UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Algoritmo memético basado en conocimiento biológico para el problema de predicción de la estructura tridimensional de la proteína

Propuesta de tema de titulación

Nombre: Camilo Farfán Pérez

RUT: 17.150.968-8

Carrera: Ingeniería Civil Informática

Año: 2013

Teléfono: (+569) 86401812

E-mail:camilo.farfan@usach.cl

Profesor Guía: Mario Inostroza Ponta

ÍNDICE DE CONTENIDOS

1. Objetivos del proyecto 1

1.1 Objetivo general 1

1.2 Objetivos específicos 1

2. Descripción del problema 1

2.1 Motivación 1

2.2 Dominio del problema 2

2.2.1 Estructuras de Proteínas 3

2.3 Enunciado del problema 5

3. Análisis de la solución 6

3.1 Estado del arte 6

3.1.1 Métodos *ab initio* sin información 6

3.1.2 Métodos *ab initio* con información 7

3.1.3 Métodos de enhebrado de secuencias 7

3.1.4 Métodos de análisis comparativo por homología 7

3.2 Alternativas de solución 7

3.3 Justificación de la solución propuesta 8

4. Descripción de la solución propuesta 9

4.1 Propósitos de la solución 9

4.2 Características de la solución 9

4.3 Alcances y limitaciones de la solución 10

4.4 Descripción de los resultados o evaluación de la solución 10

5. Metodología, herramientas y ambiente de desarrollo 11

5.1 Metodología a usar 11

5.2 Herramientas de desarrollo 12

5.2.1 Herramientas de software 12

5.2.2 Herramientas de hardware 12
5.3 Ambiente de desarrollo..... 13
6. Plan de trabajo 13
7. Referencias 15

1. OBJETIVOS DEL PROYECTO

1.1 Objetivo general

El objetivo general del proyecto es: la implementación de un algoritmo memético que incorpore conocimiento biológico y que además permita encontrar soluciones de mejor calidad para el problema 3-D PSP.

1.2 Objetivos específicos

Los objetivos específicos relacionados al proyecto son:

1. Reducción del espacio de búsqueda conformacional de la proteína, mediante el uso de una APL (*Angle Probability List*, lista de probabilidades de ocurrencia) de pares de ángulos (φ,ψ).
2. Diseño e implementación del algoritmo memético que use información biológica probable entregada por APL.
3. Diseño e implementación de un operador de búsqueda local para el algoritmo memético.
4. Comparación y análisis de resultados del experimento con datos reales conocidos.

2. DESCRIPCIÓN DEL PROBLEMA

2.1 Motivación

Con el pasar de los años la tecnología avanza a pasos agigantados, lo que ha provocado el desarrollo de distintas áreas de la ciencia. Una de estas áreas es la Biología, siendo capaz de generar grandes volúmenes de información que debe ser evaluada y estudiada. Es así como nace la Bioinformática, presentándose como el área que se centra en apoyar a la biología mediante el análisis de datos usando técnicas y

métodos propios de la informática (Cohen, 2004) (Dorn, Inostroza-Ponta, Buriol, & Verli, 2013).

Dentro de la Bioinformática, se encuentra la Bioinformática Estructural que se encarga de la investigación de la predicción de las estructuras tridimensionales de proteínas, problema conocido como 3-D PSP Problem *(Three-dimensional Protein Structure Prediction)*. El conocer la estructura de una secuencia de aminoácidos facilita la investigación de los procesos biológicos asociados. Por ejemplo, a través del análisis de la estructura 3-D de una proteína, es posible identificar las regiones de conexión en las que se podría ensamblar un inhibidor o activador de funciones biológicas, esto corresponde a la base del desarrollo de fármacos y drogas (Dorn, Inostroza-Ponta, Buriol, & Verli, 2013).

No obstante, a pesar de estos avances y de la mejora continua, este problema sigue siendo costoso computacionalmente debido a la simulación de interacciones de millones de moléculas en un medio celular, además se debe tomar en cuenta la siguiente restricción biológica: la formación de la estructura se debe realizar usando el mínimo de energía posible (Cohen, 2004). Es por ello que existen distintos enfoques para obtener predicciones que sacrifican precisión por tiempo de ejecución.

Estas estrategias pueden lograr una precisión sobre el 90% en la predicción de estructuras tridimensionales de proteínas basado en el uso de métodos heurísticos (Dor & Zhou, 2007). Sin embargo, continúan siendo costosos debido al espacio de búsqueda de soluciones.

Por lo tanto es necesario buscar un enfoque que permita desarrollar y probar algoritmos que reduzcan dicho espacio de búsqueda manteniendo o aumentando los porcentajes de precisión (calidad biológica) pero disminuyendo el tiempo de ejecución.

2.2 Dominio del problema

El problema de predicción de la estructura tridimensional de la proteína, que se resume en encontrar la conformación molecular con el mínimo de energía posible, pertenece al grupo NP-Completo (Berger & Leighton, 1998) y se enmarca como uno de los principales desafíos en el área de la Bioinformática.

Las proteínas son largas secuencias compuestas por 20 tipos diferentes de aminoácidos que están unidos por enlaces peptídicos. Los enlaces peptídicos son la consecuencia de la unión de un grupo carboxilo de un aminoácido con el grupo amino de otro produciendose la liberación una molécula de agua (Lehninger, Nelson, & Cox, 2005).

Cada vez que se realiza uno de los enlaces mencionados, la cadena comienza a crecer y a tomar forma debido a las fuerzas atómicas inherentes a estos fenómenos. En consecuencia, las estructuras comienzan a variar en base a los ángulos de torsión generados por estos enlaces, específicamente los ángulos φ y ψ, producidos por la unión de un N y un C, y un C_α y un C respectivamente.

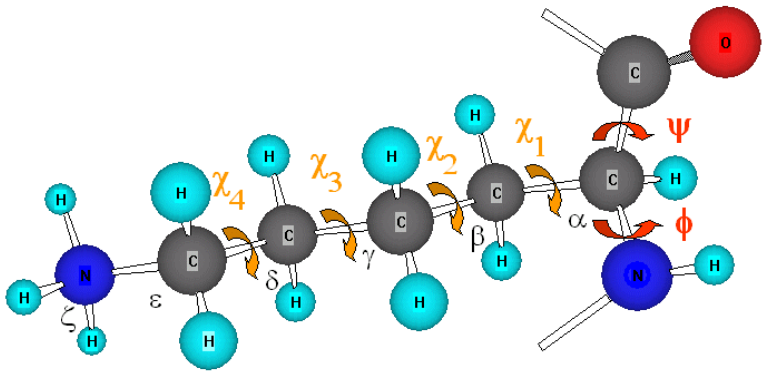

Figura 2-1: Ángulos phi y psi

A raíz de lo anterior, estas cadenas se encuentras plegadas o enrolladas con una formación específica tridimensional, que pueden ser clasificadas en los siguientes 4 niveles.

2.2.1 Estructuras de Proteínas

Estructuras Primarias

Una estructura primaria de una proteína es descrita por su secuencia linear de residuos de aminoácidos conectados a través de enlaces peptídicos (Lesk, 2002).

Estructuras Secundarias

La estructura secundaria se forma por la presencia de patrones de enlaces de hidrógeno entre los átomos del grupo amino y los átomos de oxígeno del grupo carboxílo del polipéptido.

La organización estable de residuos de aminoácidos de una proteína forman tipos de estructuras que son identificables (Lehninger, Nelson, & Cox, 2005). Las estructuras secundarias más regulares son las Hélices (*α-helix*) y las Hojas (*β-pleated sheet*) (Pauling, Corey, & Branson, 1951).

Estructuras Terciarias

Las estructuras terciarias son conocidas también como el estado nativo o funcional de la proteína (Lesk, 2002), corresponden a la unión de varias estructuras secundarias.

Esta estructura se forma producto de variaciones termodinámicas, es decir, es el resultado de las interacciones de, enlaces covalentes, hidrógeno, interacciones hidrofóbicas, electrostáticas, Van Der Waals y de fuerzas repulsivas (Lodish, y otros, 1990).

Estructuras Cuaternarias

Las estructuras cuaternarias son la unión de estructuras terciarias y secundarias, su modelamiento está dado por la interacción entre ellas.

Para Agosto del 2012, se conocían alrededor de 153.253.314 millones de secuencias de proteínas contra 83.266 estructuras tridimensionales conocidas (RefSeq: NCBI Reference Sequence Database, 2012). Esto evidencia el vacío que existe entre las secuencias y estructuras generadas debidoa la complejidad y costo de resolver el problema.

En la siguiente figura se puede apreciar los distintos niveles estructurales.

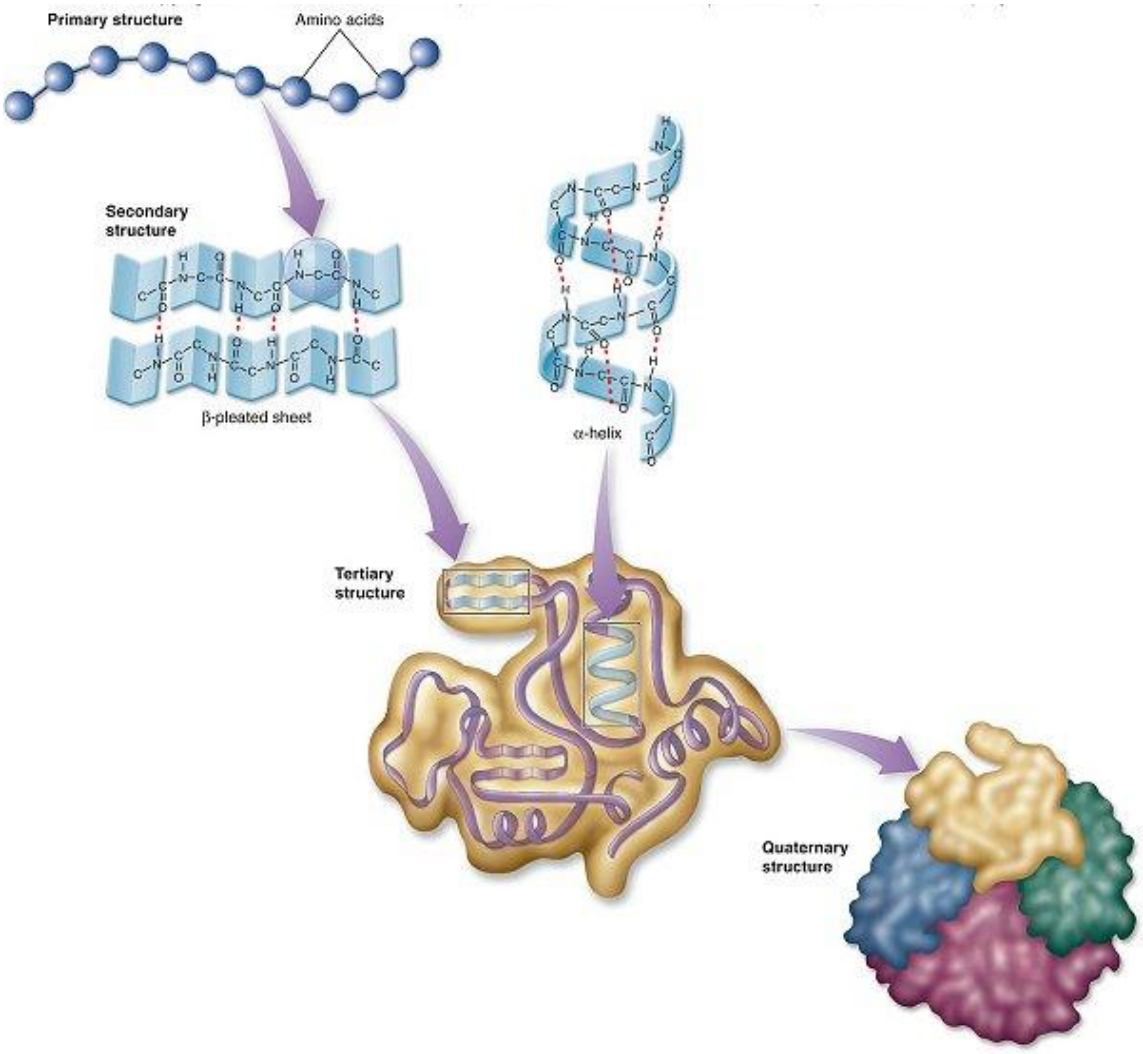

Figura 2-2: Niveles estructurales de proteínas

2.3 Enunciado del problema

Dada una secuencia de aminoácidos se debe predecir la estructura tridimensional (estructura terciaria) de la proteína que representa.

3. ANÁLISIS DE LA SOLUCIÓN

3.1 Estado del arte

El problema 3-D PSP (*Three Dimensional Protein Structure Prediction*) fue abordado por los años 60 y principios de los 70 con la introducción de métodos de predicción de estructuras secundarias (Froimowitz & Fasman, 1974), estos estaban enfocados en identificar hélices α probables basados principalmente en los modelos de transición hélice-bobinas. Posteriormente, en los años 70 se introdujo el concepto de hojas beta y de evaluaciones estadísticas sobre estructuras resueltas conocidas, estos métodos como máximo lograban un 60 a 65% de precisión (DM, 2004). No obstante, debido a la cantidad de información disponible y a los métodos de aprendizaje automático (como las redes neuronales o máquinas de vector de soporte), se puede lograr una precisión de hasta un 90% (Dor & Zhou, 2007).

Hoy en día, la predicción de estructuras terciarias es más importante que nunca debido a los avances que se han obtenido sobre el Genoma, ya que las proteínas son el resultado de la expresión genética. Existen métodos inversos que toman proteínas en su estructura cuaternaria e indican la forma y cadenas de aminoácidos que las conforman (Cristalografía de Rayos X y espectroscopia de RMN). Sin embargo estos métodos son muy costosos.

Los principales desafíos que se plantean en el problema de predicción de proteínas son el cálculo de energía libre del sistema y encontrar el mínimo global de esta energía.

Para solucionar este problema, existen 4 enfoques detallados a continuación.

3.1.1 Métodos *ab initio* sin información

Los métodos basados en *ab initio* que no utilizan información de bases de datos están fundamentados en la ley de la termodinámica, a través de la cual, se realizan simulaciones que calculan la energía interna de la proteína mediante la minimización de la función que la describe y la interacción de la proteína con el ambiente donde está inserta. El objetivo es identificar los valores de un conjunto de variables (ángulos de

torsión, posición de los átomos, etc.) que describa la conformación del polipéptido con la menor energía (Tramontano, 2006). Ejemplos de estos métodos son ASTROFOLD (Klepeis & Floudas, 2003), BAHGEERRATH (Arora & Jayaram, 1998), LINUS (Srinivasan & Rose, 2002), entre otros.

3.1.2 Métodos *ab initio* con información

Los métodos basados en *ab initio* que utilizan información de bases de datos son una variante del enfoque anterior ya que para realizar la predicción usan información existente para moldear la estructura y luego la refinan mediante simulación de partículas (Floudas, Fun, S., Moennigmann, & Rajgaria, 2006). Algunos métodos conocidos son I-TASSER (Zhang, 2008), ANGLOR (S. Wu, 2008) y FRAGFOLD (Jones, Predicting novel protein folds by using FRAGFOLD, 2001).

3.1.3 Métodos de enhebrado de secuencias

Estos métodos contrastan la secuencia con estructura tridimensional desconocida con una base de información de estructuras conocidas, mediante una función de *fitness* se escogen las estructuras más probables, posteriormente en base a estas se genera el modelo final del polipéptido (Richardson, 1981). Algunas soluciones con este enfoque son: GENTHREADER (Jones, 1999) y PROSPECT (Xu Y., 2000).

3.1.4 Métodos de análisis comparativo por homología

Estos métodos tienen el objetivo de alinear la secuencia de aminoácidos con secuencias cuyas estructuras tridimensionales sean conocidas, de manera de obtener una estructura basada totalmente en la información ya existente (Sánchez & Sali, 1997). Métodos conocidos basados en este enfoque son MODELLER (Eswar N., 2006) y PSIBLAST (Altschul S.F., 1997).

3.2 Alternativas de solución

Teniendo en cuenta la complejidad computacional del problema 3-D PSP, los métodos actuales hacen uso de una amplia gama de algoritmos de optimización y

metaheurísticas (Klepeis, Pieja, & Flouda, 2003) con el fin de proporcionar soluciones cercanas al óptimo en tiempos razonables de ejecución.

Además, al considerar las limitaciones de las cuatro clases de métodos de predicción de estructuras de proteínas, los investigadores han desarrollado recientes métodos híbridos que combinan los principios de estas cuatro clases. Por ejemplo, la exactitud que presentan los métodos de modelado por comparación por homología se combinan con la capacidad de los métodos *ab initio* de predecir (Dorn, Breda, & Souza, 2008).

Para reducir el orden y la alta dimensionalidad del espacio de búsqueda inherente a los métodos *ab initio*, se puede utilizar la información de los motivos estructurales (*motifs*) que se encuentran en las estructuras de proteínas para construir nuevas conformaciones. Estas conformaciones permiten realizar refinamientos a nivel de Mecánica Molecular (MM) y Dinámica Molecular (MD) (Gunsteren & Berendsen, 1990). En el paso de refinamiento, se evalúan las interacciones de todos los para corregir los ángulos de torsión de las estructuras secundarias. En palabras simples, se usa el conocimiento provisto de la PDB para generar una estructura inicial que será mejorada o refinada con técnicas *ab initio.*

3.3 Justificación de la solución propuesta

Dada la información disponible en la PDB y de los métodos descritos anteriormente, es que se ha decido diseñar un algoritmo memético que realice la predicción en base a la información de los ángulos de torsión presentes en la PBD, ya que uno de los grandes problemas de los métodos anteriores es el espacio de búsqueda y la calidad de la predicción.

Según investigaciones recientes, es posible reducir los espacios de búsqueda y mejorar la calidad de las predicciones usando algoritmos evolutivos que usen aquella información que es más probable (biológicamente hablando), descartando aquellas soluciones menos probables y usando el tiempo ganado en generar predicciones de mejor calidad (Dorn, Inostroza-Ponta, Buriol, & Verli, 2013).

4. DESCRIPCIÓN DE LA SOLUCIÓN PROPUESTA

4.1 Propósitos de la solución

El propósito de la solución es generar predicciones de mejor calidad biológica haciendo uso de información de ángulos de torsión conocidos mediante la reducción del espacio de búsqueda, que implicaría descartar todas aquellas soluciones biológicamente menos prometedoras.

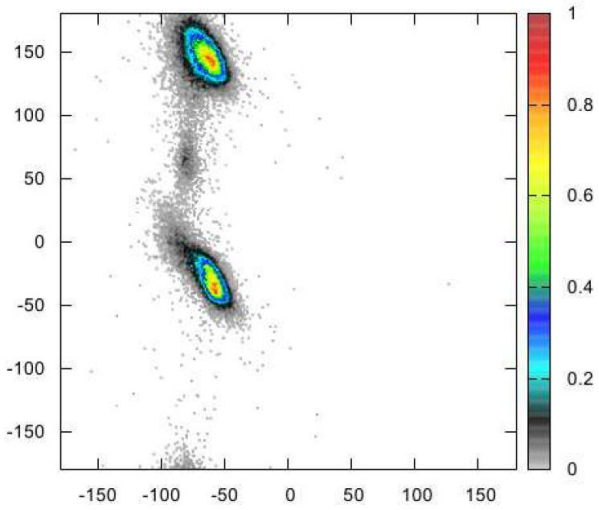

Figura 4-1: Mapa de Ramachandran para los ángulos ψ y φ del aminoácido Proline

En la figura anterior, se puede apreciar que hay 2 zonas en las que se concentran los ángulos usados por el aminoácido *Proline*, la idea es probar solo con esas áreas ya que biológicamente esos ángulos son los más usados por el aminoácido.

4.2 Características de la solución

La solución será un algoritmo memético que incorporará información de los ángulos de torsión más probables según contexto biomolecular, cuya función será tomar y generar modelos candidatos de estructuras que irán evolucionando según reglas meméticas (reproducción de la población y variación de esta mediante operadores de búsqueda local). El enfoque de la solución se sitúa en el grupo de las técnicas ab initio con información de bases de datos.

Los datos a emplear serán los provistos por la *Protein Data Bank* (F.C.Bernstein, 1977), que cuenta con la información de los ángulos de torsión φ y ψ de estructuras conocidas y comprobadas empíricamente.

La ejecución de la solución será realizada en la máquina llamada BioServer disponible como recurso computacional en el Departamento de Ingeniería Informática de la Universidad de Santiago de Chile reservado para estudios bioinformáticos.

4.3 Alcances y limitaciones de la solución

La principal limitación de la solución propuesta es que su implementación y evaluación serán llevadas a cabo solo para la simulación de estructuras terciarias. Además la solución considera comparaciones de calidad de las soluciones encontradas respecto a datos reales.

Si bien, el problema de predicción es costoso a nivel computacional, no se considera la implementación de la solución a nivel distribuido, por lo que la solución hará uso solo de los recursos provistos del dispositivo en el cual se ejecutará.

La visualización de la estructura lograda se realizará usando el software PyMOL.

4.4 Descripción de los resultados o evaluación de la solución

Respecto de los análisis de resultados y la evaluación de estos, se realizaran mediantes dos tipos de análisis:

1. **Análisis Estructural**: mediante el uso de la medida RMSD (*Root-Mean-Square-Deviation*), es decir, se contrastará el resultado obtenido de la estructura predecida tras la última generación entregada por el algoritmo memético con los datos empíricos de la PDB mediante el cálculo del error usando RMSD con la herramienta PyMOL.
2. **Análisis de estructuras secundarias**: usando la herramienta PROMOTIF, se contrastará las estructuras secundarias predecidas

por el algoritmo memético con los datos reales, con el fin de revisar la calidad de la predicción respecto de los patrones esperados.

3. **Análisis algorítmico**: se debe evaluar los tiempos y curvas de convergencia del algoritmo obtenido.

5. METODOLOGÍA, HERRAMIENTAS Y AMBIENTE DE DESARROLLO

5.1 Metodología a usar

La metodología que se utilizará consta de 3 fases, las cuales se indican a continuación:

- **Diseño**: Esta fase se divide en dos sub-fases.
  - **Concepción:** se establecen los estudios necesarios para abordar el problema revisando alternativas y estrategias de solución.
  - **Elaboración**: En esta etapa se definen modelos, esquemas y diagramas que se utilizarán para construir la solución en base a la información recopilada en la concepción.
- **Construcción**: durante la construcción se implementan todos los elementos definidos durante la elaboración.
- **Experimentos Computacionales y Análisis de Resultados**: esta etapa recoge resultados obtenidos a partir de la construcción para refinar y estabilizar el software. Se evalúa la hipótesis a acuerdo a los resultados obtenidos.

La metodología adoptada es inspirada en RUP con base en el método científico, ya que se requiere tener conciencia de todo lo necesario (necesidades a ser detectadas en la fase de concepción) para realizar la predicción de estructuras de polipéptidos mediante la construcción de la solución y su posterior análisis y validación en la etapa de experimentos computacionales.

La hipótesis del trabajo es: “Incorporar las probabilidades de ángulos de torsión φ y ψ en un algoritmo memético permite obtener predicciones de estructuras tridimensionales de proteínas de mejor calidad biológica”.

5.2 Herramientas de desarrollo

A continuación se presentan las herramientas que se utilizarán para el desarrollo del proyecto.

5.2.1 Herramientas de software

Para la escritura de la memoria se utilizarán las siguientes herramientas:

- Sistema operativo OS X Maverick 10.9
- Latex

Para el desarrollo, compilación y ejecución de las implementaciones se utilizará:

- Sistema operativo Linux Ubuntu 14.04
- PROMOTIF
- PyMOL
- C
- AmberTools 14: NAB

5.2.2 Herramientas de hardware

Las pruebas oficiales y el desarrollo serán realizados en el servidor llamado BioServer del Departamento de Ingeniería Informática.

Además serán utilizados los recursos computacionales que posee el estudiante, los cuales son:

- Computador portátil Macbook Pro Mid 2012 con:
  - 16GB de RAM DDR3.
  - Procesador Intel i5 2.5 Ghz.

- Tarjeta de video integrada Intel HD Graphics 4000.
- Unidad de almacenamiento SSD Seagate de 240GB de capacidad

5.3 Ambiente de desarrollo

El ambiente de desarrollo de la memoria será:

- Laboratorio de colaborativa, Departamento de Ingeniería Informática UdeSantiago.
- Domicilio particular del estudiante, ubicado en Renca.

6. PLAN DE TRABAJO

El plan de trabajo elaborado, considera el inicio del proyecto el día 3 de Marzo de 2014 y su finalización el día 20 de Junio de 2014, comprendiendo 16 semanas.

Se han estipulado 5 días de trabajo semanal, desde lunes a viernes, con un promedio de 8 horas de trabajo diario. Según la planificación serán 612 horas de dedicación para alcanzar la completitud del trabajo de titulación.

A continuación se muestra en detalle de la planificación en la figura 6.1.

Nombre de tarea	Duración	Comienzo	Fin
1° Etapa: Desarrollo	**80 días**	**lun 03-03-14**	**vie 20-06-14**
Diseño	**20 días**	**lun 03-03-14**	**vie 28-03-14**
Concepción	**15 días**	**lun 03-03-14**	**vie 21-03-14**
Estudio de técnicas de Predicción de Proteínas	2 días	lun 03-03-14	mar 04-03-14
Estudio de implementación de metaherística usadas en 3D PSP	4 días	mié 05-03-14	lun 10-03-14
Estudio y análisis de métodos de validación	4 días	mar 11-03-14	vie 14-03-14
Estudio de herramientas necesarias a usar con 3D PSP	4 días	lun 17-03-14	jue 20-03-14
Definición de estrategia(s) de predicción	1 día	vie 21-03-14	vie 21-03-14
Elaboración	**6 días**	**vie 21-03-14**	**vie 28-03-14**
Diseño de estrategia de predicción	2 días	vie 21-03-14	lun 24-03-14
Elaboración de modelo de clases	2 días	lun 24-03-14	mar 25-03-14
Elaboración de modelo de comunicación	2 días	mié 26-03-14	jue 27-03-14
Diseño de experimento	2 días	jue 27-03-14	vie 28-03-14
Documentación	**19 días**	**mar 04-03-14**	**vie 28-03-14**
Creación de documento	19 días	mar 04-03-14	vie 28-03-14
Construcción	**30 días**	**lun 31-03-14**	**vie 09-05-14**
Implementación de algoritmos y clases	15 días	lun 31-03-14	vie 18-04-14
Implementación de comunicación entre los componentes	15 días	lun 21-04-14	vie 09-05-14
Creación de documento	25 días	lun 07-04-14	vie 09-05-14
Experimentos computacionales	**30 días**	**lun 12-05-14**	**vie 20-06-14**
Realización de pruebas	10 días	lun 12-05-14	vie 23-05-14
Análisis y evaluación de resultados	5 días	lun 26-05-14	vie 30-05-14
Refinamiento y estabilización	5 días	lun 02-06-14	vie 06-06-14
Realización de pruebas finales	5 días	lun 09-06-14	vie 13-06-14
Análisis y evaluación de pruebas finales	5 días	lun 16-06-14	vie 20-06-14
Redacción Documento	20 días	lun 26-05-14	vie 20-06-14
2° Etapa: Confección Documento	**79 días**	**mar 04-03-14**	**vie 20-06-14**
Redacción de documento	**79 días**	**mar 04-03-14**	**vie 20-06-14**
Reunir y completar información documentada de fases anteriores	79 días	mar 04-03-14	vie 20-06-14

Figura 6-1: Planificación asociada al desarrollo del trabajo de titulación

7. REFERENCIAS

Altschul S.F., M. T. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 3389-402.

Arora, N., & Jayaram, B. (1998). Energetics of base pairs in B-DNA in solution: An appraisal of potential functions and dielectric treatments. *J. Phys. Chem. B*, 6139-6144 .

Berger, B., & Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete. *Proceedings of the second annual international conference on Computational molecular biology* (págs. 30-39). New York: ACM.

Cohen, J. (2004). Bioinformatics - An introduction for computer scientist. *ACM Computing Surveys, Vol.36*, 122-158.

DM, M. (2004). *Bioinformatics: Sequence and Genome Analysis, Vol 2.* Cold Spring Harbor Laboratory Press.

Dor, O., & Zhou, Y. (2007). Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins, 66*, 838–845.

Dorn, M., Breda, A., & Souza, O. N. (2008). A hybrid method for the protein structure prediction problem. *Lect. Notes Bioinf., vol. 5167*, 47–56.

Dorn, M., Inostroza-Ponta, M., Buriol, L., & Verli, H. (2013). A knowledge-based genetic algorithm to predict three-dimensional structures of polypeptides. *Congress on Evolutionary Computation* (págs. 1233-1240). Cancun: IEEE.

Eswar N., W. B.-R. (2006). Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics*, Chapter 5.6.

Floudas, C., Fun, H., S. M., Moennigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and de novo protein design: A review. En *Chemical Engineering Science* (págs. 966–988). Princeton: Elsevier.

Froimowitz, M., & Fasman, G. D. (1974). Prediction of the secondary structure of proteins using the helix-coil transition. *Macromolecules*, 583-589.

Gunsteren, W. v., & Berendsen, H. (1990). Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. *Angew. Chem., Int. Ed. Engl., vol. 29*, 992–1023.

Jones, D. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of Molecular Biology*, 797-815.

Jones, D. (2001). Predicting novel protein folds by using FRAGFOLD. *Proteins*, 127-132.

Klepeis, J. L., & Floudas, C. A. (2003). ASTRO-FOLD: A Combinatorial and Global Optimization Framework for Ab Initio Prediction of Three-Dimensional Structures of Proteins from the Amino Acid Sequence. *Biophys J.*, 2119–2146.

Klepeis, J., Pieja, M., & Flouda, C. (2003). Hybrid global optimization algorithms for protein structure prediction: alternating hybrids. *Biophys, Vol.84*, 869-882.

Lehninger, A., Nelson, D., & Cox, M. (2005). *Principles of Biochemistry, 4th ed.* New York: W.H. Freeman.

Lesk, A. M. (2002). *Introduction to Bioinformatics, 1st ed.* New York: Oxford University Press Inc.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., & Scott, M. (1990). *Molecular Cell Biology, 5th ed.* New York: Scientific American Books.

Pauling, L., Corey, R., & Branson, H. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37-205.

RefSeq: NCBI Reference Sequence Database. (1 de Agosto de 2012). Obtenido de NCBI: http://www.ncbi.nlm.nih.gov/refseq/

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 167.

S. Wu, Y. Z. (2008). ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. *PLOS ONE*, 3400.

Sánchez, R., & Sali, A. (1997). Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, 206-214.

Srinivasan, R., & Rose, G. (2002). Ab initio prediction of protein structure using LINUS. *Proteins*, 489-95.

Tramontano, A. (2006). *Protein structure prediction, 1st ed.* Weinheim,: John Wiley and Sons, Inc.

Xu Y., X. D. (2000). Protein threading using PROSPECT: design and evaluation. *Proteins*, 343-54.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 1471-2105.