

Computational Systems Biology

UniShare

Davide Cozzi
@dlcgold

Indice

1	Introduzione	2
2	Introduzione alla Systems Biology	3
2.1	PCNA ubiquitylation	11
2.2	I Sistemi Complessi	13
2.3	Rappresentazione Grafica	19
2.4	Tipologie di Modelli	19
2.4.1	Modelli Basati su Interazioni	25
2.4.2	Modelli Logici	26
2.4.3	Modelli Meccanicistici	27
2.4.4	Modelli Basati su Vincoli	28
2.4.5	Confronto tra i Vari Approcci	29
3	Interaction-Based Modelling	32
3.1	Introduzione alle Reti PPI	35
3.2	La Teoria dei Grafi	37

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: <https://github.com/dlcgold/Appunti>.

Capitolo 2

Introduzione alla Systems Biology

Per descrivere sistemi biologici complessi si hanno vari tipi di modelli. Kitano (il “padre” di quest’ambito), nel 2002, disse che per capire i sistemi biologici complessi bisogna integrare risultati sperimentali e metodi computazionali, ottenendo quindi la vera e propria **Systems Biology**. Tramite l’interazione di vari componenti si ottengono tali sistemi. Disse infatti:

To understand complex biological systems requires the integration of experimental and computational research — in other words a systems biology approach.

Weston, nel 2004, ha aggiunto l’importanza dello studio delle interazioni e delle regolazioni tra i vari componenti del sistema, studiando le risposte alla genetica o alle perturbazioni ambientali, al fine di capire nuove proprietà del sistema. Infatti disse:

Systems biology is the analysis of the relationships among the elements in a system in response to genetic or environmental perturbations, with the goal of understanding the system or the emergent properties of the system

Ideker (altro “padre” di quest’ambito), già nel 2001, aveva definito la System Biology come l’integrazione dei dati sperimentali con i modelli matematici che descrivono componenti e interazioni, al fine di simulare il comportamento complessivo “in silico”. Nel dettaglio, citandolo:

Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations

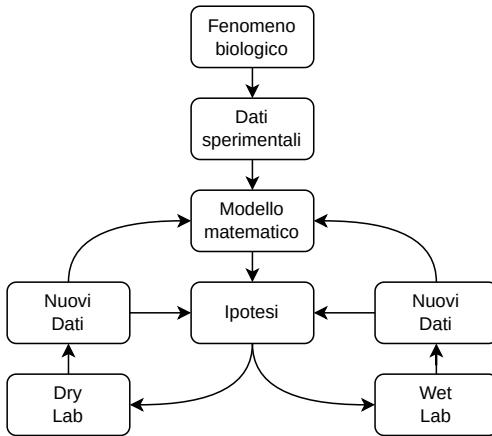


Figura 2.1: Grafico rappresentante il processo ciclico della Systems Biology.

Ai metodi standard della biologia quindi si aggiungono le teorie informatiche, quelle matematiche, quelle fisiche, quelle chimiche, quelle ingegneristiche. A partire dal fenomeno biologico quindi si effettuando esperimenti, ottenendo dei dati sperimentali relativi alle funzioni, alle strutture e alle interazioni delle varie componenti biologiche. A partire da questi dati si costruisce un **modello matematico** che porterà alla produzione di *ipotesi* a partire da esso. Inoltre l'insieme di ipotesi produrrà nuovi dati che potranno essere anche usati per rifinire il modello stesso. Inoltre tali ipotesi possono portare a sperimentazioni in **dry lab**, quindi “in silico” tramite simulazioni, ma anche in **wet lab**, quindi in laboratorio qualora possibile. Tali sperimentazioni contribuiranno a migliorare i dati stessi, producendone anche di nuovi. Si ha quindi un sistema ciclico di costante miglioramento della ricerca stessa, come visualizzabile in figura 2.1.

Un altro aspetto fondamentale del discorso è capire cosa **non** sia la *systems biology*. Citando Wolkenhauer¹:

Opening then the book, which I discovered in the London bookstore, I read the contents list: “Shotgun Fragment Assembly”, “Gene Finding”, “Local Sequence Similarities”, ... What?? ... “Protein Structure Prediction”, “Some Computational Problems Associated with Horizontal Gene Transfer” ... what on earth has this to do with systems biology, I asked myself?

...

Most important to me is however that cells and proteins are interacting in space and time, that is, we are dealing here with (nonlinear) dynamic

¹O. Wolkenhauer, Why Systems Biology is (not) called Systems Biology, BIOforum Europe 4/2007

systems. If you ask me then, systems biology is a merger of systems theory with cell biology.

...

Systems biology and bioinformatics are different but complementary.

Infatti tematiche come l'assemblaggio, l'allineamento etc... non sono tematiche della *systems biology* ma della *bioinformatica*, nonostante spesso vengano confuse e sovrapposte. L'analisi diretta dei dati biologici non è campo della *systems biology* in quanto si perde uno degli aspetti fondamentali, ovvero quello del **tempo**, che comporta lo studio di **sistemi dinamici**, che appunto di evolvono nel tempo. In bioinformatica d'altro canto si ha spesso a che fare con dati provenienti da pochi timestamp (se non direttamente da uno solo). Inoltre, sempre in bioinformatica, si studiano solitamente poche componenti biologiche, senza studiarne l'interazione tra esse.

La domanda più importante della *systems biology*, della quale possiamo vedere uno schema generale delle fasi in figura 2.2, è quindi:

dato un sistema biologico d'interesse, di cui si vogliono studiare le funzioni etc..., quale approccio modellistico è più adatto per descrivere quel sistema?

Una volta risposto a questo quesito bisogna ovviamente capire quale sia lo strumento computazionale di cui si ha bisogno per simulare e analizzare tale sistema. Bisogna infine capire quali predizioni si possono ottenere da questo modello, che comunque deve prima essere validato. Tra le cose principali che si vogliono capire abbiamo, ad esempio, se si può controllare il sistema e se si può riprodurre il tutto in laboratorio riducendo il numero di tentativi e di conseguenza anche il costo dell'esperimento in *wet lab*.

Possiamo quindi facilmente intuire che uno degli aspetti fondamentali di questo ambito è quello di fare le corrette *assunzioni*. Citando ancora Wolkenhauer²:

The modelling process itself is more important than the model. The discussion between the experimentalists and the theoretician, ro decide which variables to measure and why, how to formally represent interaction in a mathematical form is the basis for succesful interdisciplinary research in Systems Biology. In light of the complexity of molecular systems and the available experimental data, Systems Biology is the art of making the right assumptions in modelling.

si nota come il raggiungimento delle assunzioni stesse per ottenere il modello sia una fase di importanza maggiore rispetto al modello stesso. Il modello

²O. Wolkenhauer, Why Systems Biology is (not) called Systems Biology, BIOforum Europe 4/2007

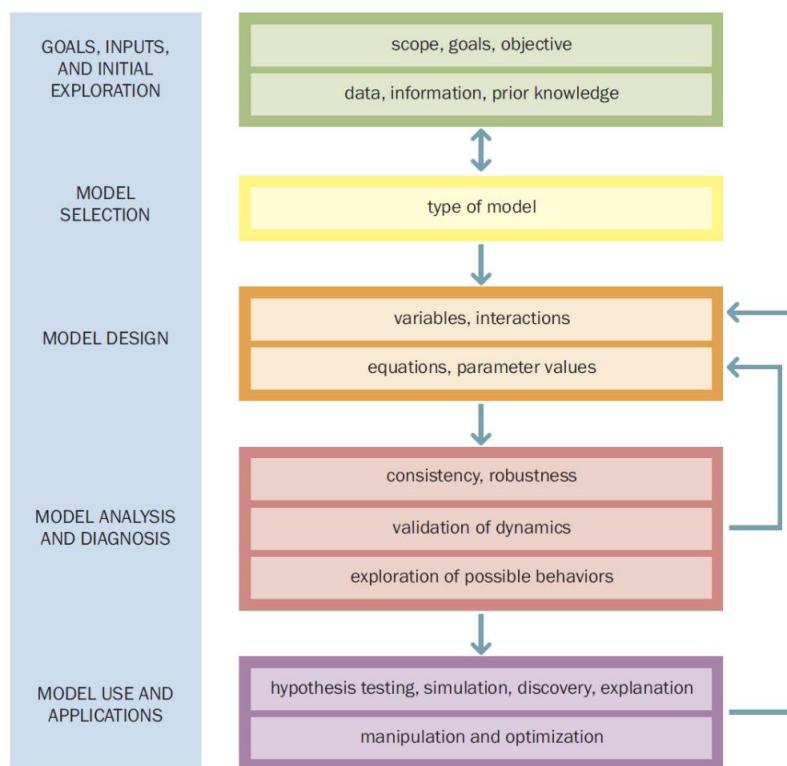


Figura 2.2: Schema generale delle fasi tipiche che compongono la systems biology.

infatti rappresenta la realtà ma non è la realtà stessa e partire da assunzioni false ed errate porterà ad un modello magari funzionante “dal punto di vista sintattico” ma non “dal punto di vista semantico”, avendo che esso non potrà mai essere validato. Nella citazione si parla inoltre di *variabili*, come elemento base dei vari modelli. Tra tali variabili si cercano relazioni, correlazioni etc... Normalmente il punto di partenza sono i *dati omici*.

Quanto qui riportato è tratto da wikipedia ³

Definizione 1. *In biologia molecolare, ci si riferisce comunemente al neologismo omica (in inglese omics) per indicare l'ampio numero di discipline biomolecolari che presentano il suffisso “-omica”, come avviene per la genomica o la proteomica. Il suffisso correlato -oma (in inglese -omes) indica invece l'oggetto di studio di queste discipline (genoma, proteoma).*

I più importanti “-oma” proposti recentemente all'interno della comunità scientifica sono:

- il **trascrittoma** è l'insieme degli mRNA trascritti nell'intero organismo, tessuto, cellula; è studiato dalla trascrittomica
- il **metaboloma** comprende la totalità dei metaboliti presenti in un organismo; è studiato dalla metabolomica
- il **metalloma** comprende la totalità delle specie di metalli e metalloidi; è studiato dalla metallomica
- il **lipidoma** comprende la totalità dei lipidi; è studiato dalla lipidomica
- l'**interattoma** comprende la totalità delle interazioni molecolari che hanno luogo in un organismo; un nome che comunemente indica la disciplina della interattomica è quello di biologia dei sistemi (systems biology)
- lo **spliceoma** (da non confondersi con lo spliceosoma, il complesso di proteine ed acidi nucleici coinvolti nello splicing) comprende la totalità delle isoforme proteiche dovute a splicing alternativo; è studiato dalla spliceomica
- l'**ORFeoma** comprende la totalità delle sequenze di DNA che iniziano con un codone ATG e terminano con un codone di stop (sequenze note come ORF, open reading frames). Queste sequenze

sono ritenute in grado di codificare per una proteina o per una parte

- **textoma:** l'insieme della letteratura scientifica disponibile alla consultazione (studiato dalla textomica)
- **kinoma:** l'insieme delle protein chinasi (dall'inglese kinase) di una cellula. Esistono pubblicazioni scientifiche che citano il termine kinomica
- **glicosiloma:** correlato alle reazioni di glicosilazione (studiato dalla glicosilomica)
- **fisioma:** correlato alla fisiologia (studiato dalla fisiomica)
- **neuroma:** l'insieme delle componenti nervose di un organismo (studiato dalla neuromica)
- **predittoma:** l'insieme delle predizioni di struttura proteica
- **reattoma:** l'insieme dei processi biologici
- **ionoma:** insieme dei nutrienti minerali e degli elementi in tracce che si trovano in un organismo
- **connettoma:** l'insieme di tutti i neuroni e le sinapsi di un cervello

Si hanno quindi vari “livelli” di studio, al variare dei dati omici, per i quali variano gli strumenti. Ad esempio:

- si ha il **genoma**, studiato tramite il *sequenziamento*, la *genotipizzazione* etc...
- il **trascrittoma**, ottenuto dopo la *trascrizione*, studiato tramite *microarrays*, *oligonucleotide chips* etc...
- il **proteoma**, ottenuto dopo la *traduzione*, studiato tramite *proteomica MS-based*, *elettroforesi* etc...
- il **metaboloma**, ottenuto tramite le *reazioni*, studiato tramite *spettroscopia di massa*, *risonanze magnetiche* etc...
- l'**interattoma**, ottenuto tramite appunto le varie *interazioni*, studiato tramite *screens yeast-to-hybrid* etc...

- il **fenomeno**, ottenuto dopo l'*integrazione* delle varie interazioni, studiato tramite *gene inactivations* etc...

Ognuno di questi “livelli” ha una panoramica diversa su quello che sta accadendo, è accaduto, potrebbe accadere o accadrà ad una certa cellula. Partendo dalle informazioni dinamiche/cinetiche, ovvero dai dati, e dalle informazioni strutturali dei vari *pathway* si riesce ad ottenere la rappresentazione matematica. Ovviamente è impensabile pensare di studiare tutti i “livelli” contemporaneamente ma si può studiare solo una parte del sistema, studiandone un paio di “livelli” o poco più. Inoltre ogni “livello” ha associato un suo formalismo matematico, legato alla singola modellazione matematica. Non sempre tali formalismi sono facilmente integrabili (magari in un caso ho delle EDO e in un altro dei grafi). Si ha quindi non solo un discorso di *data integration* ma anche di integrazione dei modelli matematici stessi e questo non sempre è possibile.

Nella realtà, inoltre, prima di scegliere un modello bisogna scegliere l'*approccio* con cui ottenerlo. Generalmente se ne hanno due in *systems biology*:

1. l’approccio **top-down**. In questo caso si parte dalle analisi omiche, solitamente con pochissimi timestamp, i cui risultati vengono trattati con tecniche bioinformatiche, che riducono anche l’influenza degli errori, per ottenere una **mappa globale di interazioni**, con le interazioni tra migliaia di componenti cellulari, dalla quale si ottiene il **modello predittivo del sistema**. Questo approccio è quindi supportato da una grande quantità di dati basati su *high-throughput* e *global profiling*
2. l’approccio **bottom-up**. In questo caso si parte dalle informazioni, prevalentemente di letteratura, le interazioni tra le componenti individuali del sistema, ceracondo magari le concentrazioni o il *kinetic-rate*, ovvero a variazione della concentrazione di un reagente o di un prodotto nel tempo misurata in moli per secondo $\left[\frac{M}{s}\right]$. Tali informazioni potrebbero non essere precise. Da queste si formalizza un modello matematico per avere poi comparazioni tra esperimenti e modelli di simulazione, ottenendo alla fine il **modello predittivo del sistema**. Questo approccio soffre quindi la mancanza di dati, specialmente di dati quantitativi. Questo approccio è più vicino a quello tipico della biologia, avvicinandosi per alcuni aspetti al *pensiero riduzionista* (che mira a studiare piccole componenti del sistema).

Tale approccio è sicuramente più complesso, per quanto si possa

limitare a studiare pathway e non l'intero metaboloma, ma per questo anche più informativo.

Ovviamente tali approcci, per quanto sarebbe fantastico, non possono essere usati in contemporanea. Detto questo solitamente l'approccio top-down studia i sistemi su larga scala per poi, a volte, procedere con uno studio bottom-up. In generale comunque la scelta dipende dalla singola situazione. Non esiste un meglio o un peggio, anche se i modelli generati dall'approccio top-down hanno generalmente una minor capacità predittiva anche se studiano sistemi più ampi rispetto all'approccio bottom-up.

Bisogna distinguere quindi quali siano le tecniche tipiche della bioinformatica (ma anche della statistica) e quali quelle della *systems biology*. L'uso di tecniche per la ricerca di similarità, correlazioni, causalità probabilistica, clustering (dove si noti che non ha un ruolo significativo il **tempo**) etc... non sono di interesse della *systems biology*, che invece è interessata allo studio delle causalità in cui il *tempo* è intrinseco e necessario. Questa necessità di avere il *tempo* comporta una maggior difficoltà nel recuperare i dati e dell'eseguire la sperimentazione ma comporta, del resto, un forte “potere di spiegazione e predizione” da parte del modello stesso.

Vediamo ora qualche definizione di base.

Definizione 2. *Definiamo **modello** come una descrizione rigorosa e assolutamente non ambigua di un sistema. Nel dettaglio tale descrizione è ottenuta tramite un adeguato formalismo matematico (l'unico per definizione non ambiguo) e un adeguato livello di astrazione (importante per non avere informazioni ridondanti o inutili nel modello).*

Definizione 3. *Definiamo **proprietà/comportamento emergente** ogni caratteristica strutturale (quindi di topologia) o dinamica (quindi in evoluzione nel tempo) di un sistema che non può essere capita e/o spiegata banalmente tramite l'enumerazione delle componenti ma che deve essere derivata unicamente come conseguenza tra le componenti stesse del sistema.*

Definizione 4. *Definiamo **simulazione** come una tecnica “computer-based” per determinare una qualsiasi caratteristica emergente e/o predire l’evoluzione temporale del sistema.*

Definizione 5. *Definiamo **metodo computazionale** come una soluzione automatica, basata su uno specifico algoritmo, usata per risolvere problemi difficili (da intendersi “difficili” anche a livello computazionale) e per analizzare sistemi in diverse condizioni.*

Si noti che, come evidenziato da Fawcett e Higginson⁴, l'uso eccessivo

⁴Tim W. Fawcett and Andrew D. Higginson, Heavy use of equations impedes communication among biologists, PNAS 2012

dei formalismi matematici rendono difficile la comunicazione con i biologi, quindi bisogna muoversi di conseguenza. I modellatori dovrebbero essere preparati a sviluppare nuovi strumenti matematici e computazionali, invece di “forzare” la descrizione e l’analisi del sistema con un framework preferito e facilmente applicabile (tipo usare le EDO per tutto a priori). I biologi sperimentali dovrebbero essere aperti a progettare nuovi protocolli di laboratorio per identificare tutte le caratteristiche qualitative e, soprattutto, quantitative che ancora mancano (per aiutare anche i modellisti). **La parte più interessante del gioco del modellismo non è ciò che il modello permette di capire, ma esattamente ciò che non è in grado di spiegare**, infatti, secondo, Box:

essentially, all models are wrong, but some are useful.

e, secondo Bower e Bolouri:

In fact, all modelers should be prepared to answer the question: “what do you know that you did not know before?” If the answer is “that i was correct”, it is best to look elsewhere.

Infatti un modello non solo deve rispondere a quello che già si sa ma deve predire qualcosa che ancora non si sa (magari anche non funzionando).

2.1 PCNA ubiquitylation

Vediamo brevemente uno studio in cui ha partecipato anche la professoressa Besozzi dove il non funzionamento del modello ha portato ad una nuova scoperta scientifica⁵.

In questo studio si cercava di studiare la **Post Replication Repair (PRR)**, ovvero il principale pathway di tolleranza al danno del DNA che bypassa le lesioni del DNA durante la *fase S*, che è in citologia (la branca della biologia che studia la cellula dal punto di vista morfologico e funzionale) una fase del ciclo cellulare, durante la quale il processo principale è la sintesi e duplicazione del materiale genetico contenuto nel DNA. Bombardando il lievito con raggi UV si è quindi studiata la proteina **PCNA**, ovvero l'*l'antigene nucleare di proliferazione cellulare*. La struttura di tale proteina (di forma a ciambella) è in grado di assumere una peculiare conformazione la quale le consente di contattare il DNA (DNA clamp) e di promuovere l’azione della

⁵Flavio Amara, Riccardo Colombo, Paolo Cazzaniga, Dario Pescini, Attila Csikász-Nagy, Marco Muzi Falconi, Daniela Besozzi, Paolo Plevani , In vivo and in silico analysis of PCNA ubiquitylation in the activation of the Post Replication Repair pathway in *S. cerevisiae*, BMC 2013

polimerasi durante la replicazione del DNA⁶. I raggi UV provocano lesioni che vengono “trattate” dalla PCNA. Se ne è quindi studiata l'**ubiquitazione**, modificazione post-traduzionale di una proteina dovuta al legame covalente di uno o più monomeri di ubiquitina. Tale legame porta, solitamente, alla degradazione della proteina stessa⁷. La *mono-ubiquitazione* avviene tramite gli enzimi *Rad6 Rad8* mentre la *poli-ubiquitazione* tramite gli enzimi *Rad5* e *Ubc13-Mms2*. La prima comporta errori di trascrizione, in quanto si aveva sintesi di DNA tra le lesioni, formando *mutageni*, mentre la seconda è “error free”.

Si conoscevano quindi i principali attori del fenomeno, ovvero la proteina e gli enzimi. C'erano varie cose che però non si conoscevano:

- l'ordine spazio temporale della cascata delle interazioni delle varie proteine, non sapendo anche i tempi di attivazione dei vari enzimi
- se il numero di lesioni influenzasse il bilanciamento tra le *mono-ubiquitazioni* e le *poli-ubiquitazioni*
- se esistesse una soglia relativa al danno che regolasse l'interazione tra i due sub-pathway

Si è quindi proceduto, in *wet lab*, irradiando il lievito in modo controllato, misurando *mono-ubiquitazioni* e le *poli-ubiquitazioni* al passare del tempo (da 0 a 300 minuti) a varie dosi di UV, e contemporaneamente studiando un modello matematico (tramite le varie reazioni, rappresentate tramite *prodotti* e *reagenti*) per effettuare le simulazioni. Si è visto, in laboratorio, che le varie forme ubiquilate di PCNA sono assenti a basse dosi di UV ($5 \frac{J}{m^2}$ e $10 \frac{J}{m^2}$), mentre ad alte dosi di UV ($50 \frac{J}{m^2}$ e $75 \frac{J}{m^2}$) entrambi i segnali sono ancora presenti dopo 5 ore nei *western blot*. La simulazione matematica confermava quanto stesse succedendo a bassi dosaggi ma non riusciva ad ottenere i risultati ad alti dosaggi. Dopo vari tentativi, rifacendo gli esperimenti (variando enzimi e geni) e sistemando il modello (tramite *parameter sweeping/estimation, analisi di sensitività*) si è sospettato che il modello fosse in realtà “corretto” ma non completo, mancava qualche ipotesi. Da qui la scoperta: si ha anche un altro pathway, il **Nucleotide Excision Repair (NER)** che “assiste” la *PCNA* quando le cellule sono gravemente lesionate. NER è infatti attivo nella *fase S* e serve alla *PRR* per funzionare correttamente *in vivo*. Risistemando il modello con *NER* ed enzimi annessi le simulazioni hanno funzionato.

⁶<https://it.wikipedia.org/wiki/PCNA>

⁷<https://it.wikipedia.org/wiki/Ubiquitina>

Questa è la prova che quando un modello non funziona si può ottenere anche una scoperta scientifica, ed è una delle situazioni (coi giusti limiti) più interessanti di questa branca di ricerca.

2.2 I Sistemi Complessi

In *systems biology* si ha quindi a che fare con sistemi che vengono definiti **sistemi complessi**, ovviamente presi nella loro “sottoclasse” relativa ai sistemi biologici.

Definizione 6. *Si definisce un **sistema complesso** con un sistema consistente di un certo numero di componenti più o meno semplici che, prese nel loro insieme, danno vita ad un comportamento emergente, grazie alle loro mutue interazioni.*

In questo contesto assumono importanza tre concetti chiave:

1. **comportamento non lineare**, quindi non facilmente prevedibile
2. **sistema aperto**, ovvero dove l’interazione con l’ambiente da parte del sistema è una delle caratteristiche da studiare e modellare
3. **sistema dinamico**, ovvero si ha che il sistema evolve nel tempo

Uno dei punti cruciali è inoltre capire che quando si procede alla modellazione di un certo sistema non si deve modellare anche cosa ci si aspetta da quel modello. Tale informazione infatti deve scaturire dalle simulazioni del modello stesso in modo completamente autonomo.

Come visto si studiano quindi insiemi di componenti. L’insieme complessivo delle funzionalità del sistema non è determinato però da una specifica funzione di ogni componente ma dalle loro interazioni. Si hanno quindi altri due concetti chiave:

1. **topologia/architettura interna**
2. **moduli funzionali**

Anche componenti molto semplici possono dare vita a un sistema complesso. Vediamo quindi qualche esempio di sistema complesso:

- un esempio “semplice” è quello di una **reazione enzimatica con feedback negativo**. In questo caso si hanno una serie di *reazioni lineari* che dal legame di un *substrato* portano ad un *prodotto*. La

complessità viene data dal *feedback negativo* in quanto la produzione del prodotto stesso porta il substrato a non legare. Si ha quindi la cosiddetta **autoregolazione** che rende questo un vero e proprio *sistema complesso*, avendo che il comportamento emergente del sistema è in realtà difficile da prevedere.

Ovviamente si potrebbe anche assumere il caso meno semplice dove si ha una serie di *reazioni non lineare*.

Si può quindi arrivare anche a parlare di casi più “estesi”, come quello ad esempio di un **pathway metabolico**. Ci sarebbe inoltre un caso, seguendo questo filo pensiero, ancora più estremo, ovvero quello del **metabolismo di un’intera cellula**, come visualizzabile nella figura 2.3, dove si hanno moltissime parti che nel dettaglio sono lineari ma che si autoregolano a vicenda, ottenendo quindi un *sistema complesso* davvero impossibile da studiare.

- un altro esempio può essere quello di una **rete di interazioni proteina-proteina (protein-protein interaction network)** dove si hanno:
 - **nodi** che rappresentano le proteine
 - **archi** che rappresentano **interazioni fisiche** e **interazioni funzionali** tra proteine (???)

In questo caso la “complessità” è data soprattutto dal numero incredibilmente alto di proteine (quindi di nodi) e di interazioni tra esse (quindi di archi) nel nostro sistema. Un esempio è visualizzabile in figura 2.4.

- un altro esempio è quello delle **reti di regolazione genica (gene regulatory network)** dove appunto si studiano le relazioni che si hanno tra le espressioni e le regolazioni tra geni. In questo caso si hanno:
 - **nodi** che rappresentano i geni
 - **archi** che rappresentano le regolazioni tra geni

Anche in questo caso si possono avere feedback e tali reti sono utili per studiare l'*over-espressione di geni*. Inoltre deve essere chiaro che la modifica in un certo gene si ripercuote, chi più chi meno, sull’intera rete anche se non si ha un singolo gene che “controlla” l’intera rete ma tutti contribuiscono alla funzionalità dell’intero sistema complesso

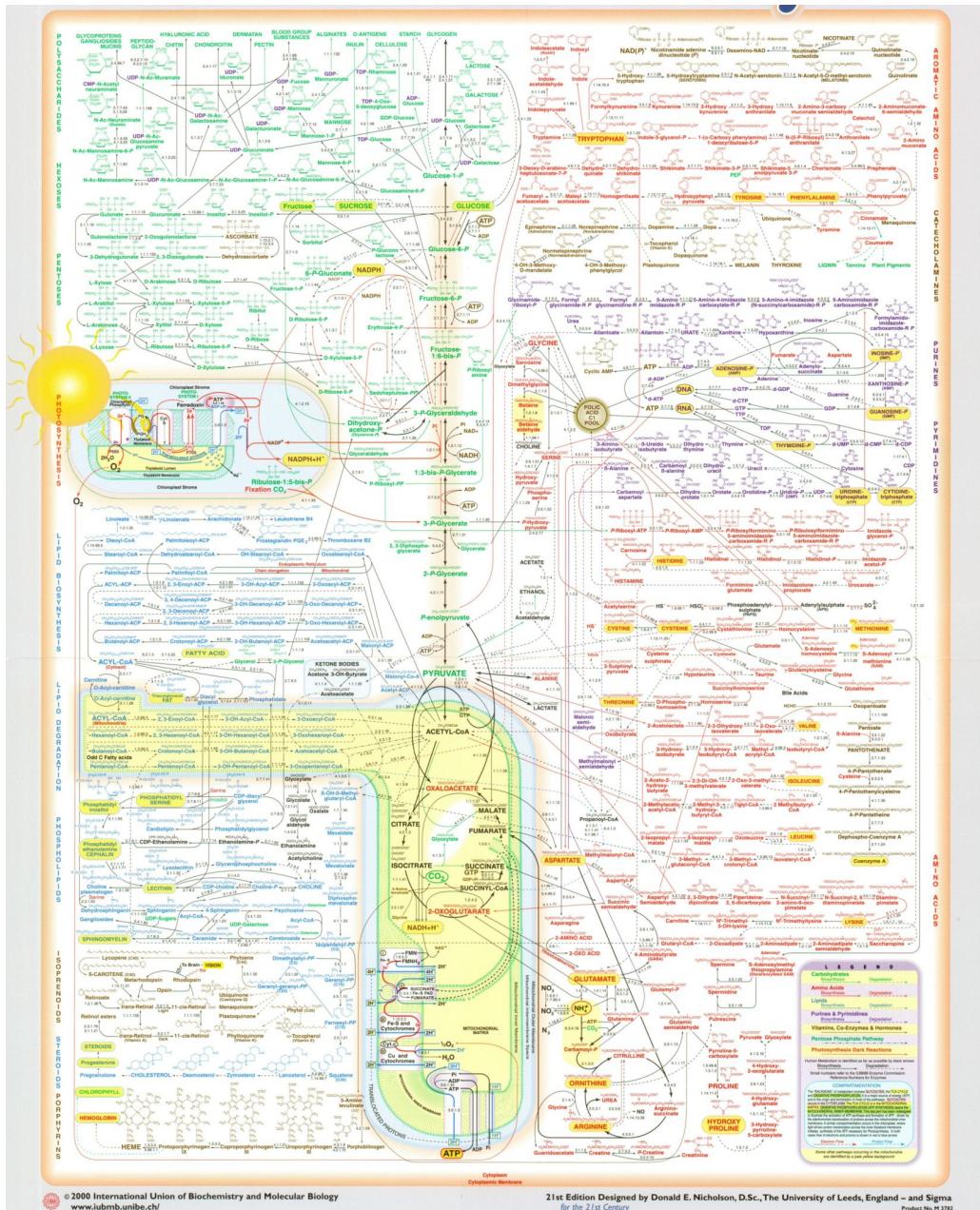


Figura 2.3: Insieme dei pathway che “compongono” il metabolismo di un’intera cellula. Tale rappresentazione è stata fatta da Donald E. Nelson, dell’università di Leeds e dall’azienda Sigma-Aldrich per la International Union of Biochemistry and Molecular Biology del 2000.

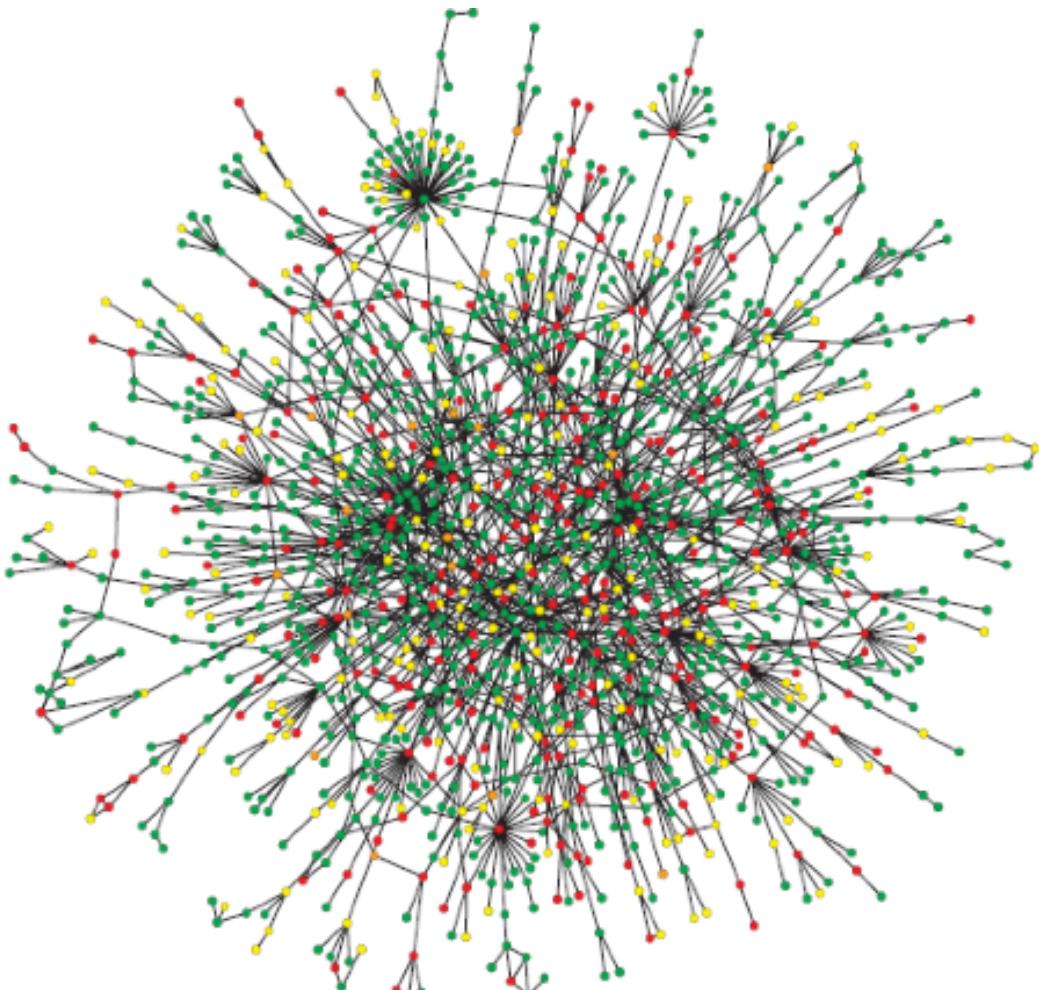


Figura 2.4: Esempio di rete di interazioni proteina-proteina, <https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/protein-protein-interaction-networks/>. In tale rete si ha lo studio sul lievito e i vari colori dei nodi rappresentano vari effetti fenotipici legati alla rimozione della proteina rappresentata dal nodo stesso. Si ha rosso per l'effetto letale, verde per l'effetto non letale, arancione per la crescita lenta e giallo per effetto sconosciuto.

- aumentando ancora il livello di complessità potremmo pensare allo studio di un certo pathway, come ad esempio il *segnale di trasduzione*, in una cellula tenendo però conto anche della *componente spaziale*, tridimensionale, della stessa, nonché le interazioni con l’ambiente. La componente spaziale, che ovviamente aggiunge complessità, è una parte rilevante del modello, come il “movimento” al suo interno (prestando sempre attenzione a non aggiungere componenti inutili al modello stesso). L’interazione con l’ambiente può portare, ad esempio, a *cascate di reazioni intra-cellulari* e a vari “input”, come *ormoni, fattori di sopravvivenza, fattori di crescita/anti-crescita, fattori di morte etc...* da considerare nel modello
- un altro esempio ancor più “complesso” può essere quello della **crescita tumorale**, magari ponendo al centro dello studio anche il rapporto tra essa e la **vascolarizzazione**, ovvero la distribuzione di vasi sanguigni in un tessuto, in quanto magari si vuole studiare la vicinanza tra il tumore e i vasi sanguigni. In questo contesto non solo lo spazio tridimensionale è di fondamentale importanza ma bisogna anche modellare cellule di vario tipo (normali, cancerogene, legate al sistema immunitario, in apoptosis, necrotiche etc...), che interagiscono in modo diverso tra loro, magari avendo anche “mutazioni” da normali a cancerogene etc... Si hanno quindi componenti eterogenee, dovendo per lo più anche modellare i vasi sanguigni e le interazioni con le cellule.
- un altro esempio, “complesso” almeno quanto il precedente, è lo studio della **formazione di biofilm**, ovvero una aggregazione complessa di microrganismi contraddistinta dalla secrezione di una matrice adesiva e protettiva. Tale barriera è comunque una struttura permeabile permettendo il passaggio dei nutrienti. In un biofilm i microrganismi, tendenzialmente batteri, non solo crescono ma, soprattutto quelli più interni e “protetti”, diventano anche più resistenti. Questa è una seria complicazione per la loro eliminazione quando fuoriescono dal biofilm. Anche qui quindi bisogna modellare lo spazio tridimensionale, l’interazione con l’ambiente, l’interazione tra i vari microrganismi (anche se si ha solitamente poca eterogeneità)
- cambiando prospettiva un altro esempio di *sistema complesso* è quello dello **sviluppo embrionale e della differenziazione cellulare** dove, a partire dall’embrione e da cellule staminali si vanno

a formare tutti i tipi di cellule che formeranno, ad esempio, i tessuti, gli organi etc... dell'uomo. In questo caso solitamente si ha un tipo di modellazione diverso, basato su componenti semplici

- un altro esempio è quello dello studio dell'**ecosistema**. Nel dettaglio uno degli aspetti studiati è quello della cosiddetta **dinamica preda-predatore**. Tale dinamica descrive il rapporto tra il numero di prede e di predatori e osserva un comportamento oscillatorio (se aumentano i predatori diminuiscono le prede fino a che non sono abbastanza per i predatori, che calano di numero, portando il numero di prede a crescere etc...)
- infine un ultimo esempio, molto attuale, di *sistema complesso* è quello dello **studio epidemiologico della diffusione di epidemie/pandemie** dove la “complessità” è incrementata anche dagli aspetti sociali e psicologici delle persone, nonché dalla loro eterogeneità anche nel dominio epidemiologico (infetti, gravemente infetti, guariti, esposti, suscettibili all'infezione, morti etc...)

In questi esempi si è spesso parlato più o meno esplicitamente di **livelli di complessità**. Per poter avere un'idea di quanti possano essere bisogna considerare vari punti di vista:

- un primo punto di vista è dato dalla **scala spaziale** dei fenomeni che si studiano. Possiamo studiare infatti eventi che accadono nel range dei nanometri, o meno, fino a pensare ad eventi in scala umana, in metri. Inoltre anche un evento che avviene in nanometri può avere conseguenze visibili in metri. Questo tipo di complessità è per lo più un problema matematico dal punto di vista della gestione della stessa
- un secondo punto di vista è dato dalla **scala temporale** dei fenomeni che si studiano. Anche in questo caso si passa dai nanosecondi, o meno, ai miliardi di anni. Un evento quasi istantaneo può avere conseguenze evolutive tra milioni di anni. La gestione di questo tipo di complessità è un grande problema dal punto di vista computazionale. La complessità aumenta all'aumentare della scala temporale
- altri livelli di complessità sono dati dai *livelli di funzione di un organismo*, avendo, ad esempio, che da *trascrittoma*, *proteoma* e *metaboloma*, in ottica pathway, si passa al *fisioma*, in ottica cellule, tessuti, organi e direttamente l'uomo

Pensando anche solo alla scala spaziale e quella temporale si ha inoltre che esse sono in sinergia ma è comunque pressoché impossibile pensare ad un modello che tenga traccia in modo completo o quasi di entrambe queste scale.

2.3 Rappresentazione Grafica

Ai biologi/biotecnologi etc... piace fare diagrammi e mappe concettuali per rappresentare graficamente le conoscenze biologiche che hanno su un sistema, ad esempio componenti molecolari e le loro mutue relazioni, formazione di complessi molecolari, presenza di feedback di regolazione positivi/negativi etc.... Come si intuisce facilmente diagrammi di questo tipo sono soggetti ad interpretazioni ambigue e limitano anche l'esplicita rappresentazione della conoscenza biologica. La matematica è l'unico linguaggio non ambiguo e fortunatamente esistono anche formalismi, come i *grafi*, le *reti di Petri* etc... che non solo sono formalmente rigorosi ma hanno anche un'interpretazione grafica (tanto amata dalle persone). Ovviamente non sempre si hanno queste soluzioni intermedie. La modellazione matematica risolve ogni errata interpretazione e descrive in modo non ambiguo quello che accade nel sistema e può potenzialmente includere ogni tipo di ipotesi che può poi essere studiata e testata in *wet lab*. In ogni caso i diagrammi possono avere utilità nella fase preliminare di discussione tra il biologo e il modellista: può essere un buon punto di partenza ma non sarà mai sufficiente per modellare il sistema, che si può ottenere solo con la formalizzazione matematica di componenti e interazioni. Un esempio è visualizzabile in figura 2.5⁸.

2.4 Tipologie di Modelli

Sistemi biologici differenti necessitano di approcci modellistici differenti, ovvero di framework matematici, quindi ad un preciso formalismo, e conseguentemente computazionali diversi. Inoltre bisogna sempre tenere in considerazione che ogni metodo computazionale legato ad un preciso modello può rispondere solo a certe tipologie di domande. Non si ha però una corrispondenza biunivoca tra ogni approccio modellistico e ogni sistema biologico, infatti diversi modelli potrebbero prestarsi bene ad un certo sistema biologico (anche se alcuni modelli sono inapplicabili per certi sistemi biologici o per

⁸Besozzi D. (2016) Reaction-Based Models of Biochemical Networks. In: Beckmann A., Bienvenu L., Jonoska N. (eds) Pursuit of the Universal. CiE 2016. Lecture Notes in Computer Science, vol 9709. Springer, Cham. https://doi.org/10.1007/978-3-319-40189-8_3

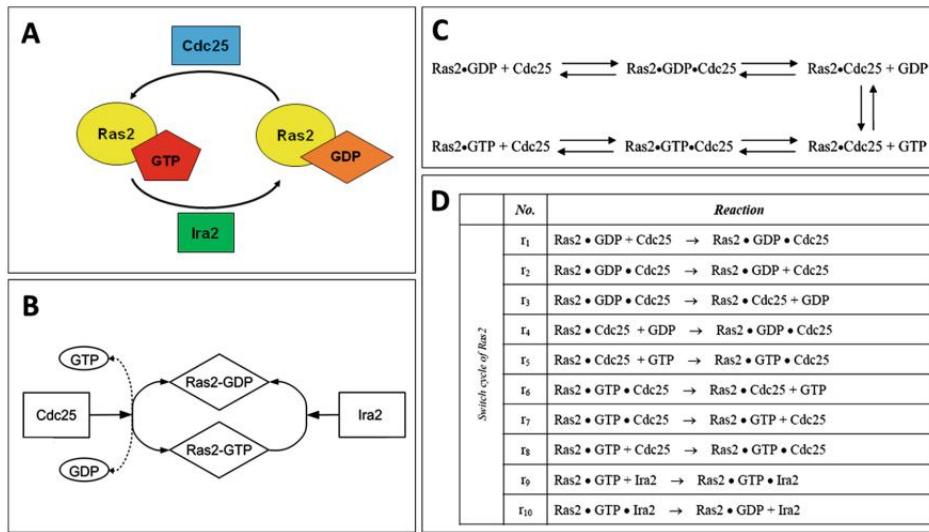


Figura 2.5: Esempio (senza entrare nei dettagli biologici che sarebbero ora superflui) di un diagramma ambiguo, in figura A, tipico dell’approccio biologico. Si hanno poi successive migliorie formali fino ad arrivare al modello matematico preciso, in figura D, e non ambiguo formato da 10 reazioni biochimiche.

certe domande su tali sistemi). La scelta del modello è quindi fortemente legata alle entità che si vogliono rappresentare e alle risposte che si vogliono ottenere dal modello. **Non si ha una strategia universalmente valida per scegliere il miglior approccio modellistico in base al sistema biologico d’interesse.**

Il primo passo è quindi l’interazione tra il biologo/biotecnologo e il modellista. Il primo deve porsi varie domande tra cui:

- cosa si sa e cosa non si sa del sistema biologico in questione?
- che tipologie di dato di laboratorio sono disponibili?
- che tipologie di dato posso misurare effettivamente in laboratorio?

Anche il modellista quindi si deve porre delle domande fondamentali, tra cui:

- quale formalismo matematico si presta meglio per questo problema?
- che strumenti computazionali sono necessari?
- che tipo di predizioni mi aspetto dal modello?

Queste questioni sono “in ciclo” tra di esse e sono la base degli studi in *systems biology*, dove farsi domande è una parte fondamentale. In merito all’ultima domanda del biologo è interessante notare che un modello **dove** essere validato in laboratorio. Qualora non sia possibile, ad esempio un “caso limite” emerso dallo studio del modello, non si può fare nulla (anche se, qualora si avessero più modelli completamente distinti che portano allo stesso risultato si può presupporre che ci sia un fondo di verità).

La domanda fondamentale resta però:

qual è la questione scientifica? Perché mi serve un modello?

La risposta a questa domanda deve essere “sicura” prima di intraprendere uno studio di modellazione.

Nel dettaglio, durante il corso, si vedranno i quattro approcci modellistici tradizionali più usati anche se si tratta di una selezione tra la moltitudine degli approcci presenti:

1. **modelli basati su interazioni** (*interaction-based models*)
2. **modelli basati su vincoli** (*constraint-based models*)
3. **modelli logici** (*logici-based models*)
4. **modelli meccanicistici** (*mechanism-based models*)

Un generale dato un certo sistema biologico d’interesse, dopo aver risposto alla domanda fondamentale e avendo quindi ben chiaro il fine di tale modello, la scelta del modello stesso viene presa considerando secondo quattro aspetti fondamentali:

1. la **dimensione del sistema**, data in primis dal numero di componenti e dal numero delle interazioni tra esse. Si distinguono, secondo questo aspetto, due grandi macro-categorie di sistemi:
 - (a) **sistemi small-scale**, se siamo nel range delle unità o delle decine di componenti/interazioni
 - (b) **sistemi large-scale**, se siamo nel range delle centinaia o migliaia (se non oltre) di componenti/interazioni

Questo è già un ottimo fattore discriminatorio per la scelta del sistema

2. il **livello di dettaglio** necessario a descrivere in modo completo le componenti del sistema e le loro interazioni. Si ricorda sempre però che formalizzare informazioni inutili e/o ridondanti comporta solo un’inutile spreco dal punto di vista computazionale

3. il **tipo** e la **qualità** dei **dati sperimentali** che sono già disponibili o che si è in grado di produrre all'evenienza con precisi protocolli al fine di supportare la fase di modellazione. Tali dati possono essere ad esempio *dati omici, western blots etc...*
4. il **carico computazionale** che l'approccio scelto comporta in fase di simulazione e analisi dei dati. Un esempio è quello della *dinamica molecolare*, che studia come interagiscono tra loro più molecole (o anche il comportamento interno di una sola). Tali studi normalmente impiegano settimane per simulare anche range temporali molto ridotti e necessitano di molte informazioni che non possono essere trascurate per ottenere un modello ed una simulazione realistici. Questa scelta è spesso un trade-off nella scelta di **approcci modellistici quantitativi** e **approcci modellistici qualitativi**.
L'uso di **super computer**, di **tecniche di calcolo parallelo su GPU** etc... sono molto comuni in *systems biology*

Una misura fondamentale è poi la **capacità predittiva del modello** che, se bassa, ci porta a preferire *modelli qualitativi*, se alta invece a *modelli quantitativi*.

Si ha quindi un comodo grafico che classifica le quattro tipologie di modelli in base a questi quattro aspetti⁹ in figura 2.6. Nel grafico notiamo come i *modelli meccanicistici*, tra quelli analizzati, siano quelli con il più alto potere predittivo, che è un aspetto fondamentale ma anche con i più alti livelli di dettaglio, costi computazionali e sfide nella misurazione dei dati richiesti. L'ovvia conseguenza è che la dimensione del sistema da studiare deve essere ridotta, avendo quindi *sistemi small-scale*.

Si noti che collocare i *modelli logici* non sia così banale in quanto il livello di dettaglio e la facilità di misurazione possono essere migliorati usando ad esempio le **logiche fuzzy**.

Un altro aspetto fondamentale da tenere in considerazione è che il processo di modellazione richiede molto tempo e si basa su continui raffinati del modello stesso, in un processo circolare, come visualizzabile in figura 2.7¹⁰. Quindi partendo dai dati biologici si abbozza un primo modello che viene

⁹Besozzi D. (2016) Reaction-Based Models of Biochemical Networks. In: Beckmann A., Bienvenu L., Jonoska N. (eds) Pursuit of the Universal. CiE 2016. Lecture Notes in Computer Science, vol 9709. Springer, Cham. https://doi.org/10.1007/978-3-319-40189-8_3

¹⁰Chou IC, Voit EO. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Math Biosci.* 2009;219(2):57-83. doi:10.1016/j.mbs.2009.03.002

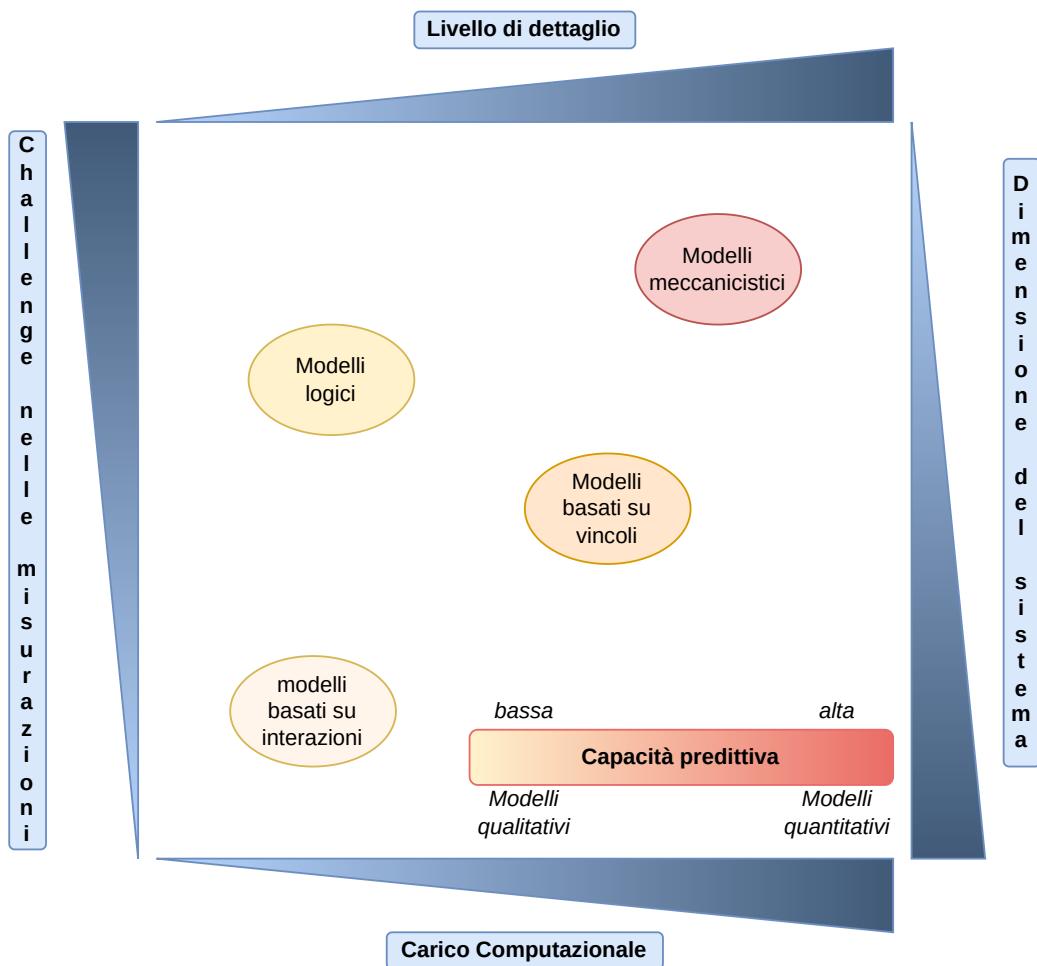


Figura 2.6: Schema riassuntivo dei quattro approcci

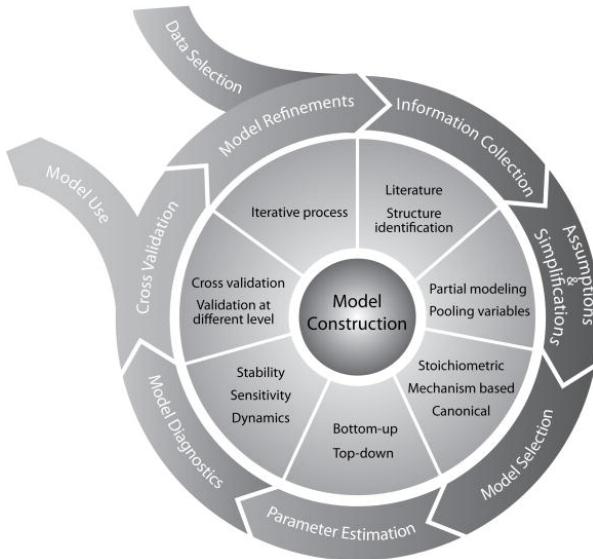


Figura 2.7: Raffigurazione che mostra i dettagli del processo ciclico di modellazione in *systems biology*. Molte, ma non tutte, delle keyword presenti verranno approfondite nel corso

poi continuamente sistemato tramite nuove ipotesi, altri dati, analisi *in silico* etc... fino all'ottenimento di un modello validato. Vediamo ora una breve introduzione ai quattro approcci elencati al fine di poter fare un confronto tra essi prima di studiarli e formalizzarli nel dettaglio.

Prima di fare ciò diamo una più precisa idea dei criteri con cui si classifica un modello.

Definizione 7. Si definisce **modello qualitativo** un modello che specifica le interazioni tra le componenti del modello stesso.

Definizione 8. Si definisce **modello quantitativo** un modello che assegna un valore ad ogni elemento che descrive e anche alle interazioni tra essi. In questo caso si possono avere o non avere cambiamenti nel modello.

Definizione 9. Si definisce **modello deterministico** un modello per il quale l'evoluzione attraverso i vari stati può essere predetta a partire dallo stato corrente, nel dettaglio anche dallo stato iniziale. Il comportamento evoluzionario del modello quindi non varierà tra una simulazione e l'altra.

Definizione 10. Si definisce **modello stocastico** un modello che descrive, a partire da uno stato corrente, uno stato futuro attraverso una distribuzione di probabilità.

Definizione 11. Un processo è detto **reversibile/irreversibile** se si può o meno procedere in avanti o indietro tra i vari stati.

Definizione 12. Con il termine **periodicità** si intende che il sistema assume una serie di stati nell'intervallo di tempo $[t, t + \Delta t]$ ma anche in:

$$[t + i\Delta t, t + (i + 1)\Delta t], \quad i = 1, 2, 3, \dots$$

2.4.1 Modelli Basati su Interazioni

Questo tipo di modelli vengono usati per *sistemi large-scale* con centinaia o migliaia di componenti che interagiscono tra loro in modo fisico o funzionale. Abbiamo vari esempi di questi modelli, tra cui:

- **reti di interazioni proteina-proteina**
- **reti di regolazione genica**
- **reti metaboliche**
- **reti di malattie**, modelli più complessi, modellati tramite un particolare tipo di grafo, che sfruttano l'integrazione tra *reti di regolazione genica* e grafi/reti rappresentanti le malattie e le relazioni tra esse. Si ottiene quindi un grafo che mette in relazione componenti genomiche e malattie

In questo caso la scelta del formalismo matematico ricade principalmente sulla **teoria dei grafi** e si ha quindi un *modello qualitativo e statico*. Infatti il *tempo* non viene considerato in tali modelli, che di conseguenza non permettono di ottenere informazioni su eventuali **comportamenti emergenti**. Non si possono nemmeno ottenere informazioni quantitative.

Parlando di *modelli basati su interazioni* non si può propriamente parlare di “simulazioni” vere e proprie in quanto in primis manca la modellazione del *tempo* ma anche di altri fattori come il *kinetic rate*. Inoltre tali modelli difettano anche di una qualsivoglia modellazione dello *spazio*. Il fulcro dello studio di tali modelli quindi solitamente si concerta sulle proprietà “architettoniche” della struttura della rete, studiando, ad esempio:

- la presenza di **hub**, ovvero nodi in cui sono entranti/uscenti un gran numero di archi rispetto agli altri nodi della rete
- misure di centralità
- presenza di *motivi (motifs)* nella rete

- la robustezza topologica

Tutte queste misure permettono anche di caratterizzare, caratterizzando la topologia stessa, una rete rispetto ad un'altra. Infatti si vedranno vari tipologie di rete, tra cui:

- **random network**
- **scale-free network**, caratterizzate da una forte *robustezza*
- **hierarchical network**

2.4.2 Modelli Logici

Questi modelli possono essere usati sia per sistemi *small-scale* che per sistemi *large-scale* e alcuni degli esempi sono:

- **reti di regolazione gene-gene**
- **pathway di trasduzione del segnale** (che si ricorda essere la capacità di una cellula di convertire uno stimolo esterno in una particolare risposta cellulare)
- **differenziazione cellulare**
- **pathway per la morte cellulare programmata**

Il primo caso è un esempio di *sistema large-scale* mentre gli altri di *sistemi small-scale*.

Dal punto di vista del formalismo matematico si ha anche qui la **teoria dei grafi**, a cui viene aggiunta la **logica booleana**, con i classici operatori logici \neg, \wedge, \vee , o anche, preferibilmente, la **logica fuzzy**, che verrà approfondita più avanti. L'idea di base è quella che lo stato delle componenti è regolato da altre componenti del sistema stesso. I nodi possono assumere o valore booleano 0/1 o, in logica fuzzy, qualsiasi valore tra 0 e 1 (con varie conseguenze nel loro studio).

Tali modelli sono in grado di simulare il tempo, rientrando quindi nella categoria dei *sistemi dinamici* ma sono anch'essi della tipologia dei *modelli qualitativi*. Tali sistemi si prestano ad essere sia *deterministici* che *non deterministici*.

Lo studio di tali modelli solitamente consiste, tramite le simulazioni e le analisi, nel determinare:

- **cicli**, ovvero sequenze finite di stati complessivi del sistema che si ripetono

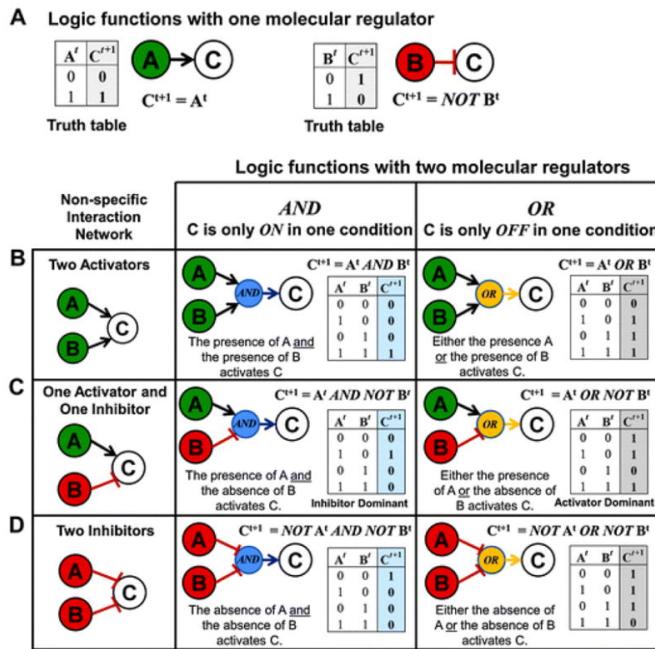


Figura 2.8: Esempio di modellazione di interazioni tra componenti tramite funzioni logiche.

- **attrattori**, ovvero degli *stati finali* che sono raggiungibili da qualsiasi stato iniziale e una volta raggiunti si resta in tali stati
- **bacini di attrattori**, ovvero percorsi che partono da stati intermedi e che conducono a degli *attrattori*

Tendenzialmente si arriva sempre ad un *ciclo* o ad un insieme di *attrattori*. La “potenza” della *logica fuzzy* permette, come detto, anche di modellare il *tempo*, derivando quindi un comportamento dinamico del sistema, descrivendo, ad esempio, la variazione nel tempo tra i valori degli stati di ogni componente.

Un esempio semplice di quello che si può ottenere con tali modelli è visualizzabile in figura 2.8¹¹.

2.4.3 Modelli Meccanicistici

Come già anticipato tali modelli si limitano a descrivere *sistemi small-scale*. Questa è la classe di approcci modellistici più complessa ed eterogenea infatti,

¹¹Wynn ML, Consul N, Merajver SD, Schnell S. Logic-based models in systems biology: a predictive and parameter-free network analysis method. Integr Biol (Camb). 2012;4(11):1323-1337. doi:10.1039/c2ib20193c

in primis, richiede una parametrizzazione completa delle componenti, con un ampio range di formalismi matematici, tra cui spiccano tra gli altri i **metodi numerici** e gli **algoritmi di simulazione stocasitca**. Un problema, già solo a questo punto, è che non si hanno spesso i dati per effettuare la parametrizzazione in quanto i biologi/biotecnologi spesso non sono interessati a misurarli.

Le simulazioni con questi modelli sono usate per studiare l'**evoluzione nel tempo**, quindi la dinamica, del sistema. Si usano metodi *deterministici*, *stocastici* e *ibridi*, insieme ad una serie infinita di altre tecniche computazionali, tra cui l'*analisi di sensitività* o il *parameters sweeping*.

Si arriva quindi a modelli *quantitativi* e *dinamici*.

La scelta tra metodi stocastici, solitamente più dispendiosi, e deterministici dipende anche dal fatto che **la vita non è deterministica**. Ad esempio modellare le interazioni tra, ad esempio tra la proteina *Mdm3* e la proteina *p53*, avendo che la prima inibisce la seconda, comporta una funzione molto pulita se studiata in modo deterministico quando in realtà, a causa di molti fattori, non si ha tale precisione se si va a studiare cosa accade realmente in natura. Da qui l'uso anche di *modelli stocastici*.

La stima dei parametri resta comunque un grandissimo problema e spesso si usano altre tecniche computazionali/modellistiche in pipeline per inferire gli stessi.

2.4.4 Modelli Basati su Vincoli

Tali modelli sono usati esplicitamente e solo per *sistemi large-scale per reti metaboliche*. Il formalismo matematico qui usato si compone di **matrici stoichiometriche**, **algebra lineare** e tecniche di **ricerca operativa** mentre le simulazioni e le analisi consistono nello studiare le variazioni nelle **distribuzione di flusso**, calcolando i valori di flusso di tutte le reazioni metaboliche, a seconda di perturbazioni/input prefissati.

Non è semplice se si può dire di ottenere dei *sistemi quantitativi* in quanto si studia il comportamento ad uno *steady state*.

Tra gli esempi di uso si hanno:

- l'**ingegnerizzazione metabolica**, ovvero l'ottimizzazione, il design e la regolarizzazione di certe strutture/funzioni metaboliche al fine di ottenere un certo fenotipo metabolico
- studiare **bersagli di farmaci**, attraverso ad esempio lo studio del *rewiring metabolico del cancro*

L'idea è quindi quella di:

- stabilire dei **vincoli**
- stabilire una **funzione obiettivo** da *massimizzare/minimizzare*
- determinare automaticamente la distribuzione dei flussi

Si parla di **Flux Balance Analysis (FBA)**.

2.4.5 Confronto tra i Vari Approcci

Viste queste prime piccole premesse sui quattro approcci possiamo fare qualche piccolo confronto.

In primis abbiamo capito come lo studio del *tempo* sia assente del tutto nei *modelli basati su interazioni* e che sia di dubbio uso nel caso dei *modelli basati su vincoli* a causa dello *steady state*. Quindi se si dovesse, ad esempio, studiare il cambio di concertazione di una certa molecola al variare del *tempo* tali approcci sarebbero da scartare a priori.

Si è anche visto come, in realtà, praticamente solo i *modelli meccanicistici* ci offrono uno studio quantitativo, al costo di una complessità sia formale, che di dati, che computazionale molto alta e riducendosi a studiare solo sistemi piccoli. Ne segue che:

I sistemi meccanicistici sono l'approccio modellistico migliore per comprendere e acquisire nuove intuizioni il funzionamento del sistema.

Nella realtà però “avere tutto” è un’utopia quindi non si ha un vero e proprio vincitore in questa “gara tra approcci modellistici”, ben riassunti nella figura 2.9¹², in quanto al variare del problema, dei dati, e di mille altri fattori potrei aver motivi validi per preferire un approccio ad un altro.

Inoltre, in questa breve introduzione, si è scoperto come ci siano moltissime **dicotomie** in *systems biology*:

- *top-down* e *bottom-up*
- *qualitativo* e *quantitativo*
- *statico* e *dinamico*
- *deterministico* e *stocastico*
- *discreto* e *continuo* (sia in ottica di rappresentazione del *tempo* che della numerazione delle componenti)

¹²Bordbar, A., Monk, J., King, Z. et al. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15, 107–120 (2014). <https://doi.org/10.1038/nrg3643>

Method	Model systems	Parameterization	Typical prediction type	Advantages	Disadvantages
Stochastic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	<ul style="list-style-type: none"> Mechanistic Dynamic Captures biological stochasticity and biophysics 	<ul style="list-style-type: none"> Computationally intensive Difficult to parameterize Challenging to model multiple timescales
Deterministic kinetic modelling	Small-scale biological processes	Detailed kinetic parameters	Reaction fluxes, component concentrations and regulatory states	<ul style="list-style-type: none"> Mechanistic Dynamic 	<ul style="list-style-type: none"> Computationally intensive Difficult to parameterize
Constraint-based modelling	Genome-scale metabolism	Network topology, and uptake and secretion rates	Metabolic flux states and gene essentiality	<ul style="list-style-type: none"> Mechanistic Large scale No kinetic information is required 	<ul style="list-style-type: none"> No inherent dynamic or regulatory predictions No explicit representation of metabolic concentrations
Logical, Boolean or rule-based formalisms	Signalling networks and transcriptional regulatory networks	Rule-based interaction network	Global activity states and on-off states of genes	Can model dynamics and regulation	Biological systems are rarely discrete
Bayesian approaches	Gene regulatory networks and signalling networks	High-throughput data sets	Probability distribution score	<ul style="list-style-type: none"> Non-biased Can include disparate and even non-biological data Takes previous associations into account 	<ul style="list-style-type: none"> Statistical Issues of over-fitting Requires comprehensive training data
Graph and interaction networks	Protein–protein and genetic interaction networks	Interaction network that is based on biological data	Enriched clusters of genes and proteins	<ul style="list-style-type: none"> Incorporates prior biological data Encompasses most cellular processes 	Dynamics are not explicitly represented
Pathway enrichment analysis	Metabolic and signalling networks	Pathway databases (for example, KEGG, Gene Ontology and BioCyc)	Enriched pathways	<ul style="list-style-type: none"> Simple and quick Takes prior knowledge into account 	<ul style="list-style-type: none"> Biased to human-defined pathways Non-modelling approach

Figura 2.9: Schema riassuntivo delle caratteristiche dei vari approcci.

- *omogeneo e eterogeneo*
- *a singolo volume e multicompartmentale*

Tutte queste dicotomie rappresentano la complessità degli studi in *systems biology*.

Integrare i vari modelli è per lo più utopia. Fare *data integration* è già di per se uno scoglio complesso ma si aggiunge anche la difficoltà di integrare vari formalismi matematici. Non si ha il modello “perfetto” ma si può scegliere bene in base al sistema biologico da studiare, magari integrando anche qualche (molto pochi) approccio modellistico diverso, come visibile in figura 2.10¹³. Nel diagramma si segnala il paper centrale di Karr et al.: *A whole-cell computational model predicts phenotype from genotype*¹⁴ cruciale nello studio di un approccio misto per modellare il sistema un’intera cellula di un piccolo batterio.

¹³Gonçalves E, Bucher J, Ryll A, et al. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. Molecular Biosystems. 2013 Jul;9(7):1576-1583. DOI: 10.1039/c3mb25489e. PMID: 23525368.

¹⁴Karr JR, Sanghvi JC, Macklin DN, et al. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150(2):389-401. doi:10.1016/j.cell.2012.05.044

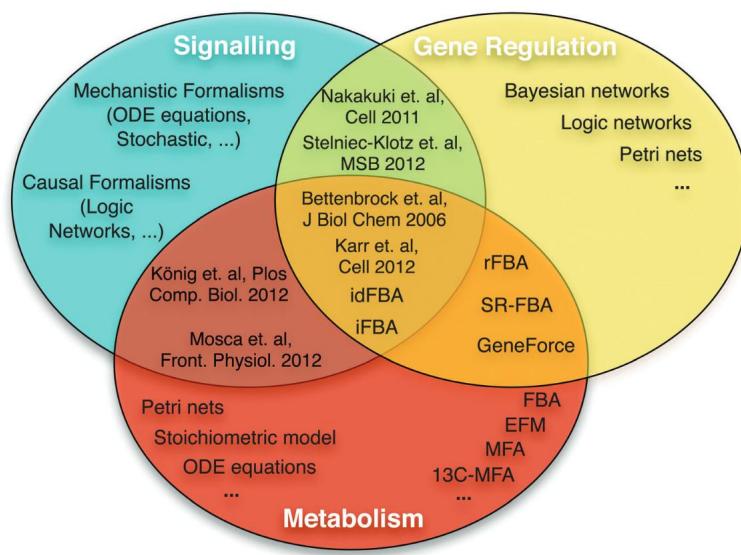


Figura 2.10: Diagramma che mostra varie soluzioni modellistiche al variare del problema biologico

Capitolo 3

Interaction-Based Modelling

Si parte con la descrizione della prima classe di modelli, parlando quindi della **interaction-based modelling**.

Ricordiamo che tali modelli, come visibile in figura 2.6:

- hanno un sistema di grandi dimensioni, con centinaia o migliaia di componenti (se non di più), essendo quindi *modelli large-scale*
- presentano tendenzialmente un basso costo computazionale per le analisi
- hanno un basso livello di dettaglio
- non presentano particolari difficoltà nella misurazione dei dati

Inoltre, sempre ricordando l'introduzione alla classe, questo approccio modelistico:

- non permette propriamente di parlare di simulazioni non modelando *tempo*, *localizzazione spaziale* e *kinetic-rates*
- indirizzano l'analisi verso lo studio topologico della rete
- si studiano proprietà emergenti prettamente strutturali quali la presenza di *hubs*, misure di centralità, presenza di *motifs* e *robustezza topologica* contro certe *perturbazioni*

Abbiamo quindi a che fare con **modelli qualitativi e statici**.

Come anticipato possiamo avere vari tipi di relazione tra i nodi della rete in questo modello, ovvero vari tipi di **interazioni**.

Vediamo qualche esempio¹ (*nel corso si approfondiranno solo i primi tre esempi, a livello genico e proteico*):

- **interazioni fisiche**, ovvero interazioni che si verificano tra biomolecole a diretto contatto. Ad esempio, le reti proteina-proteina con tali interazioni sono importanti in processi come la formazione di complessi proteici, la trasduzione del segnale e il trasporto. Sono tendenzialmente il caso di studio più semplice
- **interazioni di regolazione**, ovvero interazioni che sono eventi di attivazione o inibizione diretta. Ad esempio, nella regolazione dell'espressione genica, un fattore di trascrizione è collegato ai suoi bersagli da archi diretti nella rete. Quindi posso quindi avere sia *regolazioni positive* che *regolazioni negative* e non si hanno più *interazioni fisiche*
- **interazioni genetiche**, ovvero interazioni che connettono geni la cui simultanea perturbazione genetica porta a un risultato fenotipico diverso da quello previsto dalla combinazione di singoli effetti. Ad esempio, le interazioni letali sintetiche collegano i geni che influenzano debolmente la vitalità dell'organismo quando eliminati individualmente, ma sono letali quando eliminati in combinazione. Le *interazioni genetiche* sono utili per studiare la funzione genica e per identificare complessi e pathway che lavorano insieme per controllare le funzioni essenziali
- **relazioni di similarità**, avendo collegamenti tra oggetti biologici che sono *simili* secondo un attributo comune. È possibile utilizzare molte diverse misure di somiglianza, come la somiglianza della sequenza proteica o la coespressione genica basata su profili trascrizionali correlati (avendo sia correlazioni positive che negative). Le relazioni di somiglianza sono utili per identificare gruppi di geni o proteine funzionalmente correlati. Un altro esempio è quello di studiare se, in una certa condizione data, si possono identificare geni indipendenti con un profilo di espressione (ovvero descrizione qualitativa e quantitativa dell'insieme dei geni trascritti in un dato momento da una cellula o da un tessuto, studiabile tramite, ad esempio, l'uso dei *microarrays*)

¹Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. Nat Biotechnol. 2009 Oct;27(10):921-4. doi: 10.1038/nbt.1567. PMID: 19816451; PMCID: PMC4154490.

Quindi in generale si hanno tipi modelli di reti di interazioni associati ai vari tipi di interazione (fisica, funzionale, genica, etc...), ad esempio²:

- **Association networks**, che modellano qualsiasi tipo di relazione tra molecole, ad esempio legami, coespressione e somiglianze strutturali. Esempi di tali reti sono le **gene co-expression networks** e le **protein similarity networks**
- **Functional networks**, che modellano le relazioni funzionali tra coppie di molecole (solitamente geni o proteine). Un collegamento implica che entrambe le componenti sono coinvolte nella stessa funzione, processo o fenotipo. Un esempio sono le **genetic interaction networks** rappresentano interazioni in cui una doppia mutazione porta a un effetto epistatico (che si ha quando una coppia di alleli copre l'espressione fenotipica di un'altra coppia di alleli), ad esempio peggiore o migliore del previsto rispetto alla singola mutazione
- **Protein-Protein Interaction (PPI) networks**, che sono reti non dirette che modellano il legame proteico tra le componenti, che sono appunto proteine. Tali reti sono derivate da esperimenti ad alto rendimento che utilizzano tecniche come lo *screening di due ibridi di lievito*, la *spettrometria di massa* e la *purificazione per affinità tandem*, che sono metodi *high-throughput*. Le **signaling networks** sono correlate alle reti PPI ma in questo caso i collegamenti sono diretti in base al flusso di segnali molecolari
- **Transcription-Regulatory (TR) networks**, tali reti sono reti bipartite con un insieme di nodi che rappresentano i geni e l'altro che rappresenta i *fattori di trascrizione (TF, da "transcription factors")*. I TF sono prodotti di geni (modellati da collegamenti *gene-TF*) mentre i geni sono regolati dai TF (modellati da collegamenti *TF-gene*). I dati per tali reti sono derivati attraverso il processo di immunoprecipitazione della cromatina (detto *ChIP*). Le **gene regulatory networks** sono correlate alle reti TR ma contengono solo geni e i collegamenti rappresentano relazioni regolatorie indirette
- **Metabolic networks** che sono reti bipartite che modellano le relazioni tra le reazioni chimiche che si verificano nelle cellule e i

²Winterbach W, Van Mieghem P, Reinders M, Wang H, de Ridder D. Topology of molecular interaction networks. BMC Syst Biol. 2013;7:90. Published 2013 Sep 16. doi:10.1186/1752-0509-7-90

substrati coinvolti nelle reazioni. Spesso vengono studiate anche reti metaboliche ridotte e non bipartite contenenti solo metaboliti o solo reazioni. Parlando di metabolismo ovviamente i risultati che ottengo tramite queste reti sono diversi da quelli che otterrei usando, ad esempio, un *modello basato su vincoli*

L'analisi della topologia delle reti è formata quindi da:

- la **teoria dei grafi**, mediante la quale si misura il sistema
- le **misure di centralità**, con le quali si analizza (e non si simula) la rete
- la **classificazione della rete**, ovvero la caratterizzazione della sua topologia, e la **robustezza topologica**, che sono i risultati delle analisi. Tali risultati possono anche corrispondere a nuovi “insight” biologici

Se aggiungiamo funzioni logiche per descrivere come cambia lo stato di ogni nodo nel tempo, possiamo effettuare un'analisi dinamica di una rete. Si hanno quindi i modelli basati sulla logica, che aggiungono complessità, sia formale che computazionale, al fine di raggiungere ulteriori risultati.

3.1 Introduzione alle Reti PPI

Prima di approfondire la classe di modelli si vede un piccolo approfondimento sulle **Protein-Protein Interaction (PPI) networks**, come l'esempio in figura 2.4 (anche se si ricorda che la rappresentazione di un sistema è sempre limitante), anche al fine di capire il motivo biologico e i limiti tecnici. Gli algoritmi tipici della teoria dei grafi ci permetteranno di studiarne la topologia.

Questo sarà comunque il modello più studiato in questo corso per questa classe.

Ovviamente le reti sono formalizzate come dei **grafi** e, questo caso, si hanno:

- i *nodi* che rappresentano le proteine
- gli *archi*, che non sono orientati, che corrispondono alle interazioni di legame tra coppie di proteine

Come già anticipato le *reti PPI*, più precisamente, parlando di sistemi *large-scale*, **large-scale PPI** vengono costruite a partire da dati proteomici ottenuti tramite metodi *high-throughput*, ben definiti e con protocolli chiari (per

questo non si hanno particolari challenge dal punto di vista dell'ottenimento dati). Le *reti large-scale PPI* sono però anche caratterizzate da:

- **conflitti**, in quanto tali reti sono ottenute rappresentando tutti i casi possibili, ignorando appunto la componente temporale. Ne segue quindi che, in realtà, le interfacce proteiche possono legare molte altre diverse in modo mutuamente esclusivo, avendo magari che alcuni legami possono avvenire in modo mutuamente esclusivo in un dato tempo. Rappresentare quindi “tutti i tempi” può creare ambiguità. L’idea di togliere componenti al modello per evitare ambiguità non è una buona idea in quanto si toglierebbe senso al modello (che già ha poca capacità predittiva)
- **complessità combinatoria** in quanto si ha un’esplosione del numero di complessi distinti che possono essere formati da una rete di, appunto, “possibili” legami

Il sistema è quindi *statico*, non cambiando nel tempo, che non viene nemmeno rappresentato. Il tempo non è comunque un fattore rilevante per gli studi che si fanno su tali modelli.

Si hanno anche alte criticità, ad esempio:

- gli archi in una *rete PPI* non rappresentano necessariamente connessioni fisiche persistenti, ma piuttosto riassumono le possibilità di interazione, come già detto, e quindi si lascia spazio a:
 - *gaps* sia per i nodi che per gli archi, ovvero nodi/archi non rappresentati in quanto non si conosce una certa interazione (non ancora studiata nella letteratura scientifica)
 - interazioni che sono *false positive* o *false negative*
- le interazioni proteina-proteina nell’intera rete non si verificano, come detto, necessariamente contemporaneamente e/o utilizzando domini di legame diversi e/o nello stesso compartimento cellulare. In altre parole il non rappresentare dove accade una certa interazione può essere un problema. Si “pone tutto allo stesso livello”
- come non si hanno informazioni temporali non si hanno nemmeno informazioni quantitative riguardo al funzionamento della cellula, al suo stato, al ciclo cellulare etc. . .

- si rischia un bias dovuto al fatto che alcune proteine hanno più connessioni semplicemente perché sono meglio studiate (si parla, quando si usa molto la letteratura come base del modello, di **literature-curated networks**). Un esempio banale è pensare che la proteina *p53* compare, a data Marzo 2019, in 94552 paper secondo PubMed (34205 direttamente nel titolo) mentre *Snf1* 1037 volte (solo 318 nel titolo)
- si hanno forti limiti di modellazione. Ad esempio non si può modellare che una proteina interagisca con altre sse queste due ahnno prima avuto un'interazione tra loro (avendo che il massimo che posso avere è un “triangolo” nella rete)

Possiamo quindi concludere che **i risultati dello studio di una rete PPI (ma anche delle altre reti tipiche dell’interaction-based modelling) non sono sempre così affidabili.**

3.2 La Teoria dei Grafi

Lo studio dei **grafi** è centrale in vari problemi/tecnicologie comuni, dal web ai social, dalle reti di collaborazione (pensando ad esempio al *erdős number*) a ipotesi come la famosa *ipotesi dei sei gradi di separazione* che dimostra (come visto sia nel collegare attori a Kevin Bacon che con lo studio del 1967 più “accademico” dello psicologo Stanley Milgram) come si abbiano di media sei passaggi per collegare due persone a caso nel mondo. Si parla infatti spesso della **teoria del mondo piccolo**, che è appunto una teoria matematica e sociologica che sostiene che tutte le reti complesse presenti in natura sono tali che due qualunque nodi possono essere collegati da un percorso costituito da un numero relativamente piccolo di collegamenti.

Definizione 13. *Si definisce formalmente un **grafo** G come una coppia di insiemi:*

$$G = \langle V, E \rangle$$

dove:

- $V = \{v_1, \dots, v_n\}$ è l’insieme dei **nodi**, di cardinalità n
- $E \subseteq V \times V$ è l’insieme degli **archi**, di cardinalità m . Si ha che $E = \{e_1, \dots, e_m\}$, dove un generico arco $e_k = (v_i, v_j)$ con $k = 1, \dots, m$, è definito come una coppia di vertici $v_i, v_j \in V$

Definizione 14. *In un grafo si definisce **nodo isolato** un nodo che non è connesso a nessun altro nodo.*

Definizione 15. In un grafo si definisce **cappio** un arco tra un nodo e se stesso.

Definizione 16. In un grafo $G = \langle V, E \rangle$ si definisce **percorso/cammino** come una sequenza finita o infinita di archi che unisce una sequenza di vertici. Dati quindi due nodi $v_1, v_n \in V$ un cammino da v_1 a v_n un cammino è una sequenza di nodi:

$$(v_1, v_2, \dots, v_n) \text{ tali che } \forall i = 1, 2, \dots, n-1, \exists e = (v_i, v_{i+1}) \in E$$

Se il vertice di partenza coincide con quello di fine si parla di **ciclo**, quindi $v_1 = v_n$.

Ovviamente tra due nodi possono avere distinti cammini.

Il concetto di *cammino* e il concetto di *ciclo* assumono particolare rilevanza in biologia. I *cicli* servono, ad esempio, per rappresentare i *feedback* mentre i *cammini*, ad esempio, per le *vie di trasduzione del segnale* dove si parte da un recettore transmembrana per poi attivare una catena di reazioni.

Definizione 17. Si definisce **grafo连通的** un grafo dove esiste un percorso tra ogni coppia di vertici. In caso contrario si parla di **grafo non连通的**.

Definizione 18. Si definisce **grafo completo o grafo completamente连通的** un grafo, di n nodi, dove ogni nodo è collegato ai rimanenti $n - 1$ nodi.

Definizione 19. Si definisce **grafo totalmente sconnesso** un grafo che non presenta archi.

Definizione 20. Due nodi v_i e v_j si dicono **adiacenti** se esiste un arco, diretto o indiretto, $e = (v_i, v_j)$.

Definizione 21. Dato un nodo v in un grafo indiretto si definisce **grado/-connettività/degree** come il numero di nodi adiacenti ad esso. Tale valore solitamente si indica con k_v , che è ovviamente un numero intero non negativo. Qualora ci sia un cappio esso conta doppio.

Un **nodo isolato** presenta un grado nullo.

Definizione 22. Dato un grafo diretto si definiscono:

- **indegree** di un nodo v come il numero di archi entranti in v
- **outdegree** di un nodo v come il numero di archi uscenti da v

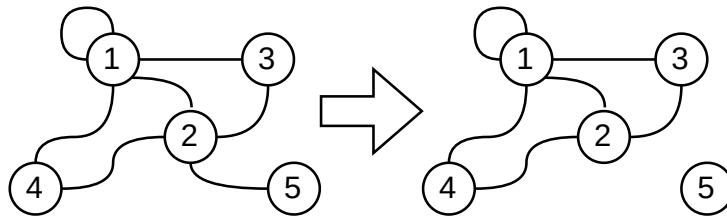


Figura 3.1: Esempio di passaggio da *grafo connesso* a *grafo non connesso* dopo un’ipotetica perturbazione, che produce il *nodo isolato* etichettato con 5.

In letteratura ha volte si definisce **grado/connettività/degree** in un grafo diretto come la somma di indegree e outdegree.

Nella modellazione di sistemi biologici mediante questa classe di modelli è interessante notare come si possa passare da un grafo connesso ad uno non connesso dopo l’influenza di una *perturbazione* sul sistema, che influisce sulle interazioni (e quindi sugli archi) tra le componenti. Tali cambiamenti hanno forte valenza dal punto di vista biologico in quanto se un sistema è propenso a produrre un grafo disconnesso dopo una perturbazione, come ad esempio in figura 3.1, allora è un sistema che è in generale propenso a “fallire”.

Definizione 23. Si definisce **grafo orientato** un grafo dove ogni arco consiste in una **coppia orientata** di vertici, altrimenti si parla di **grafo non orientato**. Formalmente si ha quindi, per un grafo orientato, con $v_1, v_2 \in V$:

$$e = (v_1, v_2) \neq e' = (v_2, v_1)$$

In tal caso, a livello grafico, gli archi sono rappresentati mediante frecce.

In questa classe modellistica si hanno:

- **grafi orientati** per *gene regulatory networks* e *signal transduction networks*
- **grafi non orientati** per *reti PPI*

Definizione 24. Dato un grafo $G = \langle V, E \rangle$ si definisce S come **sottografo** di G come una coppia di insiemi:

$$S = \langle V', E' \rangle$$

dove:

- $V' \subseteq V$

- $E' \subseteq E$

Possiamo quindi dire che, riprendendo la figura 3.1, mediante una *perturbazione*, si ottiene un sottografo del grafo di partenza.

I sottografi sono molto utili per studiare “caratteristiche” di forte interesse biologico, rappresentate appunto da sottografi³:

- **modules** che sono sottografi indotti la cui densità di archi è elevata rispetto al resto del grafo. Questa non è una vera e propria definizione in quanto al natura dei *modules* varia dal contesto e dall’algoritmo utilizzato per scoprirli
- **motifs** che sono piccoli sottografi, solitamente di tre o quattro nodi, la cui sovrarappresentazione o sottorappresentazione può indicare che le loro strutture sono “importanti” o “dannose” per il sistema. Di solito, vengono contati tutti i *motifs* distinti in una rete, ottenendo una *motif signature* per la rete che può quindi essere confrontata con le firme, ottenute campionando da un modello nullo di rete casuale appropriato, per determinare la sovrarappresentazione o sottorappresentazione. Le *motif signature* possono essere usate per caratterizzare le reti stesse
- **graphlets** che sono simili ai *motifs* ma sono *completamente connessi*. Anch’essi vengono utilizzati per costruire firme che catturano le caratteristiche locali di una rete

Questi argomenti verranno approfonditi più avanti, queste sono solo le “definizioni” recuperate nel paper indicato.

Torniamo a parlare della nozione di grado.

Definizione 25. Si definisce *distribuzione dei gradi/degree distribution* come la distribuzione di probabilità dei gradi dei nodi sull’intera rete. Tale distribuzione, denotata con $P(k)$, quindi è la probabilità che un certo nodo abbia grado esattamente pari a k . Tale probabilità si ottiene contando il numero di nodi del grafo, denotati $N(k)$, che presentano grado k e dividendo tale valore per il numero totale di nodi del grafo, che indichiamo con $N = |V|$. Si ha quindi:

$$P(k) = \frac{N(k)}{N}, \quad k = 1, 2, \dots$$

³Winterbach W, Van Mieghem P, Reinders M, Wang H, de Ridder D. Topology of molecular interaction networks. BMC Syst Biol. 2013;7:90. Published 2013 Sep 16. doi:10.1186/1752-0509-7-90

Avendo una distribuzione di probabilità ne segue che:

$$\sum_{i=1}^{k_{max}} P(i) = 1$$

La **degree distribution** ci permette di classificare un grafo, anche solo piazzando con un istogramma i valori di $P(k)$ al variare di k stesso. Ad esempio qualora si avesse un picco nel plot di tali valori allora si avrebbe che la rete ha un “grado caratteristico” (estremizzando l’esempio magari si ha una rete dove tutti i nodi hanno grado $k = 2$) e quindi non si hanno nodi fortemente connessi, con alto *degree*. Questo tipo di analisi non può essere fatta “visivamente” su reti reali ma ci si deve per forza affidare a conti precisi o al più plot della distribuzione stessa. Da tali studi posso estrarre alcune informazioni, ad esempio:

- i nodi con $k = 0$, ovvero i nodi isolati possono rilevare che ci sono probabilmente informazioni mancanti, falsi negativi etc... mentre il valore di $P(0)$ mi dice la probabilità stessa che un qualsiasi nodo della rete sia un nodo isolato
- i nodi con un k elevato sono tendenzialmente molto pochi e sono i cosiddetti **hubs**, che hanno un ruolo chiave nello studio di *reti large-scale*

Al fine di rappresentare i vari valori si usa un piano cartesiano (spesso per necessità rappresentato in *scala logaritmica*) dove:

- l’asse delle x è formato dai valori di k
- l’asse delle y è formato dai valori di $P(k)$

Rappresentando tutte le varie coppie $\langle k, P(k) \rangle$ si ottiene una “forma” che è la cosiddetta **power-law degree distribution** (che rappresenta la relazione funzionale tra k e $P(k)$ dove una variazione relativa in una delle due quantità si traduce in una variazione relativa proporzionale nell’altra quantità, indipendentemente dalla dimensione iniziale delle due quantità, avendo quindi che una quantità varia come potenza di un’altra). Lo studio della *power-law degree distribution* è spesso essenziale nello studio di sistemi biologici.

Sempre restando sullo stesso discorso si è notato che in molte reti se si ha un nodo v_i connesso con un nodo v_j , che a sua volta è connesso al nodo v_h , allora è altamente probabile che v_i sia anch’esso collegato al nodo v_h . Questo **fenomeno di clustering** può essere quantificato usando il cosiddetto **coefficiente di clustering**.

Definizione 26. Dato un nodo v si definisce il **coefficiente di clustering** del nodo v , denotato con C_v , come il numero di archi che connettono nodi adiacenti a v diviso il numero totale delle possibili connessioni che si avrebbero tra i nodi adiacenti a v . Formalmente:

$$C_v = \frac{2N_v}{k_v(k_v - 1)}$$

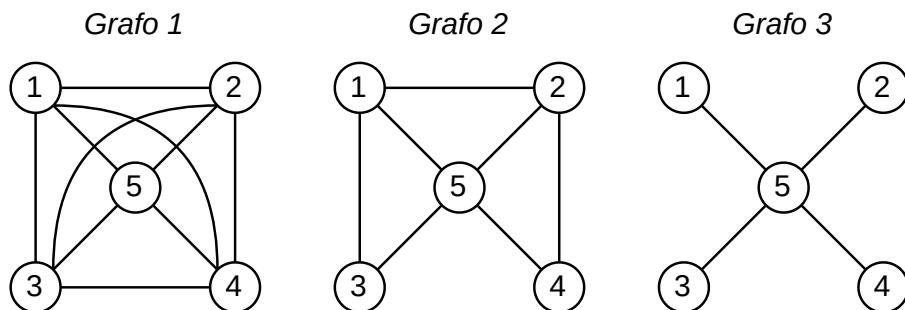
infatti si hanno:

- N_v come numero di archi che connettono coppie di nodi adiacenti a v . Questo valore è facilmente contabile avendo il grafo
- $\frac{k_v(k_v-1)}{2}$, parlando di grafo indiretto, come numero di tutti i possibili archi tra coppie di nodi adiacenti a v , che ha grado k_v , un valore conosciuto

Si osserva inoltre che, ricordando che alla fine si ha a che fare con la misura di probabilità di quanto sia probabile che si abbia o meno un cluster che include il nodo v :

$$0 \leq C_v \leq 1$$

Esempio 1. Vediamo un semplice esempio che mostra come varia il coefficiente di clustering. Si studia, nel dettaglio, il valore C_5 nei seguenti grafi:



In tutti i casi si ha $k_5 = 4$ ma:

- nel Grafo 1 si ha $N_5 = 6$ e $C_5 = \frac{12}{12} = 1$ e infatti il nodo 5 è sicuramente in cluster
- nel Grafo 2 si ha $N_5 = 3$ e $C_5 = \frac{6}{12} = 0.5$
- nel Grafo 3 si ha $N_5 = 0$ e $C_5 = \frac{0}{12} = 0$ e infatti il nodo 5 non è sicuramente in cluster

Questo tipo di coefficiente è utile soprattutto nel caso di *grafo indiretti*. Nel caso di *grafo diretti* bisogna invece ragionare in ottica di *indegree* e *outdegree*.

Un caso limite interessante in ottica di *coefficiente di clustering* è quello della **clique**.

Definizione 27. Si definisce **clique (cricca)** di un grafo non orientato $G = \langle V, E \rangle$ come un sottoinsieme $V' \subseteq V$ di vertici tale che:

$$(v_1, v_2) \in E, \forall v_1, v_2 \in V'$$

quindi un sottoinsieme di vertici con solo vertici collegati da un arco.

Il caso della *clique* è quindi il “caso migliore” parlando del fenomeno del clustering.

Il singolo *coefficiente di clustering* di un nodo comunque non è di particolare interesse se preso in modo isolato, in quanto si vuole classificare l’intera rete.

Definizione 28. Si definisce il **coefficiente di clustering medio**, denotato con $\langle C \rangle$, il valore medio di tutti i coefficienti di clustering dei nodi del grafo.

Il *coefficiente di clustering medio* permette di caratterizzare la tendenza complessiva dei nodi di una rete a formare gruppi o cluster. Questo è quindi un valore che ci permette di caratterizzare la topologia di una rete.

Definizione 29. Si definisce la funzione $C(k)$, che potremmo chiamare “**average clustering distribution**”, come la media dei coefficienti di clustering di tutti i nodi di grado pari a k nella rete.

Il valore $C(k)$ quindi fornisce un’indicazione del carattere modulare/gearchico di una rete, ovvero l’esistenza di sottografi/sottoreti caratterizzati da nodi fortemente collegati internamente, che presentano scarse connessioni con altre parti della rete.

Definizione 30. Si definisce la **lunghezza del cammino** tra due nodi v_1 e v_n come il numero di archi che si hanno nel cammino.

Si definisce **cammino minimo**, notando che potrebbe non essere unico, un cammino di lunghezza minima tra due nodi.

Anche la nozione di *lunghezza del cammino* ha un ruolo centrale nella modellazione di sistemi biologici. Basti pensare, in modo comunque approssimato e semplicistico, che più è lungo il cammino e più sono le interazioni biologiche e quindi, ad esempio, più energia è richiesta alla cellula. Infatti normalmente la natura ha fatto sì che ogni “operazione biologica” venga fatta

nel modo meno dispendioso possibile, quindi, ricollegando la modellazione a grafo, mediante *cammini minimi*. Nella realtà si vedrà il concetto di **ridondanza** in quanto si hanno vari modi, nel mondo biologico, per ottenere lo stesso risultato anche se con “cammini” di lunghezza diversa (magari per casi, ad esempio, in cui alla cellula conviene “temporeggiare”). Inoltre magari ad un percorso più lungo può comunque corrispondere un dispendio energetico minore. In ogni caso l’aggiunta di **pesi** al grafo permette una modellazione leggermente più precisa, potendo rappresentare ad esempio, i “costi” delle reazioni etc...

Un’altra cosa interessante da notare nella maggior parte delle reti è che esiste un cammino relativamente breve tra qualsiasi coppia di nodi e la lunghezza media di tale cammino è proporzionale al logaritmo della dimensione della rete, quindi al numero totale di nodi. Questa è la cosiddetta **small world property** che sembra caratterizzare la maggior parte delle reti complesse, comprese *reti metaboliche* e *reti PPI*.

Definizione 31. Si definisce un **grafo pesato** $G = \langle V, E \rangle$ come un grafo a cui viene associata anche una **funzione di peso** w :

$$w : E \rightarrow \mathbb{R}$$

Esistono, oltre a quelle già citate, anche altre metriche di studio, tra cui⁴:

- ulteriori **metriche per i cammini** come il *cammino minimo* su *grafi pesati* o il *cammino minimo medio* tra ogni coppia di nodi
- la **metrica di centralità** fornisce una classifica dei nodi in base alla loro “importanza”. La versione più semplice sfrutta appunto il grado di un nodo per misurarne la centralità, parlando quindi di **degree centrality**. Un’alternativa è la **closeness centrality** che è il reciproco della somma dei cammini più brevi verso tutti gli altri nodi (cioè un nodo la cui *closeness centrality* è alta è vicino a molti nodi). Un’ulteriore metrica è la **betweenness centrality** ovvero la frazione di cammini minimi che passano attraverso un nodo. Tra le metriche più elaborate si annoverano la **centralità per autovettori** e il **pagerank** che sono misure della frequenza con cui si arriva a un nodo quando si esegue una *random walk* su una rete.

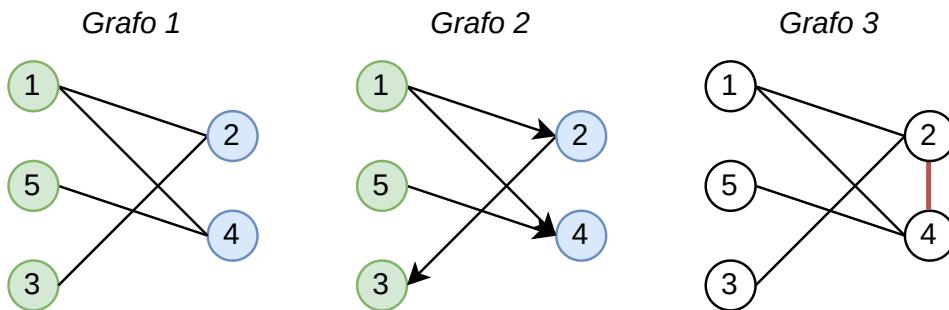
⁴Winterbach W, Van Mieghem P, Reinders M, Wang H, de Ridder D. Topology of molecular interaction networks. BMC Syst Biol. 2013;7:90. Published 2013 Sep 16. doi:10.1186/1752-0509-7-90

Queste non sono comunque metriche solitamente utili per la caratterizzazione della topologia di una rete.

Definizione 32. Si definisce un grafo $G = \langle V, E \rangle$, orientato o meno, come un **grafo bipartito** se:

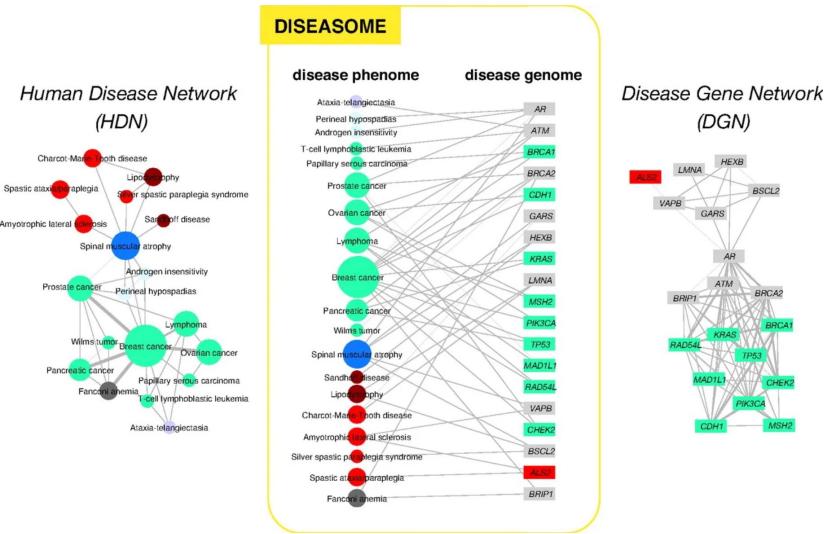
- l'insieme dei nodi V è in realtà l'unione di due sottoinsiemi di nodi V_1 e V_2 , ovvero $V = V_1 \cup V_2$, tali che la loro intersezione è nulla, ovvero $V_1 \cap V_2 = \emptyset$
- data la prima premessa si ha che ogni arco del grafo connette solo un nodo in V_1 ad un nodo in V_2

Ad esempio potremmo avere questi casi, dove i primi due grafi sono bipartiti (come evidenziato anche a livello visivo dai colori che identificano le partizioni) a differenza del terzo (a causa dell'arco in rosso):



Potenzialmente si possono anche avere **grafi tripartiti**, con 3 partizioni, o in generale **grafi multipartiti** con m partizioni.

Un esempio d'uso dei *grafo bipartito* sono le **human desaesome networks**, come ad esempio⁵:



Nella figura si hanno appunto due partizioni:

- un sottoinsieme di nodi per i geni
- un sottoinsieme di nodi per le malattie

Banalmente quindi un nodo che rappresenta una malattia è legato a un nodo che rappresenta un gene se è noto che una mutazione di quel gene induce l'insorgenza di quella malattia. A supporto posso inoltre avere due ulteriori reti solo per i geni e solo per le malattie.

Un esempio di rete basata su un *grafo multipartito* è, ad esempio, una **drug-target protein network**, come quella proposta da Nacher visualizzabile in figura 3.2⁶.

⁵Goh, Kwang-II, et al. "The human disease network." Proceedings of the National Academy of Sciences 104.21 (2007): 8685-8690.

⁶Nacher, J., Akutsu, T. Structural controllability of unidirectional bipartite networks. Sci Rep 3, 1647 (2013). <https://doi.org/10.1038/srep01647>

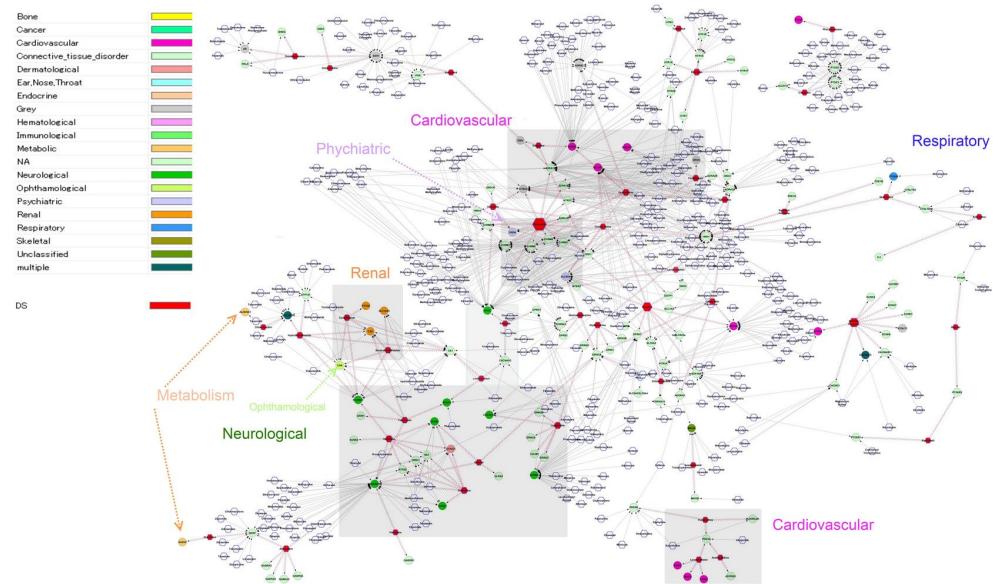
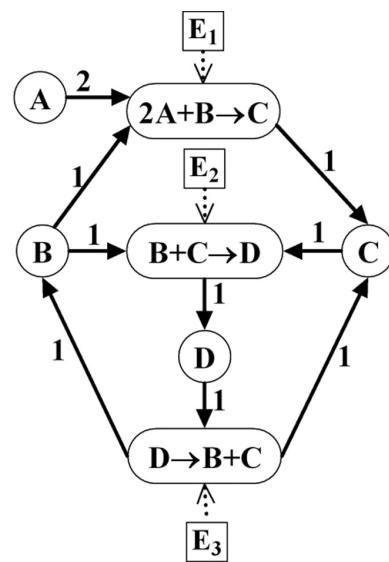


Figura 3.2: Esempio di *drug-target protein network*, rappresentata mediante grafo multipartito.

Un esempio invece di rete tripartita può essere la rappresentazione di un *pathway metabolico*⁷:



⁷Réka Albert; Scale-free networks in cell biology. J Cell Sci 1 November 2005; 118 (21): 4947–4957. doi: <https://doi.org/10.1242/jcs.02714>

dove si hanno tre tipologie di nodo:

1. nodi per i *reagenti* (nell'immagine rappresentati da cerchi)
2. nodi per le *reazioni* (nell'immagine rappresentate da ovali)
3. nodi per gli *enzimi* (nell'immagine rappresentati da quadrati)

Si hanno inoltre due tipologie di archi:

1. le linee solide per rappresentare il *mass flow*, ovvero il *tasso di turnover* delle molecole attraverso una via metabolica, tasso che serve ad indicare il dispendio energetico (???)
2. le linee tratteggiate per la *catalisi*, ovvero fenomeno chimico attraverso il quale la velocità di una reazione chimica subisce delle variazioni per l'intervento di una sostanza (o una miscela di sostanze) detta catalizzatore, che non viene consumata dal procedere della reazione stessa⁸

Inoltre i pesi degli archi indicano i *coefficienti stechiometrici*, che rappresentano infatti il rapporto tra le moli delle diverse sostanze, dei reagenti. Ovviamente l'uso delle reti in ambito biologico può essere espanso, considerando ad esempio i legami tra più reti, che rappresentano vari livelli, ad esempio:

- *reti sociali*, da usare comunque con cautela a causa della loro alta probabilità di portare *falsi positivi/negativi* anche se possono rappresentare informazioni importanti. Ormai si tende comunque a preferire il dato del singolo paziente, puntando alla *medicina personalizzata* in quanto “la media tra i pazienti” raramente è un dato utile. Esse possono rappresentare legami familiari, vicinanza tra persone, informazioni sui luoghi in cui si vive etc...
- *disease networks*
- *reti metaboliche*
- *reti PPI*
- *reti di regolazione genica*
- ...

⁸<https://it.wikipedia.org/wiki/Catalisi>

I vari livelli sono ovviamente connessi a vicenda ma non sempre è facile studiare tali connessioni a causa della mancanza di dati etc...

Altri esempi interessanti di uso sono le **reti in ecologia**, come, ad esempio, lo studio di Faust e Raes⁹ dove si studiavano le *interazioni micobiche*, tra cui il parassitismo, il mutualismo la competizione etc... tramite appunto delle reti.

Un uso recente delle reti è anche quello nelle **neuroscienze**, per capire, durante una malattia, cosa non stia funzionando bene nel cervello. Un esempio è lo studio di Chennu, Srivas et al.¹⁰ dove si è sfruttata la relazione tra reti e funzionalità del cervello per identificare quali aree del cervello/funzioni del cervello funzionassero male in pazienti in stato vegetativo e pazienti minimamente coscienti, facendo il paragone con vari soggetti controllo sani. Sono stati usati anche i concetti di grado etc... nello studio.

⁹Faust, K., Raes, J. Microbial interactions: from networks to models. Nat Rev Microbiol 10, 538–550 (2012). <https://doi.org/10.1038/nrmicro2832>

¹⁰Chennu, Srivas, et al. "Spectral signatures of reorganised brain networks in disorders of consciousness." PLoS computational biology 10.10 (2014): e1003887.