

## II. ÍNDICE DE CONTENIDOS

I. AGRADECIMIENTOS.....	i
II. ÍNDICE GENERAL.....	ii
III. ÍNDICE DE TABLAS.....	iv
IV. ÍNDICE DE FIGURAS.....	iv
V. ABREVIATURAS.....	v
VI. RESUMEN.....	vi
VII. ABSTRACT.....	vii
1. INTRODUCCIÓN.....	1
1.1. Secuencias de Inserción (IS).....	1
1.2. Nomenclatura IS.....	3
1.3. Bases de Datos IS.....	3
1.4. Desafío.....	4
1.5. Clasificación Automática.....	5
1.6. <i>Clustering</i> Jerárquico.....	5
1.7. <i>Clustering</i> Particional.....	7
2. HIPOTESIS DEL TRABAJO.....	8
2.1. Problema Observado.....	8
2.2. Solucion Propuesta.....	8
3. OBJETIVOS.....	8
3.1. Objetivo General.....	8
3.2. Objetivos Específicos.....	8
4. MATERIALES Y MÉTODOS.....	9
4.1. Secuencias y su Preprocesamiento.....	10
4.2. Matrices de Distancia.....	13
4.2.1. <i>European Molecular Biology Open Software Suite</i> (EMBOSS).....	13
4.3. Programas o Algoritmos de <i>Clusterings</i> .....	14
4.3.1. Blastclust.....	14
4.3.2. CD-HIT.....	15
4.3.3. <i>Spectral Clustering of Protein Sequences</i> (SCPS).....	16

4.3.4. <i>Markov Clustering</i> (MCL) .....	17
4.3.5. <i>Unweighted Pair Group Method with Arithmetic Mean</i> (UPGMA) .....	18
4.3.6. K-means .....	19
4.4. Algoritmo de Comparacion de <i>clusterings</i> .....	21
4.4.1. Variación de la Información (VI) .....	21
5. RESULTADOS.....	25
5.1. Resultados Blastclust.....	25
5.2. Resultados CD-HIT .....	29
5.3. Resultados SCPS .....	32
5.4. Resultados MCL .....	36
5.5. Resultados UPGMA.....	39
5.6. Resultados K-means .....	42
6. DISCUSIÓN Y CONCLUSIÓN .....	45
6.1. Discusiones y Conclusiones a partir de Tablas.....	45
6.2. Discusiones y Conclusiones a partir de Histogramas.....	47
6.3. Discusiones y Conclusiones Generales .....	49
7. REFERENCIAS.....	53
8. ANEXOS .....	55
9.1. Scripts .....	56
9.3. Creación archivo arff.....	78