Investigating residue coevolution in proteins from a structural perspective

Alexey S. Morgunov[1,2,*], P. Alexander Gunnarsson[2], Norbert K. Fehér[1,2], M. Madan Babu[2]

[1]*Trinity College, Trinity Street, Cambridge, CB2 1TQ, United Kingdom*

[2]*MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0QH, United Kingdom,* asm63@cam.ac.uk

[*]*corresponding author*

In recent years, there has been remarkable progress in the development of computational methods for detecting evolutionary couplings between residues in proteins from multiple sequence alignments (MSA) of protein families. These methods have been successfully applied in predicting three-dimensional structures from amino acid sequences, as well as in identifying functionally important residues. One of the key limitations that hinder wider applicability and higher accuracy of the so-called coevolutionary methods lies in the fact that covariation between residues is not only a function of structural and functional relationships between them – two related components that are hard to disentangle in their own right – but also of potential interactions with other proteins and ligands, of phylogeny, and of stochastic noise in the imperfect MSA. In order to begin teasing apart the contributions of various sources of coevolution to the observed signal, in the first instance we looked at the structural components. We performed a large-scale analysis of protein structure datasets, focusing on statistics that describe types, expected distributions and propensities for specific structural residue contacts, as well as contacts arising within protein complexes. This quantitative information was compared directly to the output of best performing coevolutionary methods, and the expected distributions were subsequently used as prior information to improve the performance of some of the methods in detecting important residue couplings based on previously published analyses.