Prediction of the effect of single mutations in protein stability using graph theory and Coarse Grained models.

David Ricardo Figueroa - Sergio Canales Martínez

Introduction

Mutations are a central part of the evolutionary process because of their role in diversifying genomes and proteomes by altering their protein structures and functionality. Some of these mutations could affect metabolic pathways by the decrease of catalytic efficiency and protein stability[1], leading to different neurodegenerative diseases[2]. The understanding and prediction of the effect of mutations could be used for the improvement or modification of enzymes to increase catalytic activity or propose alternative and cheaper synthesis routs without metallic catalysts. So, the importance of the predictions of changes in protein stability for mutations in amino acid sequences with fast and accurate results would be really useful in fields such as medicine or biotechnology.[3]

Currently, the experimental techniques that measure protein stability are expensive and slow,[4] therefore, the development of computational methods that can predict the effect of mutations more accurately and efficiently are necessary. The most important types of methods related to the predictions of changes in protein stability currently are: Sequence-based methods as INPS, Mustab[5] or I-Mutation 2.0[6], structure-based techniques based on potential-energy-based approaches as Fold-X[7] and Machine-Learning algorithms with combinations of sequence and 3D structure information as mCSM[8].

Molecular Dynamics Simulations and Coarse Grained Models.

Molecular dynamics simulations (MD) are techniques commonly used in the study of protein behavior with some applications for drug design, protein-protein interaction studies, and other types of molecular phenomena.[9] However, if the size of the system is considerable big ($<$100K atoms), the MD simulations are computationally expensive and therefore coarse grained (CG) models are preferred. In these models the amino acid atoms are grouped in simpler models (as shown in Figure 1) by reducing the number of interactions to calculate and thus allowing for

bigger size systems or longer simulations times. However, grouping atoms will inevitably hide some information from the original representation of the protein as hydrogen bonds or π-π stacking.[10]

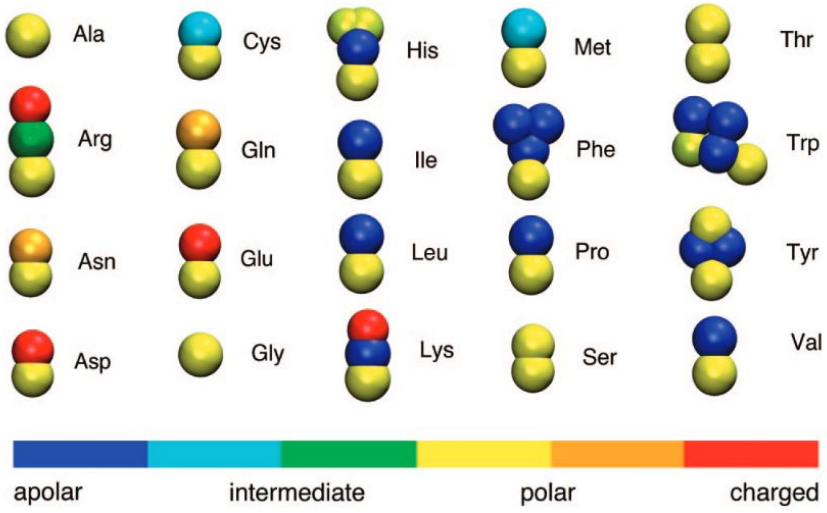

Figure 1. Coarse-grained representation of all amino acids.
Different colors represent different particle types.[11]

The CG and MD simulations use the same potential-energy-approach called the Force Field (FF), which is an empirical function that reflect the energy associated by bonds, angles, dihedral, Coulomb Interactions and Van der Waals interactions as shown in Eq 1.

$$V = \sum_{i}^{enlaces} k_{l,i}(l_i - l_{ref}) + \sum_{i}^{angulos} k_{\alpha,i}(\alpha_i - \alpha_{ref}) + \sum_{i}^{dihedros} K_d[1 + cos(n_k\theta - \theta_{ref})] + \sum_{i,j}^{N} \epsilon_{ij}\left[\left(\frac{r_a}{r_{ij}}\right)^{12} + \left(\frac{r_b}{r_{ij}}\right)^{6}\right] + \sum_{i}^{N} \frac{q_i q_j}{4\pi\epsilon r_{ij}} \quad (1)$$

Calculating the energy of the system with the FF leads to the calculation of the forces in each atom and therefore the acceleration, velocities and positions through the numerical integration of Newton's Second Law. The MD simulation allows studying protein dynamically, avoiding experimental problems involved with X-ray crystallography as unnatural clashes, specific and non-physiological conditions such as lower temperatures or different pH.[12]

The change in Gibbs free energy of the folding process (which relates with protein stability) is described as the sum of enthalpy and the entropy (Eq 2). By convention, negative values of ΔG

represent spontaneous folding processes that reflect tertiary structure with correct functionality.[13] Then, any change in this quantity could be quantified as it is shown in Eq 3.

$$\Delta G = \Delta H - T\Delta S \quad (2)$$

$$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wildtype} \quad (3)$$

The intermolecular hydrogen bonds and Van der Waals interactions that contribute favorably to the enthalpy (ΔH) and, the aggregation of water molecules around hydrophobic residues that cause a reduction in entropy (ΔS, which affects unfavorably the ΔG). The latter one, called the hydrophobic effect, depends on the character of every amino acid.[14]

Some reported approaches to calculate the change in stability as Fold-X, uses an empirical equation with parameters obtained from linear regressions and the partition of the ΔG into solvation of polar, hydrophobic groups, hydrogen bonds, Van der Waals and geometric clashes.[15] In contrast, other models as i-Mutant 2.0 use large dataset of mutations with experimental values of $\Delta\Delta$G to train a machine learning as SVM with a Radial Basis Functions Kernel.[6]

Taking advantage of the dynamical analysis allowed by the MD simulations and the driving forces quantifying the changes in protein stability, we propose an algorithm to classify if any mutation could destabilize or stabilize a protein structure. Using different conformations with explicit waters in the CG-MD, we create a graph to measure the change in the local centrality of the nodes corresponding to mutation, and then classify the effect of changes in amino acids.

Designed Algorithm to predict the effect of mutations.

The algorithm to predict the impact of mutations as a mean of changes in stability, has the following I/O:

Input: wild type protein (WT) and mutants (M)

$array\,of\,atoms: WT,$

$array\,of\,mutants: M$

an atom is defined as,

$tuple(\, name : string,\, amino\, acid : string,\, chain : char,\, amino\, acid\, number \in \mathbb{N},\, tuple(x, y, z) \in \mathbb{R}\,)$

a mutant is defined as,

$tuple(amino\, acid_1 : string,\, amino\, acid\, number \in \mathbb{N},\, amino\, acid_2 : string,\, chain : char\,)$

Precondition:

$\langle \forall m \in M \,|\, \langle \exists a \in WT \,|\, : m_{aminoacidnumber} = a_{aminoacidnumber} \wedge m_{chain} = a_{chain} \wedge m_{aminoacid1} = a_{aminoacid} \wedge m_{aminoacid2} \neq a_{aminoacid} \rangle \rangle$

Output: The difference (ΔΔG) in stability (S) between the wild type protein and each mutant

$S \to \mathbb{R},$

$array\, of\, \mathbb{R} : \Delta\Delta G$

Postcondition:

$\Delta\Delta G = \langle \forall k \in M \,|\, S(WT) - S(k) \rangle$

To carry out the algorithm, we executed molecular dynamics simulations of 300 ps after equilibrating the system using GROMACS 5.0[16] and MARTINI Force Field[11] to obtain CG solvated structures. Then we created a LJ graph, where each node is an atom in the CG model and each edge is a LJ interaction between two atoms at a certain distance r lower than 15 Å and higher than 3 Å, the weight of the LJ interaction is given by the 4th and 5th term in the Eq 1.

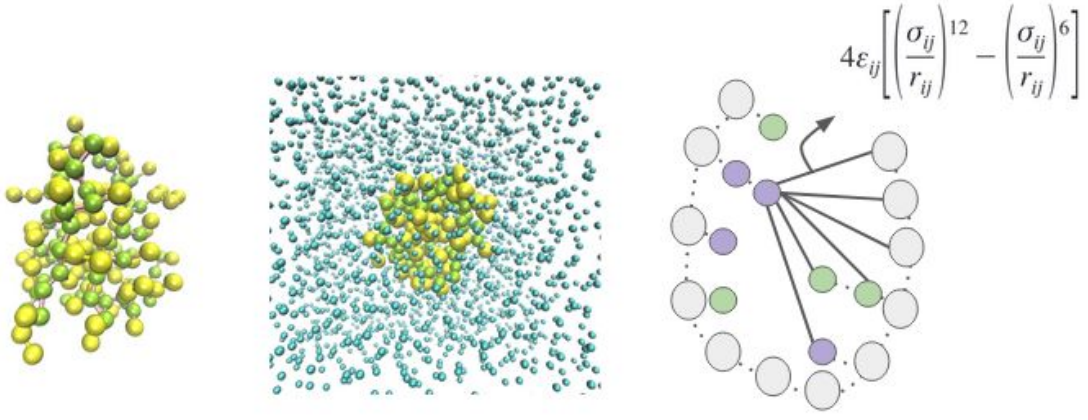

Figure 3. A) CG Model of a protein, B) CG solvated model, C) Schematic Graph Representation. Dotted lines symbolize covalent bonds and lines represent LJ interactions.

We extracted 20 conformations from the molecular dynamic simulation and proceeded by calculating the local densities as :

$$local\,densities \;=\; \langle \forall\, node \;\in\; Graph \,|\, \langle \forall\, edge \;\in\; node \,|\, (+\, subedge \;\in\; edge.node \,|\, subedge.node \;\neq\; node \wedge subedge.node \;\in\; node.edges\,)\,\rangle \,/\, |node.edges| * (|node.edges| - 1)/2\,\rangle$$

Where *Graph* is composed of *nodes* and *edges* and every *edge* has its respective associated *node* (*edge.node*). This metric shows how connected the neighbor nodes of a certain node are (assuming a neighbor node is the one connected by an edge).

Then we weighed each density depending on the energy of that specific conformation related to the energies of the other structures in the MD. For this approach, we took the maximum and minimum energy per MD trajectory and weighed the local densities with a 60%-100% scale accordingly. We used the following equation for the weighting,

$$weight(x) \;=\; \frac{x - min}{max - min} * (1 - min\%) + min\%$$

In which $min\%$ corresponds to 60% and x to the corresponding local density. Afterwards, we added all weighted densities per node and divided it by the total amount of frames.

Benchmarking

To compare the results obtained from our algorithm with another reported software, we used statistical measures of the performance of a binary classification test as sensitivity, specificity and precision extracted from the confusion matrix. We calculated this confusion matrix[17] using 20 proteins from the S264 dataset that contains curated proteins and the experimental values of $\Delta\Delta G$ of their respective mutations.[18] As shown in Figure 3, we compared against two different methodologies: Fold X, that uses an empirical energy function based on 3D structure and i-Mutant2.0 which is a support vector machine model based only on the sequence variation:

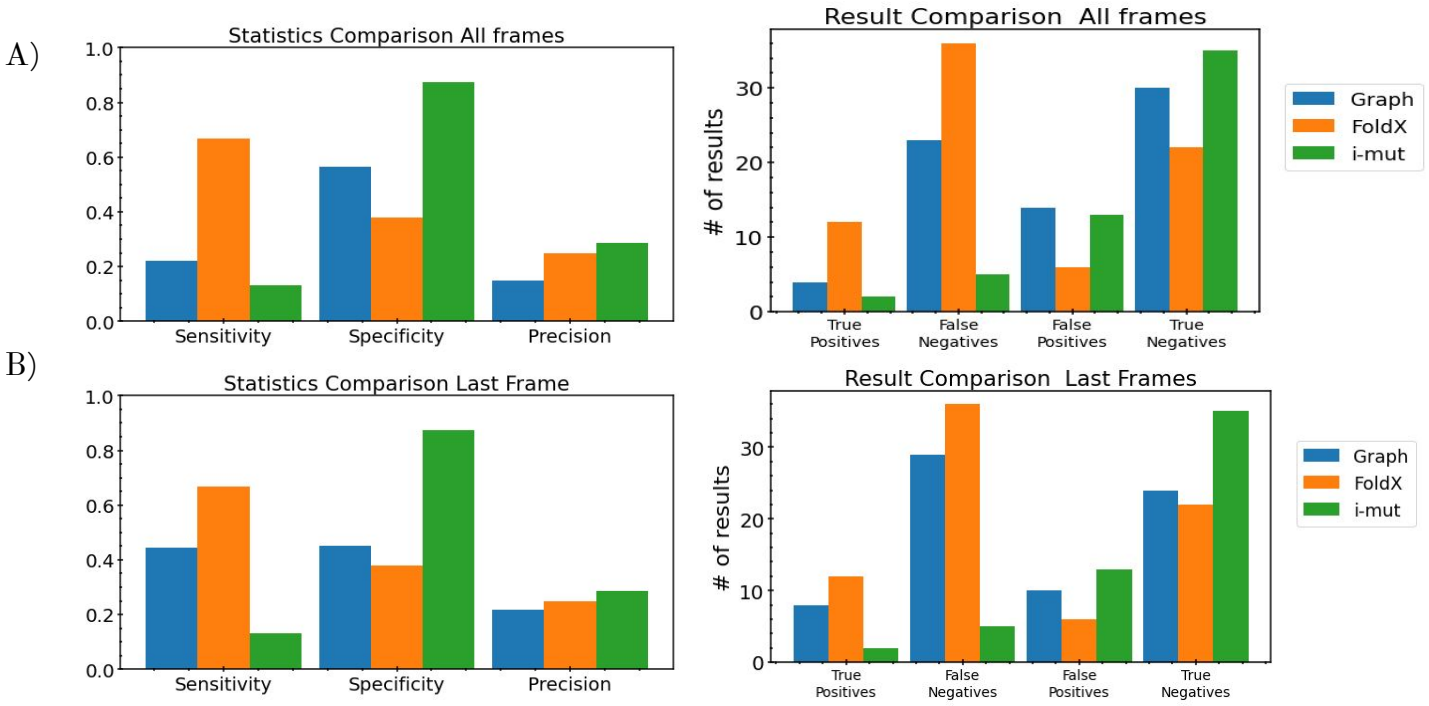

Figure 3. Benchmarking against Fold-X and i-Mutant using A) weight energy mean of the MD trajectory and B) using only the last frame from the MD simulation.

These results show that the specificity of our algorithm is better than FoldX and i-Mutant, then the prediction if a mutation destabilizes the protein was classified better (Figure 3A). These results show how the graph approach takes into account the reduction of interactions per node by changes of atoms and therefore mimics the impact of a reduction in enthalpy.

The sensitivity of Fold-X and i-Mutant outperformed our algorithm. This means that our approach was less prone to correctly determining if a mutant stabilized, because the change in local centrality should be sensitive to the distance and the strength of the LJ interaction. For example: If the interaction is attractive, it is going to be stronger at closer distances and also, if the interaction is repulsive, should be more destabilizing at closer distances. Another important difference that contributes to the higher sensitivity of Fold-X is the use of an all atoms model that includes the explicit treatment of hydrogen bonds.

In contrast, by using the calculation of change in local centrality based only on the last frame of the MD trajectory and avoiding the weighted mean based on energy of each conformation (Figure 3B), there is an increase in recall and sensitivity that could not outperform the other two software

programs and also results in a decrease of specificity. To clarify the effect of the number of conformations used for the calculation, the full confusion matrix should be calculated with a different number of structures obtained from the MD.

One important aspect about the graph construction is that the execution time increased as the number of atoms in the system as shown in Figure 4. Fold-X's execution time is independent of the number of atoms because it depends only on the energy conformation and its convergence limits. By using the same data set, the maximum and minimum execution time from Fold-X was 87 and 3.6 seconds respectively.

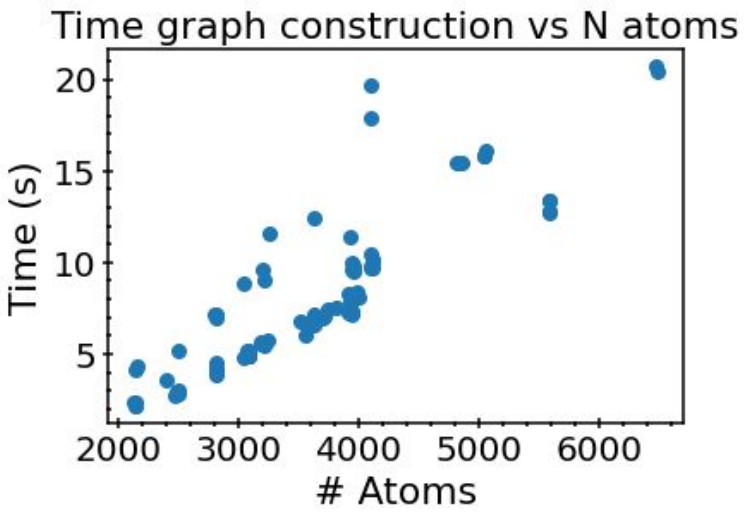

Figure 4. Dependence of time against the number of atoms in the system(including waters)

Advantages and Disadvantages.

Our approach includes two key characteristics compared to other methodologies for effect mutation prediction: The dynamical analysis of protein conformation produced by molecular dynamics simulations and, the use of explicit waters in the calculation in order to account for the hydrophobic effect around the protein topology. The advantage of the former one, is a dynamic analysis of protein structures that takes into account the energy of each conformation, where configurations with higher energies are less probable. The advantage of the latter one is the evaluation of the hydrophobic effect by repulsive LJ interactions from hydrophobic residues and closer waters.

Even if the construction of the graph calculates LJ interactions with water molecules and residues in the protein, to treat hydrogen bonds approximately, the CG model does not treat this interaction explicitly and the significant contribution to enthalpy in the Gibbs free energy is not taking into account, and this clearly reflects a disadvantage against all atoms models. Also, the time performance of our approach implies a molecular dynamic simulation that requires extra calculations that take around 60 seconds per mutant. Finally, our model does not include a special calculation for disulfide bonds, which clearly contribute to tertiary structure stability and therefore, could not predict a correct effect in single or multiple mutation of residues as methionine or cysteine.

As a perspective for future improvements and applications, the change in local centrality should be calculated taking into account the weight in the graph in order to capture not only the change in centrality but also the change in interactions and their respective dependence to distance inside the protein. In addition, the identification of water nodes with high local centrality to atoms of the protein could be indicative of depth pockets or solvent accessible areas in proteins which could be used for drug design.[19] Finally, to include the entire relationship between protein stability and function, it is necessary to include functional annotation such as active site from databases as UniprotKB.[20] Including this information could be implemented as a penalization function if any mutation is close to any important residue related to function.

References.

(1) Griffiths, A. J.; Gelbart, W. M.; Miller, J. H.; Lewontin, R. C. Protein Function and Malfunction in Cells. *Mod. Genet. Anal.* **1999**.

(2) Stefl, S.; Nishi, H.; Petukh, M.; Panchenko, A. R.; Alexov, E. Molecular Mechanisms of Disease-Causing Missense Mutations. *J. Mol. Biol.* **2013**, *425* (21), 3919–3936. https://doi.org/10.1016/j.jmb.2013.07.014.

(3) Baker, M. Protein Engineering: Navigating between Chance and Reason. *Nat. Methods* **2011**, *8* (8), 623–626. https://doi.org/10.1038/nmeth.1654.

(4) Deller, M. C.; Kong, L.; Rupp, B. Protein Stability: A Crystallographer's Perspective. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **2016**, *72* (Pt 2), 72–95. https://doi.org/10.1107/S2053230X15024619.

(5) Pandurangan, A. P.; Blundell, T. L. Prediction of Impacts of Mutations on Protein Structure and Interactions: SDM, a Statistical Approach, and MCSM, Using Machine Learning. *Protein Sci.* **2020**, *29* (1), 247–257. https://doi.org/10.1002/pro.3774.

(6) Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W306–W310. https://doi.org/10.1093/nar/gki375.

(7) Delgado, J.; Radusky, L. G.; Cianferoni, D.; Serrano, L. FoldX 5.0: Working with RNA, Small Molecules and a New Graphical Interface. *Bioinformatics* **2019**, *35* (20),

4168–4169. https://doi.org/10.1093/bioinformatics/btz184.
(8) Pires, D. E. V.; Ascher, D. B.; Blundell, T. L. MCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures. *Bioinformatics* **2014**, *30* (3), 335–342. https://doi.org/10.1093/bioinformatics/btt691.
(9) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.
(10) Noid, W. G. Perspective: Coarse-Grained Models for Biomolecular Systems. *J. Chem. Phys.* **2013**, *139* (9), 090901. https://doi.org/10.1063/1.4818908.
(11) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834. https://doi.org/10.1021/ct700324x.
(12) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.
(13) Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W. A Relationship between Protein Stability and Protein Function. *Proc. Natl. Acad. Sci.* **1995**, *92* (2), 452–456. https://doi.org/10.1073/pnas.92.2.452.
(14) Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. A Backbone-Based Theory of Protein Folding. *Proc. Natl. Acad. Sci.* **2006**, *103* (45), 16623–16633. https://doi.org/10.1073/pnas.0606843103.
(15) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33* (suppl_2), W382–W388. https://doi.org/10.1093/nar/gki387.
(16) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.
(17) Ting, K. M. Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining*; Sammut, C., Webb, G. I., Eds.; Springer US: Boston, MA, 2017; pp 260–260. https://doi.org/10.1007/978-1-4899-7687-1_50.
(18) Sanavia, T.; Birolo, G.; Montanucci, L.; Turina, P.; Capriotti, E.; Fariselli, P. Limitations and Challenges in Protein Stability Prediction upon Genome Variations: Towards Future Applications in Precision Medicine. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1968–1979. https://doi.org/10.1016/j.csbj.2020.07.011.
(19) Dias, S.; Simões, T.; Fernandes, F.; Martins, A. M.; Ferreira, A.; Jorge, J.; Gomes, A. J. P. CavBench: A Benchmark for Protein Cavity Detection Methods. *PLOS ONE* **2019**, *14* (10), e0223596. https://doi.org/10.1371/journal.pone.0223596.
(20) UniProt Consortium, T. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46* (5), 2699–2699. https://doi.org/10.1093/nar/gky092.