Cost Function Weights

John G. Archie, Martin Paluszewski, and Kevin Karplus

May 7, 2009

1 Cost Function Weights

To compute the predicted quality, we computed the linear sum of component cost functions. The combined cost, x below, was then converted to a predicted quality score by a sigmoidal function:

$$q(x) = \frac{1}{1 + e^{ax+b}}$$

where the constants a and b were fitted to the CASP7 training data.

Additionally, not all cost functions are on the same scale. For this reason, we report the pooled standard deviation (P. SD), which measures the pooled standard deviation of the cost function term times the weight for the training data. The pooled standard deviation is a better indicator of the relative importance of different cost function terms.

For SAM-T08-MQAU (Undertaker) and SAM-T08-MQAC (Undertaker+constraints), we used separate cost functions for easy and difficult targets. Difficulty was assessed by the e-value of the best template alignment.

1.1 MQAO/Alignment Constraints

For the SAM-T08-MQAO group, we used the following component cost functions. For quality prediction, the fitted sigmoidal function used constants of $a = 0.0384$ and $b = 0.901$.

Cost Function	Weight	P. SD	Description
align_bonus	55.527	9.542	selected alignment predicted constraints
rejected_bonus	42.886	3.358	rejected alignment predicted constraints
align_constraints	1.586	1.009	selected alignment predicted constraints

1.2 MQAU/Undertaker: Low E-value Targets

For the easy targets (best e-values less than 0.31687) in the SAM-T08-MQAU group, we used the following component cost functions. For quality prediction, the fitted sigmoidal function used constants of $a = 0.0175$ and $b = -0.0465$.

Cost Function	Weight	P. SD	Description
align_sheets	6.524	24.104	sheet constraints from alignments
align_constraints	32.082	22.036	selected alignment predicted constraints
pred_nb11_back	25.177	5.788	neural net predicted burial, near-backbone-11 alphabet
missing_atoms	0.012	5.233	the number of missing atoms in the model
pred_alpha_back	10.614	3.845	neural net predicted alpha torsion angle
rejected_bonus	19.142	1.662	rejected alignment predicted constraints
phobic_fit	0.218	1.184	hydrophobic radius of gyration[1]
sidechain_clashes	2.723	0.871	number of severe sidechain clashes
contact	0.602	0.773	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
pred_cb14_back	2.902	0.768	neural net predicted burial, C_β-14 alphabet

1.3 MQAU/Undertaker: High E-value Targets

For the hard targets (best e-values greater than 0.31687) in the SAM-T08-MQAU group, we used the following component cost functions. For quality prediction, the fitted sigmoidal function used constants of $a = 0.0345$ and $b = 2.10$.

Cost Function	Weight	P. SD	Description
pred_alpha_back	14.199	6.338	neural net predicted alpha torsion angle
CBContact	1.321	2.441	average number of contacts (C_β atoms within 8 Å)
nn1000	20.236	2.338	neural net predicted residue-residue distance constraints
pred_cb14_back	9.255	1.978	neural net predicted burial, C_β-14 alphabet
pred_nb11_back	7.111	1.652	neural net predicted burial, near-backbone-11 alphabet
noncontact_bonus	0.487	1.373	alignment predicted noncontacts
ehl2_constraints	0.155	1.253	neural net predicted secondary structure constraints
rejected_bonus	34.923	1.132	rejected alignment predicted constraints
hbond_geom	5.817	0.828	measure of good hydrogen bond distance and geometry
knot	3.356	0.750	detects knotted proteins
bystroff	3.133	0.577	propensity predicted Bystroff alphabet
sidechain	0.000	0.469	the negative log-probability of observing the sidechain and backbone conformation

1.4 MQAC/Undertaker+Consensus: Low E-value Targets

For the easy targets (best e-values less than 6.9768e-15) in the SAM-T08-MQAC group, we used the following component cost functions. For quality prediction, the fitted sigmoidal function used constants of $a = 0.0455$ and $b = 2.35$.

Cost Function	Weight	P. SD	Description
sim_TM	42.200	7.970	median TMscore consensus measure
sim_GDT_TS	34.939	6.760	median GDT_TS consensus measure
pred_nb11_back	6.748	1.510	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	3.663	1.231	neural net predicted alpha torsion angle
pred_CB8_sep9_back	3.864	0.810	neural net predicted burial, CB8_sep9 alphabet
pred_cb14_back	2.500	0.659	neural net predicted burial, C_β-14 alphabet
CAContact	0.455	0.574	average number of contacts (C_α atoms within 8 Å)
cb14	2.046	0.553	propensity predicted burial, C_β-14 definition
contact_order	1.571	0.254	average chain separation of contacting residues
alpha_prev	0.619	0.167	propensity predicted alpha angle of the previous residue
pred_o_sep_back	0.754	0.120	predicted H-bond sequence separation for O
pred_n_sep_back	0.635	0.096	predicted H-bond sequence separation for N

1.5 MQAC/Undertaker+Consensus: High E-value Targets

For the hard targets (best e-values greater than 6.9768e-15) in the SAM-T08-MQAC group, we used the following component cost functions. For quality prediction, the fitted sigmoidal function used constants of $a = 0.0406$ and $b = 1.79$.

Cost Function	Weight	P. SD	Description
sim_GDT_TS	62.884	6.959	median GDT_TS consensus measure
sim_TM	14.873	1.736	median TMscore consensus measure
align_bonus	12.702	1.697	selected alignment predicted constraints
pred_alpha_back	2.720	1.155	neural net predicted alpha torsion angle
pred_nb11_back	4.576	1.084	neural net predicted burial, near-backbone-11 alphabet
align_sheets	1.145	0.894	sheet constraints from alignments
rejected_constraints	0.767	0.279	rejected alignment predicted constraints
sidechain	0.000	0.134	the negative log-probability of observing the sidechain and backbone conformation
knot	0.305	0.078	detects knotted proteins
noncontact	0.024	0.073	alignment predicted noncontacts

References

1. Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. The design and characterization of two proteins with 88% sequence identity but different structure and function. Proc Natl Acad Sci USA 104(29):11963–11968, July, 2007.

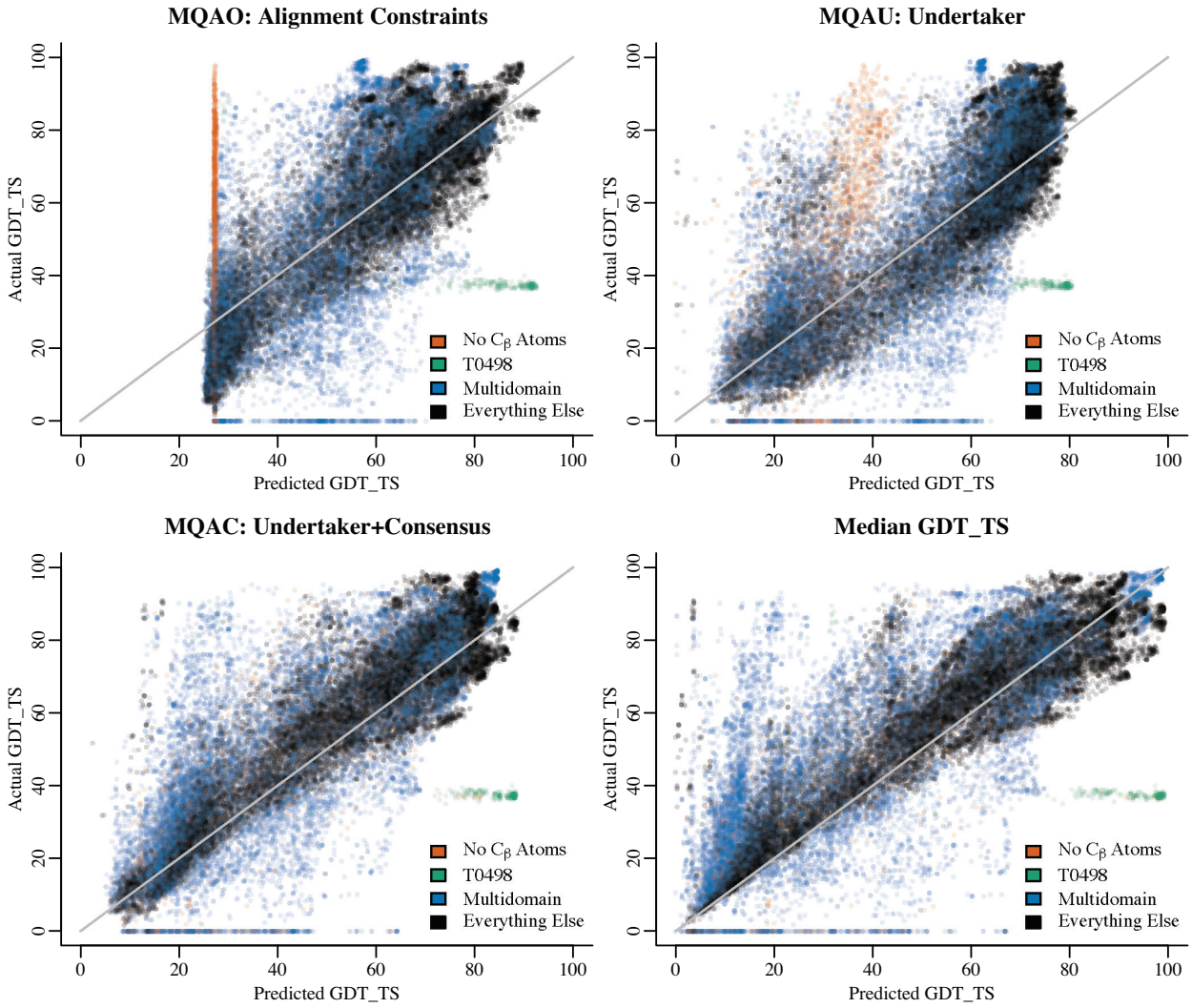

Figure 1: Predicted and actual GDT_TS scores for all target domains and all servers. Three of the plots are for submissions by groups SAM-T08-MQAO, SAM-T08-MQAU, and SAM-T08-MQAC. Median GDT_TS is a pure consensus term, which we did not submit to CASP8. T0498 and T0499 were engineered proteins, differing in only three amino acids but assuming different folds. Both structures are related to those previously described by Alexander et al.[1]

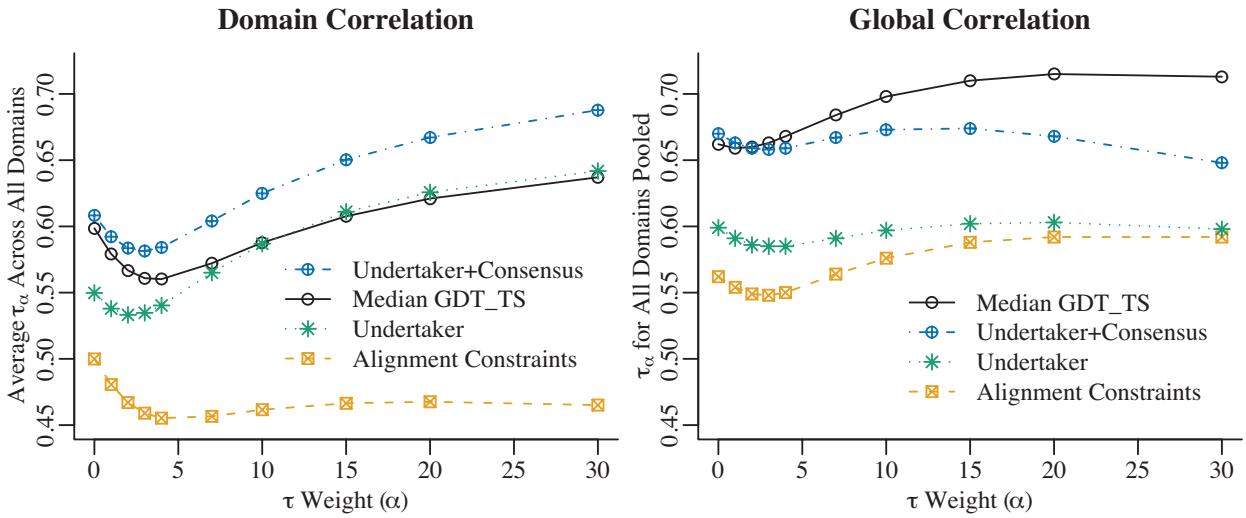

Figure 2: Correlation values for each function. Increasing values of α place increasing weight on the predicted-best set of models. An α value of 0 is equivalent to Kendall's τ, treating all models equally. Values of 0, 3, 5, 15, and 30 place half of the weight on the top 50%, 23%, 14%, 5%, and 2.3% of models. The domain correlation plot shows the average τ value over all target domains, with correlation computed separately for each target domain. The global correlation plot shows the τ value computed from combining all predictions into a single set. Adding Undertaker cost function terms to the consensus median GDT_TS method improved the ranking of models within a target, particularly when concentrating on the top-scoring models. Median GDT_TS alone is a better predictor of raw GDT_TS value, especially for picking out the easy targets, but does not do as well at ranking models for a given target. "Alignment Constraints" is SAM-T08-MQAO, "Undertaker" is SAM-T08-MQAU, "Undertaker+Consensus" is SAM-T08-MQAC, and "Median GDT_TS" is the pure consensus term, which we did not submit to CASP8.

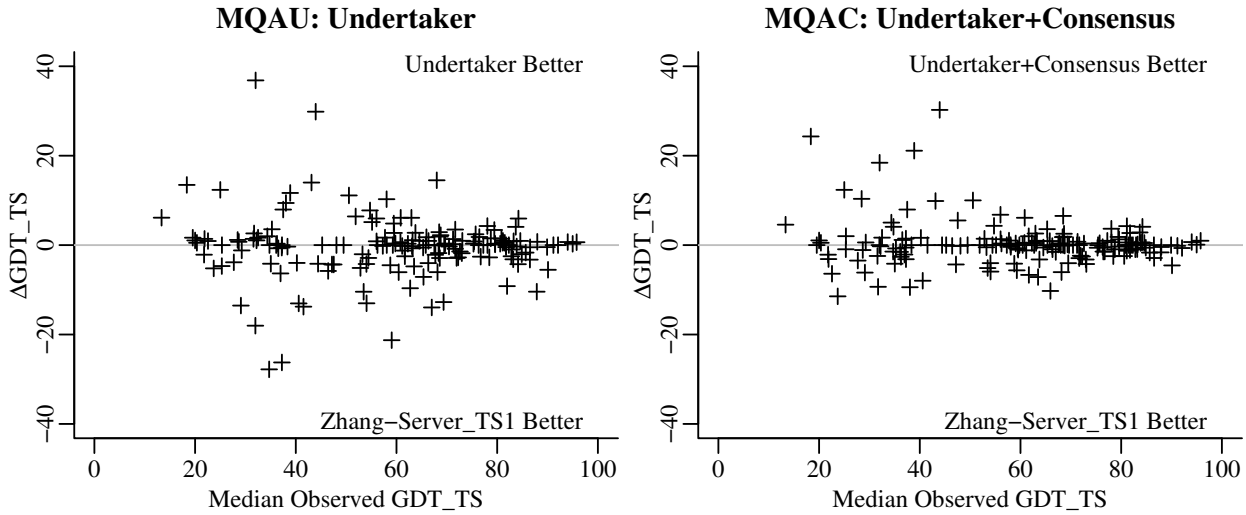

Figure 3: Undertaker (SAM-T08-MQAU) and Undertaker+Consensus (SAM-T08-MQAC) functions as metaservers compared to the best single server in the pool. The median observed GDT_TS score of all server models is used as a proxy for target difficulty. Using MQAC does slightly better than the Zhang server, but without the consensus term, the difference between the metaserver and Zhang server is quite small. Neither difference is statistically significant.