Supplemental Information

Computing the relative stabilities, and the per-residue components, in protein conformational changes

Arijit Roy[1], Alberto Perez[1], Ken A. Dill[1,2,3,*], and Justin L. MacCallum[1]

[1]Laufer Center for Physical and Quantitative Biology
[2]Departments of Physics
[3]Departments of Chemistry, Stony Brook University, Stony Brook, NY 11794.
[*]Correspondence:dill@laufercenter.org

October 28, 2013

Extended Experimental Procedures

Theory for the Confine-Convert-Release Method

We compute the total conversion free energy from the thermodynamic cycle as

$$\Delta G_{A,B} = \Delta G_{A,A^*} - \Delta G_{B,B^*} + \Delta G_{A^*,B^*}. \quad (1)$$

The free energy associated with the confine step, $\Delta G_{A,A^*}$ and release step, $\Delta G_{B,B^*}$ are computed following an identical procedure. For concreteness, let us focus on the process of computing $\Delta G_{A,A^*}$. We define a reaction coordinate λ for the confine and release steps. So, for both of those processes, we compute the total free energy using thermodynamic integration

$$\Delta G_{\mathbf{RR}^*} = \int_0^1 \langle \delta U(\mathbf{R}, \lambda)/\delta\lambda \rangle_\lambda d\lambda. \tag{2}$$

where $\mathbf{R}$ is the vector of all the atomic coordinates. The term $\langle \delta U(\mathbf{R}, \lambda)/\delta\lambda \rangle_\lambda$ indicates the ensemble average of the derivative of the potential energy with respect to λ at each particular λ along the confine or release paths. The state A is at $\lambda = 0$ and A^* is at $\lambda = 1$. The energy function $U(\mathbf{R}, \lambda)$ is

$$U(\mathbf{R}, \lambda) = U_0(\mathbf{R}) + \frac{1}{2}\lambda k_f (\mathbf{R} - \mathbf{R}^*)^2 \tag{3}$$

where U_0 is the force-field energy of conformation $\mathbf{R}$ and the second term is the harmonic potential that is used to confine the system. As a progress variable $k = \lambda k_f$ describes the strengthening of springs. We gradually increase the potential restraint such that the system travels from an almost free state to a highly confined state, freezing out most of the motion. So, Eqn. 2 can be rewritten as

$$\Delta G^{conf} = \frac{1}{2}\int_0^{k_f} < (\mathbf{R} - \mathbf{R}^*)^2 >_k \, dk = \frac{1}{2}\int_0^{k_f} \mathbf{R}_k dk \tag{4}$$

where $\mathbf{R}_k =< (\mathbf{R} - \mathbf{R}^*)^2 >_k$.

$\mathbf{R}_k$ gives the fluctuations from the reference structure. We compute $\mathbf{R}_k$ at different values of k. This can be done by simply monitoring the all-atom root-mean-square deviation (rmsd) from the reference structure. Next, following Tyka et al.[1], we perform numerical integration to get $\Delta G_{A^*,B^*}$. For the release step, $\Delta G_{B,B^*}$ we follow the same steps in a reverse order.

In the convert step, we calculate the free energy difference, $\Delta G_{A^*,B^*}$ associated with conversion of highly restrained or frozen state A^* to B^*. As both states are highly restrained, normal mode calculation applies well. After computing the G_{B^*} and G_{A^*} separately using normal mode calculations,

we calculate the free energy cost associated with this conversion using the equation, $\Delta G_{A^*,B^*} = G_{B^*} - G_{A^*}$.

Finaly, The total difference free energy, $\Delta G_{A,B}$ between the two state A and B is calculated using equation Eqn. 1.

Now, to decompose the total confinement free energy into per-residue free energies (PRFE), we use the expression

$$(\mathbf{R} - \mathbf{R}^*)^2 = \sum_{j=1}^{N} (r_j - {r_j}^*)^2 \tag{5}$$

where, r_j and ${r_j}^*$ are coordinates for residue j and its reference structure respectively, and N is the total number of residues. This decomposition is exact for $\Delta G_{A,A^*}$ and $\Delta G_{B,B^*}$. However, we can not perform the normal mode analysis for each residue. Thus, we only computed the average enthalpic contribution of each residue from the molecular dynamics trajectory of two highly restrained state. The free energy associated with convert steps now becomes $\Delta G_{A^*,B^*} \approx H_{B^*} - H_{A^*}$. The enthalpic contribution of each residue was computed using decomp module of the AMBER software package. The quantity $\Delta G_{A^*,B^*}$ is dominated by an enthalpy, since the states are sufficiently constrained that they have no conformational entropies. Since the states A^* and B^* are highly confined, the only contributor to the entropy of these states is vibrational. We found this contribution to be less than 0.40 kcal/mol for all the cases that we have studied, supporting our assumption that this entropy is small. We then use, Eqn. 1 to compute the per residue conversion free energy.

Preparation of input models for chameleon sequences

We prepared computer generated models in both α and β folds for all five chameleon sequences mentioned in the main text. To prepare the α and β folds for the GA95 sequence, we started with the crystallographic conformation of the GA95 (α fold) and GB95 (β fold) sequences. We then removed the side chains from residues at three mutated positions and from their neighbors (within $4\mathring{A}$) in both folds. Next, we called the program SCWRL4[3] to generate the side-chain conformations of the mutated residues and their neighbors. We performed extensive molecular dynamics minimizations to remove any possible bad contacts. This procedure was used to generate

all α and β folds for the five sequences that were used in the free energy calculations.

Per-residue free energy calculations for chameleon proteins

In order to better understand the switching mechanism of chameleon proteins, we approximately decomposed the total free energy into per-residue contributions. Our analysis revealed the importance of a few amino acid residues for controlling the free energy equilibrium between β and α folds. In this section, we explain some of the mechanistic details behind the per residue free energy, $\Delta\Delta G(\beta - \alpha)$ for individual amino acid residues. The first eight residues in the α fold are disordered. Some of these residues are hydrophobic and stabilize the β fold as they form part of the hydrophobic core. This effect is most prominent in the case of L7 (shown in Figure S5(A)). The next big peak that favors the β fold is at position 26, where, in a similar way, A26 is part of the hydrophobic core in the β fold, but solvent-exposed in the α fold (Figure S5(B)). In Figure S5(C), the Y45 residue stablizes the GB95 sequence in the β fold, as it has an H-bond with D47, which is absent in the α fold. On the other hand, in GA95, L45 favors the α fold, as it is part of the hydrophobic core. The first residue that favors the α fold in a major way is Q11, which forms a salt bridge with E15 (Figure S5(D)). The residue that most favors the α fold is I49, as it is part of hydrophobic core in the α fold (Figure S5(E)), but exposed to the solvent in the β fold. This effect is greater in the GA95 sequence than in GB95. We tested the stabilty of the α fold by changing the residue (I to F) at position 30 in sequences GA95 to GB95 (Figure S5(F)). The side chain of I30 is inside the hydrophobic core in the GA95 sequence and α fold. In contrast, F-30, having a larger side chain, is exposed at the protein surface in the GB95 sequence and α fold.

References

[1] Tyka, M.; Clarke, A.; Sessions, R. An Efficient, Path-Independent Method for Free-Energy Calculations. J.Phys.Chem. B 2006, 110, 17212-17220.

[2] Cecchini, M., Krivov, S.V., Spichty, M., Karplus, M. Calculation of free-energy differences by confinement simulations. Application to peptide conformers. J. Phys. Chem. B. 2009, 113, 9728-9740.

[3] Krivov, G.G.; Shapovalov, M.V.; Dunbrack, RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. (2009) 77, 778-795.

[4] Perez, A.; Yang, Z.; Bahar, I.; Dill, K.A.; MacCallum, J.L.; FlexE: Using Elastic Network Models to Compare Models of Protein Structure. J. Chem. Theory Comput., 2012, 8, 3985-3991.

Supplemental Data

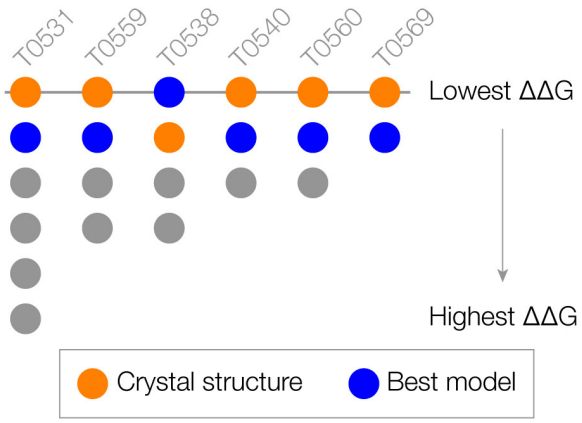


Figure S1: Related to Figure 2 and Result subsection ”CCR can often distinguish CASP-model predictions from true native structures”: The confine-convert-release method is usually able to identify the native structure and the best model (the model with the highest GDT-TS in CASP competition) from a set of decoys.

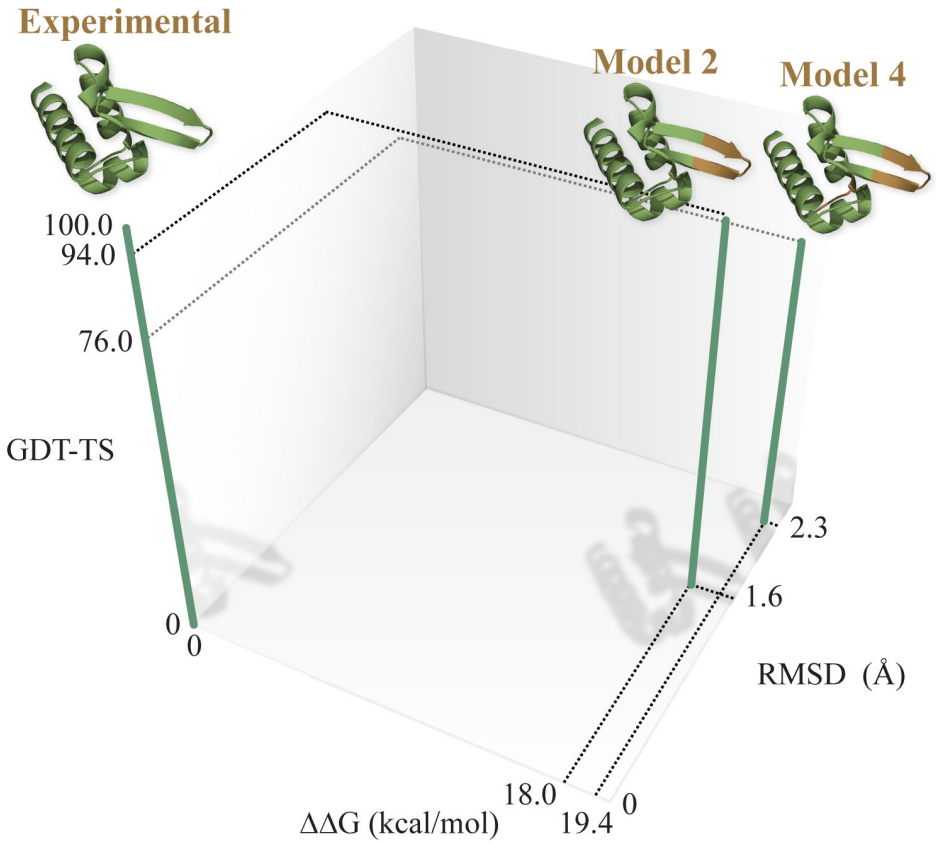
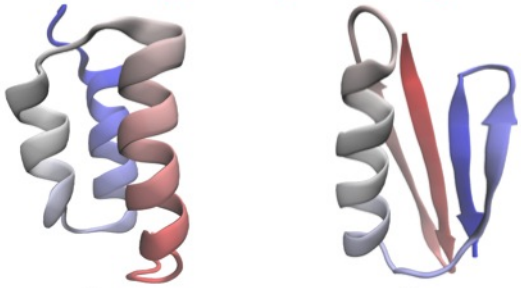


Figure S2: Related to Figure 2: Native structure and two models (from group "Splicer") for CASP target T0560.

```
      12345678901234567890123456789012345678901234567890123456
2jws: TTYKLILNLKQAKEEAIKELVDAGIAEKYIKLIANAKTVEGVWTLKDEILTFTVTE GA88
2jwu: TTYKLILNLKQAKEEAIKELVDAATAEKYFKLYANAKTVEGVWTYKDETKTFTVTE GB88
2kdl: TTYKLILNLKQAKEEAIKELVDAGTAEKYIKLIANAKTVEGVWTLKDEIKTFTVTE GA95
2kdm: TTYKLILNLKQAKEEAIKEAVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFTVTE GB95
```



Sequence	Experimental Fold		Calculated Fold		Calculated ΔG_β-ΔG_α (Kcal/Mol)
	α	β	α	β	
GA88	✔		✔		+3.94± 0.51
GB88		✔		✔	-4.36±0.46
GA95	✔		✔		+3.48±0.47
GB95		✔		✔	-5.01±0.49

Figure S3: Related to Figure 3 and Result subsection "CCR can predict the conformational preferences of 'chameleon' sequences": The CCR method correctly predicts the structural preferences (among two choices, α or β) of four chameleon sequences. Upper: the four sequences used in this study along with the protein data bank identifier. The corresponding pdb id and experimental fold are indicated in the left and right side of the sequences. Middle: the two alternative structures, colored from N (red) to C (blue). Lower: the experimentally observed fold, computationally predicted fold, and conversion free energy between the two folds is reported for each sequence.

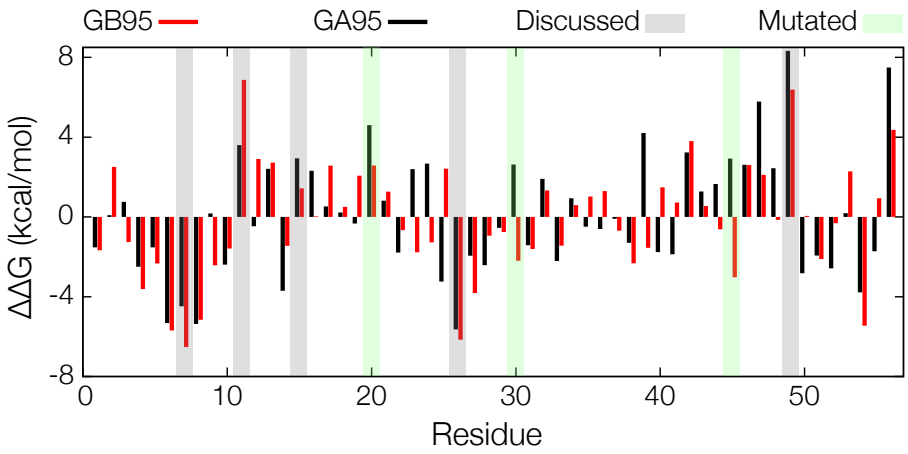


Figure S4: Related to Figure 3 and Figure 4: The per resedue free energy (PRFE) calculation give insights about how amino residues in protein G support β and α fold with GA95 and GB95 sequences. Beside those at the three mutation sites, some other important regions are also highlighted.

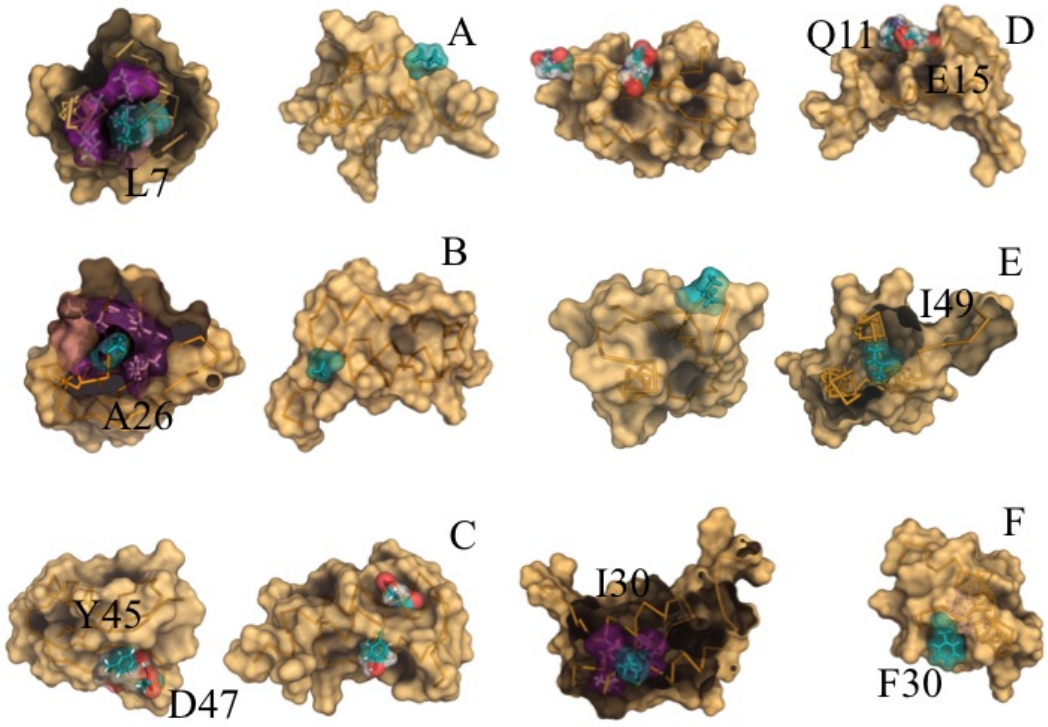


Figure S5: Related to Figure 3: (A): The side chain of L7 and (B): sidechain of A26 are stabilized by hydrophobic contacts inside protein in the β fold (left), and exposed to the solvent in the α fold (right). (C): The Y45 associated with GB95 sequence, forms a H-bond with D47 in the β fold. (D): Residues Q11 and E15 forms a salt bridge in the α fold. (E): Residue I49 is in the hydrophobic core in the α fold (right). (F): The smaller side chain of I30 is in the hydrophobic core with GA95 sequence and α fold. In contrast, F30, having a larger side chain, is exposed to the solvent in the GB95 sequence and α fold. In all the cases the left figure has β fold and right hand figure has α fold except for (F) where both the figure has α fold.

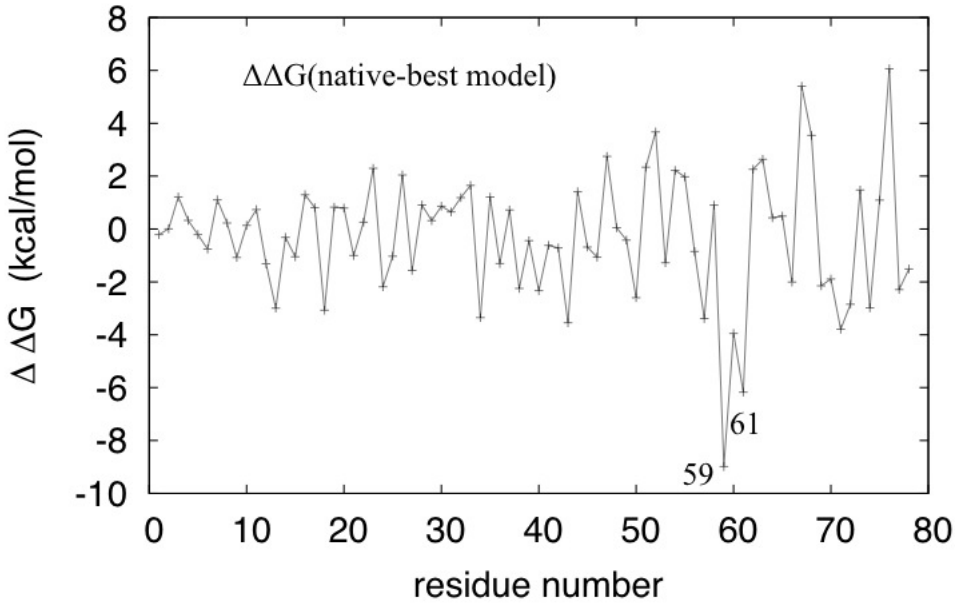


Figure S6: Related to Figure 5: Per-residue free energy between the native NMR structure and the best prediction for the CASP target T0569. Residues at position 59 and 61 stabilize the native structure in a major way.

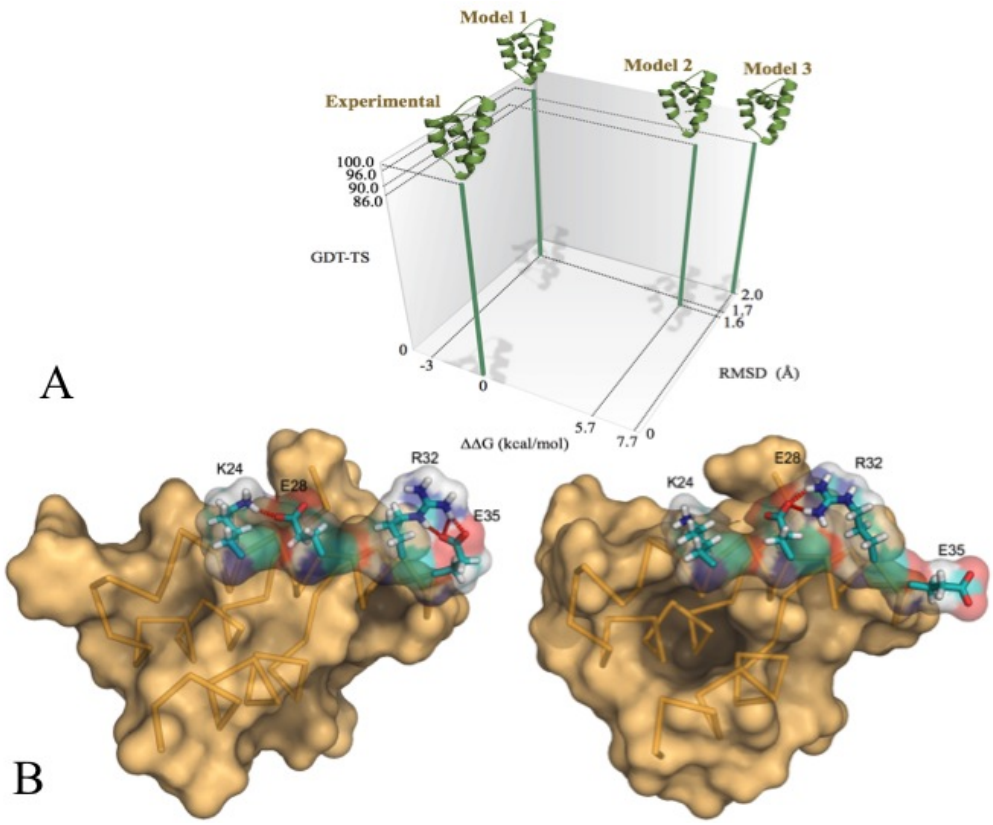


Figure S7: Related to Result subsection "Not all CCR predictions are correct": (A): The free energy-based ranking of the experimental structure and three predicted structures for an engineered protein (PDB ID: 2L09 and CASP code:T0538). Models 1, 2 and 3 are from the group PconsR, Shell and FOLDIT respectively. In this case, one of the predicted structures was found to be more stable than the experimental structure. (B): Comparison of experimental structure and a computer generated model for CASP target T0538 with per residue free energy. Two salt bridges between residues E35-R32 and E28-K24 exist in the native structure. These are absent in the computer generated model structure. Instead, a new salt bridge is formed between residues R32-E28 in model.

Table S1: Related to Figure 2 and Result subsection "CCR can often distinguish CASP-model predictions from true native structures": The CCR method assigns a more favorable free energy to the experimentally determined structure than to computer-generated predictions. For each target, we examined up to five predictions submitted by CASP participants. We report the free energy difference between the most favorable decoy and the experimentally determined structure. Positive $\Delta\Delta G (= \Delta G_{best\ decoy} - \Delta G_{native})$ values indicate that the experimental structure is predicted to be more favorable than any of the decoys.

CASP Target	PDB Identifier	$\Delta\Delta G$ (kcal/mol)	GDT-TS
T0531	2KJX	11.15 ± 0.70	44
T0538	2L09	-3.00 ± 0.47	96
T0540	3MX7	16.94 ± 0.49	70
T0559	2L01	2.10 ± 0.24	94
T0560	2L02	18.00 ± 0.49	94
T0569	2KYW	20.01 ± 0.69	78

Table S2: Related to Figure 2 and Result subsection "CCR can correctly rank-order the CASP models submitted from a given prediction team" : The six different CASP targets that are used in this study. The corresponding PDB Identifier and description are also listed.

CASP Target	PDB Identifier	Description
T0531	2KJX	extracellular domain of the jumping translocation breakpoint protein
T0538	2L09	protein asr4154 from Nostoc sp. PCC7120
T0540	3MX7	human Fas apoptotic inhibitory molecule
T0559	2L01	protein BVU3908 from Bacteroides vulgatus
T0560	2L02	protein BT2368 from Bacteroides thetaiotaomicron
T0569	2KYW	domain of adhesion exoprotein from Pediococcus pentosaceus