Flexible Backbone Design with the Sequence Tolerance Protocol

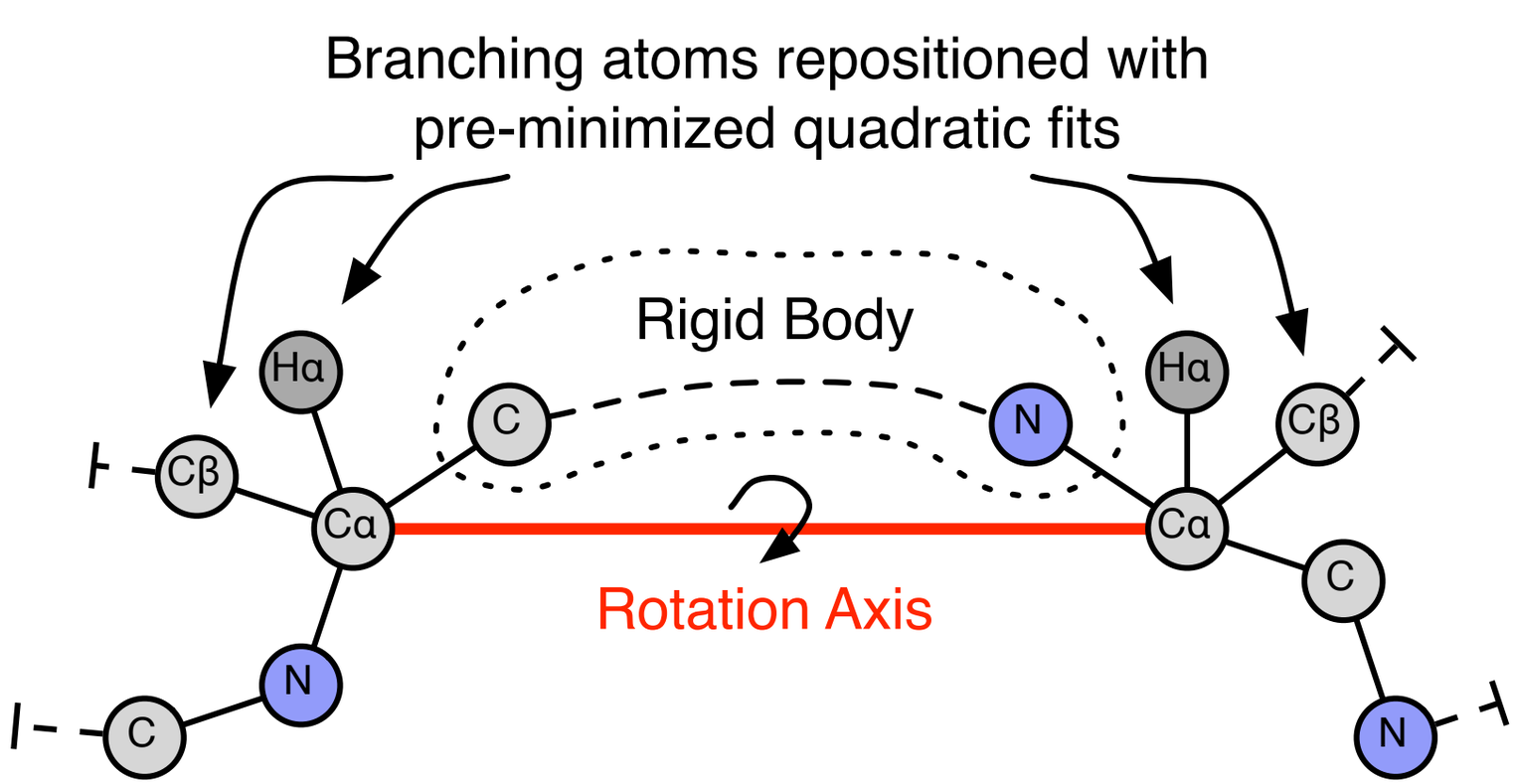
Colin Smith & Amelie Stein
RosettaCon 2012 - July 31, 2012

Why Have an Ensemble?

- Accounting for natural flexibility
- phi/psi dependence of amino acid probability
- Iterative methods often get trapped
- Adding premutation of starting structure may require structural relaxation

Backrub Ensembles

- Apply 10,000 Backrub/side chain moves (more may be necessary)
- Make an ensemble of 20-200 backbones



Starting Structure (2QMT)

Backrub Ensemble

Process Individual Backbones

...

Process Individual Backbones

...

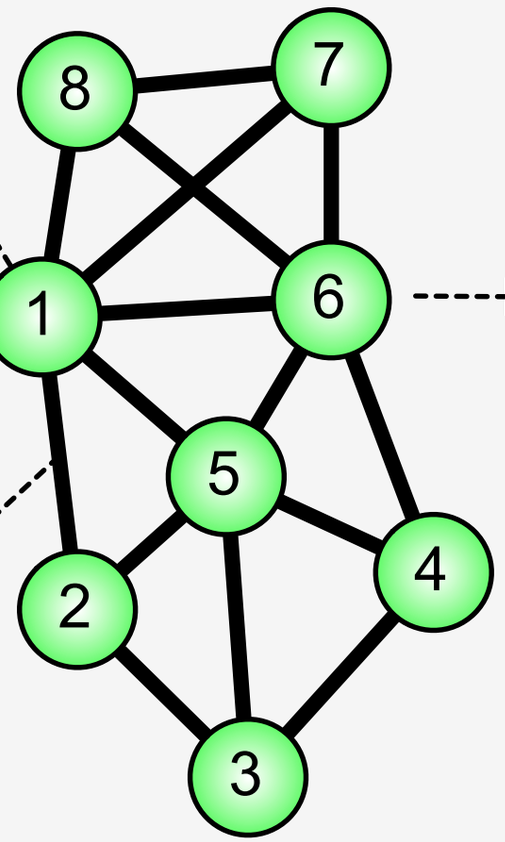
Precalculate Interaction Graph

One Body Energies

AA Type	Designed Residue Position 1								
	Ser			Thr			...		
Rotamer #	1	2	...	1	2	...	1	2	...

Two Body Energies

AA Type / Rotamer # / AA Type		Designed Residue Position 2								
		Ser			Thr			...		
		1	2	...	1	2	...	1	2	...
Designed Residue Position 1	Ser 1									
	Ser 2									
	Ser ...									
	Thr 1									
	Thr 2									
	Thr ...									
	... 1									
	... 2									
	... ...									

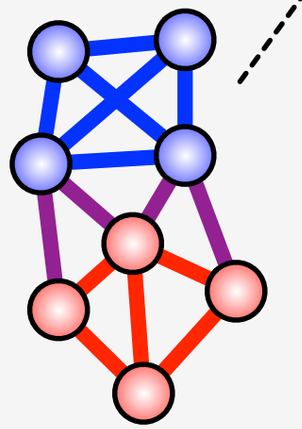


Genetic Algorithm Sequence Sampling

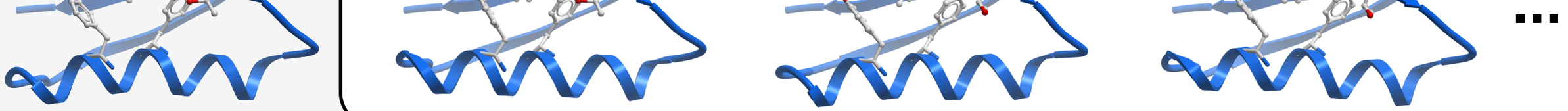
Mutation and Crossover

Simulated Annealing

Reweight Group (**A** & **B**) and Group Interaction (**A-B**) Energies



Sequence 1, Fitness 1
Sequence 2, Fitness 2
Sequence 3, Fitness 3
Sequence 4, Fitness 4
...



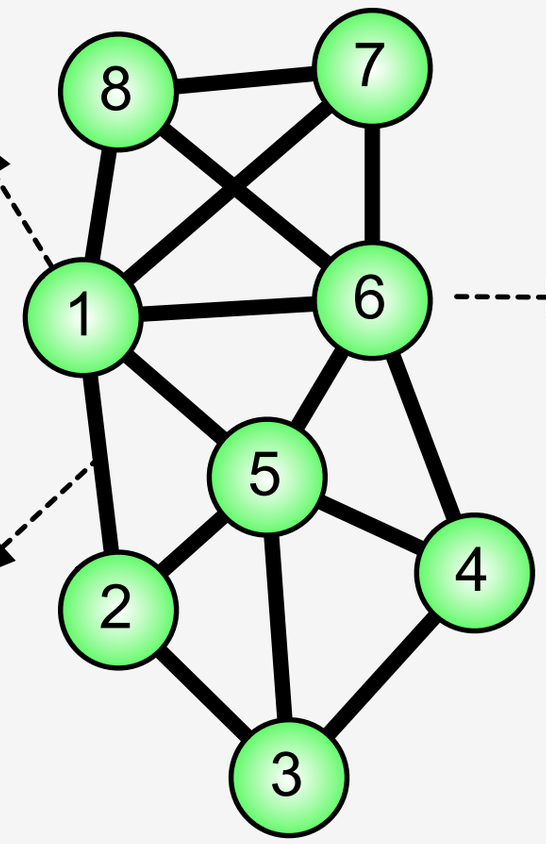
Precalculate Interaction Graph

One Body Energies

AA Type	Designed Residue Position 1								
	Ser			Thr			...		
Rotamer #	1	2	...	1	2	...	1	2	...

Two Body Energies

AA Type	AA Type	Designed Residue Position 2								
Rotamer #	AA Type	Ser			Thr			...		
		1	2	...	1	2	...	1	2	...
Designed Residue Position 1	Ser	1								
		2								
		...								
	Thr	1								
		2								
		...								
	...	1								
		2								
		...								

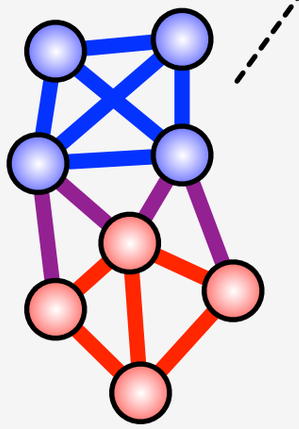


Genetic Algorithm Sequence Sampling

Mutation and Crossover

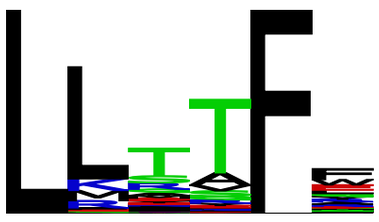
Simulated Annealing

Reweight Group (A & B) and Group Interaction (A-B) Energies

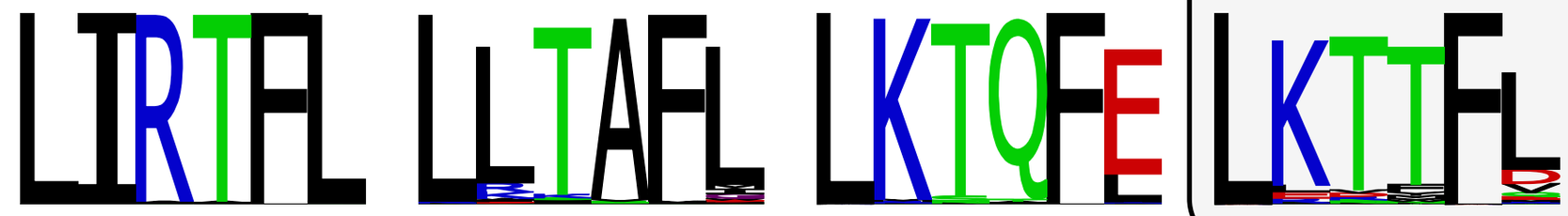


Sequence 1, Fitness 1
Sequence 2, Fitness 2
Sequence 3, Fitness 3
Sequence 4, Fitness 4
...

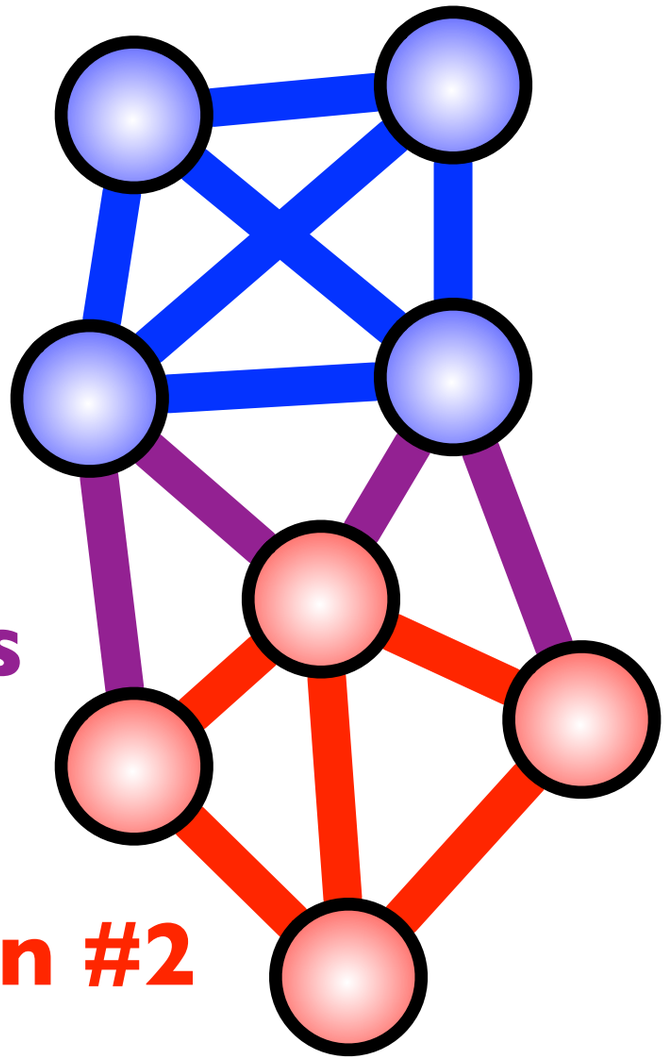
Combined Position Weight Matrix



Boltzmann Weight Sequences from Individual Backbones



Fitness Function



Backbone-Backbone Hydrogen Bonds

Monomer

BBHB, #1

Published: **0.4, 0.4**

Dimer

BBHB, #1, #2, #1-2

Published: **0.4, 0.4, 0.4, 1**

Trimer

BBHB, #1, #2, #3

#1-2, #1-3, #2-3

Published: **?**

Running the Sequence Tolerance Protocol

- Input files
- Python Script
- Parameters

```
# This function is the main data processing procedure. It takes a directory
# path which contains *.ga.entities files. It reads all those files and
# produces a set of boxplots in several different file formats. It also
# generates a position weight matrix and FASTA file for producing a sequence
# logo. By specifying plotgen=TRUE, it will produce a plot similar to
# Figure 5 in the PLoS One manuscript.

process_seqtol <- function(dirpath = ".", fitness_coef = c(1/2.5, 1/2.5, 1/2.5, 1),
                           temp_or_thresh = 0.228,
                           type = c("boltzmann", "cutoff"),
                           percentile = .5, prefix = "seqtol",
                           plotgen = FALSE, plotseq = TRUE) {
...
```

```
> source("path/to/sequence_tolerance.R")
> process_seqtol(".", c(1/2.5, 1/2.5))
```

This is also equivalent:

```
> process_seqtol(fitness_coef=c(1/2.5, 1/2.5))
```

http://cran.r-project.org/doc/manuals/R-intro.html

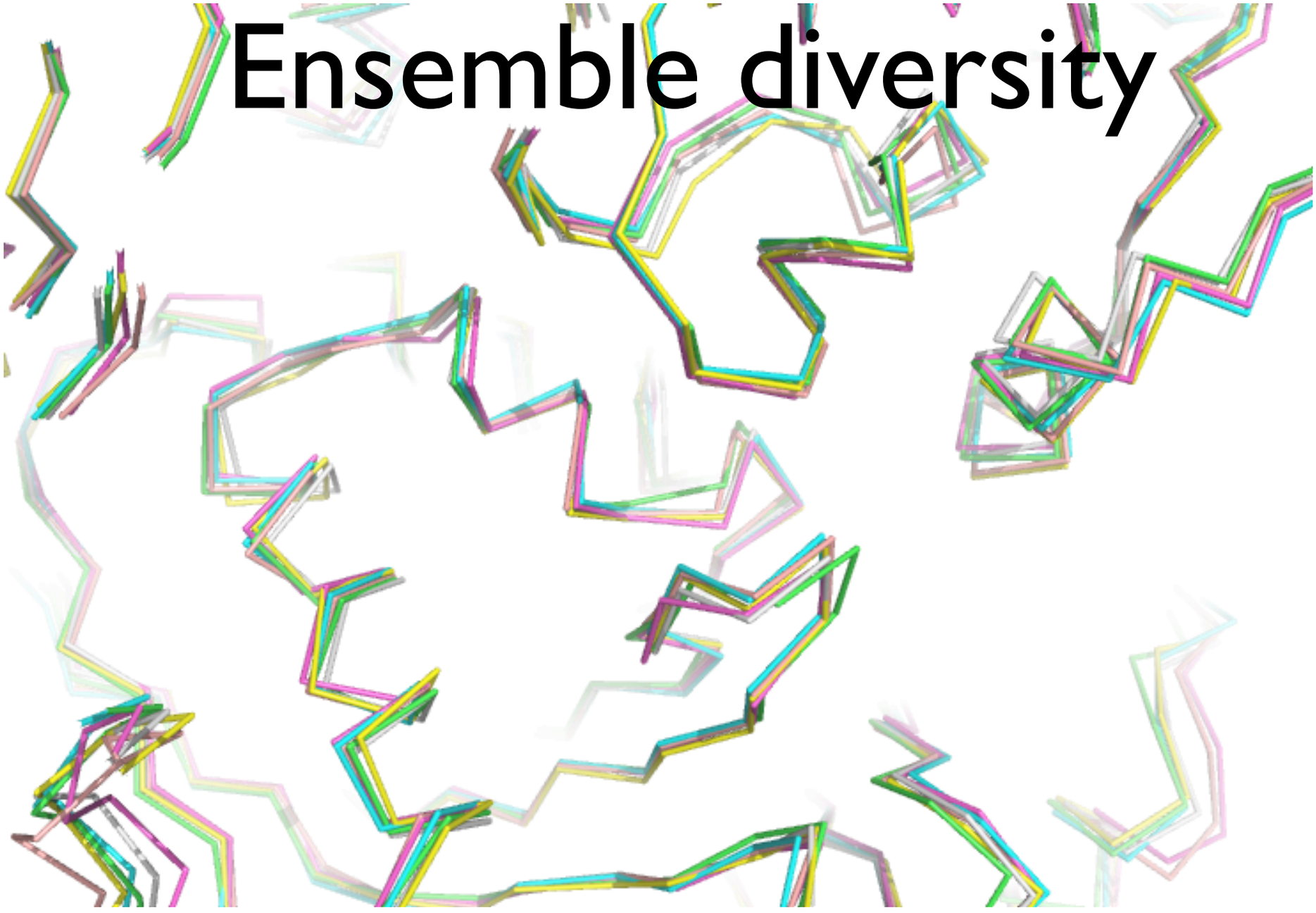
Changing Sequence Variability in Post Processing

- Temperature for Boltzmann weighting (or down-weighting fitness function coefficients)
- Percentile (probably not worth touching)

Alter Sequence Variability in Ensemble Generation

- Backrub low vs. last
- Backrub temperature
- KIC
- Relax

Ensemble diversity



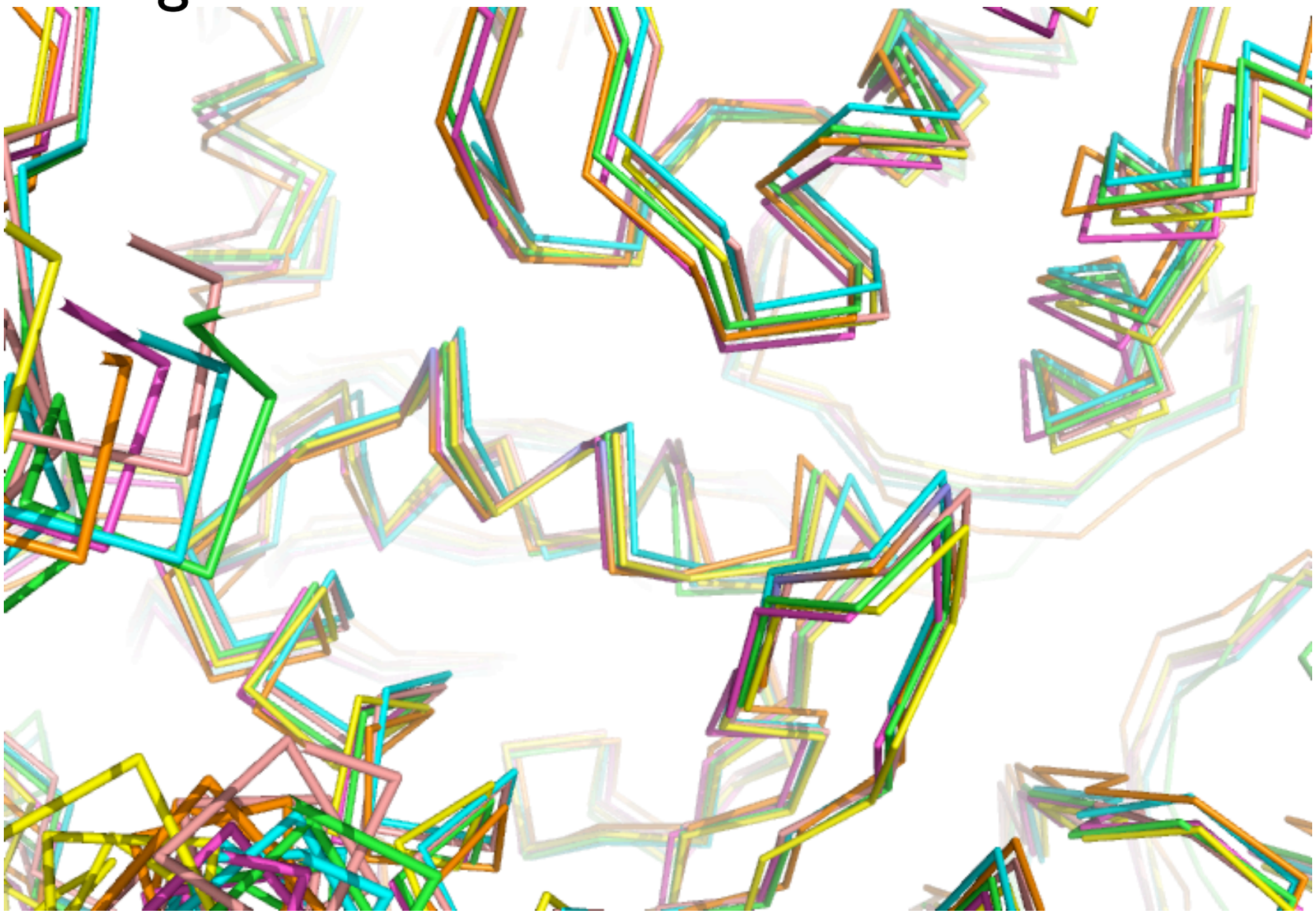
- if your results indicate very sharp residue preferences at each position or are very flat, check the ensemble variability
- large structures or badly scoring native conformations may need tweaking

Ensemble generation with vicinity KIC

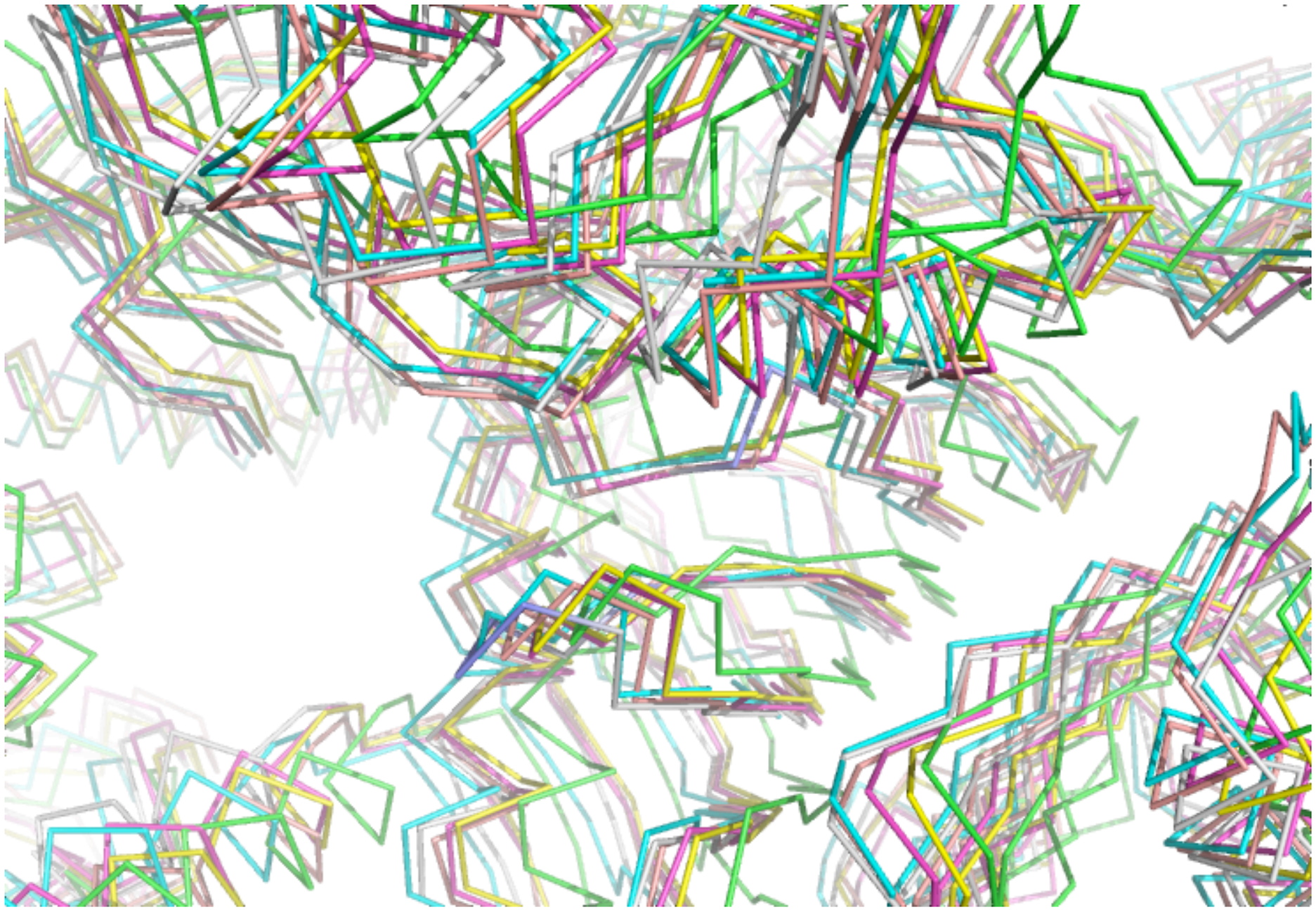
- you'll need a loop covering each protein individually:
  `LOOP (start+1) (end-1)`

- key command line changes:

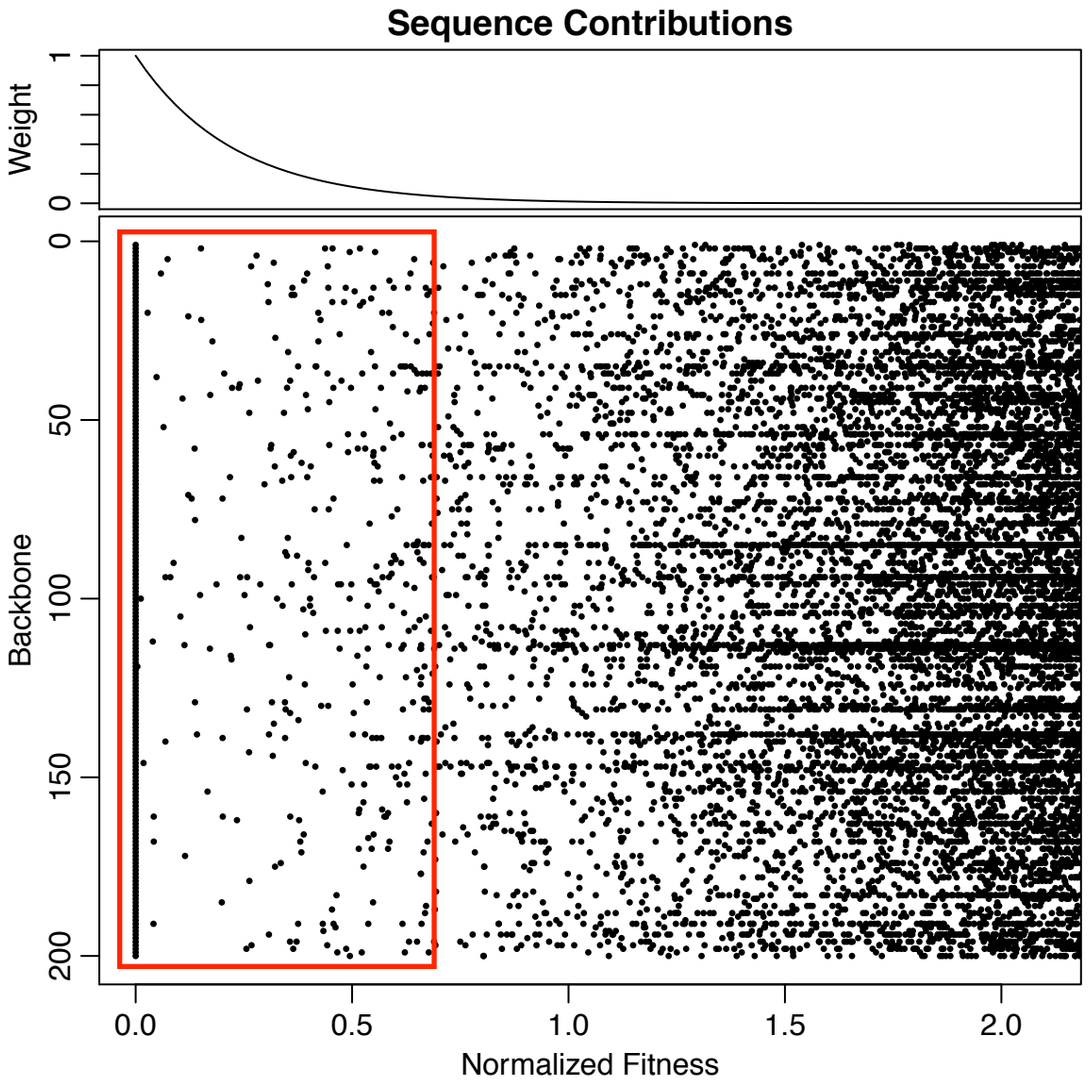
```
loopmodel.linuxgccrelease
-in:file:fullatom
-loops:refine refine_kic
-loops:outer_cycles 1
-loops:refine_init_temp 1.2
-loops:refine_final_temp 1.2
-loops:vicinity_sampling true
-loops:vicinity_degree 3
```

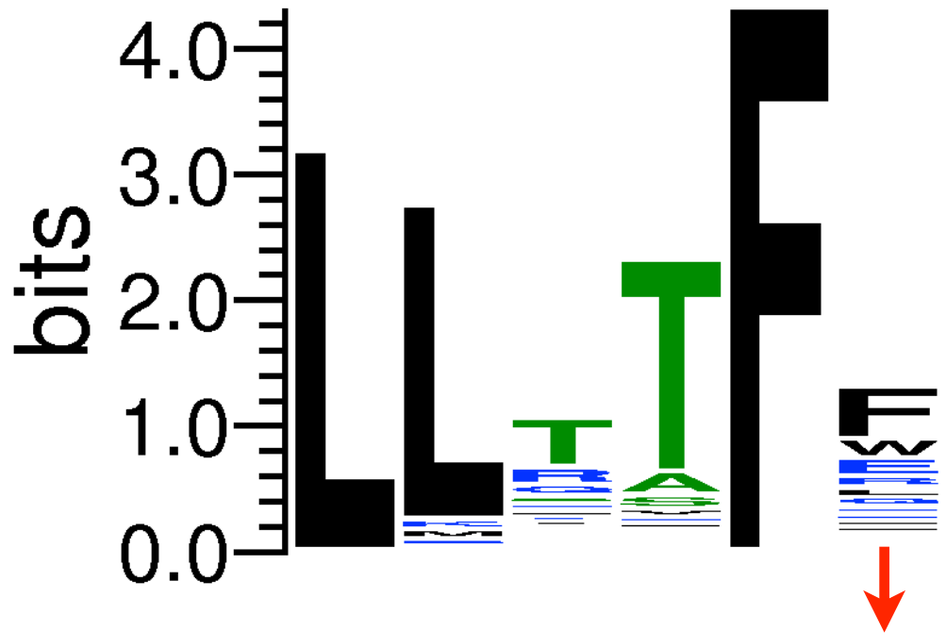


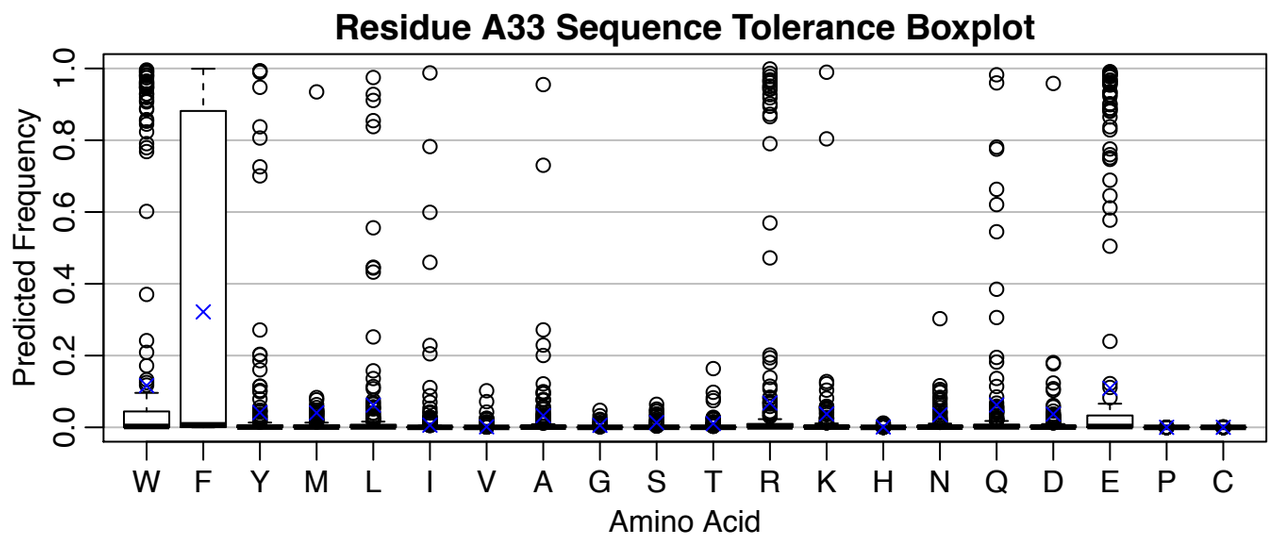
For even more diversity, try fastrelax



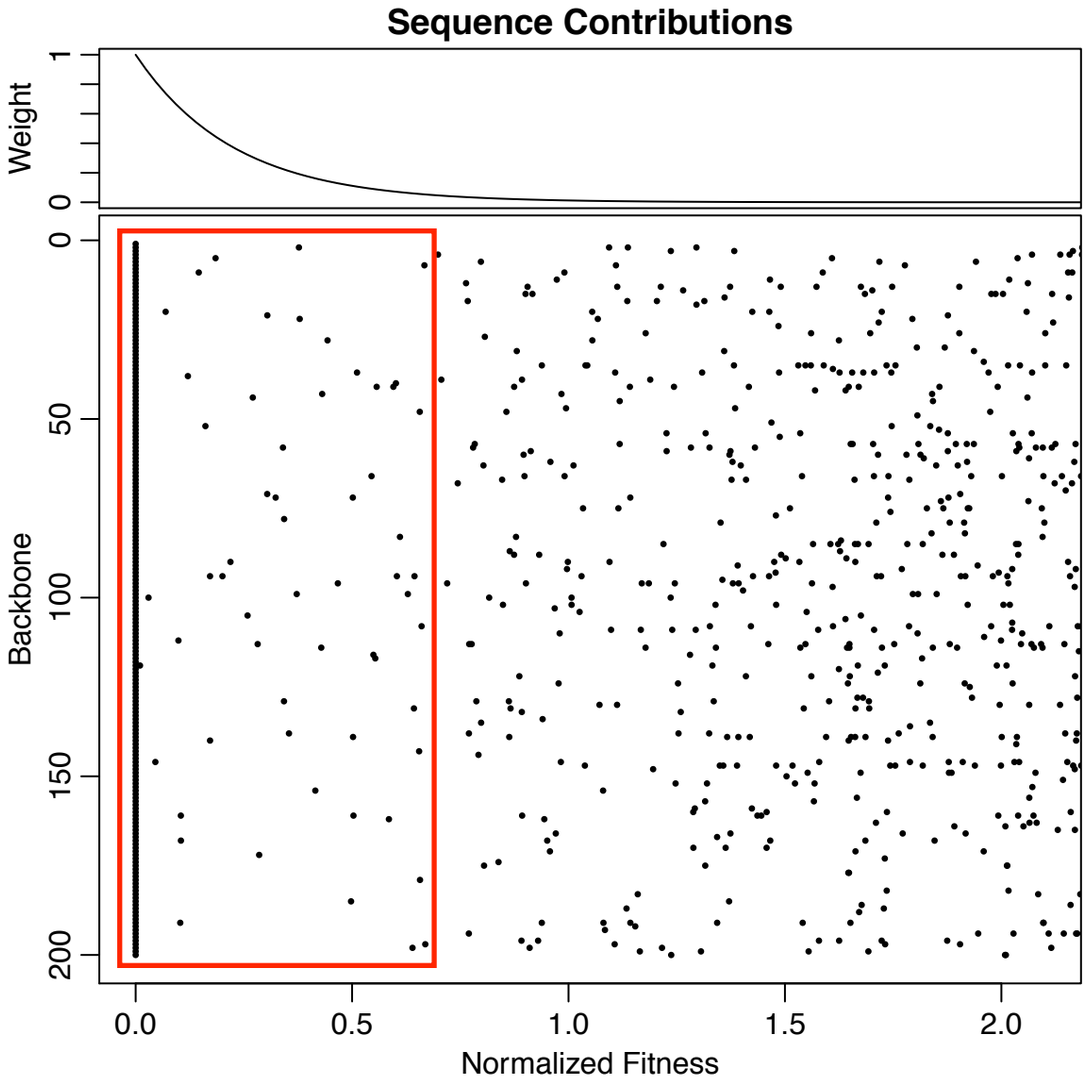
Standard Monomeric: process_seqtol(fitness_coef=c(0.4, 0.4))

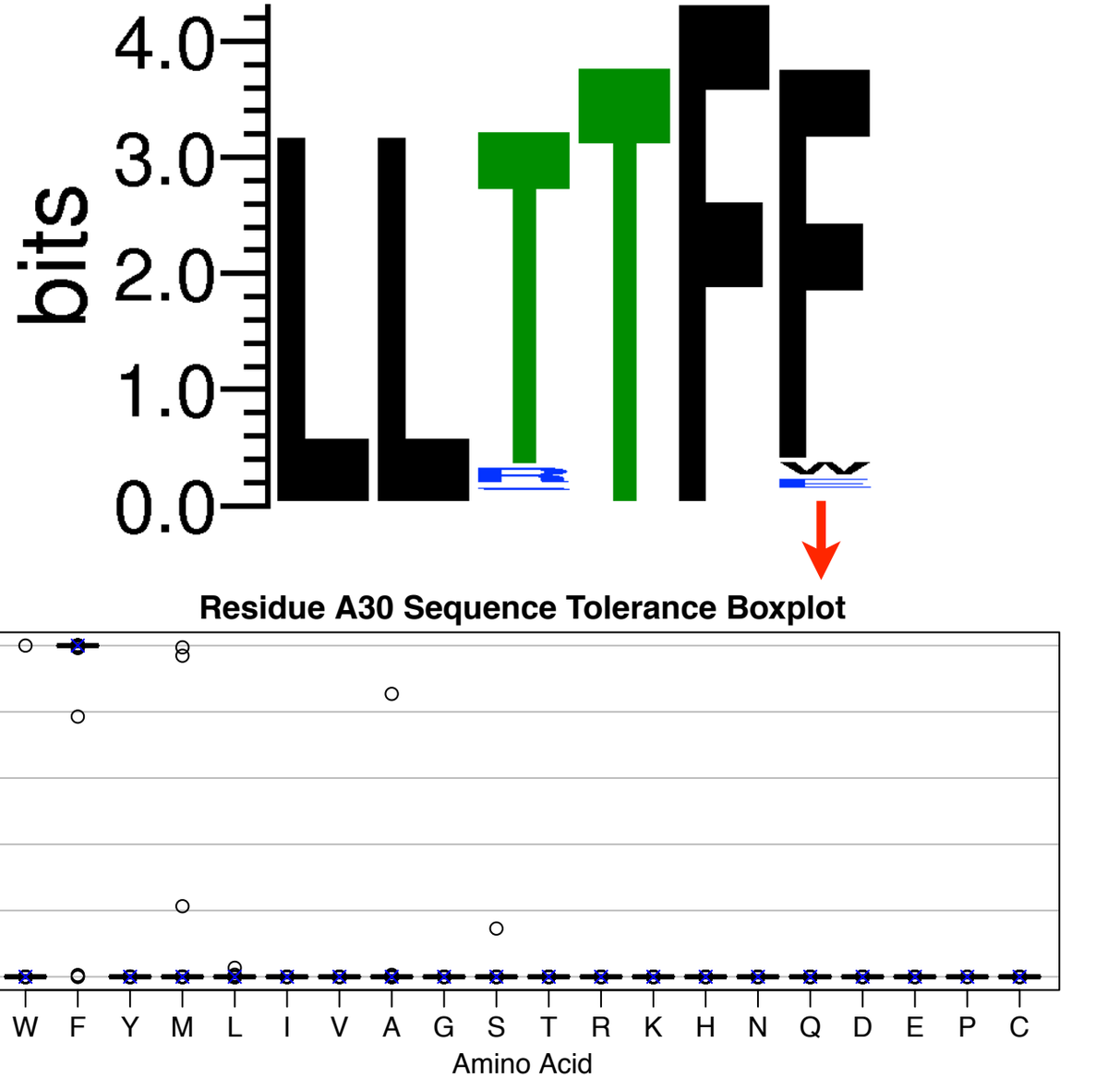






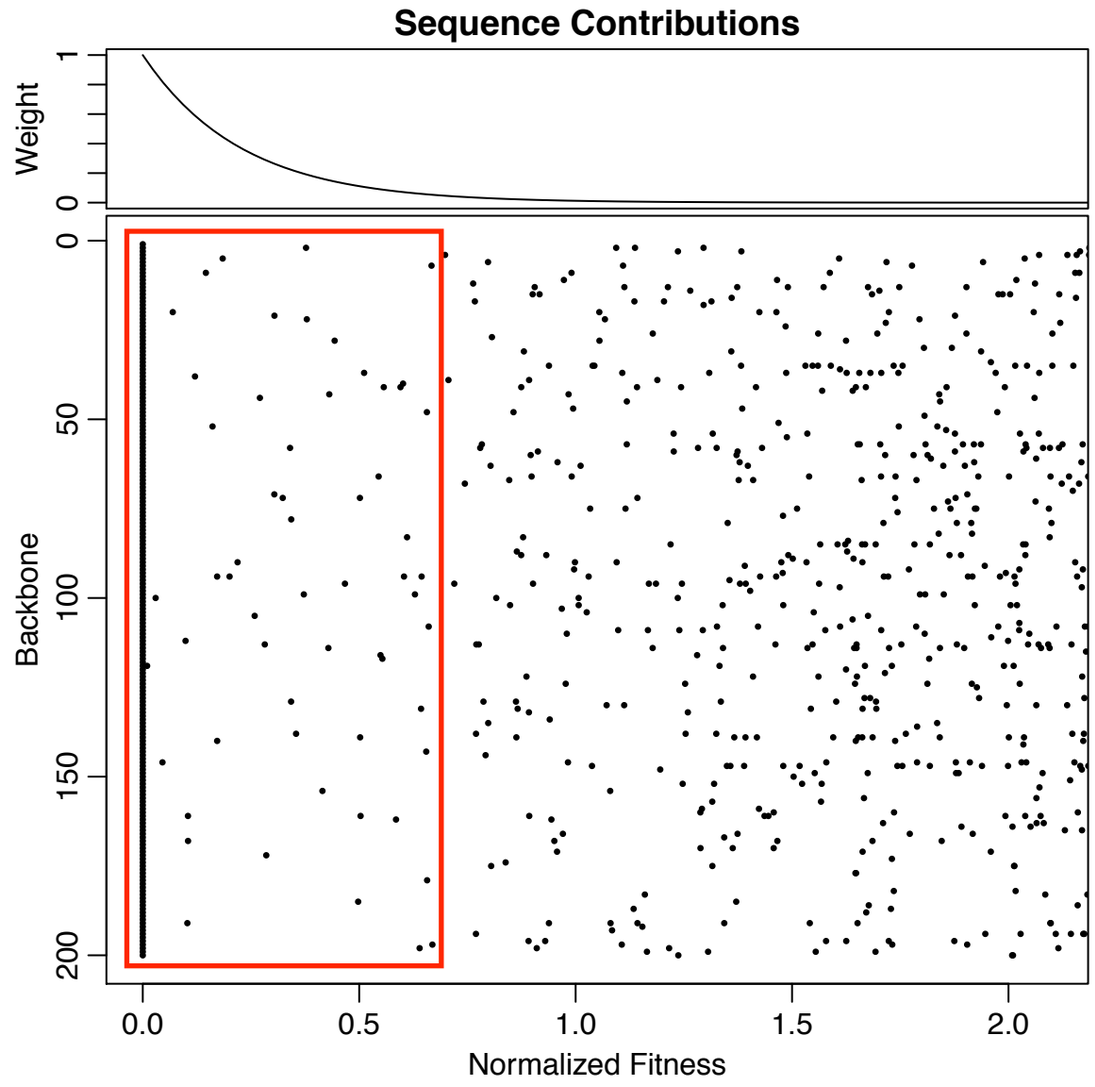
Naïve Monomeric: process_seqtol(fitness_coef=c(1, 1))



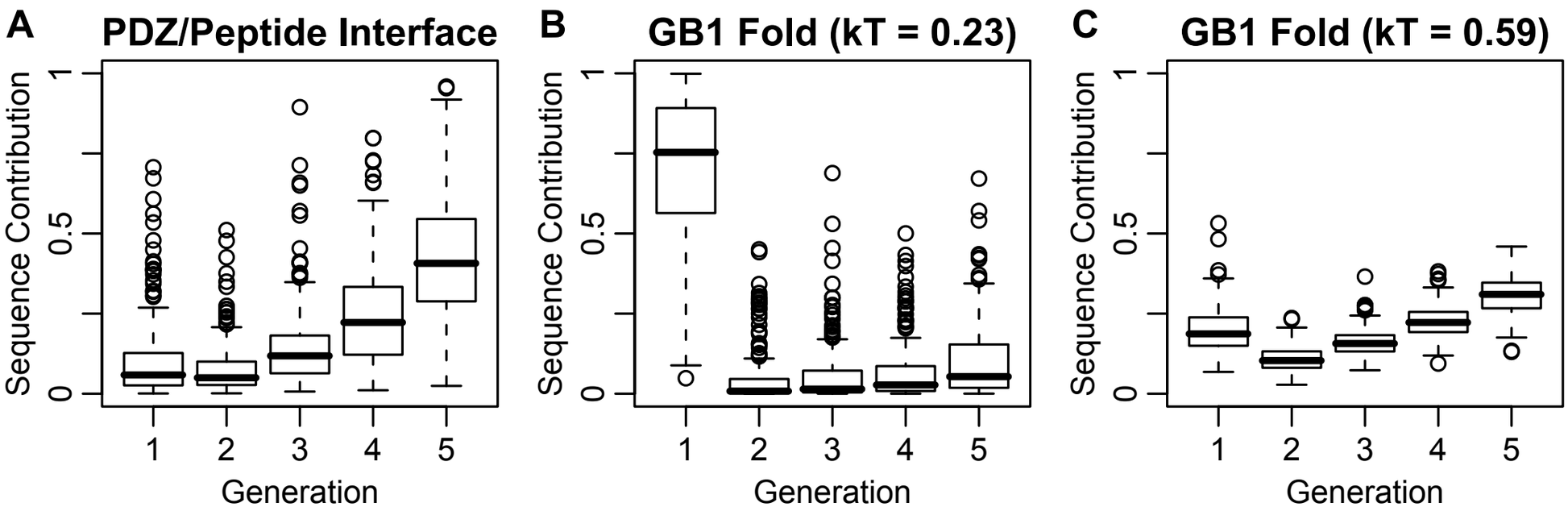


Gotcha

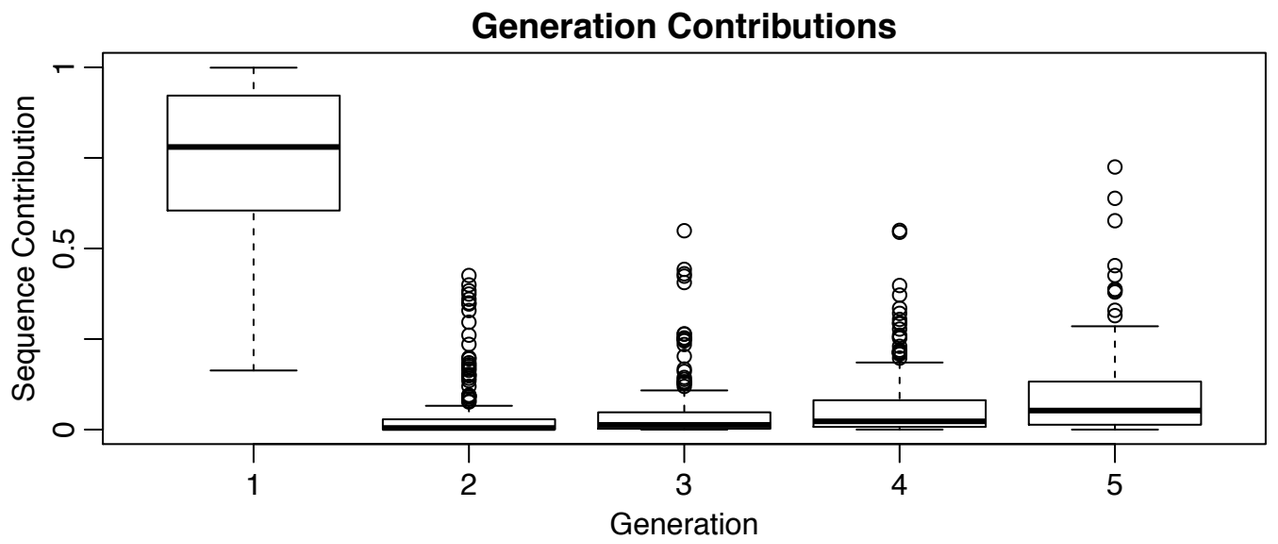
- Designing at too many positions artificially limits sequence variability



How Many Generations?




```
> process_seqtol(fitness_coef=c(1/2.5, 1/2.5),
      plotgen=TRUE)
```



Scoring Issues

- -no_his_his_pairE flag instead of his reweighting

References

- Smith, C. A. & Kortemme, T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 380, 742-756.

- Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., & Kortemme, T. (2010) RosettaBackrub--a web server for flexible backbone protein structure modeling and design. Nucleic Acids Res 38 Suppl, 569-575.

- Smith, C. A. & Kortemme, T. (2010) Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. J Mol Biol 402, 460-474.

- Smith, C. A. & Kortemme, T. (2011) Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. PLoS One 6.