DNASTAR
SOFTWARE FOR LIFE SCIENTISTS

Software for *In Silico* 3-D Protein Structure Predictions from DNA Sequence Variation Data

Matthew Keyser, Steven J. Darnell, Matt Larson, Erik Edlund, Richard Nelson, Schuyler Baldwin, Dan Nash and Tim Durfee

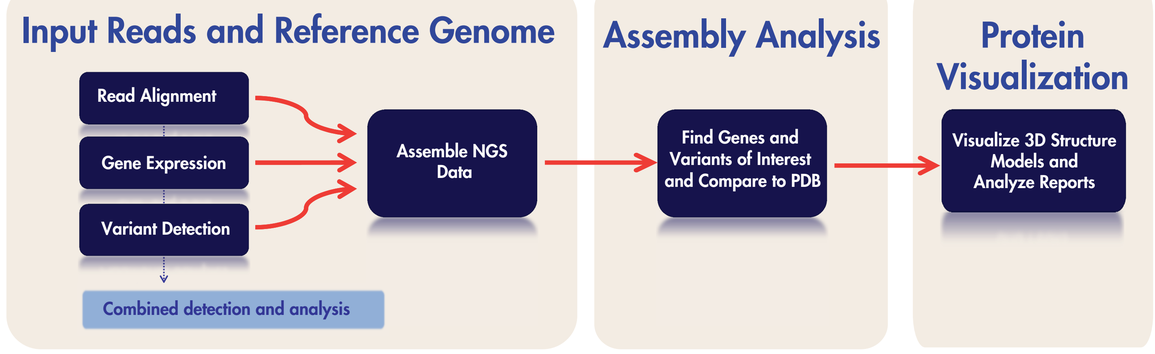
Affiliations
DNASTAR, Inc., Madison, Wisconsin, USA

Abstract

Advances in DNA sequencing have made identifying genetic variation from any individual routine, while analyzing variation across populations or cohorts has led to an explosion of candidate mutations for a variety of diseases and traits. However, these mutations are frequently non-synonymous substitutions or small indels, which make inferring their significance substantially more challenging than chain terminating nonsense and frameshift mutations. The 3-dimensional (3D) structure of a protein provides the basis not only for a mechanistic understanding of how subtle mutations affect protein function, but also the ability to rationally design countermeasures.

Since on-demand experimental determination of every mutant protein structure of interest is likely to remain infeasible for the foreseeable future, DNASTAR has developed an in silico toolkit for predicting the 3D structure of proteins harboring amino acid substitutions and/or indels. The tools are accessed through Protean 3D, DNASTAR's 3D structure viewing and analysis application, and use experimental or predicted structures as starting templates from which user-specified mutations are modelled. For substitutions, side chain conformations are optimized together with backbone perturbations and the most likely mutant structure selected based on energy minimization calculations. The use of a principled objective function to select the best structures rather than relying on sequence and knowledge-based approaches to infer structural characteristics leads to greater predictive accuracy. Moreover, the use of computationally tractable algorithms allows mutation effects to be predicted in a matter of seconds on standard desktop or laptop computers. NovaFold, DNASTAR's structure prediction application, uses template-based and ab initio techniques to model the more computationally challenging cases of insertions, deletions and homologous structure based predictions.

Here we show results from an integrated workflow where genetic variants of interest are identified from DNA sequencing data and the 3D structures of the corresponding mutant proteins are predicted using Protean 3D and NovaFold.



DNASTAR Lasergene provides integrated tools for NGS assembly with variant detection and gene expression analysis, and downstream tools for analysis of genomics and protein structure data.

Structure Visualization and Modeling

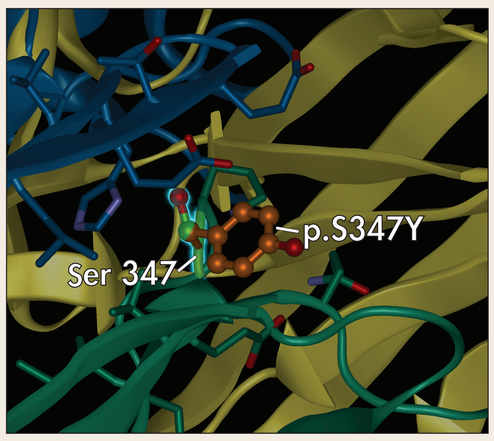


Figure 4. Repacked serine side chain with C4-A variant p.S347Y

DNASTAR's integrated structure visualization and analysis tools allow users to model mutations of interest on the PDB protein structure. Built-in energy calculations allow users to make guided hypotheses about the effect of mutations on the protein structure and function.

Protean 3D was used to model the structure of non-synonymous variants and to predict their energetic effect on the protein's fold. The side chains of the variant and residues contacting the variant were repacked using conformations defined in the Penultimate rotamer library. [6] Energetic differences between the wild type protein and the variant were evaluated with the DFIRE potential.[7]

Assembly and Variant Detection

The DNASTAR NGS Genomics pipeline provides rapid and accurate assembly of NGS sequence data for the purpose of variant detection and gene expression (RNA-seq) analysis. Combined analysis tools, along with extensive variant/gene database annotation import, provide researchers efficient means to identify candidate genes based on several criteria including mutation effects, gene expression levels, and clinical and functional importance.

Ref ID	Ref Pos	Gene Name	Uniprot_id	Kidney_normal Variant - Called Seq	Kidney_tumor Variant - Called Seq	Kidney_normal Variant - SNP %	Kidney_normal Variant - Depth	Kidney_tumor Variant - SNP %	Kidney_tumor Variant - Depth	Kidney_tumor Variant - Amino Acid Change	dbSNP ID	GERP++_NR	Kidney_tumor Variant - SIFT_pred
6GRCh38	31984403	C4A	CO4A_HUMAN	C>A	C>A	28.01%	257	37.92%	2439	p.S347Y	392610	4.34	Damaging (D)
6GRCh38	31991820	C4A	CO4A_HUMAN		G>A	------	------	11.22%	2583	p.R477Q		3.84	Tolerated (T)
6GRCh38	31997008	C4A	CO4A_HUMAN		T>G	------	------	97.14%	70	p.S1286A	201016130	3.27	Tolerated (T)

Figure 1. Variant filtering results

Name	Kidney_normal - linear total RPKM	Kidney_tumor - linear total RPKM	Fold change	Disease_description	Gene damage prediction (all cancer disease-causing genes)
C4A	420.078	5037.868	11.992 up	DISEASE: Complement component 4A deficiency (C4AD) [MIM:6...	Medium
C4B	387.885	4242.849	10.938 up	DISEASE: Systemic lupus erythematosus (SLE) [MIM:152700]: A c...	Medium
ALOX15	0.128	0.025	5.069 down	DISEASE: Note=Disease susceptibility may be associated with v...	Medium
ALOX15B	3.011	11.900	3.951 up		Medium

Figure 2. Gene filtering results

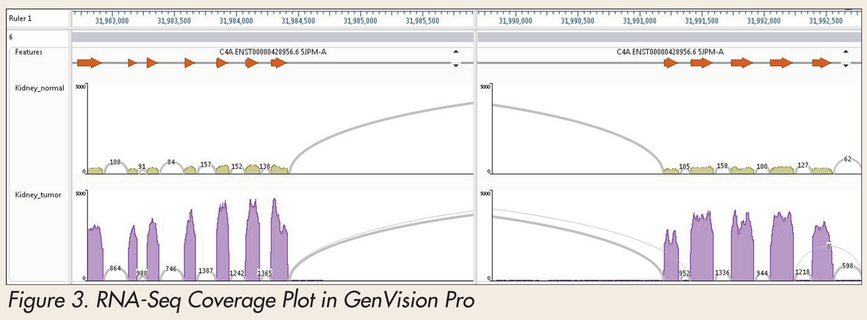


Figure 3. RNA-Seq Coverage Plot in GenVision Pro

SeqMan NGen was used to align paired samples "Kidney normal" and "Kidney tumor" (322M, 76bp Illumina paired-end reads) downloaded from RNA-Seq BioProject: PRJNA396588 *in the Short Read Archive[1] to the Human genome reference[2] containing UniProt IDs (to cross reference PDB database). Non-synonomous variants (minimum depth "10"; min "P not ref" value "10%") were identified in both normal and tumor samples (Figure 1). Additional filtering was done in ArrayStar software via imported variant annotation[3], specifically GERP ++_NR (Genome evolutionary rate profiles) and SIFT prediction to identify a candidate gene set (Figure 2). Using ArrayStar, another gene set was created from gene expression profile (fold change > 10, Figure 3) and cross compared with the variant gene candidate genes of interest. One of these genes, C4A, has been implicated in severe kidney malfunction[4] and has a resolved protein structure.[5]*

Biological Example

Gene C4A encodes protein Complement C4-A, which is essential for the propagation of the immune system's classical complement pathway. Activated C4-A is cleaved into three chains (alpha, beta, and gamma) by serine protease MASP2. Sequence analysis by the NGS Genomics pipeline predicts that C4-A variant p.S347Y will damage function. Automated protein modeling with Protean 3D suggests a structural mechanism for this predicted effect.

By combining structural bioinformatics with sequencing technologies, DNASTAR's integrated workflows can guide genomic and molecular biology researchers to create structure-based hypotheses and to investigate possibilities not evident by sequences alone.

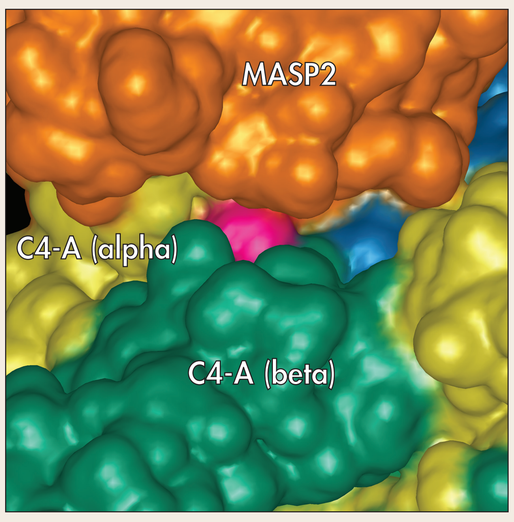


Figure 5. PDB ID: 5JTW solvent accessible surface

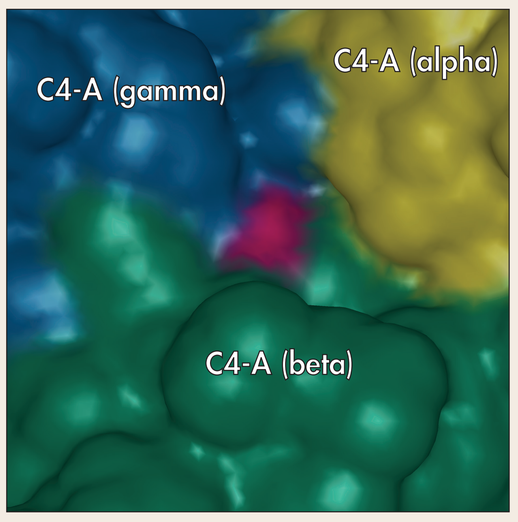


Figure 6. PDB ID: 5JPM solvent accessible surface

Energetic analysis predicts that the p.S347Y variant will stabilize the C4-A protein fold (ΔE = -1.8 units). The increased size of the Tyr residue can be accommodated without significant steric clash as Ser 347 is near the solvent-accessible surface (Figure 4). We expect that the functional damage of p.S347Y is caused by a change in binding to MASP2, which effects the activation of C4-A as well as the subsequent immune response. A significant conformational change occurs when the three chains of C4-A bind to MASP2 (Figures 5 and 6), which exposes p.S347Y to solvent (pink) and arranges it proximal to MASP2 (orange).

References:

1. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA396588
2. Human genome reference with GRCh38 and ENSEMBL annotations
3. DNASTAR Variant Annotation Database
4. https://www.ncbi.nlm.nih.gov/pubmed/15294999
5. https://www.rcsb.org/structure/5jpm
6. https://www.ncbi.nlm.nih.gov/pubmed/10861930
7. https://www.ncbi.nlm.nih.gov/pubmed/12381853

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R44GM110814. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.