Supplementary material for: Protein Structure Prediction Begins Well but Ends Badly

Rhodri Saunders[1] *Charlotte M. Deane[1]

September 30, 2009

[1] University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, England, OX1 3TG.

* Correspondence to: Rhodri Saunders or Charlotte M. Deane, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford, England, OX1 3TG. Email: saunders@stats.ox.ac.uk or deane@stats.ox.ac.uk

1 EVA

We use prediction data from EVA to assess whether secondary structure prediction accuracy along the chain is equivalent. We consider only those methods that have made over 500 predictions. These methods, their general algorithm type and the number of predictions they made through EVA are listed in table 1.

Method	Prediction type	No.Predictions	Reference
APSSP	Neural net	1252	[1]
APSSP2	Neural net	724	[2]
JPred	Combination	1294	[3]
PHD	Neural net	1871	[4]
PHDpsi	Neural net	1758	[5]
PROF_king	Neural net	1400	[6]
PROFsec	Neural net	1609	Rost, unpublished
PSIpred	Neural net	1639	[7]
SAM-T99sec	Hidden Markov model	682	[8]
SSpro2	Neural net	1304	[9]

Table 1: EVA data sets for assessing N-terminal restriction in real protein structures

2 relFRAG

2.1 The CASP data set in numbers

We analyse free-modeling data from three CASP experiments: 6, 7 and 8. Template based models from CASP7 and CASP8 are also investigated. The number of models tested, broken down by method, are given in table 2. Our data set for template based models in CASP7 and CASP8 is given in table 3.

	First Models				All Models			
	Human		Server		Human		Server	
CASP	Models	Results	Models	Results	Models	Results	Models	Results
6	888	459	497	122	2,927	1,728	2,009	562
7	1,632	1,175	1,494	783	6,105	4,728	6,130	3,305
8	587	520	721	490	4,965	3,854	2,904	2,079

Table 2: CASP data sets for free-modeling in the three most recent meetings. Models (columns 2, 4, 6, 8, 10 and 12) give the total number of free models submitted to CASP. Results (columns 3, 5, 7, 9, 11 and 13) are the number of models falling within our inclusion criteria. The number of results is smaller than the number of models because N- and C- terminal fragments may be too close in sequence space or both fragments have a root mean squared deviation greater than 2Å.

	Template Based Models	
CASP Meeting	Models	Results
7	89,604	83,937
8	66,953	62,312

Table 3: Template Based Model data from CASP7 and CASP8. Models gives the total number of models that we analysed in this study. Results gives the number of relFRAG scores that were successfully calculated. The number of results is smaller than the number of models because N- and C- terminal fragments may be too close in sequence space or the root mean squared deviation of both fragments is greater than 2Å.

2.2 Mean relFRAG score

Through our measure relFRAG, we demonstrate that fragment prediction accuracy is significantly biased towards being more accurate near the amino terminus. The mean relFRAG score correlates well with fragment length (Figure 1). Our relFRAG scores is shown to correlate with model quality (Figure 2) with better quality models generally showing a more negative relFRAG score. Here we show that as fragment length increases to cover too much of the structure the correlation breaks down. Indicating that the structural trend implied by

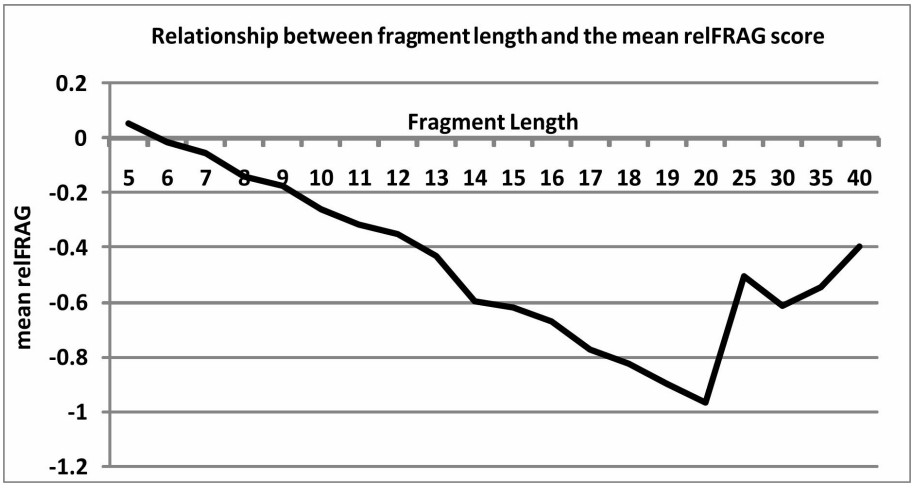

Figure 1: Fragment length is related to the mean relFRAG score, for fragment lengths 5 to 20 the correlation has an R^2 value of 0.9931. Up to length 20, as fragment length increases the bias towards better prediction accuracy near the amino terminus increases. At a fragment length of 20 the amino terminus is on average predicted over 2.5 times as accurately as the carboxy terminus. Bias toward increased prediction accuracy at the N-terminus is seen at longer fragment lengths (25, 30, 35 and 40) but it is less extreme and the line in general is heading back towards the expected value of zero. Data is taken from all valid first model predictions made in CASP6, CASP7 and CASP8.

relFRAG is concentrated on the terminal regions of proteins. The distribution of relFRAG scores is shown in figure 3. 57% of all model fragments have a negative relFRAG score; i.e. more accurate prediction near the amino terminus. This trend is also seen for fragments of length 10, 15 and 20.

When considering the actual distribution of root mean squared deviation

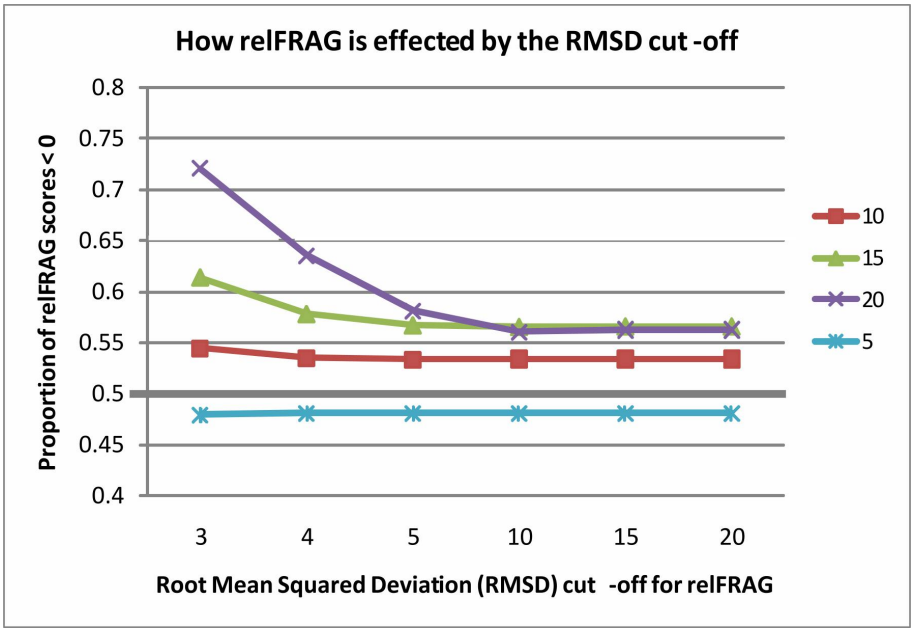

Figure 2: The effect our 2Å root mean squared deviation (RMSD) cut-off has on relFRAG scores. The RMSD cut-off has little effect on results for fragments of length 5 (blue stars). The proportion of models with a relFRAG less score than zero (better N-terminal prediction) at a fragment length of 20 (purple crosses) drops by approximately 0.14 from an RMSD cut-off of 3 to an RMSD cut-off of 5. For fragments of length 10 (red squares), 15 (green triangles) and 20 the proportion of models with a relFRAG score less than zero is significantly greater than 0.5 at all RMSD cut-offs.

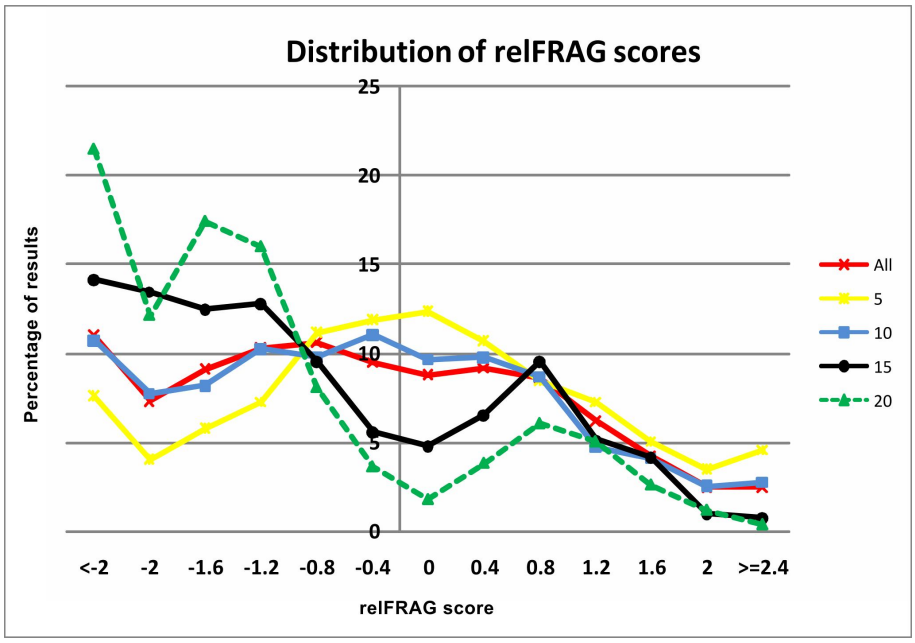

Figure 3: Distribution of relFRAG scores. Data is shown for all fragment lengths combined (all, red) and for fragments of length 5 (5, yellow), 10 (10, blue), 15 (15, black) and 20 (20, green). Data is taken from all valid first model predictions made in CASP6, CASP7 and CASP8.

(RMSD) values for model fragments (Figure 4) we find there is a higher percentage of N-terminal than C-terminal fragments in the RMSD range 0 to 1. Conversely there is a higher percentage of C-terminal than N-terminal fragments in the RMSD range 2 to 3. If we use an RMSD cut-off >3Åfor relFRAG calculations the proportion of models with a relFRAG score less than zero (better prediction at the N-terminus) is still significantly greater than 0.5 (Figure 2). Thus, the effect illustrated by relFRAG is a real, intra-model sequence-position mediated bias in prediction accuracy. That is to say, in general, the N-terminal fragment is predicted with greater accuracy than its analogous C-terminal fragment. The effect is seen at fragment lengths <5 and is more pronounced in better models.

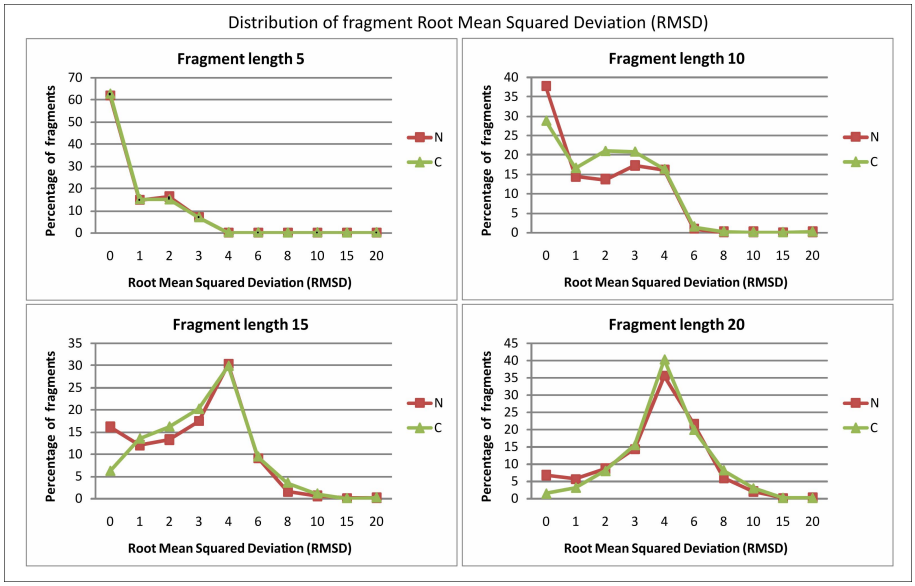

Figure 4: Distribution of root mean squared deviation (RMSD) values. Data is shown for fragments of length 5 (top left), 10 (top right), 15 (bottom left) and 20 (bottom right). The distribution of N-terminal fragment RMSD is shown in red (squares) while the distribution for C-terminal fragments is in green (triangles). The two distributions at length 5 are quintessentially identical. At lengths 10, 15 and 20 a higher percentage of N-terminal fragments than C-terminal fragments are found in the RMSD range 0 to 1. Over other RMSD ranges the distributions correlate well. Data is taken from all valid first model predictions made in CASP6, CASP7 and CASP8.

2.3 Bias is not due to one target

The number of models in our sample is very large. The number of target is considerably smaller. We perform a leave one out analysis to show that the bias observed in terminal prediction accuracy is not attributed to a single target (Table 4).

Excluded	Fragment Length			
Target	5	10	15	20
T0300	0.479	0.576	0.655	0.755
T0361_D1	0.484	0.588	0.721	0.864
T0307_D1	0.477	0.566	0.657	0.778
T0356	0.475	0.579	0.678	0.788
T0241_2	0.483	0.577	0.68	0.789
T0248_2	0.472	0.567	0.681	0.791
T0482-D1	0.49	0.577	0.68	0.789
T0443-D1	0.47	0.578	0.672	0.775
T0316	0.469	0.566	0.662	0.758
T0201	0.48	0.578	0.683	0.796
T0321_D2	0.477	0.576	0.68	0.789
T0496-D1	0.484	0.565	0.662	0.791
T0238	0.478	0.583	0.678	0.793
T0296	0.481	0.568	0.669	0.789
T0209_2	0.478	0.575	0.68	0.789
T0386	0.483	0.582	0.688	0.792
T0319	0.483	0.587	0.693	0.791
T0416-D2	0.48	0.58	0.682	0.791
T0405-D1	0.474	0.571	0.679	0.794
T0242	0.481	0.58	0.68	0.789
T0382_D1	0.476	0.578	0.665	0.768
T0397-D1	0.476	0.583	0.689	0.789
T0347_D2	0.482	0.58	0.682	0.788
T0216_1	0.479	0.579	0.687	0.786
T0321	0.475	0.569	0.68	0.789

Table 4: One target in particular is not responsible for the results observed. Column 1 indicates the target that has been excluded from the analysis. Columns 2, 3, 4 and 5 show the fraction of models that are predicted more accurately at the N-terminus than the C-terminus for each fragment length.

2.4 Algorithms

For each group we calculate the fraction of models predicted with higher accuracy at the N-terminus than the C-terminus. We consider only those groups with over 50 relFRAG results at all fragment lengths and more than 30 results

over length 10. The majority of participating groups predict the N-terminus of models more accurately than the C-terminus (Figure 5). Very few groups (17 if considering fragments of length 11 to 20) show no bias (fraction of models predicted better at the N-terminus between 0.48 and 0.52) in their predictions. These are Zhang-server, LOOPP_Server, forecast, FFASflextemplate, Pcons_multi, mariner1, Jones-UCL, GeneSilico, GeneSilicoMetaServer, Midway-Folding, SAINT1, PS2-manual (CASP 8); SSU, Sternberg, MTUNIC (CASP 7); KIST-CHOI, BAKER (CASP 6).

2.5 Domains

We find that tertiary structure prediction is generally more accurate near the amino terminus. Domains are seen as individual folding units. To investigate whether this amino terminal bias is restricted to the amino terminus of proteins or the is a general trend at the amino terminus of folding units we analyse data from domain 1 and domain 2 of multi-domain proteins. When considering all fragments more accurate prediction near the amino terminus is limited to domain 1, suggesting that the bias is concentrated at the start of a protein chain. However, when we consider only longer fragment lengths more accurate prediction of fragments near the amino terminus is evident for both domain 1 and domain 2 (Table 5).

Domain 1					
Fragment length	5	10	15	20	All
Number <0	719	673	347	120	7605
Number = 0	0	1	0	0	2
Number >0	787	445	205	53	5547
% <0	47.7	60.2	62.9	69.4	57.8
Domain 2					
Fragment length	5	10	15	20	All
Number <0	233	159	121	72	2261
Number = 0	0	0	0	0	0
Number >0	286	195	73	31	2352
% <0	44.9	44.9	62.4	69.9	49.0

Table 5: relFRAG data for free-modeling multi-domain targets broken down by domain. At longer fragment lengths, prediction accuracy is more accurate near the amino terminus. Over all lengths prediction accuracy is more biased toward better prediction at the amino terminus in domain 1 compared to domain 2. Actual data is shown for fragment lengths 5, 10, 15 and 20. Total data is the cumulative data for all fragment lengths 5 to 20 inclusive.

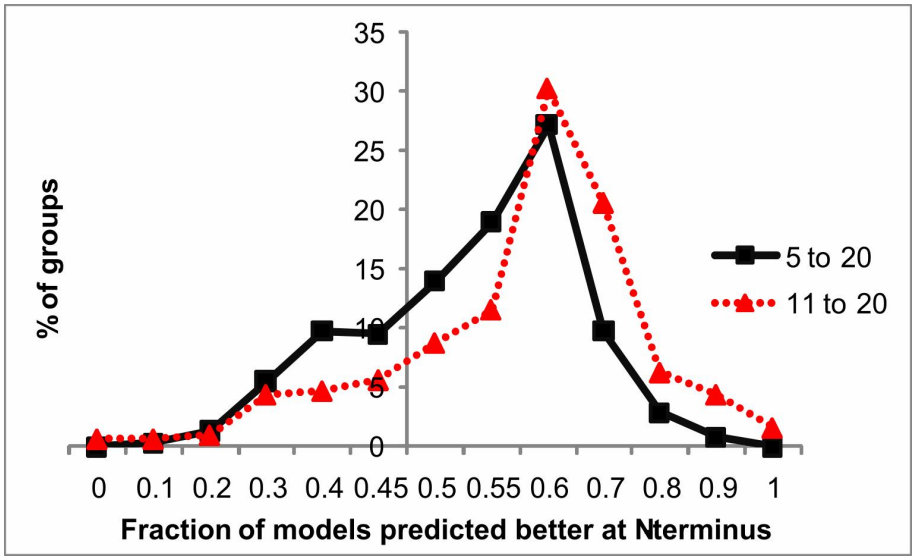

Figure 5: The majority of groups predict the N-terminus of their model with higher accuracy than the C-terminus. Considering all fragment lengths (black squares), over 70% of groups predict the N-terminus better than the C-terminus in a majority of their models. Taking fragments of length 11 and over (red triangles), the percentage of groups rises to over 80%. The X-axis is the fraction of the group's models predicted better at the N-terminus. The Y-axis indicates the percentage of groups.

2.6 Terminal Compactness

Our results suggest that N-terminal regions may be more compact and closer to their global energy minimum structure than C-terminal regions. Similarly, the C-terminal may show more subtle structural variance due to increased flexibility over the N-terminus. These suggestions run counter to the work of Laio and Micheletti [10] whose investigation of 458 proteins showed that the C-terminus is generally more compact than the N-terminus. We reran a number of their tests on a much larger data set, 2618 non-redundant proteins selected via the PISCES web server [11]. Compactness is calculated using both the Radius of Gyration (RoG) and the Moment of Inertia (MoI). As an example, MoI is the average distance of a residue (R) in the range R_N to R_C to the centre of mass (M) of all residues in the range. M is calculated as the average co-ordinates X, Y and Z of all residues in the range R_N to R_C (equation 1). Where L is the number of residues in the range R_N to R_C.

$$MoI = \frac{1}{L} \sum_{i=R_N}^{R_C} [\delta(R_i, M)]^2 \tag{1}$$

For each protein in our set we take a structural fragment of length X (where X varies from 6 to 40) from both the N-terminus and the C-terminus. The MoI and RoG of each structural fragment is then calculated and the relative terminal compactness calculated: $log(N - compactness)/(C - compactness)$. For both MoI and RoG, if the N-terminal fragment is more compact we expect a negative value to be returned. When considering our whole data set our findings support tho

se of Laio and Micheletti. However, as discussed in the main paper β-strand is more prevalent at the N-terminus and is known to be a less compact secondary structure than α-helix. When we consider only proteins with equivalent termini, that is the most N-terminal secondary structure is the same as the most C-terminal secondary structure then we find that, in general, the N-terminus is more compact than the C-terminus (Figure 6). Thus, measures of relative structural compactness are not independent of secondary structure types and it is only a fair test to compare structures with equivalent termini.

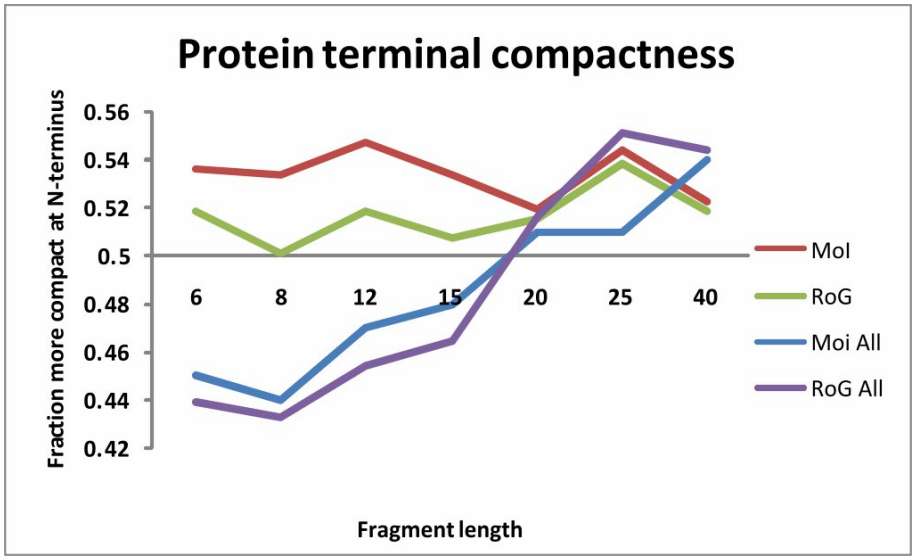

Figure 6: When comparing equivalent termini, we show that the N-terminus is generally more compact than the C-terminus (red and green lines). This is true for two different measure of compactness, Moment of Inertia (MoI) and Radius of Gyration (RoG). Equivalent termini are where both the N- and C-termini have the same type of secondary structure - e.g. either both helix or both strand. If non-equivalent termini are also included (blue and purple lines) the more compact terminus varies with fragment length.

References

[1] Raghava, G. P. S. *CASP4* , 75–76 (2000).

[2] Raghava, G. P. S. *CASP5* , A–132 (2002).

[3] Cuff, J. A. and Barton, G. J. *Proteins* **34**(4), 508–519 March (1999).

[4] Rost, B. *Methods in enzymology* **266**, 525–539 (1996).

[5] Przybylski, D. and Rost, B. *Proteins* **46**(2), 197–205 February (2002).

[6] Ouali, M. and King, R. D. *Protein science : a publication of the Protein Society* **9**(6), 1162–1176 June (2000).

[7] Jones, D. T. *J Mol Biol* **292**(2), 195–202 September (1999).

[8] Karplus, K., Barrett, C., and Hughey, R. *Bioinformatics* **14**(10), 846–856 (1998).

[9] Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. *Proteins* **47**(2), 228–235 May (2002).

[10] Laio, A. and Micheletti, C. *Proteins* **62**(1), 17–23 (2006).

[11] Wang, G. and Jr. *Nucleic Acids Res* **33**(Web Server issue), W94–8 (2005).