Use of amino acid covariation to estimate epistatic effects in the CAGI2015 SUMO Challenge

To try to predict effects of mutations on SUMO, covariation analysis was performed on a large multiple sequence alignment of sequence UBC9\_HUMAN. A sequence alignment was first built using three iterations of HHBLITS with the latest (June 2015) UNIPROT20 database (HMM inclusion E-value cutoff of 10-6 and 80% sequence coverage). A total of 2497 sequences were aligned. From this alignment conditional pairwise substitution frequencies were estimated using a BLOSUM weighting scheme and a simple pseudocount of 1. To estimate the epistatic effects, the log odds ratio was summed between the native sequence amino acid odds (at each sequence position other than the position being mutated) versus the odds of the substituted amino acid at the specified site. The log odds calculation can be conceptually thought of as a family-and-position-specific statistical potential i.e. calculated specifically for the particular protein family (as opposed to a generic potential calculated for many different protein structures).

The final scores were fitted to the given distribution of grown scores in order to scale the log odds sums to the required range of experimental values.

For the 2nd model, a simple Markov model was used as a naïve comparison i.e. with the same alignment, but no attempt at modelling pairwise epistatic effects.