Predicción de Mapas de Distancia de Proteínas basados en Vecinos más Cercanos

Conference Paper · November 2011

CITATIONS
0

READS
206

3 authors:

Gualberto Asencio Cortés
University of Pablo de Olavide
55 PUBLICATIONS **605** CITATIONS
SEE PROFILE

Jesús S. Aguilar-Ruiz
Universidad Pablo de Olavide
231 PUBLICATIONS **3,606** CITATIONS
SEE PROFILE

Alfonso Marquez-Chamorro
Universidad de Sevilla
33 PUBLICATIONS **304** CITATIONS
SEE PROFILE

Predicción de mapas de distancias de proteínas basada en vecinos más cercanos

Gualberto Asencio Cortés, Jesús S. Aguilar-Ruiz and Alfonso E. Márquez Chamorro

Grupo de Bioinformática, Universidad Pablo de Olavide, Sevilla, España
{amarcha, fdivina, aguilar, guaasecor}@upo.es

Resumen La predicción de la estructura terciaria de las proteínas consiste en determinar la conformación tridimensional de las mismas únicamente a partir de su secuencia de aminoácidos. En este trabajo se propone un método basado en ensamblaje de fragmentos de proteínas por similitud físico-química, que puede aprovechar la información extraída de estructuras conocidas de proteínas. Muchos métodos de predicción de estructura terciaria de proteínas producen un mapa de contactos. El método que se propone produce un mapa de distancias, el cual provee más información sobre la estructura de una proteína que un mapa de contactos. Muchas aproximaciones en la literatura han utilizado propiedades físico-químicas de aminoácidos para la predicción de la estructura, generalmente la hidrofobicidad, la polaridad y la carga. En nuestro método se han utilizado tres propiedades físico-químicas de aminoácidos distintas, obtenidas de diferentes trabajos de la literatura. Se han realizado predicciones sobre todas las proteínas de cápsides de virus publicadas hasta Mayo de 2011 en Protein Data Bank con máxima identidad del 30 % (67 proteínas). Se ha obtenido una precisión del 76 % y una sensibilidad del 78 % con 8 angstroms de cut-off y mínima separación en la secuencia de 7 aminoácidos. Los resultados conseguidos mejoran notablemente, con las proteínas estudiadas, los resultados de otras propuestas.

Keywords: predicción de estructura de proteínas, mapa de distancias, propiedades físico-químicas de aminoácidos, vecinos más cercanos

1. Introducción

La predicción de estructuras de proteínas es un problema actual de gran importancia en bioinformática estructural. El interés en predecir la estructura tridimensional de las proteínas se debe a que dicha estructura determina todas sus funciones y, por lo tanto, tiene una importante repercusión en medicina y biología, como ocurre en el diseño de fármacos.

Existen procedimientos experimentales para la obtención de estructuras de proteínas, tales como la cristalografía de rayos X y la resonancia magnético-nuclear (NMR). No obstante, estos procedimientos son muy costosos y de ahí el gran interés puesto en emplear computadores y algoritmos de predicción.

Desde los experimentos de Anfinsen [1], se sostiene la creencia de que toda la información de la estructura de una proteína se encuentra únicamente en su secuencia de aminoácidos. Debido a estos experimentos, los métodos de predicción de estructura

terciaria de proteínas persiguen obtener un modelo tridimensional basado únicamente en la secuencia de aminoácidos de las mismas.

Existen actualmente dos principales aproximaciones a la predicción de estructuras de proteínas. Por una parte, existen los métodos ab initio, que intentan resolver la estructura de una proteína optimizando una función de energía, basándose generalmente en principios físico-químicos y sin usar ninguna proteína como plantilla. Los métodos ab initio sólo son adecuados para proteínas de tamaño relativamente pequeño [14]. Por contra, los métodos de modelado por homologías intentan resolver las estructuras basándose en proteínas plantilla (template-modeling). Estos métodos se consideran actualmente las aproximaciones más fiables al problema de la predicción de estructuras de proteínas [12].

El modelado basado en plantillas obtiene buenos resultados cuando existen proteínas con secuencia similar a la proteína objetivo. En ausencia de éstas, se usa un modelado libre. Dentro de los métodos de modelado libres, se crearon los métodos de ensamblaje por fragmentos, que reconstruyen la estructura de una proteína a partir de fragmentos de estructuras de otras, tales como FragFold [8], Fragment-HMM [9] o ROSETTA [11]. ROSETTA utiliza una aproximación de dos etapas en la que se comienza con un modelo de baja resolución y se prosigue con una representación de todos sus átomos, para finalmente minimizar la correspondiente función de energía. Por otra parte, las propiedades físico-químicas de los aminoácidos han sido utilizadas en numerosos trabajos de predicción de estructuras de proteínas. Entre las propiedades más usadas en la literatura se encuentran la hidrofobicidad, la polaridad y la carga, utilizadas por ejemplo en los modelos HP y HPNX [7].

Existen numerosos algoritmos de predicción de estructuras de proteínas que producen un mapa de contacto para representar la estructura predicha. Nuestro método produce un mapa de distancias, que incorpora más información que el mapa de contactos, ya que contiene las distancias entre todos los aminoácidos de la molécula, no sólo si hacen contacto o no. A diferencia de los modelos 3D, tanto los mapas de contactos como los de distancias tienen una propiedad deseable y es que son insensibles a rotaciones o traslaciones de la molécula.

El método que se propone es un método de modelado libre basado en ensamblaje de fragmentos que selecciona las mejores distancias entre pares de aminoácidos a partir de fragmentos de estructuras conocidas de proteínas. Los fragmentos son escogidos mediante un proceso de búsqueda de vecinos más cercanos por similitud en su longitud y en tres propiedades físico-químicas de aminoácidos obtenidas de la literatura. Se han realizado predicciones sobre todas las proteínas de cápsides de virus publicadas en Protein Data Bank [3] con máxima identidad en sus secuencias del 30 %. Se han realizado estudios de predicciones con mínima separación en la secuencia de 7 aminoácidos. Se han indicado los resultados obtenidos con otras propuestas, que sirven de punto de referencia para valorar la calidad de las predicciones obtenidas con nuestro método.

En la sección Método se definen los elementos utilizados, se describen los procedimientos empleados por nuestro sistema de predicción y se definen las medidas de evaluación que se han utilizado. En la sección Experimentación se detallan los datos de proteínas utilizados, las condiciones de la experimentación y los resultados obtenidos.

Finalmente, en la sección Conclusiones y trabajos futuros se exponen las conclusiones principales del estudio realizado y las futuras líneas de trabajo que se pretenden abordar.

2. Método

2.1. Definición de mapa de distancias

La matriz o mapa de distancias de una secuencia de proteína es una matriz cuadrada de orden N, donde N es el número de aminoácidos que tiene dicha secuencia. El elemento (i, j), con $i < j$, de la matriz de distancias es la distancia medida en angstroms (Å) observada entre el aminoácido i-ésimo y el j-ésimo dentro de la secuencia. Para medir las distancias entre aminoácidos se utiliza un átomo de referencia. Los átomos de referencia más utilizados son el carbono alfa (CA) y el carbono beta (CB) de un aminoácido [6]. En nuestro método, se han utilizado los carbonos beta (salvo para la glicina, que se usa el carbono alfa). En la triangular inferior del mapa de distancias se almacenan las distancias predichas por el algoritmo. De esta forma, el elemento (i, j), con $i > j$, de la matriz de distancias es la distancia medida en angstroms (Å) predicha entre el aminoácido i-ésimo y el j-ésimo.

2.2. Proceso de entrenamiento

El sistema de predicción propuesto, denominado ASPF-PRED (Aminoacid Subsequences Property File Predictor), se divide en dos fases. En una primera fase, se genera un modelo de conocimiento a partir de todos los fragmentos o subsecuencias de todas las proteínas de un conjunto de entrenamiento. En una segunda fase, se realizan las predicciones de todas las proteínas de un conjunto de test usando el modelo de conocimiento generado.

El modelo de conocimiento está formado por un conjunto de vectores denominados vectores de predicción. Cada vector de predicción se obtiene a partir de una subsecuencia de una proteína de entrenamiento y contiene la longitud de la subsecuencia, valores promedio de las propiedades físico-químicas de sus aminoácidos interiores y la distancia real entre los extremos de la subsecuencia. En la Figura 1 se define formalmente el contenido de un vector de predicción de una subsecuencia de proteína.

	L	$\overline{P}_1$	$\cdots$	$\overline{P}_k$	D
$a_1a_2...a_m$	$m/lmax$	$\frac{1}{m}\sum_{i=2}^{m-1} P_1(a_i)$	$\cdots$	$\frac{1}{m}\sum_{i=2}^{m-1} P_k(a_i)$	$d(a_1, a_m)$

Figura 1. Un vector de predicción

La longitud L de cada subsecuencia se ha normalizado entre 0 y 1. Para ello, se ha dividido la longitud de cada subsecuencia entre la longitud máxima $lmax$ de todas las proteínas de training. La normalización es importante para que todos los atributos del vector de predicción estén en la misma escala y contribuyan por igual a la predicción.

Las propiedades $P_1 \ldots P_k$, atribuibles a cada aminoácido de la subsecuencia, también se normalizan, se promedian y se almacenan en el vector de predicción ($\overline{P}_1 \ldots \overline{P}_k$). Finalmente, se añade a cada vector la distancia real (D) entre los aminoácidos extremos (primero y último de la subsecuencia).

Las propiedades físico-químicas que se han utilizado son: Residue accessible surface area in folded protein [4], Average relative fractional occurrence in ER(i-1) [10] y RF value in high salt chromatography [13].

2.3. Proceso de predicción

En la segunda fase de nuestro sistema, se obtienen todos los vectores de predicción de las proteínas de test y se realiza una búsqueda secuencial completa a partir de cada uno de ellos sobre los vectores de predicción de training. El objetivo es encontrar el vector de predicción de training que guarda mayor similitud con cada vector de predicción de test. Para el proceso de búsqueda sólo se consideran los vectores de training con los mismos extremos que el vector de test. En la Figura 2 se ilustra el escenario de la búsqueda.

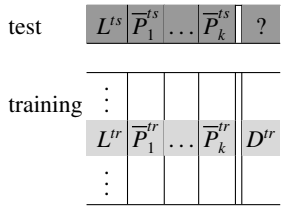

Figura 2. Búsqueda del vector de predicción más parecido

En el esquema de búsqueda de la Figura 2, L^{ts} es la longitud de la subsecuencia de test. L^{tr} es la longitud de la subsecuencia de training con mayor similitud a la subsecuencia de test. $\overline{P}_1^{ts} \ldots \overline{P}_k^{ts}$ son los valores medios de las propiedades de los aminoácidos de la subsecuencia de test y $\overline{P}_1^{tr} \ldots \overline{P}_k^{tr}$ los de la subsecuencia de training más próxima. La distancia a predecir se simboliza con ? y se le asignará la distancia D^{tr} del vector de training más parecido.

El vector de training con mayor similitud al vector de test satisface la condición expresada en la fórmula 1. Como se puede apreciar en dicha condición, para la comparación de vectores de predicción se utiliza una distancia euclídea entre los vectores de test y de training, en la que intervienen las longitudes de las subsecuencias y los valores promedios de las propiedades de sus aminoácidos, con iguales ponderaciones.

$$min \sqrt{(L^{ts} - L^{tr})^2 + (\overline{P}_1^{ts} - \overline{P}_1^{tr})^2 + \cdots + (\overline{P}_k^{ts} - \overline{P}_k^{tr})^2} \quad (1)$$

Al campo distancia D de cada vector de test se le asigna el valor del campo distancia del vector de training más próximo. La distancia asignada a cada vector de test representa la distancia predicha entre los aminoácidos extremos de la subsecuencia a la que se

refiere dicho vector. Finalmente, las distancias predichas son almacenadas en la triangular inferior de la matriz de distancias de la secuencia de test.

2.4. Medidas de evaluación utilizadas

Para evaluar la calidad de las predicciones, se han calculado varias medidas. La primera medida es la precisión, utilizada en los trabajos de Fariselli et al. [5, 6]. La segunda medida es la sensibilidad o recall, utilizada en otras propuestas [15]. En último lugar, también se han obtenido las medidas de exactitud, especificidad y Matthews Correlation Coefficient (MCC), que ofrece una evaluación más balanceada que, por ejemplo, un porcentaje [2]. Las siguientes fórmulas (2,3,4,5,6) definen estas 5 medidas.

$$Precisión = \frac{TP}{TP + FP} \tag{2}$$

$$Sensibilidad = \frac{TP}{TP + FN} \tag{3}$$

$$Exactitud = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

$$Especificidad = \frac{TN}{TN + FP} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Estas medidas se utilizan para evaluar la calidad de predicciones sobre una clase binaria; esto es, el valor a predecir es 0 ó 1. De este modo, se tienen cuatro situaciones posibles, dando lugar a cuatro recuentos de la ocurrencia de las mismas: a) tanto el valor real como predicho es 1 (TP, *true positives*), b) tanto el valor real como predicho es 0 (TN, *true negatives*), c) el valor real es 1 y la predicción 0 (FN, *false negatives*) y d) el valor real es 0 y la predicción 1 (FP, *false positives*). Dado que la clase a predecir, en el caso que nos ocupa, es un valor real (una distancia), para obtener estas medidas ha sido necesario binarizar previamente la clase utilizando una distancia umbral o cutoff.

En este trabajo se ha empleado un valor de cutoff de 8 angstroms, que es habitualmente utilizado en la literatura [5, 6, 15]. En la evaluación de las medidas se han omitido las predicciones de los pares de aminoácidos con separación mínima en la secuencia de la proteína de 7 aminoácidos, tal como se propuso en el trabajo de Fariselli et al. [6].

3. Experimentación

Se ha realizado una experimentación para comprobar la validez del método sobre todas las proteínas de cápsides de virus (Viral Capsid, GO ID: 19028) publicadas en Protein Data Bank con máxima identidad del 30 % (proteínas no homólogas) a fecha de Mayo de 2011 (67 proteínas). En la Tabla 1 se muestran los códigos PDB de las proteínas utilizadas en el estudio. Se ha utilizado una validación cruzada leave-one-out

Cuadro 1. Proteínas utilizadas en la experimentación mediante leave-one-out

$L < 150$	1TD4	1CD3	2IZW	1C8D	1MUK	3IYH
1C5E	1VD0	1EI7	2VTU	1DZL	1OPO	3IYK
1GFF	1W8X	1F15	2VVF	1EJ6	1P2Z	3IYL
1HGZ	2C0W	1F2N	2WLP	1FN9	1QHD	3JYR
1IFK	2KX4	1JS9	2ZL7	1HX6	1SVA	3KIC
1IFL	2QUD	1STM	3FMG	1IHM	1YUE	3KZ4
1IFP	2VF9	1VPS	3KML	1KVP	2BBD	1A6C
1JMU	$L150-300$	1X36	3IYP	1LP3	2JHP	1BVP
1MSC	1AUY	1ZA7	2FT1	1M1C	2TBV	3QPR
1QBE	1C8N	2BUK	$L > 300$	1M3Y	2XVR	3IZ3

Cuadro 2. Resultados obtenidos en la predicción

Set de proteínas	Sensibilidad	Precisión	Exactitud	Especificidad	MCC
Todas las proteínas (67)	0.78	0.76	0.99	0.99	0.76
$L < 150$ (16)	0.85	0.83	0.99	0.99	0.84
$150 \leqslant L < 300$ (21)	0.81	0.76	0.99	0.99	0.78
$L \geqslant 300$ (30)	0.75	0.74	0.99	0.99	0.74

para evitar el efecto de la elección de bolsas de una validación cruzada con bolsas. En la tabla 2 se presentan las medidas de evaluación obtenidas en la experimentación.

Como se muestra en la Tabla 2, se ha obtenido un valor de precisión de 0.76 y una sensibilidad de 0.78, para el grupo completo de proteínas de estudio. Para valorar la calidad de las predicciones obtenidas con nuestro método y tener unos valores de referencia, se indican los resultados obtenidos con otras propuestas de predicción de estructuras de proteínas. En concreto, en el trabajo de Zhang et al. 2005 [15] se obtuvo un valor de recall de 0.27 para 8 angstroms y 5 proteínas de test. En la propuesta de Fariselli et al. 2001 [6] se consiguió, mediante validación cruzada, un valor de precisión de 0.21 para 8 angstroms de cut-off y 7 aminoácidos de separación mínima en la secuencia.

Generalmente la precisión de la predicción de estructuras de proteínas con secuencias largas (mayor de 300 aminoácidos) suele ser menor que con secuencias más cortas, debido a la dificultad de las primeras de ser predichas. Por ejemplo, en el trabajo de Fariselli et al. 2001 [6] se obtiene un 0.11 de precisión para proteínas de 300 aminoácidos o más. Con nuestro método se obtiene una precisión de 0.74 para proteínas del mismo rango de longitud.

En la Figura 3 se muestra el mapa de distancias obtenido para la proteína 1M3Y (413 aminoácidos) del conjunto de estudio. Se ha utilizado una escala de colores para representar las distancias: desde la distancia mínima (rojo) a la máxima (azul). Según se aprecia en dicha figura, la triangular inferior de la matriz (predicción) se asemeja en gran medida a la triangular superior (observación).

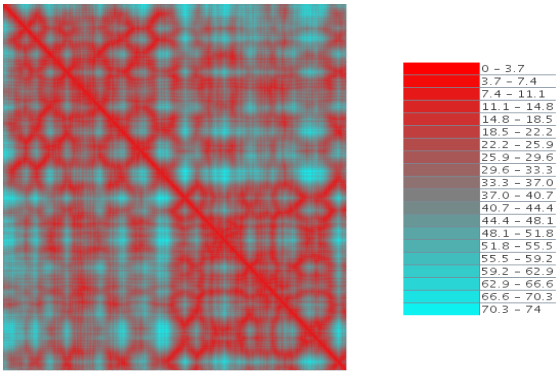

Figura 3. Mapa de distancias predicho para la proteína 1M3Y del conjunto de estudio y su escala de colores

En la Figura 4 se muestra el mapa de contactos de la misma proteína 1M3Y, obtenido a partir del mapa de distancias de la Figura 3 utilizando un umbral de corte de 8 angstroms. Tal como ocurre en el mapa de distancias, se aprecia gran similitud entre la parte real y la parte predicha. Finalmente, la estructura 3D de la proteína 1M3Y se muestra en la Figura 5.

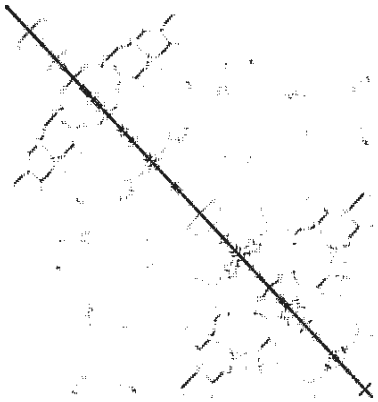

Figura 4. Mapa de contactos predicho para la proteína 1M3Y obtenido a partir del mapa de distancias con un cut-off de 8Å

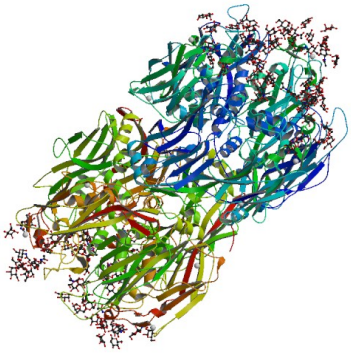

Figura 5. Representación 3D real de la proteína 1M3Y del conjunto de estudio

4. Conclusiones y trabajo futuro

La predicción de la estructura terciaria de las proteínas consiste en determinar la conformación tridimensional de las mismas únicamente a partir de su secuencia de aminoácidos. En este trabajo se ha propuesto un método basado en ensamblaje de fragmentos de proteínas por similitud físico-química, utilizando tres propiedades físico-químicas de aminoácidos. Se predicen mapas de distancias, los cuales proveen más información sobre la estructura de una proteína que los mapas de contactos. Se ha realizado una experimentación para comprobar la validez del método sobre todas las proteínas no homólogas de cápsides de virus disponibles actualmente. Se ha obtenido una precisión de 0.76 y una sensibilidad de 0.78 con mínima separación en la secuencia de 7 aminoácidos y cut-off de 8 Å, que mejora notablemente, con las proteínas estudiadas, los resultados de otras propuestas.

Como trabajo futuro, nos proponemos refinar a posteriori los mapas de distancias producidos por nuestro sistema, comprobando si satisfacen determinadas restricciones geométricas y químicas indicadas para mapas de distancias. Asimismo, nos proponemos estudiar otras propiedades físico-químicas de aminoácidos y comprobar su validez en el problema de la predicción de estructuras de proteínas.

Referencias

1. Anfinsen, C.: The formation and stabilization of protein structure. The Biochemical journal 128(4), 737–749 (1972)
2. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F., Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5), 412–424 (2000), http://bioinformatics.oxfordjournals.org/content/16/5/412.abstract
3. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. Nucl. Acids Res. 28(1), 235–242 (2000)
4. Chothia, C.: The nature of the accessible and buried surfaces in proteins. Journal of Molecular Biology 105(1), 1 – 12 (1976), http://www.sciencedirect.com/science/article/B6WK7-4FNGC6X-P/2/9ba9b95c0384b5b57b416b69641292d8

5. Fariselli, P., Casadio, R.: A neural network based predictor of residue contacts in proteins. Protein Engineering 12(1), 15–21 (1999), http://peds.oxfordjournals.org/content/12/1/15.abstract
6. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact maps with neural networks and correlated mutations. Protein Engineering 14(11), 835–843 (2001), http://peds.oxfordjournals.org/content/14/11/835.abstract
7. Hoque, T., Chetty, M., Sattar, A.: Extended hp model for protein structure prediction. Journal of computational biology : a journal of computational molecular cell biology 16(1), 85–103 (2009)
8. Jones, D.: Predicting novel protein folds by using fragfold. Proteins Suppl 5, 127–132 (2001)
9. Li, S.C., Bu, D., Xu, J., Li, M.: Fragment-hmm: a new approach to protein structure prediction. Protein science : a publication of the Protein Society 17(11), 1925–1934 (2008)
10. Rackovsky, S., Scheraga, H.A.: Differential geometry and polymer conformation. 4. conformational and nucleation properties of individual amino acids. Macromolecules 15(5), 1340–1346 (1982), http://pubs.acs.org/doi/abs/10.1021/ma00233a025
11. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., Baker, D.: Protein structure prediction using rosetta. In: Brand, L., Johnson, M.L. (eds.) Numerical Computer Methods, Part D, Methods in Enzymology, vol. 383, pp. 66 – 93. Academic Press (2004), http://www.sciencedirect.com/science/article/B7CV2-4C47J19-5/2/88812f076bde6b4f528d349c64a7f997
12. Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., Boniecki, M.: Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. Proteins 45(S5), 149–156 (2001), http://dx.doi.org/10.1002/prot.1172
13. Weber, A.L., Lacey, J.C.: Genetic code correlations: Amino acids and their anticodon nucleotides. Journal of Molecular Evolution 11, 199–210 (1978), http://dx.doi.org/10.1007/BF01734481, 10.1007/BF01734481
14. Wu, S., Skolnick, J., Zhang, Y.: Ab initio modeling of small proteins by iterative tasser simulations. BMC Biology 5(1), 17 (2007), http://www.biomedcentral.com/1741-7007/5/17
15. Zhang, G.Z., Huang, D.S., Quan, Z.H.: Combining a binary input encoding scheme with rbfnn for globulin protein inter-residue contact map prediction. Pattern Recogn. Lett. 26, 1543–1553 (July 2005), http://dx.doi.org/10.1016/j.patrec.2005.01.005