PROTEIN FOLDING PROBLEM

- How to go to the “most stable” conformation?
- Levinthal’s paradox: 100 amino acid peptide with 3 rotamers for each amino acid has $3^{198}$ wrong conformations.
- With random trial and error, folding of this protein would take billions of years... and yet, proteins in the nature can fold at sub-second scales.

THEN, HOW TO FOLD PROTEINS?

- We do not know the rules of cooperativity the Nature has so we need a human-made algorithm.

- How can you fold proteins or antibodies from secondary structure within the timescale of a PhD or a postdoc?

- How can you modify the amino acids of a known protein structure and predict the resulting structure of the modified segment?

Answer: You use Rosetta. Rosetta uses the Monte Carlo method.
MONTE CARLO ALGORITHM

Idea: Collect favorable structural changes until you reach the global minimum.

1) Sampling: Randomly generate conformations to explore the conformational space.
2) Scoring: Somehow measure how favorable the generated structures are.
   - Scores are calculated based on a score function.
   - Decision to keep or discard a conformation is made based on the Metropolis algorithm.
1) Move the structure to create a new conformation.

2) Calculate the $\Delta U$ as the energy of the $E_{\text{old conformation}} - E_{\text{new conformation}}$

3) If $\Delta U < 0$, then accept

   If $\Delta U > 0$, then calculate $W = \exp(-\Delta U/kT)$. If $W >$ a random number $R$ ($0 < R < 1$), then accept. Otherwise, reject.

In lay terms: If a structural change decreases the energy, keep the new conformation. If not, keep only if probability allows.
WHAT IS A SCORE FUNCTION?

- Score functions in the context of protein modeling consist of energy terms that are believed to represent protein interactions in a physical environment.

- The score terms can represent bonded, non-bonded, environmental, and statistical terms.

- The score terms can be given weights to adjust their contribution to the total energy of the system.
- Rosetta score function has *physical* and *statistical* terms.

- Has one-body (i.e. within the same residue) and two-body (i.e. between different residues) terms.

- The total scores are calculated as *weighted sum* of individual energy terms.

- Score units are Rosetta Energy Units (REU).

- Lower scores indicate more stable structures.
ROSETTA SCORE FUNCTION TYPES

1) **Low-resolution score function**: Uses a simplified residue representation to quickly scan conformational space. Especially useful for *ab initio* folding and loop modeling.

2) **High-resolution score function**: Uses a full-atom residue representation to score conformational terms. More realistic than the low-resolution score function.
LOW-RESOLUTION SCORE FUNCTION

- Side chains are represented as a single “centroid” pseudo-atom.

- The location of the centroid is determined using known structures from PDB as the average location of the side-chain atoms of the same residue type.

## LOW-RESOLUTION SCORE TERMS

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>env</td>
<td>Residue environment (solvation)</td>
</tr>
<tr>
<td>pair</td>
<td>Residue pair interactions (electrostatics, disulfides)</td>
</tr>
<tr>
<td>SS</td>
<td>Strand pairing (hydrogen bonding)</td>
</tr>
<tr>
<td>sheet</td>
<td>Helix arrangement into sheets</td>
</tr>
<tr>
<td>HS</td>
<td>Helix-strand packing</td>
</tr>
<tr>
<td>rg</td>
<td>Radius of gyration (vdW attraction, solvation)</td>
</tr>
<tr>
<td>cbeta</td>
<td>Cβ density</td>
</tr>
<tr>
<td>vDW</td>
<td>Steric repulsion</td>
</tr>
</tbody>
</table>

HIGH-RESOLUTION SCORE FUNCTION

- Residues are represented at full atomic level.
- Side-chains are represented explicitly as rotamers and statistical terms are calculated to assess their quality.
- Has several terms to model different types of solvation, electrostatics, hydrogen bonding energies.
## HIGH-RESOLUTION SCORE TERMS

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fa_atr</td>
<td>Lennard-Jones attractive between atoms in different residues</td>
</tr>
<tr>
<td>fa_rep</td>
<td>Lennard-Jones repulsive between atoms in different residues</td>
</tr>
<tr>
<td>fa_sol</td>
<td>Lazaridis-Karplus solvation energy</td>
</tr>
<tr>
<td>fa_intra_sol_xover4</td>
<td>Intra-residue Lazaridis-Karplus solvation energy</td>
</tr>
<tr>
<td>lk_ball_wtd</td>
<td>Asymmetric solvation energy</td>
</tr>
<tr>
<td>fa_intra_rep</td>
<td>Lennard-Jones repulsive between atoms in the same residue</td>
</tr>
<tr>
<td>fa_elec</td>
<td>Coulombic electrostatic potential with a distance-dependent dielectric</td>
</tr>
<tr>
<td>pro_close</td>
<td>Proline ring closure energy and energy of psi angle of preceding residue</td>
</tr>
<tr>
<td>hbond_sr_bb</td>
<td>Backbone-backbone hbonds close in primary sequence</td>
</tr>
<tr>
<td>hbond_lr_bb</td>
<td>Backbone-backbone hbonds distant in primary sequence</td>
</tr>
<tr>
<td>hbond_bb_sc</td>
<td>Sidechain-backbone hydrogen bond energy</td>
</tr>
<tr>
<td>hbond_sc</td>
<td>Sidechain-sidechain hydrogen bond energy</td>
</tr>
<tr>
<td>dslf_fa13</td>
<td>Disulfide geometry potential</td>
</tr>
<tr>
<td>rama_prepro</td>
<td>Ramachandran preferences (with separate lookup tables for pre-proline positions and other positions)</td>
</tr>
<tr>
<td>omega</td>
<td>Omega dihedral in the backbone. A Harmonic constraint on planarity with standard deviation of ~6 degrees</td>
</tr>
<tr>
<td>p_aa_pp</td>
<td>Probability of amino acid, given torsion values for phi and psi</td>
</tr>
<tr>
<td>fa_dun</td>
<td>Internal energy of sidechain rotamers as derived from Dunbrack’s statistics</td>
</tr>
<tr>
<td>yhh_planarity</td>
<td>A special torsional potential to keep the tyrosine hydroxyl in the plane of the aromatic ring</td>
</tr>
<tr>
<td>ref</td>
<td>Reference energy for each amino acid. Balances internal energy of amino acid terms. Plays role in design</td>
</tr>
<tr>
<td>METHOD_WEIGHTS</td>
<td>Not an energy term itself, but the parameters for each amino acid used by the ref energy term.</td>
</tr>
</tbody>
</table>

Source: [https://rosettacommons.org/demos/latest/tutorials/scoring/scoring#scoring-in-rosetta](https://rosettacommons.org/demos/latest/tutorials/scoring/scoring#scoring-in-rosetta) Taken on: 04/17/2019
SCORE FUNCTION WEIGHTS

METHOD_WEIGHTS ref 0.773742 0.443793 -1.63002 -1.96094 0.61937 0.173326 0.388298 1.0806 -0.358574 0.761128 0.249477 -1.19118 -0.250485 -1.51717 -0.32436 0.165383 0.20134 0.979644 1.23413 0.162496

fa_atr 1 fa_rep 0.55 fa_sol 0.9375 fa_intra_rep 0.005 fa_elec 0.875 pro_close 1.25 hbond_sr_bb 1.17 hbond_lr_bb 1.17 hbond_bb_sc 1.17 hbond_sc 1.1 dslf_fa13 1.25 rama 0.25 omega 0.625 fa_dun 0.7 p_aa_pp 0.4 yhh_planarity 0.625 ref 1
VAN DER WAALS ATTRACTION/REPELUTION

\[ V_{LJ} = 4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] = \varepsilon \left[ \left( \frac{r_m}{r} \right)^{12} - 2 \left( \frac{r_m}{r} \right)^{6} \right], \]

\( \varepsilon \) = energy well depths
\( r_m \) = sum of atom radii
\( r \) = atom-pair distance

---

ELECTROSTATIC INTERACTIONS

\[ E_{\text{Coulomb}}(i, j) = \frac{C_0 q_i q_j}{\varepsilon d_{ij}} \]

- \( q_i, q_j \) = charges of atoms \( i \) and \( j \)
- \( \varepsilon \) = dielectric constant
- \( d_{ij} \) = distance between atoms \( i \) and \( j \)

HYDROGEN BONDING

Acceptor – donor distance

Base-acceptor-hydrogen angle

Acceptor-hydrogen-donor angle

REFERENCE ENERGIES

- Unfolded state energies of individual amino acids.
- Residues with a large gap between folded and unfolded states are less likely to be found in proteins.
- Helps adjusting frequencies of amino acids during protein design.

\[ E_{\text{ref}} = \sum_i \Delta G_i^{\text{ref}}(\text{aa}_i) \]

Statistical terms are calculated based on the probability of finding certain rotamers/dihedrals in experimental structures deposited to PDB.

- Backbone dihedral angles
- Side-chain dihedral angles
- Side-chain rotamer probability
Rosetta also has additional score terms for both the low- and high-resolution score functions for more specialized applications. The additional terms can be used for several purposes:

1) To model non-protein biomolecules such as DNA or RNA.

2) To add constraints based on experimental data such as SAXS, NMR, and cryo-EM experiments to guide refinement procedures.

3) To add specific interactions that are not part of the default score functions.
Score functions are re-evaluated over time to add or adjust energy terms based on the emergence of new structural data.

- The current high-resolution score function in Rosetta is ref2015.

- Examples to previous score functions are talaris2013 and score12. These score functions are still used for some applications because they perform better than ref2015.
1) **Custom weights file:** A weights file with a desired set of weights for different terms can be inputted by the user. Example: your system environment is somehow significantly different than the assumptions made to generate the existing score functions.

2) **Patch file to modify existing weights:** A weights file can be used to “patch” the terms of an existing score function. Example: you want to re-scale the electrostatic terms without changing the rest of the term weights.

3) **Set weights to specific terms from the command line:** Changes to individual terms can be done using the command line instead of providing an external weights file. Example: you want to apply changes ad hoc to individual terms on the go.
SPECIALIZED SCORE FUNCTIONS

**Membrane score function(s):** Rosetta has low- and high-resolution membrane scoring functions to model proteins in a membrane environment. These functions have additional terms to include residue-membrane interactions and changes to non-bonded term weights to better represent the membrane environment.

**Orbital score function:** Uses an orbital representation to calculate partial covalent interactions such as hydrogen bonds, $\pi$-cation, $\pi-\pi$ interactions, and salt bridges more accurately.


SCORING WITH NON-PROTEIN RESIDUES

- If parameters for a molecule do not exist, you need to create parameters files to define the new atom types and conformers of the new molecule.

- The new atom types and properties are used to calculate the energy terms with the existing weights of the score function.

- Parameters for some D-amino acids, non-canonical amino acids, peptoids, sugars, DNA, and RNA are present in the Rosetta database under database/chemical/residue_type_sets.

- Rosetta score functions consist of a number of physical and statistical weighted energy terms.

- The score of a pose is a relative measure of its stability rather than a global metric such as free energy.

- Low-resolution (centroid) score function is useful when exploring conformational space with large backbone motions whereas high-resolution (full atom) score function is useful for applications involving side-chain motions.

- The term weights of the score functions can be modified (though not suggested) or additional terms can be included to account for specific interactions.

- Scoring non-protein species or unnatural amino acids is possible, but requires additional work and care.
ADDITIONAL RESOURCES

Rosetta score function overview


Rosetta scoring tutorial:

https://rosettacommons.org/demos/latest/tutorials/scoring/scoring#scoring-in-rosetta

Additional terms

https://www.rosettacommons.org/docs/latest/rosetta_basics/scoring/score-types-additional

Membrane scoring

https://www.rosettacommons.org/docs/latest/application_documentation/membrane_proteins/RosettaMP-GettingStarted-Overview

Orbital scoring

https://www.rosettacommons.org/docs/latest/rosetta_basics/scoring/NC-scorefunction-info#Partial-Covalent-Interactions-Energy-Function-(Orbitals)

Non-protein residue modeling

https://www.rosettacommons.org/docs/latest/rosetta_basics/non_protein_residues/non-protein-residues

Pyrosetta scoring examples