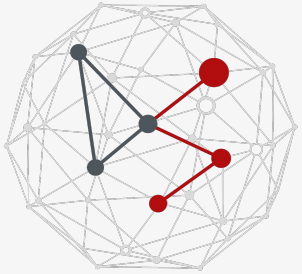
DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

STATISTICAL POTENTIALS

Master of Science in Data Science

Damiano Piovesan

Discriminatory functions

Discriminate native / non-native conformations

Types

- Simple
  - Number of atomic contacts
- Complex
  - Energy functions (molecular mechanics)
- Knowledge-based (statistical potentials)
  - AA preference

Use

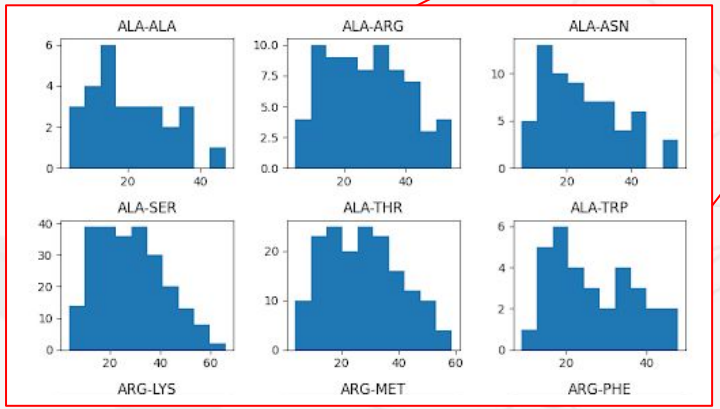
- Validate structure experiments
- Fold recognition
- Ab initio prediction

Native features

- Close packing
- Strength of electrostatic interactions
- Exposure of polar groups to solvent
- Secondary structure
- Distribution of intramolecular distances

C_α-C_α distances for all possible amino acid pairs

(PDB 1CU4, chain L)



RAPDF

DIPARTIMENTO MATEMATICA

Residue-specific **A**ll-atom conditional **P**robability **D**iscriminatory **F**unction

$$P(C|\{d_{ab}^{ij}\})$$

Correct conformation

Set of distances of **atoms *i, j*** of **amino acid type *a, b***

An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction
Ram Samudrala and John Moult. ***JMB. 1998***

Bayes' theorem

DIPARTIMENTO MATEMATICA

$$P(C|d_{ab}^{ij}) = \frac{P(d_{ab}^{ij}|C) \cdot P(C)}{P(d_{ab}^{ij})}$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Prior - Probability of correct structure

Posterior probability

$$P(C) \cdot P(d_{ab}^{ij}|C) = P(d_{ab}^{ij}) \cdot P(C|d_{ab}^{ij})$$

Likelihood - Probability of observing that distance in correct structures

Prior - Probability of observing that distance in any structure

RAPDF

- We want sets of distances not single distances
- We assume all distances are independent of one another
- Then, the joint probability can be calculated multiplying the probability of single distances

$$P(\{d_{ab}^{ij}\}|C) = \prod_{ij} P(d_{ab}^{ij}|C); \qquad P(\{d_{ab}^{ij}\}) = \prod_{ij} P(d_{ab}^{ij})$$

likelihood

prior

RAPDF

DIPARTIMENTO MATEMATICA

$$P(C|d_{ab}^{ij}) = \frac{P(d_{ab}^{ij}|C) \cdot P(C)}{P(d_{ab}^{ij})}$$

Likelihood - Probability of observing that distance in correct structures

Posterior probability

$$P(C|\{d_{ab}^{ij}\}) = P(C) \cdot \prod_{ij} \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})}$$

Prior - Constant, independent of conformation for a given sequence

Prior - Probability of observing that distance in any structure

$$S(\{d_{ab}^{ij}\}) = -\sum_{ij} \ln \frac{P(d_{ab}^{ij}|C)}{P(d_{ab}^{ij})} \propto -\ln P(C|\{d_{ab}^{ij}\})$$

Transform the product into a sum using the logarithm

RAPDF - How likelihood and prior are defined?

DIPARTIMENTO MATEMATICA

$P(d_{ab}^{ij}|C)$

Likelihood - Measure frequencies in (correct) **PDB** structures

$P(d_{ab}^{ij})$

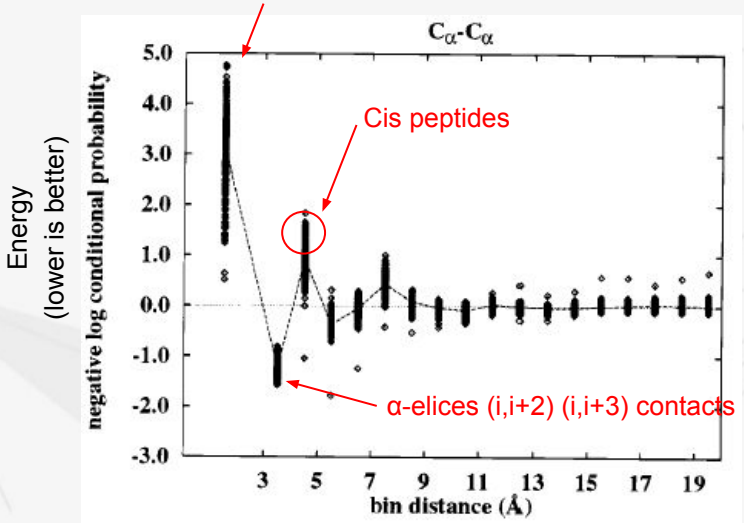
Prior - Correct + Incorrect conformations, options:

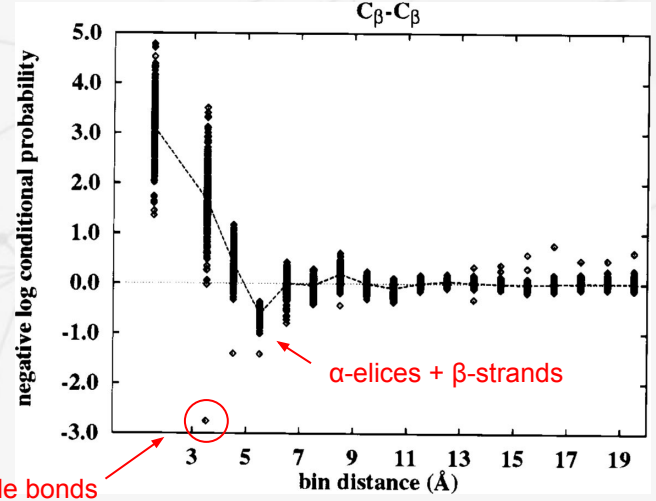
- All distances are **equally probable** → too simple
- All possible conformations of a **loop** → too simple
- Distances distribution observed in **random coils** → RAPDF would generate just compact structures
- Calculate the **average** from a set of possible compact conformations (from PDB)

RAPDF

- Different **distributions** for different **atomic combinations** (167 → atom + residue)
- Intra-residue contacts are excluded
- 17 **bins** (first bin → range [0, 3] Å; then → step of 1, up to 20)
- All counts intialized to one → to avoid log of zero

Too close →unfavored, clashes. Spread → artifact of pseudo counts

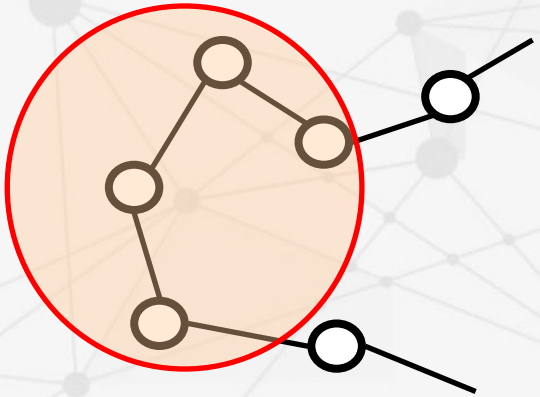




Solvation energy

Solvation energy requires calculation of **solvent-protein interactions**
It can be expressed as statistical potential

- Consider amino acids (AA) instead of atoms
- Calculate the number of residues *i* in the neighbourhood (eg 10 Å). (i = 0,…,40)
- For example, it can spot charged AA wrongly placed in the protein core → when there are a lot of surrounding atoms



$$energy = -\ln\left(\frac{P_{observed}}{P_{expected}}\right)$$

(Jones, 1999)

FRST

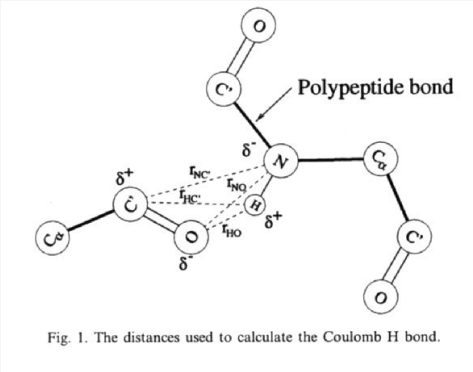
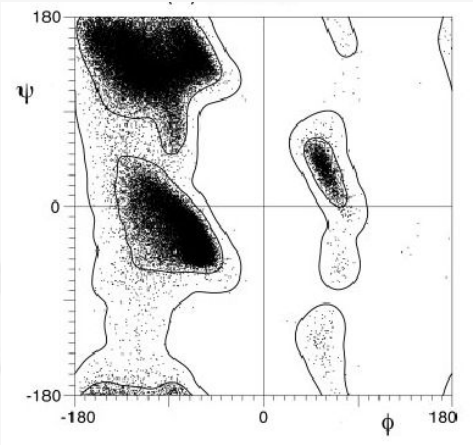


Fig. 1. The distances used to calculate the Coulomb H bond.



$$E_{FRST} = w_1 * E_{RAPDF} + w_2 * E_{SOLV} + w_3 * E_{HYDB} + w_4 * E_{TORS}$$

- **Solvation** (SOLV)
- **Hydrogen bonds** (HYDB). 3 distance measures
  - $2 \leq d(N_i, O_j) \leq 4$
  - $d(N_i, O_j) < d(N_i, C_j)$
  - $d(N_i, O_j) < d(Ca_i, O_j)$
- **Torsion angle** (TORS). Distribution of ϕ,ψ for each AA type
  - 10° bins

(Tosatto, 2005)

QMEAN

DIPARTIMENTO MATEMATICA

Table I
Short Description of the Terms and Their Combinations Used in This Work

Scoring function	Description
Torsion single	Ordinary torsion potential based on Φ and Ψ propensities of single amino acids. Bin size: 10°
Torsion three-residue	Extended torsion potential over three consecutive residues. Bin sizes: 45° for the center residue, 90° for the two adjacent residues
Pairwise Cα/pairwise Cβ	Residue-specific pairwise distance-dependent potential using Cα or Cβ atoms, respectively, as interaction centers. Range 3–25 Å, step size: 0.5 Å
Pairwise Cβ/SSE	In analogy to pairwise Cβ, but a secondary structure specific implementation was used both for the derivation and application of the potential.
Solvation Cβ	Potential reflecting the propensity of a certain amino acid for the a certain degree of solvent exposure based on number of Cβ atoms within a sphere of 9 Å around the center Cβ.
SSE X	Agreement between the predicted secondary structure of the target sequence (using method X, or consensus of three methods) and the observed secondary structure of the model as calculated by DSSP. QMEAN uses X = PSIPRED
ACCpro	Agreement between the predicted relative solvent accessibility using ACCpro (two states buried/exposed) and the relative solvent accessibility derived from DSSP (>25% accessibility => exposed)
QMEAN3	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, salvation Cβ
QMEAN4	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED
QMEAN 5	Weighted linear combination of torsion 3-residue, pairwise Cβ/SSE, solvation Cβ, SSE PSIPRED, ACCpro

(Benkert, Tosatto, Schomburg, Proteins 2008)

QMEAN

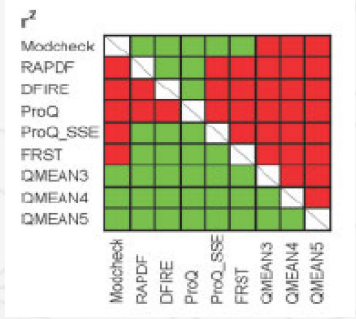
DIPARTIMENTO MATEMATICA

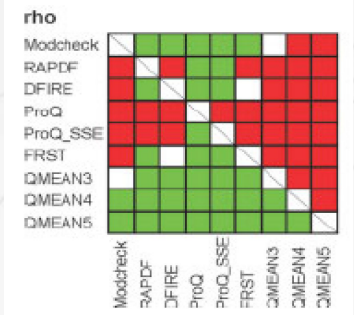
Table VI

Performance of Different Scoring Functions in Predicting the Quality of the Server Models Submitted for all 95 Targets of CASP7. Comparison of QMEAN With Other Well-Known MQAPs

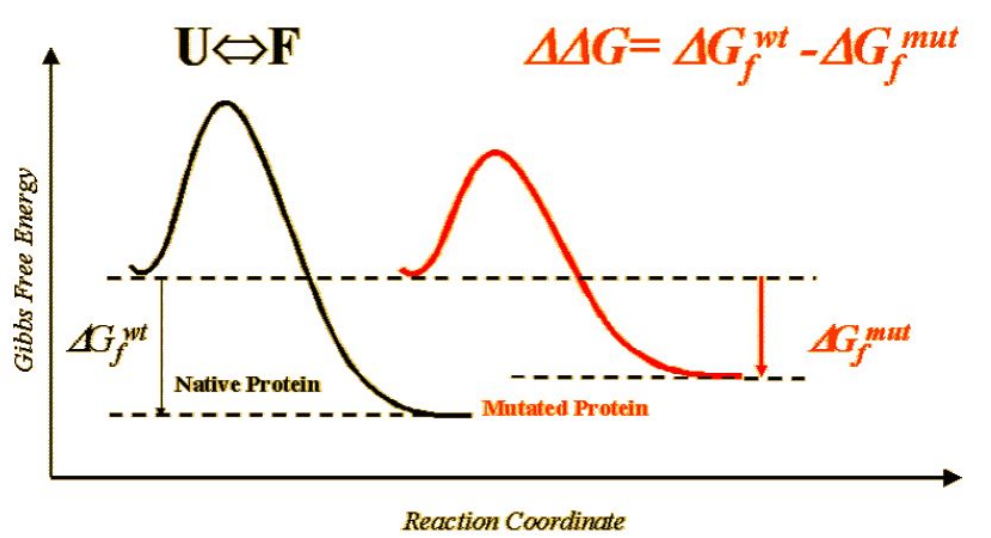
	Regression[a]		Enrichment[b]		Best predicted model[c]			Best model (GDT_TS)[d]			Native structure[d]		
Method	r^2	ρ	FE	$E_{15\%}$	Rank10	$\log P_{B1}$	$\log P_{B10}$	GDT_TS loss	Rank1	Rank10	Z_{nat}	Rank1	Rank10
Modcheck	0.64	0.59	0.33	2.70	17	−0.70	−1.67	−0.18	6	27	1.99	47	69
RAPDF	−0.50	0.50	0.31	2.44	17	−0.91	−1.67	−0.08	4	17	−2.09	55	77
DFIRE	−0.39	0.53	0.32	2.59	19	−0.93	−1.68	−0.08	5	18	−1.25	59	72
ProQ	0.36	0.26	0.13	1.22	5	−0.32	−0.99	−0.22	0	6	1.51	9	32
ProQ_SSE	0.54	0.43	0.19	1.71	8	−0.51	−1.21	−0.16	2	11	1.76	14	42
FRST	−0.57	0.53	0.30	2.36	21	−0.91	−1.74	−0.09	6	22	−2.41	56	72
QMEAN3	−0.65	0.58	0.33	2.57	16	−0.80	−1.83	−0.12	1	35	−2.27	59	75
QMEAN4	−0.71	0.63	0.38	2.76	28	−1.02	−1.90	−0.08	5	39	−1.86	55	69
QMEAN5	−0.72	0.65	0.40	2.90	30	−1.05	−1.94	−0.08	6	40	−1.89	56	71
Torsion single	−0.44	0.39	0.22	1.76	6	−0.60	−1.50	−0.13	0	13	−2.09	51	67
Torsion three-residue	−0.53	0.44	0.22	1.86	13	−0.76	−1.51	−0.11	1	10	−2.64	59	79
Pairwise Cβ	−0.58	0.51	0.30	2.51	17	−0.70	−1.70	−0.18	4	27	−1.96	39	69
Pairwise Cβ/SSE	−0.59	0.52	0.34	2.58	22	−0.84	−1.80	−0.13	5	36	−2.16	45	71
Solvation	−0.55	0.49	0.29	2.31	10	−0.55	−1.65	−0.24	2	27	−1.30	18	45
SSE PSIPRED	−0.65	0.52	0.24	2.03	9	−0.63	−1.43	−0.13	3	17	−0.89	7	25
ACCpro	−0.59	0.56	0.35	2.75	21	−0.85	−1.66	−0.11	6	33	−1.38	20	44

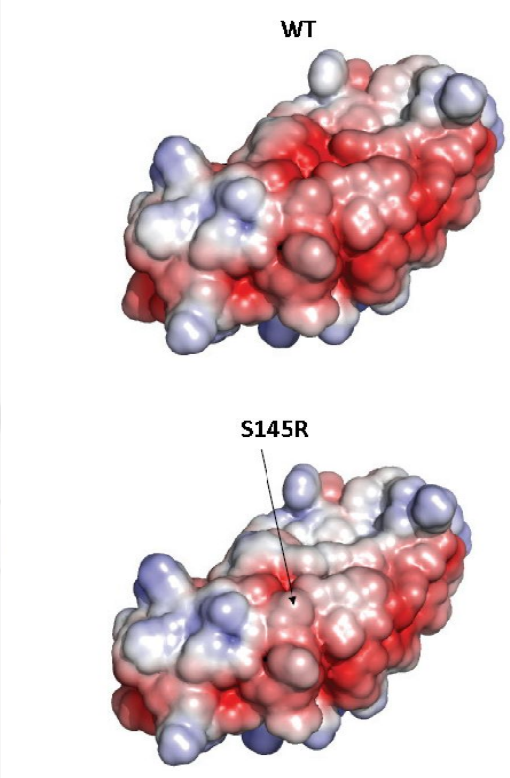
[a]Pearson's correlation coefficient r^2 and Spearmans's rank correlation coefficient ρ.
[b]*FE* stands for fraction enrichment and $E_{15\%}$ is the enrichment among the top 15% best predicted models as compared to a random selection.
[c]Rank10 are the number of targets for which the top-scoring models is among the top10 best models (based on GDT_TS). $\log P_{B1}$ and $\log P_{B10}$ are the log probability of selection the highest GDT_TS model as the best model or among the 10 best-scoring models, respectively.
[d]GDT_TS loss is the difference between the GDT_TS score of the best-scoring model and the best model in the decoy set. Z_{nat} is the *Z*-score of the native structure as compared to the ensemble of models. Rank1 *and* rank10 are the number of targets in which the native structure (or the best model based on GDT_TS, excluding the native structure) was found on the first rank or among the top 10 predictions.





Stability change upon mutation





- Predict ΔΔG
- Force fields or statistical potentials
- Machine learning