BINF 731

Protein Structure Analysis

Iosif Vaisman

2023

Structure verification and validation

Anolea
Verify3D
Procheck
WhatIf

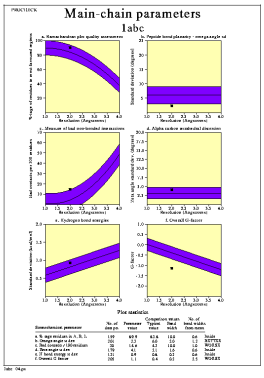
Bond lengths (Procheck)

Bond	labeling		Value	sigma
C-N	C-NH1	(except Pro)	1.329	0.014
	C-N	(Pro)	1.341	0.016
C-O	C-O		1.231	0.020
Calpha-C	CH1E-C	(except Gly)	1.525	0.021
	CH2G*-C	(Gly)	1.516	0.018
Calpha-Cbeta	CH1E-CH3E	(Ala)	1.521	0.033
	CH1E-CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E-CH2E	(the rest)	1.530	0.020
N-Calpha	NH1-CH1E	(except Gly,Pro)	1.458	0.019
	NH1-CH2G*	(Gly)	1.451	0.016
	N-CH1E	(Pro)	1.466	0.015

Bond angles (Procheck)

Angle	labeling		Value	sigma
C-N-Calpha	C-NH1-CH1E	(except Gly,Pro)	121.7	1.8
	C-NH1-CH2G*	(Gly)	120.6	1.7
	C-N-CH1E	(Pro)	122.6	5.0
Calpha-C-N	CH1E-C-NH1	(except Gly,Pro)	116.2	2.0
	CH2G*-C-NH1	(Gly)	116.4	2.1
	CH1E-C-N	(Pro)	116.9	1.5
Calpha-C-O	CH1E-C-O	(except Gly)	120.8	1.7
	CH2G*-C-O	(Gly)	120.8	2.1

Procheck output



a. Ramachandran plot quality - percentage of the protein's residues that are in the **core** regions of the Ramachandran plot.

b. Peptide bond planarity - standard deviation of the protein structure's **omega** torsion angles.

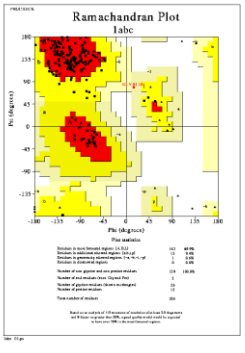
c. Bad non-bonded interactions - number of bad contacts per **100** residues.

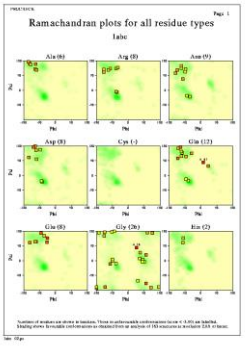
d. Cα tetrahedral distortion - standard deviation of the **ζ torsion angle** (**Cα**, **N**, **C**, and **Cβ**).

e. Main-chain hydrogen bond energy - standard deviation of the **hydrogen bond energies** for **main-chain hydrogen bonds**.

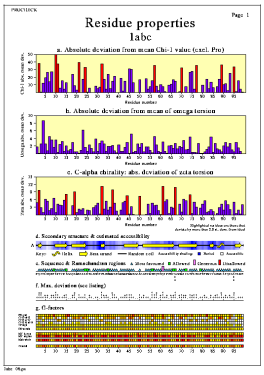
f. Overall G-factor - average of different G-factors for each residue in the structure.

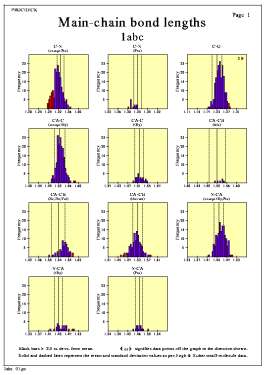
Procheck output



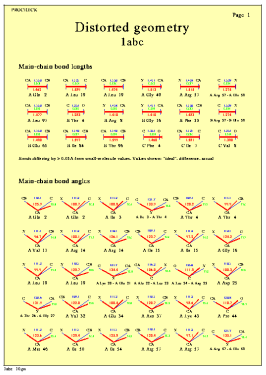


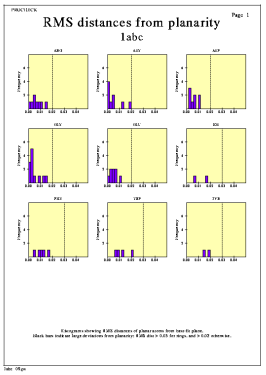
Procheck output



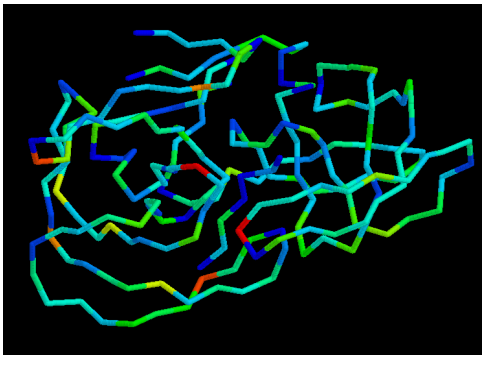


Procheck output

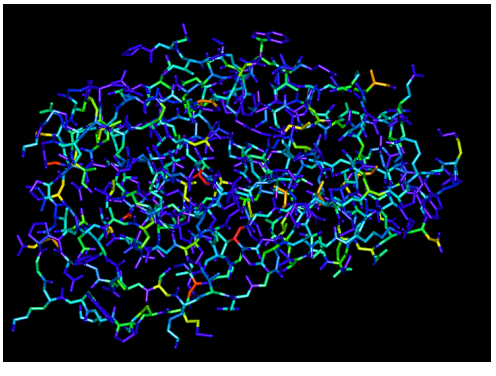




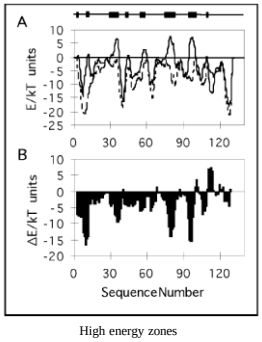
Procheck output - backbone G factors



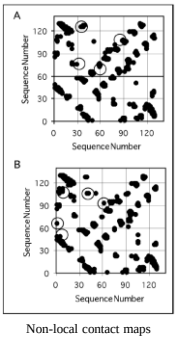
Procheck output - all atom G factors



Anolea



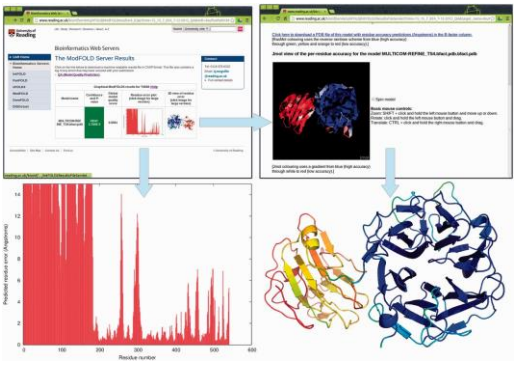
High energy zones



Non-local contact maps

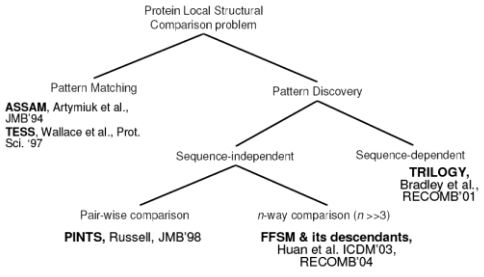
Melo et al., 1997

Model Quality Assessment



L.J.McGuffin et al., 2013

Local structure comparison algorithms



Huan, 2006

Root mean square deviation

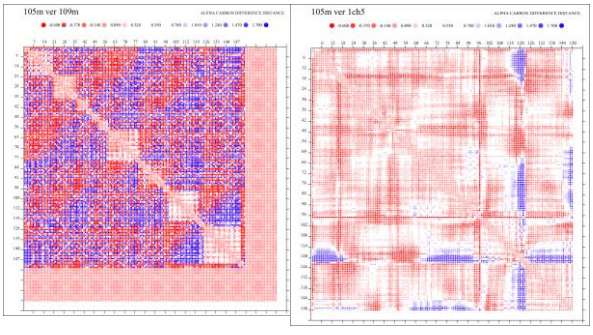
Coordinate based RMSD

$$RMSd(A,B) = \min_T \sqrt{\sum_{i=1}^{m} (A_i - TB_i)^2}$$

Distance based RMSD

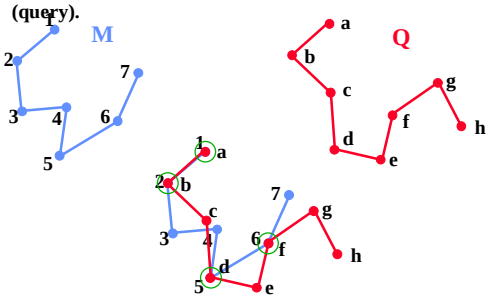
$$RMSd_D(A,B) = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \left(d_{ij}^A - d_{ij}^B\right)^2}$$

Difference Distance Matrix Plot (DDMP)



Geometric Hashing

Find common subsets, invariant under rotation and translation in two point sets M (model) and Q (query).



Finding the maximum coincidence set is an NP-hard problem

Geometric Hashing
Reference frames

- **Two points (basis pair) define a reference frame**
- **The coordinates of all points are computed in the reference frame (reference frame system)**
- **There will be pairs of points (from M and Q) with the same coordinates**
- **The number of such pairs depends on selection of reference frame and reference frame system resolution**

Geometric Hashing Algorithm

Preprocessing

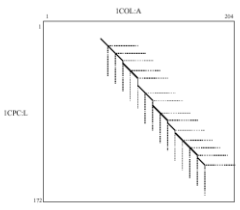
Hash table H is created. It has a bin for each cell in the frame systems. The coordinates of all points in each model frame system are calculated. If there is a point in the cell (p,q) in the frame system with basis (a_i,a_k), then (a_i,a_k) is placed in the bin H(p,q)

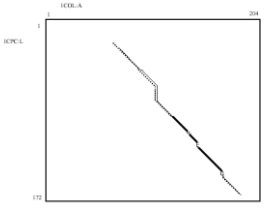
Recognition

A pair of points in the query is chosen as basis, and the coordinates of the other points are calculated. These coordinates are used as indices for H, and for each cell being indexed, a vote is given for the (model) basis pairs in the cell. The number of votes for a model basis pair is the number of coinciding points to the query (using the specified query basis pair)

Combinatorial Extension (CE) Algorithm

Structure alignment of phycocyanin (1CPC:L) to colicin A (1COL:A)

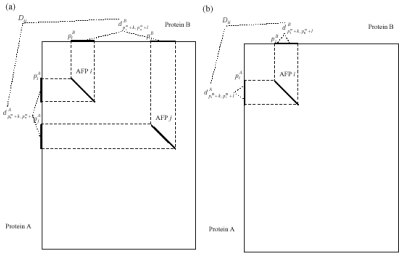




The solid line represents the optimal path built from AFPs. The dotted line represents the search area at every step of path extension.

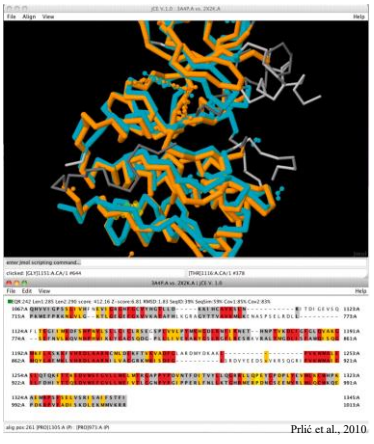
The thick solid line represents alignment overlap both before and after optimization.

Combinatorial Extension (CE) Algorithm



Calculation of distance

***Dij* for two AFPs *i* and *j* from the path** **_Dii_ for single AFP *i* from the path.**

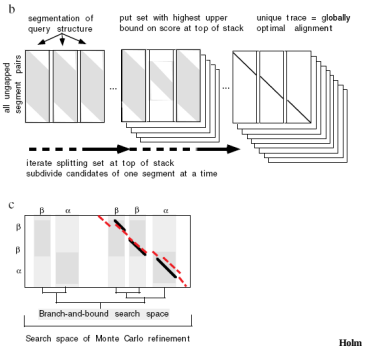


jCE and jFATCAT tool at RCSB

Pairwise structure alignment and pre-calculated 3D structure comparisons for the entire PDB

Prlić et al., 2010

DaliLight algorithm for protein structure comparison



Holm and Park, 2000

Search for common substructures by clique detection