

Glossary of (some) Rosetta Terms

ab initio - prediction of protein structure from sequence (equivalent to “de novo”)

B factor - the “temperature factor” from crystallography and seen in PDB files, the larger the value the more “flexible” the atom is

backbone - the part of an amino acid that is comprised of the C, O, C alpha, and N atoms

benchmark studies - tests done to confirm the performance of a new algorithm or method, results are compared to previous results using the same starting data

BioChemicalLibrary - BCL - a suite of programs written by the Meiler lab, contains programs for de novo protein folding, sparse data folding, ligand conformer generation, and more

Boolean - data type, can be either true or false

Calibur - a tool for clustering protein decoys written by SC Li and YK Ng

C alpha - the backbone carbon atom that the side chain is connected to

Cambridge Structure Database - CSD - a database containing small molecule crystal structures

Cartesian coordinates - the standard x,y,z description of the position of a point in space

C beta - the carbon atom of the side chain bonded to C alpha (glycine does not have a C beta atom)

centroid - a simplified representation of an amino acid, in Rosetta an amino acid is comprised of the backbone atoms, the Cbeta atom and one pseudo-atom representing the side chain

chain - a subset of amino acids that comprise a protein

chi - dihedral angles that set the 3 dimensional positioning of the side chain atoms, number sequentially chi1, chi2, etc outward from Calpha, chi1 is the angle between N-Calpha-Cbeta-Cgamma

clashes - two (or more) atoms are too close to be energetically favorable (essentially an overlap of vdW radii)

ClustalOmega - a program that aligns two (or more) sequences

cluster - grouping two (or more) protein models together based on similar 3D structure

coarse grain - initial modeling, where all atoms or energy terms may not be represented

comparative modeling - prediction of protein structure based on sequence and the structures of closely related proteins.

conformer - one of a set of 3 dimensional orientations a ligand, small molecule or amino acid side chain

constraints - actually “restraints”; adjustments to the score function to take into account additional geometric information

de novo - prediction of protein structure from sequence (equivalent to “de novo”)

decoy - a model of a protein produced by Rosetta

density map - experimental data showing where the electrons (and thus the atoms) are

design - to predict the protein sequence which has a desired function

dihedral - aka torsion; the degree of freedom of rotating around a bond

docking funnel - an energy funnel for docking

Dunbrack library - rotamer library from the Dunbrack laboratory, the standard rotamer library of Rosetta

energy function - the “score function”; the prediction of structural energy over which Rosetta operates

energy funnel - a plot showing low rmsd structures having lower energies than high rmsd structures

ensemble - a group of closely related structures

ex1/ex2 - options that specify the size of rotamer library being used

fasta - text based format describing the peptide sequence of a protein, single letter amino acid codes are used

filters - during a run, a check on the quality of the model being generated, if the model does not pass the given test (filter), it will be discarded

fixbb - option setting the backbone atoms fixed during a protocol

flags - options set by user to control the behavior of Rosetta (or other programs), can be set on the command line, on in the options file

float - In computer programming a variable which is a real number, can be a whole number or fractional

fold tree - a representation of all the residues in a protein, relates internal coordinates to cartesian coordinates, if the backbone position of one residue changes the fold tree will propagate the changes throughout the protein

fragments - 3 and 9 residue sections of protein structures, used by Rosetta to build protein models

fullatom - same as "all atom", when all atoms of a protein or molecule are individually represented

global minimum - the 3 dimensional conformation of a protein which corresponds to the lowest possible energy state, this is (usually) the conformation found in nature

hard_rep - normal Lennard Jones repulsive - used in contrast to soft_rep

heavy atom - all atoms of the backbone and sidechains except hydrogens

homology modeling - prediction of the 3D structure of a protein based on the structure of a homologous protein or proteins (typically shares 30% sequence similarity or above) equivalent to comparative modeling

I/O - input / output, usually in regards to a computer program

interaction graph - a representation of protein interactions during packing; can affect simulation speed

interface - the region of a structure where two chains interact

internal coordinates - representation of structure by bond lengths/angles/dihedrals, rather than Cartesian xyz coordinates

jump - a portion of the fold tree representing a rigid body (non-covalent) movement

knowledge-based potentials - energy function terms based on the probability of occurrence in a data set

Lenard Jones potential - LJ - A function that approximates the non-bonded interactions of neutral atoms, combines Pauli repulsion and the van der waals attractive term (also known as Lennard Jones 6-12 potential)

ligand - a molecule which binds a protein; for Rosetta a (non-polymeric) small molecule, specifically

local minimum - the lowest energy 3 dimensional state of a protein in a neighborhood of similar conformations, there may be many local minimums of a protein, but only one global minimum

low energy - A 3 dimensional model of a protein that has good packing, satisfied polar or charged residues, appropriately placed small molecules or ligands, etc

low Resolution - an experimentally determined structure of a protein where the resolution between atoms is not distinct, a crystal structure resolution above 3-4 angstroms

main chain - used interchangeably with backbone atoms

Metropolis criterion - Used by Monte Carlo methods, this equation tells whether to accept or reject a random move

minimize - optimize the protein structure by making small movements to lower energy conformations

mmCIF - macromolecular Crystallographic Information File, file format used to describe the 3 dimensional structure of a protein

model - a representation of the 3 dimensional structure of a protein

MOE - Molecular Operating Environment, a suite of programs designed for drug discovery and modeling

MOL format - a file type that contains information about the structure of a chemical; same as "SDF format"

Monte Carlo methods - a computational algorithm that uses random sampling to predict a result, in protein folding prediction Monte Carlo methods involve the random move of a side chain or backbone in 3 dimensional space

MoveMap - a list of mobile and immobile parts of a protein

Mover - Rosetta object that modifies a structure, complex movers can be built from simpler movers

native structure - the structure of a protein, ligand, etc that is found in nature, usually refers to the crystal or NMR structure of a protein

nstruct - the number of models that Rosetta will output

Octopus - a program that predicts the topology of membrane proteins from sequence alone

options - user specified directions given to Rosetta, either through the command line or through the options file, sometimes called "flags"

packing density - how close atoms are to each other; closer is better, up to a point

Packer - the part of Rosetta which does repacking; it uses Metropolis Monte Carlo Simulated Annealing to optimize rotamers

params file - a file which tells Rosetta how a residue behaves

Parser - another name for RosettaScripts

patch files - a file which makes a small adjustment to a score function

PDB - Can refer to either the Protein Data Bank, a website that contains structural information of proteins, usually determined by x-ray crystallography or NMR. Or PDB can refer to the file type used by the protein data bank to represent the 3 dimensional structure of a protein

phi - the dihedral angle describing the position of the C-N-Calpha-C atoms

Pose - Rosetta representation of a molecular system; contains the structure and associated properties

psi - the dihedral angle describing the position of the N-Calpha-C-N atoms

PyMol - Software that allows for the visualization of proteins, ligands, DNA, etc, see also RasMol, VMD, reads PDB files

refine - to take a crude structure and make it better

Relax - a protocol in Rosetta which optimizes the structure of the protein

repack - determine the conformation of sidechains which minimizes the energy

repulsive term - The part of the Lennard Jones equation which describes the effects of overlapping electron orbitals, the energy will be positive

resfile - a file which describes how to repack or design a protein

residue - the basic unit on which Rosetta operates; generally one polymeric unit (one amino acid)

restraints - adjustments to the energy function; often called "constraints" in Rosetta

REU - Rosetta Energy Units - Rosetta's arbitrary energy term, does not correspond with physical energy measurements

rigid body movement - translation and rotation of a chain as a whole, without internal bond length/angle/dihedral changes

RMSD - root mean squared deviation, the difference in 3 dimensional structure between two proteins

Robetta - an online, automated tool for protein structure prediction and analysis

RosettaScripts - an XML based interface for controlling Rosetta, allows the user greater control of methods, score functions, etc, without requiring the user to change the source code of Rosetta

rotamer - The 3 dimensional positions of amino acid side chains commonly observed in nature, can more generally refer to any position of a side chain

SASA - solvent accessible surface area – the area of a protein that can be reached by water or another solvent

score file - an output file created by Rosetta that contains a list of poses created, their energy, and the energy term components

score terms - Rosetta's energy function is a combination

scorefunction - the part of Rosetta that handles scoring (ie assigning an energy) of a given pose

scoring grid - a rapid pre-calculation of scoring for ligand docking

SDF format - a file format that describes the structure and connectivity of a molecule, used primarily for small molecules, not for proteins; also known as MOL format

side chain - The variable portion of an amino acid, the R group

simulated annealing - an optimization protocol, used by the Packer

small molecule - for Rosetta, anything that's not a polymeric biomacromolecule

soft_rep - energy function where the Lennard Jones potential is adjusted so that clashes aren't scored as badly; contrast "hard_rep"

string - in computer programming a set of alphanumeric characters, can be a single letter or many words

symmetry definitions - symdef files tell Rosetta how to treat a symmetric protein

target sequence - the sequence of the protein of unknown structure you're trying to model

TaskOperations - specifications in RosettaScripts which tell the Packer how to optimize rotamers

template structure - the known structure of a closely related protein that you're using to model your protein of interest

torsion - aka dihedral; the degree of freedom of rotating around a bond

torsion space - internal coordinates; torsion space minimization optimizes the protein by rotating dihedrals

van der Waals - describes the interactions between neutral, non-bonded atoms, in protein prediction often used interchangeably with Lennard-Jones potential

weights - specification of which score terms in which proportions should be used in the scorefunction

XML - a hierarchical data format, a custom version is used by RosettaScripts