9 October 2017

MSc Bio-IT, Integrated Bio-IT project

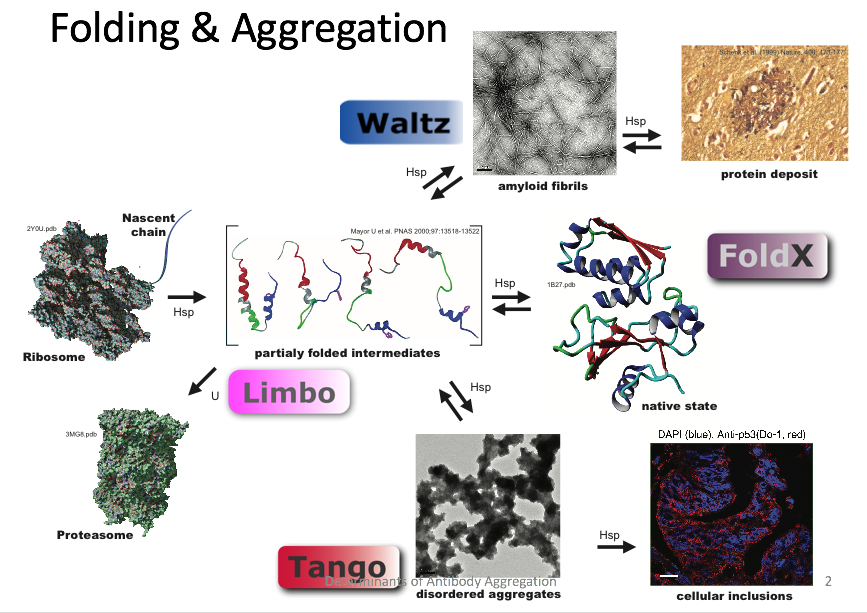
**Team SNPeffect**

**Students:** Qian Yu ([qian.yu@student.kuleuven.be)](mailto:qian.yu@student.kuleuven.be)), Xu Xiao ([xu.xiao@student.kuleuven.be](mailto:xu.xiao@student.kuleuven.be)), Colton Gowan ([colton.gowan@student.kuleuven.be)](mailto:colton.gowan@student.kuleuven.be))

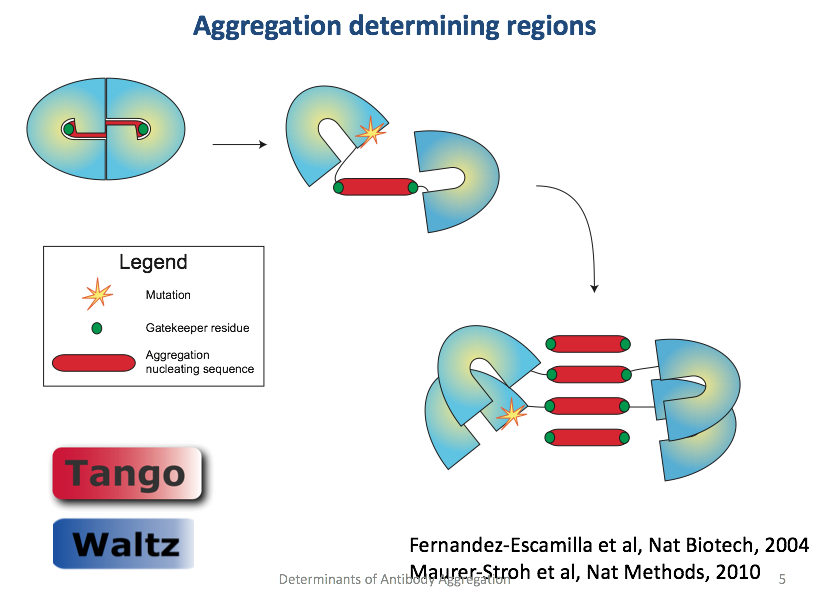
**Coaches**: Joost Schymkowitz ([joost.schymkowitz@kuleuven.vib.be](mailto:joost.schymkowitz@kuleuven.vib.be)), Mark Fiers ([mark.fiers@kuleuven.vib.be](mailto:mark.fiers@kuleuven.vib.be)), Rob Van der Kant ([rob.vanderkant@kuleuven.vib.be)](mailto:rob.vanderkant@kuleuven.vib.be))

**Background**

SNPEffect (*1-4*) is an existing online tool that uses mainly bio-informatics tools of the Switch Laboratory to predict the effect of mutations on the stability and folding of proteins. The important of these are the atomic force field FoldX (*5, 6*), the statistical thermodynamics prediction algorithm for protein aggregation TANGO (*7*) and the integrated scoring function Solubis (*8, 9*). The current version of SNPEffect only calculates predictions for a single mutation at a time, but we want to update this to the era of next generation sequencing, and allow users to start from VCF (Variant Calling Format) files. This information, which is on the transcript level, needs to be translated into protein space (i.e. map codon to amino acid), to calculate the impact on the protein. In addition to our own tools, we want to also include PolyPhen (*10*) and SIFT (*11*) predictions for each mutation. Last but not least, we will need to obtain domain boundary information from a resource like PFAM (*12*) or similar, because we want to detect disruptive mutations in protein domains with a high scoring tango region. Finally, we will iupred (*13*) to detect intrinsically disordered regions of proteins, to see if the mutations affect the aggregation propensity there.



As shown in the schematic above, (mutant) proteins need to be made on the ribosome, after which they undergo folding to the native state. If the native state is thermodynamically destabilized by the mutation, or the folding reaction is disturbed, the protein can adopt intermediately folded conformations that can lead to its degradation, or the formation of aggregates. FoldX calculates the effect of the mutation on the stability of the native state, expressed as a G value in kcal/mol. TANGO predicts the aggregation propensity of the unfolded state of the protein, which typically gives a zero score for most residues, except for short aggregation prone regions with a score of up to 100 (very rare). Since most residues are thus not part of an aggregation prone region, mutations that affect the APRs directly are rare, but a large fraction of disease mutations are destabilizing and hence cause the exposure of APRs, promoting aggregation in that way (*14, 15*). Solubis is a method design to detect APRs that are exposed in the native state, which are likely to be problematic even if the protein can fold. If a mutation would cause a strong rise in the TANGO score (e.g. by creating a new APR) in an intrinsically disordered region, that is also likely to increase aggregation.



What we want to achieve by creating a pipeline to annotate an entire mutation calling set is that we want to calculate the total ‘mutation load’ a cell caries in terms of protein damage. Typically, cancer cells have many of these mutations and they lead to an upregulation of the specialized stress response pathways, mainly upregulating molecular chaperones (*16, 17*). Inhibition of chaperones or protein degradation is used in the hospital for treating cancer, but not all tumors respond, in part perhaps because the sensitivity depends on how many damaged proteins the cell needs to handle.

**Datasets**

* We will work with a set of previously characterized patient derived melanoma cell lines for which RNA sequencing data is available (*18*), and for which we thus have mutation calling. In addition, these data provide expression levels, so we can use expression level as a weighing factor when calculating the total amount of protein damage. We know how sensitive these cells are to chaperone or degradation specific treatments.
* We have a set of 30.000 PDB files, containing homology models of about 40% of human proteins.

**Pipeline**



**Typical queries**

* (A) Find all mutations that have DDG > x in domain with TANGO > y
* (B) Find all mutations that have DTANGO > x in a domain with iupred > y
* (C) Calculate the total mutation load in a cell, e.g. by summing the expression levels of all mutations in A or B

**References**

1. G. De Baets *et al.*, SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* **40**, D935-939 (2012).

2. J. Reumers *et al.*, Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucleic Acids Res* **36**, D825-829 (2008).

3. J. Reumers, S. Maurer-Stroh, J. Schymkowitz, F. Rousseau, SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics* **22**, 2183-2185 (2006).

4. J. Reumers *et al.*, SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res* **33**, D527-532 (2005).

5. J. Van Durme *et al.*, A graphical interface for the FoldX forcefield. *Bioinformatics* **27**, 1711-1712 (2011).

6. J. Schymkowitz *et al.*, The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-388 (2005).

7. A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology* **22**, 1302-1306 (2004).

8. J. Van Durme *et al.*, Solubis: a webserver to reduce protein aggregation through mutation. *Protein Eng Des Sel*, 285-289 (2016).

9. G. De Baets, J. Van Durme, R. van der Kant, J. Schymkowitz, F. Rousseau, Solubis: optimize your protein. *Bioinformatics* **31**, 2580-2582 (2015).

10. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).

11. P. C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812-3814 (2003).

12. A. Bateman *et al.*, The Pfam protein families database. *Nucleic Acids Research* **32**, D138-D141 (2004).

13. Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-3434 (2005).

14. A. Ganesan *et al.*, Structural hot spots for the solubility of globular proteins. *Nature communications* **7**, 10816 (2016).

15. F. Rousseau, L. Serrano, J. W. Schymkowitz, How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* **355**, 1037-1047 (2006).

16. C. Dai, S. B. Sampson, HSF1: Guardian of Proteostasis in Cancer. *Trends Cell Biol* **26**, 17-28 (2016).

17. Z. Tang *et al.*, MEK guards proteome stability and inhibits tumor-suppressive amyloidogenesis via HSF1. *Cell* **160**, 729-744 (2015).

18. A. Verfaillie *et al.*, Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nature communications* **6**, 6683 (2015).