Biologia strutturale e intelligenza artificiale: una rivoluzione multidisciplinare nella comprensione delle strutture proteiche

Federico Forneris*

SUNTO – La biologia strutturale è una disciplina affascinante e in continua evoluzione, che si occupa dello studio della struttura tridimensionale delle macromolecole biologiche e della loro relazione con la funzione cellulare. Attraverso avanzate tecniche sperimentali e computazionali, si procede alla progressiva comprensione dei meccanismi molecolari che governano la vita, con implicazioni dirette sulla salute umana attraverso la comprensione dei meccanismi molecolari alla base di molteplici malattie, aprendo prospettive innovative per lo sviluppo di farmaci specifici ed efficaci. L'intervento che segue si propone di offrire un'introduzione alla biologia strutturale, con particolare attenzione al problema del *folding* delle proteine e all'impatto rivoluzionario che l'intelligenza artificiale sta avendo da qualche anno in questo settore. In relazione alla tematica e per sottolineare quanto gli strumenti basati sull'intelligenza artificiale possano rappresentare, se ben utilizzati, una preziosa risorsa per la ricerca scientifica, desidero precisare di aver fatto uso di strumenti basati sull'intelligenza artificiale nella redazione del presente testo. In particolare, la prima stesura di questa relazione è stata generata dal *software* ChatGPT a partire dalla trascrizione dell'intervento tenuto presso l'Istituto Lombardo, ottenuta anch'essa sfruttando l'intelligenza artificiale facendo analizzare la registrazione, disponibile su YouTube, al *software* Panopto.

PAROLE CHIAVE – Biologia strutturale; *Folding* delle proteine; Intelligenza artificiale.

ABSTRACT – Structural biology is a fascinating and constantly evolving discipline that focuses on studying the three-dimensional structure of biological macromolecules and their relationship with cellular functions. Through advanced experimental and computational techniques, researchers progressively gain a deeper understanding of the molecular mechanisms

* Laboratorio Armenise-Harvard di Biologia strutturale, Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Via Ferrata 9A, 27100 Pavia, http://fornerislab.unipv.it. E-mail: federico.forneris@unipv.it. Relazione tenuta il 7 marzo 2024.

Rendiconti 158.2/2025, Doi 10.3280/rndoa2025oa-21153

that govern life. This knowledge has direct implications for human health by shedding light on the molecular basis of various diseases, opening up innovative perspectives for the development of targeted and effective drugs. The following presentation aims to provide an introduction to structural biology, with particular emphasis on the problem of protein folding and the revolutionary impact that artificial intelligence has had in this field in recent years. In relation to this topic, and to highlight how AI-based tools can serve as a valuable resource for scientific research when used appropriately, I would like to clarify that I have utilized artificial intelligence tools in drafting this text. Specifically, the first draft of this presentation was generated by the ChatGPT software based on the transcription of a talk given at the Lombard Institute. This transcription was also obtained using artificial intelligence by analyzing the recording of the talk available on YouTube through the Panopto software.

KEYWORDS – Structural biology; Protein folding; Artificial intelligence.

1. IL MONDO INVISIBILE DELLA BIOLOGIA MOLECOLARE

Quando si parla di cellule e processi biologici, spesso si ha un'immagine semplificata della realtà microscopica. Nei libri di testo, le cellule vengono rappresentate come ambienti spaziosi e ordinati, con pochi elementi distinti che interagiscono tra loro. Tuttavia, la realtà è molto più complessa, e per comprendere davvero il funzionamento della cellula bisogna scendere a livello molecolare. Qui, proteine, DNA, RNA e altre biomolecole interagiscono costantemente in un intricato "balletto molecolare", che rende possibile la vita. La biologia strutturale si occupa proprio di studiare come queste molecole comunicano tra loro e come la loro forma tridimensionale influenzi la loro funzione. Comprendere questa relazione è cruciale per molteplici applicazioni, dalla biotecnologia alla medicina.

Particolarmente sfidante è poi riuscire a integrare l'informazione relativa a singole macromolecole all'interno di complesse architetture quali organelli, compartimenti sub-cellulari o intere cellule. Un esempio significativo è il modello dettagliato dell'organizzazione molecolare di una cellula di *Mycoplasma genitalium*, reso possibile attraverso sforzi congiunti di ricerca sperimentale, modellistica e grafica molecolare (Maritan *et al*., 2022). Questa rappresentazione (disponibile in modalità interattiva all'url: *https://ccsb.scripps.edu/gallery/mycoplasma_model/*, consultato il 18 settembre 2025) mostra un ambiente incredibilmente affollato, in cui proteine, acidi nucleici e altre macromolecole interagiscono costantemente per mantenere il funzionamento cellulare. Ogni singola proteina svolge una funzione specifica, contribuendo al

metabolismo cellulare, alla regolazione dell'espressione genica, alla risposta agli stimoli esterni e ad altre funzioni essenziali.

Tra tutte le macromolecole biologiche, le proteine sono particolarmente affascinanti: questi polimeri composti da lunghe catene di amminoacidi costituiscono i mattoni fondamentali della vita, svolgendo funzioni strutturali, catalitiche ed enzimatiche, regolatorie e molto altro. Ma il segreto della loro funzione sta nella loro forma tridimensionale, un aspetto che la biologia strutturale cerca di comprendere da decenni. Il processo dinamico attraverso cui una proteina raggiunge la propria configurazione funzionale è noto come *folding*. Questo meccanismo è estremamente complesso, in quanto la stessa catena di amminoacidi potrebbe teoricamente assumere un numero enorme di conformazioni. Tuttavia, in natura, ogni proteina adotta una sola o poche configurazioni stabili e funzionali.

Fin dalle prime determinazioni strutturali, la complessità nel ripiegamento della sequenza proteica in articolate strutture tridimensionali ha rappresentato una componente particolarmente sfidante della comprensione delle macromolecole. A differenza del DNA, la cui struttura a doppia elica appariva straordinariamente regolare e intuitivamente semplice da visualizzare, le proteine presentavano architetture più difficili da rappresentare e, di conseguenza, da interpretare. Un esempio evidente di questo è dato dalla presentazione della struttura cristallografica del lisozima fatta da David Phillips alla Royal Society negli anni '60, in cui per evidenziare la difficoltà del problema, Phillips mostrò due modelli metallici fianco a fianco. Il primo era il modello della proteina correttamente ripiegata. Questo modello era piccolo, compatto e altamente organizzato per via delle estese interazioni non covalenti tra gli amminoacidi ripiegati a costituire la struttura globulare. Il secondo era il modello della stessa proteina, completamente srotolato: un filo metallico che si estendeva fino al soffitto della hall della Royal Society, mostrando quanto fosse lunga la sequenza amminoacidica in assenza di un ripiegamento ordinato.

Oggi, grazie a tecniche sperimentali avanzate e all'intelligenza artificiale, decifrare la relazioni struttura-funzione di molte proteine (e quindi decifrare, almeno in parte, il problema del *folding*) è diventato una sfida sempre più possibile. I prossimi paragrafi presentano una sintetica descrizione dei momenti chiave dell'evoluzione della biologia strutturale, il problema del *folding* delle proteine, le metodiche sperimentali di più frequente utilizzo per la determinazione delle strutture molecolari delle proteine.

2. I primi studi sulle proteine e la rivoluzione della cristallografia a raggi X

I primi studi biochimici contenenti descrizioni legati all'isolamento e alla caratterizzazione delle proteine risalgono al XIX secolo, ma la loro struttura tridimensionale rimase un mistero fino alla metà del secolo successivo.

Uno dei primi a proporre un modello strutturale fu Linus Pauling, che nel 1951 propose per la prima volta, quali elementi ricorrenti nell'architettura molecolare delle proteine, le strutture secondarie a α-elica e a foglietto β (Eisenberg, 2003). Queste intuizioni furono fondamentali per comprendere la complessità delle proteine e prepararono il terreno per scoperte ancora più grandi.

La vera svolta avvenne alla fine degli anni '50, quando Max Perutz e John Kendrew riuscirono a determinare le strutture tridimensionali rispettivamente di mioglobina e dell'emoglobina usando la cristallografia a raggi X (Strandberg, 2009). Questa tecnica permette di analizzare le proteine in forma cristallina, rivelandone l'organizzazione atomica. Da allora, la cristallografia a raggi X ha permesso di risolvere migliaia di strutture proteiche, fornendo informazioni cruciali su enzimi, recettori cellulari e proteine coinvolte in malattie. A queste si sono uniti progressivamente i modelli molecolari ottenuti attraverso altre tecniche sperimentali (descritti nel §4), mettendo a disposizione dei ricercatori preziose informazioni circa l'architettura molecolare di proteine e complessi macromolecolari presenti in una moltitudine di organismi diversi.

3. Il problema del *folding* delle proteine: una sfida scientifica aperta

Una delle domande più difficili della biologia molecolare è: *come fa una proteina a ripiegarsi nella sua forma tridimensionale corretta?* Il *folding* delle proteine è uno dei più intriganti e difficili problemi scientifici aperti. In teoria, data una sequenza di 100 amminoacidi, esistono circa 10^100 possibili combinazioni tridimensionali plausibili (Regan *et al*., 2015). Tuttavia, solo una è biologicamente attiva.

Al problema combinatorio si aggiunge una componente dinamica. Il processo di biosintesi delle proteine inizia a partire dai ribosomi, da cui queste macromolecole vengono generate come lunghe catene di amminoacidi prive di una struttura tridimensionale definita. Alla fine degli anni '60, il fisico

Cyrus Levinthal sottolineò l'apparente paradosso del *folding* delle proteine. Se una proteina dovesse esplorare tutte le possibili conformazioni prima di trovare quella corretta, ci vorrebbero miliardi di anni. Bastano tuttavia pochi millisecondi perché queste catene si ripieghino spontaneamente in una singola architettura tridimensionale, precisa e funzionale (Levinthal *et al*., 1968; Id., 1969).

Tutto ciò è possibile poiché le proteine non esplorano a caso tutte le conformazioni, ma seguono percorsi energeticamente favorevoli, in cui i legami chimici e le interazioni tra amminoacidi guidano il *folding* in modo estremamente rapido ed efficiente. Tanto che errori sistematici nel *folding* proteico che portano ad architetture tridimensionali stabili differenti da quelle fisiologicamente funzionali sono causa di gravi malattie, tra le quali figurano il morbo di Alzheimer e il morbo di Parkinson, causati dall'aggregazione di proteine non correttamente ripiegate (Fitzpatrick *et al*., 2017; Yang *et al*., 2022).

4. Le principali tecniche sperimentali per studiare la struttura delle proteine

Accanto alla cristallografia a raggi X, negli ultimi decenni sono state sviluppate diverse tecniche sperimentali per studiare la struttura e il *folding* delle proteine, che rappresentano al giorno d'oggi un importante portfolio di metodologie applicabili per la caratterizzazione di complesse architetture molecolari. Ciascuna tecnica presenta alcune prerogative legate al proprio funzionamento, cui sono associati vantaggi e svantaggi applicativi. Questo fa sì che non esista una tecnica che può essere considerata universalmente migliore di altre: la scelta della migliore metodica di indagine è un passaggio fondamentale nelle indagini di biologia strutturale, poiché a seconda della tipologia di campione oggetto di studio e agli obiettivi della ricerca, un metodo potrebbe presentare indubbi vantaggi o svantaggi rispetto a un altro.

- La cristallografia a raggi X ha per decenni rappresentato l'unico approccio allo studio della struttura tridimensionale delle proteine. Indubbi vantaggi sono la possibilità di ottenere dati a risoluzione atomica (1-2 Å), permettendo la comprensione dei meccanismi di interazione tra proteine e piccole molecole, come ad esempio le relazioni enzima: substrato o i processi di interazione alla base del meccanismo di farmaci specifici per bersagli molecolari ben definiti. Le strutture tridimensionali delle proteine vengono generate attraverso l'interpretazione del dato sperimentale,

sotto forma di mappa di densità elettronica all'interno della quale è possibile modellare le posizioni dei singoli atomi presenti all'interno della struttura macromolecolare. Questa tecnica, tuttavia, richiede che le proteine vengano cristallizzate, un processo spesso complesso e non sempre possibile. Inoltre, le strutture tridimensionali ottenute attraverso la cristallografia a raggi X non consentono di decifrare molti dei processi dinamici alla base delle interazioni osservate, poiché il dato mostra quanto è stato "intrappolato" all'interno del cristallo in merito a quelle che sono le conformazioni della proteina oggetto di studio compatibili con il processo di cristallizzazione.

- La Risonanza magnetica nucleare (NMR) consente di studiare le proteine in soluzione sfruttando le proprietà magnetiche dei nuclei degli atomi presenti negli amminoacidi, evitando la cristallizzazione (Marion, 2013). Si tratta di un approccio con caratteristiche uniche in grado di analizzare proteine caratterizzate da architetture altamente dinamiche (ad esempio, proteine che presentano regioni intrinsecamente disordinate) e i loro cambiamenti conformazionali in soluzione. Da un esperimento NMR su proteine si ottiene una mappa di possibili distanze interatomiche, basata sull'identificazione di specifici segnali di risonanza associati ai nuclei delle specie chimiche presenti all'interno del campione. Questa tecnica è particolarmente adatta allo studio di proteine di piccole dimensioni. Data la complessità delle proteine (costituite da moltissimi atomi ma da pochissimi elementi, C, H, N, O, S), per poter ottenere spettri NMR decifrabili è infatti necessario procedere a complesse procedure di *labeling* isotopico, necessario per far emergere all'interno dello spettro NMR segnali interpretabili associati a specifici residui amminoacidici. Queste procedure, particolarmente onerose dal punto di vista sperimentale, sono in molti casi non applicabili a proteine di grandi dimensioni (>30 kDa), per le quali gli spettri NMR risulterebbero comunque non interpretabili a causa dell'elevato numero di atomi presenti nel campione.
- La microscopia elettronica criogenica (Cryo-EM) è stata per decenni una tecnica relegata allo studio a bassissima risoluzione di pochi e ben definiti sistemi molecolari di grandi dimensioni (ad esempio, i ribosomi). Negli ultimi 10 anni questa tecnica è andata incontro a una vera e propria rivoluzione (Bai *et al*., 2015; Kuhlbrandt, 2014), causata dall'introduzione di strumentazioni, tecniche di manipolazione del campione, automazione e procedure di analisi dati innovative che hanno permesso al Cryo-EM di diventare, oggi, l'approccio più frequentemente utilizzato per lo studio di macromolecole di grandi dimensioni (>100 kDa) e per le proteine di

membrana, notoriamente difficili da cristallizzare. Indubbio vantaggio del Cryo-EM è la possibilità di ovviare al problema della cristallizzazione e alle complesse procedure di *labeling* come nel caso del NMR, e di richiedere quantità di campione davvero ridotte rispetto alle altre tecniche qui descritte. Le moderne immagini di microscopia elettronica, rielaborate attraverso procedure di classificazione per l'individuazione delle proiezioni relative alle proteine, permettono di ottenere ricostruzioni tridimensionali sotto forma di mappe di densità, con risoluzioni locali in molti casi confrontabili con la cristallografia a raggi X. Elemento limitante degli studi Cryo-EM è al momento la dimensione delle proteine oggetto di indagine, in quanto il rapporto segnale/rumore per macromolecole non globulari o di massa inferiore ai 100 kDa risulta molto ridotto e spesso non permette di ottenere ricostruzioni tridimensionali di qualità. Il recente progresso tecnologico associato a questa tecnica (anche attraverso l'implementazione di tecniche basate sull'intelligenza artificiale (Thorn, 2022) è davvero straordinario, rendendo possibili ricostruzioni di proteine di dimensioni sempre più piccole a risoluzioni sempre maggiori (Wu *et al*., 2020).

- La spettrometria di massa, pur non essendo una tecnica "strutturale" per se, sta contribuendo in modo sempre più significativo all'ottenimento di informazioni strutturali accurate relative alle macromolecole biologiche. Moderne strumentazioni consentono infatti la determinazione di spettri di massa di proteine e complessi macromolecolari allo stato nativo, dai quali è possibile risalire a regioni coinvolte in contatti molecolari essenziali per la stabilità di queste architetture tridimensionali allo stato nativo (Heck, 2008). Ulteriori metodiche basate sull'analisi in spettrometria di massa di digeriti triptici di macromolecole trattate con *cross-linkers* chimici (Iacobucci *et al*., 2019), o attraverso lo scambio isotopico in soluzione tra acqua di solvatazione e acqua pesante (c.d. *hydrogen-deuterium exchange*) (Yan *et al*., 2009) permettono di ottenere dettagli accurati circa la posizione di specifiche porzioni di una macromolecola nello spazio tridimensionale che, combinate con altri dati strutturali provenienti da tecniche sperimentali o computazionali, facilitano la determinazione dell'architettura della macromolecola oggetto di studio.

L'importanza delle strutture tridimensionali delle macromolecole biologiche è stata oggetto di riconoscimento fin da subito, portando alla creazione di un archivio digitale, il Protein Data Bank (PDB), che raccoglie, standardizza e mette a disposizione centinaia di migliaia di strutture proteiche determinate sperimentalmente (Berman *et al*., 2003). Il PDB è una risorsa unica per la comunità scientifica internazionale e, vedremo nei prossimi paragrafi, ha contribuito in modo imprescindibile al successo delle tecnologie computazionali basate sull'intelligenza artificiale per la modellistica molecolare.

5. Gli approcci computazionali per studiare la struttura delle proteine

L'utilizzo della bioinformatica per cercare di comprendere l'architettura tridimensionale delle proteine risale a molto prima dell'avvento delle tecniche basate sull'intelligenza artificiale. Tra le metodiche *in silico* più diffuse per lo studio della biologia strutturale si annoverano:

- Le predizioni di struttura tridimensionale di proteine a partire dalla loro sequenza. Sono disponibili moltissimi algoritmi per la predizione di architetture molecolari. La maggior parte di questi metodi si basa sull'applicazione consecutiva di predizioni relative a elementi di struttura secondaria (α-eliche e β-foglietti) e di vincoli (*constraints*) di interazione tra aminoacidi presenti all'interno delle sequenze proteiche basati statisticamente su architetture determinate sperimentalmente in precedenza. Il problema principale di questi metodi è sempre stato quello relativo all'incertezza della predizione, tanto che da decenni si tiene una competizione aperta agli sviluppatori (CASP, Elofsson, 2023) per identificare le migliori metodiche e migliorare iterativamente gli algoritmi di predizione.
- Le predizioni di siti di interazione tra proteine, o tra proteine e piccole molecole. Queste metodiche (dette di *docking*) sfruttano approcci statistici basati sul posizionamento delle molecole oggetto di interazione in molteplici pose differenti, compatibili dal punto di vista dei contatti tridimensionali, della complementarietà dei contatti elettrostatici e dei contatti tra regioni idrofobiche adiacenti (Morris *et al*., 2008). Le pose testate vengono poi disposte in ordine di plausibilità in base al calcolo di molteplici parametri, tra i quali tipicamente vengono considerati l'energia libera di interazione/dissociazione e l'effettiva complementarietà sterica. Nel contesto del *drug discovery*, questi approcci rappresentano al giorno d'oggi

un passaggio fondamentale prima di affrontare qualsiasi tipo di indagine sperimentale, poiché consentono una rapida valutazione della plausibilità delle interazioni molecolari a una frazione esigua del costo di una validazione sperimentale. Il successo di queste metodiche è intrinsecamente legato alla presenza di molteplici contatti non ambigui tra le molecole oggetto di interazione: è indubbiamente più semplice svolgere un esperimento di *docking* di una piccola molecola all'interno di una tasca profonda dalla forma ben definita all'interno di una proteina rispetto all'individuare un possibile sito di interazione su di una superficie convessa priva di cavità. Analogamente, il *docking* è sicuramente di più facile applicazione in assenza di cambiamenti conformazionali e processi dinamici che possono modificare la forma dei siti di interazione, rendendoli compatibili per il contatto oggetto di studio solo quando l'interazione è di fatto avvenuta.

- Le simulazioni di dinamica molecolare, con cui è possibile effettuare predizioni del comportamento di macromolecole in soluzione per intervalli temporali nell'ordine dei micro/millisecondi, tipicamente sufficienti per valutare processi legati alla diffusione di substrati all'interno di enzimi o la stabilità di una specifica interazione in presenza di particolari condizioni di pH, di temperatura, o di forza ionica. Il progresso tecnologico associato al calcolo computazionale consente di svolgere simulazioni sempre più accurate ed estese su sistemi macromolecolari di dimensioni crescenti, rendendo queste tecniche di calcolo sempre più utilizzate per la validazione di dati sperimentali e, in congiunzione con le tecniche di *docking* discusse al punto precedente, di prendere in considerazione gli aspetti dinamici del riconoscimento molecolare per una migliore valutazione e quantificazione delle interazioni tra macromolecole.

La differenza fondamentale tra le tecniche predittive e quelle basate sul *docking* e sulla dinamica molecolare è che le prime ambiscono a ovviare all'utilizzo di metodiche sperimentali per la determinazione dell'architettura tridimensionale delle macromolecole oggetto di studio, mentre le altre fanno largo uso delle informazioni già disponibili sperimentalmente per fornire informazioni complementari e favorire sviluppi applicativi attraverso di esse.

6. La rivoluzione di AlphaFold e dei metodi di predizione di strutture tridimensionali basati sull'intelligenza artificiale

L'intelligenza artificiale ha nuovamente rivoluzionato la biologia strutturale. Uno degli sviluppi più importanti è stato AlphaFold (Jumper *et al*., 2021), un *software* sviluppato da DeepMind (Google), presentato nel gennaio 2020 e salito agli onori delle cronache per essere stato in grado di prevedere la struttura di migliaia di proteine con un'accuratezza vicina a quella sperimentale, di fatto superando di ordini di grandezza quanto ottenuto con gli algoritmi di predizione strutturali utilizzati fino a quel momento, come evidenziato dai risultati del CASP 2020 (Kryshtafovych *et al*., 2021).

L'algoritmo di AlphaFold2 utilizza reti neurali profonde per predire la struttura tridimensionale di una proteina basandosi sulla sua sequenza amminoacidica. Combinato con l'analisi delle relazioni evolutive tra le migliaia di strutture proteiche disponibili attraverso il PDB, genera predizioni accurate delle strutture tridimensionali di proteine, finora mai caratterizzate. Questo approccio, una volta messo a disposizione della comunità scientifica (dopo insistenza della stessa nei confronti di Google, che aveva inizialmente circolato versioni "chiuse" e molto limitate di AlphaFold2) ha reso più rapida, efficiente e accurata la predizione di strutture proteiche rispetto a qualsiasi altro algoritmo disponibile in quel momento. Un secondo fondamentale traguardo è stato raggiunto quando AlphaFold2 è stato applicato sistematicamente a tutte le sequenze proteiche note presenti negli organismi i cui genomi sono stati sequenziati (Tunyasuvunakool *et al*., 2021; Varadi *et al*., 2022), di fatto mettendo a disposizione della comunità scientifica modelli predittivi di milioni di architetture molecolari mai studiate in precedenza.

Sviluppi successivi, sia da parte del team di Google Deepmind che di altri gruppi di ricerca (tra i quali anche il gruppo di David Baker, tra i principali attori coinvolti nello studio computazionale delle strutture proteiche) hanno messo a disposizione svariati algoritmi basati sull'intelligenza artificiale, con i quali è oggi possibile effettuare predizioni di architetture, non solo di proteine, ma anche di complessi contenenti ligandi e modificazioni post-traduzionali (Humphreys *et al*., 2021; Thompson *et al*., 2024; Wohlwend *et al*., 2024).

7. La fine della biologia strutturale sperimentale? Tutt'altro

La narrazione di cui sopra può suggerire al lettore inesperto che gli sforzi finora condotti per determinare sperimentalmente strutture tridimensionali di proteine siano da considerarsi quasi obsoleti, in virtù dell'avvento di tecnologie computazionali con accuratezza molto elevata. Ci sono tuttavia alcuni aspetti che sottolineano come la realtà attuale sia assolutamente diversa da questo scenario.

Prima di tutto, è necessario considerare che gli algoritmi di predizione strutturale basati su intelligenza artificiale hanno ottenuto questi straordinari risultati grazie a *training* effettuato interpretando le interazioni molecolari presenti nelle strutture tridimensionali determinate sperimentalmente e messe a disposizione della comunità attraverso il PDB: senza decenni di sforzi sperimentali, nessun algoritmo di intelligenza artificiale sarebbe mai stato in grado di predire alcunché.

In secondo luogo, per quanto le predizioni risultino statisticamente piuttosto accurate, l'errore associato a queste quando si procede al confronto dettagliato con analoghe strutture sperimentali non è trascurabile. Questo fa sì che, nello studio di interazioni molecolari, così come nello sviluppo di farmaci e negli approcci di ingegneria proteica, i modelli predittivi ottenuti mediante tecniche basate sull'intelligenza artificiale risultino meno affidabili rispetto alle strutture sperimentali, soprattutto quando le molecole oggetto di studio sono particolarmente complesse (ad esempio nel caso proteine di membrana o grandi complessi macromolecolari). Scostamenti anche di pochi decimi di Å relativi a specifiche catene laterali di amminoacidi critici per interazioni molecolari possono risultare critici per la comprensione di contatti specifici e per successivi approcci di ottimizzazione di queste interazioni, ad esempio nello sviluppo di farmaci. La precisione offerta dalle metodiche sperimentali resta pertanto un elemento imprescindibile per questo tipo di approcci.

In questo contesto, la maggiore disponibilità di modelli predittivi accurati va considerata come una risorsa aggiuntiva, anziché sostitutiva, dei metodi sperimentali. Disporre di un "punto di partenza" più accurato per elaborare ipotesi di studio sperimentale offre indubbi vantaggi e consente di affrontare indagini più complesse, facendo affidamento (con le dovute cautele!) alla modellistica basata sull'intelligenza artificiale per la comprensione delle architetture "base" che caratterizzano i domini e più in generale le "forme" delle macromolecole oggetto di studio (Bai *et al*., 2024). Ed è proprio attraverso questa integrazione degli approcci che stanno emergendo i più

straordinari recenti risultati di ricerca nel campo della biologia strutturale (Mahamid *et al*., 2016; Oikonomou *et al*., 2017; van den Hoek *et al*., 2022).

Va inoltre sottolineato che il problema del *folding* descritto precedentemente è tutt'altro che risolto, poiché anche le più moderne metodiche di indagine basate sull'intelligenza artificiale riescono a fornire predizioni di quello che è lo stato finale (e quindi stabile) di una macromolecola, ma non i dettagli del processo che permette il raggiungimento di questo stato a partire da un polimero caratterizzato da una amminoacidica non ripiegata. Considerata la complessità del processo, anche in questo caso è altamente probabile che, se mai riusciremo a comprendere le basi di questo articolato processo, ci riusciremo combinando insieme una moltitudine di approcci di indagine, operando in sinergia e fornendo anche validazione ortogonale dei dati ottenuti.

Conclusioni e prospettive future

La biologia strutturale ha vissuto numerose "rivoluzioni" che hanno permesso di ampliare le possibilità di indagine attraverso l'utilizzo di una moltitudine di metodiche di indagine. Dalle prime strutture molecolari determinate con la cristallografia a raggi X all'avvento di nuove tecnologie quali l'NMR e più recentemente i "passi da gigante" del Cryo-EM, fino all'introduzione dell'intelligenza artificiale, ogni "rivoluzione" ha esteso le opportunità di comprensione molecolare della materia vivente e lo sviluppo di farmaci molecolari specifici contro bersagli sempre più complessi.

Anche il biologo strutturale è una figura in costante cambiamento: da sempre unisce competenze in ambito sperimentale e computazionale. I recenti progressi evidenziano come la combinazione e l'integrazione delle metodiche a disposizione (e delle competenze a queste associate) sia la ricetta per raggiungere i più ambiziosi risultati di ricerca. L'approccio allo studio delle macromolecole biologiche, sempre più multidisciplinare, ambisce ora a caratterizzare le relazioni struttura-funzione *in situ*, ossia nel contesto biologico in cui queste molecole si trovano a livello fisiologico. Un risultato che già è stato raggiunto attraverso approcci pionieristici di integrazione delle metodiche per un numero molto ristretto di studi, ma che nel prossimo futuro diventerà sempre più convenzionale.

Un punto fermo, nonostante le numerose "rivoluzioni", è la disponibilità delle informazioni circa le strutture tridimensionali delle macromolecole. Il PDB, ora integrato anche con i modelli predetti dall'intelligenza artificiale, e gli altri *database* ora disponibili con le informazioni strutturali determinate

attraverso i numerosi approcci sperimentali disponibili, costituiscono una risorsa fondamentale, preziosa e insostituibile per il futuro. Futuro che, per la biologia strutturale e chi vi opera, è più entusiasmante che mai.

BIBLIOGRAFIA

Bai X.C., Gonen T., Gronenborn A.M., Perrakis A., Thorn A. and Yang J. (2024). *Challenges and opportunities in macromolecular structure determination*. In: «Nat Rev Mol Cell Biol», 25: 7-12. Doi: 10.1038/s41580-023-00659-y.

Bai X.C., McMullan G. and Scheres S.H. (2015). *How cryo-EM is revolutionizing structural biology*. In: «Trends Biochem Sci», 40: 49-57. Doi: 10.1016/j.tibs.2014.10.005.

Berman H., Henrick K. and Nakamura H. (2003). *Announcing the worldwide Protein Data Bank*. In: «Nat Struct Biol», 10: 980. Doi: 10.1038/nsb1203-980.

Eisenberg D. (2003). *The discovery of the α-helix and β-sheet, the principal structural features of proteins*. In: «Proceedings of the National Academy of Sciences», 100: 11207-11210. Doi: 10.1073/pnas.2034522100.

Elofsson A. (2023). *Progress at protein structure prediction, as seen in CASP15*. In: «Curr Opin Struct Biol», 80: 102594. Doi: 10.1016/j.sbi.2023.102594.

Fitzpatrick A.W.P., Falcon B., He S., Murzin A.G., Murshudov G., Garringer H.J., Crowther R.A., Ghetti B., Goedert M. and Scheres S.H.W (2017). *Cryo-EM structures of tau filaments from Alzheimer's disease*. In: «Nature», 547: 185-190. Doi: 10.1038/nature23002.

Heck A.J. (2008). *Native mass spectrometry: a bridge between interactomics and structural biology*. In: «Nat Methods», 5: 927-933. Doi: 10.1038/nmeth.1265.

Humphreys I.R., Pei J., Baek M., Krishnakumar A., Anishchenko I., Ovchinnikov S., Zhang J., Ness T.J., Banjade S., Bagde S.R., Stancheva V.G., Li X.H., Liu K., Zheng Z., Barrero D.J., Roy U., Kuper J., Fernandez I.S., Szakal B., Branzei D., Rizo J., Kisker C., Greene E.C., Biggins S., Keeney S., Miller E.A., Fromme J.C., Hendrickson T.L., Cong Q. and Baker D. (2021). *Computed structures of core eukaryotic protein complexes*. In: «Science», 374: eabm4805. Doi: 10.1126/science.abm4805.

Iacobucci C., Gotze M. and Sinz A. (2019). *Cross-linking/mass spectrometry to get a closer view on protein interaction networks*. In: «Curr Opin Biotechnol», 63: 48-53. Doi: 10.1016/j.copbio.2019.12.009.

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P. and Hassabis D. (2021).

Highly accurate protein structure prediction with AlphaFold. In: «Nature», 596: 583-589. Doi: 10.1038/s41586-021-03819-2.

Kryshtafovych A., Schwede T., Topf M., Fidelis K. and Moult J. (2021). *Critical assessment of methods of protein structure prediction (CASP)-Round XIV*. In: «Proteins», 89: 1607-1617. Doi: 10.1002/prot.26237.

Kuhlbrandt W. (2014). *Cryo-EM enters a new era*. In: «elife», 3: e03678. Doi: 10.7554/eLife.03678.

Levinthal C. (1968). *Are there pathways for protein folding?*. In: «J. Chim. Phys.» 65: 44-45. Doi: 10.1051/jcp/1968650044.

Id. (1969). *How to fold graciously*. In: *Mössbaun spectroscopy in biological systems proceedings*. *Univ. of Illinois Bulletin*. Champaign: University of Illinois Press, 22-26.

Mahamid J., Pfeffer S., Schaffer M., Villa E., Danev R., Cuellar L.K., Forster F., Hyman A.A., Plitzko J.M. and Baumeister W. (2016). *Visualizing the molecular sociology at the HeLa cell nuclear periphery*. In: «Science», 351: 969-972. Doi: 10.1126/science.aad8857.

Marion D. (2013). *An Introduction to biological NMR spectroscopy*. In: «Molecular & Cellular Proteomics», 12: 3006-3025. Doi: 10.1074/mcp.O113.030239.

Maritan M., Autin L., Karr J., Covert M.W., Olson A.J. and Goodsell D.S. (2022). *Building structural models of a whole mycoplasma cell*. In: «J Mol Biol», 434: 167351. Doi: 10.1016/j.jmb.2021.167351.

Morris G.M. and Lim-Wilby M. (2008). *Molecular docking*. In: «Methods Mol Biol», 443: 365-382. Doi: 10.1007/978-1-59745-177-2_19.

Oikonomou C.M. and Jensen G.J. (2017). *Cellular electron cryotomography: toward structural biology in situ*. In: «Annu Rev Biochem», 86: 873-896. Doi: 10.1146/annurev-biochem-061516-044741.

Regan L., Caballero D., Hinrichsen M.R., Virrueta A., Williams D.M. and O'Hern C.S. (2015). *Protein design: past, present, and future*. In: «Biopolymers», 104: 334-350. Doi: 10.1002/bip.22639.

Strandberg B. (2009). *Chapter 1: building the ground for the first two protein structures: myoglobin and haemoglobin*. In: «J Mol Biol» 392: 2-10. Doi: 10.1016/j.jmb.2009.05.087.

Thompson B. and Petric Howe N. (2024). *Alphafold 3.0: the AI protein predictor gets an upgrade*. In: «Nature». Doi: 10.1038/d41586-024-01385-x.

Thorn A. (2022). *Artificial intelligence in the experimental determination and prediction of macromolecular structures*. In: «Curr Opin Struct Biol», 74: 102368. Doi: 10.1016/j.sbi.2022.102368.

Tunyasuvunakool K., Adler J., Wu Z., Green T., Zielinski M., Zidek A., Bridgland A., Cowie A., Meyer C., Laydon A., Velankar S., Kleywegt G.J., Bateman A., Evans R., Pritzel A., Figurnov M., Ronneberger O., Bates R., Kohl S.A.A., Potapenko A., Ballard A.J., Romera-Paredes B., Nikolov S., Jain R., Clancy E., Reiman D., Petersen S., Senior A.W.,

Kavukcuoglu K., Birney E., Kohli P., Jumper J. and Hassabis D. (2021). *Highly accurate protein structure prediction for the human proteome*. In: «Nature», 596: 590-596. Doi: 10.1038/s41586-021-03828-1.

van den Hoek H., Klena N., Jordan M.A., Alvarez Viar G., Righetto R.D., Schaffer M., Erdmann P.S., Wan W., Geimer S., Plitzko J.M., Baumeister W., Pigino G., Hamel V., Guichard P. and Engel B.D. (2022). *In situ architecture of the ciliary base reveals the stepwise assembly of intraflagellar transport trains*. In: «Science», 377: 543-548. Doi: 10.1126/science.abm6704.

Varadi M., Anyango S., Deshpande M., Nair S., Natassia C., Yordanova G., Yuan D., Stroe O., Wood G., Laydon A., Zidek A., Green T., Tunyasuvunakool K., Petersen S., Jumper J., Clancy E., Green R., Vora A., Lutfi M., Figurnov M., Cowie A., Hobbs N., Kohli P., Kleywegt G., Birney E., Hassabis D and Velankar S. (2022). *AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models*. In: «Nucleic Acids Res», 50: D439-D444. Doi: 10.1093/nar/gkab1061.

Wohlwend J., Corso G., Passaro S., Reveiz M., Leidal K., Swiderski W., Portnoi T., Chinn I., Silterra J., Jaakkola T. and Barzilay R. (2024). *Boltz-1 democratizing biomolecular interaction modeling*. In: *bioRxiv*. Doi: 10.1101/2024.11.19.624167.

Wu M. and Lander G.C. (2020). *How low can we go? Structure determination of small biological complexes using single-particle cryo-EM*. In: «Curr Opin Struct Biol», 64: 9-16. Doi: 10.1016/j.sbi.2020.05.007.

Yan X. and Maier C.S. (2009). *Hydrogen/deuterium exchange mass spectrometry*. In: «Methods Mol Biol», 492: 255-271. Doi: 10.1007/978-1-59745-493-3_15.

Yang Y., Arseni D., Zhang W., Huang M., Lövestam S., Schweighauser M., Kotecha A., Murzin A.G., Peak-Chew S.Y., Macdonald J., Lavenir I., Garringer H.J., Gelpi E., Newell K.L., Kovacs G.G., Vidal R., Ghetti B., Ryskeldi-Falcon B., Scheres S.H.W. and Goedert M. (2022). *Cryo-EM structures of amyloid-β 42 filaments from human brains*. In: «Science», 375: 167-172. Doi: doi:10.1126/science.abm7285.