Université Paris Cité

M2BI PROJECT : UE Programming 3

Topic : Design of a program to assess the quality of a 3D protein model

Student : Arnaud LY
Academic year 2022-2023

- Many ways to model proteins for studies

- Assessment of the quality of models
  - Discrete Optimized Protein Energy (DOPE)
    - Atomic distance-dependent statistical potential

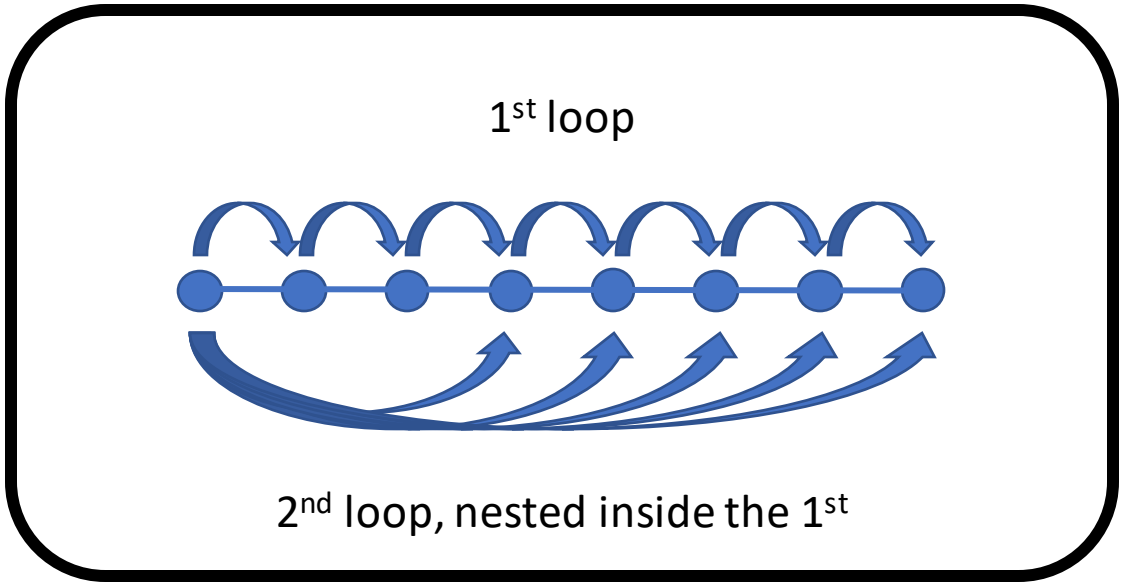
- Many ways to model proteins for studies

- Assessment of the quality of models
  - Discrete Optimized Protein Energy (DOPE)
    - Atomic distance-dependent statistical potential

Write a script to assess the quality of a protein model (Shen *et al*, 2006)

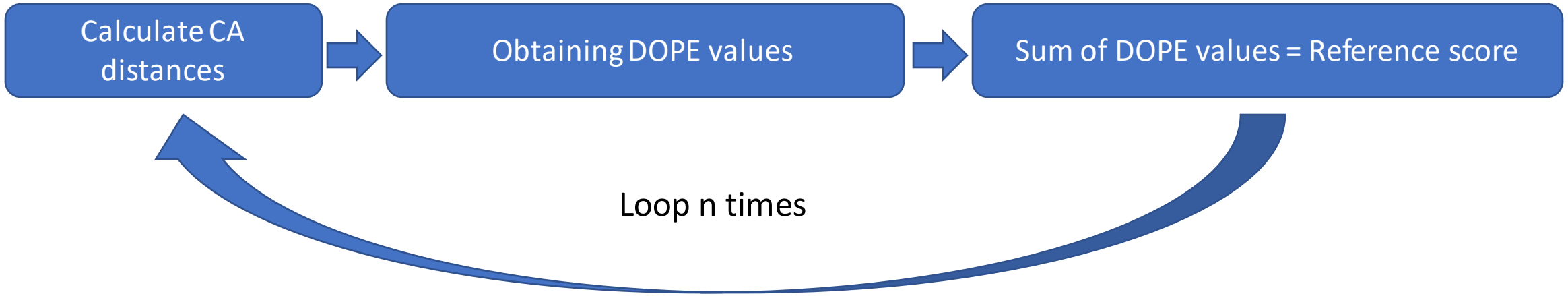
Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 Nov;15(11):2507-24. doi: 10.1110/ps.062416606. PMID: 17075131; PMCID: PMC2242414.

Loading data (from PDB file) → α carbons (CA) coordinates acquisition → Calculate CA distances



Calculate CA distances → Obtaining DOPE values → Sum of DOPE values = Protein DOPE score

	res1	res2	0.5	1.0	1.5	2.0	...	12.5	13.0	13.5	14.0	14.5	15.0
0	ALA	ALA	10.0	10.0	10.0	10.0	...	-0.08	0.01	-0.02	-0.08	-0.12	-0.02
1	ALA	ARG	10.0	10.0	10.0	10.0	...	-0.01	0.02	-0.00	-0.00	-0.10	-0.02
2	ALA	ASN	10.0	10.0	10.0	10.0	...	-0.01	0.04	0.03	-0.02	-0.05	-0.02
3	ALA	ASP	10.0	10.0	10.0	10.0	...	-0.03	-0.00	-0.02	-0.02	-0.11	-0.02
4	ALA	CYS	10.0	10.0	10.0	10.0	...	-0.06	-0.03	-0.07	-0.05	-0.11	-0.02
5	ALA	GLN	10.0	10.0	10.0	10.0	...	-0.04	0.00	0.01	-0.00	-0.11	-0.02
6	ALA	GLU	10.0	10.0	10.0	10.0	...	-0.04	-0.01	-0.03	-0.06	-0.10	-0.02
7	ALA	GLY	10.0	10.0	10.0	10.0	...	0.02	-0.02	-0.01	-0.03	-0.07	-0.02
8	ALA	HIS	10.0	10.0	10.0	10.0	...	0.01	0.03	0.03	-0.01	-0.11	-0.02
9	ALA	ILE	10.0	10.0	10.0	10.0	...	-0.09	-0.06	-0.04	-0.02	-0.08	-0.02



- Comparison of protein DOPE score : Reference DOPE score (same steps but with shuffled sequences)

- From reference score : mean (μ) and standard deviation (σ)
  - Z-score of protein DOPE score using : (μ - x) / σ

```
File name : 1pdc.pdb
Number of CA in this sequence : 45
Sum of DOPE values for original sequence = -44.83
List of ref values from shuffled sequences :
[-111.38999999999969, -112.03999999999955, -104.36999999999992, -109.2299999999997, -105.16999999999953,
-113.68999999999957, -101.95999999999982, -108.71999999999984, -110.94999999999987, -109.94000000000004]
Mean and standard deviation calculated from randomized sequence (10x) = -108.75, 3.75
The score of the original sequence compared to randomized sequences : -17.03
```

```
File name : 1gcn.pdb
Number of CA in this sequence : 29
Sum of DOPE values for original sequence = -11.4
List of ref values from shuffled sequences :
[-60.56000000000033, -56.56000000000023, -62.080000000000176, -60.68000000000027, -57.81000000000026,
-60.8400000000002, -63.36000000000025, -62.090000000000224, -63.950000000000216, -58.990000000000215]
Mean and standard deviation calculated from randomized sequence (10x) = -60.69, 2.36
The score of the original sequence compared to randomized sequences : -20.93
```

Results from file "compil_results.txt"

The higher the absolute value of this z-score is, the more likely the conformation from these coordinates is to exist *in vivo*

- Obtained expected value
- Comparison with protein conformation non-existent *in vivo*
- Improvements : Linear interpolation, not just CA data, try with bigger protein or with more complex conformation

Thank you for your attention !