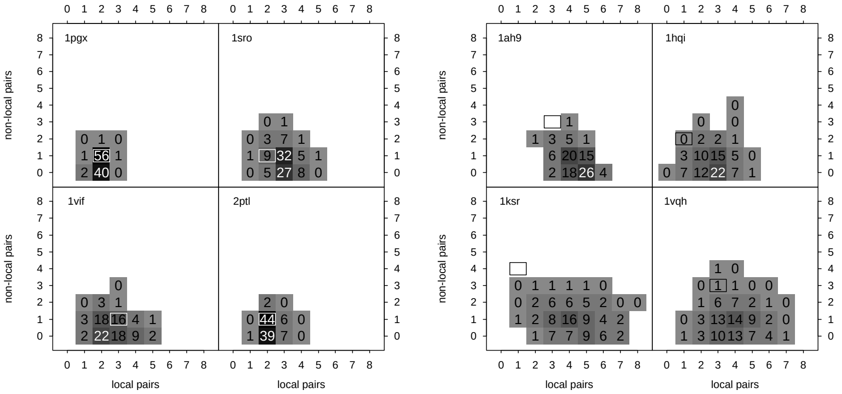
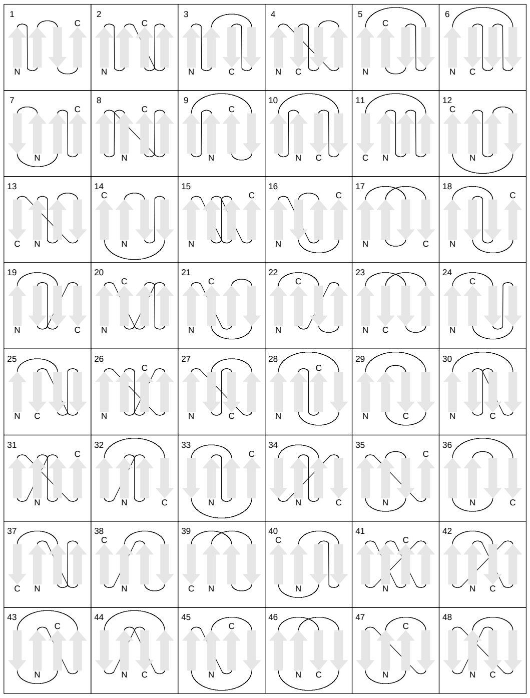
One Bad Property of Rosetta Decoys

Rosetta predominantly generates very local structures:



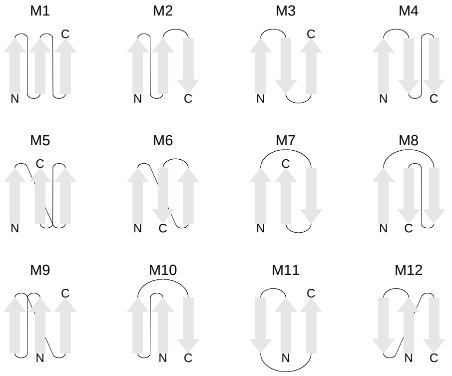
The distributions of local versus non-local strand pairs in decoy sets generated by Rosetta (size 10000-15000 each) for eight different proteins. A local strand pair is defined as a pair of strands adjacent in sequence, along the backbone of the protein, that are neighbours in the sheet. A pair of strand neighbours not adjacent in sequence is called a non-local strand pair. The number of local and non-local strand pairs for the native folds are indicated by a square in the respective panels. These numbers are rounded percentages of the frequency of decoys with the respective number of local and non-local strand pairs. A zero therefore stands for a percentage p with $0 < p < 0.5\%$, while cells without numbers represent motifs that never occurred in the decoy set.

Open Sheets in Globular Proteins



We surveyed the distribution of $\beta-$sheet motifs with two edge strands (open sheets) in the database of non-homologous globular proteins. For example, the figure to the left shows the four-stranded motifs that did not occur in the database. There are a total of 17 motifs which violate one of four absolute rules. Assuming that all loops between parallel strand pairs are right-handed, there can be a clash between two crossings connecting pairs of parallel strands in some motifs (panels 4, 13, 26 and 31). The spatial strand sequence '2413' never occurs in sheets (panels 41–48), and neither do “pretzels”, which are motifs that have crossing loops (panels 17, 23, 39 and 46). The motifs in panels 29 and 36 are named spirals.

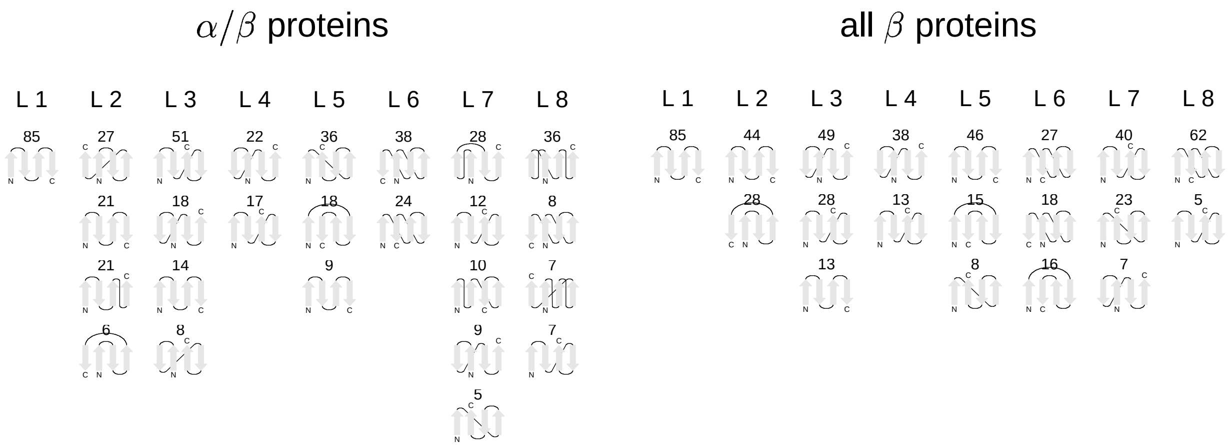
Sheets with Three Strands



	α/β				all β			
	L_1	L_2	L_3	L_4	L_1	L_2	L_3	L_4
M_1	.004	.006	.005	.049	.004	.006	.005	.042
M_2	.004	.006	**.083**	**.080**	.004	.006	**.083**	**.068**
M_3	**.897**	**.401**	**.276**	**.162**	**.897**	**.611**	**.422**	**.252**
M_4	.004	**.262**	.005	.029	.004	.042	.005	.024
M_5	.004	.006	.005	.019	.004	.006	.005	.016
M_6	.036	.012	**.547**	**.282**	.036	.012	**.401**	**.239**
M_7	.004	.006	.048	.032	.004	.006	.048	.027
M_8	.004	.006	.005	.014	.004	.006	.005	.012
M_9	.004	.006	.005	**.114**	.004	.006	.005	.012
M_{10}	.004	.006	.005	.035	.004	.006	.005	.030
M_{11}	.004	.027	.005	.032	.004	.028	.005	.027
M_{12}	.028	**.259**	.010	**.153**	.028	**.267**	.010	**.252**

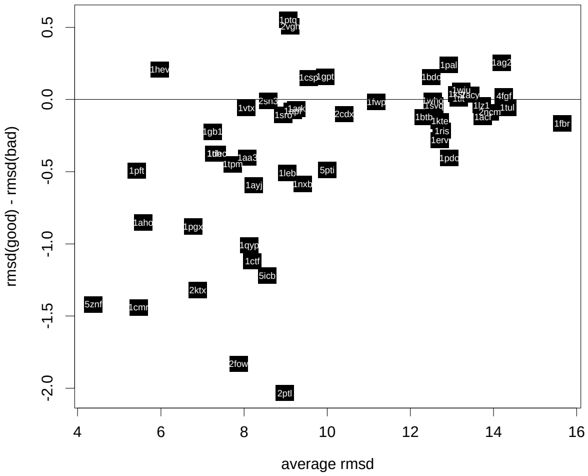
The fitted probabilities for three-stranded motifs in α/β and all β proteins, conditional on loop lengths. A short loop has ten or less residues, a long loop has more than ten residues. The loop lengths between the strands are short-short (L_1), short-long (L_2), long-short (L_3) and long-long (L_4). The motifs $M_1 - M_{12}$ are shown in the left figure. Probabilities of more than 5% are highlighted in bold fonts in the table.

Sheets with Four Strands



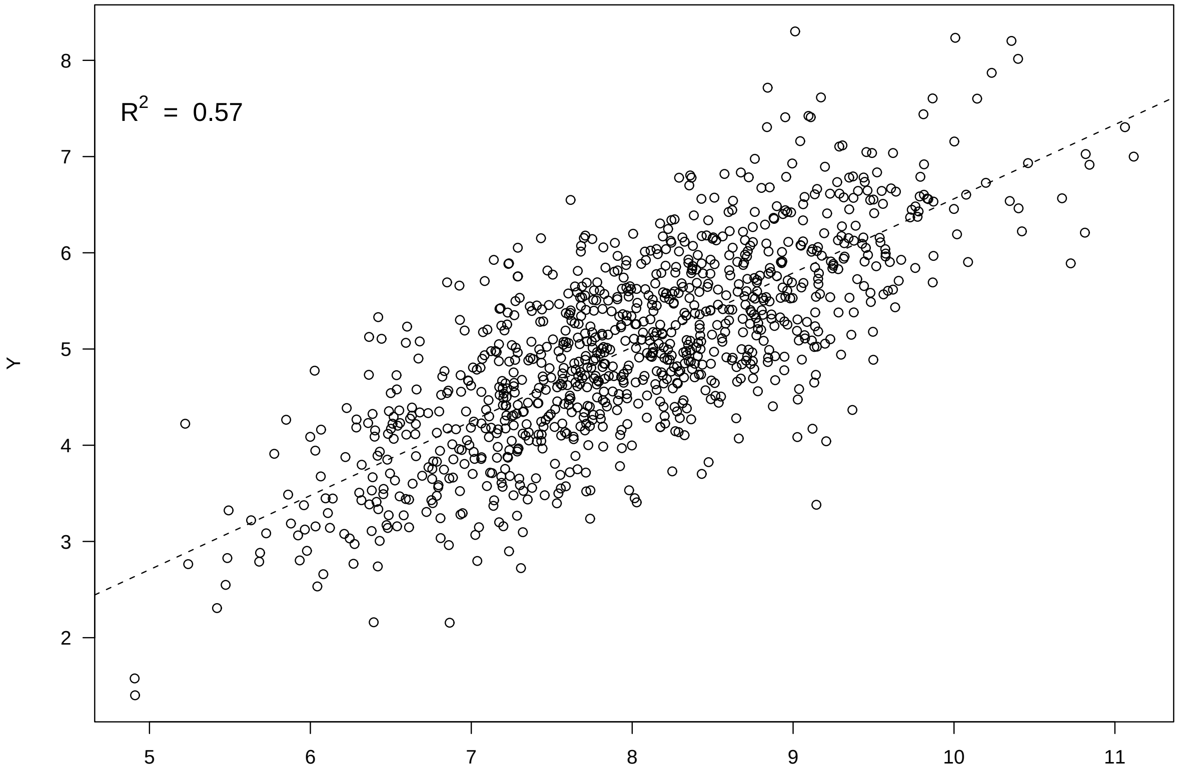
Four-stranded motifs with probabilities larger than 5%. The actual probabilities (rounded, in percent) are indicated above the motifs. L1 through L8 refers to the loop length classes, defined as SSS, SSL, SLS, SLL, LSS, LSL, LLS, LLL (S: short, L: long).

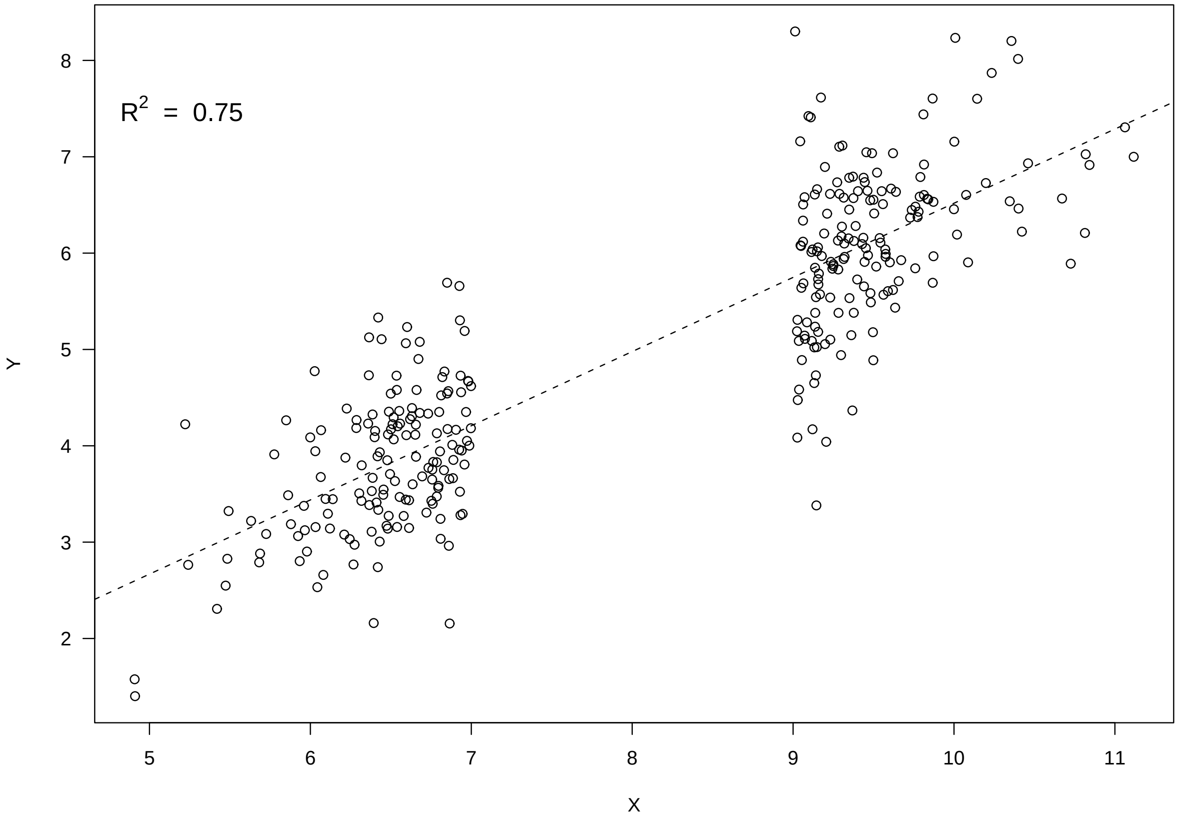
Filtering Decoys



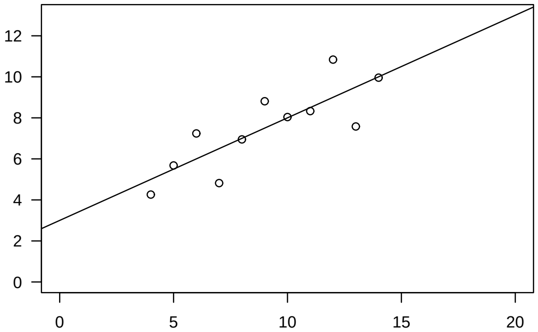
To assess the $\beta-$sheet motifs in the decoy sets, a filter was implemented that checked for proper sheet conformations (no unpaired strand, reasonable angles between strands, etc) in the generated structures. The filter alone already improved the quality of the decoy sets considerably!

The improvement in average rmsd (root mean squared deviance between the C_α atoms) achieved for 52 small proteins. On the x-axis is the average rmsd for the complete generated decoy set, on the y-axis is the difference in average rmsd between those decoys that passed the filter and those that did not. Improvements are achieved for almost all decoy sets with less than 10Å average rmsd.

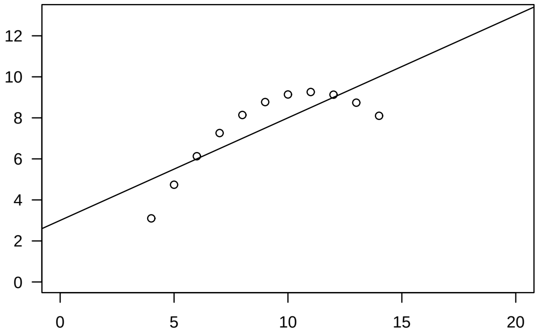




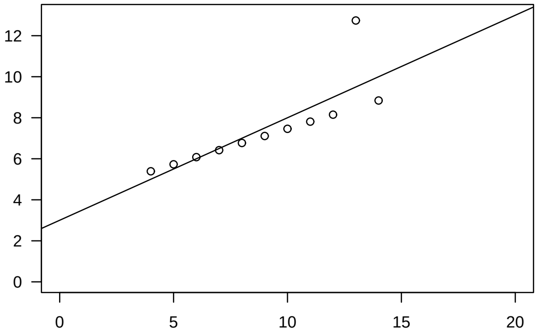
$\hat{\beta}_0$=3.0 $\hat{\beta}_1$=0.5 $\hat{\sigma}^2$=13.75 R^2=0.667



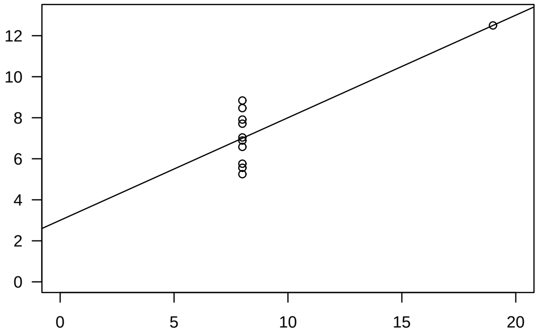
$\hat{\beta}_0$=3.0 $\hat{\beta}_1$=0.5 $\hat{\sigma}^2$=13.75 R^2=0.667

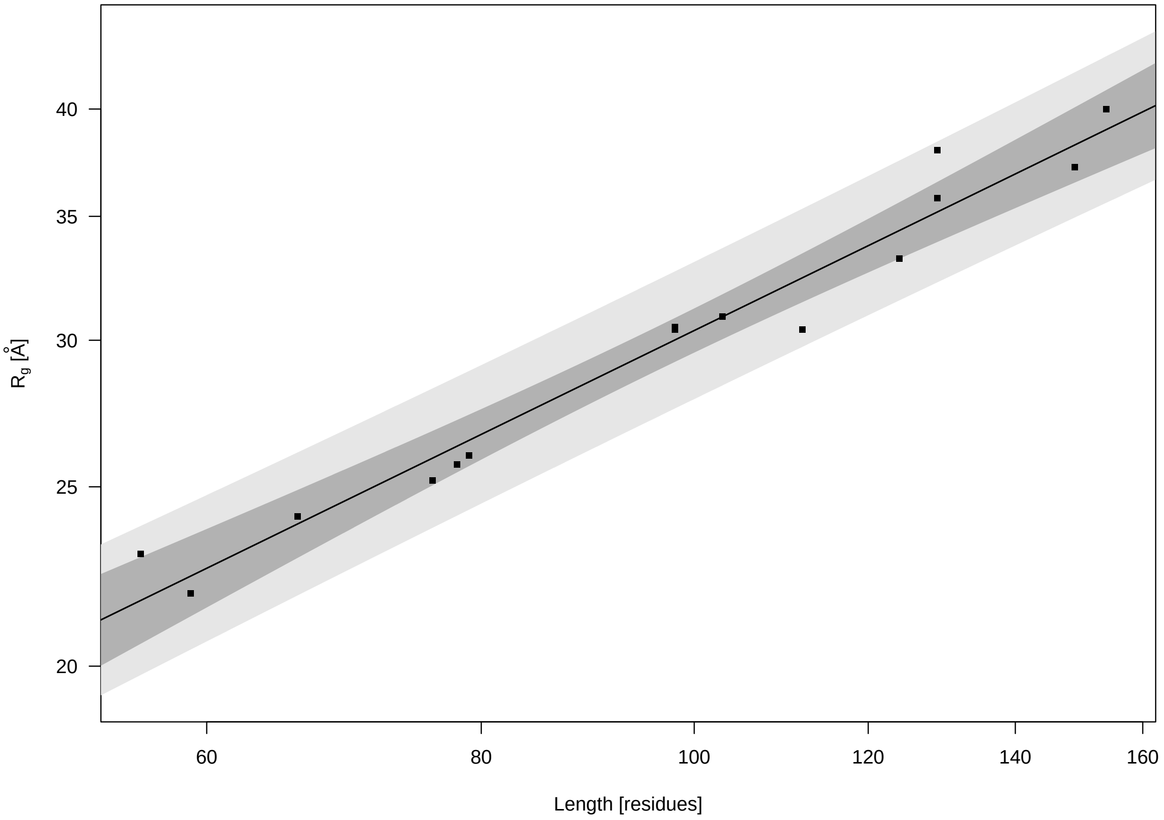


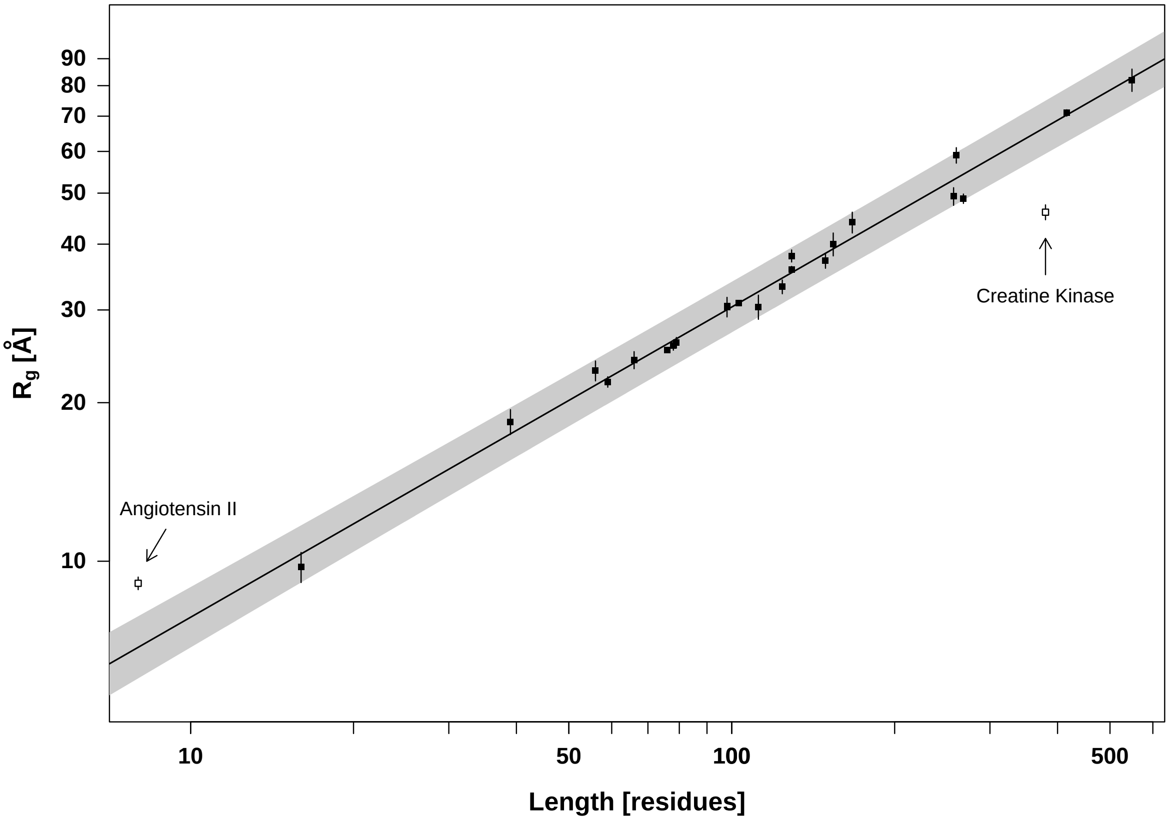
$\hat{\beta}_0$=3.0 $\hat{\beta}_1$=0.5 $\hat{\sigma}^2$=13.75 R^2=0.667



$\hat{\beta}_0$=3.0 $\hat{\beta}_1$=0.5 $\hat{\sigma}^2$=13.75 R^2=0.667




40
35
30
25
20
R_g [Å]
60
80
100
120
140
160
Length [residues]


90
80
70
60
50
40
30
20
10
R_g [Å]
10
50
100
500
Length [residues]
Angiotensin II
Creatine Kinase