Advances in DNA sequencing technology have led to an exponential growth in the number genomes that have been sequenced and have allowed for the discovery of millions of single-nucleotide polymorphisms and somatic mutations (1,2). Some mutations may be benign, but others likely affect the phenotype of the organism and may play key roles in cancers and other diseases (3,4). At present, it remains infeasible to characterise experimentally the impact of every mutation, and computational approaches often are used to analyse all mutations and prioritise them for experimental validation (5–10). However, existing computational approaches either are limited in their accuracy or are not designed for high-throughput screening, and most focus on predicting the effect of a mutation on protein folding and do not consider the impact of a surface mutation on protein-protein interactions.

Our group has developed recently a computational pipeline that, in the case of a mutation in the core of the protein, can predict the effect of that mutation on the Gibbs free energy () of protein folding, and in the case of a mutation in an interaction interface, can predict the effect of that mutation on the  of the interaction. The pipeline calculates a sequence conservation score of the mutated amino acid using SIFT (9), makes a homology model of the muted domain or a pair of interacting domains using MODELLER (11) and calculates the energetic impact of the mutation using FoldX (10). It uses a Stochastic Gradient Boosting of Decision Trees (SGB-DT) algorithm (12) trained on a dataset of experimentally-measured changes in  (8,13), to combine the sequence conservation score, the semi-empirical energy terms, and a variety of molecular details about the wild-type and mutated amino acids, in order to predict the change in  caused by the given mutation. The pipeline is fully automated, and it performs better than existing methods (5–8).

My immediate project is to extend the pipeline to include protein-peptide, protein-DNA and protein-RNA interactions, as well as interactions that involve phosphorylated residues. For protein-peptide interactions, I will incorporate into the pipeline regular expressions (14) and position-weighed scoring matrices (15) that define the preference of peptide-binding domains for different peptides. I will make homology models of all protein-peptide complexes and, in order to remove false positives, will use FoldX to filter those complexes based on affinity. I will then be able to predict changes in  caused by mutations affecting any protein-peptide interface. For protein-DNA and protein-RNA interactions, I will introduce into the pipeline new molecular descriptors that will capture the interactions between amino acids and nucleotides, and I will retrain the SGB-DT algorithm using datasets obtained for mutations at the protein-nucleic acid interface (16). The resulting algorithm will be able to make accurate predictions for mutations affecting nucleotide-binding domains. For interactions involving phosphorylated residues, I will incorporate into the pipeline a list of known phosphorylated sites in the human proteome (17–19) and will use FoldX to calculate the thermodynamic impact of phosphorylation for every protein-protein and protein-peptide interface that involves a phosphorylated residue. I will use this information to model the effect of phosphorylation on protein-protein interaction networks, and I will be able to predict changes in  caused by mutations for both phosphorylated and un-phosphorylated complexes.

Subsequent work will involve using the pipeline in the more global context of metabolomics and protein signalling networks. A machine learning algorithm such as SBG-DT will be used to combine mutation-induced changes in  with a variety of other features, such as the centrality of the protein in a protein-protein interaction network, the types of proteins with which it interacts, the pathways in which it is involved, its expression level in different tissues, its co-expression profile with other proteins and its gene ontology terms. The algorithm will be trained using large databases which list mutations that are associated with cancers and other disorders (4,20), and it will be used to predict the phenotypic effect of new and uncharacterised mutations. Such computational approaches are essential for interpreting the unprecedented amount of data that now is available, for making predictions to be used in directing future experiments, and for designing new treatments that would be targeted to specific cell abnormalities.

References:

1. Tennessen JA, Bigham AW, O’Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science. 2012 Jul 6;337(6090):64–9.

2. Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56–65.

3. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010 May 27;465(7297):473–7.

4. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005 Jan 1;33(suppl 1):D514–D517.

5. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics. 2009 Oct 1;25(19):2537–43.

6. Wainreb G, Wolf L, Ashkenazy H, Dehouck Y, Ben-Tal N. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics. 2011 Dec 1;27(23):3286–92.

7. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: prediction of changes in protein-protein binding affinity on mutations. Nucleic Acids Res. 2013 May 30;41(W1):W333–W339.

8. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res. 2004 Jan 1;32(suppl 1):D120–D121.

9. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009 Jun;4(7):1073–81.

10. Guerois R, Nielsen JE, Serrano L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. J Mol Biol. 2002 Jul 5;320(2):369–87.

11. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative Protein Structure Modeling of Genes and Genomes. Annu Rev Biophys Biomol Struct. 2000;29(1):291–325.

12. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002 Feb 28;38(4):367–78.

13. Moal IH, Fernández-Recio J. SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics. 2012 Oct 15;28(20):2600–7.

14. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Van Roey K, Altenberg B, et al. ELM--the database of eukaryotic linear motifs. Nucleic Acids Res. 2011 Nov 21;40(D1):D242–D251.

15. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 2003 Jul 1;31(13):3635–41.

16. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, et al. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res. 2006 Jan 1;34(suppl 1):D204–D206.

17. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. PROTEOMICS. 2004;4(6):1551–61.

18. Dinkel H, Chica C, Via A, Gould CM, Jensen LJ, Gibson TJ, et al. Phospho.ELM: a database of phosphorylation sites—update 2011. Nucleic Acids Res. 2011 Jan 1;39(suppl 1):D261–D267.

19. Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: the posttranslational modification database. Nucleic Acids Res. 2011 Jan 1;39(suppl 1):D253–D260.

20. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.