BIG DATA E INNOVAZIONE: PROSPETTIVE PER LA SALUTE E I DATI PERSONALI

FRANCESCO LESCAI (*)

Nota presentata dal m.e. Giuliano Gasperi
(Adunanza del 23 marzo 2023)

SUNTO. – Negli ultimi anni, il concetto di big data ha guadagnato rilevanza in vari settori, con un impatto significativo in medicina. In questo articolo esploriamo come i big data stiano trasformando la pratica medica e la protezione dei dati personali. La gestione di grandi volumi di dati genomici, ad esempio, offre nuove prospettive sulla diversità genetica e sulle malattie. Tuttavia, la vera rivoluzione non risiede nella quantità di dati, ma nella capacità di estrarre informazioni significative per comprendere fenomeni complessi. Vengono discussi esempi come il database GnomAD e studi sulle varianti genetiche legate alla schizofrenia, evidenziando l'importanza della diversità genetica e dell'analisi dettagliata. Le infrastrutture per il calcolo ad alte prestazioni e l'Intelligenza Artificiale (IA) giocano un ruolo cruciale nell'elaborazione di questi dati, con l'IA che permette di catturare relazioni non lineari tra componenti biologici. Inoltre, vengono affrontate le sfide relative alla privacy dei dati e all'impatto ambientale delle infrastrutture di calcolo. In sintesi, sottolineiamo l'importanza della qualità e della comprensione delle relazioni fra i dati, supportata da solide infrastrutture bioinformatiche e avanzate capacità analitiche dell'IA, per migliorare la comprensione dei sistemi biologici.

***

ABSTRACT. – In recent years, the concept of big data has become increasingly relevant across various sectors, with significant impact in medicine. In this article we explore how big data is transforming medical practice and personal data protection. Managing large volumes of genomic data, for instance, offers new perspectives on genetic diversity and diseases. However, the true revolution lies not in the quantity of data but in the ability to extract meaningful information to understand complex phenomena. Examples such

(*) Dipartimento di Biologia e Biotecnologie "L. Spallanzani", Università degli Studi di Pavia, Italy. E-mail: francesco.lescai@unipv.it

as the GnomAD database and studies on genetic variants linked to schizophrenia are discussed, highlighting the importance of genetic diversity and detailed analysis. High-performance computing infrastructures and Artificial Intelligence (AI) play a crucial role in processing these data, with AI enabling the capture of non-linear relationships among biological components. Additionally, the article addresses challenges related to data privacy and the environmental impact of computing infrastructures. In summary, we emphasise the importance of data quality and correlation, supported by robust bioinformatics infrastructures and advanced AI analytical capabilities, to enhance the understanding of biological systems.

Introduzione

Negli ultimi anni, il concetto di big data è diventato sempre più rilevante in tutti i campi della nostra vita quotidiana, dalla medicina all'innovazione tecnologica, alla raccolta di dati in tutti i settori. Un campo particolare in cui i big data avranno un impatto significativo però è proprio la medicina, poiché hanno implicazioni importanti non solo per la trasformazione della pratica medica, ma anche per l'impatto che la raccolta e la gestione di questi dati possono avere sulla protezione dei dati personali [1].

Prima di affrontare questo argomento, è essenziale chiarire cosa intendiamo per big data, poiché questa definizione pone grande enfasi sulla dimensionalità. Quando parliamo di big data, dobbiamo fare un confronto con la scala di esperienze che abbiamo nella vita quotidiana. Ad esempio, un'ora di film ad alta definizione occupa circa 3 gigabyte (GB) di spazio di archiviazione. La scala successiva del gigabyte è il terabyte, che corrisponde a 1.024 gigabyte. Ciò significa che un terabyte equivale a circa 341 ore di film in alta definizione. La scala successiva al terabyte è il petabyte (PB), che corrisponde a 1.024 terabyte (TB). Questa dimensione è equivalente a circa mille dischi rigidi del tipo che possono essere installati oggi su un moderno computer portatile.

Abbiamo parlato di film e se vogliamo fare un confronto tra un servizio di streaming più conosciuto al giorno d'oggi, la sua libreria di film ha una dimensione di circa 60 petabyte. Adesso, se volessimo fare un confronto più nell'ambito della genomica e quindi spostarci da un quotidiano e dall'intrattenimento verso un'ambito più medico-biologico, un genoma umano occupa su disco circa 120 GB una volta sequenziato in modo appropriato. Ciò significa che nella libreria di uno dei più

famosi servizi di streaming potremmo tranquillamente ospitare circa 520.000 genomi umani.

Continuando sulla strada della dimensionalità in ambito biomedico, possiamo volgere lo sguardo al database internazionale più comunemente utilizzato in questi anni per depositare le sequenze genomiche di coronavirus SARS-CoV-2. Al marzo 2023, questo database ospitava circa 15 milioni di sequenze. 633.200 circa di queste sequenze sono state depositate da parte della Danimarca, 3 milioni circa sono state depositate dalla Gran Bretagna, 178.130 sono state depositate dall'Italia. Se facciamo un confronto con la popolazione, quella danese è di 5 milioni e 800 mila unità, il che significa che la Danimarca, al marzo 2023, aveva sequenziato il 10.8% della propria popolazione da un punto di vista degli individui in cui era stata riscontrata l'infezione da coronavirus.

La Gran Bretagna ha una popolazione di 67 milioni e 326 mila individui, per cui le sequenze di coronavirus depositate dalla Gran Bretagna corrispondono a circa il 4,5% degli individui della propria popolazione. Se facciamo lo stesso confronto con l'Italia, che ha una popolazione di 59 milioni 109 mila 668 persone, le sequenze di coronavirus depositate nella database internazionale corrispondono allo 0.3% degli individui della nostra popolazione [2].

Continuando sulla grandezza dei database di ambito biomedico, ad esempio, un database genomico internazionalmente conosciuto è il Sequence Read Archive (SRA) negli Stati Uniti: un database che nell'ultima decade è cresciuto fino ad ospitare 25.6 petabyte di sequenze. Come abbiamo raggiunto in questi ultimi anni, queste grandi quantità di dati depositate in database accessibili a livello internazionale? Attraverso strumenti che consentono sempre più facilmente, nel caso della genomica, di ottenere sequenze dei genomi di organismi di interesse. Ad esempio, oggi ci sono almeno tre aziende che commercializzano grandi sequenziatori e queste macchine consentono in un intervallo che va dalle 24 alle 72 ore di sequenziare un equivalente di sequenze che va da 6 terabyte a 16 terabyte. Ciò corrisponde a circa 27.344 genomi umani diploidi.

Abbiamo finora discusso di dimensionalità.

In questo mio contributo, cercherò tuttavia di dimostrare che l'importanza dei big data non sta nella dimensione, nella grandezza dei dati, ma nella disponibilità di un'informazione che noi possiamo utilizzare per raccogliere nuove conclusioni, e costruire nuove prospettive su fenomeni complessi.

Informazione e big data

Un esempio significativo è rappresentato dal database internazionale di sequenze genomiche "GnomAD". Nella sua versione 3.1, questo database contiene 76.156 genomi di individui senza rapporti di parentela. La versione precedente includeva 15.788 sequenze genomiche e 125.748 sequenze di esoma, ovvero la porzione del genoma che codifica per le proteine. Sebbene questi numeri possano sembrare piccoli rispetto ai milioni di sequenze discusse in precedenza, essi sono di fondamentale importanza. Le sequenze di questi individui, infatti, hanno permesso di confrontare la diversità genomica umana a livello globale, offrendo un quadro estremamente accurato e una vasta disponibilità di dati per rappresentare la diversità ancestrale della popolazione umana [3,4].

Questo è particolarmente rilevante poiché, negli ultimi anni, si è sviluppata una conversazione significativa nella comunità scientifica internazionale riguardo al fatto che la maggior parte dei database di sequenze genomiche rappresentavano principalmente individui di origine europea appartenenti a popolazioni caucasiche, non riflettendo adeguatamente la diversità delle popolazioni umane. Grazie agli sforzi fatti per costruire una maggiore diversità ancestrale, questo database colma parzialmente questa lacuna.

Inoltre, il database ha permesso di studiare in dettaglio le variazioni di sequenza all'interno del genoma umano, incluse quelle più rare. Con l'inclusione di un numero così elevato di genomi, è stato possibile descrivere e confrontare le frequenze di tutte le variazioni genetiche scoperte. Questo ha portato a una migliore comprensione dei geni nel genoma umano maggiormente sottoposti a pressione selettiva, facendo emergere che alcune varianti tendono a scomparire attraverso le generazioni in un processo noto come "purifying selection". Ad esempio, è stato possibile determinare che i geni soggetti a mutazioni che causano una perdita di funzione mostrano solo il 48% delle variazioni attese, indicando che la maggior parte dei geni nel nostro genoma sono sottoposti a una selezione negativa che evita la trasmissione di mutazioni deleterie alle generazioni successive. Questo ha permesso di identificare geni intolleranti alle mutazioni cosiddette "loss of function" [4].

Nonostante l'importanza dei dati in termini di quantità, è la qualità dell'analisi a rivelare il vero valore di queste informazioni. La gestione di tali enormi quantità di dati pone significative sfide computazionali. Ad

esempio, conservare i dati di GnomAD occuperebbe circa 900 terabyte, ma grazie a formati di file innovativi sviluppati per la versione 3.1, l'archiviazione richiede solo 20 terabyte, riducendo drasticamente lo spazio necessario. Inoltre, gli algoritmi di calcolo innovativi hanno ridotto significativamente i costi di aggiunta incrementale di informazioni, da 13.000 dollari a circa 400 dollari per l'inclusione di 4.598 nuovi genomi.

Un altro esempio rilevante è uno studio sulle varianti rare associate al rischio di sviluppare schizofrenia, a cui abbiamo contribuito. Questo studio, inizialmente avviato dal consorzio danese chiamato iPSYCH, è stato successivamente ampliato a livello internazionale, coinvolgendo numerosi gruppi di ricerca. Lo studio ha analizzato 24.248 casi di schizofrenia e 97.322 controlli [5]. Le varianti rare coinvolte nel rischio di schizofrenia, spesso difficili da individuare, sono state identificate grazie all'elevato numero di campioni. Sono stati identificati 10 geni associati a un rischio significativo di sviluppare schizofrenia. In particolare, varianti rare che troncano la proteina conferiscono un rischio relativo tra 1.2 e 1.3 nei geni intolleranti alle mutazioni che causano perdita di funzione.

Questi risultati riflettono fenomeni osservati nelle mutazioni cosiddette *de-novo*, che compaiono spontaneamente in un individuo senza essere ereditate dai genitori. I geni identificati appartengono a tre categorie principali, corrispondenti alle ipotesi di sviluppo della malattia: geni coinvolti nella migrazione neuronale, come TRIO; geni coinvolti nel trasporto ionico, come GRIN2A, CACNA1G, e GRIA3; e geni coinvolti in altre funzioni chiave dello sviluppo cerebrale. Senza una così ampia numerosità di dati, sarebbe stato impossibile individuare questi geni e comprendere meglio i meccanismi molecolari alla base della schizofrenia [5].

Infrastrutture per i big data

Nei paragrafi precedenti è emerso chiaramente che la grandezza del dato non dovrebbe essere l'elemento più importante dei big data, ma che centrale è la capacità di estrarre informazioni da essi e i metodi utilizzati per farlo. Tuttavia, come già accennato, per processare questi tipi di dati sono necessarie infrastrutture capaci di svolgere calcoli adeguati e di conservare tali dati. Ad esempio, un petabyte di dischi occupa solitamente una parete di una stanza di medie dimensioni, oppure

richiede grandi aggregati di computer, detti cluster, di cui esistono diversi esempi.

Il cluster che abbiamo utilizzato in Danimarca per lo studio sulla schizofrenia nei campioni Danesi, ad esempio, occupava due pareti di una stanza di medie dimensioni, aveva un totale di 6.776 processori distribuiti su 213 nodi, con una capacità di archiviazione totale di 16 petabyte. Questo tipo di infrastruttura consente di processare 200 esomi in meno di 15 ore. Tuttavia, stiamo assistendo a un'evoluzione di queste infrastrutture, con una sempre maggiore centralizzazione delle risorse computazionali, come nel caso del cloud computing, dove grandi centri di calcolo concentrano le infrastrutture necessarie, permettendo a numerosi clienti di accedervi da remoto. Ciò consente di creare rapidamente l'infrastruttura necessaria come sottoinsieme di una infrastruttura fisica esistente che occupa spesso grandi capannoni industriali.

In Italia, ad esempio, grazie ai fondi del PNRR, è stato avviato un consorzio nazionale chiamato ICSC, uno dei cinque centri nazionali istituiti dal PNRR, e dedicat a High Performance Computing, Big Data e Quantum Computing, con circa 320 milioni di euro e più di 50 partner pubblici e privati.

Intelligenza Artificiale e big data

Nei paragrafi precedenti ho sottolineato più volte che la vera rivoluzione dei big data non risiede tanto nella loro dimensione quanto nella capacità di connettere le informazioni fra loro. Questo è un tema a me molto caro, poiché l'obiettivo di quest'area, su cui anche noi stiamo lavorando, è essenzialmente quello di costruire un nuovo modello della complessità. Questo approccio è profondamente diverso da quello che la biologia ha adottato fino ad ora, basato su un metodo di lavoro di tipo riduzionista. Tale metodo scompone un sistema in componenti più piccole, più facili da comprendere e su cui si possono eseguire esperimenti mirati, tentando poi di comprendere il sistema nel suo complesso ricostruendo i pezzi a partire dalle informazioni raccolte in questi frammenti dell'organismo [6].

Oggi, grazie alle grandi infrastrutture di calcolo e ai metodi disponibili, come l'intelligenza artificiale, possiamo adottare invece un approccio completamente diverso: quello di integrare la grande quantità di dati a disposizione e costruire un modello che rappresenti la

complessità biologica nel suo insieme, senza scomporre le sue parti. Un'area di grande interesse per il nostro laboratorio è, ad esempio, il deep learning, sviluppato utilizzando reti neurali artificiali. Questo approccio è estremamente importante perché il deep learning consente di catturare relazioni non lineari tra i componenti biologici [7].

Questo è cruciale poiché, fino ad oggi, la maggior parte dei modelli lineari in biologia è riuscita a catturare molte informazioni importanti, consentendo significativi progressi nella conoscenza dei sistemi biologici. Tuttavia, stiamo raggiungendo un limite oltre il quale restano ancora aspetti cruciali che l'approssimazione lineare delle relazioni all'interno dei sistemi biologici non riesce a individuare. Intendiamo utilizzare il deep learning per catturare questi pattern nascosti, applicando questi metodi non solo alla predizione del comportamento dei sistemi biologici, ma anche alla comprensione dei loro meccanismi di funzionamento.

Questa è un'area in grande sviluppo chiamata *explainability*, che consente di utilizzare i meccanismi di predizione delle reti neurali per risalire al modo in cui queste costruiscono le predizioni, studiando gli input e le relazioni scoperte dalle reti neurali. In questo modo, è possibile individuare la rilevanza dei dati analizzati dalla rete, consentendo anche di effettuare nuove scoperte. Questo aggiunge strumenti di indagine innovativi, oltre alla capacità intrinseca delle reti neurali di classificare o predire il comportamento dei sistemi biologici [8,9].

Big data, privacy e ambiente

Se inseriamo questo nel contesto biomedico, in cui la quantità di dati raccolti sui singoli individui sta aumentando giorno dopo giorno, emergono naturalmente numerosi problemi e aspetti da considerare relativi alla privacy e all'utilizzo di questi dati. Da un lato, vi è una crescente necessità di protezione dei dati personali. Dall'altro, vi è anche la necessità di sviluppare meglio il concetto di consenso dinamico, ovvero la capacità di un individuo non solo di revocare, ma anche di modificare il consenso dato per l'utilizzo dei propri dati personali.

Questa rappresenta una grande frontiera e una sfida significativa, poiché comporta numerosi problemi tecnici. Devono esistere infrastrutture e metodi di accesso per consentire la modifica del consenso sui dati personali nel tempo, e questo non è affatto semplice da realizzare.

Inoltre, esiste un importante aspetto ambientale quando si parla di

big data. La necessità di disporre di grandi infrastrutture di calcolo implica che queste strutture saranno anche grandi fonti di consumo energetico, con un impatto ambientale non trascurabile. Questo include anche l'uso delle materie prime necessarie per costruire tali infrastrutture [10].

In sintesi, mentre la gestione e l'analisi dei big data offrono enormi potenzialità per la biomedicina, è fondamentale affrontare le sfide relative alla privacy dei dati e all'impatto ambientale associato alle infrastrutture di calcolo necessarie.

Conclusioni

In questo articolo ho evidenziato diversi aspetti rilevanti quando si parla di big data. Innanzitutto, il termine "big data" tende a farci focalizzare sulla quantità dei dati, mentre l'attenzione dovrebbe essere posta sulla qualità dei dati e sulla capacità di mettere in relazione le informazioni tra loro. Ho sottolineato inoltre come le infrastrutture e la disciplina della bioinformatica siano fondamentali per dare un significato a queste grandi quantità di dati che stiamo raccogliendo.

È altresì importante sottolineare come l'intelligenza artificiale, insieme alla capacità di spiegare i risultati ottenuti attraverso di essa, possa offrire nuovi strumenti per rivoluzionare il modo in cui comprendiamo e modelliamo gli organismi biologici. In particolare, l'explainability dell'intelligenza artificiale è cruciale per interpretare e validare i modelli predittivi, contribuendo così a una comprensione più approfondita dei meccanismi biologici.

In sintesi, l'attenzione nei big data dovrebbe essere rivolta alla qualità e alla correlazione dei dati, supportata da robuste infrastrutture bioinformatiche e dalle avanzate capacità analitiche dell'intelligenza artificiale, per migliorare la nostra comprensione dei sistemi biologici.

Ringraziamenti

Programma di ricerca CN00000013 "National Centre for HPC, Big Data and Quantum Computing", finanziato dal Decreto Direttoriale di concessione del finanziamento n.1031 del 17.06.2022 a valere sulle risorse del PNRR MUR – M4C2 – Investimento 1.4 - Avviso "Centri Nazionali" - D.D. n. 3138 del 16 dicembre 2021.

BIBLIOGRAFIA

1. Goyal, I.; Singh, A.; Saini, J.K. Big Data in Healthcare: A Review. In Proceedings of the 2022 1st International Conference on Informatics (ICI); IEEE: Noida, India, April 14 2022; pp. 232-234.
2. Kontis, V.; Bennett, J.E.; Rashid, T.; Parks, R.M.; Pearson-Stuttard, J.; Guillot, M.; Asaria, P.; Zhou, B.; Battaglini, M.; Corsetti, G.; et al. Magnitude, Demographics and Dynamics of the Effect of the First Wave of the COVID-19 Pandemic on All-Cause Mortality in 21 Industrialized Countries. *Nature Medicine 2019* 2020, *26*, 1919-1928, doi:10.1038/s41591-020-1112-0.
3. Chen, S.; Francioli, L.C.; Goodrich, J.K.; Collins, R.L.; Kanai, M.; Wang, Q.; Alföldi, J.; Watts, N.A.; Vittal, C.; Gauthier, L.D.; et al. A Genomic Mutational Constraint Map Using Variation in 76,156 Human Genomes. *Nature* 2024, *625*, 92–100, doi:10.1038/s41586-023-06045-0.
4. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; ldi, J.A. x000F6; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. *Nature* 2021, 1–24, doi:10.1038/s41586-020-2308-7.
5. Singh, T.; Poterba, T.; Curtis, D.; Akil, H.; Al Eissa, M.; Barchas, J.D.; Bass, N.; Bigdeli, T.B.; Breen, G.; Bromet, E.J.; et al. Rare Coding Variants in Ten Genes Confer Substantial Risk for Schizophrenia. *Nature* 2022, *604*, 509–516, doi:10.1038/s41586-022-04556-w.
6. Jiang, Y.; Li, X.; Luo, H.; Yin, S.; Kaynak, O. Quo Vadis Artificial Intelligence? *Discov Artif Intell* 2022, *2*, 4, doi:10.1007/s44163-022-00022-8.
7. Watson, D. Interpretable Machine Learning for Genomics. 2021, doi:10.21203/rs.3.rs-448572/v1.
8. Confalonieri, R.; Coba, L.; Wagner, B.; Besold, T.R. A Historical Perspective of Explainable Artificial Intelligence. *WIREs Data Mining Knowl Discov* 2021, *11*, doi:10.1002/widm.1391.
9. Santorsola, M.; Lescai, F. The Promise of Explainable Deep Learning for Omics Data Analysis: Adding New Discovery Tools to AI. *New Biotechnology* 2023, *77*, 1–11, doi:10.1016/j.nbt.2023.06.002.
10. Lannelongue, L.; Grealey, J.; Inouye, M. Green Algorithms: Quantifying the Carbon Footprint of Computation. *Advanced Science* 2021, *8*, 2100707, doi:10.1002/advs.202100707.