Applying Undertaker Cost Functions to Model Quality Assessment (Supplementary Materials)

John Archie and Kevin Karplus

August 20, 2008

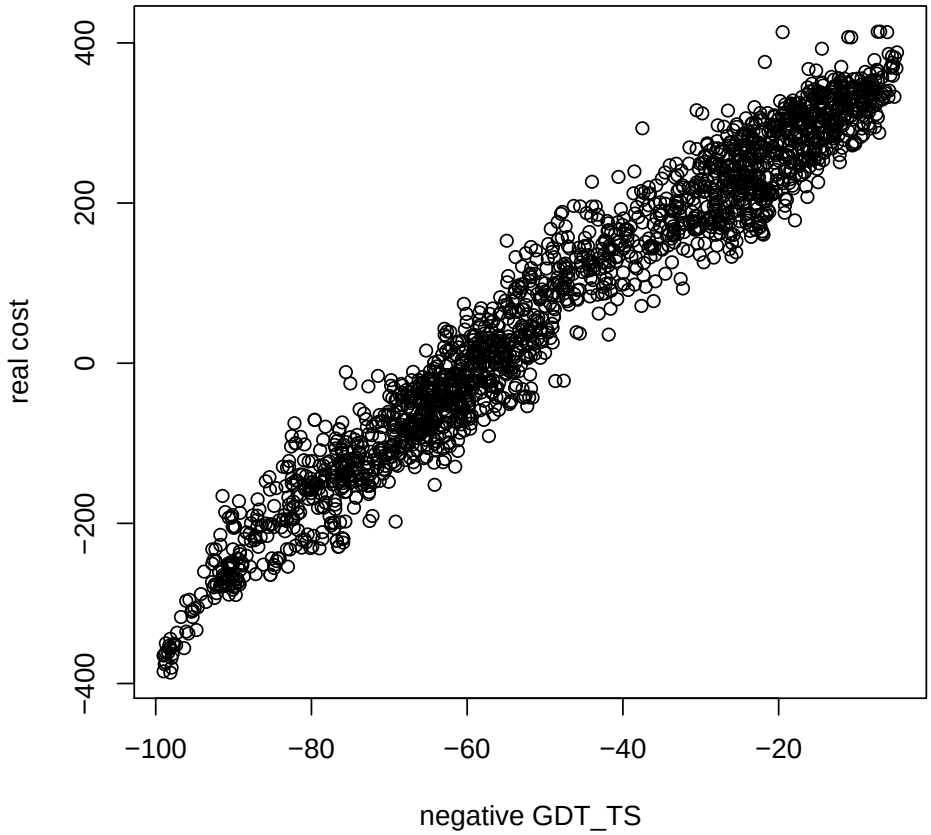


Figure 1: GDT_TS versus real cost. GDT_TS and real cost both measure model quality. After scoring all server models for all CASP7 targets with structures in the PDB, 2,000 full models were randomly selected and plotted above. As expected, there is a strong correlation between the two quality measures; good models tend to be good regardless of the scoring measure used.

(a)

Group	$\bar{r}$	$\bar{\rho}$	$\overline{\text{GDT}}$	$\bar{\tau}_0$	$\bar{\tau}_3$
under+TM	**0.90**	**0.85**	**60.4**	**0.70**	**0.68**
under	0.76	0.76	*59.8*	0.58	*0.59*
TASSER	0.63	0.69	59.7	0.54	0.51
Qiu	*0.85*	0.75	59.0	0.58	0.54
Pcons	0.82	0.75	58.3	0.56	0.52
LEE	0.82	*0.78*	57.8	*0.63*	*0.59*
ModFOLD	0.66	0.55	55.4	0.40	0.37

(b)

Group	$\bar{r}$	$\bar{\rho}$	$\overline{\text{GDT}}$	$\bar{\tau}_0$	$\bar{\tau}_3$
under+TM	**0.90**	**0.83**	**60.7**	**0.68**	**0.65**
TASSER	0.63	0.67	*59.8*	0.53	0.50
under	*0.86*	*0.77*	59.6	*0.61*	*0.57*
Qiu	0.85	0.74	59.3	0.57	0.53
LEE	0.80	0.72	58.1	0.58	0.52
Pcons	0.85	0.75	57.6	0.56	0.51
ModFOLD	0.70	0.62	55.8	0.46	0.43

(c)

Group	$\bar{r}$	$\bar{\rho}$	$\overline{\text{RC}}$	$\bar{\tau}_0$	$\bar{\tau}_3$
under+TM	**0.93**	**0.87**	**39.0**	**0.72**	**0.70**
under	*0.91*	*0.84*	*32.2*	*0.68*	*0.65*
TASSER	0.69	0.73	26.1	0.58	0.55
Qiu	0.85	0.75	25.4	0.59	0.56
LEE	0.80	0.72	10.1	0.57	0.52
Pcons	0.84	0.75	1.7	0.56	0.51
ModFOLD	0.76	0.68	4.9	0.52	0.49

Table 1: Performance of different model quality assessment. These tables show how each MQA method performs when evaluated against the CASP7 server data set which excludes Zhang Server models. "under" denotes the Undertaker cost functions only; "under+TM," the Undertaker cost functions with the median TM-score consensus term; and Qiu, data from a scoring function including the median TM-score consensus term, an atom-pairwise distance potential, and Rosetta terms (1). TASSER, LEE, Pcons, and ModFOLD indicate CASP7 groups 125, 556, 634, and 704. The correlation measures are against negative GDT_TS (a,b) or real cost (c) and are averaged over 89 CASP7 targets. The metrics are Pearson's r, Spearman's ρ, average quality of predicted best model (GDT_TS denotes GDT_TS; RC, real cost), Kendall's τ, and τ_3. Training and evaluation was done using five-fold cross-validation on all models and GDT_TS (a), complete models and GDT_TS (b); and complete models and real cost (c). Tables are sorted by the average quality of the best model. The largest value in each column is presented in bold; the second largest, italics.

The clens Contact Quality Measure

The clens quality measure was written by Tim Dreszer to assess how well the residue-residue contacts in a model match contacts observed in an experimentally determined structure. Only residue pairs in contact in the model or experimental structure are considered. When computing clens, contacts present in both structures (true positives) are counted, while contacts present in the model only (false positives) or in the experimental structure only (false negatives) are penalized. Contacts absent in both structures (true negatives) are ignored.

More precisely, contact distance is determined by measuring the distance between residue center spots (Figure 3). The distance between two residues in the model is denoted $\mathrm{D}_{\mathrm{predicted}}$, and in the experimental structure, $\mathrm{D}_{\mathrm{observed}}$. A contact score is determined by summing over all nonadjacent pairs of residues where $\mathrm{D}_{\mathrm{predicted}} \leq 9$ Å, $\mathrm{D}_{\mathrm{observed}} \leq 9$ Å, or $\mathrm{D}_{\mathrm{predicted}} + \mathrm{D}_{\mathrm{observed}} \leq 21$ Å,

$$s = \sum \frac{w(\mathrm{D}_{\mathrm{observed}} - \mathrm{D}_{\mathrm{predicted}})}{\mathrm{P}(\mathrm{contact}|\mathrm{sep})} \tag{1}$$

The weighting function for a contact (Figure 2) is defined as

$$w(x) = e^{-\left(\frac{x}{3}\right)^2} \tag{2}$$

and the probability of a contact given the sequence separation, P(contact|sep), is listed in Table 4 and is estimated from a thinned version of the PDB from Dunbrack's Pisces server[1] (2).

Finally, the contact score is converted to a cost by negating and normalizing to the range [0, 1]:

$$\mathrm{cost} = 1 - \frac{s}{s_{\max}} \tag{3}$$

where

$$s_{\max} = \sum \mathrm{P}(\mathrm{contact}|\mathrm{sep})^{-1} \tag{4}$$

The GDT_TS and Smooth GDT Quality Measures

To compute GDT_TS (3) and smooth GDT, Undertaker first computes an array ($\boldsymbol{d}$) by sampling superpositions of the model and experimental structure. A value d_i ($1 \leq i \leq N$ where N is the number of residues in the protein) indicates that a superposition exists such that i C_α points in the model are within d_i Å of the corresponding C_α points in the experimental structure. As Undertaker samples superpositions, $\boldsymbol{d}$ is updated to store the minimum encountered distance for each value of i.

After sampling superpositions, GDT_TS can then be computed as

$$\mathrm{GDT\_TS} = \frac{1}{4N} \sum_{t \in 1,2,4,8} \max(i \text{ where } d_i \leq t) \tag{5}$$

[1]The set was generated on April 5, 2005 and included structures 1.8 Å or better with an R-factor of less than or equal to 0.25, thinned to 40% sequence identity.

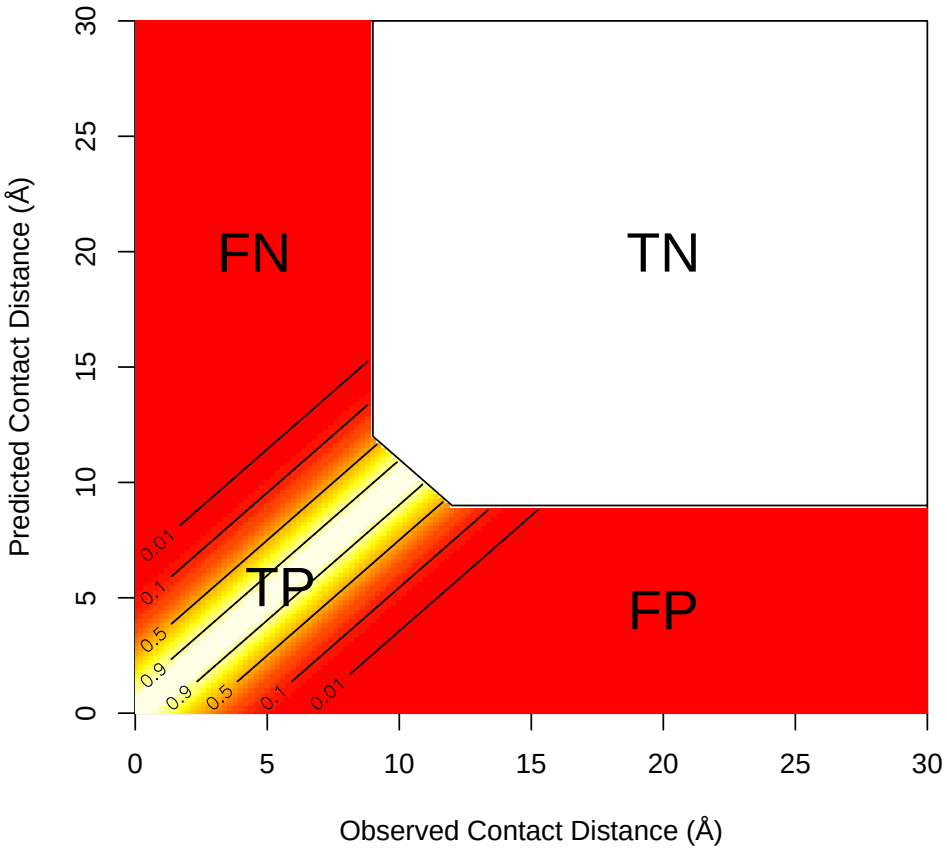


Figure 2: Individual residue pair contact score. Ideal true positives (TP) are given a score of 1, while false positives (FP) and false negatives (FN) are given a score of (practically) 0. True negatives (TN) are not considered when computing s or $s_{\max}$ (Equations 1 and 3).

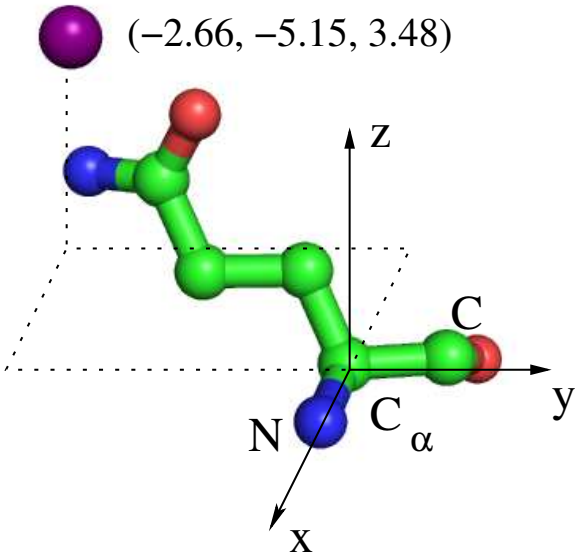


Figure 3: Residue center spot. As shown above, a coordinate system can be defined by placing the C_α at the origin, the amide N on the x-axis, and the carbonyl C on the $z = 0$ plane. For the purposes of determining residue-residue contact, a residue is defined as being located at (-2.66, -5.15, 3.48) in this coordinate system.

Alternatively, GDT_TS (and smooth GDT) can be expressed as

$$\text{GDT} = \frac{1}{N}\sum_{i=1}^{N} v(d_i) \tag{6}$$

where, in the case of GDT_TS, the value function is

$$v(d) = \begin{cases} 1 & d \leq 1 \\ 3/4 & 1 < d \leq 2 \\ 1/2 & 2 < d \leq 4 \\ 1/4 & 4 < d \leq 8 \\ 0 & 8 < d \end{cases} \tag{7}$$

Smooth GDT simply uses an alternative definition for the value function

$$v(d) = \begin{cases} 1 & \text{if } d < 3\sqrt{2}/8 \\ -\frac{1}{9}\log_2 \frac{d^2}{12^2} & \text{if } 3\sqrt{2}/8 \leq d \leq 12 \\ 0 & \text{if } 12 < d \end{cases} \tag{8}$$

These two functions are compared in Figure 4. Since smooth GDT uses a continuous range of thresholds from about 1/2 Å to 12 Å, the measure does a slightly better job when computing a score for high and low accuracy models.

The idea behind smooth GDT is very similar to TM-score (4). However, TM-score only uses only one superposition when compting $\boldsymbol{d}$—the superposition that maximizes the TM-score. Accepting that $\boldsymbol{d}$ has a slightly different meaning in the context of TM-score, the TM-score value function is defined as

$$v(d) = \frac{1}{1 + \frac{d}{c}} \tag{9}$$

where c is a constant defined in terms of the protein length.

H-bond Scaled Likelihoods

For modeling H-bond geometry, H-bonds were extracted from a thinned version of the PDB from Dunbrack's Pisces server[2] (2). H-bonds were included in this set if they met a set of geometric criteria (Table 5 and Figures 5–9). These criteria were subjectively established by visual inspection of the distributions of geometric features.

Five geometric features were identified and modeled with statistical distributions. For each feature, different models were fitted to different classes of H-bonds. For example, backbone H-bonds are those that include both the amide N and carbonyl O; backbone H-bonds with a separation (the index of the donor minus the index of the acceptor) of 3 (the

[2]The set was generated on August 15, 2003 and included structures 1.8 Å or better with an R-factor of less than or equal to 0.25, thinned to 30% sequence identity. Three structures were removed: 1nxb had 106 clashes in 62 residues; 1en2A and 1ejgA both contained microhetrogeneity.

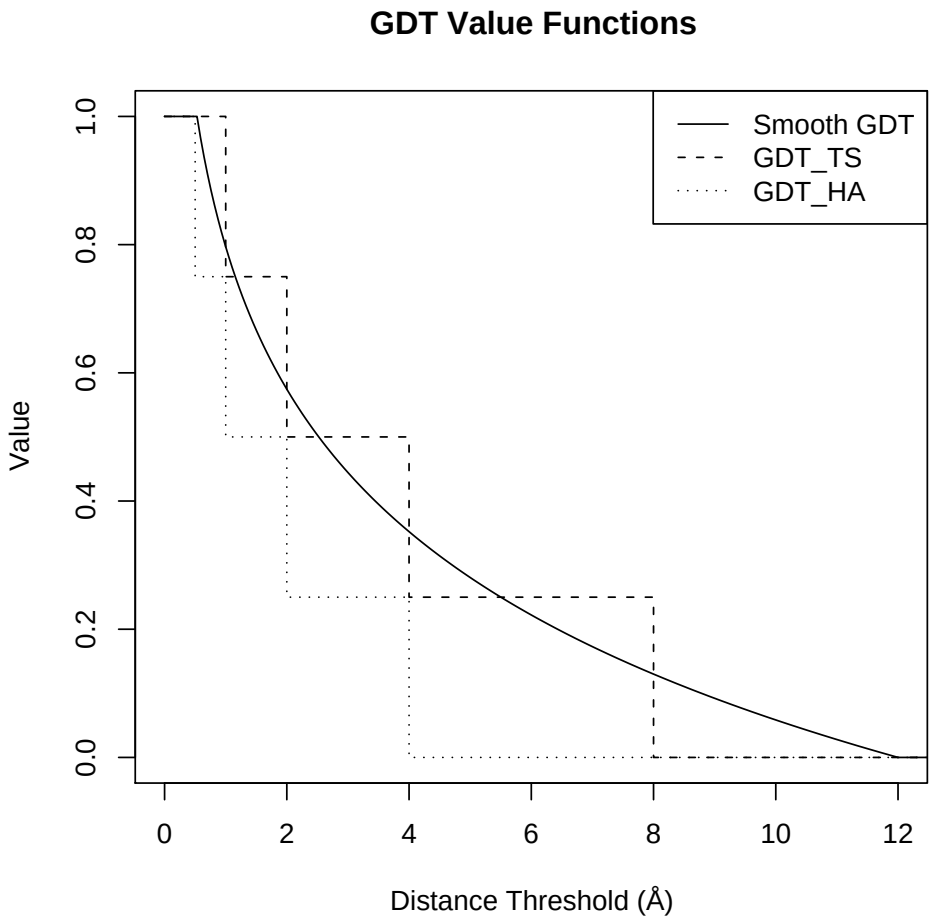


Figure 4: Smooth GDT, GDT_TS, and GDT_HA (5) value functions. All of these measures can be computed as a normalized sum of value functions (Equation 6). The value functions used for GDT and smooth GDT are plotted for comparison.

3_{10}-helix H-bond) and 4 (the α-helix H-bond) often followed different distributions. Other characteristics were also used to separate types of H-bonds where appropriate.

The scaled likelihood of an H-bond given the geometry is computed as the product of the scaled likelihoods of

- the donor-acceptor distance,
- the DAC (donor-acceptor-carbon) angle (Figure 6),
- asymmetry divided by the donor-acceptor distance (Figure 7), and
- nonplanarity divided by the donor-acceptor distance (Figure 8) or, if nonplanarity cannot be computed, the CDA (carbon-donor-acceptor) angle (Figure 5).

Each density function was scaled to have a maximum value of 1, placing each density function on a similar scale and ensuring that the product of the scaled density functions is in the range [0, 1]. These four features are almost statistically independent (data not shown).

For modeling distance, the Leonard-Jones 6-12 potential was used to model the energy of an H-bond,

$$l(x) = 2\left(\frac{r}{x}\right)^6 - \left(\frac{r}{x}\right)^{12} \tag{10}$$

These energy estimates fitted fairly well to an exponential function,

$$f(x) = e^{kl(x)+c} \tag{11}$$

where r, k, and c were fitted constants. The scaled exponential density function is then

$$f(x) = e^{kl(x)-k} \tag{12}$$

where values for r and k are listed in Table 5.

Each of the geometric features (the CDA and DAC angles, asymmetry, and nonplanarity) were fitted to normal distributions (Figures 5–8). The scaled normal density function is then expressed as

$$f(x) = e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{13}$$

For backbone H-bonds with separation of four, nonplanarity is modeled by a uniform because the distribution is too flat. When a nonbackbone H-bond has the backbone O as an acceptor, asymmetry is modeled by a uniform distribution. Whenever a measurement cannot be computed for any reason, the uniform distribution is also used as a way of ignoring that feature.

Alphabets for Quality Assessment

Burial is defined as a nonnegative integer and can be computed for a residue by counting the number of other residues that are in close proximity in a 3D coordinate system. For near-backbone-11, each residue has the same center spot (-2.66, -5.15, 3.48) as clens (Figure 3) and a count spot (1.24, 0.64, 0.23) near the backbone N. For a given residue, burial is

letter	burial range
A	$0 \leq x < 4$
B	$4 \leq x < 6$
C	$6 \leq x < 7$
D	$7 \leq x < 10$
E	$10 \leq x < 13$
F	$13 \leq x < 16$
G	$16 \leq x < 20$
H	$20 \leq x < 23$
I	$23 \leq x < 26$
J	$26 \leq x < 29$
K	$29 \leq x < \infty$

Table 2: Near-backbone-11 alphabet. Near-backbone-11 describes the degree of burial from A (least buried) to K (most buried). Burial is computed by counting the number of other residues that are in close proximity in a 3D coordinate system. For near-backbone-11, each residue has the same center spot (-2.66, -5.15, 3.48) as clens (Figure 3) and a count spot (1.24, 0.64, 0.23) near the backbone N. For a given residue, burial is defined as the number of count spots within a 9.65 Å radius of that residue's center spot.

defined as the number of count spots within a 9.65 Å radius of that residue's center spot. The near-backbone-11 alphabet is defined in Table 2.

The n_notor alphabet uses the NOtor torsion angle (Figure 9) and other H-bond properties to place the backbone N to place residues into categories; the alphabet is defined in Table 3.

Optimization of τ_3 and GDT on Complete Models

When doing the five-fold cross validation, below are the weight sets defined when using full models and GDT as a measure of model quality.

letter	NOtor range	meaning
G		separation 3 H-bonds (3_{10}-helix)
H		separation 4 H-bonds (α-helix)
P	$x < -133$ or $x > 76$	often parallel β-strand
B	$-133 \leq x \leq -17$	often antiparallel β-strand
A	$-17 < x \leq 76$	often antiparallel β-strand
S		H-bond to sidechain
M		multiple H-bonds
N		no H-bond

Table 3: N_notor alphabet. The n_notor alphabet places residues into three categories defined by the NOtor torsion angle (Figure 9) of the backbone N (P, B, and A). There are special cases for helical (G and H) H-bonds, as well as for when the backbone N does not form a single NOtor angle (S, M, and N).

Cost Function	Pooled SD	Description
align_constraint	3.278	selected alignment predicted constraints
pred_nb11_back	0.961	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	0.825	neural net predicted alpha torsion angle
noncontacts_bonus	0.771	alignment predicted noncontacts
align_bonus	0.318	selected alignment predicted constraints
dry5	0.182	propensity predicted burial, dry-5 definition
pred_o_sep_back	0.159	predicted H-bond sequence separation for O
rejected_bonus	0.147	rejected alignment predicted constraints
near_backbone	0.140	propensity predicted burial, near-backbone-11 definition
pred_cb14_back	0.139	nerual net predicted burial, C_β-14 alphabet
pred_n_sep_back	0.138	predicted H-bond sequence separation for N
ehl2_constraint	0.095	secondary structure constraints
contact	0.083	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.024	detects missing backbone atoms or chainbreaks

Cost Function	Pooled SD	Description
align_constraint	2.128	selected alignment predicted constraints
rejected_constraint	0.871	rejected alignment predicted constraints
noncontacts_bonus	0.792	alignment predicted noncontacts
pred_alpha_back	0.774	neural net predicted alpha torsion angle
pred_nb11_back	0.715	neural net predicted burial, near-backbone-11 alphabet
align_bonus	0.451	selected alignment predicted constraints
near_backbone	0.240	propensity predicted burial, near-backbone-11 definition
dry5	0.172	propensity predicted burial, dry-5 definition
pred_o_sep_back	0.115	predicted H-bond sequence separation for O
contact_order	0.101	average chain separation of contacting residues
contact	0.059	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue

Cost Function	Pooled SD	Description
align_constraint	5.286	selected alignment predicted constraints
pred_nb11_back	2.253	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	1.504	neural net predicted alpha torsion angle
align_bonus	1.078	selected alignment predicted constraints
noncontacts	0.854	alignment predicted noncontacts
ehl2_constraint	0.327	secondary structure constraints
rejected_bonus	0.288	rejected alignment predicted constraints
contact_order	0.248	average chain separation of contacting residues
pred_o_sep_back	0.247	predicted H-bond sequence separation for O
dry5	0.242	propensity predicted burial, dry-5 definition
pred_n_sep_back	0.223	predicted H-bond sequence separation for N
near_backbone	0.223	propensity predicted burial, near-backbone-11 definition
contact	0.142	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.040	detects missing backbone atoms or chainbreaks

Cost Function	Pooled SD	Description
align_constraint	5.519	selected alignment predicted constraints
pred_nb11_back	2.044	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	1.599	neural net predicted alpha torsion angle
noncontacts	0.922	alignment predicted noncontacts
dry12	0.526	propensity predicted burial, dry-12 definition
rejected_bonus	0.476	rejected alignment predicted constraints
ehl2_constraint	0.347	secondary structure constraints
pred_o_sep_back	0.245	predicted H-bond sequence separation for O
sidechain_clashes	0.193	number of severe sidechain clashes
contact_order	0.163	average chain separation of contacting residues
contact	0.079	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.060	detects missing backbone atoms or chainbreaks

Cost Function	Pooled SD	Description
align_constraint	5.636	selected alignment predicted constraints
pred_nb11_back	1.946	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	1.592	neural net predicted alpha torsion angle
noncontacts	0.961	alignment predicted noncontacts
align_bonus	0.672	selected alignment predicted constraints
dry5	0.253	propensity predicted burial, dry-5 definition
rejected_bonus	0.246	rejected alignment predicted constraints
pred_n_sep_back	0.244	predicted H-bond sequence separation for N
near_backbone	0.211	propensity predicted burial, near-backbone-11 definition
ehl2_constraint	0.186	secondary structure constraints
contact	0.171	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.030	detects missing backbone atoms or chainbreaks

Optimization of τ_3 and Real Cost on Complete Models

When doing the five-fold cross validation, below are the weight sets defined when using full models and real cost as a measure of model quality.

Cost Function	Pooled SD	Description
align_constraint	1.323	selected alignment predicted constraints
pred_nb11_back	1.061	neural net predicted burial, near-backbone-11 alphabet
align_bonus	1.014	selected alignment predicted constraints
pred_n_notor_back	0.866	neural net predicted H-bond properties, including NOtor torsion angle
pred_alpha_back	0.776	neural net predicted alpha torsion angle
noncontacts_bonus	0.733	alignment predicted noncontacts
contact	0.176	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
rejected_bonus	0.144	rejected alignment predicted constraints
near_backbone	0.130	propensity predicted burial, near-backbone-11 definition
pred_n_sep_back	0.122	predicted H-bond sequence separation for N
dry5	0.116	propensity predicted burial, dry-5 definition
sidechain	0.112	the negative log-probability of observing the sidechain and backbone conformation
ehl2_constraint	0.110	secondary structure constraints
pred_o_sep_back	0.093	predicted H-bond sequence separation for O
is_align	0.051	detects missing backbone atoms or chainbreaks
hbond_dist	0.048	H-bond distance cost function, using LJ 6-12 potential

Cost Function	Pooled SD	Description
align_constraint	1.155	selected alignment predicted constraints
pred_n_notor_back	0.789	neural net predicted H-bond properties, including NOtor torsion angle
noncontacts_bonus	0.758	alignment predicted noncontacts
pred_nb11_back	0.725	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	0.715	neural net predicted alpha torsion angle
align_bonus	0.603	selected alignment predicted constraints
rejected_constraint	0.239	rejected alignment predicted constraints
rejected_bonus	0.218	rejected alignment predicted constraints
near_backbone	0.218	propensity predicted burial, near-backbone-11 definition
sidechain	0.155	the negative log-probability of observing the sidechain and backbone conformation
contact	0.138	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
cb14	0.101	propensity predicted burial, C_β-14 definition
bystroff	0.061	propensity predicted Bystroff alphabet
hbond_dist	0.048	H-bond distance cost function, using LJ 6-12 potential
contact_order	0.046	average chain separation of contacting residues
is_align	0.044	detects missing backbone atoms or chainbreaks
way_back	0.028	propensity predicted burial, way-back

Cost Function	Pooled SD	Description
align_constraint	2.387	selected alignment predicted constraints
pred_nb11_back	1.078	neural net predicted burial, near-backbone-11 alphabet
pred_n_notor_back	0.908	neural net predicted H-bond properties, including NOtor torsion angle
pred_alpha_back	0.817	neural net predicted alpha torsion angle
noncontacts_bonus	0.776	alignment predicted noncontacts
align_bonus	0.490	selected alignment predicted constraints
rejected_bonus	0.245	rejected alignment predicted constraints
sidechain	0.239	the negative log-probability of observing the sidechain and backbone conformation
near_backbone	0.192	propensity predicted burial, near-backbone-11 definition
ehl2_constraint	0.173	secondary structure constraints
contact_order	0.069	average chain separation of contacting residues
contact	0.067	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
hbond_geom_backbone	0.065	negative log-likelihood of backbone H-bonds
pred_o_sep_back	0.047	predicted H-bond sequence separation for O
is_align	0.044	detects missing backbone atoms or chainbreaks

Cost Function	Pooled SD	Description
align_constraint	1.998	selected alignment predicted constraints
pred_nb11_back	1.292	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	0.981	neural net predicted alpha torsion angle
pred_n_notor_back	0.755	neural net predicted H-bond properties, including NOtor torsion angle
noncontacts_bonus	0.733	alignment predicted noncontacts
align_bonus	0.457	selected alignment predicted constraints
rejected_bonus	0.334	rejected alignment predicted constraints
sidechain	0.239	the negative log-probability of observing the sidechain and backbone conformation
pred_bys_back	0.162	neural net predicted bystroff alphabet
cb14	0.137	propensity predicted burial, C_β-14 definition
alpha	0.127	propensity predicted alpha angle
hbond_dist	0.100	H-bond distance cost function, using LJ 6-12 potential
dry5	0.074	propensity predicted burial, dry-5 definition
contact	0.072	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.048	detects missing backbone atoms or chainbreaks

Cost Function	Pooled SD	Description
align_constraint	1.568	selected alignment predicted constraints
pred_nb11_back	1.093	neural net predicted burial, near-backbone-11 alphabet
pred_alpha_back	0.914	neural net predicted alpha torsion angle
pred_n_notor_back	0.770	neural net predicted H-bond properties, including NOtor torsion angle
align_bonus	0.757	selected alignment predicted constraints
noncontacts_bonus	0.725	alignment predicted noncontacts
nn700_constraint	0.188	neural net predicted contact constraints
sidechain	0.181	the negative log-probability of observing the sidechain and backbone conformation
pred_bys_back	0.179	neural net predicted bystroff alphabet
near_backbone	0.159	propensity predicted burial, near-backbone-11 definition
alpha	0.135	propensity predicted alpha angle
dry5	0.089	propensity predicted burial, dry-5 definition
pred_n_sep_back	0.085	predicted H-bond sequence separation for N
rejected_bonus	0.077	rejected alignment predicted constraints
hbond_dist	0.064	H-bond distance cost function, using LJ 6-12 potential
break	0.052	penalizes chain breaks
contact	0.049	average number of contacts (centroids of the backbone and sidechain within 8 Å) per residue
is_align	0.033	detects missing backbone atoms or chainbreaks

Undertaker Availablilty

Unfortunately, we do not have the resources to provide undertaker as an easily installable package. However, the undertaker source code (with an i686 binary), some supporting files, and the scripts we are using for CASP8 MQA are available at `http://www.soe.ucsc.edu/compbio/undertaker-mqa.tgz`. We only provide the package so that people could examine the source code to see implementation details. The program comes with no documentation and will be difficult to get working. We may release a version that is easier to install in the future, but currently have no funding for development work.

References

[1] Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins: Structure, Function, and Bioinformatics. 2008;69(3):184–193.

[2] Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003;19(12):1589–1591.

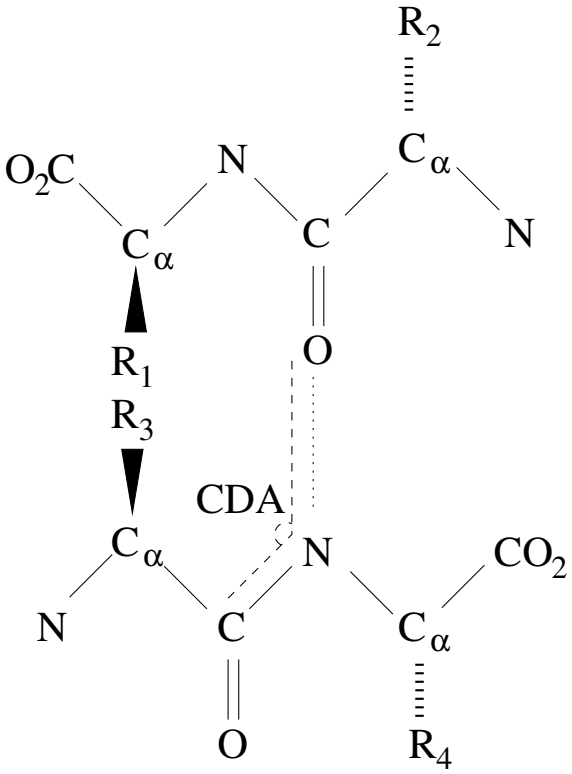
[3] Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31(13):3370–3374.

[4] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57(4):702–710.

[5] Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T, Fidelis K. New tools and expanded data analysis capabilities at the Protein Structure Prediction Center. Proteins. 2007;69(S8):19–26.

sep	probability	sep	probability	sep	probability
1	1.000000	13	0.0546504	25	0.0522208
2	0.00194422	14	0.0501612	26	0.0512311
3	0.176576	15	0.0492639	27	0.0526647
4	0.149368	16	0.0482015	28	0.0509335
5	0.278609	17	0.0475167	29	0.0518978
6	0.0788494	18	0.0484386	30	0.0515381
7	0.0658937	19	0.0481162	31	0.0504663
8	0.0690079	20	0.048748	32	0.0493822
9	0.0658018	21	0.0484083	33	0.0484552
10	0.0577989	22	0.0494527	34	0.0482571
11	0.0537069	23	0.0512322	35	0.0475389
12	0.0564388	24	0.0508363		

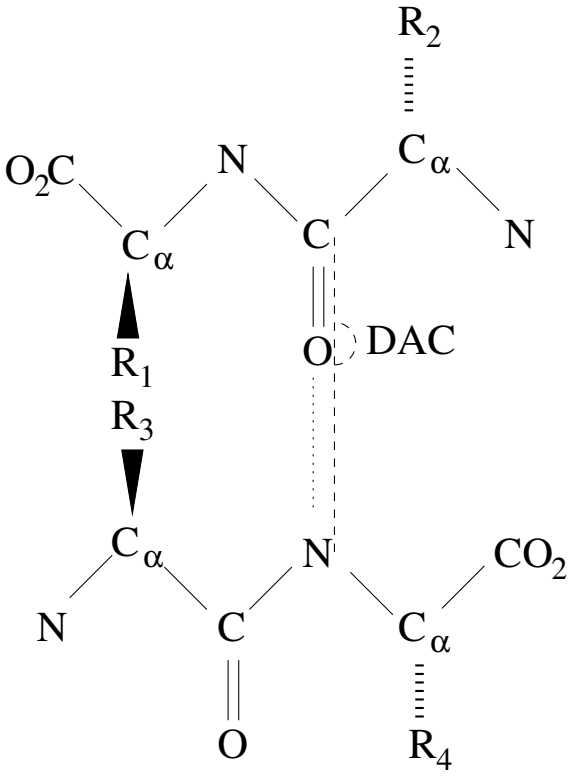
Table 4: P(contact|sep). Probabilities for contacts with a separation of greater than 35 are well estimated by $0.81(\text{sep})^{-0.8}$.



type	CDA
N-term N	[95, 140]
*	[60, 165]

type	μ	σ
one C	111.253	18.2975

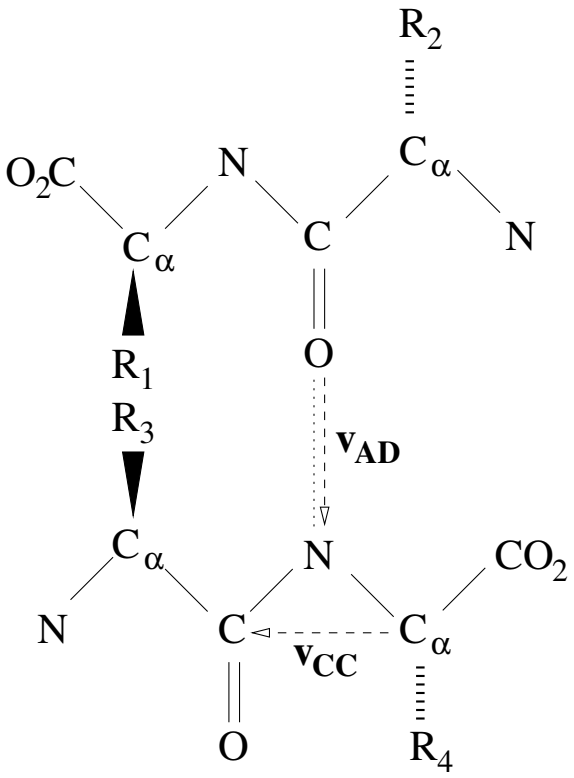
Figure 5: CDA (carbon-donor-acceptor) angle. The CDA angle (top) can be computed whenever the H-bond donor is covalently bonded to at least one C. For H-bond detection, the CDA angle must be within a certain range (center). When the N-terminal N is a donor, the permitted CDA angles have a slightly different range due to the different geometry. For determining the likelihood of an H-bond, the CDA angle is only used if one cannot compute nonplanarity (Figure 8), because only one C is bonded to the donor. The CDA angle is modeled as a normal distribution (bottom).



type	separation	DAC
bb	3	[85, 160]
bb	>3	[0, 120]
acceptor has H		[85, 160]
nonbb		[85, 180]

type	separation	μ	σ
bb	3	120.313	10.0430
bb	4	154.699	5.84436
bb	>4	154.389	11.7149
bb N, not bb O		127.207	19.0664
not bb N, bb O		133.497	18.8650
not bb N, not bb O		119.710	17.8554

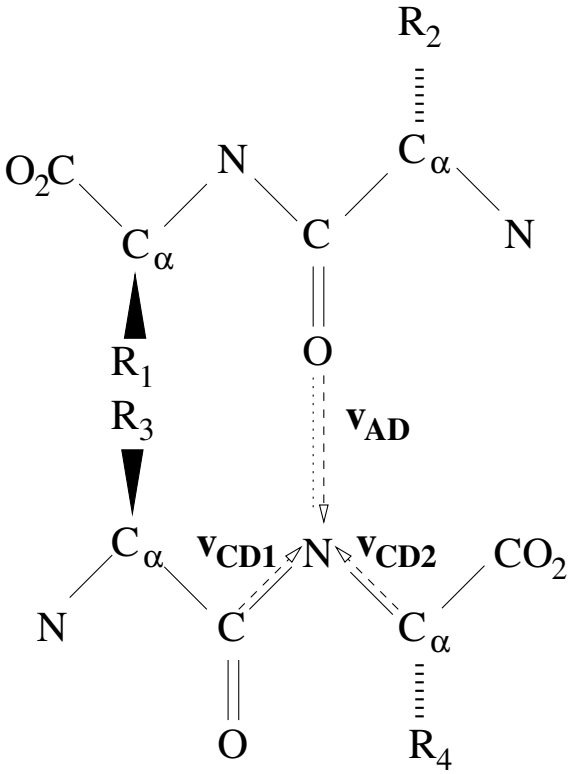
Figure 6: DAC (donor-acceptor-carbon) angle. The DAC angle (top) can be computed whenever the H-bond acceptor is covalently bonded to at least one C. For H-bond detection, the DAC angle must be within a certain range (center). For determining the likelihood of an H-bond, the DAC angle is modeled as normal distributions (bottom). Backbone is abbreviated as "bb."



type	separation	asymmetry
bb	3	[-0.4, 3.1]
bb	4	[-3.0, 3.0]
bb	>4	[-3.3, 3.3]
nonbb		[-3.3, 3.3]

type	separation	μ	σ
bb	3	0.42275	0.15418
bb	4	-0.0249	0.20357
bb	>4	-0.3355	0.31008
bb N, not bb O		0.18500	0.45128
not bb N, bb O		–	–
not bb N, not bb O		0.06418	0.52032

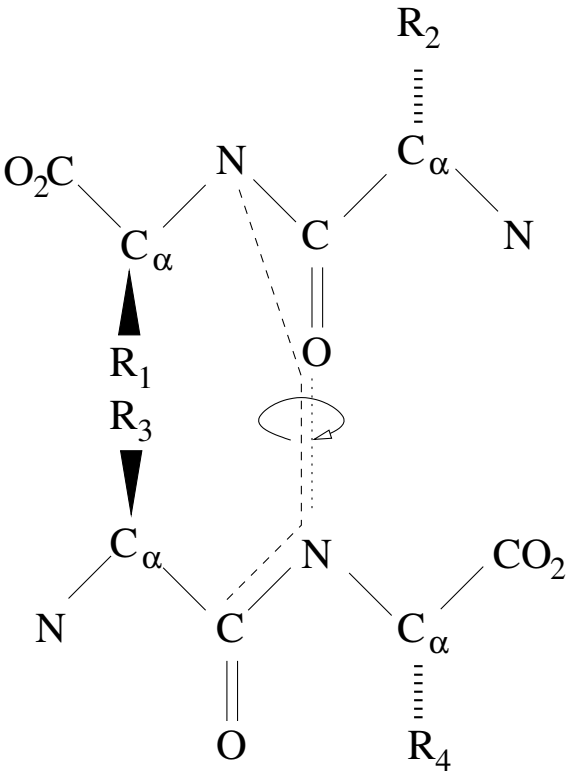
Figure 7: Asymmetry. Asymmetry (top) is computed as $\mathbf{v_{CC}} \cdot \mathbf{v_{AD}}$. Asymmetry can be computed whenever the H-bond donor is covalently bond to two C. For H-bond detection, asymmetry must be within a certain range (center). For determining the likelihood of an H-bond, asymmetry is divided by the donor-acceptor distance, and this statistic modeled as normal distributions (bottom). Backbone is abbreviated as "bb."



type	separation	nonplanarity
bb	3	[-3.5, 2.3]
bb	4	[-4.0, 3.0]
bb	>4	[-4.0, 3.6]
nonbb		[-4.5, 4.3]

type	separation	μ	σ
bb	3	0.16879	0.37696
bb	4	–	–
bb	>4	-0.0898	0.31779
nonbb		-0.0641	0.46044

Figure 8: Nonplanarity. Nonplanarity (top) is computed as $(\mathbf{v_{CD1}} \times \mathbf{v_{CD2}}) \cdot \mathbf{v_{AD}}$ and can be computed whenever the H-bond donor forms a covalent bond with two C atoms. For H-bond detection, nonplanarity must be within a certain range (center). For determining the likelihood of an H-bond, nonplanarity is divided by the donor-acceptor distance, and this statistic is modeled as normal distributions (bottom). Backbone (bb) H-bonds with a separation of greater than four have a relatively flat distribution and are therefore not modeled. Backbone is abbreviated as "bb."



type	separation	NOtor
bb	3	[-60, 35]
bb	4	[40, 140]

Figure 9: NOtor torsion angle. The NOtor torsion angle (top) is a geometric torsion angle that can only be computed for backbone H-bonds. For H-bond detection, the NOtor angle must be within a certain range for nearby backbone H-bonds (bottom). The NOtor angle is not used when computing the likelihood of an H-bond. Note that the rule for backbone H-bonds with a separation of 4 does exclude some rare antiparallel β-strand H-bonds. Backbone is abbreviated as "bb."

type	sep	dist
bb	3	[2.7, 3.6]
*	*	[2.5, 3.6]

type	separation	r	k
bb	3	2.99516	6.41089
bb	4	2.92963	9.79637
bb	>4	2.86875	10.6398
acceptor has H		2.96613	2.72088
not bb N or O, asymmetry computable		2.85315	6.51087
not bb N, bb O		2.82832	3.06984
all other cases		2.80461	3.23124

Table 5: Distance constraints. To be detected as a hydrogen bond, the distance between the donor and acceptor atoms must be within the certain limits (top). Distance is modeled as exponential distributions of Leonard-Jones energies (bottom) with fitted constants r and k (Equations 10 and 12). Backbone is abbreviated as "bb."