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

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 [26, 20, 24, 21].

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 [23, 10, 5, 15, 28, 12, 9]. 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 [13, 11], cambios en termo estabilidad [27], 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 basados en algoritmos de aprendizaje supervisado 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 [9], 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 [19].

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 la sustitución de residuos 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 *p*VHL, responsable de la enfermedad von Hippel Lindau, con el fin de exponer la versatilidad de la metodología.

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 [26]. 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 expresión matemática de aportes energéticos:

$\Delta G = Wvdw \cdot \Delta Gvdw + WsolvH \cdot \Delta GsolvH + WsolvP \cdot \Delta GsolvP + \Delta Gwb + \Delta Ghbond + \Delta Gel + \Delta GKon + Wmc \cdot T \cdot \Delta Smc + Wsc \cdot T \cdot \Delta Ssc$

Los componentes se definen a continuación.

- $\Delta Gvdw$ es la suma de las contribuciones de van der Waals de todos los átomo con respecto a la interacción con el solvente.
- $\Delta GsolvH$ y $\Delta GsolvP$ son las diferencias en energía de solvatación para grupos apolares y polares respectivamente, cuando estos cambian desde el estado no plegado a plegado.
- $\Delta Ghbond$ es la diferencia de energía libre entre la formación de un enlace de hidrógeno intra-molecular y un inter-molecular.
- ΔGwb 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 [22].
- ΔGel es la contribución electrostática de los grupos cargados, incluyendo las hélices dipolo.
- ΔSmc es el costo de la entropía de fijar el back-bone en el estado plegado; este término depende de la tendencia intrínseca de un aminoácido particular a adoptar ciertos ángulos diedros [17, 16].
- Finalmente ΔSsc es el costo de entropía de fijar una cadena lateral 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: $\Delta Gkon$ 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) [29] y ΔStr 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 $\Delta Gvdw$, $\Delta GsolvH$, $\Delta GsolvP$ y $\Delta Ghbond$ atribuidos a cada tipo de átomo se han derivado de un conjunto de datos experimentales, y ΔSmc y ΔSmc han sido considerados desde estimaciones teóricas.

- Los términos $Wvdw$, $WsolvH$, $WsolvP$, Wmc y Wsc 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[1] 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 [7, 2], evaluación de mutaciones en genomas [25], análisis de termoestabilidad [6, 14], interacciones proteínas-DNA [18], entre las principales, razón por la cual, es considerada como una herramienta común a la hora de la evaluación de mutaciones y una referente para comparar resultados de nuevas herramientas o métodos computacionales.

2.1.2 I-Mutant

I-Mutant, es una familia de software basados en algoritmos de aprendizaje supervisado para la predicción automática de estabilidad de proteínas ante cambios de residuos o sustituciones expresadas en mutaciones puntuales [8], las cuales se reflejan en los cambios de energía libre. Emplea como algoritmo para entrenamiento de modelos, Support Vector Machine (SVM), permitiendo la evaluación de las mutaciones desde la secuencia lineal de proteína o a su vez desde la estructura 3D en formato PDB.

El método fue entrenado y testeado desde la base de datos ProTherm [4], la cual representa el mayor repositorio de experimentos termodinámicos con respuestas en energía libre basados en cambios de estabilidad de la proteína para mutaciones, en diferentes condiciones.

Actualmente I-Mutant permite la clasificación de la estabilidad de la mutación y a su vez facilita la predicción de los cambios de energía libre $\Delta\Delta G$. Las medidas de desempeño se diferencian dependiendo del uso de I-Mutant y del tipo de set de datos. Si se trabaja con datos de secuencias lineales presenta una accuracy de un 77% y un coeficiente de relación de un 0.62 con un error asociado de 1.45 kcal/mol. Para el caso en que los datos procedan de información estructural, los desempeños mejoran de manera no significativa, obteniendo una accuracy de un 80% y un coeficiente de relación de un 0.71 con un error asociado de 1.30 kcal/mol.

[1] Protein Data Bank. Formato para la exposición de macromoléculas u estructuras en torno a coordenadas espaciales, obtenidas desde una cristalografía de rayos X, resnonancia magnética nuclear, o a través de modelos computacionales.

Si bien, es uno de los métodos más utilizados, el hecho de utilizar Support Vector Machine como algoritmo de aprendizaje supervisado para el entrenamiento de modelos, limita un poco para los set de datos con alta no linealidad, dado a que el algoritmo sólo traza hiperplanos. Esto podría provoca sobre ajuste o generar bajos desempeños.

2.1.3 CUPSAT

CUPSAT (Cologne University Protein Stability Analysis Tool) es una herramienta web para analizar y predecir la estabilidad de la proteína frente a cambios o sustituciones puntuales de residuos. La herramienta utiliza información estructural específica de los átomos participantes en la mutación, tales como: ángulos de torsión y potenciales de energía, con el fin de predecir los cambios en diferencia de energía que representa la sustitución, expresados en forma de $\Delta\Delta G$ [21].

Como requisitos para su uso, es necesario la estructura en formato PDB y la posición del residuo a ser mutado. Como resultado, entrega información sobre el sitio de la mutación, principalmente accesibilidad al solvente, estructura secundaria y ángulos de torsión. Además, entrega información detallada sobre las 19 posibles mutaciones para el residuo objetivo.

La herramienta fue testeada utilizado 1538 mutaciones desde denaturaciones térmicas y 1603 denaturaciones aplicando técnicas químicas. Presentado un desempeño mayor al 80% de accuracy.

2.1.4 Dmutant

Otra [30]

2.1.5 MUpro

2.1.6 MultiMutate

2.1.7 SDM

2.1.8 MOSST

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] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. (2004). Protherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research*, 32(suppl_1):D120–D121.

[5] 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].

[6] 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.

[7] 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.

[8] Capriotti, E., Fariselli, P., and Casadio, R. (2005a). I-mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, 33(suppl_2):W306–W310.

[9] Capriotti, E., Fariselli, P., and Casadio, R. (2005b). 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].

[10] 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.

[11] 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.

[12] 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].

[13] 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.

[14] Heselpoth, R. D., Yin, Y., Moult, 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.

[15] Khan, S. and Vihinen, M. (2010). Performance of protein stability predictors. *Human Mutation*, 31(6):675–684.

[16] Muñoz, V., Blanco, F. J., and Serrano, L. (1995). The hydrophobic-staple motif and a role for loop-residues in α-helix stability and protein folding. *Nature structural biology*, 2(5):380.

[17] 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.

[18] 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.

[19] 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.

[20] Pandurangan, A. P., Ochoa-Montaño, 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].

[21] 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].

[22] Petukhov, M., Cregut, D., Soares, C. M., and Serrano, L. (1999). Local water bridges and protein conformational stability. *Protein Science*, 8(10):1982–1989.

[23] 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].

[24] 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.

[25] 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.

[26] 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].

[27] 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.

[28] 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.

[29] 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.

[30] Zhou, H. and Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, 11(11):2714–2726.