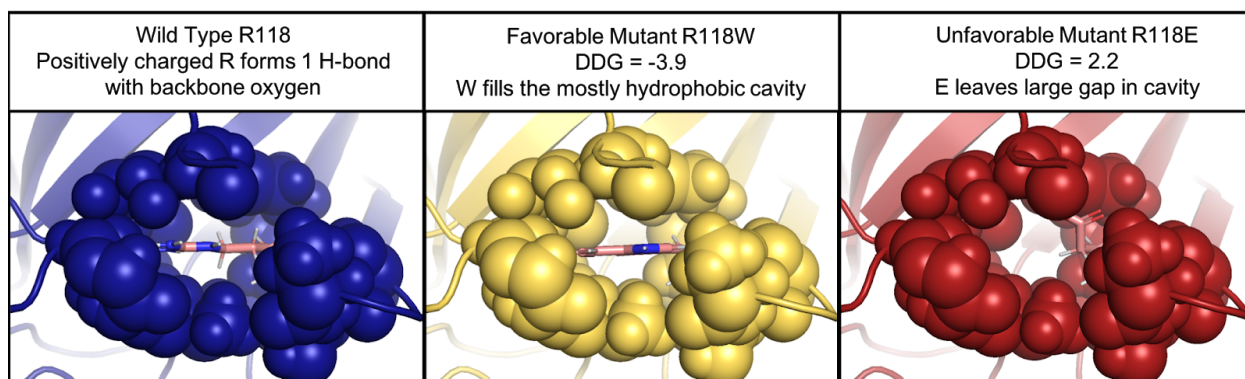


# Free Energy of mutation calculations in Cyrus Bench

Copyright Cyrus Biotechnology, 2019

## Background

Delta Delta G (DDG) is a metric for predicting how a single point mutation will affect protein stability. DDG, often referred to as  $\Delta\Delta G$ , is the change in the change in Gibbs free energy (double changes intended). DDG is a measure of the change in energy between the folded and unfolded states ( $\Delta G_{\text{folding}}$ ) and the change in  $\Delta G_{\text{folding}}$  when a point mutation is present. This has been found to be an excellent predictor of whether a point mutation will be favorable in terms of protein stability. This tool is therefore of widespread utility in computational protein engineering. Here we will introduce the theory behind DDG, discuss implementations in Cyrus Bench, and look at some example cases.



As of early 2019, the Cyrus DDG tool uses a new version of the Rosetta DDG calculation, “Cartesian DDG”, which has a variety of improvements over the previous method (Kellogg). The new method is more complex, but fully automated in Cyrus Bench, and features improvements in energy function, mutations with change in net charge, proline free energies, solvation models, side chain optimization, and structure preparation. See the support document for DDG version 2 for more scientific detail, and Park et al. below. The new version has improved accuracy, better consistency across similar input protein structures, and most notably fewer outliers with highly inaccurate calculated results.

For more information:

Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*. 2011 Mar;79(3):830-8.

Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J Chem Theory Comput*. 2016 Dec

<https://cyrusbio.com/2019/01/07/cyrus-bio-announces-full-release-of-rapid-and-accurate-protein-mutational-free-energy-tool-in-cyrus-bench%E2%93%87/>

## A Deeper Look at DDG for Proteins

As a reminder, Gibbs free energy ( $G$ ) = Enthalpy ( $H$ ) - Temperature ( $T$ ) x Entropy ( $S$ ).  $H$  is the internal energy of a protein.  $H$  decreases during protein folding because folding will cause packing of hydrophobics, optimized polar group orientation, and achieves a good proximity to ideal bond lengths and angles.  $S$  is the measure of order within a system.  $S$  of a protein becomes lower during protein folding because residue dynamics will be significantly decreased in comparison the unfolded state. However, decrease in protein  $S$  will also cause an increase in  $S$  for solvent.

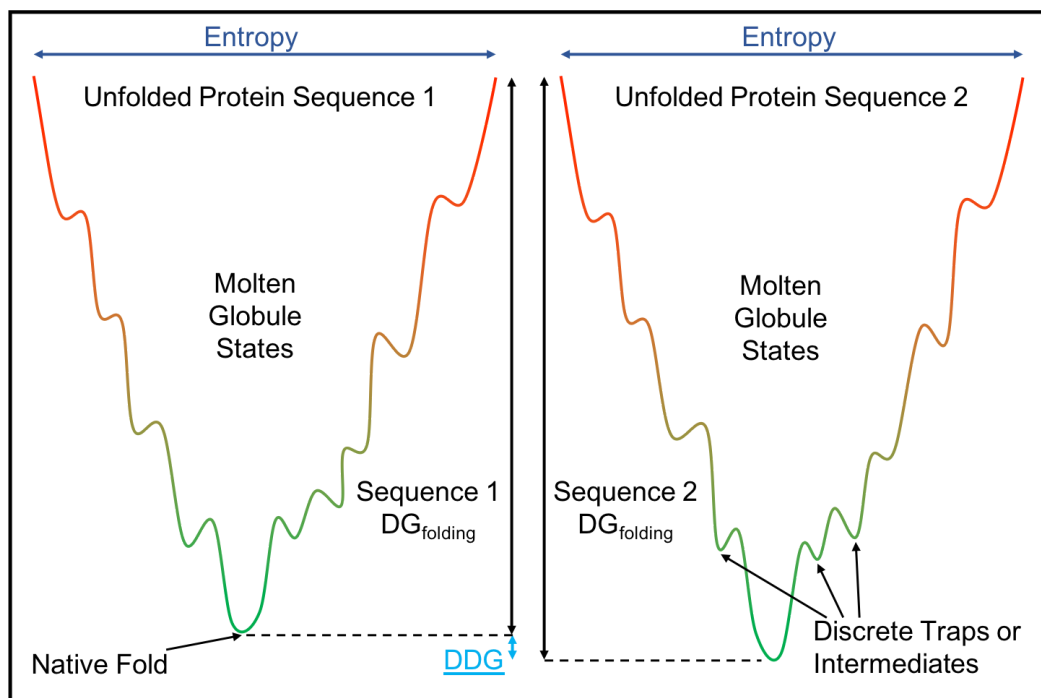
$$\Delta G_{\text{folding}} = (H_{\text{unfolded}} - H_{\text{folded}}) - T(S_{\text{unfolded}} - S_{\text{folded}})$$

$$\Delta G_{\text{mutation}} = (\Delta G_{\text{folding}})_{\text{WT}} - (\Delta G_{\text{folding}})_{\text{Mutant}}$$

DDG is calculated using Rosetta. So is calculated in Rosetta Energy Units (REU):

$$\Delta \text{REU}_{\text{mutation}} = (\text{REU}_{\text{unfolded}} - \text{REU}_{\text{folded}})_{\text{WT}} - (\text{REU}_{\text{unfolded}} - \text{REU}_{\text{folded}})_{\text{Mutant}}$$

The energy of an unfolded protein is nearly identical with a single point mutation. So, the dominant factor in DDG is the energy difference of the folded state.



The diagram above is a simplified 2D energy landscape. The x axis represents different conformations of the protein. Energy of a protein structure is measured along the y axis. The unfolded states have the highest energy. In a 3D version, the many conformations that a protein can achieve are on x/z axis, so this translates to the Entropy of a protein at an energetic state. Paths that follow the line down towards the global minimum represents conformational changes that occur during folding. The local minimas represent discrete traps or intermediate structures. Proteins must then jump the energetic hurdle to continue down the line in order to find the global minimum. In this example, sequence 2 reaches a lower minimum so is a more stable.

### DDG Calculation

The program takes a structure and runs an optimization to sample nearby conformations. Each optimization starts by repacking the side chains, then does a gradient-based minimization three times. The first time has the repulsion term weighted at 10%, then 33%, then 100%, which allows more sampling in order to move through unfavorable transition states to find a lower minimum. Minor backbone changes are allowed during optimization. After 50 repeats using the same original structure, it takes the average score of the best three structures.

The Rosetta App: ddg\_monomer, which uses ddGMover

### Running DDG

Click the folder on the left that contains your protein structure to bring it into the center window. Click the box next to the structure so that a check mark appears in the box. Then you can click the **DDG** button on the right lower panel under Metrics to bring the run into the center window.


By default, the selector will say **Selector (all)**, but you can switch to one of the selectors that you have created. The chart showing all original sequences and all potential point mutations is shown in the bottom center.

*	Orig.	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
15	L																				
19	S																				
22	L																				
26	E																				
56	F																				
60	R																				
63	L																				
64	S																				

You can select (or deselect) all mutations at all location by clicking **\***. You can also just click an entire row or column, as well as any individual box. Original sequences cannot be selected

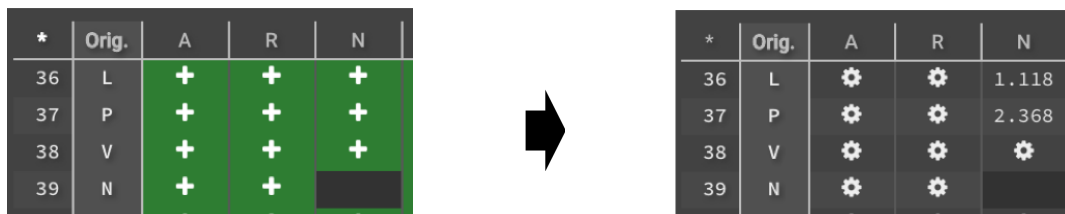
because the  $\Delta G$  is 0. You can upload your desired mutations by clicking **Upload from CSV**. Once you have selected all the mutations that you want to measure, click

 Save & Run selected

. Before beginning the run, your selections show , will show





while the calculation is running, and once it is done will show the  $\Delta G$ ...



*	Orig.	A	R	N
36	L	+	+	+
37	P	+	+	+
38	V	+	+	+
39	N	+	+	

*	Orig.	A	R	N
36	L	⚙	⚙	1.118
37	P	⚙	⚙	2.368
38	V	⚙	⚙	⚙
39	N	⚙	⚙	

Negative  $\Delta G$  indicates that the mutation will improve stability while positive  $\Delta G$  will disrupt stability. You can download the entire chart by clicking  Download as CSV . The default output in kcal/mol, you can switch to Rosetta Energy Units by clicking .

Note: Running DDG for residues at a protein-ligand interface needs to be run differently.

First, run DDG for the protein-ligand complex, then repeat the same calculation for the protein alone. This will separate the energetic changes that happen with the protein alone vs stabilization of the interface.

$$DDG_{\text{interface stability}} = DDG_{\text{complex}} - DDG_{\text{protein}}$$

DDG has not been benchmarked for this specific use.

### Understanding and Using DDG Results

DDG results will fall into three categories:

- 1)  $DDG > 0.5$ : Positive results suggest that a mutation would be destabilizing. Most mutations will be positive or close to zero because proteins have evolved to be reasonably stable. These mutations are residues that you should usually avoid during design.
- 2)  $0.5 > DDG > -0.5$ : Things that are near 0 are within the noise range so should be considered neutral. These can be included in design to allow more neutral changes in the protein that may compensate for changes in the protein made by
- 3)  $DDG < -0.5$ : Negative results suggest that the mutation would lead to a more stable protein. However, the environment at each position should be considered.
  - a) If interacting molecules are not present in the model, such as at a known zinc binding site, then a seemingly favorable mutation will not be favorable in reality.

- b) A positions that has a lot of negative DDGs could mean that this position evolved a destabilizing residue because it is necessary for its catalytic activity, for binding another molecule, or because of another functionally relevant reason.
- c) Also, consider that this measures a single point mutation. Many times it requires multiple interacting mutations in order to achieve significant stability.

DDG is only predictive for a single mutation at a time. In order to predict whether multiple mutations would provide an additive effect on stability, you must create the mutations in design and run some Relax repeats. Then compare its energy to the original sequence (which also needs Relax repeats to be comparable). This is nearly as accurate as DDG, takes longer to calculate.

### Using DDG Results, Two Examples:

Here are two examples of how DDG can be an incredibly useful tool.

#### 1) Predicting stabilizing mutations for the p53 core domain

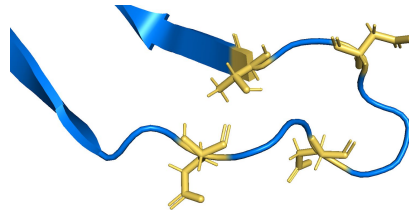
We uploaded the crystal structure, 2OCJ, into CAD and ran prepare followed by 100 repeats of Relax. Then took the lowest energy structure to run DDG, allowing all possible mutations for the entire structure. The results showed that only 23 of the 194 residues had point mutations with at least one DDG < -1.0 kcal/mol (shown in yellow below).



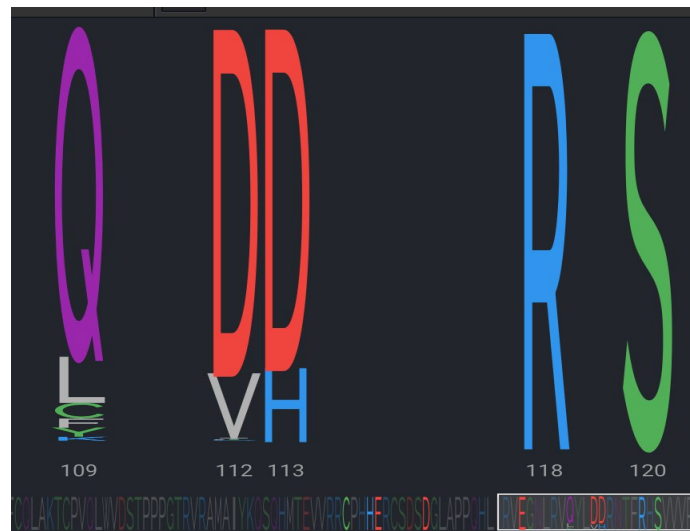
These can be used to guide design in order to stabilize the structure. For these 23 positions, there are 65 point mutations that have DDG < -1.0 kcal/mol. Always avoid making mutations at locations that would affect function. In this case, 5 of the positions are in the DNA binding region. So, eliminating them as mutable spots leaves 18 positions with 58 mutations. If the original residues are included in design, that makes over 18 billion possible sequence permutations. It would take years of calculation to properly sample them. However, if you run design for 3-6 positions at a time, you can split the job into 4 design jobs. Then take the best sequence sets from these jobs to run a full protein design. This will take just a few hours per

design and the final answer will be almost as good as the larger calculation, because local interactions have the strongest influences on mutations.

For p53, there are 4 clusters of residues. Each cluster consists of residues that are within a single secondary structure or that are in a chain of interacting residues. The cluster shown below has 4 residues on a loop and has 14 mutations. This is a very tractable design case.



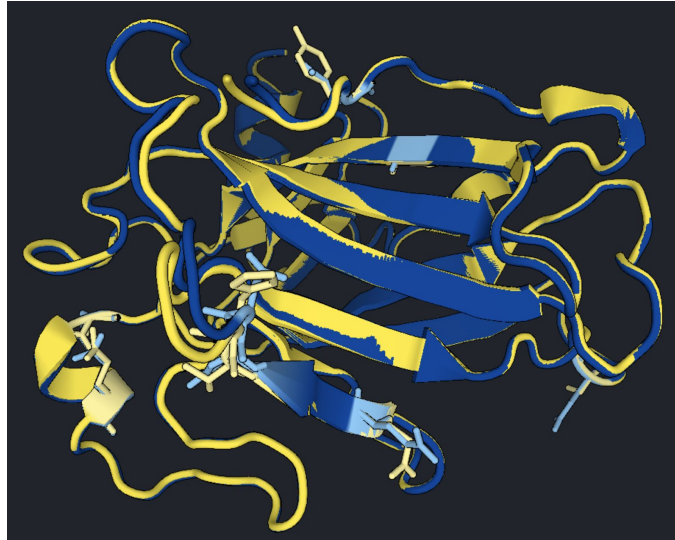
We ran up to 1,000 repeats of design for each set. One of the four designs only had 16 possible permutations so we only ran 100 repeats. But the largest design run had 3,360 permutations so 1,000 repeats was necessary to get proper sampling. The Logo for the largest design run is shown below.



From the Sequence Logo, you can see that position 109 has several potential mutations, 112 & 113 only have 2 preferred residues, and 118 & 120 only have the original residues present. Interestingly, position 109 was the only spot where the WT sequence (E) wasn't the most favored. For 118, there were 3 non-WT residues that were favored by DDG, but in design we can see that only the WT is preferred. We trust design more than DDG because it allows more conformational sampling as well as design.

After each design set is run, we look at the Sequence Logo to see what residues were most often selected. From the Sequence Logo above, we would select the 6 possible mutations at position 109, 2 for 112 & 113, and keep WT for 118 & 120. The same protocol is done for the other 3 design sets for this protein so that all desired positions can have design done simultaneously, using this smaller, more informed mutation set. In this case, we ended up with

48 permutations. We ran Flex Design with 500 repeats. The sequences of 80% of these converged on a single sequence. The new sequence had lower energy by 13 REUs, which is a significant improvement and had mutations at 8 positions.



Above is the original structure in blue and the stabilized mutant in yellow in cartoon mode. The mutation positions are shown in sticks.

Additional stabilization may be possible by starting with the stabilized structure and repeating the process. Starting with a new DDG calculation. Then using the results to guide another round of design. You can see that the backbone is nearly identical, but the slight backbone shifts and new sequence may find a few more mutations to add to the design process. Also, since this was done using Flex Design, the process can be repeated with Relax Design to sample more search space during modeling of mutations.