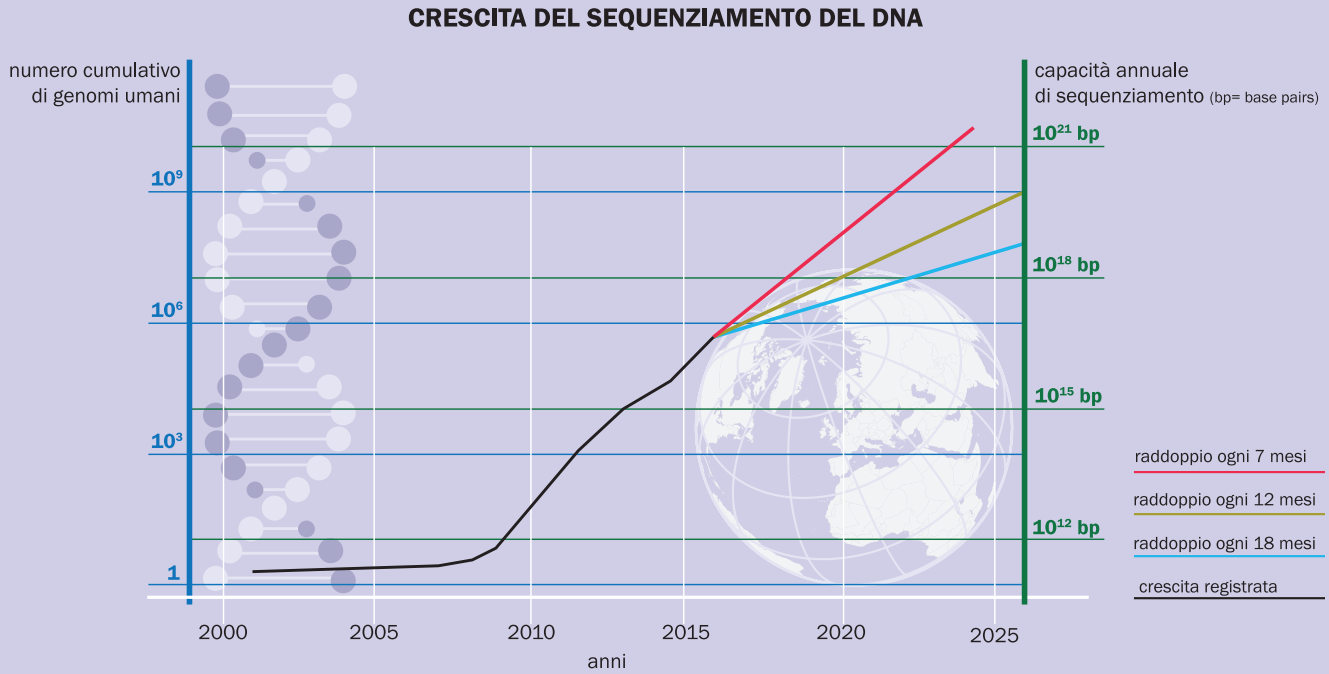
Le reti della vita

La bioinformatica per una medicina di precisione

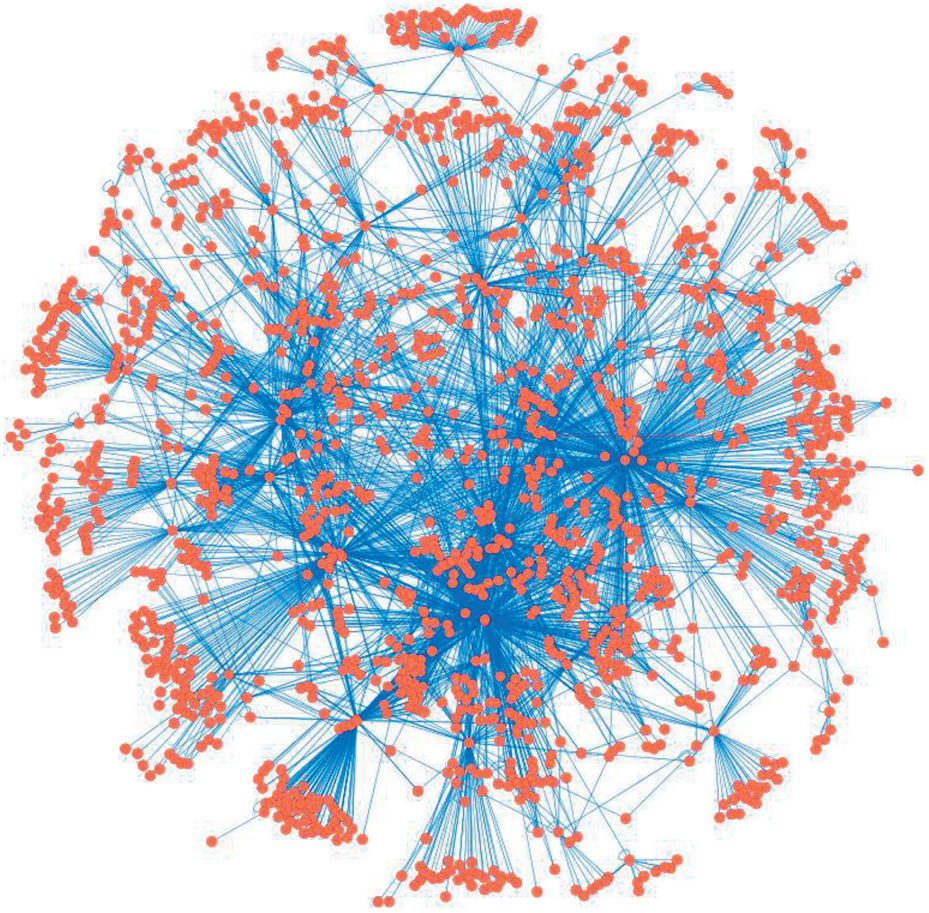
di Michele Caselle



La biologia molecolare sta attraversando in questi ultimi anni una vera e propria rivoluzione scientifica. Si tratta di un processo per certi versi simile a quello sperimentato in fisica negli anni '20 del secolo scorso quando nacque la meccanica quantistica.
Grazie alle nuove tecnologie di sequenziamento e di manipolazione del genoma i biologi hanno oggi accesso a una mole di dati e informazioni inimmaginabile fino a qualche anno fa (vd. fig. a) e questo ha portato a un radicale cambiamento di paradigma, alla nascita di nuove idee e anche di nuove discipline scientifiche. Per cercare di controllare questo diluvio di dati ed estrarne tutte le informazioni nascoste i ricercatori hanno dovuto adattare alla biologia molecolare le idee e gli strumenti tipici del *data mining*. Termini come bioinformatica, biologia computazionale o *systems biology,* che erano sconosciuti vent'anni fa, sono diventati ormai la norma nei laboratori di tutto il mondo. Si tratta di discipline giovani, fortemente interdisciplinari, dove fisici, matematici, esperti di informatica, biologi e medici lavorano assieme per arrivare a una comprensione sempre più profonda di come funzionano le cellule e gli organismi viventi.
Alla base di questo progresso sta il sequenziamento del genoma umano ottenuto all'inizio degli anni 2000. Un risultato importante non solo per le sue ricadute scientifiche, ma soprattutto per il suo valore simbolico. Con il termine "genoma" si intende la sequenza di Dna che in ogni organismo vivente codifica l'informazione genetica. Il Dna è una lunga molecola formata dalla combinazione di quattro elementi fondamentali detti "nucleotidi" o "basi": adenina, citosina, guanina e timina, solitamente indicati con l'abbreviazione A, C, G, T. La lunghezza del genoma può variare molto da specie a specie. Nell'uomo la sequenza è composta da circa 3 miliardi di basi mentre per gli organismi più semplici che esistono in natura, i batteri, può scendere fino a 500.000 basi. Nel

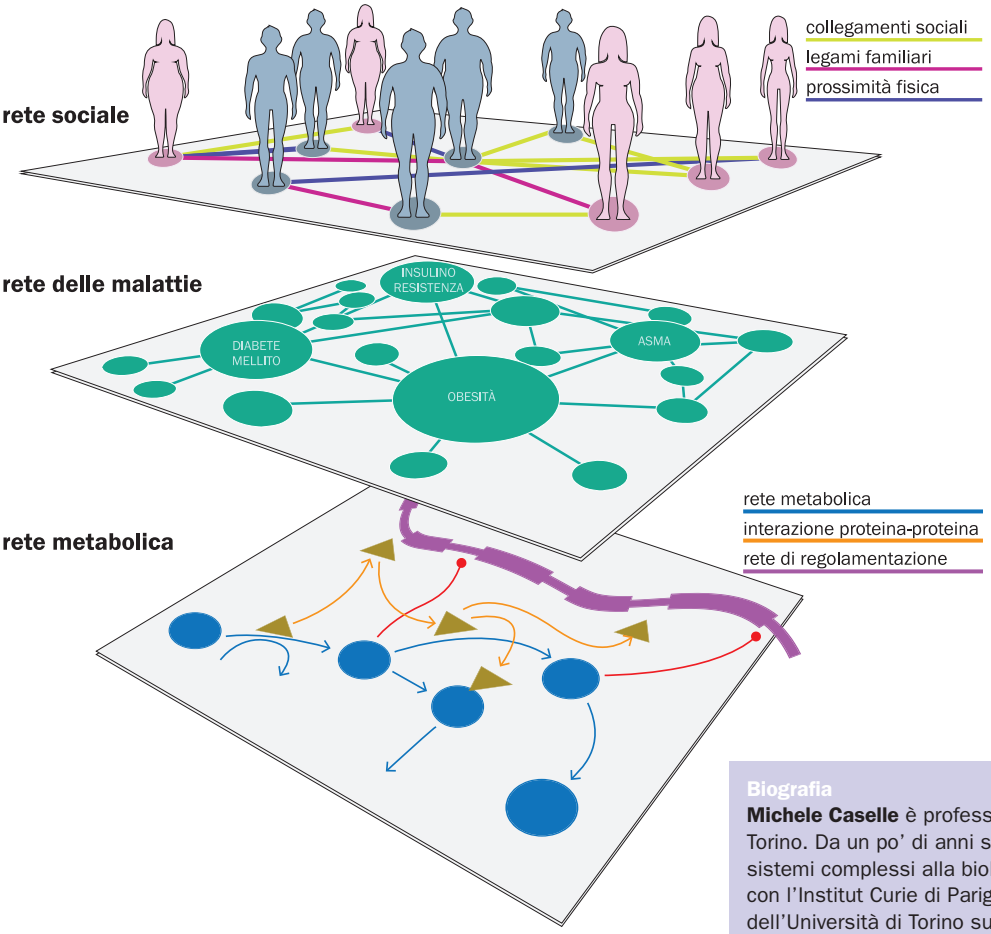
a.
Nella figura è riportata la crescita della quantità di informazioni disponibili per i ricercatori (misurata in numero di genomi umani) al passare del tempo. Come si può vedere stiamo assistendo a una crescita esponenziale (l'asse delle ordinate è in scala logaritmica, in questa scala una crescita esponenziale corrisponde a una retta), ma la cosa impressionante è la rapidità di questa crescita. La cosiddetta "legge di Moore", che ha governato la crescita della potenza di calcolo dei computer negli ultimi anni, prevede un raddoppio ogni 18 mesi (la retta blu). In ambito genomico abbiamo assistito negli ultimi 10 anni a un raddoppio ogni 7 mesi! Estrapolando questa curva (la retta in rosso nella figura) arriveremo tra un paio di anni ad avere una quantità di dati a disposizione dei ricercatori dell'ordine dello yottabyte (10^{24} byte, un milione di miliardi di miliardi di byte), paragonabile all'intera capacità di memoria disponibile sul pianeta oggi.

Dna è codificata l'intera informazione necessaria per formare l'individuo adulto e, infatti, tipicamente la lunghezza del Dna cresce con la complessità dell'organismo. La molecola di Dna è formata da una doppia elica, in cui le basi sono sempre appaiate nella combinazione (A,T) e (C,G), e proprio questa proprietà è alla base delle due caratteristiche fondamentali del Dna: la capacità di duplicarsi in modo fedele (e quindi trasferire alla progenie l'informazione genetica) e la possibilità di essere letta dalla cellula che (mediante una serie di complessi processi biochimici indicati complessivamente con il termine di "espressione genica") usa l'informazione codificata nel Dna per produrre le proteine che sono i costituenti elementari della cellula. Possiamo pensare alle proteine come ai "mattoni" con cui sono costruite tutte le componenti cellulari e che permettono alle cellule di organizzarsi in tessuti e in un organismo completo. Ogni proteina è codificata da un'unità del genoma, cioè da una porzione di una sequenza del Dna chiamata "gene". Dalla conclusione del Progetto Genoma nel 2000 siamo in grado di leggere l'intera sequenza del Dna dell'uomo e negli anni successivi lo stesso risultato si è ottenuto per tantissimi altri organismi viventi.
Negli ultimi anni grazie al miglioramento delle tecnologie siamo riusciti anche a riconoscere le differenze nel genoma da individuo a individuo, e in prospettiva sarà possibile per ogni singolo individuo conoscere la propria sequenza di Dna.
Ma leggere non significa capire! Siamo in grado di identificare tutte le proteine codificate nel genoma umano, ma siamo ancora molto lontani dal capire le loro funzioni.
Alcune cose cominciamo però a intuirle. Innanzitutto, è chiaro che per capire un sistema complesso come una cellula vivente è necessario un cambiamento di paradigma, in cui al centro dello studio deve essere posto non il singolo gene o la singola molecola, ma l'intero sistema, e soprattutto la rete di interazioni tra i vari geni. Questo cambiamento di paradigma è il primo grande regalo dei *big data* alla biologia molecolare.
È ormai chiaro che è solo a livello di rete che si possono capire tutte le complesse funzioni di un organismo vivente e anche della sua unità base, la cellula. L'altra cosa ormai chiara è che le proteine con cui sono costruiti tutti gli organismi viventi sono più o meno sempre le stesse e che il loro numero non cresce in modo significativo con la complessità dell'organismo. Il genoma umano è composto da circa 20.000 geni, più o meno lo stesso numero del topo o del moscerino della frutta. Quello che cambia al crescere della complessità è il "libretto di istruzioni", l'insieme di regole che decide con quali "mattoni" costruire una particolare cellula e in che ordine disporli. Queste istruzioni sono codificate in una rete detta "rete di regolazione" (un esempio è riportato in fig. b), le cui proprietà per gli organismi complessi sono ancora largamente inesplorate. Ricostruire questa rete di regolazione dai dati sperimentali è una delle sfide più impegnative e affascinanti della biologia computazionale moderna. Esistono essenzialmente due approcci. Il primo si basa sul fatto che i principali attori di questo sistema di regolazione (detti "fattori di trascrizione") sono in grado di riconoscere certe sequenze di Dna (dette "siti di legame") e di legarsi a queste sequenze in modo da regolare l'accensione e lo spegnimento dei geni vicini. Il problema è che identificare queste sequenze all'interno del genoma è come trovare un ago in un pagliaio. Questi siti di legame sono abbastanza corti (tipicamente da 5 a 20 nucleotidi) e possono essere degeneri (cioè accettare in alcune posizioni più basi diverse), mentre il genoma, ad esempio nel caso dell'uomo, contiene più di 3 miliardi



b.
La rete di regolazione dell'Escherichia coli, un piccolo batterio che vive abitualmente nel nostro intestino e che è uno dei più studiati organismi modello in biologia molecolare.
Nella figura ogni nodo rappresenta una proteina e le interazioni con cui una proteina regola l'espressione di un'altra proteina sono indicate da linee blu.

di basi. Esistono metodi estremamente ingegnosi per ridurre la complessità del problema. Ci sono tecniche sperimentali (dai nomi esotici, come Chip-seq, Dnase, Footprinting) che permettono di restringere la porzione di genoma in cui cercare, e si possono combinare queste tecniche con uno studio di conservazione evolutiva, basato sull'idea che se una porzione di genoma è particolarmente importante, ad esempio perché è un sito di legame, allora sarà conservato evolutivamente, cioè comparirà praticamente uguale in specie diverse.
Il secondo approccio è più indiretto. Una serie di altre tecniche sperimentali permette di misurare l'effetto finale di queste regolazioni, valutando la cosiddetta "espressione genica" (in termini più precisi, queste tecniche permettono di misurare per ogni gene la quantità di Rna messaggero, la molecola in cui viene trascritta l'informazione del Dna).
Da questa informazione, sofisticati metodi di inferenza permettono di ricostruire la rete di regolazione che ha la maggiore probabilità di aver generato lo schema di espressione genica osservato. Si tratta di metodi probabilistici, caratterizzati da un notevole grado di incertezza ma che se combinati con il metodo diretto possono dare ottimi risultati.
Siamo ancora molto lontani dall'avere una conoscenza esaustiva della rete di regolazione di organismi complessi come l'uomo. Le reti che siamo in grado di costruire oggi sono ancora piene di lacune e di errori. Inoltre, è ormai chiaro che ogni tessuto ha una propria rete di regolazione, diversa da quella degli altri tessuti. Migliorare la conoscenza di queste reti è un processo che coinvolge centinaia di gruppi di ricerca in tutto il mondo.
Si tratta di un obiettivo di grande importanza. Infatti, è ormai chiaro che molte delle patologie più complesse, dal cancro al diabete all'Alzheimer, sono in realtà dovute ad alterazioni di questa rete di regolazione. Una migliore comprensione della struttura delle reti di regolazione dei tessuti sani è il punto di partenza per riconoscere queste alterazioni e curarle. Un aspetto importante di questa linea di ricerca è stato il rendersi conto che non esistono due pazienti uguali. Ogni tumore è diverso dagli altri. Ogni tumore trova una via diversa per alterare la rete di regolazione e trovare la sua strada per invadere l'organismo. La sfida è quindi riuscire a trovare la cura giusta per ogni singolo paziente. È quella che si chiama "medicina di precisione", che è forse il regalo più grande che ci ha fatto l'era dei *big data*: la possibilità di costruire terapie personalizzate, su misura per ogni paziente, ottimizzando le possibilità di guarigione e minimizzando gli effetti nocivi dei farmaci.



c.
Alla base della medicina di precisione è l'idea di integrare diverse fonti di informazione molecolare (nella figura sono rappresentate dal livello più basso: ogni nodo potrà essere un gene o una proteina o un metabolita) associandole alle possibili patologie (nella figura è il livello intermedio, ogni nodo rappresenta una malattia diversa) e associare queste informazioni alle caratteristiche peculiari dei singoli individui (rappresentati nella figura dal livello più alto, in cui ogni singolo individuo è caratterizzato non solo dalla sua particolare sequenza di Dna e dalla sua particolare rete di regolazione genica, ma anche dalle sue interazioni con gli altri individui e con l'ambiente).

Biografia
Michele Caselle è professore di fisica teorica presso l'Università di Torino. Da un po' di anni si interessa alle applicazioni della fisica dei sistemi complessi alla biologia molecolare e in questa veste collabora con l'Institut Curie di Parigi e con il dipartimento di Oncologia dell'Università di Torino su progetti di medicina di precisione.

DOI: 10.23801/asimmetrie.2019.27.8