

Importancia de las Propiedades Físico-Químicas de los Aminoácidos en la Predicción de Estructuras de Proteínas usando Vecinos más Cercanos

Gualberto Asencio Cortés¹ Jesús S. Aguilar-Ruiz¹

¹ Universidad Pablo de Olavide, {guaasecor,aguilar}@upo.es

Resumen

La predicción de estructuras de proteínas es actualmente un importante campo de investigación dentro de la bioinformática. En esta área, existen numerosos estudios realizados en los que se ha usado la información de la separación entre los aminoácidos de una cadena para predecir la estructura de las proteínas, utilizándose en otros trabajos ciertas propiedades físico-químicas de aminoácidos. En este trabajo se han usado ambas informaciones y se ha estudiado cómo influyen en la predicción de estructuras de proteínas empleando el algoritmo de vecinos más cercanos. Hemos comprobado que la información proporcionada por las propiedades físico-químicas es de mayor interés que la separación, obteniéndose mejores tasas de acierto. Se han realizado cuatro experimentos en los que se ha usado como atributos, la separación entre aminoácidos y un conjunto determinado de propiedades físico-químicas de los mismos y, como ejemplos, todas las subsecuencias posibles encontradas en un conjunto de más de 5000 proteínas reales. Finalmente se demuestra empíricamente que la separación entre aminoácidos, ampliamente usada en la literatura, puede ser reemplazada por propiedades físico-químicas de aminoácidos, produciendo mejores predicciones. La tasa de acierto conseguida usando sólo la separación está en torno al 59%, ascendiendo este valor hasta el 79% al usarse un conjunto de propiedades físico-químicas de aminoácidos.

Palabras Clave: propiedades físico-químicas, separación, vecinos más cercanos, predicción de estructura secundaria de proteínas.

1 INTRODUCCIÓN

Este trabajo se encuentra emplazado dentro del área de la bioinformática y, dentro de la misma, en la predicción de estructuras secundarias de proteínas.

La predicción de la estructura secundaria de las proteínas consiste en averiguar un patrón de comportamiento en la formación tridimensional de las proteínas, únicamente a partir de la cadena de aminoácidos que la forman.

Hoy en día se desconoce el carácter determinista de dicho proceso de formación, estudiándose tan solo los estados inicial (cadena de aminoácidos) y final (proteína totalmente formada).

Para llevar a cabo la tarea de predicción se pueden utilizar métodos de minería de datos, los cuales trabajan, en este contexto, con bases de datos procedentes de proteínas conocidas, y mediante algoritmos de clasificación y regresión extraen el conocimiento suficiente para generar un modelo que permita averiguar de forma determinista la estructura que tendrá una proteína aún sin formarse, con un determinado grado de probabilidad de acierto. Entre estos métodos se encuentra la técnica de vecinos más cercanos, empleada en la predicción de estructuras de proteínas [4,13].

Para realizar las predicciones se pueden usar los llamados mapas de contacto, que ilustran dónde se producen las conexiones entre los aminoácidos dentro de la cadena proteica, y los llamados mapas de distancia, entre otros. En estos últimos, se representan los valores reales y predichos de distancia entre dos aminoácidos cualesquiera de la cadena. En este trabajo se predicen mapas de distancia, con lo que se hablará de regresión en lugar de clasificación, al ser la distancia un valor continuo. De esta forma, los resultados de las predicciones realizadas pueden ser expresados a posteriori en términos de contactos usando diferentes umbrales.

En numerosos estudios realizados se ha usado para la predicción la información de la separación entre

aminoácidos en las cadenas proteicas [2,3,11]. Esta separación mide el número de aminoácidos intermedios entre dos determinados en dichas cadenas. Parece claro que esta información puede ser útil para la predicción de la estructura de una proteína puesto que revela la posición relativa de los aminoácidos dentro de la cadena, y son éstos los que establecen contactos unos con otros. Además, se ha comprobado que dicha información es útil debido a que existen numerosos contactos entre aminoácidos bien distanciados en la cadena (long-range interactions) [7,12], siendo éstos ignorados en métodos de predicción locales.

Por otra parte, analicemos la naturaleza del principal fenómeno que acontece en la formación de la estructura de una proteína: la formación de un contacto entre aminoácidos. Éste es un proceso biológico donde intervienen una serie de elementos materiales, tales como los propios aminoácidos, con una serie de propiedades físico-químicas inherentes. Para determinar el lugar donde se producen estos contactos, creemos que es necesario observar los valores de dichas propiedades físico-químicas y usarlas en la predicción de la estructura de las proteínas. Existen trabajos en los que se han usado este tipo de propiedades físico-químicas [8,14] aunque son empleadas, entre otros fines, para determinar dominios de proteínas e interfaces proteína-proteína.

En consecuencia, nos planteamos predecir con ambas medidas: la separación y las propiedades físico-químicas; para luego omitir cada una y comprobar el comportamiento ofrecido tanto en conjunto como aisladamente. El objetivo es determinar cómo influyen las propiedades físico-químicas de los aminoácidos en la predicción de estructuras de proteínas, en las cuales confiamos que existe información útil.

2 MÉTODO

Se han realizado tres actividades encaminadas a probar un conjunto de datos de proteínas reales, una serie de propiedades físico-químicas conocidas y la separación entre aminoácidos, mediante un algoritmo de vecinos más cercanos, para finalmente analizar los resultados obtenidos y extraer conclusiones en relación al objetivo planteado.

Se ha escogido la herramienta Weka [5], ampliamente utilizada en procesos de minería de datos, para predecir la estructura de las proteínas. En concreto, para predecir la distancia entre cualquier par de aminoácidos de sus cadenas polipeptídicas.

Esta herramienta trabaja con archivos de datos donde deben incluirse una serie de ejemplos con un conjunto de atributos determinado y una clase o etiqueta, para la cual

se realizan las predicciones. Los archivos de datos suministrados a Weka en los experimentos realizados en este trabajo contienen los valores de separación entre aminoácidos y de determinadas propiedades físico-químicas para todas las posibles subsecuencias encontradas en un conjunto de proteínas procedentes del Protein Data Bank (PDB) [1].

En concreto, para los experimentos realizados se han tomado todas las proteínas existentes del PDB de una sola cadena polipeptídica y de longitud máxima 100; lo cual ha dado de resultado un total de 5380 proteínas (en la fecha en la que se tomaron los datos) que en adelante denominaremos set5380p. En la figura 1 se puede apreciar el procedimiento global empleado.

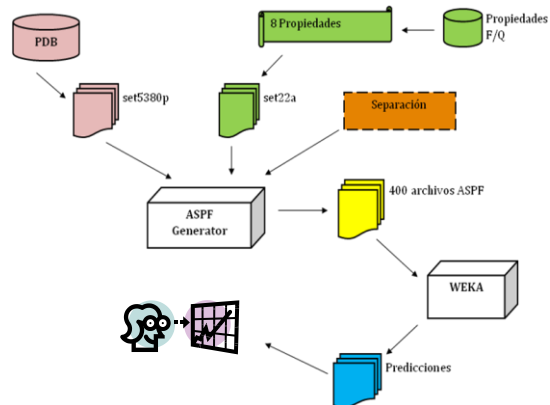


Figura 1: Proceso global empleado para obtener los resultados

2.1. GENERACIÓN DE FICHEROS DE PROPIEDADES DE SUBSECUENCIAS DE AMINOÁCIDOS (ASPF)

Para cada proteína analizada, se han considerado todos los pares posibles de aminoácidos naturales ($20 \times 20 = 400$ pares posibles: A-A, A-C, ...) y se han obtenido todas las subsecuencias incluidas en la proteína, clasificándolas según los aminoácidos de inicio y fin de cada una. Se han ignorado las subsecuencias con separación 0 entre dichos aminoácidos. Por ejemplo: si la cadena de aminoácidos es CACAC, el número de subsecuencias obtenidas para cada par de aminoácidos inicio-fin serían: A-A:1, A-C:1, C-A:1, C-C:2.

Todas las subsecuencias de todas las proteínas analizadas han sido organizadas en $20 \times 20 = 400$ ficheros de datos. Cada fichero contiene un resumen de los valores de las propiedades físico-químicas de los aminoácidos interiores, omitiendo los extremos, en todas las subsecuencias con unos aminoácidos inicio y fin determinados.

Estos ficheros reciben el nombre de Ficheros de Propiedades de Subsecuencias de Aminoácidos (ASPF). Así, por ejemplo, el fichero “A-C.aspf” contendrá un resumen de las propiedades físico-químicas de los aminoácidos que constituyen cada una de las subsecuencias con inicio en el aminoácido A y fin en el aminoácido C encontradas en todo el set de proteínas estudiado.

Se ha usado un conjunto de 8 propiedades físico-químicas descritas y evaluadas en [9], de las cuales, se han considerado siete propiedades nominales y una real. En la figura 2 se muestra un diagrama de Venn en el que se recogen varias de las propiedades usadas (hydrophobic, aliphatic, aromatic, polar, charge, small). Además, se han usado las propiedades cbetabranched y functional (numérica real), no descritas en la figura 2.

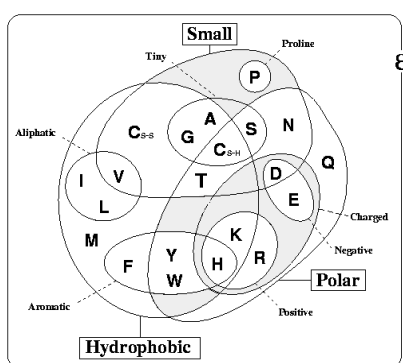


Figura 2: Diagrama de Venn de algunas propiedades físico-químicas de aminoácidos

Clasificamos los 20 aminoácidos naturales en función del grado que ofrecen para cada una de estas propiedades. Por ejemplo, para la propiedad charge, existen cuatro grados: usualmente positivo, usualmente negativo, a veces positivo y sin carga. Pueden consultarse los grados de todas las propiedades usadas en [9].

Dado que, en general, no existe ninguna relación de orden entre estos grados, se han creado 21 atributos, uno por cada grado de las 7 propiedades nominales. Cada uno de estos 21 atributos es un número real y representa el tanto por 1 de ocurrencia (en las subsecuencias) de aminoácidos que presentan cada grado de propiedad físico-química. En definitiva, se han usado 21+1 (la propiedad funcional, que es numérica real)=22 atributos físico-químicos, que denominaremos set22a, como se resume en la figura 1.

Existen numerosas propiedades físico-químicas de aminoácidos publicadas que no se han usado [6,10,15], y que nos proponemos utilizar en futuros trabajos.

Se ha implementado una aplicación denominada ASPFGenerator que permite, a partir de archivos .ENT de datos de proteínas descargados desde el PDB, generar

automáticamente los 400 ficheros ASPF descritos. Esta herramienta ha sido usada para procesar 5 GB de archivos de texto .ENT correspondientes al conjunto set5380p de estudio, generando 1,80 GB en 400 ficheros ASPF, empleando 180 hilos de ejecución y 5 minutos en el proceso.

2.2. PREDICCIÓN CON WEKA (IBK)

Una vez obtenidos los archivos de propiedades ASPF, éstos serán la entrada de la herramienta Weka para predecir las distancias entre los aminoácidos. En este trabajo se ha usado Weka 3.6.0 Experimenter y el clasificador IBk con K=1 (1-KNN).

Dado que los datos de entrada para la predicción con Weka son los archivos ASPF antes generados, se predecirá la distancia entre dos aminoácidos cualesquiera a partir del conocimiento obtenido en las propiedades físico-químicas y los valores de separación de todas las subsecuencias con origen y fin en dichos aminoácidos. De este modo, el conocimiento está basado en similitudes de naturaleza físico-química entre los aminoácidos intermedios en las cadenas polipeptídicas.

Para comprobar las predicciones resultantes y descubrir conocimiento acerca de la influencia de la separación entre aminoácidos frente a sus propiedades físico-químicas, se han realizado con Weka cuatro experimentos:

Experimento SOLO_SEP: En este experimento se ha usado únicamente el atributo separación en la obtención de vecinos más cercanos. Este experimento ha servido para sentar las bases a partir de las cuales intentar mejorar las predicciones.

Experimento SOLO_PFQ: Aquí se ha usado únicamente el set22a, antes descrito, de propiedades físico-químicas sin atributo separación en la obtención de vecinos más cercanos. Este experimento ha sido realizado para demostrar la validez de estas propiedades, aisladamente, en la predicción de estructuras de proteínas.

Experimento IGUAL_SEP: Aquí se ha utilizado como atributos el set22a, junto al atributo separación, teniendo en cuenta que se ha reducido el campo de exploración de vecinos más cercanos (subsecuencias más cercanas) a sólo aquellos con igual valor de separación que el ejemplo de test a clasificar. Se ha realizado este experimento para determinar cómo se comportan las predicciones cuando asumimos que las distancias entre aminoácidos son similares si éstos tienen la misma separación y similares propiedades.

Experimento CUALQ_SEP: En este experimento se han usado los mismos atributos que en el experimento IGUAL_SEP pero explorando entre todas las

subsecuencias disponibles. Se pretende probar la predicción de estructuras de proteínas usando la separación y las propiedades físico-químicas sin ninguna restricción.

Para la predicción en los cuatro experimentos se han usado todas las subsecuencias encontradas cuyo aminoácido de inicio es A, C, D, E, F, G, H, I, K o L.

Los atributos que se han usado en dicha distancia están todos normalizados entre 0 y 1. La distancia usada ha sido la distancia euclídea sin ponderación alguna de atributos:

$$dij = \sqrt{\underbrace{(S_i - S_j)^2}_{dSEP} + \sum_f \underbrace{(P_{fi} - P_{fj})^2}_{dPFQ}} \quad (1)$$

De ésta se deducen fácilmente los siguientes casos particulares: en el experimento IGUAL_SEP la componente dSEP = 0, en el experimento SOLO_PFQ la componente dSEP no existe, y en el experimento SOLO_SEP la componente dPFQ no existe.

Para obtener resultados adecuados al objetivo perseguido en el estudio, se han introducido algunas modificaciones al código fuente original de la herramienta Weka, de tal forma que permita obtener el error relativo cometido en las predicciones. No obstante, el comportamiento del clasificador IBk de Weka no ha sido alterado, usando para todos los experimentos K=1 (1-KNN).

2.3. ANÁLISIS DE RESULTADOS

Finalmente, una vez obtenidos los resultados de la predicción con Weka, se han analizado con Excel para descubrir, a partir de los resultados, el comportamiento de las predicciones; ilustrando mediante tablas y gráficas, las cuales se verán en el siguiente apartado, la información pertinente para justificar las conclusiones.

3 EXPERIMENTACIÓN

A continuación, se mostrarán las tablas y gráficas generadas a partir de los resultados obtenidos en la experimentación y se extraerán las conclusiones de las mismas en el marco de los objetivos planteados en este trabajo.

3.1 TASA DE ERRORES EN LOS EXPERIMENTOS

En primer lugar mostramos en la figura 3, mediante un gráfico de Boxplot, el error relativo medio cometido en cada experimento.

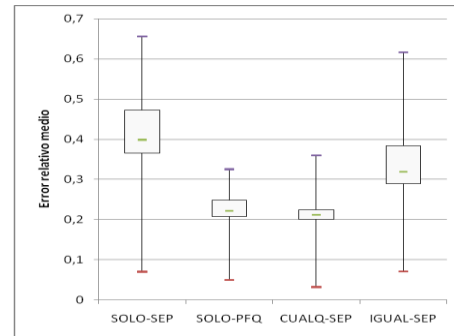


Figura 3: Gráfico Boxplot del error relativo cometido en los experimentos

Según se aprecia en la figura 3, se consigue menor tasa de errores (aprox. 0,216) cuando interviene, junto a la separación, la información de las propiedades físico-químicas de los aminoácidos en la búsqueda de vecinos más cercanos (CUALQ_SEP).

No obstante, es de interés destacar que la tasa de errores (0,227) y la desviación típica (0,272) obtenidas usando únicamente las propiedades físico-químicas (SOLO_PFQ) es similar a la conseguida en el experimento CUALQ_SEP; incluso se consigue una desviación estándar menor. Esto indica el buen comportamiento de dichas propiedades, por sí solas, a la hora de determinar las distancias entre aminoácidos.

Nótese, por otra parte, cómo el experimento IGUAL_SEP no ha arrojado tan buenos resultados como CUALQ_SEP, usando ambos la separación y las propiedades. Esto se debe a que las distancias entre aminoácidos no son siempre similares teniendo éstos la misma separación.

3.2 ANÁLISIS PORMENORIZADO DE ERRORES PARA CADA VALOR DE SEPARACIÓN

A continuación analizamos en detalle el error relativo cometido en las predicciones en función de la separación existente en las subsecuencias de proteínas estudiadas. Para cada experimento, estudiamos mediante una gráfica estos valores de error. En todas las gráficas que siguen se representa en el eje de abscisas la separación entre los aminoácidos para los cuales se ha predicho su distancia, y en el eje de ordenadas los valores de error relativo promedio cometido para cada par de aminoácidos estudiado.

En primer lugar, se presenta en la figura 4 la gráfica de errores para el experimento SOLO_SEP, donde se ha usado únicamente el atributo separación en la obtención de vecinos más cercanos.

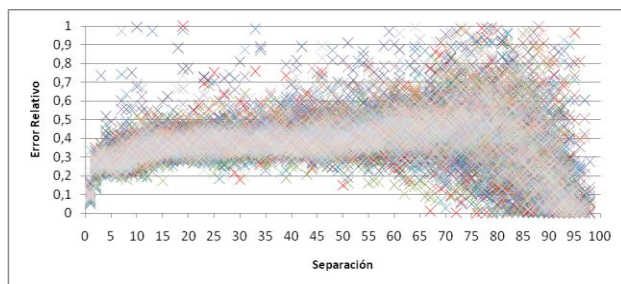


Figura 4: Gráfica de errores para el experimento SOLO_SEP

Como se puede apreciar en la figura 4, el error relativo medio aumenta conforme lo hace la separación entre aminoácidos, situándose en valores medios en torno a 0,35 y 0,4. Existe en este experimento una gran dispersión de valores de error para valores altos de separación, lo cual no es deseable. No obstante, tal como ocurrirá en el resto de experimentos, existe una tendencia clara al descenso del error relativo para valores altos de separación. En este caso, a partir del valor 80 de separación aproximadamente, los errores se hacen cada vez más cercanos a 0; realizándose perfectas predicciones en la distancia entre aminoácidos con alta separación.

Además, podemos apreciar tanto en la figura 4 como en las correspondientes al resto de experimentos, que para valores de separación pequeños (1,2) el error relativo cometido es siempre menor (entre los valores 0,05 y 0,3). Esto se debe a que la distancia entre aminoácidos próximos en la cadena proteica (separación pequeña) suele ser siempre la misma, debido a la cercanía de éstos en la cadena proteica y no debido a la formación de un contacto. De este modo, se encuentran vecinos con misma separación y distancia muy parecida y el error cometido desciende.

Mostramos a continuación en la figura 5 la gráfica de errores para el experimento SOLO_PFQ, donde se han usado únicamente las propiedades físico-químicas sin atributo separación en la obtención de vecinos más cercanos.

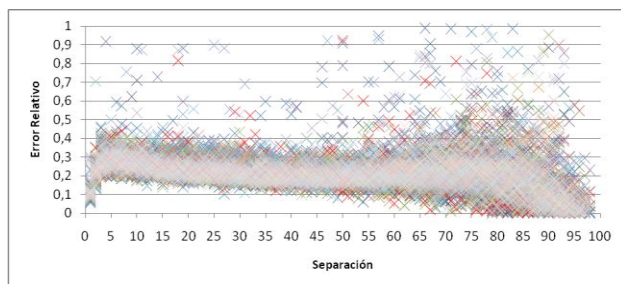


Figura 5: Gráfica de errores para el experimento SOLO_PFQ

La mejora conseguida es apreciable en varios sentidos. En primer lugar, el error relativo adopta mejores valores, situándose en valores medios en torno a 0,2 y 0,25. Por otra parte, el error relativo desciende conforme aumenta la separación, al contrario que en el experimento anterior, lo cual es altamente deseable. Además, se conserva la clara tendencia al descenso del error relativo para valores altos de separación, con la gran ventaja añadida de una considerable menor dispersión de valores de error.

A continuación se muestra en la figura 6 la gráfica de errores para el experimento CUALQ_SEP, donde se han utilizado tanto la separación como las propiedades de aminoácidos, usando en la exploración todos los ejemplos (subsecuencias de proteínas) disponibles en la base de datos. Habiendo observado los resultados conseguidos únicamente con las propiedades físico-químicas, comprobamos ahora cómo interviene la separación cuando ésta se incorpora en las predicciones.

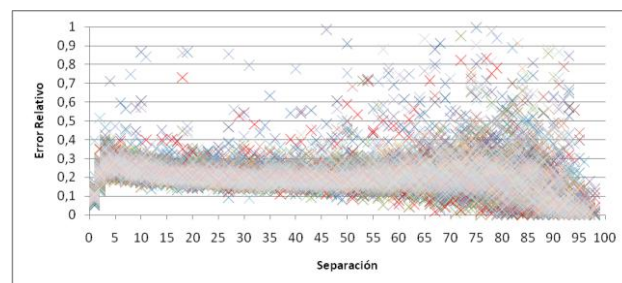


Figura 6: Gráfica de errores para el experimento CUALQ_SEP

Comprobamos que el comportamiento del error relativo es prácticamente el mismo que en el experimento SOLO_PFQ (véanse las figuras 5 y 6). Recordemos que el error relativo medio obtenido en la figura 3 era muy parecido en ambos experimentos (SOLO_PFQ y CUALQ_SEP). Además, se obtuvo un valor menor de desviación típica en el experimento SOLO_PFQ. Con ello, parece deseable usar de cualquier modo las propiedades físico-químicas, ya sea en presencia o en ausencia de la información de la separación entre aminoácidos.

En la figura 7 se presenta la gráfica de errores para el experimento IGUAL_SEP, donde se ha utilizado tanto la separación como las propiedades de aminoácidos, teniendo en cuenta que se ha reducido el campo de exploración de vecinos más cercanos a sólo aquellos con igual valor de separación que el ejemplo de test.

En este caso, como se comentó anteriormente, la componente dSEP de la distancia es siempre 0, puesto que el espacio de exploración de vecinos está formado únicamente por ejemplos con igual valor de separación que la instancia de test. Con ello, se reduce dicho espacio de exploración con respecto al total disponible y se

pueden dejar de considerar subsecuencias que pueden ser mejores, en el sentido de proporcionar mejor predicción.

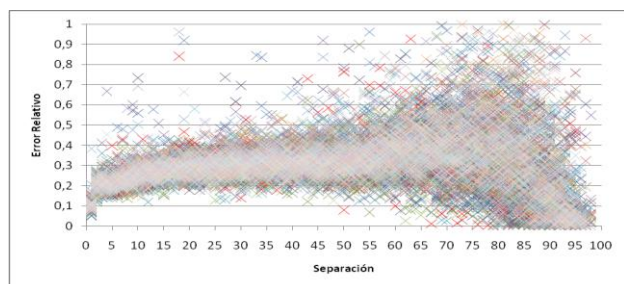


Figura 7: Gráfica de errores para el experimento IGUAL_SEP

Nótese que la gráfica de la figura 7 es similar a la de la figura 4 (experimento SOLO_SEP), salvo que, mientras en esta última la banda de errores se sitúa, en términos medios, por encima del valor 0,3, en este experimento IGUAL_SEP (donde se introducen las propiedades físico-químicas) se sitúa por encima del valor 0,2. Esto está en consonancia con el menor error relativo medio del experimento IGUAL_SEP, como revela la figura 3.

4 CONCLUSIONES

Las propiedades físico-químicas de los aminoácidos repercuten positivamente en la tasa de acierto de la predicción de estructuras proteínas mediante un clásico algoritmo de vecinos más cercanos, ya sea en presencia o en ausencia de la información de separación entre aminoácidos.

Según los resultados obtenidos en este trabajo, parece conveniente explorar más propiedades físico-químicas y añadirlas al conjunto de estudio, para determinar cuál es el conjunto o conjuntos de propiedades que mejor ayudan a la predicción de estructuras de proteínas.

Los resultados nos conducen a pensar que algún subconjunto de todas las posibles propiedades físico-químicas de aminoácidos conocidas, con una asignación oportuna de ponderaciones, permitiría realizar predicciones de mayor calidad y minimizar el error.

Referencias

- [1] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242, 2000.
- [2] Fariselli, P., Casadio, R. A neural network based predictor of residue contacts in proteins. *Protein Engineering* vol.12 no.1 pp.15-21, 1999.
- [3] Fariselli, P., Olmea, O., Valencia, A., Casadio R. Prediction of contact maps with neural networks and correlated mutations. *Protein engineering* vol.14 no.11 pp. 835-843, 2001.
- [4] Frishman, D., Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering* 9, 133-142, 1996.
- [5] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
- [6] Kawashima, S. and Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids Res.* 28, 374, 2000. <http://www.genome.jp/aaindex/>
- [7] Kihara, D., The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.*, 14:1955-1963, 2005.
- [8] Ladunga, I. PHYSEAN PHYsical SEquence ANALYSIS for the identification of protein domains on the basis of physical and chemical properties of amino acids. *Bioinformatics* vol.15 no.12 pp. 1028-1038, 1999.
- [9] M.J. Betts, R.B. Russell. Amino acid properties and consequences of substitutions. In *Bioinformatics for Geneticists*, M.R. Barnes, I.C. Gray eds, Wiley, 2003. <http://www.russell.embl-heidelberg.de/aas/>
- [10] Nakai, K., Kidera, A., Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* 2, 93-100, 1988.
- [11] Olmea, O., Valencia, A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design* 2, s25-s32, 1997.
- [12] Pan, X.M., Niu, W.D., Wang, Z.X. What is the minimum number of residues to determine the secondary structural state? *J. Protein Chem.* 18:579-584, 1999.
- [13] Salamov, A.A., Solovyev, V.V. Protein Secondary Structure Prediction Using Local Alignments. *J. Mol. Biol.*, 268, 31-36, 1997.
- [14] Surendra, S. Negi, Werner Braun. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J Mol Model.* November 2007; 13(11): 1157-1167. doi:10.1007/s00894-007-0237-0.
- [15] Tomii, K., Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* 9, 27-36, 1996.