

# **Aplicación de minería de datos y modelamiento matemático en ingeniería de proteínas**

**Diseño e implementación de nuevas metodologías para el  
estudio de mutaciones**



UNIVERSIDAD  
DE CHILE

**David Medina Ortiz**

Supervisor: Dr. Álvaro Olivera

Departamento de Ingeniería Química, Biotecnología y Materiales  
Universidad de Chile

Este trabajo es para obtener el grado de  
*Dr. en Ciencias de la Ingeniería*

June 2019



## **Abstract**

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. ...



# Tabla de contenidos

<b>Lista de figuras</b>	<b>vii</b>
<b>Lista de tablas</b>	<b>ix</b>
<b>1 Aplicaciones de la minería de datos en ingeniería de proteínas</b>	<b>1</b>
<b>2 Modelos predictivos asociados a mutaciones puntuales en proteínas</b>	<b>3</b>
2.1 Herramientas computacionales asociadas a evaluación de mutaciones . . . .	4
2.1.1 FoldX . . . . .	4
2.1.2 I-Mutant . . . . .	6
2.1.3 CUPSAT . . . . .	6
2.1.4 Dmutant . . . . .	6
2.1.5 MUpro . . . . .	6
2.1.6 MultiMutate . . . . .	6
2.1.7 SDM . . . . .	6
2.1.8 MOSST . . . . .	6
<b>3 Digitalizando propiedades fisicoquímicas de proteínas a partir de su secuencia lineal</b>	<b>7</b>
<b>4 Filogenética, propiedades fisicoquímicas y minería de datos aplicadas al diseño de mutaciones en secuencias de proteínas</b>	<b>9</b>
<b>5 Modelamiento matemático discreto aplicado al estudio de estructuras de proteínas.</b>	<b>11</b>
<b>6 Reconocimiento de patrones y extracción de información en sistemas complejos multi-dimensionales</b>	<b>13</b>

<b>7 Un caso de estudio completo: Aplicación de técnicas de minería de datos y reconocimiento de patrones para modelar el sistema de interacción antígeno anticuerpo</b>	<b>15</b>
<b>Referencias</b>	<b>17</b>

## **Lista de figuras**





## **Lista de tablas**



# **Chapter 1**

## **Aplicaciones de la minería de datos en ingeniería de proteínas**



## Chapter 2

# Modelos predictivos asociados a mutaciones puntuales en proteínas

El análisis del efecto de mutaciones puntuales en proteínas, es una de las problemáticas más estudiadas en los últimos años. Los estudios se enfocan principalmente, en la evaluación de cambios en la estabilidad de la proteína mediante la variación de energía libre que la mutación provoca [24, 18, 22, 19].

Diferentes modelos predictivos han sido desarrollados para poder predecir cambios de energía libre, en base a algoritmos de aprendizaje supervisado o mediante técnicas de minería de datos, y así, determinar el efecto de la mutación en set de proteínas de interés [21, 8, 4, 13, 26, 10, 7]. No obstante, en casos más específicos, se han desarrollado modelos para proteínas exclusivas con el fin de asociar la mutación a un rasgo clínico, particularmente, enfocado a casos de cáncer [11, 9], cambios en termo estabilidad [25], propiedades geométricas [3], entre las principales.

Sin importar el uso o la respuesta de los modelos, es necesario construir set de datos con ejemplos etiquetados, es decir, cuya respuesta sea conocida para poder entrenar modelos y así evaluar su desempeño. Los enfoques principales al desarrollo de descriptores se basan en propiedades fisicoquímicas y termodinámicas, así como también, el ambiente bajo el cual se encuentra la mutación [7], ya sea a partir de la información estructural o sólo considerando la secuencia lineal. Sin embargo, no son considerados, los componentes asociados a conceptos filogenéticos y la propensión a cambios de dicha mutación generando un gap entre ambos puntos de vista [17].

Dado a los modelos existentes y en vista a la necesidad de generar nuevos sistemas de predicción para mutaciones puntuales en proteínas, en respuesta al aumento considerable de reportes en los últimos años, se propone una nueva metodología para el diseño e implementación de modelos predictivos en mutaciones puntuales de proteínas.

Las mutaciones son descritas desde los puntos de vista estructural, termodinámico y filogenético. El desarrollo de los predictores es inspirados en el concepto de Meta Learning y es apoyado con técnicas estadísticas, tanto para la selección de modelos como para la evaluación de medidas de desempeño, entregando como resultado un conjunto de modelos para las mutaciones puntuales reportadas unificados en un único meta modelo.

Esta metodología ha sido evaluada para generar estimadores en diferentes proteínas con mutaciones reportadas con respuesta conocida, como por ejemplo: evaluando las diferencias de energía libre que provoca la mutación y clasificaciones para evaluar si el cambio aumenta o disminuye la estabilidad. A su vez, se implementaron modelos de clasificación para determinar la propensión clínica en un conjunto de mutaciones conocidas relacionados con el gen *pVHL*, responsable de la enfermedad von Hippel Lindau.

A continuación, se describen algunas herramientas computacionales y su significancia para este estudio a la hora de comparar y analizar los resultados obtenidos, seguido además, de los conceptos relacionados al aprendizaje supervisado, junto con la metodología propuesta, los resultados y discusiones del proceso, así como también su uso de esto en casos particulares.

## **2.1 Herramientas computacionales asociadas a evaluación de mutaciones**

Las herramientas computacionales asociadas a la evaluación de mutaciones puntuales se centran principalmente en el análisis de cómo ésta afecta a la estabilidad o la predicción de energía libre asociada a los residuos involucrados en la mutación. Sin embargo, a pesar de que el objetivo es el mismo, se centran en diferentes enfoques para abordar la problemática.

A continuación, se exponen algunas herramientas básicas en el estudio de estabilidad de proteínas, las cuales se aplicarán como métodos de comparación para los resultados obtenidos aplicando la metodología propuesta.

### **2.1.1 FoldX**

FoldX es una herramienta computacional, que implementa un campo de fuerza empírico desarrollado para la evaluación eficiente del efecto de las mutaciones sobre la estabilidad, el plegamiento y la dinámica de las proteínas y los ácidos nucleicos [24]. Se basa principalmente en el cálculo de energía libre a partir de estructuras 3D de macromoléculas. Sin embargo, permite además, estimar las posiciones de los protones y los puentes de hidrógeno.

La energía libre, es calculada utilizando la siguiente función

$$\Delta G = W_{vdw} \cdot \Delta G_{vdw} + W_{solvH} \cdot \Delta G_{solvH} + W_{solvP} \cdot \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{mc} \cdot T \cdot \Delta S_{mc} + W_{sc} \cdot T \cdot \Delta S_{sc}$$

- $\Delta G_{vdw}$  es la suma de las contribuciones de van der Waals de todos los átomo con respecto a la interacción con el solvente.
- $\Delta G_{solvH}$  y  $\Delta G_{solvP}$  son las diferencias en energía de solvatación para grupos apolares y polares respectivamente, cuando estos cambias desde el estado no plegado a plegado.
- $\Delta G_{hbond}$  es la diferencia de energía libre entre la formación de un enlace de hidrógeno intra-molecular y un inter-molecular.
- $\Delta G_{wb}$  es la energía libre de estabilización adicional proporcionada por una molécula de agua que hace más de un enlace de hidrógeno a la proteína que no se puede tener en cuenta con aproximaciones de solventes no explícitas [20].
- $\Delta G_{el}$  es la contribución electrostática de los grupos cargados, incluyendo las hélices dipolo.  $\Delta S_{mc}$  es el costo de la entropía de fijar la columna vertebral en el estado plegado; este término depende de la tendencia intrínseca de un aminoácido particular a adoptar ciertos ángulos diedros [15, 14].
- Finalmente  $\Delta S_{sc}$  es el costo de entropía de fijar una cadena lateran en una conformación particular [1].
- Si la estimación se desarrolla sobre proteínas oligoméricas o complejos de proteína, se adicionan dos términos a la contribución energética:  $\Delta G_{kon}$  que refleja el efecto de las interacciones electrostáticas en la constante de asociación *kon* (esto se aplica solo a las energías de enlace de la subunidad) [27] y  $\Delta S_{str}$  que es la pérdida de entropía traslacional y rotacional que se deriva de la formación del complejo. Este último término se cancela cuando observamos el efecto de mutaciones puntuales en complejos.
- Los valores de energía de  $\Delta G_{vdw}$ ,  $\Delta G_{solvH}$ ,  $\Delta G_{solvP}$  y  $\Delta G_{hbond}$  atribuidos a cada tipo de átomo se han derivado de un conjunto de datos experimentales, y  $\Delta S_{mc}$  y  $\Delta S_{sc}$  han sido considerados desde estimaciones teóricas.
- Los términos  $W_{vdw}$ ,  $W_{solvH}$ ,  $W_{solvP}$ ,  $W_{mc}$  y  $W_{sc}$  corresponden a los factores de ponderación aplicados a los términos de energía bruta. Todos son 1, excepto por la contribución de van der Waals que es de 0.33 (las contribuciones de van der Waals se derivan de la transferencia de energía de vapor a agua, mientras que en la proteína vamos de solvente a proteína).

Como entrada principal, recibe una estructura PDB<sup>1</sup> y dentro de los resultados más relevantes, se encuentran los cálculos de energía libre.

Ha sido usada en diferentes investigaciones, incluyendo el análisis de mutaciones puntuales aplicados a ingeniería de proteínas [6, 2], evaluación de mutaciones en genomas [23], análisis de termoestabilidad [5, 12], interacciones proteínas-DNA [16], entre las principales.

### **2.1.2 I-Mutant**

### **2.1.3 CUPSAT**

### **2.1.4 Dmutant**

### **2.1.5 MUpro**

### **2.1.6 MultiMutate**

### **2.1.7 SDM**

### **2.1.8 MOSST**

---

<sup>1</sup>Protein Data Bank



## **Chapter 3**

# **Digitalizando propiedades fisicoquímicas de proteínas a partir de su secuencia lineal**

Existen cerca de X proteínas reportadas en las bases de datos. Sin embargo, sólo un número limitado de ellas presentan cristal o estructura tridimensional reportada, lo cual dificulta diferentes estudios posibles a la hora de analizar mutaciones y cambios conformacionales que conlleva dicho cambio por medio de técnicas bioinformáticas que requieren la estructura de la proteína.



## **Chapter 4**

**Filogenética, propiedades fisicoquímicas  
y minería de datos aplicadas al diseño de  
mutaciones en secuencias de proteínas**



## **Chapter 5**

**Modelamiento matemático discreto  
aplicado al estudio de estructuras de  
proteínas.**



## **Chapter 6**

# **Reconocimiento de patrones y extracción de información en sistemas complejos multi-dimensionales**





## **Chapter 7**

**Un caso de estudio completo: Aplicación de técnicas de minería de datos y reconocimiento de patrones para modelar el sistema de interacción antígeno anticuerpo**



# Referencias

- [1] Abagyan, R. and Totrov, M. (1994). Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *Journal of molecular biology*, 235(3):983–1002.
- [2] Alibés, A., Nadra, A. D., De Masi, F., Bulyk, M. L., Serrano, L., and Stricher, F. (2010). Using protein design algorithms to understand the molecular basis of disease caused by protein-dna interactions: the pax6 example. *Nucleic Acids Research*, 38(21):7422–7431.
- [3] Barenboim, M., Masso, M., Vaisman, I. I., and Jamison, D. C. (2008). Statistical geometry based prediction of nonsynonymous snp functional effects using random forest and neuro-fuzzy classifiers. *Proteins: Structure, Function, and Bioinformatics*, 71(4):1930–1939.
- [4] Broom, A., Jacobi, Z., Trainor, K., and Meiering, E. M. (2017). Computational tools help improve protein stability but with a solubility tradeoff. *J Biol Chem*, 292(35):14349–14361. 28710274[pmid].
- [5] Buß, O., Muller, D., Jager, S., Rudat, J., and Rabe, K. S. (2018). Improvement in the thermostability of a b-amino acid converting o-transaminase by using foldx. *ChemBioChem*, 19(4):379–387.
- [6] Buß, O., Rudat, J., and Ochsenreither, K. (2018). Foldx as protein engineering tool: Better than random based approaches? *Computational and Structural Biotechnology Journal*, 16:25 – 33.
- [7] Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, 33(Web Server issue):W306–W310. 15980478[pmid].
- [8] Capriotti, E., Fariselli, P., Rossi, I., and Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, 9(2):S6.
- [9] Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., and Futreal, P. A. (2010). Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39(suppl\_1):D945–D950.
- [10] Getov, I., Petukh, M., and Alexov, E. (2016). Saafec: Predicting the effect of single point mutations on protein folding free energy using a knowledge-modified mm/pbsa approach. *Int J Mol Sci*, 17(4):512–512. 27070572[pmid].

- [11] Gossage, L., Pires, D., Olivera-Nappa, A., A. Asenjo, J., Bycroft, M., Blundell, T., and Eisen, T. (2014). An integrated computational approach can classify vhl missense mutations according to risk of clear cell renal carcinoma. *Human molecular genetics*, 23.
- [12] Heselpoth, R. D., Yin, Y., Moulton, J., and Nelson, D. C. (2015). Increasing the stability of the bacteriophage endolysin PlyC using rationale-based FoldX computational modeling. *Protein Engineering, Design and Selection*, 28(4):85–92.
- [13] Khan, S. and Vihinen, M. (2010). Performance of protein stability predictors. *Human Mutation*, 31(6):675–684.
- [14] Muñoz, V., Blanco, F. J., and Serrano, L. (1995). The hydrophobic-staple motif and a role for loop-residues in  $\alpha$ -helix stability and protein folding. *Nature structural biology*, 2(5):380.
- [15] Muñoz, V. and Serrano, L. (1996). Local versus nonlocal interactions in protein folding and stability—an experimentalist’s point of view. *Folding and Design*, 1(4):R71–R77.
- [16] Nadra, A. D., Serrano, L., and Alibés, A. (2011). Chapter one - dna-binding specificity prediction with foldx. In Voigt, C., editor, *Synthetic Biology, Part B*, volume 498 of *Methods in Enzymology*, pages 3 – 18. Academic Press.
- [17] Olivera-Nappa, A., Andrews, B. A., and Asenjo, J. A. (2011). Mutagenesis objective search and selection tool (mosst): an algorithm to predict structure-function related mutations in proteins. *BMC Bioinformatics*, 12(1):122.
- [18] Pandurangan, A. P., Ochoa-Montano, B., Ascher, D. B., and Blundell, T. L. (2017). Sdm: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*, 45(W1):W229–W235. 28525590[pmid].
- [19] Parthiban, V., Gromiha, M. M., and Schomburg, D. (2006). Cupsat: prediction of protein stability upon point mutations. *Nucleic Acids Res*, 34(Web Server issue):W239–W242. 16845001[pmid].
- [20] Petukhov, M., Cregut, D., Soares, C. M., and Serrano, L. (1999). Local water bridges and protein conformational stability. *Protein Science*, 8(10):1982–1989.
- [21] Quan, L., Lv, Q., and Zhang, Y. (2016). Strum: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, 32(19):2936–2946. 27318206[pmid].
- [22] Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier.
- [23] Sánchez, I. E., Beltrao, P., Stricher, F., Schymkowitz, J., Ferkinghoff-Borg, J., Rousseau, F., and Serrano, L. (2008). Genome-wide prediction of sh2 domain targets using structural information and the foldx algorithm. *PLOS Computational Biology*, 4(4):e1000052.
- [24] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The foldx web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue):W382–W388. 15980494[pmid].

- 
- [25] Tian, J., Wu, N., Chu, X., and Fan, Y. (2010). Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics*, 11(1):370.
- [26] Vaisman, I. I. and Masso, M. (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, 24(18):2002–2009.
- [27] Vijayakumar, M., Wong, K.-Y., Schreiber, G., Fersht, A. R., Szabo, A., and Zhou, H.-X. (1998). Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. *Journal of molecular biology*, 278(5):1015–1024.

