



UNIVERSITÀ DI PISA

Corso di Laurea Triennale in Informatica (L-31)

TESI DI LAUREA

Protein Folding: dai metodi classici per la predizione della struttura di proteine alla rivoluzione di AlphaFold

Relatore

Prof. Paolo Milazzo

Correlatore

Prof. Mario Pirchio

Candidato

Ludovico Venturi

ANNO ACCADEMICO 2020/2021

Riassunto

In questa tesi viene studiato il *protein folding*, ovvero il ripiegamento tridimensionale delle proteine.

In particolare si parlerà del problema della predizione della struttura di proteine attraverso metodi computazionali.

Viene prima affrontato l'argomento in chiave biologica per poi gradualmente passare ad una visione bioinformatica e successivamente legata all'intelligenza artificiale.

Si passerà per una trattazione dei metodi classici per la predizione della struttura di proteine che condurrà infine alla rivoluzione di AlphaFold.

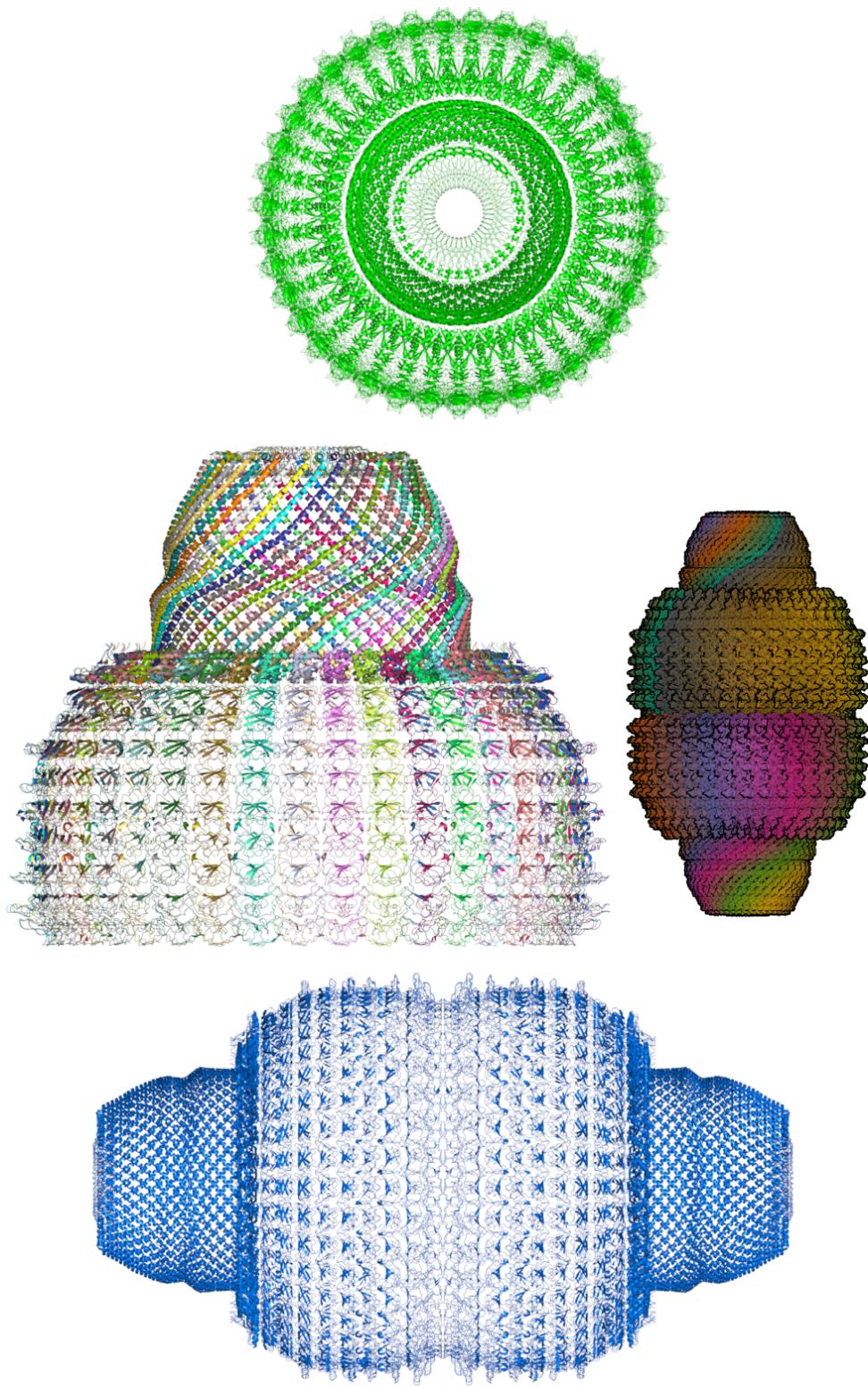


Figura 1: Rappresentazione della struttura del "vault" (ribonucleoproteina) del fegato di ratto con una risoluzione di 3.5Å, vista da varie angolazioni. La struttura di un vault è composta da 78 copie della Major Vault Protein assemblate. I "vault" sono proteine citoplasmatiche che devono il loro nome alla loro somiglianza con le "volte" di una cattedrale. Fonte[1]

«Seduto in riva all’oceano [...] ebbi la consapevolezza che tutto intorno a me prendeva parte a una gigantesca danza; [...] le mie esperienze [in fisica delle alte energie] presero vita: «vidi» scendere dallo spazio esterno cascate di energia, nelle quali si creavano e distruggevano particelle con ritmi pulsanti; «vidi» gli atomi degli elementi e quelli del mio corpo partecipare a questa danza cosmica di energia»¹

«Chiedersi il perché del big-bang, o chiedersi perché la banana sia gialla possono sembrare cose infinitamente diverse in importanza e filosofia, ma in fondo non è vero. Se si studia veramente fino in fondo perché la banana è gialla, si arriva probabilmente al perché dell’origine delle cose. [...] tutti i perché hanno un che di profondo, il punto essenziale è mettere onestamente a lavoro la mente e la fantasia»²

«[...] in Lui, le doti della mente e del cuore armonizzavano singolarmente. [...] è rimasto vivo [...] ad insegnarci ancora come sulla vetta del sapere debba ardere il puro fuoco dell’entusiasmo»

«Confessava della sua scienza l’impareggiabile funzione morale: destare nei giovani l’amore alle cose naturali, rivelarne la bellezza, discoprirne l’ordine e le leggi voleva dire per Lui potenziare ed accrescere in loro l’ordine morale interiore»³

¹F. Capra, *Il Tao della fisica*, 1975

²Pier Luigi Luisi, comunicazione personale, Luglio 2021

³G. Lambertini, “In onore di Angelo Ruffini,” in 31 mar. 1930

Indice

| | |
|--|-----------|
| Prefazione | 7 |
| 1 Introduzione | 8 |
| 2 Background | 9 |
| 2.1 Background biologico | 9 |
| 2.1.1 Organizzazione della vita: dagli atomi alle cellule | 9 |
| 2.1.2 Concetti fondamentali in biologia | 12 |
| 2.1.3 Dogma centrale della biologia | 14 |
| 2.1.4 Dai geni alle proteine | 17 |
| 2.1.5 Proteine: le macromolecole più importanti della vita | 20 |
| 2.2 Background informatico | 24 |
| 2.2.1 Bioinformatica | 24 |
| 2.2.2 Soft computing | 25 |
| 2.2.3 Intelligenza Artificiale | 26 |
| 2.2.4 Machine Learning | 27 |
| 2.2.5 Deep Learning | 28 |
| 3 Protein Folding | 34 |
| 3.1 Postulato di Anfinsen | 35 |
| 3.1.1 Denaturazione | 37 |
| 3.2 Struttura delle proteine | 38 |
| 3.2.1 Legami e interazioni molecolari | 39 |
| 3.2.2 Livelli strutturali | 41 |
| 3.2.3 Evoluzione e classificazione | 48 |
| 3.3 Dinamica del ripiegamento | 50 |
| 3.3.1 Geometria ed energetica del ripiegamento | 50 |
| 3.3.2 Ripiegamento assistito | 54 |
| 3.3.3 Misfolding, prioni e malattie | 56 |

| | | |
|----------|--|------------|
| 3.3.4 | Controllo qualità e apoptosis | 58 |
| 3.4 | Studio sperimentale delle proteine | 60 |
| 3.5 | Sfide al dogma di Anfinsen: IDP e fold switching | 67 |
| 3.5.1 | Considerazioni epistemologiche | 69 |
| 3.6 | Il problema del Protein Folding | 72 |
| 4 | Predizione della struttura di proteine | 74 |
| 4.1 | Metodi e strumenti informatici | 74 |
| 4.1.1 | Workflow e classificazione dei metodi per il PSP | 75 |
| 4.1.2 | Soft computing e deep learning | 79 |
| 4.1.3 | Output e misure di valutazione | 81 |
| 4.1.4 | Database e formati | 84 |
| 4.1.5 | Rappresentazione grafica | 88 |
| 4.1.6 | CASP ed excursus storico | 90 |
| 4.2 | Annotazioni 1D sulla struttura | 93 |
| 4.2.1 | Allineamento di sequenze | 94 |
| 4.3 | Annotazioni 2D sulla struttura | 95 |
| 4.3.1 | <i>correlated mutation</i> | 97 |
| 4.3.2 | <i>contact prediction ML-based</i> | 98 |
| 4.4 | Predizione della struttura 3D | 100 |
| 4.4.1 | <i>homology modeling</i> | 100 |
| 4.4.2 | <i>fold recognition</i> | 103 |
| 4.4.3 | <i>ab initio</i> | 105 |
| 4.4.4 | <i>fragment-based</i> | 112 |
| 4.4.5 | <i>loop modeling</i> | 114 |
| 4.4.6 | Case Study: <i>TASSER</i> | 117 |
| 5 | La rivoluzione di AlphaFold | 121 |
| 5.1 | Architettura | 126 |
| 5.1.1 | Evoformer | 127 |
| 5.1.2 | Structure Module | 131 |
| 5.1.3 | Altri dettagli | 134 |
| 5.1.4 | Analisi via ablazione | 136 |
| 5.1.5 | Differenze con AF1 | 138 |
| 5.2 | DeepMind | 139 |
| 6 | Conclusione | 141 |
| 6.1 | Sfide aperte | 141 |

| | |
|-----------------------------------|-----|
| 6.2 Etica della ricerca | 143 |
|-----------------------------------|-----|

| | |
|---------------------|------------|
| Bibliografia | 148 |
|---------------------|------------|

Prefazione

Questa tesi è il frutto di un lavoro di ricerca e studio. È stato per me⁴ un passo importante, nel quale ho messo alla prova le mie capacità di affrontare un argomento interdisciplinare e di interfacciarmi con il mondo della ricerca scientifica ricercando libri e articoli. Allo stesso tempo ho potuto dare spazio alla mia creatività, sia nello studio che durante la stesura di questo elaborato.

Non è stato immediato affrontare un tema con radici lontane dall'informatica come quello del *protein folding*. Il mio obiettivo è stato quello di acquisire un'ampia comprensione del problema e non solo degli aspetti informatici legati alla predizione della struttura di proteine. Ciò che mi affascina è quello che si può trovare al confine fra i mondi della biologia e dell'informatica, al confine fra i due linguaggi e fra i modi di ragionare. Non avevo precedenti esposizioni alla biologia prima di seguire il corso *elementi di biologia e neuroscienze* del prof. Mario Pirchio, correlatore del presente elaborato.

Oltre alle mie inclinazioni personali, trovo che considerando un problema biologico in termini esclusivamente matematici o informatici se ne riduca l'importanza. Ho provato a scendere nei dettagli su alcune questioni biologiche per portare alla luce (per quanto mi sia stato possibile e avendo cura di non andare fuori tema) alcuni "perché" dell'interesse nel ripiegamento delle proteine, come l'esistenza dei prioni e delle malattie neurodegenerative associate a problemi di ripiegamento delle proteine. Penso che comprendere la realtà biologica dell'argomento trattato e i "perché" possa instradare la ricerca interdisciplinare su vie più efficaci e collaborative.

⁴La prima persona è utilizzata solo nella prefazione per spiegare alcune scelte personali che perderebbero di "calore" se spiegate impersonalmente. Nel resto dell'elaborato verrà sempre utilizzata una forma impersonale.

Capitolo 1

Introduzione

Cos'è la vita? Da dove viene? - Fino al XVIII secolo per rispondere a tale quesito si faceva riferimento alla fede nel vitalismo: l'esistenza di una forza vitale non subordinata alle leggi della chimica e della fisica. Importanti svolte furono gli esperimenti, prima di Redi poi di Spallanzani, per dimostrare l'infondatezza della teoria della *generazione spontanea*, secondo la quale la vita poteva generarsi da materia non vivente. Un'importante passo in avanti, in concomitanza con l'affermarsi della *teoria cellulare*, fu il lavoro di Pasteur che stabilì un collegamento fra processi vitali e reazioni chimiche: per la conversione di zucchero in alcool (fermentazione) era necessaria la presenza di microorganismi.

Successivamente vi sono i lavori di Berthelot e Buchner (premio Nobel per la Chimica 1907), il quale dimostrò che era possibile ottenere la fermentazione in assenza di microorganismi, usando solamente sostanze estratte da essi. Queste sostanze furono chiamate *enzimi* (dal ted. Enzym, letteralmente «dentro il lievito»^[4]). Non si conosceva la loro natura chimica, si scoprì successivamente che tutti gli enzimi sono *proteine* (dal greco «primario», «che occupa la prima posizione»^[5]). Queste proteine agivano da catalizzatori: acceleravano le reazioni chimiche all'interno delle cellule senza cambiare la loro natura, quindi senza consumarsi e senza entrare nei prodotti finali della reazione.

La scoperta degli enzimi portò ad un cambio di paradigma nel pensiero scientifico riguardo le origini della vita: veniva ora considerata come la conseguenza di numerosi processi chimici resi possibili dalle proteine^[6]. I fondamenti del pensiero biologico si spostarono dal vitalismo al meccanicismo secondo il quale tutti i fenomeni naturali, vita compresa, sono governati dalle stesse leggi, sia per sostanze organiche che inorganiche.

L'inconcorazione delle proteine a *macromolecole più importanti della vita* si può legare ad un'altra svolta nel pensiero scientifico avvenuta nella seconda metà del XX secolo: la rivoluzione genetica. Le proteine sono i prodotti finali dei geni e sono anche coinvolte nell'espressione dell'informazione genetica. È sullo sfondo di questa rivoluzione che l'informatica si è inserita all'interno del mondo della biologia.

Capitolo 2

Background

Quello che segue è un breve viaggio tra argomenti di carattere biologico, fondamentali per comprendere il perché si vogliono studiare le proteine, e i principali settori dell'informatica toccati nel presente lavoro di tesi.

2.1 Background biologico

2.1.1 Organizzazione della vita: dagli atomi alle cellule

Nonostante le grandi differenze in dimensione, dieta, riproduzione, morfologia, comportamento, vi è un tratto comune a tutti gli organismi viventi: sono composti di cellule. Tutte le cellule sono caratterizzate da una stupefacente somiglianza chimica poiché utilizzano molecole simili e hanno ereditato tutte le stesse *intuizioni*¹ genetiche. Si pensa quindi vi sia un antenato comune a tutti i viventi: una cellula vissuta circa 3,5 miliardi di anni fa che conteneva un prototipo del macchinario universale della vita sulla Terra oggi^[7].

Prima di parlare di cellule è opportuno richiamare l'attenzione sulle strutture biologiche. L'organizzazione biologica si basa su una gerarchia di livelli strutturali², ognuno dei quali poggia su un gradino sottostante:



¹Il termine *intuizione* è qui usato creativamente per indicare le soluzioni genetiche sviluppatesi e sopravvissute ad oggi. Non si intende attribuire intelligenza, pensiero o volontà all'evoluzione.

²Questa sezione di background biologico si basa in larga parte su N. A. Campbell, J. B. Reece, L. A. Urry et al., *Biologia E Genetica*. Pearson, 2012.

Tutta la materia è costituita da 94 elementi chimici in natura^[9] (tralasciando quindi gli altri 24 elementi sintetici). La materia vivente è composta per il 96% da atomi di C, O, N, H (carbonio, ossigeno, azoto, idrogeno). Un atomo ha un nucleo composto da neutroni e protoni circondato da una nube di elettroni in rapido movimento. Il Dalton (Da) è l'unità della massa atomica, corrisponde al peso di un protone o neutrone: $1\text{Da} = 1.66 \times 10^{-24}\text{g}$. Un elettrone pesa 0.0005Da . $1\text{\AA} = 0.1\text{nm}$ e nonostante non sia un'unità di lunghezza appartenente al SI è molto utilizzata per le dimensioni di molecole, legami e atomi, il cui raggio varia da 0.25 e 3\AA . Gli elettroni più esterni sono chiamati *elettroni di valenza* e determinano il comportamento chimico di un atomo.

Lo scheletro delle molecole organiche è formato da catene carboniose, lunghe catene di atomi di carbonio legati fra loro da legami covalenti (il tipo di legame chimico più forte). Salendo di complessità si arriva alle macromolecole biologiche, fondamentali per le cellule: carboidrati, lipidi, acidi nucleici e proteine. I carboidrati sono combustibili cellulari e materiale da costruzione, i lipidi sono sia depositi di energia che i principali costituenti delle membrane cellulari, gli acidi nucleici permettono di codificare l'informazione genica e le proteine sono alla base delle funzioni vitali.

La cellula è la più piccola unità in grado di vivere. Per *vivente* si intende un essere dotato di: organizzazione interna, metabolismo, omeostasi, interazione con l'ambiente, adattamento, crescita e riproduzione.

Le cellule hanno dimensioni che variano dai *micrometri* (μm) ai *centimetri* delle uova di rana, gallina o struzzo ai *metri* di neuroni con lunghi assoni. In figura 2.1 si possono notare le diverse dimensioni di alcune cellule.

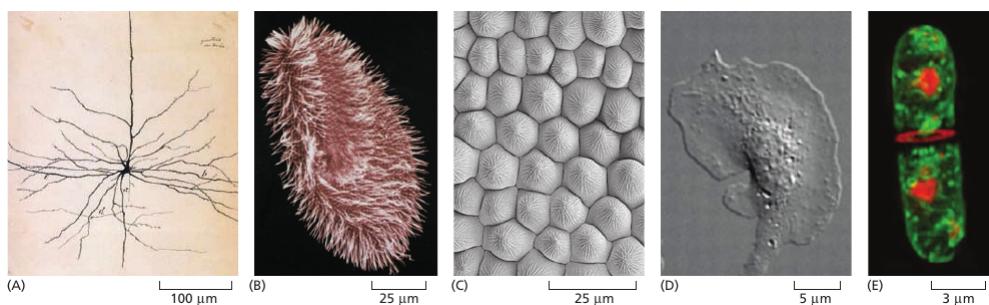


Figura 2.1: (A) disegno di un neurone. (B) Paramecium. (C) superficie di un petalo di fiore di bocca di leone. (D) Macrofago. (E) Un lievito di fissione viene catturato nell'atto di divisione cellulare. Fonte: [7]

È possibile suddividere gli esseri viventi in due domini³: *procarioti* ed *eucarioti*. Il primo include i due regni Bacteria e Archaea. Sono caratterizzati da cellule piccole, circa

³Tale classificazione è soggetta a frequenti cambiamenti.

$1\mu m$. Il secondo dominio include cinque regni: animali, piante, funghi, protisti e cromisti. Gli organismi eucarioti dispongono di cellule più grandi (circa $10-100\mu m$) dotate di compartimenti interni che separano le funzioni cellulari. La strutture tipiche di una cellula animale e di un neurone sono mostrate nelle seguenti figure:

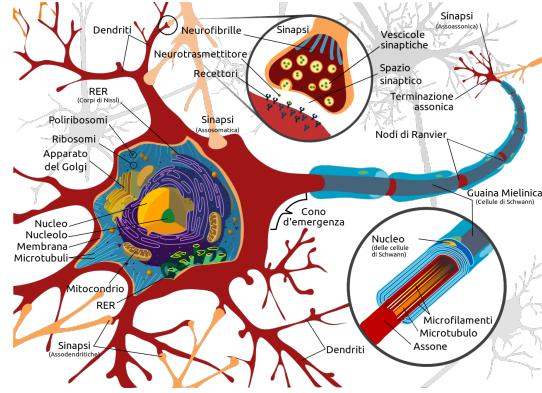
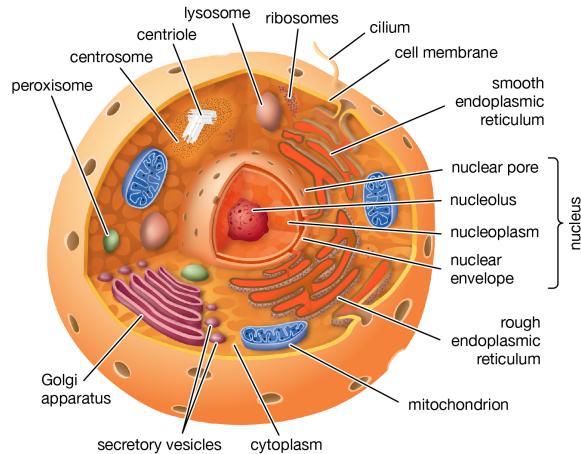


Figura 2.3: Neurone. Fonte [11]

Figura 2.2: Cellula animale. Fonte: [10]

Una cellula eucariote animale è formata innanzitutto dalla membrana cellulare, un involucro costituito da un doppio strato fosfolipidico che permette alla cellula di avere il suo "spazio vitale" in quanto la separa dall'ambiente (spesso acquoso) circostante. È attraversata da piccoli pori che le permettono lo scambio di sostanze con l'esterno. Tutto ciò che si trova all'interno della cellula è immerso nel citoplasma, gel acquoso contenente grandi e piccole molecole. Il citosol è la parte del citoplasma non contenuta all'interno delle membrane intracellulari. Il volume totale delle cellule è composto da acqua per il 70% circa. Vi è poi il citoscheletro che dà forma strutturale e permette in alcuni casi movimenti direzionati.

Il primo organello di grande importanza è il reticolo endoplasmatico, formato da tubuli e cisterne e in comunicazione con l'involucro nucleare. È rugoso quando sono presenti ribosomi (sintetizzatori di proteine). È il componente della fabbrica cellulare che si occupa di attività e sintesi di molecole fondamentali per la sopravvivenza della cellula (sintesi di steroidi, metabolismo del glucosio, eliminazione di sostanze nocive). L'apparato del Golgi produce vescicole che si fondono poi con la membrana cellulare: è una centrale di smistamento per confezionare sostanze da esportare. I lisosomi sono il centro di degradazione e riciclo della cellula. Il mitocondrio è la centrale energetica della cellula, dove avviene la respirazione cellulare: utilizza ossigeno per bruciare molecole organiche degradate nel citoplasma come *piruvato* e *acetil-coenzima A* al fine di produrre energia che verrà immagazzinata sotto forma di ATP.

Infine è presente il nucleo, custode del DNA. È formato dall'invólucro nucleare, cromatina e nucleolo. Il DNA nel nucleo è associato a delle proteine con cui forma un materiale fibroso chiamato cromatina, mostrandosi "sfilacciato" in modo da poter essere letto. Quando la cellula si riproduce la cromatina si condensa in strutture compatte e singole: i cromosomi. Il nucleolo non è provvisto di membrana e serve per la sintesi di RNA ribosomiale, cioè l'RNA che uscendo dai pori dell'invólucro nucleare andrà nel citoplasma a formare i ribosomi. L'invólucro nucleare possiede dei pori nucleari attraverso i quali possono transitare RNA e proteine, ma non DNA.

Il ciclo di vita delle cellule si basa su 4 fasi: crescita iniziale, sintesi del DNA, ulteriore crescita e mitosi (divisione cellulare). Le cellule dei mammiferi possono impiegare anche dei giorni per completare un ciclo di mitosi, mentre i lieviti solamente 90 minuti. Per questa ragione il lievito da fornaio (*Saccharomyces cerevisiae*) è molto utilizzato in citologia e genetica: è uno degli organismi eucarioti modello^[7] ed il suo genoma è stato il primo ad essere sequenziato completamente tra gli eucarioti^[12].

Le cellule hanno una durata di vita molto variabile, ad esempio alcuni organismi unicellulari come le spore possono vivere anche decenni, così come i nostri neuroni, mentre i globuli bianchi non sopravvivono oltre pochi giorni.

Gli strumenti utilizzati per indagare nel mondo microscopico riescono a mostrare dettagli che vanno dal limite di 200nm del microscopio ottico (limite imposto dalla natura ondulatoria della luce) alla precisione di 1nm del microscopio a trasmissione elettronica (che usa fasci di elettroni invece di fasci di luce ma di contro necessita di campioni molto fini):

2.1.2 Concetti fondamentali in biologia

- *Proprietà emergenti*

Ad ogni livello di indagine, ovvero passando da un livello della gerarchia strutturale al superiore, si palesano nuove proprietà non riconducibili ai livelli più semplici: le proprietà emergenti. Una singola molecola d'acqua non è né solida né liquida.

- *Teoria cellulare*

Le cellule rappresentano le unità strutturali e funzionali degli organismi.

- *Geni*

Il perpetuarsi della vita è possibile grazie alla trasmissione dei geni.

- *Forma e funzione*

Forma e funzione sono correlate a tutti i livelli biologici. Se le ali degli uccelli non fossero così come sono essi non potrebbero volare, se i mitocondri non avessero numero creste produrrebbero minori quantità di ATP, se i neuroni non avessero

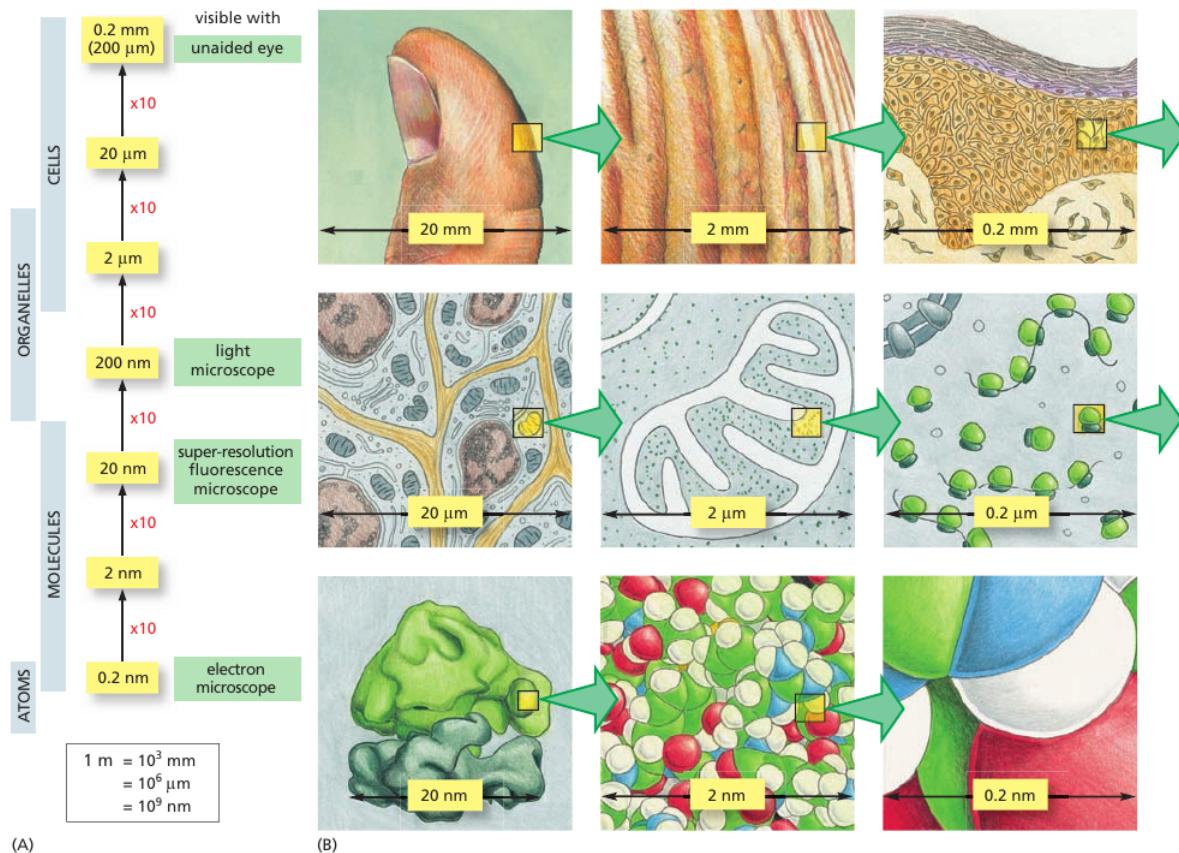


Figura 2.4: (A) Il grafico elenca le dimensioni dei livelli strutturali biologici, le unità di misura relative e gli strumenti necessari per visualizzarli. (B) Uno stesso dettaglio a varie scale di grandezza: pollice, pelle, cellule, mitocondrio, ribosomi, insieme di atomi che formano parte di una proteina. I dettagli molecolari sono oltre la potenza del microscopio elettronico. Fonte: [7]

lunghi assoni non riuscirebbero a comunicare efficientemente e se i *paramecium* non avessero le loro ciglia non potrebbero muoversi come sommersibili (vedi figura 2.1B).

- *Evoluzione*

L'evoluzione rappresenta il tema centrale ed unificante della biologia, come si è già accennato sopra. Gli organismi sono sistemi aperti che interagiscono continuamente con l'ambiente, dotati di variabilità individuale e finalizzati alla competizione per la sopravvivenza.

- *Diversità e unità*

Vi sono da 5 a 30 milioni di specie differenti eppure scendendo sempre di più nella struttura degli organismi si osserva una similitudine quasi sconcertante. Un esempio che ci riguarda è la somiglianza fra le ciglia di *paramecium* e le ciglia di una cellula epiteliale delle vie aeree degli esseri umani: presentano la stessa sezione trasversale. Il codice genetico (le triplettre) sono universali, gli amminoacidi si codificano nello stesso modo per tutti gli organismi. Diversità e unità della vita sulla Terra sono due

facce della stessa medaglia. Il sequenziamento dei genomi e il loro confronto, basato su approcci informatici, ha rivelato una conservazione evoluzionistica, un'eredità comune: è possibile infatti scambiare geni omologhi codificanti proteine del ciclo di divisione cellulare fra uomini e lievito^[7]: una cellula di lievito ha quindi tutto il macchinario molecolare necessario per leggere, interpretare e utilizzare il nostro codice genetico per la produzione di proteine umane funzionanti. Sono osservazioni simili che hanno guidato la direzione di alcune tecniche informatiche, anche per la predizione della struttura di proteine (come si vedrà successivamente).

2.1.3 Dogma centrale della biologia

Nel 1958 il premio Nobel Francis Crick introdusse il *dogma* centrale della biologia, che allo stato attuale si può considerare come l'insieme dei principali meccanismi alla base dell'espressione genica.

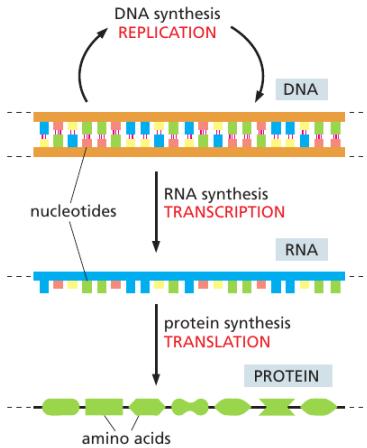


Figura 2.5: Dogma centrale in biologia. Fonte [7]

Il dogma descrive il flusso di informazione genetica: essa è conservata negli acidi nucleici DNA (RNA per alcuni virus) che possono essere duplicati, il DNA viene poi trascritto sotto forma di RNA e se codificante questo è poi tradotto in proteine, concepite come la forma operativa e terminale delle informazioni contenute nel genoma^[13].

Per avere una miglior panoramica del funzionamento di questo principio è importante approfondire la struttura del DNA (*acido desossiribonucleico*). Il DNA è una molecola composta da due catene complementari che si avvolgono l'una intorno all'altra tramite legami idrogeno formando una doppia elica. Le catene sono chiamate filamenti e sono antiparalleli. Dal punto di vista chimico è un polimero di nucleotidi, dove ogni nucleotide è composto da una base azotata, uno zucchero pentoso (*ribosio* nell'RNA e *desossiribosio* nel DNA) e un gruppo fosfato (vedi figura 2.7). Per ogni giro dell'elica vi sono 10 coppie

di basi. La struttura a doppia elica consente un'agevole meccanismo di replicazione del DNA, coadiuvato dagli enzimi DNA polimerasi, primasi e DNA ligasi. Gli accoppiamenti seguono delle regole precise: GC, AT/AU, da una parte deve esserci una pirimidina (C, T) e dall'altra una purina (A,G):

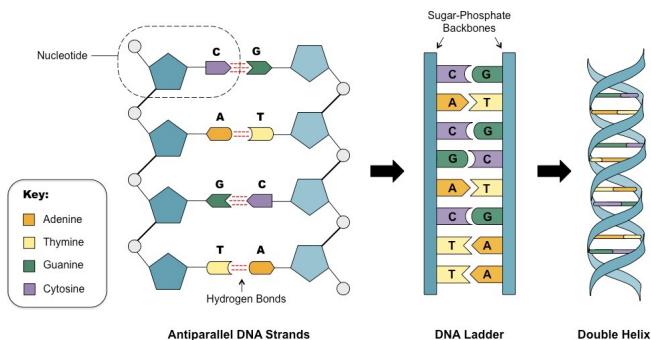


Figura 2.6: struttura del DNA. Fonte: [14]

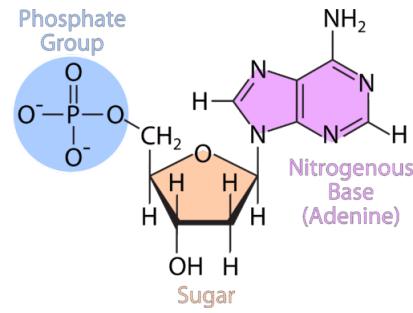


Figura 2.7: Componenti di un nucleotide con Adenina per base azotata. Fonte [15]

Il *gene* è l'unità elementare dell'informazione genetica e corrisponde al segmento di DNA (raramente di RNA) in grado di codificare la sequenza primaria di una proteina.

Geni che controllano un medesimo carattere (per esempio, il colore dei capelli) sono disposti sui cromosomi in *loci* (plurale di locus genico, posizione) identici. I cromosomi omologhi sono cromosomi morfologicamente identici che presentano, in loci corrispondenti, gli stessi geni con le stesse informazioni. Ogni gene è presente in doppia coppia nelle cellule diploidi e può risultare pertanto omozigote od eterozigote. Il termine omozigote si riferisce a un gene in cui l'informazione riportata dal gene materno è identica a quella paterna, mentre in geni eterozigoti il contributo del gene materno e paterno è diverso: in questo caso la determinazione fenotipica è legata ai concetti di dominanza e recessività genetica. Lo stesso gene nella stessa specie può esistere in varie forme, con leggere differenze nella sequenza nucleotidica: si sta parlando degli *alleli* del gene. Più precisamente l'insieme delle possibili versioni (1, 2 o più) dello stesso gene corrisponde ai suoi alleli. Un allele è quindi una delle versioni dello stesso gene nello stesso locus su un cromosoma.

Il *genoma* indica il patrimonio aploide del DNA di una cellula (compreso il DNA di altri organelli come mitocondri o cloroplasti). L'insieme di tutti i geni di un individuo determinano il suo *genotipo*; relativamente a un gene il genotipo può anche indicare il corredo di alleli che l'organismo si trova a possedere (nell'uomo al massimo 2). Il *fenotipo* indica invece l'insieme delle caratteristiche morfologiche e funzionali di un organismo, quali risultano dall'espressione del suo genotipo e dalle influenze ambientali. In un organismo, nonostante tutte le cellule condividano gli stessi geni, cellule diverse possono esprimere geni differenti (*espressione genica*).

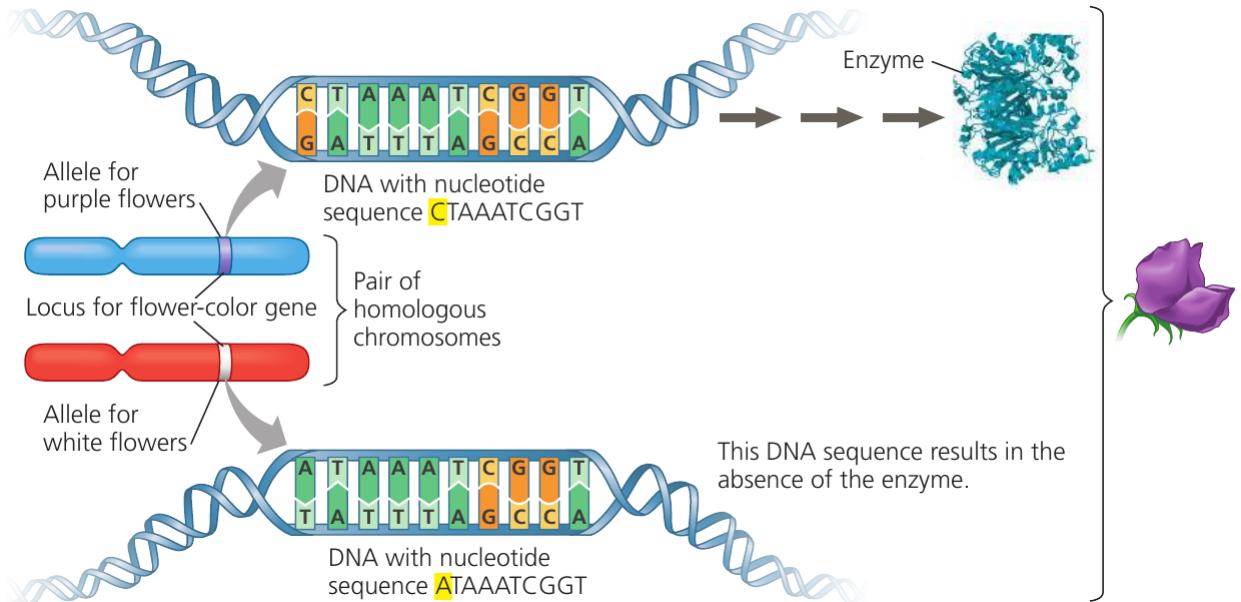


Figura 2.8: Vengono mostrati due cromosomi omologhi di una pianta di pisello con la sequenze degli alleli relativi al colore del fiore. Il cromosoma blu si assume essere ereditato dal padre e ha un allele per i fiori viola (codifica cioè un enzima che indirettamente controlla la sintesi del pigmento viola). Il cromosoma materno (rosso) ha un allele per i fiori bianchi che non codifica per nessuna proteina funzionale. Un solo allele per i fiori viola risulta essere sufficiente per sintetizzare abbastanza pigmento da far apparire il fiore viola. Fonte[16]

L'RNA (*acido ribonucleico*) esiste in varie forme. Le differenze con il DNA sono mostrate nella figura 2.9, si può notare che vi è un singolo filamento e che la base azotata timina è assente e al suo posto si trova la base uracile (U). Essendo ad un unico filamento può formare legami a idrogeno con sé stessa e assumere forme tridimensionali vantaggiose. Esistono vari tipi di RNA:

- mRNA, messaggero, contiene l'informazione per la sintesi delle proteine
- tRNA, di trasporto, necessario per la traduzione nei ribosomi
- rRNA, ribosomiale, entra nella struttura dei ribosomi
- snRNA, hnRNA

L'RNA catalitico o *ribozima*, enzima ad RNA, è una molecola di RNA in grado di catalizzare una reazione chimica similmente agli enzimi.

Il DNA dell'uomo contiene 3^9 coppie di nucleotidi (3.3Gbp , *gigabasepairs*), ha circa 21000 geni codificanti e pesa 3.56pg ⁴: se il genoma umano venisse esteso in lunghezza

⁴In termini di massa è possibile convertire il numero di paia di basi azotate in *picogrammi*, $1\text{pg}=0.978\text{Gbp}$, poiché una coppia di basi azotate pesa 650Da . Il peso del genoma umano è calcolabile come segue: $3.3 \times 10^9 \times 650 \times 1.66 \times 10^{-24} = 3.56 \times 10^{-12}\text{g}$.

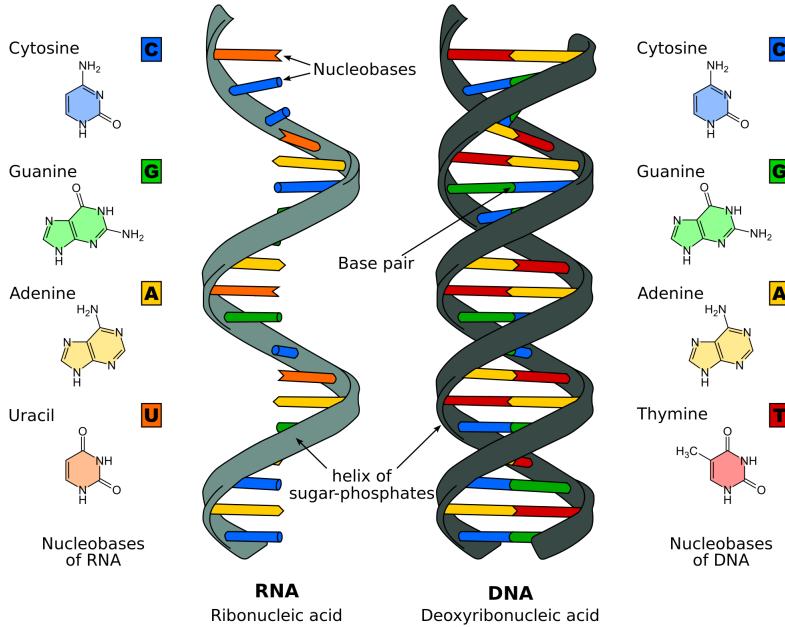


Figura 2.9: Differenze fra RNA e DNA Fonte: [17]

sarebbe lungo 2,2 metri dato che ogni nucleotide è lungo 0.34nm . Il batterio più semplice (*Nasutia deltocephalinicola*) ha un genoma di 112Kb^[18] (circa $76\mu\text{m}$ in lunghezza) contenente 137 geni codificanti, mentre il genoma maggiore ad oggi riportato è quello della pianta *Paris Japonica* con 148.8Gbp^[19], 50 volte quello dell'uomo (circa 100m in lunghezza), tanto per avere una visione quantitativa della diversità genetica tra gli organismi.

| Species | T2 phage | <i>Escherichia coli</i> | <i>Drosophila melanogaster</i> | <i>Homo sapiens</i> | <i>Paris japonica</i> |
|-------------|------------|-------------------------|--------------------------------|---------------------|-----------------------|
| Genome Size | 170,000 bp | 4.6 million bp | 130 million bp | 3.2 billion bp | 150 billion bp |
| Common Name | Virus | Bacteria | Fruit fly | Human | Canopy Plant |



Figura 2.10: Dimensioni del genoma di diverse specie a confronto. Fonte: [20] Figura 2.11: Fiore di *Paris Japonica*. Fonte [21]

2.1.4 Dai geni alle proteine

Il codice genetico lavora a sequenze di codici di 3 lettere (es. "GAA" = Glutammato), questo perché si hanno a disposizione 4 lettere (le basi azotate) e si devono codificare i 20 diversi amminoacidi. Con 2 lettere avrei 4^2 possibilità che non sono sufficienti a descrivere 20 informazioni diverse, si utilizzano pertanto 3 lettere anche se ciò causa ridondanza nei

codici. Un amminoacido è quindi codificato da una tripletta: si parla di *codice a triplette*.

Il primo passo consiste nella *trascrizione*. Un filamento di DNA fa da stampo per la creazione di mRNA, il tutto esclusivamente tramite *complementarità di forma*. Il DNA non viene aperto come una zip ma l'apertura, la trascrizione (compiuta dall'RNA polimerasi, soggetta a errori anche frequenti) e la chiusura della doppia elica avvengono di pari passo. Vi è un terminatore nel DNA per indicare la fine del gene.

Le triplette nucleotidiche dell'mRNA sono dette *codoni* e codificano un amminoacido. I codoni devono essere letti in direzione 5' -> 3'. La molecola di mRNA lascia il nucleo attraverso i pori nucleari. È importante osservare che non tutti i geni codificano proteine (lo stadio di trascrizione potrebbe risultare quello finale) e che il codice genetico è *universale*, è condiviso dai batteri, piante, animali: per tutti la prolina si codifica in "CCG".

Negli eucarioti è presente un passaggio intermedio: la *maturazione*, o fase di processamento. È composto da due sottofasi:

- *incapsulamento*, viene aggiunta una coda e un cappuccio alle due estremità al fine di proteggere l'mRNA dalla degradazione e per segnalare l'inizio ai ribosomi.
- *splicing*, il DNA possiede lunghe sequenze nucleotidiche non codificanti, gli *introni*. In questa fase vengono rimossi e gli *esoni* (sequenze codificanti) vengono riunite insieme. È in questo modo che è possibile dare origini a sequenze primarie (delle proteine) diverse a partire da un unico gene.

L'ultimo passaggio è la *traduzione*, attraverso la quale la cellula interpreta il messaggio genetico e polimerizza gli amminoacidi per costruire la relativa proteina. Il processo di traduzione è la transizione da un linguaggio a 4 lettere (basi azotate) ad un linguaggio a 20 lettere (amminoacidi). La traduzione viene realizzata dal tRNA, una sorta di adattatore da linguaggio *genetico* a linguaggio *amminoacidico*. Il tRNA è un acido nucleico a forma di L composto da circa 80 basi, a un'estremità vi è l'anticodone (interfaccia con il linguaggio genetico) e all'altra vi è il sito di legame con un singolo amminoacido. Il tRNA trasporta ai ribosomi uno specifico amminoacido contenuto nel citoplasma. Esiste di conseguenza uno specifico tipo di tRNA per ogni codone.

È interessante notare che il tRNA, proprio come le proteine, è caratterizzato dall'avere più strutture: quella primaria, costituita dalla sua sequenza nucleotidica, quella secondaria data dalla sua struttura a quadrifoglio e quella terziaria dovuta alla struttura tridimensionale a L. La differenza fra la struttura del tRNA e delle proteine sono gli elementi unitari: nel tRNA si tratta di nucleotidi mentre nelle proteine di amminoacidi.

La traduzione comincia con il primo codone (AUG, che oltre a segnalare l'inizio codifica anche la metionina, vedi figura 2.15) al quale si incassa nel ribosoma un tRNA avente il corrispondente amminoacido legato. Si formano legami idrogeno fra i nucleotidi. Arriva un

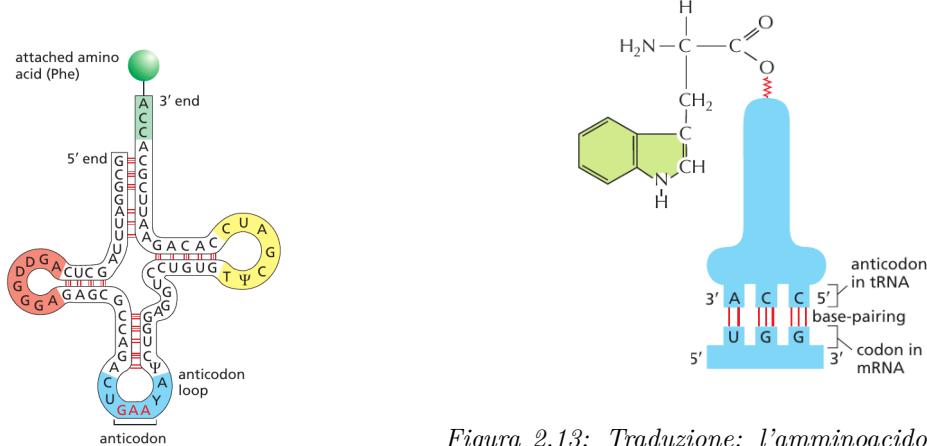


Figura 2.12: tRNA. Fonte [7]

Figura 2.13: Traduzione: l'amminoacido triptofano (*Trp*) è codificato dal codone UGG nell'mRNA e si lega al tRNA tramite un legame energetico forte. Fonte: [7]

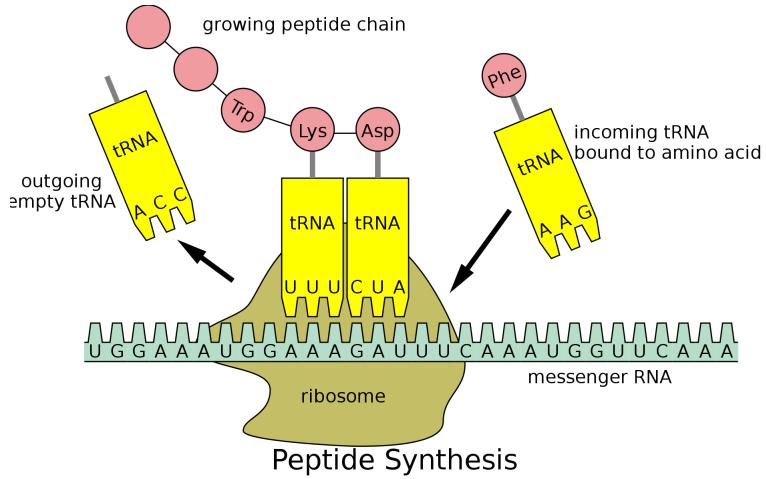


Figura 2.14: Traduzione, sintesi peptidica. Fonte: [22]

secondo tRNA combaciante con il successivo codone. I due amminoacidi si trovano vicini e formano un legame peptidico. L'mRNA scorre così che si crei posto per nuovi tRNA, nel frattempo gli amminoacidi si legano fra loro e cominciano a formare la proteina. Il ripiegamento della proteina comincia già durante la sua biosintesi. Il processo termina quando si arriva ad un codone di stop (es. UAA). Per velocizzare il processo di sintesi ribosomiale questo viene parallelizzato: tanti *poliribosomi* sono associati allo stesso mRNA attuando una rapida sintesi di copie multiple di un polipeptide a partire da un unico mRNA.

| | | | | | | | | | |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|------|
| | AGA | | UUA | | AGC | | GUA | | |
| | AGG | | UUG | | AGU | | GUC | | UAA |
| codons | GCA | CGA | GGA | AUA | CUA | CCA | UCA | ACA | |
| | GCC | CGC | GGC | CUC | | CCC | UCC | ACC | |
| | GCG | CGG | GAC | AAC | GGG | CAC | AUC | CUG | AAA |
| | GCU | CGU | GAU | AAU | GGU | CAU | AUU | CUU | AAG |
| | | | | | | | | | AUG |
| | | | | | | | | | UUU |
| amino acids | Ala | Arg | Asp | Asn | Cys | Glu | Gly | His | Ile |
| | A | R | D | N | C | E | Q | G | H |
| | | | | | | | | | I |
| | | | | | | | | | L |
| | | | | | | | | | K |
| | | | | | | | | | M |
| | | | | | | | | | F |
| | | | | | | | | | P |
| | | | | | | | | | S |
| | | | | | | | | | T |
| | | | | | | | | | W |
| | | | | | | | | | Y |
| | | | | | | | | | V |
| | | | | | | | | | stop |

Figura 2.15: Codici a tripletta degli amminoacidi. Fonte: [7]

2.1.5 Proteine: le macromolecole più importanti della vita

Le proteine sono formate dall'unione di strutture più semplici: gli amminoacidi. Un polimero amminoacidico composto da meno di 50 amminoacidi è chiamato *peptide*, se supera tale soglia *polipeptide*. Una proteina può essere quindi sia un semplice peptide⁵ che un singolo polipeptide o essere formata da più polipeptidi. La sequenza amminoacidica determina la struttura della proteina ed è proprio questo il collegamento fra il messaggio genetico nel DNA e la struttura tridimensionale che è associata alla sua funzione biologica.

Un amminoacido è una molecola organica formata da un atomo di carbonio centrale chiamato C_α circondato da 4 componenti (vedi fig. 2.17):

1. un atomo di idrogeno
2. un gruppo amminico ($\alpha - amino$), (-NH₂) in condizioni fisiologiche carico positivamente (-NH₃⁺)
3. un gruppo carbossilico ($\alpha - carboxyl$), (-COOH) carico negativamente (-COO⁻)
4. un gruppo R, gruppo laterale chiamato anche *residuo* che per sineddoche indica l'intero amminoacido una volta che questo si trova all'interno della catena proteica

Vi sono circa 20 amminoacidi proteinogenici diversi (come si può vedere nella figura 2.15 o 2.20). Il gruppo laterale non partecipa alla catena della *backbone* (spina dorsale) della proteina, resa stabile dai legami peptidici: rimane infatti libero di legarsi. È questo il "trucco" che consente alla proteina sia di ripiegarsi su sé stessa che di legarsi ad altre molecole. per tutti gli amminoacidi, tranne la glicina, l'atomo di carbonio che collega il gruppo laterale al resto della struttura, legandosi a C_α , è chiamato C_β . Gli amminoacidi possono essere polari, non polari, carichi (vedi figura 2.20) e causano differenti ripiegamenti della proteina. Di conseguenza ne influenzano la funzione, si pensi infatti al caso dell'anemia falciforme causata da 1 solo amminoacido di differenza: valina al posto del

⁵Esempi di "semplici" peptidi che svolgono funzioni biologiche sono i *neuropeptidi* che agiscono da neurotrasmettitore (ad es. endorfine) e *ormoni* quali l'insulina, il glucagone e l'ossitocina, composta da soli 9 amminoacidi e implicata nelle contrazioni uterine e nella stimolazione dei dotti lattiferi delle mammelle.

glutammato. La prima non è polare mentre il secondo è polare carico, ciò causa legami differenti, quindi ripiegamento differente e funzione biologica compromessa.

Gli amminoacidi esistono in 2 configurazioni: L e D. Essi sono infatti molecole *chirali*: le due configurazioni sono l'immagine speculare l'una dell'altra ma non sono sovrappponibili. Nella grande maggioranza degli organismi viventi le proteine sono composte solo da amminoacidi della serie L.

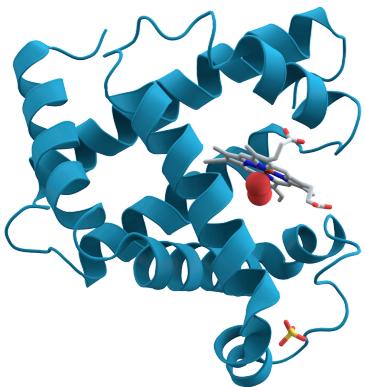


Figura 2.16: Rappresentazione a nastro della struttura tridimensionale della mioglobina. È presente un gruppo hemo al quale è legata una molecola di ossigeno (rossa). Fonte: [23]

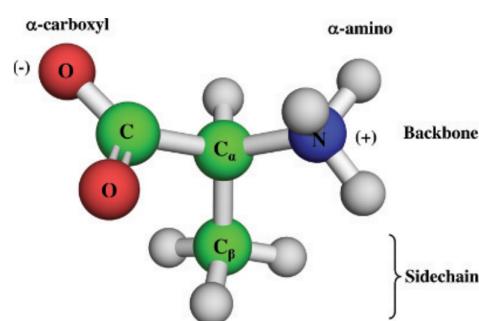


Figura 2.17: Struttura principale degli amminoacidi. Fonte [6]

Il legame peptidico è il legame che unisce tutti gli amminoacidi di una proteina: unisce il gruppo carbossilico di un amminoacido al gruppo amminico di un altro amminoacido. È un tipo di legame molto stabile, infatti l'emivita della backbone è di 400 anni a 25°C^[7]. Il legame peptidico comporta l'eliminazione della carica degli ex gruppi *amminico* e *carbossilico*.

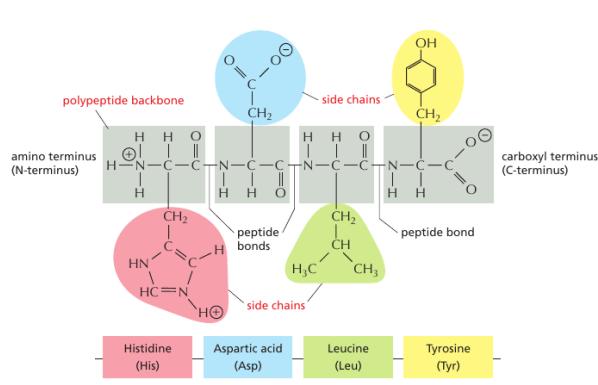


Figura 2.18: Backbone delle proteine. Fonte: [7]

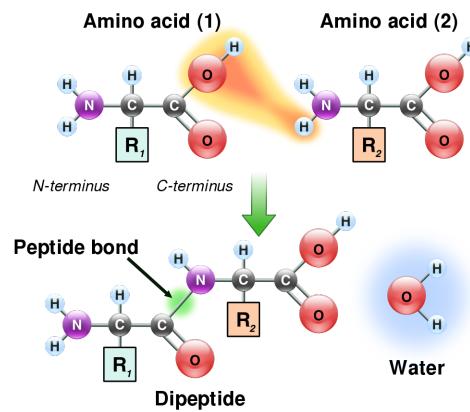


Figura 2.19: Legame peptidico. Fonte [24]

Gli unici due residui elettricamente carichi rimasti in una proteina sono quelli alle due estremità (C-terminus ed N-terminus, vedi fig. 2.18). È presente però un fenomeno che permette ai residui di interagire elettrostaticamente: la *risonanza elettronica*. Gli elettroni dei legami possono estendersi su più atomi e permettere al residuo di assumere diverse configurazioni elettroniche.

Nonostante gli amminoacidi siano solo 20, la varietà di proteine è elevatissima, in quanto gli amminoacidi si combinano tra loro in sequenze e quantità diverse. Dato un polipeptide di 100 amminoacidi si hanno 20^{100} possibili combinazioni.

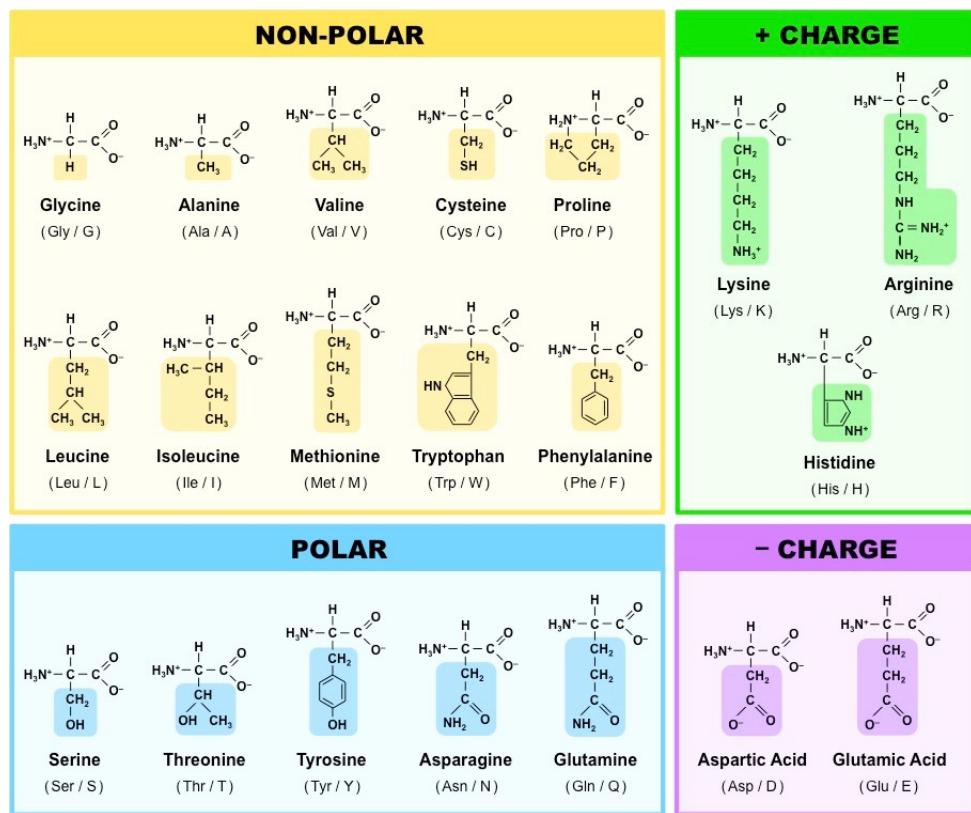


Figura 2.20: I 20 amminoacidi universali. Fonte: [25]

È possibile in realtà parlare anche di altri amminoacidi e di derivati. La *selenocisteina* è considerata il 21° amminoacido (così come la *pirrolisina* il 22°). È stata scoperta per la prima volta nel 1986 ed è codificato dal codone UGA, normalmente un codone di stop, che tuttavia in presenza di un particolare segmento di mRNA viene interpretato come elemento costitutivo. La sua struttura è identica a quella della cisteina con una sola differenza: un atomo di selenio al posto di quello di zolfo. Esistono poi una serie di derivati dagli amminoacidi. Si può dire ad esempio che la *tirosina* sia il precursore della *dopamina*, *melanina* e *adrenalina*, il *triptofano* di *serotonin* e *melatonina*. Tipicamente questi derivati sono modificati dopo la traduzione nei ribosomi: la proteina in formazione

viene modificata tramite legami covalenti da parte di enzimi e vengono a formarsi questi derivati.

Le proteine sono una classe di macromolecole con funzioni biologiche vitali, consentono infatti il funzionamento di ogni sistema vivente. Riusciamo a pensare, parlare, a digerire il cibo, a muoverci grazie alle proteine. Sono la base della vita cellulare e molecolare.

Un tipo fondamentale di proteine sono gli enzimi, come accennato inizialmente. Una loro funzione importante è correlata alla digestione negli animali. Enzimi come le *amilasi* e le *proteasi* sono in grado di ridurre le macromolecole (nella fattispecie amido e proteine) in unità semplici (maltosio e amminoacidi), assorbibili dall'intestino.

Oltre agli enzimi ci sono tante altre proteine importanti. Uno degli esempi più noti è l'emoglobina, proteina animale adibita a trasportare ossigeno dai polmoni ai tessuti così come a riportare CO₂ ai polmoni. Una molecola di emoglobina è composta da 4 polipeptidi e contiene 4 atomi di ferro che le consentono di legare reversibilmente 4 molecole di ossigeno.

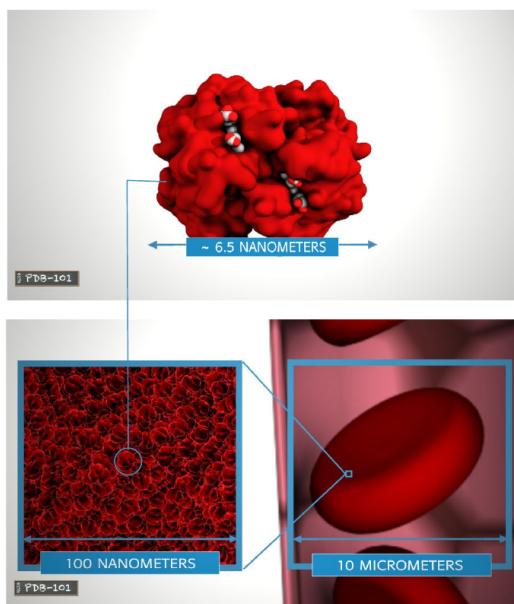


Figura 2.21: Emoglobina in diverse scale. Rapresentazione a superficie. Un globulo rosso contiene circa 280 milioni di molecole di emoglobina, per cui può portare più di 1 miliardo di molecole di ossigeno per volta. Fonte: [26]

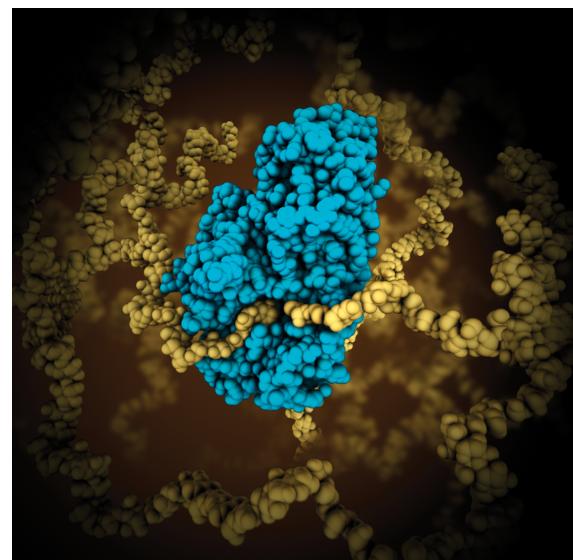


Figura 2.22: Enzima alpha Amilasi in turchese, rappresentazione di tipo space-filling. Si lega a catene di carboidrati (gialle) e le rompe in pezzi più piccoli di glucosio. Fonte [26]

Nelle cellule le proteine svolgono, fra le altre, funzioni di supporto strutturale (*collagene*), mobilità (*actina, miosina*), protezione (*anticorpi*), regolazione, ormoni (*insulina*), trasporto, catalisi, magazzino. Nel nostro corpo abbiamo un numero grandissimo di proteine: 10²⁷. Per usare una metafora di Ken Dill^[27] potremmo dire che se si potesse ingrandire

una proteina alla grandezza di un penny (diametro di 19mm) il numero di proteine che una persona ha nel corpo equivale al numero di penny che riempirebbero l’Oceano Pacifico.

Per queste e altre ragioni queste macromolecole sono il target di grandi attività di ricerca e di applicazione biotecnologiche: dal combattere malattie infettive^[28] al contrastare l’inquinamento ambientale^[29].

2.2 Background informatico

2.2.1 Bioinformatica

La *bioinformatica* ha giocato un ruolo fondamentale durante l’epidemia di COVID-19, in particolare nella realizzazione di vaccini grazie agli avanzamenti nelle tecnologie NGS (Next Generation Sequencing). La bioinformatica è una disciplina dedicata alla risoluzione di problemi biologici a livello molecolare con metodi informatici, per questa ragione viene anche chiamata *biologia computazionale*. Argomenti di interesse di questa disciplina sono:

- allineamento di sequenze genetiche
- predizione genica
- predizione della struttura di proteine
- espressione genica
- interazione proteina-proteina
- interpretazione di dati proveniente da esperimenti biochimici
- organizzazione e archiviazione conoscenze su genomi e proteomi
- modellizzazione di sistemi e reti biologiche

Come si può notare da questa lista una parte importante della bioinformatica si occupa dell’utilizzo di strumenti informatici finalizzati a manipolare, archiviare e confrontare stringhe e sequenze di caratteri. Tuttavia questa disciplina non si ferma all’analisi delle sequenze. Tra le più interessanti applicazioni bioinformatiche odierne vi sono quelle incentrate sull’analisi strutturale^[30]. Difatti la bioinformatica pone le sue fondamenta nel campo della *structural bioinformatics*: per portare un esempio il database PDB (*Protein Data Bank*) nasce negli anni ’70 per archiviare le informazioni strutturali atomiche ricavate dagli studi cristallografici sulle proteine^[31].

Non va confusa la bioinformatica (o biologia computazionale) con la *computazione bio-ispirata* (es. algoritmi genetici, reti neurali), con il *biological computing* (ossia computer composti di parti biologiche come DNA, proteine o neuroni) o con la *biological computation* (l’idea che gli organismi eseguano computazioni e che le idee di informazione e computazione possano essere la chiave per comprendere la biologia)^[32].

Il Machine Learning (ML) è uno dei paradigmi informatici che più sta influenzando il campo della bioinformatica (come la presente tesi può dimostrare). Questo è dovuto principalmente a due fattori evolutisi in parallelo negli ultimi anni: la crescita esponenziale di dataset biologici disponibili e i progressi informatici del ML. Gli strumenti di ML possono apprendere caratteristiche dei sistemi biologici inferendole direttamente dai dataset. Quando propriamente allenati questi sistemi possono fornire accurate predizioni di caratteristiche astratte, proprio come nel caso di AlphaFold per il problema della predizione della struttura di proteine.

2.2.2 Soft computing

Il *soft computing* è un paradigma che si contrappone a quello dell'*hard computing*, ovvero la risoluzione di un problema tramite l'esecuzione di un algoritmo ben definito e decidibile. Il soft computing accantona la precisione od ottimalità e innalza a obiettivo il guadagno nella comprensione del comportamento di un sistema. Il soft computing si basa su due principi:

1. l'apprendimento a partire dai dati
2. l'integrazione di conoscenza umana basata sull'esperienza, strutturata e preesistente, all'interno di modelli matematici computabili

Il ML si avvale delle tecniche del soft computing^[33] e vi entra pienamente: la stima di performance in ML è infatti l'*accuratezza predittiva*, stimata dall'errore calcolato sul test set.

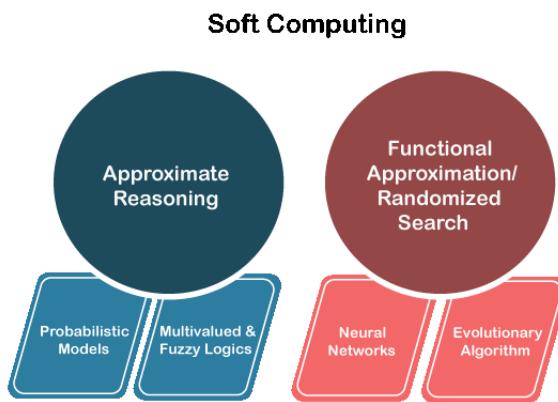


Figura 2.23: Branche del soft computing. Fonte: [34]

Evolutionary computation

La computazione evolutiva (EC) è una tecnica di ottimizzazione basata sui concetti darwiniani di evoluzione. I principali rappresentanti di questa tecnica sono gli algoritmi evolutivi (EA) e gli algoritmi genetici (GA).

Gli algoritmi genetici fanno parte del paradigma relativo alle tecniche informatiche *bio-ispirate*, così come le reti neurali. Un algoritmo genetico è un algoritmo euristico utilizzato per tentare di risolvere problemi di ottimizzazione. L'aggettivo "genetico", ispirato al principio della selezione naturale ed evoluzione biologica, deriva dal fatto che, al pari del modello evolutivo darwiniano che trova spiegazioni nella genetica, gli algoritmi genetici attuano dei meccanismi concettualmente simili a quelli dei processi biochimici genetici, come il *crossing over*.

2.2.3 Intelligenza Artificiale

Definire cosa sia l'intelligenza non è un compito semplice. Una definizione ampia e utilizzata nel mondo dell'AI è quella data da Kurzweil:

«*L'arte di creare macchine che svolgono funzioni che richiedono intelligenza quando svolte da esseri umani»*⁶

Una definizione di intelligenza proveniente da uno sfondo culturale del tutto diverso è la seguente:

«*The role of intelligence is to determine the positive and negative potential of an event or factor which could have both positive and negative results. It is the role of intelligence, with the full awareness that is provided by education, to judge and accordingly utilize the potential for one's own benefit or well-being»*⁷

Nella sua accezione più semplice, l'Intelligenza Artificiale (AI) si riferisce a sistemi che imitano l'intelligenza umana per eseguire certe attività e che sono in grado di migliorarsi continuamente in base alle informazioni raccolte. L'IA si occupa della costruzione di macchine intelligenti, della comprensione mediante modelli computazionali dei comportamenti e della psicologia di uomini, animali e agenti artificiali e può avere applicazioni innumerevoli nella società. I fondamenti dell'IA sono sin dalla nascita interdisciplinari: filosofia, matematica, economia, neuroscienze, psicologia, informatica, linguistica, cibernetica, statistica, complessità, teoria del controllo, teoria dell'informazione, robotica.

⁶R. Kurzweil, R. Richter, R. Kurzweil et al., *The age of intelligent machines*, 1990

⁷H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011

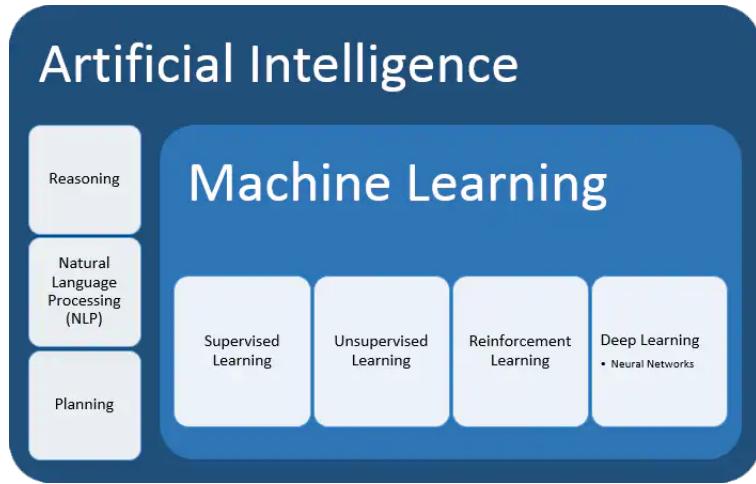


Figura 2.24: Schema riassuntivo dei campi dell'IA. Fonte: [37]

2.2.4 Machine Learning

Il Machine Learning (ML) è un sottoinsieme dell'AI che si occupa di creare sistemi che automaticamente migliorano con l'esperienza, basandosi su rigorosi fondamenti delle scienze computazionali. Utilizza metodi statistici per migliorare la performance di un algoritmo nell'identificare pattern nei dati. Domande fondamentali di questo campo sono del tipo: "come varia la performance di apprendimento al variare del numero di esempi di allenamento presentati?". Il ML è uno strumento molto potente ma è importante comprenderne i limiti. È utile quando non esiste o è difficile da formalizzare la teoria attorno ad un problema, oppure quando i dati da analizzare sono incerti, rumorosi o incompleti.

L'apprendimento è al cuore del problema dell'intelligenza, sia biologica che artificiale, ed è un principio universale comune a tutti gli organismi. Tom M. Mitchell definisce in questo modo l'apprendimento per una macchina:

«Si dice che un programma apprende dall'esperienza E con riferimento ad alcune classi di compiti T e con misurazione della performance P, se le sue performance nel compito T, come misurato da P, migliorano con l'esperienza E.»⁸

Il ML si divide in:

- *Supervised Learning*, ad es. SVM (support vector machine), in cui ai modelli vengono forniti degli esempi nella forma di possibili input e i rispettivi output desiderati e l'obiettivo è quello di estrarre una regola generale che associa l'input all'output corretto; comuni sono i task di classificazione e regressione

⁸T. Mitchell, *Machine learning*. McGraw hill New York, 1997

- *Unsupervised Learning*, in cui il modello ha lo scopo di trovare una struttura negli input forniti, come un raggruppamento naturale nei dati, senza che gli input vengano etichettati in alcun modo
- *Reinforcement Learning*, il modello interagisce con un ambiente dinamico nel quale cerca di raggiungere un obiettivo (per esempio guidare un veicolo, o imparare a giocare contro un avversario), avendo un insegnante che gli dice solo se ha raggiunto l'obiettivo
- *Deep Learning*, insieme di tecniche basate su reti neurali artificiali organizzate in diversi strati, dove ogni strato calcola i valori per quello successivo; si basa su diversi livelli di rappresentazione, corrispondenti a gerarchie di caratteristiche

Le tecniche SVM eseguono compiti di classificazione costruendo un iperpiano in uno spazio multidimensionale, e cercano di massimizzare il margine tra le diverse classi. La funzione che esegue la trasformazione dello spazio è chiamata funzione di *kernel*.

2.2.5 Deep Learning

Reti neurali artificiali (ANN)

Una rete neurale artificiale (*Artificial Neural Network*) è un modello computazionale composto da neuroni artificiali bio-ispirato alla semplificazione di una rete neurale biologica. È importante notare che l'obiettivo della modellizzazione bio-ispirata non è una comprensione delle reti neurali biologiche, data la semplicità dei modelli utilizzati, ma il tentativo di risolvere problemi ingegneristici sfruttando idee derivanti da queste. Nonostante ciò le ANN riflettono tratti di comportamento del cervello umano e consentono di riconoscere pattern e risolvere problemi difficili.

Le ANN sono composte da strati di nodi: uno strato di input, uno o più nascosti e uno di output. Ogni nodo è un neurone artificiale, si connette a tutti i nodi dello strato successivo e ha associato un peso e una soglia. Se l'output di un nodo è sopra la soglia allora il neurone è attivato, trasferendo informazioni al prossimo strato della rete. Con l'allenamento le ANN possono migliorare la loro accuratezza e rivelarsi potenti strumenti. Campi di utilizzo sono, fra gli altri, lo *speech-recognition* e l'*image recognition*⁹.

⁹La sezione seguente su alcuni aspetti del DL utili per l'argomento della tesi è basata sull'articolo M. Torrisi, G. Pollastri e Q. Le, “Deep learning methods in protein structure prediction,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, 2020.

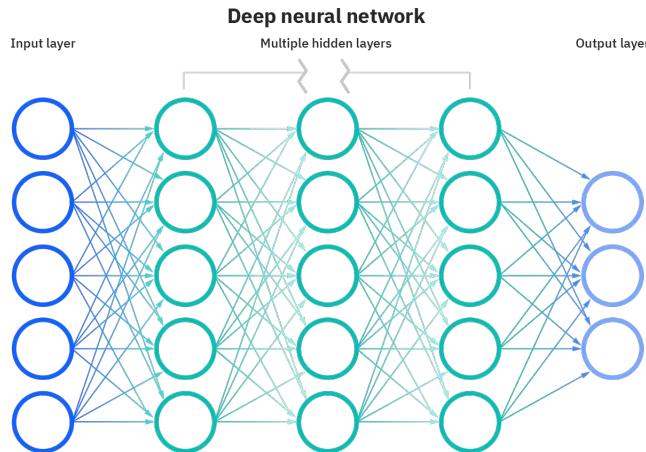


Figura 2.25: Rete neurale artificiale. Fonte: [39]

Feed Forward Neural Network (FFNN)

Una Feed Forward Neural Network (FFNN) è una rete neurale artificiale che non contiene cicli. In particolare le FFNN a strati sono FFNN i cui nodi possono essere partizionati in gruppi (strati) che sono ordinati e in cui gli output del livello i sono input solo ed unicamente per il livello $i + 1$. Il primo livello è noto come livello di input, l'ultimo come il livello di output e qualsiasi livello intermedio è un livello nascosto le cui unità formano una rappresentazione intermedia di un'istanza.

Le FFNN a strati possono essere addestrate con esempi usando l'algoritmo di *back-propagation* e hanno dimostrato di avere proprietà di approssimazione universale.

Deep Learning

La parola "deep" in *deep learning* si riferisce alla profondità degli strati in una rete neurale. Una rete neurale artificiale che consiste in almeno 4 strati (inclusi quello di input e output) può essere considerata un algoritmo di *deep learning*^[39]. Una rete neurale con 2 o 3 strati è una rete neurale semplice.

Il Deep Learning è un sottocampo del Machine Learning basato su reti neurali artificiali, che enfatizza l'uso di più livelli connessi per trasformare gli input in funzionalità suscettibili di prevedere gli output corrispondenti. Dato un set di dati sufficientemente ampio di coppie input-output, è possibile utilizzare un algoritmo di addestramento per apprendere automaticamente la mappatura dagli input agli output regolando un insieme di parametri a ogni livello della rete. Sebbene in molti casi gli elementi costitutivi elementari di un sistema di Deep Learning siano FFNN o unità elementari simili, questi vengono combinati in stack profondi utilizzando vari modelli di connettività.

Questa flessibilità architettonica consente di personalizzare i modelli di Deep Learning per qualsiasi particolare tipo di dati. I modelli di deep learning possono generalmente essere addestrati su esempi mediante *back-propagation*, che porta a rappresentazioni interne efficienti dei dati appresi per un’attività. Questo apprendimento automatico delle funzionalità elimina in gran parte la necessità di eseguirne l’ingegneria manuale. Tuttavia, i modelli di Deep Learning contengono un gran numero di parametri interni che per essere ben regolati hanno bisogno di molti dati; si dice quindi i modelli di DL siano *data-greedy*: le applicazioni di maggior successo del Deep Learning fino ad oggi sono state in campi in cui è disponibile un numero molto elevato di esempi.

Convolutional Neural Networks (CNN)

Le reti neurali convoluzionali (CNN) sono un’architettura progettata per elaborare dati organizzati con una dipendenza spaziale regolare (come i token in una sequenza o i pixel in un’immagine). Uno strato in una CNN sfrutta questa regolarità applicando lo stesso insieme di filtri convoluzionali locali tra le posizioni nei dati, acquisendo due vantaggi:

- evita il problema dell’overfitting avendo un numero molto piccolo di pesi da regolare rispetto allo strato di input e alla dimensionalità dello strato successivo
- è invariante alla traslazione

Un modulo CNN è solitamente composto da più livelli CNN consecutivi in modo che i nodi ai livelli successivi abbiano campi ricettivi più ampi e possano codificare caratteristiche più complesse. Va notato che l’FFNN ”a finestra” può essere considerata come una versione particolare e superficiale della CNN.

Recurrent Neural Networks (RNN)

Le reti neurali ricorrenti (RNN) sono progettate per apprendere caratteristiche globali dai dati sequenziali. Durante l’elaborazione di una sequenza di ingresso, un modulo RNN utilizza un vettore di stato interno per riassumere le informazioni dagli elementi elaborati della sequenza: dispone di un sottomodulo parametrizzato che prende in ingresso il vettore di stato interno precedente e l’elemento di ingresso corrente della sequenza, per produrre il vettore di stato interno corrente. Il vettore dello stato finale riassumerà l’intera sequenza di input.

Poiché la stessa funzione viene applicata ripetutamente agli elementi di una sequenza, i moduli RNN soffrono facilmente del problema della scomparsa del gradiente o dell’esplosione del gradiente quando si applica l’algoritmo di propagazione all’indietro per addestrarli.

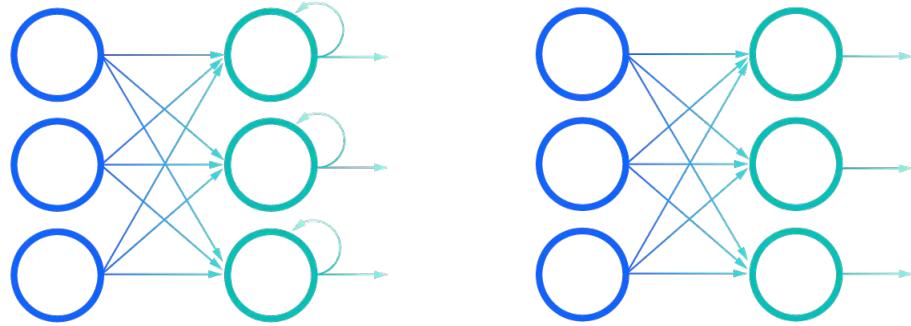


Figura 2.26: Confronto di RNN (sulla sinistra) e FFNN (sulla destra). Fonte[41]

I moduli di rete neurale ricorrenti *gated* come *Long Short Term Memory* (LSTM) o *Gated Recurrent Unit* (GRU) sono progettati per alleviare questi problemi. Sono possibili anche versioni bidirezionali di RNN (BRNN), particolarmente appropriate nelle previsioni delle *annotazioni* nei metodi per la predizione della struttura delle proteine (vedi la sez. 4.1.1), dove le istanze di dati non sono sequenze nel tempo ma nello spazio ed è desiderabile la propagazione delle informazioni contestuali in entrambe le direzioni.

Residual Neural Network (ResNet)

Anche se la profondità di un modello di Deep Learning ne aumenta l'espressività, l'aumento della profondità rende anche più difficile ottimizzare i pesi della rete a causa della scomparsa o dell'esplosione dei gradienti. Sono state proposte le Residual Network (ResNet) per risolvere questi problemi. Aggiungendo una connessione che "salta" (*skip*) da uno strato ad uno successivo, una ResNet viene inizializzata per essere vicina alla funzione di identità, evitando così grandi interazioni moltiplicative nel flusso del gradiente. Inoltre, le connessioni skip fungono da "scorciatoie", fornendo percorsi input-output più brevi affinché il gradiente scorra in reti altrimenti profonde.

I modelli ResNet tipici sono implementati con skip a doppio o triplo strato che contengono non linearità (ReLU). Una matrice di pesi aggiuntiva può essere utilizzata per apprendere i pesi di salto (HighwayNets). I modelli con diversi salti paralleli sono indicati come DenseNets.

L'idea della ResNet si basa su costrutti noti dalle cellule piramidali nella corteccia cerebrale. Il cervello ha strutture simili alle reti residue, poiché i neuroni dello strato corticale VI ricevono input dallo strato I, saltando gli strati intermedi. Nella figura 2.28 si può comparare la connessione di skip al fatto che i segnali del dendrite apicale saltino gli strati, mentre (semplificando) il dendrite basale raccoglie i segnali dello strato precedente e/o dello stesso.

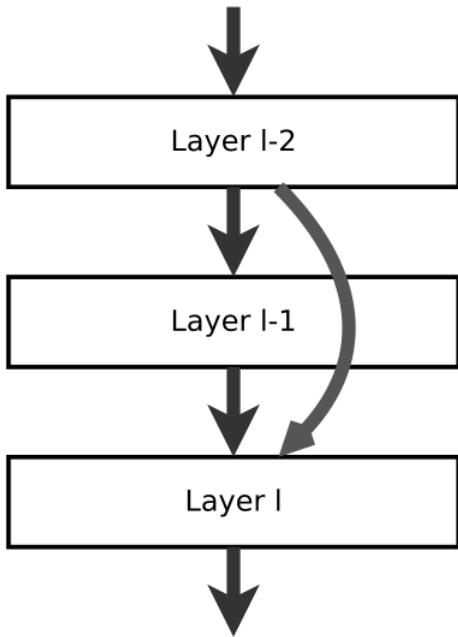


Figura 2.27: Schema della forma canonica di una ResNet. Uno strato è saltato da una connessione "skip". Fonte: [42]

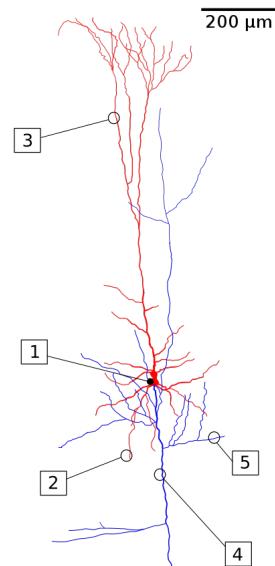


Figura 2.28: Ricostruzione di un neurone piramidale. Soma e dendriti sono etichettati in rosso, l'assone in blu. (1) Soma, (2) dendrite basale, (3) dendrite apicale, (4) assone, (5) assone collaterale. Fonte [42]

Non è chiaro quanti strati nella corteccia cerebrale siano paragonabili agli strati in una rete neurale artificiale, né se ogni area della corteccia cerebrale mostri la stessa struttura, ma su vaste aree sembrano simili. Non ci sono prove che qualcosa come la *back-propagation* abbia luogo nel cervello e non è supportata da prove neurofisiologiche nel cervello animale né l'esistenza di un "segnale di insegnamento" globale né l'ottimizzazione iterativa.

Bias induttivo per i modelli di deep learning

Nel Machine learning, il *bias induttivo* di un algoritmo è l'insieme di assunzioni che il classificatore usa per predire l'output dati gli input che esso non ha ancora incontrato. Senza bias induttivo sarebbe impossibile poter fare delle classificazioni e predizioni sui futuri dati.

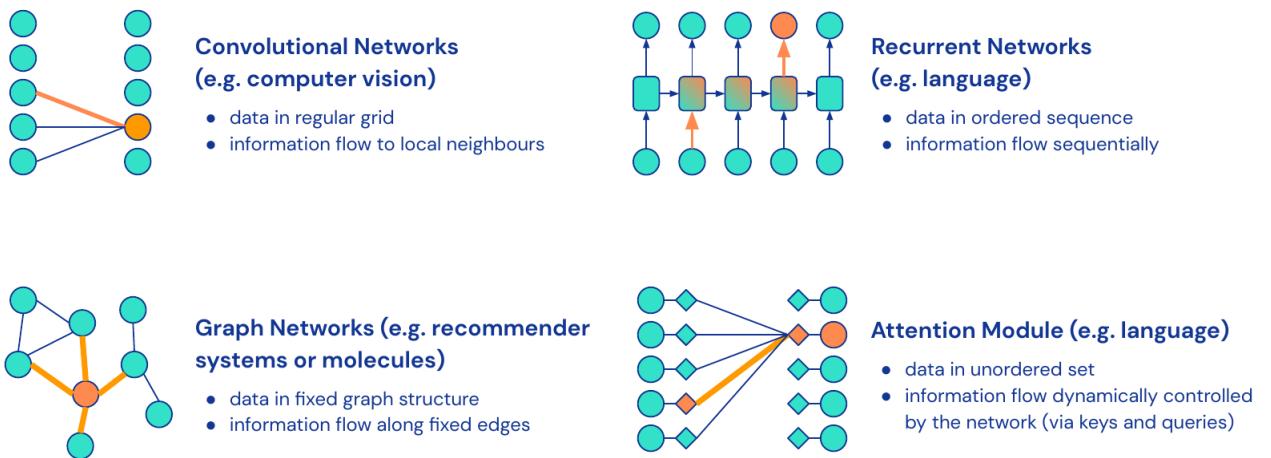


Figura 2.29: Flusso di informazione in vari modelli di deep learning. Fonte[43]

Capitolo 3

Protein Folding

«la forma è l'immagine plastica della funzione»¹

La correlazione tra forma e funzione si rivela fondamentale nel caso delle proteine. Un canale ionico neuronale permette il passaggio di ioni grazie alla sua forma a canale; una ferritina cattura e immagazzina gli ioni ferro grazie alla sua forma a sfera cava.

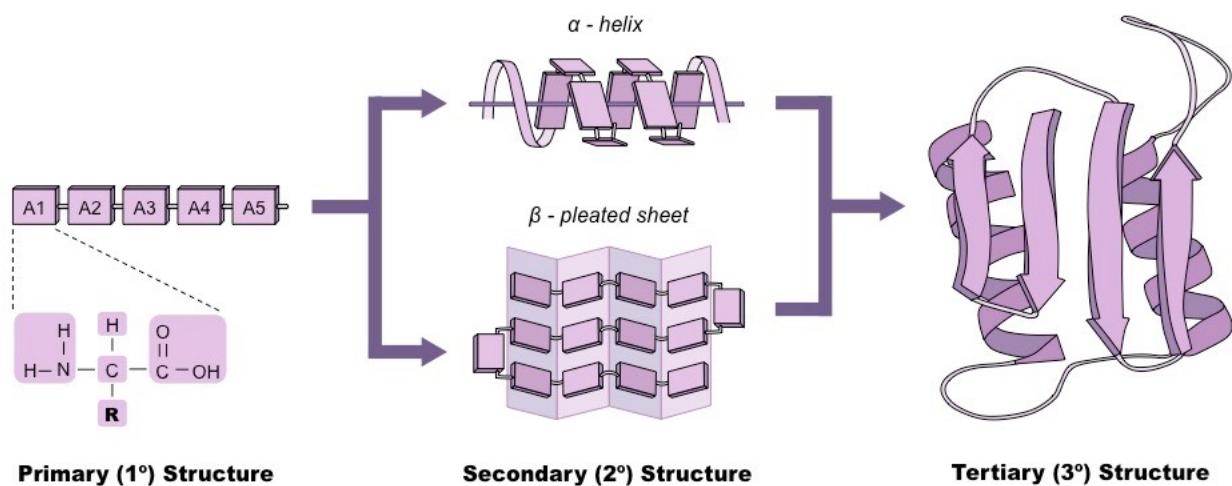


Figura 3.1: Protein folding: dagli amminoacidi alla struttura tridimensionale. Fonte: [45]

Il ripiegamento delle proteine (*protein folding*) è il processo di ripiegamento molecolare attraverso il quale a partire dalla sequenza lineare amminoacidica le proteine ottengono la loro struttura tridimensionale, chiamata forma *nativa*, che permette loro di svolgere la relativa funzione biologica².

¹A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925

²Per una trattazione superficiale della mancanza di generalità di questo paradigma si veda la sezione

Il ripiegamento nella forma tridimensionale avviene spontaneamente sia durante la sintesi proteica nei ribosomi sia al termine di questa. Una specifica proteina si ripiegherà nello stesso modo e avrà la stessa struttura finale³.

La prima teoria del ripiegamento proteico è stata proposta negli anni venti del XX secolo da Hsien Wu^[46], in relazione al processo di denaturazione (vedi sezione 3.1.1). È però Anfinsen, premio Nobel per la chimica, negli anni '60 a compiere un fondamentale passo nella comprensione del processo del ripiegamento proteico^[47].

3.1 Postulato di Anfinsen

Il postulato di Anfinsen (conosciuto anche come *dogma o ipotesi termodinamica* di Anfinsen) afferma che la struttura nativa di una proteina è determinata dalla sua sequenza di amminoacidi, sotto condizioni native. In altri termini: la struttura nativa, in ambiente fisiologico standard, corrisponde a quella struttura unica, stabile e cinematicamente accessibile avente *minima energia libera*.

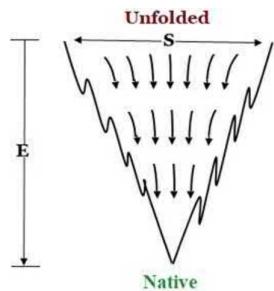


Figura 3.2: Un profilo energetico idealizzato dell'energia libera a forma di imbuto. E=energia, S=entropia.
Fonte: [48]

Vi sono quindi 3 condizioni:

1. *unicità*, la sequenza non deve possedere altre configurazioni dotate di energia libera comparabile
 2. *stabilità*, piccoli cambiamenti nell'ambiente circostante non possono produrre cambiamenti nella configurazione a energia minima. Ciò può essere descritto come una superficie parabolica di energia libera con lo stato nativo corrispondente al punto
-
- 3.5. Per una scrittura e lettura più agevole della presente tesi si è preferito accettare la visione del paradigma.

³Ciò non è vero nel 100% dei casi, alcune proteine possono avere più di una conformazione stabile per adempire funzioni diverse (vedi la sezione 3.5) e alcune proteine possono andare incontro a misfolding (vedi la sezione 3.3).

di minimo (visivamente simile ad un imbuto, vedi fig. 3.2); la superficie di energia libera nelle vicinanze dello stato nativo deve essere abbastanza ripida ed elevata

3. *accessibilità cinetica*, il percorso nella superficie di energia libera dallo stato *unfolded* a *folded* deve essere ragionevolmente piano

Esperimento di Anfinsen

L'esperimento, compiuto nel 1957^[49], consisteva nella denaturazione e rinaturazione della ribonucleasi A, dimostrando che il secondo processo era possibile senza agenti ausiliari. L'enzima in questione è formato da 124 amminoacidi, tra cui 8 cisteine che formano 4 punti disolfuro ($-CH_2 - S-S - CH_2 -$, vedi sez. 3.2.1). È stato usato un agente riducente per scindere questi punti e l'urea per denaturare la proteina: questa non mostrava più alcuna attività enzimatica. A questo punto se l'urea era rimossa prima, seguita dall'aggiunta di un agente ossidante per consentire ai punti disolfuro di riformarsi, la ribonucleasi A riacquistava spontaneamente la sua struttura terziaria e il prodotto ottenuto risultava praticamente indistinguibile dalla proteina nativa di partenza, riottenendo piena attività biologica.

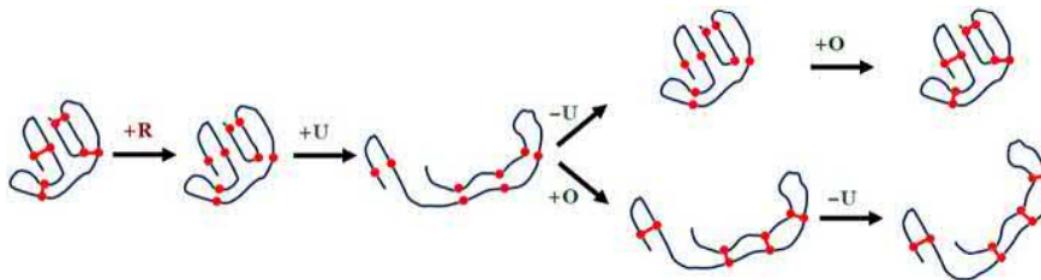


Figura 3.3: Rappresentazione schematica dell'esperimento di Anfinsen. R=reducing agent, U=Urea, O=oxidizing agent, punti rossi=cisteina, linee rosse=ponti disolfuro. Fonte: [48]

I punti disolfuro si riformano nella stessa posizione della proteina nativa nonostante ci siano 105 modi possibili per ricombinarli. Se invece veniva prima aggiunto l'agente ossidante e poi tolta l'urea il prodotto ottenuto era un miscuglio di molte delle possibili 105 configurazioni, raggiungendo solamente l'1% dell'attività enzimatica.

Dai lavori di Anfinsen è possibile trarre due ulteriori importanti conclusioni^[52]:

- ha dato via alla grande avventura della ricerca nel campo del protein folding *in vitro*⁴ piuttosto che all'interno di cellule. La struttura nativa non dipendeva quindi dal fatto che la proteina fosse sintetizzata biologicamente con l'aiuto di ribosomi (ed

⁴La locuzione latina *in vivo* significa *nel vivente*. Se il fenomeno biologico viene riprodotto in una provetta si dice *in vitro* mentre se lo si riproduce tramite una simulazione computazionale si dice *in silico*.

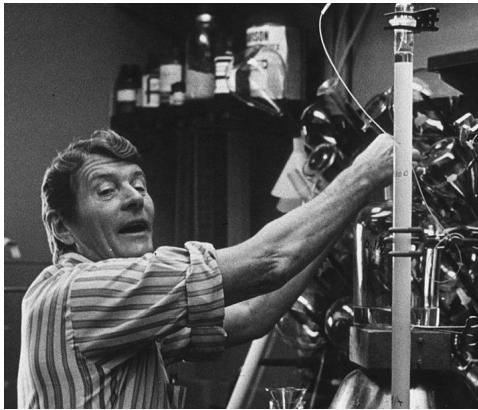


Figura 3.4: C.B. Anfinsen nel suo laboratorio.
Fonte: [50]

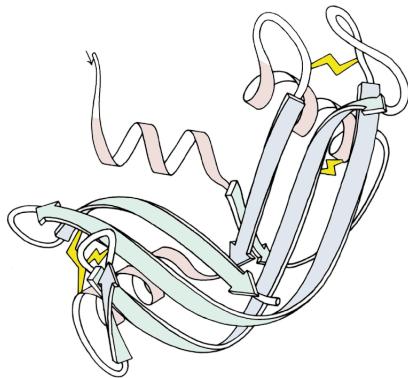


Figura 3.5: Ribonucleasi A, rappresentazione a nastro. In giallo i ponti disolfuro, rosa le α -eliche, verde e azzurro i β -foglietti. Fonte [51]

eventualmente chaperoni molecolari) o che si ripiegasse nuovamente come molecola isolata all'interno di una provetta

- l'evoluzione può agire in modo da cambiare la sequenza amminoacidica ma l'equilibrio del ripiegamento e la cinetica di una data sequenza sono materia della fisica chimica.

3.1.1 Denaturazione

La denaturazione delle proteine è il fenomeno relativo all'alterazione della struttura nativa dovuto a variazioni di temperatura, pH o contatto con determinate sostanze chimiche. La denaturazione è un processo che porta alla perdita di ordine e quindi ad un aumento di entropia. La struttura primaria rimane invariata, data la stabilità dei legami peptidici. A causa della denaturazione le proteine perdono la loro funzione biologica e possono esporre e rendere reattivi alcuni gruppi funzionali che possono causare l'aggregazione di più proteine. Può avvenire che una volta rimosso l'agente denaturante la proteina ritorni allo stato di partenza (*rinnaturazione*) ma spesso il processo è irreversibile.

La proprietà di certe sostanze chimiche (es. urea) di denaturare una molecola proteica si deve alla loro capacità di legare transientemente, attraverso legami deboli, come ad esempio legami idrogeno, i residui amminoacidici costituenti la proteina. Questi legami vengono termodinamicamente preferiti a quelli intramolecolari o intermolecolari con l'acqua. Ciò comporta l'impossibilità per la proteina di mantenere la propria struttura tridimensionale e quindi questa si denatura.

Applicazioni nella vita quotidiana di questo fenomeno sono la cottura dei cibi (basti pensare all'albumina nell'uovo) e la permanente ai capelli (denaturazione dell' α -cheratina, rompendo e riformando ponti disolfuro).

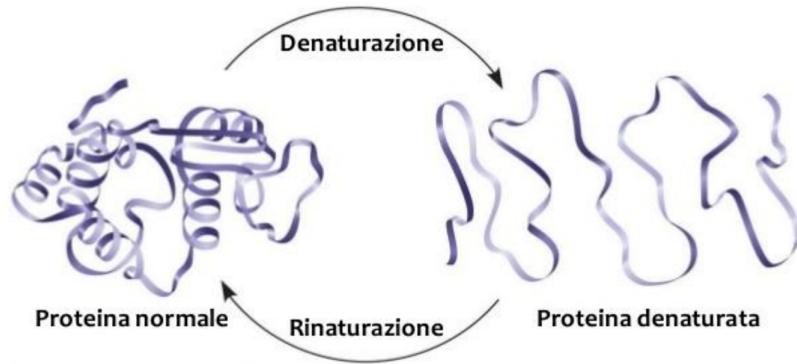


Figura 3.6: Denaturazione e rinaturazione. Fonte: [16]

3.2 Struttura delle proteine

Da un punto di vista chimico le proteine sono di gran lunga, tra quelle conosciute, le molecole strutturalmente più complesse e sofisticate funzionalmente. È possibile studiare la loro struttura individuando successivi livelli di organizzazione:

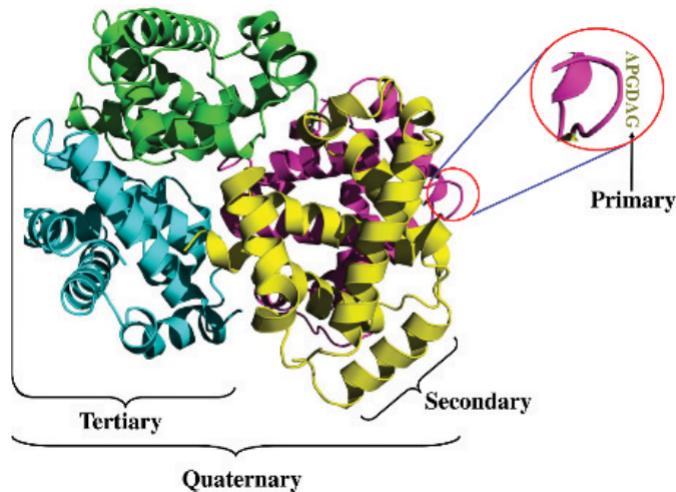


Figura 3.7: Livelli strutturali di una proteina. Fonte: [6]

- *struttura primaria*: la sequenza ordinata degli amminoacidi
- *struttura secondaria*: regioni ripetitive locali stabilizzate da legami idrogeno tra atomi della backbone (α -eliche e β -foglietti)
- *struttura supersecondaria*: combinazione di strutture secondarie e connessioni (motivi, domini, loop, giri ecc.)
- *struttura terziaria*: forma tridimensionale di una singola catena polipeptidica, risultante dalle interazioni dei residui

- *struttura quaternaria*: forma finale di proteine "assemblate" da 2 o più catene polipeptidiche già ripiegate

Prima di passare ad analizzare ogni livello della struttura delle proteine è utile un veloce sguardo ai legami chimici e alle interazioni molecolari.

3.2.1 Legami e interazioni molecolari

La chimica della vita è di un tipo speciale: è una chimica organica formata da composti carboniosi, in un ambiente acquoso, con temperature "terrestri" e complicata, basata su grandi polimeri. Gli atomi possono risultare incompleti, e grazie a questo formare legami per completarsi. Elementi puri e pienamente completi non trovano spazio nella chimica della vita. Nei viventi solo gli elettroni si spostano⁵ per ricercare stabilità, ovvero per permettere agli atomi di completare il loro guscio orbitale più esterno. Ogni atomo può avere tanti *legami* quanti elettroni gli mancano per completare il suo guscio più esterno. Le *interazioni molecolari* sono forze attrattive o repulsive tra molecole e tra atomi non legati. La *forza di legame* è la misura dell'energia necessaria per romperlo (in kJ/mol o kcal/mol). Si elencano ora i principali legami inerenti al ripiegamento delle proteine:

- *legame covalente*: prevede la compartecipazione di 2 elettroni di valenza fra più atomi ed è il tipo di legame più forte. Due o più atomi tenuti insieme da legami covalenti formano una molecola. C'è una specifica distanza di legame fra i nuclei degli atomi bilanciata tra forze attrattive e repulsive: se sono troppo vicini c'è repulsione mentre se sono troppo lontani non c'è attrazione.
 - *elettronegatività*: spesso gli elettroni in un legame sono condivisi iniquamente. Questo dipende dall'elettronegatività degli atomi, ad esempio l'ossigeno ha elettronegatività 3.4 mentre l'idrogeno 2.1. Quando la differenza di elettronegatività è compresa tra 0.5 e 1.9 la nube elettronica di legame risulta deformata verso l'atomo più elettronegativo, su cui si origina una carica parziale negativa (indicata con δ^-) mentre l'altro atomo acquisisce una carica parziale positiva di uguale valore assoluto. La molecola, divenuta *polare*, si può immaginare ora come un *dipolo* elettrico.
 - *ponti disolfuro*: i legami (o ponti) disolfuro sono legami covalenti tra due atomi di zolfo con energia di legame di 60kcal/mol. Si formano dall'accoppiamento di due gruppi tiolici (-SH). Essendo legami molto forti costituiscono un elemento architettonicale fondamentale nella struttura delle proteine. La cisteina presenta un gruppo -SH nella catena laterale e può quindi formare ponti disolfuro.

⁵Protoni e neutroni si separano solo in condizione estreme: nei reattori nucleari, nel sole, per decadimento radioattivo.

| Bond Type | Length* (nm) | Strength (kJ/mole) | |
|--|--------------|--------------------|-----------|
| | | In Vacuum | In Water |
| Covalent | 0.10 | 377 [90]** | 377 [90] |
| Noncovalent: ionic bond | 0.25 | 335 [80] | 12.6 [3] |
| Noncovalent: hydrogen bond | 0.17 | 16.7 [4] | 4.2 [1] |
| Noncovalent: van der Waals attraction (per atom) | 0.35 | 0.4 [0.1] | 0.4 [0.1] |

Figura 3.8: Distanza di legame approssimate e forza dei legami chimici. I valori della forza sono riportati in kJ/mol e in [kcal/mol]. Da notare la diminuzione di forza nel legame ionico se in ambiente acquoso. Fonte: [\[7\]](#)

- *legami non covalenti (interazioni molecolari)*
 - *attrazioni elettrostatiche*: le forze d’attrazione agiscono fra gruppi completamente carichi (legame ionico) e fra i gruppi parzialmente carichi delle molecole polari. Decresce con la distanza. Molto deboli in acqua.
 - * *legame ionico*: l’atomo più elettronegativo strappa completamente un elettrone al suo compagno, si formano due ioni (uno positivo, *catione* e uno negativo *anione*). Si ha quando la differenza di elettronegatività tra i due atomi è maggiore di 1.9.
 - *legame idrogeno*: è una forza dipolo-dipolo che si origina tra molecole contenenti un atomo di idrogeno unito covalentemente a ossigeno, fluoro o azoto. Un atomo di idrogeno elettropositivo è parzialmente condiviso da due atomi elettronegativi; ad es. nell’acqua gli atomi di idrogeno (parzialmente positivi) si trovano fra due atomi di ossigeno (parzialmente negativi). L’idrogeno, legato a uno dei due atomi di ossigeno, permette all’altro di avvicinarsi e di stabilizzare le molecole. Sono legami deboli singolarmente (1/20 della forza di un legame covalente) ma quando se ne formano simultaneamente molti sono abbastanza forti da fornire un legame stretto (l’acqua bollirebbe a -120°C senza legami idrogeno).
 - *interazioni di van der Waals*: nelle molecole apolari gli elettroni si possono accumulare in modo asimmetrico, formando regioni momentaneamente polari che permettono così una temporanea stabilizzazione fra molecole a breve distanza. Due atomi saranno attratti l’uno dall’altro fino a che la distanza fra i loro nuclei è approssimativamente uguale alla somma dei loro raggi di van der Waals (ad es. per il carbonio il raggio è di 0.2nm)

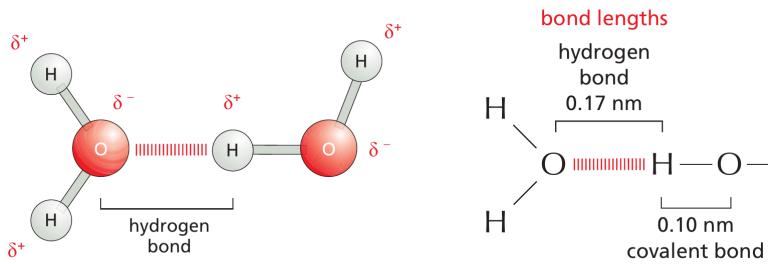


Figura 3.9: Legame idrogeno tra due molecole d'acqua. Fonte: [7]

- *forze idrofobiche*: l'acqua forza insieme i gruppi idrofobici; l'apparente attrazione è in realtà causata da una repulsione dall'acqua, che difende il suo reticolo tenuto insieme da legami idrogeno.

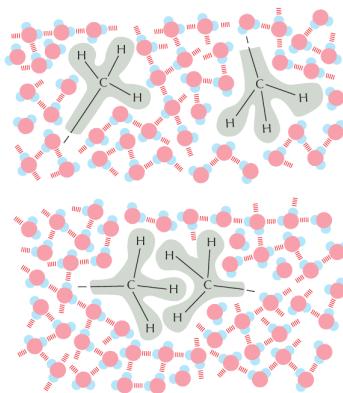


Figura 3.10: Forze idrofobiche. Fonte: [7]

Le sostanze *idrofile* si dissolvono rapidamente nell'acqua poiché le loro molecole formano legami idrogeno con le circostanti molecole d'acqua (nel caso di sostanze polari) o perché queste sono attratte dalle cariche degli ioni (nel caso di sostanze ioniche, es. cloruro di sodio, con ioni Na^+ e Cl^-). Le sostanze *idrofobiche* contengono perlopiù legami non polari e sono solitamente insolubili in acqua. Le molecole d'acqua in questo caso non sono attratte ma possono generarsi forze idrofobiche che raggruppano insieme tali sostanze (come nel nucleo idrofobico delle proteine). Per dettagli termodinamici riguardo l'idrofobicità vedi la parte finale della sez. 3.3.1.

3.2.2 Livelli strutturali

Struttura primaria

La struttura primaria delle proteine è la sequenza ordinata degli amminoacidi. La posizione nella sequenza di specifici amminoacidi è un fattore fondamentale per la determinazione di quali porzioni della proteina andranno a legarsi formando globalmente la struttura finale.

La nota importante, basata sul dogma di Anfinsen, è che la sequenza amminoacidica di ogni proteina contiene l'informazione che specifica sia la struttura nativa che la via per raggiungere quello stato. Questo comunque non vuol dire che strutture simili si ripieghino in modo simile.

Struttura secondaria

La struttura secondaria riguarda le regioni ripetitive locali stabilizzate da legami idrogeno tra atomi della backbone: α -eliche e β -foglietti.

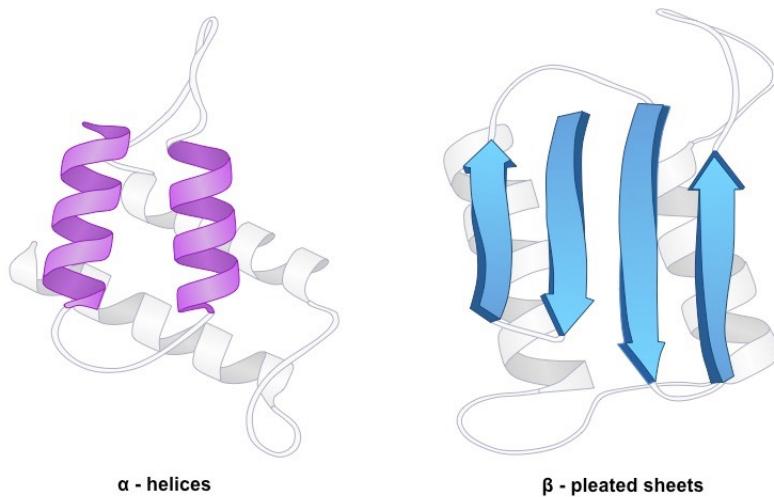


Figura 3.11: Struttura secondaria delle proteine, α -eliche e β -foglietti. Fonte: [45]

Questo livello di organizzazione è una conseguenza dei legami a idrogeno intramolecolari. All'interno della backbone del polipeptide gli atomi di ossigeno hanno una parziale carica negativa e gli atomi di idrogeno attaccati all'azoto hanno una parziale carica positiva perciò possono formarsi legami idrogeno fra questi atomi. Individualmente sarebbero deboli legami ma poiché sono ripetuti molte volte su di una regione relativamente lunga di una catena polipeptidica possono fare da supporto per una particolare conformazione.

Nella struttura ad α -elica, la struttura secondaria più comune e teorizzata già negli anni '50 da Linus Pauling, gli amminoacidi sono avvolti in una spirale tenuta insieme da legami idrogeno ogni 4 amminoacidi. Tra l'atomo di idrogeno legato all'azoto di ogni legame peptidico e l'ossigeno del gruppo carbossilico del legame peptidico sovrastante (che si trova a distanza di tre amminoacidi lungo la catena) si instaura un legame a idrogeno. Tuttavia se gli amminoacidi che si succedono lungo un tratto di catena proteica hanno gruppi R voluminosi, come avviene nella prolina, o gruppi R dotati della stessa carica elettrica, come avviene negli amminoacidi lisina e arginina, l' α -elica non può formarsi, a causa delle forze di repulsione che si generano tra i residui. Alcune proteine fibrose, come

l' α -cheratina, la proteina strutturale di capelli, lana e unghie hanno formazioni di α -eliche sulla maggior parte della loro lunghezza.

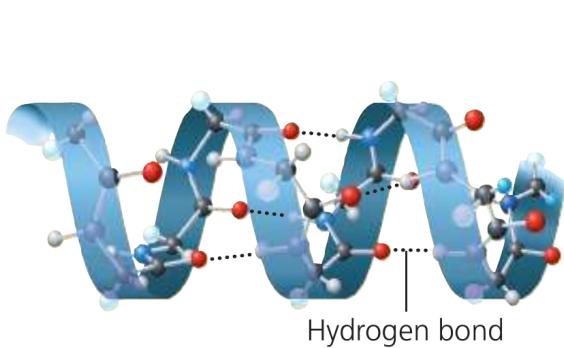


Figura 3.12: Regione di α -elica. Fonte: [16]

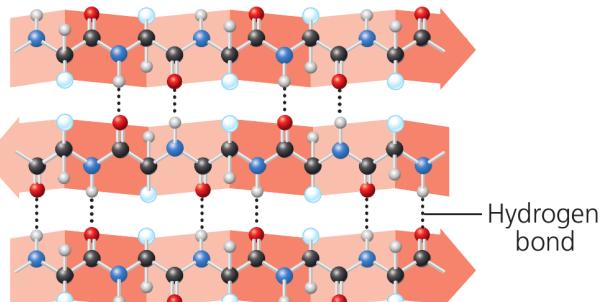


Figura 3.13: Una regione di β -foglietto composto da β -filamenti adiacenti, spesso mostrati come una freccia piegattata o piatta puntata in direzione C-terminus. Fonte [16]

Altre proteine fibrose sono invece dominate dai β -foglietti, come le proteine della seta (β -cheratina) e della tela prodotta dai ragni. In queste conformazioni due o più segmenti della catena polipeptidica giacenti lato su lato (chiamati β -filamenti) sono connessi da tre o più legami idrogeno. Si definisce β -filamento una sequenza peptidica di aminoacidi (tipicamente 5-10) che si dispone linearmente ed è in grado di formare legami idrogeno. Ciascuna delle catene è totalmente estesa e presenta una conformazione a zig-zag, dovuta alla geometria dei legami attorno a ciascun atomo di carbonio e di azoto nella catena.

I gruppi amminici di uno scheletro peptidico formano legame con quelli carbossilici del filamento opposto. In ogni singolo filamento i residui si dispongono perpendicolarmente al piano del foglietto, puntando alternativamente verso l'alto e verso il basso. I β -foglietti tendono a trovarsi all'interno del nucleo della struttura per evitare competizione con le molecole d'acqua per formare legami idrogeno e tendono a favorire residui idrofobici. Si dice che i filamenti sono paralleli quando vanno nella stessa direzione (la freccia che indica la direzione C-terminus è puntata nella stessa direzione).

Nella vita quotidiana, se tiriamo per i due estremi una fibra di lana questa si allunga: si stanno rompendo i legami idrogeno e le eliche si allontanano sempre di più, ma lasciando la presa i legami idrogeno si riformano e le eliche ricompaiono nella struttura. Se invece tiriamo la seta si può osservare che non è elastica: i foglietti di cui è composta la sua struttura non sono smantellabili senza rompere anche i legami covalenti della backbone.

Struttura supersecondaria

La struttura supersecondaria è riferita alle combinazioni spaziali di strutture secondarie in conformazioni più complesse e alle connessioni che li uniscono⁶. Può essere considerata come esempio di struttura supersecondaria la triplice elica allungata del collagene.

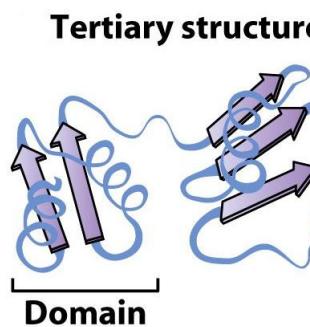


Figura 3.14: Dominio in una proteina. Fonte: [53]

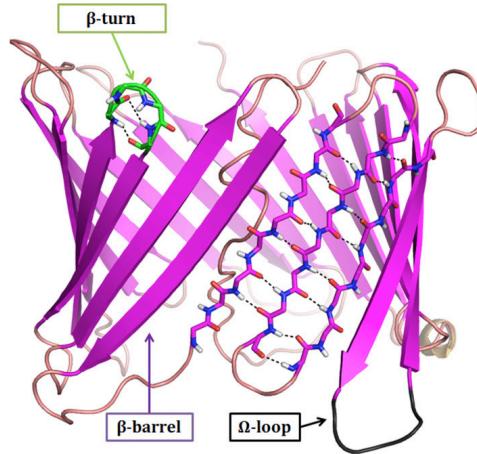


Figura 3.15: Struttura con giri, loop e motivo β -barile. Fonte: [54]

I *motivi* (motifs) e *domini* (domains) sono regioni tridimensionali della catena polipeptidica formate da differenti strutture secondarie adibite a svolgere una determinata funzione per la proteina di cui fanno parte. Tuttavia sono differenti in quanto i motivi non mantengono la loro forma se separati dalla proteina laddove i domini la mantengono. Questo perché i motivi e il resto della proteina sono più vicini e si vengono così a formare legami idrogeno che permettono ai motivi di mantenere la struttura. I domini sono sì legati alla backbone della proteina ma non abbastanza vicini alla restante parte della formazione proteica da stabilire legami, pertanto se vengono separati non perdono la loro struttura e possono mantenere la loro funzione. Una proteina con vari domini può usare questi per interazioni funzionali con differenti molecole.

Più in generale un *motivo strutturale* è una struttura tridimensionale comune che appare in una varietà di molecole differenti ed evoluzionisticamente scollegate. Nel contesto delle sequenze amminoacidiche si definisce *motivo* un pattern amminoacidico conservato in un gruppo di proteine con attività biochimica simile.

In figura 3.16 sono illustrati alcuni motivi comuni nelle strutture proteiche. Il motivo *elica-loop-elica* ad esempio consiste di due α -eliche collegate da un giro invertito. Un motivo simile è l'*elica-giro-elica* dove al posto di un loop si ha un giro che causa un cambio di direzione più netto. Questa particolare conformazione rende questo motivo in grado di

⁶Non c'è un accordo tra i vari studiosi su di una precisa classificazione di questo livello strutturale.

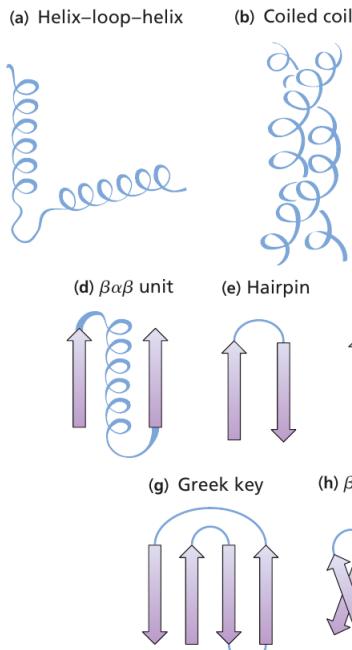


Figura 3.16: Motivi comuni. Fonte [53]

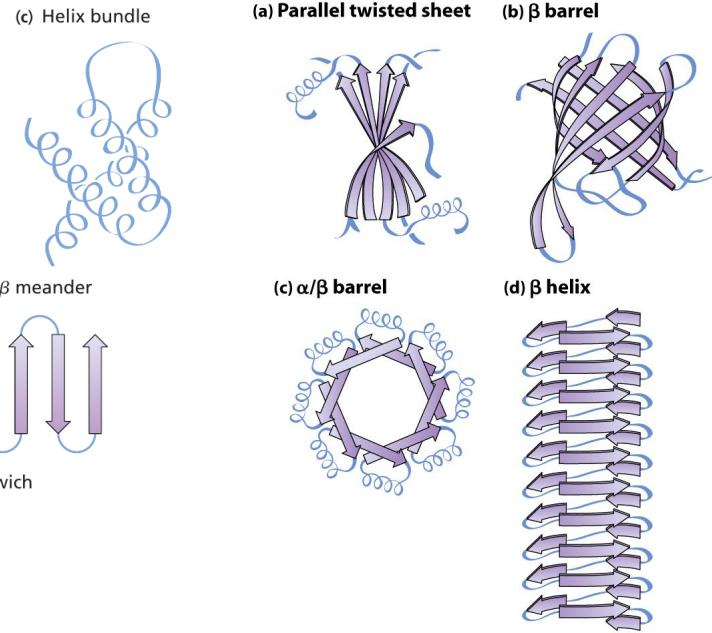


Figura 3.17: Domain folds (ripiegamenti di dominio). Fonte: [53]

legarsi alla scanalatura del DNA e infatti questo motivo si presenta in molte proteine che regolano l'espressione genica.

Il motivo a *simbolo greco* consiste di 4 β -filamenti antiparalleli in un β -foglietto dove l'ordine dei foglietti lungo la catena polipeptidica è 4,1,2,3⁷. In figure 3.15 e 3.17 è illustrato il motivo β -barile composto da β -foglietti ripiegati circolarmente a formare una struttura somigliante ad un barile comune in molte proteine di membrana. I *domain folds*, o ripiegamenti di dominio, sono grandi motivi che costituiscono il nucleo di un dominio.

Giri e *loop* causano cambi di direzione alla backbone della proteina. I loop sono regioni con una struttura tridimensionale fissa ma non regolare. Si trovano generalmente sulla superficie delle proteine. Non sono strutture casuali e non vanno confuse con regioni disordinate o dispiegate. Una loro funzione è di connettere strutture secondarie tra loro. I loop sono regioni con ruoli spesso cruciali (interazioni con altre proteine, siti di legame con molecole ecc.) ma sono anche molto variabili nella loro sequenza e struttura. È stato ipotizzato che la posizione degli introni nel DNA possa correlare con la locazione dei loop codificati nella proteina^[55].

Nelle strutture secondarie e terziarie si trovano spesso bruschi cambiamenti di direzione nella struttura: i *giri* (turns). Queste nette svolte sono possibili grazie agli amminoacidi prolina e glicina. Il gruppo R della prolina si ripiega verso il gruppo amminico, distorcendo la catena naturalmente. Si forma però uno stretto spazio a causa del giro: l'amminoacido

⁷I numeri indicano l'ordine dei filamenti ovvero la loro posizione nel β -foglietto da destra a sinistra.

con gruppo R meno voluminoso è ovviamente la glicina ed è per questo che si trovano insieme nei giri.

Struttura terziaria

La struttura terziaria è la struttura tridimensionale globale risultante dalle interazioni tra i residui successivamente alle conformazioni locali della struttura secondaria ed è quindi la descrizione del risultato del processo di ripiegamento proteico. Un tipo di interazione importante è quella idrofobica che induce i residui non polari (e quindi idrofobici) a raggrupparsi al centro della catena polipeptidica, formando un *nucleo idrofobico*. La forma della proteina può venire rinforzata dai ponti disolfuro, legami covalenti possibili solamente fra due cisteine.

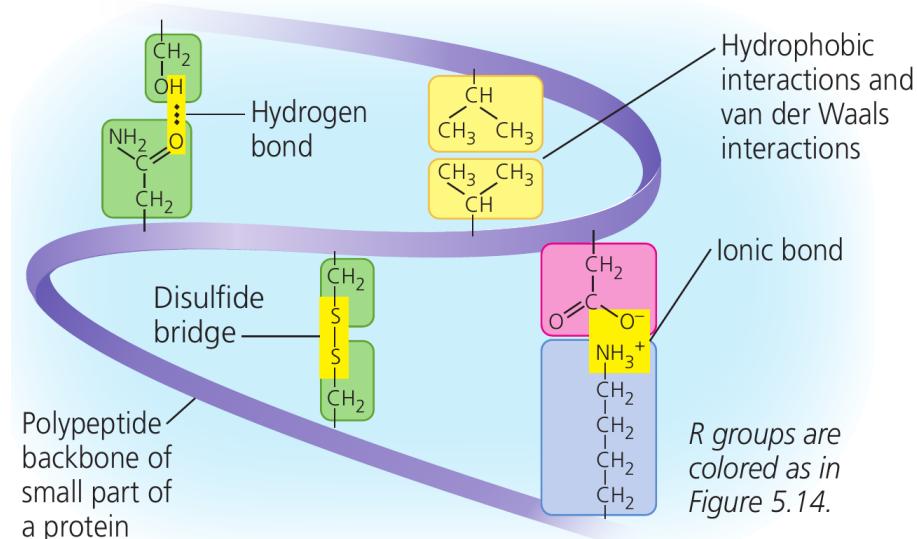


Figura 3.18: I diversi tipi di interazioni che possono contribuire alla struttura terziaria di una proteina.
Fonte: [16]

La glicina assume una speciale posizione tra gli amminoacidi dato che ha il gruppo R più piccolo, un solo atomo di idrogeno (vedi fig. 2.20): può aumentare la flessibilità locale nella struttura (come infatti accade nel caso dei *giri* sopra accennati).

Prima degli anni '80 il protein folding code (bilancio termodinamico delle forze interatomiche, vedi sez. 3.6) era visto come la somma di molte piccole interazioni (legami idrogeno, interazioni di van der Waals, attrazioni elettrostatiche) ma senza nessuna forza dominante^[52]. Negli anni '80, grazie alla modellazione basata sulla meccanica statistica, è emerso un nuovo paradigma: la componente dominante nel folding code sono le forze idrofobiche, il folding code è distribuito sia localmente che non localmente nella sequenza e le strutture secondarie di una proteina sono una conseguenza della struttura terziaria tanto quanto una causa. Poiché le strutture native sono solamente 5-10kcal/mol più stabili

dei loro stati denaturati è chiaro che nessuna forza intermolecolare può essere ignorata, e per questo la questione su quale forza sia quella dominante non è né semplice né risolta. Tuttavia risulta evidente che solo pochi residui risultano carichi nelle proteine, pertanto le forze elettrostatiche difficilmente possono essere dominanti.

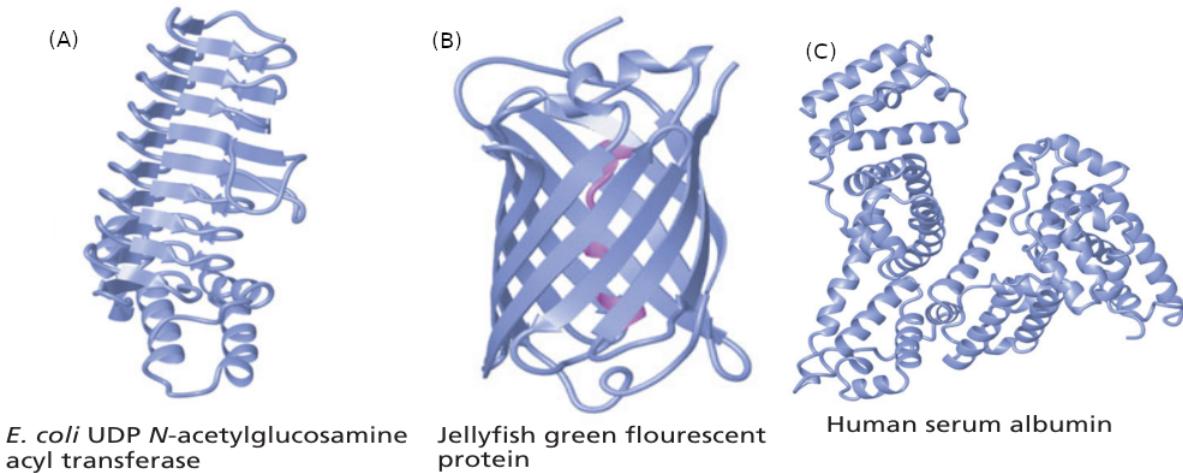


Figura 3.19: Esempi di strutture terziarie in alcune proteine. (A) (classe: all- β) La struttura dell'enzima mostra un classico esempio di β -eliche, struttura abbastanza rara. (B) (classe: all- β) Struttura a β -barile con un' α -elica centrale; i β -filamenti sono anti-parallel. (C) (classe: all- α) Albumina del siero umano. Ha molti domini costituiti da α -eliche a strati e helix bundle. Fonte [53]

Struttura quaternaria

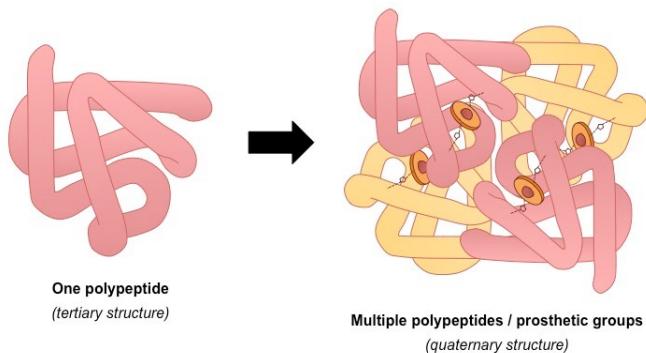


Figura 3.20: Rappresentazione di una struttura quaternaria composta da più polipeptidi e alcuni gruppi prostetici. Fonte [45]

La struttura quaternaria è la forma finale di proteine "assembrate" da 2 o più catene polipeptidiche già ripiegate. Il collagene ne è un esempio poiché è formata da 3 polipeptidi quasi interamente a spirale che si attorcigliano l'uno sull'altro formando un'elica tripla ancora più larga, dando alle lunghe fibre una grande forza (vedi anche la cheratina nella

sezione 3.2.3). Un altro esempio è l'emoglobina, proteina globulare formata da 4 subunità polipeptidiche. Le strutture terziarie delle subunità non vengono alterate.

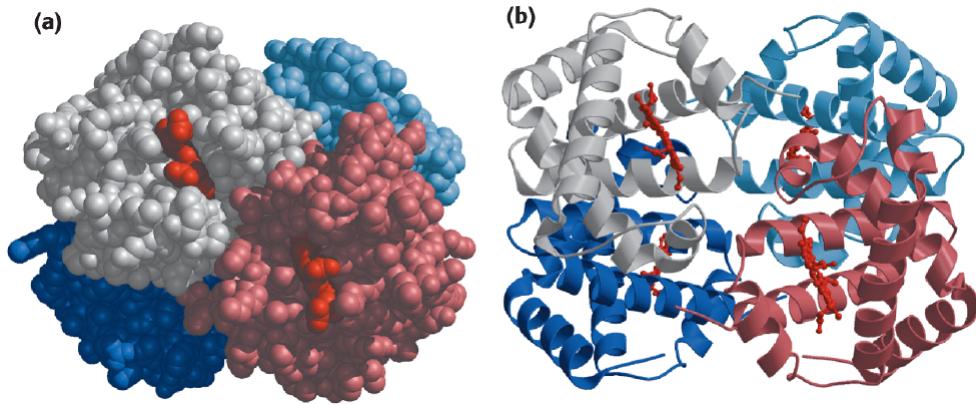


Figura 3.21: Struttura quaternaria della deossiemoglobina, ogni colore rappresenta una diversa subunità. (A) rappresentazione di tipo space-fill. (B) Rappresentazione a nastro. Fonte [53]

La struttura che ne risulta è spesso chiamata *oligomero* e le catene polipeptidiche costituenti sono dette *monomeri*, *protomeri* o *subunità*. Le subunità sono unite mediante legami idrogeno, ionici o forze idrofobiche. I rapporti spaziali tra le subunità sono fissi e la geometria della molecola globale è ben definita. L'unione delle subunità può far emergere proprietà non possedute dai singoli monomeri.

3.2.3 Evoluzione e classificazione

Evoluzione e conservazione

L'evoluzione degli organismi è legata a mutazioni spontanee che avvengono nei loro geni. Le differenze nella struttura primaria sono la “memoria” dei cambiamenti avvenuti a livello genetico nel corso dell’evoluzione. In specie legate da notevole affinità le strutture primarie delle proteine comuni sono simili. Proteine con funzioni analoghe presentano sequenze simili, è molto probabile quindi che queste sequenze si siano evolute a partire da un progenitore comune.

Al contrario di quanto si possa credere la maggior parte dei *motivi* non ha origini evolutive in comune. Motivi simili sono sorti indipendentemente e semplicemente convergono verso una struttura stabile comune. Il fatto che gli stessi motivi si presentino in centinaia di differenti strutture suggerisce l'esistenza di un numero limitato di possibili ripiegamenti nell'universo delle strutture proteiche^[56].

Classificazione

La classificazione delle proteine all'interno dei database può essere basata su somiglianze strutturali e/o di sequenza. A livello biologico, in base ai diversi livelli strutturali assunti, le proteine sono classificate in *fibrose* e *globulari*.

Le proteine fibrose sono caratterizzate dalla prevalenza di strutture secondarie rispetto a livelli di organizzazione superiore. Sono costituite da lunghe catene disposte in lunghi fasci o foglietti. La struttura è estremamente ordinata. Svolgono funzioni di protezione e sostegno. Le proteine fibrose costituiscono prevalentemente: pelle, piume, capelli, corna, unghia, squame (con funzione di protezione) e cartilagine, tendini, ossa (con funzione di sostegno). Contengono per la maggior parte residui idrofobici, pertanto le proteine fibrose risultano insolubili in acqua. Esempi di proteine fibrose sono: cheratina, collagene, elastina, fibroina.

In figura 3.22 è mostrata la struttura di un capello. La proteina dei capelli è l' α -cheratina, una struttura totalmente ad α -eliche. Un paio di queste eliche si attorcigliano per formare una doppia elica. Queste si combinano poi in strutture di ordine superiore chiamate *protofilamenti* e successivamente *protofibrille*. Circa 4 fogli di protofibrille ($4 \times 2^3 = 32$ eliche di α -cheratina in tutto) si combinano per formare un filamento intermedio. Un capello è una schiera di filamenti intermedi di α -cheratina.

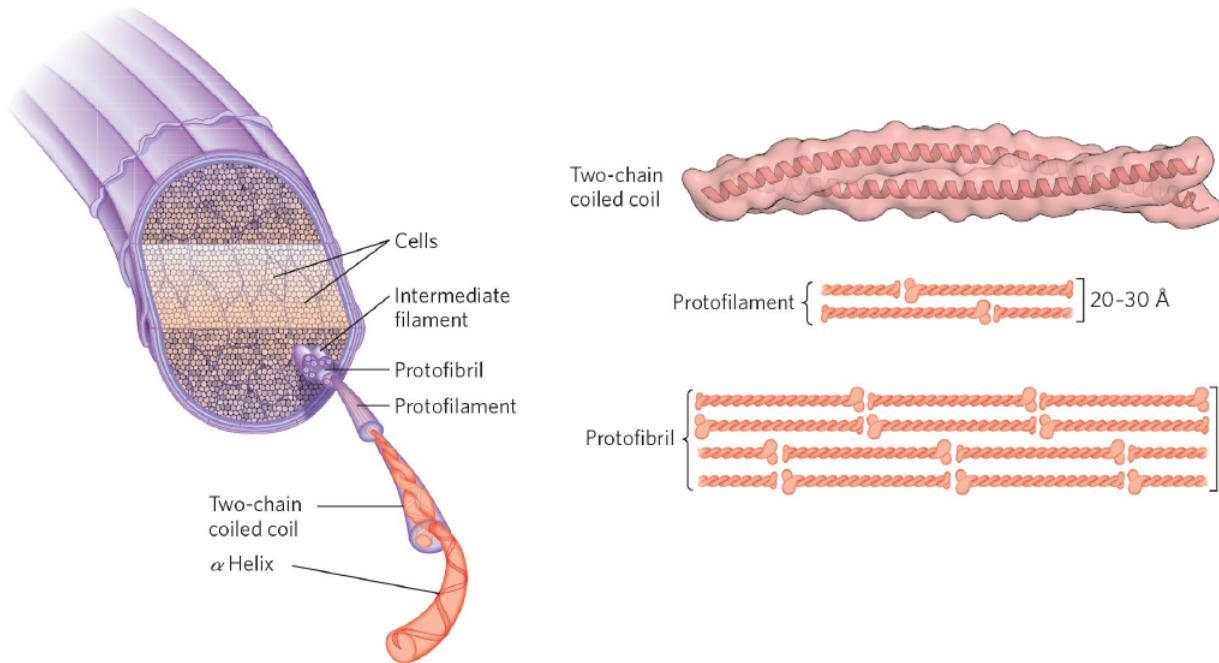


Figura 3.22: Struttura di un capello. Fonte: [57]

Le proteine *globulari* assumono una struttura terziaria e a volte quaternaria. Sono macromolecole compatte di forma all'incirca sferica. Hanno una struttura meno ordinata

rispetto a quelle fibrose. Svolgono funzioni di catalisi, trasporto e regolazione di processi cellulari. Categorie di proteine globulari sono: enzimi, trasportatori di ossigeno e lipidi, alcuni ormoni, recettori di membrana e anticorpi. La struttura è caratterizzata da brevi tratti di α -elica e struttura β , collegate da tratti non organizzati in struttura secondaria. Sono proteine solubili nel citosol.

3.3 Dinamica del ripiegamento

3.3.1 Geometria ed energetica del ripiegamento

Geometria del ripiegamento

Le tante conformazioni possibili della catena polipeptidica sono possibili grazie alla rotazione di essa attorno all'atomo C_α di ogni amminoacido.

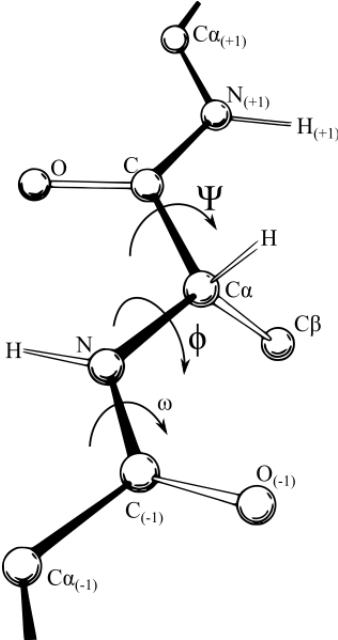


Figura 3.23: Angoli di torsione intorno all'atomo C_α

I tre angoli di torsione principali di un polipeptide sono ϕ , ψ ed ω ; i legami $N-C_\alpha$ e $C_\alpha-C$ sono relativamente liberi nella rotazione (angoli ϕ e ψ), mentre il legame $N-C$ (ω) è fisso dato il carattere di doppio legame parziale del legame peptidico alle temperature fisiologiche.

Data la relativa libertà dei due legami si ha la possibilità di isomeria: le due configurazioni possibili sono *cis* e *trans*. Delle due configurazioni possibili, la *trans* è quella favorita dal punto di vista energetico (minima repulsione sterica), infatti oltre il 99% dei legami peptidici delle proteine naturali hanno configurazione *trans*.

Grafico di Ramachandran

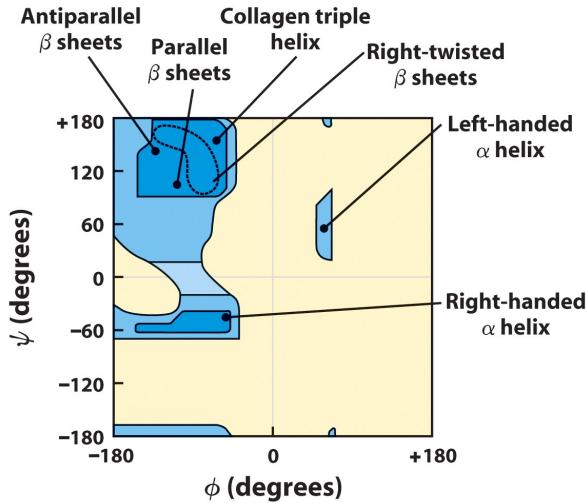


Figura 3.24: Grafico di Ramachandran. Fonte [58]

Nel grafico di Ramachandran sono riportati in ordinata i valori di Ψ ed in ascissa i valori di Φ . Ogni puntino (non presenti in figura) rappresenta la posizione di un residuo. Il grafico deriva da un modello in cui è simulata la variazione di struttura di piccoli polipeptidi e successivamente si cercano conformazioni stabili. Per ciascuna conformazione sono esaminati i contatti tra atomi, trattati come sfere solide di dimensioni determinate dai raggi di van der Waals. Le conformazioni non consentite sono quelle per le quali sono previste collisioni tra sfere. In figura 3.24 l'area gialla corrisponde a conformazioni instabili in cui atomi della catena sono ad una distanza inferiore alla somma dei raggi di van der Waals. Tali regioni sono stericamente non consentite per tutti gli aminoacidi, fatta eccezione per la glicina, priva di catena laterale. Le aree blu indicano conformazioni consentite (β -foglietti e α -eliche destrogire). Le aree azzurre mostrano le regioni consentite riformulando i calcoli con raggi di van der Waals più corti, ovvero consentendo una prossimità atomica maggiore.

Conformazioni ideali (in cui non vi sono interazioni tra catene laterali):

- α -elica: $\Phi = -57^\circ$, $\Psi = -47^\circ$
- β -parallelo: $\Phi = -119^\circ$, $\Psi = +113^\circ$
- β -antiparallelo: $\Phi = -139^\circ$, $\Psi = +135^\circ$

In conformazioni diverse, c'è un limite al numero delle combinazioni possibili dei due angoli di rotazione, perché alcune hanno effetti destabilizzanti a causa delle forze di repulsione tra gli O. Vi sono quindi conformazioni più stabili di altre perché favorite da un punto di vista energetico. Conformazioni proibite sono anche quelle in cui si svilupperebbe una forte repulsione tra le catene laterali.

Energetica del ripiegamento

La termodinamica caratterizza gli stati in natura dalla dipendenza da temperatura, pressione, volume e concentrazione chimica.

L'entropia è la quantità di disordine di un sistema, in altri termini è il numero di possibili stati configurazionali del sistema (Ω).

$$S = K_B \ln \Omega$$

Nei sistemi molecolari cambiamenti di entropia (ΔS) rappresentano cambiamenti nella libertà di movimento di atomi appartenenti sia al soluto che al solvente. L'entalpia è l'energia interna al sistema. Cambiamenti positivi di entalpia nelle macromolecole sono associate alla rottura di interazioni non covalenti favorevoli (si assume che i legami covalenti restino invariati). La formazione di legami covalenti diminuisce l'entalpia del sistema e rilascia calore verso l'ambiente. Formando legami si libera energia mentre per romperli si consuma energia.

$$\Delta H \simeq \Delta E$$

$$E = U + K$$

Dove H è l'entalpia, E l'energia interna al sistema, U l'energia potenziale (circa la somma di tutte le interazioni covalenti e non covalenti, ma non quelle non polari) e K l'energia cinetica (associata ai movimenti atomici indotti termicamente).

L'energia libera di Gibbs (G) è l'*energia utile* sotto temperatura e pressione costante. I processi spontanei raggiungono l'equilibrio decrementando la G del sistema a un minimo.

L'energia libera di Gibbs è associata all'entropia e all'entalpia dalla seguente relazione:

$$\Delta G = \Delta H - T\Delta S$$

La seconda legge della termodinamica afferma che in un sistema isolato i processi spontanei raggiungono l'equilibrio incrementando l'entropia del sistema.

Il ripiegamento delle proteine è un processo spontaneo. Esibisce una grande varietà di percorsi, meccanismi e velocità che dipendono da parametri come la composizione della proteina e le condizioni di ripiegamento. A prescindere dall'esatto meccanismo, il processo di ripiegamento segue sempre la teoria del profilo energetico (a imbuto, *energy landscape theory*), cioè il ripiegamento è sempre accompagnato da una decrescita del numero di conformazioni che possono essere testate dalla proteina, sfuggendo al paradosso di Levinthal (vedi sez. 3.6).

Una proteina che si ripiega deve procedere da uno stato ad alta energia ed alta entropia a uno stato caratterizzato da bassi valori di energia ed entropia; tale nesso è conosciuto come *imbuto di ripiegamento* (folding funnel, vedi fig. 3.25).

Le forze idrofobiche causano la creazione del nucleo idrofobico seppellendo i residui non polari nella struttura nativa, questo vuol dire che l'entropia del solvente acquoso subisce un aumento portando ad una sovra compensazione dell'entropia, ovvero l'entropia del sistema aumenta.

Le proteine non cercano una conformazione fra tutte le possibili conformazioni fino a trovare quella giusta. Piuttosto si ripiegano in una maniera cooperativa in cui ogni step limita ulteriormente le possibilità di ripiegamento degli step seguenti. Il ripiegamento è completato quando la conformazione con più basso livello di energia libera associato alla proteina è trovata. La superficie rugosa del tunnel riflette il fatto che il processo di ripiegamento passi attraverso molti minimi locali separati da barriere da alta-energia.

Una visione più accurata della struttura nativa delle proteine (anche alla luce delle questioni alzate dalle IDP, vedi sez. 3.5) può essere la seguente:

«*la struttura nativa di una proteina è quella conformazione avente minore energia libera tale da mantenere il livello di dinamicità richiesto alla proteina per svolgere la sua funzione biologica^[6]*»

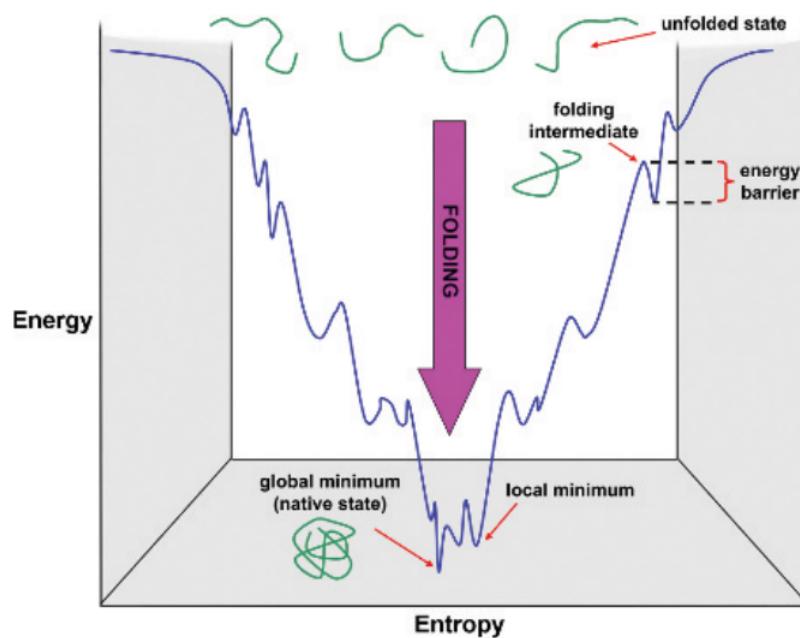


Figura 3.25: Profilo energetico a imbuto del ripiegamento. Fonte [6]

Non c'è un meccanismo di ripiegamento universale, ma una collezione di possibili meccanismi che possono essere usati. La preferenza di una proteina ad usarne uno piuttosto che

un altro può dipendere da vari fattori, uno dei quali sembra essere la struttura secondaria. Ad esempio in proteine dominate da α -eliche il ripiegamento è spesso gerarchico.

La velocità di ripiegamento correla non solo con la dimensione della proteina ma anche con la sua topologia nativa. Le proteine *fast-folding* (ripiegamento su scala temporale di nanosecondi) tendono ad avere grandi proporzioni di elementi secondari locali (α -eliche e giri) laddove quelle *slow-folding* tendono ad avere proporzioni più grandi di elementi globali (β -foglietti).

Termodinamica delle forze idrofobiche

I fattori termodinamici che danno luogo all'effetto idrofobico sono complessi e non del tutto conosciuti. L'effetto idrofobico è visto come una combinazione dell'effetto di idratazione (effetto entropico) e di interazioni di van der Waals tra molecole di soluto (effetto entalpico). Le forze idrofobiche non scaturiscono da classiche interazioni atomo-atomo ma sono un effetto indiretto risultante dalle proprietà del solvente. Per questa ragione è stato difficile conoscere la vastità delle interazioni non polari, sebbene queste siano correlate con la dimensione delle molecole interagenti^[6]:

- per piccole molecole (< 20) atomi di carbonio), l'energia libera delle interazioni non polari corrella con il numero di atomi di carbonio
- per molecole più grandi come le proteine, l'energia libera corrella l'area di superficie (ΔS) della molecola:

$$\Delta G_{np} \approx -0.025\Delta SA$$

ovvero per ogni \AA^2 della molecola interagente si ha un'energia libera di ≈ 25 cal/mol.

Questa relazione dimostra che ogni molecola con un'area di superficie può partecipare in interazioni non polari. Dato che molte molecole includono gruppi polari e hanno un'estesa area di superficie è il bilancio tra i due tipi di interazione (polare e non polare) che determina la tendenza complessiva della molecola ad essere *idrofila* o *idrofobica*.

3.3.2 Ripiegamento assistito

All'interno delle cellule le proteine più piccole si ripiegano indipendentemente, mentre proteine più grandi sono assistite principalmente da complessi chiamati *chaperoni molecolari*. È importante notare che l'assistenza è cinetica in natura: non aggiunge nuove informazioni necessarie alla proteina per ripiegarsi, pertanto il dogma di Anfinsen non viene contraddetto. Ciò che fanno questi complessi è creare un ambiente nel quale le proteine possano ripiegarsi senza "distrazioni" dovute a interazioni con altre entità (ad esempio evitando l'aggregazione con altre proteine) e senza rimanere bloccate in conformazioni intermedie

durante il loro percorso di ripiegamento. In poche parole sono misure di protezione della cellula.

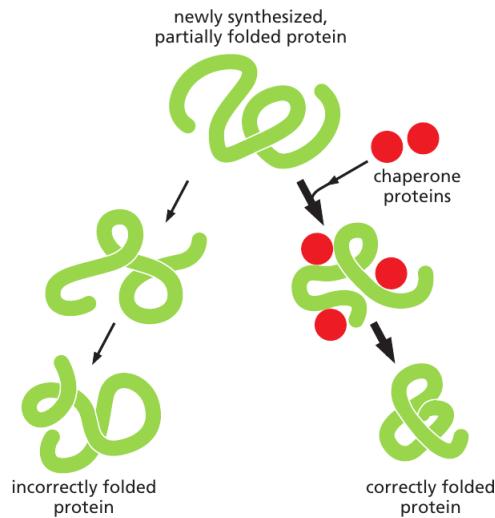


Figura 3.26: Schema della funzione dei chaperoni molecolari. Fonte: [7]

Più in dettaglio i chaperoni molecolari svolgono le seguenti funzioni:

1. assistono il corretto ripiegamento delle catene polipeptidiche (lunghe) appena sintetizzate
2. dirigono l'assemblaggio di complessi multi-enzimatici
3. donano una "seconda chance" a proteine danneggiate favorendone la rinaturazione
4. partecipano nella parziale denaturazione durante il trasporto di proteine attraverso membrane di mitocondri o cloroplasti

Tutti i compartimenti cellulari delle cellule eucariotiche (nucleo, citosol, reticolo endoplasmatico, mitocondri e cloroplasti) hanno il proprio set di chaperoni che assicura un corretto ripiegamento delle proteine. I chaperoni molecolari comprendono diverse famiglie di proteine altamente conservative, tra cui le Hsp (Heat shock protein), proteine espresse in grande quantità sotto condizioni di alto stress, per contrastarne l'effetto denaturante. Queste ultime sono state classificate in base al loro peso molecolare, ad es. Hsp60 dove "60" indica 60kDa. Le Hsp60 vengono chiamate anche *chaperonine* e sono una famiglia di chaperoni molecolari a doppio anello che agiscono da "camera di isolamento" per il ripiegamento di altre proteine^[59], famosa è la chaperonina procariotica GroEL (vedi fig. 3.27), che può essere assunta come modello di riferimento delle chaperonine.

Sebbene i mitocondri (e i cloroplasti) abbiano il loro genoma e creino le loro proteine, la maggior parte delle proteine che questi organelli usano sono codificate dai geni nel nucleo e

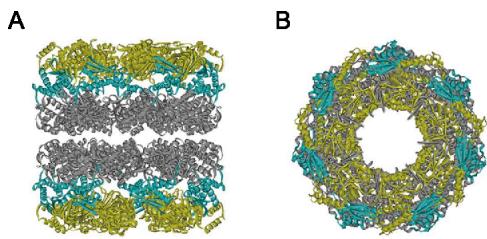


Figura 3.27: Strutture dei complessi GroEL e GroEL-GroES. (B) si può osservare la tipica forma ad anello.
Fonte: [60]

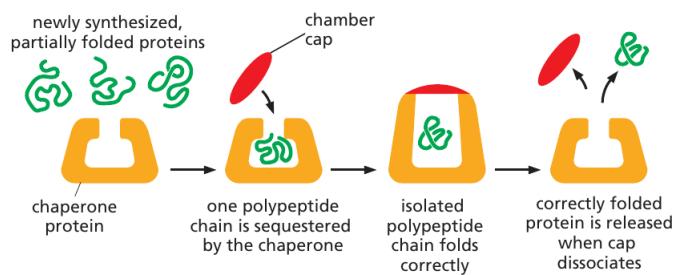


Figura 3.28: Rappresentazione schematica della funzione della camera di isolamento nelle chaperonine. Fonte [7]

importati dal citosol. Ogni proteina viene quindi parzialmente denaturata per effettuare il trasporto. I chaperoni molecolari all'interno di questi organelli aiutano a tirare le proteine attraverso le due membrane e a ripiegarle una volta all'interno^[7].

3.3.3 Misfolding, prioni e malattie

Misfolding

Il *misfolding* è il fenomeno dell'errato ripiegamento di una proteina, ovvero quando una proteina non può raggiungere il suo stato nativo. Ciò può accadere per mutazioni alla sua sequenza amminoacidica (anche per un solo amminoacido differente come nel caso dell'anemia falciforme) o per fattori esterni. Le proteine mal ripiegate tipicamente contengono β -foglietti organizzati in una struttura denominata cross- β , disposizione molto stabile e insolubile, resistente alla proteolisi. Il mal ripiegamento di alcune proteine può innescare ulteriori mal ripiegamenti e la conseguente accumulazione di proteine mal ripiegate in aggregati (od oligomeri) che possono guadagnare tossicità attraverso le interazioni intermolecolari. L'incremento dei livelli di proteine aggregate può portare alla formazione di *amiloidi*, strutture fibrillari formate da deposizioni di materiale proteico insolubile.

Malattie

L'errato ripiegamento delle proteine è alla base quindi di molte patologie umane, definite malattie da misfolding, categorizzabili in due gruppi:

- malattia causata dalla perdita o degradazione della proteina o dall'errato trasporto intracellulare
- malattie causate dall'accumulo, intra od extra-cellulare, di proteine aggregate (ad esempio le malattie da prione)

Molti tipi di tumore diventano chemio-resistenti perché iper-esprimono alcune Hsp, come la Hsp70 e la Hsp90. Le Hsp sono presenti anche in quantità elevatissime nel cervello dei pazienti con malattia di Alzheimer e morbo di Parkinson. Tuttavia si crede che la loro aumentata espressione non sia lesiva di per sé ma rappresenti piuttosto una risposta difensiva agli elevati livelli di stress che caratterizzano queste patologie. Ci sono molti morbi associati a mutazioni nei geni codificanti i chaperoni. Alterazioni genetiche delle chaperonine possono portare a patologie umane che in genere colpiscono molti organi ed apparati contemporaneamente^[61].

Prioni

I **prioni** (acronimo di "proteinaceous infective **only** particle") sono molecole di natura proteica con la capacità di trasmettere la propria forma mal ripiegata a varianti normali della stessa proteina⁸. Il ruolo ipotizzato di una proteina come agente infettivo è in contrasto con tutti gli altri agenti infettivi conosciuti, come i viroidi, virus, batteri, funghi, parassiti: tutti contengono acidi nucleici (DNA, RNA o entrambi) mentre le proteine sono composte di soli amminoacidi.

I prioni formano amiloidi che si accumulano nei tessuti e sono associati a danni di questi e alla morte cellulare. I prioni sono attualmente considerati i più probabili agenti delle encefalopatie spongiformi trasmissibili (TSE) dei mammiferi. Nel *morbo della mucca pazza* (encefalopatia spongiforme bovina), malattia neurologica degenerativa e irreversibile, vi è il ruolo di un prione a causare mal ripiegamenti di alcune proteine native causando la formazione di strutture amiloidi fatali (al microscopio le dense placche fibrose appaiono come buchi, da qui il caratteristico aspetto "a spugna"). Tutte le malattie da prione sono attualmente inguaribili e letali, con un periodo di incubazione che dura generalmente vari anni.

Gli aggregati di prioni sono stabili e questa stabilità strutturale consente loro di essere immuni alla maggior parte dei trattamenti conosciuti. L'organismo infettato non ha modo di degradarli: a differenza di virus e batteri i prioni rimangono intatti anche in presenza di trattamenti come sterilizzazione, forti dosi di radiazioni ionizzanti, uso di formaldeide, varechina, acqua bollente e a differenza delle altre proteine sono resistenti alla maggior parte delle proteasi.

La proteina di cui sono fatti i prioni, *PrP* (protease-resistant-protein, Pr per **prione**, e P per **proteina**), si trova in tutto il corpo, anche negli individui sani, ed è altamente conservata nei mammiferi. Tuttavia, la PrP trovata nel materiale infettante ha una stru-

⁸I prioni sono stati studiati e denominati in questo modo dal premio Nobel per la medicina nel 1997 Stanley Prusiner^[62].

tura diversa. Nell'uomo la PrP^c(cellulare, forma normale) è codificata da un solo gene, PRNP. La PrP^{sc}(scrapie, forma patologica) differisce dalla proteina naturale PrP^c per la conformazione tridimensionale: la PrP^c ha una struttura più aperta contenente 3 segmenti ad α -eliche e pochi β -foglietti; la PrP^{sc} invece ha una struttura più compatta e stabile e presenta un aumento di β -foglietti.

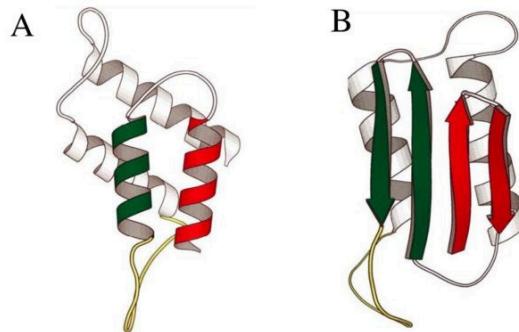


Figura 3.29: (A) Struttura della PrP^c. (B) Struttura della PrP^{sc}. Fonte: [63]

3.3.4 Controllo qualità e apoptosis

Controllo qualità

L'uscita delle proteine dal reticolo endoplasmatico (RE) è controllata per assicurare la qualità delle proteine. Sebbene alcune siano appositamente create e destinate a funzionare nel RE la maggior parte delle proteine che entrano nel RE sono destinate ad altri luoghi. Queste vengono impacchettate nelle vescicole di trasporto e gemmano per fondersi con l'apparato del Golgi. Ma l'uscita dal RE è altamente selettiva: le proteine che falliscono a ripiegarsi nella forma nativa e quelle che non si assemblano correttamente sono attivamente conservate nel RE attraverso i legami con i chaperoni molecolari che risiedono lì. Questi trattengono le proteine nel RE finché non si verifica il corretto ripiegamento o assemblaggio. Se questo non si verifica o fallisce ancora le proteine sono esportate nel citosol dove sono degradate da un *proteasoma*. Le proteine da degradare sono contraddistinte dal loro legame con l'ubiquitina⁹.

Ad esempio gli anticorpi sono composti da 4 catene polipeptidiche che si assemblano in completi anticorpi nel RE. Gli anticorpi parzialmente assemblati sono conservati nel RE finché tutte e 4 le catene non sono pronte. Le molecole di anticorpi che falliscono ad assemblarsi vengono degradate.

⁹Per "la scoperta della degradazione delle proteine mediata da ubiquitina" è stato assegnato il Premio Nobel per la chimica del 2004.

Nonostante l'indubbia utilità di questo meccanismo di controllo a volte questo può rivelarsi dannoso per l'organismo. Ad esempio una classe di mutazioni che causa la *fibrosi cistica*, comune malattia genetica che comporta seri danni polmonari, produce una proteina di trasporto della membrana plasmatica leggermente mal ripiegata. Tuttavia questa potrebbe funzionare normalmente se raggiungesse la membrana plasmatica ma, come viene suggerito da alcuni studi^[64], viene bloccata nel RE e successivamente degradata^[7] (per usare una metafora si può immaginare la situazione di un innocente condannato alla pena di morte). La nota da sottolineare in questo esempio specifico è che tale mutazione (una delle varie classi di mutazioni possibili nella fibrosi cistica) non causa un'inattivazione di una proteina importante ma la proteina attiva è scartata dalle cellule prima che questa possa avere l'opportunità di funzionare.

Proteasomi

I proteasomi sono complessi di *proteasi* (enzima in grado di catalizzare la rottura del legame peptidico delle proteine) che degradano (principalmente) proteine anomale attraverso reazioni di *proteolisi*. Sono presenti nelle cellule di tutti gli eucarioti e procarioti. La struttura e la funzione di questi complessi è altamente conservata.

A causa del ruolo dei proteasomi nella regolazione del ciclo cellulare e dell'apoptosi¹⁰, sono oggi un bersaglio rilevante nelle terapie antitumorali.

Le proteasi tuttavia possono anche essere utilizzate dai virus in maniera inversa, ovvero per produrre proteine, si parla di *proteasi virali*^[65]. Le proteine del *core virale* vengono prodotte in lunghi filamenti polipeptidici, ognuno dei quali contiene varie copie di proteine del core. Le proteasi riconoscono specifiche sequenze poste fra una futura proteina e l'altra all'interno di queste catene ed effettuano dei tagli: le singole proteine vengono separate e possono quindi progredire nella loro maturazione, andando a costruire un core virale.

Per fermare tale meccanismo sono stati progettati farmaci inibitori, specialmente nella terapia antiretrovirale (HIV, epatite C), che interferiscono con il ciclo replicativo del virus HIV proprio bloccando le proteasi. I filamenti polipeptidici non vengono più scissi se si inibisce l'attività di questi enzimi, non vengono più create le proteine del core e quindi gli stessi core. Senza core la replicazione virale si ferma.

¹⁰Il termine *apoptosi* indica una forma di morte cellulare programmata (un'auto-distruzione). Al contrario della necrosi, che è una forma di morte cellulare risultante da un acuto stress o trauma cellulare, l'apoptosi è portata avanti in modo ordinato e regolato, richiede consumo di energia (ATP) e generalmente porta a un vantaggio durante il ciclo vitale dell'organismo (è infatti chiamata da alcuni morte altruista o morte pulita). Durante il suo sviluppo, ad esempio, l'embrione umano presenta gli abbozzi di mani e piedi "palmati": affinché le dita si differenzino, è necessario che le cellule che costituiscono le membrane interdigitali muoiano.

Unfolded protein response e Apoptosi

La dimensione del RE è controllato dalla "richiesta" per il ripiegamento delle proteine. Il meccanismo di controllo nel RE, eseguito dai chaperoni molecolari, può essere sovrapposto. Quando succede le proteine mal ripiegate si accumulano nel RE. Se l'accumulo è abbastanza grande, questo innesca un complesso programma chiamato *unfolded protein response* (UPR). Questo programma incita la cellula a produrre più RE, inclusi più chaperoni molecolari, e altre proteine riguardanti il controllo qualità. L'UPR permette alla cellula di regolare la dimensione del RE per gestire propriamente il volume delle proteine in entrata. In alcuni casi tuttavia anche un RE espanso non riesce a gestire la richiesta e l'UPR indirizza la cellula verso l'*apoptosi*.

Una situazione del genere può avvenire negli adulti in cui insorge il diabete. I tessuti diventano gradualmente resistenti all'effetto dell'insulina. Per compensare questa resistenza le cellule che secernono insulina nel pancreas ne producono ancora di più. Si arriva infine alla situazione in cui il loro RE arriva ad una capacità massima e viene innescato l'UPR e di conseguenza la morte cellulare. Col tempo sempre più cellule secernenti insulina sono eliminate e la richiesta per quelle sopravvissute aumenta rendendole sempre più vulnerabili a questo meccanismo, esacerbando ulteriormente la malattia^[7].

3.4 Studio sperimentale delle proteine

Come vengono studiate le proteine

Comprendere come una particolare proteina funzioni richiede dettagli strutturali e analisi biochimiche: entrambe hanno bisogno di una grande quantità di proteine pure. Isolare però un singolo tipo di proteina dalle migliaia presenti in una cellula non è un compito semplice. Per molti anni le proteine sono state purificate direttamente dai tessuti nei quali esse erano abbondanti. Ciò comporta, oltre alla necessità fisica di procurarsi i tessuti, dover riconoscere le proteine da studiare. Queste procedure richiedono settimane e forniscono solamente pochi milligrammi di proteina pura. Oggi le proteine sono generalmente isolate da cellule coltivate in laboratorio e spesso queste sono modificate geneticamente al fine di produrre grandi quantità di una data proteina.

Il primo step in una procedura di purificazione consiste nel rompere i tessuti e le cellule in modo da far loro rilasciare il contenuto (*omogenizzazione*), chiamato *estratto* (o omogenato cellulare). Ci sono vari meccanismi, come mostrato in fig. 3.30.

Segue poi una procedura di frazionamento iniziale, tipicamente per *centrifugazione*, per separare l'omogenato in differenti parti. Per raggruppare poi le classi di molecole di

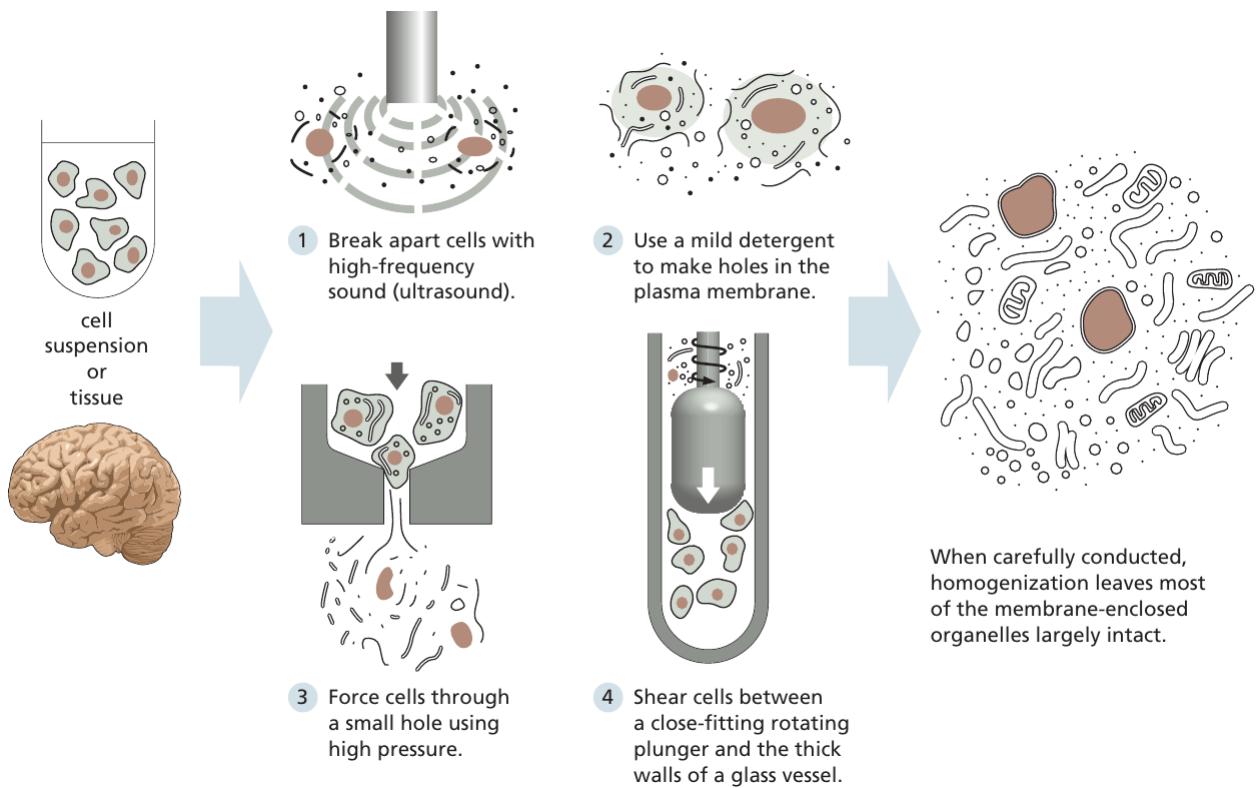


Figura 3.30: Omogenizzazione attraverso 4 diversi meccanismi, ognuno dei quali finalizzato a rompere le membrane plasmatiche. Fonte [7]

interesse si possono usare tecniche di *cromatografia* o *elettroforesi*. Una forma efficiente di cromatografia è quella per *affinità* nella quale vengono utilizzati degli anticorpi specifici per la proteina di interesse. Nel secondo metodo un insieme di proteine viene immerso in un gel e soggetto ad un campo elettrico: le proteine migreranno nel gel a differenti velocità, a seconda del loro peso molecolare e della loro carica. Proteine simili migreranno a velocità simili, pertanto potranno essere visualizzate bande o punti di aggregazione di proteine.

Una volta separate le proteine di interesse è possibile studiarne la struttura. Per quanto riguarda la struttura primaria, la prima proteina sequenziata è stata l'*insulina* nel 1955, attraverso una procedura chimica diretta. La proteina veniva prima scomposta da una determinata proteasi e successivamente ogni amminoacido, in ogni frammento, veniva determinato sperimentalmente. Un metodo molto più veloce è la *spettrometria di massa*, almeno per organismi di cui sia stato completamente sequenziato il genoma. Questa tecnica determina l'esatta massa di ogni frammento in una proteina purificata, consentendo l'identificazione della proteina nei database di sequenze genomiche.

Per eseguire la *spettrometria di massa* la proteina viene "digerita" dall'enzima *tripsina* e frammentata in peptidi o singoli amminoacidi. Questi vengono scaldati con un laser, il

che li renderà carichi e li farà evaporare. Viene poi usato un potente campo elettrico per far volare gli ioni peptidici verso un misuratore: il tempo che impiegano per arrivare è legato alla loro massa e carica (più massa = più lenti, più carichi = più veloci). L'insieme delle precise masse dei frammenti serve come "impronta digitale" (*peptide mass fingerprinting*, PMF) che verrà usata per identificare la proteina codificata dall'organismo (e i suoi geni corrispondenti) dai database il cui profilo (massa teorica) corrisponde a questa impronta peptidica.

Determinazione sperimentale delle strutture

La determinazione sperimentale della struttura delle proteine ha vissuto dei progressi significativi col passare degli anni ed è di grande importanza per i metodi computazionali di PSP, consentendo ai metodi *data-based* di affinare le loro predizioni.

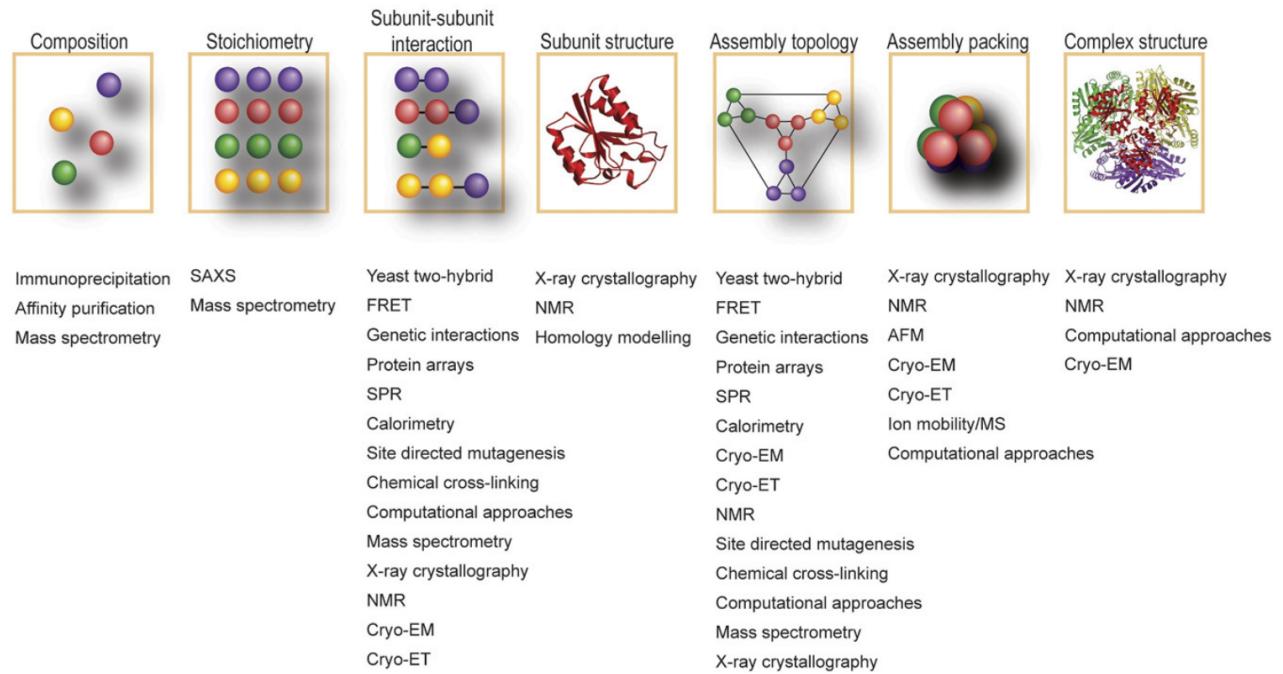


Figura 3.31: Differenti livelli di informazione ottenuta lungo il percorso della determinazione della struttura di macromolecole. AFM: atomic force microscopy; Cryo-ET: cryo-electron tomography; EM: electron microscopy; FRET: fluorescence resonance energy-transfer; NMR: nuclear magnetic resonance; SAXS: small-angle X-ray scattering; SPR: surface plasmon resonance. Fonte[66]

I metodi per la determinazione sperimentale della struttura delle proteine possono essere divisi in due gruppi:

- metodi *indiretti*, ovvero l'osservazione della proteina è possibile solo dopo sofisticate manipolazioni dei dati ottenuti:

- metodi per *diffrazione*, che si basano sulla diffrazione o sulla dispersione di particelle subatomiche o onde elettromagnetiche da parte della proteina
- metodi per *spettroscopia*, i quali si affidano all'eccitazione e susseguente rilassamento degli atomi della proteina in risposta alla radiazione elettromagnetica
- metodi *diretti*, in cui l'osservazione della proteina è diretta; al momento è possibile con la microscopia crioelettronica (Cryo-EM)

Vengono utilizzate principalmente 3 tecniche per generare informazioni strutturali sulle proteine a risoluzione atomica: *X-ray crystallography*, *nuclear magnetic resonance (NMR) spectroscopy* ed *electron microscopy*¹¹. Una volta ottenute proteine pure, queste devono essere o cristallizzate (cristallografia a raggi X), o piazzate in speciali solventi (spettroscopia NMR) o congelate (microscopia elettronica).

Il metodo di diffrazione più comune, e più anziano, è la *cristallografia a raggi X*. Produce strutture tridimensionali con la più alta risoluzione ma presenta alcune gravi carenze, la più significativa è la necessità di cristallizzare la proteina studiata. La cristallizzazione è un processo lungo e difficile e produce anche strutture di proteine al di fuori del loro ambiente. Queste strutture sono prive di qualsiasi proprietà dinamica e (raramente) possono risultare deformate. Pertanto non tutte le parti di una proteina (ad es. le parti più mobili) possono essere viste con strutture a raggi-X e di conseguenza queste regioni possono essere ignorate o aperte a interpretazioni.

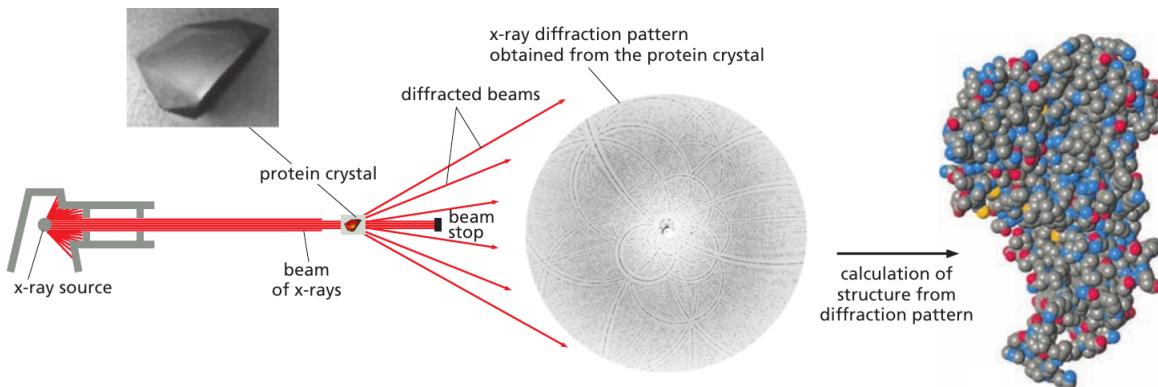


Figura 3.32: Processo di determinazione della struttura dell'enzima RuBisCO tramite cristallografia a raggi-X. Fonte[7]

I piccoli cristalli di proteine misurano meno di 1mm e sono esposti ad un'intensa esposizione ai raggi-X (i quali hanno una lunghezza d'onda pari a quella di un atomo, 1 – 2Å). I raggi X sono dispersi o diffatti dagli atomi proteici nel cristallo. Il modello

¹¹La traduzione italiana sarebbe: cristallografia a raggi-X, spettroscopia a risonanza magnetica nucleare e microscopia elettronica.

di diffrazione che ne deriva appare tipicamente come decine di migliaia di minuscoli punti disposti in complessi schemi circolari. Questi modelli di diffrazione sono registrati su una fotocamera a raggi X digitale. La posizione dei punti di diffrazione (insieme ad altre informazioni), sono effettivamente sufficienti per compiere una computazione della mappa della densità elettronica di tutti gli atomi pesanti (carbonio, azoto, ossigeno, zolfo) nella proteina di diffrazione. Da questa mappa, i cristallografi determinano le coordinate x,y,z di tutti gli atomi usando la sequenza nota della proteina. Si noti che, nella cristallografia a raggi X, anche se il pattern di diffrazione deriva da milioni di proteine contenute nel cristallo, il risultato è una struttura per una singola proteina "media".

La prima struttura determinata con questa tecnica risale al 1958; risolvere una struttura negli anni '70 richiedeva anche 6-7 anni mentre oggi è a volte possibile in soli 6-7 giorni. Il 90% delle strutture oggi determinate deriva dalla cristallografia a raggi-X^[30].

Altri metodi per diffrazione sono: *small-angle X-ray scattering* (SAXS), *neutron scattering*, *electron crystallography*. Il metodo SAXS produce strutture a risoluzione inferiore rispetto alla cristallografia a raggi X. Tuttavia, le strutture SAXS sono molto utili nell'impostare vincoli posizionali per complessi proteici di grandi dimensioni, grazie ai quali è possibile "modellare" strutture a raggi X ad alta risoluzione.

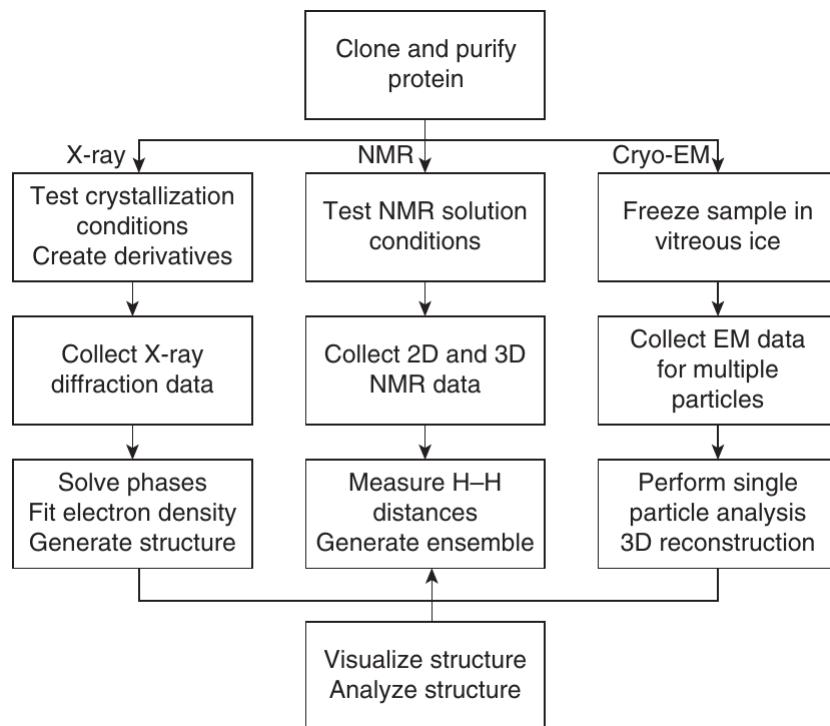


Figura 3.33: Diagramma di flusso dei principali step usati la preparazione e soluzione sperimentale della struttura 3D delle proteine. Fonte[30]

Il principale metodo spettroscopico utilizzato per la determinazione della struttura proteica è l'NMR. Questo metodo si basa sull'eccitazione dei nuclei atomici sotto un forte campo magnetico e sul loro successivo rilassamento. Viene misurato come i nuclei atomici (ad es. dell'idrogeno o dell'isotopo carbonio o azoto) assorbano la radiazione; ciò permette di determinare quanto magnetismo nucleare è trasferito da un atomo all'altro. Questo approccio consente agli scienziati di trattare la proteina nel suo ambiente naturale (soluzione, membrana) e fornisce importanti informazioni sulla sua dinamica. L'NMR è un metodo più recente della cristallografia a raggi-X: la prima struttura determinata risale al 1983. Tuttavia questo metodo ha un limite superiore alla dimensioni di molecole studiabili (40kDa) e non può studiare proteine di membrana. L'EPR, un metodo simile, si basa sull'eccitazione e sul rilassamento degli elettroni attorno agli atomi della proteina e richiede la marcatura dei residui proteici con etichette paramagnetiche.

Cryo-EM

La microscopia crioelettronica (cryo-EM) è un metodo diretto: è possibile osservare direttamente macromolecole. È una versione avanzata di microscopio elettronico, il quale fu inventato negli anni '30. Questi microscopi usano raggi di elettroni piuttosto che di luce (la lunghezza d'onda degli elettroni è molto più corta di quella della luce). Durante la metà degli anni '70 è nata la cryo-EM: l'idea è stata quella di congelare i campioni per preservarne la struttura naturale e ridurre i danni causati dai raggi di elettroni.

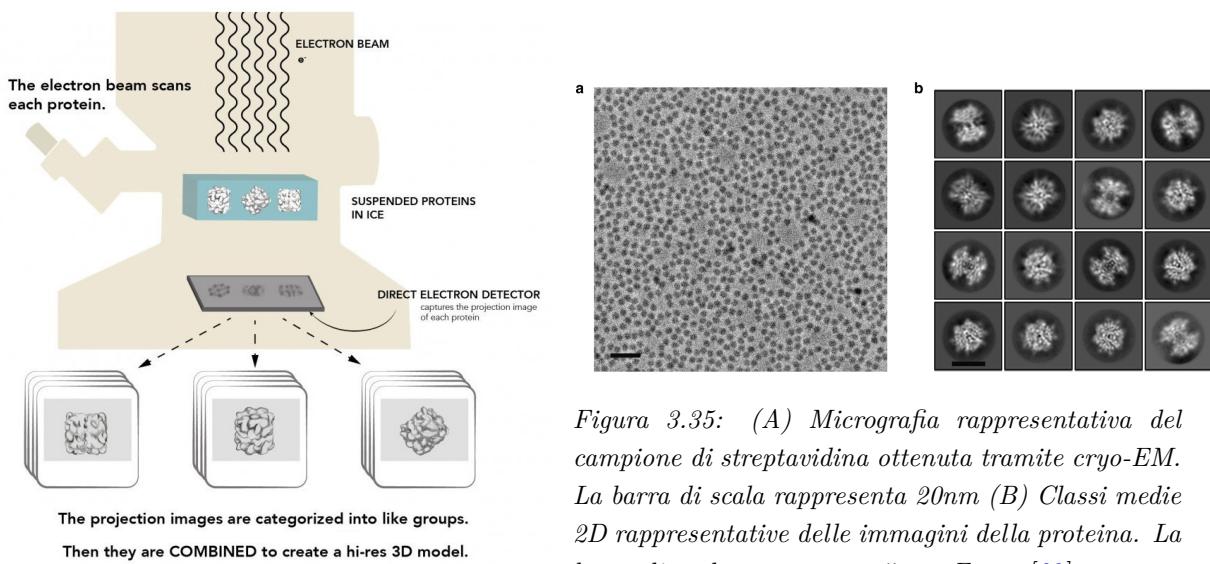


Figura 3.35: (A) Micrografia rappresentativa del campione di streptavidina ottenuta tramite cryo-EM. La barra di scala rappresenta 20nm (B) Classi medie 2D rappresentative delle immagini della proteina. La barra di scala rappresenta 5nm. Fonte [68]

Figura 3.34: Funzionamento schematico di un microscopio crioelettronico. Fonte[67]

Nella cryo-EM una goccia di acqua contenente pure proteine è inserita in una piccola griglia per EM immersa in una vasca di etano liquido a -180°C. I campioni di proteine vengono congelati velocemente (creando ghiaccio vitreo): questo assicura che le circostanti molecole d'acqua non abbiano tempo per formare cristalli di ghiaccio (che deformerebbero la forma della proteina). I campioni sono esaminati (ancora ghiacciati) da un microscopio a trasmissione elettronica (TEM) e sottoposti quindi a forti raggi di elettroni. Un rilevatore di elettroni cattura le "immagini" proiettate delle molecole e, data l'automazione odierna di simili meccanismi, vengono effettuate migliaia di micrografia per catturare il maggior numero di dettagli possibile delle molecole. Ogni micrografia conterrà centinaia di migliaia di molecole singole, ognuna orientata casualmente.

Successivamente vi è lo step di image processing: le immagini proiettate vengono categorizzate in gruppi e allineate per poi essere sovrapposte in modo da calcolare un'immagine media per ogni gruppo.

La preparazione è quindi molto più semplice rispetto alla cristallografia a raggi-X e le strutture somigliano maggiormente a quelle viste nel normale ambiente acquoso della cellula. La cryo-EM è limitata nelle dimensioni: c'è un limite inferiore che di anno in anno si abbassa (ad. es 39kDa nel 2019^[68]). Non sono un problema invece grandi proteine (anche maggiori di 100kDa). Oggi la cryo-EM può generare immagini 3D a risoluzione quasi atomica di virus e complesse macromolecole, come i ribosomi. È possibile utilizzare la cryo-EM anche per le proteine di membrana.

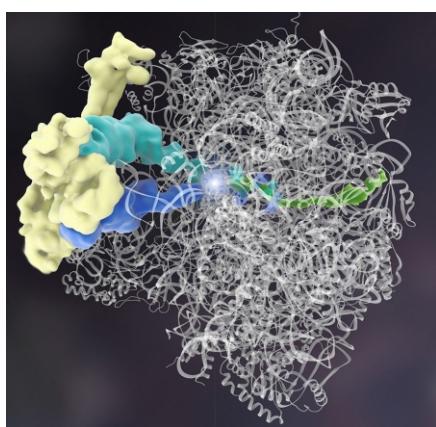


Figura 3.36: Complesso di controllo qualità dei ribosomi, basato su dati di cryo-EM. Fonte: [69]

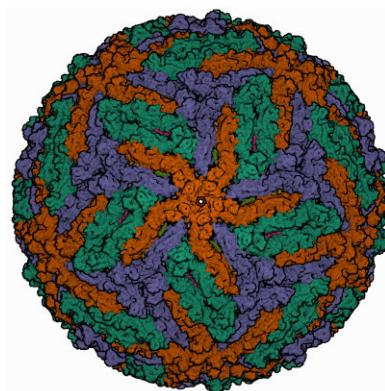


Figura 3.37: La struttura del virus Zika ottenuta tramite cryo-EM. La macromolecola ha un peso di 190kDa ed è composta da 11.000 atomi. Risoluzione di 3.8Å. Fonte [70]

È solo da pochi anni che il metodo ha fatto un grande passo in avanti, grazie ad avanzamenti nel rivelatore e nei software di *image processing*. Nel 2017 è stato assegnato

il premio Nobel per la chimica per aver contribuito a sviluppare tale metodologia¹². Si sta considerando l'adozione della microscopia crioelettronica come di una rivoluzione nel campo della biologia strutturale^{[71],[72]}. La crescita nel numero di strutture determinate è stata lenta inizialmente a causa della scarsa adozione del metodo, ma da quando si è vista la possibilità di produrre mappe dettagliate per macromolecole come i ribosomi la situazione è cambiata (vedi le figure 3.38 e 3.39). Circa l'1% delle strutture è stata determinata con questa tecnica ma i rapidi avanzamenti degli strumenti sia hardware che software potrebbero favorire la rivoluzione di cui si parla.

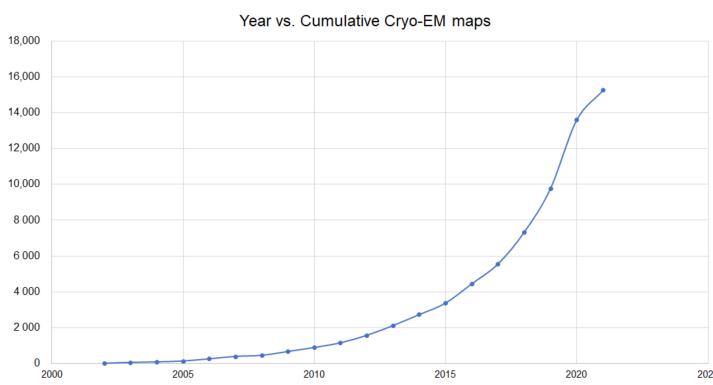


Figura 3.38: Crescita delle mappe elettroniche rilasciate nell'EMDB. Fonte [73]



Figura 3.39: Grafico della risoluzione di strutture risolte tramite cryo-EM. Fonte [71]

3.5 Sfide al dogma di Anfinsen: IDP e fold switching

Il ripiegamento delle proteine in una cellula è un processo molto complesso che riguarda il trasporto di nuove proteine sintetizzate ad appropriati compartimenti cellulari attraverso targeting, misfolding, stati dispiegati temporanei, modifiche post-traduzione, controllo qualità, aggregazione in complessi, facilitazione dei chaperoni molecolari. Come già spiegato, l'aiuto dei chaperoni molecolari non sfida il dogma di Anfinsen, in quanto non influenza la struttura nativa della proteina.

La struttura di alcune proteine è difficile da determinare per una semplice ragione: un crescente numero di ricerche biochimiche ha rivelato che un numero significativo di proteine, o regioni di proteine, non hanno una distinta struttura 3D, non la hanno finché non interagiscono con la molecola target oppure cambiano struttura nativa. La loro flessibilità e struttura indefinita è importante per la loro funzione, che potrebbe richiedere legami con differenti target in tempi diversi.

¹²A Richard Henderson, Jacques Dubochet e Joachim Frank, "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution".

Intrinsically disordered proteins

Una proteina intrinsecamente disordinata (IDP) è una proteina, o una regione di essa, a cui manca una struttura terziaria fissa od ordinata. Le IDP sono comunemente riconosciute come regioni mancanti di densità elettronica in strutture di proteine determinate a cristallografia a raggi X. Molte IDP possono adottare una struttura tridimensionale stabile dopo essersi legate ad altre macromolecole, passando per transizioni disordine-ordine e perciò risultare strutturate per un certo periodo di tempo e non strutturate per altro. Ci sono però anche IDP che svolgono la loro funzione senza assumere mai una forma ordinata attraverso la loro esistenza. Nonostante la loro mancanza di struttura stabile le IDP risultano essere una classe importante e grande di proteine.

Metodi bioinformatici suggeriscono che ?? la misura in le IDP sono prevalenti nel genoma di un organismo correla con la complessità dell'organismo in questione. Ciò implica che queste proteine giochino ruoli complessi. In accordo a studi bioinformatici su interi proteomi si stima che nei mammiferi circa il 25% di tutte le proteine siano IDP e circa il 75% di tutte le proteine di segnalazione (circa il 50% di interi proteomi) contengano lunghe regioni disordinate^[6].

Per tenere conto delle IDP in un modo che riflette la loro prevalenza sono stati aperti diversi database liberamente disponibili ad esempio DisProt e MoBiDB.

Fold switching proteins

Alcune proteine hanno multiple strutture native: possono cambiare la loro forma, rimodelando anche le loro strutture secondarie, in base a fattori esterni. Ad esempio il complesso proteico KaiB cambia ripiegamento durante la giornata agendo da orologio per i cianobatteri. Si possono immaginare le proteine *fold switching* come una sorta di transformer¹³ dove in un caso la proteina è come un robot che fa una cosa e in un altro caso, in risposta a cambiamenti ambientali, diventa un'automobile e fa qualcos'altro. Il cambio tra strutture alternative è guidato da interazioni della proteina con piccoli ligandi o altre proteine, da modificazioni chimiche (es. fosforilazione) o da cambiamenti nelle condizioni ambientali (temperatura, pH, potenziale di membrana). Ogni struttura alternativa può o corrispondere al minimo globale di energia libera della proteina in certe condizioni o essere cinematicamente intrappolata in un minimo locale di energia libera^[76] circondato da alte barriere energetiche.

Le proteine *fold switching* (FS) differiscono dalle IDP^[75]:

¹³La metafora è di Lauren Porter^[74], autrice di L. L. Porter e L. L. Looger, “Extant fold-switching proteins are widespread,” *Proceedings of the National Academy of Sciences*, vol. 115, n. 23, pp. 5968–5973, 2018.

- le FS richiedono che entrambe le loro conformazioni siano determinate, le IDP sono regioni non determinate
- le IDP sono caratterizzate da sequenze amminoacidiche caratteristiche mentre le FS sono libere da questo vincolo
- le IDP non si ripiegano cooperativamente in isolamento mentre le FS si ripiegano sia cooperativamente che indipendentemente

Per queste ragioni si può affermare che le FS non sono IDP, piuttosto sono un sottoinsieme delle proteine globulari le cui strutture stabili cambiano drasticamente in risposta al loro ambiente. I vantaggi di una proteina FS sono legati alla sua bifunzionalità:

- può affrontare velocemente richieste biologiche ovviando al bisogno di risorse cellulari aggiuntive per trascrivere e tradurre un'altra proteina. Un esempio è RfaH: funziona sia da fattore di trascrizione che di traduzione
- regolazione di inattività, può essere bloccata in uno stato di attività o inattività finché non viene innescato uno specifico segnale

È stato stimato che una percentuale tra lo 0.5 e il 4% delle proteine nel PDB cambi ripiegamento^[75].

3.5.1 Considerazioni epistemologiche

La scoperta delle IDP e delle proteine FS ha creato una spaccatura nel paradigma della struttura rigida delle proteine, secondo il quale la struttura deve essere fissa al fine di compiere la propria funzione biologica. È interessante ripercorrere velocemente la storia di questo paradigma attraverso un breve excursus storico per poterne mettere in luce le relazioni epistemologiche soggiacenti.

Nel 1894 Fischer ha proposto una metafora di *chiave e serratura* per spiegare come fosse possibile un effetto chimico tra un enzima e un glucoside:

«Per usare una metafora, vorrei dire che l'enzima e il glucoside devono adattarsi l'uno all'altro come una serratura e una chiave per esercitare un effetto chimico l'uno sull'altro^{[77][78]}»

Nel 1936 Mirsky e Pauling hanno raccolto una quantità di informazioni sufficiente per concludere:

«attribuiamo le specifiche proprietà caratteristiche delle proteine native alle loro uniche e definite configurazioni. Consideriamo le proteine denaturate essere caratterizzate dall'assenza di un'unica definita configurazione^[79]».

Né Mirsky, né Pauling né Hsien Wu (probabilmente il primo a proporre il paradigma struttura unica-funzione) citarono il lavoro di Fischer ma la sua metafora avrebbe supportato pienamente le loro tesi. Perciò ancora prima dell'esperimento di Anfinsen e delle

risoluzioni atomiche delle strutture, una specifica e ben ordinata forma tridimensionale era stata accettata come prerequisito essenziale per la funzione di una proteina.^[78]. Le prime strutture proteiche sono state determinate attraverso la cristallografia negli anni '50, dando ancora più credito all'ipotesi che una struttura fissa fosse necessaria per adempiere la funzione biologica. Queste pubblicazioni hanno contribuito a solidificare il dogma centrale della biologia molecolare. Nonostante questo già nel 1950 si parlava di molteplicità di strutture per una proteina, in particolare Karush^[80] sull'albumina del siero bovino, inferendo che le interazioni proteina-ligando stabilizzassero il membro più adatto da un insieme di strutture in equilibrio, chiamando questo fenomeno *adattabilità configurazionale*. Successivamente si può ricordare il paradosso di Levinthal negli anni '60, per arrivare negli anni '70 al dogma di Anfinsen, il cui paradigma (sequenza amminoacidica → struttura tridimensionale → funzione) è messo in discussione da evidenze di mancanza di generalità. Tuttavia quel sistema di pensiero ha preso piede ed ha pervaso quasi tutti i lavori e teorie successive. Al tempo dell'esperimento di Anfinsen e delle prime risoluzioni atomiche della mioglobina e del lisozima il prerequisito di una forma tridimensionale fissa necessaria per la funzione della proteina era già accettato. La successiva valanga di migliaia di strutture determinate sperimentalmente ha contribuito ad affossare paradigmi di pensiero alternativi.

La scoperta di segmenti intrinsecamente disordinati è stata compiuta nel 1978, quando ancora erano disponibili le strutture di sole 20 proteine. Alcuni segmenti di proteine non fornivano alcuna densità elettronica distinguibile nonostante fossero essenziali per il funzionamento. Una ragione comune è che gli atomi di quel segmento sono disordinati, ovvero la loro posizione cambia. Con l'avvento della spettroscopia NMR, sempre in quegli anni, si è arrivati a identificare intere proteine disordinate e dato che questo metodo è più preciso, la riscoperta di disordine nativo ha avuto un impatto significativo.

Fino al 2000^[81] queste idee non sono apparse nei libri di biochimica, nonostante in 50 anni fossero state pubblicate centinaia di paper sull'importanza della flessibilità e del disordine nelle strutture proteiche.

Può risultare interessante confrontare quanto accaduto con le analisi del microbiologo ed epistemologo Ludwik Fleck^[82]. Nel suo libro del 1935, *Genesi e sviluppo di un fatto scientifico*, affronta il problema della conoscenza scientifica e analizza il caso dell'evoluzione del concetto della malattia sifilide, mostrando come questo si è modificato nel tempo. Senza entrare nel merito di quell'analisi, l'epistemologo ha provato a tracciare delle linee generali su come un fatto scientifico possa svilupparsi.

Analizzando le epoche di un concetto, Fleck si esprime affermando:
«Molte teorie, per esempio, hanno due epoche nella loro vita: esse attraversano prima un'epoca classica, in cui tutto si accorda in maniera impressionante, poi una seconda

epoca, nel corso della quale si presentano solo delle eccezioni.»

Non è difficile trovare le somiglianze con il caso della struttura unica per la funzione di una proteina. Inizialmente le eccezioni sono passate inosservate, tutto si accordava perfettamente all'idea di Fischer, Pauling e poi di Anfinsen. Le eccezioni, come quella di Karush, si presentavano ma il paradigma dominante non ne subiva effetti.

Secondo Fleck, per garantire la persistenza dei sistemi d'opinione quando si presentano eccezioni:

«una contraddizione al sistema appare impensabile. Ciò che non si accorda con il sistema: non viene notato, oppure viene tacito anche se noto, oppure si fa in modo di spiegarlo, con laboriosi sforzi, come non contraddittorio rispetto al sistema: si notano, si descrivono o persino si inventano fatti che corrispondono alla concezione dominante, che cioè ne costituiscono per così dire la realizzazione»

Uno dei concetti più importanti nell'opera di Fleck è il concetto di *collettivo di pensiero e stile di pensiero*:

«Se definiamo il termine collettivo di pensiero come "la comunità degli uomini che hanno fra loro un contatto intellettuale e che si scambiano idee influenzandosi reciprocamente, noi veniamo in possesso, con questo concetto, di ciò che rappresenta lo sviluppo storico di un ambito del pensiero, di un determinato patrimonio di conoscenza e di cultura e quindi, di un determinato stile di pensiero.»

E secondo il microbiologo la condizione dell'individuo (scienziato) nei confronti dello stile di pensiero è di subordinazione inconsapevole:

«Anche se il collettivo consiste di individui, esso non è la loro semplice somma. L'individuo non ha mai - o quasi mai - la coscienza dello stile di pensiero collettivo, che quasi sempre esercita una costrizione incondizionata sul suo pensiero e che è semplicemente impensabile poter contraddirlo.»

«Viene a anche a mettersi in luce che molte idee arrivano a manifestarsi prima che ne risaltino le basi razionali e, anzi, in modo completamente indipendente da queste ultime.»
L'idea di Fischer (poi di Pauling) potrebbe avere influenzato il collettivo di pensiero al punto tale che la determinazione delle prime strutture a risoluzione atomica non potesse che dare come risposta una conferma di quelle idee. Esperimento ed esperienza non vivono entrambe nel campo dell'oggettività:

«se l'esperimento può essere interpretato come una pura e semplice domanda e risposta, l'esperienza deve essere invece intesa già come una condizione complessa, frutto di un processo di educazione che si fonda sull'interazione fra chi conosce, ciò che è conosciuto e ciò che deve essere conosciuto.»

La mancata consapevolezza di far parte di un collettivo di pensiero può purtroppo causare l'illusione che esista un nesso logico fra prove e concezioni, ma Fleck ammonisce: «*le prove si adattano alle concezioni altrettanto spesso quanto le concezioni si conformano alle prove*»

Secondo Fleck la conoscenza è sempre un processo sociale:

«"*questo libro è più voluminoso*" è incompleta. Sarebbe corretta se si aggiungesse "*di quel libro*"; [...] la frase «*qualcuno conosce qualcosa*» richiede un'aggiunta, ad es. "*sulla base di un determinato patrimonio di conoscenza*" o meglio "*come membro di un determinato ambiente culturale*" o ancora "*in un determinato collettivo di pensiero*"»

Mettendo insieme le diverse constatazioni riportate si può fare un'analisi, senza alcuna presunzione di correttezza e come semplice esercizio, del dogma di Anfinsen. Secondo Fleck una formulazione corretta sulle sue scoperte potrebbe essere: "Anfinsen propose, in conformità con le opinioni del suo tempo sul ripiegamento, denaturazione e rinaturazione delle proteine, di vedere nella sequenza amminoacidica la base necessaria e sufficiente per la determinazione della sua struttura tridimensionale nativa. Propone inoltre, sempre sulla base del collettivo di pensiero in cui era immerso, di considerare la struttura nativa di una proteina come quella struttura unica, stabile e cinematicamente accessibile avente minima energia libera". Da questo esercizio si può osservare, essendo noi immersi in un collettivo di pensiero differente, che lo stile di pensiero di Anfinsen era probabilmente ancorato alle idee di Pauling secondo il quale la conformazione stabile delle proteine era una e una soltanto.

3.6 Il problema del Protein Folding

Il problema del protein folding è la questione di *come* una sequenza amminoacidica determini la struttura atomica tridimensionale. Il processo del ripiegamento proteico non è così semplice, la maggior parte delle proteine probabilmente passa attraverso strutture intermedie sulla via per raggiungere la struttura nativa, e il semplice osservare la struttura finale non rivela i passaggi del ripiegamento richiesti per raggiungere quella forma. Il problema del protein folding consiste di 3 puzzle strettamente correlati^[52]:

- *folding code*: la questione termodinamica di quale bilancio delle forze interatomiche determini la struttura della proteina a partire da una data sequenza amminoacidica
- *folding process*: la questione cinetica di quali percorsi alcune proteine usino per ripiegarsi così velocemente
- *protein structure prediction*: si può predire la struttura nativa di una proteina dalla sua sequenza amminoacidica? In altre parole il problema computazionale di come predire la struttura nativa di una proteina dalla sua sequenza amminoacidica

Il problema del protein folding, come si può immaginare, è considerato uno dei problemi più impegnativi degli ultimi 50 anni in biochimica, ed è stato fatto riferimento alla predizione della struttura delle proteine come al *santo Graal* della biochimica computazionale^[83]. Sfortunatamente esso è stato anche rivendicato come uno dei problemi di ottimizzazione più complicati che gli informatici abbiano mai affrontato^[83].

Paradosso di Levinthal

Un aspetto importante del problema è sottolineato dal *paradosso di Levinthal*. Nel 1968 Cyrus Levinthal^[84] si rese conto che, a causa dell'elevato numero di gradi di libertà di un polipeptide non ripiegato, tale molecola presenterebbe un numero astronomico di possibili conformazioni finali. Se la proteina raggiungesse la sua conformazione finale passando via via attraverso tutte queste configurazioni, sarebbe necessario un tempo ben superiore all'età attualmente stimata dell'universo per raggiungere la configurazione corretta, anche se ogni passaggio richiedesse pochi *picosecondi*¹⁴.

Prendendo ad esempio un polipeptide di 100 residui si avranno 99 legami peptidici e di conseguenza 198 differenti angoli di legame ϕ e ψ . Assumendo che ognuno di questi angoli possa esistere in ognuna delle 3 configurazioni stabili la proteina potrebbe assumere fino a 3^{198} configurazioni (includendo ogni possibile ridondanza di ripiegamento).

In natura però molte piccole proteine si ripiegano spontaneamente in un tempo dell'ordine dei millisecondi o addirittura dei microsecondi. Il tempo di generazione di *E. coli* può essere di circa venti minuti: ciò significa che tutte le proteine essenziali per tale organismo (e presumibilmente di tutti gli altri) possono essere prodotte da zero in un tempo decisamente ristretto, al massimo nell'ordine dei minuti.

Il processo di ripiegamento non è quindi una ricerca all'interno dell'enorme spazio degli stati configurazionali possibili. La differenza enorme che esiste tra il tempo del ripiegamento prevedibile in teoria e quello osservato in realtà è appunto chiamato paradosso di Levinthal.

Come accennato nella sezione 3.3, la malattia di Alzheimer, la fibrosi cistica e altre malattie neurodegenerative sono associate al mal ripiegamento delle proteine. La conoscenza dei fattori di mal ripiegamento e la comprensione del processo di ripiegamento proteico potrebbero aiutare nello sviluppo di cure per queste malattie. Per queste ragioni è importante anche rispondere alle altre domande del problema e non fermarsi alla predizione della struttura finale, nonostante questa conoscenza fornisca un grande vantaggio per lo sviluppo di nuovi farmaci e il design di nuove proteine.

¹⁴Levinthal stima 10^{300} possibili conformazioni teoriche per una proteina di 2000 atomi^[84].

Capitolo 4

Predizione della struttura di proteine

Il protein folding problem ha sia guidato che tratto beneficio dagli avanzamenti nei metodi sperimentali e computazionali^[52]. Uno dei maggiori obiettivi della biologia computazionale è proprio il Protein Structure Prediction (PSP), ovvero la predizione della struttura nativa tridimensionale di una proteina a partire dalla sua sequenza amminoacidica. Il PSP è il problema opposto al *protein design* (la progettazione di nuove sequenze proteiche aventi delle specifiche attività).

Grazie al CASP¹, alla crescita dei database sulle proteine, allo sviluppo dei metodi per omologia, di allineamento di sequenze e all'utilizzo del Deep Learning, i metodi computazionali hanno registrato incredibili progressi, come il livello raggiunto da AlphaFold può dimostrare.

La predizione della struttura di proteine è uno strumento fondamentale: in medicina per la comprensione delle malattie da misfolding, nell'industria farmaceutica per risparmiare anni di laboriosi e costosi esperimenti correntemente richiesti per lo sviluppo di un singolo farmaco (*drug design*), in biotecnologia per il design di nuovi enzimi e in generale per acquisire maggior conoscenza sul protein folding in tutti i suoi lati.

4.1 Metodi e strumenti informatici

La piccola percentuale di strutture determinate e il gap che continua a crescere con le sequenze conosciute (vedi sotto la sezione 4.1.4) è una conseguenza della lentezza e della dispendiosità dei metodi sperimentali (e in parte anche dei progressi delle tecnologie di sequenziamento). I metodi computazionali, significativamente più veloci ed economici, potrebbero fornire una possibile soluzione a questo problema.

¹Critical Assessment of Structure Predictions, vedi la sezione 4.1.6.

4.1.1 Workflow e classificazione dei metodi per il PSP

Lo sviluppo di metodi computazionali per la predizione della struttura di proteine si è sviluppato lungo due percorsi complementari, che si concentrano sulle *interazioni fisiche* o sulla *storia evolutiva*. Esistono quindi due *paradigmi fondamentali* per affrontare il problema:

- paradigma *ab initio*
- paradigma *data-based*

Il paradigma *ab initio* (o *de novo*) si basa su un approccio puramente fisico, nel quale la struttura è predetta da zero simulando principi fisici. In questo paradigma si integra fortemente la comprensione attuale delle forze molecolari trainanti, in simulazioni termodinamiche e/o cinetiche della fisica delle proteine, o in approssimazioni statistiche della stessa.

Nel paradigma *data-based* invece si fa uso di informazioni estratte da database di sequenze o strutture di proteine.

Le proteine che esistono in natura oggi si sono sviluppate attraverso lunghi processi evolutivi, progredendo attraverso mutazioni casuali e selezione naturale. La rivoluzione genetica degli anni '50, consentendo la determinazione delle sequenze amminoacidiche, ha permesso il nascere di metodi di confronto delle sequenze. È per questo che si possono ricavare informazioni sulla struttura 3D di una sequenza amminoacidica cercando altre proteine con proprietà nella sequenza simili e una struttura nota.

È bene chiarire sin dall'inizio che metodi basati totalmente sul primo paradigma non sono computazionalmente trattabili. Per questa ragione, i metodi odierni per la PSP di sequenze senza struttura nota, sono sempre in qualche misura *data-based*. Possono essere quasi totalmente basati sui dati come nel caso della modellazione per *omologia* e *fold recognition* oppure parzialmente basati sui dati negli altri casi.

L'utilizzo, anche parziale, di tecniche *data-based* è necessario per guidare la ricerca nello spazio conformazionale tramite rappresentazioni più grossolane, in modo da superare il paradosso di Levinthal.

Varie osservazioni evolutive e strutturali supportano questo approccio. Si è visto che la struttura è più conservata della sequenza: un'identità anche solo del 50% può implicare un'identità di ripiegamento ed è possibile ricavare informazioni da mutazioni coevolute.

L'approccio *data-based* è anche supportato dall'osservazione che, sebbene il numero di famiglie di proteine multi-dominio cresca rapidamente, la scoperta di nuovi domini singoli sembra stabilizzarsi. Ciò suggerisce che la maggioranza delle proteine possa ripiegarsi in un numero limitato di domini strutturali, forse non più di 10.000-20.000. Per molte fami-

glie a singolo dominio la struttura di almeno un membro è conosciuta: si stima che ciò permetta di avere informazioni su più di 3/4 delle sequenze nei database^[7].

Una *pipeline* standard per la previsione della struttura delle proteine è basata su fasi di previsione intermedie, nelle quali vengono dedotte delle astrazioni che, risultando più semplici della struttura 3D completa, rivelano delle informazioni importanti per guidare le successive ricerche e modellazioni. Queste informazioni possono essere chiamate "annotazioni della struttura delle proteine" (Protein structure annotations, PSA). Le annotazioni sono divise in 2 categorie a seconda delle informazioni che forniscono:

- *Annotazioni 1D*: informazioni sulla backbone, caratteristiche strutturali locali (es. formazione di strutture secondarie, accessibilità al solvente)
- *Annotazioni 2D*: vincoli spaziali (es. contact map)

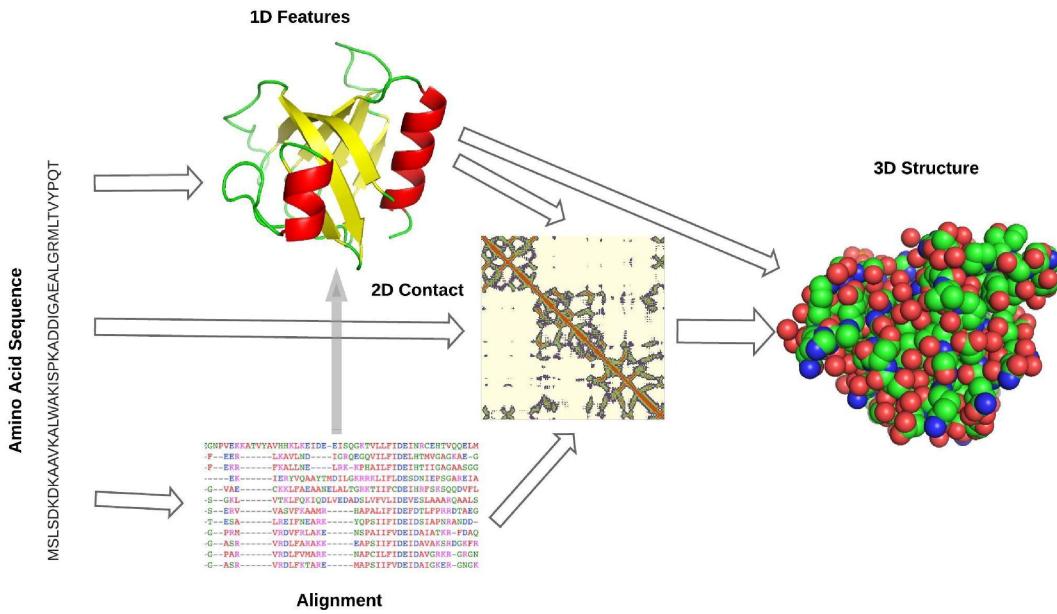


Figura 4.1: Pipeline generica per la previsione della struttura 3D di una proteina. Questo schema vuole mettere in risalto gli step intermedi relativi alle annotazioni. Fonte[40]

Ci sono due possibili situazioni in cui ci si può trovare quando si vuole modellare una proteina: si riesce a trovare almeno una proteina omologa (o con caratteristiche simili) oppure no. Nel primo caso la struttura trovata verrà chiamata *template* e si affronterà una previsione di tipo *template-based modeling* (TBM), più semplice, mentre nell'altro caso si affronterà una previsione *template-free modeling* (FM).

Come già accennato, nel panorama attuale molti degli approcci oggi utilizzati per il PSP sono prevalentemente *data-based*: i metodi puri *ab initio* vengono raramente uti-

lizzati per la predizione della struttura di proteine (per ragioni che verranno spiegate in dettaglio nella sezione 4.4.3). Tuttavia, nella pratica, alcune intuizioni del paradigma *ab initio* vengono usate in metodi prevalentemente *data-based*, ad esempio la funzione euristica di valutazione che simula il campo di forza per calcolare l'energia potenziale. Le varie tecniche vengono utilizzate in combinazione: non vi è una singola tecnica principale e i metodi migliori sono proprio quelli che riescono ad integrare vari approcci.

Quando si ha di fronte una sequenza di una proteina senza struttura nota e si vuole predire la sua forma tridimensionale, un metodo per il PSP attuale agirebbe nel seguente modo (vedi fig. 4.2)²:

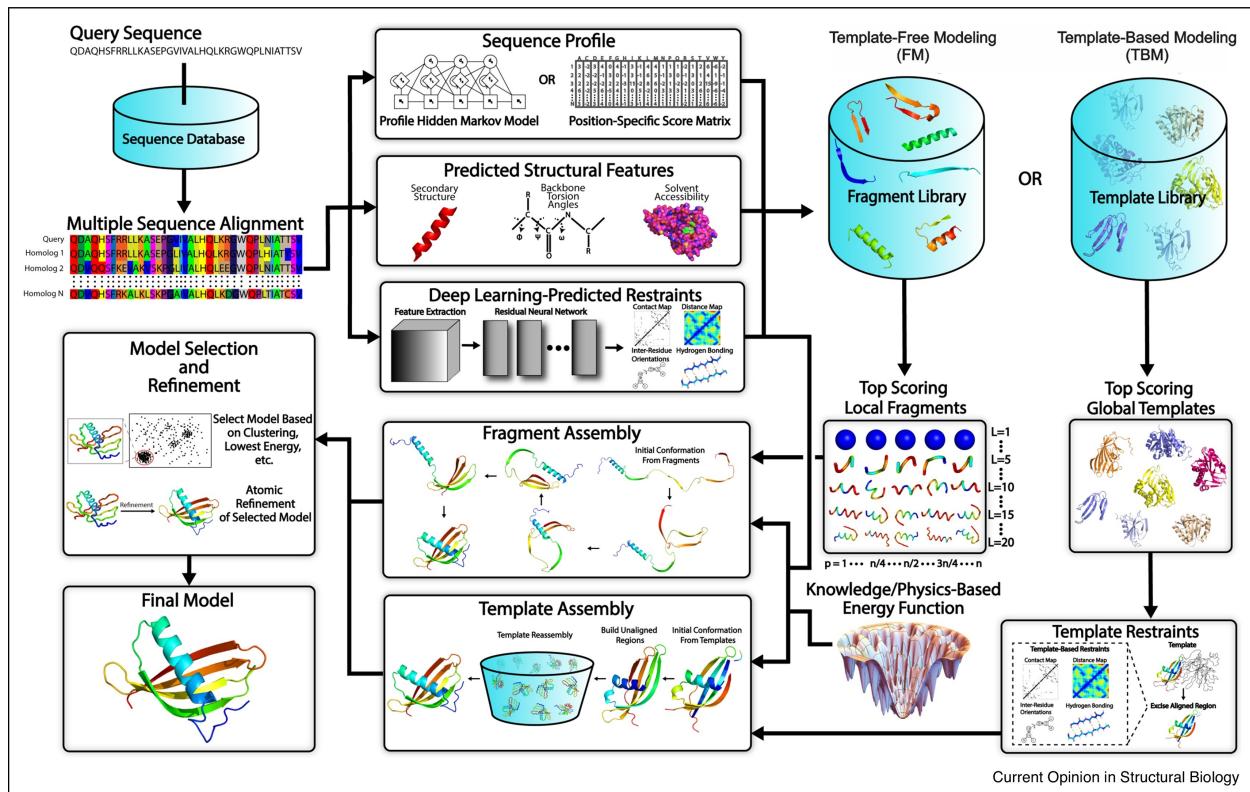


Figura 4.2: Step tipici negli approcci al PSP di tipo TBM e FM. Fonte[85]

1. viene generata una MSA per ottenere informazioni evolutive ed identificare sequenze omologhe
2. viene profilata la sequenza, sfruttando anche i risultati dell'MSA, e viene usata per dedurre le annotazioni 1D e 2D
3. viene scelto il tipo di modellazione adeguato

²Ogni argomento o metodo citato verrà spiegato nel dettaglio successivamente, in questa sezione l'obiettivo è di fornire una visione globale degli argomenti.

4. vengono assemblati i frammenti o i template
5. viene scelto il modello fra vari candidati, che sarà poi raffinato a livello atomico

Nel 1° passo l'obiettivo è ottenere informazioni evolutive che serviranno sia per identificare sequenze omologhe che per dedurre le annotazioni. Se si riesce a trovare almeno una proteina omologa allora sarà possibile procedere alla *modellazione per omologia (homology modeling)*.

Nel 2° passo si applicano delle deduzioni sulla sequenza target al fine di ricavare delle annotazioni sulla struttura, per due scopi:

- impostare dei vincoli spaziali e strutturali per guidare la modellazione
- nel caso in cui non siano state trovate proteine omologhe: per identificare template globali al fine di applicare protocolli di *fold recognition*; se non se ne trovano, tali informazioni verranno utilizzate per valutare i frammenti nella ricerca all'interno di una *fragment library*

Si rientrerà nel caso del TBM sia che venga eseguita una modellazione per *omologia* che una modellazione basata sul *fold recognition*. Si rientra invece nel caso del MF quando non si riescono a trovare dei template globali. In questo caso vengono principalmente utilizzate tecniche di modellazione *fragment-based*, ovvero basate su frammenti di proteine che verranno poi integrati. Nel 3° passo, a seconda che ci si trovi nel caso TBM o MF vengono attuate le rispettive modellazioni.

Nel 4° passo l'assemblaggio è eseguito sotto la guida di una funzione euristica del campo di forza, che può essere *energy-based* e/o *knowledge-based*, combinata con una rete neurale profonda per la predizione di determinate caratteristiche. Nel caso di una modellazione TBM si hanno anche delle restrizioni spaziali sul modello. Vengono utilizzate tecniche specifiche per regioni non allineate, come i loop (*loop modeling*).

Nel 5° passo vengono valutati i modelli, viene eseguita una valutazione della qualità (*quality assessment, QA*) stimando l'accuratezza del modello (*estimation of model accuracy, EMA*) e infine viene eseguito il raffinamento. Tipicamente viene scelto il modello con minore energia.

Nota sulla classificazione dei metodi

La predizione della struttura di proteine è stata ed è tutt'ora un campo in evoluzione. Per tale ragione risulta difficile classificare e raggruppare i metodi in nette categorie. Agli

albori i metodi erano divisi in *ab initio* e *comparative modeling*. Oggi il confine non è più così marcato (come si è potuto vedere dalla panoramica nella sezione precedente). Il CASP divide la modellazione in due classi principali in base alla difficoltà: TBM ed FM^[86], in base all'utilizzo o meno di informazioni ricavate da template (proteine con struttura 3D nota).

Nonostante questa divisione possa risultare efficace, in questo lavoro si è ritenuto opportuno seguire una classificazione diversa, che provasse a delineare invece le idee alla base degli approcci odierni e raggrupparli in più livelli secondo questo principio. La motivazione risiede nel fatto che nessun metodo preso singolarmente può dare risultati soddisfacenti: negli anni si è assistito infatti ad un utilizzo combinato dei tanti metodi trovati sfocando sempre più i margini fra le categorie. Tutto questo ha creato confusione ed un utilizzo improprio dei termini. Come si vedrà, *ab initio* indica il "puro" approccio fisico, mentre con la divisione operata dal CASP si tende ad utilizzare come sinonimi *ab initio* e *template-free modeling*³, cosa che, in fase di stesura della tesi, è stata reputata equivoca e forse addirittura erronea. Allo stesso tempo i metodi si sono evoluti negli ultimi anni, specialmente con l'avvento del Deep Learning, e le vecchie classificazioni⁴ non rispecchiano più l'attuale struttura dei metodi usati. Si è deciso di focalizzare l'attenzione sulla struttura dei metodi per il PSP odierni e da qui provare ad astrarre verso l'alto. Si è scelto per tale ragione di dividere la fase di annotazione da quella di modellazione e di rimarcare la differenza basilare tra i due grandi paradigmi per il PSP. Il quadro di riferimento è il *workflow* della figura 4.2, che delinea la struttura tipica di un metodo attuale (tralasciando i metodi end-to-end come AlphaFold2).

4.1.2 Soft computing e deep learning

Nel corso degli anni il PSP è stato affrontato anche con approcci di *soft computing*. Si sta parlando di approcci perlopiù *data-based*. I principali metodi di *soft computing* utilizzati fanno capo a queste tecniche^[87]:

- Machine Learning:
 - ANN (Artificial neural network)
 - SVM (Support vector machines), es. SVM-SEQ
 - k-Nearest Neighbors

³Un esempio è in M. Torrisi, G. Pollastri e Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, 2020. Nonostante ciò è proprio da tale lavoro che si è preso spunto per l'idea delle annotazioni.

⁴Compresa quella di A. Kessel e N. Ben-Tal, *Introduction to proteins: Structure, function and motion*, 2^a ed. Chapman e Hall/CRC, 2018, su cui il presente capitolo si basa in parte, precisamente sul capitolo 3.4.

- linear regression
- HMM (Hidden Markov Models)
- Support vector regression
- EC (Evolutionary computing), es. MECoMaP
- approcci *statistici*, basati principalmente sull’omologia e sul fold recognition
- modelli *matematici*, come un adattamento della programmazione lineare intera

I limiti della computazione evolutiva sono: la difficoltà di trovare un criterio di arresto e la possibilità di convergere verso un massimo locale come risultato di una configurazione sfavorevole dei parametri. Utilizzando questa tecnica è necessario tenere conto di una corretta scelta della rappresentazione del problema, della funzione fitness, della dimensione della popolazione e del tasso degli operatori genetici. Ad esempio, una piccola dimensione della popolazione può far sì che l’EA non possa esplorare lo spazio sufficiente per trovare una soluzione corretta.

Come limiti della tecnica SVM invece si può parlare del fatto che i modelli del kernel overfittino il criterio di selezione del modello, della difficoltà nella selezione dei parametri ottimali della funzione del kernel e della complessità algoritmica e gli ampi requisiti di memoria nei compiti su larga scala.

Le reti neurali invece offrono un elevato grado di flessibilità, nonostante la codifica dei dati di input necessariamente restrinja l’insieme delle possibili informazioni. Oltre ai vettori di input codificanti coppie di amminoacidi è possibile includere neuroni con informazioni aggiuntive, come la lunghezza della sequenza, valori di idrofobicità dell’ambiente o informazioni evolutive. Le reti neurali presentano comunque delle limitazioni di cui tenere conto, ad esempio l’uso di parametri appropriati e il possibile overfitting.

L’arrivo del Deep Learning

Il campo del PSP ha assistito a numerosi avanzamenti grazie ad approcci basati sul Deep-Learning (DL) come evidenziato dal successo di AlphaFold nell’ultimo CASP. Il DL sta diventando una delle tecnologie principali per vari domini scientifici: computer vision, natural language processing, speech recognition, guida autonoma, ecc.

Anche se le reti neurali di tipo FFNN sono state usate per prevedere annotazioni 1D sin dagli anni ’80^[40]⁵, è però solo negli ultimi 10 anni (specialmente negli ultimi 2 CASP) che si sta assistendo a vari avanzamenti nel PSP grazie al DL, in particolare nei seguenti campi^[73]:

⁵Queste reti erano tipicamente utilizzate nella loro cosiddetta versione ”a finestra”, in cui ogni segmento, composto da un numero fisso di amminoacidi in una sequenza, veniva trattato come input per un esempio separato. L’obiettivo del segmento era l’annotazione di interesse per uno degli amminoacidi in esso (solitamente quello centrale).

- generazione di MSA (es. DeepMSA)
- predizione di contatti (contact map, es. TripletRes)
- predizione di distogrammi (es. RaptorX)
- predizione della distanza fra i residui (es. PDNET)
- guidare l’assemblaggio iterativo di frammenti
- valutazione dei modelli e raffinamento (es. QDeep)
- pipeline generale del PSP (es. trRosetta o AlphaFold1)
- approcci DL-based end-to-end (es. AlphaFold2)
- pulizia dei dati nel cryo-EM (es. PIXER)
- predizione guidata sperimentalmente dalla cryo-EM (es. DeepTracer)
- predizione di strutture multidominio (es. FUpred)

Tra i modelli di Deep Learning maggiormente utilizzati nei metodi odierni vi sono le ResNet^[73].

4.1.3 Output e misure di valutazione

Modelli di output

I modelli dei dati di output hanno lo scopo di rappresentare la struttura terziaria predetta di una proteina. I principali modelli sono^[87]:

- *modello ad angolo di torsione.* Gli angoli di torsione (ϕ, ψ) sono legati alle possibilità della catena polipeptidica di assumere determinate conformazioni e data una conformazione ogni angolo di torsione è ben definito. Per tale ragione una possibile rappresentazione è

$$[(\phi_1, \psi_1), \dots, (\phi_n, \psi_n)]$$

dove n è il numero dei residui. Il grafico di Ramachandran consente di evitare le possibili collisioni fra gli atomi.

- *modello reticolare*, nel quale ogni amminoacido può essere rappresentato come una coppia (x, y) dove x e y sono le coordinate in un reticolo 2D. Considerando i possibili movimenti un’altra rappresentazione potrebbe essere tramite vettori di direzione:

$$(L_1, L_2, \dots, L_n)$$

dove $L_i \in \{UP, DOWN, LEFT, RIGHT\}$.

- *binary contact map*, nel quale vengono rappresentati i contatti fra i residui tramite una matrice $L \times L$ dove L rappresenta il numero dei residui. Un elemento (i, j) nella matrice rappresenta una coppia di amminoacidi che possono essere in contatto (1) o no (0). Si definisce contatto una distanza tra i residui inferiore ad una determinata soglia, tipicamente 8Å. Gli atomi di riferimento per tale calcolo sono in genere C_α o C_β .

Data una *contact map* è possibile ricostruire il modello 3D di una proteina risolvendo il Molecular Distance Geometry Problem (MDGP). Quando usate come modello di rappresentazione delle proteine, le mappe di contatto sono utili anche per confrontare le strutture.

- *distance matrix*, simile alla mappa dei contatti ma rappresenta le distance a valori reali invece del contatto binario
- *hydrophobic-polar* (HP), nel quale una sequenza è rappresentata come una stringa $s \in (H, P)^+$, dove H rappresenta un amminoacido idrofobico e P un amminoacido idrofilo.

Metriche di valutazione

Le misure di qualità valutano l'affidabilità dei modelli 3D rispetto alla struttura della proteina determinata sperimentalmente, le principali sono^[87]:

- RMSD (Root mean square deviation, indica la deviazione standard) rappresenta la deviazione assoluta (in Å) dei singoli atomi C_α tra il modello e la struttura conosciuta:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N |r_i^{model} - r_i^{real}|^2}$$

dove r_i^{model} indica la posizione dell'i-esimo atomo C_α nel modello. È stata la metrica di riferimento dal CASP1 al CASP4. È stato abbandonato poiché:

- il punteggio è dominato da valori anomali in regioni scarsamente previste mentre allo stesso tempo è insensibile alle parti mancanti
- dipende fortemente dalla sovrapposizione del modello con la struttura di riferimento
- GDT_TS (Global distance test_total score), è usato come maggior criterio di valutazione nel CASP e descrive le percentuali di residui ben modellati nel modello rispetto al target:

$$GDT_TS = 100 \times \frac{\sum_{d_i} \frac{GDT_i}{NT}}{4}$$

dove GDT_i è il numero di atomi C_α di una predizione che non deviano più di una soglia stabilità d_i (in Å) dai C_α della struttura conosciuta, dopo sovrapposizione ottima. NT è il numero degli aminoacidi della proteina e $d_i \in \{1, 2, 4, 8\}$. Essendo un criterio basato su sovrapposizione globale degli atomi C_α anch'esso soffre di limitazioni quando applicato a proteine flessibili e/o multi-dominio e non considera l'accuratezza nelle differenze fra atomi che non siano C_α .

- IDDT (local Distance difference test), è una metrica di valutazione non basata su sovrapposizione globale che valuta differenze di distanze locali di tutti gli atomi in un modello, includendo la validazione di plausibilità stereochimica^[88]. IDDT misura quanto sia stato riprodotto l'ambiente di una struttura di riferimento in un modello di una proteina. È calcolato su tutti gli atomi. La struttura di riferimento può essere una singola struttura o un insieme di strutture equivalenti.

Valutando tutti gli atomi è in grado di catturare l'accuratezza, ad esempio, della geometria locale di un sito di legame o il corretto ripiegamento del nucleo di una proteina. È stato introdotto nel CASP9. Assegna punteggi elevati a regioni ben previste anche se la previsione globale non è ben allineata alla struttura reale. Ciò risulta particolarmente utile nelle strutture multi-dominio, in cui i singoli domini possono essere molto accurati mentre la loro posizione relativa non lo è.

- TM_score (Template modeling score), misura la somiglianza globale tra la struttura modello e quella conosciuta in base alla distanza tra ogni paio di residui. Il punteggio è compreso tra $(0, 1]$, dove 1 indica una corrispondenza perfetta tra due strutture. Generalmente punteggi inferiori a 0,20 corrispondono a proteine non correlate mentre le strutture con un punteggio superiore a 0,5 si pensa abbiano all'incirca lo stesso ripiegamento.

- per la valutazione delle *contact map* vengono usate 3 misure:
 - *accuracy*, che rappresenta il numero di contatti correttamente predetti
 - *coverage*, che riflette la proporzione di contatti predetti diviso i contatti reali
 - X_d , distribuzione dell'accuratezza della predizione dei contatti

$$\text{accuracy} = \frac{C}{C_p}; \quad \text{coverage} = \frac{C}{C_t}; \quad X_d = \sum_{i=1}^{15} \frac{P_i - P_a}{i}$$

dove C_t rappresenta il numero dei contatti reali, C il numero di predizioni corrette, C_p il numero totale dei contatti predetti, P_i riflette il numero di coppie

stimate la cui distanza è nel range $(4(i - 1), 4i)$ e P_a rappresenta il numero di coppie reali la cui distanza è nello stesso range.

4.1.4 Database e formati

Come si è già visto è possibile descrivere una proteina attraverso la sua sequenza amminoacidica. Il risultato è una stringa di lettere alfabetiche, poiché ogni amminoacido corrisponde ad una determinata lettera (vedi fig. 2.15). Si ricorda anche che ad un amminoacido può corrispondere uno o più codoni (3 paia di basi azotate nel DNA).

Per quanto riguarda la struttura delle proteine, il formato standard è il PDB (Protein Data Bank), un esempio è quello mostrato in figura 4.3.

The diagram shows a portion of a PDB file with 18 rows of data. Each row represents an atom with the following fields: Atom Type (NH1, NH2, N, CA, C, O, CB, CG, OD1, ND2, N, LEU, CA, C, O, CB, CG, CD1), Residue Number (A 149, A 149, ASN A 150, LEU A 151, LEU A 151), Chain (A, A, A), and X,Y,Z Coordinates (31.814, 32.203, 29.346, 28.480, 28.606, 27.803, 28.732, 28.284, 27.205, 29.110, 29.629, 29.868, 29.953, 30.149, 31.208, 31.436, 32.846, -31.597, -32.934, -24.359, -23.190, -22.168, -21.276, -22.524, -23.389, -23.981, -23.463, -22.313, -21.415, -22.205, -23.422, -20.735, -19.884, -19.333, 16.995, 18.816, 18.812, 18.933, 17.808, 17.678, 20.282, 21.447, 21.430, 22.466, 16.996, 15.894, 14.597, 14.614, 16.100, 17.337, 17.256). Below the table, six labels are shown in colored boxes with arrows pointing to specific columns: 'Atom Number' (red) points to the first column; 'Atom Type' (black) points to the second column; 'Amino Acid Type' (blue) points to the third column; 'Chain' (green) points to the fourth column; 'Residue Number' (yellow) points to the fifth column; and 'X,Y,Z Coordinates' (purple) points to the last three columns.

| | | | | | | | | |
|------|------|-----|-----|---|-----|--------|---------|--------|
| ATOM | 1132 | NH1 | ARG | A | 149 | 31.814 | -31.597 | 16.995 |
| ATOM | 1133 | NH2 | ARG | A | 149 | 32.203 | -32.934 | 18.816 |
| ATOM | 1134 | N | ASN | A | 150 | 29.346 | -24.359 | 18.812 |
| ATOM | 1135 | CA | ASN | A | 150 | 28.480 | -23.190 | 18.933 |
| ATOM | 1136 | C | ASN | A | 150 | 28.606 | -22.168 | 17.808 |
| ATOM | 1137 | O | ASN | A | 150 | 27.803 | -21.276 | 17.678 |
| ATOM | 1138 | CB | ASN | A | 150 | 28.732 | -22.524 | 20.282 |
| ATOM | 1139 | CG | ASN | A | 150 | 28.284 | -23.389 | 21.447 |
| ATOM | 1140 | OD1 | ASN | A | 150 | 27.205 | -23.981 | 21.430 |
| ATOM | 1141 | ND2 | ASN | A | 150 | 29.110 | -23.463 | 22.466 |
| ATOM | 1142 | N | LEU | A | 151 | 29.629 | -22.313 | 16.996 |
| ATOM | 1143 | CA | LEU | A | 151 | 29.868 | -21.415 | 15.894 |
| ATOM | 1144 | C | LEU | A | 151 | 29.953 | -22.205 | 14.597 |
| ATOM | 1145 | O | LEU | A | 151 | 30.149 | -23.422 | 14.614 |
| ATOM | 1146 | CB | LEU | A | 151 | 31.208 | -20.735 | 16.100 |
| ATOM | 1147 | CG | LEU | A | 151 | 31.436 | -19.884 | 17.337 |
| ATOM | 1148 | CD1 | LEU | A | 151 | 32.846 | -19.333 | 17.256 |

Figura 4.3: Formato PDB. Fonte[89]

Un file PDB è essenzialmente un contenitore di coordinate X, Y, Z per ogni atomo di una struttura molecolare. I programmi di visualizzazione molecolare traducono queste coordinate X, Y, Z in immagini tridimensionali interattive. La maggior parte dei file PDB inizia con informazioni scritte sulla struttura, il laboratorio che l'ha determinata e le tecniche utilizzate nel laboratorio. Eventuali commenti come questi iniziano sempre con la parola "REMARK" e verranno ignorati dai software di visualizzazione. Esiste anche una versione XML del formato PDB chiamata PDBML.

Il formato originale (PDB) è però limitato dalla larghezza delle schede perforate per computer a 80 caratteri per riga. Intorno al 1996, il formato mmCIF (macromolecular Crystallographic Information file), un'estensione del formato CIF, è stato gradualmente introdotto. mmCIF è diventato il formato standard per l'archivio PDB nel 2014 e nel 2019, il wwPDB ha annunciato che le deposizioni per i metodi cristallografici sarebbero state accettate solo in formato mmCIF.

Database di proteine

Il *Protein Data Bank* (PDB) è un archivio di strutture tridimensionali di macromolecole biologiche determinate sperimentalmente. È il deposito principale dei centri biologici di struttura. I dati contenuti nell'archivio includono coordinate atomiche, fattori di struttura cristallografici e dati sperimentali NMR. Oltre alle coordinate, ogni deposizione include anche i nomi delle molecole, le informazioni sulla struttura primaria e secondaria, i riferimenti ai database di sequenze, ove appropriato, e le informazioni sull'assemblaggio biologico e sul ligando, i dettagli sulla raccolta dei dati e sulla soluzione della struttura e le citazioni bibliografiche.

Quando il PDB fu fondato, nel 1971, conteneva appena 7 strutture proteiche ed era frutto di una congiunzione fra il Cambridge Crystallographic Data Centre (UK) e il Brookhaven National Laboratory (USA), e da allora si è reso protagonista di una crescita pressoché esponenziale nel numero di strutture, che non mostra alcun segno di rallentamento. Nel 1998 il PDB è stato trasferito al Research Collaboratory for Structural Bioinformatics (RCSB). Nel 2003, con la formazione del *wwPDB*, il PDB è diventato un'organizzazione internazionale. I membri fondatori sono PDBe (Europa), RCSB (USA) e PDBj (Giappone). Il BMRB (Biological Magnetic Resonance Data Bank) si è unito nel 2006. Ciascuno dei quattro membri di wwPDB può fungere da centro di deposito, elaborazione dati e distribuzione dei dati PDB. Il trattamento dei dati si riferisce al fatto che il personale del wwPDB esamina e annota ogni voce presentata. I dati vengono quindi automaticamente verificati per verificarne la plausibilità.

UniProt (Universal Protein Resource) è il più grande database bioinformatico per le sequenze proteiche di tutti gli organismi viventi e dei virus. Molte informazioni derivano da progetti di sequenziamento del genoma. I database UniProt sono UniProt Knowledge-base (UniProtKB), UniProt Reference Cluster (UniRef) e UniProt Archive (UniParc). Il consorzio UniProt e le istituzioni ospitanti EMBL-EBI, SIB (Swiss Institute of Bioinformatics) e PIR (Protein Information Resource) sono impegnati nella conservazione a lungo termine dei database UniProt.

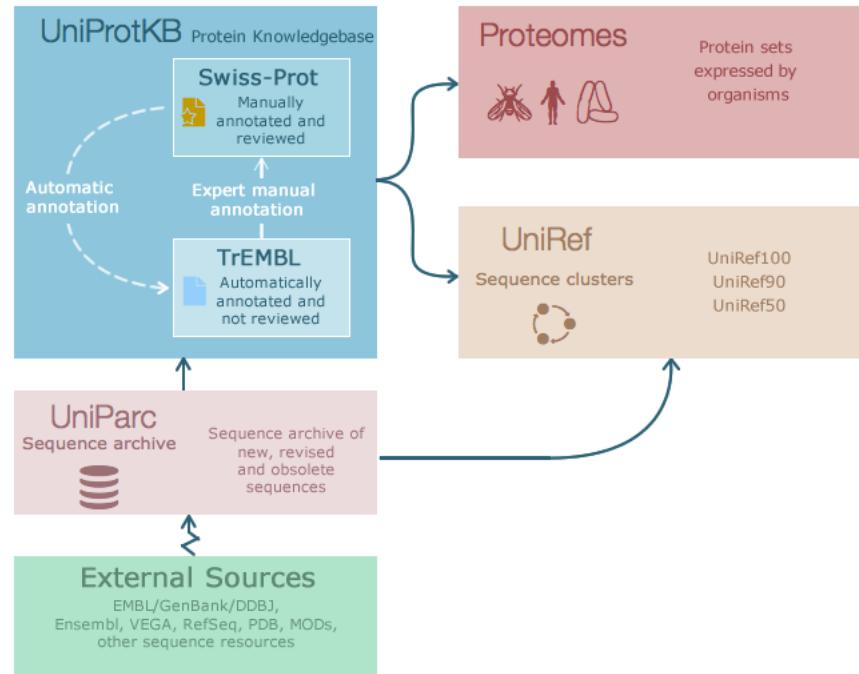


Figura 4.4: Struttura di UniProt. Fonte[90]

EMBL-EBI e SIB insieme mantenevano Swiss-Prot e TrEMBL, mentre PIR produceva il Protein Sequence Database (PIR-PSD). Questi due set di dati coesistevano con diverse priorità di copertura e annotazione della sequenza proteica. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) è stato originariamente creato perché i dati di sequenza venivano generati a un ritmo che superava la capacità di Swiss-Prot di tenere il passo. Nel frattempo, PIR ha mantenuto il PIR-PSD e i relativi database, incluso iProClass, un database di sequenze proteiche e famiglie curate. Nel 2002 i tre istituti hanno deciso di unire le proprie risorse e competenze e hanno costituito il consorzio UniProt.

EMBL (European Molecular Biology Laboratory) è un'organizzazione di ricerca intergovernativa finanziata da oltre 20 Stati membri, potenziali e associati. L'istituto europeo di bioinformatica, EMBL-EBI (EMBL-European Bioinformatics Institute) si trova a Cambridge e gestisce, insieme all'NCBI (National Center for Biotechnology Information, USA), i maggiori database di sequenze nucleotidiche e proteiche. Uno dei ruoli dell'EMBL-EBI è quello di indicizzare e mantenere i dati biologici in una serie di database, inclusi Ensembl (che ospita i dati della sequenza di interi genomi), UniProt e PDB. Fornisce anche una varietà di servizi e strumenti online, come BLAST o lo strumento di allineamento di sequenze ClustalΩ.

EMDB (Electron Microscopy Data Bank) è una banca dati di microscopia elettronica

(EMDB). È un archivio pubblico per mappe volumetriche di crio-microscopia elettronica e tomografia di complessi macromolecolari e strutture subcellulari. Copre una varietà di tecniche, tra cui l'analisi di singole particelle, la tomografia elettronica e la cristallografia elettronica. È stato fondato nel 2002 dall'EMBL-EBI e nel gennaio 2021 è divenuto un archivio gestito dalla wwPDB.

La predizione della struttura è importante per un semplice motivo: i biochimici conoscono oggi la sequenza amminoacidica per più di 225 milioni di proteine^[91] (UniProt, con circa 4.5-5 milioni aggiunte ogni mese) ma sono state determinate solamente circa 160.000 strutture tridimensionali di proteine^[91] (PDB⁶, con poco più di 10.000 strutture aggiunte ogni anno).

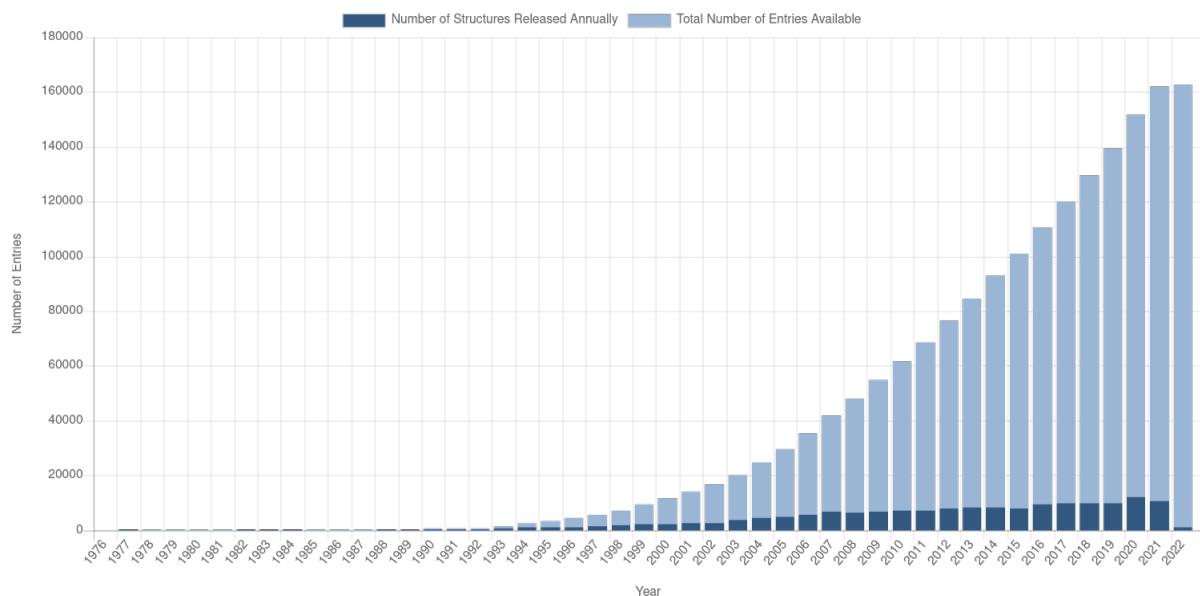


Figura 4.5: Crescita complessiva del numero di strutture di proteine pubblicate nel PDB. Fonte[92]

Sono disponibili anche database di modelli 3D di strutture proteiche, come ModBase. Un database di questo tipo è stato fondato dall'EMBL-EBI in congiunzione con DeepMind grazie al successo di AlphaFold: *AlphaFold DB*.

AlphaFold DB è un database apertamente accessibile di previsioni ad alta precisione di strutture proteiche. Basato su AlphaFold v2.0, ha consentito un'espansione senza precedenti della copertura strutturale dello spazio noto della sequenza proteica. AlphaFold DB

⁶In particolare è da notare che spesso sono presenti più strutture per una data proteina, infatti le strutture proteiche distinte determinate sono circa 55.000. Al 3 Febbraio 2022 sono presenti 162.913 strutture di proteine nella versione dell'RCSCB. Nel PDB ci sono anche strutture di altre macromolecole (complessi di acidi proteici-nucleici, DNA e RNA) per un totale, incluse le proteine, di 186.670^[92]. Nel wwPDB (database globale) vi sono in totale 197.961 strutture di macromolecole^[93].

fornisce l'accesso programmatico e la visualizzazione interattiva delle coordinate atomiche previste di una proteina, e le relative misure di confidenza stimate. La versione iniziale di AlphaFold DB contiene oltre 360.000 strutture predette in 21 proteomi di organismi modello. Il database sarà presto ampliato per coprire la maggior parte delle sequenze rappresentative del set di dati UniRef90 (oltre 100 milioni)^[94].

Nella prima versione di AlphaFold DB si è tentato di prevedere la maggior parte delle sequenze nel proteoma di riferimento UniProt nell'intervallo di lunghezza di 16-2700 amminoacidi (oltre a frammenti di 1400 residui per coprire proteine umane più lunghe) per gli organismi attualmente coperti. Sono state escluse le sequenze che contengono amminoacidi non standard. Non vengono fornite isoforme multiple.

AlphaFold DB archivia e fornisce accesso alle coordinate atomiche nei formati PDB e mmCIF, mentre fornisce il PAE (Predicted Aligned Error) in JSON.

4.1.5 Rappresentazione grafica

I file di struttura possono essere visualizzati tramite una rappresentazione grafica tridimensionale utilizzando uno dei numerosi software disponibili⁷, inclusi:

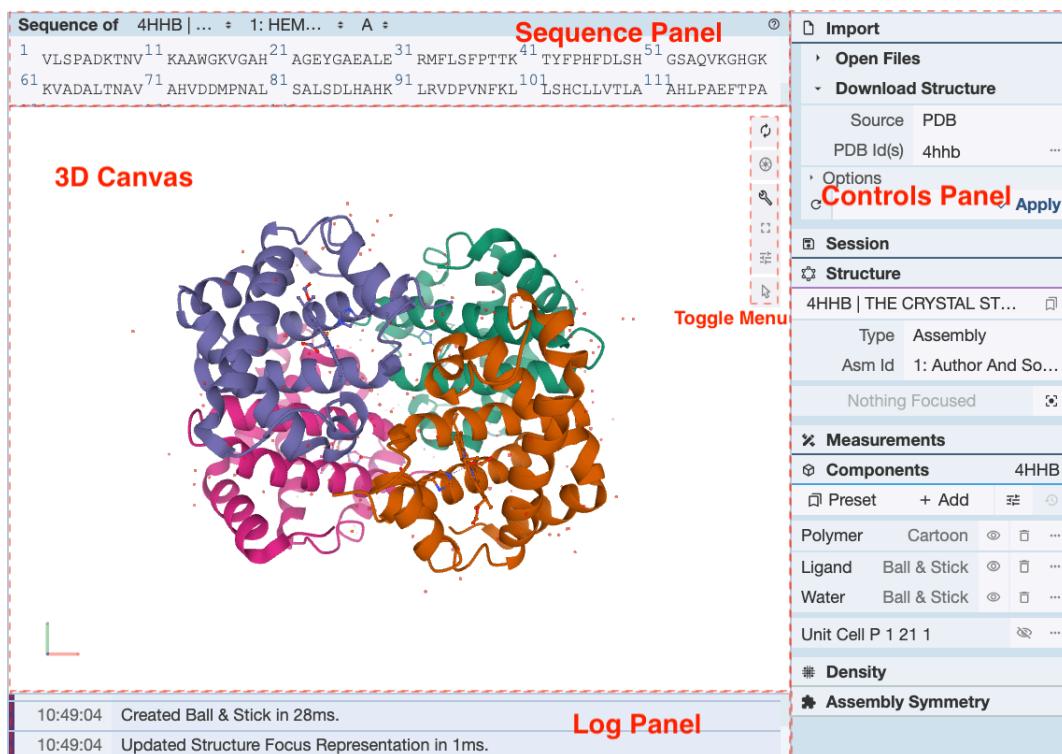


Figura 4.6: Interfaccia utente di Mol* utilizzata nel sito del PDB. Fonte[92]

⁷Una lista più completa di software grafici per macromolecole è consultabile al seguente link (RCSB): <https://www.rcsb.org/docs/additional-resources/molecular-graphics-software>.

- Mol*, frutto di una collaborazione aperta avviata da PDBe e RCSB PDB per fornire uno stack tecnologico di strumenti di analisi di macromolecole; è usato sul sito del PDB ed è possibile anche caricare i propri file
- Jmol, software open-source scritto in Java
- PyMOL
- QuteMol, sviluppato anche da Paolo Cignoni (ISTI-CNR)
- Chimera
- RasMol

Tipi di rappresentazioni grafiche

I principali tipi di rappresentazione grafica sono:

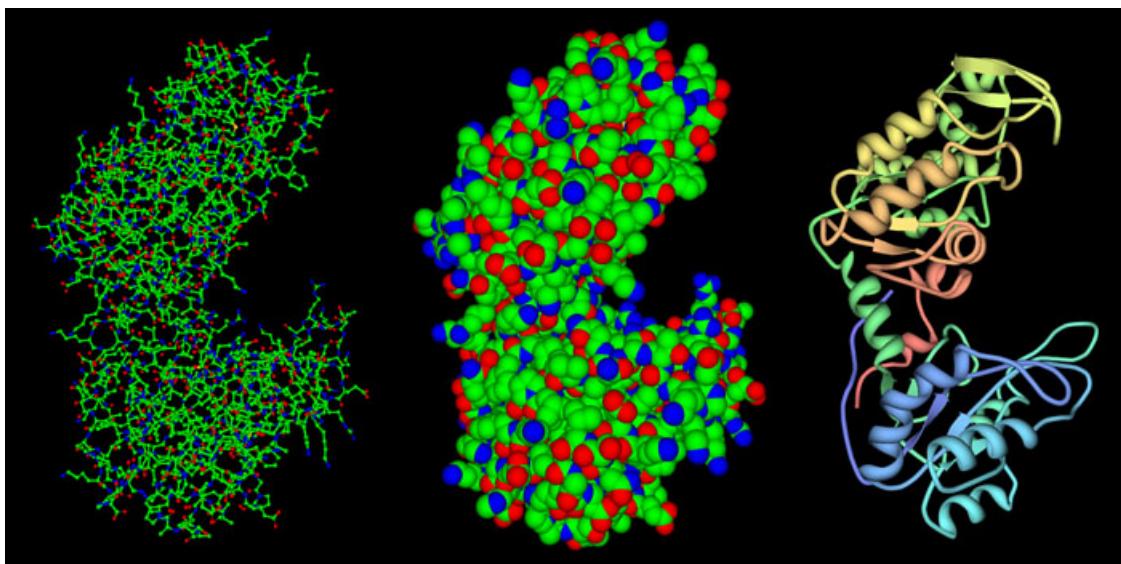


Figura 4.7: Da sinistra a destra, rappresentazioni: wire-frame, space-fill, ribbon. Fonte[92]

- *ribbon*, viene mostrata solo la backbone della proteina, rappresentata come un nastro che contiene tutti gli atomi (in realtà spesso solamente gli atomi C_α). Mette in luce sia il ripiegamento che le strutture secondarie della proteina; è una delle rappresentazioni più utilizzate
- *wire-frame*, vengono rappresentati solo i legami covalenti tra gli atomi, tracciando una linea per ciascuno dei legami. In molti casi viene utilizzata la rappresentazione *ball-and-stick* (asta e sfera) per rendere più facile la comprensione della forma tridimensionale. Rivela la connettività degli atomi nella proteina ma può sommergere il ricercato con una quantità eccessiva di dettagli

- *space-fill*, gli atomi vengono mostrati come sfere, la cui grandezza è spesso relativa al loro raggio di van der Waals. La rappresentazione è chiara e colorata, fornisce la forma generale della proteina ma non fornisce altre informazioni aggiuntive, perciò viene raramente usata nel contesto scientifico
- *surface*, utile per interpretare l'interazione della proteina con l'ambiente esterno

Ogni approccio mette in risalto informazioni differenti, e una rappresentazione combinata può risultare spesso la scelta vincente per mostrare più dettagli in una singola immagine.

È possibile anche colorare la rappresentazione sulla base di determinate caratteristiche, come il potenziale elettrostatico, la conservazione evolutiva, il grado di flessibilità della catena o l'appartenenza a una delle possibili catene polipeptidiche. Ad esempio la rappresentazione della superficie colorata in base al potenziale elettrostatico è di aiuto ai ricercatori nella scoperta di siti di legame della proteina.

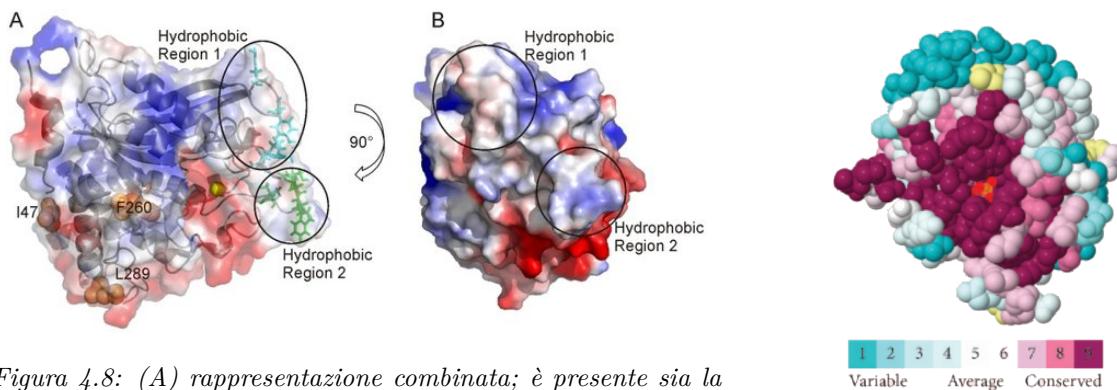


Figura 4.8: (A) rappresentazione combinata; è presente sia la rappresentazione della superficie che ribbon che ball-and-stick. (B) rappresentazione della superficie colorata in base al potenziale elettrostatico (i potenziali negativi sono rossi, quelli potenziali positivi blu e i potenziali neutri bianchi). Fonte: [95]

Figura 4.9: Rappresentazione space-fill colorata in base alla conservazione evolutiva. Fonte [6]

4.1.6 CASP ed excursus storico

Dal 1994 il campo del PSP è stato stimolato, monitorato e quantitativamente valutato dalla competizione biennale CASP (*Critical Assessment of Structure Predictions*). CASP è una sfida nella quale gruppi di ricerca si sfidano cercando di realizzare predizioni di strutture di proteine. È nota la sequenza amminoacidica di queste proteine target ma non la struttura sperimentale. Queste sequenze di proteine provengono da laboratori congiunti: è pianificata la determinazione delle loro strutture native in vitro, che verrà infine utilizzata per stabilire l'accuratezza dei metodi in gara.

Questi esperimenti a livello di comunità sono cresciuti significativamente nel corso delle edizioni: dai 33 target e i 100 modelli sottomessi nel 1994 (CASP1) agli 82 target e più di 55.000 modelli sottomessi nel 2018 (CASP13)^[83].

Ogni due anni un insieme di sequenze di proteine sono rilasciate gradualmente nel corso di un paio di mesi, durante i quali gruppi di ricerca da tutto il mondo tentano di predire le loro strutture 3D e inviano i loro modelli (fino a 5 per target).

Nei primi 6 round (CASP1-6), i target erano classificati in 3 categorie: *comparative modeling*, *fold recognition* e *ab initio*. Da allora i target sono stati classificati solo in 2 classi: le famose *template-based modeling* e *template-free modeling*. I target nella categoria TBM sono considerati "facili" mentre quelli in FM sono considerati difficili. È presente anche la divisione fra metodi completamente automatici e previsioni ottenute usando intervento umano.

È interessante notare che nei primi round del CASP, la previsione della struttura secondaria era una categoria separata. Questa categoria è stata cancellata dopo che gli organizzatori hanno notato che i vincitori di tale categoria utilizzavano un approccio circolare: prevedevano la struttura 3D e utilizzavano la struttura del modello per decifrare gli elementi della struttura secondaria.

I risultati sono valutati sulla base di vari criteri, come il numero di residui la cui posizione è stata predetta con un certo livello di accuratezza, identificazione di strutture secondarie, limiti dei domini, contatti fra residui, regioni disordinate, ecc.

Analizzando i risultati del CASP negli anni si può notare che i metodi basati su un approccio fisico, nonostante l'aumento della potenza computazionale, hanno subito solo lievi miglioramenti.

L'importanza del CASP per la biologia strutturale è alta anche per il contributo che ha dato alla creazione dei metodi automatici, accessibili anche ai non esperti.

Excursus storico

Dal punto di vista biologico e sperimentale sulle proteine, tra i punti di riferimento più importanti ci sono^[7]:

- nel 1838 è stato proposto il nome *proteina*
- durante il XIX secolo vengono scoperti la maggior parte dei 20 amminoacidi
- nel 1864 è stata cristallizzata e denominata l'emoglobina
- nel 1894, come già citato, Fischer ha proposto l'analogia chiave-lucchetto per le interazioni fra enzimi e substrato
- nel 1897 vengono gettate le basi per l'enzimologia
- nel 1926 viene dimostrato che le proteine possono essere enzimi e viene sviluppata l'ultracentrifugazione, usata per stimare il peso molecolare dell'emoglobina

- negli anni '30 vengono proposte le prime teorie sulla denaturazione delle proteine
- nel 1933 viene introdotta l'elettroforesi
- nel 1934 viene mostrata la prima diffrazione a raggi-X di una proteina
- nel 1942 viene sviluppata la cromatografia
- nel 1951 vengono proposte le strutture secondarie α -eliche e foglietti- β
- nel 1955 viene sequenziata la prima proteina (insulina)
- nel 1956 viene prodotta la prima "impronta digitale" di una proteina, mostrando la causa dell'anemia falciforme
- nel 1960 viene descritta la prima struttura tridimensionale tramite cristallografia a raggi-X a risoluzione di 2 Å
- nel 1966 viene descritta la struttura tridimensionale del lisozima, il primo enzima ad essere analizzato a risoluzione atomica
- nel 1975 Henderson e Unwin determinano la prima struttura 3D di una proteina di membrana usando una ricostruzione al computer da microografie elettroniche
- nel 1983 viene usato la spettroscopia NMR per determinare la struttura 3D di una proteina
- nel 1988 vengono sviluppati metodi per l'uso della spettrometria di massa nell'analisi di proteine e altre macromolecole
- nel 2003 è stato completato il progetto genoma umano
- tra il 1996-2013 vengono rifinati i metodi nell'uso della spettrometria di massa per identificare proteine in miscugli complessi
- tra il 1975-2013 Henderson e altri riconoscono i metodi nell'utilizzo della cryo-EM per la determinazione della struttura di larghi complessi proteici a risoluzione atomica

I primi passi della *bioinformatica* risiedono nei primi anni '60, quando ancora i desktop computer erano solo un'ipotesi e il DNA non poteva essere sequenziato. Questi passi andavano in direzione di metodi computazionali per l'analisi della sequenza delle proteine. Margaret Dayhoff è considerata la madre della bioinformatica: ha infatti sviluppato il primo software bioinformatico (COMPRESS) e il primo database per le sequenze biologiche (*Atlas of Protein Sequence and Structure*, conteneva 65 sequenze di proteine). Il codice a una lettera per gli amminoacidi si deve sempre alla Dayhoff^[96].

È in questi anni che è stato sviluppato il primo modello di sostituzione degli amminoacidi per la filogenetica. Anche l'approccio *ab initio* (non per le proteine) è emerso negli anni '60 a partire dal campo della chimica computazionale, grazie principalmente a Warshel, Levitt, Karplus⁸. Il primo programma per calcolare l'energia potenziale nelle proteine è stato sviluppato nel 1969 da Lifson e Levitt^[97].

⁸Nel 2013 il premio Nobel per la chimica è stato assegnato proprio a questi scienziati che hanno contribuito sin da quegli anni al campo della biofisica molecolare computazionale.

Le predizioni di strutture basate sui metodi *ab initio* sono emerse nella metà degli anni '80, prima per piccoli peptidi e poi per polipeptidi. La prima simulazione di MD su una proteina è stata realizzata nel 1977 da McCammon, Gelin e Karplus^[98], studiando la dinamica di ripiegamento di una proteina di 58 amminoacidi rappresentata esplicitamente ma simulata nel vuoto. Questo studio seguì il lavoro pionieristico di Levitt e Warshel del 1975 (*Computer simulation of protein folding*^[99]) sulla stessa proteina che era però rappresentata in modo più semplicistico: ogni amminoacido era rappresentato da due sfere.

Intorno all'inizio degli anni '90 è emerso dalla biologia il campo della bioinformatica, sulla base dall'analisi delle proteine, come dimostra anche l'instaurazione del CASP.

Nel 2007 si è raggiunta un'accuratezza nella predizione della struttura di proteine (a singolo dominio con massimo 90 residui) tra i 2 e i 6 Å(modellazione per omologia)^[52].

Nel 2020 AlphaFold è stato reputato vincitore del CASP riuscendo praticamente a risolvere il problema del PSP (per determinate categorie di proteine).

4.2 Annotazioni 1D sulla struttura

Le "annotazioni monodimensionali sulla struttura delle proteine" (1D PSA) sono astrazioni che descrivono la disposizione della backbone proteica.

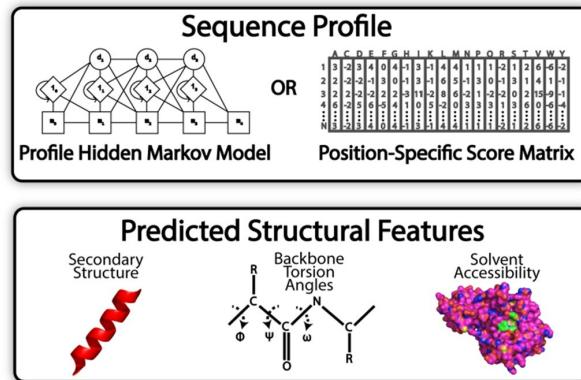


Figura 4.10: Annotazioni 1D nella pipeline generica di un metodo odierno per il PSP. Fonte^[85]

Le annotazioni monodimensionali delle sequenze possono essere legate a due tipi di proprietà:

- proprietà esplicitamente correlate con la struttura degli amminoacidi
- proprietà statistiche

Annotazioni del 1° tipo sono collegate a: formazione di determinate strutture secondarie, esposizione al solvente circostante (*solvent accessibility*), determinati angoli di torsione, interazioni con alcuni amminoacidi, ecc. (es. GenTHREADER, SPARKS-X).

Le proprietà statistiche vengono invece identificate tramite le frequenze degli amminoacidi in vari MSA. Un modo di rappresentare queste tendenze è tramite *profilo*, i quali denotano amminoacidi che frequentemente appaiono in certe posizioni dell'allineamento. Un profilo del genere è più informativo di una semplice sequenza amminoacidica: contiene informazioni evolutive (es. livelli di conservazione di specifiche posizioni nella sequenza). Il profilamento della sequenza avviene tipicamente tramite HMM (Hidden Markov Model) o PSSM (Position-specific score matrix).

Dato che tutti gli step successivi, in ogni caso (compresi metodi end-to-end), dipendono dalla qualità dell'MSA, la generazione di allineamenti di alta qualità è di incredibile importanza per il PSP. A prescindere dall'approccio e dall'utilizzo o meno del DL, sfruttare le informazioni evolutive è risultato essere lo step più importante.

4.2.1 Allineamento di sequenze

Un allineamento di sequenze multiple (MSA) è una disposizione di più di due sequenze di amminoacidi o nucleotidi allineate in modo da posizionare i residui delle diverse sequenze in colonne verticali in una maniera appropriata. I metodi di MSA sono utilizzati per l'analisi del proteoma e del genoma; sono il passo iniziale essenziale nella maggior parte dei confronti filogenetici. Sono ampiamente utilizzati per aiutare a ricercare caratteristiche comuni nelle sequenze e possono essere usati per aiutare a prevedere le strutture bi e tridimensionali di proteine e acidi nucleici. Un MSA, in essenza, mostra il grado di similarità evolutiva tra le sequenze.

In genere si assume che le sequenze che si vuole allineare siano filogeneticamente correlate e, quindi, omologhe. In questo caso, l'allineamento ideale avrà residui omologhi allineati nelle colonne.

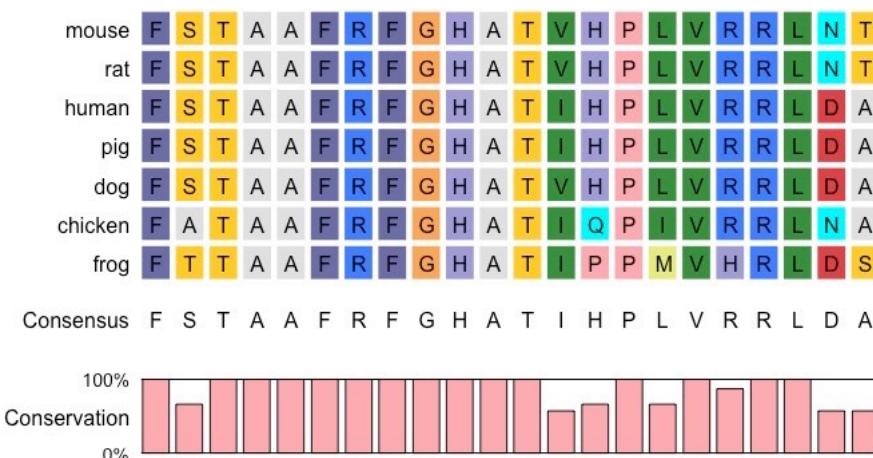


Figura 4.11: Multiple sequence alignment schematica di una sequenza proteica di varie specie. Fonte[100]

Le sequenze di proteine in genere risultano avere un grado di somiglianza maggiore rispetto alle sequenze nucleotidiche. Ciò è dovuto alla degenerazione del codice genetico e al fatto che quasi per ogni amminoacido esistono vari codoni che lo codificano, perciò differenti sequenze nucleotidiche possono codificare esattamente la stessa sequenza di amminoacidi.

Un allineamento di sequenze multiple (MSA) può essere utilizzato per tracciare l'entità della divergenza evolutiva tra sequenze correlate. Rispetto a una singola sequenza, l'MSA fornisce informazioni sulle tendenze evolutive degli amminoacidi in ciascuna posizione della sequenza, il che aiuta a caratterizzare un "profilo" della sequenza.

Un modo per generare un MSA è cercare grandi data set di sequenze proteiche e allinearli alla sequenza target. All'interno di un metodo per generare un MSA vengono utilizzati vari metodi per massimizzare i punteggi e la correttezza degli allineamenti. Ogni metodo utilizza un'euristica che cerca di replicare il processo evolutivo e ottenere un allineamento realistico. Ci sono vari tool per creare MSA, ad esempio: ClustalΩ, MAFFT, MUSCLE, T-Coffee.

4.3 Annotazioni 2D sulla struttura

Le annotazioni bidimensionali sulla struttura delle proteine (2D PSA) permettono di ricavare restrizioni spaziali come contatti inter-residuo a lunga distanza, distanze specifiche o legami idrogeno.

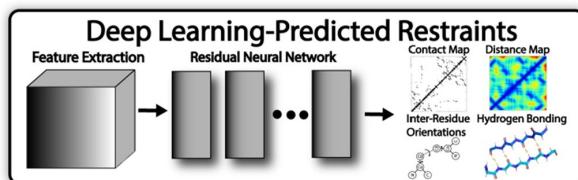


Figura 4.12: Annotazioni 2D nella pipeline generica di un metodo odierno per il PSP. Fonte[85]

Le annotazioni si basano sulla predizione dei contatti fra i residui sulla base, ad esempio, di informazioni co-evolutive (*correlated mutation*), generando o una *contact map*, o un *distogramma* o una *distance map*.

I progressi nella previsione dei contatti sono rimasti stagnanti per qualche tempo. Tuttavia, si è verificato un balzo nell'accuratezza della previsione dei contatti quando gli algoritmi hanno iniziato a utilizzare approcci di previsione globale. I primi metodi prevedevano principalmente i contatti tra le coppie di residui uno alla volta utilizzando tecniche come l'informazione reciproca, ignorando così le interazioni con altre coppie di residui e il contesto globale in cui si verificavano le interazioni; questo è in gran parte il motivo

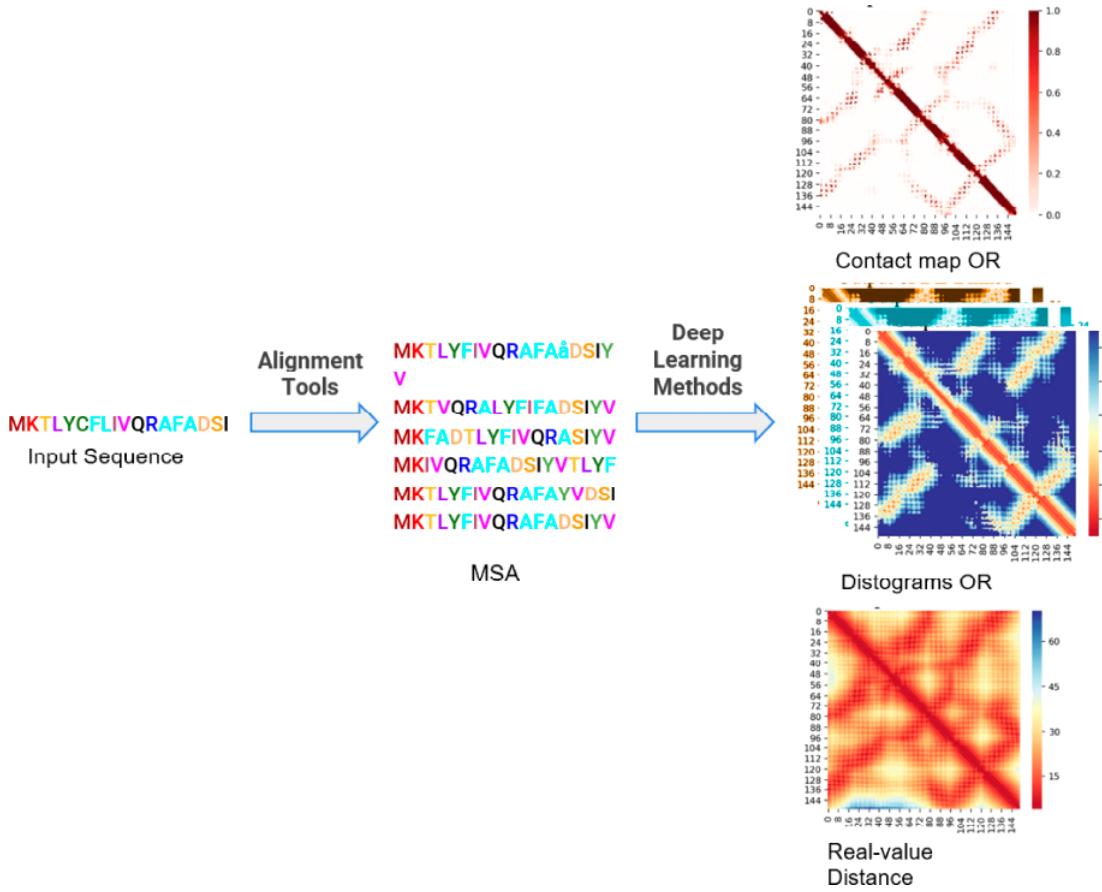


Figura 4.13: Schema generali di contact prediction basata su DL. Fonte[73]

per cui era difficile per questi metodi locali distinguere tra interazioni dirette e indirette. L'introduzione di modelli statistici globali determinati attraverso l'uso dell'analisi di accoppiamento diretto (DCA) è stata in grado di distinguere con maggiore successo tra queste interazioni dirette e indirette. Il DL ha successivamente rivoluzionato il campo della predizione dei contatti sin dal suo debutto nel CASP10, facendo progredire negli ultimi questo campo dalla predizione binaria dei contatti alla predizione della distanza a valori reali o probabilistica (distogrammi); spesso vengono utilizzate delle *deep residual neural network* (*ResNet*).

Data la natura "black-box" dei metodi di DL è stata ad esempio proposta InterPreT-ContactMap che fa riferimento al campo della xAI (explainable AI); combina reti neurali profonde con meccanismi di attenzione per aumentare la comprensibilità della predizione dei contatti.

È dal CASP11 che la predizione di una *contact map* è diventata una chiave fondamentale della pipeline dei metodi per il PSP. La predizione dei contatti può essere classificata in due categorie:

- correlated-mutation-based
- ML-based

Le *contact map* (descritte nella sez. 4.1.3) possono essere divise in 4 categorie a seconda di quanto siano lontani fra loro i residui nella sequenza:

1. local, <6 residui
2. short-range, 6-11 residui
3. medium-range, 12-23 residui
4. long-range, >24 residui

I contatti locali descrivono in realtà le strutture secondarie, pertanto i contatti che risultano essere più utili sono quelli a medio e lungo raggio.

4.3.1 *correlated mutation*

L'apparizione di mutazioni nelle sequenze delle proteine durante la loro evoluzione dipende da aspetti sia strutturali che funzionali. È interessante notare come l'analisi delle sequenze di famiglie di proteine mostri che certe posizioni tendono a *coevolvere*. In altre parole l'apparizione di una mutazione in una posizione è accompagnata da una mutazione in un'altra posizione.

È stato suggerito che un tale collegamento possa avvenire in posizioni vicine nello spazio tridimensionale e che i residui considerati interagiscano fra loro. Se una mutazione in una posizione porta a una distruzione della sua interazione con la posizione adiacente, una mutazione compensatoria di quest'ultima potrebbe rimediare al problema.

Ad esempio, se due posizioni erano originariamente polari e coinvolte in un'interazione elettrostatica favorevole, la mutazione di una posizione da polare a non polare distrugerebbe l'interazione. Se però la posizione adiacente muta anch'essa da polare a non polare, l'interazione originale può tramutarsi in un'altra interazione favorevole, stavolta non polare.

Ciò suggerisce che sia possibile inferire posizioni a contatto nella struttura ripiegata di una proteina analizzandone le variazioni nella sequenza attraverso la sua evoluzione e osservando quali posizioni sono *coevolute*. Gli accoppiamenti evoluzionisticamente inferiti possono essere usati come vincoli durante la modellazione.

Ci sono 3 principali ostacoli a questo approccio: rumore statistico (molte correlazioni sono solo frutto di rumore e quindi insignificanti), correlazioni fra residui distanti e numero insufficiente di posizioni correlate. Il metodo non risulta applicabile quando ci sono poche sequenze omologhe. Nonostante l'idea non sia nuova, solo nell'ultimo decennio tali metodi sono divenuti abbastanza accurati da poter essere usati nel PSP; metodi noti sono EVfold, Rosetta-GREMLIN, FILM3, MetaPSICOV, ecc.

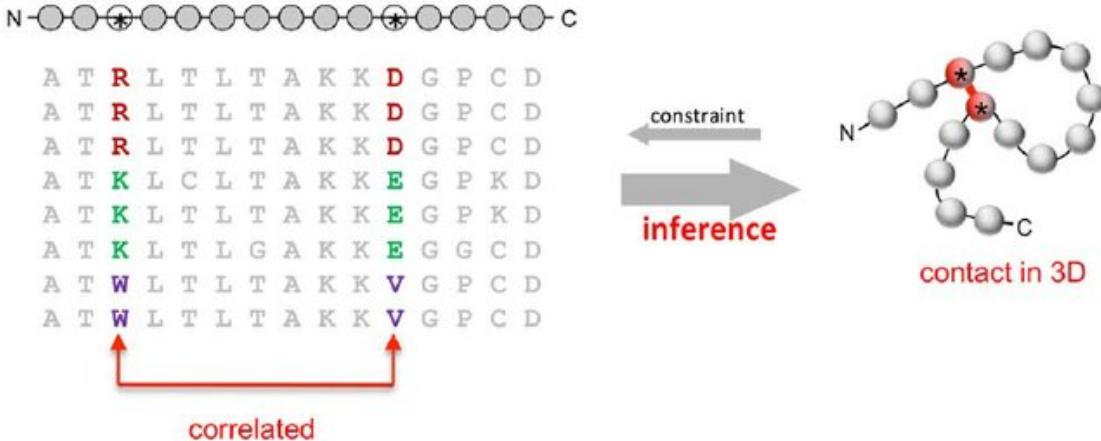


Figura 4.14: In alto, sequenza della proteina da predire, in cui ogni amminoacido è un cerchio. Sotto, un MSA con le sequenze di una famiglia correlata, tutte con lo stesso ripiegamento. A destra l'inferenza del contatto delle posizioni coevolute. Fonte[101]

4.3.2 contact prediction ML-based

La predizione di contatti tra i residui è stata ampiamente utilizzata per oltre un decennio nel campo del PSP. Tuttavia, recentemente, il paradigma si è spostato verso la previsione della probabilità di intervalli di distanza, noti anche come *distogrammi* (distanza + istogramma).

Per una sequenza di lunghezza L , un distogramma è una matrice $L \times L$, che mostra l'istogramma delle distanze tra coppie. I distogrammi sono matrici simmetriche rispetto alla diagonale. Ogni "pixel" sulla mappa rappresenta una distanza tra una coppia di residui nella sequenza. Le distanze in un distogramma sono relative, il che significa che le distanze tra i residui sono invarianti rispetto alle rotazioni e alle traslazioni 3D.

Le distanze sono "binned"⁹, pertanto il distogramma può avere tanti canali quanti sono i *bin*, ovvero $L \times L \times bins$ tensori¹⁰.

La previsione della distanza tra i residui proteici è la previsione di una matrice (2D) di distanze tra coppie a partire da una sequenza proteica (1D). Tale problema può essere confrontato con quello della stima della profondità monoculare in *computer vision* (vedi fig. 4.15).

Nell'*image depth prediction*, viene fornita una matrice dell'immagine come input e viene prevista come output una matrice di profondità in cui ogni pixel ha una profondità prevista

⁹Ciò vuol dire che i valori dei dati originali che rientrano in un dato piccolo intervallo, un *bin*, sono sostituiti da un valore rappresentativo di quell'intervallo, spesso il valore centrale. È una forma di quantizzazione.

¹⁰In matematica, la nozione di tensore generalizza tutte le strutture definite usualmente in algebra lineare a partire da un singolo spazio vettoriale.

(distanza dalla fotocamera all'oggetto). Similmente al problema della previsione della profondità, la previsione della distanza nel PSP prende in input un volume tridimensionale (altezza \times larghezza \times canali, tensore 3D) e genera una mappa della distanza con la stessa dimensione dell'input (altezza \times larghezza) ma con un singolo canale.

Tuttavia, i canali di input nei problemi di visione artificiale vanno da uno a tre (i canali RGB o HSV), ma possono arrivare alle centinaia di canali nel PSP, a seconda delle caratteristiche di input.

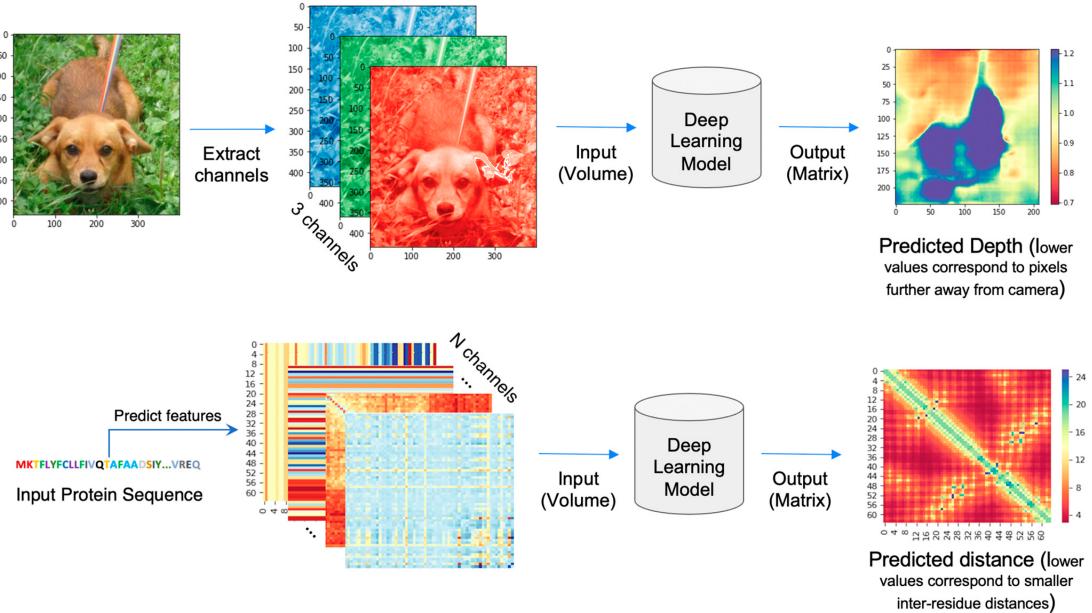


Figura 4.15: Confronto fra "image depth prediction" e "distance prediction". Fonte[73]

Fino al CASP13 i tentativi di predire dei distogrammi erano stati poco efficaci, è solo con il metodo di Xu, trRosetta e AlphaFold1 che tale tecnica è maturata. La tecnica di DL principalmente utilizzata è l'uso di ResNet.

Il metodo di Xu, implementato in Raptor X, discretizza le distanze di interazione C_β - C_β in 25 intervalli (da $< 4.5\text{\AA}$ a $< 16\text{\AA}$). È possibile ottenere una *contact map* da questa *distance map* sommando tutti valori di probabilità previsti corrispondenti a distanze $\leq 8\text{\AA}$ [73].

La predizione delle distanze a valori reali è un task ancora più difficile. Consiste nel predire le esatte distanze fisiche sull'intera *distance map* accuratamente. I metodi odierni utilizzano ResNet o GAN o meccanismi di attenzione; degli esempi sono PDNET e RealDist.

4.4 Predizione della struttura 3D

4.4.1 *homology modeling*

Nella modellazione per omologia ci si affida a somiglianze nella sequenza tra la proteina target e i template. I metodi per omologia sono perciò basati sul paradigma:

«*la sequenza codifica per la struttura*». Due proteine si definiscono *omologhe* quando hanno un progenitore comune nella loro storia evolutiva; una notevole somiglianza fra due sequenze è una forte evidenza che le queste siano correlate evolutivamente.

Sono metodi basati anche sull'osservazione che la struttura terziaria è più conservata della sequenza amminoacidica. Di conseguenza ci si aspetta una significativa similarità nella struttura fra proteine che condividono una notevole somiglianza tra le sequenze.

In altri termini, due sequenze amminoacidiche molto simili (*omologhe*), in due proteine differenti ma evolutivamente collegate, dovrebbero acquisire la stessa struttura locale.

Un approccio che utilizzi la modellazione per omologia consiste tipicamente nei seguenti passi:

1. ricerca e selezione del template
2. costruire un MSA (Multiple Sequence Alignment) che includa la proteina target e i template
3. assegnare le coordinate spaziali dei template alla sequenza della proteina target
4. raffinamento della struttura modello
5. valutazione e validazione della struttura risultante

Nel 1° step si cerca (almeno) una struttura modello tra le strutture conosciute, avente un'alta somiglianza di sequenza. È più semplice se la struttura di una proteina omologa molto simile è stata già risolta. Ci sono però alcuni gruppi di proteine, come le proteine di membrana, le cui strutture risolte sono scarse. Trovare i giusti template e caratterizzarne la loro omologia è ciò che determina in genere il successo dell'intera predizione (una somiglianza minore del 30% avrà risultati molto scarsi, mentre sopra al 50% la predizione ha buona probabilità di essere di buon livello). È possibile in ogni caso che vi sia una somiglianza *locale* anche quando la somiglianza globale è scarsa.

Nel 2° step vengono sfruttate informazioni evoluzionistiche per migliorare l'allineamento tra le sequenze dei template e del target. È difficile stabilire allineamenti fra omologhi distanti, come nel caso di target *eucarioti* e template *procaristi*. Altre tecniche usate oltre l'MSA sono: programmazione dinamica, threading, HMM ecc. Possono venire utilizzati più MSA insieme per sopprimere a problemi di disallineamento di piccole regioni.

Nel 3° step ad ogni segmento della sequenza target viene assegnato un insieme di coordinate spaziali in accordo ai risultati del MSA. Tool noti sono MODELLER (soddi-

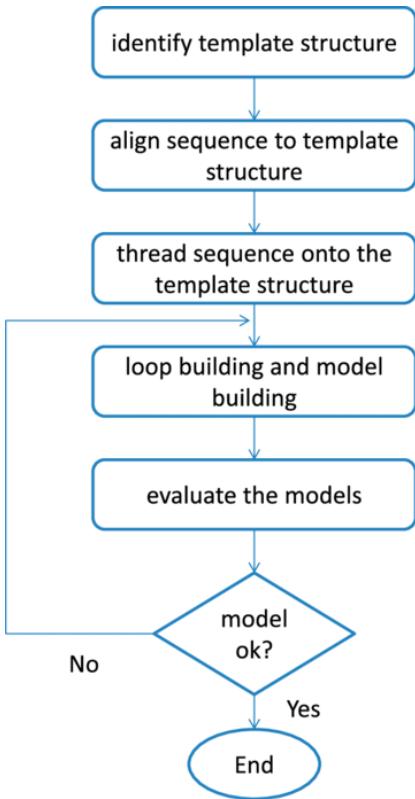


Figura 4.16: Diagramma di flusso della modellazione per omologia. Fonte: [102]

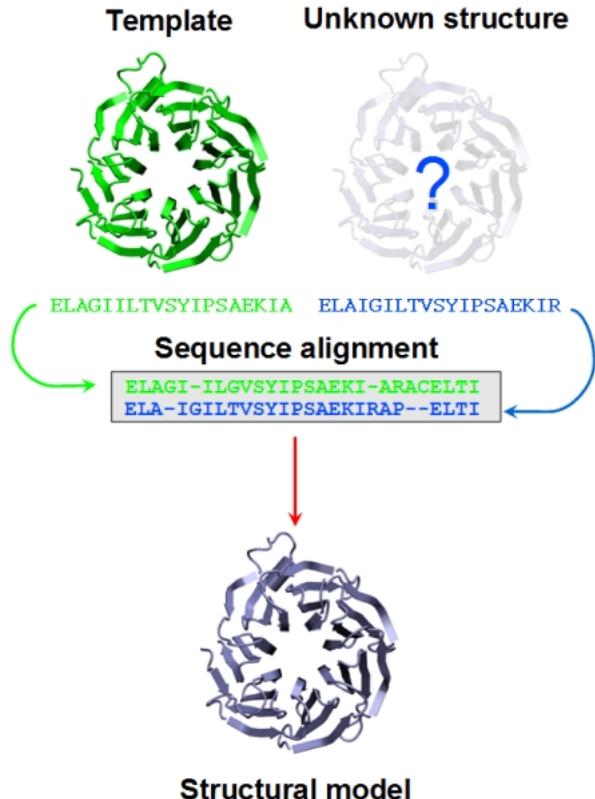


Figura 4.17: Schema esemplificativo di una modellazione per omologia. Fonte [103]

sfazione di vincoli spaziali), NEST, COMPOSER e SWISS-Model. La struttura ottenuta potrebbe essere però deformata a causa dell'utilizzo di più template e numerosi inserzioni e cancellazioni. Possono essere presenti lunghezze e angoli dei legami non ottimali e atomi sovrapposti.

Per ovviare a tali problemi nello step 4 si applica un processo di raffinamento, specialmente per quanto riguarda i loop (*loop modeling*, vedi sez. 4.4.5) e i residui (*side chain modeling*). Vengono applicati algoritmi che confrontano caratteristiche geometriche ed effettuano calcoli energetici che identificano configurazioni atomiche sfavorevoli.

Nel 5° step si valuta l'affidabilità della predizione. Si dice che un modello è affidabile quando è basato su un template corretto e un allineamento approssimativamente corretto. Si può valutare tale affidabilità in vari modi:

- alcune qualità della struttura costruita possono essere confrontate con delle tendenze statistiche
- se ci sono vari modelli predetti si calcola l'energia libera e si sceglie la struttura con minor energia libera (ad es. tramite ProSa)

- stereochimica (relativa alle proprietà spaziali delle molecole), ad esempio con PRO-CHECK
- la conservazione evoluzionistica a livello amminoacidico può essere correlata con il loro stato "esposto" o "seppellito" (l'idea di partenza è che il nucleo della proteina rimanga inalterato e la superficie sia variabile), ad esempio Profiles3D
- se si hanno a disposizione dati sperimentali della struttura nativa della proteina si può validare il modello con la consistenza a essi

Efficienza e limiti

Con una somiglianza maggiore del 50% si registra una RMSD tra 1 e 2 Å, ma è importante notare che non sempre proteine omologhe (vicine sequenzialmente) condividono la stessa funzione e struttura. Un esempio sono le proteine del lievito Gal1 e Gal3: 73% di identità e 92% di somiglianza. Queste due proteine hanno però sviluppato differenti funzioni, con Gal1 che è una galattochinasi mentre Gal3 è un induttore trascrizionale^[104].

Non c'è quindi una soglia che assicuri una sicura predizione della struttura: molte proteine con una lontana somiglianza possono svolgere la stessa funzione mentre altre altamente simili possono svolgere funzioni diverse. Una regola empirica è considerare sequenze con più del 30-40% di somiglianza come sequenze con una struttura o funzione simile.

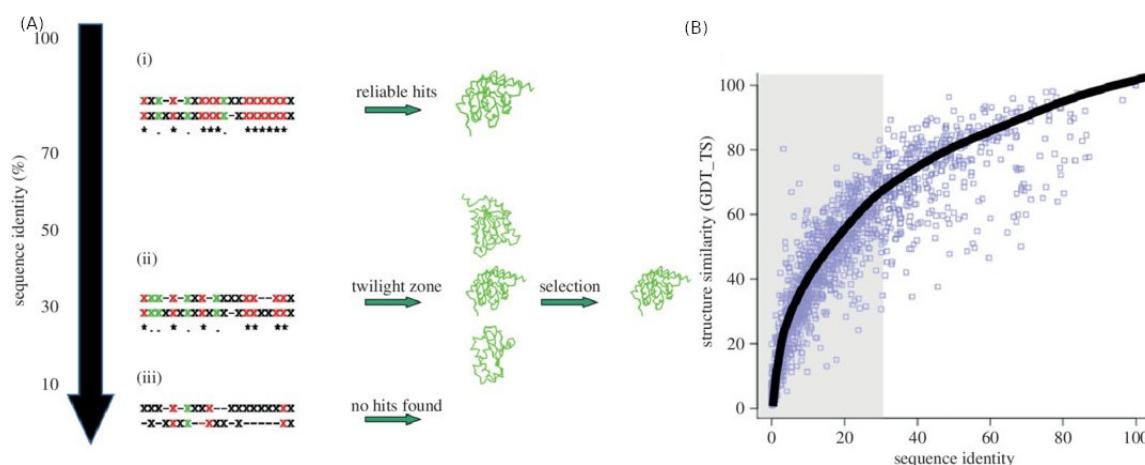


Figura 4.18: Risultati dei metodi per omologia alla variazione dell'identità nella sequenza. (A) Dimostrazione schematica dell'uso di metodi di allineamento di sequenze per identificare template. 'X' indica un qualsiasi amminoacido. (i) identità > 70%: semplici allineamenti di sequenza sono sufficienti per trovare il corretto ripiegamento. (ii) Tra il 20 e il 30% non sempre è possibile trovare il corretto ripiegamento; è necessario effettuare ulteriori raffinamenti. (iii) a bassi livelli l'utilità di questo metodo è molto bassa. (B) Somiglianza strutturale (in GDT_TS) al variare dell'identità della sequenza. Anche al 30% il livello di somiglianza è significativo. Fonte[105]

Un'osservazione fondamentale risiede sulle basi in sé del metodo: dato l'affidamento pressoché totale nella modellazione comparativa, la struttura modello è condizionata necessariamente a essere più simile ai template che alla reale struttura nativa della sequenza target, nonostante i vari processi successivi di raffinamento che, data la loro natura approssimativa, non sono perfetti.

I problemi maggiori risultano nelle regioni con bassa somiglianza, come ci si può aspettare. Si sta parlando specialmente dei *loop*, soggetti a mutazioni considerevoli durante l'evoluzione.

Si incorrono in problemi con la modellazione per omologia quando si trattano proteine che non hanno omologhe tra le strutture conosciute, come le proteine di membrana, le quali sono difficili da cristallizzare (anche se, come spiegato nella sezione 3.4, con la microscopia crioelettronica sta diventando possibile determinare le loro strutture).

Nonostante tutte le osservazioni fatte, *la modellazione per omologia, quando possibile, è correntemente il miglior metodo computazionale per predire la struttura delle proteine* e la sua applicabilità è destinata a crescere con l'aggiunta di nuove strutture determinate sperimentalmente da poter essere usate come template.

Oltre alla PSP i metodi per omologia sono anche usati nel drug design (per studiare le differenze strutturali fra le proteine bersagliate dallo stesso farmaco) e nello studio dei meccanismi catalitici.

4.4.2 *fold recognition*

Come si è già detto la struttura delle proteine è maggiormente conservata rispetto alle sequenze. Questo significa che proteine con differenti sequenze possono ancora formare strutture simili grazie a certe proprietà condivise codificate nelle loro sequenze. Identificando queste proprietà è possibile predire la struttura di una nuova proteina basandosi su un template che condivide le stesse proprietà, anche se loro sequenze sono diverse. Questa è l'idea su cui si basano i metodi di *fold recognition*. Le proprietà di cui si parla sono principalmente le annotazioni già calcolate precedentemente. In altri termini, in questo approccio si cerca una proteina con struttura conosciuta (nel PDB) che abbia alcune proprietà nella sequenza o tendenze condivise con la proteina target: le due probabilmente hanno un ripiegamento o motivi strutturali simili.

È possibile dividere il *fold recognition* in vari livelli a seconda della distanza di omologia che può intercorrere fra la proteina target e il template (come mostrato in fig. 4.19). Analizzando questa figura si possono descrivere i livelli di ricerca in un tipico algoritmo di *fold recognition*:

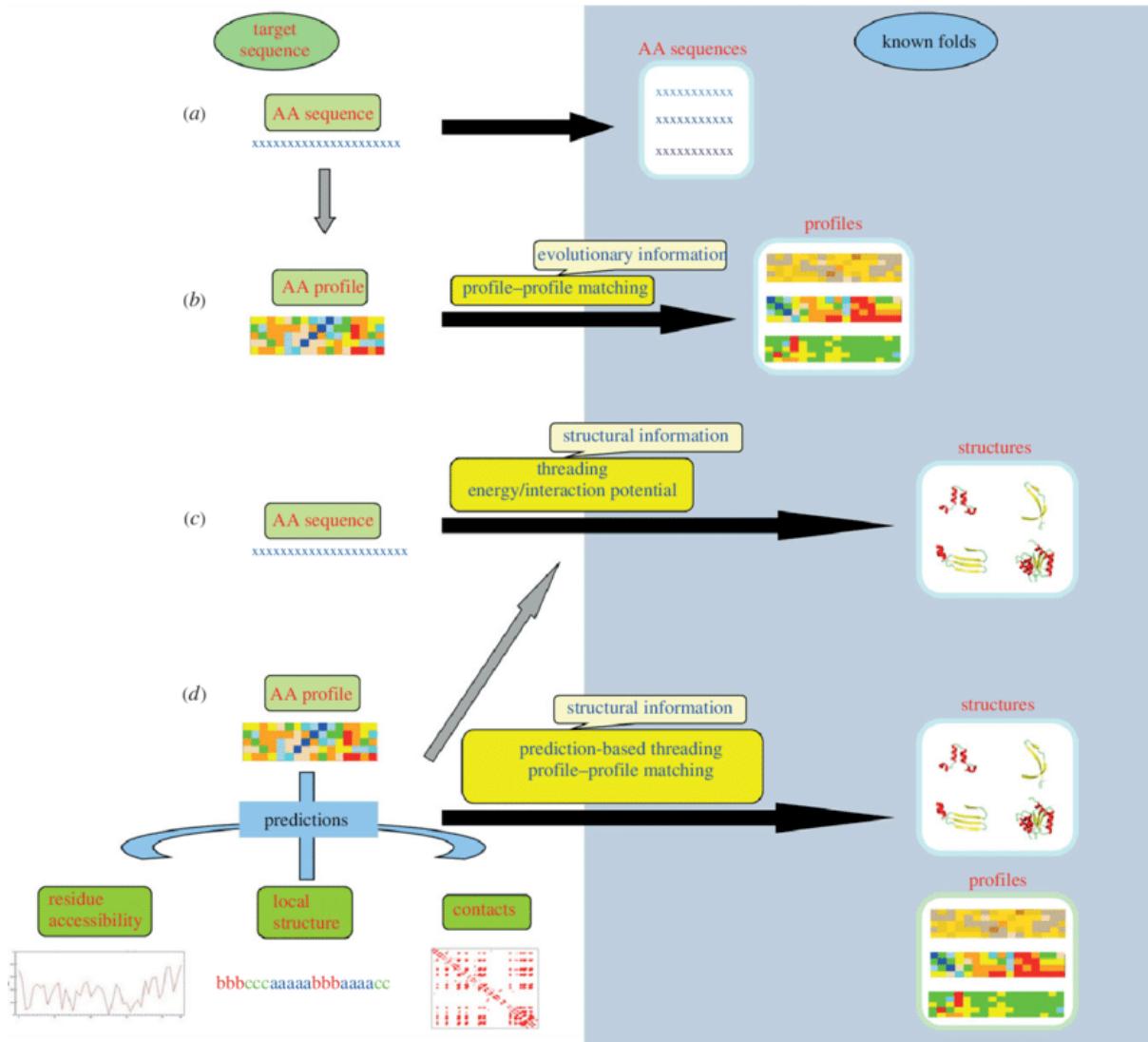


Figura 4.19: Differenti strategie per il fold recognition. A destra, su sfondo blu, c'è l'insieme delle proteine con ripiegamenti conosciuti. La lunghezza delle frecce è correlata alla distanza di relazione fra target e template. Fonte[105]

- (a) *ricerca per sequenza*, ovvero la ricerca iniziale di sequenze omologhe (relazione molto vicina)¹¹
- (b) *profile-profile matching*, vengono aggiunte informazioni evolutive usando il profilo generato nelle annotazioni 1D
- (c) *sequence-structure matching*, la sequenza target può essere allineata con strutture di proteine conosciute per valutare la compatibilità (*threading*)
- (d) *annotation search*, viene effettuata una ricerca confrontando le annotazioni 1D

¹¹Come accennato nella sezione sulla classificazione dei metodi, il confine tra le varie tecniche è sempre più sfumato. Infatti *homology modeling* e *fold recognition* sono spesso usati insieme come in questo caso, e può essere di aiuto pensare a tale metodo combinato come un *data-based modeling*.

e 2D precedentemente calcolate con le annotazioni di strutture conosciute

Nel *profile-profile matching* le proteine i cui profili sono sufficientemente simili alla proteina target possono essere usate come template globali. Un esempio di procedura profile-matching è HHpred, basata su confronti tra coppie di profili HMM target e template, e JACKHMMER. Alcuni studi riportano che i metodi di confronto fra profili basati su HMM siano più efficienti dei profili basati PSSM^[105].

Nel *sequence-structure matching* la compatibilità può essere quantificata in base all’interazione globale o al potenziale di energia. È bene notare che la procedura di *threading* può essere difficile e computazionalmente costosa.

4.4.3 *ab initio*

Il metodo più lineare e a prima vista ovvio per predire la struttura nativa di proteine è seguire la natura, simulando accuratamente come le forze fisiche guidino la proteina a ripiegarsi e usare questa simulazione per riprodurre il processo di ripiegamento su proteine con strutture sconosciute. *Ab initio*, termine latino, significa infatti ”dall’inizio”. Questo approccio si basa sul postulato di Anfinsen.

Il primo problema che sorge è superare il paradosso di Levinthal. Per farlo si assume un profilo energetico a imbuto del ripiegamento, ovvero la premessa termodinamica che la forma nativa di una proteina sia lo stato in cui risulta avere più bassa energia libera, o più precisamente (richiamando la definizione di struttura nativa data nella sez. 3.3.1) quella conformazione avente minore energia libera tale da mantenere il livello di dinamicità richiesto alla proteina per svolgere la sua funzione biologica.

Le predizioni nell’approccio *ab initio* sono pertanto *energy-based*, ovvero guidate dall’idea di minimizzare l’energia. Si può vedere il PSP secondo l’approccio *ab initio* come un problema di ottimizzazione dove una funzione di energia gioca il ruolo di *euristica* cercando di raggiungere il minimo globale di energia all’interno dello spazio di ricerca. In quanto *energy-based* usano solo informazioni sul tipo di atomi nel sistema, le loro posizioni relative nello spazio tridimensionale e le loro interazioni con gli altri atomi. Viene poi calcolato l’intero contenuto di energia del sistema e le forze agenti su ogni atomo.

L’energia totale di un sistema (*free energy*) può essere decomposta in varie componenti: cinetica, potenziale, termica ecc. È l’energia libera che determina la stabilità del sistema. Come si vedrà sotto, nell’approccio fisico non viene calcolata tutta l’energia libera ma viene approssimata con una sua parte per motivi di complessità.

Sebbene vi siano differenti metodi in questo approccio, tutti condividono due caratteristiche di base:

- calcolano il contenuto di energia del sistema in una singola configurazione

- campionano numerose configurazioni e ne trovano una con la minor energia libera

Per *configurazione* si intende la disposizione complessiva di tutti gli atomi di tutti i componenti del sistema (proteina, solvente, ioni, membrana ecc.) mentre la posizione collettiva dei soli atomi della proteina viene chiamata *conformazione*.

Molecular mechanics & dynamics

Per descrivere in maniera affidabile tutte le forze fisiche operanti sul sistema tra i differenti atomi bisognerebbe descriverne la distribuzione di tutti gli elettroni, il che richiede però calcoli di meccanica quantistica (QM). Le forze, in un sistema molecolare, risultano dalla distribuzione spaziale degli elettroni attorno agli atomi. Sfortunatamente questi calcoli sono computazionalmente molto costosi e una rigorosa caratterizzazione di un sistema macromolecolare, con milioni di atomi, è al momento insostenibile. Calcoli di QM su una singola conformazione di una piccola proteina possono richiedere mesi, tempi troppo lunghi se si ha l'obiettivo di provare tante configurazioni per sceglierne una finale.

Molecular mechanics

Per le ragioni sopra elencate gli scienziati spesso investigano sistemi macromolecolari usando approssimazioni delle reali forze in essi. Il campo da cui i calcoli per le approssimazioni sono presi è chiamato *molecular mechanics* (MM), poiché approssima sistemi molecolari usando espressioni prese dalla meccanica newtoniana classica:

- il contenuto di energia è descritto usando un nel quale gli atomi e i legami covalenti sono trattati come palline e molle
- le descrizioni che richiederebbero calcoli di QM vengono ignorate
- le rappresentazioni sono *esplicite*: prendono in considerazione tutti gli atomi (vedi fig. 4.20)

Il campo di forza sopra accennato descrive l'energia potenziale del sistema. Da notare che l'energia potenziale (intesa come entalpia) è solo una delle due componenti dell'energia libera, vedi sez. 3.3.1).

Un campo di forza è un'energia di posizione: l'energia di un oggetto in una specifica posizione all'interno di un campo (gravitazionale, elettrico, magnetico ecc.). Nelle molecole l'energia potenziale è la somma di tutti gli effetti dei campi elettrici atomici¹² in una determinata posizione. Si può approssimare l'energia potenziale all'energia risultante da tutti i legami covalenti e le interazioni non covalenti, escluse quelle non polari¹³, in una

¹²Gli atomi possiedono, in base alla loro eventuale carica, campi elettrici che influenzano gli altri atomi.

¹³Un esempio di interazione non polare è l'effetto idrofobico. Vengono escluse poiché coinvolgono principalmente cambiamenti di entropia nel solvente.

singola configurazione del sistema. In genere i campi di forza non sono una singola funzione ma una somma di più termini, ognuno corrispondente a un differente tipo di legame chimico o interazione, un esempio:

$$U_{tot} = U_{cov} + U_{elst} + U_{vdw}$$

dove per U_{tot} si intende l'energia potenziale totale, per U_{cov} l'energia potenziale dei legami covalenti, per U_{elst} quella delle interazioni elettrostatiche e per U_{vdw} quella delle interazioni di van der Waals.

L'energia potenziale delle interazioni elettrostatiche può essere calcolata con la legge di Coulomb, mentre quella delle interazioni di van der Waals tramite l'equazione di Lennard-Jones. Nelle simulazioni di sistemi biologici vengono usati: CHARMM, AMBER, GROMACS, ecc.

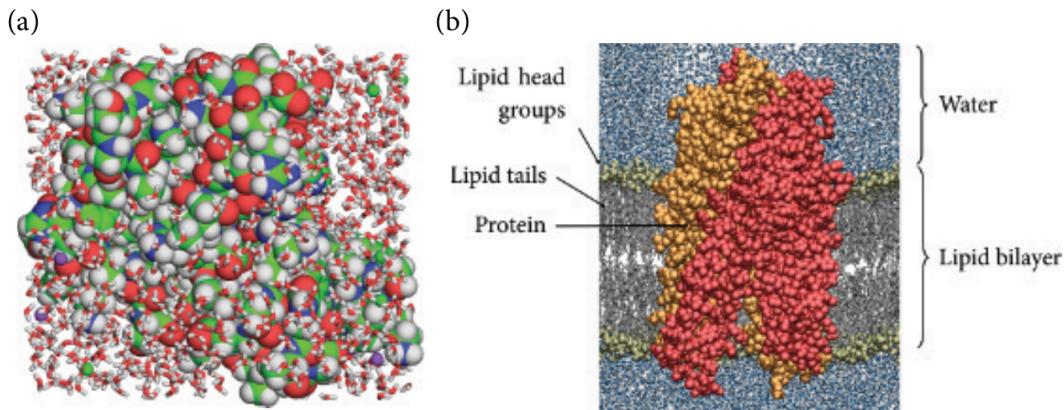


Figura 4.20: Descrizioni esplicite nei calcoli di MM. (A) una piccola proteina immersa in un solvente composto da molecole d'acqua e ioni (Na^+ , Cl^-). La proteina è rappresentata come sfere di atomi, l'acqua come bastoncini e gli ioni come piccole sfere magenta e gialle. (B) una proteina trasportatrice in un doppio strato lipidico, circondato da ambiente acquoso. La proteina e le teste dei lipidi sono rappresentate in modo space-fill, mentre l'acqua e le code dei lipidi con rappresentazione wire-frame. Fonte [6]

La descrizione approssimata fornita dal campo di forza permette di calcolare l'energia potenziale di molti sistemi macromolecolari in meno di un secondo.

Una variante del MM è la *QM/MM* nella quale i calcoli di QM sono indirizzati solamente su una piccola parte della proteina che contiene residui funzionali importanti. Le altre regioni sono soggette invece a MM, con calcoli molto più veloci¹⁴.

¹⁴Questo approccio è stato introdotto da Warshel, Levitt e Karplus.

Spazio configurazionale

Assumendo l'accuratezza del campo di forza, il calcolo dell'energia potenziale di un sistema consente di determinare (parte del)la stabilità di una configurazione. L'idea iniziale potrebbe essere quella di considerare tutte le possibili locazioni atomiche del sistema, calcolare l'energia potenziale in ogni caso e scegliere quella con la minor energia. Come si può facilmente intuire ciò risulta essere un procedimento troppo oneroso, in quanto si devono considerare anche gli atomi del solvente (ed eventuali ligandi o cofattori). Anche il solo numero delle possibili configurazioni atomiche è difficile da calcolare.

Per superare questo problema vengono usate tecniche per ridurre lo spazio di ricerca nello spazio configurazionale. Ci sono vari metodi di ricerca, ad esempio: *systematic search* (grid search basata su dettagli geometrici), *model-building model* (usa frammenti molecolari), *random approach* (movimenti random sul piano cartesiano da una configurazione iniziale), *distance geometry* (usa una matrice di distanze atomiche), *Monte Carlo method* (modifiche random e accettazione probabilistica di configurazioni a livelli energetici maggiori)^[106].

Il metodo più semplice è chiamato *energy minimization*:

1. si parte da una configurazione arbitraria
2. si calcola l'energia potenziale e viene derivato questo valore su differenti posizioni nel sistema in modo da calcolare le forze agenti su ogni atomo dalla rimanente parte del sistema
3. un piccolo cambiamento è introdotto nella posizione di ogni atomo, in risposta alle forze applicate su ognuno di essi dal resto del sistema (in accordo a quanto calcolato nel precedente step)
4. se la nuova configurazione ha un'energia minore viene adottata
5. altrimenti questa viene scartata e viene creata una nuova configurazione
6. si ritorna allo step 3 finché non si trovano più configurazioni con minor energia

Il metodo passa da una configurazione all'altra scendendo con il gradiente della superficie dell'energia potenziale finché non converge in un *punto di minimo locale*. Tutte le procedure di *energy minimization* tendono spesso a rimanere bloccate in un minimo locale di energia, non riuscendo a raggiungere il minimo globale a causa di *barriere energetiche* da scavalcare per raggiungere una configurazione con energia minore (vedi fig. 4.21 e 3.25).

Molecular dynamics

È possibile spingere l'algoritmo di minimizzazione energetica fuori da punti di minimo locale fornendo energia extra, ad esempio innalzando la temperatura del sistema (ovvero

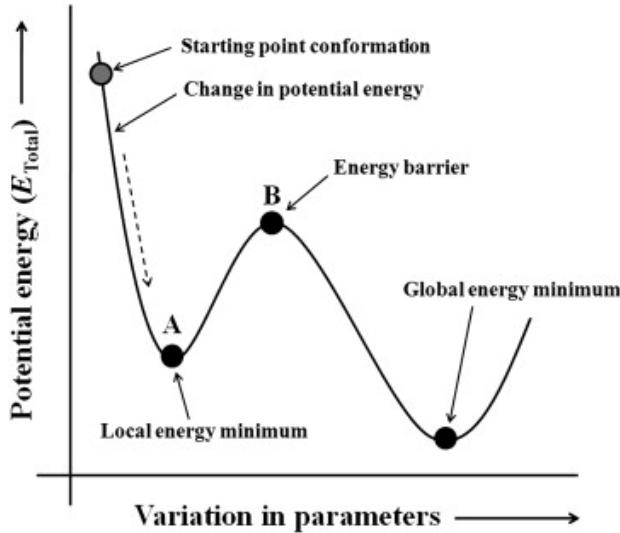


Figura 4.21: Differenti fasi energetiche di una molecola durante la sua minimizzazione energetica. Fonte[106]

aggiungendo calore virtuale). L’energia aggiunta consente agli atomi del sistema di incrementare i loro movimenti e nuove configurazioni fuori dalle barriere energetiche vengono create. Questo metodo è chiamato *Molecular dynamics* (MD) e si focalizza sui movimenti dipendenti dal tempo degli atomi nel sistema. I calcoli sono realizzati in accordo alla meccanica classica.

Agli atomi viene assegnata una velocità iniziale (proporzionale alla temperatura) e continuano a muoversi nello spazio secondo i corrispondenti cambiamenti nell’energia potenziale del sistema. Il movimento di ogni atomo nel sistema è calcolato in base alla sua energia in quel dato momento.

Le simulazione di MD sono eseguite in cicli ripetitivi di *riscaldamento* e *raffreddamento* (metodo conosciuto nel mondo informatico come *simulated annealing*, in riferimento al processo di tempra dei metalli). Nella fase di riscaldamento vengono superate le barriere energetiche mentre la fase di raffreddamento (seguita dall’*energy minimization*) consente al sistema di rilassarsi in configurazioni con minor energia.

Un metodo comune per rendere la ricerca con MD più efficiente è di spezzarla in due fasi:

- ricerca a bassa risoluzione per trovare una collezione di strutture con interazioni non polari (basato sulla nozione che il nucleo delle proteine globulari sia idrofobico)
- ricerca ad alta risoluzione fra le strutture selezionate nel primo step

Limiti dell'approccio fisico e parziali soluzioni

Sebbene simulare il ripiegamento proteico seguendo la meccanica classica possa apparire un approccio attraente, questo è pratico solo per piccole proteine e usando alte risorse computazionali: lo spazio di ricerca è enorme e il problema è computazionalmente intrattabile in modo deterministico (è NP-hard)^[83].

I metodi di MM/MD trovano difficilmente impiego in processi biologici rilevanti come il protein folding. Alcuni problemi riguardano l'approssimazione in sé del campo di forza, la sua accuratezza e i possibili doppi conteggi delle forze in gioco (es. interazioni ioniche e legami idrogeno calcolate in due espressioni differenti).

Un altro problema, sempre nell'approssimazione dell'energia libera con campi di forza, è che forniscono sì l'energia potenziale ma non l'entropia. L'unico modo per stimare l'entropia e l'energia libera dai calcoli per l'energia potenziale è eseguire questi calcoli su tutte le possibili configurazioni del sistema e poi integrarli. Il problema risiede quindi nell'impossibilità di compiere la totalità di questi calcoli a causa delle rappresentazioni esplicite usate nelle simulazioni di MD. In particolare è difficile considerare tutte le configurazioni del solvente acquoso. Ciò che si sta calcolando non è l'energia libera ma un *potenziale di forze medie* (PMF). In conclusione le simulazioni di MD non sono consigliate per descrivere gli effetti dei solventi.

Mean field approach

Per ovviare parzialmente al problema delle rappresentazioni esplicite è possibile descrivere *implicitamente* parti del sistema che vengono descritte da una proprietà media, per questa ragione tale approccio è chiamato *mean field*. Un esempio è la descrizione del solvente, la parte "meno" interessante in genere, come una massa omogenea descritta dalla sua *dielettricità*¹⁵, conosciuto anche come approccio *continuum-solvent*, vedi fig. 4.22.

Ovviamente, essendo una forte approssimazione, alcuni aspetti del sistema reale sono ignorati, come le interazioni fra gli atomi delle proteine e le molecole d'acqua. Tale problema si esacerba quando il solvente è una membrana.

Un altro compromesso è l'approccio *mixed force fields* che combina calcoli esplicativi sulla proteina e calcoli impliciti sul solvente. Viene usata l'equazione di Poisson-Boltzmann (PBE) per calcolare accuratamente l'energia libera elettrostatica, legando così l'effetto polarizzante delle cariche con il loro ambiente. Essendo però un calcolo oneroso viene in genere risolta l'equazione generalizzata di Born (GB). A partire da queste due equazioni, che calcolano la componente elettrostatica dell'energia libera, è possibile calcolare l'intera

¹⁵Proprietà di un mezzo non conduttore di essere sede di un campo elettrostatico.

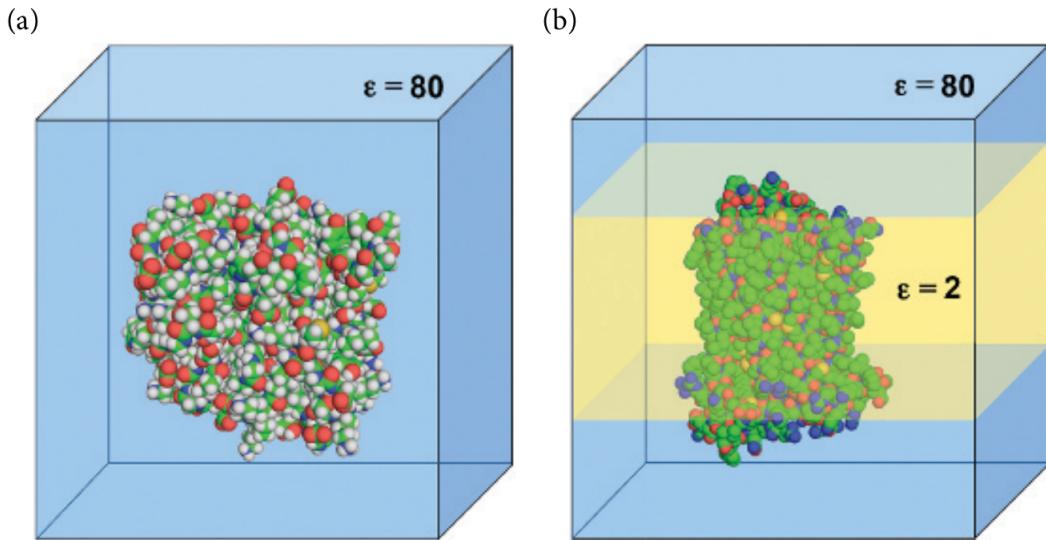


Figura 4.22: Descrizione con approccio mean-field di un sistema, il solvente è descritto implicitamente mentre la proteina esplicitamente. ϵ indica la dielettricità. (A) proteina in un solvente acquoso altamente dielettrico. (B) proteina di membrana in un ambiente eterogeneo. Il solvente acquoso è altamente dielettrico mentre la lastra semi-trasparente gialla, che rappresenta la regione biologica di doppio strato lipidico, è poco dielettrica. Fonte[6]

free energy, in modo abbastanza accurato, con calcoli che si rifanno alla surface area (SA, vedi parte finale della sez. 3.3.1).

Questi metodi sono chiamati *PBSA* e *GBSA* rispettivamente, e come si è visto permettono un calcolo più preciso dell’energia libera. Questi possono a loro volta essere combinati con la MM per rappresentare anche le interazioni del sistema (*MM-PBSA*, *MM-GBSA*) e sono oggi largamente utilizzati.

Un altro limite computazionale è il lasso temporale che si riesce a coprire. La maggior parte delle proteine si ripiega in microsecondi mentre le simulazioni riescono a coprire tempi che vanno dai pico ai nanosecondi. Grazie ad avanzamenti nelle risorse informatiche sono stati fatti dei passi avanti da questo punto di vista. Un caso interessante è *Anton*, un supercomputer progettato specificatamente per ottimizzare simulazioni di MD capace di coprire $85\mu s$ al giorno per un sistema molecolare di 23.000 atomi (180 volte più veloce di qualsiasi computer general-purpose). Altri progressi sono dovuti alla computazione parallela e alla computazione accelerata dalla GPU. Il calcolo distribuito (*grid computing*), ovvero una larga rete di computer personali dedicati volontariamente al completamento di processi, ha permesso alla rete *Folding@Home* (170.000 computer) di simulare l’intero processo di ripiegamento della proteina di legame dell’acetil coenzima A, composta da 86 residui e che richiede 10 millisecondi per ripiegarsi. Un’altra rete distribuita di calcolo è *Rosetta@Home*, con 86.000 nodi e finalizzata al PSP.

Conclusioni sull'approccio *ab initio*

I metodi *ab initio* non sono attualmente in grado di predire la struttura della maggior parte delle proteine sulla sola base della loro sequenza. Ma sono molto abili nel farlo quando il punto di partenza della predizione è una struttura vicino a quella nativa. Questi metodi sono infatti ampiamente usati per raffinare le strutture grezze ottenute dalle determinazioni sperimentali (vedi sez. 3.4).

Il loro successo dipende ampiamente dall'accuratezza della funzione di energia, dall'efficienza dell'algoritmo di ricerca nello spazio conformazionale e dall'abilità di discernere strutture native da "esche" energeticamente intrappolate. Nonostante le loro limitazioni i metodi *ab initio* sono di grande interesse perché sono gli unici, in principio, capaci di derivare la vera struttura nativa delle proteine e possono quindi fornire intuizioni importanti per il protein folding problem. Hanno infatti fornito informazioni importanti sulla dinamica delle proteine e sono utilizzati anche nel *protein engineering* e nel *drug discovery*.

Valutazione *knowledge-based*

È anche possibile rimpiazzare il campo di forza con una funzione di valutazione *knowledge-based*. Spesso queste funzioni sono composte di espressioni relative alla tendenza statistica di gruppi chimici, amminoacidi o atomi di interagire fra di loro. L'affidabilità della funzione di valutazione dipende dal database su cui si basa.

4.4.4 *fragment-based*

Laddove non siano disponibili strutture template adatte (*template-free modeling*), l'uso di metodi *fragment-based* diventa l'unica alternativa pratica poiché le tecniche *ab initio* pure richiedono enormi risorse computazionali anche per proteine molto piccole .

Questi metodi, in primo luogo, ricercano nel PDB frammenti di struttura noti (di pochi residui, ad es. 4-16) che corrispondono a sotto-sequenze della proteina di interesse. Una volta selezionati i frammenti candidati, è possibile formare strutture compatte assemblando frammenti casualmente utilizzando tecniche stocastiche come il *simulated annealing*. Quindi, viene valutata l'idoneità di ciascuna conformazione e vengono ottimizzate quelle più promettenti: mentre i metodi *ab initio* si basano sull'ottimizzazione esplicita dell'energia libera, gli approcci basati sui frammenti svolgono un compito simile utilizzando funzioni di punteggio che sono vagamente correlate alle funzioni energetiche^[83]. I metodi fragment-based sono perciò in parte *data-based* e in parte *ab initio*.

Inoltre, è stato dimostrato che l'inclusione di una misura di somiglianza tra la struttura secondaria di un frammento candidato e quella prevista nel target migliora le funzioni

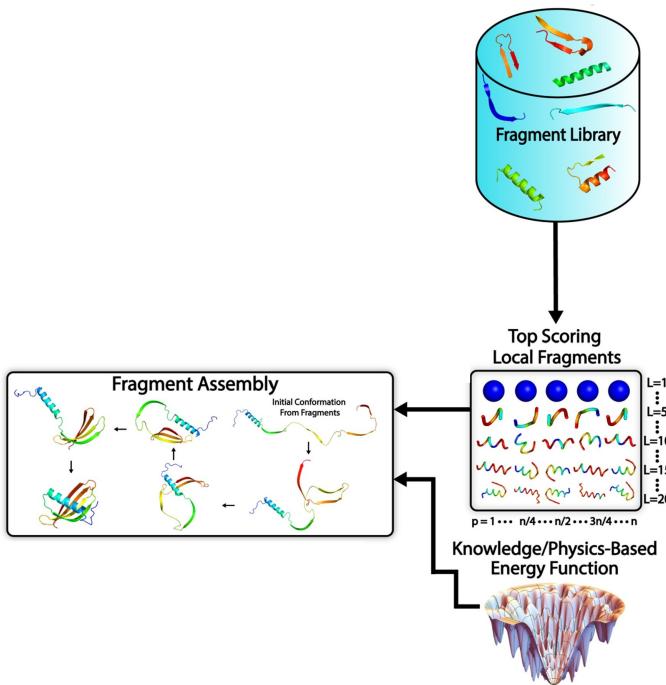


Figura 4.23: Step per le tecniche fragment-based nella pipeline generica di un metodo odierno per il PSP.
Fonte[85]

di punteggio. Per ragioni di ottimizzazione, anche in questi metodi vengono usate le annotazioni 1D e 2D calcolate precedentemente.

In questo approccio l’assemblaggio della proteina è quindi realizzato usando piccoli frammenti di proteine conosciute che verranno poi integrati. Un metodo più efficiente rispetto a cercare frammenti nel PDB risulta essere usare librerie di frammenti. Una recente libreria basata sul DL è DeepFragLib. Un principio chiave alla base del PSP *fragment-based* è che qualsiasi struttura può essere costruita dalla concatenazione di frammenti ottenuti da strutture proteiche disponibili nel PDB, per tale ragione una libreria di frammenti ideale dovrebbe essere in grado di costruire qualsiasi proteina.

Gli approcci basati sui frammenti sono molto meno dispendiosi dal punto di vista computazionale di quelli ”classici” ab initio per tre ragioni principali^[83]:

- sono *coarse grained* (a grana grossa), l’unità di calcolo è un insieme di amminoacidi anziché uno singolo e ciò diminuisce drasticamente lo spazio di ricerca conformazionale
- si utilizzano simulazioni Monte Carlo (MC) al posto della MD
- poiché i frammenti utilizzati sono già a bassa energia, non è necessario calcolare le interazioni locali all’interno del frammenti che vengono introdotti nella struttura

Per far fronte all'ampio spazio di ricerca, la maggior parte dei metodi *fragment-based* si basano sulla generazione di migliaia di strutture candidate, note come *esche*. Ciascuna di esse rappresenta, in linea di principio, una diversa traiettoria di ricerca. Tipicamente, le esche con il punteggio di energia più basso sono considerate come le migliori previsioni.

Il successo di metodi *fragment-based* per il PSP si basa su tre fondamenti:

- accuratezza della funzione energetica
- efficienza del metodo di ricerca
- qualità della libreria di frammenti

Esempi di metodi moderni *fragment-based* sono FRAGFOLD, I-TASSER, QUARK e Rosetta.

4.4.5 *loop modeling*

I loop sono regioni della struttura proteica con ruoli spesso cruciali (interazioni con altre proteine, siti di legame con molecole ecc.). Allo stesso tempo sono molto variabili nella loro sequenza e struttura rispetto alle altre regioni. Si trovano generalmente sulla superficie delle proteine e le loro strutture sono note per essere difficili da predire.

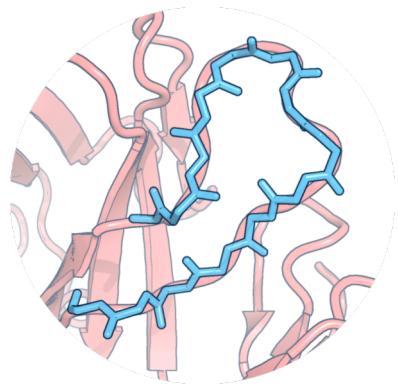


Figura 4.24: Disegno di un loop in celeste. Fonte [107]

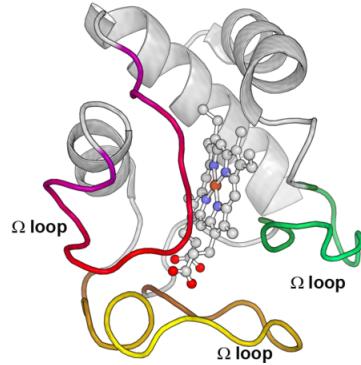


Figura 4.25: Omega loop. Sono spesso coinvolti nel riconoscimento molecolare e in funzioni regolatrici. Fonte: [108]

Il loop modeling non si applica solamente alla fase di raffinamento della modellazione per omologia della predizione di strutture proteiche. È importante anche nella predizione di frammenti mancanti nelle strutture determinate sperimentalmente. È stato stimato che in più della metà delle strutture depositate nel PDB ci siano segmenti mancanti, spesso loop^[109].

Problemi comuni nel loop modeling sono: decidere quale regione del modello sarà un loop; trovare il corretto allineamento di regioni di ancoraggio; la modellazione in sé del loop; le conformazioni di loop multipli, ecc.

Nella modellazione per omologia (nel PSP) si registrano spesso grandi deviazioni dai template omologhi: la modellazione dei loop rimane un problema aperto nella modellazione per omologia della struttura delle proteine^[110]. Le principali strategie per il loop modeling sono le stesse di quelle per la predizione dell'intera struttura:

- *data-based* (o knowledge-based), basati sull'assunzione di similarità sequenza-struttura, ovvero che loop con sequenze simili hanno anche conformazioni simili
- *ab initio*, in cui viene esplorato lo spazio conformativo
- approccio ibrido

Un protocollo comune per la modellazione, partendo da un modello della proteina senza loop e la sequenza del loop, è quello mostrato in figura 4.26, con i seguenti step:

- generazione di tutti i possibili stati del loop (*ab initio*, *data-based* o ibrido)
- valutazione e raggruppamento
- raffinamento

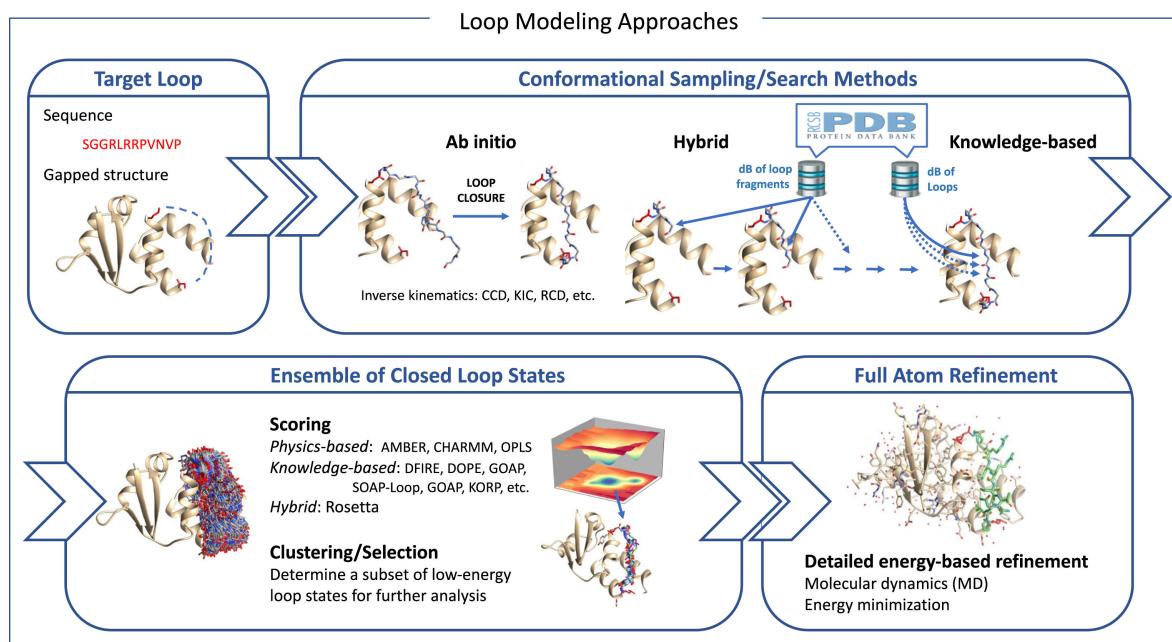


Figura 4.26: Approcci al loop modeling. Workflow schematico di un protocollo prototipo per la modellazione dei loop. Fonte[111]

La ricerca su database è efficiente per famiglie specifiche ma i loop più lunghi di 4 residui devono essere comunque ottimizzati. I residui ai fianchi della regione del loop sono chiamati residui di *ancoraggio* e sono utilizzati per effettuare la ricerca nei database. ArchPRED ad esempio considera le strutture secondarie ai fianchi del loop mancante, il loro orientamento relativo e il numero di residui mancanti per identificare conformazioni

del loop candidate. È possibile usare una funzione di valutazione basata sull'energia per valutare la modellazione dei loop, ad esempio basata sulla stereochimica (come in CHARMM).

Molti dei metodi raggiungono ottimi risultati per la predizione dei loop su strutture sperimentali in ambienti esatti (ovvero strutture cristallizzate a cui mancano le regioni dei loop). Ma nei modelli per omologia non si è ancora riusciti a raggiungere buoni risultati^[109]. Metodi allo stato dell'arte sono in grado di predire conformazioni stabili di loop relativamente corti (fino a 12 residui)^[111]. Pur con le loro limitazioni, gli approcci correnti sono pronti per essere usati in problemi impegnativi come il loop design in enzimi e anticorpi.

I metodi *ab initio* sono dipendenti dalle tecniche di ottimizzazione dell'energia e per questa ragione risultano essere lenti. Metodi *ab initio* per il completamento della struttura cristallizzata allo stato dell'arte sono Rosetta-NGK e GalaxyLoop-PS2. CODA è un esempio di metodo ibrido che combina i due approcci nel loop modeling, così come Sphinx il quale prima esegue una ricerca data-based per trovare frammenti più corti del loop di interesse in modo da ottenere informazioni strutturali, successivamente applica metodi *ab initio* per generare frammenti della corretta lunghezza.

Pochi metodi sono disponibili come web servers e quindi utilizzabili anche dai non esperti: GalaxyLoopPS2, LoopIng, Sphinx e DaReUS-Loop. Metodi locali sono invece MODELLER, Loopy, OSCAR-loop, Rosetta-NGK, LEAP e M-DISGro. Sono disponibili anche dei tool per la modellazione di loop specifici per gli anticorpi, come quelli offerti da SAbPred¹⁶: Sphinx e FREAD (knowledge-based) che effettua una ricerca su database tenendo in considerazione i vincoli spaziali dei residui di ancoraggio.

Possono essere utilizzati anche simulazioni Monte Carlo e MD per investigare proprietà termodinamiche e cinetiche dei loop. Per quanto riguarda l'utilizzo di Deep Learning o Machine Learning per la modellazione dei loop, resta da dimostrare la capacità dei metodi ML/DL di generare modelli significativi di loop flessibili, nonostante in altre aree della bioinformatica strutturale si sia rivelato uno scenario con grande potenziale^[111].

Uno dei metodi più recenti e con migliori risultati è DaReUS-Loop, un approccio *data-based* che identifica loop candidati estraendoli dal completo insieme delle strutture conosciute del PDB. Il filtraggio dei candidati si basa su confronti di conformazioni locali profilo-profilo insieme a una valutazione fisico-chimica. Applicato ai dataset del CASP11 e CASP12 mostra significativi progressi nell'accuratezza della predizione dei loop e propone una misura di confidenza che correla bene con l'accuratezza effettiva dei loop. I loro autori mostrano anche che oltre il 50% dei modelli ben riusciti sono derivati da proteine

¹⁶Collezione di tool sviluppati da Oxford Protein Informatics Group (OPIG).

non correlate: ciò suggerisce che frammenti di proteine, sotto simili vincoli, tendono ad adottare simili strutture (oltre la mera omologia)^[109].

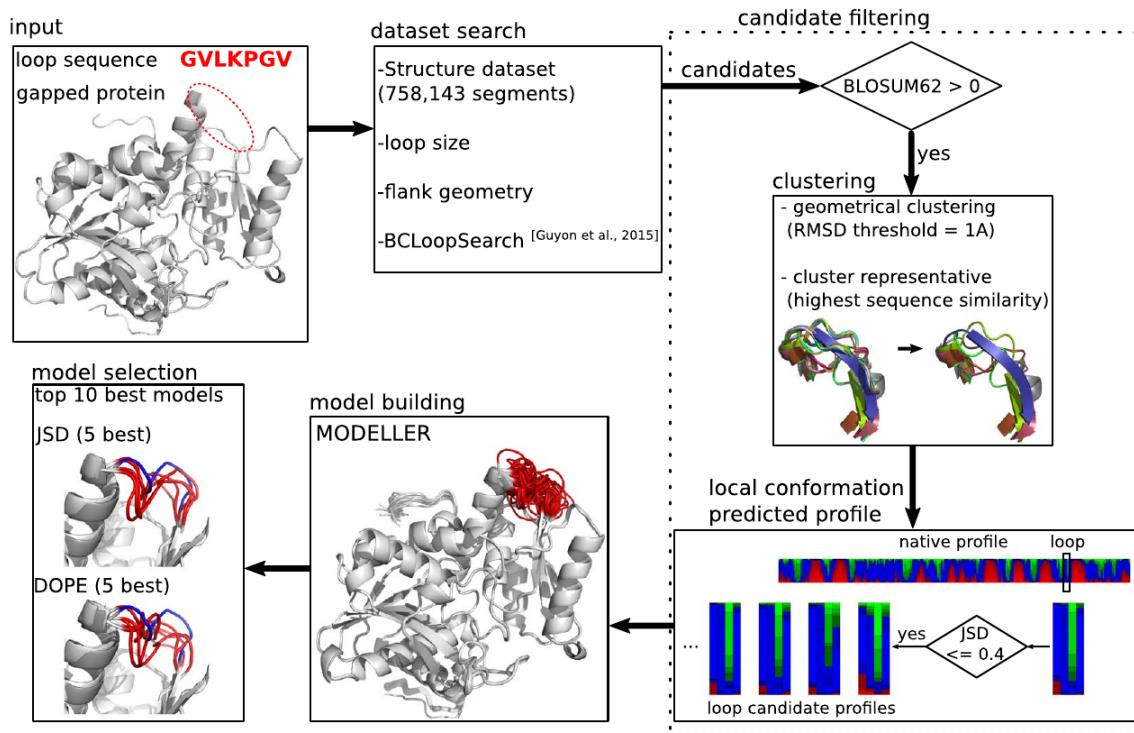


Figura 4.27: DaReUS-Loop workflow. Da notare che dal 2019^[110] nel processo di costruzione del modello non è più usato MODELLER (non free) ma GROMACS. Fonte^[109]

I principali step del metodo sono mostrati in figura 4.27 e sono:

- ricerca dei candidati del loop
- filtraggio dei candidati
- costruzione del modello
- model selection

Nell'ultimo step sono utilizzate 2 misure per valutare i modelli e vengono ritornati in output come predizioni finali i 5 migliori modelli per ogni metrica.

4.4.6 Case Study: **TASSER**

Un metodo che combina differenti approcci (*threading*, *fragment-based* e *ab initio*) è TASSER (Threading/ASSEmby/Refinement) sviluppato dal gruppo Zhang nel 2004 (negli anni ha subito vari miglioramenti). Il metodo TBM di maggior successo nel CASP è stato probabilmente I-TASSER, l'estensione iterativa di TASSER. Si è infatti costantemente

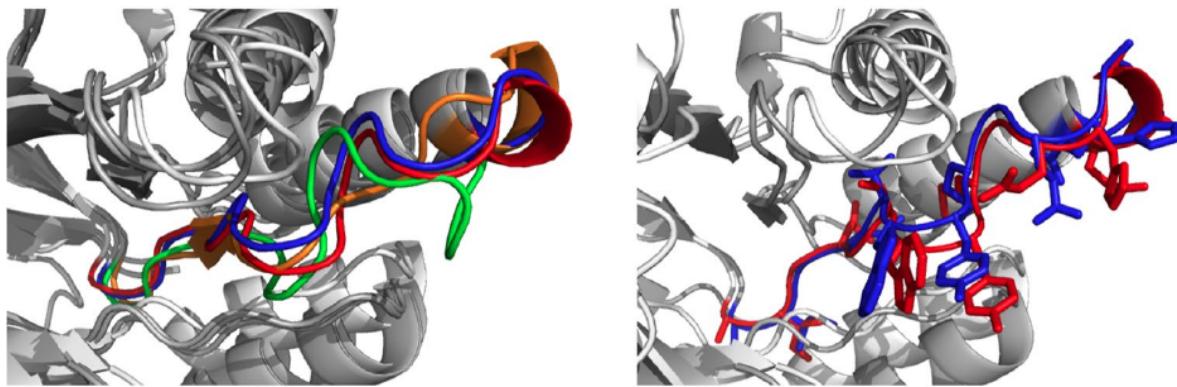


Figura 4.28: Esempi di predizione di un loop lungo (15 residui) della proteina target T0807 del CASP11 a confronto. Blu=DaReUS-Loop, verde=Rosetta NGK, arancione=GalaxyLoop-PS2, rosso=struttura cristallizzata. La RMSD di ogni loop predetto rispetto al loop nativo è riportata di seguito. DaReUS-Loop: 1.3 Å, NGK: 3 Å, PS2: 2.9 Å. Nella colonna a destra sono riportate le catene laterali della struttura nativa e di quella predetta da DaReUS-Loop. Fonte[109]

classificato come il miglior metodo automatizzato nel CASP^[85] considerando tutti i target, e anche nel CASP13 e 14, pur non competendo con AlphaFold, ha ottenuto ottimi risultati. È applicabile sia a situazioni TBM che FM in quanto non impone limiti alla lunghezza dei frammenti utilizzati (>5 residui).

I-TASSER genera inizialmente conformazioni a bassa risoluzione che sono poi rifinite; la sua pipeline di può essere riassunta nei seguenti 3 step^[83]:

- *threading*
 - vengono identificati i modelli di ripiegamento per la proteina target nella libreria di file utilizzando LOMETS
 - il threading utilizza un profilo di sequenza e assegnazioni di struttura secondaria, entrambi calcolati dalla sequenza della proteina target
 - ogni modello è classificato in base a una varietà di punteggi basati sulla sequenza e sulla struttura e i template con punteggi migliori vengono selezionati per il prossimo step
- *assemblaggio strutturale*
 - viene assemblata la struttura a grana grossa (senza gruppi laterali)
 - vengono asportati frammenti dalle regioni maggiormente allineate dei template selezionati
 - le regioni non allineate (principalmente loop) sono costruite *ab initio*
 - gli assemblaggi sono eseguiti attraverso simulazioni replica-exchange Monte Carlo (REMC) e vincolati da un campo di forza combinato *energy-based* e *knowledge-based* che include vincoli derivati dal PDB e dal threading

- le conformazioni sono strutturalmente raggruppate per produrre un insieme di rappresentanti: i *cluster centroids*
- *model selection e raffinamento*, quelle strutture sono raffinate durante un'altra fase di simulazione per produrre tutti gli atomi del modello

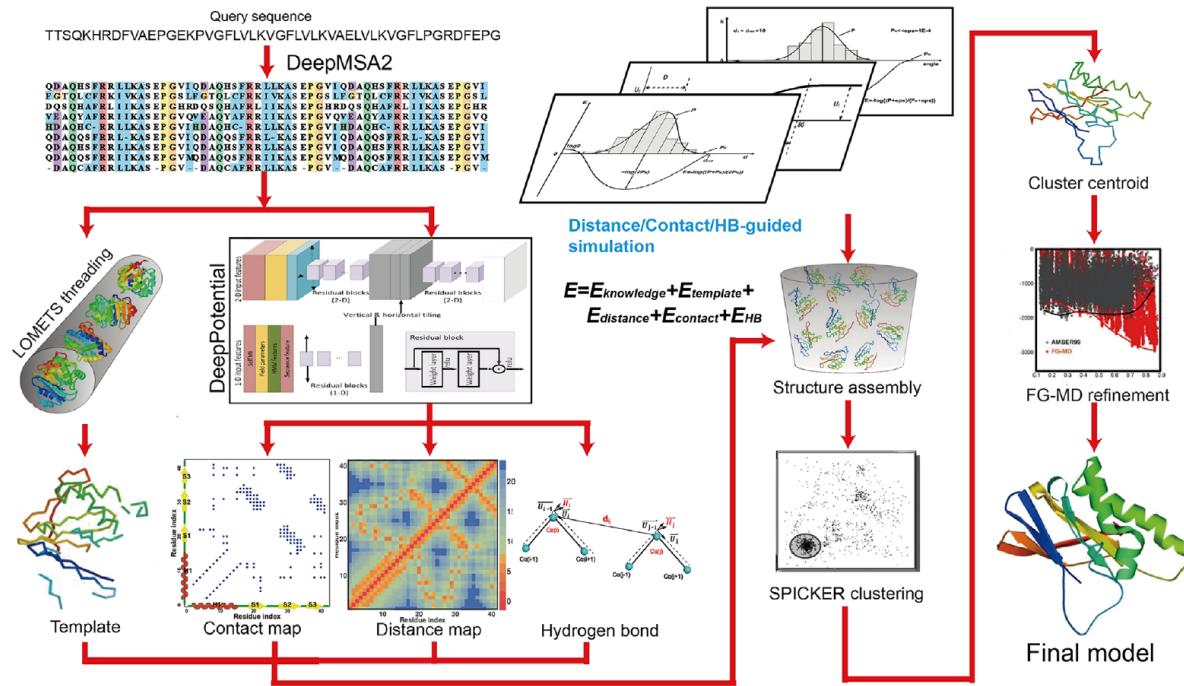


Figura 4.29: D-I-TASSER pipeline. Fonte[112]

Una delle ragioni principali del successo di I-TASSER, in particolare sul perfezionamento dei modelli, è la sua combinazione efficace di più modelli di *threading* (spesso più di 20–50) sotto la guida di un campo di forza combinato *energy-based* e *knowledge-based* ottimale i cui parametri sono stati ampiamente ottimizzati utilizzando richiami strutturali su larga scala.

Recentemente lo stesso gruppo di ricerca ha sviluppato C-I-TASSER (Contact-Iterative-TASSER), che aggiunge la predizione dei contatti tramite DL, e D-I-TASSER (la cui struttura è riportata in figura 4.29), il quale integra la predizione di distanze e legami idrogeno basata su metodi di DL con simulazioni iterative di assemblaggio di frammenti.

Lo stesso gruppo Zhang, nel 2012, ha sviluppato un altro metodo dedicato alla modellazione *ab initio*: QUARK. Questo metodo sfrutta i punti di forza di I-TASSER e Rosetta: oltre al profilo di sequenza e alla struttura secondaria, QUARK utilizza anche l'accessibilità ai solventi e gli angoli di torsione per selezionare piccoli frammenti di dimensioni

fino a 20 residui (come Rosetta ma a differenza di I-TASSER), utilizzando un metodo di threading per ciascun frammento di sequenza.

Capitolo 5

La rivoluzione di AlphaFold

Predire la struttura delle proteine è stato un importante problema di ricerca, aperto per più di 50 anni. Nonostante i vari progressi, nessun metodo è riuscito ad arrivare ad una precisione atomica, specialmente nel caso in cui non siano disponibili delle proteine omologhe. AlphaFold2 è il primo metodo computazionale che può regolarmente predire la struttura delle proteine con accuratezza atomica, anche in casi in cui nessuna struttura simile è conosciuta^[113]. AlphaFold2 (AF2) è la versione interamente ridisegnata del modello basato su reti neurali AlphaFold, entrambi sviluppati da DeepMind.

Nel CASP14 (2020) AF2 viene dichiarato vincitore del "protein structure prediction problem" per la maggior parte delle proteine a singolo dominio, dimostrando accuratezza competitiva con le strutture sperimentali nella maggioranza dei casi e superando di gran lunga tutti gli altri metodi esistenti. Alla base di AF2 c'è un nuovo approccio basato sul Machine Learning, che unisce nel design dell'algoritmo di deep learning conoscenza fisica e biologica sulla struttura delle proteine, facendo leva sugli allineamenti multi-seguenza.

Quando è iniziata a circolare la notizia che AF2 avesse risolto il problema del PSP, si pensava che avesse raggiunto un GDT_TS medio di 80^[114] (intuitivamente significa che in media l'80% della struttura delle proteine target è stato predetto). Predizioni casuali forniscono un $GDT_TS \leq 20\%$, predire la struttura grossolanamente è associato a un GDT_TS di circa il 50% mentre predire una topologia accurata porta con sé un valore di circa il 70%. Quando tutti i dettagli, comprese le conformazioni delle side-chain, sono corretti il GDT_TS supera il 90%. Alcuni, tra cui Mohammed AlQuraishi¹ suggerivano che ci sarebbero voluti almeno altri 10 anni per arrivare ad un GDT di 85-90^[115], ma AlphaFold ha riportato una mediana nel GDT_TS pari a 92.4. Un valore che AlQuraishi definisce come uno degli avanzamenti scientifici più rapidi degli ultimi decenni.

¹Assistant Professor, Department of Systems Biology della Columbia University e principale investigatore dell'AlQuraishi Laboratory. È stato anche uno dei peer reviewer del paper di AlphaFold2^[114].

Le strutture di AF2 hanno un'accuratezza riguardo la mediana della backbone² di 0.96 RMSD₉₅, mentre il prossimo miglior metodo ha dimostrato un'accuratezza di 2.8 Å. In figura 5.1 è possibile vedere il confronto dei vari metodi che hanno partecipato al CASP14 valutati secondo il Z-score³.

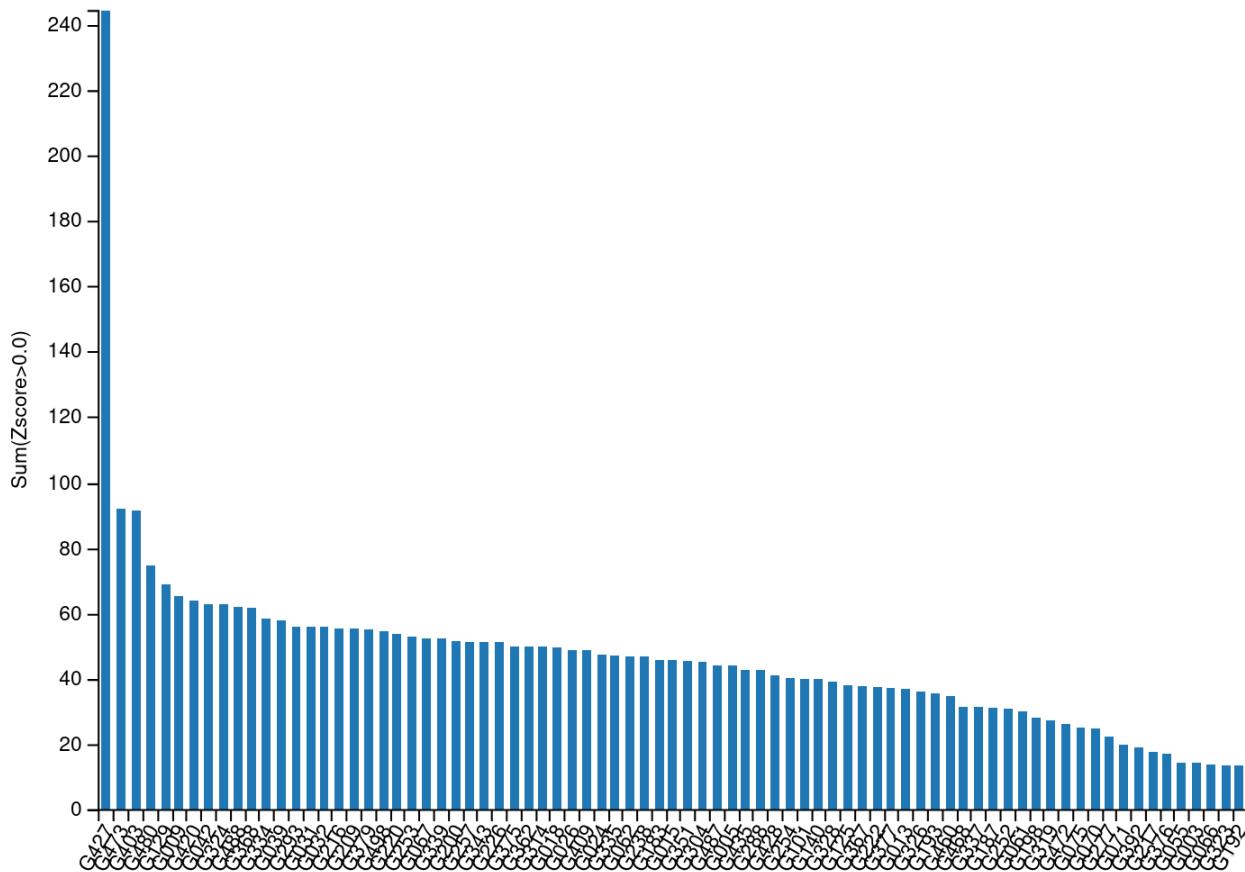


Figura 5.1: Risultati del CASP14 in base al Z-score. AlphaFold (G427) è incredibilmente avanti rispetto al secondo gruppo (473, Baker). Fonte[116]

Per quanto riguarda l'accuratezza non solo della backbone ma di tutti gli atomi, AlphaFold ha registrato un'accuratezza di 1.5 Å (per fare un confronto, un atomo di carbonio è largo approssimativamente 1.4 Å).

In alcuni casi le predizioni di AlphFold erano talmente accurate da superare i risultati sperimentali facendo mettere in discussione agli sperimentatori i risultati da loro ottenuti. Si potrebbe pensare che i target del CASP14 fossero in qualche misura più semplici rispetto

²La notazione indica C_α mean root square deviation ad una copertura del 95% dei residui. L'intervallo con 95% di confidenza riportato in questo caso da AF2 corrisponde a 0.85-1.16 Å.

³Lo Z-score è la differenza del valore di un campione rispetto alla media della popolazione, divisa per la deviazione standard; un valore alto rappresenta una grande deviazione dalla media ed è comunemente usato come procedura di rilevamento dei valori anomali.

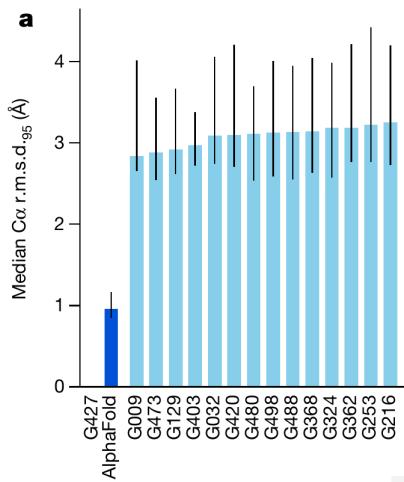


Figura 5.2: Performance di AF2 sul dataset del CASP14 ($n=87$ domini di proteine) rispetto agli altri migliori 15 metodi (su 146). Fonte [113]

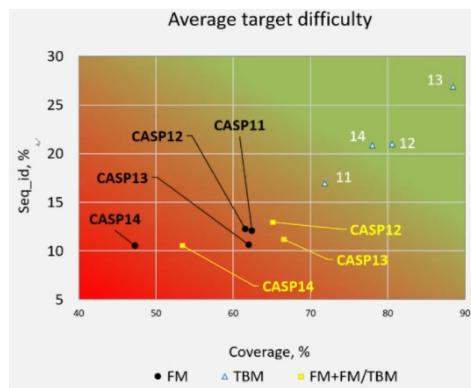


Figura 5.3: Confronto degli obiettivi degli ultimi quattro CASP in termini di copertura e identità di sequenza dei modelli disponibili. In entrambi i casi, CASP14 include gli obiettivi di modellazione libera (FM) più difficili mai forniti. Fonte [117]

a quelli degli altri anni per spiegare il successo di AlphaFold. Ma non è così, anzi, gli organizzatori hanno dimostrato che è stato il CASP più difficile (in quanto a percentuale di identità di sequenze, vedi fig. 5.3).

Un esempio in cui AF2 surclassa gli altri metodi è il target T1064. AF2 riesce ad ottenere una similitudine molto alta, con nucleo e strutture secondarie quasi perfette (nonostante una grande regione di loop sia sbagliata, ma questo potrebbe anche indicare che sia una regione flessibile).

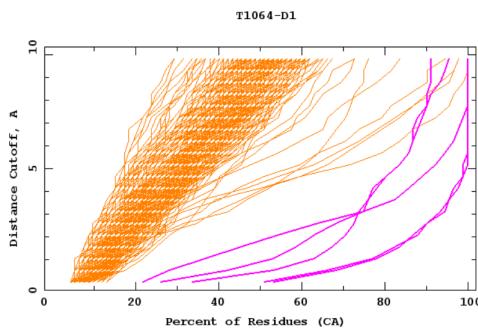


Figura 5.4: Analisi GDT dei 508 modelli inviati per la sequenza target T1064-D1. L'analisi denota il più grande insieme di atomi di C_α (percentuale della struttura modellata) che può rientrare nella distanza cutoff $\in \{0.5\text{\AA}, 1.0\text{\AA}, 1.5\text{\AA}, \dots, 10.0\text{\AA}\}$. In viola i modelli di AlphaFold. Fonte: [116]

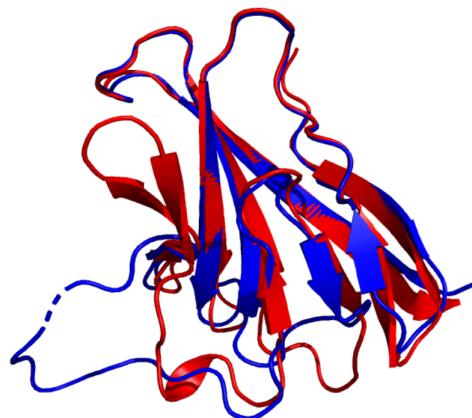


Figura 5.5: (rosso) modello di AF2 per il target T1064. (blu) struttura 7JTL_A. Fonte [117]

Gli altri metodi predicono questa struttura in modo nettamente peggiore. Prendendo in considerazione i risultati dei gruppi Baker e Zhang (i due migliori subito dopo AF2,

basati prevalentemente sulla pipeline del primo AlphaFold) si può notare che il nucleo della proteina è totalmente sbagliato e ci sono molte differenze con la struttura sperimentale (vedi fig. 5.6).

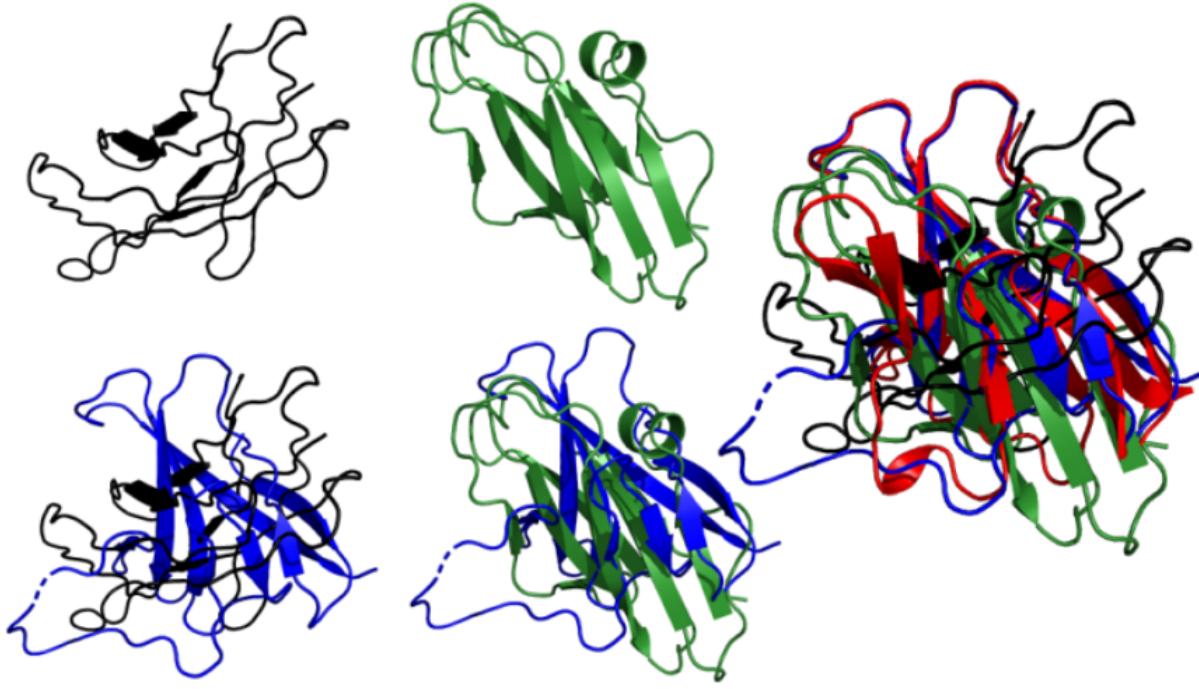


Figura 5.6: Modelli con il punteggio più alto per il target T1064 presentati dai gruppi Zhang (nero) e Baker (verde). In basso: modelli allineati con la struttura cristallina. A destra: tutti e tre i modelli (Zhang, Baker e AlphaFold 2) sono allineati con la struttura cristallina. Fonte[117]

Come è possibile vedere nel grafico in figura 5.7, AF2 vince quasi in tutti i target, ci sono addirittura casi in cui il prossimo miglior metodo raggiunge solo il 20% dell'accuratezza mentre AF è al 90%.

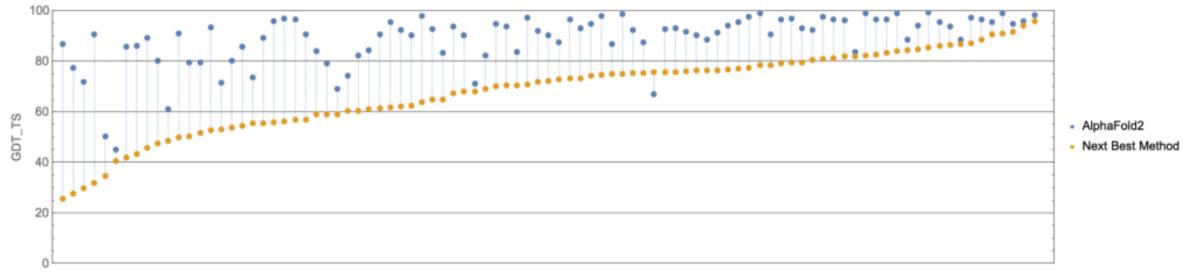


Figura 5.7: Il grafico mostra la differenza fra AF2 e il prossimo miglior metodo in tutti i target del CASP14. Fonte[115]

In generale, anche quando gli altri modelli predicono bene, è nei dettagli che AF2 si differenzia e porta la predizione ad un livello superiore (vedi fig. 5.8).

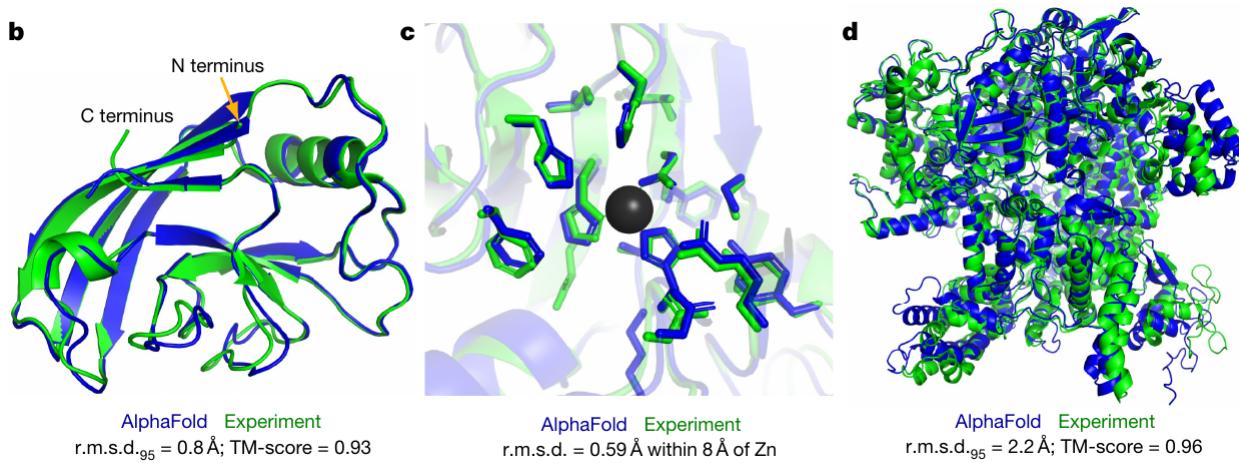


Figura 5.8: Predizioni di AF2 (in blu) sovrapposte alle strutture sperimentali (in verde) sui target: T1049, T1056, T1044. In (c) è possibile notare una corretta predizione di un sito di legame per lo zinco. La struttura in (d) è composta da 2180 residui. Fonte[113]

AlphaFold2 è in grado di fornire delle precise stime della sua affidabilità per residuo in modo da consentire un uso consapevole delle sue predizioni. La sua misura di confidenza (pLDDT) predice affidabilmente l'accuratezza IDDT-C_α della predizione corrispondente. AF2 si è dimostrato applicabile anche su proteine molto lunghe.

La rivoluzione di AlphaFold è stata assimilata alla rivoluzione avvenuta nell'ImageNet nel 2012. Ma secondo AlQuraishi le due cose non sono paragonabili. In quell'occasione il deep learning ha dimostrato di poter superare gli approcci convenzionali nel riconoscimento delle immagini sconvolgendo il mondo della computer vision. Rispetto all'avanzamento di AF2 vi è però una differenza importante: l'avanzamento nell'ImageNet è stato incrementale, quello di AF2 è invece un balzo in avanti di 10 anni, un cambiamento così profondo da mettere sottosopra un intero campo nel corso di una notte; è stato come avere l'accuratezza nell'ImageNet del 2020 già nel 2012, senza tutti i passi intermedi.

Utilizzo di AF2

Il codice sorgente di AlphaFold è stato reso pubblico da DeepMind ed è disponibile su GitHub⁴. Per funzionare AlphaFold ha bisogno di database di supporto (fino a 2.5 TB), di molta memoria e di potenza computazionale. Per questi motivi è verosimile utilizzarlo solo su server dedicati alla computazione, come quello disponibile all'Università di Pisa.

Il codice è rilasciato con un'immagine Docker e un *launcher script* associato, in modo da risultare più facilmente accessibile.

⁴<https://github.com/deepmind/alphafold>

È stata anche pubblicata una versione semplificata di AlphaFold (senza uso di template) tramite un Google Colab notebook⁵.

5.1 Architettura

Forse l'osservazione più importante da fare su AlphaFold è che DeepMind *non* ha scoperto nessun nuovo e sbalorditivo principio sul protein folding. Non ci sono sorprese di carattere biologico. Tutto si basa su un ottimo design del sistema di DL e sul livello altissimo delle abilità dei membri del team e delle risorse a disposizione. È possibile porsi domande sul perché la struttura sia proprio come è presentata nel paper. La risposta più probabile risiede nell'intensa sperimentazione, attuata grazie ad una grande quantità di risorse computazionali e guidata da una grande capacità di progettazione del team di DeepMind.

Uno dei principi su cui si basa AlphaFold è l'immersione delle intuizioni basate sulla conoscenza fisico-chimica delle proteine direttamente nella struttura della rete, non come un processo intorno ad essa. Il bias induttivo del sistema riflette la conoscenza attuale fisica e geometrica delle proteine, sminuendo l'importanza della posizione dei residui nella sequenza ed enfatizzando invece la comunicazione tra i residui vicini nella proteina ripiegata. La rete apprende iterativamente un grafo delle interazioni fra residui, ragionando su questo grafo隐式 mentre viene costruito. AlphaFold è un sistema end-to-end che produce direttamente una struttura invece di fornire come output le distanze inter-residuo.

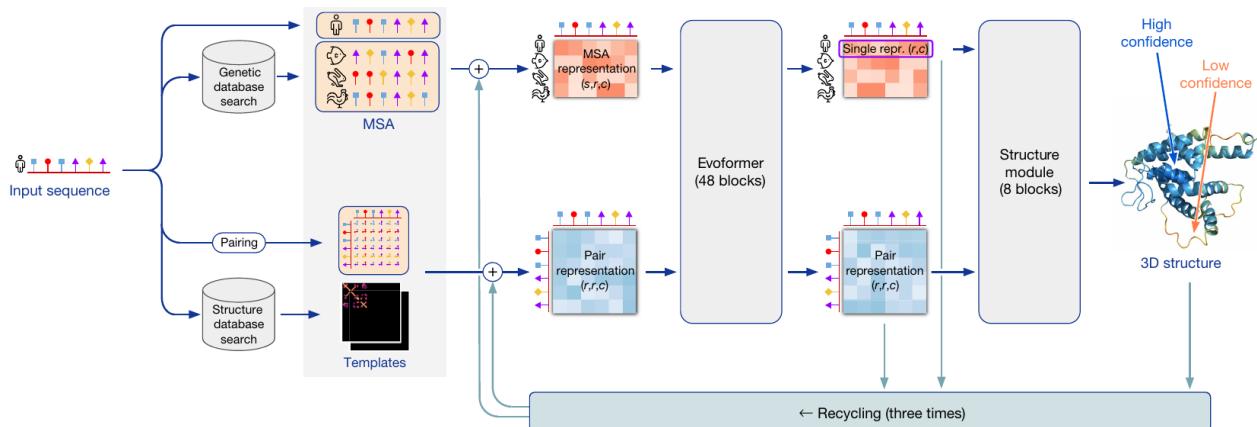


Figura 5.9: Schema architettonico di AF2. Le frecce indicano il flusso dell'informazione. Le dimensioni degli array sono riportate fra parentesi (s =numero di sequenze, r =numero di residui, c =numero di canali).
Fonte[113]

L'architettura principale di AlphaFold può essere suddivisa in 3 componenti principali. Innanzitutto vi è la parte di *preprocessing* dell'input, dove AF2 utilizza la sequenza di am-

⁵<https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb>

minoacidi in ingresso per interrogare diversi database di sequenze proteiche e costruisce un MSA e quindi una *MSA representation*. AlphaFold 2 cerca anche di identificare le proteine che possono avere una struttura simile all'input (template) e costruisce una rappresentazione iniziale dei contatti nella struttura, chiamata *pair representation*. Questo è, in sostanza, un modello di quali amminoacidi è probabile siano in contatto tra loro. La prima parte della struttura di AF2 non aggiunge niente di rivoluzionario alla pipeline dei sistemi di predizione. Vengono utilizzati anche database di metagenomica⁶ come MGnify.

Nella seconda parte del diagramma, AlphaFold 2 prende l'MSA e i template e li passa attraverso un *transformer* (lo si può, per ora, immaginare come un "oracolo" in grado di identificare rapidamente quali informazioni siano più informative). L'obiettivo di questa parte è perfezionare le rappresentazioni sia per l'MSA che per le interazioni di coppia, anche scambiando informazioni tra loro in modo iterativo. Un modello migliore dell'MSA migliorerà la caratterizzazione della geometria della rete, che contemporaneamente aiuterà a perfezionare il modello dell'MSA. Questo processo è organizzato in blocchi che vengono ripetuti in modo iterativo fino a un numero specificato di cicli (48 blocchi nel modello pubblicato).

Queste informazioni vengono portate all'ultima parte del diagramma: lo *structure module*. Questo sofisticato componente della pipeline prende la *MSA representation* e la *pair representation* e le sfrutta per costruire un modello tridimensionale della struttura. Il risultato finale è un lungo elenco di coordinate cartesiane che rappresentano la posizione di ciascun atomo della proteina, comprese le catene laterali.

L'ultima cosa da notare per farsi un'idea della struttura di AF2 è che funziona iterativamente. Una volta generata la prima struttura finale questa sarà utilizzata per raffinare ulteriormente la predizione, fino a un totale di 4 cicli di predizione.

5.1.1 Evoformer

L'Evoformer è il primo componente della struttura di AF2 a cambiare le "regole" dei sistemi di predizione classici. Il compito dell'Evoformer è di spremere ogni goccia di informazione dall'MSA, dai template e dalle altre informazioni derivanti dall'analisi di sequenze. Sono decenni che vengono estratte informazioni attraverso analisi coevolutive, ma fino al CASP13 erano perlopiù approcci statistici. Molti gruppi hanno però dimostrato che attraverso l'uso di ResNet profonde non c'era bisogno di una robusta e complicata statistica. AlphaFold2 reinventa completamente questo processo di analisi coevolutiva e la porta ad un livello di utilità molto superiore.

⁶L'applicazione di moderne tecniche di genomica senza la necessità di isolare e coltivare in laboratorio specie singole, studiandole quindi direttamente nel loro ambiente naturale.

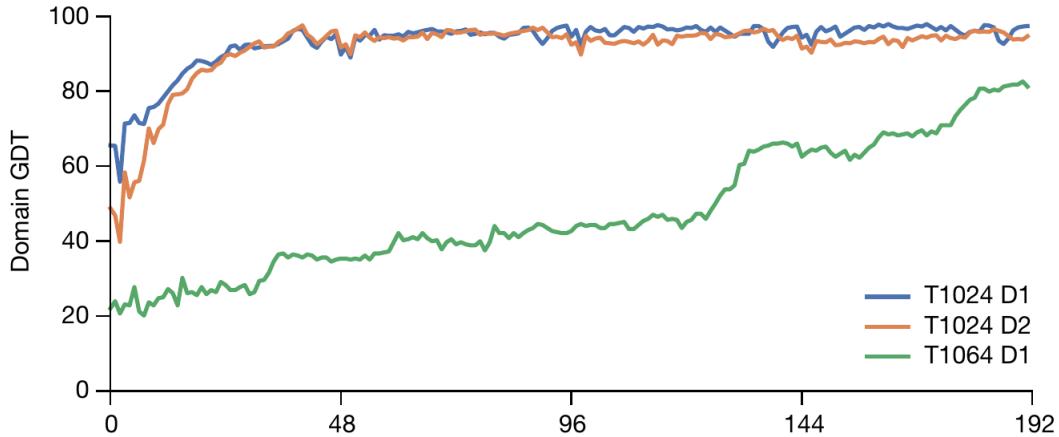


Figura 5.10: Traiettoria del valore del GDT sui domini di due target del CASP14 (T1024, composta da due domini e T1064) con 4 iterazioni del modello. Da notare che 48 blocchi dell’Evoformer costituiscono un ciclo di iterazione. I due domini T1024 ottengono la struttura corretta presto, mentre il target T1064 richiede praticamente tutta la profondità della rete per raggiungere una buona struttura finale. Fonte[113]

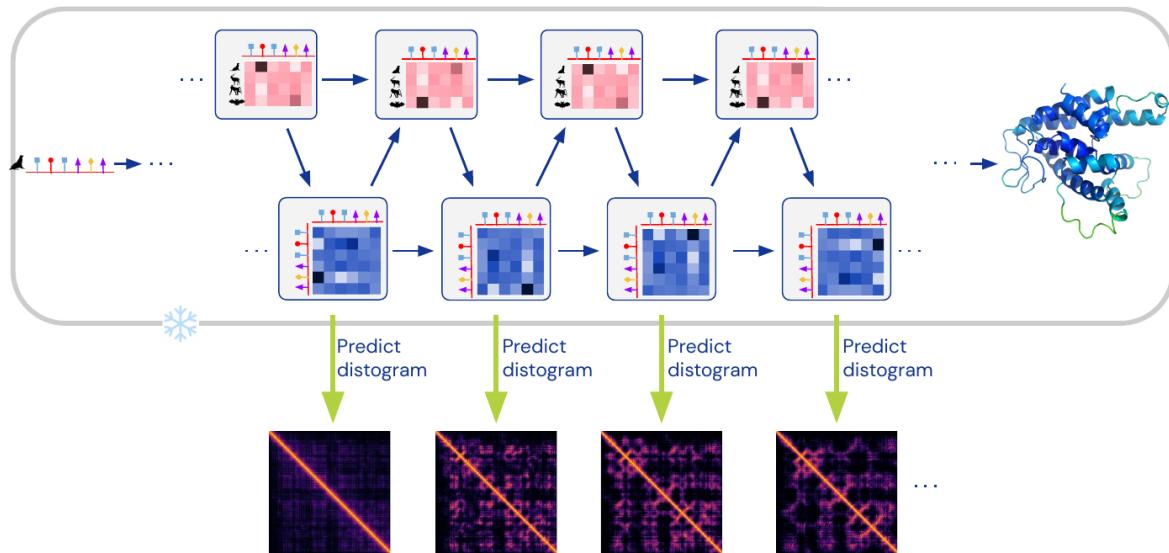


Figura 5.11: Rete di AF interrogata sui distogrammi previsti. La previsione dei distogrammi è uno dei principali passi che AF compie per “comprendere” la struttura della proteina. Fonte[43]

L’idea centrale dietro l’Evoformer è che le informazioni fluiscono avanti e indietro attraverso la rete. Prima di AlphaFold 2, la maggior parte dei modelli di deep learning richiedeva un allineamento di sequenze multiple e generava alcune inferenze sulla prossimità geometrica. L’informazione geometrica era quindi un prodotto della rete. Nell’Evoformer, invece, la *pair representation* è sia un prodotto che uno strato intermedio. Ad ogni ciclo, il modello sfrutta l’attuale ipotesi strutturale per migliorare la valutazione dell’allineamento di sequenze multiple, che a sua volta porta a una nuova ipotesi strutturale, e così via. Entrambe le rappresentazioni, sequenza e struttura, si scambiano informazioni finché la

rete non raggiunge una solida inferenza.

Il primo passo nella rete è definire gli *embeddings* (incorporamenti) per l'MSA e i template. Gli allineamenti di sequenze multiple sono in ultima istanza sequenze di simboli su un alfabeto finito e quindi un esempio di variabile discreta. Le reti neurali, invece, sono intrinsecamente continue e si basano sulla differenziazione per apprendere dal loro training set.

Un *embedding* è un "trucco" del deep learning che consente la trasformazione di una variabile discreta in uno spazio continuo (*embedded space*) in modo che la rete possa essere addestrata. È un processo molto semplice: c'è solo bisogno di definire uno strato di neuroni che riceve l'input discreto ed emette un vettore continuo. Un *embedding*, più precisamente, è uno spazio di dimensioni relativamente basse in cui è possibile tradurre vettori di dimensioni elevate. Idealmente, un *embedding* acquisisce parte della semantica dell'input posizionando input semanticamente simili vicini nello spazio di incorporamento. Un incorporamento può essere appreso e riutilizzato tra i modelli. L'embedding dettagliato che AF2 compie sulle caratteristiche di input può essere osservato in figura 5.12⁷.

L'architettura dell'Evoformer usa due *transformer*, connessi da un canale di comunicazione tra i due. Ognuno è specializzato in un certo tipo di dati: MSA o interazioni fra coppie di amminoacidi.

L'architettura *transformer* è stata introdotta nel 2017 da un gruppo del Google Brain^[119] e l'ingrediente chiave di tale architettura è chiamata *attenzione*. L'obiettivo dell'*attenzione* è identificare quali parti dell'input sono più importanti per l'obiettivo della rete neurale. I transformer hanno dimostrato empiricamente prestazioni superiori in una varietà di compiti, ad esempio nell'*image captioning* e nel *machine translation*. In quest'ultimo problema aiutano a migliorare il problema del *vanishing gradient*, un ostacolo comune durante l'allenamento. Nei modelli basati su sequenze, possono accelerare significativamente l'addestramento rispetto ai classici modelli di RNN. In particolare, sono alla base della maggior parte dei risultati più clamorosi dell'IA degli ultimi anni: ad esempio, GPT in *GPT-3* sta per "Generative Pre-training Transformer".

C'è anche un contro all'uso dei transformer: la costruzione della matrice di attenzione richiede un costo in memoria quadratico. Per questa ragione le Tensor Processing Unit (TPU) introdotte da Google apportano notevoli vantaggi.

Il primo transformer, denominabile anche *MSA transformer*, calcola l'attenzione su una matrice molto ampia. Per ridurre quello che altrimenti sarebbe un problema computazionale intrattabile, l'attenzione "fattorizza" in componenti *riga* e *colonna*. Questo

⁷I dettagli applicativi di ogni variabile e algoritmo sono consultabili nelle informazioni supplementari del paper di AlphaFold: J. Jumper, R. Evans, A. Pritzel et al., "Supplementary Information for Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, n. 7873, pp. 583–589, 2021.

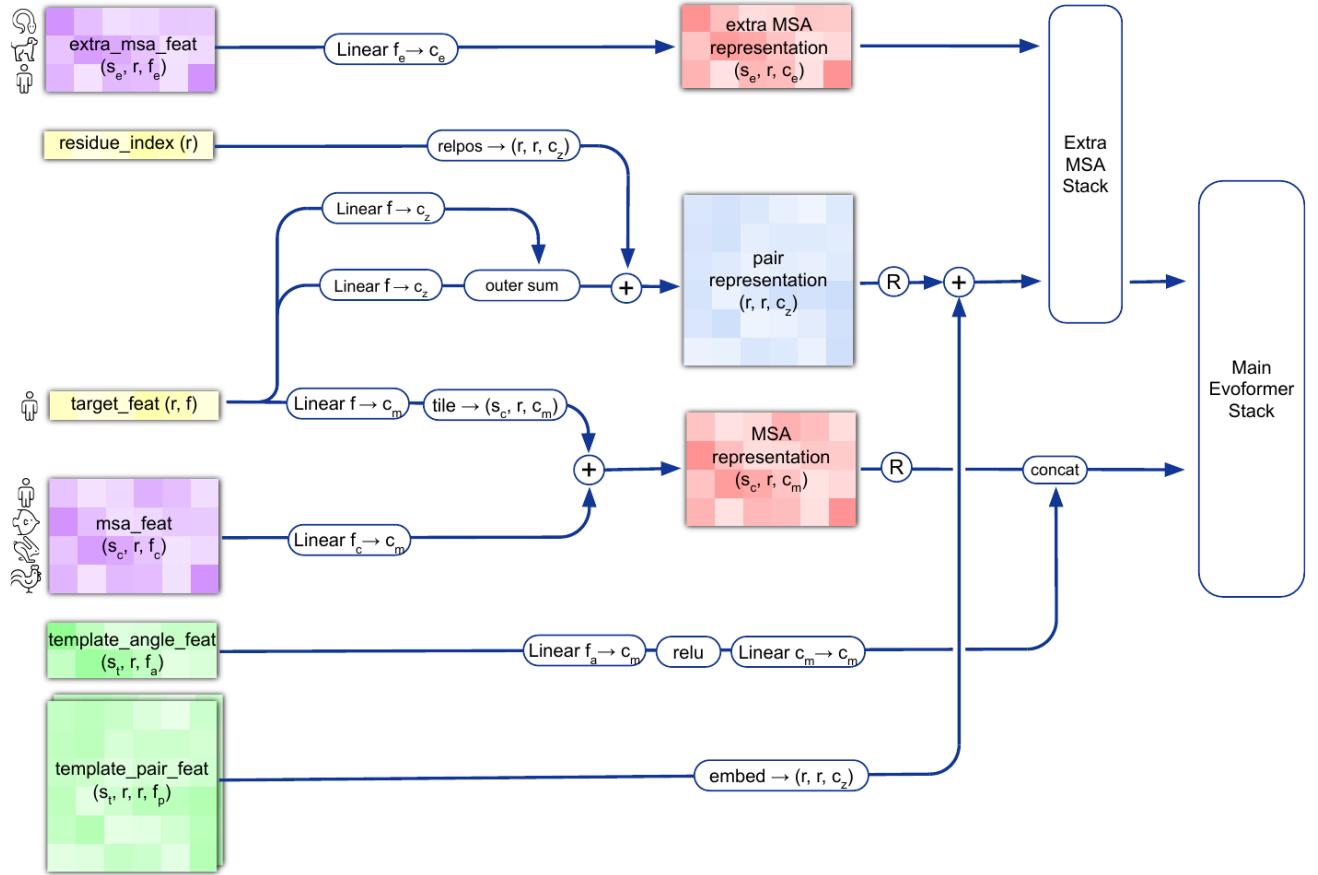


Figura 5.12: Input feature embeddings. Fonte[118]

processo prende il nome di *axial-attention*. La struttura a livelli consente di calcolare la maggioranza del contesto in parallelo durante la decodifica senza introdurre alcuna ipotesi di indipendenza. In questo contesto, semplificando, la rete calcola prima l'attenzione in orizzontale, consentendo alla rete di identificare quali coppie di aminoacidi sono più correlate; e poi in direzione verticale, determinando quali sequenze sono più informative.

La caratteristica più importante dell'MSA transformer di AF2 è che il meccanismo di attenzione *row-wise* incorpora informazioni dalla *pair representation* come si può vedere in figura 5.13, in modo da focalizzarsi sulle coppie di residui che interagiscono fra loro.

L'altro transformer, denominabile *pair transformer*, si basa su un'impostazione fondamentale: l'attenzione è disposta in termini di triangoli di residui (vedi fig. 5.15), con l'obiettivo di sfruttare la disuguaglianza triangolare:

Quest'idea permette di superare un problema classico degli approcci basati su DL per il PSP: le distribuzioni di distanze non potevano essere *embedded* nello spazio tridimensionale.

Dopo un certo numero di iterazioni, 48 nel paper, la rete ha costruito un modello delle

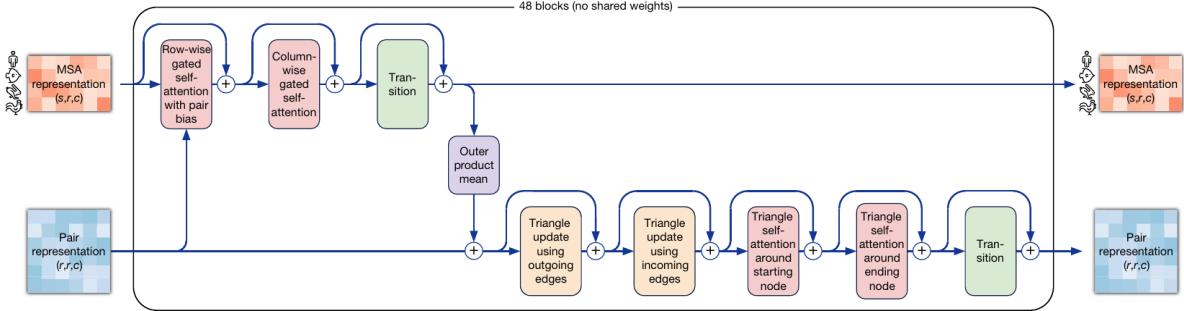


Figura 5.13: Blocco dell’Evoformer, le frecce indicano il flusso dell’informazione. Fonte[113]

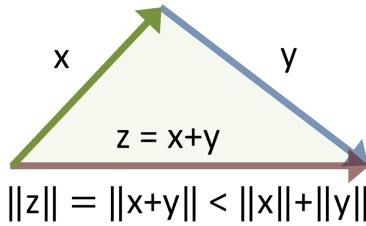


Figura 5.14: Diseguaglianza triangolare. Fonte[120]

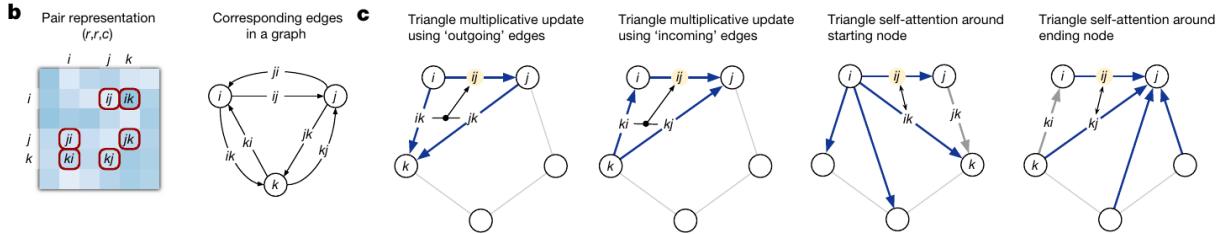


Figura 5.15: (b) Pair representation interpretata a grafo. (c) Aggiornamenti moltiplicativi ai triangoli di residui e self-attention. I dati nella pair representation sono illustrati come archi direzionati e in ogni diagramma l’arco aggiornata è "ij". Fonte[113]

interazioni all’interno della proteina.

5.1.2 Structure Module

In questo modulo viene rappresentata tridimensionalmente la struttura attraverso un ripiegamento *end-to-end* (invece di un metodo a discesa di gradiente). La proteina viene considerata come un *gas residuo*, la backbone è un gas di corpi rigidi 3D. Gli amminoacidi vengono modellati come triangoli, rappresentando i 3 atomi della backbone.

All’inizio i residui vengono tutti piazzati nell’origine (*black hole initialization*). Ad ogni step del processo iterativo vengono prodotte delle matrici di *affinità* (via matematica per rappresentare traslazioni e rotazioni in una matrice 4×4).

Vengono effettuate 8 iterazioni per ridurre le violazioni stereochimiche e raffinare la struttura. Tuttavia anche dopo il processo di raffinamento è possibile che la struttura

presenti delle violazioni. Per questa ragione vengono eseguiti dei raffinamenti ulteriori attraverso una discesa di gradiente vincolata dalle coordinate precedentemente calcolate, usando il campo di forza Amber ff99SB con OpenMM.

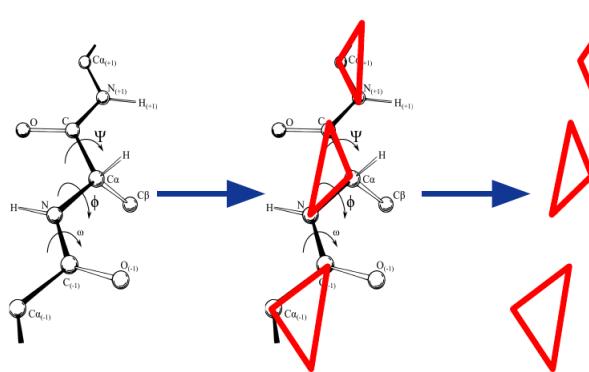


Figura 5.16: Rappresentazione come gas residuo. Fonte [43]

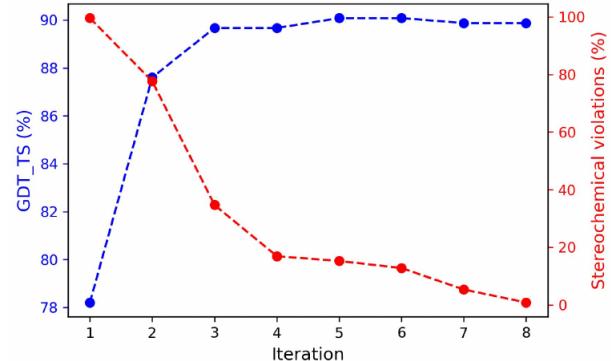


Figura 5.17: Miglioramento dell'accuratezza e diminuzione delle violazioni stereochimiche attraverso le iterazioni del modulo. Fonte [43]

I corpi rigidi vengono aggiornati da un'architettura transformer equivariante 3D, che costruisce anche i gruppi laterali (parametrizzati da una lista di angoli di torsione). Questa è una nuova architettura basata sull'attenzione ideata specificatamente per lavorare con strutture tridimensionali: *Invariant Point Attention* (IPA). Questo meccanismo di attenzione beneficia del fatto di essere invariante rispetto alle traslazioni e alle rotazioni, necessitando così di meno dati.

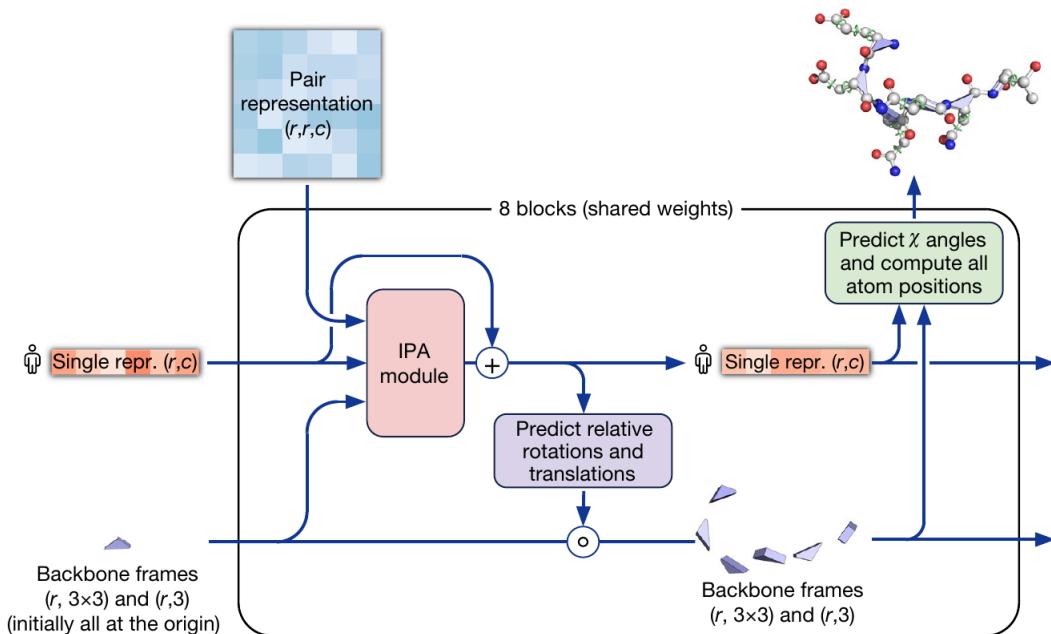


Figura 5.18: Structure module compreso IPA. Fonte [113]

Come rappresentazione iniziale viene usata la prima riga dell’Evoformer (la ”single representation” è una copia della prima riga dell’MSA representation, ovvero una sequenza) ed è chiamata s_i . La *pair representation* influenza le matrici di affinità nelle operazioni di attenzione. Le iterazioni sono 8 perché vi sono 8 strati nel modulo con pesi condivisi. Ogni strato aggiorna la rappresentazione singola astratta ($\{s_i\}$) così come la rappresentazione 3D (gas residuo).

Uno strato dello structure module è composto da 3 principali operazioni, nelle quali la singola rappresentazione astratta:

1. viene aggiornata dall’Invariant Point Attention (algoritmo 20 riga 6, vedi ??)
2. viene aggiornata da un layer di transizione
3. infine viene mappata su frame di aggiornamenti concreti che sono parti dei frame della backbone

Algorithm 20 Structure module

```

def StructureModule  $\left(\{\mathbf{s}_i^{\text{initial}}\}, \{\mathbf{z}_{ij}\}, N_{\text{layer}} = 8, c = 128, \mathbf{s}_i^{\text{initial}} \in \mathbb{R}^{c_s}\right.$ 

$$\left. \{T_i^{\text{true}, f}\}, \{T_i^{\text{alt truth}, f}\}, \{\vec{\alpha}_i^{\text{true}, f}\}, \{\vec{\alpha}_i^{\text{alt truth}, f}\}, \{\vec{\mathbf{x}}_i^{\text{true}, a}\}, \{\vec{\mathbf{x}}_i^{\text{alt truth}, a}\} \right) :$$

```

- 1: $\mathbf{s}_i^{\text{initial}} \leftarrow \text{LayerNorm}(\mathbf{s}_i^{\text{initial}})$
- 2: $\mathbf{z}_{ij} \leftarrow \text{LayerNorm}(\mathbf{z}_{ij})$
- 3: $\mathbf{s}_i = \text{Linear}(\mathbf{s}_i^{\text{initial}}) \quad \mathbf{s}_i \in \mathbb{R}^{c_s}$
- 4: $T_i = (\mathbf{I}, \vec{\mathbf{0}}) \quad \mathbf{I} \in \mathbb{R}^{3 \times 3}, \vec{\mathbf{0}} \in \mathbb{R}^3$
- 5: **for all** $l \in [1, \dots, N_{\text{layer}}]$ **do** *# shared weights*
- 6: $\{\mathbf{s}_i\} += \text{InvariantPointAttention}(\{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\}, \{T_i\})$
- 7: $\mathbf{s}_i \leftarrow \text{LayerNorm}(\text{Dropout}_{0.1}(\mathbf{s}_i))$
- 8: **# Transition.** $\mathbf{s}_i \leftarrow \mathbf{s}_i + \text{Linear}(\text{relu}(\text{Linear}(\text{relu}(\text{Linear}(\mathbf{s}_i)))))$ all intermediate activations $\in \mathbb{R}^{c_s}$
- 9: $\mathbf{s}_i \leftarrow \text{LayerNorm}(\text{Dropout}_{0.1}(\mathbf{s}_i))$
- 10: **# Update backbone.** $T_i \leftarrow T_i \circ \text{BackboneUpdate}(\mathbf{s}_i)$
- 11: **# Predict side chain and backbone torsion angles** $\omega, \phi, \psi, \chi_1, \chi_2, \chi_3, \chi_4$

Figura 5.19: Structure Module, prime 10 righe dello pseudocodice. Le righe qui fatte notare sono la 6 a la 10. Fonte[118]

L’aggiornamento dei frame della backbone (algoritmo 20 riga 10) avviene tramite la predizione di un *quaternione* per la rotazione e di un vettore per la traslazione. L’utilizzo dei quaternioni è utile perché...

5.1.3 Altri dettagli

La citazione ironica di AlQuraishi raccoglie forse al meglio le scelte strutturali in AlphaFold:

«*For AlphaFold2, the apparent answer that DeepMind gave to the question of what they should do is... yes. Self-supervision? Yes. Self-distillation? Yes. New loss function? Yes. 3D refinement? Yes. Recycling after refinement? Yes. Refinement after recycling? Yes. Templates? Yes. Full MSAs? Yes. Tied-weights? Yes. Non-tied weights? Yes. Attention over nodes? Yes. Attention over edges? Yes. Attention over coordinates? Yes. The answer, to all the questions, is yes! And this clearly paid off.*»⁸

Loss

La rete è allenata end-to-end con gradienti provenienti dalla funzione di *loss* (funzione obiettivo/di costo) FAPE (Frame Aligned Point Error) e da altre loss ausiliarie:

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}$$

dove per *aux* si intende "auxiliary loss", per *dist* "cross-entropy loss for distogram prediction", per *msa* "cross-entropy loss for masked MSA prediction", per *conf* "model confidence loss". La loss totale nella fasi di inferenza comprende anche le loss *exp resolved* che sta per "experimentally resolved loss" e *viol* che sta per "violation loss".

La loss finale in AF2 è quindi una somma pesata di loss ausiliarie multiple, che non sono necessariamente correlate con le performance ma possono fornire informazioni aggiuntive. Non viene calcolata la loss solo dell'ultima struttura finale dopo le iterazioni, ma viene calcolata per ogni iterazione. È presente anche una *distogram loss* dove la struttura predetta è usata per generare un distogramma (matrice 2D di intervalli di probabilità di distanze) da confrontare con la "realtà di base" (*ground truth*, ovvero dalle strutture PDB con più accuratezza).

Un'altra loss interessante è l'*MSA masking*. Ad ogni step al modello viene fornita una MSA con alcuni simboli "mascherati" e gli viene chiesto di predirli. È un modo per crearsi da sé un apprendimento supervisionato, come reso popolare da BERT, ed è usato sia nella

⁸"The AlphaFold2 Method Paper: A Fount of Good Ideas." (25 lug. 2021), indirizzo: <https://moalquraishi.wordpress.com/2021/07/25/the-alphafold2-method-paper-a-fount-of-good-ideas> (visitato il 11/02/2022)

fase di training che di inferenza.

Un altro dettaglio è la *self-distillation*. L’architettura di AF2 è in grado di allenarsi con grande accuratezza solamente tramite il *supervised learning* sui dati provenienti dal PDB. È stato però trovato il modo di aumentare l’accuratezza usando un approccio simile al *noisy student self-distillation*^[121]. In questa procedura viene usata una rete già addestrata per predire la struttura di circa 350.000 sequenze diverse da Uniclust30; viene poi creato un nuovo set di dati di strutture previste, filtrate in un sottoinsieme ad alta confidenza. Viene poi addestrata di nuovo la stessa architettura da zero utilizzando una combinazione di dati dal PDB e da questo nuovo set di dati di strutture previste come *training data*. Questa procedura ha lo scopo di fare un uso efficace delle sequenze non etichettate e aumenta considerevolmente l’accuratezza della rete risultante (vedi fig. 5.20).

Training

AF2 viene addestrato in un modo apparentemente strano: non su intere proteine, ma su frammenti di alcune, ciò che il team di AF2 chiama *crops*. Di solito i *crops* sono composti da un paio di centinaia di residui, quindi solo una frazione di grandi proteine.

Sorprendentemente, mentre AF2 è principalmente addestrato su frammenti fino a 256 residui (successivamente perfezionato a 384), può prevedere strutture proteiche con ben oltre 2.000 residui. Sembra un task quasi impossibile in apparenza: come si è già visto il contesto globale è fondamentale nel protein folding, due sottosequenze di amminoacidi uguali, in due differenti proteine, non si ripiegano allo stesso modo in genere. Ci sono però due fattori che consentono ad AF2 di affrontare la sensibilità al contesto:

- AF2 lavora con MSA o pattern coevolutivi, che codificano informazioni a prescindere dalla separazione nella catena
- durante la fase di inferenza AF2 usa l’intera sequenza

Quest’idea di disaccoppiare cose generalmente accoppiate stride con i modelli comuni nel ML dove le fasi di training e inferenza vengono tenute molto simili, basandosi sull’idea che più i due processi sono simili migliore sarà la predizione finale. In questo caso nella fase di allenamento è importante che il modello acquisisca informazioni con aggiornamenti dei gradienti. Nonostante AF2 non sia l’unica architettura ad aver adottato questa strategia (es. modelli generativi) essa è un’implementazione robusta dell’idea. È possibile che sia stata progettata in questo modo solamente per efficienza di memoria (sarebbe impossibile allenare su proteine intere una struttura della grandezza di AF2) ma tale scelta si è rivelata una buona idea anche dal punto di vista biofisico.

5.1.4 Analisi via ablazione

Per *ablazione* si intende la valutazione delle prestazioni del sistema rimuovendo uno o più componenti non essenziali o combinazioni di questi. Questo studio è interessante perché può rivelare l'importanza e l'efficacia di determinati componenti. Non si possono trarre conclusioni generali ma si può ipotizzare quali siano le parti più importanti per generare predizioni di qualità.

Il modello *baseline* è il modello come descritto nel paper ad eccezione del meccanismo di *self-distillation*. Viene utilizzato come base per i confronti in questi studi di ablazione.

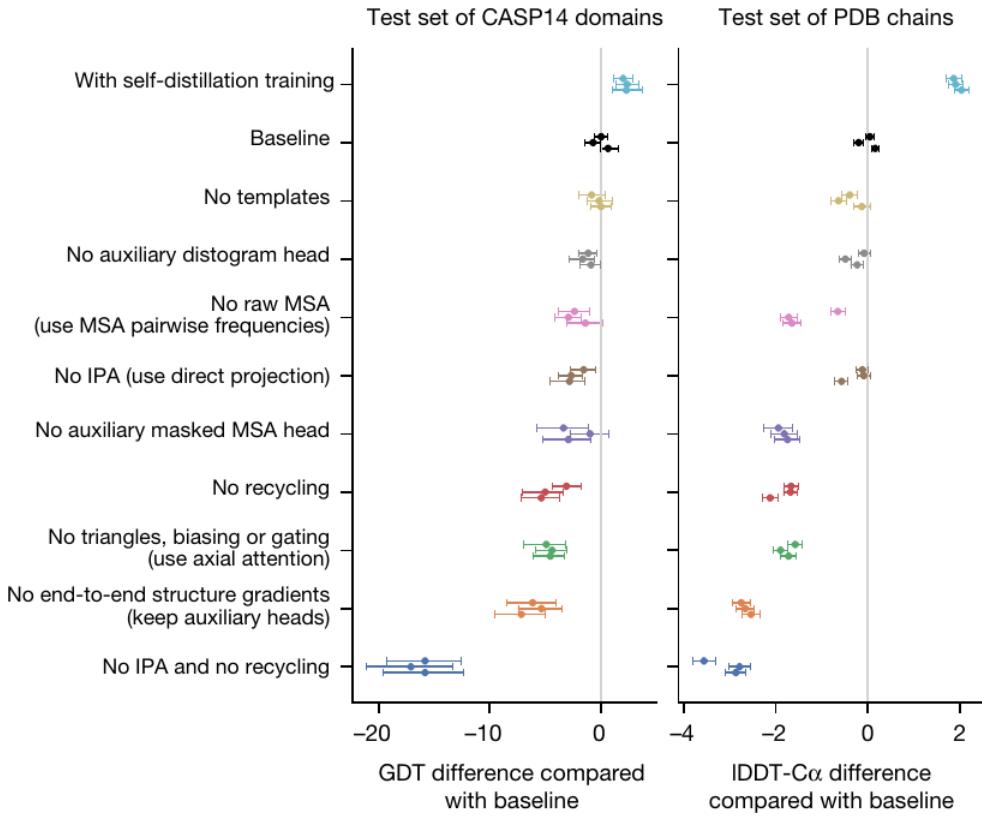
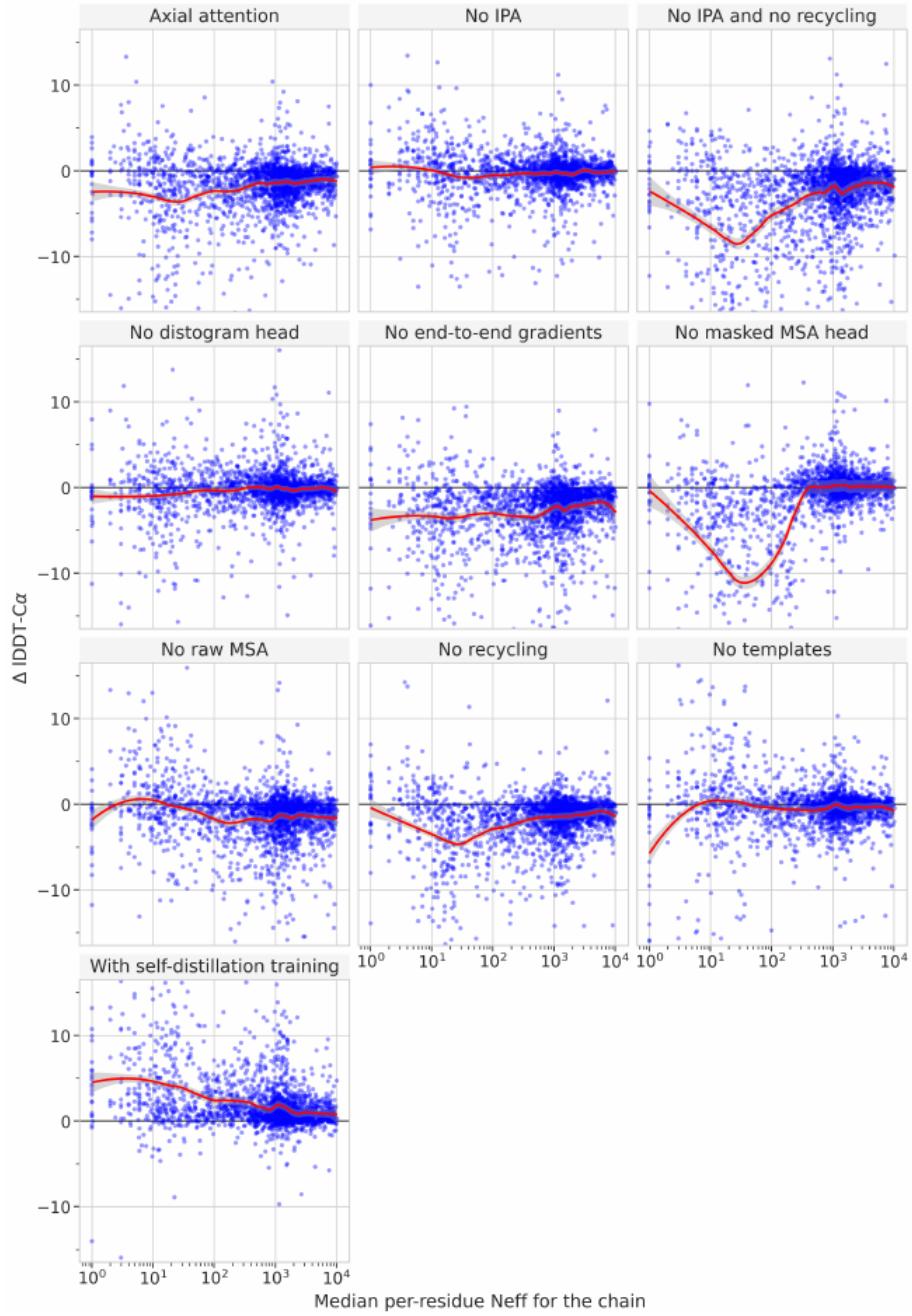


Figura 5.20: Risultati di ablazione di vari componenti su due target set: insieme di domini del CASP14 ($n=87$), PDB test set di catene con copertura di identità minore del 30% ($n=2.261$). Fonte: [113]

Ad esempio una caratteristica che può saltare all'occhio è l'apparente bassa influenza dell'IPA, nonostante sia una struttura complessa a cui il team ha dedicato molto tempo. Senza IPA lo structure module si basa solo sulla rappresentazione 1D per la generazione della struttura. Il punto fondamentale è che le performance non cambiano molto se si toglie l'IPA, a patto che il recycling venga lasciato. Quando entrambi sono tolti si può vedere (fig. 5.20) che le performance calano nettamente. Tuttavia se viene rimosso il recycling ma lasciata l'IPA le performance non subiscono grandi differenze, ed è importante notare ciò in quanto mostra che l'IPA è una struttura incredibilmente efficiente rispetto all'evoformer:



*Figura 5.21: Accuratezza in esperimenti di ablazione relativi alla baseline per differenti valori di profondità dell'MSA su recenti insiemi di proteine dal PDB, filtrati da una copertura da template <30% ($n=2.261$).
Fonte [118]*

con il recycling vengono triplicati i 48 blocchi dell'evoformer mentre l'IPA è composta di soli 8 layer.

In figura 5.21 si può notare l'accuratezza di AF2 quando vengono tolti alcuni componenti, su sequenze con MSA poco profonde. Un dettaglio che risalta è l'importanza della funzione di loss relativa al mascheramento dell'MSA e di mantenere almeno uno fra IPA

e recycling. Grazie a questi accorgimenti AF2 riesce ad avere ottima accuratezza anche quando l'MSA è poco profonda ed è forse proprio questo uno dei più grandi raggiungimenti di DeepMind.

5.1.5 Differenze con AF1

AlphaFold1 lavora sulla premessa che data una sequenza proteica, è possibile costruire un potenziale appreso e specifico per le proteine, addestrando una rete neurale profonda (DNN) per fare previsioni accurate sulla struttura e per prevedere la struttura stessa riducendo al minimo il potenziale mediante discesa del gradiente.

Le caratteristiche usate nella DNN sono caratteristiche MSA generate eseguendo HH-blits e PSI-BLAST su database di sequenze. La DNN prevede gli angoli di torsione della backbone e la distanza a coppie tra i residui. Quindi, la distanza prevista e le distribuzioni di probabilità di torsione insieme alle interazioni di van der Walls vengono combinate per formare un potenziale specifico della proteina. Infine, viene eseguita la discesa del gradiente sul potenziale specifico della proteina per ottenere il modello proteico finale.

I dati di addestramento per il modello di AF1 vengono estratti dai domini PDB e più specificamente CATH in cui sono state utilizzate 29.427 proteine per l'addestramento e 1820 proteine vengono utilizzate per i test. Le buone prestazioni di AlphaFold sono attribuite all'accuratezza delle previsioni di distanza^[73].

L'idea di ridurre al minimo il potenziale mediante la discesa del gradiente piuttosto che utilizzare l'assemblaggio dei frammenti e la successiva raffinazione del modello è piuttosto nuova.

La prima macro differenza tra i due sistemi è che AlphaFold 1 (AF1) conteneva moduli addestrati separatamente, mentre AlphaFold2 lo ha sostituito con un sistema di sottoretti accoppiate insieme in un sistema di deep learning end-to-end formato come un'unica struttura integrata.

Rispetto alla prima iterazione di AlphaFold apparsa nel CASP13, guidata dalla previsione della *distance map* basata su CNN, uno dei principali nuovi sviluppi di AlphaFold2 è l'architettura della rete neurale basata sull'*attenzione* che interviene arbitrariamente sull'intera MSA.

Inoltre, invece di utilizzare l'ottimizzazione della discesa del gradiente per costruire modelli basati sui vincoli di distanza previsti, come ha fatto AlphaFold in CASP13, AlphaFold2 utilizza un sistema di addestramento completo end-to-end dalla sequenza ai modelli di struttura, utilizzando il raffinamento strutturale iterativo basato sulla stima dell'errore locale. AF2 sostituisce le tradizionali simulazioni di ripiegamento con un modulo strutturale composto da reti neurali di transformer equivarianti 3D, che trattano ciascun

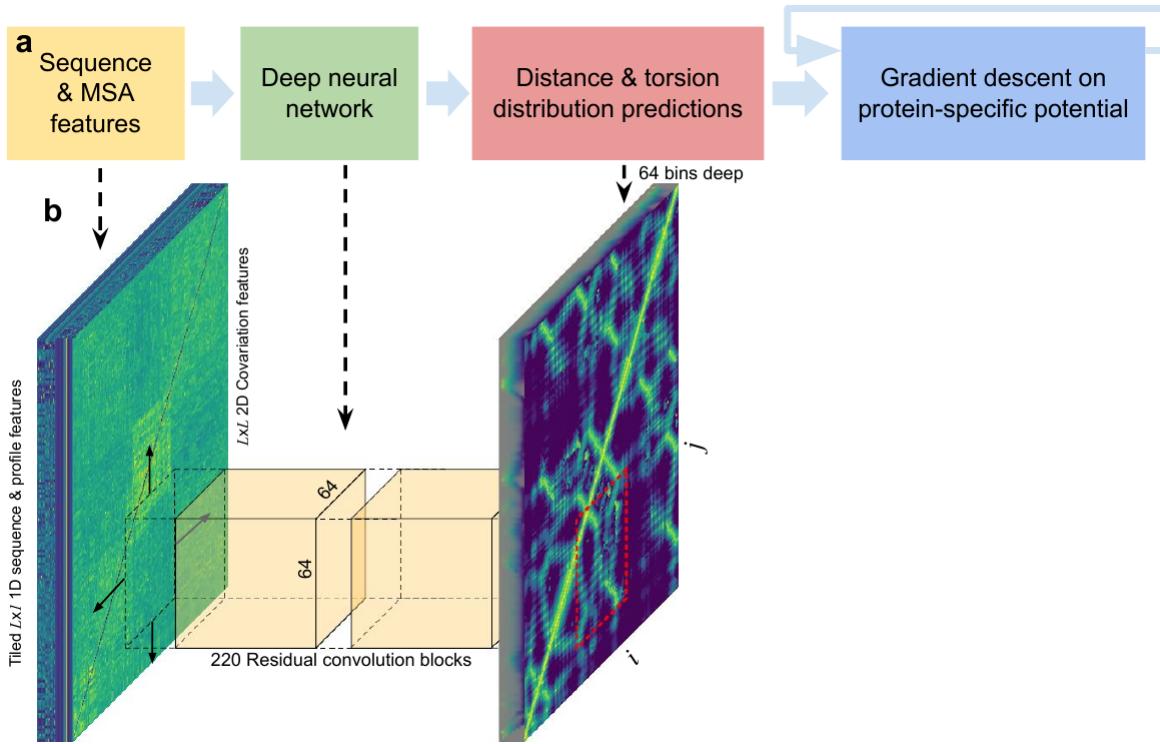


Figura 5.22: Struttura di AF1. Length $L = 155$. (a) Step della predizione della struttura. (b) La rete neurale predice l'intero distogramma $L \times L$ basato su caratteristiche dell'MSA, accumulando predizioni separate per regioni di residui 64×64 . Fonte[122]

amminoacido come un gas di corpi rigidi 3D e costruiscono direttamente la backbone proteica e le catene laterali.

5.2 DeepMind

AlphaFold è un sistema di *Artificial Intelligence* (AI) sviluppato da DeepMind che, come già visto, realizza predizioni allo stato dell'arte sulla struttura delle proteine basandosi sulle loro sequenze amminoacidiche.

DeepMind è un'azienda inglese di Intelligenza Artificiale sussidiaria di Alphabet Inc.⁹. La missione a lungo termine di AlphaFold è avanzare il progresso scientifico risolvendo problemi scientifici fondamentali attraverso l'uso di sistemi di AI.

DeepMind è stata fondata nel 2010 da Demis Hassabis, Shane Legg e Mustafa Suleyman. La società ha sede a Londra con centri di ricerca in Canada, Francia e Stati Uniti[123].

⁹In altre parole DeepMind è una società controllata: Alphabet Inc. detiene la maggioranza dei voti nell'assemblea ordinaria o un'influenza dominante sull'amministrazione.

Può risultare interessante osservare la correlazione fra i primi lavori di DeepMind e la vita di Demis Hassabis, una vita ricca di sfaccettature: bambino prodigo nel gioco degli scacchi, programmatore di videogiochi (dai 17 anni) passando per una laurea in *Computer Science*, alla fondazione del proprio studio videoludico (Elixir Studios) per poi ritornare nel mondo accademico per ottenere il suo PhD in neuroscienze cognitive nel 2009, campo nel quale ha coautorato numerosi articoli influenti su memoria e amnesia (es. rappresentazione della memoria episodica tramite *scene construction*^[124])^[125]. Per arrivare infine a fondare DeepMind e la nuovissima società Isomorphic Labs, sempre sussidiaria di Alphabet Inc. che si pone come obiettivo quello di reimaginare il processo di *drug discovery* con un approccio basato principalmente sull'AI.

DeepMind iniziò infatti a focalizzarsi sull'insegnare ad un sistema di AI come giocare a vecchi videogiochi anni '70, '80 (es. Pong, Breakout, Space Invaders), per poi passare al gioco del Go, al protein folding e recentemente alla programmazione competitiva automatizzata^[126]. DeepMind è stata acquistata da Google nel 2014 per 500 milioni di dollari^[127].

Etica

Dopo l'acquisizione di Google l'azienda ha stabilito un'*AI ethics board*. DeepMind è uno dei membri fondatori di *Partnership on AI* insieme ad Amazon, Google, Facebook, IBM e Microsoft, un'organizzazione dedicata all'interfaccia società-AI^[128].

DeepMind ha anche aperto una nuova unità denominata DeepMind Ethics and Society e si è concentrata sulle questioni etiche e sociali sollevate dall'intelligenza artificiale avendo come consulente il famoso filosofo Nick Bostrom. Nell'ottobre 2017, DeepMind ha lanciato un nuovo gruppo di ricerca per studiare l'etica dell'IA.

Alphabet

Alphabet è un'azienda statunitense fondata nel 2015 dagli stessi fondatori di Google (Larry Page e Sergey Brin) come *holding* a cui fa capo Google LLC e altre società sussidiarie: oltre a DeepMind vi sono Calico, CapitalG, Waymo, Wing, Intrinsic, Nest Labs, Sidewalk Labs, Isomorphic Labs, ecc.

Da dicembre 2019 il CEO di Alphabet è Sundar Pichai^[129]. La fondazione di Alphabet a partire da Google è stata una scelta finalizzata a rendere più trasparenti le attività inerenti a Google e concedere una maggiore autonomia alle società del gruppo che operano in settori diversi da quello dei servizi internet.

Capitolo 6

Conclusione

Il campo della predizione della struttura di proteine è stato rivoluzionato nel 2020 grazie allo sviluppo di AlphaFold2 da parte di DeepMind. Il rilascio al pubblico del codice di AlphaFold2 significa che predire la struttura di una proteina a partire dalla sua sequenza amminoacidica è, nella maggior parte dei casi, un problema risolto. Questo non vuol dire che le predizioni siano perfette o che non vi siano famiglie predette in modo poco accurato. AlphaFold2 si rivela tuttavia uno strumento fondamentale poiché permette al mondo della ricerca di ottenere informazioni strutturali su una sequenza di amminoacidi economicamente ed in poche ore.

AlphaFold2 non rivoluziona scoprendo un nuovo principio correlato al protein folding. Il segreto dietro al suo successo è un altissimo livello di ingegneria nel deep learning, sfruttando fino all'ultima goccia la mole di conoscenze disponibili sulla struttura delle proteine. La rivoluzione di AF2 è importante anche per le considerazioni sulla ricerca che innalza.

6.1 Sfide aperte

Il problema del *protein folding* non è risolto. È stato risolto, parzialmente, solo il problema della predizione della struttura di proteine.

Limiti di AlphaFold

Le predizioni di AF2 sono ragionevoli anche per alcuni casi molto complicati, come proteine ideate e anticorpi, dove i MSA non sono sempre informativi. Tuttavia AlphaFold2 non predice sempre in modo accurato (vedi fig. 6.1, dove più del 15% delle predizioni su nuove strutture nel PDB ha un'accuratezza inferiore agli 8Å). Nonostante la predizione di AF2 risulti robusta anche in caso di assenza di template e di MSA poco profondi, non risulta

altrettanto robusta su MSA quasi interamente non informativi (ad es. composti di 1 o 2 sequenze).

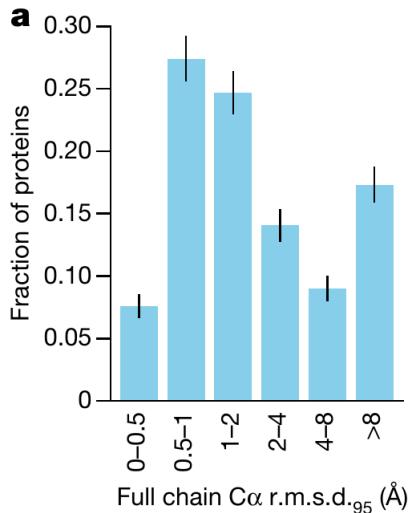


Figura 6.1: Accuratezza di AF2 su strutture recenti nel PDB. Istogrammi del backbone RMSD per catene intere, sono escluse le proteine con un template dal training set con più del 40% di identità di sequenza ($n=3.144$). Fonte[113]

È importante notare che il PDB, ovvero il training set di AlphaFold, ha un forte bias verso le proteine facilmente cristallizzabili: alcune famiglie di proteine sono sotto rappresentate e ciò di conseguenza influenza il comportamento predittivo di AlphaFold. Sono osservazioni come questa che alzano questioni sull'eccessiva centralità dei dati nella ricerca scientifica (ricerca definita appunto (*data-centric*)).

Scenari futuri

L'applicazione più diretta di AF2 è probabilmente il *drug discovery* basato sulla struttura. Fino a prima di AF2 la disponibilità di strutture era un requisito fondamentale: senza di esse non si sarebbe neanche cominciato un progetto. Isomorphic Labs, la nuova società di Demis Hassabis, si basa sugli avanzamenti di AlphaFold per reimaginare il processo di *drug discovery*.

L'ingegneria genetica può essere impiegata per produrre nuove proteine ed enzimi che contengono nuove strutture o svolgono compiti insoliti: metabolizzare rifiuti tossici o sintetizzare farmaci salvavita, ad esempio. La maggior parte dei catalizzatori sintetici non sono paragonabili, in quanto ad efficacia nella capacità di accelerare la velocità di reazioni chimiche selezionate, rispetto agli enzimi naturali. Mentre la conoscenza su come le proteine e gli enzimi sfruttino le loro conformazioni uniche per svolgere le loro funzioni biologiche cresce sempre di più, la capacità di creare nuove proteine con funzioni utili non può che migliorare^[7].

Un'altra applicazione è invece il *protein design*. Il numero di possibili proteine è astronomico e le proteine presenti in natura non sono che una piccolissima frazione di quelle generabili. Per ideare una proteina con una funzione specifica si pensa sia necessario assicurarsi che questa si ripieghi strettamente in una struttura particolare. Finora questo processo è stato rallentato dal tempo necessario per completare un ciclo di design, espressione e determinazione della struttura. Tuttavia, se le previsioni della struttura delle proteine fossero sufficientemente buone da confermare la topologia di una proteina senza una conferma sperimentale, ciò potrebbe accelerare il ciclo di test.

Per quanto riguarda il mondo della ricerca, in particolare della bioinformatica strutturale, AF2 lo ha liberato del problema del PSP. In questo campo ci si potrà ora dedicare ad altri problemi ancora più importanti e che finora non hanno ricevuto abbastanza attenzione. Degli ambiti di ricerca ora più accessibili sono:

- *protein function prediction*
- *predizione delle varianti*
- *protein dynamics*, folding e misfolding, aggregazione, regolazione allosterica, flessibilità, disordine delle proteine, fold-switching, ecc.
- *binding*, ligandi, interazioni fra proteine, interazioni fra proteine e acidi nucleici, docking macromolecolare, ecc.

È possibile che AF2 possa essere trasformativo al punto da essere per le strutture ciò che il sequenziamento del DNA è stato per la genomica. Ogni questione biologica, da quelle molecolari a quelle cellulari, potrebbe essere ora posta in termini di ipotesi strutturali, e ciò potrebbe dar vita ad un nuovo campo come la *structural systems biology*^[114].

6.2 Etica della ricerca

Come è stato possibile realizzare AlphaFold

Ci sono vari motivi di contorno al successo di DeepMind che hanno portato questo gruppo a sviluppare un sistema software in grado di risolvere il problema della predizione della struttura di proteine. È possibile elencare almeno 3 motivi principali:

- capacità del team e velocità di comunicazione
- potenza computazionale senza limiti
- tanti dati e conoscenza dalla ricerca

Prima di tutto c'è il gruppo di ricerca in sé organizzato da DeepMind. Oltre alle competenze dimostrate, un gruppo di ricerca privato ha la possibilità di sperimentare e scambiarsi idee e informazioni molto più velocemente dei gruppi accademici. L'organizzazione è differente, così come gli obiettivi a lungo termine. Il paradigma di lavoro nella ricerca privata si può definire di tipo *fast and focused*. Il motivo per cui il paper di AlphaFold ha circa 20 coautori principali è che tutti vanno nella stessa direzione.

C'è poi anche l'aspetto della (praticamente) illimitata potenza computazionale a disposizione. Nonostante possa forse aver accelerato i tempi di sperimentazione, la potenza a disposizione è solo un elemento di contorno per il raggiungimento dell'alto livello ingegneristico di AlphaFold.

DeepMind afferma di aver utilizzato approssimativamente 128 TPUv3 core in funzione per varie settimane, una quantità di risorse non eccessiva nel panorama attuale dello stato dell'arte del ML. Le TPU sono state sviluppate da Google appositamente per il deep learning e nelle giuste mani possono mostrare la loro marcia in più. Una tra le loro caratteristiche più importanti è la quantità di memoria: un chip 8-core TPUv3 ha 128GB di vRAM, dove invece le comuni GPU arrivano a circa 40-80GB. Il costo di noleggio annuale di 128 TPU può aggirarsi intorno al milione di dollari^[117].

I gruppi Baker e Zhang, i migliori dopo AlphaFold nel CASP14, per confronto hanno riferito di avere usato 4 GPU per allenare i loro modelli per un paio di settimane. Ciò che per questi due gruppi richiede un mese di sperimentazione potrebbe richiedere solo qualche ora per DeepMind. È quindi giusto considerare questo fattore come un vantaggio, che consente sia un efficace *rapid prototyping* per testare varie idee, che una base solida su cui poggiare un'architettura come quella di AlphaFold.

Un motivo invece più interessante da sottolineare è l'enorme quantità di dati e conoscenze prodotti e pubblicati dai gruppi accademici di ricerca negli ultimi decenni. AlphaFold, come tutti i sistemi di ML, può funzionare grazie alla presenza di grandi quantità di dati (milioni di sequenze e migliaia di strutture) e grazie alla conoscenza scientifica nel settore. Spesso la precedente ricerca è stata possibile grazie a fondi pubblici o iniziative accademiche finanziate dai governi, come i tool HHblits, JackHMMER e OpenMM usati da AF2. In altre parole, DeepMind è riuscita a vedere lontano e a realizzare AlphaFold perché si trova sulle *spalle di giganti*.

Ricerca accademica e industriale

Il raggiungimento di DeepMind può essere visto come un'accusa nei confronti del mondo accademico. Per un lungo periodo se si voleva fare ricerca e risolvere problemi di ricerca (come il PSP) il percorso che si delineava era trovare un lavoro nel sistema accademico.

Oggi, nel 2022, la differenza è che l'industria non è più solo un buon posto per fare ricerca applicata ma potrebbe diventare un posto dove fare quasi esclusivamente tutta la ricerca applicata. Tuttavia ci sono alcune considerazioni da fare per rendersi conto di cosa comporterebbe una tale prospettiva.

L'ultimo punto citato nella sezione precedente è particolarmente interessante poiché ci si può chiedere: se si vuole che la scienza rimanga aperta e collaborativa, fino a che punto l'informazione creata da centri di ricerca pubblici (pubblici al fine di stimolare ulteriore ricerca) appartiene alla comunità e sotto quali condizioni potrebbe essere usata per iniziative for-profit?^[117] In questo caso DeepMind ha preso la decisione di condividere al pubblico il codice di AlphaFold e di rendere lo strumento apertamente accessibile. Ma al di fuori del caso specifico, l'interrogativo rimane, specialmente con l'avvento della ricerca *data-centric*, basata sui dati e sull'intelligenza artificiale.

Riguardo allo sbilanciamento nelle risorse computazionali, una domanda che sorge è se questo impatterà la qualità della ricerca accademica in futuro. I modelli si stanno facendo sempre più complessi (vedi transformer) e la complessità cresce più di quanto cali il prezzo dell'hardware. Si potrebbe arrivare alla situazione poco sensata nella quale la ricerca accademica sia limitata a raggiungere le idee che vorrebbero inseguire dalla mancanza di disponibilità computazionale.

Per ovviare a tale problema si possono immaginare vari scenari^[117]:

- i gruppi di ricerca ricevono significativi investimenti nelle infrastrutture perché viene compreso il vantaggio che strumenti come AlphaFold potrebbero apportare
- le risorse di ricerca potrebbero essere condivise in un consorzio internazionale, come i fisici delle alte energie hanno fatto per il CERN
- si segue la strada di sviluppare strategie per ridurre l'impatto delle risorse limitate sugli strumenti di ML; ci sarebbe bisogno di più software engineer professionisti di alto livello

Il paradigma *fast and focused* dei centri di ricerca privati come DeepMind ha però dei vincoli, ad esempio^[115]:

- non pone nuove domande
- fa fiorire una sola idea

È un approccio ottimo per rispondere a delle domande precise, ma non per porre domande. Nel mondo biologico e scientifico in generale, definire domande è una parte importante del lavoro di ricerca. Non è un lavoro banale riuscire a strutturare una competizione scientifica come il CASP, e se è stato fatto è grazie allo sforzo collettivo di tanti membri della comunità scientifica. DeepMind infatti si concentra su problemi con obiettivi e metriche chiare, non sulla definizione di nuove domande.

Nonostante il paper di AF2 sia stato pubblicato, compresi alcuni studi di ablazione, internamente DeepMind potrebbe aver provato molti altri approcci ma solamente quelli vincenti vengono esplorati e condivisi. È però possibile che quei percorsi scartati, pur non contribuendo direttamente alla soluzione, avrebbero potuto produrre altri tipi di conoscenza riguardo al problema. Il paradigma in questione minimizza la circolazione ed esplorazione di idee. Non è un problema specifico di DeepMind ma di come si sta evolvendo la ricerca in sé, specialmente nel caso del ML.

Infine un'osservazione sul confronto fra ricerca accademica e industriale riguarda il livello di interessi nelle questioni pubbliche. Immaginando di dover convocare un consiglio speciale di esperti di un settore (es. epidemiologia, intelligenza artificiale) non ci si può aspettare che convocando membri di grandi aziende questi non abbiano degli interessi personali.

Grazie per la lettura

«*Il Buddha, il Divino, dimora nel circuito di un calcolatore o negli ingranaggi del cambio di una moto con lo stesso agio che in cima a una montagna o nei petali di un fiore»*¹

Ringrazio il prof. Pirchio per l'attenta revisione delle sezioni biologiche introduttive della tesi e per avermi fatto scoprire la bellezza della biologia.

Ringrazio il prof. Milazzo per avermi stimolato con un argomento interessante come quello della predizione della struttura delle proteine, per avermi lasciato molta libertà nella ricerca e nella stesura della tesi e per avermi assistito in un momento di stallo durante questo percorso.

Ringrazio la mia famiglia per la possibilità che mi ha donato di diventare la persona che sono, anche attraverso gli studi universitari.

A tutte le persone accanto a me rivolgo un particolare ringraziamento affettivo per la possibilità che mi donano di poter condividere l'armonia di fondo della vita: l'amore.

«*L'amore è la forza che trasforma e migliora l'anima del mondo»*²

¹R. M. Pirsig, *Lo Zen e l'arte della manutenzione della motocicletta*, 1974

²P. Coelho, *L'alchimista*, 1988

Bibliografia

Libri

- [6] A. Kessel e N. Ben-Tal, *Introduction to proteins: Structure, function and motion*, 2^a ed. Chapman e Hall/CRC, 2018.
- [7] B. Alberts, D. Bray, K. Hopkin et al., *Essential cell biology*, 5^a ed. W. W. Norton e Company, 2019.
- [8] N. A. Campbell, J. B. Reece, L. A. Urry, R. Brizzi, T. Niccolò e A. Bartalesi, *Biologia E Genetica*. Pearson, 2012.
- [16] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Orr e N. A. Campbell, *Campbell Biology*. Pearson, 2021.
- [30] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.
- [38] T. Mitchell, *Machine learning*. McGraw hill New York, 1997.
- [48] S. Pal, *Fundamentals of Molecular Structural Biology*. Academic Press, 2019.
- [53] L. A. Moran, H. R. Horton, K. G. Scrimgeour, M. D. Perry e D. Rawn, *Principles of biochemistry*. Pearson London, 2012.
- [57] D. L. Nelson, A. L. Lehninger e M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2017.
- [82] L. Fleck, *Genesi e Sviluppo di un Fatto Scientifico: Per Una Teoria dello stile e del collettivo di pensiero*. Il mulino, 1983.

Articoli

- [19] O. Hidalgo, J. Pellicer, M. Christenhusz, H. Schneider, A. R. Leitch e I. J. Leitch, “Is there an upper limit to genome size?” *Trends in Plant Science*, vol. 22, n. 7, pp. 567–573, 2017.

- [28] M. Batool, B. Ahmad e S. Choi, “A structure-based drug discovery paradigm,” *International journal of molecular sciences*, vol. 20, n. 11, p. 2783, 2019.
- [29] B. C. Knott, E. Erickson, M. D. Allen et al., “Characterization and engineering of a two-enzyme system for plastics depolymerization,” *Proceedings of the National Academy of Sciences*, vol. 117, n. 41, pp. 25 476–25 485, 2020.
- [31] F. C. Bernstein, T. F. Koetzle, G. J. Williams et al., “The protein data bank: A computer-based archival file for macromolecular structures,” *Journal of Molecular Biology*, vol. 112, n. 3, pp. 535–542, 1977.
- [32] M. Mitchell, “Biological Computation,” *PDXScholar*, 2010. indirizzo: https://pdxscholar.library.pdx.edu/compsci_fac/2.
- [40] M. Torrisi, G. Pollastri e Q. Le, “Deep learning methods in protein structure prediction,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, 2020.
- [46] H. Wu e E. Yang, “Studies on denaturation of proteins. XL Effect of hydrogen ion concentration on rate of denaturation of egg albumin by urea. A theory of denaturation,” *Chin J Physiol*, vol. 5, pp. 301–344, 1931.
- [47] C. B. Anfinsen, “The formation and stabilization of protein structure.,” *Biochemical Journal*, vol. 128, n. 4, p. 737, 1972.
- [49] C. B. Anfinsen, E. Haber, M. Sela e F. White Jr, “The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 47, n. 9, p. 1309, 1961.
- [52] K. A. Dill, S. B. Ozkan, M. S. Shell e T. R. Weikl, “The protein folding problem,” *Annu. Rev. Biophys.*, vol. 37, pp. 289–316, 2008.
- [54] J. Murray, N. Laurieri e R. Delgoda, “Chapter 24 - Proteins,” S. Badal e R. Delgoda, cur., pp. 477–494, 2017. DOI: <https://doi.org/10.1016/B978-0-12-802104-0.00024-X>.
- [59] N. A. Ranson, H. E. White e H. R. Saibil, “Chaperonins,” *Biochemical Journal*, vol. 333, n. 2, pp. 233–242, 1998.
- [60] R. Iizuka e T. Funatsu, “Chaperonin GroEL uses asymmetric and symmetric reaction cycles in response to the concentration of non-native substrate proteins,” *Biophysics and Physicobiology*, vol. 13, pp. 63–69, 2016.
- [62] S. B. Prusiner, M. R. Scott, S. J. DeArmond e F. E. Cohen, “Prion protein biology,” *cell*, vol. 93, n. 3, pp. 337–348, 1998.

- [63] B. Ruttkay-Nedecky, E. Sedlackova, D. Chudobova et al., “Prion protein and its interactions with metal ions (Cu^{2+} , Zn^{2+} , and Cd^{2+}) and metallothionein 3,” *ADMET and DMPK*, vol. 3, n. 3, pp. 287–295, 2015.
- [64] D. Fraser-Pitt e D. O’Neil, “Cystic fibrosis—a multiorgan protein misfolding disease,” *Future science OA*, vol. 1, n. 2, 2015.
- [66] M. Sharon, “How far can we go with structural mass spectrometry of protein complexes?” *Journal of the American Society for Mass Spectrometry*, vol. 21, n. 4, pp. 487–500, 2011.
- [68] X. Fan, J. Wang, X. Zhang et al., “Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution,” *Nature communications*, vol. 10, n. 1, pp. 1–11, 2019.
- [71] E. Callaway, “Revolutionary cryo-EM is taking over structural biology..,” *Nature*, vol. 578, n. 7794, pp. 201–202, 2020.
- [72] X.-C. Bai, G. McMullan e S. H. Scheres, “How cryo-EM is revolutionizing structural biology,” *Trends in biochemical sciences*, vol. 40, n. 1, pp. 49–57, 2015.
- [73] S. C. Pakhrin, B. Shrestha, B. Adhikari, D. B. Kc et al., “Deep learning-based advances in protein structure prediction,” *International Journal of Molecular Sciences*, vol. 22, n. 11, p. 5553, 2021.
- [75] L. L. Porter e L. L. Looger, “Extant fold-switching proteins are widespread,” *Proceedings of the National Academy of Sciences*, vol. 115, n. 23, pp. 5968–5973, 2018.
- [76] A. E. Varela, K. A. England e S. Cavagnero, “Kinetic trapping in protein folding,” *Protein Engineering, Design and Selection*, vol. 32, n. 2, pp. 103–108, 2019.
- [77] E. Fischer, “Einfluss der Configuration auf die Wirkung der Enzyme,” *Berichte der deutschen chemischen Gesellschaft*, vol. 27, n. 3, pp. 2985–2993, 1894.
- [78] A. K. Dunker, J. D. Lawson, C. J. Brown et al., “Intrinsically disordered protein,” *Journal of molecular graphics and modelling*, vol. 19, n. 1, pp. 26–59, 2001.
- [79] A. E. Mirsky e L. Pauling, “On the structure of native, denatured, and coagulated proteins,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 22, n. 7, p. 439, 1936.
- [80] F. Karush, “Heterogeneity of the binding sites of bovine serum albumin1,” *Journal of the American Chemical Society*, vol. 72, n. 6, pp. 2705–2713, 1950.
- [81] C. Bracken, M. M. Young e K. Dunker, “Disorder and flexibility in protein structure and function,” pp. 64–66, 2000.

- [83] J. Abbass e J.-C. Nebel, “Enhancing fragment-based protein structure prediction by customising fragment cardinality according to local secondary structure,” *BMC bioinformatics*, vol. 21, n. 1, pp. 1–23, 2020.
- [84] C. Levinthal, “How to fold graciously,” *Mossbauer spectroscopy in biological systems*, vol. 67, pp. 22–24, 1969. indirizzo: <https://web.archive.org/web/20110523080407/http://www-miller.ch.cam.ac.uk/levinthal/levinthal.html>.
- [85] R. Pearce e Y. Zhang, “Deep learning techniques have significantly impacted protein structure prediction and protein design,” *Current opinion in structural biology*, vol. 68, pp. 194–207, 2021.
- [86] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis e J. Moult, “Critical assessment of methods of protein structure prediction (CASP)—Round XIV,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, n. 12, pp. 1607–1617, 2021.
- [87] A. E. Márquez-Chamorro, G. Asencio-Cortés, C. E. Santiesteban-Toca e J. S. Aguilar-Ruiz, “Soft computing methods for the prediction of protein tertiary structures: A survey,” *Applied Soft Computing*, vol. 35, pp. 398–410, 2015.
- [88] V. Mariani, M. Biasini, A. Barbato e T. Schwede, “lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests,” *Bioinformatics*, vol. 29, n. 21, pp. 2722–2728, 2013.
- [94] M. Varadi, S. Anyango, M. Deshpande et al., “AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models,” *Nucleic acids research*, 2021.
- [95] C. Castaldo, S. Ciambellotti, R. de Pablo-Latorre, D. Lalli, V. Porcari e P. Turano, “Soluble variants of human recombinant glutaminyl cyclase,” *Plos one*, vol. 8, n. 8, e71657, 2013.
- [96] J. Gauthier, A. T. Vincent, S. J. Charette e N. Derome, “A brief history of bioinformatics,” *Briefings in bioinformatics*, vol. 20, n. 6, pp. 1981–1996, 2019.
- [97] M. Levitt e S. Lifson, “Refinement of protein conformations using a macromolecular energy minimization procedure,” *Journal of molecular biology*, vol. 46, n. 2, pp. 269–279, 1969.
- [98] J. A. McCammon, B. R. Gelin e M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, n. 5612, pp. 585–590, 1977.
- [99] M. Levitt e A. Warshel, “Computer simulation of protein folding,” *Nature*, vol. 253, n. 5494, pp. 694–698, 1975.

- [101] D. S. Marks, L. J. Colwell, R. Sheridan et al., “Protein 3D structure computed from evolutionary sequence variation,” *PLoS one*, vol. 6, n. 12, e28766, 2011.
- [102] G. Sliwoski, S. Kothiwale, J. Meiler e E. W. Lowe, “Computational methods in drug discovery,” *Pharmacological reviews*, vol. 66, n. 1, pp. 334–395, 2014.
- [104] A. Platt, H. C. Ross, S. Hankin e R. J. Reece, “The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase,” *Proceedings of the National Academy of Sciences*, vol. 97, n. 7, pp. 3154–3159, 2000.
- [105] A. P. Joseph e A. G. de Brevern, “From local structure to a global framework: recognition of protein folds,” *Journal of The Royal Society Interface*, vol. 11, n. 95, p. 20131147, 2014.
- [108] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio e R. Nussinov, “The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery..,” *Chemical reviews*, vol. 116 11, pp. 6391–423, 2016.
- [109] Y. Karami, F. Guyon, S. De Vries e P. Tufféry, “DaReUS-Loop: accurate loop modeling using fragments from remote or unrelated proteins,” *Scientific reports*, vol. 8, n. 1, pp. 1–12, 2018.
- [110] Y. Karami, J. Rey, G. Postic, S. Murail, P. Tufféry e S. J. de Vries, “DaReUS-Loop: a web server to model multiple loops in homology models,” *Nucleic Acids Research*, vol. 47, n. W1, W423–W428, mag. 2019.
- [111] A. Barozet, P. Chacón e J. Cortés, “Current approaches to flexible loop modeling,” *Current Research in Structural Biology*, vol. 3, pp. 187–191, 2021.
- [112] W. Zheng, Y. Li, C. Zhang et al., “Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14,” *Proteins: Structure, Function, and Bioinformatics*, vol. 89, n. 12, pp. 1734–1751, 2021.
- [113] J. Jumper, R. Evans, A. Pritzel et al., “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, n. 7873, pp. 583–589, 2021.
- [118] J. Jumper, R. Evans, A. Pritzel et al., “Supplementary Information for Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, n. 7873, pp. 583–589, 2021.
- [119] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [121] Q. Xie, M.-T. Luong, E. Hovy e Q. V. Le, “Self-training with noisy student improves imagenet classification,” pp. 10687–10698, 2020.

- [122] A. W. Senior, R. Evans, J. Jumper et al., “Improved protein structure prediction using potentials from deep learning,” *Nature*, vol. 577, n. 7792, pp. 706–710, 2020.
- [124] D. Hassabis e E. A. Maguire, “Deconstructing episodic memory with construction,” *Trends in Cognitive Sciences*, vol. 11, n. 7, pp. 299–306, lug. 2007, ISSN: 1364-6613. DOI: 10.1016/j.tics.2007.05.001.

Risorse Online

- [1] “PDB 4v60 structure summary. Protein Data Bank in Europe (PDBe). EMBL-EBI.” (9 lug. 2014), indirizzo: <https://www.ebi.ac.uk/pdbe/entry/pdb/4v60> (visitato il 07/02/2022).
- [4] “enzima nell’Enciclopedia Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/enciclopedia/enzima> (visitato il 21/01/2022).
- [5] “proteina in Vocabolario - Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/vocabolario/proteina> (visitato il 22/01/2022).
- [9] “Chemical element - Wikipedia.” (1 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Chemical_element (visitato il 31/01/2022).
- [10] “eukaryote. Definition, Structure, Facts.” (19 set. 2019), indirizzo: <https://www.britannica.com/science/eukaryote> (visitato il 22/01/2022).
- [11] “Neurone - Wikipedia.” (27 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/Neurone> (visitato il 23/01/2022).
- [12] “Saccharomyces cerevisiae - Wikipedia.” (25 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Saccharomyces_cerevisiae (visitato il 22/01/2022).
- [13] “Dogma centrale della biologia molecolare - Wikipedia.” (16 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Dogma_centrale_della_biologia_molecolare (visitato il 22/01/2022).
- [14] “DNA Structure. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-structure.html> (visitato il 22/01/2022).
- [17] “File: Difference DNA RNA-EN.svg - Wikimedia Commons.” (23 mar. 2010), indirizzo: https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg (visitato il 22/01/2022).
- [18] “Nasuia deltocephalinicola - Wikipedia.” (25 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Nasuia_deltoccephalinicola (visitato il 31/01/2022).

- [20] “Genome Size. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-3-genetics/32-chromosomes/genome-size.html> (visitato il 31/01/2022).
- [21] “Paris japonica - Wikipedia.” (31 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Paris_japonica (visitato il 31/01/2022).
- [22] “Transfer RNA - Wikipedia.” (23 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Transfer_RNA (visitato il 23/01/2022).
- [23] “Protein - Wikipedia.” (21 dic. 2021), indirizzo: <https://en.wikipedia.org/wiki/Protein> (visitato il 23/01/2022).
- [24] “Peptide bond - Wikipedia.” (4 nov. 2021), indirizzo: https://en.wikipedia.org/wiki/Peptide_bond (visitato il 23/01/2022).
- [25] “Amino Acids. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/amino-acids.html> (visitato il 23/01/2022).
- [26] “PDB101: Learn: Videos: What is a Protein?” (20 Nov. 2017), indirizzo: <https://pdb101.rcsb.org/learn/videos/what-is-a-protein-video> (visitato il 23/01/2022).
- [27] K. Dill. “The protein folding problem: a major conundrum of science: Ken Dill at TEDxSBU.” (23 ott. 2013), indirizzo: <https://www.youtube.com/watch?v=zmv3kovWpNQ> (visitato il 06/01/2022).
- [33] “Apprendimento automatico - Wikipedia.” (1 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/ApprendimentoAutomatico> (visitato il 23/01/2022).
- [34] “What is soft computing - Javatpoint.” (3 lug. 2021), indirizzo: <https://www.javatpoint.com/what-is-soft-computing> (visitato il 24/01/2022).
- [37] “Machine Learning - IBM.” (29 ago. 2020), indirizzo: <https://www.ibm.com/it-it/analytics/machine-learning> (visitato il 23/01/2022).
- [39] “What are Neural Networks?” (1 Giu. 2021), indirizzo: <https://www.ibm.com/cloud/learn/neural-networks> (visitato il 24/01/2022).
- [41] “What are Recurrent Neural Networks?” (14 Set. 2020), indirizzo: <https://www.ibm.com/cloud/learn/recurrent-neural-networks> (visitato il 07/02/2022).
- [42] “Residual neural network - Wikipedia.” (30 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Residual_neural_network (visitato il 07/02/2022).

- [43] “AlphaFold 2 presentation at CASP14, Jumper J.” (1 dic. 2020), indirizzo: https://predictioncenter.org/casp14/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf (visitato il 14/02/2022).
- [45] “Protein Structure .BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/protein-structure.html> (visitato il 27/01/2022).
- [50] “Christian B. Anfinsen - Digital Collections - National Library of Medicine.” (1 gen. 2022), indirizzo: <http://resource.nlm.nih.gov/101408166> (visitato il 25/01/2022).
- [51] “File:RibonucleaseA SS paleRib.png - Wikimedia Commons.” (16 mar. 2012), indirizzo: https://commons.wikimedia.org/wiki/File:RibonucleaseA_SS_paleRib.png (visitato il 25/01/2022).
- [55] “Protein structure prediction - Wikipedia.” (30 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Protein_structure_prediction (visitato il 27/01/2022).
- [56] L. A. Moran. “Levels of Protein Structure.” (13 mar. 2008), indirizzo: <https://sandwalk.blogspot.com/2008/03/levels-of-protein-structure.html> (visitato il 27/01/2022).
- [58] “File:Ramachandran’s Diagram.jpg - Wikipedia.” (21 giu. 2016), indirizzo: https://it.wikipedia.org/wiki/File:Ramachandran%27s_Diagram.jpg (visitato il 28/01/2022).
- [61] “Chaperonina - Wikipedia.” (21 ott. 2021), indirizzo: <https://it.wikipedia.org/wiki/Chaperonina> (visitato il 26/01/2022).
- [65] “Inibitori della proteasi - Wikipedia.” (17 ago. 2021), indirizzo: https://it.wikipedia.org/wiki/Inibitori_della_proteasi (visitato il 04/02/2022).
- [67] “What is Cryo-EM?” (21 Giu. 2018), indirizzo: <https://cryoem.slac.stanford.edu/what-is-cryo-em> (visitato il 06/02/2022).
- [69] “The Resolution Revolution: Building a Better Microscope to See at the Atomic Level.” (1 giu. 2015), indirizzo: <https://www.ucsf.edu/news/2015/06/129836/resolution-revolution-building-better-microscope-see-atomic-level> (visitato il 06/02/2022).
- [70] “RCSB PDB - 5IRE: The cryo-EM structure of Zika Virus.” (30 mar. 2016), indirizzo: <https://www.rcsb.org/structure/5IRE> (visitato il 06/02/2022).
- [74] “Lauren Porter and Fold-Switching Proteins.” (14 feb. 2020), indirizzo: <https://www.youtube.com/watch?v=IeX5ebadgiA> (visitato il 28/01/2022).

- [89] “MSOE Center for BioMolecular Modeling - Jmol Training Guide.” (24 ago. 2012), indirizzo: <https://cbm.msoe.edu/markMyweb/jmolWebsite/2-3.html> (visitato il 07/02/2022).
- [90] “About UniProt.” (2 feb. 2021), indirizzo: <https://www.uniprot.org/help/about> (visitato il 07/02/2022).
- [91] “Latest Release Information.” (27 gen. 2022), indirizzo: <https://www.ddbj.nig.ac.jp/latest-releases-e.html> (visitato il 27/01/2022).
- [92] “RCSB PDB: Homepage.” (1 feb. 2022), indirizzo: <https://www.rcsb.org> (visitato il 03/02/2022).
- [93] “wwPDB: Deposition Statistics.” (1 feb. 2022), indirizzo: <https://www.wwpdb.org/stats/deposition> (visitato il 03/02/2022).
- [100] “Sequence Alignment. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com/options/untitled/b5-bioinformatics/sequence-alignment.html> (visitato il 02/02/2022).
- [103] “Homology modeling.” (31 mag. 2014), indirizzo: <https://www.unil.ch/pmf/home/menuinst/technologies/homology-modeling.html> (visitato il 03/02/2022).
- [107] “SAbPred: FREAD.” (2 apr. 2022), indirizzo: <http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/fread> (visitato il 04/02/2022).
- [114] “The AlphaFold2 Method Paper: A Fount of Good Ideas.” (25 lug. 2021), indirizzo: <https://moalquraishi.wordpress.com/2021/07/25/the-alphafold2-method-paper-a-fount-of-good-ideas> (visitato il 11/02/2022).
- [115] “The AlphaFold2 Method Paper: A Fount of Good Ideas.” (25 lug. 2021), indirizzo: <https://moalquraishi.wordpress.com/2021/07/25/the-alphafold2-method-paper-a-fount-of-good-ideas> (visitato il 13/02/2022).
- [116] “Groups Analysis: zscores - CASP14.” (1 dic. 2020), indirizzo: https://predictioncenter.org/casp14/zscores_final.cgi (visitato il 13/02/2022).
- [117] “CASP14: what Google DeepMind’s AlphaFold 2 really achieved, and what it means for protein folding, biology and bioinformatics | Oxford Protein Informatics Group.” (10 dic. 2020), indirizzo: <https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics> (visitato il 13/02/2022).
- [120] “Disuguaglianza triangolare - Wikipedia.” (26 dic. 2021), indirizzo: https://it.wikipedia.org/wiki/Disuguaglianza_triangolare (visitato il 15/02/2022).

- [123] “DeepMind - Wikipedia.” (12 gen. 2022), indirizzo: <https://en.wikipedia.org/wiki/DeepMind> (visitato il 12/01/2022).
- [125] “Demis Hassabis - Wikipedia.” (14 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Demis_Hassabis (visitato il 22/01/2022).
- [126] “Competitive programming with AlphaCode.” (2 feb. 2022), indirizzo: <https://www.deeplearning.ai/blog/article/Competitive-programming-with-AlphaCode> (visitato il 13/02/2022).
- [127] S. Gibbs. “Google buys UK artificial intelligence startup Deepmind for £400m.” (27 gen. 2014), indirizzo: <https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmind> (visitato il 12/01/2022).
- [128] “Home - Partnership on AI.” (22 gen. 2022), indirizzo: <https://partnershiponai.org> (visitato il 22/01/2022).
- [129] L. Feiner. “Larry Page steps down as CEO of Alphabet, Sundar Pichai to take over.” (3 dic. 2019), indirizzo: <https://www.cnbc.com/2019/12/03/larry-page-steps-down-as-ceo-of-alphabet.html> (visitato il 12/01/2021).

Altre fonti

- [2] F. Capra, *Il Tao della fisica*, 1975.
- [3] G. Lambertini, “In onore di Angelo Ruffini,” in 31 mar. 1930.
- [15] S. Bewick, R. Parsons, T. Forsythe, S. Robinson e J. Dupon, “Introductory Chemistry (CK-12),” in Libretexts, 1 giu. 2021. indirizzo: [https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_\(CK-12\)](https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Book%3A_Introductory_Chemistry_(CK-12)) (visitato il 22/01/2022).
- [35] R. Kurzweil, R. Richter, R. Kurzweil e M. L. Schneider, *The age of intelligent machines*, 1990.
- [36] H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011.
- [44] A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925.
- [106] K. Roy, S. Kar e R. N. Das, “Chapter 5 - Computational Chemistry,” in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, Boston: Academic Press, 2015, pp. 151–189.
- [130] R. M. Pirsig, *Lo Zen e l'arte della manutenzione della motocicletta*, 1974.

[131] P. Coelho, *L'alchimista*, 1988.