# Distributions of Beta Sheets in Proteins with Application to Structure Prediction

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

http://biostat.jhsph.edu/∼iruczins

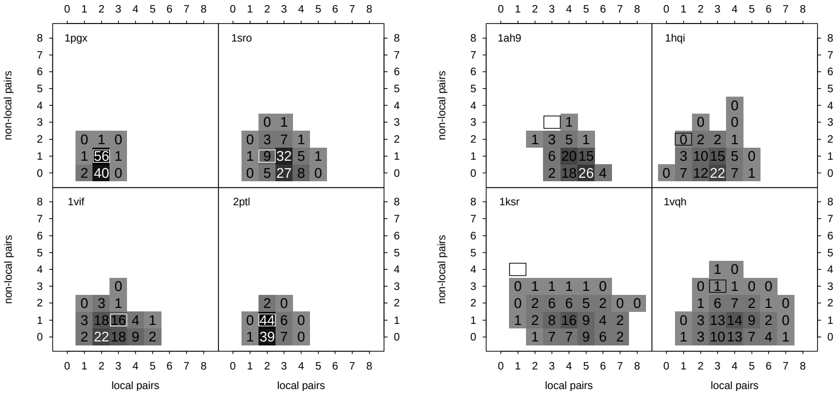
| A collaboration with: | David Baker | (University of Washington) |
|---|---|---|
| | Richard Bonneau | (Institute for Systems Biology) |
| | Charles Kooperberg | (Fred Hutchinson Cancer Research Center). |

## Abstract

We recently developed the Rosetta algorithm for ab initio protein structure prediction which generates protein structures from fragment libraries using simulated annealing. The scoring function in this algorithm favors the assembly of strands into sheets. However, it does not discriminate between different sheet motifs. After generating many structures using Rosetta, we found that the folding algorithm predominantly generates very local structures. We surveyed the distribution of $\beta-$sheet motifs with two edge strands (open sheets) in a large set of non-homologous proteins. We investigated how much of that distribution can be accounted for by rules previously published in the literature, and developed a scoring method that enables us to improve protein structure prediction for $\beta-$sheet proteins.

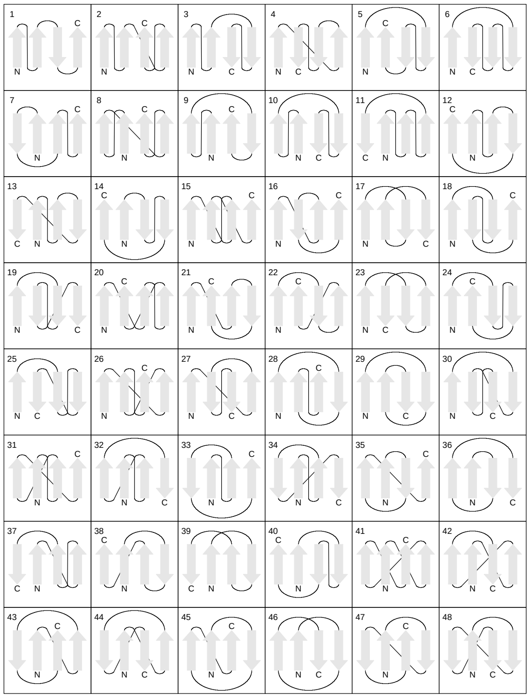
# One Bad Property of Rosetta Decoys

## Rosetta predominantly generates very local structures:



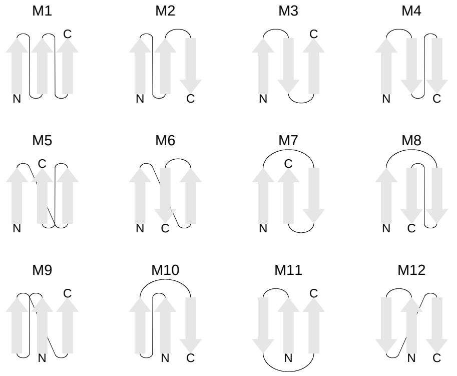
The distributions of local versus non-local strand pairs in decoy sets generated by Rosetta (size 10000-15000 each) for eight different proteins. A local strand pair is defined as a pair of strands adjacent in sequence, along the backbone of the protein, that are neighbours in the sheet. A pair of strand 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.

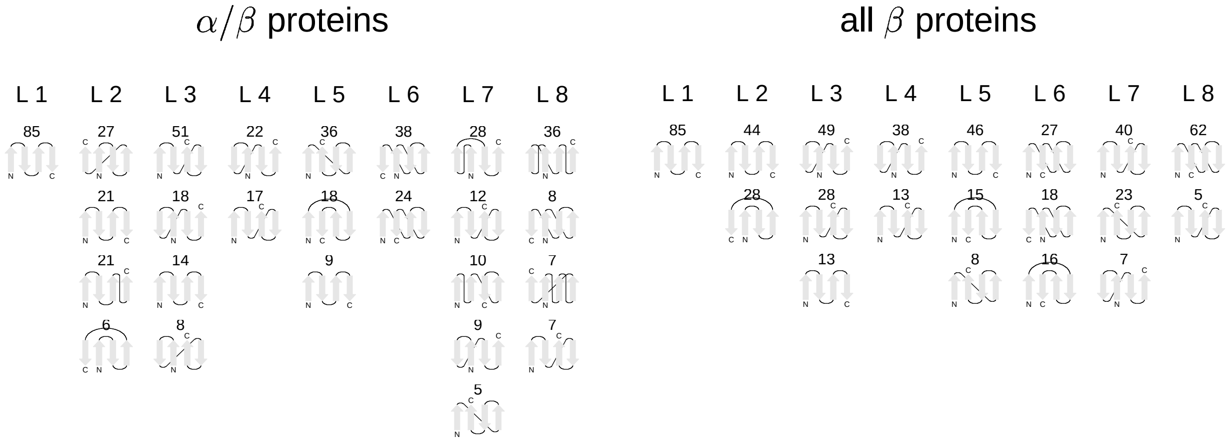
# Sheets with Three Strands



| | $\alpha/\beta$ | | | | all $\beta$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_4$ | $L_1$ | $L_2$ | $L_3$ | $L_4$ |
| $M_1$ | .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 $\alpha/\beta$ and all $\beta$ 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.

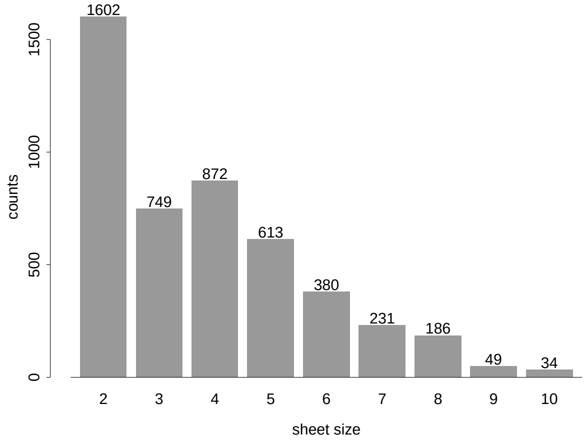
# Sheets with Four Strands



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).

# Modeling Sheets with Five or More Strands

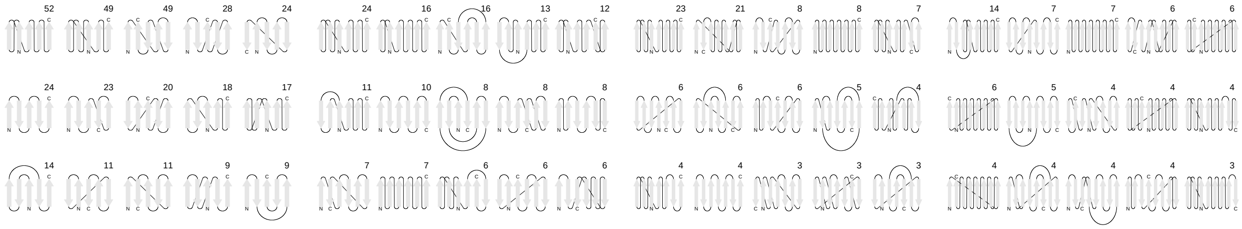
Bad news: the sheet counts decrease with the number of strands!



Shown above are the counts of sheets of sizes 2–10, observed in non-homologous proteins in the PDB data bank.

# Modeling Sheets with Five or More Strands

Good news: the motifs definitely have some regular patterns!



5 strands 6 strands 7 strands 8 strands

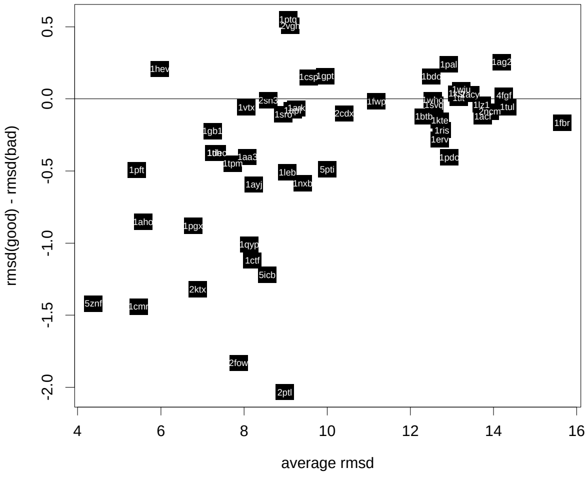
Clearly favored are parallel and anti-parallel $\beta-$sheet motifs, as well as motifs with high numbers of sequentially adjacent strands that are also spatially adjacent (motifs with many local strand pairs, i. e. few “jumps”).

# Modeling Sheets with Five or More Strands

The fine print: Let $H$ be the helical status of the protein ($\alpha/\beta$ versus all $\beta$) and let $L$ be the loop length distribution between the $n$ strands (expresses as long or short). Let $P_p$ be the number of parallel neighbour strands in a motif, $P_p^s$ the number of parallel neighbour strands in a motif with a short loop in between, $J$ the number of jumps, $J^s$ the number of jumps with a short loop between the strand pair, and $F$ the position of the first strand in the motif. Then

$$\begin{aligned}
& P(P_p, P_p^s, J, J^s, F|n, H, L) \\
= \; & P(F|n, H, L) \times P(P_p, P_p^s, J, J^s|n, H, L, F) \\
= \; & P(F|n, H, L) \times P(P_p, J|n, H, L, F) \times P(P_p^s, J^s|n, H, L, F, P_p, J) \\
\approx \; & P(F|n, H) \times P(P_p, J|n, H, L, F) \times P(P_p^s|n, H, L, P_p) \times P(J^s|n, H, L, J)
\end{aligned}$$

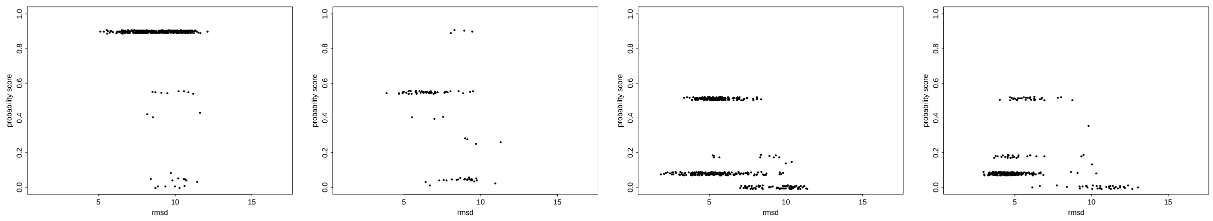
# Filtering Decoys



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 $C_\alpha$ 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.

# Scoring Decoys



SH3 domain
(1ark)

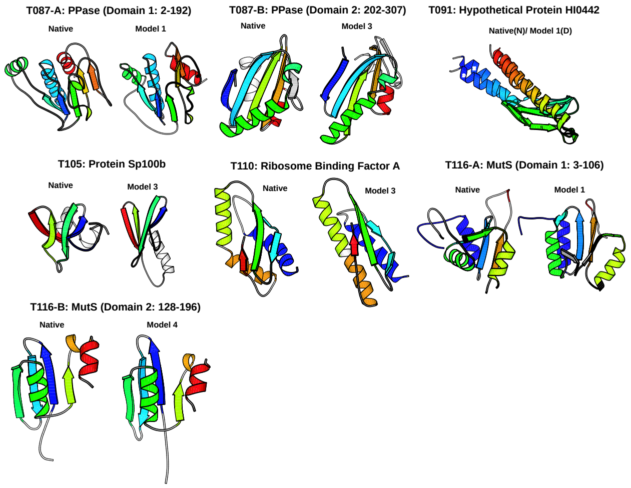
Raphanus Anti-fungal
(1ayj)

Protein G
(1gb1)

Protein L
(2ptl)

Rmsd versus motif distribution score. For ease of visualization, motif scores were slightly jittered. The SH3 domain and Rs-afp1 have a single three-stranded sheet each, Protein G and Protein L have a single four-stranded sheet each. Shown are only the decoys with the correct number of strands in the respective structure. Since the scoring function reflects the frequency how often sheet motifs occur in native proteins, motifs that score close to zero are usually in decoys with very high rmsds, and the elimination of those further improves the quality of the decoy sets. Note that all decoys with rmsd smaller than say 5.5 Å have probability scores larger than 5%.

# Predicting Protein Structure



T087-A: PPase (Domain 1: 2-192)
Native
Model 1
T087-B: PPase (Domain 2: 202-307)
Native
Model 3
T091: Hypothetical Protein HI0442
Native(N)/ Model 1(D)
T105: Protein Sp100b
Native
Model 3
T110: Ribosome Binding Factor A
Native
Model 3
T116-A: MutS (Domain 1: 3-106)
Native
Model 1
T116-B: MutS (Domain 2: 128-196)
Native
Model 4

The previously described procedures were used in the CASP protein structure prediction experiment to successfully predict $\beta-$sheet proteins with unprecedented accuracy. Above are a few examples of predicted model and native fold, determined by X-ray crystallography or NMR.

# Conclusion

We surveyed the distribution of $\beta-$sheet motifs with two edge strands (open sheets) in the database of non-homologous proteins, and examined deterministic and probabilistic rules for sheet motif distributions. We used the results of our survey to develop a full scoring function of sheet motifs for both $\alpha/\beta$ and all $\beta$ proteins, that also takes the loop lengths between the strands into account. This scoring function, paired with a filter to eliminate structures with poor sheet configurations, proved to be valuable in ab initio structure prediction. The filter and the scoring function might become even more important in the future, since we hope to be able to use the increase in computer power to create Rosetta decoys with larger and more non-local sheets. We modeled the distributions of $\beta-$sheets, but the physical origins of those distributions remain unclear. For example, we have no explanation for the fact that some motifs were not in the database we investigated, although some very similar motifs occur frequently in nature. This puzzle poses an interesting challenge for current protein design methods.

# References

- Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, and Baker D (1999), *Improved Recognition of Native-like Protein Structures using a Combination of Sequence-dependent and Sequence-independent Features of Proteins*, Proteins: Structure, Function and Genetics 34 (1) 82-95.
- Bonneau R, Tsai J, Ruczinski I, and Baker D (2001), *Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction*, Proteins: Structure, Function and Genetics 45 (S5), 119-126.
- Ruczinski I, Kooperberg C, Bonneau R, and Baker D (2002), *Distributions of Beta Sheets in Proteins with Application to Structure Prediction*, Proteins: Structure, Function and Genetics 48, 85-97.
- Bonneau R, Tsai J, Ruczinski I, and Baker D (2002), *Contact Order and Ab Initio Protein Structure Prediction*, Protein Science 11 (8), 1937-1944.