



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

Indice

1 Predizione della struttura di proteine	2
1.1 Metodi e strumenti informatici	2
1.1.1 Workflow e classificazione dei metodi per il PSP	3
1.1.2 Soft computing e Deep Learning	7
1.1.3 Output e misure di valutazione	9
1.1.4 Database e rappresentazione delle informazioni	12
1.1.5 Rappresentazione grafica	12
1.1.6 CASP	12
1.2 Annotazioni 1D sulla struttura	14
1.2.1 Allineamento di sequenze	14
1.2.2 Predizione della struttura secondaria	15
1.3 Annotazioni 2D sulla struttura	15
1.3.1 <i>correlated mutation</i>	16
1.3.2 Distogrammi	18
1.4 Predizione della struttura 3D	18
1.4.1 <i>homology modeling</i>	18
1.4.2 <i>fold recognition via threading</i>	22
1.4.3 <i>ab initio</i>	23
1.4.4 <i>fragment-based</i>	31
1.4.5 <i>loop modeling</i>	31
1.5 Case Study: <i>TASSER</i>	34
1.6 Studio sperimentale delle proteine	35
1.7 Storia della comprensione delle proteine	42
Bibliografia	44

Capitolo 1

Predizione della struttura di proteine

Il protein folding problem ha sia guidato che tratto beneficio dagli avanzamenti nei metodi sperimentali e computazionali^[50]. 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 e 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.

1.1 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 1.1.4) 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 1.1.6.

1.1.1 Workflow e classificazione dei metodi per il PSP

Quando si parla di metodi per la predizione della struttura di proteine esistono 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.

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 famiglie 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^[6].

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)

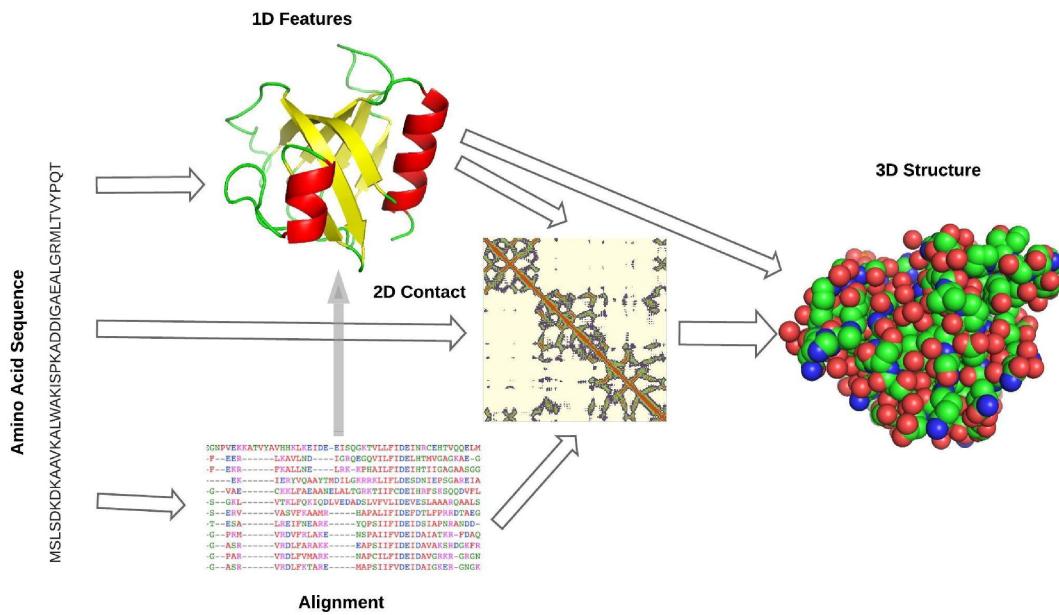


Figura 1.1: Pipeline generica per la predizione della struttura 3D di una proteina. Questo schema vuole mettere in risalto gli step intermedi relativi alle annotazioni. Fonte[39]

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 predizione di tipo *template-based modeling* (TBM), più semplice, mentre nell'altro caso si affronterà una predizione *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 utilizzati per la predizione della struttura di proteine (per ragioni che verranno spiegate in dettaglio nella sezione 1.4.3). 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. 1.2)²:

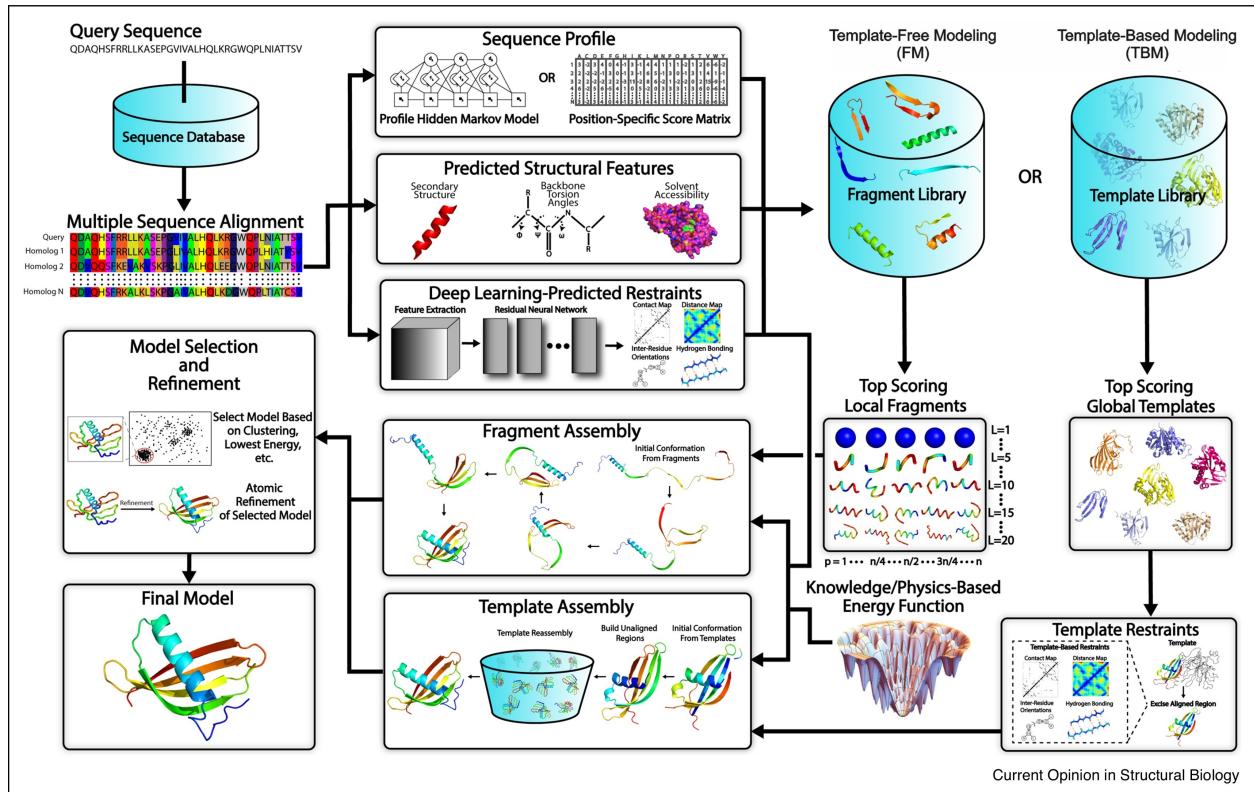


Figura 1.2: Step tipici negli approcci al PSP di tipo TBM e FM. Fonte[75]

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
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

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

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^[76], 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 1.2, che delinea la struttura tipica di un metodo attuale (tralasciando i metodi end-to-end come AlphaFold2).

1.1.2 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^[78]:

- Machine Learning:
 - ANN (Artificial neural network)
 - SVM (Support vector machines), es. SVM-SEQ
 - k-Nearest Neighbors
 - linear regression
 - HMM (Hidden Markov Models)
 - Support vector regression
- EC (Evolutionary computing), es. MECoMaP

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.

- 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. Oltre ai vettori di input codificanti di coppie di aminoacidi, è possibile includere neuroni con informazioni aggiuntive, ad es. lunghezza della sequenza, valori di idrofobicità dell’ambiente o informazioni evolutive, nonostante la codifica dei dati di input necessariamente restrin ga le possibili informazioni codificabili. Le reti neurali presentano comunque delle limitazioni di cui tenere conto, ad esempio l’uso di parametri appropriati e l’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^[39]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^[77]:

- generazione di MSA (es. DeepMSA)
- predizione di contatti (contact map, es. TripletRes)
- predizione di distogrammi (es. RaptorX)

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

- 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^[77].

1.1.3 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^[78]:

- *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 distanze 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^[78]:

- 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 amminoacidi 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^[79]. 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.
- 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

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

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.

1.1.4 Database e rappresentazione delle informazioni

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. ??). 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).

Database di proteine

—PDB— The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules that serves a global community of researchers, educators, and students. The data contained in the archive include atomic coordinates, crystallographic structure factors and NMR experimental data. Aside from coordinates, each deposition also includes the names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations.

La predizione della struttura è importante per un semplice motivo: i biochimici conoscono oggi la sequenza amminoacidica per più di 225 milioni di proteine^[80] (UniProt, con circa 4.5-5 milioni aggiunte ogni mese) ma sono state determinate solamente circa 160.000 strutture tridimensionali di proteine^[80] (PDB⁶, con poco più di 10.000 strutture aggiunte ogni anno).

1.1.5 Rappresentazione grafica

QuteMol -> cignoni!

1.1.6 CASP

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

⁶Al 3 Febbraio 2022 sono presenti 162.913 strutture di proteine nella versione dell'RCSB. 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^[81]. Nel wwPDB (database globale) vi sono in totale 197.961 strutture di macromolecole^[82].

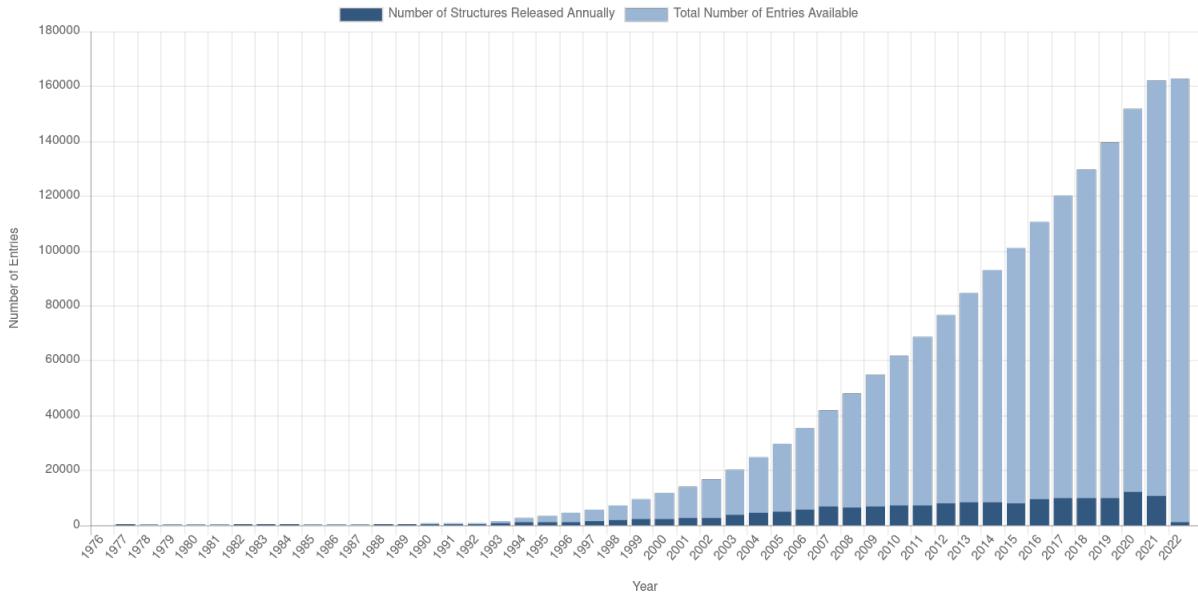


Figura 1.3: Crescita complessiva del numero di strutture di proteine pubblicate nel PDB. Fonte[81]

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)^[73].

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.

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.

Nel 2020 gli organizzatori del CASP14 hanno riconosciuto AlphaFold come soluzione del *protein-structure-prediction problem* ???.

1.2 Annotazioni 1D sulla struttura

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

Le proprietà delle sequenze possono essere raggruppate in due tipi:

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

Proprietà del 1° tipo sono collegate a: formazioni 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 *profili*, 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). È questa l'utilità dell'utilizzo di profili. Il profilo della proteina target è poi confrontato con i profili corrispondenti di tutte le proteine con riepigamenti conosciuti (*profile-profile alignment*) e le proteine sufficientemente simili alla proteina target possono essere usate come template globali. Un esempio di procedura profile-matching è HHpred, basata su confronti a coppie di profili HMM tra target e template.

1.2.1 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 in una grande varietà di analisi e pipeline nell'analisi del proteoma e del genoma e sono un 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 mostra il grado di similarità tra le sequenze.

Di solito, si dovrebbe solo tentare di allineare sequenze filogeneticamente correlate e, quindi, omologhe. In questo caso, l'allineamento ideale avrà residui omologhi allineati nelle colonne.

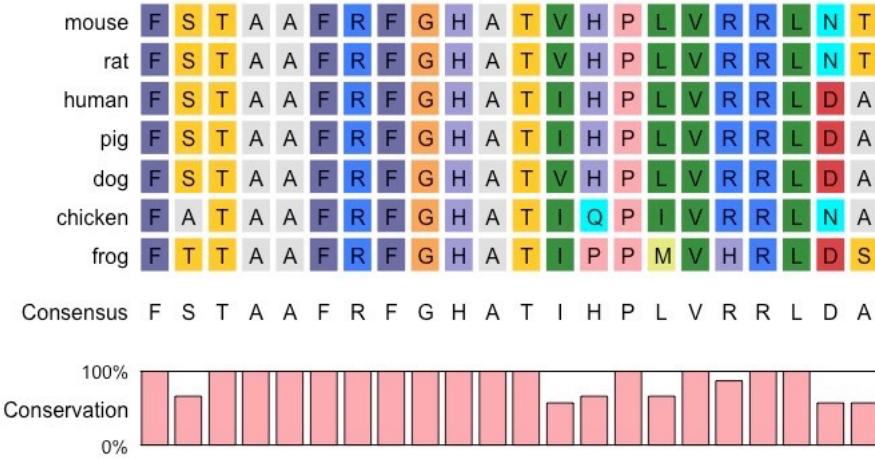


Figura 1.4: Multiple sequence alignment schematica di una sequenza proteica di varie specie. Fonte [83]

Le sequenze di proteine in genere risultano avere un grado di somiglianza maggiore rispetto alle sequenze nucleotidiche, questo 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 caratterizza un "profilo" della sequenza.

1.2.2 Predizione della struttura secondaria

1.3 Annotazioni 2D sulla struttura

Le annotazioni riguardano predizioni dei contatti fra residui sulla base di informazioni co-evolutive (*correlated mutation*), generando o una *contact map*, o un *distogramma* o una *distance map* attraverso una *deep residual neural network* (*ResNet*). Il fine è ricavare restrizioni spaziali come contatti inter-residuo a lunga distanza, distanze specifiche, legami idrogeno e angoli di torsione.

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

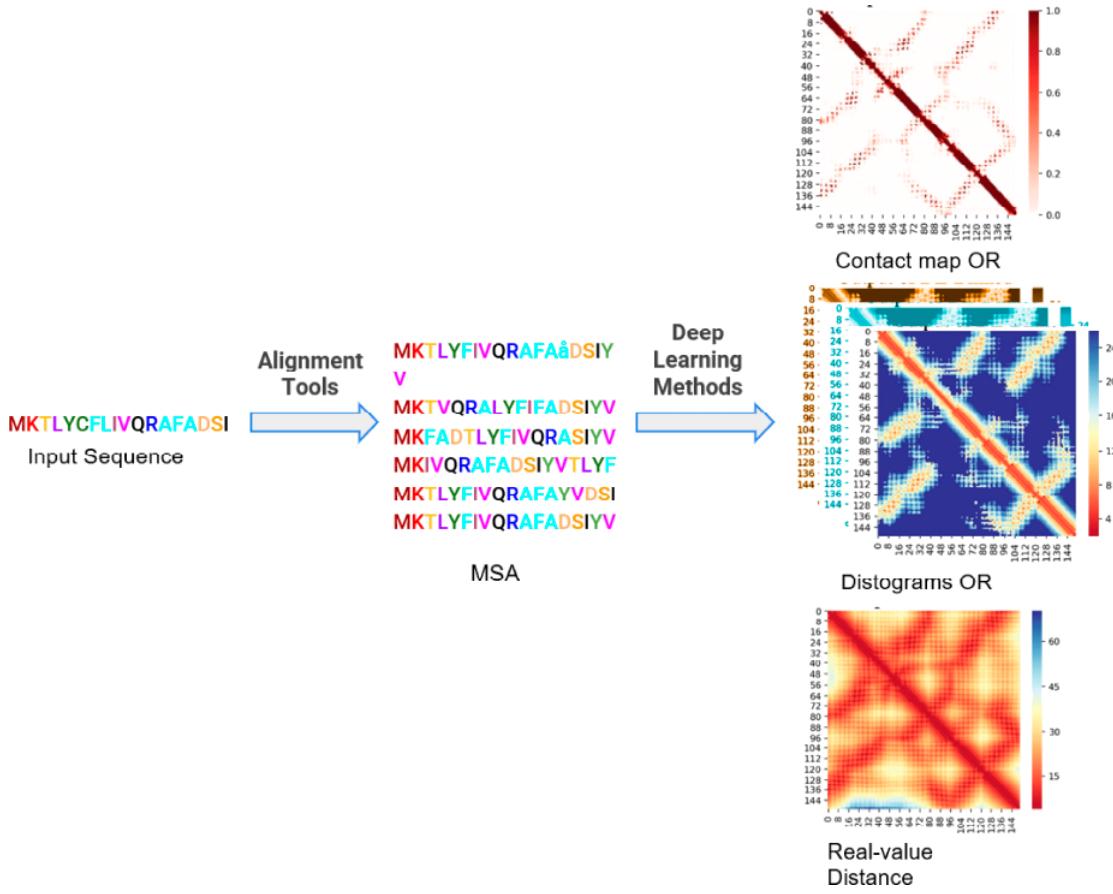


Figura 1.5: Schema generale di un workflow FM. Il DL è un passo fondamentale nel predizione di vincoli spaziali. Fonte[77]

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.

1.3.1 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.

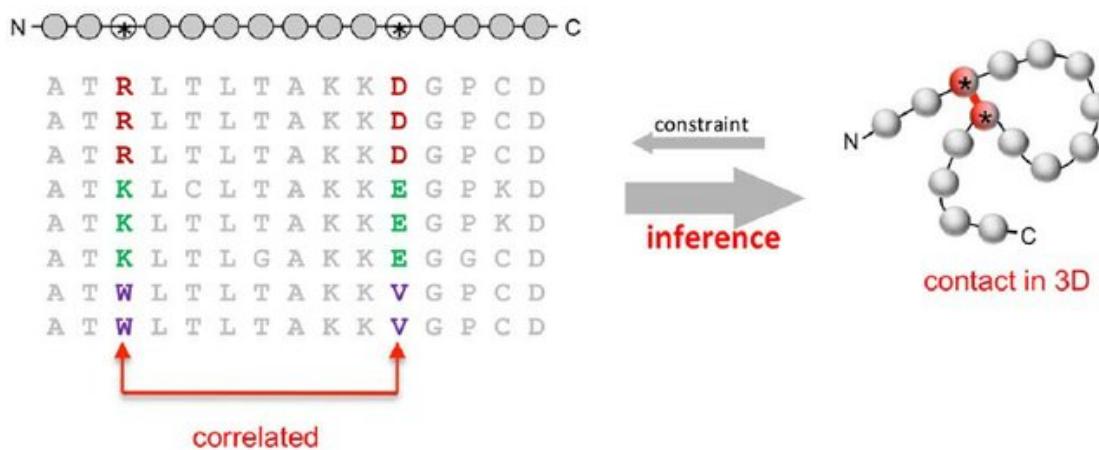


Figura 1.6: In alto, sequenza della proteina da predire, in cui ogni amminoacido è un cerchio. Sotto, un MSA con le sequenze di una famiglia correlata, tutte con lo stesso ripiegamento. A destra l'inferenza del contatto delle posizioni coevolute. Fonte[84]

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 EVfol, Rosetta-GREMLIN, FILM3, MetaPSICOV, ecc.

The hypothesis behind the approach was that if mutations that occur at two positions in an MSA are correlated, these positions are more likely to form a contact in 3D space. This is because there is evolutionary pressure to conserve the structures of proteins and a mutation at one position may be rescued by a corresponding mutation at a nearby residue. The accuracy of co-evolution-based contact map prediction remained low for many years due to the inability to distinguish between direct and indirect interactions, where indirect interactions occur when residues appear to co-evolve but do not actually

form contacts. For example, if Residues A and B are both in contact with Residue C, A and B often appear as if they co-evolve even when there is no physical contact between them. There is evidence showing that such co-evolution may have a functional cause [22] rather than a structural one, which resulted in the failure of structure-based contact derivation.^[75]

Progress in contact prediction remained stagnant for some time. However, a leap in contact prediction accuracy took place when algorithms started utilizing global prediction approaches. Early methods mainly predicted contacts between residue pairs one-at-a-time using techniques such as mutual information, thus ignoring the interactions with other residue pairs and the global context in which the interactions took place; this is largely why it was difficult for these local methods to distinguish between direct and indirect interactions. The introduction of global statistical models determined through the use of direct coupling analysis (DCA) was more successfully able to distinguish between these direct and indirect interactions [23,24].

1.3.2 Distogrammi

1.4 Predizione della struttura 3D

1.4.1 *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*».

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

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

Un approccio che utilizza 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

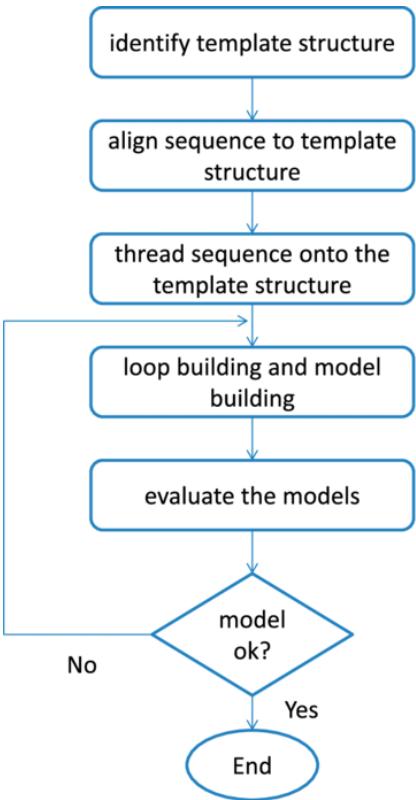


Figura 1.7: Diagramma di flusso della modellazione per omologia. Fonte: [85]

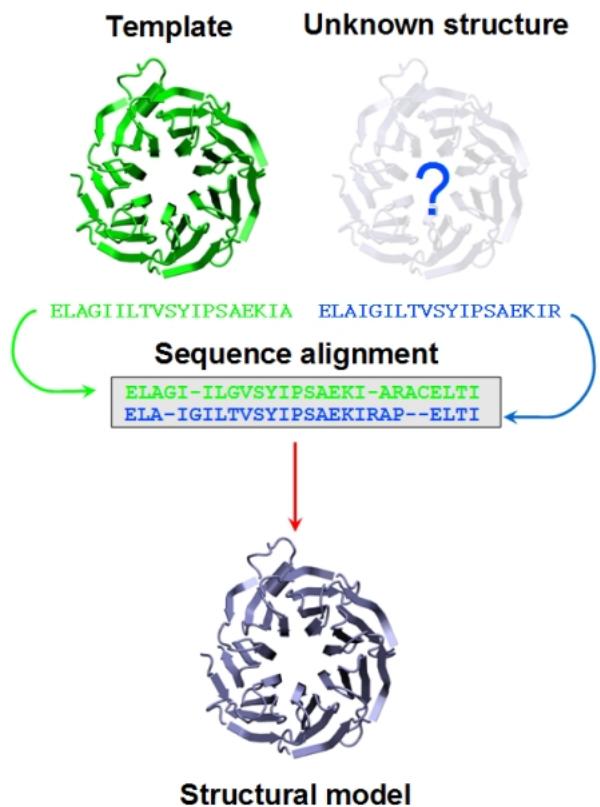


Figura 1.8: Schema esemplificativo di una modellazione per omologia. Fonte [86]

Nel 1° step si cerca una struttura modello, almeno una, tra le strutture conosciute avente un’alta somiglianza di sequenza. È più semplice se la struttura di un’una proteina omologa molto simile è stata già risolta. Ci sono però altri 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. Le principali tecniche si dividono nelle seguenti categorie:

- pattern recognition e ricerche euristiche, ad esempio BLAST (Basic Local Alignment Search Tool) e FastA
- profile and iterative alignment methods, come PSI-BLAST e HMM (Hidden Markov Models), quest’ultimo molto usato
- structure based threading, come THREADER e FUGUE

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 *procarioti*. La tecnica più usata oggi è l'MSA, altre tecniche usate sono: programmazione dinamica, MSA, threading, HMM ecc. Ci sono vari tool per creare MSA, ad esempio: ClustalΩ, MAFFT, MUSCLE, T-Coffee. Possono venire utilizzati più MSA insieme per sopperire a problemi di disallineamento di piccole regioni. Il MSA viene illustrato dopo la sezione 1.2.1.

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 (soddisfazione 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. 1.4.5), regioni fondamentali per i siti di legame, 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 in generale 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)
- stereochemica (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^[87].

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.

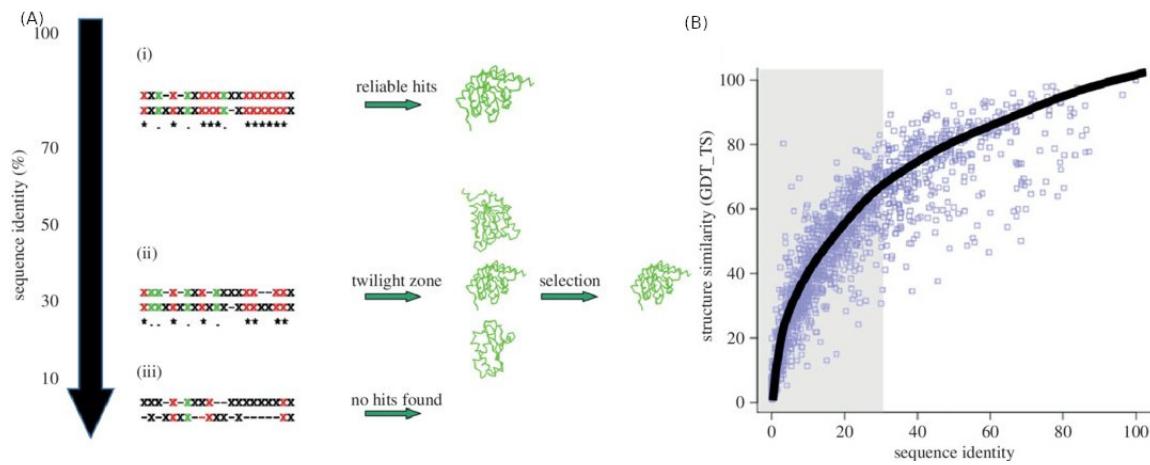


Figura 1.9: Risultati dei metodi per omologia alla variazione dell'identità nella sequenza. (A) Dimostrazione schematica dell'uso di confronti di sequenze per identificare omologie strutturali. 'X' indica un qualsiasi amminoacido. (i) identità > 70%: semplici allineamenti di sequenza sono sufficienti 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 score) al variare dell'identità della sequenza. Anche al 30% il livello di somiglianza è significativo. Fonte[88]

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.

Conclusioni sulla *modellazione per omologia*

Nonostante tutte le osservazioni fatte, *la modellazione per omologia 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.

La sfida principale che questi metodi hanno affrontato, e che AlphaFold ha "vinto", era raggiungere almeno una RMSD di 3Å. Le sfide ancora da superare riguardano l'accuratezza su grandi proteine, su proteine con un contenuto significativo di strutture β e la modellazione di proteine multi-dominio e di membrana.

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.

1.4.2 *fold recognition via threading*

Il *threading* consiste nell'allineare la sequenza della proteina target con la struttura di proteine template, che possono essere trovate sulla base di proprietà condivise (vedi sez 1.2 sulla predizione di caratteristiche strutturali a partire dalla sequenza).

an alignment between the sequence and the template structure (threading).

Identifying templates, i.e., proteins of known structure whose fold should be similar to that of the target protein. This can be done by threading, i.e., by aligning the sequence of the target protein with the structure of each fold in the PDB on the basis of shared sequence properties, and then choosing the best match. The sequence properties can be grouped into two types:

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. Se queste proprietà o ricorrenze statistiche potessero essere identificare, sarebbe 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*. 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 simile o motivi strutturali simili.

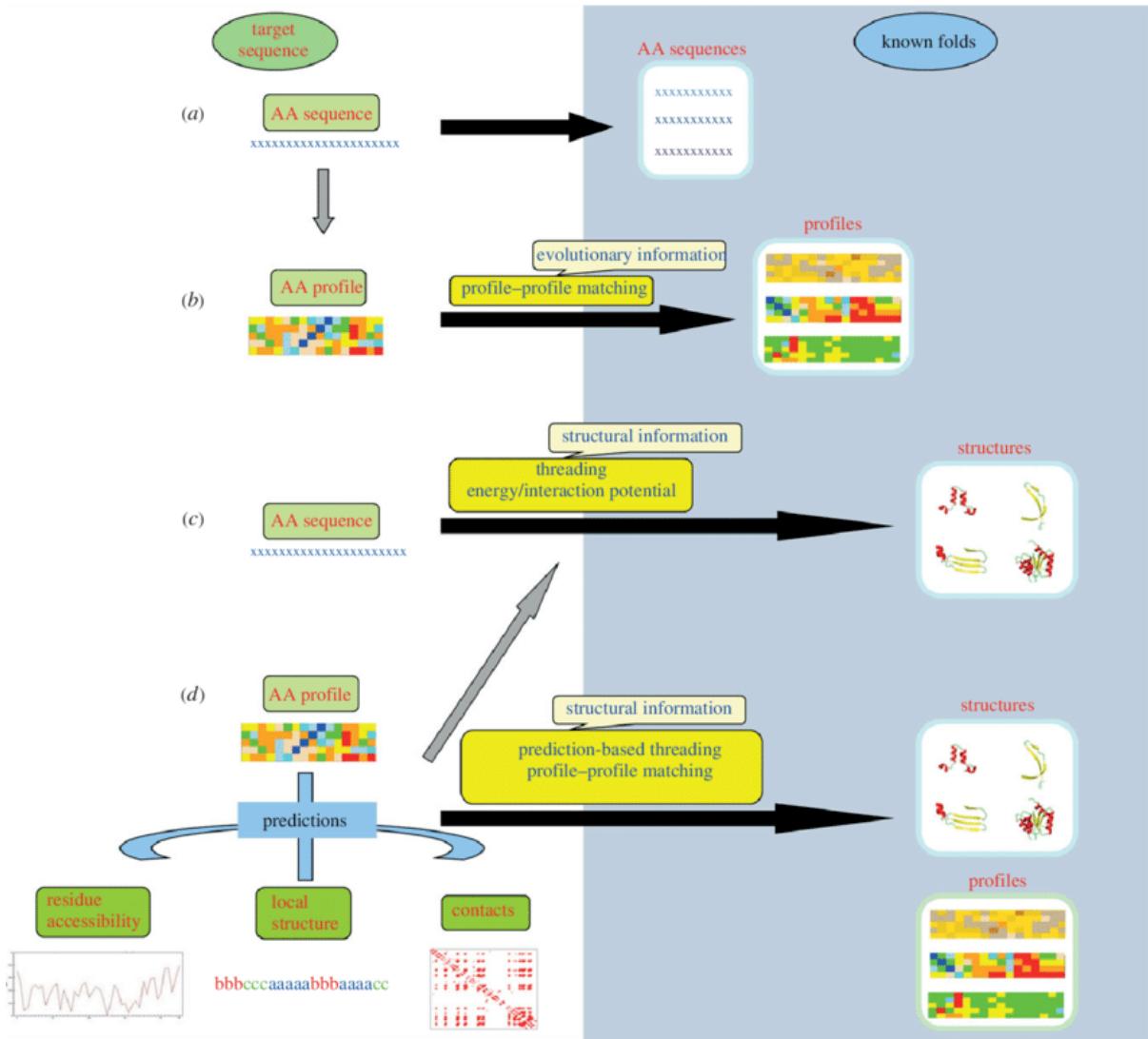


Figura 1.10: . Fonte[88]

1.4.3 *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. ??) 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 *campo di forza* 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. 1.11)

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. ??).

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 atomici⁷ 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 polari⁸, in una 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.

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 veloci⁹

⁷Gli atomi possiedono, in base alla loro eventuale carica, campi elettrici che influenzano gli altri atomi.

⁸Un esempio di interazione non polare è l'effetto idrofobico. Vengono escluse poiché coinvolgono principalmente cambiamenti di entropia nel solvente.

⁹Questo approccio è stato introdotto da Warshel, Levitt e Karplus.

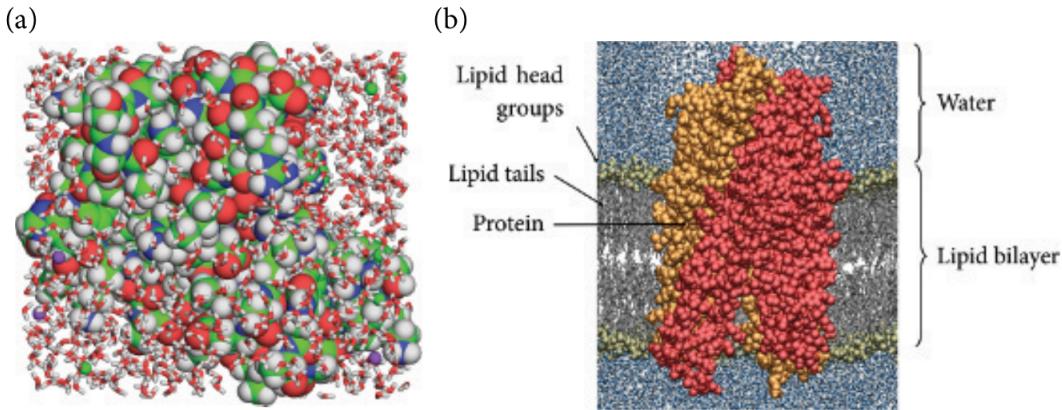


Figura 1.11: Descrizioni esplicite nei calcoli di MM. (A) una piccola proteina immersa in un solvente composto da molecole d'acqua e ioni (Na^+ , Cl^-). La proteina è rappresentata come sfere di atomi, l'acqua come bastoncini e gli ioni come piccole sfere magenta e gialle. (B) una proteina trasportatrice 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 [5]

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)^[89].

Il metodo più semplice è chiamato *energy minimization*:

1. si parte da una configurazione arbitraria
2. si calcola l'energia potenziale. 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 a rimanere bloccate in un minimo locale di energia non riuscendo spesso a raggiungere il minimo globale a causa di *barriere energetiche* da scavalcare per raggiungere una configurazione con energia minore (vedi fig. 1.12 e ??).

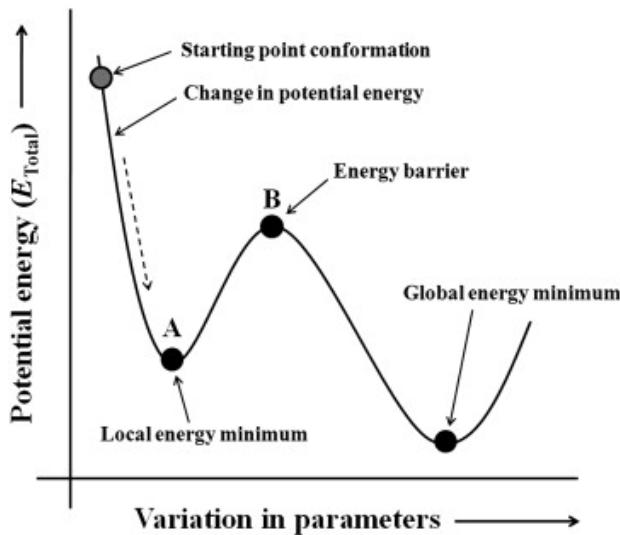


Figura 1.12: Differenti fasi energetiche di una molecola durante la sua minimizzazione energetica. Fonte[89]

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 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 simulazioni 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)^[73].

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*¹⁰, conosciuto anche come approccio *continuum-solvent*, vedi fig. 1.13.

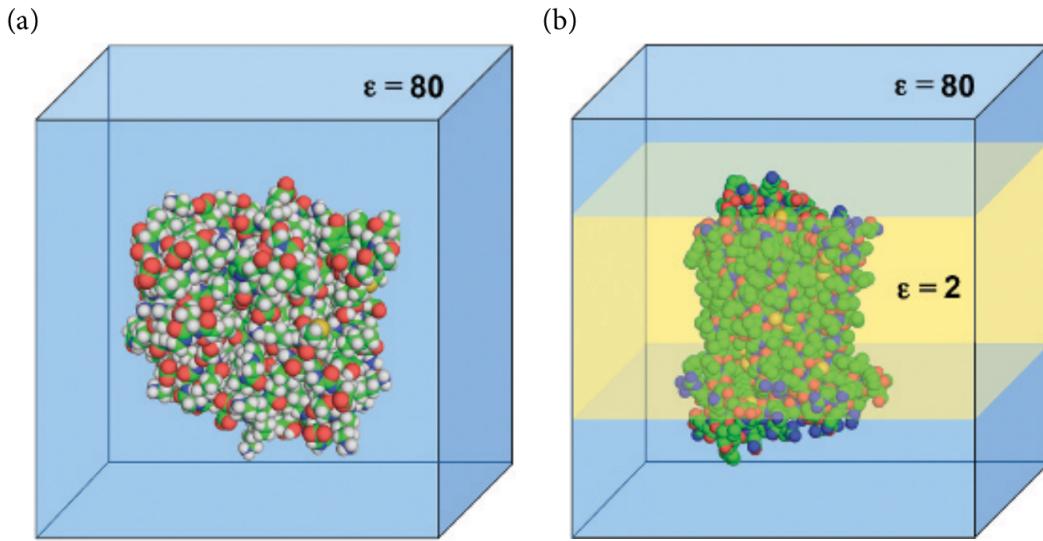


Figura 1.13: Descrizione con approccio *mean-field* di un sistema, il solvente è descritto implicitamente mentre la proteina esplicitamente. ϵ indica la dielettricità. (A) proteina in un solvente acquoso altamente dielettrico. (B) proteina di membrana in un ambiente eterogeneo. Il solvente acquoso è altamente dielettrico mentre la lastra semi-trasparente gialla, che rappresenta la regione biologica di doppio strato lipidico, è poco dielettrica. Fonte[5]

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

¹⁰Proprietà di un mezzo non conduttore di essere sede di un campo elettrostatico.

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. 1.6).

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, 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.

1.4.4 *fragment-based*

Un possibile metodo di azione è la *frammentazione delle proteine*, nel quale i calcoli sono realizzati su piccoli frammenti della proteina e poi integrati. Un esempio è il software QUARK. L'approccio *fragment-based* è usato nei metodi che combinano predizioni *template-based* con assemblaggi *ab initio*.

1.4.5 *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.

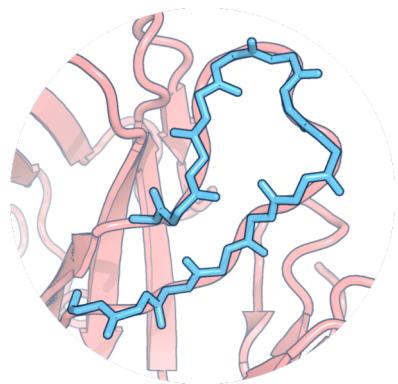


Figura 1.14: Disegno di un loop in celeste. Fonte [90]

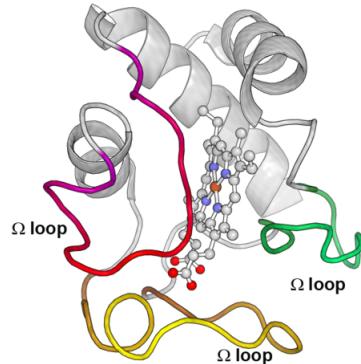


Figura 1.15: Omega loop. Sono spesso coinvolti nel riconoscimento molecolare e in funzioni regolatrici. Fonte: [91]

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^[92].

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^[93]. 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 conformazionale
- 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 1.16, con i seguenti step:

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

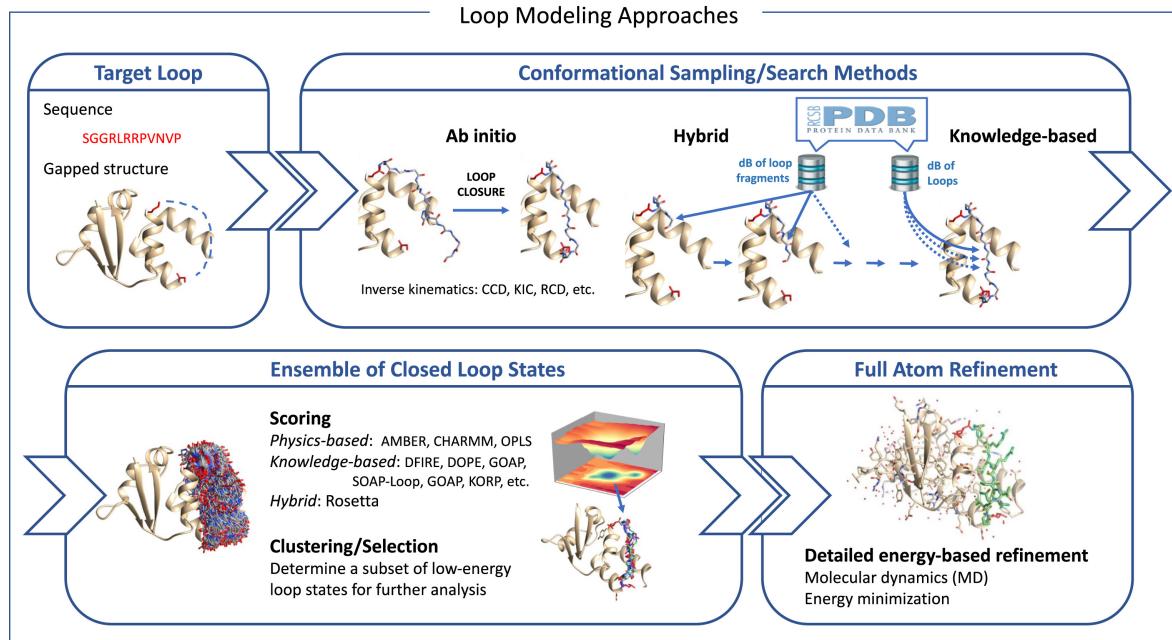


Figura 1.16: Approcci al loop modeling. Workflow schematico di un protocollo prototipo per la modellazione dei loop. Fonte[94]

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^[92]. Metodi allo stato dell'arte sono in grado di predire conformazioni stabili di loop relativamente corti (fino a 12 residui)^[94]. Seppur 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 SAbPred¹¹: 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^[94].

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

¹¹Collezione 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)^[92].

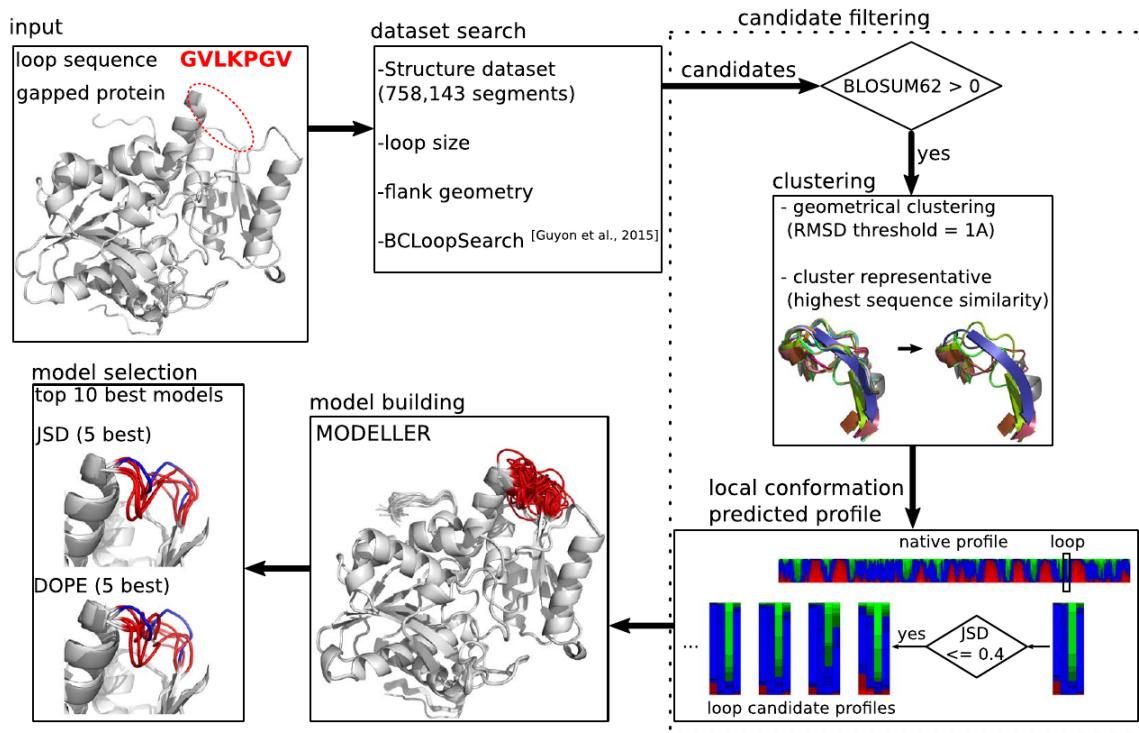


Figura 1.17: DaReUS-Loop workflow. Da notare che dal 2019^[93] nel processo di costruzione del modello non è più usato MODELLER (non free) ma GROMACS. Fonte^[92]

I principali step del metodo sono mostrati in figura 1.17 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.

1.5 Case Study: *TASSER*

-tasser papers -kessel -abbass -pearce 2021

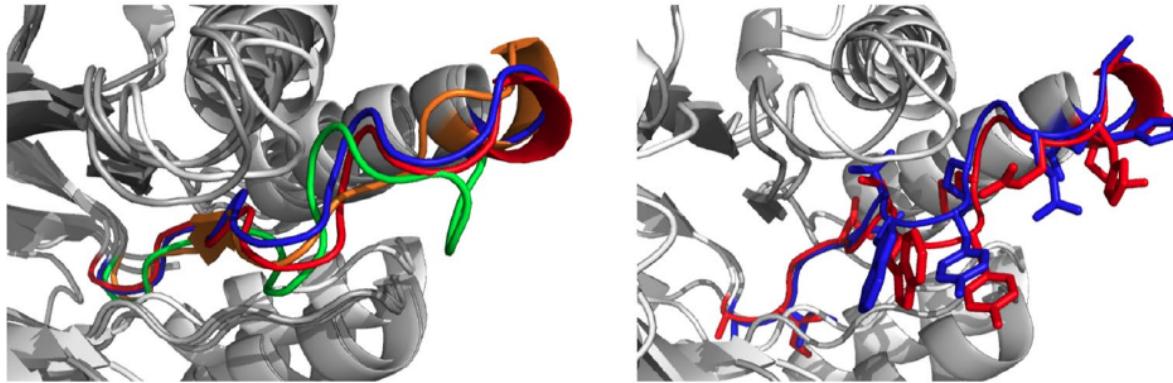


Figura 1.18: Esempi di predizione di un loop lungo (15 residui) della proteina target T0807 del CASP11 a confronto. Blu=DaReUS-Loop, verde=Rosetta NGK, arancione=GalaxyLoop-PS2, rosso=struttura cristallizzata. La RMSD di ogni loop predetto rispetto al loop nativo è riportata di seguito. DaReUS-Loop: 1.3Å, NGK: 3Å, PS2: 2.9Å. Nella colonna a destra sono riportate le catene laterali della struttura nativa e di quella predetta da DaReUS-Loop. Fonte[92]

1.6 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. 1.19.

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 interesse si possono usare tecniche di *cromatografia* o *elettroforesi*. Una forma efficiente di chromatografia 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 velo-

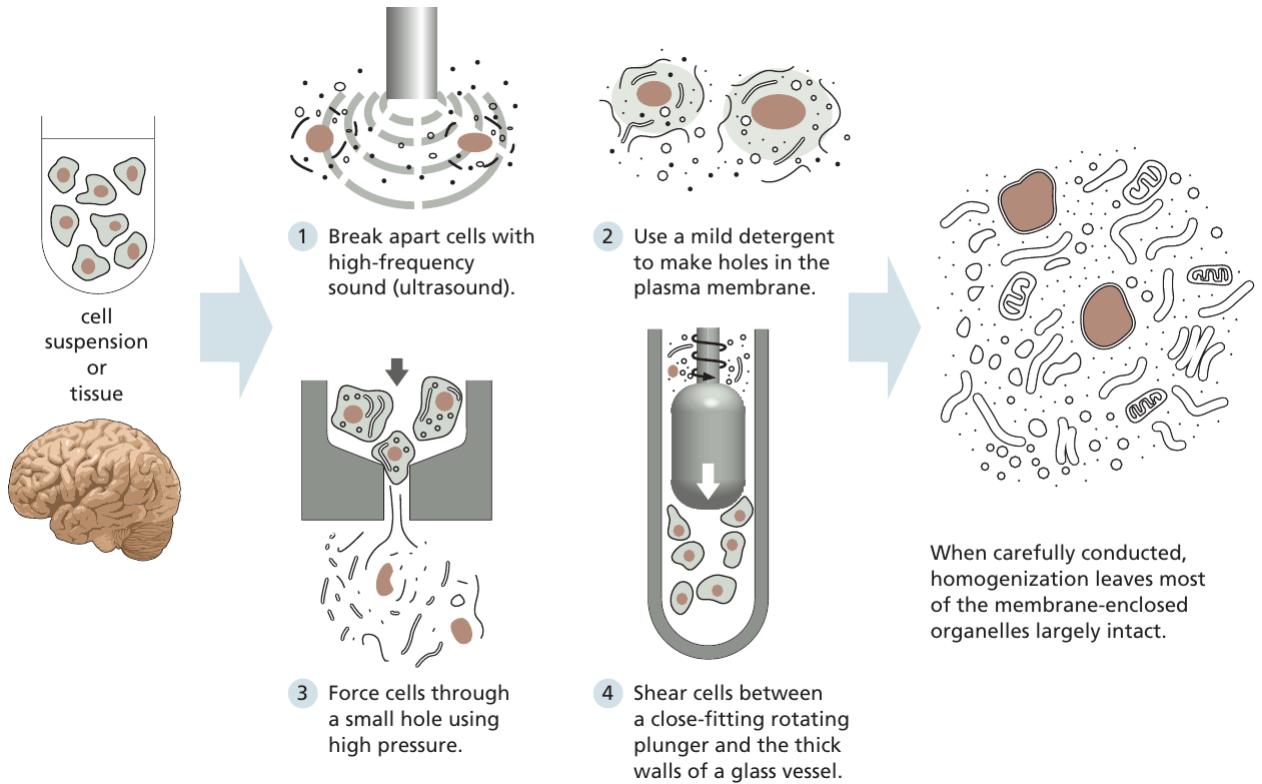


Figura 1.19: Omogenizzazione attraverso 4 diversi meccanismi, ognuno dei quali finalizzato a rompere le membrane plasmatiche. Fonte[6]

cità 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.

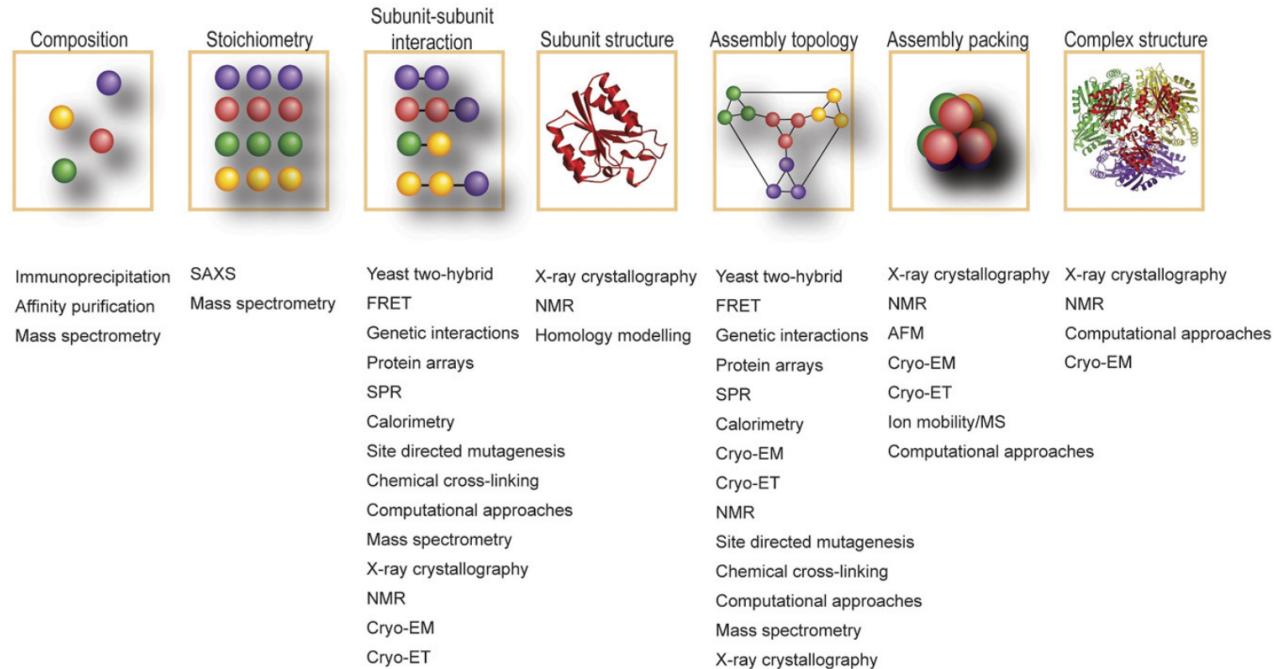


Figura 1.20: Differenti livelli di informazione ottenuta lungo il percorso della determinazione della struttura di macromolecole. AFM: atomic force microscopy; Cryo-ET: cryo-electron tomography; EM: electron microscopy; FRET: fluorescence resonance energy-transfer; NMR: nuclear magnetic resonance; SAXS: small-angle X-ray scattering; SPR: surface plasmon resonance. Fonte[95]

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*¹². 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.

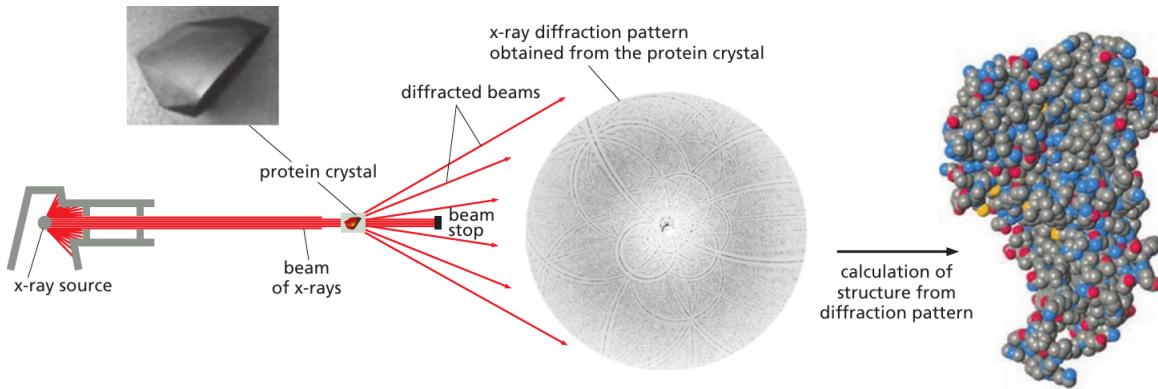


Figura 1.21: Processo di determinazione della struttura dell'enzima RuBisCO tramite cristallografia a raggi-X. Fonte[6]

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, $1 - 2\text{\AA}$). I raggi X sono dispersi o diffatti dagli atomi proteici nel cristallo. Il modello 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

¹²La traduzione italiana sarebbe: cristallografia a raggi-X, spettroscopia a risonanza magnetica nucleare e microscopia elettronica.

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^[29].

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.

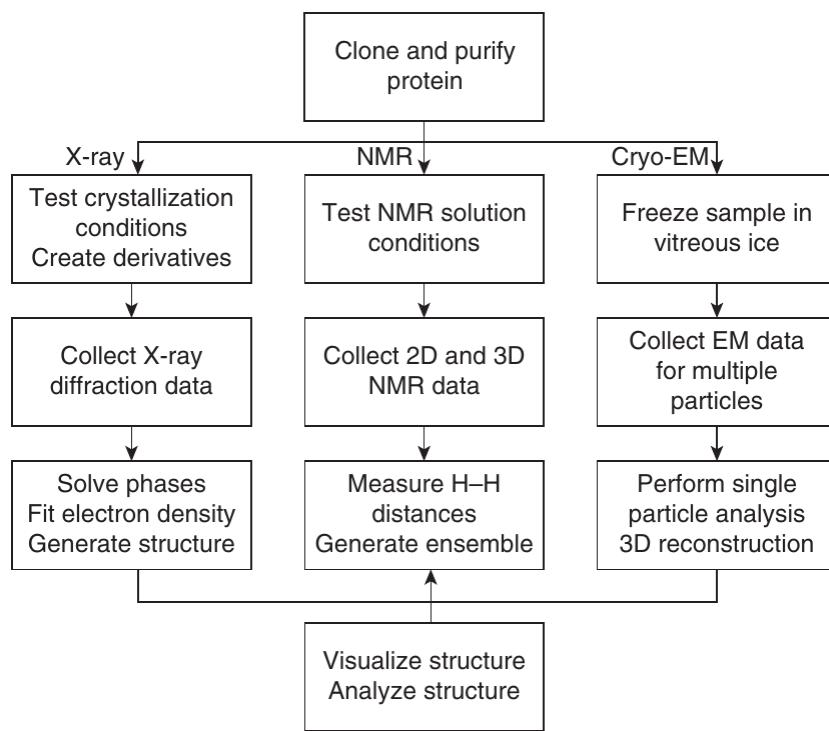


Figura 1.22: Diagramma di flusso dei principali step usati la preparazione e soluzione sperimentale della struttura 3D delle proteine. Fonte[29]

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 radiazioni; 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 (so-

luzione, 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.

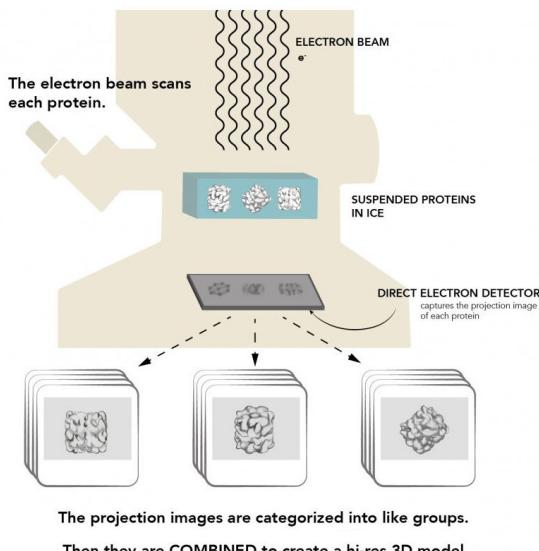


Figura 1.23: Funzionamento schematico di un microscopio crioelettronico. Fonte [96]

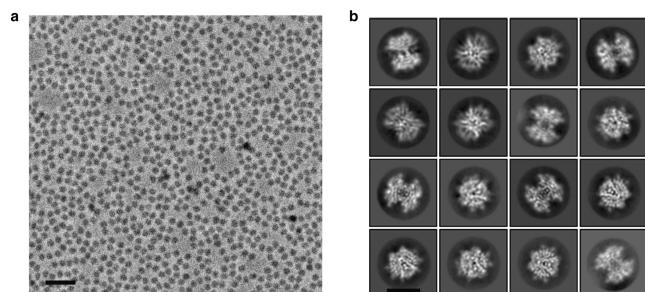


Figura 1.24: (A) Micrografia rappresentativa del campione di streptavidina ottenuta tramite cryo-EM. La barra di scala rappresenta 20nm (B) Classi medie 2D rappresentative delle immagini della proteina. La barra di scala rappresenta 5nm. Fonte [97]

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 microografie per catturare più 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^[97]). 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.

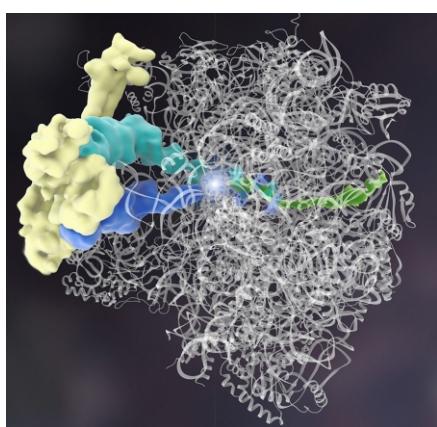


Figura 1.25: Complesso di controllo qualità dei ribosomi, basato su dati di cryo-EM. Fonte: [98]

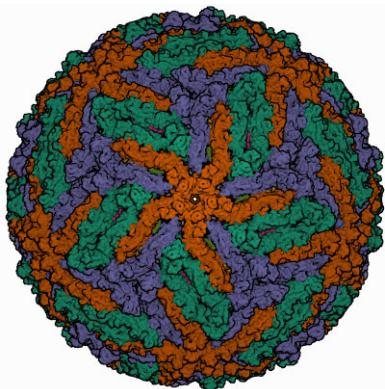


Figura 1.26: La struttura del virus Zika ottenuta tramite cryo-EM. La macromolecola ha un peso di 190kDa ed è composta da 11.000 atomi. Risoluzione di 3.8Å. Fonte [99]

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 metodologia¹³. Si sta considerando l'adozione della microscopia elettronica come di una rivoluzione nel campo della biologia strutturale^{[100],[101]}. 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

¹³A Richard Henderson, Jacques Dubochet e Joachim Frank, "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution".

è cambiata (vedi le figure 1.27 e 1.28). 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.

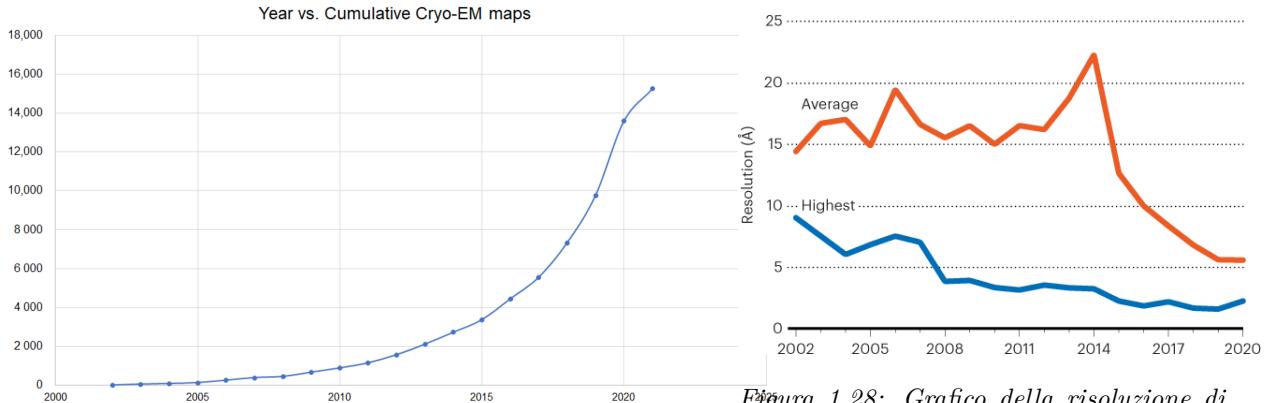


Figura 1.27: Crescita delle mappe elettroniche rilasciate nell'E-struttura risolte tramite cryo-EM. Fonte MDB. Fonte [77]

Figura 1.28: Grafico della risoluzione di

[100]

1.7 Storia della comprensione delle proteine

- alberts p.160 - psp-wiki - levitt 2001, birth of structural biology

L'approccio *ab initio* è emerso negli anni '60 a partire dal campo della chimica computazionale. Nel 2013 il premio Nobel per la chimica è stato assegnato proprio a quegli scienziati che hanno contribuito sin da quegli anni al campo della biofisica molecolare computazionale (Warshel, Levitt, Karplus).

Le predizioni di strutture basate sui metodi *ab initio* sono emerse nella metà degli anni '80, prima per piccoli peptidi e poi per polipeptidi. Il primo programma per calcolare l'energia potenziale nelle proteine è stato sviluppato nel 1969 da Lifson e Levitt^[102].

La prima simulazione di MD su una proteina è stata realizzata nel 1977 da McCammon, Gelin e Karplus^[103], 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*^[104]) sulla stessa proteina che era però rappresentata in modo più semplicistico: ogni amminoacido era rappresentato da due sfere.

anni '90, database, omologia, progetto genoma

Around the beginning of the 1990s, a new field in biology called ‘bioinformatics’ emerged, in which scientists sought to predict the characteristics of new proteins on the basis of properties of their sequences

CASP e AlphaFold

A test of homology modeling efficiency carried out in 2007 has shown that in single-domain proteins comprising 90 residues or fewer, the structures predicted by this method differed from their corresponding native structures by 2 to 6^[50].

Interestingly, in the first rounds of CASP, secondary structure prediction was a separate category. This category was cancelled after the organizers noticed that the winners in this category used a somewhat circular approach. They predicted the 3D structure and used their model structure to decipher the secondary structure elements.

Bibliografia

Libri

- [5] A. Kessel e N. Ben-Tal, *Introduction to proteins: Structure, function and motion*, 2^a ed. Chapman e Hall/CRC, 2018.
- [6] B. Alberts, D. Bray, K. Hopkin et al., *Essential cell biology*, 5^a ed. W. W. Norton e Company, 2019.
- [7] N. A. Campbell, J. B. Reece, L. A. Urry, R. Brizzi, T. Niccolò e A. Bartalesi, *Biologia E Genetica*. Pearson, 2012.
- [15] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, R. B. Orr e N. A. Campbell, *Campbell Biology*. Pearson, 2021.
- [29] A. D. Baxevanis, G. D. Bader e D. S. Wishart, *Bioinformatics*. John Wiley & Sons, 2020.
- [37] T. Mitchell, *Machine learning*. McGraw hill New York, 1997.
- [46] S. Pal, *Fundamentals of Molecular Structural Biology*. Academic Press, 2019.
- [51] L. A. Moran, H. R. Horton, K. G. Scrimgeour, M. D. Perry e D. Rawn, *Principles of biochemistry*. Pearson London, 2012.
- [55] D. L. Nelson, A. L. Lehninger e M. M. Cox, *Lehninger principles of biochemistry*. Macmillan, 2017.
- [72] L. Fleck, *Genesi e Sviluppo di un Fatto Scientifico: Per Una Teoria dello stile e del collettivo di pensiero*. Il mulino, 1983.

Articoli

- [18] 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.

- [27] 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.
- [28] 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.
- [30] 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.
- [31] M. Mitchell, “Biological Computation,” *PDXScholar*, 2010. indirizzo: https://pdxscholar.library.pdx.edu/compsci_fac/2.
- [39] 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.
- [44] 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.
- [45] C. B. Anfinsen, “The formation and stabilization of protein structure.,” *Biochemical Journal*, vol. 128, n. 4, p. 737, 1972.
- [47] 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.
- [50] 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.
- [52] 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>.
- [57] N. A. Ranson, H. E. White e H. R. Saibil, “Chaperonins,” *Biochemical Journal*, vol. 333, n. 2, pp. 233–242, 1998.
- [58] 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.
- [60] 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.

- [61] 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.
- [62] D. Fraser-Pitt e D. O’Neil, “Cystic fibrosis—a multiorgan protein misfolding disease,” *Future science OA*, vol. 1, n. 2, 2015.
- [65] 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.
- [66] 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.
- [67] E. Fischer, “Einfluss der Configuration auf die Wirkung der Enzyme,” *Berichte der deutschen chemischen Gesellschaft*, vol. 27, n. 3, pp. 2985–2993, 1894.
- [68] 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.
- [69] 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.
- [70] 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.
- [71] C. Bracken, M. M. Young e K. Dunker, “Disorder and flexibility in protein structure and function,” pp. 64–66, 2000.
- [73] 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.
- [74] 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>.
- [75] 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.
- [76] 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.

- [77] 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.
- [78] 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.
- [79] 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.
- [84] 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.
- [85] G. Sliwski, S. Kothiwale, J. Meiler e E. W. Lowe, “Computational methods in drug discovery,” *Pharmacological reviews*, vol. 66, n. 1, pp. 334–395, 2014.
- [87] 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.
- [88] 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.
- [91] 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.
- [92] 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.
- [93] 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.
- [94] A. Barozet, P. Chacón e J. Cortés, “Current approaches to flexible loop modeling,” *Current Research in Structural Biology*, vol. 3, pp. 187–191, 2021.
- [95] 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.

- [97] 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.
- [100] E. Callaway, “Revolutionary cryo-EM is taking over structural biology.”, *Nature*, vol. 578, n. 7794, pp. 201–202, 2020.
- [101] 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.
- [102] 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.
- [103] J. A. McCammon, B. R. Gelin e M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, n. 5612, pp. 585–590, 1977.
- [104] M. Levitt e A. Warshel, “Computer simulation of protein folding,” *Nature*, vol. 253, n. 5494, pp. 694–698, 1975.

Risorse Online

- [3] “enzima nell’Enciclopedia Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/enciclopedia/enzima> (visitato il 21/01/2022).
- [4] “proteina in Vocabolario - Treccani.” (13 gen. 2022), indirizzo: <https://www.treccani.it/vocabolario/proteina> (visitato il 22/01/2022).
- [8] “Chemical element - Wikipedia.” (1 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Chemical_element (visitato il 31/01/2022).
- [9] “eukaryote. Definition, Structure, Facts.” (19 set. 2019), indirizzo: <https://www.britannica.com/science/eukaryote> (visitato il 22/01/2022).
- [10] “Neurone - Wikipedia.” (27 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/Neurone> (visitato il 23/01/2022).
- [11] “Saccharomyces cerevisiae - Wikipedia.” (25 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Saccharomyces_cerevisiae (visitato il 22/01/2022).
- [12] “Dogma centrale della biologia molecolare - Wikipedia.” (16 set. 2021), indirizzo: https://it.wikipedia.org/wiki/Dogma_centrale_della_biotologia_molecolare (visitato il 22/01/2022).

- [13] “DNA Structure. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/26-structure-of-dna-and-rna/dna-structure.html> (visitato il 22/01/2022).
- [16] “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).
- [17] “Nasuia deltocephalinicola - Wikipedia.” (25 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Nasuia_deltoccephalinicola (visitato il 31/01/2022).
- [19] “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).
- [20] “Paris japonica - Wikipedia.” (31 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Paris_japonica (visitato il 31/01/2022).
- [21] “Transfer RNA - Wikipedia.” (23 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Transfer_RNA (visitato il 23/01/2022).
- [22] “Protein - Wikipedia.” (21 dic. 2021), indirizzo: <https://en.wikipedia.org/wiki/Protein> (visitato il 23/01/2022).
- [23] “Peptide bond - Wikipedia.” (4 nov. 2021), indirizzo: https://en.wikipedia.org/wiki/Peptide_bond (visitato il 23/01/2022).
- [24] “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).
- [25] “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).
- [26] 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).
- [32] “Apprendimento automatico - Wikipedia.” (1 dic. 2021), indirizzo: <https://it.wikipedia.org/wiki/ApprendimentoAutomatico> (visitato il 23/01/2022).
- [33] “What is soft computing - Javatpoint.” (3 lug. 2021), indirizzo: <https://www.javatpoint.com/what-is-soft-computing> (visitato il 24/01/2022).
- [36] “Machine Learning - IBM.” (29 ago. 2020), indirizzo: <https://www.ibm.com/it-it/analytics/machine-learning> (visitato il 23/01/2022).

- [38] “What are Neural Networks?” (1 Giu. 2021), indirizzo: <https://www.ibm.com/cloud/learn/neural-networks> (visitato il 24/01/2022).
- [40] “What are Recurrent Neural Networks?” (14 Set. 2020), indirizzo: <https://www.ibm.com/cloud/learn/recurrent-neural-networks> (visitato il 07/02/2022).
- [41] “Residual neural network - Wikipedia.” (30 gen. 2022), indirizzo: https://en.wikipedia.org/wiki/Residual_neural_network (visitato il 07/02/2022).
- [43] “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).
- [48] “Christian B. Anfinsen - Digital Collections - National Library of Medicine.” (1 gen. 2022), indirizzo: <http://resource.nlm.nih.gov/101408166> (visitato il 25/01/2022).
- [49] “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).
- [53] “Protein structure prediction - Wikipedia.” (30 dic. 2021), indirizzo: https://en.wikipedia.org/wiki/Protein_structure_prediction (visitato il 27/01/2022).
- [54] 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).
- [56] “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).
- [59] “Chaperonina - Wikipedia.” (21 ott. 2021), indirizzo: <https://it.wikipedia.org/wiki/Chaperonina> (visitato il 26/01/2022).
- [63] “Inibitori della proteasi - Wikipedia.” (17 ago. 2021), indirizzo: https://it.wikipedia.org/wiki/Inibitori_della_proteasi (visitato il 04/02/2022).
- [64] “Lauren Porter and Fold-Switching Proteins.” (14 feb. 2020), indirizzo: <https://www.youtube.com/watch?v=IeX5ebadgiA> (visitato il 28/01/2022).
- [80] “Latest Release Information.” (27 gen. 2022), indirizzo: <https://www.ddbj.nig.ac.jp/latest-releases-e.html> (visitato il 27/01/2022).
- [81] “PDB Statistics: Protein-only Structures Released Per Year.” (1 feb. 2022), indirizzo: <https://www.rcsb.org/stats/growth/growth-released-structures> (visitato il 03/02/2022).

- [82] “wwPDB: Deposition Statistics.” (1 feb. 2022), indirizzo: <https://www.wwpdb.org/stats/deposition> (visitato il 03/02/2022).
- [83] “Sequence Alignment. BioNinja.” (15 apr. 2021), indirizzo: <https://ib.bioninja.com.au/options/untitled/b5-bioinformatics/sequence-alignment.html> (visitato il 02/02/2022).
- [86] “Homology modeling.” (31 mag. 2014), indirizzo: <https://www.unil.ch/pmf/home/menuinst/technologies/homology-modeling.html> (visitato il 03/02/2022).
- [90] “SAbPred: FREAD.” (2 apr. 2022), indirizzo: <http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/fread> (visitato il 04/02/2022).
- [96] “What is Cryo-EM?” (21 Giu. 2018), indirizzo: <https://cryoem.slac.stanford.edu/what-is-cryo-em> (visitato il 06/02/2022).
- [98] “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).
- [99] “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).

Altre fonti

- [1] R. M. Pirsig, *Lo Zen e l'arte della manutenzione della motocicletta*, 1974.
- [2] F. Capra, *Il Tao della fisica*, 1975.
- [14] 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).
- [34] R. Kurzweil, R. Richter, R. Kurzweil e M. L. Schneider, *The age of intelligent machines*, 1990.
- [35] H. H. the XIV Dalai Lama, *The heart of the Buddha's path*, 2011.
- [42] A. Ruffini, *Fisiogenia, la biodinamica dello sviluppo ed i fondamentali problemi morfologici dell'embriologia generale*, 1925.
- [89] 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.