Università degli studi di Padova
Corso di Laurea in Ingegneria Informatica

Tesi di Laurea

Valutazione della predizione della struttura proteica : l'iniziativa CASP

Relatore:
Prof. Ferrari Carlo

Laureando:
Thiella Valeria

Anno Accademico 2009-2010

Ai miei genitori

Indice

Introduzione

Il contenuto di questa tesi è una breve introduzione alle proteine in generale e al problema della predizione di strutture proteiche.
Segue poi un analisi dei metodi di predizione e delle metriche di valutazione dei modelli nei confronti della proteina nativa.
L'ultima parte è un resoconto delle strategie che ebbero successo nelle ultime due edizioni del CASP.

Capitolo 1
Ripiegamento proteico.

1.1 Brevi cenni sulle proteine

Le proteine sono composti organici di fondamentale importanza per gli organismi viventi e sono caratterizzati da una sequenza ben precisa di residui aminoacidici uniti mediante un legame peptidico.
Gli aminoacidi presenti negli organismi viventi sono numerosi ma solo una piccola parte di essi (tutti della serie stereochimica L) sono sottoposti al controllo genetico, come conseguenza dei processi evolutivi, e contenuti nelle proteine: se ne utilizzano comunemente 20, indicati tipicamente attraverso una sigla di tre (o una) lettere, come riassunti in Tabella

acido aspartico	ASP (D)	acido glutammico	GLU (E)
alanina	ALA (A)	arginino	ARG (R)
asparagina	ASN (N)	cisteina	CYS (C)
fenilalanina	PHE (F)	glicina	GLY (G)
glutammina	GLN (Q)	isoleucina	ILE (I)
istidina	HIS (H)	leucina	LEU (L)
lisina	LYS (K)	metionina	MET (M)
prolina	PRO (P)	serina	SER (S)
tiroxina	TYR (Y)	treonina	THR (T)
triptofano	TRP (W)	valina	VAL (V)

Tabella: elenco degli aminoacidi con le loro sigle.

Dal punto di vista chimico, un aminoacido è caratterizzato da un carbonio centrale, detto carbonio alfa (C_α), un gruppo amminico (NH_2) ed uno carbossilico (COOH) legati al medesimo atomo di carbonio C_α. Oltre a tali gruppi “fissi”, ogni aminoacido presenta uno specifico gruppo laterale (detto gruppo R).
In funzione delle proprietà chimiche del gruppo R, un aminoacido viene classificato come acido, basico, idrofilo (o polare) o idrofobo (o apolare).
L’ingombro dei vari gruppi R che sporgono dalla catena polipeptidica, l’affinità reciproca tra gruppi polari e tra gruppi apolari, l’attrazione tra gruppi basici e gruppi acidi sono alcune delle forze che 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.

1.2 Il legame peptidico

Una caratteristica importante degli aminoacidi è che essi possono legarsi tra loro attraverso un legame forte definito legame peptidico.
Questo tipo di legame si forma tra il gruppo $–NH_2$ e il gruppo –COOH di due aminoacidi adiacenti con rimozione di una molecola d'acqua. L'insieme di più aminoacidi costituisce un polipeptide.
Gli atomi di carbonio e azoto legati con un legame peptidico costituiscono la catena principale di peptidi. La costituzione degli aminoacidi fa si che tale catena presenti ad una estremità un aminoacido con un gruppo carbossilico libero, detto C terminale e all' altra uno con un gruppo amminico libero, detto N terminale.
I residui R di ciascun aminoacido costituiscono le catene laterali.
Il legame peptidico è un legame estremamente rigido, d'altro canto i due legami contigui ad esso (il C_α-COOH e il NH- C_α) possono compiere rotazioni, formando due angoli, rispettivamente Ψ (Psi) e Φ (Phi); questi due angoli teoricamente possono variare da -180° a +180°, anche se in pratica alcuni di questi valori sono proibiti per delle interferenze che si possono venire a creare tra le catene laterali degli aminoacidi.
In figura è illustrata una catena di aminoacidi coinvolti in un legame polipeptidico.

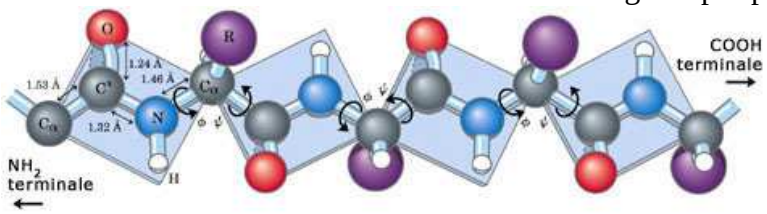


Figura: Un polipeptide.

1.3 Livelli di struttura

Nelle proteine riconosciamo fino a quattro livelli di struttura:
1. struttura primaria. Corrisponde alla specifica sequenza degli aminoacidi legati in una catena polipeptidica.
2. struttura secondaria. In base alla natura degli aminoacidi e agli angoli di legame, il polipeptide può assumere localmente conformazioni stabili più complesse, che vengono definiti motivi di struttura secondaria.
3. struttura terziaria. Rappresenta l'ulteriore ripiegamento della proteina fino a formare strutture tridimensionali prodotte a causa delle interazioni tra aminoacidi posti in punti diversi del polipeptide. Corrisponde alla struttura assunta dalla proteina quando essa si trova nel cosiddetto stato nativo.
4. struttura quaternaria. Riguarda in generale proteine molto grandi. In effetti, spesso tali proteine sono costituite da varie sub-unità essenzialmente uguali fra loro.
La struttura quaternaria riguarda la disposizione spaziale e topologica di queste sub-unità.

Strutture secondarie

Esistono diversi tipi di strutture secondarie, le più comuni sono α-eliche e foglietti-β (o β-sheet).
α-elica: è un'elica destrorsa formata dall'avvolgimento attorna ad un asse centrale immaginario del polipeptide.
In questa elica le catene laterali degli aminoacidi sono proiettate all'esterno. L' α-elica ha un'unità ripetitiva di 5.4Å, ed è stabilizzata al suo interno da interazioni non covalenti (legami idrogeno).
Foglietto-β: è una struttura più estesa e si forma quando tra le catene principali di alcuni tratti del polipeptide si stabiliscono legami ponte a idrogeno, che costringono le catene a giacere sullo stesso piano. Si distinguono due tipi di foglietto β: parallelo o antiparallelo a seconda della direzionalità dei segmenti che lo compongono. Generalmente l'antiparallelo è più stabile.

Loops

Le strutture secondarie sono unite attraverso sequenze aminoacidiche che disegnano regioni dette loops. Sebbene alcune di tali regioni possono essere molto lunghe (fino a ventuno aminoacidi), nella maggior parte dei casi esse sono composte da due fino a dieci aminoacidi. Gli hairpin loops, i loops più corti conosciuti (2-5 aminoacidi) vengono chiamati "riverse –turns" per la loro proprietà di collegare due foglietti β adiacenti eseguendo un inversione nella direzione della sequenza. I loops si trovano, generalmente, sulle regioni esterne della proteina e sono di conseguenza costituite da catene laterali per lo più idrofiliche. Inoltre i legami idrogeno tra gli aminoacidi del loop e le molecole di acqua circostanti sono in numero maggiore rispetto a quelli effettuati con gli aminoacidi adiacenti; ciò conferisce a tali regioni una maggiore flessibilità rispetto alle strutture secondarie.

Strutture terziarie

Tra le catene laterali si stabiliscono interazioni che portano a ripiegamenti ulteriori rispetto a quelli dati dalle strutture secondarie, la così detta struttura terziaria.
Tali interazioni sono di tipo debole tra aminoacidi idrofobici, interazioni dipolari tra aminoacidi con carica opposta, legami ponte idrogeno o legami a ponte disolfuro. 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.

Legami disolfuro

I legami disolfuro si formano tra atomi di zolfo facenti parte delle catene laterali di due cisteine. Il legame, che prende il nome di ponte disolfuro tipicamente non si forma sulla superficie della proteina, a causa della presenza nel citoplasma di agenti riducenti. Questi legami sono fondamentali nella determinazione della forma della proteina. Molte proteine strutturali, che richiedono un elevata stabilità, sono caratterizzate da un alto numero di legami di questo tipo. Un esempio è dato dalla cheratina che si trova nelle unghie.

1.4 Il Protein Folding

Il ripiegamento di proteine è il complesso di fenomeni attraverso cui le proteine ottengono la loro struttura tridimensionale. Tale processo può essere descritto come un auto-assemblamento intramolecolare in cui la proteina assume una specifica forma attraverso interazioni non covalenti, come legami ad idrogeno, forze idrofobiche, forze di Van der Waals, interazioni π-π.

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. La struttura primaria di una proteina è insufficiente per capire la funzione che essa svolge all' interno della cellula, nonostante il livello di complessità dello studio della struttura primaria sia già elevato. Questo è il motivo per cui il ripiegamento proteico è oggetto di ricerca.

Protein structure prediction

Le proteine presentano un ben preciso stato nativo che assumono mediante un rapido processo di ripiegamento (1μs-1s) che è in genere reversibile, esiste quindi una corrispondenza biunivoca tra la sequenza degli aminoacidi che compongono una proteina e il suo stato nativo-si dice che " lo stato nativo di una proteina è codificato nella sua sequenza primaria". Il problema del protein structure prediction si riferisce al problema di predire la struttura tridimensionale di una proteina data la sua sequenza primaria. In quest'ottica si stanno creando numerosi programmi in grado non solo di generare modelli tridimensionali ma anche di valutare l'energia libera di tali modelli.

1.5 Valutazione energetica

In natura la struttura più stabile in cui si presenta una sequenza peptidica è quella che presenta il valore energetico più basso. Quindi data una sequenza esisteranno un numero infinitamente grande di strutture con un valore energetico alto ma solo una struttura avrà un livello energetico minimo e sarà definita struttura nativa (in questo caso la nativa è rappresentata dalla struttura tridimensionale risolta con metodi sperimentali). Su questo principio si basano gli algoritmi per la valutazione energetica.

Non è corretto dire che solo una struttura avrà energia minima infatti la maggior parte delle proteine per svolgere la loro attività devono possedere una certa flessibilità strutturale, è corretto quindi dire che ci saranno una serie di strutture alternative, molto simili tra loro, che avranno un valore di energia complessiva molto basso e nell'intorno del minimo globale.

La teoria del fold prevede che una struttura possa presentare un profilo energetico molto irregolare che compreda molti minimi locali ma sia caratterizzato da un solo minimo globale. Quest'ultimo rappresenta il valore energetico della nativa, quindi, conoscendo a priori tale valore di energia sarebbe possibile valutare oggettivamente un modello.

Purtroppo questo valore non è mai noto a priori, per cui la valutazione energetica avverrà per comparazione tra un numero n di modelli, generati ad esempio utilizzando approcci di allineamento che producono diverse alternative, tra le quali sia possibile valutare la migliore, cioè quella che presenta un'energia globale minore. In effetti la tendenza attuale, soprattutto in ambito di Comparative Modeling e Fold Recognition, è quella di produrre molti modelli alternativi e di valutare tra questi quale sia quello che

più si avvicina alla nativa. Per fare ciò si sono sviluppati due metodi di valutazione energetica: il metodo fisico e il metodo statistico.
I metodi fisici si basano sulla ricerca di modelli matematici che approssimino il calcolo dell'energia reale della struttura, permettendo così di ottenere un valore reale di energia. I problemi di questo metodo sono legati principalmente ai tempi di calcolo e alla sensibilità (percentuale di predizioni corrette dal metodo sulla totalità di quelle realmente presenti).
Per quanto riguarda i metodi statistici, invece, il modello viene valutato in base alla sua vicinanza alla struttura più presente in natura.Questi metodi non restituiscono valori di energia, ma forniscono valori di pseudo –energia. Tali risultati hanno significato solo in senso comparativo (discriminano quale sia la struttura migliore in una serie di strutture alternative), inoltre sono meno sensibili rispetto ai metodi fisici. Tuttavia nonostante i limiti la loro velocità di calcolo ne ha permesso una maggiore diffusione.
La differenza in numero tra le proteine la cui formula chimica è conosciuta (sequenziate) e quelle con forma tridimensionale nota è di almeno tre ordini di grandezza. Infatti attraverso il sequenziamento di interi genomi e tecniche in silico di individuazione e traduzione dei geni è stato possibile sequenziale un gran numero di proteine prima sconosciute.
Per questo motivo la ricerca, soprattutto bioinformatica, impiega sempre più energie in questo campo.

Stati molten globule [6]
E' stata dimostrata l'esistenza di stati stabili in cui la proteina è parzialmente ripiegata che si verificano in particolari condizioni (per lo più legati a ph e temperatura). Tali stati, chiamati globuli fusi (o molten globules) evidenziano una struttura secondaria simile a quella della proteina nativa ed una struttura terziaria poco compatta, mancante di rigidità e pertanto particolarmente dinamica, come se si trovasse allo stato liquido.

Capitolo 2
Predizione di strutture proteiche

2.1 Strutture tridimensionali e CASP

Il CASP (Critical Assessment of Techniques for Protein Structure Prediction) è una competizione biennale che valuta lo stato della ricerca sul ripiegamento proteico.In sostanza laboratori e gruppi di ricerca si sfidano cercando di predire la struttura (ripiegata) di proteine i cui dati sperimentali (non teorici) non sono ancora stati pubblicati e sono noti solo agli organizzatori .
Di fatto per la risoluzione della struttura tridimensionale di una proteina esistono solo due approcci fondamentali: uno sperimentale ed uno teorico-computazionale.
I metodi sperimentali classici sono la cristallografia a raggi x e la spettroscopia a risonanza magnetica e nucleare (Nuclear Magnetic Resonance, NMR) e producono modelli molto affidabili ma richiedono spesso tempi molto lunghi, condizioni di reazione particolari e costi molto elevati.
Un trattamento computazionale gode invece di tempi e costi ridotti, inoltre non richiede condizioni di reazione particolari. Non è richiesta nemmeno la “presenza” della proteina da analizzare,ma solamente la sua sequenza primaria.
Lo svantaggio è che per il momento non sempre le predizioni che vengono prodotte sono affidabili.
I gruppi di ricerca che partecipano al CASP sono suddivisi in due categorie: quelli che prevedono l’intervento umano e quelli in cui i calcoli sono completamente automatizzati.
Alcuni di questi gruppi di ricerca fanno uso del calcolo distribuito volontario: Rosetta@home e POEM@home (Protein Optimization with Energy Methods) su piattaforma BOINC per la categoria “server” , FoldIt per quella “human”.
Alcune proteine test vengono specificatamente indirizzate per una o l’altra di queste categorie, anche se nessuno impedisce ad un gruppo di ricerca di cimentarsi in tutti i problemi sottoposti. Le diverse predizioni vengono valutate da assessors indipendenti sulla base delle strutture sperimentali. La premiazione avviene durante una conferenza in cui vengono esposti i metodi utilizzati per le predizioni. Lo scopo primo della conferenza è di individuare quali siano i metodi migliori e quindi verso quali ambiti debba essere indirizzata la ricerca bioinformatica anche in vista dell’edizione successiva. Tutti i risultati vengono pubblicati l’anno successivo nella rivista Proteins.

2.2 Accenno sugli allineamenti

Attualmente per la predizione di una struttura proteica l’approccio più utilizzato è quello di individuare una o più strutture templato (sequenze con struttura nota) da utilizzare come base per la creazione di modelli. Il tutto parte dalla generazione di vari allineamenti tra la sequenza target e la sequenza templato stessa. Generare vari allineamenti permette di ottenere un numero variabile di strutture della sequenza target, tra le quali ricercare quella più stabile, attraverso il confronto diretto con la struttura nativa o utilizzando programmi per la valutazione energetica.

L’allineamento ha lo scopo di individuare le regioni del target che sono strutturalmente uguali o identiche al templato, in modo da sfruttare direttamente le informazioni. La problematica è relativa al fatto che l’allineamento avviene tra le sequenze amminoacidiche mentre servirebbe un allineamento strutturale. L’allineamento strutturale può essere solamente approssimato dall’allineamento di sequenze, in quanto quest’ultimo è guidato da parametri che non considerano in maniera diretta elementi strutturali. Concludendo, il miglior allineamento tra sequenze non sempre coincide con il miglior allineamento strutturale.
L’allineamento consiste nella determinazione di residui equivalenti tra le sequenze target e templato, onde effettuare una valutazione si ricorre spesso a matrici di sostituzione amminoacidica che assegnino un punteggio ai diversi accoppiamenti.
Alcune regioni del target possono non venire allineate con il templato e devono quindi essere ricostruite completamente.
Una volta ricostruita la catena principale della proteina target, ove non siano state già inserite, si procede alla costruzione delle catene laterali.
I metodi di posizionamento delle catene si basano solitamente sui rotameri considerando le collisioni tra gli atomi delle catene e della backbone.
Il metodo più utilizzato per il posizionamento delle catene laterali è SCWRL, che utilizza appunto i rotameri ed un algoritmo che sfrutta la teoria dei grafi.

2.3 Metodi di predizione

Oggi il CASP è suddiviso in numerose categorie, mentre il CASP originale ne prevedeva solamente tre. Le categorie originali rappresentavano, e si può dire rappresentano tuttora, i metodi con cui si determina la struttura di una proteina.
1) Comparative Modeling: è un metodo che viene utilizzato quando sono note strutture molto simili a quella da modellare. In questa metodica non si fa altro che ricalcare le coordinate atomiche di una proteina templato (proteina stampo per la costruzione di un modello), andando eventualmente a ricostruire le parti mancanti della proteina target (proteina da modellare,la cui struttura è ignota). Il limite di identità fra le sequenze deve essere almeno intorno al 30-40 %, limite derivato da osservazioni tra percentuali di identità e sovrapposizione strutturale [9]. La ricerca dei templati, nel caso del Comparative Modeling, viene effettuata mediante algoritmi di ricerca su banca dati quali Blast o HMMER.
2) Fold Recognition: in questa tecnica, a differenza della precedente, la proteina target non presenta una similarità significativa con nessuna possibile proteina templato presente in una banca dati di strutture note. La soluzione in questo caso, è data dall’osservazione che i fold differenti, individuati nelle proteine di cui finora si conosce la struttura tridimensionale, non superano qualche centinaio (threading). Pertanto è possibile, considerando simultaneamente una serie di parametri tratti dalla struttura primaria della proteina target, come la predizione della relativa struttura secondaria, stimare le strutture tridimensionali più probabili, valutando ognuno dei parametri in termini di score parziale e utilizzando un metodo di ranking (una funzione pesata dei vari contributi al punteggio finale).
3) Ab Initio: mentre le tecniche citate precedentemente prevedono l’utilizzo di una o più sequenze templato, e sono abbastanza simili da essere ormai considerate come un

unico metodo, le metodologie Ab Initio cercano di ricostruire la struttura di una proteina utilizzando solo la sequenza primaria e applicando leggi chimico- fisiche.
Questi metodi non danno ancora risultati accettabili in quanto la teoria del fold non è ancora sufficientemente spiegata.

2.4 PDB

PDB (Protein Data Bank) è un archivio di strutture tridimensionali, sia proteiche che acidi nucleici, fondato nel 1971 dal Brookhaven National Laboratory. Rappresenta il deposito centrale dei dati biologici di strutture tridimensionali, ottenute soprattutto grazie ai metodi sperimentali.
Ad oggi si contano svariati database secondari e altrettanti progetti che si sono sviluppati per interagire e classificare i dati contenuti nel PDB in termini di struttura, funzione ed evoluzione delle proteine.
La maggior parte delle strutture proteiche depositate sono strutture proteiche risolte con raggi X; esiste anche una sezione dedicata ai modelli teorici, ma rappresenta solo una piccola percentuale sul totale delle strutture depositate.
I dati contenuti sono immagazzinati in file specificatamente creati per lo scopo, i quali contengono informazioni sull'esatta posizione spaziale di tutti gli atomi che compongono una grande biomolecola.
Il formato di questi file è il pdb: grandi file di testo organizzati in colonne dove vengono riportati tutti i dati relativi alla struttura (organizzazione, coordinate atomiche, qualità della struttura, ecc…).

2.5 Algoritmi di valutazione modello vs. nativa [1]

La questione più importante che ogni CASP assessor deve trattare è la scelta di uno schema di punteggio e delle metriche appropriate per confrontare modelli e struttura nativa. Siccome nessuna misura è migliore rispetto alle altre in tutti i casi, nel CASP se ne usano un certo numero per ottenere una stima della qualità di un modello.

- RMSD: Root-Mean-Sqare Deviation, usata in CASP 1-3, adatta al confronto di strutture simili in quanto piccole differenze locali di struttura possono risultare in alti valori di RMSD.

Questo parametro viene calcolato come: RMSD = $\sqrt{\frac{\sum (r_{ai} - r_{bi})^2}{n}}$

dove r_{ai} e r_{bi} sono le posizioni dell'atomo i nella struttura a e nella struttura b, mentre n è il numero di atomi nelle strutture. Se è uguale a zero significa che le due strutture sono identiche, contrariamente, più si discosta da questo valore più le strutture saranno differenti.

- GDT_ TS: Global Distance Test_Total Score. Usato per la prima volta dai CM (comparative modeling) assessor nel CASP 4 per superare i limiti di RMSD questo parametro misura la sovrapposizione media del modello sulla nativa e restituisce un valore compreso tra 0 e 1.

Viene calcolato come GDT_TS =(GDT_P1+ GDT_P2+GDT_P4+GDT_P8)/4

GDT_Pn= % C_α WITHIN nÅ, cioè la percentuale di residui con soglie di distanza a nÅ dopo quattro diverse sovrapposizioni LGA in modo sequence-dependent.
Assume valori elevati per modelli che riproducono perfettamente la conformazione della catena principale del target.
E' una delle misure più appropriate per la valutazione complessiva della qualità di un modello. Usato dagli assessors di tutti gli esperimenti CASP dopo CASP 4. Nel CASP6 e CASP7 una modifica di GDT_TS, GDT_HA, fu usata dagli assessors dei target della categoria high accuracy (TBM). GDT_HA usa soglie a 0.5, 1.2 e 4Å.

- ALO: percentuale di residui correttamente allineati dopo sovrapposizione LGA del modello con la struttura sperimentale del target. Un residuo del modello è considerato correttamente allineato se il suo atomo C_α è entro 3.8Å dalla posizione del corrispondente atomo della struttura sperimentale e non c'è nessun altro atomo C_α più vicino. Anche se diverso dal GDT_TS , queste due misure sono altamente correlate.

In anni recenti sono state sviluppate altre misure:

- TM-score: è l'acronimo di Template Modeling Score, è un algoritmo che misura la similarità strutturale tra la struttura nativa e il modello in esame, assegnando pesi diversi a seconda delle distanze. Il valore in output è compreso tra 0 e 1, ma già valori minori o uguali a 0.2 indicano che la struttura predetta è paragonabile ad una qualunque nativa selezionata a caso all'interno di un database come PDB.
- MaxSub-score: è un algoritmo il cui scopo ultimo è produrre un valore normalizzato tra 0 e 1, che rappresenti la qualità della predizione, attraverso l'identificazione del più vasto insieme di C_α del modello che si sovrapponga con i C_α della nativa, a meno di 3.5 Å di RMSD.

Scoring scheme [8]

Per l'estrapolazione di informazioni statistiche dai dati numerici i parametri statistici considerati sono: correlazione, z-score.
Nel CASP si usa la seguente semplice regola per assegnare uno score a un modello. Sia X il parametro selezionato ($X \in$ {GDT, RMSD, ALO }), $\bar{X}$ e σ(X) la media e la deviazione standard di X su tutte le predizioni per un dato target.
Si escludono inizialmente predizioni con valori pessimi di X:
se $X \in$ {GDT, ALO }, eliminare tutti i valori $X < \bar{X} - 2\,\sigma(X)$
se $X \in$ { RMSD}, eliminare tutti i valori $X > \bar{X} + 2\sigma(X)$
Si ricalcala poi $\bar{X}$ e σ (X) sulle rimanenti predizioni e si assegna uno score come
Score (X) = [(X- $\bar{X}$)/(0.5 σ(X))] (% PREDICTED), dove %PREDICTED è la percentuale della struttura che è presente nella predizione.

Attualmente il numero di categorie nel CASP può essere riassunto come segue:

- Modelli con templati.

- Modelli senza templati (“ Ab Initio “).
- Contatti (determinazione aminoacidi lontani nella sequenza ma vicini strutturalmente)
- Individuazione dei domini strutturali di proteine complesse.
- Predizione del disordine (regioni in cui i metodi sperimentali non forniscono modelli accurati).
- Determinazione della funzione.
- Qualità dei modelli (dato un modello riuscire a stabilirne la qualità senza avere a disposizione la struttura determinata sperimentalmente).
- Metodi di raffinamento CASP-R (la struttura determinata sperimentalmente è nota al momento della pubblicazione dei modelli da raffinare).

Capitolo 3
CASP 7-8

3.1 Valutazione di modelli Template-Based nel CASP 8 [1]

Molte misure di valutazione numerica possono dare una stima ragionevole della similarità tra un modello e la corrispondente struttura sperimentale.
Non in tutti i casi esse possono essere usate direttamente per classificare i modelli secondo la loro accuratezza. Per esempio modelli di target per i quali non possono essere identificate strutture templato possono essere piuttosto lontani dalla struttura sperimentale e perciò raggiungere score molto bassi. Però una ispezione attenta può mettere in evidenza casi dove questi modelli riproducono importanti caratteristiche della proteina target – ripiegamento complessivo, appropriati arrangiamenti di struttura secondaria, corretti contatti inter-residue e così via.
Comunque, per predizioni basate su templati, le valutazioni numeriche danno informazione sufficiente per confrontare la qualità dei modelli e perciò valutare l'efficacia dei corrispondenti metodi di predizione.

Procedure di classificazione

Nel CASP i dettagli specifici degli scoring schemes sono lasciati agli assessors. Nei CASP precedenti, gli approcci usati dai TBM assessors-formalmente CM (comparative modeling) e FR (fold recognition) assessors- differivano nella scelta delle seguenti alternative
1. usare tutti i modelli sottomessi alla valutazione per il calcolo della media e della deviazione standard (necessario per lo z-score) oppure ignorarne alcuni.
2. fissare a zero i valori di z-score negativi oppure no.
3. usare la somma degli z-score oppure la media sul numero di target predetti per la classificazione .
4. usare z-scores da una singola misura di valutazione (tipicamente GDT_TS) come base per lo schema di classificazione oppure combinare z-scores da indipendenti misure di valutazione .

Uno dei problemi nell'uso degli z-score è che la distribuzione dei valori di GDT_TS (ad esempio) relativi a tutti i modelli per un target può essere influenzata da modelli pessimi (ad esempio modelli molto corti sono irrealistici). Per eliminare l'effetto di tali modelli alcuni valori possono essere esclusi dall'insieme di dati usati per il calcolo dei valori finali di media e deviazione standard.
Per scopi di classificazione, gli z-score dei modelli sottomessi da ogni gruppo possono essere sommati o mediati sul numero di domini predetti. La somma penalizza i gruppi che non hanno sottomesso modelli per tutti i target, la media può penalizzare coloro che hanno sottomesso modelli per un gran numero di target. La Tabella 1 mostra la correlazione tra gli z-scores ottenuti usando GDT_TS , GDT_HA e ALO per i gruppi che parteciparono al CASP8. Essi sono altamente correlati per entrambi gli insiemi di target ("human and server " e "server only") perciò nel seguito sono discussi solo i risultati di GDT_TS. L'ufficiale gruppo di valutazione della categoria TBM ha sempre

usato il modello indicato come “model 1” da parte dei gruppi e continuò a farlo anche per il CASP8.

Tabella 1:

Dataset			ρ
All groups	Mean ALO Z-score	Mean GDT-TS Z-score	0.97
Human and server targets	Mean ALO Z-score	Mean GDT-HA Z-score	0.96
	Mean GDT-TS Z-score	Mean GDT-HA Z-score	0.99
Server groups	Mean ALO Z-score	Mean GDT-TS Z-score	0.97
All targets (human and server plus server only)	Mean ALO Z-score	Mean GDT-HA Z-score	0.95
	Mean GDT-TS Z-score	Mean GDT-HA Z-score	0.98

Correlazione di Spearman (ρ) tra gli z-score ottenuti da ciascun gruppo usando misure differenti.I dati sono riportati sia per il sottoinsieme “human and server” e per il completo insieme di target.

L’ufficiale gruppo di valutazione della categoria TBM ha sempre usato il modello indicato come “model 1” da parte dei gruppi e continuò a farlo anche per il CASP8.

Valutazione dei modelli TB nel CASP 8

Nell’analisi furono adottati i seguenti parametri:

1.GDT_TS fu usata come misura base per confrontare modelli e strutture sperimentali
I valori di GDT_TS sono calcolati usando LGA (Local- Global- Alignment) in
Modo sequence-dependent.

2.i modelli con meno di venti residui furono eliminati dall’insieme di dati.

3.i GDT_TS z-scores furono calcolati dopo aver eliminato i modelli con valori di GDT_TS inferiori rispetto alla media di una quantità maggiore di due volte la deviazione standard.

4.valori negativi di z-score furono fissati a zero.

5.i gruppi furono classificati secondo il mean GDT_TS z-score per i modelli designati come “model 1” dai predictors.

6.la gaussianità delle distribuzioni dei valori per ogni target fu valutata usando il Shapiro Wilk test.

7.il significato statistico delle differenze tra i valori di GDT_TS dei modelli fu valutato con un test di ipotesi per tutte le coppie di gruppi sull’insieme comune di target predetti. Un potenziale problema è che il t-test è basato sull’ipotesi di normalità delle distribuzioni che devono essere confrontate e bisognerebbe verificare che questo è il caso nell’esperimento altrimenti dovrebbe essere usato un test non parametrico – come il Wilcoxon signed rank test.

Tale test restituisce il valore della probabilità che le distribuzioni dei risultati dei due gruppi corrispondenti siano statisticamente indistinguibili.

I gruppi statisticamente indistinguibili sono poi identificati usando un cut-off di 0.01 per tale valore di probabilità.

Risultati

Tabella 2:

			"Human and Server" target subset		All targets		
Rank	Group name	Group id	No. of targets	Mean Z-score	No. of targets	Mean Z-score	Rank
1	IBT_LT	283	64	1.11			
2	DBAKER	489	64	1.03			
3	Zhang	71	64	0.94			
4	Zhang-Server	426	64	0.84	514	0.89	1
5	KudlatyPredHuman	267	18	0.83			
6	TASSER	57	64	0.83			
7	fams-ace2	434	64	0.83			
8	ZicoFullSTP	196	64	0.81			
9	SAM-T08-human	46	62	0.80			
10	Zico	299	64	0.78			
11	MULTICOM	453	64	0.78			
12	GeneSilico	371	64	0.76			
13	ZicoFullSTPFullData	138	64	0.75			
14	LEE-SERVER	293	9	0.75	97	0.80	2
15	McGuffin	379	63	0.73			
16	3DShot1	282	64	0.73			
17	Sternberg	202	64	0.72			
18	Jones-UCL	387	64	0.72			
19	mufold	310	61	0.71			
20	FAMS-multi	266	64	0.70			
21	Elofsson	200	64	0.68			
22	Chicken_George	81	64	0.67			
23	3DShotMQ	419	64	0.66			
24	Bates_BMM	178	64	0.65			
25	SAMUDRALA	34	53	0.63			
26	Hhpred5	12	64	0.61	154	0.64	5
27	LevittGroup	442	62	0.61			
28	BAKER-ROBETTA	425	64	0.60	154	0.57	8
29	RAPTOR	438	64	0.59	154	0.69	3
30	LEE	407	64	0.59			
31	MidwayFolding	208	63	0.57			
32	Phyre_de_novo	322	64	0.56	154	0.67	4
33	Ozkan-Shell	485	24	0.55			
34	Hhpred4	122	64	0.54	154	0.56	10
35	ABIpro	340	64	0.54			
36	sessions	139	4	0.52			
37	MUSTER	408	64	0.51	154	0.47	20
38	METATASSER	182	64	0.51	154	0.62	7
39	Pcons_multi	429	62	0.50	151	0.51	13
40	Pro-sp3-TASSER	409	64	0.50	154	0.63	6
41	TsaiLab	230	4	0.49			

			"Human and Server" target subset		All targets		
Rank	Group name	Group id	No. of targets	Mean Z-score	No. of targets	Mean Z-score	Rank
42	fais@hgc	198	51	0.48			
43	A-TASSER	149	64	0.47			
44	ricardo	403	12	0.46			
45	circle	396	61	0.45	150	0.40	25
46	Hhpred2	154	64	0.45	154	0.50	15
47	MULTICOM-CLUSTER	20	64	0.43	154	0.56	11
48	SAM-T08-server	256	64	0.43	154	0.48	17
49	YASARA	147	15	0.42	74	0.41	24
50	FEIG	166	64	0.41	154	0.47	18
51	GS-KudlatyPred	279	63	0.41	153	0.49	16
52	Phyre2	235	64	0.40	154	0.34	34
53	SHORTLE	253	42	0.40			
54	CBSU	353	36	0.39			
55	FAMSD	140	64	0.39	154	0.47	19
56	MULTICOM-REFINE	13	64	0.39	154	0.56	9
57	POEMQA	124	63	0.38			
58	MUProt	443	64	0.38	154	0.54	12
59	CpHModels	193	59	0.38	146	0.33	37
60	COMA-M	174	63	0.37	153	0.45	22
61	Phragment	270	64	0.37	154	0.32	40
62	FFASsuboptimal	142	60	0.36	150	0.36	32
63	EB_AMU_Physics	337	61	0.35			
64	Jiang_Zhu	369	64	0.35			
65	MULTICOM-RANK	131	64	0.35	154	0.51	14
66	TJ_Jiang	384	64	0.35			
67	reivilo	22	1	0.34			
68	FALCON	351	64	0.34	154	0.39	26
69	3D-JIGSAW_AEP	296	63	0.34	153	0.33	38
70	PS2-manual	23	61	0.34			
71	PSI	385	64	0.34	154	0.35	33
72	NirBenTal	354	11	0.33			
73	Pcons_dot_net	436	59	0.32	144	0.37	28
74	PS2-server	48	61	0.32	151	0.42	23
75	3Dshot2	427	64	0.32	154	0.34	35
76	nFOLD3	100	63	0.32	151	0.31	42
77	AMU-Biology	475	59	0.32			
78	FrankensteinLong	172	45	0.31			
79	MULTICOM-CMFR	69	64	0.31	154	0.46	21
80	jacobson	470	1	0.31			
81	FALCON_CONSENSUS	220	63	0.31	153	0.32	41
82	Softberry	113	64	0.30			
83	Poing	186	64	0.30	154	0.29	45

			"Human and Server" target subset		All targets		
Rank	Group name	Group id	No. of targets	Mean Z-score	No. of targets	Mean Z-score	Rank
84	fais-server	116	59	0.29	148	0.37	27
85	Keasar-server	415	58	0.29	140	0.37	29
86	Frankenstein	85	56	0.28	131	0.28	48
87	FFASstandard	7	60	0.28	148	0.33	39
88	taylor	356	12	0.28			
89	COMA	234	63	0.28	153	0.34	36
90	Bilab-UT	325	64	0.27			
91	FFASflextemplate	247	59	0.27	147	0.29	46
92	Pipe_int	135	60	0.26	143	0.36	30
93	Hao_kihara	284	62	0.26			
94	GeneSilicoMetaServer	297	59	0.26	147	0.27	51
95	Pcons_local	143	60	0.26	145	0.28	47
96	3D-JIGSAW_V3	449	63	0.26	153	0.31	43
97	mGenTHREADER	349	64	0.26	154	0.30	44
98	Abagyan	458	6	0.25			
99	SAINT1	119	35	0.25			
100	GS-MetaServer2	153	60	0.24	146	0.27	49
101	PRI-Yang-KiharA	39	64	0.24			
102	BioSerf	495	64	0.23	152	0.36	31
103	keasar	114	63	0.22			
104	kolinski	493	64	0.22			
105	mti	289	6	0.22			
106	POEM	207	64	0.21			
107	ACOMPMOD	2	60	0.20	143	0.17	58
108	FUGUE_KM	19	55	0.20	141	0.15	60
109	SAM-T02-server	421	60	0.19	148	0.19	56
110	Zhou-SPARKS	481	40	0.19			
111	Tripos_08	83	27	0.19			
112	fleil	70	64	0.18			
113	SAM-T06-server	477	64	0.18	154	0.21	53
114	3Dpro	157	58	0.17	147	0.18	57
115	JIVE08	330	40	0.17			
116	RBO-Proteus	479	63	0.16	153	0.19	55
117	Wolfson-FOBIA	10	7	0.15			
118	mumssp	345	5	0.14			
119	FOLDpro	164	64	0.14	154	0.09	64
120	forecast	316	64	0.13	151	0.23	52
121	Fiser-M4T	394	25	0.12	93	0.27	50
122	Sasaki-Cetin-Sasai	461	40	0.12			
123	Pushchino	243	47	0.10	127	0.21	54
124	SMEG-CCP	14	62	0.10			
125	panther_server	318	48	0.10	129	0.13	62

			"Human and server" target subset		All targets		
Rank	Group name	Group id	No. of targets	Mean Z-score	No. of targets	Mean Z-score	Rank
126	LOOPP_Server	454	56	0.09	135	0.17	59
127	Wolynes	93	27	0.08			
128	Handl-Lovell	29	18	0.07			
129	ProtAnG	110	38	0.07			
130	huber-torda-server	281	42	0.07	92	0.13	63
131	xianmingpan	463	54	0.06			
132	MUFOLD-MD	404	62	0.06	150	0.09	65
133	DelCLab	373	60	0.05			
134	mariner1	450	58	0.04	143	0.07	67
135	MUFOLD-Server	462	64	0.04	154	0.15	61
136	StruPPi	183	63	0.03			
137	TWPPLAB	420	64	0.03			
138	RPFM	5	10	0.02			
139	OLGAFS	213	43	0.02	125	0.08	66
140	NIM2	55	10	0.02			
141	POISE	170	11	0.01			
142	rehtnap	95	48	0.01	131	0.04	68
143	FLOUDAS	236	36	0.01			
144	Distill	73	62	0.01	152	0.02	69
145	ProteinShop	399	6	0.01			
146	MeilerLabRene	211	45	0.01			
147	Schenk-torda-server	262	56	0.01	136	0.00	70
148	DistillSN	272	59	0.00			
149	Mahmood-torda-server	53	39	0.00	73	0.00	71
150	Scheraga	324	35	0.00			
151	psiphifoldings	63	30	0.00			
152	igor	188	13	0.00			
153	ShakAbInitio	104	7	0.00			
154	Dill_ucsf	414	7	0.00			
155	Linnolt-UH-CMB	382	5	0.00			
156	HCA	402	5	0.00			
157	PHAISTOS	459	5	0.00			
158	BHAGEERATH	274	3	0.00	5	0.00	72
159	PZ-UAM	18	2	0.00			

La categoria TBM nel CASP8 comprende 154 unità di valutazione, 64 delle quali "human –server" e le restanti 90 "server-only ".
Nella tabella 2 tutti i gruppi sono classificati sul sottoinsieme di 64 domini human/server mentre i gruppi server sono anche classificati sulla completa lista di 154 domini.
Il test Shapiro Wilk stabilì che solo sette delle 154 distribuzioni dei valori di GDT_TS sono gaussiane all'1% del livello di confidenza. Una distribuzione non gaussiana dei valori di GDT_TS può essere dovuta al fatto che alcuni gruppi usarono differenti

strutture templato per costruire i loro modelli oppure se alcuni gruppi usarono metodi template-free.
L'applicazione di entrambi i test (t-test e Wilcoxon test) ai dati diede risultati sostanzialmente identici: i gruppi statisticamente indistinguibili risultarono tali da entrambe le analisi.
I risultati ottenuti sul sottoinsieme di target "human and server " non danno particolari informazioni sulla qualità dei diversi metodi poiché la maggior parte sono statisticamente indistinguibili (283 IBT_LT, 489 DBACKER, 71 Zhang, 426 Zhang-server, 57 TASSER, 434 fams-ace2, 196 ZicoFullSTP, 46 SAM-T08-human, 299 Zico, 453 MULTICOM, 371 GeneSilico, 138 ZicoFullSTPFullData, 379 McGuffin, 282 3DShot1).
Ciò può essere dovuto al fatto che il numero di target "human and server " non è sufficientemente alto per trarre conclusioni oppure perché la maggior parte dei metodi usati sono effettivamente molto simili.
Tra i gruppi migliori, solo il gruppo 426 (Zhang-server) si è ufficialmente registrato come server, sebbene sia possibile che qualcuno degli altri gruppi "human" abbia usato una procedura completamente automatica.
Nel confronto tra i server il gruppo 426 (Zhang-server) è il migliore.
E' statisticamente indistinguibile dal gruppo 293 (Lee-server) ma quest'ultimo sottomise predizioni solo su 97 dei 154 possibili domini TBM.
I tre successivi migliori server sono 438 Raptor, 322 Phyre_de_novo, 12 HHpred5, i quali perdono nel confronto con le predizioni human nel sottoinsieme di target " human and server ".
Questo può essere dovuto ad una effettiva miglior performance dei gruppi human, ma può anche essere dovuto a una diversa performance dei server per il sottoinsieme di target human.

3.2 Valutazione oltre i C_α di modelli TB nel CASP 8

Valutazioni precedenti dei modelli CASP basati su templati si sono focalizzate principalmente sul GDT (global distance test) dal programma LGA (local-global alignment).
Il suo potere deriva principalmente dall'uso di sovrapposizioni multiple per valutare similitudini tra strutture, in contrapposizione a metriche più quotidiane come l'RMSD (root-mean-square deviation) che usa una singola sovrapposizione. Nonostante ciò le metriche LGA considerano solo gli atomi C_α – in altre parole ignorano più del 90% della proteina.
Molti metodi di predizione attuali fanno uso di tutti gli atomi e molti modelli CASP degli ultimi anni sono sufficientemente accurati da rendere appropriata una valutazione più ampia.
Il nostro contributo (come assessors) alla valutazione della categoria TBM del CASP8 è lo sviluppo di sei metriche full-model (calcolate da Richardson e collaboratori) che sono, in un certo senso, ortogonali alle metriche basate sulla sovrapposizione delle coordinate C_α (come GDT).

Riteniamo che GDT_HA esplora un livello di dettaglio strutturale simile a quello raggiunto dalle nostre nuove misure e perciò continuiamo ad usarlo ampiamente nel seguito.
Due delle nuove metriche si focalizzano sulla geometria del sidechain: GDC-sc e la frazione di rotameri corretti (corRot, percentuale); due metriche si focalizzano sui legami idrogeno: la frazione dei legami idrogeno corretti del mainchain (HBmc, percentuale) e la frazione dei corretti legami idrogeno del sidechain (HBsc, percentuale); e due metriche si focalizzano sugli steric clashes e sulle lunghezze di legame: MolProbity, usato per validare strutture sperimentali e il mainchain reality score (MCRS). Per combinare le sei nuove metriche in una singola misura full-model, o per valutare la performance relativa tra i gruppi che effettuarono le predizioni, le metriche furono convertite in z-scores misurati in deviazioni standard sopra o sotto la media, come è pratica standard nel CASP da qualche tempo.
Lo z-score medio delle sei misure full-model è mediato con GDT z-score medio dei vari gruppi per avere la classificazione complessiva delle performance full-model high-accuracy.
Buoni full model score si correlano robustamente con alta accuratezza dei C_α. Non sono riportati qui gli z-score medi dei gruppi che sottomisero modelli per meno di venti targets.

Misura 1 – MolProbity Score (MP score)

Le prime due delle sei nuove metriche, MolProbity score e mainchain reality score, sono basate solo su proprietà del modello. MP score è usato per valutare quanto è simile ad una proteina ciascuna conformazione locale in relazione a strutture note per i rotameri delle catene laterali o per i valori Ramachandran del backbone.

MPscore è definito come:
MPscore= 0.426 × ln(1+clashscore) + 0.33 × ln(1+max(0,rota_out-1)) + 0.25 × ln(1+max (0, rama_iffy-2)) + 0.5 dove clashscore è definito come il numero di sfavorevoli sovrapposizioni steriche >0.4 Å come calcolato da Probe; valori più bassi indicano modelli migliori.
Rota_out è la percentuale di conformazioni di rotameri classificate come outliers.
Rama_iffy è la percentuale di conformazioni del backbone classificate come Ramachandran allowed or outlier (in altre parole non nelle regioni favored Ramachandran).
Valori bassi di MPscore indicano modelli più realistici. MolProbity score è stato usato per la prima volta nel CASP8 per valutare strutture non sperimentali.

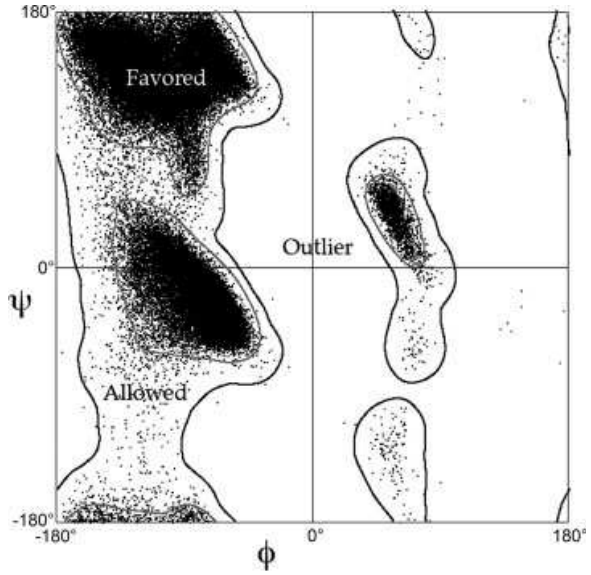


Fig 2 :Distribuzione Ramachandran empirica. I punti sono gli angoli diedri del backbone φ ,ψ .Dall'insieme delle migliori 500 strutture cristallografiche. I contorni sono calcolati con un algoritmo di smoothing dipendente dalla densità. 98%dei dati cadono all'interno della favored region (dentro il contorno grigio), 99.95% all'interno delle regioni allowed or favored (dentro il contorno nero), e lo 0.05%nella regione esterna (fuori dal contorno nero).

Misura 2 – Mainchain reality score (MCRS)

All-atom clashes, Ramachandran outliers, bond length or angle outliers for backbone.
Per completare la misura precedente fu sviluppata una valutazione di un modello che
1.usi solo atomi del backbone nella sua analisi
2.consideri le deviazioni eccessive degli angoli di legame e delle lunghezze di legame dai loro valori chimici attesi.
MCRS è definito come :
MCRS = 100- 10* spike – 5 * rama_out- 2.5 *length_out -2.5 * angle_out dove spike è la media per residuo della somma delle lunghezze da Probe (indicanti le sovrapposizioni steriche) tra coppie di atomi del mainchain.
Rama_out: è la percentuale di conformazioni del backbone classificate come Ramachandran outliers (Fig 2).
Length_out, angle_out: percentuali di residui con lunghezze di legame e angoli di legame della catana principale che sono (outliers) >4σ dall'ideale.
Il valore perfetto di MCRS è 100 e qualsiasi non idealità viene sottratta e fornisce punteggi inferiri. Il valore più basso è fissato a zero.

Misure 3,4 – Hydrogen bond correctness (HBmc e HBsc)

Le ultime quattro delle sei nuove metriche full-model sono basate su confronti tra il modello predetto e la struttura della proteina target. Gli assessors del CASP (TBM), conoscendo l'importanza dei legami idrogeno nel determinare la specificità dei

ripiegamenti proteici, esaminarono la correttezza dei legami idrogeno in relazione al target. Noi abbiamo seguito questa strada ma abbiamo separato le categorie per la catena principale (HBmc: mainchain-mainchain hydrogen bonds) e le catene laterali (HBsc: sidechain-mainchain e sidechain-sidechain hydrogen bonds), usando Probe per identificare i legami idrogeno.
Brevemente l'approccio fu di calcolare le coppie atomiche coinvolte in legami idrogeno nel target, fare lo stesso per il modello e poi vedere la percentuale di coppie coinvolte in legami idrogeno nella struttura nativa che sono correttamente riprodotte nel modello.
Probe definisce i legami idrogeno come coppie donatore –accettore più vicine (o dentro) dei contatti van der Waals.
Comunque, la tolleranza per HBsc è leggermente aumentata rispetto a HBmc e coppie donatore –accettote che sono dentro 0.5Å dei contatti van der Waals sono trattate come legami idrogeno.
Da notare che nessuna delle due metriche sta contando il totale numero di legami idrogeno; piuttosto esse contano la frazione dei corretti legami idrogeno della nativa che sono riprodotti dal modello.

Misura 5 – Rotamer correctness (corRot)

Frazione dei rotameri delle catene laterali del target corrispondenti nel modello (tutti gli angoli χ).
La metrica corROt misura la frazione delle conformazioni del sidechain del modello che si ritrovano nella struttura nativa. La procedura per determinare se i rotameri si eguagliano consiste nell'assegnare una lettera per ogni angolo χ (t= trans, m = attorno a -60º, p = attorno a +60º) e poi combinare le lettere per produrre una stringa che serve come nome del rotamero. La conformazione del modello è considerata un match se le stringhe per il modello e le strutture sperimentali sono identiche – significa cioè che tutti gli angoli χ per quel residuo devono corrispondersi. Da notare che qualsiasi catena laterale del modello che non sia in un rotamero definito (un outlier) è considerato nonmatching, a meno che il corrispondente rotamero nel target sia non definito, nel qual caso quel residuo è semplicemente ignorato da corRot.
Per i target risolti con raggi X, l'insieme dei rotameri del target consiste di tutti i residui per i quali può essere assegnato un valido nome di rotamero (punteggio percentuale assegnato da MolProbity non <1% e che non sia non definito a causa di atomi mancanti).
Per target risolti con NMR, definiamo l'insieme dei rotameri del target includere solo quei residui per i quali un rotamero con nome comprenda una specifica percentuale (85,70, 55 e 40% per catene laterali con 1,2,3 e 4 angoli χ, rispettivamente) dell'insieme.

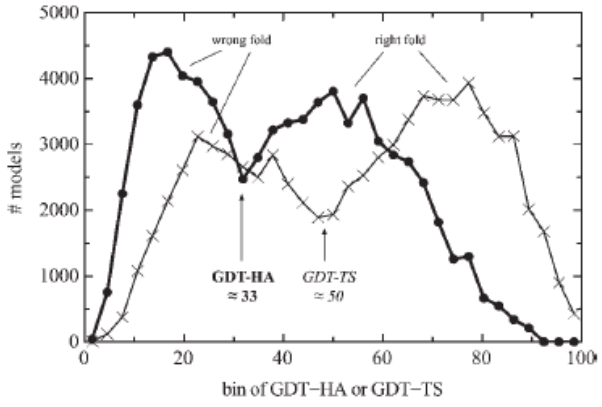
Misura 6 – Sidechain Positioning (GDC-sc)

Punteggio GDT-style per atomi alla fine di ciascuna catena laterale eccetto glicina o alanina, limiti da 0.5 a 5 Å (dal programma LGA).
La correttezza della posizione del sidechain si calcola usando la metrica GDC-sc (global distance calculation for sidechains).Questa metrica è simile a GDT-TS e si calcola usando il programma LGA. Mentre GDT usa le posizioni del carbonio alpha,

GDC-sc usa invece un atomo di riferimento da ciascun tipo di catena laterale (vedere Proteins 2009 [29-49] per la lista delle coppie atomo – residuo).
Prima di tutto si calcola la sovrapposizione ottima (per il backbone) tra il modello e la struttura nativa. Poi, per ogni residuo si calcola la distanza tra la posizione dell'atomo di riferimento nel modello e nella struttura nativa. Ciascuna distanza è assegnata a un contenitore iesimo (bin i) con i=1 per distanze < 0.5Å e i =10 per distanze < 5Å.
L'atomo di riferimento può essere assegnato a diversi bins, ad esempio se la distanza è < 0.5Å, l'atomo di riferimento sarà assegnato a tutti i bins. Il valore di GDC-sc è poi calcolato come:
GDC-sc = 100*2* (k*Pa_1 + (k-1)*Pa_2…+1* Pa_k)/(k+1)*k, dove k=10 è il numero di contenitori e Pa_i è la frazione di atomi di riferimento assegnati al contenitore i.
Un valore pari a 100 indica che tutti gli atomi di riferimento nel modello sono entro 0.5Å della loro posizione nella struttura nativa, mentre un valore pari a zero indica che tutti gli atomi di riferimento sono oltre 5Å lontani dalla posizione corretta.
Le tre misure che riguardano le catene laterali (HBsc, corRot, GDC-sc) sono valutazioni significative solo per modelli con un backbone fold approssimativamente corretto, così sono state usate solo per modelli con valori di GDT sopra la media.

Selezione dei modelli e filtraggio

La valutazione dei modelli nel CASP tramite GDT-TS ha sempre usato il modello designato come " model 1 " dai gruppi che effettuano le predizioni.
Ciò ha il vantaggio di ricompensare i gruppi che sono più abili nel selezionare i loro modelli migliori; comunque, in questo modo vengono eliminati molti dei modelli tra i migliori. Allora per la valutazione full-model (TBM) abbiamo invece scelto di valutare i modelli migliori (secondo GDT-TS) di ciascun target.
Punteggi basati sulle sovrapposizioni (GDT-HA , GDT-TS, GDC-sc) sono stati calcolati sui domini dei target perché, come nelle valutazioni TBM precedenti, non vogliamo penalizzare i gruppi che modellarono correttamente i domini ma modellarono in modo non corretto le orientazioni relative all'interno dei domini.
I punteggi basati sul confronto locale con il target e sulla qualità del modello (MPscore, MCRS, corRot, HBmc, HBsc) sono stati calcolati sull'intero target, in quanto tali punteggi sono approssimativamente additivi anche attraverso orientazioni non accurate dei domini. La distribuzione dei punteggi GDT è fortemente bimodale.
Fig 3:

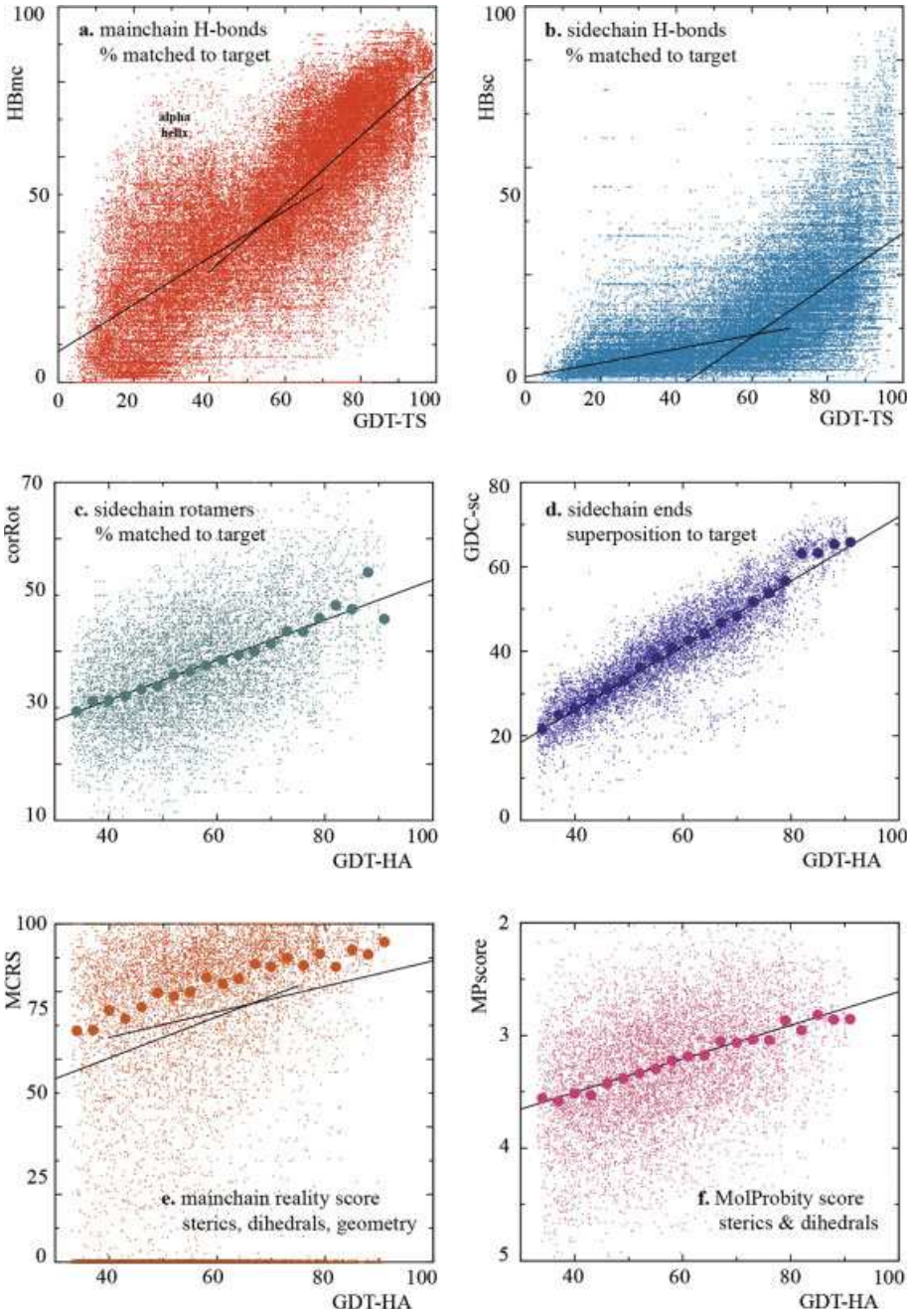


Distribuzioni bimodali di GDT-HA e GDT-TS.

Tutti i modelli TBM del CASP8 sono collocati in 33 contenitori ugualmente spaziati, separatamente per GDT-HA e GDT-TS.
La divisione tra “corretto ripiegamento “ e “ errato ripiegamento “ avviene intorno a un valore di GDT-HA di 33 e a un valore di GDT-TS di 50.
Come illustrato in figura 3 i modelli cadono sotto uno dei due picchi in GDT-HA o GDT-TS: questa proprietà delle distribuzioni suggerisce una possibile selezione dei modelli che sono più appropriati per l’analisi fornita dalle nostre nuove metriche.
Allora noi consideriamo solo i seguenti: 1)modelli con GDT-HA≥33 per le nostre metriche basate su domini e 2)modelli con almeno un dominio con GDT-HA≥33 per le nostre metriche basate sull’intero target.
(Notare che tutti i target hanno al più tre domini eccetto TO487 con cinque domini, così noi abbiamo aumentato la richiesta sul suo modello a due domini con GDT-HA ≥33).

Contenuto informativo delle misure full-model

Una qualsiasi nuova metrica appropriata che misuri la qualità di un modello dovrebbe avere una correlazione positiva con i punteggi GDT, ma dovrebbe anche fornire informazione addizionale, ortogonale con un intervallo di valori ampio e con qualche modello che raggiunge punteggi elevati.
La figura 4 riporta i valori di ciascuna delle sei misure full-model contro GDT-TS o GDT-HA e fa vedere una forte correlazione positiva in tutti i casi (notare che la correlazione è tecnicamente negativa per MPscore , ma valori più bassi di MPscore sono migliori).
I grafici 4a e 4b includono tutti i modelli attraverso l’intero range di valori di GDT e fanno vedere che i dettagli sono relativamente disaccoppiati per la metà più bassa del range dei valori di GDT ma risultano ben correlati per la metà più alta, in corrispondenza con le distribuzioni GDT bimodali in figura 3 sopra.
La figura 4 (c-f) riporta solo i modelli migliori con GDT-HA ≥33.
Grossi punti rappresentano i valori medi di ogni misura all’interno di contenitori spaziati di tre unità di GDT-HA, per migliorare la chiarezza degli andamenti, sebbene con alta variabilità alle estremità dovuta a contenitori con un minor numero di valori.



(a,b) tutti i modelli
(c-f) solo i modelli migliori con GDT-HA≥33.
Andamenti lineari doppi per modelli con GDT-TS <55 vs.≥55 in (a) e (b) e per modelli con GDT-HA <60 vs.≥60 in (e).
I grossi punti in (c-f) sono i valori medi per contenitori di tre unità di GDT-HA; contenitori ad alti valori di GDT-HA includono pochi modelli, e quindi producono alta variabilità per alcune misure (vedi corRot).Le linee sono sotto i punti medi in (e) perché molti punti giacciono a valore nullo di MCRS.

Considerati insieme questi risultati mostrano che come regola generale tutti gli aspetti migliorano insieme, ma che diversi parametri si accoppiano in modo diverso come messo in evidenza dai livelli di saturazione e dispersione che cambiano.
Senza sorpresa, GDC-sc ha la correlazione più stretta con GDT-HA. Esso misura la corrispondenza della posizione della fine delle catene laterali tra il modello e il target, per la quale la corrispondenza della posizione dei C_α è un prerequisito.
GDC-sc fa vedere il rialzo più pronunciato per alti valori di GDT-HA, un effetto individuabile anche negli altri grafici.
Richiederebbe ulteriore studio decidere in che misura questo è causato dalla disponibilità di templati più completi e in che misura c'è una soglia dell'accuratezza del backbone oltra la quale diventa più fattibile raggiungere più accuratezza.
Considerate insieme le misure GDC-sc, corRot e HBsc valutano il problema di ottimizzare il posizionamento delle catene laterali in modi distinti e permettono valutazioni future del CASP più vicine al livello atomico.
Anche le misure model-only – MCRS e MPscore – si correlano con alta accuratezza dei C_α (figura 4 e-f). MolProbity score ha alta dispersione e pendenza relativamente bassa ma è lineare sull'intero range. MCRS ha valori pessimi in modelli poveri ma satura a valori piuttosto buoni per alti GDT-HA.
La mancanza di modelli con pessimi valori di MCRS per buoni GDT-HA suggerisce che modellizzare catene principali fisicamente realistiche può essere essenziale per raggiungere predizioni accurate; comunque come notato per GDC-sc, questo legame necessita ulteriori approfondimenti.
La nuova separazione dei legami idrogeno del mainchain e del sidechain è utile, in quanto essi mostano una forte correlazione ma distribuzioni 2D diverse che quindi darebbero meno informazione se combinate.
Per bassi valori di GDT, quasi nessun legame idrogeno del sidechain è corrispondente nel target, mentre i legami idrogeno del mainchain mostrano un picco artificiale a causa della predizione di strutture secondarie di α-eliche senza corretta struttura terziaria. La metà superire di entrambe queste misure ha una correlazione molto forte e alta pendenza in relazione a GDT, ma con un ampia diffusione di valori indicativa di un contributo significativo da informazione indipendente.

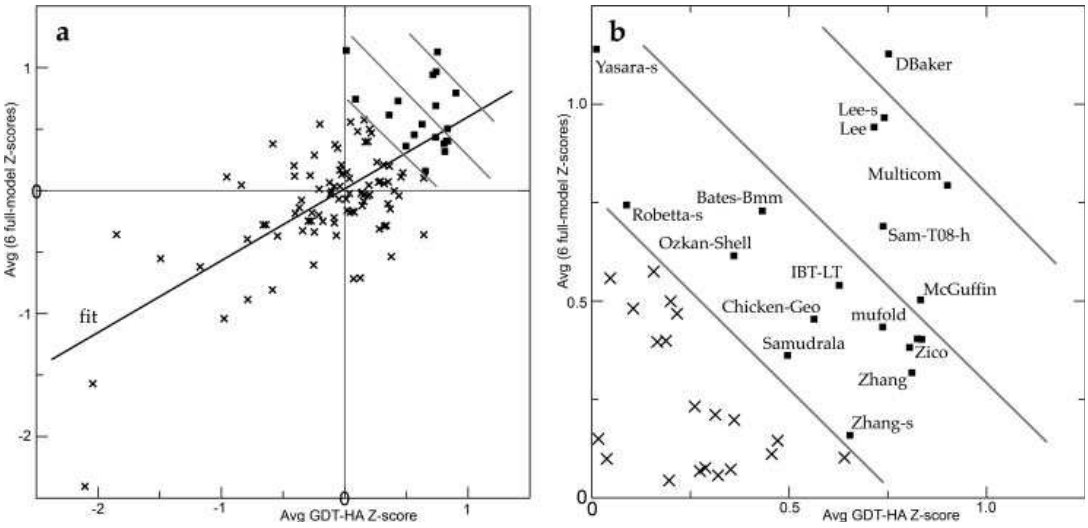
Risultati : classificazione dei gruppi

La valutazione full-model presentata qui differisce dalle classificazioni precedenti in quanto usiamo il modello migliore (secondo GDT-TS) invece del "modello 1 ", nell'uso di punteggi GDT piuttosto che GDT-z score per la scelta dei modelli (solo modelli con GDT sopra la media) e in quanto valutiamo l'intero modello.
Una ulteriore differenza dalle versioni recenti è la considerazione di dimensioni multiple delle performance: si considerano le performance combinate di GDT-TS o HA e le misure full-model.
La tabella 1 fa vedere la classificazione dei gruppi su ciascuna delle sei misure full-model.

TABELLA 1

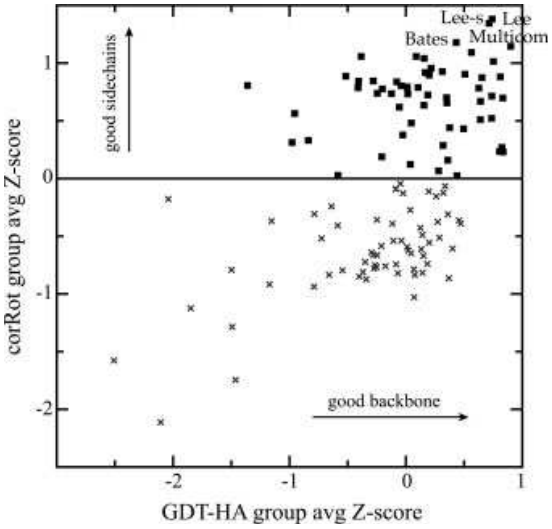
Group ID	Group name	6full +GDT[a]	6full rank	MCRS avg Z	Mpscore avg Z	HBmc avg Z	HBsc avg Z	GDC-sc avg Z	corRot avgZ
489	**DBaker**	1	2	**Yasara**	**Yasara**	LevittGroup	**Lee-s**	**Lee**	**Lee-s**
293s	**Lee-s**	2	3	**Lee**	Ozkan-Shell	**Sam-T08-h**	**DBaker**	**Lee-s**	**Lee**
453	**Multicom**	3	5	**DBaker**	**DBaker**	**DBaker**	**Lee**	**IBT_LT**	**Bates_BMM**
407	**Lee**	4	4	**Lee-s**	A-Tasser	Keasar	Keasar-s	**Multicom**	**Multicom**
046	**Sam-T08-h**	5	9	**Bates_BMM**	Robetta	Mufold	**Yasara**	McGuffin	ChickenGeo
379	McGuffin	6	16	MuProt	**Bates_BMM**	**Multicom**	LevittGroup	Zhang	Robetta
196	ZicofSTP	7	23	Robetta	**Lee-s**	Zhang	**Sam-T08-h**	LevittGroup	Sam-T08-s
138	ZicofSTPfData	8	22	**Multicom**	Keasar	PoemQA	Robetta	Zhang-s	**DBaker**
299	Zico	9	26	MulticomRef	**Lee**	**Bates_BMM**	McGuffin	ChickenGeo	Fais@hgc
310	Mufold	10	21	PoemQA	**Multicom**	Sam-T08-s	**Multicom**	Sam-T08-s	Pcons_multi
283	**IBT_LT**	11	15	Elofsson	Pcons_dot_net	Keasar-s	Sam-T08-s	**Sam-T08-h**	LevittGroup
178	**Bates_BMM**	12	8	Pcons.net	ChickenGeo	**Lee**	Ozkan-Shell	ZicofSTPfData	Zhang
147s	**Yasara**	13	1	Hhpred5	**Sam-T08-h**	McGuffin	MulticomClust	Mufold	Zhang-s
071	Zhang	14	32	Fais-s	Mufold	**Lee-s**	GeneSilico	**Bates_BMM**	**Yasara**
081	ChickenGeo	15	20	GSKudlatyPred	Sam-T08-s	**Yasara**	Keasar	**DBaker**	Pcons_dot_net
485	Ozkan-Shell	16	10	Hao_Kihara	Pcons_multi	ZicofSTP	ZicofSTP	ZicofSTP	**IBT_LT**
034	Samudrala	17	30	MulticomRank	Samudrala	ZicofSTPfData	MulticomRank	Zico	Keasar
425s	Robetta	18	7	MulticomClust	MulticomCMFR	Zico	Fams-multi	Fams-multi	MulticomCMFR
426s	Zhang-s	19	41	MulticomCMFR	Hao_Kihara	**IBT_LT**	ZicofSTPfData	Samudrala	**Sam-T08-h**
434	Fams-ace2	20	49	PS2-s	**IBT_LT**	Zhang-s	**IBT_LT**	Fams-ace2	Phyredenovo

6Full rank: classificazione basata sulla media degli Z-score delle sei mosure full-model.
[a]6Full + GDT-HA rank: classificazione basata sulla somma di (1) Avg GDT-HA Z-score e (2) media degli Z-score delle sei misure full-model. I gruppi in grassetto appaiono tra i primi quattro almeno una volta e tra i primi venti per cinque delle sei metriche full-model.
La figura 5 fa vedere la performance combinata su GDT e punteggi full-model con un grafico a due dimensioni.
Fig 5:



(a) Z-score medio dei gruppi per i sei punteggi full-model vs. z-score medio dei gruppi per GDT-HA. (b) Ingrandimento del quadrante in alto a destra dalla figura a con i gruppi migliori evidenziati.

DBaker è il vincitore di questa valutazione combinata di sovrapposizione dei C_α e correttezza strutturale.
Lee, Lee-server, MultiCom, Sam-T08-h, e McGuffin sono i successivi nella classifica (fig 5), mentre Bates-BMM, IBT-LT e Yasara sono degni di nota perché appaiono tra i primi quattro almeno una volta e tra i primi venti per cinque delle sei metriche full-model (tabella 1).
La figura 6 rappresenta lo z-score medio dei gruppi per i rotameri delle catene laterali (corRot) contro lo z-score medio dei gruppi per GDT-HA .
Fig 6:



Oltre agli andamenti visti in figura 4 in questo grafico si notano due raggruppamenti, per alti e per bassi valori.
Considerando valori di GDT-HA tra -1 e +0.5, corROt è circa indipendente da GDT-HA in entrambi i raggruppamenti.
Ciò suggerisce che molti gruppi intermedi lasciarono il posizionamento del sidechain e del backbone disaccoppiati.
Tuttavia, per i gruppi con migliori GDT-HA all'estrema destra del grafico anche corROt è eccellente, il che implica che una corretta modellizzazione del sidechain può essere necessaria per ottenere un accurato posizionamento del backbone.
Non risulta che l'eccellenza in una qualsiasi delle metriche full-model si raggiunga attraverso un compromesso con la metrica GDT; piuttosto esse tendono a migliorare insieme.

Identificazione del "right fold"

Tentiamo anche di valutare quali gruppi furono i migliori nell'identificare correttamente il templato o il "right fold "per il passo iniziale dell'homology modeling.
Per fare ciò abbiamo calcolato la percentuale di tutti i modelli di un gruppo con GDT-HA ≥ 33 considerando tali modelli sufficientemente accurati, come già discusso sopra.
Comunque abbiamo dovuto tenere conto anche della difficoltà media dei target tentati da ciascun gruppo.
Riportando in un grafico nell'asse y la percentuale di tutti i modelli di ciascun gruppo con GDT-HA ≥ 33 ("right fold ") e nell'asse x il GDT-TS medio di tutti i gruppi (media effettuata sui target tentati da ciascun gruppo, tutti i modelli) come misura della difficoltà dei target, si distinguono tre raggruppamenti: i gruppi che fecero le predizioni dei target più difficili (human targets), quelli che tentarono tutti i target e quelli che tentarono solo i target più semplici (server targets).

La tabella 2 lista i gruppi migliori in ciascuna delle tre categorie.
Tabella 2:

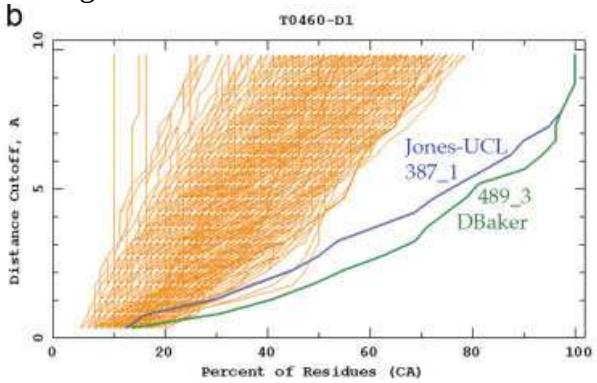
Target choice	Group ID	Group name	No.of targets attempted	Average of (target avg GDT-TS)	% Models GDT-HA≥ 33
Easier targets	147s	**Yasara**	60	72.06	92.6
	293s	**Lee-server**	78	63.40	83.6
	394s	**Fiser-M4T**	76	68.56	83.5
Average targets	379	McGuffin	119	58.09	79.2
or ≈ all targets	299	Zico	119	57.93	77.0
	138	ZicoFullSTPFullData	119	57.93	76.6
	266	FAMS-multi	120	57.59	76.1
	434	Fams-ace2	120	57.59	76.0
	196	ZicoFullSTP	119	57.93	76.0
	282	3DShot1	113	57.14	75.9
	485	Ozkan-Shell	27	56.75	75.5
	453	Multicom	119	57.93	74.7
	426s	**Zhang-server**	121	57.64	74.6
	007s	**FFASstandard**	121	58.58	74.5
	425s	**Robetta**	121	57.64	74.2
	475	AMU-Biology	98	60.64	74.0
	193s	**CpHModels**	120	59.19	73.6
	419	3DShotMQ	113	57.14	73.5
	340	ABIpro	119	57.86	73.5
	149	A-Tasser	119	57.93	73.1
	407	Lee	120	57.59	72.7
	409s	**Pro-sp3-Tasser**	121	57.64	72.4
	154s	**HHpred2**	121	57.64	72.3
	436s	**Pcons_dot_net**	117	58.14	72.1
	142s	**FFASsuboptimal**	121	58.77	72.1
	135s	**Pipe_int**	111	58.31	72.0
	122s	**HHpred4**	121	57.64	71.8
	297s	**GeneSilicoMetaServer**	119	58.43	71.1
	247s	**FFASflextemplate**	120	58.54	70.7
	443s	**MUProt**	121	57.64	70.6
	182s	**MetaTasser**	121	57.64	70.4
	438s	**Raptor**	121	57.64	70.3
	429s	**Pcons_multi**	121	58.11	70.3
Harder targets	310	Mufold	51	49.88	67.8
	283	IBT_LT	52	48.63	66.3
	489	DBaker	52	48.33	64.5
	353	CBSU	29	51.99	63.7
	200	Elofsson	53	48.51	60.8
	198	Fais@hgc	43	49.32	59.1
	178	Bates_BMM	52	48.33	58.7
	371	GeneSilico	52	48.33	58.0
	208	MidwayFolding	51	48.15	56.6
	442	LevittGroup	51	48.32	56.4

I nomi dei gruppi in grassetto indicano server.

Da notare che i gruppi server dominano per i target più semplici e i gruppi human per i target di difficoltà media e superiore.

Modelli eccellenti

Fig11b:



La figura 11(b) illusta il diagramma GDT-TS cumulativo per il FM/TBM target T0460 con due modelli molto migliori degli altri: 489_3 (DBaker) e 387_1 (Jones-UCL).
Tutti i modelli in arancio. Asse x: percentuale di atomi C_α all'interno di un limite di distanza dato dall'asse y, perciò linee più in basso e verso destra indicano predizioni che meglio coincidono con il target.
Solo i due modelli migliori raggiunsero un discreto match con il target (GDT-TS di 63 e 54, vs. il successivo gruppo a 40-44).

Per i target più semplici (server only) molti modelli hanno valori di GDT alti e molto simili, però i valori relativi alle sei misure full-model possono essere molto diversi.Tra questi citiamo il target T0494-D1: il modello 407_3 (Lee, GDT-HA=65.9, corRot=57.8%, GDCsc=56.0, HBsc=34.0%, HBmc=73.9%, MCRS=99.5, MPscore=2.76) ha il miglior full-model z-score medio su questo target; un altro modello con sostanzialmente lo stesso GDT-HA (GDT-HA=65.2, corRot=35.7%, GDCsc=45.8, HBsc=13.0%, HBmc= 61.1%, MCRS= 28.6, MPscore=3.98) ha uno z-score relativo alle misure full-model molto più basso, evidenziando una corrispondenza con il target molto inferiore per il sidechain e i legami idrogeno.
Questo caso fornisce un esempio di "valore aggiunto" oltre i C_α al fine di ottenere modelli di molta maggiore utilità.

Misure full model

I risultati riportati sopra fanno vedere che le sei nuove misure full-model hanno il comportamento corretto per valutazioni potenzialmente utili:
(1) si correlano robustamente con i punteggi GDT se misurate per modelli nella parte più alta della distribuzione GDT bimodale, ma l'estensione dei valori indica che esse danno contributo di informazione indipendente (fig 4);
(2) un numero considerevole di modelli raggiunsero buoni punteggi , ma essi non sono banalmente raggiungibili;

(3) per singoli target, l'esame di modelli con alti vs.bassi valori di punteggi full-model rivela caratteristiche di migliori vs.peggiori predizioni della struttura del target.
Quindi riteniamo che l'approccio generale di valutazione full-model sia adatto per valutare i modelli template –based del CASP, tramite queste o simili misure.

Le valutazioni high-accuracy per CASP8 riguardano un ambito definito dai modelli di predizione con GDT-HA ≥33, piuttosto che un ambito definito dai target designati come appartenenti alla categoria TBM-HA; questo approccio generale fu suggerito dopo il CASP7.
Abbiamo fatto tre tipi di valutazioni:
(1) "right fold " o identificazione del corretto templato per il passo iniziale (tabella 2);
(2) correttezza full-model, in sei componenti e complessiva (tabella 1); e (3) singoli modelli HA eccellenti.
E' importante notare che ciascuna di queste valutazioni è intrinsecamente bidimensionale, nel senso che necessita di essere considerata insieme con un'altra metrica di riferimento come GDT-TS (fig5), GDT-HA (fig 6), o le difficoltà dei target ("right fold " identification).
Molte distribuzioni dei punteggi sono bimodali (highly non –normal). Questo problema è una ragione del perché i valori GDT z-scores sono di solito troncati a zero e una ragione del perché le nostre misure full-model omettono modelli con GDT-HA <33.
I punteggi full-model che si ottengono filtrando i modelli con GDT-HA ≥33 hanno distribuzioni asimmetriche ma unimodali e possono accettabilmente essere mediati in un complessivo punteggio full-model z-score.
Notiamo poi che molti dei nuovi criteri full-model sono stati applicati in altri aspetti delle valutazioni CASP (model refinement).

3.3 Valutazione della categoria raffinamento nel CASP 8 [3]

Target usati per il raffinamento nel CASP 8

Nella categoria refinement ai gruppi che partecipano sono dati inizialmente due tipi di informazioni:
(1) come al solito la sequenza aminoacidica della proteina
(2) un modello iniziale che è tra i migliori risultati del CASP disponibili.
Il compito dei gruppi è di cercare di migliorare ulteriormente la struttura. Un tale esperimento, chiamato CASPR, fu condotto tra il CASP7 e il CASP8.
Le metriche che abbiamo usato per sapere se una struttura è stata migliorata oppure no comprendono:
(i) metodi standard basati sul backbone (GDT)
(ii) nuovi metodi dal gruppo Richardson (sei nuove misure full-model)
(iii) metodi basati su insiemi di strutture sperimentali .
Da notare che per raffinamento noi non intendiamo il raffinamento che si ottiene partendo dal miglior templato, bensì il raffinamento di strutture che sono il miglior raffinamento di quei templati.
La tabella 1 mostra i target usati per il raffinamento nel CASP8 insieme alla sorgente del modello iniziale e a misure della qualità di tale modello. I modelli iniziali furono scelti tra i migliori modelli sottomessi per ciascun target durante la normale competizione CASP.

Tabella 1:

Target	Starting model	Residues	Starting RMSD (Å)	Starting GDT-TS	Method
TR389	407_2-D1	1-134	2.63	81.3	X-Ray
TR429[a]	057_1	27-55, 75-175	6.72	47.0	X-Ray
TR432	443_5-D1	1-130	1.65	91.5	X-Ray
TR435	453_1-D1	15-58, 73-148	2.15	74.3	X-Ray
TR453	131_1-D1	5-90	1.40	88.8	X-Ray
TR454	178_1	5-196	3.24	64.3	X-Ray
TR461	253_1-D1	20-176	1.63	87.8	X-Ray
TR462[a]	198_2	1-74, 77-143	2.54	67.1	NMR
TR464	489_5-D1	18-86	2.94	77.5	NMR
TR469	426_3-D1	1-74, 77-143	2.18	88.9	NMR
TR476[b]	404_2-D1	2-88	6.85	47.1	NMR
Tr488	020_5-D1	1-95	1.43	88.2	X-Ray

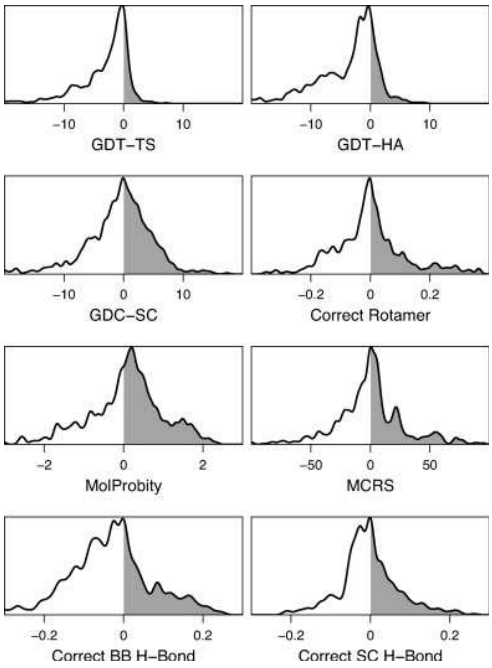
25 gruppi raffinarono un totale di dodici target.

[a] Questi target erano proteine con due domini. La maggior parte dei gruppi sottomisero una struttura combinata per i due domini. In questo caso, l'analisi fu fatta sull'intera struttura. Tre gruppi sottomisero domini separati per questi target che furono analizzati individualmente.

[b] Il modello iniziale non conteneva catene laterali. Questo target fu escluso da tutte le analisi che richiedono il side chain (GDC-sc, MolProbity, MCRS, HBsc).

Risultati

Fig 1:



Prima di tutto definiamo il raffinamento nullo come quella procedura che restituisce il modello iniziale senza nessun cambiamento. Noi giudichiamo un raffinamento "di successo" se raggiunge punteggi migliori del raffinamento nullo. Detto in altro modo, i metodi dovrebbero primum non nocere – prima di tutto non danneggiare. La figura 1 fa vedere le distribuzioni dei miglioramenti rispetto al modello iniziale per tutte le predizioni di tutti i gruppi (tutti i modelli).

Notare che è rappresentato l'inverso additivo di MolProbity score, così strutture migliori hanno ΔMolProbity >0. Valori alla destra dello zero (ombreggiati in grigio) rappresentano un miglioramento .

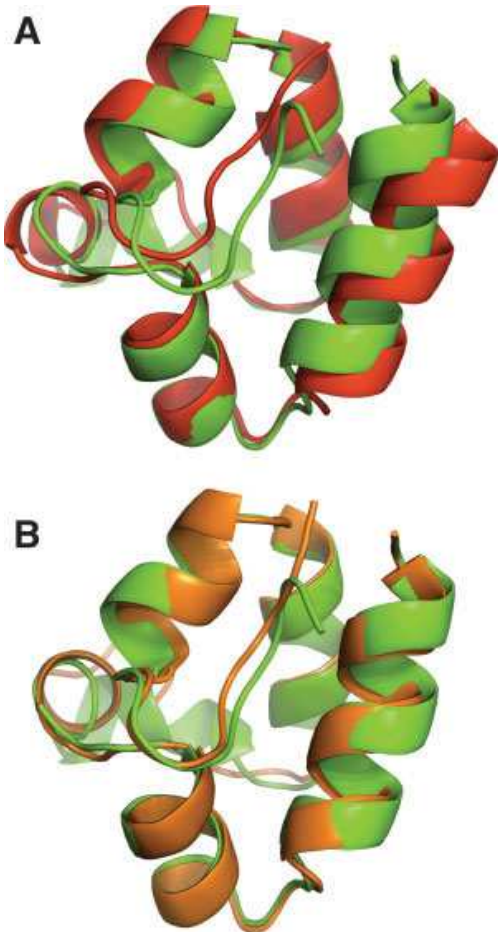
Questi risultati suggeriscono che il cambiamento medio per la maggior parte delle metriche è vicino a zero (ci sono circa tanti successi quanti fallimenti), mentre GDT-TS e GDT-HA sono

peggiori in media. Chiaramente questa analisi ignora che, come facciamo vedere nel seguito, alcuni gruppi sono molto migliori di altri. La figura due mostra un raffinamento ovviamente di successo. Il miglior risultato nella CASP8 refinement competition fu per il target TR469.

(A) Modello iniziale (in rosso , GDT-TS=82.94)sovrapposto alla struttura cristallografica (in verde)

(B) Struttura raffinata in arancio (GDT-TS =90.08) prodotta dal gruppo YASARARefine.

Fig 2:



Per confrontare i gruppi abbiamo calcolato gli z-scores: per ogni target e per ogni metrica abbiamo calcolato la media su tutte le strutture sottomesse e la deviazione standard. Poi abbiamo calcolato lo z-score di ogni modello. Questo calcolo è stato successivamente ripetuto escludendo tutti i modelli con uno z-score minore di -2.
Per ciascuna struttura, abbiamo poi calcolato una somma modificata di z-scores dove gli z-scores per GDT-TS e GDT-HA sono stati moltiplicati per un fattore pari a tre (per dare ugual peso alle metriche standard rispetto alle nuove).
Per ogni target definiamo la struttura migliore quella con la più alta somma modificata di z-scores.
Come mostrato in tabella 2 considerando solo il "modello 1 " sottomesso da ogni gruppo per ogni target, ci sono cinque gruppi migliori del raffinamento nullo (come giudicato dalla somma degli z-score): DBACKER, LEE, YASARARefine, FAMSD, e LevittGroup.
Tabella 2 :

Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct Sc H-bonds	MolProbity[a]	MCRS	Sum of Z-scores
Dbaker	-0.19	-0.69	2.88	0.07	0.04	0.09	1.18	-2.11	9.34
LEE	0.09	0.27	2.42	0.05	0.01	-0.01	0.22	16.95	7.54
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22	7.46
FAMSD	-0.73	-1.16	1.47	0.03	-0.04	0.03	0.11	-3.07	5.33
LevittGroup	-0.42	-0.99	1.38	-0.00	0.01	0.01	0.04	5.13	5.08
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.81
TASSER	-1.48	-1.54	0.78	0.00	-0.06	-0.02	0.09	1.14	4.78
FEIG_REFINE	-0.39	-0.67	1.19	-0.00	-0.07	0.03	1.05	-10.88	3.78
SAM-T08-HUMAN	-2.31	-2.88	0.73	0.02	0.02	0.02	0.07	11.55	3.38
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70	3.23
Tripos_08	-1.58	-1.92	-0.01	0.05	-0.05	0.00	0.05	8.92	2.68
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54	2.04
Jacobson	-2.98	-4.03	-2.13	-0.04	-0.07	0.04	0.11	0.57	1.21
Kolinski	-4.10	-6.71	-2.36	-0.05	-0.04	0.08	0.44	0.22	-0.34
Abagyan	-1.19	-2.31	-3.32	-0.07	-0.07	-0.01	-0.38	-4.09	-1.15
BATES_BMM	-3.68	-5.48	-1.32	0.02	-0.04	-0.02	0.07	4.53	-1.58
A-TASSER	-3.12	-4.23	-3.05	-0.04	0.00	-0.06	0.29	-0.20	-2.37
SAMUDRALA	-4.53	-3.40	-2.63	0.03	-0.08	-0.00	0.51	9.41	-4.55
FAMS-multi	-5.93	-6.82	-2.87	0.01	-0.11	0.05	-0.74	-17.08	-4.70
Keasar	-5.24	-7.84	-4.07	-0.01	0.03	0.03	0.19	-48.40	-7.05
POISE	-8.76	-11.38	-6.14	-0.02	-0.07	0.01	1.07	8.17	-8.02
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82	-8.43
MidwayFolding	-5.45	-6.66	-6.46	-0.06	-0.10	-0.03	-0.83	-14.53	-9.19
Jones-UCL	-12.69	-16.11	-11.93	-0.07	-0.12	-0.04	-0.67	-10.91	-20.68
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.01	0.09	-6.19	-30.92
POEM	-15.29	-20.15	-14.78	-0.22	-0.41	-0.13	-1.82	-31.34	-37.36

I gruppi sono ordinati secondo la somma degli z-scores. Il target 476 fu escluso da tutti i calcoli perché il modello iniziale non conteneva catene laterali. [a] Abbiamo usato l'inverso additivo di MolProbity, così i miglioramenti corrispondono a ΔMolProbity >0.

A differenza dei punteggi GDT, le altre metriche sono generalmente migliorate dalle procedure di raffinamento usate da questi gruppi.
Non ci concentriamo sulla classificazione relativa dei diversi gruppi migliori essendo essa fortemente dipendente dalla scelta delle metriche usate e dal loro peso relativo.
Risulta da alcuni calcoli (non mostrati) che i gruppi non sono in grado di giudicare i modelli sottomessi per ciascun target e, di conseguenza, la distribuzione dei risultati non cambia in modo significativo quando consideriamo l'intero insieme di modelli invece del modello 1.
Abbiamo poi ricalcolato i risultati ma questa volta esaminando solo la miglior struttura (come giudicato dalla somma degli z-scores) per ogni target (tabella 3). Considerando solo i modelli migliori ci sono otto gruppi che ebbero successo. I primi due, DBAKER e LEE, in media sono in grado di migliorare tutte le metriche.
Tabella 3:

Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct SC H-bonds	MolProbity[a]	MCRS	Sum of Z-scores
Dbaker	2.22	3.11	5.29	0.08	0.05	0.09	1.24	1.28	14.39
LEE	0.38	0.48	2.72	0.05	0.01	0.01	0.12	14.31	8.35
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22	7.46
FEIG_REFINE	0.42	0.52	1.97	0.02	-0.05	0.05	1.04	-12.97	6.95
LevittGroup	-0.18	-0.36	1.16	-0.00	0.01	0.02	0.29	6.28	6.52
TASSER	-0.41	-0.42	1.01	0.00	-0.05	-0.01	-0.01	1.18	6.28
SAM-T08-HUMAN	-0.71	-0.44	2.93	0.02	0.03	0.03	0.11	9.87	6.28
FAMSD	-0.67	-1.05	1.84	0.03	-0.04	0.04	0.19	3.70	6.06
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.81
A-TASSER	-0.36	-0.91	0.16	-0.01	0.01	-0.04	0.43	2.58	4.81
Tripos_08	-0.72	-0.90	1.10	0.06	-0.04	0.01	0.07	6.77	4.74
BATES_BMM	-1.32	-2.33	0.04	0.04	-0.03	-0.02	0.19	6.35	3.64
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70	3.23
Kolinski	-0.93	-0.93	0.57	-0.07	0.01	0.00	-0.06	7.51	3.06
Jacobson	-1.81	-2.74	-1.55	-0.04	-0.07	0.04	0.15	1.12	2.73
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54	2.04
POISE	-3.99	-5.42	-1.69	-0.02	-0.02	0.02	1.29	10.43	1.24
Abagyan	-1.07	-1.63	-2.90	-0.06	-0.06	0.01	-0.39	0.13	0.76
SAMUDRALA	-3.59	-2.36	-0.86	0.06	-0.03	0.00	0.69	11.90	-1.12
FAMS-multi	-5.83	-6.81	-2.48	-0.00	-0.10	0.05	-0.65	-12.65	-4.00
Keasar	-3.43	-5.87	-3.00	-0.01	0.04	0.03	0.19	-49.31	-4.55
MidwayFolding	-4.94	-6.20	-6.15	-0.06	-0.12	-0.03	-0.83	-11.49	-8.23
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82	-8.43
Jones-UCL	-7.14	-9.48	-7.87	-0.06	-0.07	-0.03	-0.59	-8.61	-10.30
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.01	0.09	-6.19	-30.92
POEM	-12.60	-17.00	-11.77	-0.21	-0.41	-0.12	-1.79	-27.60	-32.00

Per ciascun gruppo fu scelto un singolo modello per ogni target basato sulla somma degli z-scores. I gruppi sono ordinati secondo la somma degli z-scores dei loro modelli

migliori. In tabella 4 abbiamo scelto per ogni target il modello che dà il più grande miglioramento per ciascuna metrica separatamente.
Per esempio, la colonna GDT-TS indica il cambiamento medio in GDT-TS per ciascun gruppo, dove la media è sulla struttura con il più alto GDT-TS per ogni target.
Sulla base di questi risultati, GDT-HA, GDT-TS, e la frazione dei corretti legami idrogeno del mainchain sono le metriche più difficili da migliorare, mentre circa tutti i gruppi hanno migliorato MCRS, MolProbity e la frazione dei corretti legami idrogeno del sidechain. Complessivamente, i gruppi migliori furono in grado di migliorare la maggior parte delle metriche, di solito con una piccola perdita in GDT-TS e GDT-HA.
Mentre riteniamo che la nostra analisi fornisca un impressione generale dei gruppi che ebbero successo e della performance complessiva in questo campo, un confronto tra gruppi vicini nelle classifiche è improbabile che fornisca risultati significativi, come già notato.Calcoli statistici più significativi si potranno avere in futuro se ci saranno più gruppi o più target.

Tabella 4:

Group	GDT-TS	GDT-HA	GDC-SC	Correct rotamers	Correct MC H-bonds	Correct SC H-bonds	MolProbity[a]	MCRS
Dbaker	2.38	3.38	6.24	0.10	0.07	0.13	1.39	2.41
FEIG_REFINE	0.89	0.93	2.83	0.03	-0.03	0.07	1.25	-3.06
LEE	0.59	0.76	3.27	0.07	0.03	0.03	0.26	20.33
PS2-server	0.25	0.18	-0.05	-0.05	-0.01	-0.03	-0.20	12.70
LevittGroup	0.19	-0.01	2.03	0.01	0.04	0.05	1.37	16.16
Null	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A-TASSER	-0.01	-0.61	0.69	0.02	0.04	-0.02	0.56	5.32
SAM-T08-HUMAN	-0.12	-0.28	3.22	0.03	0.05	0.05	0.43	14.55
FAMSD	-0.38	-0.57	2.06	0.04	-0.01	0.05	0.29	6.89
TASSER	-0.41	-0.42	1.04	0.01	-0.05	-0.01	0.15	4.26
Abagyan	-0.51	-1.39	-2.12	-0.05	-0.06	0.03	-0.33	1.16
tripos_08	-0.67	-0.85	2.23	0.06	-0.04	0.02	0.07	9.66
Kolinski	-0.93	-0.93	0.57	0.00	0.01	0.14	0.52	7.51
BATES_BMM	-1.22	-2.20	0.68	0.08	-0.01	0.02	0.33	10.80
YASARARefine	-1.64	-3.10	1.02	-0.00	0.03	0.05	1.77	19.22
Jacobson	-1.76	-2.64	-1.15	-0.01	-0.06	0.08	0.33	3.63
Keasar	-3.27	-5.87	-2.04	-0.00	0.06	0.05	0.34	-42.41
SAMUDRALA	-3.36	-2.05	0.08	0.06	-0.02	0.02	0.72	14.02
Xianmingpan	-3.56	-4.89	-2.25	-0.03	-0.04	0.09	1.13	6.54
POISE	-3.83	-5.27	-1.63	-0.01	-0.01	0.03	1.44	14.27
MidwayFolding	-4.94	-6.20	-5.83	-0.04	-0.10	-0.02	-0.81	-10.93
FAMS-multi	-5.50	-6.44	-1.18	0.03	-0.09	0.07	-0.60	-8.49
EB_AMU_Physics	-6.19	-8.54	-7.32	-0.14	-0.06	0.02	0.00	-4.82
Jones-UCL	-6.99	-9.22	-6.41	-0.03	-0.05	0.00	-0.45	-0.24
POEM	-12.60	-17.00	-10.38	-0.20	-0.40	-0.10	-1.79	-21.71
Elofsson	-20.82	-19.25	-12.98	-0.00	-0.24	-0.00	0.10	-5.69

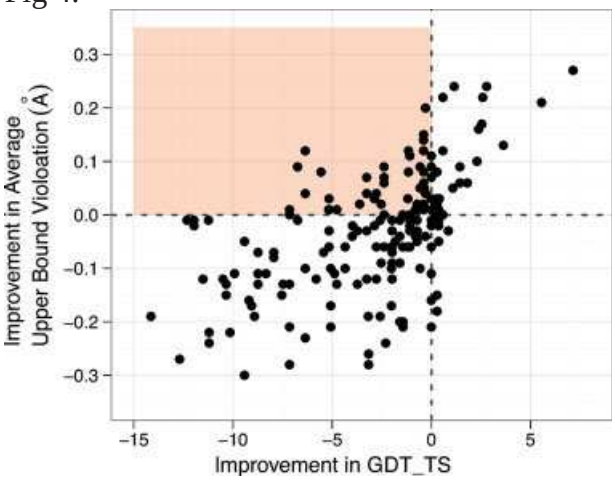
Tutte le analisi presentate finora hanno confrontato un modello di predizione con un modello derivato dai dati sperimentali. Quasi tutte le metriche che abbiamo usato nelle nostre analisi, compreso GDT-TS, richiedono la definizione di una singola struttura sperimentale.
Ovviamente, le proteine sono strutture flessibili e dinamiche e assumono diverse conformazioni, che danno origine all'insieme nativo.
Sia la cristallografia che la risonanza magnetica cercano di modellare l'eterogeneità conformazionale, ma in modi diversi. Tipicamente, le strutture NMR sono riportate come un esplicito insieme contenente 20-50 conformazioni diverse. Per CASP8, fu calcolata la struttura NMR media e poi fu scelta come struttura sperimentale la singola struttura dell'insieme che era più vicina alla media. Le strutture determinate dalla cristallografia generalmente consistono di un insieme di coordinate atomiche medie con l'eterogeneità di ciascuna posizione atomica modellata usando una distribuzione gaussiana isotropica.
Dopo la selezione di una singola struttura sperimentale, il protocollo CASP standard consiste poi nell'omettere qualsiasi regione che appare non strutturata o molto flessibile. Una qualsiasi eterogeneità rimasta, codificata da fattori B cristallografici o insiemi NMR, viene poi ignorata. Il difetto di questo protocollo è di ridurre un insieme di molte possibili conformazioni in una singola struttura.
Il modo ideale di valutare predizioni di strutture sarebbe confrontare un insieme di strutture prodotte da ciascun gruppo con insiemi di dati sperimentali medi.
Qui noi cerchiamo almeno di confrontare singoli modelli strutturali con insiemi di dati. Non trattiamo le cinque strutture sottomesse dai gruppi come un insieme, poiché ai gruppi non fu detto di trattare le loro predizioni in questo modo. In futuro, potrebbe essere interessante permettere ai gruppi di sottomettere un insieme di strutture, ciascuna delle quali essi ritengano essere ugualmente probabile e giacere all'interno del bacino della nativa.

Target con struttura nativa misurata tramite spettroscopia NMR

Ci sono strutture che sono più vicine di altre ai dati NMR ed hanno però valori più bassi di GDT-TS.Ciò suggerisce che per valutare strutture proteiche già vicine alla nativa, può essere preferibile usare misure di qualità basate su insiemi derivati sperimentalmente,, oltre a metodi, come GDT-TS, basati su una singola struttura.
La spettroscopia NMR può fornire informazioni per la struttura e per le dinamiche delle proteine.
Le intensità di NOE (nuclear overhauser effect) sono tra i parametri NMR più importanti per la determinazione della struttura perché essi forniscono informazione della distanza inter-proton.
Una coppia di atomi di idrogeno (i,j) si considera violare il limite superire di distanza NOE (NOE upper bound distance) R_{ij} quando $v_{i,j} = r_{i,j} - R_{ij}$ è positiva (dove $r_{i,j}$ è la distanza tra l'atomo di idrogeno i e j).
I limiti superiori di distanza sono stimati basandosi sui dati NOE sperimentali.
Perciò, piuttosto che confrontare con una singola struttura sperimentale stiamo invece confrontando con un insieme di limiti superiori di distanze determinate dai dati sperimentali. La violazione $v_{i,j}$ è considerata nulla se il suo valore è negativo.Il limite superire medio di violazione (Average Upper Bound Violation) è calcolato come

$v = (1/N) \Sigma_{i,j} v_{i,j}$, mediando sulle N distanze derivate dai valori NOE determinati sperimentalmente.

Fig 4:



In figura 4 come via alternativa per valutare il successo delle predizioni, l'analisi dell' Average Upper Bound Violation è confrontata con la metrica GDT-TS tradizionale. E' rappresentato il miglioramento nell' Average Upper Bound Violation per tutti i modelli raffinati ($-\Delta v$) verso il miglioramento in GDT-TS (ΔGDT) per tutti i target NMR tramite confronto con i corrispondenti modelli non raffinati. C'è correlazione tra il miglioramento in GDT-TS e l'accordo con i dati NMR, cioè più grande è il miglioramento in v per i modelli raffinati, più grande è il miglioramento nel punteggio GDT. L'accordo con i dati NMR è quantificato calcolando l'Average Upper Bound Violation per l'insieme di limiti di distanza derivati dai valori di NOE osservati sperimentalmente. Comunque, la correlazione non è perfetta; ci sono modelli che sono peggiori secondo GDT-TS ($\Delta GDT\text{-}TS<0$) che hanno miglior accordo con i dati NMR rispetto al modello non raffinato e possono essere identificati usando l'Average Upper Bound Violation ($-\Delta v>0$) (regione ombreggiata, Fig 4). Però si osserva che la maggior parte dei modelli migliori selezionati da GDT-TS sono raramente considerati peggiori dall'Average Upper Bound Violation. In futuro, il confronto con le misure NMR può essere utile nell'identificare predizioni di successo oltre a quelle selezionate dai metodi tradizionali di valutazione come GDT-TS. La tabella 5 fa vedere il miglioramento medio nel NOE upper bound violation da parte dei gruppi.

Tabella 5:

Group	Number of targets[a]	$\langle Improvement \rangle_{Bestmodel}$	$\langle Improvement \rangle_{Model1}$
Dbaker	5	0.19	0.07
Jacobson	2	0.15	0.09
Lee	5	0.13	0.10
SAM-T08-HUMAN	5	0.12	0.08
SAMUDRALA	5	0.06	0.02
YASARARefine	5	0.04	0.04
FEIG-REFINE	4	0.03	-0.05
Tripos_08	1	0.03	-0.12
Jones-UCL	5	0.03	-0.07
LevittGroup	5	0.01	-0.01

Group	Number of targets[a]	<Improvement>$_{Bestmodel}$	<Improvement>$_{Model1}$
FAMSD	5	0.00	-0.03
Null	5	0.00	0.00
BATES_BMM	5	-0.02	-0.02
POISE	5	-0.05	-0.14
Abagyan	4	-0.06	-0.07
A-TASSER	5	-0.08	-0.20
Xianmingpan	4	-0.08	-0.08
MidwayFolding	5	-0.08	-0.14
FAMS-multi	5	-0.13	-0.17
Elofsson	5	-1.93	-1.93

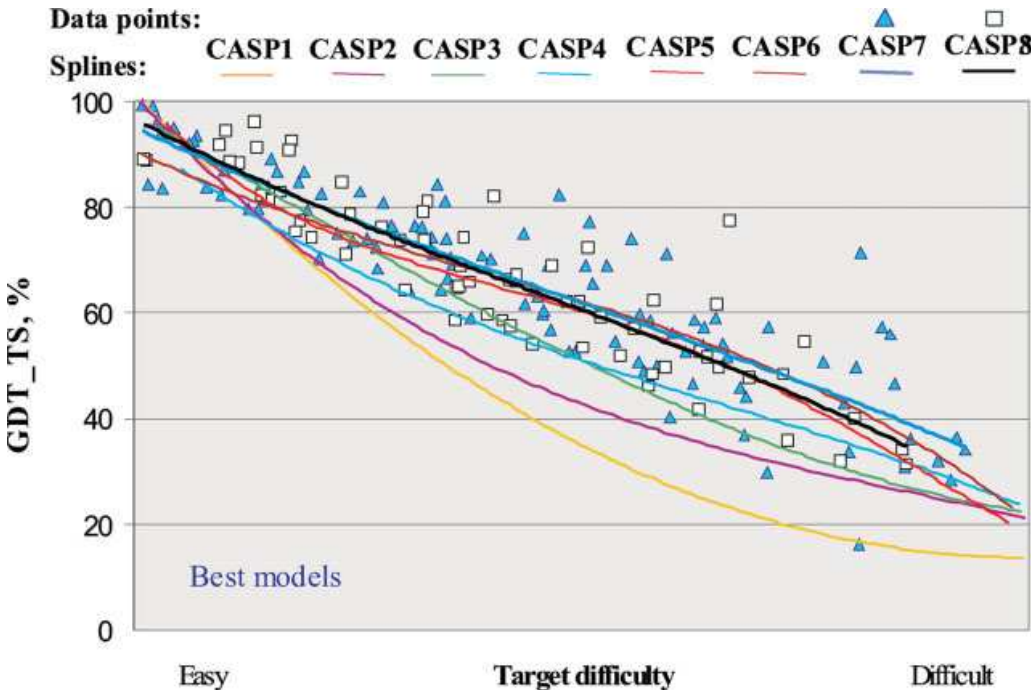
[a]Ci furono quattro target NMR, ma TR429 contiene due domini che furono analizzati separatamente.

Ci sono quattro target NMR, ma TR429 contiene due domini che sono stati analizzati separatamente. Se classificati secondo il loro modello migliore, più della metà dei gruppi sono in grado di fare meglio del raffinamento nullo. Questa performance è chiaramente migliore di quella per GDT-TS (solo quattro gruppi). Ciò dipende da un certo numero di fattori. Prima di tutto, GDT-TS usa un solo insieme di coordinate fisse del carbonio alpha, mentre l'analisi NMR usa un insieme di limiti superiori che è intrinsecamente più flessibile (tutte le distanze inferiori all'upper bound sono considerate ugualmente "buone"). In altre parole, i limiti superiori di distanza permettono la flessibilità strutturale. In secondo luogo, questa flessibilità è evidente dal fatto che il modello dei dati NMR è un insieme piuttosto che una singola struttura. Questo insieme di strutture deve ridursi a una singola struttura per usare metriche quali GDT. Per il CASP8 fu scelta la struttura più vicina alla media dell'insieme. Comunque, in nessun modo una singola struttura può codificare l'eterogeneità conformazionale che è presente nei dati sperimentali. Aggiungiamo poi che l'analisi presentata è estremamente semplicistica. Calcolare le violazioni dei limiti superiori di distanza è forse il modo più semplice di confrontare le strutture con i dati NMR. Nella refinement competition ci furono solo quattro target NMR e perciò le statistiche sono estremamente limitate. Non è significativo confrontare le classificazioni relative dei diversi gruppi per il troppo piccolo insieme di dati. Al più ogni gruppo fu giudicato solo su cinque modelli.

3.4 Risultati del CASP 8 nel contesto degli esperimenti precedenti[4]

Qualità complessiva dei modelli

Figura 2:



Negli anni la correttezza della previsioni è aumentata ma purtroppo si nota una assenza di miglioramento nelle ultime tre edizioni.

Il grafico va letto in questo modo:

sull'asse verticale c'è un indice di correttezze della miglior previsione GDT_TS % (100 % indica che la previsione equivale alla realtà sperimentale), sull'asse orizzontale c'è la difficoltà della previsione da effettuare (complessità della proteina).

Già nelle prime edizioni del CASP (ad esempio la prima in arancio) si avevano delle ottime previsioni per i target semplici , ma dei pessimi risultati per quelli complessi.

Con l'andare degli anni, e quindi delle edizioni del CASP, le linee (cioè l' andamento medio) si alzano sulla destra: questo significa che i modelli predittivi funzionano sempre meglio anche con le proteine complesse. Nelle ultime tre edizioni questo trend al rialzo si è fermato anche se alcune singole previsioni ,rappresentate dai singoli quadrati o triangoli ,sono eccellenti. I laboratori Baker con i loro progetti Roretta@home e simili si sono dimostrati sempre tra i migliori alle varie edizioni del CASP a cui hanno partecipato. Anche POEM@home ha fatto una discreta figura al CASP8.

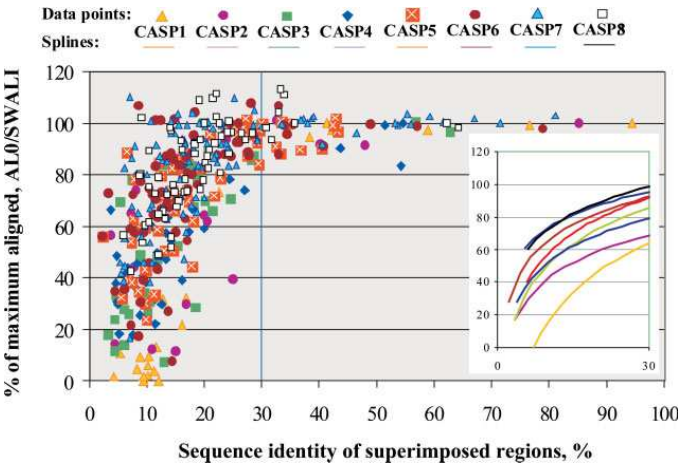
Ci furono tre sviluppi positivi:

1. per l'insieme dei modelli migliori relativi a ciascun target, CASP8 registrò il più alto numero di modelli dove l'accuratezza dell'allineamento va oltre il massimo ottenibile dal miglior templato

2. miglioramento dell'accuratezza nel modellare regioni che non sono presenti nel miglior templato (structurally nonconserved regions)
3. diminuzione della perdita della qualità dei modelli che deriva dalla selezione di modelli non ottimi come i migliori.

1. Accuratezza nell'allineamento

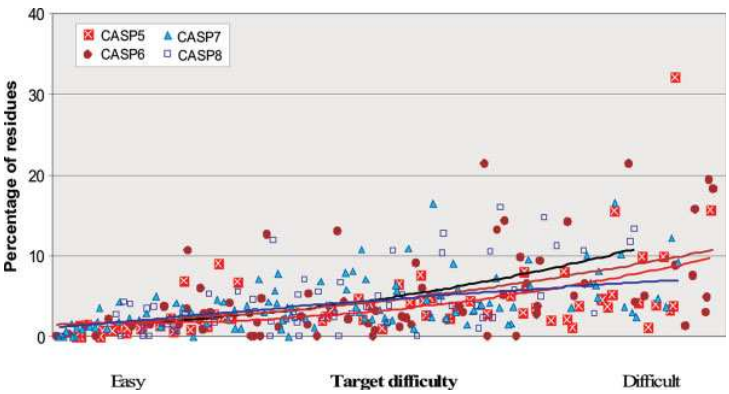
I modelli ottenuti con metodi template –based sono più accurati di ciò che può essere ottenuto semplicemente copiando il miglior templato a disposizione.
Fig 3:



La figura 3 mostra l'accuratezza nell'allineamento (ALO) dei migliori modelli per tutti i CASP, come una percentuale del maximum alignability (SWALI, Smith-Waterman alignment score). Accuratezza nell'allineamento più grande di quel limite riflette dettagli addizionali nel modello che non si trovano nel miglior templato.
L'andamento logaritmico delle linee illustra continuo progresso in questo aspetto. Ci sono 22 target su tutti i CASP dove l'accuratezza supera per più del 2% il maximum alignability.Tra questi, 9 target provengono dal CASP8 (13,6 %dei target totali), 8 target dal CASP7 (7,8%), 4 target dal CASP6 (5,6%), 1 target dal CASP5 (1,6%).
Secondo questi dati quindi c'è stato progresso.

2.Percentuale di residui correttamente posizionati nei migliori modelli ma non presenti nel singolo miglior templato.

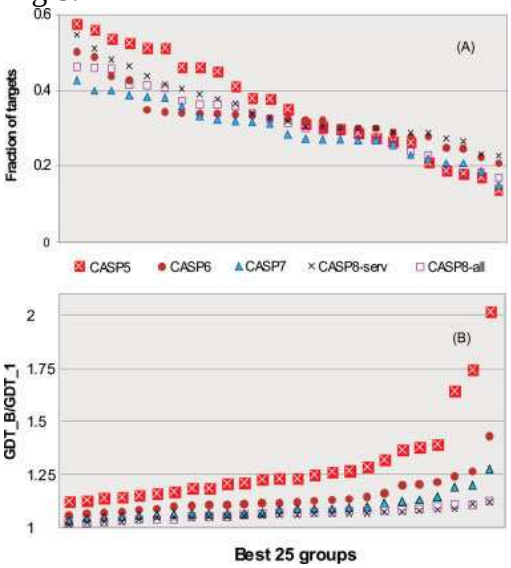
Fig6:



Le regioni non presenti nel miglior templato possono essere modellate in tre modi: combinando caratteristiche di altre strutture templato disponibili, con metodi di raffinamento o con l'uso di metodi template – free.
Un residuo è correttamente allineato se il suo atomo C_α è entro 3,8Å dalla posizione sperimentale. L'andamento quadratico indica che c'è stato progresso per i target più difficili. Ci sono 11 targrt nel CASP 8 per i quali i migliori modelli coprono più del 10% dei residui nel target rispetto al miglior templato.

3. Riconoscimento dei modelli migliori tra il decoy set

Fig 8:



Nel CASP i gruppi possono sottomettere alla valutazione al più cinque modelli per ogni target. Essi devono indicare il modello numero 1 come il migliore dei cinque.Vediamo negli ultimi 4 CASP l'abilità dei predictors di selezionare il modello migliore e la perdita nella qualità dei modelli dovuta all'incapacità di riconoscere il migliore.
Fig 8 rappresenta la frazione dei target dove i migliori 25 gruppi (in termini di GDT_TS medio) effettuarono una selezione corretta del loro modello migliore.
I gruppi sono ordinati secondo la loro abilità di selezione. Per il CASP8 i risultati sono dati per tutti i gruppi su 66 domini human/server (CASP8-all) e per i gruppi server su tutti i 164 domini (CASP8-serv).
I gruppi server mostrano migliore abilità di selezione dei gruppi human.
I migliori selectors furono in grado di riconoscere i loro modelli migliori per il 46% dei target nel CASP8. Nel CASP5 si osserva una frazione del 58%. Questa apparente performance inferiore nel CASP8 può essere fuorviante –nel CASP5 ci fu più variabilità tra i modelli sottomessi dai vari gruppi e quindi risultò più facile scegliere.
La fig 8b rappresenta il rapporto medio GDT_TS tra il miglior modello e il modello scelto come numero 1 per tutti i target nei CASP5-8. I punti nel grafico relativi al CASP8 sono più bassi di quelli relativi agli altri CASP e quindi c'è una diminuzione nella perdita della qualità dei modelli dovuta ad una selezione non ottima. Il rapporto peggiore nel CASP 8 è pari a 1,125 (il punto più a destra nel grafico) che è molto migliore rispetto al risultato degli altri CASP.

Nel passato portò molto progresso lo sviluppo di nuove tecniche quali metodi di allineamenti multipli di sequenze e procedure di assemblamento di frammenti.
Chiaramente oggi servirebbero nuovi metodi.

3.5 Valutazione delle predizioni nella categoria QA del CASP 7[5][7]

Nel CASP 7 fu introdotta una nuova categoria (qualità assessment category, QA): a coloro che parteciparono fu dato un insieme di modelli prodotti dai server e fu richiesto di predire la "qualità" di ciascun modello (la sua distanza dalla struttura nativa).
Dopo che le strutture furono rese disponibili, Andriy Kryshtafovych (Università della California, Davis, CA, USA) e Anna Tramontano (Università di Roma ' La Sapienza') analizzarono la correttezza delle predizioni in questo campo.
La qualità di un modello di struttura proteica determina il suo corretto uso (ad esempio modelli di moderata qualità sono di limitato uso per applicazioni come il docking). Per modelli TB esistono regole per valllutare prima la qualità di un modello sulla base della similitudine di sequenza tra la proteina target e il templato omologo o sulla base della somiglianza a coppie tra gli elementi di un allineamento di sequenze multiple. Tuttavia il fold recognition e i metodi basati sui frammenti possono produrre modelli di qualità anche nel caso di assenza di rivelabile similitudine di sequenza con proteine di struttura nota.
In questi casi la similitudine di sequenza non può dunque essere usata per stimare in anticipo la qualità di un modello. Metodi capaci di predire la qualità di un modello solo sulla base delle sue coordinate sarebbero perciò di grande valore per chi li usa.
Questo è il contesto all'interno del quale gli organizzatori del CASP decisero di introdurre una nuova categoria di predizione . Per condurre l'esperimento model qualità prediction, gli organizzatori si avvantaggiarono del fatto che i modelli prodotti dai server che partecipano al CASP erano resi disponibili a tutti i predictors nel corso dell'esperimento, poco dopo il rilascio dei target.
I risultati dei server, distribuiti pubblicamente , sono tradizionalmente usati dai gruppi human come punto di partenza per il loro lavoro. Nel CASP 7 , i modelli dei server furono usati anche come target per la predizione della qualità dei modelli. Ai gruppi che parteciparono fu chiesto di produrre stime della qualità di questi modelli prima che la corrispondente struttura sperimentale fosse disponibile. Ai predictors fu data l'opportunità di sottomettere predizioni di qualità per le strutture considerate complessivamente (Model Qualità Mode 1, QM1, un valore per modello) e/o su una base residuo per residuo (Model Qualità Mode 2, QM2).

QM1 : qualità complessiva di un modello

23.864 modelli per 95 target furono soyyomessi dai server e resi disponibili ai predictors nel sito web del CASP 7 (http://predictioncenter.org/casp7) .
Ventotto gruppi parteciparono all'esperimento QM1, nove dei quali sottomisero anche predizioni per QM2. Nel QM1 fu richiesto ai partecipanti di sottomettere un punteggio di qualità compreso tra 0.0 e 1.0 per ciascun modello. La qualità si intende su scala assoluta e direttamente correlata con la qualità del modello cioè un punteggio pari a 1.0 identifica un modello coincidente perfettamente con la proteina target. Comunque, alcuni gruppi sottomisero valori relativi per ciascun target, cioè assegnarono un valore di 1.0 al miglior modello per ciascun target , indipendentemente da quanto buono fosse. Come mostriamo dopo , la nostra valutazione delle predizioni prende in considerazione entrambe le possibilità .Nel QM2 fu richiesto di assegnare una stima dell'errore (in Å)

a ciascun residuo di ciascun modello, o un valore nullo (indicato con "X") nel caso si scelga di non sottomettere la stima per uno o più residui .Per valutare il grado di successo dei vari gruppi abbiamo calcolato il coefficiente di correlazione di Pearson tra la qualità predetta e il valore di GDT-TS per ciascun modello (calcolato e reso disponibile dal Prediction Center).
Abbiamo calcolato il coefficiente di Pearson sia su base target per target (per considerare casi dove i predictors hanno normalizzato i loro assegnamenti di qualità nel range 0.0-1.0 per ciascun target) sia considerando tutti i modelli per tutti i target insieme per valutare se la qualità predetta fosse indicativa dell'effettiva accuratezza di ciascun modello, indipendentemente dal target specifico.
I risultati discussi qui furono ottenuti usando il coefficiente di correlazione di Pearson stimato (r) tra le variabili di interesse. Comunque, la classificazione finale dei gruppi non è influenzata dalla scelta dello schema di punteggio (risultati non mostrati).
Abbiamo calcolato la distribuzione dei valori di r per ciascun target, la sua media e la deviazione standard, e abbiamo assegnato uno z-score a ciascuna delle predizioni.
La distribuzione dei valori di r è solo approssimativamente normale, come verificato attraverso lo Shapiro-Wilk test, perciò gli z-score non hanno in teoria un significato statistico corretto. Ciò nonostante pensiamo che lo z-score è ancora appropriato come sistema di punteggio in questo contesto, dato che esso considera la difficoltà di predire la qualità dei modelli per ciascun target. Perciò, la somma su tutti i target degli z-score di un metodo è una stima ragionevole della sua performance complessiva . La somma degli z-score su tutti i target, per ciascun gruppo partecipante all'esperimento QM1 è mostrata in figura 2. Valori negativi di z-score furono fissati a zero con lo scopo di non penalizzare metodi più innovativi.
E' evidente che i gruppi 634 (Pcons) e 556 (Lee) sono molto migliori degli altri.

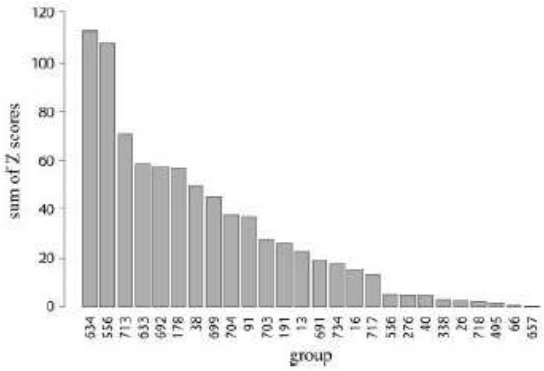
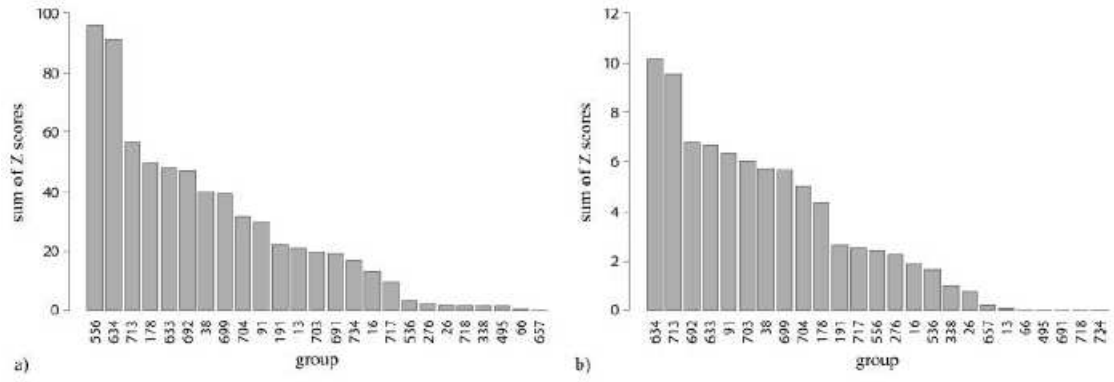


Figura 2

Il significato delle differenze tra i punteggi per diversi gruppi fu valutato attraverso un t-test a coppie sull'insieme comune di modelli predetti. I risultati del t-test stabiliscono che la differenza tra le predizioni ottenute dai gruppi 634 e 556 e quelle ottenute dai rimanenti gruppi è statisticamente significativa (valore di probabilità ≤ 0.01) .
La procedura descritta sopra fu ripetuta dividendo i target nelle loro rispettive categorie di predizione: template- based (TBM) e template-free modelling (FM). Siccome

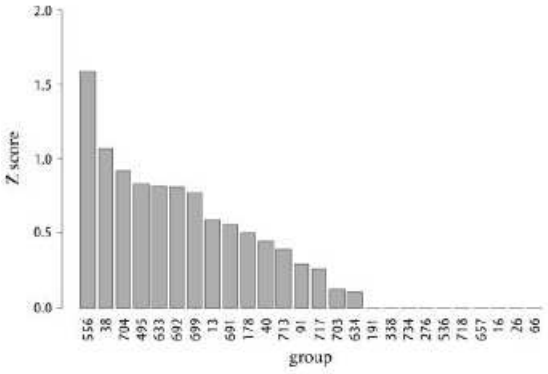
l’assegnamentoi di categoria è basato sui domini mentre le predizioni di qualità considerano l’intero target, l’analisi basata sulle categorie fu limitata a proteine di singolo dominio (68) e a proteine i cui domini sono stati assegnati alla stessa categoria (22). I risultati sono mostrati in figura 3.
Fig 3:



(a) Punteggi dei gruppi che parteciparono all’esperimento QM1 per target TBM.
(b) Classificazione per target FM. I punteggi sono calcolati su base target per target.

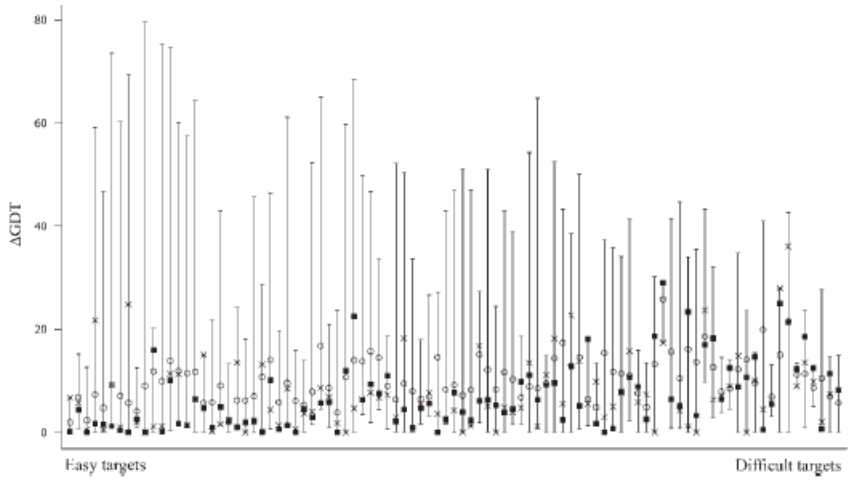
Non ci sono stati molti cambiamenti nei risultati. Il gruppo 713 (Circe-QA) si distingue dagli altri per i modelli FM, sebbene bisogna ricordare che ci sono solo dieci target FM e perciò non è chiaro se questa conclusione possa essere generalizzata.
Successivamente abbiamo calcolato il coefficiente di correlazione globale per tutti i modelli di tutti i target considerati insieme (figura 4).

Fig 4:



In questo caso il coefficiente di correlazione complessivo indica l’abilità dei metodi di assegnare punteggi non relativi. Il significato statistico della differenza tra i coefficienti di correlazione raggiunti da due diversi gruppi fu calcolato come segue:
dati due coefficienti di correlazione di Pearson r_1 e r_2 abbiamo applicato la trasformazione di Fisher $z' = (\ln(1+r) - \ln(1-r)) \times 0.5$.

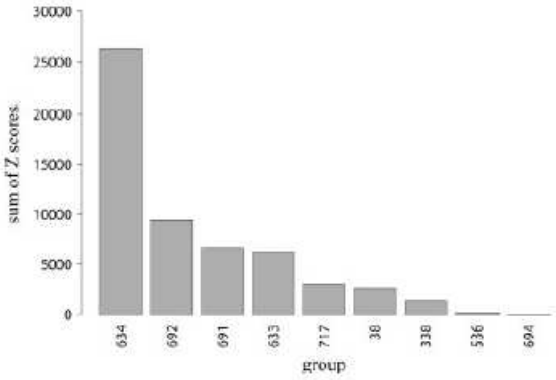
La variabile z'_1-z'_2 ha distribuzione normale con varianza $s^2=1/(n_1-3)+1/(n_2-3)$, dove n_1e n_2 rappresentano il numero di modelli valutati dai due gruppi.
Il valore associato di probabilità (p value) indica la probabilità che le due predizioni con coefficienti di correlazione r_1 e r_2 sono statisticamente indistinguibili. Come mostrato in figura 4 solo il gruppo 556 è migliore rispetto agli altri. C'è un ulteriore parametro di valutazione che noi consideriamo possa essere rilevante dal punto di vista di un utilizzatore, cioè quanto perderebbe un utilizzatore se selezionasse il modello classificato come primo da ciascun metodo.In altre parole cerchiamo quale sia la differenza in GDT-TS (ΔGDT) tra il modello classificato come primo da un metodo e il miglior modello per ciascun target. I risultati sono piuttosto diversi per target diversi (figura 5);
Fig 5:



Valori di ΔGDT per il gruppo 556 (quadrati neri) e 634 (croci) a confronto con il ΔGDT medio di tutti i gruppi (cerchi vuoti) nell'esperimento QM1. Le linee verticali indicano i valori massimo e minimo di ΔGDT medio per ciascun target.
Comunque il ΔGDT medio su tutti i target è compreso tra 3.18 e 27.74. I valori di ΔGDT medio per i gruppi 556 e 634 sono rispettivamente 6.77 e 7.28. I ΔGDT dei gruppi 556 e 634 sono più bassi dei valori medi per 72 e 65 target, rispettivamente, e questa differenza è statisticamente significativa ($P<0.01$).
Se un miglioramento medio di alcune unità di GDT-TS sia significativo dipende ovviamente dall'applicazione specifica del modello.

QM2 : predizioni di qualità residue-based

La valutazione delle predizioni QM2 si basa ancora sul coefficiente di correlazione di Pearson. Abbiamo calcolato la somma degli z-score come descritto sopra (figura 6).
Fig 6:



Punteggi dei nove gruppi che parteciparono all'esperimento QM2.

Il significato statistico delle differenze tra questi punteggi fu valutato attraverso un t test a coppie usando i residui comuni dei modelli comuni tra i vari gruppi. Abbiamo scartato dall'analisi 511 modelli (circa il 2.2% del numero totale di modelli) per i quali furono sottomesse meno di quattro predizioni. Il gruppo 556 non partecipò all'esperimento QM2. Il gruppo 634 è il migliore (figura 6). Abbiamo anche classificato i nove gruppi separatamente per le categorie TBM e FM. Tutti i domini appartenenti a entrambe le categorie (T0304, T0321_1, T0348 e T0382) furono inclusi nella categoria TBM.I risultati sono mostrati in figura 8.

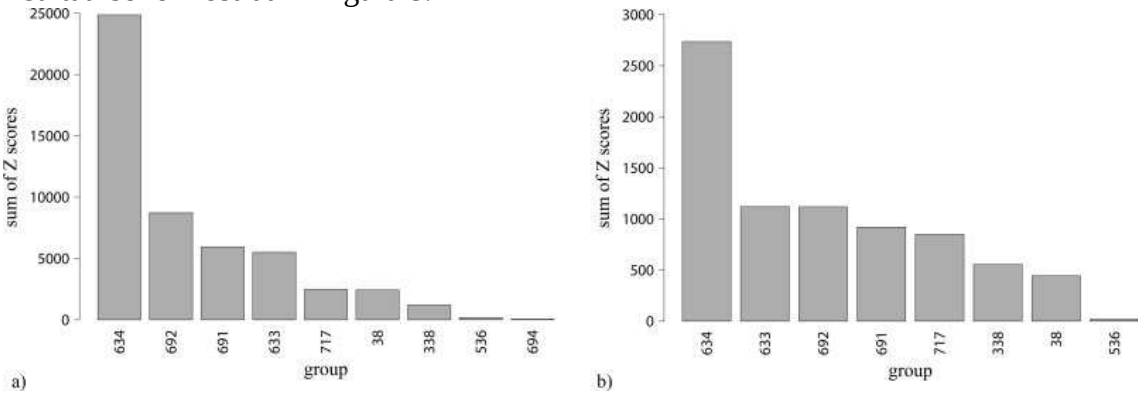


Fig 8

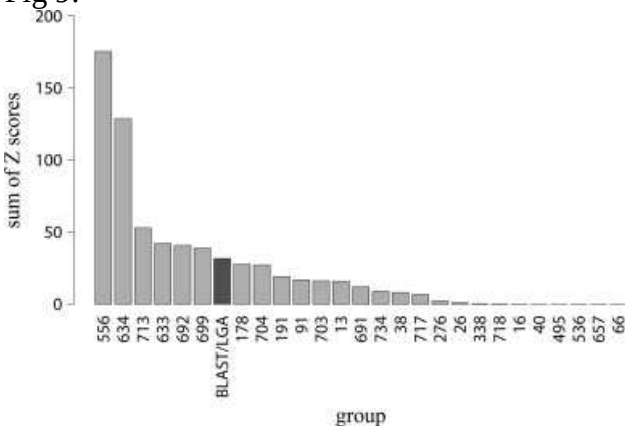
La figura 8 rappresenta i punteggi dei gruppi che parteciparono all'esperimento QM2 per i target nelle categorie TBM e FM (figura a e b, rispettivamente).Il gruppo 634 è migliore rispetto agli altri in entrambi i casi.

Predittore naïve

Abbiamo deciso di verificare se i risultati raggiunti nell'esperimento avrebbero potuto essere riprodotti, o persino migliorati, tramite metodi semplici basati sulla nostra esperienza nei CASP precedenti.

Abbiamo ideato un "predittore naïve" chiamato BLAST/LGA. Questo metodo identifica la proteina di struttura nota più vicina alla proteina target tramite una ricerca PSI-BLAST. Poi esso sovrappone ciascun modello del target al templato identificato usando LGA. La predizione di qualità è semplicemente il punteggio LGA-S diviso per 100 (per normalizzare il risultato nel range 0.0-1.0).

Abbiamo confrontato i risultati ottenuti dal predittore naïve con quelli di tutti gli altri metodi nelle categorie TBM/FM. Modelli per target multi-dominio non furono considerati.
Fig 9:



Confronto della performance del metodo naïve BLAST/LGA con tutti gli altri metodi che hanno sottomesso predizioni QM1 per target TBM.
Per modelli TBM solo 556 e 634 ottennero risultati migliori del predittore naïve, e solo pochi altri metodi raggiunsero risultati di qualità confrontabile (figura 9).

Strategia e metodologia dei gruppi migliori

La metodologia del gruppo 634 (Pcons) è un aproccio di consenso basato dulle predizioni multiple sottomesse da diversi gruppi, e perciò è adatta per selezionare buoni modelli tra un insieme di modelli alternativi, ma non può essere usata per valutare in assoluto la qualità di un modello.
La strategia adottata dal gruppo 556 (Lee), sebbene di successo, è di scarso uso al di fuori degli esperimenti CASP.
Questo gruppo produsse modelli molto buoni per la maggior parte dei target e, successivamente, confrontò tutti i modelli del target con le proprie predizioni, assegnando un valore che dipende dalla distanza del modello analizzato dal proprio.
Di interesse è la strategia usata dal gruppo 713 (Circe-QA) perché basata solo sulle coordinate del modello analizzato. Per questo motivo riteniamo sia interessante seguire i suio sviluppi nel futuro.

3.6 Valutazioni delle predizioni nella categoria high accuracy template-based modeling del CASP7 [10]

Una questione nel comparative modeling è quella del “valore aggiunto” ossia fino a che punto il modello aggiunge informazione oltre l’affermazione che il target assomiglia al templato.
Nel CASP 6 fu concluso che per i target più semplici la difficoltà era relativa ai metodi di raffinamento per migliorare rispetto alla struttura del templato e fu suggerito che più attenzione dovesse essere posta in questo campo. Per questa ragione nel CASP 7 è stata introdotta la categoria high accuracy template-based modeling (HA/TBM). I target sono stati assegnati a questa categoria , quando le predizioni furono chiuse, sulla base di due criteri:

1. per garantire che fosse disponibile una buona struttura templato nel PDB quando furono effettuate le predizioni, sovrapposizioni strutturali devono identificare almeno un templato con punteggio LGA più grande di 80.
2. per garantire che è possibile costruire un buon modello, fu richiesto che almeno un modello avesse un punteggio GDT-TS più grande di 80.

Criteri di valutazione

I punteggi usati in questo lavoro sono stati calcolati usando risultati dal programma LGA: ALO (punteggio di allineamento basato sulla sovrapposizione ottenuta con LGA), LGA-S (sequence independent superposition score), e punteggi GDT (sequence-dependent superposition scores).
Abbiamo calcolato i punteggi z-score per le valutazioni numeriche con la versione high accuracy di GDT (GDT-HA). La qualità delle predizioni del side chain è stata valutata confrontando gli angoli di torsione tra il modello e il target. Differenze negli angoli di torsione furono calcolate usandoi il programma LSQMAN. Dove le catene laterali furono omesse, gli angoli di torsione furono classificati come incorretti. Sono stati valutati quattro punteggi: la frazione di residui con angoli χ_1 predetti entro 15º o 30º e la frazione di residui con entrambi gli angoli χ_1 e χ_2 predetti entro 15º o 30º. Sebbene ci sia maggiore incertezza negli angoli di torsione di residui superficiali , tutti i residui furono inclusi nell'analisi per aumentare il numero di osservazioni.
Come nuovo criterio, abbiamo introdotto una misura che esprime quanto adatto sia un modello per risolvere strutture cristallografiche tramite sostituzione molecolare. Il programma Phaser, usato per risolvere strutture cristallografiche , riporta per ogni potenziale soluzione un punteggio LLG (log-likelihood-gain) il quale misura quanto il modello concorda con i dati.

Risultati

La tabella 1 presenta i risultati di classificazione numerica per i migliori venti gruppi che sottomisero predizioni per i target HA/TBM nel CASP 7. TS indica che furono sottomesse coordinate atomiche da parte del gruppo. n_{HA} è il numero di predizioni sottomesse per domoni di target appartenenti alla categoria HA/TBM, e n_{MR} è il numero di predizioni sottomesse per target usati per sostituzione molecolare. I gruppi sono ordinati in base alla somma dei loro z-score medi per GDT-HA , corrette coppie χ_1/χ_2 e LLG.
Tabella 1:

Group	n_{HA}	Mean GDT-HA Z-score	Mean ALO Z-score	Mean χ_1 Z-score	Mean χ_1/χ_2 Z-score	n_{MR}	Mean LLG Z-score	Sum
TS556	26	0.995	0.727	1.427	1.290	12	0.842	3.127
TS020	26	0.746	0.684	1.242	1.307	12	0.738	2.792
TS249	6	0.590	0.351	0.348	0.349	4	1.731	2.670
TS186	27	0.349	0.289	1.280	1.311	12	0.874	2.534

TS004	28	0.432	0.382	1.405	1.290	12	0.792	2.515
TS671	28	0.654	0.657	0.876	0.933	12	0.616	2.203
TS010	28	0.464	0.562	1.187	1.185	12	0.487	2.136
TS234	2	0.414	0.338	0.865	0.672	2	1.028	2.115
TS664	28	0.588	0.630	0.907	0.924	12	0.510	2.022
TS209	26	0.447	0.353	0.997	0.687	12	0.883	2.016
TS568	28	0.574	0.636	0.768	0.752	12	0.688	2.015
TS559	4	0.448	0.484	0.396	0.449	2	1.105	2.001
TS338	28	0.604	0.522	0.271	0.333	12	1.016	1.954
TS024	28	0.838	0.795	0.561	0.679	12	0.411	1.928
TS650	5	0.502	0.448	0.406	0.558	3	0.862	1.923
TS191	6	0.461	0.395	1.542	1.438	0	—	1.899
TS046	28	0.491	0.412	0.777	0.912	12	0.457	1.859
TS026	26	0.751	0.587	0.813	0.610	12	0.497	1.858
TS047	28	0.373	0.333	1.002	0.969	12	0.426	1.768
TS064	11	0.298	0.231	0.435	0.263	4	1.196	1.758

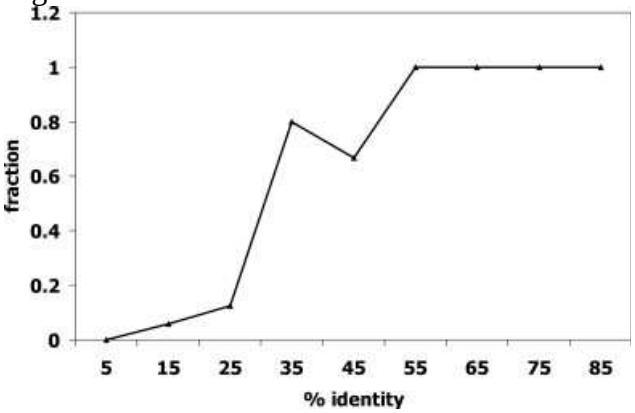
Il gruppo 556 (LEE) è il solo gruppo che raggiunse risultati ottimi secondo tutti i criteri considerati: qualità del fold (in particolare GDT-HA), qualità dei rotameli del side-chain, e qualità dei modelli per sostituzione molecolare.

Targets

Nel CASP 7 sono state valutate dagli assessors predizioni per un totale di 95 target. Un certo numero di questi avevano domini multipli, così ci furono in totale 123 domini. Tra questi 28 domini da 24 target sono stati assegnati alla categoria HA/TBM.

Sebbene la maggior parte dei target hanno potenziali strutture templato con un alto livello di identità di sequenza, come ci si potrebbe aspettare dati i criteri per entrare nella categoria HA/TBM, ci fu un dominio (dominio 2 di T0303)per il quale il templato più vicino aveva identità di sequenza solo del 13%. E' interessante osservare l'effetto dell'identità di sequenza sulla probabilità che un templato mostri un alto livello di similarità strutturale (misurato in questo contesto da un punteggio LGA-S maggiore di 0.8). La figura 1 fa vedere, come funzione dell'identità di sequenza per il templato più vicino, la frazione di domini assegnati alla categoria HA/TBM.

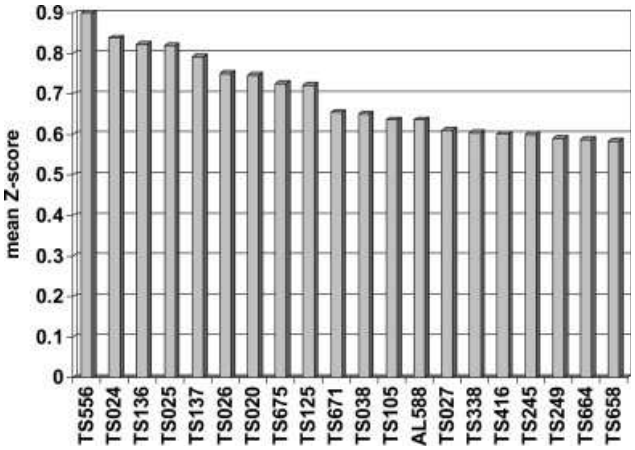
Fig 1:



Come ci si potrebbe aspettare, c'è una buona correlazione tra l'identità di sequenza e la probabilità di un alto punteggio LGA-S.
C'è un salto in corrispondenza del 30% di identità di sequenza, che va d'accordo con l'evidenza che c'è una eccellente possibilità di risolvere una struttura cristallografica tramite sostituzione molecolare con un modello con identità di sequenza pari al 30% o migliore. Inoltre, è stato affermato che generalmente aumenta molto l'accuratezza della modellizzazione template-based per un livello di identità di sequenza pari al 30% [11].

Qualità della predizione del ripiegamento

La valutazione della predizione del fold di concentra sul punteggio GDT-HA.
Giudicati in base al mean GDT-HA z-score, il gruppo 556 (LEE) ha la performance migliore (figura 2).
Fig 2 :



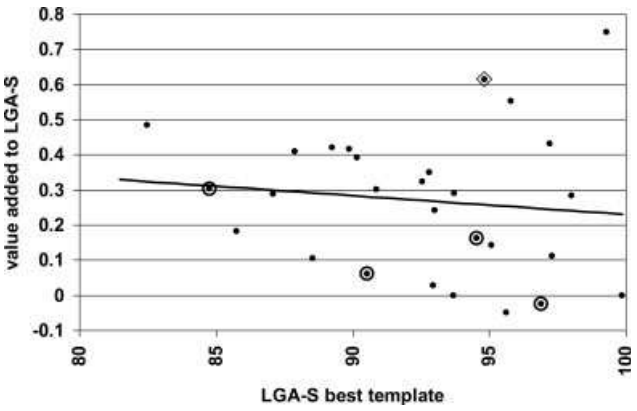
I migliori 20 mean z-score per il criterio GDT-HA.

Valore aggiunto al templato

Un modo per valutare il valore aggiunto al templato è considerare se i punteggi di valutazione sono migliori per i modelli rispetto ai templati sui quali essi sono basati.Un problema è che il modello è basato su un allineamento di sequenza esplicito, mentre l'allineamento di sequenza per il templato deve seere dedotto, preferibilmente dall'allineamento strutturale. Per tener conto di questo fatto, abbiamo guardato i punteggi LGA-S (sequence independent) per i modelli. I dati in figura 3 dimostrano che il miglior modello migliora considerevolmente il miglior templato.

E' possibile aggiungere valore al singolo miglior templato in vari modi. Uno potrebbe essere attraverso metodi di raffinamnento, che sarebbe difficile valutare da questi dati.Un altro potrebbe essere l'uso di templati multipli per costruire il modello. Un indicazione che questo è un importante fattore è che i target con un singolo templato (evidenziati in figura 3) hanno un miglioramento inferiore nel punteggio LGA-S in media rispetto a quelli con più di un templato. Ad esempio è evidente che l'uso di templati multipli ha avuto un impatto significativo nella costruzione dei migliori modelli per il target T0315 (anche evidenziato in figura 3). I modelli migliori per questo target assomigliano alla PDB entry 1J6o per i residui 10-15 e alla struttura 1YIX per i residui 20-25 e alcune annotazioni indicano che entrambi i templati sono stati effettivamente usati simultaneamente in ciascuno dei modelli.

Fig 3 :

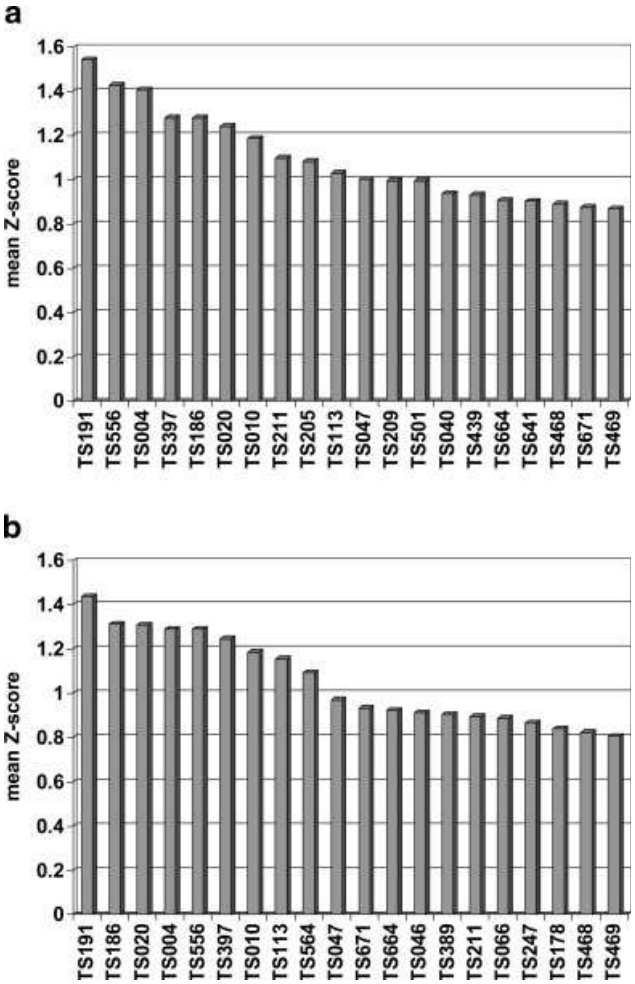


Misura del "valore aggiunto" attraverso la frazione del potenziale miglioramento nel punteggio LGA-S che fu raggiunto.Un modello perfetto avrebbe un punteggio LGA-S pari a 100, così la frazione del miglioramento potenziale è definita come ($LGA\text{-}S_{model} - LGA\text{-}S_{template}$)/($100 - LGA\text{-}S_{template}$). Questo è graficato come funzione del punteggio LGA-S per il templato più vicino. I punti corrispondenti ai target con un singolo templato sono evidenziati con cerchi, mentre il punto corrispondente al target T0315 è evidenziato con un diamante.

Accuratezza degli angoli di torsione

La figura 6 fa vedere lo z.score medio per la predizione degli angoli χ_1 o di entrambi gli angoli χ_1 e χ_2 .Il gruppo 191 (Schomburg-group) ha i risultati migliori per l'accuratezza dei rotameli, ma bisogna notare che questo gruppo sottomise predizioni solo per sei domini su ventotto domini (Tabella 1).

Fig 6:



(a) Migliori venti z-score medi per la frazione di residui per i quali gli angoli χ_1 sono predetti entro 30º.

(b) Migliori venti z-score medi per la frazione di residui per i quali entrambi gli angoli χ_1 e χ_2 sono predetti entro 30º.

Qualità dei modelli per sostituzione molecolare

Una grande frazione delle strutture cristallografiche depositate nel PDB sono risolte usando il metodo della sostituzione molecolare.

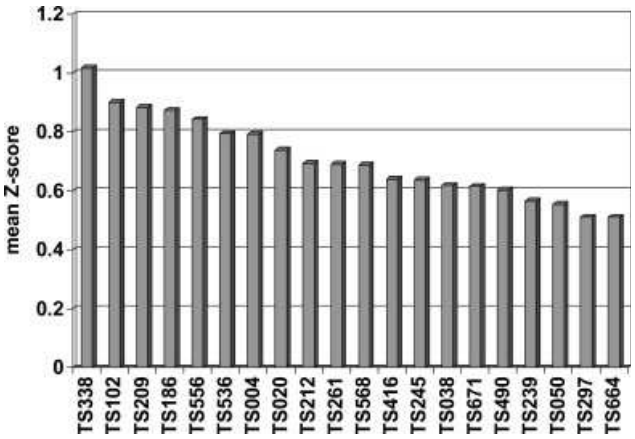
Nella sostituzione molecolare un modello atomico è ruotato e traslato per essere posizionato nell'unità cellulare del cristallo della proteina target, permettendo di avere stime dell'informazione di fase tramite fasi calcolate dal modello. La qualità del modello atomico influenza il successo in due modi. Prima di tutto, modelli migliori danno un segnale più forte nelle ricerche di rotazione e traslazione. In secondo luogo, modelli migliori danno fasi più accurate, dalle quali possono essere calcolate mappe della densità elettronica più chiare così un modello finale può essere ottenuto più facilmente.

Nel passato l'evidenza suggeriva che quando è disponibile un buon templato piuttosto che aggiungere valore, i modelli spesso riducevano il valore delle strutture proteiche omologhe per lo scopo della sostituzione molecolare.

Comunque, ci sono stati segni che gli algoritmi di modellizzazione son migliorati dal punto di vista che essi possono adesso essere utili per la sostituzione molecolare.

Per esempio, il server web CaspR (server web per sostituzione molecolare automatica usando l'omology modeling) ha generato un certo numero di modelli potenziali tramite allineamenti alternativi che sono stati usati poi come input a MODELLER, ed è stato spesso trovato che almeno uno di questi modelli è migliore rispetto al templato originale. Analogamente il gruppo Tramontano ha constatato che un certo numero di modelli sottomessi dai CASP precedenti forniscono migliori modelli per sostituzione molecolare rispetto al singolo miglior templato. Abbiamo perciò esaminato i modelli sottomessi per un certo numero di target HA/TBM, per valutare quale fosse migliore per sostituzione molecolare e se sono oppure no migliori rispetto al miglior templato disponibile.
I modelli sono stati classificati usando il punteggio LLG fornito dal programma Phaser. Fu deludente trovare che solo 33 dei 1588 modelli che furono valutati diedero un punteggio LLG più alto rispetto al singolo miglior templato.Per sette dei dodici target, nessuno dei modelli furono migliori rispetto al singolo miglior templato.
Contrariamente, il gruppo Tramontano trovò miglioramenti rispetto al templato per cinque di sette target selezionati dal CASP 5 e dal CASP 6. La differenza probabilmente è nel criterio di selezione imposto per entrare nella categoria HA/TBM dove fu richiesto che ci fosse un buon templato. Ciò lascia meno spazio per il miglioramento nella modellizzazione.

Fig 7:



Migliori venti mean z-score per il criterio LLG, per gruppi che hanno predetto almeno 10 dei 12 target usati come test per sostituzione molecolare.

Il più alto rate di successo nel migliorare rispetto al miglior templato fu per il gruppo 249 (taylor), che fornì un modello migliore rispetto al miglior templato per tre dei quattro target per i quali essi sottomisero modelli. La figura 7 fa vedere che dei gruppi che sottomisero predizioni per almeno 10 dei 12 target, la migliore performance complessiva fu del gruppo 338 (UCB-SHI).

3.7 Strategie di successo per predizioni template-free (CASP7)

Il gruppo Baker con Rosetta@HOME e il gruppo Zhang furono i gruppi che ebbero maggior successo nella categoria template-free nel CASP7. Entrambi questi gruppi ebbero diverse strategie di predizione per diverse categorie di predizione. Dichiarare che un gruppo è il migliore è difficile, in quanto alcuni gruppi effettuarono predizioni di molti target in più rispetto ad altri, rendendo difficile confrontare i risultati complessivi.
Rosetta@HOME fu il migliore, nel gruppo template-free e nella categoria template-based per proteine con meno di 200 residui e nessuna struttura templato con più del 30%di identità di sequenza. Secondo i punteggi GDT_TS, Baker e Zhang furono entrambi i migliori e sono indistinguibili, ma Baker fu molto migliore di alcuni gruppi che furono indistinguibili (né migliori né peggiori) di Zhang.
Perciò, il gruppo Baker fu il migliore per le predizioni template-free. Il gruppo Zhang fu migliore rispetto al gruppo Baker nelle predizioni template-based di proteine più lunghe con identità di sequenza più grande del 20% con un templato di struttura nota e iun predizioni di proteine con identità di sequenza maggiore del 30% con un templato di struttura nota. Il gruppo Baker ebbe due gruppi separati che parteciparono alla competizione CASP7: Robetta Server e Rosetta@HOME.
Per predizioni template-free, entrambi i gruppi usarono la stessa strategia basata sull'assemblamento di frammenti per generare un gran numero di conformazioni, ma Rosetta@HOME effettuò poi un raffinamento su ciascuna conformazione.
Rosetta@HOME fu in grado di effettuare una computazione così estesa su ciascuna conformazione grazie al suo enorme potere computazionale.
Una media di 500.000 CPU l'ora furono disponibili per ciascun target attraverso il calcolo distribuito. Rosetta@HOME ebbe più successo di Robetta Server per predizioni template-based, ma entrambi i gruppi furono classificati tra i migliori 20 secondo i punteggi GDT_TS.
L'algoritmo sviluppato dal laboratorio Zhang è chiamato TASSER, significa "Thresding, Assembly, Refinement". Nel CASP7 usarono una versione modificata di TASSER chiamato I-TASSER. Il gruppo Zhang partecipò come gruppo human, "Zhang", e come gruppo server, "Zhang Server", nel CASP7.
Il gruppo human fu un po' migliore del gruppo server, specialmente nella categoria free-modeling. Entrambi i gruppi usarono sostanzialmente lo stesso sistema per predire la struttura proteica, ma ci furono alcune differenze. L'algoritmo I-TASSER usa una vasta varietà di tecniche per produrre predizioni eccellenti: programmazione dinamica, modelli lattice, threading, allineamento, un campo di forze molto complicato, simulazioni Monte Carlo, reti neurali, e separazione di domini. La programmazione dinamica è usata per trovare le migliori strutture templato per il threading. Sono anche inclusi termini dipendenti dalle predizioni di struttura secondaria, come termini legati ai legami idrogeno tra gli atomi del backbone. Questo potenziale è poi usato per mettere insieme le strutture su un modello lattice con simulazioni Monte Carlo. Dopo di che, le strutture sono raffinate. I-TASSER predice solo gli atomi C_α. Usa altri algoritmi per aggiungere gli atomi del backbone e per costruire le catene laterali.
Il gruppo HUMAN usò ispezione visiva per separare i domini delle proteine, mentre il gruppo server separò i domini tramite metodi computazionali, e questa sembra essere una delle ragioni per cui il gruppo human ebbe risultati migliori per i modelli template-free.

3.8 Conclusioni

Si può dire che il lavoro di valutazione delle predizioni e dei metodi (svolto da assessors indipendenti per le diverse categorie) sia molto importante per almeno tre ragioni:

• per evidenziare le aree dove la comunità che partecipa non è stata in grado di progredire in modo significativo;

• per rimuovere false affermazioni di successo da un importante campo come la predizione di strutture ;

• spingere la comunità che partecipa a sviluppare e migliorare i metodi.

Bibliografia

[1] Proteins 2009; 77(Suppl 9):18-28.Evaluation of template-based models in CASP8 with standard measures.Domenico Cozzetto, Andriy Kryshtafovych, Krzysztof Fidelis, John Moult, Burkhard Rost, Anna Tramontano.

[2] Proteins 2009; 77(Suppl 9):29-49.The other 90% of the protein: Assessment beyond the Cαs for CASP8 template-based and high-accuracy models.Daniel A. Keedy, Christopher J. Williams, Jeffrey J. Headd, W. Bryan Arendall, Vincent B. Chen, Gary J. Kapral, Robert A. Gillespie, Jeremy N. Block, Adam Zemla, David C. Richardson, Jane S. Richardson.

[3] Proteins 2009; 77(Suppl 9):66-80.Assessment of the protein-structure refinement category in CASP8.Justin L. MacCallum, Lan Hua, Michael J. Schnieders, Vijay S. Pande, Matthew P. Jacobson, Ken A. Dill.

[4] Proteins 2009; 77(Suppl 9):217-228.CASP8 results in context of previous experiments.Andriy Kryshtafovych, Krzysztof Fidelis, John Moult.

[5] Proteins 2007; 69(Suppl 8):175-183.Assessment of predictions in the model quality assessment category.Domenico Cozzetto, Andriy Kryshtafovych, Michele Ceriani, Anna Tramontano.

[6] T.Kiefhaber and R.L. Baldwin. Intrinsic stability of individual alpha-helices modulates structure and stability of the apo-myoglobin molten globule.J. of Mol. Biol.,252:122-132,1995.

[7] FEBS Journal 274 (2007) 1651-1654.An account of the Seventh Meeting of the Worldwide Critical Assessment of Techniques for Protein Structure Prediction.Anna Tramontano.

[8] Proteins 2001;Suppl 5:22-38.Analysis and assessment of comparative modeling predictions in CASP4.Anna Tramontano, Raphael Leplae, Veronica Morea.

[9] Burkhard Rost.Twilight zone of protein sequence alignments. Protein Engineering,(12):85-94,1999.

[10] Proteins 2007;69(Suppl 8):27-37.Assessment of CASP7 predictions in the high accuracy template-based modeling category.Randy J. Read and Gayatri Chavali.

[11] Science 2001;294:93-96.Protein structure prediction and structural genomics.Baker D, Sali A.
Protein Structure Prediction Center.8 th community wide experiment on the critical assessment of techniques for protein stucture prediction, 2008. http://www.predictioncenter.org/casp8.