Reti Neurali e Medicina di Precisione

Michele Caselle *Dipartimento di Fisica - Università di Torino*
Istituto Nazionale di Fisica Nucleare, Sez. di Torino, Torino, Italy

La complesità e ricchezza dei dati a disposizione di biologi e medici è aumentata in modo impressionante in questi ultimi anni. Questa ricchezza di informazione rappresenta una sfida affascinante per chi fa ricerca in questo campo ed impone l'uso di metodi nuovi ed originali di analisi. In questo contributo vedremo due esempi di questi approcci innovativi, basati sull'uso delle reti neurali e di strumenti tipici della teoria delle reti e della meccanica statistica dei sistemi disordinati. Questo tipo di analisi può avere ricadute importanti in molti ambiti e soprattutto in quella che si chiama `Medicina di Precisione` e che è destinata a cambiare il modo di curare i pazienti e personalizzare le terapie nei prossimi anni.

Introduzione

Negli ultimi anni stiamo assistendo ad una sempre più forte integrazione tra Medicina e discipline quali Fisica, Matematica ed Informatica. Le motivazioni di questa integrazione sono essenzialmente due. La prima è che la grande quantità di dati a disposizione oggi richiede per essere analizzata metodi di inferenza che sono da sempre bagaglio dei fisici e dei matematici e strumenti informatici per manipolare questi dati che coinvolgono inevitabilmente le competenze degli informatici. La seconda è che si sta capendo con sempre maggiore chiarezza che patologie complesse quali il cancro richiedono approcci complessi. Ad esempio, nel caso dei tumori, è ormai sempre più chiaro che ogni paziente ha un suo proprio modo di sviluppare il cancro con un insieme di alterazioni genetiche (i cosiddetti geni `driver`) che lo rendono in qualche modo unico. Riuscire ad identificare il più rapidamente possibile i geni `driver` del singolo paziente può aiutare il medico ad adattare i protocolli terapeutici e migliorare grandemente l'efficacia delle terapie. Questo nuovo approccio alle patologie complesse è noto come `Medicina di Precisione`[1] [1].

Parallelamente questo processo di integrazione ha permesso a fisici e matematici di allargare i loro orizzonti e scoprire nuovi ambiti in cui applicare gli strumenti di modellizzazione ed inferenza che sono il bagaglio tipico di ogni bravo fisico e matematico.

Tra i vari strumenti che i fisici hanno messo in campo in questi ultimi anni un ruolo speciale lo hanno avuto il vasto insieme di metodi ed idee che fanno uso della teoria delle reti e si appoggiano a tecniche di inferenza basate sull'intelligenza artificiale.

[1] In origine si parlava a questo proposito di `Medicina Personalizzata`, ma si è giustamente osservato che da sempre il bravo medico è attento alle singole persone ed alle loro caratteristiche uniche ed irripetibili e la sua attività è, in questo senso, da sempre personalizzata. Si è quindi preferito il termine `Medicina di Precisione` che rende in modo più efficace il livello di accuratezza e, appunto, di precisione a cui sono arrivati oggi gli strumenti diagnostici

In questo contributo vorrei provare a raccontare due esempi di questi approcci che ho scelto sia per la rilevanza delle ricadute di tipo clinico che per la interdisciplinarietà dei metodi usati. Il primo si basa sull'utilizzo di una tecnica nota come `topic modelling` [2] per l'analisi di dati genomici (più precisamente, come vedremo tra poco, per l'analisi del livello di espressione dei singoli geni) in pazienti affetti da cancro. Il secondo utilizza una versione leggermente modificata di un modello molto noto in teorie delle reti, il cosiddetto `modello di Hopfield` [3], per identificare il particolare sottotipo tumorale di un dato paziente e permettere così al medico di adattare al singolo paziente la terapia. In entrambi i casi vedremo come si possa trasferire in ambito medico una conoscenza sviluppata in contesti diversi e apparentemente lontani (classificazione automatica dei testi per il `topic modelling` e studio dell'organizzazione delle memorie in una rete neurale per il `modello di Hopfield`) ma accomunati al nostro problema dall'alto livello di complessità ed eterogeneità dell'oggetto di studio.

Genoma, Trascrittoma e tipi cellulari.

Tutte le cellule del corpo umano condividono lo stesso genoma, cioè la stessa sequenza di DNA e gli stessi geni, ma le cellule dei diversi tessuti sono molto diverse tra loro. Questa differenza è dovuta al fatto che solo alcuni geni (quelli giusti!) in ogni tessuto sono `espressi` cioè sono tradotti in proteine, che sono i mattoni fondamentali di cui è fatta la cellula. Questo processo noto come `espressione genica` è finemente regolato ed è proprio la presenza di questa complessa rete di regolazione, che è uno dei principali oggetti di studio della moderna Biologia Molecolare, a spiegare la incredibile varietà e ricchezza di possibili tipi cellulari all'interno del corpo umano.

Il processo di espressione dei geni passa attraverso uno stadio iniziale in cui i geni (di nuovo, solo quelli giusti) vengono trascritti in una molecola, nota come RNA messaggero (mRNA), che viene successivamente tradotta nella corrispondente proteina[2]. L'insieme degli RNA messaggeri presenti in una cellula è noto come `Trascrittoma` e rappresenta una fotografia fedele dello stato della cellula e del tipo cellulare cioè, nel caso di tessuti sani, del tipo di tessuto a cui appartiene e, nel caso di tessuti tumorali, del particolare tipo di tumore. Grazie alle tecnologie sviluppate negli ultimi anni è oggi possibile ottenere il `Trascrittoma` di una cellula in modo abbastanza veloce ed economico e rapidamente questa tecnologia dai centri di ricerca si sta spostando in corsia. Lo studio dei livelli di espressione dei geni attraverso la quantificazione degli RNA messaggeri è destinato a diventare nel prossimo futuro il principale strumento diagnostico a disposizione dei medici ed è esattamente questo l'oggetto principale del nostro studio.

Tipicamente il risultato di un'analisi del `trascrittoma` si presenta come una lista di circa 20.000 numeri (quanti sono i geni che codificano le proteine nel genoma umano[3]) da cui vorremmo poter estrarre le informazioni che ci interessano. Ad esempio, nel nostro caso, il particolare tipo di tumore da cui è affetto il paziente che stiamo esaminando.

Per aiutare i medici in questo lavoro sono state create delle grandi banche dati dove sono raccolti molti di questi `Trascrittomi` con l'annotazione, eseguita da oncologi esperti, del particolare tipo di tumore da cui è stata estratta la biopsia e tutta una serie di altre importanti informazioni, quali il livello di gravità del tumore, le terapie a cui è stato sottoposto il paziente, quanto è sopravvissuto e cosi' via. Il più importante di questi *database*, `The Cancer Genome Atlas` (TCGA) [4], contiene più di 30.000 di questi `Trascrittomi` ed è continuamente aggiornato con nuovi dati sui pazienti e con nuove biopsie.

Il problema per il medico che ha in cura un paziente sembra a questo punto abbastanza facile: si tratta solo di capire, dato il `Trascrittoma` del paziente a quale è più simile tra quelli contenuti nel database TCGA. In realtà il problema è molto

[2] Il processo in realtà è molto più complesso, prevede una serie di passaggi intermedi finemente regolati e in genere da un singolo gene è possibile produrre più proteine diverse, ma per la nostra discussione possiamo ignorare queste complicazioni.

[3] Negli esperimenti generalmente vengono anche valutati i livelli di espressione di tutta una serie di altri geni che non codificano per proteine ma, di nuovo, questo è un ulteriore livello di complessità che possiamo ignorare per la nostra discussione.

più complesso di quanto sembri perché questi dati sono estremamente etereogenei e bisogna cercare la risposta in uno spazio a 20.000 dimensioni (tante quanti i geni del `Trascrittoma`) la cui geometria ci è totalmente sconosciuta [5, 6].

Qui entrano in scena le reti neurali e diventa importante il contributo dei fisici e dei matematici abituati da tempo ad affrontare problemi di inferenza di questo livello di complessità

Il problema si può spezzare in due parti: innanzitutto una fase di analisi del database in cui si cercano le cosidette `signatures` cioè collezioni di geni i cui livelli di espressione possano essere usati come firme (`signatures`) dei vari tipi e sottotipi tumorali e successivamente la costruzione di un classificatore che sia in grado di associare il `Trascrittoma` del paziente a cui siamo interessati alla più simile di queste `signatures`. Come già detto, entrambi questi passaggi sono meno semplici di quanto possa sembrare: la costruzione delle `signatures` richiede di interpretare correttamente le correlazioni esistenti tra i dati all'interno del *database*; il secondo passaggio richiede una nozione di somiglianza che dipende dalle caratteristiche dello spazio in cui questi dati sono immersi.

Esistono in letteratura moltissimi algoritmi per affrontare questi ostacoli. Come anticipato nell'introduzione, in questo contributo vorrei concentrarmi in particolare su due approcci che, oltre ad essere particolarmente efficaci, sono interessanti in sè come esempi di trasferimento di conoscenze tra ambiti diversi: l'uso del `topic modelling` per l'identificazione delle signatures e del `modello di Hopfield` per associare un dato `Trascrittoma` al più vicino sottotipo tumorale.

Gene signatures e topic modelling

Col termine `topic modelling` si intende quell'insieme di algoritmi che cercano di individuare in modo automatico l'argomento (`topic`) di un testo all'interno di un *corpus* testuale (ad esempio Wikipedia) a partire dalla frequenza delle parole che lo compongono [2]. L'uso di questo tipo di algoritmi per il nostro problema si basa sull'osservazione che la classificazione automatica dei testi è in realtà molto simile al tipo di classificazione a cui noi siamo interessati. In questa analogia, il *corpus* di testi è rappresentanto dal *database* TCGA, i testi sono i `Trascrittomi` dei singoli pazienti, i geni sono le parole da cui sono composti i testi e il livello di espressione di un gene in un particolare `Trascrittoma` sarà l'analogo della frequenza con cui una data parola appare in un dato testo. Infine, la cosa più importante, i `topic` sono le `signatures` a cui siamo interessati. Così come nell'applicazione ai testi il risultato della analisi con il `topic modelling` sarà identificare gli argomenti di cui parla un certo testo e allo stesso tempo raggruppare assieme tutti i testi che trattano dello stesso argomento, nell'applicazione al TCGA il risultato sarà associare ad ogni `Trascrittoma` un insieme di `signatures` che lo caratterizzano e raggruppare assieme `Trascrittomi simili`, caratterizzati dallo stesso insieme di `signatures` e quindi, si suppone, appartenenti allo stesso sottotipo tumorale.

Il più popolare degli algoritmi di `topic modelling` si chiama `Latent Dirichlet Allocation` (LDA) [7] e, come suggerisce il nome, assume come *prior* del processo di inferenza una distribuizione di Dirichlet per la probabilità dei `topic` nei testi. La scelta di questa distribuzione è particolarmente comoda per velocizzare gli algoritmi di inferenza usati, ma non ha nessuna ulteriore giustificazione. In particolare non c'è nessun motivo per cui debba essere la *prior* giusta per il problema a cui siamo interessati.

Negli ultimi anni c'è stato un interessante progresso in questo campo, che ha permesso di superare le limitazioni di LDA. In particolare si è capito che l'identificazione dei `topic` è equivalente al problema di trovare le comunità in un *network* bipartito [8] (si veda la figura 1) e si sono così potuti adattare al `topic modelling` una serie di algoritmi molto potenti elaborati negli anni scorsi nell'ambito della Teoria delle Reti [9]. In particolare uno di questi, una versione gerarchica dello `Stochastic Block Modelling` [10, 11], si è dimostrato lo strumento giusto per superare le limitazioni di LDA (in particolare per ricostruire i topic senza alcuna particolare assunzione a priori sulla loro distribuzione all'interno dei testi) e si adatta perfettamente alla applicazio-

ne a cui siamo interessati, caratterizzata da una grandissima eterogeneità dei dati [12].

L'aspetto più interessante di questo algoritmo è la sua natura probabilistica. Come risultato del processo di inferenza l'algoritmo restituisce una partizione ottimale dei geni e separatamente dei pazienti in gruppi (che chiameremo `topic` per i geni e *cluster* per i pazienti) e delle distribuzioni di probabilità $P(t|s)$ che un dato `topic` t sia associato ad un certo paziente s e la probabilità $P(g|t)$ che un dato gene g faccia parte di un `topic` t. Da queste due distribuzioni si possono ricostruire tutte le quantità di interesse ed in particolare le `signatures` di cui parlavamo prima [12].

Come funziona lo *Stochastic Block Modelling gerarchico*?

Lo *Stochastic Block Modelling gerarchico* (hSBM) è un algoritmo che cerca di massimizzare la probabilità che il `modello` θ (dove per `modello` si intende una particolare partizione di geni in `topic` e pazienti in *clusters*, convenzionalmente indicata con θ) descriva i dati $\mathcal{A}$

$$P(\theta|\mathcal{A}) = P(\mathcal{A}|\theta)P(\theta) \qquad (1)$$

senza fare alcuna assunzione a priori.

Nella applicazione in cui siamo interessati $\mathcal{A}$, è la matrice di espressione genica i cui ingressi $\mathcal{A}_{ij}$ rappresentano il numero (opportunamente normalizzato) di RNA messaggeri del gene i nel paziente j. L'algoritmo si basa sul fatto che $\mathcal{A}$ può anche essere interpretata come la matrice di adiacenza di un *network* bipartito pesato i cui nodi nei due *layer* sono rispettivamente i geni ed i pazienti ed i *link* del *network* (che essendo bipartito possono solo congiungere geni con pazienti e non geni o pazienti tra loro) hanno un peso dato proprio dal livello di espressione del gene i-esimo nel paziente j-esimo.

L'idea alla base dell'algoritmo è di associare ad ogni ingresso della matrice $\mathcal{A}$ (che, lo ricordiamo, è un numero intero positivo che corrisponde al numero di RNA messaggeri di quel particolare gene in quel particolare paziente) un numero di legami, che chiameremo `edges` d'ora in poi, pari all'elemento di matrice.

A questo punto l'algoritmo che massimizza la probabilità è un semplice Montecarlo (cioè una catena di Markov le cui mosse sono pesate dalla probabilità $P(\theta|\mathcal{A})$) che assume all'inizio una distribuzione casuale di geni nei `topic` e pazienti nei *cluster* e ad ogni mossa della catena sposta uno dei nodi (ad esempio un gene) da un `topic` all'altro accettando o rifiutando la mossa a seconda di come cambia la probabilità $P(\theta|\mathcal{A})$. Nell'algoritmo si studia, per comodità, non la probabilità ma il suo logaritmo cambiato di segno Σ che viene abitualmente chiamato `description length`

$$\Sigma = -lnP(\mathcal{A}|\theta) - lnP(\theta)\ .$$

L'obiettivo dell'algoritmo è minimizzare Σ, cioè trovare la configurazione di *cluster* e `topic` che descrive in modo più compatto i dati contenuti nella matrice $\mathcal{A}$[4]

Ma la cosa interessante dell'algoritmo è che alla fine del processo di ottimizzazione non abbiamo semplicemente la partizione ottimale,

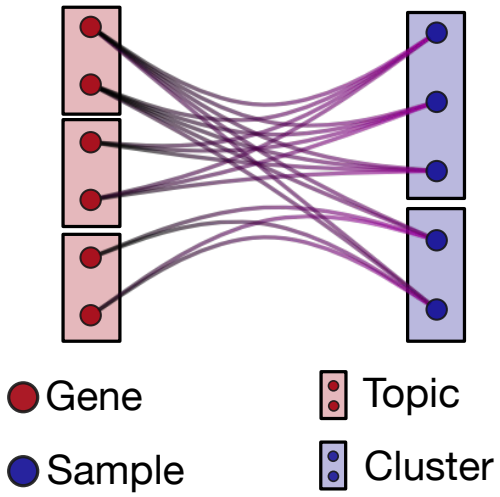

Figura 1: *I dati di espressione genica di un set di pazienti possono essere rappresentati come un grafo bipartito. Nel layer di sinistra ci sono i geni, in quello di destra i pazienti ed i link che li collegano hanno un peso proporzionale al livello di espressione dei geni nei vari pazienti (ovvero gli ingressi della matrice $\mathcal{A}$ che è interpretata come una matrice di adiacenza). Nella figura si vede anche il risultato dell'analisi con l'algoritmo di topic modelling, con i geni divisi in topics e i pazienti in clusters.*

[4] Chi ha dimestichezza con l'entropia di Shannon riconoscerà immediatamente in questa definizione il fatto che per ogni distribuzione di probabilità P è possibile costruire un codice C tale che la lunghezza in bit di C è uguale a $-\log P$ e che il modello θ che meglio descrive i dati è sempre quello che li comprime meglio, cioè quello associato al più breve codice C.

ma possiamo anche ricavare facilmente le probabilità $P(t|s)$ e $P(g|t)$ che saranno date dalle relazioni:

$$P(t|s) = \frac{N_{t,s}}{N_s} \tag{2}$$

dove $N_{t,s}$ è il numero di `half-edges` (estremità degli `edges`) del grafo provenienti dal `topic` t che puntano sul paziente s e N_s è il numero totale di `half-edges` che puntano sul paziente s, mentre

$$P(g|t) = \frac{N_{t,g}}{N_t}. \tag{3}$$

dove $N_{t,g}$ è il numero di `half-edges` del `topic` che puntano sul gene g, e N_t è il numero totale di `half-edges` che puntano sul `topic` t.

Sono queste probabilità che ci permettono di identificare i `topic` da associare ad un particolare gruppo (*cluster*) di pazienti (che saranno i `topic` con la più alta $P(t|s)$ mediata su tutti i pazienti del *cluster*) e di identificare all'interno dei `topic` i geni più importanti (quelli che massimizzano $P(g|t)$) che formeranno le `signatures` che stavamo cercando.

Una volta definite queste `signatures` possiamo affrontare la seconda parte del problema: costruire un classificatore che in modo affidabile ci dica qual è il particolare sottotipo tumorale del paziente che stiamo studiando.

I sottotipi tumorali come memorie di un cervello in cui i geni giocano il ruolo di neuroni

Il `modello di Hopfield` [3] è forse uno dei più eleganti esempi di applicazione della Meccanica Statistica alle Neuroscienze. È un modello che cerca di spiegare il funzionamento della memoria associativa cioè della capacità del cervello di ricostruire un intero ricordo partendo solo da un sottoinsieme degli elementi che lo compongono. Il modello è definito da un insieme di N variabili binarie $S_i(t) = \pm 1$, $i = 1...N$ che evolvono nel tempo e che interagiscono tra loro mediante una matrice di costanti di accoppiamento $J_{i,j}$ che possono prendere sia valori positivi che negativi. Il modello vuole essere una rappresentazione molto semplificata del cervello: gli S_i rappresentano i neuroni, il valore $S = +1$ indica che il neurone in quell'istante è attivo (`firing`), mentre $S = -1$ indica che il neurone è inattivo e i pesi J_{ij} rappresentano l'intensità delle connessioni sinaptiche tra i neuroni S_i e S_j che possono essere sia attivatorie (J_{ij} positivo) che inibitorie (J_{ij} negativo) . In questa realizzazione le memorie che si vogliono ricordare sono opportuni *pattern* (etichettati dall'indice μ) $S_i = \xi_i^\mu$ di neuroni attivi o silenti . Nel seguito assumeremo di voler memorizzare nella rete P *pattern*, quindi $\mu = 1...P$. La configurazione dei neuroni evolve nel tempo secondo la regola:

$$S_i(t+1) = sign\left(\sum_j J_{ij} S_j(t)\right) \tag{4}$$

L'idea alla base del modello è scegliere i pesi sinaptici J_{ij} in modo che le memorie ξ_i^μ siano punti fissi della dinamica descritta dall'eq. (4). In questo modo se la configurazione iniziale è sufficentemente simile alla memoria che si vuole ricordare (e quindi è nel bacino d'attrazione del corrispondente punto fisso) dopo poche iterazioni dell'eq.(4) verrà attratta dal punto fisso. In altri termini, se un certo numero di neuroni è già allineato secondo la memoria a cui siamo interessati, la dinamica porterà anche tutti gli altri neuroni ad allinearsi nella direzione giusta e quando la configurazione memorizzata sarà raggiunta il processo si fermerà. Questo è esattamente ciò che si intende quando si parla di memoria associativa. Nelle applicazioni pratiche il numero di neuroni che devono essere allineati alla memoria nella configurazione iniziale per raggiungere il punto fisso giusto dipende da P e può essere anche molto basso.

La bellezza del modello sta nel fatto che esiste una regola molto semplice, detta `regola di Hebbs`, che permette di far sì che i punti fissi coincidano con le memorie

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{5}$$

È facile rendersi conto che il modello, con poche modifiche (tre valori invece di due: +1 se il gene è molto espresso, -1 se è poco espresso, 0 se non appartiene a nessuna delle signatures utilizzate per classificare i tumori), si presta perfettamente al nostro scopo. Nell'applicazione come classificatore di tumori i neuroni del modello

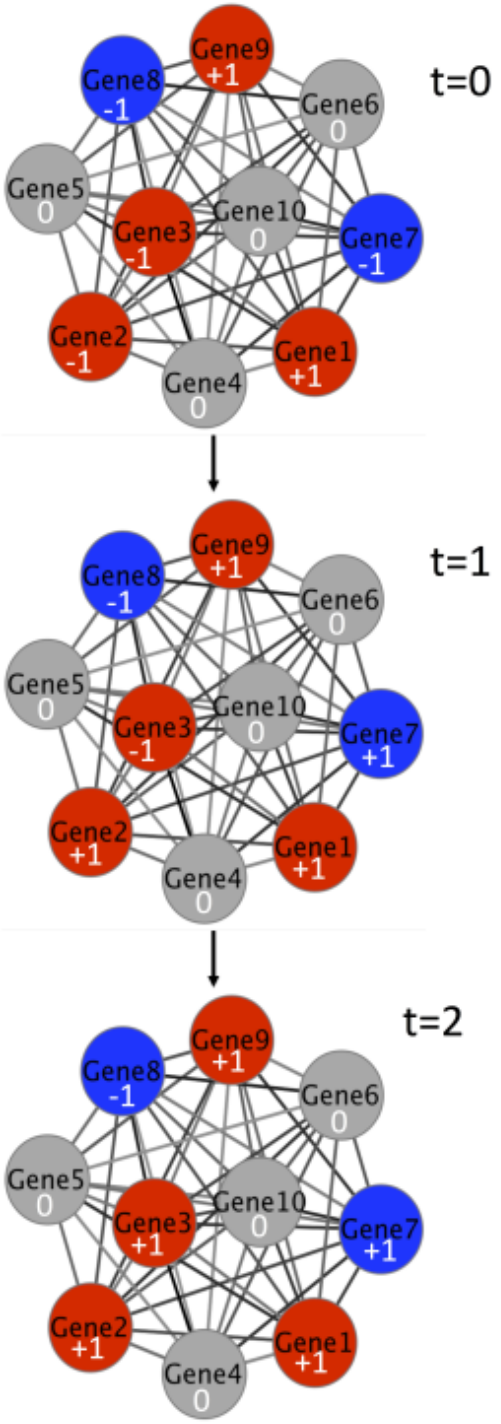

Figura 2: *Evoluzione verso la configurazione di minimo del modello di Hopfield applicato alla classificazione dei tumori [13].*

saranno i geni e le memorie (che definiscono le interazioni tra geni attraverso la `regola di Hebbs`) saranno le `signatures` dei vari tipi e sottotipi tumorali [13]. Inizializzando la rete con i valori dei livelli di espressione dei geni di un particolare paziente, la rete evolverà verso una delle sue memorie permettendoci così di classificare rapidamente il paziente associandolo al particolare sottotipo tumorale che corrisponde alla memoria verso cui è stata attratta la configurazione iniziale [13]. *Test* su vari tipi diversi di tumori, con una ampia varietà di sottotipi da identificare, mostrano che l'algoritmo è veloce ed estremamente affidabile [13]. In fig.(2) si può vedere un esempio schematico del funzionamento del modello in un caso particolarmente semplice di classificazione binaria. Le memorie (`signatures`) sono identificate dai colori rosso e blu (per semplicità abbiamo assunto memorie che non si sovrappongono). Nella parte alta della figura si vede la configurazione iniziale (definita dai valori +1 e -1) all'istante $t = 0$ e nei riquadri successivi l'evoluzione del sistema verso il minimo. Il paziente sarà associato alla `signature` rossa e quindi al corrispondente sottotipo tumorale.

Conclusioni

In questo contributo abbiamo visto due esempi di contaminazione tra discipline diverse. Metodi tipici dell'analisi dei testi (`topic modelling`) basati sulla meccanica statistica dei sistemi complessi (`stochastic block modelling`) e sulla teoria delle reti (ricostruzione di comunità su *network* binari) possono essere usati per risolvere problemi in Biologia Molecolare che sono estremamente difficili da affrontare con metodi di analisi più *standard*. Abbiamo anche visto come una rete neurale pensata per modellizzare la memoria associativa (il `modello di Hopfield`) possa essere adattata al problema di identificare in modo affidabile il particolare tipo di tumore da cui è affetto un dato paziente. Entrambi questi esempi mostrano la ricchezza e la potenza di un approccio interdisciplinare alla ricerca scientifica, approccio che è la cifra distintiva del modo di fare scienza in questo terzo millennio e che , ne sono sicuro, ci riserverà ancora bellissime sorprese nel prossimo futuro.

[1] E. A. Ashley, *Towards precision medicine*, Nature Reviews Genetics 17 (SEP, 2016) 507–522.

[2] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, *An overview of topic modeling and its current applications in bioinformatics*, Springerplus 5 (SEP 20, 2016).

[3] J. J. Hopfield. *Neural networks and physical systems with emergent collective computational abilities*, Proc. Nat. Acad. Sci. U.S.A., 79(8) (1982) 2554.

[4] K. Chang, C. Creighton, C. Davis, et al., *The cancer genome atlas pan-cancer analysis project*, Nature Genetics 45 (2013), no. 10 1113.

[5] N. Andor, T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley, *Pan-cancer analysis of the extent and consequences of intratumor heterogeneity*, Nature Medicine 22 (JAN, 2016) 105+.

[6] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, *Outcome signature genes in breast cancer: is there a unique set?*, Bioinformatics 21 (JAN 15, 2005) 171–178.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent dirichlet allocation*, Journal of Machine Learning Research 3 (2003), no. Jan 993–1022.

[8] M. Gerlach, T. P. Peixoto, and E. G. Altmann, *A network approach to topic models*, Science Advances 4 (2018), no. 7 eaaq1360.

[9] S. Fortunato and D. Hric, *Community detection in networks: A user guide*, Physics Reports 659 (NOV 11, 2016) 1–44.

[10] T. P. Peixoto, *Efficient monte carlo and greedy heuristic for the inference of stochastic block models*, Physical Review E 89 (Jan, 2014) 012804.

[11] T. P. Peixoto, *Hierarchical block structures and high-resolution model selection in large networks*, Physical Review X 4 (Mar, 2014) 011047.

[12] F. Valle, M. Osella, and M. Caselle, *A Topic Modeling Analysis of TCGA Breast and Lung Cancer Transcriptomic Data*, Cancers 12 (2020), no. 12 3799.

[13] L. Cantini and M. Caselle, *Hope4Genes: a Hopfield-like class prediction algorithm for transcriptomic data*, Scientific Reports 9 (JAN 23, 2019).

Michele Caselle: è Professore di Fisica Teorica presso l'Università degli Studi di Torino. Si occupa da alcuni anni di applicazioni della Fisica dei Sistemi Complessi alla Biologia Molecolare.