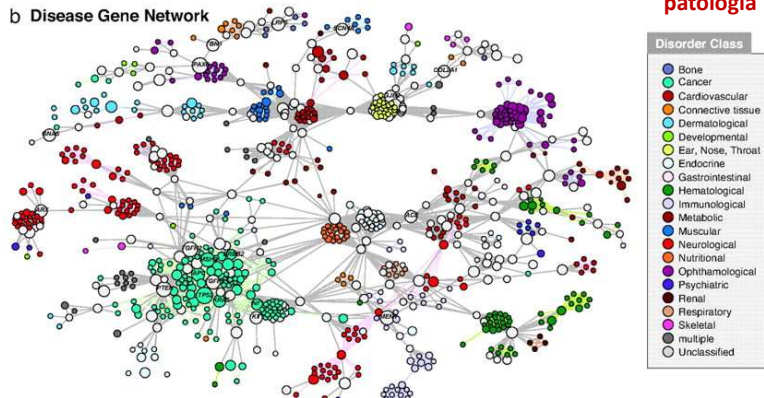


Human Disease Gene Network

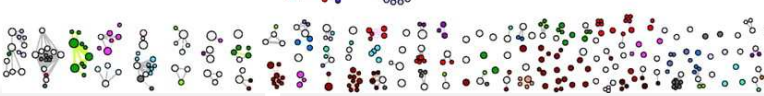
OMIM
Online Mendelian Inheritance in Man

Nodes: **Geni**
Links: **se associati con la stessa patologia**

b Disease Gene Network



Dei 1777
geni
↓
1377
hanno
almeno un
link

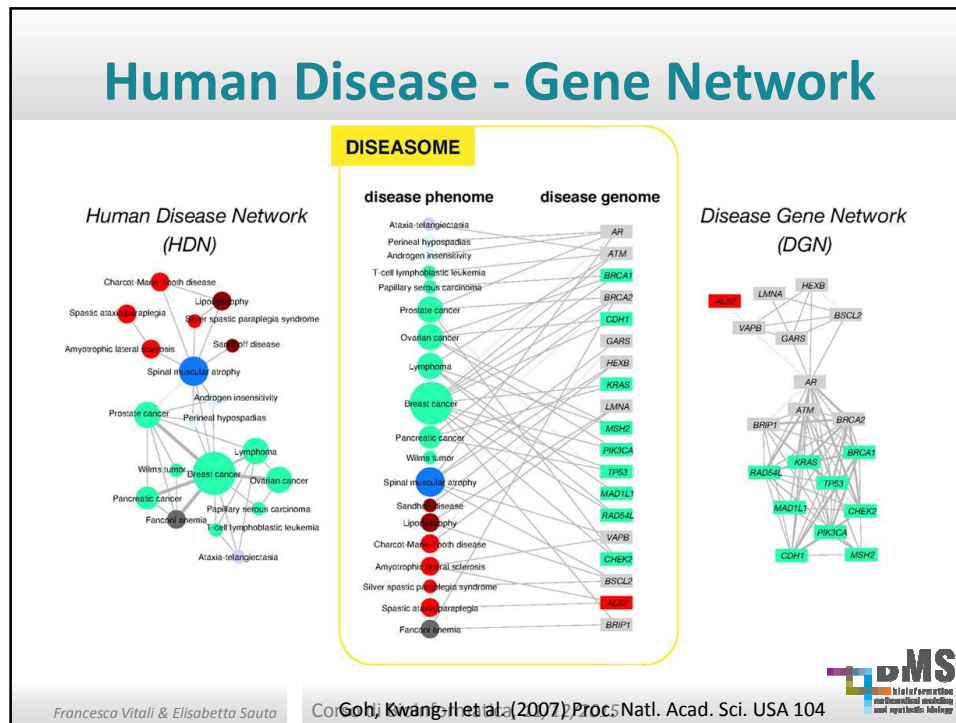


Francesca Vitali & Elisabetta Sautu

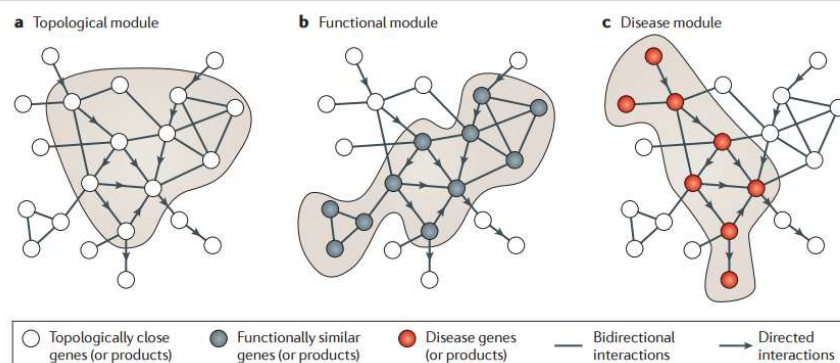
Corso di Bioinformatica, Goh, Kwang-I et al. (2007) Proc. Natl. Acad. Sci. USA 104



Human Disease - Gene Network



Studio dei moduli di una rete



Studio dei vicini

Nodi coinvolti nelle stesse funzioni cellulari:

- Tendono ad interagire tra loro
- Sono localizzati nel vicinato

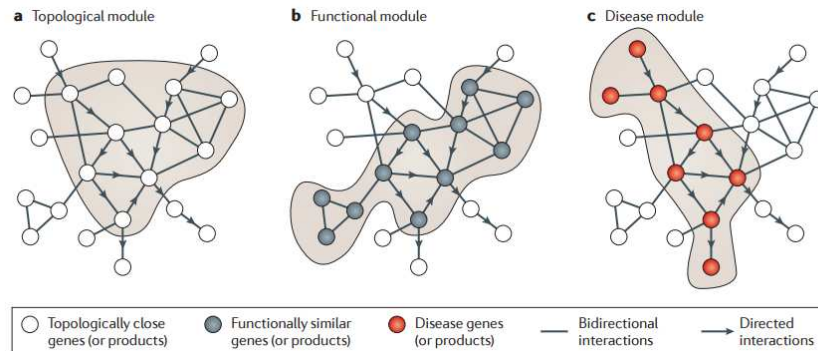
Gruppi di nodi la cui perturbazione può essere dovuta ad una particolare patologia

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015

BMS Bioinformatics and Molecular Simulation

Studio dei moduli di una rete



ASSUNZIONI DELLA "NETWORK MEDICINE"

- Moduli funzionali corrispondono a moduli topologici
- Una patologia può essere vista come un "rottura" dei moduli funzionali

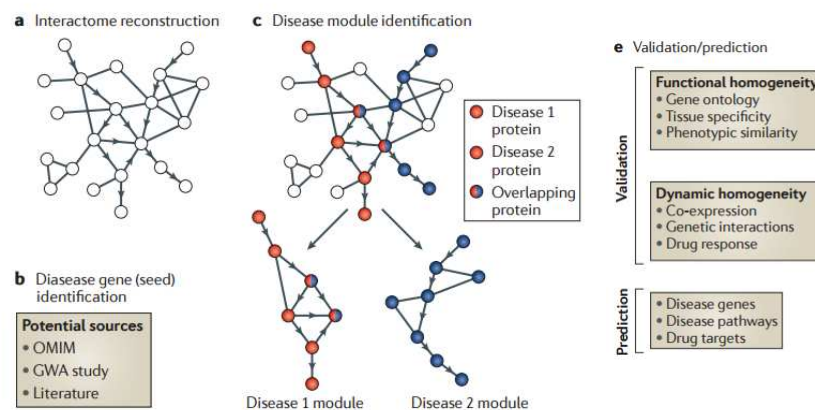
→ 3 MODULI COINCIDONO

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Come identificare i moduli malattia?

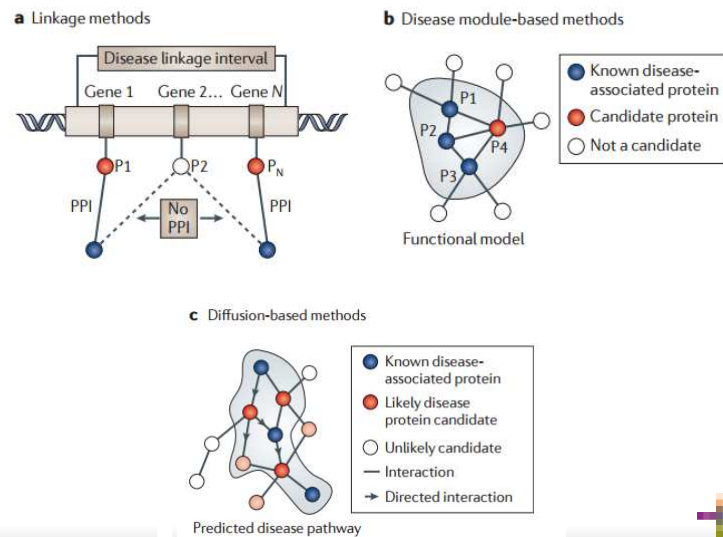


Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Selezione dei geni malattia

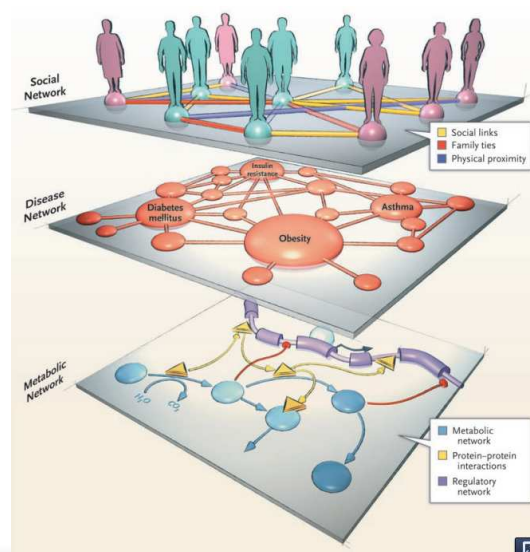


Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Network Medicine

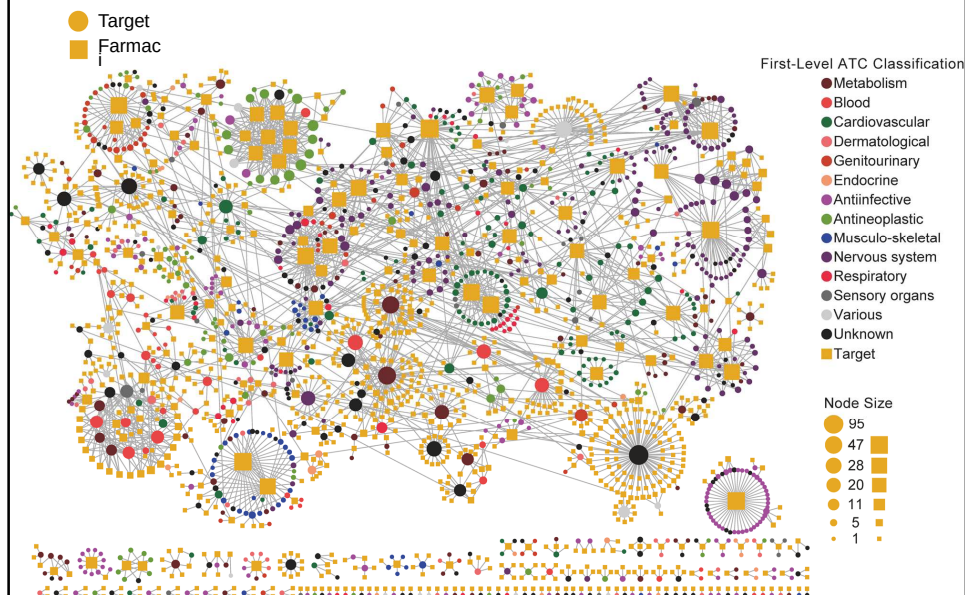


Francesca Vitali & Elisabetta Sautu

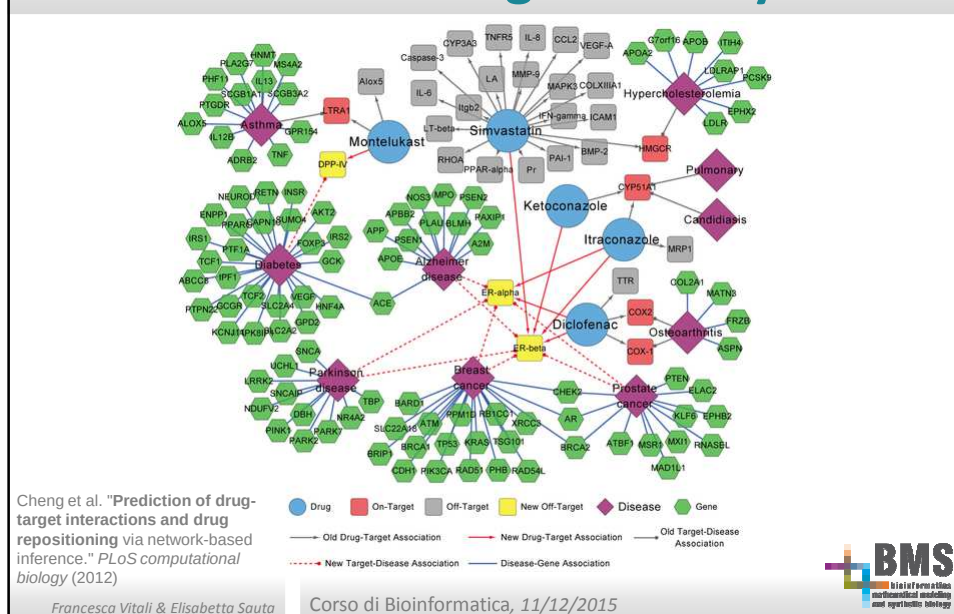
Corso di Bioinformatica, 11/12/2015



Network drug discovery



Network drug discovery



Dati e conoscenza

Pou5f1
Klf4
Sox2
Sox3

My data

BMS
Bioinformatics
and
Molecular
Systems

Francesca Vitali & Elisabetta Sauta Corso di Bioinformatica, 11/12/2015

Dati e conoscenza

ci sono articoli scientifici che li citano?

BIOINFORMATICA

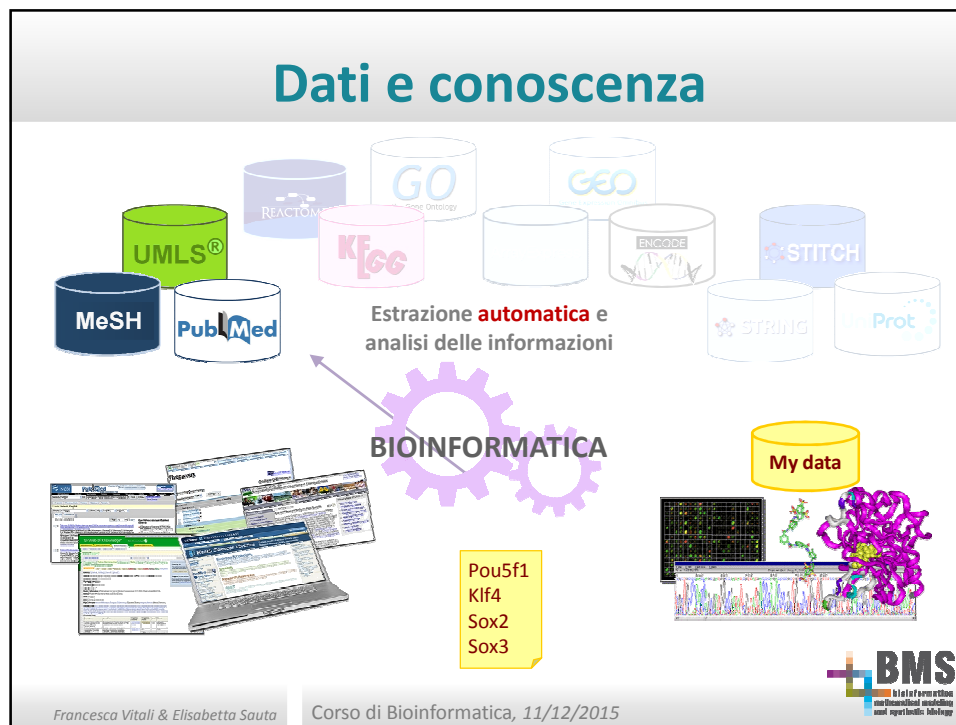
Pou5f1
Klf4
Sox2
Sox3

My data

BMS
Bioinformatics
and
Molecular
Systems

GENI o PROTEINE di interesse

Francesca Vitali & Elisabetta Sauta Corso di Bioinformatica, 11/12/2015



Letteratura scientifica: MeSh

PubMed

MeSh

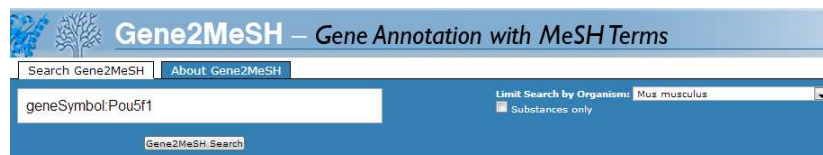
Informazioni sull'articolo provenienti da altri database

BMS
Bioinformatics
and
Molecular
Simulation

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

Annotazione MeSH dei geni



```
genes = ['Sox2', 'Pou5f1', 'Nanog']
# per ciascun gene salvo i MeSH associati e, per ciascun MeSH term,
# il numero dei paper in cui compare
gene2mesh = {}
for k, gs in enumerate(genes):
    if k % 100 == 0:
        print k
    url = "http://gene2mesh.ncbi.org/fetch?genesymbol=%s" % gs
    data = urllib.urlopen(url).read()
    dom = xml.dom.minidom.parseString(data)
    res_xml = dom.getElementsByTagName("Result")

    for i, el in enumerate(res_xml):
        h = el.getElementsByTagName("MeSH") # mesh
        j = h[0].childNodes[0].getElementsByTagName("Name")
        name = str(j[0].childNodes[0].data)
        p = el.getElementsByTagName("DocumentSet")
        np = len(p[0].childNodes) # numero di paper associati
        dictg = gene2mesh.get(gs, {}) # salvo
        dictg[name] = np
    gene2mesh[gs] = dictg
```

Codice Python (o Matlab, Java, ...):

- Effettua una query con il nome del gene *Pou5f1* sulla banca dati Gene2MeSH
- Scarica i risultati
- Annota il gene con i MeSH trovati e il numero di articoli

gene annotazioni (MeSH)

Pou5f1 [*Embryonic Stem Cells*: 98
Pluripotent stem cells: 57 ...]

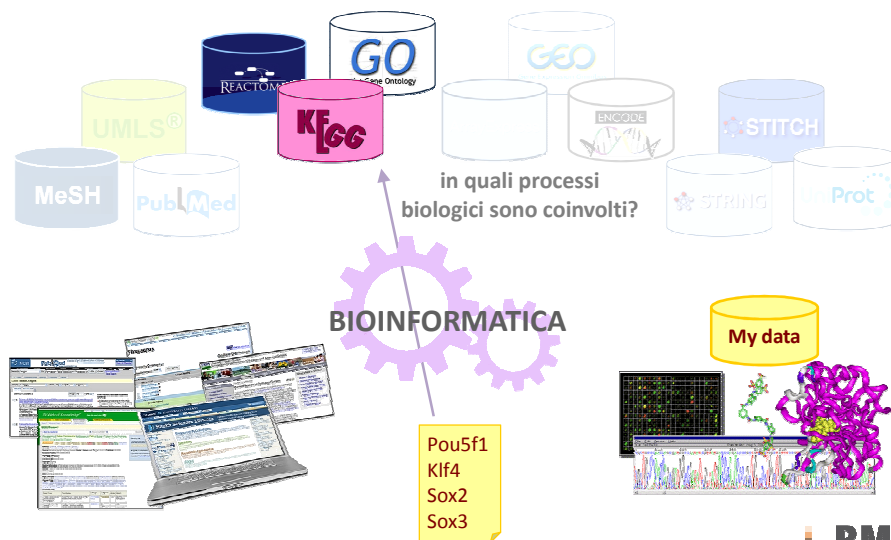
Profilo di annotazione

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Dati e conoscenza

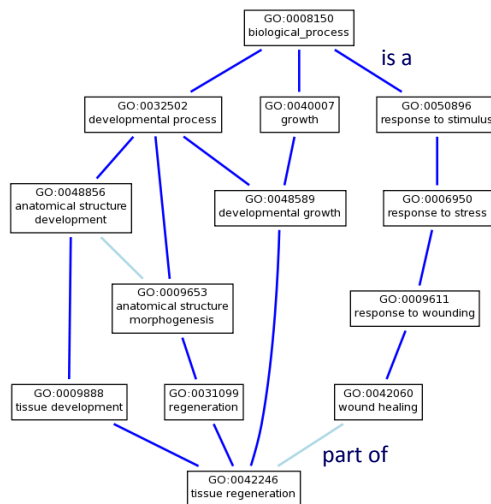


Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



La Gene Ontology



- Vocabolario controllato di termini

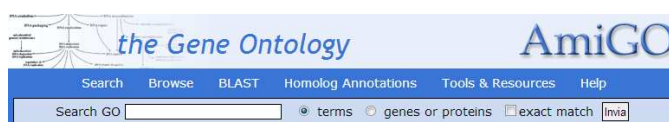
- Diverse classi di relazioni

- “è un”
- “è parte di”
- “è sinonimo di”

- Multidimensionalità

- Biological process
- Molecular function
- Cellular component

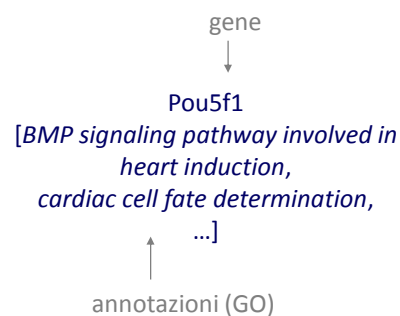
La Gene Ontology



Pou5f1

Accession, Term	Ontology
GO:0003130 : BMP signaling pathway involved in heart induction	28 gene products view in tree biological process
GO:0060913 : cardiac cell fate determination	21 gene products view in tree biological process
GO:0045165 : cell fate commitment	2522 gene products view in tree biological process
GO:0060795 : cell fate commitment involved in formation of primary germ layer	287 gene products view in tree biological process
GO:0001712 : ectodermal cell fate commitment	17 gene products view in tree biological process
GO:0001711 : endodermal cell fate commitment	131 gene products view in tree biological process
GO:0001714 : endodermal cell fate specification	74 gene products view in tree biological process

Profilo di annotazione



Analisi di enrichment

- Trovare termini di annotazione “sovrarappresentati” in una lista di geni
- I geni della mia lista associati con un pathway sono più di quelli che mi aspetterei di trovare se selezionassi una lista casuale di geni?

DAVID Tool



Enrichment p-value: test di Fisher

- Distribuzione ipergeometrica: p value=probabilità di ottenere la configurazione osservata o una “più estrema”
- Se estraggo casualmente 300 geni da 30000 (di cui 40 annotati con “p53 signaling pathway”), qual è la probabilità di ottenere almeno 3 geni annotati come “p53 signaling pathway”?

	dati	genoma
nel pathway	3	40
non nel pathway	297	29960

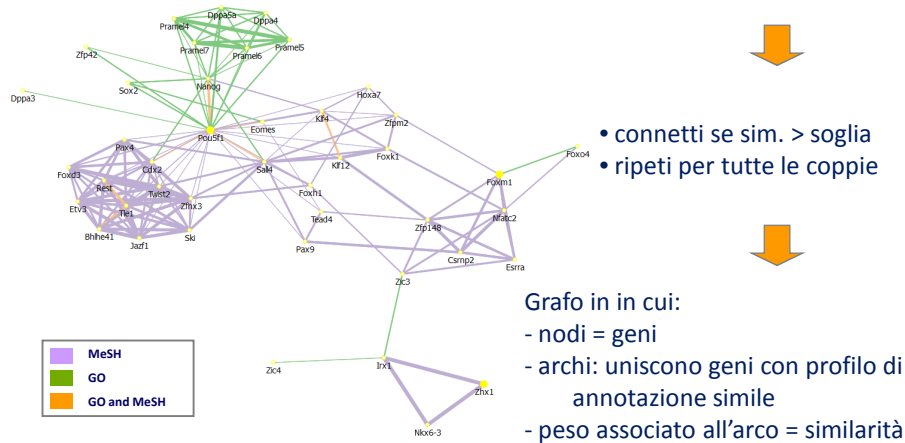
Term	RT	Genes	Count	%	P-Value
Amyotrophic lateral sclerosis (ALS)	RT		4	2,6	2,3E-2
MAPK signaling pathway	RT		8	5,2	3,4E-2
Calcium signaling pathway	RT		6	3,9	5,2E-2
Neurotrophin signaling pathway	RT		5	3,2	5,5E-2
Small cell lung cancer	RT		4	2,6	7,3E-2
Focal adhesion	RT		6	3,9	8,2E-2
Pathways in cancer	RT		8	5,2	8,3E-2
Prostate cancer	RT		4	2,6	8,4E-2

Reti di associazione genica

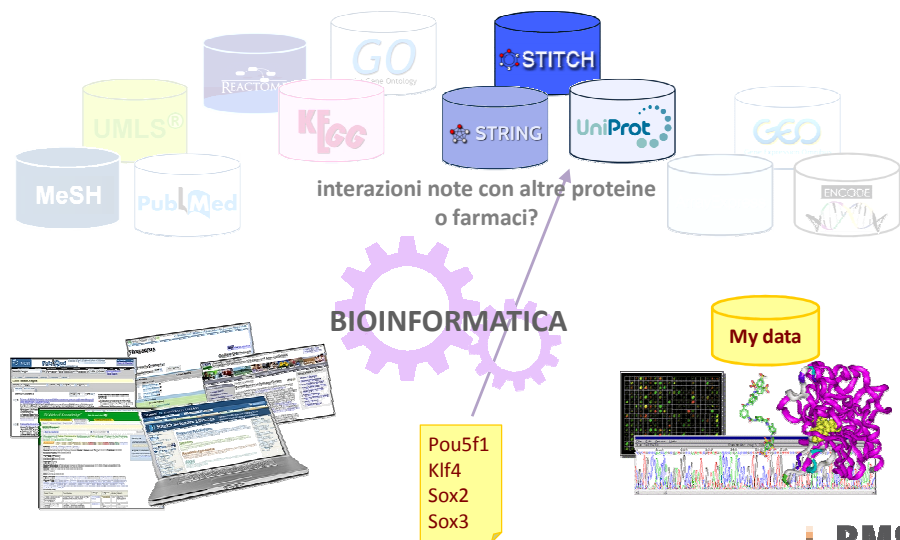


Gene1: [termini GO enriched], [termini MeSH enriched]
Gene2: [termini GO enriched], [termini MeSH enriched]

similarità
(e.g. indice di jaccard)



Dati e conoscenza



STRING

- STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) contiene associazioni note e predette fra proteine

search by name search by protein sequence multiple names multiple sequences

protein name: (examples: #1 #2 #3)
POU5F1

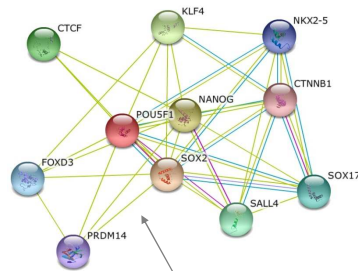
(STRING understands a variety of protein names and accessions; you can also try a [random entry](#))

organism:
auto-detect

interactors wanted:
COGs Proteins

Reset GO

please enter your protein of interest...



confidence score
per ogni link



Tipo di evidenza

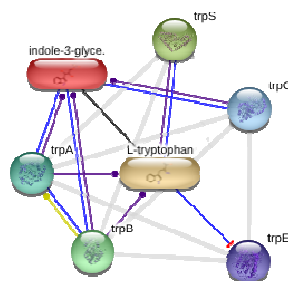
Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



STITCH

- STITCH (Search Tool for Interactions of Chemicals) contiene associazioni note e predette fra proteine e vari tipi di molecole, inclusi farmaci



- Approvati
- Neutraceutici
- Sperimentali
- Metaboliti

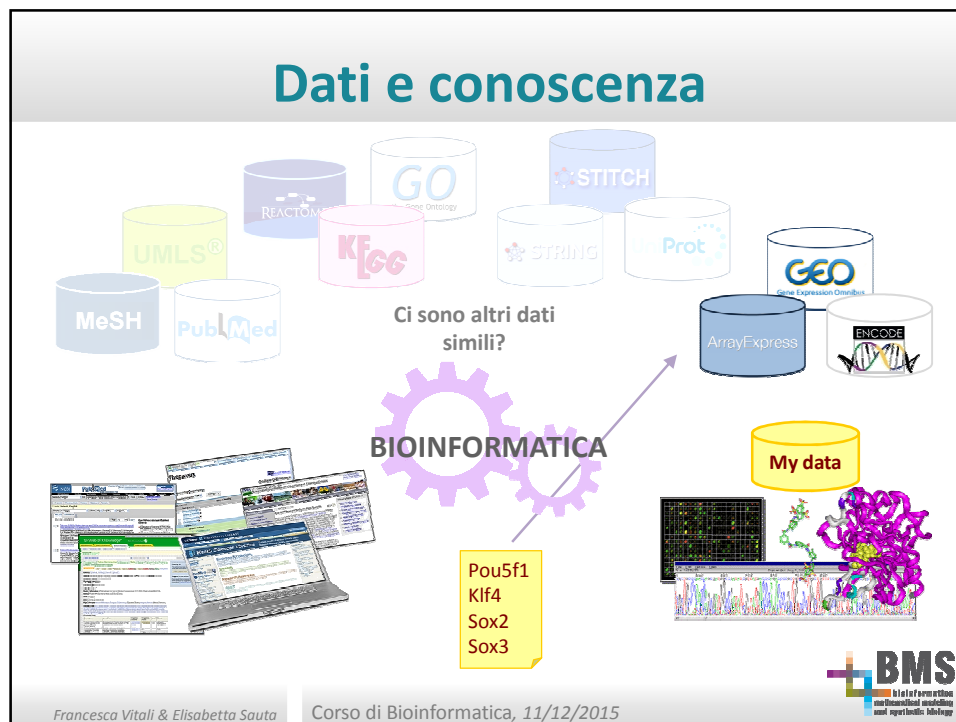
come in STRING per ogni link:

- Confidence score
- Tipo di evidenza

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015





Gene Expression Atlas

EMBL-EBI Enter Text Here Find Terms of Use | Privacy | Cookies

Databases Tools Research Training Industry About Us Help Site Index

ATLAS home | about the project | faq | feedback | das | api | help

Pou5f1 Mus musculus

Pou5f1 is differentially expressed in 123 experiments: 57 organism parts: liver, testis, ... (55 more); 2 disease states: normal, tumor; 41 cell types 2 cell lines 12 compound treatments and 13 other conditions.

Experimental Factors

organism part
studied in E-TABM-1039, E-MEXP-114, E-TABM-865, E-GEOD-24940, E-GEOD-3963, ... (35 experiments)

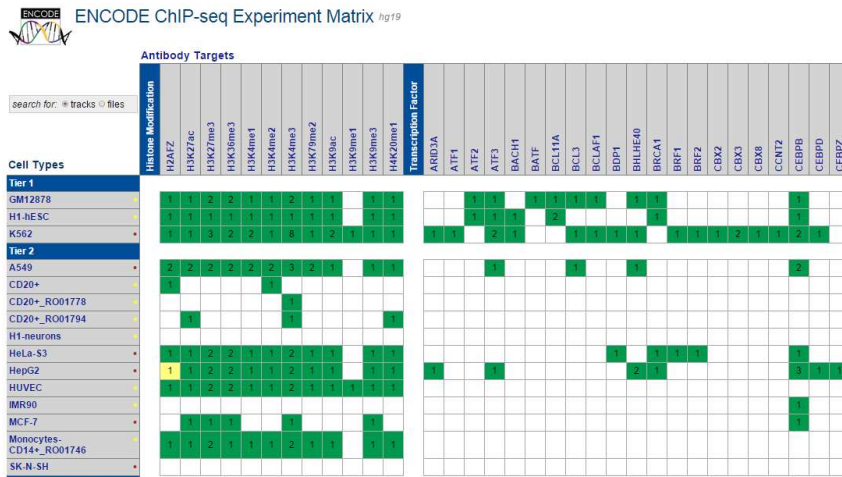
Number of published studies where the gene is over/under expressed compared to the gene's overall mean expression level in the study.

- Esperimenti
- Annotazione dei geni a tessuti, malattie o linee cellulari in base ai dati di espressione contenuti nel database

Francesca Vitali & Elisabetta Sauta Corso di Bioinformatica, 11/12/2015

BMS
Bioinformatics in the context of systems biology and experimental biology

ENCODE ChIP-seq

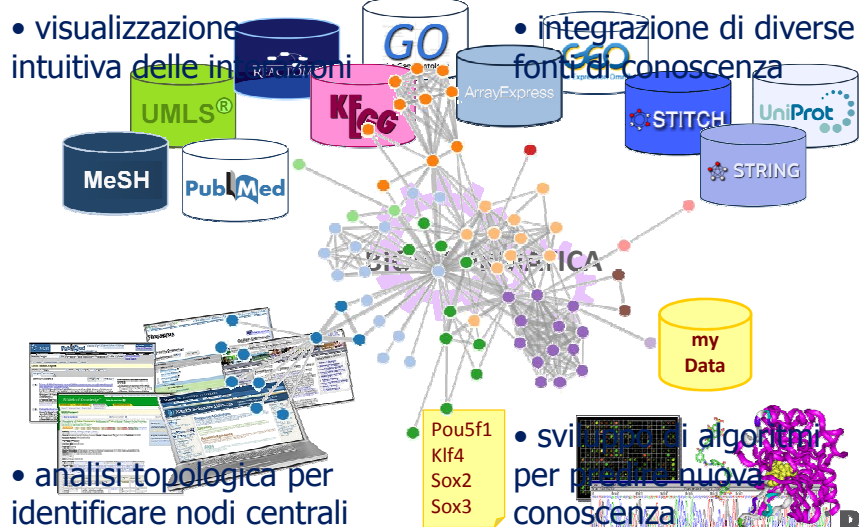


Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Bioinformatica integrativa



Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



Attività e collaborazioni

- Reti di interazione proteica per la selezione di potenziali terapie per I tumori al seno di tipo triplo negativo



Fondazione IRCCS Salvatore Maugeri

Pavia

AstraZeneca 

Astra Zeneca

Sweden

- Reti di associazione genica per lo studio delle leucemie e per l'identificazione automatica di target



Fondazione IRCCS Policlinico San Matteo

Pavia



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

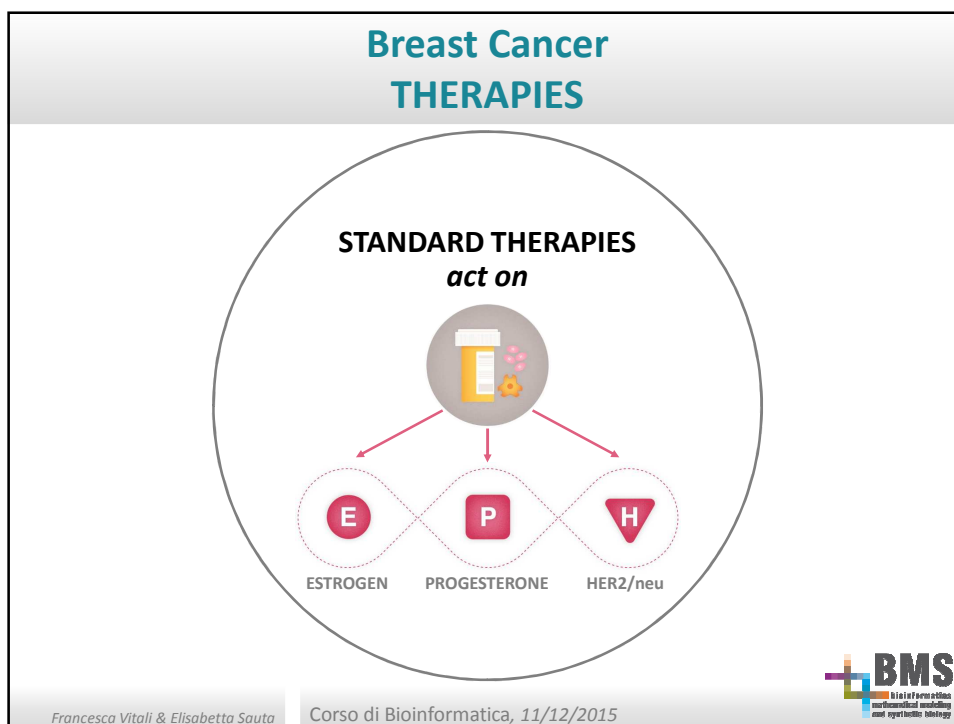
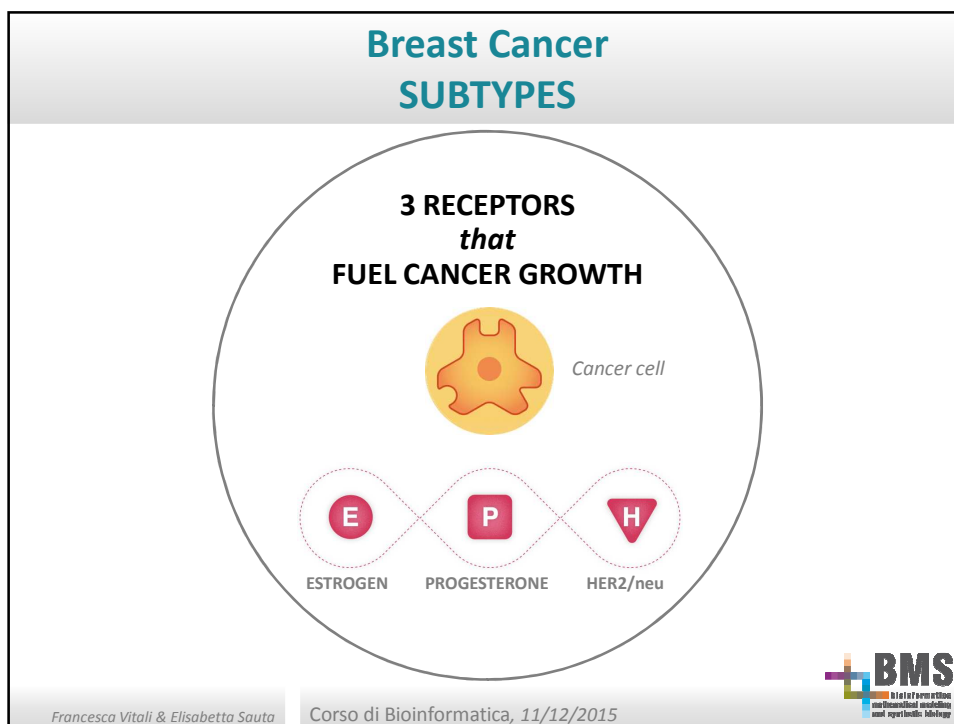
Attività e collaborazioni

- Reti di interazione proteica per la selezione di potenziali terapie per I tumori al seno di tipo triplo negativo
- Reti di associazione genica per lo studio delle leucemie e per l'identificazione automatica di target

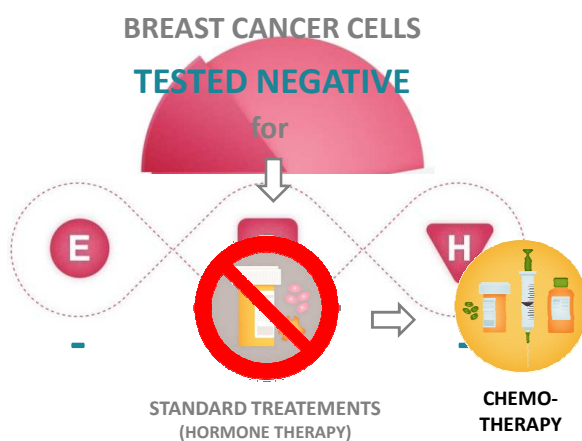


Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



Triple Negative Breast Cancer - TNBC



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



NETWORK BASED APPROACH

SEARCH FOR NEW THERAPIES



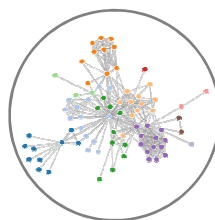
BY EXPLOITING THE PRIOR KNOWLEDGE

DATA SOURCES



+

INTERACTION
NETWORKS



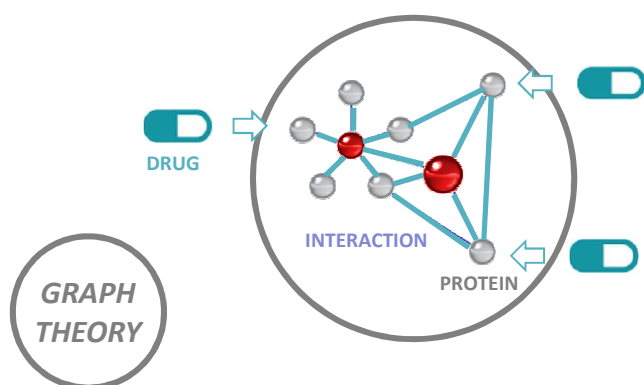
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



MULTI-TARGET APPROACH

PROTEIN-PROTEIN INTERACTION (PPI) NETWORK



Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



PIPELINE



Selection of
**CANDIDATE
TARGET COMBINATIONS**



Analysis of
DRUGS



Simulations of
DRUGS ACTIONS



Plan
**IN-VITRO
EXPERIMENTS**

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



PIPELINE



Selection of
**CANDIDATE
TARGET COMBINATIONS**



Analysis of
DRUGS



Simulations of
DRUG ACTIONS



Plan
**IN-VITRO
EXPERIMENTS**

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



PIPELINE



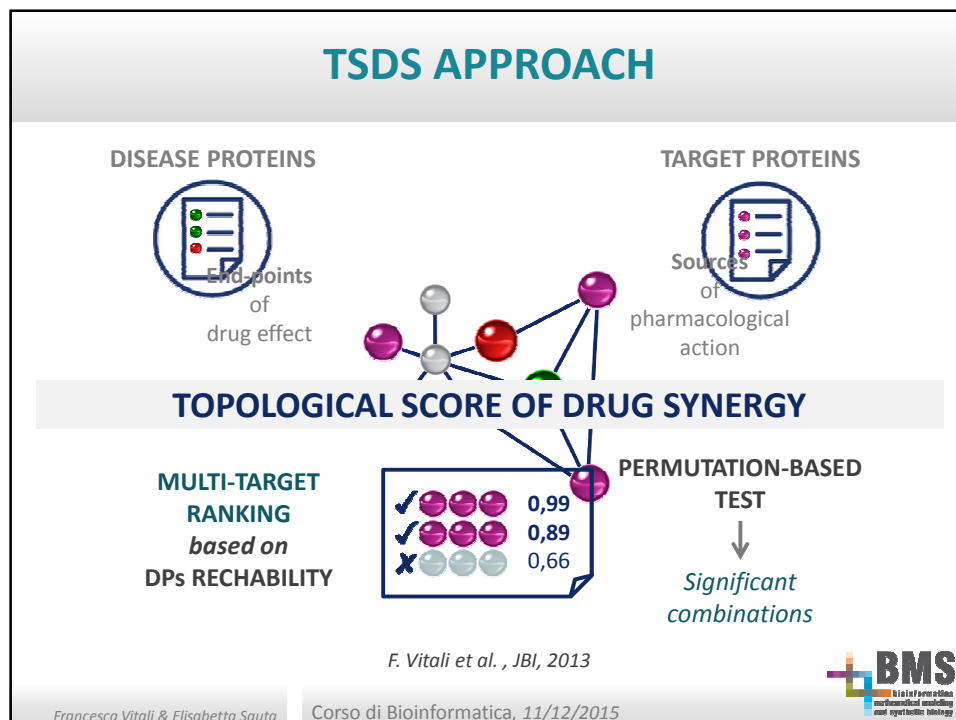
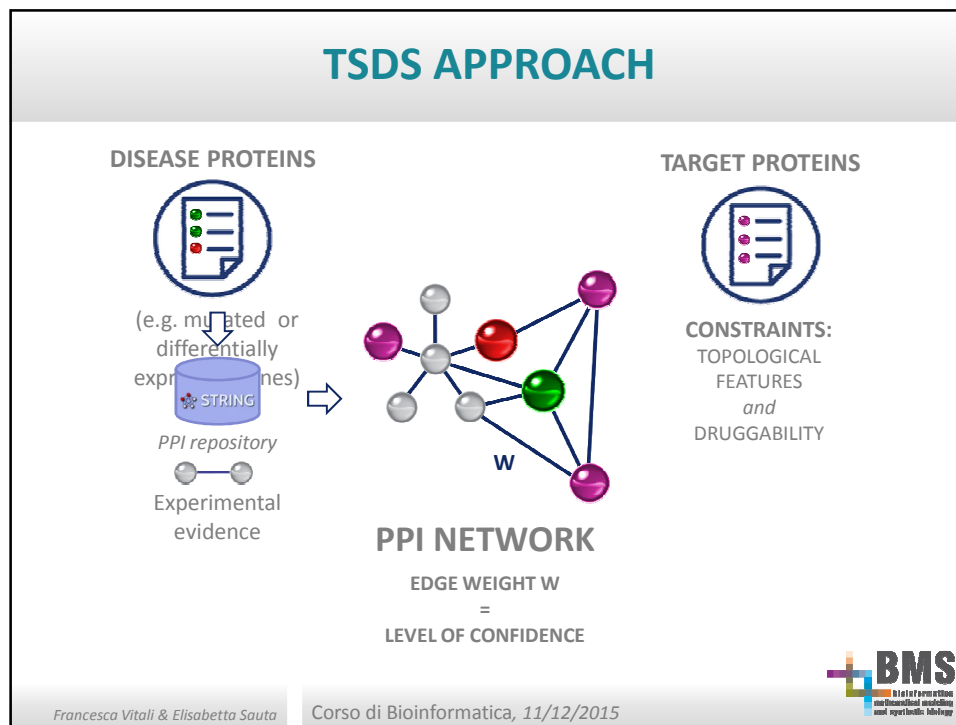
Selection of
**CANDIDATE
TARGET COMBINATIONS**

through
**TOPOLOGICAL SCORE OF DRUG SYNERGY
(TSDS)**

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015





PIPELINE



Selection of
CANDIDATE
TARGET COMBINATIONS



Analysis of
DRUGS



Simulations of
DRUG ACTIONS



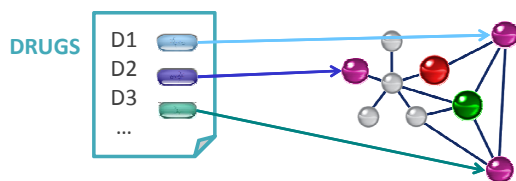
Plan
IN-VITRO
EXPERIMENTS

Francesca Vitali & Elisabetta Sauta

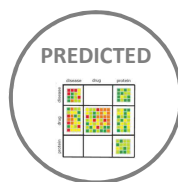
Corso di Bioinformatica, 11/12/2015



DRUG ANALYSIS



Drug-Target Interactions
DTI

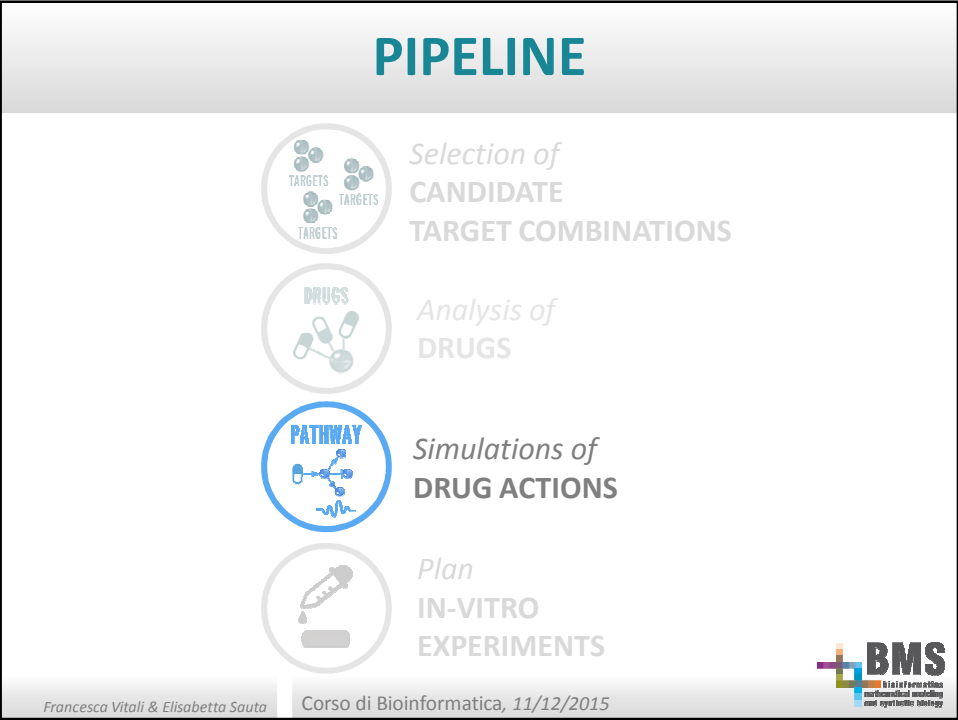


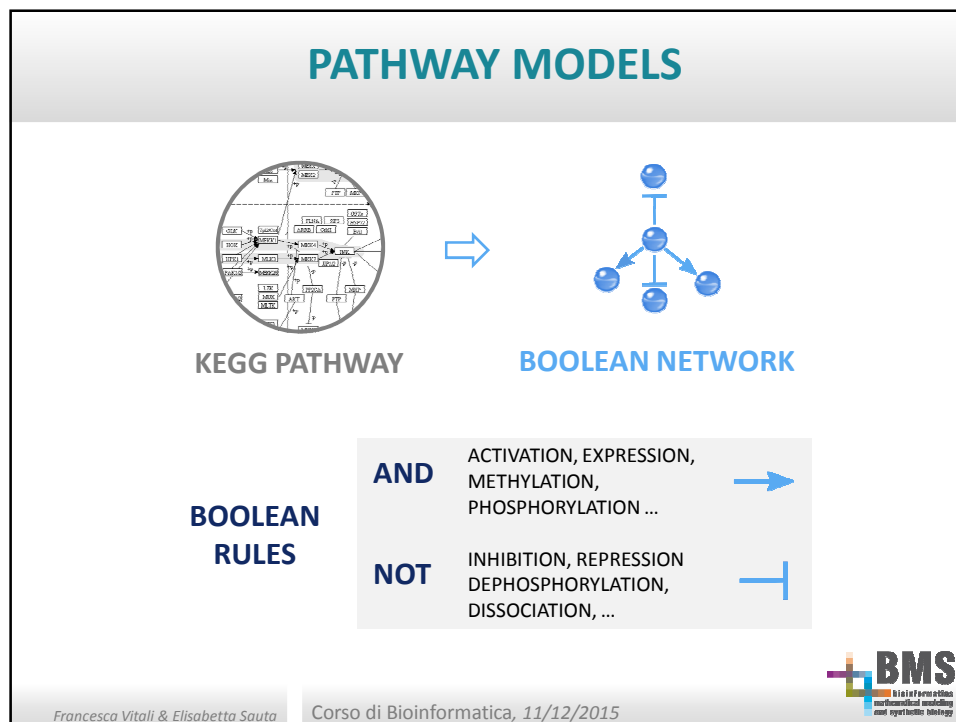
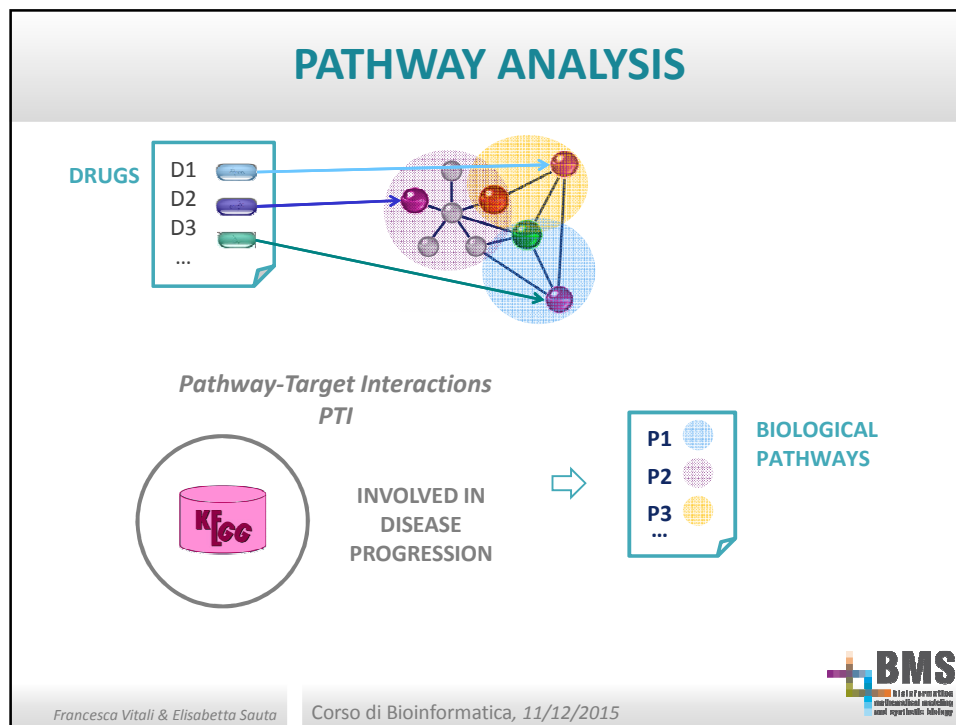
Matrix Tri-factorization

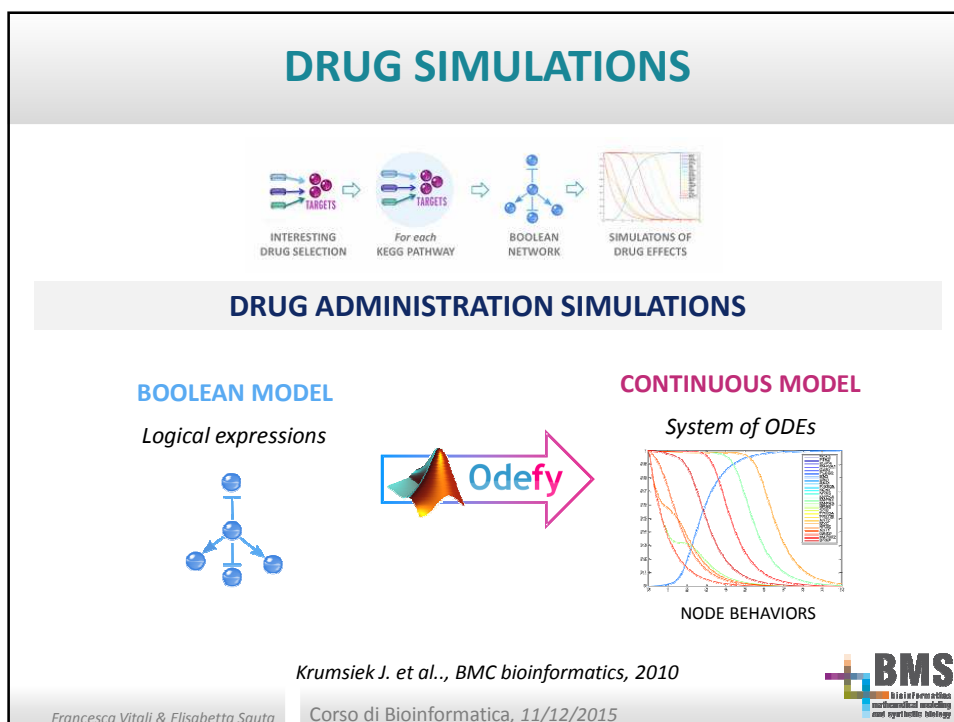
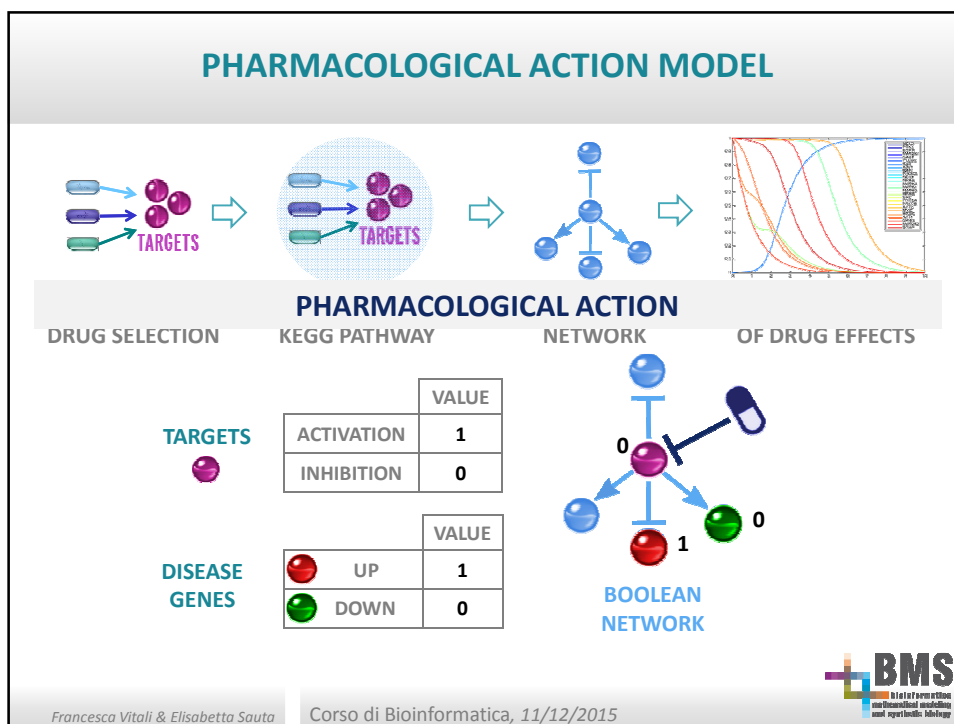
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015





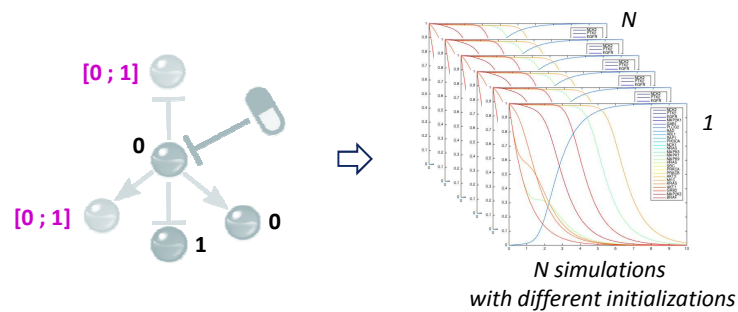




DRUG SIMULATIONS



MONTECARLO SIMULATIONS

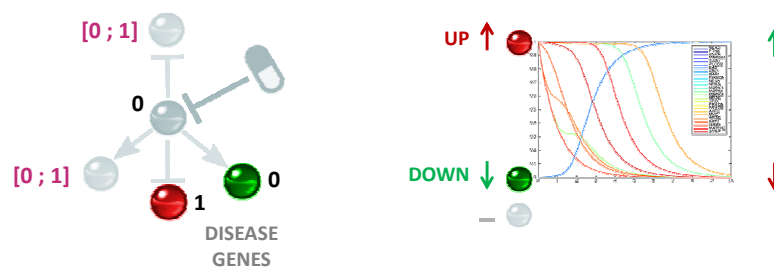


Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULT ANALYSIS



DRUG/DRUG COMBINATION

has to
REGULARIZE as much as possible Disease Genes
while
other nodes must NOT be PERTURBED

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULT ANALYSIS

CONFUSION MATRIX

	<i>Disease Genes</i>	<i>Disease Genes</i>	
<i>Regularized Genes</i>	TP	FP	<i>Tot RG</i>
<i>Regularized Genes</i>	FN	FN	<i>Tot RG</i>
	<i>Tot DG</i>	<i>Tot DG</i>	<i>N</i>

$$\text{Sensitivity} = \frac{TP}{\text{Tot DG}}$$

$$\text{Precision} = \frac{TP}{\text{Tot PG}}$$

AIM: Maximization of regularized disease genes (TP)

Index of
THERAPEUTIC
EFFICACY

$$\text{EFFECT index} = 2 * \frac{\text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



PIPELINE



Selection of
CANDIDATE
TARGET COMBINATIONS



Analysis of
DRUGS



Simulations of
DRUG ACTIONS



Plan
IN-VITRO
EXPERIMENTS

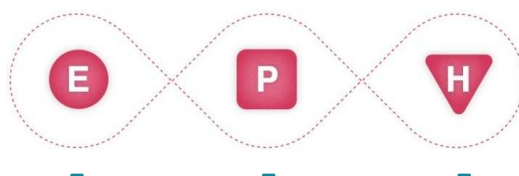
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULTS - TNBC

BREAST CANCER CELLS
TESTED NEGATIVE
for



Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



RESULTS - TSDS APPROACH

Shah et al., 2012

LETTER

The clonal and mutational evolution spectrum of primary triple-negative breast cancers

32 TARGET PROTEINS

43 DISEASE PROTEINS

involved in the

TOPOLOGICAL SCORE OF DRUG SYNERGY

MULTI-TARGET
RANKING

PPI repository

134 combinations of
16 different proteins
PPI NETWORK

None is currently used in
ongoing clinical trials

Francesca Vitali & Elisabetta Sautu

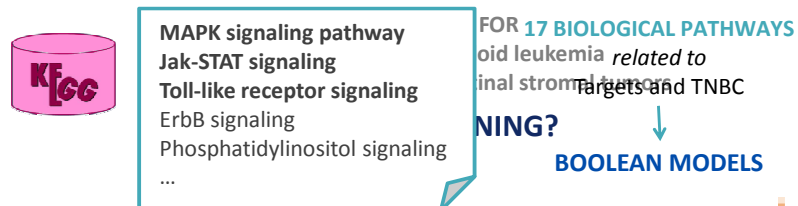
Corso di Bioinformatica, 11/12/2015



RESULTS - DRUG AND PATHWAY ANALYSIS



IMATINIB PATHWAY ANALYSIS



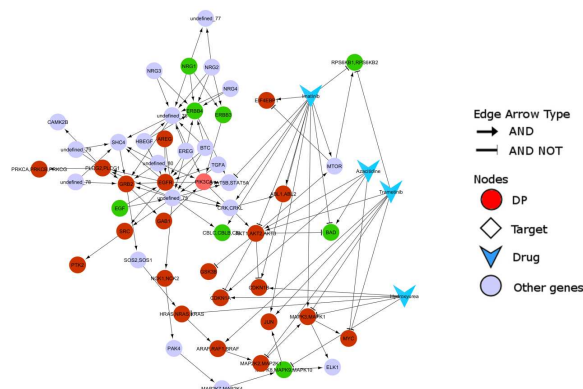
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULTS - PATHWAY ANALYSIS

MAPK signaling pathway – BOOLEAN NETWORK



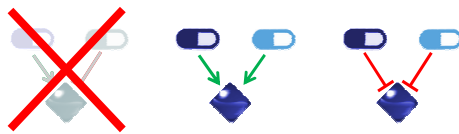
Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULTS - PATHWAY ANALYSIS

MULTI-TARGET APPROACH = multiple drugs acting on different targets



NO OPPOSITE DRUG-TARGET INTERACTIONS

DRUG ADMINISTRATION SIMULATIONS

1. Imatinib
2. Imatinib + Vemurafenib
3. Imatinib + Flucytosine
4. Imatinib + Vemurafenib + Flucytosine



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULTS – MONTECARLO SIMULATIONS

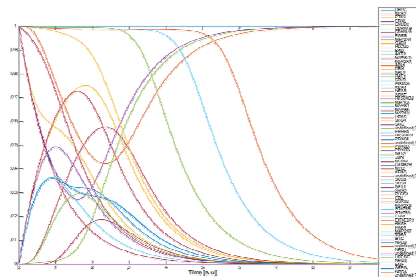
for each **DRUG**:

for each **PATHWAY**:

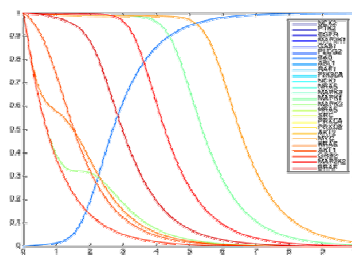
MONTECARLO SIMULATIONS by applying **Odefy**



Example: Imatinib – MAPK pathway



ALL NODE BEHAVIOURS



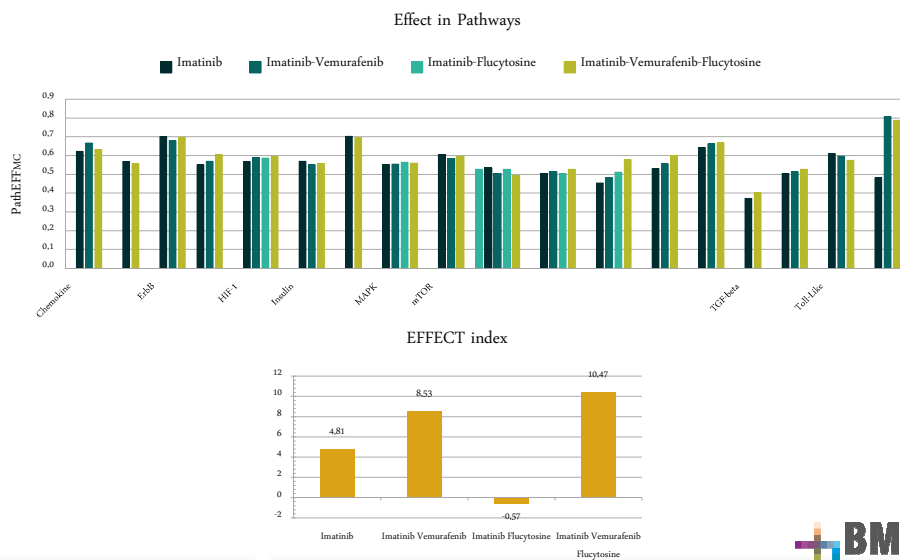
DISEASE NODE BEHAVIOURS

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



RESULTS – MONTECARLO SIMULATIONS



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



IMATINIB IN-VITRO EXPERIMENTS

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

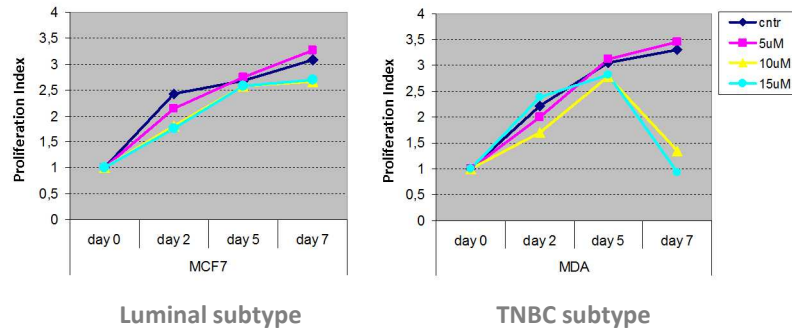


IN VITRO RESULTS



MTT assay – Evaluation of cell viability

2-5-7 days of treatment with different concentration of Imatinib



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

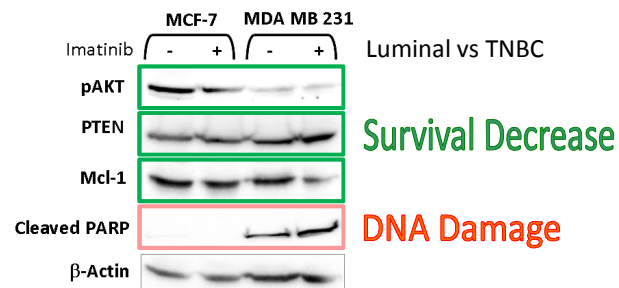


IN VITRO RESULTS



Western Blot assay – Disease gene evaluation

One week treatment with 10μM Imatinib



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

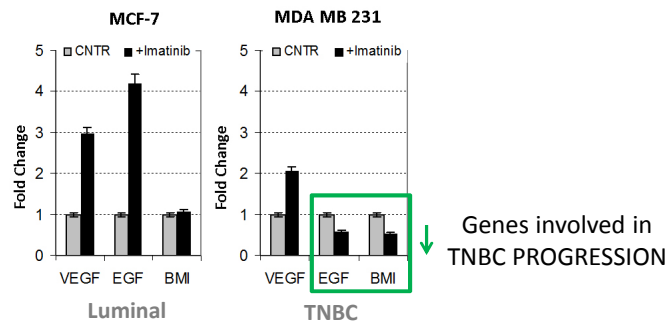


IN VITRO RESULTS



Gene expression assay – Disease gene evaluation

One week treatment with 10 μ M Imatinib



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

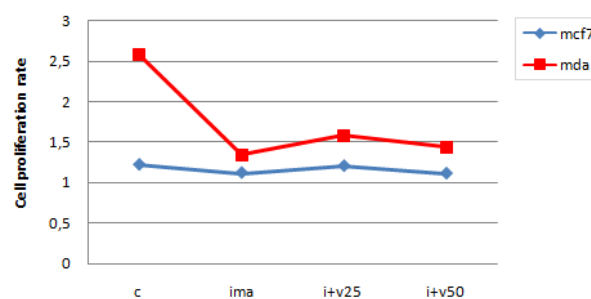


IN VITRO RESULTS - COMBINATION



MTT assay – Evaluation of cell viability

IMATINIB + VEMURAFENIB



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015





IN-VITRO EXPERIMENTS ON GOING of drug combinations

Francesca Vitali & Elisabetta Sautu

Corso di Bioinformatica, 11/12/2015



CONCLUSIONS

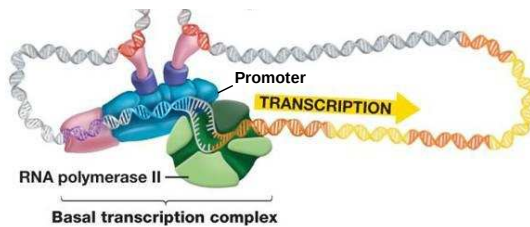
- The application to breast cancer
 - highlights **potential targets and drugs**
 - enables the selection of top candidate pathways and potential **drugs** in a variety of multifactorial diseases
- The developed approach can easily be applied to other complex diseases
 - Network construction based on a list of mutated proteins in patient's specific proteomic or genomic background – precision medicine

Francesca Vitali & Elisabetta Sautu

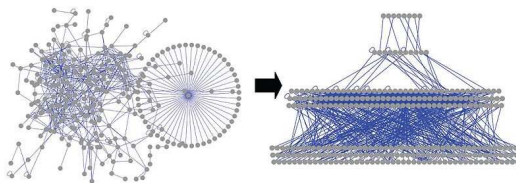
Corso di Bioinformatica, 11/12/2015



PIPELINE



Identification of
**TRANSCRIPTION FACTORS
REGULON**



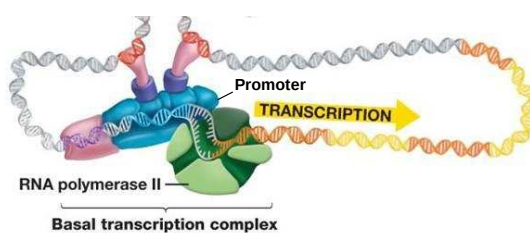
Analysis of
TRANSCRIPTIONAL MODULES



Francesca Vitali & Elisabetta Sauta

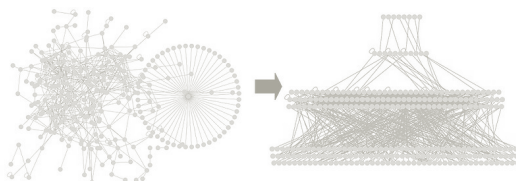
Corso di Bioinformatica, 11/12/2015

PIPELINE



Identification of
**TRANSCRIPTION FACTORS
REGULON**

through
ChIP-seq Data



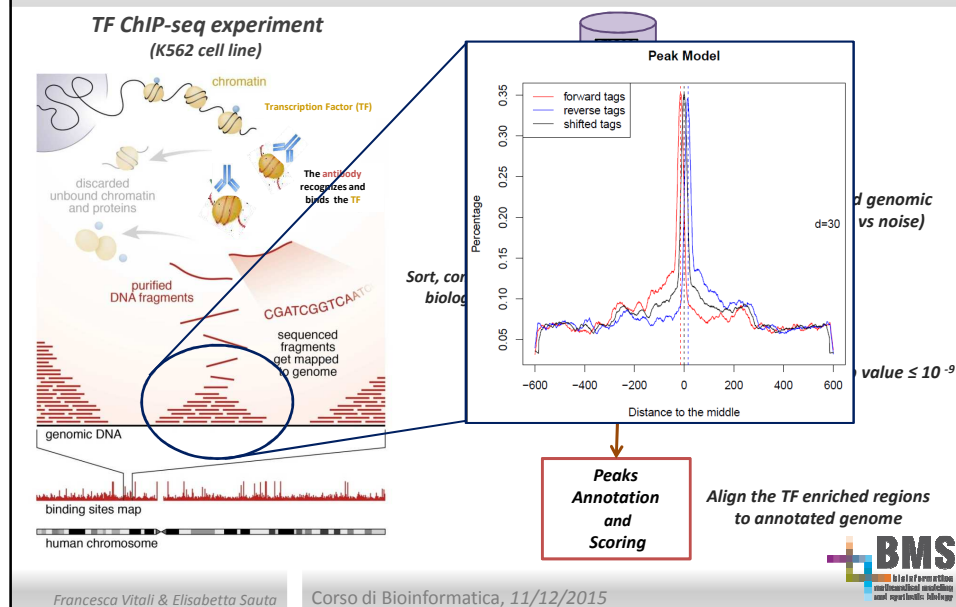
Analysis of
TRANSCRIPTIONAL MODULES



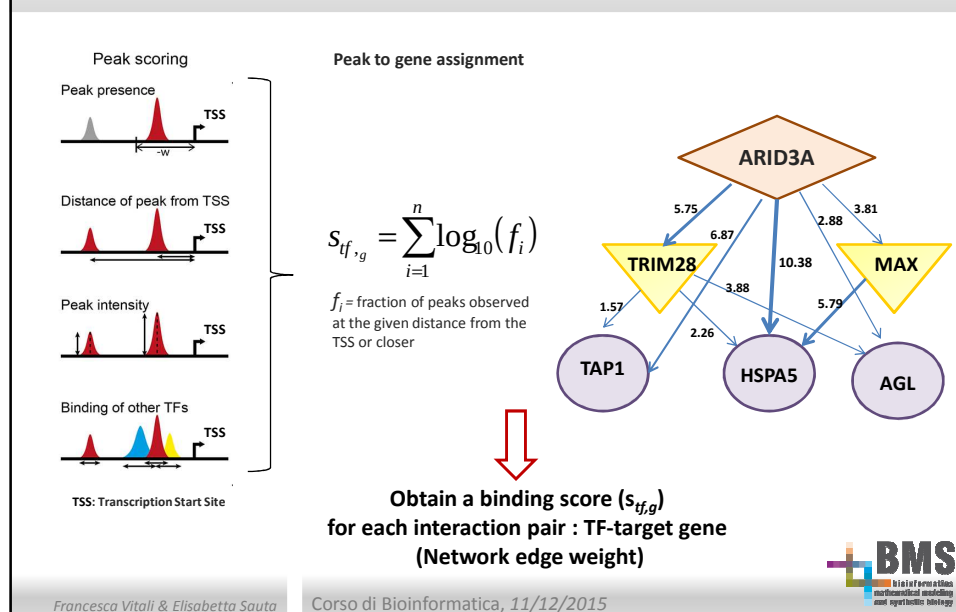
Francesca Vitali & Elisabetta Sauta

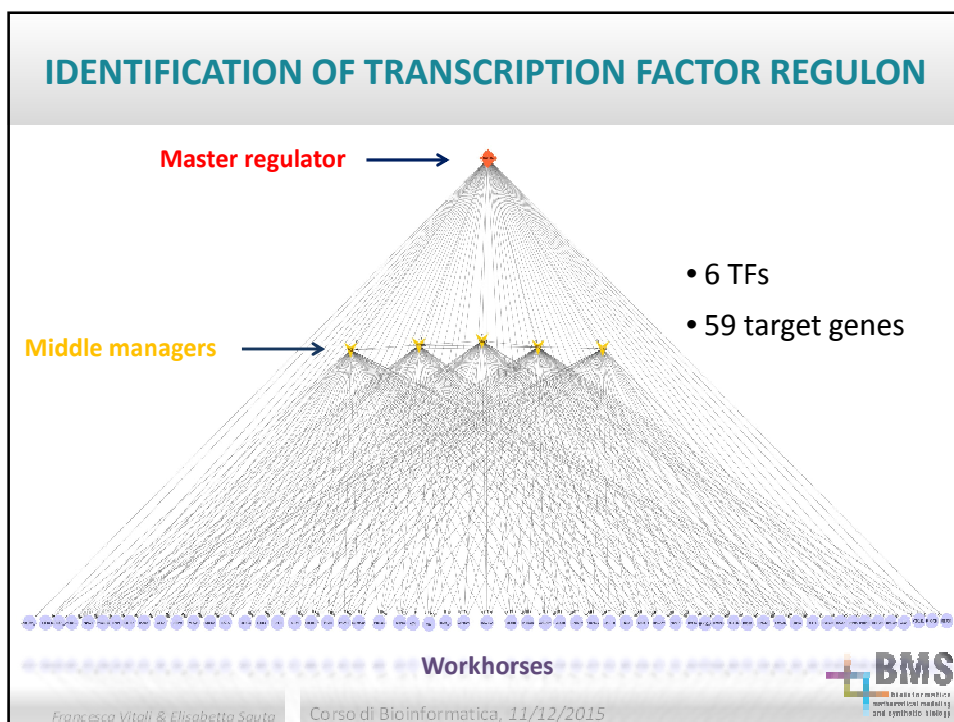
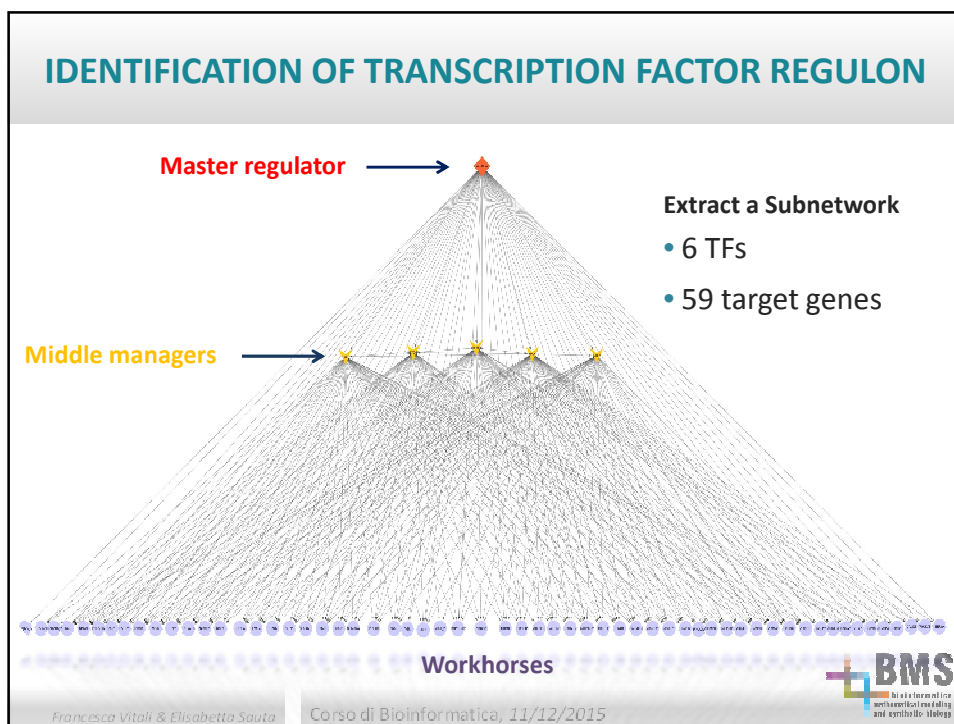
Corso di Bioinformatica, 11/12/2015

IDENTIFICATION OF TRANSCRIPTION FACTORS REGULON

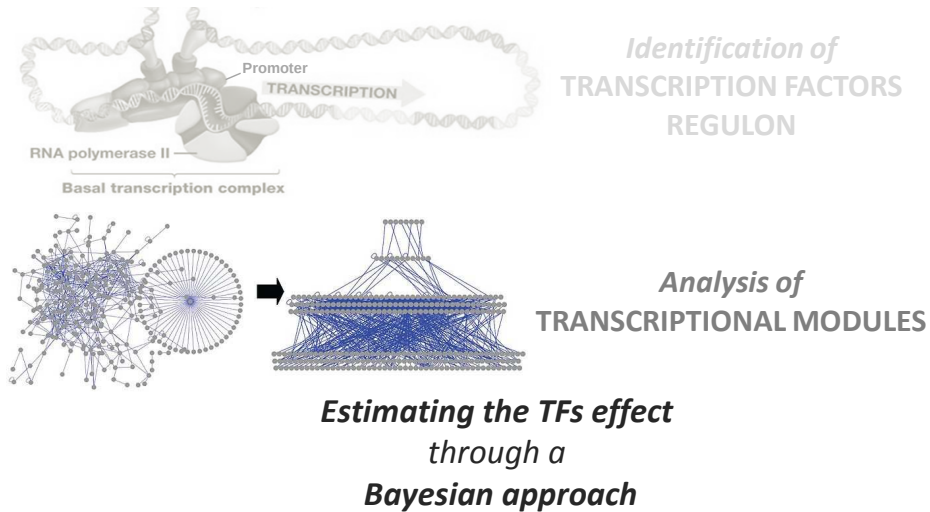


IDENTIFICATION OF TRANSCRIPTION FACTORS REGULON





PIPELINE



Francesca Vitali & Elisabetta Sauta

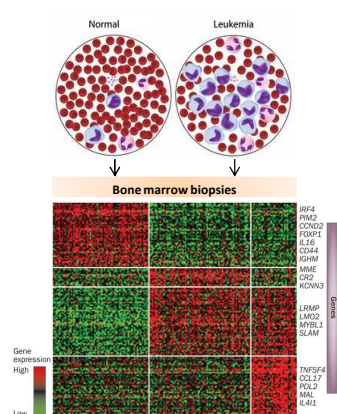
Corso di Bioinformatica, 11/12/2015



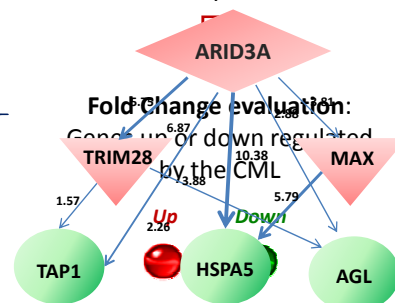
ESTIMATING THE TF EFFECT: INTEGRATE THE TRN



Integrate the TRN with gene expression profiles from K562 cell line



Fold Change evaluation:
Gene expression profiles by array
 healthy vs CML



Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



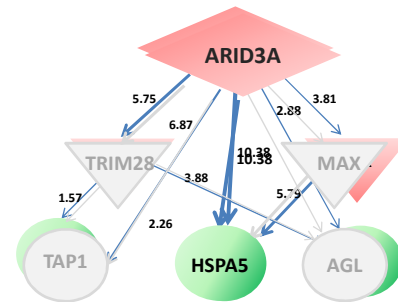
ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

Bayesian Network

1. Model Structure:

Directed Acyclic Graph (DAG)

	ARID3A	TRIM28	MAX	TAP1	HSPA5	AGL
ARID3A	0	1	1	1	1	1
TRIM28	0	0	0	1	0	1
MAX	0	0	0	0	1	0
TAP1	0	0	0	0	0	0
HSPA5	0	0	0	0	0	0
AGL	0	0	0	0	0	0



2. Model Parameters:

Conditional Probability Tables (CPT)

$$\text{odds}_{tf,g} = \frac{p_{\text{binding}}}{1 - p_{\text{binding}}}$$

$$p_{tf,g} = \frac{\text{odds}_{tf,g}}{1 + \text{odds}_{tf,g}} = \frac{s_{tf,g}}{1 + s_{tf,g}}$$

e.g.

$$p_{ARID3A, HSPA5} = 10.38 / (1 + 10.38) = 0.912$$

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

2. Model Parameters:

Conditional Probability Tables (CPT)

joint probability

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i))$$

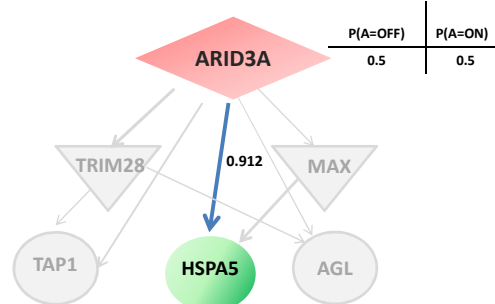
$X_1 \dots X_n$ random variables

$pa(X_i)$ set of parents variables

Each variable X_i is independent of its non-descendants (ND)

$$X_i \perp ND(X_i) | pa(X_i)$$

e.g. $TRIM28 \perp MAX | ARID3A$



Noisy OR		Binary state to each node	
ARID3A	P(HSPA5=OFF)	P(HSPA5=ON)	
P(OFF)	On = 2	Off = 1	-
P(ON)	0.5 * (1 - 0.912)	0.912	

$$P(X_i = ON) | pa(X_i) = 1 - \prod_j q_{i,j}$$

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

- Matlab Bayes Net toolbox (BNT)
- Using the experimental evidence (from the FC evaluation), each node has two possible states: Up Down
● = 2 ● = 1
- 2 different approaches:



Estimate the posterior probability of the parents nodes (TFs) for each state



Estimate the posterior probability of the child nodes for each state, using the opposite of the experimental evidence state for the TFs nodes

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

1. Bottom Up

Estimate the posterior probability of the parents nodes (TFs) for each state

TF	ESTIMATED STATE	EXP. EVIDENCE
ARID3A	UP ($p=1$)	UP
TRIM28	UP ($p=1$)	UP
MAX	DOWN ($p=1$)	DOWN

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

2. Top Down

- Using the opposite of the **estimated states** for each parent node (TF)

TF	ESTIMATE STATE	<u>OPPOSITE STATE</u>
ARID3A	UP ($p=1$)	DOWN
TRIM28	UP ($p=1$)	DOWN
MAX	DOWN ($p=1$)	UP

- Using all possible TF combinations

e.g. ARID3A-TRIM28; ARID3A - MAX; TRIM28-ARID3A; ARID3A-TRIM28-MAX..

Estimate the posterior probability of the child nodes (target genes) for each state

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: BAYESIAN APPROACH

Identify the best TFs combination as the most effective regulators

$$\text{log-likelihood score} \quad \log(L(p)) = -\sum_{i=1}^n \log(p)$$

p : the posterior probability of each child node state



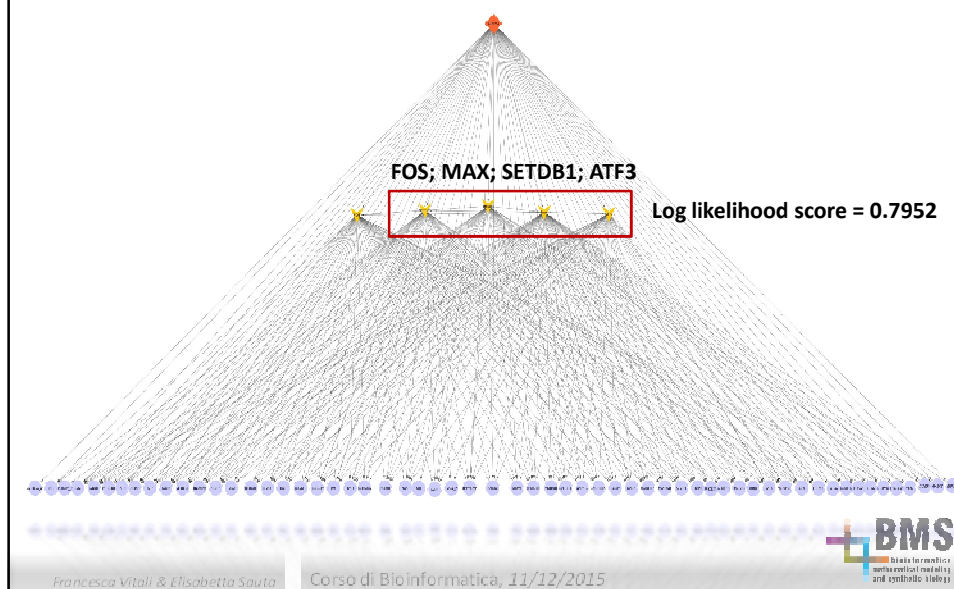
> posterior probability → < log-likelihood score → > TFs combination efficacy

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015



ESTIMATING THE TF EFFECT: PRELIMINARY RESULTS



ESTIMATING THE TF EFFECT: PRELIMINARY CONCLUSIONS



Down

🟢 **ATF3**: member of the ATF/CREB family of transcription factors that functions as a **tumor suppressor** → **potential drug target**

Down

🟢 **MAX**: member of the leucine family of zipper transcription factors plays a key role in myelopoietic differentiation program

🟢

SETDB1: acts as transcriptional repressor

Down

🟢 **FOS**: member of the leucine family of zipper transcription factors is involved in cell cycle regulation and progression

Francesca Vitali & Elisabetta Sauta

Corso di Bioinformatica, 11/12/2015

