Advances in DNA sequencing technology have led to an exponential growth in the number of organisms whose genomes have been sequenced, and nucleotide sequences of over 30 million proteins now are publically available. While advances also have been made in structural genomics, with almost 90 thousand protein structures deposited in the protein data bank, this still covers only a small fraction of all protein families. Since protein structure determines protein function, it is highly desirable to have accurate structural information when predicting protein interaction networks and when designing pharmaceuticals that would alter protein function or would regulate protein-protein interactions.

Computational protein structure prediction through homology modelling can be reasonably accurate for proteins that share over 30% sequence homology with other proteins with known 3D structures. If no such proteins can be found, reference structures for small fragments of the protein of interest can still be used to assist in de novo structure prediction, by using protein fold recognition and threading approaches. Predicting protein structure *de novo* without relying on reference structures is also possible, since the genetic sequence contains all the information required to describe the 3D structure of a protein. The primary obstacle to *de novo* protein structure prediction is the large number of conformations that have to be sampled to arrive at the optimal structure of the protein, which limits this approach only to very small proteins. Furthermore, energy functions that are used to evaluate the quality of the predicted structures are only approximations to the actual quantum mechanical laws that govern molecular and sub-molecular interactions, leading to inaccuracies in the final predicted structure of a protein.

A recent algorithm called evo-fold predicts protein structure by looking at evolutionary constraints between amino acid residues in a protein sequence [(Marks et al. 2012;](http://wizfolio.com/?citation=1&ver=3&ItemID=651&UserID=14332&AccessCode=4870F096D12A4BFAB925C026D374371B&CitationSuffix=)[Hopf et al. 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=650&UserID=14332&AccessCode=AD6AE867FBE7486DBBD02CFAFDD92694&CitationSuffix=). Amino acids that are important for protein structure and function are likely to be conserved through evolution, and a mutation in one amino acid may lead to a selection of a compensatory mutations in nearby or interacting amino acids in order to rescue the functional phenotype. By looking at extensive sequence alignments for related proteins, a list of those evolutionarily-constrained amino acids can be obtained. The assumed proximity between those amino acids can be used to set geometric constraints during the computational folding of the protein, in the same way that chemical shifts obtained through NMR experiments can be used as distance constraints to predict protein structure. While the structures predicted by evo-fold would usually deviate by less than 4 Å from the reference crystal structure, the accuracy of the approach could potentially be improved even further by using *de novo* protein structure prediction, comparative modelling and structural threading. For instance, using ROSETTA it was possible to refine a coarse-grained NMR structure obtained only from the chemical shift assignments of the backbone atoms and residual dipolar coupling alignment of backbones from distant areas of the protein, in some cases to within 1.1 Å accuracy with the reference crystal structure [(Lange et al. 2012)](http://wizfolio.com/?citation=1&ver=3&ItemID=698&UserID=14332&AccessCode=0D83347A999B487F86ECE86D0B7A24C3&CitationSuffix=). Similar improvements may be possible by using ROSETTA to refine the structures produced by evo-fold.

Evo-fold not only was able to find evolutionarily constrained amino acids that govern the folding , stability and function of the protein, but in the case of dimeric proteins it could also find the key side-chains responsible for dimerisation. Presumably, such evolutionary constraints would also be present between different but interacting proteins. Mapping intermolecular constraints for putative interacting partners potentially could be used to reveal specific sites of interaction and to predict the structure of the multi-protein complex. This information could be used to design pharmaceuticals that would prevent or stabilise such interaction, either by binding to one of the proteins and blocking the interaction motif or by binding to both proteins and preventing their dissociation, respectively.

My immediate project is to create a protein engineering pipeline that would combine previously-developed comparative modelling and *de novo* protein structure prediction algorithms with information provided by evo-fold about the evolutionary co-variation in the nucleotide sequence. One possible application of this pipeline would be to design more stable scaffolding proteins. Typically this is done by predicting the effect of various mutations in the protein using molecular force fields. However if the mutations are confined only to those residues that do not show evidence of evolutionary constraint, this would both reduce the space of potential mutations that have to be sampled and decrease the chance that a mutation would disrupt the basic structure of the scaffold protein. Another application would be to design specific protein-protein interaction interfaces. While experimental techniques such as phage display still outperform computational techniques in producing molecules that bind their target with a high affinity, it is difficult to control the region on the target that such a molecule will bind. If we map the evolutionary constraints between the target molecule and other proteins that bind to it a particular region, and if we include those evolutionary constraints in our scaffolding protein and randomise the region outside those constraints, then we may be able to maintain the specificity to a particular region on the target protein while at the same time experimentally optimising the affinity and specificity of association.

[**References**](https://wizfolio.com/?style=1&ver=3&UserID=14332&StyleName=Medline-AuthorYear%20(Alexey%20mod%20Final))

Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;149(7):1607-1621.

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc Natl Acad Sci U S A* 2012;109(27):10873-10878.

Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol* 2012;30(11):1072-1080.

Park SJ, Cochran JR. Protein Engineering and Design. CRC Press; 2009. 416 p.