ANÁLISIS DE AUTOREACTIVIDAD DE ANTICUERPOS LEUCÉMICOS SOPORTADO POR ESTRATEGIAS DE INTELIGENCIA ARTIFICIAL

YASNA ISABEL BARRERA SAAVEDRA
INGENIERO CIVIL EN BIOINFORMÁTICA

RESUMEN

Los antígenos son moléculas externas reconocidas por el organismo de variada estructura y naturaleza. El sistema inmune ha desarrollado técnicas de reconocimiento para estos agentes patógenos, representando diferentes mecanismos de defensa contra una posible infección, siendo los anticuerpos los responsables de esta detección. Predecir qué anticuerpo reconocerá a un antígeno, o estimar a nivel cualitativo la intensidad de la interacción que se producirá, es una tarea ardua y compleja, representando un gran desafío en el área inmunológica. Debido a que los antígenos pueden ser distintos tipos de moléculas, y tener procedencia en diferentes patógenos, la forma en la cual un anticuerpo reconoce un conjunto de antígenos con diversas intensidades de interacción, es una pregunta que se ha abordado desde diferentes perspectivas. Por otra parte, el organismo ha desarrollado estrategias para reconocer moléculas externas de aquellas propias. Esto evita que se genere una respuesta inmune sobre tejidos en el organismo. Las moléculas propias del organismo que desencadenan esta respuesta son denominadas auto antígenos, y al proceso de presentar defensas contra estas moléculas se le denomina auto reactividad. El análisis de auto antígenos es de gran relevancia, tanto para el estudio de enfermedades auto inmunes, como para enfermedades relacionadas a células propias del organismo. En el caso de la leucemia, un tipo de cáncer que afecta a células del tejido sanguíneo, el estudio de la auto reactividad y la interacción entre auto antígenos y anticuerpos es de gran relevancia para el diseño y propuestas que permitan diagnosticar y tratar esta enfermedad. Gran parte de los estudios de interacción entre auto antígenos y anticuerpos se han realizado utilizando técnicas experimentales. No obstante, diversos enfoques in-silico han sido desarrollados empleando diferentes herramientas computacionales como docking o simulación molecular para cálculos de energía libre y visualización de interacciones. Pese a

su gran utilidad, estas técnicas poseen un alto costo asociados a la necesidad de material experimental, necesidad de poseer estructuras definidas o modelos confiables, elevados tiempos de simulación, entre otros. De esta forma la aplicación de técnicas de machine learning y diversos métodos de codificación representan una alternativa potente al problema de reconocimiento de interacción entre proteína-proteína, particularmente, a secuencias de auto antígenos y anticuerpos de leucemia. A partir de la información de interacciones entre 45 secuencias de cadena pesada de anticuerpos y cerca de 8000 secuencias de auto antígenos, Se diseñó e implemento un sistema predictivo ensamblado cualitativo del nivel de intensidad de la interacción entre auto antígenos y cadenas pesadas de anticuerpos. Como estrategias de entrenamiento de modelos predictivos, se combinaron variados métodos de representación de proteínas, principalmente Natural Language Processing y propiedades fisicoquímicas, con diferentes algoritmos de aprendizaje supervisado logrando un predictor ensamblado con un rendimiento del 81% de accuracy. Se aplicaron diferentes estrategias de validación que permiten demostrar la robustez del sistema predictivo propuesto, incluyendo sistemas de validación cruzada y métodos propios basados en estrategias Leave One Antibody Out. Adicionalmente, se diseñó e implemento un conjunto de colecciones de moléculas inmunológicas integradas en un único sistema de base de datos, al cual acoplado a una estrategia de clasificación filogenética, se diseña e implementa una estrategia de clasificación de secuencias de autoantígenos basado en propiedades descriptivas, funcionales y componentes filogenéticos. La combinación del conjunto de colecciones con el sistema de clasificación, en conjunto con el sistema ensamblado predictivo, facilita el diseño de estrategias de identificación de secuencias autoantígenos y su evaluación contra anticuerpos leucémicos, brindando los soportes iniciales para herramientas de diseño y descubrimiento de antígenos/anticuerpos que cumplan con características relevantes para el problema de la leucemia, denotando la usabilidad de métodos computacionales en problemas complejos de la ingeniería médica.

ABSTRACT

Antigens are external molecules of varied structures and nature recognized by the body. The immune system has developed recognition techniques for these pathogens, representing different defense mechanisms against possible infection, the antibodies responsible for this detection. Predicting which antibody will recognize an antigen or estimating the intensity of the interaction that will occur at a qualitative level is an arduous and complex task, representing a significant challenge in the immunological area. Because antigens can be different types of molecules and have origins in various pathogens, how an antibody recognizes a set of antigens with varying intensities of interaction is a question that has been approached from different perspectives. On the other hand, the organism has developed strategies to recognize external molecules on its own. This prevents an immune response from being generated on tissues in the body. The body's own molecules that trigger this response are called self-antigens, and the process of presenting defenses against these molecules is called self-reactivity. The analysis of self-antigens is of great relevance, both for studying autoimmune diseases and for diseases related to the body's own cells. In leukemia, a type of cancer that affects cells of the blood tissue, the study of self-reactivity and the interaction between self-antigens and antibodies is of great relevance for the design of proposals that allow the diagnosis and treatment of this disease. Much of the interaction studies between self-antigens and antibodies have been carried out using experimental techniques. However, various in-silico approaches have been developed using different computational tools such as docking or molecular simulation for free energy calculations and interactive visualization. Despite their great utility, these techniques have a high cost associated with the need for experimental material, the need to have defined structures or reliable models, high simulation times, among others. In this way, the application of machine learning techniques and various coding methods represent a powerful alternative to the problem of protein-protein interaction recognition, particularly to leukemia self-antigen and antibody sequences. From the information of interactions between 45

sequences of antibodies heavy chain and about 8000 sequences of self-antigens, a qualitative assembled predictive system for the level of intensity of the interaction between self-antigens and heavy chains of antibodies was designed and implemented. As predictive model training strategies, various protein representation methods were combined, mainly Natural Language Processing and physicochemical properties, with different supervised learning algorithms, achieving an assembled predictor with a performance of 81% accuracy. Different validation strategies were applied to demonstrate the robustness of the proposed predictive system, including cross-validation systems and proprietary methods based on Leave One Antibody Out strategies. Additionally, a set of collections of immunological molecules integrated into a single database system was designed and implemented. Coupled with a phylogenetic classification strategy, a method for classifying self-antigen sequences based on descriptive properties was designed and implemented. This method uses different functional properties and phylogenetic components to estimate the relation of new sequences with the set of self-antigen sequences. The combination of the group of collections with the classification system, in association with the assembled predictive system, facilitates the design of strategies for the identification of selfantigen sequences and their evaluation against leukemic antibodies, providing the initial supports for tools of creation and discovery of antigens/antibodies that meet relevant characteristics for the leukemia problem, denoting the usability of computational methods in complex issues of medical engineering.