

ÍNDICE

ÍNDICE	4
ÍNDICE DE FIGURAS	6
INDICE DE TABLAS	7
RESUMEN	8
ABSTRACT	9
INTRODUCCIÓN	10
1. Metaloproteínas	10
1.1 Superfamilia Fur	10
1.2 Tipos de proteínas Fur	11
1.2.1 Ferric uptake regulator (Fur)	12
1.2.2 Zinc uptake regulator (Zur)	12
1.2.3 Peroxide stress sensing Regulator (PerR)	13
1.2.4 Manganese uptake regulator (Mur)	14
1.2.5 Niquel-responsive regulator (Nur)	14
1.3 Sitio de unión al ADN	15
2. Predicción de sitios de unión al ADN	19
2.1 Machine Learning	20
2.2 Métodos de predicción basados en machine learning	20
3. Descriptores Moleculares	22
3.1 Basados en secuencia aminoacídica	23
3.2 Basados en estructura tridimensional	24
HIPÓTESIS Y OBJETIVOS	26
1. Hipótesis	26
2. Objetivo general	26
3. Objetivos específicos	26
MATERIALES Y MÉTODOS	27
1. Construcción del set de datos	27
1.1 Descargar estructuras de proteínas unidas al ADN.	28
1.2 Eliminación de redundancia del set de datos	29
1.3 Selección de fragmentos con y sin capacidad de unión al ADN.	29
2. Caracterización del set de datos	30
2.1 Descriptores de secuencia aminoacídica	31
2.2 Descriptores de estructura tridimensional	32
2.3 Preprocesamiento y limpieza de datos	33
3. Desarrollo de los modelos de predicción	33
3.1 Selección de características	33

3.2 Modelos de clasificación SVM Y RF	34
RESULTADOS	37
1. Resultados objetivo específico 1: Construir un set de datos que contenga información de secuencia aminoacídica y/o de estructural de los fragmentos de unión al ADN en proteínas de la superfamilia Fur y otros factores de transcripción.	37
1.1 Estructuras utilizadas en el estudio.	37
2. Resultados objetivo específico 2: Caracterizar el set de datos con descriptores moleculares a nivel de secuencia aminoacídica y estructural.	40
3. Resultados objetivo específico 3: Evaluar modelos Máquinas de Vectores de Soporte y Random Forest para su aplicación en la predicción de sitios de unión al ADN.	41
3.1 Selección de características y aplicación de modelos.	41
3.1.1 Atributos de secuencia	42
3.1.2 Atributos de estructura	46
DISCUSIÓN	50
CONCLUSIONES	54
REFERENCIAS	56
ANEXOS	64
1. Anexo 1: Ranking de atributos de secuencia.	64
2. Anexo 2: Selección de atributos de secuencia.	67
3. Anexo 3: Ranking de atributos de estructura.	68
4. Anexo 4: Selección de atributos de secuencia.	70

ÍNDICE DE FIGURAS

<i>Figura 1. Representación del modelo estructural de una proteína Fur dimerica.</i>	<i>11</i>
<i>Figura 2. Procesos celulares modulados por reguladores de absorción férricos.</i>	<i>15</i>
<i>Figura 3. Representación de la estructura MRS2-Fur (código PDB: 4RB2).</i>	<i>18</i>
<i>Figura 4. Número de complejos proteína-ADN liberadas en PDB cada año.</i>	<i>19</i>
<i>Figura 5. Número de estructuras proteicas liberadas en PDB cada año.</i>	<i>25</i>
<i>Figura 6. Diagrama de la metodología general que se aplicará en esta investigación.</i>	<i>27</i>
<i>Figura 7. Estructuras utilizadas en la investigación.</i>	<i>37</i>
<i>Figura 8. Alineamiento estructural de la zona de unión al ADN.</i>	<i>39</i>
<i>Figura 9. Comparación estructural de zona de unión.</i>	<i>40</i>
<i>Figura 10. Atributos de secuencia acumulados.</i>	<i>43</i>
<i>Figura 11. Rendimiento de modelos SVM y RF en atributos de secuencia.</i>	<i>45</i>
<i>Figura 12. Atributos de estructura acumulados.</i>	<i>46</i>
<i>Figura 13. Rendimiento de modelos SVM y RF en atributos estructurales.</i>	<i>48</i>

INDICE DE TABLAS

<i>Tabla 1. Matriz de confusión.</i>	<u>35</u>
<i>Tabla 2. Matriz de distancia p de aminoácidos.</i>	<u>38</u>
<i>Tabla 3. Resultado de los modelos SMV y RF.</i>	<u>48</u>