Validazione del modello molecolare

Laurea Magistrale in Biotecnologie Mediche

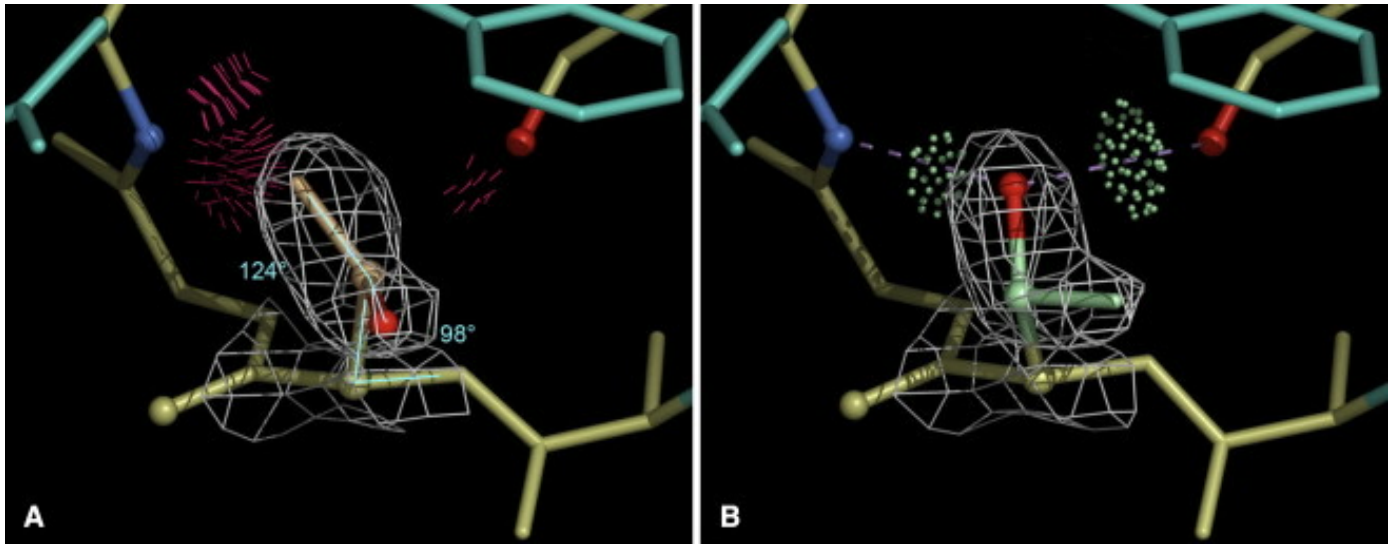
Curriculum Nanobiotecnologie

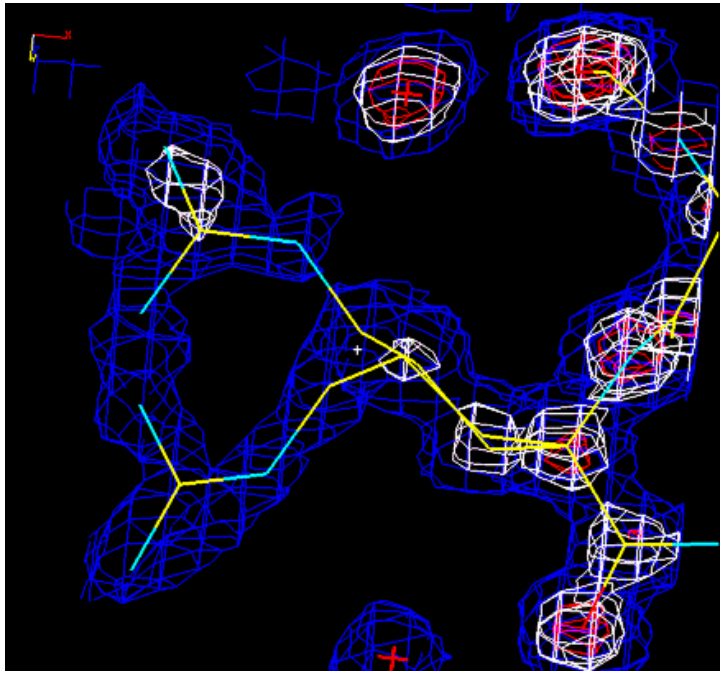
A.A. 2022-23

Introduzione alla Validazione

Limiti del model building

Per quanto abbia una base fisica estremamente solida, il processo di model building a partire dalle intensità diffratte è sempre, almeno in parte, ***soggettivo.***



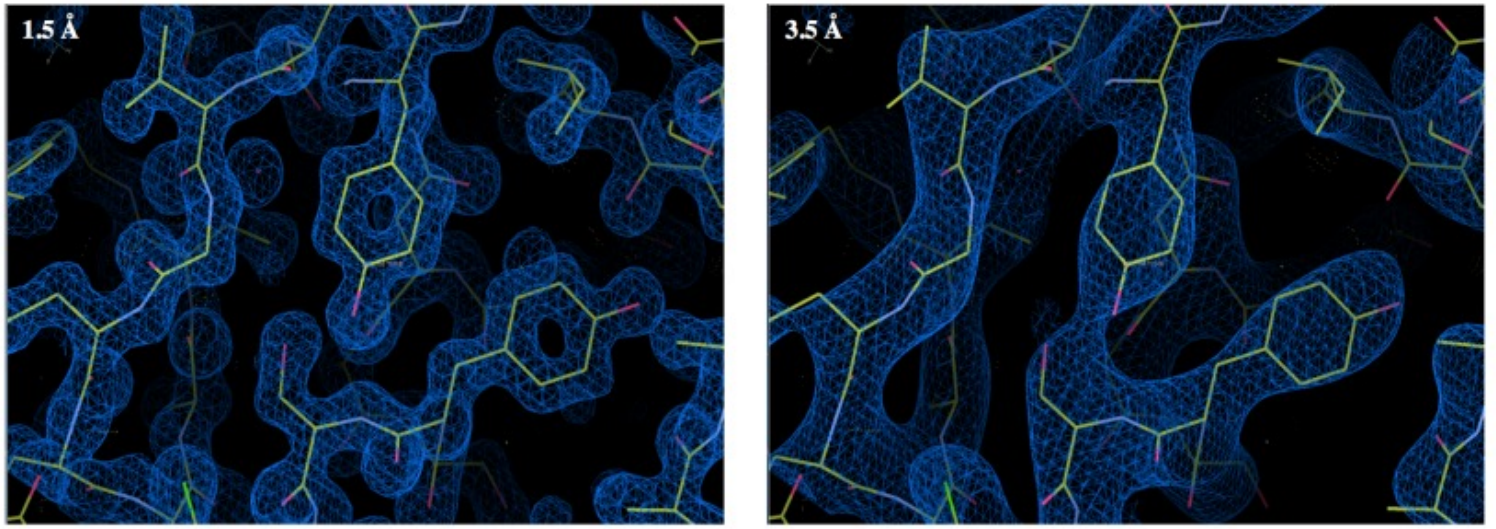


Sulla base della sola mappa di densità elettronica, la scelta tra i due ‘rotameri’ non è così ovvia.

Interpretazione delle mappe

Due fattori influenzano principalmente la ‘qualità’ delle mappe (e quindi la costruzione del modello molecolare).

- Risoluzione
- Qualità dei dati (errori: casuali o sistematici)



When high resolution matters: comparison of a high and a low resolution electron density map

La qualità delle mappe può indirizzare le scelte fatte durante il model building.

Fattori che influenzano la qualità dei dati:

Risoluzione:

- Qualità del cristallo (fondamentale)
- Crioprotezione (molto importante)
- Intensità della sorgente (non è più un problema, anzi!)

Qualità dei dati:

- Radiation damage
- Rumore nei dati sperimentali (radiazione diffusa, elettronica del sistema di rivelazione)
- Errori sistematici nel corso dell'esperimento (tempo di esposizione effettivo (shutter), criostato, monocromatore…)

Inoltre la qualità delle fasi iniziali influenza notevolmente le modalità di costruzione del modello

Oggettività e Soggettività

La combinazione di errori statistici e risoluzione non ottimale può condurre a scelte errate nella costruzione del modello. Questi errori non sempre sono così ovvi e possono condurre interpretazioni funzionali errate.
Verso la fine degli anni 80 questa problematica comincia a diventare chiara:

COMMENTARY

Between objectivity and subjectivity

Carl-Ivar Brändén and T. Alwyn Jones

Protein crystallography is an exacting trade, and the results may contain errors that are difficult to identify. It is the crystallographer's responsibility to make sure that incorrect protein structures do not reach the literature.

NATURE · VOL 343 · 22 FEBRUARY 1990

Freedom and Liberties…

GERARD J KLEYWEGT AND T ALWYN JONES

Where freedom is given, liberties are taken

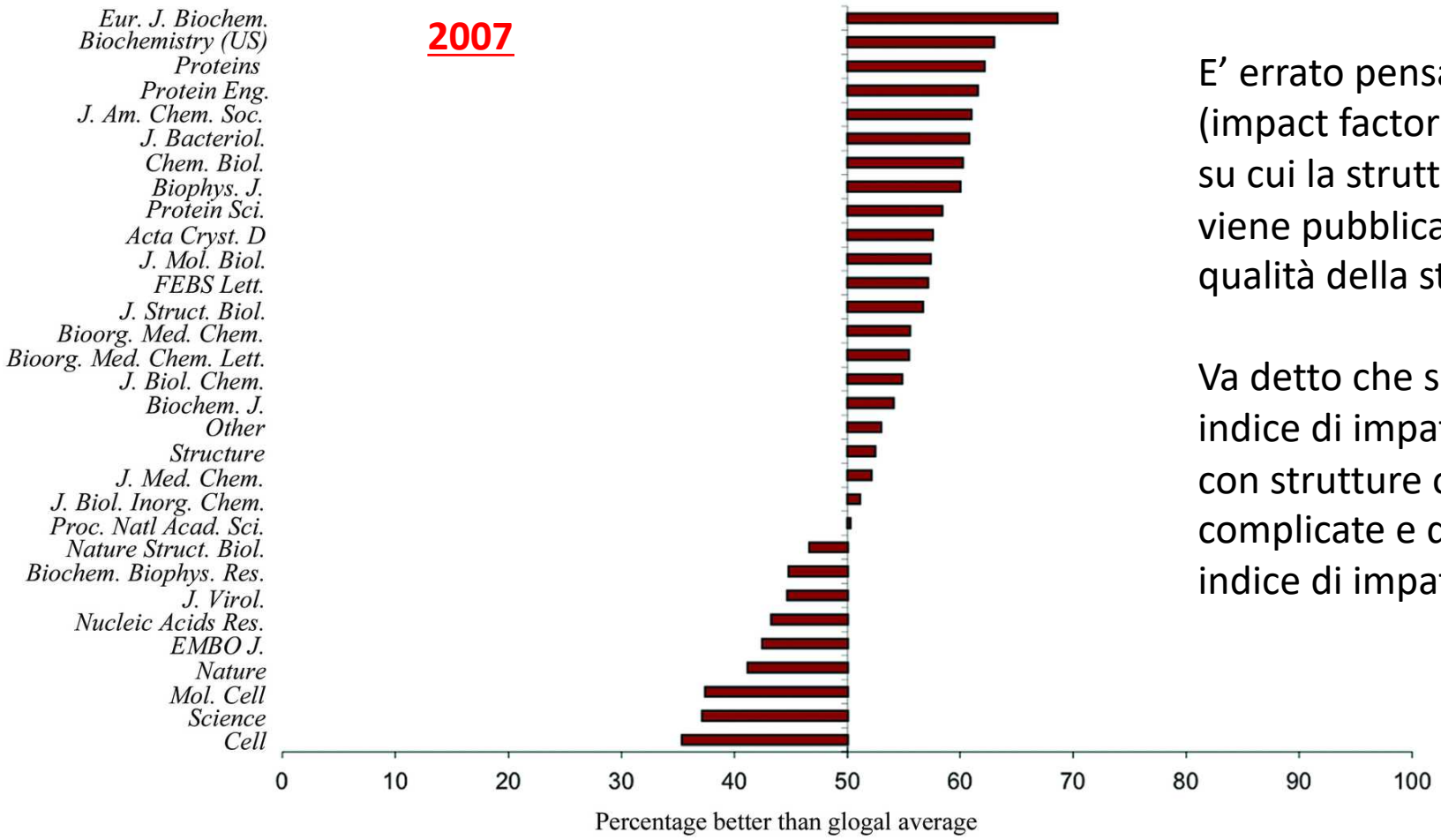
Structure 15 June 1995, **3**:535–540

Here, we argue that, overall, the community is still doing a poor job in its treatment of structures whose crystals diffract poorly. In the worst cases, even if there are no 'errors', biological results are being interpreted with a precision that is not warranted by the information contained in the diffraction data. In the best of cases, a low R-factor (a measurement of how well a model fits the measured diffraction data) is being waved around as proof of the correctness of the structure. In their publication [2], Brändén and Jones warned that structures with R-factors around 25% or higher could indicate problem structures.

La costruzione del modello deve essere guidata dalla qualità dei dati di diffrazione.

Un errore molto comune è quello di ***trarre conclusioni sulla struttura che non sono supportate dalla qualità dei dati***, sia per la limitata risoluzione che per la scarsa qualità della mappa di densità elettronica.

Qualità dei dati e qualità della pubblicazione



E’ errato pensare che la qualità (impact factor) della pubblicazione su cui la struttura cristallografica viene pubblicata, sia garanzia di qualità della struttura. Anzi…

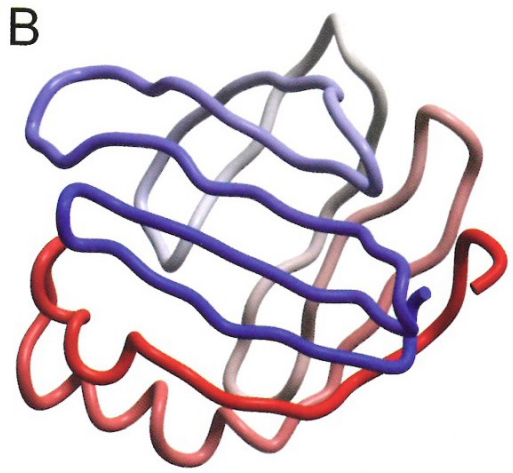
Va detto che spesso riviste di alto indice di impatto pubblicano articoli con strutture oggettivamente complicate e difficili (da cui l’alto indice di impatto!)

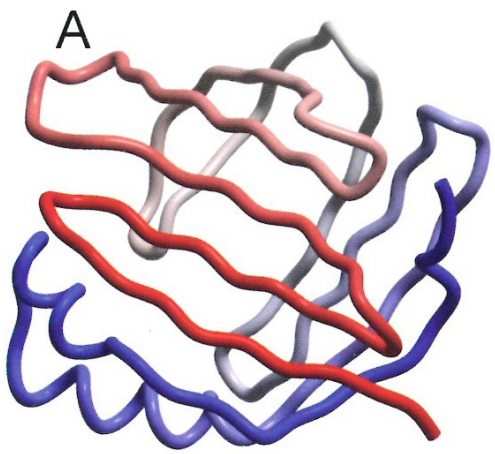
Sbagliare non è così difficile

CRAB

Sbagliata!

Tracciata volontariamente a ritroso(!!!)





CRAB

Corretta!

Come esempio, è stata modellata volontariamente a ritroso nella sua densità elettronica la proteina CRAB, ovvero nella densità dell'N-terminale è stato modellato il C-terminale e così via…
Il fattore R non era poi così male (25%) per una proteina modellata (tracciata) in modo totalmente erroneo…

Necessità di strumenti di validazione

La necessità di essere critici verso le strutture cristallografiche ha portato all'introduzione di nuovi criteri di valutazione della qualità sia dell'affinamento cristallografico che del modello finale.

Per questo motivo sono stati introdotti nuovi indicatori di qualità, che possiamo dividere in:

- Indicatori di qualità del processo di raffinamento (riguardano le intensità/fattori di struttura)
- Indicatori di qualità delle mappe di densità elettronica
- Indicatori di qualità della geometria della struttura (riguradano la 'sensatezza' sterochimica del modello finale)

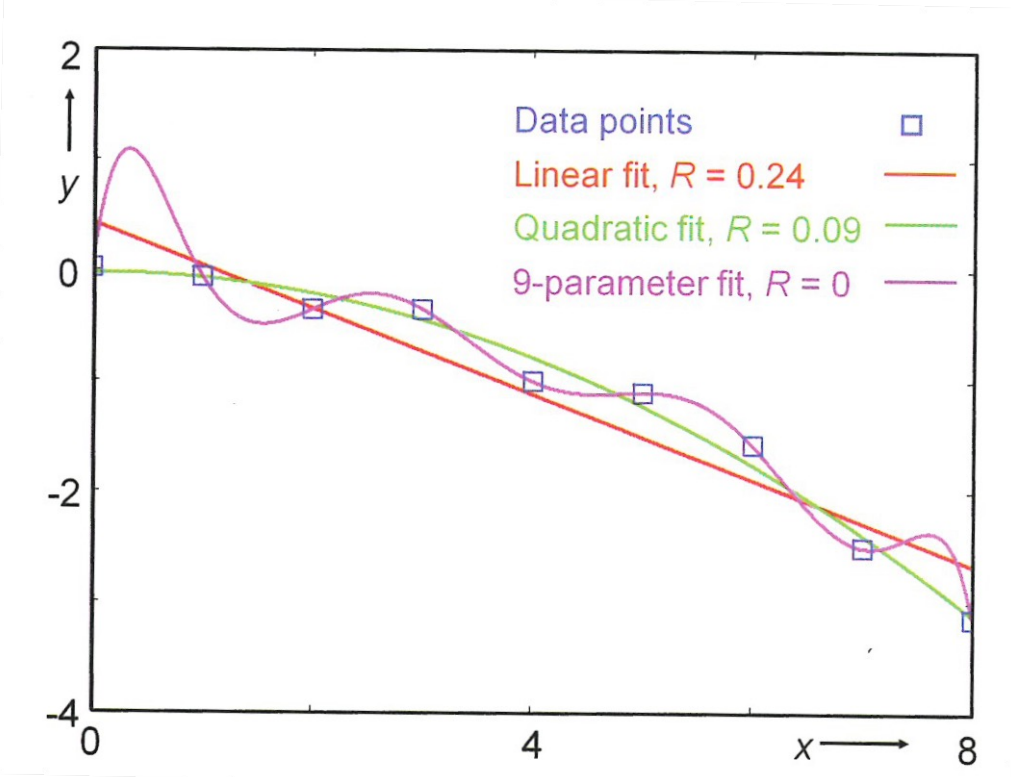
Qualità del processo di raffinamento

Overfitting

In questo esempio i dati sperimentali si riferiscono ad un fenomeno fisico lineare e quindi descritto da un'equazione del tipo: $y=ax + b$.
I dati sono affetti da errore statistico e chiaramente non coincidono con i valori aspettati (ideali).

Posso fare un fit dei dati con equazioni sempre più complesse, la qualità del fit migliora, come testimoniato dalla diminuzione del fattore *R*, ma i modelli di ordine superiore a quello lineare non hanno senso (fisico).

Il miglioramento artificialmente ottenuto con l'introduzione di ulteriori parametri, privi di senso fisico, prende il nome di overfitting.



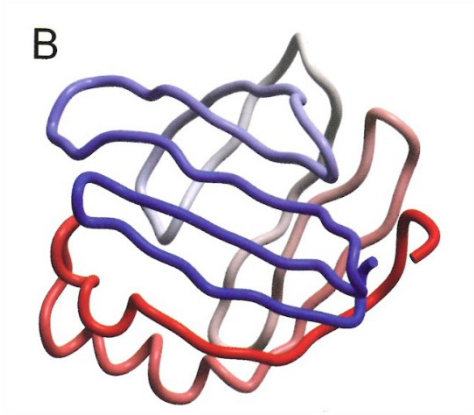
Overfitting e costruzione del modello

Il problema dell'overfitting si pone anche in cristallografia, nel corso del processo di model building e affinamento.

In generale l'osservazione di mappe 2Fo-Fc e Fo-Fc può suggerire la necessità di correggere il modello, eventualmente introducendo nuovi atomi o gruppi di atomi.
La correzione apportata può essere giusta o sbagliata poiché le mappe Fo-Fc e nella sua densità elettronica sono affette da rumore (errore statistico), ma talvolta anche una correzione sbagliata può portare ad una diminuzione (artificiale) del fattore R.

L'abbassamento del fattore R a seguito di una modifica erronea (priva di senso 'fisico') del modello è il classico esempio di overfitting cristallografico.

Un esempio classico ì l'introduzione di 'nuovi' atomi in posizioni in cui non sono davvero presenti.



CRAB
Completamente sbagliata
R = 25%

Cross validation

Al fine di limitare quanto più possibile il problema dell'overfitting, è stato introdotto un nuovo indicatore di qualità: il fattore **R_{free}** basato sul concetto di Cross Validation

L'idea di base della Cross Validation è quella di effettuare il fit su un data-set (***working-set***) e di validare il risultato del fit su un altro data-set (***test-set***) non utilizzato per il fit ma che sarà comunque affetto da un diverso errore casuale (ma con la stessa distribuzione dell'errore).

Se il fit è corretto, gli indicatori di qualità del working-set (R_{work}) e del test-set (R_{free})deve migliorare, se invece il fit è sbagliato, l'indicatore di qualità del test-set peggiorerà (quello del working-set può peggiorare ma anche migliorare).

In cristallografia non abbiamo un secondo data-set di dati, per cui il nostro insieme di fattori di struttura è diviso in due parti:

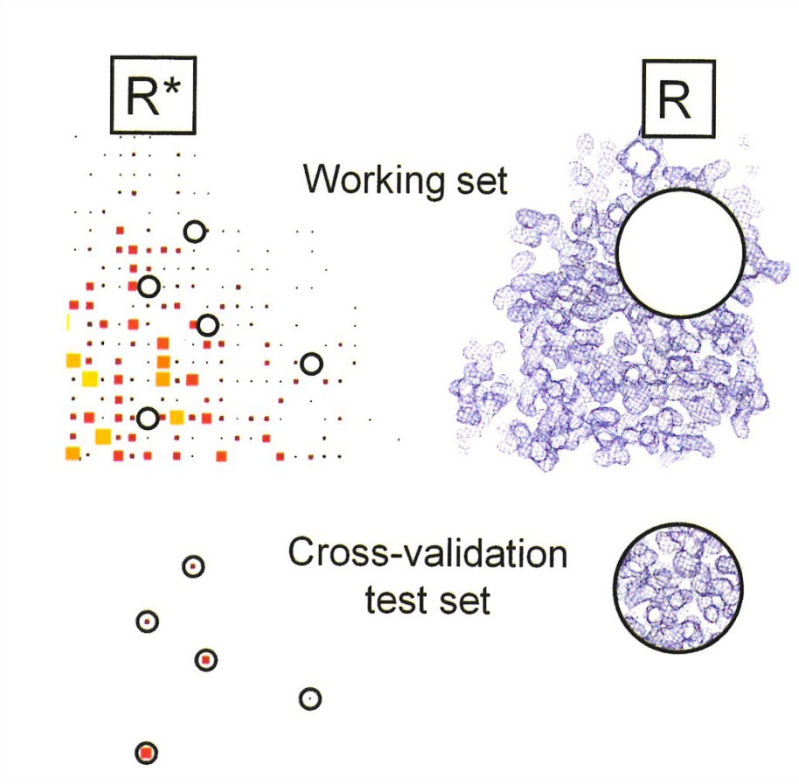
- **Working-set**: in genere il 95% dei dati disponibili e utilizzati nel corso dell'affinamento con minimi-quadrati e quindi usati per calcolare il fattore **R_{work}**.

- **Test-set**: in genere il 5% dei dati; non vengono utilizzati nel corso dell'affinamento ma utilizzati esclusivamente per calcolare il fattore R_{free}.

Working-set e Test-set

I fattori di struttura del test-set sono estratti casualmente ma in modo che siano equamente distribuiti in termini di risoluzione e di indici h, k, l.

Si deve evitare di campionare solo una regione specifica dello spazio reciproco



Fattore R_{free}

Al termine dell'affinamento oltre al normale fattore R, calcolato con i fattori di struttura appartenenti al working-set, viene calcolato anche il fattore R_{free}, calcolato invece con i fattori di struttura appartenenti al test-set.

$$R_{work} = \frac{\sum_{(hkl)\in work-set} ||F_o| - |F_c||}{\sum_{(hkl)\in work-set} |F_o|}$$

Se le modifiche apportate al modello sono state coerenti con il modello 'vero' (quello che produce i dati sperimentali Fo), anche l'R_{free} deve migliorare (deve diminuire), non solo R_{work}.

$$\boldsymbol{R_{free} = \frac{\sum_{(hkl)\in test-set} ||F_o| - |F_c||}{\sum_{(hkl)\in test-set} |F_o|}}$$

L'**R_{free}** è quindi un indicatore della validità delle ipotesi fatte nel corso dell'affinamento strutturale, ovvero se le modifiche introdotte nel nostro modello strutturale sono dovute alla corretta interpretazione del dato sperimentale o se invece sono dovute a scelte errate (causate dal rumore nelle mappe). **Il calcolo di R_{free} è quindi una guida nel corso dell'ottimizzazione del modello e da maggiore oggettività alle scelte fatte.**

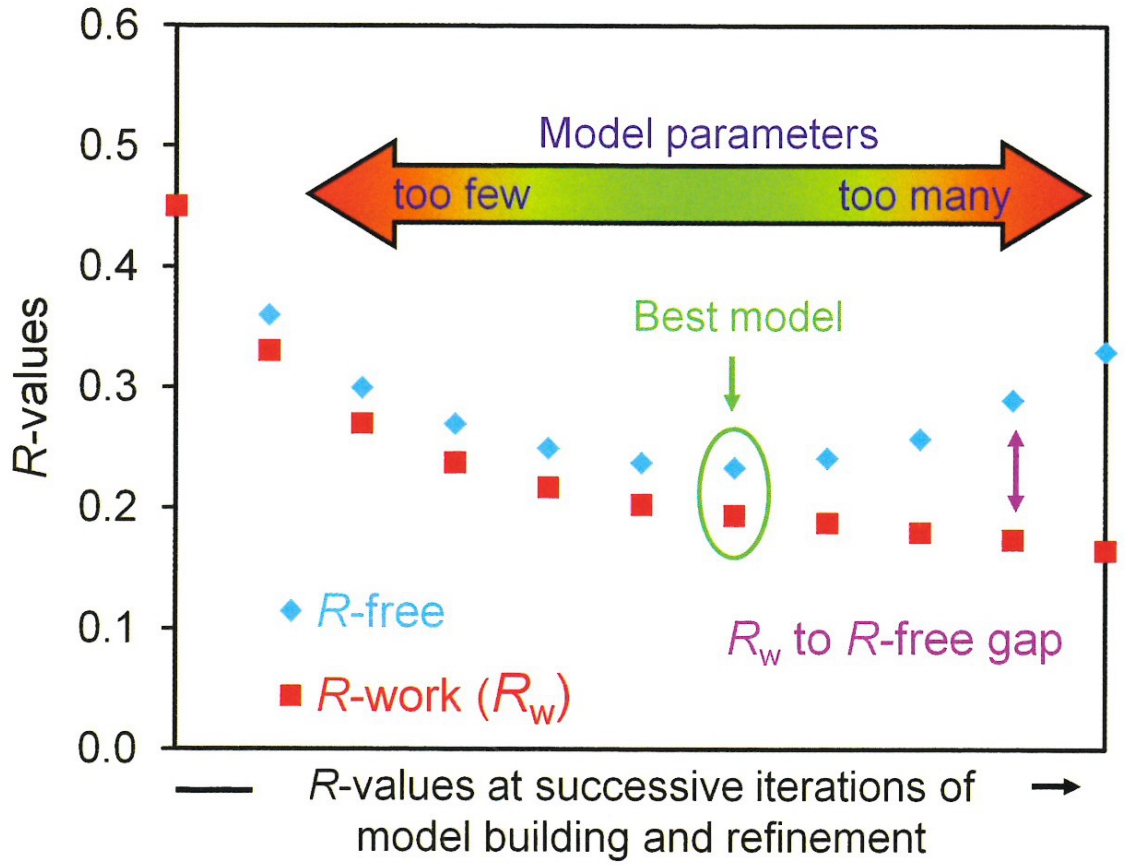
Le scelte errate possono essere dovute alla soggettività nelle scelte o più semplicemente al rumore intrinseco nelle mappe di densità elettronica (che può indurre all'errore).

R_{free} e model building

L'uso dell'R_{free} è un validissimo aiuto nel corso della costruzione del modello e riduce grandemente il pericolo di overfitting.
Il calcolo del fattore R_{free} limita la soggettività nelle scelte ed è correlato con l'accuratezza del modello finale.

Per quanto non elimini completamente la possibilità di errore (soggettività!) la riduce notevolmente.

Il fattore R_{free} è generalmente più grande del fattore R, però è più importante il suo andamento ovvero la sua diminuzione (insieme a al fattore R) a seguito di una modifica del modello e del conseguente affinamento.



Qualità delle mappe di densità elettronica

La qualità della mappa ρ(r)

I fattori R e R_{free} sono indicatori '***globali***' nel senso che ci danno informazioni sulla **qualità globale** del modello, tuttavia non ci danno informazioni **locali**, ovvero non ci dicono se esistono e quali sono le parti 'problematiche' del modello.

Per la valutazione della qualità di parti specifiche del modello si deve far ricorso a indicatori di qualità che forniscono una informazione di tipo locale.

Questi indicatori sono in generale basati sulla correlazione tra la mappa di densità calcolata (|Fc| e fasi dalle coordinate del modello) e densità osservata (|Fo| e fasi dalle coordinate del modello).

Densità elettronica osservata

$$\rho_{obs}(r) = \frac{1}{V} \sum_{h,k,l=-\infty}^{\infty} |F_o| \exp[-2\pi i\,(hx + ky + lz) - i\varphi_{calc}]$$

Densità elettronica calcolata

$$\rho_{calc}(r) = \frac{1}{V} \sum_{h,k,l=-\infty}^{\infty} |F_c| \exp[-2\pi i\,(hx + ky + lz) - i\varphi_{calc}]$$

Indicatori di qualità della densità locale

***Real Space Correlation Coefficent*: RSCC**

$$\boldsymbol{RSCC(x)} = \frac{\sum[(\rho_{obs}(x) - <\rho_{obs}(x)>)(\rho_{calc}(x) - <\rho_{calc}(x)>)]}{[\sum(\rho_{obs}(x) - <\rho_{obs}(x)>)][\sum(\rho_{calc}(x) - <\rho_{calc}(x)>]}$$

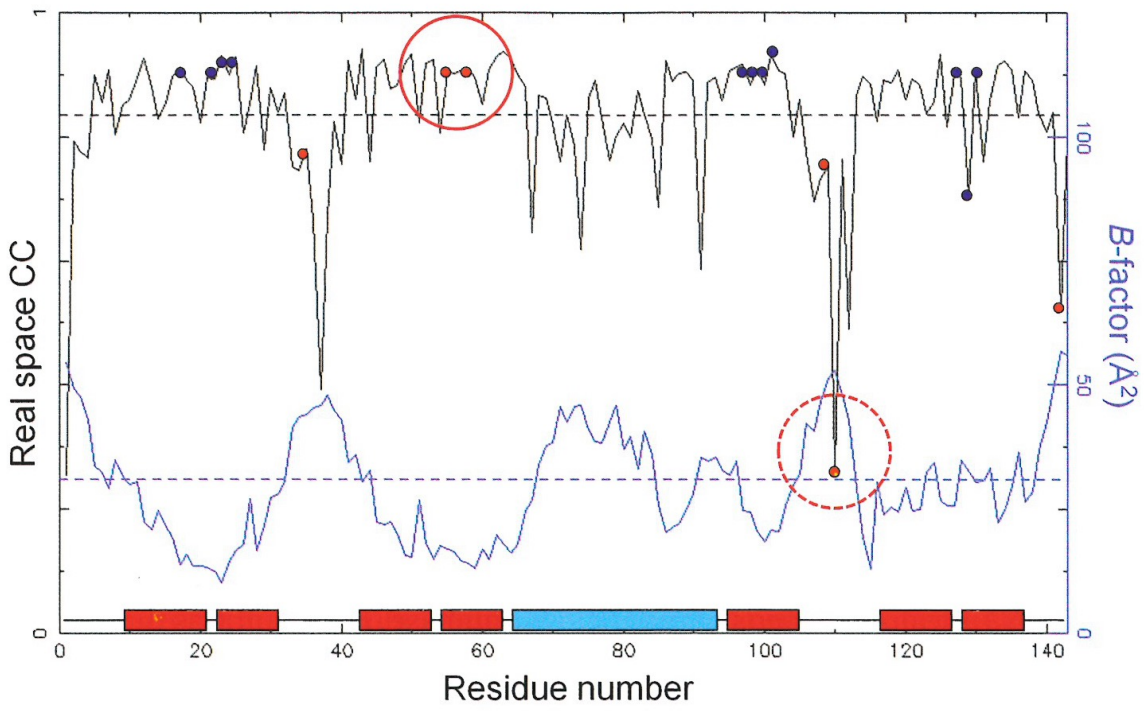
L'$RSCC(x)$ non è altro che un coefficiente di correlazione tra le densità elettronica calcolata e quella osservata in un certo punto dello spazio.

Correlazione tra mappe calcolate e osservate

Gli indicatori di qualità che comparano le mappe di densità elettronica calcolata e osservata sono degli utili strumenti per valutare localmente (residuo per residuo) e velocemente se esistono parti del modello in cui la densità elettronica è poco definita.

Una basso valore di RSCC (è buono quando è sopra 0.9) può essere indicativo di errori nel modello.

Una bassa RSCC può anche essere dovuto ad un elevato valore dei fattori termici, indice che quella parte della proteina è molto flessibile e con un elevato moto termico e quindi con una densità 'debole'.



Validazione della geometria

Validazione Geometrica

Gli indicatori di qualità più utilizzati per la valutazione del modello cristallografico , sia durante l'affinamento che al termine di esso sono di natura geometrica.

Questi indicatori si basano sul fatto che le macromolecole biologiche e le proteine in particolare obbediscono a stringenti regole di natura stereochimica.

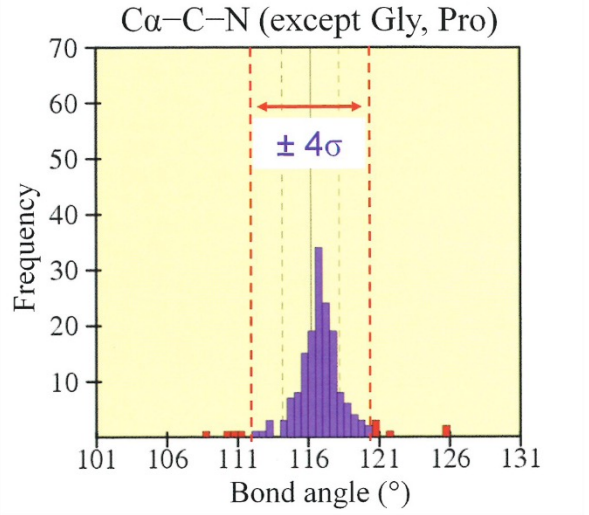
Indicatori di qualità utilizzati in tutte le fasi dell'affinamento e di natura globale sono i valori quadratici medi (rmsd) delle deviazioni dai valori di aspettazione (dati di letteratura) dei parametri stereochimici .

Questi valori sono calcolati per distanze e angoli di legame ma anche per altre grandezze geometriche

$$rmsd_{bond} = \frac{1}{N}\sqrt{\sum (d_{obs} - d_{calc})^2}$$

$$rmsd_{angle} = \frac{1}{N}\sqrt{\sum (\vartheta_{obs} - \vartheta_{calc})^2}$$

Laddove un parametro geometrico si discosta eccessivamente dal valore aspettato (*outliers*), è probabile che ci sia un problema con il modello molecolare costruito.

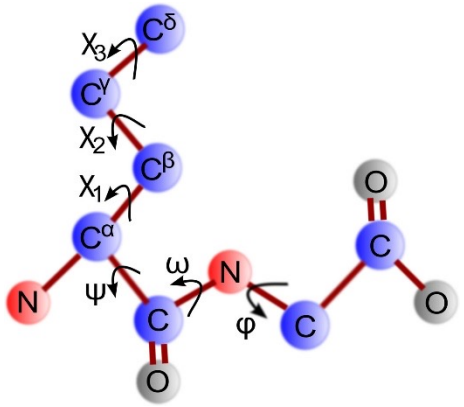


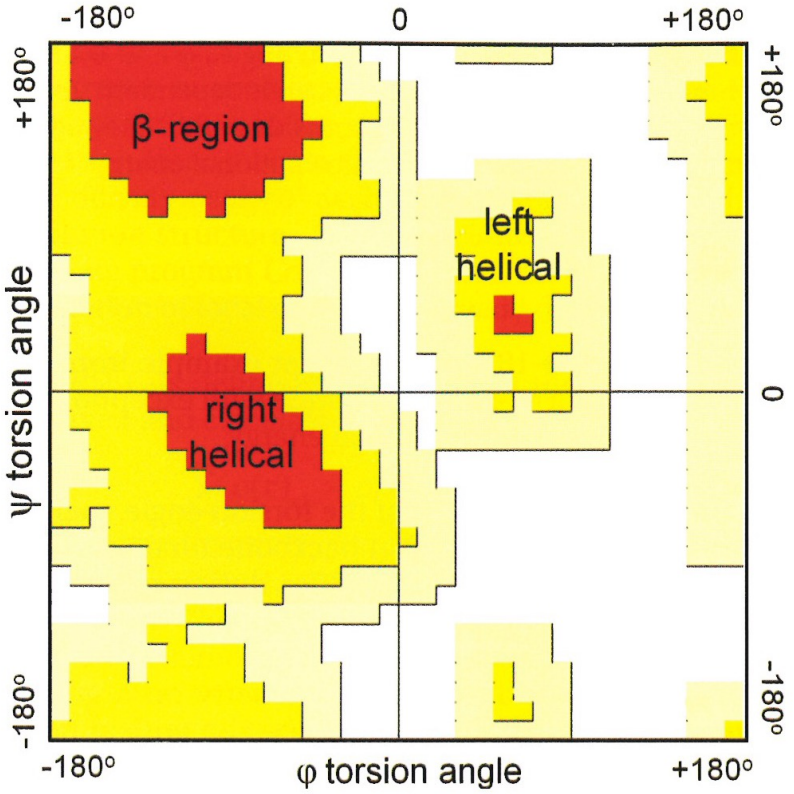
Plot di Ramachandran

E’ noto che **in una catena polipeptidica non tutte le conformazioni sono energeticamente favorevoli.**

Solo particolari combinazioni degli angoli φ e ψ sono energeticamente favorevoli, alcune combinazioni sono decisamente sfavorite e quindi poco probabili (**Plot di Ramachandran**).

In un modello costruito correttamente il numero di residui in posizioni sfavorite dovrebbe essere assente o molto limitato.





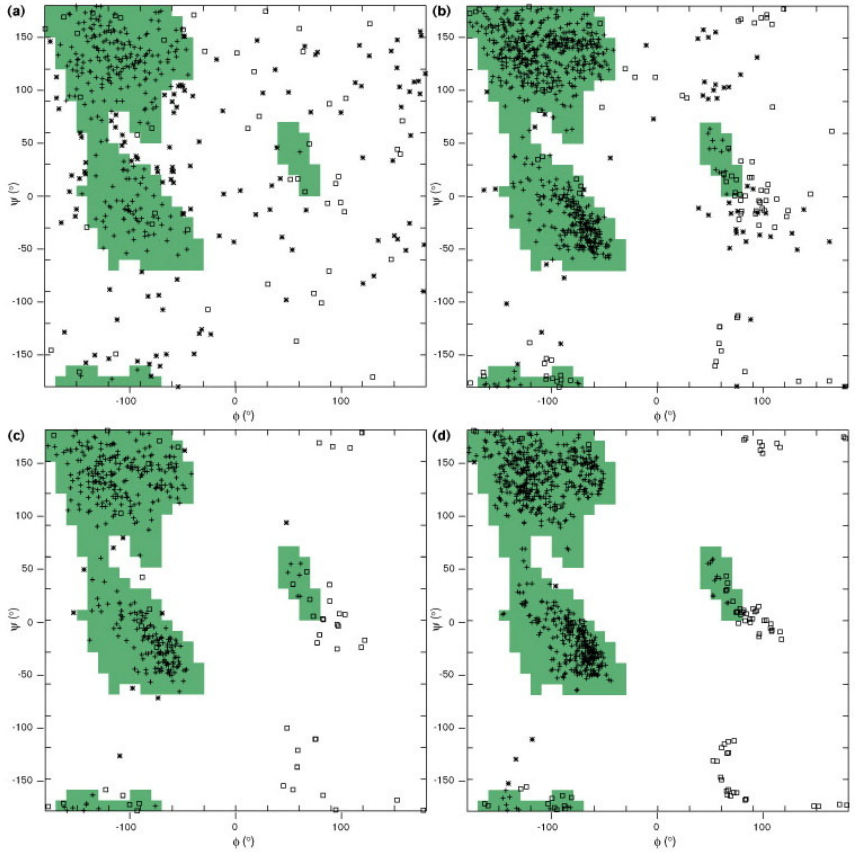
Plot di Ramachandran

L'analisi del plot di Ramachandran è uno degli strumenti di validazione più potenti.

Dal numero di ***outliers*** è facile capire se il modello è stato costruito con attenzione o se invece contiene errori, talvolta grossolani.

In modelli costruiti correttamente la larga maggioranza dei residui è in zone ampiamente favorite.

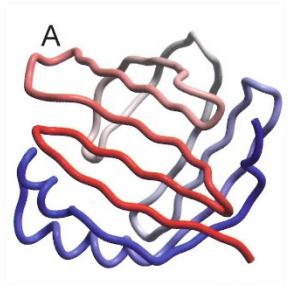
Possono esistere residui in zone sfavorite, ma questi devono essere esaminati con attenzione.
Talvolta, valori non favorevoli di φ e ψ sono associati ad una funzione biochimica precisa.

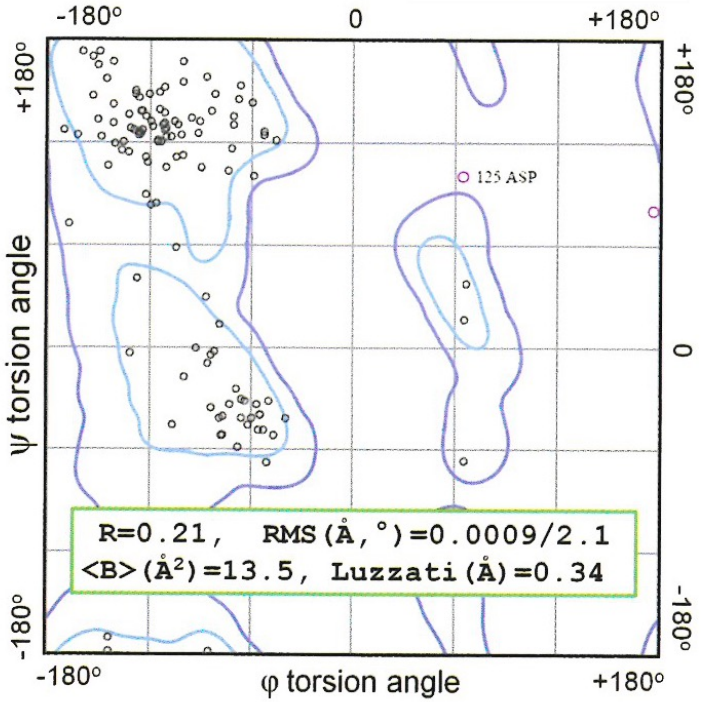


Ogni punto indica la combinazione di φ e ψ per un dato residuo

Ramachandran Plot - esempio

Struttura Vera

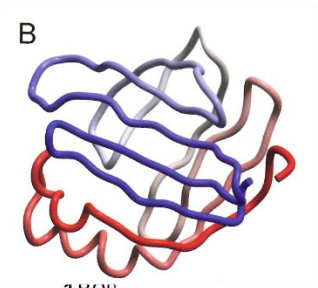




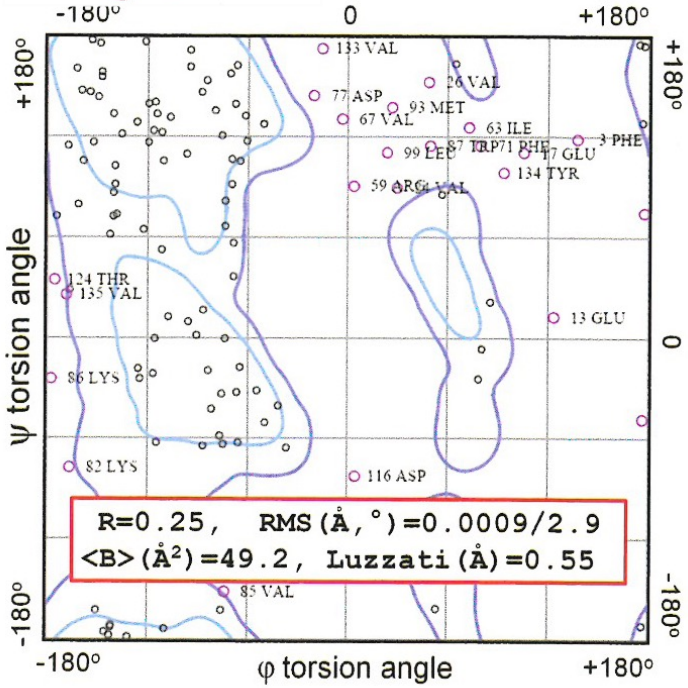
CRAB

Il valore del fattore R non è molto diverso tra le due strutture.
Anche le deviazioni RMS su distanze e angoli di legame non sono molto diverse.

I plot di Ramachandran sono però molto diversi!



Struttura tracciata a ritroso

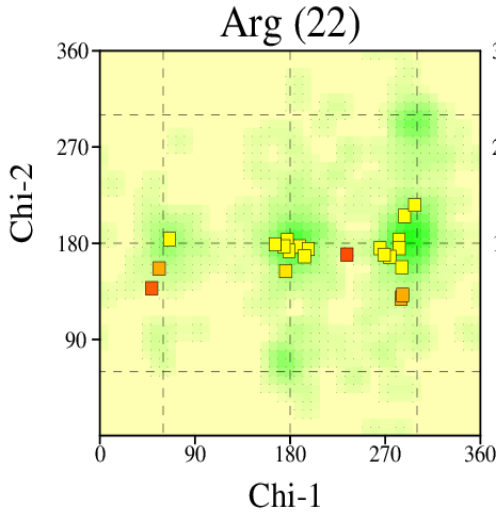


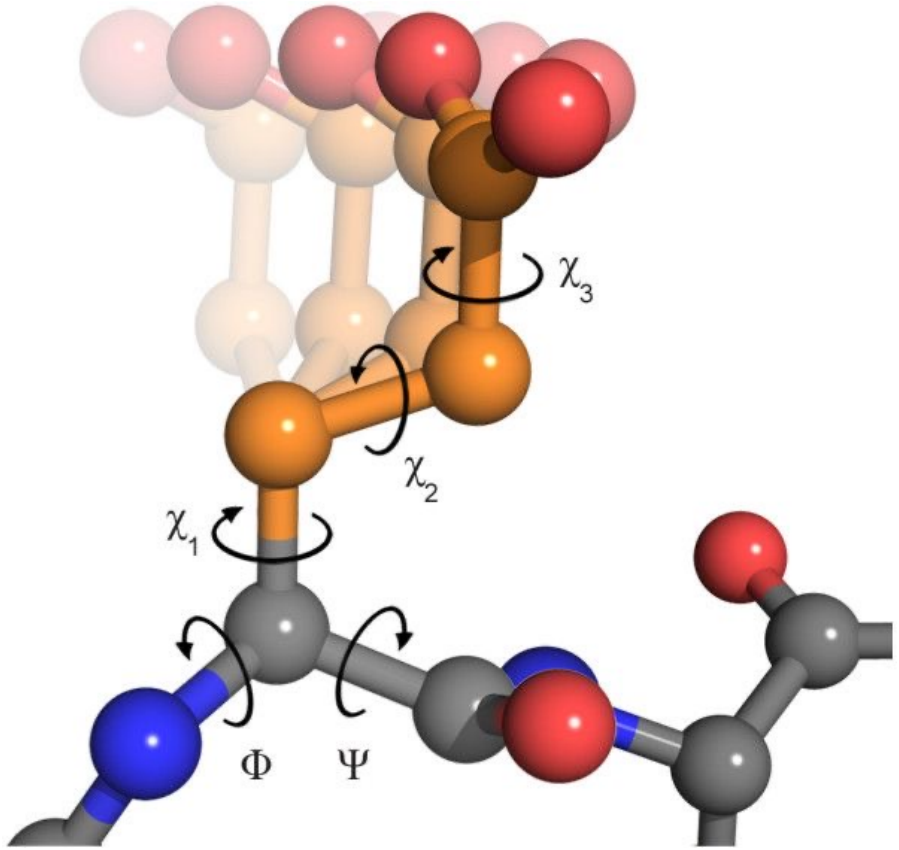
Conformazione delle catene laterali

Come per gli angoli φ e ψ nella catena principale, così anche **nelle catene laterali esistono conformazioni favorite e sfavorite. Le conformazioni favorite sono frequenti e prendono il nome di *rotameri*.**

In generale **ci si aspetta che la catena laterale assuma una conformazione simile a quella di uno dei rotameri consentiti per quel residuo.**

Eventuali residui con conformazione sfavorita devono essere controllati e soggetti ad esame critico.



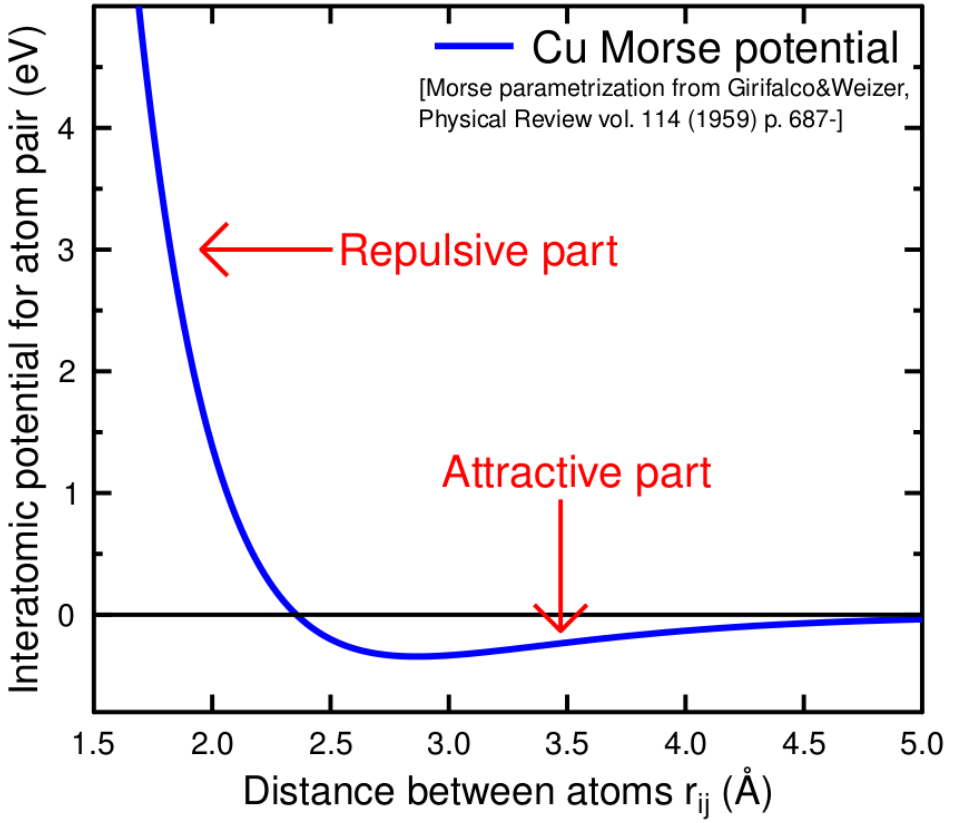


Collisioni (clashes)

In una molecola complessa come una proteina, gli atomi si trovano a determinate distanze dai loro primi vicini.

Può accadere che le distanze interatomiche tra atomi non legati, siano inferiori a quanto consentito dalla tipologia di atomi (*raggi di Van der Waals*) e dal contesto molecolare (*attrazione o repulsione*).

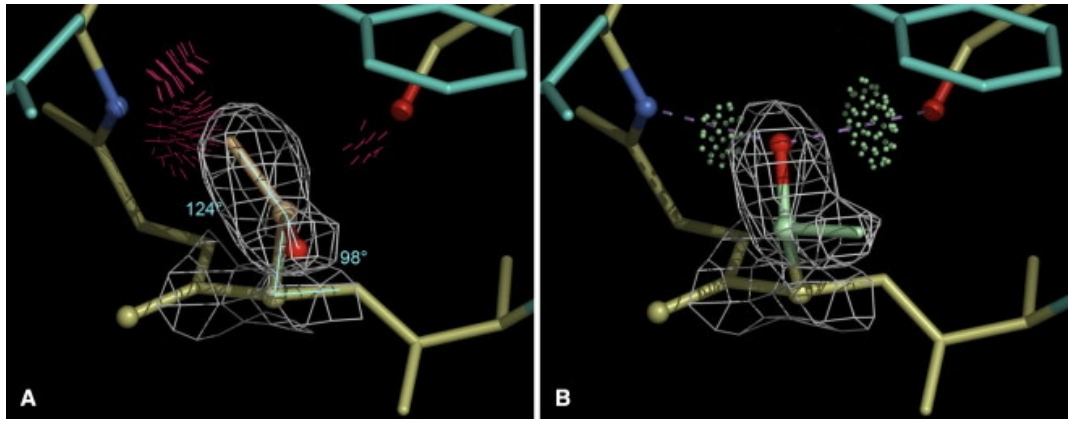
In questi casi parliamo di collisioni (clashes) tra atomi.



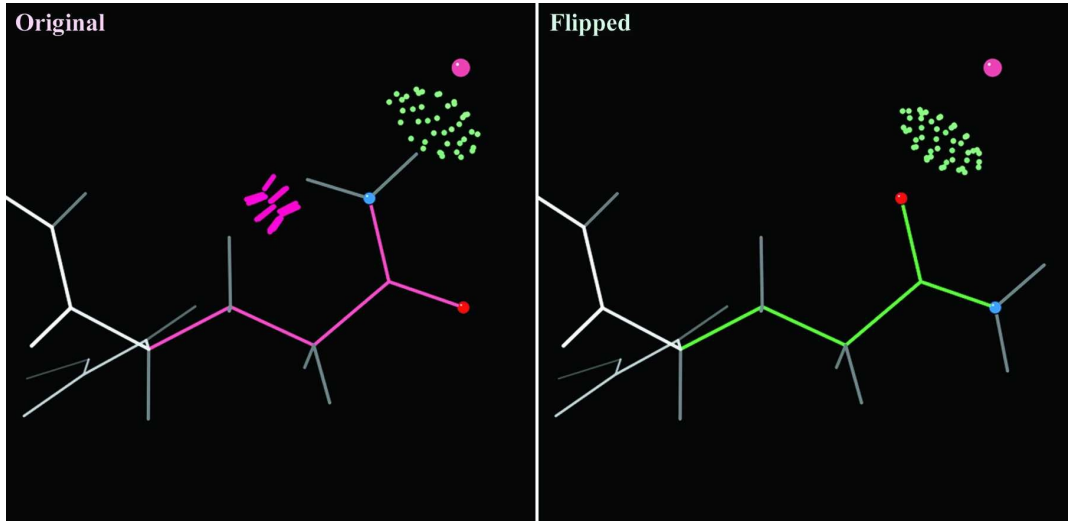
[Kai Nordlund 26.12.2014]

Collisioni in pratica

Le collisioni sono in genere il sintomo di errori nel modello in quanto descrivono situazioni non fisiche e **dovrebbero essere ridotte quanto più possibile** con un'opportuna modifica al modello.



Conformazione errata della catena laterale di una treonina



[*i pallini verdi indicano interazioni favorevoli, quelli rossi indicano interazioni sfavorevoli*]

Conformazione errata di una catena laterale di un'asparagina

Collisioni

Le collisioni, in quanto indice di un modello molecolare errato, si accompagnano in genere ad altri problemi del modello medesimo.

CRAB, modello corretto!

```
REMARK  40  MOLPROBITY OUTPUT SCORES:
REMARK  40  ALL-ATOM CLASHSCORE      :   19.08   95th percentile*
REMARK  40  BAD ROTAMERS             :    4.2%    5/118     (TARGET  0-1%)
REMARK  40  RAMACHANDRAN OUTLIERS    :    1.5%    2/134     (TARGET  0.2%)
REMARK  40  RAMACHANDRAN FAVORED     :   91.0%  122/134     (TARGET 98.0%)
```

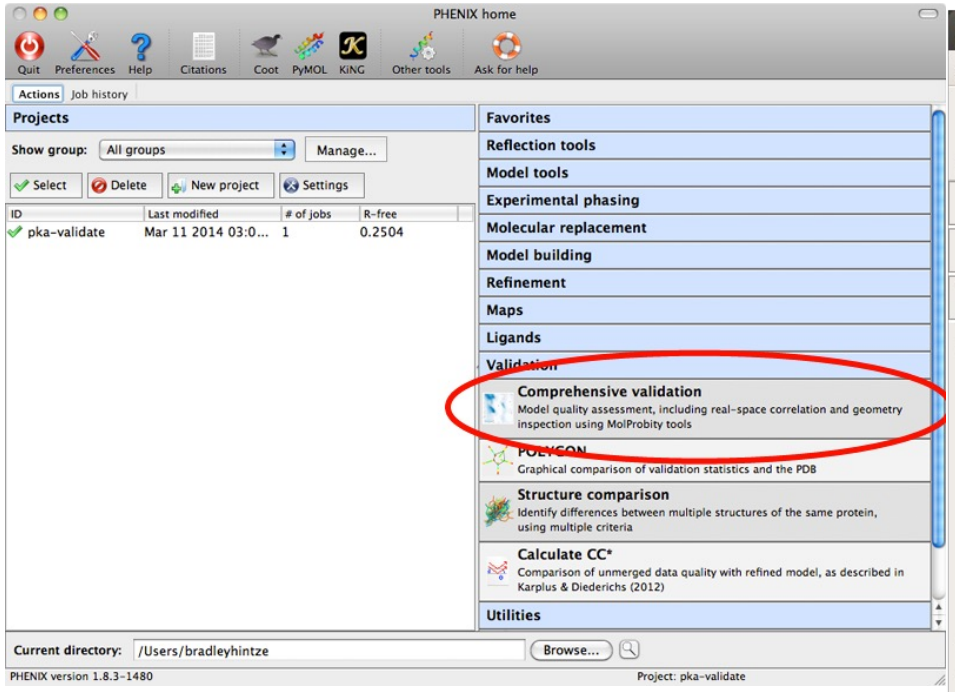
```
REMARK  40  MOLPROBITY OUTPUT SCORES:
REMARK  40  ALL-ATOM CLASHSCORE      :   73.64   0th percentile*
REMARK  40  BAD ROTAMERS             :   29.3%   36/123     (TARGET  0-1%)
REMARK  40  RAMACHANDRAN OUTLIERS    :   20.7%   28/135     (TARGET  0.2%)
REMARK  40  RAMACHANDRAN FAVORED     :   52.6%   71/135     (TARGET 98.0%)
```

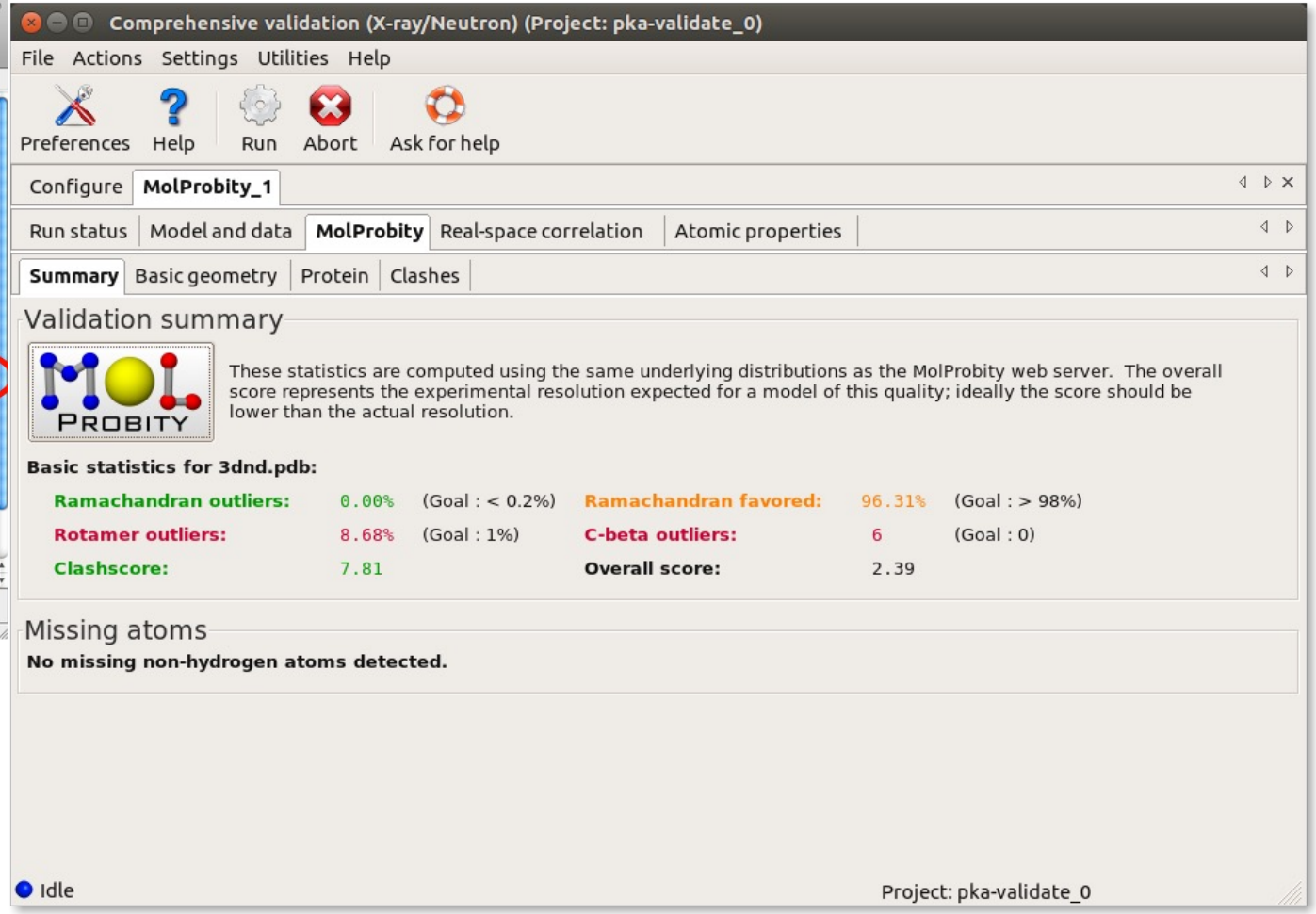
Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.
*** 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.**

CRAB modello errato! (Tracciato a ritroso)

Software per la validazione geometrica

Validazione





La validazione è riconosciuta come strumento essenziale nel corso dell'ottimizzazione del modello

Molprobity

PDB REDO

Il servere di PDB_REDO (https://pdb-redo.eu/) ottimizza il protocollo di raffinamento su base automatica.
PDB_REDO sceglie il modello più opportuno e applica le procedure di raffinamento, sulla base della struttura in input, della risoluzione dei dati di diffrazione e della qualità dei dati.

Significant model changes	
Description	**Count**
Rotamers changed	2
Side chains flipped	0
Waters removed	35
Peptides flipped	0
Chiralities fixed	0
Residues fitting density better	231
Residues fitting density worse	1

Validation metrics from PDB-REDO		
	PDB	**PDB-REDO**
Crystallographic refinement		
R	0,2016	0,1355
R-free	0,2121	0,1612
Bond length RMS Z-score	0,597	0,559
Bond angle RMS Z-score	0,780	0,795
Model quality (raw scores | percentiles)		
Ramachandran plot appearance	60	66
Rotamer normality	77	88
Coarse packing	N/A	N/A
Fine packing	32	34
Bump severity	98	93
Hydrogen bond satisfaction	77	57

In genere PDB REDO migliora un po' i valori degli indicatori di qualità anche se raramente modifica in modo sostanziale le cose.

Deviazioni Standard (?)

Le coordinate presenti nel PDB non hanno deviazioni standard!

Si possono calcolare per strutture ad alta risoluzione.

La procedura dei minimi quadrati con restraints rende complicata la valutazione delle deviazioni standard.

Sono state proposte varie formule per esprimere le incertezze medie (di una struttura cristallografica) sulle coordinate, che utilizzano il fattore-R, ma anche i fattori termici.

Il cosiddetto Luzzati plot, mette in relazione l'andamento del fattore-R con la risoluzione, da questo andamento si può capire qual è l'incertezza media sulle posizioni atomiche, del modello cristallografico.