Extensive Protein and DNA Backbone Sampling Improves Structure-based Specificity Prediction for C_2H_2 Zinc Fingers: Supplementary Information

1 Methods

1.1 Knowledge-based protein-DNA interaction potential

Using a database of high-resolution co-crystal structures [1], we parameterized a simple protein-DNA interaction potential for use in the low-resolution phase of our fragment assembly simulations. This knowledge-based potential has two components: an environment term, which captures amino acid preferences for being at the protein-DNA interface, and a pair term, which reflects amino acid-nucleotide contact propensities. To fit these two terms, all amino acid-nucleotide pairs within a distance of 12 Å were recorded, and the total number of such nucleotide neighbors for each protein residue was totaled and written out. Distances were calculated by defining a single interaction center for each residue. For amino acids, this position was taken to be the C_β (C_α for Gly); for nucleotides, it is calculated by averaging two atoms in the base: N7+N6 for A, C5+N4 for C, N7+O6 for G, and C5M+O4 for T. The neighbor counts were binned (0, 1-2, ..., 11-12, > 12), and a propensity was calculated for each amino acid to have a given number of neighbors by comparing the actual counts for that amino acid with the expected number based on the overall frequency of that amino acid and the frequency of that neighbor bin. To parametrize the environment term, the amino acid-base distances were binned (0-4, 4-6, 6-8, 8-10, 10-12 Å), and a propensity for each (amino acid, base, distance-bin) triple was calculated by comparing the actual counts for that triple to expected counts estimated based on the independent counts for that amino acid and base in the given bin. All propensities were converted to scores by taking their negative logarithm.

$$\begin{aligned}
E_{\text{env}}(aa, b) &= -\log\left(\frac{N(aa,b)}{P(aa)N(b)}\right) \\
&= -\log\left(\frac{N(aa,b)N}{N(aa)N(b)}\right) \\
E_{\text{pair}}(aa, na, d) &= -\log\left(\frac{N(aa,na,d)}{P(aa|d)P(na|d)N(d)}\right) \\
&= -\log\left(\frac{N(aa,na,d)N(d)}{N(aa,d)N(na,d)}\right)
\end{aligned}$$

Here aa is one of the 20 amino acids, na is one of the 4 bases, b represents a neighbor-count bin, d represents a distance bin, and $N(\cdot)$ indicates the total number of occurrences in the database (N in formula S1 stands for the total number of amino acids in the database). To compute the score for a model, we find all protein-DNA residue pairs within 12 Å, compute neighbor counts for each amino acid, score each amino acid according to its neighbor propensities, and sum the environment scores for each protein-DNA contact.

1.2 Orientation-dependent implicit solvation model

To better model stacking interactions at protein-DNA interfaces, we developed a simple, orientation-dependent variant of the Lazaridis-Karplus (LK) implicit solvent model [2]. In the standard LK model, the interaction energy for two atoms a_1 and a_2 is the sum of two terms, one capturing the desolvation of a_1

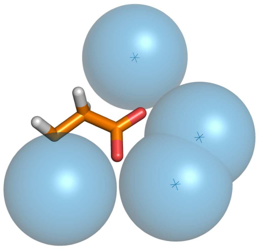

Figure S1: Locations of virtual water atoms used to calculate orientation-dependent desolvation energy of an Asp sidechain.

by a_2, and one capturing the desolvation of a_2 by a_1. The magnitude of each contribution depends on the distance between the two atoms and their respective atom types. We modified the desolvation contributions for polar atoms to make them dependent on the relative orientation of the desolvating atom. Virtual water atoms are placed at optimal locations relative to the polar atom based on its expected hybridization state (Fig. S1), and the distance from the desolvating atom to each of these optimal waters is computed. A scaling term for the LK desolvation energy is computed based on the minimal water distance d using the equation

$$\lambda(d) = \begin{cases} 1 & d < R - w \\ \lambda_0 + (1-\lambda_0) S\left(\frac{R-d}{w}\right) & R - w \leq d < R \\ \lambda_0 & R \leq d \end{cases}$$

where $R = 1.4 + R_{\mathrm{LJ}}$ is a cutoff distance equal to sum of the Lennard-Jones radii of a water molecule and the desolvating atom, w is the width of a ramping zone in which the scaling factor interpolates between λ_0 and 1 as the desolvating atom approaches the optimal water location, and S is a sigmoidal interpolation function with $S(0) = 0$ and $S(1) = 1$. We took $\lambda_0 = 0.5$ and $w = 0.9$ based on visual inspection of various stacking geometries; it is likely that these parameters could be more systematically optimized.

1.3 Comparison with other methods

We compared our structure-based approach with three previously described and publicly accessible algorithms for predicting ZF-DNA interactions: a structure-based approach incorporating family-specific amino acid-nucleotide interaction preferences learned from experimental binding data [3] (ttp://compbio.cs.uji.ac.il/Zinc/; “Kaplan05”); ZIFIBI, which uses a hidden Markov model to generate binding site predictions [4] (ttp://bioinfo.anyang.ac.kr/ZIFIBI/); and a recent machine-learning approach that incorporates data on binding and non-binding DNA sites through the use of a support vector machine [5] (ttp://compbio.cs.princeton.edu/zf/}; ‘‘Persikov09’’).Giventat experimental binding data for the proteins of known structure were likely used to train one or more of these methods, we restricted the comparison to the six ZF proteins without solved structures whose specificities were recently profiled by protein binding microarrays [6]. As it was not straightforward to generate full PFMs for each algorithm, we focused on the simple metric that counts the number of positions at which the preferred base in prediction and experiment agree. Predictions for the Kaplan05 algorithm were generated through the web server. For several of the sequences, the algorithm only generated a binding prediction for a single finger, although the results suggested that other fingers were identified in the sequence. For these targets, we resubmitted subsequences to generate predictions for all fingers. ZIFIBI predictions only depend on the amino acids at helix positions −1, 3 and 6. These amino acid triplets were entered into the web form to generate predictions for the benchmark proteins. The Persikov09 algorithm does not generate binding specificity profiles for a target zinc finger protein; instead, it searches an input DNA sequence for potential binding sites and ranks them according to an energy model. To compare with the experimental profiles, we downloaded the SVM models and performed searches against DNA sequences containing all 4-mers in order to identify optimal binding sites for each finger. The top-scoring site was compared to the experimental data. The results reported in the main text are for the linear kernel model, which recovered 31 out of 36 positions. Interestingly, the polynomial kernel model, which was reported to be superior in prediction accuracy [5], recovered only 30 of 36 positions.

2 Results

2.1 Variation in Protein-DNA interface geometry

The interface fragment assembly protocol depends on the existence of template structures with similar interface geometries from which to extract interface fragments. If interface geometry is more conserved in the zinc finger protein family than in other families of DNA-binding proteins, the protocol might not be expected to perform as well on these other families. To test this, we collected representative protein structures from three other eukaryotic transcription factor families: the homeodomains, the b-ZIP proteins, and the b-HLH proteins. We aligned each family by defining a core interface consisting of 7 residues in an alpha-helix, and three base-pairs of DNA. We then computed pairwise RMSD values over these 13 residues, and compared the distribution of RMSDs to the distribution obtained for the ZF proteins in our benchmark (using the 7 helix positions −1 through 6 and the corresponding DNA triplet). The results are given in Figure S2, which shows that the ZF protein-DNA interfaces are, in fact, slightly more diverse than those of these other protein families. This result suggests that the interface fragment assembly algorithm may yield useful predictions for these families as well.

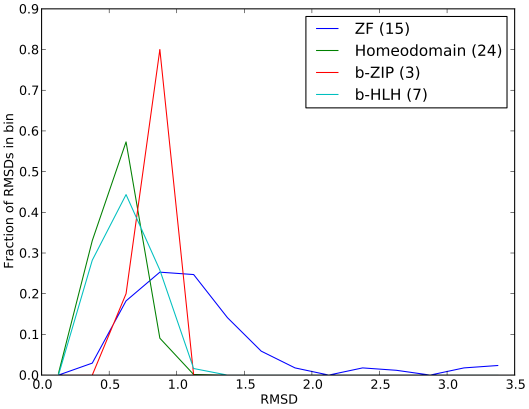

Figure S2: Interface RMSD distributions for DNA-binding protein families. For each family, we aligned the indicated number of representative protein structures using a core interface (7 protein and 6 DNA residues) and computed pairwise RMSD values. The RMSD values were binned with a bin size of 0.25Å; the resulting distributions are plotted here using the midpoints of the bins.

References

[1] Havranek, J. J., Duarte, C. M., and Baker, D. (2004) A simple physical model for the prediction and design of protein-DNA interactions. *J Mol Biol,* **344**(1), 59–70.

[2] Lazaridis, T. and Karplus, M. (1999) Effective energy function for proteins in solution. *Proteins,* **35**(2), 133–52.

[3] Kaplan, T., Friedman, N., and Margalit, H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol,* **1**(1), e1.

[4] Cho, S. Y., Chung, M., Park, M., Park, S., and Lee, Y. S. (2008) ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. *Biochem Biophys Res Commun,* **369**(3), 845–8.

[5] Persikov, A. V., Osada, R., and Singh, M. (2009) Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics,* **25**(1), 22–9.

[6] Zhu, C., Byers, K. J., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M., Philippakis, A. A., Hu, Y., De Masi, F., Pacek, M., Rolfs, A., Murthy, T., Labaer, J., and Bulyk, M. L. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res,* **19**(4), 556–66.

[7] Siggers, T. W., Silkov, A., and Honig, B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J Mol Biol,* **345**(5), 1027–45.

[8] Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F., and Joung, J. K. (2009) Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. *Nat Protoc,* **4**(10), 1471–501.

[9] Fu, F., Sander, J. D., Maeder, M., Thibodeau-Beganny, S., Joung, J. K., Dobbs, D., Miller, L., and Voytas, D. F. (2009) Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Res,* **37**(Database issue), D279–83.

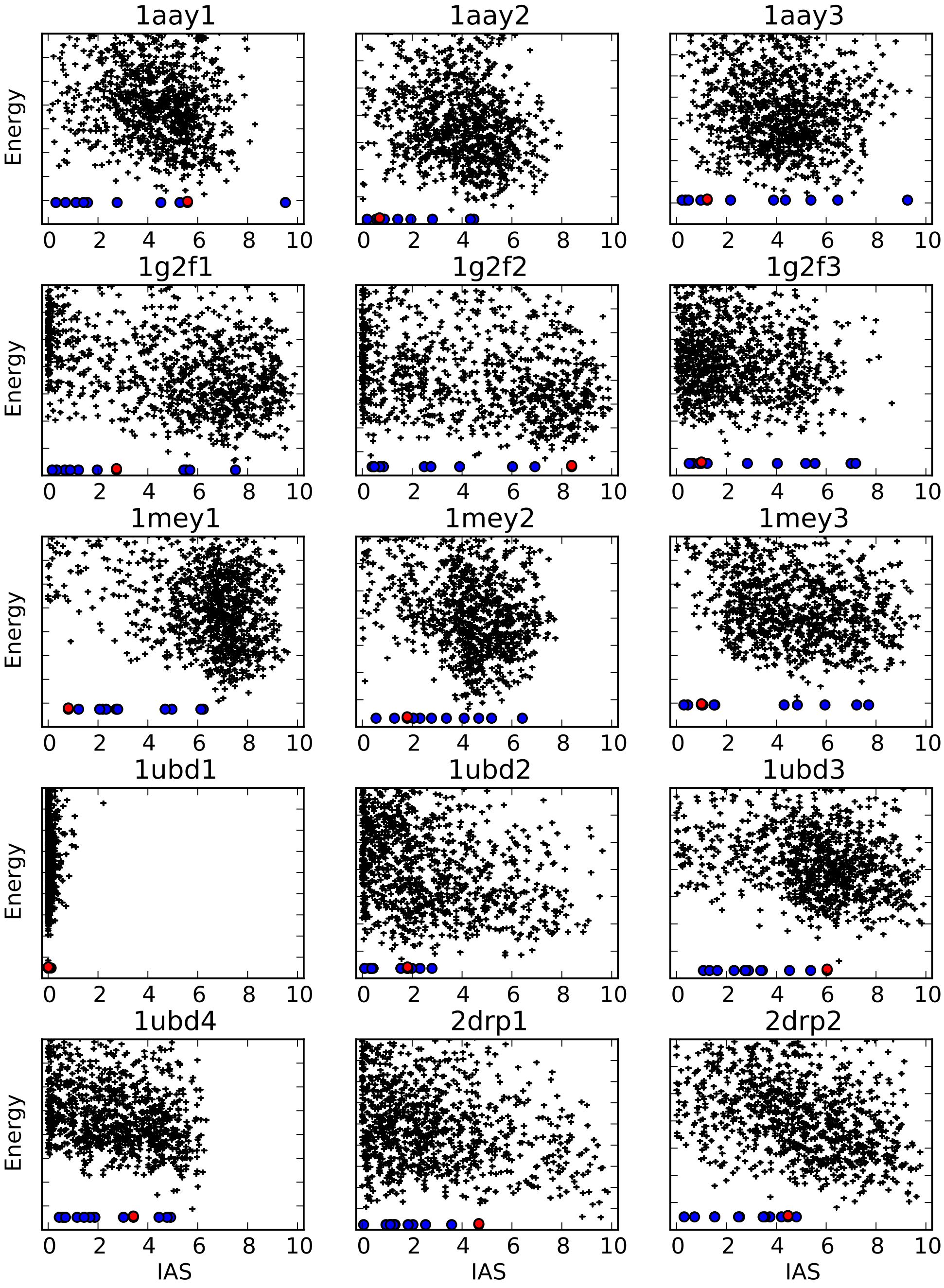

Figure S3: Scatter plots of IAS score [7] (x-axis) against all-atom energy (y-axis) for fragment assembly models built for the 15 individual zinc fingers with solved structures in the benchmark set. Blue circles along the bottom indicate the IAS similarity values for all templates used in fragment selection. The red circle marks the template with highest sequence similarity. The IAS score ranges from 0 to 10 and increases with increasing interface similarity. Y-axis tick marks are shown at 5.0 energy unit spacing (~6.5 kcal/mol).

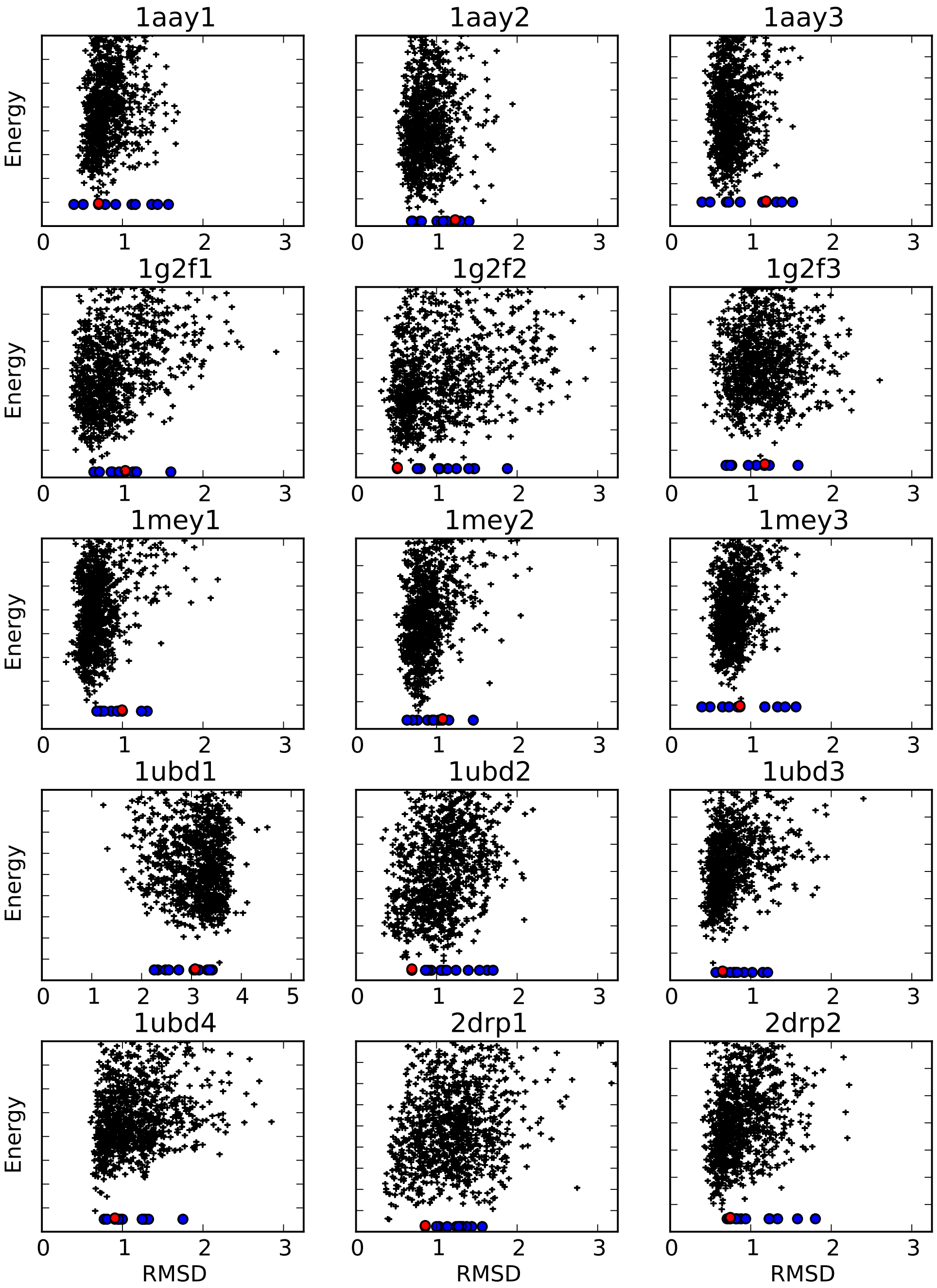

Figure S4: Scatter plots of RMSD (x-axis; computed over C_α atoms for helix positions -1 to 6 and C1′ atoms for both strands of the triplet binding site) against all-atom energy (y-axis) for fragment-assembly models built for the 15 individual zinc fingers with solved structures in the benchmark set. Blue circles along the bottom indicate the RMSD values for all templates used in fragment selection. The red circle marks the template with highest sequence similarity. Y-axis tick marks are shown at 5.0 energy unit spacing (~6.5 kcal/mol).

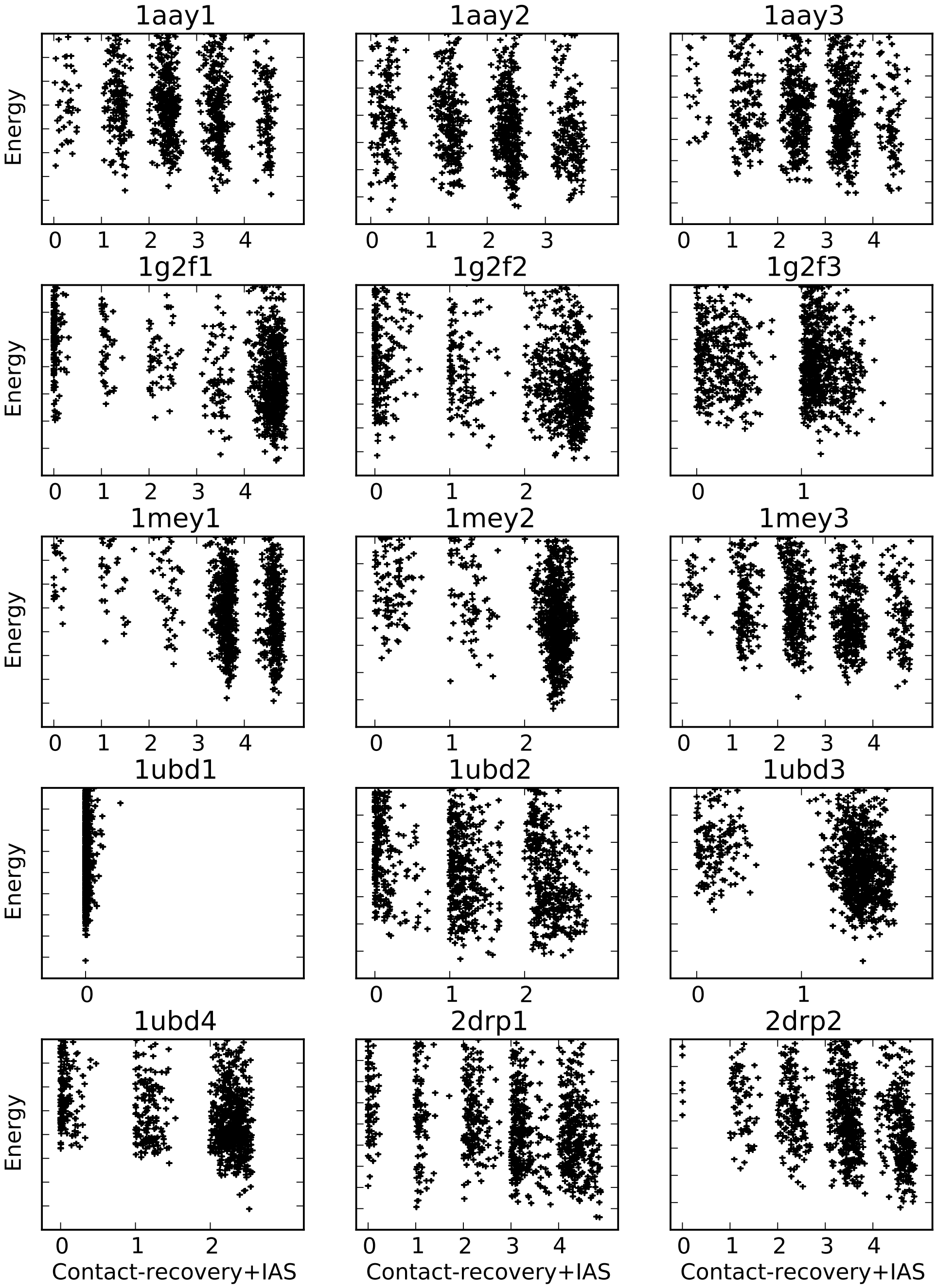

Figure S5: Scatter plots of all-atom energy (y-axis) versus a similarity score that combines contact recovery and IAS score [7] (x-axis). The IAS score is multiplied by 0.09 so that models with different numbers of recovered native contacts can be differentiated. Contacts correspond to protein-DNA hydrogen bonds to major groove atoms in the triplet binding site (1ubd finger 1 has no such contacts). The IAS score ranges from 0 to 10 and increases with increasing interface similarity. Y-axis tick marks are shown at 5.0 energy unit spacing (~6.5 kcal/mol).

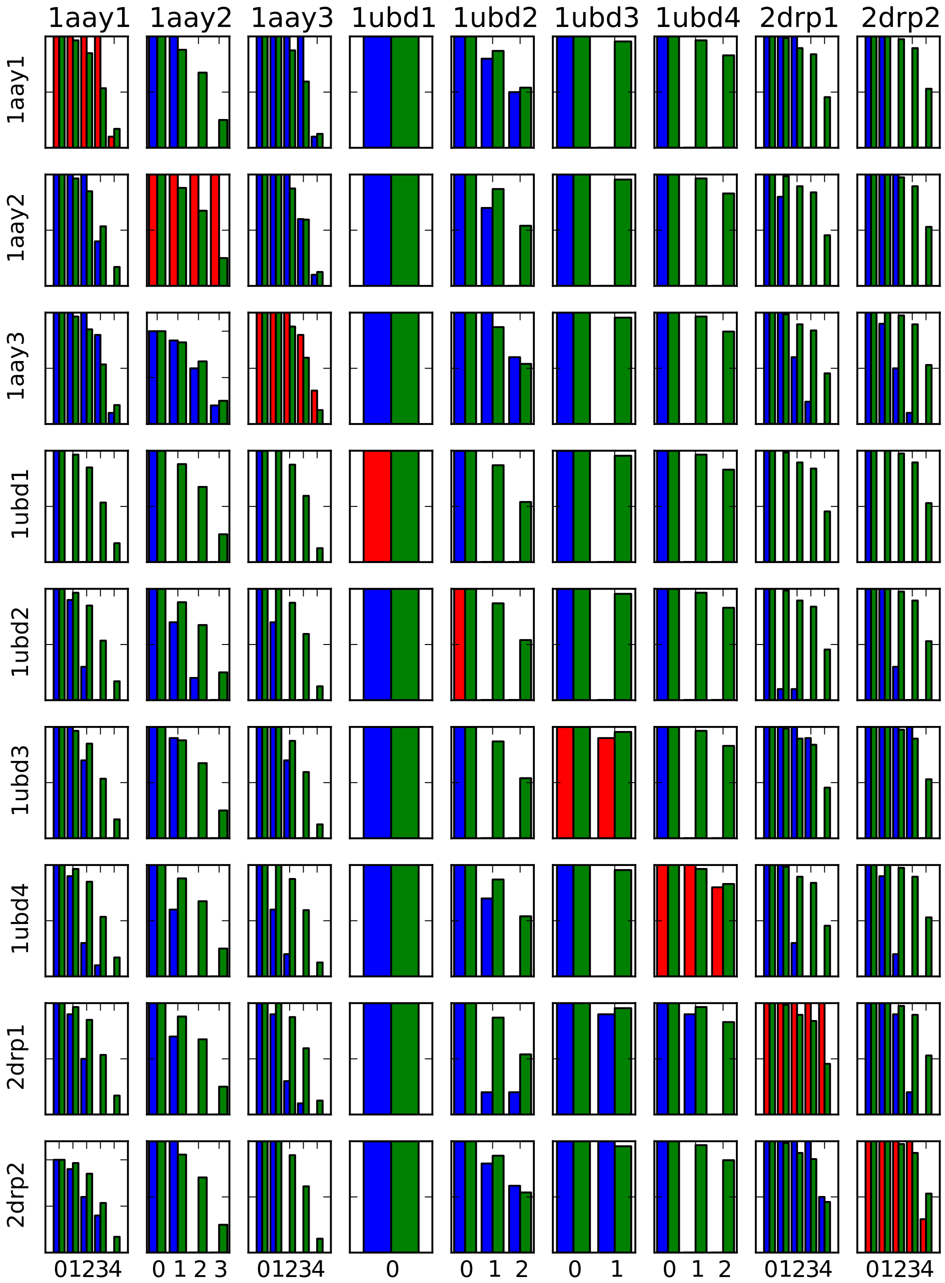

Figure S6: Cumulative contact recovery histograms comparing fragment assembly models to fixed-backbone homology models built by template-based sidechain prediction. Red/blue bars in a given row and column represent models built for the target corresponding to that column using the template backbone corresponding to that row; blue bars represent non-self-template models, and the red bars represent the models built with the target backbone itself. Green bars represent fragment assembly models. The height of each bar represents the fraction of all models recovering at least that many native contacts. Contacts correspond to protein-DNA hydrogen bonds to major groove atoms in the triplet binding site (1ubd finger 1 has no such contacts).

ID	Target site	Predicted site		ID	Target site	Predicted site		ID	Target site	Predicted site	
211	GTCGGGGTA	GCCGGGGCA	(7)	212	GTCGGGGTA	GCCGGGGCA	(7)	213	GTCGGGGTA	TAATGGGCA	(4)
214	GTCGGGGTA	GCCGGGGCA	(7)	215	GTCGGGGTA	GCCGGGGCA	(7)	216	GTCGGGGTA	GCCTGGGCA	(6)
217	GTCGGGGTA	GCCGGGGCA	(7)	218	GTCGGGGTA	GTCGGGGCA	(8)	219	GTCGGGGTA	GTCGGGGCA	(8)
220	GTCGGGGTA	GCCGGGGCA	(7)	221	GTCGGGGTA	GCCGGGGCA	(7)	222	GTCGGGGTA	GCCGGGGCA	(7)
223	GAAGCAGCA	GAAGCAGGC	(7)	224	GAAGCAGCA	GAAGCAGCT	(8)	225	GAAGCAGCA	GAAGCAGCT	(8)
226	GAAGCAGCA	GAAGCAGCG	(8)	227	GAAGCAGCA	GAAGCAGGA	(8)	228	GAAGCAGCA	GAAGCAGCT	(8)
229	GAAGCAGCA	GAAGCAGCT	(8)	230	GAAGCAGCA	GAAGCAGCG	(8)	231	GAAGCAGCA	GAAGCAGCG	(8)
232	GAAGCAGCA	GAAGCAGCG	(8)	233	GAAGCAGCA	GAAGCAGAA	(8)	234	GAAGCAGCA	GAAGCAGCG	(8)
235	GAAGATGGT	GAAGATGTA	(7)	236	GAAGATGGT	GAAGATTTA	(6)	237	GAAGATGGT	GAAGATGTT	(8)
238	GAAGATGGT	GAAGAAGCC	(6)	239	GAAGATGGT	GATGATGTT	(7)	240	GAAGATGGT	GAAGATGTT	(8)
241	GAAGATGGT	GATGATGTA	(6)	242	GAAGATGGT	GAAGATGTT	(8)	243	GAAGATGGT	GAAGCATTT	(5)
244	GAAGATGGT	GAAGATGTT	(8)	245	GAAGATGGT	GATGCTGTG	(5)	246	GAAGATGGT	GAAGATGTT	(8)
247	GACGACGGC	GAAGAAGGC	(7)	248	GACGACGGC	GAAGATGGC	(7)	249	GACGACGGC	GAAGATGGC	(7)
250	GACGACGGC	GACGACGGA	(8)	251	GACGACGGC	GAAGACGGC	(8)	252	GACGACGGC	GACGATGGC	(8)
253	GACGACGGC	GACGACGGC	(9)	254	GACGACGGC	GACGATGGC	(8)	255	GACGACGGC	GACGATGGC	(8)
256	GACGACGGC	GAAGACGGC	(8)	257	GTCGATGCC	GTGGCTGAC	(6)	258	GTCGATGCC	GCCGATGCC	(8)
259	GTCGATGCC	GCCGATGCT	(7)	260	GTCGATGCC	GCCGACGCC	(7)	261	GTCGATGCC	GCCGATGCC	(8)
262	GTCGATGCC	GCCGATGCC	(8)	263	GTCGATGCC	GCCGATGCC	(8)	264	GTCGATGCC	GCCGATGCC	(8)
265	GTCGATGCC	GCCGATGCC	(8)	266	GAGGACGGC	GTGGACGGC	(8)	267	GAGGACGGC	GGGGAAGGC	(7)
268	GAGGACGGC	GACGACGGC	(8)	269	GAGGACGGC	GAGGATGGC	(8)	270	GAGGACGGC	GAGGACGGC	(9)
271	GAGGACGGC	GACGACGGC	(8)	272	GAGGACGGC	GGGGAAGGC	(7)	273	GAGGACGGC	GGGGACGGC	(8)
274	GAGGACGGC	GAGGATGGC	(8)	275	GAGGACGGC	GTGGACGGC	(8)	276	GAGGACGGC	GAGGACTTA	(6)
277	GTGGCGGAT	GTGGGGTGT	(6)	278	GTGGCGGAT	GTGGGGTAT	(7)	279	GTGGCGGAT	GTGGCGGTT	(8)
280	GTGGCGGAT	GTGGCGGTT	(8)	281	GTGGCGGAT	GTGGCGTAT	(8)	282	GTGGCGGAT	GAGGCGGAC	(7)
283	GTGGCGGAT	GGGGCGGTT	(7)	284	GTGGCGGAT	GTGGCGGTT	(8)	285	GAGGACGGC	GAGGACGGC	(9)
286	GAGGACGGC	GAGGACGGC	(9)	287	GAGGACGGC	TTGGACGGC	(7)	288	GAGGACGGC	GAAGACGGC	(8)
289	GAGGACGGC	GAGGACGGC	(9)	290	GAGGACGGC	GAGGACGGC	(9)	291	GAGGACGGC	GAGGATGGC	(8)
292	GAGGACGGC	GAGGACGGC	(9)	293	GCCGTCGCC	GACGCCGCC	(7)	294	GCCGTCGCC	GCCGGCGCC	(8)
295	GCCGTCGCC	GCCGCCGCC	(8)	296	GCCGTCGCC	GCCGCCGCC	(8)	297	GCCGTCGCC	GTCGCCGCC	(7)
298	GCCGTCGCC	GTCGCCGCC	(7)	299	GCCGTCGCC	GACGCCGCC	(7)	300	GCCGTCGCC	GCCGCCGTC	(7)
301	GCCGTCGCC	GCCGCCGCC	(8)	302	GCCGTCGCC	GACGTGGCC	(7)	303	GCCGTCGCC	GACGTGGCC	(7)
304	GCCGTCGCC	GACGTGGCC	(7)	305	GCCGTCGCC	GACGTGGCC	(7)	306	GCCGTCGCC	GACGTGGCC	(7)
307	GCCGTCGCC	GACGTGGCC	(7)	308	GCCGTCGCC	GCCGCCGTA	(6)	309	GCTGCTGCC	GCTGCTGTC	(8)
310	GCTGCTGCC	GCTGCTGCT	(8)	311	GCTGCTGCC	GCCGATGCC	(7)	312	GCTGCTGCC	GCCGATGCC	(7)
313	GCTGCTGCC	GCCGATGCC	(7)	314	GCTGCTGCC	GCCGATGCC	(7)	315	GCTGCTGCC	GCCGCCGCC	(7)
316	GCTGCTGCC	GCCGATGCC	(7)	317	TTAGAAGTG	TTAGAATGG	(7)	318	TTAGAAGTG	TTAGAATTG	(8)
319	TTAGAAGTG	TTAGGGGCG	(6)	320	TTAGAAGTG	TTAGAAGAG	(8)	321	TTAGAAGTG	TTAGAATTG	(8)
322	TTAGAAGTG	TTAGGGGCG	(6)	323	TTAGAAGTG	TTAGAATTG	(8)	324	TTAGAAGTG	TTAGAAGAG	(8)
325	TTAGAAGTG	TTAGAAGAG	(8)	326	TTAGAAGTG	TTAGAAGGG	(8)	327	TTAGAAGTG	TTAGAATTG	(8)
328	TTAGAAGTG	TTAGAAGAG	(8)	329	TTATGGGAG	TTATAGGAG	(8)	330	TTATGGGAG	TTATGGGAG	(9)
331	TTATGGGAG	TTATAGGAG	(8)	332	TTATGGGAG	TTATGGGAC	(8)	333	TTATGGGAG	TTATGGGAC	(8)
334	TTATGGGAG	TTATGGGAC	(8)	335	TTATGGGAG	TTATGGGAC	(8)	336	TTATGGGAG	TTATAGGAC	(7)
337	TTATGGGAG	TTATAGGAG	(8)	338	TTATGGGAG	TTATAGGAC	(7)	339	TTATGGGAG	TTATAGGAG	(8)
340	TTATGGGAG	TTATGGGAG	(9)	341	TTATGGGAG	TTATAGGAC	(7)	342	TTATGGGAG	TTATGGGAG	(9)
343	TTATGGGAG	TTATGGGAC	(8)	344	TTATGGGAG	TTATGGGAG	(9)	345	TTATGGGAG	TTATAGGAC	(7)
346	TTATGGGAG	TTATAGGAC	(7)	347	TTATGGGAG	TTATGGGAG	(9)	348	TTATGGGAG	TTATAGGAC	(7)
349	TTATGGGAG	TTATAGGAC	(7)	350	TTATGGGAG	TTATGGGAC	(8)	351	TTATGGGAG	TTATAGGAA	(7)
352	TTATGGGAG	TTATAGGAC	(7)	353	GAAGACGCT	GAAGAAGCA	(7)	354	GAAGACGCT	GAAGATGCC	(7)
355	GAAGACGCT	GAAGACGCA	(8)	356	GAAGACGCT	GAAGACGTT	(8)	357	GAAGACGCT	GAAGAAGCA	(7)
358	GAAGACGCT	GAAGACGCA	(8)	359	GAAGACGCT	GAAGATGCC	(7)	360	GAAGACGCT	GAAGACGTT	(8)
361	GAAGACGCT	GAAGATGCC	(7)	362	GAAGACGCT	GAAGATGCC	(7)	363	GAAGACGCT	GAAGACGTC	(7)
364	GAAGACGCT	GAAGATGCC	(7)	365	GAGGACGTG	GAGGACTTG	(8)	366	GAGGACGTG	GAGGACGTG	(9)
367	GAGGACGTG	GAGGACGTG	(9)	368	GAGGACGTG	GGGGACGAG	(7)	369	GAGGACGTG	GAGGACGTG	(9)
370	GAGGACGTG	GAGGACGTG	(9)	371	GAGGACGTG	GAGGACGTG	(9)	372	GAGGACGTG	GAGGACTGG	(7)
373	GAGGACGTG	GAGGAATTG	(7)	374	GAGGACGTG	GAGGAATTG	(7)	375	GAGGACGTG	GTGGACGAG	(7)
376	GAGGACGTG	GGGGACGAG	(7)	377	GGAGGTGGT	GGAGGCTGT	(7)	378	GGAGGTGGT	GGAGGTTGT	(8)
379	GGAGGTGGT	GGAGGCTGT	(7)	380	GGAGGTGGT	GGAGGTTGT	(8)	381	GGAGGTGGT	GGAGGCGTT	(7)
382	GGAGGTGGT	GGAGGCGGC	(7)	383	GGAGGTGGT	GGAGGCTGT	(7)	384	GGAGGTGGT	GGAGGCTGG	(6)
385	GGAGGTGGT	GGAGGCTGT	(7)	386	GGAGGTGGT	GGAGGCTGT	(7)	387	GGAGGTGGT	GGAGGCTGT	(7)
388	GGAGGTGGT	GGAGGCTGT	(7)	389	GCCGGCGGC	GACGGTGGG	(6)	390	GCCGGCGGC	GACGGTGGG	(6)
391	GCCGGCGGC	GACGGTGGG	(6)	392	GCCGGCGGC	GACGGCGGC	(8)	393	GCCGGCGGC	GACGGTGGG	(6)
394	GCCGGCGGC	GACGGCGGC	(8)	395	GCCGGCGGC	GACGGCGGC	(8)	396	GCCGGCGGC	GACGGTGGG	(6)
397	GCCGGCGGC	GACGGCGGC	(8)	398	GGAGGAGGT	GGAGGAGGC	(8)	399	GGAGGAGGT	GGAGGAGGC	(8)

Table S1: Specificity predictions for OPEN [8] zinc finger arrays: ZiFDB [9] array ID; target site for which the 3-finger array was selected; binding site predicted by structural simulations; count of the number of positions at which the two sites agree (in parentheses). Positions in the predicted site that match the experimental site are underlined.

ID	Target site	Predicted site	ID	Target site	Predicted site	ID	Target site	Predicted site
400	GGAGGAGGT	GGAGGAGGC (8)	401	GGAGGAGGT	GGAGGATGT (8)	402	GGAGGAGGT	GGGGAGGAC (4)
403	GGAGGAGGT	GGAGGAGGC (8)	404	GGAGGAGGT	GGAGGAGGC (8)	405	GGAGGAGGT	GGAGGAGGC (8)
406	GGAGGAGGT	GGAGGAGGC (8)	407	GGAGGAGGT	GGAGGAGGC (8)	408	GGAGGAGGT	GGAGGAGGC (8)
409	GGAGGAGGT	GGAGGAGGC (8)	410	GGCGGCGGA	GCGGGGGAA (5)	411	GGCGGCGGA	GCGGGGGAA (5)
412	GGCGGCGGA	GGGGCGGAC (4)	413	GGCGGCGGA	GGGGCGGAC (4)	414	GGCGGCGGA	GGGGCGGAC (4)
415	GGCGGCGGA	GGGGTGGAA (5)	416	GGCGGCGGA	GGGGCGGAC (4)	417	GGCGGCGGA	GGGGTGGAA (5)
418	GGCGGCGGA	GGGGCGGAC (4)	419	GGCGGCGGA	GGGGCGGAC (4)	420	GACGCTGCT	GACGCCGCT (8)
421	GACGCTGCT	GACGCCGCT (8)	422	GACGCTGCT	GACGCCGCT (8)	423	GACGCTGCT	GACGCTGCC (8)
424	GACGCTGCT	GACGCCGCT (8)	425	GACGCTGCT	GACGCTGCC (8)	426	GACGCTGCT	GACGCAGCC (7)
427	GACGCTGCT	GACGCCGCC (7)	428	GACGCTGCT	GACGCCGCT (8)	429	GACGCTGCT	GACGCCGTC (6)
430	GAGTGAGGA	GAGTGAGGG (8)	431	GAGTGAGGA	GAGTGAGGG (8)	432	GAGTGAGGA	GAGTGAGGG (8)
433	GAGTGAGGA	GAGTGAGGG (8)	434	GAGTGAGGA	GAGTGAGGC (8)	435	GAGTGAGGA	GAGGGAGGC (7)
436	GAGTGAGGA	GAGTGAGGG (8)	437	GAGTGAGGA	GAGTGAGGG (8)	438	GAGTGAGGA	GAGTGAGGC (8)
439	GAGTGAGGA	GAGTGAGGC (8)	440	GAGTGAGGA	GAGTGAGGC (8)	441	GAGTGAGGA	GAGTGAGGC (8)
442	GGGGAGGAG	GGGGAGGAC (8)	443	GGGGAGGAG	GGGGAGGAG (9)	444	GGGGAGGAG	GGGGAGGAT (8)
445	GGGGAGGAG	GGGGAGGAC (8)	446	GGGGAGGAG	GGGGAGGAC (8)	447	GGGGAGGAG	GTGGAGGAT (7)
448	GGGGAGGAG	GTGGAGGAG (8)	449	GGGGAGGAG	GTCGAGGAG (7)	450	GGGGAGGAG	GGGGAGGAC (8)
451	GCGGCGGAC	GCGGCGGAA (8)	452	GCGGCGGAC	GGGGCGGAA (7)	453	GCGGCGGAC	GGGGCGGAA (7)
454	GCGGCGGAC	GGGGCGGAA (7)	455	GCGGCGGAC	GGGGCGGAA (7)	456	GCGGCGGAC	GGGGCGGAA (7)
457	GCGGCGGAC	GGGGCGGAA (7)	458	GCGGCGGAC	GGGGCGGAA (7)	459	GCGGCGGAC	GAGGCGGAA (7)
460	GCGGCGGAC	GGGGCGGAA (7)	461	GCGGCGGAC	GGGGCGGAA (7)	462	GCGGCGGAC	GGGGCGGAA (7)
463	GCCGCCGGC	GCCGCCGGC (9)	464	GCCGCCGGC	GCCGCCGGC (9)	465	GCCGCCGGC	GCCGCCGGC (9)
466	GCCGCCGGC	GCCGCCGGC (9)	467	GCCGCCGGC	GCCGCCGGC (9)	468	GCCGCCGGC	GCCGGCGGC (8)
469	GCCGCCGGC	GCCGGCGGC (8)	470	GCCGCCGGC	GCCGGCGGC (8)	471	GCCGCCGGC	GCCGGCGGC (8)
472	GCCGCCGGC	GCCGCCGGC (9)	473	GTGGACGCG	GTGGAAGGG (7)	474	GTGGACGCG	GTGGACGGG (8)
475	GTGGACGCG	GTGGACGGG (8)	476	GTGGACGCG	GTGGAAGGG (7)	477	GTGGACGCG	GTGGATGGC (6)
478	GTGGACGCG	GTGGACGGG (8)	479	GTGGACGCG	GTGGAATGT (5)	480	GTGGACGCG	GTGGACGGG (8)
481	GTGGACGCG	GAGGAAGGG (6)	482	GTGGACGCG	GTGGAAGCG (8)	483	GTGGACGCG	GGGGAAGCC (6)
484	GCCGCTGGG	GCCGCTTGG (8)	485	GCCGCTGGG	GCCGCTTGG (8)	486	GCCGCTGGG	GCCGCTTGG (8)
487	GCCGCTGGG	GCCGCTTGG (8)	488	GCCGCTGGG	GACGCTTGG (7)	489	GCCGCTGGG	GACGCTTGG (7)
490	GCCGCTGGG	GACGCTTGG (7)	491	GCTGATGCC	GCCGATGCC (8)	492	GCTGATGCC	GCCGATGCC (8)
493	GCTGATGCC	GCCGATGCC (8)	494	GCTGATGCC	GCCGATGCC (8)	495	GCTGATGCC	GCCGATGCC (8)
496	GCTGATGCC	GCCGATGCC (8)	497	GCTGATGCC	GCCGACGCT (6)	498	GCTGATGCC	GCCGATGCC (8)
499	GCTGATGCC	GCCGATGCC (8)	500	GCTGATGCC	GCCGATGCC (8)	501	GCTGATGCC	GCCGATGCC (8)
502	GCGGCTGGG	GGGGCTTGG (7)	503	GCGGCTGGG	GCGGCTTGG (8)	504	GCGGCTGGG	GTGGCTTGG (7)
505	GCGGCTGGG	GGGGCTGGG (8)	506	GCGGCTGGG	GTGGCTTGG (7)	507	GCGGCTGGG	GGGGCTTGG (7)
508	GCGGCTGGG	GGGGCTTGG (7)	509	GCGGCTGGG	GGGGCTTGG (7)	510	GCGGCTGGG	GGGGCTGGG (8)
511	GCGGCTGGG	GGGGCTGGG (8)	512	GCGGCTGGG	GGGGCTGGG (8)	513	GAGTTTGCC	GACTTTGCC (8)
514	GAGTTTGCC	GATTAAGCC (6)	515	GAGTTTGCC	GACTTTGCC (8)	516	GAGTTTGCC	GACTGAGCC (6)
517	GAGTTTGCC	GACTTTGCC (8)	518	GAGTTTGCC	GATTTAGCT (6)	519	GAGTTTGCC	GATTTAGCC (7)
520	GAGTTTGCC	GACTGAGCC (6)	521	GAGTTTGCC	GATTTAGCC (7)	522	GAGTTTGCC	GACTGAGCC (6)
523	GAGTTTGCC	GATTTAGCC (7)	524	GAGTTTGCC	GATTTAGCC (7)	525	GTGGCTGGT	GTGGCTGGG (8)
526	GTGGCTGGT	GAGGCTGTA (6)	527	GTGGCTGGT	GAGGCTGTA (6)	528	GTGGCTGGT	GTGGCTGGG (8)
529	GTGGCTGGT	GAGGCTGTA (6)	530	GTGGCTGGT	GAGGCTGTA (6)	531	GTGGCTGGT	GTGGATGTT (7)
532	GTGGCTGGT	GAGGCTGGG (7)	533	GTGGCTGGT	GAGGCTGTA (6)	534	GTGGCTGGT	GAGGCTGTA (6)
535	GTGGCTGGT	GAGGCTGTA (6)	536	GGCGCCTAC	GTTGCCGAC (6)	537	GGCGCCTAC	TTGGTCGGG (2)
538	GGCGCCTAC	TTGGACTTT (3)	539	GGCGCCTAC	TTGGCCGGG (3)	540	GGCGCCTAC	TTGGCCTTC (5)
541	GGCGCCTAC	TTGGCCGGG (3)	542	GGCGCCTAC	GGAGCCTAC (8)	543	TGGGTGGCA	TGGGGGGCC (7)
544	TGGGTGGCA	TGGGGGGCC (7)	545	TGGGTGGCA	TGGGGGGCC (7)	546	TGGGTGGCA	TGGGGGGCC (7)
547	TGGGTGGCA	TGGGGGGCC (7)	548	TGGGTGGCA	TGGGGGGCC (7)	549	TGGGTGGCA	TAGGTGGCC (7)
550	TGGGTGGCA	TGGGGGGCC (7)	551	TGGGTGGCA	TAGGTGGCC (7)	552	TGGGTGGCA	TGGGCGGCC (7)
553	TGGGTGGCA	TAGGTGGCC (7)	554	TGGGTGGCA	TGGGTGGTC (7)	555	TGGGGTGCC	TGGGGCGCC (8)
556	TGGGGTGCC	TGGGGTGCC (9)	557	TGGGGTGCC	TGGGATGCT (7)	558	TGGGGTGCC	TGGGGCGCC (8)
559	TGGGGTGCC	TGGGGCGAC (7)	560	TGGGGTGCC	TGGGGCGCC (8)	561	TGGGGTGCC	TGGGGAGCT (7)
562	TGGGGTGCC	TGGGGCGAC (7)	563	TGGGGTGCC	TGGGGCGCC (8)	564	TGGGGTGCC	TGGGGCGCC (8)
565	TGGGGTGCC	TGGGGCGCC (8)	566	TGGGGTGCC	TGGGGAGCC (8)	567	TGGGAGTCT	TGGGAGTAA (7)
568	TGGGAGTCT	GTGGAGTAA (5)	569	TGGGAGTCT	GTGGAGTAA (5)	570	TGGGAGTCT	GTGGAGTAA (5)
571	TGGGAGTCT	TGGGAGTAT (8)	572	TGGGAGTCT	GGGGAGTAA (6)	573	TGGGAGTCT	TGGGAGTAA (7)
574	TGGGAGTCT	TTGGAGTAC (6)	575	TGGGAGTCT	GTGGAGGTA (4)	576	TGGGAGTCT	GTGGAGTAC (5)
577	GGGGAAGAG	GGGGAAGAC (8)	578	GGGGAAGAG	GGGGAAGAC (8)	579	GGGGAAGAG	GGGGAAGAC (8)
580	GGGGAAGAG	GGGGAAGAC (8)	581	GGGGAAGAG	GGGGAAGAC (8)	582	GGGGAAGAG	GGGGAAGAC (8)
583	GGGGAAGAG	GGGGAAGAC (8)	584	GGGGAAGAG	GGGGAAGAC (8)	585	GGGGAAGAG	GGGGAAGAC (8)
586	GGGGAAGAG	GGGGAAGAG (9)	587	GGGGAAGAG	GGGGAAGAC (8)	588	TCTGGCGCT	TTAGACGCT (6)
589	TCTGGCGCT	TTAGGCGCT (7)	590	TCTGGCGCT	TTAGCCGCA (5)	591	TCTGGCGCT	ACAGGTGTT (5)
592	TCTGGCGCT	GTTGGCGTA (5)	593	TCTGGCGCT	TTAGGCGCC (6)	594	TCTGGCGCT	TTAGATGCT (5)
595	TCTGGCGCT	ACAGGCGTC (5)	596	TCTGGCGCT	TTTGATGTC (4)	597	TCTGGTTTC	GTTGGCTGG (4)
598	TCTGGTTTC	TTAGGCGTT (4)	599	TCTGGTTTC	TTAGACTGG (3)	600	TCTGGTTTC	ACAGACTGG (3)
601	GAAGGATTC	GAAGGATGG (7)	602	GAAGGATTC	GATGGATGG (6)	603	GAAGGATTC	GAAGGATGG (7)
604	GAAGGATTC	GATGGATGG (6)	605	GGCGGAGAT	GGCGGATTT (7)	606	GGCGGAGAT	GGCGGATGT (7)
607	GGCGGAGAT	GGCGGATTT (7)	608	GGCGGAGAT	TTGGGAGTT (5)	609	GGCGGAGAT	GGCGGATTT (7)
610	GGCGGAGAT	GGCGGATGT (7)	611	GGCGGAGAT	GGCGGATTT (7)			

Table S1: (cont.)