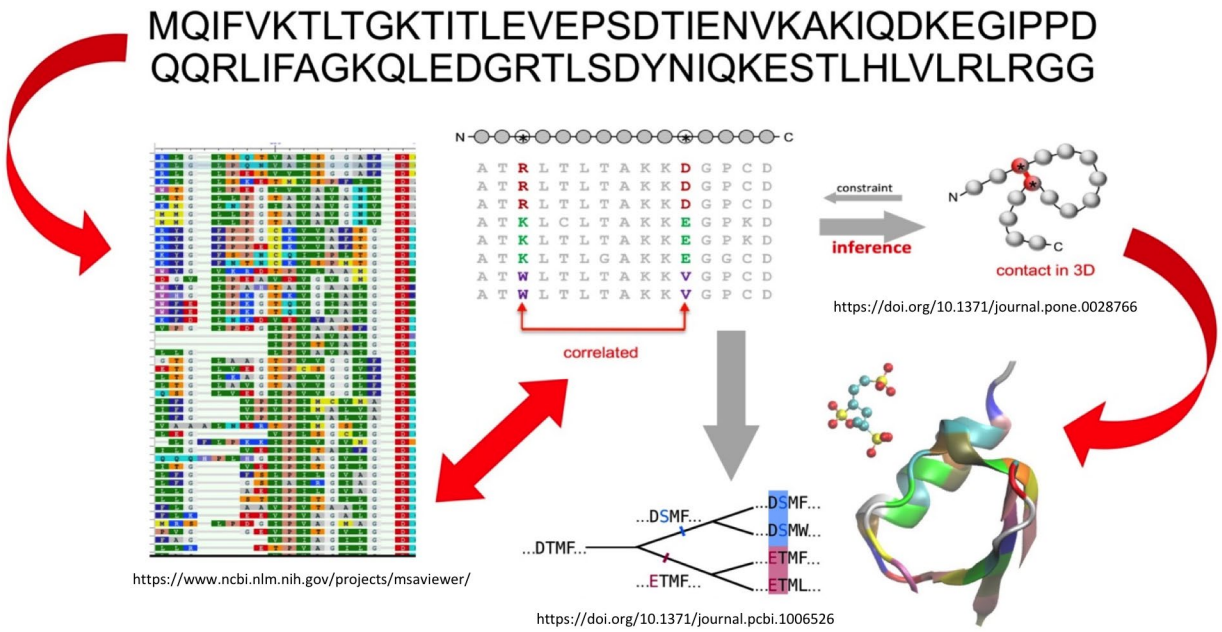
Genome to Structure: From Multiple Sequence Alignment to Virtual Ligand Screening Using Co-Evolutionary Protein-Residue Contact-Prediction in KBase

Ada Sedova[1]* (sedovaaa@ornl.gov), James G. Jeffryes[2], Loukas Petridis[1], Brian H Davison,[1] Christopher S. Henry[2], and **Julie Mitchell**[1]

[1]Oak Ridge National Laboratory, Oak Ridge, Tennessee
[2]Argonne National Laboratory, Argonne, Illinois

http://cmb.ornl.gov/index.php/research/bioenergy/dynamic-visualization-of-lignocellulose

Project Goals: The annotation of gene function benefits greatly from information on protein structure and binding. In this regard we have proposed the development of new functionality and workflows for KBase, in support of the Dynamic Visualization of Biological Structures SFA at Oak Ridge National Laboratory. The new structural branch of KBase includes tools to perform hidden-Markov model-based sequence searches with HHSuite[1], and CCMPredPy/CCMGen[2,3] tools to perform coevolutionary analysis of a protein sequence against metagenomic data, in order to predict residue contacts. These contact maps can be used to create phylogenetic trees, refine an existing multiple sequence alignment, and as restraints for use in protein folding algorithms. Finally, a protein-ligand docking workflow will allow for the use of the AutodockVINA[4] software to screen databases of metabolites, connecting metabolic-level function via physical molecular interactions to genomic information.



The annotation of gene function benefits greatly from information on protein structure and binding. In this regard we are developing a new structural branch for KBase, in support of the Biofuels SFA at Oak Ridge National Laboratory. The first workflow in this structural-biology

functionality starts from a protein sequence and uses HHSuite to perform a Multiple Sequence Alignment (MSA), followed by CCMPredPy to perform coevolutionary analysis to predict residue contact maps. These contacts can then be used as restraints in protein folding algorithms to generate a three-dimensional structural model for the protein from sequence. Information from the contact map will also provide value independent of any structure prediction, as the results indicate amino-acid residues that drive folding stability which is useful for experimental strategies in protein redesign and can also be used to refine the MSA and to create a phylogenetic tree. Several apps will be made available in KBase to utilize the HHSuite tools and the CCMPredPy/CCMPredGen tools. A second workflow will allow for the upload of a model protein structure or download of an experimental structure from the RCSB Protein Data Bank, followed by protein-ligand docking. The Ligand Screening App will use the AutodockVINA protein-ligand docking software to allow the protein structure to be screened against a set of metabolites. A potential use of this app will be in functional annotation of uncharacterized proteins: while sequence-based bioinformatics approaches have helped in annotating many genes to date, a structure-based treatment will facilitate additional insights into protein function that are obscured by lack of detectable sequence homology. In particular, for newly discovered enzymes, structural similarity along with ligand screening will help researchers anticipate likely functions. Along with the three KBase apps, we will develop sample narratives to illustrate their use.

Lignocellulosic biomass is a complex substrate that requires the synergistic action of a variety of enzymes for its efficient deconstruction. Biomass pretreatment generates byproducts, including solubilized lignin-derived aromatics, that inhibit enzymatic hydrolysis of cellulose[5]. Which bioproducts are formed depends on the biomass feedstock as well as the details of the pretreatment process. Applying the Ligand Screening App to predict which specific byproducts affect which particular enzymes can lead to an optimal selection of cellulolytic enzyme cocktails that minimize inhibition. A further barrier to biofuels and bioproduct production is that fermentation products, pretreatment solvents and byproducts can be toxic to microorganisms. We envision using CCMPred to predict the structures of novel proteins of any newly-sequenced microorganism and then docking solvents/byproducts on all proteins. Determining which proteins, and which protein residues, the small molecules bind to may lead to rational genetic engineering of those proteins and to microbes exhibiting improved tolerance to toxic pretreatment byproducts and solvents.

References

1) Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2004 Nov 5;21(7):951-60.

2) Vorberg S, Seemayer S, Soeding J. Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction. bioRxiv. 2018 Jan 1:344333.

3) Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014 Jul 26;30(21):3128-30.

4) Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry. 2010 Jan 30;31(2):455-61.

5) Ximenes E, Kim Y, Mosier N, Dien B, Ladisch M. Deactivation of cellulases by phenols. Enzyme and microbial technology. 2011 Jan 5;48(1):54-60.

Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract no. DE-AC05-00OR22725. This program is supported by the U. S. Department of Energy, Office of Science, through the Genomic Science Program, Office of Biological and Environmental Research, under FWP ERKP917.