UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN INGEGNERIA INFORMATICA

TESI DI LAUREA

SVILUPPO DI UN METODO PER LA COSTRUZIONE AUTOMATICA DI STRUTTURE PROTEICHE DA ALLINEAMENTI DI SEQUENZA

RELATORE:
PROF. CARLO FERRARI

CORRELATORE:
PROF. SILVIO C. E. TOSATTO

LAUREANDO:
FRANCESCO LOVO

A.A. 2012-2013
Padova, 12 Marzo 2013

Ai miei genitori

Sommario

Scopo di questa tesi è quello di presentare un'articolata libreria C++ che implementa un metodo per la costruzione automatica di strutture proteiche a partire dalle sequenze amminoacidiche. Tale lavoro, svolto presso il laboratorio di BioComputing al Dipartimento di Biologia dell'Università degli Studi di Padova, si inserisce in un'area di grande interesse per la bioinformatica: il problema del protein folding (ripiegamento proteico).

Conoscere la struttura tridimensionale delle proteine è essenziale nella comprensione delle loro funzioni e dei fenomeni biologici a queste correlati, ma ottenere tali strutture per via sperimentale è un processo lungo e dispendioso. Per la presente tesi si è assemblato un metodo computazionale di previsione della struttura a partire da componenti in parte sviluppati all'interno dello stesso laboratorio, con il non secondario obiettivo di automatizzare l'intero processo di predizione della struttura. La strategia generale seguita nella progettazione si basa sull'approccio knowledge-based: il problema viene risolto andando ad estrarre informazioni da banche dati di sequenze con strutture risolte, alla ricerca di una proteina che faccia da guida nella costruzione della struttura della sequenza in analisi. Si parla quindi di modellazione per omologia: è noto infatti che l'evoluzione tende a conservare la struttura piuttosto che la sequenza, pertanto maggiore è l'identità tra due sequenze e tanto maggiore è la probabilità che queste condividano la stessa struttura. Trovato un buon allineamento tra due sequenze amminoacidiche è quindi possibile inferire con ragionevole sicurezza anche un buon allineamento strutturale.

Automatizzare le scelte da compiere o costruire algoritmi che simulino l'intervento umano in tale processo di predizione è molto difficile. L'approccio che si è deciso di seguire è quello di realizzare più soluzioni possibili a partire da un'unica struttura templato attraverso la produzione di più allineamenti profilo-profilo alternativi, stato dell'arte in materia di allineamenti proteici. Dopo un'eventuale ulteriore raffinamento con la modellazione delle catene laterali e dei loops (le parti meno conservate nell'allineamento) viene scelto il modello migliore attraverso un metodo di valutazione energetica, la cui implementazione ha richiesto un profondo lavoro di re-ingegnerizzazione di una soluzione pre-esistente.

Al termine della fase di progettazione e realizzazione del codice sono stati svolti approfonditi test di verifica rispetto ad una precedente versione, simulando la partecipazione all'ultima edizione del CASP: una delle più importanti competizioni internazionali nel campo della predizione e modellazione di strutture proteiche. I risultati hanno evidenziato la bontà delle scelte fatte ai fini dell'automatizzazione (procedura di ricerca del templato, parametri del processo di allineamento in particolare) e il sensibile miglioramento delle strutture prodotte.

Indice

1 Introduzione

Le proteine sono strutture molto complesse ed eterogenee che rappresentano il risultato della traduzione dei geni, e sono i costituenti fondamentali di tutte le cellule animali e vegetali. Ad esse sono associate diverse funzioni sia di tipo prettamente strutturale, sia di tipo enzimatico o regolativo, con compiti di trasporto dentro e fuori dalla cellula o di difesa da sostanze esterne ed estranee all'organismo.

L'assunzione della funzione fisiologica di una proteina, sia essa un enzima, un trasportatore, un recettore o una proteina strutturale, è resa possibile dalla sua struttura tridimensionale. La conoscenza della struttura tridimensionale è pertanto essenziale per le importanti ricadute in vari campi quali per esempio:

Biologia Molecolare: l'analisi delle strutture, con l'identificazione della posizione e dimensione di un sito attivo è il primo passo nella comprensione dei meccanismi alla base del funzionamento degli organismi.

Biologia Evoluzionistica: dall'analisi delle strutture proteiche di due specie diverse si possono ottenere informazioni sul loro grado di parentela dato che trovare proteine analoghe, cioè con la stessa funzione ma conformazione diversa, è indice di grande distanza evolutiva.

Biotecnologie: da un punto di vista teorico è possibile progettare la forma di una proteina al fine di ottenere un comportamento desiderato.

Medicina: mutazioni nei geni possono determinare la formazione di proteine con strutture anomale che svolgono funzioni scorrette spesso causa di patologie anche gravi. Lo studio della loro forma può quindi portare a comprendere come tali anomalie siano alla base dei meccanismi patologici.

Farmacologia: le proteine sono spesso gli obiettivi dei farmaci. Conoscere la loro forma permette quindi di progettare farmaci più specifici e mirati, minimizzando gli effetti collaterali.

Ad oggi esiste un enorme divario, destinato ad aumentare, fra il numero di strutture note determinate sperimentalmente e di sequenze conosciute. Tale gap si è allargato molto negli ultimi anni a causa dei notevoli progressi nel sequenziamento di interi genomi, ai quali non è corrisposto un altrettanto significativo miglioramento dei metodi di determinazione delle strutture tridimensionali.

Gli attuali metodi sperimentali, quali la cristallografia a raggi X e la spettroscopia a risonanza magnetica nucleare (NMR), permettono di avere informazioni piuttosto accurate sulla struttura tridimensionale delle proteine, ma sfortunatamente richiedono spesso tempi lunghi, hanno elevata complessità e soffrono ancora di limiti applicativi tali da impedire l'analisi completa di tutte le sequenze.

Nasce quindi la necessità di metodi computazionali veloci ed efficaci in grado di ricostruire la struttura 3D a partire da informazioni che possono limitarsi alla sola sequenza amminoacidica, o spaziare in un insieme più o meno ampio di vincoli strutturali determinati empiricamente.

Esistono differenti approcci e tecniche risolutive, il lavoro presentato in questa tesi segue quello che ad oggi è il più promettente: il *comparative modeling.* Sfruttando la similarità fra sequenze, il comparative modeling utilizza la struttura di proteine note (d'ora in poi chiamate templato) come sistema di riferimento nella costruzione del modello 3D della sequenza (d'ora in poi indicata come target) di cui si cerca la struttura.

Il progetto realizzato è stato chiamato HOMER (acronimo di HOmology ModellER), ed è stato implementato sottoforma di server web i cui servizi saranno a breve accessibili al pubblico[1]. Sviluppato sulla base di una precedente versione, ne rappresenta però una completa rivisitazione sia a livello concettuale che implementativo: l'intero processo di costruzione del modello è stato infatti ridefinito e le soluzioni adottate nei singoli passi aggiornate o completamente riviste.

Il normale utilizzo prevede la costruzione di un modello strutturale a partire dalla sola sequenza sequenza di amminoacidi, ma è anche possibile fornire un proprio allineamento (in formato FASTA) e una singola struttura templato (in formato PDB). Quest'ultima può essere caricata direttamente, o selezionata dal database PDB locale. A richiesta HOMER può modellare le regioni di loop e le catene laterali, e in genere segue una serie di protocolli che si sono affermati nelle edizioni bi-annuali del CASP: una competizione mondiale sulla predizione di struttura. Il risultato del programma, che include il modello prodotto e una valutazione residuo per residuo del profilo energetico della struttura, è accessibile attraverso pagine web dinamiche.

Un elemento di novità rispetto ad altri servizi analoghi è la possibilità di poter includere nella struttura finale informazioni sui cofattori presenti nella struttura usata come templato, in particolare ioni-metallici che spesso sono di estremo interesse per l'influenza che hanno proprio sulla funzione svolta dalla proteina a cui sono legati. Infine in HOMER è prevista anche la possibilità di modellare particolari proteine chiamate omodimeri: strutture formate dall'unione di sub-unità di identica natura chimica. Per questa loro caratteristica un buon allineamento anche solo per una piccola porzione della sequenza target può essere replicato in altre sezioni della stessa, ottenendo quindi un modello molto più esaustivo. In genere i programmi di comparative modeling trascurano situazioni di questo tipo.

Un ulteriore obiettivo di questa tesi, e per nulla secondario, è quello di automatizzare l'intero processo di costruzione dei modelli. Il campo della predizione di strutture proteiche ha un numero molto elevato di potenziali utenti e riscuote un sempre crescente interesse. Efficaci e consolidati strumenti esistono già da tempo, ma il loro corretto e proficuo utilizzo richiede esperienza e conoscenze precise del problema. Un server automatico di predizione in cui l'intervento umano possa essere totalmente escluso dovrebbe poter avvicinare tutta la comunità scientifica e consentire l'uso di tali strumenti bioinformatici anche ai non esperti.

Nei capitoli successivi, dopo una breve introduzione sui principali aspetti strutturali delle proteine e delle metodologie più usate nella determinazione delle strutture terziarie, verranno illustrate nel dettaglio le scelte fatte ai fini dell'automatizzazione, le soluzioni tecniche adottate nella fase di ricerca del templato, nell'allineamento, nella costruzione dei modelli "grezzi", nel loro raffinamento attraverso la modellazione delle catene laterali e dei loops, nella selezione della struttura più plausibile sulla base di valutazioni energetiche, ed infine i risultati ottenuti.

1.1 Contenuto dei capitoli:

Capitolo2: Proteine. Il capitolo introduce i principali aspetti strutturali delle proteine: tipo e struttura degli amminoacidi proteici, livelli strutturali, relazione tra forma e funzione, folding proteico, metodi di analisi sperimentali e computazionali. La trattazione ha carattere meramente

[1]Il servizio è accessibile al seguente indirizzo: http://biocomp.bio.unipd.it/homer/auto.html

generale e si propone di fornire una panoramica di concetti di base che verranno ripresi nei capitoli successivi.

Capitolo3: Allineamento di sequenze. Il capitolo illustra brevemente il problema dell'allineamento di coppie di sequenze proteiche e la sua importanza nei processi comparativi di modellazione. Si spiega il significato delle matrici di sostituzione e dei profili, e vengono descritti i principali algoritmi (esatti ed euristici).

Capitolo4: PDB. Il capitolo elenca le maggiori banche dati biologiche riportando statistiche ed altre informazioni di interesse. Si sofferma sulla Protein Data Bank e sul formato PDB.

Capitolo5: Stato dell'arte: CASP. Il capitolo descrive uno dei più importanti esperimenti nella predizione di strutture proteiche, nato con lo scopo di valutare oggettivamente lo stato dell'arte e i miglioramenti conseguiti in questo campo, e la sua importanza come mezzo di direzionamento della ricerca biologica.

Capitolo7: La libreria Biopool. Il capitolo riporta le classi e i programmi necessari a rappresentare la struttura di una proteina all'interno del progetto.

Capitolo8: Ricerca del templato. Il capitolo illustra la procedura di ricerca del templato e descrive la strategia PDB_BLAST.

Capitolo9: GenSubAli: allineamento. Il capitolo mostra la struttura della libreria responsabile dell'allineamento e le varie opzioni che mette a disposizione. Descrive poi le scelte prese ai fini dell'automatizzazione della procedura.

Capitolo10: Homer: costruzione del modello grezzo. Il capitolo descrive come avviene il processo di creazione del modello della proteina target usando come riferimento le coordinate atomiche degli amminoacidi della struttura templato, dato un allineamento delle loro sequenze.

Capitolo11: Modellazione delle catene laterali. Il capitolo illustra il passo che si occupa di aggiungere le catene laterali a tutti quegli amminoacidi per i quali, a causa di un allineamento solo parziale, non è stato possibile recuperare le posizioni dal templato.

Capitolo12: QMEAN: valutazione energetica. Il capitolo spiega come avviene la valutazione su base energetica dei vari modelli prodotti, e quindi la scelta del modello più valido tra i vari candidati.

Capitolo13: Modellazione dei loop. Il capitolo illustra il processo di modellazione dei cosiddetti "loop": le parti più variabili di una proteina che non vengono allineate con il templato.

Capitolo14: Risultati. Il capitolo riporta e discute i risultati ottenuti modellando i target del CASP10, e confronta i valori ottenuti con la versione precedente del programma.

Capitolo15: Conclusioni. Il capitolo riassume il lavoro svolto e propone alcuni possibili sviluppi futuri.

Appendice A: Materiale CASP10. L'appendice contiene i dati relativi ai target dell'esperimento CASP10 utilizzati per validare il programma.

2 Proteine

Le proteine sono polimeri lineari composti da amminoacidi uniti mediante un legame peptidico. La lunghezza di tali sequenze varia da circa 40 a più di 1000 amminoacidi.

Il lavoro fondamentale per attivare la funzione fisiologica di una proteina è svolto dal processo di ripiegamento (folding), durante il quale la proteina in soluzione si assesta in una struttura tridimensionale. Il loro ruolo nella regolazione della maggior parte delle attività cellulari è intrinsecamente legato alle numerose conformazioni con cui queste macromolecole possono presentarsi.

Nonostante le peculiarità delle strutture 3D di singole proteine, l'osservazione globale della loro composizione permette di astrarre forme di carattere generale utili per evidenziare classi di proprietà simili.

Le principali funzioni svolte riguardano:

- trasporto (es. mioglobina, emoglobina);
- catalisi[2]: enzimi (es. proteasi, cellulasi);
- metabolismo: ormoni (es. insulina, glucagone);
- sostegno: proteine strutturali (es. collageno, cheratina, fibrotina);
- movimento: proteine contrattili (es. miosina, actina);
- "difesa": anticorpi (es. immunoglobuline);
- "attacco": tossine batteriche, veleni dei serpenti.
- riserva di amminoacidi (es. ovoalbumina, caseina).

La maggior parte delle proteine interagisce con piccole molecole, chiamate ligandi, o altre proteine per assolvere ai propri compiti.

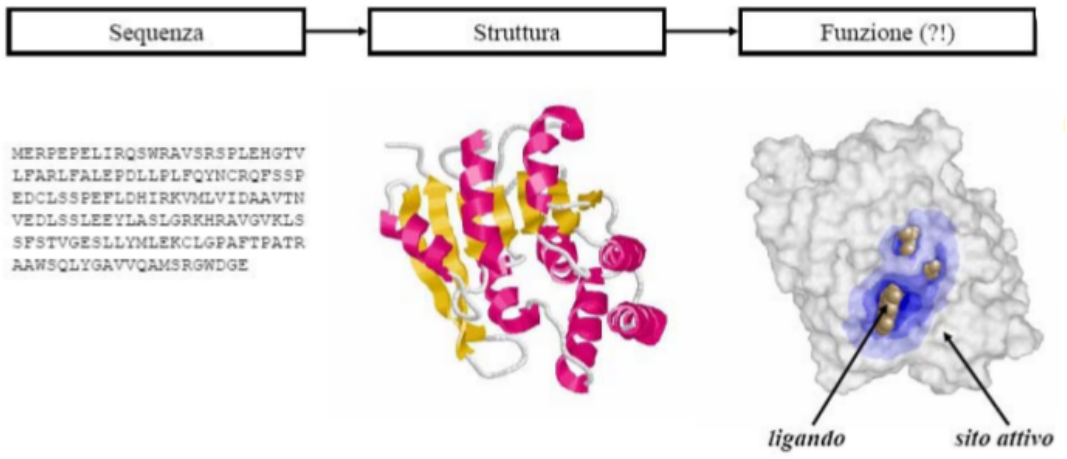

Figura 1: Una proteina

[2]Con il termine catalisi si intende il controllo della velocità delle reazioni biochimiche

2.1 Amminoacidi

Nelle proteine in natura troviamo 20 diversi tipi di amminoacidi.

Gli amminoacidi, che si chiamano anche residui, hanno una struttura di fondo comune costituita da un atomo di carbonio centrale, denominato carbonio alfa (Cα) , un gruppo amminico (NH2) ed uno carbossilico (COOH) legati al medesimo atomo di carbonio Cα. Tale struttura standard si lega alla struttura di altri amminoacidi, andando così a formare una catena che prende il nome di backbone.

Amminoacido	Sigle
Alanina	Ala (A)
Arginina	Arg (R)
Asparagina	Asn (N)
Acido Aspartico	Asp (D)
Cisteina	Cys (C)
Glutammina	Gln (Q)
Acido Glutammico	Glu (E)
Glicina	Gly (G)
Istidina	His (H)
Isoleucina	Ile (I)
Leucina	Leu (L)
Lisina	Lys (K)
Metionina	Met (M)
Fenilalanina	Phe (F)
Prolina	Pro (P)
Serina	Ser (S)
Treonina	Thr (T)
Triptofano	Trp (W)
Tirosina	Tyr (Y)
Valina	Val (V)

Figure 2: Amminoacidi: nomi e sigle

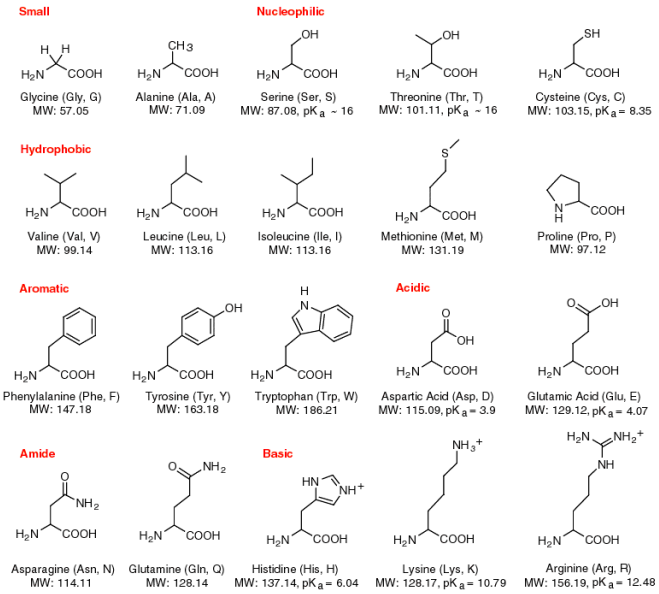

Figure 3: Amminoacidi: struttura e proprietà chimiche

Oltre a tali gruppi “fissi”, ogni amminoacido presenta uno specifico gruppo laterale o catena laterale che lo caratterizza funzionalmente. L’unica eccezione è rappresentata dalla Glicina, che è priva di catena laterale.

Dato che questa è la parte che varia per ogni amminoacido, essa viene indicata con una R, che sta ad indicare il “resto” della molecola. Per lo stesso motivo il termine residuo viene usato come sinonimo di amminoacido.

In funzione delle proprietà chimiche del gruppo R, un amminoacido viene classificato come acido, basico, idrofilo (o polare) o idrofobo (o apolare). L’ingombro dei vari gruppi R che sporgono dalla catena polipeptidica e le loro caratteristiche chimiche concorrono a modellare la conformazione della proteina nello spazio (la struttura terziaria), conformazione dalla quale dipende in modo essenziale l’attività biologica della proteina stessa.

Gli amminoacidi si possono unire tra loro attraverso legami peptidici (un tipo di legame covalente), quindi polimerizzare e formare proteine. Questo tipo di legame si forma tra il gruppo -NH2 e il gruppo -COOH di due amminoacidi adiacenti, con rimozione di una molecola d’acqua. Poiché

il legame si forma tra un atomo di azoto ed uno di carbonio, che delimitano le estremità della molecola, essi prendono il nome di n-terminale (atomo di azoto) e c-terminale (atomo di carbonio).

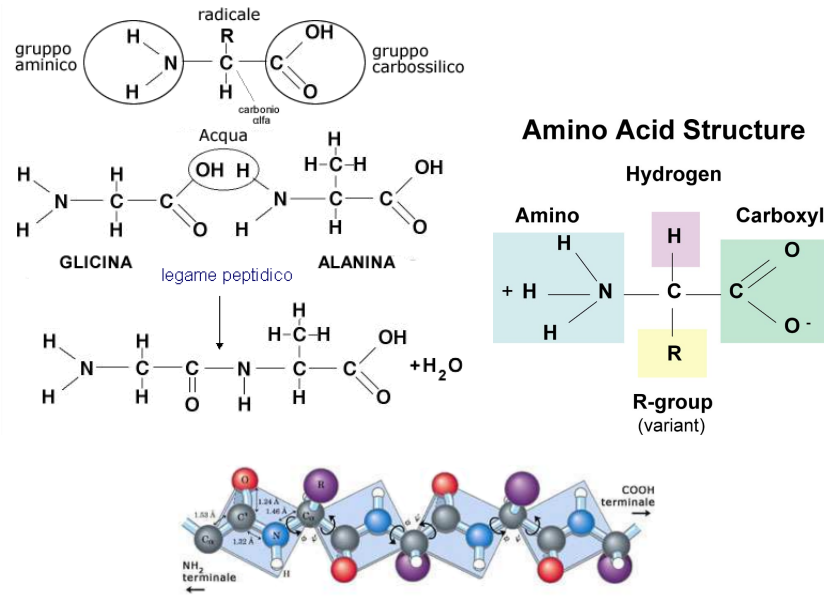

Figura 4: Il legame peptidico

Alcuni legami possono ruotare su se stessi permettendo una certa flessibilità alla struttura di una proteina. I legami che permettono questa libertà vengono chiamati angoli torsionali e sono essenziali per permettere alla proteina di assumere la propria forma definitiva.

Il legame peptidico è un legame estremamente rigido, d'altro canto i due legami ad esso contigui (il C-COOH e il NH-C) possono compiere rotazioni, formando due angoli, rispettivamente (Psi) e (Phi). Questi due angoli teoricamente possono variare da -180° a +180° anche se in pratica la libertà effettiva che una struttura può avere è limitata: ad esempio dalla possibilità di collisioni che si possono creare tra le catene laterali degli amminoacidi con altri elementi della molecola stessa.

Una proteina, essendo una macromolecola formata da decine di migliaia di atomi, potrebbe potenzialmente assumere un numero incredibilmente grande di possibili ripiegamenti. Tuttavia considerazioni fisiche limitano di molto le possibili conformazioni finali di una proteina.

Gli atomi non si possono mai sovrapporre e si comportano a grandi linee come sfere con un raggio definito detto raggio di Van Der Waals, ciò limita non poco il numero di angoli ammessi in una catena polipeptidica.

Riportando in un grafico φ in funzione del corrispondente ψ, si ottiene il cosiddetto grafico o mappa di Ramachandran in cui si evidenziano tre regioni di coppie consentite in cui gli amminoacidi tendono a ripiegarsi in base all'ingombro delle catene laterali. Per ogni singolo amminoacido (tranne che per la Glicina che non ha ingombro sferico a causa della mancanza di una catena laterale) è possibile plottare una mappa di Ramachandran che ne descriva le possibili conformazioni.

Come si vede in figura , è possibile notare delle zone con gradazione differente:

1. ZONA SCURA : nessuna collisione (regioni favorite)

2. ZONA INTERMEDIA : basso rischio di collisione (regioni ammesse)

3. ZONA CHIARA : collisioni tra atomi certe (valori di phi e psi non ammessi)

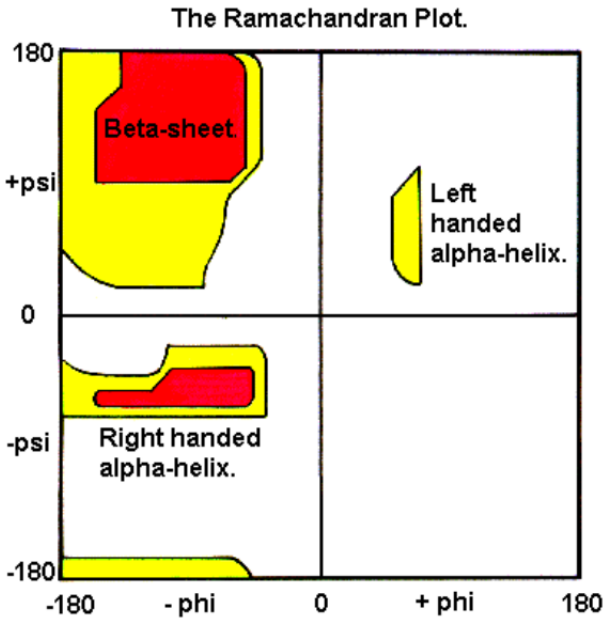

Figura 5: Mappa di Ramachandran

Gli angoli torsionali non sono tuttavia presenti solo nella backbone, ma anche nella catena laterale, aumentando la flessibilità nel ripiegamento della proteina.

Gli angoli torsionali necessari per il corretto posizionamento delle catene laterali (rotameri , χ o chi) sono presenti in numero variabile fino a cinque, e in letteratura sono disponibili tabelle di valori ideali che suggeriscono gli angoli di torsione più probabili a fronte di specifiche conformazioni del backbone proteico.

2.2 Livelli di struttura

Nella struttura proteica si riconoscono più livelli di organizzazione, all'interno delle quali nel processo di folding vengono sviluppate parti delle strutture che andranno poi a determinare proprietà e forma ultima della proteina.

Grazie a criteri essenzialmente gerarchici si possono distinguere in quattro differenti tipologie:

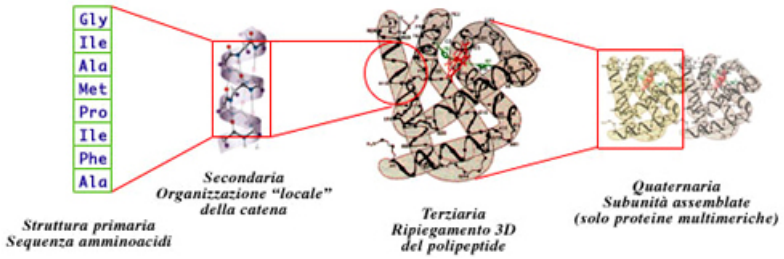

Figura 6: Livelli strutturali

1. STRUTTURA PRIMARIA

   La struttura primaria corrisponde alla specifica sequenza degli amminoacidi del backbone.

   Essa non descrive la struttura biologicamente attiva della proteina, ma ne determina tutte le proprietà chimiche e contiene l'informazione necessaria e sufficiente a definire gli ordini di struttura superiori.

   Nella struttura primaria vi è quindi l'informazione per guidare il processo di ripiegamento della proteina verso la propria conformazione funzionale attiva.

   Estrarre tale informazione per trovare lo stato nativo partendo dalla sequenza lineare di amminoacidi è proprio ciò che costituisce il *Protein Folding Problem*.

   Gli amminoacidi possono presentarsi in tutte le combinazioni possibili, ripetendosi più volte[3]. La lunghezza della catena peptidica può variare da pochi residui ad diverse centinaia.

   Anche grazie ai miglioramenti nelle tecniche di sequenziamento dei genomi sono ormai milioni le sequenze proteiche note.

   La grande disponibilità di informazioni ha mostrato come in organismi diversi esistono sequenze diverse che però codificano per proteine che da un punto di vista strutturale e funzionale sono sovrapponibili.

   Questo è strettamente legato al concetto che la sequenza si evolve molto più rapidamente rispetto alla struttura.

2. STRUTTURA SECONDARIA

   Il primo passo nel processo di ripiegamento della proteina passa attraverso la formazione di semplici conformazioni locali ordinate formate da legami ad idrogeno.

   In base alla natura degli amminoacidi e agli angoli di legame, il polipeptide può assumere localmente conformazioni più complesse tra cui riconosciamo le α-eliche (alfa eliche) e i β-sheets (foglietti beta).

   I fattori fondamentali che intervengono nella creazione della struttura secondaria sono la minimizzazione dell'ingombro sterico delle catene laterali e la loro carica: se catene laterali che si trovano in posizioni molto vicine tra loro hanno cariche omologhe, soprattutto in soluzione, si potrà generare della repulsione che impedirà così la formazione del legame ad idrogeno all'interno della catena principale.

   Le conformazioni che portano a questo arrangiamento regolare sono presenti sulla mappa di Ramachandran (figura 5) dove abbiamo la rappresentazione degli angoli torsionali Phi (φ) e Psi (ψ) e sulla base di questo sappiamo che alcune zone sono preferite.

   Quindi già dall'analisi della sequenza è possibile stabilire una preferenza a formare delle α eliche piuttosto che un foglietto β.

   Le alfa eliche formano delle strutture spiralizzate regolari che sono stabilizzate da ponti idrogeno locali.

[3] In realtà solo poche fra le combinazioni possibili corrispondono a proteine: le differenze fondamentali fra una sequenza proteica ed una casuale sono ancora sconosciute.

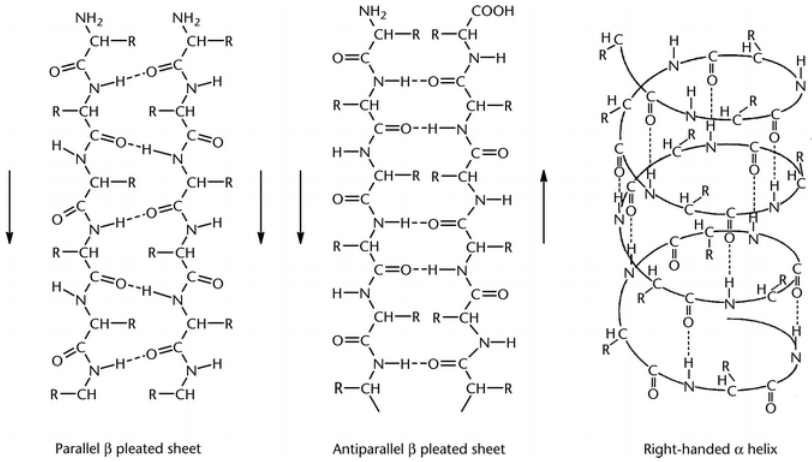

Figura 7: Esempi di struttura secondaria

Le catene R dei residui si posizionano verso l'esterno della struttura a spirale a causa del loro impedimento sterico all'interno della catena.

Un'α-elica è presente quasi sempre nella forma destrogira, con il lato interno che tende ad accumulare amminoacidi idrofobici, e quello esterno amminoacidi idrofilici.

I foglietti beta assumono un ripiegamento molto più disteso in confronto a quello descritto in precedenza, e la struttura in questo caso è stabilizzata da ponti idrogeno tra amminoacidi che sono lontani in sequenza.

In un foglietto beta la catena polipeptidica è ripiegata con andamento a zig-zag (filamento β) ed i gruppi R sono posti perpendicolarmente al piano dei legami peptidici con direzione opposta.

La catena così ha una distanza assiale tra due residui adiacenti molto più distesa, che passa da 1,5 Å dell'α elica a 3,5 Å nella struttura β a pieghe.

I β piani possono formarsi: tra catene polipeptidiche parallele (foglietto β parallelo: con lo stesso orientamento ammino-terminale e carbossi-terminale del polipeptide); tra catene polipeptidiche antiparallele (foglietto β antiparallelo: con orientamento in senso contrario dei gruppi ammino-terminali e carbossi-terminali); in una sola catena polipeptidica che si ripiega su se stessa formando tratti paralleli o antiparalleli.

Oltre alle due strutture regolari appena descritte, nelle proteine sono presenti tratti di catena apparentemente disorganizzati, detti loops, che collegano le strutture secondarie ed hanno un ruolo importante nell'organizzazione 3D della struttura molecolare.

I loops si trovano generalmente nelle regioni esterne della proteina e presentano di conseguenza catene laterali per lo più idrofiliche.

Inoltre i legami idrogeno tra gli amminoacidi del loop e le molecole d'acqua circostanti sono in numero maggiore rispetto a quelli effettuati con gli amminoacidi adiacenti. Tale peculiarità conferisce una relativa flessibilità a tali regioni e consente cambi di direzione anche repentini alle sequenze con conformazione α e β che vanno a collegare.

Sebbene alcune di queste regioni possano essere molto lunghe (fino a venti amminoacidi), nella maggior parte dei casi sono composte da due fino a dieci amminoacidi.

Gli hairpin loops, i loop più corti conosciuti (2-5 amminoacidi) vengono anche chiamati "reverse turns " per la loro proprietà di collegare due foglietti-β adiacenti eseguendo una inversione nella direzione della sequenza.

Considerando questi loops, negli ultimi anni è emerso il concetto di struttura *super-secondaria*.

Si è visto che una buona parte delle strutture proteiche tende ad essere composta da elementi regolari che vanno oltre la singola α-elica o il singolo foglio-β.

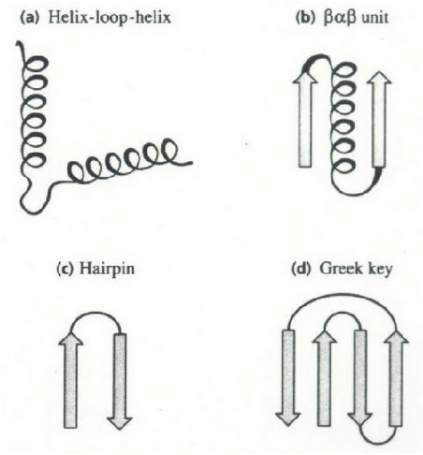

Figura 8: Motivi strutturali

I motivi strutturali più ricorrenti sono:

- elica-loop-elica: due α-eliche collegate fa un turn;
- β-turn: è un elemento molto comune nelle strutture proteiche. Abbiamo un filamento β esteso, un turn che serve per invertire la conformazione della catena principale, e poi un altro filamento β della stessa lunghezza che va ad accoppiarsi e a formare ponti di idrogeno;
- β-α-β: due filamenti β paralleli intercalati da un'α-elica. I due loop di collegamento possono avere lunghezze molto variabili e funzioni specifiche diverse. In genere i filamenti β sono relativamente corti, e l'asse dell'elica è parallelo a quello dei filamenti-β.
- chiave greca: quattro filamenti-β, due brevi loop e un loop più lungo. La caratteristica del motivo a chiave greca è il diverso ordine dei filamenti-β antiparalleli componenti la struttura rispetto alla posizione nella catena peptidica.

3. STRUTTURA TERZIARIA

Si parla di struttura terziaria quando si ha il ripiegamento completo della proteina in una conformazione tridimensionale unica che ne determina la funzione.

Tale organizzazione, detta anche stato nativo, viene descritta attraverso le coordinate spaziali di tutti gli atomi del polipeptide.

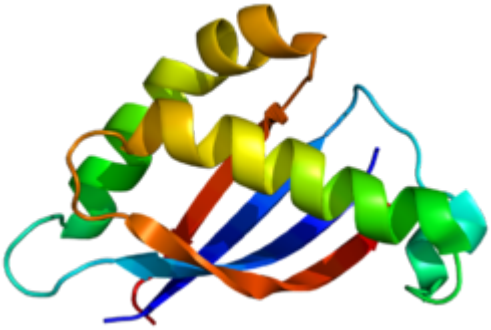

Figura 9: Struttura terziaria di una generica proteina

Questo ordine di struttura è determinato da una serie di interazioni di varia natura che si stabiliscono tra le catene laterali e che portano a ripiegamenti ulteriori rispetto a quelli dati dalle strutture secondarie.

Tali interazioni sono di tipo debole tra amminoacidi idrofobici, interazioni dipolari tra amminoacidi con carica opposta, legami a ponte idrogeno o legami a ponte disolfuro.

Visto il loro grande numero, forniscono un contributo talvolta più stabilizzante di un legame covalente.

Complessivamente questo insieme di legami porta ad esporre al solvente (in condizioni fisiologiche l'acqua) le parti polari della catena, ospitando all'interno della proteina o del peptide le parti non polari.

DOMINI PROTEICI

Un altro concetto utile è quello dei *domini proteici,* delle regioni compatte ed uniformi che ripiegano in modo autonomo e quindi potrebbero presumibilmente esistere anche in assenza delle parti restanti della proteina.

Per motivi sperimentali le strutture che vengono risolte molto spesso contengono soltanto un singolo dominio, questo perché le tecniche sperimentali (cristallografia a raggi X e risonanza magnetica nucleare) consentono di analizzare un singolo dominio per volta.

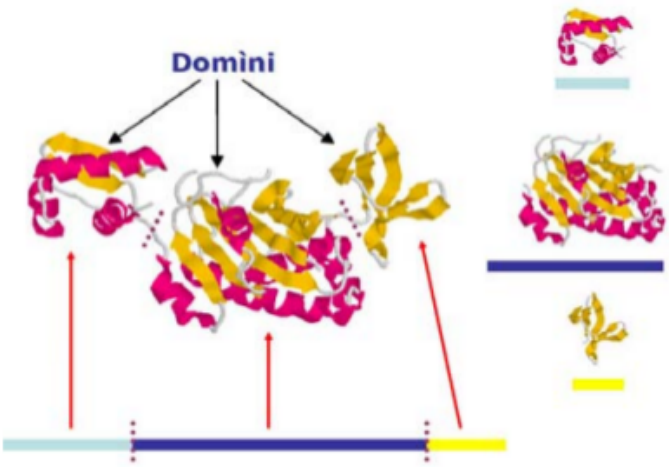

Figura 10: Domini proteici

4. STRUTTURA QUATERNARIA

Per buona parte delle proteine la struttura terziaria rappresenta l'ultimo livello di organizzazione strutturale. E' il caso delle proteine monomeriche costituite da un'unica unità funzionale biologicamente attiva.

Una singola proteina può però interagire con altre proteine per andare a formare dei complessi macromolecolari. Spesso tali proteine sono costituite da varie sub-unità essenzialmente uguali tra loro, come nel caso dell'emoglobina.

La struttura quaternaria riguarda la disposizione spaziale e topologica di queste sub-unità.

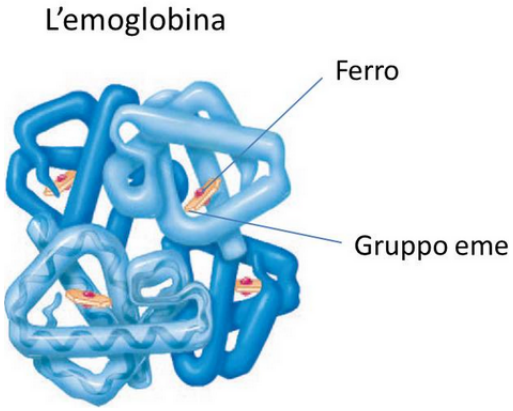

Figura 11: L'emoglobina è un esempio di struttura quaternaria

2.3 Metodi di analisi

A fronte di metodi sperimentali per la determinazione di sequenze proteiche ormai estremamente rapidi[4] e relativamente economici, la risoluzione empirica delle corrispondenti strutture 3D presenta ancora diversi ostacoli.

Le tecniche sperimentali permettono di ottenere modelli accurati e affidabili, ma non sono sempre applicabili e richiedono strumenti più complessi e talvolta mesi di lavoro.

In particolare si fa ricorso a:

- *cristallografia a raggi X* o *spettroscopia a risonanza magnetica nucleare (NMR).*

L'enorme importanza ricoperta dalla determinazione della struttura di una proteina nell'analisi e comprensione delle sue funzioni, ha portato a dedicare molte risorse ed energie allo sviluppo di metodi informatici per la predizione della struttura proteica che potessero se non sostituire, almeno indirizzare la ricerca svolta con i metodi tradizionali.

I processi computazionali infatti sono in genere semplici e veloci, anche se i risultati prodotti sono approssimazioni talvolta soggette ad errori notevoli. Metodi sperimentali e computazionali sono in ogni caso fortemente correlati. Un modello computazionale può essere convalidato da una serie di dati sperimentali e, viceversa, approcci empirici sono spesso guidati dalla costruzione di modelli virtuali che permettono di discriminare tra determinate soluzioni.

[4]la determinazione degli amminoacidi in una sequenza proteica richiede meno di un giorno

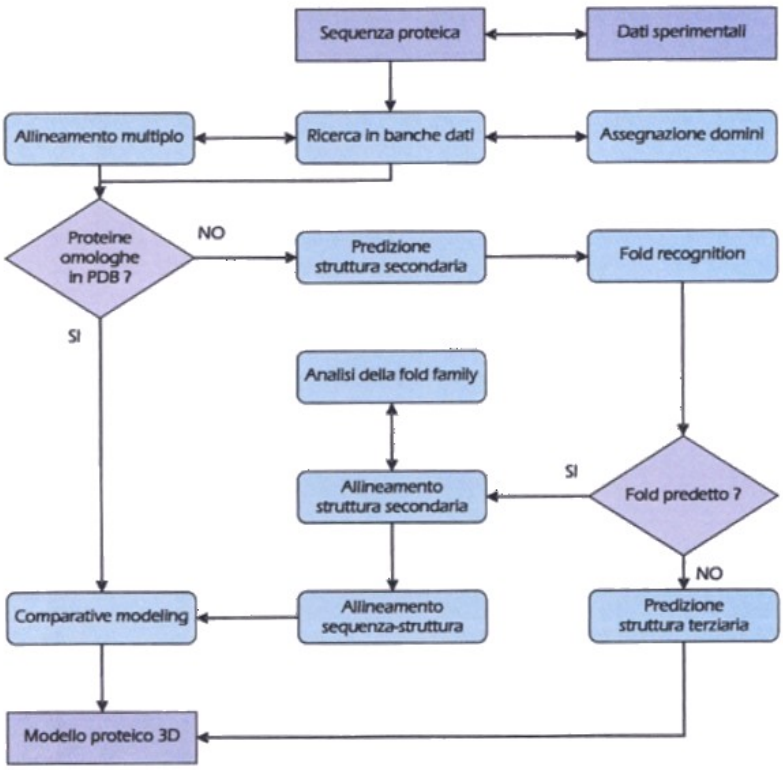

Figura 12: Metodi computazionali nella predizione della struttura proteica

In biologia computazionale la simulazione del processo di ripiegamento della catena polipeptidica in una conformazione 3D stabile costituisce una sfida nella quale, basandosi sul grado di similarità tra la sequenza sconosciuta (target) e le strutture raccolte nei database (templati), si distinguono tre principali approcci:

- *Comparative Modeling*, *Fold Recognition*, *Ab initio*.

Di seguito verranno approfondite entrambe le tipologie presentate.

2.3.1 Metodi sperimentali

Cristallografia ai raggi X

La cristallografia è la scienza che indaga la disposizione degli atomi nei solidi.

E' il metodo sperimentale usato più di frequente nella determinazione della struttura di una proteina. E' anche il più accurato in quanto capace di determinare strutture ad una risoluzione inferiore ai 2Å[5].

La risoluzione con cui una proteina viene risolta è una importantissima misura di qualità: più questa è bassa e maggiore sarà il numero di errori contenuti nella struttura[6].

Il primo passo nella cristallografia a raggi X è la cristallizzazione della proteina da analizzare.

[5]Un Angström (Å) è definito come 10^{-10} m. Una tipica lunghezza di legame varia tra 1.1 e 1.5 Å.
[6]Ad oggi strutture risolte con risoluzioni inferiori ai 3Å non sono considerate attendibili.

La formazione del cristallo si ottiene tramite congelamento, aggiunta di sale o in qualche altro modo, finché non si ottiene una struttura rigida e ordinata che possiamo esporre ad un fonte di raggi X per averne quindi una proiezione.

La mancanza di regole precise e la variabilità delle condizioni sperimentali (temperatura, concentrazione, presenza di soluti e cofattori, etc.) rendono questo passaggio lungo e problematico, ed il successo non è sempre garantito.

Ottenuto un numero sufficiente di cristalli, si colpisce il materiale con un fascio di raggi X che viene diffratto in direzioni specifiche.

A seconda degli angoli e dell'intensità di questi raggi diffratti un cristallografo può produrre un'immagine tridimensionale della densità di elettroni nel cristallo.

Da questa è infine possibile ricavare le posizione media degli atomi, così come anche i loro legami chimici ed altre informazioni.

Spettroscopia a risonanza magnetica nucleare

La risonanza magnetica nucleare (Nuclear Magnetic Resonance, NMR) si basa su proprietà quanto-meccaniche della materia immersa in campi magnetici.

Fornisce informazioni strutturali esaminando l'influenza dell'ambiente locale circostante sulla risposta ai campi magnetici degli atomi, derivando importanti informazioni sulle distanze inter-atomiche e sugli angoli torsionali.

Le strutture NMR non sono accurate tanto quanto quelle ottenute ai raggi X, ma hanno il vantaggio di usare la proteina in soluzione, che è il suo ambiente naturale.

Operare su proteine in soluzione, quindi non sempre nella stessa conformazione rigida, permette di valutare meglio la flessibilità della proteina.

Nell'NMR abbiamo pertanto una serie di istantanee della molecola, tipicamente 20-30 strutture simili, tutte comunque consistenti con i dati sperimentali raccolti.

Questi modelli saranno poi tutti inseriti nel relativo file PDB (sezione 4.3), semplicemente separandoli con una riga che contiene la parola chiave MODEL.

Il limite fondamentale di questa tecnica è costituito dalla soglia massima imposta alla dimensione della macromolecola analizzata (circa 100-300 residui).

2.3.2 Metodi computazionali

Comparative (or Homology) modeling[7]

La modellazione comparativa o per omologia si applica quando sono note strutture (templati) con sequenza molto simile alla proteina (target) da modellare, ed è anche il metodo che è stato sviluppato per questa tesi.

L'idea fondamentale è che proteine con un buon livello di similarità di sequenza risultano anche strutturalmente equivalenti[8].

[8] Le strutture si conservano molto più delle sequenze durante l'evoluzione [1].

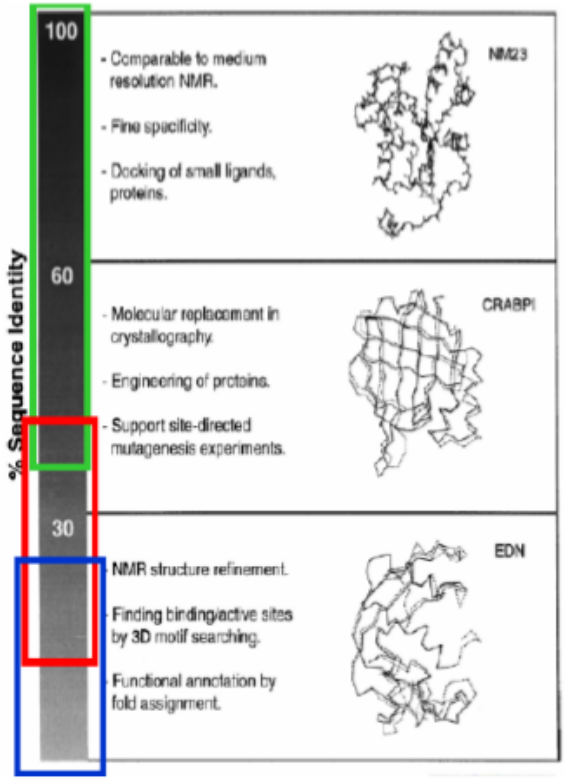

Figura 13: Folding e similarità di sequenza: relazione

Nelle proteine esistono poi regioni più o meno conservate: il confronto fra macromolecole omofunzionali con similarità di sequenza decrescente mostra come le regioni funzionalmente importanti siano in posizioni strutturalmente più conservate, in sequenza e struttura, rispetto ad altri tratti.
Questo tipo di approccio sostanzialmente richiede una ricerca di sequenze cristallizzate in banca dati, per esempio utilizzando il protocollo PSI-BLAST che verrà descritto nel capitolo 8.

Il modello del target viene costruito sulla base del templato così identificato, copiando le coordinate atomiche e ricostruendo le eventuali parti mancanti.

I risultati che possono essere ottenuti con il comparative modeling sono molto accurati.

Vengono posizionate anche le catene laterali e sui modelli ottenuti si possono fare considerazioni anche di carattere funzionale.

Per contro il limite richiesto è che vi sia un elevato grado di similarità tra la sequenza target e la sequenza templato, non inferiore al 30-35%.

Il protocollo generalizzato per la modellazione comparativa si compone generalmente dei seguenti passi:

- Identificazione di un insieme di templati. Si ricercano una o più proteine con struttura nota che presentino una similarità di sequenza superiore al 30% con la sequenza target.

- Allineamento della sequenza target alle sequenze templato. L'allineamento multiplo permette di individuare le regioni della proteina target più o meno conservate in tutte le strutture templato. Allineamenti fra sequenze con similarità pari o superiore al 70% sono in genere privi di complicazioni e possono essere tranquillamente affidati a procedure automatiche; in caso contrario spesso è necessario un intervento manuale esperto per ottenere un buon risultato. [9]
- Costruzione del modello grezzo. Una volta individuate le regioni strutturalmente conservate, la catena principale della struttura bersaglio viene allineata a questi frammenti, e vengono copiate tutte le coordinate atomiche utili alla formazione del nucleo del modello.
- Modellazione dei loop. Se sono disponibili più templati un approccio utile è quello di cercare modelli per i loop tra le strutture che condividono zone pre e post-loop simili.
- Posizionamento delle catene laterali. Terminato il modello della catena principale, si possono aggiungere gli atomi delle catene laterali facendo ricorso a librerie di rotameri (come è stato fatto in questa tesi) o sfruttando approcci di dinamica molecolare.
- Raffinamento e valutazione del modello. Manualmente o tramite calcoli di minimizzazione energetica si cerca di risolvere possibili problemi strutturali, come ad esempio collisioni tra catene laterali. Se si sono prodotti più modelli si sceglie il migliore attraverso valutazioni di energia.

Fold recognition

Quando la proteina target non manifesta una significativa similarità di sequenza con proteine a struttura nota (al di sotto del 35-40% di identità, fino al 15-20%), la tecnica del comparative modeling non può essere applicata.

Fold recognition sfrutta la conoscenza del fatto che il numero dei naturali ripiegamenti proteici (fold) è limitato [2].

Pertanto è plausibile che una sequenza con nessuna significativa similarità di sequenza, possa comunque avere una struttura simile a quella di una seconda sequenza.

Gli approcci maggiormente seguiti riguardano la ricerca di templati evolutivamente lontani tramite l'uso di tecniche di allineamento complesse (tipo allineamenti profilo contro profilo) o il threading.

Si possono inoltre usare informazioni sulla struttura secondaria o altri accorgimenti per limitare il numero di falsi positivi.

Il risultato che si può ottenere con la Fold Recognition è in genere un modello molto più approssimato rispetto al comparative modeling. Le catene laterali ad esempio non sempre vengono inserite.

[9] Proteine con similarità di sequenza superiore al 50% mantengono circa il 90% dei residui in posizioni strutturalmente conservate.

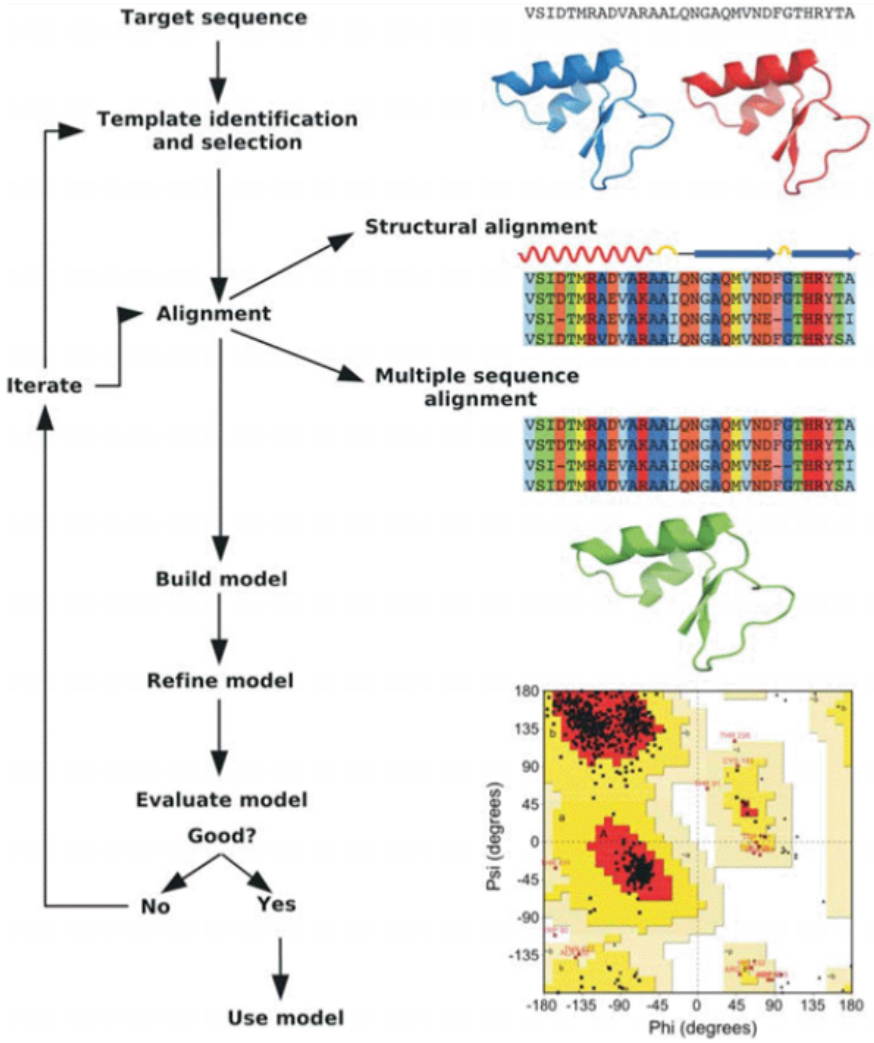

Figura 14: Protocollo seguito nel Comparative Modeling

Ab initio (o novel fold)

Ab initio è un metodo che non utilizza templati, ma cerca di costruire una struttura modello basata sulle proprietà fisico-chimiche della catena amminoacidica.

I calcoli sono basati su complesse funzioni di energia e per questo in genere richiedono lunghi tempi di calcolo.

Si tratta di una metodologia che per il momento non da ancora risultati accettabili in quanto la teoria del folding non è ancora sufficientemente spiegata.

3 Allineamento di sequenze

L'allineamento di sequenze biologiche è un passo imprescindibile per capire la funzione di molte proteine. Oltre a servire allo scopo di allineare due sequenze tra loro, è anche il presupposto per analisi più complesse come le ricerche di similarità in banche dati, la costruzione di alberi filogenetici o il riconoscimento di pattern specifici e domini funzionali.

Le metodologie di allineamento adottate per il progetto HOMER sono più elaborate di quelle che vengono introdotte in questo capitolo, ma i concetti espressi verranno comunque ripresi.

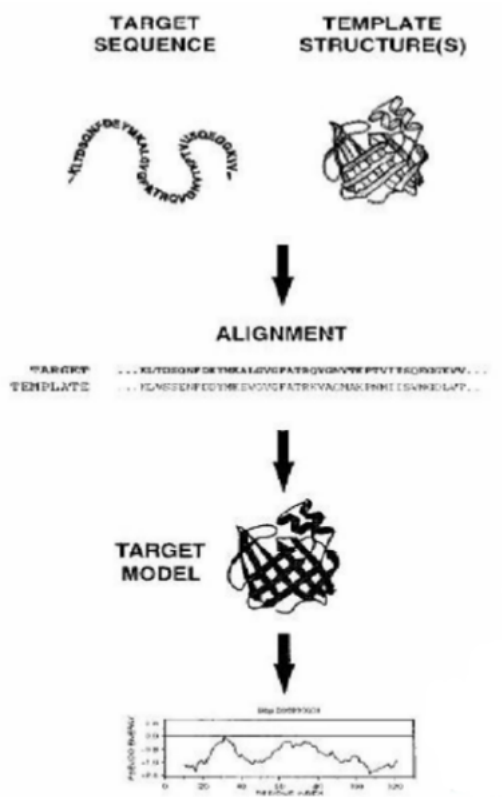

Figura 15: Un buon allineamento è condizione necessaria nella produzione di un modello

3.1 Metodi esatti

Nell'allineamento "sequenza contro sequenza" l'algoritmo che da i migliori risultati è la programmazione dinamica.

Questo approccio si basa sul principio di Bellman (1975), nel quale si afferma che che i problemi complessi si possono opportunamente decomporre in sotto problemi più semplici.

L'allineamento sequenza contro sequenza dipende dalla scelta di tre parametri fondamentali: la matrice di sostituzione, il *gap open* e il *gap extension*. Questi serviranno per generare una tabella chiamata matrice di allineamento.

La matrice di sostituzione viene costruita a partire dal fatto che le proteine sono costituite da 20 amminoacidi i quali presentano caratteristiche chimico-fisiche diverse tra loro. Pertanto le singole sostituzioni amminoacidiche non avranno lo stesso peso.

Nonostante le matrici di similarità possano basarsi direttamente sulle proprietà chimico-fisiche dei singoli amminoacidi, quelle attualmente più utilizzate sono state sviluppate con metodi statistici.

I vari punteggi indicano la frequenza con cui un amminoacido si sostituisce ad un altro in famiglie di proteine omologhe.

Le più usate oggi per gli allineamenti sono BLOSUM e PAM [3].

Gli altri elementi fondamentali per la costruzione di una tabella di allineamento sono i valori associati ai gap. Con il termine gap si intende un mancato accoppiamento di un amminoacido in una sequenza con un amminoacido nell'altra.

Si tratta in pratica di una inserzione o delezione (a seconda di quale delle due sequenze presenta il gap) la cui lunghezza può essere variabile.

Nella costruzione delle tabelle di allineamento si distingue tra gap open e gap extension assegnando nel primo caso un punteggio più negativo: si è osservato infatti che da un punto di vista statistico e biologico, l'apertura di un nuovo gap è molto più difficile della continuazione di un gap già presente.

Una volta definiti i tre parametri si procede con la compilazione della tabella di allineamento, che avrà dimensioni pari alla lunghezza delle due sequenze da allineare.

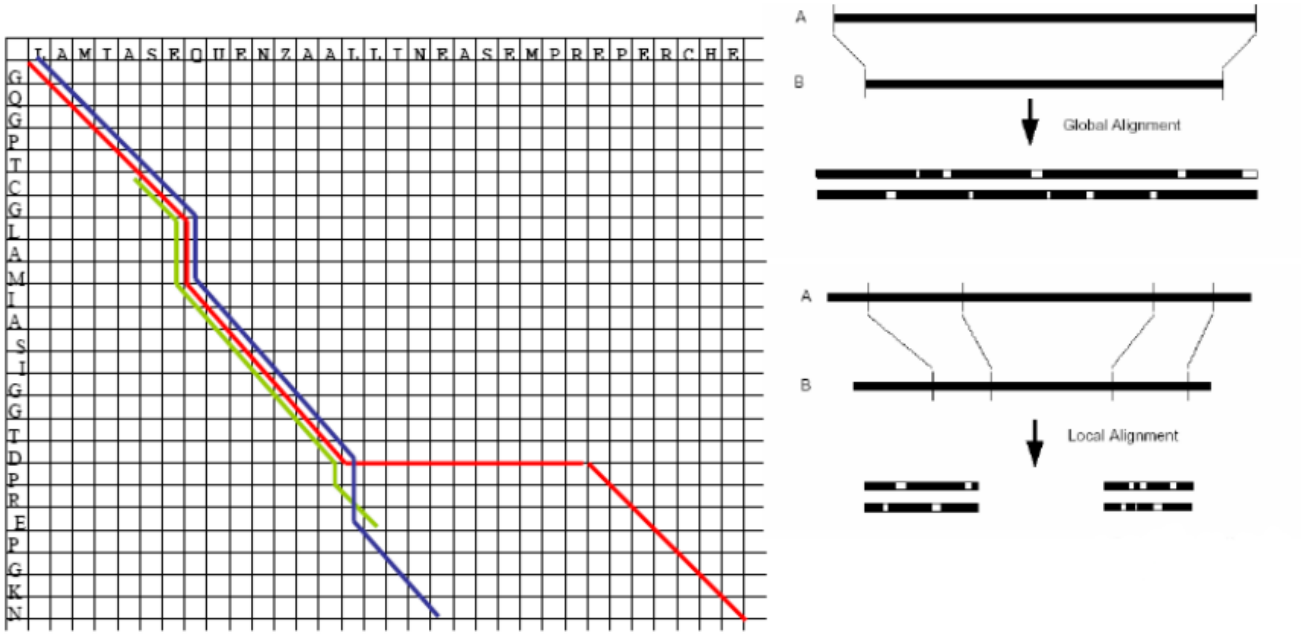

Figura 16: Tabella di allineamento

La tabella verrà completata inserendo in ogni casella un punteggio che deriverà dai parametri prestabiliti e dalla somma dei punteggi delle caselle di provenienza.

Esistono tre metodi per calcolare un percorso all'interno di una tabella dei punteggi:

- ALLINEAMENTO GLOBALE: noto anche come algoritmo di Needlemann & Wunsch [4], impone sempre un allineamento che comprende tutti i residui delle due sequenze, indipendentemente dalla loro similarità. Il problema maggiore con questo tipo di approccio è di forzare

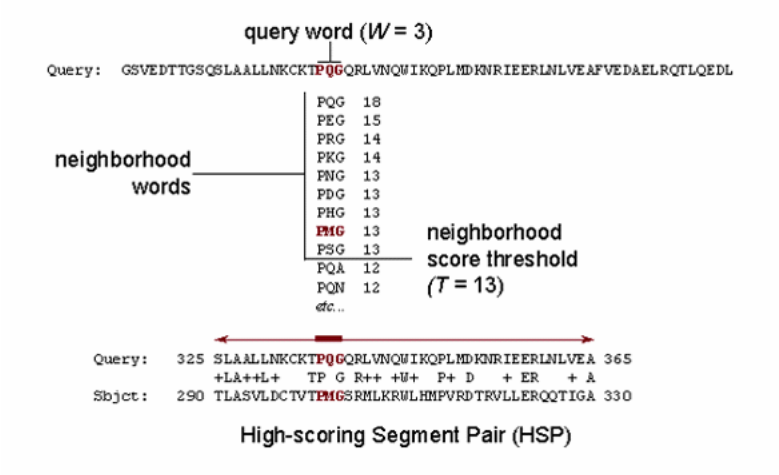

Figura 17: L'algoritmo BLAST

troppo un allineamento rendendolo non significativo: le sequenze verranno sempre allineate da N-term a C-term indipendentemente dalla loro lunghezza e dal tipo di sequenza. Per questo motivo gli allineamenti globali vengono applicati quando si vogliono confrontare accuratamente due sequenze in cui la similarità sia estesa per tutta la lunghezza. Graficamente significa andare sempre dalla prima posizione in alto a sinistra sino all'ultima in basso a destra della matrice.

- ALLINEAMENTO LOCALE: noto con il nome di algoritmo di Smith &Watermann [5], questo tipo di allineamento, a differenza del precedente, permette di allineare soltanto i residui più simili di due sequenze. Il limite è che la regione allineata sia più corta rispetto alla regione realmente significativa. Si concentra solo su una parte molto breve della matrice che avrà il punteggio di similarità più elevato.

- ALLINEAMENTO FREESHIFT: Si ottiene combinando gli algoritmi globale e locale, e nasce per far fronte alle loro limitazioni. Identifica regioni di effettiva similarità, ottenendo un allineamento più ricco di quello locale. Visivamente significa procedere da un'estremità all'altra della matrice.

3.2 Metodi euristici

Gli algoritmi di programmazione dinamica sono perfetti per allineare due sequenze in modo esatto, ma sono troppo lenti e quindi non adatti ad effettuare ricerche di similarità su grandi banche dati.

In questi casi si ricorre ad algoritmi euristici, fondati su assunzioni non certe ma estremamente probabili, in grado di fornire in tempi accettabili una soluzione vicina a quella ottima.

BLAST [6] (Basic Local Alignment Search Tool) è il tool più popolare per questo tipo di ricerche. Esegue confronti fra coppie di sequenze alla ricerca di regioni di similarità, piuttosto che un allineamento globale tra le intere sequenze.

BLAST può eseguire migliaia di confronti fra sequenze in pochi minuti e in poco tempo è possibile confrontare una sequenza query con l'intero database per ricercare tutte le sequenze simili ad essa.

L'algoritmo si compone di più fasi:

- Si estraggono tutte le possibili *word* di m lettere dalla sequenza query (in genere m=3 per le proteine, 11 per il DNA).

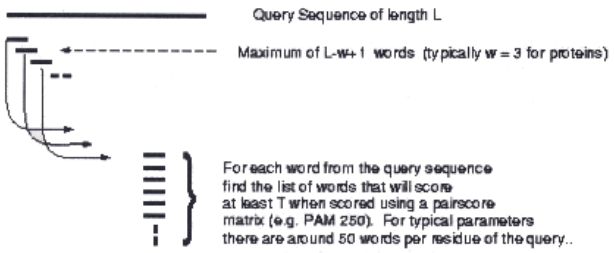

Figura 18: Blast: passo 1 e 2

- Per ogni word della sequenza da esaminare viene costruita una lista di possibili words che, se confrontate con la sequenza in questione, hanno un punteggio superiore ad un valore soglia T (compreso tra 11 e 15) calcolato di volta in volta in base alla composizione e alla lunghezza della sequenza in esame.
- Si confronta la lista di words con le sequenze contenute nel database alla ricerca di match esatti.

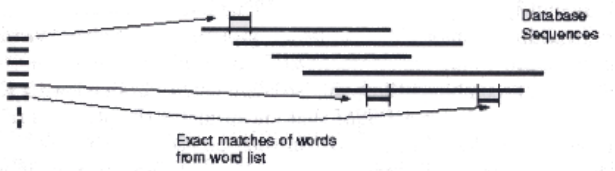

Figura 19: Blast: passo 3

- Quando viene riscontrata una corrispondenza (hit), essa viene estesa a monte e a valle per vedere se è possibile definire un tratto di sequenza in grado di raggiungere un punteggio superiore ad un valore-soglia S.

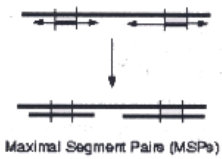

Figura 20: Blast: passo 4

Per la presente tesi verrà utilizzata una sua particolare implementazione detta PSI-BLAST, che sfrutta un approccio iterativo i cui le sequenze trovate ad ogni ciclo sono usate per costruire un modello di punteggio per la ricerca del ciclo successivo.

L'argomento verrà discusso nel capitolo 8.

4 Banche dati e PDB

I recenti progressi della biologia molecolare e dell'ingegneria genetica hanno prodotto un'enorme quantità di materiale scientifico, portando alla ribalta la necessità di nuovi sistemi di organizzazione, accesso e fruizione delle informazioni.

Questa esigenza da dato un forte impulso allo sviluppo di imponenti banche dati; strumenti oggi fondamentali per la ricerca e la divulgazione dei risultati.

4.1 Banche dati di sequenze proteiche: Swiss-Prot

Le banche dati di sequenze proteiche raccolgono sequenze proteiche ottenute sia dalla determinazione sperimentale di sequenze amminoacidiche, sia dalla traduzione di sequenze nucleotidiche (DNA e RNA) per le quali è stata individuata o predetta la funzione di gene codificante per una proteina.

I dati vengono accuratamente validati e arricchiti di informazioni specifiche.

Una delle più importanti banche dati di sequenze proteiche è Swiss-Prot.

Creata nel 1986 da Amos Bairoch, è sviluppata in Svizzera a Ginevra dallo Swiss Institute of Bioinformatics (SIB) e dallo European Bioinformatics Institute (EBI).

L'obiettivo di Swiss-Prot è quello di fornire sequenze proteiche affidabili corredate di un buon numero di informazioni addizionali, come la funzione della proteina, i suoi domini funzionali, la presenza di amminoacidi modificati, regioni peptidiche, siti di splicing proteici, polimorfismi e altri segnali e dati rilevanti per la struttura della proteina.

In SWISS-PROT sono riportate anche le informazioni relative ad alterazioni della proteina e si cerca di garantire una ridondanza minima ed un alto livello di integrazione con le altre banche dati bioinformatiche.

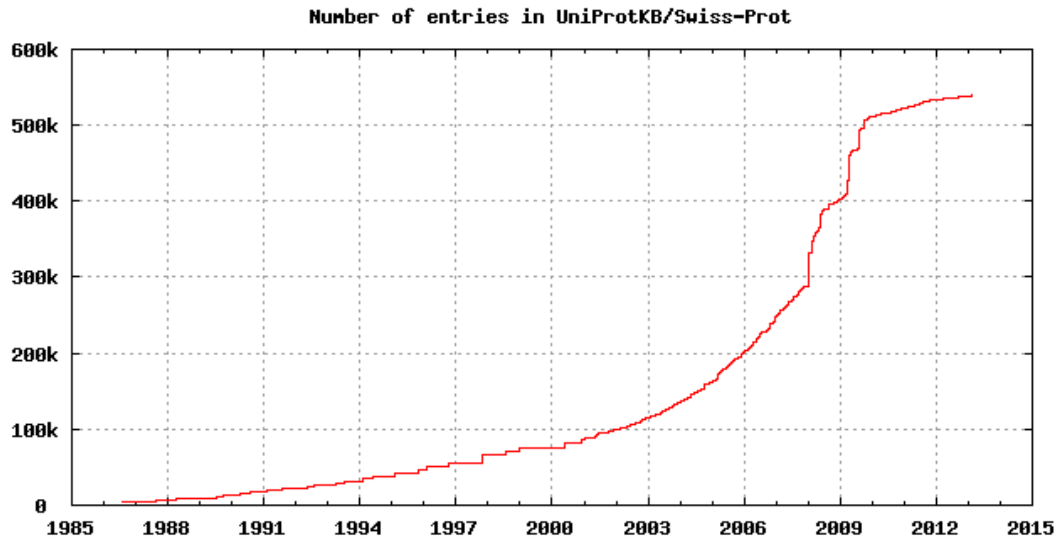

Figura 21: Crescita del numero di sequenze proteiche depositate nella banca dati Swiss-Prot

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	72610	1435	3730	3	77778
NMR	8599	1022	192	7	9820
ELECTRON MICROSCOPY	342	41	123	0	506
HYBRID	46	3	2	1	52
other	147	4	5	13	169
Total	81744	2505	4052	24	88325

Tabella 1: Tavola riassuntiva delle strutture depositate nella banca dati PDB

4.2 Banche dati di strutture proteiche: PDB

Le informazioni strutturali di una proteina riguardano la distribuzione spaziale degli atomi di tutti i suoi amminoacidi. Tali dati corrispondono alle coordinate atomiche determinate attraverso vari metodi sperimentali di analisi strutturale.

La più importante banca dati mondiale di strutture proteiche è senz'altro la PDB (Protein Data Bank).

Fondata nel 1971 dal BNL (Brookhaven National Laboratory), dal 1998 è ospitata presso l'RCSB (Research Collaboratory for Structural Bioinformatics con sede alla Rutgers University, negli USA).

Ospita tutte le strutture che sono state risolte sperimentalmente (principalmente con le tecniche a raggi X o NMR) e che sono disponibili al pubblico:

- **strutture proteiche** (inclusi complessi proteici, capsidi virali, cofattori, substrati etc), il 92% del totale delle strutture presenti nella banca dati;
- **strutture di acidi nucleici** (DNA e RNA), il 4% del totale;
- **strutture di complessi proteici/nucleici** (es. fattori di trascrizione legati al DNA e ribosomi), 4% del totale;
- **strutture di altre macromolecole** (es. carboidrati), poche decine.

PDB è un database ridondante e può contenere più versioni di una stessa proteina depositate in tempi successivi.

Questo perché la struttura di una proteina può essere stata risolta con metodi e risoluzioni differenti, nella sua forma libera o co-cristallizzata con altri ligandi e cofattori, o presentare svariate mutazioni con struttura 3D pressoché identica.

Ogni struttura ha un suo codice identificativo di quattro simboli: un numero e tre caratteri alfanumerici. Il numero rappresenta la versione del file, mentre i caratteri ricordano, quando possibile, i nomi delle strutture a cui sono associati.

Si stima che almeno tre quarti delle strutture depositate siano molto simili tra loro [1], riducendo il numero di strutture "uniche" a meno di 5000; fatto su cui poggiano la loro validità i vari strumenti di modellazione per comparazione.

Utilizzando la PDB bisogna tenere conto che non si tratta di una banca dati ideale.

Molto spesso infatti le informazioni delle coordinate non sono omogenee tra loro, e può verificarsi il caso che manchino le informazioni sulla posizione di alcuni atomi o di un intero gruppo

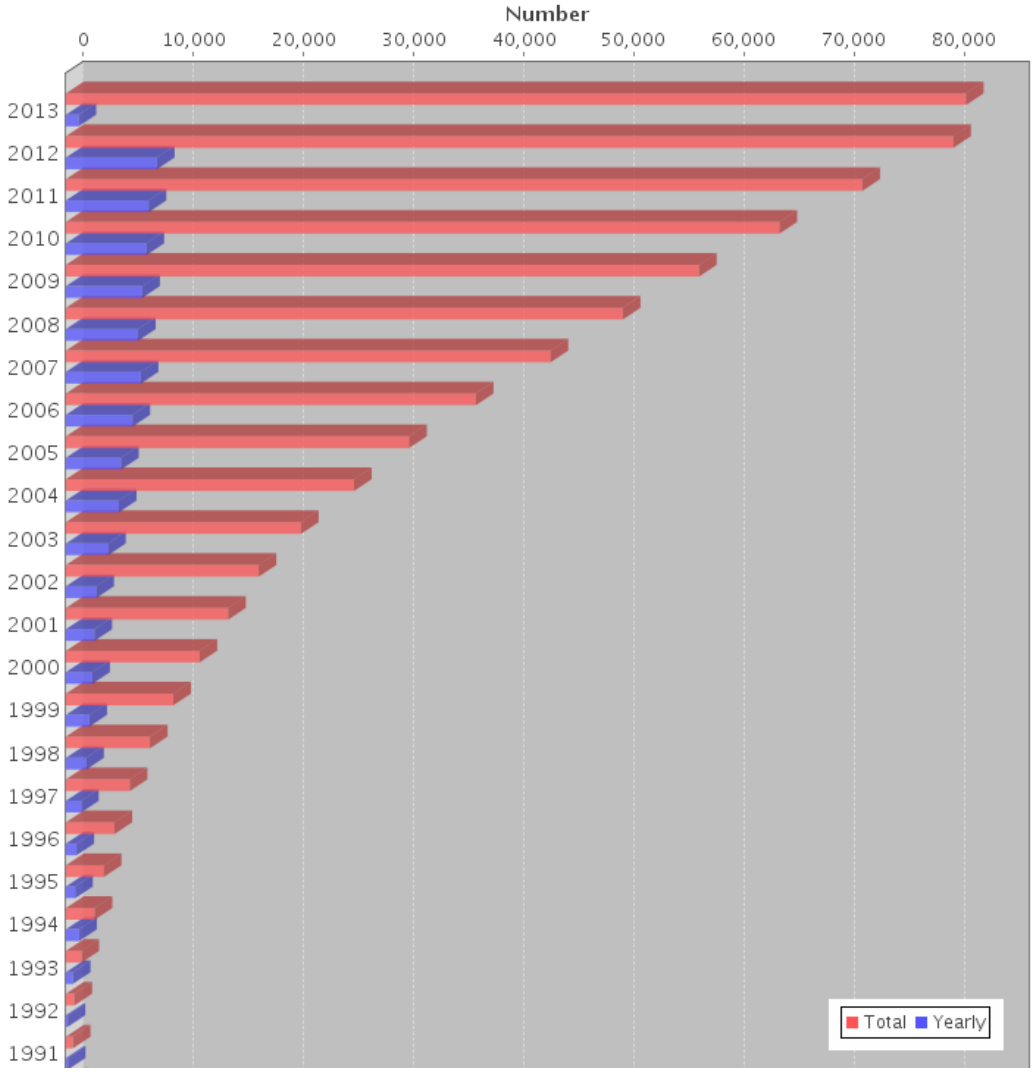

Figura 22: Crescita del numero di strutture depositate nella PDB. Il dato del 2013 è aggiornato al 19 febbraio

di residui a causa dei limiti dei metodi sperimentali di diffrazione e di spettroscopia NMR nella risoluzione delle strutture.

4.3 Formato PDB

Nella Protein Data Bank le informazioni sulle strutture depositate sono organizzate secondo l'omonimo formato.

Il formato PDB è un file di testo organizzato in colonne, e viene utilizzato dai vari programmi di visualizzazione per poter visualizzare la molecola nelle sue varie sfaccettature.

Ogni file PDB è essenzialmente diviso in due parti: la prima parte contiene la descrizione della molecola contenuta, gli autori, i dettagli sperimentali della risoluzione della struttura, la risoluzione, la lista dei residui della proteina, la descrizione della struttura secondaria e così via.

La seconda parte riporta invece le coordinate atomiche dei residui della macromolecola.

All'inizio di ogni riga è presente una parola chiave che definisce il tipo di informazione che segue. Di seguito si riportano le principali:

- HEADER: identifica l'intestazione della proteina.
- COMPD: abbreviazione di compound (composizione), è la definizione esatta della proteina.
- AUTHOR: gli autori.
- REMARC: ogni commento che gli autori abbiano ritenuto necessario.
- SEQRES: sequenza amminoacidica della proteina. Questa informazione è importante perché non sempre coincide con quello che effettivamente si riesce a vedere della struttura in modo sperimentale.
- HELIX, SHEET, TURN: è l'informazione sulla struttura secondaria.
- ATOM: sezione che riporta le coordinate atomiche: per ogni amminoacido viene riportato un numero progressivo in relazione alla sequenza amminoacidica, il tipo, la catena di appartenenza, gli atomi che lo compongono e le relative coordinate tridimensionali (X,Y,Z), l'occupancy e il B_factor. L'occupancy in particolare è un valore che esprime quanto è certa la posizione di un determinato atomo: di solito è il 100%, meno se si hanno più conformazioni alternative (spesso è il caso di strutture risolte tramite NMR). In alternativa può essere utilizzato il B-factor, o fattore di temperatura, che indica nelle strutture cristallografiche quanto era mobile quell'atomo (più è basso il valore, più è sicura la posizione di quell'atomo).
- HETATM: è l'informazione riguardante tutto ciò che non è proteina e che è presente nell'informazione sperimentale (co-fattori metallici, pezzi di DNA, molecole d'acqua, etc.).

```
Nome              HEADER    LYASE (CARBON-OXYGEN)                   05-JUN-95   1PDZ
                  TITLE     X-RAY STRUCTURE AND CATALYTIC MECHANISM OF LOBSTER ENOLASE
Composizione      COMPND    MOL_ID: 1;
                  COMPND    2 MOLECULE: ENOLASE;
                  COMPND    3 CHAIN: A;
Sorgente          SOURCE    MOL_ID: 1;
                  SOURCE    2 ORGANISM_SCIENTIFIC: HOMARUS GAMMARUS;
                  SOURCE    3 ORGANISM_COMMON: EUROPEAN LOBSTER;
Parole Chiave     KEYWDS    LYASE (CARBON-OXYGEN)
Metodo            EXPDTA    X-RAY DIFFRACTION
Autore            AUTHOR    J.JANIN,S.DUQUERROY,C.CAMUS,G.LE BRAS
                  REVDAT    3   13-JUL-11 1PDZ    1       VERSN
                  REVDAT    2   24-FEB-09 1PDZ    1       VERSN
                  REVDAT    1   14-NOV-95 1PDZ    0
Note              REMARK    1
                  REMARK    1 REFERENCE 1
                  REMARK    1  AUTH   C.DUMAS,S.DUQUERROY,J.JANIN
                  REMARK    1  TITL   PHASING WITH MERCURY AT 1 ANGSTROM WAVELENGTH
                  ...
Sequenza          SEQRES    1 A  434  ACE SER ILE THR LYS VAL PHE ALA ARG THR ILE PHE ASP
residui           SEQRES    2 A  434  SER ARG GLY ASN PRO THR VAL GLU VAL ASP LEU TYR THR
                  SEQRES    3 A  434  SER LYS GLY LEU PHE ARG ALA ALA VAL PRO SER GLY ALA
                  SEQRES    4 A  434  SER THR GLY VAL HIS GLU ALA LEU GLU MET ARG ASP GLY
                  ...
Struttura         HELIX     1   1 TYR A   56  GLY A   58  5                                   3
secondaria        HELIX     2   2 PHE A   62  ASP A   70  1                                   9
                  HELIX     3   3 ILE A   72  SER A   79  1                                   8
                  ...
                  SHEET     1   A 3 LYS A   4  PHE A  11  0
                  SHEET     2   A 3 PRO A  17  THR A  25 -1  N  TYR A  24   O  LYS A   4
                  SHEET     3   A 3 GLY A  28  ALA A  33 -1  N  ALA A  32   O  VAL A  21
                  ...
                  CRYST1  110.800  110.800   73.400  90.00  90.00 120.00 P 31 2 1      6
Riferimenti       ORIGX1      1.000000  0.000000  0.000000        0.00000
atomici           ORIGX2      0.000000  1.000000  0.000000        0.00000
                  ORIGX3      0.000000  0.000000  1.000000        0.00000
                  SCALE1      0.009025  0.005211  0.000000        0.00000
                  SCALE2      0.000000  0.010421  0.000000        0.00000
                  SCALE3      0.000000  0.000000  0.013624        0.00000
Coordinate        ATOM      4  N   SER A   1      17.116  38.085  15.047  1.00 24.12           N
atomiche          ATOM      5  CA  SER A   1      18.102  38.089  16.116  1.00 28.37           C
                  ATOM      6  C   SER A   1      19.523  38.443  15.615  1.00 24.34           C
                  ATOM      7  O   SER A   1      19.884  38.179  14.468  1.00 24.86           O
                  ATOM      8  CB  SER A   1      18.084  36.756  16.862  1.00 29.91           C
                  ATOM      9  OG  SER A   1      17.293  36.859  18.039  1.00 40.25           O
                  ...
                  ATOM   3296  N   SER A 433       6.260  68.088   4.842  1.00 53.41           N
                  ATOM   3297  CA  SER A 433       5.449  66.927   5.215  1.00 57.51           C
                  ATOM   3298  C   SER A 433       4.863  67.024   6.633  1.00 62.86           C
                  ATOM   3299  O   SER A 433       3.872  66.313   6.913  1.00 64.58           O
                  ATOM   3300  CB  SER A 433       6.299  65.648   5.091  1.00 53.60           C
                  ATOM   3301  OG  SER A 433       7.068  65.630   3.893  1.00 40.17           O
                  ATOM   3302  OXT SER A 433       5.415  67.787   7.450  1.00 69.30           O
                  TER    3303       SER A 433
Etero-atomi       HETATM 3304  P   PGA A 439      31.884  69.406  13.709  0.50 28.44           P
                  HETATM 3305  O1P PGA A 439      31.147  70.209  12.544  0.50 24.59           O
                  HETATM 3306  O2P PGA A 439      31.170  69.896  14.892  0.50 29.65           O
                  HETATM 3307  O3P PGA A 439      31.666  67.975  13.310  0.50 30.23           O
                  HETATM 3309  C2  PGA A 439      29.746  70.173  12.524  0.50 23.69           C
                  ...
                  HETATM 3313 MN    MN A 440      26.634  74.962  12.870  1.00 42.19          MN
                  HETATM 3314  O   HOH A 441      26.513  73.398  11.547  1.00 29.12           O
                  HETATM 3315  O   HOH A 442      28.502  74.232  13.907  1.00 62.21           O
                  END
                                                    [X]     [Y]     [Z]
```

Figura 23: Il formato PDB

5 Stato dell'arte: CASP

Il CASP (Critical Assesment of Tecniques for Protein Structure Prediction) è una competizione internazionale che valuta lo stato della ricerca sul ripiegamento proteico [7,8].

Agli inizi degli anni 90 molti gruppi di ricerca presentavano pubblicazioni in cui dichiaravano di aver risolto il problema della predizione della struttura delle proteine, mentre in realtà ogni metodo riuscita a risolvere bene solo un limitato numero di proteine.

Per mettere ordine in questo frangente ed evitare che la ricerca scientifica prendesse direzioni sbagliate, nel 1994 John Moult indisse la prima edizione del CASP, che da allora viene organizzato ogni due anni con lo scopo di incentivare il miglioramento delle strategie computazionali predittive.

I gruppi che partecipano alla competizione sono chiamati a generare modelli tridimensionali di una serie di proteine la cui struttura non sia ancora nota se non agli organizzatori del concorso.

Lo scopo è quello di verificare se i metodi sviluppati possono funzionare anche alla cieca, misurando lo stato dell'arte e i miglioramenti in tutti i maggiori settori della predizione di strutture proteiche.

Le diverse predizioni vengono valutate da assessors indipendenti sulla base delle strutture sperimentali. La premiazione avviene durante una conferenza in cui vengono illustrati i metodi utilizzati per le predizioni.

Lo scopo della conferenza è quello di stimolare la competizione tra i gruppi di ricerca, individuando quali siano i metodi migliori e quindi verso quali ambiti debba essere indirizzata la ricerca bioinformatica anche in vista dell'edizione successiva.

Oggi il CASP è suddiviso in numerose categorie di competizione, mentre il CASP originale ne prevedeva soltanto tre, una per ognuno dei metodi con cui si determina la struttura di una proteina: Comparative Modeling, Fold Recognition ed Ab Initio.

Le prime due categorie nelle ultime edizioni sono state raggruppate in una unica (almeno per il CASP), ossia quella delle predizioni basate su templati.

Si è visto che un approccio di homology modeling molto "naif" tende a commettere una serie di errori che si possono facilmente correggere usando delle informazioni in più. Quindi per migliorar la qualità dei modelli si integrano una serie di tecniche tipiche della fold recognition quali ad esempio l'uso delle informazioni sui profili o sulla struttura secondaria.

Nel 1998 nasce anche il CAFASP (Critical Assesment Fully Automated of Tecniques for Protein Structure Prediction) con lo scopo di valutare separatamente una nuova tipologia di predittori i cui calcoli sono completamente automatizzati.

La costruzione di modelli proteici richiede tutta una serie di step cui l'intervento è spesso fondamentale per riuscire ad avere buoni risultati, in particolare nella ricerca del templato e nell'allineamento delle due sequenze target-templato. Ne segue che spesso i ricercatori meno esperti in bioinformatica tendono a non utilizzare gli strumenti di predizione che potrebbero aiutare le loro ricerche.

Per risolvere questa situazione e avvicinare tutta la comunità scientifica all'uso di questi strumenti, si tenta da qualche anno di sviluppare dei server automatici di predizione in cui l'intervento umano non sia assolutamente necessario.

In quest'ottica separare all'interno del CASP le valutazioni dei metodi automatici da quelli che prevedono l'intervento umano ha servito per avere una valutazione più mirata delle criticità nelle scelte compiute per simulare un intervento umano.

Il progetto di questa tesi si inserisce proprio in questo campo: a fronte di metodologie e strumenti ormai consolidati per affrontare i vari passaggi che portano dalla sequenza di amminoacidi alla costruzione della struttura tridimensionale, trovare il modo più efficace per combinarli e rendere automatiche le scelte da compiere ad ogni passo è una sfida molto più importante.

Nel corso degli anni alle categorie tradizionali si sono aggiunte numerose categorie minori:

- Predizione dei contatti: una categoria che ha avuto molto successo e che si occupa di determinare quali amminoacidi della sequenza saranno vicini in struttura, ossia in contatto tra loro.
- Predizione dei domini strutturali: data una sequenza si tratta di stabilire quali determinate parti hanno la capacità di ripiegarsi in maniera autonoma dal resto della proteina.
- Predizione del disordine: consiste nell'identificare quelle regioni delle proteine che non si ripiegano secondo le regole classiche, ovvero parti il cui ripiegamento in struttura terziaria è di difficile previsione.
- Predizione della funzione: ovvero derivare qualche informazione sulla funzione della proteina una volta identificata la sua struttura.
- Qualità dei modelli: aggiunta di recente, si pone come obiettivo quello di dare una valutazione di affidabilità in termini numerici alle predizioni effettuate.

Una questione molto importante all'interno del CASP è la scelta di uno schema di punteggio e delle metriche appropriate per confrontare modelli e struttura nativa.

Il più semplice e comune algoritmo di valutazione modello vs. nativa nella misura di similarità fra strutture proteiche è RMSD (Root Mean Square Deviation).

RMSD rappresenta la deviazione quadratica media e serve per paragonare strutture identiche, eccetto rotazioni e traslazioni. Ciò significa che se abbiamo le stesse proteine nello stesso orientamento, possiamo sovrapporle e misurare la distanza di ogni coppia di carboni alfa (ma il calcolo può essere esteso anche agli altri atomi della backbone, in genere i più pesanti quali N e C) che condividono la stessa posizione in sequenza, per determinare la loro similarità di struttura.

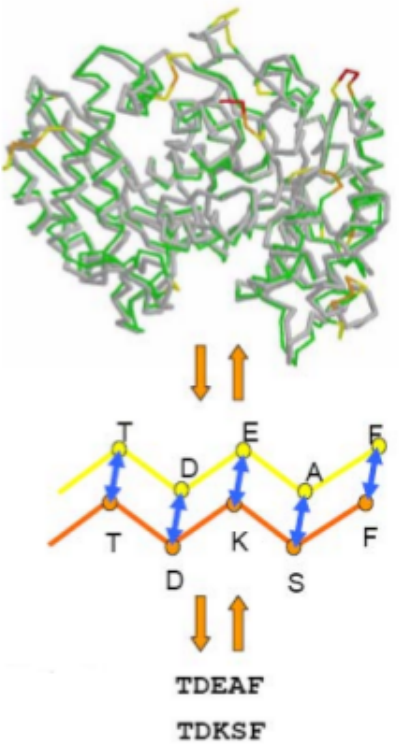

Figura 24: Sovrapposizione della struttura target e del templato

La distanza tra le due strutture è calcolata a partire dalla distanza degli atomi e si chiama *distanza euclidea*.

$$RMSD(a,b) = \sqrt{\frac{\sum(\bar{r}_{ai} - \bar{r}_{bi})}{n}} = \sqrt{\frac{\sum(distanza(a_i,b_i)^2)}{n}}$$

$$distanza(a,b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2}$$

$$RMSD(a,b) = \sqrt{\frac{\sum((a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2)}{n}}$$

dove $\bar{r}_{ai} - \bar{r}_{bi}$ sono le posizioni dell'atomo i nelle strutture a e b, ed n è il numero di tomi nelle strutture.

Quando il valore di RMSD è pari a 0 significa che le due strutture sono identiche, contrariamente, più si discosta da questo valore e più le strutture saranno differenti.

Nelle ultime edizioni del CASP si è però preferito adottare una diversa misura del grado di similarità tra modello e nativa: il GDT_TS (Global Distance Threshold Tertiary Structure).

Questo parametro misura la sovrapposizione media del modello sulla nativa, restituendo un valore compreso tra 0 (nessuna similarità) e 1 (massima similarità). Generalmente valori minori o uguali a 0.2 sono sintomi di una non corrispondenza tra modello e nativa. Questa metrica è stata sviluppata per essere una misura più accurata rispetto a RMSD, in quanto ha il vantaggio di essere

meno sensibile a errori locali come brevi regioni di loop modellate in modo errato all'interno di un modello altrimenti ragionevolmente accurato.

Il punteggio GDT descrive la massima percentuale di residui che possono essere strutturalmente allineati entro una definita soglia di distanza rispetto alla posizione nella struttura sperimentale. E' tipico nel calcolo del punteggio GDT utilizzare diverse soglie di distanza crescenti, ad esempio 1,2,4 e 8 Å, calcolando la media delle percentuali p_x di residui allineati nel seguente modo:

$$GDT - TS = (p_1 + p_2 + p_4 + p_8)/4$$

Assume valori elevati per modelli che riproducono perfettamente la conformazione della catena principale del target. Esiste anche una versione ad alta precisione della misura chiamata GDT-HA. Utilizza soglie di distanza più restrittive ed è quindi più rigorosa.

6 HOMER: web server per la modellazione comparativa

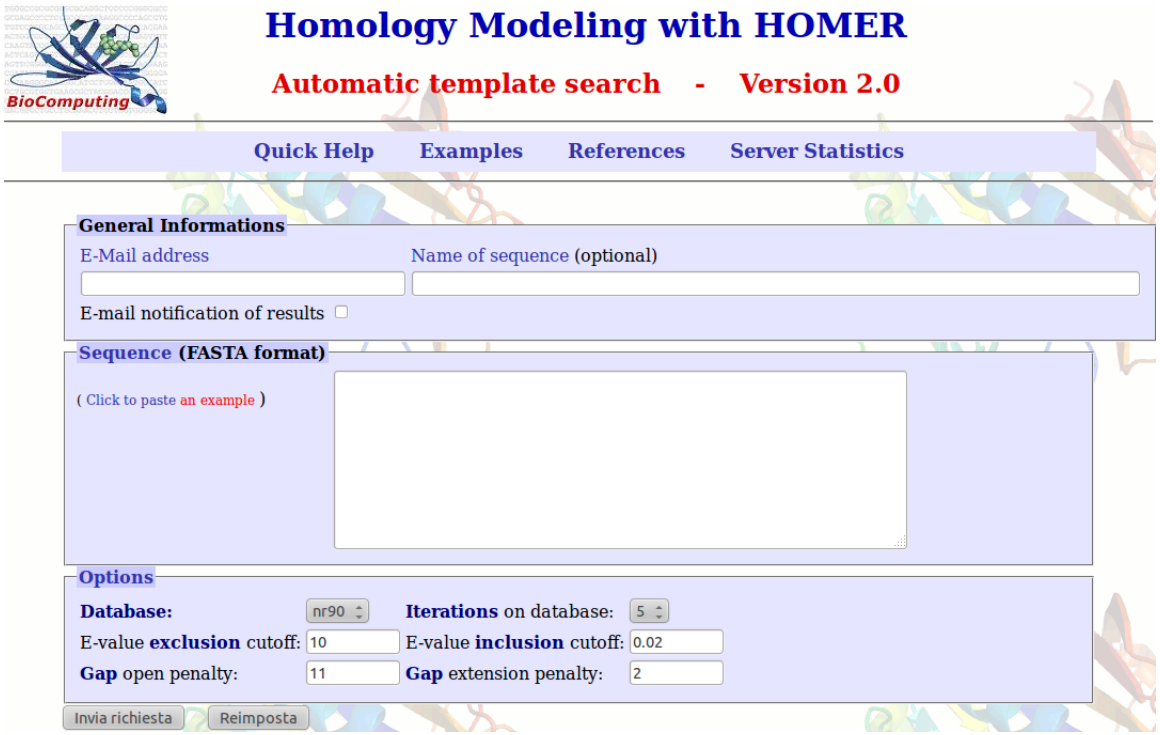

Figura 25: L'interfaccia web del servizio Homer

L'approccio al problema del protein folding può essere affrontato in molti modi.

Homer segue l'approccio del comparative modeling, i cui passi passi principali sono già stati illustrati nei capitoli precedenti (in particolare nella sezione 2.3.2). Per ognuno di questi verrà spiegato in modo dettagliato come sono stati affrontati all'interno di questo progetto e quali particolari scelte implementative sono state adottate.

La strategia generale prevede di individuare il miglior templato possibile per la sequenza target, produrre un certo numero di allineamenti alternativi tra le sequenze target e templato, usare questi allineamenti come guida nella produzione di modelli grezzi sui quali eventualmente modellare le catene laterali mancanti, trovare il miglior modello sulla base di alcune valutazioni energetiche ed eventualmente, infine, operare su quest'ultimo la modellazione dei loop.

In HOMER in particolare la ricerca del miglior templato consiste nel selezionare sempre il primo risultato fornito dalla procedura PDB-BLAST (descritta nel capitolo 8), cioè quello con il miglior e-value.

Nella produzione degli allineamenti, dopo una prima fase in cui si è cercato di identificare la miglior combinazione di scoring function, weighting scheme e gap penalty function per la produzione di allineamenti profilo contro profilo, si è preferito fornire più soluzioni alternative in modo da sfruttare meglio le potenzialità della libreria di allineamento. A favore di questa seconda soluzione vi è anche una considerazione sul tempo computazionale richiesto: poiché il calcolo dei profili e della struttura secondaria viene fatto solo una volta, produrre più tipologie di allineamento non è molto più costoso che concentrarsi su di un'unica soluzione, ed inoltre garantisce migliori risultati.

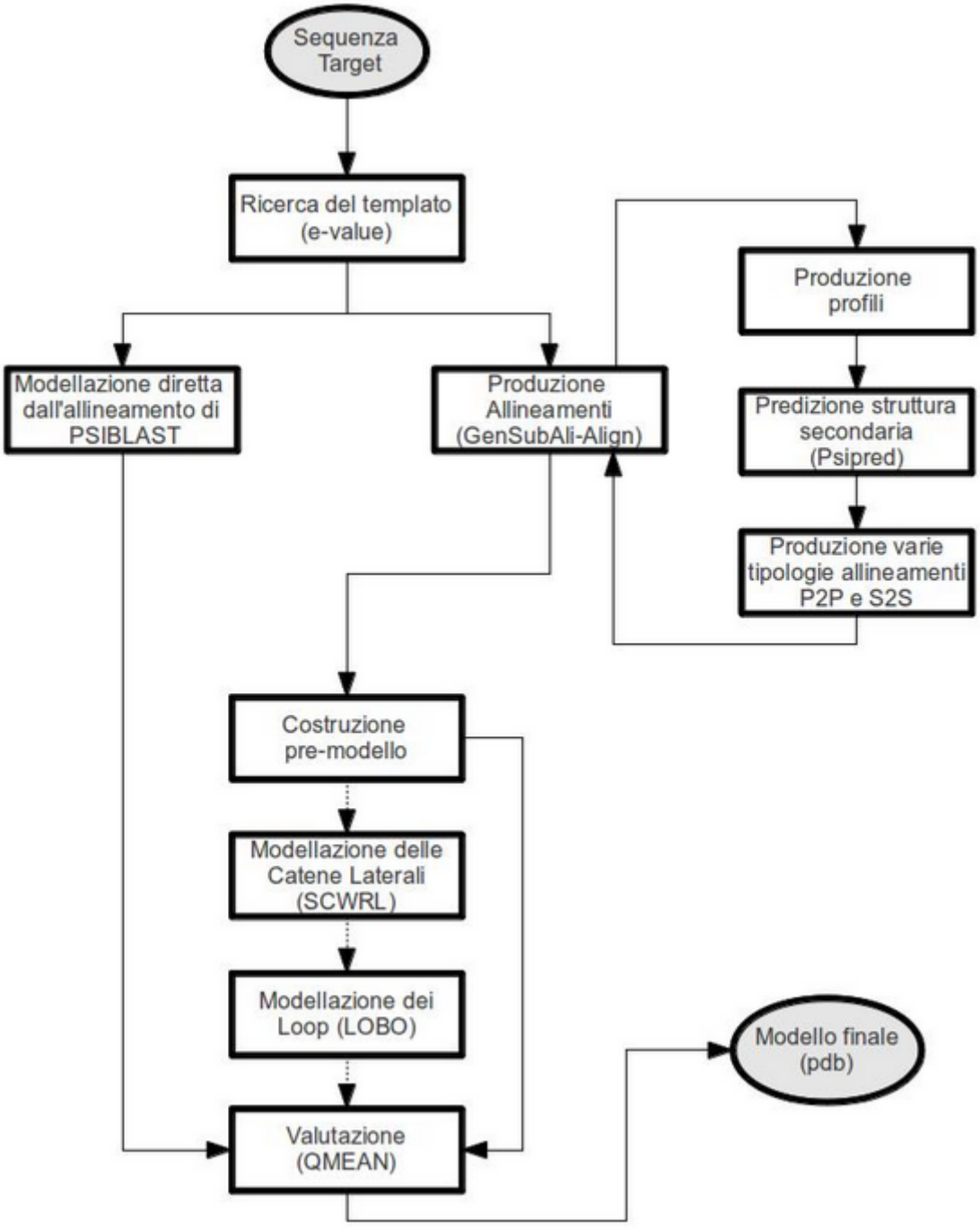

Figura 26: Schema di funzionamento di Homer

In quest'ottica, a fianco degli allineamenti profilo contro profilo che sfruttano l'algoritmo free-shift, sono stati aggiunti anche alcuni allineamenti globali sequenza contro sequenza, e anche l'allineamento prodotto da PSIBLAST nella procedura di ricerca del templato viene riutilizzato.

Va sottolineato che nello sviluppo di HOMER lo scopo prefisso non era solamente quello di ottenere uno strumento che possa rapportarsi a soluzioni che rappresentano lo stato dell'arte nel campo della predizione e modellazione di strutture proteiche, ma fondamentale era rendere il processo totalmente automatico, veloce e facile da usare.

Lo sforzo maggiore è quindi stato quello di trovare il giusto compromesso tra accuratezza e velocità, il che ha portato ad esempio alla scelta di utilizzare un solo templato dal quale ricavare più allineamenti (grazie ad approcci di programmazione dinamica) sui quali costruire più modelli per poi scegliere il migliore.

Il risultato finale del lavoro svolto è un server web che a partire da una sequenza di amminoacidi è in grado di fornire una previsione sulla sua struttura tridimensionale impiegando generalmente meno di un'ora, e affrontando in completa autonomia i vari passaggi.

Nonostante l'intervento umano non sia necessario, un utente esperto ha comunque la possibilità di intervenire a fondo sul funzionamento del programma: andando a modificare i parametri degli algoritmi di allineamento, cambiando la scelta predefinita del templato, editando manualmente gli allineamenti trovati, indicando in quali banche dati effettuare la ricerca di sequenze simili, scegliendo di modellare o meno catene laterali o loops nella costruzione del modello.

Homer inoltre non si limita alla sola possibilità per l'utente di fornire la sequenza target di cui vuole conoscere la possibile struttura; consente infatti di saltare la parte di ricerca del templato qualora questo venga indicato, o ancora di limitarsi ai soli passaggi di costruzione del modello nel caso l'utente abbia già un proprio allineamento da sottoporre.

Alcuni elementi interessanti che differenziano Homer rispetto ad analoghi servizi server, sono la possibilità di includere cofattori metallici o di altro tipo nel modello finale (fig. 27), o la capacità di modellare correttamente particolari tipologie di proteine dette omodimeri.

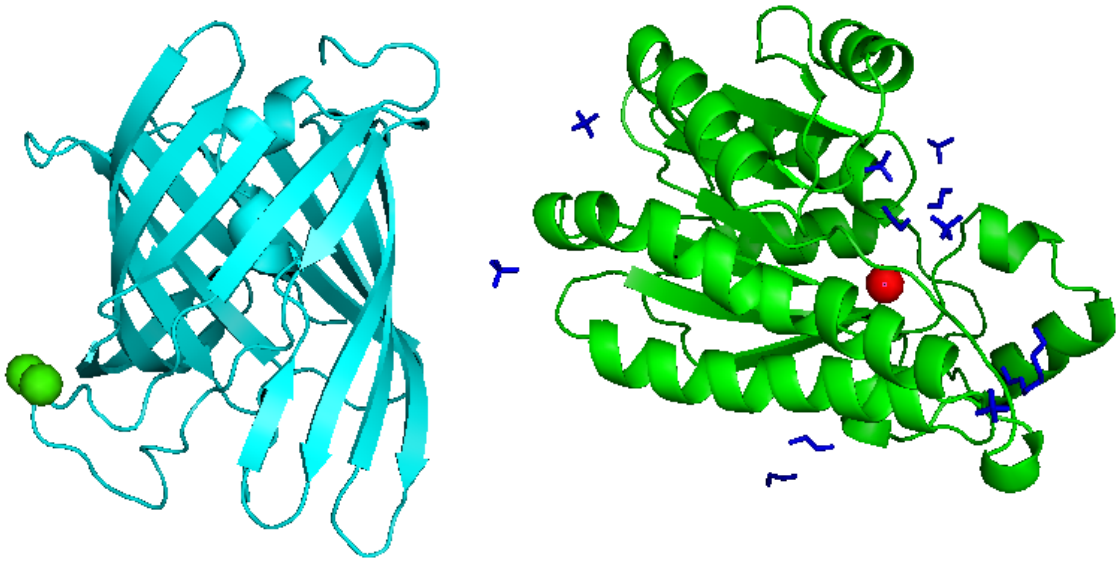

Figura 27: T0738 e T0689: i modelli dei target includono ligandi e cofattori del templato

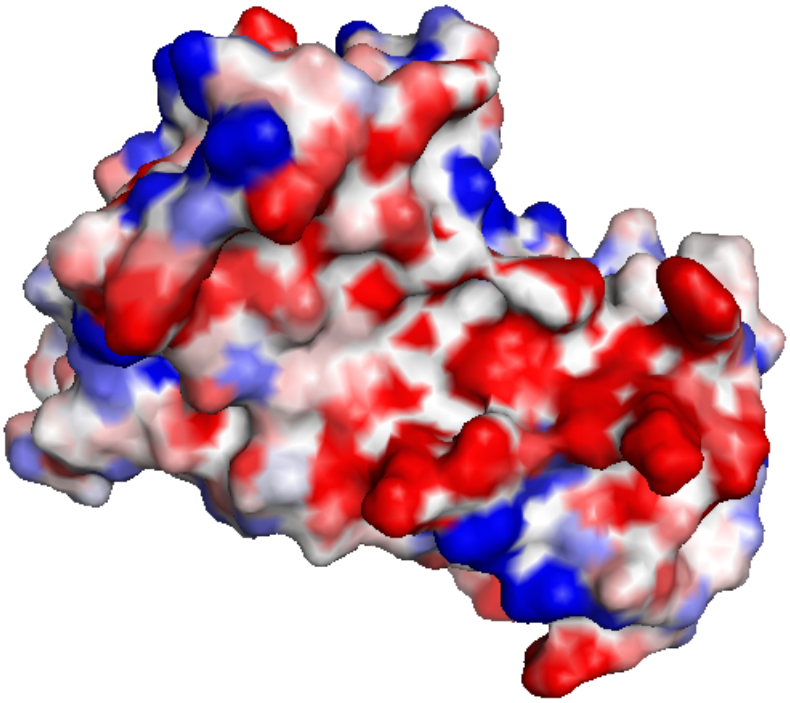

Figura 28: Visualizzazione della superficie elettrostatica di un modello prodotto con Homer

Si tratta di strutture che presentano sezioni ripetitive di identica natura chimica; pertanto mentre è difficile trovare un buon allineamento che copra una parte consistente della sequenza, trovato un buon modello per una di queste sub-unità questo può essere facilmente esteso al resto della proteina.

Homer inoltre fornisce anche un'analisi elettrostatica del modello prodotto (figura 28) utilizzando il web server Bluues nel calcolo del potenziale elettrostatico [9].

7 La libreria Biopool

La libreria Biopool2000 (Biopolimer Object Oriented Library, sviluppata a partire dal 2000) è la parte centrale del progetto HOMER.

Al suo interno sono definite tutte le classi di base per rappresentare dal punto di vista informatico una struttura proteica in tutte le sue componenti, con tutti i relativi metodi per manipolarli.

Lo scopo principale è quello di rappresentare una catena amminoacidica in modo efficace, includendo la capacità di leggere sequenze lineari di amminoacidi o di processare una struttura in formato PDB; uno dei più diffusi standard nella descrizione della struttura di una proteina di cui si è parlato nel capitolo 2.2.

Le posizioni degli atomi vengono espresse con due diverse rappresentazioni: oltre alle classiche coordinate cartesiane che ne esprimono la posizione nello spazio a tre dimensioni (rispetto ad un'origine arbitraria) proprio come avviene nei file PDB, viene utilizzato anche un sistema di coordinate interne che descrive la posizione di un atomo in termini delle sue relazioni con gli atomi posizionati in precedenza, in termini di lunghezza di legame, angolo di legame e angolo di torsione.

Il motivo della duplice rappresentazione è semplice: il sistema 3-D di coordinate cartesiane è subito comprensibile a chiunque, ed estremamente utile nel calcolo delle energie in quanto rende immediato il calcolo delle distanze tra una qualsiasi coppia di atomi, ma è d'intralcio quando occorre modificare con frequenza la struttura della proteina.

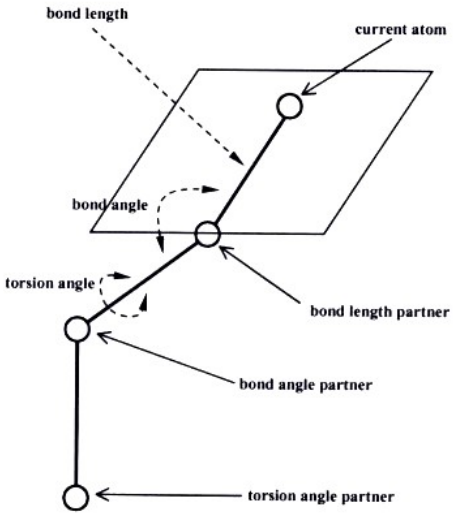

Figura 29: Il sistema di coordinate interne

La sola modifica di un angolo di torsione nella catena principale (backbone) richiederebbe infatti l'immediato ricalcolo delle posizioni di tutti gli atomi successivi.

Poiché le modifiche alla struttura dipendono strettamente da valutazioni di angoli di legame e angoli di torsione tra amminoacidi, anziché convertire di volta in volta tali informazioni in coordinate cartesiane, risulta molto più efficiente effettuare tutte le elaborazioni necessarie sulla struttura utilizzando la notazione interna, per poi ottenere solo alla fine le posizione tridimensionali attraverso una opportuna conversione.

Entità elementari quali atomi, catene laterali, amminoacidi sono rappresentati rispettivamente dalle classi Atom, SideChain, AminoAcid.

Vi sono poi un certo numero di classi astratte che hanno il compito di descrivere alcune particolari relazioni tra queste entità, quali ad esempio le classi SimpleBond e Bond che rappresentano i legami covalenti tra atomi nel primo caso, e gruppi di amminoacidi nel secondo.

Le classi AtomCode e AminoAcidCode come si può facilmente intuire definiscono i codici con cui vengono indicati atomi (es. CA, NZ) e amminoacidi, prevedendo per questi ultimi sia il simbolo convenzionale ad una lettera che il simbolo convenzionale a tre lettere (es. A o Ala per l'alanina, K o Lys per la lisina).

Le classi AminoAcid e Sidechain discendono gerarchicamente dalla classe Group in quanto rappresentano di fatto un insieme di atomi legati tra loro (relazione 1 a N).

Più in alto nella gerarchia troviamo poi la classe Monomer che serve a definire una importare proprietà compositiva nella definizione delle sottoclassi, assicurando che oggetti di tipo amminoacido e catena laterale non possano contenere ricorsivamente altri oggetti simili. Questa possibilità è invece consentita a oggetti che ereditano le proprietà della classe Polymer.

Nella versione originale della libreria il principale attore in questo ramo del diagramma era l'oggetto Spacer. Entità di questo tipo sono state concepite per rappresentare collezioni di amminoacidi legati tra loro, quali catene amminoacidiche, singoli domini o, potenzialmente, intere proteine.

Discendendo dalla classe Polymer un oggetto Spacer ha infatti la possibilità di contenere ricorsivamente altri Spacer, permettendo quindi di rappresentare con un singolo oggetto le diverse catene di una proteina.

Tale schema, che ricorda il funzionamento delle matriosche, non è stato però ritenuto sufficientemente preciso e specializzato per modellare in modo efficace e ordinato la struttura di una proteina.

La prima parte del lavoro di questa tesi è stato quindi quello di progettare e implementare un modo più efficace per rappresentare la struttura di una proteina. Un primo elemento di forte novità è l'introduzione di una nuova classe elementare: la classe Ligand.

Questa nuova classe, che condivide con AminoAcid la proprietà di essere sostanzialmente un gruppo di atomi, serve ad introdurre nel processo di predizione l'informazione riguardante tutto ciò che non è proteina, e che è presente nell'informazione sperimentale.

In particolare si è deciso di distinguere tra tre categorie: co-fattori metallici, co-fattori di altro tipo, acqua.

La presenza di tali elementi, specialmente nel caso degli ioni metallici, influenza il funzionamento della proteina ed è quindi utile includerli nel processo di modellazione. Inoltre la loro presenza aiuta a produrre modelli più precisi permettendo di tener conto del loro ingombro in particolare nella fase di piazzamento delle catene laterali.

Altro importante cambiamento è la radicale ridefinizione del ruolo di uno Spacer: da contenitore pluripotente a contenitore di una singola catena di amminoacidi.

In modo del tutto simile è stato poi definito anche un contenitore per ligandi che appartengono ad una stessa catena: la classe LigandSet.

Per rappresentare l'intera struttura di una proteina è stata infine aggiunta una ulteriore classe denominata Protein. Un oggetto di questa classe ha un struttura interna decisamente più ordinata rispetto all'originale approccio tramite Spacer.

Al suo interno troviamo tanti oggetti Polymer quante sono le catene da modellare, associati in modo biunivoco grazie all'aggiunta di una opportuna proprietà.

Trattandosi di una classe astratta, un Polymer ha a sua volta un ruolo di semplice contenitore: al suo interno ci sarà sempre lo Spacer che organizza le informazioni della particolare catena associata al Polymer, ed eventualmente il LigandSet (a seconda delle richieste dell'utente che può essere o meno interessato a queste informazioni, o al fatto che la proteina presenti effettivamente o meno dei cofattori per la data catena).

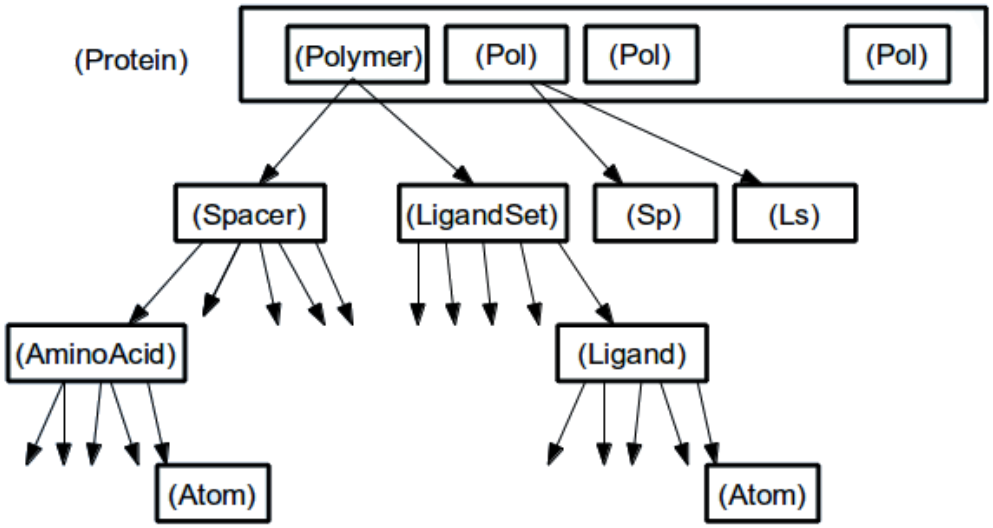

Figura 30: Schema della classe Protein

Ulteriori modifiche hanno poi riguardato le classi PdbLoader e PdbSaver, per ampliare la capacità di gestire il formato standard PDB da e verso la nuova classe Protein.

8 Ricerca del templato

L'individuazione del templato è uno passaggio molto delicato nell'ambito della costruzione di un modello per una proteina.

Errori commessi in questa fase non possono essere corretti successivamente e pregiudicano quindi in maniera irrecuperabile il risultato finale della predizione portando alla costruzione di strutture spesso anche del tutto sbagliate.

Il templato nella maggior parte dei casi è una proteina omologa al target ed in genere tanto minore è la distanza evolutiva e tanto maggiore è la probabilità che le due strutture siano simili (e quindi sovrapponibili).

Il problema nel comparative modeling è quindi quello di trovare un omologo di cui sia nota la struttura e i cui dati di allineamento con il target siano significativi.

Usare un tool di ricerca su database allo stato dell'arte riduce significativamente la possibilità di avere dei falsi positivi.

Per questa tesi si è scelto di fare ricorso al protocollo PDBBLAST [10]. L'implementazione di tale protocollo consiste nel cercare di collezionare quanta più informazione possibile sulla famiglia proteica a cui la sequenza target appartiene, al fine di migliorare la ricerca di una struttura omologa.

PDBBLAST prevede due passi:

1. nel primo PSI-BLAST è utilizzato nella ricerca di sequenze omologhe in un database non ridondante NR (NCBI non redundant)[10]. Il database NR contiene tutte le sequenze proteiche pubblicamente disponibili conosciute fornite dalle principali banche dati (Swiss-Prot, GenBank etc.). PSI-BLAST è un'estensione del metodo BLAST per la ricerca contro una banca dati di sequenze e sfrutta l'idea di utilizzare un profilo si frequenza. E' un allineamento progressivo che tiene conto della traccia evolutiva di una sequenza e quindi di come questa può variare senza che si modifichi la struttura e la funzione della proteina. Utilizza una procedura iterativa per cui tutte le sequenze che superano la soglia minima imposta di similarità partecipano alla creazione di un modello detto PSSM (Position Specific Scoring Matrix, o matrice di peso) utilizzata nei cicli successivi per cercare sequenze evolutivamente più distanti rispetto a quelle che erano state trovate al passo precedente. La PSSM è il "prodotto" della matrice di sostituzione (come BLOSUM o PAM: amminoacidi simili sono trattati in modo diverso da amminoacidi non simili) con la matrice di frequenza calcolata dagli allineamenti della sequenza query contro gli hit che hanno superato la soglia imposta (profilo). Dopo la prima fase, che avviene come un BLAST normale, la ricerca di nuovi hit prosegue utilizzando la matrice PSSM al posto delle generiche matrici di sostituzione 20x20. In questo modo i valori sono specifici per ogni posizione dell'allineamento. Con PSI_BLAST migliorano i livelli di affidabilità, l'e-value[11] è molto più significativo e allo stesso tempo vengono identificate molte più sequenze che non quelle considerate soltanto da BLAST. Il punto centrale è che la

[10] Nel nostro caso la scelta è ricaduta su NR90: derivata da NR clusterizzando ad una soglia del 90% di identità di sequenza.

[11] L'e-value valuta la bontà di un allineamento indicando la probabilità di avere lo stesso punteggio effettuando un allineamento con una sequenza casuale. In generale per i database proteici un e-value viene considerato significativo quando è minore di 10^{-6}.

matrice PSSM va a rappresentare sempre meglio la variabilità di quella data famiglia proteica, e questa procedura viene ripetuta finché non si raggiunge un numero prestabilito di iterazioni, tipicamente 4 o 5[12], o finché la ricerca non va a convergenza. In genere non si compiono mai più di sei iterazioni per evitare fenomeni di "deriva": la sequenza iniziale si può perdere durante le iterazioni se esiste una seconda famiglia proteica numerosa simile.

2. Nel secondo step la matrice PSSM generata è usata per effettuare una ricerca sul database di strutture PDB. In questo modo le sequenze trovate saranno quasi certamente associate ad una struttura e successivamente su questi possibili templati si possono effettuare delle considerazioni di affidabilità. Nel nostro caso ci limitiamo ad utilizzare un solo templato, selezionando sempre quello che presenta il miglior e-value, ma è bene considerare anche il rapporto tra dimensione della regione allineata e la percentuale di identità.

Poiché l'identificazione di un buon templato è una condizione fondamentale ai fini del corretto svolgimento del processo di previsione della struttura del target, HOMER prevede un output intermedio alla fine di questo step.

L'utente può quindi prendere visione della struttura scelta e di alcune importanti indicazioni sulla sua qualità: l'e-value, il bit score, l'allineamento prodotto da PSIBLAST nella fase di selezione di quel determinato templato, la presenza di eventuali regioni disordinate.

L'utente, se ne ha le competenze, a questo punto può eventualmente decidere di selezionare un altro templato, oppure di procedere scegliendo tra le varie tipologie di allineamento offerte. Come opzione predefinita si ricorre ad un allineamento profilo contro profilo, integrando anche informazioni sulla struttura secondaria.

[12]Non essendosi rivelato eccessivamente pesante del punto di vista del tempo computazionale, nel nostro caso si è scelto di eseguirlo per 5 iterazioni.

Homology Modeling with HOMER

Automatic template search

Title: t06504round
emailaddress: francesco.lovo@gmail.com
pid: 994519590

Status: finished

***NB:** The data will be kept for **two weeks**, after which it will be **deleted** without further notice.*

Available files:

Input parameters: HTML

PSI-BLAST best Template with annotations.

Sequence.

Sequence alignments with links to National Center for Biotechnology Information here

Function.

Gene Ontology information here

Structure.

Local qualities.

PDB ID:106SA
Resolution (-10 for NMR):-10
Local sequence identity:0.374631268436578
Score(bits):277
E-value:6e-89,

Global qualities.
PSI-BLAST sequence identity:0.367052023121387
Homer sequence identity (127/346): 0.3670520231213873
Query length:346
Disorder (0/346): 0.0
DSSP Undetermined residue (usually marked with 'X' and can be considered modified): (0/346): 0.0
PDB numbering breaks: 0
If the Homer and PSI-BLAST sequence identies are very different there may be a alignment shift due to disorder.
The alignment and model might be incorrect.
Given 0 'X'. This should be true to proceed:
(Homer sequence identity + disorder) = (PSI-BLAST sequence identity).
(0.3670520231213873+0.0) = 0.367052023121387
(0.3670520231213873) = 0.367052023121387

```
Residue numbering   :                 20                40                60                80                100
PDB numbering       :                 97                117               137               157               177
Full Query          :AATLATLPAPINQIFPDADLAEGIRAVLQKASVTDVVTQEELESITKLVVAGEKVASIQGIEYLTNLEYLNLNGNQITDISPLSNLVKLTNLYIGTNKITDISALQNLTNLRELY
Aligned Query       :..TLATLPAPINQIFPDADLAEGIRAVLQKASVTDVVTQEELESITKLVVAGEKVASIQGIEYLTNLEYLNLNGNQITDISPLSNLVKLTNLYIGTNKITDISALQNLTNLRELY
Aligned template    :..TLQADRLGIKSIDGVEYLNNLTQINFSNNQLTDITPLKNLTKLVDILMNNNQIADITPLANLTNLTGLTLFNNQITDIDPLKNLTNLNRLELSSNTISDISALSGLTSLQQLS
Aligned 2nd struct. :...EEECCSSCCCCCTTGGGCTTCCEEECCSSCCCCCGGGTTCTTCCEEECCSSCCCCCGGGTTCTTCSEEECCSSCCCCCGGGTTCTTCSEEEEEEEECCCGGGTTCTTCCEEE
Residue numbering   :    120               140               160               180               200               220
PDB numbering       :    196               216               235               255               275               295
Full Query          :LNEDNISDISPLANLTKMYSLNLGANHNLSDLSPLSNMTGLNYLTVTESKVKDVTPIANLTDLYSLSLNYNQIEDISPLASLTSLHYFTAYVNQITDITPVANMTRLNSLKIGN
Aligned Query       :LNEDNISDISPLANLTKMYSLNLGANHNLSDLSPLSNMTGLNYLTVTESKVKDVTPIANLTDLYSLSLNYNQIEDISPLASLTSLHYFTAYVNQITDITPVANMTRLNSLKIGN
Aligned template    :F.GNQVTDLKPLANLTTLERLDISSNK.VSDISVLAKLTNLESLIATNNQISDITPLGILTNLDELSLNGNQLKDIGTLASLTNLTDLDLANNQISNLAPLSGLTKLTELKLGA
Aligned 2nd struct. :E.EESCCCCGGGTTCTTCCEEECTTSC.CCCCGGGGGCTTCCEEECCSSCCCCCGGGGGCTTCCEEECCSSCCCCCGGGGGCTTCCEEECCSSCCCCCGGGTTCTTCSEEECCS
Residue numbering   :         240               260               280               300               320               340
PDB numbering       :         315               335               355               .                 393               413
Full Query          :NKITDLSPLANLSQLTWLEIGTNQISDINAVKDLTKLKMLNVGSNQISDISVLNNLSQLNSLFLNNNQLGNEDMEVIGGLTNLTTLFLSQNHITDIRPLASLSKMDSADFANQV
Aligned Query       :NKITDLSPLANLSQLTWLEIGTNQISDINAVKDLTKLKMLNVGSNQISDISVLNNLSQLNSLFLNNNQLGNEDMEVIGGLTNLTTLFLSQNHITDIRPLASLSKMDSADFAN..
Aligned template    :NQISNISPLAGLTALTNLELNENQLEDISPISNLKNLTYLTLYFNNISDISPVSSLTKLQRLFFYNNKVS..DVSSLANLTNINWLSAGHNQISDLTPLANLTRITQLGLND..
Aligned 2nd struct. :SCCCCCGGGTTCTTCSEEECCSSCCSCCGGGGGCTTCSEEECCSSCCCCCGGGGGCTTCCEEECCSSCCC..CCGGGGGCTTCCEEECCSSCCCBCGGGTTCTTCCEEECCC..
Residue numbering   :
PDB numbering       :
Full Query          :IKK
Aligned Query       :...
Aligned template    :...
Aligned 2nd struct. :...
```

Key : **red**: matched residue, **green**: other aligned residues, **blue**: gap in template and **purple**: disorder (removed from model)

Figura 31: Output intermedio dopo la fase di ricerca del templato (A)

Figura 32: Output intermedio dopo la fase di ricerca del templato (B)

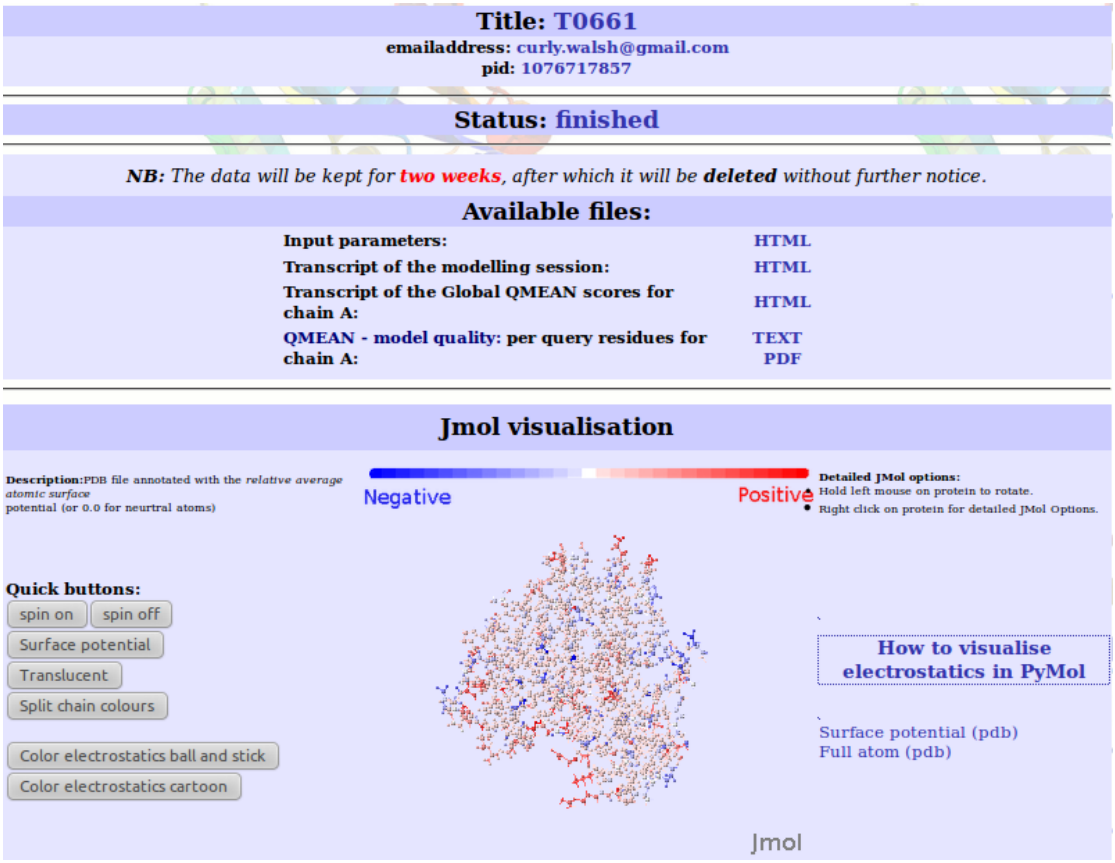

Figura 33: Output finale

9 GenSubAli: Allineamento

Il secondo passo nell'ambito della costruzione di un modello proteico è l'allineamento fra target e templato.

Anche in questo caso si tratta di un passaggio fondamentale e molto delicato per la bontà del risultato finale, in quanto si tratta di allineare target e templato da un punto di vista strutturale.

Il problema maggiore è che l'allineamento strutturale di cui si avrebbe bisogno può solo essere approssimato usando un allineamento fra le due sequenze. Una immediata conseguenza è che il miglior allineamento tra sequenze non corrisponde necessariamente al miglior allineamento strutturale.

Da un punto di vista teorico i protocolli di allineamento seguiti in questa fase dovrebbero essere più rigorosi rispetto a quelli utilizzati nell'ambito della ricerca in banca dati. Nel secondo caso si vuole solo identificare il templato, mentre nel primo si vogliono identificare le regioni strutturalmente simili.

In realtà le due fasi spesso possono sovrapporsi e usare gli stessi metodi.

In parte questo è vero anche per la particolare implementazione seguita in questa tesi, nella quale il ricorso ad allineamenti profilo-profilo prevede l'utilizzo del protocollo PSI_BLAST già descritto al capitolo precedente.

Vedremo inoltre che si farà uso anche di informazioni strutturali del templato combinate da predizioni strutturali del target ricavabili direttamente dalla sequenza, e di come si seguirà l'attuale tendenza a privilegiare la produzione di allineamenti alternativi che poi vengono valutati in termini energetici sui modelli prodotti.

9.1 Concetti fondamentali nell'allineamento P2P

Allineamenti multipli L'informazione biologica contenuta in un allineamento multiplo è certamente superiore a quella di tutti i possibili allineamenti a coppie. In un allineamento semplice si possono osservare posizioni più o meno conservate, ma non si ha alcuna indicazione circa la rilevanza funzionale di questi residui. Tale definizione può essere dedotta dall'osservazione di un allineamento multiplo, dove i residui più importanti dal punto di vista funzionale risultano fortemente conservati fra tutte le sequenze dell'allineamento.

profilo di un allineamento multiplo I dati biologici racchiusi in un allineamento multiplo possono essere riportati nel suo **profilo** che attribuisce a ciascuna colonna dell'allineamento il punteggio relativo ad ognuno dei venti amminoacidi proteici. Similmente, soppesando la frequenza di gap nelle differenti colonne, è possibile accordare una diversa penalità per l'inserimento di gap in funzione della maggiore o minore propensione ad accettarli. Il profilo di un allineamento multiplo è rappresentato da una tabella dove le righe sono pari al numero degli amminoacidi proteici (più eventuali due ulteriori righe per coefficienti posizione-specifici relativi all'inserimento e all'estensione di gap) e le colonne alla lunghezza dell'allineamento. Il valore complessivo nella cella (i,j) corrisponde alla frequenza di ricorrenza normalizzata dell'i-esimo amminoacido nella j-esima colonna dell'allineamento multiplo. La ricerca in banca dati di sequenze omologhe molto divergenti è fra le applicazioni preponderanti dei pro-

fili. Una ricerca iterativa con i profili è implementata dal programma PSI-BLAST in modo del tutto analogo a quanto avviene nella fase di ricerca del templato.

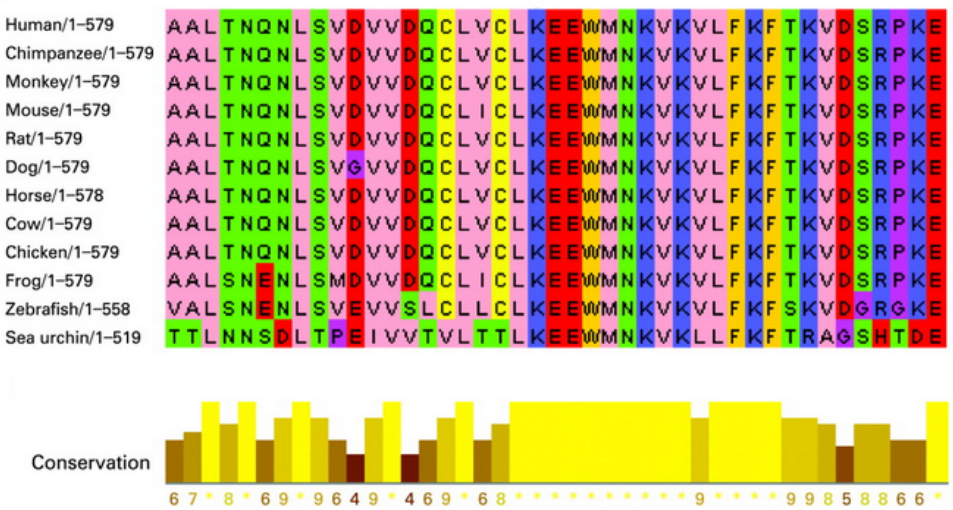

Figura 34: Esempio di allineamento multiplo

Allineamenti profilo contro profilo Il confronto fra profili è una tecnica molto potente per la creazione di allineamenti accurati e rappresenta lo stato dell'arte in materia di allineamenti proteici [11]. Elaborando le informazioni biologiche racchiuse nei profili costruiti a partire dalle sequenze di input, è possibile superare (in termini di performance) i tradizionali allineamenti sequenza contro sequenza e profilo contro sequenza. La procedura di allineamento profilo contro profilo richiede di:

- scegliere quali sequenze includere nell'allineamento multiplo;
- definire uno schema per assegnare un peso alle sequenze contenute nell'allineamento multiplo;
- definire un metodo per calcolare le frequenze degli amminoacidi a partire dall'allineamento pesato;
- definire una funzione per assegnare un punteggio agli abbinamenti fra coppie di colonne dei due profili.
- Definire una funzione di penalizzazione dei gap;
- decidere se e come aggiungere eventuali informazioni strutturali (nell'ambito di questa tesi si è scelto di aggiungere informazioni tratte dalla previsione di struttura secondaria).

Di seguito verrà illustrato come come tale procedura è stata implementata nel programma sviluppato.

9.2 GenSubAli e la libreria Align

Nella produzione di allineamenti a partire dal templato, si è fatto ricorso ad una libreria C++ sviluppata internamente al gruppo di lavoro presso cui è stata svolta la presente tesi: ALIGN [12,13].

Align implementa diverse tecniche di allineamento:

- permette di effettuare allineamenti sequenza contro sequenza (S2S), profilo contro sequenza (P2S) e profilo contro profilo (P2P).
- Implementa tutti i principali algoritmi di allineamento (3): globali, locali e semi-globali o freeshift.
- Utilizza diverse tecniche per la penalizzazione dei gap lineari, affini e variabili.
- Utilizza tre differenti weighting schemes nella costruzioni dei profili,
- Utilizza undici differenti scoring functions negli allineamenti profilo contro profilo.
- Utilizzare informazioni strutturali di varia natura.

Nel corso del capitolo si limiterà la descrizione alle classi e ai metodi più importanti evidenziando i legami fra i blocchi della libreria.

Ai fini dell'automazione dell'intero processo di produzione degli allineamenti, l'uso di tale libreria è stato affidato ad uno script perl appositamente realizzato: GenSubAli.

Tale script si occupa di implementare l'intera strategia di allineamento sviluppata per questo progetto, ed eventualmente di modificarla qualora un utente esperto preferisca definire un proprio workflow interagendo con Homer tramite l'apposita interfaccia web.

In ogni caso la strategia generale delineata è quella di ricorrere ad allineamenti profilo contro profilo calcolati tramite l'uso dell'algoritmo freeshift, arricchiti dalle informazioni sulle previsioni di struttura secondaria.

GenSubAli si occupa dunque tramite PSI_BLAST della produzione dei profili a partire dalle due sequenze input target e templato, utilizzando PSI-PRED[13] nella produzione della struttura secondaria [14].

Poiché circa il 50% della catena polipeptidica è strutturato in α-eliche o in filamenti-β, la predizione della struttura secondaria facilita la conferma di relazioni strutturali o funzionali tra fra proteine con bassa similarità di sequenza, contribuendo quindi a migliorare con queste informazioni la ricerca dell'allineamento ottimale fra sequenze nella modellazione comparativa, e quindi la predizione della struttura terziaria attraverso la giustapposizione degli elementi di struttura secondaria.

[13]PSIPRED utilizza un sistema di reti neurali e si basa sulla PSSM generata da PSI-BLAST sulla sequenza di input. La procedura è divisa in tre passi:

- la costruzione del profilo di sequenze
- la predizione della struttura secondaria operata da una prima rete neurale
- la predizione di struttura secondaria ottenuta filtrando il primo output con una seconda rete neurale.

L'affidabilità media è stimata intorno all'80%.

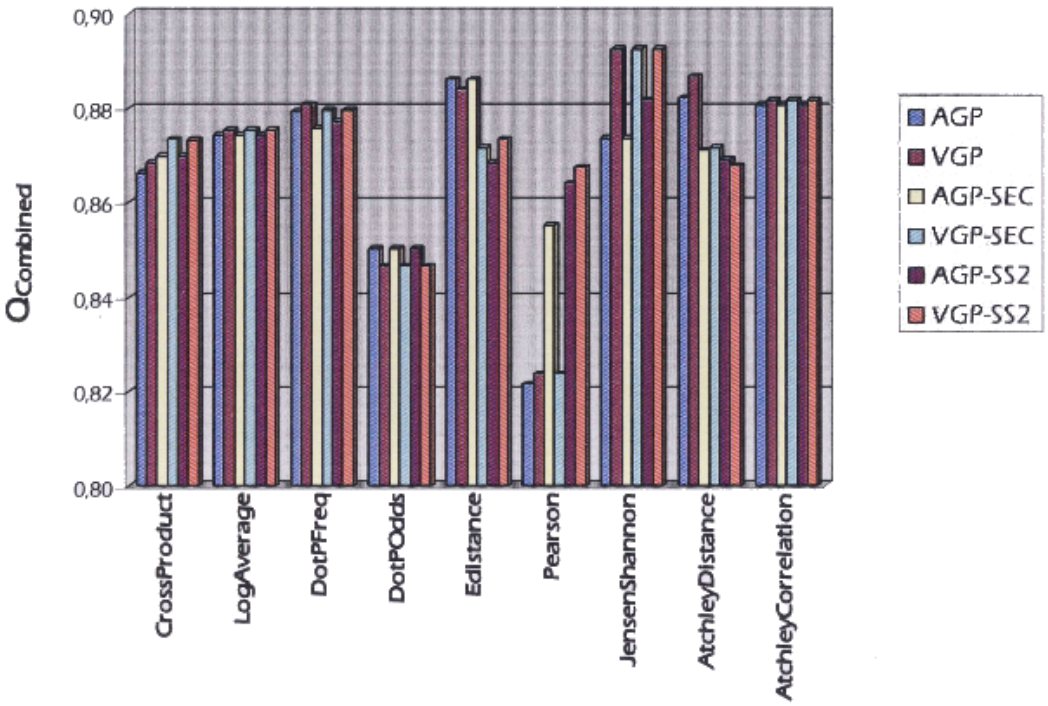

Figura 35: Performance attese dalle varie scoring e gap-penalty function, in base ad un precedente lavoro di analisi

A fronte dell'elevato numero di possibili combinazioni offerto dalla libreria Align, la strategia di utilizzo adottata si è basata sul lavoro di analisi svolto da altri colleghi che hanno collaborato all'estensione della libreria introducendo proprio la possibilità di effettuare allineamenti P2P [13].

In base ai risultati, in GenSubAli un primo approccio è stato quello di identificare il migliore metodo di allineamento disponibile: si è deciso di accostare lo schema di peso Henikoff, la scoring function Jensen Shannon e la funzione VGP di penalizzazione variabile dei gap con utilizzo dell'informazione di struttura secondaria del templato.

Ognuna di queste opzioni selezionate verrà presentata nel dettaglio, mentre per le altre opzioni fornite dalla libreria Align verrà fatta solo una breve presentazione, demandando ulteriori approfondimenti alla tesi sopra citata. Una successivo rivisitazione della procedura ha poi portato ad affiancare agli allineamenti prodotti con questa strategia, alcuni allineamenti rappresentativi di tutte le altre varie combinazioni possibili.

Oltre ad individuare l'approccio migliore all'utilizzo di tale libreria, il lavoro di tesi ha richiesto anche una profonda revisione in particolare degli schemi di peso utilizzati nella produzione dei profili.

A causa di implementazioni troppo naive, in particolare nello schema Henikoff [15], il tempo computazionale di questo singolo passaggio poteva richiedere fino ad un paio d'ore, rappresentando una criticità di fatto inaccettabile nell'adozione di questa soluzione in quello che nelle intenzioni vuole essere un tool automatico veloce ed efficiente.

La libreria Align, implementando tecniche di programmazione dinamica, può produrre un numero variabile di possibili allineamenti.

L'utente ha comunque la facoltà di indicare qual'è il massimo numero di allineamenti di sequenza tra target e templato a cui è interessato; numero che può variare da 1 a 500, che è il numero massimo di sequenze che si è scelto di utilizzare per questo progetto nella fase di produzione dei profili. Tale parametro va ad influire sul risultato di PSI-BLAST e, a differenza del numero massimo di allineamenti, si è scelto di renderlo del tutto trasparente all'utente.

In genere tuttavia il numero di allineamenti utili non è molto alto: come scelta di default tale parametro è impostato a 25 per la strategia principale, a cui si aggiungono altri 20-25 allineamenti rappresentativi delle altre combinazioni.

GenSubAli inoltre, prima di fornire tali allineamenti alla successiva parte del programma che si occuperà di costruire su questi dei modelli, controlla che tra questi non vi siano dei duplicati, andando quindi a filtrare l'insieme prodotto.

In questo modo si evita di andare a produrre più volte lo stesso modello, con conseguente spreco di tempo computazionale.

9.3 Struttura della libreria Align

9.3.1 AlignmentData

La classe astratta AlignmentData gestisce la ricostruzione dell'allineamento trovato.

In particolare sono state implementate due differenti classi derivate:

- SequenceData, per gli allineamenti che non tengono conto della struttura secondaria,
- SecSequenceData, per gli allineamenti che tengono conto della struttura secondaria. Sarà quindi questa la classe utilizzata nella produzione di allineamenti in HOMER.

9.3.2 GapFunction

La classe GapFunction è un'altra classe astratta con il compito di gestire il calcolo delle penalità per l'introduzione e l'estensione di gap nell'allineamento.

Per consentire il confronto fra più funzioni di penalizzazione, sono state implementate diverse classi derivate:

- AGPFunction per impiegare una funzione lineare[14]o affine[15].
- VGPFunction (utilizzato in questa tesi) per ricorrere ad un approccio più avanzato che modula le penalità in virtù di specifiche caratteristiche strutturali. In particolare, è possibile modificare la penalità di apertura valutando la configurazione strutturale osservata in prossimità dell'inserzione/delezione:

$$O(i) = o + (W_H \cdot H_i + W_s \cdot S_i + W_B \cdot B_i + W_C \cdot C_i + W_D \cdot P_i)$$

[14]Impostando penalità di apertura ed estensione allo stesso valore costante.

[15]Una funzione AGP (affine Gap Penalty) potrebbe diminuire progressivamente il valore della penalità con l'estensione del gap.

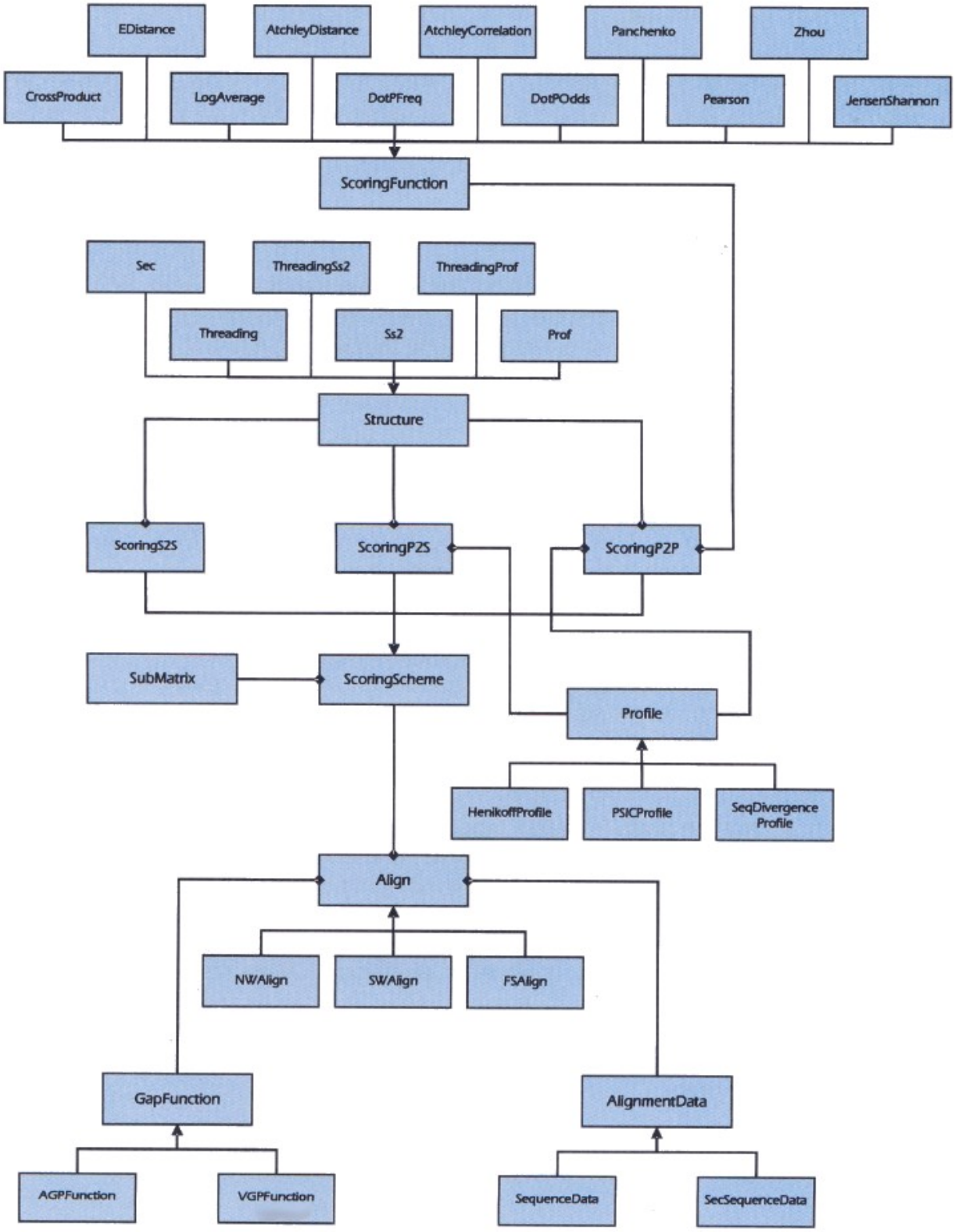

Figura 36: Schema della libreria Align

Il valore restituito è almeno o, ma può aumentare all'interno specialmente se ricade all'interno di eliche o strutture β. I parametri H,S,B e C variano tra 0 e 1. H_iè il parametro di propensione conformazionale per l'alfa elica, S_i è una grandezza analoga per le strutture β. B_i è il coefficiente di sepoltura dell'i-esimo residuo. C_i è la linearità del backbone proteico per l'i-esimo residuo. Infine le variabili W_i sono i pesi dei cinque parametri. Annullando tutti i pesi la funzione VGP si riduce ad una tradizionale funzione AGP.

9.3.3 Profile

La classe Profile gestisce la costruzione di un profilo a partire da un allineamento multiplo.

Nel caso più semplice, il profilo viene generato senza assegnare un particolare peso alle sequenze contenute nell'allineamento multiplo.

Per avvalersi di particolari weighting schemes, sono state implementate tre classi derivate:

- Henikoff Profile per impiegare lo schema Henikoff [15].
- PSICProfile, per impiegare lo schema PSIC [16].
- SeqDivergenceProfile, per impiegare lo schema SeqDivengence.

Come già anticipato, la scelta in questo caso è ricaduta sullo schema Henikoff, il quale:

- determina, per l'm-esimo amminoacido dell'i-esima sequenza, il sottoinsieme delle sequenze che hanno un amminoacido nella stessa colonna dell'allineamento multiplo, indipendentemente da dove queste inizino o terminino rispetto alle altre.
- individua la prima colonna in cui tutte le sequenze del sottoinsieme sono rappresentate da un amminoacido o un gap interno, denotando questa colonna come C_{left}.
- in modo analogo, individua l'ultima colonna in cui tutte le sequenze del sottoinsieme sono rappresentate da un amminoacido o da un gap interno, denotando questa colonna come C_{right}.

Il peso dell'm-esimo amminoacido dell'i-esima sequenza è dato da:

$$W_i^m = \frac{1}{C_{right} - C_{left} + 1} \cdot \sum_{j=C_{left},C_{right}} \frac{1}{N_{diff}^j \cdot n_i^j}$$

dove N_{diff}^j è il numero di amminoacidi differenti nella j-esima colonna che fanno parte del sottoinsieme dell'allineamento multiplo trovato al punto precedente, e n_i^j è il numero di ricorrenze del j-esimo amminoacido dell'i-esima sequenza nella j-esima colonna nello stesso sottoinsieme dell'allineamento multiplo.

9.3.4 ScoringFunction

La classe astratta scoring function gestisce il calcolo dei punteggi per gli abbinamenti fra le colonne dei due profili.

Per valutare l'efficacia di numerose scoring functions, nella libreria Align sono state implementate varie classi derivate che seguono diversi approcci:

- CrossProduct, per impiegare la somma dei prodotti delle frequenze per tutte le combinazioni degli amminoacidi (log-odds della matrice BLOSUM).
- LogAverage, per impiegare il logaritmo della somma dei prodotti delle frequenze per tutte le combinazioni degli amminoacidi (frequenze della matrice BLOSUM).
- DotFreq, per impiegare il prodotto scalare (frequenze).
- DotPOdds, per impiegare il prodotto scalare (log-odds).
- Edistance, per impiegare la distanza euclidea.
- Pearson, per impiegare il coefficiente di correlazione di Pearson.
- JensenShannon, per impiegare l'omonima funzione.
- AtchleyCorrelation, per impiegare il coefficiente di correlazione di Pearson sui fattori delle metriche di Atchley.
- Panchenko, per impiegare l'omonima funzione.
- Zhou, per impiegare l'omonima funzione.

In base ai dati forniti in [13], in HOMER si è deciso di adottare la funzione LogAverage, sviluppata da Von Ohsen e Zimmer nel 2003.

Tale funzione moltiplica i prodotti delle frequenze Q_a^1 e Q_b^2 di una coppia di colonne, per le rispettive frequenze q_{ab} della matrice di sostituzione BLOSUM62 e restituisce il logaritmo naturale del punteggio finale:

$$S_{1,2} = ln\left(\sum_{a=1}^{20}\sum_{b=1}^{20} Q_a^1 \cdot Q_b^2 \cdot q_{ab}\right)$$

9.3.5 Structure

La classe astratta Structure gestisce le eventuali informazioni strutturali sulle sequenze di input.

Per servirsi di molteplici informazioni strutturali, sono state implementate le classi derivate:

- Sec, per impiegare informazioni sulla struttura secondaria nel formato FASTA[16] (predizione a 3 stati conformazionali). Il punteggio S_{ij} per una coppia di posizioni è stabilito grazie

[16]Derivabile dal formato HORIZ oSS2 di PSI-PRED, soluzione adottata nel presente lavoro di tesi.

ad una matrice di sostituzione 3x3 simile alle matrici utilizzate per la similarità fra residui amminoacidici:

$$S_{sse(i,j)} = s(state_{target(i)}, state_{templato(j)})$$

e quindi moltiplicato per un opportuno coefficiente cSec.

$$S_{i,j} = cSec \cdot S_{sse(i,j)}$$

Questa per inciso è anche la soluzione adottata in questo progetto.

- Threading, per impiegare informazioni di threading della sequenza templato[17].
- Ss2, per impiegare informazioni sulla struttura secondaria nel formato SS2 di PSI-PRED.
- Prof, per impiegare informazioni sulla struttura secondaria e sull'accessibilità al solvente nel formato PROF.
- ThreadingSs2, per impiegare contemporaneamente le informazioni strutturali di Threading e di Ss2.
- ThreadingProf, per impiegare contemporaneamente le informazioni strutturali di Threading e Prof.

9.3.6 ScoringScheme

La classe astratta ScoringScheme gestisce lo schema di punteggio dell'allineamento.

Il metodo scoring, comune a tutte le sottoclassi, restituisce il punteggio (compresa la componente dovuta alle informazioni strutturali) per una coppia di posizioni.

Tra le varie classi derivate l'implementazione dello schema di punteggio è strettamente correlata al tipo di allineamento usato.

Si distinguono pertanto tre diverse classi:

- ScoringS2S, per gli allineamenti sequenza contro sequenza. Il punteggio S_{ij} per una coppia di posizioni è stabilito grazie ad una matrice di sostituzione standard:

$$S_{AA(i,j)} = s\left(amino_{target(i)}, amino_{templato(j)}\right)$$

moltiplicato per un coefficiente cSeq e sommato all'eventuale punteggio strutturale:

$$S_{AA(i,j)} = cSeq \cdot S_{AA(i,j)} + S_{STR(i,j)}$$

- ScoringP2S, per gli allineamenti profilo contro sequenza. Il punteggio S_{ij} fra l'i-esima colonna del profilo per una coppia di posizioni è stabilito grazie ad una matrice di sostituzione standard:

$$S_{AA(i,j)} = s\left(amino_{target(i)}, amino_{templato(j)}\right)$$

[17] Le informazioni di threading impiegate in ALIGN sono tabelle 20xN, con N lunghezza della sequenza templato, dove ogni colonna contiene un punteggio di qualità (fitness score) per ciascun amminoacido in quella posizione nel templato.

moltiplicato per un coefficiente cSeq e sommato all'eventuale punteggio strutturale.

- ScoringP2P, per gli allineamenti profilo contro profilo[18]. In questo caso il punteggio S_{ij} fra l'i-esima colonna del profilo della sequenza target e la j-esima colonna del profilo della sequenza templato, viene espresso dalle varie scoring function implementate, moltiplicato per un opportuno coefficiente cSeq e sommato all'eventuale punteggio strutturale.

9.3.7 Align

La classe astratta Align gestisce l'algoritmo di allineamento utilizzato.

Esistono quindi tre classi derivate che coprono le tipologie di allineamento già presentate nel capitolo 3:

- NWAlign, per gli allineamenti globali (algoritmo di Needleman-Wunsch)
- SWAlign, per gli allineamenti locali (algoritmo di Smith-Waterman)
- FSAlign, che è quello utilizzato nel corrente progetto, per gli allineamenti semi-globali (algoritmo Freeshift).

[18]Dal momento che Homer in automatico fa ricorso ad allineamenti profilo-profilo, questa è la classe effettivamente utilizzata nel progetto.

10 Homer: costruzione del modello grezzo

La costruzione di un pre-modello è la parte più semplice di tutto il processo e consiste nel costruire la struttura del target sulla base delle coordinate atomiche del templato.

La parte implementativa consiste in un insieme di script in Perl e di classi in C++ raccolte in un pacchetto chiamato Homer.

Il suo sviluppo storicamente è strettamente legato a quello della libreria Biopool e anche le modifiche fatte per questo progetto rispecchiano le novità introdotte al capitolo 7, in particolare nella gestione della nuova classe introdotta: la classe Protein.

Per inciso questi due strumenti sono gli unici che il predittore realizzato per questa tesi ha in comune con la precedente versione, sviluppata sempre all'interno del laboratorio di BioComputing [17].

I passi da compiere qui sono molto semplici: si tratta in sostanza di sfruttare le regioni allineate copiando le posizioni degli amminoacidi del templato sui relativi residui del target. Ciononostante si tratta dell'operazione più caratteristica del processo di modellazione comparativa. E' quindi naturale che questo componente dia il nome all'intero progetto.

Abbiamo detto che la struttura templato viene usata come "stampo" per costruire il modello seguendo l'allineamento. Se vi è identità di sequenza questo si traduce nella possibilità di copiare direttamente tutte le coordinate cartesiane, comprese quelle delle catene laterali.

Al contrario, se i residui non sono identici, l'informazione sulla catena laterale, che è diversa per ogni amminoacido, viene persa in questo passaggio. In questo caso quindi, ad essere utilizzate sono solo le coordinate atomiche della backbone.

Gli approcci di costruzione del modello grezzo possono variare nel caso in cui si abbiano più templati a disposizione.

In tali situazioni si hanno due alternative:

- mediare le varie posizioni ricavabili da tutti i templati per ciascun residuo e applicarle al residuo (approccio restrained-based)
- scegliere per ciascuna regione del target il templato che si allinea in maniera migliore e usare solo queste coordinate atomiche (approccio fragment-based).

Homer implementa il secondo metodo descritto, ma per il momento si limita a considerare un solo templato: quello che offre l'allineamento migliore.

Estendere l'implementazione dell'approccio fragment-based può portare a modellare in modo efficace anche regione del target che con un solo templato rimarrebbero escluse, con ovvie ricadute positive sulle valutazioni dei modelli prodotti (più è ampia la porzione che si riesce a modellare e migliore è la valutazione).

Per questo motivo rappresenta senz'altro uno degli sviluppi che si dovranno considerare per questo progetto.

I modelli ottenuti come risultato di questo step non rappresentano mai delle strutture complete, e vengono chiamanti modelli grezzi o pre-modelli proprio per indicare il fatto che riman-

gono porzioni della sequenza target da modellare, più o meno numerose a seconda delle qualità dell'allineamento trovato.

Le catene laterali mancanti e le regioni variabili della struttura (generalmente loop) che presentano inserzioni o delezioni, dovranno essere predette e modellate con altri metodi.

Nei prossimi capitoli vedremo in che modi e con quali strumenti queste problematiche sono state affrontate in HOMER.

11 Modellazione delle catene laterali

La costruzione delle catene laterali è un processo che si è guadagnato una certa autonomia rispetto agli altri passi di modellazione.

Spesso infatti si adotta una ragionevole semplificazione computazionale per la quale le catene laterali vengono considerate indipendenti dal backbone, il cui sviluppo nello spazio è mantenuto fisso.

Ne segue che l'RMSD della struttura cambia relativamente poco, e in competizioni come il CASP la loro presenza non influenza la qualità del modello, quasi sempre basata su valutazioni del grado di sovrapposizione dei soli carboni α.

Ciò non di meno in HOMER questo passaggio è stato incluso in quanto, oltre ad ottenere una struttura più completa, le catene laterali concorrono a definire la struttura dei siti attivi di una proteina, e conoscerne le conformazioni diventa quindi assai importante.

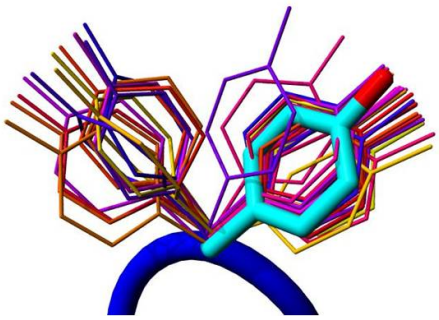

Figura 37: Possibili configurazioni alternative per una catena laterale

Il metodo più usato per modellare tutte quelle catene laterali che non possono essere costruite a livello di modello grezzo, e a cui si è fatto ricorso anche in questa tesi, è SCWRL [18,19].

SCWRL usa un approccio di tipo knowledge-based dove vengono considerate le conformazioni più probabili per ciascuna catena laterale sulla base dell'osservazione delle strutture proteiche note.

Tali conformazioni preferite sono note con il termine di **rotameri** e fortunatamente sono in numero molto limitato.

SCWRL posiziona le catene laterali scegliendo le conformazioni più probabili al fine di porre gli atomi il più possibile distanti fra loro e di minimizzare l'ingombro sterico. Nel caso di incompatibilità strutturali, ad esempio sovrapposizioni, crea dei cluster comprendenti un certo numero di catene laterali e sceglie tra le varie combinazioni strutturalmente compatibili, quella più probabile.

In ogni caso dove è possibile è sempre meglio mantenere le conformazioni delle catene laterali del templato in quanto tra proteine omologhe gli amminoacidi conservati mantengono la stessa conformazione della catena laterale.

Un discorso analogo può essere fatto anche per ligandi e cofattori-metallici: il programma di modellazione a cui si è scelto di fare ricorso può sfruttare questo tipo di informazioni contenute nel templato per migliorare il posizionamento delle catene laterali nel target. Anche questi elementi hanno infatti un loro ingombro spaziale, e tenerne conto permette quindi di avere risultati più precisi.

In particolare SCWRL richiede che le informazioni sulle coordinate cartesiane degli atomi del templato gli vengano fornite in input separatamente da quelle su ligandi e cofattori.

Fortunatamente ciò rispecchia proprio il modo in cui è stata implementata la classe Protein in Biopool, al cui interno le due tipologie di informazioni sono già naturalmente separate nelle classi Spacer e LigandSet.

12 Qmean: valutazione energetica

Nella predizione di strutture proteiche normalmente viene prodotto un gran numero di modelli alternativi e la selezione del più accurato è un passaggio cruciale.

Ovviamente non conoscendo la struttura reale non è possibile fare delle valutazioni oggettive, ma si punta piuttosto ad utilizzare metodi statistici che forniscano valori di pseudo-energia legati al concetto di probabilità e propensione.

Più precisamente verrà premiato un modello che comprende soluzioni strutturali molto frequenti in natura, mentre se presenta soluzioni strutturali rare o mai viste verrà penalizzato.

QMEAN (acronimo di Qualitative Model Energy ANalisys) è il valutatore energetico che è stato inserito in HOMER [20,21,22,23]. Sviluppato in collaborazione con il laboratorio di BioComputing, è stato sviluppato anche con l'intenzione di migliorare le performance di un altro analogo tool sviluppato nel laboratorio in cui è stata svolta questa tesi: FRST [24].

Essendo stato sviluppato in parte esternamente, il codice sfruttava una libreria di terze parti per rappresentare i vari elementi biologici che si è rivelata essere molto più pesante e farraginosa di Biopool.

All'atto di integrarlo in HOMER si è quindi deciso, ai fini dell'ottimizzazione, di utilizzare un radicale approccio di reverse engineering e di riscrivere interamente il programma per far si che potesse usare la libreria Biopool. Questa decisione che ha richiesto una discreta parte del tempo di questo progetto, ha portato ad una versione molto più leggera e performante del programma, garantendo nel futuro una più semplice manutenzione ed estrema facilità nell'interfacciarlo con i numerosi tool sviluppati all'interno del laboratorio e che sono tutti basati sulla medesima libreria.

QMEAN comprende 6 potenziali:

- Torsion potential: è il modulo che fornisce un valore che considera la propensione di un residuo ad assumere una determinata conformazione degli angoli torsionali. Questo potenziale considera gli amminoacidi in triplette, e in questo modo la descrizione della geometria locale di un determinato residuo viene migliorata andando a considerare anche gli angoli torsionali dei residui adiacenti.

- Pair residue: fornisce un potenziale statistico a coppie che considera le distanze tra diversi residui usando i carboniiβ come centri di interazione. Un intervallo di distanze compreso tra 3 e 25 Å ha rivelato dare i migliori risultati.

- Pair all-atom: simile al precedente, considera i questo caso tutti gli atomi dei diversi residui, ad eccezione degli idrogeni.

- Solvation: potenziale statistico che, calcolando per ciascun amminoacido il numero di altri residui racchiusi in una sfera di raggio 9 Å, permette di approssimare il valore di energia di solvatazione.

- SSEagreement: corrispondenza tra la predizione di struttura secondaria della sequenza target (ottenuta con PSIPRED) e la struttura secondaria del modello (calcolata con DSSP)

- ACCagreement: corrispondenza tra la previsione di accessibilità al solvente (calcolata con ACCpro) e l'informazione di accessibilità al solvente ottenuta con DSSP.

Il valore di pseudo-energia globale è definito nel seguente modo:

$QMEAN = W_{torsion} \cdot E_{torsion} + W_{solvation} \cdot E_{solvation} + W_{pair,residue} \cdot E_{pair,residue} + W_{pair,all-atom} \cdot E_{pair,all-atom} + W_{SSEagreement} \cdot S_{SSEagreement} \cdot W_{ACCagreement} \cdot S_{ACCagreement} + intercept$

dove $W_{torsion} = -0.00185$, $W_{solvation} = -0.00054$, $W_{pair,residue} = -0.00062$, $W_{pair,all-atom} = -0.00108$, $W_{SSEagreement} = 0.38072$, $W_{ACCagreement} = 0.57997$, $intercept = -0.28663$

E' inoltre possibile ottenere una valutazione residuo per residuo, e questo viene fatto per il modello finale al fine di fornire un'indicazione di quali parti della proteina risultano ben modellate e quali invece probabilmente sono indicatrici di errori nella costruzione del modello.

12.1 Struttura della libreria QMEAN

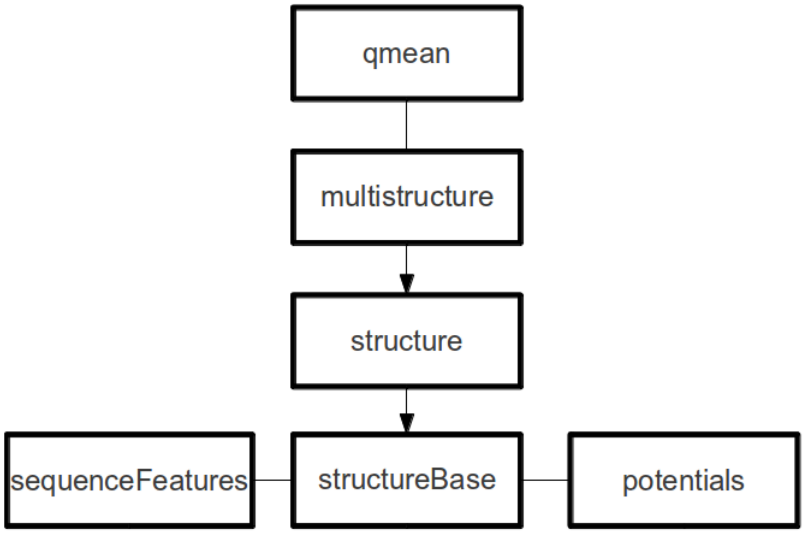

Figura 38: Principali classi della libreria QMEAN

Nella nuova implementazione la gerarchia delle classi è stata rivista: alcune sono state accorpate, altre radicalmente modificate. In ogni caso si è cercato di mantenere simile la struttura principale basata sulla classe structureBase, e l'implementazione dei 6 potenziali che questa va a calcolare.

Di seguito viene riportata una breve descrizione delle principali componenti, e a seguire verrà mostrato il confronto con la complessità della versione originale effettuato con CCCC: un tool automatico per l'analisi della qualità del codice prodotto.

12.1.1 qmean

Il punto d'accesso all'intera libreria è però la classe qmean: questa svolge il compito di interfaccia verso l'utente, raccogliendo le informazioni necessarie all'esecuzione del programma quali:

- la directory contenente i modelli da analizzare
- il file con la previsione della struttura secondaria (ottenuto con PSIPRED)

- il file con la previsione dell'accessibilità al solvente (ottenuta con ACC_pro)
- la directory di output dove scrivere i risultati

12.1.2 multistructure

Questa classe riceve le informazioni di input e crea un oggetto di tipo *structure* per ciascuno dei modelli da analizzare. Si occupa inoltre di preparare l'output, fornendo in particolare un file in formato ods con le informazioni sui singoli potenziali calcolati per ogni struttura in analisi, ed un ulteriore file in cui riporta il punteggio globale calcolato come media dei sei risultati, pesando i differenti valori secondo i parametri illustrati in precedenza.

12.1.3 structure

Structure è la classe con la quale si rappresenta e valuta il modello.

Il nucleo è rappresentato da sei metodi che si occupano del calcolo di ciascuno dei potenziali presentati in precedenza:

- getPairwise_SSE_Energy che si occupa del potenziale basato sulla distanza di coppie di amminoacidi
- getCombined_torsion_Energy che calcola il potenziale torsionale
- get_all_atom_Pairwise_SSE_Energy che estende il calcolo del primo potenziale a coppie dagli amminoacidi (il cui centro di riferimento è il $C\beta$, agli atomi. Questo potenziale quindi discrimina tra strutture più complete e meno complete.
- get_SSE_Q3_PSIPRED_score che raccoglie le informazioni sulla previsione di struttura secondaria e le confronta con quelle ricavate direttamente dal modello in esame
- get_ACC_conservation_SSpro_score analogo al precedente con la differenza che in questo caso la caratteristica in analisi è l'accessibilità al solvente.

12.1.4 structureBase

Concettualmente simile alla precedente, implementa anch'essa sei metodi che tuttavia si occupano di calcolare i potenziali per un determinato amminoacido. E' compito della classe structure richiedere di volta in volta al relativo oggetto della classe structureBase i potenziali per ogni singolo amminoacido presente nel modello.

Di fatto quindi la vera implementazione di tali potenziali è compito di questa classe, che rappresenta quindi la parte più importante dell'intera libreria.

12.1.5 sequenceFeatures

E' una classe di supporto al lavoro di structureBase: si occupa di estrarre le informazioni utili sulla previsione di struttura secondaria e sull'accessibilità al solvente dai file in formato PSIPRED e ACCRO rispettivamente, compiendo un preliminare lavoro di verifica della congruità di tali informazioni.

Module Name	LOC	MVG	COM	L_C	M_C
Alignment	*885*	148	456	1.941	0.325
AminoAcidMaps	*916*	**1409**	352	2.602	4.003
Modeling	*980*	167	717	1.367	0.233
MultiStructure	*1460*	169	330	4.424	0.512
Potentials	407	8	36	*11.306*	0.222
SequenceFeatures	471	149	60	*7.850*	2.483
Structure	**4533**	*963*	1149	3.945	0.838
StructureBase	**3140**	*557*	1545	2.032	0.361
StructureEnergyProfile	*567*	65	165	3.436	0.394
TMscore	423	98	64	6.609	1.531
Training	**3144**	*610*	639	4.920	0.955
anonymous	*1348*	*265*	249	5.414	1.064

Figura 39: Analisi della complessità per la versione originale di QMEAN

12.1.6 potentials

Altra classe di supporto a structureBase: si occupa di recuperare informazioni dalle proprie librerie di potenziali, ad esempio dalla libreria di rotameri per il calcolo dei potenziali torsionali.

12.1.7 Analisi della complessità

La riscrittura di QMEAN ai fini della sostituzione della originale libreria di base con Biopool, ha permesso di snellire in modo notevole il programma e di diminuirne la complessità con effetti positivi sia per la leggibilità del codice, sia per il tempo di calcolo.

A riprova di questo fatto si è utilizzato un tool di analisi automatica (CCCC) per valutare il codice sorgente in C++ delle due versioni. La qualità viene espressa attraverso varie metriche, in figura 39 e 40 vengono riportate alcune delle più utili:

- LOC = righe di codice. Linee vuote o righe di commento non sono conteggiate.
- COM = linee di commenti. E' il numero di linee di commento identificate dall'analizzatore. Non vengono considerati commenti che non si estendono all'intera riga.
- MVG = Complessità ciclomatica di McCabe. Esprime la complessità nel livello di decisioni nelle funzioni che compongono il programma. Una definizione più rigorosa consiste nel numero di percorsi linearmente indipendenti attraverso un grafo orientato aciclico che mappa il flusso di controllo di un sottoprogramma. L'analisi quindi considera come valore il massimo numero possibile di combinazioni delle varie decisioni che si possono prendere nello svolgimento del programma.
- L_C = righe di codice per linea di commento. Indica la densità dei commenti rispetto alla dimensione testuale del programma.
- M_C = complessità ciclomatica per riga di commento. Indica la densità dei commenti rispetto alla complessità logica del programma.

Module Name	LOC	MVG	COM	L_C	M_C
MultiStructure	127	6	26	4.885	0.231
Potentials	77	3	9	*8.556*	------
SequenceFeatures	187	61	26	*7.192*	2.346
Spacer	0	0	0	------	------
Structure	258	51	48	5.375	1.062
StructureBase	*966*	197	321	3.009	0.614
anonymous	76	16	7	*10.857*	2.286

Figura 40: Analisi della complessità per l'attuale versione di QMEAN

Come si può vedere il numero di righe di codice per le classi principali, come ad esempio structureBase, è notevolmente diminuito, come lo è anche la complessità. Altre classi poi sono state del tutto eliminate principalmente perché sostituite da analoghe controparti nella libreria Biopool.

13 Modellazione dei loop

Al pre-modello possono mancare interi frammenti di catena principale.

Il motivo solitamente è dovuto al fatto che queste regioni sono meno importanti per la proteina e risultano meno conservate da un punto di vista evolutivo, pertanto non vengono allineate con il templato.

Di questi amminoacidi non si ha quindi alcuna informazione strutturale per cui ci si trova a dover ricostruire totalmente una parte di proteina ignota, che per giunta rappresenta la sua regione maggiormente flessibile.

La difficoltà per la ricostruzione di un loop, che è direttamente proporzionale alla sua lunghezza, fa si che il Loop Modeling sia una delle fasi più lunghe in termini di tempi di calcolo e che spesso il risultato ottenuto non sia molto affidabile.

Ciononostante, dato che in natura non può esistere una struttura con dei buchi, modelli incompleti vengono solitamente penalizzati in competizioni quali il CASP.

Esistono diversi approcci al problema del loop modeling:

- Costruzione ex-novo dell'inserzione (metodi Ab Initio): si usano tecniche di Novel Fold dove la struttura di una regione viene calcolata a partire dalle caratteristiche chimico-fisiche dei residui che la compongono. Si generano molti frammenti alternativi provando a creare delle combinazioni casuali tra cui selezionare la soluzione migliore tramite una scoring function. L'unico vincolo strutturale è basato su considerazioni geometriche (angoli torsionali) legate alla posizione degli amminoacidi pre e post-loop.

- Uso di librerie di frammenti (approccio knowledge-based): si sfruttano le conoscenze che si hanno a disposizione andando ad estrarre frammenti di loop con struttura conosciuta dalla banca dati PDB. Sulla libreria così realizzata vengono fatte delle ricerche in base alla dimensione e alla sequenza della regione da modellare. Si sceglie il frammento che rispetta meglio i vincoli geometrici e che non crea problemi di sovrapposizione con il resto della struttura.

- Metodo Divide & Conquer: se il problema richiede di collegare il punto A con il punto B che stanno all'inizio e alla fine del loop che vogliamo modellare, trovando il punto C che sta al centro abbiamo di fatto scomposto il problema in due sotto problemi più semplici. Si procede in modo ricorsivo finché non si arriva al singolo amminoacido, a quel punto le soluzioni trovate vengono ricomposte a formare la soluzione per il loop originale.

Il problema principale ovviamente è trovate il punto centrale C. La soluzione a questo problema consiste nel creare delle lookup table: un database di possibili punti centrali e posizioni finali di tutte le lunghezze.

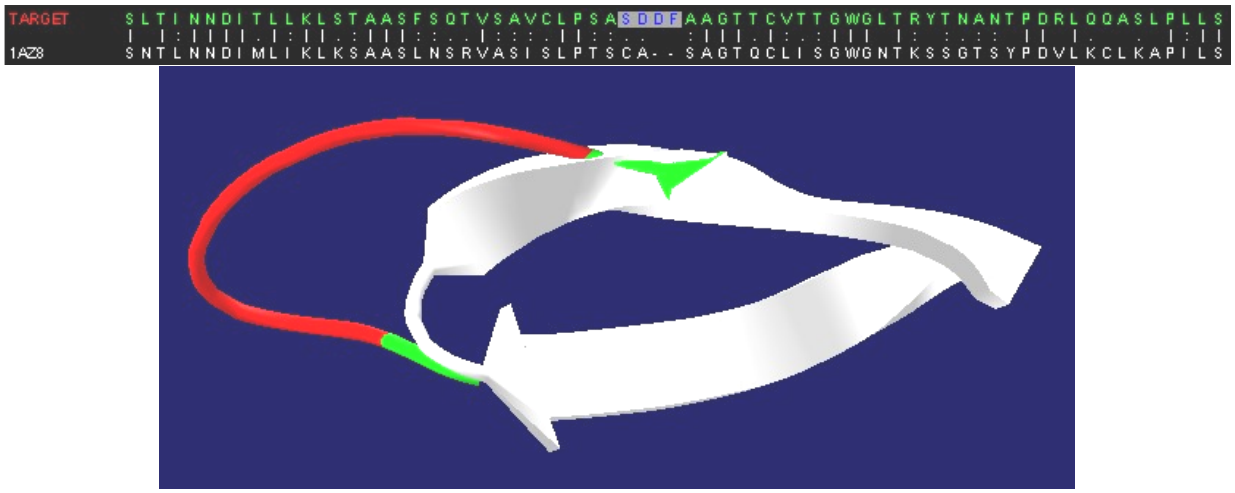

Figura 41: Esempio di modellazione di un loop

Selezione e ranking delle soluzioni trovate si basano su considerazioni di:

- sequenza: si eliminano le configurazioni impossibili;
- geometria: la geometria prima e dopo il loop deve essere buona;
- energia, tramite un potenziale statistico che esprima quanto gli atomi siano in una buona posizione.

Per il problema del loop modeling si è scelto di utilizzare un altro strumento sviluppato al laboratorio: LOBO.

LOBO (acronimo di LOop Build-up and Optimization) implementa una strategia di tipo divide&conquer e utilizza look-up tables (LUTs) calcolate a priori per minimizzare i tempi di calcolo. Queste tavole sono costruite a partire da una distribuzione di Ramachandran di angoli torsionali ricavati da strutture PDB e contengono loop di varia lunghezza.

I vari candidati loop sono soggetti a vari criteri di valutazione: filtri di Van Der Waals e di continuità della catena etc.

L'algoritmo e le strutture dati necessarie per il processo di loop modeling sono implementate nel pacchetto Nazgûl; per ulteriori informazioni si rimanda a [25].

In HOMER il ricorso alla modellazione dei loop, come avviene anche per la ricostruzione delle catene laterali, è del tutto facoltativo. In ogni caso, poiché si tratta di un passaggio che richiede solitamente parecchio tempo, il suo utilizzo viene sempre limitato al modello finale.

In pratica quindi il loop modeling in homer non concorre alla scelta del modello migliore tra le varie strutture prodotte, ma tenta solamente di migliorarlo una volta che sia stato individuato.

Va però aggiunto che nei risultati mostrati in questa tesi si è deciso di non includere il passaggio di modellazione dei loop. Sfortunatamente LOBO presenta ancora alcune problematiche non risolte che possono compromettere il funzionamento di HOMER, mentre l'obiettivo di questa tesi era ottenere una versione del programma stabile e completamente automatica. Il suo effettivo utilizzo richiede dunque ancora una ulteriore fase di analisi.

14 Risultati

Per verificare le capacità del programma realizzato si è deciso di simulare la partecipazione all'ultima edizione del CASP: il CASP10. La partecipazione di server automatici e dei gruppi di ricerca è mantenuta separata, ed il test si è dunque svolto solo sui target dedicati alla prima categoria. Una lista completa è stata inclusa in appendice.

Nel momento in cui si scrive questa tesi, il CASP si è concluso da qualche mese e parte dei modelli delle sequenze oggetto di competizione sono già stati inseriti nelle relative banche dati. Queste inoltre nel frattempo si sono arricchite di molte altre strutture che potenzialmente possono rappresentare templati migliori di quelli che erano disponibili al momento della competizione.

Per garantire la validità dei dati ottenuti si sono pertanto preparate copie locali delle banche dati di sequenze e strutture proteiche necessarie (nr90 e fold 98) che risalgono alla fine del mese di aprile 2012: pochi giorni prima dell'inizio del CASP.

Nello sviluppo di HOMER una prima fase di test ha riguardato l'influenza del numero di iterazioni usate in PSI-BLAST nella ricerca del templato. Dalla letteratura si evince che in genere 4 round sono sufficienti a trovare un buon templato, e che non è consigliabile andare oltre i 6 per evitare fenomeni di deriva. Poiché al crescere del numero di iterazioni cresce anche il tempo di calcolo, si è fatto veloce confronto tra i due valori più promettenti: 4 e 5 round.

Per tutti gli altri parametri quali il costo di apertura ed estensione dei gap, la lunghezza di parola, i valori di soglia, la matrice di sostituzione usata etc. si è preferito affidarsi alla solidità delle scelte implementate di default del programma.

I risultati sono riportati in figura 2:

Nella tabella sono riportati per ogni target il rispettivo templato individuato utilizzando prima 4 e poi 5 iterazioni. Anziché comparare i relativi e-value, si è preferito utilizzare come riferimento il rapporto tra lunghezza del templato e lunghezza del target. Tale valore fornisce una stima, seppur approssimativa, della "copertura" che il templato può assicurare sulla sequenza in analisi. Un buon modello infatti non dipende solo dalla qualità del templato, ma anche dall'estensione della porzione di target che su questo si riesce a modellare. Ovviamente non c'è garanzia che nella fase di allineamento tutti i residui del templato vengano utilizzati, ma in questo caso il valore trovato è sufficiente a delineare il comportamento generale dell'algoritmo.

Come si può notare, in alcuni casi l'utilizzo di 5 iterazioni ha portato ad un miglioramento (celle evidenziate in blu) che, cosa importante, spesso è legato all'individuazione di un nuovo e più utile templato (celle evidenziate in giallo). Vi è anche un caso in cui il rapporto in esame è invece diminuito (celle rosse).

Se consideriamo ad esempio il target T0738 notiamo che l'incremento è di un solo amminoacido nella sequenza templato, mentre per il T0702 è notevole: si passa da un rapporto tra il numero di residui di 0,90 ad un rapporto di 1,00. Il nuovo templato in questo caso potrebbe quindi permetterci di modellare l'intera struttura del target. Data l'importanza di questa fase nel processo di modellazione anche piccoli miglioramenti sono diventano importanti, facendo quindi preferire l'utilizzo di 5 iterazioni. Un buon esempio dell'importanza di tenere in considerazione la lunghezza del templato oltre al parametro di e-value in base a cui PSI-BLAST ordina i risultati, è dato dal target T0652 (figura 42).

TARGET	residui	PSI-BLAST 4R	TPLen	TPLen/TGLen	PSI-BLAST 5R	TPLen	TPLen/TGLen
T0645	537	3CGHA	515	0,96	3CGHA	515	0,96
T0648	102	1HX5G	97	0,95	1HX5G	97	0,95
T0650	346	1O6SA	336	0,97	1O6SA	336	0,97
T0652	292	3LFRB	136	0,47	3LFRB	136	0,47
T0654	166	2F9SB	129	0,78	2F9SB	129	0,78
T0657	154	1B55B	163	1,06	1B55B	163	1,06
T0659	85	2KPPA	88	1,04	2KPPA	88	1,04
T0661	215	2QGUA	211	0,98	2QGUA	211	0,98
T0662	79	3EJBG	75	0,95	3EJBG	75	0,95
T0664	540	3CGHA	514	0,95	3CGHA	514	0,95
T0667	194	3LLCA	230	1,19	3LLCA	230	1,19
T0669	109	1PQXA	79	0,72	1PQXA	79	0,72
T0672	335	1A8SA	272	0,81	1A8SA	272	0,81
T0675	75	1X6EA	55	0,73	1X6EA	55	0,73
T0677	153	1A5JA	102	0,67	1A5JA	102	0,67
T0679	223	1IVNA	178	0,80	1IVNA	178	0,80
T0681	224	3HN5B	213	0,95	3HN5B	213	0,95
T0683	403	3H2GA	368	0,91	3H2GA	368	0,91
T0685	253	3D36A	218	0,86	3D36A	219	0,87
T0688	196	1M9SA	206	1,05	1O6SA	211	1,08
T0689	234	3FZXA	216	0,92	3FZXA	216	0,92
T0692	473	3JZ4A	458	0,97	3JZ4A	458	0,97
T0694	315	1WAWA	341	1,08	1WAWA	341	1,08
T0696	111	1LQKB	109	0,98	1F9ZB	127	1,14
T0697	483	3RBFB	479	0,99	3RBFB	478	0,99
T0698	119	2L2CA	102	0,86	2L2CA	102	0,86
T0699	234	1VGTB	222	0,95	1VGTB	222	0,95
T0701	322	2ZDSF	320	0,99	2ZDSF	324	1,01
T0702	271	3GT0A	245	0,90	3TRIB	270	1,00
T0703	272	2J7VD	259	0,95	2J7VD	259	0,95
T0706	217	3IBSA	206	0,95	3IBSA	206	0,95
T0708	196	3LQYA	181	0,92	3LQYA	181	0,92
T0710	220	2XEED	148	0,67	2XEED	148	0,67
T0712	223	3U22A	200	0,90	3U22A	200	0,90
T0714	88	1FHGA	86	0,98	1FHGA	86	0,98
T0715	462	3JZ4A	467	1,01	3JZ4A	463	1,00
T0716	71	3A02A	55	0,77	3A02A	55	0,77
T0721	301	2A87B	305	1,01	2A87B	305	1,01
T0731	79	2KZ5A	69	0,87	2KZ5A	69	0,87
T0733	390	3V5NC	382	0,98	3V5NC	382	0,98
T0736	168	3B8LF	129	0,77	3EF8A	137	0,82
T0738	249	3SJ7B	241	0,97	3RROB	242	0,97
T0747	121	3D33B	97	0,80	3D33B	97	0,80
T0749	449	3EU8D	420	0,94	3EU8D	420	0,94
T0750	188	1IVNA	174	0,93	1IVNA	174	0,93
T0752	156	3B8LF	129	0,83	3EF8A	139	0,89
T0753	109	2KNRA	111	1,02	2KNRA	111	1,02
T0755	264	4ECFA	126	0,48	4ECFA	151	0,57
T0756	179	1ULYA	123	0,69	1ULYA	123	0,69
T0757	247	2OWNA	244	0,99	2OWNA	244	0,99
T0758	388	1JQIB	380	0,98	1JQIB	380	0,98

Tabella 2: Analisi del numero di residui dei templati al variare del numero di iterazioni di PSIBLAST

```
PSIBLAST 2.2.26+
...                                                                  Score     E
Sequences producing significant alignments:                          (Bits)  Value

pdb|3LFRB|   unnamed protein product                                   217    9e-71
pdb|3OI8B|   unnamed protein product                                   207    2e-66
...

>pdb|3LFRB|
Length=136

 Score =  217 bits (555),  Expect = 9e-71, Method: Composition-based stats.
 Identities = 80/136 (59%), Positives = 106/136 (78%), Gaps = 1/136 (1%)

Query  65   ADQRVRDIMIPRSQMITLKRNQTLDECLDVIIESAHSRFPVISEDKDHIEGILMAKDLLP  124
            AD +VRDIM+PRSQMI++K  QT  E L  +I++AHSR+PVI E  D + G+L+AKDLLP
Sbjct  1    ADLQVRDIMVPRSQMISIKATQTPREFLPAVIDAAHSRYPVIGESHDDVLGVLLAKDLLP  60

Query  125  FMRS-DAEAFSMDKVLRQAVVVPESKRVDRMLKEFRSQRYHMAIVIDEFGGVSGLVTIED  183
             +   D ++  + K+LR A  VPESKR++ +L+EFR+   HMAIVIDE+GGV+GLVTIED
Sbjct  61   LILKADGDSDDVKKLLRPATFVPESKRLNVLLREFRANHNHMAIVIDEYGGVAGLVTIED  120

Query  184  ILELIVGEIEDEYDEE  199
            +LE IVG+IEDE+D E
Sbjct  121  VLEQIVGDIEDEHDVE  136

>pdb|3OI8B|
Length=156

 Score =  207 bits (528),  Expect = 2e-66, Method: Composition-based stats.
 Identities = 80/154 (52%), Positives = 107/154 (69%), Gaps = 1/154 (1%)

Query  33   NRDELLALIRDSGQNDLIDEDTRDMLEGVMDIADQRVRDIMIPRSQMITLKRNQTLDECL  92
            + +++L L+R + + ++ D DT   LE V+D +D  VRD MI RS+M  LK N +++
Sbjct  4    SAEDVLNLLRQAHEQEVFDADTLLRLEKVLDFSDLEVRDAMITRSRMNVLKENDSIERIT  63

Query  93   DVIIESAHSRFPVISEDKDHIEGILMAKDLLPFMRSDAEAFSMDKVLRQAVVVPESKRVD  152
              +I++AHSRFPVI EDKD + GIL AKDLL +M  + E F +  +LR AV VPE K +
Sbjct  64   AYVIDTAHSRFPVIGEDKDEVLGILHAKDLLKYMF-NPEQFHLKSILRPAVFVPEGKSLT  122

Query  153  RMLKEFRSQRYHMAIVIDEFGGVSGLVTIEDILE  186
             +LKEFR QR HMAIVIDE+GG SGLVT EDI+E
Sbjct  123  ALLKEFREQRNHMAIVIDEYGGTSGLVTFEDIIE  156
```

Figura 42: PSI-Blast output per T0652

L'attuale procedura utilizzata nella fase di ricerca del templato va a selezionare il templato 3LRF, e in particolare la catena B. Il secondo candidato, il templato 30I8 presenta molti più residui e un e-value comunque molto buono. Potrebbe quindi rivelarsi più utile.

Per verificare queste ipotesi al momento è allo studio una nuova procedura di selezione che vada a considerare entrambe le caratteristiche.

Nella tabella 3 si analizza invece il comportamento di HOMER (HomerP2P nella versione che sfruttava una sola tipologia di allineamento, HomerP2PManyAlignments nella versione che ne utilizza più di una) in confronto ai modelli prodotti con una sua precedente versione con la quale si è effettivamente partecipato all'ultima edizione del CASP con la sottomissione di due modelli (CASPOld1 e CASPOld2), in base a quanto concesso dal regolamento. Come ulteriore riferimento si riportano i risultati di un ulteriore partecipante (Distill) che rappresenta bene la qualità media dei partecipanti.

TARGET	HomerP2P	HomerP2PManyAlignents	CASPOld1	CASPOld2	Distill
T0645	0,6571	0,7068	0,2525	0	0,7681
T0648	0,7878	0,6831	0	0	0,8140
T0650	0,5863	0,7625	0,7633	0	0,6202
T0652-D1	0,7699	0,7699	0	0	0,8786
T0652-D2	0	0	0	0	0,9066
T0654	0,5858	0,6437	0	0,5485	0,6959
T0657	0,7951	0,7970	0,8083	0,8233	0,8703
T0659	0,8209	0,8209	0,1926	0,8547	0,9155
T0661	0,7149	0,7203	0,7149	0,7149	0,7365
T0662	0,6809	0,6809	0,7138	0,7237	0,7401
T0664	0,6566	0,7626	0,1280	0,2435	0,7982
T0667	0,5521	0,5625	0,5391	0,4622	0,6354
T0669	0,5464	0,5619	0,5876	0,5129	0,616
T0672	0,5705	0,5649	0,5425	0,5369	0,6042
T0675	0,8426	0,8426	0,8426	0	0,8426
T0675-D2	0,7417	0,7417	0,7667	0	0,8167
T0677	0,6136	0,6136	0,6591	0,6420	0,8693
T0677-D2	0,4931	0,5035	0,5	0,4444	0,5556
T0679	0,4987	0,5352	0,4749	0,5616	0,6633
T0681	0,6904	0,6865	0,0838	0,0647	0,6954
T0683	0,2170	0,2264	0,2756	0,2540	0,7224
T0685	0,6250	0,6563	0,2778	0,5104	0,7014
T0685-D2	0,6040	0,6040	0,4234	0,4799	0,5785
T0688	0,5730	0,5270	0,4149	0,4743	0,2821
T0689	0,8424	0,8483	0	0	0,8732
T0692	0,7415	0,7420	0,7346	0,7415	0,7755
T0694	0,5761	0,6346	0,5401	0,5088	0,6434
T0696	0,4300	0,4100	0,4775	0	0,4775

Tabella 3: GDT_TS score di Homer (HomerP2P: allineamenti di un solo tipo, HomerMultiP2P: varie combinazioni di allineamento), di due modelli di una precedente versione e di Distill (che rappresenta la qualità media dei partecipanti al CASP).

TARGET	HomerP2P	HomerP2PManyAlignents	CASPOld1	CASPOld2	Distill
T0697	0,6114	0,6124	0,6447	0,6567	0,7937
T0698	0,3825	0,4167	0,4765	0,4594	0,5897
T0699	0,6489	0,6722	0	0	0,8167
T0701	0,6582	0,7025	0,3877	0,4343	0,6250
T0702	0,7043	0,6203	0	0	0,7379
T0703	0,4706	0,6287	0	0	0,7298
T0706	0,7306	0,7176	0	0	0,6956
T0708	0,7768	0,7500	0,7041	0,7474	0,8240
T0710	0,5193	0,4562	0,5464	0,5206	0,6740
T0712	0,8629	0,8696	0,8199	0	0,9140
T0714	0,6222	0,6818	0,6335	0,7017	0,8097
T0715	0,5443	0,5437	0,5121	0,4989	0,5351
T0716	0,8775	0,8775	0,8824	0,9216	0,9461
T0721	0,6087	0,6263	0,4883	0,4540	0,6522
T0731	0,7955	0,7955	0,8091	0	0,8318
T0733	0,5957	0,6410	0,5253	0	0,6270
T0736	0,5964	0,5919	0,5120	0,4985	0,7139
T0738	0,8032	0,7902	0,7922	0,7861	0,8614
T0747	0,2083	0,2083	0,6250	0,6000	0,2667
T0749	0,8926	0,8902	0,1183	0,1086	0,8981
T0750	0,4849	0,489	0,4780	0,4931	0,6470
T0752	0,6233	0,7027	0,6486	0,5709	0,8091
T0753	0,6644	0,6852	0,7176	0,7269	0,7292
T0755	0,3808	0,3479	0,1017	0,2936	0,4157
T0756	0,7582	0,7582	0,6648	0,6676	0,6896
T0756-D2	0,2878	0,2645	0,3227	0,2762	0,2558
T0757	0,7490	0,7470	0,7379	0,7429	0,7713
T0758	0,6749	0,6749	0,6872	0,6844	0,6919
Media	**0,6125**	**0,6316**	**0,4562**	**0,3919**	**0,7080**

Tabella 4: GDT_TS score di Homer (HomerP2P: allineamenti di un solo tipo, HomerMultiP2P: varie combinazioni di allineamento), di due modelli di una precedente versione e di Distill (che rappresenta la qualità media dei partecipanti al CASP).

Come si può notare HOMER si comporta mediamente molto meglio della versione da cui ha preso spunto, anche grazie al fatto che risulta molto più robusto riuscendo a costruire una struttura per ogni target. Inoltre nella versione in cui vengono utilizzate varie tipologie di allineamento anziché una sola (selezionata in quanto mediamente la più promettente), garantisce maggiori probabilità di trovare un modello più efficace. Meno favorevole risulta invece il confronto con Distill, da cui si evince che Homer non raggiunge ancora un livello tale da poter virtualmente essere inserito nella prima metà della classifica.

Un esempio che può spiegare la differenza di prestazioni è la mancata modellazione da parte di HOMER del secondo dominio della sequenza T0652-D2 (fig 45). Alcuni dei test forniti dal CASP riguardano infatti proteine che presentano più unità che ripiegano indipendentemente.

Come si è già avuto modo di dire nel capitolo 2, normalmente in un singolo file PDB troviamo un solo dominio. Dato che Homer utilizza un solo templato nella previsione della struttura prote-

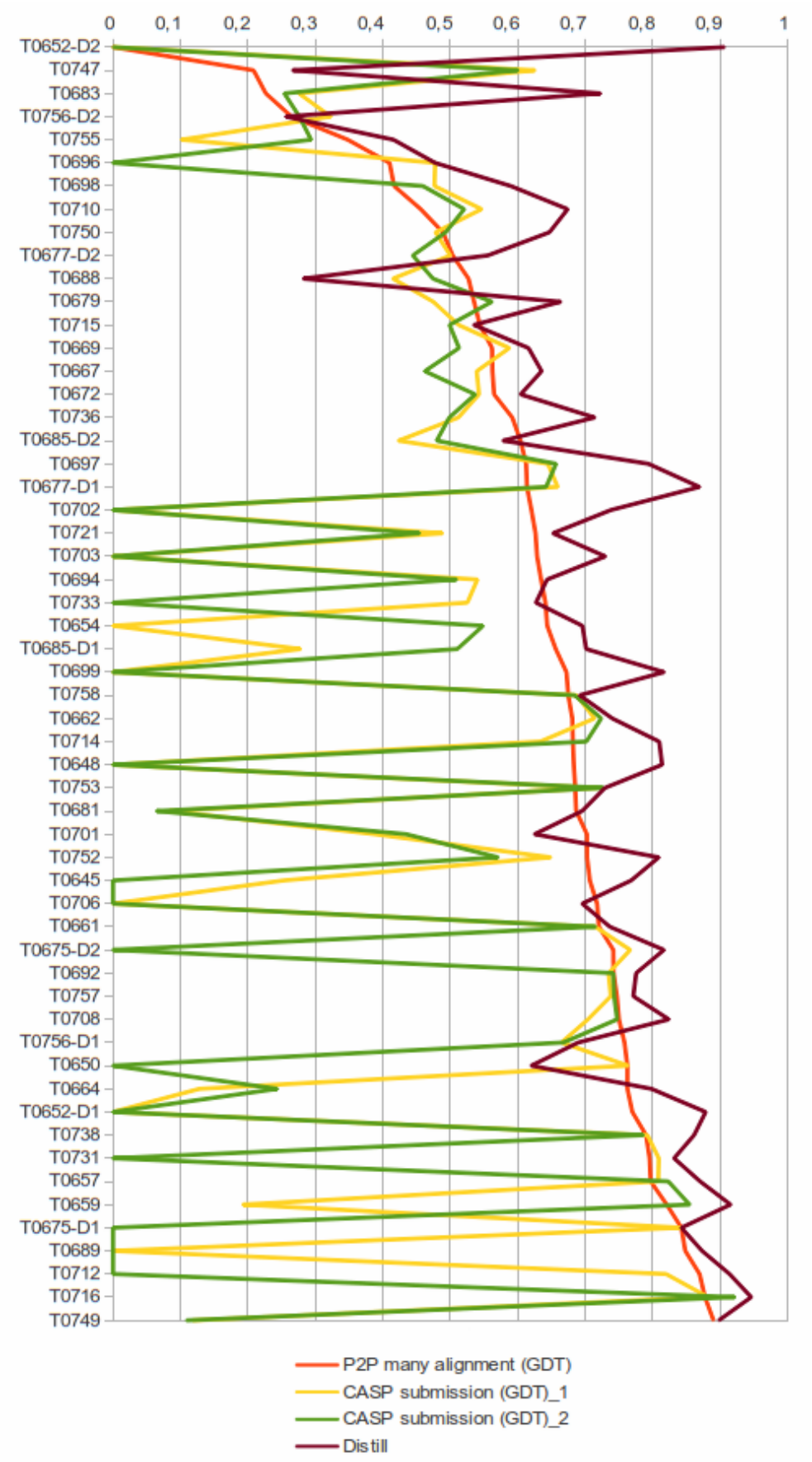

Figura 43: Grafico dei valori di GDT_TS di tabella3

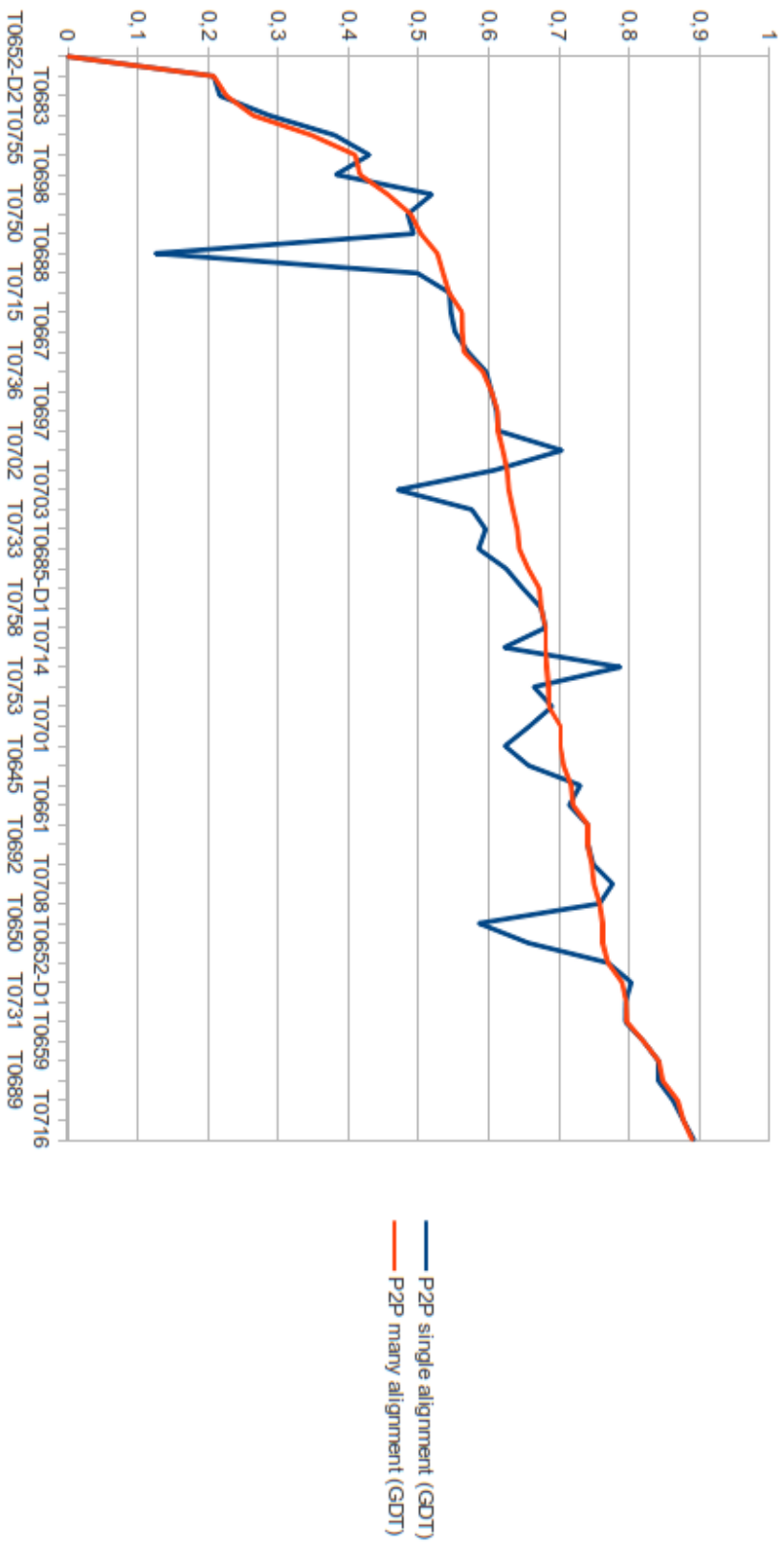

Figura 44: Confronto tra le due versioni di Homer

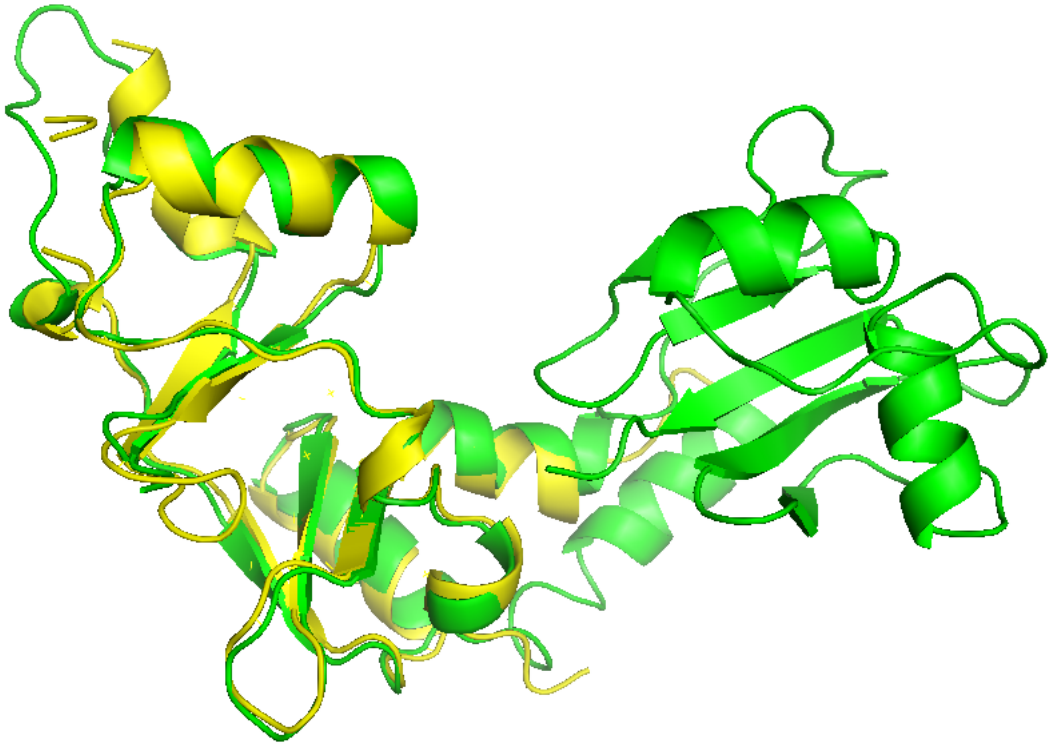

Figura 45: Target T0652 e modello generato da Homer: solo il primo dominio viene modellato

ica, è evidente che quello scelto permette di modellare (bene) solo il primo dominio (GDT_TS di 0,7699), mentre è del tutto inadatto per il secondo (GDT_TS 0). Distill invece riesce in questo caso a modellarli entrambi ed il motivo è che utilizza più di un templato alla volta, in base al cosiddetto approccio *fragment-based*.

15 Conclusioni

Gli obiettivi posti sono stati raggiunti ed il sistema si è dimostrato autonomo e stabile: per ogni sequenza target è stato prodotto un modello. La qualità media dei risultati è sensibilmente migliorata rispetto alla versione a cui è ispirato. La strategia di generare varie combinazioni di allineamento, in particolare profilo contro profilo utilizzando vari weighting scheme, scoring function e gap penalty function si è dimostrata più vantaggiosa rispetto all'iniziale tentativo di individuare la singola tipologia con le migliori performance. Sono inoltre stati individuati i punti per in quali vi è spazio per ulteriori miglioramenti: nella ricerca del templato, nella fase di allineamento e di modellazione dei loop.

15.1 Sviluppi Futuri

Un primo interessante aggiornamento, che sulla carta promette buone possibilità di migliorare la qualità media dei modelli, potrebbe essere quello di considerare anche il numero di residui, oltre all'e-value, nella scelta del templato. In un orizzonte più ampio sarà però necessario garantire il pieno sviluppo all'approccio fragment-based per compiere un ulteriore sostanziale salto di qualità.

Il ricorso a più templati, anziché uno solo come avviene ora, dovrebbe permettere di ampliare la copertura della sequenza target negli allineamenti, con notevoli benefici nella fase di produzione dei pre-modelli soprattutto nel caso di target definiti "difficili", o di proteine che presentano più domini.

Vi è inoltre l'esigenza di una più ampia e completa analisi delle capacità della libreria Align. I benchmarking su cui si sono basate le scelte implementative volte ad automatizzare il processo di allineamento non risultano infatti sufficientemente esaurienti.

Alcune caratteristiche della libreria sono state testate solo in parte o non sono state valutate affatto. In particolare è necessario considerare:

- gli weighting schemes PSIC e SeqDivergence negli allineamenti profilo contro profilo
- le scoring function Patchenko e Zhou
- le costanti e le variabili elaborate nella funzione VGP (con i vari pesi quali wH associato alla propensione strutturale per l'α-elica, wS per i β-sheet, wB per l'accessibilità al solvente, wC per la linearità del backbone, wD per il coefficiente di sepoltura, che sono di fatto non utilizzati).
- le informazioni strutturali Threading e Prof.

Per tutte queste funzioni non sono state infatti trovate considerazioni sulle performance.

In particolare si consiglia di analizzare la classe Prof: essa integra nella produzione degli allineamenti sia informazioni sulla struttura secondaria (attualmente già sfruttate), sia sull'accessibilità al solvente (finora non si vi si è fatto ricorso). Poiché quest'ultimo tipo di informazione viene comunque prodotto per QMEAN, utilizzarlo anche nella fase di allineamento non richiederebbe alcuno sforzo aggiuntivo.

Infine è necessario completare il lavoro di integrazione della libreria LOBO e garantirne l'affidabilità, in quanto anche la modellazione dei loop dovrebbe garantire un ulteriore miglioramento delle strutture prodotte.

A Materiale CASP10

In questa appendice vengono riportate le principali informazioni sui target relativi al CASP10, che sono stati utilizzati in questa tesi per valutare le performance di HOMER.

Nella lista seguente da sinistra a destra vengono riportati: la sigla identificativa del target, il nome del modello reale inserito nella banca dati PDB (non tutte le strutture sono già state inserite), il numero di residui, l'indicazione del metodo sperimentale usato per determinare la struttura, la data di sottomissione nel sito del CASP e quella di termine per la sottomissione dei modelli da parte dei partecipanti.

Target	Name	Nres	Method	Entry	Expiry
T0645	4F7A	537	X-RAY	1/5	4/5
T0648	-	102	X-RAY	2/5	5/5
T0650	4FMZ	346	X-RAY	3/5	6/5
T0652	-	292	X-RAY	4/5	7/5
T0654	4FO5	166	X-RAY	7/5	10/5
T0657	2LUL	154	NMR	8/5	11/5
T0659	4ESN	85	X-RAY	9/5	12/5
T0661	4FCZ	215	X-RAY	10/5	13/5
T0662	2LTE	79	NMR	10/5	13/5
T0664	4F53	540	X-RAY	11/5	14/5
T0667	4FLE	194	X-RAY	14/5	17/5
T0669	2LTL	109	NMR	15/5	18/5
T0672	4F0J	335	X-RAY	16/5	19/5
T0675	2LV2	75	NMR	17/5	20/5
T0677	-	153	NMR	18/5	21/5
T0679	4H08	223	X-RAY	21/5	24/5
T0681	4FXT	224	X-RAY	22/5	25/5
T0683	4EZI	403	X-RAY	23/5	26/5
T0685	4FMT	253	X-RAY	24/5	27/5
T0688	4EZQ	196	X-RAY	25/5	28/5
T0689	4FVS	234	X-RAY	28/5	31/5
T0692	4H7N	473	X-RAY	29/5	1/6
T0694	-	315	X-RAY	30/5	2/6
T0696	-	111	X-RAY	31/5	3/6
T0697	-	483	X-RAY	1/6	4/6
T0698	-	119	X-RAY	1/6	4/6
T0699	-	234	X-RAY	1/6	4/6
T0701	-	322	X-RAY	4/6	7/6
T0702	-	271	X-RAY	5/6	8/6
T0703	4HES	272	X-RAY	6/6	9/6

Target	Name	Nres	Method	Entry	Expiry
T0706	-	217	X-RAY	7/6	10/6
T0708	4H17	196	X-RAY	8/6	11/6
T0710	-	220	X-RAY	11/6	14/6
T0712	4GBS	223	X-RAY	12/6	15/6
T0714	2LVC	88	NMR	13/6	16/6
T0715	-	462	X-RAY	14/6	17/6
T0716	2LY9	71	NMR	14/6	17/6
T0721	4FK1	301	X-RAY	19/6	22/6
T0731	2LZ1	79	NMR	27/6	30/6
T0733	4GGA	390	X-RAY	28/6	1/7
T0736	-	168	X-RAY	2/7	5/7
T0738	-	249	X-RAY	3/7	6/7
T0747	4G5A	121	X-RAY	11/7	14/7
T0749	4GL3	449	X-RAY	12/7	15/7
T0750	-	188	X-RAY	12/7	15/7
T0752	4GB5	156	X-RAY	13/7	16/7
T0753	4GOG	109	X-RAY	13/7	16/7
T0755	4H1X	264	X-RAY	16/7	19/7
T0756	4G6G	179	X-RAY	16/7	19/7
T0757	4GAK	247	X-RAY	17/7	20/7
T0758	-	388	X-RAY	17/7	20/7

Tabella 5: Lista dei target del CASP10 assegnata a predittori automatici (“server only”)

Riferimenti bibliografici

[1] **Orengo C., Michie A., Jones S., Jones D, Swindells M., Thornton JM.** (1997) CATH - a hierarchic classification of protein domain structures. Structure 5:1093-1108.

[2] **Soding J., Remmert M.** (2011) Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Current Opinion in Structural Biology 21:404-411.

[3] **Henikoff S., Henikoff J.G.** (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci USA, 85, 2444-2448.

[4] **Needlemann S.B. and Wunsh C.D.** (1970) A general method applicable to search for similarities in the amino acid sequence of two proteins. J. Mol. Biol, 48:443-453.

[5] **Smith T.F. and Watermann M.S.** (1981) Identification of common molecular subsequences. J. Mol. Biol, 1981.

[6] **Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J.** (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. 25:3389-3402.

[7] **Moult J., Fidelis K., Kryshtafovych A., Tramontano A.** (2011) Critical assessment of methods of protein structure prediction (CASP) - round IX. Proteins 79(Suppl 10):1-5.

[8] **Thiella V.** (2009/2010) Valutazione della predizione della struttura proteica. l'iniziativa CASP. Padua@thesis, http://tesi.cab.unipd.it/.

[9] **Walsh I., Minervini G., Corazza A., Esposito G., Tosatto S.C.E., Fogolari F.** (2012) Bluues Server: electrostatic properties of wild-type and mutated protein structures. Bioinformatics. 28(16):2189-90.

[10] **Rychlewski L., Jaroszewski L., Li W., Godzik A.** (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci. 9:232-241.

[11] **Wang G., Dunbrack R.L.jr.** (2004) Scoring profile-to-profile sequence alignments. Protein Science, 13:1612-1626.

[12] **Tosatto S.C.E., Albiero A., Mantovan A., Ferrari C., Bindewald E., Toppo S.** (2006) Align: a C++ Class Library and Web Server for Rapid Sequence Alignment Prototyping. Curr Drug Discov Technol 3(3):167-73.

[13] **Negri E.** (2007/2008) Metodologie profilo-profilo per l'allineamento di sequenze proteiche. Padua@thesis, http://tesi.cab.unipd.it/.

[14] **Jones D.T.** (1999) Protein secondary structure prediction based on position-specific scoring matrices. J.Mol.Bio. 292(2):195-202.

[15] **Henikoff S., Henikoff J.G.** (1994) Position-based sequence weights. J. Mol. Biol. 243:574-578

[16] **Sunyaev S.R., Eisenhaber F., Rodchenkov I.V., Eisenhaber B., Tumanyan V.G., Kuznetsov E.N.** (1999) PSIC: Profile extraction from sequence alignments with position-specific counts of indipendent observations. Protein Eng. 12:387-394.

[17] **Tosatto S. C. E., Toppo S.** (2006) Large-Scale Prediction of Protein Structure and Function from Sequence. Current Pharmaceutical Design, 12, 2067-2086.

[18] **Shapovalov M.V., Dunbrack R.L.jr.** (2001) A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. Structure 19, 844-858.

[19] **Canutescu A.A., Shelenkov. A.A., Dunbrack R.L.jr.** (2003) A graph theory algorithm for rapid protein side-chain prediction. Protein Science 12(9):2001-2014.

[20] **Benkert P., Tosatto S. C. E., Schwede T.** (2009) Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. Proteins 77 (Suppl 9): 173-180.

[21] **Benkert P., Schwede T., Tosatto S. C. E.** (2009) QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Structural Biology 9:35.

[22] **Benkert P., Tosatto S. C. E., Schomburg D.** (2008) QMEAN: A comprehensive scoring function for model quality assessment. Proteins 71: 261-277.

[23] **Benkert P., Kunzli M., Schwede T.** (2009) QMEAN server for protein model quality estimation. Nucleic Acids Research, Vol 37, Web Server Issue W510-W514.

[24] **Tosatto S. C. E.** (2005) The Victor/FRST Function for Model Quality Estimation. Journal of Computational Biology, 12(10): 1316-1327.

[25] **Tosatto S.C.E., Bindewald E., Hesser J., Männer R.** (2002) A divide and conquer approach to fast loop modeling. Protein Engineering 15(4):279-286.

[26] **Ohsen N., Sommer I., Zimmer R., Lengauer T.**(2004) Arby: automatic protein structure prediction using profile-profile alignment and confidence measures. Bioinformatics Vol. 20, No. 14, p. 2228-2235.

[27] **Durbin R., Eddy S.R., Krogh A., Mitchison G.** (1998) Biological sequence analysis. Cambridge University Press.

[28] **Zhang Y., Skolnick J.** (2004) Scoring Function for Automated Assessment of Protein Structure Template Quality. Proteins 57:702-710.

[29] **Albiero A.** (2004/2005) Allineamenti inversi e selezione energetica di strutture proteiche. Tesi, Università degli Studi di Padova.

[30] **Moro A.** (2006/2007) Valutazione su larga scala di predizioni alternative di strutture proteiche. Tesi, Università degli Studi di Padova.

[31] **Fantato M.** (2010) Un algoritmo genetico per la predizione della configurazione spaziale del nucleo idrofobico di proteine. Padua@thesis, http://tesi.cab.unipd.it/.