RIDAA
Repositorio Institucional
Digital de Acceso Abierto de la
Universidad Nacional de Quilmes

Universidad
Nacional
de Quilmes

Teppa, Rosana Elin

Identificación de redes de residuos con alta información mutua implicancias para la detección de residuos funcionalmente importantes

Esta obra está bajo una Licencia Creative Commons Argentina.
Atribución - No Comercial - Sin Obra Derivada 2.5
https://creativecommons.org/licenses/by-nc-nd/2.5/ar/

Documento descargado de RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes de la Universidad Nacional de Quilmes

Cita recomendada:
Teppa, R. E. (2017). Identificación de redes de residuos con alta información mutua implicancias para la detección de residuos funcionalmente importantes. (Tesis de doctorado). Universidad Nacional de Quilmes, Bernal, Argentina. Disponible en RIDAA-UNQ Repositorio Institucional Digital de Acceso Abierto de la Universidad Nacional de Quilmes http://ridaa.unq.edu.ar/handle/20.500.11807/338

Puede encontrar éste y otros documentos en: https://ridaa.unq.edu.ar

Universidad
Nacional
de Quilmes

Roque Sáenz Peña 352 // Bernal
Buenos Aires // Argentina
t.: (+41 11) 4365 7100
f.: (+54 11) 4365 7101
info@unq.edu.ar

Teppa, Rosana Elín, Repositorio Institucional Digital de Acceso Abierto,
Septiembre del 2014, 153 pp.,
http://ridaa.unq.edu.ar,
Universidad Nacional de Quilmes, Secretaría de Posgrado,
Doctorado en Ciencias Básicas y Aplicadas

Identificación de redes de residuos con alta información mutua implicancias para la detección de residuos funcionalmente importantes

Rosana Elín Teppa

elinteppa@gmail.com

Resumen

Los aminoácidos en una proteína pueden variar, sin embargo hay ciertas restricciones funcionales y estructurales a estos cambios para mantener una proteína funcional. Frecuentemente las posiciones de residuos funcionalmente importantes se encuentran conservadas. Sin embargo, estudios mutacionales han demostrado que muchas posiciones con baja conservación pueden tener relevancia funcional. Particularmente se encuentran mutaciones compensatorias, dadas por residuos que varían de manera coordinada para preservar o restaurar la función/estructura proteica. Estas posiciones coevolucionadas han sido de interés por su uso en la identificación de residuos que interactúan dentro de una proteína y en la predicción de sitios funcionalmente importantes.

Una manera de inferir coevolución en una proteína es cuantificar la covariación de los residuos a partir de un alineamiento múltiple de secuencias homólogas. Se han desarrollado una gran cantidad de métodos para cuantificar esta covariación, uno de ellos es la Información Mutua derivada de la teoría de la información.

El presente trabajo tiene por objetivo el estudio del uso de la Información Mutua, y medidas derivadas de ella, para la predicción de sitios funcionalmente importantes en proteínas. Luego de un capítulo introductorio, el capítulo 2 está dedicado al desarrollo de un método predictivo de residuos catalíticos en enzimas utilizando diferentes puntajes derivados de la Información Mutua, la conservación de residuos e información estructural.

En el capítulo 3 se incluye el análisis de un tipo particular de posiciones importantes, llamadas determinantes de especificidad. Estos residuos, al igual que los residuos con alta Información Mutua, pueden encontrarse en la proximidad espacial de los residuos catalíticos. Se evaluó el grado de coincidencia entre ambos tipos de residuos funcionalmente importantes y su utilidad combinada para la predicción de residuos catalíticos.

En el capítulo 4 se realiza un análisis de secuencias exhaustivo que incluye la predicción de sitios coevolucionados y determinantes de especificidad para una familia particular de enzimas, de relevancia en el área de la glicobiología. Los resultados incluyen la identificación de nuevas secuencias y la detección de nuevas subfamilias. Además ayudaron a obtener una mejor comprensión de la divergencia funcional y a proponer un modelo evolutivo de la familia.

El capítulo final contiene el estudio de la interfaz de interacción proteína-proteína, mediante puntajes derivados de la Información Mutua y conservación. El objetivo de este estudio es aportar información a fin de comprender la evolución de los sitios responsables de la interacción y de su entorno estructural. Creemos que los resultados obtenidos pueden guiar el desarrollo de un método predictivo de la interfaz de interacción proteína-proteína.

PROPUESTA DE TESIS DOCTORAL

IDENTIFICACIÓN DE REDES DE RESIDUOS CON ALTA INFORMACIÓN MUTUA

IMPLICANCIAS PARA LA DETECCIÓN DE RESIDUOS FUNCIONALMENTE IMPORTANTES

LICENCIADA ELÍN TEPPA

DIRECTOR
DR. CRISTINA MARINO BUSLJE
CO-DIRECTOR
DR. MORTEN NIELSEN

COMITÉ DE EVALUACIÓN
DR. JULIO CARAMELO
DR. IGNACIO PONZONI
DR. SEBASTIÁN FERNANDEZ ALBERTI

FEBRERO, 2014
ARGENTINA

"It takes all the running you can do,
to keep in the same place.
If you want to get somewhere else,
you must run at least twice as fast as that!"

— Lewis Carroll,Through the Looking-Glass.

AGRADECIMIENTOS

En primer lugar quiero agradecer a la directora de este trabajo, Dr. Cristina Marino Buslje por su dedicación y paciencia en todos estos años. Ha guiado mi trabajo con entusiasmo, compromiso y excelencia científica. Su optimismo y calidez humana han hecho de mi aprendizaje un camino placentero.

También quiero agradecer al Dr. Morten Nielsen por haber aceptado co-dirigir mi tesis, por su buena predisposición para responder mis innumerables preguntas y reorientar mi investigación oportunamente. Su claridad para desarrollar ideas y transmitirlas de manera sencilla, han facilitado enormemente mi trabajo.

Igualmente agradezco a los miembros del comité de seguimiento tesis, los doctores Julio Caramelo, Ariel Chernomoretz e Ignacio Sánchez, por el tiempo dedicado a seguir mi trabajo, sus valiosos comentarios y discusiones.

Me alegra agradecer a mis compañeros del laboratorio de Bioinformática y a los integrantes del laboratorio de Biología de Sistemas Integrativa del Instituto Leloir, por su compañerismo, apoyo constante y espíritu colaborativo. Especialmente a Diego Zea y Ariel Berenstein por su sincera amistad durante todos estos años y por compartir conmigo sus conocimientos. Así mismo agradezco a Diego Vadell por su ayuda técnica, que fue siempre más allá de su obligación profesional.

Quiero expresar también mi más sincero agradecimiento a la investigadores Dr. Anne Harduin-Leppers, por introducirme en el fascinante campo de la glicobiología, por su confianza y su cálido recibimiento en su laboratorio. Así mismo agradezco al laboratorio de Bioinformática Estructural del GRIB, especialmente al Dr. Baldo Oliva, por haber guiado parte de mi trabajo de doctorado y permitirme desarrollarlo en total libertad.

Agradezco por el financiamiento a la fundación Richard Lounsbery, CONICET, Erasmus Mundus y al Instituto Leloir por haberme facilitado los medios suficientes para llevar a cabo todas las actividades propuestas durante el desarrollo de esta tesis.

Finalmente quiero agradecer a mi familia por su constante apoyo y comprensión.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

DECLARACIÓN

Por la presente declaro que el trabajo propuesto como tesis ha sido escrito por mí, que es el registro de los trabajos realizados durante los años 2010-2013 y que no ha sido presentado previamente en ninguna otra Universidad para la obtención de un título de doctorado.

Buenos Aires, Febrero 2014

Elin Teppa

PUBLICACIONES

Como resultado del trabajo realizado en el período de mi tesis doctoral se han publicado los siguientes artículos (incluídos como Anexo) y capítulos de libro.

Artículos:

1. *Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification.* Marino Buslje C [§] , Teppa E [§] , Di Doménico T, Delfino JM, Nielsen M. PLoS Comput Biol. 2010.[§] Contribución equitativa.

2. *Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction*. Teppa E, Wilkins AD, Nielsen M [§] , Buslje CM [§] . BMC Bioinformatics. 2012. [§] Contribución equitativa.

3. *MISTIC: Mutual information server to infer coevolution*. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. Nucleic Acids Res. 2013 (Web Server issue).

4. *Integrative view of β-galactoside α2,3-sialyltransferases (ST3Gal) molecular and functional evolution in Deuterostomes: significance of lineage specific losses.* Petit D [§] , Teppa E [§] , Mir AM, Vicogne D, Thisse C, Thisse B, Filloux C, Harduin-Lepers A. [§] Contribución equitativa. Artículo en revisión.

Capítulos de libros:

1. *Identification of coevolving amino acids using mutual information. Residues coevolution networks and their implication for functional important site prediction.* Teppa E, Zea D, Marino Buslje C. Libro: Introduction to Sequence and Genome Analysis. ISBN: 9781477554913. iConcept Press. 2012.

2. *A Practical Approach to Reconstruct Evolutionary History of Animal Sialyltransferases and Gain Insights into the Sequence-Function Relationships of Golgi- glycosyltransferases.* Petit D, Teppa E, Petit JM, Harduin-Lepers A. Libro: Glycosyltransferases: Methods and Protocols. Springer. ISBN: 9781627034647. 2013

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

ÍNDICE DE TABLAS

NOMENCLATURA

APC	Average Product Correction
AUC	Área Bajo la Curva ROC
CCS	Coeficiente de correlación de Spearman
Cls	Puntaje de Versomilitud de Catalítico (por Catalitic Likeliness Score)
CSA	Base de datos Catalytic Site Atlas
ivET	Integer value ET
KL	Conservación Kullback-Leibler
MB	Mutational behavior. Método de predicción de sitios determinantes de specificidad.
MI	Información Mutua
MSA	Alineamiento Múltiple de Secuencias
pC	Conservación en la proximidad (por proximity Conservation)
pCexp	Conservación en la proximidad considerando residuos expuestos
PDB	Protein Data Bank
pMI	Información Mutua en la proximidad (por proximity Mutual Information)
pMIexp	Información Mutua en la proximidad considerando residuos expuestos
RCW	Row Column Weighting
ROC	acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor
RSA	Area Relativa Accesible al Solvente
rvET	real value Evolutionary Trace
SDP	Posición Determinante de Specificidad
ST3Gal	alpha2,3-sialiltransferasas

RESUMEN

Los aminoácidos en una proteína pueden variar, sin embargo hay ciertas restricciones funcionales y estructurales a estos cambios para mantener una proteína funcional. Frecuentemente las posiciones de residuos funcionalmente importantes se encuentran conservadas. Sin embargo, estudios mutacionales han demostrado que muchas posiciones con baja conservación pueden tener relevancia funcional. Particularmente se encuentran mutaciones compensatorias, dadas por residuos que varían de manera coordinada para preservar o restaurar la función/estructura proteica. Estas posiciones coevolucionadas han sido de interés por su uso en la identificación de residuos que interactúan dentro de una proteína y en la predicción de sitios funcionalmente importantes.

Una manera de inferir coevolución en una proteína es cuantificar la covariación de los residuos a partir de un alineamiento múltiple de secuencias homólogas. Se han desarrollado una gran cantidad de métodos para cuantificar esta covariación, uno de ellos es la Información Mutua derivada de la teoría de la información.

El presente trabajo tiene por objetivo el estudio del uso de la Información Mutua, y medidas derivadas de ella, para la predicción de sitios funcionalmente importantes en proteínas. Luego de un capítulo introductorio, el capítulo 2 está dedicado al desarrollo de un método predictivo de residuos catalíticos en enzimas utilizando diferentes puntajes derivados de la Información Mutua, la conservación de residuos e información estructural.

En el capítulo 3 se incluye el análisis de un tipo particular de posiciones importantes, llamadas determinantes de especificidad. Estos residuos, al igual que los residuos con alta Información Mutua, pueden encontrarse en la proximidad espacial de los residuos catalíticos. Se evaluó el grado de coincidencia entre ambos tipos de residuos funcionalmente importantes y su utilidad combinada para la predicción de residuos catalíticos.

En el capítulo 4 se realiza un análisis de secuencias exhaustivo que incluye la predicción de sitios coevolucionados y determinantes de especificidad para una familia particular de enzimas, de relevancia en el área de la glicobiología. Los resultados incluyen la identificación de nuevas secuencias y la detección de nuevas subfamilias. Además ayudaron a obtener una mejor comprensión de la divergencia funcional y a proponer un modelo evolutivo de la familia.

El capítulo final contiene el estudio de la interfaz de interacción proteína-proteína, mediante puntajes derivados de la Información Mutua y conservación. El objetivo de este estudio es aportar información a fin de comprender la evolución de los sitios responsables de la interacción y de su entorno estructural. Creemos que los resultados obtenidos pueden guiar el desarrollo de un método predictivo de la interfaz de interacción proteína-proteína.

1

INTRODUCCIÓN A LA COEVOLUCIÓN SITIO ESPECÍFICA

El fenómeno de Coevolución fue primeramente descripto en términos macroscópicos, una de las definiciones de coevolución más ampliamente aceptada fue dada por Thompson [121] como cambios evolutivos que ocurren de manera recíproca entre especies interactuantes. Refiriéndose a que el cambio en una población modifica la presión de selección sobre una segunda población y vice versa, un ejemplo muy estudiado a nivel de organismos es la coevolución patógeno/huésped.

De manera análoga se define la coevolución molecular como el cambio en un locus que afecta la presión de selección de otro locus, siendo este cambio recíproco Atchley et al. [5]. La evolución de muchos genes y proteínas no es independiente sino que están ligadas a la evolución de otros componentes. Lo mismo sucede a un nivel inferior, residuos individuales coevolucionan dentro de una misma proteína, lo que se denomina coevolución sitio específica.

Aunque los cambios en diferentes posiciones en una proteína pueden ser pasados de una generación a la siguiente, la función y la conformación espacial de las proteínas, imponen restricciones en la variaciones de residuos que pueden ser aceptados para que la proteína siga siendo funcional. Es esperable que los sitios que se encuentran próximos en la estructura tridimensional sean los que impongan restricciones en los cambios unos a otros, es decir, que cambien de manera concertada o coevolucionen Altschuh et al. [3], Clarke [28], Atchley et al. [5], Gloor et al. [48], Yeang and Haussler [132], Halabi et al. [52], Aguilar et al. [1]. En la Figura 1 se esquematiza el concepto de coevolución entre residuos interactuantes.

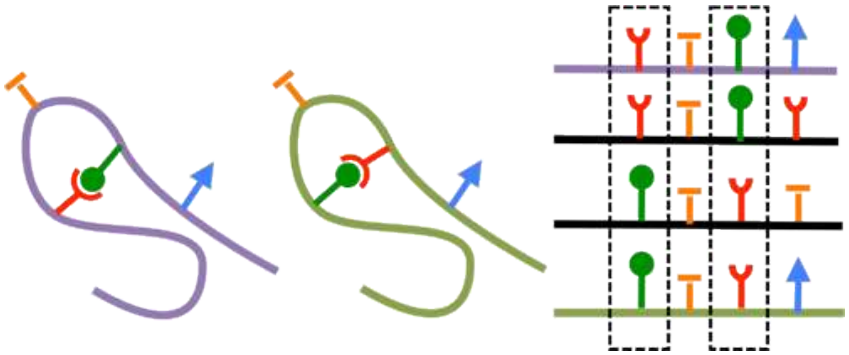


Figura 1: Representación esquemática de la coevolución sitio específica. A la izquierda se representan aminoácidos con formas complementarias (verde y rojo), que han intercambiado sus posiciones en la cadena polipeptídica. A la derecha se representa el alineamiento múltiple de secuencias correspondiente donde se puede observar la covariación de residuos en las posiciones marcadas en linea punteada. Imagen extraída de http://gremlin.bakerlab.org/gremlin_faq.php

Los aminoácidos funcionalmente relacionados se encuentran vinculados evolutivamente ya que una mutación en alguno de ellos puede causar un efecto dramático en la función y en la presión de selección sobre otras posiciones de aminoácidos. En estos casos para que una mutación sea fijada, son necesarias mutaciones compensatorias en otro u otros sitios, para mantener o restaurar una función. Este concepto de coevolución proviene de la idea de covariación propuesta por Fitch and Markowitz [45], donde se propone que en un momento evolutivo dado, una región particular de la proteína es invariable (debido a restricciones espaciales o funcionales) mientras que otras regiones acumulan mutaciones. Debido a que las mutaciones se fijan en alguna parte de la secuencia a lo largo de la evolución, esto indica que las restricciones selectivas de las regiones invariables pueden cambiar.

1.1 COVARIACIÓN Y COEVOLUCIÓN

Es importante distinguir entre la mera covariación de residuos, que puede observarse en un Alineamiento Múltiple de Secuencias (MSA por Multiple Sequence Alignment) y la coevolución. Mientras que la primera se refiere al cambio simultáneo de aminoácidos sin importar las causas de éstos cambios, la coevolución implica que el cambio sea recíproco y debido a una presión de selección en común Codoner and Fares [29].

La covariación entre una posición x e y en un alineamiento múltiple de secuencias puede ser descompuesta en diferentes términos como se muestra en la ecuación (1) y se esquematiza en la Figura 2.

$$C_{xy} = C_{filogenia} + C_{estructura} + C_{función} + C_{interacción} + C_{estocástica} \quad (1)$$

La covariación filogenética ($C_{filogenia}$) fue primeramente explicada por Felsenstein donde se señala la dependencia histórica entre las especies y además que esta dependencia es extensible a nivel de residuos proteicos. Los componentes $C_{estructura}$ y $C_{función}$ dan cuenta de la covariación debida a una presión de selección en común para mantener la estructura o función Atchley et al. [5] y son difíciles de distinguir ya que no son mutuamente excluyentes y grupos de aminoácidos pueden coevolucionar debido a la combinación de diferentes dependencias. El término $C_{interación}$ se refiere a la coevolución existente entre posiciones de aminoácidos que interactúan, usualmente refleja un componente estructural y/o funcional, lo que hace aún más difícil su distinción de los otros términos. La covariación estocástica ($C_{estocástica}$) puede deberse a covariación convergente de los sitios debido a su dinámica de mutación. Las dificultades para quitar este componente se deben a las limitaciones para producir un modelo de análisis para dar cuenta de la covarianza estocástica, por lo que muchos métodos se basan en simulaciones de MSAs. Estos MSAs simulan los parámetros evolutivos de alineamientos biológicos y se pueden usar para producir una distribución de las probabilidades de detectar coevolución bajo un cierto modelo de sustitución de aminoácidos. La identificación de la coevolución estocástica está muy condicionada por las propiedades estadísticas del MSA. Los MSAs de baja calidad y poco poblados son más propensos a producir sitos falsos de coevolución funcional, como resultado del efecto significativo de la estocasticidad en la detección de la coevolución Fares and Travers [44].

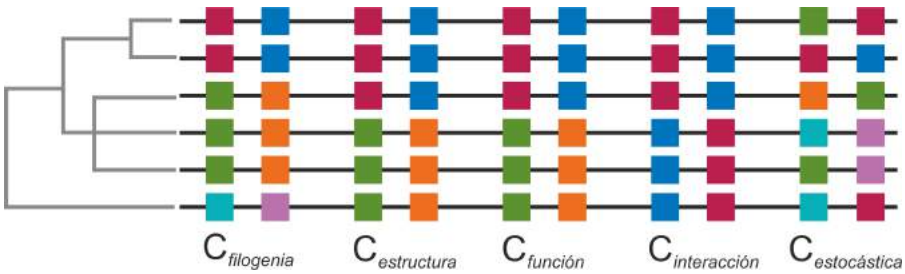


Figura 2: Componentes de la señal de coevolución.
Los cuadrados de diferentes color representan diferentes aminoácidos. Las diferentes secuencias (lineas horizontales) están filogenéticamente relacionadas siguiendo la topología mostrada (izquierda).

1.2 INTRODUCCIÓN A LOS MÉTODOS GENERALES

1.2.1 *Información Mutua*

En términos biológicos para el análisis de secuencias, la Información Mutua (MI por Mutual Information) entre dos posiciones (dos columnas de un MSA) refleja la disminución de la incertidumbre en una posición a partir del conocimiento de otra posición. Intuitivamente, la MI mide la información compartida por dos columnas de un alineamiento múltiple de secuencias. Dada dos posiciones cualesquiera en un MSA que pueden considerarse variables aleatorias x e y, la MI entre ellas está dada por la relación que se muestra en al ecuación (2)

$$MI_{xy} = \sum P\left(a_x, b_y\right) \cdot \log\left(\frac{P\left(a_x, b_y\right)}{P\left(a_x\right) \cdot P\left(b_y\right)}\right) \tag{2}$$

Donde $P\left(a_x, b_y\right)$ es la frecuencia de aparición del aminoácido a en la posición x y el aminoácido b en la posición y en la misma secuencia; $P\left(a_x\right)$ es la frecuencia del aminoácido a en la posición x, $P\left(b_y\right)$ es la frecuencia de aparición del aminoácido b en la posición y en el alineamiento.

Un valor de MI igual a cero indica independencia evolutiva entre los sitios x e y. Cabe mencionar que dos posiciones invariables representarían un caso extremo de co-aparición, y la MI entre ellas es cero. Es decir, se asume independencia evolutiva al no tener evidencia de la co-variación entre ellas. Para los valores positivos de MI, su magnitud depende de la fuerza de covariación entre ambos sitios Blahut, RE [12]. Como consecuencia el poder de la MI para predecir coevolución real depende del nivel de conservación en el MSA Fodor and Aldrich [46], Martin et al. [83].

La MI permite inferir sitios que varían de manera correlacionada en secuencias homólogas. Aunque en principio el cálculo de MI es simple, su interpretación biológica y evolutiva ha sido discutida y diferentes enfoques han testeado su utilidad Dunn et al. [38], Cover and Thomas [30], del Sol et al. [35], Halabi et al. [52].

La aplicación de la MI a alineamientos múltiples de secuencias fue introducido por primera vez por Korber et al. [70] para la identificación de sitios que covariaban en péptidos virales. Luego el uso de la MI fue extendido como medida de coevolución por Giraud et al. [47]. La ventaja del uso de MI para cuantificar la coevolución radica en la aplicabilidad del método sin requerir conocimientos acerca de la relación entre los residuos en el MSA y su dinámica evolutiva. Los valores de MI se ven afectados por la historia filogenética común entre las secuencias ($C_{filogenia}$), a menos que esa dependencia sea corregida explícitamente en el modelo de detección de coevolución. Así, un gran desafío para la detección de la coevolución fue la distinción

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

de correlaciones filogenéticas del resto de las correlaciones. Muchos estudios han intentado corregir los valores de MI quitando el efecto de la dependencia filogenética Tillier and Lui [122], Wollenberg and Atchley [129], Gouveia-Oliveira and Pedersen [49], Dunn et al. [38], Buslje et al. [16], Dutheil [39].

La dependencia entre las posiciones de un MSA se debe a que las secuencias proteicas no son independientes, sino que contienen una señal inherente dada por su historia evolutiva en común. Esto ha sido claramente demostrado por Gouveia-Oliveira and Pedersen [49], donde muestran cómo secuencias relacionadas pueden resultar en una señal de covariación que se parece a la coevolución. En la Figura 3 se esquematiza el fenómeno mostrando variaciones independientes que dan una señal que imita a la señal de coevolución.

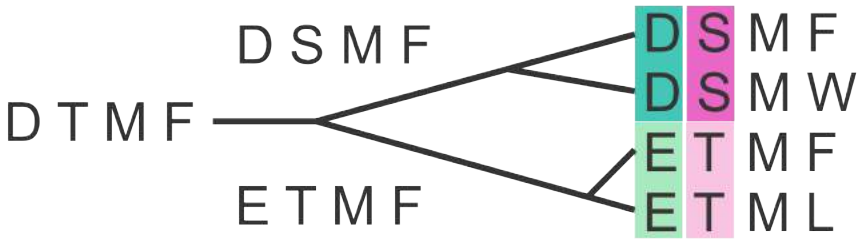


Figura 3: Señal filogenética que imita coevolución
Dos mutaciones independientes pueden resultar en grupos de secuencias que muestran una señal que puede confundirse con coevolución. Las posiciones 1 y 2 del alineamiento presentan alta MI aunque no coevolucionan, ya que las mutaciones se produjeron de manera independiente sobre las distintas secuencias. (Figura basada en Figura 1 de Gouveia-Oliveira and Pedersen [49])

Se han propuesto diferentes abordajes para disminuir la señal impuesta por la filogenia a fin de mejorar la identificación de posiciones que coevolucionan. Una manera simple de tener en cuenta los ancestros compartidos es medir el grado de correlación que puede esperarse solamente por filogenia y estocasticidad. Esto se realiza mediante un procedimiento de randomización o *bootstrap*, el cual elimina cualquier correlación funcional y permite obtener una estimación empírica de la distribución nula de las correlaciones.

Un obstáculo adicional para la detección de coevolución es la falta de un set de datos de referencia construido a partir de datos reales para evaluar los métodos, ya que nunca accedemos a la historia de coevolución real en los datos biológicos. Como aproximación se asume que los pares de residuos que están en contacto (por ejemplo distancia $C\beta$ menor a 8Å) coevolucionan Korber et al. [70], Gouveia-Oliveira and Pedersen [49], Del Sol et al. [36], Buslje et al. [16]. Es evidente que es una aproximación a la realidad, ya que hay muchos sitios que están en contacto que no coevolucionan y a la vez muchos

sitios que coevolucionan no necesariamente están en contacto directo. Sin embargo para la mayoría de los pares que coevolucionan se puede asumir que están en contacto, considerando que pueden influenciarse mutuamente de manera directa.

La sensibilidad de la mayoría de los métodos desarrollados para detectar coevolución utilizando MI depende de ciertas característica del MSA como ser: (*a*) su calidad ; (*b*) el número de secuencias ; y (*c*) el nivel de divergencia.

Aunque el cálculo de MI usualmente se realiza entre pares de residuos, el fenómeno de coevolución se da entre grupos de residuos que varían de manera conjunta. Un residuo dado puede tener alta MI con diferentes residuos, definiendo verdaderas redes de coevolución. A modo de ejemplo se presenta en la Figura 4 una red de MI para una familia de enzimas de sialiltransferasas. En la red de coevolución cada nodo representa un aminoácido y se trazan lineas entre los nodos cuando el puntaje de MI es significativo.

Figura 4: Red de MI para una familia de sialiltransferasas

Aunque la MI es calculada entre pares de residuos, los residuos pueden coevolucionar de grupos formando redes de coevolución. La red de MI se define como un grafo donde cada nodos representa un residuo en una posición particular del MSA y se trazan lineas entre los nodos cuando el puntaje de MI es significativo. La longitud de las lineas es relativa al valor de MI entre los nodos, las lineas más cortas representan un puntaje de MI mayor y las más largas un valor de MI menor. Los nodos están coloreados según su conservación, en escala de rojo (mayor) a azul (menor). Imagen generada con el servidor MISTIC Simonetti et al. [115].

1.2.2 *Correcciones al algoritmo de Información Mutua*

Dos ejemplos de correcciones para afrontar el problema de la señal filogenética y la entropía inherente a cada columna de un MSA son, el *row column weighting* (RCW) Gouveia-Oliveira and Pedersen [49] y el Average Product Correction (APC) Dunn et al. [38]. El método RCW divide el valor de MI entre dos posiciones por el promedio de

MI que tienen esas posiciones con cualquier otra, como se muestra en la ecuación 3.

$$RCW = \frac{MI_{xy}}{(\overline{MI}_{x\bullet} \cdot \overline{MI}_{\bullet y}) \div 2} \tag{3}$$

Donde $\overline{MI}_{x\bullet}$es el valor promedio de MI entre la posición x y el resto de las posiciones ($\bullet$) en el MSA, análogamente se define $\overline{MI}_{\bullet y}$.

El método Average Product Correction sustrae el término definido como APC para generar un puntaje de MI corregida como se muestra en las ecuaciones 4 y 5.

$$MI_{APC} = MI_{(x,y)} - APC_{(x,y)} \tag{4}$$

$$APC_{(x,y)} = \frac{MI_{x\bullet} \cdot MI_{\bullet y}}{\overline{MI}_{\bullet\bullet}} \tag{5}$$

Donde $\overline{MI}_{\bullet\bullet}$ es el promedio del valor de MI sobre todos los pares de posiciones del MSA.

En el trabajo de Buslje et al. [16] se ha comparado el desempeño predictivo de ambas correcciones, RCW y APC en dos set de datos comprensivos, uno compuesto por secuencias biológicas y otro construido con datos artificiales por Gouveia-Oliveira and Pedersen [49]. Se ha demostrado que el método con mejor desempeño incluye la corrección APC. Además se han propuesto dos nuevas correcciones que mejoran aún más el desempeño del método: el pesado de secuencias y la corrección por bajo numero de observaciones (pseudo-cuentas). A la vez se introduce una permutación de secuencias para el cálculo de MI que permite cambiar de escala el puntaje de MI y hacerlo comparable entre diferentes familias de proteínas. El algoritmo de MI utilizado en el presente trabajo proviene del desarrollado en Buslje et al. [16], estas nuevas correcciones introducidas se explican a continuación.

Pesado de secuencias

Al trabajar con datos biológicos, frecuentemente los MSAs presentan un sesgo no natural y redundancia de secuencias, por ejemplo, debido a la secuenciación en múltiples cepas o a la selección de especies a secuenciar. Por lo tanto es esperable que un algoritmo de clusterización mejore el cálculo de MI. Se empleó el algoritmo Hobohm 1 Hobohm et al. [59] para definir grupos de secuencias. Los grupos se definieron para un valor de identidad de secuencia$\geqslant$62 % Shackelford and Karplus [111]. Luego se le asignó un peso a cada secuencia dentro de un grupo correspondiente a la inversa del número de secuencias del grupo al que pertenece.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

Corrección por pseudo-cuentas

Para MSAs con bajo número de secuencias el cálculo de la ocurrencia de aminoácidos en una posición dada es estimado a partir de un bajo número de observaciones, y su contribución a la MI puede ser muy ruidosa. Por otro lado, una combinación (un par) de aminoácidos que no se observa en un MSA puede deberse meramente a que el muestreo no fue suficiente.

Por ello se introdujo una corrección para el bajo número de observaciones. La probabilidad de los aminoácidos, $P\left(a_x, b_y\right)$, se calcula a partir de $N(a,b)$, el número de veces que el par (a,b) es observado en las posiciones x e y en el MSA más una constante λ. La frecuencia del aminoácido a en la posición x y del aminoácido b en la posición y se calcula como se muestra en la ecuación 6

$$P(a_x,b_y) = \frac{\lambda + N(a_x, b_y)}{N} \tag{6}$$

Donde

$$N = \sum_{a,b}(\lambda + N(a_x, b_y))$$

$$P(a_x) = \sum_{b} P(a_x, b_y)$$

$$P(b_y) = \sum_{a} P(a_x, b_y)$$

Se le da un valor inicial $N(a_x, b_y) = \lambda$ para todo par de aminoácidos. Es decir, todo par posible de aminoácidos será observado al menos λ veces. Solamente para MSAs con un bajo número de secuencias, donde una gran fracción de pares de aminoácidos no son observados, el parámetro λ influencia el cálculo de probabilidades. Contrariamente, para MSAs con un gran número de secuencias, la mayoría de los pares de aminoácidos serán observados al menos una vez, y la influencia de λ será menor. Se evaluó el desempeño del método a diferentes valores de λ en el rango [0-2] con pasos de 0.01 y se obtuvo el mejor desempeño con $\lambda = 0{,}05$, encontrándose resultados similares en el rango 0.025-0.075. El valor óptimo se encontró de manera consistente siendo independiente del tamaño del set de datos usado, el modelo evolutivo y la tasa de evolución.

Transformación a Z-score

Cada valor de MI entre un par dado de posiciones es comparado con la distribución de valores de MI obtenidos a partir de un grupo de MSAs aleatorizados. El Z-score es calculado como el número de desviaciones estándar que el valor de MI observado se aparta de la media obtenida con los MSAs aleatorios. Para cada MSA se realizaron

100 permutaciones, manteniendo los gaps fijos en sus posiciones originales. Se ensayaron dos métodos de permutación, uno basado en columnas (aleatorización vertical) y otro basado en secuencia (aleatorización horizontal). El primero desafía la hipótesis de que las secuencias son homólogas y están correctamente alineadas, pero que las columnas no están correlacionadas. Mientras que el segundo método desafía la hipótesis de que las secuencias no son homólogas. El mejor resultado fue obtenido con el segundo método de permutación, por lo que el Z-score no testea de manera adecuada la hipótesis nula (que las columnas no sean correlacionadas) por lo que debe ser interpretado solamente como un puntaje predictivo más, que permite la comparación entre familias.

1.2.3 *Algoritmo de MI utilizado*

En el presente trabajo se utilizará como medida de coevolución la MI con las correcciones APC, el pesado de secuencias por clusterización, la corrección por pseudo-cuentas y la transformación a Z-score como se describe en Buslje et al. [16], donde también se establece como valor de Z-score significativo si el puntaje $\geqq$ 6,5. Además se evaluó la dependencia del desempeño del método con el número de secuencias del MSA, se encontró que los MSAs con menos de 400 secuencias únicas (clusters al 62 % de identidad), tienden a mostrar un desempeño predictivo bajo, AUC $\leq$ 0,75. De aquí en adelante se hace referencia al puntaje de Z-score obtenido con las correcciones mencionadas, como puntaje de MI.

1.2.4 *Puntajes derivados de MI: cMI y pMI*

El cálculo de MI genera un puntaje para cada par de residuo en el MSA. Se buscó generar un puntaje por residuos que caracterice la cantidad de información mutua en su proximidad física. Este puntaje se definió en dos pasos. Primero se calculó un puntaje de información mutua acumulada (cMI por Cumulative MI) para cada residuos como la suma de los valores de MI mayores a un valor umbral en el que el residuo en particular participa, como se muestra en la ecuación 7.

$$cMI_i = \sum_{j, MI(i,j) > u} MI_{(i,j)} \tag{7}$$

Este puntaje de cMI captura la información de hasta qué punto un determinado amino ácido participa en una red de información mutua. Luego, a partir del puntaje de cMI, se calculó la MI en la proximidad para cada residuo como el promedio de la cMI de todos los residuos que se encuentran dentro de una distancia física determinada. Finalmente se normalizaron los valores de proximidad para cada MSA para que los puntajes pertenezcan al rango $[0 - 1]$, al que llamamos

puntaje por proximidad pMI, como se muestra en la ecuación 8 y se esquematiza en la Figura 5 .

$$pMI_i = \frac{1}{N} \cdot \sum_{j, d_{ij} < u} cMI_{(j)} \tag{8}$$

Donde la suma se realiza sobre todos los residuos j en una proteína dada, dentro de una distancia $d_{ij} < u$. La distancia d_{ij} se calculó como la distancia mínima entre cualquier par de átomos diferentes de H entre los residuos i y j , cMI_j es la información mutua acumulada del residuos j, y u es la distancia umbral utilizada.

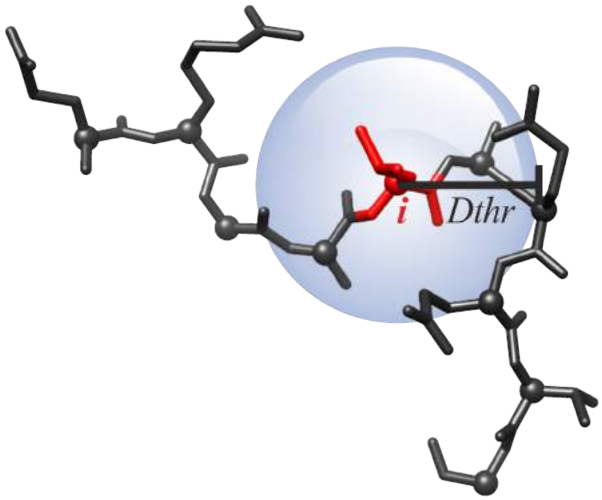


Figura 5: Puntaje de MI en la proximidad
Para cada residuo i se calcula la pMI como la sumatoria de las cMI de los residuos que se encuentran a una distancia menor que una Distancia de corte (Dthr por Distance threshold).

1.2.5 *Evaluación del desempeño de un método predictivo*

La curva ROC (Receiver Operating Characteristic) ilustra el desempeño de un clasificador binario como su discriminación variando los valores de cortes. La curva ROC se genera graficando la tasa de verdaderos positivos (TPR: True Positive Rate) contra la tasa de falsos positivos (FPR: false positive rate) variando el puntaje del predictor. Las curvas que se encuentran por encima de la diagonal representan métodos con cierto poder de discriminación, teniendo mayor poder cuando se acercan a la esquina superior izquierda del gráfico. Usualmente se calcula el área bajo la curva ROC (AUC) como un cuantificador global del desempeño del predictor, un clasificador aleatorio da un valor de AUC de 0.5, mientras que un clasificador perfecto corresponde a un AUC de 1.0, como se muestra en la Figura 6.

$$TPR = Sensibilidad = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{9}$$

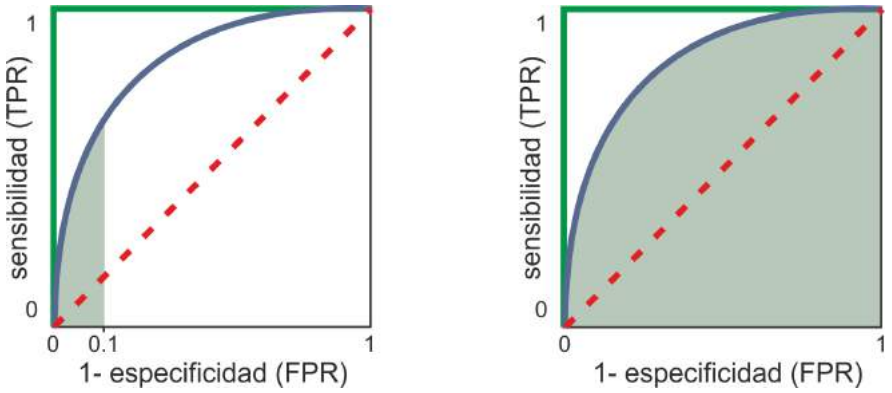


Figura 6: Curvas ROC

Las lineas punteadas rojas representan la curva ROC para un predictor aleatorio. La curva verde representa un método predictivo perfecto, donde los valores de sensibilidad y especificidad son máximos. La curva azul representa un ejemplo de un método de clasificación real. El área verde debajo de la curva representa el AUC. En la imagen de la izquierda AUC01: es el área resultante de la integración para una tasa de falsos positivos del 0.1 (10 %FPR). A la derecha se representa el AUC completo.

$$FPR = 1 - Especificidad = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{10}$$

Donde TP , FP, TN y FN corresponden a verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos respectivamente, de acuerdo a un valor de corte determinado. El área bajo la curva ROC es utilizada para medir la habilidad global de un predictor para distinguir los positivos de los negativos. En la Figura 6 se grafican dos curvas ROC distinguiendo la diferencia entre AUC y AUC01.

2

RESIDUOS CATALÍTICOS PRESENTAN UNA RED DE RESIDUOS CON ALTA MI.

IMPLICANCIAS PARA SU DETECCIÓN

2.1 RESUMEN

Los residuos catalíticos presentan una alta conservación y se encuentran localizados en el sitio funcional de las enzimas para llevar a cabo su función. Sin embargo muchos residuos que no son catalíticos también presentan alto grado de conservación e inversamente no todos los residuos catalíticos están altamente conservados en un familia determinada, haciendo de esta manera que sea difícil su detección.

La hipótesis de este trabajo es que los residuos catalíticos presentan una firma definida por una red de residuos con alta información mutua en su proximidad espacial, y que esta firma serviría para predecir los residuos catalíticos.

Para testear esta hipótesis se usaron 434 familias de enzimas de la base de datos Pfam que poseen anotación de sus residuos catalíticos en la base de datos catalytic site atlas. Corroboramos la hipótesis propuesta y demostramos que la MI puede complementar la medida de conservación para la detección de residuos catalíticos, concluyendo con el desarrollando de un método con alto poder predictivo.

2.2 INTRODUCCIÓN

Los residuos catalíticos (RC) juegan un rol fundamental en las enzimas por lo que se encuentran bajo una fuerte presión de selección, presentando generalmente un alto grado de conservación.

Diferentes métodos computacionales han sido desarrollados para la detección de RC basados en diferentes principios. Algunos de ellos están basados en características fisicoquímicas de los aminoácidos Bartlett et al. [9], densidad de grupos funcionales conservados Innis et al. [61], análisis de secuencia (conservación, identificación de patrones, conservación de bloques a lo largo de la secuencia y evolución, entre otros) Zhang et al. [134], Chien et al. [25], Erdin et al. [43], Mihalek et al. [86], Manning et al. [82], Sterner et al. [117], propiedades de secuencia y de estructura Petrova and Wu [96], Bernardes et al. [10], Cilia and Passerini [27], evolución e información de estructura tridimensional Kristensen et al. [72], Sankararaman and Sjolander [109], Sankararaman et al. [110], Ward et al. [128], redes neuronales

Tang et al. [120] y combinación de los diferentes métodos mencionados Alterovitz et al. [2].

La conservación es una manera sencilla e intuitiva de predecir residuos funcionalmente importantes en proteínas. Sin embargo, muchos residuos que no son catalíticos están altamente conservados e inversamente, no todos los residuos catalíticos están altamente conservados dentro de una familia de proteínas dada. Por otro lado se ha demostrado que los residuos involucrados en redes de coevolución son funcionalmente importantes Lee et al. [74], Kuipers et al. [73], Gloor et al. [48] y varios estudios dieron evidencia que son importantes para la especificidad y la regulación alostérica Lockless and Ranganathan [81], Shi et al. [114], Chakrabarti and Panchenko [23].

El ambiente estructural y fisicoquímico de un sitio activo debe preservarse para que la proteína mantenga su función a lo largo de la evolución. Esto impone limitaciones en la diversidad de amino ácidos que se encuentran en la cercanía del sitio activo y esta presión de selección podría llevar a una coevolución de los aminoácidos.

La hipótesis de este trabajo es que los residuos catalíticos llevan una señal definida por una red de residuos que coevolucionan en su proximidad espacial.

Algunos autores han sugerido una relación entre sitios funcionalmente importantes y residuos vecinos que coevolucionan Rausell et al. [103], Chakrabarti and Panchenko [24], Little and Chen [78], aunque hasta el presente no han mostrado cómo la presencia de residuos que coevolucionan puede proveer información cuantitativa para la identificación de residuos catalíticos.

En este trabajo se realizó un estudio a gran escala con el objetivo de desafiar la hipótesis propuesta. Introducimos dos nuevos puntajes derivados de la MI: cMI y pMI. El puntaje cMI que mide el grado de MI compartida por un residuo dado considerando los valores de MI a partir de cierto umbral. Por otro lado, el puntaje pMI da cuenta de las redes de alta MI en la proximidad estructural de un residuo particular, dentro de un cierto umbral de distancia. Estos puntajes fueron introducidos en la sección 1.2.4 y los parámetros utilizados para su cálculo se indican más adelante en Materiales y Métodos sección 2.3.5. De una manera análoga calculamos un valor de conservación en la proximidad de un residuo dado, llamado pC (por proximity Conservation). Luego analizamos el desempeño predictivo de cada puntaje y finalmente, integramos los diferente puntajes derivados de la información mutua y conservación para crear un método capaz de predecir residuos catalíticos.

Es importante notar que el objetivo de este trabajo no es desarrollar un método de detección de catalíticos que supere a los existente, sino demostrar la existencia de una red de información mutua en la proximidad de los residuos catalíticos, y además demostrar que ésta

información puede ser usada de manera complementaria a las medidas de conservación y análisis de secuencia convencionales.

2.3 MATERIALES Y MÉTODOS

2.3.1 *Set de datos*

El set de datos fue construido basado en la base de datos Catalytic Site Atlas (CSA, versión 2.2.11 publicada en Agosto del 2009) 99. Esta base de datos provee anotaciones de los residuos catalíticos para enzimas depositadas en PDB. Los residuos catalíticos fueron definidos por estar directamente involucrados en algún aspecto de la reacción catalizada por la enzima. La base de datos posee dos tipos de anotaciones: (i) originales, anotaciones realizadas manualmente y (ii) por homología, que contiene anotaciones inferidas por homología, a partir de una búsqueda Psi-BLAST y alineamiento con una entrada original de CSA. En este trabajo se utilizaron solamente las anotaciones originales que comprenden 968 entradas, pertenecientes a enzimas contenidas en 455 familias diferentes de Pfam. Debido a inconsistencias en algunas entradas entre las bases de datos CSA y PDB, algunas familias fueron excluidas del análisis. El set de datos resultante se compone de 434 familias de Pfam cada una conteniendo al menos una proteína de estructura conocida (entrada en PDB) y anotación original en CSA. Este set de datos contiene en total 1212 residuos catalíticos anotados en CSA. Para 9 de las 434 familias el PDB elegido como referencia fue obtenido por RMN, en estos casos se utilizó el primer modelo para representar la estructura. Las 434 familias incluidas en el análisis cubren un amplio rango de plegamientos, siendo en SCOP Murzin et al. [89] 8 clases, 199 plegamientos, 249 superfamilias y 389 familias.

Cuando se encontraba disponible más de una entrada de PDB con anotación en CSA para una misma familia de Pfam, la estructura de referencia fue seleccionada según el siguiente criterio: secuencia con mayor cobertura del MSA de Pfam; año de determinación de la estructura (preferentemente posterior al 2000) y resolución. Los MSA fueron pre-tratados eliminando los gaps en la secuencia de referencia. Además fueron excluidos del análisis las posiciones con >50 % de gaps y las secuencias con una longitud menor al 50 % de la secuencia de referencia, como se describe en Buslje et al. [16].

2.3.2 *Conservación*

La conservación para cada posición del MSA fue calculada con tres medidas diferentes: entropía de Shannon Shannon [112], entropía relativa Kullback-Leibler (KL) Cover and Thomas [30] utilizando la distribución de frecuencia background obtenida de la base de datos Uni-

Prot UniProt Consortium [125] y Frecuencia máxima (frecuencia del aminoácido más representado). Cada una de estas medidas fueron calculadas en (i) MSA crudo, (ii) MSA con redundancia de secuencias corregido usando pesado de secuencia mediante una clusterización al 62 % de identidad, (iii) MSA incluyendo corrección por pseudo-cuentas y (iiii) MSA aplicando clusterización y pseudo-cuentas. De esta manera, fueron 12 el número total de medidas de conservación utilizadas.

2.3.3 *Información Mutua*

La Información Mutua fue calculada como se describe en Buslje et al. [16] y se explica en la Introducción sección 1.2.3.

2.3.4 *Puntaje de Verosimilitud de Catalítico*

Se definió el puntaje de Versosimilitud de catalítico (Cls) como la suma pesada de la conservación definida en términos de la entropía relativa de KL, la información mutua en la proximidad (pMI) y la conservación en la proximidad (pC), como se indica en la ecuación 11

$$Vc = (1 - w_{MI} - w_C) \cdot KL + w_{MI} \cdot pMI + w_C \cdot pC \tag{11}$$

Donde w_C y w_{MI} son pesos relativos ajustables.

2.3.5 *Optimización de parámetros*

El cálculo del puntaje Cls depende de tres parámetros y de dos pesos relativos:

Z_{thr} : Valor umbral del Z-score de MI para incluir un par de aminoácidos en el puntaje cMI

D_{MI}: Valor umbral de distancia para incluir un aminoácido en el puntaje pMI

D_C: Valor umbral de distancia para incluir un aminoácido en el puntaje pC

w_{MI}: Peso relativo del puntaje pMI

w_C: Peso relativo del puntaje pC

Estos parámetros fueron estimados usando validación cruzada de 5 iteraciones (five-fold cross validation), donde los valores óptimos fueron obtenidos usando un muestreo de fuerza bruta en 4/5 del set de datos para optimizar el valor promedio de AUC. Mientras que los datos restantes, 1/5 del set de datos, fueron evaluados utilizando el grupo de parámetros óptimos calculados. Este procedimiento fue repetido 5 veces generando 5 juegos de parámetros óptimos y valores de evaluación del desempeño para cada MSA del set de datos.

2.3.6 *Medición del desempeño predictivo*

Para los puntajes de conservación, pMI y Cls se evaluó el desempeño predictivo en la detección de residuos catalíticos en términos del área bajo la curva ROC (AUC) por familia. La medida de AUC puede no ser óptima si el set de datos tiene una proporción alta de negativos y una alta especificidad puede traducirse en un gran número de falsos positivos. En este trabajo se complementó la medida de AUC con AUC01, calculada como el área bajo la curva ROC integrada desde una tasa de falsos positivos de 0 hasta una tasa de falsos positivos de 0.1 dividido por 0.1 (el área máxima para esa fracción de la curva). Para ambas medidas AUC y AUC01 un valor de 1.0 indica una predicción perfecta, mientras que una predicción aleatoria da una valor de 0.5 y 0.05 para AUC y AUC01 respectivamente. Se tomaron los residuos anotados como catalíticos de la base de datos CSA como el set positivo, y todos los demás residuos con coordenadas en PDB fueron asignados como negativos. El desempeño final fue determinado como el promedio de AUC sobre las 434 familias de Pfam.

2.4 RESULTADOS

2.4.1 *Conservación de secuencia*

En este trabajo se investigaron tres medidas de conservación en cuatro condiciones distintas, generando en total doce puntuaciones de conservación. Las medidas de conservación investigadas fueron: Kullback-Leibler, Entropía de Shannon y Máxima Frecuencia; en las siguientes condiciones:

A. MSA crudo

B. Incluyendo clusterización y pesado de secuencias

C. Corrección por pseudo-cuentas

D. Incluyendo clusterización con pesado de secuencias y pseudo-cuentas.

El desempeño predictivo de RC para las doce puntuaciones de conservación se muestra en la tabla 1.

La medida de conservación con mejor desempeño predictivo en términos de AUC fue Kullback-Leibler crudo, con un promedio de AUC calculado sobre todas las familias de 0.892 y AUC01 de 0.485 (el cálculo crudo se refiere a la exclusión del pesado por secuencias y la corrección por pseudo cuentas). En términos de AUC01 la inclusión del pesado de secuencias mejoró el desempeño predictivo en las tres medidas de conservación. La medida Máxima Frecuencia tuvo un desempeño significativamente peor que las dos medidas basadas

Medida de Conservación	Max-Freq		Shannon		Kullback-Leibler	
	AUC	AUC01	AUC	AUC01	AUC	AUC01
Crudo	0.874	0.458	0.880	0.464	**0.892**	0.485
Clusterización	0.870	0.461	0.876	0.465	0.890	**0.502**
Pseudo-cuentas	0.857	0.380	0.852	0.371	0.877	0.437
Clusterización y Pseudo-cuentas	0.847	0.353	0.837	0.335	0.868	0.411

Tabla 1: Desempeño predictivo de las medidas de conservación en cuatro condiciones distintas

El desempeño predictivo se reporta como promedio de AUC sobre las 434 familias de Pfam, para las medidas de conservación Máxima Frecuencia (Max-Freq), Shannon y Kullback-Leibler. Se resalta en negrita el valor más alto de desempeño para AUC y AUC01.

en información ($p< 0.0001$, test binomial excluyendo extremos). Aunque el desempeño es similar entre Shannon y Kullback-Leibler para los puntajes crudos, la diferencia es significativa ($p<0.005$, test binomial excluyendo extremos). La diferencia entre Kullback-Leibler (KL) crudo y con pesado de secuencias es marginalmente significativa con un p-valor de 0.05 a favor de KL crudo si se considera AUC y de KL con pesado de secuencias si se considera AUC01. Para mantener la simplicidad en los análisis subsiguientes se emplea el puntaje de Kullback-Leibler crudo como medida de conservación.

También se analizó el grado de dependencia del desempeño predictivo de la medida KL con la cantidad de secuencias del MSA utilizado. La Figura 7 muestra que son necesarias al menos 10 secuencias para obtener una predicción confiable utilizando KL. La diferencia en el desempeño predictivo entre las familias con menos de 10 secuencias y las familias con más de 10 secuencias, es altamente significativo ($p < 0.001$, t-test).

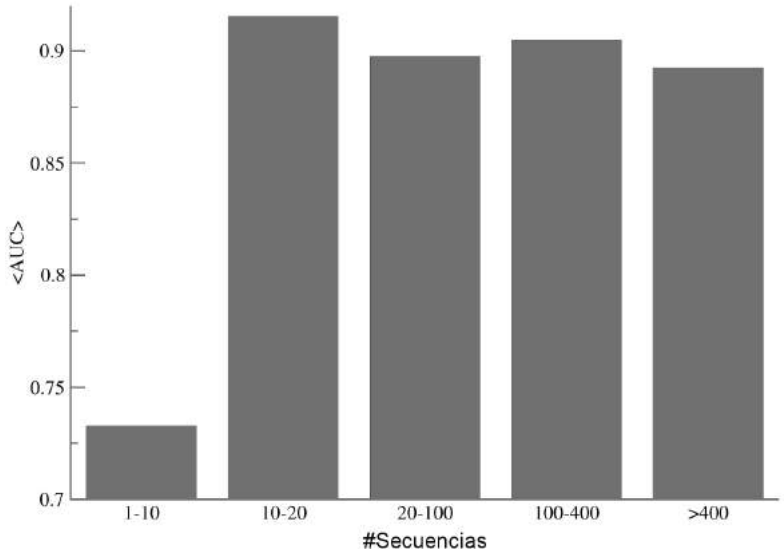


Figura 7: Histograma del desempeño predictivo del puntaje de conservación KL crudo en función del número de secuencias del MSA

El número de familias de Pfam en cada intervalo es de 9, 9, 36, 66 y 314 respectivamente.

2.4.2 *Información Mutua acumulada en la proximidad de los residuos catalíticos*

Para analizar la MI en el entorno de los RC, se generaron dos nuevas puntuaciones llamadas cMI y pMI con la intención de capturar la información mutua al rededor del residuo. Primeramente se notó que los residuos cercanos en la estructura tridimensional a los residuos catalíticos tienden a tener altos puntajes de cMI. Por ello se generó un puntaje para medir la MI en la proximidad de un residuo dado (pMI), que representa la cMI que se encuentra en la proximidad, dentro de un umbral de distancia determinado. Se observó que los residuos catalíticos presentan mayor pMI que otros residuos conservados.

Esta observación fue estudiada utilizando 434 familias de enzimas de la base de datos Pfam y calculando el desempeño de la medida de pMI como predictor de residuos catalíticos. Se definió un valor de corte de 7.5Å para definir la proximidad estructural de un residuo, y un valor umbral de Z-score de 6.0 para definir la MI confiable. El valor promedio del desempeño predictivo de pMI en la predicción de residuos catalíticos para las 434 familias, medido en términos de AUC y AUC01 fue de 0.843 y 0.342 respectivamente, ambos valores difieren significativamente del valor random ($p<0.0001$, test binomial excluyendo extremos). El número de residuos próximos puede variar para cada aminoácido dependiendo de su ubicación relativa en la estructura tridimensional, por ejemplo si un residuo se encuentra en el core de la proteína tiene mayor número de residuos en su proximidad espacial frente a un residuo ubicado en la superficie. Para evitar un

sesgo debido a la localización, el valor de pMI fue normalizado por el número de interacciones.

2.4.3 *Conservación en la proximidad de un residuo*

Debido a que el sitio activo puede estar compuesto por varios residuos catalíticos cercanos en la estructura tridimensional, se exploró la posibilidad de que un puntaje de proximidad basado en conservación sea un buen predictor de residuos catalíticos. Se utilizó el mismo valor umbral de distancia que para la medida pMI. Se encontró que el puntaje de conservación en la proximidad (pC) presenta un promedio de desempeño predictivo en AUC y AUC01 de 0.854 y 0.379 respectivamente. Estos valores son mayores que los obtenidos con pMI, pero tanto para AUC como para AUC01 la diferencia entre pC y pMI no es estadísticamente significativa ($p<0.05$, test binomial excluyendo extremos).

2.4.4 *Puntaje combinado de verosimilitud de catalítico (Cls)*

Se generó un puntaje de verosimilitud de catalítico combinado (Cls por Catalytic Likeliness Score) para identificar los residuos catalíticos. Este puntaje se calculó como la suma pesada de los puntajes de conservación de Kullback-Leibler, y los puntajes de proximidad pMI y pC. Los parámetros óptimos de los puntajes fueron identificados usando una validación cruzada de 5 veces (five-fold cross validation) como se describe en Materiales y Métodos. Los valores óptimos encontrados para los parámetros Z_{thr}, D_{MI}, D_c, w_{MI} y w_c se muestran en la tabla 2. Los valores bajos de desviación estándar para cada uno de los parámetros estimados indica que la optimización es robusta en los diferentes set de datos de validación cruzada. El promedio de desempeño para el puntaje Cls en la detección de catalíticos en términos de AUC y AUC01 es de 0.927 y 0.594 respectivamente. Este desempeño es significativamente mayor que el obtenido con los puntajes de conservación KL, pMI y pC individualmente ($p<0.001$ para cada caso usando test binomial excluyendo extremos).

Para investigar la contribución individual de los puntajes de proximidad en el desempeño del puntaje Cls, se buscaron los parámetros óptimos para un puntaje combinado incluyendo solo uno de los dos puntajes de proximidad en combinación con el puntaje de conservación KL. Los parámetro óptimos fueron estimados usando validación cruzada de 5 veces, los resultado se muestran en la tabla (2).

El cálculo de valores de MI confiables requiere que el número de grupos de secuencias, agrupadas al 62 % de identidad sea $\geq$ 400 Buslje et al. [16], por este motivo se analizó el sub grupo de 172 familias que cumplían con este criterio. En este sub grupo el puntaje combinado que incluye la pMI, alcanza un desempeño de AUC=0.920 y

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

Método	KL+pMI	KL+pC
Parámetros	$w_{MI} = 0{,}80 \pm 0{,}0$	$w_c = 0{,}6 \pm 0{,}0$
	$D_{MI} = 7{,}9 \pm 0{,}2$	$D_c = 8{,}0 \pm 0{,}0$
	$Z_{thr} = 5{,}5 \pm 0{,}32$	
AUC	0,922	0,910
AUC01	0,514	0,562

Tabla 2: Parámetros óptimos y promedio del desempeño predictivo para la combinación de conservación con un puntaje de proximidad

El método KL+pMI combina el puntaje de conservación KL con la medida de información mutua pMI. KL+pC, es la combinación del puntaje de conservación KL, con la medida de conservación en la proximidad pC. w_{MI} es el peso relativo de pMI, D_{MI} es el valor umbral de distancia de proximidad para pMI, Z_{thr} es el valor umbral del Z-score de MI, w_c es el peso relativo de pC y D_C es el valor umbral de distancia de pC.

AUC01=0.597. Estos valores superan significativamente el desempeño del método combinado con la Conservación en la proximidad, pC, con un AUC=0.889 y AUC01=0.559 ($p<0.05$, test binomial excluyendo extremos). Esto indica que la pMI aporta información que no está contenida en el puntaje de conservación.

2.4.5 *Desempeño predictivo en la familia PF00890 como caso de estudio*

Para ilustrar las contribuciones de las diferentes medidas, se muestra en la Figura 8 el rol de las cuatro medidas de predicción: KL, pMI, pC y Cls, para la familia PF00890 representada por la enzima Fumarato reductasa de *Shewanella putrefaciens* MR-1 (código PDB= 1D4C). Esta familia fue elegida del grupo de las 172 familias que contienen al menos 400 clusters de secuencias.

La actividad catalítica de la fumarato reductasa es llevada a cabo por los residuos His364, Arg401, His503 y Arg544 Leys et al. [76]. Se puede observar que el puntaje de conservación KL de los residuos catalíticos es relativamente bajo Figura 8A, mientras que para pC y pMI los puntajes son altos en la proximidad de los RC (Figura 8B y C). Comparando estas dos figuras es evidente que las dos medidas de proximidad aportan información diferente al puntaje combinado Cls. Finalmente el resultado obtenido con el puntaje Cls se muestra en la Figura 8D. Los valores de AUC para las cuatro medidas de predicción de la Figura 8 son 0.92, 0.94, 0.98 y 0.99 respectivamente. Estos valores traducidos en números de falsos positivos predichos a 100 % de sensibilidad son 47, 39, 15 y 4 respectivamente. Es decir,

que estos valores corresponden el número de residuos no catalíticos que obtuvieron un puntaje predictivo mayor a los puntajes de los residuos catalíticos. Este resultado ejemplifica el alto poder predictivo del puntaje Cls para identificar residuos catalíticos y disminuir el número de falsos positivos.

La ganancia en el desempeño predictivo para detectar residuos catalíticos es consistente en las familias independientemente del nivel de conservación de los RC, sin embargo la mayor ganancia en desempeño predictivo al incluir pMI se encuentra en familias con baja conservación de los catalíticos. Si se toman por ejemplo las 217 familias de Pfam con menor desempeño predictivo para el puntaje de conservación KL y nos preguntamos para cuántas de estas familias el puntaje pMI mejoran la predicción, encontramos que este número es significativamente mayor que el correspondiente número para el grupo de 217 familias con mayor desempeño predictivo para KL (p-value<0.001, test binomial excluyendo extremos). Esta diferencia en la mejora del desempeño entre los dos sub-set de familias no se debe a diferencias en la cantidad de datos, ya que el promedio del tamaño de las familias es comparable (p>0.1, t-test). El entorno catalítico de un sitio activo necesita estar conservado dentro de una familia de proteínas para mantener su función, y se podría especular que cuando la conservación del RC es débil, el entorno catalítico se mantiene en gran medida por coevolución.

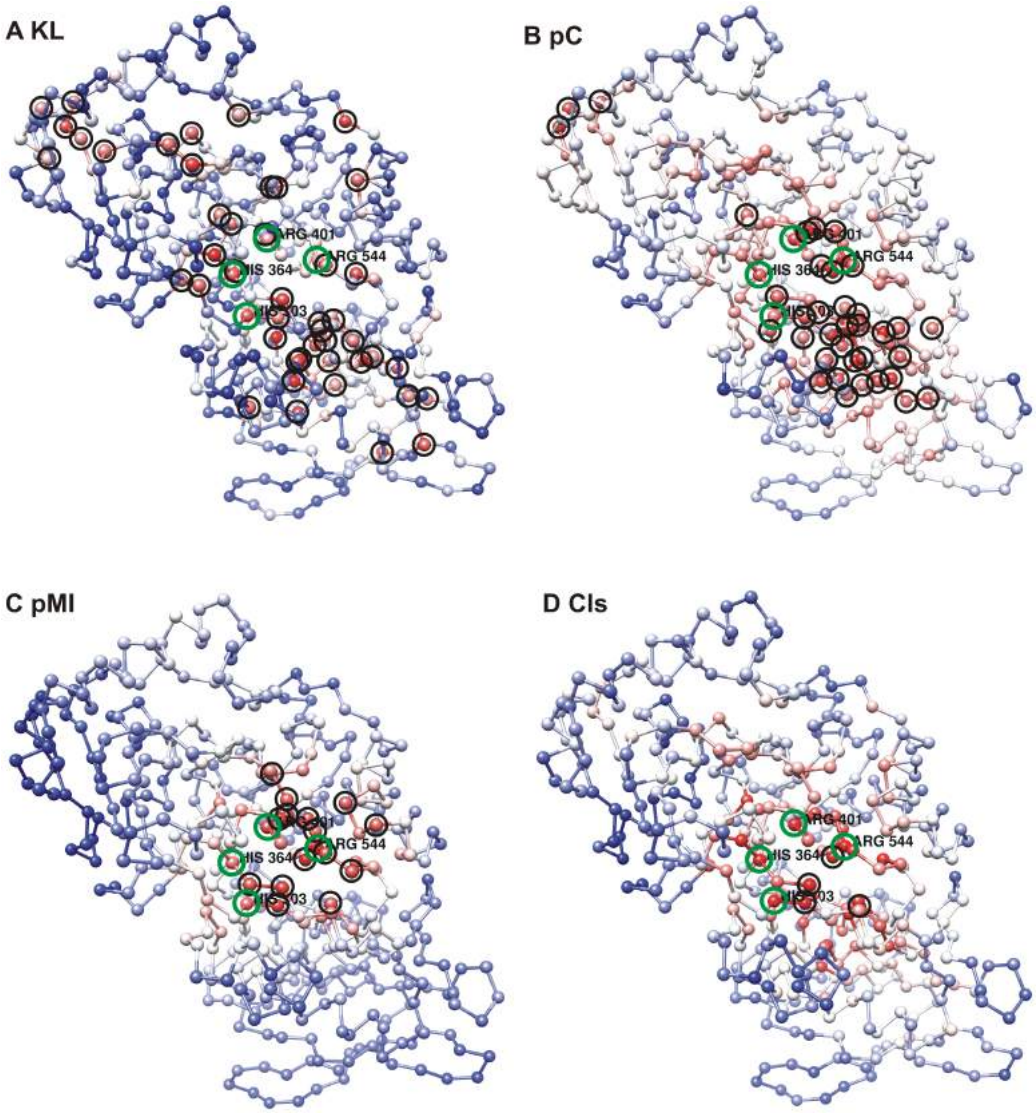


Figura 8: Identificación de residuos catalíticos utilizando cuatro puntajes de predicción distintos

Se muestra en representación de Cα el PDB:1D4C perteneciente a la familia de Pfam PF00890.Los residuos catalíticos se encuentran encerrados en círculos verdes. Los cuatro puntajes de predicción son A) Conservación KL, B) Conservación en la proximidad, C) MI en la proximidad y D) Verosimilitud de catalítico. Se señalan con círculos negros los residuos falsos positivos, que alcanzan los valores de : 47, 39, 15 y 4 respectivamente. Los puntajes de predicción se representan en escala del azul al rojo (azul: menor; rojo: mayor).

2.4.6 *Sensibilidad de los diferentes puntajes para predecir residuos catalíticos*

También examinamos la sensibilidad de los diferentes métodos a diferentes valores de corte de especificidad, los resultados se muestran en la tabla 3. Los valores obtenidos confirman la ganancia predictiva

	Sensibilidad					
Especificidad	KL	pMI	pC	KL+pMI	KL+pC	Cls
0.99	0.222	0.122	0.159	0.300	0.282	**0.315**
0.95	0.544	0.375	0.423	0.646	0.637	**0.667**
0.90	0.716	0.560	0.604	0.802	0.774	**0.816**
0.85	0.798	0.666	0.703	0.861	0.835	**0.862**

Tabla 3: Sensibilidad de los métodos de detección de residuos catalíticos a diferentes valores de especificidad

La sensibilidad fue calculada como promedio sobre las 434 familias de Pfam a diferentes valores de corte de especificidad. Se resaltan en negrita los mejores valores obtenidos para cada especificidad.

sobre el set de datos completos, debida a la inclusión del puntaje pMI en el puntaje combinado Cls. Para los diferentes valores de corte de especificidad Cls presenta la sensibilidad más alta, siendo la diferencia con todos los otros métodos estadísticamente significativa ($p<0.05$, test binomial excluyendo extremos). El valor umbral de Cls para las 434 familias y una especificidad de 0.90, es de $1{,}44 \pm 0{,}26$. La baja desviación estándar obtenida indica que el puntaje Cls es estable para el set de datos y sugiere que el método puede ser utilizado para cualquier familia de enzimas, independientemente de su plegamiento, requiriendo un MSA con un número de secuencias mayor a 10 (ver Figura 7).

2.5 DISCUSIÓN

Es esperable que los residuos catalíticos se encuentren en general conservados y localizados en el sitio funcional de la proteína para llevar a cabo su función. Sin embargo, muchos residuos que no son catalíticos se encuentran altamente conservados e inversamente no todos los residuos catalíticos se encuentran altamente conservados dentro de una familia proteica dada. El requerimiento de mantener una determinada función catalítica durante el curso de la evolución impone grandes limitaciones en la diversidad de aminoácidos del entorno estructural del sitio activo. En este trabajo desafiamos la hipótesis que los residuos catalíticos tienen una firma particular definida por una red de residuos cercanos espacialmente que comparten alta MI, y comprobamos la utilidad de esta firma para la identificación de los residuos catalíticos.

La hipótesis fue corroborada usando un set de datos de 434 familias de Pfam, donde cada una de ellas posee al menos una estructura

tridimensional conocida y uno o más residuos catalíticos asignados por la base de datos CSA [99].

El trabajo demuestra en términos cuantitativos directos, a través de la ganancia en el desempeño predictivo, la contribución de la señal de coevolución en la detección de residuos catalíticos.

Adicionalmente, se ha sugerido que en la cercanía de los residuos catalíticos se encuentran un tipo particular de residuo que le otorga especificidad en la reacción catalizada, estos residuos son llamados posiciones determinantes de especificidad (SDP por Specificity Determining Position) Capra and Singh [19], Ubersax and Ferrell Jr [124], Kalinina [64]. Estos residuos podrían definir la especificidad del sustrato o el tipo de reacción catalizada, aunque también pueden localizarse lejos del sitio activo, definiendo especificidad de interacción con otra molécula o involucrados en mecanismos control alostérico.

Es esperable que las posiciones determinantes de especificidad también posean cierta conservación dentro de una subfamilia, por su importancia funcional. También se ha propuesto a la coevolución como uno de los mecanismos evolutivos para generar nuevas especificidades de reacción. Cabe preguntarse entonces si la red con alta MI que se encuentra en la cercanía de los residuos catalíticos se corresponde con los residuos determinantes de especificidad, o hasta qué grado ambos tipos de residuos funcionalmente importantes coincide. Esta cuestión abordada en el capítulo siguiente.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

3

SITIOS DETERMINANTES DE ESPECIFICIDAD, CONSERVACIÓN Y COEVOLUCIÓN.

APLICACIÓN EN LA PREDICCIÓN DE RESIDUOS CATALÍTICOS

3.1 RESUMEN

Utilizando un set de datos comprensivo de familias de enzimas y medidas basadas en conservación, conservación dentro de grupos de secuencias y coevolución, inspeccionamos los diferentes componentes del contenido de información para investigar su potencial predictivo y su grado de concordancia en las predicciones. Los resultados obtenidos demuestran que los diferentes métodos desafiados pueden dividirse a partir de una concordancia limitada de sus predicciones en los siguientes tres grupos:

1) Método Evolutionary Trace real value (rvET) y conservación (C);

2) Métodos derivados de MI;

3) Métodos diseñados específicamente para la predicción de posiciones determinantes de la especificidad: integer-value Evolutionary Trace (ivET), SDPfox y XDET.

Encontramos, usando puntajes de proximidad, que sólo los métodos de los primeros dos grupos presentan un desempeño confiable para la predicción de residuos catalítico. Luego investigamos si los métodos de estos dos primeros grupos aportan información complementaria, obteniendo el mejor desempeño predictivo para la combinación del puntaje de conservación con los puntajes de proximidad de MI y rvET. Sin embargo la señal de rvET provee una aumento limitado en la capacidad predictiva al ser combinado con los puntajes de proximidad de conservación y MI. Estas observaciones demuestran que rvET y cMI proveen información complementaria al sistema predictivo.

Este trabajo contribuye a comprender las diferentes señales de evolución y demuestra que es posible mejorar la predicción de residuos catalíticos integrando información secuencial simple como la conservación, información secuencial compleja e información estructural.

3.2 INTRODUCCIÓN

La mayoría de los métodos desarrollados para la predicción de sitios funcionalmente importantes en proteínas se basan en la detección de alguna señal relacionada con la evolución de la familia de secuencias. Tres señales claras de evolución son: (i) la conservación, (ii) la conservación dentro de grupos de secuencias que comparten una especificidad y (iii) la coevolución entre los residuos.

Es bien conocida la relación entre importancia funcional y conservación evolutiva, resulta intuitivo pensar que un residuo funcionalmente importante deba mantenerse conservado a lo largo de la evolución. El cálculo de conservación es simple y su interpretación biológica es directa. Es esperable que un cambio en una posición conservada, incluso para grupo de proteínas muy diversas, tenga un efecto deletéreo en su función. El cambio en un residuo que se mantuvo conservado por otro residuo que también se conserva, puede indicar un cambio en la función de la proteína; de allí el uso de medir la conservación dentro de grupos de secuencias para predecir posiciones determinantes de especificidad (SDPs). Estas posiciones son aquellas que en un MSA están conservadas dentro de un grupo de proteínas que poseen la misma especificidad, pero difieren entre los grupos, pudiendo estar involucradas en la unión específica a un sustrato/inhibidor o en la interacción con otras moléculas Oliveira L et al. [91], Pirovano et al. [98], Chakrabarti and Panchenko [23].

Se han desarrollado diversos métodos para predecir SDP, la mayoría de ellos requieren una clasificación previa de las secuencias en grupos de especificidad Pirovano et al. [98], Casari et al. [20], Brown et al. [15]. Esta clasificación impone un límite en la aplicabilidad de los métodos ya que para la gran mayoría de las proteínas se desconoce su especificidad. Para resolver este problemas otros métodos de predicción de SDPs realizan una clasificación previa de las secuencias en los grupos de especificidad siguiendo algún criterio ad-hoc. Por ejemplo Capra and Singh [19] arman los grupos de especificidad usando una combinación de la clasificación en familia de Pfam, los números EC (Enzyme Commission numbers) y similitud de secuencias. Otros métodos definen los grupos de secuencias basados solamente en similitud de secuencia Mazin [85] o estadística Bayesiana Marttinen [84]. Muchos de estos métodos realizan una aproximación a la clasificación de las secuencias utilizando información filogenética Lichtarge et al. [77], Mihalek et al. [86], Pei [94] o una combinación de información filogenética y análisis de entropía. Otros métodos clasifican de manera ordenada los residuos según su importancia relativa en el MSA Lichtarge et al. [77], Mihalek et al. [86], Morgan [88], Sankararaman and Sjolander [109]. Estos abordajes son diferentes en su diseño pero todos ellos buscan patrones específicos de conservación como indicador de importancia funcional.

Existe una gran variedad de métodos dedicados a identificar residuos con impacto crítico en la función proteica basados en medidas del contenido de información extraído de un MSA. Sin embargo no está claro hasta qué punto el poder predictivo de los diferentes métodos coinciden. Tampoco se ha investigado si alguno de ellos supera a los otros en su capacidad predictiva cuando se emplean para la identificación de un tipo particular de sitio con importancia funcional. Este trabajo tiene como objetivo abordar esta cuestión, evaluar el grado de

concordancia en las predicciones de los diferentes métodos y comprara su capacidad predictiva en la identificación de residuos catalíticos en enzimas. La elección de los residuos catalíticos para realizar el ensayo se debe a la anotación precisa y no ambigua disponible a nivel de residuos.

Dentro de los métodos disponibles para predecir SDPs se eligieron aquellos que no requieren una clasificación previa de las secuencias por especificidad. Se incluyeron dos tipos de métodos:

1. Los métodos que asignan un valor a un residuo según su importancia funcional:
   a) Conservación
   b) MI
   c) Evolutionary trace real value (ETrv) que utiliza información evolutiva y entrópica Mihalek et al. [86]
2. Los métodos destinados a predecir SDPs:
   a) Puntaje de evolutionary trace integer value (ETiv), que representa la conservación dentro de grupos de manera cualitativa
   b) SDPfox Mazin [85] que predice SDP de manera independiente de la filogenia
   c) XDET Pazos et al. [93] que está basado en la comparación del comportamiento mutacional

La comparación de estos métodos nos permite descomponer el contenido de información de un MSA, investigar el grado de coincidencia en las predicciones y estimar su potencial predictivo en la identificación de residuos catalíticos.

3.3 MATERIALES Y MÉTODOS

3.3.1 *Construcción del set de datos*

El set de datos utilizado consta de MSAs de enzimas extraídos de Pfam, con al menos un representante de estructura conocida y con anotaciones en la base de datos CSA. Se utilizó como datos iniciales prácticamente el mismo set de datos que en el capítulo anterior, exceptuando 10 familias de Pfam que fueron descartadas por contener errores en alguna de las bases de datos empleadas. Para calcular el efecto de la redundancia de secuencias en los diferentes métodos, se evaluó el desempeño de los métodos MI y ET (Evolutionary Trace) usando alineamientos múltiples tal como se recuperan de Pfam, así como también en un set de alineamientos a los cuales se les reduce la redundancia al 62 % de identidad de secuencias. Los métodos

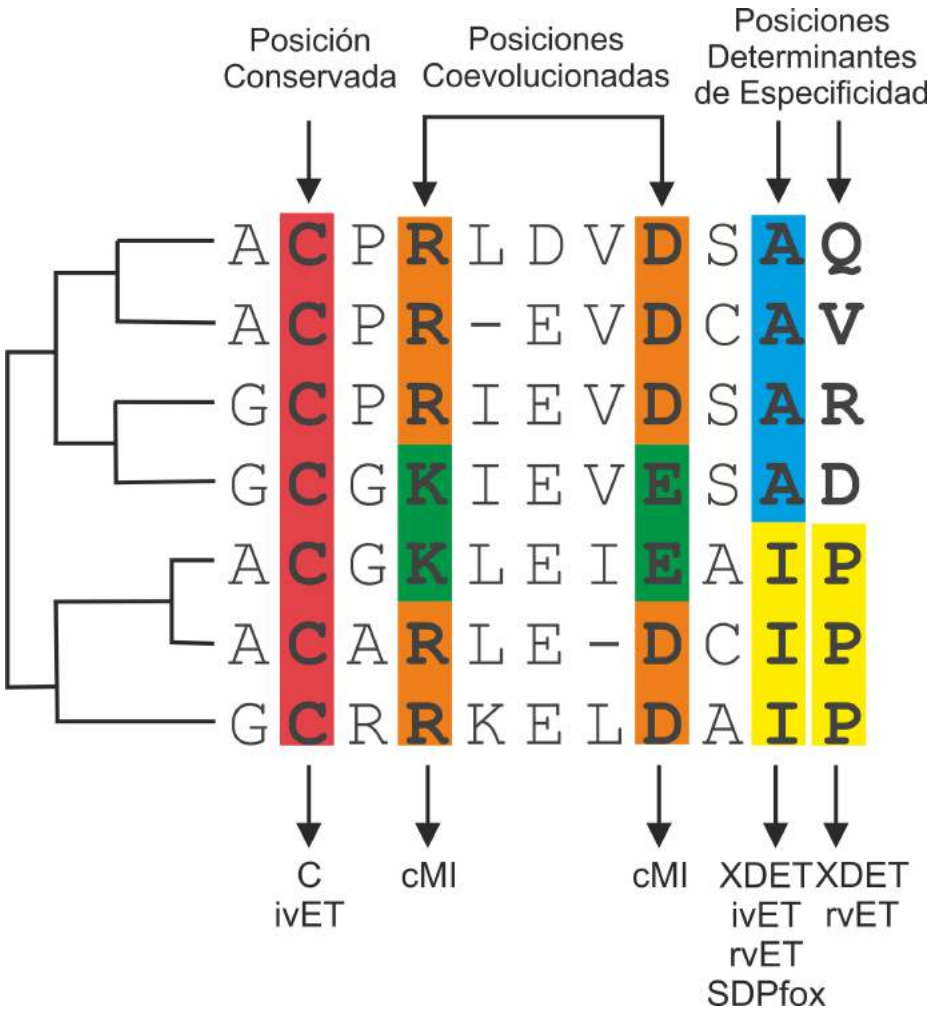


Figura 9: Diferentes patrones de columnas en un MSA
Representación esquemática de un MSA y el árbol filogenético correspondiente (a la izquierda). La posición conservada se resalta en rojo, las posiciones coevolucionadas en naranja y verde, y los SDPs en amarillo y azul. En la parte inferior del MSA se indican los métodos desarrollados para detectar cada tipo de posición.

SDPfox y XDET fueron evaluados solamente en el set de MSAs de redundancia reducida al 62 % y al 50 % de identidad de secuencia respectivamente, debido a limitaciones de los métodos en el número de secuencias y el tiempo de cálculo requerido, como se detalla en la sección siguiente. Los diferentes set de datos fueron llamados MSA100 (sin reducción de redundancia); MSA62 y MSA50 respectivamente. Para distinguir los resultados de métodos obtenidos en los diferentes grupos de MSAs, se indica el nombre del método seguido del número correspondiente al MSA utilizado, por ejemplo MI62 hace referencia al método de MI calculado en el set de MSAs con redundancia reducida al 62 % de identidad.

La reducción de redundancia en los alineamientos fue realizada con el programa T-Coffee Notredame et al. [90]. El set de datos completo de MSAs para las 424 familias de Pfam, incluyendo la anotación de los residuos catalíticos esta disponible en el sitio `http://www.cbs.dtu.dk/suppl/immunology/CSA`. En la Figura 10 se muestra un histograma del número de secuencias por familias para cada set de MSAs.

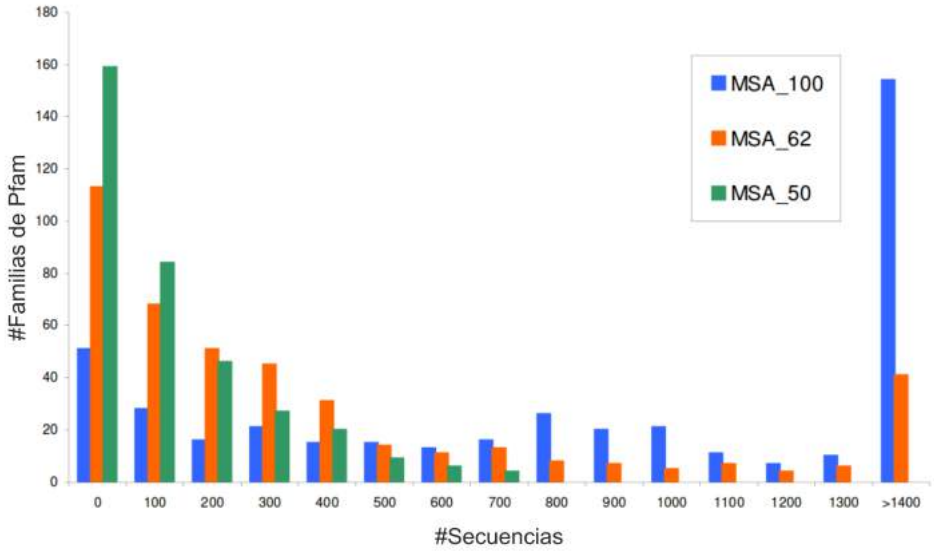


Figura 10: Distribución del número de secuencias en las familias Pfam para cada grupo de MSAs.
MSA100 sin reducción de redundancia; MSA62 y MSA50 con reducción de redundancia al 62 y 50 % de identidad de secuencia respectivamente.

3.3.2 *Programas para predecir SDPs*

Los programas utilizados realizan las predicciones a partir de un MSA y dan como resultado un puntaje por residuo. La predicción de SDPs se llevo a cabo con:

A. Integer Value ET (ivET) que representa la conservación dentro de grupos de una manera cualitativa Mihalek et al. [86]. La Figura 11 se representa el funcionamiento del método ET. El puntaje ivET representa el número de ramas resultante al cortar el árbol en un determinado nodo, para que esa posición sea determinante de especificidad. Una posición completamente conservada tendrá un puntaje ivET=1, y una posición completamente variable solo será clase específica en el último corte sobre las hojas del árbol y su puntaje será más elevado. Un menor valor del puntaje ivET indica mayor relevancia de la posición.El cálculo de este puntaje para una posición i se realiza como se muestra en la siguiente ecuación

$$ivET_i = 1 + \sum_{n=1}^{N} X \begin{cases} 0 & \textit{si la posición i está conservada dentro de cada grupo g} \\ 1 & \textit{en cualquier otro caso} \end{cases}$$

B. Método SDPfox para la predicción de SDPs de manera independiente de la filogenia Mazin [85]. Este programa consta de dos algoritmos, uno dedicado a la separación de las secuencias en grupos, y el otro diseñado para predecir SDP que calcula la información contenida para una posición p como se muestra en la ecuación 12. El programa fue descargado de `http://bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp` y fue ejecutado de manera local con los parámetros por defecto. Este método presenta una limitación en el número de grupos de especificidad por familia que pueden ser analizados, el límite es entre 2 y 200 grupos de especificidad. También posee un límite para la longitud total de las secuencias (<500 residuos). Por estas limitaciones las predicciones con este modo fueron hechas en el set de datos de MSA62.

$$Ip = \sum_{i=1}^{N} \sum_{\alpha=1}^{20} f_p(\alpha, i) log \frac{f_p(\alpha, i)}{f_p(\alpha)} \tag{12}$$

Donde $\alpha = 1, ..., 20$ es el tipo de residuo, $f_p(\alpha, i)$ es la relación entre el número de ocurrencias de residuo α en el grupo i en la posición p sobre la longitud de la columna en el MSA completo, $f_p(\alpha)$ es la frecuencia del residuo α en la columna del MSA, $f_{(i)}$es la fracción de proteínas que pertenecen al grupo i.

C. Programa XDET basado en la comparación del comportamiento mutacional de una posición contra el comportamiento mutacional del alineamiento completo Pazos et al. [93], del Sol Mesa et al. [37]. El programa está compuesto de dos métodos para detectar las posiciones relacionadas con la especificidad funcional. Se utilizó el método llamado mutational behavior (MB)

de XDET ya que no requiere una clasificación previa de las secuencias en los grupos de especificidad, en la Figura 12 se esquematiza su funcionamiento. Debido al alto costo de tiempo computacional requerido por este método, donde el tiempo crece cuadráticamente con el número de secuencias, solo fue posiblemente ejecutar el programa XDET en el set de datos MSA50. El código fuente de XDET fue provisto por los autores.

3.3.3 *Métodos de predicción de sitios funcionalmente importantes*

La predicción de sitios funcionalmente importantes fue llevada a cabo con los siguientes métodos:

A. Conservación de secuencia, calculada en el MSA100 como la entropía relativa Kullback-Leibler usando la distribución de frecuencias de aminoácidos background obtenida de la base de datos UniProt http://www.uniprot.org.ar/.

B. Información mutua calculada en términos de la información mutua acumulada (cMI), que mide el grado de información mutua compartida para un residuo dado.

C. Puntaje real value Evolutionary Trace (rvET) Mihalek et al. [86], Rodriguez [105] que incorpora la entropía como medida cuantitativa de la conservación, generando como resultado un ranking de posiciones según su importancia relativa, como se muestra en la ecuación 13.

$$rvET = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} (- \sum_{a=1}^{20} f_{ia}^{g} \ln f_{ia}^{g}) \tag{13}$$

Donde f_{ia}^{g} es la frecuencia de aparición del aminoácido a en la posición i en el grupo g, N es el número de secuencias (número total de ramas en el árbol) y n es el número de nodo numerado según el método UPGMA (ver numeración de nodos en Figura 11) .

3.3.4 *Derivación de puntajes para predecir residuos catalíticos*

Para integrar los diferentes puntajes con la medida de conservación, se definió un puntaje combinado como se muestra en la ecuación 14

$$P = (1 - w) \cdot C + w \cdot X \tag{14}$$

Donde w es un peso relativo que pertenece al rango $[0 - 1]$. Para cada MSA de familias de proteínas los puntajes fueron normalizados para que pertenezcan al rango $[0 - 1]$. Esta combinación se realizó

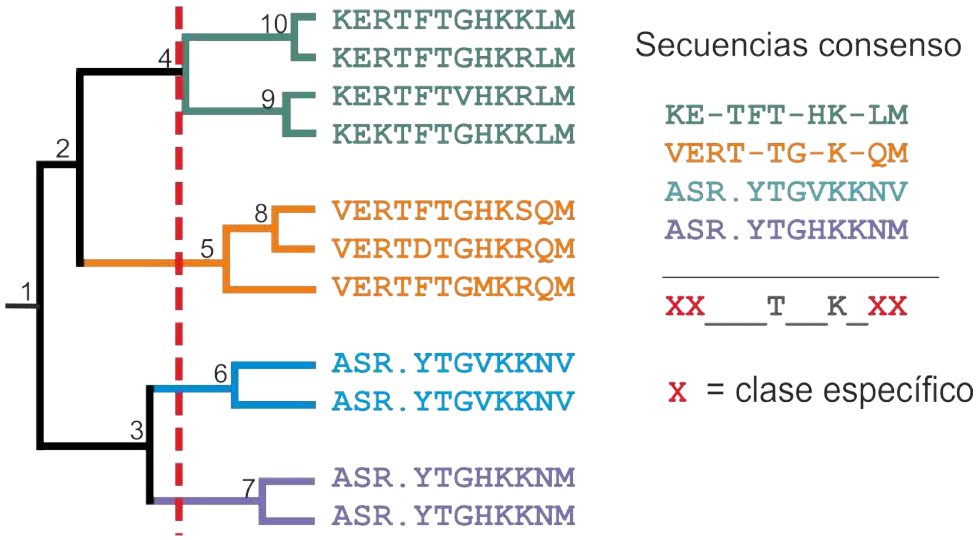


Figura 11: Método Evolutionary Trace

A partir del MSA el programa crea un árbol filogenético por similitud de secuencias, luego el árbol es cortado verticalmente sobre distintos nodos un determinado número de veces. Con cada corte del árbol se definen diferentes grupos de secuencias a las que se le calcula una secuencia consenso. En la figura el corte es representado por la linea punteada roja y quedan definidos cuatro grupos de secuencias (cuatro ramas en el árbol). Para una posición dada la secuencia consenso incluye un aminoácido si este está completamente conservado en el grupo de secuencias. Si la posición en variable dentro de un grupo, se considera que esa posición es neutral y en la secuencia consenso se representa con "-".

Finalmente se comparan las secuencias consenso, si un residuo es conservado dentro de cada grupo y difiere en al menos un grupo, entonces se considera que puede determinar la especificidad. Estas posiciones se señalan con una X. En el ejemplo de la figura, las posiciones predichas como clase específica poseen un puntaje ivET = 4, el número de ramas generadas a partir del corte.

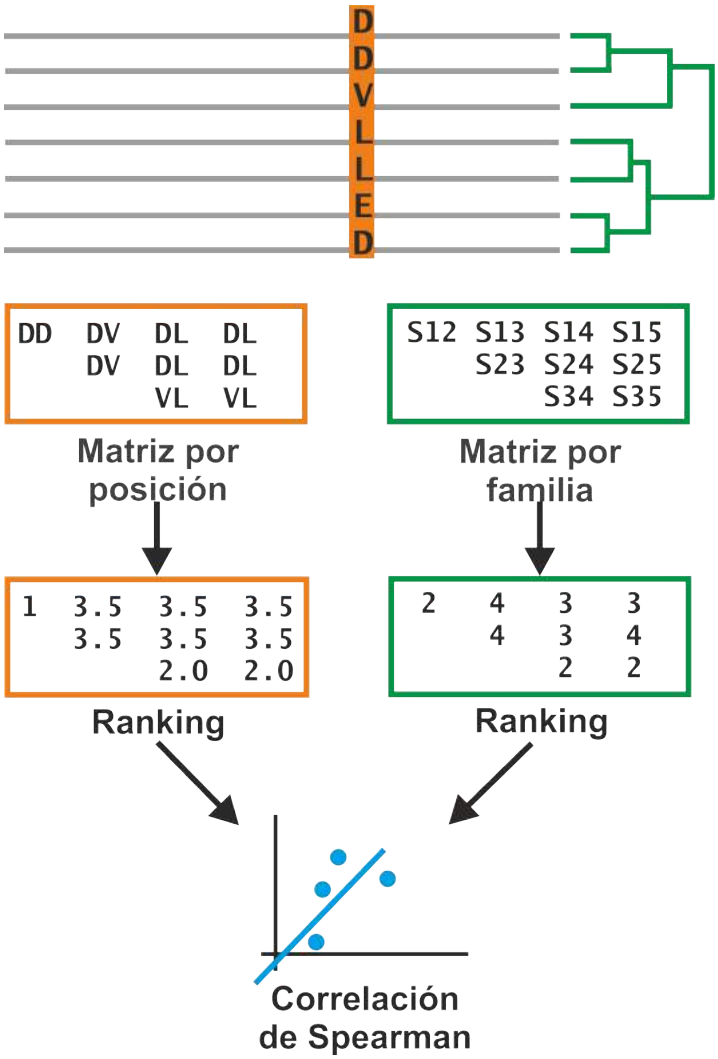


Figura 12: Esquema del método Mutational Behavior de XDET

En este método se compara el comportamiento mutacional de cada posición en el MSA con el comportamiento mutacional de la familia completa. El comportamiento mutacional de una posición se representa por una matriz que contiene un puntaje de similitud para todos los pares de residuos presentes en una posición dada. El comportamiento mutacional de una familia completa se representa con una matriz de similitud de a pares de secuencias. Finalmente se calcula la correlación de spearmen entre ambas matrices. Una alta correlación indica que la posición es un buen candidato de SDP. Figura basada en la figura 6 de del Sol Mesa et al. [37].

para X={pC , prvET62 , pMI62}. Para el puntaje p(ET), la fórmula cambia a la ecuación 15

$$P = (1 - w) \cdot C - w \cdot prvET \tag{15}$$

Debido a que para ET un menor puntaje indica mejor posición en el ranking.

También se evaluó la combinación de los diferentes puntajes integrados a la conservación y conservación en la proximidad, el puntaje combinado se calculo como se muestra en la ecuación 16

$$P = (1 - w_1 - w_2)C + w_1 pC + w_2 X \tag{16}$$

Donde w_1y w_2 son pesos relativos en el rango $[0 - 1]$ y $w_1 + w_2 < 1$. Con la ecuación 16 la combinación fue hecha para X={prvET62, pMI62}. También aquí el signo del último término fue negativo para X=prvET62.

Finalmente se combinaron los métodos como se muestra en la ecuación 17

$$P = (1 - w_1 - w_2 - w_3)C + w_1 pC - w_2 prvET62 + w_3 pMI62 \tag{17}$$

Donde w_1, w_2 y w_3 son pesos relativos en el rango de $[0 - 1]$ y $w_1 + w_2 + w_3 < 1$

3.4 RESULTADOS

El análisis está basado en el mismo set de datos de familias enzimáticas de Pfam con anotación original en la base de datos CSA descripto previamente. Con estos datos se calcularon medidas relacionadas a la evolución para los diferentes métodos de interés y luego se analizó la correlación entre estas medidas y su potencial predictivo para la identificación de RC. Todos los métodos estudiados se han desarrollado para la identificación de sitios funcionalmente importantes dentro de familias de proteínas, sin embargo pueden dividirse en dos grupos:

1. Los métodos que ordenan las posiciones en un MSA de acuerdo a su importancia funcional relativa, sin importar cuales son las causas de esa importancia. En esta categoría se incluyen los métodos cMI, real-value evolutionary trace (rvET) y conservación de secuencia (C).

2. El grupo que comprende los métodos destinados a predecir posiciones determinantes de especificidad (SDPs) en una familia de proteínas: XDET, ivET y SDPfox.

3.4.1 *Concordancia de los diferentes métodos predictivos*

Para determinar la influencia que la redundancia de datos puede tener en las predicciones para los diferentes métodos, se midió la correlación entre los puntajes calculados en los MSAs recuperados de Pfam (MSA100) y en un set de MSAs de redundancia reducida. Si un método fuera insensible a la redundancia de secuencias los puntajes producidos en los diferentes MSAs deberían estar altamente correlacionados. Esto sucede con cMI (Coeficiente de correlación de Spearman, CCS = 0.76) y rvET (CCS = 0.93). Sin embargo para ivET la correlación entre los puntajes obtenidos en los dos set de MSAs fue débil (CCS=0.21) indicando que la redundancia de los datos para este método tiene un fuerte impacto en la predicción (ver Figura 13). Los valores de correlación obtenidos y las desviación estándar se indican en la tabla suplementaria 9 en la página 85.

Los métodos dedicados a la predicción de SDP estiman un puntaje relacionado con la importancia funcional del residuos como determinantes de la especificidad proteica. Un tema relevante para analizar es el grado de concordancia entre ellos. Los métodos de predicción de SDP (ivET, SDPfox y XDET) muestran un solapamiento mutuo limitado (Figura 13), los valores de correlación son bajos para cada comparación, siendo el mayor valor obtenido 0.34 de la comparación entre SDPfox y XDET.

Luego se investigó el grado de coincidencia entre la información extraída de los métodos desarrollados para la detección de SDP (ivET, SDPfox y XDET) y la señal de información de cMI, que indica posiciones con un alto grado de información mutua compartida. Se encontró que la cMI tiene una baja coincidencia con todos los otros métodos (CCS < 0.28 para cada comparación, ver Figura 13).

Luego se analizó la correlación entre los métodos destinados a ordenar las posiciones por importancia funcional (rvET, cMI y conservación). Como se esperaba, la conservación está altamente correlacionada con rvET para los set de datos MSA100 y MSA62 (CCS > 0.7, en ambos casos). Se encontró que la cMI está débilmente correlacionada con la conservación (CCS 0.16 para MSA100 y MSA62), finalmente la correlación entre rvET y cMI fue moderadamente débil con una correlación máxima de 0.41.

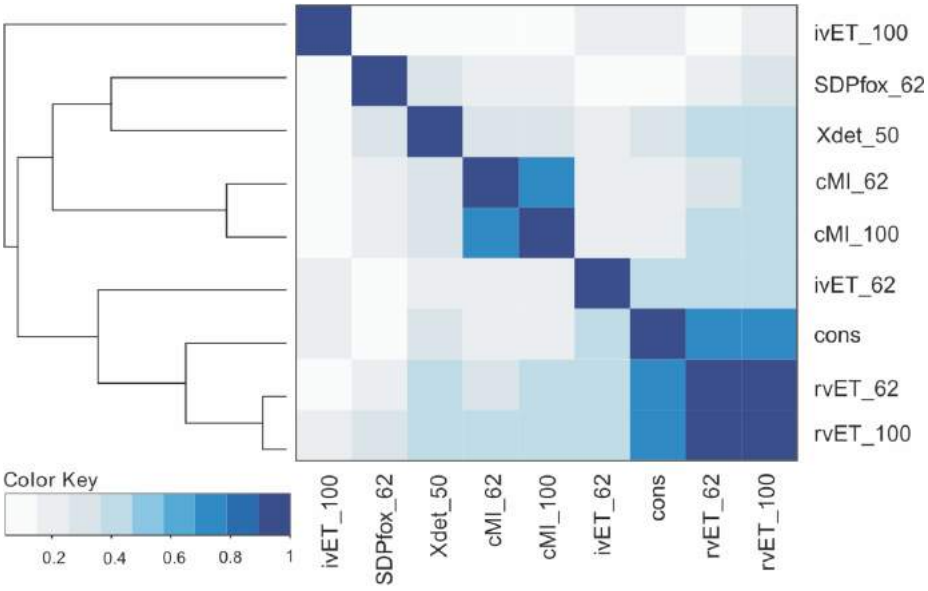


Figura 13: Representación en HeatMap de la correlación de Spearman entre métodos.

Los números luego del nombre de los métodos (50, 62 y 100) indican la redundancia de las secuencias en el MSA. El dendograma a la izquierda indica la distancia entre los métodos. Los colores indican la correlación en escala de blanco (0, sin correlación) a azul (1, correlación perfecta). Todas las correlaciones son estadísticamente diferentes de cero (T-test, p-value de corte 0.05)

Los resultados anteriores demuestran que los diferentes métodos ensayados pueden ser divididos en los siguientes tres grandes grupos con solapamiento limitado:

1. Métodos con una señal altamente correlacionada con la conservación, compuesto por rvET, la conservación en sí misma e ivET cuando es evaluado en un set de datos de redundancia reducida.

2. Métodos cuya señal es derivada de la información mutua, compuesto solamente por cMI.

3. Métodos diseñados explícitamente para la identificación de SDPs, compuesto por SDPfox y XDET. Estos métodos presentan baja correlación con cualquier otro método ensayado.

Finalmente el método ivET evaluado en el set de datos MSA100 (ivET100) aparece como un outlier en el análisis y no muestra solapamiento con ninguno de los otros métodos. El resultado general señala que el solapamiento entre los diferentes métodos en la mayoría de los casos es limitado, sugiriendo que los residuos con mayor puntaje de cMI y SDPs no forman necesariamente el mismo grupo de residuos. Además es notable que los métodos creados para detectar el mismo tipo de posiciones como SDPs (ivET, SDPfox y XDET) muestran entre ellos baja concordancia en los puntajes de predicción.

3.4.2 *Medida de suma de información en la proximidad para predecir residuos catalíticos*

Como se describió previamente en el capítulo anterior, los residuos catalíticos están caracterizados por una alta MI en su proximidad estructural, es decir que los residuos que se encuentran dentro de una determinada distancia umbral a los RC son ricos en cMI. Para investigar si existe una observación similar con los otros métodos analizados calculamos una medida de proximidad para cada método e investigamos en qué grado esta medida contribuye a la identificación de RC. Para cada residuos calculamos un puntaje de proximidad como la suma de los puntajes de los residuos localizados dentro de una determinada distancia del residuo en cuestión como se muestra en la ecuación 18.

$$pScore_i = \frac{1}{N} \cdot \sum_{j, d_{ij} < u} Score_j \tag{18}$$

Donde la suma se realiza sobre todos los residuos j en una proteína dada dentro de una distancia $d_{ij} < u$ del residuo i, d_{ij} es calculada como la distancia mínima entre cualquier par de átomos diferentes de H de los residuos i y j, $Score_j$ es el puntaje para el residuos j de un método dado, y u es la distancia umbral. Los umbrales de distancia u fueron optimizados para cada método predictivo.

Estas medidas fueron designadas anteponiendo la letra "p" al nombre del método, por ejemplo p(rvET) por el puntaje de proximidad de rvET.

La tabla 4 muestra el desempeño predictivo de los diferentes métodos con las correspondientes distancias óptimas. Todos los métodos con la excepción de p(ivET) evaluado en MSA100, SDPfox y XDET, pueden ser utilizados como predictores razonables de RC (AUC > 0.8 para todos los casos). Nuevamente se observa una gran diferencia en el rendimiento del método ivET al ser evaluado en los diferentes MSAs. Todos los métodos con AUC > 0.8 presentan el valor umbral de distancia entre 5 y 7Å. Los valores umbral de distancia fueron mayores sólo para los métodos de predicción de SDPs que tienen un rendimiento predictivo bajo.

Método	Promedio AUC	Umbral de Distancia (Å)
p(SDPfox62)	0.703	10
p(XDET50)	0.736	8
p(ivET62)	0.835	7
p(ivET100)	0.640	7
p(rvET62)	0.878	5
p(rvET100)	0.875	7
p(MI62)	0.823	7
p(MI100)	0.833	7
p(C)	0.854	5

Tabla 4: Evaluación de puntajes de proximidad
Desempeño predictivo de los puntajes de proximidad, se indica la distancia umbral con la que se obtuvo el mejor desempeño predictivo para la detección de RC medido en términos de AUC.

Luego se investigó en qué grado el desempeño predictivo de los métodos es estadísticamente diferente. Se obtuvo el siguiente orden entre los métodos:
$p(rvET62) \backsimeq p(rvET100) \simeq p(C) > p(ivET62) \backsimeq p(MI100) \backsimeq p(MI62) > p(XDET50) \backsimeq p(SDPfox62) > p(ivET100)$
El símbolo " $\backsimeq$ " significa que el valor precedente es mayor pero no estadísticamente diferente; el símbolo ">" significa que el valor precedente es significativamente mayor, realizando test binomial excluyendo extremos con un valor de corte de p-value de 0.05. Considerando el desempeño predictivo de los métodos para identificar RC, éstos pueden dividirse en tres grupos:

1. p(rvET62), p(rvET100) y p(C)
2. p(ivET62), p(MI100) y p(MI62)
3. p(XDET50) y p(SDPfox62)

Considerando a el método p(ivET100) como outlier.

3.4.3 *Puntaje combinado para la predicción de residuos catalíticos*

En el capítulo anterior demostramos que los puntajes pC y pMI mejoran la mejora predictiva de RC al ser combinados con la conservación. Ahora se busca investigar hasta qué grado esta observación se mantiene cuando se integran los otros métodos aquí estudiados al combinarse con la conservación. De esta manera podemos investigar

hasta qué grado cada método aporta información complementaria al modelo predictivo. Se definieron diferentes modelos, combinando uno o más puntajes de proximidad con la medida de conservación. Se evaluaron los métodos p(MI62) (la medida usada previamente para la detección de RC), el puntaje de ET con mejor desempeño p(rvET62) y p(C).

La tabla 5 muestra los valores de desempeño en términos de AUC y AUC0.1 y los pesos relativos para los diferentes modelos. Por ejemplo la segunda fila corresponde al modelo $0{,}2 \cdot C + 0{,}8 \cdot p(C)$ y se indica el desempeño óptimo para el modelo definido como combinación del puntaje de conservación y el puntaje de proximidad de la conservación, y se indica que el peso relativo para ambos términos es 0.2 y 0.8 respectivamente. En la tabla, un peso relativo igual a cero indica que ese puntaje no contribuye al desempeño del modelo.

Método	AUC	AUC0.1
C	0.881	0.491
0.2 C + 0.8 p(C)	0.898	0.553
0.15 C + 0.85 p(rvET62)	0.913	0.567
0.25 C+ 0.75 p(MI62)	0.912	0.555
0.15 C +0.0 p(C) + 0.85 p(rvET62)	0.913	0.567
0.15 C +0.3 p(C) +0.55 p(MI62)	0.916	0.571
0.15 C+ 0.0 p(C)+0.45 p(rvET62)+0.4p(MI62)	**0.921**	**0.586**

Tabla 5: Desempeño de los diferentes métodos en términos de AUC
La columna Método especifica el modelo combinado con los pesos óptimo, el modelo incluye el puntaje de conservación combinado con diferentes puntajes de proximidad. AUC y AUC01 son los valores promedios sobre las 424 familias.

A partir de estos resultados notamos que todos los puntajes de proximidad contienen información complementaria a la conservación, ya que mejoran el desempeño predictivo. Es decir, los métodos C+p(X) donde X es igual a C, rvET62 o MI62, mejoran significativamente el desempeño predictivo del puntaje de conservación ($p<0.05$, test binomial excluyendo extremos). También es interesante observar que los pesos de relativos de p(C) son cero en todos los modelos que incluyen p(rvET62). Esto sugiere que el alto desempeño de p(rvET), mostrado en la tabla 4, se debe a la señal de conservación contenida en el puntaje rvET, como también se sugirió a partir del análisis de correlaciones presentado en la Figura 13.

El desempeño predictivo para el modelo C+p(C)+p(MI) no presenta diferencia estadísticamente significativa del obtenido con C+p(C)+p(rvET)

($p<0.1$, test binomial excluyendo extremos). Finalmente el modelo con mejor desempeño fue C+p(rvET)+p(MI), que mejora significativamente todos los modelo excepto C+p(C)+p(MI) ($p<0.05$, test binomial excluyendo extremos).

Los resultados obtenidos demuestran que los puntajes rvET y cMI capturan distintas señales de un MSA y aportan información complementaria al sistema predictivo.

3.5 DISCUSIÓN

Muchos algoritmos han sido propuestos para la identificación de residuos que son críticos para el funcionamiento general de una proteína, y de manera particular para la especificidad. En el presente capítulo se comparó un grupo de estos métodos para predecir sitios determinantes de especificidad en términos de concordancia en sus predicciones y también en la habilidad para identificar residuos catalíticos en enzimas mediante un puntaje de proximidad. A partir de los resultados se desprende que lo métodos desafiados pueden dividirse en tres grupos, con limitado solapamiento mutuo. Un grupo compuesto por los métodos cuya señal predictiva está altamente correlacionada con la conservación de secuencia (rvET y la propia conservación); un segundo grupo compuesto por lo métodos cuya señal predictiva es derivada de la información mutua (cMI); y el tercer grupo conteniendo los métodos desarrollados específicamente para la detección de sitios determinantes de especificidad (SDPfox, XDET y ivET).

Con respecto a la influencia de la redundancia de secuencias sobre las predicciones, se encontró que el método ivET es altamente sensible a la redundancia, debido a que la correlación de los puntajes obtenidos en los dos set de MSAs es débil (CCS=0.21). Una posible explicación de este efecto se representa en la Figura 14.

Figura 14: Representación esquemática del impacto de la redundancia de secuencia en el puntaje ivET

La imagen superior representa un alineamiento múltiple de secuencias redundante, y la inferior uno con redundancia reducida. Las lineas punteadas en color rojo representan las posibles particiones del árbol filogenético, generando en cada caso dos grupos de secuencias resaltados en naranja y en celeste. Para cada grupo el método construye una secuencia consenso, representada a la derecha de cada grupo. La secuencia consenso indica el aminoácido invariable dentro del grupo, y se indica como gap ("-") la posición variable. Seguidamente las secuencias consensos son comparadas para generar la secuencia "Evolutionary Trace" (representada en color negro en la figura). Solo las posiciones con una conservación grupo-específica son predichas como SDPs (indicadas como **X** en la figura), es decir posiciones con un aminoácido invariable en un grupo y un aminoácido también invariable pero distinto del primero, en el otro grupo. La diferencia en la predicción para los dos alineamientos se puede observar en la novena posición, resaltada en fondo gris.

A partir de definir puntajes de proximidad para cada método y desafiándolos para la predicción de residuos catalíticos, encontramos que solo los métodos de los primeros dos grupos mencionados muestran un desempeño predictivo confiable (media de AUC mayor a 0.8), indicando que los puntajes de proximidad derivados de los métodos de predicción de SDPs tienen una capacidad limitada para la identificación de residuos catalíticos.

Comparando los diferentes métodos de predicción de SDPs se encontró que coinciden de manera muy limitada en sus predicciones, a pesar de ser métodos diseñados para capturar la misma señal funcional.

Finalmente investigamos hasta qué grado la señal de información de conservación, rvET y cMI (que pertenecen a los dos grupos con

mejor desempeño para la predicción de RC) son complementarias, de manera que la señal combinada pueda arrojar una mejora significativa en la capacidad predictiva. Se encontró que el desempeño predictivo mejora significativamente al combinar la señal de conservación con los puntajes de proximidad de rvET y MI. Este resultado confirma la noción que los métodos rvET y cMI son distintos en naturaleza y que ambos aportan información que es complementaria al sistema predictivo. Sin embargo, en el set de datos utilizados, también se demuestra que la mejora en el desempeño predictivo con la señal de rvET es limitada e insignificante si se combina con el puntaje de conservación en la proximidad.

Cabe aclarar que las conclusiones obtenidas son con respecto a la identificación de residuos catalíticos, y a pesar de que los resultados de los diferentes métodos para predicción de SDPs no correlacionan entre sí, algunos de ellos han demostrado ser exitosos en set de datos pequeños, con un limitado número de secuencias y pocos grupos de especificidad Chakrabarti and Panchenko [22], Kalinina [64], Ye [131]. También se obtuvieron resultados predictivos satisfactorios en el trabajo Rodriguez [105], donde se demostró mediante verificación experimental la capacidad predictiva de rvET de los residuos responsables de la especificidad entre los ligandos de dopamina y serotonina en el receptor de bioamina de Clase A.

El resultado principal de este análisis es que los puntajes predictivos para los diferentes métodos evaluados tienen una coincidencia mutua limitada, y particularmente los métodos de identificación de SDPs y los métodos basados en información mutua capturan una señal distinta de la información evolutiva.

En conclusión el trabajo contribuye a una mejor comprensión de las diferentes señales evolutivas de una proteína. De una manera cuantitativa se caracterizaron similitudes y diferencias entre diferentes medidas de información capturadas a partir de un alineamiento múltiple de secuencias y se demostró que es posible mejorar de manera significativa la capacidad para detectar residuos catalíticos a partir de la integración de diferentes tipos de medidas.

4

ST3GAL COMO CASO DE ESTUDIO.

PREDICCIÓN DE SITIOS FUNCIONALMENTE IMPORTANTES MEDIANTE MI Y CONSERVACIÓN DENTRO DE GRUPOS

4.1 RESUMEN

En este trabajo se realizó un análisis de secuencia de la familia de proteínas α2,3-sialiltransferasas (ST3Gal), que son enzimas claves en la biosíntesis de glicoproteínas y glicolípidos, se encuentran clasificadas hasta el momento en 6 subfamilias numeradas de ST3Gal I a ST3Gal VI. La exploración de base de datos genómicas y la identificación de motivos funcionales de secuencias, permitió la anotación de 121 secuencias génicas nuevas de ST3Gal. *Se realizó un análisis de redes de similitud de las secuencias proteicas que, combinado con estudio filogenético, nos permitió proponer una nueva clasificación de las secuencias en subfamilias, distinguiendo 3 nuevas subfamilias, llamadas ST3Gal VII, ST3Gal VIII y ST3Gal IX. Estos análisis también nos permitieron distinguir grupos de subfamilias relacionadas. Para cada grupo de subfamilias y para cada subfamilia, se predijeron las posiciones determinantes de especificidad. Adicionalmente para la familia ST3Gal se predijeron las posiciones que coevolucionaron. Ambos tipos de residuos funcionalmente importante fueron analizados en una estructura tridimensional representativa de la familia.*

En conjunto, los resultados obtenidos ayudaron a construir una visión integral de la evolución funcional y molecular de la familia ST3Gal.

4.2 INTRODUCCIÓN

Las sialiltransferasas son enzimas responsables de catalizar la transferencia de un residuo de ácido siálico desde un donor activado (CMP-β-Neu5Ac, CMP-β-Neu5Gc, CMP-β-KDN) al extremo no reductor de cadenas de glicoproteínas y de glicolípidos Audry et al. [7]. Estos ácidos siálicos, de glicoproteínas o glicolípidos, están involucrados en diversas funciones biológicas como ser el proceso de reconocimiento celular, la regulación de la vida media de las células y proteínas plasmáticas y la modulación del sistema inmune y apoptosis Varki [127]. La reacción catalizada por esta superfamilia de enzimas se esquematiza en la Figura 15.

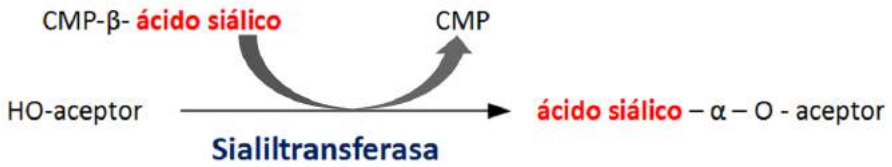


Figura 15: Esquema de la reacción catalizada por las sialiltransferasas

Las sialiltransferasas son una superfamilia de enzimas que catalizan la transferencia de ácido siálico desde un donor activado (CMP-β-ácido siálico) al extremo oxidrilo no reductor de un aceptor, formando glicoproteínas o glicolípidos sialilados.

Particularmente, para la familia ST3Gal (α2, 3-*sialiltransferasa*) el ácido siálico es transferido al residuo terminal de galactosa de disacáridos de tipo I, II ó III (Galβ1,3GlcNAc; Galβ1,4GlcNAc o Galβ1,3GalNAc) resultando en la formación de un enlace glicosídico α2-3.

El análisis de secuencias de sialiltransferasas bien caracterizadas permitió la identificación de 4 motivos funcionales que se han utilizado para la identificación de nuevas secuencias de esta superfamilia. Estos motivos se localizan en el dominio catalítico y son llamados L (large), S (Small), III y VS (Very Small). Estudios de mutagénesis dirigida sugieren que los motivos S, III y VS están involucrados en la unión al sustrato aceptor Datta [34], Datta et al. [33], Datta and Paulson [32]. Adicionalmente, cada subfamilia de sialiltransferasa descripta contiene diferentes motivos funcionales que las caracterizan. En la figura 16 se muestran los motivos de secuencia de la superfamilia y de la familia ST3Gal.

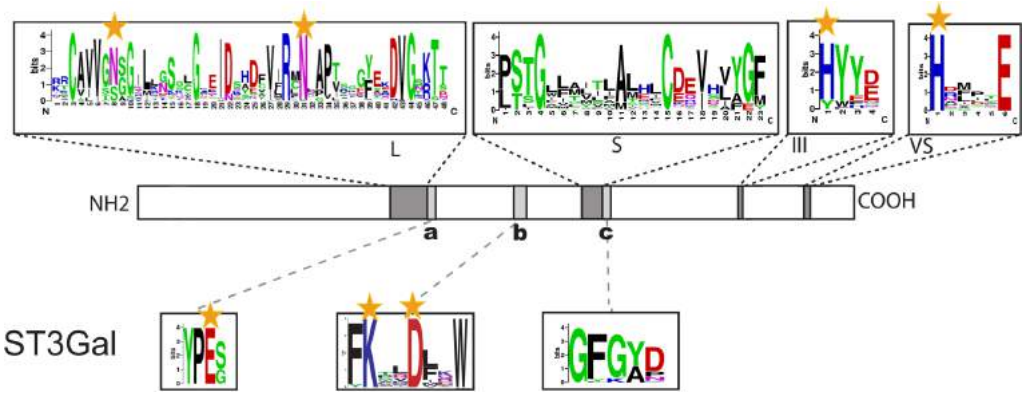


Figura 16: Logo de secuencia de los motivos de la superfamilia Sialiltransferasa y de la familia ST3Gal
Las estrellas amarillas indican los residuos de unión al sustrato. En la parte superior se muestra un primer nivel de conservación representado por logos de secuencias de los motivos de superfamilia L, S, III y VS. Estos motivos caracterizan las sialiltransferasas de vertebrados. En el medio de la figura se muestra una representación esquemática de la secuencias de ST3Gal, donde se indica la posición relativa de los motivos. En la parte inferior se muestran los logos de secuencias correspondiente a los 3 motivos funcionales de la familia ST3Gal, implicados en el reconocimiento de sustrato aceptor (Galβ1,3GalNAc). Figura adaptada de la Figura 1 de Harduin-Lepers [54].

Dentro de la familia ST3Gal, las 6 subfamilias descriptas hasta el momento presentan una especificidad enzimática solapada Ellies et al. [42], Priatel et al. [100], Yang et al. [130], todas ellas guiando la biosíntesis de varios glicotopes[1] sialilados en O-glicanos, N-glicanos y gangliósidos Kitagawa and Paulson [67], Kojima et al. [68], Kono et al. [69], Rohfritsch et al. [106]. Las subfamilias ST3Gal I y ST3Gal II sintetizan preferentemente estructuras α2,3 sialiladas en disacáridos de tipo III (Neu5Acα2-3Galβ1-3GalNAc-R) encontrados en O-glicanos y gangliósidos. Las subfamilias ST3Gal II y ST3Gal III son responsables de prácticamente todas las sialilaciones de los gangliósidos cerebrales en ratón Sturgill et al. [118]. Las subfamilias ST3Gal III, ST3Gal IV y ST3Gal VI comparten una actividad enzimática similar, catalizando la transferencia de ácido siálico al residuo de galactosa del disacárido Galβ1,3GlcNAc de glicoproteínas y glicolípidos. Finalmente ST3Gal V usa como sustrato aceptor principalmente lactosil-ceramida y en menor medida el glicoesfingolípido Gal-Cer, sintetizando los gangliósidos G_{M3} y G_{M4} respectivamente. Los gangliósidos desempeñan un papel fundamental en la formación de los cuerpos celulares neuronales Kotani et al. [71] y la transmisión del impulso nervioso Lo et al. [79]. Además se ha demostrado la existencia de secuencias de ST3Gal,

1 Glicotope: el carbohidrato constituye la región donde el antígeno interacciona con un anti-cuerpo específico, juegando un rol decisivo en el proceso de reconocimiento inmunológico.

que contienen a la vez motivos de secuencias típicos de las subfamilias ST3Gal I y II, por lo que no es posible asignarlas a ninguna de estas dos subfamilias. Se ha sugerido que representan secuencias ancestrales ortólogas al ancestro en común de ambas subfamilias, consecuentemente se las clasifica como ST3Gal I/II Lehmann et al. [75].

La desregulación en la expresión de ST3Gal ha sido reportada en varias enfermedades humanas Dall'Olio and Chiricolo [31], Harduin-Lepers et al. [55], Hennet [58], y más recientemente, se ha reportado que variaciones genética en mutaciones en los genes ST3GAL3 y ST3GAL5 conducen a discapacidad intelectual y defectos congénitos de la glicosilación de tipo II Boccuto et al. [13], Edvardson et al. [41], Hu et al. [60], Simpson et al. [116].

En el presente trabajo se estudia la relación entre las secuencias de ST3Gal mediante redes de similitud. Mostramos que estas secuencias están organizados en cuatro grupos distintos llamados GR1, GR2, GR3 y GRx, este último grupo compuesto por secuencias encontradas exclusivamente en genomas de invertebrados.

También se predijeron los residuos conservados dentro de los grupos propuestos, que podrían estar involucrados en la divergencia funcional y en la especificidad diferencial del aceptor de las enzimas ST3Gal y se predijeron sitios de coevolución mediante el cálculo de MI. Estas posiciones funcionalmente importantes fueron mapeadas en la estructura 3D de una secuencia representativa de ST3Gal a fin de dilucidar su rol funcional.

En conjunto los diferentes resultados permitieron obtener una vista integradora de la evolución funcional y molecular en Deuterostomados y proponer un modelo de eventos de duplicación por el que los genes st3gal han sufrido divergencia funcional en vertebrados superiores y perdida de gene de manera linaje-específica en mamíferos.

4.3 MATERIALES Y MÉTODOS

4.3.1 *Recuperación de secuencias de Sialiltransferasas*

Para el presente estudio se realizó sobre seceuncias de ST3Gal pertenecientes a eucariotas, la identificación de estas secuencias se realizó como se describe en Petit et al. [95]. Como primera fuente de información, se utilizaron las secuencias de la familia GT-29 de la base de datos CAZy `http://www.cazy.org/` Cantarel et al. [18]. Luego se buscaron secuencias homólogas de ST3Gal explorando todas las bases de datos genómicas y EST disponibles, como las mantenidas por NCBI, DDBJ, ENSEMBL y las bases de datos especializadas JGI para *Branchiostoma floridae*, the genome sequencing center de Whasington Univeristy School of medicine, MO para la lamprea *Petromyzon marinus*, KEGG GENES Hashimoto et al. [57, 56], Kanehisa and Goto [65] usando BLASTN, TBLASTN y PsiBLAST Altschul et al. [4] con

un valor de corte de E-value de 0.01 para todas las búsquedas. La asignación de las secuencias encontradas a la familia ST3Gal fue determinada por la presencia de motivos de secuencias específicos que sirven como firma de esta familia Harduin-Lepers [54], Patel and Balaji [92].

4.3.2 *Red de similitud de secuencias*

La red de similitud de secuencias se construyó como se describe en Atkinson et al. [6]. Se creó una base de datos personalizada utilizando formatdb del programa BLAST stand-alone. La base de datos incluye 336 secuencias de ST3Gal relacionadas y 27 de ST6Gal I como grupo de control negativo. La relación de a pares de secuencias fue calculada realizando una búsqueda por cada secuencia con el algoritmo blastall en la base de datos personalizada, y el puntaje E-value para cada comparación se utilizó como medida de similitud entre secuencias. La red fue visualizada utilizando el programa Cytoscape Shannon et al. [113], donde cada secuencia es representada como un nodo y las líneas fueron definidas entre cada par de nodos con un E-value menor que el valor de corte. Los nodos fueron coloreados según la subfamilia a la que pertenecen, ya sea conocida (ST3Gal I- ST3Gal VI) o predicha (ST3Gal VII-ST3Gal IX).

Los puntajes E-values dependen de la base de datos donde se realiza la búsqueda, en el presente análisis al utilizar un grupo de secuencias relacionadas como base de datos, el E-value debe considerarse simplemente como un tipo de puntaje y no como un verdadero valor esperado por azar.

4.3.3 *Conservación y predicción de SDPs*

Para medir la conservación para cada subfamilia de ST3Gal y grupos de subfamilias, se programó una búsqueda automática de secuencias consenso como se describe a continuación. Dentro de una ventana de 4 aminoácidos continuos en la secuencia, una secuencia peptídica fue considerada conservada si más del 70 % de las posiciones presentan un aminoácido con al menos el 60 % de presencia dentro de las secuencias seleccionadas, de lo contrario la posición es considerada variable. El tamaño de la ventana y los valores de corte son parámetros a seleccionar del programa. Para remover posiciones no informativas, las columnas que contuvieran más del 50 % de gaps fueron eliminadas. El último paso consiste en alinear los bloques de secuencias conservadas obtenidos para los diferentes grupos de secuencias. El alineamiento construido contiene 382 posiciones de aminoácidos. El programa escrito en C++ se encuentra disponible a solicitud.

Para cada subfamilia se generó una secuencia consenso usando el programa descripto y fueron alineadas entre sí usando el programa MEGA 5.0 Hall [53]. El alineamiento resultante fue utilizado para detectar las posiciones que son conservadas dentro de grupos de proteínas que se predice que llevan a cabo la misma función (grupos de especificidad).

Las predicciones se realizaron considerando los patrones de conservación entre los 3 grupos de subfamilias (GR1, GR2 y GR3) y también dentro de cada grupo. Las posiciones predichas fueron mapeadas en la estructura 3D de referencia.

4.3.4 *Predicción de sitios que coevolucionan y modelado de una región en la estructura 3D*

Las posiciones que coevolucionan fueron predichas utilizando el servidor web MISTIC Simonetti et al. [115]. La predicción está basada en el cálculo de MI entre pares de residuos de un MSA. Los resultados incluyen un puntaje de MI entre pares, los puntaje por posición cMI y pMI. Para el cálculo de este último se utilizó la única estructura conocida de sialiltransferasa de mamíferos. La estructura pST3Gal I porcina fue parcialmente resulta por cristalografía de rayos X (PDB: 2WNB) Rao et al. [102] y completada modelando el dominio flexible lid usando el método de modelado de loops de MODELLER Sali and Blundell [108]. Se generaron 50 loops en la región 305-316, que luego fueron rankeados por el puntaje DOPE. Los modelos con mejor puntación fueron inspeccionados visualmente y evaluados con plots de Ramachandran. El modelo seleccionado presenta todos los ángulos torsionales de la región modelada en la zona permitida. Como se trata de un loop móvil, el modelo solo trata de representar una posición posible de los átomos en el espacio para visualizar los resultados de la predicción en el contexto de la proteína completa y mejorar la comprensión de la importancia funcional de esta región.

4.4 RESULTADOS

4.4.1 *Identificación de secuencias de ST3Gal*

Para encontrar los genes con similitud significativa a los genes conocidos *st3gal* de mamíferos, se realizó una búsqueda BLAST en varias bases de datos nucleotídicas de vertebrados e invertebrados usando las secuencias de ST3Gal conocidas. La búsqueda fue basada en el hecho de que las secuencias de los motivos L, S, III y VS son altamente conservados en todas las sialiltransferasas animales y sirven como firmas para su identificación. Adicionalmente se buscaron los motivos de la familia ST3Gal a, b y c para la identificación de homología. Se encontró una amplia distribución filogenética de los genes *st3gal*

en el reino Metazoa. Varios marcadores de secuencia expresada o EST (acrónimo del inglés expressed sequence tag) de la esponja *Oscarella carmela* fueron atribuidas a una única secuencia se *st3gal1/2* (ver Figura 17).

Sin embargo no se encontraron secuencias de *st3gal* en cnidarios o protostomados. En Ambulacraria se encontraron dos copias de *st3gal* para el Hemicordado *S. kowalevskii* y el Equinodermo *Strongylocentrotus purpuratus* (erizo de mar púrpura) y una copia de *st3gal* en el genoma del erizo de mar *Hemicentrotus pulcherrimus*. Para Vertebrados la mayoría de los genomas de mamíferos examinados, incluyendo *Homo sapiens* y *Mus musculus*, contienen los seis miembros de ST3Gal previamente descriptos. El ave *Gallus gallus* presenta ocho secuencias correspondientes a ST3Gal, mientras que el genoma del anfibio *Silurana tropicalis* (rana) contiene siete secuencias y el ajolote *Ambystoma mexicanum* presenta tres. En la rama de los actinopterigios (clase de peces óseos), el pez primitivo catán pinto *Lepisosteus oculatus* presenta ocho secuencias de ST3Gal, mientras que en genoma de Teleósteos se encuentran más de doce secuencias. Finalmente seis secuencias correspondientes a ST3Gal fueron encontradas en la lamprea *Petromyzon marinus.*

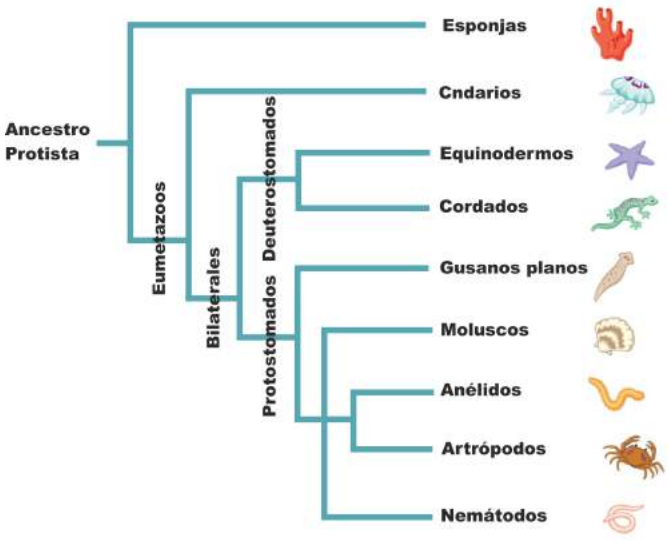


Figura 17: Filogenia basada en la morfología tradicional de los filos animales

El número de acceso a las 121 secuencias de *st3gal* identificadas y analizadas en este estudio se muestran como material suplementario en la sección 7.2 en la página 87.

4.4.2 *Red de Similitud de secuencias*

La visualización de redes de similitud permite explorar las relaciones entre secuencias y la formación de grupos de acuerdo a diferentes valores de corte definidos por el usuario. Cada relación entre secuencias se calcula a partir de un alineamientogenerado con BLAST de a pares de secuencias y se toma el puntaje de E-value como medida de similitud entre ellas. En la Figura 18 se presentan las redes obtenidas a diferentes valores de corte de E-value y se indican los grupos de subfamilias propuestos (GR1, GR2, GR3 y GRx).

El mayor grado de similitud se encuentra entre las secuencias pertenecientes al GR1, que incluye las secuencias de la nueva subfamilia ST3Gal VIII propuesta. Estas secuencias se agrupan incluso a valores de corte estrictos ($1e^{-90}$ y $1e^{-100}$) (ver Figura 18E y F). Las secuencias del GR2 forman grupos distintos para cada subfamilia utilizando valores de corte de E-values permisivos ($1e^{-60}$), destacando el grado de similitud entre sus miembros comparativamente bajo. Sin embargo, las secuencias de la subfamilia propuesta ST3Gal IX muestran una relación más estrecha con ST3Gal IV que con la subfamilia ST3Gal VI (Figura 18B). Por otro lado, las secuencias pertenecientes al grupo GR3 forman una subred separada a E-value$\geqslant 1e^{-70}$ y se evidencia que las secuencias de ST3Gal VII tienen mayor similitud con las secuencias de ST3Gal V, que con otras subfamilias de su grupo. Utilizando un valor de corte estricto, E-value $= 1e^{-100}$, las secuencias se separan en grupos de subfamilias distinguibles, a excepción de las secuencias del GR1 que permanecen conectadas formando un solo cluster. Además, las secuencias ST3Gal de peces forman sub-redes independientes, indicando que tienen un menor grado de similitud con las secuencias de su misma subfamilia. Valores de corte permisivos (E-values de $1e^{-55}$ a $1e^{-70}$) permiten distinguir la relación entre las secuencias de los grupos GR2 y GR3 (es decir ST3Gal IV / VI / IX y ST3Gal III / V / VII) y la aparición de los grupos individuales GR1 y GRx (ver Figura 18A-C). En cada panel de la figura 18, la ausencia de lineas entre el grupo de control (ST6Gal) y el resto las secuencias de ST3Gal permite la elección de los valores de corte de E-value que implica una relación de similitud. Las relaciones observadas entre las subfamilias y los grupos de subfamilias con las redes de similitud, concuerdan con los resultados obtenidos por análisis filogenético (Material suplementario Figura 36 en la página 86).

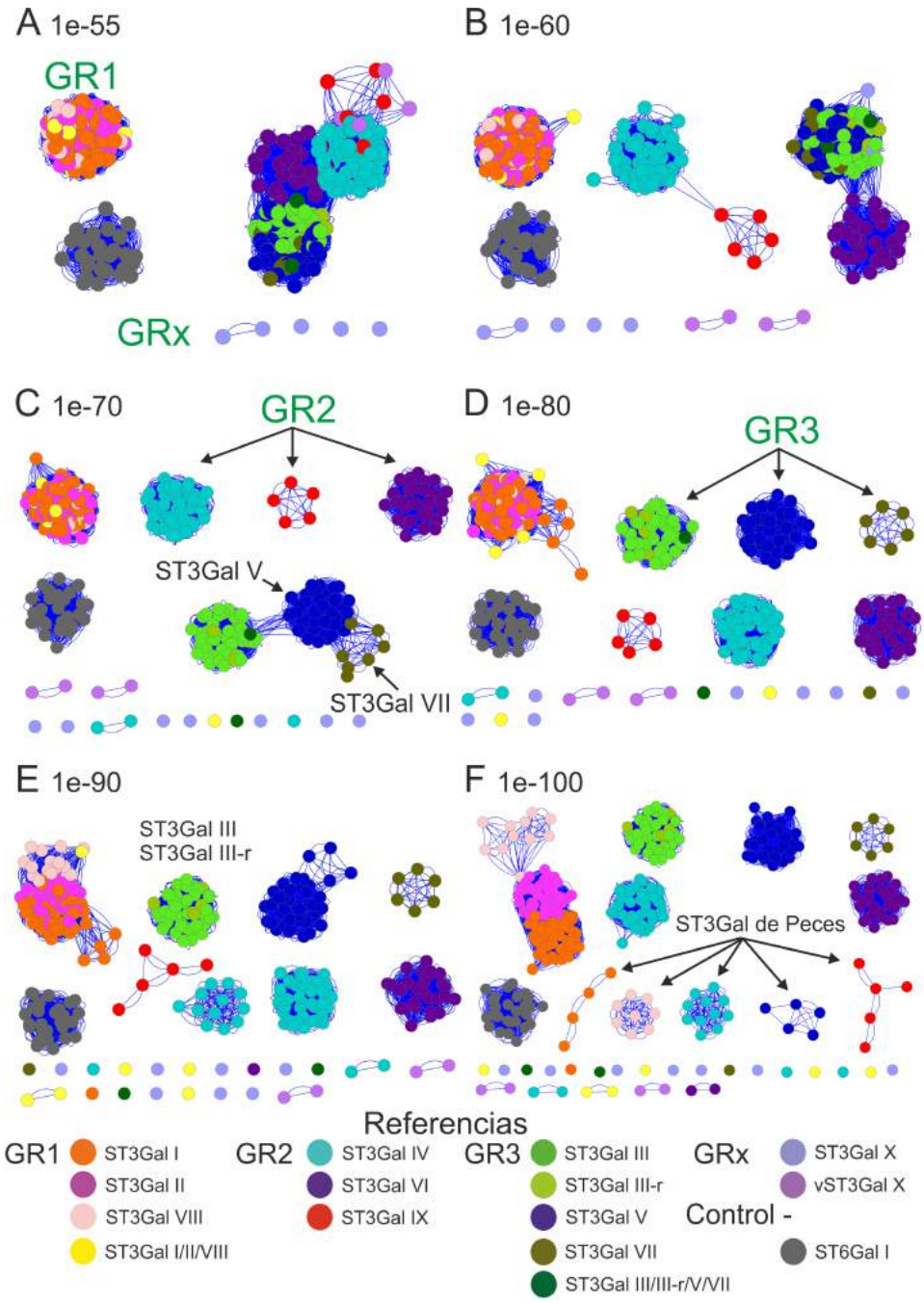


Figura 18: Red de similitud de secuencias

Cada secuencia es representada como un nodo y se definen las lineas entre ellas cuando el E-value es menor al valor de corte. Los nodos fueron coloreados según la subfamilia a la que pertenece la secuencia. En cada panel se representa la relación entre 336 secuencias de ST3Gal, más las secuencias de ST6Gal como control negativo, y se indica el valor de corte utilizado. A) Con valor de corte permisivo se observan dos cluster de secuencias de ST3Gal, uno formado por las secuencias de GR1 y el segundo por el resto de las secuencias de ST3Gal.B) Las secuencias de GR2 comienzan a formar cluster por subfamilia, separándose de las secuencias de GR3. C) Comienzan a separarse las secuencias de la subfamilia ST3Gal VII de ST3Gal V. D) Las secuencias de GR3 forman 3 clusters, permaneciendo unidas las secuencias de ST3Gal III, ST3Gal III-r. También se separa del GR1 las secuencias de ST3Gal I, II y VII. E) Las secuencias de ST3Gal IV forman dos cluster, correspondiendo uno a secuencias de peces. F) Comienzan a distinguirse las subfamilias del GR1, y todas las secuencias de peces forman nuevos clusters.

4.4.3 *Predicción de SDPs en ST3Gal*

El cambio en el grado de conservación de una posición en particular puede reflejar divergencia funcional luego de un evento de duplicación génica, debido a que una de las copias puede evolucionar bajo restricciones relajadas, lo que permite acumular cambios y desarrollar nuevas funciones y especificidades. Este mecanismo ha sido previamente estudiado y se distinguen dos tipos de divergencia funcional. El Tipo I hace referencia a las posiciones conservadas en un grupo de secuencias y variable en otro(s), mientras que las posiciones de Tipo II presentan diferentes aminoácidos conservados en cada grupo, como se muestra en la Figura 19.

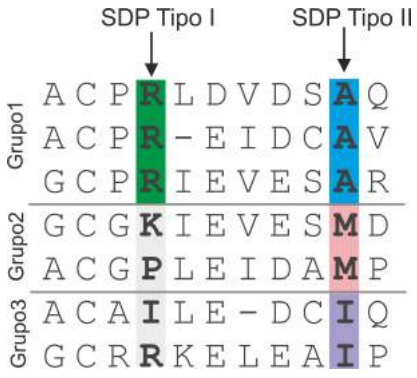


Figura 19: SDPs de Tipo I y Tipo II
Representación de un MSA donde se distinguen tres grupos de secuencias con diferente especificidad. Se resaltan en color los SDPs de Tipo I (posición conservada en grupo de secuencias y variable en los otros), y de Tipo II (residuos diferentes conservados dentro de cada grupo).

En la Figura 20 se muestra el MSA generado a partir de las secuencias consenso obtenidas con nuestro programa. El grupo GR1 se distingue del resto de los grupos, como se evidencia a partir de la presencia de numerosos SDPs de tipo I. Los SDPs fueron mapeados en la estructura tridimensional de referencia y se continuó el análisis con los 5 SDPs que se encuentran en el sitio activo de la proteína. En la tabla 6 se resumen los 5 SDPs predichos y su ubicación en la superficie proteíca se muestra en la Figura 21.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

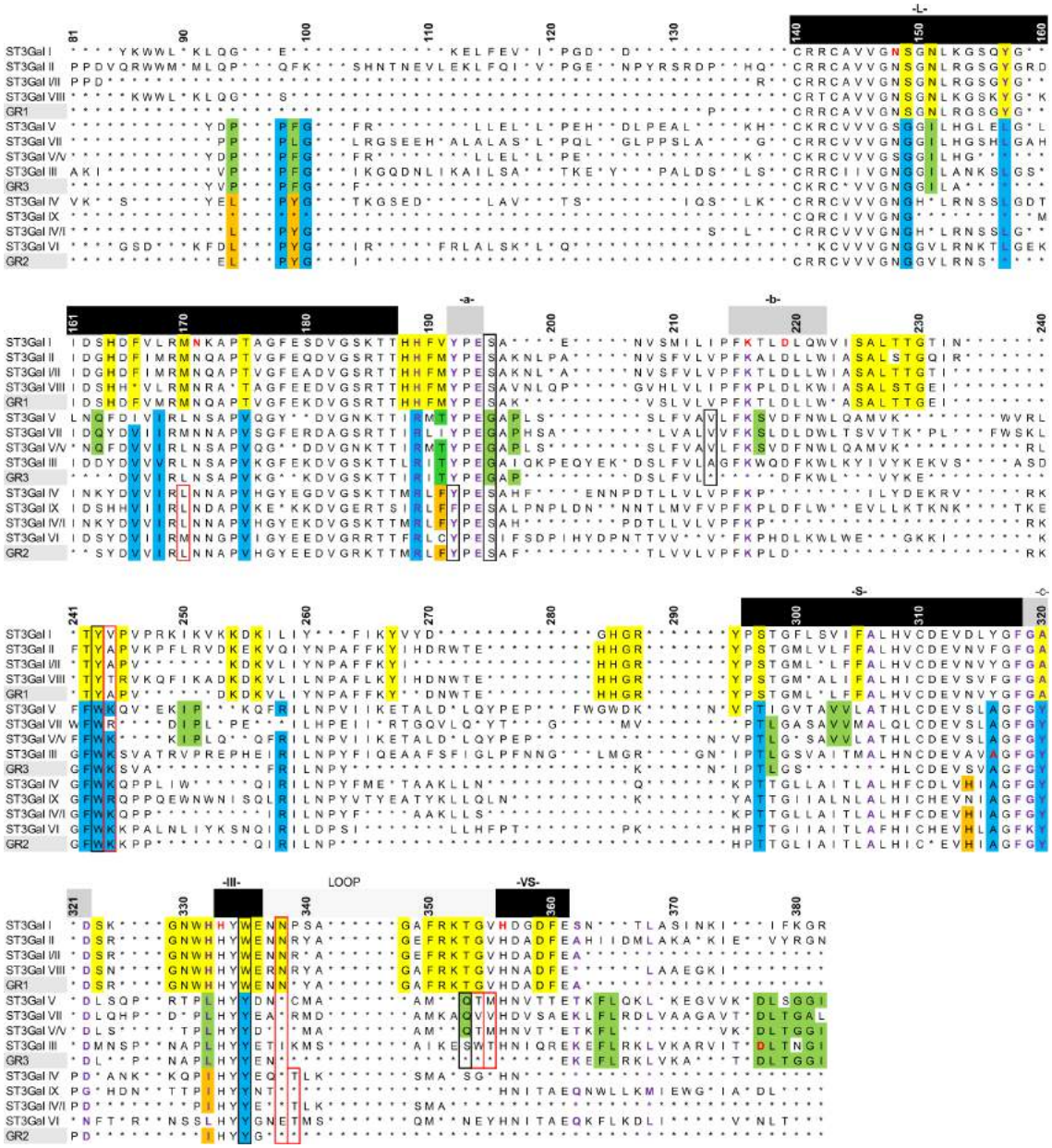


Figura 20: MSA usado para detectar SDPs

Para cada subfamilia se construyó una secuencia consenso que luego fue alineada utilizando MEGA 5.0. El MSA resultante es usado para detectar SDPs. Las posiciones marcadas con fondo amarillo identifican aminoácidos conservados dentro secuencias del GR1, con fondo naranja las del GR2, con fondo verde las del GR3, y con fondo azul las de GR2 y GR3. La banda superior muestra los motivos de superfamilia L, S, III y VS en negro y los motivos específicos de la familia ST3Gal a, b y c en gris. Para la secuencia consenso de la subfamilia ST3Gal I, se resaltan los aminoácidos implicados en unión al sustrato con fuente roja, demostrado por cristalografía de rayos X para ST3Gal porcina.

PDB	MSA	GR1	GR2	GR3	Descripción
S197	195	S	G	S	Próximo al sustrato donor
Y233	243	Y	W	W	Puente de H con Gal OH-6
V234	244	V/A/T	K	K/R	Próximo al sustrato aceptor
W304	335	W	Y	Y	Próximo al sustrato donor
N307	338	N	*	*	Próximo al sustrato donor

Tabla 6: Resumen de los SDPs predichos entre los grupos de secuencias de ST3Gal

Para los SDPs encontrados en el sitio activo de la enzima se indica en la primer columna la posición correspondiente en la estructura de referencia; en la segunda columna la posición correspondiente en el MSA; en las subsiguientes el residuo mayormente conservado para cada grupo, se marca con * si la posición es variable; en la última columna se incluye una descripción que puede relacionarse con la función de la posición.

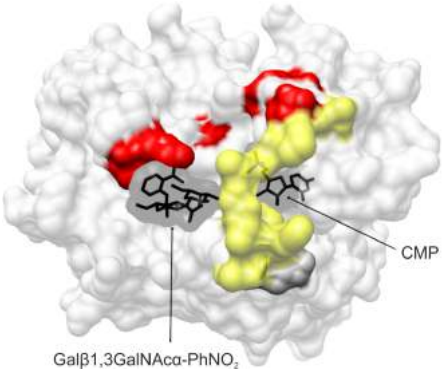


Figura 21: Ubicación de los 5 SDPs en la estructura de referencia

Representación en superficie de la estructura pST3Gal I. Se resalta el amarillo el dominio lid modelado, en rojo los 5 SDPs predichos de los grupos GR1, GR2 y GR3, los sustratos donor (CMP) y aceptor (Galβ1,3GalNAcα-$PhNO_2$) se representan en stick en color negro. El dominio flexible puede afectar el sitio de unión de ambos sustratos. Imagen generada con UCSF Chimera Pettersen et al. [97].

En la posición 195 del alineamiento, una Ser se encuentra altamente conservada para los grupos GR1 y GR3, mientras que para GR2 se encuentra conservada una Gly. En la estructura de referencia pST3Gal I, esta posición corresponde a la Ser-197 localizada a distancia de contacto del grupo fosfato del CMP (ver Figura 22). El grupo -OH de la Ser puede estar implicado en la formación de un puente de H contribuyendo a la estabilización del fosfato, mientras que la Gly no tendría ninguna implicancia en la estabilización. El residuo Tyr en la posición 243 (Y-233 en la estructura de referencia) está altamente conservado en el GR1, con la excepción de la secuencia perteneciente a *O. carmela* que presenta un Trp, al igual que las secuencias de los grupos GR2 y GR3. Este residuo Tyr está involucrado en la forma-

ción de un puente de H con el residuo de galactosa y fue reportado como determinante de especificidad del aceptor Rao et al. [102], Rakic et al. [101]. En la posición 244 se predijo un SDP de tipo I (V-234 en la estructura). Se trata de una posición variable en el GR1, mientras que para GR2 y GR3 se mantiene conservada una Lys. En la estructura pST3Gal, esta posición correspondiente al residuo V-234 se encuentra a distancia de contacto del hidroxi (2-hidroxifenil) oxo amonio (CG3), análogo del sustrato aceptor. Dos SDPs correspondientes a las posiciones 335 y 338 (residuos W-304 y N-307 en la estructura, figura D) fueron predichos dentro del dominio flexible lid. El anillo aromático del residuo en posición 335 puede participar en una estabilización pi-pi (pi-staking) con el sustrato donor. En la posición 338 se encuentra Asn conservada para el GR1, mientras que esta posición es altamente variable para el resto de los grupos.

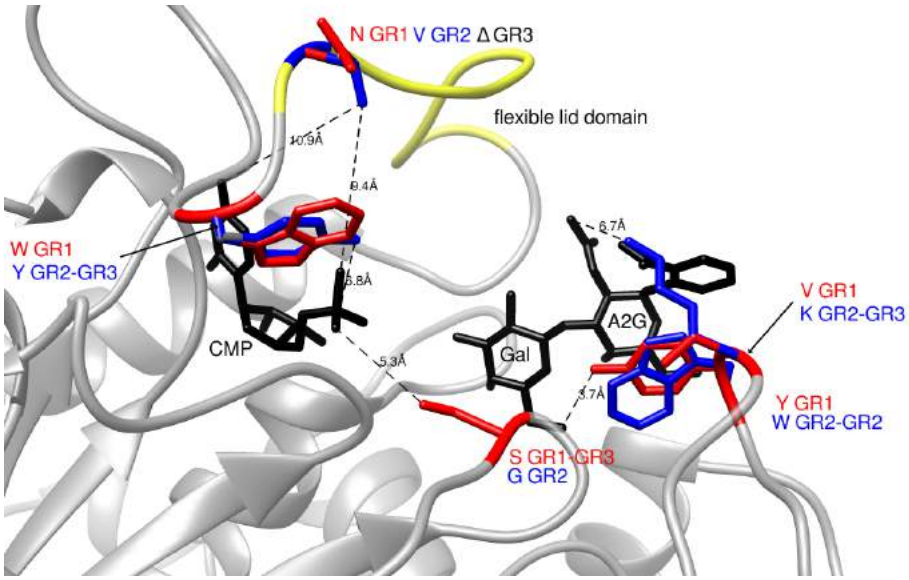


Figura 22: Representación del sitio activo con los 5 SDPs predichos para GR1, GR2 y GR3

El dominio lid modelado se muestra en amarillo, los sustratos aceptor y donor se representan como stick en color negro, los SDPs se representan como stick en rojo para GR1 y en azul para GR2, para GR3 se encuentra en rojo o en azul según coincida el aminoácido con los otros grupos. Para cada SDP se indica el aminoácido conservado y el grupo al cual pertenece. Por ejemplo, próximo al CMP se encuentra un W conservado para GR1 representado en rojo, mientras que una Y es conservada para GR2 y GR3 en la misma posición, y se encuentra en azul. Cuando no hay un residuo conservado para un grupo, la posición se señala con el símbolo Δ en negro. Imagen generada con UCSF Chimera Pettersen et al. [97].

Los SDPs predichos dentro de cada grupo de secuencias de ST3Gal de vertebrados (GR1, GR2 y GR3) se muestran en la tabla 7, para

GR2 y GR3 se encontraron dentro del sitio activo 3 y 4 SDPs respectivamente. En el alineamiento de secuencias de GR2, se encuentra conservado un residuo Leu en la posición 170 en ST3Gal IV y ST3Gal IX, mientras que una Met se encuentra conservada para ST3Gal VI. Esta posición se ubica cerca del sustrato donor en la estructura de referencia (T-129). En la posición 192, se encuentra conservada una Tyr en ST3Gal IV y ST3Gal VI, mientras que una Phe está conservada para ST3Gal IX. Aunque se trata de un cambio conservativo de residuos aromáticos, el cambio puede impactar en la especificidad ya que el O7 de la Tyr está involucrado en la estabilización del donor CMP-NeuAc Rao et al. [102], Rakic et al. [101] y la falta de un grupo OH en la Phe impide cumplir este rol. En la posición 304 del MSA, una Thr es conservada en las subfamilias ST3Gal IV y ST3Gal VI y una Asn en ST3Gal IX. Esta posición pertenece al dominio flexible lid y puede ubicarse cerca del sustrato donor.

Con respecto a los SDPs dentro del GR3, se encuentra conservada una Ser en la posición 353 en ST3Gal III y ST3Gal III-r, mientras que para ST3GalV y ST3Gal VII se preserva Gln. Este SDP corresponde a T-316 en la secuencia de referencia y se encuentra en el dominio lid. La Ala-213 esta conservada en ST3Gal III y ST3Gal-IIIr, mientras que se conserva Val en ST3Gal V y ST3Gal VII. Estos resultados muestran la similitud existente a nivel de residuos funcionales entre las subfamilias ST3Gal III y ST3Gal III-r por un lado, y las subfamilias ST3Gal V y ST3Gal VII por otro. Esta similitud concuerda con los resultados obtenidos por la red de similitud de secuencias. Se encontraron 2 SDPs en las posiciones 354 y 355 que presentan un residuo conservado diferente para cada subfamilia. En la posición 354 los aminoácidos W, T y V se encuentran conservados para las subfamilias ST3Gal III/ST3Gal III-r, ST3Gal V y ST3Gal VII respectivamente. Cabe destacar que 3 de los 4 SDPs predichos son consecutivos en secuencia y se encuentran en el dominio flexible lid. La comparación de las posiciones conservadas para las secuencias del GR1 muestran que la subfamilia ST3Gal VII tiene una mayor variación en las posiciones conservadas del GR1 que ST3Gal I y ST3Gal II. Sin embargo ninguna de estas posiciones se encuentran cercanas al sitio activo de la enzima.

4.4.4 *Posiciones coevoluciondas en ST3Gal*

La red de MI para la familia ST3Gal muestra que los puntajes más altos (top 10 % de MI) se encuentran en posiciones que pertenecen a los motivos de superfamilia (L, S, III o VS) o los motivos de familia (a, b ó c), como se observa en la Figura 23. Los puntajes más altos de cMI, que caracteriza el grado de información mutua acumulada para cada posición, se encuentra principalmente en las posiciones pertenecientes a los motivos de superfamilia. Las posiciones 318 y 324, por

PDB	MSA	Aminoácido consenso en GR1		
		ST3Gal I	ST3Gal II	ST3Gal III
E110	116	E	Q	*
P128	121	P	P	*
F168	166	F	F	*
T225	228	T	S	S
Y250	260	Y	Y	L
V258	268	V	I	I

PDB	MSA	Aminoácido consenso en GR2			Descripción
		ST3Gal IV	ST3Gal VI	ST3Gal IX	
T172	170	L	M	L	Cerca donor
Y194	192	Y	Y	F	Estab. Neu5Ac
P308	304	T	T	N	Lid, cerca donor

PDB	MSA	Aminoácido consenso en GR3			Descripción
		ST3Gal III	ST3Gal V	ST3Gal VII	
T316	353	S	Q	Q	Lid, cerca aceptor
G317	354	W	T	V	Lid, cerca aceptor
V318	355	T	M	V	Lid, cerca aceptor
V210	213	A	V	V	Cerca aceptor

Tabla 7: Predicción de SDPs dentro de cada grupo de secuencias (GR1, GR2 y GR3)

Se indica en la primer columna la posición de los aminoácidos en la estructura de referencia y en la segunda la posición en el MSA. Se indica con * si la posición es variable. En la última columna se detalla si pertenece al dominio Lid, si estabiliza una molécula o si se encuentra cerca de algún sustrato.

ejemplo, presentan valor alto de cMI esto podría indicar que poseen una importancia elevada dentro del motivo VS. Cabe destacar el alto valor de cMI de los residuos R-60, F-124, L-331 y F-340 a pesar de no pertenecer a ningún motivo descrito ya que pueden desempeñar un rol importante en el mantenimiento de la red de coevolución, y su importancia no puede ser encontrada mediante la medición de la conservación. Se consideraron para el análisis las 10 posiciones con mayor puntaje de cMI, la sub-red MI resultante se muestra en la figura 24. El residuo G-273 muestra el mayor valor de cMI y el mayor número de interacciones de MI (55 líneas de la red completa). Curiosamente, las diez posiciones con mayor cMI forman un clique, es decir una red donde cada residuo está conectado a todos los otros (9 líneas en todos los casos)(Figura 24A). Cada residuo en esa red pertenece a un motivo, 8 pertenecen a los motivos de superfamilia L, S y III, y 2 pertenecen a los motivos de familia b y c. El residuo N-173 comparte el valor más alto MI con Y-303, ambos localizados en el sitio activo de la enzima. N-173 está localizado cerca de sustrato donor y participa en la estabilización de grupo fosfato, tal como se describe anteriormente Rao et al. [102]. El amino ácido D-216 pertenece al motivo de familia b y se ha informado de interactúa con Gal-OH-6. Esta posición es la más distante en la estructura 3D del resto de los residuos de la red; y posee sus valores más altos de MI con las posiciones C-145, G-160, y G-185 del motivo de superfamilia L y C-284 perteneciente al motivo S.

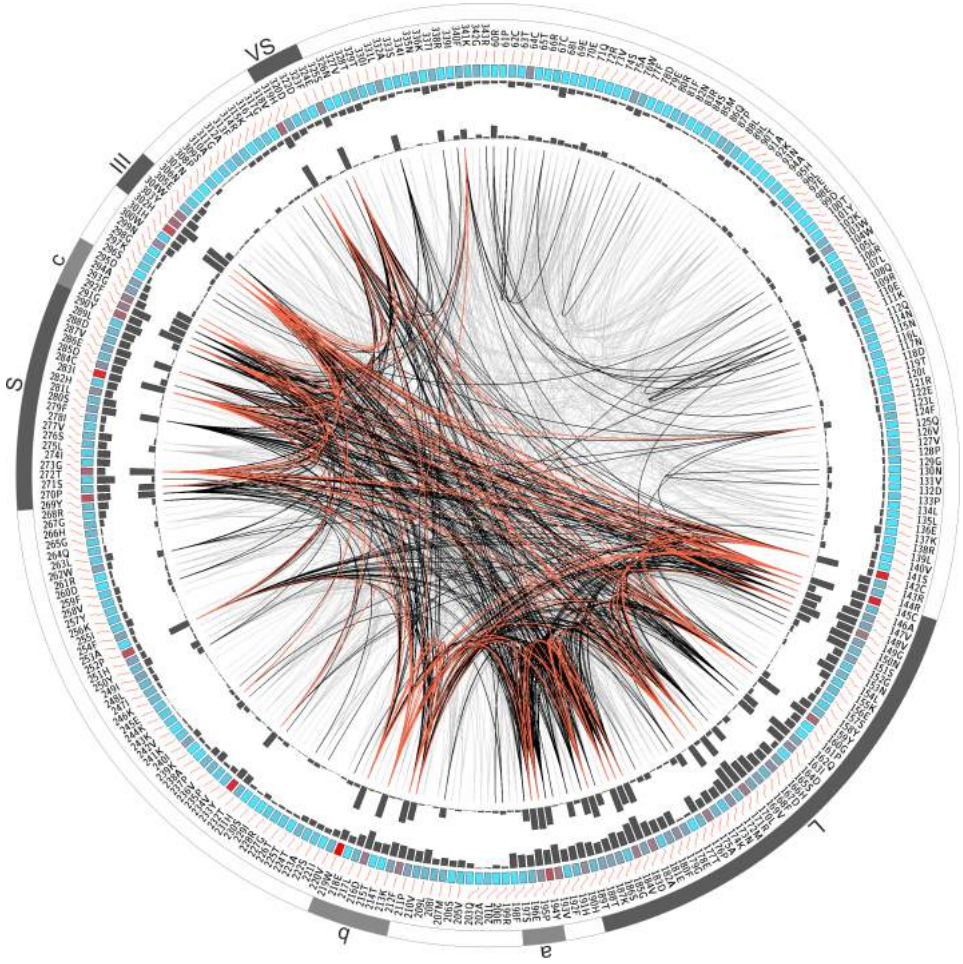
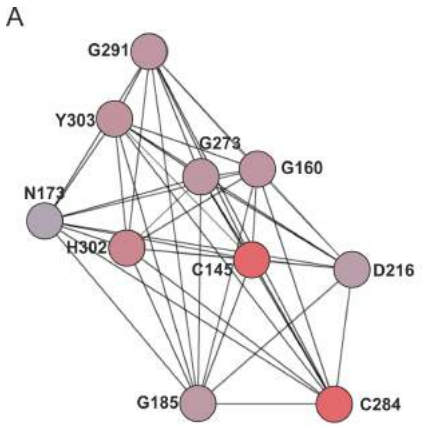


Figura 23: Representación en circos de la MI, cMI y pMI de ST3Gal
La secuencia se encuentra representada de forma circular. La información contenida en cada círculo, yendo del exterior al interior de la figura es la siguiente: el círculo exterior indica la posición de los motivos de secuencia de la superfamilia Sialiltransfersa (en gris oscuro), y los motivos de la familia ST3Gal (en gris claro). El siguiente círculo indica la identidad y el número de resiuo según la numeración del PDB de la secuencia de referencia. En el siguiente círculo se encuentran rectángulos coloreados según el grado de conservación de la posición, la escala de mayor a menor va del rojo al cian. El cuarto y quinto círculo representan como histogramas los puntajes de cMI y pMI respectivamente. En el centro del círculo se encuentran líneas que conectan pares de posiciones con puntaje de MI significativo (MI >6.5). En rojo, el 5 % de mayor puntaje de MI (top 5 %), en negro el siguiente 25 % mayor y en gris el 70 % restante. Imagen generada con el servidor MISTIC Simonetti et al. [115].

Figura 24: Red de alta MI en ST3Gal y mapeo en la estructura 3D



A) Los nodos representan aminoácidos y las líneas entre ellos valores de MI significativos. La longitud de las líneas es inversamente proporcional al valor de MI. Los nodos están coloreados por el puntaje cMI en escala de violeta a amarillo, de mayor a menor.

B) Posiciones predichas como coevolucionadas mapeadas en la estructura pST3Gal. Se representa en esferas rojas los Cα de los 10 residuos con mayor puntaje cMI. Imagen generada con UCSF ChimeraPettersen et al. [97].

Los resultados obtenidos también incluyen el análisis de sintenia y paralogía conservada alrededor de los genes *st3gal* en diferentes especies animales (resultado no mostrado). Así mismo se estudió el patrón de expresión de los genes *st3gal* en varios organismos modelos a partir de información recuperada de la base de datos Expressed Sequence Tag y complementada con experimentos en *B. taurus* y *D. rerio* (resultados no mostrados). El conjunto de resultados nos permitió proponer un modelo evolutivo de lo genes *st3gal* en animales que se incluyen como material suplementario Figura 37 en la página 87.

4.5 DISCUSIÓN

En los últimos años numerosos estudios se dedicaron a determinar el dominio catalítico de ST3Gal Rao et al. [102], Vallejo-Ruiz et al. [126] y de identificar y caracterizar firmas funcionales de las sialiltransferasas, los motivos llamados L, S, III y VS mediante mutagénesis dirigida a sitios conservados y análisis cinéticos del sitio activo de las mutantes Takashima et al. [119], Rakic et al. [101], Audry et al. [7]. Recientemente una mutación con cambio de sentido (missense)[2] en el gen humano *ST3GAL3* situada por fuera de los motivos de superfamilia y familia descriptos ha sido asociada con Discapacidad Intelectual no sindró-

2 Mutación con cambio de sentido: mutación puntual no sinónima en la cual se produce un cambio en un único nucleótido, provocando la aparición de un codón que codifica para un aminoácido diferente

mica Hu et al. [60] (donde la discapacidad intelectual es el único rasgo evidente de la enfermedad) y otra mutación dentro del motivo de superfamilia S causa encefalopatía epiléptica llamada síndrome de West Edvardson et al. [41]. En este estudio se predice la existencia de posiciones conservadas en grupos de secuencias de ST3Gal que pueden estar implicadas en aspectos críticos de la función general de la proteína y diferencial en la especificidad de los sustratos donor y aceptor. Se realizó un primer análisis de SDPs entre los grupos de ST3Gal de vertebrados GR1, GR2 y GR3 identificando 5 SDPs localizados cerca del sitio activo de la enzima. Los primeros 2 SDPs corresponden a S-197 e Y-233 en la estructura de referencia pertenecen a los motivos L y S respectivamente. Los otros 3 (V-324, W-304 y N-307) se localizan dentro del loop (dominio lid) no observado por cristalografía de rayos X y modelado en este trabajo. Se cree que forma parte del sitio de unión y que cambia de conformación cerrándose sobre la estructura (como una tapa) acomodando el sustrato donor Rao et al. [102]. A partir de este estudio, se sugiere que estos 3 SDPs desempeñan un rol de modulación de la especificidad de los sustratos donor y aceptor y de la actividad enzimática de ST3Gal durante la evolución en Vertebrados. Es importante notar que ST3Gal III y ST3Gal IV de mamíferos que muestran el mismo requerimiento hacia Gal OH-6 Rohfritsch et al. [106] presentan el mismo SDP W-243 en el MSA, mostrado en la Figura 22, una posición que se sugiere implicada en la formación de puente de H con Gal HO-6 Rao et al. [102].

También se predijeron los SDPs para cada grupo de secuencias de ST3Gal de vertebrados. Es sabido que ST3Gal I y ST3Gal II pertenecientes al grupo GR1, son los principales responsables de sialilación de O-glicanos y glicolípidos en mamíferos. Debido a que no fueron encontrados SDPs localizados en el sitio activo dentro del grupo GR1 (tabla 7), se sugiere que esta nueva subfamilia descripta ST3Gal VIII, conserva la misma actividad enzimática que ST3Gal I y II.

En el grupo GR3, ST3Gal III y ST3Gal V de mamíferos y ST3Gal VII de pez presentan actividades enzimáticas muy diferentes. ST3Gal III usa principalmente como sustrato aceptor disacáridos de tipo I (Galβ1-3GlcNAc) encontrados en glicoproteínas y disacáridos de tipo III (Galβ1-3GalNAc) encontrados en glicolípidos cerebrales Sturgill et al. [118]. Por otro lado ST3Gal V (G_{M3}sintasa) usa casi exclusivamente lactosil-ceramida (Galβ1,4Glc-Cer) y la nueva subfamilia descripta llamada ST3Gal VII, utiliza Gal-Cer conduciendo a la formación de GChisada et al. [26]. Es interesante que se encontraron 4 SDPs cerca del sustrato aceptor, de los cuales 3 forman parte del loop funcional, lo que podría explicar la divergencia funcional de estas enzimas, aunque aún no se han evaluado experimentalmente.

Finalmente en el grupo GR2, ST3Gal IV y ST3Gal VI usa disacáridos de tipo I o tipo II (Galβ1-3/4GlcNAc) dirigiendo la formación de epitopos sialilados de Lewis encontrados en glicoproteínas y gli-

colípidos expresados en la superficie celular.Se espera que la nueva subfamilia descripta ST3Gal IX presente la misma actividad enzimática. Se describieron 3 SDPs en el GR3, todos ellos localizados cerca del sustrato donor y uno de ellos dentro de motivo de familia a que fue propuesto como proveedor de especificidad del donor del ácido siálico Rao et al. [102].

El análisis de coevolución revela que existen sitios específicos con alta MI dentro de los motivos de superfamilia (L, S o VS) y de los motivos de familia sugiriendo que, al menos parte de estos motivos han evolucionado de manera concertada. Este análisis identifica residuos distantes en la secuencia proteica que podrían haber evolucionado bajo una presión de selección en común. Debido a que cada uno de los motivos descriptos presenta residuos con alta MI, el resultado puede interpretarse como un grupo de motivos altamente conservados unidos por una red de residuos que coevolucionan. Los 10 residuos con mayor cMI forman un Clique, lo que podría indicar un alto nivel de redundancia y cohesividad, cada residuo estaría co-evolucionando junto con los otros. Todos estos residuos pertenecen a los motivos previamente descriptos y 2 de ellos se encuentran implicados en interacción con los sustratos. Aunque es difícil establecer conclusiones generales a cerca del rol funcional de los residuos que coevolucionan, los puntajes de MI identifican residuos funcionalmente importantes que son diferentes a los detectados por conservación o conservación dentro de grupos (SDPs).

5

ESTUDIO DE LA INTERFAZ DE INTERACCIÓN PROTEÍNA-PROTEÍNA

5.1 RESUMEN

En el presente trabajo se utiliza un set de datos comprensivo de interacción proteína-proteína basado en estructuras tridimensionales resultas de las cuales se generó un MSA. Se analizaron las secuencias de las unidades de interacción con medidas de conservación, MI, accesibilidad al solvente y otras derivas de ellas. Como resultado del estudio no se encontró ninguna medida capaz de distinguir los residuos de interacción del resto de los residuos de la proteína o de la superficie.Sin embargo los resultados obtenidos aportan información que puede ser valiosa para el desarrollo de futuros métodos predictivos.

5.2 INTRODUCCIÓN AL ESTUDIO DE INTERACCIÓN PROTEÍNA-PROTEÍNA

Uno de los desafíos de la era post-genómica es la caracterización e identificación de las interacciones proteína-proteína en los organismos, debido a su importancia durante el comportamiento normal y patológico de las células. Se estima que el interactoma humano contiene entre 100.000 y 600.000 interacciones, de las cuales fueron descritas experimentalmente cerca de 50.000. Es decir que la diferencia entre las interacciones experimentalmente validadas y las predichas es como mínimo del 50 %, siendo la proporción mucho mayor para especies menos estudiadas. Por este motivo la predicción computacional y caracterización de las interacciones proteína-proteínas es un área de gran interés y continuo desarrollo.

Uno de los puntos de partida para la caracterización de un complejo proteico, es la descripción de su superficie de contacto; es decir, la interfaz de interacción. Una interfaz de interacción queda definida por los contactos físicos establecidos entre dos o más moléculas, típicamente se distingue del resto de los residuos de superficie a partir del ΔASA entre el monómero y el complejo. En la Figura 25 se muestra a modo de ejemplo la estructura de un complejo proteico y su interfaz de interacción.

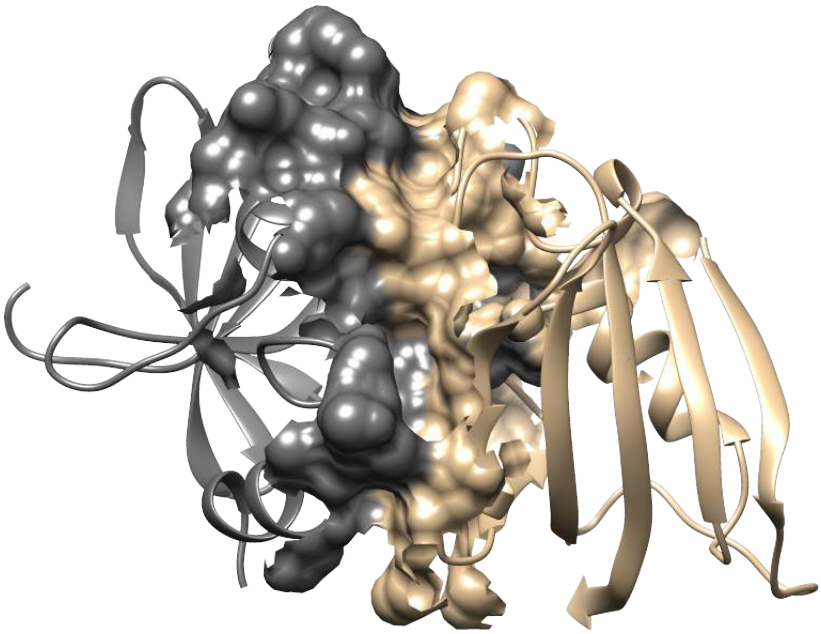


Figura 25: Interfaz de interacción proteína-proteína
Se muestra en representación de cinta la estructura del complejo y como superficie la región de interacción.

Bogan y Thorn notaron que dentro de los residuos de la interfaz, había algunos que contribuían más a la energía de unión y éstos se encontraban ocultos al solvente en la estructura del complejo Bogan and Thorn [14]. Propusieron un modelo donde la interfaz queda compuesta por una región central desolvatada rodeada por residuos energéticamente menos importantes, que quedan en contacto con el agua. Este modelo coincide con otro modelo que fue propuesto en paralelo, donde se postula que la interfaz queda compuesta por un Core de átomos ocultos bordeado por otros accesibles Lo Conte et al. [80]. A partir de estas descripciones se propuso el modelo core-rim Bahadur et al. [8], Chakrabarti and Janin [21], donde los residuos que posean al menos un átomo completamente oculto en el complejo pertenecen a la región core, y los residuos restante de la interfaz, que mantienen una accesibilidad parcial al solvente durante la interacción, son asignados a la región rim.

La distinción entre estas dos regiones de la interfaz se ejemplifica en la Figura 26.

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

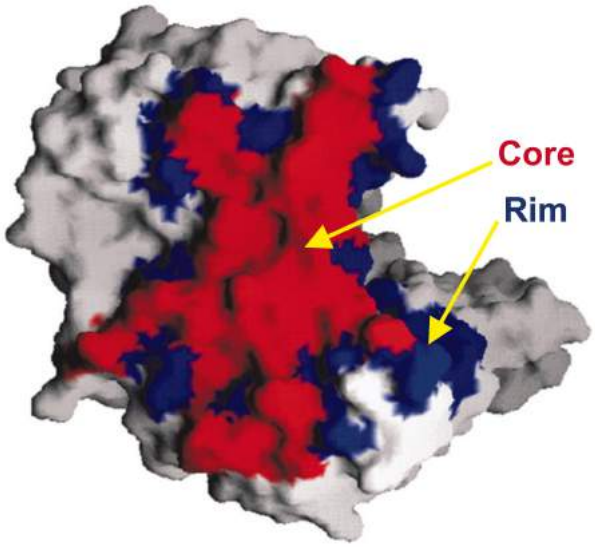


Figura 26: Región core y rim para una cara de una interfaz de interacción La región core (rojo) es más grande que la rim (azul), típicamente representa el 77 % de la interfaz Bahadur et al. [8]

Se ha descripto que la región core de la interfaz posee mayor cantidad de aminoácidos hidrofóbicos que lo distinguen de la región rim y del resto de la superficie. Desde el punto de vista funcional, los residuos del core de la interfaz y del rim tienen diferentes contribuciones a la energía de unión y consecuentemente diferentes presiones evolutivas. La coevolución sitio específica entre proteínas que interaccionan fue mayormente detectada en la región rim Travers and Fares [123], Yeang and Haussler [132], Kann et al. [66], mientras que el core se mantiene mayormente conservado Hakes et al. [51], Eames and Kortemme [40], Mintseris and Weng [87], Guharoy and Chakrabarti [50], Caffrey et al. [17]. Estos datos de conservación concuerdan con la mayor presencia de los llamados residuos "hot spot", en el core de la interfaz. Los residuos "hot spot" son lo que contribuyen de manera dominante a la energía libre de unión ($\Delta\Delta G$>2Kcal/mol) y pueden entorpecer la interacción al ser mutados por alanina. La interfaz de interacción típicamente constituye el 10 % de los residuos de una proteína, y solo una minoría de éstos contribuye significativamente a la energía de unión Clarke [28].

El objetivo de este trabajo es estudiar el grado de conservación y de coevolución de los residuos responsables de la interacción proteína-proteína en un set de datos comprensivo. Para ello se desarrollaron diferentes puntajes derivados de la conservación, MI y del área accesible al solvente.

5.3 MATERIALES Y MÉTODOS

5.3.1 *Set de datos*

PICCOLO Bickerton et al. [11] es una base de datos comprensiva que captura los detalles estructurales de la interacción proteína-proteína a nivel atómico. Esta construida a partir de todas las estructuras depositadas en el PDB que presentan más de una cadena polipeptídica y que haya sido resueltas por cristalografía de rayos X. Se define como unidad de interacción una cadena de PDB, y cada interacción es estudiada de a pares de cadenas.

La identificación de los residuos de interacción fue definida en dos pasos:

1) Se tomó un valor umbral de distancia de 6.05Å (máxima distancia de un puente de hidrógeno mediado por agua Robert and Janin [104]) para identificar los átomos presentes en las diferentes cadenas polipeptídicas que se encuentran próximos entre sí y se los considera potencialmente interactuantes. Aunque la definición de contactos usando un valor umbral de distancia es muy extendida, este método es considerado sensible pero no específico debido a que muchos átomos próximos pueden no estar implicados de manera directa en una interacción energéticamente significativa, por ello se implementó el paso siguiente.

2) Se evaluó si los átomos próximos, hallados en el punto anterior, pueden efectivamente participar en la interacción de manera directa y energéticamente significativa. Para ello se consideró la naturaleza química de cada par de átomos, sus radios atómicos y definiciones precisas de interacciones moleculares, que incluyen criterios de ángulos y distancias.

El proceso completo provee un set de interacciones átomo-átomo intermoleculares bien definidos.

Debido a la redundancia del PDB se realizó un agrupamiento considerando el par de interacción para generar un set de interfaces no redundantes. Primero las interfaces fueron agrupadas por el identificador de UniProt de ambos componentes. Luego se compararon los residuos de interacción entre aquellos pares con identificadores idénticos. Si ambas caras de las interfaces compartían más del 75 % de las posiciones de interacción, entonces las interfaces eran consideradas redundantes y asignadas a un mismo grupo. Finalmente para cada grupo se eligió un miembro representativo según la calidad de la estructura, la resolución del cristal y el números de residuos ausentes en el PDB.

Las coordenadas atómicas depositadas en los archivos del PDB resueltos por cristalografía de rayos X refleja el contenido de una unidad asimétrica (el mínimo grupo de átomos que generan la celda unidad). Aunque la unidad asimétrica puede representar el ensamblado

biológico funcional de la proteína, frecuentemente contiene múltiples moléculas biológicas o incluso una porción de la molécula biológica. Por este motivo existen dos versiones de PICCOLO, una de ellas construida a partir de las unidades asimétricas del PDB y otra a partir de los ensamblados biológicos.

En el presente trabajo se utilizaron todos los hetero complejos de PICCOLO, de la versión construida a partir del ensamblado biológico generado por PISA, ya que son más propenso a reflejar el ensamblado oligomérico de mayor relevancia biológica. El set de datos obtenido de esta forma está compuesto por 2471 cadenas de PDB. La base de datos contiene las anotaciones de los residuos próximos a la interfaz y los residuos de interacción proteína-proteína. Adicionalmente se le solicitó a los autores las estructuras de los complejos con los que se construyó la base de datos, ya que previamente debieron solucionar inconsistencias propias del PDB en un proceso que llamaron de "saneamiento". Algunos de los problema resueltos fueron:

A. Estructuras con múltiple ocupancia de átomos

B. Modelos múltiples de RMN

C. Numeración de residuos con inserción alfabética

D. Ausencia de residuos

E. Estructuras con resolución solo a nivel de Cα

F. Contiene en total más de 300 aminoácidos no estándar (naturales o modificados)

G. Estructuras de baja resolución con tipo de residuo no asignado

H. Identificación de cadena fuera del formato estándar (numéricas y mayúsculas)

5.3.2 *Alineamiento Múltiple de Secuencias*

Para realizar los alineamientos se extrajeron las secuencias de cada unidad de interacción y se descartaron aquellas con una longitud menor a 50 aminoácidos. Con cada secuencia se realizó una búsqueda en la base de datos UniRef90 con el algoritmo phmmer. Se consideró la secuencia extraída de PICCOLO como referencia y se quitaron las columnas con gaps preservando su integridad. Adicionalmente fueron removidas las secuencias con una cobertura menor al 50 % de la secuencia de referencia, quedando el set de datos compuesto por 2491MSAs.

5.3.3 *Cálculos de puntajes por residuo*

A partir de los MSAs se calculó el puntaje de conservación Kulback-Leibler para cada posición y el puntaje obtenido se corrigió para que perteneciera al rango [0-1] como se muestra en la ecuación 19. Adicionalmente se calculo la MI entre pares de posiciones, y la cMI para cada posición y se corrigió de manera análoga.

$$P = \frac{P - Pmínimo}{Pmáximo - Pmínimo} \tag{19}$$

Para este análisis los puntajes de proximidad pMI y pC fueron calculados como se indica en la ecuación 20. Notar que el puntaje para un residuo dado incluye la información de los residuos próximos pero excluye la información de sí mismo.

$$pMI_i = \sum_{j, d_{ij} < u} cMI_j \tag{20}$$

Donde la suma se realiza sobre todos los residuos j de una proteína dada dentro de una distancia $d_{ij} < u$, d_{ij} es calculada como la distancia mínima entre cualquier par de átomos diferentes de H de los residuos i y j siendo $i \neq j$.

También se generaron dos nuevos puntajes llamados pMIexpuestos y pCexpuestos (pMIexp y pCexp) donde se agrega como condición para la sumatoria que el residuo posea una valor de RSA > 0.2.

$$pMIexp_i = \sum_{j, d_{ij} < u, RSA_j > 0{,}2} cMI_j \tag{21}$$

Con el programa DSSP Kabsch and Sander [63], Joosten et al. [62] se cálculo la área accesible al solvente de cada residuo en el monómero. Es necesario considerar que los residuos tienen diferentes dimensiones y por lo tanto diferentes áreas accesibles al solvente. Por ejemplo, triptófano es mucho más grande y tiene una superficie mucho mayor que la alanina. Por este motivo se relativizó el puntaje obtenido con DSSP al tripéptido de Alanina A-X-A, para calcular el Área Relativa Accesible al Solvente (RSA). Con respecto al valor de corte del RSA para distinguir residuos de superficie, existen diferentes criterios, algunos autores considera que un residuo se encuentra expuesto si su RSA es mayor a 0.25 , 0.05 Zellner et al. [133], y 0.16 Rost and Sander [107]. En este trabajo utilizamos 0.2 como valor intermedio entre los diferentes valores de literatura.

5.3.4 *Optimización de parámetros*

Para los puntajes de proximidad pMI y pC se evaluó el desempeño predictivo de los puntajes calculados a valores de corte de distancia

entre 5-10Å, con pasos de 0.5. No se encontraron diferencias significativas en el desempeño predictivo.

Utilizando una grid de entre 0-1 con pasos de 0.1 se optimizaron los pesos del modelo que se muestra en la ecuación 22

$$S = (1 - w_1 - w_2 - w_3 - w_4)RSA + w_1 C + w_2 cMI + w_3 pC + w_4 pMI \tag{22}$$

Donde S representa el puntaje resultante de la combinación pesada de los diferentes puntajes, w son los pesos de cada término y $w_1 + w_2 + w_3 + w_4 = 1$.

Se evaluó el modelo análogo con los puntajes pMIexp y pCexp.

5.4 RESULTADOS

Análisis del set datos

En este trabajo se utilizan datos de interacción extraídos de la base de datos PICCOLO, donde se define como unidad de interacción una cadena de PDB y las interacciones son estudiadas de a pares de cadenas. Se trabajó con el subset de interacciones de heterocomplejos, compuesto por 2471 cadenas de PDB. PICCOLO provee un set de datos no redundante a nivel de interfaces, pero redundante a nivel de dímeros y de unidades de interacción. Es decir que un mismo par de unidades de interacción, puede interactuar más de una vez, definiendo diferentes interfaces. En la Figura 27 se muestra como ejemplo las diferentes interfaces definidas para un mismo par de interacción correspondiente al PDB 3LKT.

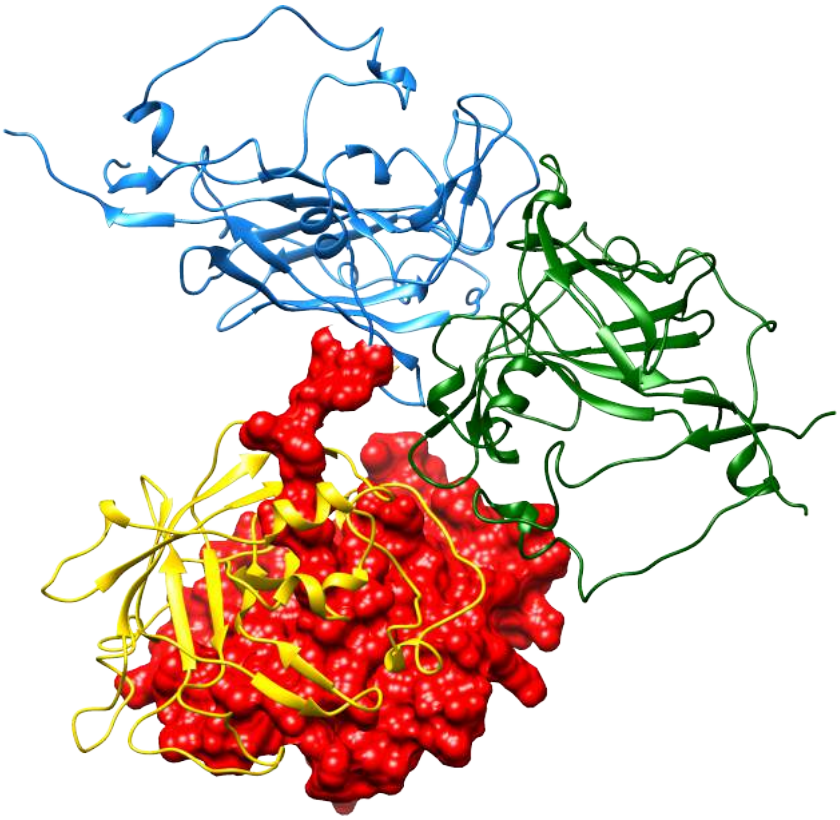


Figura 27: Diferentes interfaces de interacción definidas por un mismo par de unidades de interacción
La cadena de PDB representada como superficie en rojo (Uniprot Accession Number P00436) interactúa con 3 cadenas idénticas entre sí representadas en ribbon en azul, verde y amarillo (todas corresponden al Uniprot Accession Number P00437), definiendo tres interfaces de interacción diferentes.

En la Figura 28 se muestra el número de ocurrencias de un mismo par de unidades interactuantes en la base de datos, es decir el número de interfaces de interacción distintas definidas para cada par de unidades de interacción.

Otra característica a considerar de los datos es el número de veces que una misma cadena de PDB, participa en diferentes interacciones, esta vez considerando diferentes compañeros de unión. En la Figura 29 se muestra el número de veces de que una cadena de PDB participa en una interacción.

Definición de residuos de interacción

Dado que la base de datos PICCOLO provee dos niveles de distinción de residuos de interfaz, en este trabajo se llamarán residuos próximos a aquellos que satisfagan la primer condición (distancia entre átomos <6.05Å); y residuos de interacción a aquellos que además de ser próximos poseen las características necesarias para interactuar de manera directa. En la Figura 30 se muestran las distribuciones de

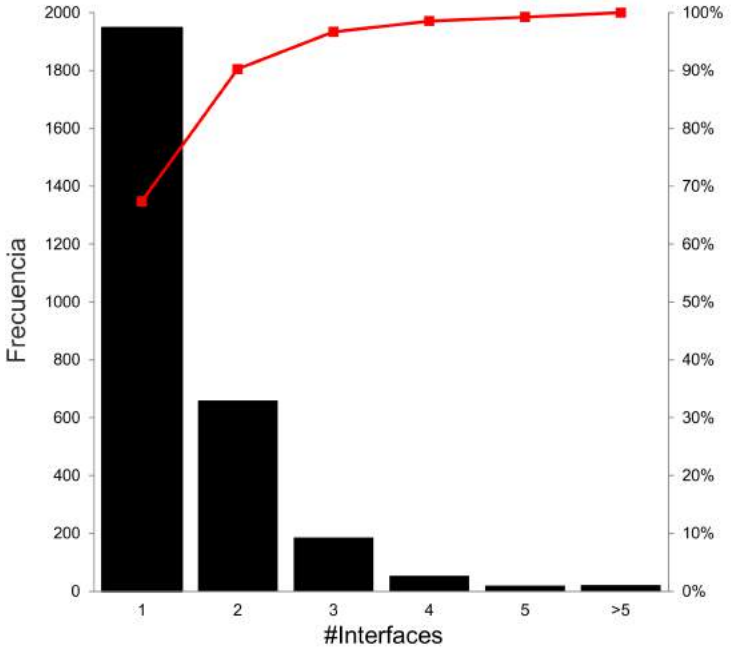


Figura 28: Histograma del número de interfaces definidas vs frecuencia de heterodímeros en PICCOLO
La línea gris clara representa el porcentaje acumulado.

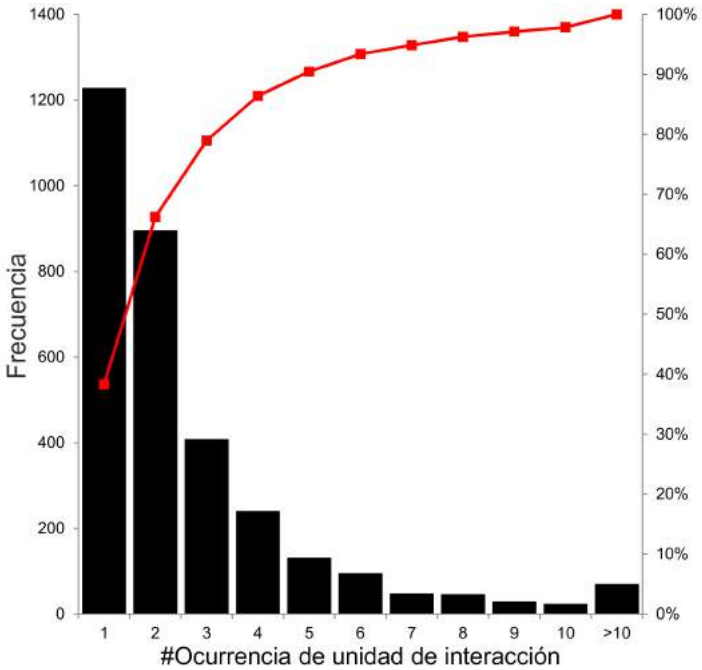


Figura 29: Histograma de ocurrencia de unidades de interacción de heterodímeros de PICCOLO
La línea gris clara representa el porcentaje acumulado.

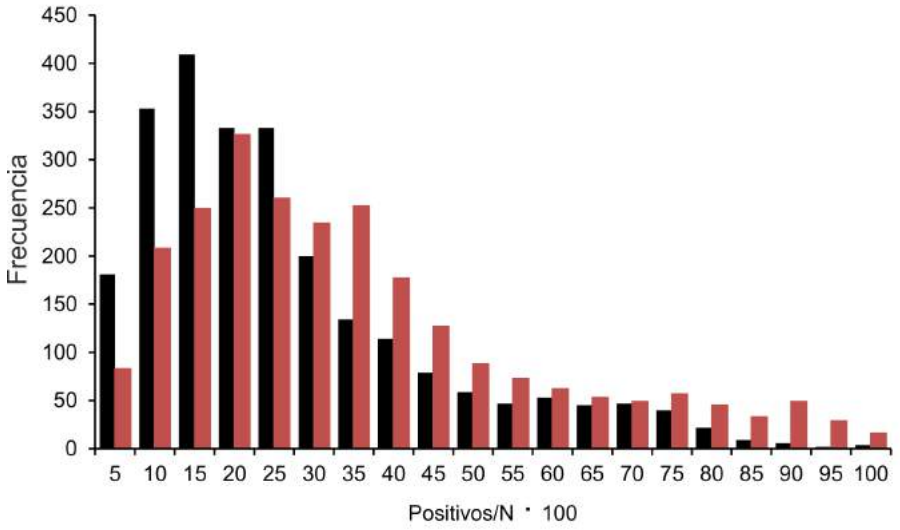


Figura 30: Histograma de positivos sobre el total de residuos para heterodímeros de PICCOLO
Los residuos positivos fueron considerados utilizando dos criterios: los de interacción (en negro) y los de proximidad (rojo). N representa el número total de residuos de la unidad de interacción.

ambos tipos de residuos de interfaz para todos los heterodímeros de PICCOLO.

Puntaje RSA y derivados de conservación y de MI en la interfaz de interacción

Tomando como positivos a los residuos de interacción y como negativos a todo el resto de los residuos de la base de datos, se evaluó la distribución de los diferentes puntajes, el resultado se muestra en la Figura 32. Se encontraron distribuciones similares entre los residuos de interacción y el resto de los residuos de las proteínas, para cada puntaje ensayado.

Por otro lado, es sabido que los residuos de interacción se encuentran en la superficie proteica por lo que su RSA será considerable. La distribución de la RSA se muestra en la Figura 31.
A fin de integrar la información contenida en el puntaje de RSA, se analizó la distribución de los puntajes considerando solamente los que presentaran un RSA > 0.2 (solo residuos de superficie). El resultado se muestra en la Figura 33. Se observa que las distribuciones son similares para ambos tipos de residuos, en todos los puntajes ensayados. Para el puntaje pCexp se puede observar una mayor abundancia de residuos de interacción a partir de cierto valor de corte, pCexp > 0.4.
Adicionalmente se evaluó el desempeño predictivo de cada puntaje en términos de AUC para la predicción de residuos de interacción, los resultados se resumen en la tabla 8. Los puntajes de proximidad

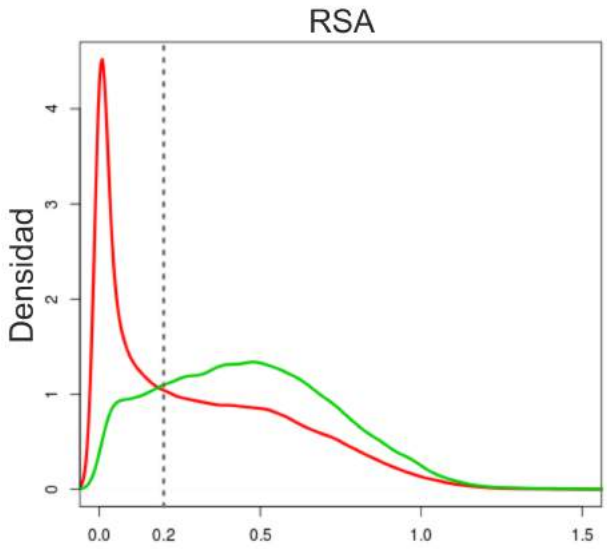


Figura 31: Plot de densidad para RSA
La línea verde representa los residuos de interacción (positivos) y la roja el resto de los residuos de las unidades de interacción (negativos). La línea puntada indica el criterio utilizado para considerar un residuo expuesto (RSA = 0.2).

pC y pMI mostraron un desempeño predictivo cercano a 0.5, sin diferencia estadísticamente significativa para los diferentes valores de corte de distancia testeados. Todos los puntajes ensayados presentan un desempeño predictivo cercano al random.

Combinación de diferentes puntajes

A fin de evaluar si los puntajes proporcionan información complementaria, se ensayó la combinación de los puntajes en el modelo que se muestra en la ecuación23.

$$(1 - w_1 - w_2 - w_3) \cdot RSA + w_1 \cdot C + w_2 \cdot pC + w_3 \cdot cMI + w_4 \cdot pMI \tag{23}$$

En el modelo combinado, para los puntajes de proximidad se utilizó un valor de corte de 6.0Å.

En la ecuación 24 se muestra el modelo con mejor desempeño predictivo, promedio de AUC $0{,}64610 \pm 0{,}11573$.

$$0{,}6RSA + 0{,}1C + 0{,}1pC + 0{,}2pMI \tag{24}$$

Notar que el término cMI no contribuye en el desempeño predictivo, por lo que su peso óptimo fue de 0.0 en el modelo. La diferencia en el desempeño obtenido utilizando con el puntaje de RSA

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

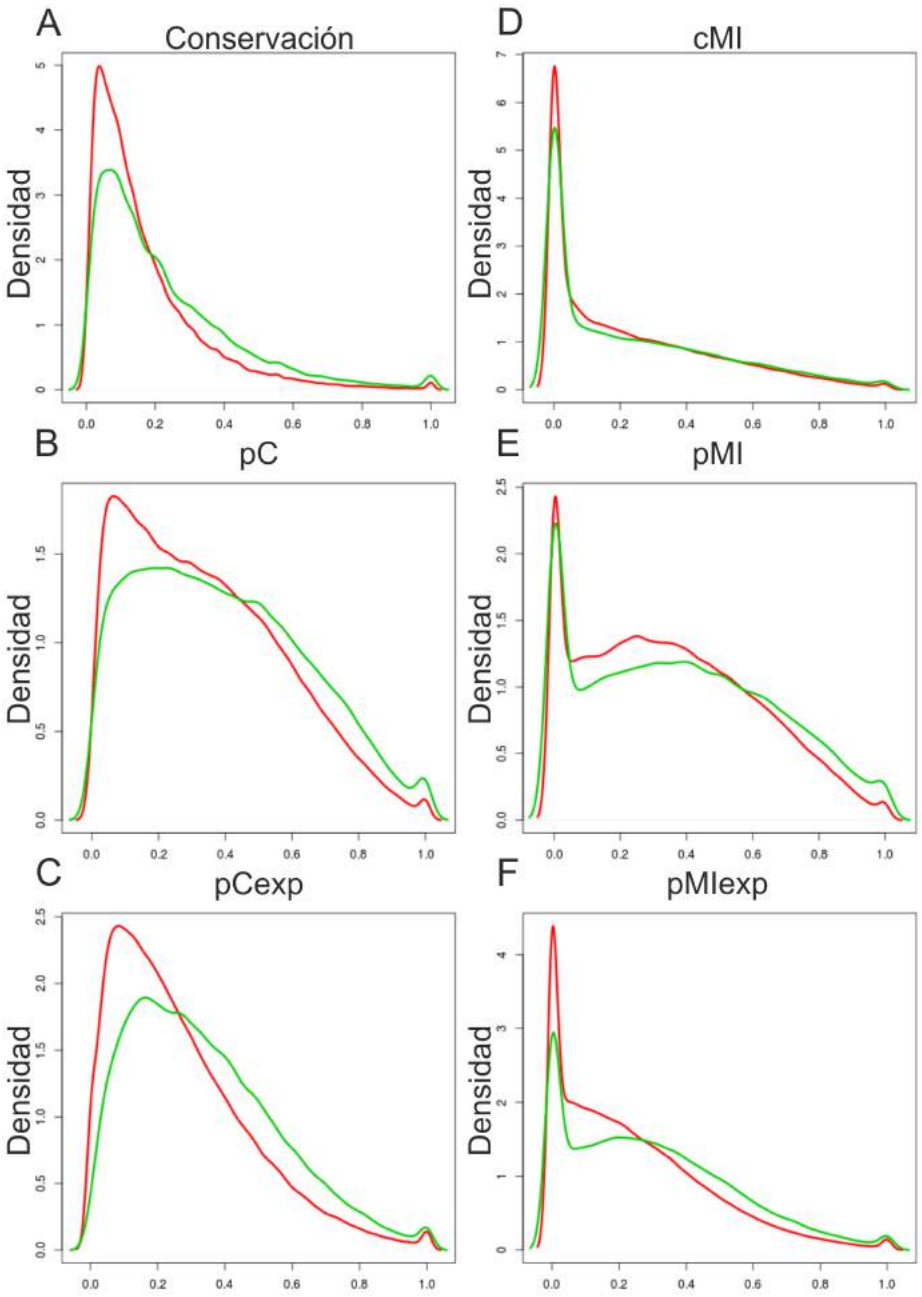


Figura 32: Plot de densidad
En cada panel la línea verde representa los residuos de interacción (positivos) y la roja el resto de los residuos de las unidades de interacción (negativos).

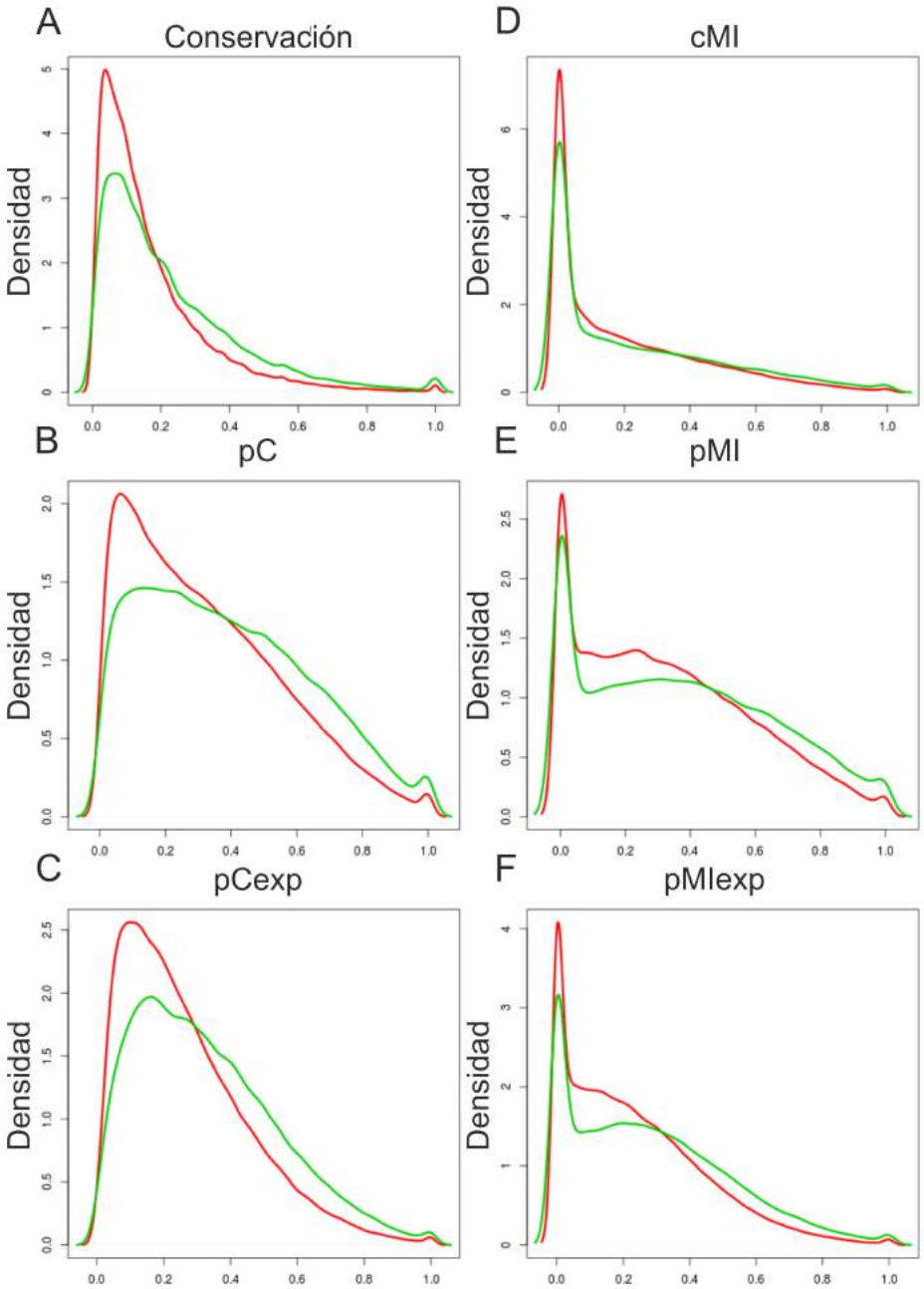


Figura 33: Plot de densidad para residuos expuestos
Se grafican los puntajes para residuos con un RSA>0.2. La línea verde representa los residuos de interacción (positivos) y la roja el resto de los residuos de superficie (RSA > 0.2) de las unidades de interacción (negativos).

Puntaje	promedio AUC
RSA	0.63408
C	0.53431
cMI	0.53171
pC	0.53195
pMI	0.53963

Tabla 8: Desempeño predictivo de los diferentes puntajes para la predicción de residuos interacción proteína-proteína.
Área relativa accesible al solvente (RSA), conservación (C), Información Mutua acumulada (cMI), conservación en la proximidad (pC) e información mutua en la proximidad (pMI). Promedio de AUC calculando sobre las 2471 cadenas de PDB analizadas.

y el modelo combinado es estadísticamente significativa (Wilcox test p-value=1.036e-06). Sin embargo, el desempeño es cercano a un predictor aleatorio, y el mayor peso relativo es para el puntaje de RSA.

Los valores de AUC para el mejor modelo encontrado varían entre 0.10526 y 1.00000, en la Figura 34 se muestra la distribución del desempeño predictivo para el mejor modelo combinado obtenido.

El desempeño predictivo para el subset de unidades de interacción que participan en la formación una sola interfaz es de 0.63648.

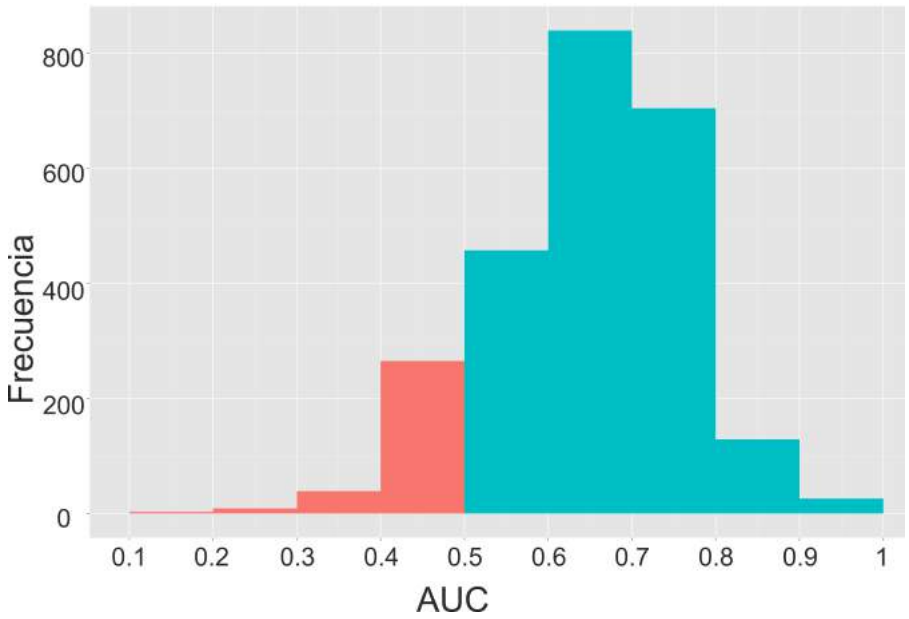


Figura 34: Histograma de AUCs para el mejor modelo combinado obtenido. Coloreado en rojo los valores menores a 0.5 y en verde los mayores.

Adicionalmente se analizó el subset de 995 MSAs que contienen más de 400 clusters de secuencias, necesarios para que el cálculo de MI sea confiable Buslje et al. [16], el promedio de AUC para el mejor modelo combinado es de 0.65214 y la distribución se muestra en le Figura 35.

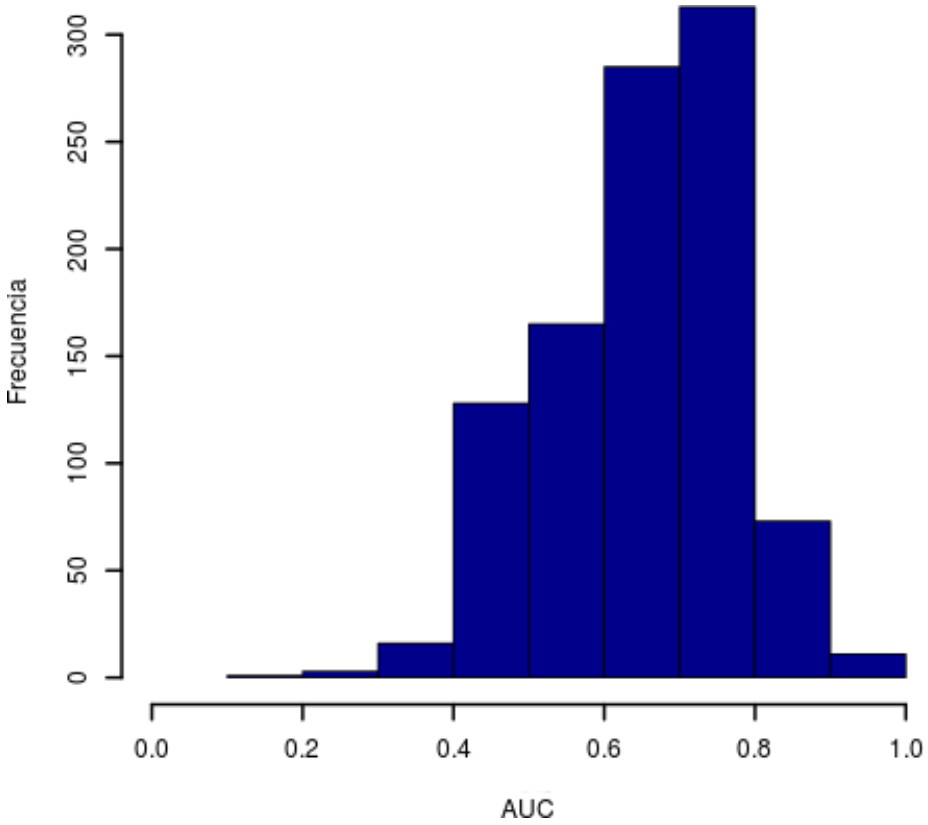


Figura 35: Histograma de AUC en el subset con más de 400 clusters de secuencias

Histograma generado a partir del mejor modelo combinado

5.5 DISCUSIÓN

Se estudiaron los residuos implicados en interacción proteína-proteína mediante diferentes señales evolutivas como son la conservación y la información mutua, se derivaron puntajes combinados con características estructurales de los residuos como el RSA y los puntajes por proximidad.

A partir de los resultados obtenidos concluimos que de haber una señal de coevolución intramolecular y conservación en la región de interacción, estas son débiles y no son suficientes para distinguir los residuos de interacción del resto de los residuos de la superficie.

Cabe destacar que existe una gran variedad de métodos predictivos de interacción proteína-proteína, pero sólo unos pocos dedicados a

predecir los residuo de interacción y que no incluyan información del complejo.

Posiblemente una de las mayores dificultades para predecir correctamente los residuos de interacción se debe a que i) la interacciones múltiples de una misma proteína pueden definir distintas interfaces y residuos de interacción, incluso interaccionando con el mismo compañero de interacción; ii) el cambio conformacional que sufren las proteínas al interaccionar no es considerado, ya que para ellos sería necesario contar con información estructural de cada unidad de interacción como monómero y formando el complejo. Los puntajes que consideran la estructura (puntajes de proximidad y RSA), son calculados a partir de la conformación que la unidad de interacción presente en el complejo. Estos puntajes pueden ser diferentes en la unidad de interacción libre, lo que sería un problema para el desarrollo de un método predictivo; iii) la naturaleza distinta de las interacciones (por ejemplo interacciones transitorias y permanentes) puede imponer diferentes presiones de selección haciendo que la detección de señales evolutivas en un set variado de complejos sea más dificultosa.

6 CONCLUSIONES

En el presente trabajo se ha explorado el uso de diferentes medidas que pueden extraerse a partir de un MSA, como ser la conservación, conservación dentro de grupos de secuencias, la MI y puntajes derivados de estos, para predecir sitios funcionalmente importantes. Para el caso particular de los residuos catalíticos hemos demostrado en un set comprensivo de datos, que poseen una red con alta información mutua en su proximidad espacial. Demostramos en términos cuantitativos directos, a través de la ganancia en el desempeño predictivo, la contribución de la señal de coevolución en la detección de residuos catalíticos.

Considerando que se ha reportado que en la proximidad espacial de los residuos catalíticos pueden encontrarse un tipo particular de residuos que determinan la especificidad de una enzima, hemos encontrado que los método predictivos de SDPs ensayados tienen una concordancia limitada en sus predicciones. Cabe mencionar que los método de SDPs utilizados no requieren información previa de los grupos de especificidad de la familia a analizar, siendo este un paso difícil y crítico para la predicción.

También encontramos que los residuos con elevado puntaje el puntaje cMI, no concuerdan con los predichos por los métodos de SDPs, conservación ni con el puntaje iv de Evolutionary Trace que ordena los residuos según su importancia funcional.

Además generando puntajes de proximidad con los métodos de predicción de SDPs utilizados, no se obtiene un método predictivo razonable para la identificación de residuos catalíticos. El mejor método predictivo encontrado surgió de la combinación de la señal de conservación con los puntajes de proximidad de rvET y cMI. Lo que confirma que ambos métodos son distintos en naturaleza y que aportan información complementaria al sistema predictivo.

Tomando como caso de estudio la familia ST3Gal, se realizó un análisis de conservación, predicción de SDPs y coevolución. Se encontró una red de coevolución, compuesta por residuos que pertenecen a los motivos funcionales importantes de la familia y de la superfamilia a la que pertenece. Sugiriendo que al menos parte de estos motivos funcionales evolucionan de manera concertada. Adicionalmente se predijeron SDPs, entre subfamilias y grupos de subfamilias de especificidad conocida. En este caso tampoco hubo concordancia entre las posiciones con alta MI y las predichas como SDPs. Al conocer los grupos de especificidad, la predicción de SDPs es sencilla, ya que el patrón de conservación entre grupos puede observarse a partir del

MSA. La predicción de estas posiciones ha sido de relevancia, porque pueden aportar información sobre el mecanismo de reacción de una enzima, el mecanismo evolutivo que lleva al surgimiento de una nueva especificidad y además permiten inferir funciones a secuencias aún no caracterizadas. El conjunto de resultados permitió comprender la evolución molecular de la familia ST3Gal y su diversidad funcional.

Por último, con respecto a la interfaz de interacción proteína-proteína se observó que la distribución de puntaje de conservación, MI y puntajes derivados de ellos, fue similar en los residuos de interacción que en el resto de la proteína. Indicando que de existir señal de conservación y coevolución en esta región, esta es muy débil y de difícil detección. Consideramos que los resultados obtenidos contribuyen a una mejor caracterización de la interface de interacción y pueden ser útiles para guiar el desarrollo de futuros predictores de interacción proteína-proteína.

7

FIGURAS Y TABLAS SUPLEMENTARIAS

7.1 MATERIAL SUPLEMENTARIO DEL CAPÍTULO 3

	SDPfox62	XDET50	ivET62	ivET100	cMI62	cMI100	rvET62	ivET100	cons
SDPfox62	1								
XDET50	0.34±0.2	1							
ivET62	0.10±0.13	0.22±0.13	1						
ivET100	0.07±0.14	0.08±0.13	0.21±0.21	1					
cMI62	0.16±0.26	0.26±0.21	0.17±0.18	0.04±0.16	1				
cMI100	0.16±0.27	0.28±0.21	0.18±0.19	0.05±0.18	0.76±0.15	1			
rvET62	0.23±0.16	0.38±0.15	0.41±0.24	0.14±0.26	0.36±0.23	0.38±0.21	1		
rvET100	0.26±0.17	0.38±0.15	0.41±0.22	0.21±0.29	0.36±0.22	0.41±0.20	0.93±0.10	1	
cons	0.05±0.12	0.31±0.16	0.44±0.17	0.17±0.22	0.16±0.22	0.16±0.19	0.74±0.15	0.73±0.15	1

Tabla 9: Correlación de Spearman entre los métodos y su desviación estándar.

El puntaje de correlación de Spearman varía entre 0 (sin correlación) y 1 (correlación perfecta).

7.2 MATERIAL SUPLEMENTARIO DEL CAPÍTULO 4

Figura 36: Árbol filogenético por máxima verosimilitud para 124 secuencias de ST3Gal

Se muestra el árbol con máxima verosimilitud, la longitud de las ramas es proporcional al número de sustituciones por sitio. El análisis comprende 124 secuencias y 228 posiciones, todas las posiciones con menos del 95 % de cobertura fueron descartadas. Los valores de bootstrap fueron calculados a partir de 500 replicaciones, se muestran los valores >80 % a la izquierda de cada punto de divergencia.

Figure 11 Petit *et al.* Evolutionary study of *st3gal* gene expression in Deuterostomes

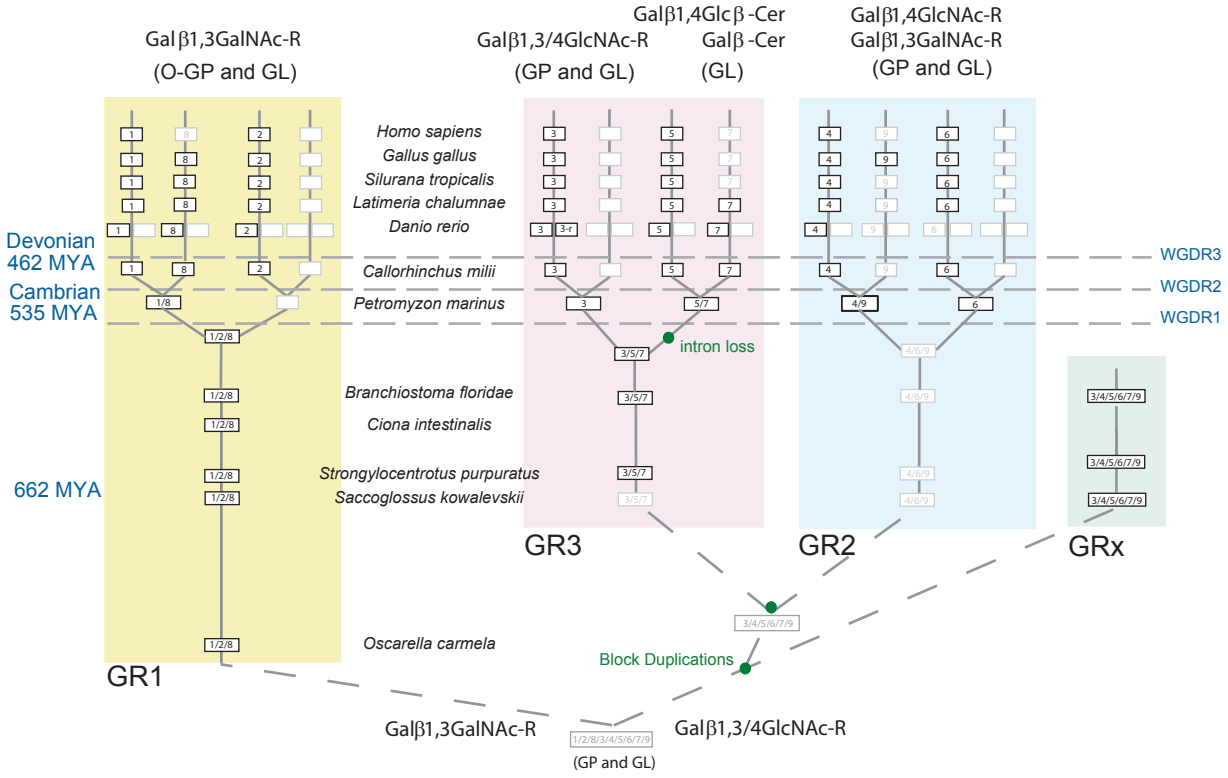


Figura 37: Diagrama esquemático de la evolución de ST3Gal en varias especies animales

Esta figura representa la evolución de los genes ST3Gal, representados por rectángulos negros cuando se ha detectado en el genoma y por rectángulos grises cuando se han perdido. La divergencia funcional entre cada grupo GR1, GR2, GR3 y GRx se adquirió muy pronto durante la evolución en deuterostomados y nuevas actividades enzimáticas propias de cada subfamilia se adquirieron después de eventos de duplicación del genoma completo, señalados como WGD (whole genome duplication). Se indican con un círculo verde los eventos de duplicación y pérdida de intrones. GP=Glicoproteína, GL=Glicolípido.

Table S1: Nomenclature, name and accession number in GenBank/EMBL/DDBJ of *st3gal* gene sequences identified and used in this study. The name of the organism and length of the predicted protein are indicated.

subfamily	**organism**	**Acc Nber GenBank**	**length**
ST3Gal8 (GR1)			
ST3Gal8	*Proteropterus aethiopicus*	FL668363	265
ST3Gal8	*Anolis carolinensis*	XM_003220475	326
ST3Gal8	*Chrysemys picta bellii*	HG531973	329
ST3Gal8	*Danio rerio*	AJ783740	341
ST3Gal8	*Latimeria chalumnae*	HG531974	332
ST3Gal8	*Melopsittacus undulatus*	XM_005147589	319
ST3Gal8	*Silurana tropicalis*	AJ585763	332
ST3Gal8A	*Xenopus laevis*	AJ585762	332
ST3Gal8B	*Xenopus laevis*	HG531975	332
ST3Gal8	*Gallus gallus*	XM_417321	313
ST3Gal8	*Python bivittatus*	HG531976	328
ST3Gal8	*Lepisosteus oculatus*	HG531977	340
ST3Gal2 (GR1)			
ST3Gal2	*Anolis carolinensis*	GADO01065579	350
ST3Gal2	*Bos taurus*	AJ748841	350
ST3Gal2	*Chrysemys picta bellii*	HG531978	347
ST3Gal2	*Danio rerio*	AJ783741	374
ST3Gal2	*Gallus gallus*	AJ585761	349
ST3Gal2	*Homo sapiens*	X96667	350
ST3Gal2	*Latimeria chalumnae*	HG531979	349
ST3Gal2	*Loxodonta africana*	XM_003417075	350
ST3Gal2	*Monodelphis domestica*	XM_001362558	350
ST3Gal2	*Ornithorhynchus anatinus*	XM_003431271	350
ST3Gal2	*Petromyzon marinus*	HG531980	>309
ST3Gal2	*Python bivittatus*	HG531981	347
ST3Gal2	*Silurana tropicalis*	XM_002931660	351
ST3Gal2	*Taeniopygia guttata*	XM_002188528	348
ST3Gal2	*Takifugu rubripes*	AJ626817	376
ST3Gal1 (GR1)			
ST3Gal1	*Anolis carolinensis*	XM_003219376	342
ST3Gal1	*Bos taurus*	NM_001097983	339
ST3Gal1	*Chrysemys picta bellii*	HG531982	360
ST3Gal1	*Gallus gallus*	X80503	342
ST3Gal1	*Homo sapiens*	L29555	340
ST3Gal1A	*Latimeria chalumnae*	HG531983	342
ST3Gal1B	*Latimeria chalumnae*	HG531984	342
ST3Gal1	*Lepisosteus oculatus*	HG531985	329
ST3Gal1	*Loxodonta africana*	XM_003408191	340
ST3Gal1	*Monodelphis domestica*	XM_001381537	342
ST3Gal1	*Oryzias latipes*	AJ871407	334
ST3Gal1	*Python bivittatus*	HG531986	>335
ST3Gal1	*Silurana tropicalis*	FN550106	334

ST3Gal1	*Taeniopygia guttata*	XM_002188126	341
ST3Gal1	*Takifugu rubripes*	AJ626816	333
ST3Gal1	*Tetraodon nigroviridis*	AJ744802	334
ST3Gal5 (GR3)			
ST3Gal5	*Anolis carolinensis*	XM_003226613	370
ST3Gal5	*Callorhinchus milii*	HG532005	>301
ST3Gal5	*Lepisosteus oculatus*	HG531987	>316
ST3Gal5	*Danio rerio*	AJ619960	364
ST3Gal5	*Gallus gallus*	AY515255	363
ST3Gal5	*Homo sapiens*	AB018356	362
ST3Gal5	*Monodelphis domestica*	XM_001379503	362
ST3Gal5	*Mus musculus*	Y15003	358
ST3Gal5	*Ornithorhynchus anatinus*	HG531988	384
ST3Gal5	*Petromyzon marinus*	HG531989	>331
ST3Gal5	*Silurana tropicalis*	FN550108	372
ST3Gal5	*Takifugu rubripes*	AJ865087	316
ST3Gal5	*Taeniopygia guttata*	XM_002186654	368
ST3Gal7 (GR3)			
ST3Gal7	*Danio rerio*	AJ783742	383
ST3Gal7	*Oryzias latipes*	AJ871411	386
ST3Gal7	*Takifugu rubripes*	AJ865347	386
ST3Gal7	*Lepisosteus oculatus*	HG531990	365
ST3Gal3 and ST3Gal3-r (GR3)			
ST3Gal3	*Anolis carolinensis*	XM_003220202	390
ST3Gal3	*Danio rerio*	AJ626821	355
ST3Gal3-r	*Danio rerio*	AJ626820	372
ST3Gal3	*Gallus gallus*	AJ865086	374
ST3Gal3	*Gasterosteus aculeatus*	HG531992	370
ST3Gal 3-r	*Gasterosteus aculeatus*	HG531991	356
ST3Gal3	*Lepisosteus oculatus*	HG531993	333
ST3Gal3	*Latimeria chalumnae*	HG531994	375
ST3Gal3	*Homo sapiens*	L23768	444
ST3Gal3	*Xenopus laevis*	BC169739	360
ST3Gal3	*Meleagris gallopavo*	XM_003208778	390
ST3Gal3	*Mus musculus*	X84234	374
ST3Gal3	*Python bivittatus*	HG531995	369
ST3Gal3	*Oryzias latipes*	HG531996	371
ST3Gal3	*Silurana tropicalis*	AJ626823	374
ST3Gal3	*Taeniopygia guttata*	XM_002192118	374
ST3Gal3	*Takifugu rubripes*	AJ626818	356
ST3Gal6 (GR2)			
ST3Gal6	*Homo sapiens*	AF119391	331
ST3Gal6	*Mus musculus*	AF119390	331
ST3Gal6	*Gallus gallus*	AJ585767	329
ST3Gal6	*Anolis carolinensis*	GAFZ01198187	327
ST3Gal6	*Latimeria chalumnae*	HG531997	329
ST3Gal6	*Silurana tropicalis*	AJ626744	331

ST3Gal6	*Taeniopygia guttata*	XP_002194132	356
ST3Gal6	*Meleagris gallopavo*	XP_003202812	329
ST3Gal4 (GR2)			
ST3Gal4	*Anolis carolinensis*	XP_003225763	329
ST3Gal4	*Callorhinchus milii*	HG531998	>243
ST3Gal4	*Danio rerio*	AJ744809	332
ST3Gal4	*Gallus gallus*	AJ866777	328
ST3Gal4	*Homo sapiens*	L23767	333
ST3Gal4	*Latimeria chalumnae*	HG531999	>181
ST3Gal4	*Meleagris gallopavo*	XM_003212645	328
ST3Gal4	*Mus musculus*	X95809	333
ST3Gal4	*Silurana tropicalis*	AJ622908	330
ST3Gal4	*Takifugu rubripes*	AJ865346	272
ST3Gal4	*Python bivittatus*	HG532000	342
ST3Gal4	*Lepisosteus oculatus*	HG532001	381
ST3Gal9 (GR2)			
ST3Gal9	*Gallus gallus*	XM_004945803	394
ST3Gal9	*Meleagris gallopavo*	HG532002	312
ST3Gal9	*Ornithorhynchus anatinus*	XM_001508863	322
ST3Gal9	*Pelodiscus sinensis*	HG532003	365
ST3Gal1/2/8 (GR1)			
	Ciona intestinalis	AJ703817	379
	Ciona savignyi	AJ626814	374
	Branchiostoma lanceolatum	HG532006	312
	Branchiostoma floridae	XM_002604587	377
	Saccoglossus kowalevskii	XM_002741450	350
	Strongylocentrotus purpuratus	AM420340	357
	Oscarella carmela	HG532004	374
ST3Gal3/3-r/5/7 (GR3)			
	Strongylocentrotus purpuratus A	XM_001184759	398
	Strongylocentrotus purpuratus B	XM_003730192	472
	Strongylocentrotus purpuratus C	XM_776047	335
	Branchiostoma floridae	XM_002601289	480
ST3Gal3/3-r/5/7/4/6/9 (GRx)			
	Saccoglossus kowalevskii	HG532008	393
	Branchiostoma lanceolatum	HG532007	365
	Branchiostoma floridae	XM_002606112	371

BIBLIOGRAFÍA

[1] Daniel Aguilar, Baldo Oliva, and Cristina Marino Buslje. Mapping the mutual information network of enzymatic families in the protein structure to unveil functional features. *PloS one*, 7 (7):e41430, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0041430. PMID: 22848494. (Cited on page 3.)

[2] Ron Alterovitz, Aaron Arvey, Sriram Sankararaman, Carolina Dallett, Yoav Freund, and Kimmen Sjölander. ResBoost: characterizing and predicting catalytic residues in enzymes. *BMC bioinformatics*, 10:197, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-197. PMID: 19558703. (Cited on page 16.)

[3] D Altschuh, A M Lesk, A C Bloomer, and A Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of molecular biology*, 193(4):693–707, February 1987. ISSN 0022-2836. PMID: 3612789. (Cited on page 3.)

[4] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, September 1997. ISSN 0305-1048. PMID: 9254694 PMCID: PMC146917. (Cited on page 50.)

[5] W R Atchley, K R Wollenberg, W M Fitch, W Terhalle, and A W Dress. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular biology and evolution*, 17(1):164–178, January 2000. ISSN 0737-4038. PMID: 10666716. (Cited on pages 3 y 5.)

[6] Holly J. Atkinson, John H. Morris, Thomas E. Ferrin, and Patricia C. Babbitt. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, 4(2):e4345, February 2009. doi: 10.1371/journal.pone.0004345. URL http://dx.doi.org/10.1371/journal.pone.0004345. (Cited on page 51.)

[7] Magali Audry, Charlotte Jeanneau, Anne Imberty, Anne Harduin-Lepers, Philippe Delannoy, and Christelle Breton. Current trends in the structure-activity relationships of sialyltransferases. *Glycobiology*, 21(6):716–726, June 2011. ISSN 1460-2423. doi: 10.1093/glycob/cwq189. PMID: 21098518. (Cited on pages 47 y 64.)

[8] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3):708–719, November 2003. ISSN 1097-0134. doi: 10.1002/prot.10461. PMID: 14579361. (Cited on pages 68 y 69.)

[9] Gail J Bartlett, Craig T Porter, Neera Borkakoti, and Janet M Thornton. Analysis of catalytic residues in enzyme active sites. *Journal of molecular biology*, 324(1):105–121, November 2002. ISSN 0022-2836. PMID: 12421562. (Cited on page 15.)

[10] Juliana S Bernardes, Jorge H Fernandez, and Ana Tereza R Vasconcelos. Structural descriptor database: a new tool for sequence-based functional site prediction. *BMC bioinformatics*, 9: 492, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-492. PMID: 19032768. (Cited on page 15.)

[11] George R Bickerton, Alicia P Higueruelo, and Tom L Blundell. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC bioinformatics*, 12:313, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-313. PMID: 21801404. (Cited on page 70.)

[12] Blahut, RE. *Principles and Practice of Information Theory*. Addison-Wesley, 1987. (Cited on page 6.)

[13] Luigi Boccuto, Kazuhiro Aoki, Heather Flanagan-Steet, Chin-Fu Chen, Xiang Fan, Frank Bartel, Marharyta Petukh, Ayla Pittman, Robert Saul, Alka Chaubey, Emil Alexov, Michael Tiemeyer, Richard Steet, and Charles E Schwartz. A mutation in a ganglioside biosynthetic enzyme, ST3GAL5, results in salt & pepper syndrome, a neurocutaneous disorder with altered glycolipid and glycoprotein glycosylation. *Human molecular genetics*, 23(2):418–433, January 2014. ISSN 1460-2083. doi: 10.1093/hmg/ddt434. PMID: 24026681 PMCID: PMC3869362. (Cited on page 50.)

[14] A A Bogan and K S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, July 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1998.1843. PMID: 9653027. (Cited on page 68.)

[15] D. P. Brown, N. Krishnamurthy, and K. Sjolander. Automated protein subfamily identification and classification. *PLoS Comput Biol*, 3:e160, 2007. (Cited on page 30.)

[16] Cristina Marino Buslje, Javier Santos, Jose Maria Delfino, and Morten Nielsen. Correction for phylogeny, small number of observations and data redundancy improves the identification of

coevolving amino acid pairs using mutual information. *Bioinformatics (Oxford, England)*, 25(9):1125–1131, May 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp135. PMID: 19276150. (Cited on pages 7, 10, 12, 17, 18, 22 y 81.)

[17] Daniel R Caffrey, Shyamal Somaroo, Jason D Hughes, Julian Mintseris, and Enoch S Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein science: a publication of the Protein Society*, 13(1):190–202, January 2004. ISSN 0961-8368. doi: 10.1110/ps.03323604. PMID: 14691234. (Cited on page 69.)

[18] Brandi L Cantarel, Pedro M Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. The carbohydrate-active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic acids research*, 37(Database issue):D233–238, January 2009. ISSN 1362-4962. doi: 10.1093/nar/gkn663. PMID: 18838391 PMCID: PMC2686590. (Cited on page 50.)

[19] J. A. Capra and M. Singh. Characterization and prediction of residues determining protein functional specificity. 24:1473 – 1480, 2008. (Cited on pages 27 y 30.)

[20] G. Casari, C. Sander, and A. Valencia. A method to predict functional residues in proteins. *Nat Struct Mol Biol*, 2(2):171 – 178, 1995. (Cited on page 30.)

[21] Pinak Chakrabarti and Joël Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–343, May 2002. ISSN 1097-0134. PMID: 11948787. (Cited on page 68.)

[22] S. Chakrabarti and A. Panchenko. Ensemble approach to predict specificity determinants: benchmarking and validation. 10 (1):207, 2009. (Cited on page 46.)

[23] S. Chakrabarti and A. R. Panchenko. Coevolution in defining the functional specificity. 75:231 – 240, 2009. (Cited on pages 16 y 30.)

[24] Saikat Chakrabarti and Anna R Panchenko. Structural and functional roles of coevolved sites in proteins. *PloS one*, 5(1): e8591, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0008591. PMID: 20066038 PMCID: PMC2797611. (Cited on page 16.)

[25] Ting-Ying Chien, Darby Tien-Hao Chang, Chien-Yu Chen, Yi-Zhong Weng, and Chen-Ming Hsu. E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. *Nucleic acids research*, 36(Web Server issue):W291–296, July 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn324. PMID: 18524800. (Cited on page 15.)

[26] Shin-ichi Chisada, Yukihiro Yoshimura, Keishi Sakaguchi, Satoshi Uemura, Shinji Go, Kazutaka Ikeda, Hiroyuki Uchima, Naoyuki Matsunaga, Kiyoshi Ogura, Tadashi Tai, Nozomu Okino, Ryo Taguchi, Jinichi Inokuchi, and Makoto Ito. Zebrafish and mouse alpha2,3-sialyltransferases responsible for synthesizing GM4 ganglioside. *The Journal of biological chemistry*, 284 (44):30534–30546, October 2009. ISSN 1083-351X. doi: 10.1074/jbc.M109.016188. PMID: 19542236 PMCID: PMC2781608. (Cited on page 65.)

[27] Elisa Cilia and Andrea Passerini. Automatic prediction of catalytic residues by modeling residue structural neighborhood. *BMC bioinformatics*, 11:115, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-115. PMID: 20199672. (Cited on page 15.)

[28] N D Clarke. Covariation of residues in the homeodomain sequence family. *Protein science: a publication of the Protein Society*, 4(11):2269–2278, November 1995. ISSN 0961-8368. doi: 10.1002/pro.5560041104. PMID: 8563623. (Cited on pages 3 y 69.)

[29] Francisco M. Codoner and Mario A. Fares. Why should we care about molecular coevolution? *Evol Bioinform Online*, 4:29–38, February 2008. ISSN 1176-9343. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2614197/. PMID: 19204805 PMCID: PMC2614197. (Cited on page 4.)

[30] Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley-Interscience, August 1991. ISBN 0471062596. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0471062596. (Cited on pages 6 y 17.)

[31] F Dall'Olio and M Chiricolo. Sialyltransferases in cancer. *Glycoconjugate journal*, 18(11-12):841–850, December 2001. ISSN 0282-0080. PMID: 12820717. (Cited on page 50.)

[32] A K Datta and J C Paulson. The sialyltransferase "sialylmotif"participates in binding the donor substrate CMP-NeuAc. *The Journal of biological chemistry*, 270(4):1497–1500, January 1995. ISSN 0021-9258. PMID: 7829476. (Cited on page 48.)

[33] A K Datta, A Sinha, and J C Paulson. Mutation of the sialyltransferase s-sialylmotif alters the kinetics of the donor and acceptor substrates. *The Journal of biological chemistry*, 273(16): 9608–9614, April 1998. ISSN 0021-9258. PMID: 9545292. (Cited on page 48.)

[34] Arun K Datta. Comparative sequence analysis in the sialyltransferase protein family: analysis of motifs. *Current drug targets*, 10

(6):483–498, June 2009. ISSN 1873-5592. PMID: 19519350. (Cited on page 48.)

[35] Antonio del Sol, Hirotomo Fujihashi, Dolors Amoros, and Ruth Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, 2:2006.0019, 2006. ISSN 1744-4292. doi: 10.1038/msb4100063. PMID: 16738564. (Cited on page 6.)

[36] Antonio Del Sol, Marcos J Araúzo-Bravo, Dolors Amoros, and Ruth Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome biology*, 8(5):R92, 2007. ISSN 1465-6914. doi: 10.1186/gb-2007-8-5-r92. PMID: 17531094. (Cited on page 7.)

[37] A. del Sol Mesa, F. Pazos, and A. Valencia. Automatic methods for predicting functionally important residues. 326(4):1289 – 1302, 2003. (Cited on pages 34 y 37.)

[38] S D Dunn, L M Wahl, and G B Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics (Oxford, England)*, 24(3):333–340, February 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm604. PMID: 18057019. (Cited on pages 6, 7 y 9.)

[39] Julien Y Dutheil. Detecting coevolving positions in a molecule: why and how to account for phylogeny. *Briefings in bioinformatics*, 13(2):228–243, March 2012. ISSN 1477-4054. doi: 10.1093/bib/bbr048. PMID: 21949241. (Cited on page 7.)

[40] Matt Eames and Tanja Kortemme. Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure (London, England: 1993)*, 15(11):1442–1451, November 2007. ISSN 0969-2126. doi: 10.1016/j.str.2007.09.010. PMID: 17997970. (Cited on page 69.)

[41] Simon Edvardson, Anna-Maria Baumann, Martina Mühlenhoff, Oliver Stephan, Andreas W Kuss, Avraham Shaag, Liqun He, Shamir Zenvirt, Raimo Tanzi, Rita Gerardy-Schahn, and Orly Elpeleg. West syndrome caused by ST3Gal-III deficiency. *Epilepsia*, 54(2):e24–27, February 2013. ISSN 1528-1167. doi: 10.1111/epi.12050. PMID: 23252400. (Cited on pages 50 y 65.)

[42] Lesley G Ellies, David Ditto, Gallia G Levy, Mark Wahrenbrock, David Ginsburg, Ajit Varki, Dzung T Le, and Jamey D Marth. Sialyltransferase ST3Gal-IV operates as a dominant modifier of hemostasis by concealing asialoglycoprotein receptor ligands.

Proceedings of the National Academy of Sciences of the United States of America, 99(15):10042–10047, July 2002. ISSN 0027-8424. doi: 10.1073/pnas.142005099. PMID: 12097641 PMCID: PMC126621. (Cited on page 49.)

[43] Serkan Erdin, R Matthew Ward, Eric Venner, and Olivier Lichtarge. Evolutionary trace annotation of protein function in the structural proteome. *Journal of molecular biology*, 396(5):1451–1473, March 2010. ISSN 1089-8638. doi: 10.1016/j.jmb.2009.12.037. PMID: 20036248. (Cited on page 15.)

[44] Mario A Fares and Simon A A Travers. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics*, 173(1):9–23, May 2006. ISSN 0016-6731. doi: 10.1534/genetics.105.053249. PMID: 16547113. (Cited on page 5.)

[45] WalterM. Fitch and Etan Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593, October 1970. ISSN 0006-2928. doi: 10.1007/BF00486096. URL http://dx.doi.org/10.1007/BF00486096. (Cited on page 4.)

[46] Anthony A Fodor and Richard W Aldrich. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins*, 56(2):211–221, August 2004. ISSN 1097-0134. doi: 10.1002/prot.20098. PMID: 15211506. (Cited on page 6.)

[47] B. G. Giraud, Alan Lapedes, and Lon Chang Liu. Analysis of correlations between sites in models of protein sequences. *Physical Review E*, 58(5):6312–6322, November 1998. doi: 10.1103/PhysRevE.58.6312. URL http://link.aps.org/doi/10.1103/PhysRevE.58.6312. (Cited on page 6.)

[48] Gregory B Gloor, Louise C Martin, Lindi M Wahl, and Stanley D Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, May 2005. ISSN 0006-2960. doi: 10.1021/bi050293e. PMID: 15882054. (Cited on pages 3 y 16.)

[49] Rodrigo Gouveia-Oliveira and Anders G Pedersen. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms for molecular biology: AMB*, 2:12, 2007. ISSN 1748-7188. doi: 10.1186/1748-7188-2-12. PMID: 17915013. (Cited on pages 7, 9 y 10.)

[50] Mainak Guharoy and Pinak Chakrabarti. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15447–15452, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0505425102. PMID: 16221766. (Cited on page 69.)

[51] Luke Hakes, Simon C Lovell, Stephen G Oliver, and David L Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (19):7999–8004, May 2007. ISSN 0027-8424. doi: 10.1073/pnas.0609962104. PMID: 17468399. (Cited on page 69.)

[52] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, August 2009. ISSN 1097-4172. doi: 10.1016/j.cell.2009.07.038. PMID: 19703402. (Cited on pages 3 y 6.)

[53] Barry G Hall. Building phylogenetic trees from molecular data with MEGA. *Molecular biology and evolution*, 30(5):1229–1235, May 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst012. PMID: 23486614. (Cited on page 52.)

[54] Harduin-Lepers. Comprehensive analysis of sialyltransferases in vertebrate genomes. *Glycobiology Insights*, page 29, February 2010. ISSN 1179-2515. doi: 10.4137/GBI.S3123. URL http://www.la-press.com/comprehensive-analysis-of-sialyltransferases-in-vertebrate-genomes-article-a1870. (Cited on pages 49 y 51.)

[55] Anne Harduin-Lepers, Marie-Ange Krzewinski-Recchi, Florent Colomb, Francois Foulquier, Sophie Groux-Degroote, and Philippe Delannoy. Sialyltransferases functions in cancers. *Frontiers in bioscience (Elite edition)*, 4:499–515, 2012. ISSN 1945-0508. PMID: 22201891. (Cited on page 50.)

[56] Kosuke Hashimoto, Susumu Goto, Shin Kawano, Kiyoko F Aoki-Kinoshita, Nobuhisa Ueda, Masami Hamajima, Toshisuke Kawasaki, and Minoru Kanehisa. KEGG as a glycome informatics resource. *Glycobiology*, 16(5):63R–70R, May 2006. ISSN 0959-6658. doi: 10.1093/glycob/cwj010. PMID: 16014746. (Cited on page 50.)

[57] Kosuke Hashimoto, Toshiaki Tokimatsu, Shin Kawano, Akiyasu C Yoshizawa, Shujiro Okuda, Susumu Goto, and Minoru Kanehisa. Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans. *Carbohydrate research*, 344(7):881–887, May

2009. ISSN 1873-426X. doi: 10.1016/j.carres.2009.03.001. PMID: 19327755. (Cited on page 50.)

[58] Thierry Hennet. Diseases of glycosylation beyond classical congenital disorders of glycosylation. *Biochimica et biophysica acta*, 1820(9):1306–1317, September 2012. ISSN 0006-3002. doi: 10.1016/j.bbagen.2012.02.001. PMID: 22343051. (Cited on page 50.)

[59] U Hobohm, M Scharf, R Schneider, and C Sander. Selection of representative protein data sets. *Protein science: a publication of the Protein Society*, 1(3):409–417, March 1992. ISSN 0961-8368. doi: 10.1002/pro.5560010313. PMID: 1304348 PMCID: PMC2142204. (Cited on page 10.)

[60] Hao Hu, Katinka Eggers, Wei Chen, Masoud Garshasbi, M Mahdi Motazacker, Klaus Wrogemann, Kimia Kahrizi, Andreas Tzschach, Masoumeh Hosseini, Ideh Bahman, Tim Hucho, Martina Mühlenhoff, Rita Gerardy-Schahn, Hossein Najmabadi, H Hilger Ropers, and Andreas W Kuss. ST3GAL3 mutations impair the development of higher cognitive functions. *American journal of human genetics*, 89(3):407–414, September 2011. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.08.008. PMID: 21907012 PMCID: PMC3169827. (Cited on pages 50 y 65.)

[61] C Axel Innis, A Prem Anand, and R Sowdhamini. Prediction of functional sites in proteins using conserved functional group analysis. *Journal of molecular biology*, 337(4):1053–1068, April 2004. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.01.053. PMID: 15033369. (Cited on page 15.)

[62] Robbie P Joosten, Tim A H te Beek, Elmar Krieger, Maarten L Hekkelman, Rob W W Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of PDB related databases for everyday needs. *Nucleic acids research*, 39(Database issue):D411–419, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1105. PMID: 21071423. (Cited on page 72.)

[63] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, December 1983. ISSN 0006-3525. doi: 10.1002/bip.360221211. PMID: 6667333. (Cited on page 72.)

[64] O. V. Kalinina. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. 13(2):443 – 456, 2004. (Cited on pages 27 y 46.)

[65] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000. ISSN 0305-1048. PMID: 10592173 PMCID: PMC102409. (Cited on page 50.)

[66] Maricel G Kann, Benjamin A Shoemaker, Anna R Panchenko, and Teresa M Przytycka. Correlated evolution of interacting proteins: looking behind the mirrortree. *Journal of molecular biology*, 385(1):91–98, January 2009. ISSN 1089-8638. doi: 10.1016/j.jmb.2008.09.078. PMID: 18930732. (Cited on page 69.)

[67] H Kitagawa and J C Paulson. Differential expression of five sialyltransferase genes in human tissues. *The Journal of biological chemistry*, 269(27):17872–17878, July 1994. ISSN 0021-9258. PMID: 8027041. (Cited on page 49.)

[68] N Kojima, Y C Lee, T Hamamoto, N Kurosawa, and S Tsuji. Kinetic properties and acceptor substrate preferences of two kinds of gal beta 1,3GalNAc alpha 2,3-sialyltransferase from mouse brain. *Biochemistry*, 33(19):5772–5776, May 1994. ISSN 0006-2960. PMID: 8180204. (Cited on page 49.)

[69] M Kono, Y Ohyama, Y C Lee, T Hamamoto, N Kojima, and S Tsuji. Mouse beta-galactoside alpha 2,3-sialyltransferases: comparison of in vitro substrate specificities and tissue specific expression. *Glycobiology*, 7(4):469–479, June 1997. ISSN 0959-6658. PMID: 9184827. (Cited on page 49.)

[70] B T Korber, R M Farber, D H Wolpert, and A S Lapedes. Covariation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 90(15):7176–7180, August 1993. ISSN 0027-8424. PMID: 8346232. (Cited on pages 6 y 7.)

[71] M Kotani, I Kawashima, H Ozawa, T Terashima, and T Tai. Differential distribution of major gangliosides in rat central nervous system detected by specific monoclonal antibodies. *Glycobiology*, 3(2):137–146, April 1993. ISSN 0959-6658. PMID: 8490240. (Cited on page 49.)

[72] David M Kristensen, R Matthew Ward, Andreas Martin Lisewski, Serkan Erdin, Brian Y Chen, Viacheslav Y Fofanov, Marek Kimmel, Lydia E Kavraki, and Olivier Lichtarge. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC bioinformatics*, 9:17, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-17. PMID: 18190718. (Cited on page 15.)

[73] Remko K P Kuipers, Henk-Jan Joosten, Eugene Verwiel, Sjoerd Paans, Jasper Akerboom, John van der Oost, Nicole G H Leferink, Willem J H van Berkel, Gert Vriend, and Peter J Schaap. Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins*, 76(3):608–616, August 2009. ISSN 1097-0134. doi: 10.1002/prot.22374. PMID: 19274741. (Cited on page 16.)

[74] Byung-Chul Lee, Keunwan Park, and Dongsup Kim. Analysis of the residue-residue coevolution network and the functionally important residues in proteins. *Proteins*, 72(3):863–872, August 2008. ISSN 1097-0134. doi: 10.1002/prot.21972. PMID: 18275083. (Cited on page 16.)

[75] Friederike Lehmann, Sørge Kelm, Frank Dietz, Mark von Itzstein, and Joe Tiralongo. The evolution of galactose alpha2,3-sialyltransferase: Ciona intestinalis ST3GAL I/II and takifugu rubripes ST3GAL II sialylate Galbeta1,3GalNAc structures on glycoproteins but not glycolipids. *Glycoconjugate journal*, 25(4):323–334, May 2008. ISSN 1573-4986. doi: 10.1007/s10719-007-9078-4. PMID: 17973185. (Cited on page 50.)

[76] D Leys, A S Tsapin, K H Nealson, T E Meyer, M A Cusanovich, and J J Van Beeumen. Structure and mechanism of the flavocytochrome c fumarate reductase of shewanella putrefaciens MR-1. *Nature structural biology*, 6(12):1113–1117, December 1999. ISSN 1072-8368. doi: 10.1038/70051. PMID: 10581551. (Cited on page 23.)

[77] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2):342 – 358, 1996. (Cited on page 30.)

[78] Daniel Y Little and Lu Chen. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PloS one*, 4(3):e4762, 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004762. PMID: 19274093 PMCID: PMC2651771. (Cited on page 16.)

[79] Y L Lo, T H Leoh, Y F Dan, L L Lim, A Seah, S Fook-Chong, and P Ratnagopal. Presynaptic neuromuscular transmission defect in the miller fisher syndrome. *Neurology*, 66(1):148–149, January 2006. ISSN 1526-632X. doi: 10.1212/01.wnl.0000191400.77080.21. PMID: 16401873. (Cited on page 49.)

[80] L Lo Conte, C Chothia, and J Janin. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285 (5):2177–2198, February 1999. ISSN 0022-2836. PMID: 9925793. (Cited on page 68.)

Repositorio Institucional Digital de Acceso Abierto, Universidad Nacional de Quilmes

[81] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (New York, N.Y.)*, 286(5438):295–299, October 1999. ISSN 0036-8075. PMID: 10514373. (Cited on page 16.)

[82] Jonathan R Manning, Emily R Jefferson, and Geoffrey J Barton. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC bioinformatics*, 9:51, 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-51. PMID: 18221517. (Cited on page 15.)

[83] L C Martin, G B Gloor, S D Dunn, and L M Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics (Oxford, England)*, 21(22):4116–4124, November 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti671. PMID: 16159918. (Cited on page 6.)

[84] P. Marttinen. Bayesian search of functionally divergent protein subgroups and their function specific residues. 22:2466 – 2474, 2006. (Cited on page 30.)

[85] P. Mazin. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms for Molecular Biology*, 5(1):29, 2010. (Cited on pages 30, 31 y 34.)

[86] I Mihalek, I Res, and O Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of molecular biology*, 336(5):1265–1282, March 2004. ISSN 0022-2836. doi: 10.1016/j.jmb.2003.12.078. PMID: 15037084. (Cited on pages 15, 30, 31, 34 y 35.)

[87] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10930–10935, August 2005. ISSN 0027-8424. doi: 10.1073/pnas.0502667102. PMID: 16043700. (Cited on page 69.)

[88] D. H. Morgan. ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, 22(16):2049 – 2050, 2006. (Cited on page 30.)

[89] A G Murzin, S E Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247 (4):536–540, April 1995. ISSN 0022-2836. doi: 10.1006/jmbi.1995.0159. PMID: 7723011. (Cited on page 17.)

[90] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: a novel method for fast and accurate multiple sequence alignment. 302 (1):205 – 217, 2000. (Cited on page 33.)

[91] W. Oliveira L, G. Vriend, and A. P. Ljzerman. Identification of class-determining residues in g protein-coupled receptors by sequence analysis. 5(3-4):159 – 174, 1997. (Cited on page 30.)

[92] Ronak Y Patel and Petety V Balaji. Identification of linkage-specific sequence motifs in sialyltransferases. *Glycobiology*, 16 (2):108–116, February 2006. ISSN 0959-6658. doi: 10.1093/glycob/cwj046. PMID: 16207893. (Cited on page 51.)

[93] F. Pazos, A. Rausell, and A. Valencia. Phylogeny-independent detection of functional residues. 22(12):1440 – 1448, 2006. (Cited on pages 31 y 34.)

[94] J. Pei. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics*, 22: 164 – 171, 2006. (Cited on page 30.)

[95] Daniel Petit, Roxana Elin Teppa, Jean-Michel Petit, and Anne Harduin-Lepers. A practical approach to reconstruct evolutionary history of animal sialyltransferases and gain insights into the sequence-function relationships of golgi-glycosyltransferases. *Methods in molecular biology (Clifton, N.J.)*, 1022:73–97, 2013. ISSN 1940-6029. doi: 10.1007/978-1-62703-465-4_7. PMID: 23765655. (Cited on page 50.)

[96] Natalia V Petrova and Cathy H Wu. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC bioinformatics*, 7:312, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-312. PMID: 16790052. (Cited on page 15.)

[97] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, October 2004. ISSN 0192-8651. doi: 10.1002/jcc.20084. PMID: 15264254. (Cited on pages 58, 59 y 64.)

[98] W. Pirovano, K. A. Feenstra, and J. Heringa. Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res*, 34(22):6540 – 6548, 2006. (Cited on page 30.)

[99] C. T. Porter, G. J. Bartlett, J. M. Thornton, and The Catalytic Site Atlas. volume 32. 2004. (Cited on pages 17 y 27.)

[100] J J Priatel, D Chui, N Hiraoka, C J Simmons, K B Richardson, D M Page, M Fukuda, N M Varki, and J D Marth. The ST3Gal-I sialyltransferase controls CD8+ t lymphocyte homeostasis by modulating o-glycan biosynthesis. *Immunity*, 12(3):273–283, March 2000. ISSN 1074-7613. PMID: 10755614. (Cited on page 49.)

[101] Bojana Rakic, Francesco V Rao, Karen Freimann, Warren Wakarchuk, Natalie C J Strynadka, and Stephen G Withers. Structure-based mutagenic analysis of mechanism and substrate specificity in mammalian glycosyltransferases: porcine ST3Gal-I. *Glycobiology*, 23(5):536–545, May 2013. ISSN 1460-2423. doi: 10.1093/glycob/cwt001. PMID: 23300007. (Cited on pages 59, 60 y 64.)

[102] Francesco V Rao, Jamie R Rich, Bojana Rakić, Sai Buddai, Marc F Schwartz, Karl Johnson, Caryn Bowe, Warren W Wakarchuk, Shawn Defrees, Stephen G Withers, and Natalie C J Strynadka. Structural insight into mammalian sialyltransferases. *Nature structural & molecular biology*, 16(11):1186–1188, November 2009. ISSN 1545-9985. doi: 10.1038/nsmb.1685. PMID: 19820709. (Cited on pages 52, 59, 60, 62, 64, 65 y 66.)

[103] Antonio Rausell, David Juan, Florencio Pazos, and Alfonso Valencia. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(5): 1995–2000, February 2010. ISSN 1091-6490. doi: 10.1073/pnas.0908044107. PMID: 20133844. (Cited on page 16.)

[104] C H Robert and J Janin. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *Journal of molecular biology*, 283(5):1037–1047, November 1998. ISSN 0022-2836. doi: 10.1006/jmbi.1998.2152. PMID: 9799642. (Cited on page 70.)

[105] G. J. Rodriguez. Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. 107(17):7787 – 7792, 2010. (Cited on pages 35 y 46.)

[106] Philippe F Rohfritsch, John A F Joosten, Marie-Ange Krzewinski-Recchi, Anne Harduin-Lepers, Benoit Laporte, Sylvie Juliant, Martine Cerutti, Philippe Delannoy, Johannes F G Vliegenthart, and Johannis P Kamerling. Probing the substrate specificity of four different sialyltransferases using synthetic beta-d-galp-(1–>4)-beta-d-GlcpNAc-(1–>2)-alpha-D-Manp-(1–>O) (CH(2))7CH3 analogues general activating effect of replacing n-acetylglucosamine by n-

propionylglucosamine. *Biochimica et biophysica acta*, 1760(4):685–692, April 2006. ISSN 0006-3002. doi: 10.1016/j.bbagen.2005.12.012. PMID: 16439063. (Cited on pages 49 y 65.)

[107] B Rost and C Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20(3):216–226, November 1994. ISSN 0887-3585. doi: 10.1002/prot.340200303. PMID: 7892171. (Cited on page 72.)

[108] A Sali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234 (3):779–815, December 1993. ISSN 0022-2836. doi: 10.1006/jmbi.1993.1626. PMID: 8254673. (Cited on page 52.)

[109] S. Sankararaman and K. Sjolander. INTREPID - INformation-theoretic TREe traversal for protein functional site IDentification. 24:2445 – 2452, 2008. (Cited on pages 15 y 30.)

[110] Sriram Sankararaman, Fei Sha, Jack F Kirsch, Michael I Jordan, and Kimmen Sjölander. Active site prediction using evolutionary and structural information. *Bioinformatics (Oxford, England)*, 26(5):617–624, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq008. PMID: 20080507. (Cited on page 15.)

[111] George Shackelford and Kevin Karplus. Contact prediction using mutual information and neural nets. *Proteins*, 69 Suppl 8: 159–164, 2007. ISSN 1097-0134. doi: 10.1002/prot.21791. PMID: 17932918. (Cited on page 10.)

[112] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1948.tb01338.x/abstract. (Cited on page 17.)

[113] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, November 2003. ISSN 1088-9051. doi: 10.1101/gr.1239303. PMID: 14597658 PMCID: PMC403769. (Cited on page 51.)

[114] Zhengshuang Shi, Katheryn A Resing, and Natalie G Ahn. Networks for the allosteric control of protein kinases. *Current opinion in structural biology*, 16(6):686–692, December 2006. ISSN 0959-440X. doi: 10.1016/j.sbi.2006.10.011. PMID: 17085044. (Cited on page 16.)

[115] F. L. Simonetti, E. Teppa, A. Chernomoretz, M. Nielsen, and C. Marino Buslje. MISTIC: mutual information server to infer coevolution. *Nucleic Acids Research*, 41(W1):W8–W14, May 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt427. URL http://nar.oxfordjournals.org/content/41/W1/W8.abstract?keytype=ref&ijkey=40HGK0KRqtu20UI. (Cited on pages 9, 52 y 63.)

[116] Michael A Simpson, Harold Cross, Christos Proukakis, David A Priestman, David C A Neville, Gabriele Reinkensmeier, Heng Wang, Max Wiznitzer, Kay Gurtz, Argyro Verganelaki, Anna Pryde, Michael A Patton, Raymond A Dwek, Terry D Butters, Frances M Platt, and Andrew H Crosby. Infantile-onset symptomatic epilepsy syndrome caused by a homozygous loss-of-function mutation of GM3 synthase. *Nature genetics*, 36(11):1225–1229, November 2004. ISSN 1061-4036. doi: 10.1038/ng1460. PMID: 15502825. (Cited on page 50.)

[117] Beckett Sterner, Rohit Singh, and Bonnie Berger. Predicting and annotating catalytic residues: an information theoretic approach. *Journal of computational biology: a journal of computational molecular cell biology*, 14(8):1058–1073, October 2007. ISSN 1066-5277. doi: 10.1089/cmb.2007.0042. PMID: 17887954. (Cited on page 15.)

[118] Elizabeth R Sturgill, Kazuhiro Aoki, Pablo H H Lopez, Daniel Colacurcio, Katarina Vajn, Ileana Lorenzini, Senka Majić, Won Ho Yang, Marija Heffer, Michael Tiemeyer, Jamey D Marth, and Ronald L Schnaar. Biosynthesis of the major brain gangliosides GD1a and GT1b. *Glycobiology*, 22(10):1289–1301, October 2012. ISSN 1460-2423. doi: 10.1093/glycob/cws103. PMID: 22735313 PMCID: PMC3425327. (Cited on pages 49 y 65.)

[119] Shou Takashima, Takumi Matsumoto, Masafumi Tsujimoto, and Shuichi Tsuji. Effects of amino acid substitutions in the sialylmotifs on molecular expression and enzymatic activities of α2,8-sialyltransferases ST8Sia-I and ST8Sia-VI. *Glycobiology*, 23(5):603–612, May 2013. ISSN 1460-2423. doi: 10.1093/glycob/cwt002. PMID: 23315426. (Cited on page 64.)

[120] Yu-Rong Tang, Zhi-Ya Sheng, Yong-Zi Chen, and Ziding Zhang. An improved prediction of catalytic residues in enzyme structures. *Protein engineering, design & selection: PEDS*, 21(5):295–302, May 2008. ISSN 1741-0126. doi: 10.1093/protein/gzn003. PMID: 18287176. (Cited on page 16.)

[121] John N. Thompson. *The Coevolutionary Process*. University of Chicago Press, November 1994. ISBN 9780226797595. (Cited on page 3.)

[122] Elisabeth R M Tillier and Thomas W H Lui. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics (Oxford, England)*, 19(6):750–755, April 2003. ISSN 1367-4803. PMID: 12691987. (Cited on page 7.)

[123] Simon A A Travers and Mario A Fares. Functional coevolutionary networks of the hsp70-hop-hsp90 system revealed through computational analyses. *Molecular biology and evolution*, 24(4): 1032–1044, April 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm022. PMID: 17267421. (Cited on page 69.)

[124] Jeffrey A. Ubersax and James E. Ferrell Jr. Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, 8(7):530–541, July 2007. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm2203. URL http://www.nature.com/scitable/content/mechanisms-of-specificity-in-protein-phosphorylation-14018840. (Cited on page 27.)

[125] UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic acids research*, 38(Database issue):D142–148, January 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp846. PMID: 19843607. (Cited on page 18.)

[126] V Vallejo-Ruiz, R Haque, A M Mir, T Schwientek, U Mandel, R Cacan, P Delannoy, and A Harduin-Lepers. Delineation of the minimal catalytic domain of human galbeta1-3GalNAc alpha2,3-sialyltransferase (hST3Gal i). *Biochimica et biophysica acta*, 1549(2):161–173, October 2001. ISSN 0006-3002. PMID: 11690653. (Cited on page 64.)

[127] Ajit Varki. Sialic acids in human health and disease. *Trends in Molecular Medicine*, 14(8):351–360, August 2008. ISSN 1471-4914. doi: 10.1016/j.molmed.2008.06.002. URL http://www.cell.com/trends/molecular-medicine/abstract/S1471-4914(08)00133-0. (Cited on page 47.)

[128] R Matthew Ward, Eric Venner, Bryce Daines, Stephen Murray, Serkan Erdin, David M Kristensen, and Olivier Lichtarge. Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. *Bioinformatics (Oxford, England)*, 25(11):1426–1427, June 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp160. PMID: 19307237. (Cited on page 15.)

[129] K R Wollenberg and W R Atchley. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of*

Sciences of the United States of America, 97(7):3288–3291, March 2000. ISSN 0027-8424. doi: 10.1073/pnas.070154797. PMID: 10725404. (Cited on page 7.)

[130] Won Ho Yang, Claudia Nussbaum, Prabhjit K Grewal, Jamey D Marth, and Markus Sperandio. Coordinated roles of ST3Gal-VI and ST3Gal-IV sialyltransferases in the synthesis of selectin ligands. *Blood*, 120(5):1015–1026, August 2012. ISSN 1528-0020. doi: 10.1182/blood-2012-04-424366. PMID: 22700726 PMCID: PMC3412327. (Cited on page 49.)

[131] K. Ye. Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a machine-learning approach for feature weighting. 24:18 – 25, 2008. (Cited on page 46.)

[132] Chen-Hsiang Yeang and David Haussler. Detecting coevolution in and among protein domains. *PLoS computational biology*, 3(11):e211, November 2007. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030211. PMID: 17983264. (Cited on pages 3 y 69.)

[133] Hermann Zellner, Martin Staudigel, Thomas Trenner, Meik Bittkowski, Vincent Wolowski, Christian Icking, and Rainer Merkl. PresCont: predicting protein-protein interfaces utilizing four residue properties. *Proteins*, 80(1):154–168, January 2012. ISSN 1097-0134. doi: 10.1002/prot.23172. PMID: 22038731. (Cited on page 72.)

[134] Tuo Zhang, Hua Zhang, Ke Chen, Shiyi Shen, Jishou Ruan, and Lukasz Kurgan. Accurate sequence-based prediction of catalytic residues. *Bioinformatics (Oxford, England)*, 24(20):2329–2338, October 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn433. PMID: 18710875. (Cited on page 15.)

ANEXO

A

ARTÍCULOS PUBLICADOS

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification

Cristina Marino Buslje[1☯]*, Elin Teppa[1☯], Tomas Di Doménico[2], José María Delfino[2], Morten Nielsen[3]

1 Fundación Instituto Leloir, Buenos Aires, Argentina, **2** Institute of Biochemistry and Biophysics (IQUIFIB), School of Pharmacy and Biochemistry, University of Buenos Aires, Buenos Aires, Argentina, **3** Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Lyngby, Denmark

Abstract

Identification of catalytic residues (CR) is essential for the characterization of enzyme function. CR are, in general, conserved and located in the functional site of a protein in order to attain their function. However, many non-catalytic residues are highly conserved and not all CR are conserved throughout a given protein family making identification of CR a challenging task. Here, we put forward the hypothesis that CR carry a particular signature defined by networks of close proximity residues with high mutual information (MI), and that this signature can be applied to distinguish functional from other non-functional conserved residues. Using a data set of 434 Pfam families included in the catalytic site atlas (CSA) database, we tested this hypothesis and demonstrated that MI can complement amino acid conservation scores to detect CR. The Kullback-Leibler (KL) conservation measurement was shown to significantly outperform both the Shannon entropy and maximal frequency measurements. Residues in the proximity of catalytic sites were shown to be rich in shared MI. A structural proximity MI average score (termed pMI) was demonstrated to be a strong predictor for CR, thus confirming the proposed hypothesis. A structural proximity conservation average score (termed pC) was also calculated and demonstrated to carry distinct information from pMI. A catalytic likeliness score (Cls), combining the KL, pC and pMI measures, was shown to lead to significantly improved prediction accuracy. At a specificity of 0.90, the Cls method was found to have a sensitivity of 0.816. In summary, we demonstrate that networks of residues with high MI provide a distinct signature on CR and propose that such a signature should be present in other classes of functional residues where the requirement to maintain a particular function places limitations on the diversification of the structural environment along the course of evolution.

Citation: Marino Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M (2010) Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification. PLoS Comput Biol 6(11): e1000978. doi:10.1371/journal.pcbi.1000978

Editor: Burkhard Rost, Columbia University, United States of America

Received March 19, 2010; **Accepted** September 27, 2010; **Published** November 4, 2010

Copyright: © 2010 Marino Buslje et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was partially funded by the Argentinean National Research Council-CONICET grant number PIP 112200801_01936 (http://www.conicet.gov.ar/); CMB, MN and JMD are researchers of CONICET. ET is funded by the Leloir Institute Foundation (http://www.leloir.org.ar/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cmb@leloir.org.ar

☯ These authors contributed equally to the work.

Introduction

Catalytic residues play a fundamental role in enzymes and are generally expected to be conserved and located in the functional site of proteins. Even though characterization of catalytic residues (CR) is critical for the understanding of enzyme function, their identification remains a daunting task. To guide the identification of CR, several computational approaches have been developed based on different principles. To cite some examples: catalytic site features, amino acid physicochemical character [1], conserved functional groups density [2], sequence analysis (conservation, patterns, conserved blocks along the sequence, evolution, entropy, among others) [3,4,5,6,7,8], sequence and structure properties [9,10,11], evolution and 3D structure information [12,13,14,15], neural networks [16], 3D structure combined with ionization properties of a residue and its vicinity in the structure [17] and combinations of several of the above mentioned [18]. Conservation is the natural and intuitive way to predict functional residues in proteins. However, many non-catalytic residues are highly conserved and conversely, not all CR are fully conserved throughout a given protein family. On the other hand, residues involved in coevolving networks have been postulated to be functionally important [19,20,21] and several studies have provided evidence that they are important for specificity or allosteric regulation [22,23,24].

The structural environment of an active site must be highly conserved in order for the protein to maintain its function during the course of evolution. This places strict limitations on the amino acid diversity in the proximity of an active site, and it therefore seems plausible to hypothesise that catalytic residues would carry a particular signature defined by a network of close proximity of residues with high mutual information.

Although earlier published methods have suggested a linkage between functionally important sites and neighbouring coevolving residue [21,25,26] at present, to the best of our knowledge, no method explicitly show how the presence of such coevolving residues can provide quantitative information useful for catalytic sites identification beyond what is captured by conservation. Several methods have been proposed for identifying specificity defining positions (SDP) aiming at locating positions that are

Author Summary

Enzymes are responsible for several critical cellular functions. The so-called catalytic residues are fundamental to attain the enzyme function. Those residues are often highly conserved within protein families sharing similar structure and function. Characterization of catalytic residues is essential for the understanding of enzyme function. However, this is a difficult task because conservation is a poor discriminator of catalytic residues due to the fact that many non-catalytic residues are highly conserved in a given protein family. We anticipate that variations in the structural environment of a catalytic site should be highly restrained in order for the protein to maintain its function along the course of evolution, and hypothesise that catalytic residues, due to these restrains, must carry a particular signature defined by networks of proximity sharing high mutual information (MI). We validated this hypothesis on a large data set of protein sequences with known catalytic residues, and demonstrated that catalytic sites are indeed surrounded by networks of coevolved residues. Such networks should also be present in other classes of proteins and we suggest that MI networks could be a novel feature of general importance beneficial for the prediction of functional residues.

specific for a given subfamily and hence potentially could define its specificity [25,26]. These residues are suggested to be located in the proximity of the active residues in order to carry out their role of defining the substrate specificity. The signal from such evolutional signatures could at first resemble co-evolution, and the overlap between the methods predicting SDPs and the method proposed here could seem substantial. However, the subfamily specific positions may not be coevolving, in fact they might be fully conserved within each subfamily, and Gouveia-Oliveira and Pedersen have described in details that such subfamily defining residues do not carry signatures of co-evolution but rather a phylogenetic signal that mimics coevolution [27]. The methods put forward by Gouveia-Oliveira et al, [27], Dunn et al. [28], and Buslje et al. [29] all attempt to reduce this phylogenetic bias in the signal for MI calculation aiming at identifying truly coevolving residue-pairs. Moreover, the method proposed here is hypothesis-free, and can be applied without any prior functional cluster classification of the input multiple alignment.

Here, we perform a large-scale benchmark analysis aiming at testing the hypothesis that catalytic residues carry a signature defined by networks of close proximity of residues with high mutual information. An investigation on the relationship between conservation, coevolution networks and catalytic residues is carried out on a dataset of 434 families of enzymes. We introduce a new concept, Mutual Information Proximity (pMI) that characterizes the mutual information network in the proximity of a given residue and analyse whether this measurement can complement the conventional conservation score for the detection of catalytic residues. The goal of this work is two-fold. First, we aim to validate the hypothesis stated above and demonstrate that proximity residue networks of high mutual information characterize functional residues. In doing this, we also aim at addressing the issue on the correlation between residues defined as SDP and residues carrying high signals of being part of the mutual information network. Secondly, we seek to integrate this mutual information signature to create a method able to identify catalytic residues useful for guiding the identification of functional sites in proteins.

Note, that in this work, we do not suggest that the proposed method should be more accurate than the other methods developed earlier for prediction of functional residues. We merely seek to demonstrate the existence of a mutual information network signature in the proximity of functional residues, and show that this signature is complementary to the conventional sequence conservation measurement, hence most likely would benefit any functional residue prediction method.

Results

The main focus of this work was to investigate if mutual information could contribute beyond sequence conservation to the identification of catalytic residues. The result section naturally falls in three parts. First, we investigated how different measurements of sequence conservation could be used for the identification of catalytic residues. Next, a similar analysis was performed using different measurements of mutual information, and finally the analysis was carried out using a combined measurement of conservation and mutual information. Performance details of all methods included in the analysis are shown in supplementary table S2.

Sequence conservation

As catalytic residues are highly conserved, a natural measure used to detect them is the conservation score in a MSA. Here, we investigated three conservation measurements in four different conditions leading to twelve different conservation scores (for details see material and methods). The conservation measurements are all per-residue measurements, and their predictive performance for a given protein sequence is readily measured in terms of the AUC value. The results of this analysis on the 434 CSA Pfam families are shown in table 1.

The conservation measurement with the highest predictive performance in terms of AUC was the raw KL score with an average AUC value of 0.892 and an AUC01 value of 0.485. In terms of AUC, the raw calculation excluding both sequence weighting and pseudo count correction did perform best for all three conservation measurements. In terms of AUC01, the inclusion of sequence weighting in all cases did improve the predictive performance. The Max-Freq measurement performed significantly worse than both information-based measurements ($p<0.0001$, binomial test excluding ties). Although the performance is very similar between the raw Shannon and raw KL scores, the difference is highly significant ($p<0.005$, binomial test excluding ties). The difference between the raw and sequence weighted (c) KL score is borderline significant with a p-value of 0.05 in favour of the raw KL score for AUC and in favour of KL including sequence weighting when using AUC01. In order to make the subsequent analyses as simple as possible, for the remaining part of the work we used the raw KL score as a conservation measurement.

We analysed to what degree the predictive performance of the raw KL measurement depended on the number of sequences in the multiple sequence alignment (MSA) used as the source to estimate the conservation score (see figure 1). This figure clearly demonstrates that at least 10 sequences are required in order to make any meaningful predictions using the KL conservation measurement (similar results were observed for the other two conservation measurements). Note, that the variation in performance for each bar in the histogram is large and error-bars are not included (the raw data included in the figure are available in Supplementary table S2). The difference in predictive performance between the families with less than or more than 10

Table 1. Average performance in terms of the AUC and AUC01 values of the three methods: Max-Freq, Shannon, and Kullback-Leibler described to measure conservation.

Conservation measure	Max-Freq		Shannon		Kullback-Leibler	
	AUC	AUC01	AUC	AUC01	AUC	AUC01
Raw	0.874	0.458	0.880	0.464	**0.892**	0.485
C	0.870	0.461	0.876	0.465	0.890	**0.502**
L	0.857	0.380	0.852	0.371	0.877	0.437
CI	0.847	0.353	0.837	0.335	0.868	0.411

Each measurement is applied under four conditions defined by sequence weighting using clustering (c); pseudo count correction using low counts (l), the combination of the two (cl), and no correction (raw). In bold is highlighted the method with the highest performance for each performance measure.
doi:10.1371/journal.pcbi.1000978.t001

sequence members is however statistically highly significant ($p<0.001$, t-test).

Mutual information

We next turned to mutual information and analysed the environment of a catalytic residue by means of the mutual information carried by the surrounding residues. We introduced a cumulative Mutual Information concept (cMI) that measures the degree of shared mutual information of a given residue (above a certain significance threshold as measured in terms of the MI Z-score, see material and methods). We noticed that residues in close proximity with CR tend to have high cMI scores (see figure 2b). Furthermore, when measuring the proximity Mutual Information (pMI), which tells about the networks of mutual information in the proximity of a residue (within a certain distance threshold), the catalytic residues were observed to have higher pMI than other conserved residues (see figure 2c for an example).

We exploited this observation on the complete Pfam benchmark dataset, and calculated the performance of the pMI measurement as a predictor of catalytic residues. Using a distance cut-off of 7.5 Å to define the structural proximity, and a Z-score threshold of 6.0 to define reliable mutual information interactions (see [29]), the average predictive performance of the pMI measurement in terms of the average AUC and AUC01 values on the 434 Pfam entries was 0.843 and 0.342, respectively which in both cases is significantly different from random ($p<0.0001$, binomial test excluding ties). As the number of proximity interactions is used to normalize the pMI measurement, this predictive performance does not stem from any implicit bias in the data imposed by catalytic residues being in a particular state of solvent exposure.

Comparison between SDPs and cMI

To investigate how the mutual information measure (cMI) proposed in this work correlates to earlier proposed measures for SDP, we compared in terms of the Spearmans rank correlation the SDR Z-score values given in the SDR database (http://paradox.harvard.edu/sdr/) [30] to the cMI values. In doing this, we obtained a mean correlation value over the 158 Pfam families covered by both methods of $0.29+/-0.20$ (for details see materials and methods). Even though this correlation is significantly different from random ($p<0.01$, binomial test excluding ties), it is far from perfect. This highly suggests that the cMI and SDR measures carry

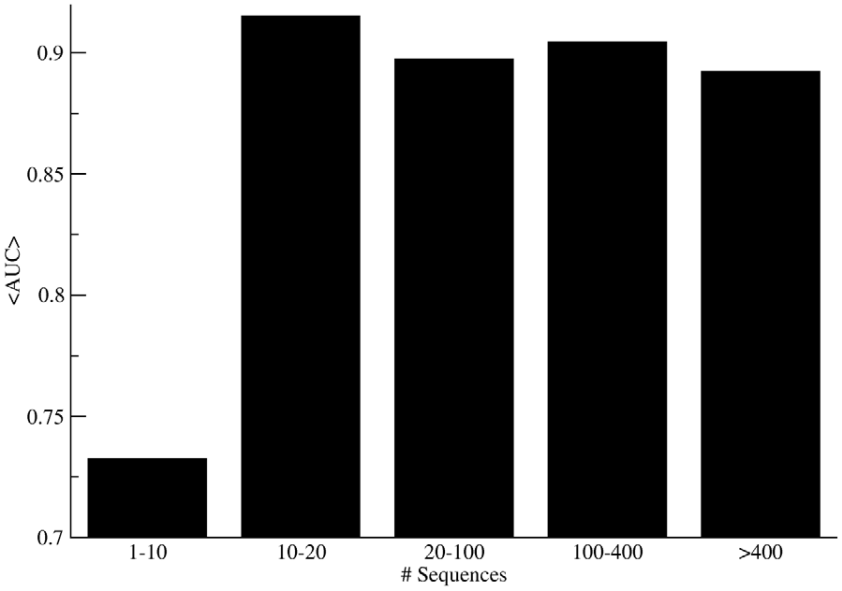


Figure 1. Histogram over predictive performance of the raw KL scores as a function of the number of sequences in the MSA. The number of Pfam entries in each sequence bin is 9, 9, 36, 66, and 314, respectively.
doi:10.1371/journal.pcbi.1000978.g001

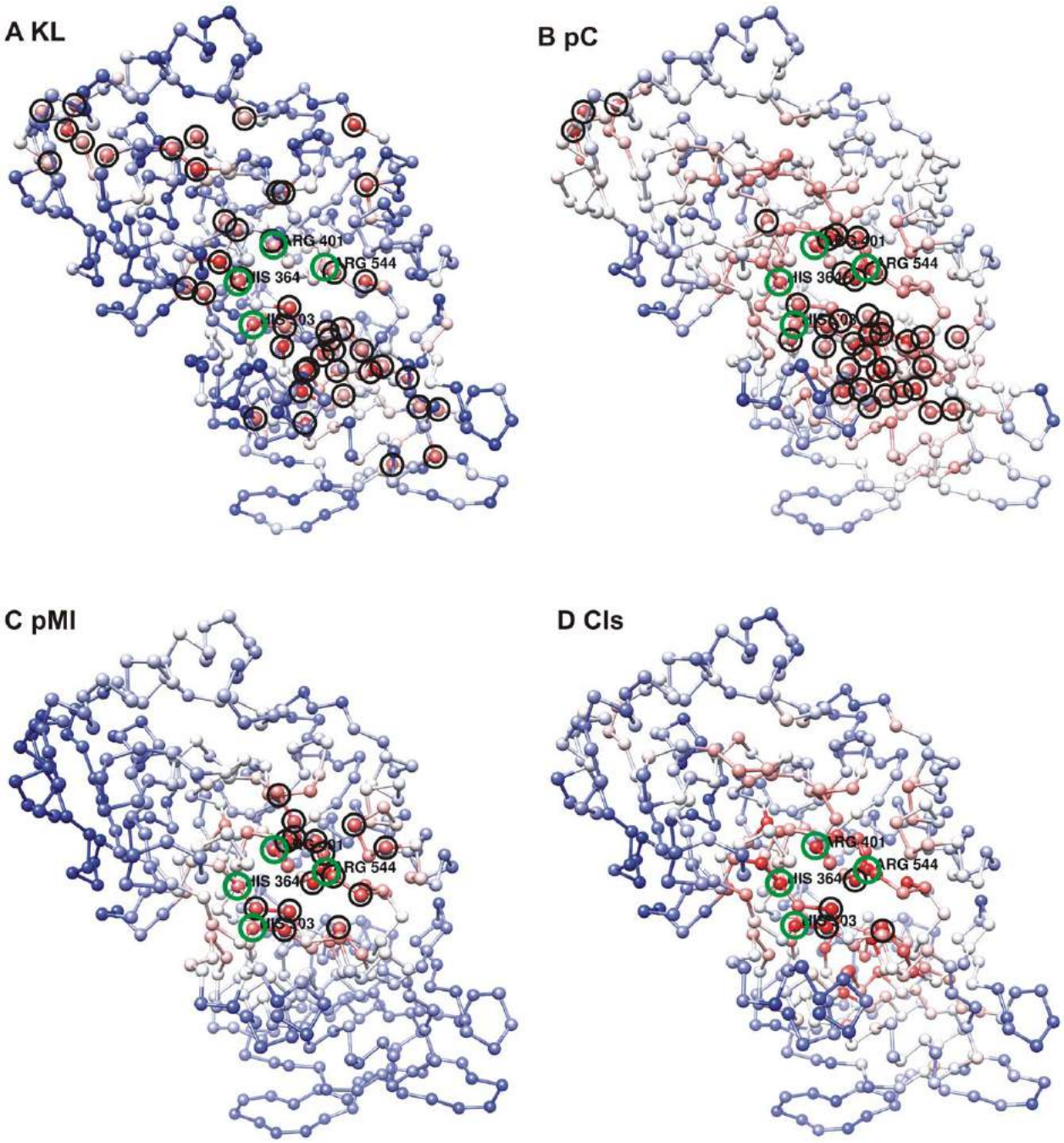


Figure 2. Identification of catalytic residues using four different prediction scores. Plotted is the Cα representation of the PDB entry 1D4C representing the Pfam PF00890 entry. Catalytic residues are encircled in green. The four different prediction scores are shown A) KL Conservation, B) Proximity conservation (pC), C) proximity MI (pMI) and D) Catalytic likeliness score (CIs). Highlighted with black circles are the predicted false positive residues: 47, 39, 15 and 4 respectively. The prediction scores are represented in blue to red scale (blue: lowest; red: highest). Molecular graphics image was produce with UCSF Chimera package. (University of California, San Francisco).
doi:10.1371/journal.pcbi.1000978.g002

distinct information. We next calculated the correlation between the two measures and the KL (Kullback-Leibler) conservation score. Here, we obtained an average Spearmans rank correlation values of 0.64 ± 0.21, and -0.04 ± 0.17 for the SDR Z-score and cMI measure respectively. These results further demonstrate that the SDP and cMI measures are different in nature, and that SDR Z-score is highly related to sequence conservation whereas the cMI score is independent of the latter. This strongly suggests that the cMI measure is more information rich compared to SDP when combined with sequence conservation.

Conservation of the residue proximity

As the active site in most cases is defined in terms of multiple catalytic residues in close proximity, it is natural to suggest that a

proximity score based on sequence conservation would be a strong catalytic residue predictor. Using the same distance cut-off as for the mutual information proximity score, we find that the proximity conservation score, pC, achieves an average predictive performance of 0.854 and 0.379 in terms of AUC and AUC01, respectively. These values are greater than what was obtained using the pMI score, but for both AUC and AUC01, the difference between the two methods is not statistically significant ($p<0.05$, binomial test excluding ties).

Combined catalytic likeliness score

We finally applied the combined catalytic likeliness score (Cls) to identify catalytic residues. The Cls is calculated as a weighted sum of the KL conservation the pMI mutual information and the pC scores. The optimal parameters defining the score were identified using 5-fold cross validation as described in Materials and Methods. The parameters Z_{thr}, D_{MI}, D_C, W_{MI} and W_C were found to have the following optimal values $Z_{thr} = 5.5 \pm 0.2$, $D_{MI} = 8.0 \pm 0.1$, $D_C = 5.6 \pm 0.5$, $w_{MI} = 0.6 \pm 0.0$, and $w_C = 0.2 \pm 0.0$. The low standard deviation value on each parameter-estimate indicates that the parameter optimization is robust across the different cross-validation data sets. The average performance in terms of the AUC and AUC01 of the Cls score to detect catalytic residues was 0.927, and 0.594, respectively. This performance is significantly higher than the KL conservation, the pMI and the pC individual scoring functions ($p<0.001$ in all cases using binomial test excluding ties).

To investigate the individual contribution to the performance of the Cls score of the pMI and pC measures, we next searched for optimal parameters for a combined score including only one of the two proximity measures in combination with the KL conservation score. Estimating the optimal parameters using 5 fold cross-validation as described above, we find the following results (see table 2).

The AUC values for both of these methods are significantly lower that what was obtained using the Cls score combining the conservation score with both proximity measures ($p<0.01$ in both cases, binomial test excluding ties) demonstrating that the two proximity measures contribute distinct information to the combined Cls score. The difference between the two scores including only one proximity measure is not statistically significant when looking at the complete data set of 434 PF families.

Table 2. Optimal parameters and average predictive performance in terms of AUC and AUC01 for the two combined prediction methods including only one proximity measure.

Method	KL+pMI	KL+pC
Parameters	$w_{MI} = 0.8 \pm 0.0$	$w_C = 0.6 \pm 0.0$
	$D_{MI} = 7.9 \pm 0.2$	$D_C = 8.0 \pm 0.0$
	$Z_{thr} = 5.5 \pm 0.32$	
AUC	0.922	0.910
AUC01	0.574	0.562

KL+pMI is the method combining KL conservation with the pMI mutual information measure. KL+pC is the method combining KL conservation with the pC conservation measure. w_{MI} is the relative weight on pMI, D_{MI} is the proximity distance threshold for the pMI measure, Z_{thr} is the MI Z-score threshold, w_C is the relative weight on pC, and D_C is the proximity distance threshold for the pC measure. Parameters and standard deviations were identified using five-fold cross validation as described in Materials and Methods.
doi:10.1371/journal.pcbi.1000978.t002

However, when looking at the subset of 172 PFam families that are covered by more than 400 unique sequences/clusters (corresponding to the number of clusters needed to provide reliable estimates of MI as shown by Buslje et al. [29]), the combined method including proximity mutual information, pMI, achieves a performance of AUC = 0.920, and AUC01 = 0.597. These values significantly outperform the performance values AUC = 0.889 and AUC01 = 0.559 of the combined method including proximity conservation, pC ($p<0.05$, binomial test excluding ties). This further underlines the observation that the pMI measure contributes information not included in the conservation scores.

To further illustrate that the two proximity measures contribute different information to the combined Cls-score, we in figure 2 display the role of the four prediction measurements, KL, pMI, pC and Cls for the identification of the catalytic residues in the Pfam entry PF00890 represented by fumarate reductase of Shewanella putrefaciens MR-1 (PDB entry 1D4C). This family was chosen from the subset of 172 Pfams entries mentioned above covered by more than 400 unique sequences/clusters (similar results are obtained for most other families in this set). The function of fumarate reductase is carried out by the active cite residues His364, Arg401, His503 and Arg544 [31]. It can be seen that the KL conservation score of the catalytic residues is relatively low (figure 2a) while both the pC, and pMI scores are high in the catalytic residue proximity (figure 2b, and 2c). Comparing the figures 2b and 2c, it is evident that the two proximity measures contribute different information to the combined, Cls, prediction score. Finally, the combined catalytic likeliness score (Cls) is depicted in figure 2d. The AUC values for the four prediction measurements shown in figure 2 are 0.92, 0.94, 0.98 and 0.99 (KL, pC, pMI and Cls respectively). These values translate into a number of false positive predictions at 100% sensitivity (corresponding to the number of non-catalytic residues with a prediction score higher than the lowest score obtained by a CR) of 47 (figure 2a), 39 (figure 2b), 15 (figure 2c), and 4 (figure 2d), again underlining the strong predictive power of the Cls measurements in identifying catalytic residues and eliminating false positive predictions.

The gain in predictive performance for detecting catalytic residues is consistent for families independently on the level of conservation of the catalytic residue, however the most dramatic gain in performance when including pMI is observed for families where the conservation of the catalytic residues is poor. If we for instance take the 217 Pfam families with the lowest predictive performance when using the KL conservation score and ask how many of these families gain in performance when including the pMI score, we find that this number is significantly higher compared to the corresponding number of families in the group of 217 Pfam families with the highest predictive performance using the KL conservation score ($p<0.001$, binomial test excluding ties). This difference in performance gain between the two subsets of Pfam families is not imposed by a difference in data size between the two sets as the average family size in the two set is comparable ($p>0.1$, t-test). The catalytic environment of an active site needs to be conserved in order for a protein family to maintain its function, and one might speculate that when the conservation of a catalytic residue is weak, the catalytic environment is maintained in great measure by coevolution.

We next determined the sensitivities of the different methods at different specificity thresholds. This analysis is summarized in table 3. The analysis clearly confirms the strong improvement across the entire benchmark data set of the predictability of catalytic residues imposed by the inclusion of the pMI score in the combined catalytic likeliness score. At all specificity thresholds, the

Table 3. Sensitivity of the catalytic residue identification methods at different specificity thresholds.

	Sensitivity					
Specificity	**KL**	**pMI**	**pC**	**KL+pMI**	**KL+pC**	**Cls**
0.99	0.222	0.122	0.159	0.300	0.282	**0.315**
0.95	0.544	0.375	0.423	0.646	0.637	**0.667**
0.90	0.716	0.560	0.604	0.802	0.774	**0.816**
0.85	0.798	0.666	0.703	0.861	0.835	**0.862**

KL is the Kullback-Leibler conservation score, pMI is the proximity averaged mutual information score. pC is the proximity averaged conservation score, KL+pMI is the combined score of KL and pMI, KL+pC is the combined score of KL and pC, and Cls is the Catalytic likeliness score, The sensitivity is determined as an average over the 434 CSA families at the different specificity thresholds. In bold is highlighted the best performing method at each specificity level.
doi:10.1371/journal.pcbi.1000978.t003

Cls method did achieve the highest sensitivity. The difference in sensitivity between the Cls and the other methods is statistically significant ($p<0.05$, binomial test excluding ties) for all comparisons. The Cls score threshold corresponding to a specificity of 0.90 for the 434 CSA families is 1.44 ± 0.26. This low standard deviation of the threshold score indicates that the Cls approach is stable across the different CSA families and suggests that the method can be applied universally to any enzyme protein family independently of diversities in structure, composition and size of the MSA, as long as the number of sequences is greater than 10 (see figure 1).

Discussion

Catalytic residues are in general expected to be conserved and located in the functional site of a protein in order to attain their function. However, many non-catalytic residues are highly conserved as well and conversely, not all catalytic residues are conserved throughout a given protein family, making identification of catalytic residues a big challenge. The requirement to maintain a given catalytic function during the course of evolution places great limitations on the diversity of the structural environment of an active site. Therefore, here we put forward the hypothesis that catalytic residues carry a particular signature defined by networks of close spatial proximity residues sharing high mutual information, so that this signature could be applied to differentiate functional from other non-functional conserved residues.

We tested this hypothesis using a data set of 434 Pfam families each characterized by a PDB structure and one or more catalytic residues assigned from the CSA database, and investigated whether mutual information could complement conventional amino acid conservation scores and improve the ability to detect catalytic residues. Three methods to calculate sequence conservation were considered and the KL relative entropy (KL) was shown to significantly outperform both the Shannon entropy and maximal frequency measurements. We observed that sequence-weighting and low count correction do not improve the predictive performance for any of the methods. Additionally, in order to achieve reliable predictions the number of sequences required in the MSA was found to be relatively small. Only 10 sequences in the MSA were needed to reach AUC values of 0.89.

We observed that in the proximity of a catalytic site, residues are rich in shared mutual information (calculated as the cumulative mutual information, cMI): therefore, we defined a residue specific score characterizing this fact in terms of a structural proximity average (termed pMI) score. The pMI score was demonstrated to be a strong predictor for catalytic residues, suggesting that catalytic residues indeed carry a particular signature imposed by networks of mutual information. We compared the predictive performance of the pMI measure to that of a proximity measure based on sequence conservation and demonstrated that the two measures achieved comparable predictive performance but more importantly that they carried distinct information suitable as predictor of catalytic residues. Finally, we demonstrated that the conventional KL relative entropy sequence conservation, the pC and pMI measurements are complementary and that a combined catalytic likeliness score (Cls) of the three leads to significantly improved prediction accuracy. For instance, we found that, at a specificity threshold of 0.90, the KL, pMI, pC and Cls methods have a sensitivity of 0.716, 0.560, 0.604 and 0.816, respectively.

This work thus demonstrates in direct quantitative terms (gain in predictive performance) the contribution of the coevolution signal in determining catalytic residues, and hence goes beyond earlier published papers in the field [20,21,25,26] and not only describe the observation that such signals might be present near functionally important residues but in details demonstrate how such information can be applied to guide their identification.

We also analyzed to what extent the score characterizing specificity defining positions (SDPs) and the mutual information derived score defined in this work carry distinct information on the functional neighbor of catalytic residues. We used data from the Paradox database to carry out the comparison, and compared SDP and cMI scores for a set of 158 families covered by both methods. The obtained results clearly demonstrated that the SDP and cMI measures are different in nature, and that SDR Z-score is highly related to sequence conservation whereas the cMI score is independent of the latter. This observation strongly suggests that the cMI measure is more information rich for the identification of functional residues compared to SDP when combined with sequence conservation.

In summary, we have demonstrated that mutual information provides a distinct proximity signature that can be applied to determine catalytic residues. The approach outlined is general, and we suggest that the method should be applicable to the identification of other classes of functional residues where the requirement to maintain a particular function places limitations on the diversity of the structural environment along the course of evolution.

Materials and Methods

Dataset

The dataset was constructed based on the CSA database (version 2.2.11, released August 2009) [32]. CSA provides catalytic site annotation for enzymes in the PDB. Catalytic residues were defined as those residues thought to be directly involved in some aspect of the reaction catalysed by an enzyme (for a detailed description of the classification see [1]). The database consists of two types of annotated sites: an original, hand annotated set and an additional homologous set, containing annotations inferred by Psi-Blast and sequence alignment to one of the original entries. CSA contains 968 original literature entries, which belong to 455 Pfam families [33]. Due to some inconsistency between CSA and PDB, a few families were eliminated, so that we ended up with a dataset of 434 protein families (each of one containing at least one PDB entry), which in turn include a total of 1212 CSA, annotated catalytic residues. For 9 of the 434 families the selected PDB representative was an NMR structure. For these PDB entries the first model was

selected to represent the structure. The 434 Pfam families included in the benchmark data set cover 8 SCOP classes, 199 folds, 249 super families and 389 families.

When more than one PDB entry with catalytic site annotation was available for a given family, one reference PDB entry was selected following the criteria: highest sequence coverage of the Pfam MSA, the year of structure determination (preferably later than 2000) and resolution (Supplementary table S1 provides the Pfam family and reference PDB). In all cases, MSAs were gap trimmed to remove positions with gaps in the reference sequence. In addition, all positions with >50% gaps, as well as sequences covering <50% of the reference sequence length were removed, as described in [29]. Supplementary figure S1 shows the distribution of the number of sequences and sequence clusters in the dataset.

Conservation

Conservation of each position in the MSA's was calculated with three different measurements: Shannon entropy [34], KL relative entropy [35] calculated using an amino acids background frequency distribution obtained from the Uniprot database [36] and the maximal frequency (the frequency of the most represented amino acid). Each of these measurements were calculated from the raw MSA, from the MSA corrected for sequence redundancy using sequence weighting by 62% identity clustering (c), from the MSA including pseudo-counts to correct for low counts (l) [37,38] and from the MSA applying both clustering and pseudo-count correction (cl). The total number of conservation measurements investigated was hence twelve.

Mutual information

Mutual information (MI) was calculated as described in [29]. In short, the MI is calculated between pairs of columns in the MSA. The frequency for each amino acid pair is calculated using techniques of sequences weighting and low count corrections and is compared to the expected pair-frequency assuming that the amino acids are non-correlated. Next, the MI is calculated as a weighted sum of the log-ratios between the observed and expected amino acids pair frequencies. The APC method of Dunn et al. [28] was applied to reduce the background mutual information signal for each pair of positions and the MI scores were finally translated into MI Z-scores by comparing the MI values for each pair of position to a large set of MI values calculated from permutated MSA. MI gives a value for each pair of residues in a MSA. We sought a mutual information score per residue that characterizes the extent of mutual information "interactions" in its physical neighbourhood. This score was defined in two steps. First, we calculated a cumulative mutual information score (cMI) for each residue as the sum of MI values above a certain threshold for every amino acid pair where the particular residue appears. This value defines to what degree a given amino acid takes part in a mutual information network. Next, we defined a proximity average for each residue as the average of cMI of all the residues within a certain physical distance to the given amino acid. Finally, we normalized the proximity average values for a given MSA to fall in the range [0–1] to obtain the proximity MI (pMI) score. The distance between each pair of residues in the structure was calculated as the shortest distance between any two atoms different from H belonging to each of the two residues.

Combined catalytic likeliness score

We define a combined catalytic likeliness score (Cls) as a weighted sum of the conservation (defined in terms of the KL relative entropy), the proximity mutual information (pMI) and the proximity conservation (pC) scores.

$$Cls = (1 - w_{MI} - w_C) \cdot KL + w_{MI} \cdot pMI + w_C \cdot pC$$

Here, pC is the average conservation score of residues within a given proximity distance, and w_C, and w_{MI} are adjustable relative weights.

Parameter optimization

The calculation of the combined catalytic likeliness score depends on three parameters; Z_{thr} (Z-score threshold for including an amino acids pair in the cMI score), D_{MI} (distance threshold to include an amino acid in the pMI average score), D_C (distance threshold to include an amino acid in the pC average score), and the relative weights, w_{MI} and w_C, on pMI and pC, respectively. These parameters were estimated using five-fold cross validation, where optimal values were obtained using brute force grid-sampling on 4/5 of the data set to optimize the average AUC value and the remaining 1/5 of the data was evaluated next using this set of optimal parameters. This procedure was repeated five times leading to five sets of optimal parameters and evaluation performance values for each MSA in the data set.

Measurement of predictive performance

The predictive performance in detecting catalytic residues, by way of conservation, pMI and Cls, was evaluated in terms of the area under the ROC curve (AUC) [39] per family. The AUC measure might not be optimal if the benchmark data set has a high ratio on negative data, and a high specificity in actual number could translate into a large number of false positive. In such situations, it might be beneficial to use only the high specificity part of the ROC curve to calculate the predictive performance. Here, we hence complement the AUC measure with AUC01 calculated including only the specificity range for 1 to 0.9 when calculating the AUC. For both measures will a value of 1 indicate a perfect prediction while a value of 0.5 indicates a random prediction. Annotated catalytic residues in the CSA were taken as the positive set, and all other residues with annotated PDB-ATOM coordinates were assigned as negative. The final performance was determined as the average AUC over the 434 CSA Pfam families.

Comparison between SDPs and cMI scores

We downloaded the entire Paradox SDR database (specificity-determining residues in protein families database; http://paradox.harvard.edu/sdr/), and identified the subset of families present in our benchmark dataset where the reference sequence from the CSA database was also member of the paradox multiple sequence alignment (MSA). This gave us a set of 158 families. The Paradox database provides SDR Z-scores only for a subset of the positions in the MSA [30]. Residues with undefined SDR Z-score were assigned a Z-score of 0 to allow for complete sequence coverage. Next, we compare for each family the SDR Z-score value to our cMI (cumulative mutual information) value of each position in the alignment in terms of the Spearmans rank correlation. We also calculate the Spearmans rank correlation between KL and both SDR Z-score and cMI values of each position for each family in the dataset.

Supporting Information

Figure S1 Histogram of the number of families in the Pfam benchmark data set. A) number of sequences B)number of clusters. The insets show a zoom from 0 to 1,000 sequences/clusters.

Found at: doi:10.1371/journal.pcbi.1000978.s001 (0.02 MB PDF)

Table S1 Pfam PDB correlation. Pfam accession, PDB taken as reference for that family, and pdb region included in the analysis.
Found at: doi:10.1371/journal.pcbi.1000978.s002 (0.03 MB PDF)

Table S2 Performance details of all methods included in the analysis. Cons and C means conservation; pMI: proximity MI; pC: proximity conservation, Cls: catalytic likeliness score; Nseq: number of sequences; Ncluster:number of clusters; pdb: pdb taken as reference.

Found at: doi:10.1371/journal.pcbi.1000978.s003 (0.16 MB XLS)

Author Contributions

Conceived and designed the experiments: CMB MN. Performed the experiments: CMB ET TDD MN. Analyzed the data: CMB ET TDD MN. Contributed reagents/materials/analysis tools: CMB MN. Wrote the paper: CMB ET JMD MN.

References

1. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM (2002) Analysis of Catalytic Residues in Enzyme Active Sites. J Mol Biol 324: 105–121.
2. Innis CA, Anand AP, Sowdhamini R (2004) Prediction of Functional Sites in Proteins Using Conserved Functional Group Analysis. J Mol Biol 337: 1053–1068.
3. Zhang T, Zhang H, Chen K, Shen S, Ruan J, et al. (2008) Accurate sequence-based prediction of catalytic residues. Bioinformatics 24: 2329–2338.
4. Chien T-Y, Chang DT-H, Chen C-Y, Weng Y-Z, Hsu C-M (2008) E1DS: catalytic site prediction based on 1D signatures of concurrent conservation. Nucl Acids Res 36: W291–296.
5. Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol 396: 1451–1473.
6. Mihalek I, Res I, Lichtarge O (2004) A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance. J Mol Biol 336: 1265–1282.
7. Manning J, Jefferson E, Barton G (2008) The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. BMC Bioinformatics 9: 51.
8. Sterner B, Singh R, Berger B (2007) Predicting and Annotating Catalytic Residues: An Information Theoretic Approach. J Comput Biol 14: 1058–1073.
9. Petrova N, Wu C (2006) Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics 7: 312.
10. Bernardes J, Fernandez J, Vasconcelos A (2008) Structural descriptor database: a new tool for sequence-based functional site prediction. BMC Bioinformatics 9: 492.
11. Cilia E, Passerini A (2010) Automatic prediction of catalytic residues by modeling residue structural neighborhood. BMC Bioinformatics 11: 115.
12. Kristensen D, Ward RM, Lisewski A, Erdin S, Chen B, et al. (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinformatics 9: 17.
13. Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K. Active site prediction using evolutionary and structural information. Bioinformatics 26: 617–624.
14. Matthew Ward R, Venner E, Daines B, Murray S, Erdin S, et al. (2009) Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25: 1426–1427.
15. Sankararaman S, Sjolander K (2008) INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24: 2445–2452.
16. Tang Y-R, Sheng Z-Y, Chen Y-Z, Zhang Z (2008) An improved prediction of catalytic residues in enzyme structures. Protein Eng Des Sel 21: 295–302.
17. Tong W, Wei Y, Murga LF, Ondrechen MJ, Williams RJ (2009) Partial Order Optimum Likelihood (POOL): Maximum Likelihood Prediction of Protein Active Site Residues Using 3D Structure and Sequence Properties. PLoS Comput Biol 5: e1000266.
18. Alterovitz R, Arvey A, Sankararaman S, Dallett C, Freund Y, et al. (2009) ResBoost: characterizing and predicting catalytic residues in enzymes. BMC Bioinformatics 10: 197.
19. Byung-Chul L, Keunwan P, Dongsup K (2008) Analysis of the residue-residue coevolution network and the functionally important residues in proteins. Proteins: Structure, Function, and Bioinformatics 72: 863–872.
20. Kuipers RK, Joosten HJ, Verwiel E, Paans S, Akerboom J, et al. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins 76: 608–616.
21. Gloor GB, Martin LC, Wahl LM, Dunn SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry 44: 7156–7165.
22. Lockless SW, Ranganathan R (1999) Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science 286: 295–299.
23. Shi Z, Resing KA, Ahn NG (2006) Networks for the allosteric control of protein kinases. Curr Opin Struct Biol 16: 686–692.
24. Chakrabarti S, Panchenko AR (2009) Coevolution in defining the functional specificity. Proteins: Structure, Function, and Bioinformatics 75: 231–240.
25. Rausell A, Juan D, Pazos F, Valencia A. Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc Natl Acad Sci U S A 107: 1995–2000.
26. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257: 342–358.
27. Gouveia-Oliveira R, Pedersen AG (2007) Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms Mol Biol 2: 12.
28. Dunn SD, Wahl LM, Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24: 333–340.
29. Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics 25: 1125–1131.
30. Donald JE, Shakhnovich EI (2005) Determining functional specificity from protein sequences. Bioinformatics 21: 2629–2635.
31. Leys D, Tsapin AS, Nealson KH, Meyer TE, Cusanovich MA, et al. (1999) Structure and mechanism of the flavocytochrome c fumarate reductase of Shewanella putrefaciens MR-1. Nat Struct Biol 6: 1113–1117.
32. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl Acids Res 32: D129–133.
33. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucl Acids Res 36: D281–288.
34. Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–423.
35. Cover TM, Thomas JA (1991) Elements of information theory , Inc EJWS, editor.
36. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142–148.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25: 3389–3402.
38. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, et al. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics 20: 1388–1397.
39. Swets J (1988) Measuring the accuracy of diagnostic systems. Science 3: 1285–1293.

Teppa et al. BMC Bioinformatics 2012, 13:235
http://www.biomedcentral.com/1471-2105/13/235

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction

Elin Teppa[1], Angela D Wilkins[2], Morten Nielsen[3,4†] and Cristina Marino Buslje[1*†]

Abstract

Background: A large panel of methods exists that aim to identify residues with critical impact on protein function based on evolutionary signals, sequence and structure information. However, it is not clear to what extent these different methods overlap, and if any of the methods have higher predictive potential compared to others when it comes to, in particular, the identification of catalytic residues (CR) in proteins. Using a large set of enzymatic protein families and measures based on different evolutionary signals, we sought to break up the different components of the information content within a multiple sequence alignment to investigate their predictive potential and degree of overlap.

Results: Our results demonstrate that the different methods included in the benchmark in general can be divided into three groups with a limited mutual overlap. One group containing real-value Evolutionary Trace (rvET) methods and conservation, another containing mutual information (MI) methods, and the last containing methods designed explicitly for the identification of specificity determining positions (SDPs): integer-value Evolutionary Trace (ivET), SDPfox, and XDET. In terms of prediction of CR, we find using a proximity score integrating structural information (as the sum of the scores of residues located within a given distance of the residue in question) that only the methods from the first two groups displayed a reliable performance. Next, we investigated to what degree proximity scores for conservation, rvET and cumulative MI (cMI) provide complementary information capable of improving the performance for CR identification. We found that integrating conservation with proximity scores for rvET and cMI achieved the highest performance. The proximity conservation score contained no complementary information when integrated with proximity rvET. Moreover, the signal from rvET provided only a limited gain in predictive performance when integrated with mutual information and conservation proximity scores. Combined, these observations demonstrate that the rvET and cMI scores add complementary information to the prediction system.

Conclusions: This work contributes to the understanding of the different signals of evolution and also shows that it is possible to improve the detection of catalytic residues by integrating structural and higher order sequence evolutionary information with sequence conservation.

Keywords: Coevolution, Mutual information, Specificity determining position, Catalytic residues, Functional sites, Sequence analysis

* Correspondence: cmb@leloir.org.ar
†Equal contributors
[1]Fundación Instituto Leloir, Avda. Patricias Argentinas 435, CABA C1405BWE, Argentina
Full list of author information is available at the end of the article

BioMed Central

© 2012 Teppa et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

A number of methods have been developed to predict functionally important sites in protein families based on sequence and structure information. The importance of a particular residue in a protein can be due to many different factors, including structural stability, protein-protein interaction, protein-DNA/RNA interaction, ligand binding site and maintenance of protein functions.

In most cases, it is difficult to assign a particular function to a particular residue or group of residues, as function is determined by a subtle interplay between multiple residues and mutation to any of them might impact the protein function and/or structure. In some cases however, the association between a particular residue and a protein function can be readily recognized. One such example being catalytic residues, where large data set exist defining residues within a given protein sequence linked to a given catalytic function [1].

Most methods developed to predict functionally important sites in protein families rely on some signal related to protein evolution. Three clear signals of evolution are: conservation, conservation within specific groups of sequences sharing a common function, and coevolution between residues (see Figure 1). Conservation is straightforward to calculate and interpret. A change in a conserved position (even when proteins are highly diverse) should have a deleterious effect on the protein function. Specificity determining positions (SDPs) are those positions within multiple sequence alignments (MSAs) that are conserved within groups of proteins that perform the same function (specificity groups) and varying between groups with different functions/specificities. These sites generally determine protein specificity either by binding specific substrate/inhibitor or through interaction with other protein [2-4].

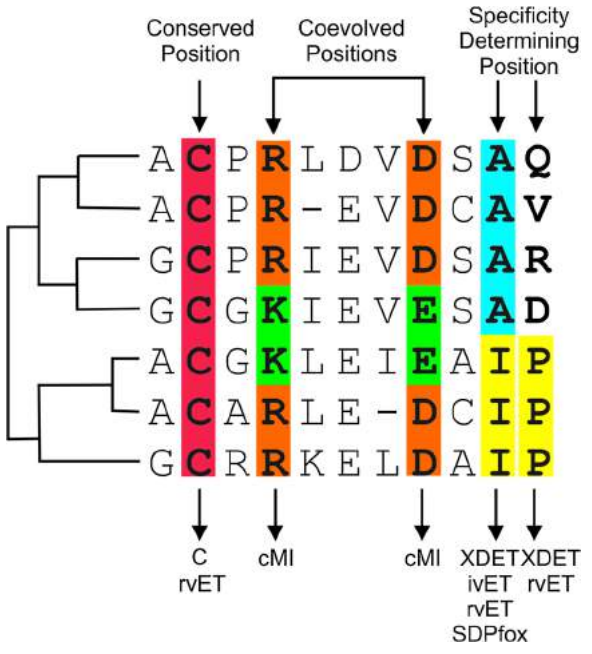


Figure 1 Different column patterns in an MSA. Schematic representation of an MSA and its phylogenetic tree (left). Conserved position is highlighted in red, coevolved positions in green and orange and putative SDPs in yellow and blue. On the top are indicated the column pattern and on the bottom, the suitable method to detect each kind of position (C: conservation score; cMI: cumulative MI; ivET: integer value ET; rvET: real value ET; XDET and SDPfox are also indicated).

The degree of co-evolution between pairs of residues is commonly estimated using a measure of mutual information (MI). If two residues share high signal of mutual information, the two residues most likely are co-evolving, meaning that in order to maintain a given protein function a mutation of one residue is linked to a specific compensatory mutation of the other residue.

Several methods to predict specificity-determining positions have been developed. Many of these require a previous classification of the proteins into functional groups [3,5,6], which is a problematic limitation since the specificity of a given protein is unavailable in the great majority of cases and is non-trivial to calculate and validate. To overcome this problem, methods have been developed that group the sequences in a MSA upon certain ad-hoc criteria [7,8]. As an example, Capra and Singh [9] addressed the classification problem using a combination of Pfam, EC numbers and sequence similarity. There are also methods where the clustering is based on sequence similarity alone [10] or Bayesian statistics [11]. Many of these methods approximate the classification of sequences using phylogeny [12-14] or a combination of phylogenetic information and entropies analysis [15]. Other methods rank residues by their relative importance in the MSA [12,13,16-18]. These approaches differ in design, but all look for specific patterns of amino acids conservation as indicators of likely functional importance.

Finally, inter-relationship between two or more positions (estimated using mutual information) can contribute a different type of biological information related to protein function and functional importance of specific residues. We have earlier introduced a cumulative mutual information concept (cMI) that measures the degree of shared mutual information of a given residue and the proximity mutual information (p(MI)) which measures the amount of shared mutual information in the proximity of a given residue [16]. In a large benchmark data set of enzymatic protein families, we showed that whereas identification of catalytic residues (CR) is strongly guided by sequence conservation, mutual information (or coevolution) provides an additional and complementary signal that significantly improves the predictive power.

A large panel of methods thus exists aiming at identifying residues with critical impact on protein functionally relying on measures of information content extracted from multiple sequence alignments. However, it is not clear to what extent the predictive power of the different methods overlap, and if any of the methods have higher predictive potential compared to others when it comes to the identification of a particular type of functional important sites. Here, we aim at addressing this question by comparing the ability to identify CR in enzymatic proteins of different information-based methods. Although CR clearly do not constitute the sole test-case to perform such an investigation we have chosen this test-case due to the large data sets of unambiguous annotations of functionally important residues available.

Using this test-case of CR identification, we seek to decompose and compare the predictive signal of a series of unsupervised (i.e. methods that do not require prior functional clustering) information-based predictions method. The analysis includes on the one hand, methods aim at ranking residues by their functional importance using i) conservation; ii) mutual information [16] and iii) evolutionary trace real value (rvET) that incorporates evolutionary and entropic information from multiple sequence alignments [13]. On the other hand, we include methods aimed at detection of specificity positions i.e. i) the evolutionary trace integer value (ivET) score that represents conservation within groups in a qualitative manner [12]; ii) SDPfox [10] that predicts SDPs in a phylogeny-independent manner and iii) XDET [19] that is based on the comparison of the mutational behavior of a position with the mutational behavior of the full-length protein MSA, by directly comparing the corresponding distance matrices.

Comparing these methods will allow us to break up the different components of information content included in a MSA, investigate to what degree they overlap and estimate their predictive potential for the identification of active site residues in catalytic proteins.

Results and discussion

The analysis is based on a set of 424 enzymatic Pfam families earlier described by Marino Buslje (2010) [16] (for details see Methods). In short, each family is characterized by a (MSA) taken from Pfam [20], has an annotated set of CR taken from the CSA database [1] and by having a known three-dimensional structure for at least one of its members. Given this data set, we calculated measures related to evolution for the different methods included in the benchmark, and next analyzed the overlap/correlation between these measures and their predictive potential for identification of CR in proteins.

Although all the methods are intended to identify functionally important sites within protein families, they can be divided into two major groups: the methods that rank positions in the MSA according to their relative functional importance within the protein family, no matter what this importance might be due to. In this category falls the cumulative mutual information (cMI), real-value evolutionary trace (rvET) and sequence conservation (C). The other group consists of methods intended to predict specificity determining positions in a family of proteins and includes XDET, ivET and SDPfox.

Concordance of the different predictions methods

To determine the influence data redundancy might have on the prediction scores for the different methods, we measured the correlation between scores calculated on the MSAs as retrieved from Pfam (MSA100) and on a set of sequence redundancy reduced MSAs (for details see materials and methods). If a given method is insensitive to sequence redundancy, the scores produced from the different MSAs should be highly correlated. This is true for cMI (Spearman's rank correlation coefficient, SCC = 0.76) and rvET (SCC = 0.93). However, for ivET we found only a weak correlation between the scores obtained using the two data sets (SCC = 0.21) indicating that data redundancy for this method strongly impacts the predictive output (see Figure 2, Additional file 1 Table S1 and Additional file 2 Figure S1).

Methods for prediction of SDPs aim at estimating a score that correlates with the functional importance of a given residue in terms of protein specificity. Another critical question to analysis is therefore the degree of concordance between different prediction methods aiming at identifying SDPs. From Figure 2, it is clear that the methods for SDP identification (ivET, SDPfox and XDET) show limited mutual overlap. The correlations values are low for all comparisons, with the highest value of 0.34 being between SDPfox and XDET.

Next, we investigated to what degree the information extracted by the methods developed for detection of SDPs (ivET, SDPfox, XDET) overlapped with the information signal of cMI, which points out positions with a high degree of shared mutual information. We found that cMI has a low overlap with all these methods (SCC < 0.28 for every comparison, see Figure 2 and Additional file 1 Table S1).

We next analyzed the correlation between methods aimed to rank the residues by functional importance (rvET, cMI and conservation). As expected, conservation was strongly correlated with rvET [13] for both the MSA100 and MSA62 (redundancy reduction at 62%) data sets (SCC > 0.7, in both cases). cMI was found to be weakly correlated with conservation (SCC 0.16 for both MSA100 and 62), and finally the overlap between the

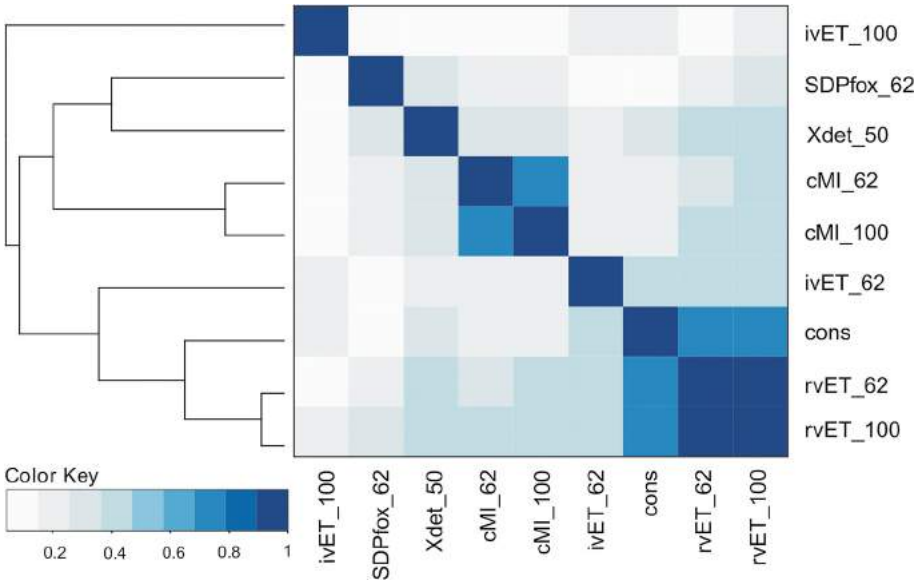


Figure 2 Heat map representation of the Spearman rank correlation coefficient between methods. cMI: cumulative MI, ivET: integer value evolutionary trace, rvET: real value evolutionary trace, cons: conservation. Numbers following the methods name (100, 62 and 50) indicate the redundancy of the sequences in the MSA (100, 62 and 50% redundancy reduced). The dendrogram indicates the distance between methods. Correlation colour key goes from white (0, no correlation) to blue (1, perfect correlation). All correlations are statistically different from zero (T-test, p-value threshold of 0.05).

rvET and cMI methods was moderately weak with a maximal correlation of 0.41.

The above results demonstrate that the different methods included in the benchmark in general can be divided into major groups with a limited mutual overlap. One group containing the methods with a signal highly correlated to sequence conservation (rvET, conservation and ivET when was evaluated on redundancy reduced data). Another group containing the methods which signal is derived from mutual information (cMI). The methods designed explicitly for the identification of SDPs (SDPfox and XDET) have low correlation to any other method included in the benchmark. The ivET method evaluated on the MSA100 data set (ivET100) appears as an outlier in this analysis and does not show overlap with any other method. The results hence in general underline that the overlap between the different methods in most cases is limited, suggesting that high ranked cMI and SDPs do not necessarily form the same group of residues. Also, it is noticeable that methods aimed to detect the same kind of positions as SDPs (ivET, SDPfox, and XDET) display rather low mutual concordance in prediction scores.

Proximity summed information measures for predicting catalytic residues

We have earlier demonstrated that CR are characterized by a structural proximity with high mutual information, i.e. that residues within a certain distance threshold of CR are rich in shared MI [16]. To investigate if similar observations can be made for the other information measures included in the benchmark, we calculated a proximity measure of each method and investigated to what degree this measure contributed to the identification of CR. For each residue, we calculated the proximity score as the sum of the scores of the residues located within a certain distance from the residue in question (see equation 1)

$$pMI_i = \sum_{j,d_{ij}<t} cMI_j \qquad (1)$$

where the sum is over all residues j in the given protein within a distance $d_{ij} < t$ from the residue i, where d_{ij} is the shortest distance between any pair of heavy atoms of two residues i and j, cMI_j is the cumulative mutual information score of residue j, and t a distance threshold. Those measures are designated with a "p" preceding the name of each method, i.e.: p(rvET) for proximity rvET, p(ivET) for proximity ivET, p(C) for proximity C, and p(MI) for proximity cMI. The threshold distance t was optimized for each prediction method.

Table 1 gives the results of the calculation, and shows that all methods with the exception of p(ivET) evaluated on the MSA100 data set, SDPfox and XDET could be used as reasonable predictors of CR (AUC > 0.8 for all the methods). Note that we also here observe a large difference in the performance of the ivET method when

Table 1 Performance and optimal distance threshold of the proximity measures for detecting catalytic residues

Method	AUC average	Distance cutoff (Å)
p(SDPfox62)	0.703	12
p(XDET50)	0.736	8
p(ivET62)	0.835	7
p(ivET100)	0.640	7
p(rvET62)	0.878	5
p(rvET100)	0.875	7
p(MI62)	0.823	7
p(MI100)	0.833	7
p(C)	0.854	5

The number of protein families included in SDPfox is 289, 298 in XDET and 424 in all other methods. "p" before the method's name denotes "proximity". The number following the method's name denotes the MSA data set on which the method was evaluated (ie: 50 = MSA50). The optimal distance cut-off for the proximity sum was found using a grid-search as described in Methods.

evaluated on the different sets of MSAs. The results in Table 1 shows that all high performing methods (AUC > 0.8) have a optimal proximity distance threshold between 5 and 7 Å. Only for the SDPs prediction methods which all have poor predictive performance is the distance threshold larger.

We can next investigate to what degree the performance of the different methods is statistically different. In doing this, we obtain the following rank of the methods:

$$\begin{aligned} p(rvET62) \approx p(rvET100) \approx p(C) \quad &> p(ivET62) \\ \approx p(MI100) \approx p(MI62) &> p(XDET50) \\ \approx p(SDPfox62) &> p(ivET100) \end{aligned}$$

Here a "≈" means that the preceding value is higher but not statistically different and ">" means significantly higher, where statistical tests were conducted as binomial tests excluding ties using a p-value threshold of 0.05. The different methods hence fall in three different groups a) p(rvET62), p(rvET100) and p(C), b) p(ivET62), p(MI100) and p(MI62), c) p(XDET50) and p(SDPfox62), with p(ivET100) as an outlier.

Combined catalytic likeliness Score (Cls) with the best performing distance threshold for each method and optimizing the weight for each term

We have demonstrated in a previous work how the p(C) and p(MI) scores when integrated with conservation enhance the predictive performance for identification of CR [16]. Here, we aimed at demonstrating to what degree this observation is maintained when integrating the other methods included in the benchmark with the conservation score. In this way, we can in a simple manner investigate to what degree each method adds complementary information to the final prediction model. We defined different combined models by adding one or more proximity scores to the conservation score. For each Pfam family, the additional feature was normalized so that the values fell in the range [0–1] (for details see Methods). We included in this benchmark p(MI62) (previously used for CR detection [16]), the best performing ET method p(rvET62), and p(C).

Table 2 gives the performance values in terms of the AUC (area under the ROC curve) and AUC0.1 (area under the ROC curve integrated up to a false positive rate of 0.1) and optimal relative weights (estimated using 5 fold cross-validation) for the different models. The $0.2 \cdot C + 0.8 \cdot p(C)$ row hence gives the optimal performance for the model defined as a combination of conservation (C) and the proximity sum of conservation (p(C)), and states that the optimal relative weight of the two terms is 0.2 on conservation and 0.8 on p(C), respectively. In the table, a weight equal to 0 indicates that a given score did not contribute to the performance of the model.

Several observations can be made from these results. First of all, it is clear that all proximity scores contain complementary information that when combined with conservation (C) leads to an improved predictive performance (all models C + p(XX), where XX equals C, rvET62, or MI62 significantly outperform the model based on conservation only, $p < 0.05$ one-tailed binomial test excluding ties). Also, it is striking to observe that the relative weight on the p(C) score in all models including p(rvET) is zero. This strongly suggests that the high performance of the p(rvET) method shown in Table 1 is driven by the signal of sequence conservation contained within the rvET score (as also suggested from the correlation analysis in Figure 2). The model C+p(C) +p(MI), achieved a higher performance than the corresponding model C + p(C) + p(rvET), the difference is however not statistically significant ($p > 0.1$, one-tailed binomial test excluding ties). Finally, the model C + p(rvET) + p(MI) integrating both the cMI and rvET proximity scores had the highest performance of all models

Table 2 Performance of different methods in terms of the AUC

Method	AUC	AUC0.1
C	0.881	0.491
0.2 C + 0.8 p(C)	0.898	0.553
0.15 C + 0.85 p(rvET62)	0.913	0.567
0.25 C + 0.75 p(MI62)	0.912	0.555
0.15 C + 0.0 p(C) + 0.85 p(rvET62)	0.913	0.567
0.15 C + 0.3 p(C) + 0.55 p(MI62)	0.916	0.571
0.15 C + 0.0 p(C) + 0.45 p(rvET62) + 0.4 p(MI62)	0.921	0.586

Methods give the optimal combined model including conservation and the different proximity scores. The relative weights were determined using fivefold cross validation as described in the text. AUC and AUC0.1 are the average performance values over the 424 protein families.

included in the benchmark, and significantly outperformed all other models, except C + p(C) + p(MI) ($p < 0.05$ in all case, one-tailed binomial test excluding ties). In terms of specificity and precision, the C + p (rvET) +p(MI) method had average performance values of 0.88 and 0.25 respectively at a sensitivity level of 0.98 when evaluated on the 424 Pfam data sets. At a sensitivity level of 0.55 these values are altered to 0.96 and 0.45, respectively. For comparison, these values are reduced to 0.82/0.20 and 0.93/0.35, respectively, using the model defined from conservation only. Note, that these conclusions are maintained integrating multiple residue-specific information measures (conservation, rvET, and cMI) with the corresponding proximity scores. Doing this, we confirm that both the rvET and cMI measures carries complementary information, and that this complementarity is captured both at the per-residue and proximity level (data not shown).

Taken together, these observations demonstrate that the rvET and cMI scores capture distinct signals from the MSA and add complementary information to the prediction system.

Conclusions

Many algorithms have been proposed for the identification of residues critical for protein function in general and protein specificity in particular. Here, we have compared a series of such methods in terms of both the concordance between their predictions and their ability to identify catalytic sites in proteins with enzymatic function. From our results, we find that the methods included in the benchmark can be divided in three groups with limited mutual overlap. One group consists of methods which predictive signal is strongly correlated to sequence conservation (rvET, and sequence conservation itself), one group consists of the methods whose predictive signal is derived from mutual information (cMI), and the last group consists of the methods developed for prediction of specificity determining positions (SDPfox, XDET and ivET).

Defining a proximity score for each method as suggested by [16] and benchmarking for the ability to identify CR, we find that only methods from the first two of the above three groups displayed a reliable predictive performance (mean AUC value above 0.8), indicating that the methods from the SDP group has limited value for the identification of residues critical for protein function. Comparing the different methods for prediction of specificity determining positions we found that they shared limited mutual overlap despite the fact that they are designed to capture a common functional signal.

Finally, we investigated to what degree the information signal of conservation, rvET and cMI methods (belonging to the two well-performing groups of methods) was complementary so that the combined signal could significantly improve the predictive capacity. Here, we found that the predictive performance could be significantly improved when combining conservation with the signal from the proximity scores of the different methods. The best performing method was found to consist of a combination of sequence conservation and proximity scores for both rvET and cMI. This finding confirms the notion that the rvET and cMI methods are distinct in nature, and that the two methods add informative and complementary information to the prediction system. The benchmark however also demonstrated that the gain in predictive performance of the rvET signal is limited and insignificant if combined with the conservation proximity scores.

It is critical to emphasize that the conclusions obtained in this work are strictly related to the identification of CR in enzymatic proteins. Albeit the different methods for predicting SDPs do not correlate strongly in our dataset, some have proven to be successful in the predictions in small size benchmark data sets with a limited number of sequences and few specificity groups [21-23]. Capra et. al (2008) [9] obtained reasonable results in predicting SDPs using as true positive, predicted instead of experimentally determined SDPs. Also, successful prediction results were obtained by Rodriguez GJ (2010) [24], demonstrating with experimental verification that they were able to predict, with rvET, residues responsible for the specificity between dopamine and serotonine ligands in bioamine receptors of the Class A G-Protein coupled receptors family.

What remains an unquestionable result from our analysis is that prediction scores for the different methods evaluated share a limited overlap, and in particular that the methods for SDP identification and the method based on mutual information capture a very distinct signal of evolutionary information.

In conclusion, we believe this work contributes to: i) a better understanding of the different signals of evolution of a protein; ii) in a highly quantitative manner characterize similarities and differences between different information measures captured within a multiple sequence alignment and iii) demonstrates that it is possible to significantly improve the ability to detect CR by integrating these different types of information measures.

Methods

Data set construction

The data set comprise 424 Pfam multiple sequence alignments (MSA) with CR annotation in Catalytic Site Atlas database (version 2.2.11, released August 2009) [1] earlier published by [16]. CSA provides catalytic site annotation for enzymes in the PDB. Catalytic residues are defined as those residues directly involved in some

aspect of the reaction catalysed by an enzyme. Note, that the data is different from the original publication [16] due to parsing errors for 10 MSAs. We ended up with a dataset of 424 Pfam families which in turn include a total of 1328 CSA annotated catalytic residues. Each family has on average 3 CR (standard deviation 1.72) with a minimum of 1 and a maximum of 23 CR per family. The distribution of the number of sequences per families is shown in Additional file 3: Figure S2.

For each family at least one three-dimensional structure is known and this protein sequence was taken as a reference. When more than one PDB entry with catalytic site annotation was available for a given family, one reference PDB entry was selected following the criteria: highest sequence coverage of the Pfam MSA, the year of structure determination (preferably later than 2000) and resolution. Multiple sequence alignments were taken from Pfam [20] and pretreated by trimming deletions and insertions across the whole alignment so as to preserve the continuity of the reference sequence.

In order to investigate the effect of sequence redundancy on the different methods, we tested the performance of the MI and ET methods using the full multiple alignments (as retrieved form Pfam) as well as in a set of redundancy reduced alignments (reduced at 62% identity). SDPfox and XDET were tested only with a set of MSA 62% and 50% identity redundancy reduced respectively, due to their limitation in the number of input sequences allowed and the large runtime requirements (see below). The different benchmark data sets are named MSA100 (no redundancy reduction applied), MSA62 and MSA50 respectively. Redundancy reduced alignment were generated with T-Coffee software [25]. The complete data sets of MSAs for the 424 Pfam families, including catalytic site annotations is available at http://www.cbs.dtu.dk/suppl/immunology/CSA.

SDP prediction software

SDPs predictions were performed with: a) Integer value ET (ivET) score that represents conservation within groups in a qualitative manner [12]; b) SDPfox method that predicts SDPs in a phylogeny-independent manner [10]. The software was downloaded from http://bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp and run locally with default parameters. This method has a limitation on the number of specificity groups per family (between 2 and 200 specificity groups) and total length of the sequence (<500 residues), so the predictions for this method were hence made on the MSA62 data set; and c) XDET software is based on the comparison of the mutational behavior of a position with the mutational behavior of the whole alignment [19,26]. It furnishes two methods for detecting position related to functional specificity. Here, we used the mutational behavior (MB) method of XDET that does not use external arbitrary functional classification. Due to the high cost of the computer time of the method (the running time grows quadratic with the number of sequences) it was only possible to run XDET on the MSA50 data set. Source code for XDET was obtained from the authors.

Methods of functionally important residues prediction

Prediction of functionally important residues was performed with the following methods: a) Sequence conservation, was calculated from the MSA100 as the Kullback–Leibler relative entropy [27] using an amino acids background frequency distribution obtained from the UniProt database (http://www.uniprot.org/); b) Mutual information was calculated in terms of the cumulative Mutual Information (cMI) score, that measures the degree of shared mutual information of a given residue [16]; c) Evolutionary Trace real value score (rvET) [24], which incorporates entropy as a quantitative measure of conservation giving a rank of positions by their relative importance.

Predictive performance

The predictive performance in detecting CR using the proximity scores was evaluated in terms of the area under the ROC curve (AUC) per family. Annotated CR in the CSA were taken as the positive set, and all the other residues were assigned as negative. Both the full AUC value and the value integrated for specificities from 1 to 0.9 were included to capture the high specificity performance of the different measures [28]. The overall predictive performance was evaluated as a simple average of the per-family obtained AUC values. Parameters for each model were optimized using fivefold cross validation.

Proximity summed scores

We calculated proximity scores for each method as a sum of the scores within a certain physical distance to the given amino acid. The distance between each pair of residues in the structure was calculated as the minimum distance between two heavy atoms. The optimal distance threshold for each proximity measure was found using a grid of 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 Å.

Derived scores to predict catalytic residues

To integrate different score with sequence conservation, a combined score was defined as (equation 2)

$$S = (1 - w)\ C + wX \qquad (2)$$

where w is a relative weight in the range [0–1]. For each protein family MSA, the additional feature was normalized so that the values fell in the range [0–1]. This single combination was made for X = {p(C), p(rvET62) and

p(MI62)}. Note that for p(ET), the formula is $S = (1 - w)\ C - w\ p(rvET62)$, as ET best rank is the smallest number.

When two features are added, the combined score was calculated as equation 3.

$$S = (1 - w_1 - w_2)C + w_1\ p(C) + w_2\ X \quad (3)$$

Where w_1 and w_2 are relative weights both in the range [0–1], and $w_1 + w_2 < 1$. Here the combination was made for X = {p(rvET62) and p(MI62)}. Also here the sign for the last term was negative when X = p(rvET62).

Finally, the complete combination of all methods was calculated as equation 4.

$$S = (1 - w_1 - w_2 - w_3)\ C + w_1\ p(C) - w_2\ p(rvET62) + w_3\ p(MI62) \quad (4)$$

where w_1, w_2 and w_3 are relative weights in the range [0–1] and $|w_1 + w_2 + w_3 < 1$.

Additional files

Additional file 1: Table S1. Spearman rank correlation between methods and their standard deviation.

Additional file 2: Figure S1. Schematic representation of the sequence redundancy impact on ivET predictions.

Additional file 3: Figure S2. Distribution of the number of Pfam families vs number of sequences per family.

Competing interests

The author(s) declare that they have no competing interests.

Author's contributions

ET: Contributed to the acquisition, analysis and interpretation of data, results discussion and drafting the manuscript. AW: acquisition of data, discussion and manuscript revision; MN and CMB: conceived the study, participated in its design and coordination, results analysis, discussions and writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgments

ET, was partially supported by a fellowship from the Lounsbery Foundation. ADW, was supported by training fellowships from the National Library of Medicine to the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NLM grant 5T15LM07093).
MN and CMB are supported by the National Research Council of Argentina (CONICET).

Author details

[1]Fundación Instituto Leloir, Avda. Patricias Argentinas 435, CABA C1405BWE, Argentina. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas. [3]Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark. [4]Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, San Martín, B 1650 HMP, Buenos Aires, Argentina.

Received: 8 May 2012 Accepted: 5 September 2012
Published: 14 September 2012

References

1. Porter CT, Bartlett GJ, Thornton JM, The Catalytic Site Atlas: *Nucleic Acids Res* 2004, **32**:129–133. Database issue.
2. Oliveira L W, Vriend G, Ljzerman AP: **Identification of class-determining residues in G protein-coupled receptors by sequence analysis.** *Receptors Channels. 5th edition.* 1997, **5**(3-4):159–174.
3. Pirovano W, Feenstra KA, Heringa J: **Sequence comparison by sequence harmony identifies subtype-specific functional sites.** *Nucleic Acids Res* 2006, **34**(22):6540–6548.
4. Chakrabarti S, Panchenko AR: **Coevolution in defining the functional specificity.** *Proteins* 2009, **75**:231–240.
5. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Mol Biol* 1995, **2**(2):171–178.
6. Hannenhalli SS, Russell RB: **Analysis and prediction of functional sub-types from protein sequence alignments.** *J Mol Biol* 2000, **303**(1):61–76.
7. Brown DP, Krishnamurthy N, Sjolander K: **Automated protein subfamily identification and classification.** *PLoS Comput Biol* 2007, **3**:e160.
8. Wicker N, *et al*: **Secator: A Program for Inferring Protein Subfamilies from Phylogenetic Trees.** *Mol Biol Evol* 2001, **18**(8):1435–1441.
9. Capra JA, Singh M: **Characterization and prediction of residues determining protein functional specificity.** *Bioinformatics* 2008, **24**:1473–1480.
10. Mazin P, *et al*: **An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies.** *Algorithms for Molecular Biology* 2010, **5**(1):29.
11. Marttinen P, *et al*: **Bayesian search of functionally divergent protein subgroups and their function specific residues.** *Bioinformatics* 2006, **22**:2466–2474.
12. Lichtarge O, Bourne HR, Cohen FE: **An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families.** *J Mol Biol* 1996, **257**(2):342–358.
13. Mihalek I, Res I, Lichtarge O: **A Family of Evolution-Entropy Hybrid Methods for Ranking Protein Residues by Importance.** *J Mol Biol* 2004, **336**(5):1265–1282.
14. Pei J, *et al*: **Prediction of functional specificity determinants from protein sequences using log-likelihood ratios.** *Bioinformatics* 2006, **22**:164–171.
15. Ye K, Vriend G, Ljzerman AP: **Tracing evolutionary pressure.** *Bioinformatics* 2008, **24**(7):908–915.
16. Marino Buslje C, *et al*: **Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification.** *PLoS Comput Biol* 2010, **6**(11):e1000978.
17. Morgan DH, *et al*: **ET viewer: an application for predicting and visualizing functional sites in protein structures.** *Bioinformatics* 2006, **22**(16):2049–2050.
18. Sankararaman S, Sjolander K: **INTREPID - INformation-theoretic TREe traversal for Protein functional site IDentification.** *Bioinformatics* 2008, **24**:2445–2452.
19. Pazos F, Rausell A, Valencia A: **Phylogeny-independent detection of functional residues.** *Bioinformatics* 2006, **22**(12):1440–1448.
20. Finn RD, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38**(suppl 1):D211–D222.
21. Ye K, *et al*: **Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting.** *Bioinformatics* 2008, **24**:18–25.
22. Chakrabarti S, Panchenko A: **Ensemble approach to predict specificity determinants: benchmarking and validation.** *BMC Bioinforma* 2009, **10**(1):207.
23. Kalinina OV, *et al*: **Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families.** *Protein Sci* 2004, **13**(2):443–456.
24. Rodriguez GJ, *et al*: **Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors.** *Proc Natl Acad Sci* 2010, **107**(17):7787–7792.
25. Notredame C, Higgins DG, Heringa J: **T-coffee: a novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302**(1):205–217.
26. del Sol Mesa A, Pazos F, Valencia A: **Automatic Methods for Predicting Functionally Important Residues.** *J Mol Biol* 2003, **326**(4):1289–1302.
27. Kullback S, Leibler R: **On Information and Sufficiency.** *Ann. Math. Statist* 1951, **22**(1):7.
28. Stranzl T, *et al*: **NetCTLpan: pan-specific MHC class I pathway epitope predictions.** *Immunogenetics* 2010, **62**(6):357–368.

doi:10.1186/1471-2105-13-235
Cite this article as: Teppa *et al.*: Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. Implication for catalytic residue prediction. *BMC Bioinformatics* 2012 **13**:235.

W8–W14 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 28 May 2013*
doi:10.1093/nar/gkt427

MISTIC: mutual information server to infer coevolution

Franco L. Simonetti[1], Elin Teppa[1], Ariel Chernomoretz[1,2], Morten Nielsen[3,4] and Cristina Marino Buslje[1,*]

[1]Bioinformatics Unit, Fundación Instituto Leloir, Av. Patricias Argentinas 435, C1405BWE, Buenos Aires, Argentina, [2]Departamento de Física, FCEN, UBA and IFIBA (UBA-CONICET), Pabellón 1, Ciudad Universitaria, 1428 Buenos Aires, Argentina, [3]Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800, Lyngby, Denmark and [4]Instituto de Investigaciones Biotecnológicas, Universidad de San Martín, Martín de Irigoyen 3100 CP 1650, San Martín, Buenos Aires, Argentina

Received January 23, 2013; Revised April 18, 2013; Accepted April 27, 2013

ABSTRACT

MISTIC (mutual information server to infer coevolution) is a web server for graphical representation of the information contained within a MSA (multiple sequence alignment) and a complete analysis tool for Mutual Information networks in protein families. The server outputs a graphical visualization of several information-related quantities using a circos representation. This provides an integrated view of the MSA in terms of (i) the mutual information (MI) between residue pairs, (ii) sequence conservation and (iii) the residue cumulative and proximity MI scores. Further, an interactive interface to explore and characterize the MI network is provided. Several tools are offered for selecting subsets of nodes from the network for visualization. Node coloring can be set to match different attributes, such as conservation, cumulative MI, proximity MI and secondary structure. Finally, a zip file containing all results can be downloaded. The server is available at http://mistic.leloir.org.ar. In summary, MISTIC allows for a comprehensive, compact, visually rich view of the information contained within an MSA in a manner unique to any other publicly available web server. In particular, the use of circos representation of MI networks and the visualization of the cumulative MI and proximity MI concepts is novel.

INTRODUCTION

Multiple sequence alignments (MSA) of homologous proteins carry at least two types of information: one given by the conservation of amino acids at certain positions, and another given by the interrelationship between two or more positions. Mutual Information (MI) from information theory can be used to estimate the extent of the coevolutionary relationship between two positions in a protein family (1–3). Mutual information is therefore often applied to predict positional correlations in an MSA guiding the identification of structurally or functionally important positions in a given protein fold or family. For example, mutations of essential residues in a protein sequence may occur, only if a compensatory mutation takes place elsewhere within the protein to preserve or restore activity (2). However, it should be taken into account that a covariation signal is made up of phylogeny, structure, function, interactions and stochastic components (4), and that high MI values are hence not a proof of coevolution, they are rather suggestive of it. MI values can in particular be misleading if sequences are not collected properly or the underlying sequence alignment not built correctly.

Several servers have been developed to calculate MI including (5–8). The main output from these servers is a plain text file with the results. The servers described in (5) and (6) provide different scoring functions, and the analysis is limited to 800 and 1000 sequences in the MSA, respectively. If a reference structure is available, the methods described in (6) and (7) provide a static image with the coevolving residue pairs highlighted in the structure.

In contrast to these tools, the MISTIC (mutual information server to infer coevolution) server offers an interactive platform to analyze and visualize MI and distance networks, perform network analyses, filtering results by different scores simultaneously at both nodes and edges and different options for graphical representation of the information contained within an MSA. The MI calculation implemented in MISTIC is described in (9). In short, the calculation includes corrections for phylogeny and

*To whom correspondence should be addressed. Tel: +54 11 52387500; Fax: +54 11 52387501; Email: cmb@leloir.org.ar

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

entropy biases, low number of observations and sequence weighting to correct for data redundancy (for details see 'Materials and Methods' section). We will through out the manuscript refer to this MISTIC mutual information value as MI score. Once the MI score is calculated between residues, a network is created where nodes are residues and links between nodes represent a significant coevolutionary signal (9). The results from the server include two sections. On the one hand, a static output page is provided with (i) the information of the MSA condensed into a circos representation (10) (a way of visualizing data in a circular layout), (ii) a MI network and (iii) a distance network if a reference protein structure is supplied. On the other hand, an interactive network interface is given where several nodes, edges and network properties can be displayed and analyzed. Node data visualization includes the amino acid frequency at a given position (if one node is selected) or a Kullback–Leibler (KL) sequence logo providing information about enrichment/depletion of amino acids (11) when several positions are selected. Also, the secondary structure, the Kullback-Leibler conservation score (12), cumulative Mutual Information (cMI) that measures the degree of shared mutual information of a given residue and the proximity Mutual Information (pMI), which tells about the networks of mutual information in the proximity of a residue (within a certain distance threshold) (13) are also available as part of the node information in the MI network. Edges (MI scores between two residues) are selectable on the net and displayed in the edges tab. Several filters can be applied on the nodes and edges to highlight any subnetwork of interest. Nodes can be filtered by conservation, cMI and pMI scores. Edges can be filtered by MI score, spatial distance and the sequential distance between residues. Examples could be selecting the highest scoring N edges, the highest pMI scored nodes, the MI between residues i and i+n (where n is the distance in the sequence from residue i).

In summary, the MISTIC server allows to integrate sequence and structure information contained in an MSA in a comprehensive, compact, visually rich manner that enables the user to extract essential information in terms of networks, conservation and structure for any subset of residues of interest guiding the identification of functionally important residues in a protein. MISTIC is available at http://mistic.leloir.org.ar.

MATERIALS AND METHODS

Corrected MI score

The Mutual Information score implemented in the MISTIC method is calculated between pairs of columns in the MSA as described in (9). Briefly, the frequency for each amino acid pair is calculated using sequence weighting and low count corrections and compared with the expected frequency assuming that mutations between amino acids are uncorrelated. Next, the MI score is calculated as a weighted sum of the log ratios between the observed and expected amino acid pair frequencies. The Average Product Correction (APC) method of Dunn *et al.* (14) is applied to reduce the background mutual information signal for each pair of residues, and the MI scores are finally translated into MI *z*-scores by comparing the MI values for each pair of position with a distribution of prediction scores obtained from a large set of permutated versions of the MSA. In earlier work, we have found that a *z*-score threshold of 6.5 defines a sensitivity of 0.4 and a specificity of 0.95 (9). Based on this work, MISTIC reports every MI value between two residues >6.5.

The server allows analyzing the cMI that defines to what degree a given amino acid takes part in a mutual information network and the pMI that characterizes the mutual information network in the proximity of a given residue. The cMI score for each residue is calculated as the sum of corresponding MI Z-score values (>6.5 threshold) over all residues within the MSA. The pMI is a proximity average calculated for each residue as the average of cMI of all the residues within 5 Å to the given amino acid (13). The distance between each pair of residues in the structure is calculated as the shortest distance between any two atoms (excluding H atoms).

Inputs

The main input file consists of an MSA of protein sequences in FASTA, Nexus, Phylip, PIR or ClustalW format. Users can upload their own MSA or alternatively can be guided to upload the suggested MSA from Pfam database (9) providing a protein sequence, a Uniprot ID or a Pfam accession number (15). Once loaded, the file format is checked, and if correct, additional parts of the submission page become available. A reference sequence can be set by typing a sequence identifier in the input box, which autocompletes based on the sequence IDs from the uploaded MSA. If no sequence is specified, the first sequence in the alignment will be used as reference sequence. Also, a Protein Data Bank (PDB) code can be specified or a PDB file uploaded to allow mapping the information contained in the MSA onto the selected PDB structure. Users can optionally provide an e-mail address and a job description to receive a notification of job completion. Advanced options are available for algorithm parameters modification, such as sequence clustering, low count correction and gap removal.

Outputs

After job completion, the results web page can be accessed through the link sent by e-mail, the bookmarked page or by using the jobID. Several representations of the information contained in the MSA are displayed. First, an MI Circo is displayed (Figure 1 and Supplementary Figure S1), which is a circular representation of the reference sequence mapped with the different information measures calculated by the MISTIC server. The information of each circular track from the outer to inner circles is the following: labels in the first (outer) circle indicate the position in the alignment and the amino acid code of the reference sequence. If a structure file was provided, numbering will refer to the PDB structure. The colored square boxes of the second circle indicate the MSA position

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

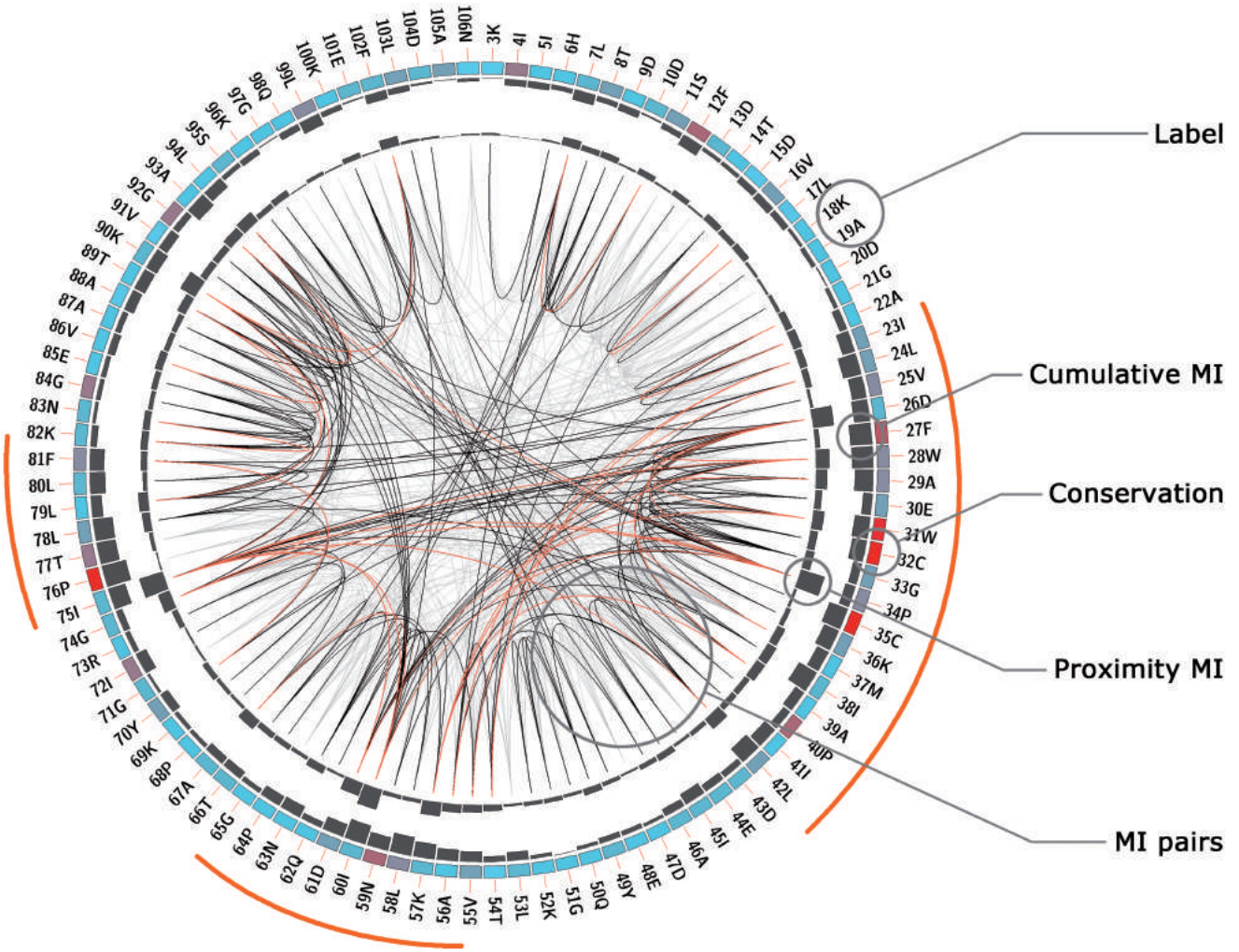


Figure 1. Circos representation of the Thioredoxin protein family (PF00085). Labels in the first (outer) circle indicate the amino acid code and the PDB number of the reference sequence. Colored square boxes of the second circle indicate the KL conservation score (from red to cyan, red: highest; cyan: lowest). The third and fourth circles show the cMI and pMI scores as histograms, facing outward and inward, respectively. Lines in the center of the circle connect pairs of positions with MI score >6.5. Red lines represent the top 5%; black ones are between 70 and 95%, while gray ones account for the last 70%. Orange bars indicate the three regions mentioned in the text.

conservation (highly conserved positions are in red, while less conserved ones are in blue). The third and fourth circles show the cMI and pMI scores (13) as histograms, facing outward and inward, respectively. Lines in the center of the circle connect pairs of positions with MI score >6.5 (9). Red lines represent the top 5% percentile; black lines are MI values between the 70 and 95% percentile, while gray lines account for the remaining MI values.

To complete the description of the analyzed protein family, an MI network and a distance network (if PDB provided) are built (Figure 2 and Supplementary Figure S1). Network graphs are composed of nodes joined by edges. Each node represents a residue, and the edge between two nodes indicates an MI value >6.5 (in the MI network) or a distance <5 Å (in the distance network). Default node coloring represents amino acid conservation for the given residue.

An interactive interface to further explore and characterize the MI network using Cytoscape Web (16) is provided (Figures 3–5) based on Flash and Java plug-ins. Node and edge information is displayed if selected. Clicking each node, amino acid frequencies of the given residue position are displayed, while by clicking several nodes, a sequence logo representation with amino acid enrichment and depletions is shown (Figure 3). Nodes of interest can be selected and shown onto a reference structure if available allowing for mapping of network characteristics onto the protein 3D structure. Structures are displayed using a Jmol Applet (http://www.jmol.org/). Several filtering tools are offered for selecting specific subsets of nodes from the network for visualization. Also, first neighbors and the maximal associated subnetwork of a node can be selected. Selections can be mapped onto the reference structure with different labels, colors and types of representation (structure and

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

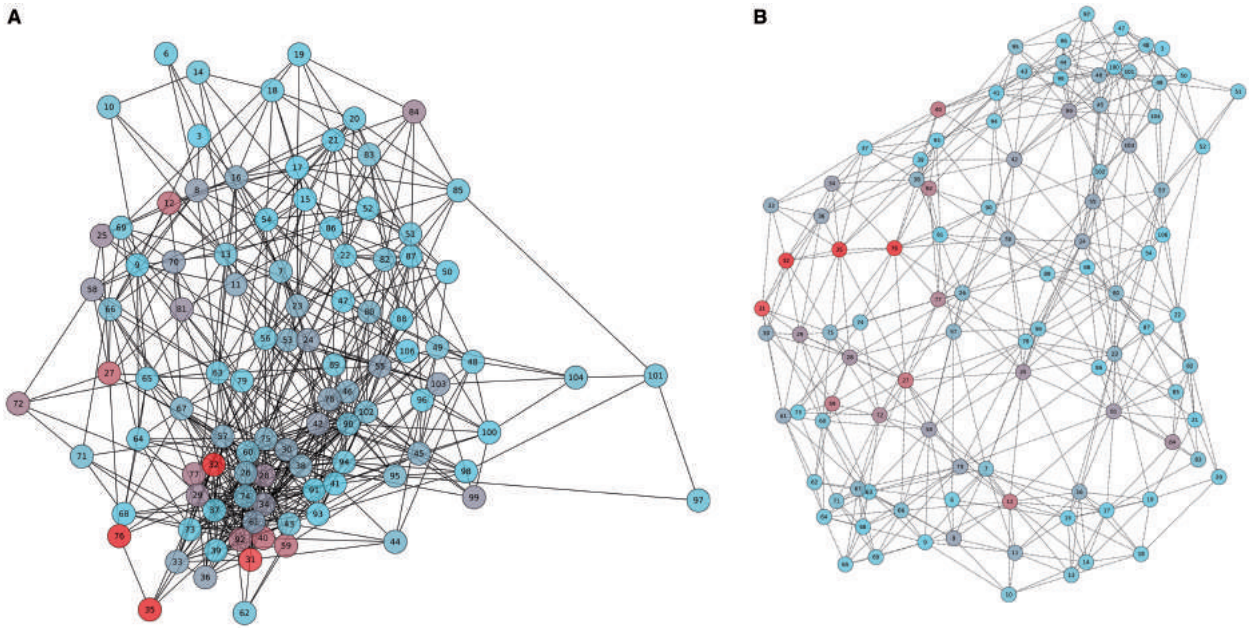


Figure 2. The MI network and distance network of the Thioredoxin protein family (PF00085). Panel (**A**) Mutual information network (PF00085). Panel (**B**) Distance network of the reference structure (PDB code: 2trx). Amino acids are represented as circles colored from red to cyan on conservation (from higher to lower). Edges are represented as lines binding nodes if they have a significant MI value (MI score > 6.5) or are closer than 5 Å (panels A and B, respectively).

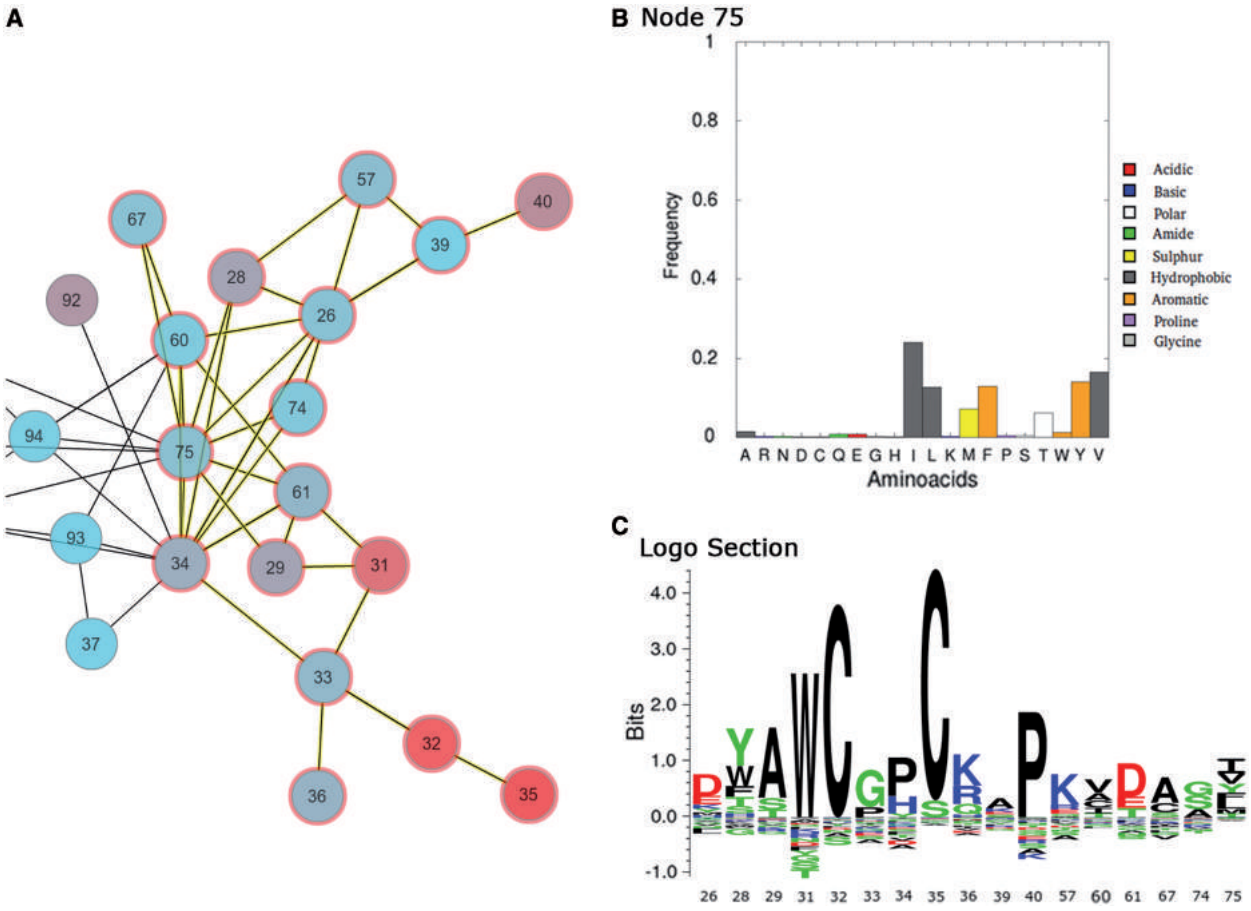


Figure 3. MISTIC interactive framework showing in panel (**A**) part of the PF00085 MI network with several nodes and edges selected (selected circles are highlighted in red and selected edges in yellow). Panel (**B**) shows the amino acid frequency when only one node is selected (e.x: 75). Panel (**C**) shows the KL sequence logo of the (several) selected nodes.

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

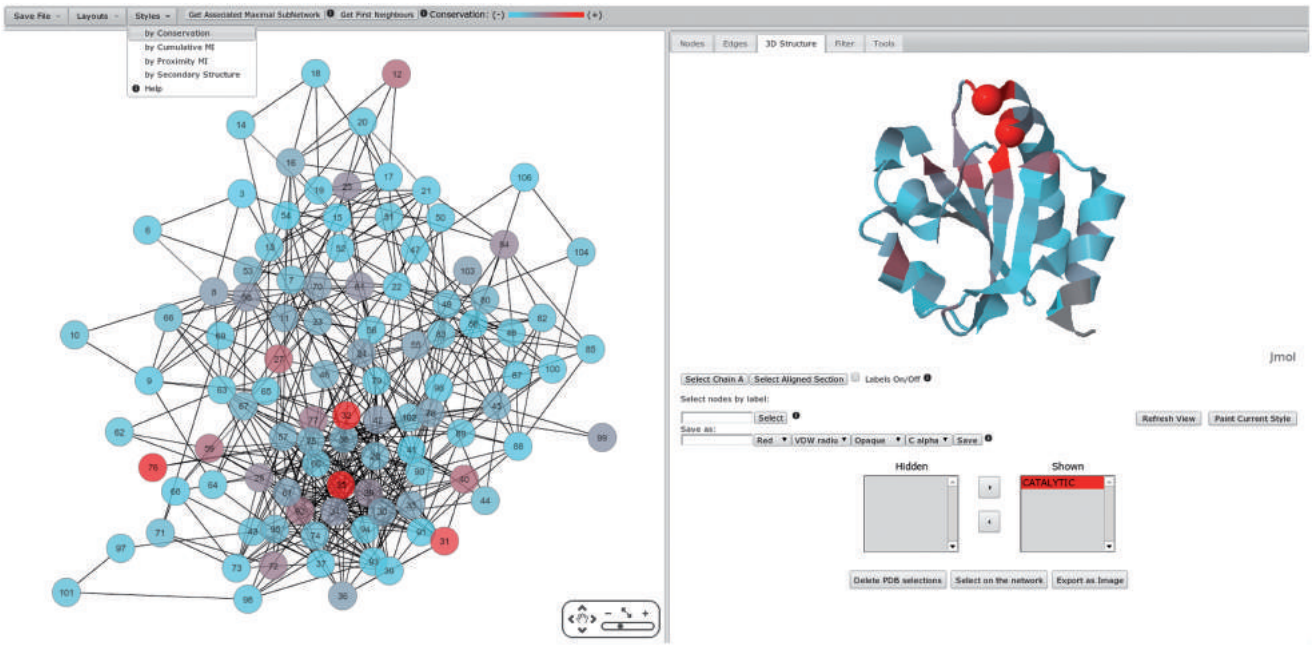


Figure 4. Left: Mutual information network (PF00085) colored from red to cyan on conservation (from higher to lower). Right: ribbon representation of the reference structure (PDB code: 2trx) colored as the network style (conservation). Also, catalytic residues are highlighted with Cα VDW style.

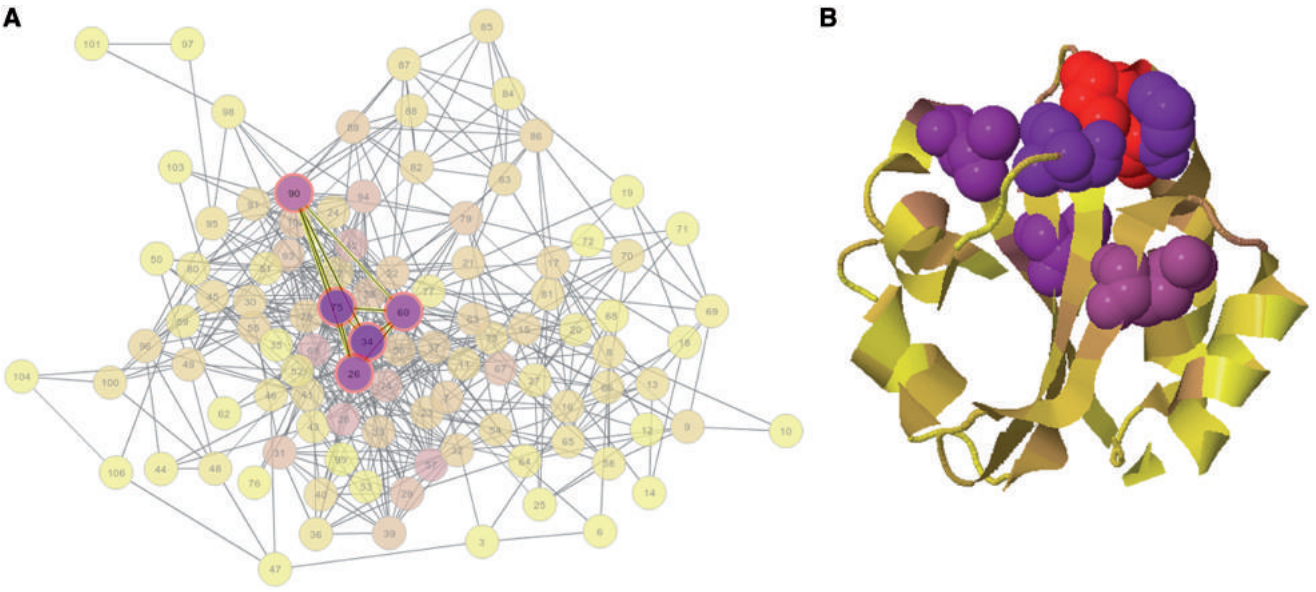


Figure 5. (**A**) MI network colored on residue cMI scores from violet to yellow (higher to lower). Residues with highest cMI were selected by filtering the network (upon cMI). (**B**): ribbon representation of the reference structure (PDB code: 2trx) colored as the network style (cMI score). Selected subnetwork as well as the catalytic residues are highlighted with VDW style.

reference sequence are aligned with the Smith–Waterman algorithm). In addition, node coloring can be set to match different attributes (network style), such as conservation, cMI, pMI and secondary structure (the last two if a reference structure is available). This coloring scheme can automatically be transferred to the structure by clicking 'paint current style' in the structure tab. This allows an easy visualization of the different position attributes onto the 3D structure (Figure 4). Finally, a zip file containing all results, including the MSA, MI score and conservation data in raw format, logos and circos images and network files to load on Cytoscape's desktop version, can be downloaded for further user manipulation.

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

Example of use: The Thioredoxin protein family (PF00085)

The Thioredoxin protein family is used to illustrate some of the server functionalities. Thioredoxins are small enzymes that participate in redox reactions via the reversible oxidation of an active center disulfide bond (residues C32 and C35). They are proteins known to be present in all organisms and play their role in many important biological processes. The MSA of the Thioredoxin family was uploaded from Pfam (Pfam accession PF00085). The full alignment (16 281 sequences, ~100 residues long) is handled by the server with a calculation time of ~430 s. The reference sequence and structure were set as THIO_ECOLI and PDB code 2trx. Figure 1 shows in a circos representation that the most conserved positions are the catalytic residues C32 and C35. Also, it can be observed that information is accumulated in three main regions: residues 23–42, 55–65 and 75–81 (outer histogram pMI) and inner lines (MI connections). Within those regions, individual residues (hubs) with the high cMI values (a large number of MI connections) can be found. Figure 2 shows the complete MI network (panel A) and distance network (panel B). All in all this analysis illustrates the intuitive use of the MISTIC server as a guide to point out functionally important residues in a protein.

Further analyzing the interactive network, Figure 3 displays the frequency of amino acids at a particular position and the KL logo when more than one position is selected. Figure 4 shows the MI network colored by conservation and mapped onto the reference structure with the same scheme of colors, and the catalytic residues C32 and C35 are shown with Cα Van Der Waals radius (VDW) representation. It can be observed how conservation is distributed on the protein structure. It has earlier been demonstrated that residues in the close proximity of catalytic residues are enriched in cMI scores (13). By mapping the MI network by cumulative MI (upper bar: style/by Cumulative MI) and filtering by cMI (in the filters tab), this finding can easily be confirmed. This subnetwork together with the most conserved residues can be displayed on the 3D structure and in such a way the relative location of both kinds of residues (catalytic and rich in cumulative information) can be easily visualized (Figure 5).

The observations made in (9) by calculating the MI score, mapping onto the PDB structure, selecting top-scoring MI pairs and measuring their distance to the catalytic residues can thus be reproduced in few simple steps with MISTIC (Supplementary Figure S2).

DISCUSSION AND CONCLUSIONS

We have developed an interactive web server providing the end users with a highly intuitive view of the information contained within an MSA. The server allows for an in-depth analysis of the evolutionary signal contained within protein families providing the user with a unique view of the interrelationship between the different information signals (conservation, MI, MI networks and physical distance networks) contained within an MSA.

To the best of our knowledge, MISTIC is the only publicly available method that offers an interactive platform to analyze MI and distance networks, perform network analysis by filtering results by different scores simultaneously at nodes and edges, as well as different options for graphical representation of the different information signals.

The server has no restrictions on protein length and number of sequences in the alignment. This is a critical feature of the MISTIC method as the accuracy of the evolutionary analysis (and sequence analysis in general) is highly influenced by the number and divergence of the sequences in the MSA (9,14). Therefore, limiting the analysis to MSAs containing a few hundred sequences in our view will limit the use to data sets where the calculation of the coevolutionary signal is inaccurate, hence making the use of MI score improper.

We hence believe that the functionality of MISTIC is unique, and trust that the server will provide a powerful tool for non-bioinformatics end users to analyze the information signal contained with protein families and guide the search for residues essential for protein function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1 and 2.

FUNDING

F.L.S., E.T., A.C., M.N. and C.M.B. are researchers at the Argentinean national research council (CONICET). CONICET grants [PIP1936 and PIP0087] (in part). Funding for open access charge: [PIP1936] (CONICET).

Conflict of interest statement. None declared.

REFERENCES

1. Gloor,G.B., Martin,L.C., Wahl,L.M. and Dunn,S.D. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, **44**, 7156–7165.
2. Martin,L.C., Gloor,G.B., Dunn,S.D. and Wahl,L.M. (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, **21**, 4116–4124.
3. Tillier,E.R. and Lui,T.W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**, 750–755.
4. Atchley,W.R., Wollenberg,K.R., Fitch,W.M., Terhalle,W. and Dress,A.W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.*, **17**, 164–178.
5. Yip,K., Patel,P., Kim,P.M., Engelman,D.M., McDermott,D. and Gerstein,M. (2008) An integrated system for studying residue coevolution in proteins. *Bioinformatics*, **24**, 290–292.
6. Gouveia-Oliveira,R., Roque,F., Wernersson,R., Sicheritz-Ponten,T., Sackett,P., Molgaard,A. and Pedersen,A. (2009) InterMap3D: predicting and visualizing co-evolving protein residues. *Bioinformatics*, **25**, 1963–1965.
7. Fares,M. and McNally,D. (2006) CAPS: coevolution analysis using protein sequences. *Bioinformatics*, **22**, 2821–2822.

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014

8. Kozma,D., Simon,I. and Tusnády,G.E. (2012) CMWeb: an interactive on-line tool for analysing residue–residue contacts and contact prediction methods. *Nucleic Acids Res.*, **40**, W329–W333.
9. Buslje,C.M., Santos,J., Delfino,J.M. and Nielsen,M. (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, **25**, 1125–1131.
10. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S. and Marra,M. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
11. Thomsen,M. and Nielsen,M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, 25.
12. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*, John Wiley & Sons, Inc.
13. Marino Buslje,C., Teppa,E., Di Doménico,T., Delfino,J.M. and Nielsen,M. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification. *PLoS Comput. Biol.*, **6**, e1000978.
14. Dunn,S.D., Wahl,L.M. and Gloor,G.B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.
15. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
16. Lopes,C., Franz,M., Kazi,F., Donaldson,S., Morris,Q. and Bader,G. (2010) Cytoscape web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.

Downloaded from http://nar.oxfordjournals.org/ by guest on January 29, 2014