

## TITULO DEL PROYECTO

“Aplicación de algoritmos de cómputo inteligente a problemas en Bioinformática.”

## DESCRIPCIÓN DEL PROYECTO

### Resumen

Describe de manera general la problemática que abordará en su proyecto de investigación, cómo la pretende resolver y sus posibles resultados, máximo una cuartilla.

En general, hay varias tareas fundamentales en el análisis de secuencias biológicas [Yasubumi Sakakibara, 2005]: 1) Cálculo de alineaciones múltiples; 2) Representación de motivos; 3) Clasificación de proteínas y predicción de sus estructuras y funciones. 4) Proporcionar un marco de trabajo unificado para varios modelos de diferencias para alineaciones pares de secuencias. 5) Encontrar regiones de codificadas de genes en secuencias de genomas. 6) Analizar regiones originadoras y sus regulaciones de transcripción. 7) Identificar secuencias de señales para localización de proteínas y sus interacciones. En este proyecto nos enfocaremos a clasificar las proteínas y predecir estructuras de RNA utilizando algoritmos de cómputo inteligente como los algoritmos evolutivos, además de aplicar algoritmos del lenguaje natural al problema de análisis de secuencias biológicas. Uno de los objetivos de la bioinformática es *detectar regularidades* en secuencias de cadenas biológicas. Nuevos descubrimientos dentro del área relacionada con la adquisición del lenguaje natural ha probado que los niños de ocho meses de edad pueden detectar regularidades lingüísticas y aprender estadísticas simples para el reconocimiento de los límites de las palabras en una secuencia de voz continua [Saffran et. Al., 96]. Puesto que el lenguaje que un niño tiene que aprender es desconocido y complejo como la secuencia del ADN o ARN sería para nosotros, tal vez no nos sorprendería que técnicas de lenguaje puedan ser útiles para revelar regularidades similares en los datos genómicos.

Un problema recurrente regularmente en el análisis de proteínas y secuencias de ADN es la *redundancia de los datos*. Muchas cadenas dentro de las proteínas o bases de datos genómicas representan miembros de familias de genes o proteínas, o versiones de genes homólogos encontrados en diferentes organismos. Varios grupos pueden tener la misma secuencia, por lo tanto, las cadenas pueden ser más o menos parecidas, si no idénticas; sin embargo, diferencias significativas pueden reflejar organismos genuinos o variaciones específicas de tejidos. El uso de conjuntos de datos redundantes implica una fuente potencial de errores estadísticos. Si existen muchos datos redundantes en una secuencia de datos, los algoritmos de predicción tendrán problemas para generalizar y se sobre-ajustarán a los datos de entrenamiento, esto provocarán

que el algoritmo de predicción o clasificación falle al predecir o clasificar nuevas secuencias. Lo que sugeriría es realizar una limpieza de datos redundantes por medio de algoritmos de aprendizaje automático.

Otro problema importante en el análisis de secuencias biológicas es el *tamaño* de dichas secuencias. Los algoritmos clásicos de búsqueda de patrones por lo regular son efectivos para cadenas cortas y fallan al tratar de explorar conjuntos de datos muy grandes. Algunos algoritmos de aprendizaje automático pueden tratar con grandes conjuntos de datos, sin embargo, aquí se presenta un área de oportunidad para generar nuevos algoritmos de aprendizaje automático para trabajar con grandes conjuntos de datos.